Heterogeneous Uncertainty: The Impact of Quantitative and Qualitative Uncertainty in Data Pipelines

ANONYMOUS AUTHOR(S)

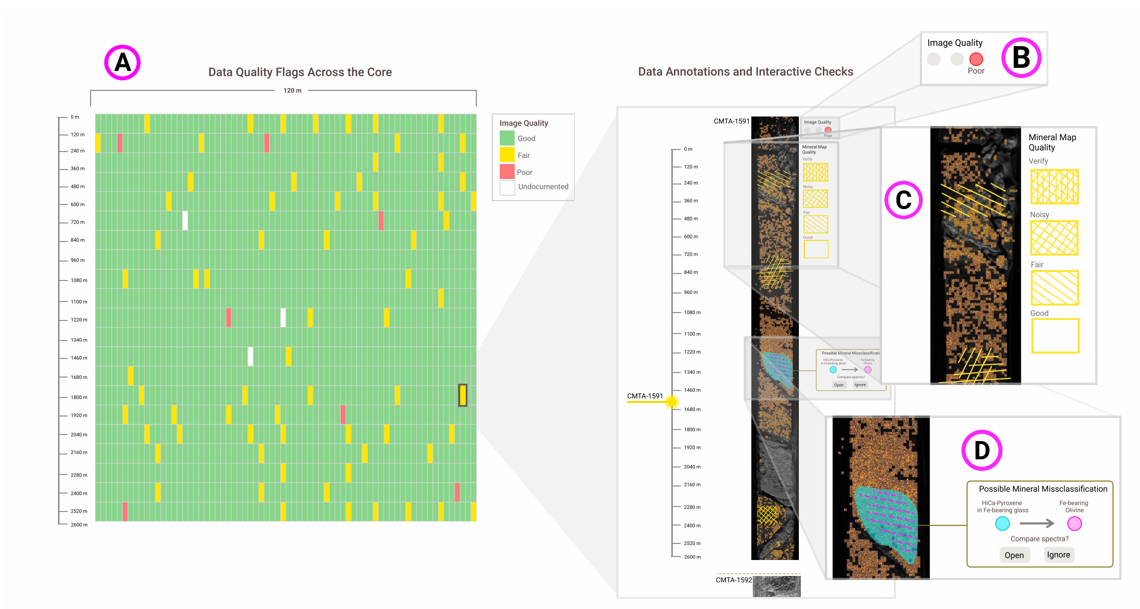

Fig. 1. Presenting system and sensor-based error. (A) An overview of data quality flags for the entire dataset where one core section is selected and viewed with mineral map. (B) A data quality flag for the opened core section. (C) Annotations of the mineral map. (D) Annotations for a region of the mineral map where a mineral *may* have been misclassified. Overview and selected core section possess axes showing depth where the section was drilled.

Effective reasoning about uncertainty remains challenging for scientific and machine learning (ML) communities in part due to its heterogeneity. Multiple sources of uncertainty contribute to imprecision in downstream analysis, yet existing approaches often bucket these distinct sources into a single measurement. This paper argues for more nuanced treatment of heterogeneous uncertainty in research and data pipelines. Through a case study of a large-scale, collaborative geophysics research project, we document the sources of heterogeneous uncertainty and identify how they contribute to “research debt”. We present an initial exploration of how these heterogeneous sources of uncertainty might be communicated beyond aggregate encodings, and demonstrate that doing so can offer greater transparency for downstream analysis.

CCS Concepts: • **Data Science** → **Uncertainty**; • **User Interviews** → *Subject Matter Experts.*

Additional Key Words and Phrases: uncertainty, data visualization, heterogeneous uncertainty, case study, subject matter experts

Manuscript submitted to ACM

ACM Reference Format:
Anonymous Author(s). 2022. Heterogeneous Uncertainty: The Impact of Quantitative and Qualitative Uncertainty in Data Pipelines . 1, 1 (July 2022), 23 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Including notions of uncertainty in lay communication, scientific discourse, and even machine learning explanations has received increasing attention [36, 37, 62]. Data visualization plays a key role in this work as uncertainty communication has a rich history in the data visualization research community [37, 63, 65, 66]. But the scope and nature of this work is often limited to a particular framing of uncertainty; in seeking to effectively communicate and understand uncertainty, information about its underlying nature is often lost. As Skeels et al. [77] point out, uncertainty is complex, multifaceted, and not always quantifiable, making it difficult to compute. Similarly, this variability introduces unique challenges to uncertainty visualization.

Heterogeneous uncertainty—wherein multiple sources of ambiguity contribute to the overall uncertainty of a system or model—offers a clear example of this difficulty. When uncertainty is presented visually, it is most frequently through a singular, cumulative encoding, such as in the height of an error bar or boxplot [64], an animation [1, 21], the width of a line chart's ribbon [65], or text and glyphs on or near a graph [7]. These binned encodings are visually simple not just for readability, but also because of the form of uncertainty we as a community have focused on: total uncertainty, a quantified, numerical value representing an apparently complete measure of data's ambiguity. Yet as [42] describe, the "drive to reduce uncertainty can lead to unwarranted expressions of certainty." Even when uncertainty arises across a data pipeline [66], this underlying heterogeneity is not often communicated, introducing potentially misleading abstractions.

For the visualization community, the absence of work navigating complex, heterogeneous uncertainty may be tied to insufficient examples where heterogeneity exists and requires unique treatment. Uncertainty is well-described in its potential introduction along a data pipeline, but most examples illustrating uncertainty provenance are not comprehensive to one dataset. To this end, we demonstrate the ways in which heterogenous uncertainty arises and might be communicated through a case study of a large, collaborative, scientific dataset of drilled oceanic core samples. The dataset is part of a collaboration with a multi-national, large-scale geophysics research project [3], and contains multiple sources of uncertainty that necessitate expert discourse. We detail four such sources of uncertainty within the dataset, contextualized by interviews with expert stakeholders, supporting the need for more extensive approaches to uncertainty communication. These touch on stochastic and epistemic uncertainty, algorithmic and interpolation uncertainty, and experimenter bias. Our work aims to explore *how* and *what* uncertainty measures might be presented to different stakeholders. Finally, we reflect on our findings and describe potential trade-offs different communities face regarding heterogeneous uncertainties. We argue that by providing more "surface"—more levels, measures, and facets of uncertainty—to interact with and query against, knowledge is extended.

2 RELATED WORK

2.1 Types of Uncertainty

Uncertainty arises from "incomplete information" about data, systems, models, or simply the state of the world [44]. It confounds stakeholders, introducing ambiguity that must be rationalized or reduced to ensure better decision-making. A number of frameworks and taxonomies of uncertainty have been proposed across a range of different domains, including economics [75], statistics [31], life sciences [71], and medicine [28]. These taxonomies can be general, but are more frequently domain-specific; at some point, practically every basic or applied research field will publish a review of the particular forms of uncertainty they face. Targeted taxonomies help communities understand the implications of *specific* uncertainties they must manage. This is useful because strategies for reducing uncertainty are shaped by the type of uncertainty and the context of the task [84]. In contrast, generalized taxonomies offer broad direction but rarely get into the minutia of underlying causes.

Within machine learning communities, uncertainty is often distinguished as either epistemic uncertainty (e.g., due to an intrinsic lack of knowledge), or aleatoric uncertainty (e.g., natural variation) [61]. These distinctions are not overly descriptive, and uncertainty taxonomies often take a step further in characterizing uncertainty's complexity. For example, Smithson [78] and Han et al [28] both propose variants of a three-dimensional taxonomy in detailing uncertainty that arises when information is characterized by probability (e.g., stochasticity), ambiguity (e.g., multiple interpretations of a single event or variable) or vagueness (e.g., imprecision or fuzziness in definitions or measurement). Other taxonomies describe uncertainty as it arises when progressing through an analysis pipeline.

For example, Pang et al. [66] explored where uncertainty may be introduced along a visualization pipeline—from measurements, data transformations, models, and even the visualization process. In the visualization literature, commonly cited types of uncertainty include error, accuracy, precision, validity, quality, variability, noise, completeness, confidence, and reliability [19]. Skeels et al. [77] classified uncertainty similarly, but added two unique measures to the list: measurement precision, completeness, inferences, *credibility*, and *disagreement*.

Uncertainty has multiple working definitions. For some domains, the term references quantifiable, measurable ambiguity. But uncertainty as defined by "incomplete information" can also be subjective, difficult to abstract into a concise measurement. Popular frameworks for uncertainty rarely addresses this qualitative form of uncertainty directly, but there are some exceptions. McCurdy et al. [57] coined the term "implicit error" to describe a type of measurement error not explicitly recorded or communicated, but inherent to an expert's interpretation of data. Through a disparately collected, heterogeneous public health dataset, McCurdy et al. [57] noted that many data discrepancies are often *not* reflected in a dataset. Instead, this unrecorded error is accounted for qualitatively by experts during analysis, based on their implicit domain knowledge. McCurdy et al. [57]'s two-part formalized framework details characteristic traits of implicit error—source, type, magnitude, direction, confidence, and extent of the error—and proposes a method to uncover these traits via domain expert interactions. Their work highlighted the value of externalizing implicit error for supporting more effective data analysis, despite the difficulty faced in capturing it. Ultimately, implicit error is pervasive in all data, as the simple act of observation—automated or not—is informed in some way by the observer.

2.2 Machine Learning and Uncertainty

Recent work within the machine learning community has discussed how different sources of uncertainty may have unique implications for model outputs [20]. Thus far, this work has been limited to differences between mislabeled (noisy) data, and unique, poorly represented data. But uncertainty is a growing question for the machine learning community, and has been related to questions of model transparency and appropriate trust [4].

Machine learning (ML) models are historically poor self-evaluators, tending to be over confident when incorrect—a result of miscalibrated uncertainty measures and under-specification [17]. As a result, recent work has explored models that self-report and self-calibrate uncertainty measures [46, 47, 67]. Recognizing sources of uncertainty in these contexts is a key to uncovering the causes for model overconfidence.

