Designing Visual Decision Support Systems to Foster Trust and Support Expertise Diversity through Human-Centered Data Science

Andrea Figueroa

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

Doctor of Philosophy

University of Washington

2026

Reading Committee:

Cecilia Aragon, Chair

Mark Haselkorn

Brock Craft

Program Authorized to Offer Degree:

Human Centered Design and Engineering

University of Washington

Abstract

Designing Visual Decision Support Systems to Foster Trust and Support Expertise Diversity through Human-Centered Data Science

Andrea Figueroa

Chair of the Supervisory Committee:
Cecilia Aragon
Human Centered Design and Engineering

Decision support systems (DSSs) increasingly shape how people interpret complex data and make consequential decisions. Although these systems are often evaluated through technical performance, usability, or efficiency, their effectiveness also depends on whether users can make sense of system outputs, judge their reliability, and engage with tools in ways that reflect their knowledge, roles, and goals. In visual DSSs, design choices about how data is shown directly shape how users interpret system outputs and decide what to do with them.

This dissertation examines how visual DSSs can be designed and evaluated to foster trust and support expertise diversity. Grounded in Human-Centered Data Science and Munzner's nested model of visualization design, it develops the Dual-Lens Evaluation Framework, which analyzes visual DSSs through two independent perspectives: what the system assumes about its users, and what users independently bring to the interaction. The gap between these two perspectives is treated as the site where trust holds or breaks down depending on whether expertise diversity is supported.

The framework is developed and refined across three studies in distinct decision-making domains. The first, set in a traffic incident coordination platform, establishes the value of the user-side perspective and points to visualization as a promising direction for making system reasoning visible to users. The second, set in a public-facing transportation planning platform, examines how visualization features shape understanding, confidence, and engagement across users with varied expertise, and contributes the first refinement of the framework. The third, set in a qualitative coding environment, applies the full framework and surfaces tensions between system assumptions and user perspectives that would otherwise remain invisible, including a field-level divide in how the domain problem itself was understood.

This dissertation contributes the Dual-Lens Evaluation Framework for visual DSSs, empirical insights from three domains into how users with varied expertise interpret, trust, and engage with visual decision tools, and four design priorities for visual DSSs that account for the expertise their users bring: system transparency, communication, layered engagement, and reflexive adaptability. The work also contributes a refinement that moves expertise diversity beyond a novice-to-expert competence range toward the different kinds of expertise users bring through their particular relationship to the domain problem.

Table of Contents

Chapter 1: Introduction

1.1 Motivation and Problem Statement

Decision-making is a fundamental activity in both personal and professional contexts, from choosing a commuting route to allocating public resources during a crisis. As our access to data increases, the decisions we make are increasingly influenced, or directly supported, by data-driven tools. Decision Support Systems (DSSs) have become essential for navigating this complexity. These computer-based interactive systems assist decision-makers by presenting relevant data and insights rather than replacing their judgment (Eom & Kim, 2006). The role of DSSs has been researched and implemented across multiple domains, including business, policy, healthcare, and more, showcasing their versatility and potential impact.

While many DSSs are designed to provide accurate and efficient outputs, existing approaches tend to prioritize technical performance and usability for defined use cases, often without explicitly addressing the communication of system reasoning or the accommodation of diverse user needs. Even systems that are technically sound may fail when users do not trust them or when key perspectives are excluded from the design process. These design tendencies can unintentionally limit engagement, particularly for users who hold different roles, levels of expertise, or decision-making responsibilities.

In particular, designing DSSs that cater to the diverse needs of both novices and experts remains challenging. Novices may struggle with cognitive overload when engaging with complex visualizations, while experts often require advanced tools that allow for in-depth exploration and analysis. Balancing simplicity for novices and depth for advanced users is a persistent challenge for DSSs designers.

How a DSS presents information matters as much as what it presents. Central to the success of these systems is their ability to transform vast and intricate datasets into actionable insights. Data visualization offers a powerful means to accomplish this, and many DSSs rely on it. Data visualization aligns with the human cognitive and perceptual strengths, providing intuitive and digestible formats that support the identification of patterns, trends and anomalies (Munzner, 2014). Because data visualization mediates how users perceive and interpret system outputs, it is also the layer where trust is built or broken, and where design choices most directly shape whether a system is accessible across expertise levels.

However, DSS design and data visualization, while providing the structure and mechanism for effective decision support, do not on their own ensure that systems serve users across expertise levels or that users trust what the system presents. To address this, I draw from Human-Centered Data Science (HCDS), which applies human-centered design methods to the entire data lifecycle, from collection and analysis to interpretation and use. HCDS emphasizes the inclusion of stakeholder perspectives, contextual awareness, reflection, and the use of qualitative and mixed methods to ensure data-driven systems serve human needs and values. Together, these three areas provide the foundation for this work: DSSs define the decision-making context, visualization provides the representational mechanism, and HCDS offers the evaluative lens for centering user trust and expertise diversity.

Drawing on these three areas, this dissertation investigates how visualizations within DSSs can be designed to foster trust and support expertise diversity. This work centers on two constructs: trust, understood as the user's confidence in what the system is doing and why, and expertise diversity, the system's capacity to serve users with varying roles, knowledge backgrounds, and goals. Together, they reframe the evaluation of visual DSSs around the people who use them, beyond metrics like accuracy, efficiency, and time to decision, enabling users not only to

interpret outputs confidently but also to see their perspectives and needs reflected in the design of the system.

1.2 Research Goals and Questions

This dissertation explores three core goals:

1. Investigate how design choices in visual DSSs shape trust and support expertise diversity across distinct decision-making domains and user populations.
2. Identify design priorities for visual DSSs that support trust and expertise diversity across users with different roles, knowledge backgrounds, and goals.
3. Develop a human-centered evaluation framework that extends the evaluation of visual DSSs beyond usability to surface whether systems support trust and expertise diversity.

To achieve these goals, this dissertation addresses the following research questions:

1. How do the design choices embedded in visual DSSs support or undermine trust and expertise diversity?
2. What design priorities emerge from evaluating visual DSSs through the lens of trust and expertise diversity?
3. How can human-centered evaluation of visual DSSs be extended beyond usability to assess whether systems support trust and expertise diversity?

1.3 Dissertation Overview and Contributions

This dissertation consists of three interconnected studies. The first surfaces the problem within a specific operational context; the second and third develop and refine the framework across distinct decision-making domains to assess its applicability beyond the initial setting:

Study 1 examines the use of the Virtual Coordination Center (VCC), a collaborative platform used by regional transportation agencies to detect and manage traffic incidents. By analyzing how operators interpret data and system-generated alerts, this study surfaces key challenges around system transparency and the loss of expert knowledge in decision processes, identifying visualization as a direction for making system logic transparent and preserving contextual expertise across operator roles and experience.

Study 2 builds on these findings by evaluating how expert and non-expert users interact with scenario-based visualizations in the Transportation 2050 (T2050) platform. T2050 presents interactive visualizations of long-term transportation scenarios for the Puget Sound region to inform about transportation challenges. This study explores which design features enhance users' understanding, trust, and support for expertise diversity, and initiates the development of a human-centered evaluation framework aligned with trust and expertise diversity.

Study 3 tests and refines this framework within TextPrizm, a qualitative coding environment. Using the Generalized Cohen's Kappa (GCK) metric, this study investigates how visualizations of coder agreement and divergence support trust and expertise diversity, and operationalizes the full evaluation framework, surfacing where design assumptions hold or break down across the expertise range.

The contributions of this dissertation include a human-centered evaluation framework for DSSs, developed and refined across three studies; empirical insights into how users engage with visual decision tools across varied domains and expertise levels; and practical design priorities for designing systems that foster trust and support expertise diversity.

Chapter 2: Background and Related Work

Designing effective Decision Support Systems (DSSs) requires more than technical functionality. It demands careful attention to how users understand, trust, and interact with complex data. This dissertation draws on three foundational areas (DSSs, Data Visualization, and Human-Centered Data Science) to explore how visual tools can foster trust and expertise diversity by making system logic transparent and usable for people with varying levels of expertise.

2.1 Decision Support Systems

2.1.1 Origins and Definitions

Decision Support Systems have been a subject of research since the 1970s (Morton, 1971). In 1978, Keen and Morton defined DSSs as computer-based systems designed to enhance decision-making by assisting and supporting decision-makers rather than replacing their judgment (Keen & Morton, 1978). These systems aim to improve the effectiveness and quality of decisions by providing tools to process and analyze relevant information.

In 1980, Keen proposed a task-centered perspective for DSSs, emphasizing the importance of adaptive design to address the unpredictability of their usage (Keen, 1980). He argued that a DSS can only provide effective support when it incorporates an adaptive process in both its design and usage, enabling it to address the complexities of task identification. These complexities include insufficient knowledge to define task procedures, users' unclear and evolving needs, the influence of the DSS on shaping tasks, and variations in how different users handle tasks. To overcome these challenges, he suggested that an initial system could help users identify their needs, and as tasks and requirements evolve, the system must adapt accordingly.

Additionally, the DSS should be flexible enough to accommodate diverse user approaches and usage patterns.

A foundational distinction in the DSS literature concerns the nature of the problem these systems are designed to support. Simon (1960) described decision-making as a process of three phases: *intelligence*, which involves identifying conditions that call for a decision; *design*, which involves developing and evaluating possible courses of action; and *choice*, which involves selecting among them. Gorry and Morton (1971) built on this framework to classify decision problems along a spectrum according to how much structure exists across these three phases, a spectrum that has since been extended to include wicked problems that resist structured resolution entirely (Pretorius, 2017):

- **Structured problems** have clearly defined parameters and solutions. All three phases can be specified through algorithms or decision rules, requiring little to no human judgment.
- **Semi-structured problems** involve elements that can be computationally supported alongside elements that require human judgment. The three phases are partially structured, meaning the system can assist but cannot replace the decision-maker.
- **Unstructured problems** are undetermined across all three phases and rely entirely on human judgment for their resolution. No algorithms or decision rules can fully specify the problem or its solution.
- **Wicked problems** cannot be clearly defined, involve multiple stakeholders with conflicting values, and have no definitive solution (Rittel & Webber, 1973). They are continuously addressed rather than solved. Regional and city planning are among their canonical examples (Skaburskis, 2008).

Gorry and Morton (1971) argued that structured decisions are part of the domain of Management Information Systems, which handle routine data processing with little to no human judgment involved. DSSs were designed to support semi-structured and unstructured decisions, where human reasoning and contextual judgment cannot be replaced or automated.

Keen and Scott Morton (1978) identified three core components across DSSs: the **knowledge base**, which stores data and rules for decision-making; the **model**, which provides analytical or computational tools; and the **user interface**, which enables user interaction with the system. Several distinct types of DSSs have been identified in the literature, each reflecting a different configuration of these components (Burstein et al., 2008):

- **Model-driven:** Focuses on providing access to and manipulation of analytical models. These systems rely on simulation, optimization, and mathematical models to support decision-making by analyzing different scenarios.
- **Data-driven:** Centers on the access, manipulation, and analysis of large datasets. These systems often use data warehouses and tools for querying and online analytical processing to help users extract insights from historical or real-time data.
- **Communication-driven:** Utilizes communication and collaboration technologies, such as groupware and video conferencing, to facilitate decision-making among groups or teams by enabling structured and efficient information sharing.
- **Document-driven:** Employs document management systems to provide decision support by retrieving and analyzing unstructured documents, such as policies, procedures, or historical records, using search engines or content analysis tools.

- **Knowledge-driven:** Relies on expert knowledge and inference mechanisms to provide tailored recommendations or solutions. These systems use knowledge bases, rules, and reasoning algorithms to guide decision-making, particularly in specialized or complex problem domains.

Across the literature, these approaches have been extended to address increasingly complex and collaborative decision-making environments. In collaborative DSSs, the emphasis shifts to collective decision-making, where the preferences of multiple agents are considered and aggregated to reach a shared outcome. However, effective preference aggregation requires more than simply combining individual choices. It should incorporate contextual information and ethical considerations for more robust and morally aligned outcomes. Greene et al. (2016) advocate for moving beyond traditional aggregation methods, proposing a fusion that integrates constraints, values, ethics, and preferences. This fusion ensures that system outcomes reflect ethical principles, align with agents' moral values, and adhere to safety constraints, fostering trust and accountability in the decision-making process.

Evaluation methodologies for assessing the quality and effectiveness of DSSs have expanded significantly from early measures like time efficiency, perceived usefulness, and decision confidence to more comprehensive indicators such as ease of use, understanding, and user satisfaction. Across the literature, multiple evaluation frameworks have been proposed, each tailored to specific fields and perspectives, reflecting the breadth of domains in which DSSs have been deployed.

2.1.2 Evaluating DSSs

Early approaches to DSS evaluation focused on methodologies to measure "better" decisions, which could be personalized to each scenario (Keen & Scott Morton, 1978). "*Service*

Measures" is one of these methodologies and it involves the metrics used to assess the quality and effectiveness of DSSs from the user's perspective, such as:

- *Responsiveness of the system*
- *Availability and convenience of access*
- *Reliability*
- *Quality of system support, such as documentation and training*

Phillips-Wren (2004) et al. proposed a multiple-criteria framework for evaluating DSSs. Until then, evaluation measures focused on efficiency or effectiveness. Efficiency measures the use of resources by the DSS. Effectiveness is about decision outputs. This work aims to shift the focus of only measuring one or the other, to include both in the same framework for a more complete approach. Their framework included process measures such as proficiency in phases and steps of the decision-making and changes in the organization, while outcome measures included organizational performance and structure or culture. These measures are individualized to each decision problem domain and could include metrics such as the amount of data and variables, time to establish criteria, and closeness to previous models.

DSSs have been applied across various fields, such as medicine, finance, and education, each requiring evaluation metrics suited to their specific needs and contexts. The myriad of evaluation approaches for DSSs typically fall into three main categories (Boukhayma & ElManouar, 2015):

Multi-faceted Approach: Includes three facets, technical, empirical, and subjective. Technical evaluation focuses on system performance, including aspects such as algorithms, data flow, system development costs, and testing. Empirical evaluation examines the effectiveness of the DSS while also providing helpful feedback through methods like surveys, case studies, and

experimental methods. Subjective evaluation measures the effectiveness of the system from the different perspectives of developers, users, and organizations. The multi-faceted approach combines objective and subjective metrics within the same framework to achieve a more comprehensive and realistic evaluation. Table 2.1 provides a breakdown of metrics associated with each facet and their placement on the objectivity scale.

Objectivity of Criteria	Technical Aspects	Empirical Aspects	Subjective Aspects
Objective ↕ **Subjective**	• Data flow • Application control	• Cost/benefit analysis • Utilization information economics	
		• Decision makers' confidence • Time taken	
			• Ease of use • User interface • Understanding

Table 2.1. Multi-faceted approach evaluation framework criteria.

Sequential Approach: This approach is inspired by the decision-making process of humans, involving four main steps: identification of criteria, formative evaluation, evaluation of the outcome, and summative evaluation. The first step aims to identify the system goals, guiding the entire evaluation process. In the second step, a model is developed to be later verified and tested. Third, the appropriate solution is chosen to be later implemented in the fourth phase. The sequential approach is proposed to fit alongside the development process of the system.

General Approach: This approach is proposed as a global model to evaluate the effectiveness of DSSs in any field or domain. It employs a framework known as *decisional guidance*, which addresses four key dimensions of guidance: targets, forms, models, and scope. The evaluation focuses on three main measures: quality, efficiency, and satisfaction. These measures are applied

across two main dimensions: decision value and decision-maker. Figure 2.1 illustrates this approach.

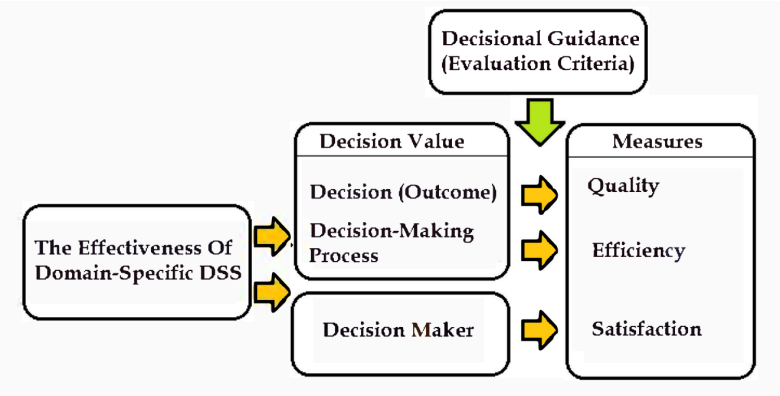

Figure 2.1. General approach framework.

Given that decision-makers are the primary users of DSSs, previous evaluation approaches incorporate their input but often overlook more human-centered factors. While the human-centered design of DSSs has been explored, the evaluation metrics remain largely focused on the output, emphasizing metrics like financial benefit, error reduction, accuracy improvement, and time efficiency (Smith et al., 2009). These measures are effective in assessing the tangible benefits of implementing a DSS but may neglect critical human components that influence these outcomes, such as trust in the system, understanding its limitations, ethical challenges faced by the decision maker, and other contextual and subjective factors.

Ongoing research has explored the moral dilemmas that DSSs present to decision-makers, designers, and developers of DSSs (Poszler & Lange, 2024; Greene et al., 2016). Proposed solutions emphasize the integration of ethical considerations into their design and development, stressing the need to address these aspects. Additionally, studies on the effects of decision stressors on decision quality (Phillips-Wren & Adya, 2020) show that factors such as

information overload, time pressure, complexity, and uncertainty negatively impact decision outcomes. These findings highlight the need to incorporate measures like self-reported stress and risk propensity into the design and evaluation of DSSs (Adya & Phillips-Wren, 2019).

Furthermore, ethical decision-making, such as in healthcare, military, or criminal justice, has been studied from the standpoint of what is called *Intelligent Decision Support Systems (IDSSs)*, which integrate artificial intelligence (AI) components into DSSs. In their research (Poszler & Lange, 2024), Poszler and Lange examine the implications of AI's inclusion, particularly its influence on the decision-making process, proposing a theoretical framework to evaluate it. Although their framework lacks empirical validation, it serves as a starting point for ethical questioning of these tools. Their work highlights the potential risks of IDSSs and discusses the consequences these technologies could have, advocating to "pursue a more reflected, human-centered approach to the use, governance, and development of IDSSs."

Despite these advances, a critical gap remains in how DSSs address the human factors critical to decision-making across diverse user groups: how users understand, trust, and make sense of what the system is doing and why. In high-stakes or interdisciplinary environments, users need to grasp not just what the system recommends, but how it arrived at that recommendation. This calls for a shift toward systems that make their logic visible, preserve expert reasoning, and accommodate diverse ways of making sense of data, a challenge I will address in this dissertation.

2.2 Data Visualization

Data visualization offers a powerful approach to decision support. Tools like dashboards and interactive graphs transform complex datasets into actionable insights, bridging the gap between technical data and diverse user expertise.

Data visualization is the graphical representation of data, transforming raw data into visual formats like charts, graphs, and maps to reveal patterns, trends, and insights. Its history spans centuries, with roots in early cartography and statistical graphics. Modern visualization techniques have been shaped by key contributions integrating principles of design, technology, and human perception. In 1983, Edward Tufte published The Visual Display of Quantitative Information (Tufte & Graves-Morris, 1983), emphasizing clarity, precision, and the elimination of "chartjunk", unnecessary decorative elements that distract from data insights. In the 1990s, Ben Shneiderman introduced the "Visual Information-Seeking Mantra" (Shneiderman, 2003), establishing principles for interactive data exploration and user-centric design.

In 2003, Shneiderman proposed the *Visual Information-Seeking Mantra* as a guiding principle for designing effective interactive data visualizations: "*Overview first, zoom and filter, then details on demand*" (Shneiderman, 2003). This mantra emphasizes a structured process for exploring datasets, allowing users to begin with a broad understanding and progressively refine their focus. By leveraging interactivity and aligning with human perceptual strengths, the mantra provides a foundation for creating intuitive and actionable visualization tools.

The mantra identifies seven tasks that visualization systems should support:

- **Overview:** Gain a high-level understanding of the dataset to identify areas of interest.
- **Zoom:** Focus on a specific subset of the data for closer inspection.
- **Filter:** Narrow the dataset by excluding data based on specific criteria.

- **Details-on-Demand:** Provide additional, in-depth information about selected items as needed.
- **Relate:** Explore connections or relationships between data elements.
- **History:** Retrace and review exploration steps, supporting iterative refinements and undoing.
- **Extract:** Save or isolate subsets of data for further use or analysis

While Shneiderman's visual information-seeking mantra provides a strong foundation for designing interactive visualizations, Craft and Cairns (2005) critique its limitations in addressing complex and dynamic design challenges. They highlight the mantra's strengths as an intuitive heuristic but point out the lack of empirical validation and specificity for diverse contexts. To address these gaps, authors advocate for the integration of design patterns as a complementary approach, offering greater flexibility and adaptability to meet varied user needs and data complexities in modern visualization systems.

Building on these principles, Card, Mackinlay, and Shneiderman co-authored Readings in Information Visualization: Using Vision to Think (Card et al., 1999) in 1999. This foundational text connected human perception with effective visualization design. It explored how visual representations leverage the brain's ability to process shapes, colors, and spatial relationships quickly. Stephen Few expanded on these ideas in 2009 with Now You See It (Few, 2009), focusing on the power of visual perception and offering practical techniques for exploring data through "thinking with our eyes." In 2014, Tamara Munzner advanced these concepts further in Visualization Analysis and Design (Munzner, 2014), providing structured frameworks for creating visualizations tailored to specific user tasks and human perceptual strengths. While some of these approaches have been critiqued and refined to meet evolving needs and

technologies, they remain foundational to the field of data visualization, providing the basis for continued innovation.

Another approach comes from the work of Tamara Munzner, who proposes a foundational framework for visualization design, offering a systematic approach to creating effective visual systems (Munzner, 2014). A key contribution of this work is that it pairs each design level with appropriate validation methods, addressing a gap in the field where visualization research had largely relied on algorithm-level benchmarks or anecdotal case studies to justify design choices.

Munzner's framework introduces four nested levels of visualization design, as shown in Figure 2.2, each addressing a specific aspect of creating effective and efficient visual systems:

1. **Domain situation**: This top-level focuses on understanding the specific domain and its users, including their goals, questions, and the data they interact with, by identifying specific workflows and tasks to inform visualization design. The goal is to identify the needs of the target audience and the domain-specific problems the visualization aims to solve. Methods such as interviews, observations, and contextual inquiries are used to define the requirements accurately.

2. **Task and Data Abstraction**: At this level, domain-specific problems are transformed into generalizable tasks and data representations. The designer abstracts user questions and actions into broader tasks such as summarizing, comparing, or identifying trends. Data abstraction involves structuring the data into forms like tables, networks, or temporal sequences that align with these tasks. This step ensures the visualization is not overly tied to domain-specific details, making it adaptable.

3. **Visual Encoding and Interaction Idiom**: This level defines how the data will be visually represented and how users will interact with it. Visual encoding involves choosing graphical elements like position, color, or spatial layouts that best communicate the data. Interaction idioms determine how users can manipulate the visualization, such as filtering, zooming, or adjusting parameters dynamically. This step integrates principles of human perception to optimize clarity and usability.

4. **Algorithm**: The innermost level focuses on the computational techniques required to implement the visual encoding and interaction idioms. Algorithms ensure that the system operates efficiently, handling large datasets and providing a seamless user experience. Key considerations include computational speed, memory usage, and scalability.

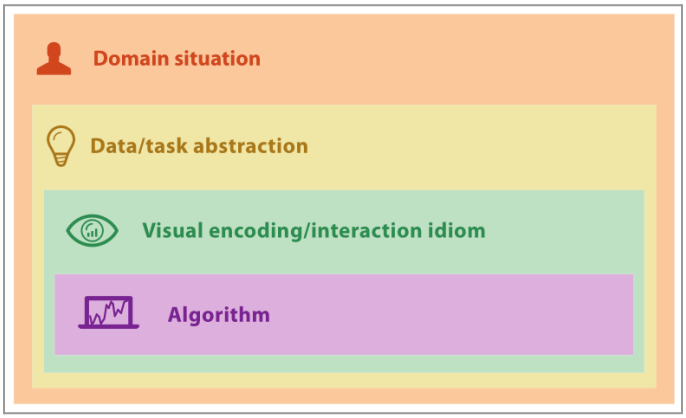

Figure 2.2. The four nested levels of visualization design.

Munzner organized the decisions at these levels through a What-Why-How typology: “What” characterizes the data type (e.g., tables, networks, temporal sequences), “Why” specifies the user’s task (e.g., compare, summarize, discover trends), and “How” defines the visual encoding and interaction choices. This typology provides a precise descriptive vocabulary for the abstraction and idiom levels, connecting what is being represented to why and how it is shown.

This layered approach emphasizes iteration, ensuring that decisions at each level refine and inform subsequent choices, and that a mischaracterization at an upper level cascades through those below it. Munzner's framework addresses this through distinct threats to validity at each level and two modes of validation: immediate validation, which examines design choices directly without requiring user input, and downstream validation, which tests whether those choices hold through user studies of the implemented system (Munzner, 2009). This framework is particularly relevant to the design of visual DSSs, as it provides the tools for balancing simplicity for novices and depth for experts while fostering user confidence through interactive, task-aligned visualizations.

Integrating narrative visualization techniques into visual DSSs enhances usability by combining structured storytelling with interactive exploration. A survey exploring approaches to merging these fields highlights various visualization structures suited for DSS design (Komenda & Karolyi, 2016).

Author-driven visualizations lead users through a clear, structured narrative using static or animated content to ensure key points are easily understood. In contrast, *User-driven visualizations* enable free exploration with interactive tools like filters, sliders, and customizable maps, catering to diverse user needs and encouraging engagement. A hybrid approach combines the strengths of both styles. The *martini glass structure* starts with a tightly guided narrative (the stem) to introduce key insights and transitions into open-ended exploration (the glass body). Similarly, *interactive slideshows* combine static content with interactive elements, guiding users through predefined sections while allowing deeper data exploration. These hybrid structures allow users to understand essential insights while exploring data in ways that suit their needs, improving flexibility, engagement, and decision-making.

In this dissertation, visualization is explored not only as a vehicle for insight but as a central mechanism for improving trust and expertise diversity. Well-designed visualizations can make system behavior transparent, reveal context-sensitive reasoning, and enable users to engage meaningfully with complex decision space, regardless of their expertise.

2.3 Human-Centered Data Science

Visualization makes data accessible and interpretable, but does not itself ensure that the systems it serves are designed with all users in mind. Human-Centered Data Science (HCDS) addresses this by centering human needs and values throughout the data science process.

HCDS was first discussed in 2016 (Aragon et al., 2016), defined as a field of study that includes human-computer interaction, social sciences, statistics, and computational skills. HCDS draws from human-centered design practices, such as qualitative research and mixed methods, to enrich data science practices through principles such as contextual knowledge, ethical responsibility, transparency, and user-centered evaluation (Aragon et al., 2022).

This field interrogates the practices of data science, questioning each step of the process. It starts with the design focusing on understanding the needs and context of our users. It continues through data collection, management, and analysis, ensuring that the data is fair, ethical, representative, and respects privacy while considering its broader impacts. Finally, it addresses how information is presented, aiming to deliver useful information to our users, supporting them in gathering meaningful and actionable insights (Aragon et al., 2022).

2.3.1 HCDS as a Framework for Data Science

Data science is an interdisciplinary field that focuses on computational approaches to analyzing large-scale data, combining methods from multiple fields, such as mathematics, statistics,

computer science, and information sciences. The data science cycle often begins with a question, followed by data collection, processing, and analysis, culminating in the representation and interpretation of results. These steps aim to transform raw data into actionable insights that are meaningful for the user (Aragon et al., 2022).

While traditional data science has been largely technical, HCDS builds on this foundation by integrating user needs, values, and contextual understanding at every stage of the data science cycle. This approach incorporates qualitative and mixed methods, recognizing the critical role of human intervention by data scientists, which is often overlooked. For example, during the data collection phase, a data scientist might *discover* a relevant dataset, *capture* it intentionally, *curate* or modify it to suit project needs, *design* categories to organize it, or even *create* new values. Each of these steps involves human judgment, making it essential to study these interventions to better guide data scientists in making ethical and well-informed choices.

In *Human-Centered Data Science: An Introduction (Aragon et al., 2022)*, four key practices are proposed to guide any data science project. These practices are concerned with not only the process of doing data science but also the importance of considering the broader context in which the project is situated.

Asking Good Questions: The starting point should always be a well-thought-out and meaningful question or questions. Even though we might have a good and large dataset, it might not be useful to answer our questions. Good questions should be novel, clear, specific, and have a well-defined, doable scope. Questions should be empirical, falsifiable, focused, and important. This step in the process will guide the design of our project and be a constant reminder of what the direction of our project is, even if the questions are revisited.

Ethics: Data science has the potential to create powerful tools that drive significant change, enabling discovery, the development of new services, and making everyone's day-to-day activities more efficient. Data is power, and with this power comes the responsibility to adhere to ethical guidelines, as data can be used for both good and harm. *The Belmont Report,* created by the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research*,* provides three general principles to guide research ethics with human subjects: respect for persons, beneficence, and justice. In HCDS, ethics serves as a crucial point of reflection throughout the process. Each stage must be examined through an ethical lens assessing fairness and considering who may be harmed or benefit. Ethical challenges should be addressed continuously, ensuring that the process remains fair at every stage of design, development, and implementation.

Designing for Others: One of HCDS's key practices involves considering how others might use our project, emphasizing *reproducibility*, *accessibility*, *openness*, and *readability*. Reproducibility allows others to replicate the project and achieve the same results, requiring us to provide everything necessary for them to do so. Accessibility focuses on ensuring that the work is inclusive and accounting for various impairments that users might have. Openness refers to making our work accessible for others to learn from and build upon, though this may be limited if data privacy concerns are involved. Finally, readability and legibility ensure that the process and decisions leading to our results are clearly described, enabling others to understand and engage with the work effectively.

Reflection: Reflexivity is a key distinction of the HCDS field, encouraging continuous self-reflection on decisions and practices throughout the data science process. Since human input is involved at every stage of the data science cycle, incorporating reflection allows us to reevaluate our choices and decisions. This critical thinking can uncover overlooked issues that

might impact the performance of our tools, ensuring that potential biases, ethical concerns, and unintended consequences are addressed.

2.3.2 HCDS Methods

In addition to the considerations for the design, development, and implementation of data science systems, methods from social science, design studies, and critical theory provide valuable perspectives on the way we approach these tools. These methods enable critical viewpoints that can uncover issues that might otherwise remain hidden.

Qualitative Social Science Methods: While quantitative approaches analyze numerical data through mathematical models, qualitative research methods focus on *qualities*, capturing subjective experiences and contextual insights that numbers cannot measure. Methods such as observation, interviews, focus groups, participatory analysis, ethnography, and grounded theory provide meaningful ways to explore complex social dimensions.

Value Sensitive Design: Value Sensitive Design (VSD) aims to integrate human values into the design of systems from an ethical standpoint (Friedman, 1996). This includes prioritizing values such as user autonomy, the minimization of bias, and the incorporation of data statements (Bender & Friedman, 2018). VSD also highlights the importance of considering both *direct* stakeholders who directly interact with the system and *indirect* stakeholders who are impacted by or have an interest in the system's outcome. This approach provides a moral framework for examining designs, fostering the integration of ethical human values into technological development.

Data Feminism: Feminist theory seeks to challenge sexism and other forms of oppression, advocating for a future that is just, equitable, and livable. Data feminism builds on these principles by first examining how conventional practices in data science reinforce existing

inequalities. It then aims to leverage data science as a means to challenge and transform the distribution of power (D'Ignazio & Klein, 2020). This framework acknowledges the often-overlooked reality that "***power is not distributed equally in the world***."

In *Data Feminism*, D'Ignazio and Klein propose seven core principles to guide efforts in using data to challenge and change power structures: (1) *examine power* by analyzing its dynamics; (2) *challenge power* to work towards justice; (3) *elevate emotion and embodiment* by valuing diverse, lived knowledge; (4) *rethink binaries and hierarchies* that perpetuate oppression; (5) *embrace pluralism* by integrating multiples, especially local and experiential, perspectives; (6) *consider context* as data is shaped by the context in which it exists; and (7) *make labor visible* to acknowledge the collaborative efforts behind data work.

Critical Race Theory: Critical race theory applied to data science focuses on identifying and addressing racial inequalities embedded in technologies, which can perpetuate and automate racial discrimination (Benjamin, 2019). As racial biases exist in everyday life, the data collected from society often reflect these biases, reinforcing discriminatory practices through technology. By examining the biases inherent in both the data and design choices, we can take proactive measures to mitigate harmful outcomes.

For this dissertation, HCDS offers a critical lens through which to evaluate how visual DSSs are designed and whom they serve. Of the methods discussed in this section, qualitative social science methods and Value Sensitive Design are most directly suited to examining trust and expertise diversity, as they provide the tools to capture users' independent perspectives and ensure that the full range of stakeholders is considered.

2.4 Expertise Diversity and Trust in DSS Design

Sections 2.1, 2.2, and 2.3 establish both a foundation and a gap. DSSs define the decision-making context, data visualizations surface meaning from complex data, and HCDS centers the human experience. However, integrating them in ways that center users across expertise levels and support trust remains an open challenge. Addressing this begins with examining how expertise and trust have each been approached in the visual DSS literature.

2.4.1 Expertise Diversity in Visualization

Users of visual DSSs vary widely in their domain knowledge, roles, and familiarity with data visualization. Understanding how these differences shape user needs is essential for designing systems that are effective across expertise levels. Non-experts may feel overwhelmed by complex visualizations, whereas expert users often need advanced tools to explore detailed data.

Previous studies (Maltese et al., 2015) (Tomasi et al., 2023) have shown that individuals with higher domain expertise exhibit superior data visualization literacy, achieving greater accuracy and experiencing lower cognitive load, particularly with simpler designs. These findings highlight the importance of novice-friendly formats that simplify information while still allowing experts to engage with complex data through layered or customizable features.

Interactivity has proven valuable in bridging the literacy gap between novice and expert users. Interactive features such as filtering, zooming, and adjustable parameters enhance engagement and improve comprehension, making visualizations more accessible to less experienced users. Simultaneously, these tools provide the flexibility experts need to perform in-depth analyses (Oliveira et al., 2021).

Additionally, training plays a critical role in reducing disparities between expertise levels. Training helps novices develop the skills and cognitive strategies that experts use when interpreting visualizations, thereby improving judgment accuracy, efficiency, and confidence (Eberhard, 2023). Structured training programs and repeated exposure to visualization tasks effectively narrow the performance gap.

2.4.2 Trust and User Confidence

Beyond expertise, trust and user confidence shape how effectively people engage with visual DSSs. User confidence refers to the level of trust or certainty a user has in their decision-making process or the advice provided by a DSS (van der Waa et al., 2020). User confidence in themselves and the system are closely linked. Distrust in the system can undermine self-confidence, even if the user initially feels certain.

Studies in visualization and decision-making reveal that visualizations can boost user confidence without necessarily improving decision quality, leading to overconfidence. Incorporating explicit uncertainty representations in visualizations helps users better evaluate data reliability, reducing overreliance on potentially misleading or incomplete information and promoting more cautious decision-making (Eberhard, 2023). The relationship between uncertainty awareness and trust calibration in visual analytics systems has been further examined by Sacha et al. (2016), who argue that users who are unaware of uncertainties propagating through a system are prone to overtrust or undertrust its outputs, undermining effective decision-making.

Furthermore, user confidence and trust, especially in AI-based DSSs, are strongly influenced by the system's transparency, explicability, and interpretability. Clear explanations and

understandable reasoning are essential to foster user confidence, especially in high-stakes domains (Nicodeme, 2020).

