Human-AI Interaction for Exploratory Search & Recommender Systems with Application to Cultural Heritage

Benjamin Charles Germain Lee

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2023

Reading Committee:

Daniel S. Weld, Chair

Nicholas Weber

Noah Smith

Program Authorized to Offer Degree:
Computer Science & Engineering

University of Washington

Abstract

Human-AI Interaction for Exploratory Search & Recommender Systems with Application to Cultural Heritage

Benjamin Charles Germain Lee

Chair of the Supervisory Committee:
Professor Emeritus Daniel S. Weld
Computer Science & Engineering

Exploratory search and recommender systems are ubiquitous and central to information navigation. Yet, many pressing challenges remain surrounding the development of robust systems, from producing high-quality data and metadata to answering fundamental questions in human-AI interaction concerning the interactive affordances for search and recommendation. These challenges are exacerbated by 1) the ever-expanding wealth of information to be searched, and 2) the widespread incorporation of increasingly opaque and complex machine learning models into deployed systems. This thesis explores these challenges and investigates how we can improve interaction mechanisms in exploratory search and recommendation. Much of this dissertation adopts the setting of digital cultural heritage collections, where impoverished metadata redoubles challenges of searchability, with implications across disciplines.

This dissertation introduces three primary contributions through publicly deployed systems and datasets. First, we demonstrate how the construction of large-scale cultural heritage datasets using machine learning can answer interdisciplinary questions in library & information science and the humanities (Chapter 2). Second, based on the feedback of users of these cultural heritage datasets, we introduce open faceted search, an extension of faceted search that leverages human-AI interaction affordances to empower users to define their own facets in an open domain fashion (Chapter 3). Third, encountering similar challenges with

the deluge of scientific papers, we explore the question of how to improve recommender systems through human-AI interaction and tackle the broad challenge of advice taking for opaque machine learners (Chapter 4).

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Number | Page

ACKNOWLEDGMENTS

The greatest joy of completing the research detailed in this dissertation is not the work itself but rather the communities I have been fortunate to join as a result. I am forever grateful for the relationships and friendships that have grown out of this document. Accordingly, I would like to thank so many people who have made this work – and my journey – possible.

First, I thank my advisor, Dan Weld, for a truly wonderful five years. I began my Ph.D. as an outsider to the world of computer science, and I cannot thank Dan enough for shepherding me through a half-decade intellectual and professional journey within this world. I am, of course, deeply grateful for everything that Dan has taught me from an academic perspective: selecting interesting research questions (a skill that I contend Dan is uniquely gifted with), developing experiments, managing projects, and articulating contributions. More importantly, I will always be grateful for Dan's generosity, including his guidance and patience in helping me to find my way in computer science, thoughtfulness in incorporating my interdisciplinary interests, and joy that he has brought to our work together. This dissertation is first and foremost a celebration of this.

I have been truly fortunate to work so closely with colleagues at the Library of Congress during my time as Innovator in Residence and beyond. I thank my close friends Jaime Mears and Eileen Jakeway in LC Labs for their collaboration over many years surrounding Newspaper Navigator – without their expertise and thoughtfulness, the project would never have been possible. It has been wonderful to collaborate with Trevor Owens surrounding web archives, and I have learned so much through our work together, which I very much look forward to continuing. In LC Labs and Digital Strategy, I would like to thank Laurie Allen, Meghan Ferriter, Leah Weinryb Grohsgal, Abbey Potter, Kate Zwaard, and Brian Foo. In Serials, I thank Robin Butterhof, Nathan Yarasavage, and Deb Thomas. In IT Design &

Development, I thank Chris Adams, John Foley, and Tong Wang. In Web Archives, I thank Abbie Grotke. I also thank Deputy Librarian of Congress, Mark Sweeney. Lastly, I thank the entire National Digital Newspaper Program staff at the Library of Congress, as well as Molly Hardy at the National Endowment for the Humanities, whose expertise I am so grateful to be able to continue to draw from on a regular basis.

At the Allen Institute for Artificial Intelligence, I thank Doug Downey and Kyle Lo for their continued collaboration surrounding LIMEADE and beyond. I have learned so much working with them and am so fortunate to be a part of the Semantic Scholar community, which I look forward to continue being a part of as I transition into my new role at the University of Washington. I also thank Shannon Shen for his friendship and willingness to collaborate on all things layout-related, which I very much look forward to continuing.

In the Spring of 2020, I was incredibly lucky to join a zoom call with Jim Casey, Joshua Ortiz Baco, and Sarah Salter related to digital newspapers. Little did I know that this zoom call would develop into a three-year research collaboration that is in many senses still beginning. I am so thankful to count all three as not only close colleagues but dear friends, and I look forward to continuing to work together for many years to come.

I would like to thank my committee members for their input and guidance throughout my Ph.D.: Noah Smith, James Fogarty, and Dan Weld of the Paul G. Allen School for Computer Science & Engineering at the University of Washington, Nic Weber in the Information School at the University of Washington, and Marti Hearst of UC Berkeley School of Information.

Within the Paul G. Allen School for Computer Science & Engineering, I would like to thank all of my friends in the Lab for Human-AI Interaction: Gagan Bansal, Sherry Wu, Jim Chen, Marissa Redensky, Raymond Fok, and Madeleine Grunde-McLaughin. I thank Kurtis Heimerl and Katharina Reinecke for their courses, "Computing for Social Good" and "Computing Ethics," respectively – in both of these courses, I developed work that appears in this document, and I am grateful for their thoughtful feedback and guidance. I would like to thank Elise deGoede Dorough for all of her advice, guidance, and kindness surrounding navigating my Ph.D. Across campus, I would like to thank Anna Preus and Aashna Sheth –

our work together has been a highlight of my Ph.D. I would also like to thank the students in my graduate seminar, "Computing Cultural Heritage," whose enthusiasm for the subject material continues to be a source of motivation for me. I thank them for everything I have learned from them.

At the Stroum Center for Jewish Studies at the University of Washington, I would like to thank Devin Naar, Makena Mezistrano, Sarah Zaides Rosen, Noam Pianko, and the 2020-21 Graduate Fellows cohort. It has been a true joy to collaborate with the Center on work surrounding the Sephardic Studies Digital Library, and I am grateful to be able to be an active member of the community in the coming years.

I would also like to thank additional external collaborators whom I have been fortunate enough to work with over the years and continue to work with moving forward: the editorial team at *Digital Humanities Quarterly*, Robert Ehrenreich, Michael Haley Goldman, Andrew Kloes, Gabe Pizzorno, Ilene Berson, Michael Berson, Paul Fyfe, Thomas Smits, Melvin Wevers, Katherine McDonough, Daniel Wilson, Kaspar Beelen, S.J. Klein, and Nathan Robinson.

I am grateful to have been supported by a National Science Foundation Graduate Research Fellowship under Grant DGE-1762114, as well as the Library of Congress Innovator-in-Residence Position and the Richard and Ina Willner Memorial Fellowship from the Stroum Center in Jewish Studies at the University of Washington. The work in this dissertation would not have been possible without the support of the University of Washington WRF/Cable Professorship, the Allen Institute for Artificial Intelligence (AI2), Office of Naval Research grant N00014-18-1-2193, and National Science Foundation RAPID grant 2040196.

On a personal note, I am beyond fortunate to have such close friends who enrich my life every day. In CSE, I am grateful to Ashrujit Ghoshal, Galen Weld, Joe Janizek, Johan Michelove, and Timothy Akintilo in particular for their friendship as we have navigated grad school and beyond. I would also like to thank Andrew White, Arjun Byju, David Cromwell, Eli Lee, Grant Ringelman, Greg Parker, Johan Michelove, Luke Ross, Patrick Rollo, Saahil

Mehta, Stephen Portillo, Val Bolotnyy, Vishnu Razdan – I am forever grateful to count you as friends.

Lastly, I thank my family. How do I begin to thank my parents for their love and support throughout my life? Let this dissertation be a testament to our family's love. Thank you for always being there for me, and for always reminding me of the purpose of life: the people with whom we surround ourselves. Though my Zeyda, Sumner Joseph Paul Germain, and Oma, Rita Schorr-Germain, both passed away before my completion of this dissertation, I am confident that the completion of this document means just as much to them as it does to me. I have dedicated the dissertation to my Oma. It was in pursuit of my family history with her that first brought me to the archive. In many senses, I have never left. This dissertation is the fruit of the life and family she cultivated, and I owe my professional path and personal journey to her. I carry her love with me always and will strive to be a *mensch*.

DEDICATION

To Rita Schorr-Germain

"Though I lack the art
to decipher it,
no doubt the next chapter
in my book of transformations
is already written.
I am not done with my changes."
– Stanley Kunitz, *The Layers*

Chapter 1

INTRODUCTION

1.1 Motivation

Search and recommender systems are ubiquitous to the online experience, heavily mediating the information that we encounter on a daily basis. Consider, for example, how we navigate the World Wide Web, find new music, discover new videos, identify new restaurants, retrieve new academic papers, and browse library collections. Though many users of such systems have well-defined search needs with specific end destinations in mind, there are just as many users with entirely different search needs. For example, a user might desire to explore and make sense of a collection of information without a clear end goal, or the user may be unable to formulate an initial query to quickly reach their desired end destination [137]. *Exploratory search and recommendation* comprise this range of settings. Much research across computer science and library & information science (LIS) over the past quarter century has focused on problems pertaining to exploratory search and recommendation, from improving the underlying machine learning algorithms to designing new user interfaces and affordances in order to facilitate search and discovery.

This dissertation examines these challenges through the lens of *human-AI interaction*. Situated at the intersection of artificial intelligence and human-computer interaction, human-AI interaction foregrounds a macroscopic view of the broader human-AI system in lieu of a focus on the AI system in isolation [12]. By adopting this holistic framing and evaluating systems with real users, human-AI interaction as an area of research seeks to improve interaction mechanisms toward a human's desired goals. Human-AI interaction is an essential consideration for a wide range of domains and tasks, from medical diagnosis [225] to model debugging [184].

Significantly, exploratory search and recommendation fit squarely in this paradigm of human-AI interaction because end-users iteratively interact with machine learning-powered search and recommender systems to retrieve relevant results – by design, these systems are intended to guide users to content of interest and thus necessitate user-centric consideration and evaluation. Significantly, modes of interaction remain limited in deployed search and recommender systems, with few affordances for high-level advice or facet specification. The importance of human-AI interaction is redoubled in this context by two considerations. First, search & recommendation as tasks will only become more important as the wealth of digital information continues to grow rapidly. Second, given that search and recommender systems increasingly rely on opaque machine learning models to generate results, understanding human-AI interaction patterns and developing new modes of interaction will remain an important area of focus.

Much of this dissertation adopts the setting of *digital cultural heritage*, namely, digital collections produced by libraries, archives, and museums. Because cultural heritage collections are regularly navigated by wide audiences with different needs and expectations, researchers in exploratory search have historically utilized these collections to implement and evaluate systems [408]. Notably, search and discovery in this context is particularly challenging, as metadata are often impoverished due to limitations of digitization pipelines. Such limitations include unreliable optical character recognition, poor scan quality, lossy intermediate steps such as microfilming, degraded physical artifacts themselves, and large size, both in cardinality and bits. In this context, just creating datasets that are functional for a range of end-users is challenging in its own right. Digital cultural heritage collections are thus an ideal setting for studying new modes of human-AI interaction within exploratory search, where improvements have the capacity to unlock these collections to scholars and the public alike. Along these lines, this thesis pays particular attention to addressing the real needs of historians, humanists, educators, students, and genealogists who utilize digital cultural heritage collections by incorporating a series of multidisciplinary collaborations with these end-users in mind.

1.2 Contributions

Research on cultural heritage is necessarily inter-disciplinary. Motivated by these real-world challenges in exploratory search & recommendation, human-AI interaction, and cultural

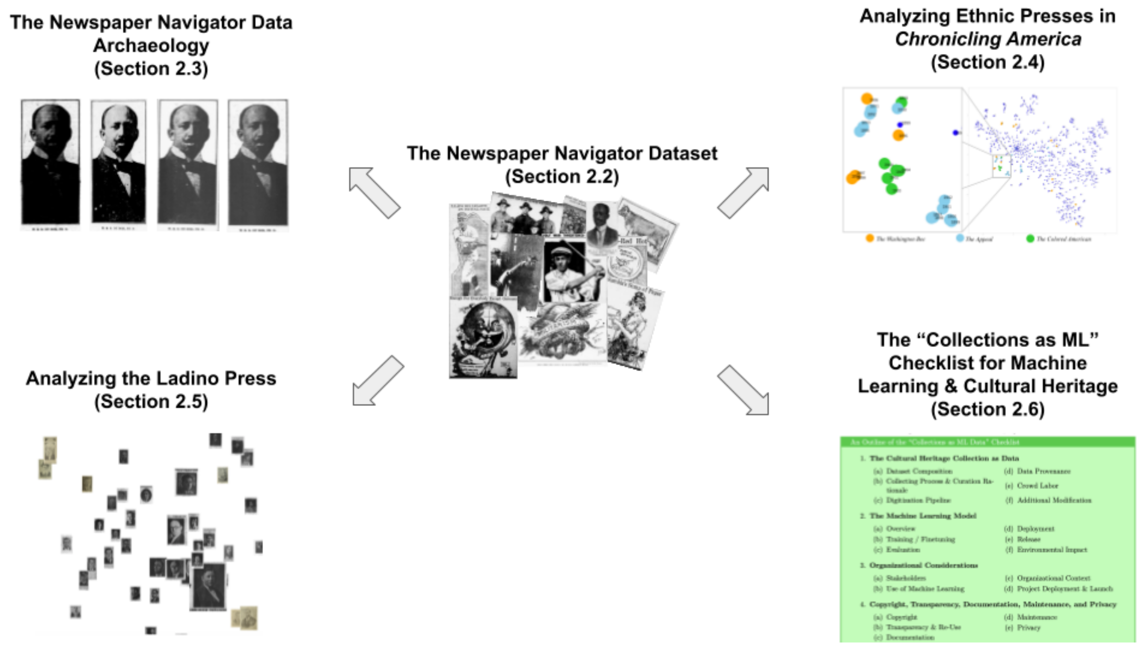

Figure 1.1: An overview of the contributions of Chapter 2. In particular, we demonstrate how the construction of the Newspaper Navigator dataset (Section 2.2) enables multidiscplinary contributions including the Newspaper Navigator data archaeology (Section 2.3), novel analysis of the page layouts of ethnic presses in *Chronicling America* (Section 2.4) and of the visual culture within the Ladino press (Section 2.5), and the construction of the "Collections as ML Data" checklist for machine learning and cultural heritage (Section 2.6).

heritage, this dissertation introduces three primary contributions.

1.2.1 Contribution 1

We demonstrate how the construction of large-scale cultural heritage datasets using machine learning offer answers to interdisciplinary questions in library & information science and the humanities (Chapter 2). Large-scale cultural heritage collections raise a wealth of questions across disciplines, including library & information science and the humanities. For example, how do scholars search these collections effectively? How do existing search affordances mediate what scholars discover? To what extent can emerging machine learning methodologies applied to these collections answer downstream humanities questions? Significantly, answering these questions meaningfully requires interdisciplinary approaches that unify computational methodologies with humanistic inquiry.

In this chapter, we examine the 16+ million digitized historic newspapers within the *Chronicling America* collection as a case study for reimagining discoverability and access for digital collections using an interdisciplinary approach grounded in machine learning. We begin by introducing the Newspaper Navigator dataset comprising photographs, illustrations, comics, editorial cartoons, maps, headlines, and advertisements within these 16 million newspaper pages in *Chronicling America* [202]. We detail the finetuning of a visual content model for this task and present the pipeline utilized to process all ̃100 terabytes of data. Moreover, we describe how the publicly released dataset can be accessed and queried.[1]

We then turn to a series of studies that utilize the Newspaper Navigator dataset to answer interdisciplinary questions in library & information science and the humanities (an overview can be found in Figure 1.1). First, we introduce a data archaeology of Newspaper Navigator: an autoethnographic reflection surrounding the construction of the dataset with a particular focus on the dataset's sociotechnical implications. In particular, this document investigates the many ways that a newspaper is transmuted, mediated, and decontextualized in its journey from a physical artifact to a series of probabilistic photographs, illustrations, and beyond in the Newspaper Navgiator dataset. Second, we detail a multidisciplinary collaboration to study the editorial practices within multiethnic newspapers. In this collaboration, we center our attention on page layout, as inferred from the dataset, as a method for determining similarity between multiethnic newspaper titles as a new approach to excavate editorial decisions and practices across communities. Third, we describe an effort to perform the first macroscopic analysis of the Ladino press using the Newspaper Navigator visual content recognition model. We consider the Sephardic Studies Digital Library at the University of Washington, one of the world's largest collections of Ladino material. Because Ladino suffers from poor optical character recognition (itself a form of algorithmic marginalization), and scholars are restricted to close reading and manual search methodologies, we offer this work surrounding visual culture in the Ladino press as a path forward for improving discoverability. Fourth, we introduce the "Collections as ML Data" checklist,

[1] The full dataset can be accessed at `https://news-navigator.labs.loc.gov`.

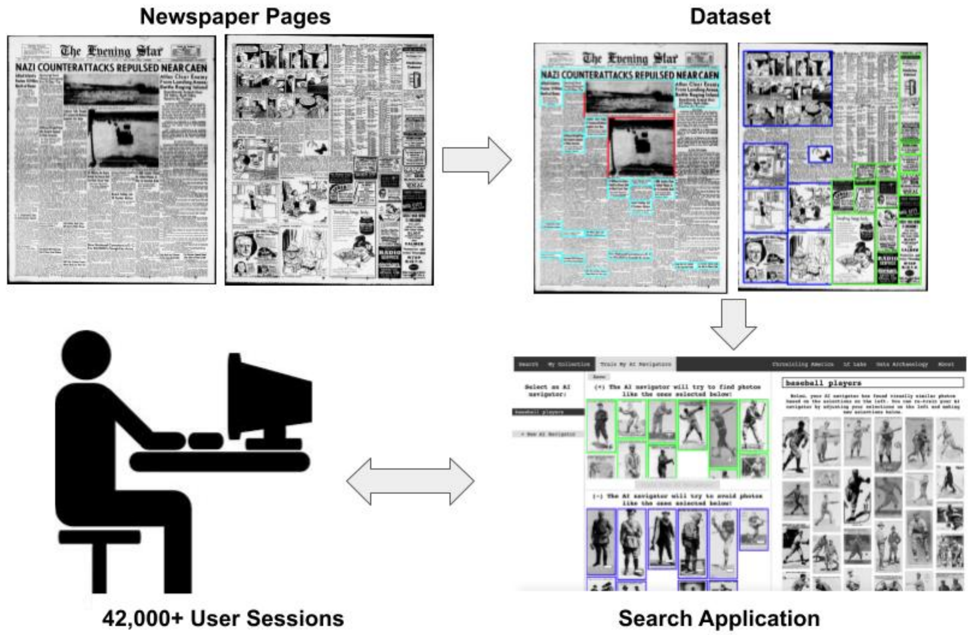

Figure 1.2: An overview of our project, Newspaper Navigator. The project begins with digitized newspaper pages from the *Chronicling America* collection. In the first step (Chapter 2), we extract visual content from the 16+ million pages using a finetuned object detection model, resulting in the Newspaper Navigator dataset. In the second step (Chapter 3), we introduce an open faceted search interface from 1.5+ million photos in the Newspaper Navigator dataset, empowering users to define and refine their own facets in an open domain fashion via interactive machine learning. In the third step, we evaluate the search application's 42,000+ user sessions and identify impact across user groups including scholars, teachers and students in the classroom, genealogists, and beyond.

a curated list of guidelines and considerations for researchers engaging in projects at the intersection of machine learning and cultural heritage, as inspired by the construction of the Newspaper Navigator dataset. We justify the checklist components by drawing from existing checklists and best practices from both machine learning and cultural heritage, as well as surveying and iteratively refining the checklist using a series of relevant projects in this space. We conclude the chapter by emphasizing the importance of interdisciplinary approaches to digital collections, with a particular emphasis on how the further development of such approaches is essential to improving discoverability.

1.2.2 Contribution 2

Based on the feedback of users of these humanities datasets, we introduce *open faceted search*, an extension of faceted search that leverages human-AI interaction affordances to empower users to define their own facets in an open domain fashion (Chapter 3). While the dataset approach presented in Chapter 2 provides one mode of facilitating search and discovery for digital cultural heritage, feedback from users of the Newspaper Navigator dataset made it clear that improvements to the search affordances themselves was a necessary step in order to improve discoverability. In particular, users had specific visual facets that they wished to search for, with no single taxonomy comprehensive enough to cover them all. Even with a faceted search interface, these users would face limitations surrounding discoverability due to the inevitable incompleteness of the taxonomy. In response to this limitation of requiring a facet taxonomy to be pre-defined, we introduce open faceted search as a framework.

We instantiate open faceted search with the Newspaper Navigator search application, a publicly-deployed search interface for 1.5+ million photographs in the Newspaper Navigator dataset.[2] In addition to supporting keyword search and basic faceted metadata search over the photographs, the application supports an affordance called "Train My AI Navigators," which surfaces an interactive machine learning interface and enables a user to iteratively train an AI navigator (facet learner) to retrieve photographs pertaining to an open facet by defining positive and negative examples. This affordance is highly responsive, performing the machine learning inference and re-ranking in under a second. In Figure 1.2, we provide an overview for the full Newspaper Navigator project, beginning with raw newspaper pages in the *Chronicling America* collection and resulting in over 42,000 user sessions for the Newspaper Navigator search application. User groups include a wide range of communities, such as scholars, teachers and students in the classroom, and genealogists. In this regard, the project is an ideal testbed for open faceted search, with organic users who have real exploratory search needs.

---

[2]The Newspaper Navigator search application can be accessed at: `https://news-navigator.labs.loc.gov/search`

To evaluate open faceted search, we analyze the logs of these 42,000 user sessions of the Newspaper Navigator search application. By studying hundreds of named and trained open facets, we demonstrate that canonical image recognition taxonomies do not support the diversity of facets desired by real users of the search application, in turn motivating the importance of open facets. By evaluating 54 training curves, we show evidence of facet learning for many of these open facets. In response to the open facets that do not show evidence of learning, we demonstrate the capacity for greatly improved learning with state-of-the-art multimodal CLIP embeddings. We then turn to ongoing work surrounding the development of *zero-metadata open faceted search* for image collections, an extension of open faceted search that facilitates the interactive bootstrapping of faceted search interfaces using large language models, even in the absence of any descriptive metadata for the images. We present an initial system that supports zero-metadata open faceted search for large-scale image collections with tens of facets. We conclude by articulating next steps and future work for open faceted search as a general framework.

1.2.3 Contribution 3

Encountering similar challenges with the deluge of scientific papers, we explore the question of how to improve canonical recommender systems with human-AI interaction, in particular, through the broad challenge of advice taking (Chapter 4). Whether in the context of searching scientific papers or debugging an image classification model, the problem of *advice taking* – updating an AI via high-level human feedback – is a ubiquitous challenge. While opaque models ranging from neural networks to boosted decision forests have seen widespread adoption in operational contexts, and methods of generating post-hoc explanations for decisions made by opaque models have been developed, the problem of advice taking for opaque models remains understudied. In this chapter, we ask two primary questions:

1. Can one translate high-level human advice into a correction to an arbitrary, opaque, machine-learned model which uses a different set of features than those used to express the advice?

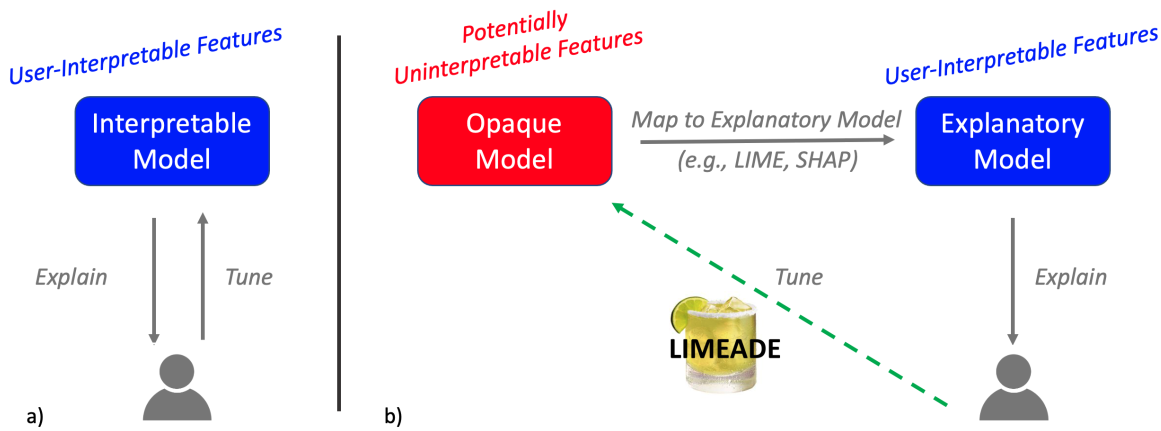

Figure 1.3: An interpretable model, such as a GA^{2}M (left), is by definition explainable and tunable. With an opaque model (right), methods such as LIME and SHAP enable the user to receive an approximate explanation, often using a new vocabulary (e.g superpixels instead of pixels). LIMEADE allows users to provide feedback — using features of the explanatory model — and then modifies the original, opaque model by retraining.

2. Do these methods allow end-users to improve the accuracy of natural, real-world models more easily than by simply annotating more instances?

To answer these questions, we introduce LIMEADE, a new framework for AI advice taking extensible to general opaque machine learners. LIMEADE considers post-hoc explanatory methods such as LIME [320] and SHAP [226], which generate explanations consisting of feature weights for high-level, interpretable features. As shown in Figure 1.3, LIMEADE completes the feedback loop by enabling the user to provide feedback using these high-level features and translating this feedback into an update to the opaque model. This update is accomplished through the construction of pseudo-instances. LIMEADE is (1) a general, model-agnostic framework, (2) applicable to a wide range of domains and tasks, and (3) capable of accepting different types of high-level advice.

We evaluate LIMEADE by considering two real-world settings: image classification debugging and text recommendation. For image classification debugging, we evaluate the strength of a LIMEADE update in comparison to a baseline update of labeling more examples. Comparing the performance of LIMEADE to the performance of the baseline for 20

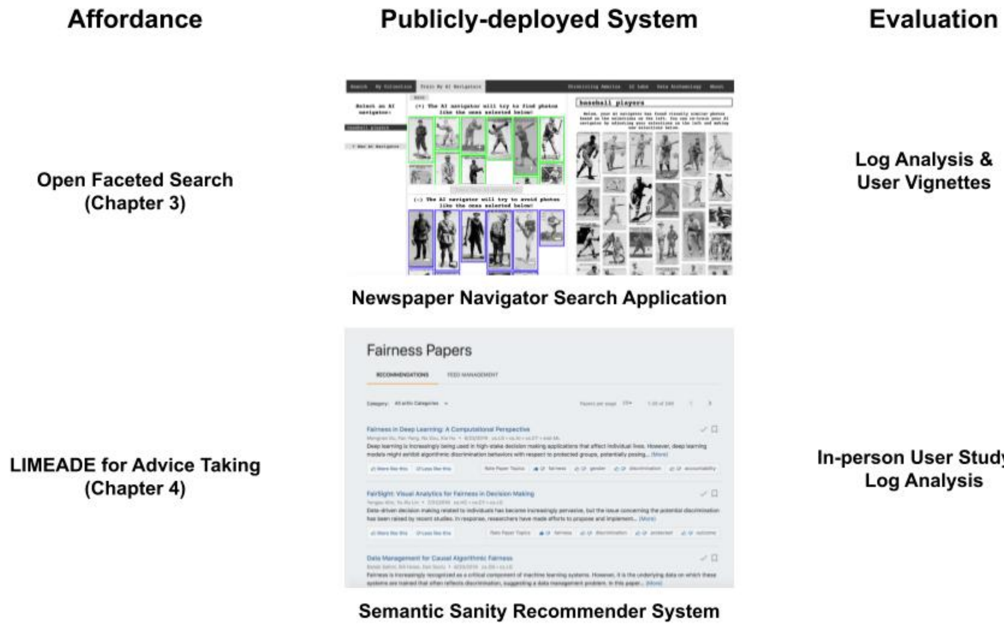

Figure 1.4: A roadmap for evaluating human-AI interaction affordances presented in this dissertation using publicly deployed systems. While additional approaches are utilized to evaluate these affordances, we highlight publicly-deployed systems as an important contribution of this dissertation.

different object classes in the few-shot setting, we find that LIMEADE significantly improves accuracy.

For the text modality, we apply LIMEADE to a neural recommender system for scientific papers on a public website. In the context of this recommender system, LIMEADE enables users to specify advice in terms of unigrams and bigrams that are surfaced by an interpretable explanatory model approximating the opaque neural recommender. As shown in Figure 1.4, this builds the work in Chapter 3 of this dissertation by implementing new affordances in publicly-deployed systems with organic users. In an in-person user study of this recommender system, we demonstrate that LIMEADE leads to significantly higher perceived user control, trust, and satisfaction. However, we do not show evidence of improved accuracy in relation to the baseline adopted in our study. In this regard, our evaluations clearly demonstrate the generality of LIMEADE as a framework, thus answering question 1, but more work remains to be done surrounding question 2.

Our work in the text modality also uncovers a phenomenon that we call the explanation-action tradeoff: a tension between explanation quality and advice diversity caused by user interface limitations. When a user is restricted to providing advice with a small number of explanation terms, we observe a clear feedback loop in which giving more advice causes explanation diversity to decrease, further restricting the number of terms with which a user can specify advice.

We conclude this chapter with a "Call to Action" surrounding the problem of advice taking and articulate a series of future directions to answer questions raised throughout our work.

1.3 Prior Publications and Authorship

Though I am the primary author of the research appearing in this dissertation, my research is the result of years of collaboration with my advisor, Daniel S. Weld, as well as collaborators at the Allen Institute for Artificial Intelligence, the Library of Congress, and beyond. Here, I summarize the research included in this dissertation that has been previously published, along with my co-authors surrounding this work. The Newspaper Navigator dataset (Section 2.2) is in collaboration with Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld, and is based on a publication that appeared in CIKM 2020 [202]. The Newspaper Navigator data archaeology (Section 2.3) is based on a single-author publication that appeared in *Digital Humanities Quarterly* [196]. The work surrounding Newspaper Navigator and multi-ethnic periodicals (Section 2.4) is in collaboration with Joshua Ortiz Baco, Sarah H. Salter, and Jim Casey, and is based on a publication that appeared in *Computational Humanities Research* (CHR) Conference 2021 [199]. The work surrounding Newspaper Navigator and the Ladino press (Section 2.5) is based upon a single-author chapter that appeared in the book *Jewish Studies in the Digital Age*, published by De Gruyter Press [197]. The "Collections as ML Data" Checklist (Section 2.6) is based upon a single-author publication that forthcoming in the *Journal of the Association for Information Science and Technology* (JASIST) Special Issue: "Conceptual Models of the Sociotechnical" [193]. The Newspaper Navigator search application and the initial work surrounding open faceted search (Section 3) is in collaboration with Daniel S. Weld and is partially based on a publication that appeared as a demo at UIST 2020 [205]. LIMEADE (Chapter 4) is in collaboration with Doug Downey, Kyle Lo, and Daniel S. Weld, and is based on a publication that forthcoming in the *ACM Transactions on Interactive Intelligent Systems* (TiiS) Special Issue: "Human-centered Explainable AI" [201]. To reflect my collaborators' contributions in multi-author works, I use the first-person plural in the appropriate places in these chapters.

Chapter 2

COMPUTING CULTURAL HERITAGE: INTERDISCIPLINARY QUESTIONS SURROUNDING SEARCH & DISCOVERY

2.1 Introduction

This chapter investigates how the construction of large-scale humanities datasets using machine learning can inform research questions across disciplines, from library & information science to history to Jewish studies. In this chapter, we introduce our project, Newspaper Navigator, as a case study for improving discoverability and access. We begin by introducing our publicly-released Newspaper Navigator dataset, consisting of extracted visual content including photographs, illustrations, comics, editorial cartoons, maps, headlines, and advertisements from the 16+ million pages in the *Chronicling America* collection of digitized historic American newspapers. We then introduce a number of questions from library & information science and the humanities that can be answered with the dataset. A road map for the chapter is presented below:

1. Section 2.2 introduces the Newspaper Navigator dataset.

2. Section 2.3 presents the Newspaper Navigator data archaeology, an autoethnography surrounding the construction of the dataset with a particular focus on the dataset's sociotechnical implications.

3. Section 2.4 details a multidisciplinary collaboration with print scholars to study the editorial practices embedded within multiethnic newspapers using page layouts derived from the Newspaper Navigator dataset.

4. Section 2.5 presents an effort to perform the first macroscopic analysis of the Ladino press using the Newspaper Navigator visual content recognition model.

5. Section 2.6 introduces the "Collections as ML Data" checklist, a checklist for researchers engaging in projects at the intersection of machine learning and cultural

heritage, as inspired by the guidelines and considerations undertaken while constructing the dataset.

6. Section 2.7 concludes the chapter and articulates future directions for this line of research.

2.2 The Newspaper Navigator Dataset: Extracting Headlines and Visual Content from 16 Million Historic Newspaper Pages in Chronicling America

Chronicling America is a product of the National Digital Newspaper Program, a partnership between the Library of Congress and the National Endowment for the Humanities to digitize historic newspapers. Over 16 million pages of historic American newspapers have been digitized for Chronicling America to date, complete with high-resolution images and machine-readable METS/ALTO OCR. Unfortunately, Chronicling America only supports page-level browsing, without the capacity to browse at the level of headlines or isolated visual content – a feature of considerable interest to its users.To provide this functionality, we introduce a novel machine learning model trained on bounding box annotations to accomplish this task (Section 2.2.5). We describe our pipeline that utilizes this deep learning model to extract 7 classes of visual content: headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements, complete with textual content such as captions derived from the METS/ALTO OCR, as well as image embeddings for fast image similarity querying. We report the results of running the pipeline on 16.3 million pages from the Chronicling America corpus and describe the resulting Newspaper Navigator dataset, the largest dataset of extracted visual content from historic newspapers ever produced. The Newspaper Navigator dataset, finetuned visual content recognition model, and all source code are placed in the public domain for unrestricted re-use.

This work was done in collaboration with Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, Daniel S. Weld, and is based on a publication that appeared in CIKM 2020 [202].

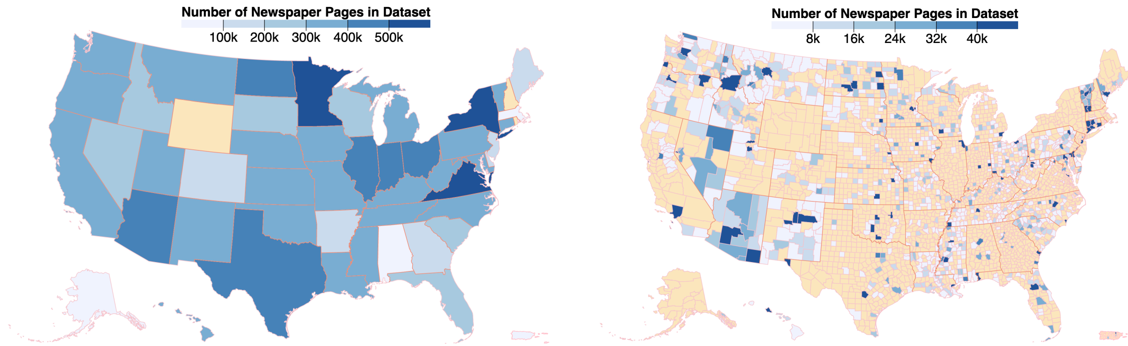

Figure 2.1: Choropleth maps at the state and county level showing the geographic coverage of the 16.3 million *Chronicling America* historic newspaper pages included in the Newspaper Navigator dataset. Yellow coloring indicates that no pages cover the corresponding region. Puerto Rico is pictured in the bottom-right of each map.

2.2.1 Introduction

Chronicling America, an initiative of the National Digital Newspaper Program - itself a partnership of the Library of Congress and the National Endowment for the Humanities - is an invaluable resource for academic, local, and public historians; educators and students; genealogists; journalists; and members of the public to explore American history through the uniquely rich content preserved in historic local newspapers. Over 16 million pages of newspapers published between 1789 to 1963 are publicly available online through a search portal and public API. Among the page-level data are 400 DPI images, as well as METS/ALTO OCR, a standard maintained by the Library of Congress that includes text localization [301].

The 16.3 million *Chronicling America* pages included in the Newspaper Navigator cover 174 years of American history, inclusive of 47 states, Washington, D.C., and Puerto Rico. In Figure 2.1, we show choropleth maps displaying the geographic coverage of the 16.3 million *Chronicling America* newspaper pages included in the Newspaper Navigator dataset. In Figure 2.2, we show the temporal coverage of these pages. The coverage reflects the selection process for determining which newspapers to include in *Chronicling America* [69, 298]. The selection process should be considered in the methodology of any research performed using the Newspaper Navigator dataset.

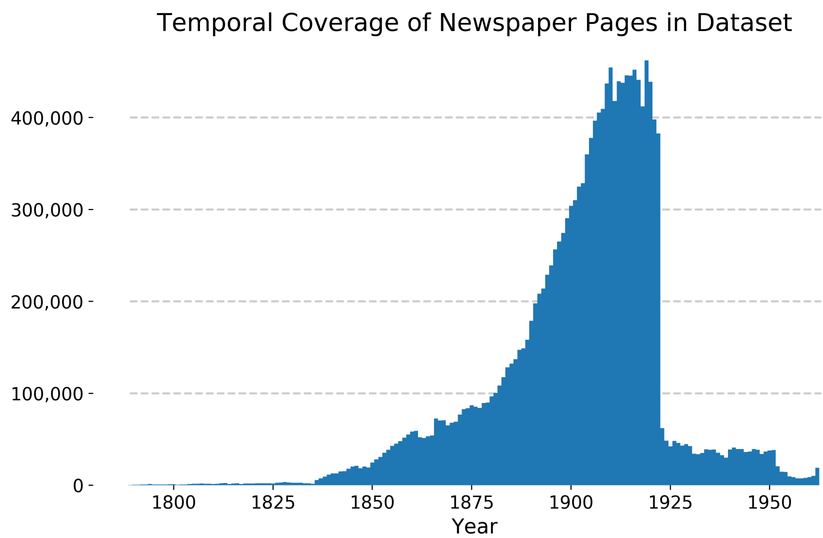

Figure 2.2: A histogram showing the temporal coverage of the 16.3 million *Chronicling America* historic newspaper pages included in the Newspaper Navigator dataset.

While the images and OCR in *Chronicling America* provide a wealth of information, users interested in extracted visual content, including headlines, are currently restricted to general keyword searches or manual searches over individual pages in *Chronicling America*. For example, staff at the Library of Congress have produced a collection of Civil War maps in historic newspapers to date, but the collection is far from complete due to the difficulty of manually searching over the hundreds of thousands of *Chronicling America* pages from 1861 to 1865 [253]. A complete dataset would be of immense value to historians of the Civil War. Likewise, collecting all of the comic strips from newspapers published in the early 20th century would provide researchers with a corpus of unprecedented scale. In addition, users currently have no reliable method of determining what disambiguated articles appear on each page, presenting challenges for natural language processing (NLP) approaches to studying the corpus. A dataset of extracted headlines not only gives researchers insight into the individual articles that appear on each page but also enables users to ask questions such as, "Which news topics appeared above the fold versus below the fold in which newspapers?" Indeed, the digital humanities questions that could be asked with such a dataset abound.

And yet, the possibilities extend beyond the digital humanities to include public history, creative computing, educational use within the classroom, and public engagement with the Library of Congress's collections.

To engage the American public and begin the construction of datasets of visual content within *Chronicling America*, the Library of Congress Labs launched the Beyond Words crowdsourcing initiative in 2017.[1] Volunteers were asked to draw bounding boxes around photographs, illustrations, comics, editorial cartoons, and maps in World War 1-era *Chronicling America* newspapers; they were also asked to transcribe captions by correcting the OCR within each bounding box, as well as record the content creator. Approximately 10,000 verified annotations have been collected to date.

Our research builds on Beyond Words by utilizing the bounding boxes drawn around photographs, illustrations, comics, editorial cartoons, and maps, as well as additional annotations including ones marking headlines and advertisements, to finetune a pre-trained Faster-RCNN implementation from Detectron2's Model Zoo [318, 405]. Our visual content recognition model predicts bounding boxes around these 7 different classes of visual content in historic newspapers. This section presents our work training this visual content recognition model and constructing a pipeline for automating the identification of this visual content in *Chronicling America*. Drawing inspiration from the Beyond Words workflow, we extract corresponding textual content such as headlines and captions by identifying text from the METS/ALTO OCR that falls within each predicted bounding box. This method is effective at captioning because Beyond Words volunteers were asked to include captions and relevant textual content within their bounding box annotations. Lastly, to enable fast similarity querying for search and recommendation tasks, we generate image embeddings for the extracted visual content using ResNet models pre-trained on ImageNet. This resulting dataset, which we call the Newspaper Navigator dataset, is the largest collection of extracted visual content from historic newspapers ever produced. Our contributions are as follows:

1. We present a publicly available pipeline for extracting visual and textual content

[1] https://labs.loc.gov/work/experiments/beyond-words/

from historic newspaper pages, designed to run at scale over terabytes of image data. Visual content categories include headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements.

2. We release into the public domain a finetuned Faster-RCNN model for this task that achieves 63.4% bounding box mean average precision (mAP)[2] on a validation set of World War 1-era *Chronicling America* pages. This model has been incorporated into Layout Parser, a unified toolkit for deep learning based document image analysis, enabling layout analysis with the model in just a few lines of Python code [338].
3. We present the Newspaper Navigator dataset, a new public dataset of extracted headlines and visual content, as well as corresponding textual content such as titles and captions, produced by running the pipeline over 16.3 million historic newspaper pages in *Chronicling America*. This corpus represents the largest dataset of its kind ever produced. The dataset can be found at `https://news-navigator.labs.loc.gov`.

2.2.2 Related Work

Corpora & Datasets

Over the past 15 years, efforts across the world to digitize historic newspapers have been remarkably successful [248]. In addition to *Chronicling America*, examples of large repositories of digitized newspapers include Trove [52], Europeana [289, 396], Delpher [81], The British Newspaper Archive [207], OurDigitalWorld [280], Papers Past [275], NewspaperSG [40], newspapers.com [246] and Google Newspaper Search [57]. These repositories have inspired the construction of datasets for related supervised learning tasks. In addition to Beyond Words, datasets for historic newspaper recognition include the National Library of Luxembourg's historic newspaper datasets [243] that include segmented articles and advertisements; KBK-1M, a dataset of 1,603,396 images with captions extracted from historic Dutch newspapers; CHRONIC, a dataset of 452,543 images in historic Dutch newspapers [348]; and SIAMESET, a dataset of 426,777 advertisements in historic Dutch newspapers

[2]Mean average precision is the standard metric used for benchmarking object detection models, incorporating intersection over union to assess precision and recall. We describe the metric in more detail in subsection 2.2.5.

[393]. Datasets for machine learning tasks with historical documents include READ-BAD [124] and DIVA-HisDB [343]. However, all of these datasets are subsets of visual content rather than comprehensive datasets of extracted content from full corpora. Our work uses the Beyond Words dataset to train a visual content recognition model in order to process the visual content in the *Chronicling America* corpus comprising 16+ million historic newspaper pages.

Visual Content Extraction

Other researchers have built tools and pipelines for extracting and analyzing visual content from historic documents using deep learning.[3] PageNet utilizes a Fully Convolutional Network for pixel-wise page boundary extraction for historic documents [367]. dhSegment is a deep learning framework for historical document processing, including pixel-wise segmentation and extraction tasks [17]. Liebl and Burghardt benchmarked 11 different deep learning backbones for the pixel-wise segmentation of historic newspapers, including the separation of layout features such as text and tables [209]. The AIDA collaboration has applied deep learning techniques to newspaper corpora including *Chronicling America* and the Burney Collection of British Newspapers [219, 220, 221] for tasks such as poetic content recognition [222, 349] and visual content recognition using dhSegment [223]. Instead of a pixel-wise approach, we utilize bounding boxes, resulting in higher performance. In addition, our pipeline recognizes 7 different classes of visual content, extracts corresponding OCR, and generates image embeddings. Lastly, we deploy our visual content recognition pipeline at scale.

Article Disambiguation

Article disambiguation for historic newspaper pages has long been of interest to researchers, including the IMPRESSO project [303], NewsEye project [313], and Google Newspaper Search [57]. Of particular note is the approach taken by Google Newspaper Search, which extracted headline blocks using OCR font size and area-perimeter ratio as features and

[3]For approaches to historic document classification that do not utilize deep learning, see for example [194].

utilized the extracted headlines to perform article segmentation [57].[4] We, too, focus on headline extraction because it serves as its own form of article disambiguation. However, unlike previous approaches, we treat headline extraction as a *visual* task at the image level, rather than a *textual* task at the OCR level. Our novel approach is to leverage the visual distinctiveness of headlines and train a classifier to predict bounding boxes around headlines on the page. The headline text within each bounding box is extracted from the METS/ALTO OCR.

Lastly, proper article disambiguation requires the ability to filter out text from advertisements due to their ubiquity. As with headlines, we treat advertisement identification as a visual task rather than a textual task because the advertisements are so naturally identified by their visual features. Because our visual content recognition model robustly identifies advertisements, we are able to disambiguate newspaper text from advertisement text.

Image Embeddings and Cultural Heritage

In recent years, researchers have utilized image embeddings for visualizing and exploring visual content in cultural heritage collections. The Yale Digital Humanities Lab's PixPlot interface [90] and the National Neighbors project [213] utilize Inception v3 embeddings [364]. Google Arts & Culture's t-SNE Map utilizes embeddings produced by the Google search pipeline [84]. The Norwegian National Museum's Principal Components project [129] uses finetuned Caffe image embeddings [157]. Olivia Vane utilizes VGG-16 embeddings to visualize the Royal Photographic Society Collection [380]. Likewise, Brian Foo has created a visualization of The American Museum of Natural History's image collection [106] using VGG-16 embeddings [344]. Refik Anadol uses embeddings to visualize the SALT Research collection [15]. Regarding visual content in historic newspapers in particular, Wevers and Smits utilize Inception v3 embeddings to analyze the CHRONIC and SIAMESET datasets described in subsection 2.2.2. Their work includes deploying SIAMESE, a recommender system for historic newspaper advertisements, and analyzing the training of a new classification layer on top of the Inception embeddings to predict custom categories [394]. Indeed,

[4]To our knowledge, the extraction and classification of visual content was outside of the scope of the project.

in addition to supporting visualizations of latent spaces that capture semantic similarity, image embeddings are desirable for visual search and recommendation tasks due to the ability to perform fast similarity querying with them. Using ResNet-18 and ResNet-50 [134] models pre-trained on ImageNet, we generate image embeddings for the extracted visual content, which are included in the Newspaper Navigator dataset.

2.2.3 Code

All code can be found in the public GitHub repository: `https://github.com/LibraryOfCongress/newspaper-navigator`. All code is open source, placed in the public domain for unrestricted re-use. In addition, included in the repository are the finetuned visual content recognition model, the training set on which the model was finetuned, a Jupyter notebook for experimenting with the visual content recognition model, and a slideshow of predictions.

2.2.4 Constructing the Training Set

Repurposing Beyond Words Annotations

To create a training set for our visual content recognition model, we repurposed the publicly available annotations of photographs, illustrations, maps, comics, and editorial cartoons derived from Beyond Words, a crowdsourcing initiative launched by the Library of Congress to engage the American public with the visual content in World War 1-era newspapers in *Chronicling America*. In order to finetune the visual content recognition model, we first reformatted the crowdsourced Beyond Words annotations into a proper data format for training a deep learning model. We chose the Common Objects in Context (COCO) dataset format [210], a standard data format for object detection, segmentation, and captioning tasks adopted by Facebook AI Research's Detectron2 deep learning platform for object detection [405]. The verified Beyond Words annotations used as training data were downloaded from the Beyond Words public website on December 1, 2019.

We reiterate that the instructions for the "Mark" step asked users to "enclose any caption or text describing the picture and the illustrator or photographer" [187]; therefore, a model trained on these annotations learns to include relevant text within the bounding boxes for

Table 2.1: A breakdown of Beyond Words annotations included in the training data for the visual content recognition model, as well as all annotations constituting the training/validation dataset.

Performance (Validation)		
Cateogry	**Beyond Words Annotations**	**Total Annotations**
Photograph	4,193	4,254
Illustration	1,028	1,048
Map	79	215
Comic/Cartoon	1,139	1,150
Editorial Cartoon	293	293
Headline	-	27,868
Advertisement	-	13,581
Total	6,732	48,409

visual content, which can then be extracted from the corresponding METS/ALTO OCR in an automated fashion.

Adding Annotations

Because headlines and advertisements were not included in the Beyond Words workflow, we added annotations for these categories for all images in the dataset. These annotations are not verified, as each page was annotated by only one person. Due to the low number of annotated maps in the Beyond Words data (79 in total), we also annotated 122 pages containing maps, which we retrieved by performing a keyword search of “map” on the *Chronicling America* search portal restricted to 1914-1918. We downloaded the pages on which we identified maps and annotated all 7 categories of visual content on each page. Like the headline and advertisement annotations, these annotations are not verified.

Training Set Statistics

The augmented Beyond Words dataset in COCO format can be found in the Newspaper Navigator repository and is available for unrestricted re-use in the public domain. It contains $3,559$ World War 1-era *Chronicling America* pages with $48,409$ annotations. The annotation category breakdown appears in Table 2.1.

2.2.5 Training the Visual Content Recognition Model

To train a visual content recognition model for identifying the 7 classes of different newspaper content, we chose to finetune a pre-trained Faster-RCNN object detection model from Detectron2's Model Zoo using Detectron2 [405] and PyTorch [286]. Because model inference was the bottleneck on runtime in our pipeline, we chose the Faster-RCNN R50-FPN backbone, the fastest such backbone according to inference time. Though we could have utilized the highest performing Faster-RCNN backbone, which achieved ∼5% higher mean average precision on the COCO [210] pre-training task at the expense of 2.5x the inference time, qualitative evaluation of predictions with the finetuned R50-FPN backbone indicated that it was performing sufficiently. Furthermore, we conjecture that the performance of our model is limited by noise in the training data, rather than model architecture and selection. First, the ground-truth Beyond Words labels were not complete because volunteers were only required to draw one bounding box per page (though more could be added). Second, there was non-trivial disagreement between Beyond Words annotators when marking bounding boxes due to the complexity of visual content layouts.[5]

We performed all finetuning using PyTorch on a g4dn.2xlarge Amazon EC2 instance with a single NVIDIA T4 GPU. Finetuning the R50-FPN backbone was evaluated on a held-out validation set according to an 80%-20% split; the JSON files containing the training and validation splits are available for download in the GitHub repository. We used the following hyperparameters: a base learning rate of 0.00025, a batch size of 8, and 64 proposals per image. `RESIZE_SHORTEST_EDGE` and `RANDOM_FLIP` were utilized as data augmentation techniques.[6] Using early stopping, we finetuned the model for 77 epochs, requiring 17 hours of runtime on the NVIDIA T4 GPU. The model weights file is publicly available and can be found in the GitHub repository for this project.

We report a mean average precision on the validation set of 63.4%; average precision (AP) for each category, as well as the number of validation instances in each category, are reported in Table 2.2. We chose AP because it is the standard metric in the computer

[5]Beyond Words was launched as an experiment, without interventions in workflow or community management; the annotation accuracy should be assessed accordingly.

[6]These were the only supported data augmentation methods at the time of training.

Table 2.2: Average precision (AP) on validation data for the finetuned visual content recognition model on the different categories of content, as well as the number of instances of each category in the validation set. *Averaged* is the mean average precision (mAP) across the 7 classes. *One Class* refers to the average precision when combining all visual content into a single class, capturing how much error is introduced by the detection of visual content versus the classification.

Performance (Validation)

Category	**AP**	**# in Val. Set**
Photograph	61.6%	879
Illustration	30.9%	206
Map	69.5%	34
Comic/Cartoon	65.6%	211
Editorial Cartoon	63.0%	54
Headline	74.3%	5,689
Advertisement	78.7%	2,858
Averaged (mAP)	63.4%	N/A
One Class	75.1%	9,931

vision community for benchmarking object detection tasks. Given a fixed intersection over union (IoU) threshold to evaluate if a prediction is correct, AP is computed by sorting all classifications according to prediction score, generating the corresponding precision-recall curve, and modifying it by drawing the smallest-area curve containing it that is monotonically decreasing. For the COCO standard, AP is then computed by averaging the precision interpolated over 101 different recall values and 10 IoU thresholds from 50% to 95%. For our calculations, we utilized all predictions with confidence scores greater than 0.05, the default threshold in Detectron2.

2.2.6 The Pipeline

Building the Manifest

In order to create a full index of digitized pages for the pipeline to process, we used a forked version of the AIDA collaboration's `chronam-get-images` repository[7] to generate a manifest of filepaths for each newspaper batch.[8] Manifests consisting of 16,368,424 *Chronicling*

[7] https://github.com/bcglee/chronam-get-images

[8] For more information on the batches, see https://chroniclingamerica.loc.gov/batches.

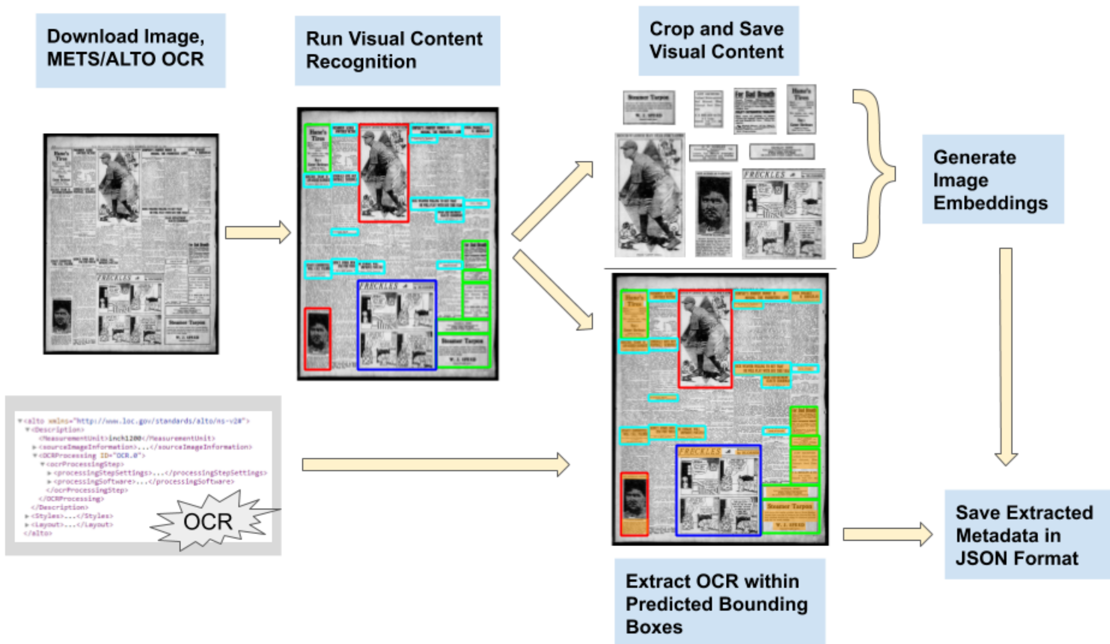

Figure 2.3: A diagram showing the steps of our pipeline.

America pages were compiled in total on March 17, 2020.

Steps of the Pipeline

In Figure 2.3, we show the pipeline workflow. Each manifest was processed in series by our pipeline, which consists of 6 steps:

1. *Downloading the image and METS/ALTO XML for each page and downsampling the image by a factor of 6 to produce a lower resolution JPEG.* Downsampling was performed to reduce I/O and memory consumption, as well as to avoid the overhead introduced by the downsampling that Detectron2 would have to perform before each forward pass during model inference. This step was run in parallel across all 48 CPU cores on each EC2 instance. The files were pulled down from the Library of Congress's public AWS S3 buckets.
2. *Running the visual content recognition model inference on each image to produce bounding box predictions complete with coordinates, predicted classes, and confidence*

scores. This step was run in parallel across all 4 GPUs on each EC2 instance. Predictions with confidence scores greater than 0.05 were saved. We chose to save predictions with low confidence scores in order to allow a user to select a threshold cut based on the user's ideal tradeoff between precision and recall.

3. *Extracting the OCR within each predicting bounding box.* This step required parsing the METS/ALTO XML and was run in parallel across all 48 CPU cores on each EC2 instance.
4. *Cropping and saving the extracted visual content as downsampled JPEGs (for all classes other than headlines).* This step was run in parallel across all 48 CPU cores on each EC2 instance.
5. *Generating ResNet-18 and ResNet-50 embeddings for the cropped and saved images with confidence scores of greater than 0.05.* This step was implemented using a forked version of img2vec[9] [328]. This step was run in parallel across all 4 GPUs on each EC2 instance. The ResNet-18 and ResNet-50 embeddings were extracted from the penultimate layer of each respective architecture after being trained on ImageNet.[10] The 2,048-dimensional ResNet-50 embeddings were selected due to ResNet-50's high performance and fast inference time relative to other image recognition models [39]. The 512-dimensional ResNet-18 embeddings were generated due to their lower dimensionality.
6. *Saving the extracted metadata and cropped images.* The format of the saved metadata is described thoroughly on the dataset's landing page, `https://news-navigator.labs.loc.gov`.

Running the Pipeline at Scale

All pipeline processing was done on 2 g4dn.12xlarge Amazon AWS EC2 instances, each with 48 vCPUs (Intel Cascade Lake) and 4 NVIDIA T4 GPUs. All pipeline code was written in Python 3. The pipeline successfully processed 16,368,041 pages (99.998%) in 19 days of wall-clock time. The manifests of the processed pages, as well as the 383 pages that failed,

[9] `https://github.com/bcglee/img2vec`

[10] We downloaded the pre-trained models from `torchvision.models` in PyTorch [286].

Table 2.3: A breakdown of extracted content in the Newspaper Navigator dataset. Three cuts on confidence score are presented to show the effects when favoring precision or recall.

Newspaper Navigator Dataset Statistics

Category	**Count $\geq$ Confidence Score**		
	≥ 0.9	≥ 0.7	≥ 0.5
Photograph	1.59×10^6	2.63×10^6	3.29×10^6
Illustration	8.15×10^5	2.52×10^6	4.36×10^6
Map	2.07×10^5	4.59×10^5	7.54×10^5
Comic/Cartoon	5.35×10^5	1.23×10^6	2.06×10^6
Editorial Cartoon	2.09×10^5	6.67×10^5	1.27×10^6
Headline	3.44×10^7	5.37×10^7	6.95×10^7
Advertisement	6.42×10^7	9.48×10^7	1.17×10^8
Total	1.02×10^8	1.56×10^8	1.98×10^8

are in our GitHub Repository.

2.2.7 The Newspaper Navigator Dataset

Statistics & Visualizations

A statistical breakdown of extracted content in the Newspaper Navigator dataset is presented in Table 2.3. Because the choice of cut on confidence score affects the cardinality of the resulting visual content, we include statistics for three different threshold cuts of 0.5, 0.7, and 0.9. In Figure 2.4, we show visualizations of the number of photographs, illustrations, maps, comics, editorial cartoons, headlines, and advertisements in the Newspaper Navigator dataset according to year of publication. These visualizations show the average number of appearances per page, as well as the average fraction of the page covered, for each of the seven classes from 1850 to 1950. As in Table 2.3, we show three different cuts. In Figure 2.4, we observe trends such as the rise of photographs at the turn of the 20th century and the gradual increase in the amount of page space covered by headlines from 1880 to 1920.

Dataset Access

The Newspaper Navigator dataset can be accessed via the Newspaper Navigator GitHub repository, as well as the webpage `https://news-navigator.labs.loc.gov/`. This landing

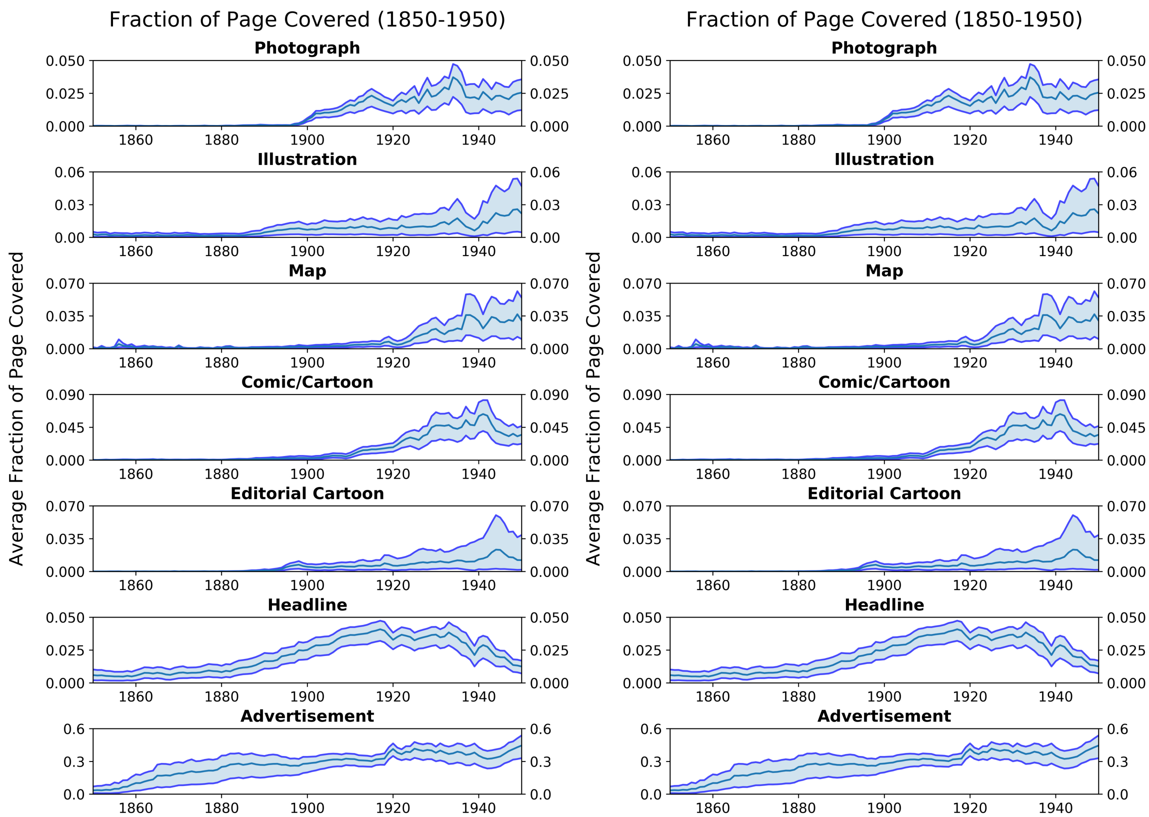

Figure 2.4: Multipanel plots visualizing the visual content in the Newspaper Navigator dataset over time (*left*: number per page; *right*: fraction of each page covered). In each plot, the middle line corresponds to a cut of 0.7 on confidence score, and the upper and lower bounds of the confidence interval in light blue correspond to cuts of 0.5 and 0.9, respectively.

page contains a detailed descripton of the data format, as well as instructions for how to query the dataset. A search user interface is in development.