Bias, a specific form of uncertainty, has been a common thread in machine learning fairness research. This work tends to emphasize the "human" sources of bias as these can cause discriminative harm when magnified by a model [12, 69]. Such bias is difficult to accurately assess, and ML researchers outside fairness instead tend to focus on uncertainty as it relates to probabilistic outcomes and model accuracy. This is changing as the community recognizes that deployed models *fail* when implications of human bias are not considered carefully enough [40], but work separating *heterogeneous* uncertainties in modeling contexts—bias, error, and beyond—remains limited.

2.3 Uncertainty Visualization and Analysis

The responsibility of not only understanding where uncertainty exists, but also *how* to communicate it falls largely on researchers and journalists. Despite broad recognition of the need to include measures of uncertainty in visualizations, many authors are hesitant to do so—for example, in Hullman's 2019 study on uncertainty visualization authorship (which used a visualization-literate convenience sample), nearly half of respondents admitted to considering but ultimately not including uncertainty measures in their charts [36].

There are multitude of reasons why visualization authors avoid portraying uncertainty in their work. Hullman [36] describe concerns for chart comprehensibility, the quality of a reader's experience, the risks of wrongly encouraging data distrust, and the limited number of high-quality uncertainty visualization examples. Even when uncertainty is presented visually, it is often through a singular, cumulative encoding, such as in the height of an error bar or boxplot [64], an animation [1, 21], the width of a line chart's ribbon [65], or text and glyphs on or near a graph [7]. The nature of these visual artifacts hints at the balancing act visualization authors face in communicating uncertainty—they must weigh the relative trade-offs of a graph's comprehensibility with the desire to communicate nuance. These binned encodings are visually simple not just for readability, but also due to the uncertainty we as a community have focused on: total uncertainty, a quantified, numerical value representing an apparently complete measure of data's ambiguity. Yet as Kale et al. [42] describe, the *"drive to reduce uncertainty can lead to unwarranted expressions of certainty, which has consequences for decision-making individually and at an organizational level."*

2.4 Technical and Research Debt

Technical debt refers to the long-term costs of allowing insufficient artifacts within systems to remain [10, 45]. Historically, the term was used when convenient decisions in the short-term led to downstream "debt" that developers must pay back either through extra work or loss of product quality and functionality. Machine learning (ML) systems are particularly sensitive to hidden technical debt as developers must cope with both traditional code maintenance *and* debt accumulated as a feature of ML data dependencies [74]. Modeling-specific debt can be difficult to detect because it exists at system and organization levels, rather than in code. Incurring debt—by not refactoring code, minimally considering training data, forgetting documentation, or allowing unnecessary dependencies—may expedite development at first, but at the expense of compounding work in the future. For modeling, this debt can be hidden, compounding silently overtime.

Here, we use the term *research debt* in reference to decisions and artifacts within a data pipeline which also incur debt but with the added complexity that such debt may lead to imprecise or skewed scientific findings—a challenge exacerbating the "replication crisis" within research, where scientific results are found to be unreproducible [51, 52]. Reproducibility problems are often blamed on researchers' communication of procedure and analysis. However, the lack of tools supporting deliberation of alternative decisions or contextualizing ambiguity may also be blamed [42]. Software development has established best practices to mitigate technical debt, but these are not so clear in data work.

Researchers must reason with heterogeneous uncertainty when developing conclusions. This uncertainty, when not mitigated appropriately, is a form of technical debt accumulated over the course of collection, analysis, and modeling.

3 PROBLEM DOMAIN AND BACKGROUND

There are many moving parts to manage within large, collaborative research projects, each contributing to the complexity and introducing new sources of uncertainty researchers must contend with. For this reason, these collaborative projects are ideal case studies for understanding heterogeneous uncertainty across data pipelines.

We explore heterogeneous uncertainty from the framing of one such collaboration, the ICSDP Oman Drilling Project (OCDP). OCDP is a scientific research collaboration studying how the oceanic crust was formed. The project was funded to recover and analyze 3.2 km of earth core—cylindrical rock drilled and removed from the earth—recovered from the "Rosetta Stone of complex tectonic settings", a region in Oman where rock that was once ocean floor and upper mantle and has since been thrust up through the continent [38, 70]. Multiple types of data were collected, each helping researchers understand the geological events involved in the crust formation. The datasets are used to map minerals within the earth at different spatial resolutions.

Multiple methods of data collection were conducted to analyze rock composition, including: building detailed core descriptions (Figure 2D); physical sampling (Figure 2C); and X-ray, CT, and microspectroscopy scanning across the entire drilled core (Figure 2A, D). Of these methods, time-intensive compositional analyses such as thin section petrographic analysis (small physical samples viewed under microscopes for close, detailed descriptions shown in Figure 2C and X-ray diffraction were collected in areas of high interest. These samples act as ground truth for the portion of the core they were removed from, but they are not collected contiguously across the entire core. In contrast, the microspectral images are taken for every core section—creating a comprehensive view of all 3.2km of drilled earth—and are used to create "mineral maps" showing which minerals are present in a given core section. By referring to the spectral mineral maps, a geophysicist can interpolate mineral composition between physical samples. Core descriptions are critical to geological research, acting as an overview to direct research initially. Dozens of researchers collect, analyze, and publish on the Oman core data, many with differing agendas and data needs [23, 26]; here, we focus on the work and data interrelated to spatially-resolved reflectance spectra collected across the entirety of the drilled core sections.

4 METHODS

Over a span of six months, we conducted forty-eight unstructured and semi-structured interviews, participatory and co-design processes, interactive workflow observations, cognitive walk-throughs, and think-aloud sessions with hyperspectral and geophysics researchers. These sessions focused on the challenges researchers faced in their current workflows and analysis software. Sessions were recorded via audio, video, or careful notes. Recordings were transcribed, and thematic analysis was initially conducted to synthesize findings. We do not report our findings from this initial thematic analysis as many of these themes relate to systems requirements. Instead, we explore a common thread across interviews: uncertainty, in its many forms.

We group heterogeneous uncertainty uncovered in our interviews through light coding and our own expertise. In section 5, we describe sources of uncertainty specific to Oman Core. As part of our documentation process, we illustrate the nuance of disparate sources of uncertainty and how they may—in some cases—necessitate distinct forms of presentation and counteracting measures to avoid compounding research debt. We generalize our findings to the broader scientific and machine learning communities. Finally, we present a selection of initial exploratory interface and

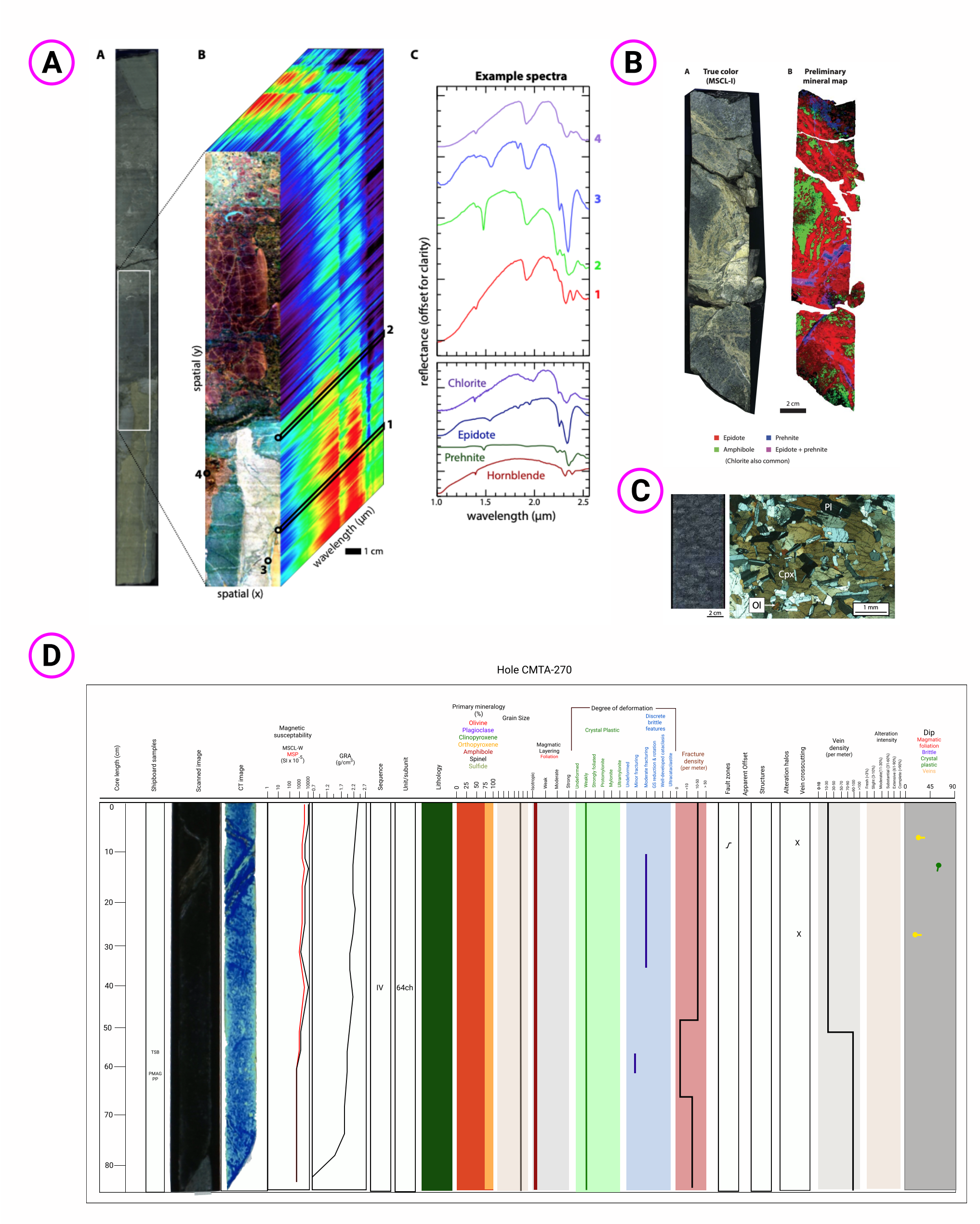

Fig. 2. Examples of data types used by Oman core researchers [26, 43]. (A) Depiction of HSI data for a singe section: (X, Y) represents pixel location, spectral bands are mapped to Z. An example spectral graph is shown with absorption features of minerals. (B) An RGB image and associated mineral map built with HSI data. (C) Petrographic thin section and section where it was extracted. (D) Abbreviated view of a core description log.

visualization designs for communicating heterogeneous uncertainty, developed through participatory and co-design sessions. We discuss the motivation behind these motifs and how they might generalize beyond Oman core data.

