Fred Hohman Research Statement

Interactive Scalable Interfaces for Machine Learning Interpretability

Data-driven machine learning paradigms now solve the world's hardest problems by automatically learning from data. Unfortunately, what is learned is often unknown to both the people who train models and the people they impact. This has led to a rallying cry for *interpretability*. But what is interpretability? How do we scale up explanations for modern, complex models? And how can we best communicate them to people?

Through my experience working closely with researchers, designers, and practitioners at Apple Machine Intelligence, Microsoft Research, NASA JPL Human Interfaces, and Pacific Northwest National Lab, it is clear applying machine learning is a people problem. From data collection to model deployment, people make decisions using models that impact people. Therefore, I take a *human-centered approach* to machine learning interpretability by **designing and developing interactive interfaces to enable interpretability at scale and for everyone**.

My thesis **operationalizes** interpretability in practice (GAMUT [CHI2019a], TELEGAM [VIS2019a]), **scales** explanations to complex models (Survey [TVCG2018], SUMMIT [TVCG20]), and **communicates** explanations to people (Interactive Articles [PP19, VISxAI18, VisComm19, Distill20]). My interdisciplinary research contributes to human-computer interaction, machine learning, and more importantly *their intersection*, including **open-source interactive interfaces, scalable algorithms, and new, accessible communication artifacts**.

Figure 1: An overview of my interdisciplinary research where I design and develop interactive interfaces to **enable** interpretability **at scale** and **for everyone**.

My work thus far has resulted in **23 peer-reviewed publications** (11 conference; 12 workshop, poster, and demo papers) at the premiere venues within human-computer interaction (CHI) and visualization (VIS, TVCG). I have won a Best Paper award at CHI 2019, a **NASA PhD Fellowship**, the **Georgia Tech President's Fellowship**, and a **Microsoft AI for Earth Award**. My work has made significant impact to industry and society: my visualization systems have been deployed at Microsoft, demoed to executive leadership at their internal TechFest, and inspired the visualization design of the widely-used InterpretML interpretability toolkit (2.4K+ Github stars); my visualization explanations scale to large models and datasets (e.g., InceptionNet + ImageNet with 1.2M images); and my interactive articles have been read by 250,000+ people.

Operationalizing Machine Learning Interpretability

In machine learning, practitioners risk making decisions based on hidden biases, spurious correlations, and false generalizations. Discovering such problems is the purpose and promise of machine learning interpretability. Yet the concept of interpretability remains nebulous, such that people lack guidance for how to incorporate interpretability into models and tools.

GAMUT: Understanding How Practitioners Understand Models

To solve these problems, I sought to *operationalize interpretability*, creating a collection of capabilities interactive interfaces should support to enable interpretability in practice. Through an iterative design process with 9 expert machine learning researchers and practitioners at Microsoft, I designed GAMUT [CHI19a], a first-of-its-kind visual analytics system that instantiated the operationalization (subset seen in Fig. 2A-B). Using GAMUT as a design probe, I investigated why and how professional data scientists interpret models, and how the operationalization and its interface affordances could help them during analysis.