Pre-packaged Datasets

To make the Newspaper Navigator dataset accessible to those without coding experience, we have pre-packaged hundreds of smaller datasets as zip files, along with metadata in JSON and CSV formats. The pre-packaged datsets are grouped by year and visual content type, enabling users to download all of the 1921 headlines or 1864 maps, for example. Instructions for downloading the pre-packaged datasets can be found at the dataset landing page.

Table 2.4: Average precision (AP) on test sets of 500 annotated pages from 1850 to 1875 and from 1875 to 1900. Due to the rarity of the other classes in the labeled data, only headlines, advertisements, and illustrations are included. As in Table 2.2, *One Class* refers to AP when combining all visual content into one class, capturing how much error is introduced by the detection of visual content versus the classification.

Performance for 19th Century Newspaper Pages

Category	**AP (1850-1875)**	**AP (1875-1900)**
Headline	21.2%	51.6%
Advertisement	7.3%	44.7%
Illustration	N/A	36.4%
One Class	12.1%	48.1%

2.2.8 Discussion

Generalization to 19th Century Newspapers

Given that the visual content recognition model has been trained on World War 1-era newspapers, it is natural to question how the model generalizes to 19th century newspapers. Though Figure 2.4 reveals trends consistent with intuition, such as the emergence of photographs in historic newspapers around 1900, it is still worthwhile to quantify generalization. To do so, we randomly selected and annotated 500 newspaper pages from 1850 to 1875 and 500 pages from 1875 to 1900. In Table 2.4, we present the average precision for headlines, advertisements, and illustrations in the test sets using our annotations as the ground truth. Comparing the results to those in Table 2.2, we observe a moderate dropoff in performance for pages published between 1875 and 1900, as well as a more major dropoff for pages published between 1850 and 1875. However, the extracted visual content from these pages in the Newspaper Navigator dataset is still of sufficient quality for novel analysis.

Partnering with Volunteer Crowdsourcing

Our work is a case study in partnering machine learning projects with volunteer crowdsourcing initiatives, a promising paradigm in which annotators are volunteers who learn about a new topic by participating. With the growing efforts of cultural heritage crowdsourcing initiatives such as the Library of Congress's By the People [274], Smithsonian's

Digital Volunteers [152], the United States Holocaust Memorial Museum's History Unfolded [242], Zooniverse [345], the New York Public Library's Emigrant City [208], the British Library's LibCrowds [206], the Living with Machines project [401], and Trove's newspaper crowdsourcing initiative [24], there are many opportunities to utilize crowdsourced data for machine learning tasks relevant to cultural heritage, from handwriting recognition to botany taxonomic classification [287, 333]. These partnerships have the potential to provide insight into project design, decisions, workflows, and the context of the materials for which crowdsourcing contributions are sought. We hope that our project encourages more machine learning researchers to partner with volunteer crowdsourcing projects, especially on topics pertinent to cultural heritage.

2.2.9 Conclusion & Future Work

We have described our pipeline for extracting, categorizing, and captioning visual content in historic newspapers, including headlines, photographs, illustrations, maps, comics, editorial cartoons, and advertisements. We present the Newspaper Navigator dataset containing these 7 types of extracted visual content from 16.3 million pages from *Chronicling America*. This is the largest dataset of its kind ever produced. In addition to releasing this dataset, we have released our visual content recognition model for historic newspapers, as well as a new training dataset for this task based on annotations from Beyond Words, the Library of Congress Labs's crowdsourcing initiative for annotating and captioning visual content in World War 1-era newspapers in *Chronicling America*. All code has been placed in the public domain for unrestricted re-use.

Future work on the pipeline itself includes improving the visual content recognition model's generalization ability for pre-20th century newspaper pages, especially for the 10.4% of the pages in the Newspaper Navigator dataset published before 1875. This could be accomplished by finetuning on a more diverse training set, which could be constructed by partnering with another volunteer crowdsourcing initiative. One could also imagine training an ensemble of visual content recognition models on different date ranges. Given that only 10.4% of pages in the Newspaper Navigator dataset were published before 1875, it is straightforward to re-run the pipeline with an improved visual content recognition model on this subset.

To improve the extracted OCR, future work includes training a pipeline to correct systematic errors. In the second step of the Beyond Words pipeline, volunteers corrected the OCR appearing in each marked bounding box, resulting in approximately 10,000 corrected textual annotations to date. It is straightforward to construct training pairs of input and output in order to train a supervised model to correct OCR. Other approaches to OCR postprocessing include utilizing existing post-hoc OCR correction pipelines [14, 247], which could be benchmarked on the Beyond Words training pairs.

The future work that excites us most consists of the many ways that the Newspaper Navigator dataset can be used. We are currently building a new search user interface that will be user tested to evaluate new methods of exploratory search. Future work also includes investigating a range of digital humanities questions. For example, the Viral Texts [72] and Oceanic Exchanges [71] projects have studied text reproduction patterns in 19th century newspapers; the Newspaper Navigator dataset allows us to study photograph reproduction in 20th century newspapers. In addition, using the headlines in Newspaper Navigator, we can study which news cycles appeared in different regions of the United States and when. These examples are just a few of many to be explored with the Newspaper Navigator dataset. We hope to inspire a wide range of digital humanities, public humanities, and creative computing projects.

2.3 Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset

2.3.1 Preface: Contextualizing This Work

This work is based upon a single-author publication that appeared in *Digital Humanities Quarterly* [196]. This "data archaeology" was written for an audience distinct from much of the work that appears in this dissertation: specifically, those in the humanities. In particular, this work represents an attempt to communicate the Newspaper Navigator to a wider audience and participate in emerging discussions within humanistic disciplines, including critical data studies and the digital humanities, surrounding the application of machine learning to cultural heritage. As a result, this work dispenses with a technical overview (found in other sections) in favor of an emphasis on participating in these discussions in the humanities. Due to the nature of much of this text serving an explanatory function for these new audiences, it may appear redundant to computer science audiences who have read the technical overview of the project. This data archaeology is autoethnographic because it represents my own attempt to document and self-reflect upon a project that I have created, rather than auditing or excavating a project created by another person or organization. My hope is that this data archaeology offers a new perspective not captured within the technical work toward a more holistic description of Newspaper Navigator.

2.3.2 Introduction

The increasing roles of machine learning and artificial intelligence in the construction of cultural heritage and humanities datasets necessitate critical examination of the myriad biases introduced by machines, algorithms, and the humans who build and deploy them. From image classification to optical character recognition, the effects of decisions ostensibly made by machines compound through the digitization pipeline and redouble in each step, mediating our interactions with digitally-rendered artifacts through the search and discovery process. As a result, scholars within the digital humanities community have begun advocating for the proper contextualization of cultural heritage datasets within the socio-technical systems in which they are created and utilized. One such approach to this contextualization is the data archaeology, a form of humanistic excavation of a dataset that Paul Fyfe defines as "recover[ing] and reconstitut[ing] media objects within their changing ecologies" [113]. Within critical data studies, this excavation of a dataset - including its construction and mediation via machine learning - has proven to be a capacious approach. However, data archaeologies have yet to be adopted as standard practice among cultural heritage practitioners who produce such datasets with machine learning.

I present a data archaeology of the Library of Congress's Newspaper Navigator dataset, which I created as part of the Library of Congress's Innovator in Residence program [202]. The dataset consists of visual content extracted from 16 million historic newspaper pages in the *Chronicling America* database using machine learning techniques. In this case study, I examine the manifold ways in which a *Chronicling America* newspaper page is transmuted and decontextualized during its journey from a physical artifact to a series of probabilistic photographs, illustrations, maps, comics, cartoons, headlines, and advertisements in the Newspaper Navigator dataset [113]. Accordingly, I draw from fields of scholarship including media archaeology, critical data studies, science and technology studies, and autoethnographic approaches throughout.[11]

To excavate the Newspaper Navigator dataset, I consider the digitization journeys of four different pages in Black newspapers included in *Chronicling America*, all of which

[11] For a representative example of an autoethnography, see [26].

reproduce the same photograph of W.E.B. Du Bois in an article announcing the launch of The Crisis, the official magazine of the NAACP. In tracing the newspaper pages' journeys, I unpack how each step in the *Chronicling America* and Newspaper Navigator pipelines, such as the imaging process and the construction of training data, not only imprints bias on the resulting Newspaper Navigator dataset but also propagates the bias through the pipeline via the machine learning algorithms employed. Along the way, I investigate the limitations of the Newspaper Navigator dataset and machine learning techniques more generally as they relate to cultural heritage, with a particular focus on marginalization and erasure via algorithmic bias, which implicitly rewrites the archive itself.

In presenting this case study, I argue for the value of the data archaeology as a mechanism for contextualizing and critically examining cultural heritage datasets within the communities that create, release, and utilize them. I offer this autoethnographic investigation of the Newspaper Navigator dataset in the hope that it will be considered not only by users of this dataset in particular but also by digital humanities practitioners and end users of cultural heritage datasets writ large.

2.3.3 Why a Data Archaeology?

As machine learning and artificial intelligence play increasing roles in digitization and digital content stewardship, the Libraries, Archives, and Museums ("LAM") community has repeatedly emphasized the importance of ensuring that these emerging methodologies are incorporated ethically and responsibly. Indeed, a major theme that emerged from the "Machine Learning + Libraries Summit" hosted by LC Labs in September, 2019, was that "there is much more 'human' in machine learning than the name conveys" and that transparency and communication are first steps toward addressing the "human subjectivities, biases, and distortions" embedded within machine learning systems [156]. This data archaeology has been written in support of this call for transparency and responsible stewardship, which is echoed in the Library of Congress's Digital Strategy [192], as well as the recommendations in Ryan Cordell's report to the Library of Congress "ML + Libraries: A Report on the State of the Field" [70], Thomas Padilla's OCLC position paper "Responsible Operations: Science, Machine Learning, and AI in Libraries" [284], and the University of Nebraska-Lincoln's report on machine learning to the Library of Congress [223]. I write this data archaeology from my perspective of having created the dataset, and although I am not without my own biases, I have attempted to represent my work as honestly as possible. Accordingly, I seek not only to document the construction of the Newspaper Navigator dataset through the lens of data stewardship but also to critically examine the dataset's limitations. In doing so, I advocate for the importance of autoethnographic approaches to documenting a cultural heritage dataset's construction from a humanistic perspective.

This article draws inspiration from recent works in media and data archaeology, including Paul Fyfe's "An Archaeology of Victorian Newspapers" [113]; Bonnie Mak's "Archaeology of a Digitization" [230]; Kate Crawford and Trevor Paglen's "Excavating AI: The Politics of Images in Machine Learning Training Sets" [75]; and, most directly, Ryan Cordell's "Qi-jtb the Raven: Taking Dirty OCR Seriously," in which Cordell traces the digitization of a single issue of the Lewisburg Chronicle from its selection by the Pennsylvania Digital Newspaper Project to its ingestion into the *Chronicling America* online database, with a focus on the distortive effects of OCR [69, 112, 230]. As argued by Trevor Owens and Thomas Padilla, it is essential to "document how digitization practices and how the affordances of particular sources . . . produce unevenness in the discoverability and usability of collections" [282]. Recent works within the machine learning literature have analogously emphasized the importance of documenting the collection and curation efforts underpinning community datasets and machine learning models. Reporting mechanisms include "Datasheets for Datasets" [115], "Dataset Nutrition Labels" [145], "Data Statements for NLP" [37] "Model Cards for Model Reporting" [240], and "Algorithmic Impact Assessments" [317]. This case study adopts a similar framing in stressing the importance of reporting mechanisms, with a particular focus on the data archaeology in the context of cultural heritage datasets.

In the following subsections, I trace the digitization process and data flow for Newspaper Navigator, beginning with the physical artifact of the newspaper itself and ending with the machine learning predictions that constitute the Newspaper Navigator dataset, reflecting on each step through the lens of discoverability and erasure. In particular, I study four different *Chronicling America* Black newspaper pages published in 1910, each depicting the same photograph of W.E.B. Du Bois, as the pages move through the *Chronicling America* and Newspaper Navigator pipelines. All four pages reproduce the same article by Franklin F. Johnson, a reporter from *The Baltimore Afro-American* [97]; the headline is as follows:

NEW MOVEMENT
BEGINS WORK
Plan and Scope of the Asso-
ciation Briefly Told.
Will Publish the Crisis.
Review of Causes Which Led to the
Organization of the Association in
New York and What Its Policy Will
Be-Career and Work of Professor
W.E.B. Du Bois

The article describes the creation of the National Association for the Advancement of Colored People (NAACP), details W.E.B. Du Bois's background, and announces the launch of *The Crisis*, the official magazine of the NAACP, with Du Bois as Editor-in-Chief. The four pages comprise the front page of the October 14th, 1910, issue of the *Iowa State Bystander* [267]; the 16th page of the October 15th, 1910, issue of *Franklin's Paper the Statesman* [262] and the 2nd and 3rd pages of the October 15th, 1910, and November 26th, 1910, issues of *The Broad Ax*, respectively [260, 261]. All four digitized pages are reproduced at the end of the chapter.

2.3.4 Chronicling America: A Genealogy of Collecting, Microfilming, and Digitizing

Any examination of Newspaper Navigator must begin with the genealogy of collecting, microfilming, and digitizing that dictates which newspapers have been ingested into the *Chronicling America* database. The question of what to digitize is, in practice, answered and realized incrementally over decades, beginning at its most fundamental level with the question of which newspapers have survived and which have been reduced to lacunae in the historical record, compiling bibliographies of serials published after 1820 remains an immensely difficult task [131]. Historic newspapers present challenges for digitization in part due to the ephemerality of the physical printed newspaper itself: many newspapers were microfilmed and immediately discarded due to a fear that the physical pages would deteriorate.[12] Indeed, almost all of the pages included in *Chronicling America* have been digitized directly from microfilm. In the next subsection, I will examine the microfilm imaging process in more detail; however, in most cases, librarians selected newspapers for collecting and microfilming decades before the National Digital Newspaper Program was launched in 2004. These selections were informed by a range of factors including historical significance - itself a subjective, nebulous, and ever-evolving notion that has historically served as the basis for perpetuating oppression within the historical record. In "Chronicling White America," Benjamin Fagan highlights the paucity of Black newspapers in *Chronicling America*, in particular in relation to pre-Civil War era newspapers [95]. It is imperative to remember that this paucity can directly be traced back decades to the collecting and preserving stages.[13]

In regard to collecting, the newspaper page is both an informational object (i.e., the newspaper page as defined by its content) and a material object (i.e., the specific printed copy of the newspaper page) [281]. At some point in time, librarians accessioned a specific copy of each printed page and microfilmed it or contracted out the microfilming. The materiality of that specific printed page is a confluence of unique ink smudges, rips, creases, and page alignment, much of which is captured in the microfilm imaging process. Though we may not make much of a crease or a smudge on a digitized page when we find it in the *Chronicling America* database, it can very well take on a life of its own with a machine learning algorithm in Newspaper Navigator. The machine learning algorithm might deem two newspaper photographs as similar simply due to the presence of creases or smudges,

[12] The extent to which newspaper microfilming was driven by credible fear of deterioration versus other factors, such as microfilm marketing, is an important question that is rightly debated. For more on this topic, see [28].

[13] For example, a 2017 article describing the West Virginia University Libraries' West Virginia & Regional History Center and its participation in the National Digital Newspaper Program states: "By August 2017, all known issues of West Virginia's African-American newspapers from the 19th and early 20th centuries will have been digitized" [235]. The article describes Curator Stewart Plein's efforts to locate surviving copies of three Black West Virginia newspapers in order to digitize and include them in *Chronicling America*.

even if the photographs are easily discernible to the naked eye, or the smudges are of entirely different origin (i.e., a printing imperfection versus a smudge from a dirty hand).

It is only by foregrounding these subtleties of the collection, preservation, and microfilming processes that we can understand the selection process for *Chronicling America* in its proper context. The grant-seeking process dictates selection criteria for *Chronicling America* by which state-level institutions including state libraries, historical societies, and universities apply for two years of grant funding from the National Digital Newspaper Program via the Division of Preservation and Access at the National Endowment for the Humanities. With the awarding of a grant, a state-level awardee then digitizes approximately 100,000 newspaper pages published in their state for inclusion in *Chronicling America* [276, 300]. The grant-seeking and awarding process is nuanced, but salient points include that state-level applicants must assemble an advisory board including scholars, teachers, librarians, and archivists to aid in the selection of newspapers, and grants are reviewed by National Endowment for the Humanities staff, as well as peer reviewers.[14]

Regarding selection criteria for newspaper titles, the National Digital Newspaper Program defines the following factors for state-level awardees to consider for content selection after a newspaper is determined to be in the public domain [299]

1. Image quality in the selection of microfilm
2. Research value
3. Geographic representation
4. Temporal coverage
5. Bibliographic completeness of microfilm copy
6. Diversity (i.e., "newspaper titles that document a significant minority community at the state or regional level")
7. Whether the title is orphaned (i.e., whether the newspaper has "ceased publication and lack[s] active ownership" [301])

[14] For a thorough case study of this process, I direct the reader to "Qi-jtb the Raven," in which Ryan Cordell walks through an example with the Pennsylvania Digital Newspaper Program [69].

8. Whether the title has already been digitized.

Though factors such as research value are considered by each state awardee's advisory board, as well as by the National Endowment for the Humanities and peer review experts, the titles included in *Chronicling America* are largely dictated by which exist on microfilm and are of sufficient image quality within a state-level grantee's collection. Thus, the significance of the collection and microfilming practices of decades prior cannot be understated.

I also highlight that assessing microfilmed titles based on image quality is a complex procedure in its own right. The National Digital Newspaper Program has made publicly available a number of resources devoted specifically to this task, including documents and video tutorials [31, 302]. They articulate factors such as the microfilm generation (archive master, print master, or review copy), the material (polyester or acetate), the reduction ratio, and the physical condition. The detailed resources made available by the National Digital Newspaper Program, the Library of Congress, and the National Endowment for the Humanities for navigating this process are testaments to the multidimensional complexity of the selection process for *Chronicling America* [276, 296, 299].

We have not yet investigated the topic of digitization, and we have already encountered a profusion of factors from collection to digitization that mediate which artifacts appear in *Chronicling America* and thus Newspaper Navigator. Let us now examine the microfilm itself.

2.3.5 The Microfilm

In "What Computational Archival Science Can Learn from Art History and Material Culture Studies," Lyneise Williams shares a powerful anecdote of coming across a physical copy of a 1927 issue of the French sports newspaper *Match L'Intran* that featured accomplished Black Panamanian boxer, Alfonso Teofilo Brown, on the front cover [398]. Williams describes Brown as "glowing. He looked like a 1920s film star rather than a boxer" [398]. Curious to learn more about the printing process, Williams discovered that the issue of Match L'Intran was produced using rotogravure, a specific printing process that could "capture details in dark tones" [398]. However, when Williams found a version of the same newspaper cover that had been digitized from microfilm, it was apparent that the microfilming process had washed out the detail of the rotogravure, reducing Brown to a “flat black, cartoonish form” [398]. Williams relays the anecdote to articulate that the microfilming process itself is thus a form of erasure for communities of color [398].

The grayscale saturation of photographs induced by microfilming is widely documented and recognizable to most researchers who have ever worked with the medium [28]. however, Lyneise Williams’s article affords us a lens into what precisely is lost amongst the distortive effects of the microfilming process. This erasure via microfilming can be seen in *Chronicling America* directly. In Figure 2.5, I show the same photograph of W.E.B. Du Bois as it appears in 4 different *Chronicling America* newspaper pages published during October and November of 1910 and digitized from microfilm [260, 261, 262, 267]. The phenomenon described by Williams is immediately recognizable in these four images: Du Bois’s facial features are distorted by the grayscale saturation. In the case of the *Iowa State Bystander*, Du Bois has been rendered into a silhouette.

Moreover, each digitized reproduction reveals unique visual qualities, varying in contrast, sharpness, and noise - a testament to the confluence of mediating conditions from printing through digitization that have rendered each newspaper photograph in digital form. Even in the case of the two images reproduced in the *The Broad Ax*, which were digitized from the very same microfilm reel (reel #00280761059) by the University of Illinois at Urbana-Champaign Library, variations are still apparent. To understand how these subtle differences between images are amplified through digitization, we now turn to optical character recognition.

2.3.6 The OCR

Optical character recognition, commonly called OCR, refers to machine learning algorithms that are trained to read images of typewritten text and output machine-readable text, thereby providing the bridge between an image of typewritten text and the transcribed text itself. Because OCR algorithms are “trained and evaluated using labeled data: examples with ground-truth classification labels that have been assigned by another means,” the algorithms are considered a form of supervised learning in the machine learning literature

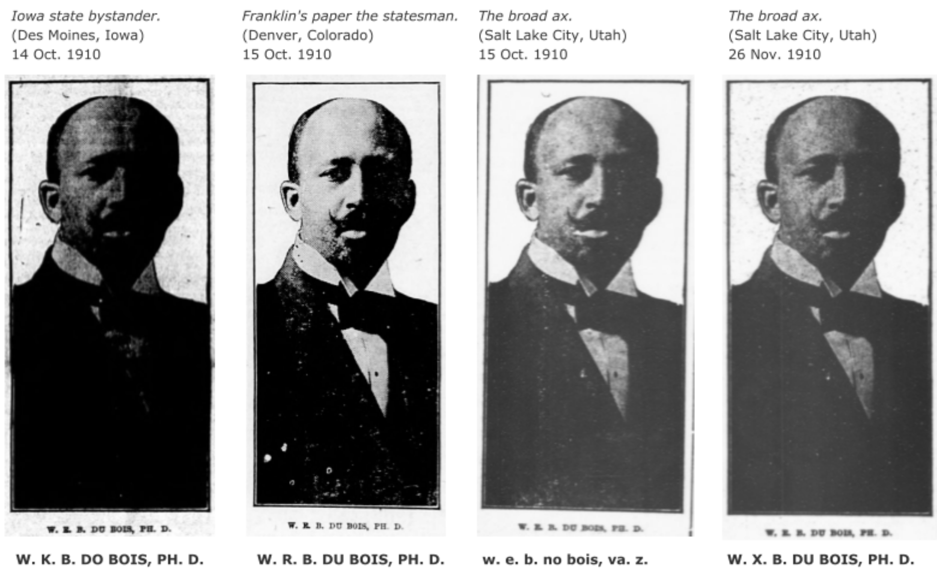

Figure 2.5: The same image of W.E.B. Du Bois reproduced in 4 different digitized Black newspapers in *Chronicling America* from 1910. Note that the combined effects of printing, microfilming, and digitizing have led to different visual effects in each image, ranging from contrast to sharpness. The OCR transcriptions of the caption "W. E. B. DU BOIS, PH. D." appearing in the image of W.E.B. Du Bois reproduced in 4 different digitized Black newspapers in *Chronicling America* are shown below each image. These OCR transcriptions are provided by *Chronicling America*.

[194]. OCR engines are remarkably powerful in their ability to improve access to historic texts. Indeed, OCR is a crucial form of metadata for *Chronicling America*, enabling keyword search in the search portal and making possible scholarship with the newspaper text at large scales.[15] However, OCR is not perfect. Although humans are able to discern an "E" from an "R" on a digitized page even if the type has been smudged, an OCR engine is not always able to do so: its performance is dependent on factors ranging from the sharpness of text in an image to printing imperfections to the specific typography on the page.

In Figure 2.5, I show the OCR transcriptions of the captions of all four images, as provided by *Chronicling America*. All four transcriptions fail to reproduce the true caption

[15]For exemplary research collaborations that utilize the *Chronicling America* bulk OCR, see the Viral Text Project [72] and the Oceanic Exchanges Project [71].

with 100% accuracy, differing from one another by at least one character. Consequently, a keyword search of "W. E. B. Du Bois" over the raw text would not register the caption for any of the four photographs (the *Chronicling America* search portal utilizes a form of relevance search to alleviate this problem). These examples reveal how sensitive OCR engines are to slight perturbations, or "noise," in the digitized images, from ink smudges to text sharpness to page contrast. Though the NDNP awardees who contributed these pages may have utilized different OCR engines or chosen different OCR settings, the OCR for the two image captions from *The Broad Ax* that have been from the very same microfilm reel was in all likelihood generated using the same OCR engine and settings. Put succinctly, OCR engines amplify the noise from both the material page and the digitization pipeline.[16]

Though OCR engines have become standard components of digitization pipelines, it is important to remember that OCR engines are themselves machine learning models that have been trained on sets of transcribed typewritten pages. Like any machine learning model, OCR predictions are thus subject to biases encoded not only in the OCR engine's architecture but also in the training data itself. Though it is often called algorithmic bias, this bias is undeniably human, in that the construction of training data machine learning models are imprinted with countless human decisions and judgment calls. For example, if an OCR engine is trained on transcriptions that consistently misspell a word, the OCR engine will amplify this misspelling across all transcriptions of processed pages.[17] A recurring theme of algorithmic bias is that it is a force for marginalization, especially in the context of how we navigate information digitally. In *Algorithms of Oppression*, Safiya Noble describes how Google's search engine consistently marginalizes women and people of color by displaying search results that reinforce racism [250]. This bias is not restricted to Google: in *Masked by Trust: Bias in Library Discovery*, Matthew Reidsma articulates how library search engines suffer from similar biases [316]. Despite the fact that knowledge of algorithmic bias in

---

[16]For other examinations of how OCR mediates our interactions with digital archives, see [145, 239, 357, 372, 402].

[17]For a concrete example of a similar phenomenon in the image domain, see [194], in which a machine learning algorithm was trained to classify digitized images but consistently misclassified images that had been misoriented 180 degrees in the scanning bed - a consequence of the classifier not having seen enough instances of these misoriented scans during training.

relation to search engines and image recognition tools is becoming increasingly widespread among the cultural heritage community, the errors introduced by OCR engines are often accepted as inevitable without critical inquiry from this perspective. However, algorithmic bias is a useful framework for examining OCR engines [5].

Perhaps the most significant challenge to studying OCR engines is that the best-performing and most widely-used OCR engines are proprietary. Though ABBYY FineReader and Google Cloud Vision API offer high performance, the systems fundamentally are black boxes: we have no access to the underlying algorithms or the training data. The ability to audit a system is crucial to developing an understanding of how it works and the biases it encodes. The fact that many OCR engines are opaque prevents us from disentangling whether poor performance on a particular page is due to algorithmic limitations or due to a lack of relevant training data. The distinction is significant: the former may reflect an algorithmic upper bound, whereas the latter reflects decisions made by humans.

Indeed, algorithmic bias distorts and occludes the historical record, as it is made discoverable through OCR. Discrepancies in OCR performance for different languages and scripts is a consequence of human prioritization, from the collection of training data and lexicons to the development of the algorithms themselves. As articulated by Hannah Alpert-Abrams in "Machine Reading the Primeros Libros," "the machine-recognition of printed characters is a historically charged event, in which the system and its data conspire to embed cultural biases in the output, or to affix them as supplementary information hidden behind the screen" [5]. Alpert-Abrams's work reveals how the OCR inaccuracies for indigenous languages recorded in colonial scripts perpetuate colonialism. For other languages such as Ladino, typically typeset in Rashi script, the lack of high-performing OCR has presented consistent challenges for digitization and scholarship.

In the case of *Chronicling America*, the National Digital Newspaper Program is exemplary in its efforts to support OCR for non-English languages. In the Notice of Funding Opportunity for the National Digital Newspaper Program produced by the Division of Preservation of Access at the National Endowment for the Humanities, OCR performance in different languages is explicitly addressed: "Applicants proposing to digitize titles in languages other than English must include staff with the relevant language expertise to review the quality of the converted content and related metadata" [276]. I have included this discussion of OCR and algorithmic bias to offer a broader provocation regarding machine learning and digitization: how much text in digitized sources has been transmuted by this effect and thus effectively erased due to inaccessibility when using search and discovery platforms?

2.3.7 The Visual Content Recognition Model

I will now turn to the Newspaper Navigator pipeline itself, in particular the visual content recognition model. Trained on annotations from the Beyond Words crowdsourcing initiative, as well as additional annotations of headlines and advertisements, the visual content recognition model detects photographs, illustrations, maps, comics, editorial cartoons, headlines, and advertisements on historic newspaper pages.

As described in the previous subsection, examining training data is an essential component of auditing any machine learning model, from understanding how the dataset was constructed to uncovering any biases in the composition of the dataset itself. For the visual content recognition model, this examination begins with Beyond Words. Launched in 2017 by LC Labs, Beyond Words has collected to-date over 10,000 verified annotations of visual content in World War 1-era newspaper pages from *Chronicling America*. The Beyond Words workflow consists of the three steps listed below:

1. A "Mark" step, in which volunteers are asked to draw bounding boxes around visual content on the page [187]. The instructions read as follows:

> "In the Mark step, your task is to identify and select pictures in newspaper pages. For our project, 'pictures' means illustrations, photographs, comics, and cartoons. You'll use the marking tool to draw a box around the picture using your mouse. After you have marked all pictures on the newspaper page, click the 'DONE' button. Skip the page altogether by clicking the 'Skip this page' button. If no illustrations, photographs, or cartoons appear on the page, click the 'DONE' button. Not sure if a picture should be marked? Select the 'Done for now, more left to mark' button so another

volunteer can help finish that page. Please do not select pictures within advertisements."

2. A "Transcribe" step, in which volunteers are asked to transcribe the caption of the highlighted visual content, as well as note the artist and visual content category ("Photograph," "Illustration," "Map," "Comics/Cartoon," "Editorial Cartoon") [188]. The transcription is pre-populated with the OCR falling within the bounding box in question. The instructions for this step state:

   > "Most pictures have captions or descriptions. Enter the text exactly as you see it. Include capitalization and punctuation, but remove hyphenation that breaks words at the end of the line. Use new lines to separate different parts of captions and descriptions. You can zoom in for better looks at the page. You can also select 'View the original page' in the upper right corner of the screen to view the original high resolution image of the newspaper."

   An example of this step can be seen in Figure 2.6.

3. A "Verify" step, in which volunteers are asked to select the best caption for an identified region of visual content from at least two examples; alternatively, a volunteer can add another caption [189]. The instructions state: "Choose the transcription that most accurately captures the text as written. If multiple transcriptions appear valid, choose the first one. If the selected region isn't appropriate for the prompt, click 'Bad region."'

For the purposes of Newspaper Navigator, only the bounding boxes from the "Mark" step and the category labels from the "Transcribe" step were utilized as training data; however, understanding the full workflow is essential because annotations are considered "verified" only if they have passed through the full workflow.

A number of factors contribute to which *Chronicling America* pages were processed by volunteers in Beyond Words. First, the temporal restriction to World War 1-era pages affects the ability of the visual content recognition model to generalize: after all, if the model is trained on World War 1-era pages, how well should we expect it to perform on 19th century

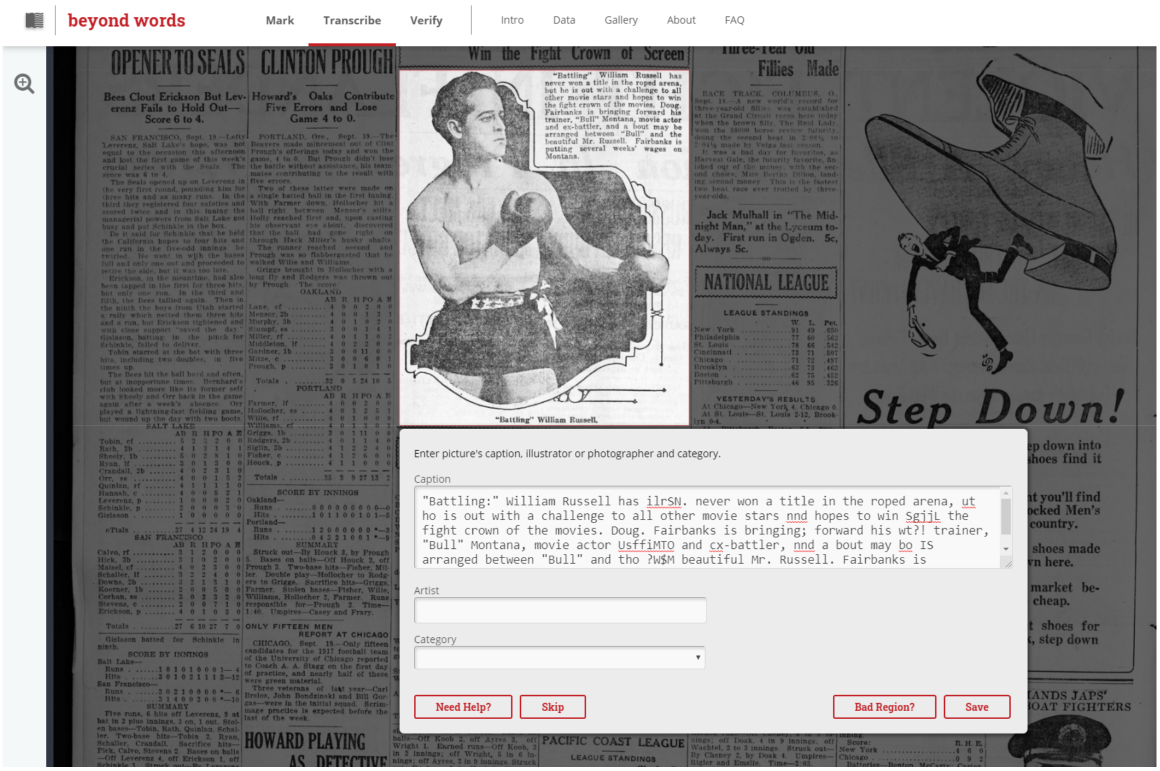

Figure 2.6: A screenshot showing an example of the "Transcribe" step of the Beyond Words workflow. Note that the photograph caption is pre-populated using the OCR falling within the bounding box [188].

pages? I will return to this question later in the subsection. Moreover, Beyond Words volunteers could select either an entirely random page or a random page from a specific state, an important affordance from an engagement perspective, as volunteers could explore the local histories of states in which they are interested. But this affordance is also imprinted on the training data, as certain states - and thus, certain newspapers - appear at a higher frequency than if the World War-1 era *Chronicling America* pages had been drawn randomly from this temporal range in *Chronicling America*.

Furthermore, it should be noted that the "Mark" and "Transcribe" steps - specifically, drawing bounding boxes and labeling the visual content category - are complex tasks. Because newspaper pages are remarkably heterogenous, ambiguities and edge-cases abound. Should a photo collage be marked as one unit or segmented into constituent parts? What precisely is the distinction between an editorial cartoon and an illustration? How much relevant textual content should be included in a bounding box? Naturally, volunteers did not always agree on these choices. In this regard, the notion of a ground-truth, a set of perfect annotations against which we can assess performance, is itself called into question. Moreover, with thousands of annotations, mistakes in the form of missed visual content, as well as misclassifications, are inevitable. These ambiguities and errors are natural components of any training dataset and must be taken into account when analyzing a machine learning model's predictions.

A breakdown of Beyond Words annotations included in the training data can be found in the second column of Table 2.1. I downloaded these 6,732 publicly-accessible annotations as a JSON file on December 1, 2019. Table 1 reveals an imbalance between the number of examples for each category; in the language of machine learning, this is called class imbalance. While the discrepancy between maps and photographs is to be expected, the fact that so few maps were included was concerning from a machine learning standpoint: a machine learning algorithm's ability to generalize to new data is dependent on having many diverse training examples. To address this concern, I searched *Chronicling America* and identified 134 pages published between January 1st, 1914, and December 31st, 1918, that contain maps. I then annotated these pages myself.

In addition, during the development of the Newspaper Navigator pipeline, I realized the value in training the visual content recognition model to identify headlines and advertisements. Consequently, I added annotations of headlines and advertisements for all 3,559 pages included in the training data. The statistics for this augmented set of annotations can be found in the third column of Table 2.1. Though I attempted to use a consistent approach to annotating the headlines and advertisements, my interpretation of what constitutes a headline is certainly not unimpeachable: I am not a trained scholar of periodicals or of print culture; even if I were, the task itself is inevitably subjective. Furthermore, I made decisions to annotate large grids of classified ads as a single ad to expedite the annotation process. Whether this was a correct judgment call can be debated. Lastly, annotating all 3,559 pages for headlines and advertisements required a significant amount of time, and there are inevitably mistakes and inconsistencies embedded within the annotations. My own decisions in terms of how to annotate, as well as my mistakes and inconsistencies, are embedded within the visual content recognition model through training. For those interested in examining the training data directly, the data can be found in the GitHub repository for this project [195].

Beyond the construction of the training data, I made manifold decisions regarding the selection of the correct model architecture and the training of the model. Because this discussion surrounding these choices is quite technical, I refer the reader to [202] for an in-depth examination. However, I will state that the choice of model, the number of iterations for which the model was trained, and the choice of model parameters are all of significant import for the resulting trained model and consequently, the Newspaper Navigator dataset.

I will now turn to the visual content recognition model's outputs in relation to the Newspaper Navigator pipeline. The model itself consumes a lower-resolution version of a *Chronicling America* page as input and then outputs a JSON file containing predictions, each of which consists of bounding box coordinates,[18] the predicted class (i.e., "photograph," "map," etc.), and a confidence score generated by the machine learning model.[19] Cropping out and saving the visual content required extra code to be written. Because the high-resolution images of the *Chronicling America* pages, in addition to the METS/ALTO OCR, amount to many tens of terabytes of data, questions of data storage became major considerations in the pipeline. I chose to save the extracted visual content as lower-resolution JPEG images in order to reduce the upload time and lessen the storage burden. Though the Newspaper Navigator dataset retains identifiers to all high-resolution pages in *Chronicling America*, the images in the Newspaper Navigator dataset are altered by the downsampling procedure. This downsampling procedure should be free of any significant biasing effects.

For visual content recognition, Newspaper Navigator utilized an object detection model, which is a type of widely-used computer vision technique for identifying objects in images. The performance for computer vision techniques is regularly measured using metrics such as average precision. For Newspaper Navigator, the model's performance on a specific page, as

[18]Bounding box coordinates refer to the positions of the corners of the predicted bounding box, relative to the image coordinates.

[19]The confidence score is examined in more detail in the next subsection.

measured by average precision, is dependent on a confluence of factors. These factors include the page's layout, artifacts and distortions introduced in the microfilming and digitization process, and - most importantly - the composition of the training data. Thus, each image is "seen" differently by the visual content recognition model. In Figure 2.7, I show the four images of W.E.B. Du Bois, as identified by the visual content recognition model and saved in the Newspaper Navigator dataset. Each image is cropped slightly differently. In the case of the image from the *Iowa State Bystander*, extra text is included, while in the case of the images from The Broad Ax, the captions are partially cut off. The loss in image quality is due to the aforementioned downsampling performed by the pipeline. This downsampling leads to artifacts such as the dots appearing on Du Bois's face in the image from the *Iowa State Bystander*, as well as the streaks in the image from *Franklin's Paper the Statesman*, that are not present in Figure 2.5.

Returning to the question of the visual content recognition model's performance on pages published outside of the temporal range of the training data (1914-1918), it is possible to provide a quantitative answer by measuring average precision on test sets of annotated pages from different periods of time. In [202], I describe this analysis in detail and demonstrate that the performance declines for pages published between 1875 and 1900 and further declines for pages published between 1850 and 1875. This confirms that the composition of the training data directly manifests in the model's performance. While it is certainly the case that the Newspaper Navigator dataset can still be used for scholarship related to 19th century newspapers in *Chronicling America*, any scholarship with the 19th century visual content in the Newspaper Navigator dataset must consider how the dataset may skew what visual content is represented.

Let me conclude this subsection with a discussion of the act of visual content extraction itself in relation to digitization. While this extraction enables a wide range of affordances for searching *Chronicling America*, it is also an act of decontextualization: visual content no longer appears in relation to the mise-en-page. At the end of the chapter, the full pages containing the photographs of W.E.B. Du Bois are reproduced, showing each photograph in context. Only by examining the full pages does it become clear that the article featuring W.E.B. Du Bois was printed with a second article in the *Iowa State Bystander* and *The*

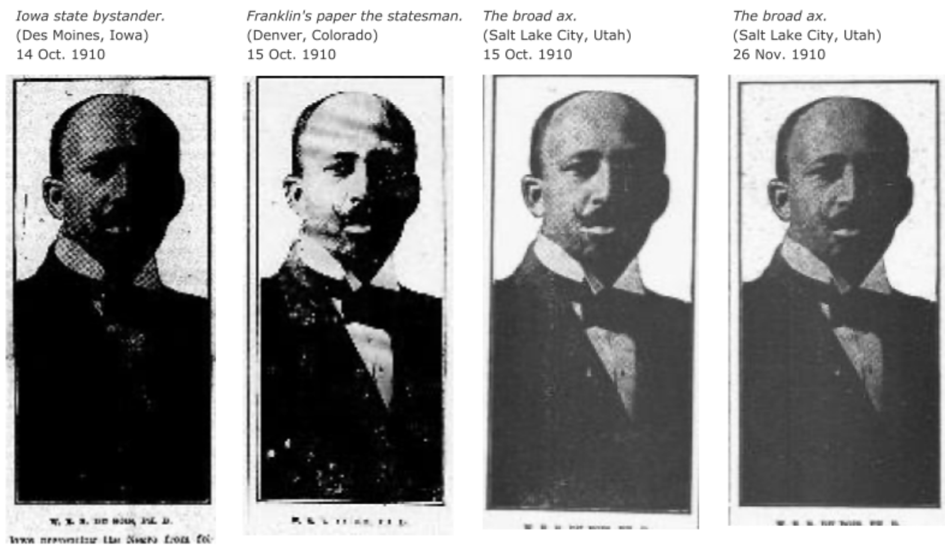

Figure 2.7: The four images of W.E.B. Du Bois, as identified by the visual content recognition model and included in the Newspaper Navigator dataset [263, 264, 265, 266].

Broad Ax, the headline of which reads: "ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON — Afro-American Women Unite For Active Campaign Against Injustice." Furthermore, upon examination, the *Iowa State Bystander* front page features the article on The Crisis and W.E.B. Du Bois as the most prominent article of the issue. Though links between the extracted visual content and the original *Chronicling America* pages are always retained, this decontextualization inevitably transmutes how we perceive and interact with the visual content in *Chronicling America.* Indeed, all uses of machine learning for metadata enhancement are a form of decontextualization, centering the user's discovery and analysis of content around the metadata itself.

2.3.8 Prediction Uncertainty

Perhaps the most fundamental question to ask of the Newspaper Navigator dataset is: "How many photographs does the dataset contain?" Because the dataset has been constructed using a machine learning model, predictions are ultimately probabilistic in nature, quantified by the confidence score returned by the model. This begs the question of what counts as an identified unit of visual content: a user is much more inclined to tally a prediction of a map if it has an associated confidence score of 99% rather than 1%. However, choosing this cut is fundamentally a subjective decision, informed by the user's end goals with the dataset. In the language of machine learning, picking a stringent confidence cut (i.e., only counting predictions with high confidence scores) emphasizes precision: a prediction of a photograph likely corresponds to a true photograph, but the predictions will suffer from false negatives. Conversely, picking a loose confidence cut (i.e., counting predictions with low confidence scores) emphasizes recall: most true photographs are identified as such, but the predictions will suffer from many false positives. In this regard, the total number of images in the Newspaper Navigator dataset is dependent on one's desired tradeoff between precision and recall. In Table 2.3, I show the dynamic range of the dataset size, as induced by three different cuts on confidence score: 90%, 70%, and 50%. Figure 2.8 shows the effects of different cuts on confidence score for the page featuring W.E.B. Du Bois in the November 26,1910, issue of *The Broad Ax*.

Rather than pre-selecting a confidence score threshold, the Newspaper Navigator dataset contains all predictions with confidence scores greater than 5%,[20] allowing the user to define their own confidence cut when querying the dataset. However, the website for the Newspaper Navigator dataset also includes hundreds of pre-packaged datasets in order to make it easier for users to work with the dataset. In particular, users can download zip files containing all of the visual content of a specific type with confidence scores greater than or equal to 90%, for any year from 1850 to 1963. I made this choice of 90% as the threshold cut for these pre-packaged datasets based on heuristic evidence from inspecting sample pre-packaged datasets by eye. However, as articulated above, based on different use cases, this cut of 90% may be too restrictive or permissive: relevant visual content may be absent from the pre-packaged dataset or lost in a sea of other examples. In Figure 2.9, I show the visual content recognition

[20]This modest cut is provided to remove the large number of predictions with confidence scores between 0% and 5%, which have high false-positive rates, and thus reduce the size of the Newspaper Navigator dataset.

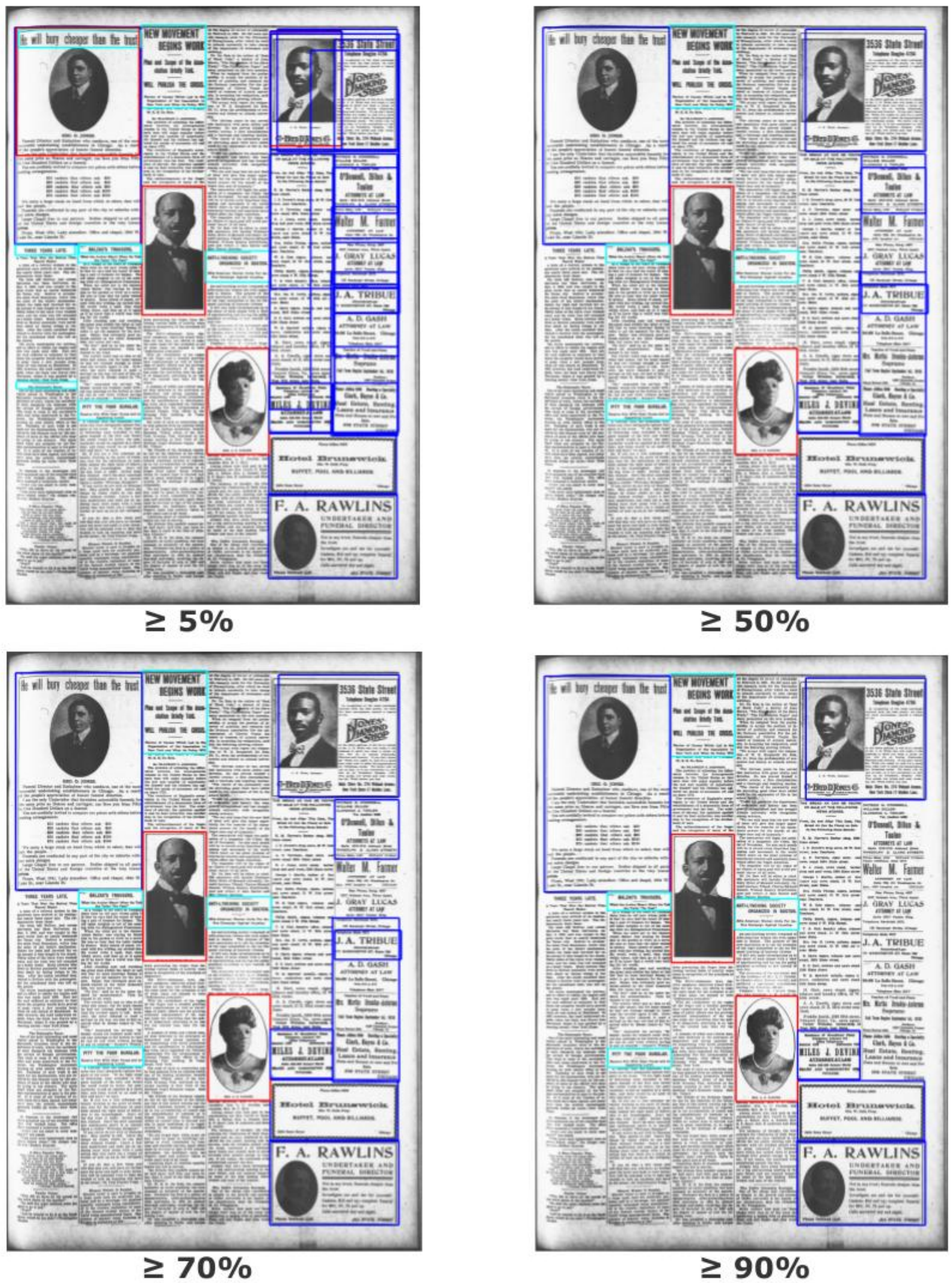

Figure 2.8: The same page of *The Broad Ax* from November 26, 1910, along with predictions from the visual content recognition model, thresholded on confidence score at 5%, 50%, 70%, and 90% [265, 270]. Note that red corresponds to a prediction of "photograph," cyan corresponds to a prediction of "headline," and blue corresponds to a prediction of "advertisement."

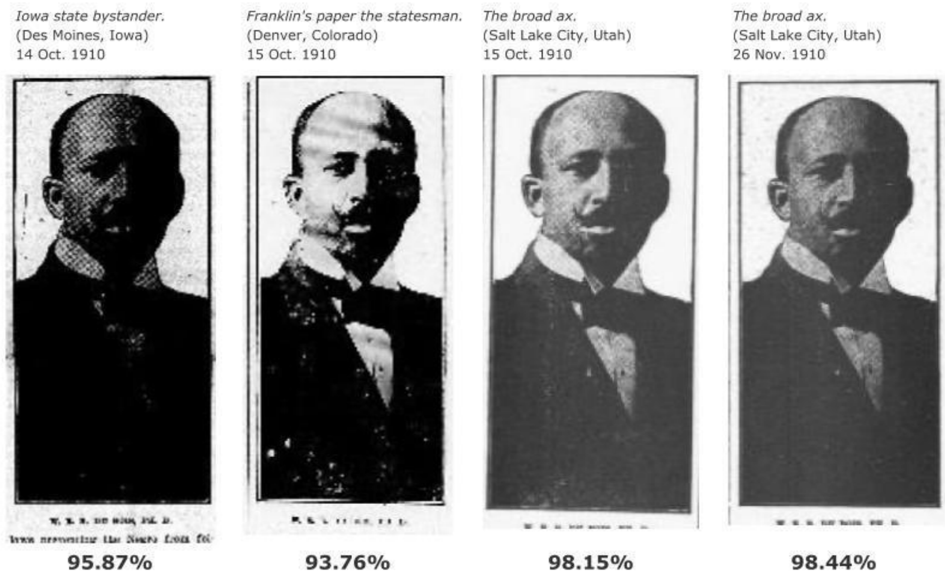

Figure 2.9: The visual content recognition model's confidence score for each of the four images of W.E.B. Du Bois. Note how the model assigns a different confidence score to each identified image [268, 269, 270, 271].

model's confidence scores for the four images of W.E.B. Du Bois described throughout this data archaeology. The effect of a cut on confidence score can be seen here: selecting a cut of 95% would exclude the image from *Franklin's Paper the Statesman*. I raise this point to emphasize that even this seemingly innocuous choice of 90% for the pre-packaged datasets alters the discovery process and thus can have an impact on scholarship.

Just as the bounding box predictions themselves are affected by the training data, as well as newspaper page layout, date of publication, and noise from the digitization pipeline, so too are the confidence scores. In particular, the visual content recognition model suffers from high-confidence misclassifications, for example, crossword puzzles that are identified as maps with confidence scores greater than 90%. High-confidence misclassifications pose challenges for machine learning writ large, and the field of explainable artificial intelligence is largely devoted to developing tools for understanding this type of misclassification [392]. However, these high-confidence misclassifications can often be traced back to the composition of the training set. For example, the fact that the visual content recognition model sometimes identifies crossword puzzles as maps with high confidence is likely due to the fact that the training data did not contain enough labeled examples of maps and crossword puzzles for the visual content recognition model to differentiate them with high accuracy.

The questions surrounding confidence scores and probabilistic descriptions of items is by no means restricted to the Newspaper Navigator dataset. I echo Thomas Padilla's assertion that "attempts to use algorithmic methods to describe collections must embrace the reality that, like human descriptions of collections, machine descriptions come with varying measure of certainty" [284]. Machine-generated metadata such as OCR are also fundamentally probabilistic in nature; this fact is not immediately apparent to end users of cultural heritage collections because cuts on confidence score are typically chosen before surfacing the metadata. Effectively communicating confidence scores, probabilistic descriptions, and the decisions surrounding them to end users remains a challenge for content stewards.

2.3.9 OCR Extraction

In the Newspaper Navigator pipeline, a textual description of each prediction is obtained by extracting the OCR within each predicted bounding box. The resulting textual description is thus dependent on not only the OCR provided by *Chronicling America* but also the exact coordinates of the bounding box: if the coordinates of a word in the localized OCR extend beyond the bounds of the box, the word is excluded. I experimented with utilizing tolerance limits to allow words that extend just beyond the bounds of the boxes to be included, but doing so ultimately introduces false positives as well, as words from neighboring articles or visual content were inevitably included some fraction of the time. Once again, the tradeoff between false positives and false negatives is manifest.

In Figure 8, I show the textual descriptions of the four images of W.E.B. Du Bois, as identified by the Newspaper Navigator pipeline. Significantly, in the Newspaper Navigator dataset, the OCR is stored as a list of words, with line breaks removed; these lists are what appear in Figure 2.10. These four examples provide intuition as to how the captions are altered. While the examples from the *Iowa State Bystander* and *Franklin's Paper the*

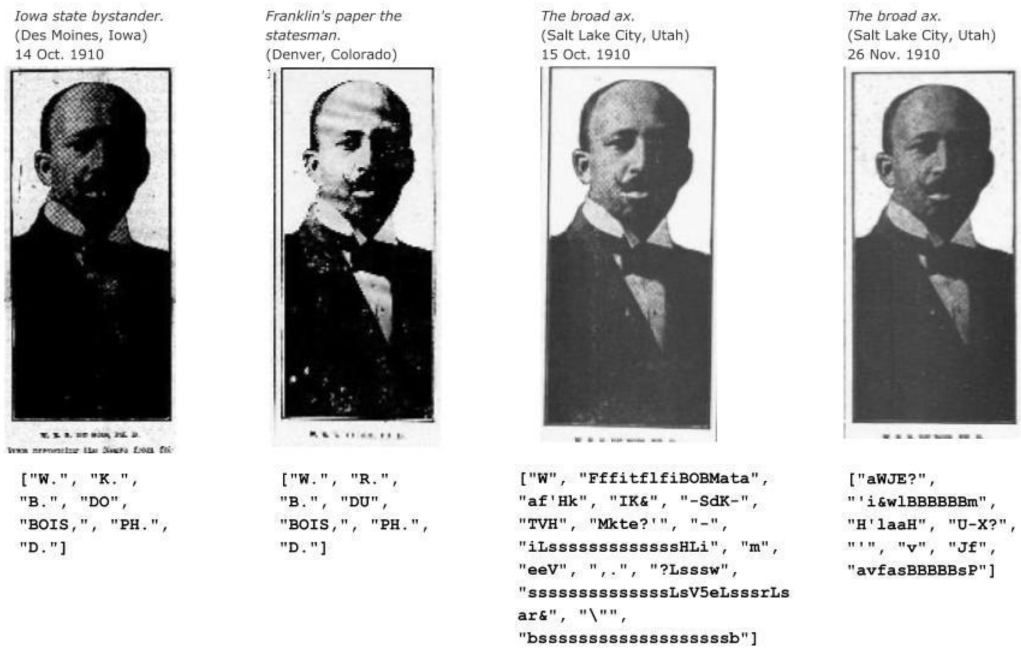

Figure 2.10: The textual descriptions of each image, as extracted from the OCR and saved in the Newspaper Navigator dataset [268, 269, 270, 271].

Statesman both have the same captions as shown in Figure 2.5, the captions for both of the examples from *The Broad Ax* are unrecognizable. Because the bounding boxes have clipped the caption, none of the characters from the proper OCR captions from Figure 2.5 are present. Furthermore, the captions contain OCR noise due to the OCR engine attempting to read text from the photographs. Consequently, the mentions of W.E.B. Du Bois are erased from the textual descriptions in the Newspaper Navigator dataset. The visual content in the Newspaper Navigator dataset is thus decontextualized not only in the sense that the visual content is extracted from the newspaper pages but also in the sense that the OCR extraction method further alters the textual descriptions. While the images from the *Iowa State Bystander* and *Franklin's Paper the Statesman* are still recoverable with fuzzy keyword search, the two images from *The Broad Ax* are impossible to retrieve with any form of keyword search, revealing another instance in which employing automated techniques for collections processing affects discoverability.

Fortunately, visual content can still be recovered using similarity search over the images themselves; these methods are discussed in detail in the next subsection. However, in the case of headlines, the errors introduced by OCR engines and the subsequent OCR extraction have no recourse, as similarity search for images of headlines would only capture similar typography and text layout.[21]

To illustrate the effects of this OCR extraction on headlines, I reproduce in Table 2.5 the extracted OCR as it appears in the Newspaper Navigator dataset for Franklin F. Johnson's headline:

NEW MOVEMENT
BEGINS WORK
Plan and Scope of the Asso-
ciation Briefly Told.
Will Publish the Crisis.
Review of Causes Which Led to the
Organization of the Association in
New York and What Its Policy Will
Be-Career and Work of Professor
W.E.B. Du Bois

The full pages are reproduced in the appendix for reference. Notably, all four extracted headlines contain OCR errors, as well as missing words due to the OCR extraction. The visual content recognition model consistently fails to include the last line of the headline, "W.E.B. Du Bois," revealing another case in which Du Bois's name is rendered inaccessible by keyword search in the Newspaper Navigator dataset.

2.3.10 Image Embeddings

The persistent question of biases in OCR, which motivated much my earlier analysis in this data archaeology surrounding discoverability, encourages scholars to consider other modes

[21] The Newspaper Navigator dataset does not retain the cropped images of headlines, as the textual content is more salient than visual snippets in the case of headlines.

Table 2.5: The extracted OCR associated with each of the four photographs of W.E.B. Du Bois [268, 269, 270, 271].

Iowa State Bystander (14 Oct. 1910)	**Franklin's Paper the Statesman (15 Oct. 1910)**	**The Broad Ax (15 Oct. 1910)**	**The Broad Ax (26 Nov. 1910)**
98.72%	99.57%	99.76%	99.70%
[“NEW”, “MOVEMENT”, “BEGINS”, “WORK”, “and”, “Plan”, “Scope”, “of”, “the”, “Asso\u00ad”, “ciation”, “Briefly”, “Told.”, “WILL”, “PUBLISH”, “THE”, “CRISIS.”, “Review”, “of”, “Causae”, “Which”, “Lad”, “to”, “the”, “Organisation”, “of”, “the”, “Auooiation”, “In”, “Naw”, “York”, “and”, “JWhat”, “It*”, “Polioy”, “Will”, “Ba\u2014Career”, “and”, “Wark”, “of”, “Profeasor”]	[“NEW”, “MOVEMENT”, “BEGINS”, “WORK”, “Plan”, “and”, “Scope”, “of”, “the”, “Asso”, “ciation”, “Briefly”, “Told.”, “WILL”, “PUBLISH”, “THE”, “CRISIS.”]	[“NEW”, “MOVEMENT”, “BEGINS”, “WORK”, “Plan”, “and”, “Sep”, “if”, “the”, “Asso”, “ciation”, “Briefly”, “Told.”, “WILL”, “PUBLISH”, “THE”, “CRISIS,”, “Be”, “Career”, “nnd”, “Work”, “of”, “Professor”, “W.”, “E.”, “B.”, “Du”, “Bois.”, “Review”, “of”, “Causes”, “Which”, “Led”, “to”, “the”, “Oraantealion”, “of”, “th.”, “A.Me!.!?n”, “i”, “i”, “New”, “York”, “and”, “What”, “IU”, “Policy”, “Will”]	[“NEW”, “MOVEMENT”, “BEGINS”, “WORK”, “Plan”, “and”, “Scope”, “of”, “the”, “Asso”, “ciation”, “Briefly”, “Told.”, “WILL”, “PUBLISH”, “THE”, “CRISIS.”, “Review”, “of”, “Causes”, “Which”, “Lad”, “to”, “tha”, “Organization”, “of”, “the\””, “Association”, “In”, “New”, “York”, “and”, “What”, “Its”, “Policy”, “Will”]

of search, including visual search with image embeddings. An image embedding canonically refers to a low-dimensional representation of an image, often a list of a few hundred or a few thousand numbers, that captures much of the image's semantic content. Image embeddings are typically generated by feeding an image into a pre-trained neural image classification model (i.e., a model that takes in an image and outputs a label of "dog" or "cat") and extracting a representation of the image from one of the model's hidden layers, often the penultimate layer.[22] Image embeddings are valuable for three reasons:

1. Image embeddings are remarkably adept at capturing semantic similarity between images. For example, images of dogs tend to be clustered together in embedding space, with images of bicycles in another cluster and images of buildings in yet another. These clusters can be fine-grained: sometimes, the red bicycles are grouped closer together than the blue bicycles.

