DESIGNING FROM DATA TO DISCOVERY: HUMAN-CENTERED MACHINE LEARNING FOR INTERPRETABLE SCIENTIFIC DATA EXPLORATION

A Dissertation Proposal
Presented to
The Academic Faculty

By

Austin P Wright

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in Machine Learning in the
College of Computing
School of Computational Science and Engineering

Georgia Institute of Technology

April 2025

DESIGNING FROM DATA TO DISCOVERY: HUMAN-CENTERED MACHINE LEARNING FOR INTERPRETABLE SCIENTIFIC DATA EXPLORATION

Thesis committee:

Dr. Duen Horng Chau
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. B. Aditya Prakash
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Kai Wang
School of Computational Science and Engineering
Georgia Institute of Technology

Dr. Alex Endert
School of Interactive Computing
Georgia Institute of Technology

Dr. Scott Thomas Davidoff
Human-Computer Interaction Instititute
Carnegie Mellon University

Date approved: April 22, 2025

How often people speak of art and science as though they were two entirely different things, with no interconnection. [...] That is all wrong. The true artist is quite rational as well as imaginative and knows what he is doing; if he does not, his art suffers. The true scientist is quite imaginative as well as rational, and sometimes leaps to solutions where reason can follow only slowly; if he does not, his science suffers.

Isaac Asimov, The Roving Mind

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

It is often ignored how scientific discovery is a social activity done by people, and thus designing the statistical and computational tools that these people use to explore novel and complex data requires a human-centered approach. While modern data mining methods purport to be able to assist in this endeavor by making more data types more amenable to visualization and analysis; very frequently these methods, when straightforwardly applied, do not solve the right problems that actual scientists face in their workflows. What is needed are better *human-centered principles of applied data-science* that takes into account this divide between scientific users and existing machine learning problem formulations.

This thesis contributes towards precisely that goal, using extensive embedded field work to identify and understand the needs of specific groups of scientists across multiple domains, and designing new machine learning tools to address them. From this concrete basis I develop frameworks for the centering of people in the meta-process of the *design of machine learning models within their total context of actual scientists' processes of scientific discovery*; a synthesis of **machine learning** (ML) theoretic and **human-computer interaction** (HCI) methodological frameworks. This work is therefore structured into two interrelated thrusts:

(1) **Human-Centered Discovery Frameworks** where based on embedded user research on scientists working collaboratively in context, I develop frameworks for understanding the human processes of scientific discovery, model how ML systems interact with these processes, and create guidelines for improving the design of such systems.

(2) **Interpretable ML for Exploratory Science** in which I utilize these frameworks to collaborate with scientists and develop novel interpretable ML methods that address the particular problems of scientific users doing exploratory data analysis. Altogether this work contributes to scholarship in in ML, HCI, and scientific domains.

CHAPTER 1
THESIS OVERVIEW

1.1 Thesis Goal: Vision and Motivation

With the advent of modern machine learning, epitomized by large neural networks, recent years have seen an explosion of powerful new tools in a wide range of domains, all of which are predicated upon the framework of data-centric modeling [1], a style of modeling far removed from traditional perspectives of scientific inquiry [2]. Such scientific perspectives are necessarily centered on explaining phenomena using theoretical models that can themselves be understood, at least in part and/or in principle, by people [3, 4]. Against this, data-centric models specifically eschew the restrictions of theoretical scaffolding in order to unlock greater expressive modeling capability, and thus are able through sheer volume of data and computing to encode extremely complex patterns in data with astounding accuracy [5, 6, 7]. Despite the profound difference in perspective, the empirical success of modern machine learning has nevertheless generated a great deal of interest in the sciences towards finding a way to utilize these techniques [8, 9, 10].

This utilization, however, is a conceptually difficult task. While in some sense modern machine learning models have identical structure to existing and ubiquitous statistical models (learning a function with a set of parameters optimized to predict a dependent output variable given some independent input variables over some dataset), they differ more fundamentally in the context of what they actually produce once optimized. Existing statistical models are generally integrated within science precisely with regards to how the parameters map to explanatory models within the scientific domain. Interpretation of the statistical model is not merely a 'nice to have' or auxiliary function, but rather is the whole raison d'être of the endeavor and is fundamental to the properties of the models themselves.

On the other hand while there certainly has been a great deal of progress on machine learning interpretability, all of these methods of interpretation necessarily fall short in many contexts of precise scientific interpretation. This is the case because, if there did exist a precise (without loss of information) transformation between the parameters of a neural network and the parameters of a scientific model; then a truly data-centric model such as a large neural network would simply not be learning anything beyond the model and thus would perform no better than the model on empirical predictive accuracy. In situations where this is the case clearly there is nothing to be gained by using the neural network. In fact what does very often occur is that a neural network will perform *better* than the scientific model precisely *because* there is no such direct mapping; being able to more flexibly encode all (or at least more) of the regularity present in the data without external restriction of being limited to encoding understood scientific phenomena.

Thus we have both the fundamental problem, and at the same time the fundamental opportunity, present in modern machine learning models: by removing the restrictions of theory, the model is able to learn precisely that which scientists have yet to conceptualize, in other words it is powered by *discovery*[1]. The challenge then is in translating this discovery back into scientific frameworks in order to actually generate knowledge, since the standard for such knowledge requires a much greater degree of follow-up confirmatory work that can only be done if new phenomena are integrated into the scientific conceptual structure.

While there have been cases where direct application of machine learning methods in the sciences have proven effective[11], this can only occur in the select few cases where interpretation of models is not the goal, but rather utilization of pure predictive power is looked for, for instance in cases of medical diagnosis [12]. Thus while work on interpretation has proliferated, it has broadly been in the context of helping to design tools that allow these predictive models to be more trustworthy in the hands of end users [13, 14,

[1] As will be elaborated throughout this work, in practice it is not so simple to distinguish between the non-conceptualized features of data that represent a scientifically impactful discovery from the non-conceptualized features of data that represent the inevitable idiosyncrasies of any particular dataset that are not relevant to the broader scientific enterprise.

15, 16], and thus the threshold of rigor regarding what constitutes an interpretation is different. Such methods thus, while very useful in these limited contexts, are insufficient for achieving genuine scientific discoveries in their own rights.

Nonetheless, when considering the utilization of AI in scientific software tools, talk of design has heretofore been largely limited to the design of the *interface* between model and user to improve interpretability and trust in some statically determined model [17, 18, 19, 20]. This approach, while meaningful in the context of improving the application of existing ML tools, is limited. For the creation of new kinds of tools where existing machine learning model formulations are more inextricably misaligned with scientific user priorities, a more radical conceptual reworking is needed, where situating models in terms of how their outputs can fit within the ontology of scientific users and their dynamic understanding of the underlying domain of study is presented as the core problem formulation. This thesis, thus, works towards the centering of people in the process of the design of *machine learning models themselves*; a synthesis of machine learning theoretic and human-computer interaction methodological frameworks which aims to enable a transformation of data analytic methods into genuine tools of discovery for actual scientists.

1.2 Thesis Overview

The large scale structure of this thesis, outlined in Figure 1.1, is divided into two interrelated thrusts: **Part 1: Human-Centered Discovery Frameworks** where, grounded in the theories of participatory design and methodologies of HCI, I develop new frameworks for understanding the human processes of scientific discovery in the context of how ML systems interact with these processes, and create guidelines for improving the design of such systems; and **Part 2: Interpretable ML for Exploratory Science** in which I collaborate with multidisciplinary scientific teams across multiple domains and develop novel interpretable ML methods that address the particular problems of scientific users in the process of exploratory data analysis. These two thrusts build on each-other, where applied work

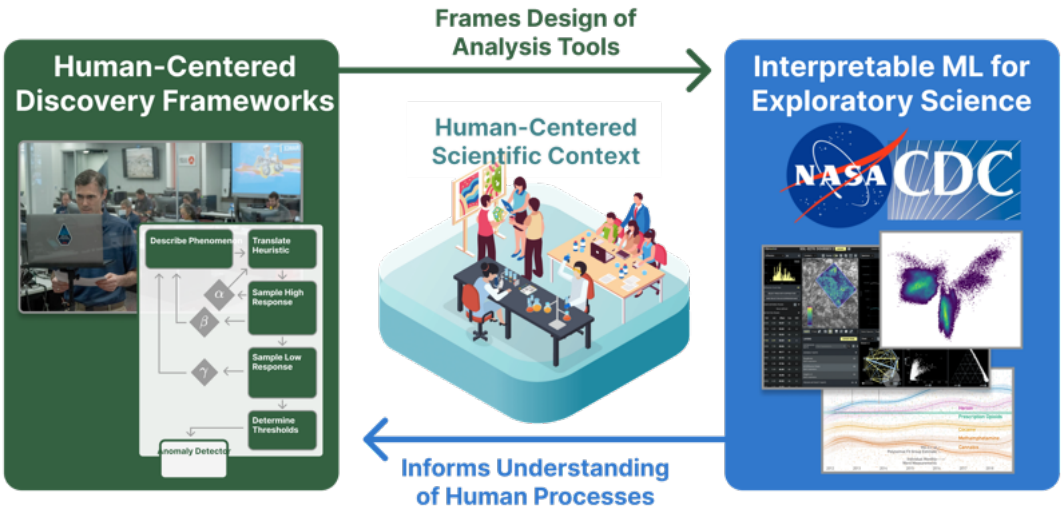

Figure 1.1: Overview of the dialectical structure of my thesis work. Applied work in embedded collaboration develops new ML methods that solve scientists' problems, and this collaboration forms the empirical basis that informs new abstracted design frameworks and principles which in turn are utilized to frame development of new machine learning methods. This structure takes very seriously both the low-level and concrete problems of specific scientists and domains while simultaneously working to generalize and expand these insights into a wide range of other new concrete contexts.

in Part II forms the empirical basis for grounding the design of frameworks in Part I in concrete contexts, while the frameworks developed in Part I form scaffolding to structure the applied collaboration and design of tools in Part II. Thus while in reality these aspects of my work cannot be completely separated and were completed contemporaneously, for the purposes of this thesis I will start by introducing the new theoretical design frameworks in Part I followed by their application in Part II.

Additionally, while I am the principal author of all of the research included in this thesis, the research is the result of years of collaboration with my PhD advisor, Duen Horng (Polo) Chau, as well as many mentors and colleagues at Georgia Institute of Technology, NASA Jet Propulsion Lab, and the Centers for Disease Control. To reflect my collaborators' contributions, I will use the first-person plural through the relevant thesis chapters. The research in this thesis that has been published previously is listed in Table 1.1.

Part 1: Human-Centered Discovery Frameworks A. P. Wright et al., "A comparative analysis of industry human-ai interaction guidelines," arXiv preprint arXiv:2010.11761, 2020 [21] A. P. Wright et al., "Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization," Proceedings of the ACM on Human-Computer Interaction, vol. 5, no. CSCW1, pp. 1–26, 2021. [22] A. P. Wright, P. Nemere, A. Galvin, D. H. Chau, and S. Davidoff, "Lessons from the development of an anomaly detection interface on the mars perseverance rover using the ISHMAP framework," in Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 91–105. [23]
Part 2: Interpretable ML for Exploratory Science A. P. Wright, C. M. Jones, D. H. Chau, R. M. Gladden, and S. A. Sumner, "Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media," Journal of Biomedical Informatics, vol. 119,p. 103 824, 2021. [24] A.P. Wright, S. Davidoff, and D. H. Chau. "Nested Fusion: A Method for Learning High Resolution Latent Structure of Multi-Scale Measurement Data on Mars." Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024. [25]

Table 1.1: Outline of thesis parts and the existing work that they are based on. Some of this work has been completed and published in leading HCI and AI peer reviewed venues. The thesis sections based on prior work will be expanded and elaborated upon within the context of their contribution towards my larger thesis.

1.3 Overview of Part 1: Human-Centered Discovery Frameworks

In Part I, I will address the primary research questions of:

- **RQ1**: What kinds of processes of data analysis and engagement with machine learning models do different kinds of scientific teams partake in, and how do models interact and integrate with domain conceptualizations through these processes?
- **RQ2**: How can integrating an understanding of these processes guide the design of models that better align with the concrete needs of specific scientists?

In order to answer these question I develop the concept of *Discovery Frameworks*, which model a class of scientific data analysis workflows to outline and can guide the design process of visual analytics and machine learning systems to ensure *interpetability* and *applicability* for specific scientific users engaged in similar workflows.

Based on the application of HCI and Human-Centered Design principles and methodologies towards **RQ1**, discovery frameworks provide the conceptual structure through which we can convert these results towards **RQ2**. While the concept of discovery frameworks enables a commitment to extremely concrete analysis of specific scientific users in a particular context of analysis; rather than forming a generalization based on domain-ontological considerations (such as a kind of method useful in one medical subfield being applied in a different but related medical subfield), discovery frameworks form structures that can be generalized based on *workflow*. Thus, while a discovery framework may not generalize even within a specific scientific discipline, if a workflow in in entirely unrelated discipline shares the same key characteristics defined in the discovery framework it can be applicable. This understanding of the importance of studying the workflow mechanics of highly specific applied problems and generalizing in this more 'horizontal' fashion is an essential component of the thesis of this work.

The first steps towards answering these questions and in developing discovery frameworks must be found in synthesizing existing design guidelines for human-centered AI

more generally, in order to form a baseline of conceptual considerations to integrate with later specific user research. In Chapter 3 I present such an integrative survey covering the design principles used in practice within industry [21]. The guidelines that are used in practice are chosen and emphasized above the large number of guidelines being published in other venues in order to maintain the focus of this work on grounding our design principles in the concrete needs of people actually interacting with models regularly.

Next, in order to develop these guidelines towards the ultimate goal of designing models for scientific end users, we must also study how existing models integrate with these guidelines and how resultant systems affect users' internal understanding of the topics being modeled. I do this in Chapter 4, where I develop the language classification visualization system RECAST [22]. RECAST takes existing language classification models and allows users to directly explicate and modify language in line with the classification. In this study we find how direct application of standard machine learning methods using human-centered principles undermines the initial model's effectiveness; as users, when understanding the heuristics of the model, very quickly are able to discover the ways in which the model is not indexed on the true phenomena of interest underlying the classification, but rather its empirical correlates, which quickly cease to be effective predictors when introducing interactivity. Thus, this study forms the fundamental insight which grounds all subsequent discovery frameworks: *that the entire process of the development of models must take into account how models and their problem formalization schema interact with the conceptual schemata of the actual people they are used by* ***in an interrogative context***. Thus in order for models to provide useful results that hold up to interrogative scrutiny, they must integrate at every stage of their development the interactive conceptualization of users and their determination of phenomena, and must be situated such that their outputs enable continuous revision and reinterpretation rather than imposing static classifications and predictions. From this insight we can then turn to specific scientific users and the developing of a novel discovery framework that enables alignment with their understanding.

Our human-centered approach is fundamentally grounded in addressing the actual analytic needs of scientists working in practice, developing our framework based on specific kinds of analytic workflows. Thus, it is vital to focus primarily on a depth of integration with real science teams rather than on breadth of applicability through a-priori abstraction to include as many domains as possible. In fact, by more concretely focusing on the unique needs of specific scientists we are able to develop methods that generalize *more than* methods that over-abstract problems and miss the needs of many actual users. In Chapter 5 I present the result of extensive embedded user research within the NASA JPL Mars Perseverance Rover PIXL Science Team, a world leading scientific collaboration between many disciplines including astrobiology, geology, geochemistry, planetary science, and physics, among many others [23]. By collaboration among such a diverse team of scientists, working on the many related problems of the search for extraterrestrial life and the geologic history of Mars, we can learn about a wide array of concrete problems that can form the baseline of a practical framework for human centered machine learning that may be applicable well outside of any of the individual scientists' specific domain. Through this collaboration I present the novel design and discovery framework of Iterative Semantic Heuristic Modeling of Anomalous Phenomena (ISHMAP) which forms the first major step in creating a technique for situating machine learning models within scientific workflows.

1.4 Overview of Part 2: Interpretable ML for Exploratory Science

After introducing our human-centered design frameworks, a key aspect of the interplay between ML and HCI must be addressed by actualizing and evaluating discovery frameworks through their application in developing novel machine learning models and tools both in their original as well as additional contexts. In Part II we do just that, by applying our discovery frameworks towards the re-framing of classic machine learning problems in new terms and with new problem formulations and show how resultant models are more able to address the end goals of actual scientists than direct application of existing state of the art

machine learning models, answering the research question:

> **RQ3**: Do machine learning models and tools developed using human-centered design frameworks more effectively solve scientific problems and enable scientific interpretation, compared to existing machine learning methods?

I address this with two examples of machine learning methods developed through our collaborative design frameworks with scientists. In Chapter 6 in further collaboration with NASA JPL PIXL scientists I develop a new model, Nested Fusion, for dimensionality reduction and multi-scale latent visualization on scientific datasets. Utilizing and building off the discovery frameworks developed in Part I, Nested Fusion utilizes a Bayesian modeling approach to learn an approximate latent distribution under the assumption of a singular high-resolution latent structure that is observed through arbitrarily many different nested measurement modalities at different scales. Nested Fusion outperforms existing state of the art dimensionality reduction and representation learning methods on standard unsupervised learning benchmarks, having the highest reconstruction fidelity across all visualizable latent dimensionality when compared to standard VAE based models and even outperforming much less scalable nonparametric methods like UMAP. More importantly, Nested Fusion learns a better structure that is visualizable, interpretable, and better corresponds to scientists' understanding of mineral structure, allowing Nested Fusion to integrate into scientific workflows structured and defined in Part I, and showing how such integration concretely enabling scientific discovery.

In Chapter 7, I showcase how the methods developed thus far can extend into other scientific domains. I present work in collaboration with epidemiologists and social science researchers at the Centers for Disease Control and Prevention introducing a new method to quantify linguistic shifts in online communities based on social media posts and comments [24]. In this work I show how diachronic word embeddings, trained on specifically chosen online sub-communities, can be used to calculate linguistic shift of concepts over time within these communities. I introduce a novel metric of the Relative Similarity Ratio (RSR)

to calculate such shifts over time, taking into account global changes in the community word usage distribution, and importantly frame the problem as that of a 'scientific query' that is dynamic based on the particular context that a scientist might be interested in. We show how, in this discovery framework re-contextualization, surprisingly small word2vec embeddings are able to discover reliable trends over time and validate the results of the model through historical post-hoc comparison to known trends in drug overdoses between 2010 and 2020. This work provides a strong exemplar of the orthogonality of standard machine learning metrics and scientific goals for interpretation; where a comparatively simple statistical model can enable better scientific insights because of the analytic framing engendered through human-centered discovery frameworks.

1.5 Thesis Statement

These parts when taken together, and in answering the research questions so far posed, my thesis shows that:

> A human-centered approach to the design and application of machine learning towards the concrete problems of working scientists provides a structured method for transforming regularity in complex data into patterns of *perceptible scientific phenomena*, enabling the creation of novel tools that can become powerful engines for discovery; tools situated within scientific workflows, that hold up to scientific scrutiny, and enable scientists to better perceive the structures underlying their data in the terms they can understand.

1.6 Research Contributions and Impact

- In collaboration with NASA JPL, I introduced Nested Fusion [25], a new method for latent visualization of complex multi-scale data, learning representations at the highest possible resolution that are much more scientifically interpretable than existing approaches. This work was recognized as **Best Paper Runner-Up (Applied Data Science Track) at the world's top Data Mining Conference ACM KDD 2024**.

- With Public Health Researchers at the CDC, I pioneered a novel method for interactive extraction of temporal semantic trends from diachronic word embeddings, enabling detection of trends in the opioid epidemic **over a year earlier** than traditional methods [24] which has been recognized with the **2022 CDC Excellence in Quantitative Sciences Award.**

- I spear-headed a **first-of-its-kind design framework**, ISHMAP, which introduces a radical new approach for human-centered model development [23], work which was awarded the **2023 NASA Space Act Board Award**. This work was **used by NASA scientists** as an essential component of the scientific discoveries published in the world leading journal, Science [26, 27] and covered by **over two dozen global news outlets** [28].

- I introduced the **first comparative survey and taxonomy of Human-Centered AI design guidelines used in practice** that forms a fundamental structure for studying and designing AI systems [21] which has been used for teaching Machine Learning Interaction at the University of Washington, a top computer science gradate program [29].

CHAPTER 2
BACKGROUND AND RELATED WORK

2.1 Conceptual Foundations of Scientific and Data-Analytic Methods

This thesis aims to understand how to utilize machine learning to develop practically useful tools for scientists. Naturally then, the first question that must be addressed in this endeavor is: "what do we even mean specifically by science and scientists?". In this section I will provide a brief introduction to some of the major scholarship working to understand and delineate science, and thus provide the background to allow us to conceptualize the key design considerations when looking to apply machine learning in scientific domains.

This background, while being strictly well outside of the domains of computer science, HCI, and ML, is all the more essential to contextualize this work specifically because it is frequently ignored. In so far as previous work has attempted to generalize concepts of 'AI for Science' (AI4Science) beyond the particularity of individual domain applications, much of it is based on fairly simple generalizations of what constitutes 'the scientific method' [11]. For instance, prominent recent work purporting to develop an 'AI Scientist' [30] that utilizes LLM's for a 'full-stack' approach from reading papers to writing papers does not engage with or implement any structure whatsoever regarding experimentation and methodology and unsurprisingly as a result produces output that is frequently entirel

However there exists a vast literature of scholarship which undermines many of the assumptions and structures used in such generalizations. This literature has gone on to further develop or criticize aspects of the naive model, with particular focus on what *actual scientists do* to determine what *science is*. Drawing primarily from the work of contemporary philosophy of science [3, 31] this section will start by providing a brief background of some of the most important conceptual schemes relevant for understanding more pre-

cisely the variety of activities constitutive of science, and conclude by outlining a working understanding of how science will be conceptualized and delineated for the purpose of this work.[1]

Epistemology and the Scientific Method The first crucial conceptual background for delineating precisely the activities this thesis aims to understand and assist under the category of 'Science', is to attempt to differentiate science from epistemology – the process of gaining knowledge – more generally. Despite ideas similar to those discussed in this section dating back as far as Aristotle [31], the most appropriate starting point for a model of what we today call 'The Scientific Method' is Francis Bacon and his *Novum Organum* (1620) [32]. Bacon's method formed a specific normative outline of how to conduct inquiry of the natural world and form reliable inductive inferences from observation while avoiding certain kinds of logical errors. Simplified, in the Baconian method one systematically generates observations of a phenomenon, as well as similar observations where a phenomenon is not present, and aims to find the common structures underlying these observations that can explain the phenomena observed. While Bacon's model in reality was ultimately too restrictive to be carried out generally in practice, its structure forms the starting point from which later models of methodology would develop.

Verificationism and Falsificationism After attributing at least the general ethos of systematic observation as perhaps the primary source of 'The Scientific Method', no other philosopher is more often considered (at least among working scientists if not working philosophers) when discussing norms of science than Karl Popper and his doctrine of Falsi-

[1]This extremely brief section, when considering the extraordinary breadth of the topic at hand, could far outstrip my actual dissertation in length and detail, as the very notion of unifying 'Science' in any neat form has not only been a project extending back centuries historically, but is a project fundamentally complicated by the huge variety of activities called science across time, place, discipline, and even among individual scientists. Thus, while I aim to provide a more engaged overview of many of the most important ideas in this tradition for an audience of computer scientists, at least when compared to much of the other work under the banner of 'AI4Science', it is important to note the extremely limited scope of the goals of this engagement to helping contextualize my original research and what is required to understand how it differs from other approaches specifically in contemporary scientific machine learning.

fication [33]. Popper's theory, outlined first in *The Logic of Scientific Discovery* (1959) [34], aimed to provide a better way to demarcate between science and pseudoscience. Unlike the inductive models following in the Baconian and Empiricist tradition, Popper proposed a hypethetico-deductive (HD) model where observation and experimentation cannot provide positive evidence of a theory, but rather can only show that a theory is false if the theory and observation disagree. Thus Popper's criterion for a theory to be considered scientific rather than pseudoscience is that it must, in principle, logically be able to be falsified by some conceivable observation. That is, a theory is scientific if and only if along with the theory there could exist some way to test if the theory is wrong. It is important to emphasize (as it is often lost in the popularizations of falsificationism) that this is an extremely radical framework for understanding science, under which no amount of observation consistent with a theory ever provides any evidence *for* a theory, instead theories are only ever are falsified. What this implies is that, for Popper, science is characterized by a 'permanent openness, a permanent and all- encompassing critical stance, even with respect to the fundamental ideas in a field'[3].

Despite its influence, Popper's position has faced much criticism, such as the fact that the falsification criterion is itself a non-falsifiable claim, as well as the fact that it simply does not align particularly well with how scientists in practice actually evaluate evidence. However, despite these criticisms Popper's model has become extremely influential, and softened versions of the falsification criterion are utilized ubiquitously in scientific disciplines as a barometer of valid theory formation [33].

Scientific Revolutions and Incommensurablity Both Baconian inductive empiricism and Popperian falsificationism provide highly influential normative methodologies meant to influence how science 'ought' to operate. However these models, despite their large differences, have in common a certain kind of abstraction of the process of science away from how it is carried out historically in practice. This trend in the study of science dramatically

shifted after the publication of Thomas Kuhn's *The Structure of Scientific Revolutions* in 1962 [35]. While sometimes the radicality of this work has been overstated, it certainly still does undermine at least some of the presuppositions of earlier logical empiricist work [3], and it undeniably has had a profound effect on all later work in the conceptual study of science.

For our purposes there are two key aspects of this work (and the works of contemporaries in the same tradition) that are important to consider, the model of 'Normal Science' operating under a paradigm, and the incommensurability of different paradigms emblematic of scientific revolution. In Kuhn's model the overwhelming majority of scientific work primarily consists in scientists performing what he terms 'puzzle solving exercises' of refining the details of a scientific paradigm[2]. Here a paradigm essentially refers to some exemplary work or set of works that outline the standard methodology, worldview, and delineation of the specific project the scientist doing 'Normal Science' is working on. An essential implication of this is that observations under a paradigm are 'theory-laden' [3], where the presuppositions of the paradigm determine 'what the scientist can see' in so far as it determines valid selection of phenomena to consider. Theory-ladeness substantially weakens traditional inductive empiricist conceptions under which observation comes first and can form a neutral base from which to form theories, since if the very observations themselves contain and encode theoretical content the resultant knowledge can only be conditional.

Similarly, and in important contradistinction to Popper, Kuhn finds that in order for Normal Science to operate smoothly, the core aspects of the paradigm being worked within are not regularly questioned. Only when there is a sufficient accumulation of anomalies – being the observations that are difficult or impossible to reconcile within a paradigm – does a 'revolution' occur where some alternative paradigm supersedes the old one and normal

[2]It is important to add a note of caution here, as the term 'paradigm' might be one of the most often used, and often misconstrued, concepts in the history of the philosophy of science. So much so that Kuhn himself in much of his later work specifically aimed at rebutting many of the popular interpretations of his work and in particular misuse of the term 'paradigm'.

science can resume under the new paradigm.

The other key aspect of Kuhn's theory is that the revolutionary process of paradigm shift operates categorically differently than normal science. While for other models of science substantial and revolutionary change in theories occurs by the same mechanisms as the rest of science, for Kuhn the choice *between* different paradigms is radically different than normal science which operates *within* a paradigm. It is in this respect that different paradigms are said to be *incommensurable*, where there is no common measure between them and scientists operating on different sides of the divide can meaningfully said to be "living in different worlds" [35]. Alongside Kuhn, contemporaries such as Paul Feyerabend showed how, when viewing science historically, incommensurability precludes the possibility for any consistent method to reliably choose between paradigms, and that thus the whole history of science in practice fundamentally relies upon at least a few key 'non-scientific-methodological' choices [36, 37].

While, much like Popper, the most radical interpretations of Kuhn and his contemporaries have largely been rejected, his initiation of the 'historical turn' and questioning of the normative relevance of logical empiricism on science *as it actually occurs*, is profoundly influential and important if we wish to develop tools and methods that are applicable to concrete and thus historically and paradigmatically situated scientists.

Sociological Approaches Following Kuhn and the general turn away from studying science as purely an abstract logical and epistemological process but rather as more of a historically contingent and socially situated activity, much of the scholarship on conceptualizing science up to today has been led by the sociology of science[38]. The most prominent work in this tradition, such as the work by Bruno Latour [39, 40] and Steve Woolgar [41, 4] have emphasized the simple fact that scientific practice is a *collaborative human process*, and thus studying the particular ways that people work with each other, including all of the many biases and social dynamics involved, is absolutely indispensable to understanding

science as an activity. Work in this tradition has thus methodologically expanded 'beyond the armchair' to show that embedded research among actual working scientists and tracing their interactions can develop much better models of the complex and idiosyncratic activities of how science operates concretely [42].

Science in the Era of Big Data A great deal of the most recent work in outlining scientific method has specifically focused on the impact of rapidly advancing machine learning and data science[43]. As briefly discussed in the introduction, the data-centric style of contemporary machine learning modeling is radically different than any existing theory-based scientific method [1]. Some scholars have posited that the empirical success of data-centric modeling makes hypothesis based scientific methods "obsolete" [2], or at the very least poses challenges as a viable alternative in some fields[44]. Others have emphasized the similarities between data-centric approaches and more classical Baconian inductivist approaches, where the ability for machine learning models to effectively model data with essentially no theoretical background knowledge[3] addresses many of the previous critiques of inductive scientific reasoning based on the theory-ladeness of observation [45].

However some are less optimistic, seeing many of the promises of big data be less scientifically productive in practice than advertised [10], criticizing a structure of 'fishing expeditions' to find spurious correlations [46], that claims of atheoretical neutral data are false since the structures of data collection and processing are frequently predicated on particular unexamined and unsophisticated 'classificatory theory' [1], and highlighting the challenges presented when the insights provided by large and distributed models are so far outside of how human cognition works to be considered 'alien' and thus potentially impossible to integrate into the traditional human-conceptualization and interpretable body of scientific knowledge [47].

Nevertheless, as data-centric methods have become increasingly common, their success

[3]Consider the famous aphorism about developing natural language processing that "every time I fire a linguist my performance goes up" [5]

is ”strongly dependent on participants' ability to critically assess the embodied and propositional knowledge involved in data journeys”, further emphasizing the importance of the sociological approach ”focusing more on the processes through which research is carried out than on its ultimate outcomes” [1]. Most vitally, when tracing the concrete relationships of data, evidence, discovery, and theory formation, we find an essential conceptual distinction between *data* and *phenomena*. Most famously outlined by James Bogen and James Woodward [48], data being the raw measurements of experiments and observations do not directly relate to eh objects referred to in classical theories of scientific evidence. Such theories are built upon more abstracted, theory dependent, and relational objects that delineate the 'true' (as determined within the scientific paradigmatic context) phenomena that structure the more contingent data and are the objects about which scientific theories are actually constructed. There is a wide spectrum of views relating to how directly connected data are to phenomena, from strong 'representational views' that sees data as direct and reliable representations of real phenomena, all the way to strongly 'relational views' which posit that the transformations from data to phenomena are negotiated in a socially and theoretically dependent way that can vary throughout the process of inquiry [43, 49]. Understanding this process of transformation from data to phenomena, and how machine learning and visual analytics software affects the process is a central focus of this thesis.

Upon understanding the distinction between data and phenomena among the objects that are 'out-there', there similarly introduces a dual distinction between the objects 'in-the-mind' that are used to understand them. Thus precisely while in the traditional machine learning lexicon we have the relationship of data to models, there is the parallel relationship of phenomena to theories[4]. While models are able to learn precise and computable structures characterized by an ability to make *predictions about data*, theories in this sense refer to the set of human mental structures that define domain knowledge characterized

[4]When compared to the terms for the distinction between data and phenomena, the terms model and theory have much more baggage and are used in a wider variety of senses both in computer science as well as in the philosophy of science. Thus is is important to emphasize the potentially idiosyncratic use of the terms here, and that for the purposes of this thesis theory refers specifically to this more restricted concept

by an ability to *explain phenomena*. This distinction has been described as the difference between 'knowing-how' (i.e. a model being able to make predictions) and 'knowing-that' (i.e. a theory being able to explain what is occurring) [50].

As shown in Figure 2.1, one defining factor of the data-centric framework is how phenomena are structural excluded in favor of pure data modeling. How then can data-centric science be actually doing any science, in the sense of producing knowledge and communicating results between people rather than just sharing model weights and code, which requires theory which itself as discussed previously requires phenomena? The answer is that theoretical knowledge can only be formed through the mediation of the interpretation of the model, which turns the model into a unique kind of singular phenomena to be explained. However, because frequently the design of the model is so far removed from theoretical structures (as opposed to traditional phenomena which develop more naturally in conjunction with theory), this interpretation and explanation mediation can be extremely difficult and error prone; requiring the development of many different kinds of specific machine learning interpretability and explanation systems which are the focus of Section 2.4.

2.2 Summarizing Frameworks of Scientific Inquiry

In the previous section I have briefly outlined the key conceptual frameworks of scientific method and practice that the rest of this thesis utilizes. Figure Figure 2.1 illustrates these structures as well as the more generalized *Human-Centered Framework* which builds upon the key conceptual distinctions developed in the other main traditional frameworks and extends their outline into how they are integrated into a paradigmatically contextualized social process that can vary substantially from practice to practice resembling one or more of the other frameworks dynamically, as "attempting to draw a sharp distinction between hypothesis-driven and data-intensive science is misleading; these modes of research are not in fact orthogonal and often intertwine in actual scientific practice." [46]

In the human-centered framework rather than starting with phenomena, theory, or data,

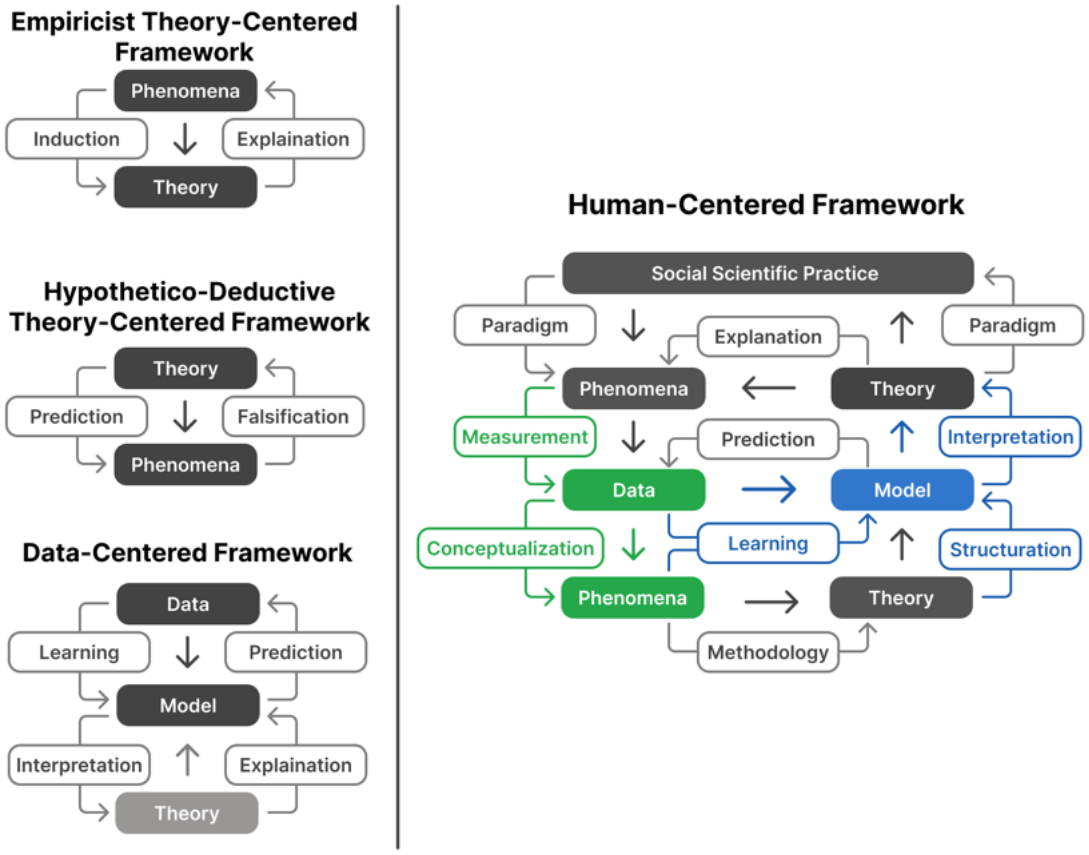

Figure 2.1: Illustration of different frameworks of scientific inquiry, where the general flow of reasoning is outlined with annotation of the modes of mediation between conceptual categories that different frameworks emphasize. Left (from top to bottom): Baconian Inductive Empiricist framework which starts with phenomena and works to develop theories which explain them, the Popperian Hyothetico-Deductive Framework which instead starts with hypothetical theories that can produce refutable predictions about phenomena that can be falsified, and the Data-Centric Framework which starts with raw data and learns models that predict patterns in data and can thus only create theory through axillary functioning of interpreting the model without access to domain phenomena. Right: The Human-Centered Framework built upon in this thesis which provides a flexible scaffolding within which the large variety of highly distinct scientific practices can be mapped within. Green highlighted categories form the concepts of emphasis in Part I, while the blue highlighted categories mark the concepts of emphasis in Part II.

we start with the context of the specific social scientific practice. Through paradigmatic norms this produces the space of possible phenomena[35], from which data is created by a 'data-journey' [1] which consists of the variable processes of measurement – either experimental or observational – which must be considered before simply 'taking the dat and running'. Data can then be directly modeled in a data-centric fashion or may be further conceptualized into phenomena directly to be applicable in theory formation through either inductive or deductive methodologies. Theory then may structure choices relating to data-modeling, and the resultant models must then be interpreted back into theory in order to provide explanations as well as to integrate fully into the scientific paradigm. The plurality of styles of reasoning compatible with this framework allow it to be used to contextualize more precisely the myriad ways in which different aspects of machine learning, human-computer interaction, and design can be integrated into scientific workflows.

2.3 Contemporary Applications of Machine Learning in the Sciences

The extremely fast pace of development in AI and ML, in particular in natural language processing and LLMs have garnered a large amount of interest in applications of these technologies in science. For instance, prominent recent work purporting to develop an 'AI Scientist' [30] that utilizes LLM's for a 'full-stack' approach from reading papers to writing papers. However this work, in general, engage with or implement any epistemological structure whatsoever regarding experimentation and methodology and unsurprisingly as a result produces output that is frequently entirely disconnected from reality. Therefore more specified tools that engage with specific domain problems enabling integration with structures and norms of theory formation and evaluation are more promising, such as utilizing methods from unsupervised and generative models to inverse problems with strong inductive biases[11].

These applications of machine learning systems in the sciences can be broadly separated into two very broad kinds: *Exploratory Systems* and *Predictive Systems*. Exploratory

systems aim to assist scientists in exploring complex datasets broadly through various unstructured visualization and data wrangling processes. Predictive systems take more statically formed problems such as diagnostics [51], material synthesis[52], protein synthesis [53], or simulations[54] from consistent data types and trains a classification or regression model to assist scientists in this particular task.

Structurally, predictive systems are pure data-centric models in the framework introduced in Figure 2.1 by providing predictive power over regularities in data *directly* without requiring recourse to existing delineations of phenomena (such as with feature extraction) or existing theory. However, for such a system to contribute towards any kind of larger scientific enterprise interpreting the model is essential, and such interpretations must be able to produce explanations precisely of the same scrutiny as traditional scientific theory. This is a high bar, which is sometimes possible to meet when data-journeys are relatively transparent with respect to domain phenomena, but much more difficult if there is any complication in data-journeys relative to relevant phenomena. Nonetheless, there exist various methods for visual analytics for deep learning interpretability that have been developed [18]. These methods can be model architecture dependent, such as as SANVis [55] and exBert [56] which allow for interactive exploration of the attention mechanisms in the very commonly used Transformer models; or model independent such as LIME[57] and SHAP[58]. Furthermore, dynamic visualizations that allow for user experimentation can assist in the understanding of a machine learning model [59]. Some tools take a more active approach to explain models through counterfactuals [60]. Others, like Errudite [61], allow users to test their own hypotheses with respect to the true error distribution on the entire dataset. Others use a human-in-the-loop design [62].

Exploratory systems on the other hand do not generally aim to replace a scientific model, but rather aim to assist scientists in *discovering the phenomena that require explanation*, and thus do not require the same kinds of explanation in and of themselves, as their role is more limited within the scientific structure to the task of conceptualization in

Figure 2.1, deferring the scrutiny of explanation to better suited domain models. Among exploratory systems, dimensionality reduction techniques such as UMAP [63], T-SNE [64], MDS [65], Isomap [66], and the most commonly used PCA [67] are fairly ubiquitous in a variety of scientific domains [68, 69, 70, 71, 72, 73, 74]. Another technique that can produce comparable visualizations is the approach of latent analysis which takes a more probabilistic, generally Bayesian, approach to the problem of learning low dimensional representations. These approaches mostly stem from the development of variational autoencoders (VAE) [75, 76, 77], and different latent models have been introduced to handle many scientific problems [78, 79, 80] including planetary science[81] among many other domains.

2.4 HCI and Tools for Interpretable Machine Learning

As many of the explainable machine learning systems have found, dynamic and integrated systems with humans in the loop is an extremely powerful technique for effective utilization by end users where trust is essential such as in science. Thus, while having more independent origins, the recent history of hybrid disciplines such as HCI-AI[82], HCML [83], and human-guided ML [84] reflect an interest in drawing on knowledge generated across fields to jointly inform the development of systems that touch on each discipline. HCI researchers, for example have looked to understand the challenges to design AI systems that fit with user needs[85], and to use new properties exposed by these systems as a resource for designers[86]. The complexity of dealing with some form of embedded intelligence led other researchers to introduce particular methods to structure ideation and iteration of AI- and ML-systems for ubicomp[87, 88] or dialog systems[89].

Alternatively, researchers in AI and ML have drawn upon expert knowledge to inform ML models [90], or to bring interactivity into learning systems [91], looking to leverage interaction to define ML model rules [92]. These methods have been applied to anomaly detection in cybersecurity, spatio-temporal, and behavior modeling contexts[93].

In the sciences, this crossover has looked at interactivity and glass-box models as a way to support interpretable and configurable deployed machine learning models [94]. However, fieldwork in disciplines such as oceanography have shown that while interpretation and understanding of code and models is essential, it is insufficient to contributing within a larger scientific workflow as the primary driver of change will often come from anomalous "moments of flux" which naturally lead to reconceptualizations that are not amenable to fixed data perspective implicit in any traditional data science or machine learning model, as "a singular focus on problem-solving may marginalize opportunities for innovation that could drive community engagement, and, therefore, momentum and adoption"[95]. Therefore this work looks to expand the tradition of utilization of participatory design practices in the context of AI approaches to science by enabling a more flexible modeling approach while maintaining established key aspects of interpretability.

