Introduction to the Special Issue on Human-Centered Machine Learning

REBECCA FIEBRINK, Goldsmiths, University of London, UK
MARCO GILLIES, Goldsmiths, University of London, UK

Machine learning is one of the most important and successful techniques in contemporary computer science. Although it can be applied to myriad problems of human interest, research in machine learning is often framed in an impersonal way, as merely algorithms being applied to model data. However, this viewpoint hides considerable human work of tuning the algorithms, gathering the data, deciding what should be modeled in the first place, and using the outcomes of machine learning in the real world. Examining machine learning from a human-centered perspective includes explicitly recognizing human work, as well as reframing machine learning workflows based on situated human working practices, and exploring the co-adaptation of humans and intelligent systems. A human-centered understanding of machine learning in human contexts can lead not only to more usable machine learning tools, but to new ways of understanding what machine learning is good for and how to make it more useful. This special issue brings together nine papers that present different ways to frame machine learning in a human context. They represent very different application areas (from medicine to audio) and methodologies (including machine learning methods, HCI methods, and hybrids), but they all explore the human contexts in which machine learning is used. This introduction summarizes the papers in this issue and draws out some common themes.

CCS Concepts: • **Human-centered computing**; • **Computing methodologies → Machine learning**;

Additional Key Words and Phrases: human-centered machine learning, interactive machine learning

1 INTRODUCTION

Machine learning research often centers on impersonal algorithmic concerns, removed from human considerations such as usability, intuition, effort, and human learning; it is also too often detached from the variety and deep complexity of human contexts in which machine learning may be ultimately applied. However, considerable human work is always a part of gathering training data, tuning algorithms, and integrating machine learning into real-world systems. And human values, goals, and social structures always underpin decisions about what should be modeled in the first place.

It is our position that examining machine learning from a human-centered perspective includes explicitly recognizing both human work and the human contexts in which machine learning is used. Human-centered machine learning thus involves aligning algorithms and systems to human goals and capabilities, creating hybrid human-machine systems capable of achieving better results than either humans or algorithms working alone, and designing and evaluating machine learning systems using human-centered methods. A human-centered approach to machine learning demands making machine learning more usable and effective for a broader range of people, and designing new systems in full recognition of the agency and complexity of human users—including both people employing machine learning and those using or being affected by systems driven by machine learning.

Authors' addresses: Rebecca Fiebrink, Goldsmiths, University of London, London, SE14 6NW, UK, r.fiebrink@gold.ac.uk; Marco Gillies, Goldsmiths, University of London, London, SE14 6NW, UK, m.gillies@gold.ac.uk.

This special issue of TiiS follows a CHI 2016 workshop on Human-Centered Machine Learning [Gillies et al. 2016]. Our motivation for that workshop, and for this special issue, is to highlight research that employs human-centered approaches to the conception, design, and evaluation of machine learning systems. The papers presented in this issue illustrate how human-centered practices can be employed to make systems that are more usable and useful, and how a deeper understanding of human practices and contexts can ultimately lead to machine learning systems that have greater impact and address a wider variety of human concerns.

The papers presented in the following pages describe applications of machine learning to a wide span of human endeavors such as medical assessment, qualitative coding in the social sciences, audio annotation, creation of movemnet-sound interactions, and product recommendation. The research presented herein addresses a variety of human-relevant goals: from identifying and handling ambiguity in data to improving the accuracy and usability of machine learning systems to supporting human decision-making.

Human-centered machine learning brings together machine learning and human-computer interaction (HCI), two fields that have very different methodologies. Machine learning research works with pre-existing, often standardized datasets, using standard measures such as accuracy, precision and recall (for example the paper by Smith *et al.* in this issue), while HCI works directly with people, using quantitative or qualitative user studies (e.g., papers in this issue by Zeitz-Self *et al.* or Kim and Pardo). Crowd computing methods as used by Chen *et al.*, Dumitrache *et al.* and Zhang *et al.* in this issue sit in between the two: while human-generated data underpins this work, the data is often closely integrated into a machine learning process (e.g., used as training data).