2. Image embeddings can be constructed by feeding images into an image classification model already trained on another dataset (such as ImageNet), meaning that generating image embeddings is a useful method for comparing images without having to construct training data by labeling images.

3. Image embeddings are low-dimensional and thus much smaller in size than the images themselves (i.e., on the order of kilobytes instead of megabytes). As a result, image embeddings are much less computationally expensive to compare to one another when conducting similarity search, clustering, or related tasks. In short, image embeddings speed up image comparison.

Utilizing image embeddings to visualize and explore large collections of images has become an increasingly common approach among cultural heritage practitioners [394]. Projects and institutions that have utilized image embeddings for visualizing cultural heritage collections include the Yale Digital Humanities Lab's PixPlot interface [90], the National Neighbors project [213], Google Arts and Culture [84], The Norwegian National Museum's Principal Components project [129], the State Library of New South Wales's Aero Project [116], the Royal Photographic Society [380], The American Museum of Natural History [106], and

[22] If these words are unfamiliar, the three takeaways listed are more important.

The National Library of the Netherlands [217, 394]. These visualizations provide insights into broader themes in the collections, thereby allowing curators, researchers, and the public to explore collections at a scale previously only possible by organizing images by color or other low-level features.[23] In this regard, image embeddings provide new affordances for searching over images that complement canonical faceted and keyword search.

Because these image embeddings enable these visualization approaches and open the door to similarity search and recommendation, I opted to include image embeddings as part of the Newspaper Navigator pipeline. Indeed, these image embeddings power the similarity search functionality in the Newspaper Navigator user interface and, in this regard, are crucial to the broader vision of the project [205].[24] To generate the embeddings, I utilized ResNet-18 and ResNet-50, two variants of a prominent deep learning architecture for image classification, both of which had already been pre-trained on ImageNet [134].

ImageNet is perhaps the most well-known image dataset in the history of machine learning. Constructed by scraping publicly available images from the internet and recruiting Amazon Mechanical Turk workers to annotate the images, ImageNet contains approximately 14 million images across 20,000 categories [82]. Kate Crawford and Trevor Paglen's essay "Excavating AI: The Politics of Images in Machine Learning Training Sets" offers a history and incisive critique of the classification schema of ImageNet; here, I will summarize the most salient critiques. First, many of the categories in the taxonomy utilized are themselves marginalizing [75] Though many of the classes relating to people were removed in 2019, ImageNet had previously bifurcated the "Natural Object > Body > Adult Body" category into "Male Body" and "Female Body" subcategories. Second, ethnic classes were included, implying that 1) classification into rigid categories of ethnicity is possible and appropriate and 2) a machine learning system could learn how to classify ethnicity from these images. Diving deeper, the classifications become horrifying in their supposed granularity: until 2019, an image of a woman in a bikini was accompanied with the tags "slattern, slut, slovenly woman, trollop" [75] Though many embedding models are pre-trained on subsets

[23]For an introduction to some of these methods with lower-level features, see [231].

[24]The search application can be found at: `https://news-navigator.labs.loc.gov/search.`

of ImageNet categories included in the ImageNet Large Scale Visual Recognition Challenge that elide these particularly troubling classifications, these classifications nonetheless necessitate a reckoning with our use of ImageNet writ large, especially in regard to how the semantics of ImageNet is projected onto any image embedding generated with such a model [82].[25]

However, questions probing the data in ImageNet fail to critique the ethically questionable practices on which ImageNet is built. Though the researchers responsible for the dataset scraped all 14 million images from public URLs, ImageNet does not provide any guarantees on image copyright, as only the URLs are provided in the database: "The images in their original resolutions may be subject to copyright, so we do not make them publicly available on our server" [150]. It is highly unlikely that a photographer with an image in the dataset could have known that a photograph could be used this way, much less actively consent to the image's inclusion, as is the case with subjects in the photographs. Furthermore, the labels themselves were collected using Amazon's Mechanical Turk platform, which has been repeatedly criticized for its exploitative labor practices: as of 2017, workers earned a median wage of approximately $2 an hour on the platform [130]. Scholars including Natalia Cecire, Bonnie Mak, and Paul Fyfe have highlighted how outsourced marginalized labor underpins digitization efforts, and the reliance on Mechanical Turk for the production of ImageNet further entrenches the digitization and discovery process within a system of labor exploitation [53, 113, 230]. As cultural heritage practitioners and humanities researchers, we must acknowledge these exploitative practices, and we must reckon with how we perpetuate them through the use of ImageNet as a training source for image search and discovery.

In offering these critiques, my intention is not to dismiss ImageNet in a wholesale manner. Certainly, the benefits of utilizing ImageNet are manifold, as evidenced by widespread community adoption, as well as new affordances for searching cultural heritage collections enabled by the dataset that are shaping the contours of digital scholarship. In the case of my own scholarship with Newspaper Navigator, I have elected to utilize machine learning models pre-trained on ImageNet precisely for these reasons. I offer these provocations instead to

[25] The specific categories used in the challenge can be found at: `http://image-net.org/challenges/LSVRC/2010/browse-synsets`.

question how we can do better as a community, not only in imagining alternatives but in bringing them to fruition. Classification is an act of interpretive reduction, whether by human or machine, and thus manifests all too often as an act of oppression.[26] And yet, the structure imposed by classification constitutes the very basis for search and discovery systems. The salient question is thus not how we dispense of these systems but rather how we progressively realize a more inclusive vision of these systems, from the labor practices behind their construction to the very classification taxonomies themselves.

How, then, do image embeddings derived from ImageNet mediate our interactions with the photographs in Newspaper Navigator? Figure 2.11 shows a visualization of 1,000 photographs from the Newspaper Navigator dataset published during the year 1910. This visualization was created using the ResNet-50 image embeddings, as well as a dimensionality reduction algorithm known as T-SNE [376]. With T-SNE, a cluster of photographs indicates that the photographs are likely semantically similar, but the size of the cluster and distances from other clusters bear no meaning [390]. With this in mind, we can examine the clusters. Despite the fact that the high-contrast, grayscale photographs in Newspaper Navigator are markedly different, or "out-of-sample," in comparison to the clear, color images in ImageNet, the clusters nonetheless capture semantic similarity. In Figure 2.11, we observe the clustering of photographs depicting crowds of people, as well as photographs depicting ships and the sea. This visualization technique with the image embeddings is thus powerful in helping to navigate large collections of photographs by their semantic content.

What about the photographs of W.E.B. Du Bois? In Figure 2.12, I show the clusters containing these four photographs. This visualization affords us a lens into the limitations of image embeddings. First, it is evident that image embeddings are directly impacted by the distortions of the digitization process: while the three photographs from *Franklin's Paper the Statesman* and *The Broad Ax* are clustered together with other portraits, the photograph from the *Iowa State Bystander* is located in an entirely different cluster - a consequence of the fact that the *Iowa State Bystander* photograph is saturated and that W.E.B. Du Bois's facial features are obscured (notably, neighboring photographs suffer from similar distor-

[26] For more reading on this topic, see [45].

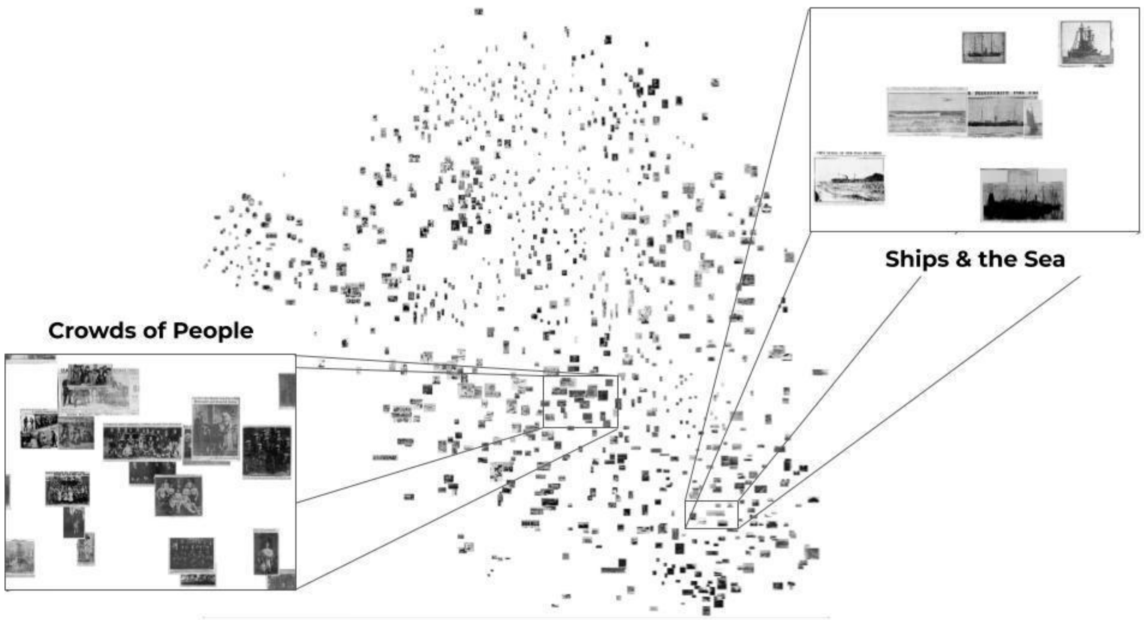

Figure 2.11: A visualization of 1,000 photographs from the year 1910 in the Newspaper Navigator dataset, generated using the Newspaper Navigator ResNet-50 image embeddings.

tions). A search engine powered with these image embeddings would in all likelihood return the three photographs from *Franklin's Paper the Statesman* and *The Broad Ax* together, but the fourth photograph would effectively be lost. This algorithmic mediation is particularly troubling because, as described earlier, the microfilming digitization process causes newspaper photographs of darker-skinned people to lose contrast. While this loss in image quality is marginalizing in its own right, image embeddings perpetuate this marginalization: digitized newspaper portraits of darker-skinned individuals are more likely to suffer from saturated facial features, in turn resulting in these photographs being lost during the discovery and retrieval process, as is the case with the saturated *Iowa State Bystander* photograph of W.E.B. Du Bois in Figure 2.12. Understanding these limitations of image embeddings are particularly relevant in the case of Newspaper Navigator, as these image embeddings power the visual similarity search affordance within the publicly-deployed Newspaper Navigator search application [205]. Though machine learning methods are often offered as panaceas

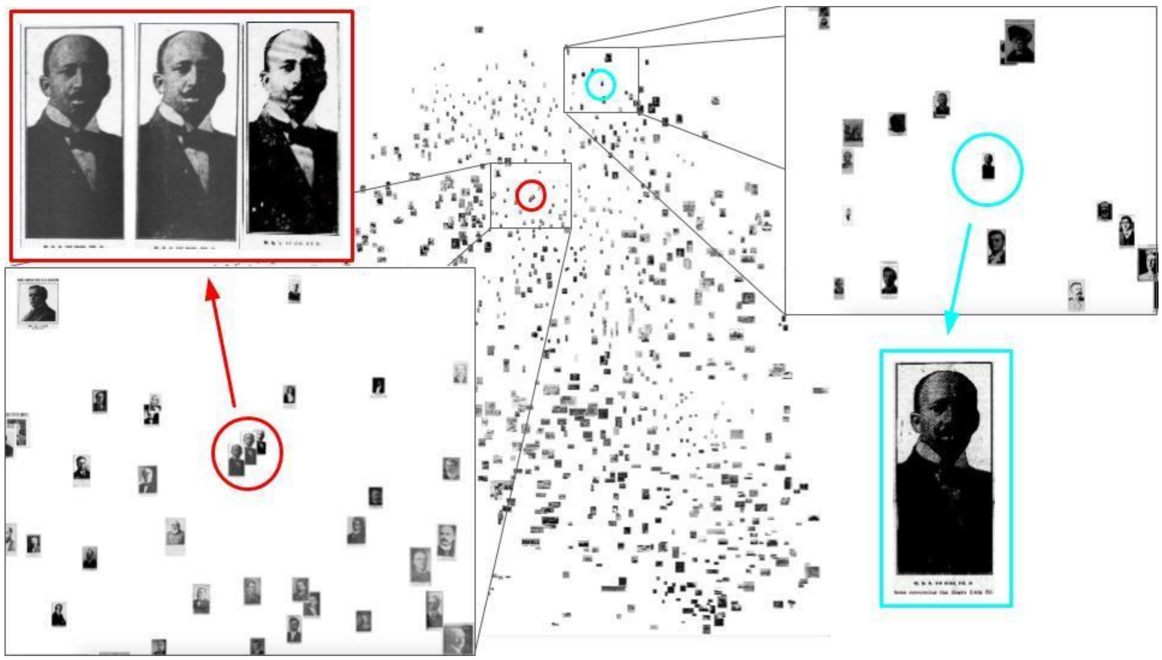

Figure 2.12: Figure 10. The same visualization as in Figure 2.11, this time showing the locations of the four photographs of W.E.B. Du Bois. .

for automation, this algorithmic erasure reminds us that traditional methods of scholarship and historiography, such as detailed analyses and close readings of Black newspapers in *Chronicling America*, are more important than ever to counter algorithmic bias.

2.3.11 Environmental Impact

Any examination of a dataset whose construction required large-scale computing would be remiss in not investigating the environmental impact of the computation itself. The carbon emissions generated from training a state-of-the-art machine learning model such as BERT is comparable to a single flight across the United States; however, factoring in experimentation and tuning, the carbon emissions can quickly amount to the carbon emissions of a car over its entire lifetime, including fuel [334]. OpenAI's GPT-3 model required several thousand petaflop/s-days to train; without specific numbers, the carbon emissions are not possible to calculate exactly, but they are nonetheless substantial [46]. In response, machine learning

Table 2.6: Carbon emissions from the GPU usage for Newspaper Navigator, broken down by project component. Note that all computation was done on Amazon AWS g4dn instances in the zone "us-east-2." The carbon emissions were calculated using the Machine Learning Impact Calculator [190].

GPU Emissions

Activity	# of NVIDIA T4 GPUs	GPU Hours (each)	Carbon Emissions
Training	1	19	0.96 kg CO2
Pipeline Processing	8	456	144.56 kg CO2
Experimentation for Training and Pipeline Processing (estimate)	8	24	7.66 kg CO2
Total	-	-	153.18 kg CO2

researchers have recommended ideas such as "Green AI," with the goal of encouraging the community to value computational efficiency and not just accuracy [334].

In the case of Newspaper Navigator, most of the compute time was devoted to processing all 16.3 million *Chronicling America* pages with the visual content recognition model, as opposed to training the model itself. In Tables 2.6 and 2.7, I report details on training the model and running the pipeline, as well as the carbon emissions generated by each step, computed using the Machine Learning Impact Calculator [190]. In total, approximately 380 kg CO2 were emitted during the construction of the Newspaper Navigator dataset, including development, experimentation, training, pipeline processing, and post-processing. It should be noted that this number is an estimate, as the statistics for experimentation and post-processing are difficult to quantify exactly. Nonetheless, this is approximately equivalent to the carbon emissions incurred by a single person flying from Washington, D.C., to Boston [107]. I include these numbers in the hope that cultural heritage practitioners will consider the environmental impact of utilizing machine learning and artificial intelligence for digital content stewardship. Doing so is essential to the data archaeology: given that climate change will disproportionately affect cultural heritage institutions in regions unable to develop proper infrastructure to withstand rapid temperature fluctuations and unprecedented flooding, even the environmental impacts of utilizing machine learning within digital content stewardship has the capacity to contribute to erasure and marginalization.

Table 2.7: Carbon emissions from the CPU usage for Newspaper Navigator, broken down by project component. Note that all computation was done on Amazon AWS g4dn instances in the zone "us-east-2." The CPU processors are all 2nd generation Intel Xeon Scalable Processors (Cascade Lake) [151]. The 48-core processor outputs approximately 350 W; the 4-core processor outputs approximately 104 W [153, 154]. The carbon emissions were calculated using the Machine Learning Impact Calculator [190]. Note that the energy consumption by RAM is not factored in, but it is insignificant in comparison to the CPU and GPU energy consumption.

CPU Emissions

Activity	**CPU Processor (#)**	**# of Processor CPU Cores**	**CPU Hours (each)**	**Carbon Emissions**
Training	1	4 CPUs	19	1.13 kg CO2
Pipeline Processing	2	48 CPUs	456	181.9 kg CO2
Experimentation for Training and Pipeline Processing (estimate)	2	48 CPUs	24	9.57 kg CO2
Extra Computation (dataset post-processing, etc., estimate)	1	48 CPUs	168	33.52 kg CO2
Total	-	-	-	226.12 kg CO2

2.3.12 Conclusion

In this data archaeology, I have traced four *Chronicling America* pages reproducing the same photograph of W.E.B. Du Bois as they have traveled through the *Chronicling America* and Newspaper Navigator pipelines. The excavated genealogy of digital artifacts has revealed the imprintings of the complex interactions between humans and machines. Indeed, the journey of each newspaper page through the *Chronicling America* and Newspaper Navigator pipelines is one of refraction, mediation, and decontextualization that is compounded upon with each step. Decisions made decades ago when microfilming a newspaper page inevitably affect how the machine learning models employed for OCR, visual content extraction, and image embedding generation ultimately process the pages, render them as digital artifacts in the Newspaper Navigator dataset, and mediate their discoverability.

As articulated by Trevor Owens in *The Theory and Craft of Digital Preservation*, machine learning and artificial intelligence are the "underlying sciences for digital preservation"

[281]. Though machine learning techniques provide us with new affordances for searching and studying cultural heritage materials, they have the power to perpetuate and amplify the marginalization and erasure of entire communities within the archive. This erasure, coupled with the labor practices involved in creating training data as well as the environmental impact of training and deploying machine learning models in large-scale digitization pipelines, necessitates that we continue to examine the broader socio-technical ecosystems in which we participate. In doing so, we can work toward a more inclusive vision of the digital collection and the ways in which we render its contents discoverable.

How, then, is Newspaper Navigator situated within this vision? In reimagining how we search over the visual content in *Chronicling America*, one explicit goal of the project is to engage the public with the rich history preserved within historic American periodicals and thus build on *Chronicling America* as a free-to-use, public domain resource for scholars, educators, students, journalists, genealogists, and beyond [200, 203]. With Newspaper Navigator, it is my belief that the new modes of interacting with *Chronicling America* have the capacity to not only enable a breadth of new scholarship but also foster engagement in and reckoning with America's multilayered history of oppression. In documenting the different components of the project with this data archaeology and corresponding technical paper [202], as well as releasing the full dataset and all code into the public domain, I have intended to be as transparent as possible with the tools and methodologies employed. Newspaper Navigator is not without its shortcomings, but my hope is that the project contributes to this vision of the digital collection through transparency and inclusivity, as well as the scholarship and pedagogy that it has enabled.

I offer this case study not only to contextualize the Newspaper Navigator dataset but also to advocate for the autoethnographic data archaeology as a valuable apparatus for reflecting on a cultural heritage dataset from a humanistic perspective. Though the digital humanities community has yet to adopt the data archaeology as standard practice when creating and releasing cultural heritage datasets, doing so has the capacity to improve accountability and context surrounding applications of machine learning for both practitioners and end users. Given the manifold ways in which machine learning mediates access to the archive and perpetuates erasure, reflecting critically on these systems is not only urgent but essential


IOWA STATE BYSTANDER.

VOL. XVII, No 18. DES MOINES, IOWA, FRIDAY, OCTOBER 14, 1910. Price Five Cents.

CITY NEWS.

NEW MOVEMENT BEGINS WORK

Plan and Scope of the Association Briefly Told.

WILL PUBLISH THE CRISIS.

Restaurants Change Hands.

A New Undertaking Firm.

THE VARIETY

ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON.

The New Wardrobe

Figure 2.13: *Iowa state bystander. [volume]* (Des Moines, Iowa), 14 Oct. 1910. *Chronicling America*: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn83025186/1910-10-14/ed-1/seq-1/

for transparency and inclusivity.

PAGE 16 THE STATESMAN, DENVER, COLORADO.

From Union Comes Strength

An Injury to One is the Concern of All

Member Representatives Wanted

INTERNATIONAL LIBERTY UNION

370

NEW MOVEMENT BEGINS WORK

Plan and Scope of the Association Briefly Told.

WILL PUBLISH THE CRISIS.

Review of Causes Which Led to the Organization of the Association in New York and What Its Policy Will Be—Career and Work of Professor W. E. B. Du Bois.

By FRANKLIN F. JOHNSON.

The problem of adjusting the differences between the heterogeneous masses in the United States so that each man will enjoy equality before the law and equality of opportunity for himself and his children has agitated the minds of statesmen oft and on since 1776.

The overthrow of England's sovereignty in the United States and the establishment of a democratic form of government was the first. The existence of slavery, the agitation against it and its final extinction was another step in the recognition of the brotherhood of man.

The enfranchisement of the Negro and the abrogation of many of the laws preventing the Negro from following various fields of activity were steps in recognition of the principles of democracy.

The disfranchisement laws, jim-crow measures, restricted school facilities for the race, are efforts to undo the start made for the extension to all of what many are pleased to term "Jeffersonian principles."

With the curtailment of the rights of the Negro and a steadily growing race consciousness as well as knowledge of public affairs and the growing economic independence of the race has come the conviction that the spirit of democratic principles is no less for the colored man than for the white.

A number of white and colored men, all firm believers in the principles of democracy and strong opposers to all temporizing with social prejudices and its manifestation in discriminating laws, have organized the National Association For the Advancement of Colored People, with offices in the Evening Post building, 20 Vesey street, New York.

The need of such an association can be best gleaned from an address which one of its leading promoters, Oswald Garrison Villard, editor of the New York Evening Post, delivered at the recent meeting in New York of the National Negro Business league. Said Mr. Villard:

W. E. B. DU BOIS, PH. D.

"My friends of the Business league, do not let the historian of the future say that in this money making age the colored American became so engrossed in the accumulation of means and the establishment of business as to forget those higher things without which he cannot hope to succeed and rise to great usefulness, to be really worthy of his heritage of American citizenship.

"Do not forget that, while every effort must be spent in fortifying ourselves in every community by business and material success, equal effort must be devoted to that of far greater movement, the insistence upon the Negro's equal and political rights in every place and at all times.

"Anything short of absolute equality before the law is slavery.

"My friends, this republic cannot exist half slave, half free, any more today than it could in the time of Abraham Lincoln."

Dr. W. E. B. Du Bois, the eminent authority on racial questions, has been engaged as secretary of the association. Professor Du Bois is a native of Massachusetts. After graduating from Fisk university in 1888 he entered Harvard university, receiving the degree of bachelor of arts in 1890 and the degree of master of arts the following year.

He was awarded a fellowship and after studying in Berlin, was award

President Young of the Florida Agricultural and Mechanical college is calling special attention of the colored citizens of the state to the fact that it is no longer necessary for them to send their children out of the state for advanced industrial and academic training.

The state is now furnishing them such training tuition free. All thoughtful, taxpaying citizens will avail themselves of this exceptional opportunity to educate their children.

As a matter of fact, the state is now offering them at the Agricultural and Mechanical college, Tallahassee, free instruction in twenty-three industries, including typewriting and stenography, a collegiate course in science and a course in pedagogy.

PHONE MAIN 5554.

Do you work for money? Why not let your money work for you. Ours works night and day, and we can place yours in the same position, talk it over with

THE COLORED AMERICAN LOAN & REALTY CO.

913 TWENTY-FIRST ST.

A. A. WALLER, Secretary and Manager

CURTIS M. HARRIS

Funeral Director

Phone Main 6743

A. M. LAWHORN

UNDERTAKER

Mrs. J. J. Stafford, Lady Assistant

A First-Class Mortuary Establishment.

FIRST AID TO THE BEREAVED IN THE TIME OF THE DEATH OF THEIR LOVED ONES

Parlors 1921 Arapahoe Street

Licensed Embalmer

Southern View of Northern Politics.

Politics all over the country is now the topic. Several states will hold important elections. The congressional elections take place in November. All eyes are turned to New York, awaiting the outcome of the fight of the progressives and the old guard. The outcome will have much to do with the future of the Republican party.—Savannah (Ga.) Tribune.

T. S. Rector is serving chili piping hot in two ways at his store, 1916 Arapahoe street.

This office is prepared to do all kinds of job printing. Estimates furnished. It is the only race enterprise of its kind operating, and if you desire its service call phone Main 7905. There is a difference between merely soliciting printing and actually doing the work. Get our prices and you will see that difference.

For Rent advertisements appear on page two except such as come to us too late for such positioning. These appear at a cost of 50c per month, or 5c per line if run by the week.

J. R. CONTEE, PRESIDENT,
R. E. HARDY, LICENSED EMBALMER.

The Douglass Undertaking Company

Phone Main 6123

Incorporated-Bonded to the city

1023 19th Street

Denver, Colorado

Figure 2.14: *Franklin's paper the statesman. [volume]* (Denver, Colo.), 15 Oct. 1910. *Chronicling America*: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn91052311/1910-10-15/ed-1/seq-16/


THE BROADAX

NEW MOVEMENT BEGINS WORK

CHIPS

Figure 2.15: *The broad ax. [volume]* (Salt Lake City, Utah), 15 Oct. 1910. *Chronicling America*: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-10-15/ed-1/seq-2/

He will bury cheaper than the trust

GEO. G. JONES.

NEW MOVEMENT BEGINS WORK

WILL PUBLISH THE CRISIS.

THREE YEARS LATE.

BALZAC'S TROUSERS.

PITY THE POOR BURGLAR.

ANTI-LYNCHING SOCIETY ORGANIZED IN BOSTON.

3536 State Street

O'Donnell, Dillon & Toolen

Walter M. Farmer

J. GRAY LUCAS

J. A. TRIBUE

A. D. GASH

MILES J. DEVINE

Hotel Brunswick

BUFFET, POOL AND BILLIARDS.

F. A. RAWLINS

UNDERTAKER AND FUNERAL DIRECTOR

Figure 2.16: *The broad ax. [volume]* (Salt Lake City, Utah), 26 Nov. 1910. *Chronicling America*: Historic American Newspapers. Lib. of Congress. https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-11-26/ed-1/seq-3/

2.4 Navigating the Mise-en-Page: Interpretive Machine Learning Approaches to the Visual Layouts of Multi-Ethnic Periodicals

This section presents a computational method of analysis that draws from machine learning, library science, and literary studies to map the visual layouts of multi-ethnic newspapers from the late 19th and early 20th century United States. This work departs from prior approaches to newspapers that focus on individual pieces of textual and visual content. Our method combines *Chronicling America*'s MARC data and the Newspaper Navigator machine learning dataset to identify the visual patterns of newspaper page layouts. By analyzing high-dimensional visual similarity, we aim to better understand how editors spoke and protested through the layout of their papers.

This work was in collaboration with Joshua Ortiz Baco, Sarah H. Salter, and Jim Casey, and is based on a publication that appeared in *Computational Humanities Research* (CHR) Conference 2021 [199]. A second article on this work is forthcoming in *Critism: A Quarterly for Literature and the Arts* [25].

2.4.1 Introduction

This section presents a set of developing methods for analyzing Black, Latinx, and other ethnic newspapers in the late nineteenth and early twentieth-century United States. We address these newspapers in this essay because multi-ethnic newspapers faced challenges above and beyond their era's respective forms of mainstream journalism. Contending with forces of racism and xenophobia often required editors to tailor their paper's visual grammar, communicating in carefully guarded ways that were still capable of circulating to distant readers and communities. Those traditions of complex editorial craft in the multi-ethnic presses represent an especially rich archive for analysis.

Editors in the historical multi-ethnic press often chose to proceed circumspectly by focusing on their newspaper's "mise-en-page." Rather than publish bold condemnations in headlines or editorials, it was frequently safer to rely on other affordances of newspapers provided by the aggregate, often invisible techniques collapsed under the heading of editing. [117]. Editorial techniques include the labor required to gather, arrange, and issue a newspaper as well as the organization of those labors behind the scenes. They might vary greatly by paper, time, and place. Readers might recognize traces of editorial techniques in a paper's use of headlines for emphasis, photographs to dramatize articles, or the separation of content into subsections. At times, these techniques are performed by dedicated editors: news editors, copyeditors, managing editors, or photo editors. We focus on the design of front pages, where many of the above-named editorial techniques intersect to create meaningful relationships between items on the page. This feature of multi-ethnic editorial communication makes it untenable to extract and analyze isolated textual or visual items from these newspapers.

Computational methods make it possible to analyze the visual patterns of multi-ethnic newspapers at scale. While prior research has approached these questions through textual analysis, a pivot to computational visual analysis compensates for the frequently noisy OCR text produced from historically neglected multi-ethnic newspapers.

Our working method uses machine learning to map the layouts of hundreds of thousands of newspaper pages, quantified by visual similarity, using the Newspaper Navigator dataset

Figure 2.17: A visualization of the front page of the May 22, 1917, issue of *La Prensa* [272], annotated with bounding boxes of visual content derived from the Newspaper Navigator dataset [202]. These bounding boxes are predictions made by a machine learning model showing, in this case, the locations of headlines, photographs, and maps (each bounding box includes the predicted class, as well as a confidence score, in the top left corner). In this section, we utilize these bounding boxes to compare newspaper titles according to visual similarity.

(as pictured in Figure 2.17) [202]. Visualizing and comparing layouts at scale provides a wayfinding function, helping to direct archival researchers to distinctive and meaningful places in large collections of digitized newspapers and cultural heritage collections writ large.

In this section, we describe work in progress that brings machine learning, library science, and literary studies into conversation. This work includes: (1) a dataset of visual features from 16.3 million pages of historical newspapers; (2) remediation in 207 catalog records for 309 ethnic editors; and (3) modes for reading the practical languages of editing in specific historical contexts. Prior research efforts blend together in our exploratory machine learning prototypes for quantitatively and qualitatively measuring the visual similarities of the formats of a newspaper page. These visual similarities can help illuminate the ideas and activism that flowed through the editorial craft and conduct of historical multi-ethnic newspapers. Moreover, our work formulating a "similarity score" affirms the promise of multi-disciplinary collaborations for improving access and analysis of larger-scale digitized newspaper collections.

2.4.2 Related Prior Work on Developing Computational Approaches to Multi-Ethnic Periodicals

Multiple fields are eager to explore the new possibilities for analyzing digitized collections of historical newspapers. Complementary conversations in computational, archival, and literary studies have begun to pursue those possibilities in rich but largely disparate ways. Silos persist even as many scholars make parallel use of the Library of Congress's *Chronicling America* and other national newspaper digital initiatives [52, 81, 181, 207, 275, 280, 289, 297, 396].

This work builds on multi-ethnic press scholarship and archival research in *Chronicling America* and other repositories. The prevailing view of political activism in historical multi-ethnic newspapers has tended to focus on the textual and visual content that drove demands for social justice [65, 244, 353, 397]. A complementary approach has focused on social histories, tracing the networks of a paper's contributors and readers [95, 109, 114, 319, 352]. What can be easy to miss, however, is that editors can use the format of a paper itself to voice critiques and protests. The formal nature of editors' craft is difficult to grasp in the meticulous approaches to archival research on a case-by-case basis. Our work aims to develop experimental machine learning applications to map the larger patterns of editorial practices and editing conventions that span languages, communities, and eras.

Conversations in library and archival studies are exploring how best to account for histories of race and ethnicity in metadata. Rather than being reduced to a theme or topic within a collection, specialists have rethought how race and ethnicity might work as organizing principles for collecting and describing historical materials. Race and ethnicity as categories of identity are historically contingent and socially constructed. That fluidity challenges the rigors of knowledge organization systems that impose stable categories and schemas. For some archivists, however, categories of race and identity productively serve as a form of provenance to guide collection development and access policies [32, 77, 406]. Many of these changes at the collection level are being addressed in the metadata and, more recently, through linked open data. Because the development of controlled vocabularies, such as the Library of Congress Subject Headings, and Machine Readable Cataloging (MARC) largely predate sensitivity to these histories in libraries and archives, metadata has become a vital space for rediscovering the historical editors of multi-ethnic newspapers [382].

At the same time, an overabundance of work with more "computationally robust," institutionally privileged white print collections has prompted preliminary conversations regarding the directions of computational humanities research. Indeed, from OCR engines that do not support indigenous languages or different dialects, to entire subfields of natural language processing and textual analysis that assume English as the default language of study, algorithmically-mediated erasure continues to be uncovered and detailed by scholars [5, 95, 128, 131, 161, 250, 324, 398, 407]. Many have advocated for the development of computational approaches for those collections that remain marginal. This work concerns one such corrective for digitized periodicals: utilizing machine learning to study visual content and page layout rather than textual content, circumventing the use of OCR engines entirely.

Computational scholars have isolated visual content within periodicals in order to produce datasets of isolated content. Humanistic scholars have utilized these atomized items for interpretation. We are instead interested in the visual layouts themselves and the relations they imply on the page. By treating relations and patterns within layouts as forms of interpretive content – reading them as a "visual grammar" of photographs, advertisements, headlines, and beyond – we are in effect exploring a publication's creative history, local contexts, and circulation networks.

2.4.3 Methods

Newspaper Navigator coincides with current trends in periodical studies. As Sean Latham and Robert Scholes argue in an influential essay, "We have often been too quick to see magazines [and other periodicals] as containers of discrete bits of information rather than autonomous objects of study" [191]. In practice, this means that periodical scholars tend to examine not only the individual pieces of content in a given publication, but also its overall construction, taking into account such architectural elements as a publication's formats, columns, and visual layout. The typical methods to apply this approach require intensive reading of individual newspaper pages. Researchers gradually develop an intuition about the general style and flavor of a certain publication. A publication's implicit patterns can be relatively intangible because periodical formats are necessarily fixed—to preserve continuity—and flexible—to accommodate new and changing scenarios. A paper looks like itself almost always, but not always. Even beyond academic inquiry, dedicated readers of any periodical often develop this kind of intuitive mental map of a publication's page designs and formats.

We saw the possibility of combining these two types of reading. On one hand, Newspaper Navigator makes it possible to chart visual features across the entirety of *Chronicling America*. On the other hand, the methods in periodical studies encourage attention to the juxtaposition of visual features on a page. Starting with these two methods, we developed a set of preliminary working questions. How do the visual features on the page of a newspaper interlock and mutually inform each other? If we look at all of those features holistically, what kinds of styles and ideas can we detect over the lifespan of a single newspaper? More broadly, what added perspective can we gain by using the Newspaper Navigator to trace the sum of newspaper layouts over hundreds of publications? Finally, can we establish a "baseline" for newspaper production in the later 19th and early 20th century United States that would allow us to see the critical departures made by multi-ethnic newspapers in the service of social justice during moments of elevated racism or xenophobia?

Methods I: Adapting MARC 21 Metadata for Research

Adapting the Newspaper Navigator dataset to answer these questions required us to collect additional data about the subsets of multi-ethnic newspapers within *Chronicling America*. This information was partially extracted from the bibliographic data of newspaper titles in the machine-readable cataloging (MARC) format. All records for newspapers in *Chronicling America* follow the cataloging standards of the Cooperative ONline SERials Program (CONSER) [273], providing a somewhat normalized set of features for sorting newspapers. Additionally, *Chronicling America*'s partnering institutions provide a "title essay" that describes and contextualizes each newspaper. These unstructured data has recently been the focus of efforts for metadata enhancement in remediation projects geared to identifying ethnic editors [279].

The last step in collecting and organizing our data was combining MARC fields and structured data generated from the text of title essays to find relevant publications. Metadata from the title essays allowed us to select only ethnic newspapers with identifiable editors. From this subset, we selected MARC fields used by cataloguers to record title changes, race, ethnicity, and languages associated with each publication. We initially hoped to find an indicator of visual change within the metadata for newspapers. We considered that serials with more than one language edition, language notes, or different titles over their publication runs would feature changes to visual and layout elements of newspapers to mark these relationships or changes.[27] Title changes, for example, are a cataloguing practice for linking preceding and succeeding titles of a publication that has had a substantial name change, which we assumed would also produce modifications to the graphic elements on the page. We were able to parse these metadata fields where the change inherent in

[27]The 780 and 785 fields indicate all preceding and succeeding titles for a publication, respectively. The 546 field describes the language of a resource while the 765 indicates the original language of a translated publication and the 775 captures other language editions of a publication.

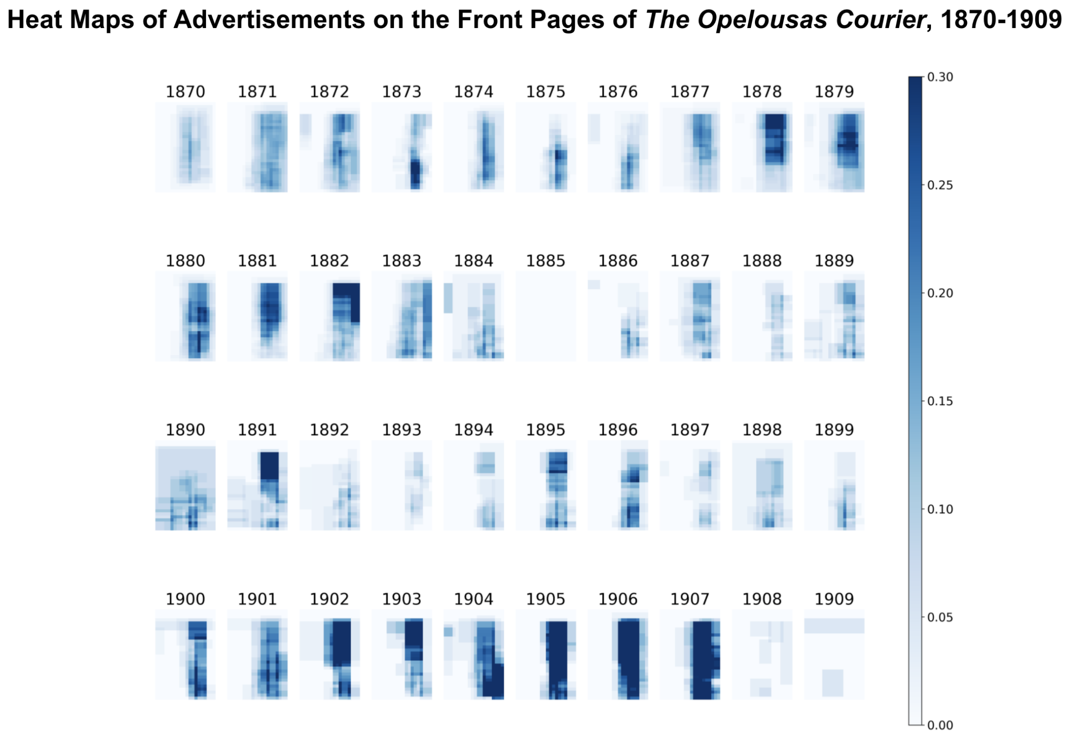

Figure 2.18: A heatmap of advertisements appearing on the front page of the Opelousas Courier, 1870-1909. Darker regions on the heatmaps correspond to a higher concentration of advertisement pages appearing in that region in aggregate over a given year.

the seriality of newspapers is recorded, which in turn provided parameters for capturing difference, similarity, and linking in an otherwise static knowledge organization system.

Methods 2: Quantifying and Visualizing the Mise-en-Page

The first visual output we experimented with were heat maps, created with Newspaper Navigator data that encodes the positions of visual features on a newspaper page. Inspired by PageOneX [73], these heat maps concisely summarize visual content patterns in an interpretable fashion. Figure 2.18 shows the locations and relationships of advertisements across four decades of *The Opelousas Courier*.[28] As an initial experiment in visual layout representation, the heat maps enabled us to trace the development of discrete visual features,

[28] *The Opelousas Courier* is described in *Chronicling America* as a Louisiana multilingual newspaper, with content in French and English. https://chroniclingamerica.loc.gov/lccn/sn83026389/marc/.

otherwise impossible to quantify by examining front pages by eye.

This approach facilitates several comparative opportunities. We rendered multiple visual categories into heat maps to understand the interplay between visual components in a single publication. We also utilized the heat map graphics to compare different newspapers using the same visual categories. The heat maps offered a useful preliminary system for studying newspaper layout as historical information.

Heat maps added greater nuance to our historical questions about newspaper layout. For example, what are the relationships between visual content, editorial practices, and print technology? What are the roles of regional and local context in predicting layout? How do we account for journalistic trends, such as the rise of syndicated content or nationally-circulating advertisements?

The heat maps showed the need for further exploration. In analyzing one publication at a time, the heat maps reproduced the methods of archival periodical studies, even in visually abstracted ways. While the MARC records helped us to identify the titles we were most interested in rendering into heat maps, this piecemeal approach to newspaper metadata and layouts proved only a first step.

Methods 3: Toward a Visual Constellation of Multi-Ethnic Newspapers

Our next experiment attempted to scale up possible representations to refocus on illustrating change and similarity across tens, hundreds, and even thousands of publications. Our resulting exploratory model utilized the quantified heat maps to define a metric over visual similarity. In particular, the distance between two newspaper titles can be captured by the residuals from subtracting two heatmaps from one another. As an initial approach, we segmented each newspaper title's run into individual years. For each title and year, we then treated the seven resulting heat maps (one for advertisements, one for headlines, etc.) collectively as a single, high-dimensional vector. In order to interpret the clusters in this high-dimensional space, we utilized T-SNE for dimensionality reduction [376].

This process resulted in exploratory constellations of newspaper titles, as shown in Figure 2.19. In this figure, we also show a close-up of one such "similarity cluster," revealing that

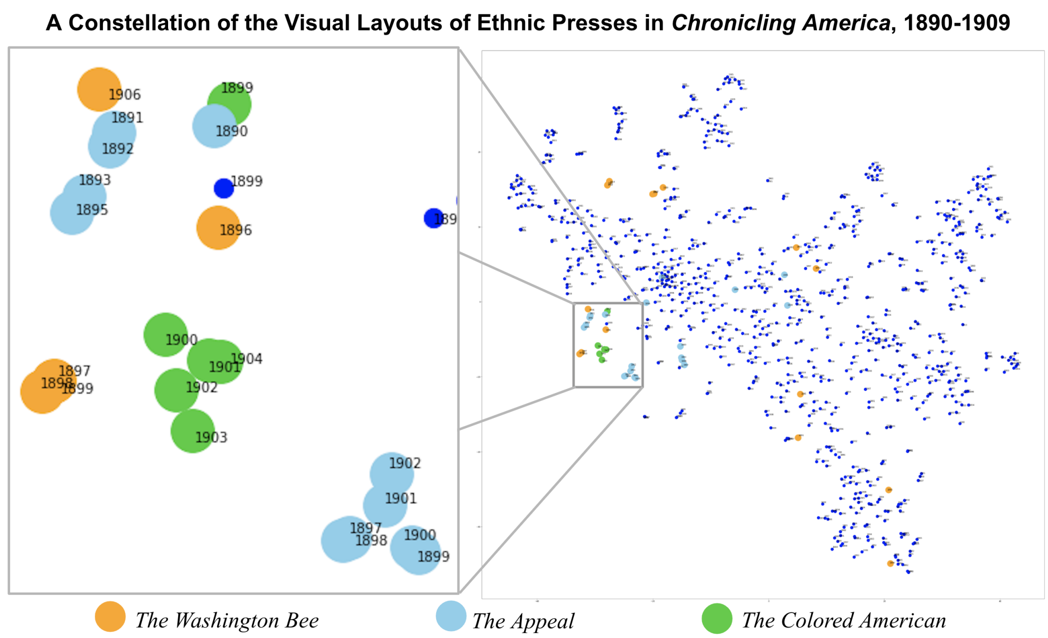

Figure 2.19: A 2-dimensional map of visual similarity across all ethnic titles in *Chronicling America*, filtered using [279]. Each point on the similarity map represents the composite front page of a given newspaper title for a given year, from 1890 to 1909 (individual years are labeled on the visualization). The magnified cluster reveals that *The Washington Bee*, *The Appeal*, and *The Colored American* are largely grouped together across years (individual front pages from each title are shown in Figure 2.20). Notably, all three titles are from the Black press.

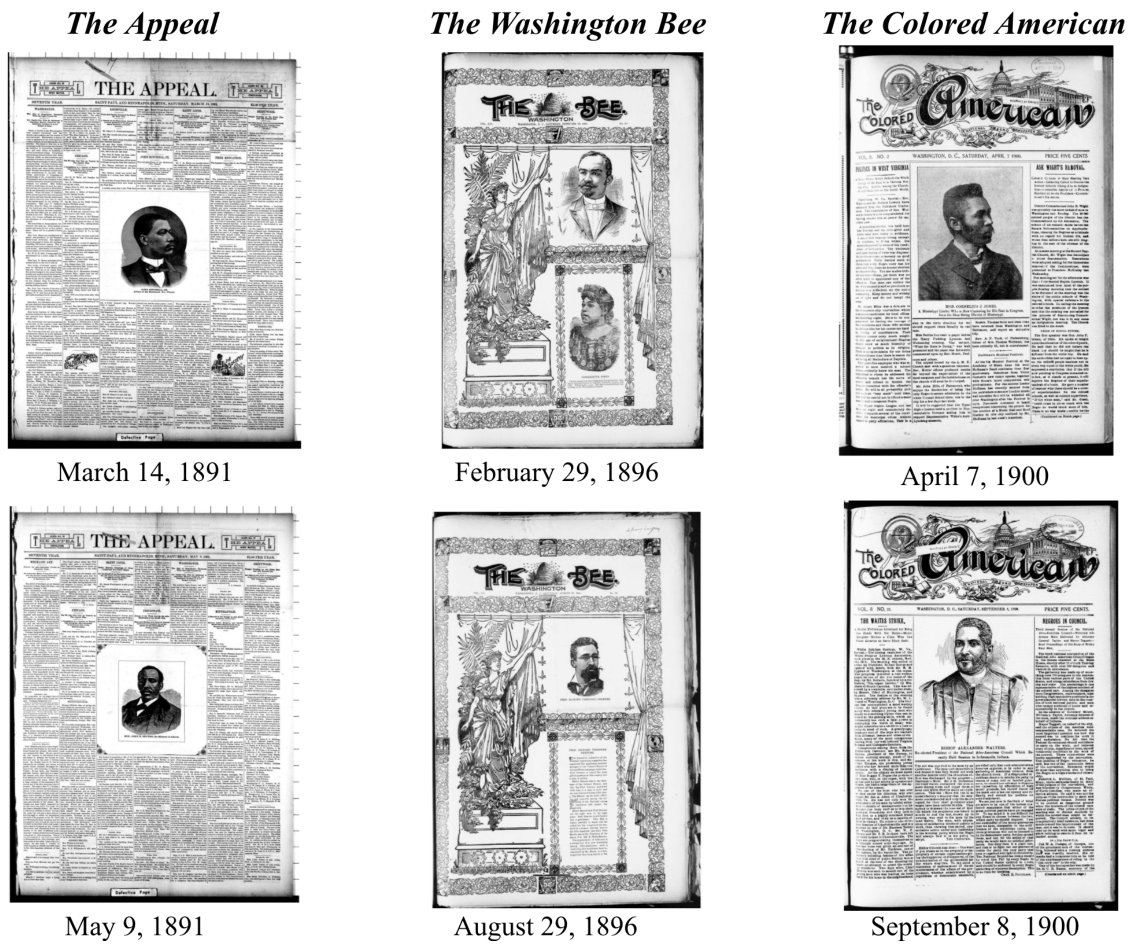

Figure 2.20: Sample front pages from *The Appeal*, *The Washington Bee*, and *The Colored American* from 1891, 1896, and 1900, respectively – all of which are title-year pairs appearing in the same cluster in Figure 2.19 [254, 255, 256, 257, 258, 259]. All six front pages feature visual content – namely, illustrations and photographs of individuals – prominently in the center of the page.

three Black newspapers (*The Washington Bee*, *The Appeal*, and *The Colored American*) are largely grouped together across years. In Figure 2.20, we show example front pages for these three titles from years included within the cluster. All six pictured front pages feature visual content on the center of the page. Notably, all six feature photographs and illustrations of individuals, a trend that is even more apparent when browsing these front pages across different issues. These images demonstrate that the similarity clusters, generated using Newspaper Navigator's rubric of seven content classes, accurately reflect common visual patterns of layout in distinct newspapers across the Black newspapers held in *Chronicling America*. The shared qualities across each of the three newspapers pictured in Figure 2.20 are suggestive. They are roughly contemporaneous African American newspapers that locate specific types of images near the center of the front page. These results indicate the capacity of the method to produce a multifaceted research question: what can we learn from the use of visual culture and portraiture in the material pages of late nineteenth-century African American newspapers?

These constellations of similarity provide a means to uncover small networks to focus on and interpret at a much more granular level. This is not a fully quantitative representation, but rather an exploratory framework for identifying and interpreting networks of related publications and the editorial practices that shape them. Given our overarching interest in the historical trajectories and meanings of newspaper formats, these visualizations show great promise. The constellation map is a model for how to convert raw bounding box information into interpretable provocations about historical similarity. Similarity clusters introduce visual layouts as clues to a shared context, inviting new research into the different historical traditions, technological moments, or political causes that comprise the histories of editorship.

2.4.4 Findings & Future Work

Our ongoing collaboration offers three central findings to date. First, this work finds that full level CONSER MARC records provide useful faceted metadata for accurately selecting multi-ethnic newspapers in machine-learning applications. Specifically, the most value comes from fields that register change over time, such as preceding and succeeding titles, and describe relationships between different editions of the same publication. Further metadata enhancement focused on identifying ethnic editors will significantly improve the interpretative value of visualizations and pattern recognition from large-scale collections like *Chronicling America*.

Second, this collaboration reveals how computational notions of layout similarity can be leveraged for humanistic inquiry. Using the machine learning-constructed Newspaper Navigator dataset of visual content, we have generated heat maps and constellation maps for exploring trends of visual layout within and across newspaper titles, respectively. These visualizations provide us with new visual grammars and affordances for excavating editorial practices.

Third, this work required us to find a collaborative working process for creating multidisciplinary methods of analysis that engage with machine learning, library science, and the humanities. Rather than trying to blend or transplant our respective research methods, we focused on creating a shared vocabulary that could speak to current research agendas in each of the relevant fields. This set of theoretical exchanges required iterative conversations, frequently pedagogical in nature, to find a method that is not merely extractive and static but generative and dynamic. This process will help guide the future growth of conversations in computational periodicals research.

Our roadmap for future work is guided by the provocations offered throughout this section. First, we plan to further utilize MARC data and the constellation maps in concert with one another in order to inform our understanding of both. The MARC data has the capacity to guide our navigation and understanding of the constellation maps; conversely, the constellation maps provide mechanisms for record enhancement by allowing scholars, catalogers, and librarians alike to sift through incomplete legacy records of ethnic publications. We offer this direction of future work in pursuit of uncovering editorship in ethnic presses. We plan to refine our similarity metrics and visualizations from a computational perspective in pursuit of this goal as well. The results of this research will drive new inquiry into the longer and varied histories of radical dissent and veiled protest in the multi-ethnic press.

Lastly, we note that a continuation of this collaboration is forthcoming in *Critism: A Quarterly for Literature and the Arts* [25].

2.5 *The Digital Humanities and the Ladino Press: Using Machine Learning to Extract and Analyze Visual Content in Historic Ladino Newspapers*

La Vara, *El Tiempo*, and *La Boz De Oriente* represent three of the major historic Ladino newspapers published across the diasporic Sephardic Jewish world in the 20th century – from New York to Constantinople to Istanbul. While Sephardic Jewish history and culture have received increasing scholarly attention in recent years, the vast corpus of Ladino newspapers largely remains unmined, and the field remains marginal from the perspective of Jewish Studies. A new collaboration between the Stroum Center for Jewish Studies' Sephardic Studies Program and the Paul G. Allen School for Computer Science & Engineering at the University of Washington seeks to draw on innovative machine learning techniques to render the visual content of Ladino newspapers more accessible to scholars and students alike and, in so doing, change the trajectory of Sephardic Studies writ large. This chapter reports the findings of this research.

Many Ladino titles have been digitized by the Sephardic Studies Program at the University of Washington. These Ladino newspapers contain not just articles and editorials but also an abundance of rich visual content that sheds light on Sephardic Jewish experiences in modernity. The advertisements appearing within Ladino newspapers have received attention from scholars within Sephardic Studies, and analysis thereof has revealed connections between the American Ashkenazic and Sephardic communities [238], as well as the ways in which advertisers' attempts to provide remedies speak to "readers' anxiety about the fragility of life under Ottoman rule" [353]. Indeed, the visual content within newspapers has proven to be a capacious source for humanists. Within periodicals studies, scholars have utilized the visual content in newspapers to investigate topics as far ranging as the evolution of comedic sensibilities within comic strips to hidden editorial practices embedded within newspaper layout [30, 65]. This collective body of work is bolstered by new methodologies being employed within the digital humanities to extract and analyze visual content in historic newspapers [112, 293, 394].

In this chapter, I scale up this analysis of visual content to explore the Ladino press at a macroscopic level. Using a machine learning model that I developed as part of my project, Newspaper Navigator, I have constructed a dataset of extracted photographs, illustrations, maps, comics, editorial cartoons, and advertisements from over 15,000 digitized pages of these Ladino newspapers [194, 194]. This approach of utilizing a machine learning model to extract visual content represents an emerging methodology for digital humanities research with periodicals and presents opportunities to facilitate access and research within Jewish Studies.

With the extracted visual content from the Ladino newspapers, it is possible to study the transnational dynamics shaping Sephardic print culture and the broader Sephardic experience at an unprecedented scale. Accordingly, I describe my results related to analyzing this visual content using emerging visualization techniques in order to provide insights related to recurring motifs and temporal trends. I offer this work as a case study in interdisciplinary research in the digital humanities and Jewish Studies. Throughout the chapter, I offer methodological reflections related to applying emerging computational techniques to Jewish Studies. I conclude the chapter with a reflection on the ethical considerations of applying machine learning and computer vision techniques to these Ladino newspapers and, more generally, to Jewish cultural heritage.

This work is based upon a single-author chapter that appeared in the book *Jewish Studies in the Digital Age*, published by De Gruyter Press [197].

2.5.1 The Digital Humanities and Visual Analysis of Newspapers

The visual culture preserved within historic newspapers has proven to be a fruitful and capacious source among scholars across diverse research areas. For example, scholars have studied the embedded editorial cartoons to understand the invocation of historical analogies [400]; comic strips to understand the evolution of humor about ethnicity [65]; illustrations to study the portrayal of identity [397]; maps to assess cartographic practices as well as the spatial thinking abilities of readers [251, 329]; and photographs to study the history of photojournalism [123]. Within Jewish Studies, Sarah Stein's book *Making Jews Modern: The Yiddish and Ladino Press in the Russian and Ottoman Empires* makes a compelling case for the significance of visual culture within Ladino and Yiddish newspapers [353]. Stein's detailed analysis of the advertisements within the Constantinople-based Ladino newspaper El Tiempo traces recurring motifs in order to argue how readers sought remedies to the anxieties of modernity under Ottoman rule: advertisements for medicines, insurance, clothes, travel, and lotteries all targeted readers concerned about stability and class. This chapter builds on this already significant body of work in order to consider advertisements and other visual content in the Ladino press at the macroscopic scale.

Scholarship making use of visual culture in historic newspapers has been redoubled by the growing interest in visual analysis within the digital humanities. Though research in the digital humanities has historically centered around text as the primary medium of interest, the field's "visual digital turn" over the past decade has begun foregrounding the analysis of visual media, including images and video [231, 232, 394]. This visual digital turn has coincided with methodological advances in machine learning approaches to image analysis due to deep learning. With the democratization of deep learning approaches to image recognition over the past few years via open source libraries and pre-trained models, digital humanities practitioners have begun utilizing these approaches for a wide range of research goals, from enriching the metadata of digitized collections to analyzing sitcoms [21, 120, 212, 223]. As these approaches continue to improve, it is clear that machine learning will occupy an increasingly important role within the digital humanities and the humanities writ large, as well as within the cultural heritage sector, including libraries and archives.

Within periodicals studies, researchers have begun utilizing machine learning approaches to study the visual components of newspaper pages, from analyzing visual layouts to classifying and searching the visual content embedded within the pages [112, 202, 293, 394]. Indeed, as a mode of humanistic inquiry, the application of machine learning to the visual analysis of newspapers has much to offer to both Jewish Studies and the digital humanities.

In the case of the Ladino press, the utilization of machine learning methodologies to study the visual content embedded within the newspaper pages is even more urgent due to the extant challenges surrounding the application of optical character recognition (OCR) algorithms to transcribing Ladino texts. Off-the-shelf OCR algorithms have yielded poor performance to date because these algorithms interpret Ladino texts printed in Rashi script as Hebrew; the poor OCR quality in turn restricts the ability of scholars to perform reliable keyword searches or apply digital humanities methodologies for textual analysis. Though new OCR engines are being developed specifically for Ladino texts, a fundamental challenge remains at this time: how do we study the Ladino press at a macroscopic scale beyond close reading [366]? Applying machine learning to study the visual content in these pages affords us a path forward.

The next subsections of this chapter concern the methodology employed to extract and analyze this visual content within 15,820 Ladino newspaper pages using machine learning. The chapter then turns to analyzing the extracted visual content.

2.5.2 Constructing the Dataset of Excavated Visual Content

In this subsection, I describe the process of utilizing this visual content recognition to extract visual content from Ladino newspapers in more detail. The corpus of Ladino newspapers consists of 15,820 Ladino newspaper pages from eight titles published between 1890 and 1948, amounting to 63.3 gigabytes (GB) of image data [194]. Table 2.8 presents a breakdown of the Ladino corpus included in this analysis according to title and date of publication. Manually assembling a dataset of extracted visual content across this full corpus would require hundreds of human annotation hours. However, the automated Newspaper Navigator visual content recognition model can process multiple pages per second on a single graphics

Table 2.8: Ladino newspaper titles with corresponding number of images and digitized pages processed using Newspaper Navigator. In the case of *La Vara*, each image contains two newspaper pages. In bold are the statistics for all digitized pages for a given title..

Newspaper Title	# of Digitized Images	# of Pages
El Instruktor revista siyentifika i literaria	**331**	**331**
El jugeton, Jurnal umoristiko	**4**	**4**
***El Kirbatch Americano* (1915–1917)**	**206**	**206**
El Luzero Sefaradi (October 1926)	28	28
El Luzero Sefaradi (May 1927)	28	28
***El Luzero Sefaradi* (Total)**	**56**	**56**
El Progresso / Yosef Daat	**332**	**332**
El Tiempo (1890–1891)	926	926
El Tiempo (1896–1897)	1,138	1,138
El Tiempo (1900)	584	584
El Tiempo (1900–1901)	981	981
***El Tiempo* (Total)**	**3,629**	**3,629**
***La Boz de Oriente* (April, 1931–April, 1932)**	**860**	**860**
La Vara (January 9, 1922 – June 4, 1923)	141	282
La Vara (May 1, 1927 – December 17, 1929)	675	1,350
La Vara (January 3, 1930 – December 20, 1932)	701	1,402
La Vara (December 7, 1932 – April 25, 1941)	667	1,334
La Vara (January 6, 1933 – December 27, 1935)	636	1,272
La Vara (January 3, 1936 – August 26, 1938)	704	1,408
La Vara (September 2, 1938 – April 25, 1941)	681	1,362
La Vara (May 2, 1941 – December 29, 1944)	648	1,296
La Vara (January 5, 1945 – February 13, 1948)	348	696
***La Vara* (Total)**	**5,201**	**10,402**
Total	**10,619**	**15,820**

processing unit (GPU), making it possible to process the full Ladino corpus in just a few hours.

To begin the processing of these pages, I first moved the high-resolution images of digitized Ladino newspaper pages to a private Amazon AWS S3 bucket, a form of cloud storage that facilitates fast computing against the corpus. I then wrote code to process these pages using the existing Newspaper Navigator visual content recognition model that had been trained on the Beyond Words annotations. Because machine learning models can be stored as single files known as "weights files," which can be loaded onto a computer and utilized for processing data with just a few lines of code, the majority of this code was devoted to handling the downloading of images from the cloud and processing the images in parallel. To deploy this code, I ran the processing pipeline on a rented Amazon AWS g4dn.12xlarge EC2 instance consisting of 48 CPUs and four NVIDIA T4 GPUs. In total, the pipeline extracted six classes of visual content across the corpus of Ladino newspapers: photographs, illustrations, maps, comics, editorial cartoons, and advertisements. I then saved the resulting extracted images, as well as metadata from the machine learning model, to the AWS S3 bucket, making it straightforward for us to download the full dataset and relevant subsets as necessary. I am currently in the process of investigating options for making this dataset of extracted visual content available to researchers and the public alike.

A breakdown of identified visual content is presented in Table 2.9. Because the machine learning model returns a confidence score with each predicted bounding box, and because one's choice of threshold cut on confidence score affects one's tradeoff between false positives and false negatives and thus changes the size of the resulting dataset, I include three cuts on confidence score in the table: 90%, 70%, and 50%.

Notably, this visual content recognition model was trained on annotated World War I-era newspaper pages in Chronicling America, rather than annotated Ladino pages. Consequently, the resulting dataset contains a nontrivial number of false positives and false negatives, as evidenced by the map class, which largely consists of false positives. It should be noted that the performance of the visual content recognition model is dependent on a confluence of factors, ranging from page layout to time period, typeface, language, and even subtleties of the digitization pipeline, such as the scanner used to image the pages. For an analysis of the effects of time period on the generalization performance of the visual content recognition model, I refer the reader to the Newspaper Navigator dataset paper [194]. It is undoubtedly the case that the utilization of a model trained on annotations for Chronicling America pages, rather than Ladino pages, impacts the resulting dataset. This effect can be quantified by evaluating the performance of the visual content recognition model on a hand-labeled test sample of Ladino pages (an evaluation left for future work). However, as evidenced by the analysis of the dataset presented in the next subsection, it is clear that the resulting dataset is of more than sufficient quality for supporting downstream exploration and research pertaining to questions of humanistic inquiry.