Part I

Human-Centered Discovery Frameworks

Overview of Part I

In this part I will present the theoretical development of the design principles and methodologies which result in *discovery frameworks*. These are the human-centered design practices, based in the tradition of HCI user research, for the design and creation of tools of scientific discovery. As discussed in Chapter 2, the actual practices of scientists are wildly variable, and therefore in order for such a framework to be sensible it must be based on a practice of user research to outline the specific needs of specific scientists. Discovery frameworks provide structure to that process in scientific contexts, both in terms of considering the universal properties of designing for interpretable machine learning as well as providing the scaffolding for determining more precisely what aspects of the specifics of an embedded application are relevant. Without such structure such a framework would consist of nothing more than the imperative to: "listen to your collaborators". While that is a good starting point[5] the theoretical virtues of enabling more 'active listening' under some kind of structure of what to listen for and how to integrate the answers into designs is what discovery frameworks look to provide.

Towards this end, Part I consists of three chapters which are slightly edited versions of work completed through the course of my PhD and previously published in leading HCI peer reviewed venues[6] [21, 22, 23]. Chapter 3 and Chapter 4 present work describing relevant components of general purpose machine learning interpretation design. Chapter 5 presents both a case study of applying these design methods towards a specific scientific use case, and introduces the first discovery framework of ISHMAP which is a specific framework for structuring anomaly detection based workflows. Aspects of ISHMAP can then be generalized and all of the design principles of Part I applied towards Part II which focuses on how applying these principles leads to the development of more effective and interpretable scientific analysis tools.

[5] And indeed more than many applied machine learning work does currently

[6] As this is the case, the specific problems that each chapter addresses will be more limited in scope and not focus very much on how each chapter works together towards the general thrust of the thesis.

CHAPTER 3

COMPARATIVE ANALYSIS OF AI DESIGN GUIDELINES

3.1 Introduction

If we wish to understand how best to design AI systems with respect to particular scientific user considerations, we must first take into account the more general design principles of such systems as a starting point. To form this baseline of design considerations we take queues from the design guidelines for human-AI interaction published by the developers of the most widely used AI infused software products, including Apple [96], Google [97],and Microsoft [98]. While these guidelines from industry sources have a great deal in common, they differ in key aspects from their methodology, emphasis on different principles, dissemination venues, and target audiences.

This work differs from other surveys [99, 100, 101, 18] on explainability and trust for AI in two ways. First, this work focuses in particular on guidelines put forth by industry; which, while they are informed by academia, do not necessarily match exactly. Furthermore, the scope of these guidelines tends to be broader beyond ensuring explainability or trust in AI but also includes guidelines for all of the other aspects that AI systems differ from other software. Therefore these guidelines can provide a more practically useful overall tool for AI developers.

No existing work has yet synthesized the information from all of these sources, leading to potentially fragmented—or even competing—standards for AI developers and designers. In this chapter, we provide an overall analysis of all of the different guidelines being proposed by different companies. Our work makes the following key contributions:

1. This work provides a **comparative analysis** of design guidelines for human-AI interaction proposed by different companies. This includes a survey of the guidelines,

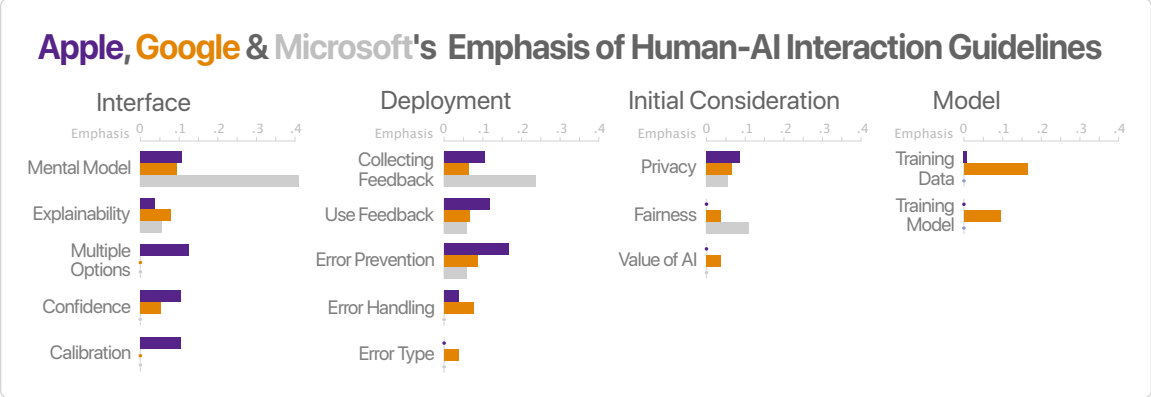

Figure 3.1: Relative emphasis of human-AI interaction guidelines by Apple, Google and Microsoft. A "dot" indicates no emphasis for a guideline subcategory. We see that the largest difference is that Google gave much more emphasis to model considerations for training data and processes, while Apple and Microsoft spent very little or no emphasis specifically on the model. Interface and Deployment categories dominated in roughly equal proportions at all three companies. When considering subcategories each company emphasized, the most notable is that Microsoft used nearly 40% of its emphasis on the specific category of mental models. We can also see that Apple seems to have comparatively more emphasis in categories relating to smooth user experiences such as Error Prevention, Calibration, Confidence, and Multiple Options.

comparison of each company's methodology for developing the guidelines (in so far as it is made public), and comparison to current academic work in human-computer interaction. With our comparative study, AI practitioners can more easily understand both which guidelines have a consensus by the overlap and common design recommendations proposed by different companies, and also multi-faceted views from the differences in the companies' approaches.

2. Also, this work introduces a larger **inclusive taxonomy** of how all of these guidelines fit together **in a unified guideline structure** for Human-AI Interaction. This unified structure is valuable even though each of the individual sets of guidelines has its own proprietary hierarchy, as it provides an external overarching structure against which each set of guidelines can be compared, allowing us to measure the differing emphasis of each company. Furthermore, our unified structure can serve as an extensible base for future development in Human-AI Interaction.

3.2 Survey of Guidelines

The three most recent publications from major technology companies of guidelines have been from Google, Apple, and Microsoft. Each of these released guideline systems has a substantially different structure, emphasis, and perspective on the same central question of how to build products that use AI in a human-centered manner. This section surveys each of these industry publications and examines the context in which they were produced.

3.2.1 Microsoft

The first set of guidelines is Microsoft's "Guidelines for Human-AI Interaction"[98], published at CHI 2019 in May of 2019. In that work, researchers from Microsoft surveyed over 168 potential guidelines originating from internal and external industry sources, public articles, and academic literature. These guidelines were combined and re-organized into a coherent set of 18 guidelines; all of which have a common style of "a rule of action, containing about 3-10 words and starting with a verb" [98]. Guidelines were then structured over when the guideline is relevant to the *user* over the course of their interactions with the product.

Finally, Microsoft conducted a user study with HCI practitioners to evaluate the applicability and clearness of the guidelines. This academic approach resulted in the fewest number of guidelines of the companies surveyed, however, it was the only set to outline the process explicitly on how the guidelines were developed and evaluated.

3.2.2 Google

At roughly the same time in May 2019, Google released their comprehensive set of AI interaction guidelines: Google's "People + AI Guidebook"[97]. This guidebook is based both on "data and insights from Google product teams and academic research"[97]. While it lacks the open experimental validation of a user study, it does contain extended refer-

ences to the academic literature. Instead of organizing the guidelines around the process for a *user*, Google breaks its content down into distinct concepts that a *developer* has to continuously keep in mind. These are: User Needs + Defining Success, Data Collection + Evaluation, Mental Models, Explainability + Trust, Feedback + Control, and Errors + Graceful Failure. Furthermore, the Google Guidebook takes a much longer form consisting of 113 individual guidelines, with each including more content and extensions when compared to the other publications.

3.2.3 Apple

Finally, in June of 2019 at WWDC'19, Apple announced its Human Interface Guidelines for Machine Learning[96]. These guidelines differed from the "bottom-up" approach of the academic literature collation and user study refinement present in other guidelines. Instead, this document is a primary source of "practitioner knowledge", foregoing references or data, and thus is seemingly based entirely upon standing design principles within the Apple organization; which helps provide a unique and different perspective from the other more academic style works. While this style may present potential issues as a standalone document, it may help result in a greater overall synthesis of knowledge when considering all three sets of guidelines together by bringing a diversity of perspectives[102]. The document is focused on specifying how Apple's design principles are applied in the case of machine learning infused products. Making it comparatively more focused on aspects of user interfaces rather than AI model functionality. The 59 guidelines in the document are broken up into two main themes, the *inputs* of a system and the *outputs* of a system. Within each of these categories, there are further subcategories. For inputs, the guidelines focus on Explicit Feedback, Implicit Feedback, Calibration, and Corrections. Guidelines in each of these sections aim to help design the processes by which AI products ask for, collect, use, and apply user data and interactions. The sections on outputs cover Mistakes, Multiple Options, Confidence, Attribution, and Limitations. These sections all contain

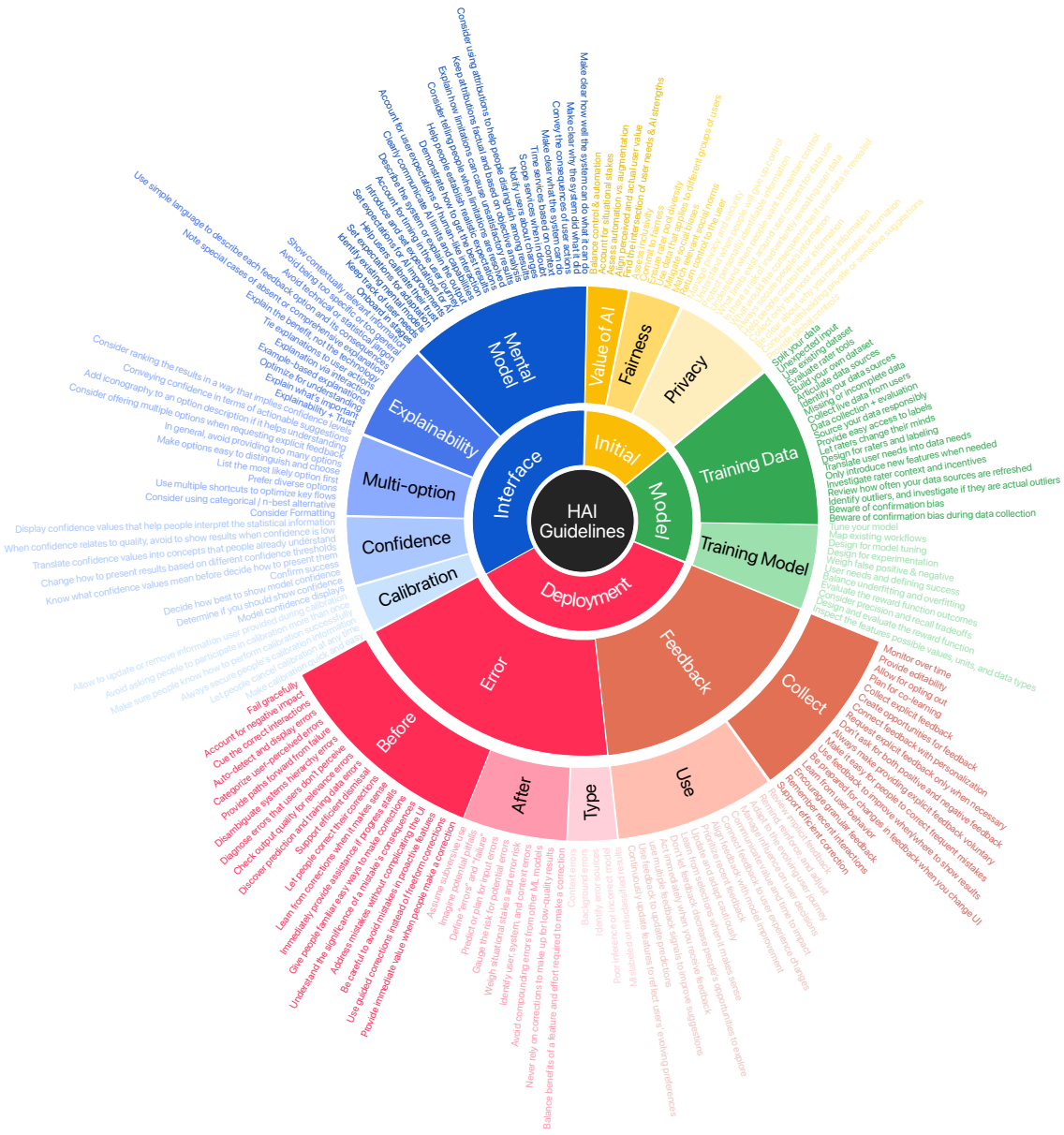

Figure 3.2: Unified Guideline Structure. The inner ring consists of the higher level categorizations, and further sub-categorizations developed during the affinity diagram process are shown concentrically. The outermost rays consist of the specific guidelines colored based on their categorizations. Further references on each guideline and its corresponding categorization and source document can be found in Appendix A.

Company	Categories	Subcategories	Guidelines
Microsoft	4	N/A	18
Google	6	20	113
Apple	2	9	59

Table 3.1: Total number of guidelines and categories from each company surveyed.

guidelines focused on taking the output of a model and displaying it to a user in a way that is understandable and actionable for the ultimate purpose of the product.

3.3 Unified Guideline Structure

While there are significant differences between the individual sets of guidelines, there is also substantial overlap. Furthermore, the huge amount of competing standards for desired AI systems can wear a developer down when they try to learn and adhere to many different guidelines. Developing a synthesis of all of the major AI guideline systems into a single comprehensive structure may make learning all of the important guidelines more straightforward, and future extensions and changes more possible. By fitting each company's guidelines within a larger consistent structure, the differences in emphasis between companies become readily apparent. This paves the way forward for the development of new guidelines and better AI-infused products.

In an affinity diagram process similar to that done by Microsoft [98], we separated all of the guidelines from all three sources, excluding guidelines that were not meaningful without the context of higher-level categorizations from their source document, resulting in 194 individual guideline statements. We then conducted a card sorting exercise to find similar groups of guidelines and sort them into an affinity diagram. This resulted in twelve distinct categories of guidelines. We then repeated the process among these categories to define four high-level categories. The resulting categories are outlined below and the full hierarchy of guidelines is shown in Figure 3.2.

3.3.1 Initial Considerations

These categories are generally relevant in the initial design phase of a system as things that must be considered before other development can proceed.

Value of AI The 5 guidelines within this category focus on understanding the value that AI may be able to bring to a product in addressing a user's need before going headfirst into development. This includes statements like "Find the intersection of user needs and AI strengths" and "Balance control and automation".

Privacy / Security These 13 guidelines focus on ensuring that user data is always secure and that users have the ability to control their own data. These guidelines are also broadly applicable to software more generally such as "Always secure people's information" and "Collect only the most essential information".

Fairness Issues of fairness in AI are an important and emerging topic in the study of human-AI interaction. However, there are comparatively few widely adopted techniques to help ensure fairness. The surveyed companies published 7 guidelines in this category, many of which are comparatively vague due to this underdeveloped area such as: "Mitigate social biases" and "Commit to fairness".

3.3.2 Model

These categories focus on the design of the machine learning model itself in the model design, data collection, and training procedure.

Training Process These 11 guidelines cover how best to handle training the model, avoiding certain kinds of errors or data issues such as "Consider precision and recall tradeoffs" and "Balance underfitting and overfitting", and ensuring the model is best optimized for the actual purpose of the product.

Training Data While some guidelines focused on the training process, these 20 guidelines focused specifically on training datasets. These guidelines differ as questions about the data often are more high level and involve different actions, such as "Review how often your data sources are refreshed" or even analyzing what data is being used to "Beware of confirmation bias".

3.3.3 Deployment

These categories focus on the deployment of trained models and how to handle continuous fine-tuning and errors.

Errors Given that AI systems fundamentally lack the kind of determinism and predictability of other software, being able to handle errors becomes especially important in this context. Therefore all three companies included extensive guidelines covering the handling of errors, resulting in a category of 26 guidelines. We broke down these guidelines into additional subcategories of considerations for design made before an error occurs, and systems for handling after an error occurs ("Support efficient dismissal" and "Learn from corrections when it makes sense"), as well as guidelines enumerating types of errors ("Mislabeled or misclassified results" vs "Background errors").

User Feedback / Personalization There are several guidelines for how to design systems that learn specific user preferences over time, and how to give users control over this process. These are internally divided into 17 guidelines about how to collect information ("Remember recent interactions" and "Allow for opting out"), and 15 guidelines on using that information ("Prioritize recent feedback" and "Don't let implicit feedback decrease people's opportunities to explore").

3.3.4 Interface

These categories focus on the design of user interactions and interfaces with AI systems.

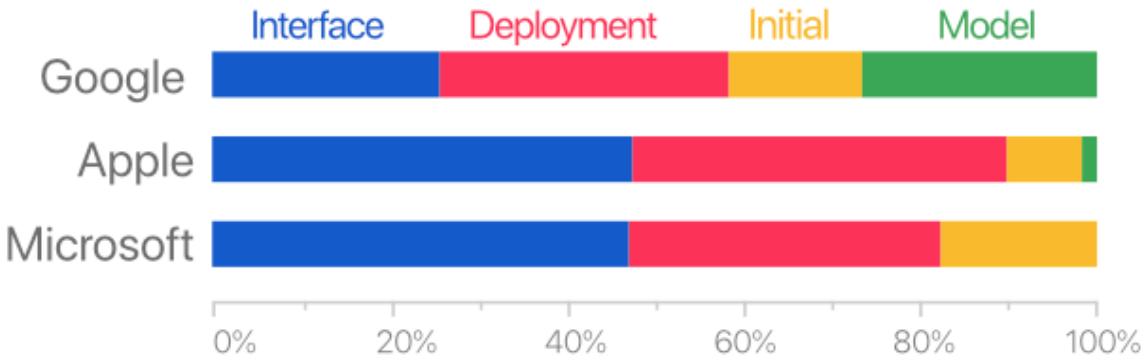

Figure 3.3: Relative emphasis of human-AI interaction guidelines by Apple, Google and Microsoft.

Expectations / Mental Models 24 guidelines covered how best to design systems with user expectations in mind. Building intuitive mental models of how the system works is a key point of emphasis for building user trust. This includes "Make clear why the system did what it did" and "Set expectations for adaptation".

Explainability This category of 12 guidelines focuses on explainability. This includes trying to make the results and process of models more transparent for users. Examples include "Show contextually relevant information" and "In general, avoid technical or statistical jargon".

Multiple Options A key component of designing interfaces for systems with a range of potentially incorrect outputs is giving users multiple options based on model outputs. Of the 11 guidelines in this category some focus on how to display options to the user as either "Categorical / N-Best Alternatives" as well as what set of options to show users who "Prefer diverse options".

Confidence A unique component of AI systems is the ability to make predictions of variable confidence. This is an important feature to expose to users the level of uncertainty in model output. The 9 guidelines covering the ways to handle uncertainty include "avoid

showing results when confidence is low" as well as "Decide how best to show model confidence as Categorical, N-best alternatives, or Numeric".

Calibration Apple included 6 guidelines about calibrating models, which no other company talked about. Calibration differs from model training as these guidelines consider the user experience of initially fine-tuning existing models for each specific user. These include guidelines like "Avoid asking people to participate in calibration more than once" and "Let people cancel calibration at any time".

3.4 Emphasis Differences

After developing a unified structure for guidelines, we can then look back at each company's set of guidelines and compare them within this new context. To see the difference in emphasis between different companies we calculated the percentage of each company's total guidelines falling within each of our high and low-level categories. This is to control for the large difference in the number of guidelines between companies and to see how much relative emphasis each placed on different areas.

Figure 3.3 shows the distribution of high-level categories between each of the companies. This shows that the largest difference is that Google gave much more emphasis to model considerations for training data and processes, while Apple and Microsoft spent very little or no emphasis specifically on the model. Beyond that, Interface and Deployment categories dominated in roughly equal proportions at all three companies. Finally, Apple spent marginally less emphasis on initial considerations, although this effect is somewhat small.

Figure 3.1 goes into more detail on which specific categories each company emphasized. The most notable is that Microsoft used nearly 40% of its emphasis on the specific category of mental models. We can also see that Apple seems to have comparatively more emphasis in categories relating to smooth user experiences such as Error Prevention, Cali-

bration, Confidence, and Multiple Options.

These differences may help us understand the effects of the different methodologies used to generate these guidelines, where academic style work will tend to emphasize areas of established academic study in HCI such a mental models, while engineering-driven efforts such as Google's may focus more on the model side, and the culture and values of an organization such as Apple on seamless user experience will affect the kinds of guidelines present when developed from institutional experience. Only when surveyed together these differences become apparent, showing the need for comparative analysis and synthesis of all of these sources of knowledge.

3.5 Discussion

3.5.1 Applications to Visualization

Of particular note to this work are the guidelines that can be informed by visualization research. While most of the guidelines outlined fit within a much more broad HCI context, some may gain from known guidelines within visualization specifically. The interface category is the category most closely related to visualization; with subcategories such as multi-option interfaces having direct parallels in visualization concepts such as small multiples. Furthermore, data visualization research has been a hub of the best methods to achieve explainability such as the guidelines "Explanation via interaction" and "Example-based explanation" being motivating paradigms in interpretability research within visualization[103, 104]. Furthermore, there has been work within visualization assessing how best to implement many of these guidelines such as visualizing uncertainty [105] and understanding biases[106, 107]. As modern AI systems are inherently data-centric, future developments in data visualization clearly inform methods of interfacing with AI products more generally.

3.5.2 Limitations and Future Work

This work has focused on published guidelines from three major companies, however, others such as IBM[108] have developed more sets of guidelines as well. Specific emphasis has been placed on fairness in particular, with guidelines coming from international organizations such as the European Union[109], resulting in many separate specific sets of ethics guidelines that some companies are producing independently from other usability and general-purpose AI guidelines[110]. Further work is needed to integrate these guidelines into the structure put forth in this work. Moreover, control over these guidelines is currently being held by these few very large companies, which may have incentives to emphasize different aspects of AI than the rest of the community. Therefore these guidelines must be augmented by the community. Toward this end the guidelines developed in this work can be found at https://ai-open-guidelines.readthedocs.io/, which puts forth an open call to collect a community-driven set of Human-Centered AI guidelines. Finally, many of these guidelines are clearly aspirational rather than practical, and thus study on the degree to which these and other companies adhere to these guidelines would be of great interest in understanding the actual effectiveness of these guidelines in the development of real products.

3.6 Conclusion

In this work, we have surveyed nearly 200 guidelines for building AI systems from three major technology companies. We have then compared them and developed a single unified taxonomy of AI guidelines. This structure allowed us to see the effects of the different approaches of these companies on what they emphasize as important. Furthermore, this structure can provide a basis of analysis for future work in developing new guidelines from industry, academia, and individuals; and synthesizing information from all of these sources to best provide a more complete reference for anyone looking to build AI systems. We

have taken this work and made it open for extension, so that these guidelines are always available and determined by the community instead of solely by large companies. These guidelines can subsequently be used to structure the design processes developed for more specific use cases throughout the rest of this work.

CHAPTER 4
UNDERSTANDING USER RECOURSE AND INTERPRETABILITY OF LANGUAGE CLASSIFICATION MODELS WITH INTERACTIVE VISUALIZATION

4.1 Introduction

Following the development of the baseline general purpose design guidelines for AI systems, we can then turn to studying the specific context of scientists and how these existing guidelines interact with the peculiarities of scientific workflows described in Section 2.2. Scientists form a distinctly difficult group to study when looking to perform traditional HCI user studies, due to the fact that there are simply not very many of them, doing very different tasks, and their time is extremely valuable, which often makes large scale evaluations of designs prohibitive. However, if by utilizing our understanding of scientific processes (see Section 2.1) when compared to the processes that general purpose design guidelines are built around, we can conceptually extract some aspects of design space relevant to scientists which are transferable. While the specific background knowledge of scientists vary widely between domains and even individuals, as Section 2.1 explains we have strong reason to believe that certain methodological aims of these users are fairly consistent. In particular, as outlined in Figure 2.1, scientific interaction with models involves an interrogative stance towards models involving interpretation and explanation mediated through domain expertise. This negotiation is understudied in the existing explainability literature which generally focuses on how explanations of models assist in users' trust in model predictions[101] rather than interpreting the model in domain terms. Therefore by abstracting the specific domain knowledge, if we can study generally how users interrogate models through interactive visualizations (built in accordance with guidelines from Chapter 3) to

inform their own internal understandings of a topic, we can gain essential data on the design considerations for scientific use cases.

In this chapter we consider a paradigmatic example of a domain where large scale user studies are much more tractable and where users recruited on the internet have meaningful expertise, social media language. All social media users are subject to content moderation practices which are increasingly being performed by automated models which can be flawed [111, 112, 113], thus leading to the imperativeness of users being able to connect their behaviour with the actual standards they are subject to [114, 115]. This interpretive task has the precise structure of interrogative auditing of models to gain domain insight that we are interested in studying, and so we can evaluate with large scale user studies how users interact with a visual machine learning model interpretation system built with existing design guidelines to discover guidelines applicability towards scientific interpretation focused workflows.

4.2 Overview of Ideas and Contributions

In this chapter we will study the effects of model interpretation on user understanding of toxic language moderation by developing an interactive tool called **RECAST** (Figure 4.1), which allows for the interrogation of content moderation models through counterfactual alternative wording and attention visualization. RECAST's design does not require any expertise in machine learning from users, and enables them to visualize their language through the eyes of the algorithm. To this end, the primary focus of RECAST is allowing users to visualize where and how a model detects toxicity within a specific piece of text, make actionable changes to their language to reduce toxicity *as determined by the model*, and thus gain generalizable insights into how the model works to inform future language and spur potential changes if flaws are found. Furthermore, we study the effects of allowing users to experiment with RECAST, analyzing how their language changes as they are more aware of the model. The contributions of this chapter are:

RECAST: an interactive system allowing users to dynamically interrogate the classification of toxicity by visualizing its sources within a piece of text and examining alternative wordings. This provides users a method to understand and interact with toxicity detection models.

Experimental Findings: We find that using RECAST as users learn to optimize language for the model human labeled toxicity can *increase* compared to naïve editing. Having major implications for the robustness of models to interrogative analysis.

Open Source Implementation: of RECAST that enables broad access and future work. RECAST provides a model agnostic framework for analysis of language classification models by not only model developers, but domain users. RECAST can be used so that all relevant parties can become better aware of emerging issues in these models. We have uploaded our source code as supplementary material. We will immediately make the code public on GitHub upon publication of this work.

4.3 Overview of Results and Impact

As Figure 4.5 shows, when users utilize RECAST to apply their understanding of the model towards the actual phenomena of toxic language we find an extremely revealing inversion of results. When using RECAST to optimize language for the model users are able to substantially reduce the toxicity scores according to the model, showing an effectiveness in the visual analytics system in transferring knowledge of the model to end users. However what we find is that in such an optimization, the resultant language performs **worse** with respect to post-hoc human annotated toxicity when compared to naive plain text editing. This finding points out two essential considerations for our continuing work. First, that visualization systems can provide a powerful means for exposing the functioning of models when adhering to the design guidelines found in Chapter 3, with a particular emphasis on interactivity. Secondly that such visualization is *insufficient* for the underlying scientific

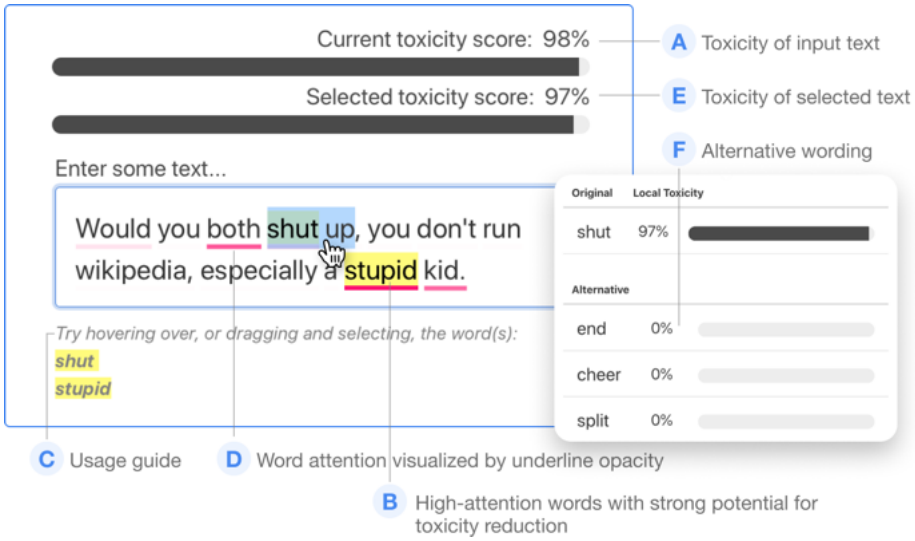

Figure 4.1: The RECAST user interface. **A.** Toxicity score of overall input text shows edits' effect on toxicity in real time. **B.** Words whose possible alternatives have strong potential for toxicity reduction are highlighted in yellow. **C.** Usage guide for RECAST's capabilities. **D.** Underline opacity visualizes model's attention on words, including those without alternatives, to inform users about which words contribute important context **E.** Showing the toxicity score of selected text allows users to localize the sources of toxicity and search for the regions most important to edit. **F.** Hovering over highlighted toxic text displays alternative wording in a pop-up.

interpretation of phenomena, as what we have found is that when introducing humans directly into the process loop for the example model used in RECAST, the model ceases to be effective at differentiating the actual underlying linguistic phenomenon of toxicity. This implies that it must be essential that we consider the participation of users within the design of the underlying models, datasets, and the systems they power in order for them to have hope of producing outputs that are robust in actual interrogative scientific use cases. This insight forms the key differentiation of our approach from the existing design guidelines and most common explainability approaches, and will be utilized in formation of the novel design frameworks introduced in Chapter 5.

4.4 Online Toxicity and Content Moderation

Toxicity online is widespread: a 2015 user survey on online social network platform Reddit found that 50% of negative responses were attributed to hateful or offensive content [116]; however, addressing toxicity through automated means is not trivial—as there must always be a choice of determining what language should be removed and what should not. The same survey *also* found that 35% of complaints from extremely dissatisfied users were about heavy handed moderation and censorship. With the inherent trade-offs baked into content moderation, it is challenging to find a middle ground for this issue. Furthermore, the extreme scale of social media interactions [117] exacerbates these challenges. These issues have lead to the development of automatic toxicity detection models such as the Google Perspective API [111].

Introducing automation, however, raises its own challenges. Machine learning models, responsible for detecting and moderating toxic language, can themselves be flawed [112, 113]. When in the past users could rely on clear community standards from human moderation (or at the least an ability to communicate with a moderator), the adoption of fully automated systems make human-facilitated moderation much more difficult. Moderators who do not understand how automated tools work may not be able to contribute as much

after these tools are adopted. As platforms rely more heavily on automated systems for moderation, users also receive less feedback and might not be able to clearly connect their behaviors with community standards [114]. This fundamentally reduces the effectiveness of automated moderation, as users cannot learn what they did wrong especially if they are unfamiliar with the language or social norms of a platform.

Not only do these models lack feedback for users, but a great deal of work has highlighted inherent biases against certain subgroups based on gender and race [118, 112, 119] within modern machine learning and natural language processing (NLP) models. Although there has been an up-swell of approaches attempting to understand or mitigate these issues [120, 121, 122, 123, 124], the majority of these techniques have been designed from the perspective of developers of machine learning models, with comparatively little emphasis placed on developing tools that are useful to those people who are most affected by these systems: end users.

Without feedback or explanation, users of online forums that use toxicity detection systems based on black-box NLP models might question how their language is being examined. In such scenarios, there is no way to interpret *why* the model considers language toxic. Without tools providing an avenue of recourse [115] designed for actual end users to audit what is being detected and make actionable changes to their language, people are disempowered to participate in discourse online. Furthermore, without the ability to detect when a model is falsely flagging language due to either linguistic limitations or social biases, the work of finding inaccuracies and correcting models is left entirely to the unrepresentative population of machine learning researchers and software engineers. Therefore, providing end-users the ability to interactively audit the models that affect them will help democratize the improvement of these models.

Finally, interactive auditing opens an avenue for recourse. Given an explanation for how a model works, users can re-evaluate writing toxic text, increase awareness of potential limitations in toxicity detection models, and inform people who are unaware of certain

toxic jargon. Black box models, however, are impossible to interrogate. These models provide end-users with little capacity to observe their underlying decision-making processes. Highlighting features that contribute to a model's output provides users with concrete evidence when pursuing recourse.

4.5 Related Work

4.5.1 Content Moderation

There has been extensive research on human-managed online content moderation [125, 126], especially relating to the effect of transparency from the perspective of the end user [127, 128] and the importance of explanations [129]. Various works have called for the development of tools that support, rather than supplant, the efforts of human moderators [129, 130, 126]. Furthermore, several works have focused on studying the development of new automated moderation systems [131, 132], and their effect on the dynamics of the platforms that utilize them [133]. However, most of this research has focused on rule based systems, such as the Reddit Auto-Moderator [114]. In contrast, work that has used statistical machine learning approaches, possibly with the exception of [132], has often focused on accuracy and over transparency or moderator experience [134]. In this work we aim to study the effect of these systems from *the perspective of the end user* as opposed to moderators managing the system. In particular, we look at the effects of deep neural network based moderation systems, which are notoriously difficult to interpret, in comparison to rule based systems. This also introduces issues regarding the perceived legitimacy of platforms, as neural models lack many of the core procedural values required for a platform to be seen as fair, such as due process, transparency, and openness [135]. In some cases, linguistic variation also allows users to bypass automated content moderation systems [136]. Understanding what a model perceives as toxic is integral to evaluating its efficacy. Therefore, a key goal of this work **is to provide or outline implicit procedures used by automated toxicity detection models that can bring higher levels of transparency in these systems.**

Recent work has shown that even limited transparency and explanations from automated systems can be as effective as explanations from human moderators, which "suggest an opportunity for deploying automated tools at a higher rate for the purpose of providing explanations" [129]. We also hope that, similar to [137], this platform can serve as a host for experiments in visualizing the predictions of different types of algorithms and the impact of these visualizations on user behaviour.

4.5.2 Toxicity Reduction Interventions

Many researchers have explored and built both social and technical approaches to identifying, reducing, and combating hateful content online [138, 139, 126, 140, 141, 142]. One of the main solutions these research or platforms propose is to block, ban or suspend the message or the user account. Although removing content or banning relevant users who are perceived as toxic by automated models may reduce their impact to some extent, it may also eliminate legitimate or important speech. A number of interventions have been developed in the CSCW/CHI community to combat hateful content and harassment [143]. For instance, [126] designed a system using psychologically "embedded" CAPTCHAs containing stimuli intended to prime positive emotions and mindsets, influencing discussion positively in online forums, while [144] explored cues that could encourage bystander interventions. [145] raises concerns about conflict between automated moderation systems and human guidelines, studying Wikipedia's current automated moderation system. [132] introduced a sociotechnical moderation system for Reddit called Crossmod to help detect and signal comments that would be removed by moderators. In contrast, our proposed tool RECAST aims at influencing discourse more directly at the end-user stage to promote fairness and interpretability. [146] examined a number of efforts on hate speech and identified three ways of responding to hate speech: (1) removing hateful content, (2) directly rebutting hate speech, and (3) educating and empowering community users. Our proposed tool aligns well with (3), as RECAST enables users to see the toxicity levels of their content transpar-

ently, offers examples of instances when content is and is not toxic, and provides model visualization and reasoning. Furthermore, RECAST helps users with alternative wording, helping them express similar ideas in non-toxic ways.

4.5.3 Natural Language Models and Toxicity Detection

Driving the development of many new toxicity detection methods are recent advances in the field of NLP. In particular, the introduction of massive pretrained Transformer [147] models such as BERT [148] have accelerated the state of the art in many downstream tasks like toxic language classification. The learned representation from these models are often used as the input for a relatively simple model which can be trained on a specific task like toxicity classification. This process is very common as it hugely reduces the amount of training and data required to achieve good results. However, this does mean that any potential issues present in BERT (or other pretrained Transformer models) are inherited by the fine-tuned model. Issues include bias [118, 112, 119], as well as a general lack of semantic understanding [149]. Such issues are particularly important when considering notions of toxicity, which can take many forms, some of which are linguistically pleasant but semantically abhorrent. In fact, work has shown that the Perspective API is susceptible to the same adversarial attacks that fool other NLP models [113]. To address this, there has been some progress on building frameworks to automatically mitigate bias in text directly [150]; however many problems in this space remain open. Despite these issues, deep neural network based models outperform more traditional rule based models in detecting biased language [151], and are increasingly being deployed and thus carrying these flaws into socially-important real-world situations.

4.5.4 Visual Language Interpretability Systems

In order to express information about the model to non-technical end users, our tool fits generally within the tradition of visual analytics for deep learning explainability [18]. From

a visual analytics perspective, various interactive tools have been built for understanding the internal mechanisms of general purpose natural language processing systems. However, (a) there has been little work on understanding the function and impact of toxicity detection systems specifically and (b) these tools are aimed mainly towards developers. Some tools such as SANVis [55] and exBert [56] allow for interactive exploration of the attention mechanisms in Transformer models. The attention scores associated with each word allow connections between words and emphasis of certain words in context to be highlighted. These approaches take a generally static view of the data, where the tool is viewed as a method to explore existing data.

However, dynamic visualizations that allow for user experimentation can assist in the understanding of a machine learning model [59]. Some tools take a more active approach to explain models through counterfactuals [60]. Others, like Errudite [61], allow users to test their own hypotheses with respect to the true error distribution on the entire dataset. Some techniques attempt to adversarially perturb language input to a model [152, 153] in order to change its classification; other methods use a human-in-the-loop design [62]. In this work, we synthesize these two paradigms by passively visualizing model attention, while also enabling interactivity with AI-guided and human-driven counterfactuals. Importantly, we design our tool in the context of automated moderation systems, prioritizing usability from the perspective of non-specialists.

Finally, tools like The Perspective API also offer limited interpretability, allowing platforms to embed text editing areas with a small widget that notifies users with a binary output (toxic/non-toxic) when their input exceeds a toxicity threshold. It also updates as users edit, and allows easy feedback for users to notify that they think the model was incorrect. This provides some of the benefits of RECAST, however it lacks more extensive visual information to help lead users to specific problems in their text, which maintains the black box nature of the model.

4.6 Design of RECAST

In this section, we motivate the design of the RECAST tool through a formative user survey, and formalize a series of design goals from the concerns raised in our user survey.

4.6.1 Formative Survey for Understanding Automated Content Moderation

To understand difficulties relating to toxicity moderation, and to outline user needs (especially those related to recourse), we surveyed 100 Amazon Mechanical Turk Workers with social media accounts in the United States about their experience surrounding online toxicity and moderation.[1]

We asked how often users noticed toxic language on the internet and found that 62% responded noticing toxicity either 'often' or 'always' compared to only 6% responding with either 'rarely' or 'never.' Our survey also highlights the clear effect of racism as an undeniable component, with 16% of non-white users reporting 'always' noticing toxicity compared to only 4% of white users (and only 2% for white males). Toxicity is not only noticed but caustic: we asked about the effect toxicity had on users lives both online and offline, and found that overall 43% responded that toxic language online had a 'somewhat negative' or 'very negative' effect on their lives offline. This effect was also influenced by gender, with only 20% of males reporting negative effects while 52% of non-males reporting negative effects offline. We find that not only is toxic language pervasive but it disproportionately and intersectionally harms already underprivileged groups.

Given the groups' overall clear negative experience of toxicity, we also wanted to understand how they felt about the trade-off between free speech on online spaces and toxic language. In an open ended question, 36 responses were highly supportive of strong mod-

[1]The survey took an average of 6 minutes to complete with compensation of $0.80, above the US Federal minimum wage. Workers were selected from the Amazon Mechanical Turk pool with filters to ensure they were within the United States, and held Reddit and Twitter accounts to ensure a certain familiarity with online discourse. Respondents had an average age of 34 (standard deviation of 8 years), identified as 66% male, 33% female, and 1% nonbinary, and 75% White, 9% Asian, 8% black, 2% Latino, 1% Native American, and 5% other/unspecified.

eration to reduce toxicity, with responses along the lines of *"I think the tradeoff is well worth it. I am so tired of hearing foul language all the time."* At the same time 31 were very skeptical of any moderation and highly supportive of user freedom of speech, responding *"I think free speech is important. So people can say what they want even if it is toxic."* These divided responses highlight the inherent difficulty of balancing the standards of content moderation and freedom of speech.

Many neutral responses were additionally skeptical of automated systems in particular. For example, one participant noted that *"in many platforms moderation is biased. The automatic tools are not sufficient to remove toxic material from the site. These tools end up removing quality content instead of toxic ones."* A common complaint from these responses claimed that addressing the issue of toxicity through simple models would exacerbate existing tensions between moderators and users. To understand the specific problems users had with these automated systems we then asked about feedback. Among participants who have had a post removed either by a moderator or an automated system, we found that when removed by a human, 52% reported receiving feedback 'often' or 'always', while only 36% reported receiving feedback from automated systems.

Finally, to understand which areas users felt need the most improvement, we asked participants how they would improve existing systems for online content moderation, and what features they would want in a tool. Concretely, users noted wanting familiarity and similarity to existing tools for modifying language in spellcheck and auto-complete interfaces **"*like auto correct but for language vs. typos*"** that could simply **"*suggest other words*"**.

This aligned with a desire that any tool should be *"something friendly and approachable that isn't too intrusive or annoying."* Many of the responses were outright hostile to the concept of a tool for reducing toxicity for fear of censorship, meaning that any such tool would be most effective the less visible it is—similar to the notion of nudging in behavioural economics [154].

Stemming from the same well justified fear, users emphasized that such a tool requires

"an appeals process, information about why a post is removed, a rejection of a post with advice for fixing it before posting something"; and *"the ability to give feedback on the tool since it will almost certainly have failure scenarios. The tool should also be able to work in real time, and have an excellent understanding of English."* From our survey, we found that users indirectly recognize that current machine learning systems do not have a perfect understanding of natural language. Thus, a more concrete way to provide feedback is very important for such a system. Finally one user noted that they would like to see the sources for a model, *"guidelines.. lots of written comments of what is deemed toxic"*, which helps justify the need for these models and datasets to be open source.

4.6.2 Design Goals

We synthesized the information from the study to identify the main design goals for RECAST.

G1 Interpretation. Users often feel as if they are unsure of the specific guidelines and requirements they are asked to maintain, and may not know what about their language may cause a post to be removed. Therefore it is important to provide explanations to users that are useful not only for a specific comment but more generally for how the classifiers work. Ease of interpretability will enable users to build appropriate mental models for planning how they use language in the future. These interpretations can provide the bedrock of any potential action a user may want to make either on or off of the platform.

G2 User Driven. In order to make sure users do not feel overly censored, we ensure that no decision about editing text is to be made without the explicit choice of the user, and that a wide variety of options be presented to maximize the capability of the user to say what they mean. This differentiates RECAST from fully automatic, end-to-end approaches, which have been presented for reducing bias or toxicity [150]. A trade-off with this design principle concerns users who are determined to use toxic

language. However, such users would not use a tool like this in the first place if it prevented their ultimate desired language. In order to understand how real users might use similar tools, and to study how even toxic users interact with moderation models, we prioritize a user-centered design.

