Do Machine Teachers Dream of Algorithms?

Gonzalo Ramos
Microsoft Research
Redmond, WA
goramos@microsoft.com

Felicia Ng
CMU
City, State
fng@cs.cmu.edu

Nicole Barbosa Sultanum
University of Toronto
Toronto, ON
nicolebs@cs.toronto.edu

Christopher Meek
Microsoft Research
Redmond, WA
meek@microsoft.com

Jina Suh
Microsoft Research
Redmond, WA
jinsuh@microsoft.com

Soroush Ghorashi
Microsoft Research
Redmond, WA
sorgh@microsoft.com

Abstract

Machine Teaching (9) is an emerging approach for incrementally creating semantic machine learning (ML) models, by focusing on improving the productivity of the human teacher during an interactive ML process. One of the challenges of this approach is developing the support for the process of eliciting subject-matter knowledge relevant for a machine learner. This process we call *Knowledge Decomposition* not only encompasses knowing what type of knowledge to articulate, but also the language, one articulates it with. In turn, language and articulation influence and are influenced by the teacher's mental models. We share our findings on people's mental models about learning systems from two formative studies where people teach to hypothetical learning systems. These findings are important for practitioners creating teaching experiences, as the wrong notion of how the machine learner works and what it needs can negatively affect the overall teaching task. We argue that addressing this mismatch is core to the success of machine teaching and a multidisciplinary opportunity for HCI, and ML practitioners alike.

1 A Brief Introduction to Machine Teaching

Because of its potential to amplify human-like decision making at scale, there is an increasing interest in incorporating ML into many aspects of existing and new systems at all levels of society. However, realizing this vision of ML everywhere is not straightforward. In particular, creating ML models is a task requiring skills beyond those of the general population.

Several approaches aim at making ML models accessible to non ML-experts, with different roles and skills. These approaches include AutoML (4), toolkits like as scikit-learn (8) or PyTorch, or online environments such as those provided by Microsoft's Azure, Amazon's AWS or Google Cloud. Interactive Machine Learning (IML) (12) is another approach that has made tangible progress in bringing ML model creation closer to the subject-matter expert that has the knowledge that needs to be incorporated into the model. While all ML process where humans are involved are interactive under a certain lens, we use the term the way (3; 5; 1) describe it: "an interaction paradigm in which a user or user group iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review", where humans in the loop are engaged in a "rapid, focused and incremental learning cycles". Examples of IML systems include the Wekinator (6), Gestalt (7), and Crayons (5). Still, these IML solutions, as a family, are not specific about the skills or knowledge they require from their users, both in terms of the underlying ML algorithms at play, the parameters users can or should affect, or how transparent or semantic the resulting ML model is.

Workshop on Human-Centric Machine Learning at the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. (non-archival)

Machine teaching as presented in (9) builds on IML by taking a point of view regarding the skills and roles of teachers, and by focusing on the exchange of knowledge between a human teacher and a machine learner. [1] Machine teaching is complementary and it contrasts ML in the following way: *if ML is about extracting knowledge from labeled data, machine teaching is about extracting knowledge from people*. Machine teaching focuses on processes where ML models are the result of incremental iteration, and is principled about the form and types of knowledge a human teaches to a learning system. Machine teaching emphasizes teaching productivity and accessibility, while producing ML models that are semantic by design. Furthermore, it proposes an abstract interface layer that hides the complexities of the ML algorithm and its parameters, making it accessible to end-users that need only subject-matter expertise and the inherent capacity to teach.

1.1 How and What Teachers Teach

Teaching leverages innate human capabilities, which can be amplified through experience. Wall et al.'s (11) provide a complimentary introduction to machine teaching while providing insight into how to enable people to be effective machine teachers. This work illustrates a *machine teaching loop* (Figure 1) that shows the activities in which a teacher takes part while teaching.