Table 2.9: A breakdown of extracted visual content in the Ladino newspaper titles processed. Three different cuts on the visual content recognition model's confidence score (90%, 70%, and 50%) are presented to show the effect of the cut choice when favoring false positives or false negatives.

Visual Content Type	$\geq$ **90%**	$\geq$ **70%**	$\geq$ **50%**
Photographs	348	770	1,060
Illustrations	52	378	960
Maps	27	182	300
Comics	10	39	184
Editorial Cartoons	8	31	111
Advertisements	18,381	31,523	42,505
Total	18,826	32,923	45,120

2.5.3 Analyzing the Excavated Visual Content

To begin the macroscopic analysis of the excavated visual content, I created cluster-based visualizations of advertisements and photographs, grouped according to their semantic content. In this step, I generated image embeddings for all of the extracted visual content; to accomplish this, I modified the Newspaper Navigator pipeline code, available in the Library of Congress GitHub repository for Newspaper Navigator, and utilized img2vec, a library for the streamlined generation of image embeddings from image files [195, 328]. The image embeddings utilized in this analysis are lower-dimensional representations of the images extracted from the hidden layers of ResNet-18 and ResNet-50, two neural image classification models [134]. Originally trained on ImageNet, these models can classify images according to their content (e.g., "dog" or "cat") [82]. Because these models capture the semantics of images, image embeddings generated by feeding images into these models capture semantic similarity: if the distance between two image embeddings (vectors that are each hundreds or thousands of dimensions in length) is small, the corresponding images likely have similar semantics. Thus, generating visualizations of photos and advertisements clustered based on the image embeddings can provide an informative summary of the landscape of visual content. To make the high-dimensional clustering visible in two-dimensional visualizations, I have utilized T-SNE, a dimensionality reduction algorithm that preserves close clusters of points, meaning that clustered points in the visualization are also clustered in the high-dimensional embedding space. It should be noted that long distances are not preserved in T-SNE, so the relative positions of clusters should not be taken into consideration [376, 390].

To start, I generated visualizations of the 348 photographs identified by the Newspaper Navigator visual content recognition model with confidence scores greater than 90%. Figure 2.21 presents the cluster visualization of these 348 photographs. In this visualization, I present a summary view that would have ordinarily required manually inspecting and analyzing 15,820 pages. Examining the visualization, it is immediately apparent that many photographs depict people. Portrait shots, such as the ones clustered together and depicted in Figure 2.22, are one of the most common types of image. Other notable clusters include wartime photographs (shown in Figure 2.23) and crowds and groups of people (shown in Figure 2.24).

As shown in Table 2.9, the critical mass of identified visual content consists of advertisements. The advertisements embedded within the Ladino press attest to daily life within Sephardic culture from Constantinople to New York. By studying these extracted advertisements at a macroscopic scale, it is possible to augment the extant historiography, including Sarah Stein's detailed analysis of advertisements in the Ladino and Yiddish Press [353].

Given the number of advertisements identified, I chose to generate cluster visualizations of advertisements for *La Vara* for different temporal regions. Figure 2.25 shows a cluster visualization of 2,812 advertisements extracted from *La Vara* issues published between January 3, 1936, and August 26, 1938. The visualization reveals numerous distinct clusters of the same advertisements reproduced multiple times throughout multiple issues of *La Vara* within the given temporal range. Many of these clusters contain dozens of the same advertisement, reflecting businesses that chose to advertise consistently within the pages of *La Vara*. Figure 2.26 shows six of these clusters, along with magnified versions of the reproduced advertisements. Examples of advertisers with reproduced advertisements in the dataset include Brockman Monument Works, Meyer London's Matzos, Standard Truss Co., Golden Wine & Liquor Co., Paradise Interior Decorators, Aristocratic Imported Virgin Olive Oil, Harem Oriental Pastry, Constantinople Oriental Pastry Shop, Macedonia Importing Co., Royal Hall, Mid-Bronx Used Car Exchange, the Luxor Food Market, the Luxor Restaurant, the Sephardic Jewish Center, Inc., Joseph Levy (furniture, radios, and

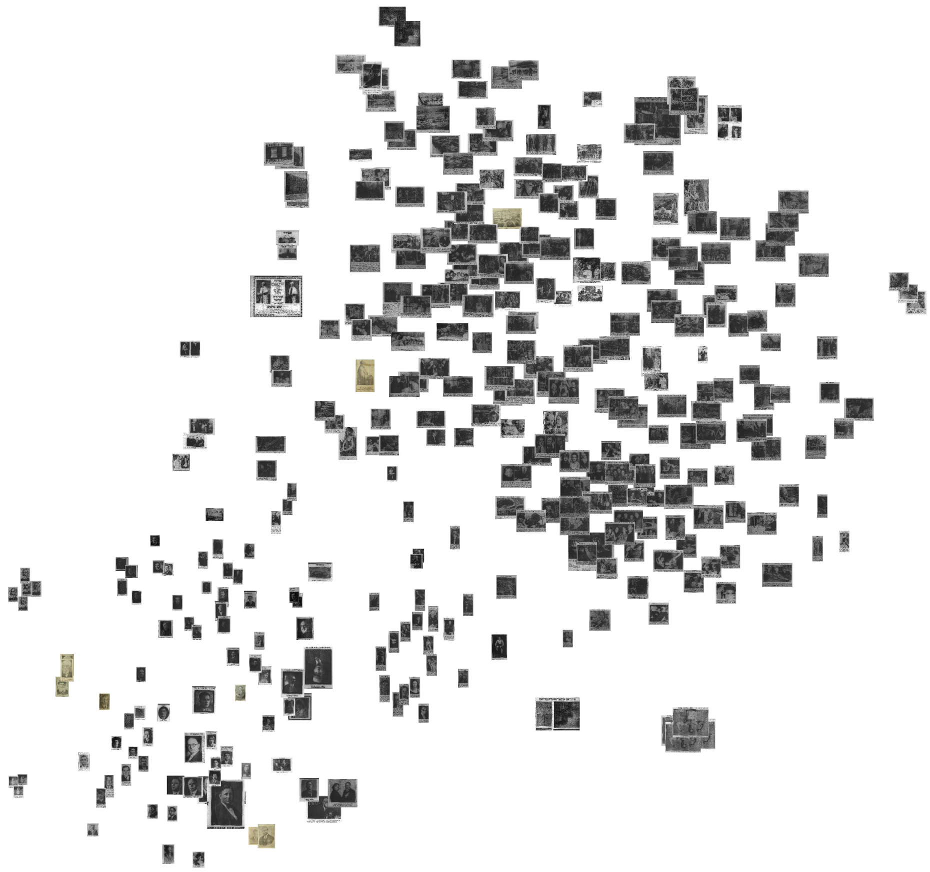

Figure 2.21: A cluster visualization of the 348 photographs identified by the visual content recognition model with confidence scores greater than 90%. I constructed this visualization using ResNet-50 embeddings and T-SNE for dimensionality reduction.

oilcloths), Louis J. Opal (counselor at law), Simon S. Nessim (counselor at law), Dr. J. Feitelson (dental surgeon), Irving Matalon, P. Vladeff, and Madame Gilda Malky. The overwhelming majority of these advertisements are for local New York City businesses and also feature prominent English text, both of which reflect *La Vara*'s role as an American Sephardic press within New York City.

Analysis of the surfaced advertisements that have been reproduced many times over reveals similar advertising patterns to those uncovered within *El Tiempo* by Sarah Stein: a preponderance of advertisements of sartorial nature, as well as for doctors, dentists, medical treatments for ailments, and legal counsel [353]. In the case of Constantinople-based *El*

Figure 2.22: A magnified cluster within Figure 2a consisting of portrait shots of people.

Tiempo, Stein argues that these advertisements speak to readers' anxieties under the precarity of Ottoman rule, including class and economic anxieties and aspirations. The apparent resonances between the advertisements in *El Tiempo* and those in the New York-based *La Vara* suggest an even broader pattern of Sephardic Jewish experiences in response to social and economic uncertainty and change during the late 19th and early 20th centuries, whether in the United States or the Ottoman Empire.

Of particular interest are the recovered advertisements for Meyer London's Matzos (entry b in Figure 2.26), which appeared concurrently in American Yiddish newspapers. As described by Makena Mezistrano, "Matsa advertisements in the American Yiddish and Ladino presses offer a rare opportunity to place these two communities in dialogue with one another, instead of only positioning them as separate or in bitter conflict – two com-

Figure 2.23: A magnified cluster within Figure 2a consisting of wartime photographs.

mon assumptions about intra-Jewish relationships in twentieth century New York" [238]. Thus, the advertisements uncovered through this cluster-based analysis speak to not only individual Sephardic communities but also relationships between and across communities, as embedded within cultural practices.

However, not all clusters correspond to advertisements. In Figure 2.27, magnified clusters of extracted photographs and full newspaper pages from *La Vara* are shown. These clusters are false positives, reflecting the imperfect performance of the visual content recognition model utilized for visual content extraction. These clusters are an important reminder that algorithmic approaches to extracting visual content are inevitably imperfect. However, using clustering and other machine learning techniques, it is possible to remove many of these false positives quickly.

2.5.4 Future Work

Ongoing work consists of continuing to explore the extracted visual content via both macroscopic analysis and close analysis within the page-level context. In terms of macroscopic

Figure 2.24: A magnified cluster within Figure 2a consisting of photographs of crowds and groups of people.

analysis, I plan to expand the study of the reprinting patterns of advertisements in order to examine the network of advertisers that funded the Ladino press. By building on the provocations offered in this chapter, one can ask questions such as: did advertisers purchase advertising space in different titles? And what does this tell us about the interconnectedness of the Ladino press? Moreover, I plan to expand this analysis to different temporal slices of *La Vara* along with quantitative assessments of different photograph types in order to understand the evolution of the visual content. I will also expand this analysis to include a greater exploration of the other Ladino titles present in the dataset (as enumerated in Table 2.9). With this future analysis, I can begin to ask questions surrounding the intended audiences of the advertisements and how they changed over time as well as by title, building

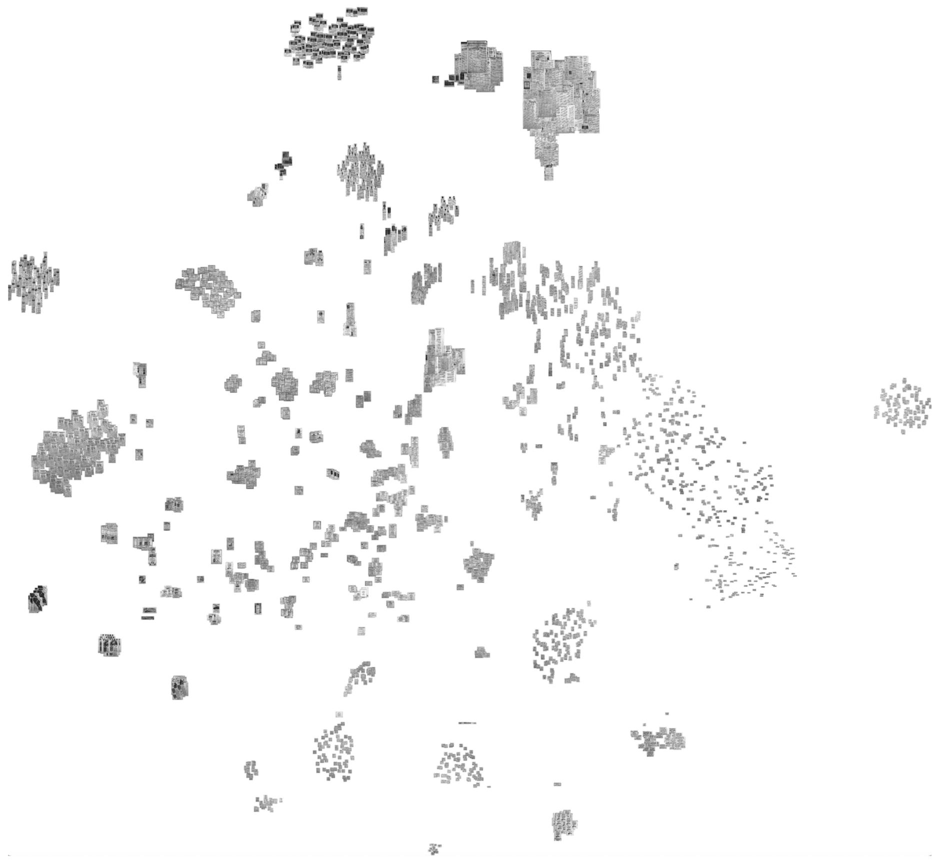

Figure 2.25: A cluster visualization of 2,812 advertisements identified by the visual content recognition model with confidence scores greater than 90% within issues of *La Vara* published between January 3, 1936, and August 26, 1938. I constructed this visualization using ResNet-50 embeddings and T-SNE for dimensionality reduction.

on Sarah Stein's analysis of the advertisements in *El Tiempo* [353]. Because the analysis of the visual content in this chapter has focused on the extracted dataset, future work also entails understanding the visual content by recontextualizing it at the page level within the broader mise en page. What types of articles accompany visually similar photographs? What advertisements appear next to one another? What do the captions reveal about the visual content?

From a computational perspective, future work with the dataset of extracted visual content entails evaluating the generalization of the visual content recognition model on the Ladino newspaper pages. This evaluation will require manually annotating enough pages

Figure 2.26: Six different magnified clusters within Figure 2.25 showing advertisements reprinted throughout different issues of *La Vara* (top), along with magnified versions of the reproduced advertisements (bottom). The advertisements are for Brockman Monument Works (a), Meyer London's Matzos (b), Standard Truss Co. (c), and Aristocratic Imported Virgin Olive Oil (d, e, f).

across different titles and temporal slices in order to derive reliable statistics. With this in-depth evaluation across many different newspapers at varying time periods, it is possible to better understand the bias of the visual content recognition model, which will, in turn, inform the results of this macroscopic analysis even further. Lastly, given that so many of the advertisements had captions written in English, future work entails running English

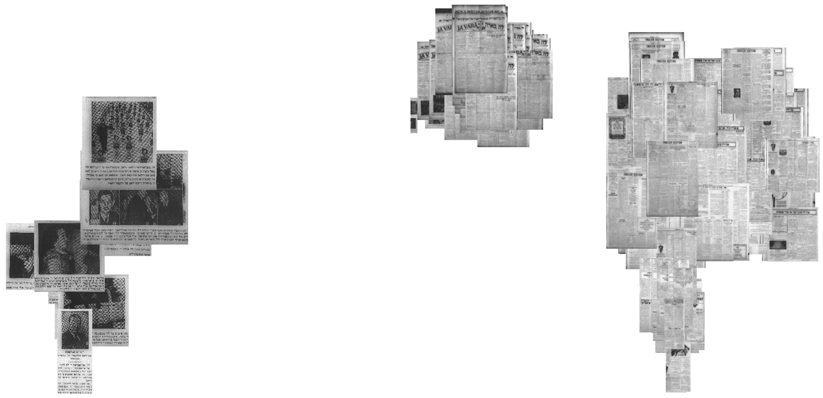

Figure 2.27: Magnified clusters of photographs and full pages, showing false positives among the identified advertisements from *La Vara*.

OCR engines on the advertisements utilizing the results for textual analysis of the captions.

Other potential work includes cross-matching the visual content in the Ladino press with the extracted visual content from other newspaper corpora, such as the visual content from 16 million pages in *Chronicling America* contained within the Newspaper Navigator dataset. Identifying reproduction patterns among the visual content within these different presses could indeed speak to the proximity or marginal position of the Ladino press in relation to broader American newspaper syndicates. Moreover, Makena Mezistrano's discovery of Meyer London's Matzos advertisements in both the American Sephardic press and American Ashkenazic press suggests that this future direction of cross-matching visual content across different presses has the capacity to enrich our understanding of how cultural practices change across different communities.

Lastly, in regard to the dataset of extracted visual content from Ladino titles, I have two primary goals for future work. First, as articulated earlier in this chapter, I plan to make this dataset of extracted visual content publicly available to encourage re-use among scholars and the public. Second, I hope to expand the Ladino titles that have been processed in order to further excavate the visual content across the Ladino press.

2.5.5 Ethical Considerations

Given the profound implications of machine learning perpetuating marginalization and erasure through algorithmic bias and other mechanisms, any application of machine learning to cultural heritage collections would be remiss without a discussion of the ethical considerations surrounding doing so. Within the library, archive, and museum ("LAM") community, there has been a growing effort to consider a critical, sociotechnical lens surrounding machine learning, data science, and cultural heritage. This effort has culminated in the development of responsible operations and best practices, as well as surveys of projects in this liminal space [70, 156, 283]. In the case of difficult and understudied histories, extra precaution must be taken, and a rich discourse in the scholarly community has explored the ethics of datafication and the application of machine learning methodologies within this context [194, 295]. In this subsection, I draw from these emerging bodies of work and explicitly build on the Newspaper Navigator data archaeology, which I wrote in order to detail the implications of machine learning for search and discovery from a sociotechnical perspective.

Though often overlooked, the marginal position of Sephardic Studies within Jewish Studies has been amplified by machine learning, having altered the discoverability of the Sephardic historical record via digitization. As detailed earlier in this chapter, off-the-shelf OCR algorithms consistently fail to transcribe Ladino texts with a high enough degree of fidelity to facilitate reliable keyword search and textual analysis. This effective erasure of Ladino texts from search and discovery platforms is the result of a confluence of factors, from the availability of training data to the monetary value of preferentially selecting widely studied languages for inclusion in proprietary OCR engines. Significantly, this linguistic erasure is not limited to Ladino: a similar systemic problem has been documented for Yiddish and indigenous languages, which speaks to a specific form of algorithmic bias, in which human decisions surrounding which languages should be prioritized when training OCR algorithms have a profound impact on resulting scholarship [5, 54].

In this chapter, I seek not only to foreground the algorithmic marginalization of Sephardic history but also to offer an alternative approach to recover the voices that have been lost through digitization. Certainly, the utilization of machine learning to excavate visual content in Ladino newspapers is not without its own challenges. The Newspaper Navigator visual content recognition model performs better on pages that more closely resemble the training data, and the extracted dataset presented in this chapter suffers from a nontrivial number of false positives and false negatives as a result [194]. These false positives and false negatives motivate methodological improvements, such as training a visual content recognition model specifically for the Sephardic press in order to better capture the nuances of Sephardic visual culture. Moreover, image recognition algorithms used to evaluate image similarity have been shown to perpetuate their own forms of bias and marginalization [196]. Because these algorithms have been trained by machine learning practitioners with specific objectives and categories in mind, a fundamental question is raised as to whether the groupings identified by the algorithms capture the relationships most valuable to scholars. While these methods have the capacity to expose new groupings, they inevitably distort the viewer's perceptions of what constitutes similarity. I therefore offer these methodological approaches with such considerations in mind, a reminder of the importance of canonical historiographic approaches that can be used in concert with machine learning. And yet, this chapter has provided the first macroscopic view of the Ladino press via the excavated visual content and thus serves as a corrective to the algorithmic marginalization of Sephardic Studies.

2.6 *The "Collections as ML Data" Checklist for Machine Learning and Cultural Heritage*

Within cultural heritage, there has been a growing and concerted effort to consider a critical sociotechnical lens when applying machine learning techniques to digital collections. Though the cultural heritage community has collectively developed an emerging body of work detailing responsible operations for machine learning in galleries, museums, archives, and libraries at the organizational level, there remains a paucity of guidelines created for researchers embarking on machine learning projects with digital collections. The manifold stakes and sensitivities involved in applying machine learning to cultural heritage underscore the importance of developing such guidelines. This section contributes to this need by formulating a detailed checklist with guiding questions and practices that can be employed while developing a machine learning project that utilizes cultural heritage data. I call the resulting checklist the "Collections as ML Data" checklist, which, when completed, can be published with the deliverables of the project. By surveying existing projects, including my own project, Newspaper Navigator, I justify the "Collections as ML Data" checklist and demonstrate how the formulated guiding questions can be employed by researchers.

This work is based upon a single-author publication that is forthcoming in the *Journal of the Association for Information Science and Technology* (JASIST) Special Issue: "Conceptual Models of the Sociotechnical" [193].

2.6.1 Introduction

Over the past few years, the field of machine learning has seen a growing movement to develop algorithmic impact assessments, checklists, and best practices that machine learning researchers and practitioners (those who utilize machine learning in research, experiments, and operational contexts) can consult while creating datasets, training models, and operationalizing these systems [37, 240]. These efforts show great promise, having seen widespread adoption across the field – from conference submission requirements [67, 292] to researchers' own utilization of these resources to guide their work and communicate their decisions throughout a project's development.

Concurrently, the cultural heritage sector is critically investigating the sociotechnical implications of applying machine learning to digital collections. Indeed, machine learning has a long history among the gallery, library, archive, and museum (GLAM) communities, as well as the digital humanities [378]: for example, optical character recognition (OCR) algorithms have been used in digitization pipelines for decades [70]. Informed by critical cataloging [1, 45, 111, 175, 278], critical data studies [75], science & technology studies [250], the "Collections as Data" movement [245, 283, 304], digital strategies among GLAMs [192], computational archival science [64, 142, 155, 325], and decades of scholarship, this effort has produced "state of the field" reports and best practices for cultural heritage. For example, reports such as Thomas Padilla's "Responsible Operations: Data Science, Machine Learning, and AI in Libraries" and Ryan Cordell's "ML + Libraries: A Report on the State of the Field" have been foundational in articulating principles and provocations to guide the operationalization of machine learning at GLAMs [70, 283]. Yet, such efforts have largely focused their attention on guidance at the organizational level, rather than for individual researchers and scholars, who face additional challenges.

There is much to be gained by developing a checklist for researchers carrying out machine learning projects in the context of treating cultural heritage collections as data. Even with good intentions, such projects risk kitschifying or exploiting those represented in the digitized collections in question; glossing over digitization subtleties that impact the performance and output of machine learning models; utilizing machine learning when it is not necessary due to organizational agendas surrounding emerging technologies; beginning machine learning projects with no plans for sustainability; or violating the privacy of end-users of systems that are built. Proper usage of a checklist facilitates thoughtful engagement with such challenges and therefore has the potential for profound impacts for researchers studying digital cultural heritage collections.

This section draws from these emerging movements in the machine learning and cultural heritage communities in order to produce the "Collections as ML Data" checklist that researchers can utilize when embarking on projects in machine learning and cultural heritage. The completed checklist can be published with a project's deliverables in order to provide the audience with a structured description of the project's considerations and limitations. To justify each component of the checklist and demonstrate how it can be applied, this work examines a series of case studies. Though the "Collections as ML Data" checklist has been designed for researchers applying machine learning to cultural heritage collections, the checklist offers guidelines and considerations pertinent to other communities and disciplines as well, including computational social science and non-cultural data work. Moreover, the checklist's cultural heritage-specific lens offers a new perspective to computer science researchers surrounding data practices and beyond.

2.6.2 Related Work

This work contributes the first detailed checklist for projects at the intersection of machine learning and cultural heritage. The "Collections as ML Data" checklist builds on existing work surrounding best practices within the respective machine learning and cultural heritage communities, as well as the growing scholarship at the intersection of these fields. It also draws from the literature on conceptual models and scholarship on the digital humanities, critical data studies, and science & technology studies within the broader umbrella of information studies. This subsection serves to highlight this related work in more detail, contextualize the "Collections as ML Data" checklist, and highlight its contributions.

Checklists, Toolkits, and Best Practices within the Machine Learning Community

This work draws from existing checklists, toolkits, and best practices surrounding machine learning. Relevant work includes [19, 22, 37, 85, 86, 87, 110, 115, 127, 146, 167, 190, 228, 240, 307, 317, 321, 336, 361].[29] In the subsection "A Taxonomy of ML & AI Toolkits, Checklists, and Impact Assessments," a detailed taxonomy of these guidelines is provided, and I refer the reader to this subsection for a more thorough analysis of these works' contributions.

Collectively, these checklists, toolkits, and impact assessments emphasize a range of considerations for different stages of machine learning projects, including the construction of a dataset, the training and auditing of a model, and the deployment of a model in an operational sense. In addition, the checklists target a range of different audiences, from researchers and practitioners themselves to downstream users of the constructed datasets or trained models. As highlighted in [37], such considerations are significant because "there are both scientific and ethical reasons to be concerned. Scientifically, there is the issue of generalizability of results; ethically, the potential for significant real-world harms." Indeed, from a scientific standpoint, incomplete documentation of a dataset or model can lead to unintentional misuse by other researchers; insufficient evaluation of a dataset or model can lead to unforeseen issues of generalization when operationalized; and lack of clarity surrounding copyright can hinder the adoption of a dataset or model. From an ethical standpoint, machine learning datasets risk exploiting personal data and raising questions of privacy; labor practices behind data annotations are not always foregrounded, leading to questions of labor exploitation; machine learning models can perpetuate marginalization and oppression; and operationalized systems can be fragile, failing without warning to end-users. Though no set of guidelines can comprehensively cover all such scientific and ethical questions, a checklist nonetheless represents a first significant step toward systematizing shared practices surrounding ethical and responsible decision making. In this regard, the checklist represents a conceptual model for analysis with emerging machine learning methodologies.

While all of the aforementioned checklists apply to machine learning research, it is im-

[29] It is worth noting that domain-specific machine learning checklists and best practices have been published as well, for example with chemistry [23] and medicine [335].

portant to note that none of them have been designed with the particular setting of cultural heritage in mind, which requires special consideration. For example, researchers must pay particular attention to their intended audience, which may range greatly from scholars to students in the classroom; the provenance of the digital cultural heritage collections with which they are working, including the nuances of digitization; and potential discrepancies surrounding technical fluency in interdisciplinary collaborations. It is clear that much can be gained by developing a checklist to guide researchers through such subtleties. Indeed, this is precisely the central contribution of this work with the "Collections as ML Data" checklist.

In addition, this checklist offers guidelines that are relevant to computer science researchers more generally, even if they already have experience with existing machine learning checklists. For example, by considering cultural heritage-specific sensitivities surrounding curation rationale and collection provenance articulated in this checklist, machine learning researchers may seek out a more nuanced understanding of the origins of their datasets. Moreover, by engaging with the question of why machine learning must be utilized in order for the project to succeed, computer science researchers may further interrogate the sociotechnical implications of their work.

Best Practices from Cultural Heritage

This work draws from an emerging body of literature devoted to best practices surrounding machine learning and data science within cultural heritage. For example, the Library of Congress (LC) Labs team's report for the ML + Libraries summit summarizes the findings of a one-day conference hosted with three goals: "survey[ing] the range of ongoing projects in the broader cultural heritage landscape; surfac[ing] major possibilities and barriers for applying machine learning in a library setting; and demonstrat[ing] the possibilities of machine learning for use at the Library of Congress to internal audiences" [156]. While this report highlights many important directions for machine learning and cultural heritage data, one shared challenge that emerged from the summit was the "lack of a clear roadmap for the use of machine learning in cultural heritage," indicating the importance of checklists within the field. Ryan Cordell's report "Machine Learning + Libraries: A Report on the State of the Field," commissioned by LC Labs at the Library of Congress, offers a detailed history of machine learning within libraries, shared challenges, and recommendations [70]. Thomas Padilla's OCLC report "Responsible Operations: Data Science, Machine Learning, and AI in Libraries" details "responsible operations refers to individual, organizational, and community capacities to support responsible use of data science, machine learning, and AI" [284]. Other reports, such as the Europeana "AI in Relation to GLAMs" Task Force's 2021 report [108], Colavizza et al.'s "Archives and AI: An Overview of Current Debates and Future Perspectives" [64], and Fiorucci et al.'s "Machine Learning for Cultural Heritage: A Survey" [101] offer taxonomies of existing projects in machine learning and cultural heritage, along with analysis of field-level trends.

Collectively, these reports draw from a long history of critical work within library, archival, and information sciences, as well as the humanities, in order to articulate the specific subtleties that arise when applying emerging technologies to cultural heritage in particular: how to uphold the library community's standards for privacy, how to consider the provenance of collections and their digitization, and how to foster technical fluency among cultural heritage researchers are just a few such considerations [284]. This body of work in cultural heritage thus proposes conceptual models for ethical practices when applying new technologies to digital collections.

While these reports foreground best practices and recommendations, they are intended primarily as guiding principles at the organizational and field level. For example, Padilla's paper addresses audiences including library administrators, faculty, and staff; university administrators and disciplinary faculty; professionals; and funders [283]. Likewise, the Library of Congress's "ML + Libraries Summit" included participants with a range of backgrounds beyond digital humanities research, including librarians and archivists with expertise on topics such as metadata generation and preservation [156]. Only Ryan Cordell's report offers a checklist for practitioners, which takes the form of "25 Questions for Structuring an ML Project" [70]. Here, Cordell provides a set of guiding questions for library teams working on projects involving machine learning, a crucial development in developing project-oriented guidelines [70]. While Cordell's report emphasizes many salient guidelines, it is important to recognize that the target audience of library teams is distinct from the audience considered in this work. Just as Cordell's work is an adaptation of the aforementioned machine learning reports for the library community, especially within an institutional context, the "Collections as ML Data" checklist in this work can be seen as a further adaptation of Cordell's template for individual researchers at the intersection of machine learning and cultural heritage. As one example, a researcher may not have the same context for a digital collection of interest as the library team responsible for it does. Consequently, checklist questions surrounding a cultural heritage dataset, its provenance, and the digitization pipeline used to produce it are instructive for researchers as they familiarize themselves with the dataset. Likewise, even though considerations surrounding copyright, maintenance, and documentation may be standard at the institutional level, individual researchers may not have this institutional context surrounding the significance of such considerations. The "Collections as ML Data" checklist incorporates such questions geared toward researchers at this intersection.

This work builds on Cordell's "25 Questions for Structuring an ML Project" in three more ways. First, the "Collections as ML Data" checklist incorporates a computer science focus by drawing heavily from literature in machine learning and related fields. For example, the "Collections as ML Data" checklist includes a detailed examination of the machine learning model being utilized. By bringing together perspectives from both machine learning and cultural heritage, this work offers new ideas and perspectives to both communities. Second, this work demonstrates how these questions can be utilized in practice by detailing case studies. Third, the "Collections as ML Data" checklist expands on Cordell's 25 questions by providing even more granularity in walking through each step of a relevant project, for example, documenting a machine learning model that has been utilized and deployed. Indeed, the "Collections as ML Data" checklist proposed in this work builds on this work to produce a detailed checklist that, when completed by researchers, can be distributed with project deliverables in order to convey the considerations and limitations surrounding the project to its audience and auditors.

Projects Involving Machine Learning & Cultural Heritage

This work builds on the rich collective body of projects and research at the intersection of machine learning and cultural heritage. Surveys and state of the field reports such as [64, 64, 70, 101, 108, 126, 156], as well as resources such as the *Reviews in DH* project registry [126] and the *Debates in the Digital Humanities* series [119], describe the history of such work and enumerate current projects. In order to develop, refine, and evaluate the proposed checklist with case studies, this work draws from such surveys and state of the field reports to identify relevant projects at the intersection of machine learning and cultural heritage. The chosen projects are described in more detail in the subsection "Selecting Representative Projects for Case Studies."[30]

Documentation Practices in Information Science and the Digital Humanities

More generally, it is important to situate the "Collections as ML Data" checklist within the broader landscape of research in information science and the digital humanities [409]. Both disciplines have embraced novel information artifacts as worthy of publication alongside traditional journal papers. Such artifacts include datasets, online visualizations and exhibitions, and computational replay and provenance systems [237]. These efforts represent innovative deliverables for digital humanities projects that foreground transparency, making it possible for research communities to more easily reconstruct research findings as well as pursue new research on top of existing scaffolding. The "Collections as ML Data" checklist aligns squarely with this pursuit of transparency within information science and the digital humanities by serving as a deliverable that, when completed and published, addresses precisely these goals. In this regard, the "Collections as ML Data" checklist is a conceptual model of the sociotechnical considerations undertaken by a researcher applying machine learning to cultural heritage.

[30] I note that I utilize my project, Newspaper Navigator, as a case study because experiences developing the dataset [202] and launching the search application [205] served as initial provocations for developing this work. For more information on the autoethnographic approach that I adopted for Newspaper Navigator, I refer the reader to [196].

2.6.3 Overview of Methodology

Surveying Relevant Literature from Machine Learning

In order to formulate initial questions and suggestions for the "Collections as ML Data" checklist from a machine learning perspective, I started by surveying the relevant machine learning literature articulated in the subsection "Checklists, Toolkits, and Best Practices within the Machine Learning Community." To make sense of this capacious space of checklists, toolkits, impact assessments, and beyond, I produced a taxonomy of such work. One goal of creating this taxonomy was to develop a more comprehensive list of guiding questions to be included in the "Collections as ML Data" checklist from the perspective of machine learning. A second goal of producing this taxonomy was to help situate researchers with a guiding roadmap for existing work from the machine learning literature. Accordingly, I present this taxonomy in the subsection "A Taxonomy of ML & AI Toolkits, Checklists, and Impact Assessments."

Identifying Representative Projects as Case Studies

Next, I utilized the reports and surveys on machine learning projects in cultural heritage in order to select a representative grouping of projects as case studies for developing and testing the "Collections as ML Data" checklist. In particular, I refined desiderata for selecting projects that emphasized project diversity: institutional setting, digitized collection medium, digitized collection topic, machine learning methodology employed, field of study, intended audience, and final deliverable form. This process yielded five projects to serve as case studies. Though many papers in the machine learning checklist literature utilize case studies to evaluate the proposed checklists, the methodology behind selecting the case study projects is not always detailed. Accordingly, I elaborate on this process in the subsection "Selecting Representative Projects for Case Studies" with the goal of transparent documentation.

Surveying Relevant Literature from Cultural Heritage

The next step entailed surveying best practices and responsible operations from cultural heritage in order to develop checklist questions specific to cultural heritage projects. Here, I primarily drew from Ryan Cordell's report [70], Thomas Padilla's report [284], LC Labs's ML + Libraries summit report [156], and the "EuropeanaTech AI in relation to GLAMs" Task Force's report [108]. As described in the subsection "Best Practices from Cultural Heritage," these papers contain various guidelines and principles for machine learning and cultural heritage, touching on the special sensitivities to be considered by cultural heritage researchers, as well as the challenges faced by collaborations among stakeholders with different levels of fluency and training with machine learning and cultural heritage. I translated these guidelines and provocations into questions within the "Collections as ML Data" checklist.

Creating and Refining the Initial "Collections as ML Data" Checklist

I next turned to refining the checklist utilizing existing projects in this space. Drawing from my own experiences with my project, Newspaper Navigator [196, 202, 205], which served as the initial motivation for developing this checklist, I formulated a series of additional questions relevant to working with cultural heritage collections, resulting in an initial version of the "Collections as ML Data" checklist. To test this initial version, I utilized the projects selected as case studies. Inspired by the methodology utilized in [277], these case studies resulted in vignettes for justifying and detailing each checklist item. In analyzing each project, I identified redundancies within the checklist questions as well as subtleties raised by the project that were not yet covered by the checklist. Accordingly, I iteratively revised and refined the checklist with each case study.

Incorporating Feedback from Researchers in Machine Learning and Cultural Heritage

The last step in refining the checklist entailed incorporating feedback from researchers in both machine learning and cultural heritage collections. In particular, I further tested the efficacy and extensibility of the checklist by workshopping it with colleagues at the University of Washington and other institutions. I began by soliciting feedback from Professor Katharina Reinecke and the other students enrolled in the CSE 599 "Computer Ethics" graduate course at the University of Washington. This included two rounds of soliciting written feedback surrounding the checklist, as well as an additional round of workshopping the checklist during an oral presentation. During this phase, I focused on feedback from the perspective of machine learning and refined the checklist accordingly. To receive feedback from the perspectives of cultural heritage and digital humanities research, I subsequently solicited written comments from senior researchers with primary affiliations outside of academic computer science, including information schools. These researchers had extensive experience with cultural heritage data. During these rounds of feedback, I provided earlier drafts of the paper, and the written feedback that I received primarily concerned the comprehensiveness and organization of the checklist, as well as my justifications for its components. In conjunction with the previous steps – including my own repeated refinement of the checklist using the case studies – these iterations culminated in the final checklist presented in the subsection "The "Collections as ML Data" Checklist: An Overview."

2.6.4 A Taxonomy of ML & AI Toolkits, Checklists, and Impact Assessments

As described in the subsection "Overview of Methodology," in order to develop initial questions for the "Collections as ML Data" checklist based on existing work, I performed a literature review for checklists, impact assessments, toolkits, and best practices in machine learning. I began the review with papers that I had already encountered and followed the citation graphs. In addition, I searched paper repositories and consulted colleagues within machine learning. Lastly, I sought recommendations from colleagues in machine learning. With the literature identified, I then categorized papers into four discrete groups. The collected works in each of these four categories offer a different perspective for machine learning projects, and one explicit aim of the "Collections as ML Data" checklist is to draw from all of these perspectives. It should be noted that the set of papers that I have included is not comprehensive, as work in this space is evolving quite quickly, and I limited my survey primarily to academic publications. As such, this taxonomy should be treated not as a full survey but rather a guiding resource for researchers looking to situate themselves within the landscape of machine learning assessments.

Dataset Assessments

Though datasets have always been essential to the field of machine learning, it has been only in recent years that the discipline has begun to adopt a critical lens toward examining their construction, composition, and utilization. Often motivated in the literature through invocations of examples of machine learning models being deployed in high stakes decision-making processes (i.e., medical diagnosis, legal recidivism, and credit score determination), an emerging body of work is calling for researchers to publish dataset assessments along with the datasets themselves. The dataset assessment can serve as a compliance checklist (such as ensuring IRB & GDPR compliance [121]); a practical guide for consumers of the dataset (such as other machine learning researchers using the dataset to train a model); a pedagogical tool for the public; and a means for recourse for those whose data are contained within the dataset. Data assessments are often inspired by and modeled after regulatory efforts such as the nutrition label for food and beverage packaging. Five related approaches to data assessments that have received widespread attention in the machine learning community to date are:

1. "Datasheets for Datasets" [115]
2. "Data Statements for NLP" [37]
3. "The Dataset Nutrition Label" [146]
4. "Data Cards" [307]
5. "Comprehensive and Comprehensible Data Catalogs" [361]

Model Assessments

A concurrent thread emerging from the machine learning literature is the development of model assessment rubrics and checklists for practitioners to complete. Motivated in a similar fashion to dataset assessments, model assessments concern the analysis of the training, evaluation, deployment, and operationalization of the model itself. Existing work in this space has advocated for these model assessments to be published along with the models themselves, much akin to dataset assessments. Examples of model assessments include:

1. "Model Cards for Model Reporting" [240][31]
2. "CheckList for NLP Models" [321]
3. "FactSheets" [19]
4. "Consumer Labels" for machine learning models [336]
5. "System Cards" for AI decision-making for public policy [127]
6. Microsoft Research's AI fairness checklist [41]
7. The "AI Blindspot" discovery process [49]

Though all of these papers make mention of the importance of scrutinizing training data, the emphasis is on model training, deployment, and operationalization.

Algorithmic Impact Assessments

Inspired by environmental impact statements that construction programs must produce, algorithmic impact statements offer an accountability framework for those who operationalize algorithms [317, 340]. Just as "the environmental impact statement process combines a focus on core values with a means for the public, outside experts, and policymakers to consider complex social and technical questions" [317], the algorithmic impact statement advocates for an iterative process of development between agencies, the public, and knowledgeable outside parties. Examples of algorithmic impact assessments include:

1. The AI Now Institute's "Algorithmic Impact Assessments" [317]
2. The ACM Conference on Fairness, Accountability, and Transparency in Machine Learning's "Principles for Accountable Algorithms" and "Social Impact Statement for Algorithms" [85]
3. Nick Diakopoulos's "Algorithmic Accountability Reporting" [86, 87]

It should be noted that the algorithmic impact assessment has a strong emphasis on policy applications and thus tends to be written with policy makers in mind.

[31] Closely related work, such as "Interactive Model Cards" [76], is worth noting as well.

Toolkits & Ethics-based Approaches

A fourth category of assessment is the algorithmic "toolkit," which does not fall cleanly into the three aforementioned categories. Two examples from this category are summarized below:

1. Microsoft Research's FairLearn toolkit [41], which could be considered part of the model assessment class, but which I have differentiated because it is a codebase that can be used to audit systems from the perspective of fairness. Using the toolkit does not result in a deliverable to be shared with the public but rather helps the practitioner modify the system itself.
2. The Washington State ACLU's Algorithmic Equality Toolkit [167], which is differentiated from the model assessment category because it is intended primarily for activists and community advocates in order to "promote public understanding of algorithms and artificial intelligence" and increase accountability and regulation.

Related to the toolkits in this category are ethics-based approaches to tech project design, which offer slightly different but valuable perspectives to the categories previously enumerated. Examples include "Envisioning Cards" [110], "Tarot Cards for Tech" [22], and "Surveying the Landscape of Ethics-Focused Design Methods" [61], a survey of 63 such methods.

2.6.5 Selecting Representative Projects for Case Studies

In order to develop a checklist for machine learning and cultural heritage, it was next necessary to identify a representative selection of existing projects in this emerging body of work in order to serve as case studies. To identify relevant projects, I consulted four reports that survey examples in this space: the Library of Congress Labs team's summary report for the ML + Libraries summit [156], Ryan Cordell's report "Machine Learning + Libraries: A Report on the State of the Field" [70], Thomas Padilla's OCLC report "Responsible Operations: Data Science, Machine Learning, and AI in Libraries" [284], and the "EuropeanaTech AI in relation to GLAMs" Task Force's report [108]. In addition, I consulted the digital humanities project registry published by *Reviews in the Digital Humanities* [126], as well as the *Debates in the Digital Humanities* series [119]. I also informally surveyed colleagues in machine learning, the digital humanities, and cultural heritage to generate candidate projects.[32]

It is important to clarify that the very notion of a "representative" subsample of projects is a fraught one, as it is highly subjective and entirely dependent on one's criteria for what constitutes representativity. For example, these surveys reflect an American perspective and are focused on the library setting. To address this, I refined dimensions and criteria along which I evaluated each project. These criteria are enumerated below and are inspired by dimensions facets in the reports that I consulted:

– *Institutional setting* (galleries, libraries, archives, museums, academic departments)
– *Digitized collection medium* (images, text, video, audio)
– *Digitized collection topic, including time period, subject matter, geographic location, and language*
– *Machine learning methodology employed* (image classification, facial recognition, named entity recognition, etc.)
– *Discipline or field of study* (history, computer science, data art, etc.)
– *Intended audience* (historians, educators, the public, etc.)
– *Deliverable form* (paper, visualization, interface, exhibit)

This process yielded five representative projects:

1. **The Real Face of White Australia**, a project created by Kate Bagnall and Tim Sherratt at the University of Tasmania and the University of Canberra, respectively [339]. The project utilizes facial recognition to uncover photographs of non-white Australians as preserved within the National Archives of Australia in order to "explore the records of the White Australia Policy through the faces of those people."
2. **Citizen DJ**, a project by Brian Foo, an Innovator in Residence at the Library of Congress and a Data Visualization Artist at the American Museum of Natural History [104]. Citizen DJ uses machine learning to extract and sort audio samples from

[32]Relevant surveys including [64] have since been published but were unavailable during this step in the research process.

the Library of Congress's collections and allows the American public to explore the collections by remixing the samples using a hip-hop sampler interface.

3. **The Transkribus platform**, initially "developed by the University of Innsbruck in cooperation with leading research groups from all over Europe as part of the Horizon 2020 EU research project READ" [162, 314]. Transkribus empowers users to train their own OCR models in an interactive machine learning fashion using iterative training with custom sets of typewritten and handwritten documents.
4. **The Distant Viewing project** by Lauren Tilton and Taylor Arnold at the University of Richmond [21]. The Distant Viewing project utilizes facial recognition and image classification to study sitcoms such as "Bewitched" and "I Dream of Genie" through the lenses of media studies and the digital humanities.
5. **Newspaper Navigator** [196, 202, 205], my own project developed in conjunction with the Library of Congress, which utilizes object detection to extract visual content from 16 million historic newspaper pages and reimagines exploratory search by empowering users to train their own interactive machine learners to retrieve images according to user-defined facets. I selected my own project primarily because I could reflect on the subtleties involved as a primary stakeholder.

If a central goal of the "Collections as ML Data" checklist is to aid researchers, one condition for the success of the checklist is applicability across a range of projects. In Table 2.10, I compare the selected projects according to the criteria articulated earlier in this subsection. The range of the selected projects' collection media, machine learning tasks, audiences, and deliverable forms not only speak to the diverse nature of projects within machine learning and cultural heritage but also collectively serve as an important test surrounding the "Collections as ML Data" checklist's relevance.

Iteratively refining the checklist by applying it to these projects as described in the subsection "Overview of Methodology" served an important method for conceptualizing the responsible practices detailed and improving the comprehensiveness of the checklist's questions and suggestions. In the subsection "Applying the "Collections as ML Data" Checklist," I provide vignettes describing the application of the refined "Collections as ML

Project	Collection Medium	ML Task	Audience	Deliverable Form
The Real Face of White Australia [339]	Document Scans (Gov't Documents)	Facial Recognition	Public	"Wall of Faces" Online Interface
Citizen DJ [104]	Audio (Music, Field Recordings, etc.)	Audio Extraction & Similarity	Public	Online Hip Hop Sampler & Exploratory Interface
The Distant Viewing project [21]	Video (TV Sitcoms)	Facial Recognition & Image Classification (at frame level)	Scholars & ML Researchers	The Distant Viewing Toolkit & Scholarly Output
The Transkribus platform [162, 314]	Document Scans (Handwritten & Typewritten)	OCR & Handwriting Recognition	Scholars, Librarians & Archivists, ML Researchers	Online Platform
Newspaper Navigator [196, 202, 205]	Document Scans (Newspapers)	Visual Content Extraction & Similarity	Scholars, ML Researchers & Public	The Newspaper Navigator Dataset & Search Interface + Scholarly Output

Table 2.10: A table categorizing the selected projects as case studies for the "Collections as ML Data" checklist developed in this section.

Data" checklist to each project, revealing the ways in which each project tests the checklist in a different manner.

2.6.6 The "Collections as ML Data" Checklist: An Overview

In this subsection, I provide an overview of the "Collections as ML Data" checklist. An outline of the checklist can be found below. The overview in this subsection is structured around the four central components of the checklist: the cultural heritage collection as data; the machine learning model; organizational considerations; and copyright, transparency, documentation, maintenance, and privacy. The full "Collections as ML Data" checklist is enumerated in the last subsection of this chapter (" The Full "Collections as ML Data" Checklist"). The checklist is partitioned into four components. Each checklist question that is directly inspired by related work is accompanied by corresponding citations. In this subsection, I elaborate each of the four components. In the subsection "Applying the "Collections as ML Data" Checklist," I provide use cases of the checklist.

An Outline of the "Collections as ML Data" Checklist

1. **The Cultural Heritage Collection as Data**
   - (a) Dataset Composition
   - (b) Collecting Process & Curation Rationale
   - (c) Digitization Pipeline
   - (d) Data Provenance
   - (e) Crowd Labor
   - (f) Additional Modification
2. **The Machine Learning Model**
   - (a) Overview
   - (b) Training / Finetuning
   - (c) Evaluation
   - (d) Deployment
   - (e) Release
   - (f) Environmental Impact
3. **Organizational Considerations**
   - (a) Stakeholders
   - (b) Use of Machine Learning
   - (c) Organizational Context
   - (d) Project Deployment & Launch
4. **Copyright, Transparency, Documentation, Maintenance, and Privacy**
   - (a) Copyright
   - (b) Transparency & Re-Use
   - (c) Documentation
   - (d) Maintenance
   - (e) Privacy

The Cultural Heritage Collection as Data

In this component of the "Collections as ML Data" checklist, researchers are encouraged to interrogate and reflect on the cultural heritage collection(s) being utilized in the project as data. In drawing from the dataset assessments described in the subsection "Dataset Assessments," this component emphasizes a clear understanding of the dataset's composition and provenance, including who is represented in the dataset, how the data was collected, and beyond. In accordance with the nuances required in the context of cultural heritage, this component also draws from best practices in cultural heritage in order to encourage researchers to excavate curation, digitization, and any crowd labor utilized in augmenting the dataset. By articulating the contours of the cultural heritage data in question, researchers engage with the sociotechnical implications of treating cultural heritage collections as data, producing documentation that foregrounds these considerations for the project's audience.

The Machine Learning Model

In this component of the checklist, researchers are asked to engage with the subtleties of the machine learning model(s) utilized. The questions in this subsection draw from the model assessments described in the subsection "Model Assessments" in order to address canonical questions surrounding a model's documentation, including its training, evaluation, deployment, and release. This component is tailored to the specifics surrounding cultural heritage collections, including whether the machine learning model has been utilized to make a single pass over a collection or has been continuously deployed, and whether the model has applications outside of cultural heritage. Moreover, this component asks researchers to consider the environmental impact of training and deploying the model. Collectively, the questions in this subsection emphasize both scientific and ethical best practices.

Organizational Considerations

In this component of the checklist, researchers are encouraged to consider the broader organizational considerations of the project, including documenting the stakeholders and organizational context. Because projects may involve stakeholders with different fluencies and experience levels with both machine learning and cultural heritage, this component emphasizes considerations surrounding the subtleties that emerge in this context: do stakeholders have access to gain expertise in these domains? Can this project be utilized to build data fluency at the organization? Moreover, stakeholders are asked to consider a fundamental question that is often overlooked in the machine learning literature: is it necessary to use machine learning in this context, and if so, why? Lastly, the stakeholders are asked to consider the target audiences of the project. Here, stakeholders are asked to return after a project's launch in order to reflect on which audiences were reached and what feedback was received. More generally, this component encourages stakeholders to engage critically with the organizational complexities introduced at the intersection of machine learning and cultural heritage.

Copyright, Transparency, Documentation, Maintenance, and Privacy

The fourth and final component of the "Collections as ML Data" checklist concerns five project components that are essential to a successfully launched project: copyright, transparency, documentation, maintenance, and privacy. Given the complexities raised by copyright in the context of cultural heritage, stakeholders are asked to engage with the ways in which copyright impacts the project's scope and deliverables. Stakeholders are also encouraged to walk through the project's efforts toward transparency, including support for outside audits, as well as the availability of code documentation for reproducibility and reuse. Because GLAMs tend to require stringent considerations surrounding user privacy, stakeholders are asked to report what data on visitors will be collected and whether consent will be requested. Lastly, stakeholders are asked to address plans for maintenance after launch. This final component of the "Collections as ML Data" checklist emphasizes a holistic approach to project development that is often overlooked when considering only the machine learning elements in isolation.

2.6.7 Applying the "Collections as ML Data" Checklist

In order to justify four components of the "Collections as ML Data" checklist and provide concrete case studies, I have included vignettes from each of the projects articulated in Table 2.10. These vignettes motivate the importance of each checklist component and illustrate how adopters of the "Collections as ML Data" checklist might begin utilizing the guiding questions.

Case Studies 1: The Cultural Heritage Collection as Data

Building on the subsection "The Cultural Heritage Collection as Data," the following vignettes demonstrate the subtleties raised by treating cultural heritage collections as data for machine learning projects, whether for training a machine learning model or processing with one.

1. *Dataset Composition*: Though the questions in this subsection are inspired by machine learning-oriented checklists [37, 115, 146], understanding the composition of a

cultural heritage dataset is crucial to *any* cultural heritage project and research. Machine learning only redoubles the considerations that must be made. "Citizen DJ" is an exemplary project from the perspective of how a collection's composition must be taken into consideration. To help guide visitors using the project's hip-hop sampler, Brian Foo has created an ethics guide that is available on the project's main site [105]. The guide walks a visitor through the process of considering a dataset's composition in order to properly address attribution and compensation, as well as cultural and historical contexts. Foo's ethics guide not only educates visitors but also documents the considerations made during the project's development, thereby serving as a valuable project artifact.

2. *Collecting Process & Curation Rationale*: Many cultural heritage collections have decades-long, complex histories surrounding their creation and curation. First, let us consider a collection's creation. In the case of "The Real Face of White Australia," the government documents in consideration were originally produced under the Australian Immigration Restriction Act, namely, certificates granting exemption to the Dictation Test [339]. The project's goal of recovering the people marginalized by these policies can only be understood when considering the documentation's role within the oppressive system. The project is notable for how the project leaders pay close attention to the origins of a collection that documents a difficult history. Next, let us turn to curation. In the case of *Chronicling America*, the newspaper corpus on which Newspaper Navigator is built, the selection process for including a newspaper title within the collection is a highly nuanced process, dependent on criteria enumerated by the Division of Preservation and Access at the National Endowment for the Humanities, as well as state-level contributors [196]. Understanding these curation decisions are essential to properly contextualizing the abundances and lacunae of representation within the corpus.

3. *Digitization Pipeline*: The digitization pipeline can have a profound impact on a machine learning model's predictions, as evidenced by the Newspaper Navigator dataset [196, 202]. Lyneise Williams has documented how the distortive effects of the mi-

crofilming process can lead to erasure of people of color by saturating darker skin tones [398]. This phenomenon is present within *Chronicling America* and digital cultural heritage collections writ large. In particular, the Newspaper Navigator data archaeology demonstrates how embeddings generated by a ResNet model pre-trained on ImageNet fail to retrieve the same photograph of W.E.B. Du Bois among four different digitized *Chronicling America* newspaper pages due to precisely this effect of microfilming distortion. Documenting the digitization process is thus crucial to understanding how a machine learning model processes a cultural heritage dataset [196].

4. *Data Provenance*: Though much of a cultural heritage dataset's provenance will have been articulated in the previous two subsections ("Collecting Process & Curation Rationale" and "Digitization Pipeline"), there may very well be additional salient details regarding the genealogy of the data. Consider, for example, the Library of Congress's "American English Dialect Recordings: The Center for Applied Linguistics Collection" included within Citizen DJ. The recordings in the collection were originally obtained by over 200 collectors [369]. In 1983, the Center for Applied Linguistics obtained these recordings from the collectors in order to improve access, as funded by a grant from the National Endowment for the Humanities. In 1986, the Center for Appplied Linguistics donated 405 audio recordings to the Library of Congress, 350 of which have been made available online (with the remaining 55 withheld due to copyright considerations). The Citizen DJ website incorporates a summary of the data provenance for this collection (as well as for the other collections included) as part of the browsing experience, thereby contextualizing which recordings are available for remixing and foregrounding the mediating factors that have shaped the collection. The complex provenance of collections must therefore be considered when assessing a digital collection's contours.

5. *Crowd Labor*: Many GLAMs are pursuing volunteer crowdsourcing initiatives to engage the public with their collections. Because many volunteers find crowdsourcing opportunities to be enriching and educational, such experiences have something to

offer to both GLAMs and the volunteers who participate [322]. These initiatives stand in sharp contrast to the utilization of outsourced, contracted laborers to improve metadata with datasets, an approach common among datasets in the machine learning community. Crowd workers such as Mechanical Turk workers have been paid extremely low hourly wages [130]. This checklist subsection is provided in order to encourage project stakeholders to consider the project's relationship to labor through the lens of data labeling, to credit those who contributed their labor, and to improve transparency surrounding the project's audience. In the case of Newspaper Navigator, the bounding boxes utilized as training data for the visual content recognition model was derived from *Beyond Words*, a crowdsourcing initiative launched in 2017 by LC Labs that asked volunteers to identify photographs, illustrations, maps, comics, and editorial cartoons on World War 1-era newspaper pages in *Chronicling America*. *Beyond Words* was essential to the project, and volunteers are acknowledged for their contributions.

6. *Additional Modification*: A cultural heritage dataset might require additional modifications in order to compute against it. For example, researchers might have to conduct extra cleaning to address "dirty" OCR or other imperfect metadata. Likewise, researchers may have to make decisions surrounding how to represent uncertainty in dates when specific days, months, or even years are unknown for certain items. However, these alterations made during data cleaning are not always foregrounded. The question in this subsection is intended to provide the project stakeholders with an opportunity to document any such changes made.

Collectively, these vignettes provided in response to the "Collections as ML Data" checklist draw out nuances that would not necessarily be surfaced by responding to a canonical dataset assessment from the machine learning literature. As demonstrated by these examples, data provenance, curation, and processing via digitization pipelines represent significant steps within cultural heritage and must be foregrounded in a project's documentation – as the "Collections as ML Data" checklist encourages.

Case Studies 2: The Machine Learning Model

The following vignettes elaborate on the subsection "The Machine Learning Model" in order to justify the importance of documenting the machine learning model(s) employed across each facet of this "Collections as ML Data" checklist component. All five case studies are highlighted with the goal of foregrounding the diversity of responses that the "Collections as ML Data" checklist elicits surrounding the machine learning components of projects.

1. *Model Details*: The questions in this subsection are intended to provide an overview of the machine learning model(s) being utilized, including: *What architectures are being used? What tasks are they being used for?* These details surrounding a machine learning model are essential to document.For example, all five case studies included in this work utilize neural models. While neural models are noteworthy for their high performance, they are also inscrutable, meaning that it is fundamentally difficult to understand each model's decision-making process. Indeed, the research fields of explainable AI and human-AI interaction are devoted to understanding the implications of relying on such models [392]. Noting that each of the five case studies utilizes an inscrutable neural model would therefore be important context to provide within a checklist response. The questions included in this subsection are standard within the machine learning literature, and all of the case studies address the questions in this subsection within their documentation.

2. *Training / Finetuning*: A unifying theme across the machine learning checklist literature is an emphasis on documenting the training of a machine learning model, which is essential for reproducibility and context surrounding its performance. Projects at the intersection of machine learning and cultural heritage often differ from machine learning research, in that the former do not always emphasize methodological advances in machine learning. As a result, these projects often successfully utilize off-the-shelf algorithms or models, as is the case with Citizen DJ [104], "The Real Face of White Australia" [339], and the Distant Viewing project [21]. In the case of Newspaper Navigator, a pre-trained object detection model was finetuned for the specific task of identifying visual content on newspaper pages [202]. In the case of Transkribus,

end-users are empowered to train their own OCR and handwriting recognition models on datasets that they themselves have curated [162]. All of these projects elaborate on the considerations surrounding the specific approach chosen.

3. *Evaluation*: A model's evaluation is already standard practice in machine learning, and this subsection contains questions that have been operationalized by the field writ large. Within the case studies, evaluation is a common theme. For example, with Transkribus, end-users are provided with extensive documentation surrounding how to evaluate the performance of an OCR model on a corpus of digitized documents. For scholars working with specific collections, performance can vary greatly, and understanding a Transkribus model's accuracy is essential in order to determine whether it should be relied upon to generate transcriptions. The provocations in this subsection are also intended to encourage project stakeholders to consider auditing their systems for fairness and utilizing tools for generating explanations for predictions if necessary. As described in an earlier vignette, neural models pose significant challenges and risks when utilized because they are fundamentally opaque. However, tools for generating post-hoc explanations exist for neural models, and utilizing them represents responsible machine learning practice. The remaining questions in this subsection encourage such considerations surrounding explainability, as well as the analogous considerations for fairness.

4. *Deployment*: In the context of cultural heritage projects, the deployment of a machine learning model can take many forms: from the one-off utilization of a machine learning model to produce metadata or pre-process data (as utilized by all projects considered in this subsection) to the deployment of machine learning models that users can continuously train and use to transcribe scan documents (as is the case with Transkribus [314]) or re-rank image search results (as is the case with the Newspaper Navigator search application [205]). With these different deployments, various considerations must be made: what computational resources will be required to maintain a machine learning model that must be run continuously? Is the model responsive enough to serve predictions in real time? How will the machine learning model's predictions be

utilized? These varied uses motivate the guiding questions in this subsection.

5. *Release*: This subsection serves to elicit elaborations surrounding the model's release, from availability to extensibility: it is important to make models available not only for auditing purposes but also for re-use by the broader community with other projects. Transkribus's model for public release is particularly notable. Transkribus users have made 122 trained models publicly available as of October, 2022 [315]. These models are contributed to the central repository along with corresponding documentation, including who trained each model, what training dataset was utilized, benchmark evaluation scores, and the types of script that the model can process. In the case of "low resource" languages that lack equitable funding surrounding OCR research, Transkribus's models may provide a path forward for scholars. The Newspaper Navigator finetuned visual content recognition model weights have been incorporated into LayoutParser, a pip-installable Python package for document layout analysis [338], presenting yet another option for release and re-use.

6. *Environmental Impact*: It has been well-documented that the training and deployment of machine learning models can have significant carbon footprints, commensurate with vehicular emissions [334, 360]. In the case of the Newspaper Navigator dataset construction, over 5 years of wall clock computing time was required to produce the dataset [202]. The Newspaper Navigator data archaeology documents the emissions produced by the project from training through deployment [196]. In total, this amounted to approximately 226 kg of CO_2 emissions, equivalent to a single cross-country flight. As machine learning projects continue to become more computationally intensive, considering the carbon emissions of a machine learning model and pursuing alternative, less resource-intensive approaches are responsibilities for which all project stakeholders must be held accountable.

Given the range of uses of machine learning demonstrated by the case studies and the different responses that the "Collections as ML Data" checklist could elicit, it is evident that proper documentation using the checklist can help guide stakeholders, project audiences, and external auditors regarding the specifics of how machine learning has been utilized.

Case Studies 3: Organizational Considerations

Following the subsection "Organizational Considerations," these examples serve to demonstrate the value of documenting a project's organizational considerations using the "Collections as ML Data" checklist. As demonstrated below, this component surfaces organizational context that might otherwise not be documented.