What are we to make of the diversity of methodologies in this special issue? Should we conclude that human-centered machine learning is not yet a mature discipline because it does not have its own established methodology? Should we attempt to combine machine learning and HCI methods into a common methodology? Or should we celebrate this methodological diversity? Should human-centered machine learning remain, as Blackwell suggested of HCI [Blackwell 2015], an "inter-discipline" that consciously draws strength and creativity from the different disciplines that feed into it? These are just some of the questions we hope to spark with this special issue.

Below, we provide a brief summary of each paper selected for inclusion in this issue, accompanied by our own thoughts about how each paper contributes to an understanding of what human-centered machine learning can look like.

2 PAPERS IN THIS SPECIAL ISSUE

2.1 A Review of User Interface Design for Interactive Machine Learning

Dudley and Kristensson present an overview of prior research in one of the most important approaches to Human-Centered Machine Learning: Interactive Machine Learning (IML), and in particular user interface design for IML. They authors define IML as follows:

> *Interactive Machine Learning* is an interaction paradigm in which a user or user group iteratively trains a model by selecting, labeling and/or generating training examples to deliver a desired function.

IML is highly relevant to any discussion of human-centered machine learning, as it places particular importance on human interactions with the machine learning process, and much IML research combines approaches to system design and evaluation from human-computer interaction with machine learning techniques. This paper provides a valuable contribution by helping define and understand what is now becoming an established research area. It synthesizes the prior research by defining a set of elements common in most IML interfaces (sample review, feedback assignment, model inspection and task overview); describing a generalized workflow for IML (feature selection,

model selection, model steering, quality assessment, termination assessment and transfer); and outlining a number of emergent solution principles for the challenges of IML (Make task goals and constraints explicit, Support user understanding of model uncertainty and confidence, Capture intent rather than input, Provide effective data representations, Exploit interactivity and promote rich interactions and Engage the user). This synthesis provides guidance for future research in IML as well as a language in which to talk about and compare IML solutions.

2.2 Using Machine Learning to Support Qualitative Coding in Social Science: Shifting The Focus to Ambiguity

Chen *et al.* present work about the use of machine learning for the social sciences and begin their paper with a broad overview of the challenges involved. Social scientists deal with data sets that are complex, heterogeneous and often large. Machine learning has the potential to speed and support research processes with such data. However, the authors draw our attention to a number of mismatches between the machine learning experts' understanding of data analysis and the understanding of social scientists. Machine learning and social science not only have different technical methods, but also often have different goals and philosophies. For instance, social science often employs a theory- and hypothesis-driven approach to data, while machine learning typically deals with data that is gathered and analyzed without reference to a theoretical framework (or more worrying, with a theoretical framework that is entirely implicit and unstated even by the researcher). This is one of many important issues raised by the paper, which shows us the difficulties of applying machine learning to the processes of an established discipline and of communicating across disciplines.

This paper looks particularly at the use of machine learning for the qualitative data analysis practice of "coding", which the authors describe as *"a process of arranging qualitative data in a systematic order by segregating, grouping and linking it in order to facilitate formulation of meaning and explanation"*. Chen *et al.* identify ambiguity as a key challenge in the coding process (it is also an important theme of other papers in this special issue). Some of the textual data may be ambiguous in terms of its meaning and therefore the appropriate code. This can result in human coders being more uncertain of their code as well as in disagreement between coders. In a Mechanical Turk study, the authors found that non-expert crowd workers were consistent with experts in their judgment of ambiguity of data and that ambiguity was likely to result in disagreement between both expert and non-expert coders.

2.3 Predicting User's Confidence During Visual Decision Making

Smith *et al.* address a key problem raised by many papers in this special issue: humans providing training data for machine learning systems may be acting with uncertainty, leading to training data that is inaccurate or inconsistent. Information about human uncertainty has the potential to improve machine learning outcomes, for instance by reducing reliance on labels that may not be correct. However, obtaining information about uncertainty can be difficult: while it is possible to ask people how confident they are, this can impose a large overhead of human time and effort, and making confidence judgments can itself be a difficult task.