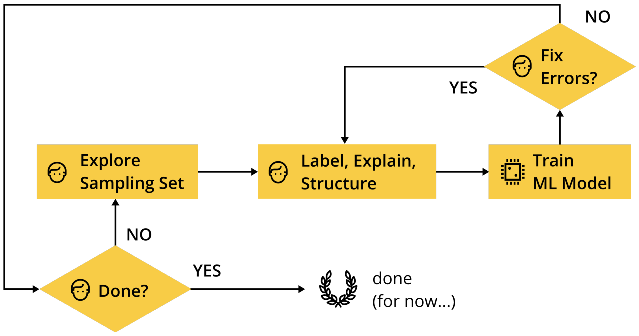

Figure 1: Machine Teaching Loop. It often starts by exploring the set of documents to label or review.

While in this loop, teachers use a *teaching language* allowing the expression of meaningful knowledge such as labels, features, schema as well as actions such as sampling or schema specification. In this paper, we share some of our findings when seeking to characterize this language, by studying the types and ways people express rich knowledge while teaching to a learning system. In particular, we focus on how teachers form mental models about the systems they interact with, and pose questions relevant to designers and researchers working at the intersection of HCI and ML.

In the next section, we will share partial results from two parallel formative studies that observed how people teach hypothetical learning systems to create a classifier for text documents; and an entity extractor for images, respectively. Full details about these studies and their full findings are outside of the scope of this paper, and are currently part of longer submissions to peer-reviewed conferences.

2 Observing Teachers of a Classifier for Text Documents

In one formative study, we observed 20 participants teaching to a hypothetical learning system how to assign the labels *business* and/or *food* to rich text format documents (where the text font could have attributes such as size and weight) for 45 minutes each. The hypothetical learning system consisted of (1) researchers playing the role of the learning system, and (2) physical materials that participants (or teachers) could interact with such as documents, markers, highlighters, post-it notes, tags, etc.

The study began by researchers setting the context of the teaching task of assigning labels to text documents. Participants labeled documents and articulated knowledge that *they thought would be useful for the system to make decisions*. Researchers told participants that they were free to provide useful knowledge in the form of concepts, relationships and/or rules. Researchers instructed participants to commit useful knowledge into a post-it note. Participants were also instructed to produce a knowledge summary/map that organized and laid out all committed knowledge during the session in any representation they desired. As participants produced knowledge, the system (via the

[1] This is a different interpretation from (13) which uses the same name to denote the inverse problem to ML.

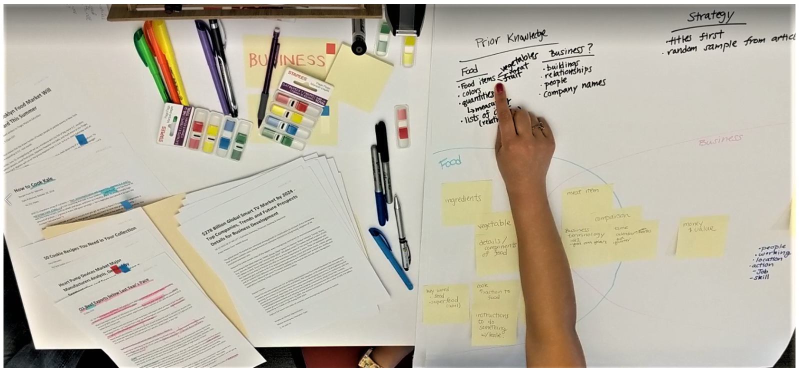

Figure 2: An example for the deliverables of the formative study on teaching a classifier for text documents. Labeling and annotation task (left), and knowledge summary task (right).

researcher's role-play) provided feedback about what it understood or what it did not. This feedback dialog encouraged participants to explain their knowledge, sometimes by further decomposing their knowledge (e.g., to explain the concept of "header", participants introduced "font size" or "content order"). Lastly, researchers explained to the participants that they should not be concerned about how the learner worked or made predictions, other than the fact that the learner will take the given knowledge and decide how to use it to predict the labels. Figure 2 illustrates results from a session.

3 Observing Teachers of an Extractor for Images

In another formative study, we observed participants engaging in a 60-minute machine teaching task where they taught a hypothetical learner system how to extract or segment a concept or entity from an image. The hypothetical learner setup was similar to the formative study described above. The study began by researchers setting the context of the task of teaching a system to segment an entity, such as *person playing tennis*, *person riding a bicycle*, or *occupied nest* from an image.