1. *Stakeholders*: The questions in this subsection are intended to capture the intricacies surrounding project stakeholders, from considering stakeholder backgrounds to reflecting on whether all relevant stakeholder groups have been included. Given that projects at the intersection of machine learning and cultural heritage often involve researchers from different disciplinary backgrounds, documenting each project member's familiarity with both machine learning and cultural heritage is one key consideration. "The Real Face of White Australia" project team provides this context within their accompanying book chapter, "The People Inside" [339]. In particular, the authors detail the experience of being historians utilizing machine learning within the context of the project. Of course, potential stakeholders beyond the project team must also be reflected upon. For example, within "The Real Face of White Australia," the project team chose to present individuals' photographs even though the individuals themselves cannot consent, as the photos are from over a century ago. The explicit surfacing of this question surrounding the consultation of stakeholder groups foregrounds the challenges that a project in this space must confront, especially surrounding communities represented within the underlying cultural heritage collections.

2. *Use of Machine Learning*: Within the cultural heritage community, special considerations are taken when applying computational methodologies [103]. Within this context, it is imperative to consider why machine learning must be applied. Is the motivation driven by real need or by organizational pressures? In the case of Newspaper Navigator, the data archaeology [196] motivates the utilization of machine learning: not only to improve access at scale but also to re-imagine the search affordances by empowering users to train their own machine learners to retrieve relevant visual content. With "The Real Faces of White Australia," faces could have been extracted

by hand, without the use of facial recognition. However, the project team contends, "We could have manually cropped images from an assortment of files to create an exhibition of faces, but machine processing added the power of scale and the possibility of serendipity. As reactions to the wall have highlighted, the sheer number of faces, arranged in a seemingly endless array, carried both political critique and emotional engagement" [339]. Such reflection clarifies the reasons why the project team chose to apply machine learning in this context and demonstrates the importance of addressing this question.

3. *Organizational Context*: While this checklist is designed specifically with researchers in mind, it is important to note that projects are still undertaken within the context of the broader organization. The questions in this subsection are motivated by the recommendations and guiding questions in the reports detailing best practices for machine learning and cultural heritage that address this larger context. Here, the intent is to ask stakeholders to reflect on how the project can serve the broader institution or organization by improving data fluency and training surrounding both machine learning and cultural heritage. Doing so can have significant longitudinal effects toward the proper operationalization of machine learning in GLAMs and proper ethical considerations surrounding cultural heritage at organizations specializing in machine learning. In the case of Newspaper Navigator, the project's organizational context is documented in an article in *EuropeanaTech Insight* [203], detailing how the project fits into the Library of Congress's digital strategy [192].

4. *Project Deployment & Launch*: While checklists have shown great promise for encouraging researchers across disciplines to document a project during development, it is also essential to encourage reflection after the project has been launched. The questions in this subsection are therefore split into two categories: pre-launch and post-launch. The intent is to ask project stakeholders to reflect honestly on the goals surrounding the project's launch and the successes and failures relative to these goals. In "The People Inside," Sherratt and Bagnall describe the project's reception and respond candidly [339]. This context is valuable not only for those who encounter

"The Real Face of White Australia" but also for researchers who are inspired by the project.

As revealed through these case studies, additional nuances emerge for projects in this space surrounding fluency in both machine learning and cultural heritage at the individual project and organizational levels. With the "Collections as ML Data" checklist, project stakeholders are encouraged to surface these subtleties and critically engage with their motivations for utilizing machine learning as well, both before and after the project's release.

Case Studies 4: Copyright, Transparency, Documentation, Maintenance, and Privacy

Building on the subsection "Copyright, Transparency, Documentation, Maintenance, and Privacy," the following vignettes demonstrate the importance of considering a project's deliverables beyond the typical framings of data assessments and model assessments.

1. *Copyright*: Copyright is an essential consideration of any ML project involving cultural heritage collections, from the collections themselves to the machine learning models, code, and final deliverables. Permissive copyright encourages re-use and opens possibilities for projects to be adapted, whereas restrictive copyright may hinder audiences from even interacting with a project. Citizen DJ illustrates the subtleties introduced by copyright. Brian Foo has created a copyright checklist for users to assess how samples created within Citizen DJ can be re-used [105]. As a project that offers an interactive experience surrounding remixing, articulating how visitors can use the music that they have created not only calibrates expectations but also provides an important educational opportunity surrounding the legal history of sampling. Moreover, all of the project's code is open source and placed in the public domain for unrestricted re-use, meaning that other researchers as well as members of the general public can utilize the code for their own projects. The questions in this subsection of the checklist are included in order to have project stakeholders document the subtleties introduced by copyright in its many manifestations.

2. *Transparency & Re-use*: Making a project's code and deliverables transparent and extensible for re-use is a valuable contribution to the "Collections as Data" commu-

nity. Given that many research projects are undertaken without professional software engineers, the value of sharing code is redoubled. The Distant Viewing Toolkit is exemplary in this regard [20]. The toolkit not only powers the Distant Viewing Project but also makes it possible for other digital humanities and media studies scholars to utilize the toolkit for their own research projects. Indeed, "The Real Faces of White Australia," Transkribus, Newspaper Navigator, Citizen DJ, and the Distant Viewing project all have code documentation in the form of publicly-available GitHub repositories, enabling external auditors to evaluate the systems, as well as researchers to consult their code.

3. *Documentation*: Documentation is essential to promoting transparency and facilitating re-use. While code may be available, it might not be possible to reconstruct the project, given the state of documentation. Additionally, documentation from a project management perspective makes it possible to understand the logistical challenges, coordination, and timelines required to bring the project to fruition. "The Real Faces of White Australia," Newspaper Navigator, Citizen DJ, and the Distant Viewing projects have documented the methodologies employed and the socio-technical contexts for the projects, in addition to making code available. [21, 105, 196, 202, 339]. This documentation makes it possible for the project's audiences and potential auditors to understand the context and considerations taken by the project stakeholders.

4. *Maintenance*: Because the deliverables of many projects at the intersection of machine learning and cultural heritage are digital artifacts, projects oftentimes disappear without warning or no longer function properly due to a wide range of reasons: domain names expiring, codebases becoming deprecated, and data hosting being terminated are just a few examples. For end-users, such outcomes are especially jarring when there is little or no warning. Therefore, providing clear documentation on expected maintenance and upkeep is important for the project's audience. Citizen DJ details the project's expected lifecycle on its 'About' page as an experiment, thereby helping site visitors to set realistic expectations surrounding sustained usage of the tool and

to plan for the project's eventual transformation.

5. *Privacy*: For any project deployed on the internet, it is important to provide users with an understanding of data collection and ask for consent if necessary. While collecting user data may provide a compelling opportunity to improve a project, doing so must be done in a manner that respects user privacy. For example, at academic institutions, researchers are required to seek approval or exemption from institutional review boards surrounding any research involving human subjects, including the collection of online data. For projects involving digital collections, careful consideration of user privacy is particularly pressing, given the long history of libraries' dedication to preserving patron privacy. Will project members utilize analytics tools to track engagement? If so, what kinds of data will be collected? For both Citizen DJ and Newspaper Navigator, no personally-identifiable information is collected on any site visitor, in compliance with the Library of Congress's privacy policy.

As revealed through these case studies, the "Collections as ML Data" checklist encourages stakeholders to document all aspects of a project's deliverables, especially through the lenses of transparency, sustainability, and privacy. By centralizing this documentation, the checklist encourages researchers, as well as the broader community, to foreground these considerations.

2.6.8 Discussion & Future Work

In this section, I have introduced the "Collections as ML Data" checklist, a detailed set of guiding questions for projects that utilize machine learning in the context of cultural heritage collections. The checklist brings together perspectives from two distinct communities and serves as a conceptual model for ethically responsible decisions in the context of applying machine learning to cultural heritage collections. When completed by researchers, the checklist answers can be published along with the deliverables of the project in order to increase transparency and foreground responsible practices. A conceptual model is particularly important in this space, where ethical failures of machine learning projects can be redoubled due to the sensitivities required with cultural heritage collections in particular.

In the subsection "Related Work," I began the paper by contextualizing the "Collections as ML Data" checklist within two bodies of emerging work: the machine learning movement to produce checklists, toolkits, and algorithmic impact statements, as well as the formation of best practices and state-of-the-field reports from the cultural heritage community for utilizing machine learning. In the subsection "Overview of Methodology," I then described my methodology in producing the "Collections as ML Data" checklist. This process included detailing a taxonomy of existing machine learning guidelines ("A Taxonomy of ML & AI Toolkits, Checklists, and Impact Assessments") in order to inform the contributions of the "Collections as ML Data" checklist, as well as to provide researchers with a guide to existing work within the machine learning literature. I also documented the process by which I selected five projects to serve as case studies for developing and evaluating the checklist ("Selecting Representative Projects for Case Studies"). I then provided an overview of the "Collections as ML Data" checklist itself ("The "Collections as ML Data" Checklist: An Overview") and provided concrete examples of applying components of the checklist to the five case studies in order to demonstrate the value of the checklist ("Applying the "Collections as ML Data" Checklist"). This remaining subsection serves to reflect on the "Collections as ML Data" checklist's use in practice, enumerate future work, and conclude the paper.

Applying the "Collections as ML Data" Checklist in Practice

The "Collections as ML Data" checklist is intended to be utilized as a resource throughout all stages of a machine learning project with cultural heritage data, from the initial steps of identifying a digital collection to the final steps of publishing the project's deliverables. Per the recommendations of [85], consultation and engagement with the checklist questions would ideally take place during the design phase, pre-launch, and post-launch in order to allow for the checklist's considerations to be incorporated into the project's development. Moreover, a completed version can be published along with the project's deliverables in order to promote transparency, as well as communicate decisions made and shortcomings faced during the project's development with both the project's audience and potential auditors.

Reflecting on Machine Learning and Cultural Heritage

Though the "Collections as ML Data" checklist has been formulated for projects involving cultural heritage, the considerations within this checklist are relevant to a range of other communities. For example, computational social scientists within political science, economics, sociology, and beyond share many of the same questions surrounding data ethics and machine learning applications. More generally, the checklist offers guidelines relevant to non-cultural data collections, especially in the context of machine learning being applied. It should also be noted that cultural heritage is, in a sense, the substrate of machine learning research and practice. From the text on the web to the photos on Flickr, our collective cultural heritage is utilized ubiquitously by machine learning researchers and product teams across the world as machine learning training datasets [46, 229]. In this regard, the specific considerations surrounding cultural heritage that are offered in the "Collections as ML Data" checklist can be interpreted much more broadly.

Future Work

I will monitor adoptions of the "Collections as ML Data" checklist by researchers in order to understand its broader usage and impact. Additional future work includes the development of different versions of this checklist that can be utilized by different communities within cultural heritage, such as the computational archival science community in the context of automated or semi-automated recordkeeping [142, 325]. Along these lines, I believe one particularly fruitful area of future work to be adapting this checklist specifically for collaborations between digital humanities researchers and professionals at galleries, libraries, archives, and museums who manage, preserve, and provide access to digital collections. Though the "Collections as ML Data" checklist has been subject to many iterations, it is nonetheless not comprehensive. Of course, no such checklist can be entirely comprehensive, and just because a checklist has been utilized does not mean the project should not be interrogated further or documented more extensively. Here, I will paraphrase Ryan Cordell's first guiding question in his "25 Questions for Structuring an ML Project": what is *missing* from the "Collections as ML Data" checklist [70]?

2.6.9 The Full "Collections as ML Data" Checklist

The Cultural Heritage Collection as Data

Here, a distinction is drawn between the cultural heritage collection being studied and the training dataset being utilized for the machine learning model. For example, a project might utilize a pre-trained model to generate embeddings for a photo collection. In this subsection, we consider the cultural heritage collection itself; in the subsection "The Machine Learning Model," we consider the machine learning model's training data.

1. **Dataset Composition**
   - (a) Who or what is depicted in the dataset? [115]
   - (b) If the dataset depicts people, are any specific subgroups of people represented? Are any specific individuals personally identifiable? [115]
   - (c) If the dataset depicts people, are any individuals still living? Does this project comply with privacy laws in countries where it will be shared?
   - (d) What medium is the dataset? (image, video, text, web archive, etc.)
   - (e) How large is the dataset, both in cardinality and in disk storage?
   - (f) What metadata is available for the dataset items? [146]
   - (g) Does copyright impact this dataset? If so, how? [70, 115, 156, 283]
   - (h) Does this dataset pertain to a difficult history? If so, what extra precautions are being taken?

2. **Collecting Process & Curation Rationale** (*language borrowed from* [37])
   - (a) Who curated the cultural heritage collection from which this dataset is derived?
   - (b) What organization or institution was the collection created for?
   - (c) What funding was utilized (if known)?
   - (d) What collection process was utilized? [37]
   - (e) When was the collection assembled? (i.e., when were the photographs taken or ethnographies recorded?)
   - (f) What instruments were utilized to create the collection? (i.e., a recording device, camera, etc.)
   - (g) If people are included, did individuals consent at the time of collection?
   - (h) What were the decision-making processes behind the collection's curation? [37]
   - (i) What is unknown about the collection process & curation rationale?

3. **Digitization Pipeline** (only applicable if the dataset is a digitized version of a physical collection)
   (a) Who selected what was digitized?
   (b) What organization or institution oversaw the digitization?
   (c) What funding was utilized?
   (d) What criteria were utilized for determining what was digitized? [70]
   (e) What were the steps in the digitization pipeline? (For example, in the case of photos, what scanners were used to digitize the documents? In the case of documents, what OCR engines were utilized?)
   (f) What metadata was algorithmically produced?
4. **Data Provenance**
   (a) What is the provenance of the dataset, from collection through digitization? [37, 85, 146]
   (b) Is any part of the provenance unknown?
5. **Crowd Labor**
   (a) Have volunteers or crowd workers added metadata to the dataset? [70, 156, 283]
   (b) If so, how were they recruited and compensated?
   (c) If so, what metadata did they produce? (i.e., transcriptions, annotations, etc.)
6. **Additional Modification**
   (a) Were any additional steps taken after collection curation and digitization in order to produce the dataset in question? (i.e., Were any items removed? Were any additional metadata added? etc.)

The Machine Learning Model

Note: if multiple machine learning models were utilized in the project, this step should be completed for each model.

1. **Overview**
   (a) What model architecture has been utilized? [240]
   (b) What is the task that the model is being deployed to perform?
   (c) Who trained, finetuned, and/or deployed this model? [240]

   (d) Across what organizations or institutions did this training, finetuning, and/or deployment take place? [240]
   (e) What funding was utilized? [115]

2. **Training / Finetuning**
   (a) Was the model trained from scratch?
   (b) If so, what data was used to train the model? [240]
   (c) If not, was a pre-trained model utilized? Where can more information on the pre-trained model be found?[240]
   (d) Was the pre-trained model finetuned? If so, what data was utilized for finetuning?
   (e) If training or finetuning was performed, what computational resources were utilized?

3. **Evaluation**
   (a) How was the model's performance evaluated? [240]
   (b) What data was used for evaluation? [19, 240]
   (c) If the model involves data pertaining to people, has the model been audited for fairness and bias using tools such as FairLearn? [19, 41, 85, 156, 228, 317]
   (d) Have any tools been utilized to generate explanations for predictions (i.e., LIME [321], SHAP [226], TCAV [172]) and modify the model in response? [19, 70, 85, 284, 321]

4. **Deployment**
   (a) How was the model deployed? Was it used to make a single pass over the cultural heritage dataset in question, or will it be continuously deployed?
   (b) What computational resources were utilized for deployment?
   (c) Are the metadata generated by the machine learning model (embeddings, classifications, etc.) available as project deliverables?

5. **Release**
   (a) Has the resulting model been made available for download? (*if no, the following questions can be skipped*)
   (b) What license has been provided? [240]

(c) Who are the primary intended users, and what are the intended use cases? [240]

(d) Does this model have applicability outside of cultural heritage collections?
(e) What are ways that this model could be misused, either intentionally or unintentionally? [227, 240]

6. **Environmental Impact**
   (a) What were the carbon emissions produced by training, finetuning, and/or deploying this model? [70, 190, 360]
   (b) How does the environmental impact of this model compare to that of other components of the project, such as a collection's digitization or stakeholders' flights to relevant conferences?

Organizational Considerations

1. **Stakeholders**
   (a) What stakeholder groups are involved in this project? [70]
   (b) What is each project member's familiarity with machine learning? [70, 156]
   (c) What is each project member's familarity with cultural heritage collections as data?
   (d) Has the project notified and sought input from all potentially relevant stakeholder groups, such as those included within the cultural heritage dataset itself? [227, 317]
   (e) Do groups affected by the project, such as individuals and communities directly represented within the cultural heritage dataset, have an avenue for contacting project staff and seeking recourse? If so, whom should they contact? If not, why not? [85, 240, 317]

2. **Use of Machine Learning**
   (a) Was it necessary to use machine learning for this project?
   (b) If so, why?
   (c) If not, why was machine learning still utilized?
   (d) What are potential critiques of applying machine learning in this context?

3. **Organizational Context**

(a) Can this project be used to build data fluency within the organization or institution? [284]

(b) Do there exist programs or paths for training staff affiliated with the project to develop machine learning skillsets? [70, 284]

(c) Do there exist programs or paths for training staff affiliated with the project to develop fluency with cultural heritage collections?

4. **Project Deployment & Launch**

(a) Who is the target audience of this project? [227]

(b) How does the target audience align with the audiences that the institution or organization is hoping to engage?

(c) If the target audience of the project is the public, does it make an attempt to educate the public regarding the machine learning approaches employed?

(d) *Did the project launch reach the intended audience?**

(e) *Has the project received feedback from stakeholders, including the audience? If so, what feedback has been received?**

(f) *Has the launch of the project resulted in any changes to the project?**

(* = *to be completed post-launch*)

Copyright, Transparency, Documentation, Maintenance, and Privacy

1. **Copyright**

(a) Building on question 1.1.g, does copyright impact the dataset, model, code, or deliverables for the project? [70, 115, 156, 240, 283]

(b) If they are made available, what licenses have been chosen?

(c) If they are proprietary, how does this impact re-use?

2. **Transparency & Re-use**
   (a) Can the project be audited by outsiders? If so, is there funding available to support outside audits? [240, 317]
   (b) Is the code created for the project extensible for other cultural heritage researchers? [284]
   (c) If so, does the project provide any tutorials or toolkits for re-use?

3. **Documentation**
   (a) Does the project have documentation? [167]
   (b) If so, is the documentation interpretable by the project's audience?
   (c) Is the project reproducible to an outside researcher, given the documentation available?

4. **Privacy**
   (a) If the project is hosted online, are data on visitors collected? If so, what kinds of user data are collected? [70]
   (b) Is visitor consent gained before gathering online data? [70]

5. **Maintenance**
   (a) Will the project and code be maintained? [115]
   (b) If so, how frequently, and who will be responsible for maintaining it?

2.7 Conclusion & Future Work

In this chapter, we have demonstrated how large-scale cultural heritage datasets can be constructed using machine learning and how such datasets can be utilized to answer research questions in library & information science and the humanities. Adopting *Chronicling America* as a case study, we first introduced the Newspaper Navigator dataset of extracted visual content across 16+ million digitized historic newspaper pages. We then detailed how the dataset can be utilized for downstream scholarship in library and information science, the digital humanities, critical data studies, print history, Jewish studies, and beyond. With these studies, we have demonstrated the capacity for interdisciplinary scholarship and collaboration. Notably, the Newspaper Navigator dataset has implications for the field of document layout analysis as well: the Newspaper Navigator visual content recognition model has been incorporated into Layout Parser, a "unified toolkit for deep learning based document image analysis," enabling layout analysis with the model in just a few lines of Python code [338].

As more and more cultural heritage collections continue to be digitized, the challenges of rendering these collections searchable only continue to increase. Opportunities to apply emerging methodologies from computer vision, natural language processing, and multimodal machine learning are manifold, with the potential to drastically improve impoverished metadata and thus render these digital collections discoverable. Future directions include not only continuing to apply such methodologies but also developing interdisciplinary research programs that can both build these datasets and answer scholarly questions using them. Let this be a call to action for the development of such programs.

Chapter 3
OPEN FACETED SEARCH

Research in human-computer interaction has repeatedly demonstrated the value of facets in supporting exploratory search and sensemaking. While the faceted search interface remains ubiquitous, it requires a facet taxonomy to be pre-defined and applied to a collection of items prior to deploying a search system. This challenge is particularly pronounced when descriptive metadata is impoverished. We introduce *open faceted search*, a framework in which end-users can define their own facets in an open-domain fashion using interactive machine learning. Under this framework, a user trains a "facet learner" to retrieve items belonging to an open facet by iteratively refining the facet learner's predictions. To instantiate open faceted search, we introduce the publicly-deployed Newspaper Navigator search application, an open faceted search system for 1.5 million photographs extracted from historic newspapers launched in collaboration with the Library of Congress. We report on our ongoing evaluation of open faceted search. To do so, we perform detailed log analysis of the search application's public usage over 46 months of deployment and report on over 42,000 unique user sessions. Our analysis of hundreds of open facets created by users reveals that users desire facets outside of canonical image recognition taxonomies. We further investigate and report on organic user interactions with open faceted search surrounding the definition, training, and application of open facets. We conclude this chapter by introducing zero-metadata open faceted search, an extension of open faceted search that enables the full bootstrapping of open faceted search interfaces for image collections using large language models and multimodal embeddings, even when the images have no associated metadata.

This work surrounding open faceted search was in collaboration with Daniel S. Weld and is partially based on a publication that appeared as a demo at UIST 2020 [205].

3.1 Introduction

Search systems are ubiquitous to the online experience, heavily mediating the information that users encounter on a daily basis. Though some users of search systems have well-defined search needs with specific end destinations in mind (e.g., lookup tasks & question-answering), many users have entirely different search needs, broadly characterized as *exploratory search*. For example, a user might desire to make sense of a collection of information without a specific end destination [233, 341, 395]. In *faceted search*, a taxonomy of metadata facets are applied to items across different attributes, enabling end-users to traverse the facet taxonomy and filter items while browsing. Research in human-computer interaction over the past two decades has consistently demonstrated the utility of facets for facilitating exploratory search in comparison to alternative methods, such as basic keyword search and cluster-based search. Today, faceted search systems remain the default interface type for e-commerce sites, library discovery systems, and beyond.

Unfortunately, faceted search suffers from a primary limitation: it requires the facet taxonomy to be constructed in advance and applied to the items before deploying the search system. Historically, constructing and applying such a taxonomy has required manual intervention, which can be both expensive and time consuming. While approaches to automated and semi-automated facet and taxonomy construction have been offered as solutions to the problem of manual intervention [35, 355, 359], and advances in machine learning have improved the potential for automated facet labeling, the question of how to enable *open facets* to be defined by users remains unsolved. In the case of image search, the old adage of "a picture is worth a thousand words" rings true: captions and metadata inevitably fail to fully describe an image, and the importance of open facets is redoubled. Likewise, open facets could be especially useful for collections with impoverished metadata, as well as those with a wide range of users with drastically different search needs.

In this chapter, we present and evaluate *open faceted search*, a new search framework in which users can define their own facets in an open domain fashion during the search process [205]. Users are empowered to define their own facets through interactive machine learning affordances by iteratively training a "facet learner" to retrieve relevant items described by the desired open facet. Like standard facets, open facets can be applied during a search session.

To demonstrate open faceted search's feasibility, we have successfully built the Newspaper Navigator search application, an open faceted search application for 1.5 million photos extracted from historic newspaper pages and found within the Newspaper Navigator dataset [202]. We chose this setting in particular because these photos had little to no descriptive metadata, and their textual captions contain copious errors due to the open challenge of newspaper text for optical character recognition (OCR). Consequently, constructing a canonical faceted search interface would require significant manual intervention.

To evaluate open faceted search, we analyze two years of user logs for the Newspaper Navigator search application, launched for the general public in September, 2020, in collaboration with the Library of Congress. Studying over 42,000 user sessions, we are able to evaluate organic usage of open faceted search by a wide range of end-users, including scholars, genealogists, schoolgroups, and beyond. We demonstrate the breadth of open facets defined by end-users, indicating a need for openness in faceted search. Moreover, in analyzing open facet training patterns, we show evidence of facet learning for many open facets and demonstrate the capacity for advances in multimodal embeddings to further improve these results.

Lastly, we discuss ongoing research surrounding zero-metadata open faceted search. With this extension, users can interactively bootstrap a full facet taxonomy using large language models, which can then be applied to image collections in a zero-shot fashion. The user can then add open facets to the taxonomy, enabling continual refinement during the search process. Significantly, this makes it possible to generate an interactive open faceted search interface for image collections with no associated metadata. We introduce an initial system for zero-metadata open faceted search and describe future work to fully realize the vision.

In summary, the contributions of this chapter are:

1. We introduce open faceted search, a novel framework that empowers users to define and apply their own facets using interactive machine learning during the exploratory search process.

2. We instantiate open faceted search with the Newspaper Navigator search application, an open faceted search system for 1.5 million historic newspaper photographs publicly deployed with the Library of Congress.
3. We evaluate open faceted search by analyzing the logs of 42,406 user sessions in the Newspaper Navigator search application, resulting in hundreds of trained open facets.
4. We demonstrate the importance of "openness" by showing that 85.5% of user-defined open facets are not covered by ImageNet-1000, revealing a long tail of open facet construction.
5. We introduce ongoing work toward zero-metadata open faceted search, enabling an end-user to interactively bootstrap a full faceted search system for image collections that is open as well, even when the images have no associated metadata.

3.2 Related Work

3.2.1 Faceted Search

With faceted search, users are able to filter search results according to different metadata facets, which describe items across many orthogonal dimensions, including categories and attributes [137]. With hierarchical faceted search, facets are structured hierarchically to enable users to toggle between different levels of granularity. User studies with projects such as Flamenco and beyond have demonstrated that faceted search interfaces facilitate exploratory search, provide fluid interaction between the user and the search engine, and improve the ease of use in comparison to standard keyword search interfaces [92, 93, 136, 138, 140, 182, 249, 354, 411]. Of particular relevance to this chapter is [408], which demonstrates the utility of hierarchical faceted search in the context of exploratory image search using a cultural heritage collection from the Fine Arts Museum of San Francisco. Indeed, hierarchical faceted search has remained an active area of research and has seen widespread adoption across the World Wide Web, ranging from e-commerce sites to travel sites to library discovery systems [391]. The central limitation of faceted search is that the facet taxonomy must be pre-defined and applied to a collection before deploying the search interface [139]. Open faceted search addresses this limitation by augmenting faceted search in order to empower users to define their own facets.

3.2.2 Automated & Semi-Automated Facet Taxonomy Construction

Given that faceted search requires a facet taxonomy to be constructed and applied in advance of deployment, a longstanding area of focus among researchers has been the development of systems for automated and semi-automated facet taxonomy construction. For example, the CastaNet system utilizes an external lexical database to bootstrap up a hierarchical facet taxonomy [141, 355, 356]. The CASCADE system provides an automated workflow that crowd workers can utilize to quickly construct a taxonomy [60]. Kong et al. propose faceted search for the web by automatically extracting facets in an open domain fashion using retrieval, candidate extraction, and facet refining [178, 179]. Other systems that automate the construction of facets include ImageSieve, an exploratory search system for museum archives that utilizes named entity recognition to bootstrap facets [211]; MediaFaces, a system that automates the construction of facets and that was integrated into Yahoo's production image search system (as of 2010) [379]; Strong et al.'s system for generating facets using concept extraction from Wikipedia [359]; CIDER, a system that utilizes Wikipedia for automatic query expansion [148]; and Begelman et al.'s facet construction via tag clustering [35]. Though this body of research proposes methods for enriching facets for exploratory search, none of these approaches provide methods for directly empowering users to dynamically define their *own* facets.

The crowdsourcing literature also includes related work such as REVOLT, a system for collaborative labeling of concepts [55]; Sun et al.'s framework for generating concept hierarchies via crowdsourcing [362]; and Deng et al.'s framework for crowdsourced multi-label annotation [83]. The structured labeling paradigm, which facilitates more consistency during the labeling process by allowing crowd workers more expressivity beyond binary labels, is particularly valuable for open faceted search [183]. Halevy et al.'s framework for "Pay-As-You-Go Integration Systems" for databases describes a methodology for combining heterogeneous data sources into a unified taxonomy and thus is also applicable to this notion of automated taxonomy construction [79].

3.2.3 Image Search Systems & Content-based Image Retrieval

Decades of computer science research has focused on image search systems, many of which leverage content-based image retrieval (CBIR) techniques [16, 50, 96, 143, 149, 173, 176, 290, 294, 308, 350, 358, 377, 388, 413]. Relevant systems include WhittleSearch, which enables the user to provide feedback on returned images by making natural language requests, such as "less pointy" or "brighter" [180]; visual query suggestion, a framework that utilizes a combination of keyword search and visual similarity search to combat the heterogeneity within keyword search results [410]; Fauqueur et al.'s system for providing relevance feedback by combining boolean operations on image regions to search [98]; and FutureView, an image search interface that shows predictive views of potential search refinements in order to support sensemaking [149]. Open faceted search re-imagines content-based image retrieval in the interactive machine learning setting and empowers the user to steer the machine learner to define facets of interest via the rapid retraining.

Of particular note is the image search system CueFlik. Developed throughout a series of papers [7, 8, 9, 102], CueFlik is a system for "interactive concept learning," a form of interactive machine learning that, in the words of the authors, "allows end-users to quickly create their own rules for re-ranking images based on their visual characteristics" via interactive machine learning [102]. The authors performed various user evaluations in order to identify the optimal interface for surfacing examples for the user to train interactive concept learners [102]; the best active learning methodologies for selecting examples for the user to rate [7]; and new affordances for the user to surf through their labeling history while training and re-training the interactive concept learners [8]. Though the interactive concept learning paradigm initially proposed by CueFlik facilitated learning concepts that could be captured with low-level features, open faceted search extends this paradigm to high-level image features as well, enabling the user to define facets that capture high-level, semantic concepts. Moreover the authors comment that "an interesting possibility ... is supporting end-user interactive specification of facets based on the visual properties of images" [102]. Open faceted search addresses precisely this goal.

Additionally, a number of relevant studies consider the text domain for the exploratory search setting. For example, the SearchLens system enables users to define their own "lenses" via keywords that can be combined via complex interactions to refine a search feed in the context of exploratory search [56]. The closely related FeedLens is also relevant in this context [169]. Open faceted search builds on this work to explicitly empower users to rapidly define their own facets in the form of topics or concepts by defining examples at the instance level.

3.2.4 Analysis of Users' Needs in Image and Photo Archives

Across the disciplines of computer science, library & information science, and archival science, researchers have studied user needs surrounding image and photo archives [16, 18, 59, 94, 99, 100, 160, 171, 234, 291, 350, 399]. This body of work reveals the heterogeneity of needs and expectations expressed by users, further motivating user empowerment when searching image collections. Moreover, [58, 62, 66, 94, 100, 160] all provide taxonomies for characterizing user search queries with photo and image archives. Of particular relevance to this chapter are [62], which surveys such taxonomies and considers an in-person, recruited user study of queries for the Library of Congress's American Memory photo archive, and [66], which provides a useful taxonomy of patrons' real queries at the North Carolina Collection at the University of North Carolina at Chapel Hill and the North Carolina State Archives in Raleigh.

3.2.5 Search Log Analysis

Search engine logs have long been utilized to evaluate the performance of search systems [137]. Unlike other evaluation methods, such as qualitative studies, in-person evaluations, and surveys, log analysis provides a non-obtrusive method for studying searches [29]. Moreover, search logs enable analysis at scale. In this chapter, we evaluate open faceted search using the logs of the Newspaper Navigator search application, which cover 46 months of the application's online deployment. In Section 3.4, we describe these search logs in more detail, with a particular emphasis on privacy considerations.

3.3 Introducing the Newspaper Navigator Search Application

3.3.1 Background

To instantiate open faceted search, we built the Newspaper Navigator search application, a publicly-deployed search interface for 1.56 million photographs from the Newspaper Navigator dataset (all of the photos from the dataset published between 1900 and 1963 with confidence scores $\geq 90\%$). Because the photographs in the Newspaper Navigator dataset have limited metadata, and their corresponding captions suffer from manifold errors due to bounding box imprecision and highly-varying optical character recognition, open faceted search has the capacity to improve discoverability for the collection in a meaningful way. Significantly, open faceted search itself was inspired by conversations with users of the Newspaper Navigator dataset, who had specific visual facets that they wished to search for, but no way to formulate relevant keyword searches. Moreover, early users of the Newspaper Navigator dataset varied greatly, from scholars to students in the classroom to genealogists, meaning that their needs and expectations were far-ranging. All of these considerations motivated openness as a key area for extending faceted search.

The Newspaper Navigator search application was built using Flask, Python, and vanilla JavaScript; all machine learning components were implemented with scikit-learn [288]. The application was fully containerized in Docker and hosted using AWS Fargate in collaboration with Chris Adams and John Foley of IT Design & Development at the Library of Congress.

Launched on September 15, 2020, the search application has been continuously deployed without any major service outages. The public search application can be found at: `https:news-navigator.labs.loc.gov/search`, and all code for the search application can be found at: `https://github.com/LibraryOfCongress/newspaper-navigator`.

Here, we detail a sample exploratory search experience with open faceted search. A full demonstration video showing these affordances can be found at: `https://vimeo.com/454247544`.

3.3.2 A Walkthrough of the Newspaper Navigator Search Application

Upon landing on the home page (Figure 3.1), a user can perform a keyword search against the captions of the 1.5 million photos as an initial entry point. Figure 3.2 shows a keyword search of `baseball` and the returned results. In total, 5,427 photographs are returned, the captions of all of which contain the string `baseball`. These results demonstrate both the strengths and weaknesses of keyword search. On the one hand, thousands of results are returned, ranging from portraits of baseball players to action shots to team photographs. On the other hand, this heterogeneity can present challenges in its own right, especially if a user is interested in isolating one specific type of photograph. Doing so is exceedingly difficult with keyword search. Moreover, many relevant results are being lost due to imperfect OCR, as well as the intrinsic limitation of keyword search only returning images whose captions contain the literal phrase `baseball`. Let us consider a scenario in which the user would like to study action shots of baseball players, which will serve as our motivating example for

Figure 3.1: The landing page for the Newspaper Navigator search application (https://news-navigator.labs.loc.gov/search).

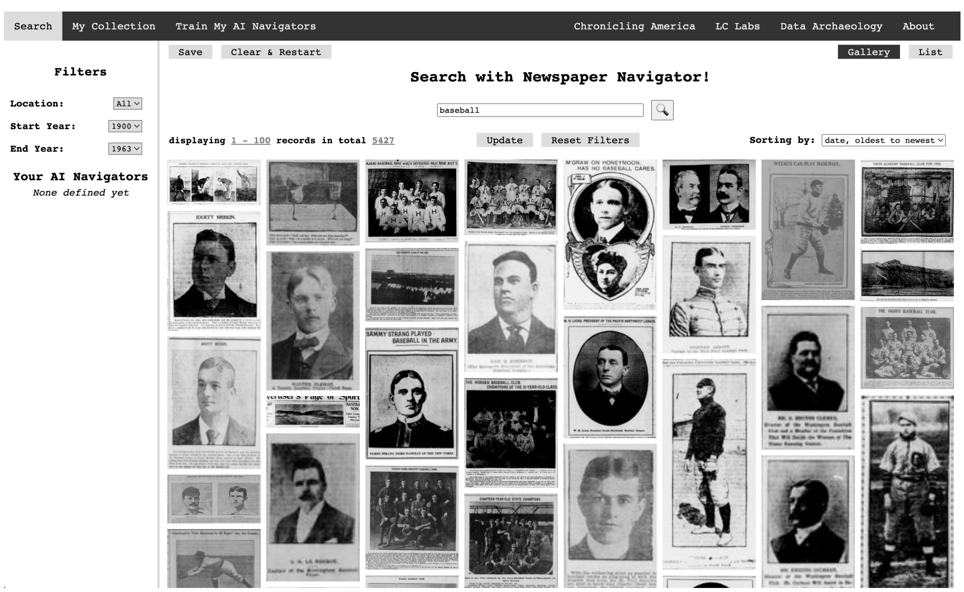

Figure 3.2: The results returned from a keyword search of `baseball`. Note that 5,427 photographs are returned, including portraits of baseball players, action shots, team photographs, and beyond.

Figure 3.3: An example of the modal that appears when selecting for more information on a photograph.

open faceted search in this walkthrough.

Before proceeding to open faceted search, it is instructive to highlight other functionalities in the search application. For example, the user can sort the returned photographs by date of publication or filter by state or year. Upon hovering over a photograph, the user is presented with the option to "+ collection" (add the photograph to "my collection") or "info" (view more information on the photograph). Selecting for more information (Figure 3.3), a modal is surfaced containing additional information about the photograph, including its date of publication, its caption with the keyword highlighted in context, the newspaper title in which it appeared, and options to download a high-resolution version of the photograph, view the photograph in context in the original newspaper page, learn more about the newspaper title, or generate a citation for the photograph.

Upon selecting "+ collection," the photograph is now outlined with a red border. Navigating to the "My Collection" tab (Figure 3.4), the user can view the photographs that they have selected. The metadata for these photographs can be downloaded as a CSV. Moreover, the user can copy a URL that saves the state of the collection to be returned to

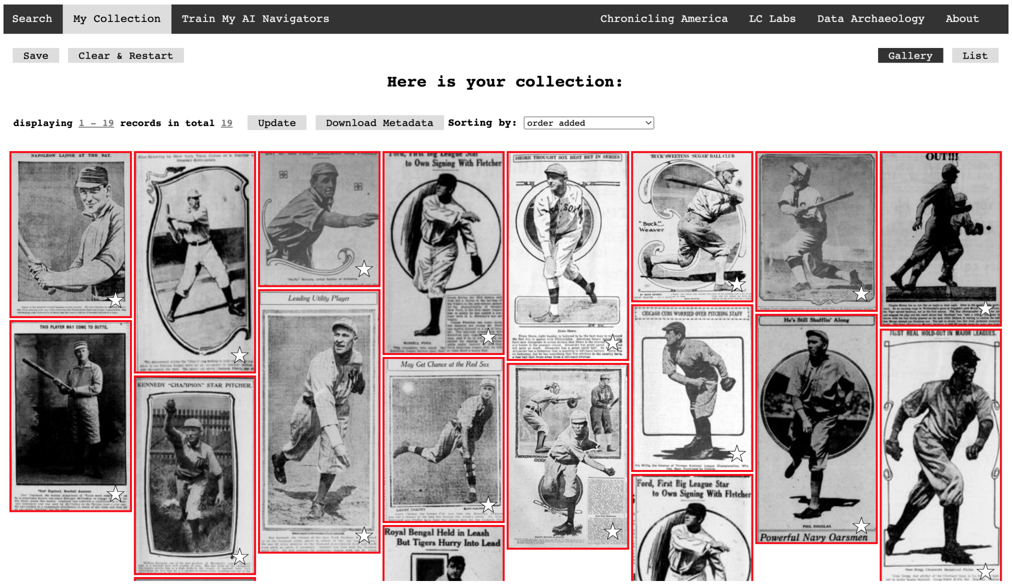

Figure 3.4: Selected images appear in the "My Collection" page, where metadata can be downloaded, and a persistent link can be generated in order to return to the collection or share with friends.

later or shared with friends.

Navigating to the "Train My AI Navigators" tab (Figure 3.5), the open faceted search interface is surfaced. On the top left of this interactive machine learning interface, photographs from "My Collection" appear, at first with a single photograph highlighted with a green border. On the right-hand side, the nearest neighbors to this photograph are displayed in rank order, as determined using pre-computed ResNet-18 embeddings [134]. When hovering over the images on the right, the user has the opportunity to either select "+" and steer the AI navigator toward this photograph, or select "-" and steer the AI navigator away from this photograph. In the language of interactive machine learning, the user is defining positive and negative training examples for their open facet of interest. This open facet (or AI navigator) can be named using the text box in the upper right. When the user clicks the "Train My AI Navigator" button, a linear model is trained on all of the positive and negative examples, along with 1000 randomly-drawn examples from the unlabeled pool, which are treated as negative pseudoexamples. The system then re-ranks all 1.5+ million photos according to prediction score, and the ranked results appear on the right-hand side. Of note

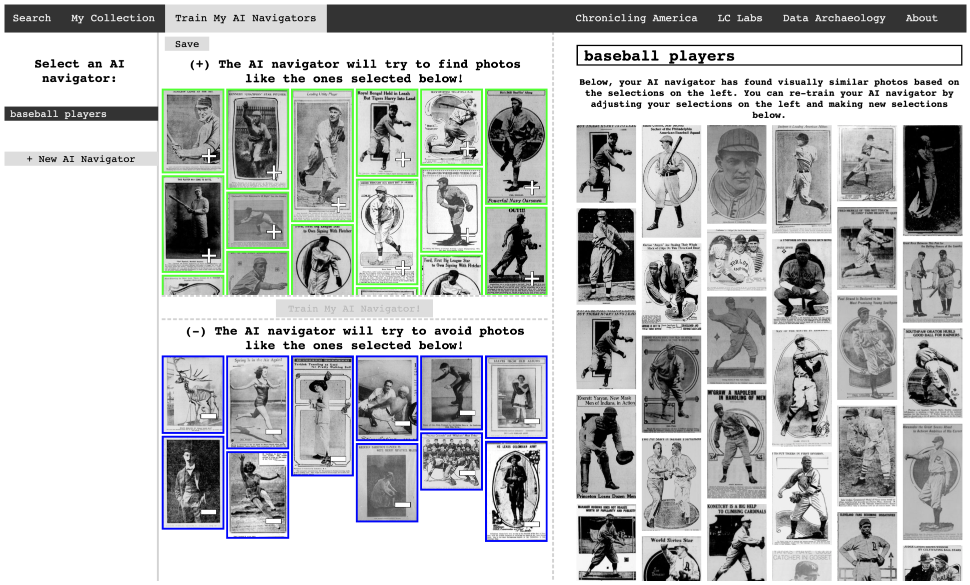

Figure 3.5: The open faceted search interface for the Newspaper Navigator search application, showing an example user session curating a facet learner for a “baseball players” open facet. Newspaper Navigator supports user-defined facets of semantic content and lower-level image features.

is the fact that this process introduces very little latency: from clicking the button, results are returned within a second on the deployed site. Figure 3.5 shows one such example after multiple rounds of training. The user has named this AI navigator “baseball players,” and the large majority of returned results are, indeed, action shots of baseball players. Notably, these photographs have been retrieved entirely based on their visual features, and their captions do *not* need to contain the string `baseball`.

Once the AI navigator has been trained, it appears as a facet on the left-hand side of the main search page, to be applied just like a standard facet. It can be used to formulate complex queries with keyword search or filtering against the other metadata facets.

3.4 The Newspaper Navigator Search Logs

The Newspaper Navigator search logs obtained for analysis contain information on sessions spanning 46 months of public deployment, from September 15th, 2020 (the public launch of the search application) to July 26th, 2022. In order to comply with privacy mandates at

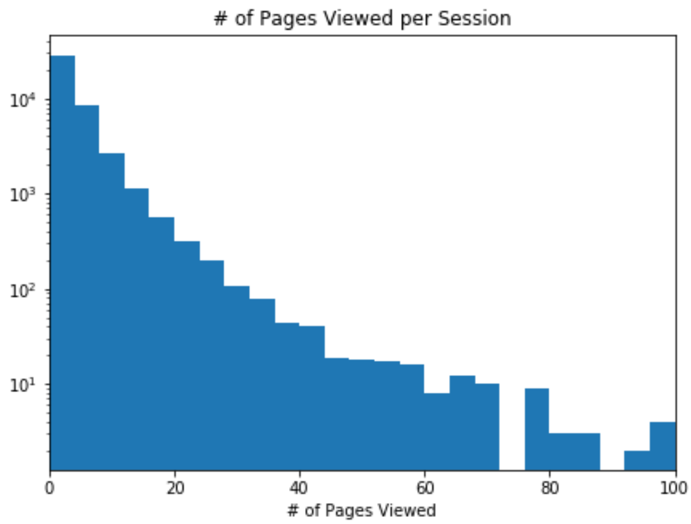

Figure 3.6: A histogram of pageviews per session for all 42,403 user sessions for the Newspaper Navigator search application between September 15th, 2020, and July 26th, 2022. Note the log-scale on the y-axis.

the Library of Congress, these search logs that we received have been modified by staff so that the logs do not contain personally-identifiable information, instead consisting solely of timestamps and URLs of fully anonymized user sessions. This study is in full compliance with the University of Washington's Human Subjects Division.

3.4.1 Overall Usage Patterns

Between September 15th, 2020, and July 26th, 2022, the Newspaper Navigator search application supported 42,403 user sessions that included at least one non-empty keyword search on the application's homepage. Figure 3.6 shows a histogram of pageviews per session. Of these 42,403 user sessions, 821 sessions (1.9%) performed at least one training iteration of an open facet, resulting in 867 trained open facets. Of note is that the Newspaper Navigator search application only included a tutorial video on the homepage and did not include an explicit onboarding tutorial with tooltips – important for contextualizing this 1.9% conversion rate of users to open facet training. Of the 867 trained open facets, 258 (29.8%) were given names by end-users (naming facets was an optional step in the training process). Of the 821 sessions, 50 (6.1%) trained multiple open facets. This section focuses on evaluating these 867 trained open facets in order to understand the benefits and limitations of open faceted search, as implemented within the Newspaper Navigator search application.

3.5 Evaluating Open Faceted Search

3.5.1 Taxonomizing Open Facets

One essential question surrounding open faceted search is to what extent existing facet taxonomies provide sufficient coverage for user interests. If these taxonomies are incomplete, the utility of open faceted search is clear. In the case of the Newspaper Navigator search application, we can analyze the 258 named open facets in relation to existing image recognition taxonomies in order to answer this question. In this section, we consider the fraction of open facets that are already covered by ImageNet-1000, a canonical image recognition taxonomy. We report on additional summary statistics pertaining to the open facets as well.

Open Facet Name Diversity

After de-duplicating the 258 named facets, we find 205 unique facet names, meaning that 53 defined open facets are duplicates. This diversity among open facets being defined by end-users suggests that users have varying search interests, providing initial evidence for the utility of openness.

Open Facet Composition

What is the composition of these facet names? Of these 205 unique names, 45 (22.0%) are proper nouns, 12 (5.8%) contain proper nouns, and 148 (72.2%) consist entirely of common nouns. Significantly, the open faceted search affordance included in the Newspaper Navigator search application is unable to capture most proper nouns (e.g., disambiguating buildings in different towns or recognizing specific people). This suggests that some users have unrealistic expectations for the open faceted search affordance. This could be improved in future iterations through clearer communication surrounding what types of open facets can be defined using the facet learning affordance. Moreover, 51 of the 205 unique facet names (24.9%) concern people, many of which pertain to professions. This interest in searches related to people among users is an important consideration surrounding the Newspaper Navigator photographs.

Open Facet Names and ImageNet-1000

Comparing the 205 unique facet names to the categories in ImageNet-1000, we find that only 22 (10.7%) are covered by the canonical image recognition taxonomy. Of the 258 total facets (including duplicates), 40 (15.5%) are covered by ImageNet-1000. This result shows clear evidence surrounding the importance of open faceted search: existing image recognition taxonomies fail to cover the breadth of user interests. It is important to qualify this with the fact that facet names are imperfect proxies because they may not be clear specifications of what users actually want to search for. A broad facet name may be a confound with the precision of the facet learner (i.e., named as such because the returned results are so varied). Also, facet names may be aspirational in terms of what the user wants to search for but did not actually achieve via training. In the next section, we discuss facet training patterns, but additional user evaluation could be of assistance in clarifying these questions. Lastly, it is important to note that many of the open facets defined within the Newspaper Navigator search application concerned terms of historical relevance that are not included in ImageNet-1000, designed with born-digital images scraped from the World Wide Web in mind. However, this only further motivates open faceted search, where user desires and image collections may diverge greatly from existing taxonomies.

3.5.2 Understanding Open Facet Training Patterns

Figure 3.7 shows a histogram of the number of training iterations for each open facet, both for the 258 named open facets and 867 total trained open facets. Over half of the sessions trained an open facet for more than one iteration, suggesting that users understand the interactive nature of the open facet training process.

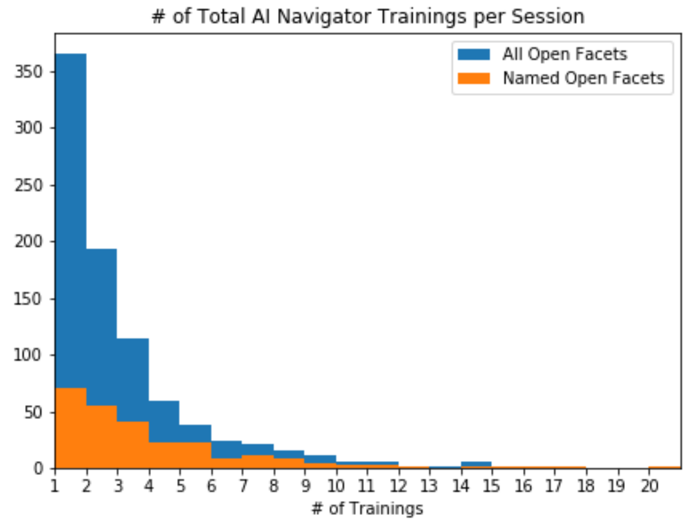

Figure 3.7: A histogram of open facet ("AI navigator") trainings per session. The orange bars correspond to training sessions in which the open facets were also named; the blue bars correspond to all open facet training sessions.

In total, 18 different users changed the name of an open facet during the training process. While some of these sessions reflect a desire to define multiple open facets in the same session, some also reflect a decision to update the facet name as a result of the photographs surfaced during the training process. This indicates that in some cases, the facet names are less reflective of the user's desired query but rather reflective of the results being returned.

3.5.3 Evaluating Open Facet Performance

Using the search logs, it is possible to evaluate open facet performance at each stage of the interactive training process for many of the named open facets. Of the 258 total named open facets, we first remove all such ones that appear in ImageNet-1000 for two reasons: 1) the ResNet-18 embeddings used for visual similarity were trained on ImageNet-1000, artificially biasing the embeddings toward higher performance on these facets, and 2) we are most interested in open facets that do *not* appear in existing image classification taxonomies. We then remove facets containing proper nouns. Lastly, we remove facets that are sufficiently

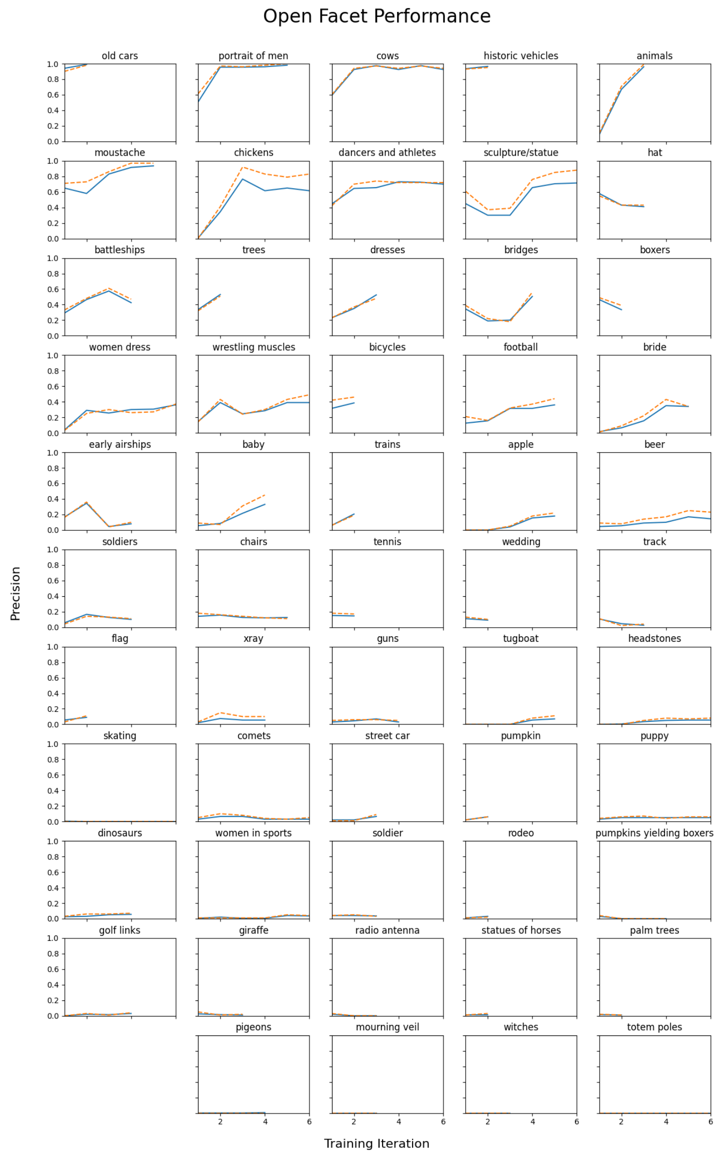

Figure 3.8: Facet performance for 54 open facets defined by users of the Newspaper Navigator search application. In each plot, the training curve shows precision as a function of training iteration, i.e., the performance of the open facet learner at the each iteration of the user's training/re-training of the facet learner. Some curves have higher training iteration values than others because users were free to train each open facet learner for as many iterations as was desired. Precision at 100 is shown as an orange dotted line, and precision at 200 is shown as a blue solid line (200 represents the number of images shown on the right-hand side of the facet training interface in the search application). None of these 54 facets appear in ImageNet-1000. Note that these facets have been trained for variable number of iterations, based on user preference.

subjective, making them difficult to establish ground truth. Of the remaining facets, we evaluate precision at 100 and precision at 200 for each open facet at each stage of the training process. Here, precision at 200 is chosen because the interactive facet learning interface in the search application shows 200 photographs on the right-hand side during training. We label relevant photographs by hand for this evaluation.

In Figure 3.8, we show the individual facet training curves for 54 different open facets. These 54 represent 56% of the unique named open facets (current work surrounds the evaluation of the remaining 41 unique open facets). The individual training curves reveal a range in facet performance, from some facets with near-perfect precision (e.g., "old cars," "moustache"), to others with no facet learning ("mourning veil," "witches," "totem poles"). For 28 of the 54 open facets, we observe precision at 200 of greater than 20% (i.e., at least 40 of the top 200 images). While it is important to recognize that the large majority of open facets do not achieve 50%, relevant results are being returned that can be quickly filtered by end-users during the search process; without open faceted search, this would require manually scanning over a million images or relying on captions, imperfect representations made further unreliable due to OCR mistranscriptions. For facets with higher precision, we observe clear evidence of facet learning, i.e., training curves that improve with more training, suggesting that the interactive training functionality is understood by some fraction of end-users.

Significantly, for a number of open facets including "chickens," "dancers and athletes," "early airships," and "bride," we observe an effect in which the end-users *overtrain* the open facets, resulting in precision to decline in the last stages of training. Amershi et al. observed a similar phenomenon with the CueFlik system for interactive concept learning [7]. This suggests that further work remains surrounding how best to communicate overtraining to end-users.

3.5.4 CLIP Embeddings Improve Open Facet Performance

A central question surrounding the performance of open facets in the Newspaper Navigator search application is to what extent improvements in computer vision would lead to higher precision. Though ResNet embeddings were state-of-the-art when we launched the search application in 2020, much progress has been made. In this section, we consider multimodal embeddings from CLIP, OpenAI's contrastive language–image pre-training model [311]. In particular, we can perform zero-shot open facet learning by querying with a facet name and retrieving the nearest image neighbors in rank order.

As an initial simulated experiment, we consider the final training state of all 49 evaluated open facets in Figure 3.8 with non-zero precision at 200. For the top 200 photos retrieved by open facet, we compute CLIP embeddings and perform the zero-shot facet ranking using the facet name provided by the user (i.e., sorting by inverse distance between the name embedding and photograph embeddings). Utilizing the ground-truth labels from the previous experiment, we evaluate the normalized discounted cumulative gain (NDCG) of the re-ranked images using the CLIP approach [170]. We then compare the CLIP NDCG to the baseline NDCG using the ResNet-18 embeddings. Of the 49 open facets, we find that CLIP has a higher NDCG for 38 and an equal NDCG for the remaining 11 – never lower than the baseline.

To demonstrate the capacity for CLIP embeddings to enable open facet learning even in the most difficult cases presented in Figure 3.8, we perform zero-shot facet ranking for the worst-performing open facets from the Newspaper Navigator search application. Here, we rank all 1.5+ million photos in the search application using the zero-shot approach. In Table 3.1, we report CLIP precision at 10 for these open facets. In all cases except "mourning veil" (0% for both CLIP and the baseline), the CLIP results have many more examples in the top 10 rankings than the original open facets did in the top 200.

3.6 Newspaper Navigator Search Application Impact

In order to complement the log analysis and demonstrate broader impact, we provide an overview of communities who have utilized the Newspaper Navigator search application. The search application has seen usage among scholars, teachers and students in the classroom, genealogists conducting family and local history research, and more broadly, members of the public. A detailed description of the organizational considerations behind the project can be found in a *EuropeanaTech Insight* article co-authored with members of the LC Labs

Open Facet Name	**CLIP Precision at 10**
Golf links	80%
Giraffe	100%
Radio antenna	30%
Statues of horses	60%
Palm trees	100%
Pigeon	80%
Mourning veil	0%
Witches	50%
Totem poles	60%

Table 3.1: CLIP precision at 10 for the worst-performing open facets from the Newspaper Navigator search application. Here, we report precision at 10 because the CLIP results have many more examples in the top 10 rankings than the original open facets did in the top 200 (with the exception of "mourning veil").

team [203].

3.6.1 Scholars

Examples of the broader impact of the Newspaper Navigator project within a scholarly context can be found in Chapter 2. An overview of the search application from a digital humanities perspective can be found in Lorella Viola's review of Newspaper Navigator in *Reviews in DH* [385].

3.6.2 Teachers and Students in the Classroom

The Newspaper Navigator search application has been utilized in classrooms ranging from social studies to computer science. For an overview of usage in the context of social studies, we refer the reader to [200], an article in *Social Education* co-authored with Ilene Berson and Michael Berson, Professors of Education at the University of South Florida. Educators including Jacqueline Katz and Peter DeCraene have utilized the search application in their STEM classes, and Katz's Library of Congress webinar surrounding teaching applications is a valuable resource [168]. Moreover, in his podcast, *The Primary Source*, Tom Bober provided an educator's view of the search application [42].

3.6.3 Genealogists

Among genealogy communities, users have adopted the Newspaper Navigator search application in the context of research surrounding family history and local history. Podcasts such as Lisa Louise Cooke's *Genealogy Gems* [68] and Shamele Jordon's *Genealogy Quick Start* [159] provide overviews of using the search application for genealogical research.

We look forward to continuing to track use cases among different communities during the Newspaper Navigator search application's continued deployment.

3.7 Log Analysis: Next Steps

While the included analysis provides clear evidence for the potential of open faceted search, additional analysis of the Newspaper Navigator search logs would be instructive. First, in taxonomizing the open facets defined by end-users, it would be valuable to apply existing image archive taxonomies such as [58, 62, 66]. In particular, [62] considers an in-person, recruited user study of queries for the Library of Congress's American Memory photo archive, and [66] considers patrons' queries at the North Carolina Collection at the University of North Carolina at Chapel Hill as well as the North Carolina State Archives in Raleigh. Understanding any differences in open facet queries in Newspaper Navigator in relation to this related work would help to clarify expectations surrounding open faceted search and keyword search. Along these lines, taxonomizing the keyword searches within the Newspaper Navigator search application and comparing them with the defined open facets could shed light on this.

In evaluating the performance of open facet learning, future work includes assessing the ceiling of open facet performance through additional training, both with the existing ResNet-18 embeddings in the search application and with CLIP embeddings. Assessing upper bounds in performance is helpful in determining the limits of this implementation of open faceted search, and doing so also allows us to evaluate the gaps between what users achieved through training and what is possible. This in turn sheds light on ways that the implementation of open faceted search can be improved. Moreover, determining how open facet performance varies based on facet type would be interesting to study further.

Additional trends surrounding open faceted search experiences could be interesting to explore. For example, using timestamps, we can evaluate how much time each user spent training each open facet in order to make a case for the responsiveness during a training session. We can also evaluate how many users applied an open facet in a complex query with keyword searches or other metadata facets on the main page. From this analysis, we can learn the combinations of keywords applied with open facets.

Lastly, we can incorporate the findings of this log analysis into the development of future open faceted search systems. Indeed, we begin this work in Section 3.8 toward zero-metadata open faceted search.

3.8 Toward Zero-metadata Open Faceted Search

The previous sections of this chapter have concerned the development and evaluation of open faceted search as a framework for enabling the definition of new facets in an open domain fashion. This begs a natural question in human-AI interaction: can open faceted search be extended to accommodate the bootstrapping and iterative refinement of facet taxonomies as well? This section introduces *zero-metadata open faceted search*, an extension that incorporates precisely these affordances, enabling users to interactively build a facet taxonomy for image collections – in addition to facilitating canonical "openness" through the definition and addition of new facets. Significantly, with the marked advances surrounding large language models (LLMs) and multimodal learning, it is now feasible to interactively generate facet taxonomies and apply such taxonomies to image collections in a zero-shot fashion. Thus, we can build faceted search systems for image collections from scratch *even when the images have no descriptive metadata*. Zero-metadata open faceted search therefore addresses a ubiquitous problem among image collections ranging from digitized cultural heritage (which often have impoverished metadata) to personal photo libraries (which rarely have metadata beyond geotagging and dates).

In this section, we introduce progress surrounding zero-metadata open faceted search. Namely, we introduce an initial system capable of generating fully articulated faceted search interfaces for image collections with no metadata necessary. This system is (1) bootstrapped using large language models (LLMs) and (2) capable of being refined in an open domain fashion through user interaction. This section then turns to current limitations as well as future directions of this work that are necessary to realize the broad vision of open faceted search offered throughout this chapter, especially in relation to human-AI interaction.

3.8.1 Implementing Zero-metadata Open Faceted Search: A High-level Overview

We begin by providing a high-level overview of our implementation of zero-metadata open faceted search.