Smith *et al.* address these problems by using machine learning to estimate users' confidence at a visual decision making task (similar to a labeling task). Rather than explicitly asking users about their confidence, the researchers use implicit information gathered during the task: specifically, the user's eye gaze patterns while performing the task. The paper presents a novel representation of gaze data over time as a 2D image. This data is amenable to learning techniques that are commonly used for image data, in this case convolutional neural networks. This paper demonstrates the effectiveness of using implicit information for understanding user confidence and points to future

research that could make measures such as gaze tracking an integral part of a human-centered machine learning workflow.

2.4 Crowdsourcing Ground Truth for Medical Relation Extraction

Dumitrache *et al.* study the use of crowd workers to label medical texts. The workers' goal is to identify relations between terms within the text; for example, the sentence "fever induces dizziness" demonstrates the relation "causes" between the terms "fever" and "dizziness". Relation extraction of this type is a key data mining problem and one that, like many machine learning tasks, requires large quantities of labeled training data. Having medical experts label the data is expensive if it is possible at all, given these experts' time constraints, so using crowd-sourcing to collect ground truth annotations on which a machine learning system can be trained is an attractive alternative.

Like work by Chen *et al.* and Smith et al. in this special issue, a key issue in this paper is recognizing and dealing appropriately with ambiguity in the human labeling task. Dumitrache *et al.* found that many sentences in their corpus were ambiguous, and this was reflected in disagreement between crowd workers (in some cases, in sentences for which the authors themselves could not resolve the ambiguity). Rather than discarding instances for which crowd workers disagree or attempting to resolve the ambiguity to a single "correct" label as most researchers do, they actively make use of the ambiguity by weighting labels based on the agreement between crowd annotators. The authors then compare a medical relation extraction classifier trained on the weighted, crowd-supplied labels to a classifier trained on expert-provided labels, as well as to a fully automated baseline method. The results show that the crowd-labeled data approach was competitive with the results of experts and outperformed the automated approach.

2.5 Visualizing Ubiquitously Sensed Measures of Motor Ability in Multiple Sclerosis: Reflections on communicating machine learning in practice

Morrison *et al.* explore the use of machine learning to aid clinical decision making, specifically in the context of measuring human motor ability in the assessment of multiple sclerosis. Computer-based sensing systems and machine learning have the potential to improve on human assessments of motor ability. However, making such systems useful in clinical practice requires that algorithms provide more information than a simple assessment score: clinicians work in a complex decision-making landscape, in which they must integrate information about algorithmic assessment with their own knowledge, as well as with collaborative assessment with human colleagues.

This paper describes work to understand how visualization can be used to aid in the interpretability of machine learning systems for multiple sclerosis assessment. In an application of HCI methods to machine learning system design, the authors use a series of design iterations with clinicians to arrive at a better understanding of the challenges and user needs in this space. These reveal that simply making the algorithmic decision-making process more transparent to users is insufficient for supporting human decision-making. On the other hand, visualization can reveal useful aspects of the algorithmic decision-making context, including information about data quality and relationships within the data that are interpretable by and relevant to clinicians.

The contributions of this paper are of interest far beyond medical assessment. For instance, the method of employing a series user-centered design workshops with domain experts can be applied to better understand design challenges and principles for many machine learning application contexts. Further, the authors present a compelling argument against treating the design of machine learning algorithms and the design of user-facing applications as distinct tasks that can be done independently.

2.6 A Human-in-the-loop System for Sound Event Detection and Annotation

Kim and Pardo leverage machine learning's ability to learn from human-provided examples not to replace a human annotator, but to speed up human annotation. They focus on the application domain of annotating sound files to indicate the locations and durations of sounds of interest (e.g., spoken words, musical instruments, environmental sounds). Doing such annotation manually is labor-intensive, yet training a machine learning algorithm to accurately perform this annotation can require many examples (possibly more example sounds than exist in the corpus of interest), and may still yield insufficient accuracy.