Building on the challenges identified in Section 2.1, and drawing from the frameworks established in Sections 2.2 and 2.3, this dissertation focuses on two constructs that are central to designing visual DSSs that support expertise diversity and user trust. I conceptualize *expertise diversity* as the system's capacity to serve users with varying roles, knowledge backgrounds, and goals. A DSS designed for expertise diversity accommodates multiple ways of seeing and interpreting data, rather than designing for a single "average" user. I define *trust* as the user's ability to understand what the system is doing, why it is doing it, and how reliable the outputs are given the data sources and decision rules in use. Trust in DSSs is not solely a function of accuracy, but about communication of reasoning, uncertainty, and limitations, enabling users to make informed decisions with confidence.

This dissertation investigates how visual decision tools can be designed and assessed not just for performance, but for how well they help users interpret decisions, retain expert knowledge, and build trust across diverse levels of experience. Through three studies, on design needs (Chapter 3), visualization features (Chapter 4), and evaluation strategies (Chapter 5), it contributes towards a new, human-centered approach to supporting decision-making through systems that foster trust and support expertise diversity.

2.5 A Theoretical Framework for Evaluating Visual DSSs

Designing visual DSSs that support users across expertise levels requires a careful balance between simplicity and depth. Novices need guidance and intuitive interactions, while experts

require advanced tools for deeper explorations. How to evaluate whether a system achieves this balance, and whether it promotes trust in doing so, remains an open question.

This dissertation proposes that Munzner's nested levels of visualization design and HCDS methods, used together, offer a productive basis for approaching that question, one that combines the structural lens of visualization design theory with the critical and ethical tools needed to center trust and expertise diversity in the evaluation. Building on Munzner's validation logic, this section extends that framework to the evaluation of visual DSSs by specifying evaluative questions at each nested level and by introducing an independent user-side lens that captures what users bring outside of the system's framing.

2.5.1 Munzner's Framework as an Evaluative Lens

Applied as an evaluative lens, Munzner's framework offers a systematic way to assess visual DSSs across four nested levels, from how well the system understands its users to how effectively it encodes and presents information. Its layered, human-centered approach ensures that the design of the DSS is deeply rooted in the domain context, allowing user needs and workflows to guide every stage of the development. While the framework was originally proposed for both design and validation, this dissertation draws on its evaluative dimension, using its nested structure to assess whether a system's design choices, at each level, adequately account for users with varying expertise and communicate system reasoning in ways that support trust.

Following Munzner's distinction between immediate and downstream validation (Munzner, 2009), the evaluative questions specified at each level can be answered at two points: by examining the system's design choices directly, and by testing whether those choices hold

against user evidence. The following paragraphs examine each level, proposing specific threats and validation questions that are developed further through the empirical work in the chapters that follow.

At the **domain situation** level, the system's design reflects assumptions about who its users are, what they need, and what problem they are trying to solve. Evaluating this level involves examining whether user research methods such as interviews, focus groups, and surveys were used to capture user goals and challenges, and whether contextual inquiries and observational studies were applied to obtain insights into workflows and decision-making contexts across varying expertise levels.

Insights gained at this stage inform subsequent layers, meaning that a mischaracterization here cascades downstream. When the system is aligned with domain-specific needs, however, it ensures relevance and usability, directly addressing the goal of fostering user confidence and supporting expertise diversity.

- **Threat***: The domain problem is mischaracterized, failing to account for the goals and expertise range of its users.*
- **Immediate validation:**
    - Does the system's domain characterization explicitly account for users across a range of expertise levels?
    - Was user research conducted with the target users prior to or during development, capturing their goals, challenges, and decision-making workflows?
- **Downstream validation*:***
    - Do users across expertise levels recognize the system as relevant to their own goals and decision-making context?

At the **task and data abstraction** level, the system's design choices reflect assumptions about which tasks users need to perform and how data should be structured to support them. Evaluating this level involves examining how decision-making processes have been broken down into core tasks such as comparing data points, identifying trends, or evaluating scenarios, and how data has been structured into usable formats, such as time-series data for business trend analysis or hierarchical models for organizational insights. Critically, it also asks whether tasks cater to both novice users (e.g., summarizing trends) and experts (e.g., performing in-depth analyses).

Abstracting tasks and structuring data ensures that the DSS supports expertise diversity, as identifying the common and specific tasks of different users allows for a more informed assessment of representations, enhancing accessibility and decision-making across user groups.

- **Threat***: The abstraction fails to reflect the tasks and data needs of users across expertise levels.*
- **Immediate validation:**
    - Do the data types require prior domain knowledge to interpret?
    - Do the tasks the system supports assume a particular level of domain expertise, or do they accommodate users across the expertise range?
- **Downstream validation*:***
    - Do users across expertise levels interpret the data structures accurately, or do certain formats create confusion or misinterpretation?
    - Are users across expertise levels able to complete the tasks the system was designed to support?

At the **visual encoding and interaction idiom** level, the system's design choices determine how data is visually represented and how users interact with it, directly shaping whether the

system is interpretable and trustworthy across the expertise range. Evaluating this level involves examining how the design accounts for the full range of user needs, through distinct formats, layered complexity, or interactive features that allow users to engage at varying depths, and how effectively it communicates data in ways that support trust in system outputs.

Drawing on Shneiderman's mantra, interactions such as filtering, zooming, and details on demand can serve as criteria for assessing how well the system supports users across expertise levels in navigating and exploring data at their own pace. Narrative patterns, such as the *martini glass* approach, provide structured guidance for novices while allowing experts to customize their exploration pathways, fostering user confidence and supporting expertise diversity.

- **Threat***: The visual encoding and interaction choices fail to make data interpretable and trustworthy across the expertise range.*
- **Immediate validation:**
    - Are the visual encoding choices appropriate for the data and clearly communicated so users can interpret them without domain-specific knowledge?
    - Does the visualization communicate data provenance, limitations, and uncertainty explicitly?
    - Do the interaction mechanisms allow users to engage at varying depths, or do they assume a single exploration path?
- **Downstream validation*:***
    - Do users correctly interpret the visual encodings, or does confusion emerge at particular expertise levels?
    - Do users express appropriate confidence in the system's outputs, or do patterns of over- or undertrust appear?

- Do users across expertise levels engage with the interaction mechanisms at varying depths, or do certain features remain inaccessible to particular user groups?

At the **algorithm** level, the system's computational performance and output accuracy determine whether the system supports or disrupts user engagement. Efficient algorithms are critical for maintaining engagement and trust, as performance issues during computationally intensive tasks such as filtering or processing large-scale data can disrupt the user experience and undermine confidence in the system.

- **Threat***: The algorithm fails to support user engagement through insufficient performance or incorrect outputs.*
- **Immediate validation:**
  - Does the system respond to interactions without disruptive delays under expected use conditions?
  - Does the algorithm produce accurate and consistent outputs across different user interactions?
- **Downstream validation*:***
  - Do users maintain confidence in the system during computationally intensive operations, or do delays trigger uncertainty or disengagement?
  - Do users across expertise levels trust the system's outputs, or do inaccuracies surface and undermine confidence?

Together, these four levels offer a systematic lens, rooted in visualization theory, for assessing whether a visual DSS supports expertise diversity and trust across its full design. An upstream mischaracterization of users at the domain level will cascade through every level below it, meaning that evaluation at each level is not independent but deeply interconnected. Munzner's

framework is already human-centered in its approach, keeping user needs and workflows central to evaluating the system. However, at its core, it is the system that is being evaluated, with users serving as input to that process rather than as the frame of reference. HCDS provides the tools to extend that frame, centering the users' experience and asking whether the system addresses users across expertise levels and in ways that support trust.

This dissertation repurposes Munzner's framework. It retains the nested four levels, the cascade logic in which a mischaracterization at an outer level propagates through every level below it, the distinction between immediate and downstream validation, and the threat-and-validation scaffolding that structures each level. It replaces Munzner's specific per-level threats and validations, which assess the validity and effectiveness of a visualization's design choices, with the threats and questions specified above, tailored to whether a system supports trust and accounts for the expertise its users bring. This framework offers a way to locate where problems originate. A breakdown in trust or a failure to support expertise diversity can be traced to the level where the design choice was made, and the cascade logic helps explain how that choice may shape the levels below it.

2.5.2 HCDS Methods Across the Nested Levels

HCDS is particularly suited to complement Munzner's framework because visual DSSs sit at the intersection of data science and human-centered design, where understanding user characteristics, tasks, and context are all inherently intertwined with system design (Smith et al., 2012). As a field that explicitly integrates human-centered practices into the data science process, HCDS provides the tools to evaluate not just how a system performs, but how it is understood and trusted by the people who use it (Aragon et al., 2022). Four core practices orient this approach: asking good questions, ethics, designing for others, and reflection. Applied at

each level of Munzner's framework, these practices extend the evaluative focus to include what users bring independently of the system's assumptions, asking who the users are on their own terms, what they actually need, and how they make sense of data outside of how the system defines it.

This dissertation approaches evaluation at each level by seeking to capture both the system's design choices and the user's independent perspective, then examining the alignment between them. Munzner's framework is already human-centered in its approach, keeping user needs and workflows central to the evaluation, but the system remains the primary unit of analysis, with users serving as the reference point against which design choices are assessed. HCDS provides the tools to capture what users bring independently in the context of data-driven decision making, their goals, expectations, mental models, and ways of making sense of data. The gap between what the system assumes about its users and what users actually need is where trust breaks down and expertise diversity is either supported or excluded. Where the two perspectives diverge, misalignments in how the system was designed and how users actually experience it may become visible.

The following paragraphs examine the independent user-side lens at each level, identifying the goal of applying HCDS methods and exploring how that might look in practice. These examples are illustrative rather than definitive, as many methodologies could serve the same goal at each level, and the appropriate choice will depend on the research context and user population.

Domain: Users bring their own understanding of the domain problem, their role within it, and what decisions they are responsible for making. Interviews and contextual inquiry, alongside VSD's stakeholder analysis are the primary tools here. VSD is particularly valuable because it prompts consideration of both direct stakeholders, such as users and developers, and indirect stakeholders, those affected by the system's outcomes without directly interacting with it,

providing valuable perspectives on broader societal and organizational implications that are often absent from system design. Examining both perspectives surfaces questions about whose expertise was centered in defining the problem and whether certain user groups were overlooked or assumed to share the same needs.

- **Goal:** *Identify the full range of users present in the domain, their roles, stakes, and how they understand the problem and their place within it.*
- **Illustrative Applications:**
    - Stakeholder mapping to identify who is directly involved in decision-making in this domain and who is affected by those decisions without direct involvement.
    - Semi-structured interviews asking users to describe their role in the domain, the decisions they make, and what information they consider essential.

Task and Data Abstraction: Users hold their own mental models of what tasks matter to them and how they conceptualize the data relevant to their decisions. Task analysis and think-aloud protocols can reveal how users actually approach their goals, which may differ across expertise levels. An expert may need detailed comparative analysis while a novice may need trend identification, and if the abstraction only serves one profile, that misalignment becomes visible here.

- **Goal:** *Understand the tasks users consider most meaningful in their domain and how they think about and structure the information relevant to their decisions.*
- **Illustrative Applications:**
    - Card sorting exercises where users group and rank types of domain information by relevance to their decisions, surfacing how they naturally structure and conceptualize data.

- Semi-structured interviews asking users to walk through how they approach a typical decision in their domain, what steps they take, and what information they prioritize.

Visual Encoding and Interaction: Users bring existing visual literacy and trust criteria to any data representation they encounter, shaped by their expertise and position in the domain. Think-aloud protocols and interviews capture what users actually perceive, understand, and trust, and where confusion or lack of confidence emerges across expertise levels. VSD's data statements are relevant here because transparency about data sources, quality, and limitations shapes whether users can trust what the system shows them independently of how well it was designed at the levels above.

- **Goal:** *Understand how users across the expertise range interpret and evaluate visual information in their domain, and what they consider trustworthy or meaningful in data representations.*
- **Illustrative Applications:**
  - Think-aloud protocols where users engage with visual representations they currently use in their domain, verbalizing what they find trustworthy, meaningful, or confusing.
  - Participatory envisioning where users sketch or describe what a meaningful and trustworthy visual representation of their domain problem would look like to them, surfacing their visual literacy and trust criteria.

Algorithm: While this is the most removed level from the independent user-centered perspective, performance issues such as lag or processing delays can affect trust differently across expertise levels. An expert may recognize a delay as a normal computational process while a novice may interpret it as system failure, undermining confidence in the system's reliability.

- **Goal:** *Understand users' existing expectations of reliability, accuracy, and transparency of computational processes in data-driven outputs within their domain context.*
- **Illustrative Applications:**
    - Wizard of Oz prototyping where the researcher controls simulated data-driven outputs to probe at what point users interpret computational behaviour as unreliable or untrustworthy in their domain context.
    - Scenario-based questioning presenting users with hypothetical data-driven outputs that conflict with their domain intuition, asking them to reason through whether they would trust the result and why.

Data Feminism and Critical Race Theory offer important critical lenses for evaluating DSSs. Data Feminism's emphasis on pluralism aligns with this dissertation's focus on expertise diversity, and CRT provides a framework for examining how DSSs may perpetuate or mitigate racial inequalities through their data and design choices. However, fully operationalizing either framework would introduce dimensions of power structures and racial inequality that require their own dedicated analytical approaches, expanding the scope of this dissertation well beyond its focus on expertise diversity and trust. This dissertation deliberately centers expertise diversity as the construct through which pluralism is examined, and trust as the outcome through which system alignment with user needs is assessed. Examining how racial dimensions and power structures intersect with expertise and trust remains a productive and necessary direction for future work.

Qualitative social science methods and Value Sensitive Design are particularly well suited to capturing the independent user perspective in the context of trust and expertise diversity. Interviews and think-aloud protocols surface user's goals, expectations, and ways of making sense of data on their own terms, while VSD's stakeholder analysis ensures the full range of

users affected by the system is considered, and data statements address transparency about data sources and limitations. The studies that follow draw primarily on these methods to capture the user perspective, adapting their application to each research context.

2.5.3 Toward Human-Centered Design Priorities for Visual DSSs

The literature reviewed in this chapter points towards four areas that kept emerging across these bodies of work as likely to matter most when evaluating whether a visual DSS fosters trust and supports expertise diversity. Grounded in the theoretical integration above, these are introduced here as preliminary design priorities that will be further refined and revisited in Chapter 6.

Customization: Achieved through interactivity, this priority allows the system to adapt to the specific needs of users, recognizing that different levels of expertise require different ways of exploring and interacting with data. Craft and Cairns (2005) identified limitations in Shneiderman's interaction principles for diverse user contexts, pointing toward the need for systems that support different ways of navigating and exploring data rather than assuming a single exploration path. Through visual encoding and interaction design, interactivity empowers users to tailor their experience, ensuring that users across expertise levels can engage with the system in ways that match their goals and workflows.

Training: Support for data literacy and user confidence in data-driven decision making is essential for users navigating visual DSSs. When users lack the background to interpret certain representations, the system needs mechanisms to bridge that gap (Eberhard, 2023). By leveraging narrative patterns (Segel & Heer, 2010), the system can guide users through learning processes, making complex data and visualizations more accessible. This approach ensures that users develop the skills needed to engage effectively with the DSS, fostering a sense of competence and confidence in their decision-making abilities, connecting directly to both the

abstraction and encoding levels, where the gap between what the system assumes users know and what they actually bring to the interaction is most visible.

Transparency: Clear communication of system processes, including the limitations and uncertainties inherent in the data and its representations, is the foundation of this priority. Users need to understand not just what the system shows but why, meaning the logic and reasoning behind system outputs, and how reliable it is, including the quality, limitations, and uncertainty of the underlying data (Sacha et al., 2016), a need that connects directly to trust and is operationalized through VSD's data statements (Bender & Friedman, 2018). By openly sharing this information, the system fosters accountability and empowers users to make well-informed decisions. This is particularly relevant at the encoding and interaction level, where users with less domain knowledge may be more vulnerable to misplacing trust in system outputs they cannot fully evaluate.

Adaptability: A system's capacity to adapt is essential to ensure the DSS remains effective as user needs and data evolve. Keen (1980) argued that a DSS can only provide effective support when it incorporates an adaptive process in both its design and usage, recognizing that user needs are unclear and evolving. Evaluation must be integrated throughout the design process to assess the system's effectiveness while considering contextual and ethical aspects, connecting to HCDS's reflective practice and the domain level, where a deeper understanding of users should feed back into how the system evolves, ensuring continuous improvement and long-term relevance.

The system-side lens established in Section 2.5.1 and the independent user-side lens introduced in 2.5.2 work together to surface where visual DSSs support or exclude trust and expertise

diversity. How these lenses are applied, and in what sequence, is explored through the studies that follow and consolidated in Chapter 6.

Chapter 3: Contextual Knowledge and System Transparency in Traffic Incident Decision Support

3.1 Project Background and Context

The Virtual Coordination Center (VCC) is a shared digital operational environment where diverse agencies engaged in regional transportation management can share data, communicate, and collaboratively respond to stresses on the transportation system. Partner agencies include King County Metro (KCM), Seattle Police Department (SPD), Seattle Fire Department (SFD), Washington State Patrol (WSP), Seattle Department of Transportation (SDOT) and Washington State Department of Transportation (WSDOT). The VCC functions as a knowledge-driven decision support system in a high-stakes operational context, where timely coordination across agencies is essential for managing major traffic incidents effectively. The VCC was developed by the University of Washington's Center for Collaborative Systems for Safety, Security, and Resilience (CoSSaR) and by May 2023 had nearly 200 active users across seven partner agencies (Haselkorn & Phelps, 2024).

A central challenge for VCC operators is the early identification of events likely to escalate into what the system designates as "VCC-level incidents": situations requiring cross-agency coordination and shared situational awareness. Figure 3.1 shows the VCC interface as operators encounter it, with the dispatch feed, the map view, and the panel of active Incident Models.

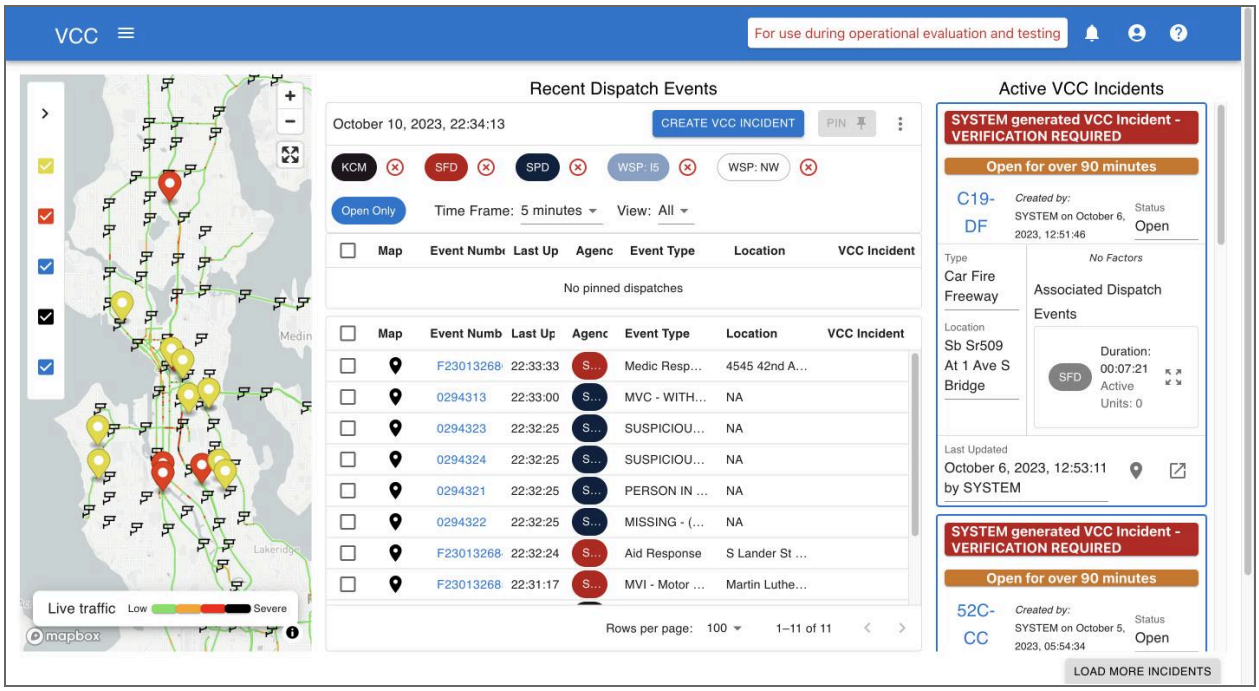

Figure 3.1. The VCC operator interface, showing the dispatch feed (center), the map view (left), and the Active VCC Incidents panel (right)

As dispatches arrive in the integrated dispatch feed, they are evaluated by transportation managers who can launch an Incident Model if they assess the event as a likely VCC-level incident. In addition to this human review, the VCC applies a rules engine that evaluates incoming dispatch data and automatically generates an Incident Model (IM) when certain criteria are met. The goal of this mechanism is latency reduction, not decision support. It targets dispatches that are obvious enough to be identified through a small set of yes-no criteria, reducing the time between their arrival and their classification. Dispatches whose status depends on interpretation are left to operator judgment by design. These criteria, developed through consultation with experienced operators, flag the following types of dispatches:

- Events from the Seattle Fire Department that include: "*Tunnel MVI*", "*Car Fire Freeway*", or "*Fire Response Freeway*" in the event type.

- Events from the Washington State Patrol in Area "*I5*" that include: "*Road Closure*", "*Fatal Traffic Collision*", "*Disabled Vehicle Fire*", or "*Possible suicidal pedestrian on bridge or overpass*" in the event type.
- Any event that includes "*bridge*" in the location and "*blocking*" in the event type.

When a dispatch matches one or more of these criteria, the system generates an IM and sends an email alert to all users. All system-generated IMs are marked unverified until confirmed by a human incident manager, who may modify or delete the model based on additional contextual knowledge. What counts as additional contextual knowledge, and who holds it, is not always clear to the system or to newer operators.

The project team described the rules engine as "*a promising start*" and "*a good first step*" (Haselkorn & Phelps, 2024). Its criteria were kept deliberately narrow because classification beyond the obvious cases is complex for two reasons. First, misclassifying a dispatch as a VCC-level incident when it does not warrant that status can mobilize valuable agency resources unnecessarily. Second, dispatch data alone does not always reveal whether an event will escalate, as incidents that appear routine at the outset can evolve into situations requiring full cross-agency coordination.

To investigate these challenges and explore opportunities for design improvements, this study was conducted from June 2022 through Fall 2023, beginning with semi-structured interviews of VCC operators in July 2022 and followed by quantitative analysis of dispatch and incident data collected between February and September 2023. The goal was to understand how operators reason through incident classification, identify what factors inform their judgment, and determine design directions that could better support decision-making across varying levels of operator experience and role.

3.2 Research Objectives

The complexity of incident classification in the VCC environment raises questions that go beyond the performance of the rules engine. Operators bring contextual knowledge to their assessments, such as time of day, location patterns, agency-specific priorities, weather conditions, or proximity to key infrastructure, that are not systematically captured in the system. As operators turn over and institutional memory shifts, that expertise risks being lost.

This study set out to examine how operators reason through classification decisions, what factors shape their judgment, and what that reveals about the design of decision support tools intended to serve users across varying levels of experience. Because the rules engine was scoped to the obvious cases, comparing its logic to operator reasoning is not a test of whether the system matches human judgment. It is a way of examining where automated detection ends and what reasoning operators apply beyond it. By combining analysis of incident data with semi-structured interviews of VCC operators, we aimed to answer the following questions:

RQ1. What reasoning processes and contextual knowledge shape how operators make classification decisions?

RQ2. How does the VCC's rules engine classification logic compare to how operators reason through incident assessment?

RQ3. What does the division between system classification logic and operator reasoning suggest for designing decision support that serves operators across experience levels?

This study surfaced a foundational design challenge that is not unique to the VCC: when decision support systems do not make their logic visible or preserve the reasoning of experienced users, they risk losing the very expertise their operations depend on.

3.3 Methods

This study used a mixed-methods approach to understand how operators reason through incident identification and to examine the classification work that falls outside the rules engine's detection criteria. This approach was selected because neither data source alone could fully address the research questions. Following Tashakkori and Creswell (2007), mixed methods research is understood here as an approach in which the investigator collects and analyzes data, integrates the findings, and draws inferences using both qualitative and quantitative methods. Data collection and operator interviews were conducted as part of the federally funded VCC project, supported by the Federal Highway Administration's Advanced Transportation and Congestion Management Technologies Deployment Program (Haselkorn & Phelps, 2024). This study constitutes a secondary analysis of the operational data and interview records.

Quantitative analysis of incident records reveals patterns in how the system and operators classified incidents, but it cannot capture the reasoning behind those decisions. Qualitative interviews surface the contextual knowledge operators bring to their assessments, knowledge that exists in practice but is not explicitly represented in system logic. Together, the two methods provide a grounded view of both what the system did and how operators reasoned independently of it, which is central to the human-centered evaluation approach used in this study.

The qualitative component consisted of semi-structured interviews with five VCC operators across partner agencies, each lasting approximately one hour. Participants were selected through purposive sampling, prioritizing operational expertise and agency representation over sample size. Because the VCC is used by operators and transportation managers in defined operational roles across specific partner agvencies, selection was based on agency and role. The five participants represented three distinct partner agencies: a state level agency, a regional level

agency, and a city level agency, each with different operational scopes and criteria for what constitutes a VCC-level incident. In expert interview research, coverage across roles and perspectives is a stronger indicator of data quality than sample size alone (Patton, 2002). To protect participant confidentiality while maintaining analytical transparency, participants are referred to throughout this chapter using assigned identifiers (P1 through P5). Table 3.1 provides an overview of participant roles and agency affiliations.

Participant	Agency	Role
P1	State level agency	Operations Manager
P2	Regional level agency	Operations Chief
P3	City level agency	Operations Supervisor
P4	City level agency	Operations lead
P5	City level agency	Operations lead

Table 3.1 Participant Overview.

Interviews began with open-ended questions about how operators assess incoming dispatches and decide whether to launch an Incident Model. Participants then walked through a set of actual past incidents, reflecting on whether they considered each one a VCC-level event and what factors informed that judgment. This structure was designed to surface the contextual reasoning operators apply in practice, beyond what dispatch data alone can capture. The full interview protocol is included in Appendix A.

Interview data was synthesized through close reading and cross-participant comparison, focusing on recurring variables, shared or individual challenges, and agency-specific patterns in how operators define and respond to VCC-level incidents. These variables then guided the quantitative analysis of Incident Model data collected between February and September 2023.

The quantitative component drew on all Incident Model reports generated in the VCC between February 22 and September 30, 2023. Each report contains information about the incident including location, event type, start time, associated dispatches, who created it, operator updates, clearance time, collaborative actions, and map annotations. This data enabled analysis of patterns in how incidents were detected, verified, and resolved across the study period.

3.4 Findings

The findings reveal a complementary relationship between system detection and operator judgment in incident classification. Notably, approximately 72% of closed Incident Models were identified solely by operators, raising a key question: what reasoning do operators apply to the classification work the rules engine leaves to them, and what would it take to support it? To explore this, we first examine overall patterns in Incident Model creation and outcomes. During the study period, 354 Incident Models were generated in the VCC. Of these, 302 were closed as verified incidents and 52 were deleted. Table 3.2 shows the breakdown of closed and deleted Incident Models by origin, distinguishing between those created by the user and those generated by the system.

Incident Model Final Status	**Created by**	**Quantity**
Closed	User	217
	System	85
Deleted	User	29
	System	23
TOTAL Incident Models	n/a	354

Table 3.2. Number of closed and deleted Incident Models.

Every Incident Model in the VCC is ultimately either closed, when the situation is cleared, or deleted, if it was created incorrectly, was a test, or was system-generated and never verified by an operator. Figure 3.2 shows the cumulative count of all IMs created over the study period. Human-generated incidents are represented in blue and system-generated incidents in red, with dashed lines indicating deleted IMs. The visualization shows that the number of deleted incidents remained relatively constant over time, while closed incidents, both human and system-generated, increased almost linearly. The pattern of deleted system-generated incidents is particularly informative, as it reflects cases where the rules engine flagged a dispatch that operators did not confirm as a VCC-level event.

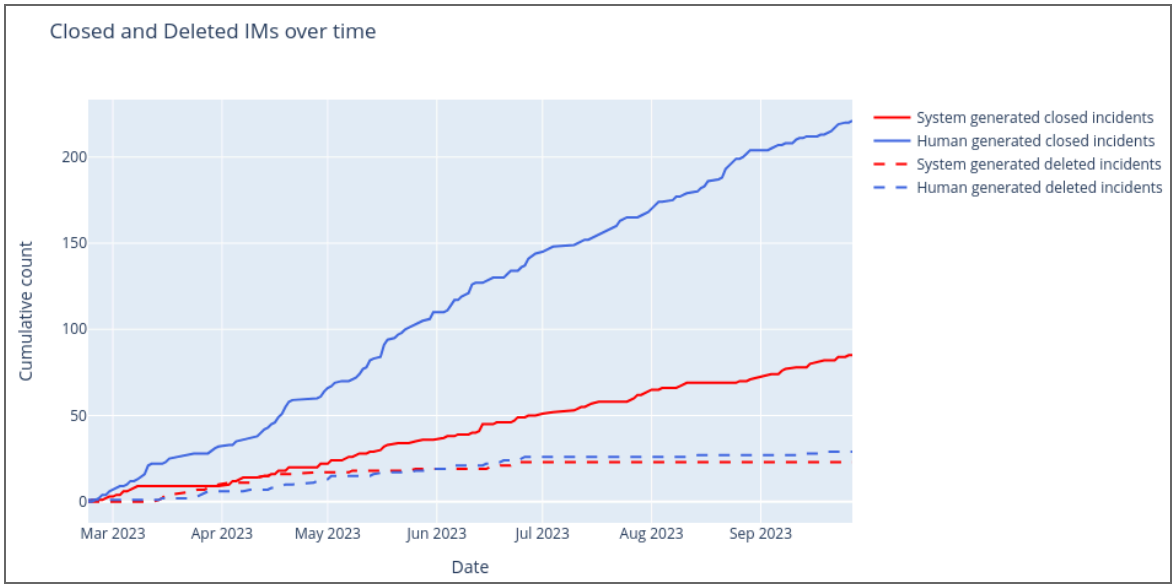

Figure 3.2. Closed and Deleted Incident Models

System-generated IMs represent 28.15% of all closed Incident Models, while system-generated deleted IMs represent 44.23% of all deleted models. Of the 108 system-generated IMs, only 21.29% were deleted, suggesting that nearly 80% were verified by an operator or deemed worthy of maintaining as a record. This indicates that the rules engine is effective at avoiding misclassification, a critical consideration given that unnecessary alerts can mobilize valuable

agency resources. Figure 3.3 shows the distribution of Incident Models over time, illustrating that incidents were not created every day and that some days had significantly higher activity than others, with an average of 42.43 closed incidents per month, calculated across full months in the study period. Thirty percent of all data collection days had zero incidents while another 30% had only one. The month of May was a notable exception, with 60 Incident Models launched.

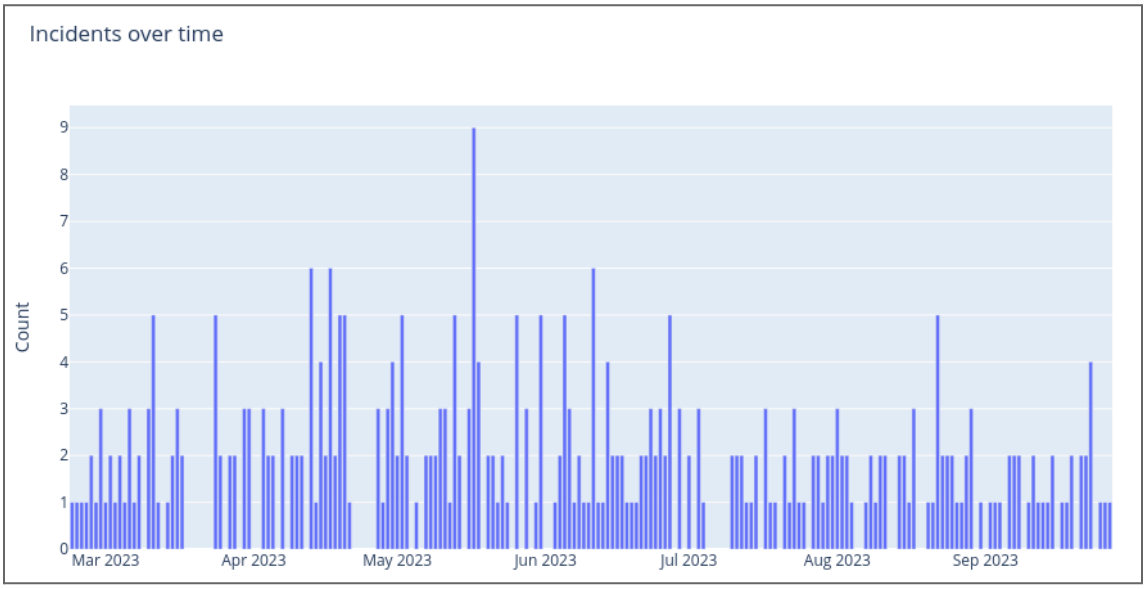

Figure 3.3. Incident count over time.

To better understand this pattern, operator interviews conducted in July 2022 provided insight into how operators reason through dispatch assessment, identifying the contextual variables they rely on.

These interviews revealed that operators rely on a range of contextual variables to assess whether an incident may escalate, and their accounts illustrate how multidimensional that assessment is in practice. P4, city level agency, described the components they consider:

> *"Some of the components that go into assessing the traffic impact are definitely going to be location, and then at that location, how many lanes are being blocked, how many directions are being blocked too... And another component is also like*

> *what time of day is this occurring? Is this occurring during an AM or PM peak, or is it just the middle of the day lunchtime and there's not much traffic going on? So those are all pretty big components that are going to determine whether we're going to have to coordinate with other agencies."*

This account illustrates how operators assess incidents not through a single trigger but through a layered reading of spatial, physical, and temporal conditions simultaneously. P5, city level agency, captured the nature of that assessment directly:

> *"Most of the time it's a judgmental call, but I mean, I think the question to be answered is: what's the traffic impact? I think that's when we should respond quickly."*

This framing, classification as judgment rather than rule-following, marks the boundary the rules engine was designed around. The obvious cases are automated; everything that requires this judgment is left to operators. Many operators also noted that certain types of events evolve differently depending on surrounding conditions. For example, a seemingly minor incident near a freeway entrance during rush hour may have very different consequences than the same event at an off-peak time. The following variable analysis draws on both Incident Model data and operator accounts to examine where automated detection ends and what reasoning operators apply beyond it.

3.4.1 Variable Analysis

The following analysis examines four variables that emerged as most significant in both the Incident Model data and operator accounts: location, event type, time of day, and additional contextual factors. For each variable, quantitative patterns from the data are presented alongside operator accounts that explain the reasoning behind those patterns.

Location

Location emerged as the most critical variable in operator assessment of dispatch events. All five participants identified location as a key factor in their classification decisions, and four of them, P1, P2, P4, and P5, named it as their first consideration when reviewing a dispatch event. Figure 3.4 displays all Incident Models in the Seattle Area, with user-generated incidents shown in blue and system-generated incidents in red. Incident Models are distributed across different areas from Seattle to Everett, mostly on highways. System-generated incidents are concentrated closer to the I-5 corridor, consistent with the geographic constraints of the current rules.