Step 1: Generating an Initial Taxonomy through Prompt Generation

Recent advances in large language models show great promise for an astonishingly wide range of tasks, including taxonomy generation and enrichment [365]. While previous methods exist for the automated or semi-automated construction of facet taxonomies, large language models offer the potential for the seamless and interactive curation of tailored facet taxonomies for specific collection compositions. For example, we can query a large language model with the following prompt:

> Prompt: "Could you give me a subject matter-based facet taxonomy for browsing a collection of historic photographs?"

The returned taxonomy (in this case from GPT 3.5 [46]), reproduced below, is a two-level hierarchical taxonomy with 5 parent facets and 30 children facets. It is indeed a useful initial taxonomy for exploring historic photographs.

- People
  - Portraits (individuals or groups)
  - Family and social relationships
  - Community life
  - Occupations and professions
  - Sport and leisure activities
  - Special events and celebrations
  - Fashion and clothing
- Places
  - Buildings and architecture
  - Urban landscapes and cityscapes
  - Rural landscapes and scenery
  - Parks and recreation areas
  - Historical sites and landmarks
  - Transportation hubs and infrastructure
- Historical Events & Eras
  - Wars and conflicts
  - Political and social movements

  - Scientific and technological advancements
  - Economic changes and industrializations
  - Cultural and artistic movements
  - Natural disasters and environmental changes

- Cultural and Social Aspects
  - Religious practices and traditions
  - Ethnic and cultural diversity
  - Education and intellectual life
  - Entertainment and popular culture
  - Food and drink culture
  - Health and medicine

- Nature and Environment
  - Landscapes and seascapes
  - Wildlife and animals
  - Plants and vegetation
  - Climate and weather patterns
  - Geological features and natural phenomena

This taxonomy can be further expanded through prompt engineering, such as expanding the hierarchy depth. We explore one such refinement in Step 4.

Step 2: Generating Embeddings for Images

The next step toward zero-metadata open faceted search entails pre-computing multimodal embedddings over the images in the image collection of interest. In this specific instance, we consider embeddings from CLIP, OpenAI's contrastive language–image pre-training model [311]. Significantly, these embeddings are inexpensive to compute and only need to be computed once, so this can be performed upon system launch. Notably, CLIP can be finetuned for specific image collections (such as historic photographs), but we consider out-of-the-box embeddings in this chapter.

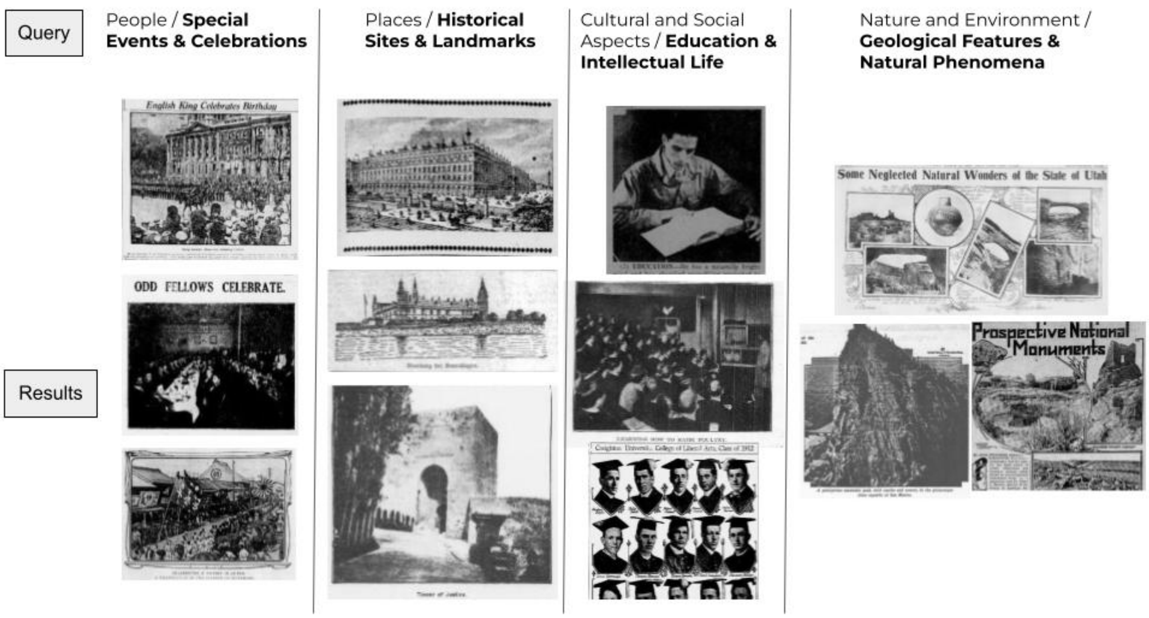

Figure 3.9: Using CLIP embeddings, we query for four different facets from the facet taxonomy generated by an LLM. The displayed results are nearest neighbors, as determined using the pre-computed CLIP embeddings.

Step 3: Compute Zero-shot Facets

The next step is to compute facet-image relevance for all pairs ($n_{facets} \times n_{images}$). As a baseline, we compute the cosine similarity between the text embedding for a specific facet name and the image embedding within CLIP's embedding space. This metric enables us to rank images according to facet relevance, which can in turn be used to populate a facet taxonomy. Significantly, cosine similarity is remarkably inexpensive to compute (a single dot product), so a full facet taxonomy can be applied to hundreds of thousands of images in seconds. In Figure 3.9, we show examples of images retrieved using this method of facet ranking for four different facets: "Special Events & Celebrations," "Historical Sites & Landmarks," "Education & Intellectual Life," and "Geological Features and Natural Phenomena." The returned images demonstrate a clear semantic understanding of these facets. Interestingly, the CLIP embedding space captures an understanding of the text appearing within the images: two of the "Special Events & Celebration" images contain the string "celebrate," and one of the "Geological Features and Natural Phenomena" images contains the text "natural."

Step 4: Adding Open Facets & Refining the Facet Taxonomy

Once an initial facet taxonomy has been created with a large language model and applied to an image collection, the taxonomy can be expanded via open faceted search. A facet of interest to an end-user can be defined using the interactive machine learning affordance introduced earlier in this chapter, or using the zero-shot CLIP retrieval method introduced in Step 3. In either case, once the facet has been applied to the image collection in question, it can be incorporated into the facet taxonomy using prompt generation. For example, providing the following prompt in relation to the taxonomy introduced earlier:

> Prompt: "Where does 'wedding' belong in this taxonomy?"

GPT 3.5 generates the following response:

> "'Wedding' would likely belong under the 'Special events and celebrations' category within the 'People' facet of the taxonomy. This category could include various types of events and celebrations, such as weddings, birthdays, anniversaries, graduations, and religious ceremonies. So, photographs of weddings could be categorized under 'Special events and celebrations' along with other types of celebratory occasions."

Notably, through further prompt engineering, this response can be converted into structured output, such that the returned facet taxonomy is in a usable hierarchical format such as JSON.

3.8.2 An Initial Zero-metadata Open Faceted Search System

Our initial zero-metadata open faceted search system takes as input an arbitrary image collection. As an example image collection, we utilize all of the photographs from the

Newspaper Navigator dataset appearing in newspapers published between January 1st, 1900, and December 31st, 1909. In total, this amounts to 284,651 images. We pre-compute CLIP embeddings for these images.

Bootstrapping an Initial Search System

For an initial taxonomy, we use the two-level hierarchical taxonomy introduced in the previous section, generated using the provided prompt and containing 35 facets (5 parent facets and 30 children facets). If a user wishes to use a different taxonomy, a different prompt can be utilized in this step.

To apply the facets to the image, the aforementioned process detailed in Step 3 is utilized for zero-shot facet learning and re-ranking. Using a single core of a 2nd generation Intel Xeon scalable processor, generating the initial faceted search interface for 35 facets and 284,651 images requires only 18.1 seconds. Notably, the operations could easily be further optimized and parallelized, making it possible to scale this approach to hundreds of facets and millions of images with minimal latency during initial launch.

Figure 3.10 shows screenshots of the system, as implemented in a Jupyter notebook. In this initial interface, any of the 35 facets can be selected, triggering the 284,651 images to be re-ranked according to relevance to the selected facet. In the figure, we show two such facets: "Sport & Leisure Activities" and "Natural Disasters & Environmental Changes."

Putting the "Open" in "Zero-metadata Open Faceted Search"

In addition to supporting browsing over the 35 existing facets, the user can also add open facets of their choosing by entering the facet names in the "New Facet" field. As shown in Figure 3.11, the system automatically (1) queries a large language model to add the new facet to the existing taxonomy, and (2) applies the facet to all of the images using the same methodology utilized for the 35 initial facets. In the case shown in the figure, the open facet "Gothic Architecture" returns salient images and is added to the proper place within the hierarchy, under "Places → Buildings & Architecture → Gothic Architecture." Thus, this implementation of zero-metadata open faceted search supports the interactive expansion of

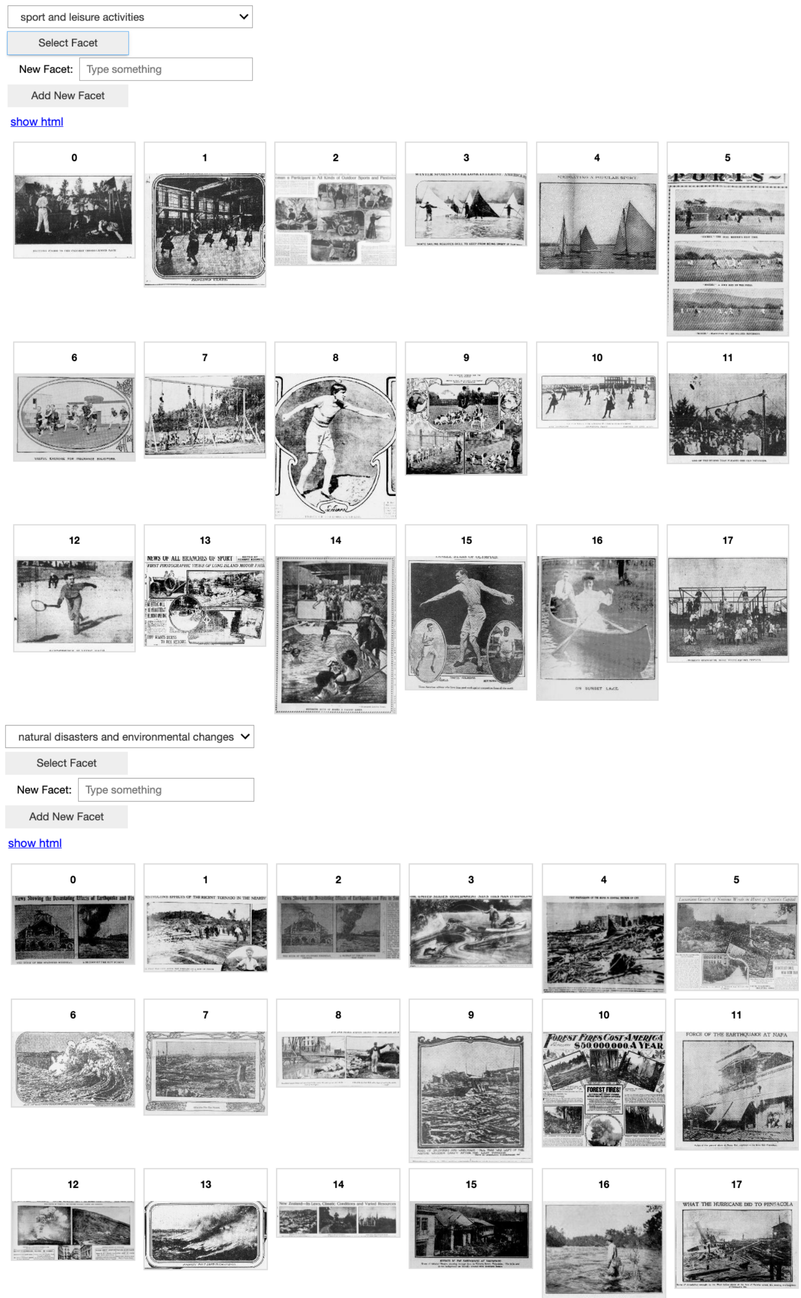

Figure 3.10: Screenshots of our zero-metadata open faceted search interface for browsing 284,651 images from the Newspaper Navigator dataset. In the top screenshot, the top 18 results are shown for the "Sport & Leisure Activities" facet; in the bottom screenshot, the top 18 results are shown the "Natural Disasters & Environmental Changes" facet.

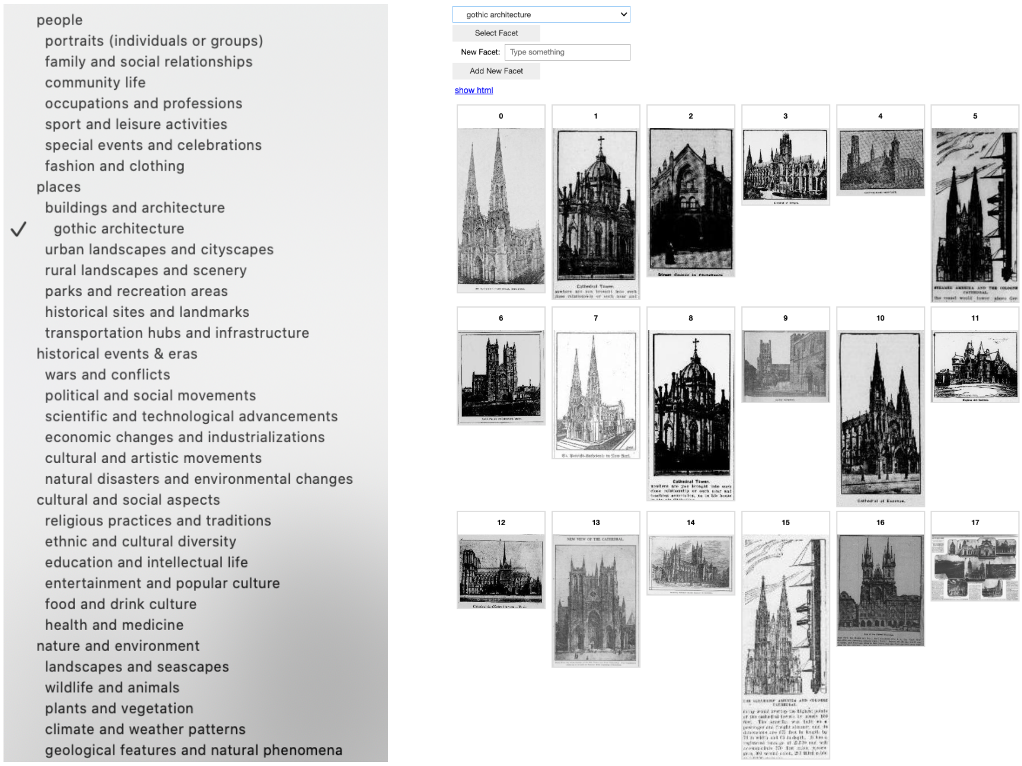

Figure 3.11: An example of open facet functionality within zero-metadata open faceted search. The user adds the facet "Gothic architecture," triggering the system to add this facet to the existing facet taxonomy under the parent facet "buildings and architecture" (left) and apply the facet to all 284,651 images (right).

a facet taxonomy during the browsing experience.

3.9 Future Work

To realize the full vision of open faceted search, especially in the zero-metadata context, much future work remains to be performed. This section serves to detail this work.

3.9.1 Current Limitations

We begin with current limitations. While the initial zero-metadata open faceted search system shows great promise, the current implementation faces a number of open challenges that are important to enumerate.

First, while this system and the Newspaper Navigator search application provide a rank-ordering of facet relevance for images, does not incorporate a notion of facet cardinality, i.e., a binary decision as to whether or not a facet applies to a specific image. Facet cardinality is important because clearly communicating it to end-users is an important part of the exploratory search process and standard functionality among deployed faceted search interfaces online.

Second, while the facet taxonomy itself is hierarchical and encodes relationships between facets (e.g., parent-child and sibling-sibling relationships), the current zero-shot facet ranking process treats each facet independently and does not take these relationships into account. With facet cardinality, doing so is necessary for coherence. For example, every image falling under a child facet must also fall under the parent facet. Moreover, assuming sibling facets are mutex, no image can fall under two sibling facets. Indeed, adhering to principles of faceted search is essential to the user experience.

Third, this system does not allow for facet composition: namely, filtering simultaneously with two or more facets – a core feature of faceted search. This, too, requires a notion of facet cardinality to be implemented. Even with facet cardinality, adjusting how the images are ranked requires further investigation, especially with real users.

Fourth, the interactive process by which open facets are added to the existing facet taxonomy suffers from frequent errors returned by the large language model, placing the facets in incorrect locations. While this process will inevitably suffer from errors, incorporating affordances by which end-users can correct the facet taxonomy interactively is essential.

Lastly, the search interface presented in this section is rudimentary due to the constraints of the current implementation in a Jupyter notebook. Fully developing the interface remains an important next step. Important user interface affordances include functionalities for adding, removing, and moving facets with the facet taxonomy.

3.9.2 Future Work

We begin the future work surrounding paths forward for existing limitations. Concerning the challenge of facet cardinality, active learning methods could enable end-users to quickly establish boundaries for open facet labels. Methods of communicating uncertainty surrounding facet labels could also be explored. These methods have implications for the challenges of facet composition and preserving parent-child and sibling-sibling facet relationships. For example, active learning for selecting joint facet boundaries over sibling facets under the constraint that sibling facets are mutex would not only preserve coherence but perhaps also lead to higher-quality facet labels as well. The same is true for parent-child facets.

While this chapter has adopted photographs from the Newspaper Navigator dataset as the primary image collection of interest, future work with open faceted search includes implementation and evaluation with other image collections, such as personal photo libraries. Significantly, users in this context have very different needs and expectations than users of the Newspaper Navigator photographs, presenting an important opportunity to expand open faceted search's utility. Along these lines, expanding beyond the image domain is another direction of future work. For example, developing open faceted search for large document collections is an exciting opportunity to explore the generality of the framework and further expand the framework's utility. As explored in the next chapter, the challenges surrounding searching large document collections such as the Semantic Scholar scientific paper corpus have resonances with the challenges articulated in this chapter.

Of course, evaluation remains an essential component of future work. While log analysis of 42,000 user sessions of the Newspaper Navigator search application showed clear evidence for the utility of open faceted search, addressing further refinements require proper evaluation, whether through in-person studies or online analysis. In particular, a key step in any future work surrounding the aforementioned current limitations will require proper assessment with real users.

To ensure that this future work surrounding open faceted search addresses real challenges within cultural heritage, we have begun a collaboration with Nic Weber to understand existing limitations and challenges surrounding machine learning-powered exploratory search for digital collections. Across cultural heritage, researchers and practitioners are enriching exploratory search interfaces for image collections using machine learning. However, there remains a paucity of documentation surrounding the processes by which these interfaces have been created and deployed. What are the motivations behind interface choices and interaction affordances? How have these systems been evaluated? What visual facets do end-users desire to search for using machine learning? The goal of this ongoing work is to address precisely these questions through interviews with practitioners and researchers who have created such systems. We have begun by identifying relevant projects at the intersection of cultural heritage, exploratory search, and machine learning. We have also conducted interviews with the creators of eleven such search systems in order to understand the motivations, evaluations, and limitations of existing systems. We are currently in the process of refining recommendations toward the future development of machine learning-based search and discovery for cultural heritage.

Lastly, it is clear that large language models will play an increasingly important role in exploratory search, and open faceted search is no exception. Utilizing large language models to further expand zero-metadata open faceted search toward interactivity via natural language communication is an exciting prospect. The goal of fully boostrapped and articulated open faceted search systems for large collections of images and text remains an exciting one, and the pursuit of these directions will be in service of realizing this vision.

Chapter 4

LIMEADE: FROM AI EXPLANATIONS TO ADVICE TAKING

Research in human-centered AI has shown the benefits of systems that can explain their predictions. Methods that allow an AI to take advice from humans in response to explanations are similarly useful. While both capabilities are well-developed for *transparent* learning models (e.g., linear models and GA^{2}Ms), and recent techniques (e.g., LIME and SHAP) can generate explanations for *opaque* models, little attention has been given to advice methods for opaque models. This chapter introduces LIMEADE, the first general framework that translates both positive and negative advice (expressed using high-level vocabulary such as that employed by post-hoc explanations) into an update to an arbitrary, underlying opaque model. We demonstrate the generality of our approach with case studies on seventy real-world models across two broad domains: image classification and text recommendation. We show our method improves accuracy compared to a rigorous baseline on the image classification domains. For the text modality, we apply our framework to a neural recommender system for scientific papers on a public website; our user study shows that our framework leads to significantly higher perceived user control, trust, and satisfaction.

This work was done in collaboration with Doug Downey, Kyle Lo, and Daniel S. Weld, and is based on a publication that is forthcoming in the *ACM Transactions on Interactive Intelligent Systems* (TiiS) Special Issue: "Human-centered Explainable AI" [201].

4.1 Introduction

A long-standing vision in AI is the construction of an *advice taker*, a system whose behavior, in the words of John McCarthy, "will be improvable merely by making statements to it, telling it about its symbolic environment and what is wanted from it. To make these statements will require little if any knowledge of the program or the previous knowledge of the advice taker" [236]. Indeed, today's guidelines for human-AI interaction dictate that ML systems should be able to explain their predictions to end-users and accept advice and corrections from them [6, 11, 163]. Both explanation and advice-taking methods exist for transparent models, such as linear classifiers or generalized additive models (GA^{2}Ms) [51, 389, 403], and their benefits for transparent recommenders have been demonstrated within the human-in-the-loop machine learning and human-AI interaction literature [43, 185]. These advice-taking approaches allow the human to provide high-level feedback on how specific input features should be driving the transparent model's behavior. In our related work (Section 4.6), we elaborate on such approaches.

However, opaque models, such as boosted decision forests and deep neural networks, are a different story. Because they often provide the highest performance and are widely used, numerous researchers have investigated methods for generating post-hoc explanations of opaque ML models — typically by creating a transparent approximation to the opaque model, called an explanatory model [125]. Several researchers have developed methods for translating high-level *human advice* into specific classes of differentiable, neural models [88, 215, 323, 327, 332], but to our knowledge only Schramowski *et al.* [332] have introduced a method that works for *arbitrary* opaque models, and it is not capable of handling advice that corrects an agent's erroneous predictions (Section 4.6.3).

Furthermore, even the advice-taking methods whose application is restricted to specific opaque model classes [215, 323, 327] have limited empirical evaluation, often restricted to datasets that have been artificially biased (*e.g.*, Decoy MNIST and Iris-Cancer [327]) in a way that a simple human tip (*e.g.*, "Ignore the artifact in the lower right corner") can correct the problem. To demonstrate that advice-taking methods are useful in actual practice, experiments with large, real-world domains seem essential.

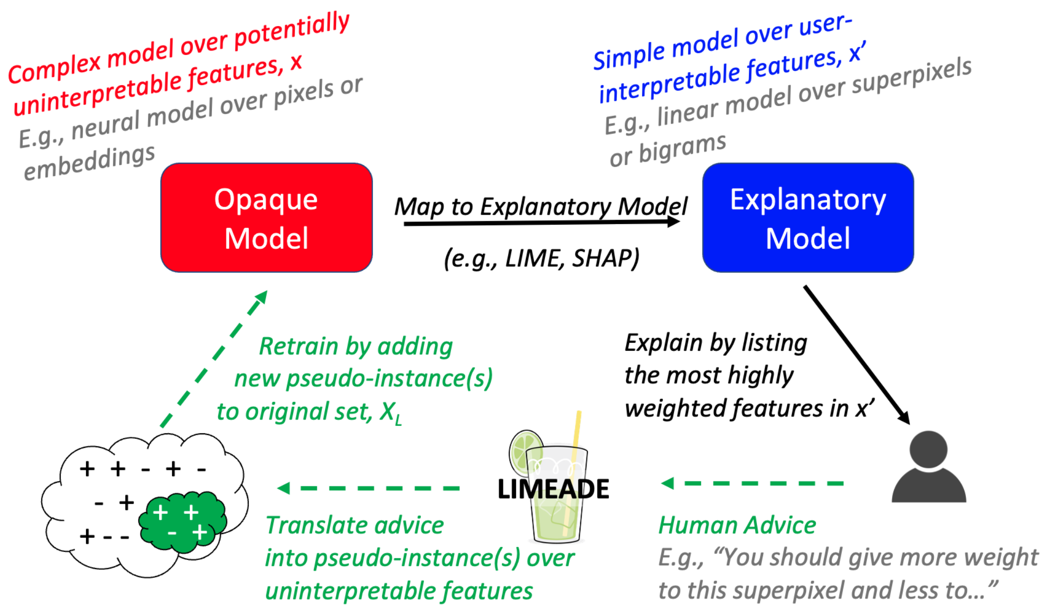

Figure 4.1: LIMEADE takes a user's advice – given in terms of features of the explanatory model — and then modifies the original, opaque model by retraining. This is challenging because the mapping from opaque to explanatory model is typically many-to-one and hence not invertible.

Thus, two central questions for human-AI interaction remain unanswered:

1. Can one translate high-level human advice into a correction to an arbitrary, opaque, machine-learned model which uses a different set of features than those used to express the advice?

2. Do these methods allow end-users to improve the accuracy of natural, real-world models more easily than by simply annotating more instances?

This chapter answers the first question affirmatively, but presents mixed results on the second. Specifically, we present LIMEADE, a general framework for updating an arbitrary, opaque machine learned model given high-level human advice, e.g. phrased in the same vocabulary used by a posthoc explanation of its behavior. As shown in Figure 4.1, our approach builds upon explanatory approaches such as LIME [320] and SHAP [226] that describe the local behavior of a model in the region of a given instance. Given a trained model and an instance to be classified, these post-hoc approaches output an explanation in the form of a weighted list of *interpretable* features (typically distinct from the features utilized in the opaque model) that influence the instance's classification. With LIMEADE, a user can then provide feedback in the same high-level terms as the explanation in order to modify the original, opaque model. LIMEADE converts this user advice back into the original feature space of the opaque classifier by generating pseudo-instances representative of these features and retraining. Unlike other methods intended for machine learning practitioners and model developers, LIMEADE empowers end-users with little or no machine learning expertise to tune the system.

LIMEADE builds on the longstanding research areas of human-in-the-loop machine learning and interactive machine learning to provide a framework that is sufficiently general to address a wide range of model architectures, tasks, and modes of advice. We emphasize that LIMEADE is a general framework in three distinct senses:

1. LIMEADE can be utilized for a wide range of advice-taking applications, from explanatory debugging to personalized recommendation.
2. LIMEADE is architecture-agnostic and enables advice taking for different types of opaque machine learning models, including both classifiers and rankers.
3. LIMEADE accepts different types of human advice (in this chapter, we focus on advice given as binary feedback in terms of high-level features).

Accordingly, we show that our framework is general by demonstrating its success on seventy real-world models across two broad domains: image classification and text ranking. For our first case study, we use LIMEADE to give advice to twenty binary image classifiers (*e.g.*, models predicting "giraffe" or "not giraffe") that are built on precomputed neural embeddings [134]. Our implementation of LIMEADE in the image domain translates a human's simulated advice to the classifier in response to a LIME explanation expressed using superpixel features. Using this simulated advice, we demonstrate that this implementation significantly improves system accuracy, compared to a strong baseline, in a few shot setting. To accelerate future research, we are releasing our LIMEADE image domain code at: `https://github.com/uw-hai/LIMEADE`.

To establish the generality of our approach, we perform a second case study in a very different domain with a different task. In this second case study, we incorporated LIMEADE within Semantic Sanity, a publicly-deployed research paper recommender system with hundreds of users. While recommendations are made using an opaque neural model built on top of precomputed paper embeddings [63], LIMEADE allows humans to provide advice in terms of unigrams and bigrams (*e.g.*, marking them as of interest or not) that are suggested by an approximate, linear explanatory model. In a simulation study based on organic user feeds in the log data, we show that explanation-based advice taking improves recommender quality, but we fail to find a significant improvement compared to adding a comparable number of labeled instances. We also perform an in-person user study showing that users feel that the ability to provide high-level feedback significantly improves their sense of trust, control and system transparency.

Moreover, our work reveals that some ways of soliciting user advice may cause tension between explanation quality and advice diversity, potentially limiting the user's ability to adjust the ML model. We observed this *explanation-action tradeoff* in our second case study, where constraints on the user interface allowed us to accept advice on just a small number of the potential explanation terms. Such advice created a feedback loop, powered by iterative applications of advice, that reduced explanation diversity and hence limited users' future opportunities to further improve the classifier.

Significantly, our chapter leaves a number of questions surrounding the advice-taking problem unanswered. First, we do not conclusively answer the question of whether advice-taking methods allow end-users to improve the accuracy of real-world, opaque models more easily than by simply annotating more instances. Moreover, in our image domain experiment, we uncover that the effectiveness of advice-taking methods may decrease with more supervision. Finally, it is important to further study how advice-taking fits into broader human-in-the-loop machine learning frameworks that incorporate human interventions surrounding design parameters, model and algorithm choice, error tolerance, and beyond [6]. In many ways, we view this chapter as a "Call to Action" to galvanize more researchers to study the advice-taking problem for opaque machine learners, as it is a rich area of study within human-AI interaction with many questions still to answer.

4.2 LIMEADE: Advice Taking for Opaque Models

In this section, we provide a formal overview of the LIMEADE framework and detail how it can be applied to opaque machine learning models to enable advice taking. With LIMEADE, we assume that the human would like to give advice to an opaque machine learning model. By *opaque*, we mean that the model architecture may be completely unknown, or (if known), it may have too many parameters and nonlinearities for a human to understand. However, we assume that the model's inputs and outputs are available and that the model can be retrained on new instances. We work in a semi-supervised learning setting, in which the goal is to learn a hypothesis that maps an s-dimensional real-valued input vector to a label (for classification) or a real-valued output score in $[-1, 1]$ (*e.g.*, for recommendation). We are given a set $\mathcal{X}_L$ of labeled training instances (x, y, w), where $x \in \mathbb{R}^s$, y is the value to be learned, and w is the weight assigned to the instance when training. Additionally, we optionally have a large, dense pool $\mathcal{X}_U$ of unlabeled instances (x). Our explainable machine learning problem setting closely follows that of previous work in explainable ML [226, 320]. We assume that each instance x can be represented as a binary-valued vector x' that lies in an *interpretable* space. For example, in the text domain, the dimensions of x might contain embeddings produced by a transformer [381], whereas the dimensions of x' would correspond to interpretable features such as term frequency-inverse document frequency (TF-IDF) values for n-grams.[1] In the image domain, the dimensions of x would be pixels, while the dimensions of x' might be superpixels [320] or fine-grained features [4, 177].

Given an instance x to explain, our approach uses an *explanatory model* g in the interpretable space that locally approximates the opaque classifier f, i.e., $g(h'(z)) \approx f(z)$ for z' nearby x'. The model g can be any interpretable model, such as a decision tree or linear model, produced using LIME or a comparable method. We refer to the method that produces g as $\textsc{Explain}(f, x, h')$.

Algorithm 1 details LIMEADE's approach to enabling a model to take advice, and Figure

[1]Term frequency-inverse document frequency, or TF-IDF, is a method of text featurization. Each document is featurized according to a fixed set of n-grams, where the feature value for the n-gram is given by the n-gram's frequency in the specific document in question, times a weight that puts more emphasis on terms that are less common across the corpus [351].

4.2 illustrates a concrete example of applying LIMEADE on the paper recommendation domain. Given an instance of interest, x, we obtain an explanation $g(x')$ of the model's output $f(x)$ using $\textsc{Explain}(f, x, h')$. The human can then provide a label on a feature of x'. Informally, a positive label on feature j of x' represents the human's assessment that instances z' near x' should tend to be positive when $z'[j] = 1$. For example, a user of our paper recommendation system might give a positive label to the term "BERT" in a natural language processing paper to indicate interest in papers about the technique. Notably, a user's feedback is provided in terms of the high-level vocabulary of the explanatory model, not by eliciting new features to be added.

Algorithm 1 Enabling an opaque model to take advice using LIMEADE. Given a set of required inputs, LIMEADE solicits human advice in response to an explanation of a classified instance and retrains the opaque model accordingly. EXPLAIN is a function that generates an explanation for a given model and instance.

Inputs:
$\mathcal{X}_L, \mathcal{X}_U$ // sets of labeled and unlabeled instances
$f_t : \mathbb{R}^s \to [-1, 1]$ // opaque classifier, version at time t
$x \in \mathbb{R}^s$, $x' \in \{0,1\}^{s'}$ // instance & instance in interpret. rep.
$h' : \mathbb{R}^s \to \{0,1\}^{s'}$ // mapping s.t. $x' = h'(x)$
$\pi_{x'} : \{0,1\}^s \to \mathbb{R}_+$ // weighting based on distance
$k \in \mathbb{N}$ // number of pseudo-instances
$g_t = \textsc{Explain}(f_t,\, x,\, h')$ // obtain explanatory model
$\textsc{Display}(g_t,\, x')$ // display key features of $g_t(x')$ to end-user,
// who then selects one feature (indexed j) as + or − indicator of instance
receive $a \in \{-1, 1\}$ **and** $j \in \{1, \ldots, s'\}$
// select k instances, label them using action a, and weight according to distance from x'
$\mathcal{N}_x \leftarrow \{\}$
for $1, \ldots, k$ **do**
  $\tilde{x} = \textsc{GetInstance}(x, x', \mathcal{X}_U)$
  if $h'(\tilde{x})[j] = 1$ **then**
    $\mathcal{N}_x \leftarrow \mathcal{N}_x \cup \{(\tilde{x}, a, \pi_{x'}(h'(\tilde{x})))\}$
  end if
end for
$\mathcal{X}_L \leftarrow \mathcal{X}_L \cup \mathcal{N}_x$
$f_{t+1} \leftarrow \textsc{Retrain}(\mathcal{X}_L,\, f_t)$
return f_{t+1}

LIMEADE uses the human's action to improve the opaque model f by creating a set of k training pseudo-instances with repeated calls to $\textsc{GetInstance}(x, x', \mathcal{X}_U)$. We experiment with two implementations of GETINSTANCE: sampling and generative. Sampling from the unlabeled pool is effective when the unlabeled pool is relatively dense, meaning one can acquire many instances with interpretable features similar to those of x'. Generative approaches can be helpful when data are less dense. For example, with images, LIMEADE can create synthetic pseudo-instances by greying out random subsets of the superpixels in the input image, essentially reversing LIME's process for generating the explanatory model, g. The generative approach also works in the textual domain, *e.g.*, by creating a synthetic document with nothing but the tokens selected by the user.

LIMEADE only retains the pseudo-instances that contain the acted-upon feature j, i.e. those $\tilde{x}$ for which $h'(\tilde{x})[j] = 1$. LIMEADE then assigns a value to each pseudo-instance according to the user action: $+1$ if the user assigned a positive feature label, and -1 otherwise.

LIMEADE assigns each pseudo-instance a weight based on its proximity to x', with instances more similar to x' given higher weight.[2] The reasons to weight local instances more highly are twofold: the explanatory method may only be locally correct [320], and the human actions may only be locally applicable. For example, the positive label on "BERT" discussed earlier is helpful within the local scope of natural language processing papers, but could become misleading if applied globally—in biology papers for example, the term "BERT" often refers to a different meaning (the "BERT gene"). After selecting and weighting the pseudo-instances, LIMEADE can optionally condense the selections (e.g., collapsing the instances into a single centroid). Finally, LIMEADE adds the resulting pseudo-instances to the labeled training set $\mathcal{X}_L$ and calls Retrain to train the classifier f on the new data set.

While Algorithm 1 is written in terms of binary classification, our approach generalizes naturally to the multiclass setting. This would entail that step 3 in Algorithm 1 solicit not only which feature was a positive or negative indicator, but also which class—pseudo-instances would then be labeled in step 8 with respect to the chosen class. In the case of negative indicators in the multiclass setting, the pseudo-instances could be assigned random classes other than the chosen class.

We reiterate that LIMEADE is general in many senses: our framework is model-agnostic,

[2] We measure proximity in the interpretable space, but it is equally possible to measure in the original space instead.

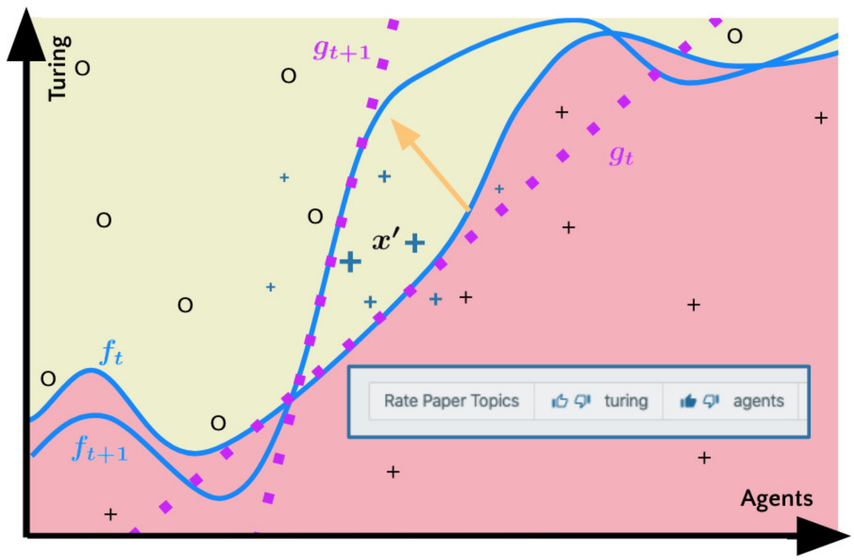

Figure 4.2: LIMEADE updates an arbitrary opaque ML model by creating pseudo-instances. Here, we consider a recommender system for papers. Small black o's and +'s show the original training set (here, a user's ratings of papers), and shaded regions denote the complex boundary of the opaque classifier f_t. In order to explain a prediction, $h(x')$, the system generates a locally faithful explanatory model using LIME or an alternative method. This is g_t, shown as a purple dotted line. In practice, the explanatory model likely has many more than the two dimensions shown above, but suppose 'Turing' and 'agents' are highly weighted terms, hence used in the explanation. When the human specifies feature-level advice, e.g., 'I want more papers about "agents"', it could be used to directly alter a linear explanatory model (creating the new purple dotted line g_{t+1}); however, no simple update exists for an arbitrary, opaque classifier, which may be nonlinear and use completely different features, such as word embeddings. Instead, LIMEADE generates positive pseudo-instances (shown as blue +'s) that have the acted-upon feature and are similar to the predicted instance. The pseudo-instances are weighted (shown by relative size) by their distance to the predicted instance x' that was used to elicit feedback. By retraining on this augmented dataset, LIMEADE produces an opaque classifier that has taken the advice, shown as a changed nonlinear decision boundary f_{t+1}.

applies to a diverse range of advice-taking applications, and enables different forms of advice taking. In the next sections, we present two case studies that highlight the general applicability of LIMEADE.

4.3 Case Study 1: LIMEADE for Image Classification

We now present our evaluation of LIMEADE in the image domain in order to study whether LIMEADE allows humans to update real-world models more effectively than simply labeling more instances. In particular, we use LIMEADE to enable updates based on simulated end-user advice for twenty deep neural image classifiers, *e.g.*, a skateboard detector or fire hydrant detector. In Figure 4.3, we illustrate an example of how LIMEADE is used to process high-level advice in this context. With this simulated image domain experiment, we wanted to study the following research questions:

1. Does advice taking with LIMEADE further improve classifier performance as compared to a rigorous baseline of adding more labeled instances?
2. How do LIMEADE-powered improvements change as a function of supervision?

4.3.1 Experimental Setup

In order to determine whether LIMEADE can support advice taking in the image domain, we evaluated on binary image classifiers, each comprising a logistic regression model trained on pre-computed image embeddings. As a base image dataset, we utilized $20,000$ images from the COCO dataset [210]. In order to create superpixel features for LIMEADE feedback, we leveraged the same segmentation algorithm [288] used by LIME to compute superpixels for all 20,000 images. To generate embeddings for all full images and corresponding individual superpixels, we retrieved their representations from the penultimate layer of a ResNet-50 backbone pre-trained on ImageNet [82, 134]. For a given superpixel, we computed the corresponding embedding by feeding the mimimum bounding box containing the superpixel to the embedding model. Pre-computing these embeddings resulted in a bank of embeddings for 20,000 images along with embeddings for all corresponding individual superpixels.

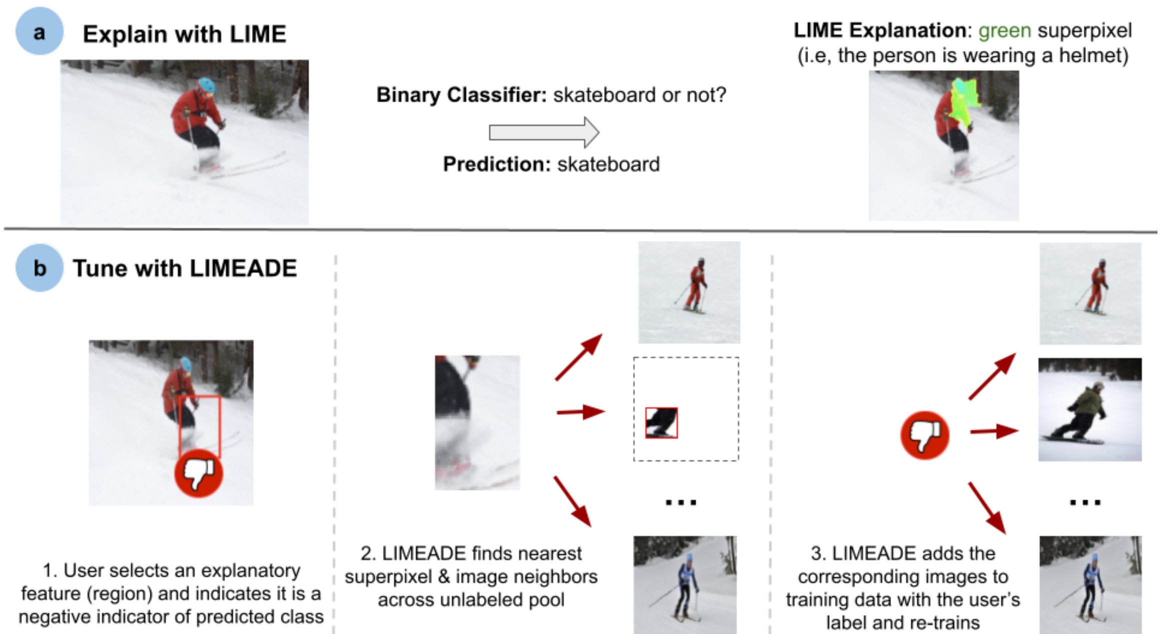

Figure 4.3: **a)** Suppose an opaque classifier incorrectly classifies an image of a skier as a positive instance of skateboarding. Suppose further that LIME returns an explanation showing a superpixel containing the skier's helmet as a positive indicator of the skateboarding class. Having seen this explanation, the user realizes that the classifier is predicting "skateboard" based on a spurious confound and should be looking elsewhere (we note that the end-user, such as a crowd worker, must understand the classification task but needs neither domain-specific knowledge nor an understanding of machine learning). **b)** While a helmet is an appropriate positive indicator for the skateboarding class, the user gives the advice that another superpixel, containing skis and ski poles, is a negative indicator. LIMEADE translates this advice by updating the opaque model and retrieving unlabeled images and superpixels most similar to this ski superpixel (in our experiment, we retrieve the 50 most similar). The corresponding full images are then added to the training data — with negative labels — and the model is retrained, completing the LIMEADE update. In general, a false positive classification will lead to negative feedback, and a false negative classification will lead to positive feedback (as illustrated in Figure 4.2).

In order to ensure that our embeddings had not already been trained on the target classes in our experiment, we tested binary classifiers only on all 20 classes that are in COCO but not in ImageNet-1000.[3] We wanted to measure the performance of a LIMEADE update relative to a baseline update, so we completed 100 randomized initial configurations for each class. Moreover, for each configuration, we randomly constructed an initial training set of one positive and one negative instance (experiments in the 10-shot setting were less-effective, as described in Section 4.5.2). We evaluated the two-shot accuracy of a logistic regression model on a held-out validation set and then performed one of the following two updates with both a randomly-drawn positive instance and a randomly-drawn negative instance simultaneously to preserve class balance:

1. Baseline: We update the model by adding the positive and negative instances to the training data and retraining.

2. LIMEADE: First, we generate LIME explanations of the opaque classifier for both the positive and negative instance. In the positive case, we simulate a human's advice in response to the explanation by utilizing the COCO segmentation masks to automatedly give the superpixel(s) indicative of the class a positive label (i.e., in the case of "giraffe," we select all superpixels containing giraffes using the COCO segmentation masks in the image labeled as "giraffe"). In the negative case, we give the superpixel most influencing the LIME explanation a negative label. We then generate embeddings of these labeled regions and use the embeddings to retrieve the nearest superpixels and full images across the unused pool (consisting of $19,996$ images, along with their individual superpixels). We append the embeddings of the nearest neighbors' corresponding full images to the training data along with $+$ and $-$ labels, respectively, and retrain. This simulated approach to human advice enabled us to study the effectiveness of LIMEADE updates by testing many initial configurations across a range of image classes.

We wanted to evaluate LIMEADE across different hyperparameter settings, so we varied

[3] The 20 classes are: baseball glove, snowboard, giraffe, carrot, surfboard, fork, sink, cow, donut, toothbrush, knife, bed, horse, cake, motorcycle, frisbee, skateboard, fire hydrant, scissors, and suitcase.

the number of nearest neighbors included in the update ($n_{neighbors} = \{1, 5, 10, 25, 50, 100\}$), as well as the relative sample weight of the update ($w = \{0.25, 0.5, 1, 2, 4\}$), and performed a grid search. We evaluated performance on a balanced, held-out validation set of 400 positive instances and 400 negative instances for each class and selected the hyperparameters with the highest validation accuracy. This process yielded a relative sample weight of 0.25, as well as 50 nearest neighbors included in the update. With these hyperparameters selected, we then evaluated final performance on a separate, held-out test set of 400 positive instances and 400 negative instances for each class.

4.3.2 LIMEADE Feedback is More Effective than the Baseline

In Table 4.1, we report the net changes in classifier accuracy when making updates with LIMEADE and the baseline across all 20 classes and 100 runs per class, as evaluated on the test set. We find that LIMEADE updates with simulated advice outperform the baseline for 16 of 20 classes, giving an average boost of 9.33% compared to the baseline's average boost of 8.21%. These results are statistically significant: a paired t-test of LIMEADE against the baseline yields a p-value of 2.3×10^{-9} across all 2,000 runs.

4.3.3 Diminishing Returns as Supervision Increases

While conducting our case study with the image domain, our empirical results indicated that the LIMEADE-powered improvements decrease as a function of more supervision. For example, repeating the experiments in the 10-shot setting, we find that LIMEADE gives an overall boost of 0.63%, whereas the baseline gives an overall boost of 0.88%. It is important to note, however, that there is a fundamental entanglement between training data and supervision with respect to LIMEADE. LIMEADE is most valuable when the original supervisory data is subject to spurious correlations (i.e., when teaching "cat," if all cats seen in the training data happen to be black, a LIMEADE update communicating that color does not matter has high utility). If the training data is representative (because the training data contains more instances or because the instances themselves are better-selected), we expect a LIMEADE update to provide less utility, as there are fewer potentially

Class	2-Shot Accuracy	Δ Baseline	Δ LIMEADE	p-value	Winner
Baseball Glove	76.73%	10.64% ± 1.25%	10.75% ± 1.52%	0.91	LIMEADE
Snowboard	75.97%	10.67% ± 1.08%	10.74% ± 1.42%	0.93	LIMEADE
Giraffe	74.54%	12.30% ± 1.28%	16.42% ± 1.51%	9.3×10^{-8}	**LIMEADE***
Carrot	72.82%	7.37% ± 1.15%	9.76% ± 1.39%	4.9×10^{-4}	**LIMEADE***
Surfboard	72.59%	8.23% ± 1.19%	8.96% ± 1.36%	0.34	LIMEADE
Fork	72.14%	7.70% ± 1.24%	6.96% ± 1.64%	0.53	Baseline
Sink	71.55%	10.38% ± 1.12%	10.44% ± 1.49%	0.95	LIMEADE
Cow	69.90%	8.66% ± 1.00%	11.53% ± 1.07%	5.5×10^{-6}	**LIMEADE***
Donut	67.51%	8.23% ± 0.96%	9.65% ± 1.11%	0.093	LIMEADE
Toothbrush	65.85%	5.26% ± 0.82%	4.86% ± 1.02%	0.63	Baseline
Knife	65.47%	7.31%± 1.10%	7.43% ± 1.43%	0.86	LIMEADE
Bed	65.16%	9.50% ± 0.93%	11.73% ± 1.16%	5.7×10^{-3}	**LIMEADE***
Horse	63.66%	8.49% ± 0.90%	10.20% ± 1.31%	0.050	**LIMEADE***
Cake	63.48%	8.53% ± 1.02%	9.07% ± 1.33%	0.49	LIMEADE
Motorcycle	63.16%	9.37% ± 0.98%	15.97% ± 1.08%	3.7×10^{-11}	**LIMEADE***
Frisbee	62.67%	7.80% ± 0.85%	7.03% ± 1.18%	0.32	Baseline
Skateboard	61.09%	6.93% ± 0.82%	7.52% ± 1.03%	0.48	LIMEADE
Fire Hydrant	59.21%	6.31% ± 0.75%	8.38% ± 0.92%	8.7×10^{-3}	**LIMEADE***
Scissors	57.38%	6.51% ± 0.79%	4.92% ± 1.05%	0.037	**Baseline***
Suitcase	55.65%	4.04% ± 0.63%	4.23% ± 0.67%	0.71	LIMEADE
Total	66.83%	8.21%	9.33%	2.3×10^{-9}	**LIMEADE***

Table 4.1: Updates using LIMEADE boost the accuracy of an opaque image classifier more than the baseline. Results are shown for 20 classes averaged over 100 randomly-initialized runs each, and the accuracy boosts are reported relative to an average initial, 2-shot accuracy on a test set. For the updates, standard errors are reported, and a $*$ indicates p-value ≤ 0.05. LIMEADE outperforms the baseline on 16 of 20 classes and provides an overall boost of 9.33%, as opposed to the baseline's overall boost of 8.21%.

spurious correlations for a human to correct via a LIMEADE update. Indeed, as the quality of the originally-trained classifier approaches perfection, the value of LIMEADE goes to 0, much as the value of more training data also decreases. Our empirical evidence thus agrees with our intuition that a LIMEADE update is most valuable in the low supervision setting.

4.4 Case Study 2: LIMEADE for Paper Recommendation

For our second domain, we selected text ranking both for variety and importance. The overwhelming influx of new scientific publications poses a daily challenge for researchers [38, 91, 135, 164, 346]. However, based on Beel *et al.* [34]'s survey of 185 publications on academic paper recommendation, only a few systems explain why papers have been recommended or respond to user feedback other than liking/disliking specific papers, and all such systems rely on interpretable recommenders [27, 47, 165, 285, 383]. The ability to explain and take advice for higher-performance paper recommenders, therefore, fills an important void.

Furthermore, a complete evaluation of a human-AI interaction approach requires testing it with real users in the loop [11]. For LIMEADE, we wanted human users who were authentically motivated to understand and improve an ML classifier. In this regard, we built Semantic Sanity, a computer science (CS) research-paper recommender system based on Andrej Karpathy's arXiv Sanity Preserver [166]. Deployed as a publicly-available platform, Semantic Sanity enables users to curate feeds from over 150,000 CS papers recently published on arXiv.org. With this testbed, users are implicitly incentivized to understand and improve the recommender system powering their feed in order to receive more interesting papers. Note further that each user is a task expert, since the users determine their own preferences.

Lastly, this study complements the first case study presented in the previous section. We intentionally selected two case studies with very different settings: while our first case study considered image classification, this study surrounding text ranking considers a different domain and task. Thus, studying and evaluating our implementation of LIMEADE with Semantic Sanity provides evidence for the generality of our framework.

4.4.1 Neural Recommender

To generate individual recommendations, we utilize a neural model consisting of a linear SVM on top of neural paper embeddings pre-trained on a similar papers task [63]. Each paper is represented by the first vector (i.e., the [CLS] token typically chosen for text classification) after encoding the paper title and abstract using SciBERT [36]. The neural embedding model is finetuned on a triplet loss $\mathcal{L} = max(0, v_i^T v_+ - v_i^T v_- + m)$ where m is a margin hyperparameter and v_i, v_+ and v_- are the vectors representing a query paper, a similar paper to the query paper, and a dissimilar paper to the query paper, respectively. The similar paper triples are heuristically defined using citations from the Semantic Scholar corpus [13], treating cited papers as more similar than uncited papers. Recommendations are generated by training the model on a user's annotation history, with additional negative instances randomly drawn from the full corpus of unannotated papers.

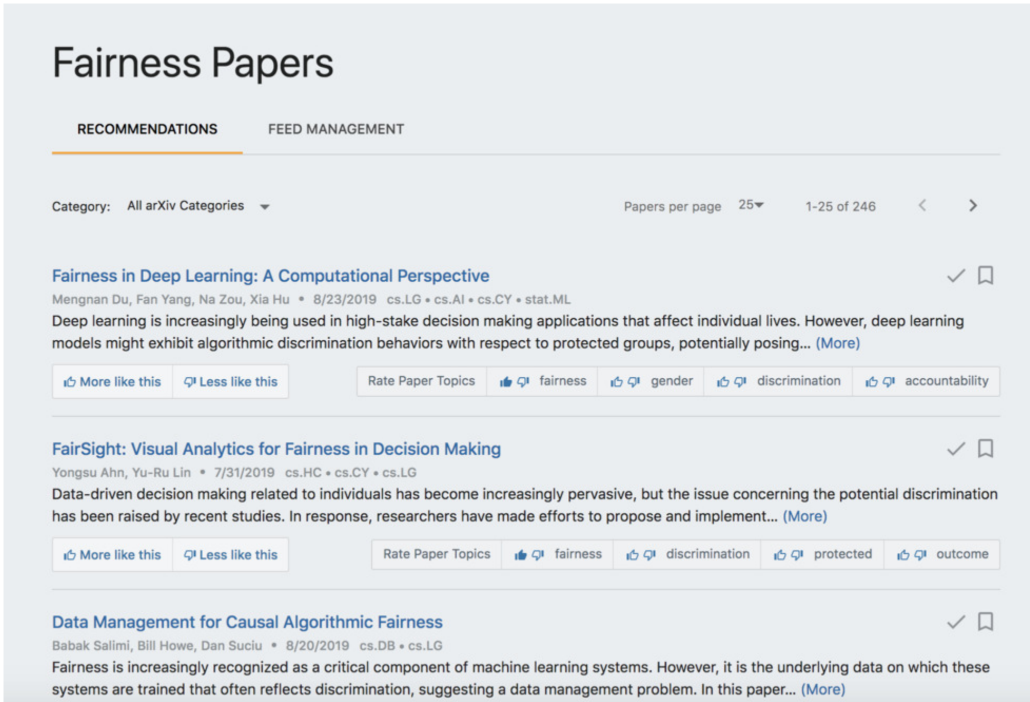

Figure 4.4: The UI for a feed in Semantic Sanity. Users can rate the papers themselves with the "More like this" and "Less like this" buttons, a standard feed affordance. Under each paper, the system also presents four terms to explain why it was recommended and solicits feedback with "Rate Paper Topics" — by clicking thumbs up or down, the user can give advice by requesting that the feed include more or less of the specified topic.

A user begins the process of curating their feed by either selecting a specific arXiv CS category or issuing a keyword search and then rating a handful of the resulting papers. A feed consists of a list of recommended papers sorted by predicted recommendation score (see Figure 4.4). Each paper can be rated using traditional "More like this" or "Less like this" buttons underneath each paper description.

4.4.2 Implementation of Explanations and Feedback

The UI for Semantic Sanity (Figure 4.4) displays a list of recommended papers and adorns each with an explanation comprising four terms; to the left of each term are thumbs-up and thumbs-down buttons, enabling the user to not only rate the papers themselves but also *give advice* in response to the explanation and indicate if they would like to see more or fewer papers related to that term. We refined our user interface design through iterative informal user testing. The explanatory terms are generated using a simple, explanatory model (LIMEADE's Explain function), which we implement as a linear model over uni- and bigram features. In particular, our linear model is defined as $g(x') = w_0 + \sum_i w_i x'_i$, and the explanation surfaced for $g(x')$ consists of high-impact terms in the model, i.e., those with high values for the product $w_i x'_i$. Specifically, we select the 20,000 features with the highest term frequency across our corpus. Our approach of using a post-hoc explanatory model is similar to that used by LIME, except to enable real-time performance our explanatory model is trained as a global, rather than local, approximation of the neural model [320]. This global approximation was chosen because testing on an early prototype revealed that generating explanations for a feed using LIME was too computationally expensive, since LIME requires sampling nearby instances and training a model for *each* recommendation on the page; this latency negatively impacted the recommendation experience.[4]

Given the explanatory model, LIMEADE's Display function chooses explanations to display by computing each term's contribution to the output of the linear model for the given paper, which is equal to the product of the term's TF-IDF value for the paper with the term's feature weight in the linear model. We note here that even though the explanatory model is a global approximation, the explanations are local ones, as this product encodes instance-specific information on why the paper has been recommended.[5] Next to each explanatory term are thumbs up and down buttons (see Figure 4.4). When the user provides feedback with these buttons, LIMEADE generates pseudo-instances and retrains the neural recommender. We use a generative approach within GetInstance that leverages the unlabeled pool of papers. We select the top 100 papers from the full corpus with the highest TF-IDF value for the feedback term and generate a single synthetic pseudo-instance (i.e., we use $k = 1$) equal to the centroid of these papers' embeddings with a weight of 1. The instance is appended to the user's history and labeled with the user's annotation of the term (+/-).

[4] Note that in a LIME-style approach to instance-level explanation, one needs to either sample neighbors around the instance (which may be quite distant) or generate synthetic instances, which in our case requires generating paper embeddings on the fly. Both approaches introduced unacceptable latency.

[5] In a later implementation of Semantic Sanity, the platform also included a global explanation surfaced at the top of each feed, which is the subject of another study [310].

4.4.3 Online Traffic

In the next section, we describe a controlled user study comparing Semantic Sanity with and without LIMEADE. However, since its public launch, Semantic Sanity has also attracted considerable organic traffic: users with accounts have constructed 2,478 feeds and have logged 21,713 paper annotations and 1,320 topic annotations (we note that annotating topics was only possible after the LIMEADE-based implementation was introduced on November 11, 2019, five months after the initial launch of Semantic Sanity). The target user base was computer science researchers, and the platform was advertised through social media and email lists. In Sections 4.4.7 and 4.5.3, we analyze a subset of the organic user logs as a complementary part of our evaluation.

4.4.4 User Study: Experimental Setup

In order to evaluate the effectiveness of our LIMEADE-based system for recommending papers with real users, we performed an in-person user study. With this user study, we wanted to address the following research questions:

1. Do participants prefer LIMEADE over a baseline of just explanations according to self-reported ratings of trust, control, transparency, intuition, paper coverage, and the overall system?
2. Does LIMEADE increase participants' feed quality, evaluated quantitatively with blind ratings of recommended papers?
3. How do participants utilize the topic-rating affordances powered by LIMEADE?
4. What constructive feedback do participants have surrounding our particular instantiation of LIMEADE with Semantic Sanity?

We recruited 21 participants through a public university's computer science email lists. All participants were adults who reported experience with reading computer science research papers in our screening questionnaire. Each session lasted one hour, and each participant was compensated with a $25 Amazon gift card. Our IRB application did not include a plan to collect and share participants' demographic data, and therefore, we could not include it in the study results.

Participants were asked to curate feeds of computer science papers pertaining to a topic of their choice using two different recommendation user interfaces (UIs), one that used LIMEADE to provide advice-taking explanations, and one that did not present explanations, instead only allowing users to rate the papers themselves (the baseline); other than this difference, the UIs were the same. The participants were asked to choose a topic that they were interested in following over time as new papers are added to the arXiv, but not so general that it is already covered by an existing arXiv CS category (e.g., artificial intelligence). Once a topic was selected, each participant was asked to name the desired feed, which served as the goal for curation using both UIs.

Each participant began curation by selecting exactly three seed papers that were then used to initialize the feeds in both UIs. Both systems surfaced the same initial recommendations in response to the participant's three seed papers and thus had identical initial states. Each participant was then presented with one of the two UIs and given instructions on how to use it. 11 participants received the baseline system first, and 10 received the LIMEADE system first. They were then presented with the second UI. For both UIs, the participants were told to use as many or as few annotations as desired until their feed was curated to their liking, or a maximum of 10 minutes was reached. We recorded the participants' annotations for both feeds. After using each UI, the participants were asked to complete a short survey. They were then asked to rate a blind list of combined recommendations from the two feeds that they had curated, according to whether they would like to see each paper in their desired feed. These recommendations were generated on a held-out paper corpus, disjoint from the papers available within the feed UIs.

Data were successfully collected for all 21 participants. The participants' chosen topics varied greatly, including "Spiking Neural Networks," "Moderation of Online Communities," and "Dialogue System Evaluation."