G3 Minimalism. In order to ensure accessibility for end users who are not familiar with complex data visualization paradigms, and to make sure that the tool is not overbearing or irritating, we aim to build a tool that minimizes extraneous views. This also has a trade-off of precluding more advanced or comprehensive user interfaces; for the purposes of this study, however, providing greater accessibility for users (for instance users whose first language is not English) is an important consideration and thus is prioritised.

G4 Easy Feedback. Anticipating that any model for detecting toxicity will be flawed, it is valuable for users to be able to highlight erroneous classification easily to ensure they feel they are being heard, and to improve the underlying model when possible.

G5 Accessibility. To develop a tool that is accessible for users without specialized computational resources, we deploy our tool using lightweight modern web technologies, and place emphasis on ensuring our system runs efficiently for low-resourced users. We also open-source our code to support reproducible research.

4.6.3 Ethical Considerations

Outside of the primary design goals of RECAST, there are special considerations we need to give to potential ethical issues[2]. RECAST enables users to edit text so that it is no longer be detected as toxic by a classification model. However it is highly possible that the resulting text can in reality remain toxic. Aspects of this work are similar in functionality to work done to generate adversarial examples for text classification [155, 62]. Bad actors could

[2]These considerations are in addition to standard practice considerations with anonymous data collection and annotations. This research study has been approved by the Institutional Review Board (IRB) at the researchers' institution.

potentially use RECAST to pass truly toxic language past existing filters.

To avoid these scenarios, we have included controls within RECAST that prevent it being used in such a scenario where it detects explicit hate speech or uneditable/irredeemable toxicity. Many bad actors already have the ability to bypass models through the method of trial and error. The end result of proliferation of a tool like RECAST is not to improve the ability for bad actors to 'game the algorithm,' as it is already being gamed [136]. For example, the #thyghgapp vs #thighgap phenomena, outlined by [136], highlights how bad actors already change syntax to avoid detection from automated models. Rather, RECAST aims to provide a novel service to users who are already acting in good faith to provide transparency about automatic moderation methods. This set of issues is not unique to RECAST, instead these issues implicate the toxicity detection models themselves, and the platforms deploying them. Users cannot demand or instigate change if they do not understand these issues, which further motivates the development of RECAST despite its initial risks.

Finally, a primary contribution of this work is not just the development of RECAST for its own sake, but also as a means of understanding how toxicity detection models affect people's choice of language when compared to human labeled toxicity. We suspect that optimizing language for automated models are an inevitable consequence of their increased deployment (as language approved by such models will become the only kind of language made visible). Because of these consequences, and in the effort to prevent any malicious use, our released open source contribution will include an informal consent form that discusses appropriate ethical issues, regulations, and best practices.

4.7 RECAST

RECAST (Figure 4.1) is an online interactive tool with the primary focus of allowing users to visualize toxicity within a text, make changes to reduce toxicity, and gain generalizable insights into how toxicity classifier models work.

At the same time, it is important to note what RECAST is *not*. RECAST is *not* designed

to necessarily change, explain, or interpret anything regarding "real" toxicity as far as it even can be directly or objectively identified. Rather RECAST allows users' access to, interpretation of, and interaction with, *toxicity detection models*. As we have established, the current state of the art of NLP can be linguistically naïve [149], thus we expect that notions of "true" toxicity and detected toxicity will diverge, especially in the scenario when users try to edit toxic language to comply with these models. Furthermore it is also important to note that the contribution of RECAST is not novel NLP or visualization methods (in fact the RECAST architecture is model agnostic), but rather in the synthesis of existing tools in addressing specific user issues and then studying the effect that the systems underlying these issues have on online discourse and user experience.

4.7.1 Visualizing Comment Toxicity

The primary information RECAST expresses is the toxicity classification score of user input. Users can input text and view the toxicity of the overall sentence, along with which words contribute most to the output score. RECAST displays a score between 0 and 100, which represents the probability that the model assigns to the input for whether or not it is toxic. This is shown at the very top of the tool as a bar (Figure 4.1A). The minimal bar design, in monochrome and removed somewhat from the text, is noticeable enough to provide informational feedback, while subdued enough to not distract from the text itself. The toxicity bar dynamically updates as the user edits the text. This allows users to experiment with their own or suggested edits and get real time feedback for any counterfactual scenarios. This enables both effective exploration to choose the best possible wording and easy iteration to test hypothesis for how the model works. The inclusion of this metric visualization may incentivize a local optimization or 'hill-climbing' approach. However this approach better aligns with the user desire for limited intervention, as a user may be more likely to make an edit if it is a small change to their text rather than complete rewrite. Furthermore the reactive visual provides immediate feedback, facilitating experimentation

Explanation Method	Human Annotation Overlap	Average Compute Time
Integrated Gradient Based	$.86 \pm .06$	1880 ms
Attention Based	$.87 \pm .07$	99 ms

Table 4.1: Comparing gradient and attention based methods for flagging toxic words in RECAST. Evaluations were run a computer with an Intel i7 2600K and a single NVIDIA GTX 1070.

which has shown to be effective in building understanding [59].

4.7.2 Explaining Toxicity Classification

A key component of RECAST involves identifying the tokens in a text that are indicative of toxicity, and attributing importance to these tokens. There are two widely-used automated techniques that perform attribution: gradient based explanation and attention based explanation [147, 156]. Both these techniques offer a numeric value for each word in an input sentence, where the magnitude of the numeric value corresponds to relative importance in a model's prediction. However, the capacity for these methods to explain model predictions may differ across tasks [157]. In this subsection, we compare gradient and attention based explanations to human annotations and to each-other, selecting an appropriate technique for RECAST's backend.

Although there is much debate as to whether attention is a good proxy for explanation [157], when interpreted carefully, attention can be a *rough and weak proxy* for explanation [158, 159]. As a precaution, we compared our attention based metric (denoted as $attn$) to typical gradient/saliency based techniques (denoted as $grad$). For $attn$, we computed the average attention score over last layer heads in our Transformer model (before the linear classification layers) using the end/CLS token on the input text. On the other hand, $grad$ was calculated using an integrated gradient approach documented by [156], using the standard gradient operation on the input to the model to evaluate importance. For evaluation purposes, we collected two different metrics: speed of inference, and set overlap ($\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$) for flagged words in sentences. In the set overlap metric, X

and Y are sets of flagged words from a given sentence.

To effectively compare both our techniques, we aligned differing output ranges for $grad$ and $attn$, since $attn$ is bounded between $[0, 1]$ and $grad$ between $[0, \infty]$. We manually tuned cutoffs for attention (i.e. $attn > .2$) and found cutoffs for gradient based approaches by collecting the distribution of $attn$ scores, $p(x)$, and $grad$ scores, $q(x)$, over a random 5% subset of our training dataset (7979 instances). Then we took the percentile value of our .2 $attn$ cutoff at $P^{-1}(.2) = .9$, and we identified the corresponding gradient cutoff using the percentile value from the attention distributions, $q^{-1}(P^{-1}(.2)) = .02$. Finally, we selected words from a smaller subset (50 instances) to compare attention and gradient based flagging to human annotations.

We conducted two analyses to compare gradient and attention based methods for explainability:

1. Analyze word overlap and inference speed on a random subset of our training data (7979 instances), comparing $grad$ and $attn$.

2. Analyze word overlap on a smaller random subset of our training data (50 instances), comparing $grad$, $attn$, and human annotations. The guidelines for this task required annotators to flag all words contributing to toxicity either implicitly or explicitly.

For analysis 1, we found that the average $overlap(grad, attn) = .82 \pm .02$ at a 95% confidence interval, for 7979 instances. For inference times, saliency methods required significantly more time due to the added back-propagation step. We recorded flagging speed per batch, with $grad$ at 1.88 s $\pm$ 9.92 ms (mean $\pm$ std. dev. of 7 runs, 1 loop each), and $attn$ at 98.8 ms $\pm$ 2.46 ms per loop (mean $\pm$ std. dev. of 7 runs, 10 loops each). **For analysis 2,** the authors who are familiar with the task setup annotated 50 random examples manually, highlighting tokens they considered toxic. We recorded average $overlap(grad, attn) = .79 \pm .12$, $overlap(grad, human) = .86 \pm .06$, and $overlap(attn, human) = .87 \pm .07$ – all values are at 95% confidence intervals. Re-

gardless of flagging technique, we find that each method highlights core toxic elements in text that align with human annotations.

In order to achieve real-time explanations with minimal latency, gradient explanations are prohibitively slow ($1.88s$ vs $98.8ms$). In our token flagging task, both attention and gradient techniques perform similarly, and have reasonable overlap with human annotations. Therefore, we use the attention to explain which words are highly associated with our model's predictions. We importantly understand that attention, in some contexts, may not explain model prediction. However, in our specific scenario (flagging toxic phrases), attention is both **significantly faster and flags similar tokens** when compared with human annotation (overlap at .82 and .87 for both task 1 and 2, respectively). Because our selected techniques perform similarly, and $attn$ provides improved inference speeds, we utilize $attn$ for this work.

4.7.3 Visualizing Model Attention

Various visualization concepts have been used to show the relative importance of words and visualize attention, such as highlighting and opacity [160]. However, we utilized an underline on every word, where the opacity of each underline would be controlled by the magnitude of attention placed on each word (Figure 4.1D). This method was chosen as it helped with legibility of the text, which is vital for users understanding differences in textual classifications. Like the classification bar, its design is purposefully simple, mirroring the text interaction techniques of underlining to note where editing is required—something end users are familiar with from common text editing software. This accessibility allows RECAST to effectively communicate the complexities of toxicity detection models using a visual language users are already fluent in.

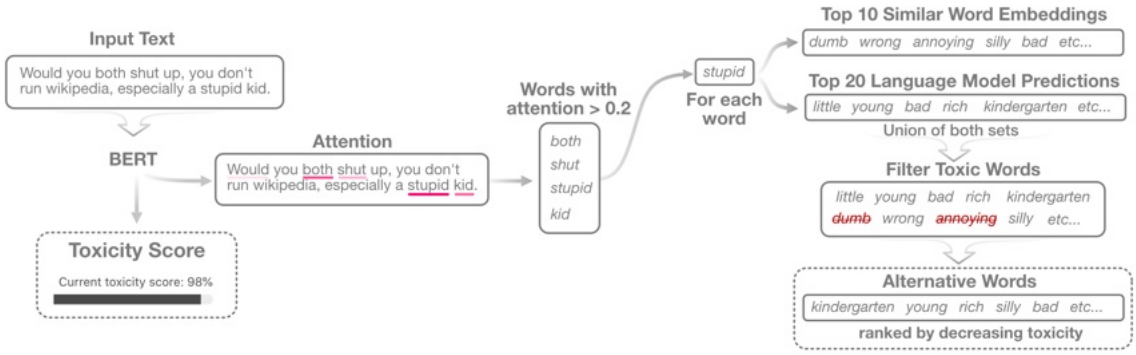

Figure 4.2: Process for generating alternative words

4.7.4 Alternative Wording

Beyond passive visualizations, RECAST suggests concrete, actionable edits to text, assisting users in lowering toxicity through alternative wording. The alternative wording feature provides users with options to swap or delete words in a sentence that are responsible for high toxicity scores. A word in the input is highlighted (Figure 4.1B) to draw particular attention to it when it meets the criteria of:

1. An attention score greater that 0.2
2. RECAST can find alternative words with individual toxicity less than 0.4
3. The alternatives have a positive impact on the overall input toxicity.

These thresholds were determined during the analysis to best match the attention highlights to human annotation. The requirements of alternatives having a positive impact both globally (for the whole text) and locally (for the replacement word on its own to be benign) provides a safeguard against malicious use.

When the user hovers over any of these words, suggested substitutions are shown and ranked in a popup (Figure 4.1F). Selecting one of these alternatives replaces the word and the new toxicity score is updated. This mode of interaction is also easy and intuitive for users due to its similarity to familiar spellcheck or thesaurus tools (motivated by our

survey), requires little retyping of edits, and gives options if users cannot immediately think of an alternative word. Furthermore, it displays a range of options which gives the end user agency in maintaining the original meaning as closely as possible. Finally, beyond the act of making the sentence less toxic, the technique allows users to learn which words tend to be highlighted, and what common synonyms the algorithm tends to suggest. This allows people to learn about the model and use this knowledge while writing future comments.

Figure 4.2 illustrates how the set of alternative words for a given toxic input word is calculated. Given a word for which we are calculating potential alternatives we first find its nearest neighbors in a Glove [161] word embedding space [162]. We limit nearest neighbor search to 10, balancing the amount of user choice in word options while reducing the cognitive load of choosing among too many [163]. These words will match the original word closely in meaning, and provide a solid base of options for users to reword from. Furthermore these vectors are not dependent on a manually curated list of synonyms and extends to find similar but non-synonym words. Next we use a BERT language model to find other words that may fit within the context of the word to be replaced. We do this by feeding the original input into the language model with the selected word masked, causing the language model to output a probability distribution of likely words that fit within that context. From these words, we select the 20 most likely. Then, we take the union of these two sets and filter out any words with individual toxicity greater than 0.4, along with words that do not have a positive impact on the overall input toxicity.

Using current state of the art NLP models does not ensure that every option will be a good replacement. To this end, RECAST highlights several alternatives so the user will likely find at least one good replacement that they can select. Our controls ensure that no replacement makes the result worse. This gives the user the most amount of control over the process while still leveraging all of the potential options and power of modern NLP systems.

4.7.5 Multiple Alternatives

In addition to replacing single words, sometimes toxic words come in groups or phrases where the toxicity is not individually attributable to any one of the words, this may require a different editing paradigm for the end user. To support this, RECAST allows a user to select a contiguous text phrase, and alternatives are generated for the n-gram of toxic words contained within that phrase. Sets of words are chosen from the same universe of words as for single word replacements through word embedding similarity. Furthermore, the language model masks the entire set of words to be replaced and gets the most likely tuples from the resulting joint distribution. This allows RECAST to encode the linguistic coherence of not just each word in context but the whole set of words within context. Finally sets of words are ranked by the resultant toxicity of the edit on the selection as before.

4.7.6 User Feedback

Knowing that the classification model is expected to make mistakes, and that a goal is to provide user recourse for handling those mistakes, we have also included an integrated feedback form within the tool. If a user feels the model has made a mistake, there is an included text box below the main input space for comments to submit to the developers. In a deployed system, this would forward complaints on to the relevant platform. This can be used as a means for re-training and improving the model as well as providing a direct way for users to pressure platforms when models exhibits bias or other issues. By logging inaccuracies highlighted using RECAST, end users can concretely identify when models utilize tokens that should not be attributed with a toxic prediction. Furthermore, visual feedback from RECAST provides developers and researchers with identifiable sources of errors in their models.

4.7.7 System Implementation

While RECAST as a tool is built to be model agnostic (as long as a model uses attention or similar method), for our evaluation we needed to include a backend implementation of current state of the art toxicity detection models as a useful proxy for deployed systems.

Dataset

The dataset we used for training the backend model for RECAST was sourced by the Kaggle competition run by Google's Jigsaw [164], which is based on the dataset used by Google's Perspective API [111]. The Perspective API is used as one of the most commonly used and openly available content moderation tools. Therefore, its underlying dataset was a suitable proxy for RECAST's goal. By benchmarking against this dataset, we can compare our performance directly against that of the Perspective API, and thus be well justified in the representativeness of our model.

We also chose to use this dataset because it was pre-cleaned, openly available, and contains a wide variety of baseline models through the Kaggle competition. The dataset consists of a set of 312735 comments from Wikipedia's talk page edits, along with multiple labels that characterise the form of toxicity (toxic, severe toxic, obscene, threat, insult, and identity hate). We chose to only use the toxic label in the dataset for modeling purposes, as the other labels were subsets of toxicity and we wanted a sharper focus.

A noteworthy limitation of this dataset is its focus on explicit hate speech, or speech that directly insults through the use of particular keywords and phrases (like "shut up" and "stupid," as seen in Figure 4.1). Implicit hate speech, however, tends to focus on stereotypes, avoiding explicit phrases (e.g., "you're smart for a girl" containing no individually toxic components yet expressing a misogynist meaning). Future work on collecting implicit hate speech is needed to help extend RECAST and other toxicity detection systems to support such examples.

Model	ROC-AUC
Logistic Regression	0.963
Naïve-Bayes SVM	0.972
LSTM	0.977
Fine Tuned BERT (RECAST)	**0.982**
Large Ensemble (Kaggle Leader)	0.989

Table 4.2: Kaggle Reported Toxicity Detection Performance. Highlighting the fine tuned BERT model used in this work.

Model Architecture

To detect toxicity in text, we fine-tuned a state-of-the-art Transformer based model (BERT) that performs reasonably well across various language modeling tasks. Transformer based models rely extensively on self-attention to predict text [147]. Concretely, self-attention allows a model to detect toxicity based on the context of a word. Transformer models work by applying self-attention mechanisms to the input several times, over several layers. A final output is selected by propagating attention across these layers. Although there are a wide range of possible models, such as those proposed in the original 2017 Jigsaw Kaggle competition [164], we decided to utilize Transformer models due to their prevalence in most modern NLP tasks, as RECAST aims to be generally useful for current models.

Table 4.2 summarizes performances across several baseline classifiers using the ROC-AUC metric [165], which is a standard metric in machine learning classification problems and the one reported in the Kaggle leaderboard. The BERT model we used performs on par with the current state of the art Kaggle leader-board by utilizing a single model as opposed to an extremely opaque ensemble of models. Furthermore, it remains representative of the current trends in NLP.

Software

All deep learning based models used in our system were implemented in PyTorch [166], a library for building deep neural networks. We also utilized the HuggingFace package

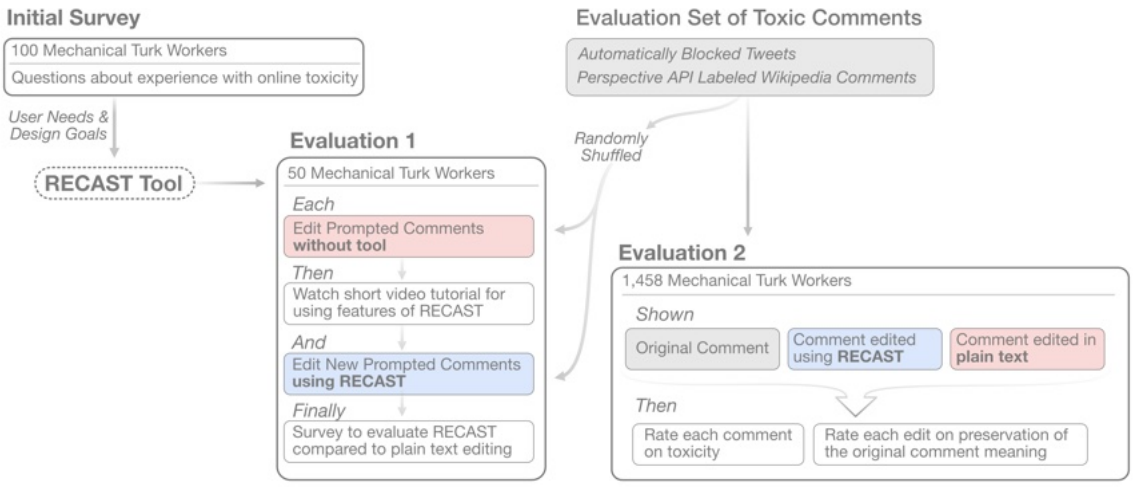

Figure 4.3: Multiple Evaluation Procedure

for pretrained Transformer models. Although we use Transformers across RECAST, we built our frontend to be model agnostic, and so the backend model can easily be swapped out without major code changes (provided the replacement model supports generating attention-like explanations for its predictions). Our frontend was written using Svelte for compartmentalizing our code, and D3.js for miscellaneous visual elements. Because RECAST's predictions occur on the backend, our client itself can run on systems with reduced computational power. RECAST does not present any novel NLP architectures, techniques or methods. Instead, it is built upon the current state of the art, with consideration for real time performance. A contribution of this work is the usable tool itself as an implementation designed for this specific use case, and the insights gained from being able to study users interacting with the tool.

4.8 Evaluation

We conducted two coordinated evaluations of RECAST, as outlined in Figure 4.3, to study how well it addressed our design goals as well as to study the effect that user interpretability of toxicity detection models might have on online discourse.

4.8.1 Evaluation 1: Editing Toxic Comments with RECAST

Methodology

In order to evaluate how users would use a tool like RECAST, we considered the task of editing toxic comments on social media. This scenario is representative of a situation a user would be presented with when using a tool like RECAST—users would be interested in interacting with the model and potentially making changes to their original comment only after their commented is detected as toxic. Since one of our insights from the initial study was that users would prefer a tool to be lightweight, we imagine that RECAST would only be deployed in these kinds of editing situations, where a comment is already recognized as toxic instead of being used every time a user writes from scratch. As a result, we present annotators with potentially toxic comments compared to asking annotators to come up with some on their own. While there may be some difference between users editing provided comments instead of editing their own comments, having a common set of comments for users to edit and thus comparable resulting outputs provides a larger benefit in terms of reproducible analysis.

Sample Pool

For this evaluation we conducted a within-subject study which compared editing comments using RECAST to normal editing without any help from RECAST. We recruited 50 users (with a mean age of 38, consisting of 32 self-identified males and 18 self-identified females, all within the United States), from Amazon Mechanical Turk, an online microtasking platform. Users were paid $1.80 per task (matching the United States federal minimum wage).

Task Description

The task users were given was to edit a set of provided comments scraped from two sources. The Perspective Kaggle dataset [164], consisting of Wikipedia comments manually labeled

Survey Question	Proportion Agrees	95% CI
The tool is easy to use.	78%	±12%
The tool is helpful in reducing the toxicity of a comment.	70%	±13%
The tool is helpful in understanding the criterion of labeling comments as toxic.	80%	±11%

Table 4.3: User evaluations of RECAST. Both "Agree" and "Strongly Agree" are included as agreement.

as toxic, provided a strong source of comments for model training. To supplement our study with comments outside of the Kaggle dataset, we took replies on Twitter scraped on May 9, 2020, found by looking at top trending hashtags in the US. We collected responses that were hidden below all other replies, behind the following filter warning: "Show additional replies, including those that may contain offensive content."[3] This mixture of sampled comments allowed us to examine both examples within and outside of the training dataset. In the case of Twitter, we provided users with contextual thread information to help them decide how to edit the comments. For a given editing task, users were provided a description of the context of the original comment, as well as preceding comments in the cases where they were available.

Users were initially prompted to edit 4 comments using a standard text editor, then shown a video describing the features of RECAST. Next, users were asked to edit a second set of four comments using RECAST. In order to compare the resulting texts, we randomized the set of comments provided to the RECAST enabled and RECAST disabled groups. Finally users were asked to rate on a five-point Likert scale the degree to which they thought the addition of the tool did or did not help them reduce comment toxicity and understand the mechanics of the toxicity detection model. Table 4.3 shows that strong majorities found the tool to be easy to use and helpful, and provided a good understanding of the model's heuristics.

[3]Specific comment text for each of these cases used in the user study can be found in the appendix.

Open Ended Responses

To further validate these results and account for positivity bias in the responses, and to analyze the effectiveness of RECAST in user understanding, we asked open-ended questions about what they learned about the model to gauge the generalizability of the patterns they learned through the study. We asked: *"After using the tool, how would you characterize by what criterion language gets labeled as toxic versus benign?"*.

Many users noticed the tendency of the model to focus more on specific keywords than overall sentiment, as two users noted:

> *"For the most part they get labeled by individual words with a negative connotation."*
>
> *"I think that it tends to pick keywords that can be considered highly offensive."*

Some users pointed out how keywords extended beyond just directly toxic words but common co-occurrences as well:

> *"I think that language that is obviously offensive (slurs, etc.) is labeled as toxic, as well as words that, with a high frequency, occur often close to other offensive words to make up larger phrases."*

However some users noticed the cultural influence and flaws in which words were highlighted:

> *"I think slang words and curse words are flagged more than negative opinions."*
>
> *"I think that, especially slang, gets misconstrued within the tool and they falsely label it as toxic, when in reality it's not."*

One user also noticed the flaws of the underlying model by experimenting themselves with the tool, as they noticed differences between toxic language that the model highlights, and toxic meaning which it does not:

> *"I think that 'language' is a bad way to determine what's toxic. You can write terrible things in cordial proper language, and also be kind in crass harsh language. In many cases I couldn't make something not toxic without changing the entire premise, as people were just trying to be rude no matter what."*

They later went on to describe the experimenting they had done:

> *"I tested 'I love this motherfucker, I'd take a bullet for him any day of the week' it comes back as 100% toxic. Saying 'I genuinely hope you just donť wake up tomorrow' is 2%. There's clearly a flaw in the system"*.

Takeaways

These responses showcases some generic takeaways users gained through using RECAST:

1. Current automated moderation models focus mostly on individual words, not higher level meaning.

2. Words that are considered toxic are sometimes influenced by dialect and slang.

These findings are consistent with how linguists describe modern NLP behaviour[149], showing how **RECAST enables nontechnical users to quickly understand heuristically how highly complicated language classification models work in practice**. Furthermore, RECAST provides a canvas for easy experimentation that enables users to effectively find flaws in the system and generate meaningful specific critiques, empowering users to potentially take action where they otherwise would not be able to.

4.8.2 Evaluation 2: Toxic Comment Editing Comparison

In our second evaluation, we compare the resulting edited text provided by users in study 1 to analyze the effect that editing with or without RECAST has on the final comments. This will help us not only understand the effect of RECAST on online discourse, but also

study, more generally, how users might interact and fine tune their language when using an automated toxicity filter.

Methodology

To do this, we recruited 1,458 participants from Amazon Mechanical Turk and asked them to compare two edited comments—one generated by a participant in study 1 in the RECAST-enabled condition, and another generated by a participant in study 1 in the RECAST-disabled condition. These comment pairs were randomly selected from the same prompt for each condition. Three different participants were asked to assess each set of: original comment, RECAST-enabled edit, and RECAST-disabled edit (anonymized as *Edit 1* and *Edit 2* randomly). Participants were asked to rate how much they view each comment version as toxic on a five-point scale, and were also asked how well they perceived each edit preserved the general content and intent of the original comment.

Results

We found that both the RECAST enabled and disabled case were statistically indistinguishable with respect to maintaining the original general meaning, both $60\% \pm 2\%$ of the time. RECAST-enabled and RECAST-disabled are comparable conditions to look at toxicity, since neither is disproportionately changing the input so as to be incomparable. As expected, the original comments which had already been labeled as toxic, either within the Wikipedia dataset or by the Twitter content filter, were mostly considered toxic, though a significant minority of these original comments were also labeled as not-toxic, highlighting how people perceive toxicity as non-binary. As shown in Figure 4.5, the edits made with RECAST disabled were generally classified as less toxic than comments made with RECAST enabled.

Figure 4.4 showcases the joint distributions of the original comment toxicity label in each of the edit conditions. A closer look at the joint distribution suggests that in the RECAST enabled case, the labeled toxicity is more highly correlated with the original toxicity

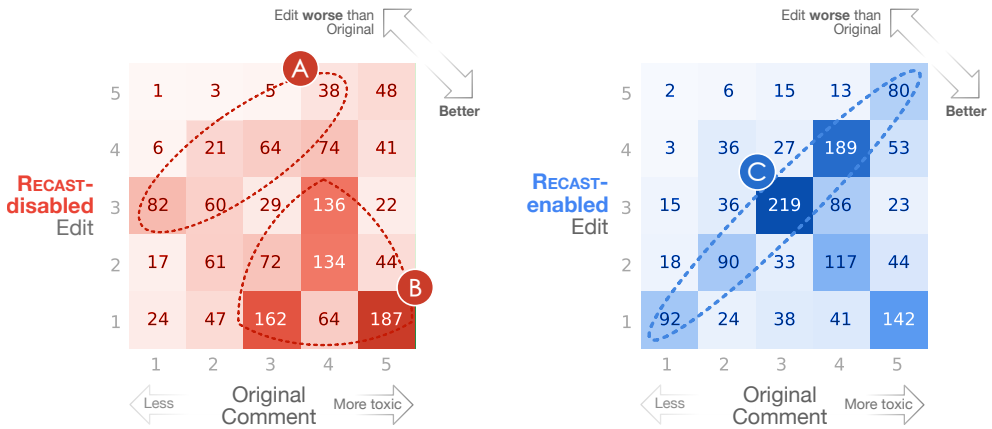

Figure 4.4: Joint distributions of enabled and disabled edits vs original comment toxicity. Note (A) which showcases the upper diagonal of the disabled case where the resulting toxicity is higher than original toxicity. This region is higher populated than in the enabled case, showing that there is a higher risk of increasing toxicity when writing edits without RECAST. However (B) showcases that without RECAST, even high toxicity comments are often reduced. Overall the disabled case shows that without a tool the resultant toxicity is independent of the original. (C) highlights the opposite effect in the enabled case, where the strong representation along the diagonal shows that the resulting toxicity of edits generated using RECAST is more likely to be similar to the original toxicity, which is a benefit in that there are fewer cases where toxicity is increased in the upper diagonal, but a cost in the lower effectiveness of reducing toxicity in the lower diagonal.

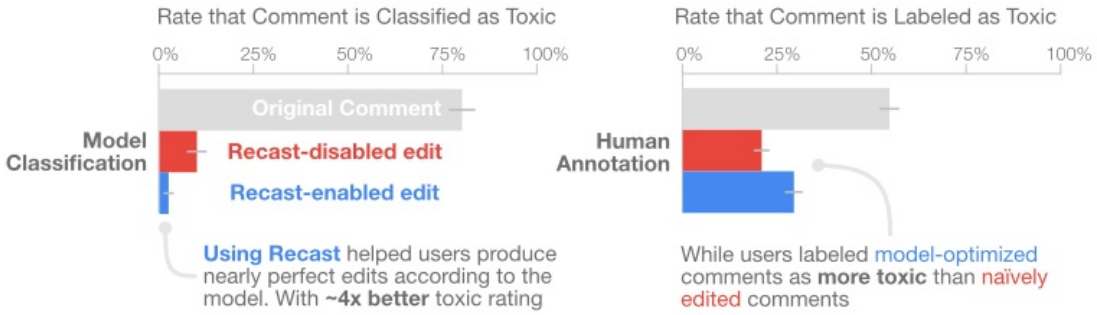

Figure 4.5: Distribution of toxic labels (either 'Agree' or 'Strongly Agree' with statement that a comment is toxic for human annotation, or classification by model) for unedited comments, comments edited without RECAST, and comments edited with RECAST. Error bars show the 95% binomial proportion confidence interval under the asymptotic normal approximation. We find that RECAST does produce optimal comments according to the model. However we find that the model systematically under reports toxicity among edited comments, and that model optimized comments are labeled as on average more likely to be toxic by human annotators.

when compared to the disabled case (Kendall Tau[167] of $\tau = 0.15$ with $p < 0.1$ for enabled and $\tau = -0.03$ with $p > .1$ for the disabled case). This shows that when not using RECAST, users' ability to write less toxic versions were independent of the original toxicity. However, when using RECAST, the original toxicity heavily informed the resulting text, where users were less likely to make highly toxic comments much better but much less likely to make comparatively less toxic comments worse.

Comparing Human Detected Toxicity to Model Classified Toxicity in Edited Comments

Finally, we looked at the difference between the human annotations of toxicity and model classifications of these edits. We reclassified the outputs of each edit with our fine tuned toxicity classification model, and compared the resulting classifications to the corresponding human labels. As models become more impactful gatekeepers of what language is and is not allowed, any heuristic which works better for the model will be structurally incentivized and potentially become more common. By examining the difference between the moderation using human determined toxicity and by model determined toxicity, we can

potentially hypothesize future directions of online discourse as such models become more prevalent.

In Figure 4.5 we see that the RECAST enabled comments remained classified as toxic $2.4 \pm 1.4\%$ of the time, while with RECAST disabled the model classified the result as toxic $9.9 \pm 2.6\%$ of the time. At the same time, edits made using RECAST do not reliably decrease the human-annotated toxicity of comments, especially for comments with already high toxicity that need the most editing. However, edited comments with originally high toxicity are still classified as toxic by the detection model only a quarter as often as the same comments edited without RECAST.

Takeaways

This analysis shows that:

1. RECAST was highly effective at helping users reduce toxicity *as detected through the model*, but not as effective at reducing human annotated toxicity.

2. Therefore when language is optimized for the model (which is what is implicitly incentivized by the deployment of these models), the model ceases to be a good judge of toxicity as determined by human annotators. Language that is less toxic to the model can be more toxic to humans. "When a measure becomes a target, it ceases to be a good measure"[168].

4.9 Discussion and Implications

4.9.1 RECAST as a tool for reducing toxicity

When discussing the use of RECAST as a tool for reducing toxicity, we have shown that there are two, potentially competing, meanings of the task. There is the underlying notion of toxicity as language with an adverse effect on people, and there is toxicity as the output

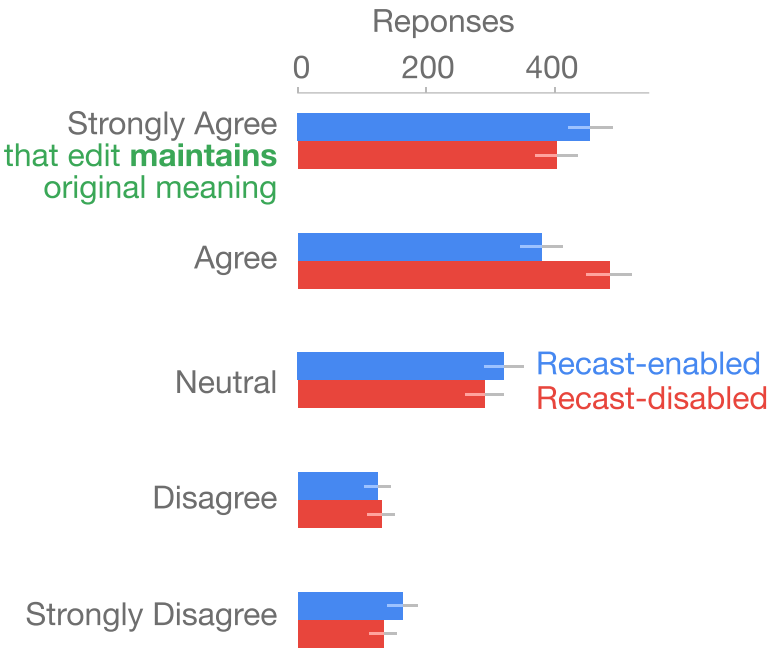

Figure 4.6: Distribution of assessments of how well Enabled and Disabled condition edits maintain the original meaning. χ^2 test shows no statistically significant different between the distributions (with high confidence $p > .99$).

of the models used to moderate platforms. The development of models like the Perspective API are predicated on the idea that these two concepts are—if not the same—at least asymptotically close as models improve. However, our work has shown that when users have direct access to the toxicity classification models, their optimized language to match the model does **not** also optimize human labeled toxicity.

RECAST is clearly effective at allowing users to reduce the model toxicity of their comments, and does this with good ease of use and accessibility, while maintaining the original meaning. RECAST also provides a path of least resistance between language that is classified as toxic and language not classified as toxic, and then gives users the power to choose how to use that knowledge. This is useful for a variety of people. Users who may not have an strong fluency of English may inadvertently say things considered toxic without knowledge, and RECAST provides a frictionless way for these users to make changes and learn what is acceptable.

On the other hand, some users are better informed than the model when identifying

toxic language. We have highlighted how implicit definitions of toxicity used by models are fundamentally different than what humans consider toxic; toxicity is only a meaningful concept in so far as it has an effect on people, not computers. As users pointed out in subsubsection 4.8.1, slang or dialect may be misclassified as toxic. RECAST allows these users to both circumvent potentially unjust/biased models, and raise awareness of these issues through explained examples.

While RECAST does not seem to reduce human labeled toxicity due to users optimizing for the model output, this further emphasizes the need for user recourse and model oversight in this space. RECAST is an initial effort towards this goal. However, we recognize that there is substantial work that needs to be done from the developer side of these platforms—and potentially on the policy side—in order for the recourse that RECAST provides to be impactful.

4.9.2 RECAST as a tool for interpreting toxicity models

RECAST was also designed to allow exploration and visualization of toxicity models in order to allow users to understand them and provide a degree of transparency. A large majority of study participants in our evaluation reported that RECAST helped them understand the toxicity model (Table 4.3), and many users were able to produce insightful comments about the patterns presented by the model through RECAST. This is a significant component of recourse that RECAST helps facilitate. When toxicity classification models are used to police language, then users are subject to often biased rules that they may not even know are being applied. Providing actionable recourse is predicated on those affected by those systems understanding *how* they are affected, which is nearly impossible when the system is a black box algorithm. RECAST hopes to break into that black box and allow end-users and non-experts to then see these rules as they are applied.

4.9.3 Future of Online Discourse

As machine learning systems are more often deployed to manage moderation online, users will be required to tune their language to match the standards set forth by these models. This will occur regardless of the availability of tools like RECAST, as only language meeting these models' standards will be visible or filtered (creating a form of survivorship bias). By introducing RECAST as a way to more directly optimize for these models, we can study the long term evolutionary effect of misapplied filters on future online discourse. Our evaluation quantitatively highlights how the standards required to succeed with a model diverge from human perceived toxicity; **as model based standards become more prevalent, toxicity according, to human standards, may in fact increase**. This is a troubling trend in our large scale quantitative evaluation that warrants further study.

4.9.4 Limitations and Future Work

The design of the RECAST interface is also meant to be generalizable as models improve. Based on prior research in the toxicity detection space, RECAST utilized the state-of-the-art BERT model to estimate the degree of toxicity in messages and suggest alternative wordings. Despite using a specific type of toxicity detection model, RECAST is agnostic to BERT specifically and can be easily combined with other machine learning models. Similarly, though the alternative wording suggestion component currently relies on Word2Vec, it serves as a generic framework and is compatible with other embeddings or techniques to generate as broad a space of options for users as possible. RECAST is also agnostic to model explanation techniques, provided explanations are based on individual words or phrases within a text. Because of our justification in subsection Subsection 4.7.2, we expect gradient-based model explanations to yield similar results due to flagging of similar tokens. Finally, RECAST will enable future to work validate the effectiveness of new explainability techniques on model assisted intervention for content moderation.

With respect to toxicity detection models themselves, our work highlights the numerous

challenges associated with automated moderation systems. For example, hateful content may be expressed in multiple ways, e.g., sarcasm, irony, coded text. Users may even use hateful words or phrases to refer to themselves. Instead of investigating different forms of hateful content, we work with a large-scale benchmark corpus with a pre-defined set of toxicity labels. RECAST can be further extended to handle various formats of toxic speech. Furthermore, RECAST could allows users to interact with models to take various definitions of toxicity to their limits by optimizing their language. As such, RECAST may provide a useful backbone for the study of different notions of toxicity.

Our evaluation of RECAST was mainly conducted on Amazon Mechanical Turk with annotators in a lab-like environment. As a result, we could not assess the long-term effect introduced by RECAST. Future work could build upon our research to further investigate whether users will be likely to use tools like RECAST in their daily interactions on different online platforms, and how RECAST's involvement affects users' subsequent participation.

This also relates to another area of future work: building out implementations of RECAST that may run in the browser. We actively made decisions to ensure RECAST is light weight and can be run without significant computational resources on the front-end. By making the code open source, we open an avenue for future work to expand the functionality of RECAST into a browser extension. This would help both validate the results of the studies we have run by embedding RECAST in a more realistic scenario, but also make RECAST accessible to the users who may benefit from it.

However, we caution future researchers to carefully weigh the ethical implications of widely deploying RECAST, as such functionality may be highly useful for those users working in good faith, but potentially harmful if used by bad actors. These risks are inherent in any functionality helping users navigate these systems, as explained in Subsection 4.6.3. Our study finds that there is an important distinction between two notions of toxicity: (1) Language detected by a model as toxic, and (2) Language that has adverse effects on real people. An inherent risk of visual analytics tools is their ability to only optimize for (1),

which while we may hope better aligns with (2) in the future; we find that currently it does not. Thus while RECAST is effective at helping users acting in good faith to reduce (1), its inability to consistently reduce (2) elucidates flaws in current NLP models rather than the specific design of RECAST. Finally, an important limitation of any tool is that it requires users to *want* to lower toxicity; which of course is often not the case. However, explicitly handling malicious users is outside the scope of this work, and future work studying when users act maliciously and how to better design human-AI interfaces to **de-escalate** toxic behaviour before suggesting alternatives—may yield better systems for automated content moderation .

4.10 Conclusion

In this chapter we considered the problem of end-user interpretation of machine learning toxicity detection models by introduced the interactive tool, RECAST. RECAST provides users the ability to interact with toxicity detection models and visualize how they work. Through these interactions, users are able to make actionable changes to their language in order to reduce toxicity, while gaining generalizable insights about toxicity models. Through multiple large scale user evaluations, we showed the effectiveness of RECAST in helping users edit text to decrease model defined toxicity while providing interpretable explanations to users. At the same time we highlighted the pitfalls of using toxicity detection models for moderation, as toxicity defined by the model differed from human labeled toxicity. This provides an empirical example of the limitations of direct machine learning model interpretation (along the lines of data-centric scientific framework from Section 2.2), where the mental theories induced by interpreting a purely data-centric model can substantially diverge from theories regarding the underlying phenomena; and thus that even if machine learning models can be made to be less opaque and easily interpreted, these interpretations may have less explanatory power than hoped for.

CHAPTER 5

ISHMAP : A HUMAN-CENTERED DESIGN FRAMEWORK FOR SCIENTIFIC ANOMALY DETECTION MODELS

5.1 Introduction

Now that we have established core structures (Chapter 2), design principles (Chapter 3), and limitations (Chapter 4) of designing machine learning systems for scientific applications, in this chapter we will take these insights and apply them in a concrete scientific application in order to develop a design framework to structure scientific machine learning model development that is actually useful to world leading working scientists. Recall from Chapter 2 that according to Kuhn [35], revolutionary discoveries occur with the accumulation of anomalies, which are identified when perceptible phenomena cannot be integrated into an existing paradigm. However when looking at Figure 2.1, under a more data-centric framework new data is never conceptualized into paradigmatic phenomena in a form that is amenable to the accumulation of anomalies *as such*. Therefore enabling new methods of anomaly *conceptualization* is the stage of scientific inquiry as described in Section 2.2 most essential to develop new tools to facilitate discovery.

Even while novel machine learning models are rapidly improving the state of the art in standard formulations of anomaly detection[169], scientific applications present an interesting and complex problem for different from the standard. While research has explored how algorithms can be made to be explain their reasoning[170, 58, 57], this chapter investigates how HCI methods can be deeply integrated within the framing of model development to enable anomaly phenomena conceptualization (see Figure 2.1) to drive interpretability as a user-defined quality that is considered a first class objective rather than a post-hoc computed explanation.