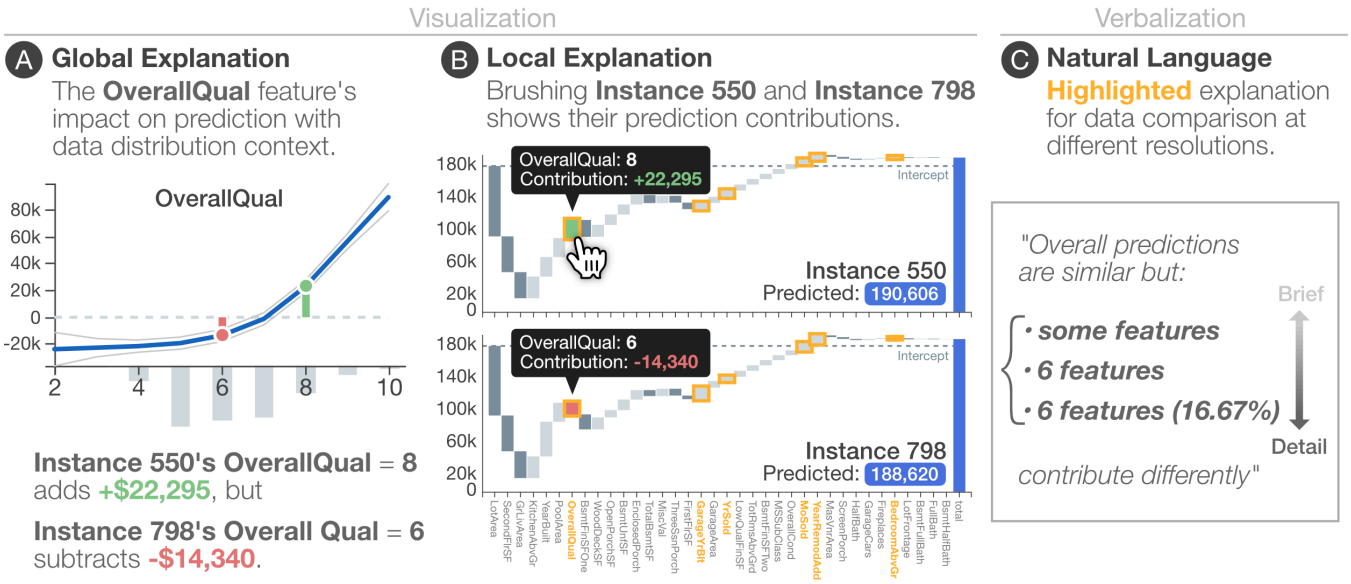

Figure 2: GAMUT + TELEGAM allow practitioners to interactively and scalably explain generalized additive models. For example, a practitioner can (**A**) visualize global feature explanations, (**B**) compare local instance predictions with data distribution context, and (**C**) control natural language explanation resolution, e.g., comparing prediction explanations.

My investigation showed that interpretability is not a monolithic concept: practitioners have different reasons to interpret models and tailor explanations for specific audiences, often balancing competing concerns of simplicity and completeness. Since our original work, GAMUT has been **deployed at Microsoft**, demoed for executive leadership at their internal TechFest, and incorporated into their open-source library InterpretML (2.4K+ Github stars).

TELEGAM: Combining Visualization and Verbalization for Interpretability

Despite GAMUT's usefulness for interpreting models, practitioners must still analyze visualizations which can require additional expertise. To reduce this cognitive load, I investigated using other mediums, such as natural language (i.e., verbalization), to provide a simple, yet effective way to communicate and summarize key aspects about a model, such as the overall trend in its predictions or comparisons between pairs of data instances. I extended my work in GAMUT and designed and developed TELEGAM [VIS19a], an interactive interface that *combines visual and verbal explanations*. TELEGAM generates natural language explanations while letting a user interactively specify the resolutions (e.g., "brief" to "detailed," Fig. 2C). Together, GAMUT and TELEGAM demonstrate how interactive interfaces can better communicate model explanations, based on user-specified resolutions, to differing stakeholders invested in machine learning systems.

Scaling Deep Learning Interpretability

Standardized toolkits for building neural networks have democratize deep learning. Since neural networks exhibit challenging and sometimes mystifying behavior, it is crucial that we equip practitioners with tools for identifying when a model works correctly, when it fails, and ultimately how to improve its performance.

Interrogative Survey for Visualization in Deep Learning

To obtain an overview of visualization and interpretability techniques for deep learning, I conducted the first survey

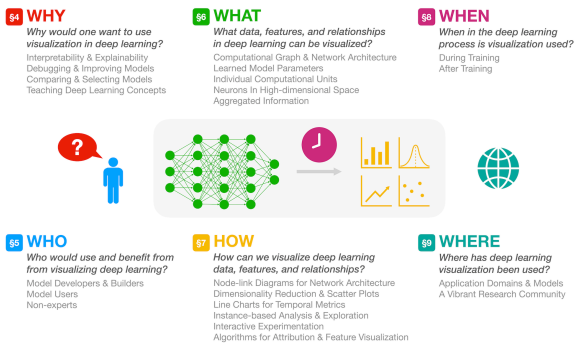

Figure 3: My interrogative survey uses the 5 Ws & H to help researchers and practitioners quickly learn about deep learning visualization.

which thoroughly summarizes the field using a *human-centered interrogative framework* (the Five W's and How: Why, Who, What, How, When, and Where; Fig. 3) to help people quickly learn key aspects of this rapidly growing body of research [TVCG18]. When examining literature, I discovered that existing work on interpreting neural network predictions for vision tasks typically focuses on explaining predictions for single images or neurons. But as predictions are often computed from millions of weights optimized over millions of images, such explanations can easily miss a bigger picture.

SUMMIT: Visualizing Activation and Attribution Summarizations

To scale-up deep learning interpretability, I designed and developed SUMMIT [TVCG20], an interactive system that summarizes and visualizes *what* features a deep learning model has learned and *how* those features interact to make predictions. SUMMIT introduces two new scalable summarization techniques, both based on aggregation, that discover important neurons and identify relationships among such neurons. SUMMIT combines these techniques to create an *attribution graph* (Fig. 4) that reveals and summarizes crucial neuron associations and substructures that contribute to a model's outcomes.

