Understanding and Supporting Knowledge Decomposition for Machine Teaching

Felicia Ng
Carnegie Mellon University
Pittsburgh, PA, United States
fng@cs.cmu.edu

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

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

ABSTRACT
Machine teaching (MT) is an emerging field that studies non-machine learning (ML) experts incrementally building semantic ML models in efficient ways. While MT focuses on the types of knowledge a human teacher provides a machine learner, not much is known about how people perform or can be supported in this essential task of identifying and expressing useful knowledge. We refer to this process as *knowledge decomposition*. To address the challenges of this type of Human-AI collaboration, we seek to build foundational frameworks for understanding and supporting knowledge decomposition. We present results of a study investigating what types of knowledge people teach, what cognitive processes they use, and what challenges they encounter when teaching a learner to classify text documents. From our observations, we introduce design opportunities for new tools to support knowledge decomposition. Our findings carry implications for applying the benefits of knowledge decomposition to MT and ML.

Author Keywords
Machine teaching; Interactive machine learning; Knowledge decomposition; Sensemaking

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** *Empirical studies in HCI;*

INTRODUCTION
Machine teaching (MT), as defined by [27, 34], is a field of study that aims to enable anyone who has knowledge in an application domain, but not necessarily in machine learning (ML), to incrementally and iteratively teach that knowledge to a machine in an efficient way. MT is a perspective on the human-in-the-loop ML process used to create ML models that focuses on optimizing the way in which humans transfer subject-matter knowledge to machines instead of focusing on optimizing the learning algorithms as in traditional ML research. MT leverages peoples' ability to teach and can lead to models that are semantic, reusable and easy to maintain. Some commercial products like LUIS [21] and bonsai [5] are examples of systems that apply nascent MT approaches.

In this paper, we define users as subject-matter experts in the application that they want to build an ML model for, but not necessarily experts in ML. A subject-matter expert could be someone with a highly specialized knowledge (e.g., neurology) or someone with a highly subjective definition of a concept (e.g., interior design that I like); teaching ML to these subject-matter experts and extracting models can be costly or mining knowledge about these subjects can be impossible through crowd sourcing. We refer to the person creating an ML model through MT as a *machine teacher*, or *teacher* for short, and we refer to the learning algorithm that produces an ML model from the given knowledge input as a *machine learner*, or *learner*. For example, consider a person who wants to build an image classifier for dogs; they have knowledge about how to identify dogs but no expertise in ML. The main form of knowledge that they can provide through traditional ML approaches is labels. However, one of the challenges of this approach is that it often requires a large dataset of labeled examples, which, for some subject matters, may or may not be easily available. Another challenge of traditional ML solutions is that they lead to black box models that provide predictions based on opaque features with non-relatable semantics. For example, after providing a large dataset of labeled examples, a model-in-training may incorrectly predict that an image of a muffin is a dog. The human is then left wondering why such a prediction was made and how to fix the model.

MT provides answers to some of these challenges by leveraging the fact that human teachers possess and can offer richer forms of knowledge than just labels. For example, a teacher may know about features, concepts, relationships, rules, and other strategies that are useful for recognizing dogs. In addition to selecting specific examples of images and labeling them as "Dog" or "Not Dog", the teacher can also provide semantic explanations about why an image is labeled as such. They can explain specific concepts that they look for, such as the body parts of a dog (i.e., features like the eyes, ears, nose, legs, etc.), relationships between those concepts, such as how the nose is typically centered beneath the eyes, and many others. Being able to express these richer forms of knowledge to the machine can make the teaching process efficient; the teacher can explicitly tell the machine about these semantically meaningful features through a few examples, rather than requiring the machine to learn arbitrary features through a large dataset of labeled examples. Through this approach, the

DIS '20, July 6–10, 2020, Eindhoven, Netherlands.