Kim and Pardo thus use an iterative human-in-the-loop approach in which a small number of annotations provided by a person inform a nearest-neighbor-based relevance weighting, which identifies the unlabeled regions of audio most likely to contain the sound of interest. The human annotator adjusts the machine-generated labels on these regions if necessary, the relevance weighting is updated, and the process repeats with the next set of machine-suggested annotation regions. In the authors' evaluation with human annotators, this approach yielded a two-fold speed-up in completing annotation of sound files, compared to manual annotation alone.

This paper demonstrates how understanding the user's interaction capabilities and true goal can lead to a different formulation of the machine learning component of a system than one might expect. If the user's goal were to create a truly automatic annotation system, for instance, a more conventional active learning approach (in which the system recommends the user label data the algorithm finds ambiguous or informative) may have been appropriate. Instead, though, this system focuses on requesting labels for the audio segments that it most confidently believes are examples of the target sound, in order to most efficiently direct the user's attention to these areas.

The paper's evaluation of this approach is notable in its differentiation between the time incurred by the requirement that the user listen to the audio recommended for labeling by the algorithm (time which would be shortened, for instance, by a more accurate algorithm), and the extra time incurred by human interactions with the interface (the "interaction overhead" which might be improved, for instance, by a better user interface). By measuring the interaction overhead incurred by people using a particular interface, the time to label a new corpus (or the time to annotate using a different algorithm) with this interface can thus be estimated using a simulation experiment rather than additional user studies.

2.7 Evaluation and Refinement of Clustered Search Results with the Crowd

Zhang *et al.* describe a hybrid human-machine approach for clustering search results. While automatically clustering items can help people navigate and understand large datasets, clustering algorithms do not always group items in ways that people find most coherent or useful. This is the case for the application domain considered by Zhang *et al.*, who focus on clustering of search results returned for a user's query in the Google Play Store. One approach to improving the usefulness of these machine-generated clusters of search results is to have human experts manually refine them, but this does not scale to large numbers of search queries. Zhang *et al.* show that crowd-sourcing the task of refining clusters can produce good clusters in a more scalable way.

Their solution combines algorithmic and crowd-driven components. Multiple clustering algorithms are applied in parallel to the search results, due to the observation that different algorithms may work well for different types of searches. Crowd workers then evaluate these alternative clusterings, refine the best, and assign good titles to the final clusters. The implementation of crowd-driven components of the system thus entails both the decomposition of the overall task of improving clusters into small tasks that can be assigned to individual workers, as well as the

design of user interfaces that enable workers without specialized expertise to easily and accurately perform these tasks.

2.8 Observation-Level and Parametric Interaction for High-Dimensional Data Analysis

A number of human-in-the-loop systems for exploratory data analysis or model-building use one (or both) of the following interaction strategies: user manipulation of model parameters, or user manipulation of data from which new model parameters are inferred. Zeitz-Self *et al.* propose that these two strategies support different types of human tasks. They explore both strategies in the context of a data visualization system that uses weighted multidimensional scaling to project a dataset with multiple attributes into a two-dimensional visualization. This system allows users to directly manipulate the attribute weights used in the distance metric underlying the projection, or to move data points within the visualization to demonstrate an implied distance relationship (for which weights are then inferred). A controlled user study revealed that manipulating weights and manipulating data points supported different types of analytical activities, such as attribute-based filtering and observation comparison, respectively.

Notably, this paper considers a type of interactive task in which the goal is not to identify the "right" model weights for a particular dataset or application. Rather, the goal is to use exploratory manipulation of an underlying model in conjunction with a dynamic visualization of the current model to better understand the data. As in the paper by Kim and Pardo, then, the primary purpose of machine learning here is not to build a reusable model or replace human judgment, but to aid in a human task. Here, this task is understanding data. Machine learning is well placed to help with this task, but it is notable that improved human understanding does not result from the successful output of the machine learning process. Rather, understanding is facilitated by iterative interactions with the data and the model. This opens up an interesting possibility that, in interactive machine learning, human interactions and learning can be ends in themselves.