Researchers asked participants to first think about a target concept (e.g., *person riding a bicycle*) before looking at any images, and verbally describe this target concept by articulating sub-concepts, relationships and/or rules. After this foresight task, researchers asked participants to, given an image, extract the target concept through *Descriptions* – by articulating/marking sub-concepts, relationships and/or rules that a learner could see in an image, and that *they thought would be useful to the system to make decisions*. After this initial teaching task, researchers presented participants with images for which the learning system provided feedback (via the researchers' role-play, e.g. markings on an image, or verbal prompts like "the system sees a person") about what it could see and understand. Researchers then asked participants to reflect on the cases where there was a disagreement between their judgment and the system's, and to refine their descriptions. As participants produced descriptions, the system feedback triggered a dialog that inspired participants to further decompose their concepts and entities. Similarly to the former study, researchers did not explicitly describe to teachers how the learner worked or made predictions.

4 On Teachers' Mental Models

As designers of machine teaching experiences, we want to craft experiences that reinforce or lead to useful mental models leading to good teaching results. It is important then to characterize the ways people think about how learning systems work and be controlled, so that we can leverage mental models people naturally gravitate towards, and discourage the ones that could lead to undesirable results.

In our two formative studies, participants formed ideas about how the learning system worked, despite being instructed not to make assumptions. This is not surprising, since as humans, we cannot stop form forming theories about how the world works. These mental models incorporate prior knowledge and language, and influence how people form perspectives, make decisions, and act. People express these models through the interfaces they use. During our formative studies, we tried to provide as much freedom of expression as we could to participants, with the objective of not leading them into a particular teaching grammar or language. As Teachers, participants' actions had the goal of affecting the behaviors of the system, not only to alter the *prediction* behaviors, but also to influence the *learning* behaviors. While observing these actions was not a direct measure on the participants' mental models about the learning system, they still suggest what these models might be:

Monolithic Structure. Participants did not verbalize an explicit distinction between components such as a *learner*, that computes the parameters of a model, and a *model*, that computes predictions. For teachers, the learner-model tuple was seen as a monolithic structure that they influenced either through labels, features, and/or rules.

Rule-Based System. In both studies, teachers articulated some form of rules that the learning system should follow to make predictions. Participants used "rules" to both define concepts and to fix incorrect predictions. These rules took the form of (1) presence of an entity or a concept ("[For a *person riding a bicycle* extractor,] there's a person and there's a bicycle"), (2) procedures ("first look at the title and see if it has any food-related words", "If the frequency of these keywords is greater than 5, then label it as *Food*"), or (3) probabilistic ("person wears a helmet *sometimes*").

Knowledge is Exclusive. Despite participants knowing that a document could have more than one label, many thought in terms of *exclusive features*: knowledge that only applied to a particular target label (e.g., *Meats* for food or *Banks* for business). E.g., some participants thought concepts such as *money* were too general to be useful. The knowledge maps they produced reflected this exclusivity as many showed non-overlapping sub-concepts regions. Having knowledge exclusivity promotes conceptualizing rule-based systems and straightforward decision boundaries for predictions.

Evidence has Weight. Participants wanted to express *importance* for certain words or concepts. For example, participants used terms such as ''strong" or "weak", marked a certain number of "stars" to indicate that the learner should weigh some things more than others. Other participants prescribed how weights should affect predictions: "If you see this set of words, then 90% sure it's this label". This type of mental model is consistent with observations from (10).

Power of Examples. Independently of their background on ML, participants used examples to steer the system's behaviors. They made decisions about the quality and quantity of example to teach. E.g., participants browsed the available text documents before starting to label, to get a general sense of what the target concept looks like.[2] Others used prior ML knowledge to underline the importance of examples: "Because I know [with ML] is important to have good examples [...] the algorithm will be better based on the examples it has".

We described five distinct mental models observed from our two formative studies. We hope these stimulate further discussions about how to harness them as HCI/ML designers and researchers.

