SALEEMA AMERSHI | RESEARCH STATEMENT

Fundamentally, machine learning is done by people, for people. As a human-computer interaction researcher, I create technologies to make people better at building and using machine learning systems.

Rapid advancements continue in algorithms and systems to improve the performance and reliability of machine learning models. Yet, the success of these algorithms for any practical purpose depends on the capabilities of the people applying them. People collect the necessary data over which those algorithms are run. People design the features or architectures over which those algorithms learn. People determine if the learned models are sufficiently reliable to deploy, and, if not, people are tasked with diagnosing failures and taking action to improve subsequent iterations. Regardless of how superior our community's algorithms are, if people fail at any of these steps, our models will fail to have their intended impact on driving the intelligent applications and services of the future.

Study after study has shown that most of the difficulties people face in realizing machine learning based solutions are in progressing through the end-to-end applied machine learning process (Figure 1), not in executing specific algorithms to train models. For example, interviews with data scientists reveal that much of their time and effort is spent collecting and preparing data before it can even be trained on [12]. Similarly, studies show practitioners must iterate through various steps of the machine learning process multiple times before models are deemed reliable enough to deploy [15]. The consequences of not supporting the effective practice of machine learning can range from embarrassing mishaps to potential harm to the people relying on those models for automation or decision making [14, 18, 8].

CURRENT RESEARCH

My research examines the end-to-end machine learning process in real-world applications and scenarios. In doing so, I identify challenges and opportunities for significant advancement and invent technologies for making people better at machine learning while making the models they build more robust and reliable. Throughout my work, I distill guiding principles applicable in a broader context, providing a foundation for future machine learning systems.

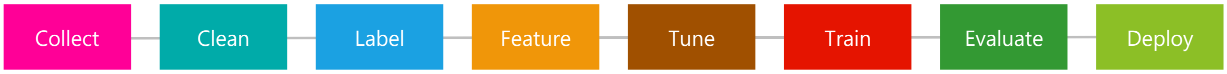

Figure 1. Key steps in the end-to-end applied machine learning process. Note that this is a fully connected graph where the result of any step may suggest looping back to previous steps to improve model performance.

Helping people label. In supervised machine learning, algorithms learn to reproduce and generalize from the mapping of training data inputs to outputs (or "labels"). High quality training data is therefore critical to producing effective models. Because people are often recruited to provide training data labels, considerable research has gone into developing algorithms and workflows to address data quality issues including labeler misunderstanding and inattentiveness or ambiguity inherent in the data itself.

Our research identified a distinct problem people face in labeling data, unsupported by existing solutions. *Concept evolution* refers to a labeler's mental process of defining and refining their desired mapping from data input to output (their "concept") as they observe data, typically resulting in inconsistent labels and poor quality models. A formative study we conducted showed that even skilled practitioners were only 81% consistent with themselves at labeling the exact same set of data over two sessions (with 67% of those participants' labels changing significantly from one session to the next).

Structured labeling is an interaction technique we introduced to help people consistently evolve their concepts as they label data. Structured labeling supports the explicit grouping, tagging, and organization of data within a traditional labeling scheme (mutually exclusive label targets). Building a visible structure while labeling helps people recall their decisions and consistently revise their decision boundaries as their concepts evolve. Our controlled evaluation showed that structured labeling was used and preferred by participants and helped them label significantly more consistently than traditional labeling. Moreover, the structures produced using our technique contain valuable information that can improve other phases of the machine learning process including revealing data distribution deficiencies and explaining model behaviors by supporting performance analysis on semantically grouped sub-structures.

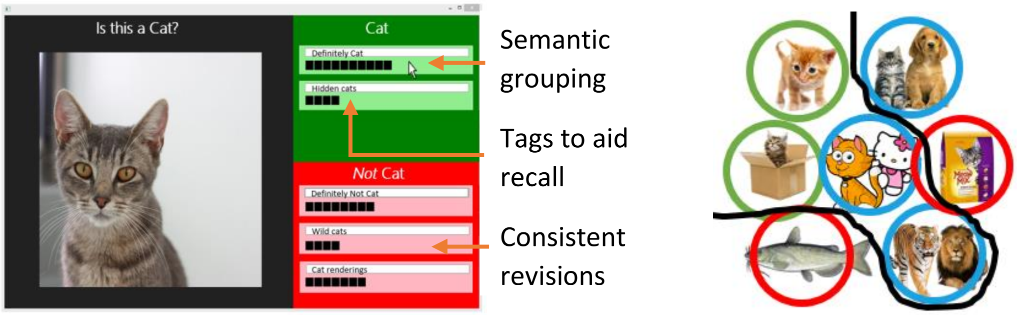

Figure 2. Structured labeling supports the explicit grouping, tagging and organization of data to help people consistently evolve their target concepts (e.g., "what is a cat?"). The resulting structures contain information that can improve other steps in model building (e.g., featuring, evaluation) and support experimentation with different label decision boundaries (right).

Our work on structured labeling received a best paper award at ACM CHI 2014, the premier conference on human-computer interaction [11]. In 2017, we extended this work to scale to large datasets labeled by crowdworkers by creating a new collaborative crowd workflow for generating structured output [7].