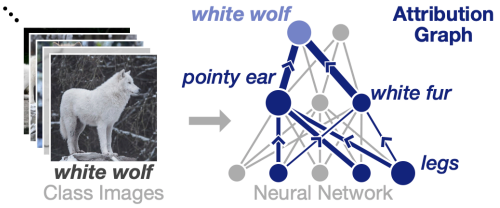

Figure 4: My attribution graph approach in SUMMIT summarizes thousands of images and reveals how a model's learned high-level features are composed from low-level ones.

SUMMIT (Fig. 5) scales to large data (e.g., ImageNet with 1.2M images) and uses neural network feature visualization and dataset examples to help users distill complex neural networks into compact visualizations. Using SUMMIT with GPU clusters, I discovered multiple surprising insights into a prevalent, large-scale image classifier's (InceptionNet) learned representations that prompted reexamining the training data and network architecture design. SUMMIT is web-based, open-sourced, and is accessible through a **live demo + embedded article**.

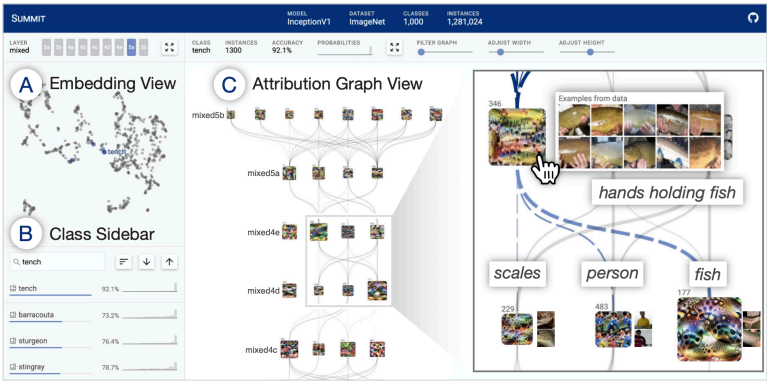

Figure 5: SUMMIT discovers surprising associations in deep models and their data, such as the *tench* fish using parts of people for classification (this is because all tench images in ImageNet have people holding the fish!).

Communicating Interpretability to Everyone

Most interpretability interfaces are designed for data literate people with machine learning expertise. But machine learning now impacts everyone, therefore it is important that everyone knows how to interact with it. Without technical overhead, how can we teach and make machine learning more approachable and accessible for people such as non-experts, students, and underrepresented groups in computing?

Parametric Press: Interactive Articles in Practice

On the web, a new medium for communication is emerging. *Interactive articles* interleave text with animations, visualizations, and simulations and leverage active reading to engage learners. To teach people about machine learning, I have written multiple interactive articles on interpretability, fairness, and bias (Fig. 6) [PP19], common data science techniques such as dimensionality reduction [VISxAI18, Best Paper Honorable Mention], and launched a new open-source publishing initiative called The Parametric Press to test this medium in the wild—while giving a platform to authors to tell data-driven stories and create explorable explanations [VisComm19]. These articles went viral (250,000+ views and media coverage), which

allowed me to analyze reader patterns at scale to evaluate how these artifacts are read in practice, a critical yet underexamined aspect of publishing interactive content.

Figure 6: My interactive article that discusses machine learning's impact on society includes: *descriptive text, interactive graphics, data visualizations, bespoke animations, and live user-controlled simulations.*

Critical Analysis of Interactive Article Design

With my experience in interactive publishing at scale, these articles, while still relatively rare, have been shown to be more engaging, attract broad readership and acclaim, and may help improve recall and learning—yet we do not know that much about them. Unfortunately, achieving this level of exposition in research dissemination is challenging: interactive articles are highly time consuming to author, require a diverse set of skills (e.g., editorial, design, and programming), and currently lack any formal incentive structure in research. Furthermore, there is no formalization for why they are useful and to what extent they can benefit readers.

Given these challenges, I examined the design of interactive articles by synthesizing theory from human-computer interaction, journalism, education, and multimedia learning. This work *ties together the theory and practice* of authoring and publishing interactive articles and surveys their design space and interactive techniques. To demonstrate and reinforce the power of interactive articles, this work is written as an interactive article [Distill20, in submission], and includes interactive graphics highlighting the techniques surveyed and discussed.

Future Research

I believe that data-driven technology should *empower people, augmenting human intelligence and decision-making.* My continued mission is to create close collaborations between multiple diverse disciplines to both advance our understanding of human-machine collaboration and create practical tools so that people can confidently interact with and trust machine learning.

Mixed-initiative Methods for Model Development

In my previous work I have shown the importance of *designing data* in the machine learning process [CHI20] and how mixed-initiative methods can help practitioners *discover intersectional bias* in their data and models [VIS19b]. Continuing this focus on the importance of data in machine learning, I am interested in creating mixed-initiative methods for machine learning's development cycle and its forward-facing challenges.