Which system...	Baseline	LIMEADE	p-value
...trust more?	4	**17***	0.043
...more control?	0	**21***	≈ 0
...more transparent?	3	**18***	0.012
...more intuitive?	**12**	9	0.664
...not missing relevant papers?	3	**18***	0.012

Table 4.2: Among 21 participants, most prefer our system over the baseline when prompted with these questions. *(*) indicates a statistically significant result under a two-sided binomial test against a null hypothesis of no preference between the systems.*

Likert scale rating	Baseline	LIMEADE	p-value
Overall system	3.38 ± 0.59	$\mathbf{3.85 \pm 0.57}$*	0.043
Would use again?	3.38 ± 1.16	$\mathbf{3.90 \pm 0.94}$	0.257

Table 4.3: Mean $\pm$ Standard Deviation of 21 participant ratings of each system. Ratings were on a scale from 1 (worst/no) to 5 (best/yes). *(*) indicates a statistically significant result under a two-sided paired t-test against a null hypothesis of zero mean difference between the systems.*

4.4.5 User Study: Quantitative Results

User Experience: Participants Prefer LIMEADE

In the surveys administered after using each UI, we asked each participant to provide overall ratings for each system and to state which system they preferred along dimensions such as trust and intuitiveness. The results are summarized in Tables 4.2 and 4.3.[6]

Overall, participants rated our approach significantly higher than the baseline. They also rated it significantly higher on trust, control, transparency, and on confidence that their recommendations were not missing relevant papers. Understandably, our LIMEADE system appeared less intuitive to participants than the baseline due to the increased complexity of the UI, though this result was not statistically significant. Finally, while not statistically significant due to small sample size, participants indicated more likelihood to use our system again over the baseline. In aggregate, these results indicate a higher-quality user experience with the LIMEADE system than with the baseline system.

[6]For all statistical significance tests, we report adjusted p-values using the Holm-Bonferroni procedure for multiple comparisons [147] in R's p.adjust library [309].

Mixed Results for Feed Curation Time

In analyzing the times required by each participant to complete feed curation using the two systems, we observe that eight participants finished feed curation with the baseline system first, seven finished with the LIMEADE system first, and the remaining six utilized all ten minutes for the curation of each system.

Most Participants used Both Paper and Topic Ratings

To explore the breakdown of participants' rating habits with the baseline system and the LIMEADE system, we present Figure 4.5. In the left plot in Figure 4.5, we observe that participants displayed a high degree of variance in the number of ratings applied during feed curation, ranging from 7 annotations to 61 annotations with the baseline system. Comparing the total number of annotations made using the system with LIMEADE vs. the number of annotations made with the baseline, we find a best-fit slope of 0.913. This suggests that the participants made approximately the same number of annotations across both systems.

In the right plot, we observe that there is significant diversity in how participants applied topic annotations, ranging from 2 annotations to 27. However, most participants utilized a combination of paper and topic ratings, with more paper ratings than topic ratings on average. Interestingly, five out of the twenty-one participants provided more negative paper ratings than positive ones in the baseline; when presented with the LIMEADE affordances, no participants provided more negative paper ratings than positive ones, but four participants applied more negative topic ratings than positive ones.

Blind Ratings of Recommendations: No Significant Difference in Feed Quality

We also investigated whether the topic-level feedback provided by LIMEADE measurably increased the quality of participants' feed. We showed participants the top 20 recommendations generated by both systems on the held-out corpus of papers and measured their ratings. Specifically, we computed the discounted cumulative gain (DCG)[7] and average

[7]We did not use NDCG because the participants liked different numbers of papers between the two systems.

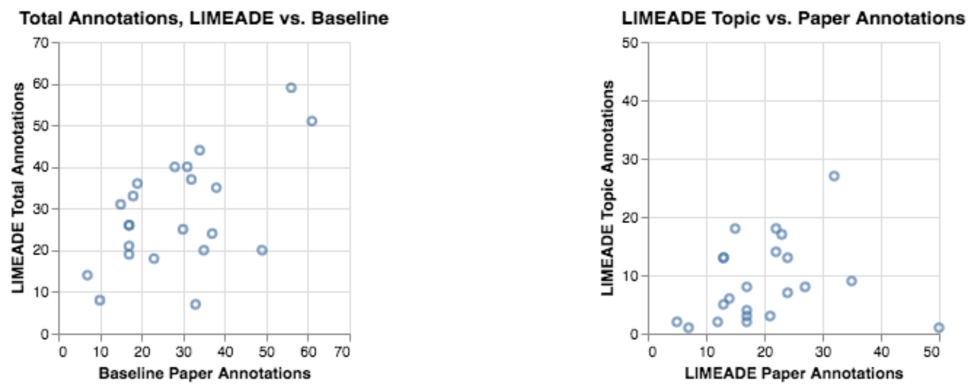

Figure 4.5: Scatter plots showing (left) the total number of annotations used to curate a feed with LIMEADE (paper annotations + topic annotations) vs. the number of baseline paper annotations per user, and (right) the number of topic annotations vs. the number of paper annotations in the LIMEADE system. Most participants used LIMEADE-powered topic-level feedback as well as paper-level feedback.

precision (AP), common metrics for assessing recommendation feed quality. For DCG, we observe a mean difference of 0.259 in favor of the baseline system recommendations; however, the corrected p-value for the two-sided, paired t-test for mean differences is 0.218, indicating no significant difference in feed quality between the two systems under DCG. For AP, we observe a mean difference of 0.0412 in favor of the baseline system, with a corrected p-value of 0.257, also indicating lack of significant difference in quality under AP. Based on the constructive feedback that we received, we speculate that this result could be improved by making implementation-specific adjustments to Semantic Sanity.

4.4.6 User Study: Qualitative Feedback

We analyze participants' text responses and provide a sample of quotes that complement the quantitative results. After using each system, participants were asked to provide free text responses to the question, "Would you like to share anything else about using the system?" At the end of the study, participants were also asked, "Do you have any last thoughts that you would like to share regarding actionable explanations?" Overall, participants found the advice-taking affordance granted by our system helpful: *"The explanations here were especially useful in their capacity as decisions rather than just explanations. I would have found them really really annoying if they were presented only as an explanation of why you* *thought I would like a paper, rather than an attribute I could ask for more or less of."* In particular, participants stressed the importance of the LIMEADE affordance as a filtering mechanism: *"The topics feature was excellent, because there are many papers which cover *some* topics I like but also some that I don't, and this let me pick that out."*

The constructive feedback received in the qualitative responses illustrate a number of implementation-specific improvements that could be made to Semantic Sanity. The most common category of constructive feedback concerned the quality of terms in the explanations, mentioned by 10 participants. Though we utilized stemming to eliminate these redundancies in each paper explanation, we did not eliminate synonyms from the list of terms. For example, three of the ten participants specifically requested that abbreviations in explanations be removed or linked to full terms. These issues reflect the negative consequences of utilizing 20,000 TF-IDF terms for our explanatory model featurization. In addition, five of these users also stated that the terms were too general. We speculate that the term quality in the explanations negatively impacted the users' ability to give advice to the model via the LIMEADE affordance.

Similarly, three participants directly addressed what we term the explanation-action tradeoff in the next section, noting that the lack of diversity of terms in the explanations was limiting. One participant commented: *"After a few minutes, almost all the same terms that I had liked were coming up, so there were few new terms for me to thumbs up or down. I think if the system could focus on bringing up relevant papers that have a new term or two to which I can react, that would make the curation even better."* This suggests tuning the system to favor explanation diversity even more than we did in our initial implementation.

Interestingly, two users believed that the set of topics surfaced was too restrictive, one thought that the terms were too diverse, and one thought the diversity was a good feature. This provides some evidence that different users have different preferences for explanation diversity, suggesting that it should perhaps be tuned in a user-specific manner. Additionally, four participants commented on topic annotation strength, all of whom indicated that it was too potent, revealing the importance of empirically evaluating the optimal strength of an update. Based on this feedback, we reduced the annotation strength in our application following the evaluation.

4.4.7 Feed Quality Revisited Using Log Data: Term Annotations in LIMEADE can Improve Performance

We also investigated the effect of high-level advice on a different set of users — those who used Semantic Sanity in the wild, rather than as part of a laboratory user study — using the log data of the online deployment. Specifically, we compiled a data set of 1,636 rated papers across 30 feeds, where each feed had at least one annotated explanation (the average number of annotations for these feeds was 4.4 terms). We evaluate two recommenders: a baseline ranker that uses only the rated papers, and a LIMEADE ranker that uses both the rated papers and the annotated terms processed by LIMEADE. We evaluate at three different training sizes (2, 5, and 10 labeled papers), and to maximize the contrast between LIMEADE and the baseline, we always provide LIMEADE with *all* of the explanation annotations for the feed (4.4 terms per feed on average). Thus, this experiment measures whether LIMEADE's pseudo-instance approach can be effective given sufficiently informative term annotations, but is not an accurate simulation of the system in practice (in which term annotations would arise only from explanations on papers in the limited training set). For each feed and size we compute the average normalized discounted cumulative gain (NDCG) ranking performance for up to ten sampled training sets, testing on the remainder. The average of the NDCG statistics across feeds is our final evaluation measure.

Number of labeled papers	Base Ranking Performance	Δ LIMEADE
2	0.884	**0.015**
5	0.901	**0.009**
10	0.908	**0.005**

Table 4.4: Simulated evaluation of ranking performance (NDCG) based on log data from actual usage in case study 2. LIMEADE improves performance over the baseline system, which does not use the annotated explanations.

Table 4.4 shows that LIMEADE does improve performance over the baseline, but the benefits of the annotated explanations diminish as the number of initially rated papers increases. The individual differences shown in the table are not statistically significant, but the aggregate performance over all three sizes shows LIMEADE performing significantly better than the baseline (p-value 0.017, two-tailed paired t-test, after Holm-Bonferoni correction). LIMEADE with 2 and 5 annotated papers performs comparably to the baseline with 5 and 10 annotated papers, respectively, meaning that LIMEADE reduces the number of paper labels required to achieve a given level of performance by an amount roughly equal to its number of term annotations in this experiment. The experiment is inconclusive regarding whether giving advice via term annotations would be *preferable* to obtaining a similar number of labeled instances in this domain. Experiments with more users and feeds are necessary to resolve these questions.

4.5 Discussion

Evaluating on real-world domains with real human interactions is crucial in order to make progress in human-centered AI broadly, and for advice taking in particular. This section considers broader questions of the connections between our two case studies, the effectiveness of human feedback, and the interactions between the fidelity of explanations and the affordances provided for action.

4.5.1 Demonstrating the Generality of LIMEADE

In order to demonstrate that LIMEADE is a universal mechanism for applying high-level advice to an arbitrary ML model, we chose our case studies to span a diverse range of dimensions. Table 4.5 summarizes the differences, which include the source domain (image vs. text), type of model (classification vs. ranking), nature of the explanatory vocabulary, and method for generating pseudo-instances. There are many options for creating pseudo-instances, and future work will be necessary to uncover the best methods. For example, is it better to generate one instance (as we did in the text domain) or several (as we did in the image domain)? Is it better to label naturally-occurring (unlabeled) instances as we did in the image domain, or to create a synthetic pseudo-instance as we did by computing the centroid of matching examples in the text domain?

Usage differed across the case studies as well. In the image domain, we evaluated the effect of a single piece of high-level advice on the accuracy of the classifier. In the text

	Case Study 1	Case Study 2
Domain	Image	Text
Task	Object detection	Paper recommendation
ML Setting	Classification	Ranking
Explanatory Features	Superpixels (≈10−25), instance-specific	n-gram topics (3 shown of 20,000 possible), same vocabulary across instances
Feedback Mechanism	↑ or ↓ vote any superpixel	↑ or ↓ vote any of 3 displayed n-grams
LIMEADE GetInstance Method (for selecting pseudo-instance)	Retrieve nearest unlabeled images & superpixels	Generate a synthetic instance equal to the centroid of matching unlabeled instances
Evaluation	Simulated study with clear ground truth	In-person user study & simulated study using log data from online usage
Results	Advice taking with LIMEADE improves accuracy more quickly than adding labeled instances, though this effect diminishes with more initial training data.	Advice taking with LIMEADE leads to increased control, trust, satisfaction, & system transparency. Inconclusive w.r.t. improving accuracy more easily than adding labeled instances.

Table 4.5: A comparison of our two case studies.

domain, however, users interacted repeatedly to improve the ranker by providing a sequence of high-level advice and labeled examples in the way that seemed most natural to them.

4.5.2 When Does High-level Advice Improve Learner Accuracy?

When tested on numerous domains, we obtained positive to indeterminate results about the effectiveness of LIMEADE processing high-level human advice. Does this reflect a weakness in the LIMEADE approach or limitations of our LIME explanations? Or is it intrinsic — perhaps human-interpretable vocabulary is simply too dissimilar to the features learned by modern neural methods for *any* human advice to be useful. Maybe getting more data is the only or the most effective way for humans to help out?

One thing seems clear — in order to answer this question, the research community must conduct more experiments on real world domains, rather than toy domains with artificial confounds, such as Decoy MNIST.

While they only simulate interactions, our image domain experiments (Section 4.3) reflect actual human judgments about which regions contain the object in question. LIMEADE-processed advice about which regions contained an object significantly improved classifier accuracy in the two shot case. However, when we conducted similar experiments after training the twenty classifiers with ten instances, we found no significant improvement. Perhaps this is because the model had already learned where the objects were located. More likely, it had found the context imparted from background information to be useful in the classification decision. It also may stem from the segmentation algorithm that induced the 'advice vocabulary,' or perhaps the LIMEADE method weighted instances incorrectly.

While users clearly liked the ability to provide high-level advice and felt it increased their sense of trust and control, we found mixed results with respect to improving ranking accuracy as measured with DCG. Our controlled study over 21 users (Section 4.4.5) found no significant difference between feed accuracy incorporating LIMEADE advice *vs.* feeds created with simple labeled instances. In contrast, we did find significant improvements stemming from LIMEADE advice in our simulated log study on 30 different users (Section 4.4.7). The differences could stem from our LIMEADE mechanism, the bi-gram vocabulary chosen as features in the explanatory model, the size of our study, or some other reason.

We strongly believe that much more research should be devoted to this important question. LIMEADE is an important first step, but our chapter should be considered a "Call to Action" for more investigation. To this end, we will release the code for LIMEADE and our image experiments, including our modified COCO dataset with the precomputed superpixel vocabulary and corresponding embeddings.

It is also important to contextualize our findings within prior work on advice taking. While studies such as [184] demonstrated a clear improvement in classifier accuracy in the setting of explanatory debugging, other studies have found the opposite. Of particular interest are the results from [2] and [386], which demonstrated that tunability for search and recommendation tasks can negatively impact feed quality when it takes the form of adding or removing terms from the featurization. Likewise, [78] shows that advice taking with interpretable models can lower accuracy. Lastly, [403] is another datapoint that indicates that letting people into the interactive machine learning loop can be problematic. These concerns are especially problematic, given users' clear expectations that feedback will lead to ML improvement [347]. For this reason, we reiterate that our chapter is a "Call to

Action" for more research surrounding high-level advice and learner accuracy.

4.5.3 Exposing the Explanation-Action Tradeoff

Semantic Sanity chooses explanations to display by computing each term's contribution to the output of the linear model for the given paper, which is equal to the product of the term's TF-IDF value for the paper with the term's feature weight in the linear model. The natural choice is to surface the terms with the highest-magnitude contributions in the linear model [320]; we call this a *greedy* approach. Users could then react to these presented terms, thumbing them up or down.

Comments from early users of our paper recommender indicated that there is a tradeoff between using the greedy explanation approach and the explanatory terms' uses as affordances for feedback, which we call the *explanation-action tradeoff.* In particular, user action on an explanatory feature will lead the model to place increasing importance on it and correlated features. With the greedy approach, these terms will begin to dominate both the model and the explanations, limiting the number of unique explanation terms and thus subsequent opportunities for feedback. For example, 'thumbs-up'ing the term "fairness" causes papers about fairness to rise in the feed; under the greedy approach, these papers will contain the term "fairness" in their explanations, thereby crowding out new terms for the user to act on.

Based on the feedback we received, our final implementation of Display in Semantic Sanity uses a diversity-biased approach that samples explanatory features in a way that prevents previously suggested terms from dominating subsequent explanations.[8] However, as noted in Section 4.4.6, three participants commented in their qualitative feedback that they would have still liked even more diversity — further evidence that properly considering and calibrating the explanation-action tradeoff is important for advice taking.

To illustrate the impact of the explanation-action tradeoff and the distinction between

[8]Specifically, Display uses a parameter γ and samples terms proportionally to the magnitude of term contribution, raised to the γ power (higher values of γ result in a more greedy approach; lower values increase diversity). We selected $\gamma = 4$ for our implementation. To further reduce term redundancy in each recommended paper's explanation, we used the Python `NLTK` PorterStemmer [218] to deduplicate terms with the same stems (e.g., "fair" and "fairness") from each explanation.

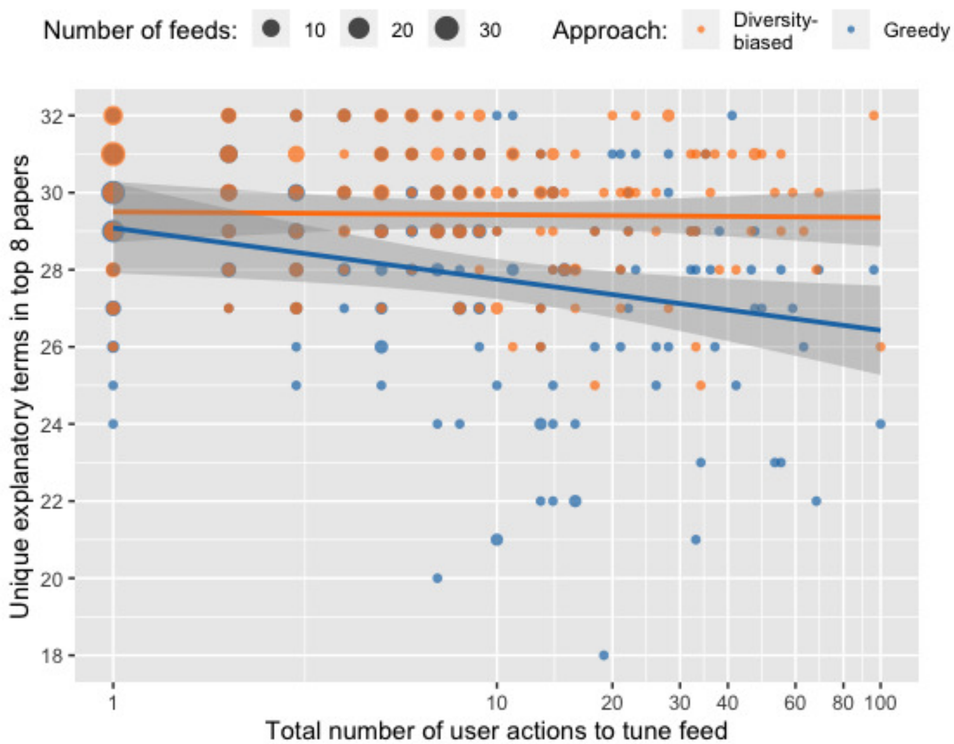

Figure 4.6: A scatter plot showing the number of unique explanation terms in the first page of the feed vs. the number of actions taken by the user in order to give advice to their their feed. Orange dots correspond to diversity-biased explanations currently used in the system. Blue dots correspond to greedy explanations, where the most important terms are surfaced without stochasticity. The size of each dot corresponds to the number of feeds in that bin. Note that greedy explanations (blue) display a stronger negative correlation between unique terms and term annotations than diversity-biased explanations (orange). Thus, the greedy approach limits opportunities for advice taking with topics as the feed curation process evolves, while the diversity-biased approach continues to facilitate advice taking with topics.

our diversity-biased approach and the canonical greedy method, we perform an analysis on the logs of 300 users' feeds from Semantic Sanity's online deployment. For each user, we compute *(i)* the total number of actions the user has taken on displayed explanatory terms, and *(ii)* the number of unique explanation terms among the latest top eight recommended papers under our diversity-biased DISPLAY implementation. We then repeat *(ii)* but with DISPLAY with $\gamma = \infty$ to simulate what explanatory terms the users would see today under a greedy approach.

In accordance with the explanation-action tradeoff, we observe in Figure 4.6 that the number of unique explanation terms (i.e. advice-taking affordances) tends to be lower under a greedy approach. Furthermore, this effect grows stronger as users give advice to their feeds to be increasingly specific to a particular topic.[9] In contrast, the number of affordances remains relatively constant under our diversity-biased approach. Though some explanation terms with lower contribution weight are included within the explanatory model, our diversity-biased approach thus successfully mitigates the crowding effect observed with the greedy approach.

The explanation-action tradeoff is related to, but distinct from, the classical *explore-exploit tradeoff* faced by recommender systems and other machine learners [363]. The explore-exploit tradeoff entails deliberately passing up a known reward in the hopes of learning more about the reward structure in order to have better long-term gains. Thus, the explore-exploit tradeoff encourages taking a chance in executing an action in the hopes that it will provide a big reward, leading to frequent execution of the action in the future. The explanation-action tradeoff is similar to the explore-exploit tradeoff, in the sense that it entails deliberately declining to provide the most accurate explanation in the hope that providing an affordance for the user to execute a feedback action will lead to better long term recommendations. However, with the explanation-action tradeoff, even if the system is fortunate when taking a chance by providing a less faithful explanation that successfully solicits user feedback, the system will *never* want to repeat the specific explanation-action

[9] Figure 4.6 likely understates the impact of the tradeoff, as users had been exposed to explanations under the diversity-biased approach prior to this analysis. Had they been exposed to explanations under the greedy approach for their entire sessions, we likely would observe an even stronger crowding-out effect.

in the future. We therefore highlight the explanation-action tradeoff as an important consideration when implementing an advice-taking system.

While the explanation-action tradeoff was observed in our second case study, it is important to clarify why we did not observe a similar tradeoff in our first case study in the image domain. We believe this to be a consequence of fundamental differences between our case studies, as detailed in Table 4.5. First, the explanatory vocabulary is not fully *discoverable* with papers (we only surface 3 terms out of thousands), but it is fully discoverable with images (all constituent superpixels can easily be viewed simultaneously). Second, advice taking in Semantic Sanity was *iterative* because users repeatedly refined their feeds, whereas only one round of advice taking was performed in our image domain experiment, providing no chance for a feedback cycle to develop. Lastly, the explanatory vocabulary is *shared* across papers, but not across images: when providing advice multiple times in the image domain, a different instance will be surfaced each time, meaning that a new set of superpixel features is available for eliciting advice. Given these differences, we would not expect to observe the explanation-action tradeoff in the image domain. However, it is evident that the explanation-action tradeoff may arise in enough advice-taking settings that documenting and investigating it is an important contribution.

4.5.4 Decoupling the Effect of Explanations & Advice Taking

Previous studies have shown that users prefer recommendations with explanations over recommendations alone [370, 412]. In our user study, we did not include an "explanations only" baseline, which would have helped to isolate the contribution of explanations in the preference for our LIMEADE system among participants. However, we did analyze the user study results post-hoc to investigate this question. In particular, we studied the results in Tables 4.2 and 4.3 in order to assess whether participants' self-reported preferences for our LIMEADE system over the baseline system correlated with utilization of the LIMEADE affordance for rating topics. The participants who voted LIMEADE higher on trust, transparency, intuitiveness, and confidence in not missing papers performed 5.4, 4.6, -0.5, and 3.8 more topic annotations, on average, than those who voted the baseline higher, respec-

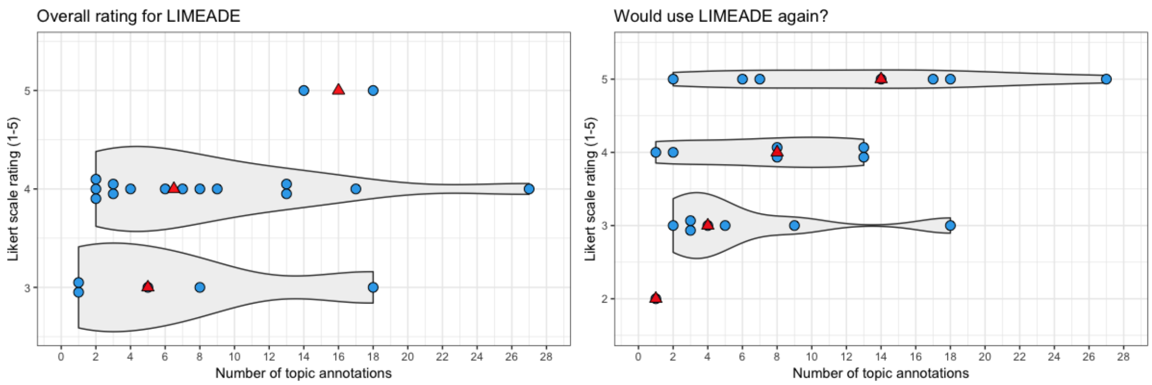

Figure 4.7: Plots showing participants' Likert scale evaluations of our overall LIMEADE system (left) and the likelihood that they would use our system in the future (right) as functions of the number of topic annotations made when using our LIMEADE system. The red triangles show the median number of annotations for each rating level.

tively. This suggests that the positive outcomes for those metrics were *not* a result of the explanations alone, but were influenced by the advice-taking affordance of LIMEADE.

In Figure 4.7, we investigate how the number of topic ratings used by each participant varies as a function of their Likert scale ratings in Table 4.3. We find that a higher overall rating of our LIMEADE system and a higher self-reported likelihood of using our LIMEADE system in the future are correlated with using more topic annotations (i.e., giving more advice). This indicates that more usage of the LIMEADE affordance correlates with a more positive perception of the LIMEADE system.

4.6 Related Work

Space precludes a discussion of work on explanation generation; we focus our description of prior work on approaches for incorporating human advice in machine learning models and on approaches for creating pseudo-instances by labeling features. Some work transcends these distinctions, however; Smith-Renner *et al.* [347] show that many users expect that an ML model will improve over time and that users are frustrated with imperfect AI systems that provide explanations without supporting the ability to receive corresponding feedback. Furthermore, there are other general-purpose ways to improve model accuracy besides high-level advice: labeling new training instances, altering the weights of training instances, and providing an 'undo' button to remove a label that has just been added are a few such methods.

4.6.1 Enabling Machine Learners to Take Human Advice: Interpretable Models

Research from interactive machine learning and human-AI interaction has shown the benefits of enabling learning models, including both recommender systems and classifiers, to take advice from humans [10, 342]. For example, Lou *et al.* [224], Lou *et al.* [225], Caruana *et al.* [51], and Wang *et al.* [389] have demonstrated the value of GAMs and GA^{2}Ms, which can be directly modified by humans via the alteration of shape functions. Likewise, Kulesza *et al.* [184] have shown the power of explanatory debugging of models. However, this research has focused on transparent, interpretable models, where the models can be adjusted directly [392]. LIMEADE extends the paradigm of interactive machine learning and advice taking to opaque models. Moreover, these evaluations often focus on user ratings rather than quantitative demonstrations of a model's improvement in accuracy via advice taking. As argued by [368], benchmarking and evaluation remain open problems in leveraging explanations in interactive machine learning, one form of advice taking. In our second case study, we directly quantitatively evaluate LIMEADE accuracy improvements relative to a baseline.

Recommender systems are a common domain for studying explainability and advice taking due to the feedback loop and interactivity essential to the task of recommendation [3, 48, 133, 216, 306, 371, 373, 374, 386, 412]. Some recommender systems take a human's advice via affordances other than rating content [133]. The majority of these systems enable advice taking in response to a global explanation of the system's behavior [27, 43, 44, 47, 122, 158, 165, 174, 252, 285, 326, 330, 375]. Others enable advice taking in response to instance-level explanations or no explanations at all [2, 132, 185, 384]. The combined affordances of advice taking and explainability can lead to a higher degree of user satisfaction [144]; more trust in and perceived control of the system [74, 144, 305, 383]; and better mental models, without significantly increasing the cognitive load [184, 186, 326]. In contrast to LIMEADE, however, all of this work either relies on interpretable models or implements advice taking in an algorithmic-specific fashion that is not extensible to an arbitrary opaque machine-learned model.

4.6.2 Enabling Machine Learners to Take Human Advice: Architecture-Specific Models

Other work has explored the extension of advice taking to specific classes of opaque models, such as neural architectures. Like LIMEADE, the methods proposed by both Rieger *et al.* [323] and Ross *et al.* [327] accept human input in response to advice given in terms of an explanatory vocabulary, but their methods are restricted to differentiable models whose gradients can be accessed. Rieger *et al.* modify the loss function in order to incorporate a "contextual decomposition explanation penalization" that encodes a human's domain knowledge in response to an explanation; and Ross *et al.* modify the loss function through input gradient penalization as a form of regularization. However, both methods are largely evaluated with simulated experiments on small, artificial datasets, where the confounds are often synthetically generated. With DECOY-MNIST, for example, the training data is artificially colored systematically, leading a learner to recognize color rather than shapes. Some methods can effectively adjust the loss function to reflect advice like "ignore color" yielding more robust behavior, but in the real world confounds are much more complex, and it is not clear that these methods generalize well, even for their specific architecture classes.

Liu and Avci [215] present an NLP-specific method that allows a developer to introduce a term into the loss function that can counteract biases exposed by explanations. Specifically, the method can be used to guide a hate-speech detection model away from overly relying on tokens (such as 'gay') associated with protected groups. This is different from the feedback accepted by LIMEADE, since it says "Ignore this feature," rather than "Consider this feature to be positive/negative," but it is an important type of high-level advice. Liu and Avci tested their approach on both a synthetic and real-world domain, showing modest improvement on the latter. Unlike LIMEADE, however, their approach works only for neural models and has only been tested on an NLP toxicity detection task.

In computer vision models, researchers have created methods for analyzing the behavior of specific neurons, *e.g.* discovering one that produces foliage in a generative model; follow-on research has developed methods for similarly editing these models by rewriting the behavior of those neurons [33, 241]. While impressive, these models are both domain and architecture specific, and require great expertise on the part of the user — far from McCarthy's dream.

4.6.3 Enabling Machine Learners to Take Human Advice: Arbitrary Opaque Models

Dasgupta *et al.* [80] consider the problem of teaching an opaque learner whose representation and hypothesis class are unknown. The authors show that by interacting with the black-box learner, a teacher can efficiently find a good set of teaching instances. However, Dasgupta *et al.*'s approach is highly theoretical and assumes a noiseless version-space formulation of learning, where the concept is perfectly learnable. Most importantly, in contrast to LIMEADE, their method doesn't enable the teacher to provide advice in a high-level language.

Broadly speaking, the advice-taking interaction in LIMEADE is similar to classical human-in-the-loop active learning (AL) [337], which includes techniques that are applicable to opaque models. However, LIMEADE is distinct from typical AL in that the user is not limited to labeling instances, but can give advice on how the interpretable features should be driving model behavior (which are converted into pseudo-labeled instances using our approach). Further, AL work focuses on algorithms to select informative instances for labeling, whereas LIMEADE creates affordances for feedback on top of explanations that *the user* may choose to act upon.

Closest to our work, Schramowski *et al.* [332] present a method for adding a user into the ML training loop in order to see the AI's explanations and provide feedback to improve decision making. Like LIMEADE, their method works with an arbitrary opaque classifier, requiring only the ability to add new instances to the training set. Furthermore, they also interpret human feedback in the vocabulary used in an arbitrary, explanatory model, such as that produced by LIME [320]. However, unlike our work, Schramowski *et al.* do not provide a way for the human to explain to the AI why it made a mistake. Instead, they focus on corrections for when the model is "right for the wrong reason." Like LIMEADE, their method generates pseudo-instances, called "counter-examples," that are created by altering the selected feature of the explained instance in order to reduce confounds (including through randomization, a change to an alternative value, or a substitution with the value for that component appearing in other training instances of the same class). Furthermore, Schramowski *et al.* include only a single experiment to demonstrate their model-agnostic method: on a version of the toy MNIST dataset that was artificially biased to include decoy pixels (Table 1a [332]); their other experiments used a version of Ross *et al.*'s neural-specific loss [327].

4.6.4 Labeling Features and Creating Pseudo-instances

While canonical methods of feedback involve providing additional labeled instances [387], one approach to semi-supervised learning involves training a machine learner on labeled instances as well as labeled *features* [89, 118, 214, 312, 331, 404]. In the text classification setting, this often takes the form of labeling n-gram features. These features are then used to construct pseudo-instances (*e.g.*, documents containing just the labeled n-gram itself, labeled according to the feature's assigned label) or to power methods such as the generalized expectation criteria [404]. LIMEADE extends this semi-supervised approach by translating feature labels in an explanatory model into pseudo-instances for retraining a much more complex opaque model, which is represented using different features.

4.7 Conclusion & A Call to Action

To be an effective partner in a human-AI team, an AI system must be able to not only explain its decisions but also take advice given by humans in terms of that explanation. While interpretable classifiers such as GAMs support explanation-based advice taking, and post-hoc methods such as LIME provide *explanations* for opaque ML models, we present the first method for *updating* an arbitrary opaque model using positive and negative advice given in terms of a high-level vocabulary (such as the featurization of an explanatory model). Furthermore, we are the first to evaluate such a method on a large number (71) of real-world models and with user studies.[10] In our first case study, we used LIMEADE to implement advice taking on twenty image classification domains. We showed significant

[10]This includes 20 image classifiers in our first case study (Section 4.3), 21 paper recommenders in our user study (Section 4.4.5), and 30 paper recommenders in our simulated log study (Section 4.4.7).

improvement over a strong baseline in the two-shot case. In our second case study, we incorporated LIMEADE into Semantic Sanity, a publicly-available computer science research paper recommender. Significantly, this case study adopted a different domain and different task, demonstrating how LIMEADE is a general framework for advice taking. Our user study over 21 participants demonstrated that users strongly prefer our advice-taking system, lauding perceived control and their sense of trust. While we failed to show improved accuracy of the resulting recommender for these users, as measured with DCG, a study of the long-term logs of 30 different, organic users did show significantly improved NDCG. Furthermore, another log study uncovered a fundamental tension between canonical explanation approaches that greedily select the most influential features and those that provide the best affordances for advice taking.

Much work remains to be done. We hope to develop improved methods for interpreting human advice and better understand when such advice is useful. Experiments using different explanatory vocabularies would also be useful. Additional questions, such as simultaneous advice taking from multiple people in the non-personalized setting, are worth pursuing. Furthermore, developing other forms of advice taking remains a fruitful area for exploration. For example, enabling humans to give advice by adding features, or communicating through natural language or other forms of communication, are understudied challenges. Moreover, understanding various “hyperparameters” surrounding advice taking, such as the proper strength of an update, remains an important question both empirically and theoretically. For example, in the case of recommenders, should the strength of an advice-taking update be personalized? Should it change during different stages of updating a model? Are other methods effective at combating the explanation-action tradeoff, such as using arbitrary English feedback to generate a pseudo-instance, rather than restricting to advice written using the features surfaced in greedy explanations? While we did not evaluate LIMEADE according to improvements in fairness, robustness, or model compliance, advice taking could be used for these purposes, and another compelling direction of research concerns refining and evaluating advice-taking frameworks in this context. Lastly, it is important to further investigate the entanglement between training data and supervision with respect to advice taking, as described in Section 4.3.3.

We consider our chapter a "Call to Action" for researchers in human-AI interaction to study the advice-taking problem for opaque machine learners. From search & recommendation to image recognition to medical diagnosis, opaque machine learners are ubiquitous. End-users deserve new methods for adjusting these machine learning systems by giving advice in terms of a high-level vocabulary. To aid future research, we will release the code written for our image domain experiments, including our modified COCO dataset with the precomputed superpixel vocabulary and corresponding embeddings, as well as our functioning implementation of LIMEADE. We hope that this work will contribute to opening a new direction of research in human-AI interaction devoted to this challenging and pressing problem.

Chapter 5

CONCLUSION

This dissertation has examined the challenges raised by exploratory search and recommendation and offers several paths forward in the context of human-AI interaction. Recognizing the particular challenges of discoverability surrounding digital cultural heritage collections, this dissertation has adopted this setting throughout much of the work presented.

In Chapter 2, we demonstrated the value of large-scale cultural heritage datasets in advancing interdisciplinary research in the humanities and library & information science. In particular, we described our construction of the Newspaper Navigator dataset, consisting of extracted visual content from over 16 million digitized historic newspaper pages. We then detailed how the dataset could answer questions in fields including critical data studies, print history, Jewish studies, sociotechnical systems, and beyond.

In Chapter 3, we introduced open faceted search, a framework that empowers users to define facets in an open domain fashion during the exploratory search process. We described our implementation of open faceted search within the publicly-deployed Newspaper Navigator search application containing over 1.5 million photos from the dataset. We evaluated open faceted search using the search logs of over 42,000 user sessions in the application, demonstrating clear need for open facets, as well as promising evidence of facet learning. We also detailed broader impact of the search application. To expand open faceted search as a framework, we introduced ongoing work surrounding zero-metadata open faceted search, which enables the bootstrapping of full open faceted search interfaces using large language models, even when image collections have no associated metadata.

In Chapter 4, we introduced LIMEADE, a general framework for advice taking for opaque machine learners. We evaluated LIMEADE in two distinct settings: image classification debugging and text recommendation. For the first study, we demonstrated that

LIMEADE outperforms a strong baseline. For the second study, we implemented LIMEADE within Semantic Sanity, a publicly-deployed paper recommender system. We conducted an in-person user study demonstrating significantly higher perceived user control, trust, and satisfaction with LIMEADE. We also conducted log analysis of the publicly-deployed website. We presented this work as a "Call to Action," arguing that users should have new methods for giving advice using a high-level vocabulary in order to tune machine learning systems.

Collectively, this body of work demonstrates the importance of novel interaction affordances for exploratory search and recommendation. As end-users interact with increasingly large corpora of information using increasingly inscrutable machine learning systems, the importance of further work in human-AI interaction surrounding exploratory search and recommendation is especially important.

Moving forward, born-digital cultural heritage collections including web archives present unique challenges, ranging from unprecedented scale to interdisciplinary collaboration required to make sense of these collections. Though I have begun this line of work surrounding born-digital content with Trevor Owens in an article in the *International Journal of Digital Humanities* [204], the emerging questions are manifold and exciting. While each chapter offers concrete future work, we end this document toward a horizon of interdisciplinary work in computing cultural heritage.

Navigating and making sense of digital cultural heritage collections presents distinctly interdisciplinary challenges. From enriching metadata to developing new modes of search & discovery to excavating the sociotechnical implications of these systems, pressing questions abound. Developing new approaches in machine learning, human-computer interaction, library and information science, and the digital humanities in order to navigate and make sense of these digital collections successfully is essential. To accomplish this goal, we must foreground ethical stewardship of digitized and born-digital cultural heritage at both local and global scales. This dissertation offers the challenge of computing cultural heritage as an increasingly crucial area of research and pedagogy for scholars across disciplines over the coming years [198].

BIBLIOGRAPHY

[1] Melissa Adler. Classification Along the Color Line: Excavating Racism in the Stacks. *Journal of Critical Library and Information Studies*, 2017. URL: https://journals.litwinbooks.com/index.php/jclis/article/view/17.

[2] Jae-wook Ahn, Peter Brusilovsky, Jonathan Grady, Daqing He, and Sue Yeon Syn. Open user profiles for adaptive news systems: Help or harm? In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 11–20, New York, NY, USA, 2007. ACM. URL: http://doi.acm.org/10.1145/1242572.1242575, doi:10.1145/1242572.1242575.

[3] Jae-wook Ahn, Peter Brusilovsky, and Shuguang Han. Personalized search: Reconsidering the value of open user models. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 202–212, New York, NY, USA, 2015. ACM. URL: http://doi.acm.org/10.1145/2678025.2701410, doi:10.1145/2678025.2701410.

[4] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2927–2936, 2015. doi:10.1109/CVPR.2015.7298911.

[5] Hannah Alpert-Abrams. Machine Reading the Primeros Libros. *Digital Humanities Quarterly*, 010(4), October 2016.

[6] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, December 2014. URL: https://aaai.org/ojs/index.php/aimagazine/article/view/2513, doi:10.1609/aimag.v35i4.2513.

[7] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Overview based example selection in end user interactive concept learning. In *Proceedings of the 22nd Annual ACM Symposium on User Interface Software and Technology*, UIST '09, page 247–256, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1622176.1622222.

[8] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Examining multiple potential models in end-user interactive concept learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, page 1357–1360, New York, NY, USA, 2010. Association for Computing Machinery. doi:10.1145/1753326.1753531.

[9] Saleema Amershi, James Fogarty, Ashish Kapoor, and Desney Tan. Effective end-user interaction with machine learning. In *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, AAAI'11, page 1529–1532. AAAI Press, 2011.

[10] Saleema Amershi, James Fogarty, and Daniel Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 21–30, New York, NY, USA, 2012. ACM. URL: http://doi.acm.org/10.1145/2207676.2207680, doi:10.1145/2207676.2207680.

[11] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, and et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3290605.3300233.

[12] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19,

page 1–13, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3290605.3300233.

[13] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 84–91, New Orleans - Louisiana, June 2018. Association for Computational Linguistics. URL: https://aclanthology.org/N18-3011, doi:10.18653/v1/N18-3011.

[14] Chantal Amrhein and Simon Clematide. Supervised ocr error detection and correction using statistical and neural machine translation methods. *JLCL*, 33(1):49–76, 2018. URL: https://doi.org/10.5167/uzh-162394.

[15] Refik Anadol. Archive dreaming. http://refikanadol.com/works/archive-dreaming/, 2020.

[16] Paul André, Edward Cutrell, Desney S. Tan, and Greg Smith. Designing novel image search interfaces by understanding unique characteristics and usage. In Tom Gross, Jan Gulliksen, Paula Kotzé, Lars Oestreicher, Philippe Palanque, Raquel Oliveira Prates, and Marco Winckler, editors, *Human-Computer Interaction – INTERACT 2009*, pages 340–353, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[17] Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan. dhsegment: A generic deep-learning approach for document segmentation. In *ICFHR '18*. IEEE, 2018.

[18] Linda H. Armitage and Peter G.B. Enser. Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299, 1997. arXiv:https://doi.org/10.1177/016555159702300403, doi:10.1177/016555159702300403.

[19] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilovic, Ravi Nair, Karthikeyan Natesan Ramamurthy, Darrell Reimer, Alexandra Olteanu, David Piorkowski, Jason Tsay, and Kush R. Varshney. Factsheets: Increasing trust in ai services through supplier's declarations of conformity, 2019. arXiv:1808.07261.

[20] Taylor Arnold and Lauren Tilton. Distant viewing toolkit: A python package for the analysis of visual culture. *Journal of Open Source Software*, 5(45):1800, 2020. doi:10.21105/joss.01800.

[21] Taylor Arnold, Lauren Tilton, and Annie Berke. Visual Style in Two Network Era Sitcoms. *Journal of Cultural Analytics*, 2019. Publisher: Department of Languages, Literatures, and Cultures. doi:10.22148/16.043.

[22] Artefect Group. Tarot cards for tech, 2018. URL: http://tarotcardsoftech.artefactgroup.com/.

[23] Nongnuch Artrith, Keith T. Butler, François-Xavier Coudert, Seungwu Han, Olexandr Isayev, Anubhav Jain, and Aron Walsh. Best practices in machine learning for chemistry. *Nature Chemistry*, 13:505 – 508, 2021.

[24] Marie-Louise Ayres. 'singing for their supper': Trove, australian newspapers, and the crowd. In *IFLA WLIC 2013*, 2013.

[25] Joshua Ortiz Baco, Benjamin Charles Germain Lee, Jim Casey, and Sarah H. Salter. Toward an experimental bibliography of hemispheric reconstruction newspapers. *criticism: A Quarterly for Literature and the Arts*, 64(3-4):29 – 52, 2022. URL: https://digitalcommons.wayne.edu/criticism/vol64/iss3/15.

[26] Moya Bailey. #transform(ing)dh writing and research: An autoethnography of digital humanities and feminist ethics. *Digital Humanities Quarterly*, 9(2), 2015.

[27] Fedor Bakalov, Marie-Jean Meurs, Birgitta König-Ries, Bahar Sateli, René Witte, Greg Butler, and Adrian Tsang. An approach to controlling user models and per-

sonalization effects in recommender systems. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 49–56, New York, NY, USA, 2013. ACM. URL: http://doi.acm.org/10.1145/2449396.2449405, doi:10.1145/2449396.2449405.

[28] Nicholson Baker. *Double Fold: Libraries and the Assult on Paper.* Penguin Random House, 2001.

[29] Judit Bar-Ilan. Position paper: Access to query logs - an academic researcher's point of view. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.

[30] Kevin Barnhurst and John Nerone. *The Form of News: A History.* The Guilford Press, 2002.

[31] Kate Barrall and Christine Guenther. Microfilm selection for digitization, 2005. URL: https://www.loc.gov/ndnp/guidelines/NEH_MicrofilmSelectionNDNP.pdf.

[32] Jeannette A Bastian. Documenting Communities Through the Lens of Collective Memory. In Dominique Daniel and Amalia S. Levi, editors, *Identity Palimpsests: Archiving Ethnicity in the U.S. and Canada*, number book 5 in Archives, archivists and society, pages 15–34. Litwin Books, Sacramento, CA, 2014.

[33] David Bau, Steven Liu, Tongzhou Wang, Jun-Yan Zhu, and Antonio Torralba. Rewriting a deep generative model. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 351–369, Cham, 2020. Springer International Publishing.

[34] Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338, November 2016. doi:10.1007/s00799-015-0156-0.

[35] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space, 2006.

[36] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. URL: https://aclanthology.org/D19-1371, doi:10.18653/v1/D19-1371.

[37] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. URL: https://www.aclweb.org/anthology/Q18-1041, doi:10.1162/tacl_a_00041.

[38] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. Content-based citation recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 238–251, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/N18-1022, doi:10.18653/v1/N18-1022.

[39] Simone Bianco, Remi Cadene, Luigi Celona, and Paolo Napoletano. Benchmark analysis of representative deep neural network architectures. *IEEE Access*, 6:64270–64277, 2018. doi:10.1109/ACCESS.2018.2877890.

[40] Mazelan bin Anuar, Cally Law, and Soh Wai Yee. Challenges of digitizing vernacular newspapers & preliminary study of user behaviour on newspapersg's multilingual ui. In *IFLA 2012 Pre-Conference*, Mikkeli, Finland, 2012.

[41] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. Fairlearn: A toolkit for assessing and improving fairness in ai. Technical Report MSR-TR-2020-32, Microsoft, May 2020. URL: https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/.

[42] Tom Bober. The primary source podcast s1e3: The newspaper navigator search app, an educator's view, 2020. URL: https://podcasts.apple.com/us/podcast/s1e3-the-newspaper-navigator-search-app-an-educators-view/id1530140365?i=1000492297315.

[43] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. Tasteweights: A visual interactive hybrid recommender system. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, pages 35–42, New York, NY, USA, 2012. ACM. URL: http://doi.acm.org/10.1145/2365952.2365964, doi:10.1145/2365952.2365964.

[44] Svetlin Bostandjiev, John O'Donovan, and Tobias Höllerer. LinkedVis: exploring social and semantic career recommendations. In *Proceedings of the 2013 international conference on Intelligent user interfaces - IUI '13*, page 107, Santa Monica, California, USA, 2013. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2449396.2449412, doi:10.1145/2449396.2449412.

[45] Geoffrey C Bowker and Susan Leigh Star. *Sorting things out: classification and its consequences*. MIT Press, 2000. OCLC: 45846910.

[46] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[47] Simon Bruns, André Calero Valdez, Christoph Greven, Martina Ziefle, and Ulrik Schroeder. What Should I Read Next? A Personalized Visual Publication Recommender System. In Sakae Yamamoto, editor, *Human Interface and the Management*

of Information. Information and Knowledge in Context, pages 89–100, Cham, 2015. Springer International Publishing.

[48] Peter Brusilovsky, Marco de Gemmis, Alexander Felfernig, Pasquale Lops, John O'Donovan, Giovanni Semeraro, and Martijn C. Willemsen. Interfaces and human decision making for recommender systems. In *Proceedings of the 14th ACM Conference on Recommender Systems*, RecSys '20, page 613–618, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3383313.3411539.

[49] Ania Calderon, Dan Taber, Hong Qu, and Jeff Wen. Ai blindspot discovery process. URL: https://aiblindspot.media.mit.edu/.

[50] Yang Cao, Hai Wang, Changhu Wang, Zhiwei Li, Liqing Zhang, and Lei Zhang. Mindfinder: interactive sketch-based image search on millions of images. In Alberto Del Bimbo, Shih-Fu Chang, and Arnold W. M. Smeulders, editors, *ACM Multimedia*, pages 1605–1608. ACM, 2010. URL: http://dblp.uni-trier.de/db/conf/mm/mm2010.html#CaoWWLZZ10.

[51] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery. doi:10.1145/2783258.2788613.

[52] Steve Cassidy. Publishing the trove newspaper corpus. In *LREC'16*, pages 4520–4525, Portorož, Slovenia, 2016. URL: https://www.aclweb.org/anthology/L16-1715.

[53] Natalia Cecire. Works cited: The visible hand, 2011. URL: http://nataliacecire.blogspot.com/2011/05/visible-hand.html.

[54] Yiddish Book Center. Yiddish ocr is live! URL: https://www.yiddishbookcenter.org/about/news/yiddish-ocr-live.

[55] Joseph Chee Chang, Saleema Amershi, and Ece Kamar. Revolt: Collaborative crowdsourcing for labeling machine learning datasets. In *Proceedings of the 2017 CHI*

Conference on Human Factors in Computing Systems, CHI '17, page 2334–2346, New York, NY, USA, 2017. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/3025453.3026044, doi:10.1145/3025453.3026044.

[56] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. Searchlens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, page 498–509, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3301275.3302321.

[57] Krishnendu Chaudhury, Ankur Jain, Sriram Thirthala, Vivek Sahasranaman, Shobhit Saxena, and Selvam Mahalingam. Google Newspaper Search: Image Processing and Analysis Pipeline. In *ICDAR '09*, pages 621–625, Barcelona, Spain, 2009. IEEE. URL: http://ieeexplore.ieee.org/document/5277572/, doi:10.1109/ICDAR.2009.272.

[58] Hsin-liang Chen. An analysis of image queries in the field of art history. *Journal of the American Society for Information Science and Technology*, 52(3):260–273, 2001. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/1532-2890%282000%299999%3A9999%3C%3A%3AAID-ASI1606%3E3.0.CO%3B2-M. URL: http://onlinelibrary.wiley.com/doi/abs/10.1002/1532-2890%282000%299999%3A9999%3C%3A%3AAID-ASI1606%3E3.0.CO%3B2-M, doi:10.1002/1532-2890(2000)9999:9999<::AID-ASI1606>3.0.CO;2-M.

[59] Hsin-Liang Chen. Image retrieval knowledge and art history curriculum in the digital age. *Proceedings of the 2001 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing*, 2001.

[60] Lydia B. Chilton, Greg Little, Darren Edge, Daniel S. Weld, and James A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, page 1999–2008, New

York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2470654.2466265.

[61] Shruthi Sai Chivukula, Ziqing Li, Anne C. Pivonka, Jingning Chen, and Colin M. Gray. Surveying the landscape of ethics-focused design methods, 2021. arXiv:2102.08909.

[62] Youngok Choi and Edie M. Rasmussen. Searching for images: the analysis of users' queries for image retrieval in American history. *Journal of the American Society for Information Science and Technology*, 54(6):498–511, April 2003. doi:10.1002/asi.10237.

[63] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online, July 2020. Association for Computational Linguistics. URL: https://aclanthology.org/2020.acl-main.207, doi:10.18653/v1/2020.acl-main.207.

[64] Giovanni Colavizza, Tobias Blanke, Charles Jeurgens, and Julia Noordegraaf. Archives and ai: An overview of current debates and future perspectives. *J. Comput. Cult. Herit.*, 2021. doi:10.1145/3479010.

[65] Jean Lee Cole. *How the Other Half Laughs: The Comic Sensibility in American Culture, 1895-1920.* U Press of Mississippi, Jackson, 2020.

[66] Karen Collins. Providing Subject Access to Images: A Study of User Queries. *The American Archivist*, 61(1):36–55, January 1998. URL: https://meridian.allenpress.com/american-archivist/article/61/1/36/23894/Providing-Subject-Access-to-Images-A-Study-of-User, doi:10.17723/aarc.61.1.b531vt5q0q620642.

[67] NAACL Conference Committee. Reproducibility Checklist, 2021. URL:

https://iuliaturc-google.github.io/naacl-org/naacl-2021-website/calls/reproducibility-checklist/.

[68] Lisa Louise Coooke. Genealogy gems: How to use chronicling america's newspaper navigator to find photos and images - free tool, 9 2020. URL: https://www.youtube.com/watch?v=Yp1syVr3vf4.

[69] Ryan Cordell. "Q i-jtb the Raven": Taking Dirty OCR Seriously. *Book History*, 20(1):188–225, 2017. Publisher: Johns Hopkins University Press. URL: http://muse.jhu.edu/article/674968, doi:10.1353/bh.2017.0006.

[70] Ryan Cordell. Machine Learning + Libraries: A report on the state of the field, 2020. URL: https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

[71] Ryan Cordell, M. H Beals, Isabel G Russell, Julianne Nyhan, Ernesto Priani, Marc Priewe, Hannu Salmi, Jaap Verheul, Raquel Alegre, Tessa Hauswedell, and et al. Oceanic exchanges, 2019. URL: osf.io/wa94s, doi:10.17605/OSF.IO/WA94S.

[72] Ryan Cordell and David Smith. Viral texts: Mapping networks of reprinting in 19th-century newspapers and magazines, 2017. URL: http://viraltexts.org.

[73] Sasha Costanza-Chock and Pablo Rey-Mazón. Pageonex: New Approaches to Newspaper Front Page Analysis. *International Journal of Communication*, 10:28, 2016.

[74] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, August 2008. doi:10.1007/s11257-008-9051-3.

[75] Kate Crawford and Trevor Paglen. Excavating ai: The politics of images in machine learning training sets, 2019. URL: https://excavating.ai/.

[76] Anamaria Crisan, Margaret Drouhard, Jesse Vig, and Nazneen Rajani. Interac-

tive model cards: A human-centered approach to model documentation. *ArXiv*, abs/2205.02894, 2022.

[77] Dominique Daniel. Archival Representations of Immigration and Ethnicity in North American History: From the Ethnicization of Archives to the Archivization of Ethnicity. *Archival Science*, 14(2):169–203, June 2014. URL: http://link.springer.com/10.1007/s10502-013-9209-6, doi:10.1007/s10502-013-9209-6.

[78] Shubhomoy Das, Travis Moore, Weng-Keen Wong, Simone Stumpf, Ian Oberst, Kevin Mcintosh, and Margaret Burnett. End-user feature labeling: Supervised and semi-supervised approaches based on locally-weighted logistic regression. *Artif. Intell.*, 204:56–74, nov 2013. doi:10.1016/j.artint.2013.08.003.

[79] Anish Das Sarma, Xin Dong, and Alon Halevy. Bootstrapping pay-as-you-go data integration systems. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 861–874, New York, NY, USA, 2008. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/1376616.1376702, doi:10.1145/1376616.1376702.

[80] Sanjoy Dasgupta, Daniel Hsu, Stefanos Poulis, and Xiaojin Zhu. Teaching a black-box learner. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1547–1555, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL: http://proceedings.mlr.press/v97/dasgupta19a.html.

[81] Delpher. About delpher, N/A. URL: https://www.delpher.nl/over-delpher/wat-is-delpher/maak-kennis-met-delpher.

[82] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi:10.1109/CVPR.2009.5206848.

[83] Jia Deng, Olga Russakovsky, Jonathan Krause, Michael S. Bernstein, Alex Berg, and Li Fei-Fei. Scalable multi-label annotation. In *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems, CHI '14, page 3099–3102, New York, NY, USA, 2014. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/2556288.2557011, doi:10.1145/2556288.2557011.

[84] Cyril Diagne, Nicolas Barradeau, and Simon Doury. t-sne map. https://experiments.withgoogle.com/t-sne-map, 2018.

[85] Nicholas Diakapoulos, Sorelle Friedler, Marcelo Arenas, Solon Barocas, Michael Hay, Bill Howe, H. V. Jagadish, Kris Unsworth, Arnaud Sahuguet, Suresh Venkatasubramanian, Christo Wilson, Cong Yu, and Bendert Zevenbergen. Principles for accountable algorithms and a social impact statement for algorithms. URL: https://www.fatml.org/resources/principles-for-accountable-algorithms.

[86] Nicholas Diakopoulos. Algorithmic Accountability Reporting: On the Investigation of Black Boxes, 2014. doi:10.7916/D8ZK5TW2.

[87] Nicholas Diakopoulos. Accountability in algorithmic decision making. *Commun. ACM*, 59(2):56–62, January 2016. doi:10.1145/2844110.

[88] Finale Doshi-Velez and Been Kim. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]*, March 2017. arXiv: 1702.08608. URL: http://arxiv.org/abs/1702.08608.

[89] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, page 595–602, New York, NY, USA, 2008. Association for Computing Machinery. doi:10.1145/1390334.1390436.

[90] Douglas Duhaime. Pixplot. https://github.com/YaleDHLab/pix-plot, 2020.

[91] Michael Ekstrand, Praveen Kannan, James Stemper, John Butler, Joseph Konstan, and John Riedl. Automatically building research reading lists. In *RecSys '10*, page 159–166. ACM, 2010. doi:10.1145/1864708.1864740.

[92] Ame Elliott. Flamenco image browser: Using metadata to improve image search during architectural design. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, page 69–70, New York, NY, USA, 2001. Association for Computing Machinery. doi:10.1145/634067.634112.

[93] Jennifer English, Marti Hearst, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. Hierarchical faceted metadata in site search interfaces. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '02, page 628–639, New York, NY, USA, 2002. Association for Computing Machinery. doi:10.1145/506443.506517.

[94] P. Enser and C. McGregor. Analysis of visual information retrieval queries: Report on project g16412 to the british library research and development department, 1992.

[95] Benjamin Fagan. Chronicling white america. *American Periodicals: A Journal of History & Criticism*, 26(1):10 – 13, 2016. URL: https://muse.jhu.edu/article/613375.

[96] Jianping Fan, Yuli Gao, Hangzai Luo, Daniel A. Keim, and Zongmin Li. A novel approach to enable semantic and visual image summarization for exploratory image search. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, MIR '08, page 358–365, New York, NY, USA, 2008. Association for Computing Machinery. doi:10.1145/1460096.1460155.

[97] Hayward Farrar. *The Baltimore Afro-American*. Greenwood Publishing Group, 1998.

[98] Julien Fauqueur and Nozha Boujemaa. Mental image search by boolean composition of region categories. *Multimedia Tools Appl.*, 31(1):95–117, October 2006. URL: https://doi-org.offcampus.lib.washington.edu/10.1007/s11042-006-0033-3, doi:10.1007/s11042-006-0033-3.

[99] Kathleen Fear. User Understanding of Metadata in Digital Image Collections: Or, What Exactly Do You Mean by "Coverage"? *The American Archivist*, 73(1):26–60, 05 2010. arXiv:https://meridian.allenpress.com/american-archivist/

article-pdf/73/1/26/2055035/aarc_73_1_j00044lr77415551.pdf, doi:10.17723/aarc.73.1.j00044lr77415551.

[100] Raya Fidel. The image retrieval task: implications for the design and evaluation of image databases. *New Review of Hypermedia and Multimedia*, 3(1):181–199, 1997. arXiv:https://doi.org/10.1080/13614569708914689, doi:10.1080/13614569708914689.

[101] Marco Fiorucci, Marina Khoroshiltseva, Massimiliano Pontil, Arianna Traviglia, Alessio Del Bue, and Stuart James. Machine learning for cultural heritage: A survey. *Pattern Recognition Letters*, 133:102–108, 2020. URL: https://www.sciencedirect.com/science/article/pii/S0167865520300532, doi:https://doi.org/10.1016/j.patrec.2020.02.017.

[102] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. Cueflik: Interactive concept learning in image search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 29–38, New York, NY, USA, 2008. Association for Computing Machinery. doi:10.1145/1357054.1357061.

[103] Claudio Fogu, Wulf Kansteiner, and Todd Samuel Presner. *Probing the ethics of Holocaust culture*. Harvard University Press, 2016. OCLC: 946610654.

[104] Brian Foo. Citizen dj, 2020. URL: https://citizen-dj.labs.loc.gov/.

[105] Brian Foo. Citizen dj copyright and ethics guide, 2020. URL: https://citizen-dj.labs.loc.gov/guide/.

[106] Brian Foo. Visualizing amnh image collection with machine learning. https://github.com/amnh-sciviz/image-collection, 2020.

[107] Carbon Footprint. Carbon footprint calculator, 2020. URL: https://calculator.carbonfootprint.com/calculator.aspx?lang=en-GB&tab=3.

[108] Europeana AI for GLAMs Task Force. Europeanatech ai in relation to glams report & recommendations, 2021. URL: https://pro.europeana.eu/files/Europeana_

Professional/Europeana_Network/Europeana_Network_Task_Forces/Final_reports/AI%20in%20relation%20to%20GLAMs%20Task%20Force%20Report.pdf.

[109] Frances Smith Foster. A Narrative of the Interesting Origins and (Somewhat) Surprising Developments of African-American Print Culture. *American Literary History*, 17(4):714–740, 2005. <p>Volume 17, Number 4, Winter 2005</p>. URL: http://muse.jhu.edu/journals/american_literary_history/v017/17.4foster.html.

[110] Batya Friedman and David G. Hendry. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '12*, pages 1145–1148, 2012.

[111] Jonathan Furner. Dewey Deracialized: A Critical Race-Theoretic Perspective. *KNOWLEDGE ORGANIZATION*, 34(3):144–168, 2007. URL: https://www.nomos-elibrary.de/index.php?doi=10.5771/0943-7444-2007-3-144, doi:10.5771/0943-7444-2007-3-144.

[112] Paul Fyfe and Qian Ge. Image analytics and the nineteenth-century illustrated newspaper. *Journal of Cultural Analytics*, 3(1), 10 2018. doi:10.22148/16.026.

[113] Paul K. Fyfe. An archaeology of victorian newspapers. *Victorian Periodicals Review*, 49:546 – 577, 2016.

[114] Eric Gardner. *Black Print Unbound: The Christian Recorder, African American Literature, and Periodical Culture*. Oxford University Press, September 2015. Google-Books-ID: 76n2CQAAQBAJ.