ACM ISBN 978-1-4503-6974-9/20/07 ...$15.00.
http://dx.doi.org/10.1145/3357236.3395454

learner can also provide semantic explanations back to the user. For example, after the user teaches the machine to recognize the concept of "eyes", it might show that the reason why it labeled an image of a muffin as "Dog" is because it recognized the chocolate chips on it as "eyes". That would help the user identify how to fix the model (e.g., by limiting the maximum number of eyes to 2). This example is illustrative of main teaching interactions (sampling, featuring, labeling, debugging) and how ML models created through a MT process are semantic and thus explainable by design.

We define *knowledge decomposition*[1] as a process of identifying and expressing useful knowledge by breaking it down into its constituent parts or relationships. As the example above highlighted, knowledge decomposition is critical to a successful and effective MT process in which a teacher can articulate richer forms of explanations for their labels to the learner. [34] identified moments during a MT loop where it is appropriate to add semantic features, but they do not indicate the specific decomposition that should take place. Furthermore, because humans use a lot of tacit knowledge to make decisions, it is unclear how to build tools that help people select and express that knowledge to a machine.

MT represents a shift in the research and design of ML model creation tools from technical systems (focusing on what the learner needs and can offer) to human-AI collaboration systems (focusing on what the human teacher needs and can offer in addition to the what the learner needs and can offer). We explore the unique challenges of such a collaborative approach to ML model creation by building foundational frameworks for understanding and supporting the knowledge decomposition process in MT. Understanding knowledge decomposition process would help designers of future MT systems to support integrating and engaging ML novices in the model building loop. We conducted a formative study to observe what types of knowledge people teach, how they organize/represent this knowledge, what cognitive processes they use, and what challenges they encounter when teaching a learner to classify text documents. (In this study, we use rich text format documents, in which the text font has attributes such as size and weight.) We synthesize our observations into a set of design opportunities for MT tools to support knowledge decomposition and discuss how these design opportunities can inspire new interventions beyond existing tools.

Our work offers the following contributions to the fields of MT research and HCI at large:

- An empirical study of knowledge decomposition during a MT task (i.e. building a classifier for text documents).
- Frameworks for characterizing the types of knowledge that people want to teach and how they organize/represent this knowledge, as well as the steps and challenges they encounter during the knowledge decomposition process.
- A set of design opportunities for new MT tools to support user needs during the knowledge decomposition process.

[1]We use the term "knowledge decomposition" in MT to indicate something different from the term "knowledge decomposition" in cognitive tutors literature (i.e., [10])

BACKGROUND AND RELATED WORK

Interactive Machine Learning and Machine Teaching

People interact with ML processes from different roles (e.g., data scientist, producer of labels, consumer of predictions, etc.), and at different stages (e.g., algorithm selection and tuning, labeling, data cleaning, etc.). While the above processes may be called "interactive," we specifically use the term interactive machine learning (IML) as described and reviewed in the literature such as [12, 13, 31]: "an interaction paradigm in which a user [...] iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review", where humans in the loop take part in "rapid, focused and incremental learning cycles" [1].

MT [2] builds on IML by specifying the contract (the information and knowledge exchanged) between the human in the loop - in this case, the teacher - and the machine learner. [27] presents MT's fundamentals and describes it as a process where any information processing skill teachable to a human should be as easily taught to a machine. [34] introduces a MT environment, MATE, and uses it to formalize aspects of the MT process. Furthermore, they provide insight into expert teaching patterns to help novice users become expert-like teachers. Microsoft's LUIS [21] service supports creation of simple language-understanding ML models on short sentences. During a LUIS teaching session, users input labels, semantic features and entities, and take advantage of predefined models, features and entities. While both of these systems allow users to engage in teaching, they do not provide comprehensive support for knowledge decomposition.

We advance the MT research agenda by diving into understanding the knowledge decomposition process in the context of a common MT task (i.e., building a classifier for text documents). Furthermore, we seek to identify design opportunities for how MT tools can be designed to support the knowledge decomposition process more explicitly (e.g., by providing technological interventions at the proper teaching moments).