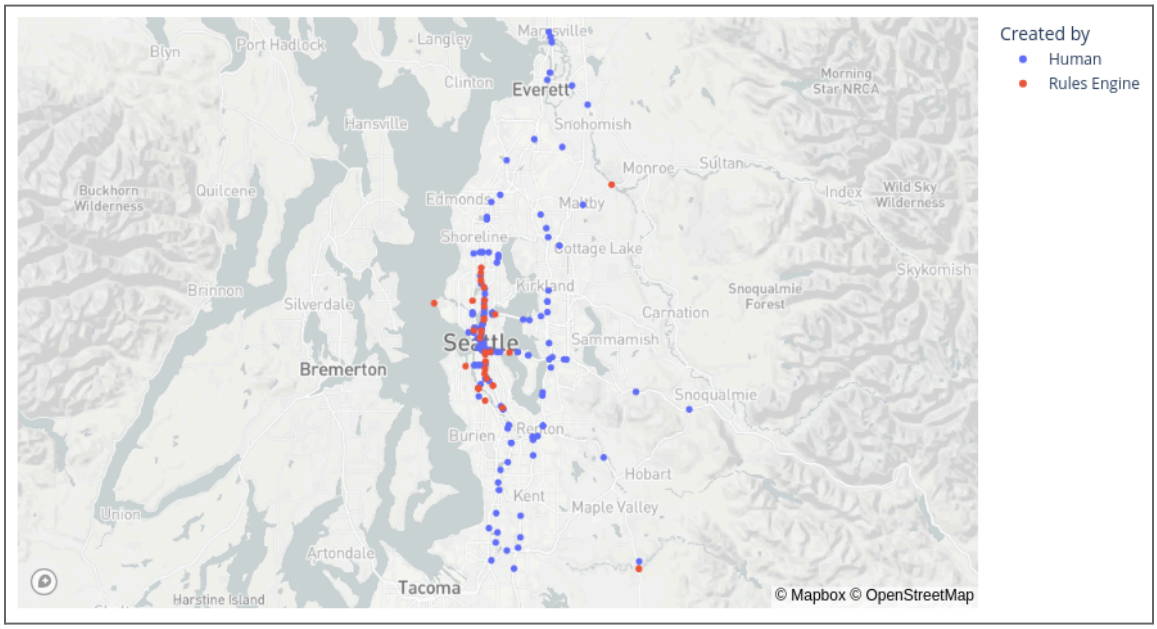

Figure 3.4 Incident Models in the Seattle Area

Figure 3.5 shows a heatmap of Incident Model locations across the study period, illustrating which areas had the highest density of incidents. The concentration of user-generated incidents extends well beyond the fixed geographic boundaries encoded in the current rules, suggesting that operators apply broader location-based reasoning when assessing whether a dispatch event warrants cross-agency coordination.

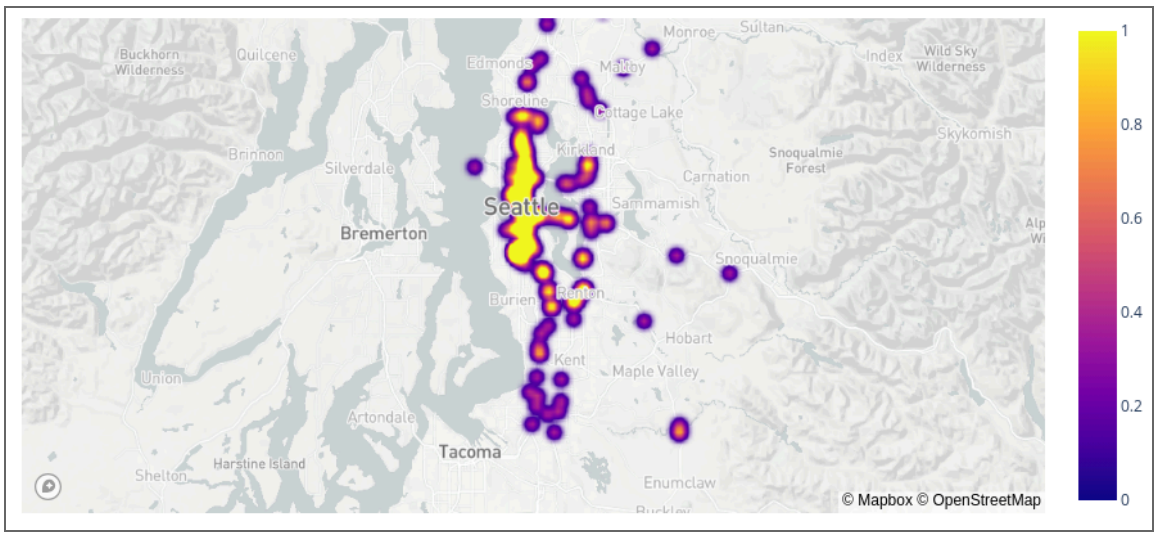

Figure 3.5 Heatmap of Incident Models

Operators described location not as a fixed point but as a spatial relationship requiring active interpretation. P1, state level agency, explained their process when reviewing a dispatch event:

> *"And then I'm usually scanning for... if it has any lane description, like which lanes are blocked. Because when I look at the location... it makes me wonder, is it on the CD [Collector-Distributor] or is it just like a cross street that spilled onto the main line? So the lane information is what I'm looking for next."*

This response illustrates how operators read locations through a layered lens, considering not just where an event is occurring but how its position relative to key infrastructure, lanes, and access points shapes its likely impact. The current rules flag events through fixed geographic criteria such as Area I-5 or bridge locations. Incidents in high-density or high-impact locations beyond those boundaries fall to operator detection. The map gives operators the spatial data to do that work, but the system neither supports the relational reasoning they apply to it nor communicates which locations its rules monitor and which it leaves to them.

Event Type

System-generated incidents are concentrated around the event types specified in the current rules, including Traffic Hazard Blocking Roadway, Car Fire Freeway, Fire Response Freeway, Road Closure, Fatal Traffic Collision, and Disabled Vehicle Fire. These incidents are typically closed or deleted within 20 to 30 minutes of creation, while VCC-level incidents last 90 minutes or more, illustrating why the rules were kept narrow: the event label identifies the clearly critical cases, while escalation beyond them depends on surrounding conditions.

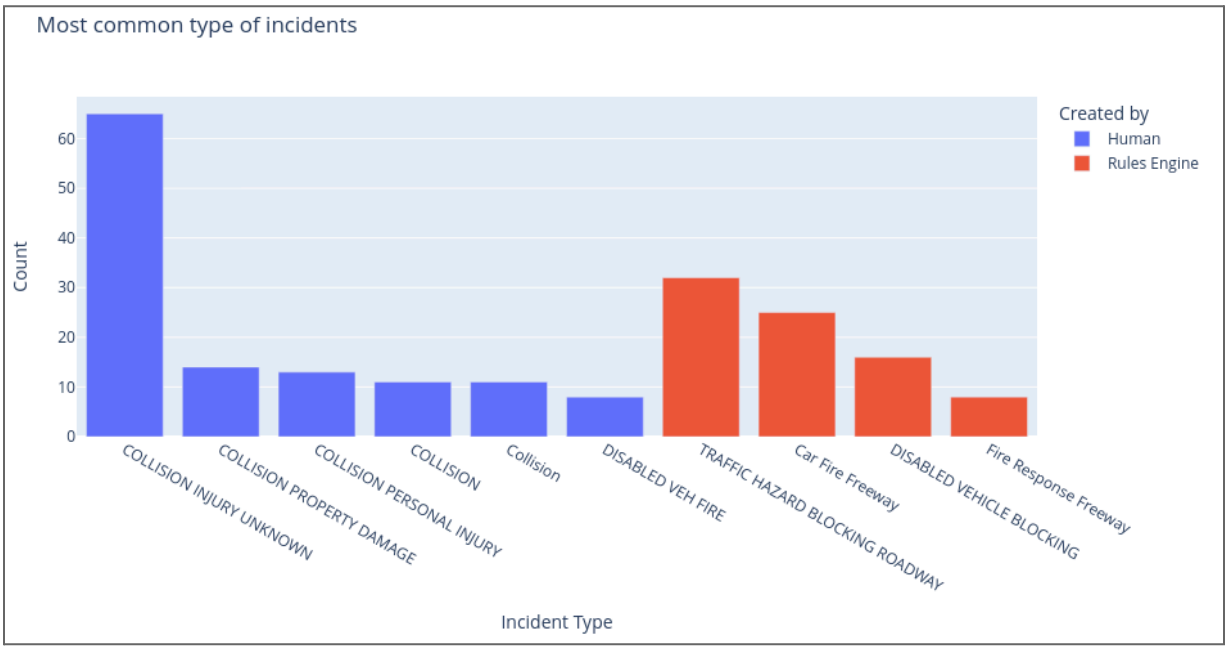

Figure 3.6. Most Common Incident Types

Figure 3.6 shows the most common event types for user-generated and system-generated incidents. For user-generated incidents, the most common types are collision-focused: Collision injury unknown, collision property damage, and collision personal injury. None of these collision types are currently included in the rules, consistent with their design: a collision's VCC-level status depends on context, not category. The data shows how much of the

classification work this leaves to operators, who assessed these events by weighing location, unit response, and time of day together.

Operators described event type as one of several interdependent signals rather than a standalone criterion. P2, regional level agency, explained how these signals work together:

> *"Location, where it is on the freeway is going to be important to us... and then what it is. So event type, location, those are going to be our most important things. And then we would look at... active units responding... In general, what it is, where it is, and how other agencies are treating it as a response, so how many units."*

This account illustrates that event type is never assessed in isolation. Its significance depends on where the event is occurring and how other agencies are responding to it. A collision on a major freeway interchange carries different implications than the same event type on a side street, a distinction that depends on the contextual judgment the rules leave to operators.

Time of Day

Interviews and qualitative analysis of Incident Models revealed that time of the day plays an important role in the urgency of a dispatch event. Traffic conditions vary widely across the day, and the same event carries different consequences depending on when it occurs.

Figure 3.7 shows the frequency of Incident Models at different hours of the day, revealing two notable concentrations. A smaller cluster appears in the late night and early morning hours, between midnight and 2am. Incident frequency then drops significantly during the late morning hours between 9am and 11am. A more sustained and substantial peak begins around noon and reaches its highest point in the mid-afternoon around 3pm, consistent with PM commute traffic patterns and operator accounts of how peak hour traffic amplifies the impact of dispatch events.

These patterns suggest that time of day shapes not just the frequency of incidents but the conditions under which they escalate.

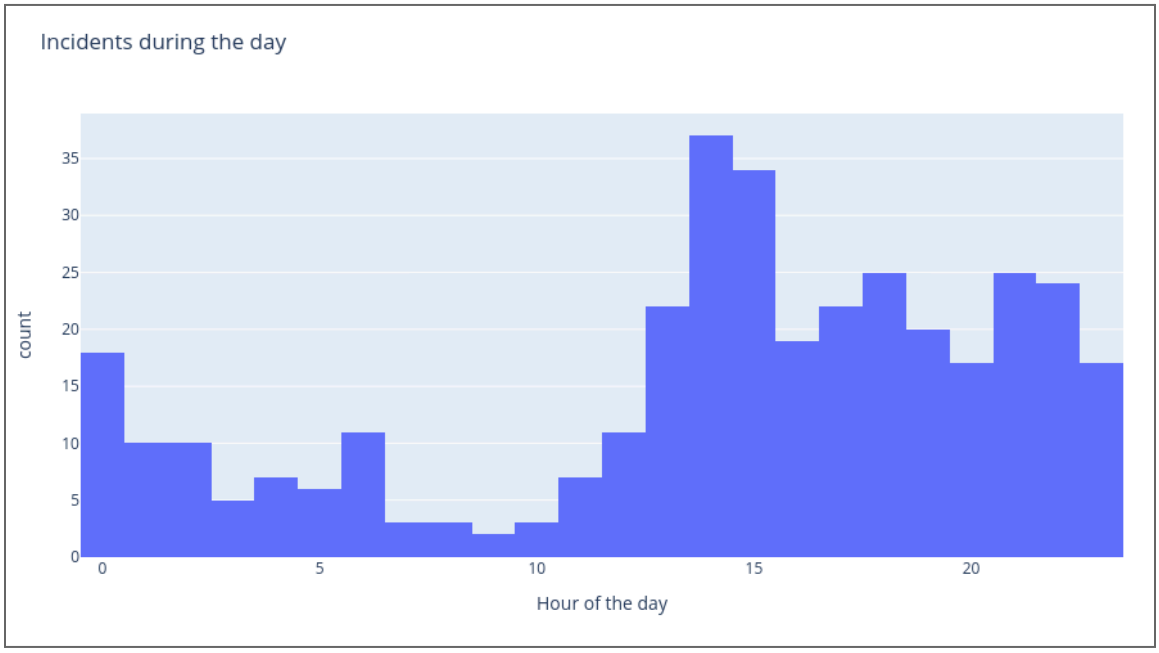

Figure 3.7. Frequency of Incident Models

Operator accounts further illustrate how time of day shapes the assessment of incident severity. As noted in the opening of this section, P4, city level agency, explicitly identified AM and PM peak hours as key factors in determining whether an event warrants cross-agency coordination, consistent with the afternoon concentration visible in Figure 3.7. P5, city level agency, echoed this directly, noting that a multi-lane freeway closure at peak hour would inevitably draw in the regional level agency. This account illustrates how time of day is not assessed as an isolated variable but as a condition that amplifies the consequences of other factors, particularly location and event type.

Additional Contextual Factors

Analysis of Incident Model data and operator interviews surfaced additional variables beyond location, event type, and time of day that inform how operators assess the seriousness of a

dispatch event. Three of these variables, multiple dispatches, dispatch duration, and number of updates, were examined quantitatively in the Incident Model records. A fourth, the availability of CAD notes, emerged exclusively from operator accounts as central to how operators assess incidents.

Multiple Dispatches: Operators noted that when a significant incident begins, multiple dispatches referring to the same event often arrive within a short window of time, sometimes from different agencies and nearby locations. Analysis of past Incident Models confirmed that clusters of related dispatches within defined time and distance ranges are a meaningful signal of VCC-level events.

Duration of the Dispatch: The duration of a dispatch event is an important factor in determining the likelihood that it indicates a VCC-level incident. Figure 3.8 shows the distribution of dispatch event durations across the study period, with the 90-minute threshold marked by a dashed line. The majority of events fall well below this threshold, but a meaningful tail of longer-duration events extends beyond it, consistent with operators' accounts that incidents requiring cross-agency coordination tend to persist significantly longer than routine dispatches. This 90-minute threshold is not arbitrary. P1, state level agency, identified it explicitly as part of their operational definition of a VCC-level incident: "*Any incident that blocks the lane for over 90 minutes.*"

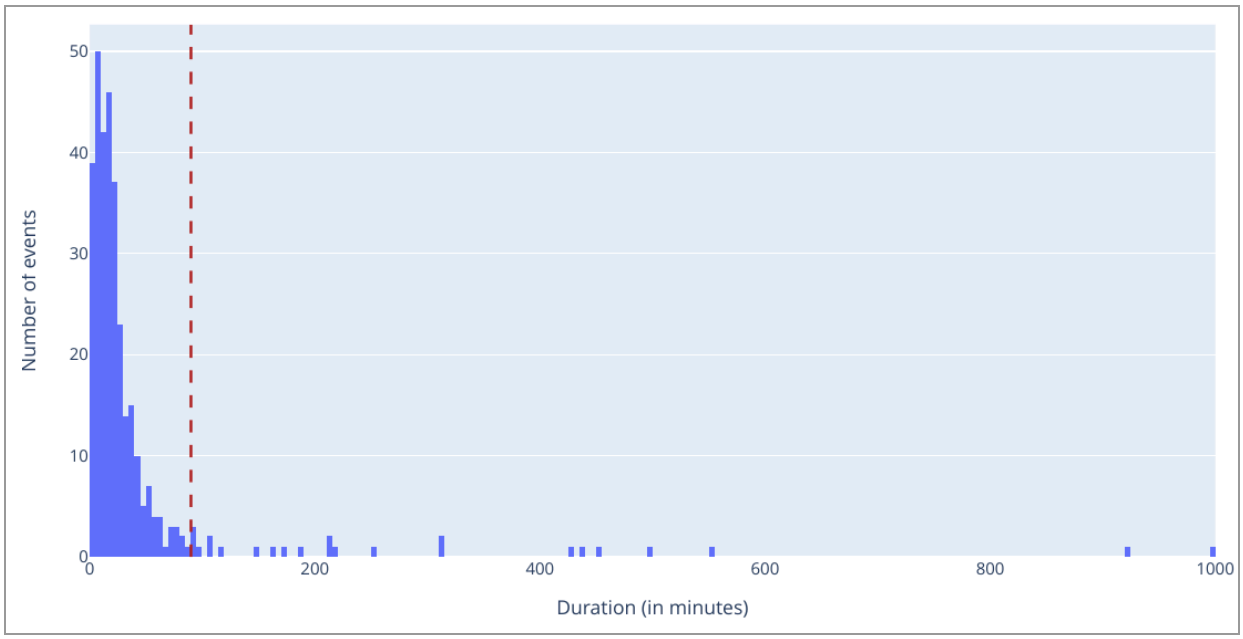

Figure 3.8 Distribution of dispatch event durations.

Analysis of closed Incident Models further illustrates this pattern. Of the 217 user-generated closed IMs, 109 (50%) lasted longer than 90 minutes. Of the 85 system-generated closed IMs, 46 (54%) lasted longer than 90 minutes, suggesting that within its detection scope, the rules engine identified incidents of similar duration to those captured by operators.

Among WSP events on the I-5 corridor, approximately 62 lasted more than 90 minutes, while the rules engine detected about 21 of them, consistent with criteria scoped to event types rather than outcomes. Duration is not a straightforward classification variable. Some Incident Models remain open for extended periods not because the incident is ongoing but because they have not been closed or reviewed. However, duration does serve as a useful retrospective signal. Incidents that appear moderate at the outset but persist beyond 90 minutes may warrant VCC-level attention, particularly when lane blockages continue on main highways or arterial roads and cross-agency coordination becomes necessary. In this sense, duration is most

meaningful when considered alongside real-time variables such as the number of updates and the level of cross-agency activity.

Number of Updates: While dispatch duration is informative, the number of updates to an Incident Model provides a more reliable signal of whether an incident is genuinely ongoing. Analysis of Incident Models with long duration showed two distinct patterns: models that remained open with few updates, suggesting the incident had cleared but the model had not been closed, and models with a high number of updates, reflecting active coordination and ongoing resource needs. Where incidents were genuinely ongoing, the frequency of updates corresponded with the need for continued cross-agency collaboration. This variable is not part of the rules engine but represents a meaningful indicator of incident complexity that could complement event type, location, and other contextual variables.

CAD Notes: A consistent finding across operator interviews was the critical importance of Computer-Aided Dispatch (CAD) notes for incident assessment and cross-agency situational awareness. CAD notes capture real-time field information including injuries, lane status, and responding unit details that operators rely on to assess severity and coordinate responses. P4, city level agency, captured this directly:

> *"The notes are just crucial in providing situational awareness as it unfolds."*

This account illustrates that CAD notes are not supplementary information but a core input to how experienced operators reason through classification decisions. CAD notes are where much of the situational reasoning operators rely on actually lives. Making that information more visible and synthesizable within the shared environment would support the classification work the rules leave to operators.

3.4.2 Themes from Operator Interviews

Beyond the variable-level findings, operator interviews surfaced three cross-cutting themes that speak directly to what any decision support system in this context should account for. These interviews captured how operators reason through incident classification in their existing workflows, surfacing expert knowledge and operational practices that structured dispatch data alone cannot capture. Together they characterize the reasoning the system leaves to operators and the expertise that reasoning depends on.

Agency-Specific Definitions of VCC-Level Incidents

A consistent pattern across all five interviews was that operators did not share a single definition of what constitutes a VCC-level incident. Instead, each agency applies criteria shaped by its operational scope, service responsibilities, and risk tolerance. This variation is not a failure of coordination, it reflects the legitimate differences in what each agency manages and who it serves.

P1, state level agency, applied the most structured definition, grounded in numerical thresholds:

> *"Anything that requires multiple agencies to respond. And anything that has a large impact to the traveling public… more than 50% of the lanes blocked on the [*state level agency*] freeways, or there's a fatality, or there's a hazmat incident like a chemical spill, or an incident involving a semi… or any incident that blocks the lane for over 90 minutes."*

P1 also noted that even these thresholds are not absolute, acknowledging that political factors can override operational criteria:

> *"There's always the political nature of things, right? Like if the mayor of a city is going through that particular incident and is stuck in traffic, and that particular person has to go to this very politically hyped-up event... we have to start a*

> *VCC-level incident for it, just so that everybody's in the know, and that everything's documented, and we're doing everything we can to get that particular person through this congestion. So there's always wrinkles to every incident."*

This observation reveals that even within a single agency, classification involves judgment that no rule can fully encode.

P3 and P4, city level agency, framed VCC-level incidents through the lens of compounding effects and multi-agency coordination rather than fixed thresholds. P3 described an incident as VCC-level when it represents "*a significant reduction in the capacity of a major facility*" requiring multi-agency response, while P4 emphasized that coordination needs vary by location; an incident on a transit corridor may require regional level agency involvement but not necessarily state level agency.

P2, regional level agency, applied a service-based definition directly tied to their agency operational responsibilities:

> *"For a VCC-level incident for us, it would be anything that is blocking a major arterial and that would be for a longer period of time… affecting our service in a way that would delay [vehicles] or cause them to be rerouted for probably longer than 15 or 20 minutes."*

P2 also noted that the same event type can range from routine to VCC-level depending on additional details. A disturbance might be unremarkable, but if it involves a weapon, it can immediately stop a vehicle in an intersection, escalating to a high-priority situation requiring cross-agency coordination.

P5, city level agency, framed VCC-level incidents through their direct traffic impact, emphasizing the cascading effect on the broader network:

> *"If you see that there is a multi-lane closure on a freeway at a peak hour heading out of downtown, that's definitely going to impact major traffic and [regional level agency] is going to be involved."*

For P5, the threshold is less about the event itself and more about its downstream consequences for city streets and cross-agency coordination.

Taken together, these accounts illustrate that VCC-level incident classification is not a single judgment but a family of judgments shaped by agency role, service geography, and operational priorities. The VCC definition of a VCC-level incident as a situation requiring multi-agency coordination provides a shared umbrella, but how each agency determines when that threshold has been reached differs by operational context. These differences in definitions are not a problem to be solved, each agency's criteria reflect legitimate operational needs that must be preserved. What this finding points to is the value of a decision support system that can accommodate multiple valid definitions simultaneously, surfacing relevant information for each agency's context without imposing a single classification standard across all of them.

Classification as Contextual Judgment

A second theme across all five interviews was that incident classification is not a single-step decision based on one data source, but a contextual judgment that draws on multiple streams of information simultaneously. Operators described synthesizing dispatch data, camera feeds, CAD notes, field command updates, and direct agency-to-agency communication to build a picture of what is happening and whether it warrants cross-agency coordination. This multi-source reasoning process is a form of expertise that any decision support system in this context must be designed to support.

P1, state level agency, described using cameras not as a primary information source but as a verification layer on top of dispatch data:

> *"We don't really gather information from the camera itself. It just helps us verify the lanes that were reported to us from dispatch, like which lanes are blocked. Is the location accurate? It helps us find it and verify that it is there"*

The verification step itself requires judgment, knowing which lanes matter, what the location implies, and whether the dispatch description matches what the camera shows.

P2, regional level agency, described a unique advantage their agency brings to this process. With approximately 1,300 vehicles on the road during peak hours, they often have early situational awareness of blockages before other agencies do. However, P2 also noted a fundamental limitation in their own internal system:

> *"And I'm not sure that anything that we put into the CSR [Coordinator Service Records] that the VCC could see would actually be able to pick up on that, because it may just be that they called the on-call chief, and that to us internally means that's a major event and we have to kind of go up the chain of command. But for the VCC to recognize something like that... what we have right now doesn't show necessarily the severity of it."*

This account reveals that at the time, critical escalation signals such as a phone call or an internal chain of command decision existed entirely outside structured data systems.

P3, city level agency, identified field command decisions as a critical information gap. When decisions are made on scene, such as reopening a lane, that information does not flow quickly into the shared environment, leaving operators unable to update their responses in real time:

> *"That information sometimes doesn't flow quickly from the field to everybody. Getting that sort of incident command information is useful because it lets us change our DMS [Dynamic Message Signs]."*

P4 and P5, city level agency, both pointed to CAD notes as essential for situational awareness, as discussed in the variable analysis. P5 also described navigating the dispatch feed visually, piecing together what is happening by clicking on map pins near known congestion areas:

> *"I might look at a map to see where the pins are, to see where it's close to the freeway. And then I might click on a few to see if it's causing traffic."*

Across all five participants, the picture that emerges is of operators who are skilled at synthesizing incomplete and heterogeneous information to make confident classification decisions. This is not a workaround, it is their workflow and expertise. A decision support system that surfaces only structured dispatch data will always present operators with an incomplete picture. The design opportunity this finding points to is a system that actively integrates and makes visible the full range of contextual signals operators rely on, supporting and extending the multi-source reasoning operators already bring to classification.

Operator Expertise as Institutional Memory

A third theme that emerged across interviews was that the expert knowledge operators bring to classification decisions is not captured or preserved anywhere in the system. The rules engine was originally developed through consultation with experienced operators, meaning that expert reasoning shaped its initial design. However, the logic of those rules is static and not visible to users, and the contextual reasoning behind them is not documented or communicated to operators who were not part of that process. When a system-generated Incident Model appears, operators have no way to see why it was created, which limits how confidently even experienced users can act on it.

This gap has different implications depending on operator experience and agency role. P1, state level agency, demonstrated throughout the interview a highly structured classification

framework built over years of operational experience, including numerical thresholds, political considerations, and agency-specific protocols.

P2, regional level agency, brought 20 years of experience to the interview, having started in a direct service operator role and worked up to Chief. His account of how their agency reads escalation signals reflects deep institutional knowledge that is not recorded anywhere in the system. Recognizing that a call to the on-call chief signals a major event is expertise that lives entirely in practice.

P3, city level agency, described compounding effect reasoning that comes from operational experience rather than any documented protocol. Knowing that an incident on a major arterial becomes more serious when I-5 is already closed, because the arterial serves as the detour route, requires familiarity with the regional network that newer operators may not have.

P4 and P5, city level agency, both described using notes within dispatch data as a primary mechanism for sharing and understanding situational awareness and reasoning across agencies in real time. Usage data from the study period confirms this pattern, with 316 notes attached to Incident Models, suggesting that operators actively use available mechanisms to communicate contextual knowledge that may not fit into structured data fields.

Across all five participants, expert knowledge emerged as something that is actively practiced, informally shared, and deeply contextual. The design opportunity this finding points to is a system that makes this reasoning more visible, more transferable, and more accessible across varying levels of operator experience.

3.5 Design Implications

Incident classification in the VCC is divided between automated detection and operator judgment. The division is intentional: the rules engine handles a narrow set of high-certainty events, and operators carry everything that requires contextual reasoning shaped by agency role, experience, and real-time information. What this study found is that the division is invisible. The system does not communicate which classifications its rules cover, why a system-generated Incident Model was created, or how it differs from one an operator launched. At the same time, the reasoning that operators apply beyond the rules exists only in practice. Without that visibility, even experienced users cannot fully apply their knowledge to interpret system behavior, and the expertise the operation depends on goes uncaptured. Both gaps carry implications for trust and expertise diversity.

From our findings, the following design priorities emerge:

- **Design for context integration:** Future iterations should incorporate contextual factors, such as time, location, and dispatch frequency into the system to support operator assessment of incident severity.
- **Augment existing visual reasoning:** Operators already engage in complex, multi-source reasoning, cross-referencing map pins, scanning CAD notes, and monitoring camera feeds to build context. Design interventions should aim to synthesize these fragmented data streams into unified, interactive, visual spaces.
- **Enable system interpretability:** Users should be able to see why the system generated an alert or did not flag an incident, with clear explanations that support decision confidence and learning.

- **Embed and share expert reasoning:** Decision support tools should retain and communicate the strategies developed by experienced operators to support knowledge transfer across users.
- **Support adaptive feedback:** The system should be responsive to ongoing use, allowing both its detection criteria and its support features to be revisited as operational needs evolve.

Given these needs, visual tools offer a promising design direction, not as an interface add-on, but as an extension of the work operators already perform. As this study revealed, operators naturally engage in complex, multi-source visual reasoning, pointing toward the need to design visual interfaces that synthesize these data sources into unified, interactive visual systems. By serving as an interface between the system's rules and the operator's contextual judgment and expert knowledge, visualizations can leverage the strengths of the human visual system to reveal underlying patterns and make system logic more accessible without disrupting established workflows. Visualization stands out as a compelling path toward building more transparent, interpretable decision support systems that serve a range of operator expertise and roles.

Improving decision support in the VCC points toward augmenting the rules engine alongside a new visualization layer. The rules engine could be extended to communicate its own scope, surfacing which event types it covers, why a system-generated Incident Model was created, and where its detection ends. A visualization layer could then bring together the contextual signals operators already track across different sources, location, time of day, dispatch clusters, and update frequency, into a unified interface that supports the operators' judgment.

These findings establish the foundational need this dissertation responds to: decision support systems that cannot explain their logic or surface expert reasoning will struggle to build trust and serve users across varying levels of expertise. This study also demonstrates the value of

evaluating decision support from two perspectives simultaneously, examining what the system does and what operators independently bring to the process, and identifying where the rules engine's scope ends and operator judgment begins. The studies that follow apply this evaluative approach in new domains and with different user populations, exploring how interactive visualizations within DSSs can enhance this complex, multi-source reasoning, supporting trust and expertise diversity in practice.

Chapter 4: Visual Design Features in Long-Term Planning Visualizations

4.1 Project Background and Context

4.1.1 Regional Growth and the T2050 Initiative

By 2050, the Central Puget Sound Region is projected to grow by 1.5 million people, bringing the population to approximately 5.8 million. Supporting this growth will require major infrastructure decisions: improvements to the I-5 corridor, the introduction of ultra-high-speed rail, and the addition of a third commercial airport. These choices carry long-term consequences for mobility, equity, and environmental impact across the region.

The Transportation 2050 (T2050) platform was developed through the Mobility Innovation Center (MIC) at the University of Washington to help the public and policymakers explore these future transportation challenges. Rather than relying on static reports or abstract planning models, T2050 uses interactive, scenario-based visualizations to enable deeper public understanding and informed planning, inviting users to engage directly with data by comparing conditions across time and exploring the consequences of current trends and proposed interventions. The platform embeds these visualizations within a narrative sequence, providing contextual statistics and framing that guides users through the regional planning challenges the data illustrates.

Development of the platform began in Fall 2024 with data collection across multiple transportation agencies, followed by a series of stakeholder meetings to identify key planning challenges, system constraints, and user information needs. These early engagements directly

shaped the platform's direction. Subsequent phases included data cleaning, integration, and initial visualization design, with frequent feedback loops from stakeholders throughout. Visualizations were refined based on ongoing review and input from partners. The project was conducted in collaboration with WSDOT, King County, Challenge Seattle, Microsoft, Boeing, and Alaska Airlines, whose participation ensured that the platform addressed the priorities of both public agencies and major regional employers.

This human-centered development process, grounded in stakeholder engagement and iterative design, established the foundation for the formal user evaluation described in this chapter. That evaluation turns from how the platform was built to whether it actually works for the people it was designed to serve, examining which visualization features support understanding, confidence, and engagement across user groups with varying roles and levels of expertise.

4.1.2 Platform Overview

At the time of this study, the T2050 platform brought together a set of interactive visualizations, each drawing on regional planning data to illustrate a distinct dimension of Washington's transportation future. This chapter focuses on four visualizations selected for analysis, covering demographic change, commute conditions, bridge infrastructure, and regional travel mode alternatives, chosen to represent a range of design complexity, interactivity, and intended user targets across the platform.

The Population Growth visualization (Viz 1) uses county-level population projections from the Washington State Office of Financial Management (2022a, 2022b) to illustrate how population growth has evolved across the state and how it is expected to continue through 2050. A color-coded choropleth map with a year slider (1962-2050) enables exploration of spatial and temporal patterns over time (Figure 4.1). Two linked line graphs show population and growth

rate over time, highlighting steady statewide population increases and underscoring the mobility challenges that will intensify without proactive improvements.

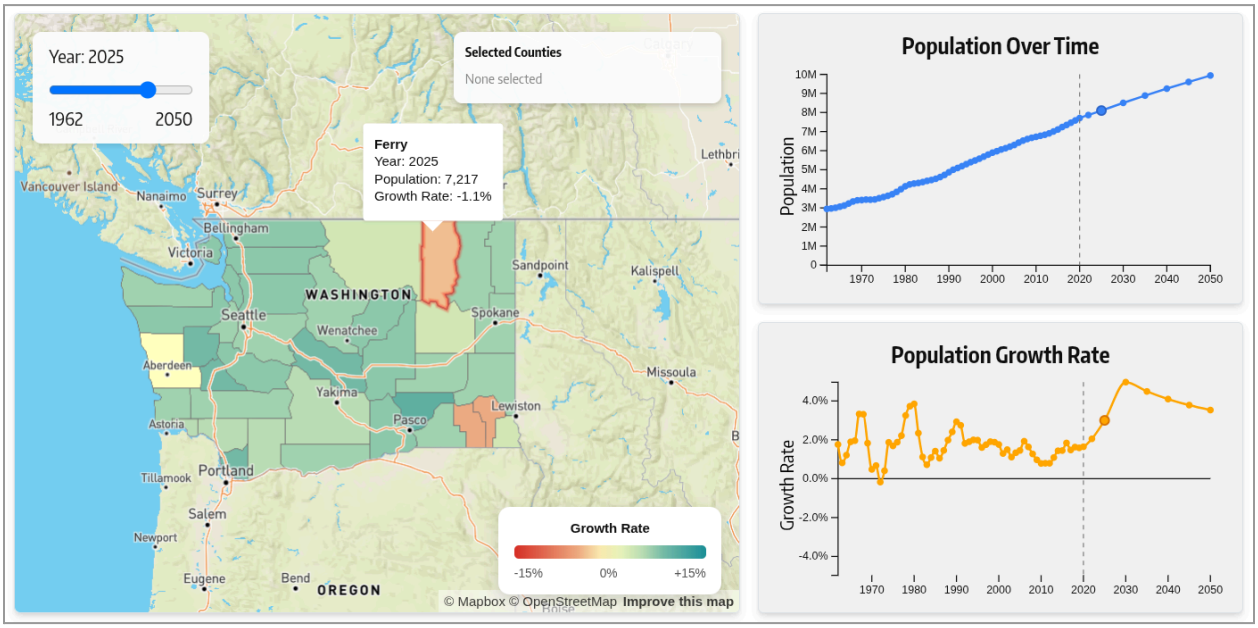

Figure 4.1. Viz 1: Population Growth Visualization.

The Commute Travel Times visualization (Viz 2) integrates travel demand models from Puget Sound Regional Council (2024) and Thurston Regional Planning Council (2022), to provide a personalized view of current and projected 2050 commute conditions. Users can select a model from a dropdown menu and enter origin and destination addresses to generate a customized route. Two side-by-side maps display present-day and forecasted travel routes and times (Figure 4.2). By grounding the data in users' own travel context, this visualization highlights the individual-level impacts of future mobility conditions.

The Bridge Conditions visualization (Viz 3) uses bridge inventory and needs data from WSDOT (2024) to illustrate the current condition of bridges across the state and their associated improvement costs. A scatter map displays every bridge in Washington State (Figure 4.3), color-coded by condition (Good, Fair, or Poor), with county-level filtering available. A right-hand panel displays the total number of bridges and total improvement costs for the

selected area, while two bar charts below show the distribution of bridge conditions and detour distances. This visualization highlights the scale of current maintenance needs and the mobility impacts of deteriorating infrastructure, underscoring the importance of timely investment decisions.