5 SOURCES OF UNCERTAINTY

Uncertainty was often *not* a term used by researchers. Instead, our participants described their decision making processes when working with ambiguity, and how they attempted to reduce or characterize any uncertain factors. This often implied validating the data and their work through cross-referencing other sources of information (e.g. other people, other datasets, similar examples within their own dataset, and previously published libraries or work). This triangulation occurred regularly—an effect of the complexities faced when working with multiple stakeholders on big, novel data.

The ambiguity—or uncertainty—these scientists were responding to shared many commonalities with uncertainty in other domains, but often *wasn't* communicated or recognized as a form of uncertainty. And the steps taken in response to uncertainty were often not documented (the exception being measurement imprecision or uncertainty in statistical evaluations, such as what is presented by confidence intervals). This is because reporting *quantified* uncertainty is conventional for almost all disciplines, but reporting institutional knowledge [32] or implicit error [57]—the "stuff around the edges"—is not. Through a detailed description of the Oman core dataset, we unpack how heterogeneous uncertainty is introduced across a full data pipeline. We broadly group these sources of uncertainty as follows: Human Factors in Collection Bias, Measurement Error, and Modeling. Within each category, we discuss how multiple sources of uncertainty contribute to overall imprecision. While we frame our discussion to generalize across many data pipelines, our emphasis is on how heterogeneous uncertainty presents within a single research project.

5.1 Human Factors in Collection Bias

We broadly describe systematic error accrued during collection as *Collection Bias*. There are multiple opportunities for uncertainty to be introduced as data is collected. The particular impact of these early sources of ambiguity depend on the type of data, the subject of observation, and the method of collection, each with potentially varied outcomes: biases may reduce an experimental or observed effect, amplify it, or even offset other biases [27]. In other cases, collection bias may lead to skew in what is observed or added to the dataset.

In HCI and ethical ML contexts, we often assume this type of uncertainty is a feature of human involvement. Collection bias is *not* just caused by people. In practice, researchers in pure science domains regularly account only for non-human sources of collection bias. Human factors also impact pure science, but historic desires for objectivity [18] faced by all research communities, and the difficulty of characterizing these factors in a tractable, fieldable way have led to it being regularly overlooked. As implications of these two forms of collection bias are distinct, we discuss human factors in data collection bias here and touch on non-human sources of uncertainty in subsection 5.2. Regardless, undocumented collection bias in both pure science and applied machine learning may distort reality, invalidating scientific conclusions and introducing harmful priors.

5.1.1 Oman Core. For Oman core researchers, working in a large-scale collaboration meant human factors were even more influential to data collection. The researchers worked in twelve hour shifts to extract core sections, develop core description logs, secure physical samples, and collect various imaging data. Of these artifacts, *core description logs* (Figure 2D) frame future research directions, helping researchers navigate the core by building an overview of its features. The core descriptions offer pivotal insight into how these researchers approached their work, "connecting the dots” between apparently disparate research workflows. In discussing Collection Bias, we focus on the interplay between researchers, core logs, and the other data.

Core descriptions are a fundamental to geological research as much of the work relies on visual evaluations and expert descriptions. As researchers examine the drilled rock samples, they systematically record *all* available information such as texture, patterns, grain size, apparent type, and locations where samples are taken. This information acts as an overview for the core and is then used to determine the lithology (rock type), mineralogy, potential geological history, structure and alteration zones (changes in mineral composition of rock caused by physical or chemical means). Core descriptions are sometimes published as a first contribution of larger geophysics research projects as they provide such rich sources of information. They are often saved as large excel files, filled with data, annotations, and simple visualizations documenting the entire length of the drilled core sections. This level of detail is significant, offering external viewers an intimate view of the researchers’ progression over time (in both the geological and temporal sense).

While the core descriptions are useful, they do not perfectly capture the state of the drilled rock or the granularity researchers might need. To supplement this, physical samples were simultaneously collected semi-regularly across the cores and in regions of particular interest. These samples—such as thin sections (Figure 2C) or X-ray crystallography—are used to evaluate the efficacy of initial log descriptions and pursue new research questions. They act as “ground truth” references for complex portions of the core where other methods may prove unreliable.

5.1.2 Sources of Uncertainty. We uncovered three human sources of collection bias introduced during data collection and initial core documentation: 1) evolving understanding, 2) diverging research agendas, and 3) experimenter fatigue. Largely falling into the category of implicit error, each form of collection bias may have implications for efficacy and ease of subsequent research discoveries. Our examples here emphasize the need for documenting uncertainty *provenance* (the history of data) and how we might build it retrospectively.

Evolving Understanding. During logging, researchers describe the core in great detail *as they understand it*—when deeper sections of the core are drilled, new discoveries from the physical samples occur, or novel patterns are uncovered, community understanding of these geological processes *evolves*. While these changes may be hinted at in the core documentation, the underlying shifts in knowledge leading to these changes in documentation are not clearly documented, and prior logging may not always revisited. When comparing the deepest sections of the core to the top sections, researchers must be aware of this natural learning curve because it influences what features of the rock are attended to—and where more physical samples are taken. As an example, an uncommon mineral was detected by researchers collecting hyperspectral images during core extraction. This mineral occurrence was unexpected by the geologists visually inspecting the core sections, and would have remained uncaught if not for the additional reference point. It had unique implications for the geological events for that region of oceanic crust, and the ensuing discussion led to deliberate references of the mineral within the core description. As the extraction continued, the mineral remained a focal point for the geologists to attend to and additional samples were taken in the area where it occurred. This discovery remained salient to the researchers, but—beyond documentation of sampling and the mineralogy of the rock—is not so obvious from the core descriptions.

Limited meta-descriptions of analysis provenance and pivotal shifts in knowledge are pervasive in data-driven domains [32], creating spaces for implicit bias [57]. In this context, time and resource constraints introduced by working long shifts over a short two-week period challenges how much retrospective work is reasonable. Instead, future research involving the logs often relies on researchers’ communal memories of the data collection process and any hints that can be inferred by the log notes.

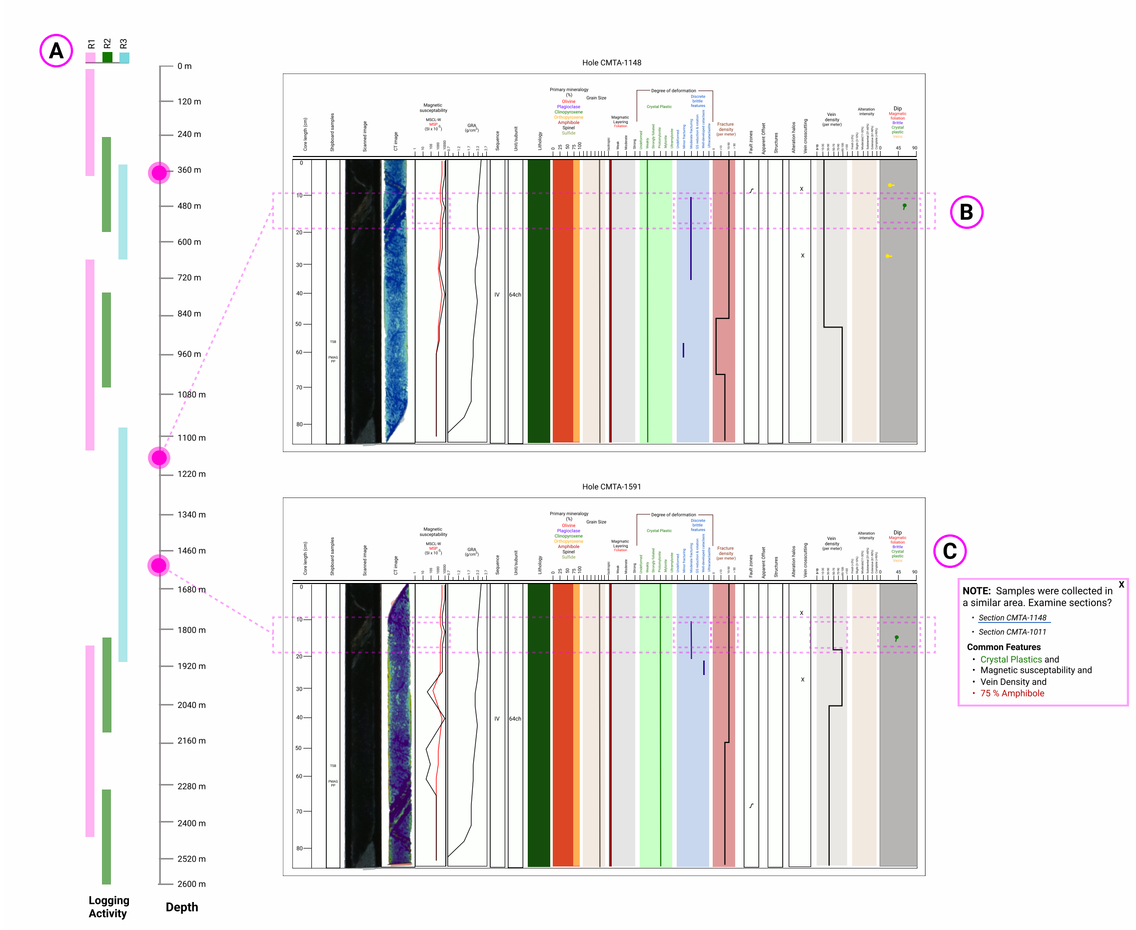

Fig. 3. Presenting human factors in collection. (A) an timeline overview of different stakeholder documentation and sampling activity. (B) selected core descriptions of sections with particular patterns of sampling based on clustered features in descriptions. (C) Highlighted similarities in descriptions marking where sampling behavior deviated, with a prompt to open related views.