In this chapter we present an application of just such a design process, we engaged in a multi-year collaboration with a team of scientists at NASA who analyze data from the PIXL instrument to understand Martian geochemistry, and thus the possibility of extra-terrestrial life [171, 172, 173, 174]. Our collaboration set forth with the following research questions:

1. **RQ1:** Within the specific scientific workflow of PIXL scientists, what are the particular requirements that a modeling approach must satisfy?

2. **RQ2:** In what ways and to what degree does the existing standard approach to anomaly detection through machine learning satisfy and violate these requirements?

3. **RQ3:** How might a different anomaly detection modeling framework to enable the development of more effective systems given these requirements?

In the course of addressing these questions we explored broader approaches to anomaly detection in a scientific context and report the following four contributions:

1. **Formative design study**. We present the findings of an 18 month long study where through a series of contextual inquiry interviews we outlined the analytic workflow of the NASA PIXL science team, found key challenges faced by scientists in detecting and interpreting spectral anomalies in X-ray florescence (XRF) data, and developed a comprehensive model of how PIXL scientists approach anomaly detection. This study revealed three key design goals used to guide the development of tools assisting in this workflow.

2. **Novel spectroscopy anomaly detection algorithm**. We describe a new method to automatically detect and classify diffraction and other spectral anomalies accurately (93.4% test accuracy), directly within raw unquantified spectra, providing a significant improvement over existing methods.

3. **Deployed algorithm and visualization system** . We embedded this algorithm within a visualization tool that has been deployed and has now become a regular and impor-

tant component of all PIXL based analysis conducted over the past year, used daily by over 97 NASA scientists and NASA-affiliated scientists around the globe. We evaluate the success of the application further by examining some examples of novel planetary science enabled by the tool.

4. **We introduce a novel framework, ISHMAP, for the collaborative development of anomaly detection tools for scientific teams like PIXL**. Finally we present a framework, Iterated Semantic Heuristic Modeling of Anomalous Phenomena (ISHMAP), which was developed from the success of this application that can serve as a useful tool in itself for future collaborations in similar scientific anomaly detection settings. This framework integrates HCI and AI perspectives on model development and presents a different formulation of the problem of anomaly detection specifically designed to fulfill the needs of scientific users. ISHMAP introduces a process for how to produce anomaly detection models that provide first-class scientific interpretations by default, ensuring scientific users buy-in, and also can be tightly integrated with existing modeling techniques (including deep learning approaches).

This in diving deep into this specific application, this paper looks showcase an example of a possible way to synthesize AI anomaly detection research with methods developed in HCI. We believe that by combining methods across disciplines, researchers may be better able to take on high priority problems like anomaly detection, in partnership with scientific communities, and to help drive discovery. We offer evidence of how this bridging supported our own inquiry in the case of the NASA PIXL science team. In this next section, we situate this work within the larger landscape of HCI, AI, and Astrobiology research.

5.2 Background and Related Work

5.2.1 The Search for Extraterrestrial Life

The search for extraterrestrial life is among the great contemporary scientific endeavors [171], and Mars forms a key component of that search [172]. A principal way scientists around the world build an understanding of Mars as a possible host for life is to study its planetary-scale geology and geochemistry over time [173, 174, 175]. NASA's PIXL instrument supports that ongoing investigation by capturing co-aligned visible imaging and thousands of spatially localized pairs of *X-ray fluorescence* (XRF) spectra in a single experiment [176].

The PIXL instrument represents a generational change in the scale and sensitivity of extra-terrestrial XRF measurements[177, 178, 179]. While this can bring analytical leaps, it also means that the data are more sensitive to spectral anomalies that can lead scientists to misinterpret data.

Anomalies in XRF spectra have been historically identified manually. But with each experiment site including thousands of spectra to manually investigate, each with hundreds of peaks, anomalies become increasingly difficult and time consuming to manually identify. And missing spectral anomalies could lead to the misinterpretation of the elemental chemistry, mineralogy and ultimately the planetary scale environment that acted upon the samples under investigation. Therefore there is substantial scientific interest in performing reliable and interpretable anomaly detection on the incoming data from PIXL.

5.2.2 Bridging HCI and AI methods for Interpretable Modeling

While having more independent origins, the recent history of hybrid disciplines such as HCI-AI[82], HCML [83], and human-guided ML [84] reflect an interest in drawing on knowledge generated across fields to jointly inform the development of systems that touch on each discipline. HCI researchers, for example have looked to understand the challenges

to design AI systems that fit with user needs[85], and to use new properties exposed by these systems as a resource for designers[86]. The complexity of dealing with some form of embedded intelligence led other researchers to introduce particular methods to structure ideation and iteration of AI- and ML-systems for ubicomp[87, 88] or dialog systems[89].

Alternatively, researchers in AI and ML have drawn upon expert knowledge to inform ML models [90], or to bring interactivity into learning systems [91], looking to leverage interaction to define ML model rules [92]. These methods have been applied to anomaly detection in cybersecurity, spatio-temporal, and behaviour modeling contexts[93].

In the sciences, this crossover has looked at interactivity and glass-box models as a way to support interpretable and configurable deployed machine learning models [94]. However, fieldwork in disciplines such as oceanography have shown that while interpretation and understanding of code and models is essential, it is insufficient to contributing within a larger scientific workflow as the primary driver of change will often come from anomalous "moments of flux" which naturally lead to reconceptualisations that are not amenable to fixed data perspective implicit in any traditional data science or machine learning model, as "a singular focus on problem-solving may marginalize opportunities for innovation that could drive community engagement, and, therefore, momentum and adoption"[95]. Therefore this work looks to expand the tradition of utilization of participatory design practises in the context of AI approaches to science by enabling a more flexible modeling approach while maintaining established key aspects of interpretability.

5.3 Formative Study

This section presents the findings of our inquiry into the analytic workflow of PIXL data by scientists focusing primarily on anomalies. We started this work with a series of contextual inquiry interviews [180] conducted in a cadence of approximately every two weeks over the course of 18 months to understand the different users of PIXL data. While we spoke to and collaborated with many dozens of scientists working with PIXL data, we focused our

attention with five primary users: three spectroscopists who we will refer to as R1, R2, and R3; a sedimentologist, R4; and a geochemist, R5.

Our research question entering into this study was to find what are the primary constraints that any modeling intervention in this context must satisfy in order to be useful within the scientists analytic workflow (RQ1). Through these interviews we were able to define three primary design constraints which elucidate the requirements for any anomaly detection method to be useful to PIXL scientists. While each design goal is firmly situated within the context of PIXL analysis, the underling rationale and aspects of PIXL data that inform each goal are not unique to PIXL and likely are applicable in a broad range of scientific applications.

5.3.1 Background on PIXL Science Workflow

In order to understand the context of the design goals for anomaly detection for PIXL, we must first establish some basic background of the data formats and processing steps scientists work with. During an experiment, PIXL operates on a specially designated sample known as a *target*. PIXL sends an X-ray beam into a location on the rock's surface. At each location, PIXL's X-ray beam causes the chemical elements in the target to fluoresce, which is captured in a data type called a *spectrum* [181], a 4096-index array of *channels*. Each channel records the count of electrons, sensed at a distinct energy level measured in kilo electron-volts, or *keV*. PIXL's *A* and *B* detectors each record the distinctive fluorescence patterns emitted by each point on a target from two different phase angles. These distinctive responses take the form of *fluorescence peaks*, which are Gaussian peaks of a fixed width of channels over background measurements which are dependant on the chemical composition of the X-ray beam location. During an experiment, PIXL's camera also captures visible light images of a target. When data are returned to Earth, the X-ray beam geometry is reconstructed, and each spectral point is localized within each of the returned images. Overall, an experiment at a target returns a series of visible light images, and

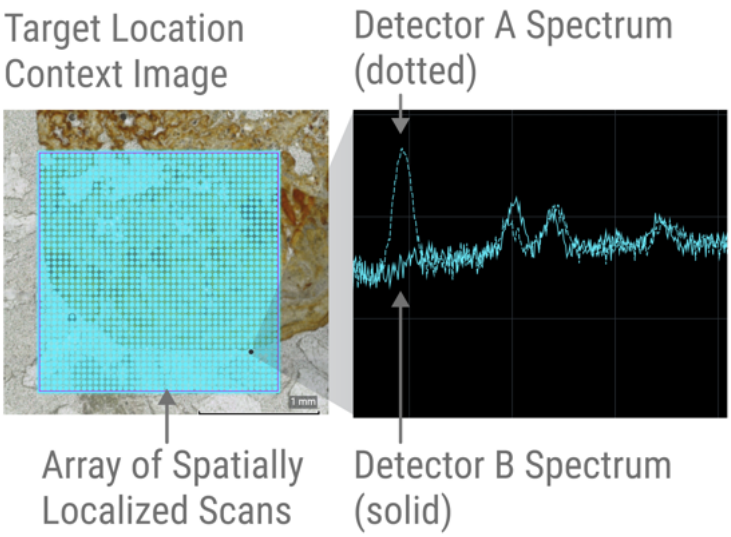

Figure 5.1: Overview of data provided by PIXL instrument

around 4,000 spectral points, each with an A and B spectra, and x and y coordinates within each image [176] (see Figure Figure 5.1).

Once an experiment is conducted with the PIXL instrument on Mars, the spectral and image data from that experiment is sent back to earth and analyzed primarily through the PIXLISE analysis user interface [182]. This data is analyzed and transformed in multiple steps by different subject matter experts. The first step in analysis, spectroscopists translate the peaks in the spectra into elements that they believe are present in the target [183].

The spectroscopist then uses their list of elements to *quantify* the spectra, translating the intensity of the various peaks into an empirical estimate of the percent of the total elemental mass each element constitutes in the target [183]. PIXL uses the PIQUANT quantification algorithm [184, 185], and exposes those capabilities through the PIXLISE user interface. Each spectra is quantified independently by PIQUANT, while spectroscopists determine the set of elements to quantify using the *bulk sum* of all of the spectra in the dataset.

After the spectra have been quantified into elemental weight percents the broader science team begins to analyze the dataset. Since pure elements are uncommon in nature, the next task of the science team is to determine how the elements that have been detected

and quantified combine to form some combination of the currently known 5700+ minerals [186]. The primary driver of this analysis is the quantified elemental weight percent map of the dataset, which is visualised using many standard Geology and Geochemistry data visualization techniques such as *ternary diagrams* and *heatmaps* combined with visualizations unique to PIXLISE such as *chord diagrams*. These quantitative signals are importantly augmented with additional signals from the color, shape, and texture of the rock in the images, or its *morphology*, to consider mineral candidates.

As candidate minerals emerge, scientists then designate them as *regions of interest*, or ROI's, and evolve theories of the historical geologic processes that brought the rock and its individual ROIs into existence and altered them over time [173]. These theories become increasingly dependent on the contextual and visual information and how minerals are situated as the discussion broadens to include Astrobiologists who can aggregate the details of the geochemistry, geophysics, and climate, to build a long-term theory of the broader context of the target, site and region and implications towards biological habitability.

5.3.2 Design Goal 1: Focus on Raw Data Over Processed and Quantified Data

When considering the problem of anomaly detection in this workflow we first sought to understand what data structure to analyze, the raw spectra or the quantification. What we observed was that for many scientists the information from the PIXL instrument spectra was almost entirely mediated through elemental quantification of fluorescence phenomena and the visualization of these quantifications within PIXLISE. This means that for the most part, anomalies are found by discovering unexpected results in a quantification and backtracking to find some non-fluorescent spectral phenomena that is causing an erroneous quantification. As R2 described:

> " ... I'm getting quite a bit of activity in a barium oxide map, should I be getting that much... ? ... So I might then ... look at those individual spectra "

This method has obvious downsides, as any anomaly that causes an error in quantification

not unusual enough to merit deeper scrutiny can propagate misleading information. Furthermore, the total amount of time and effort spent on downstream correction can be much greater than early detection, especially given the scale of data that PIXL produces. Therefore we formed our first design constraint (*DG1*) that we should **perform analysis on raw and unquantified spectra** in the hopes of catching phenomena that may be obfuscated through bulk sum quantification.

5.3.3 Design Goal 2: Robustness to Limited Ground Truth Labeling

The next constraint we found was that the actual amount of reliable ground truth labels currently existing in PIXL datasets is very limited. This is due to the requirement of manual processing by a small number of expert users to reach reliable conclusions, paired with the novel scale of data being produced by PIXL. R1 expressed their desire to help winnow down potential anomalies before digging into deeper analysis:

> "What we want ... in the automation phase is reduce ... one data set a day with 5000 spectra down to ... a few spectra a week. Because ... this is a multi-day interrogation [for] one data set, and so ... with all of the other science outputs ... ... you then want to go flagged for looking at later."

Thus we formed our second constraint (*DG2*) to be that our method must be **robust to a small number of ground truth labels** and thus provide a reasonable number of flags for experts to be able to precisely analyze.

5.3.4 Design Goal 3: Allow Differentiation of Anomalies by Scientific Causal Process

Another major constraint that was emphatically expressed to us throughout our initial interviews was the vital importance of understanding why given anomalies may be presented within the context of scientific models. For instance R1 described why spectra are looked at manually currently due to the huge number of different ways to model a spectra and thus the requirement of background knowledge:

> "Fluorescence data is fitted manually for a reason. ... you could have something with ... 15 lines from rare earth elements in the spectra. There's always some expert user who knows something about the sample fitting the data. ... because the combinations are infinite, that it's not something ... automatically done."

Furthermore, anomalous phenomena may contain useful information in themselves about the physical processes that causes them, and thus may be worthy of analysis in their own right. We found that the non-fluorescent phenomena most often discovered when manually investigating quantification anomalies is diffraction, which is an effect often investigated on its own in the context of purpose built X-ray diffraction instruments [187], but which has a signal response sometimes overlapping fluorescence peaks. Thus we hypothesized that such spectral anomalies, if sufficiently differentiated, could be used as another source of auxiliary information similar to the visual context imaging within the mineral identification process. What is currently lacking is a way to find and characterize anomalies early in the pipeline and then utilize these detections for improved downstream analysis. R1 stated the goal similarly:

> "The ultimate test would be if fluorescence yields an ambiguous mineral that the diffraction can make unambiguous."

Therefore we determined for our final design goal (*DG3*) that it is very important for scientists to be able to not only find anomalies but **interpret and differentiate different kinds of anomalies** in the context of helping them understand the underlying science.

5.3.5 Comparison to the Existing Approach: Evaluating Standard Machine Learning Based Anomaly Detection Methods Using Design Goals

Once we had understood the domain and relevant design goals we sought to evaluate standard approaches to the problem of anomaly detection and take into account any potential

issues (RQ2), and to guide the specific problems we need to address in out solution (RQ3). These standard approaches for anomaly detection and machine learning can be broadly broken down into traditional methods and deep learning based methods. Traditional methods tend to be designed for tabular data and are thus generally not well suited for analysis of raw data as required by design goal Section DG1, while deep learning models are well known for their adaptability to complex non-tabular data formats. Deep learning based anomaly detection methods tend to follow the general structure of training a model to encode data into a compact representation to find structural patterns and outliers [169]. The main classes of these methods that do not need extensive labeling (as required by design goal Section DG2) are feature extraction methods and normality representation methods[169], both of which contain explicit assumptions that conflict with our design goals. Feature extraction methods assume that "The feature representations extracted by deep learning models preserve the discriminative information that helps separate anomalies from normal instances." This violates our finding from goal Section DG2, where measurement is expensive and thus sufficient data to form a comprehensive feature set that includes rare and nuanced anomaly classes without explicit labels is not available. Normality representing models take the form of latent space encoders such as auto-encoders or generative adversarial networks and assume that non-anomalous instances can be better represented and reconstructed by these encoding models than anomalous instances. We found this to not be true in practice, when considering the constraint of goal Section DG2 we had no way to preemptively sort out normal as opposed to anomalous instances for semi-supervision meaning anomalous instances must be included in the training data. This can cause problems as the model then will learn to represent those instances just as well as normal instances. This makes further sense when considering that the choice of reconstruction loss function is "designed for dimension reduction or data compression, rather than anomaly detection. As a result, the resulting representations are a generic summarization of underlying regularities, which are not optimized for detecting irregularities." This property, when paired with

the lack of supervised labels from Section DG2, fundamentally violates goal Section DG3 as general purpose patterns explicitly do not prioritize truly rare or categorically different anomalies but will always prioritize either single point anomalies or simple sub-samples of classes of normal data which happen to be more rare.

Our goal of anomaly differentiation by scientific interpretation Section DG3 shows clearly how scientific end users care most about the *underlying causal processes* as opposed to the most clear surface level empirical patterns. Lacking robust labeling, unsupervised methods all ultimately can do nothing but optimize in different ways for surface level empirical similarity without considering different causal processes. Thus even when such methods are able to discover some of the anomalies that are present, they fundamentally keep the task of sorting through the important vs unimportant classes of anomaly as a manual process. We hope to be able to reverse this order of operations, and allow scientists to differentiate the kinds of anomalies they care about first, and then detect them directly allowing for pre-sorted interpretations when processing new data.

5.4 Method

When considering the weaknesses of deep learning and traditional data science based anomaly detection, a root cause of issues is that the problem framing involves either no direct input from scientists (which inevitably violates design goal Section DG3) or only has input mediated through labeling (which in order to communicate required nuance would require a scale that violates design goal Section DG2). Thus we set out to develop an alternative model development framework which can more effectively and efficiently incorporate scientists' prior knowledge into anomaly detection models (RQ3).

A key insight in developing this framework is the differentiation between *phenomena* and *data*. Most existing methods focus purely on the space of data, and thus anomalies must be defined as individual data points. However scientists work in an ontology that is more abstracted from data space, where there are many underlying processes that can occur in

the physical world and these processes can be measured in many different, or incomplete manners. What is important is not tied to a single datum, but rather what any subset or superset of data can imply about the underlying phenomena. Thus what we consider as phenomena can occur at multiple levels of scale. A single data point of a high dimension or complexity may contain within it multiple instances of different phenomena.

This form of analysis is the one side of a trade-off. Data space analysis optimizes for *completeness* by encoding all measurable information contained samples from dataset up to the limits of the size of the dataset without regard for *correctness* or why any given datum has a particular set of features. By modeling phenomena we inherently limit completeness as only phenomena considered explicitly can be modeled, which must by necessity be less than the innumerable number of underlying factors that could be modeled given perfect knowledge. However what we gain is correctness, where phenomena that are of interest can be modeled more fittingly to their natural scale and be separated for interpretation by default rather than through post-hoc analysis of data space latent encodings.

Here we present a design framework, *Iterative Semantic Heuristic Modeling of Anomalous Phenomena* (ISHMAP, see Figure Figure 5.2), that can provide a template for developer and scientist collaboration to perform phenomena based anomaly analysis. Guided by the design goals laid out previously it can produce heuristic raw data feature extractors based on scientifically determined meaningful anomalous phenomena, and iterate adaptively based on the limited amount of scientist time available. We utilized this framework within the context of PIXL science and here discuss both the details of its was application within PIXL science to enable novel scientific discoveries as well as laying out the general principles as a potential resource for other collaborations guided by similar design goals.

5.4.1 Scientist Description of Anomalous Phenomena

The entry point into ISHMAP and the key differentiation between phenomena as opposed to data centered analysis is starting with scientist driven explicit identification of

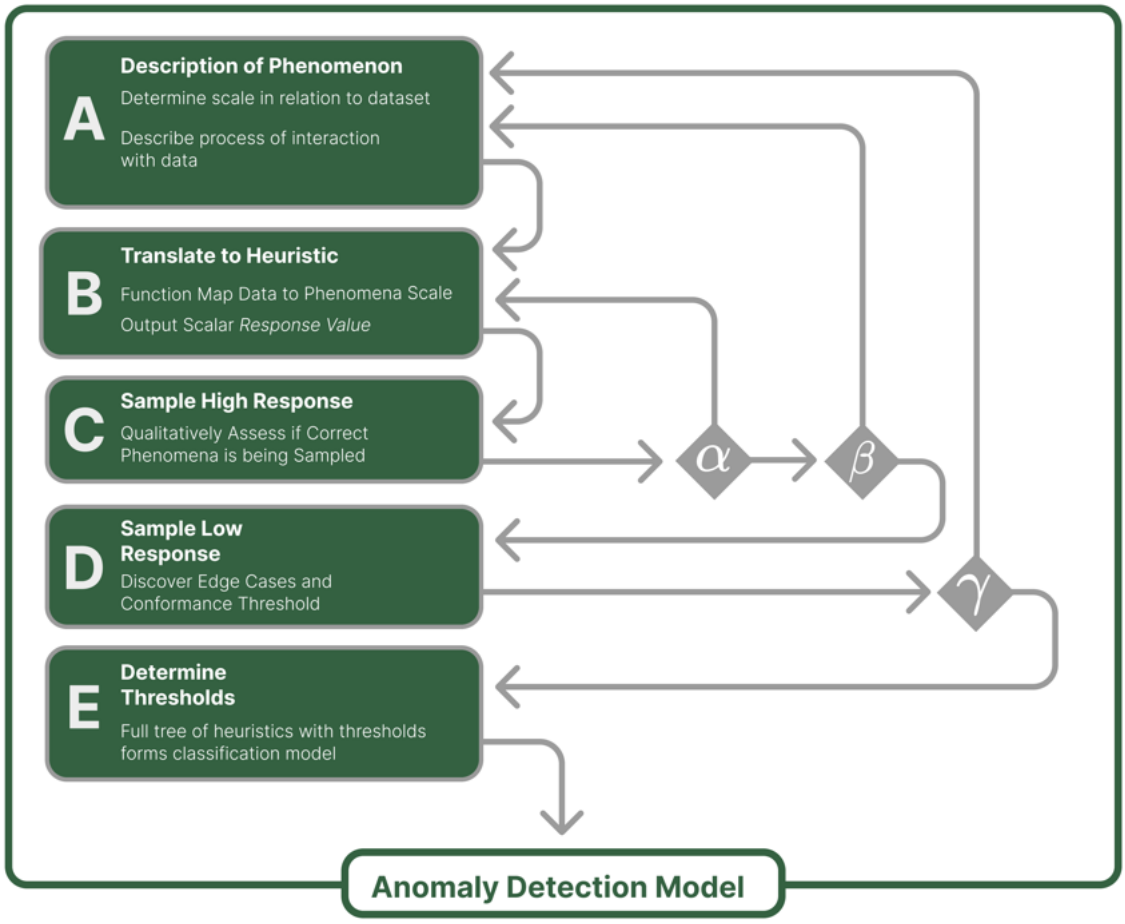

Figure 5.2: Overview of the ISHMAP Design Framework. Starting with scientist lead descriptions of phenomena (A), translated by developers into a computable heuristic function (B), which enables sampling of archetypal data instances (C), iterated until heuristics can provide a clear signal for the high response samples, followed by sampling of low response samples to determine a classification threshold (D), which when determined allows the return of a finalized model of the target phenomenon (E). An example of such a model as a result of this process can be seen in figure Figure 5.3. ISHMAP contains three distinct iteration cycles. Cycle α is homologous to the standard machine learning training loop and iterates the heuristic function. Cycle β recursively disambiguates between distinct phenomena that are co-selected by a given heuristic. Cycle γ detects and models cases of scientific ambiguity as a distinct class of phenomena in order to detect and model ambiguitities independently and improve model performance as well as build up detections of ambiguities sufficient to eventually build more reliable scientiifc conceptualizations.

a specific anomalous phenomenon (Fig. Figure 5.2A). The process begins by outlining a semantic class of anomalous phenomena within the ontology of the given scientific domain. This specific class is chosen by scientists as conveying some important information. With PIXL we started with the phenomena of diffraction. Scientists chose to start with this phenomenon because diffraction only occurs with a particular and theoretically well-understood set of conditions involving crystal structure. This means that the presence of diffraction can tell scientists very important information about the physical structure of a sample that elemental composition alone cannot differentiate.

Once a phenomenon is decided upon the first characteristic that must be determined is the scale at which the phenomena is measured by the available data. Does the phenomenon occur as a specific kind of data point? Does is manifest as a pattern between adjacent points? Or does it occur, potentially multiple times, as a subset within a single data point? The scale of analysis is determined based on the prior understanding both of the physical phenomenon as well as the characteristics of the measurement methods producing the available data. This scale determination is absolutely essential to enabling properly interpretable heuristics, as it determines the *input* of the heuristic function and thus the way it can express a phenomenon as the most natural explanation for how a phenomenon exhibits itself in a larger dataset. What must be decided for a scale determination is a **sampling procedure to extract from the primal dataset all potential instances of the target phenomenon**, as well as a map back to the primal dataset determining which parts of specific data point or points are being included in a sample. For our example scientists know that diffraction occurs as distinct peaks within spectra, and based on the known resolution properties of the PIXL instrument, these peak responses are assumed to be discrete signals within a window of size 0.2 KeV which is the full-width-half-max size of detectable gaussian peaks for the detector [188]. This means that the sampled input for a heuristic will be all contiguous windows of width 0.2 KeV which can be individually evaluated as potential diffraction peaks.

Once the correct scale of analysis is determined, scientists should then describe the differential causal process of the phenomenon with respect to the data measurement. What this entails is describing how the given phenomenon interacts with the measurement process and thus the differential between the described anomalous and non-anomalous data with respect to their underlying known or hypothesized causality. This description does not need to be extremely thorough, as it will later be translated through various lossy processes, it merely has to describe they primary ways in which this phenomenon differs from the default assumptions of the model, and how these differences manifest in the data. A well chosen scale determination will tend to greatly decrease the complexity of such descriptions when compared to descriptions that must work purely in the primal dataset. This description will form the basic starting point from which heuristics can be designed and iterated upon. For diffraction the causal process is well understood, diffraction is an effect that occurs when the PIXL instrument is particularly aligned with a crystal structure in the sample and PIXL sends X-rays of the correct frequency that resonates with the lattice the response will scatter with constructive interference forming a response peak at that resonant frequency. These response peaks are similar in shape to florescence as gaussian peaks with width determined by the detector resolution, however their causal process differs such that they can occur at arbitrary frequencies as opposed to solely at elemental florescent frequencies and the spatial dependence of the effect is sensitive enough that a diffraction response in one of PIXL's two detectors is very unlikely to occur at the same frequency in the other detector. Once scientists have formed such a description of the scale of a phenomenon and the causal forward process that generates differential data measurement, developers can proceed to the next step of ISHMAP (Fig. Figure 5.2B).

5.4.2 Translation into Heuristic Model

After a definition of the phenomenon is provided by the scientists, it is then the job of the developer to translate this definition into a computable heuristic model (Fig. Figure 5.2B).

This stage of the framework can take many different forms depending on the nature of the description provided. The only requirement being that the end result of an iteration of development be a program that takes as input a sample of data of the form set forth by the phenomenon scale characterization and output a scalar value proportional to how well a given sample conforms to the forward process of the anomaly characterization. The form of this program can depend on a number of factors including the format of forward process description, the amount and format of available data in the phenomenon scale, and compute resources available. When designing the initial heuristic for diffraction we chose to manually implement a statistical test of the assumptions provided in the previous step. We defined a function where given a window of spectrum counts of width 0.2 KeV, we test the hypothesis that one of the two PIXL detectors contains a statistically significant response peak above the spectrum's noise threshold while the other detector does not. This is done using a paired difference t-test [189] between the two detectors pairwise over each channel in the window. This is is used as the counts in each channel are not independent since the underlying count for each channel is dependent on the X-ray frequency of that channel. We calculate and return the absolute value (since it does not matter which of the two detectors is the one where the diffraction is detected) of the t-statistic for the window as the measure of the statistical effect of potential diffraction.

While in the PIXL case the heuristic model took the form of hypothesis testing based on the assumptions provided by the scientists, there are many possible methods of formulating the heuristic. If the given scientific description is more easily expressed as Baysean priors then models (including deep learning based models) that utilize such probabilistic formulation may be useful. Alternatively if direct simulation techniques for particular anomalies are provided then direct similarity comparisons based on forward process simulation may be a better fit. The role of the heuristic model in the ISHMAP framework is not to enforce a single modeling paradigm for all phenomena, but rather to provide an interface for different models to work in ensemble within the larger framework. The important components are

that models must be chosen to have the most appropriate phenomenon-scale inputs, have comparable scalar heuristic outputs, and provide some way of parameterization based on scientist priors with possibility of fine tuning. If all of these requirements are met then the heuristic can be calculated for each phenomenon-scale datum in the dataset and these pairs of data and heuristic value can be utilized in the next step (Fig. Figure 5.2C).

5.4.3 Heuristic Model Evaluation from Sampling High Response Archetypes

Given a version of the heuristic model the next step in ISHMAP is to evaluate the heuristic and subsequently determine the kind of iteration for model refinement (Fig. Figure 5.2C). This is done first by sampling data from the high end of the distribution of heuristic responses. If the heuristic model is performing well these high response samples should form strong archetypes of the phenomena to be modeled. In this phase of the process scientists should be given an opportunity to inspect the class of model archetypes and determine the coherence of the class. If the high response samples are consistently determined to be good examples of the desired phenomenon then the heuristic model can be confirmed and moved on to the next threshold tuning phase. Otherwise the set of high response samples will contain instances of phenomena other than the target.

In this case scientists must then determine what else is being included. If the set contains a meaningful amount of instances of a distinct anomalous phenomenon then we can recursively iterate the ISHMAP procedure to model this other class and thus form a differentiation. This is exactly what occurred when evaluating the initial diffraction heuristic. During this phase R2 pointed out:

> " And, okay, yeah, this is something actually that I think is good for me to point out to you, because the algorithm identifies this as a diffraction. And this is not diffraction. This is actually, I would say, an intensity mismatch that we're seeing these not just in that peak. But in some other locations, other peaks in the spectra, I think there's a little bit of intensity mismatch has to do with measuring the rough surface, like

measuring like these larger grains. ”

What we had found was an additional class of anomalous phenomena due to *surface roughness*. This phenomenon was then subsequently modeled using ISHMAP, where the scientists' background knowledge informed us that surface roughness effects are frequency independent and thus the roughness phenomenon-space included whole spectra, and that the effect is expected to be a constant attenuation of the signal in a single detector. This background informed the heuristic roughness detector of calculating the mean detector difference across the whole frequency range of a spectrum, being essentially the maximum likelihood estimate of an assumed constant attenuation factor. This heuristic proved to be highly effective in the first iteration at distinguishing roughness effects. Upon completion of the deeper level of ISHMAP iteration with an effective roughness detector heuristic we could then separate out diffraction from roughness effects in the high response region leaving a now coherent class of diffraction instances using the original diffraction heuristic. Note that the phenomenon scale for roughness is not the same as that for diffraction, an example of how two different phenomena can have overlapping effects at one scale but can be easily differentiated at another scale. The recursive iteration and phenomenon centric structure of ISHMAP ensures that all heuristics model phenomena at their native scale while still being able to provide information between scales.

After differentiating all distinct classes of anomalies present in the high response sample set the primary heuristic model can be directly iterated to optimize differentiation with non-anomalous normal data and associated background noise. This iteration loop is flexible to the amount of scientist labeling available, as the baseline modeling assumptions assure a certain minimum semantic coherence of the initial heuristic making additional optimization optional while the flexibility of the heuristic modeling interface allow for models that can benefit from additional feedback when available. Thus the expected availability of downstream scientist labeling is an important constraint to consider when choosing a heuristic model class. This phase of ISHMAP is considered completed and we can con-

tinue to the next phase (Fig. Figure 5.2D) after sufficient iteration to ensure the the high response samples of the heuristic model consistently represent archetypal instances of the target phenomenon.

5.4.4 Heuristic Threshold Tuning from Sampling Moderate Response Edge Cases

Once the heuristic model has been determined to be responding to the correct features there remains the task of determining a threshold for the heuristic response value for the purpose of categorization. This process is very similar to the previous phase where samples with heuristic values around a potential threshold are generated and then evaluated by scientists. If the samples remain highly coherent then the threshold can be lowered, while if there are insufficient numbers of the correct anomalies in a set of samples the threshold can be raised. This process can be iterated and tuned depending on the comparative importance placed by scientists to false positives and false negatives. This phase (Fig. Figure 5.2D) still contains the same opportunity for iteration present in the high response phase (Fig. Figure 5.2C), where if additional phenomena are found in the boundary region they can be recursively modeled providing cleaner edge case regions. Furthermore in this phase we additionally consider a special class of phenomenon we call ambiguities. In the region of a decision boundary for phenomena detection, scientists will often find instances which are ambiguous in certain respects making an underlying label difficult or impossible to assign. These ambiguous instances should not be confused with well defined samples whose heuristic values happen to be near the decision boundary. Rather, ambiguities refer to examples where even the scientific ground truth may not be extremely clear. These phenomena introduce challenges for any classification scheme. In ISHMAP they are addressed by simply modeling ambiguity as a distinct phenomenon, where even if scientific priors will be by definition less robust, the features that scientists use to justify differences in interpretation are used as the basis for the heuristic modeling. Such features must necessarily exist as otherwise scientists would not have any empirical basis for uncertainty. This additional it-

erative phase is repeated until either no more ambiguous instances are found in which case a threshold can be determined exactly based on error tolerances, or is limited by scientist availability as often the ambiguous class may contain infinitely deep amount of different and idiosyncratic features that could be modeled.

When sampling edge cases of diffraction scientists found two different ambiguous phenomena. The first were cases of single detector response where the potentially diffracting window was had a significant difference in detectors, but the difference was not strongly peaked as the prior expectation would be for a diffraction response and was instead more flat. This was then modeled by measuring the relative height of the detector difference peak in the center of the window to differentiate strongly vs weakly peaking instances. The second class of ambiguous phenomena were cases where instead of a consistent background in the non-diffracting detector, the baseline detector (detector with lower average count over the window) contained a small peak as well but attenuated compared to the other detector. This introduced different interpretations by different scientists and thus was modeled as ambiguous using a heuristic of looking at the coefficient of variation of the baseline detector in a window. After differentiating these primary classes of ambiguous phenomena the boundary region of the diffraction heuristic threshold became sufficiently clean to determine a threshold to classify diffraction, in this case with a bias placed on reducing false positive determinations.

Once all interfering phenomena have been modeled and differentiated, and a classification threshold is determined, the final output of ISHMAP is a deployable classification model (Fig. Figure 5.2E). This model has the same fundamental classification structure as a decision tree, shown in figure Figure 5.3, since the aforementioned phenomena differentiation is powered by similar recursive iterations of ISHMAP returned classifiers.

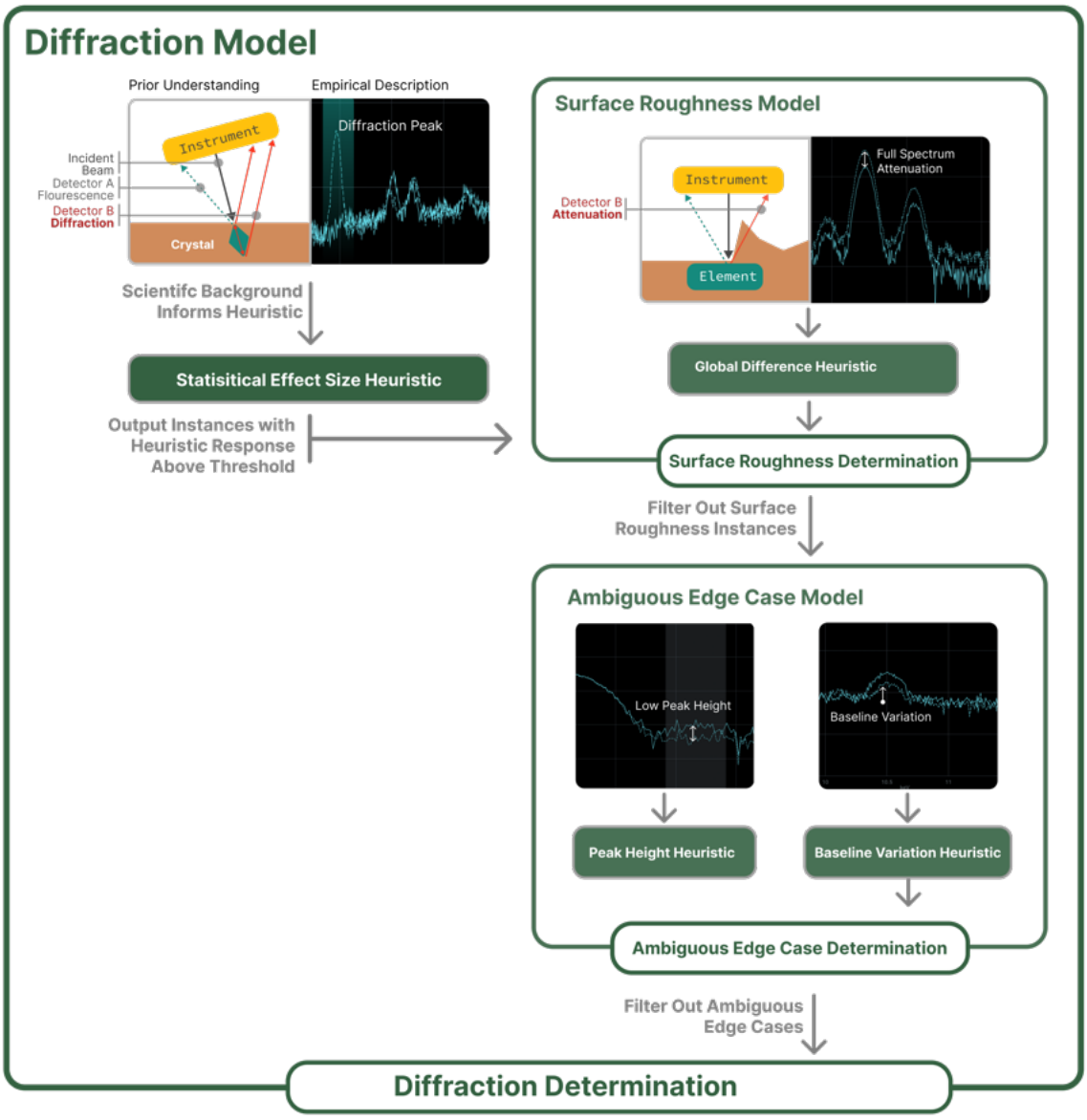

Figure 5.3: The architecture of the output model (Fig. Figure 5.2E) of utilizing ISHMAP to detect diffraction peaks.

5.5 Evaluation

In order to evaluate our approach we both conducted a standard quantitative error analysis for the model component of the work (Section Subsection 5.5.1), as well as a qualatative evaluation of the success of the final deployed system (Section Subsection 5.5.2). In doing so we do not aim to show that ISHMAP as a design framework is verifiably superior to other design approaches (as this would require counterfactual information of the success of many different design frameworks on this same problem which is well outside of the scope of this work). Instead we merely wish to show how ISHMAP *is capable* of producing strong scientific outcomes in **this** application, and thus *may* provide useful in further applications.

5.5.1 Evaluation of Anomaly Detection Model

After completing the ISHMAP procedure for the target phenomenon of diffraction, we are left with a classification model (Fig. Figure 5.2E / Fig. Figure 5.3). In order to evaluate the real world accuracy of the model within the context of its use in mineral identification, and to avoid information leakage and test our model's generalizability, after developing our diffraction model using input from R2 and R3, we set to test the model using labels from additional scientists (R1, R4, and R5) on different datasets. Multiple different scientists were used in order to ensure reliable labels. We presented the three different scientists a representative random sample of 213 spectra that were not used in the training process. The sample was balanced to include 107 spectra uniformly randomly sampled among spectra determined by the model as containing diffraction peaks, and 106 sampled uniformly from spectra determined as not containing diffraction (through either exhibiting Surface Roughness, Ambiguity, or no anomaly). The scientists were then left to determine if any of the presented spectra contained diffraction. While all three scientists were presented with all 213 spectra, some spent considerably less time providing only a few labels on the presented spectra due to their time constraints. Furthermore we also found that the individual scien-

tists had varying sensitivities for positively determining diffraction in cases of uncertainty and thus would often disagree on their determinations. Therefore we could not form reliable ground truth labels for 16 spectra which had only two labels from different scientists who disagreed, as well as for 45 spectra with only a single label from an individual scientist (which as we found is an unreliable indicator without a second opinion). This left 152 spectra (of which 144 had a total consensus and 8 had a majority determination) which we were able to use as a basis for evaluation. Of this reliable ground truth set, the model correctly predicted the presence or absence of diffraction with **93.4%** accuracy.

These results match the qualitative experience that scientists expressed when examining the outputs of the model, as in an interview with R2 they expressed that:

> "Your tool works very well, and is finding [diffraction peaks] in almost all cases, ... And so there wasn't really anything that was being identified incorrectly."

This qualitative reliability and satisfaction from the perspective of scientist domain expert end users forms the most relevant evaluation of the efficacy of the tool when comparing to previous methods which do not present any systematic alternative baseline, and thus a strong example of the effectiveness of the ISHMAP Framework for designing useful models for scientific users.

5.5.2 Impact of Deployed Interface

An important benefit of the ISHMAP collaborative design process is that once a model detecting a particular anomalous phenomena is complete, there is already assurance that the model is answering an actual problem of interest for scientists and collaborating scientists have a built in degree of ownership and buy-in to the technique[190]. This means that practical deployments of tools that utilize such a model are much more likely to result in adoption into the scientific workflow. To showcase how this collaborative design and scientific interpretation-first modeling approach can not only perform well with regards to general classification benchmarks but additionally result in meaningful new capabilities

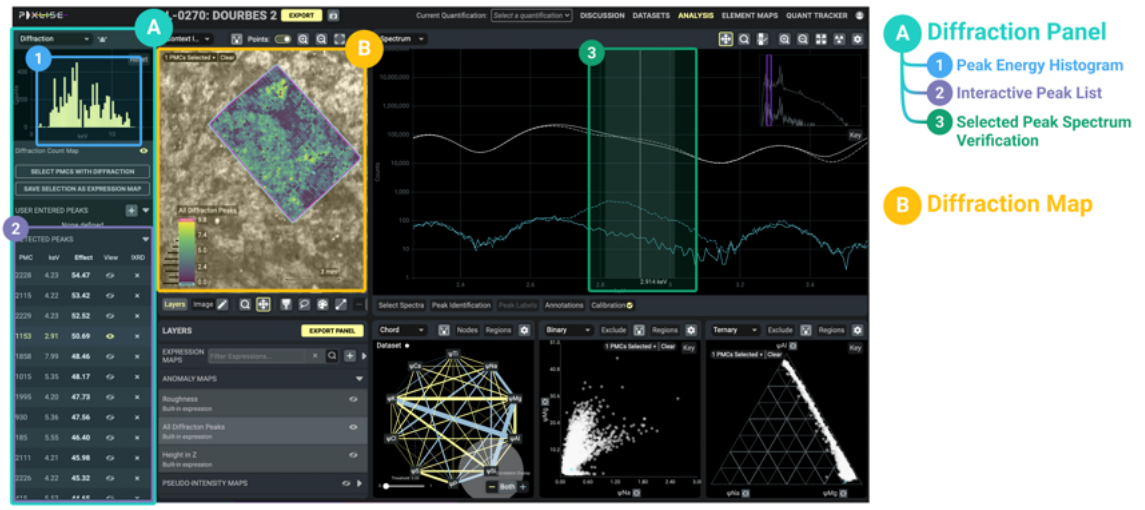

Figure 5.4: Overview of interface components within PIXLISE Application displaying the Guillaumes Mars dataset. (A) The **Diffraction Panel** enables quick identification and verification of individual diffraction peaks or grouped similar peaks. (B) The **Diffraction Map** displays the spatial distribution of diffraction peaks either over the whole spectrum range, for particular energy subsets, or in combination with custom defined expressions from other PIXLISE analysis tools.

for scientific end users we can consider the deployment of our ISHMAP diffraction model within the PIXL science workflow. After finalizing the model we integrated the model outputs with the existing primary visual analytics toolkit used by PIXL science: PIXLISE [191, 192, 193].