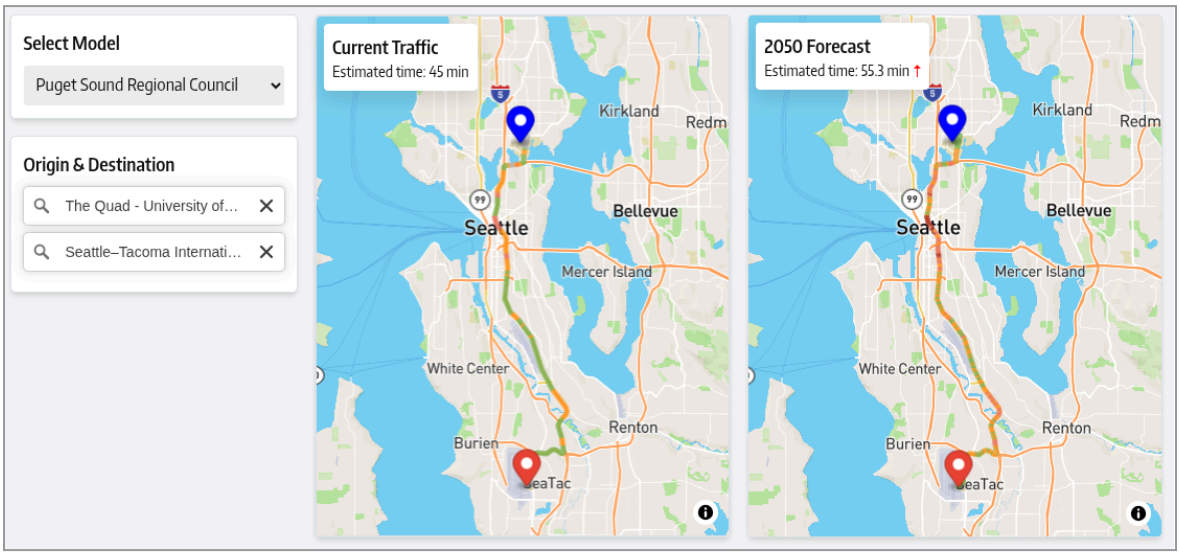

Figure 4.2. Viz 2: Commute Travel Times Visualization.

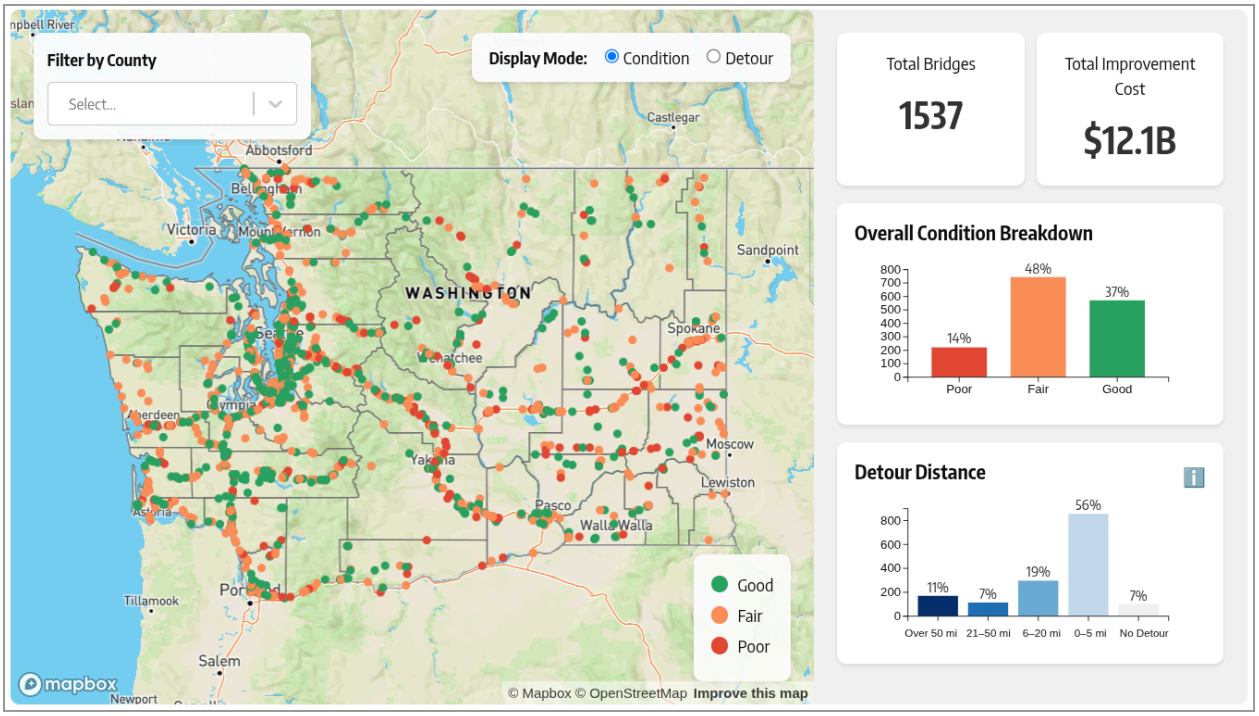

Figure 4.3. Viz 3: Bridge Conditions Visualization.

The High-Speed Rail visualization (Viz 4) uses data from WSDOT's Ultra-High-Speed Ground Transportation Study (WSDOT, 2019) to compare travel times between train, car, air travel, and high-speed rail for routes connecting Vancouver, Seattle, and Portland. Users can select origin and destination cities, and the visualization updates to display a horizontal bar graph showing travel times for each mode of transportation (Figure 4.4). This visualization illustrates the potential impact of implementing high-speed rail on regional travel times, highlighting how this infrastructure investment could improve mobility between these major population centers.

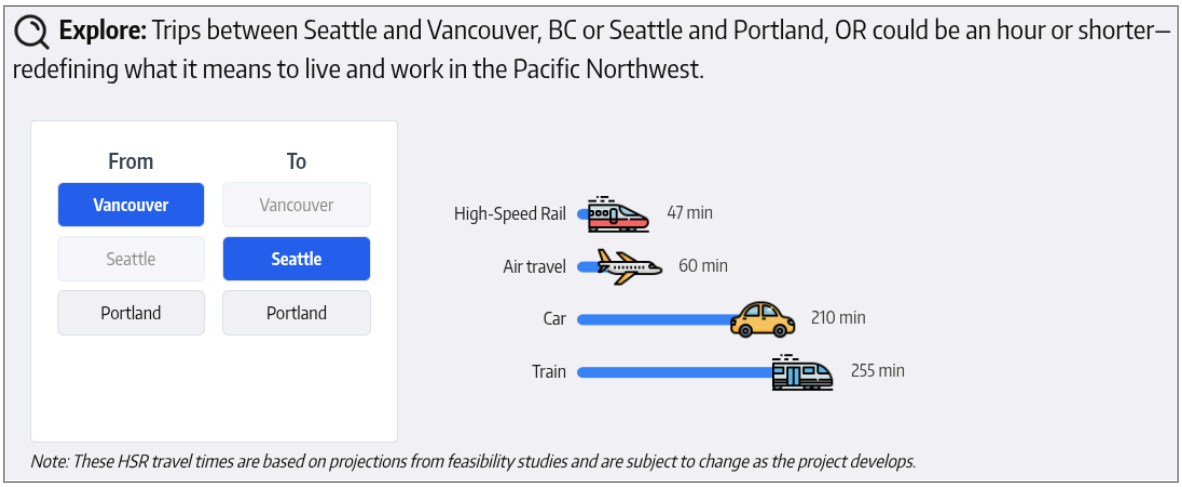

Figure 4.4. Viz 4: High-Speed Rail Visualization.

Together, these four visualizations represent a range of design complexity and interactivity, from a simple dynamic comparison chart to a multi-layered spatial analysis environment, providing a rich basis for examining how users with different levels of expertise engage with long-term planning data.

4.1.3 T2050 as a Site for Studying Visual DSSs

Regional transportation planning falls at the wicked end of the decision problem spectrum, as discussed in Section 2.1.1, characterized by multiple stakeholders with conflicting values, no

discrete solution, and problems that are never fully resolved but continuously addressed as conditions evolve (Rittel & Webber, 1973).

T2050 was designed to support this kind of problem. Rather than producing a single recommendation or optimal solution, it enables users to interpret projected conditions, explore alternative scenarios, and form judgments about long-term transportation challenges. This positions T2050 as a procedural DSS, a system that facilitates navigation of a complex decision space rather than resolving it (Pretorius, 2017). The decision it supports is not made at the moment of interaction but through the understanding that interaction builds over time.

This makes T2050 a particularly productive site for studying trust and expertise diversity in visual DSSs. Because the platform was designed for a wide public audience rather than specialized operational stakeholders, it draws in users with very different levels of domain knowledge, familiarity with data visualization, and personal stakes in the decisions being explored. This is precisely the condition under which design choices are most likely to reveal where trust is built or broken and where expertise diversity is supported or excluded.

4.2 Research Objectives

T2050's design as a public-facing platform draws in users across a wide range of domain expertise, creating exactly the evaluative conditions this dissertation addresses. Examining how that range of users engages with the platform's visualization features requires looking at both what the system was designed to do and what users bring to the interaction.

Study 1 pointed toward visualization as a promising direction for making system logic more accessible and surfaced the value of evaluating decision support from two perspectives simultaneously, what the system does and what users independently bring. Study 2 extends this to a broader expertise range, from transportation professionals to members of the general public,

focusing specifically on examining the visualization features that support or undermine trust and expertise diversity.

This study pursues three interrelated questions:

RQ1. How do T2050's visualization features support or create barriers for interpretation and trust across the expertise range?

RQ2. What does the system-side component of the Chapter 2 preliminary theoretical framework reveal when applied to T2050?

RQ3. What does combining the system-side and empirical analyses reveal about how to evaluate trust and expertise diversity in visual DSSs?

This study contributes to the understanding of how visualization features shape trust and expertise diversity in visual DSSs and establishes initial evaluation criteria for a human-centered evaluation framework through the combination of a system-side analysis and empirical user data. The findings also offer practical insights for the continued refinement of the T2050 platform.

4.3 Methods

To examine how visualization features support or undermine trust and expertise diversity in a public facing planning platform, this study combines think-aloud user testing with semi-structured interviews and pre- and post-study surveys. Survey data provides a structured baseline for understanding who participants are before they interact with the platform and how their attitudes shift afterward.

While the T2050 platform includes multiple visualizations, the empirical and system-side analyses focused on four that together represent the range of design complexity, interactivity, and intended user targets across the platform:

Visualization	Primary User Target	Visual Complexity	Interactivity
Viz 1: Population Growth	Both	Moderate	High
Viz 2: Commute Travel Times	Non-expert	Low	Moderate
Viz 3: Bridge Conditions	Expert	High	High
Viz 4: High-Speed Rail	Non-expert	Low	Low

Table 4.1. T2050's visualizations selected for analysis.

4.3.1 Study Design and Participants

Participants were recruited primarily from the Puget Sound region through the University of Washington's Mobility Innovation Center (MIC) and social media platforms, with residence within Washington State as the only eligibility requirement. The study distinguished between expert and non-expert users based on domain expertise, defined here as participants' professional background, institutional knowledge, and familiarity with transportation planning data and infrastructure. Participants were classified as experts or non-experts based on self-rated transportation expertise and occupational background collected through a screening survey. Where needed, assessment from the transportation expert who facilitated participant recruitment informed classification.

The sample included:

- **Expert users** (n=6, P6 through P11): transportation professionals, urban planners, and policymakers with high or expert-level transportation knowledge.

- **Non-expert users** (n=5, P1 through P5): members of the general public whose transportation knowledge is grounded in lived experience as commuters and residents, without formal training or professional practice in the domain.

Visualization literacy varied independently within both groups and is not controlled for in the study design. Patterns attributed to domain expertise reflect differences in professional knowledge rather than general data literacy.

Nielsen (1994) demonstrates that small samples, often around five participants, can capture a substantial proportion of usability issues in relatively homogeneous user populations. When user groups are heterogeneous, however, additional participants are needed, as smaller mixed samples may underrepresent group-specific patterns (Caulton, 2001). With 11 participants including expert and non-expert users, this study sample size was selected to capture variation across both groups while remaining appropriate for in-depth qualitative analysis.

Data collection consisted of semi-structured interviews combined with think-aloud user testing, conducted via Zoom between June and August 2025. Sessions ranged from 30 to 60 minutes. Think-aloud protocols are particularly well suited to visualization evaluation, as they surface how users interpret and interact with visual representations in real time rather than retrospectively (Ericsson & Simon, 1980; Carpendale, 2008). Participants shared their screens for navigation tracking and provided informed consent for recording and transcription prior to participation. The first two sessions served as pilots, informing protocol refinement and are not included in the analytical sample; one additional participant was excluded due to prior exposure to the platform through involvement in its iterative development. The remaining 11 participants form the analytical sample reported here.

The study protocol consisted of participants independently exploring the visualizations and completing structured navigation tasks while verbalizing their thoughts and decision-making processes. Each visualization was paired with a specific task and follow-up questions designed to surface participants' interpretations of the data and their perspectives on long-term transportation challenges. The session concluded with three debrief questions reflecting on their navigation experiences and personal concerns about future congestion and travel time changes in the region (Appendix B.1).

For the purpose of this study, the platform was presented to participants in a stripped-down form. The full T2050 website embeds the visualizations within a narrative sequence, with contextual framing that guides users through the content. The study site removed this scaffolding, presenting each visualization with only a title and a brief exploration prompt, isolating the visualizations as the unit of analysis and allowing the study to examine how users engage with the tool independently of the narrative context in which it was originally designed to operate.

This study received an exempt determination from the University of Washington Human Subjects Division (IRB Study ID: STUDY00022911) under Category 2 of federal human subjects regulation (45 CFR 46.104). All participants were adults who provided informed consent prior to participation. Data was anonymized and stored on UW-approved systems; participation was voluntary.

4.3.2 Analytical Approach

This study analyzes the T2050 visualizations through two complementary analytical phases. The first applies reflexive thematic analysis to the user study data to surface how users across expertise levels made sense of and engaged with the visualizations. The second applies the

system-side evaluative questions developed in Section 2.5.1 across Munzner's four nested levels. The comparison between the two phases, presented in Section 4.5, produces an initial empirical refinement of the system-side evaluative approach.

Phase 1: Empirical Analysis

Think-aloud and interview data was analyzed using reflexive thematic analysis (Braun & Clarke, 2019). Reflexive TA was selected for its fit with interpretive research questions that seek to understand how people make sense of their experiences. This approach is well suited to the exploratory nature of this study and the rich, context-dependent data generated through think-aloud and interview methods. Analysis progressed through familiarization with the dataset, iterative coding, and theme development, oriented by the research questions around trust and expertise diversity, with ongoing reflection on developing patterns and interpretive decisions throughout.

As the designer and developer of the T2050 visualizations, I bring a specific positionality to this analysis, including familiarity with the platform's design decisions and the underlying data, and a potential bias toward interpreting user responses favorably in relation to those decisions. Another researcher led the formal user study sessions, providing distance between data collection and my design role. Throughout analysis, I actively reflected on how my design background shaped my interpretations and remained open to findings that contradicted my expectations.

The codes that resulted from this process are presented in Appendix B.2, organized into four sensitizing concepts that capture how users make sense of the data, how they evaluate the credibility of system outputs, how their expertise and lived experience shape interpretation, and how visualization design features support or hinder engagement.

Phase 2: System-Side Analysis

This evaluation directly addresses the second research question of this study, examining what the system-side evaluative questions reveal when applied to a public-facing visual DSS. Drawing on the evaluative questions introduced in Chapter 2.5.1, each visualization is examined across Munzner's four nested levels through immediate validation, examining the system's design choices, and downstream validation, synthesizing participant responses per Munzner level to test whether those design choices held against user experience. The immediate validation was conducted independently of the empirical analysis. The downstream validation then synthesized findings from the empirical analysis and interview data against the evaluative questions to assess their scope.

4.4 Findings

4.4.1 Phase 1: Empirical Findings

Findings by Theme

Five themes were constructed from the reflexive thematic analysis of the interview data, each capturing a different pattern in how participants engaged with and made sense of the T2050 visualizations. Across these five themes, the distinction between expert and non-expert participants was not in their capacity for engagement with the visualizations, but in the forms that engagement took, shaped by the contextual knowledge and domain familiarity each brought to the interaction.

1. Lived Experience as the Test of the Data

The most consistent pattern across the interviews was that participants, regardless of expertise level or domain background, drew on their own experiences with the transportation system when evaluating the visualizations, using personal trips, commutes, and knowledge of local infrastructure to confirm or question the data displayed in the visualization. For most participants, the first question was not what the data showed, but whether it matched what they already knew.

For non-expert participants, lived experience was the primary instrument for assessing whether the data was credible. P4 did not simply doubt the displayed air travel time to Vancouver, they also itemized every step the visualization had skipped:

> *"Air travel, this is a false number. So I would not agree with that number, because I've done that flight several times… when you factor in all of the time we're driving from my house to Seatac, parking, getting through security, walking to the gate, waiting… getting there an hour ahead of time, and then landing in Vancouver, getting out of the Vancouver airport and onto the train. It is more of a 2 hours. It's faster for us to drive than to take a plane."* (P4, non-expert)

P2 applied the same reasoning to the bridge conditions visualization, connecting the displayed condition ratings to a specific infrastructure failure they had experienced directly:

> *"I live in West Seattle, and when the pandemic happened I know that the upper bridge broke down because it wasn't maintained. So if it had been sitting in that fair condition for an extremely long time, and they weren't monitoring it. That's actually pretty scary."* (P2, non-expert)

The West Seattle bridge had sat in "fair condition" until it failed, and that experience shaped how P2 read every "fair" label on the map. Their concern was not just about one bridge but about what the category actually meant and if it could be trusted.

Expert participants drew on lived experience as well, but it was grounded in both personal use of the system and years of professional work within it. Where non-expert participants reached for a trip they had taken, experts reached for data they worked with directly, studies they had run, and lived experiences within the transportation domain. For some, that history became a direct basis for assessing credibility. P7, who had built a travel demand model for Thurston County, evaluated the population projections against what they had experienced through that process:

> "*I did travel demand modeling in this county. The first travel demand model that was built in Thurston County. I did that back in 91, and I remember the population projections then, and people feeling like that was shocking… so I do think that looks quite accurate.*" (P7, expert)

P10 brought the same logic to the population growth rate, testing the displayed numbers against a benchmark carried from their own professional work:

> *"When I was in traffic operations, we always used a standard growth rate of like 3%. So 5 seems higher than 3. But again, probably 3% was kind of the national average when I was in traffic operations 10 years ago, and my guess is that in Washington it probably is higher than 3%."* (P10, expert)

For expert participants, the personal and professional archives were inseparable. They did not just know the transportation system as users, they had worked inside it, and that shaped what they trusted.

2. The Numbers Don't Speak for Themselves

A number without context is a number without meaning. Across all visualizations, participants encountered numbers, rates, and condition labels that they understood visually but could not

evaluate. Without benchmarks, definitions, data sources, or modeling assumptions, they had no way to judge whether what they were seeing was significant, normal, or worth acting on.

Non-expert participants often requested a point of reference to evaluate what they were seeing. P3 saw the population growth rate but had no way to assess its significance:

> *"It isn't obvious initially, that it's 4.1. Is it 4.1% of what? What's the benchmark? You know what I mean?"* (P3, non-expert)

P1 went further, asking not just for context but for access to the data source and methodology:

> *"I want to see how it's calculated… it would be nice if, like underneath, you could have a link to the original data source or something."* (P1, non-expert)

The bridge conditions visualization illustrated this pattern clearly. Three of the non-expert participants asked what "fair" meant, and none found an answer through the interactions.

Expert participants asked different questions about the same problem. Where non-experts needed context and transparency, experts interrogated what produced it. P8 questioned the lack of clarity in the calculation of population growth rate, as different approaches would change their interpretation of the data:

> *"Is that growth rate from now? Or is that the annual growth rate in 2035. Really different outcomes depending on what they're doing."* (P8, expert)

As a transportation professional, P9 understood the engineering terminology but recognized that it would be opaque to most users, identifying what was missing:

> *"The very difficult piece of this is, what is fair condition, general structure, deterioration, or inadequate strength? These engineering terms are really tricky."* (P9, expert)

Both groups pointed toward the same finding; without knowing how a number was calculated or what it was measured against, participants could not move from reading the visualization to deciding whether to trust it.

3. When Data Lands as a Feeling

Engagement with the visualizations was often emotional before it was analytical. For many participants, the first reaction to a projection or a condition rating was not a question or an evaluation but a feeling: dread, fear, anger, excitement, disbelief. In these cases, the analytical work that followed was not separate from the reactions but shaped by them. Participants responded more strongly when the data reflected conditions they recognized from their own lives.

This was particularly visible among non-expert participants. P1's response to the commute travel time projections was immediate:

> *"The fact that I would still have to drive to get to the airport in 2050 is really really depressing, like just makes me want to gouge my eyes out because it doesn't make sense."* (P1, non-expert)

P2's reaction to the bridge conditions visualization followed a similar pattern, moving from alarm at the data to worry about what it meant for their own safety and daily commutes:

> *"It causes me fear, because I worry about the safety of our public infrastructure and the age of the infrastructure, and then it also disrupts my life."* (P2, non-expert)

Some expert participants also responded emotionally, but the reaction was shaped by professional context. P8's reaction to the population growth data was immediate but quickly turned toward the consequences:

> *"Oh, dear God, where is this gonna go, right? Where are we going to densify?" (P8, expert)*

P7 responded to the bridge conditions visualization not with fear but with excitement at what the visualization made visible:

> *"Oh, very good! This is so important!... Okay, that's blown my mind."* (P7, expert)

When a visualization landed emotionally, it shaped what participants looked at next, what they questioned, and what they trusted. This pattern was present across user groups and visualizations.

4. "What Does That Look Like?"

When the visualizations communicated the data clearly, participants wanted to go further, asking to compare scenarios, adjust assumptions, and explore trade-offs that the current design did not support. Participants were not confused by the data but engaged by it, proposing new ways it could be presented or questions that could be answered if connected to additional data sources.

For some non-expert participants, this meant filtering the data to match their own context. P5 wanted the commute travel time estimates to reflect the travel reality that actually mattered to them:

> *"I mean, I would almost only want to see peak time, because that's the time that really affects me."* (P5, non-expert)

For others, the proposals were more ambitious. P1 suggested a detailed what-if scenario around the high-speed rail visualization, unprompted:

> *"If I was to run a train from Seattle to Portland, and there's no scheduling constraints from BNSF… How long would that take?… Is there a 25 million*

> *dollar upgrade I can make to a section so I can increase the speed by 15 miles per hour? What does that look like?"* (P1, non-expert)

Expert participants wanted to go further as well, but through policy and funding trade-offs. P8 illustrated this directly on the bridge conditions visualization:

> *"What if you only funded the poor ones? You know. What would that cost? Be able to play a little bit of "what if." Oh yeah, by the way, that's another 2% on your gas tax."* (P8, expert)

Whether the question was personal or professional, the pattern was consistent. Participants who understood the data wanted to reason with it, and the strongest engagement came from those who could imagine using the visualizations to answer questions the current design had not yet addressed.

5. A Future for Whom?

When participants could not find themselves in the projections, whether because the time horizon was too distant or the geography too narrow, the data lost relevance regardless of its accuracy or design quality. This theme surfaced most clearly on the commute travel time visualization, which was the most customizable of the four, directly inviting participants to locate themselves in the data. That invitation made personalization possible, but it also made exclusion visible.

Every instance of temporal skepticism in the dataset came from a non-expert participant. P5 described the mismatch between the visualization's planning horizon and their own:

> *"2050 is so far away. When I'm thinking about what my future is in Seattle, it's like I'm making decisions 5 years out. Am I still going to be here in 5 years?"* (P5, non-expert)

P3 was more direct, noting that 2050 fell beyond their own life expectancy and therefore beyond personal relevance. For non-expert participants, the future being visualized had to fall within a personally meaningful horizon to feel relevant.

Participants located outside the visualization's coverage area also experienced exclusion. P10 connected their own location to a broader institutional critique:

> *"I guess it's great for the Northwest of the state, but one of the criticisms we get in our agency is that you represent the state, but a lot of your resources are either in Olympia, in headquarters, or in the Northwest region. And then I live in Yakima, I live in Spokane, and I don't get the level of transportation information."* (P10, expert)

P9, based outside the major urban corridors, raised the same concern from a personal standpoint:

> *"The difficulty for me is that I'm not in one of these cities, so I don't know that it would benefit me. I think it might benefit me, because there might be less cars on the road, so my driving trip might be better. But coming from a more rural area, it won't really benefit me much."* (P9, expert)

The temporal and geographic dimensions were different, but the underlying pattern was the same. When participants could not locate themselves in what the visualization showed, they disengaged from it.

Visualization Design Features

Across the four T2050 visualizations, participants generally engaged with the data successfully. Most features supported interpretation without prior instruction, and both expert and non-expert participants were able to navigate and reason with what they saw. Where specific design choices created barriers, those barriers did not always affect both groups in the same way.

The route selection interface on the high-speed rail visualization illustrates this directly. The visualization launched with Vancouver and Seattle pre-selected, as shown in Figure 4.5. When participants were asked to reverse the route to Seattle to Vancouver, clicking Seattle in the origin column did not work, as the button was disabled because Seattle was already selected as the destination. The required workaround was selecting a third city in the destination column first, a logic that was not communicated anywhere in the interface. Participants described the experience as a puzzle. Notably, non-expert participants were more likely to attribute the difficulty to themselves, which affected how confidently they continued engaging with the visualization, while experts identified it as a problem with the design. This points to how visual design choices may carry different implications for trust across expertise levels.

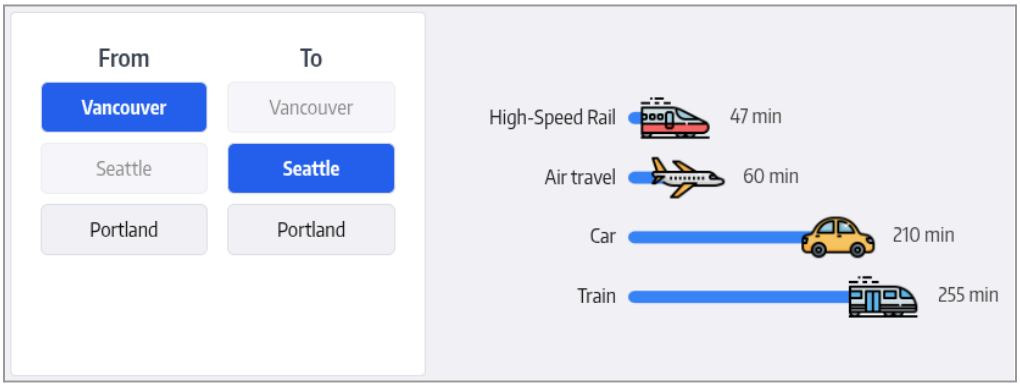

Figure 4.5. High-Speed Rail Visualization: Route selection interface.

The five themes from the previous section describe how participants made sense of the visualizations, but that sense-making depended on whether the interface let them get there. Across all four visualizations, usability breakdowns interrupted the interpretive work participants were doing and, in some cases, ended it entirely. P10's response was direct: *"Well, I don't see instructions. So I just stopped playing with the map"* (P10, expert). The specific features that supported or undermined this engagement are examined below.

1. Interactivity and depth of engagement

The interactive features that worked most effectively across both groups shared a common characteristic: they made data accessible at different levels of depth without requiring participants to engage at any particular one. Hover interactions and tooltips gave immediate access to values without requiring an understanding of the underlying data structure first. P11 (expert) captured the underlying logic, noting that hover interactions felt intuitive because they followed conventions from prior tool use: *"Pretty intuitive knowing that typically if you hover over a map, values will show up."* Non-expert participants used the same feature just as successfully. P5, after hovering over a bar on the bridge condition visualization, noted: *"I just hovered over that fair bar, and it then gave me a little window that gave me the exact number."*

Pop-ups opened on click extended this pattern. On the bridge conditions visualization, as shown in Figure 4.6, clicking on a bridge opened a detailed pop-up with condition ratings, dimensions, and cost estimates.

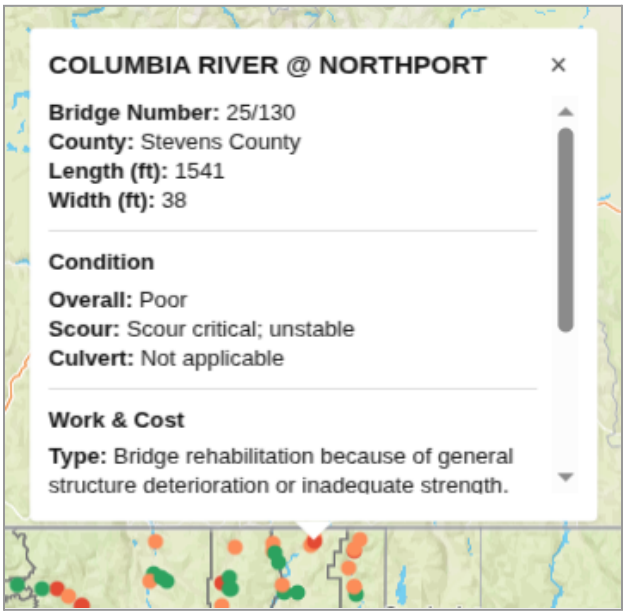

Figure 4.6. Bridge conditions visualization: pop-up opened on click.

While participants from both groups accessed the same pop-up, what they did with that information differed. Non-expert participants tended to read the top-level data while expert participants scrolled down, using the detail to refine initial reactions and arrive at more precise judgments. P9's response illustrates this directly:

> *"When I first looked at it, I felt 'Oh, Yikes, that's terrible. We should fix this.' And when I dig into it, these are like little bridges that nobody really goes on, but the one over Black Lake is an important one."* (P9, expert)

The pop-up transformed an initial reaction of alarm into a more precise judgment. Across both interactive features, the detail was accessible without being imposed on those who did not need it, supporting different levels of engagement.

2. Dashboard and exploration

Two of the four visualizations were linked dashboards where selections on one component updated others. On the population growth visualization, a year slider updated the map, pop-up info, and charts simultaneously, while clicking on a county or year in the chart filtered all the underlying data. The bridge conditions visualization connected a county selection dropdown, condition and detour bar interactions, and a display mode toggle to the map and summary statistics. This structure enabled participants to explore the data according to their own priorities, moving between geographic, temporal, and conditional views of the same data.

Both groups engaged with this structure successfully, using filtering interactions to narrow the data to what was most relevant to them, whether that was a specific county, a region, or a particular condition category. P1 discovered that selecting multiple counties produced a cumulative regional view:

> *"So now, if I select all, oh, that's cool. So I select these like, basically the Puget Sound region and it changes it. It's cumulative."* (P1, non-expert)

How participants found these interactions, however, varied across expertise levels. Expert participants approached dashboards with an expectation that hovering over components would reveal values and clicking would filter the data displayed, and in most cases those expectations were correct. Non-expert participants relied more heavily on explicit visual cues to identify what was interactive, and where those cues were absent, they either missed the interaction entirely or discovered it by accident. One non-expert participant noted that if they had not clicked on a county, they would have misread the graphs entirely. The linked structure of the dashboard supported exploration, but when participants could not identify what was interactive, the consequences were not just usability failures but potential sources of analytical error.

3. Visual encoding

Most visual encoding choices across the visualizations were received positively, including proportional bar charts, geographic mapping, and icon-based comparisons, all of which supported interpretation. Two specific encoding choices created credibility issues as shown in Figure 4.7.

On the commute travel times visualization, the visual difference between current and future conditions implied a more significant change than the underlying numbers supported. The map representation suggested a larger increase than the six minutes actually projected, leading one non-expert participant to question whether the data was accurate. The visualization presented the correct number, but the visual representation was exaggerated relative to the actual increase, producing doubt about whether what was shown could be trusted.

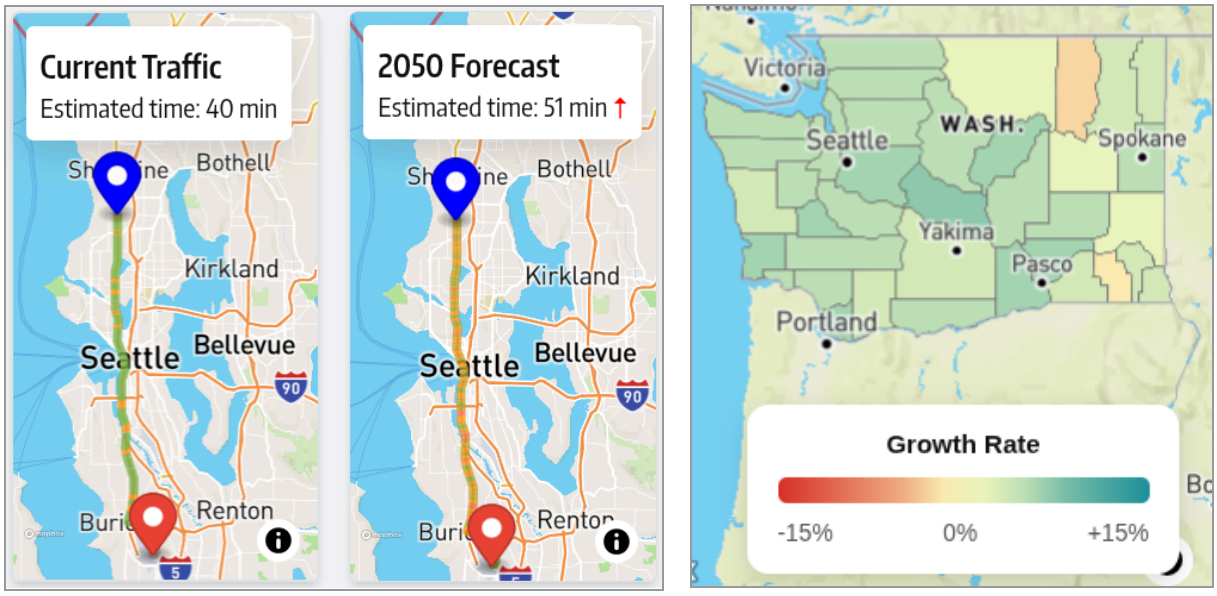

Figure 4.7. Left: Side-by-side map comparison encoding current and projected commute conditions through road color saturation, with estimated travel times displayed as text labels. Right: Choropleth map encoding county-level population growth rate through a diverging color scale centered at zero, ranging from -15% to +15%.

On the population growth visualization, the choropleth map used a diverging color scale running from crimson through pale yellow to teal, centered at zero. As shown in Figure 4.7, the values between these end points could be interpreted as a standard green-to-red scale. In most Western cultural conventions, green signals positive and red signals negative, and participants brought that association to the data. P2 articulated the tension this created:

> *"Green is good, but high growth rate may not be good."* (P2, non-expert)

High growth rates signal both development opportunities and increased demand on infrastructure. The diverging color scale implied that high growth was inherently good, which was not the intended message of the visualization, creating uncertainty about whether it was presenting information or making an argument.

4. Labels, legends, and annotations

Across all four visualizations, participants consistently encountered information that the interface did not explain. Geographic labels were incomplete, leaving participants to guess whether a city name referred to a location on one side of the border or the other. Selection state labels were too vague to confirm what was actually selected. Color legends were absent or ambiguous, with participants unsure whether a color represented a value for a specific year or an average over time. Participants repeatedly requested small inline cues that would have resolved these gaps: a note indicating that a bar could be clicked for more detail, or that double-clicking would zoom the map. What was missing in each case was not additional functionality but a brief explanation.

The absence of these elements did not affect all participants equally. Expert participants could draw on transportation domain knowledge to fill in what the interface did not explain. Non-expert participants did not have the same resources and were more likely to guess, misread, or stop engaging. The missing labels, legends, and annotations implicitly required domain knowledge to compensate for what the visualization did not communicate.

The visual features that supported engagement consistently, such as hover interactions, pop-ups, and linked components, did so by making depth available without requiring it, allowing participants to engage at the level that supported their own knowledge and priorities. Where the design created barriers, it did so by leaving the visualization's logic unexplained, whether through missing labels, ambiguous color conventions, or hidden features. This underscores the importance of clear communication of what is being shown and what interactions are available, while supporting multiple levels of engagement.

4.4.2 Phase 2: System-Side Analysis

The following analysis applies the system-side component of the preliminary framework developed in Chapter 2 to four T2050 visualizations. Each visualization is evaluated through the validation questions specified at each nested level to assess how the system's design choices account for users across expertise levels and communicate system reasoning in ways that support trust. The domain situation is evaluated once at the platform level; the remaining three levels are evaluated per visualization.