Knowledge Engineering

Expert Systems (ES) [14] were introduced in 1980's as a promising technology that implemented automated decision-making and consisted of a knowledge base and an inference engine that could reason over it. To build these systems, Knowledge Engineers [29] interview subject-matter experts and translate expert knowledge into the system's knowledge base. Although the task of collecting subject-matter knowledge has been named Knowledge Elicitation (in ES context) or Knowledge Decomposition (in MT context), Knowledge Elicitation and Knowledge Decomposition differ in important ways. Specifically, building an ES follows a waterfall model where knowledge acquisition happens before its representation and the system's deployment. This process demands technical expertise and considerable resources from a knowledge engineer [11, 16]. In contrast, when building a model through MT, a teacher is the source of subject-matter knowledge and takes part in an iterative, incremental knowledge exchange,

[2][36] uses the term "machine teaching" to describe something different: finding the optimal training set, given a particular learning algorithm and a target model.

through knowledge decomposition, with an always-learning system. In a MT process, subject-matter experts are in control of creating the ML model, while doing tasks previously reserved to a knowledge engineer. These distinctions are critical because they introduce new challenges with understanding and supporting what non-ML experts need and can offer when teaching directly to a learning algorithm, rather than to a human engineer. Our work explores these challenges by studying the knowledge decomposition process during a MT session.

Knowledge and Decomposition

There are 2 ways that human knowledge can be used to train learning algorithms. Unsupervised learning requires indirect human knowledge as raw data, from which patterns and structures are extracted. Supervised learning requires direct human knowledge as inputs, generally presented as labels that map between pre-configured or automatically-extracted features and prediction outputs. In both contexts, humans are treated as data generators or label oracles. However, humans naturally want to provide more than just raw data and labels [1], and taking advantage of their domain knowledge throughout the process of training a model has benefits. For example, humans can jump-start an active learning system by explicitly searching for and discovering positive items before learning starts [3, 7], suggest new or alter existing features [30], and directly manipulate weights of features [18] or label features [8].

Knowledge decomposition is a cognitive process. Our research builds on cognitive science theories of processes that humans use to perform classification tasks in everyday life. For example, the Prototype theory, the Exemplar theory, and the Knowledge theory all posit that humans develop "concepts" or mental representations of categories as they encounter things throughout life, and every new thing is classified by comparing it to the "concepts" that are stored in memory [22]. These concepts are composed of many different forms of knowledge, including features or dimensions, values along each dimension, weights (e.g., how important each feature or dimension is), criterion levels, and schemas (e.g., relations or constraints between features or dimensions). Neuroscience has shown that the processes of learning, representing, and retrieving concepts are automatically activated in the brain by recognizing, naming, imagining, and answering questions about the concepts [4, 20, 23, 33]. Some cognitive research studies have prompted people to explicitly decompose the knowledge they use to mentally represent concepts by using a "property generation task," in which they are asked to list as many properties as they can think of related to a given concept [25].

Our research expands on these prior works by extracting the implicit knowledge and processes that people use when teaching a machine how to perform an ML task (i.e., multi-classification of text documents). Specifically, we sought to answer the following research questions:

RQ1. What **types of knowledge** do people want to teach?

RQ2. What **types of decomposition structures** do people use to represent the knowledge they want to teach?

RQ3. What **cognitive processes** do people use to identify and express the knowledge they want to teach?

RQ4. What **challenges** do people face during knowledge decomposition?

METHOD

The purpose of this exploratory study is to inform the design of new MT tools by observing the knowledge and decomposition processes that people naturally want to use when teaching a machine. Thus, we used a pen-and-paper wizard-of-oz-styled task with a think aloud and interview procedure in order to not constrain participants' behaviors to the limitations of existing tools. This allowed them to express knowledge that they want to teach in ways that current systems do not support yet by explaining it to us through natural language.

To simulate a MT task, we asked participants to teach a machine how to perform a multi-label classification task on text documents. We chose the task of creating a classifier, because it is a common one that has a variety of everyday applications (e.g., organizing emails, news articles, etc.) and can be complex enough to elicit a rich set of knowledge from participants.