Disparate Research Agendas. Variability in the researchers' agendas is similarly influential to core descriptions and physical sampling: petrologists interested in geothermal intrusions might collect thin section petrographic data largely in rock veins, while another focused on water interactions cares little about, leading to cases of sampling-based datasets biased toward a subset of rock characteristics. As cross-referencing various sampled data with descriptions and hyperspectral (HSI) images is used to evaluate the reasonableness of findings, these trends may cause frustration for those not influencing the sampling decisions, a point we re-address in subsection 5.3.

These examples of implicit error [57]—evolving understanding and diverging research agendas—are communicated only as institutional knowledge [32], and are incredibly difficult to quantify or visualize. While researchers "in the know" are aware of this implicit error in collection and modify their work accordingly, external researchers may not be cognizant of these nuances. This contributes to research debt, as new generations of researchers must either relearn prior knowledge, seek aid from others, or incorrectly interpret data without awareness of to its conception.

Experimenter Fatigue. Experimenter fatigue is well described in prior literature [39, 55] and can greatly impact the accuracy of collected data. In instances where a significant amount of data is quickly collected over short periods of time, the risks of mistakes occuring as a result of fatigue increases. Experimenter fatigue in Oman Core is difficult to measure as people switched between logging the core and other duties—no one person was responsible for logging the core information, making it difficult to account for or even recognize when the influence of fatigue is present. Further, who was responsible for documenting different sections of the core is not explicitly included within the descriptions. Researchers collecting the data mitigated fatigue through regular twelve hour shifts, but this is still an exhausting amount of time to be attentive.

5.1.3 Implications. There is dirth of examples highlighting changes in data collection methods and disparities in working agendas, yet we know that better documentation of data and data analysis is beneficial across many domains, particularly those with complex or high risk environments [58]. Evolving understanding and variable research agendas are interrelated—a research agenda is shaped by new discoveries *and* prior experiences. And the influence of these priors, which perhaps resonates most with our cultural understanding of human bias, generalizes beyond research agendas. We know bias in datasets may cause harm [69], and that it can arise from something as simple as the order in which data is presented to labelers [56]. This is true outside Oman core. For example, crowdsourced labels are often biased by personal opinions; this is unavoidable. Even *expert* annotators are not able to objectively label data without careful prompting [35]. Yet historically, these priors are not communicated with the dataset, possibly because its inclusion entails additional work for collectors. When biases *are* documented, however, it becomes possible to mitigate their downstream effects, such as through post-collection interventions [53].

Failure to properly adjust expectations to the limitations of data can have serious consequences. Jordan [41] described a now well-known case of this: a model which used "white spots" as a predictor of Down syndrome was trained on lower resolution CT scans than the real world scans the model was deployed on. Because of this discrepancy, noise in higher resolution images were incorrectly labeled as precursors for down syndrome—an avoidable consequence of poorly communicated changes to data specifications.

Even with the level of detail found in our core descriptions example, information was missing about decisions that ultimately led to observable changes in the data. This missing information contributes to overall research debt, as the "proverbial garden of forking paths" taken by researchers [25] is obscured to outsiders. We will illustrate this using an example described by our interviewees. During core extraction and early analysis, researchers did not distinguish two minerals in descriptions or other sampling documentation. At the time, the differences between the minerals did not, *"inform the scientific goal of determining trends in hydration, formation temperatures, and water chemistry with depth,"* as they had similar implications to the researchers [23]. These minerals were spectrally distinct however—a term we unpack in subsection 5.2 meaning the minerals presented differently within the dataset—and this distinction was important for a downstream modeling task. Later, when the researchers ran their classification models, this distinction needed to be revisited (which we discuss in subsection 5.3), but doing so required the expertise of involved researchers and their knowledge of what their decision implied for core documentation and research findings.

When introducing new or external researchers to datasets, much of this nuance can be lost, creating opportunities for misunderstanding and increasing the difficulty of working with the data. Outside Oman core data, evolving understanding influences the composition of labeled datasets, and research processes more generally. Questions related to how labeling hierarchies are developed and why certain decisions are made (such as for rare outliers) is important information for downstream analysis, particularly when a model incorrectly learns a concept. In order to fix these errors, we must know why it ocurred; even pervasive skew can be missed when its origins are not obvious in documentation.

5.1.4 Communication. Outside geology and similar disciplines, it is uncommon for researchers to have such information-rich documentation. Even within the Oman core dataset, information about experimenter fatigue, evolving understanding, and the different researcher agendas is not explicitly documented. Because so much information *is* detailed, however, and because the data, collection, and analysis maintains an innately sequential structure, biases present within the Oman core descriptions and sampling may be inferred retrospectively. For example, free-hand notes about interesting features of the rock are common in the core descriptions—different comment styles (such as brevity or grammar) can be distinguished using known natural language processing techniques to create timelines of the research process and researchers involved. And because the core data is extracted and evaluated in a linear fashion, changes in sampling patterns can be tracked as a surrogate measure of evolving understanding, different research agendas, and experimenter fatigue. Within the descriptions, co-occurrences of minerals, features, rock types, and experimental sampling can be grouped to into behavior profiles. From these behavior profiles, we uncover discrepancies in the data—places where an unexpected decision was made (e.g. samples were not taken when expected, or were taken unexpectedly) can be highlighted for the researchers. In doing this, we create a proxy for collection and analysis provenance; although likely imperfect, these measures provide a starting place for researchers to reflect on the project's development.

The question remains then of how we communicate these profiles and the behavior discrepancies such that they *contextualize the data*. This is important, as the presentation will influences how researchers respond to any highlighted discrepancies. From on our interviews, co-design, and participatory design sessions, we noted that this contextualization must reference the data in way that is meaningful to the users *without* impeding analysis or exploration. Further, presentation should facilitate comparisons and follow-ups by the researchers. This allows them to evaluate the validity of highlighted discrepancies. Finally, these methods of communication should not imbue artificial authority to the system or recommendation. This final point was not uncovered through our interviews, but instead borrows from known concerns for data visualizations and machine learning explanations to *discourage* critique [33] and *encourage* overtrust [13, 24]. While we as a research community are still unclear how to reduce the assumed authority of these modalities, careful, non-prescriptive language and visual "sketchiness" [8, 33] may help.

In Figure 3 and Figure 4, we illustrate a selection of visual representations communicating diverging research agendas and evolving understanding. The first, Figure 3, shows an interface where a user is evaluating the efficacy of physical sampling across the core. The user is currently considering a section where a sample *wasn't* extracted.

The interface shows four elements: a depth axis highlighting the location of the most similar sections; an adjacently aligned activity timeline showing when different stakeholders were likely logging core descriptions (Figure 3A); core descriptions with annotations hinting at similarities between sections with disparate sampling behavior (Figure 3B); and an interactive comment box noting sections may be similar and the shared features between the different sections (Figure 3C). Within the depth axis, three pink dots highlight where similar core sections are located. The activity timeline was built using the previously described behavioral profiles, allowing viewers to note who was involved in documenting which regions of the core. While not shown, this timeline must be editable as researchers are implicitly aware of *who* was documenting the core and whether the profiles correctly characterized them.

Annotations—pink, rectangular boxes—on the core log next to Figure 3C highlight an area where physical samples were *not* extracted, breaking expectations built from prior sampling in core sections with similar features. A comment box is open next to the annotation asking if the user would like to examine sections with similar characteristics. The

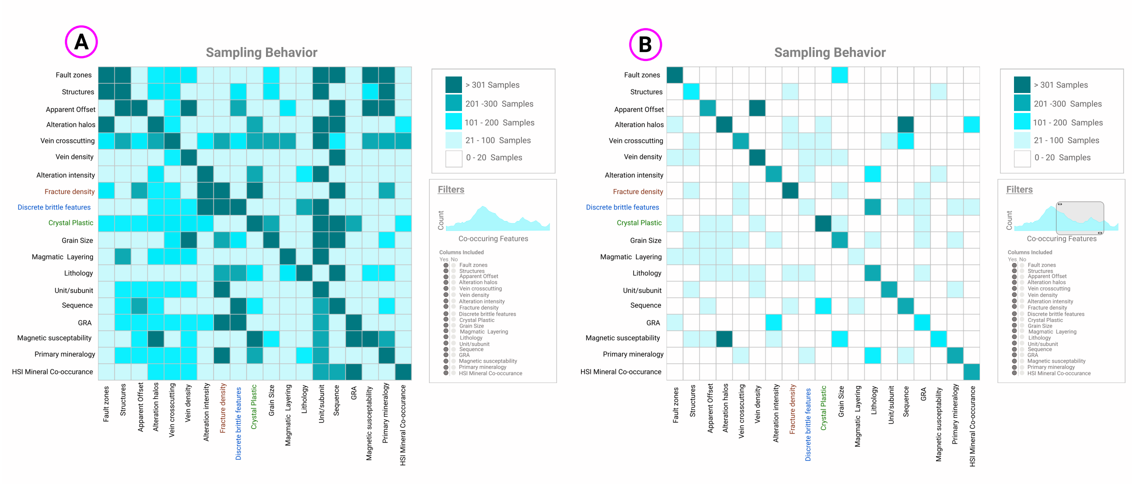

Fig. 4. An alternative approach to presenting possible human biasing effects. (A) Characteristics within core descriptions are sorted by X, Y axes. Sampling frequency where features co-occur are mapped to discrete colors. (B) Filtering by description characteristics and number of overlapping features (co-occurrences) highlights relationships researchers responsible for collection cared about in the data.

most similar regions are listed under the query, along with the features of note. These features are highlighted once more with simple annotations. The user has opened one of the sections—located at a different dot on the depth axis—and the shared features between the two documents are highlighted for the user to examine Figure 3B.