The first full version of the model and associated visualizations was deployed in November 2021, with preliminary versions available to scientists as early as June 2021. At the time of writing, it is used daily by over 97 NASA scientists and NASA-affiliated scientists across the globe collaborating on the PIXL science mission. Within the PIXL science group discussion board the functionality of our tool was mentioned over 80 times in the context of working group discussions, with 39 instances highlighting diffraction informed geo-science interpretations that would not have otherwise been easily visible and 28 instances of corrected spectroscopy and quantification error detection. Now, whenever new data is beamed down from Mars, our model allows scientists to instantaneously discover diffraction peaks that can inform downstream mineral identification.

5.5.3 Diffraction Panel Interface

Immediately once a dataset is loaded into PIXLISE, scientists can use the *diffraction panel* (Fig. Figure 5.4A) to identify particular diffraction peaks. The diffraction panel includes a histogram of all of the energy levels where diffraction peaks have been detected. The scientists can choose a set of energy ranges in the histogram to in turn select all locations which contain diffraction peaks in those ranges (Fig. Figure 5.4A.1). This allows scientists to quickly discover where different kinds of diffraction peaks, and thus minerals, are located (as the diffraction energy is a direct function of the crystal structure of the underlying mineral), as can be seen in figure Figure 5.5. This is a particularly valuable feature for scientific interpretation enabled by the fact that our diffraction model works at the correct diffraction scale framing codified by ISHMAP as opposed to the default framing of anomalous spectra which a machine learning based model would utilize. Scientists can then further verify individual peak identifications from a sortable list of the detected diffraction peaks (Fig. Figure 5.4A.2). Scientists can select a peak and then see its corresponding spectrum and energy location on the PIXLISE spectrum view plot (Fig. Figure 5.4A.3), and then confirm if this classification is a correct determination of diffraction or a false positive within the interface. This is required as even a very accurate model contains errors and allowing users to further refine the available data allows the model to be updated and improved continuously, as well as helping build trust with the scientists by ensuring that they always have the ability to override the interpretation of the model.

5.5.4 Diffraction Map Visualization

In addition to the peak-specific workflow enabled by the diffraction panel, we have also implemented a more high-level visualization of anomaly structure via the *diffraction map* interface (Fig. Figure 5.4B). Within the PIXL science team, diffraction maps have become an accessible, shareable, and invaluable piece of information for the process of mineral identification as new data comes down from PIXL.

The diffraction map visualization overlays a heatmap of the density of diffraction peaks or surface roughness anomalies at each beam location on top of the visual context image. This allows scientists to quickly find clusters of diffraction that are indicative of a crystal grain, group the locations within that cluster, and create regions of interest that can be applied towards further analysis. The diffraction map has become the preferred method of scientists to share findings of crystal structure, as R3 commented when looking at the diffraction map for the Beaujeu [194] dataset:

> "I found an interesting correlation between the regions marked with a high number of diffraction peaks, and the regions that we have geochemically identified as plagioclase... It is great to see the usefulness of the diffraction peak detection algorithm in practice."

These maps can be customized in a number of ways. The default view when starting is to show the density of diffraction present at all energy levels, this is the most broadly applicable when lacking a particularly strong prior about the specific crystallography of the target. However if a scientist has a hypothesis about a particular crystal configuration which would predict diffraction at predictable frequencies there are two ways to visualise diffraction with more specificity. PIXLISE contains a custom domain-specific language for custom maps of expressions. The output of the diffraction model is a supported query within this language and allows scientists to integrate anomaly information with other existing analysis. Additionally, the diffraction panel histogram selection supports the creation of maps directly. By being able to see the distribution of diffraction peak energy, scientists can analyze the distribution and find clusters without a-priori knowledge, creating maps that rather than showing the overall crystallographic structure of the sample can isolate particulate grains (Fig.. Figure 5.5).

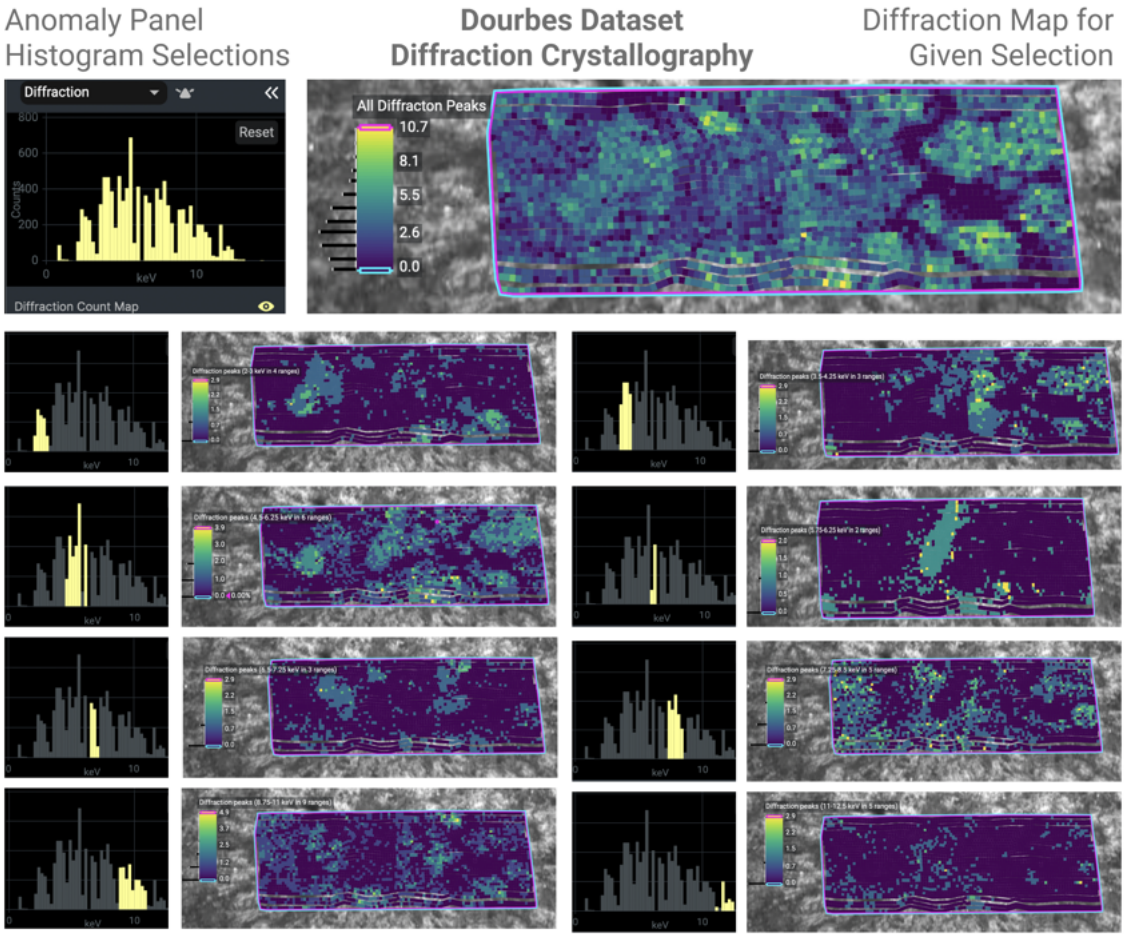

Figure 5.5: Screenshots of diffraction maps of the Dourbes[195] dataset formed from different selections in the diffraction panel of diffraction peak energies. What can be see is rich information regarding the spatial distribution of diffraction peaks at different energy levels, with peaks in close energy clusters also clustering spatially. This implies the presence of unique crystal grains which can be readily seen using these diffraction maps.

5.5.5 Enabling Ongoing Scientific Discoveries about Martian Geology

In November of 2021, the PIXL conducted a series of XRF scans of a sample with the codename **Dourbes** [195] at the Séítah formation [196] in the floor of the Jezero crater on Mars. This location presented an acute issue for the problem of mineral identification. Due to weathering, it is impossible to clearly identify crystal grains from the context imaging, and XRF information cannot sufficiently differentiate between all relevant physical properties. This information is extremely important to make inferences about the geological history of the site and has formed a significant challenge to previous Mars missions in similar situations.

Fortunately, due to the additional data collected by the PIXL instrument and the development of our suite of tools scientists were rapidly able to visualize the crystal structure of the sample with the diffraction map (Fig. Figure 5.5). The diffraction map functionality enabled robust spatial comparison of diffraction with elemental analysis, and thus scientists could make strong claims about particular grains of elements, their crystal properties, and thus identified mineralogy. By going beyond the information available in standard fluorescence and elemental quantification, this comparison provided decisive evidence about the mineralogy of Séítah formation rocks, as expressed by a PIXL scientist in a recent paper published at a top-tier scientific journal[26]:

> "Collocated crystal sizes and mineral identities are critical for interpreting textural relationships in rocks and testing geological hypotheses, but it has been previously impossible to unambiguously constrain these properties ... Here we demonstrate that diffracted and fluoresced x-rays detected by the PIXL instrument ... provide information about the presence or absence of coherent crystalline domains in various minerals."

This finding has formed the central component of continuing research within the PIXL science team and is functionally enabled by the diffraction detection and visualization capa-

bilities powered by our model. The effectiveness of the model in integrating with existing scientific workflows and assisting in high impact analysis immediately upon deployment further provides strong evidence of the effectiveness of the ISHMAP design framework.

5.6 Discussion

5.6.1 Generalizability of Design Goals

While this work has showcased a single specific successful application, we hope that there are a number of useful insights from our solution that that can be utilized in other applications as well. In order to evaluate the applicability of our framework for other use cases the key deciding factors should be through the alignment of our outlined design goals. While we developed these goals strictly within the context of embedded user research for the particular domain of PIXL scientists, we justified the formulation of each goal based on aspects of scientists' workflows that we found to be common among the different individuals and specialities within the fairly diverse PIXL team as well as on aspects that, while exhibited in this specific workflow, are not necessarily unique to it. For instance our design goal Section DG1 is based on the observation that anomalies are more easily missed the further a dataset is from the 'native' scale of the anomalous phenomenon. Since essentially by definition anomalies are classes of phenomena that default assumptions do not apply to, any processing steps are likely to introduce errors with respect to anomaly detection. We can then say that a focus on raw data is not just an important design consideration for PIXL, but for any analytic workflow that includes steps that processes data in an irreversible manner based on violable assumptions. Our design goal Section DG2 essentially formalizes an extremely common design constraint of limited data and label availability. The entire branch of unsupervised machine learning studies various implications for modeling with such a restriction. So while discovering that this constraint was applicable to our specific domain is extremely important, it is also clear that there are many other domains that share a similar requirement. Finally our design goal Section DG3 similarly expresses a well es-

tablished and known weakness of deep learning based anomaly detection[169] and shows why it is important in our use specific case. What this all implies is that while the goals we developed are in one respect specifically tied to the PIXL science domain, we expect that there are very likely a large number of other domains who share these goals as well, and it is in these domains where ISHMAP may be a useful tool to structure model development.

5.6.2 Limitations of ISHMAP

While we present the substantial potential effectiveness of the ISHMAP framework, like all frameworks it is not universally applicable and has limitations to where is should be utilized. Of course the primary drivers of whether ISHMAP is appropriate is whether the design goals laid out are of relevance. In prioritizing these goals other potential priorities are de-emphasized. In particular the ISHAMP process can only help in discovering known anomaly classes, as well as anomaly classes that are discovered during the ISHMAP process. There are many domains, including scientific domains, where data sources are sufficiently novel to contain many potential 'unknown unknown' classes of anomaly that users do not have a prior expectation for. Since ISHMAP relies on explicit description of known phenomena, such unknown anomalies cannot be captured reliably. In such applications it has been shown that more pure deep learning based methods can be highly effective in discovering such unique or point anomalies [169].

Furthermore the collaborative process may have substantial organizational overhead due to the requirement for consistent iteration and feedback between developers and scientists. Depending on the nature of a collaboration this overhead may occasionally present a greater manual effort burden than the effort of just providing more ground truth labels, undermining design goal Section DG2. Thus when making the choice of whether to undertake an ISHMAP collaboration both sides must consider both the technical and organizational nature of the problem to determine if it is the right fit.

5.6.3 Opportunities for Future Work

In presenting the ISHMAP framework we have only taken a first step in improving scientific anomaly detection with human-centered-AI methodology. While the results of our implemented detection tool with the PIXL team was a clear success, it only formed a proof-of-concept for the methodology. We encourage researchers to replicate, evaluate, and refine the methodology in additional domains. Additionally the framework itself presents clear opportunities for development. The flexibility of the framework makes it amenable to many different forms of utilization and deployment, and so future work to fine tune the most effective processes both technical and procedural for scientific prior encoding, heuristic generation, and sample evaluation may greatly assist in more efficient and effective utilization of ISHMAP. Indeed, while ISHMAP was developed in order to address the shortcomings of pure deep learning based approaches, studying how to integrate deep learning models and all their expressive power within this more interpretation-focused framework may allow for the best of both methods. Finally, for the current formulation of ISHMAP, while the expertise of scientists is absolutely essential, the role of the model designer/developer is comparatively procedural. This leaves a potential opportunity to develop automated tools or interactive interfaces for scientists to engage in their side of the ISHMAP procedure entirely independently, which would massively increase the potential for science teams to develop their own robust and interpretable anomaly detection models.

5.6.4 Reproducibility

The code for discussed in this work is distributed across a number of repositories in the broader PIXLISE project that is open-sourcing all of its constituent repositories at https://github.com/pixlise[197, 198]. Additionally the deployed PIXLISE tool itself is publicly accessible. Anyone can request an account at https://www.pixlise.org/ and use all of the functionality of PIXLISE, including the anomaly detection functionality discussed in this work, on any public datasets. Most of the datasets used as examples in this work are

publicly available either on PIXLISE or in raw data form directly from the NASA Planetary Data System (PDS) at https://pds-geosciences.wustl.edu/m2020/urn-nasa-pds-mars2020_pixl/.

5.7 Conclusion

Many of history's most important scientific discoveries can be attributed to the thorough analysis of anomalies[35]. Today, as scientific datasets get larger and more complex, so too do the methods used to find the anomalies within them, sometimes at the expense of the ability to interpret and explain the anomalies[169] which is a fundamental component of scientific analysis. In this work we sought to integrate human-computer interaction methodologies with the state of the art in AI to develop a method for anomaly detection that is both effective and interpretable. In collaboration with a world leading science team at NASA, we conducted extensive user research to understand their specific analytic workflow, and developing design goals that represent the needs of scientists with respect to interpretable anomaly detection. Based on these design goals we introduced ISHMAP, a novel design framework for the development of scientific anomaly detection models. By utilizing ISHMAP to develop an anomaly detection toolkit used daily by NASA scientists around the world and contributing to ongoing scientific discoveries, we showcased a proof of concept for a method to enable better science by taking a human-centered approach to both the technical and scientific problems of anomaly detection which we hope can assist scientists and researchers looking to not only detect, but understand anomalies in their data.

Part II

Interpretable ML for Exploratory Science

Overview of Part II

In Part I we developed frameworks and principles for the human-centered design of machine learning tools in scientific contexts. In particular Part I culminated in the development of the ISHMAP discovery framework (Chapter 5), which is a method to structure collaboration and design of intelligent systems specifically in the process of anomaly detection and phenomena conceptualization as described in Section 2.2. This framework specifically focuses on the aspects of phenomena conceptualization in the 'context of discovery' as opposed to the 'context of justification'[199], as different tools both mathematical and statistical are emphasized differently when either trying turn 'known unknowns' into 'knowns' (analysis/justification) as opposed to turning 'unknown unknowns' into 'known unknowns' (exploration/discovery). Part II will focus on identifying further scientific exploratory workflows with the relevant properties aligned with the ISHMAP discovery framework and utilizing this meta-methodological framework to develop new interpretable machine learning methods.

An essential component of this thesis is that purely abstracted knowledge, even when formed on the basis of concrete user research, can only be fully validated through successful application in further concrete contexts. Therefore in this part, I will apply the insights and frameworks from Part I toward additional specific scientific problems and show how through utilizing these methods we can develop novel interpretable machine learning techniques that better address scientific problems. Much like Part I, the chapters in this part are slightly edited versions of scholarship already published in top peer reviewed data science [25] and bioinformatics [24] venues, and as such will primarily focus on the more specific domain problems each new method solves.

In Chapter 6 I present an extension of my initial collaboration in Chapter 5 with NASA JPL scientists where I developed the ISHMAP design framework. This foray showcased the importance of an iterative, scientist driven, approach to exploratory data analysis. Fur-

thermore a key aspect of the ISHMAP framework involved determining the appropriate measurement scale of scientific phenomena, and relies on an ability to visualize such phenomena that may span multiple modalities and scales. If we wish to empower scientists to discover underlying phenomena in multi-scale measurement datasets through ISHMAP, they must be able to visualize the underlying joint distribution at the correct scale. Therefore in Chapter 6 I propose a new technique, NestedFusion, which can perform latent dimensionality reduction on arbitrarily nested multi-scale datasets to allow scientists to effectively visualize the joint patterns in underlying scientific phenomena that span multiple dataset scales and modalities at the highest possible resolution, enabling better integration with frameworks such as ISHMAP.

In Chapter 7 I showcase an important generalization of the design frameworks and modeling techniques developed thus far into an entirely different domain, showcasing the flexibility and generalizability of the approach. In collaboration with epidemiologists at the CDC, I apply frameworks from Part I by developing an interactive technique for public health researchers to conceptualize and query emerging trends in language use patterns in social media data in order for detection of emerging public health threats. This method is validated against the known pattern of the emergence of synthetic opioids such as fentanyl in the opioid epidemic, where this method is able to detect the emerging lethality of fentanyl 1-2 years before it was detected by traditional methods.

CHAPTER 6

NESTED FUSION : A METHOD FOR MULTI-SCALE LATENT DIMENSIONALITY REDUCTION VISUALIZATION AT HIGH RESOLUTION

6.1 Introduction

In scientific data analysis the initial exploratory phase of visualizing and conceptualizing the relevant empirical phenomena in a dataset is both an essential aspect for effective work and comparatively under studied in the context of scientific applications, where skipping such inductive explorations in favor of immediately utilizing known models for analysis is the de facto standard (See Section 2.2). However, in Chapter 5 we showed how unanticipated or anomalous phenomena can often mislead such analysis, motivating a workflow that at least starts with purely empirical exploration of data in the initial phases of work after making measurements in order to have a more informed prior of the distribution of actual phenomena within a dataset before applying the more rigorous scientific models to ensure the chosen models are appropriate. While common data-centric techniques of exploratory analysis such as dimensionality reduction visualization have proven to be very effective in many domains of scientific inquiry [68, 69, 70, 71, 72, 73, 74, 78, 79, 80, 81], in domains with multiple measurement apparatuses of different resolutions and scales, existing techniques can fail to model some of the phenomena we wish to discover. This is because the standard formalization for dimensionality reduction techniques is that of a single dataset of measurements of identical shape which corresponds one to one with the set of objects and patterns between objects that the analysis aims to visualize. However it is often the case that underlying phenomena are differentiated at levels that do not align with the resolutions of measurement each apparatus perfectly [23]. Rather, there may be multiple methods of measurement which each elucidate different aspects of an underlying

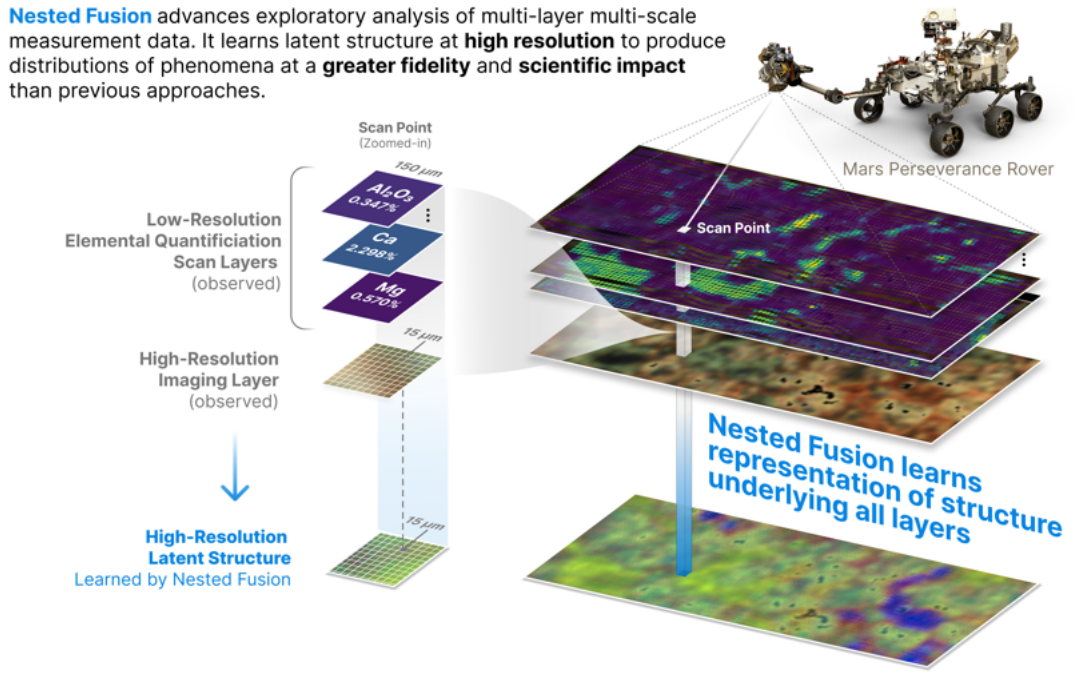

Figure 6.1: Our method, Nested Fusion, radically accelerates the exploratory analysis of Nested Measurement Datasets by learning the latent structure at high resolution to produce distributions of phenomena at a greater fidelity and scientific impactfulness than previous approaches. In the figure the DOURBES target location is shown, out of over a hundred locations on Mars scanned by the Perseverance Rover.

structure but which all have varying resolution scales and thus are sensitive to the different properties of various aggregations of the structure.

One such domain where scientists require more powerful exploratory analysis tools is the work done by the PIXL Science team with the Mars Perseverance Rover at NASA (National Aeronautics and Space Administration). In service of the high-level goal of searching for signs of a history of life on Mars, scientists are interested in the fine-grain mineral structure of *target* locations on the Martian surface [200]. The Perseverance Rover contains two (among many) scientific instruments to assist in this task: the Planetary Instrument for X-ray Lithochemistry (PIXL) instrument [201], which includes an X-ray fluorescence (XRF) spectrometer, and a Micro-Context Camera (MCC) for multi-spectral imaging. When observing a specific target location of geological interest, the rover will use both of these instruments to conduct two co-aligned scans as shown in Figure Figure 6.1. While both of the instruments scan over the same physical location, their resolutions are much different, where for each scan point, a single XRF spectrum corresponds to a larger patch of approximately 100 MCC imaging pixels. At the same time, each instrument elucidates different aspects of the underlying mineralogy of the target. While the spatial precision of each MCC pixel corresponds much more closely to individual homogeneous mineral grains, it lacks a nuanced depth of information to accurately differentiate minerals based on chemistry. On the other hand, each XRF spectrum produces a detailed quantified distribution of the chemical composition of the scan point, but the larger diameter of this point may encompass multiple grains of different minerals thus producing an aggregate chemical distribution. The ultimate scientific question is about understanding the distribution of underlying minerals. While both measurements offer extremely powerful signals concerning this distribution, neither alone encompasses all the possible information to explore, leading to the need for modeling these different measurement scales together. To tackle these significant scientific challenges, we present the following major contributions:

1. **A novel problem formulation** tailored to exploratory analysis of nested measurement

datasets, which consist of irregularly overlapping measurements of multiple scales (Sec Subsection 6.3.1). This formulation is rooted in addressing the practical needs of PIXL scientists at NASA who analyze XRF and MCC data collected by the Mars Perseverance Rover.

2. **The Nested Fusion algorithm**, a new model for latent analysis and dimensionality reduction for nested measurement datasets (Sec Subsection 6.3.2), This method is significantly more effective than alternatives, yielding latent encodings at a resolution far higher than what existing dimensionality reduction techniques can achieve. We evaluate the effectiveness of Nested Fusion both qualitatively within the context of initial data exploration and quantitatively in data reconstruction fidelity. Nested Fusion outperforms the state of the art in dimensionality reduction for nested measurement datasets, providing more interpretable and practically useful results (Sec Section 6.4).

3. **Deployment of Nested Fusion** in scientific practice within the PIXL team for the Mars Perseverance Rover, enabling scientifically meaningful visual interpretation and efficient discovery of cross-modal patterns (Sec Section 6.5). We analyze how Nested Fusion is utilized in practice and how it fits within the scientists' existing analytic workflows. To ensure reproducibility of our technique and findings, we have open-sourced it at https://github.com/pixlise/NestedFusion

6.2 Background and Related Work

In this section, we introduce the scientific problem statement and dataset overview from the PIXL instrument on the Mars Perseverance Rover, define our formalization of nested measurement datasets, and go over related work in scientific exploratory data analysis and dimensionality reduction.

6.2.1 Mars Perseverance PIXL Data

The PIXL instrument aims to measure the mineral structure of small rock samples (called *targets*) on the surface of Mars contributing toward the larger inquiry towards any potential evidence of a history of life on Mars. For each individual target on the martian surface multiple scans are taken. First is the MCC Multi-spectral imaging camera, which takes a series of four images illuminated by specific wavelengths of near-visible light: Near-Infrared (NIR), Green, Blue, and Ultraviolet (UV). This produces a single color image for each target with 4 primary channels, as opposed to the standard 3 channel RGB, and is often analysed using the 16 distinct ratios between them. Each image will contain on average about 500,000 of these 16 channel pixels, spanning a region of approximately 100 square centimeters with each pixel corresponding to a resolution of approximately 15 microns. At roughly the same time a scan is taken of the same target with the PIXL instrument for X-Ray spectroscopy. This instrument produces much more detailed quantitative data, consisting of a grid of X-Ray fluorescence spectra which are quantified to represent the distribution of elemental weight percentages at each scan point, we call this distribution a *quantification*. Each scan can consist of between 1000 and 10,000 individual spectra (depending on the particular shape of the target) covering a smaller region of approximately 30 square millimeters. Each *scan point* is measured with a beam diameter of 50-200 microns[1], thus corresponding to a region covering approximately 100 MCC pixels as shown in Figure Figure 6.1.

Thus far, at the time of writing, during the time that the Perseverance Rover has been in operation, there have been 103 target locations scanned producing a total of 295,602 52-dimensional (the number of unique elements included in all quantifications) quantified spectra, as well as 26,966,169 MCC pixels. However, not all scans include both data types

[1]This beam diameter is energy dependent and since there is a nonlinear transformation between the energy levels of the spectrum and the final quantified elemental distribution where each element is quantified using the full energy range, we treat the upper range of the beam diameters as representing the region encompassed in a quantified scan point.

and so for this work focusing on combining information from both measurements, we are restricting to a total of 103,005 scan points which each contain a single quantification as well as 100 corresponding MCC pixels.

6.2.2 Related Work

Previous work in collaboration with PIXL scientists has shown how data science techniques can form an essential component of their scientific workflow by focusing specifically on modeling anomalies and visualizing distinct empirical phenomena[23]. This work focuses on the problem of initial visualization and thus on dimensionality reduction as an effective technique for enabling such visualization for the high dimensional PIXL data.

Dimensionality reduction techniques such as UMAP [63], T-SNE [64], MDS [65], Isomap [66], and the most commonly used PCA [67] are fairly ubiquitous in a variety of scientific domains [68, 69, 70, 71, 72], and even specifically XRF spectroscopy [73], as well as Mars multi-spectral imaging[74].

Another conceptualization that can produce comparable visualizations is the approach of latent analysis which takes a more, generally Bayesian, probabilistic framework to the problem of learning low dimensional representations. These approaches mostly stem from the development of variational autoencoders (VAE) [75], and different latent models have been introduced to handle many scientific problems [78, 79, 80] including planetary science[81] among many other domains.

6.3 Proposed Method: Nested Fusion

Grounded in understanding from previous work with PIXL scientists [23] our aim is to develop a method for visualizing and determining the distribution of mineral phenomena within each PIXL target, and to assist in their identification based on their relationship between the past history of targets. Focusing on targets where both XRF and MCC data are present and overlapping, we hope to enable work to discover new patterns that each individ-

Symbol / Term	Meaning
Nested Measurement Dataset (M)	A class of dataset which combines multiple kinds measurements that cover a common area
Data Scale (X)	The set of measurements of a particular kind that define a data layer of a specific resolution
Nested Scale (X_S)	A scale at a higher resolution which has a correspondence where multiple measurements in the nested scale correspond to a single measurement in the lower resolution scale
Nesting Function (η)	A function which maps a specific data point at a scale to the set of data points in the corresponding nested scale that cover the same physical space.
Maximum Resolution Latent Scale ($X_\varnothing$)	The scale for which no further nested scales exist, defining the highest resolution available in the dataset and thus the resolution at which latent structure can be modeled
Latent Base Scale Correspondence (β)	A function which maps a specific data point at any scale to the set of data points at the maximum resolution latent scale that cover the same space as defined by repeated nesting.
$x_i \in X$	A specific data point at some data scale X
$x_i^\varnothing \in X_\varnothing$	A specific data point at the maximum resolution latent scale
z_i	The latent encoding corresponding to $x_i^\varnothing$

Table 6.1: Notations and terminology used in this paper

ual instrument cannot differentiate independently. While scientific interpretation is the end goal, the specific interpretations (i.e., "we see a grain of olivine here or a potential aqueous intrusion there") enabled by the method are out of the scope of this work. Therefore we introduce a precise formalized problem statement which aims to properly encode the scientific priors and goals of the problem with specific consideration to the non-standard mixed scale measurements present in PIXL data, while simultaneously laying the foundation for how such methods can be more easily generalized to new domains. Finally after introducing the problem formulation we will describe our proposed method, Nested Fusion, which looks to solve this problem.

6.3.1 Problem Formalization of Nested Measurements

As Figure Figure 6.1 shows the nested hierarchical structure of PIXL data is not immediately amenable to standard data science techniques barring some flattening operation which leads to over aggregation and loss of resolution (see *Joint Models* in Section Subsection 6.3.3). Thus we introduce a formalization of **nested measurement datasets** which we will use to model this structure and subsequently perform better analysis on the data in a more natural manner, while also outlining precisely the requirements that any other dataset must meet in order to utilize the methods introduced in Section Section 6.3 in other domains. Table Table 6.1 summarizes the notations and terminology introduced in this section and used throughout this paper.

We recursively define a nested measurement dataset M as consisting of a tuple of two components: $M = (X, S)$. The first component $X = \{x_i \in \mathbf{R}^d\}_{i=1}^N$ is simply a standard dataset of N independent and identically distributed samples of d dimensional data representing the particular measurements at some specific scale. Then S is what we define as the *nested scale*. The nested scale is a tuple (M', η) of another nested measurement dataset $M' = (X', S')$ as well as a *nesting function* ($\eta : X \to 2^{X'}$) which maps each data point in X to a set of corresponding data points in X' that cover the same underlying physical

and latent area. In order to terminate this regress there must be a final scale $X^{\varnothing}$ which has no further nested scale and thus is notated as $\varnothing$. Having no further nested scale means that $X^{\varnothing}$ is the highest resolution available in the nested measurement dataset, and so we refer to it as the *maximum resolution latent scale* since our aim to model latent structure at this maximum resolution.

The key assumption is that all of the information at lower resolution scales supervenes on latent information at the maximum resolution. That is, that there is some more basic structure underlying the dataset that is approximately modeled at the maximum resolution as an unobserved latent variable, where each sample $x_i^{\varnothing} \in X_{\varnothing}$ is generated from a random process involving the latent value z_i that has a prior probability distribution $p(z)$, producing some conditional distribution $p(x^{\varnothing}|z)$ that we aim to learn[2]. For all other scale samples with nesting function η we then define the β correspondence which returns the set of all latents at the base maximum resolution that correspond to a sample x_i :

$$\beta(x_i^{\varnothing}) := \{z_i\} \tag{6.1}$$

$$\beta(x_i) := \bigcup_{x'_j \in \eta(x_i)} \beta(x'_j) \tag{6.2}$$

The supervenience assumption then can be restated probabilistically that all lower resolution scale variables are generated from the conditional distributions $p(x|\beta(x))$, and thus are conditionally independent of measurements at any scale other than the maximum. This structure is outlined in the graphical model for the PIXL dataset in Figure Figure 6.3.

While seemingly fairly abstract and obscure, this underlying structure and supervenience assumptions of a nested measurement dataset is in fact pervasive in the sciences [202, 203]. The natural sciences in particular commonly share the physicalist reduction assumption (at least within a single domain), that any given composite object of study is

[2]Note for notation, we include indices for actual measurement samples, while not including indices when referring to the random variable that generates the samples.

fully reducible to the set of underlying physical objects of which it is composed [204, 205]. This assumption necessitates that if multiple kinds of measurement apparatus measure an overlapping subject in time and space, then there must be some correspondence relation between the two measurement modalities. Furthermore this assumption enables us to study the intersections between these different layers of composed abstraction, as each class of composite structure is often best observed using separate kinds of measurement that very often do not have perfectly aligned scale and resolution. More complex composite structures will tend to exhibit additional complexity and depth (note the high dimensionality of the PIXL quantified spectra) however at the expense of necessarily being more spatially diffuse. While higher resolution measurements may be possible at the expense of more limited depth.[3]

6.3.2 The Nested Fusion Algorithm

The previous section describes the formalized problem of learning latent maximum resolution scale variables from nested data. One important aspect to note when introducing our solution is that the formulation of the latent variables at this scale is itself already a modeling approximation. In reality we expect fundamental structures within a domain to exist at finer scales than are directly accessible, and so we simply use the highest resolution available in any given nested measurement dataset as a proxy scale for a 'true' latent z. What this lends support to is the use of variational inference as a method to efficiently learn approximate distributions of z, which is acceptable as we do not in general actually have strong enough priors about the structure and properties of a 'true' z to justify other methods which have significant computational and other drawbacks when compared to widespread empirical success of variational auto-encoding models. Therefore the approach taken in this work, **Nested Fusion**, is a variational auto-encoder model[75] structured to work on nested measurement datasets.

[3]Think of the accuracy vs precision distinction, where here generally increasing resolution increases spatial precision but makes it more difficult for each individual measurement to be accurate.

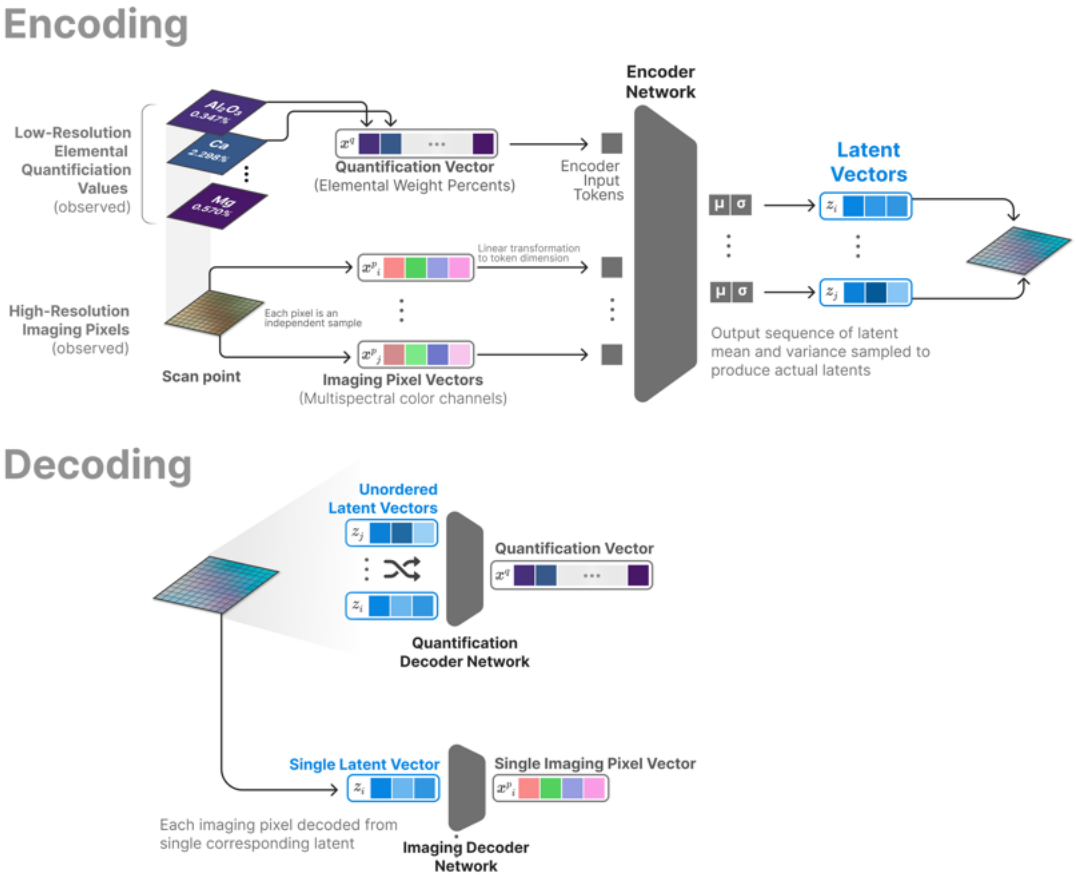

Figure 6.2: Model architecture and data processing pipeline for Nested Fusion as applied to PIXL data. High resolution latent vectors are encoded given a scan point containing an XRF quantification vector and collection of MCC imaging pixels.

Figure Figure 6.2 describes Nested Fusion's architecture. Without loss of generality, we explain how the framework is applied to the PIXL data/scenario presented in Figure Figure 6.1. Specifically we show how a scan point consisting of both low-resolution elemental quantification values and nested high-resolution imaging pixels, is jointly used to learn high-resolution latent vectors. The latents are learned though optimizing via stochastic variational inference[206] both encoder and decoder models to maximally reconstruct the original scan points[4]. The 1-, 2-, or 3-dimension latents would then be used for visualization by PIXL Scientists.

[4]In addition to the other components of the KL Divergence loss used in variational inference which integrates some priors on the latents as well

First, let us consider the encoder step. For the encoder model, which estimates the conditional latent distribution given the data $q(z|M)$, we must choose a class of distributions for the latent prior $p(z)$ and specify the relevant class of distributions for the data type of each measurement scale. We focus on a basic prior model of latents being standard normal ($z \sim \mathcal{N}(0, \mathbf{I})$) which we can use to compare to other methods of dimensionality reduction. However it is important to note other latent structures are possible to model, including mixing categorical and other distributions as relevant for the visualization technique and kind of analysis being done.

The task for the encoder then is to take the nested structure M as input, and output the reparamtetrized latent distribution parameters μ_z and σ_z for each z_i at the maximum resolution latent scale. In order to do this, we must choose a network architecture that can adequately handle the structure of M and/or perform some transformations on M to ensure it is compatible with the chosen encoder network structure. The approach taken by Nested Fusion is to convert the hierarchical set of heterogeneous data points into a single sequence of tokens that can be used as input to an encoder model.[5] This is done by first using a learned mapping T_X which is a linear transformation for each data scale to a common high dimensional token dimension (determined as the sum of all dimensionalities of each data modality to ensure no bottleneck at this stage) that can be used as a common shape for the encoder sequence. Then a sequence is built where starting at the lowest resolution dataset X in M, for each data point $x \in X$ we append the corresponding token at the front of the sequence, and then find the sequence for the nested scales of that token recursively and append the resultant nested sequence (here using addition/summation notation to represent

[5]The ordered sequence is an effective encoding for the nested structure as it can maintain locality within each scale and sequence models in language are perhaps the best examples of contemporary models which effectively encode nested structures (grammar in the case of language).

sequence concatenation).

$$Seq(x_i^{\varnothing}) := [T_X(x_i^{\varnothing})] \tag{6.3}$$

$$Seq(x_j) := [T_X(x_j)] + \sum_{x' \in \eta(x)} Seq(x') \tag{6.4}$$

Once a sequence of tokens is generated this sequence is passed into some sequence-to-sequence encoder model which outputs a sequence of corresponding estimated latent parameterization means and variances. However only the output positions actually corresponding to $x_i^{\varnothing}$ inputs are then taken to sample a latent from the reparamtetrized distribution $z_i \sim \mathcal{N}(\mu_z, \sigma_z)$.

For decoding, remember the conditional distribution for data points defined as $p(x|\beta(x))$. Thus, what is required for decoding is a unique model for each scale in M, where a model either takes as input a single latent in the case of the maximum resolution scale or a set of latents as defined by the correspondence set β. For the latent scale decoder, a simple multi-layer perceptron is an appropriate architecture, while for the higher levels needing to decode sets of latents we can use transformers [147]. Importantly, in order to the prevent the potential pitfall of the model merely using positional information to encode information only used in the aggregate decoding step not corresponding to the actual specific latent at each point, our approach uses a transformer without positional embeddings in this step as they are order invariant, thus ensuring that the full distribution of latents, rather than a few arbitrary picked out latents, properly encodes lower resolution aggregate information.

Finally, given the encoder and decoder models, as well as the latent prior distributions, the models are trained using stochastic variational inference on the evidence lower bound as is standard for a VAE based architecture[75]; implemented in our case using the probabilistic programming framework, Pyro[207].

To evaluate our method of Nested Fusion we test the model performance on the real, large-scale Mars Perseverance PIXL dataset introduced in Section Subsection 6.2.1 com-

paring to existing dimensionality reduction and latent analysis techniques. As analysis of this unique dataset representing the frontier of Mars exploration is the raison d'être for this work as a whole, we specifically focus on evaluation with direct relevance towards the scientific goals and capabilities of scientists actively working at NASA JPL and around the globe on this data.