Domain Situation

- ***Immediate: Does the system's domain characterization explicitly account for users across a range of expertise levels?***

Prior to development, the team explicitly identifies two user groups: decision-makers, characterized as policymakers and transportation experts, and the general public. This distinction acknowledges the expertise range the platform is intended to serve and shaped the direction of the visualizations with both user groups.

- ***Immediate: Was user research conducted with the target users prior to or during development, capturing their goals, challenges, and decision-making workflows?***

Stakeholder meetings were conducted throughout development with policymakers and transportation experts. The primary goal captured was consolidating planning data that exists in silos across agencies into a single visual platform. Challenges and workflows were not directly gathered; given the wicked nature of the problem, they were inferred. Expert users were assumed to need analytical depth and access to multiple variables; non-expert users, personalization and simpler entry points. The general public was considered throughout the

process, but their needs were inferred by the design team rather than gathered directly. No formal user research was conducted prior to this study.

- ***Downstream: Do users across expertise levels recognize the system as relevant to their own goals and decision-making context?***

Expert participants whose work aligned with the platform's planning focus recognized the data as relevant to their decision-making contexts, while several also framed the platform's main value as public communication for non-expert audiences. Non-expert participants found the platform relevant when it matched their own geography and planning horizon, but the 2050 timeframe felt too distant for some, and they differed in whether they saw the platform as useful for their own independent use. Geographic scope also limited relevance across both groups, as participants outside the model coverage area or major urban corridors found the platform less applicable to their own contexts.

Task and Data Abstraction

Immediate: Do the data types require prior domain knowledge to interpret?

Viz 1	The underlying data is both categorical and numerical: counties as nominal and year as interval, alongside population total and growth rate as ratio measures, including both historical and projected values. The first three require no domain knowledge to interpret. Population growth rate is interpretable without transportation domain knowledge but assumes prior knowledge of percentage based metrics.
Viz 2	The data includes origin and destination addresses and regional models as nominal categories, with each model corresponding to a geographic coverage area. Current and projected travel times are ratio measures, traffic intensity is ordinal, and the route is a spatial path. Travel times and traffic intensity require no domain knowledge to interpret, as both draw on conventions familiar from everyday mapping applications.

	Regional models assume familiarity with both the geographic areas and the regional planning organizations they represent.
Viz 3	The primary data types are bridge location as spatial, county as nominal, bridge condition and detour distance as ordinal, and improvement cost as ratio. Secondary data includes bridge description, scour and culvert assessments, and work and cost details, spanning ratio, ordinal, and nominal measures. Bridge location and county require no domain knowledge. Improvement cost is broadly interpretable as a financial measure. Bridge condition is generally accessible, though its meaning depends on engineering criteria. Detour distance assumes familiarity with the concept of infrastructure detour routing. The secondary data largely requires engineering domain knowledge to interpret meaningfully.
Viz 4	Two data types are present: transportation mode as nominal and travel time as a ratio measure, requiring no domain knowledge to interpret.

Immediate: *Do the tasks the system supports assume a particular level of domain expertise, or do they accommodate users across the expertise range?*

Viz 1	The primary tasks are to explore and compare spatial and temporal patterns of county-level population growth, identify county-specific values, and summarize broader trends. These tasks support users across the expertise range: trend exploration and spatial comparison require no domain expertise, while derived measures and linked components offer additional depth.
Viz 2	The primary task is to compare current and projected 2050 traffic conditions for a personalized route, identifying how much longer a trip will take without future transportation investment. This task requires no domain expertise and is grounded in users' own travel experience. No broader tasks such as regional overview or pattern identification are supported, offering limited analytical depth.
Viz 3	Tasks that do not require domain knowledge include exploring the spatial distribution

	of bridge conditions and detour distances, identifying bridges by location and condition, and summarizing total bridges and improvement costs. Additional tasks such as interpreting technical details of specific bridges and identifying county-level and condition- or detour-based patterns assume domain knowledge. This range of tasks supports users across multiple levels of expertise.
Viz 4	The primary task is to compare travel times across four transportation modes for predefined city pairs. This task requires no domain expertise to perform, though it offers no analytical depth beyond the comparison itself.

Downstream*:* *Do users across expertise levels interpret the data structures accurately, or do certain formats create confusion or misinterpretation?*

Viz 1	Counties and year were interpreted accurately across expertise levels, and population totals were read correctly by most participants, though some initially confused the population total with the growth rate when navigating between the two measures. Growth rate created interpretive difficulty across both groups: non-expert participants wanted a benchmark or reference point to evaluate the displayed value, while expert participants questioned how the rate was calculated and over what reference period.
Viz 2	Travel times and traffic intensity were interpreted accurately across expertise levels. The regional model as a data category created confusion for non-expert participants, who did not recognize what it represented until they experimented with the dropdown; expert participants recognized the models immediately from professional experience. Participants across both groups noticed that the travel time values did not explain their assumptions about travel mode and time of day.
Viz 3	Bridge location, county, and improvement cost were interpreted accurately across expertise levels. Bridge condition categories were read as ordinal labels by both groups, but their underlying criteria were not understood: non-expert participants asked what "*fair*" meant in practical terms, while expert participants noted that the

	engineering terminology used to define condition ratings would be opaque to general users. Detour distance was the most consistently misunderstood data type, with participants across both groups uncertain whether it represented existing routes or hypothetical ones if a bridge were to close.
Viz 4	Transportation mode and travel time were interpreted correctly across expertise levels, but the distinction between current and projected values was not clearly communicated, leaving some participants unable to determine which modes reflected existing conditions and which were hypothetical. Participants in both groups understood travel time more broadly than the visualization represented, noting that the displayed values excluded access time to airports and train stations.

Downstream: *Are users across expertise levels able to complete the tasks the system was designed to support?*

Viz 1	Participants across expertise levels successfully located the expected growth rate for 2035 in their county, completing the core task through the year slider and county hover interactions. The connection between county selection and linked chart filtering was not communicated and had to be discovered; one non-expert participant noted that without discovering the county click, they would have risked misreading the linked data. Expert participants completed the task fluently and, during free exploration, extended their engagement to broader temporal and spatial patterns unprompted, suggesting that the visualization supported deeper engagement without requiring it.
Viz 2	Participants within the model's coverage area completed the task successfully across expertise levels. Two expert participants based outside the covered region could not use their home addresses and completed the task with alternative addresses with interviewer assistance. Browser autocomplete blocked successful address entry for two participants, and the interface provided no error feedback to explain what had gone wrong.

Viz 3	Most participants successfully identified the number of fair condition bridges in their county through hover interactions on the bar chart. Retrieving the improvement cost for fair condition bridges required clicking the bar to filter, an interaction that was not communicated in the interface, and most participants needed interviewer assistance to complete this part of the task regardless of expertise level. During free exploration, expert participants engaged more deeply with bridge-level detail and display mode switching.
Viz 4	Both tasks were completed across expertise levels, but the route selection logic created friction in Task 2 because participants could not select Seattle as a destination while it remained selected as the origin, with non-expert participants more likely to attribute the issue to their own error and expert participants more likely to identify it as a design problem.

Visual Encoding and Interaction Idiom

Immediate: Are the visual encoding choices appropriate for the data and clearly communicated so users can interpret them without domain-specific knowledge?

Viz 1	The choropleth map uses area marks with color hue and saturation encoding growth rate through a diverging color scale, appropriate for a ratio measure across geographic units. Selected counties are additionally encoded with a diagonal hatch texture. The linked charts use line marks with position encoding population count and growth rate over time, appropriate for temporal data. A dashed vertical line marks the boundary between historical and projected data. The year slider encodes year through linear position, appropriate for sequential temporal data. Color legend, axis labels, titles, and hover tooltips are present to communicate these visual encoding choices to users. However, caption or definition is missing for growth rate, the dashed line is not labeled, and the texture encoding for selected counties has no legend, leaving users to infer the meaning of key variables and visual elements if they don't have prior knowledge of them.

Viz 2	Two side-by-side maps encode time period through spatial position, appropriate for direct temporal comparison. The route is encoded as a line mark, with color hue along the line encoding traffic intensity along it, and origin and destination as point marks differentiated by color hue, all following conventions familiar with everyday mapping applications. The selected regional model is encoded as a boundary line mark delineating its geographic coverage area. Travel time, the primary ratio measure, is communicated through a text label in each map. Communication components include map titles, estimated travel time, and a directional arrow indicating travel time increase when appropriate. However, no color legend for traffic intensity is provided, the directional arrow is not labeled, and the model boundary is not labeled on the map.
Viz 3	Bridge locations are encoded as point marks of uniform size on the scatter map. County boundaries are encoded as spatial region line marks, with unselected counties overlaid with a semi-transparent layer when a selection is active. In condition mode, color hue encodes bridge condition using a traffic light convention (red=Poor, orange=Fair, green=Good), appropriate for an ordinal measure with clear directional meaning. In detour mode, color hue and luminance encode detour distance through a continuous blue range from dark to light, appropriate for an ordinal sequential measure, though lighter categories have reduced visibility against the map background. Bar marks with length encode bridge counts in both charts, with color hue matching the map encoding, appropriate for categorical comparison. Improvement cost and total bridge count are communicated as text labels with no visual encoding. Communication components include map legends for both display modes, bar chart titles, axis labels, percentage labels on bars, and an info button providing a definition for the No Detour category. However, condition and detour distance categories are not defined within the visualization, leaving users to interpret their meaning without criteria or context.
Viz 4	The visualization uses bar marks with length encoding travel time, appropriate for comparing a ratio measure across categorical modes. Transportation mode is encoded

	through categorical position on the vertical axis and reinforced through shape via mode icons. Mode labels, travel time values, and icons are present, making the encoding self-explanatory without requiring prior domain knowledge.

Immediate: Does the visualization communicate data provenance, limitations, and uncertainty explicitly?

Viz 1	Data provenance is not communicated within the visualization. Uncertainty in the projected values is not addressed, and no confidence intervals or explicit limitations statements are present.
Viz 2	Mapbox is attributed through info buttons on each map and a "powered by Mapbox" note in the search interface, covering map tiles and routing functionality. The underlying travel demand model sources and the forecast methodology are not referenced within the visualization. No uncertainty ranges are included in the projected travel times and no limitations statement is present.
Viz 3	Data provenance is not communicated within the visualization. The criteria used to classify bridge conditions as Good, Fair, or Poor are not explained. The methodology behind improvement cost estimates is not explained either. An info button provides a partial definition for the No Detour category, but detour distance as a measure is not explained. No uncertainty ranges or data currency information are present.
Viz 4	Data provenance is not communicated, the underlying report is not referenced within the visualization. Limitations are partially addressed through a note at the bottom stating that high-speed rail travel times are based on projections from feasibility studies and subject to change, but this applies to only one of the four modes.

Immediate: Do the interaction mechanisms allow users to engage at varying depths, or do they assume a single exploration path?

Viz 1	Several interaction mechanisms are present: selection and filtering through county

	clicks and chart mark selection; linked views that update the choropleth map and line charts simultaneously; temporal navigation through the year slider; and details-on-demand through hover tooltips. Together, these allow engagement at varying depths, from broad spatial-temporal overview to focused county-level comparison and interpretation.
Viz 2	Interaction mechanisms include selecting through the regional model dropdown and autocomplete address search boxes, and navigating, zooming, and panning across both maps. Beyond model and route selection, no additional mechanisms such as filtering, linking, or details-on-demand are present. The visualization supports limited variation through route and model changes but assumes a single exploration path.
Viz 3	Several interaction mechanisms are present: selecting and filtering through the county multi-select dropdown and bar chart click filtering; view switching between condition and detour through the display mode toggle; details-on-demand through bridge point clicks revealing scrollable pop-ups and bar chart hover highlighting corresponding map points; and navigating, zooming, and panning across the map. County and bar chart filters can be applied simultaneously and independently of the display mode. These mechanisms support engagement at varying depths, from broad statewide exploration to individual bridge detail. However, bar chart filtering, condition and detour filter reset, and bridge point click are undiscoverable without prior knowledge, as no visual cues or instructions are provided.
Viz 4	The only interaction mechanism is selection through city button clicks. The visualization assumes a single exploration path: select a route, read the comparison. There is no filtering, linking, or details-on-demand, offering no additional depth for expert users who may want to explore beyond the predefined options.

Downstream: Do users correctly interpret the visual encodings, or does confusion emerge at particular expertise levels?

Viz 1	Line marks and position encoding in the linked charts were correctly interpreted across expertise levels. The choropleth's diverging color scale produced two distinct interpretive problems: non-expert participants applied conventional associations between green and positive and red and negative that conflicted with the scale's neutral diverging intent, while one expert participant identified a discrepancy between the numerical value displayed and the color shown for their county. The legend did not accurately reflect the color mapping applied to the map and was therefore an unreliable reference for interpreting county values.
Viz 2	Traffic intensity colors and the side-by-side comparison were correctly interpreted across expertise levels. One non-expert participant noted that the map showed a bigger change than the reported travel time difference suggested, and questioned whether the data was accurate.
Viz 3	Condition color encoding, bar marks, and point marks were correctly interpreted across expertise levels. One non-expert participant questioned the use of circle icons for bridge locations instead of lines. The display mode toggle labels did not clearly communicate what switching between condition and detour views would show, leaving participants uncertain about the difference until they experimented. The legends for both display modes identified the categories but did not define them, leaving participants without criteria to interpret what condition ratings or detour distance ranges actually represented.
Viz 4	Bar marks and mode icons were interpreted correctly across expertise levels, but the visualization did not distinguish current from projected travel times, leaving the same ambiguity around existing and hypothetical conditions. One expert participant also identified ambiguity in the city labels, noting that Vancouver could refer to either Vancouver, Canada, or Vancouver, Washington.

Downstream: Do users express appropriate confidence in the system's outputs, or do patterns of over- or undertrust appear?

Viz 1	Expert participants questioned the calculation methodology and reference period behind the growth rate, using domain benchmarks and prior modeling experience to assess whether the projections seemed plausible. Non-expert participants were more varied, with some accepting the outputs without question and others flagging the missing data source as a barrier to trust.
Viz 2	Several participants across both groups found the projected travel time increases too small to be believable, drawing on their own experience and professional knowledge to question the outputs. Unstated assumptions about travel mode and time of day limited confidence further, as participants could not tell what conditions the projections actually reflected. Expert participants also questioned the methodology and data coverage, noting they had no way to evaluate what the system could reliably show.
Viz 3	Non-expert participants generally accepted the outputs without questioning their basis, though some asked how improvement costs were calculated and whether the data reflected current conditions. One non-expert participant questioned whether condition ratings reflected actual bridge safety or administrative classification. Expert participants raised concerns about data completeness and methodology, questioning whether all bridge types were represented and how key measures were defined.
Viz 4	First- and last-mile exclusion was the main trust concern, with participants across both groups questioning whether the displayed values accurately represented total travel time. Non-expert participants drew on personal travel experience to challenge the air travel value specifically, while expert participants identified the exclusion as a methodological limitation of the comparison. One non-expert participant also noted that the visualization felt framed toward promoting high-speed rail, raising concern about whether the data was being presented politically.

Downstream: Do users across expertise levels engage with the interaction mechanisms at varying depths, or do certain features remain inaccessible to particular user groups?	
Viz 1	Hover interactions and the year slider were accessible across expertise levels, with both groups successfully using these features to engage with the data. County click filtering and linked chart behavior were less consistently discovered by non-expert participants. One non-expert noted that without discovering the county click, they would have misread the linked charts. Expert participants engaged with a broader range of interactions, drawing on prior experience with similar tools, though one expert had difficulty identifying which county was selected due to the volume of simultaneous movement across visualization components.
Viz 2	Both groups successfully completed the available interactions, including address entry and map navigation, though browser autocomplete blocked address entry for some participants and the interface provided no error feedback to explain what had gone wrong. The single exploration path felt limited across expertise levels: non-expert participants wanted simpler adjustments such as a different year or peak-time filtering, while expert participants wanted more analytical depth through mode comparison and scenario planning that the visualization did not support.
Viz 3	The county dropdown caused initial confusion for some participants, as it differed from the map-click selection used in the previous visualization. Bar chart hover was accessible to both groups, but bar chart click filtering and bridge point pop-ups were undiscoverable without prior knowledge, with non-expert participants less likely to find them independently. One non-expert participant accidentally rotated the map and could not reset it, reflecting the absence of map control options. Expert participants engaged with the full range of mechanisms fluidly, combining filtering, display mode switching, and bridge-level detail.
Viz 4	City button selection was accessible to both groups, though the route selection logic required an unclear workaround that made the task more difficult for most

	participants. The single exploration path was identified as a limitation by expert participants, who wanted to explore cost, carbon impact, and scenario trade-offs that the visualization did not support.

Algorithm

Immediate: Does the system respond to interactions without disruptive delays under expected use conditions?

Viz 1	The year slider, county selection and chart mark interactions trigger updates across linked components. These operations are computationally moderate and no disruptive delays are expected under normal use conditions.
Viz 2	Route generation requires fetching real-time traffic data, producing a brief delay. All other interactions, including model selection, map navigation, zooming, and panning, respond without disruptive delays under normal use conditions.
Viz 3	All data is loaded client-side. However, county filtering, bar chart hover highlighting, and bar chart click filtering involve noticeable delays and laggy behavior under normal use conditions.
Viz 4	No disruptive delays are expected under normal use conditions.

Immediate: Does the algorithm produce accurate and consistent outputs across different user interactions?

Viz 1	Outputs across linked components are consistent and accurate.
Viz 2	Outputs are consistent when users select from the Mapbox autocomplete dropdown. However, browser autocomplete can interfere with the Mapbox search, causing users to select an unresolved location and producing no route output.
Viz 3	Outputs are generally accurate and consistent across county filtering, display mode

	toggling, and summary statistic updates. Bar chart click filtering occasionally produces incorrect values and bar sizes. The county filter dropdown does not consistently close when clicking on the map, producing inconsistent behavior across interactions.
Viz 4	Outputs are consistent and accurate across city selections.

Downstream: Do users maintain confidence in the system during computationally intensive operations, or do delays trigger uncertainty or disengagement?

Across all four visualizations, participants did not report delays or express uncertainty related to system performance, suggesting that interaction speed did not meaningfully affect confidence or engagement during the study.

Downstream: Do users across expertise levels trust the system's outputs, or do inaccuracies surface and undermine confidence?

Viz 1	No computational inaccuracies were identified by participants, and confidence concerns centered on the absence of provenance and methodological transparency rather than errors in what the system calculated. The color encoding inconsistency described above was an exception, leading one expert participant to question whether the values being displayed could be trusted.
Viz 2	Outputs were consistent for participants who successfully entered addresses within the model's coverage area. When address entry failed, whether from browser autocomplete or an out-of-region address, the interface provided no detailed explanation of what had gone wrong or how to proceed.
Viz 3	Summary statistics were trusted across expertise levels. The bar chart click filtering inconsistency was identified by one non-expert participant, who noticed incorrect values and bar sizes after clicking between condition categories and questioned whether the displayed numbers could be relied upon.

Viz 4	Several participants questioned the displayed travel times because they did not match their own experience, noting that train travel times did not account for known delays, air travel times excluded access and security time, and high-speed rail projections seemed difficult to trust given the infrastructure and funding challenges involved.

Across all four visualizations, the most consistent pattern was the absence of transparency around data provenance, methodology, and definitional criteria, a gap that shaped how both groups evaluated what they were seeing regardless of how well the visual encodings worked. Design barriers such as undiscoverable interactions and missing labels affected both groups, but differently: expert participants drew on domain conventions to compensate, while non-expert participants were more likely to disengage or misinterpret.

Non-expert participants found relevance contingent on geographic, temporal, and personal fit, and even when those conditions were met, they often did not see themselves using the platform independently. As one non-expert participant described it, the data was interesting "*at a glance,"* but meaningful only *"if it appeared on a postcard of some political campaign in the right context"* (P3, non-expert). Expert participants, by contrast, consistently recognized its communicative value, with one noting, "*As a professional, I don't see this as an analytical tool, but I would certainly see it as a public communications tool*" (P8, expert). This contrast raises questions about the alignment between the platform's design assumptions and the needs of its intended audience.

4.5 Analytical Synthesis

The empirical and system-side analyses in Section 4.4 surface different aspects of how the T2050 visualizations supported or undermined trust and expertise diversity. This section brings

them together to examine where they aligned, where they diverged, and what those gaps reveal about how the evaluative approach should be refined.

4.5.1 Alignment and Gaps

Together, the two analytical phases offered complementary perspectives on how the T2050 visualizations supported or undermined trust and expertise diversity, revealing patterns at different levels of detail. The reflexive thematic analysis surfaced broad patterns in how participants made sense of the visualizations, including how they drew on lived experience to evaluate outputs, where transparency gaps created interpretive barriers, how engagement varied across expertise levels, and which visual features supported or hindered that engagement. The system-side analysis examined the visualizations through a structured evaluation approach, identifying where issues emerged across each one and showing how specific design conditions shaped interpretation, trust, and engagement across both groups.

The system-side evaluative questions captured the major themes identified through the thematic analysis. *Lived Experience as the Test of the Data* and *The Numbers Don't Speak for Themselves* were reflected in questions about data interpretation, data sources and assumptions, and output credibility. *What Does That Look Like?* and *A Future for Whom?* appeared in questions about domain relevance, task support, and interaction depth. The visualization design features described in Section 4.4.1 mapped onto questions about data types, encoding clarity, interaction mechanisms, and interaction depth.

Together, these phases showed not only how participants made sense of the visualizations, but also where the design supported, constrained, or complicated that sense-making. At the same time, the comparison revealed a gap: the questions captured what users did and experienced, including whether they interpreted the data, trusted outputs, completed tasks, engaged with

interactions, and recognized relevance, but were less explicit about how users arrived at those responses. More specifically, they did not fully capture how users assessed trust in system outputs, how they determined whether the visualizations were relevant, and how prior knowledge, lived experience, and domain familiarity shaped differences in interpretation and engagement across expertise levels.

When Data Lands as a Feeling was the clearest example of this limitation. Emotional response appeared in the downstream validation findings through questions of relevance, confidence, and trust, but it was not treated as a primary focus. The thematic analysis showed why this mattered: affect was one mechanism through which participants decided what to trust, what to question, and whether to engage more deeply with the data. This points to a refinement in the evaluative questions. Rather than asking only whether users trusted system outputs, the evaluative approach should ask what mechanisms users drew on to assess the data's trustworthiness.

A second gap emerged when translating the task downstream question into the study protocol. The question asks whether users across expertise levels are able to complete the tasks the system was designed to support, but it does not distinguish between tasks that should be accessible to all users and tasks that vary by expertise, goals, and context. In T2050, some tasks functioned as common entry points, while others supported deeper exploration for users with more specific goals or domain knowledge. Testing every possible task may not be feasible or even useful when not every task is intended for every user. The downstream task question should therefore move beyond whether users complete the tasks the system supports, asking instead how prior knowledge shapes engagement with each task across the expertise range.

4.5.2 Toward an Evaluation Framework

The gaps identified in Section 4.5.1 point to several refinements to the system-side evaluative questions. These refinements apply to both the immediate and downstream validation questions, though in different ways. The immediate validation questions can be strengthened by asking how the system accounts for expertise diversity in its design choices, rather than only whether multiple expertise levels are acknowledged. The downstream questions can be refined to capture a more nuanced understanding of how expertise shapes the processes through which users assess trust, determine relevance, and decide how deeply to engage.

Domain Situation: The refinement shifts the focus from whether the system accounts for expertise diversity and whether user research was conducted to how it is reflected in the system's understanding of users: what assumptions it makes about their prior knowledge, goals, and stakes, and what reasoning processes or contextual knowledge were captured during development.

Immediate	How does the system's domain characterization account for users across a range of expertise levels, and what assumptions does it make about prior knowledge?
Immediate	What user research was conducted with target users prior to or during development, and what did it capture regarding users' goals, challenges, reasoning processes, contextual knowledge, and criteria for trust?
Downstream	What factors shape users' assessment of the system's relevance to their own goals and decision-making contexts across expertise levels?

Task and Data Abstraction: This refinement examines how tasks and data types are designed to support users across the expertise range, asking what each assumes about prior knowledge,

who it is intended to serve, and how those assumptions shape interpretation and engagement in practice.

Immediate	What prior domain context is assumed for interpreting each data type, and how is each data type intended to be useful across expertise levels?
Immediate	What tasks does the system support, what prior knowledge do they assume, and which expertise levels are they intended to serve?
Downstream	Do users across expertise levels interpret each data type accurately, and how does prior knowledge and domain context shape differences in interpretation?
Downstream	How do users across the expertise range engage with and complete the system's tasks, and how does prior knowledge shape that engagement?

Visual Encoding and Interaction Idiom: At this level, the questions are refined to ask how visual choices communicate across different levels of expertise. The main refinement concerns trust: instead of asking whether users express appropriate confidence in outputs, the questions should ask what users draw on to assess the trustworthiness of what the system shows.

Immediate	How well do the visual encoding choices fit the data being represented, how are those choices communicated, and what domain knowledge is assumed for interpreting them?
Immediate	How are data provenance, limitations, and uncertainty communicated?
Immediate	How do the interaction mechanisms support different depths of engagement across users' goals, contexts, and expertise levels, and how clearly are those interaction options communicated?
Downstream	How do users interpret the visual encodings, and how do prior knowledge or reasoning strategies shape differences in interpretation across expertise levels?

Downstream	How do users assess the trustworthiness of what the visualization shows, and how is that judgment shaped by user knowledge and system transparency?
Downstream	How do users across expertise levels engage with interaction mechanisms, and what shapes the depth of engagement they are able to access?

Algorithm: The refinement at this level consolidates the immediate and downstream questions around computational reliability, including response time, accuracy, and consistency. The downstream question shifts from whether delays or inaccuracies affect trust to how users interpret system behavior and how that interpretation shapes confidence or engagement.

Immediate	How reliable is the system's computational behavior across interactions, including response time, output accuracy, and consistency?
Downstream	How do users across expertise levels interpret the system's computational behavior, and how does that interpretation shape confidence and engagement?

These refinements make the system-side evaluative questions developed in Chapter 2 more precise by shifting from whether trust, relevance, task completion, and engagement occur across expertise levels to how those outcomes are produced. Together, these refinements provide the basis for the design implications that follow, translating the findings from Study 2 into practical guidance for designing visual DSSs that support trust and expertise diversity.

4.6 Design Implications

Participants across both expertise levels engaged with the T2050 platform and described finding value in what it showed. What shaped that engagement was not visual complexity but what each participant brought to the interaction: their domain expertise, their lived experience with the transportation system, and how closely the platform's scope matched their own contexts. The

design choices that supported engagement were those that communicated their own logic; where that communication broke down, gaps in interpretation, trust, and relevance followed. From these findings, the following design priorities emerge:

- **Support layered engagement:** Interactive features and dashboards supported engagement across both expertise levels by making detail accessible without imposing it. Design should support multiple depths of engagement without assuming a single path through the data, since the level of depth that supports engagement differs depending on the expertise users bring.
- **Make exploration paths visible:** Linked dashboards and exploration features extended what participants could do with the data, but many were not discovered without prompting. Available exploration paths should be communicated at the start of interaction rather than left for participants to find.
- **Account for cultural values in encoding:** Color conventions and other encoding choices carry cultural associations that shape interpretation. These choices should be examined for the values they embed and communicated when shared conventions cannot be assumed.
- **Contextualize the data:** Participants across both groups needed to know where data came from and what the numbers meant in context before they could assess what the visualizations showed. Data sources, methodological context, and reference points for interpreting values should be accessible within the visualization.
- **Support communication and system feedback:** When the interface did not explain what elements meant or respond clearly to user actions, participants were left without guidance; non-expert participants were more likely to misread elements or attribute

interaction issues to themselves. Labels, legends, and clear feedback on system behavior should be present throughout the visualization.

Beyond specific design gaps, the evaluation surfaced a question about whether the platform fit the users it was designed to serve. Expert and non-expert participants understood the platform's communicative value differently: expert participants recognized its potential as a public communications tool, while non-expert participants were less certain how they would engage with it independently. Exploring that mismatch further requires understanding what each group brings to the domain before encountering the system. Study 3 introduces a pre-exposure baseline as an independent analytical step to capture that directly.

The expert/non-expert classification used to organize these findings was productive at this site, where professional domain knowledge and its absence tracked a meaningful difference. Study 3 will return to that classification and reframe it.

Chapter 5: Evaluating a Visual DSS for Trust and Expertise Diversity in Qualitative Research Visualizations

5.1 Project Background and Context

Qualitative coding is a foundational component of qualitative research, through which researchers assign nominal categories to textual or other human-generated data for later interpretation and analysis (Charmaz, 2006). Although coding can be organized through structured codebooks and predefined categories, it remains an interpretive and labor-intensive process that requires researchers to read, assess, and make judgments about the meaning of the data.

In collaborative research teams, this challenge is amplified. When multiple researchers code the same dataset, differences in contextual background, methodological experience, and familiarity with the data and categories can produce inconsistent code application. These inconsistencies are often difficult to detect, and when left unaddressed, they can affect the quality of findings and the confidence of team members in the analysis (Krippendorff, 2011). This study is situated within a collaborative coding environment developed to address these coordination needs, and draws on an agreement metric designed for this coding context. The outputs of that metric form the basis for the visual decision support system evaluated in this chapter.

5.1.1 Collaborative Coding in TextPrizm

TextPrizm is a web-based application designed to support the collaborative annotation of short texts. Originally developed by Dr. Michael Brooks (2016) and redeveloped in 2019, it was used in the Human-Centered Data Science Lab from Winter 2020 to Fall 2022 to support the coding of online community data. Figure 5.1 shows the TextPrizm coding interface.

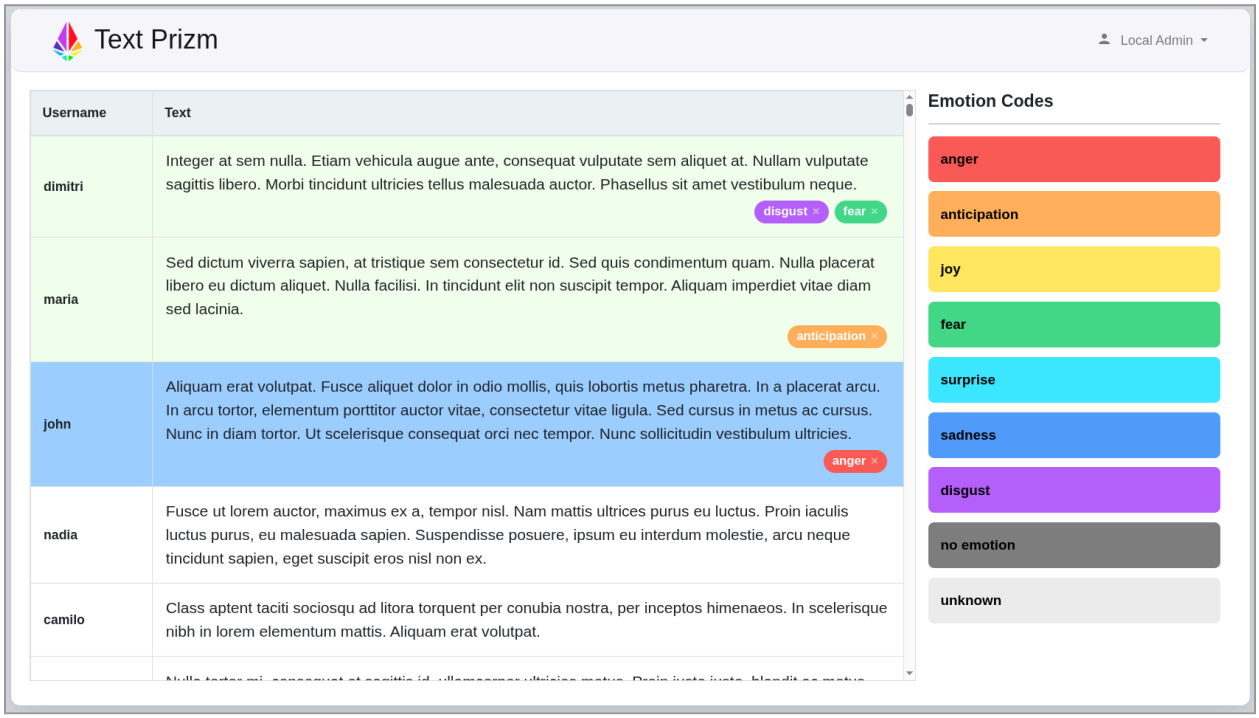

Figure 5.1. TextPrizm coding interface.

The data coded was fanfiction reviews from an online community, annotated for emotion using a non-mutually exclusive taxonomy, meaning a single review could receive more than one emotion code. The coding team included researchers with varying levels of qualitative experience, and coders worked with overlapping but not identical subsets of the data.

5.1.2 Agreement Assessment in Qualitative Coding

Assessing agreement across this team presented several challenges. Commonly used Inter-Rater Reliability (IRR) metrics, including Cohen's kappa (Cohen, 1960), Fleiss' kappa (Fleiss, 1971), and Krippendorff's alpha (Krippendorff, 2011), each address different team configurations but share a common limitation: they assume mutually exclusive categories. The non-mutually exclusive emotion taxonomy, the variable team size, and the unequal distribution of coding assignments across the team meant that none of these metrics could be directly applied, leaving the team without a reliable way to assess where disagreement was occurring or why.

To address this, we introduced Generalized Cohen’s Kappa (GCK) in 2023 (Figueroa, 2023). GCK extends the logic of Cohen's kappa by redefining agreement for non-mutually exclusive categories, introducing coder combination definitions, and calculating chance agreement through Monte Carlo simulations. As a result, GCK can be applied to variable team sizes, unequal coding distributions, and non-mutually exclusive taxonomies.

Unlike existing metrics, which produce a single overall score, GCK generates intermediate outputs, including agreement scores broken down by category and coder combination. For a team trying to understand and resolve disagreement, this added detail is significant. Rather than only knowing that agreement is low, researchers can identify which categories are contested, which coder combinations are most misaligned, and how agreement changes over time. As a result, teams can use agreement assessment not only to evaluate consistency, but also to identify specific areas of disagreement that may need further discussion.

5.1.3 Agreement Visualizations as Decision Support

Collaborative qualitative coding offers a productive site for studying visual DSSs. Agreement assessment is a semi-structured decision problem: the process is defined and the metrics are

computable, but determining where disagreement is significant and how to respond requires situated judgment. Coding teams also routinely include researchers at different career stages and with varying levels of familiarity with qualitative methods, making expertise diversity a structural feature of the setting.

Visualizations are well suited to support this kind of decision-making, as they can make agreement patterns across codes, coders, and time visible in ways that numerical scores alone cannot. The visualizations examined in this study were developed to address this need, building on early low- and mid-fidelity prototypes created in 2021 with initial feedback from qualitative coders, and refined into the two interactive visualizations described below.

The Code Agreement Timeline Visualization (Viz 1) shows how agreement varies across emotion codes over time, using kappa values on the y-axis and time on the x-axis, with each colored line representing a different emotion code and a dashed line representing overall agreement (Figure 5.2). Background bands show Cohen's kappa agreement levels, from poor to almost perfect, including a small negative range for agreement below chance. Users can switch between a cumulative view, which shows how agreement develops across accumulated coding data, and a weekly view, which isolates agreement patterns for each coding interval. A side panel summarizes per-code agreement values, while hover details, highlighting, filtering, and reset controls support overview-level interaction. Together, these features allow users to compare agreement patterns across codes and time, making visible both longer-term agreement trends and week-to-week fluctuations that may need closer interpretation.