These human-sourced sources of uncertainty are often subjective, thus presenting them as abstracted values may overtly influence researchers' interpretation of the data, *further adding to research debt*. Our use of annotations to communicate these human-sourced biases borrows is a response to this. Rather than visualizing quantified metrics (which we explored extensively in our design sessions, an example of which can be seen in Figure 4), our goal was to create a modality of *support* rather than tell. We noted that many abstractions made it difficult to relate to the data, even when they offered opportunities to discover new relationships within the data. For this context, supporting required methods allowing researchers to contribute to the body of knowledge about the implicit error—a feature McCurdy et al. [57] described as an critical to characterizing the causes of error—and recommending areas where human review might be necessary.

5.2 System & Sensor-based Error

In subsection 5.1, we focused on uncertainty introduced by humans. Here, we turn to uncertainty caused by technical, computational, or sensor errors. In some cases, how people mitigate human and non-human types of uncertainty might have overlap. How we uncover and communicate the causes of uncertainty generally does not. Here, we describe three non-human sources of uncertainty: Data Corruption, Noise, and Miscalibration. Each requires a different response from researchers—in some cases, the solution is to delete or replace data. In others, there's a systematic error which *if known* can be resolved through adjustments to the dataset.

5.2.1 Oman Core. Moving forward, we focus on hyperspectral imaging (HSI) data. While Oman Core researchers collected many types of data, their HSI dataset is large, complex, and challenging to model. As a brief overview, the

Oman Core HSI dataset included images for the entire 3.2km length of drilled core—over 4,000 individual core sections. For a single section, a hyperspectral image can contain over a million pixels, with each pixel containing hundreds of spectra. Unprocessed, the raw dataset is ~31 TB. HSI data consists of hundreds of adjacent mid-infrared wavelengths of light [14]. These wavelengths (or spectral bands, shown in Figure 2A) are collected as three-dimensional image cubes $I(x, y, \lambda)$ where x and y represent pixel coordinates and λ represents the spectral bands. Researchers characterize features within the spectra to determine what materials are present within a given image. These characteristics are *spectral absorption features*, distinctive peaks and valleys in the spectral graph where light is reflected. Absorption features are unique to a material's structural and elemental composition, allowing researchers to rapidly determine the rock core's mineral composition per pixel at a micro (~85 μm/pixel visible-near infrared and ~250 μm/pixel for shortwave infrared) scale by comparing an image's absorption features to previously documented absorption features.

5.2.2 Sources of Uncertainty. Because hyperspectral data captures light reflectance, the quality of the image and the image subject is fundamental to trustworthy analysis. Data that is too noisy or inaccurate will produce sub-optimal mappings to minerals, distorting downstream analysis [6]. Trust in spectral analysis builds on the expectation that the rock face is cleanly cut, there is minimal shadowing, and that little noise is introduced during collection. We describe three sources of uncertainty tied to data quality below. Generally, each of these sources is recognizable through an expert visual check, but such manual work is arduous.

Data Corruption. The simplest example of error is when data is corrupted. For Oman core researchers, these are relatively straightforward cases to catch as the images "don't look like we'd expect", as one interviewee described, appearing similar to white noise on a screen. This is generally true for many domains—corrupted data isn't usable, and that is often immediately apparent. But when evaluating large, complex datasets like Oman Core, some instances may still slip through despite best efforts. These corrupted examples contribute to a dataset's noisiness and may harm downstream models. The reasonable response is to remove the corrupted data, assuming it is caught. In the best case scenario, the data is unimportant or easily replaced. But for Oman core researchers, replacing images is not always feasible, and removing an image disrupts the HSI continuity. In regions of the core that are geologically uninteresting, that loss might be acceptable. In others, losing rare exemplar sections could hinder research agendas. Thankfully, image corruption was not common in the Oman core dataset.

Noise. Noise—unwanted additions to the data—is harder to detect than corrupted data, but is still achievable through careful audits. To minimize noise within their HSI dataset, Oman core researchers normalized the HSI images during pre-processing to avoid uncharacteristic spikes in spectral graphs. Yet even with these processing steps, noise may still be problematic. In our interviews, there were two instances where noise was particularly problematic. We will use examples from the Oman core HSI dataset to explain. In HSI data, thresholds for what is "normal" data are pre-determined by experts. But HSI data measures how photons interact with materials, and these thresholds do not always account for unusual properties of light. The clearest example of this occurs when two minerals mix such that resulting spectra exceeds the thresholds set by researchers. This is a result of the physical mechanisms by which light (photons) excites molecules within a mineral or mineral mixture. When minerals mix together, resulting spectra may not present a linear or additive relationship—for instance, reflected light may even jump from infrared wavelengths (invisible to the human eye) to the visible spectrum []. We briefly describe the complexity this introduces to statistical analysis further in subsection 5.3. Because these thresholds are set a priori, instances where legitimate spectra surpass these threshold may be listed as noise and removed from the dataset. In this instance, legitimate features of data were treated as noise *despite* their accuracy, leading to the loss of a potentially valuable sample.

Alternatively, when HSI images containing a calibration reference are normalized, the newly normalized reflectance can present spectral features resembling minerals in normalized radiance values [9]. This is rare, but will lead to mislabeling the reference as a mineral. Here, the data *is* noisy, but because the noise wasn't distinguishable using the threshold set by the researcher it has been included in reporting. Both of these scenarios—noise appearing as signal, and signal appearing as noise—are difficulties faced by both Oman core researchers and the scientific community at large.

Artifacts in the spectra can be introduced from something as simple as the angle an image was captured at, or how the rock face was cut: if it is not reasonably flat, ridges might introduce shadows affecting how the light scatters. As HSI sensors measure photons, increased scattering *not* caused by the rock composition bungles this measurement. Some automated tools can catch artifacts by checking for images with spectra outside an expected range. Automated checks like this are not perfect, however, as the quality of raw, unprocessed HSIs is inconsistent. Oman core HSI data has a higher spectral variance compared to other hyperspectral imagery, making it technically simpler to distinguish minerals, but contains rare mixtures that automated checks are likely to incorrectly flag [23]. Spectral geologists develop an intuition for a given dataset based on prior experiences, the context in core descriptions, and similarly located samples. This allows the geologists to determine the reasonableness of absorption features in a given image for a specific region of the core—an evaluation not easily replicated by automated tools.

Miscalibration. Measurements have inherent inconsistency—we could repeatedly observe the same subject using the same method and yet still find differences in our results. This may be caused by natural variation in observed subjects, in methods of measurement, or in both [5]. For sensor users, measurement error is a particularly common phenomenon that must be constantly attend to. This is because the physical properties of sensors change over time—gravity, movement, changes in environment, and the effects of time passing all may cause a "drift" in measurement specifications. Readers might have experienced a simple example of this: a telescope is a wonderful tool to explore the night sky, but even a gentle nudge can shift it out of focus. Only by refocusing the telescope can viewers see with the clarity previously experienced. Similarly, sensors must be calibrated regularly to ensure recorded values are accurate.

Oman core researchers regularly calibrated their sensitive HSI sensors. These calibrations were two-fold: they calibrated sensors using certified laboratory light source, checking for inconsistent recordings and adjusting the sensor accordingly. Then, each spectral image is captured with a calibration reference during collection. These references are made of simple, certified materials with well-known optical properties [76]. This two-step calibration ensures that the sensor is calibrated for internal consistency *and* that the recording itself will be calibrated to the given environment at collection. If the first calibration is done poorly, a sensor may no longer be internally consistency in reporting. This type of miscalibration can not be modeled by researchers, rendering the data uninformative.

In contrast, this second calibration allows researchers to retroactively calibrate their data—these recordings are internally consistent (thus reliable signal), but are not yet translatable to other sensor measurements. This secondary calibration is neccessary because ambient light at collection may vary in lumination, distorting absorption features. By comparing the calibration reference's reflectance in the spectral image to its expected value, researchers discover how much the sensor and light conditions deviated from their laboratory settings. Once measured, this variance is used to build a 'noise' profile to normalize the image against in downstream analysis [49]. If the calibration reference is not included—or if the material in the reference is contaminated—future spectral data processing may result in incorrect minerals being associated with the image subject.

5.2.3 Implications. Measurement errors are ubiquitous, but there are several ways researchers may eliminate or minimize them. These fall into two buckets: building systems designed to reduce the effects of error *after the fact*, and dealing with error *as it arises*. Each solution exhibits trade-offs. Models and systems built to operate under uncertainty are still breakable under certain conditions, as shown by a significant amount of ML research on adversarial examples [54], and building robust systems entails a loss of accuracy [81].

Responding to error where it occurs is also not a simple task. Automated systems (such as our example of checking spectral absorption features using expected thresholds) can fail to account for cases that are unexpected by the system designer, or may mislabel real, valuable outliers within datasets as noise. Yet manual review is often intractable for large datasets. Instead, most domains have found a blending of both automated and manual approaches necessary, but this is only applicable for error that is known. Because many errors—sometimes serious ones—may go undetected for surprisingly long times [50], these automated systems must include clear descriptions of the problem. Ideally, this description includes possible causes and remedies, even tools to make correction easier.

These challenges are not unique; literature from over two decades ago also explored how to design for error [50], but new developments in computation, graphics, and human-computer interaction, and the massive need for better tools to assess big data offers rich opportunities for improvement. Mediocre responses to technically-sourced data errors, just like human biasing factors, continues to effect downstream analysis in ways that we are often not aware of. By informing expert practitioners in an *understandable* way of these varied sources when necessary, closer examination and improvement of these effects is possible.

5.2.4 Communication . One of the biggest challenges that our interviewees faced was not in *managing* system and sensor-based error, but in communicating it to their collaborators. This is understandable—measurement error is a well-described problem with known means of mitigation, but methods to communicate its implications are not. Not all of the people working on the project had expertise in interpreting HSI data, but many still relied on this data to help them interpolate findings across the drilled core. Distinguishing which portions of the data could be used for this—because it was reliable—and which could not—because there was too much noise—was important for ensuring valid research conclusions.