First, in order to utilize nested fusion we have to define the relevant nested measurement dataset formulation for the PIXL dataset, which we define as:

$$M_{PIXL} := (X_q, ((X_p, \varnothing), \eta_{qp})) \tag{6.5}$$

This includes X_q which consists of 103,005 of quantified spectra which are represented as 52 dimensional non-negative real valued vectors whose elements are the elemental weight percentage values produced from PIXL XRF scan points. Here X_p is the set of 1,983,506 MCC multispectral imaging pixels which are 16 dimensional non-negative real valued vectors[6]. Finally we have η_{qp} which is the nesting function of XRF scan points to corresponding pixels. This is generated by utilizing the known range of XRF beam diameters of the PIXL instrument being approximately 150 microns, as well as the calibrated location alignment of MCC images with XRF scan points. This alignment allows us to have a shared coordinate system and thus calculate physical distance between scan-point centroids and MCC pixels. Thus we can define the nesting function to select all pixels within 75 microns of an XRF scan point, which results in the 100 pixel aggregations previously discussed:

$$\eta_{qp}(x_q) = \{x_p \in X_p \mid \text{distance}(\text{loc}(x_q), \text{loc}(x_p)) \leq 150\mu m\} \tag{6.6}$$

[6]This number is less than what you would expect given that each scan point with a quantification covers an area of 100 pixels, however in reality many of these areas overlap, meaning the same pixel can be included in multiple different scan points. Our formalization of nested measurement datasets allows this without issue and in fact it is preferred to strict partitioning as we can better models the actual resolution of dependency for each measurement. The only issue occurs when converting back into physical space such as with the color plot from Figure Figure 6.1. We address this by simply averaging the multiple produced pixel level decoded inferences for overlapping pixels, however introducing more sophisticated techniques of dis-aggregation is a very promising direction for future work

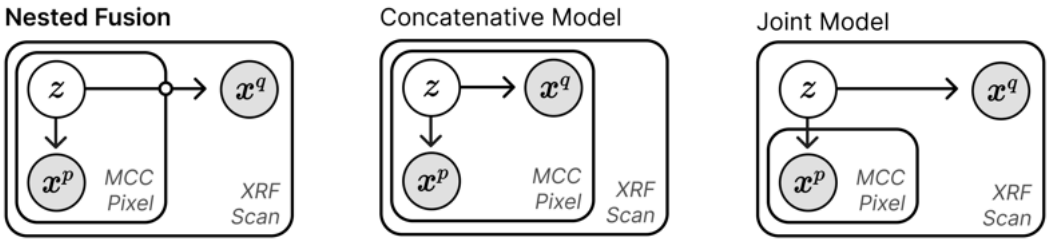

Figure 6.3: Plate Notation for Graphical Models representing different latent variable formulations for the PIXL MCC nested measurement dataset. From left to right we have: (Left) Nested Fusion, representing the latent corresponding to the maximum resolution datascale and informing higher level measurements through aggregated functions; (Center) the concatenative model where there is a latent at the maximum resolution scale which affects higher level corresponding measurements not in aggregate but independently; and (Right) the joint model where a latent exists at low resolution and determines the whole distribution of all high resolution measurements.

6.3.3 Comparing with Alternative Models

To demonstrate the effectiveness of Nested Fusion, we compare it with alternative dimensionality reduction models that can combine both scales of data. Since this problem is non-standard we must introduce the set of alternative models that allows utilization of existing methods to our problem. We categorize these models into three types based on how they handle the nested structure of the PIXL nested measurement dataset, Nested Fusion (our method), Concatenative Models, and Joint Models. We describe these three classes of models in Figure Figure 6.3 using the language of Bayesian graphical models, which illustrates how these classes encompass a full taxonomy of problem conceptualizations for nested measurement datasets[7]. However within each of these classes any particular model type (e.g., UMAP or VAE) can be used. For our comparisons we took both alternative modeling frameworks and for each trained three representative models. First representing the most common approach to dimensionality reduction used ubiquitously in practise is Principle Component Analysis (PCA). Then to represent state of the art dimensionality

[7]This set only covers all possible alternatives when limited to two nested scales such as PIXL, when increasing the number of nestings the number of alternative methods produced by combining Joint and Concatenative models become combinatoric

reduction we used UMAP[63] over t-SNE[64] as it provides state of the art performance, has a well documented history of applications in science, and is among the techniques least sensitive to hyperparameters, and is much more computationally efficient for our scale of data. Finally we also trained a variational autoencoder to represent the most standard approach to generative latent analysis. Since Nested Fusion and the variational autoencoder methods are agnostic to the specific neural network sizes and architectures used, for our evaluation we trained multiple networks using simple multi-layer perceptron models (with the exception of using a transformer encoder for the Nested Fusion decoding step as described in Section Section 6.3) with hidden layer sizes from 64 to 256 and a number of hidden layers from 4 to 16 and selected the best-performing models at each latent dimensionality. Nested Fusion's open-source repository provides the pretrained tested models at https://github.com/pixlise/NestedFusion. These methods together cover the most common latent analysis and dimensionality reduction techniques used in practice, including both parametric and non-parametric methods. Furthermore, as PIXL scientists are the ultimate users who visualize the latents in 1-, 2-, and 3- dimensions we compare Nested Fusion with these alternatives at such dimensions.

Joint Models The first class of alternative model we will consider are joint models which attempt to model the joint distribution of a low resolution data point and it's entire corresponding nested scales in a single latent. For the PIXL dataset we can describe this framework as trying to find a single latent for each XRF scan point:

$$\forall x_i^q \in X_q.\ \ z_i \leftrightarrow (x_i^q, \eta_{qp}(x_i^q)) \tag{6.7}$$

Concatenative Models The other class of model considered are concatenative models, where each high resolution data point is used as the latent scale, and lower resolution corresponding measurements are simply concatenated to the high resolution sample vector. For PIXL we describe this as taking each XRF scan quantification and duplicating it and

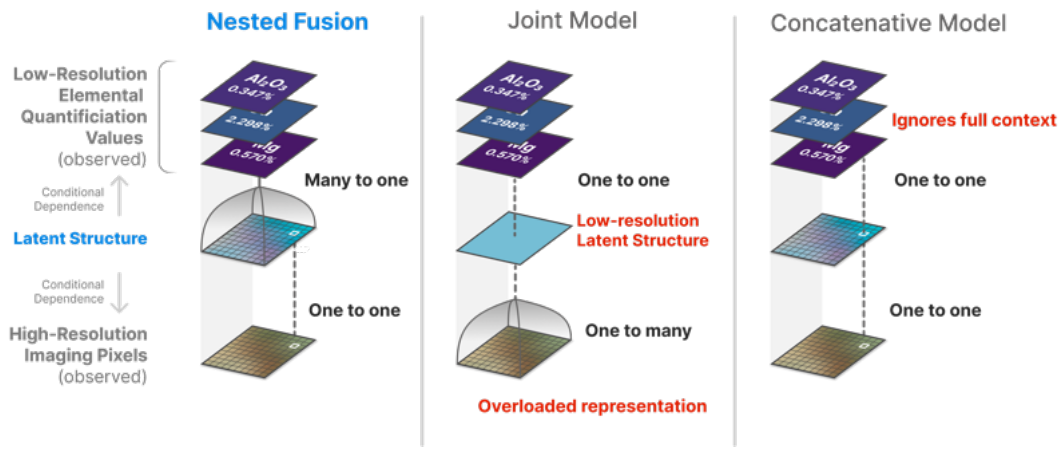

Figure 6.4: Comparison between alternate models and their relative downsides. The left column shows the dependence mappings from the learned latent spaces to the two measurement spaces for Nested Fusion. The center column shows how a joint encoding learns a lower resolution representation which overloads the decoder for high resolution imaging data. The right column shows how a concatenative model ignores to full spatial context of the low resolution measurements by only forming a mapping from a single high resolution point.

concatenating on top of each individual MCC pixel and using this to learn a high resolution latent:

$$\forall x_i^q \in X_q.\ \ \forall x_j^p \in \eta_{qp}(x_i^q).\ \ z_j \leftrightarrow (x_i^q, x_j^p) \tag{6.8}$$

6.4 Evaluation:

Nested Fusion Effectiveness

6.4.1 Conceptual Drawbacks of Alternative Methods Compared to Nested Fusion

Despite covering the full set of possible alternative approaches (given the nested measurement dataset framework), each of these method classes has substantial conceptual drawbacks, illustrated in Figure Figure 6.4. A joint model has a much more difficult encoding task where each latent value is overloaded with encoding the whole set of $\eta(x_i^q)$ making fidelity with low latent dimensionality very difficult. Furthermore it will also only produce

a latent at the lowest possible resolution, the exact opposite of the high resolution latents in Nested Fusion. Concatentative models can perform somewhat better, as they produce latents as similarly high resolution to Nested Fusion. However the concatenative method of combining layers erases all scale contextualization of each high resolution data point, thus encoders and decoders do not have access to more complex distributional information within each nesting scale, which potentially can have an effect on the accuracy of final low resolution estimates when such information is important. For instance, if we consider a case where two scan points includes the same kinds of minerals but in different proportion, this will affect the values of x_i^q and x_j^q in such a way that any concatenative model must necessarily produce different embeddings even for the exact same kind of mineral! This false encoding of the confounding distributional information on the individual scale is inextricable from the concatenative method. However since Nested Fusion has access to this distributional information for its encoder and decoder, in principle it could learn something close to a 'true embedding' which the concatenative model strictly could not. Therefore, given these conceptual drawbacks of the entire range of alternate models to Nested Fusion, we have reason to prefer it based on our prior and theoretical understanding of what the different techniques can learn in principle.

6.4.2 Qualitative Evaluation

It is important to restate that the success or failure of any of the presented latent analysis and dimensionality reduction techniques is determined entirely within the context of their actual use, for the purposes of this paper being in their application within PIXL science. Previous work has outlined the basic structure of how machine learning techniques have been successfully applied within the PIXL science team, by enabling an iterative semantic phenomena modeling process[23] that helps scientists map out the space of considerations before continuing with standard domain modeling. Therefore we begin our evaluation of the different methods of latent analysis at the same point that PIXL scientists begin

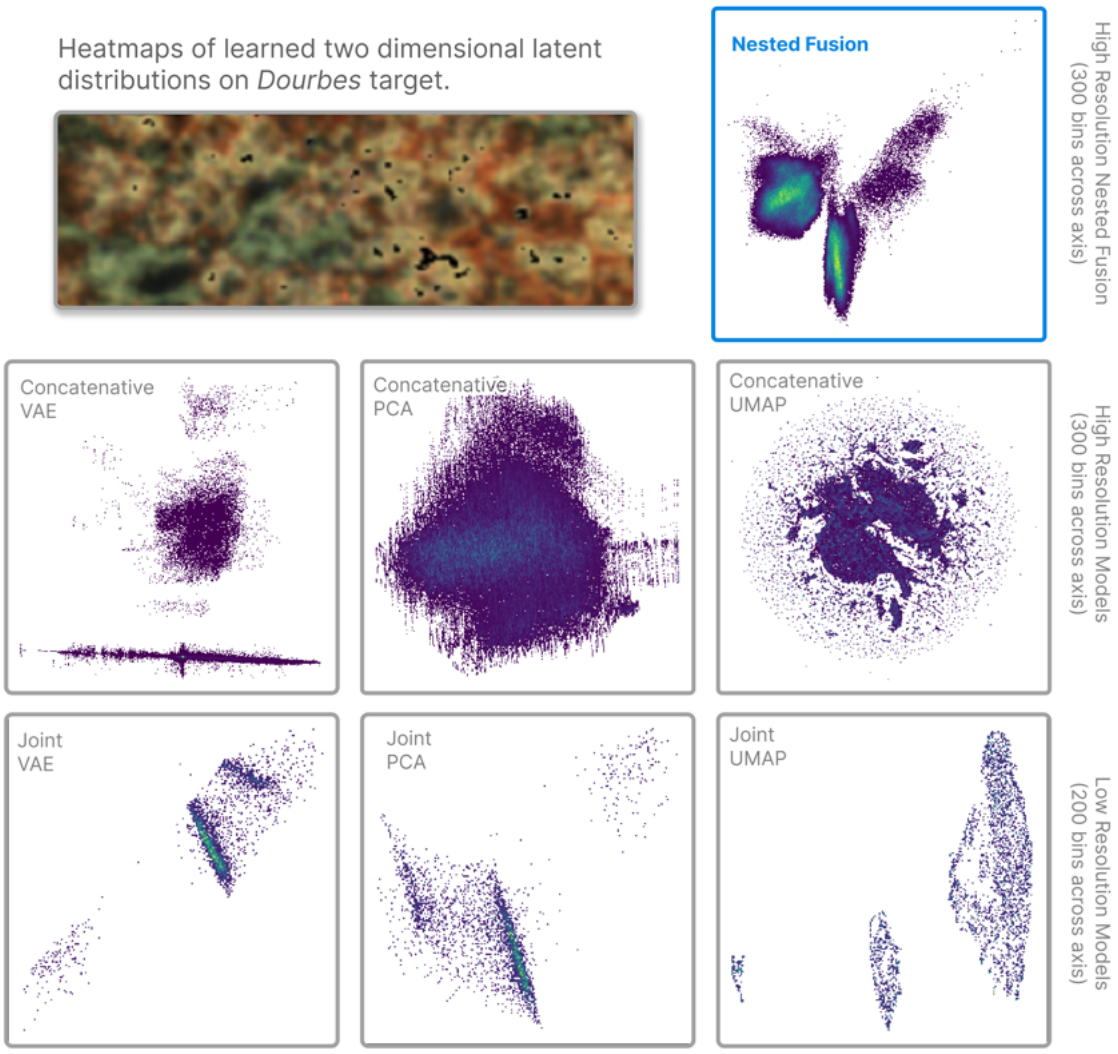

Figure 6.5: Comparison of 2D Latent Distributions from different methods applied to *Dourbes* target (RGB map of MCC Image shown in top right). Axes are unitless latent values. High resolution models (left column: Nested Fusion and concatenative models) displayed with 300 bins across each axis, while low resolution joint models (right column) has 200 bins due to the differing number of samples in each model type. Note that especially for high resolution distributions further structure can be visualized by changing the bin threshold to differentiate modes that overlap, in this figure for simplicity we display the full distribution with no minimum threshold.

their analysis by visualizing the resultant latent distributions produced by each method directly, as two dimensional heatmaps, in order to try to discover the distinct phenomena to consider in their later modeling. Figure Figure 6.5 shows the output of each of the methods applied to the *Dourbes* target from Figure Figure 6.1. Specifically, for such a two dimensional heatmap plot of the latents scientists expect to see a small number of distinguishable regularities which can either be regions visualized in the heatmap as distinct areas of higher density in bright green or as separable clusters which need not be high density but otherwise must be otherwise identifiable as a standalone feature to consider.

In Figure Figure 6.5, notice how all of the joint methods (right column) learn a comparatively small set of regularities, each showing only three distinct modes. While this regularity and differentiability is certainly a positive, we know from previous authoritative analysis on this specific target [26] that there are at least more than three relevant phenomena that must be distinguished and so we have reasonably high confidence that these representations are overly abstracting. The high resolution methods (left column) are more varied. Concatenative VAE, like the joint models, produces three primary clusters, while Concatenative PCA encodes a continuous global structure with limited local differentiation, which in this context makes mineral identification much more difficult. Finally concatenative UMAP produces an extremely complex distribution which shows no consistent high-density regions. Like a Rorschach inkblot, such complexity cannot serve as a reliable basis for building trustworthy shared interpretations between scientists focused on finding specific, repeatable, and understandable regularities. Indeed, the UMAP visualization produced in this context is perhaps the least scientifically helpful of all the options for PIXL scientists working on mineral identifications. Finally, we see that Nested Fusion (top left) produces the most distinguishable structure consisting of two large high-density regions on the left (which each are themselves clearly composed of a mixture of multiple overlapping but non-identical modes) accompanied by two more lower-density clusters on the right and another on the left. The distribution produced by Nested Fusion matches the

scientific priors much more closely, where a reasonable number (more than three and less than a few hundred) of identifiable regularities likely corresponding with minerals can be clearly seen.

To further explore this effect, we compare the latent sub distributions of the highest performing methods (UMAP and Nested Fusion) when selecting known mineral grains to see the reliability of how well the latent space can be used to identify minerals. Based on existing well analyzed data in the Dourbes target[26] we looked to compare methods based on how well they could differentiate known distinct minerals. In Figure Figure 6.6 the red region corresponds to known olivine while the green region corresponds to known pyroxene, two highly distinct mineral types present at the Dourbes target. We then can select the sets of latent samples corresponding to these two spatial regions in the dataset and compare each latent sub-distributions. What we want to see is a high degree of differentiability between the distributions of these two classes of mineral. Using the Wasserstein Distance metric[208] for empirical distributions, we found the Nested Fusion distance to be 1.416 while UMAP was 1.057. This shows Nested Fusion performing nearly **50% better** than UMAP on this metric of mineral differentiability. Owing to the diffuse structure of the UMAP embedding when compared to the highly dense structure of Nested Fusion, the space was less clearly able to form distinct modes of different minerals, which is the primary goal of utilizing dimensionality reduction in this application context.

Additionally, we ran HDBSCAN[209] to extract clusters from within the olivine subset to visualize the regularities in each of the latent spaces, ideally corresponding to the unique mineral grains illustrated in Fig.9 in Tice et al. [26] which is exactly the scientific prior conceptualization of the underlying phenomena at this target. We show in Figure 6.7 the latent response in the original data-space of the detected clusters that most closely align with the scientific prior distributions of minerals. What we see is that the clusters found by the dense regions in the Nested Fusion space can be much more reliably correlated with the final scientific determinations of the Dourbes sample with a cluster correspond-

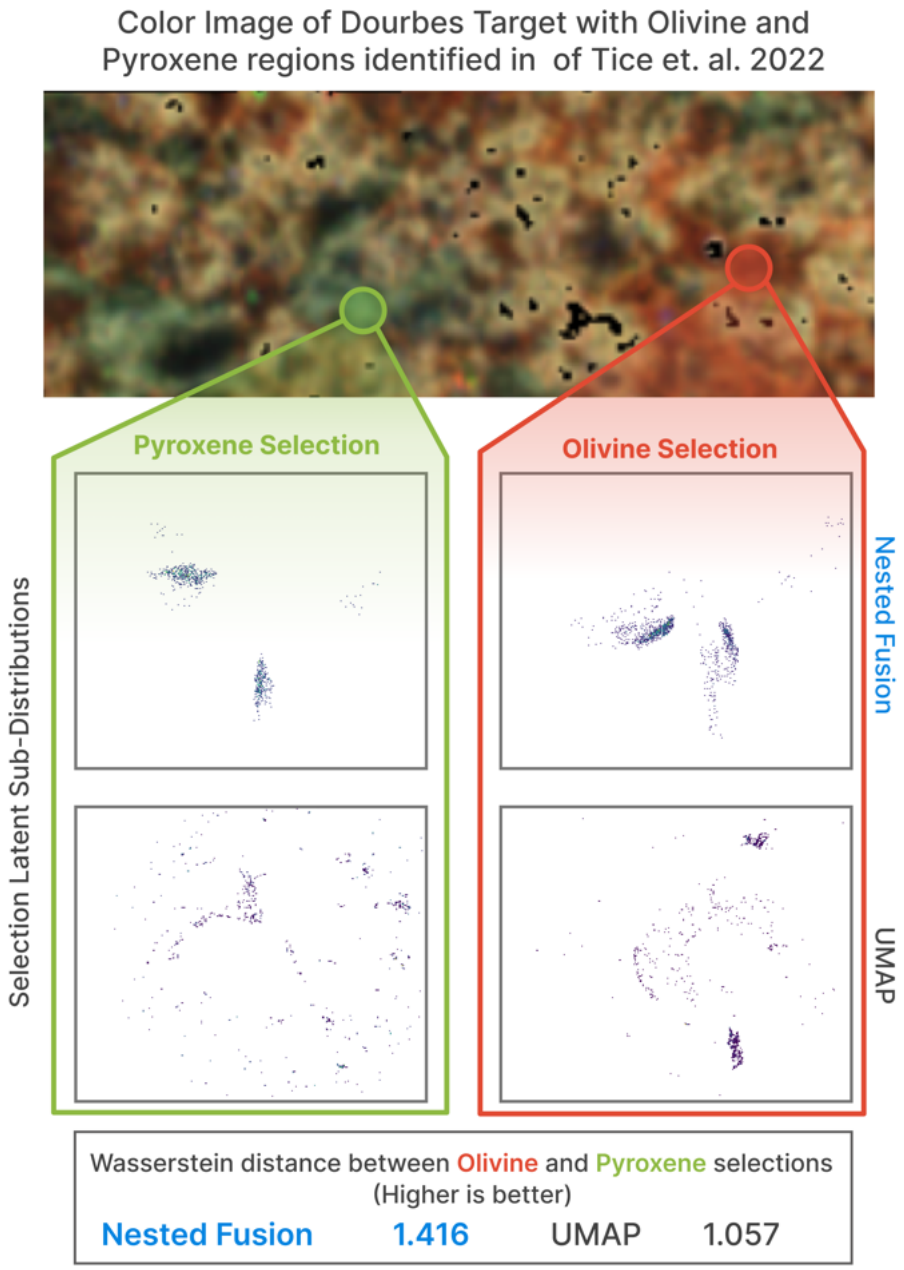

Figure 6.6: Comparison of Nested Fusion and Concatenative UMAP with latent dimension 2 in differentiating distinct minerals in the Dourbes target. In green is shown a region of the target identified as Pyroxene while in red is a region identified as Olivine based on existing analysis[26]. Comparing the latent sub-distributions of these two samples, Nested Fusion produces a distribution which has a greater degree of separation between the different minerals.

ing to low-(Fe,Mg) silicates, a cluster corresponding to Silicate+carbonate, and a cluster corresponding to (Fe,Mg)-carbonate. While on the other side we see in the UMAP space the detected clusters are much less dense and uniform, and showcase a much weaker and noisier loose correlation to the underlying mineralogy.

These results show how Nested Fusion produces a distribution more effective at identifying and representing distinct minerals or other phenomena which aligns precisely with what PIXL scientists hope to achieve in the scientific workflow of exploratory analysis, showing qualitatively the clear superiority of Nested Fusion to the alternative methods in assisting effective science.

6.4.3 Quantitative Evaluation

Besides the qualitative properties of the distributions that make them practically scientifically useful, PIXL scientists also require that the latent models are trustworthy enough in retaining most of the meaningful information present in the underlying data, and since we do not know a-priori what is or is not meaningful we must ensure that a representation retains as much information as possible about the original data to reconstruct it completely. Good fidelity then is a *necessary* but *not sufficient* condition for effective utilization, in particular considering the fidelity of quantifications which scientists trust as more authoritative when grounding mineral identification. Thus, we compare Nested Fusion with alternative models using reconstruction fidelity, a standard metric in evaluating auto-encoding models, to quantify how much information is preserved in the latent encodings. For each model we calculate the coefficient of determination R^2 for both $\hat{X}_q$ as well as $\hat{X}_p$ reconstructions in Table Table 6.2.

Our results show that Nested Fusion significantly outperforms all joint models (Joint VAE, Joint UMAP, and Joint PCA) at each reduced latent dimensionality used by PIXL scientists. This is expected, as explained Section Subsection 6.4.1, because the same dimensional latent values are tasked with a much greater amount of encoding and thus would

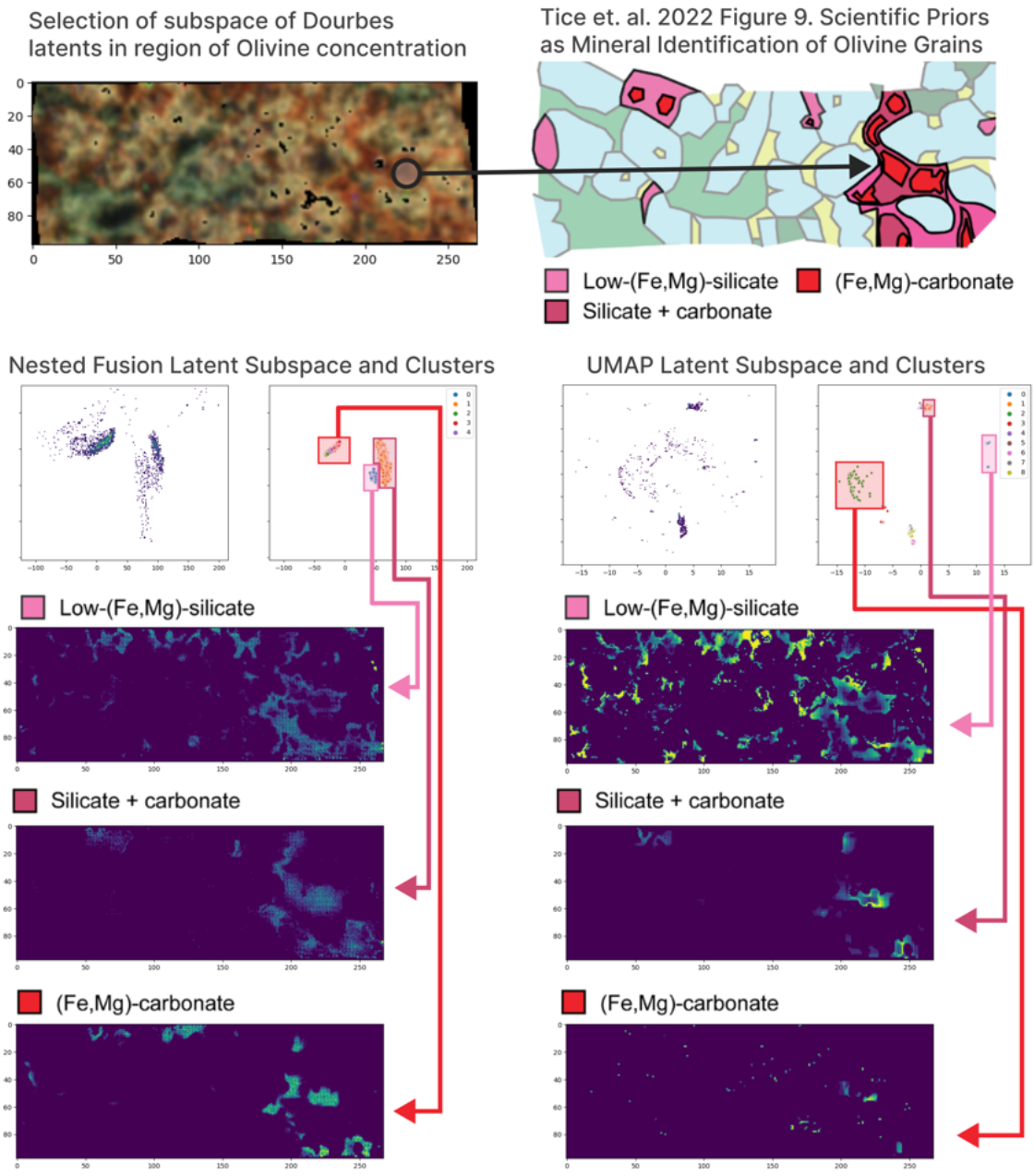

Figure 6.7: Comparison of Nested Fusion and Concatenative UMAP with latent dimension 2 in differentiating distinct mineral grains within the olivine vein in the Dourbes target. Clusters were automatically generated from the latent spaces using HDBSCAN, and the spatial response distributions of the top closely aligned clusters are shown below each space. Clusters extracted from the Nested Fusion space much more closely align with the mineral structure determined by Tice et.al.

be expected to perform worse at the low dimensionalities tested, and it confirms the observations from the qualitative evaluation that important information is likely being lost in the encoding. Concatenative models however tend to perform relatively better in these metrics. Among the concatenative models, concatenative PCA performs universally worst across all metrics, which is not surprising due to PCA being a linear model with limited modeling capacity. Concatenative VAE and UMAP both perform similarly in reconstructing the imaging layer as effectively as Nested Fusion. However, this layer contributes significantly less towards building trust for scientific interpretations as a standalone measurement but is most effective only when augmented with the more solid source of scientific semantic grounding in the XRF quantifications. When considering then the quantification reconstructions, what we find is that as predicted in Section Subsection 6.4.1 Nested Fusion significantly outperforms concatenative VAE in reconstructing the XRF quantification layer. Finally, concantenative UMAP's $\hat{X}_q$ reconstruction fidelity is lower but comparable to Nested Fusion's — however, given the other significant drawback of UMAP's inability to use this accuracy to practically assist in scientific exploration, its reconstruction performance is essentially irrelevant.

In summary, Nested Fusion attains higher reconstruction fidelity than the state of the art in dimensionality reduction and latent modeling while producing substantially more useful latent codes for scientific analysis.

6.5 Scientific Deployment and Impact

The ultimate importance of Nested Fusion is not found in its evaluation metrics but in its ability to have scientific impact by assisting PIXL scientists in visualizing and exploring combinations of datasets they simply could not easily or efficiently do otherwise. Towards this end, we deployed Nested Fusion in multiple capacities within the PIXL science team. The primary method thus far scientists have been able to utilize Nested Fusion is through its standalone implementation which is now open source at https://github.com/

Model	$\dim(z) = 1$		$\dim(z) = 2$		$\dim(z) = 3$	
	R_p^2	R_q^2	R_p^2	R_q^2	R_p^2	R_q^2
Nested Fusion	0.88	**0.92**	*0.97*	**0.98**	0.97	**0.99**
Joint VAE	0.63	0.59	0.81	0.86	0.84	0.93
Joint PCA	0.74	0.02	0.75	0.44	0.75	0.64
Joint UMAP	-	-	0.68	0.65	0.70	0.63
Concatenative VAE	*0.89*	0.81	0.94	0.90	*0.99*	0.93
Concatenative PCA	0.87	0.02	0.88	0.47	0.89	0.65
Concatenative UMAP	-	-	0.96	0.96	0.98	0.97

Table 6.2: Model reconstruction fidelity, measured as reconstruction fidelity R^2 values for both the MCC imaging layer X_p (denoted as R_p^2) and the XRF quantification layer X_q (deonted as R_q^2) for latent dimensions of 1,2, and 3 needed by PIXL scientists. Nested Fusion outperforms all models across all latent dimensions on R_q^2 (highlighted in bold font), the crucial metric used by PIXL scientists when assessing the scientific trustworthiness of a model.

pixel NestedFusion. This implementation directly works on existing and continuously incoming PIXL data, and pre-trained models are also available. Pre-trained models include multiple different latent dimensionalities as well as models that include latent categorical class assignments that have the latent prior distribution being a sample of some latent class from a Dirichlet prior as well as a regular continuous latent code vector which allows the model to differentiate automatically seemingly categorically distinct regions.

With this implementation, PIXL scientists are able to easily visualize the distribution of multiple kinds of latent encodings across many targets at once. PIXL scientists choose to visualize these distributions in a number of ways, including direct distributions in latent space (see Figure Figure 6.5) as well as visualizing various mappings into color overlaying the target image such as the plot in Figure Figure 6.1. These two methods together allow scientists to see both abstract as well as spatial patterns and regularities in the data. These visualization techniques help PIXL scientists firstly discover quick heuristic understandings of the distribution of empirical phenomena present in a single target as well as commonalities in phenomena across multiple targets.

Through participatory design sessions over 6 months with nearly a dozen scientists, we have discovered a primary (though not exclusive) workflow that Nested Fusion enables

within the context of exploratory data analysis. When a new dataset is generated, the process by which PIXL scientists begin to come to a consensus on its mineral composition is highly iterative, involving bringing forward various hypotheses and then coming up with ways to test these hypotheses given the data. The mechanics of how a hypothesis is tested can be complex and difficult, and so a method that could assist in having a more informed starting position in this iteration can greatly increase the efficiency of the whole process, saving a huge amount of extremely valuable and limited time. By forming a latent space over the whole history of PIXL data, and an encoder that can efficiently process this new data before having to retrain, new data can be quickly visualized and broken down into a few key regularities which can be compared to historical precedent of regions or even individual grains which bare a strong resemblance to regions or grains in the new dataset. This then helps to form a better initial assessment of the minerals present at a target and thus substantially speed up the overall identification process. This transforms the workflow of initial exploratory analysis, which historically would take the roughly 10-person team of spectroscopists approximately 21 days in collaboration to come to an initial determination of minerals into one which a single scientist can generate instantly a latent distribution and through refinement generate an identification of comparable quality in a matter of hours.

We found how the combination of the high fidelity of Nested Fusion along with its computational efficiency at inference time were both essential components compared to existing or alternative models to achieve buy-in by scientists. Furthermore we found that non-parametric alternative methods such as UMAP proved to be ineffective despite competitive fidelity due to the inability to form distributions for new data efficiently and producing distributions that are difficult or impossible to reliably interpret within the context of looking to understand specific phenomena, and thus does not help solve the scientific workflow problem that Nested Fusion addresses.

Nested Fusion provides a fundamentally new way for PIXL scientists to quickly visualize distributions of phenomena that span multiple measurement types and scales and

thus explore new data more efficiently and effectively than was previously possible. This has provided a lesson for any interested applied data scientist: increasing the alignment between machine learning problem statement and scientific problem ontology, in this case by more accurately modeling multiple scale relationships, is an absolutely essential component of achieving genuine impact with these tools. Therefore we hope that future work will continue to develop ways in which we can improve the very frame from which we pose data science problems just as much as improving the methods for how we solve them, in order to make sure we can not only do better data science, but just do great science.

CHAPTER 7
DETECTION OF EMERGING DRUGS INVOLVED IN OVERDOSE VIA DIACHRONIC WORD EMBEDDINGS OF SUBSTANCES DISCUSSED ON SOCIAL MEDIA

7.1 Introduction

As discussed in the introduction to Part II, a key component of designing methods that purport to enable better scientific exploratory workflows broadly must be able to extend concrete examples of doing so in multiple different domains. From Chapter 2 we have established that despite differences in domain ontology, when analyzing science from a human-process standpoint different divisions can be drawn, making methods designed for specific workflows in one domain potentially applicable in an entirely different domain as long as certain workflow similarities persist.

In this chapter we will introduce the application of frameworks from Part I towards the new domain of public health epidemiological surveillance in collaboration with scientists at the CDC. This would, on first inspection, and by traditional divisions of discipline, be extremely far removed from the primary scientific use case so far explored in this thesis of planetary geochemistry. However it is just that difference in appearance that makes the fact that the underlying analytic workflows can be extremely similar more striking. The specific problems facing astrogeologists of experimentally inaccessible observational data and the problem of requiring new machine learning models to better conceptualize (see Figure 2.1) are also faced at a conceptual workflow level by public health researchers looking for better ways to measure large scale and fast moving population social dynamics of particular public health interest such as drug use and overdose. Therefore we can apply the same discovery framework to structure the way we develop machine learning models

for this different domain to produce more meaningful scientific results. As the bulk of this chapter was originally published in the domain science Journal of Biomedical Informatics [24], the discovery framework design process underlying the work was not discussed for reasons of scope. However note the design of thew methods introduced in this chapter and how they do in fact align within the design framework of developing machine learning modeling systems that enable scientists to interactively interrogate and negotiate with data (here social media text data) in order to extract phenomena of scientific import (relationship of drugs with overdose).

7.1.1 Emerging Causes of Overdose

Reducing overdose deaths is a top public health priority in the United States.[210] In 2018, 67,367 Americans died from a drug overdose, nearly double the number of individuals dying from a drug overdose in 2009. [211] Indeed, the rapid rise in deaths due to drug overdose over the past decade contributed to decreases in overall life expectancy in the U.S. in 2015, 2016, and 2017.[212, 213, 214, 215] One major challenge to reducing fatalities from drug overdose is that the substances involved in fatalities are continually shifting, creating substantial challenges with matching evidence-based interventions and resource deployment to shifting on-the-ground epidemiology in communities.[216, 217] Research analyzing overdose fatalities over a multi-decade period has demonstrated that the U.S. has experienced various sub-epidemics primarily driven by different drugs over time.[217] In recent years, deaths from overdose in the U.S. have experienced a rapid epidemiologic transition from deaths involving prescription opioids to heroin to highly-potent synthetic opioids.[216] Illicitly manufactured synthetic opioids, namely fentanyl and its analogs, are now the most common drug class involved in overdose deaths in the U.S.[218]

7.1.2 Challenges for Early Detection

The recent and rapid emergence of fentanyl as the drug most commonly involved in overdose deaths in the U.S. illustrates the challenges with prevention of drug overdose as a result of emerging and evolving substances in the illicit drug supply.[219] Historically, fentanyl, which was primarily used as a prescription product in patch or oral lozenge form for the control of chronic pain[220, 221] or as an injectable product in the operative setting for analgesia,[222] had relatively lower rates of misuse compared to other prescription opioids [223] and was not a primary driver of overall overdose mortality.[224] However, since 2013 there has been a rapid, substantial influx of illicitly manufactured fentanyl into the illicit drug supply in the U.S. that has corresponded with a rapid rise in overdose deaths involving synthetic opioids, especially fentanyl. [224, 225, 226] National information on which substances are involved in overdose deaths is derived from death certificates and published by CDC. Unfortunately, owing to the complexity of post-mortem toxicologic testing, the increased time needed for investigation of deaths not from natural causes, rising numbers of overdose deaths needing investigation, inadequately staffed medical examiner and coroner offices, and the decentralization of death records, national information on which substances are involved in overdose deaths has traditionally lagged by 1 or more years. [227] Hence, major nationwide attention to the emergence of illicitly manufactured fentanyl and fentanyl analogs in the illicit drug supply and in overdose deaths did not occur as rapidly as possible given information delays on substances involved in overdose deaths. [228]

7.1.3 Novel data sources and emerging drug detection

Within the past decade there has been a growing body of literature examining the potential of novel social media data sources to better understand substance use patterns and trends.[229, 230, 231, 232, 233, 234] However, major challenges persist in the development of a quantitative approach to successfully identify emerging drugs, with two predominant approaches existing. The first approach seeks to identify conversations about substances

through keyword lists or lexicons and primarily focuses on examining the frequency or proportion that a given drug is mentioned. [228, 235] The major limitation of this approach is that it requires knowledge of which drugs to search for by name and that examining counts or proportions is often inadequate as emerging substances are, by definition, mentioned infrequently during the early years of their emergence. The second leading approach to emerging drug detection focuses on building machine learning classifiers to detect types of posts—such as those discussing substance misuse—and then observe which substances are being mentioned in such conversations.[236] While machine learning models have progressed markedly in their ability to be able to accurately classify posts, the majority of substance misuse related posts discuss popular drugs. Emerging drugs, because they are initially rarer in mention, become hard to detect in the larger volume of conversations about commonly misused substances. Therefore, a need exists to identify improved quantitative approaches for emerging drug detection; in this manuscript we adapt and translate recent methodological approaches in computational linguistics to the task of emerging drug detection.

7.1.4 Measuring Semantic Shifts in Natural Language

Recent work in computational linguistics has explored "diachronic word embeddings," for quantifying shifts in the meaning of words over time.[237, 238] "Word embeddings" are complex numerical representations of words generated from a neural network model and "diachronic" refers to the fact that shifts in these numerical representations are being measured over time. For example, the meaning or context of the word "broadcast" has evolved over time, moving from a family of words used in farming (i.e., sowing seeds) to one nearer to media entertainment, and this shift can be both discovered in an automated fashion and quantitively described. [238] Researchers have recently shown that beyond studying language evolution, use of diachronic word embeddings can enable study of social phenomenon, such as change in gender bias over time through study of natural language.[239]

We hypothesized that use of diachronic word embeddings to measure semantic shifts in drugs over time from large volumes of public social media posts about substance use could potentially enable improved detection of emerging drugs involved in overdose faster than is currently possible with traditional public health surveillance systems. Using the emergence of fentanyl as a primary example, we demonstrate the applicability of diachronic word embeddings for such early detection. These findings can inform future innovative surveillance strategies and the development and implementation of more timely and targeted prevention and response strategies.

7.2 Methods

7.2.1 Data Source

We analyzed anonymous, public posts from Reddit, the largest forum messaging site, which has over an estimated 400 million users and has been rated the third most visited website in the United States. [240] On the platform, users create various public message boards known as 'subreddits', where posts and comments are made. There exist a large number of forums dedicated to substance use-related topics. We identified 225 drug-related subreddits from public lists of such forums (https://www.reddit.com/r/Drugs/wiki/subreddits) and manual qualitative searching of the site. Publicly available, historic Reddit data is accessible from multiple sources, including the Google BigQuery repository [241] and the Pushshift Reddit repository. [242] We downloaded and utilized data from January 2011 through 2018 through the application programming interface (API) of https://pushshift.io/.

7.2.2 Text Processing and word embeddings

As the intent of the project was to measure finely grained temporal shifts in the meaning of words, we separated posts on Reddit into one-month intervals. We used all posts made to Reddit on 225 drug related forums, including top-level posts, known as "submissions" on the platform, and all replies to those posts, known as "comments". Each post was pre-

Year	Total Post Volume	Total Words	Unique Words
2011	3,836,916	74,884,763	383,261
2012	6,915,352	140,974,586	545,869
2013	3,990,656	98,962,910	463,896
2014	7,265,859	176,198,238	652,449
2015	8,538,950	223,211,789	748,023
2016	9,821,619	262,394,595	806,880
2017	10,346,390	269,206,729	838,493
2018	13,704,634	360,365,525	912,460

Table 7.1: Drug Subreddits Corpora Information by Year

Year	Fentanyl[1]	Rx Opioids[2]	Methamphetamine	Cocaine	Heroin	Cannabis
2011	0.54	9.81	26.36	34.29	17.16	354.45
2012	1.43	18.55	27.13	36.12	24.08	340.47
2013	2.63	29.15	39.79	46.42	39.95	227.26
2014	1.93	15.14	36.44	43.14	26.45	271.29
2015	5.96	24.98	42.54	45.95	39.51	230.12
2016	11.45	28.36	50.02	48.84	41.82	201.28
2017	17.63	30.70	49.74	50.50	44.65	161.68
2018	15.57	29.73	51.58	54.41	41.51	208.20

Table 7.2: Proportion of Words from a Given Category per 100,000 Words by Year. [1] The following spelling variants are used to identify fentanyl posts: fent, fents, fentanyl, fentynal, fentanil, fentanils, fentanyls, fentnyl, fetanyl, fentantyl. [2] oxycodone, oxy, roxy, percocet, oxycontin, hydrocodone, vicodin, hydromorphone, dilaudid, morphine

Year	Synthetic Opioid Deaths[1]	Prescription Opioid Deaths[2]
2011	2666	15,140
2012	2628	14,240
2013	3105	14,145
2014	5544	14,838
2015	9580	15,281
2016	19,413	17,087
2017	28,466	17,029
2018	31,335	14,975

Table 7.3: Opioid Deaths by Year

processed using standard natural language processing techniques, including lowercasing text, removing hyperlinks, and removing punctuation.

We then constructed word embeddings using the word2vec model, a dominant method in natural language processing for this task. [243] Specifically, each word is represented by a string of numbers (in our case, 50 numbers) that specify an exact coordinate position or vector in 50-dimensional space. Prior studies in this field of research have used vector lengths ranging from 20 to 1000 dimensions [243]; we chose a 50-dimensional embedding for two main reasons. Firstly, the dataset used in our analysis—while containing tens of millions of posts—is significantly smaller than the original word2vec datasets using a vector length of 300 or more dimensions; reducing dimensionality in these cases helps to avoid overfitting. In addition to increased computational time with larger vector lengths, the dimensionality reduction techniques we use to visualize these embeddings (i.e., to reduce the information to just 2 dimensions for plotting as discussed below) are more challenging to compute in higher dimensions. This model, being relatively simple in comparison to larger pretrained language models or transformer architectures is actually preferable in this context. Since we are looking to detect changes over time, we cannot use pretrained models that can have included training data outside of our temporal scope for a specific embedding space. Additionally since we want to be able to detect trends as soon as possible having better temporal resolution is of maximum importance, which necessarily reduces the size of the training data for each temporal embedding. Therefore, larger models which need more data in order to outperform simpler word2vec models will never be able to achieve the same degree of temporal resolution for the same level of modeling accuracy. Thus word2vec with our cross-validated embedding dimension of 20 is in-fact the optimal model choice for this problem.