Helping people feature. Features are machine-understandable representations of data that must capture relevant information in order for machines to learn desired concepts. Accordingly, featuring is considered one of the most critical factors in building effective models [9]. Yet, little guidance or best practices exist for how people should conceive of useful features (*feature ideation*). While features can sometimes be obtained from related literature or automatically generated, practitioners still report spending considerable time thinking of and developing custom features [15, 9].

We developed *FeatureInsight*, an interactive tool that supports comparison-based, error-driven feature ideation. At any given iteration of model building, FeatureInsight automatically examines the model's errors, selects comparison items (items the model considers similar although they received differing user-provided labels), and then summarizes the errors and comparison items for further inspection by people. This approach turns the feature ideation problem into one of identifying salient differences between data appearing side-by-side on screen.

We examined FeatureInsight's support for feature ideation via a controlled experiment using a 2 (comparison of *pairs* vs. *sets*) x 2 (displaying *raw data* vs. *summaries*) within-subjects design. Our results show that summaries significantly improve the feature ideation process (i.e., resulted in significantly better model performance), especially when used in combination with set comparison which helped with generalization. This work appeared in the IEEE VAST conference in 2015 [6].

Pair Comparison

Set Comparison

Summaries

Positives Predicted as Negative
Terms related to: kittens
Negatives Predicted as Positive

Figure 3. FeatureInsight supports interactive error-driven, comparison-based feature ideation.

Helping people evaluate. Machine learning is an iterative process. These iterations are often driven by evaluating a model's current performance (e.g., poor performance may suggest looping back to collect additional data or create new features). Performance analysis typically begins with inspecting summary statistics of common metrics (e.g., accuracy, precision/recall). However, while summary statistics can efficiently convey the existence of errors, they do not indicate error severity or potential root causes. This type of error debugging then typically requires a disruptive cognitive switch to different tools or modes where practitioners can locate relevant data to gain insights and inform subsequent iterations.

ModelTracker is an interactive visualization we developed to bridge the gap between performance inspection and error debugging. In ModelTracker's unit visualization (see Figure 4), each training item is individually represented as boxes color-coded by their user-given labels (e.g., green as positive and red as negative). These boxes are laid out on screen according to the model's prediction scores in order to convey performance (e.g., agreement between the model and user is indicated when all green boxes are to the right while all red boxes are to the left). This single compact display conveys numerous metrics (including accuracy and precision/recall at all thresholds simultaneously) while showing item-level performance and diagnostic information (e.g., severity is indicated by unit position and annotations are used to suggest possible causes of errors including data or feature deficiencies). ModelTracker's unit-based display also enables direct access to data for further error analysis and debugging.

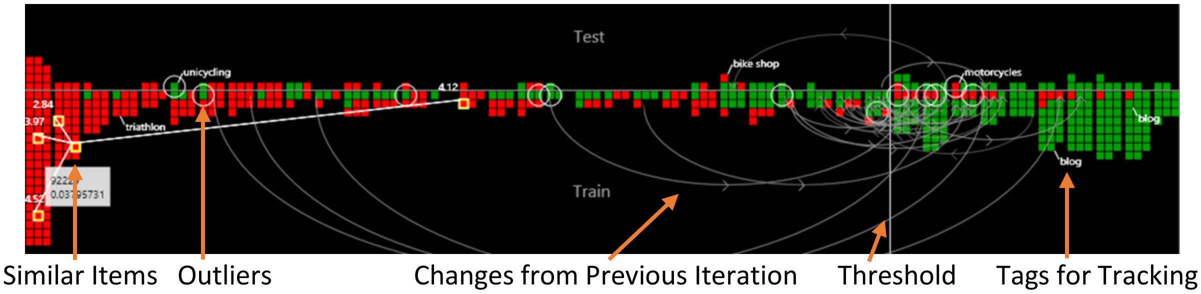

Figure 4. ModelTracker conveys information about numerous performance metrics while providing direct access to data.

Our investigations with real practitioners building models with ModelTracker for production use [17, 13] and experiments comparing ModelTracker with traditional techniques demonstrate that ModelTracker supports efficient and accurate performance analysis and debugging of machine learned models. ModelTracker for binary classification appeared in CHI 2015 [1] while a variant of ModelTracker (we call Squares) for multiclass classification appeared in IEEE TVCG 2016 [16].

FUTURE PLANS

I envision everyday people effectively harnessing the power of machine learning to achieve their goals and improve their lives. Realizing this future requires investment in the study and development of technologies to support the practice and use of machine learning.

Reaching new classes of people. Machine learning researchers are already working to expand the reach of machine learning to new domains and devices. Yet, most approaches to reaching new classes of people involve either providing access to a few pre-baked models or to infrastructures and APIs that assume users have a high-level of technical know-how. My work in supporting practitioners as well as non-expert end-users (e.g., people interacting with automation technologies and recommender systems) aims to lower the barriers to entry in machine learning. Still, much work remains. For example, how can efforts in tooling be best combined with automated support to make applied machine learning more efficient? How can people be taught to be better teachers of machine learning systems? How can algorithms and processes be made more interpretable so people can confidently incorporate machine learned models into social infrastructures? Questions like these must be answered if we are to reach the masses of people who stand to be empowered by machine learning technologies.