Instead, they expressed a desire to “flag” images that weren’t necessarily reliable. This flagging wasn’t a computed value, but rather qualitative measure developed by the experience and expertise of the spectroscopists. Their concern was a recurring theme in our design sessions, and as a result of this externalization, the researchers began including simple meta-data descriptions about the quality of the underlying HSI data and the downstream mineral maps built from them within their dataset. We incorporated this new value *and* their need for simple, contextual cues into our approaches to uncertainty communication, a sampling of which is shown in Figure 1.

In Figure 1A, we show an overview of these data flags. Because Oman core data is a thin, very long, continuous sequence of hyperspectral images, we re-order it to fit on a screen. Both x and y axes represents depth measurements. The x-axis measures the entire length of the core, with tick marks every 120m. The y-axis represents the 120m of core sections in between ticks. Data quality flags are color-coded by the label determined by researchers applied to the mineral maps: Green is a “good” data, “yellow” is fair, and red is poor, unreliable, or noisy.

Prior visualization work has explored how to communicate missing data [79]. In the Oman dataset, there were instances where data was also missing. A low-risk example of this was in the qualitative data quality labels the researchers had begun adding. In several sections, there were cases where some data did not have this label. While likely not harmful, Figure 1A includes a simple method of highlighting which sections had not received a flag, allowing researchers to easily note what portions of the core have not been labeled yet.

5.3 Modeling and Analyses

There are a multitude of ways that uncertainty arises during modeling and analysis. Here, we describe types of uncertainty introduced by modeling and statistical analysis. We ground our discussion using our Oman core example, but generalize beyond Oman core to similar contexts across domains. Our depiction of uncertainty sources is intended to highlight the variability of uncertainty and its outcome, not be a comprehensive review of all possible causes of "incomplete information". We connect the sources of uncertainty presented in subsection 5.1 and subsection 5.2 to downstream impacts on modeling, and highlight the cost of not addressing these causes earlier in the data pipeline.

5.3.1 Oman Core. Typical classification of hyperspectral data is a time-consuming, semi-manual interpretive process. Researchers must define class requirements based on the application's context as new data introduces new research agendas and new variables to consider. As an example, what is interesting and expected for the Oman core HSI data will be dramatically different when compared to HSI data recovered by the Mars rover. For Oman HSI data, these researchers might care how igneous rock occurs in the core [72]. By examining where, how, and what type of igneous rocks are present, they discover how cycles of hydrothermal activity are involved in the formation and cooling of the earth's crust. Contrasted to Oman HSI data, these same researchers might instead care about magnesium carbonates within the Martian HSI data [72]—minerals often formed by microbial activity.

This matters because the common approach to geological HSI classification begins with experts determining the likely geologic processes involved in formation of the imaged rock, using this knowledge to target probable absorption features within the new spectra. **Note**: we discussed how core description logs are developed in in subsection 5.1. Here, we can directly see the downstream influence of this bias. If the core description logs include inaccuracies, geologists analyzing HSI data may mistarget important absorption features. Once the researchers have determined which important absorption features to isolate, researchers will map the presence and relative amount of minerals within each pixel. This mapping is be accomplished through comparisons against established absorption features published in spectral libraries [15], use of unsupervised or supervised classifiers, mixture modeling, or simple algebraic operations with expected threshold values.

5.3.2 Sources of Uncertainty. Uncertainty is introduced anywhere probability, statistics, or inference is applied. This is certainly true of unsupervised or supervised learning. Here, we describe instances where uncertainty is introduced in modeling, including: 1) Library Dependencies, 2) Problems with Dimensionality, 3) Overlapping Features (non-unique absorption), 4) Overlapping Meaning, (Substitutions), 4) Variable Noise Profiles, and 5) Blindspots. If left unaddressed, each of these may contribute to research debt.

Library Dependencies. Interpolating between a gold standard and the real world is a common challenge in research. HSI data is particularly complex, but many communities experience similar difficulties. HSI classification is semi-manual *because* there is so much natural variability in a spectra that algorithmic classification cannot solely be relied on. This logic can be applied to machine learning applications. When supervised models are trained on a labeled dataset, they develop expectations from this data. The model will perform badly in deployment if real data is out of distribution to the training data. Instead, datasets must be regularly updated with new examples to fill in missing data.

As we previously described, different minerals have different chemical compositions, and thus different spectra. Sampled spectra are compared to libraries of known mineral reflectance spectra, allowing researchers to build mineral maps across the entire core (shown in Figure 2B). These libraries provide a baseline against which all naturally occurring materials can be compared against, but using them comes with a drawback: they were built using the reflectance of minerals in a pure, powdered form. We've described how sensitive HSI sensors are to changes in the rock or ambient light. They are also sensitive to changes in material—powdered minerals are not a perfect spectral match for minerals in solid rock. Features from these gold standard examples must be translated, often through trial and error, to match the specific application needs.

In a similar vein (pun intended), spectral libraries measure absorption features for *pure* minerals, but minerals do not occur in that form in nature. Instead, minerals are often mixed together. Spectra are complex, nonlinear functions of particle size, abundance, material opacity, and surface type [30, 68]. When multiple minerals are mixed together, photons of light can interact with the materials such that the resulting spectra far exceeds the thresholds set by mineral libraries and the geologists. These mixtures are difficult to characterize, a task that grows increasingly complex when more than two component minerals are involved [83]. Researchers must do some form of "spectral unmixing" to measure presence and abundance of minerals, typically through a linear model, or by incorporating principle component (PCA) or Gaussian mixture models. Because of the modeling complexity, algorithms are often be finely tuned to measure a specific mineral. This may negatively impact recognition of other minerals. In our example, there are dozens of minerals in the core and their presence is not consistent—optimizing for one mineral must be done carefully.

Overlapping Features. Misclassification often occurs when there is not enough signal within by a dataset for a model to correctly learn to label or group a sample. This type of uncertainty can usually be reduced by increasing the number of examples within the dataset that match the confusing data point [20, 32, 73], but there are instances where the error still can not be reduced. When the method of observation used to develop the dataset cannot capture distinguishing features—because the classes share *overlapping features*—classification using said data becomes intractable.

We'll illustrate this using an example shared by an interviewee: before this point, we've described absorption features as specific to a mineral, but this is not always the case. There are occasionally minerals that are indistinguishable in the spectra, even to an expert, but these spectrally similar minerals can have dramatically different implications. Instances of indistinguishable minerals being misclassified are not uncommon, leading to systemic over- or under-reporting [34]. When minerals are indistinguishable, researchers are faced with a conundrum. Using the data that they have, how do they differentiate the minerals?

Algorithmic approaches fail to surface potential misclassifications under these contexts, and the researchers likely will not catch the incorrect label unless prompted. Because the HSI data does not capture a meaningful signal, the researchers must turn to external resources—e.g., thin sections and core descriptions—to correctly relabel the core HSI data. Similar misclassification behaviors do occur in other science domains, and these also require manual overview and external data to uncover.

Overlapping Meaning. In contrast to overlapping features, some classes actually have *overlapping meaning*. This source of ambiguity includes places where apparently distinct features or concepts have a shared interpretation, and should thus be grouped or treated the same downstream. One interviewee described an instance of this within HSI data. Occasionally, an element within an mineral will be substituted for another (such as iron and magnesium, which often switch). This will change the spectra. The minerals with substitution will then deviate from the prototypical presentation of spectral libraries, but this change will not impact the implications of the mineral's presence. This is generally more of an annoyance than a concern, as simple heuristics for post-classification interventions can resolve this discrepancy. These guidelines must be developed by a priori by researchers, however, otherwise they will not be distinguished in the model outputs and may lead to missteps in researchers' interpretation.

Blindspots. There are times where the sensor or tool of observation simply does not work for a given subject. Cases of this might include cameras that are too low resolution to catch small changes in the environment, or measuring non-conductive fluid (e.g. pure water) flow rates with a magnetic flow meter (a popular meter used in plumbing that measures the voltage of fluids). Mismatches between the sensing technology and the material of interest can lead to skewed measurements—in our plumbing example, downstream effects may include ruptured pipes and water damage. In research, these mismatches might be an effect of best fit.

In Oman core HSI data, this mismatch means there are *invisible minerals*. The researchers are aware of this, but chose to use the HSI data because it is effective at capturing the majority of mineral presences in high resolution. Quartz is spectrally featureless in visible, near-infrared and shortwave-infrared HSI sensing [22]. These minerals do not show up in the spectra but are known to be present throughout the core. Wrangling with this unknown is difficult for modeling and analysis, is easily missed by researchers, and yet is still important for valid results.

Problems with Dimensionality. Because there are few training datasets likely to generalize to the specific context, unsupervised learning is a popular means for grouping mineral spectra. Spectral data is highly dimensional, and there are "infinitely many" naturally-occurring infrared spectra, a well know weakness for clustering approaches [23]. Unsupervised clustering involves translating a sample into vector space, then grouping samples via distance measures. In highly dimensional data, samples appear dissimilar in too many ways, impeding the grouping.

Typically, dimensionality reduction is applied to dataset to mitigate this problem, but performing dimensionality reduction on HSI data is difficult—clusters in HSI are typically nonlinear, and may have class overlap, rendering the reduction an overly lossy function [60]. This is true of any dataset with nonlinear relationships, requiring careful reflection and evaluation when applying these techniques lest they lead to misclassification from signal loss.

Noise Profiles Vary. As we discussed in subsection 5.2, noise profiles in hyperspectral data vary widely depending on target, imaging conditions, instrumentation, and calibration, and can have distinct spatial or spectral structures [16, 48, 59]. Because of this, classification can be difficult—researchers must either curate parameters and thresholds to match the influence of the noise, or they must normalize the data using the differences between the target and known conditions. When the noise varies, these adjustments will not match all of the different needs in the data. If the same noise profile is incorrectly applied across all HSI images, the results may be incomparable, and the resulting classifications imprecise.

5.3.3 Implications. Sources of uncertainty introduced early in a data pipeline accumulate and compound within the modeling stage. When possible, these uncertainties are best responded to *where they arise*, as the complexity of characterizing the uncertainty grows as time passes, memory fades, documentation is lost, and people stop iterating on the dataset. Overtime, unaddressed uncertainty can become a permanent feature. Examples of this can be found even in published training datasets [69, 80], where failures to appropriately evaluate uncertainty within the data can become very public.