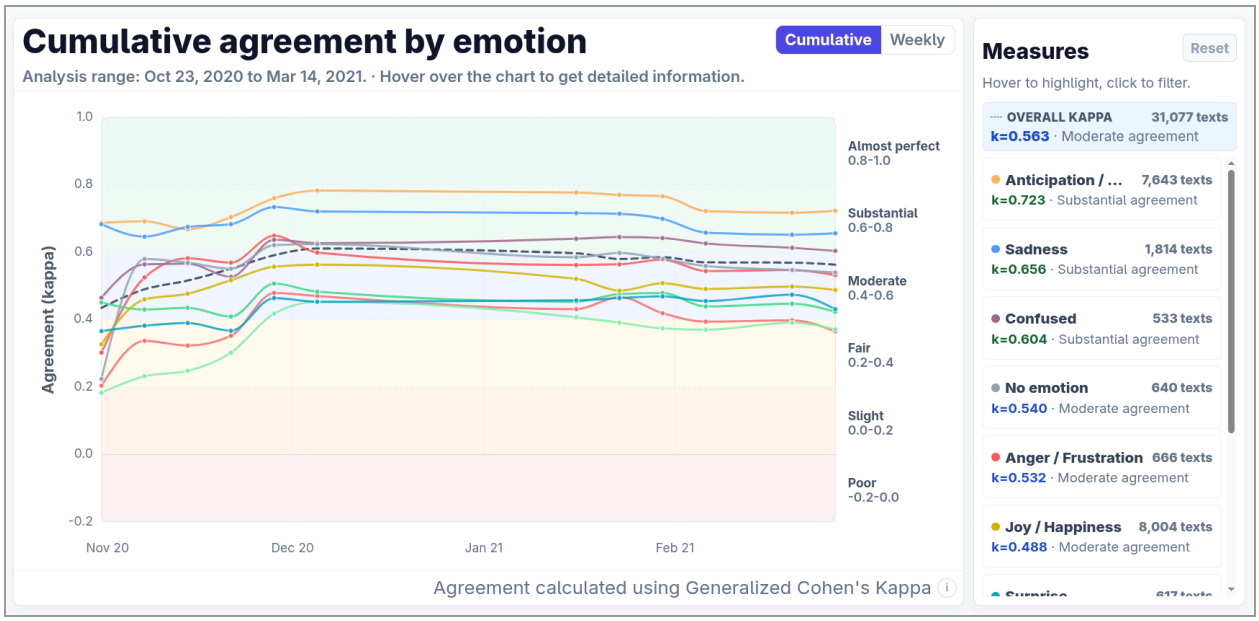

Figure 5.2. Viz 1: Code Agreement Timeline Visualization.

The Coder Agreement Network Visualization (Viz 2) represents how agreement is distributed among coders (Figure 5.3). In this network layout, each node represents a coder, with node size indicating the amount of text coded by that person. Edges connect coders who coded overlapping text, with edge thickness representing the amount of co-coded text and edge color representing the strength of agreement between the pair. The visualization includes filters for coders and emotion codes, a kappa range filter, and a side panel that summarizes per-code agreement for the current view. Users can hover over coders or edges for additional detail, click to focus the visualization on specific coders or coder pairs, and rearrange the network layout as needed. Together, these features support exploration of overall team-level agreement patterns and more focused comparisons of how agreement varies between specific coders and codes.

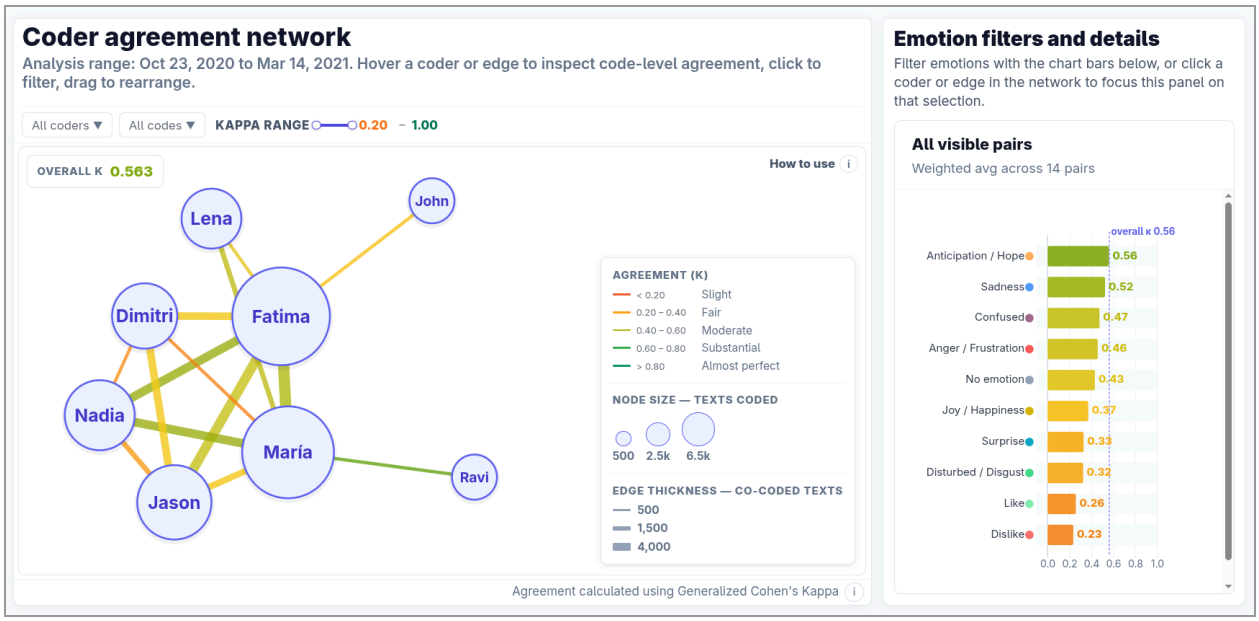

Figure 5.3. Viz 2: Coder Agreement Network Visualization.

These visualizations are designed to make agreement distribution across codes and coders visible and interpretable, directly supporting team discussions about where and why disagreement occurs. This setting, centered on interpretive research practice rather than transportation planning, offers a different context in which to examine how visual DSSs support trust and expertise diversity, and to test whether the evaluation approach developed in Chapter 2 and 4 holds beyond the domain in which it was developed.

5.2 Research Objectives

Studies 1 and 2 established the need for human-centered evaluation of visual DSSs and refined the system-side component of the theoretical framework introduced in Chapter 2. Study 3 extends this work by bringing the system-side analysis together with the independent pre-exposure user baseline from Section 2.5.2 to examine where the agreement visualizations support or undermine trust and expertise diversity.

While Study 2 focused on a public-facing transportation platform with a broad expertise range, this study shifts to a domain centered on interpretive research practice. This shift makes it possible to examine how the framework applies in a different decision context, where users vary not only in domain familiarity but also in their understanding of inter-rater reliability and qualitative coding workflows.

This study is guided by three research questions:

RQ1. Which features of the agreement visualizations support or create barriers for interpretation and trust across users with varying roles and backgrounds in qualitative coding?

RQ2. What does comparing the system-side analysis with the independent pre-exposure user baseline reveal about the design of the agreement visualizations?

RQ3. What does the combined application of the system-side evaluative questions and independent user baseline reveal about how the evaluation approach transfers across domains?

This study contributes to the development of the evaluation framework by operationalizing both lenses together and testing the comparison in a different domain, while also offering practical insights for designing visual DSSs that support trust and expertise diversity.

5.3 Methods

5.3.1 Study Design and Participants

Participants were recruited through University of Washington researcher and alumni networks during Spring 2026. Recruitment targeted individuals with direct experience in collaborative qualitative coding, as the study required familiarity with the team-based coding context the

visualizations were designed to support. The recruitment pool was informed by the VSD-informed stakeholder mapping exercise further described in Sections 5.3.2 and 5.4.2, expanding it beyond IRR-familiar lead coders to include qualitative coders involved in agreement discussions.

To be eligible, participants had to have participated in a qualitative coding project as part of their research or professional work and to have coded as part of a team of at least two coders. Formal inter-rater reliability experience was not required; participants only needed experience with a collaborative coding process in which coder agreement or alignment was discussed, measured, or otherwise assessed.

Participants were classified as experts or non-experts based on self-reported familiarity with IRR metrics, collected through a screening survey (Appendix C.1). Data visualization literacy varied within both groups and was not controlled for.

The sample included:

- **Expert users** (n=4, P1 through P4): participants who had used IRR metrics, participated in assessing them, or regularly used them in their work.
- **Non-expert users** (n=4, P5 through P8): participants who were not familiar with IRR metrics or had heard of them but had not applied them in their own work.

The sample size was selected to support qualitative analysis rather than statistical generalization. Consistent with usability research on small samples and heterogeneous user groups (Nielsen, 1994; Caulton, 2001), the final sample of eight participants was intended to capture recurring patterns within and across the two expertise groups.

Data collection consisted of semi-structured interviews combined with think-aloud user testing, conducted via Zoom during Spring 2026. Each session had two parts. The first was a

pre-exposure interview designed to capture participants' prior experience with collaborative qualitative coding, agreement assessment, and expectations for what would make agreement information useful or trustworthy. The second part involved interaction with the two agreement visualizations described in Section 5.1.3. Participants completed visualization-specific tasks while thinking aloud and then reflected on their interpretation, trust, confusion, and perceived fit with their own coding background. Think-aloud protocols were used because they are well suited to visualization evaluation, surfacing how users interpret and interact with visual representations in real time rather than only retrospectively (Ericsson & Simon, 1980; Carpendale, 2008). The full study protocol is included in Appendix C.2.

This study received an exempt determination from the University of Washington Human Subjects Division (IRB Study ID: STUDY00015698) under Category 2 of federal human subjects regulation (45 CFR 46.104). All participants were adults who provided informed consent prior to participation. Data was anonymized and stored on UW-approved systems; participation was voluntary.

5.3.2 Analytical Approach

My positionality in this study is shaped by my role as both designer and evaluator. I designed and developed TextPrizm, developed the Generalized Cohen's Kappa agreement metric, and designed the agreement visualizations evaluated in this Chapter. I also contributed to the qualitative coding project represented in the visualizations. These roles gave me direct knowledge of the coding workflow, agreement calculations, design rationale, and project context, while also creating the possibility that my interpretations could be shaped by my involvement in the system. I addressed this through reflexive memoing and by grounding findings in recurring patterns across participant interviews.

System Design Evaluation

The system-side analysis applies the refined immediate validation questions developed in Section 4.5.2 to each agreement visualization. The goal of this analysis is to examine the design choices embedded in each visualization prior to user exposure, generating threat hypotheses about where those choices may fail to support trust or expertise diversity in practice. The immediate validation was conducted independently of the empirical analysis and before any recruitment or participant sessions took place.

Pre-Exposure User Study

Before recruitment, VSD stakeholder mapping (Friedman & Hendry, 2019) was used at the domain level to identify the range of users present in the collaborative coding context and clarify their roles and stakes. The pre-exposure interviews then captured those stakeholder perspectives empirically, with the resulting data analyzed using the goals specified across Munzner's four nested levels in Section 2.5.2 as an organizing framework. Responses were mapped to each level to synthesize what participants brought independently to the study: their understanding of the domain problem, the tasks they expected agreement visualizations to support, the visual or interaction features they anticipated, and their criteria for trusting computational agreement outputs. Responses were organized into a participant-by-level matrix and then synthesized across participants to establish an independent user baseline for comparison with the system-side analysis.

Post-Exposure User Study

Post-exposure think-aloud and interview data was analyzed using the downstream validation questions developed in Section 4.5.2 across Munzner's four nested levels. Responses were examined per visualization to assess whether the threat hypotheses generated in the system-side

analysis materialized during live interaction, and to surface where design assumptions held or broke down across expert and non-expert participants. The post-exposure interview also included reflection questions in which participants compared their pre-exposure expectations with their post-exposure experience; this data is used to further inform the alignment and gaps analysis in Section 5.5.3.

5.4 Findings

5.4.1 System-Side Analysis

The system's design choices were analyzed before participant interaction, focusing on how the visualizations characterize the domain, structure the task and data, encode agreement information, support interaction, and implement computational behavior. This step identifies design assumptions and potential threats to trust and expertise diversity that will then be tested through downstream analysis.

Domain Situation

How does the system's domain characterization account for users across a range of expertise levels, and what assumptions does it make about prior knowledge?

The visualizations are primarily designed for lead coders, coding managers, or researchers responsible for monitoring agreement across a collaborative qualitative coding team. This domain characterization positions agreement as something interpreted at the team level, supporting oversight of how codes, coders, and coding discussions evolve during the process. Individual coders may also use the visualizations to understand how their interpretations align with others, but they are not the main target users. The design assumes that users are members of, or closely familiar with, the coding project being visualized. Project-specific context, such

---

as the meaning of code categories and the relevant coding periods, is therefore expected to be familiar to users. However, the visualizations also rely on methodological and system-specific concepts that may not be equally familiar across the team. While concepts such as coders and codes are common in qualitative coding, terms such as codings, co-coded texts, kappa, cumulative agreement, and weekly agreement require additional familiarity with agreement assessment and with how the system structures coding data. As a result, the visualizations account for different levels of methodological expertise to some extent, but still assume enough prior knowledge for users to connect agreement patterns to coding team management and decision-making.

What user research was conducted with target users prior to or during development, and what did it capture regarding users' goals, challenges, reasoning processes, contextual knowledge, and criteria for trust?

The visualizations were developed through prior lab experience with collaborative qualitative coding and informal formative feedback from qualitative coders. Early and mid-fidelity prototypes were presented in qualitative coding group meetings, where coders gave real-time feedback on the visualization concepts and their usefulness for interpreting agreement. This process shaped the design direction, but it was not a structured user research study and did not systematically capture users' goals, challenges, reasoning processes, contextual knowledge, or criteria for trust.

Task and Data Abstraction

What prior domain context is assumed for interpreting each data type, and how is each data type intended to be useful across expertise levels?

Both	The primary categorical variable is the emotion code, a nominal category from the project's codebook. Numerical agreement data includes per-code kappa values and overall kappa values. Ordinal categorical data includes agreement ranges, which

	classify kappa values into conventional interpretation levels. Text counts function as a numerical ratio value indicating how much coded data contributes to each agreement value. Emotion codes provide the categorical structure of the agreement analysis; because this data is project-specific, it is assumed to be familiar to users embedded in the coding project. Per-code and overall kappa values support comparison of agreement across codes, coders, or time depending on the visualization; interpreting these measures requires knowledge of agreement assessment metrics and aggregation. Text counts contextualize each agreement value by showing how much coded data contributes to it; interpreting these values requires users to understand what is being counted.
Viz 1	The Code Agreement Timeline adds temporal interval data, including the overall analysis period, weekly coding intervals within that period, and average weekly kappa summaries. This data supports comparison of agreement across time; interpreting them requires users to understand the distinction between weekly, cumulative, and overall agreement.
Viz 2	The Coder Agreement Network adds coder-level and relational data: coders are nominal categorical entities, and coder pairs function as relational data indicating that two coders coded at least one shared coding block. The total number of texts coded by each coder and the number of co-coded texts between each coder pair function as numerical ratio counts. Pairwise kappa values provide numerical agreement measures for each connected coder pair, and the per-code summary provides weighted average kappa values across visible coder pairs. This data supports comparison of how agreement is distributed across coders and coder pairs; interpreting it requires users to understand how coding work was assigned across the team and that agreement can only be calculated between coders with overlapping coded data. Coded text and co-coded text counts contextualize coder-level and pairwise agreement values by showing how much data supports each

	relationship. Weighted average per-code summaries support comparison across emotion codes in the current network view; interpreting them requires users to understand that these summaries aggregate across visible coder pairs rather than representing a single pairwise relationship.

What tasks does the system support, what prior knowledge do they assume, and which expertise levels are they intended to serve?

Viz 1	There are two main tasks: assessing how agreement changes across the project timeline and comparing agreement across emotion codes, both are contextualized through exact kappa values, interpretation ranges, and text counts. These tasks treat agreement as something that can be examined at both overview and detail levels, including broader weekly or cumulative trends and specific codes or time periods, assuming that users understand the coding timeline, the project codebook, and the meaning of kappa-based agreement values. Agreement ranges and summary information reduce the expertise required to read some values, but the tasks are most directly oriented toward lead coders or researchers who can identify and interpret agreement patterns and bring them into coding review and discussion.
Viz 2	The Coder Agreement Network is structured around tasks of locating agreement patterns within a coding team, comparing coder relationships, and identifying where agreement is stronger or weaker across the team. These tasks assume that users can interpret pairwise agreement and understand that coder relationships depend on overlapping coded data. They are most directly oriented toward lead coders or researchers monitoring a collaborative coding process, rather than individual coders working only with their own assigned data.

Visual Encoding and Interaction Idioms

	How well do the visual encoding choices fit the data being represented, how are those choices communicated, and what domain knowledge is assumed for interpreting them?
Viz 1	Time is encoded along the x-axis, fitting the temporal structure of weekly coding intervals, and kappa is encoded through vertical position, fitting its numerical interval structure. Emotion codes are encoded through color hue, which fits their nominal categorical structure but becomes harder to interpret when several categories use visually similar hues. Overall kappa is distinguished with a dashed line, separating the aggregate measure from the per-code lines. Agreement ranges use hue in both the background bands and right-side kappa values, which fit their ordinal structure and provide a visual reference for interpreting kappa values, but may conflict with hue also encoding emotion codes. These encodings are communicated through axis labels, legends in the right-side measures panel, titles and subtitles, and agreement range labels. Because the main encodings are explicitly labeled, the primary domain knowledge assumed is that users can interpret changes in kappa values as meaningful changes in agreement across codes and coding intervals.
Viz 2	Coders are represented with nodes and edges represent coder pairs with overlapping coded data. Node size encodes the number of texts coded by each coder, edge thickness encodes the number of co-coded texts, and edge hue encodes pairwise kappa. The right-side summary uses bar length to represent weighted average kappa values across visible coder pairs, with hue also indicating agreement range. These encodings are communicated through a panel with all the encoding definitions, titles, and subtitles. However, the visualization assumes that users can interpret pairwise kappa as agreement between coders and understand that coder relationships depend on overlapping coded data. Additionally, the network representation introduces ambiguity as spatial proximity is visually salient but does not encode any variable, which may lead users to interpret node position.

How are data provenance, limitations, and uncertainty communicated?	
Both visualizations state that agreement is calculated using Generalized Cohen's Kappa and cite the GCK paper through an information popup, but they do not explain how the displayed agreement measures are calculated. Limitations and uncertainty are not directly addressed.	
How do the interaction mechanisms support different depths of engagement across users' goals, contexts, and expertise levels, and how clearly are those interaction options communicated?	
Viz 1	Users can switch between cumulative and weekly views to examine agreement at different temporal ranges, filter or highlight emotion codes to focus on specific categories, and use hover details to inspect exact kappa values and text counts for specific time intervals. These interactions support different depths of engagement: users with less familiarity with kappa values can rely on overview patterns and agreement ranges, while users with more methodological knowledge can inspect specific values and compare aggregation levels. The interactions are communicated through visible controls, short instructional text, and hover/click cues in the side panel, although the conceptual meaning of cumulative versus weekly views and the distinction between highlighting, filtering, and resetting may still require some exploration.
Viz 2	Users can filter by coders, pairs, emotion codes, and kappa ranges through clicking in the network nodes/edges, clicking the bars in the right panel, or by filter dropdowns to examine agreement across different subsets of the coding team. Hovering over a coder shows their average agreement, total texts, and code-level text counts, while hovering over an edge shows total pairwise agreement and code-level agreement values for that coder pair. Additionally, the nodes can be rearranged to overcome occlusion. These interactions provide multiple levels of detail, from an overview of coder relationships to specific pairwise and code-level values. However, using this depth depends on users' familiarity with pairwise agreement, co-coded text

	counts, and the meaning of kappa values. The interactions are mostly visible through controls, legends, short instructions, and hover cues, but users may still need to explore the interface to understand how filters, selections, and hover details change what is shown.

Algorithm

How reliable is the system's computational behavior across interactions, including response time, output accuracy, and consistency?

Viz 1	The visualization displays precomputed agreement values calculated with the GCK algorithm, including per-code kappa, overall kappa, cumulative values, and weekly values derived from the same underlying coding data at different aggregation levels. Because interactions only change which precomputed values are displayed, response time is quick and interaction-based computational errors are unlikely. The main reliability concern is communicative consistency rather than computational behavior, as the interface uses both "texts" and "codings" to describe the amount of coded data and requires users to distinguish between overall kappa and average weekly kappa summaries.
Viz 2	The network displays precomputed GCK agreement values, so interactions respond quickly and only change what is shown rather than recalculating agreement. However, the kappa range filter starts at 0.20, causing relationships below that threshold to be hidden. As a result, some coders disappear when filtering or hovering by emotion, and weighted average values can change when those users are excluded. This creates inconsistent behavior and risks obscuring analytically important low-agreement relationships.

The system-side analysis reveals an assumption that runs across all four levels: the visualizations are designed for users who can interpret and connect kappa-based agreement values to coding team management. This shapes the domain characterization, the task structure,

and the interaction design, limiting the expertise range the system is designed to serve. Neither visualization explains how the displayed agreement values are calculated, leaving users with no way to verify what they see beyond the GCK citation; data provenance and uncertainty are similarly unaddressed. In the network visualization, an additional threat emerges at the algorithm level: the kappa range filter creates inconsistent behavior that can hide analytically important low-agreement relationships.

5.4.2 Pre-Exposure Baseline Findings

The pre-exposure baseline provides the independent user-side counterpart to these design assumptions, establishing what participants understood about the problem and their role within it before seeing the system. The subsections below trace how those understandings vary across Munzner's four nested levels; a cross-level synthesis at the end identifies the patterns that run across them and frames the analysis that follows.

Domain Situation

Goal: ***Identify the full range of users present in the domain, their roles, stakes, and how they understand the problem and their place within it.***

The independent user baseline began with a VSD-informed stakeholder mapping exercise to clarify who was implicated by agreement discussion support before participants were recruited. This mapping expanded the assumed user group identified in Section 5.4.1, from IRR-familiar lead coders to a broader set of stakeholders involved in or affected by agreement discussions, identifying two groups beyond the initial assumption: lead coders who assess agreement without formal metrics, and individual coders whose work is represented in the visualization. Two additional stakeholders were identified: the developer, who made the design choices embedded in the system and is accountable for how it performs, and communities represented in the data,

whose language and experience is categorized through the coding process. The full mapping is included in Appendix C.3.

All eight participants described some form of agreement process as part of their collaborative coding work, ranging from formal IRR metrics to codebook discussion and consensus-based review. P2 (expert) described what that work involved in practice:

> *"A lot of disagreement handling came around calibrating our coders to really understand what we were going after and what the framework was."*

Their roles included coding managers, coding leads, study designers, overseeing researchers, and solo or paired coders working through discussion, reflecting the hierarchical structure of collaborative coding teams. When disagreements could not be resolved through discussion, participants described turning to someone with greater authority or experience. P7 (non-expert) described bringing in "*a third person who was more knowledgeable*" to "*act as the tiebreaker.*" Participants described different understandings of what agreement is. This difference generally aligned with the expert and non-expert classification, but not consistently. One perspective, more common among expert participants, understood agreement as making coding reliable enough to support confidence in analysis, reporting, or downstream use. P2 (expert) described using coded data as "*a ground-truth dataset for training a machine learning algorithm*," while P1 (expert) described agreement as necessary for deciding whether categories with unclear boundaries could be used at all. P4 (expert) described this approach directly while also acknowledging its limits:

> *"We used IRR for that because it felt like, when we're talking about components of a video, this is more objective. We can trust an algorithm to handle figuring out how close things are. [...] For these interviews, we were very clear about not wanting to use an IRR because it felt very subjective. We wanted to have the*

> *discussion decide for us, and we didn't use an IRR because we didn't feel like the numerical value would tell us anything."*

A second perspective understood agreement as developing shared interpretation rather than producing a measurable result, more common among non-expert participants. P5 (non-expert) described coding as "*understanding the essence of the phenomenon*," where disagreement could reflect different lenses on what mattered in the data. P6 (non-expert) connected this to methodological stance, explaining that their analytical approach drew on a specific epistemological tradition, which in their words "*creates different ways of how we even analyze*." P7 (non-expert) similarly framed their work around "*the narratives in the data,*" adding that they were not "*too concerned with if they're completely in agreement.*" The stakes differed accordingly: for those who understood agreement as making coding reliable enough, poor agreement meant the process lacked the methodological rigor to move forward; for those who understood it as shared interpretation, the concern was reaching surface-level agreement without genuinely understanding what drove the disagreement.

Task and Data Abstraction

***Goal:** Understand the tasks users consider most meaningful in their domain and how they think about and structure the information relevant to their decisions.*

Across both groups, participants described an iterative process involving three main tasks: locating disagreements and tracing them back to the underlying data, diagnosing the type of disagreement, and resolving it through discussion. Locating where disagreements had occurred meant identifying which categories had diverged across coders or which specific items had been

coded differently. With larger teams and datasets, locating became increasingly difficult. P1 (expert) described what that step involved:

> *"I would run, in effect, a script to find where the points of disagreement were between coders, and then bring up those cases and ask people to explain their reasoning."*

P2 (expert) described a similar challenge, building downstream visualizations manually to surface patterns that were not immediately apparent. Non-expert participants, working with smaller teams, could more often locate disagreements through direct comparison, whether by placing codes side by side, reviewing independently coded transcripts, or working through the data together in discussion.

Once disagreements were located and traced back to the data, the next task was diagnosing what type of disagreement it was. Across both groups, participants described two types: disagreement that reflected a coder's misunderstanding of the codebook, and disagreement that reflected a gap in the codebook itself. P2 (expert) described this directly:

> *"What we have to figure out is, did they disagree because someone just doesn't understand the taxonomy... Or did they disagree because this [item] didn't fit into the taxonomy in one way or another?"*

P8 (non-expert) described the same distinction:

> *"Are we thinking that this code is misapplied, or are we thinking that we need to revise the codebook?"*

P1 (expert) approached diagnosis through pairwise patterns across coders, noting that when specific coder combinations consistently diverged on the same codes, this pointed to ambiguity in the codebook rather than an individual calibration problem. P5 (non-expert) described a

related distinction: disagreement over which aspect of a phenomenon mattered most required analytic discussion, while disagreement caused by definitional overlap between codes was easier to resolve because coders could return to anchor examples or reassess the scope of the category.

Resolving disagreements meant bringing specific items and categories to the team and working on them through discussion. What participants were trying to achieve through that discussion differed: expert participants described working toward a codebook stable enough to proceed, while non-expert participants described trying to understand what drove the disagreement, sometimes including interpretive variation in the analysis rather than eliminating it. When discussion could not resolve a disagreement, both groups described deferring to the lead or a more knowledgeable or senior researcher, though this was the exception.

Visual Encoding and Interaction

Goal: ***Understand how users across the expertise range interpret and evaluate visual information in their domain, and what they consider trustworthy or meaningful in data representations.***

Participants described a range of visual approaches to qualitative coding work. Spreadsheets were the most common across both groups, used consistently across years of academic and industry work, often with conditional formatting used to color-code categories or surface disagreement. P2 (expert) described the experience:

> *"When you're staring at a spreadsheet, your eyes are tired […] you can't really look at what's on screen and take it all in and be like, okay, what's the composition that we're seeing here?"*

Some expert participants also described building agreement visualizations separately, including pairwise agreement charts and code prevalence summaries. Others structured their visual work spatially. P6 (non-expert) described writing codes on color-coded sticky notes on easel boards, placing them side by side to make interpretive differences visible. Several participants described using tools such as *FigJam* and *Miro* for similar spatial organization. Purpose-built tools such as *ATLAS.ti* and *Dedoose* were also used, though primarily for their color-coded annotation features rather than their built-in visualizations. Across all approaches, color was the consistent visual strategy, used to encode different things depending on the context: categories, interviewees, themes, or disagreements.

When asked to imagine a visualization for assessing agreement, participants focused on different aspects of the coding process. Non-expert participants were less focused on agreement scores and more interested in code distribution, coder tendencies, and access to the underlying data. P5 (non-expert) described wanting to see how much each coder contributed to each code, normalized against others:

> *"If I see how much a person is contributing to that specific code compared to others, kind of normalized compared to others... maybe there's a lean toward one specific code from one coder, I think that helps me understand how they're understanding the phenomenon."*

Expert participants described more agreement-specific representations. P4 (expert) imagined a bubble chart where size encoded agreement strength, with the ability to access the underlying data:

> *"Maybe here's where people are agreeing on... a really nice big circle, and then you click on that, and then it's all of the things that have been coded as entertainment, with people's names on it."*

Across both groups, the imagined visualizations shared a consistent structure: an overview that made patterns visible at a glance, with the ability to navigate directly into the underlying data. P7 (non-expert) described wanting to see agreement along the narrative arc of the transcript and to click directly into the relevant section, while P8 (non-expert) framed this in terms of traceability, imagining a visualization like "*two groups of tree roots coming together*," where the shared outcome remained traceable back to each coder's original interpretation. When the imagined visualization included agreement scores, several participants described wanting to see how agreement values were calculated as a condition for trusting what the visualization showed.

Algorithm

Goal: Understand users' existing expectations of reliability, accuracy, and transparency of computational processes in data-driven outputs within their domain context.

Expert participants often described agreement metrics as legitimate tools, grounded in field standards and peer-reviewed practice. P1 (expert) described Cohen's kappa as "*what the standard is for the field.*" Non-expert participants more consistently questioned whether metrics were appropriate for qualitative work at all. P7 (non-expert) described a principled rejection grounded in Braun and Clarke's reflexive thematic analysis framework, which explicitly discourages inter-rater reliability. P6 (non-expert) described a deeper epistemological objection:

> *"For me, true qualitative data can stand on its own without the quantitative validation. [...] The moment you quantify qualitative data, it's like you're not trusting qualitative data, right?"*

Among participants who engaged with agreement metrics, how trust was formed varied considerably. P2 (expert) described understanding what a metric meant as a basic scientific requirement:

> *"Well, I think, as a scientist, you need to understand what these metrics mean."*

P1 (expert) took this further, using both raw percent and Cohen's kappa together as a verification check rather than relying on either alone. At the other end, P4 (expert) described delegating trust entirely to the recommendations of senior collaborators, noting that they trusted *"the people that trust it."* For P8 (non-expert), the concern was not about trust in a metric but about whether it could accurately represent their coding situation at all: overlapping transcript segments and partially overlapping code sets meant the agreement boundaries were not well defined enough for a standard metric to capture.

Across the four levels of the pre-exposure baseline, several patterns appeared consistently. Participants described needing to navigate from a summary view back to the specific items or transcripts where disagreements had occurred, a need that shaped both how they described their tasks and what they imagined a visualization would show. Trust in agreement processes and metrics was shaped by hierarchy, methodological tradition, and institutional authority. Participants also consistently described the same diagnostic challenge: when a disagreement occurred, was it a coder who had misunderstood the codebook, or a codebook that was not clear enough to apply consistently? Distinguishing between the two was a central part of agreement work across both groups.

However, a more fundamental tension ran through all four levels: participants differed not in familiarity with IRR metrics but in how they understood qualitative coding agreement itself. For expert participants, agreement was about making coding reliable enough to report with

confidence. For non-expert participants, it was about developing a shared understanding of what the data meant. This tension, surfaced before any participant had seen the visualizations, reflects a field-level divide rooted in epistemological and methodological traditions rather than technical familiarity.

This tension called into question the study's classification of experts and non-experts, pointing toward a reframing: all participants brought expertise, varying not in level but in methodological tradition, research context, and relationship to the problem the system was designed to solve. This reframing shapes the analysis that follows: the post-exposure findings examine not only whether participants could use the visualizations but whether the system's framing of agreement as a measurable, visual outcome resonated with the range of expertise the pre-exposure baseline established.

5.4.3 Post-Exposure Findings

The analysis of think-aloud and post-exposure interview data tested the threat hypotheses generated in Section 5.4.1 against what participants actually did and said during interaction, surfacing where design assumptions held or broke down across the expertise range. The subsections below trace those findings across Munzner's four nested levels.

Domain Situation

What factors shape users' assessment of the system's relevance to their own goals and decision-making contexts across expertise levels?

Lead coders and participants who managed collaborative coding teams connected both visualizations to their practice during interaction. P1 (expert) described the system as helpful for identifying which coders to check in with and which codes needed attention. P2 (expert)

described the system as serving a question they were already working on, wanting to understand "*how trained this coder is*" and how well each had internalized the taxonomy, while P3 (expert) described the system as useful for uncovering coder confusions on the taxonomy or identifying differences in perspectives. P4 (expert) described the network as matching what they had imagined, appreciating the macro-to-micro structure that moved from team-level patterns to specific pairwise disagreements, and P8 (non-expert) described it as matching the conversational structure of how their coding project had actually unfolded.

For some participants, fit was more conditional. P6 (non-expert) described how familiarity with the metric shaped what participants could take from the visualizations:

> *"It's not about how you came to the numbers, but what can we really say about it? [...], at least to me, as someone who doesn't work with Cohen's kappa, it detracts from moving the work forward to written work. But if I'm familiar with Cohen's kappa, I pick things up. I can see the visual. It's like, okay, I need to have a conversation with [coder a] and [coder b] to see what's going on between their coding of sadness."*

However, P7 (non-expert) described a mismatch in fit, noting they did not typically work on projects with that many coders and "*don't know that I would use the network graph as much*."

Participants who found both visualizations useful also described specific features and contextual information they would want added. P8 (non-expert) described the absence of a temporal dimension in the network directly:

> *"The time-based aspect isn't represented here at all. This is just showing us over the course of the whole project [...] having this, the cumulative version, and maybe just a final-week version, might be sufficient for me to feel like I understood the effect on the final results."*

Most participants described wanting a path back to the underlying texts and codebook. P4 (expert) described wanting codebook definitions and sample codes accessible within the same interface, noting they would not "*have to go open a different thing*." P1 (expert) described wanting context for what the disagreements meant in practice, and P7 (non-expert) described the same direction more directly, noting "*if I could click this and somehow get to the text itself*."

Task and Data Abstraction

Do users across expertise levels interpret each data type accurately, and how does prior knowledge and domain context shape differences in interpretation?

Both	Participants noted that project-specific details such as codebook definitions and coder identities were unfamiliar, but described them as information they would have from their own project, confirming the assumptions identified in the system-side analysis: P3 (expert) described it as something they "*would probably know if [they] were doing this [themselves]*." While kappa values were accurately interpreted as measures of agreement across all participants, uncertainty about the generalized version appeared regardless of kappa familiarity, with P1 (expert) noting they wanted to understand the distinction before citing the values and P7 (non-expert) indicating that a link to more information would help. The agreement ranges that contextualized those values were interpreted correctly across participants, but as P1 (expert) noted, "*different fields have different perspectives on what counts as high agreement,*" and the thresholds imposed a single perspective. Responses to that single perspective ranged from accepting the classification without interrogating it to P2 (expert) reading moderate-range values as insufficient and P7 (non-expert) describing low agreement as not a problem within their practice. Most participants read text counts as information about how much data supported

each value, though P6 (non-expert) expressed uncertainty about what the counts meant. P8 (non-expert) connected the difference in texts coded across emotions to how reliable the kappa values were across codes:

> *"If you compare something like anticipation/hope, which has 7,000 applications, to dislike, which only has 400, [...] a couple of people disagreeing on a couple of pieces of dislike data could have a bigger effect on this visualization than a similar thing happening with anticipation/hope."*

This observation was unique to P8, whose role as lead coder and codebook creator shaped how they read the counts.

Viz 1	Interpretation of cumulative and weekly values varied, with most participants connecting time intervals and gaps to their own coding workflows. P7 (non-expert) described the distinction clearly, noting that cumulative is "*showing [them] how it seems to be getting better over time*" and weekly is "*just showing [them] the variations in the week*." Others needed clarification to reach the same understanding, with P4 (expert) describing the distinction as "*a little bit confusing*."
Viz 2	Participants generally interpreted coders, total texts coded, and pairwise kappa values correctly, but differed in how they connected them to their coding practices. Participants with supervisory roles read the data as useful for team management, with P1 (expert) describing it as a way to check in with coders doing less work, and P4 (expert) using pairwise kappa to identify "*exactly the codes that they're having difficulties reaching agreement on*." P2 (expert) read the coder-level data as an indication of how well each coder had internalized the taxonomy, while P8 (non-expert) connected coder disagreement to differences in background and interpretation, describing it as useful for surfacing "*if there are differences in how people are coming into this project*." The co-coded text count and weighted average per-code summary were interpreted in

ways that reflected different levels of engagement with how collaborative coding data is structured. P6 (non-expert) did not know the concept, asking "*what does it mean to have co-coded texts?"* P5 (non-expert) understood the count and applied it as context for evaluating disagreements, describing how if "*two people had really high disagreement, and they co-coded a good amount of data, I would want to hear their perspectives*." P8 (non-expert) was the only participant to independently distinguish the weighted average per-code summary from pairwise kappa, noting that the numbers differed and describing the summary as useful for identifying which codes were generally difficult across the team.