Grounding theory and algorithms in reality. Research in machine learning algorithms and theory often makes simplifying assumptions about the capabilities and limitations of people [19]. Active learning research, for example, often estimates user effort via label counts required to achieve a certain level of model performance. Label counts, however, are a poor proxy for user effort (expertise, priming, and interfaces can all impact efficiency and accuracy in labeling). If we are to trust machine learning systems and their robustness guarantees, algorithms and theories should be grounded in a realistic understanding of the people involved. For example, my work in structured labeling, which demonstrates that people struggle to label subjective data and even flip labels over time, informed a new theoretical analysis showing a penalty in the complexity bounds of active learning under these conditions [10]. Appropriately modeling people in our algorithms and theory requires collaboration across the boundaries of machine learning and human computer interaction research.

Establishing common principles for interaction. Decades of research and practice have resulted in established guidelines for how people should interact with computing systems. To accelerate the vision of empowering people through machine learning, we need guidelines for how people should effectively interact with learning algorithms and systems. We have learned, for example, that providing people with feedback about what happened [1] and allowing them to undo their actions [3] mitigates the difficulties of interacting with inherently unpredictable learning-based systems. We have also learned that people adopt a variety of model building strategies with varying levels of success [1, 2] and could therefore use more guidance in progressing through the machine learning process. And we have learned that while most end-user facing machine learning systems limit interaction to simple preference elicitation, in many situations people desire and can benefit from richer controls for steering machine learning systems towards desired behaviors (e.g., [2, 3, 4, 5, 6]). Establishing a comprehensive set of principles requires systematic experimentation in a wide variety of scenarios—defined by the specific people, data types, and target concepts involved—and generalization therefrom.

My work on making people better at building and using machine learning aims to achieve these goals.

REFERENCES

1. **Amershi, S.,** Chickering, M., Drucker, S., Lee, B., Simard, P., and Suh, J. (2015) ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. CHI 2015.
2. **Amershi, S.,** Fogarty, J., Kapoor, A., and Tan, D. (2009) Overview-Based Example Selection in Mixed-Initiative Interactive Concept Learning. UIST 2009.
3. **Amershi, S.,** Fogarty, J., Kapoor, A., and Tan, D. (2010) Examining Multiple Potential Models in End-User Interactive Concept Learning. CHI 2010.
4. **Amershi, S.,** Fogarty, J., and Weld, D. S. (2012) ReGroup: Interactive Machine Learning for On-Demand Group Creation in Social Networks. CHI 2012.
5. **Amershi, S.**, Lee, B., Kapoor, A., Mahajan, R., and Christian, B. (2011) CueT: Human-Guided Fast and Accurate Network Alarm Triage. CHI 2011. Best paper honorable mention.
6. Brooks, M., **Amershi, S.**, Lee, B., Drucker, S., Kapoor, A., and Simard, P. (2015) FeatureInsight: Visual Support for Error-Driven Feature Ideation in Text Classification. VAST 2015.
7. Chang, J.C., **Amershi, S.**, and Kamar, E. (2017) Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets. CHI 2017.
8. Crawford, K. (2016) Artificial Intelligence's White Guy Problem. The New York Times, June 25, 2016.
9. Domingos, P. (2012) A few useful things to know about machine learning. Communications of the ACM, 2012.
10. Huang, T.K., Li, L., Vartanian, A., **Amershi, S.**, and Zhu, J. (2016). Active Learning with Oracle Epiphany. NIPS 2016.
11. Kulesza, T., **Amershi, S.**, Caruana, R., Fisher, D., and Charles, D. (2014) Structured Labeling to Facilitate Concept Evolution in Machine Learning. CHI 2014. Best paper award.
12. Lohr, S. (2014) For Big-Data Scientists, 'Janitor Work' is Key Hurdle to Insights. The New York Times, August 17, 2014.
13. Microsoft Cognitive Services, Language Understanding Intelligent Service (LUIS). www.luis.ai
14. Miller, C. (2015) When Algorithms Discriminate. The New York Times, July 9, 2015.
15. Patel, K., Fogarty, J., Landay, J., A., and Harrison, B. (2008) Examining Difficulties Software Developers Encounter in the Adoption of Statistical Machine Learning. AAAI 2008.
16. Ren, D., **Amershi, S.**, Lee, B., Suh, J., and Williams, J.D. (2016) Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers. TVCG 2016.
17. Simard, P., Chickering, M., Lakshmiratan, A., Charles, D., Bottou, L., Garcia Jurado Suarez, C., Grangier, D., Amershi, D., Verwey, J., and Suh, J. (2014) ICE: Enabling Non-Experts to Build Models Interactively for Large-Scale Lopsided Problems. arXiv:1409.4814 [cs.AI].
18. Tett, G. (2014) Mapping Crime – Or Stirring Hate? The Financial Times, August 22, 2014.
19. Wagstaff, K. (2012) Machine Learning that Matters. ICML 2012.