Sources of imprecision, noise, and ambiguity modeling are common, particularly in domains that require significant human overview. For Oman Core, automating this labor-intensive workflow faces many challenges. Recent work by [2] discussed how people *want* to closely interact with their data, and that abstractions can be frustrate attempts at sensemaking. Similarly, Oman core researchers wanted to frequently examine the raw data, switching between multiple views to understand the context and implications of analysis and giving special attention to portions of data or analysis that likely introduce more uncertainty. For big data, this can be an impossibly large task. Instead, tools that facilitate targeting human oversight are helpful; characterizing uncertainty introduces more considerations for what such automated systems might target, and should hint at possibly appropriate mitigation responses.

One nuanced source of research debt has particular overlap with the technical debt faced in software development: library dependencies. In our HSI example, researchers used gold standard spectral libraries to set hard-coded thresholds and values within their analysis scripts. These libraries will often be replaced in the future with new, more accurate or context-appropriate gold standards (for example, spectral libraries of powdered minerals will be replaced with their solid counterparts). When this happens, researchers must contend with the remnants of prior work and answer the following questions: how do we compare our future analyses to our old? In some cases, the work is important enough to repeat classification using new spectra. In others, trial and error may provide a means to map between the two eras of work. This translation will only be possible if there is data (and uncertainty) provenance to contextualize the dataset.

Finally, library or data dependencies often do not translate to the real world, in all its variability. These dependencies introduce many opportunities to misclassify real data that is out of expected distribution. Recent work is exploring how to surface cases where data *is* out of distribution and not noisy or mislabeled [29, 82]. These methods may ameliorate the effects of uncertainty which can only be addressed *after* modeling.

In software, updates to package dependencies require extensive refactoring. Technical debt is accumulated when these updates are not completed. Paralleling this, data dependencies—an aspect of any project where protocols or data specifications change—accrue research debt when poorly managed.

5.3.4 Communication . Many sources of uncertainty are present within modeling tasks. Many of these are downstream uncertainties are caused by earlier steps in a data pipeline. This was reflected in our design sessions and interviews, where we found that communicating downstream uncertainty was *also* beneficial for modeling tasks as it facilitated reasoning. For this reason, Figure 1 includes uncertainty from prior steps *and* uncertainty introduced in modeling.

For each of our discussed sources of uncertainty, researchers' needs differed—in the case of invisible minerals, a spectroscopist must account for absent quartz post-analysis by relying on other data sources, e.g. core descriptions and thin section sampling, but the presence or lack thereof of quartz may not immediately influence their research process. This discrepancy can be managed later as long as it is properly documented and communicated to other researchers interpreting the data.

Random sources of uncertainty have errors that vary according to each analysis, affecting the relative precision of different results. This stochastic uncertainty is familiar, and existing methods for communicating it may be effective enough when data is presented through common graphs and charts. However, hyperspectral images and their subsequent mineral maps are not a typical format for data visualization. When quantified, individual pixels may present unique error values. These errors were important to the researchers, as the mineral maps documented the entirety of the core, augmenting or "filling in the gaps" of data collected more sparsely.

To this end, our design goals focused on how to situate uncertainty measures within the particular context of the data. Here, we used "sketchy" [33] crosshatching to annotate uncertainty on the images, as shown in Figure 1 C. These sketchy annotations highlight regions within the HSI data where uncertainty in mineral map classification—either numerical uncertainty based on algorithmic parameters, or the qualitative meta-labels included by the researchers—is present. In this particular case, we use the researchers' labels. The level of crosshatching in the region is based on the significance of the uncertainty: areas where the data is less trustworthy will have more annotations.

6 DISCUSSION AND FUTURE WORK

Existing uncertainty taxonomies document multiple sources of uncertainty, yet this heterogeneity is rarely reflected in the presentations we use to direct next steps. This may be tied to how we think about reducing uncertainty—actions taken to reduce uncertainty within a system depend on the source and context, but we tend to compress uncertainty to cumulative measures for both modeling and communication purposes. This creates a useful, though artificial, simplicity. For researchers in the midst of developing tools or deploying models, simplicity may obfuscate important information.

We know that the presence of uncertainty in visualizations, which is marked as part and parcel of either data or associated statistical models, is intended to contextualize known ambiguity and improve understanding between readers and the data. Our discussions of heterogeneous uncertainty, and our motivations in communication methods, are intended to support dialogue not just between the analyst and the data, but between experts, analysts, and data more broadly.

We posit that total uncertainty represented through a singular encoding channel may limit *experts* in fundamental ways. We have discussed how heterogeneous uncertainty may unfold within a single data pipeline and the research debt it may accrue. Just as implicit error is one facet of uncertainty requiring deeper engagement, compounding effects of heterogeneous uncertainty requires expert dialogue to build clarity. Externalizing the process of synchronizing, validating, and enhancing interpretation across multiple sources of uncertainty can inform error mitigation, and is particularly critical in cases where multiple experts collaborate around a dataset. Research debt, just like technical debt, may be paid down by refactoring, documenting, iterating, and refining data and analyses processes.

Tools supporting researchers in building empirical certainty and replicability, we argue, may benefit from exposing individual facets of uncertainty. Similarly, Future work may explore tools for developing uncertainty provenance. Through unpacking heterogeneous uncertainty, we begin to ensure appropriate trust is placed in data [11].

REFERENCES

[1] Gregor Aisch. 2019. Why we used jittery gauges in our live election forecast. https://www.vis4.net/blog/2016/11/jittery-gauges-election-forecast/

[2] Lyn Bartram, Michael Correll, and Melanie Tory. 2021. Untidy Data: The Unreasonable Effectiveness of Tables. *arXiv preprint arXiv:2106.15005* (2021).

[3] Andreas Beinlich, Oliver Plümper, Esmée Boter, Inigo A Müller, Fatma Kourim, Martin Ziegler, Yumiko Harigane, Romain Lafay, Peter B Kelemen, and Oman Drilling Project Science Team. 2020. Ultramafic rock carbonation: Constraints from listvenite core BT1B, Oman Drilling Project. *Journal of Geophysical Research: Solid Earth* 125, 6 (2020), e2019JB019060.

[4] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 401–413.

[5] J Martin Bland and Douglas G Altman. 1996. Measurement error. *BMJ: British medical journal* 312, 7047 (1996), 1654.

[6] Barbara Boldrini, Waltraud Kessler, Karsten Rebner, and Rudolf W Kessler. 2012. Hyperspectral imaging: a review of best practice, performance and pitfalls for in-line and on-line applications. *Journal of near infrared spectroscopy* 20, 5 (2012), 483–508.

[7] Georges-Pierre Bonneau, Hans-Christian Hege, Chris R. Johnson, Manuel M. Oliveira, Kristin Potter, Penny Rheingans, and Thomas Schultz. 2014. *Overview and State-of-the-Art of Uncertainty Visualization.* Springer London, London, 3–27. https://doi.org/10.1007/978-1-4471-6497-5_1

[8] Nadia Boukhelifa, Anastasia Bezerianos, Tobias Isenberg, and Jean-Daniel Fekete. 2012. Evaluating Sketchiness as a Visual Variable for the Depiction of Qualitative Uncertainty. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2769–2778. https://doi.org/10.1109/TVCG.2012.220

[9] Anna Brook and Eyal Ben Dor. 2011. Spectral quality indicators for hyperspectral data. In *2011 3rd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS)*. 1–5. https://doi.org/10.1109/WHISPERS.2011.6080934

[10] Nanette Brown, Yuanfang Cai, Yuepu Guo, Rick Kazman, Miryung Kim, Philippe Kruchten, Erin Lim, Alan MacCormack, Robert Nord, Ipek Ozkaya, et al. 2010. Managing technical debt in software-reliant systems. In *Proceedings of the FSE/SDP workshop on Future of software engineering research*.

47–52.

[11] Peter Buneman, Sanjeev Khanna, and Tan Wang-Chiew. 2001. Why and where: A characterization of data provenance. In *International conference on database theory*. Springer, 316–330.

[12] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.

[13] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. https://doi.org/10.1109/ICHI.2015.26

[14] Chein-I Chang. 2003. *Hyperspectral imaging: techniques for spectral detection and classification*. Vol. 1. Springer Science & Business Media.

[15] Nikita V Chukanov. 2013. *Infrared spectra of mineral species: extended library*. Springer Science & Business Media.

[16] Roger N Clark, Trude VV King, Matthew Klejwa, Gregg A Swayze, and Norma Vergo. 1990. High spectral resolution reflectance spectroscopy of minerals. *Journal of Geophysical Research: Solid Earth* 95, B8 (1990), 12653–12680.

[17] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395* (2020).

[18] Lorraine Daston and Peter Galison. 2021. *Objectivity*. Princeton University Press.

[19] Suzana Djurcilov, Kwansik Kim, Pierre Lermusiaux, and Alex Pang. 2002. Visualizing scalar volumetric data with uncertainty. *Computers & Graphics* 26, 2 (2002), 239–248.

[20] Daniel D'souza, Zach Nussbaum, Chirag Agarwal, and Sara Hooker. 2021. A Tale Of Two Long Tails. *arXiv preprint arXiv:2107.13098* (2021).

[21] Charles R. Ehlschlaeger, Ashton M. Shortridge, and Michael F. Goodchild. 1997. Visualizing spatial data uncertainty using animation. *Computers & Geosciences* 23, 4 (1997), 387–395. https://doi.org/10.1016/S0098-3004(97)00005-8 Exploratory Cartograpic Visualisation.

[22] Nicolas Francos, Gila Notesco, and Eyal Ben-Dor. 2021. Estimation of the Relative Abundance of Quartz to Clay Minerals Using the Visible–Near-Infrared–Shortwave-Infrared Spectral Region. *Applied Spectroscopy* (2021), 0003702821998302.