How do users across the expertise range engage with and complete the system's tasks, and how does prior knowledge shape that engagement?

Viz 1 Participants began assessing agreement changes across the timeline during free exploration before they were prompted to do so, describing the break between coding periods, broader trend patterns, and changes in agreement over time. Some participants connected those patterns to possible reasons, with P2 (expert) interpreting the post-break agreement dip as suggesting that "*everyone forgot how to code*" and P4 (expert) connecting it to a possible member change or the need for "*re-ramping up the codebook."* P8 (non-expert) interpreted specific moments of disagreement and agreement across codes as evidence that "*something happened*," describing the task as "*a very helpful diagnostic tool"* for identifying what to look at next.

Participants approached comparing agreement across emotion codes with different goals. P1 (expert) connected low-agreement codes to codebook decisions, describing how a problem code would lead them to consider "*separating this category into two different categories*" or to "*shift the names of these categories*," and noted that changes in agreement could reflect "*discussions about what the code meant*." P3 (expert) described how consistently low-agreement codes would lead them to "*get*

rid of [them], or change the phrasing." P6 (non-expert) found value in the comparison for surfacing where disagreement occurred but questioned what it contributed beyond what their practice already produced:

> *"I would hope [...] this would raise those types of conversations in research meetings where it's like, 'hey, y'all, we're kind of not agreeing in this space.' But I also think those are the conversations that would emerge without Cohen's kappa, because I know that's the type of conversations we were having, even without Cohen's kappa."*

For P6, the comparison task identified where disagreement existed, but the conversations it was meant to prompt were already part of their research practice.

Viz 2 Participants quickly identified agreement patterns within the team before any tasks were prompted. Some participants with supervisory roles connected these patterns to specific next steps, with P1 (expert) describing focusing on coders with the lowest pairwise agreement and P4 (expert) describing the visualization as "*really helpful as a leader*" for identifying where to focus meetings. P2 (expert) read the patterns as indicators of how well coders had internalized the taxonomy. P7 (non-expert) engaged with the patterns but found the network less relevant to their own context, noting they "*don't usually work on projects with this many coders*."

Participants extended that engagement to specific pairs and codes through the comparison tasks. P4 (expert) described how the pairwise structure changed what the task output made possible:

> *"In my previous experience, when you're just looking at Cohen's kappa and seeing that someone's really low, it was a really difficult conversation, because it almost felt like we were talking to this person and telling them that they were doing it wrong. But with this, we can see, oh, actually, typically [coder A] and [coder B] have a lot of agreement, except for these situations, and then we can talk about those."*

The conversation shifted from judging a coder's performance to addressing a specific disagreement, and P8 (non-expert) described being able to say 'look' and point to it directly. P6 framed this visibility as both useful and potentially risky:

> *"How does someone with more power in the team come to recognize, I'm more detached from the coding, that when I come in, I create this [...] sort of awkwardness. At the same time, this could be harmful to people starting out with coding. [...] It creates an insider/outsider, potentially."*

This kind of visibility could surface disagreements that power dynamics might otherwise leave unspoken, while also exposing and isolating those new to the work.

Visual Encoding and Interaction Idioms

How do users interpret the visual encodings, and how do prior knowledge or reasoning strategies shape differences in interpretation across expertise levels?

Viz 1

Time and kappa position encodings were interpreted correctly across all participants, while color was challenging for all. P2 (expert) described having "*a lot of colors, a lot of codes*" making it "*a little hard to tell the difference,*" and P1 (expert) noted being "*not actually 100% sure*" which line represented which code.

Prior knowledge shaped how participants engaged with the background bands and the overall kappa line. For most, both encodings served as useful reference points, with P1 (expert) appreciating the background hue gradations and P8 (non-expert) using them to navigate directly to low-agreement codes. Where participants diverged was not in reading the encodings but in whether they needed what the encodings communicated. P4 (expert) did not initially notice the bands and misread the dashed line as missing data. P7 (non-expert) noticed and understood both but described wanting to remove the bands and not needing the overall kappa, asking "*why do I care about the overall kappa?*" when individual codes told them what they needed to

	know.
Viz 2	The legend communicated the encodings clearly, and participants who consulted it before engaging read node size, edge thickness, and color accurately. P4 (expert) described appreciating "*this little instruction*," and P7 (non-expert) read it explicitly before exploring. P3 (expert) and P5 (non-expert) both engaged with the network before consulting the legend and carried the timeline's color scheme into their reading. P3 noted being "*a little confused about the colors at first*," and P5 noted they had "*kind of connected sadness to this green*" before correcting themselves. Spatial positioning was the most significant encoding challenge, particularly for participants without prior experience reading network graphs. P7 (non-expert), P6 (non-expert), and P3 (expert) all treated node proximity as a data encoding, reading closer nodes as indicating stronger agreement. P7 tested this intuition and identified the repositioning feature as the source of that confusion, noting it "*doesn't actually change anything*" but that users might "*think that this is somehow different than this, but it's still showing the exact same data*."

How do users assess the trustworthiness of what the visualization shows, and how is that judgment shaped by user knowledge and system transparency?

Viz 1	Trust in the computational outputs was conditional rather than absolute. The GCK citation provided a reference point, with P5 (non-expert) noting it gave them "*some confidence that there [was] some theory behind it,"* but their trust was grounded in the overall pattern of stable agreement rather than the citation itself. P3 (expert) described trust grounded in a different condition, noting they would trust it *"more or less based on experience."* P4 (expert) described a more active assessment, checking outputs for signs of "*double counting*" or missing data before accepting the values. For participants who engaged more specifically with the values, what the system did not communicate shaped how far trust could extend. P1 (expert) described wanting to

click through to the raw data to verify the values, noting it would increase their trust. P6 (non-expert) expressed wanting to know how the calculation was done. P2 (expert) described needing context about what the outputs should look like in a multi-tag context in order to evaluate them. P8 (non-expert) identified that uneven text counts across codes could affect comparability, a limitation the visualization did not address. P7 (non-expert) presented a different case, drawing a clear line between trust in the outputs and how those outputs shaped their engagement:

> *"I think it's important for me to distinguish right now between trust and impact. I would trust that the information it's giving me is correct. But I personally don't think that, from the kinds of analyses I do, low agreement is a bad thing."*

For P7, what the system communicated or not was not what shaped their engagement; their own methodology was.

Viz 2

Some participants maintained similar levels of trust as in the previous visualization, while others found the network format introduced different conditions for establishing trust. P2 (expert) evaluated it against professional publishing standards and P4 (expert) treated it as a diagnostic to verify against raw codes. P5 (non-expert) described trusting it more than the previous visualization, noting it provided "*a little more detail about the coders*" and more agency to explore. P3 (expert) trusted it less, describing the network structure as harder to follow and conditional on understanding it better. P7 (non-expert) described a similar barrier from a different angle, trusting the values but not their own ability to read the network. P8 (non-expert) reframed the question entirely, describing the network as more trustworthy because it reduced their need to speculate: "*I think this is slightly more useful for me in terms of not making up fictions of my own*."

How do users across expertise levels engage with interaction mechanisms, and what shapes the depth of engagement they are able to access?

Viz 1	Because color encoding made identifying individual codes difficult in the chart, the hovering and filtering in the side panel were the most helpful interactions across all participants. P7 (non-expert) described the panel as their "*number one favorite feature*," noting they "*would not have been able to tell*" which code was which without it. P8 (non-expert) discovered its full utility mid-session, realizing the values they were zooming in to find were directly accessible in the panel. Depth of engagement with other interactions was shaped by whether participants encountered a specific need. The hover tooltip produced split responses, with P2 (expert) finding it useful for seeing "*all of the information*" about a specific week, while P5 (non-expert) found it disruptive to reading the overall chart. Both the cumulative and weekly toggle and the highlight and filter interaction were available, but the instructions did not make them easy to find. P3 (expert) did not discover the weekly view until prompted, noting they "*didn't realize that it was there at first*" despite finding it more useful for identifying disagreements once they had it. P1 (expert) found the highlight interaction mid-session, noting "*I didn't know I could do that*," despite the option being present in the instructions.
Viz 2	Hovering and filtering were the central interaction mechanisms across participants, used to examine agreement at different levels of detail. P4 (expert) described how hovering "*visualizes the change,*" showing where agreement held and where it fell apart. P7 (non-expert) used hover to retrieve exact values for task completion, describing finding the lowest agreement "*by hovering over the lines and seeing which one was lowest.*" Code filtering was used by most participants to isolate specific emotions for the comparison tasks. P8 (non-expert) described the interaction model as intuitive from the start, noting it felt "*similar to what I would have expected from using tools like Tableau.*" Depth of engagement varied across participants. P7 (non-expert) found the network graph "*hard to interpret,*" preferring the summary panel and engaging primarily with the overview level rather than specific pairwise values. On discoverability, several

	interactions were found mid-session despite the instructions. P4 (expert) discovered the center-by-coder feature during a task, noting "*Oh, I can also center by [coder]*." P8 (non-expert) discovered the reordering behavior during exploration, responding "*Oh, they reorder. Okay, wait, I have new stuff to do*." P6 (non-expert) did not discover the bar filtering feature independently, learning about it from the interviewer mid-session.

Algorithm

How do users across expertise levels interpret the system's computational behavior, and how does that interpretation shape confidence and engagement?

Viz 1	Response time was consistent across sessions and no participant described delays or unexpected computational behavior during their interaction.
Viz 2	The inconsistent behavior identified in the immediate validation appeared across participants. P1 (expert) noticed values shifting without explanation, describing numbers that "*keep moving*" and noting it made them "*more nervous*" and would prevent them from using the outputs or visualization in a formal report. P4 (expert) noticed a coder disappearing when filtering, attributing it to data absence rather than the filter. P6 (non-expert) arrived at the same concern independently, asking whether "*the denominator [was] the same across all of them*."

Across the four levels, participants drew from their professional roles and methodological contexts rather than their IRR familiarity. This shaped engagement at every level: which data types felt meaningful, which tasks connected to their own work, which encodings they used, and what criteria shaped their trust in the system's outputs. Participants whose roles aligned with the system's assumed user engaged with the depth the system offered. Where that alignment was partial or absent, participants connected the same information to different questions, or found the

system's framing of agreement as a measurable, visual outcome only conditionally relevant to how they worked.

The post-exposure data also tested the threat hypotheses generated in Section 5.4.1, and several materialized. Color encoding challenged all participants in the timeline. Spatial positioning in the network was misread as a data encoding across both groups. The kappa range filter's inconsistency caused coders to disappear and values to shift without explanation; multiple participants noticed, and one expert described the outputs as unreliable for a formal report as a result. These breakdowns point to specific design problems that Section 5.5.2 examines directly.

5.5 Analytical Synthesis

5.5.1 Expertise and Methodological Context

The study's initial classification of expertise around IRR familiarity did not hold across the three analytical steps: what varied was not familiarity level but methodological tradition and research context. The system-side analysis identified a mischaracterization at the domain level: the visualizations assumed a primary audience of IRR-familiar lead coders, framing agreement as something interpreted through kappa values and connected to coding team management. This cascaded downstream through the task structure, interaction design, and encoding choices, each of which required methodological familiarity to engage with meaningfully.

Before recruitment, VSD stakeholder mapping expanded the study's assumed user group beyond IRR-familiar lead coders, identifying a broader set of stakeholders involved in or affected by agreement discussions. This shaped the eligibility criteria: participants needed experience with collaborative qualitative coding in which agreement was discussed or assessed, but IRR familiarity was not required. The pre-exposure interviews then captured what that broader

sample brought independently. Participants described two different understandings of what agreement meant in their work: for some, it was about making coding reliable enough to report with confidence; for others, it was about developing shared understanding of what the data meant. That difference did not map cleanly onto the expert/non-expert classification. It reflected a field-level divide rooted in epistemological and methodological traditions, and it was present before any participant had seen the visualizations.

The mischaracterization at the domain level shaped how participants engaged with and formed trust in the system during interaction. Participants whose work aligned with the system's assumed user connected the outputs directly to coding management decisions from their own practices. Where that alignment was absent, participants engaged on different terms, evaluating the system against criteria it did not address or connecting it to questions it had not been designed to answer. For them, the system's framing of agreement as a measurable, visual outcome did not match how agreement functioned in their own work.

What the pre-exposure baseline surfaced is a reframing of how expertise diversity is approached, considering a range of valuable and diverse expert perspectives, each shaped by methodological tradition and research context, each bringing criteria the system had not accounted for. Without this baseline, the domain-level design choices would appear as reasonable assumptions about a defined user group. Section 5.5.4 takes it up as a refinement.

5.5.2 Visualization Design and Interpretability

The findings in Section 5.4 point to specific design features that shaped how participants across different qualitative coding backgrounds interpreted the agreement data and decided what to trust. The analysis below addresses RQ1 by examining those features.

Domain assumptions and downstream effect

The system-side analysis identified the kappa-based agreement model as a design assumption. The pre-exposure baseline revealed it as a mischaracterization: participants understood agreement differently and described visual approaches that reflected those differences, including code distribution over time, comparison across coders, and tracing disagreement back to the specific items where it occurred. That mismatch cascaded through the levels: in the task structure, which assumed users would interpret kappa values as meaningful; in the encoding, which made those values the primary visual output; and in the interaction design, which was built around exploring agreement patterns rather than accessing what produced them. The threshold ranges carried the same assumption, classifying kappa values against a single standard that participants did not share: some read moderate agreement as insufficient while others described low agreement as unproblematic within their practice. For participants whose practice did not center on kappa-based assessment, the visualization had less value for their workflows at every level. Offering multiple comparison dimensions alongside kappa-based views, with consistent access to the underlying data, would allow users to engage with agreement in ways that fit their practice.

> **Recommendation**: Design visualization systems to support multiple models of what the domain problem means rather than assuming users share a single one.

Depth of engagement and discoverability

The interactive features that worked most consistently were those that made detail available through interactions. The side panel in Viz 1 was the clearest example: color encoding made individual code lines difficult to distinguish in the main chart, and most participants used the panel to identify specific codes and read exact kappa values instead. Hover interactions and

code/coder filtering played a similar role in Viz 2, giving participants a way to move between overview patterns and specific pairwise comparisons. This pattern extended across both visualizations: the weekly toggle, the highlight interaction, the center-by-coder feature, and bar filtering were all found late or not at all without prompting, even when instructions were present. The features that most supported depth were also the least visible. A brief onboarding animation or guided walkthrough before independent exploration begins could draw attention to features participants would otherwise miss.

> **Recommendation**: Communicate available interactions at the start of the session, and ensure that features supporting depth are visible rather than discoverable.

Output transparency

The agreement visualizations cited Generalized Cohen's Kappa and linked to the GCK paper. For some participants, the citation was enough to establish that the outputs had a methodological basis. For others, it was not. Participants wanted to know how the values were calculated, what the metric assumed, and where it might not apply. Some wanted access to the raw data to verify the outputs; others identified limitations the visualization did not address, including how uneven text counts affected comparability across codes. The citation communicated the method that was used, but participants needed to understand how the values were calculated in order to interpret them correctly. A plain-language or visual explanation of how values are calculated, what GCK assumes, and where outputs should be interpreted with caution, accessible from the visualization, could address this directly.

> **Recommendation**: Communicate not only what computational outputs are but how they are produced, what they assume, and where they may not apply.

While the themes above were surfaced through the trust and expertise diversity lens, the interviews also identified concrete design issues. These findings reveal that the approach surfaces not only where system assumptions diverge from user needs but also specific problems that affect interpretability directly.

- In Viz 2, spatial proximity was consistently read as encoding agreement strength. Node position is determined by a draggable layout rather than any data variable, but participants treated closer nodes as more aligned coders. Spatial proximity could be used to double encode co-coded texts or agreement strength, or be explicitly labeled as not encoding any data.
- Color consistency across the two visualizations created a related problem. Participants carried Viz 1's color scheme into Viz 2, associating colors with emotion codes that did not hold in the network context. Establishing a consistent encoding system across both visualizations, or clearly signaling when color is being used differently, could prevent this.
- The kappa range filter in Viz 2 caused values to shift and coders to disappear without explanation. Participants noticed the inconsistency, and for at least one it directly affected their trust in the outputs. Removing the minimum threshold from the filter default, or clearly communicating when relationships are being excluded, could resolve this.

5.5.3 Alignments and Gaps

This analysis examines where the system-side and user-side analyses converge or diverge at each of Munzner's four nested levels. The gaps point to where trust broke down or expertise diversity was excluded. Reflection questions at the end of each interview session asked participants to compare their pre-exposure expectations with their post-exposure experience; those responses serve as an additional validation of the findings at each level.

At the **domain level**, both system and user-side analyses aligned on the domain problem: collaborative coding requires a way to understand where and why disagreement occurs across a team. Agreement processes were part of every participant's workflow, approached differently but present across all of them. The fundamental gap was centered in how agreement itself was understood. The system framed it as a measurable, kappa-based outcome interpreted through team management. The user-side baseline revealed a field-level divide about what agreement is, a mismatch that cascaded downstream to every level. Reflection responses confirmed where it surfaced: P5 described the system as "*definitely designed with someone else in mind*," noting "*it's not for where I come from*"; P6 described it as suited for "*mixed methods*" researchers; P7 noted the network "*felt like it was designed with someone else in mind*."

At the **task and data** level, the user-side baseline surfaced three tasks: identifying where agreement was low, diagnosing why it occurred, and resolving it. Both analyses aligned on the first: the system supported identification across codes, coders, and time, which participants recognized as valuable. The gap centered on the second and third, where the range of expertise shaped what users needed to do next. How participants understood the source of disagreement and what they would do about it depended on their methodological context and role, and the system's task abstraction did not account for that range. P3 described wanting to know "*where were the strongest discrepancies between codes*," and P1 described needing access to "*the raw data of codes*" and codebook definitions to understand what a disagreement actually meant in context.

At the **visual encoding and interaction** level, both analyses aligned on the value of interactive features that supported different depths of engagement: the system provided hover details, filtering, and the side panel, and participants expected and valued that kind of overview and details exploration. The system-side analysis centered the visualization as the unit of analysis,

identifying where encoding choices created a misfit, while the user-side baseline shifted the center to the user, explaining why that misfit happened and surfacing what visualizations would be useful to them. P7, for instance, wanted to see "*how the agreement shifts throughout the data itself*" rather than across calendar time, a need rooted in how their own work was structured rather than in the system's logic. The baseline also revealed something the system-side had framed differently: access to underlying data was not an interaction depth feature but a condition for trust. Participants across the full range described needing to "*click through to the raw codes*" to trust what the visualization was showing them.

At the **algorithm** level, the system-side analysis evaluated whether GCK was implemented reliably, assuming it was the right metric for the users the system was designed to serve. The user-side baseline revealed a different question: whether GCK was an appropriate way of measuring agreement at all. Participants brought a range of perspectives on numerical agreement metrics the system had not accounted for, from accepting them as legitimate field standards, to questioning whether a numeric metric could represent their specific coding situation, to a more fundamental objection rooted in whether measuring agreement as a number was appropriate in qualitative coding practice.

5.5.4 Framework Refinement

The alignment and gap analysis, combined with the expertise reframing from Section 5.5.1, produces three refinements to the evaluation approach that are applied in Chapter 6.

The first concerns the methodological sequence. Study 3 is the first to operationalize the full sequence: system-side analysis before recruitment, pre-exposure baseline before interaction, and post-exposure analysis after. Conducting the system-side analysis before recruitment allows for a clear understanding of the design assumptions embedded in the system and the users it was

designed for, which then informs the pre-exposure step. At the domain level, that step began with VSD stakeholder mapping, which expanded the assumed user group before recruitment took place and made it possible to bring in the range of perspectives that shaped this study's findings. The pre-exposure baseline then continued in sequence with the post-exposure phase with the same participants in the same session, enabling comparison both within each participant and across the full range. This study demonstrates that each step built on what the previous one established: the system-side analysis established what the system assumed, the pre-exposure baseline captured what users brought independently, and the post-exposure phase tested where those assumptions held.

The second concerns expertise diversity. The framework approached expertise through levels, and this study operationalized those levels through IRR familiarity as the defining criterion. The pre-exposure baseline showed that what varied across participants was not familiarity level but methodological tradition and research context, and that each position along that range represented a valid expert perspective rather than a point on a competence scale. What the evaluation should ask is not whether users with different expertise levels can engage with the system, but whether the system's framing of the problem resonates with the diversity of expertise users bring to it.

The third concerns the role of domain constructs. Study 3 showed that agreement meant different things to different participants before any of them had seen the system, a difference the study directly addressed but that the existing question did not make explicit. The refinement makes the role of domain constructs explicit at both levels: the immediate question should ask what the system assumes about the core domain constructs, and the downstream question should ask how those assumptions map onto how users across the expertise range experience them.

5.6 Design Implications

The most consequential finding of this study is a reframing of expertise diversity itself. The classification of participants as experts or non-experts, organized around IRR familiarity, did not hold: what varied was not familiarity level but methodological tradition and relationship to the domain problem, and each position along that range represented a valid expert perspective rather than a point on a competence scale. This reframing looks back at Study 2 as well, where the binary was useful because professional transportation knowledge and its absence tracked a real difference, but where the participants labeled non-experts were always bringing expertise of their own, grounded in lived experience and personal relationship to the system. What varies across users is not how much expertise they have but what kind, and Chapter 6 develops this as a core contribution of the framework.

Qualitative coding agreement proved to be a productive domain for testing the evaluative approach precisely because of what it surfaced. How participants understood agreement in their own work determined its relevance, which tasks they brought to it, and what conditions they needed to trust what it showed. For participants whose understanding of agreement aligned with the system's framing, the visualizations were directly useful. For those whose understanding did not, trust broke down not because the visualizations failed to solve the problem, but because the domain characterization had already misframed what the problem was. That mischaracterization excluded their expertise and that exclusion cascaded through every design decision below it. From these findings, the following design priorities emerge:

- **Provide access to underlying data**: Participants across the expertise range needed to verify what the visualizations showed and to act on what they found. Visual DSSs should treat access to underlying data as a core feature, not an optional depth of engagement.

- **Communicate computational assumptions**: Citing the methodology behind computational outputs was not sufficient for participants to assess whether it applied to their situation. Systems should communicate not only what a method is but what it assumes and what contexts it is designed for.
- **Make interaction options discoverable**: Several interaction features were not found without prompting, limiting how far participants could engage with the data. Interaction options should be communicated clearly and consistently throughout the visualization rather than left for participants to discover.
- **Maintain consistency across visualizations**: Inconsistent computational behavior caused values to shift and elements to disappear without explanation, undermining trust in the outputs. Computational behavior should be consistent and predictable across interactions and across visualizations within the same system.

These priorities are grounded in findings that only became visible through the combination of both analytical steps. However, the findings point beyond specific design fixes toward a more fundamental recommendation: the domain characterization needs to be revised to account for the diverse backgrounds, methodological traditions, and contextual knowledge users bring. The next step is assessing whether those needs can be addressed through the existing visualizations, through additions to them, or through new features and representations entirely.

Chapter 6: A Human-Centered Evaluation Framework for Visual DSSs

6.1 The Three-Study Arc

The three studies of this dissertation were situated in domains selected for what each could contribute to the approach's development: together they span the decision problem spectrum, three distinct expertise configurations, and three levels of methodological completeness.

Study 1 was situated in a semi-structured operational decision context with expert users. It surfaced what those users needed from decision support: contextual knowledge and system transparency the existing platform did not provide. That finding pointed toward visualization as a productive direction and established the analytical value of the user-side perspective: what expert operators brought to their decisions could not be inferred from the system alone.

Study 2 was selected because it offered conditions the first study could not: a wicked long-term planning problem, a built visual DSS, and a broad expertise range spanning transportation professionals and members of the general public. It was the appropriate site for testing whether the theoretical evaluative questions from Chapter 2 could surface meaningful findings. Comparing the system-side analysis against reflexive thematic analysis confirmed they could, and the gaps between the two produced the first round of refinements to the evaluative questions. Study 2 also revealed what was still missing: an independent user-side baseline established before interaction rather than drawn from post-interaction data alone. The breaks in expertise support that Study 2 surfaced pointed toward the user-side baseline as a necessary next step: understanding what users brought independently, before any interaction with the system.

Study 3 was selected to operationalize the pre-exposure baseline and to test the combined approach in a domain centered on interpretive research practice. It was the first study to run the full sequence: system-side analysis before recruitment, pre-exposure baseline before interaction, and post-exposure analysis after. The system-side analysis established what the system assumed about its users; the pre-exposure baseline revealed those assumptions as a mischaracterization, surfacing a field-level epistemological divide about what agreement meant in qualitative coding practice that cascaded through every design decision below it.

Taken together, the three studies provide the empirical basis for assessing the approach against qualitative trustworthiness criteria (Guba & Lincoln, 1989). In Study 2, comparing the system-side evaluative questions against reflexive thematic analysis showed that two independent approaches converged on the same patterns, supporting credibility. Study 3 demonstrated transferability: applied in a domain and expertise configuration the prior studies had not addressed, it produced findings consistent with what the approach was designed to surface. Across all three studies, the structured methodology, documented analytical steps, and reflexive memoing conducted throughout provide the audit trail that supports dependability and confirmability. Together, these establish the trustworthiness of the approach as a human-centered evaluation method for visual DSSs.

6.2 The Dual-Lens Evaluation Framework

The Dual-Lens Evaluation Framework is the central contribution of this dissertation, a human-centered evaluation approach for visual DSSs, built through the three studies and presented here in full. It analyzes a visual DSS through two independent perspectives: what the system assumes about its users, and what users independently bring to the domain problem. The system-side perspective examines the design assumptions embedded in the system and tests

whether those assumptions hold under actual use. The user-side perspective establishes what users bring to the domain problem on their own terms, before any exposure to the system. That gap between the two perspectives is the framework's analytical focus: where a system's assumptions about its users diverge from what users actually bring, trust breaks down because expertise diversity is not supported.

6.2.1 Core Architecture: The Two Lenses

The system-side lens examines the system at each level through two sets of questions. Immediate validation questions analyze what the system assumes about users' prior knowledge, goals, and domain context, and how those assumptions shape what the system supports and what it excludes. Downstream validation questions test whether those assumptions hold in practice, through analysis of how users with varying expertise interpret, engage with, and form trust in the system during interaction.

The user-side lens establishes what users bring to the domain problem at each level, independently of the system. Applied in a pre-exposure phase, it centers users' existing knowledge, reasoning processes, goals, and relationship to the domain problem before any encounter with the system. At each level, the lens is guided by a goal that specifies what the baseline should establish: who the users are, what they understand about the problem, what tasks they bring to it, and how they assess the trustworthiness of information in the domain. The specific methods used to achieve each goal are not prescribed. This flexibility is intentional: different domains present different user populations, problem structures, and contextual conditions, and prescribing a specific methodology would limit the framework's applicability across them. What the framework specifies is the goal at each level, not the instrument.

6.2.2 Expertise Diversity

Within this framework, expertise diversity does not refer to a competence scale from novice to expert. It refers to the expertise every user brings through their particular relationship to the domain problem, shaped by how they encounter, reason about, and act within it. What varies is not how much expertise they have, but what kind, and whether the system's design accounts for it. The evaluation is not about whether users can engage with the system, but whether the system's framing of the problem resonates with the expertise they bring to it. That question runs through both lenses: the system-side lens asks what expertise the system's design assumptions account for, and the user-side lens establishes what expertise is actually present in the domain.

This framing builds on what the studies surfaced about expertise. Study 2 used an expert/non-expert binary organized around professional transportation knowledge, and that division tracked a real difference in how the two groups engaged with the visualizations. The participants classified as non-experts were not lacking expertise, however; they brought expertise of their own, grounded in lived experience and their relationship to the transportation system. Study 3 made this explicit, surfacing participants who differed not in familiarity level but in methodological tradition and epistemological stance, differences a competence-based binary would have obscured. What the framework establishes is not a classification of users but what kind of expertise a given domain holds and whether the system accounts for it.

6.2.3 Operational Methodology

The evaluation process consists of three sequential steps: system-side analysis, pre-exposure baseline, and post-exposure analysis. Each step is conducted at a specific point in the process: system-side analysis before recruitment, pre-exposure baseline before interaction, post-exposure analysis after.

Step 1. **System-side analysis:** Examines the design choices embedded in the visual DSS across Munzner's four nested levels. At each level, immediate validation questions map what the system assumes about its users: who they are, what prior knowledge they are expected to bring, what tasks the system is designed to support, and what conditions it creates for trust. The output is a set of threat hypotheses: specific points where the system's design choices are likely to create barriers to trust or exclude the expertise users in the domain bring. Table 6.1 presents the immediate validation questions at each level.

Level	Questions
Domain Situation	Q1. How does the system's domain characterization account for the expertise diversity of its users, what assumptions does it make about prior knowledge, and what does it assume about how the core domain constructs are understood? Q2. What user research was conducted with target users prior to or during development, and what did it capture regarding users' goals, challenges, reasoning processes, contextual knowledge, and criteria for trust?
Task & Data Abstraction	Q1. What prior domain context is assumed for interpreting each data type, and how is each data type intended to be useful to users with varied expertise? Q2. What tasks does the system support, what prior knowledge do they assume, and which users are they intended to serve?
Visual Encoding & Interaction	Q1. How well do the visual encoding choices fit the data being represented, how are those choices communicated, and what domain knowledge is assumed for interpreting them?

	Q2. How are data provenance, limitations, and uncertainty communicated? Q3. How do the interaction mechanisms support different depths of engagement across users' goals, contexts, and expertise, and how clearly are those interaction options communicated?
Algorithm	Q1. How reliable is the system's computational behavior across interactions, including response time, output accuracy, and consistency?

Table 6.1. Immediate validation questions.

Step 2. **Pre-exposure baseline:** Captures the expertise users hold through their particular engagement with the domain problem, across all four levels and independently of the system. Before recruitment, the full range of users implicated by the system is identified, which expands the assumed user group beyond what the system-side analysis captured and informs eligibility criteria. The pre-exposure phase then captures what those participants understand the problem to be, what tasks they find meaningful, and how they assess the trustworthiness of information in their domain. Table 6.2 presents the goals that guide the pre-exposure baseline at each level.

Level	**Goal**
Domain Situation	Identify the full range of users present in the domain, their roles, stakes, and how they understand the problem, its core constructs, and their place within it.
Task & Data Abstraction	Understand the tasks users consider most meaningful in their domain and how they think about and structure the information relevant to their

	decisions.
Visual Encoding & Interaction	Understand how users with varied expertise interpret and evaluate visual information in their domain, and what they consider trustworthy or meaningful in data representations.
Algorithm	Understand users' existing expectations of reliability, accuracy, and transparency of computational processes in data-driven outputs within their domain context.

Table 6.2. Pre-exposure baseline goals.

Step 3. **Post-exposure analysis:** Tests whether the threat hypotheses from Step 1 materialized during live interaction. Participants interact with the system, and downstream validation questions are applied across all four levels to examine where design assumptions held and where they broke down against what users actually brought to the interaction. Table 6.3 presents the downstream validation questions at each level.

Level	**Questions**
Domain Situation	Q1. How does the system's characterization of the domain problem and its constructs map onto how users across the expertise diversity experience them?
Task & Data Abstraction	Q1. Do users with varied expertise interpret each data type accurately, and how does prior knowledge and domain context shape differences in interpretation? Q2. How do users across the expertise range engage with and complete the system's tasks, and how does prior knowledge shape that engagement?

Visual Encoding & Interaction	Q1. How do users interpret the visual encodings, and how do prior knowledge or reasoning strategies shape differences in interpretation across the expertise range? Q2. How do users assess the trustworthiness of what the visualization shows, and how is that judgment shaped by user knowledge and system transparency? Q3. How do users across the expertise range engage with interaction mechanisms, and what shapes the depth of engagement they are able to access?
Algorithm	Q1. How do users across the expertise range interpret the system's computational behavior, and how does that interpretation shape confidence and engagement?

Table 6.3. Downstream validation questions.

6.2.4 The Analytical Site: The Gap

The alignment and gap analysis compares what each lens surfaced at each of Munzner's four levels. Figure 6.1 presents the framework's structure: the system-side lens on one side capturing what the system assumes about its users, the user-side lens on the other capturing what users independently bring to the domain problem, and the four nested levels each lens examines. The space between them is the gap. Where the two lenses converge, the system's assumptions hold against what users bring; where they diverge, trust breaks down because expertise diversity is not supported.

Domain situation: The gap surfaces when the system's framing of the domain problem does not hold for all users. Users whose expertise is shaped by a different understanding of what the problem is find the system was not built for them. Because the domain level sets the foundation for every level below it, that gap cascades downstream.

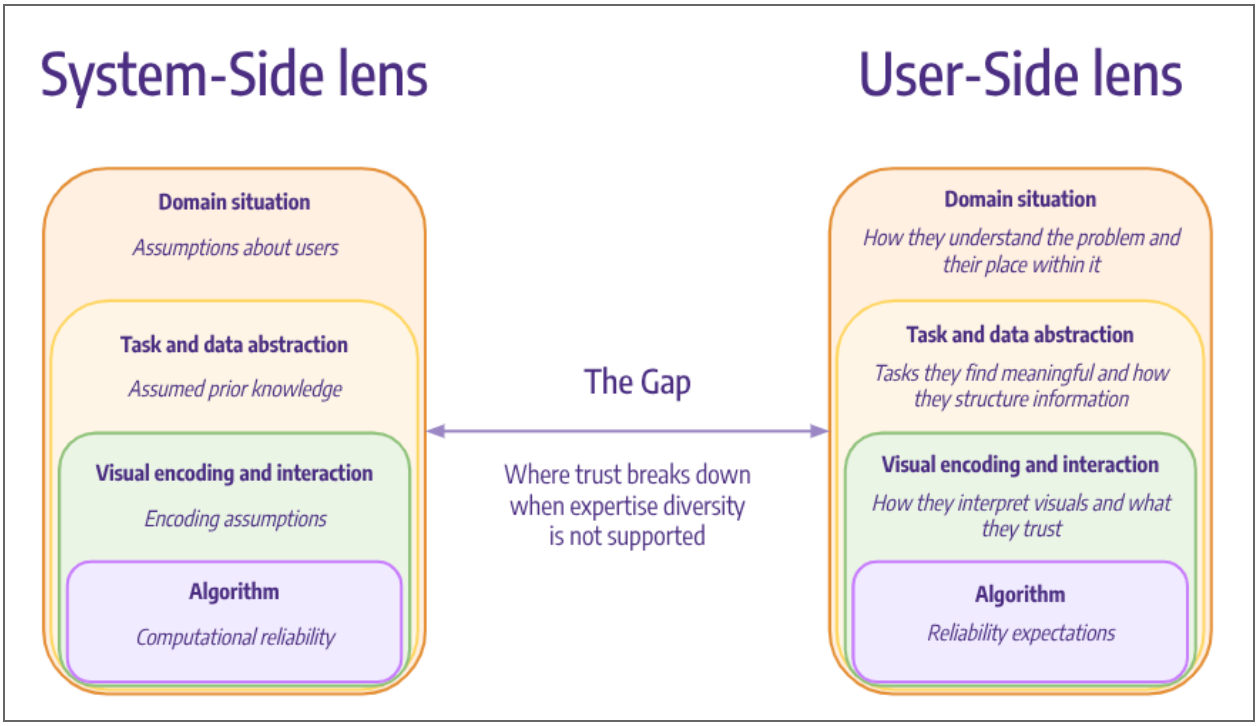

Figure 6.1. The Dual-Lens Evaluation Framework.