[115] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III au2, and Kate Crawford. Datasheets for datasets, 2020. arXiv:1803.09010.

[116] Mauricio Geraldo. Building aero — dx lab - state library of nsw, 2020. URL: https://dxlab.sl.nsw.gov.au/blog/building-aereo/.

[117] Paula Giddings. *Ida: A Sword Among Lions: Ida B. Wells and the Campaign against Lynching*. HarperCollins, New York, N.Y., 2009.

[118] Shantanu Godbole and Sunita Sarawagi. Discriminative methods for multi-labeled classification. In Honghua Dai, Ramakrishnan Srikant, and Chengqi Zhang, editors, *Advances in Knowledge Discovery and Data Mining*, pages 22–30, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg.

[119] Matthew K. Gold and Lauren Klein. Debates in the digital humanities, 2012.

[120] Joshua Gomez, Kristian Allen, Mark Matney, Tinuola Awopetu, and Sharon Shafer. Experimenting with a machine generated annotations pipeline. *The Code4Lib Journal*, (48), May 2020. URL: https://journal.code4lib.org/articles/15209.

[121] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, October 2017. arXiv: 1606.08813. URL: http://arxiv.org/abs/1606.08813, doi:10.1609/aimag.v38i3.2741.

[122] Brynjar Gretarsson, John O'Donovan, Svetlin Bostandjiev, Christopher Hall, and Tobias Höllerer. Smallworlds: Visualizing social recommendations. *Computer Graphics Forum*, 29(3):833–842, 2010. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8659.2009.01679.x, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-8659.2009.01679.x, doi:10.1111/j.1467-8659.2009.01679.x.

[123] Michael Griffin. The great war photographs: Constructing myths of history and photojournalism. In Bonnie Brennan and Hanno Hardt, editors, *Picturing the Past: Media, History, and Photography*, pages 122–157. University of Illinois Press, 1999.

[124] T. Grüning, R. Labahn, M. Diem, F. Kleber, and S. Fiel. Read-bad: A new dataset and evaluation scheme for baseline detection in archival documents. In *13th IAPR Intl. Workshop on Document Analysis Systems*, pages 351–356, 2018. doi:10.1109/DAS.2018.38.

[125] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Comput. Surv.*, 51(5), August 2018. doi:10.1145/3236009.

[126] Jennifer Guiliano and Roopika Risam. Reviews in the digital humanities, 2021. URL: https://reviewsindh.pubpub.org/project-registry-a-z.

[127] Furkan Gursoy and I. Kakadiaris. System cards for ai-based decision-making for public policy. *ArXiv*, abs/2203.04754, 2022.

[128] Kristin L. Gustafson. Ethnic Newspaper Producers Face Archiving Challenges. *Newspaper Research Journal*, 36(3):314–327, September 2015. doi:10.1177/0739532915600744.

[129] Francoise Hanssen-Bauer, Magnus Bognerud, Dag Hensten, Gro Benedikte Pedersen, Even Westvang, and Audun Mathias Øygard. t-sne map. https://github.com/nasjonalmuseet/propinquity, 2018.

[130] Kotaro Hara, Abi Adams, Kristy Milland, Saiph Savage, Chris Callison-Burch, and Jeffrey Bigham. A data-driven analysis of workers' earnings on amazon mechanical turk, 2017. arXiv:1712.05796.

[131] Molly Hardy and DiCuirci Lindsay. Critical cataloging and the serials archive: The digital making of "mill girls in nineteenth-century print". *Archive Journal*, 2019. URL: https://www.archivejournal.net/essays/critical-cataloging-and-the-serials-archive-the-digital-making-of-mill-girls-in-ninet

[132] F. Maxwell Harper, Funing Xu, Harmanpreet Kaur, Kyle Condiff, Shuo Chang, and Loren Terveen. Putting Users in Control of their Recommendations. In *Proceedings of the 9th ACM Conference on Recommender Systems - RecSys '15*, pages 3–10, Vienna, Austria, 2015. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2792838.2800179, doi:10.1145/2792838.2800179.

[133] Chen He, Denis Parra, and Katrien Verbert. Interactive recommender systems: A survey of the state of the art and future research challenges and op-

portunities. *Expert Systems with Applications*, 56:9–27, September 2016. URL: https://linkinghub.elsevier.com/retrieve/pii/S0957417416300367, doi:10.1016/j.eswa.2016.02.013.

[134] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi:10.1109/CVPR.2016.90.

[135] Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. Context-aware citation recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 421–430, New York, NY, USA, 2010. ACM. URL: http://doi.acm.org/10.1145/1772690.1772734, doi:10.1145/1772690.1772734.

[136] M. A. Hearst. Design recommendations for hierarchical faceted search interfaces. In *Proc. SIGIR 2006, Workshop on Faceted Search*, pages 26–30, August 2006.

[137] Marti Hearst. *Search user interfaces*. Cambridge University Press, Cambridge; New York, 2009. OCLC: 401146420.

[138] Marti Hearst, Ame Elliott, Jennifer English, Rashmi Sinha, Kirsten Swearingen, and Ka-Ping Yee. Finding the flow in web site search. *Commun. ACM*, 45(9):42–49, September 2002. doi:10.1145/567498.567525.

[139] Marti A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, April 2006. doi:10.1145/1121949.1121983.

[140] Marti A. Hearst. Uis for faceted navigation recent advances and remaining open problems, 2008.

[141] Marti A. Hearst and Emilia Stoica. NLP support for faceted navigation in scholarly collection. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries (NLPIR4DL)*, pages 62–70, Suntec City, Singapore, August 2009. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/W09-3608.

[142] Mark Hedges, Richard Marciano, and Eirini Goudarouli. Introduction to the special issue on computational archival science. *J. Comput. Cult. Herit.*, 15(1), jan 2022. doi:10.1145/3495004.

[143] D. Heesch. A survey of browsing models for content based image retrieval. *Multimedia Tools and Applications*, 40:261–284, 2008.

[144] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work - CSCW '00*, pages 241–250, Philadelphia, Pennsylvania, United States, 2000. ACM Press. URL: http://portal.acm.org/citation.cfm?doid=358916.358995, doi:10.1145/358916.358995.

[145] Tim Hitchcock. Confronting the digital. *Cultural and Social History*, 10(1):9–23, 2013. arXiv:https://doi.org/10.2752/147800413X13515292098070, doi:10.2752/147800413X13515292098070.

[146] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards, 2018. arXiv:1805.03677.

[147] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL: http://www.jstor.org/stable/4615733.

[148] Enamul Hoque, Orland Hoeber, and Minglun Gong. Cider: Concept-based image diversification, exploration, and retrieval. *Inf. Process. Manage.*, 49(5):1122–1138, September 2013. doi:10.1016/j.ipm.2012.12.001.

[149] Sayantan Hore, D. Glowacka, Ilkka Kosunen, K. Athukorala, and G. Jacucci. Futureview: Enhancing exploratory image search. In *IntRS@RecSys*, 2015.

[150] ImageNet. What about the images?, 2020. URL: https://image-net.org/download-faq.

[151] Amazon Web Services Inc. Amazon ec2 instance types, 2020. URL: https://aws.amazon.com/ec2/instance-types/.

[152] Smithsonian Institution. Smithsonian digital volunteers: About. https://transcription.si.edu/about, N/A.

[153] Intel. Intel©xeon©platinum 8256 processor (16.5m cache, 3.80 ghz) product specifications, 2020. URL: https://ark.intel.com/content/www/us/en/ark/products/192467/intel-xeon-platinum-8256-processor-16-5m-cache-3-80-ghz.html.

[154] Intel. Intel©xeon©platinum 9242 processor (71.5m cache, 2.30 ghz) product specifications, 2020. URL: https://ark.intel.com/content/www/us/en/ark/products/194145/intel-xeon-platinum-9242-processor-71-5m-cache-2-30-ghz.html.

[155] Lise Jaillant and Annalina Caputo. Unlocking digital archives: Cross-disciplinary perspectives on ai and born-digital data. *AI Soc.*, 37(3):823–835, sep 2022. doi:10.1007/s00146-021-01367-x.

[156] Eileen Jakeway, Lauren Algee, Laurie Allen, Meghan Ferriter, Jaime Mears, Abigail Potter, and Kate Zwaard. Machine learning + libraries summit event summary, 2020. URL: https://labs.loc.gov/static/labs/meta/ML-Event-Summary-Final-2020-02-13.pdf.

[157] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[158] Yucheng Jin, Nava Tintarev, and Katrien Verbert. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems - RecSys '18*, pages 13–21, Vancouver, British Columbia, Canada, 2018. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=3240323.3240358, doi:10.1145/3240323.3240358.

[159] Shamele Jordon. Genealogy quick start tv, 9 2020. URL: https://www.youtube.com/watch?v=EUkZLAgckXc.

[160] Corinne Jörgensen. Attributes of images in describing tasks. *Inf. Process. Manag.*, 34:161–174, 1998.

[161] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.560, doi:10.18653/v1/2020.acl-main.560.

[162] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. Transkribus - a service platform for transcription, recognition and retrieval of historical documents. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 04, pages 19–24, 2017. doi:10.1109/ICDAR.2017.307.

[163] Subbarao Kambhampati, Sarath Sreedharan, Mudit Verma, Yantian Zha, and Lin Guan. Symbols as a lingua franca for bridging human-ai chasm for explainable and advisable ai systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(11):12262–12267, Jun. 2022. URL: https://ojs.aaai.org/index.php/AAAI/article/view/21488, doi:10.1609/aaai.v36i11.21488.

[164] Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. A scalable hybrid research paper recommender system for microsoft academic. In *The World Wide Web Conference*, WWW '19, pages 2893–2899, New York, NY, USA, 2019. ACM. URL: http://doi.acm.org/10.1145/3308558.3313700, doi:10.1145/3308558.3313700.

[165] Antti Kangasrääsiö, Dorota Glowacka, and Samuel Kaski. Improving controllability and predictability of interactive recommendation interfaces for exploratory search. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 247–251, New York, NY, USA, 2015. ACM. URL: http://doi.acm.org/10.1145/2678025.2701371, doi:10.1145/2678025.2701371.

[166] Andrej Karpathy. Arxiv sanity preserver, 2015. URL: http://www.arxiv-sanity.com/.

[167] Michael Katell, Meg Young, Bernease Herman, Dharma Dailey, Aaron Tam, Vivian Guetler, Corinne Binz, Daniella Raz, and P. M. Krafft. An Algorithmic Equity Toolkit for Technology Audits by Community Advocates and Activists. *arXiv:1912.02943 [cs]*, December 2019. arXiv: 1912.02943. URL: http://arxiv.org/abs/1912.02943.

[168] Jacqueline Katz. Innovations with digitized newspapers, 4 2023. URL: https://www.loc.gov/static/programs/teachers/professional-development/webinar/documents/Chronicling%20America%20Webinar.pdf.

[169] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel Weld, and Jonathan Bragg. Feedlens: Polymorphic lenses for personalizing exploratory search over knowledge graphs. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*, UIST '22, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3526113.3545631.

[170] Jaana Kekäläinen. Binary and graded relevance in ir evaluations—comparison of the effects on ranking of ir systems. *Information Processing & Management*, 41(5):1019–1033, 2005. URL: https://www.sciencedirect.com/science/article/pii/S0306457305000075, doi:https://doi.org/10.1016/j.ipm.2005.01.004.

[171] W. Kienreich, V. Sabol, M. Granitzer, W. Klieber, M. Lux, and W. Sarka. A visual query interface for a very large newspaper article repository. In *16th International Workshop on Database and Expert Systems Applications (DEXA'05)*, pages 415–419, 2005. doi:10.1109/DEXA.2005.35.

[172] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In Jennifer G. Dy and Andreas Krause, editors, *ICML*, volume 80 of *Proceedings of Machine Learning Research*,

pages 2673–2682. PMLR, 2018. URL: http://dblp.uni-trier.de/db/conf/icml/icml2018.html#KimWGCWVS18.

[173] Yanir Kleiman, Joel Lanir, Dov Danon, Yasmin Felberbaum, and Daniel Cohen-Or. Dynamicmaps: Similarity-based browsing through a massive set of images. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, page 995–1004, New York, NY, USA, 2015. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/2702123.2702224, doi:10.1145/2702123.2702224.

[174] Bart P. Knijnenburg, Niels J.M. Reijmer, and Martijn C. Willemsen. Each to his own: How different users call for different interaction methods in recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 141–148, New York, NY, USA, 2011. ACM. URL: http://doi.acm.org/10.1145/2043932.2043960, doi:10.1145/2043932.2043960.

[175] Steven A. Knowlton. Three decades since prejudices and antipathies: A study of changes in the library of congress subject headings. *Cataloging & Classification Quarterly*, 40(2):123–145, 2005. arXiv:https://doi.org/10.1300/J104v40n02_08, doi:10.1300/J104v40n02_08.

[176] Janin Koch, Nicolas Taffin, Michel Beaudouin-Lafon, Markku Laine, Andrés Lucero, and Wendy E. Mackay. Imagesense: An intelligent collaborative ideation tool to support diverse human-computer partnerships. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1), May 2020. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/3392850, doi:10.1145/3392850.

[177] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *ArXiv*, abs/2007.04612, 2020.

[178] Weize Kong. Extending Faceted Search to the Open-Domain Web. *ACM SIGIR*

Forum, 50(1):90–91, June 2016. URL: https://dl.acm.org/doi/10.1145/2964797.2964814, doi:10.1145/2964797.2964814.

[179] Weize Kong and James Allan. Extending Faceted Search to the General Web. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 839–848, Shanghai China, November 2014. ACM. URL: https://dl.acm.org/doi/10.1145/2661829.2661964, doi:10.1145/2661829.2661964.

[180] Adriana Kovashka, Devi Parikh, and Kristen Grauman. Whittlesearch: Image Search with Relative Attribute Feedback. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[181] Kelley Kreitz. American Alternatives: Participatory Futures of Print from New York City's Nineteenth-Century Spanish-Language Press. *American Literary History*, 30(4, 1):677–702, 2018. doi:https://doi.org/10.1093/alh/ajy032.

[182] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. What do exploratory searchers look at in a faceted search interface? In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09, page 313–322, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1555400.1555452.

[183] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. Structured labeling for facilitating concept evolution in machine learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, page 3075–3084, New York, NY, USA, 2014. Association for Computing Machinery. doi:10.1145/2556288.2557238.

[184] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, New York, NY, USA, 2015. ACM. URL: http://doi.acm.org/10.1145/2678025.2701399, doi:10.1145/2678025.2701399.

[185] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI '12*, page 1, Austin, Texas, USA, 2012. ACM Press. URL: http://dl.acm.org/citation.cfm?doid=2207676.2207678, doi:10.1145/2207676.2207678.

[186] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, 2013. doi:10.1109/VLHCC.2013.6645235.

[187] LC Labs. Beyond words, mark, 2017. URL: beyondwords.labs.loc.gov/#/mark.

[188] LC Labs. Beyond words, transcribe, 2017. URL: beyondwords.labs.loc.gov/#/transcribe.

[189] LC Labs. Beyond words, verify, 2017. URL: beyondwords.labs.loc.gov/#/verify.

[190] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning, 2019. arXiv:1910.09700.

[191] Sean Latham and Robert Scholes. The Rise of Periodical Studies. *PMLA/Publications of the Modern Language Association of America*, 121(2):517–531, 2006. doi:10.1632/003081206X129693.

[192] LC Labs. The fy 2019-2023 digital strategic plan of the library of congress version 1.1.2, 2019. URL: https://www.loc.gov/static/portals/digital-strategy/documents/Library-of-Congress-Digital-Strategy-v1.1.2.pdf.

[193] Benjamin Charles Germain Lee. The "collections as ml data" checklist for machine learning and cultural heritage. *Journal of the Association for Information Science and Technology*, n/a(n/a). URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.24765, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.24765, doi:https://doi.org/10.1002/asi.24765.

[194] Benjamin Charles Germain Lee. Machine learning, template matching, and the International Tracing Service digital archive: Automating the retrieval of death certificate reference cards from 40 million document scans. *Digital Scholarship in the Humanities*, 34(3):513–535, 2018. doi:10.1093/llc/fqy063.

[195] Benjamin Charles Germain Lee. Newspaper navigator github repository, 2020. URL: https://github.com/LibraryOfCongress/newspaper-navigator.

[196] Benjamin Charles Germain Lee. Compounded Mediation: A Data Archaeology of the Newspaper Navigator Dataset. *Digital Humanities Quarterly*, 15(4), 2021. URL: http://www.digitalhumanities.org/dhq/vol/15/4/000578/000578.html.

[197] Benjamin Charles Germain Lee. The digital humanities and the ladino press: Using machine learning to extract and analyze visual content in historic ladino newspapers. In Gerben Zaagsma, Daniel Stökl Ben Ezra, Miriam Rürup, Michelle Margolis, and Amalia S. Levi, editors, *Jewish Studies in the Digital Age*, pages 189–214, Berlin, Boston, 2022. De Gruyter Oldenbourg. URL: https://doi.org/10.1515/9783110744828-010, doi:doi:10.1515/9783110744828-010.

[198] Benjamin Charles Germain Lee. Computer science research and digital humanities questions. In Simon Appleford, Gabriel Hankins, and Anouk Lang, editors, *Digital Futures of Graduate Study in the Humanities*. University of Minnesota Press, 2023.

[199] Benjamin Charles Germain Lee, Joshua Ortiz Baco, Sarah H. Salter, and Jim Casey. Navigating the mise-en-page: Interpretive machine learning approaches to the visual layouts of multi-ethnic periodicals. In Maud Ehrmann, Folgert Karsdorp, Melvin Wevers, Tara Lee Andrews, Manuel Burghardt, Mike Kestemont, Enrique Manjavacas, Michael Piotrowski, and Joris van Zundert, editors, *Proceedings of the Conference on Computational Humanities Research, CHR2021, Amsterdam, The Netherlands, November 17-19, 2021*, volume 2989 of *CEUR Workshop Proceedings*, pages 49–61. CEUR-WS.org, 2021. URL: https://ceur-ws.org/Vol-2989/short_paper3.pdf.

[200] Benjamin Charles Germain Lee, Ilene R. Berson, and Michael J. Berson.

Machine learning and the social studies. *Social Education*, March 2021. URL: https://www.socialstudies.org/social-education/85/2/machine-learning-and-social-studies.

[201] Benjamin Charles Germain Lee, Doug Downey, Kyle Lo, and Daniel S. Weld. Limeade: From ai explanations to advice taking. *ACM Trans. Interact. Intell. Syst.*, mar 2023. Just Accepted. doi:10.1145/3589345.

[202] Benjamin Charles Germain Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, Chris Adams, Nathan Yarasavage, Deborah Thomas, Kate Zwaard, and Daniel S. Weld. The newspaper navigator dataset: Extracting headlines and visual content from 16 million historic newspaper pages in chronicling america. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 3055–3062, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3340531.3412767.

[203] Benjamin Charles Germain Lee, Jaime Mears, Eileen Jakeway, Meghan Ferriter, and Abigail Potter. Newspaper navigator: Putting machine learning in the hands of library users. *EuropeanaTech Insight*, 16, 2020. URL: https://pro.europeana.eu/page/issue-16-newspapers.

[204] Benjamin Charles Germain Lee and Trevor Owens. Grappling with the scale of born-digital government publications: Toward pipelines for processing and searching millions of pdfs. *International Journal of Digital Humanities*, 3:91–114, 2022. doi:10.1007/s42803-022-00042-x.

[205] Benjamin Charles Germain Lee and Daniel S. Weld. Newspaper navigator: Open faceted search for 1.5 million images. In *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, page 120–122, New York, NY, USA, 2020. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/3379350.3416143, doi:10.1145/3379350.3416143.

[206] LibCrowds. Libcrowds documentation. https://docs.libcrowds.com/, N/A.

[207] British Library and Findmypast. About the british newspaper archive, N/A. URL: http://www.britishnewspaperarchive.co.uk/help/about.

[208] New York Public Library. About emigrant city. http://emigrantcity.nypl.org/#/about, N/A.

[209] Bernhard Liebl and Manuel Burghardt. An evaluation of dnn architectures for page segmentation of historical newspapers, 2020. arXiv:2004.07317.

[210] Tsung-Yi Lin, Michael Maire, Serge Belongie, and et al. Microsoft coco: Common objects in context. In *ECCV '14*, pages 740–755. Springer Intl. Publishing, 2014.

[211] Yiling Lin, Jae-wook Ahn, Peter Brusilovsky, Daqing He, and William Real. Imagesieve: Exploratory search of museum archives with named entity-based faceted browsing. In *Proceedings of the 73rd ASIS&T Annual Meeting on Navigating Streams in an Information Ecosystem - Volume 47*, ASIS&T '10, USA, 2010. American Society for Information Science.

[212] Matthew Lincoln, Julia Corrin, Emily Davis, and Scott B. Weingart. Campi: Computer-aided metadata generation for photo archives initiative, 2020. URL: https://doi.org/10.1184/R1/12791807.v1.

[213] Matthew Lincoln, Golan Levin, Sarah Reiff Conell, and Lingdong Huang. National neighbors: Distant viewing the national gallery of art's collection of collections. https://nga-neighbors.library.cmu.edu, 2019.

[214] Bing Liu, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Text classification by labeling words. In *Proceedings of the 19th National Conference on Artifical Intelligence*, AAAI'04, page 425–430. AAAI Press, 2004.

[215] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy, July 2019. Association for

Computational Linguistics. URL: https://www.aclweb.org/anthology/P19-1631, doi:10.18653/v1/P19-1631.

[216] Benedikt Loepp, Catalin-Mihai Barbu, and Jürgen Ziegler. Interactive Recommending: Framework, State of Research and Future Challenges. In *Proceedings of the 1st Workshop on Engineering Computer-Human Interaction in Recommender Systems (EnCHIReS '16)*, pages 3–13, 2016. URL: http://ceur-ws.org/Vol-1705/02-paper.pdf.

[217] Juliette Lonij and Melvin Wevers. SIAMESE | KB LAB. URL: https://lab.kb.nl/tool/siamese.

[218] Edward Loper and Steven Bird. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*, 2002.

[219] Elizabeth Lorang. Patterns, Collaboration, Practice: Algorithms as Editing for Historic Periodicals, 2018.

[220] Elizabeth Lorang and Leen-Kiat Soh. Application of the Image Analysis for Archival Discovery Team's First- Generation Methods and Software to the Burney Collection of British Newspapers, 2019.

[221] Elizabeth Lorang and Leen-Kiat Soh. Using Chronicling America's Images to Explore Digitized Historic Newspapers & Imagine Alternative Futures, 2019.

[222] Elizabeth Lorang, Leen-Kiat Soh, Maanas Varma Datla, and Spencer Kulwicki. Developing an image-based classifier for detecting poetic content in historic newspaper collections. *D-Lib Mag.*, 21, 2015.

[223] Elizabeth Lorang, Leen-Kiat Soh, Yi Liu, and Chulwoo Pack. Digital libraries, intelligent data analytics, and augmented description: A demonstration project, 2020. URL: https://labs.loc.gov/static/labs/work/reports/UNL-final-report.pdf.

[224] Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 150–158, New York, NY, USA, 2012. Association for Computing Machinery. doi:10.1145/2339530.2339556.

[225] Yin Lou, Rich Caruana, Johannes Gehrke, and Giles Hooker. Accurate intelligible models with pairwise interactions. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 623–631, New York, NY, USA, 2013. Association for Computing Machinery. doi:10.1145/2487575.2487579.

[226] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL: http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

[227] Michael Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *CHI Conference on Human Factors in Computing Systems*. ACM, March 2020. URL: https://www.microsoft.com/en-us/research/publication/co-designing-checklists-to-understand-organizational-challenges-and-opportunities-arou

[228] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, Honolulu HI USA, April 2020. ACM. URL: https://dl.acm.org/doi/10.1145/3313831.3376445, doi:10.1145/3313831.3376445.

[229] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. Exploring the Limits of

Weakly Supervised Pretraining. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 185–201, Cham, 2018. Springer International Publishing.

[230] Bonnie Mak. Archaeology of a digitization. *Journal of the Association for Information Science and Technology*, 65(8):1515–1526, 2014. URL: https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/asi.23061, arXiv:https://asistdl.onlinelibrary.wiley.com/doi/pdf/10.1002/asi.23061, doi:https://doi.org/10.1002/asi.23061.

[231] Lev Manovich. How to compare one million images? In David M. Berry, editor, *Understanding Digital Humanities*, pages 249–278, London, 2012. Palgrave Macmillan UK. doi:10.1057/9780230371934_14.

[232] Lev Manovich. Data science and art history. *International Journal for Digital Art History*, (1), June 2015. URL: https://doi.org/10.11588/dah.2015.1.21631.

[233] Gary Marchionini. Exploratory search: From finding to understanding. *Commun. ACM*, 49(4):41–46, April 2006. doi:10.1145/1121949.1121979.

[234] Marjo Markkula and Eero Sormunen. End-user searching challenges indexing practices in the digital newspaper photo archive. *Information retrieval*, 1:259–285, 1999.

[235] Monte Maxwell. Wvu today — wvrhc seeking copies of rare african-american newspapers, 2017. URL: https://wvutoday.wvu.edu/stories/2017/01/19/wvrhc-seeking-copies-of-rare-african-american-newspapers.

[236] John McCarthy. Programs with common sense, 1960.

[237] Barbara McGillivray, Paola Marongiu, Nilo Pedrazzini, Marton Ribary, Mandy Wigdorowitz, and Eleonora Zordan. Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences. *Publications*, 10(4), 2022. URL: https://www.mdpi.com/2304-6775/10/4/39, doi:10.3390/publications10040039.

[238] Makena Mezistrano. Why are these passover ads different from all other ads?, 2021. URL: https://jewishstudies.washington.edu/sephardic-studies/why-are-these-passover-ads-different-from-all-other-ads/.

[239] Ian Milligan. Illusionary order: Online databases, optical character recognition, and canadian history, 1997–2010. *Canadian Historical Review*, 94(4):540–569, 2013. arXiv:https://doi.org/10.3138/chr.694, doi:10.3138/chr.694.

[240] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model Cards for Model Reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 220–229, January 2019. arXiv: 1810.03993. URL: http://arxiv.org/abs/1810.03993, doi:10.1145/3287560.3287596.

[241] Jesse Mu and Jacob Andreas. Compositional explanations of neurons. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17153–17163. Curran Associates, Inc., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/c74956ffb38ba48ed6ce977af6727275-Paper.pdf.

[242] United States Holocaust Memorial Museum. History unfolded: Us newspapers and the holocaust. https://newspapers.ushmm.org/about/project, N/A.

[243] Bibliothèque nationale du Luxembourg. Bibliothèque nationale du luxembourg historical Newspapers, N/A. URL: https://data.bnl.lu/data/historical-newspapers/.

[244] John Nerone and Kevin G. Barnhurst. US Newspaper Types, the Newsroom, and the Division of Labor, 1750–2000. *Journalism Studies*, 4(4):435–449, November 2003. Publisher: Routledge _eprint: https://doi.org/10.1080/1461670032000136541. doi:10.1080/1461670032000136541.

[245] Clemens Neudecker. Cultural heritage as data: Digital curation and artificial intelligence in libraries.

[246] Newspapers.com. About newspapers.com, N/A. URL: http://www.newspapers.com/about/.

[247] T. Nguyen, A. Jatowt, M. Coustaty, N. Nguyen, and A. Doucet. Deep statistical analysis of ocr errors for effective post-ocr processing. In *JCDL '19*, pages 29–38, 2019. doi:10.1109/JCDL.2019.00015.

[248] Bob Nicholson. The digital turn: Exploring the methodological possibilities of digital newspaper archives. *Media History: Special Issue: Journalism and History: Dialogues*, 19(1):59–73, 2013. URL: http://www.tandfonline.com/doi/abs/10.1080/13688804.2012.752963.

[249] Xi Niu, Xiangyu Fan, and Tao Zhang. Understanding faceted search from data science and human factor perspectives. *ACM Trans. Inf. Syst.*, 37(2), January 2019. doi:10.1145/3284101.

[250] Safiya Umoja Noble. *Algorithms of oppression: how search engines reinforce racism.* NYU Press, 2018. OCLC: 987591529.

[251] André Reyes Novaes. *Maps in Newspapers: Approaches of Study and Practices in Portraying War since 19th Century.* Brill, Leiden, The Netherlands, 2019. URL: https://brill.com/view/title/54806, doi:https://doi.org/10.1163/9789004398832.

[252] John O'Donovan, Barry Smyth, Brynjar Gretarsson, Svetlin Bostandjiev, and Tobias Höllerer. Peerchooser: Visual interactive recommendation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1085–1088, New York, NY, USA, 2008. ACM. URL: http://doi.acm.org/10.1145/1357054.1357222, doi:10.1145/1357054.1357222.

[253] Library of Congress. Civil war maps (new-york daily tribune): Topics in chronicling america. URL: https://www.loc.gov/rr/news/topics/civilwarmaps.html.

[254] Library of Congress. The Appeal, 14 March 1891. Chronicling America: Historic American Newspapers, 1891. URL: https://chroniclingamerica.loc.gov/lccn/sn83016810/1891-03-14/ed-1/seq-1/.

[255] Library of Congress. The Appeal, 9 may 1891. Chronicling America: Historic American Newspapers, 1891. URL: https://chroniclingamerica.loc.gov/lccn/sn83016810/1891-05-09/ed-1/seq-1/.

[256] Library of Congress. The Washington Bee, 29 August 1896. Chronicling America: Historic American Newspapers, 1896. URL: https://chroniclingamerica.loc.gov/lccn/sn84025891/1896-08-29/ed-1/seq-1/.

[257] Library of Congress. The Washington Bee, 29 February 1896. Chronicling America: Historic American Newspapers, 1896. URL: https://chroniclingamerica.loc.gov/lccn/sn84025891/1896-02-29/ed-1/seq-1/.

[258] Library of Congress. The Colored American, 7 April 1900. Chronicling America: Historic American Newspapers, 1900. URL: https://chroniclingamerica.loc.gov/lccn/sn83027091/1900-04-07/ed-1/seq-1/.

[259] Library of Congress. The Colored American, 8 September 1896. Chronicling America: Historic American Newspapers, 1900. URL: https://chroniclingamerica.loc.gov/lccn/sn83027091/1900-09-08/ed-1/seq-1/.

[260] Library of Congress. The broad ax. [volume] (salt lake city, utah), 15 oct. 1910. chronicling america: Historic american newspapers, 1910. URL: https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-10-15/ed-1/seq-2/.

[261] Library of Congress. The broad ax. [volume] (salt lake city, utah), 26 nov. 1910. chronicling america: Historic american newspapers, 1910. URL: https://chroniclingamerica.loc.gov/lccn/sn84024055/1910-11-26/ed-1/seq-3/.

[262] Library of Congress. Franklin's paper the statesman. [volume] (denver, colo.), 15 oct. 1910. chronicling america: Historic american newspapers, 1910. URL: https://chroniclingamerica.loc.gov/lccn/sn91052311/1910-10-15/ed-1/seq-16/.

[263] Library of Congress. Image of w.e.b. du bois from franklin's paper the statesman (15 october 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america,

1910. URL: https://news-navigator.labs.loc.gov/data/cohi_abbeyville_ver01/data/sn91052311/00279550730/1910101501/2272/001_0_93.jpg.

[264] Library of Congress. Image of w.e.b. du bois from the broad ax (15 october 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910101501/0538/002_0_98.jpg.

[265] Library of Congress. Image of w.e.b. du bois from the broad ax (26 november 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910112601/0564/004_0_98.jpg.

[266] Library of Congress. Image of w.e.b. du bois from the iowa state bystander (14 october 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iahi_ames_ver01/data/sn83025186/00202198417/1910101401/1015/001_0_95.jpg.

[267] Library of Congress. Iowa state bystander. [volume] (des moines, iowa), 14 oct. 1910. chronicling america: Historic american newspapers, 1910. URL: https://chroniclingamerica.loc.gov/lccn/sn83025186/1910-10-14/ed-1/seq-1/.

[268] Library of Congress. Newspaper navigator metadata for franklin's paper the statesman (15 october 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/cohi_abbeyville_ver01/data/sn91052311/00279550730/1910101501/2272.json.

[269] Library of Congress. Newspaper navigator metadata for the broad ax (15 october 1910). from the library of congress, newspaper navigator

dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910101501/0538.json.

[270] Library of Congress. Newspaper navigator metadata for the broad ax (26 november 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iune_charlie_ver01/data/sn84024055/00280761059/1910112601/0564.json.

[271] Library of Congress. Newspaper navigator metadata for the iowa state bystander (14 october 1910). from the library of congress, newspaper navigator dataset: Extracted visual content from chronicling america, 1910. URL: https://news-navigator.labs.loc.gov/data/iahi_ames_ver01/data/sn83025186/00202198417/1910101401/1015.json.

[272] Library of Congress. La Prensa, 22 may 1917. Chronicling America: Historic American Newspapers, 1917. URL: https://chroniclingamerica.loc.gov/lccn/sn83045395/1917-05-22/ed-1/seq-1/.

[273] Library of Congress. About CONSER - Program for Cooperative Cataloging (Library of Congress), N/A. URL: https://www.loc.gov/aba/pcc/conser/about/.

[274] Library of Conress. About by the people. https://crowd.loc.gov/about/, N/A.

[275] National Library of New Zealand. Papers past. https://natlib.govt.nz/collections/a-z/papers-past, N/A.

[276] NEH Division of Preservation and Access 2020. Notice of funding opportunity, national digital newspaper program, 2020. URL: https://www.neh.gov/sites/default/files/inline-files/National-Digital-Newspaper-Program-NOFO-January-2020.pdf.

[277] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. Critical race theory for hci. In *Proceedings of the 2020 CHI Conference on*

Human Factors in Computing Systems, CHI '20, page 1–16, New York, NY, USA, 2020. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/3313831.3376392, doi:10.1145/3313831.3376392.

[278] Hope A. Olson. The power to name: Representation in library catalogs. *Signs*, 26(3):pp. 639–668, 2001. URL: http://www.jstor.org/stable/3175535.

[279] Joshua Ortiz Baco. Title Essays, Linked Data, and the Ethnic Press in Chronicling America, 2019. URL: https://www.neh.gov/blog/title-essays-linked-data-and-ethnic-press-chronicling-america.

[280] OurDigitalWorld. Ourdigitalworld: Digital newspapers, N/A. URL: https://ourdigitalworld.net/what-we-do/digital-newspapers/.

[281] Trevor Owens. *The Theory and Craft of Digital Preservation*. Johns Hopins University Press, Baltimore, 2018.

[282] Trevor Owens and Thomas Padilla. Digital sources and digital archives: historical evidence in the digital age. *International Journal of Digital Humanities*, 1:325–341, 2020.

[283] Thomas Padilla. On a collections as data imperative, 2018. URL: https://digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf?loclr=blogadm.

[284] Thomas Padilla. Responsible Operations: Data Science, Machine Learning, and AI in Libraries, August 2020. Journal Abbreviation: Responsible Operations: Data Science, Machine Learning, and AI in Libraries Last Modified: 2020-5-12 Publisher: OCLC. URL: https://www.oclc.org/research/publications/2019/oclcresearch-responsible-operations-data-science-machine-learning-ai.html, doi:10.25333/xk7z-9g97.

[285] Denis Parra and Peter Brusilovsky. User-controllable personalization: A case study with SetFusion. *International Journal of Human-Computer Studies*,

78:43–67, June 2015. URL: https://linkinghub.elsevier.com/retrieve/pii/S1071581915000208, doi:10.1016/j.ijhcs.2015.01.007.

[286] Adam Paszke, Sam Gross, Francisco Massa, and et al. Pytorch: An imperative style, high-performance deep learning library. In *NIPS '19*, pages 8024–8035. 2019. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[287] Katelin D Pearson, Gil Nelson, Myla F J Aronson, and et al. Machine Learning Using Digitized Herbarium Specimens to Advance Phenological Research. *BioScience*, 2020. doi:10.1093/biosci/biaa044.

[288] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[289] Aleš Pekárek and Marieke Willems. The Europeana Newspapers: A Gateway to European Newspapers Online. In *Progress in Cultural Heritage Preservation*, pages 654–659, Berlin, 2012. Springer.

[290] S. Petridis, H. Shin, and L. Chilton. Symbolfinder: Brainstorming diverse symbols using semantic networks. In *preprint*, 2020.

[291] H. Pflüger and T. Ertl. Sifting through visual arts collections. *Comput. Graph.*, 57(C):127–138, June 2016. doi:10.1016/j.cag.2016.03.009.

[292] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily B. Fox, and H. Larochelle. Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program). *ArXiv*, abs/2003.12206, 2021.

[293] Andrew Piper, Chad Wellmon, and Mohamed Cheriet. The page image: Towards a

visual history of digital documents. *Book History*, 23(1):365–397, June 2015. URL: https://doi.org/10.1353/bh.2020.0010.

[294] B. A. Plummer, M. H. Kiapour, S. Zheng, and R. Piramuthu. Give me a hint! navigating image databases using human-in-the-loop feedback. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2048–2057, 2019. doi:10.1109/WACV.2019.00222.

[295] Todd Presner. The ethics of the algorithm: Close and distant listening to the shoah foundation visual history archive. In Fogu Claudio, Wolf Kansteiner, and Todd Presner, editors, *Probing the Ethics of Holocaust Culture*, pages 175–202. Harvard University Press, 1999.

[296] National Digital Newspaper Program. About the program - national digital newspaper program (library of congress), 2019. URL: https://www.loc.gov/ndnp/about.html.

[297] National Digital Newspaper Program. Chronicling America | Library of Congress, 2020. URL: https://chroniclingamerica.loc.gov/about/api/.

[298] National Digital Newspaper Program. Chronicling America guidelines & resources, 2020. URL: http://www.loc.gov/ndnp/guidelines/.

[299] National Digital Newspaper Program. Content selection - national digital newspaper program (library of congress), 2020. URL: https://www.loc.gov/ndnp/guidelines/selection.html.

[300] National Digital Newspaper Program. The national digital newspaper program (ndnp) technical guidelines for applicants, 2020. URL: https://www.loc.gov/ndnp/guidelines/.

[301] National Digital Newspaper Program. About chronicling america, N/A. URL: https://chroniclingamerica.loc.gov/about/.

[302] National Digital Newspaper Program and the University of Kentucky Libraries.

Meta — morphosis: Tutorials. URL: https://www.uky.edu/Libraries/NDNP/metamorphosis/tutorials.html.

[303] Impresso Project. Impresso project, 2017. URL: impresso-project.ch/project/overview/.

[304] Project Team. The santa barbara statement on collections as data, 2017. URL: https://collectionsasdata.github.io/statement/.

[305] Pearl Pu and Li Chen. Trust building with explanation interfaces. In *Proceedings of the 11th International Conference on Intelligent User Interfaces*, IUI '06, pages 93–100, New York, NY, USA, 2006. ACM. URL: http://doi.acm.org/10.1145/1111449.1111475, doi:10.1145/1111449.1111475.

[306] Pearl Pu, Li Chen, and Rong Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, pages 157–164, New York, NY, USA, 2011. ACM. URL: http://doi.acm.org/10.1145/2043932.2043962, doi:10.1145/2043932.2043962.

[307] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. Data cards: Purposeful and transparent dataset documentation for responsible ai. *ArXiv*, abs/2204.01075, 2022.

[308] Till Quack, Ullrich Mönich, Lars Thiele, and B. S. Manjunath. Cortina: A system for large-scale, content-based web image retrieval. In *Proceedings of the 12th Annual ACM International Conference on Multimedia*, MULTIMEDIA '04, page 508–511, New York, NY, USA, 2004. Association for Computing Machinery. doi:10.1145/1027527.1027650.

[309] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL: https://www.R-project.org/.

[310] Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. Exploring the role of local and global explanations in recommender systems. In *Ex-*

tended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems, CHI EA '22, New York, NY, USA, 2022. Association for Computing Machinery. doi:10.1145/3491101.3519795.

[311] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL: https://proceedings.mlr.press/v139/radford21a.html.

[312] Hema Raghavan and James Allan. An interactive algorithm for asking and incorporating feature feedback into support vector machines. In *SIGIR '07*, page 79–86, 2007. URL: https://dl.acm.org/doi/10.1145/1277741.1277758.

[313] Juha Rautiainen. Opening digitized newspapers for different user groups - successes and challenges. In *IFLA WLIC 2019*, Athens, Greece, 2019.

[314] READ COOP. Transkribus: Transcribe. collaborate. share., 2020. URL: https://readcoop.eu/transkribus/.

[315] READ COOP. Public models in transkribus, 2022. URL: https://readcoop.eu/transkribus/public-models/.

[316] Matthew Reidsma. *Masked by Trust: Bias in Library Discovery.* Litwin Books, Sacramento, 2019.

[317] Dillon Reisman, Jason Schultz, Kate Crawford, and Meredith Whittaker. Algorithmic impact assessments: A practical framework for public agency accountability, 2018. URL: https://ainowinstitute.org/aiareport2018.pdf.

[318] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In

NIPS '15, pages 91–99, 2015. URL: http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf.

[319] Jane Rhodes. *Mary Ann Shadd Cary: The Black Press and Protest in the Nineteenth Century*. Indiana University Press, Bloomington, IN, November 1998.

[320] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM. URL: http://doi.acm.org/10.1145/2939672.2939778, doi:10.1145/2939672.2939778.

[321] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.442, doi:10.18653/v1/2020.acl-main.442.

[322] Mia Ridge, Samantha Blickhan, Meghan Ferriter, Austin Mast, Ben Brumfield, Brendon Wilkins, Daria Cybulska, Denise Burgher, Jim Casey, Kurt Luther, Michael Haley Goldman, Nick White, Pop Willcox, Sara Carlstead Brumfield, Sonya J. Coleman, and Ylva Berglund Prytz. *The Collective Wisdom Handbook: Perspectives on Crowdsourcing in Cultural Heritage*. 2021. URL: https://doi.org/10.21428/a5d7554f.1b80974b.

[323] Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge, 2019. arXiv:1909.13584.

[324] Roopika Risam. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy*. Northwestern University Press, Evanston, Illinois, 2018.

[325] Gregory Rolan, Glen Humphries, Lisa Jeffrey, Evanthia Samaras, Tatiana Antsoupova, and Katharine Stuart. More human than human? artificial intelligence in the archive. *Archives and Manuscripts*, 47(2):179–203, 2019. arXiv:https://doi.org/10.1080/01576895.2018.1502088, doi:10.1080/01576895.2018.1502088.

[326] Stephanie L. Rosenthal and Anind K. Dey. Towards maximizing the accuracy of human-labeled sensor data. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*, IUI '10, page 259–268, New York, NY, USA, 2010. Association for Computing Machinery. doi:10.1145/1719970.1720006.

[327] Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, page 2662–2670. AAAI Press, 2017.

[328] Christian Safka. img2vec, 2019. URL: https://github.com/christiansafka/img2vec.

[329] Pınar Sarın and Necla Uluğtekin. Analyzing newspaper maps for earthquake news through cartographic approach. *ISPRS International Journal of Geo-Information*, 8(5), 2019. URL: https://www.mdpi.com/2220-9964/8/5/235, doi:10.3390/ijgi8050235.

[330] James Schaffer, Tobias Höllerer, and John O'Donovan. Hypothetical recommendation: A study of interactive profile manipulation behavior for recommender systems, 2015. URL: https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS15/paper/view/10444.

[331] Robert E. Schapire, Marie Rochery, Mazin G. Rahim, and Narendra Kumar Gupta. Boosting with prior knowledge for call classification. *IEEE Transactions on Speech and Audio Processing*, 13:174–181, 2005.

[332] Patrick Schramowski, Wolfgang Stammer, Stefano Teso, Anna Brugger, Xiaoting Shao, Hans-Georg Luigs, Anne-Katrin Mahlein, and Kristian Kersting. Right for

the wrong scientific reasons: Revising deep networks by interacting with their explanations, 2020. arXiv:2001.05371.

[333] Eric Schuettpelz, Paul B. Frandsen, Rebecca B. Dikow, Abel Brown, Sylvia Orli, Melinda Peters, Adam Metallo, Vicki A. Funk, and Laurence J. Dorr. Applications of deep convolutional neural networks to digitized natural history collections. *Biodiversity Data Journal*, 5, 2017. doi:10.3897/BDJ.5.e21139.

[334] Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. Green ai. *Commun. ACM*, 63(12):54–63, nov 2020. doi:10.1145/3381831.

[335] Ian A. Scott, Stacy M. Carter, and Enrico W. Coiera. Clinician checklist for assessing suitability of machine learning applications in healthcare. *BMJ Health & Care Informatics*, 28, 2021.

[336] Christin Seifert, Stefanie Scherzinger, and Lena Wiese. Towards generating consumer labels for machine learning models. *2019 IEEE First International Conference on Cognitive Machine Intelligence (CogMI)*, pages 173–179, 2019.

[337] Burr Settles. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2012. doi:10.2200/S00429ED1V01Y201207AIM018.

[338] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. In *Document Analysis and Recognition – ICDAR 2021: 16th International Conference, Lausanne, Switzerland, September 5–10, 2021, Proceedings, Part I*, page 131–146, Berlin, Heidelberg, 2021. Springer-Verlag. doi:10.1007/978-3-030-86549-8_9.

[339] Tim Sherratt and Kate Bagnall. The people inside, January 2019. doi:10.5281/zenodo.3579530.

[340] Ben Shneiderman. Opinion: The dangers of faulty, biased, or malicious algorithms requires independent oversight. *Proceedings of the National Academy of*

Sciences, 113(48):13538–13540, 2016. Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/113/48/13538.full.pdf. URL: https://www.pnas.org/content/113/48/13538, doi:10.1073/pnas.1618211113.

[341] Ben Shneiderman, Benjamin B. Bederson, and Steven M. Drucker. Find that photo! interface strategies to annotate, browse, and share. *Commun. ACM*, 49(4):69–71, April 2006. doi:10.1145/1121949.1121985.

[342] Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *arXiv:1707.06742 [cs, stat]*, July 2017. arXiv: 1707.06742. URL: http://arxiv.org/abs/1707.06742.

[343] Foteini Simistira, Mathias Seuret, Nicole Eichenberger, Angelika Garz, Marcus Liwicki, and Rolf Ingold. DIVA-HisDB: A Precisely Annotated Large Dataset of Challenging Medieval Manuscripts. In *ICFHR '16*, pages 471–476, Shenzhen, China, 2016. IEEE. URL: http://ieeexplore.ieee.org/document/7814109/, doi:10.1109/ICFHR.2016.0093.

[344] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR '15*, 2015.

[345] Robert J. Simpson, Kevin R. Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *WWW '14 Companion*, 2014.

[346] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. An overview of microsoft academic service (mas) and applications. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 243–246, New York, NY, USA, 2015. ACM. URL: http://doi.acm.org/10.1145/2740908.2742839, doi:10.1145/2740908.2742839.

[347] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. No explainability without accountability:

An empirical study of explanations and feedback in interactive ml. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3313831.3376624.

[348] T. Smits and W.J. Faber. Chronic (classified historical newspaper images), 2018. URL: http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images.

[349] Leen-Kiat Soh, Elizabeth Lorang, and Yi Liu. Aida: Intelligent image analysis to automatically detect poems in digital archives of historic newspapers. In *IAAI '18*, 2018. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16880.

[350] Eero Sormunen, Marjo Markkula, and Kalervo Järvelin. The perceived similarity of photos: A test-collection based evaluation framework for the content-based image retrieval algorithms. In *Proceedings of the 1999 International Conference on Final Mira*, MIRA'99, page 8, Swindon, GBR, 1999. BCS Learning & Development Ltd.

[351] Karen Sparck Jones. *A Statistical Interpretation of Term Specificity and Its Application in Retrieval*, page 132–142. Taylor Graham Publishing, GBR, 1988.

[352] Derrick Ramon Spires. *The Practice of Citizenship: Black Politics and Print Culture in the Early United States*. University of Pennsylvania Press, Philadelphia, 2019.

[353] Sarah Abrevaya Stein. *Making Jews Modern: The Yiddish and Ladino Press in the Russian and Ottoman Empires*. Indiana University Press, 2004.

[354] Robin Stewart, Gregory Scott, and Vladimir Zelevinsky. Idea navigation: Structured browsing for unstructured text. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, page 1789–1792, New York, NY, USA, 2008. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/1357054.1357332, doi:10.1145/1357054.1357332.

[355] Emilia Stoica and Marti Hearst. Demonstration: Using wordnet to build hierarchical

facet categories. In *Proceedings of the International ACM SIGIR Workshop on Faceted Search*. ACM Press, 2006.

[356] Emilia Stoica, Marti Hearst, and Megan Richardson. Automating creation of hierarchical faceted metadata structures. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 244–251, Rochester, New York, April 2007. Association for Computational Linguistics. URL: https://aclanthology.org/N07-1031.

[357] Carolyn Strange, Daniel McNamara, Joshua Wodak, and Ian Wood. Mining for the meanings of a murder: The impact of ocr quality on the use of digitized historical newspapers. *Digital Humanities Quarterly*, 8:1–17, 2014.

[358] Grant Strong, Orland Hoeber, and Minglun Gong. Visual image browsing and exploration (vibe): User evaluations of image search tasks. In Aijun An, Pawan Lingras, Sheila Petty, and Runhe Huang, editors, *Active Media Technology*, pages 424–435, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[359] Grant Strong, Enamul Hoque, Minglun Gong, and Orland Hoeber. Organizing and browsing image search results based on conceptual and visual similarities. In George Bebis, Richard Boyle, Bahram Parvin, Darko Koracin, Ronald Chung, Riad Hammound, Muhammad Hussain, Tan Kar-Han, Roger Crawfis, Daniel Thalmann, David Kao, and Lisa Avila, editors, *Advances in Visual Computing*, pages 481–490, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[360] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, July 2019. Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/P19-1355, doi:10.18653/v1/P19-1355.

[361] Pranav Subramaniam, Yintong Ma, Chi Li, Ipsita Mohanty, and Raul Castro Fernan-

dez. Comprehensive and comprehensible data catalogs: The what, who, where, when, why, and how of metadata management. *ArXiv*, abs/2103.07532, 2021.

[362] Yuyin Sun, Adish Singla, Dieter Fox, and Andreas Krause. Building hierarchies of concepts via crowdsourcing. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, page 844–851. AAAI Press, 2015.

[363] Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: an introduction*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018.

[364] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR '16*, pages 2818–2826, 2016.

[365] Kunihiro Takeoka, Kosuke Akimoto, and Masafumi Oyamada. Low-resource taxonomy enrichment with pretrained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2747–2758, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL: https://aclanthology.org/2021.emnlp-main.217, doi:10.18653/v1/2021.emnlp-main.217.

[366] DiJeSt Team. Our text recognition ground truth and model – dijest. URL: https://dijest.net/gtmodel/.

[367] Chris Tensmeyer, Brian Davis, Curtis Wigington, Iain Lee, and Bill Barrett. Pagenet: Page boundary extraction in historical handwritten documents. In *Proceedings of the 4th Intl. Workshop on Historical Document Imaging & Processing*, pages 59–64, New York, USA, 2017. ACM. URL: http://doi.acm.org/10.1145/3151509.3151522, doi:10.1145/3151509.3151522.

[368] Stefano Teso, Öznur Alkan, Wolfang Stammer, and Elizabeth Daly. Leveraging explanations in interactive machine learning: An overview, 2022. URL: https://arxiv.org/abs/2207.14526, doi:10.48550/ARXIV.2207.14526.

[369] The Library of Congress. American english dialect recordings: The center for applied linguistics collection. about this collection. URL: https://www.loc.gov/collections/american-english-dialect-recordings-from-the-center-for-applied-linguistics/about-this-collection/.

[370] Nava Tintarev and Judith Masthoff. A survey of explanations in recommender systems. In *2007 IEEE 23rd International Conference on Data Engineering Workshop*, pages 801–810, 2007. doi:10.1109/ICDEW.2007.4401070.

[371] Nava Tintarev and Judith Masthoff. Designing and Evaluating Explanations for Recommender Systems. In Francesco Ricci, Lior Rokach, Bracha Shapira, and Paul B. Kantor, editors, *Recommender Systems Handbook*, pages 479–510. Springer US, Boston, MA, 2011. doi:10.1007/978-0-387-85820-3_15.

[372] Myriam C. Traub, Jacco van Ossenbruggen, and Lynda Hardman. Impact analysis of ocr quality on research tasks in digital archives. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Research and Advanced Technology for Digital Libraries*, pages 252–263, Cham, 2015. Springer International Publishing.

[373] Chun-Hua Tsai and Peter Brusilovsky. Beyond the ranked list: User-driven exploration and diversification of social recommendation. In *23rd International Conference on Intelligent User Interfaces*, IUI '18, pages 239–250, New York, NY, USA, 2018. ACM. URL: http://doi.acm.org/10.1145/3172944.3172959, doi:10.1145/3172944.3172959.

[374] Chun-Hua Tsai and Peter Brusilovsky. Explaining recommendations in an interactive hybrid social recommender. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, pages 391–396, New York, NY, USA, 2019. ACM. URL: http://doi.acm.org/10.1145/3301275.3302318, doi:10.1145/3301275.3302318.

[375] Chun-Hua Tsai and Peter Brusilovsky. The effects of controllability and explainability

in a social recommender system. *User Modeling and User-adapted Interaction*, pages 1–37, 2020.

[376] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008. URL: http://www.jmlr.org/papers/v9/vandermaaten08a.html.

[377] Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, page 341–350, New York, NY, USA, 2009. Association for Computing Machinery. doi:10.1145/1526709.1526756.

[378] Daniel van Strien, Mark Bell, Nora Rose McGregor, and Michael Trizna. An introduction to ai for glam. In Katherine M. Kinnaird, Peter Steinbach, and Oliver Guhr, editors, *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop*, volume 170 of *Proceedings of Machine Learning Research*, pages 20–24. PMLR, 08–13 Sep 2022. URL: https://proceedings.mlr.press/v170/strien22a.html.

[379] Roelof van Zwol, Börkur Sigurbjornsson, Ramu Adapala, Lluis Garcia Pueyo, Abhinav Katiyar, Kaushal Kurapati, Mridul Muralidharan, Sudar Muthu, Vanessa Murdock, Polly Ng, Anand Ramani, Anuj Sahai, Sriram Thiru Sathish, Hari Vasudev, and Upendra Vuyyuru. Faceted exploration of image search results. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 961–970, New York, NY, USA, 2010. Association for Computing Machinery. URL: https://doi-org.offcampus.lib.washington.edu/10.1145/1772690.1772788, doi:10.1145/1772690.1772788.

[380] Olivia Vane. Visualising the royal photographic society collection: Part 2. https://www.vam.ac.uk/blog/digital/visualising-the-royal-photographic-society-collection-part-2, 2018.

[381] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N

Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[382] Gabriela Baeza Ventura, Lorena Gauthereau, and Carolina Villarroel. Recovering the US Hispanic Literary Heritage: A Case Study on US Latina/o Archives and Digital Humanities. *Preservation, Digital Technology & Culture (PDT&C)*, 48(1):17–27, April 2019. URL: https://www.degruyter.com/document/doi/10.1515/pdtc-2018-0031/html, doi:10.1515/pdtc-2018-0031.

[383] Katrien Verbert, Denis Parra, Peter Brusilovsky, and Erik Duval. Visualizing recommendations to support exploration, transparency and controllability. In *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, IUI '13, pages 351–362, New York, NY, USA, 2013. ACM. URL: http://doi.acm.org/10.1145/2449396.2449442, doi:10.1145/2449396.2449442.

[384] Jesse Vig, Shilad Sen, and John Riedl. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Transactions on Interactive Intelligent Systems*, 2(3):1–44, September 2012. URL: http://dl.acm.org/citation.cfm?doid=2362394.2362395, doi:10.1145/2362394.2362395.

[385] Lorella Viola. Review: Newspaper Navigator. *Reviews in Digital Humanities*, III(6), jun 21 2022. https://reviewsindh.pubpub.org/pub/newspaper-navigator.

[386] Annika Wærn. User Involvement in Automatic Filtering: An Experimental Study. *User Modeling and User-Adapted Interaction*, 14(2):201–237, June 2004. doi:10.1023/B:USER.0000028984.13876.9b.

[387] Eric Wallace, Pedro Rodriguez, Shi Feng, Ikuya Yamada, and Jordan Boyd-Graber. Trick me if you can: Human-in-the-loop generation of adversarial examples for question answering. *Transactions of the Association for Computational Linguistics*,

7:387–401, 2019. URL: https://aclanthology.org/Q19-1029, doi:10.1162/tacl_a_00279.

[388] J. Z. Wang, Jia Li, and G. Wiederhold. Simplicity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001. doi:10.1109/34.955109.

[389] Zijie Jay Wang, Alex Kale, Harsha Nori, Peter Stella, Mark E. Nunnally, Duen Horng Chau, Mihaela Vorvoreanu, Jennifer Wortman Vaughan, and Rich Caruana. Gam changer: Editing generalized additive models with interactive visualization. *ArXiv*, abs/2112.03245, 2021.

[390] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 2016. URL: http://distill.pub/2016/misread-tsne, doi:10.23915/distill.00002.

[391] Bifan Wei, Jun Liu, Qinghua Zheng, Wei Zhang, Xiaoyu Fu, and Boqin Feng. A survey of faceted search. *J. Web Eng.*, 12(1–2):41–64, February 2013.

[392] Daniel S. Weld and Gagan Bansal. The challenge of crafting intelligible intelligence. *Commun. ACM*, 62(6):70–79, may 2019. doi:10.1145/3282486.

[393] M. Wevers and J. Lonij. Siameset, 2017. URL: http://lab.kb.nl/dataset/siameset.

[394] Melvin Wevers and Thomas Smits. The visual digital turn: Using neural networks to study historical images. *Digital Scholarship in the Humanities*, 35(1):194–207, 2019. doi:10.1093/llc/fqy085.

[395] Ryen W. White, Bill Kules, Steven M. Drucker, and m.c. schraefel. Introduction. *Commun. ACM*, 49(4):36–39, April 2006. doi:10.1145/1121949.1121978.

[396] Marieke Willems and Rossitza Atanassova. Europeana Newspapers: searching digitized historical newspapers from 23 European countries. *Insights*, 28:51–56,

2015. URL: http://insights.uksg.org/articles/10.1629/uksg.218/, doi:10.1629/uksg.218.

[397] Andreá N. Williams. Cultivating Black Visuality: The Controversy over Cartoons in the Indianapolis Freeman. *American Periodicals: A Journal of History & Criticism*, 25(2):124–138, 2015. URL: http://muse.jhu.edu/article/601461, doi:10.1353/amp.2015.0032.

[398] L. Williams. What computational archival science can learn from art history and material culture studies. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3153–3155, 2019. doi:10.1109/BigData47090.2019.9006527.

[399] F. Windhager, P. Federico, G. Schreder, K. Glinka, M. Dörk, S. Miksch, and E. Mayr. Visualization of cultural heritage collection data: State of the art and future challenges. *IEEE Transactions on Visualization and Computer Graphics*, 25(6):2311–2330, 2019. doi:10.1109/TVCG.2018.2830759.

[400] Betty H. Winfield and Doyle Yoon. Historical images at a glance: North korea in american editorial cartoons. *Newspaper Research Journal*, 23(4):97–100, 2002. arXiv:https://doi.org/10.1177/073953290202300411, doi:10.1177/073953290202300411.

[401] Living with Machines Collaboration. Living with machines: About us. https://livingwithmachines.ac.uk/about/, N/A.

[402] Rebecca K Wright. Typewriting Mass Observation Online: Media Imprints on the Digital Archive. *History Workshop Journal*, 87:118–138, 02 2019. arXiv:https://academic.oup.com/hwj/article-pdf/doi/10.1093/hwj/dbz005/28296183/dbz005.pdf, doi:10.1093/hwj/dbz005.

[403] Tongshuang Wu, Daniel S. Weld, and Jeffrey Heer. Local decision pitfalls in interactive machine learning: An investigation into feature selection in sentiment analysis. *ACM Trans. Comput.-Hum. Interact.*, 26(4), jun 2019. doi:10.1145/3319616.

[404] Xiaoyun Wu and Rohini Srihari. Incorporating prior knowledge with weighted margin support vector machines. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, page 326–333, New York, NY, USA, 2004. Association for Computing Machinery. doi:10.1145/1014052.1014089.

[405] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

[406] Joel Wurl. Ethnicity as Provenance: In Search of Values and Principles for Documenting the Immigrant Experience. *Archival Issues*, 29(1):65–76, 2005. Publisher: Midwest Archives Conference. URL: http://www.jstor.org/stable/41102095.

[407] Elizabeth Yakel. Archival representation. *Archival Science*, 3(1):1–25, March 2003. doi:10.1007/BF02438926.

[408] Ka-Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. Faceted metadata for image search and browsing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, page 401–408, New York, NY, USA, 2003. Association for Computing Machinery. doi:10.1145/642611.642681.

[409] Marcia Lei Zeng, Chris Alen Sula, Karen F. Gracy, Eero Hyvönen, and Vania Mara Alves Lima. Jasist special issue on digital humanities (dh). *Journal of the Association for Information Science and Technology*, 73, 2022.

[410] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Trans. Multimedia Comput. Commun. Appl.*, 6(3), August 2010. doi:10.1145/1823746.1823747.

[411] Junliang Zhang and Gary Marchionini. Evaluation and evolution of a browse and search interface: Relation browser++. In *Proceedings of the 2005 National Conference on Digital Government Research*, dg.o '05, page 179–188. Digital Government Society of North America, 2005.

[412] Yongfeng Zhang and Xu Chen. Explainable Recommendation: A Survey and New Perspectives. *arXiv:1804.11192 [cs]*, April 2018. arXiv: 1804.11192. URL: http://arxiv.org/abs/1804.11192.

[413] Xiang Zhou and Thomas Huang. Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.*, 8:536–544, 04 2003. doi:10.1007/s00530-002-0070-3.