2.9 Motion-Sound Mapping through Interaction An Approach to User-Centered Design of Auditory Feedback using Machine Learning

Françoise *et al.* apply machine learning to gestural interaction design. This follows a recent trend in human computer interaction to look to new modes of interaction that are based on fuller body movements than a traditional mouse and keyboard or touch screen. They look at movement as a means of controlling sound synthesis, based on the insight that humans naturally interact with sound, and particularly music, with body movement such as dance or foot tapping. Their system allows end users to design *mappings* between movement and sound. Machine learning makes it possible to define complex, non-linear mappings by demonstrating example.

Françoise *et al.* use machine learning as a tool for interaction design. They address the challenge that traditional rapid prototyping approaches used in HCI, such as paper prototypes and wireframes are not well suited to movement based interaction. These techniques focus on the visual design of interfaces on assumption that the physical actions of the user will be relatively standard. For movement interaction, on the other hand the focus is on people's movements, not visual displays, particularly on the audio based application in this paper, where a visual interface can be entirely missing. Another challenge is that movement interaction is an embodied skill in the sense that we know how to dance or gesture by doing, rather than explicitly in an intellectual or symbolic form. This makes it hard to explicitly define gestures, for example in code, because know by doing, without being able to explain how we do it. Interactive machine learning makes it possible to prototype gestural interfaces by demonstrating movements themselves, rather than on paper or in symbolic form. Françoise *et al.* have developed an approach called *mapping through interaction*

in which users demonstrate examples of interactions with sound, and these examples are used to train a machine learning model.

The authors used a mixed approach to evaluation methods. They have a qualitative study in which users could interact with their system as part of a game. This allowed an HCI based evaluation of user experience. This HCI approach was supplemented with an off-line, machine learning style evaluation in which the accuracy of different machine learning algorithms was compared with a standard gesture data set. This ability to work across different disciplinary methodologies is likely to be an important part of Human-Centered Machine Learning research in future.

3 CONCLUSION

As we have seen, the papers in this special issue represent many different ways to bring human considerations into the creation and use of machine learning systems. Perhaps this is the most interesting outcome of this special issue: human-centered machine learning is not a single approach, but a wide diversity of problems, methods, technologies and theories, all of which could, and should, be explored by researchers for years to come. As machine learning moves out of the research lab and into more real-world systems, the question of how to ensure that these systems are usable and useful for people becomes increasingly urgent. We hope the papers in this special issue will help inform readers as they contemplate where and how we might continue to make machine learning research more human-centered, foregrounding human goals and experiences in both the development and application of new machine learning technologies.

ACKNOWLEDGMENTS

The authors would like to thank the co-organizers and attendees of the CHI 2016 workshop on Human-Centered Machine Learning [Gillies et al. 2016] for their initial work articulating a vision for human-centered machine learning and building a community around this topic. We are indebted to TiiS Editor-in-Chief Michelle Zhou and assistant Anbang Xu for their guidance in the preparation of this special issue, and we are grateful for the hard work and thoughtful feedback of the anonymous peer reviewers who assisted with the paper selection and revision process. Finally, we would like to thank the authors whose hard work has made this special issue as diverse and stimulating as it is.

REFERENCES

Alan F. Blackwell. 2015. HCI As an Inter-Discipline. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '15)*. ACM, New York, NY, USA, 503–516. https://doi.org/10.1145/2702613.2732505

Marco Gillies, Rebecca Fiebrink, Atau Tanaka, Jérémie Garcia, Frédéric Bevilacqua, Alexis Heloir, Fabrizio Nunnari, Wendy Mackay, Saleema Amershi, Bongshin Lee, Nicolas d'Alessandro, Joëlle Tilmanne, Todd Kulesza, and Baptiste Caramiaux. 2016. Human-centred machine learning. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3558–3565. https://doi.org/10.1145/2851581.2856492