Task and data abstraction: The gap surfaces when the system's assumptions about what tasks matter and what data is relevant do not hold for all users. Users whose expertise leads them to different tasks or data needs are left without a way to act on what they know.

Visual encoding and interaction: The gap surfaces when what the system shows and how it allows exploration do not account for the representations users find meaningful and the interactions they need. Those expectations are shaped by expertise; where they diverge from the system's choices, users cannot engage with the data on their own terms.

Algorithm: The gap surfaces when the system's computational approach does not meet users' criteria for what reliable outputs look like. Those criteria are shaped by prior experience and expertise; where the system's outputs or behavior diverge from them, users cannot place trust through what they know.

6.3 Design Priorities for Visual DSSs

Chapter 2 introduced four provisional design priorities for visual DSSs: customization, training, transparency, and adaptability. These were grounded in the theoretical integration of visualization design, HCDS methods, and the DSS literature, and introduced as lenses to be tested through the empirical work that followed.

Transparency was introduced as communicating the limitations and uncertainty inherent in data and system outputs. All three studies confirmed this and extended it. In Study 1, operators could not see why the rules engine flagged an incident or why it did not. That invisibility limited both their confidence in system outputs and their ability to learn from system behavior. In Study 2, participants indicated that data sources and reference points were necessary before they could assess what the visualizations showed. In Study 3, citing the methodology behind computational outputs was not sufficient. Participants needed to understand what the method assumed and where it might not apply. Participants also indicated that access to underlying data was necessary to verify what the visualizations showed. System transparency, as a design priority, extends beyond data limitations and uncertainty to cover system reasoning, computational assumptions, and the reliability of system behavior.

> ***System Transparency:*** *Visual DSSs should make accessible the reasoning behind system outputs, the assumptions embedded in computational methods, the underlying data on which outputs are based, and the reliability of system behavior across interactions.*

Training was introduced as the system supporting users to bridge knowledge gaps, using narrative patterns and guided processes to make complex data accessible. The studies confirmed this and located where that communication most often broke down. Across all three studies,

when labels, legends, and feedback were absent, when narrative structure was not present, when available interactions were not surfaced, or when expert reasoning that existed in practice was not made accessible within the system, participants were left to interpret what the system showed without guidance. The absence of communication did not fall on users equally: those whose expertise aligned with what the system left unsaid could recover it from their own background, while those whose expertise diverged were left without a way forward. Communication, as a design priority, is not about what the system discloses but how it speaks to users: through structure, feedback, narrative, and the explicit surfacing of what is available to them.

> ***Communication:*** *Visual DSSs should actively communicate what they assume and what they offer, through narrative structure, labels, feedback, and the explicit surfacing of available options, supporting users regardless of where their expertise sits relative to the system's assumptions.*

Customization was introduced as supporting users across expertise levels through interactivity, allowing them to tailor their engagement in ways that matched their goals and workflows. Studies 2 and 3 confirmed that users across the expertise range engaged with the same systems differently: what served as an overview for some supported detailed analysis for others, and what was sufficient for one user's priorities left another needing more depth. The systems that supported this range were those that made detail accessible without imposing it, allowing users to engage at the level their knowledge and priorities required. Layered engagement, as a design priority, is supported through that flexibility: systems should offer multiple depths of engagement rather than assume all users share the same goals or follow the same path through the data.

Layered Engagement: *Visual DSSs should be designed to support multiple depths of engagement, allowing users across the expertise range to explore data at the level their knowledge, goals, and priorities require.*

Adaptability was introduced as a system's capacity to evolve with user needs and data over time. The gaps identified across the three studies traced back to assumptions embedded at the domain level: who the system was designed for, what reasoning it considered valid, and what contextual factors it incorporated. These assumptions were not visible within the systems themselves, and without mechanisms to surface and revise them, the gaps persisted. Domains and user populations also change; assumptions that hold at one point in time may not hold later. Reflexive adaptability, as a design priority, requires building in mechanisms to surface and revise domain-level assumptions, and treating that process as continuous rather than resolved at deployment.

Reflexive Adaptability: *Visual DSSs should build in mechanisms for surfacing and revising their assumptions about users at the domain level, and should treat evaluation and revision as an ongoing process as domains, user populations, and practices continue to evolve.*

Together, these four priorities offer practical guidance for both designing and evaluating visual DSSs. Used in design, they provide direction on what the system should disclose, how it should communicate to users, what depths of engagement it should support, and how it should be built to evolve. Used in evaluation, they function as criteria for identifying where a system falls short and what revision is needed.

6.4 Limitations

Several limitations shape what the framework can and cannot establish. Some concern the conditions under which it was developed and tested; others concern the scope of what it was designed to evaluate.

Across Studies 2 and 3, I was the designer and developer of the systems under evaluation. The system-side analysis, the interpretation of the pre-exposure baseline data, and the gap analysis all drew on system and domain knowledge built through design, development, and practice. Munzner's framework intends the system-side analysis to be conducted by someone with direct knowledge of the design choices. The depth of the analysis demonstrated here, however, may be partly contingent on a level of familiarity that the framework's methodology does not fully account for. The framework has not been tested by a researcher approaching a system and domain from the outside, and whether it produces comparable analytical depth in that context is not established.

The framework's evaluative scope reflects similar choices. The evaluative questions, the analytical approach, and the gap analysis are calibrated to surface where systems do not support trust and expertise diversity. Whether the same approach holds when the evaluative focus shifts to other constructs such as equity, accessibility, or racial inequalities has not been tested. The framework also evaluates the visualization layer of a visual DSS: how the system represents data, structures interaction, and communicates its computational outputs to users. It does not assess whether the underlying system produces accurate, equitable, or well-reasoned outputs. A visual DSS can support trust and expertise diversity at the visualization layer while the underlying model, data pipeline, or recommendation logic remains unexamined.Both are deliberate scope boundaries, but together they leave a portion of what makes a decision support system trustworthy unexamined.

The three studies used qualitative methods and small purposive samples, consistent with the research questions and the framework's human-centered orientation. The findings are not statistically generalizable. The framework also does not assess quantitative system performance metrics such as task completion time, error rates, or usability scores. These are acknowledged boundaries rather than gaps the framework was designed to fill.

6.5 Future Work

The framework was developed and refined as an evaluative tool, applied to systems after they were built. Applying it earlier, during the design process, opens a different set of possibilities. The evaluative questions at each level identify what a system needs to account for; the same questions, used earlier, could guide design decisions before those choices are made.

That earlier application also raises a question about sequence. In Study 3, the three components followed a specific, deliberate order. What the framework surfaces when the user-side lens comes first, when user-side findings feed into the system-side analysis, or when components run in parallel, has not been tested. Different configurations could change not just what the framework produces but what it is for.

The four design priorities were derived from empirical findings; they were not used as inputs into the design of the systems evaluated here. Using them as design criteria from the start, rather than deriving them after the fact, is a direction the framework naturally opens. As it moves into new domains, decision problem types, and constructs, those priorities will likely evolve.

Visual DSSs were the site for developing this framework because the gap it surfaces matters most there: these systems support consequential decisions that depend on human judgment, yet their evaluation has been dominated by performance and usability metrics that do not capture whether the system's framing of the problem fits the expertise users bring. Its applicability to

other system types, from visual analytics to exploratory data interfaces, and to fields beyond decision support, is an interesting direction to explore.

Chapter 7: Conclusion

This dissertation began from a concern that visual DSSs are often evaluated through technical performance, usability, or efficiency, while the conditions under which users trust them and engage with them in ways that reflect their expertise remain unexamined. Three goals followed from that concern: to investigate how design choices in visual DSSs shape trust and support expertise diversity, to identify design priorities that account for the expertise users bring, and to develop a human-centered evaluation framework that extends the evaluation of visual DSSs beyond usability across distinct decision-making domains.

Three studies addressed these goals at different points along the decision problem spectrum. Study 1, set in a traffic incident coordination platform, established the analytical value of examining decision support from two perspectives simultaneously, what the system does and what users independently bring, and identified visualization as a promising direction for making system reasoning visible to users across experience levels. Study 2, set in a public-facing transportation planning platform, applied the preliminary system-side evaluative questions developed in Chapter 2 alongside reflexive thematic analysis, examined how visualization features shape understanding, confidence, and engagement across users with varied expertise, and produced the first refinement of those evaluative questions. Study 3, set in a qualitative coding environment, operationalized the full sequence of the framework and surfaced tensions between system assumptions and user perspectives that would otherwise have remained invisible, including a field-level epistemological divide about what the domain problem meant that the system's design had not anticipated.

Taken together, the studies answered the three research questions guiding this work. Design choices in visual DSSs supported or undermined trust and expertise diversity through how they

characterized the domain problem, structured tasks and data, encoded information and supported interaction, and behaved computationally. Gaps at each level shaped what users could interpret, verify, and act on, and gaps at the domain level cascaded through every level below. The design priorities that emerged from evaluating visual DSSs through the lens of trust and expertise diversity are system transparency, communication, layered engagement, and reflexive adaptability, each synthesizing the conditions that supported these goals and gaps that limited them across the three studies. The human-centered evaluation of visual DSSs can be extended beyond usability through the Dual-Lens Evaluation Framework, which builds on Munzner's nested model of visualization design and the methods of Human-Centered Data Science.

The contributions of this dissertation are empirical insights from three domains into how users with varied expertise interpret, trust, and engage with decision support systems and visual decision tools; four design priorities for visual DSSs that account for the expertise their users bring; and the Dual-Lens Evaluation Framework for visual DSSs. The work also contributes to a refinement of how expertise diversity is approached in the evaluation of visual DSSs. The literature this dissertation builds on treats expertise diversity primarily as a competence range from novice to expert. The empirical work, in particular the pre-exposure baseline introduced in Study 3, surfaced a different framing: expertise diversity refers to the expertise users bring through their particular relationship to the domain problem, shaped by methodological tradition, lived experiences, or professional background, and how they encounter, reason about, and act within the problem. What varies across users is not how much expertise they have, but what kind, and whether the system's design accounts for it. This reframing shaped the framework, the design priorities, and what the evaluation surfaces.

As discussed in Chapter 6, the framework was developed and applied retrospectively, through my own direct knowledge of the systems evaluated in Studies 2 and 3. Whether the same

analytical depth holds when the framework is applied by an outside researcher, or used earlier in the design process rather than as an evaluation tool, remains open. These directions point toward where the framework can be tested, refined, and extended.

This dissertation set out to examine how visual DSSs can be designed and evaluated to foster trust and support expertise diversity. The Dual-Lens Evaluation Framework offers a structured way to do that work: by treating what the system assumes about its users and what users independently bring as two perspectives to be analyzed independently, and by making the gap between them the place where trust is built or broken. The expertise users bring to a visual DSS is not a problem to be managed but a condition to be designed for, and the systems that account for it are better positioned to earn their users' trust.

References

1. Adya, M., & Phillips-Wren, G. (2019). Stressed decision makers and use of decision aids: a literature review and conceptual model. Information Technology & People, 33(2), 710–754. doi:10.1108/itp-04-2019-0194

2. Aragon, C., Guha, S., Kogan, M., Muller, M., & Neff, G. (2022). Human-centered data science: An introduction. MIT Press.

3. Aragon, C., Hutto, C., Echenique, A., Fiore-Gartland, B., Huang, Y., Kim, J., ... & Bayer, J. (2016, February). Developing a research agenda for human-centered data science. In Proceedings of the 19th ACM conference on computer supported cooperative work and social computing companion (pp. 529–535).

4. Bender, E. M., & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. Transactions of the Association for Computational Linguistics, 6, 587-604.

5. Benjamin, R. (2019). Assessing risk, automating racism. Science, 366(6464), 421-422.

6. Boukhayma, K., & ElManouar, A. (2015, December). Evaluating decision support systems. In 2015 15th International Conference on Intelligent Systems Design and Applications (ISDA) (pp. 404-408). IEEE.

7. Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. Qualitative research in sport, exercise and health, 11(4), 589-597.

8. Brooks, M. (2016). Human centered tools for analyzing online social data (Doctoral dissertation).

9. Burstein, F., Holsapple, C. W., & Power, D. J. (2008). Decision support systems: A historical overview. In Handbook on decision support systems 1: Basic themes (pp. 121–140).

10. Card, S. K., Mackinlay, J. D., & Shneiderman, B. (1999). Using vision to think. In Readings in information visualization: Using vision to think (pp. 579, 1).

11. Carpendale, S. (2008). Evaluating information visualizations. In Information visualization: Human-centered issues and perspectives (pp. 19-45). Berlin, Heidelberg: Springer Berlin Heidelberg.

12. Caulton, D. A. (2001). Relaxing the homogeneity assumption in usability testing. Behaviour & Information Technology, 20(1), 1-7.

13. Charmaz, K. (2006). Constructing grounded theory: A practical guide through qualitative analysis. sage.

14. Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20(1), 37–46.

15. Craft, B., & Cairns, P. (2005, July). Beyond guidelines: what can we learn from the visual information seeking mantra?. In Ninth International Conference on Information Visualisation (IV'05) (pp. 110-118). IEEE.

16. D'Ignazio, C., & Klein, L. F. (2020). Data Feminism. MIT Press.

17. Eberhard, K. (2023). The effects of visualization on judgment and decision-making: a systematic literature review. Management Review Quarterly, 73(1), 167-214.

18. Eom, S., & Kim, E. (2006). A survey of decision support system applications (1995–2001). Journal of the Operational Research Society, 57(11), 1264–1278.

19. Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. Psychological review, 87(3), 215.

20. Few, S. (2009). Now you see it: Simple visualization techniques for quantitative analysis.

21. Figueroa, A., Ghosh, S., & Aragon, C. (2023, July). Generalized Cohen's kappa: a novel inter-rater reliability metric for non-mutually exclusive categories. In International Conference on Human-Computer Interaction (pp. 19-34). Cham: Springer Nature Switzerland.

22. Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological bulletin, 76(5), 378.

23. Friedman, B. (1996). Value-sensitive design. Interactions, 3(6), 16-23.

24. Friedman, B., & Hendry, D. G. (2019). Value sensitive design: Shaping technology with moral imagination. MIT Press.

25. Gorry, G. A., & Scott Morton, M. S. (1971). A framework for management information systems. Sloan Management Review, 13(1), 1–22.b

26. Greene, J., Rossi, F., Tasioulas, J., Venable, K., & Williams, B. (2016, March). Embedding ethical principles in collective decision support systems. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 30, No. 1).

27. Guba, E. G., & Lincoln, Y. S. (1989). Fourth generation evaluation. Sage.

28. Haselkorn, M., & Phelps, T. (2024). Model deployment of the Virtual Coordination Center for multimodal integrated corridor management. Washington State Department of Transportation.

29. Keen, P. G. (1980). Decision support systems: A research perspective. Decision Support Systems: Issues and Challenges, International Institute for Applied Systems Analysis (IIASA) Proceedings Series, 11, 23–27.

30. Keen, P. G., & Scott Morton, M. S. (1978). Decision support systems: an organizational perspective.

31. Komenda, M., & Karolyi, M. (2016). A Survey on Data-driven Decision Support Systems Using Effective Narrative Pattern.

32. Krippendorff, K. (2011, January). Computing Krippendorff's alpha-reliability. Retrieved from https://repository.upenn.edu/asc_papers/43/

33. Maltese, A. V., Harsh, J. A., & Svetina, D. (2015). Data visualization literacy: Investigating data interpretation along the novice-expert continuum. Journal of College Science Teaching, 45(1), 84-90.

34. Munzner, T. (2009). A nested model for visualization design and validation. IEEE transactions on visualization and computer graphics, 15(6), 921-928.

35. Munzner, T. (2014). Visualization analysis and design. CRC Press.

36. Nicodeme, C. (2020, June). Build confidence and acceptance of AI-based decision support systems-Explainable and liable AI. In 2020 13th international conference on human system interaction (HSI) (pp. 20-23). IEEE.

37. Nielsen, J. (1994). Usability engineering. Morgan Kaufmann.

38. Oliveira, M., Cabral, P., & Taibi, D. (2021). Interactivity to improve visual analysis in groups with different literacy levels.

39. Patton, M. Q. (2002). Qualitative research and evaluation methods (3rd ed.). SAGE.

40. Phillips-Wren, G. E., Hahn, E. D., & Forgionne, G. A. (2004). A multiple-criteria framework for evaluation of decision support systems. Omega, 32(4), 323-332.

41. Phillips-Wren, G., & Adya, M. (2020). Decision making under stress: The role of information overload, time pressure, complexity, and uncertainty. Journal of decision systems, 29(sup1), 213-225.

42. Poszler, F., & Lange, B. (2024). The impact of intelligent decision-support systems on humans' ethical decision-making: A systematic literature review and an integrated framework. Technological Forecasting and Social Change, 204, 123403.

43. Pretorius, C. (2017). Exploring procedural decision support systems for wicked problem resolution. South African Computer Journal, 29(1), 191–219.

44. Puget Sound Regional Council. (2024). SoundCast activity-based travel model [Data set]. https://www.psrc.org/activity-based-travel-model-soundcast

45. Rittel, H. W. J., & Webber, M. M. (1973). Dilemmas in a general theory of planning. Policy Sciences, 4(2), 155–169.

46. Sacha, D., Senaratne, H., Kwon, B. C., Ellis, G., & Keim, D. A. (2016). The role of uncertainty, awareness, and trust in visual analytics. IEEE Transactions on Visualization and Computer Graphics, 22(1), 240–249.

47. Segel, E., & Heer, J. (2010). Narrative visualization: Telling stories with data. IEEE transactions on visualization and computer graphics, 16(6), 1139-1148.

48. Scott Morton, M. S. (1971). Management Decision Systems: Computer-based Support for Decision Making. United States: Division of Research, Graduate School of Business Administration, Harvard University.

49. Shneiderman, B. (2003). The eyes have it: A task by data type taxonomy for information visualizations. In The craft of information visualization (pp. 364–371). Morgan Kaufmann.

50. Simon, H. (1960). The new science of management decision. Prentice-Hall.

51. Skaburskis, A. (2008). The origin of "wicked problems." Planning Theory & Practice, 9(2), 277–280.

52. Smith, P. J., Beatty, R., Hayes, C. C., Larson, A., Geddes, N. D., & Dorneich, M. C. (2012). Human-centered design of decision-support systems. In J. A. Jacko (Ed.), Human computer interaction handbook (3rd ed.). CRC Press.

53. Smith, P. J., Geddes, N. D., & Beatty, R. (2009). Human-centered design of decision-support systems. In Human-Computer Interaction (pp. 263-292). CRC Press.

54. Tashakkori, A., & Creswell, J. W. (2007). The new era of mixed methods. Journal of Mixed Methods Research, 1(1), 3-7.

55. Thurston Regional Planning Council. (2022). Regional travel demand model [Data set]. https://www.trpc.org/319/Travel-Demand-Modeling

56. Tomasi, S. D., Liu, J., Cheng, F., & Han, C. (2023). The role of individual characteristics: How thinking style and domain expertise affect performances on visualization. Information visualization, 22(3), 265-276.

57. Tufte, E. R., & Graves-Morris, P. R. (1983). The visual display of quantitative information (Vol. 2, No. 9). Graphics Press.

58. Van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. International Journal of Human-Computer Studies, 144, 102493.

59. Washington State Department of Transportation. (2019). Ultra-high-speed ground transportation study: Business case analysis [Technical report]. https://wsdot.wa.gov/sites/default/files/2021-11/Ultra-High-Speed-Ground-Transportation-Study-Business-Case-Analysis-Full-Report-with-Appendices-2019.pdf

60. Washington State Department of Transportation. (2024). WSDOT all bridge and tunnel inventory (state & local) [Data set]. Retrieved October 31, 2024, from https://geo.wa.gov/datasets/WSDOT::wsdot-all-bridge-and-tunnel-inventory-state-local

61. Washington State Office of Financial Management. (2022a). Intercensal estimates of April 1 population for the state and counties: 1960–2020 [Data set]. https://ofm.wa.gov/data-research/population-demographics/estimates/archive/

62. Washington State Office of Financial Management. (2022b). Growth Management Act population projections for counties: 2020 to 2050 [Data set]. https://ofm.wa.gov/data-research/population-demographics/forecasts-projections/growth-managment-act/2022-projections

Appendices

Appendix A: Study 1 Materials

Before starting: Ask for recording consent.

> *Are you okay with me/us recording the conversation?*
>
> *Again, this recording is used for research purposes to ensure accurate capture of your responses. It is to make sure that I/we capture your answers accurately and avoid misunderstanding. It's also good for me/us to focus on the conversation and not on taking notes. Your responses will be anonymized in any reporting of this research.*

This is a semi-structured interview. This set of questions are designed to guide the conversation and they can change during the interview.

Introduction

1. Introductions
2. Explanation of the purpose of the research
    a. *As part of this initiative we would like to interview experts, both because they have the expertise that we are interested in modeling and because they represent VCC users who are likely to take advantage of the support that the component will provide.*
    b. *The main goal of this project is to provide support and guidance to users of the VCC in the form of early identification of a VCC-level incident and notifications to stakeholders. In this interview we have two goals: Understanding how this component could be helpful for them and gathering their expert knowledge to design the model.*
    c. *We acknowledge that you may have answered variations of these questions before, but we are interested in gathering your knowledge specifically for the creation of a model that will help as an assistant in the future of VCC.*
3. Do you have questions about the procedure before we start the interview?

Interview questions

The machine learning component we want to create will serve as an assistant on the VCC early incident detection. This assistant will have vast knowledge of past incidents and when an event can become one. Our goal in this interview is to understand what information this assistant needs in order to achieve its goals and that's where your expertise is crucial.

What is a VCC Incident?

1. What would you consider to be a VCC level incident?
    - What would it take for a set of dispatch and/or events to be considered as a VCC level incident?

2. What are the steps and expert knowledge you use when identifying an incident?
    - What happens after you identify an incident?
3. The current definition of a VCC incident is: “a transportation situation that requires collaboration across agencies and roles (e.g., traffic incident management, congestion management, and population movement).”
    - What would you add to this definition?

Decision-support component (Assistant)

This component in the VCC will be able to use previous historical data to do early identification of a VCC incident.

4. What would be useful information for you to obtain from this component?
5. What could you do differently and better with an early incident identification?
6. Are there any historical data sources that can be used to train the model? Examples of previous events that could be considered as incidents for example.

What information would be useful to identify an incident?

7. What information about the incident are you interested in?

VCC Event:

This is an example of a [agency] event. This is the current information being shown in the VCC system:

8. What data would be relevant to you and your agency?
9. What data would not be relevant to you?
    - What would make it relevant?
10. What information would be useful in order to classify this event as a VCC Incident?
11. What extra information would be useful to you to make these decisions? Even if not available today.

Debrief

12. Is there something that you might not have thought about before that occurred to you during this interview?
13. Is there anything you would like to ask me?

Appendix B: Study 2 Materials

B.1 Usability Testing / Interview Session

By: Susanna Lamervo
Estimated Duration: 30 minutes
Format: Virtual, Zoom
Recording & Transcript: Audio transcript and screen record with consent

1. Intro & Consent, 2 min

- Thank you for being here today, and taking part in our study!
- Do you live in WA? What is your zip code? State ID?
- Today I will show you visualizations and I'm looking to hear your feedback as you try to use them and understand them. Your responses are confidential and there are no right or wrong answers, just your personal experiences.
- Is it okay if we record this session for note-taking, and delete the recording after analysis?

2. Share the website link(s) and explain procedure, 3 min

- I'll drop a link in the chat to the website we're testing. Once it's open, please share your screen so I can follow along.
- I'll send tasks in the chat as we go, and as you complete them, please talk out loud what you're doing and thinking while navigating the site.
- Any questions or shall we start?

3. Tasks & Questions (25 minutes)

- **Map of Population Growth**
    - **Task:** Can you find out what the expected growth rate is for 2035 in the county you currently live in?
    - **Question:** How do you feel about such a growth rate?
- **Map of Projected Traffic Levels**
    - **Task:** Explore how your travel from home to SeaTac airport will be changed from current to 2050 travel time?
        - If outside Puget: **Task:** Explore how travel from Mountlake to SeaTac airport will be changed from current to 2050 travel time?
    - **Question:** Does the travel time change between different modes of transportation? Increase/decrease?
        - If there's an increase, ask: How do you feel about such an increase in traffic time? How would that personally affect you?
- **Map of Bridge Conditions**
    - **Task:** How many fair condition bridges are there in your county and what is estimated to be the improvement costs of those bridges in total?
    - **Question:** How do you think the condition of bridges will affect your daily life in the future?

- **Traffic Mode Travel Times**
    - **Task 1:** What is the fastest projected travel mode to get from Seattle to Vancouver, Canada?
    - **Task 2:** How long does it take to travel from Portland to Vancouver by car, versus traveling by high-speed rail?
    - **Question:** What do you think will be the most efficient way for you to travel long distances in the future, and why? What mode would you prefer?

4. Final questions

1. How did you feel navigating through these tasks? Ask "why" as a follow-up
2. How do you think congestion on the roads will affect you in the future? Are you concerned? Ask "why" as a follow-up
3. How do you think rising travel times will affect you in the future? Are you concerned? Ask "why" as a follow-up

B.2 Reflexive Thematic Analysis Codes

Sensitizing Concept	Code	Definition
Making sense of the data	Term or measure clarity	Captures moments where the meaning of a displayed number, label, term, or visual variable is unclear or ambiguous to the user.
	Benchmark or baseline needed	Captures moments where a displayed value lacks a reference point or baseline that would allow users to judge whether it is high, low, normal, or meaningful.
	Source and method transparency	Captures user interest in where the data came from, how it was calculated, what benchmark or method was used, or what evidence supports the displayed values.
	Projection and scenario assumptions	Captures how users interpret, question, or accept assumptions built into future projections or modeled scenarios, including questions about what inputs, conditions, or policy choices the model assumes.
Trust and credibility	Output feels credible	Captures moments where users perceive the visualization's outputs or projections as believable, reasonable, expected, or trustworthy.
	Output feels misleading or incomplete	Captures moments where users perceive the visualization's outputs or projections as unrealistic, partial, overly simplified, or missing important real-world conditions.

	Temporal relevance and skepticism	Captures users questioning whether a given projection horizon is meaningful, credible, or personally relevant as a basis for engagement or decision-making.
	Affective engagement	Captures strong affective reactions (alarm, distress, frustration, or excitement) prompted by the visualization or the scenario it represents, where the emotional response reflects or shapes how the user engages with the credibility or significance of what is shown.
Situated knowledge and decision relevance	Professional/domain knowledge applied	Captures participants using transportation, planning, modeling, policy, or infrastructure knowledge to interpret or critique the visualization.
	Situated relevance and lived experience	Captures participants interpreting the visualization through their own mobility practices, local knowledge, past experiences, or sense of whether the issue matters to them personally.
	Concept accessibility	Captures whether technical concepts, terms, visual encodings, and interactions are presented in ways that users with different levels of domain expertise, visualization literacy, or perceptual ability can access and interpret.
	Geographic fit and representation	Captures whether the visualization represents the places, regions, routes, or communities that users care about, including feelings of exclusion, misrepresentation, or rural/urban imbalance.
	Real-world consequence reasoning	Captures moments where users move from interpreting the visualization to reasoning about what it implies for decisions, priorities, or actions, whether personal, institutional, or collective.
	Perceived personal utility	Captures moments where users reflect on whether they would use the visualization themselves, how often, and in what circumstances, including assessments of whether the tool fits their needs, habits, or decision-making context.
Visualization design and interaction	Visualization design supports interpretation and reasoning	Captures moments where the visualization's design, structure, or interactivity helps users understand what is being shown, identify patterns or comparisons, notice surprises, connect findings to prior knowledge, or form an interpretation of what the data means.

	Visualization design creates barriers	Captures design or interaction features that make users unsure what is selected, what changed, what is clickable, what a visual encoding means, or how to recover from an action.
	Requests additional data	Captures moments where users request additional data dimensions, variables, or contextual information that would enrich or extend what the visualization already shows.
	Comparative and scenario flexibility	Captures moments where users want the visualization to support comparisons, alternative views, or scenario-based exploration in order to interpret differences, test assumptions, or examine how outcomes change under different conditions.

Appendix C: Study 3 Materials

C.1 Screening Survey

#	Question	Response format/options
1	Email	Short answer
2	What is your current role?	Short answer
3	Have you participated in a qualitative coding project as part of your research or work?	Yes; No
4	What field or discipline does your qualitative coding work fall within	Short answer
5	How many qualitative coding projects have you participated in?	1–2; 3–5; More than 5; Other
6	What was your role in those projects?	Coder; Lead Coder; Codebook Designer; PI or Project Lead; Other
7	How many coders, including yourself, were present in those projects?	1; 2; 3–5; 6+
8	How familiar are you with inter-rater reliability metrics such as Cohen's kappa?	- Not familiar at all - Heard of them but never used them Understand what they measure but have not used them - Used them or been involved in assessing them - Regularly use or assess agreement metrics
9	Have you ever worked on a project where agreement across multiple coders was formally assessed with an Inter-rater reliability Metric?	Yes; No; Unsure

C.2 Interview Protocol

Estimated duration: 45–60 minutes. Conducted via Zoom with screen sharing.

Opening: Consent and Introduction

Before starting the session:

Thank you for taking the time to participate today. Before we begin, I want to walk you through a few things.

This study is conducted under the UW's IRB, and has received an exempt determination. Your participation is voluntary and you may stop at any time without any consequence.

Everything we discuss today will be anonymized. Your name will not be attached to any data, and any identifying details will be removed before analysis or reporting.

I would like to record this session for transcription purposes only. The recording will not be shared outside the research team and will be deleted after analysis.

Do you consent to being recorded today?

[Wait for verbal confirmation before starting the recording.]

Do you have any questions before we begin?

[Begin recording, then proceed to Part 1.]

Part 1: Pre-Exposure

"This session is structured in two parts. In the first part I am interested in learning more about your qualitative coding experience, so I will ask you some questions about projects you have worked on"

Goal: Capture the independent user-side baseline across all Munzner levels. **Time:** ~20 minutes

1. How long have you been doing qualitative coding?
2. Tell me about a qualitative coding project you have worked on. What was your role and what were you coding?
    - *If two coders only:* Did you work on any projects with a larger team, three or more coders? If so, can you tell me more about it?
3. What decisions were made based on the coding in that project, and who made them?
4. How did you define agreement in that project? What did it mean for coders to be aligned or not?
5. When you did find disagreement, what did you do with that information?
6. What information about agreement would have made your disagreement discussions more productive?
    - *For example, was there a moment where you wished you knew something you didn't?*
7. What visual tools or representations, if any, have you used specifically for understanding coder agreement patterns? What made them useful or trustworthy to you?
8. When you get a score or measure of agreement, what makes you trust or question it?
9. How much do you need to understand how it works to trust it?

10. Have you ever encountered a situation where an agreement score or result didn't match your own sense of what was happening in the data? What did you do?
11. If you were going to visualize coder agreement, what would that look like, and how would you want to interact with it?

Part 2: Post-exposure

Context Introduction

Before sharing the visualizations, briefly describe the context.

I am going to show you two visualizations built from a real qualitative coding project. A team of coders annotated online text data with emotion categories, and they could apply more than one emotion to the same text, so it could be tagged as both sad and angry at the same time. These visualizations are meant to be used during the coding process, not only at the end of it.

As you explore each visualization, please think out loud; tell me what you notice, what you are trying to figure out, and what is or is not clear to you. There are no right or wrong answers.

Ask participant to share screen to their browser

Part 2.a: Agreement Timeline

Focus: How disagreement is distributed across time and categories. **Time:** ~15 minutes

Share the weekly agreement by emotion visualization. Ask participants to think aloud throughout.

Open-Ended

12. Take a moment to look at this visualization. What do you notice first, and what do you think it is showing you?
13. Looking at how agreement changes over time, what patterns stand out to you?
14. How would something like this fit into the kind of coding work you described earlier?

Tasks

If the participant has not switched to the weekly view or has not tried clicking on a category to filter, point these out and ask them to explore before moving to the reflection questions.

15. Which emotion code had the lowest agreement across the full project period? How did you find it?
16. Can you find a week where a specific emotion code dropped significantly compared to the weeks around it? How did you find that?

Reflection

17. How much do you trust what this visualization is showing you? Walk me through what makes you more or less confident in what it presents.
18. Were there any features of this visualization (ways of interacting with it or information it showed) that felt accessible and useful to you?
19. Were there any features of this visualization (ways of interacting with it or information it showed) that felt like they required background knowledge you may or may not have had?
20. Is there anything here that confuses you?
21. Is there anything you would want to know more about?

Part 2.b: Coder Agreement Network

Focus: How disagreement is distributed across coders. **Time:** ~15 minutes

Share the coder agreement network visualization. Ask participants to think aloud throughout.

Open-Ended

1. Take a moment to look at this visualization. What do you notice first, and what do you think it is showing you?
2. Looking at how agreement varies across coders and emotions, what patterns stand out to you?
3. How would something like this fit into the kind of coding work you described earlier?

Tasks

If the participant is unclear on what the edges, node sizes, or colors represent, explain the legend before they attempt the tasks. If they are struggling with the K range slider or have not noticed it, point it out and ask them to explore it.

4. Can you tell me what coder had the lowest agreement with Maria? How did you find that?
5. Can you find which coders disagreed most on Sadness? How did you find that?

Reflection

6. How much do you trust what this visualization is showing you? Walk me through what makes you more or less confident in what it presents.
7. Were there any features of this visualization (ways of interacting with it or information it showed) that felt accessible and useful to you?
8. Were there any features of this visualization (ways of interacting with it or information it showed) that felt like they required background knowledge you may or may not have had?
9. Is there anything here that confuses you?

10. Is there anything you would want to know more about?

Part 2.c: Cross-Visualization Reflection

Time: ~10 minutes

No screen sharing needed for this part. Conversational.

1. Earlier you described what coder agreement information matters to you and what you imagined a visualization of it might look like. How do these visualizations compare to that?
2. Was there anything you expected to find that was missing?
3. Thinking about your own background and experience with qualitative coding, did either visualization feel well suited to how you work, or did you feel it was designed with a different kind of user in mind?
    - *Follow-up if design mismatch:* Was there a specific moment where that came through?

C.3 VSD Stakeholder Mapping

Stakeholder group	Direct / indirect	Role in agreement discussion support	Values at stake	Risk if overlooked	How this shaped the study
Developer	Direct	Makes the design choices embedded in the system and is accountable for how agreement is framed and presented.	Design accountability, interpretive validity, system credibility	Design assumptions embedded in the system may never be made visible or challenged as part of the evaluation.	Addressed through positionality statement in Section 5.3.2 and independent system-side analysis in Section 5.4.1.
IRR-familiar lead coders (assumed user)	Direct	Interpret agreement patterns and make decisions about codebook revisions, training, or further discussion.	Trust, transparency, methodological rigor, accountability	Evaluation reflects only users already equipped to interpret kappa values, missing the broader range involved in agreement discussions.	Included participants with coding leadership or team coordination experience.
IRR-unfamiliar lead coders	Direct	Facilitate agreement discussions without a formal reliability metric.	Accessibility, interpretability, methodological confidence	The system could be unusable for leads who assess agreement through discussion rather than metrics, excluding a valid form of expertise.	Did not require IRR familiarity even for participants in leadership roles.
Individual coders	Both	Participate in agreement discussions; their coding work is represented in the visualization.	Fairness, autonomy, psychological safety, interpretive legitimacy	Individual coders' interpretive perspectives would be absent from the evaluation, and the visualization's potential to expose or judge their work would go unexamined.	Included coders across the role hierarchy, not only leads.
Communities represented in data	Indirect	Their language, emotion, or experience is categorized through the coding process.	Representational accuracy, representational responsibility	Agreement optimization may flatten interpretive complexity or erase meaningful ambiguity.	Informed the analytical framing; disagreement was interpreted as potentially reflecting genuine complexity in the data, not only coder error.