We implemented the word2vec model using the skip-gram architecture, given its generally better performance when working with less common words, which is of particular interest when trying to identify emerging trends. Additional parameters utilized for the

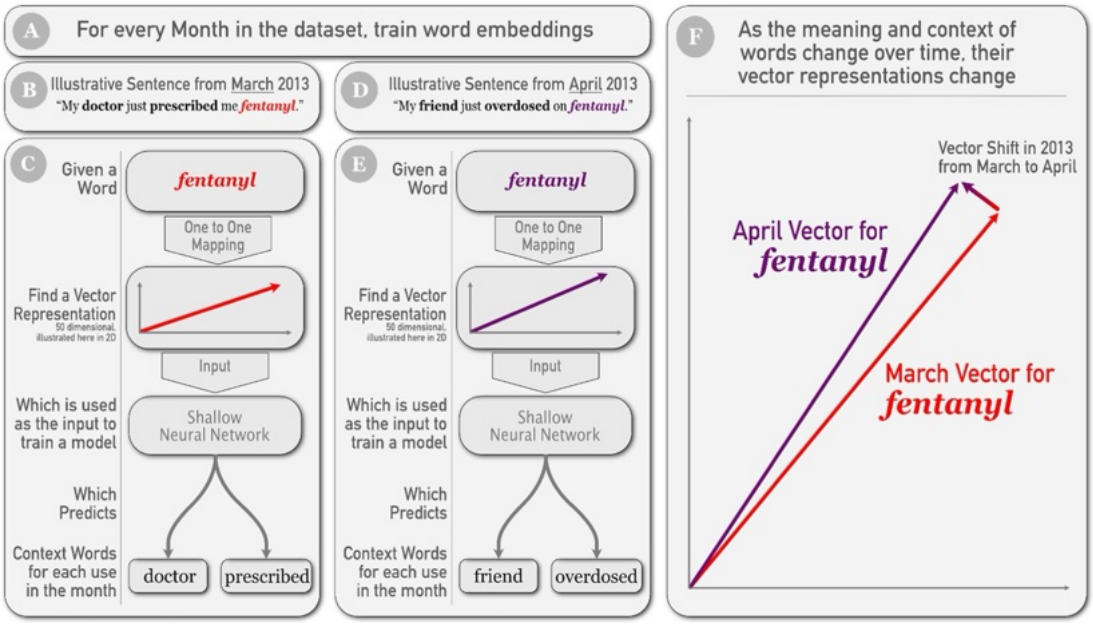

Figure 7.1: Visual Schematic of Word2Vec Processes for Measuring Temporal Shifts in Word Embeddings. Note: Figure presents a visual explanation of how word embeddings are constructed from natural language and used to assess temporal shifts in word meaning

word2vec model included a window size of 5, minimum required word count of 5 to compute a vector, 5 epochs for training, and a negative sampling value of 0. Word2vec is a neural network model that scans though the text of a given post and trains a model to predict the surrounding context given a word, as illustrated in Figure 7.1. This process produces a vector representation for each word (the "word embeddings"). The resulting embeddings are shown to capture semantic similarity through distance metrics in vector space, as discussed below. We trained embeddings for every month from January 2011 through December 2018, resulting in a set of 50 dimensional vectors representing each word's meaning in that month.

7.2.3 Measuring Semantic Change

The next challenge in analyzing semantic shift of words is ensuring that the embedding space that is learned for each time interval is comparable across time. To accomplish this we use a Procrustes transformation for each month to align with the previous month; this

approach is widely used in computational linguistics for this task. [238] Once we have developed this embedding space of word meaning over time, we can then develop ways to further extract specific insights related to words of interest. In particular, we hypothesized that measuring the semantic shift over time of a given drug word, such as 'fentanyl', could be a feasible approach to automate detection of emerging substances causing overdose, as drugs that are increasingly discussed in the context of overdose events would demonstrate increasing semantic proximity to 'overdose' related words.

In order to measure this type of shift, we define a new metric, which we refer to as the Relative Similarity Ratio (RSR). The RSR represents how the similarity of a given word or set of words to another group of words can be calculated, taking into account a reference group. This general quantitative approach forms the basis of linguistic research using diachronic word embeddings. [239] However, while many other methods for analysis of semantic shift over time use generally static vocabularies and pure cosine similarity between words as an analysis metric, in our data the number of new unique words increases significantly over time. This causes pure cosine similarity to undergo 'inflation' as more words in the same space affects the relative angle between words, which requires correction to appropriately measure change over time. Our approach explicitly takes this into account by calculating a relative metric that finds similarity of words compared to a reference group. This method controls for rapidly changing dynamics in the size of the datasets used to populate the diachronic embedding space, such as constantly shifting, social media forums.

In order to calculate the RSR for a given query word (Q), the formula takes a given set of n reference words (R), and a set of m target words (T), as shown:

$$RSR(Q) = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\left(\frac{1}{|T|}\sum_{j=1}^{|T|}\frac{\langle Q_i, T_i\rangle}{\|Q_i\|\,\|T_j\|}\Big/\left(\frac{1}{|T||R|}\sum_{j=1}^{|T|}\sum_{k=1}^{|R|}\frac{\langle T_j, R_k\rangle}{\|T_j\|\,\|R_k\|}\right)\right) \tag{7.1}$$

For each query word q in a given timespan, the RSR is the average cosine similarity of the embedding of word q to a set of target words, which is then compared to the average cosine similarity of a set of reference words to the set of target words. This allows us to generate a flexible single metric on how similar any given word is to a set of target words, relative to a specified baseline. Our main analysis for validation purposes examines semantic shift in the drug 'fentanyl'. In the formula presented above for the RSR the following set of ten fentanyl-related words, drawn from published literature [228] (fent, fents, fentanyl, fentynal, fentanil, fentanils, fentanyls, fentnyl, fetanyl, fentantyl), are each entered as w in the formula to capture semantic usage of the word fentanyl in posts. We calculated the RSR of each word in the fentanyl set every month and used this distribution of values in subsequent analyses using the statistical tests described below. The following set of overdose related terms (od, overdose, ods, oding, overdosing, overdosed, overdoses) are entered as T, the target words to which to cosine similarity of the fentanyl words is measured. Lastly, we created a reference group of words R by which the cosine similarity between these words and the target words are also calculated and then compared to the cosine similarity of fentanyl to the target words, to yield a ratio as presented in the RSR. For the reference group we selected ten common prescription opioid terms that were present in each month of the entire dataset and broadly covered common generic/ brand formulations (oxycodone, oxy, roxy, percocet, oxycontin, hydrocodone, vicodin, hydromorphone, dilaudid, morphine); these are entered as R, the reference words by which the cosine similarity between these words and the target words are also calculated and then compared as a ratio

to the cosine similarity of fentanyl to the target words as presented in the RSR. Of note, for the common prescription opioid terms, we did not enter extensive lists of spelling variants given the sizeable number of drugs already represented in the list and the fact that previous research on prescription opioids has revealed that correct spellings are the most prevalent variant and reference words must be present in the text over the entire time period studied. [234] Furthermore, spelling variants/errors already exist in close semantic proximity to the base words and thus do not meaningfully alter our measures of semantic distance. Sensitivity analyses testing the inclusion of additional variants revealed no change in results. Including potential spelling variants mainly aids in identifying a sufficient number of mentions of a given drug if that substance is relatively rare in the overall corpus, which aids in tightening confidence intervals around estimates for rare substances. For illustrative purposes and to show the semantic movement of fentanyl relative to drugs other than prescription opioids, we calculate RSR values over time for the following additional substances/categories. Up to 10 spelling or colloquial variants for each word are included, drawn from previous research on these substances. [233, 235, 244, 245]

- Heroin: heroin, heroins, herion, h, heroine, heorin, tar
- Cannabis: cannabis, cannibis, marijuana, mj, marajuana, marihuana, ganja, pot, weed
- Methamphetamine: methamphetamine, methamphetamines, methamp, methampetamine, crank, meth, speed, ice, shards, crystal
- Cocaine: cocaine, cocain, blow, coke, crack, crakc, coca, yayo

7.2.4 Statistical analysis

To describe crude trends in word frequency over time in Reddit posts, we first calculated the proportion of words in a particular drug category out of all words used in a given year from 2011 to 2018 (i.e., how many times fentanyl words are used out of all words in a given year). For added context, we also calculated overdose deaths involving prescription

opioids and overdose deaths involving synthetic opioids annually from 2011 to 2018, the most recent year for which data are available, based on data from the National Center for Health Statistics' National Vital Statistics System. [227] Overdose deaths were those with an ICD-10 underlying cause of death code: X40-44, X60-64, X85, Y10-14. Overdoses involving prescription opioids had T40.2-T40.3 in the multiple-cause-of-death field. Overdoses involving synthetic opioids had T40.4 in the multiple-cause-of-death-field. In order to verify that the signal we measure is a meaningful indicator of emerging overdose trends, we examined the case of fentanyl and compared it to other prescription opioids, with the hypothesis being that the RSR of fentanyl to overdose would initially be within the range of other prescription opioids, and over time it would diverge from the reference group as illicitly manufactured fentanyl is increasingly involved in overdoses. We calculated the RSR of each word in the fentanyl set every month and compared this distribution over the course of each year to the RSR values generated from the prescription opioid set using a Mann Whitney U Test. P values ¡ 0.05 were considered statistically significant.

7.2.5 Visualizing word shifts

Diachronic embeddings contain a large amount of information regarding the changing meaning of words over time, however accessing the full breadth of that information is still difficult since we cannot visualize vectors in 50 dimensions. We created 2 plots to visually explore information from word embeddings. The first (Figure 7.2) presents a higher-level exploration of the movement of the word vector for fentanyl over the study period. This plot uses a leading dimensionality reduction algorithm called Uniform Manifold Approximation and Projection (UMAP) to compress the 50 dimensional vectors into 2 dimensions. [246] For ease of visualization in this figure, we plot movement in the word "fentanyl" and do not include spelling variants as these exist in close proximity to the correctly spelled word. The trajectory for "fentanyl" is calculated using a smoothed univariate cubic spline over the set of points for every month. For general context, in this figure we also plot a

subset of other drug words representing some of the most common prescription opioids and illicit substances for general context. The second plot we create (Figure 7.3) shows changes in the RSR values over time, which is a more complex calculation that takes into account multiple drug words and spelling variants, along with a reference group, as described above. Plots are made using the Python Seaborn library and the distribution of the RSR for fentanyl and its spelling variants are plotted with a polynomial trendline and 95% confidence interval.

7.3 Results

A total of 64,420,376 drug-related posts between January 2011 and December 2018 were included in our final analysis (Table 1). Over the time period examined, the total number of posts made in each year increased from 3,836,916 in 2011 to 13,704,634 in 2018. As a proportion of all words used in drug-related posts in a given year, fentanyl increased several fold from 0.54 per 100,000 words in 2011 to 15.57 per 100,000 words in 2018. Common prescription opioid words also increased in usage from 9.81 per 100,000 words in 2011 to 29.73 per 100,000 words in 2018. Word usage trends are also presented in Table 1 for the additional drug categories studied. Fig. 2 plots the trajectory of the UMAP dimensionality-reduced word vectors for fentanyl from 2011 to 2018. As shown by the blue arrow, fentanyl shows two important patterns during the study period. First, fentanyl moves from close proximity to other prescription opioids such as oxycodone and Percocet, toward illicitly used substances, such as heroin and cocaine. Second, fentanyl moves more closely to overdose and overdose-related terms. Fig. 3 plots the Relative Similarity Ratio (RSR) displaying the semantic proximity of various substances to overdose over time, in relation to the reference category of common prescription opioids. The figure show both the individual measurements of the RSR for each word in the substance groups, as well as the overall group trendline calculated by a 10th order polynomial. Both fentanyl and heroin exhibit an upward trend in RSR over time, suggesting an increasing semantic prox-

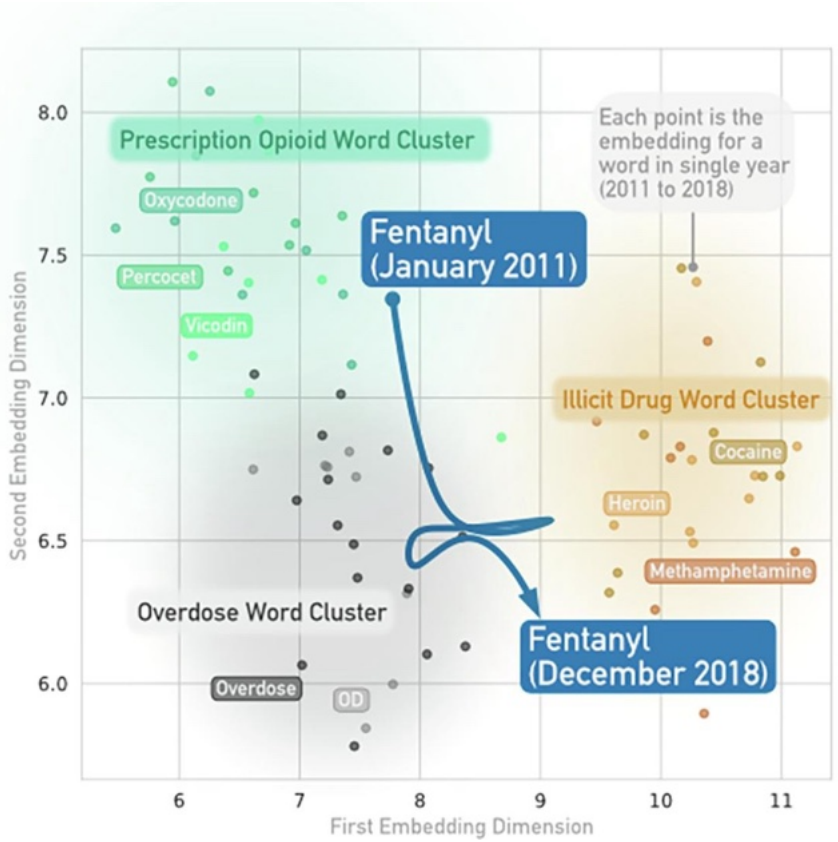

Figure 7.2: Semantic Movement of Fentanyl in Two-Dimensional Space, 2011 to 2018. Note: Fig. 2 plots the trajectory of the word 'fentanyl' (blue arrow) from 2011 to 2018, as calculated from the positions of each month's word embeddings by the Uniform Manifold Approximation and Projection (UMAP) algorithm. UMAP compresses the 50 dimension word vectors into just 2 dimensions for visualiation on a standard coordinate plane; each UMAP axis is an arbitrary unitless number representing position on the best fitting two dimensional structure and is intended to allow for the visualization of the relative position of nearby words and their clusters. For ease of visualization, only select drug words are plotted and spelling variants are not included as they exist in very close vector space to the correctly spelled substance word. Words illustritive of prescription opioids (such as oxycodone and Percocet), other illicitly used substances (cocaine, methamphetamine, heroin), and overdose (overdose, od) are also plotted with their yearly values to reveal the general semantic space in which these words exist. Fentanyl moves from close proximity to other prescription opioids toward illicitly used substances and overdose.

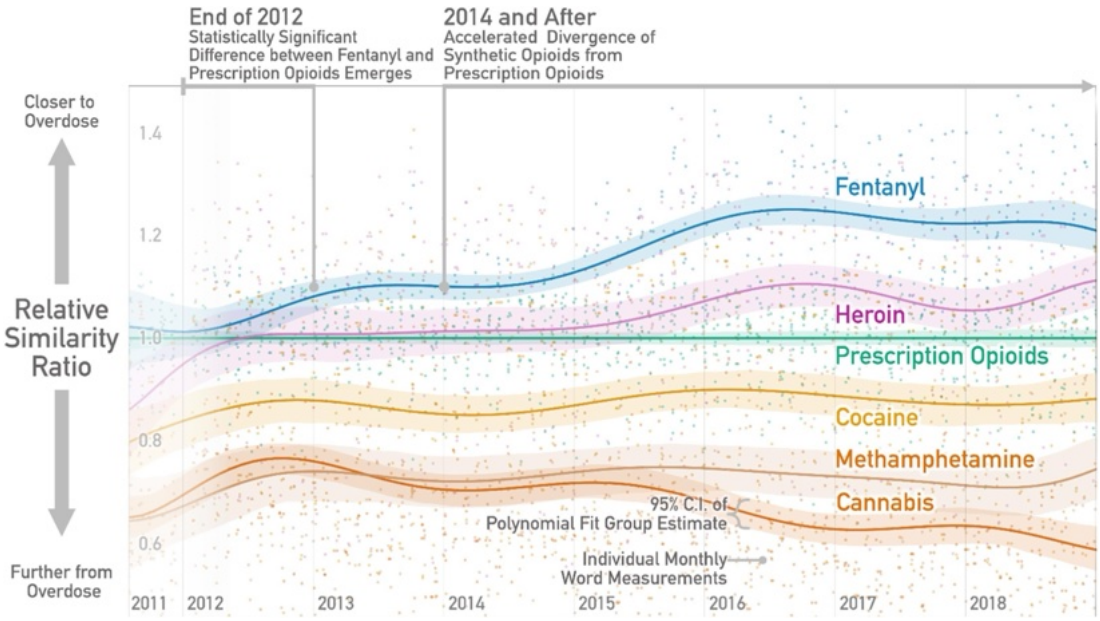

Figure 7.3: Relative Similarity Ratio Displaying Semantic Proximity of Various Substances to Overdose Over Time. Note: Y-axis displays the Relative Similarity Ratio (RSR) metric, which compares the strength of the semantic association between a given substance (i.e., fentanyl) and overdose to the strength of the semantic association between the reference group (common prescription opioids, centered at 0) and overdose. The degree to which a substance is above the green horizontal center line reveals its semantic proximity to overdose, relative to the common prescription opioid terms. Trendlines are drawn through the monthly RSR values for each word in a given category (i.e., fentanyl and its spelling variants) with a 10th order polynomial approximation for the group trend and shading displaying a surrounding 95% confidence interval.

imity to overdose. The RSR for fentanyl increases beyond that of heroin, indicating that semantic proximity of fentanyl to overdose is greater than hat of heroin to overdose. The RSR of fentanyl peaks in 2016 at 1.25 (95% CI 1.22–1.28) and plateaus thereafter, relative to the reference group of prescription opioids. The RSR for other illicit drugs including cocaine, methamphetamine, and cannabis is lower than the reference category of common prescription opioids. Annual differences in the distribution of the RSR for fentanyl and overdose compared to prescription opioids and overdose are depicted in Table 2. On an annual basis, statistically significant differences between fentanyl and other prescription opioids emerge by 2012. After 2012, the strength of the statistical difference markedly increases. This emergence of fentanyl from real-time online data from 2012 was more than 1–2 years earlier than information available to public health professionals at the time; the first increase in deaths from fentanyl occurred in 2013 but this information was not observed by public health professionals until 2014 owing to the 1 year lag in mortality data reporting nationally.

7.4 Discussion

In this chapter, we adapt a mathematical approach from computational linguistics based on diachronic word embeddings and demonstrate that such a measure holds promise for enabling automated detection of emerging drugs involved in overdose. Indeed, our findings yield intelligible results on substance word shifts that can be assessed both quantitatively and visually. Although multiple studies have indicated that real-time social media data on illicit substances reveal potentially useful correlations with findings from slower, traditional epidemiologic systems, [228, 235, 247, 248] there have been a number of challenges to-date in fully realizing the potential of social media data for emerging drug detection. As noted in the introduction, drug detection efforts using a purely lexicon-based approaches rely on a keyword list of drugs of interest to search large volumes of social media text for matching mentions and track frequencies and proportions of drug terms overtime. Lexicon-

based approaches, while widely used, present challenges for emerging drug detection as emerging drugs are rare in frequency earlier in their emergence. [228, 235]

Machine learning based approaches, which attempt to build linguistic classifiers that can detect whether a given post is discussing substance use, [229] also struggle with emerging drug detection as initially the raw frequency of such substances may be very small relative to commonly used substances, causing initially rarer, novel compounds to be difficult to identify among the large volume of discussions about other drugs. The diachronic word embedding approach which we apply in this manuscript aims to further techniques to overcome a number of these challenges. Perhaps most importantly, this approach is independent of the frequency of mentions of a given substance. Although a larger number of drug mentions helps to increases confidence around estimates, the semantic distance between a given substance and overdose can be calculated for any quantity of posts. Secondly, through developing our measure as a ratio and with the inclusion of a reference group, we are able to examine the substances most closely associated with overdose in spite of myriad semantic shifts that occur over time.

While the current study focuses on retrospectively validating the ability of diachronic word embeddings to identify changes in the semantic context of fentanyl, a key emerging drug of the past decade, ultimately prospective deployment and detection of emerging drugs is needed. While a longitudinal, prospective validation is beyond the scope of the current manuscript, there are a couple approaches by which we believe such a system could operate in a semi-supervised way. First, because emerging drug names are often not known a priori, we could extract a large corpus of potential drug words by querying for cosine similarity to known substances. Although emerging drugs could conceivably arise in a completely new pharmacologic class, in general, emerging drugs are variants of existing drug classes and compounds, such as carfentanil and other fentanyl analogs emerging following fentanyl. After extracting a large number of potential candidate drugs, these drug names would then be passed through the RSR metric to measure their semantic proximity to overdose over-

time. A human with expertise in the subject matter, such as a toxicologist would then examine the drugs which show the closest proximity to overdose or with rapid movement toward overdose and use this information to inform further study of these compounds using traditional public health and laboratory systems.

It should be noted that it is also possible to pass every word in the corpus (which would encompass drug words as well as non-drug words) through the RSR calculation. This approach would be expected to increase sensitivity of detection of new substances but at the potential added cost of additional person-time to inspect an increased number of candidate words. Substances of concern from social media derived signals, could be verified by early field studies, [249] undergo validation by inspection of trends from other data systems such as prescribing data, Emergency Department based syndromic surveillance, or illicit drug seizures by law enforcement, [225] or aid in the development and deployment of relevant laboratory assays.

Some limitations of this work and areas for future research should be noted. Overall, generalizability of these models to future emerging substances and applicability to other social networking sites given the unique nature of language on each site is not known and merits future research. Further prospective validation is also needed to fully refine the most successful target word(s) for emerging drug detection. In our analyses, we used a set of overdose related terms as the target words to which to cosine similarity of emerging drugs was measured. In our testing, this target set was intuitive and worked well for synthetic opioids such as fentanyl; however, detection of other emerging or reemerging substances (which may not be primary causes of overdose but rather are co-used with opioids or simply used for psychoactive effects but are not highly fatal) will likely require an alternate or expanded set of target words. For example, if one wished to identify emerging drugs of use that are not necessarily associated with overdose, the target set of words might include the terms “high”, “amped”, “stoned”, etc. Nonetheless, our measure is flexible to these adjustments. Additionally, macro-level secular changes in word usage has the potential

to alter the results of such an approach. For example, towards the end of the time period studied, we note a decrease in the RSR of fentanyl, though it still is significantly higher than the common prescription opioids reference group. Although fentanyl remains the single leading drug causing excess overdose mortality in the U.S., fentanyl is now widely appreciated as such and is no longer an "emerging drug." Thus, posts discussing fentanyl have likely, to some degree, shifted to other topics beyond overdose concerns, causing a slight decrease in RSR. It is also important to note that beyond the semantic proximity of fentanyl to overdose, the strength of the association of the reference words with overdose also influences the RSR for fentanyl. We suspect that our approach is most useful for detecting emerging drugs or detecting shifts in drug use patterns.

Nonetheless, this research helps advance computational approaches to substance use and overdose prevention and may have broader applicability for other public health use cases. For example, better quantifying language change over time may aid in the study of public perception of health-related policies or in understanding large-scale public shifts in norms, behaviors, and beliefs. For example, understanding public shifts in beliefs about corporal punishment for children is a key area of interest for violence prevention researchers that lacks large scale quantitative information and could be further explored through these methods. Similarly, understanding changing levels of stigmatization and stigmatizing language over time surrounding discussion of mental health could be explored through these methods. The use of prescription and illicit drugs continues to undergo evolution in the U.S., including among opioids, stimulants, cannabinoids, and other substances. [250, 251]While social media data hold promise for emerging threat detection, the development of robust and scalable mathematical approaches are urgently needed to extract insights in a way that is manageable and actionable for public health professionals. Automated assessment of semantic shifts in substances as detected in large volumes of unstructured text may improve efforts at early detection of emerging drugs and thereby accelerate early recognition and prevention and response efforts by public health and clinical professionals.

CHAPTER 8
CONCLUSION AND FUTURE DIRECTIONS

In summary, my dissertation presents a new paradigm for the application of machine learning techniques into scientific workflows by bridging methodologies from ML and HCI to frame the whole process of model development and tool design as a socially contextualized human-centered process, focusing particularly on transforming problems of *data* into problems of concrete *perceptible phenomena*. This thesis thus advances knowledge in AI/ML and HCI, improving the capability of machine learning systems to be actually utilized by scientific experts by taking seriously the human-centered process of scientific discovery and developing methods to integrate the specific understandings of users into every aspect of the design and development process.

8.1 Research Contributions Revisited

The research that comprises this thesis has contributed to the study of how users utilize visualization systems to interact with machine learning models in an interrogative context and developed novel design frameworks to maximize how scientific users can integrate themselves more tightly into the creation of intelligent systems in order to more effectively solve their actual problems. I have utilized these design insights towards the development of novel machine learning algorithms and methods which better integrate into these frameworks and thus improve over the state of the art with respect to addressing concrete user needs. This work has had a major impact both within CS/ML/HCI, as well as in domain science as highlighted by:

- In collaboration with NASA JPL, I introduced Nested Fusion [25], a new method for latent visualization of complex multi-scale data, learning representations at the high-

est possible resolution that are much more scientifically interpretable than existing approaches. This work was recognized as **Best Paper Runner-Up (Applied Data Science Track) at the world's top Data Mining Conference ACM KDD 2024**.

- With Public Health Researchers at the CDC, I pioneered a novel method for interactive extraction of temporal semantic trends from diachronic word embeddings, enabling detection of trends in the opioid epidemic **over a year earlier** than traditional methods [24] which has been recognized with the **2022 CDC Excellence in Quantitative Sciences Award.**

- I spear-headed a **first-of-its-kind design framework**, ISHMAP, which introduces a radical new approach for human-centered model development [23], work which was awarded the **2023 NASA Space Act Board Award**. This work was **used by NASA scientists** as an essential component of the scientific discoveries published in the world leading journal, Science [26, 27] and covered by **over two dozen global news outlets** [28].

- I introduced the **first comparative survey and taxonomy of Human-Centered AI design guidelines used in practice** that forms a fundamental structure for studying and designing AI systems [21] which has been used for teaching Machine Learning Interaction at the University of Washington, a top computer science gradate program [29].

Furthermore, the elaboration of of Discovery Frameworks and Human-Centered methodologies as a unifying conceptual framework that reinterprets these individual research accomplishments within a more comprehensive and theoretically cohesive structure provides the foundational contribution toward enabling future scholarship that can similarly empower scientists.

8.2 Future Research Vision

This thesis and the research that it is composed of has been able to develop tools, technologies, frameworks, and methods to bring people into the center of the whole context of the design of machine learning systems, and to understand and improve the human process of statistically and computationally mediated scientific discovery. At the same time it opens many opportunities for future research and enables many exciting kinds of future applications. In this final section I will outline some of the future research directions I am most excited about.

8.2.1 Fostering Additional Domain Collaborations

Each different scientific domain, and in-fact each different lab within a domain faces unique, interesting, and surprising problems from which we can learn. For instance, one such group is earth scientists studying climate and biomarker signals from satellite data, who face similar observational mutli-modal analysis problems as Mars scientists discussed in Chapter 6. In-fact nearly all domains of observational science face the same general problem of finding a way to represent and explore complex data. My human-centered approach provides the best way to address these common problems in the way that is well suited to each individual domain and group of scientists, and continuing to extend this approach to more new cases is likely to provide many years worth of fruitful and impactful research.

8.2.2 Experimental Evaluations of Scientific Discovery Processes and Machine Learning Tools

While qualitative work with a small group of scientific expert collaborators is highly generative of new approaches to scientific machine learning problems, future work could augment this with developing a human-centered test-bed for advanced methods of evaluation

and measurement to fully explore the mechanics of discovery frameworks and how they work. Bringing to bear the insights forged from qualitative work and domain collaboration into larger scale controlled experimental human-subjects research, including advanced psychophysical and neuroimaging measurements such as eye tracking and EEG could experimentally determine the precise analytic inductive biases present when people, with all of their perceptual and cognitive biases, utilize data analysis technologies in the processes of conceptualization as well as the other cognitive process of scientific inquiry. This may help to understand the causes of the essential ”Eureka!” moments of discovery, and the aspects of design that best enable them.

8.2.3 Developing New Human-Centered Discovery Frameworks

While my work in developing design frameworks such as ISHMAP [23] provides a promising start to human centered theories of design for machine learning and data analysis; there are still many questions that remain to be answered. What thought processes and unconscious factors are important for the discovery of novel phenomena? How do scientists differentiate meaningful versus unimportant patterns? How does the process of communication of preliminary findings affect the conceptualization of data both within and across domains? Furthermore, while the frameworks I have discussed have primarily focused on exploration and discovery based phases of inquiry such as conceptualization, further study to structure systems involved in the normative justificatory side of inquiry are also required. In-fact such frameworks are especially important with decreasing trust in science, and an increase in LLM powered AI applications in science based on less scientifically grounded methodologies.

Answering these questions requires not only the empirical approaches I have thus-far utilized, but also requires more well developed theoretical structures within which to situate these findings. As these questions are fundamentally focused on questions of human understanding, perception, and communication, this must entail deeper engagement with the

social sciences and humanities which presents an extremely exciting and novel approach to the study of machine learning and scientific data analysis moving forward.

8.3 Conclusion

It is often overlooked that fundamentally the goal of data analysis and it's accompanying technologies in machine learning and data science is not just better benchmarks but the discovery of new knowledge. Knowledge being something that is not some abstract entity, but rather something existing concretely within a human and social context. In order to do good science it seems almost tautological to say that we have to genuinely care about what scientists *as people* are actually doing. It is my primary goal as a researcher to situate scientists in the center of how we design, understand, and evaluate machine learning and data science technology. This involves a bridging of methods and frameworks from ML/AI, HCI, Human-Centered Design, and scientific domain expertise. I hope this thesis can be a catalyst for a future where by synthesizing these fields, we can better design our systems of data so that the actual people doing science will be better equipped to make groundbreaking scientific discoveries.

REFERENCES

[1] S. Leonelli, "Data-centric biology: A philosophical study," in *Data-centric biology*, University of Chicago Press, 2016.

[2] C. Anderson, "The end of theory: The data deluge makes the scientific method obsolete," *Wired magazine*, vol. 16, no. 7, pp. 16–07, 2008.

[3] P. Godfrey-Smith, *Theory and reality: An introduction to the philosophy of science*. University of Chicago Press, 2009.

[4] C. Coopmans, J. Vertesi, M. E. Lynch, and S. Woolgar, *Representation in scientific practice revisited*. MIT Press, 2014.

[5] J. Hirschberg, "Every time i fire a linguist, my performance goes up, and other myths of the statistical natural language processing revolution. invited talk," in *Fifteenth National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[6] S. J. Russell and P. Norvig, *Artificial intelligence a modern approach*. London, 2010.

[7] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, http://www.deeplearningbook.org.

[8] e. a. Nestor Maslej, *The ai index 2025 annual report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, 2025.

[9] *Nsf announces 7 new national artificial intelligence research institutes*, https://new.nsf.gov/news/nsf-announces-7-new-national-artificial, Accessed: 2024-03-05.

[10] S. Succi and P. V. Coveney, "Big data: The end of the scientific method?" *Philosophical Transactions of the Royal Society A*, vol. 377, no. 2142, p. 20 180 145, 2019.

[11] H. Wang *et al.*, "Scientific discovery in the age of artificial intelligence," *Nature*, vol. 620, no. 7972, pp. 47–60, 2023.

[12] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, "Artificial intelligence in disease diagnosis: A systematic literature review, synthesizing framework and future research agenda," *Journal of ambient intelligence and humanized computing*, vol. 14, no. 7, pp. 8459–8486, 2023.

[13] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: An analytical review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 5, e1424, 2021.

[14] F. K. Došilović, M. Brčić, and N. Hlupić, "Explainable artificial intelligence: A survey," in *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, IEEE, 2018, pp. 0210–0215.

[15] P. Chau, A. Endert, D. A. Keim, and D. Oelke, "Interactive visualization for fostering trust in ml," 2023.

[16] F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim, "Should we trust (x) ai? design dimensions for structured experimental evaluations," *arXiv preprint arXiv:2009.06433*, 2020.

[17] F. Hohman, H. Park, C. Robinson, and D. H. P. Chau, "S ummit: Scaling deep learning interpretability by visualizing activation and attribution summarizations," *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1096–1106, 2019.

[18] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau, "Visual analytics in deep learning: An interrogative survey for the next frontiers," *IEEE transactions on visualization and computer graphics*, vol. 25, no. 8, pp. 2674–2693, 2018.

[19] Z. J. Wang *et al.*, "Cnn explainer: Learning convolutional neural networks with interactive visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1396–1406, 2020.

[20] H. Park *et al.*, "Neurocartography: Scalable automatic visual summarization of concepts in deep neural networks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 1, pp. 813–823, 2022.

[21] A. P. Wright *et al.*, "A comparative analysis of industry human-ai interaction guidelines," *arXiv preprint arXiv:2010.11761*, 2020.

[22] A. P. Wright *et al.*, "Recast: Enabling user recourse and interpretability of toxicity detection models with interactive visualization," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CSCW1, pp. 1–26, 2021.

[23] A. P. Wright, P. Nemere, A. Galvin, D. H. Chau, and S. Davidoff, "Lessons from the development of an anomaly detection interface on the mars perseverance rover using the ishmap framework," in *Proceedings of the 28th International Conference on Intelligent User Interfaces*, 2023, pp. 91–105.

[24] A. P. Wright, C. M. Jones, D. H. Chau, R. M. Gladden, and S. A. Sumner, "Detection of emerging drugs involved in overdose via diachronic word embeddings of substances discussed on social media," *Journal of Biomedical Informatics*, vol. 119, p. 103 824, 2021.

[25] A. P. Wright, S. Davidoff, and D. H. Chau, "Nested fusion: A method for learning high resolution latent structure of multi-scale measurement data on mars," in *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024, pp. 5969–5978.

[26] M. M. Tice *et al.*, "Alteration history of séítah formation rocks inferred by pixl x-ray fluorescence, x-ray diffraction, and multispectral imaging on mars," *Science Advances*, vol. 8, no. 47, eabp9084, 2022.

[27] Y. Liu *et al.*, "An olivine cumulate outcrop on the floor of jezero crater, mars," *Science*, vol. 377, no. 6614, pp. 1513–1519, 2022.

[28] *Altmetric imact report for tice et al*, https://scienceadvances.altmetric.com/details/138878908/news, Accessed: 2024-03-05.

[29] K. Patel, *University of washington, computer science 510 fall 2022 sylabus*, 2022.

[30] C. Lu, C. Lu, R. T. Lange, J. Foerster, J. Clune, and D. Ha, "The ai scientist: Towards fully automated open-ended scientific discovery," *arXiv preprint arXiv:2408.06292*, 2024.

[31] B. Hepburn and H. Andersen, "Scientific Method," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Summer 2021, Metaphysics Research Lab, Stanford University, 2021.

[32] F. Bacon, *Novum organum*. Clarendon press, 1878.

[33] M. Mulkay and G. N. Gilbert, "Putting philosophy to work: Karl popper's influence on scientific practice," *Philosophy of the Social Sciences*, vol. 11, no. 3, pp. 389–407, 1981.

[34] K. Popper, *The logic of scientific discovery*. Routledge, 2005.

[35] T. S. Kuhn, *The structure of scientific revolutions*. Chicago University of Chicago Press, 1970, vol. 111.

[36] P. Feyerabend, *Against method: Outline of an anarchistic theory of knowledge*. Verso Books, 2020.

[37] E. Oberheim and P. Hoyningen-Huene, "The Incommensurability of Scientific Theories," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Spring 2025, Metaphysics Research Lab, Stanford University, 2025.

[38] D. Bloor, "The heirs to the subject that used to be called philosophy," in *Wittgenstein: A social theory of knowledge*. London: Macmillan Education UK, 1983, pp. 182–184, ISBN: 978-1-349-17273-3.

[39] B. Latour, *Reassembling the social: An introduction to actor-network-theory*. Oup Oxford, 2007.

[40] B. Latour and S. Woolgar, *Laboratory life: The construction of scientific facts*. Princeton university press, 2013.

[41] M. Lynch and S. Woolgar, "Representation in scientific practice," 1990.

[42] B. Latour, *Science in action: How to follow scientists and engineers through society*. Harvard university press, 1987.

[43] "Scientific Research and Big Data," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta, Ed., Summer 2020, Metaphysics Research Lab, Stanford University, 2020.

[44] S. Carrol and D. Goodstein, "Defining the scientific method," *Nature methods*, vol. 6, no. 4, p. 237, 2009.

[45] W. Pietsch, "Aspects of theory-ladenness in data-intensive science," *Philosophy of Science*, vol. 82, no. 5, pp. 905–916, 2015.

[46] K. C. Elliott, K. S. Cheruvelil, G. M. Montgomery, and P. A. Soranno, "Conceptions of good science in our data-rich world," *BioScience*, vol. 66, no. 10, pp. 880–889, 2016.

[47] T. Nickles, "Alien reasoning: Is a major change in scientific research underway?" *Topoi*, vol. 39, no. 4, pp. 901–914, 2020.

[48] J. Bogen and J. Woodward, "Saving the phenomena," *The philosophical review*, vol. 97, no. 3, pp. 303–352, 1988.

[49] S. Leonelli, "What distinguishes data from models?" *European journal for philosophy of science*, vol. 9, no. 2, p. 22, 2019.

[50] N. Cartwright, *Nature, the artful modeler: Lectures on laws, science, how nature arranges the world and how we can arrange it better*. Open Court Publishing, 2019, vol. 23.

[51] M. Mirbabaie, S. Stieglitz, and N. R. Frick, "Artificial intelligence in disease diagnostics: A critical review and classification on the current state of research guiding future direction," *Health and Technology*, vol. 11, no. 4, pp. 693–731, 2021.

[52] E. O. Pyzer-Knapp *et al.*, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, no. 1, p. 84, 2022.

[53] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with alphafold 3," *Nature*, vol. 630, no. 8016, pp. 493–500, 2024.

[54] L. Zhang, J. Han, H. Wang, R. Car, and E. Weinan, "Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics," *Physical review letters*, vol. 120, no. 14, p. 143 001, 2018.

[55] C. Park *et al.*, *Sanvis: Visual analytics for understanding self-attention networks*, 2019. arXiv: 1909.09595 [cs.CL].

[56] B. Hoover, H. Strobelt, and S. Gehrmann, *Exbert: A visual analysis tool to explore learned representations in transformers models*, 2019. arXiv: 1910.05276 [cs.CL].

[57] M. T. Ribeiro, S. Singh, and C. Guestrin, "" why should i trust you?" explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[58] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[59] V. Lai, H. Liu, and C. Tan, "" why is' chicago'deceptive?" towards building model-driven tutorials for humans," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–13.

[60] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2019.

[61] T. Wu, M. T. Ribeiro, J. Heer, and D. Weld, "Errudite: Scalable, reproducible, and testable error analysis," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 747–763.

[62] B. Laughlin, C. Collins, K. Sankaranarayanan, and K. El-Khatib, *A visual analytics framework for adversarial text generation*, 2019. arXiv: 1909.11202 [cs.HC].

[63] L. McInnes, J. Healy, and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. arXiv: 1802.03426 [stat.ML].

[64] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[65] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[66] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[67] H. Hotelling, "Analysis of a complex of statistical variables into principal components.," *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[68] E. Becht *et al.*, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[69] F. O. Bagger, S. Kinalis, and N. Rapin, "Bloodspot: A database of healthy and malignant haematopoiesis updated with purified and single cell mrna sequencing profiles," *Nucleic acids research*, vol. 47, no. D1, pp. D881–D885, 2019.

[70] K. A. Oetjen *et al.*, "Human bone marrow assessment by single-cell rna sequencing, mass cytometry, and flow cytometry," *JCI insight*, vol. 3, no. 23, 2018.

[71] X. Li *et al.*, "Manifold learning of four-dimensional scanning transmission electron microscopy," *npj Computational Materials*, vol. 5, no. 1, p. 5, 2019.

[72] C. Lin, C. Griffith, K. Zhu, and V. Mathur, "Understanding vulnerability of children in surrey," *The University of British Columbia: Vancouver, BC, Canada*, 2018.

[73] D. R. Thompson *et al.*, "Automating x-ray fluorescence analysis for rapid astrobiology surveys," *Astrobiology*, vol. 15, no. 11, pp. 961–976, 2015.

[74] A. Pletl, M. Fernandes, N. Thomas, A. P. Rossi, and B. Elser, "Spectral clustering of crism datasets in jezero crater using umap and k-means," *Remote Sensing*, vol. 15, no. 4, p. 939, 2023.

[75] D. P. Kingma and M. Welling, *Auto-encoding variational bayes*, 2013. arXiv: 1312.6114 [stat.ML].

[76] A. Van Den Oord, O. Vinyals, *et al.*, "Neural discrete representation learning," *Advances in neural information processing systems*, vol. 30, 2017.

[77] G. E. Moran, D. Sridhar, Y. Wang, and D. M. Blei, "Identifiable deep generative models via sparse decoding," *arXiv preprint arXiv:2110.10804*, 2021.

[78] A. J. Gayoso, "Deep generative modeling for single-cell omics data," Ph.D. dissertation, University of California, Berkeley, 2023.

[79] E. N. Weinstein and D. Marks, "A structured observation distribution for generative biological sequence prediction and forecasting," in *International Conference on Machine Learning*, PMLR, 2021, pp. 11 068–11 079.

[80] C. Xu, R. Lopez, E. Mehlman, J. Regier, M. I. Jordan, and N. Yosef, "Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models," *Molecular systems biology*, vol. 17, no. 1, e9620, 2021.

[81] S. Singh, S. Daftry, and R. Capobianco, "Planetary environment prediction using generative modeling," in *AIAA SCITECH 2022 Forum*, 2022, p. 2085.

[82] K. Inkpen, S. Chancellor, M. De Choudhury, M. Veale, and E. P. S. Baumer, "Where is the human? bridging the gap between ai and hci," in *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI EA '19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–9, ISBN: 9781450359719.