First, since data labeling is a critical bottleneck for kick-starting machine learning projects, I envision approaches for *interactive data programming* that allow people to easily write labeling functions for large datasets and interactively visualize and test their results. Once data is labeled, I am interested in formalizing test-driven machine learning by constructing a framework to enable people to quickly specify *unit tests for model evaluation* that could ensure models trained for specific tasks are fair and pass both human-defined (e.g., task-specific, societal) and computational checks (e.g., metrics). With this evaluation criteria, in a scenario where a model may underperform on specific subgroups of data, I foresee future systems that identify such groups, alert and present them to a user, and *suggest potential solutions* such as data augmentation or applying a new "patch" model to treat underrepresented data differently.

Making Interpretability Common Practice

Knowing how to represent and communicate machine learning explanations requires a deep understanding of the target users. Since the design and development of interpretability tools is still in its infancy, little work exists on evaluating such tools in practice. In my own work [CHI19a], I have noticed practitioners are eager to trust explanations, neglecting their typical healthy skepticism about their data and models. Interpretability in practice must instill confidence in a user while they are taking their next action, but not mislead. Related to my work on enhancing data science tooling [CHI19b, Best Paper], future interpretability tools should be usable and lower a users cognitive load when evaluating a model. This future interaction between people and machine learning will not be achieved without *understanding how practitioners use interpretability tools in their own work*. I am interested in integrating interpretability tools on deployed models, running user studies to investigate how practitioners incorporate interpretability into their workflow, with the goal of making interpretability common practice.

Beyond Dissemination: Accessible Research Distillation

In all my research, I make an effort to disseminate my work in multiple mediums to make it as approachable and accessible as possible to people of all backgrounds. This includes live demos, video walkthroughs, open-source code, blog posts, recorded talks, slides, and the papers themselves. Although I plan to continue this practice, I want to amplify it with more ambitious approaches to make research accessible [TVCG20], particularly to non-technical people and underrepresented groups [PP19]. I wish to be the person who is pointed to not only for excellence in research, but in *research dissemination, exposition, and distillation*. Organizations have only begun testing this practice, but I wish to further legitimize and co-publish research artifacts (e.g., interactive articles [VisComm19]) alongside traditional research announcements. As someone who had little exposure to computer science research until late in college, I believe these efforts will help motivate students and underrepresented people to learn about computing. The potential here is untapped, and I imagine a future where the broader public wields numeric and graphical literacy and understand's machine learning's impact on their daily lives in an increasingly quantified and data-driven world.

References

[CHI19a] **GAMUT: A Design Probe to Understand How Data Scientists Understand Machine Learning Models.** Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, Steven Drucker. *ACM CHI 2019*.

[VIS19a] **TeleGam: Combining Visualization and Verbalization for Interpretable Machine Learning.** Fred Hohman, Arjun Srinivasan, Steven Drucker. *IEEE VIS 2019*.

[TVCG18] **Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers.** Fred Hohman, Minsuk Kahng, Robert Pienta, Duen Horng (Polo) Chau. *IEEE TVCG (Proc. VAST) 2018*.

[TVCG20] **SUMMIT: Scaling Deep Learning Interpretability by Visualizing Activation and Attribution Summarizations.** Fred Hohman, Haekyu Park, Caleb Robinson, Duen Horng (Polo) Chau. *IEEE TVCG (Proc. VAST) 2020*.

[PP19] **Myth of the Impartial Machine.** Alice Feng, Shuyan Wu, Fred Hohman, Matthew Conlen, Victoria Uren. *Parametric Press, 2019*.

[VISxAI18] **The Beginner's Guide to Dimensionality Reduction.** Matthew Conlen, Fred Hohman. *VISxAI at IEEE VIS, 2018*.

[VisComm19] **Launching the Parametric Press.** Matthew Conlen, Fred Hohman. *VisComm at IEEE VIS 2019*.

[Distill20] **Communicating with Interactive Articles.** Fred Hohman, Matthew Conlen, Jeffrey Heer, Duen Horng (Polo) Chau. *Distill 2020, in submission*.

[CHI20] **Understanding and Visualizing Data Iteration in Machine Learning.** Fred Hohman, Kanit Wongsuphasawat, Mary Beth Kery, Kayur Patel. *ACM CHI 2020*.

[VIS19b] **FAIRVIS: Visual Analytics for Discovering Intersectional Bias in Machine Learning.** Angel Cabrera, Will Epperson, Fred Hohman, Minsuk Kahng, Jamie Morgenstern, Duen Horng (Polo) Chau. *IEEE VAST 2019*.

[CHI19b] **Managing Messes in Computational Notebooks.** Andrew Head, Fred Hohman, Titus Barik, Steven Drucker, Robert DeLine. *ACM CHI 2019*.