[23] Angela F Gao, Brandon Rasmussen, Peter Kulits, Eva L Scheller, Rebecca Greenberger, and Bethany L Ehlmann. 2021. Generalized Unsupervised Clustering of Hyperspectral Images of Geological Targets in the Near Infrared. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4294–4303.

[24] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.

[25] Andrew Gelman and Eric Loken. 2013. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University* 348 (2013).

[26] Rebecca N Greenberger, Michelle Harris, Bethany L Ehlmann, Molly Crotteau, Peter B Kelemen, Craig E Manning, and Damon AH Teagle. 2020. Hydrothermal Alteration and Mineralogy of the Basaltic/Gabbroic Ocean Crust: Insights from Microimaging Spectroscopy of the Oman Drilling Project Cores. In *AGU Fall Meeting Abstracts*, Vol. 2020. P079–0005.

[27] Paul Gustafson. 2003. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press.

[28] Paul KJ Han, William MP Klein, and Neeraj K Arora. 2011. Varieties of uncertainty in health care: a conceptual taxonomy. *Medical Decision Making* 31, 6 (2011), 828–838.

[29] Dan Hendrycks and Kevin Gimpel. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136* (2016).

[30] Rob Heylen, Mario Parente, and Paul Gader. 2014. A review of nonlinear hyperspectral unmixing methods. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 7, 6 (2014), 1844–1868.

[31] James S Hodges. 1987. Uncertainty, policy analysis and statistics. *Statistical science* (1987), 259–275.

[32] Aspen Hopkins and Serena Booth. 2021. Machine Learning Practices Outside Big Tech: How Resource Constraints Challenge Responsible Development. (2021).

[33] Aspen K Hopkins, Michael Correll, and Arvind Satyanarayan. 2020. VisuaLint: Sketchy in situ annotations of chart construction errors. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 219–228.

[34] Briony HN Horgan, Edward A Cloutis, Paul Mann, and James F Bell III. 2014. Near-infrared spectra of ferrous mineral mixtures and methods for their identification in planetary surface spectra. *Icarus* 234 (2014), 132–154.

[35] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.

[36] Jessica Hullman. 2019. Why authors don't visualize uncertainty. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 130–139.

[37] Jessica Hullman, Xiaoli Qiao, Michael Correll, Alex Kale, and Matthew Kay. 2018. In pursuit of error: A survey of uncertainty visualization evaluation. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 903–913.

[38] Adrian Immenhauser and Robley K Matthews. 2004. Albian sea-level cycles in Oman: the 'Rosetta Stone'approach. *GeoArabia* 9, 3 (2004), 11–46.

[39] Michael D Johnson, Donald R Lehmann, and Daniel R Horne. 1990. The effects of fatigue on judgments of interproduct similarity. *International Journal of Research in Marketing* 7, 1 (1990), 35–43.

[40] Dustin Jones. 2021. Facebook apologizes after its AI labels black men as 'primates'. *NPR* (2021).

[41] Michael I. Jordan. 2019. Artificial Intelligence—The Revolution Hasn't Happened Yet. *Harvard Data Science Review* 1, 1 (1 7 2019). https://doi.org/10.1162/99608f92.f06c6e61 https://hdsr.mitpress.mit.edu/pub/wot7mkc1.

[42] Alex Kale, Matthew Kay, and Jessica Hullman. 2019. Decision-making under uncertainty in research synthesis: Designing for the garden of forking paths. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[43] P.B. Kelemen, J.M. Matter, D.A.H. Teagle, J.A. Coggon, and the Oman Drilling Project Science Team. 2020. Proceedings of the Oman Drilling Projec. *International Ocean Discovery Program* (2020).

[44] Mykel J Kochenderfer. 2015. *Decision making under uncertainty: theory and application.* MIT press.

[45] Philippe Kruchten, Robert L Nord, and Ipek Ozkaya. 2012. Technical debt: From metaphor to theory and practice. *Ieee software* 29, 6 (2012), 18–21.

[46] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International Conference on Machine Learning*. PMLR, 2796–2804.

[47] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 6405–6416.

[48] EK Leask, BL Ehlmann, MM Dundar, SL Murchie, and FP Seelos. 2018. Challenges in the search for perchlorate and other hydrated minerals with 2.1-μm absorptions on Mars. *Geophysical research letters* 45, 22 (2018), 12–180.

[49] Jeremy M Lerner, Nahum Gat, and Elliot Wachman. 2010. Approaches to spectral imaging hardware. *Current protocols in cytometry* 53, 1 (2010), 12–20.

[50] Clayton Lewis and Donald A Norman. 1995. Designing for error. In *Readings in Human–Computer Interaction*. Elsevier, 686–697.

[51] Eric Loken and Andrew Gelman. 2017. Measurement error and the replication crisis. *Science* 355, 6325 (2017), 584–585.

[52] Helen Longino. 2002. The social dimensions of scientific knowledge. (2002).

[53] Lingyu Lyu, Mehmed Kantardzic, and Tegjyot Singh Sethi. 2019. Sloppiness mitigation in crowdsourcing: detecting and correcting bias for crowd scoring tasks. *International Journal of Data Science and Analytics* 7, 3 (2019), 179–199.

[54] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[55] Michael R Maniaci and Ronald D Rogge. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* 48 (2014), 61–83.

[56] Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2017. Sequence effects in crowdsourced annotations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2860–2865.

[57] Nina McCurdy, Julie Gerdes, and Miriah Meyer. 2018. A framework for externalizing implicit error using visualization. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 925–935.

[58] Michael Muller, Christine T Wolf, Josh Andres, Michael Desmond, Narendra Nath Joshi, Zahra Ashktorab, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Evelyn Duesterwald, et al. 2021. Designing Ground Truth and the Social Life of Labels. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[59] Scott Murchie, R Arvidson, Peter Bedini, K Beisser, J-P Bibring, J Bishop, J Boldt, P Cavender, T Choo, RT Clancy, et al. 2007. Compact reconnaissance imaging spectrometer for Mars (CRISM) on Mars reconnaissance orbiter (MRO). *Journal of Geophysical Research: Planets* 112, E5 (2007).

[60] James M Murphy and Mauro Maggioni. 2018. Unsupervised clustering and active learning of hyperspectral images with nonlinear diffusion. *IEEE Transactions on Geoscience and Remote Sensing* 57, 3 (2018), 1829–1845.

[61] Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

[62] Priyanka Nanayakkara and Jessica Hullman. 2020. Toward Better Communication of Uncertainty in Science Journalism. (2020).

[63] Gregory M. Nielson and Bernd Hamann. 1991. The Asymptotic Decider: Removing the Ambiguity in Marching Cubes. In *Proc. Visualization*. IEEE Computer Society, Los Alamitos, 83–91. https://doi.org/10.1109/VISUAL.1991.175782

[64] Chris Olston and Jock D Mackinlay. 2002. Visualizing data with bounded uncertainty. In *IEEE Symposium on Information Visualization, 2002. INFOVIS 2002*. IEEE, 37–40.

[65] Lace Padilla, Matthew Kay, and Jessica Hullman. [n.d.]. Uncertainty visualization. ([n. d.]).

[66] Alex T Pang, Craig M Wittenbrink, Suresh K Lodha, et al. 1997. Approaches to uncertainty visualization. *The Visual Computer* 13, 8 (1997), 370–390.

[67] Nicolas Papernot and Patrick McDaniel. 2018. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765* (2018).

[68] F Poulet and S Erard. 2004. Nonlinear spectral mixing: Quantitative analysis of laboratory mineral mixtures. *Journal of Geophysical Research: Planets* 109, E2 (2004).

[69] Vinay Uday Prabhu and Abeba Birhane. 2020. Large image datasets: A pyrrhic win for computer vision? *arXiv preprint arXiv:2006.16923* (2020).

[70] Oman Drilling Project. 2004. nternational Continental Scientific Drilling Program. *GeoArabia* 9, 3 (2004), 11–46.

[71] Helen M Regan, Mark Colyvan, and Mark A Burgman. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. *Ecological applications* 12, 2 (2002), 618–628.

[72] Eva L Scheller, Carl Swindle, John Grotzinger, Holly Barnhart, Surjyendu Bhattacharjee, Bethany L Ehlmann, Ken Farley, Woodward W Fischer, Rebecca Greenberger, Miquela Ingalls, et al. 2021. Formation of Magnesium Carbonates on Earth and Implications for Mars. *Journal of Geophysical Research: Planets* 126, 7 (2021), e2021JE006828.

[73] Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Mądry. 2018. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285* (2018).

[74] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015), 2503–2511.

[75] George Lennox Sharman Shackle. 2010. *Uncertainty in economics and other reflections.* Cambridge University Press.

[76] Muhammad Saad Shaikh, Keyvan Jaferzadeh, Benny Thörnberg, and Johan Casselgren. 2021. Calibration of a Hyper-Spectral Imaging System Using a Low-Cost Reference. *Sensors* 21, 11 (2021), 3738.

[77] Meredith Skeels, Bongshin Lee, Greg Smith, and George G Robertson. 2010. Revealing uncertainty for information visualization. *Information Visualization* 9, 1 (2010), 70–81.

[78] Michael Smithson. 2012. The many faces and masks of uncertainty. In *Uncertainty and risk*. Routledge, 31–44.

[79] Hayeong Song and Danielle Albers Szafir. 2018. Where's my data? evaluating visualizations with missing data. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 914–924.

[80] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. 2020. From imagenet to image classification: Contextualizing progress on benchmarks. In *International Conference on Machine Learning*. PMLR, 9625–9635.

[81] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* (2018).

[82] Weikai Yang, Zhen Li, Mengchen Liu, Yafeng Lu, Kelei Cao, Ross Maciejewski, and Shixia Liu. 2020. Diagnosing concept drift with visual analytics. *arXiv preprint arXiv:2007.14372* (2020).

[83] Hengqian Zhao and Xuesheng Zhao. 2019. Nonlinear unmixing of minerals based on the log and continuum removal model. *European Journal of Remote Sensing* 52, 1 (2019), 277–293.

[84] Torre Zuk and Sheelagh Carpendale. 2007. Visualization of uncertainty and reasoning. In *International symposium on smart graphics*. Springer, 164–177.