[83] M. Gillies *et al.*, "Human-centred machine learning," in *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, ser. CHI EA '16, San Jose, California, USA: Association for Computing Machinery, 2016, pp. 3558–3565, ISBN: 9781450340823.

[84] Y. Gil *et al.*, "Towards human-guided machine learning," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, pp. 614–624.

[85] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman, "Re-examining whether, why, and how human-ai interaction is uniquely difficult to design," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, ser. CHI '20, Honolulu, HI, USA: Association for Computing Machinery, 2020, pp. 1–13, ISBN: 9781450367080.

[86] J. J. Benjamin, A. Berger, N. Merrill, and J. Pierce, "Machine learning uncertainty as a design material: A post-phenomenological inquiry," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, ser. CHI '21, Yokohama, Japan: Association for Computing Machinery, 2021, ISBN: 9781450380966.

[87] S. Davidoff, M. K. Lee, A. K. Dey, and J. Zimmerman, "Rapidly exploring application design through speed dating," in *Proceedings of the 9th International Confer-*

ence on Ubiquitous Computing, ser. UbiComp '07, Innsbruck, Austria: Springer-Verlag, 2007, pp. 429–446, ISBN: 9783540748526.

[88] W. Odom, J. Zimmerman, S. Davidoff, J. Forlizzi, A. K. Dey, and M. K. Lee, "A fieldwork of the future with user enactments," in *Proceedings of the Designing Interactive Systems Conference*, ser. DIS '12, Newcastle Upon Tyne, United Kingdom: Association for Computing Machinery, 2012, pp. 338–347, ISBN: 9781450312103.

[89] Q. Yang, J. Cranshaw, S. Amershi, S. T. Iqbal, and J. Teevan, "Sketching nlp: A case study of exploring the right things to design with language intelligence," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, Glasgow, Scotland Uk: Association for Computing Machinery, 2019, pp. 1–12, ISBN: 9781450359702.

[90] M.-H. Chung, M. Chignell, L. Wang, A. Jovicic, and A. Raman, "Interactive machine learning for data exfiltration detection: Active learning with human expertise," in *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2020, pp. 280–287.

[91] J. A. Fails and D. R. Olsen, "Interactive machine learning," in *Proceedings of the 8th International Conference on Intelligent User Interfaces*, ser. IUI '03, Miami, Florida, USA: Association for Computing Machinery, 2003, pp. 39–45, ISBN: 1581135866.

[92] L. Guo, E. M. Daly, O. Alkan, M. Mattetti, O. Cornec, and B. Knijnenburg, "Building trust in interactive machine learning via user contributed interpretable rules," in *27th International Conference on Intelligent User Interfaces*, 2022, pp. 537–548.

[93] L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *Journal of Visualization*, vol. 22, no. 2, pp. 401–417, 2019.

[94] Z. J. Wang *et al.*, "Interpretability, then what? editing machine learning models to reflect human knowledge and values," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 4132–4142.

[95] K. Kuksenok, "Influence apart from adoption: How interaction between programming and scientific practices shapes modes of inquiry in four oceanography teams," Ph.D. dissertation, 2016.

[96] *Human interface guidelines for machine learning*, https://developer.apple.com/design/human-interface-guidelines/machine-learning/overview/introduction/, Accessed: 2019-10-17.

[97] *People + ai guidebook: Designing human-centered ai products*, https://pair.withgoogle.com/, Accessed: 2019-10-17, May 2019.

[98] S. Amershi *et al.*, "Guidelines for human-ai interaction," in *CHI 2019*, Best Paper Honorable Mention, ACM, May 2019.

[99] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren, "The state of the art in enhancing trust in machine learning models with the use of visualizations," in *Computer graphics forum (Print)*, 2020.

[100] S. Mohseni, N. Zarei, and E. D. Ragan, *A multidisciplinary survey and framework for design and evaluation of explainable ai systems*, 2020. arXiv: 1811.11839 [cs.HC].

[101] F. Sperrle, M. El-Assady, G. Guo, D. H. Chau, A. Endert, and D. Keim, *Should we trust (x)ai? design dimensions for structured experimental evaluations*, 2020. arXiv: 2009.06433 [cs.HC].

[102] M. Issitt and J. Spence, "Practitioner knowledge and the problem of evidence based research policy and practice.," *Youth and policy.*, no. 88, pp. 63–82, Jan. 2005.

[103] N. Das *et al.*, *Bluff: Interactively deciphering adversarial attacks on deep neural networks*, 2020. arXiv: 2009.02608 [cs.LG].

[104] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson, "The what-if tool: Interactive probing of machine learning models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 56–65, 2020.

[105] J. Hullman, "Why authors don't visualize uncertainty," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 130–139, 2020.

[106] E. Wall, J. Stasko, and A. Endert, "Toward a design space for mitigating cognitive bias in vis," in *2019 IEEE Visualization Conference (VIS)*, 2019, pp. 111–115.

[107] Á. A. Cabrera, W. Epperson, F. Hohman, M. Kahng, J. Morgenstern, and D. H. Chau, *Fairvis: Visual analytics for discovering intersectional bias in machine learning*, 2019. arXiv: 1904.05419 [cs.LG].

[108] *Ibm design for ai*, https://www.ibm.com/design/ai/.

[109] *Ethics guidelines for trustworthy ai*, https://ec.europa.eu/futurium/en/ai-alliance-consultation, Apr. 2019.

[110] A. Cutler, M. Pribić, and L. Humphrey, "Everyday ethics for artificial intelligence," *PDF, IBM Corporation*, 2019.

[111] G. LLC, *Perspective api*, https://www.perspectiveapi.com/, 2017.

[112] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The risk of racial bias in hate speech detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 1668–1678.

[113] H. Hosseini, S. Kannan, B. Zhang, and R. Poovendran, *Deceiving google's perspective api built for detecting toxic comments*, 2017. arXiv: 1702.08138 [cs.LG].

[114] S. Jhaver, I. Birman, E. Gilbert, and A. Bruckman, "Human-machine collaboration for content regulation: The case of reddit automoderator," *ACM Trans. Comput.-Hum. Interact.*, vol. 26, no. 5, Jul. 2019.

[115] S. Venkatasubramanian and M. Alfano, "The philosophical basis of algorithmic recourse," 2020.

[116] C. Buni and S. Chemaly, *The secret rules of the internet*, Apr. 2016.

[117] K. Smith, *53 incredible facebook statistics and facts*, 2019.

[118] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai, *Man is to computer programmer as woman is to homemaker? debiasing word embeddings*, 2016. arXiv: 1607.06520 [cs.CL].

[119] T. Manzini, L. Yao Chong, A. W. Black, and Y. Tsvetkov, "Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 615–621.

[120] G. Harrison, J. Hanson, C. Jacinto, J. Ramirez, and B. Ur, "An empirical study on the perceived fairness of realistic, imperfect machine learning models," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 392–402, ISBN: 9781450369367.

[121] K. Sokol and P. Flach, "Explainability fact sheets: A framework for systematic assessment of explainable approaches," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 56–67, ISBN: 9781450369367.

[122] I. D. Raji *et al.*, "Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20, Barcelona, Spain: Association for Computing Machinery, 2020, pp. 33–44, ISBN: 9781450369367.

[123] H. Kaur, H. Nori, S. Jenkins, R. Caruana, H. Wallach, and J. Wortman Vaughan, "Interpreting interpretability: Understanding data scientists' use of interpretability tools for machine learning," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[124] M. A. Madaio, L. Stark, J. Wortman Vaughan, and H. Wallach, "Co-designing checklists to understand organizational challenges and opportunities around fairness in ai," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–14.

[125] T. Gillespie, *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[126] J. Seering, T. Fang, L. Damasco, M. Chen, L. Sun, and G. Kaufman, "Designing user interface elements to improve the quality and civility of discourse in online commenting behaviors," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–14.

[127] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, "" did you suspect the post would be removed?" understanding user reactions to content removals on reddit," *Proceedings of the ACM on human-computer interaction*, vol. 3, no. CSCW, pp. 1–33, 2019.

[128] S. Jhaver, S. Ghoshal, A. Bruckman, and E. Gilbert, "Online harassment and content moderation: The case of blocklists," *ACM Trans. Comput.-Hum. Interact.*, vol. 25, no. 2, Mar. 2018.

[129] S. Jhaver, A. Bruckman, and E. Gilbert, "Does transparency in moderation really matter?: User behavior after content removal explanations on reddit," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, 2019.

[130] J. Seering, "Reconsidering community self-moderation: The role of research in supporting community-based models for online content moderation," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Oct. 2020.

[131] K. Long *et al.*, ""could you define that in bot terms"? requesting, creating and using bots on reddit," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ser. CHI '17, Denver, Colorado, USA: Association for Computing Machinery, 2017, pp. 3488–3500, ISBN: 9781450346559.

[132] E. Chandrasekharan, C. Gandhi, M. W. Mustelier, and E. Gilbert, "Crossmod: A cross-community learning-based system to assist reddit moderators," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019.

[133] R. S. Geiger and D. Ribes, "The work of sustaining order in wikipedia: The banning of a vandal," in *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, ser. CSCW '10, Savannah, Georgia, USA: Association for Computing Machinery, 2010, pp. 117–126, ISBN: 9781605587950.

[134] P. Burnap and M. L. Williams, "Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making," *Policy & Internet*, vol. 7, no. 2, pp. 223–242, 2015.

[135] N. Suzor, T. Van Geelen, and S. Myers West, "Evaluating the legitimacy of platform governance: A review of research and a shared research agenda," *International Communication Gazette*, vol. 80, no. 4, pp. 385–400, 2018.

[136] S. Chancellor, J. A. Pater, T. Clear, E. Gilbert, and M. De Choudhury, "# thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities," in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, ser. CSCW '16, San Francisco, California, USA: Association for Computing Machinery, 2016, pp. 1201–1213, ISBN: 9781450335928.

[137] J. N. Matias and M. Mou, "Civilservant: Community-led experiments in platform governance," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, Montreal QC, Canada: Association for Computing Machinery, 2018, pp. 1–13, ISBN: 9781450356206.

[138] A. L. Strachan, "Interventions to counter hate speech," *GSDRC Applied Research Services*, vol. 23, 2014.

[139] B. Mathew *et al.*, "Thou shalt not hate: Countering online hate speech," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, 2019, pp. 369–380.

[140] J. Cai and D. Y. Wohn, "What are effective strategies of handling harassment on twitch? users' perspectives," in *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*, ser. CSCW '19, Austin, TX, USA: Association for Computing Machinery, 2019, pp. 166–170, ISBN: 9781450366922.

[141] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, and A. Vakali, *Mean birds: Detecting aggression and bullying on twitter*, 2017. arXiv: 1702.06877 [cs.CY].

[142] M. Filippo *et al.*, "Misogynistic language on twitter and sexual violence," 2015.

[143] K. Mahar, A. X. Zhang, and D. Karger, "Squadbox: A tool to combat email harassment using friendsourced moderation," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–13.

[144] S. H. Taylor, D. DiFranzo, Y. H. Choi, S. Sannon, and N. N. Bazarova, "Accountability and empathy by design: Encouraging bystander intervention to cyberbullying on social media," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, no. CSCW, Nov. 2019.

[145] C. E. Smith, B. Yu, A. Srivastava, A. Halfaker, L. Terveen, and H. Zhu, "Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1–14, ISBN: 9781450367080.

[146] D. K. Citron and H. Norton, "Intermediaries and hate speech: Fostering digital citizenship for our information age," *BUL Rev.*, vol. 91, p. 1435, 2011.

[147] A. Vaswani *et al.*, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[148] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[149] A. Ettinger, *What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models*, 2019. arXiv: 1907.13528 [cs.CL].

[150] R. Pryzant, R. D. Martinez, N. Dass, S. Kurohashi, D. Jurafsky, and D. Yang, *Automatically neutralizing subjective bias in text*, 2019. arXiv: 1911.09709 [cs.CL].

[151] C. Hube and B. Fetahu, "Neural based statement classification for biased language," in *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, ser. WSDM '19, Melbourne VIC, Australia: Association for Computing Machinery, 2019, pp. 195–203, ISBN: 9781450359405.

[152] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, *Adversarial attacks on deep learning models in natural language processing: A survey*, 2019. arXiv: 1901.06796 [cs.CL].

[153] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, *How can we fool lime and shap? adversarial attacks on post hoc explanation methods*, 2019. arXiv: 1911.02508 [cs.LG].

[154] R. H. Thaler and C. R. Sunstein, *Nudge: Improving decisions about health, wealth, and happiness*. Penguin, 2009.

[155] J. Li, S. Ji, T. Du, B. Li, and T. Wang, "Textbugger: Generating adversarial text against real-world applications," *Proceedings 2019 Network and Distributed System Security Symposium*, 2019.

[156] M. Sundararajan, A. Taly, and Q. Yan, *Axiomatic attribution for deep networks*, 2017. arXiv: 1703.01365 [cs.LG].

[157] S. Wiegreffe and Y. Pinter, *Attention is not not explanation*, 2019. arXiv: 1908.04626 [cs.CL].

[158] K. Clark, U. Khandelwal, O. Levy, and C. D. Manning, "What does bert look at? an analysis of bert's attention," *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, 2019.

[159] S. Serrano and N. A. Smith, "Is attention interpretable?" In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 2931–2951.

[160] H. Valdivieso, D. Parra, A. Carvallo, G. Rada, K. Verbert, and T. Schreck, *Analyzing the design space for visualizing neural attention in text classification*, 2019.

[161] R. JeffreyPennington and C. Manning, "Glove: Global vectors for word representation," in *Conference on Empirical Methods in Natural Language Processing*, Citeseer, 2014.

[162] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, ser. NIPS'13, Lake Tahoe, Nevada: Curran Associates Inc., 2013, pp. 3111–3119.

[163] E. R. Tufte, *The visual display of quantitative information*. Graphics Press, 2018.

[164] Kaggle, *Jigsaw toxicity dataset*, 2017.

[165] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006, ROC Analysis in Pattern Recognition.

[166] A. Paszke *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., Curran Associates, Inc., 2019, pp. 8024–8035.

[167] M. G. Kendall, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[168] M. Strathern, "'improving ratings': Audit in the british university system," *European Review*, vol. 5, pp. 305–321, 1997.

[169] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep learning for anomaly detection: A review," *ACM Comput. Surv.*, vol. 54, no. 2, Mar. 2021.

[170] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, eaay7120, 2019.

[171] C. F. Chyba and K. P. Hand, "Astrobiology: The study of the living universe," *Annual Review of Astronomy and Astrophysics*, vol. 43, no. 1, pp. 31–74, 2005.

[172] A. G. Fairén *et al.*, "Astrobiology through the ages of mars: The study of terrestrial analogues to understand the habitability of mars," *Astrobiology*, vol. 10, no. 8, pp. 821–843, 2010.

[173] J. F. Banfield, J. W. Moreau, C. S. Chan, S. A. Welch, and B. Little, "Mineralogical biosignatures and the search for life on mars," *Astrobiology*, vol. 1, no. 4, pp. 447–465, 2001, PMID: 12448978. eprint: https://doi.org/10.1089/153110701753593856.

[174] J. L. Bishop, E. Murad, M. D. Lane, and R. L. Mancinelli, "Multiple techniques for mineral identification on mars:: A study of hydrothermal rocks as potential analogues for astrobiology sites on mars," *Icarus*, vol. 169, no. 2, pp. 311–323, 2004.

[175] S. McMahon *et al.*, "A field guide to finding fossils on mars," *Journal of Geophysical Research: Planets*, vol. 123, no. 5, pp. 1012–1040, 2018. eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2017JE005478.

[176] A. C. Allwood *et al.*, "Pixl: Planetary instrument for x-ray lithochemistry," *Space Science Reviews*, vol. 216, no. 8, pp. 1–132, 2020.

[177] R. Rieder *et al.*, "The chemical composition of martian soil and rocks returned by the mobile alpha proton x-ray spectrometer: Preliminary results from the x-ray mode," *Science*, vol. 278, no. 5344, pp. 1771–1774, 1997.

[178] R. Gellert *et al.*, "Alpha particle x-ray spectrometer (apxs): Results from gusev crater and calibration report," *Journal of Geophysical Research: Planets*, vol. 111, no. E2, 2006.

[179] S. J. VanBommel *et al.*, "Deconvolution of distinct lithology chemistry through oversampling with the mars science laboratory alpha particle x-ray spectrometer," *X-Ray Spectrometry*, vol. 45, no. 3, pp. 155–161, 2016.

[180] D. Wixon, K. Holtzblatt, and S. Knox, "Contextual design: An emergent view of system design," in *CHI*, 1990, pp. 329–336.

[181] B. Beckhoff, B. Kanngießer, N. Langhoff, R. Wedell, and H. Wolff, *Handbook of practical X-ray fluorescence analysis*. Springer Science & Business Media, 2007.

[182] D. Schurman *et al.*, "Pixelate: Novel visualization and computational methods for the analysis of astrobiological spectroscopy data," American Geophysical Union, Jun. 2019.

[183] M. Haschke, "Laboratory micro-x-ray fluorescence spectroscopy," *Cham: Springer International Publishing*, vol. 10, pp. 978–983, 2014.

[184] W. T. Elam, B. D. Ravel, and J. Sieber, "A new atomic database for x-ray spectroscopic calculations," *Radiation Physics and Chemistry*, vol. 63, no. 2, pp. 121–128, 2002.

[185] C. M. Heirwegh, E. W. T., and L. P. O'Neil, "The focused beam x-ray fluorescence elemental quantification software package piquant," *Spectrochimica Acta Part B: Atomic Spectroscopy*, vol. 196, p. 106 520, 2022.

[186] S. W. Ruff and J. D. Farmer, "Silica deposits on mars with features resembling hot spring biosignatures at el tatio in chile," *Nature Communications*, vol. 7, no. 1, p. 13 554, Nov. 2016.

[187] D. L. Bish *et al.*, "X-ray diffraction results from mars science laboratory: Mineralogy of rocknest at gale crater," *science*, vol. 341, no. 6153, p. 1 238 932, 2013.

[188] A. Allwood *et al.*, "Texture-specific elemental analysis of rocks and soils with pixl: The planetary instrument for x-ray lithochemistry on mars 2020," in *2015 IEEE Aerospace Conference*, 2015, pp. 1–13.

[189] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[190] E. B.-N. Sanders and P. J. Stappers, "Co-creation and the new landscapes of design," *Co-design*, vol. 4, no. 1, pp. 5–18, 2008.

[191] D. Flannery *et al.*, "Increasing efficiency of mars 2020 rover operations via novel data analysis software for the planetary instrument for x-ray lithochemistry (pixl)," *Proceedings of the 2021 Committee on Space Research (COSPAR) Scientific Assembly*, vol. 43, B0.2 2021.

[192] C. Ye, L. Hermann, S. Bhat, N. Yildirim, D. Moritz, and S. Davidoff, "Pixlise-c: Exploring the data analysis needs of nasa scientists for mineral identification," in *ACM Conference on Human Factors in computing systems Workshop on Human-Computer Interaction for Space Exploration*, ser. SpaceCHI 2021, Honolulu, HI, USA: Association for Computing Machinery, 2021, ISBN: 9781450367080.

[193] NASA, *Pixlise application*, 2021.

[194] NASA. "Pixl sol 140." (2021).

[195] NASA. "Pixl's view of dourbes." (2021).

[196] NASA. "Two perspectives of séítah rocks." (2021).

[197] P. Nemere, R. Stonebraker, A. Galvin, T. Barber, S. M. Fedell, and S. Davidoff, *Pixlise/pixlise-ui: Release 2.0.16*, version v2.0.16-ui, Jan. 2023.

[198] A. P. Wright, P. Nemere, R. Stonebraker, A. Galvin, and S. Davidoff, *pixlise/diffraction-peak-detection: 2.0 open source migration release*, version v2.0, Aug. 2022.

[199] H. Reichenbach, "Experience and prediction: An analysis of the foundations and the structure of knowledge," 1938.

[200] K. A. Farley *et al.*, "Mars 2020 mission overview," *Space Science Reviews*, vol. 216, no. 8, p. 142, Dec. 2020.

[201] A. C. Allwood *et al.*, "Pixl: Planetary instrument for x-ray lithochemistry," *Space Science Reviews*, vol. 216, no. 8, pp. 1–132, 2020.

[202] I. Brigandt and A. Love, "Reductionism in Biology," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Summer 2023, Metaphysics Research Lab, Stanford University, 2023.

[203] C. Craver and J. Tabery, "Mechanisms in Science," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Fall 2023, Metaphysics Research Lab, Stanford University, 2023.

[204] D. Stoljar, "Physicalism," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Spring 2024, Metaphysics Research Lab, Stanford University, 2024.

[205] R. van Riel and R. Van Gulick, "Scientific Reduction," in *The Stanford Encyclopedia of Philosophy*, E. N. Zalta and U. Nodelman, Eds., Winter 2023, Metaphysics Research Lab, Stanford University, 2023.

[206] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.

[207] E. Bingham *et al.*, "Pyro: Deep Universal Probabilistic Programming," *Journal of Machine Learning Research*, 2018.

[208] A. Ramdas, N. García Trillos, and M. Cuturi, "On wasserstein two-sample testing and related families of nonparametric tests," *Entropy*, vol. 19, no. 2, p. 47, 2017.

[209] L. McInnes, J. Healy, S. Astels, *et al.*, "Hdbscan: Hierarchical density based clustering.," *J. Open Source Softw.*, vol. 2, no. 11, p. 205, 2017.

[210] O. of National Drug Control Policy, *National drug control strategy*, 2020.

[211] W.-b. I. S. Query, "Reporting system (wisqars). 2015," *Centers for Disease Control and Prevention, National Center for Injury Prevention and Control. Available from: http://www.cdc.gov/injury/wisqars/index. html [Accessed 14 9 2016]*, 2019.

[212] S. L. Murphy, K. D. Kochanek, J. Xu, and E. Arias, "Mortality in the united states, 2020," 2021.

[213] J. Xu, S. L. Murphy, K. D. Kochanek, and E. Arias, "Mortality in the united states, 2015. nchs data brief, no 267. hyattsville, md: Us department of health and human services, cdc," *National Center for Health Statistics*, 2016.

[214] K. D. Kochanek, S. L. Murphy, J. Q. Xu, and E. Arias, "Mortality in the united states, 2016. nchs data brief, no 293," *National Center for Health Statistics*, 2017.

[215] D. Dowell *et al.*, "Contribution of opioid-involved poisoning to the change in life expectancy in the united states, 2000-2015," *Jama*, vol. 318, no. 11, pp. 1065–1067, 2017.

[216] W. M. Compton and C. M. Jones, "Epidemiology of the us opioid crisis: The importance of the vector," *Annals of the New York Academy of Sciences*, vol. 1451, no. 1, pp. 130–143, 2019.

[217] H. Jalal, J. M. Buchanich, M. S. Roberts, L. C. Balmert, K. Zhang, and D. S. Burke, "Changing dynamics of the drug overdose epidemic in the united states from 1979 through 2016," *Science*, vol. 361, no. 6408, eaau1184, 2018.

[218] N. Wilson, "Drug and opioid-involved overdose deaths—united states, 2017–2018," *MMWR. Morbidity and mortality weekly report*, vol. 69, 2020.

[219] H. Hedegaard, B. A. Bastian, J. P. Trinidad, M. Spencer, and M. Warner, "Regional differences in the drugs most frequently involved in drug overdose deaths: United states, 2017," 2019.

[220] *Duragesic prescribing information. food and drug administration.* https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/019813s079lbl.pdf.

[221] *Fentora prescribing information. food and drug administration.* https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/021947s029lbl.pdf.

[222] *Fentanyl citrate prescribing information. food and drug administration*, https://www.accessdata.fda.gov/drugsatfda_docs/label/2019/019115s033lbl.pdf.

[223] S. F. Butler, R. A. Black, T. A. Cassidy, T. M. Dailey, and S. H. Budman, "Abuse risks and routes of administration of different prescription opioid compounds and formulations," *Harm reduction journal*, vol. 8, pp. 1–17, 2011.

[224] M. R. Spencer, M. Warner, B. A. Bastian, J. P. Trinidad, and H. Hedegaard, "Drug overdose deaths involving fentanyl, 2011-2016.," *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System*, vol. 68, no. 3, pp. 1–19, 2019.

[225] *Nflis brief: Fentanyl, 2001–2015. u.s. department of justice, drug enforcement agency, diversion control division*, https://www.nflis.deadiversion.usdoj.gov/DesktopModules/ReportDow

[226] R. Gladden and P. Seth, "Trends in deaths involving heroin and synthetic opioids excluding methadone, and law enforcement drug product reports, by census region-united states, 2006-2015," *MMWR Morb Mortal Wkly Rep*, vol. 66, no. 34, pp. 897–903, 2017.

[227] https://wonder.cdc.gov/mcd.html.

[228] D. A. Bowen, J. O'Donnell, and S. A. Sumner, "Increases in online posts about synthetic opioids preceding increases in synthetic opioid death rates: A retrospective observational study," *Journal of general internal medicine*, vol. 34, pp. 2702–2704, 2019.

[229] A. Sarker *et al.*, "Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter," *Drug safety*, vol. 39, pp. 231–240, 2016.

[230] J. Kalyanam, T. Katsuki, G. R. Lanckriet, and T. K. Mackey, "Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the twittersphere using unsupervised machine learning," *Addictive behaviors*, vol. 65, pp. 289–295, 2017.

[231] T. Katsuki, T. K. Mackey, and R. Cuomo, "Establishing a link between prescription drug abuse and illicit online pharmacies: Analysis of twitter data," *Journal of medical Internet research*, vol. 17, no. 12, e280, 2015.

[232] T. K. Mackey, J. Kalyanam, T. Katsuki, and G. Lanckriet, "Twitter-based detection of illegal online sale of prescription opioid," *American journal of public health*, vol. 107, no. 12, pp. 1910–1915, 2017.

[233] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L. S. Nelson, and A. F. Manini, "Epidemiology from tweets: Estimating misuse of prescription opioids in the usa from social media," *Journal of Medical Toxicology*, vol. 13, pp. 278–286, 2017.

[234] A. Sarker, G. Gonzalez-Hernandez, Y. Ruan, and J. Perrone, "Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter," *JAMA network open*, vol. 2, no. 11, e1914672, 2019.

[235] S. A. Sumner, T. M. Haegerich, and C. M. Jones, "Temporal trends in online posts about vaping of cannabis products," *Journal of addiction medicine*, vol. 15, no. 2, pp. 173–174, 2021.

[236] A. Sarker *et al.*, "Utilizing social media data for pharmacovigilance: A review," *Journal of biomedical informatics*, vol. 54, pp. 202–212, 2015.

[237] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Cultural shift or linguistic drift? comparing two computational measures of semantic change," in *Proceedings of the conference on empirical methods in natural language processing. Conference on empirical methods in natural language processing*, vol. 2016, 2016, p. 2116.

[238] W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Diachronic word embeddings reveal statistical laws of semantic change," *arXiv preprint arXiv:1605.09096*, 2016.

[239] N. Garg, L. Schiebinger, D. Jurafsky, and J. Zou, "Word embeddings quantify 100 years of gender and ethnic stereotypes," *Proceedings of the National Academy of Sciences*, vol. 115, no. 16, E3635–E3644, 2018.

[240] C. Nguyen, "Reddit beats out facebook to become the third-most-popular site on the web," *Retrieved September*, vol. 15, p. 2019, 2018.

[241] *Google cloud. bigquery public datasets*, https://cloud.google.com/bigquery/public-data.

[242] *Pushshift*, https://pushshift.io/.

[243] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[244] S. Chancellor, G. Nitzburg, A. Hu, F. Zampieri, and M. De Choudhury, "Discovering alternative treatments for opioid use recovery using social media," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, 2019, pp. 1–15.

[245] *Drug facts. drug enforcement agency*, https://www.dea.gov/factsheets.

[246] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[247] M. J. Paul, M. S. Chisolm, M. W. Johnson, R. G. Vandrey, and M. Dredze, "Assessing the validity of online drug forums as a source for estimating demographic and temporal trends in drug use," *Journal of addiction medicine*, vol. 10, no. 5, pp. 324–330, 2016.

[248] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L. S. Nelson, and A. F. Manini, "Epidemiology from tweets: Estimating misuse of prescription opioids in the usa from social media," *Journal of Medical Toxicology*, vol. 13, pp. 278–286, 2017.

[249] M. C. Mercado, S. A. Sumner, M. B. Spelke, M. K. Bohm, D. E. Sugerman, and C. Stanley, "Increase in drug overdose deaths involving fentanyl—rhode island, january 2012–march 2014," *Pain medicine*, vol. 19, no. 3, pp. 511–523, 2018.

[250] C. M. Jones, "Patterns and characteristics of methamphetamine use among adults—united states, 2015–2018," *MMWR. Morbidity and mortality weekly report*, vol. 69, 2020.

[251] M. Kariisa, "Drug overdose deaths involving cocaine and psychostimulants with abuse potential—united states, 2003–2017," *MMWR. Morbidity and mortality weekly report*, vol. 68, 2019.

Appendices

APPENDIX A

FULL SET OF HAI GUIDELINES

Category Level 1	Category Level 2	Company	Guideline
Initial Considerations	Value of AI	Google	Find the intersection of user needs & AI strengths.
Initial Considerations	Value of AI	Google	Balance control & automation.
Initial Considerations	Value of AI	Google	Assess automation vs. augmentation
Initial Considerations	Value of AI	Google	Align perceived and actual user value
Initial Considerations	Value of AI	Google	Account for situational stakes
Initial Considerations	Fairness	Google	Consider bias in the data collection and evaluation process
Initial Considerations	Fairness	Google	Assess inclusivity
Initial Considerations	Fairness	Google	Use data that applies to different groups of users
Initial Considerations	Fairness	Google	Commit to fairness
Initial Considerations	Fairness	Microsoft	Match relevant social norms.
Initial Considerations	Fairness	Microsoft	Mitigate social biases.
Initial Considerations	Fairness	Google	Ensure rater pool diversity
Initial Considerations	Privacy	Google	Is there a risk of inadvertently revealing user data? What would the consequence be?
Initial Considerations	Privacy	Google	Protect personally identifiable information
Initial Considerations	Privacy	Google	Understand when people want to maintain control
Initial Considerations	Privacy	Google	Understand when people will give up control
Initial Considerations	Privacy	Google	What limits exist around user consent for data use
Initial Considerations	Privacy	Google	Return control to the user
Initial Considerations	Privacy	Google	Manage privacy and security
Initial Considerations	Privacy	Apple	Help people control their information
Initial Considerations	Privacy	Apple	Always secure people's information
Initial Considerations	Privacy	Apple	Collect only the most essential information.
Initial Considerations	Privacy	Apple	Be clear about why you need people's information
Initial Considerations	Privacy	Apple	Consider withholding private or sensitive suggestions
Initial Considerations	Privacy	Microsoft	Provide global controls

Category Level 1	Category Level 2	Company	Guideline
Model	How to train your model	Google	Design for experimentation
Model	How to train your model	Google	Inspect the features possible values, units, and data types
Model	How to train your model	Google	Evaluate the reward function outcomes
Model	How to train your model	Google	Weigh false positive & negative
Model	How to train your model	Google	Consider precision and recall tradeoffs
Model	How to train your model	Google	Balance underfitting and overfitting
Model	How to train your model	Google	Tune your model
Model	How to train your model	Google	Map existing workflows
Model	How to train your model	Google	Design and evaluate the reward function
Model	How to train your model	Google	User needs and defining success
Model	How to train your model	Google	Design for model tuning
Model	Training Data	Google	Review how often your data sources are refreshed
Model	Training Data	Google	Collect live data from users
Model	Training Data	Google	Provide easy access to labels
Model	Training Data	Google	Use existing dataset
Model	Training Data	Google	Translate user needs into data needs
Model	Training Data	Google	Only introduce new features when needed
Model	Training Data	Google	Data collection + evaluation
Model	Training Data	Google	Identify your data sources
Model	Training Data	Google	Identify any outliers, and investigate whether they are actual outliers or due to errors in the data
Model	Training Data	Google	Source your data responsibly
Model	Training Data	Google	Build your own dataset
Model	Training Data	Google	Design for raters and labeling
Model	Training Data	Google	Split your data
Model	Training Data	Google	Let raters change their minds
Model	Training Data	Google	Evaluate rater tools
Model	Training Data	Google	Missing or incomplete data
Model	Training Data	Google	Unexpected input
Model	Training Data	Google	Investigate rater context and incentives
Model	Training Data	Google	Articulate data sources
Model	Training Data	Apple	Beware of confirmation bias

Category Level 1	Category Level 2	Company	Guideline
Interface	Explainability	Google	Explain the benefit, not the technology
Interface	Explainability	Google	Use simple, direct language to describe each explicit feedback option and its consequences
Interface	Explainability	Google	Optimize for understanding
Interface	Explainability	Google	Explainability + Trust
Interface	Explainability	Google	Note special cases of absent or comprehensive explanation
Interface	Explainability	Google	Explanation via interaction
Interface	Explainability	Google	Example-based explanations
Interface	Explainability	Google	Explain what's important
Interface	Explainability	Google	Tie explanations to user actions
Interface	Explainability	Apple	In general, avoid technical or statistical jargon
Interface	Explainability	Apple	Avoid being too specific or too general
Interface	Explainability	Microsoft	Show contextually relevant information
Interface	Confidence	Google	Model confidence displays
Interface	Confidence	Google	Decide how best to show model confidence
Interface	Confidence	Google	Categorical
Interface	Confidence	Google	N-best alternatives
Interface	Confidence	Google	Numeric
Interface	Confidence	Google	Determine if you should show confidence
Interface	Confidence	Apple	When you know that confidence values correspond to result quality, you generally want to avoid showing results when confidence is low.
Interface	Confidence	Apple	Consider changing how you present results based on different confidence thresholds
Interface	Confidence	Apple	In general, translate confidence values into concepts that people already understand.
Interface	Confidence	Apple	Know what your confidence values mean before you decide how to present them
Interface	Confidence	Apple	In scenarios where people expect statistical or numerical information, display confidence values that help them interpret the results.
Interface	Confidence	Apple	Confirm success

Category Level 1	Category Level 2	Company	Guideline
Interface	Expectations / Mental Models	Google	Onboard in stages.
Interface	Expectations / Mental Models	Google	Help users calibrate their trust.
Interface	Expectations / Mental Models	Google	Introduce and set expectations for AI
Interface	Expectations / Mental Models	Google	Set expectations for AI improvements
Interface	Expectations / Mental Models	Google	Account for timing in the user journey
Interface	Expectations / Mental Models	Google	Keep track of user needs
Interface	Expectations / Mental Models	Google	Identify existing mental models
Interface	Expectations / Mental Models	Google	Clearly communicate AI limits and capabilities
Interface	Expectations / Mental Models	Google	Set expectations for adaptation.
Interface	Expectations / Mental Models	Google	Describe the system or explain the output
Interface	Expectations / Mental Models	Google	Account for user expectations of human-like interaction.
Interface	Expectations / Mental Models	Apple	Consider using attributions to help people distinguish among results.
Interface	Expectations / Mental Models	Apple	Keep attributions factual and based on objective analysis.
Interface	Expectations / Mental Models	Apple	Help people establish realistic expectations.
Interface	Expectations / Mental Models	Apple	Explain how limitations can cause unsatisfactory results
Interface	Expectations / Mental Models	Apple	Consider telling people when limitations are resolved
Interface	Expectations / Mental Models	Apple	Demonstrate how to get the best results
Interface	Expectations / Mental Models	Microsoft	Make clear what the system can do.
Interface	Expectations / Mental Models	Microsoft	Make clear how well the system can do what it can do
Interface	Expectations / Mental Models	Microsoft	Make clear why the system did what it did.
Interface	Expectations / Mental Models	Microsoft	Convey the consequences of user actions.
Interface	Expectations / Mental Models	Microsoft	Notify users about changes.
Interface	Expectations / Mental Models	Microsoft	Scope services when in doubt.
Interface	Expectations / Mental Models	Microsoft	Time services based on context.

Category Level 1	Category Level 2	Company	Guideline
Interface	Calibration	Apple	Avoid asking people to participate in calibration more than once.
Interface	Calibration	Apple	Make calibration quick and easy
Interface	Calibration	Apple	Make sure people know how to perform calibration successfully.
Interface	Calibration	Apple	Let people cancel calibration at any time.
Interface	Calibration	Apple	Give people a way to update or remove information they provided during calibration.
Interface	Calibration	Apple	Always secure people's calibration information
Interface	Multiple Options	Google	Categorical / N-Best Alternatives
Interface	Multiple Options	Google	Consider Formatting
Interface	Multiple Options	Google	Use multiple shortcuts to optimize key flows
Interface	Multiple Options	Apple	Whenever possible, help people make decisions by conveying confidence in terms of actionable suggestions.
Interface	Multiple Options	Apple	List the most likely option first.
Interface	Multiple Options	Apple	In situations where attributions aren't helpful, consider ranking or ordering the results in a way that implies confidence levels
Interface	Multiple Options	Apple	Consider offering multiple options when requesting explicit feedback.
Interface	Multiple Options	Apple	In general, avoid providing too many options
Interface	Multiple Options	Apple	Prefer diverse options
Interface	Multiple Options	Apple	Make options easy to distinguish and choose
Interface	Multiple Options	Apple	Add iconography to an option description if it helps people understand it.

Category Level 1	Category Level 2	Company	Guideline
Deployment	Error Prevention	Google	Account for negative impact
Deployment	Error Prevention	Google	Auto-detect and display errors
Deployment	Error Prevention	Google	Disambiguate systems hierarchy errors
Deployment	Error Prevention	Google	Diagnose errors that users don't perceive
Deployment	Error Prevention	Google	Check output quality for relevance errors
Deployment	Error Prevention	Google	Fail gracefully
Deployment	Error Prevention	Google	Discover prediction and training data errors
Deployment	Error Prevention	Google	Cue the correct interactions
Deployment	Error Prevention	Google	Categorize user-perceived errors
Deployment	Error Prevention	Google	Provide paths forward from failure
Deployment	Error Prevention	Apple	Understand the significance of a mistake's consequences
Deployment	Error Prevention	Apple	As you work on reducing mistakes in one area, always consider the effect your work has on other areas and overall accuracy
Deployment	Error Prevention	Apple	When possible, address mistakes without complicating the UI
Deployment	Error Prevention	Apple	Learn from corrections when it makes sense
Deployment	Error Prevention	Apple	When possible, use guided corrections instead of freeform corrections
Deployment	Error Prevention	Apple	Let people correct their corrections
Deployment	Error Prevention	Apple	Provide immediate value when people make a correction
Deployment	Error Prevention	Apple	Give people familiar easy ways to make corrections
Deployment	Error Prevention	Apple	Immediately provide assistance if progress stalls
Deployment	Error Prevention	Apple	Be especially careful to avoid mistakes in proactive features
Deployment	Error Prevention	Microsoft	Support efficient dismissal
Deployment	Error types	Google	Identify error sources.
Deployment	Error types	Google	Background errors.
Deployment	Error types	Google	Context errors
Deployment	Error types	Google	Mislabeled or misclassified results
Deployment	Error types	Google	Poor inference or incorrect model
Deployment	Error Handling	Google	Assume subversive use
Deployment	Error Handling	Google	Imagine potential pitfalls
Deployment	Error Handling	Google	Gauge the risk for potential errors
Deployment	Error Handling	Google	Identify user, system, and context errors
Deployment	Error Handling	Google	Weigh situational stakes and error risk
Deployment	Error Handling	Google	Avoid compounding errors from other ML models
Deployment	Error Handling	Google	Define "errors" and "failure"
Deployment	Error Handling	Google	Predict or plan for input errors
Deployment	Error Handling	Apple	Never rely on corrections to make up for low-quality results
Deployment	Error Handling	Apple	Always balance the benefits of a feature with the effort required to make a correction

Category Level 1	Category Level 2	Company	Guideline
Deployment	Collecting Feedback	Google	Collect explicit feedback.
Deployment	Collecting Feedback	Google	Monitor over time.
Deployment	Collecting Feedback	Google	Allow for opting out.
Deployment	Collecting Feedback	Google	Plan for co-learning.
Deployment	Collecting Feedback	Google	Connect feedback with personalization.
Deployment	Collecting Feedback	Google	Create opportunities for feedback.
Deployment	Collecting Feedback	Google	Provide editability.
Deployment	Collecting Feedback	Apple	Be prepared for changes in implicit feedback when you make changes to your app's UI.
Deployment	Collecting Feedback	Apple	Don't ask for both positive and negative feedback.
Deployment	Collecting Feedback	Apple	Make it easy for people to correct frequent or predictable mistakes.
Deployment	Collecting Feedback	Apple	Always make providing explicit feedback a voluntary task.
Deployment	Collecting Feedback	Apple	Request explicit feedback only when necessary.
Deployment	Collecting Feedback	Apple	Consider using explicit feedback to help improve when and where you show results.
Deployment	Collecting Feedback	Microsoft	Remember recent interactions.
Deployment	Collecting Feedback	Microsoft	Encourage granular feedback.
Deployment	Collecting Feedback	Microsoft	Support efficient correction.
Deployment	Collecting Feedback	Microsoft	Learn from user behavior.

Category Level 1	Category Level 2	Company	Guideline
Deployment	Addressing / using Feedback	Google	Review implicit feedback.
Deployment	Addressing / using Feedback	Google	Adapt to the evolving user journey.
Deployment	Addressing / using Feedback	Google	Remind, reinforce, and adjust.
Deployment	Addressing / using Feedback	Google	Communicate value and time to impact.
Deployment	Addressing / using Feedback	Google	Align feedback with model improvement.
Deployment	Addressing / using Feedback	Google	Manage influence on user decisions.
Deployment	Addressing / using Feedback	Google	Connect feedback to user experience changes.
Deployment	Addressing / using Feedback	Apple	When possible, use multiple feedback signals to improve suggestions and mitigatie mistakes.
Deployment	Addressing / using Feedback	Apple	Prioritize recent feedback.
Deployment	Addressing / using Feedback	Apple	Learn from selections when it makes sense.
Deployment	Addressing / using Feedback	Apple	Update and adapt cautiously.
Deployment	Addressing / using Feedback	Apple	Use feedback to update predictions on a cadence that matches the user's mental model of the feature.
Deployment	Addressing / using Feedback	Apple	Act immediately when you receive explicit feedback and persist the resulting changes.
Deployment	Addressing / using Feedback	Apple	Don't let implicit feedback decrease people's opportunities to explore.
Deployment	Addressing / using Feedback	Microsoft	Continuously update your feature to reflect people's evolving interests and preferences.

VITA

Austin P. Wright was born on July 12, 1996 to parents Heather Smith and David Wright. In 2018 he earned his BA in Physics and Computer Science from the University of California, Berkeley and in 2019 he earned his Msc in Machine Learning from Imperial College London. Since 2019 his research at Georgia Tech and in collaboration with NASA JPL and the CDC has earned acclaim and awards both at home and abroad. He enjoys spending quite time alongside his partner Arielle Spencer and their three cats: Pierre, Marie, and Ada.