5 Changing Minds and Mental Models

What is a good mental model? We propose the notion that good mental models are *useful* mental models. They do not need to be a 1:1 representation of a learning system. Instead, a useful mental model needs to have just enough detail to support a person's task and goals. Thus the fitness of a mental model is affected not only by the system's interface and the parameters of a system's components, but also the success metrics of the task at hand.

Designers of teaching experiences, not only have a say on how a visual interface looks, but also on how a UI/X instantiates a teaching language: its building blocks and syntax. This includes how one visualizes explanations for predictions and a system's state. Systems like MATE (11) do not allow for teachers to express rules, a design decision that might lead to simple mental models that are easy to onboard, at the cost of teaching expressivity.

Designers of learning algorithms have a say on their interfaces and the semantics of their learning parameters. Techniques such as GAMs (2) can be used to invite non-ML experts to change a predictor's parameters directly. For other learning systems, such as deep neural networks, changing a network's hyperparameters is out of scope for non-ML experts. There are opportunities for ML researchers to think about how to leverage the subject-matter knowledge that teachers provide beyond labels and features.

[2] We hypothesize that people preview the dataset to quickly form a simple ad-hoc, rule-based system in their head using back-of-the-envelope features they ideate through their browsing.

Lastly, experience and algorithm designers can, and should, work together towards delivering thoughtful human-centered machine learning solutions that support and guide useful mental models. Experience designers need literacy on what algorithms can do and what input they require. Algorithm designers need literacy on the rich knowledge beyond labels that people can provide, as well as how such input can be used in learning.

References

[1] Amershi, S., Cakmak, M., Knox, W.B., Kulesza, T.: Power to the people: The role of humans in interactive machine learning. AI Magazine **35**(4), 105–120 (2014)

[2] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N.: Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1721–1730. ACM (2015)

[3] Dudley, J.J., Kristensson, P.O.: A review of user interface design for interactive machine learning. ACM Transactions on Interactive Intelligent Systems (TiiS) **8**(2), 8 (2018)

[4] Elshawi, R., Maher, M., Sakr, S.: Automated machine learning: State-of-the-art and open challenges (2019)

[5] Fails, J.A., Olsen, Jr., D.R.: Interactive machine learning. In: Proceedings of the 8th International Conference on Intelligent User Interfaces. pp. 39–45. IUI '03, ACM, New York, NY, USA (2003)

[6] Fiebrink, R., Cook, P.R.: The wekinator: a system for real-time, interactive machine learning in music. In: Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR) (2010)

[7] Patel, K., Bancroft, N., Drucker, S.M., Fogarty, J., Ko, A.J., Landay, J.: Gestalt: integrated support for implementation and analysis in machine learning. In: Proceedings of the 23nd annual ACM symposium on User interface software and technology. pp. 37–46. ACM (2010)

[8] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)

[9] Simard, P.Y., Amershi, S., Chickering, D.M., Pelton, A.E., Ghorashi, S., Meek, C., Ramos, G., Suh, J., Verwey, J., Wang, M., et al.: Machine teaching: A new paradigm for building machine learning systems. arXiv preprint arXiv:1707.06742 (2017)

[10] Stumpf, S., Rajaram, V., Li, L., Burnett, M., Dietterich, T., Sullivan, E., Drummond, R., Herlocker, J.: Toward harnessing user feedback for machine learning. In: Proceedings of the 12th International Conference on Intelligent User Interfaces. pp. 82–91. IUI '07, ACM, New York, NY, USA (2007). https://doi.org/10.1145/1216295.1216316, `http://doi.acm.org/10.1145/1216295.1216316`

[11] Wall, E., Ghorashi, S., Ramos, G.: Using expert patterns in assisted interactive machine learning: A study in machine teaching. In: Proceedings of the 17th IFIP TC 13 International Conference on Human-Computer Interaction. Springer International Publishing (2019)

[12] Ware, M., Frank, E., Holmes, G., Hall, M., Witten, I.H.: Interactive machine learning: letting users build classifiers. International Journal of Human-Computer Studies **55**(3), 281–292 (2001)

[13] Zhu, X.: Machine teaching: an inverse problem to machine learning and an approach toward optimal education. In: The Twenty-Ninth AAAI Conference on Artificial Intelligence (2015)