Participants

In order to collect generalizable observations on knowledge decomposition across people with different types of expertise, we recruited participants from a diverse range of professional roles and years of experience in each role. 20 employees from a large technology company participated in our study (12 female, 8 male). All participants were between 18-59 years of age (6 18-29 years, 5 30-39 years, 6 40-49 years, 3 50-59 years), and held some form of university degree as their highest completed level of education (5 Bachelor's, 12 Master's, 3 Doctoral). Many participants had multiple years of experience in multiple types of professional roles throughout their career, including 10 Engineers (2-28 years), 6 People Managers (1-25 years), 7 Program/Project manager (1-21 years), 5 Designers/Creatives (4-16 years), 10 Researchers/Scientists (1-10 years), 3 Administrators (2-30 years), and 1 Other (6 years). Most participants were not ML experts: 12 reported that they had heard of ML but never built a model, 4 reported that they had taken some classes about ML but never built a model in practice outside of those classes, 3 reported that they occasionally build ML models in practice, and 1 reported that they frequently build ML models in practice.

Materials

To simulate a multi-label classification task, we collected 60 recent online articles from a news aggregation service [35] – we crawled 20 articles from news sites in each of the following categories: "Food", "Business", and "Any." We restricted all queries to articles in the English language. We chose "Food" and "Business" as the 2 primary labels for our classification task, because they are common news article topics that participants are likely to be familiar with, and subjective enough that participants would bring their own subject-matter expertise to labeling an article to be "Food", "Business", both, or none.

In addition to printed copies of the articles, we gave participants colored pens, markers, highlighters, Post-it flags, Post-it notes, and large poster papers to help them communicate any knowledge that they wanted to teach the machine.

Figure 1. Examples of participants' completed deliverables for the labeling and annotation task (left), and knowledge summary task (right).

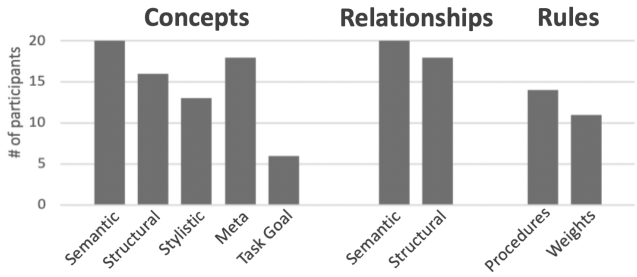

Figure 2. The types of knowledge and number of participants who wanted to teach each type for a text-based multi-label classification task in our exploratory study.

Procedure

Each participant's study session lasted for 1 hour, during which they performed a MT task, followed by a semi-structured interview about their task experience.

For the MT task, participants were asked to teach a hypothetical learning machine how to assign the labels of "Food" and "Business" to news articles. They were told that the two labels are not mutually exclusive and that there were no right or wrong answers, so they could use their own definitions for "Food" articles and "Business" articles. As such, participants were inherently experts in the subject-matter for the task: their own opinion on article topic.

To elicit knowledge from participants, we asked them to perform 2 types of teaching tasks:

Labeling and annotation task. We asked participants to find at least 2-3 examples of articles for each of the 2 labels (i.e., "Food", "Business"), and annotate which parts of the articles helped them determine how to label them. We allowed participants to count a multi-label document as both an example of "Food" and an example of "Business".

Knowledge summary task. We asked participants to create a summary of all the useful knowledge they think the machine needs in order to label articles as "Food" and/or "Business" by writing and/or drawing them on Post-it notes and large poster paper. We allowed participants to include any knowledge from the articles or from their own memory and to structure their knowledge summary in any way.

Figure 1 shows examples of a participant's completed deliverable from each task. Since a primary goal of this study is to observe participants' natural processes for these MT tasks, we instructed them to complete them in any order they would like rather than prescribing a specific procedure.

During the tasks, we simulated the machine's reactions to participants' teaching by asking think aloud questions. To avoid researcher bias and maintain consistency across all participants while eliciting knowledge during think aloud, we developed a set of systematic rules and prompts to moderate study sessions. For example, whenever a participant assigned a label to an article, we asked "How do you know that this article is about [label]?", and whenever a participant articulated a concept that is not explicitly written in an article, we asked "How do you know if an article is about or contains [concept]?"

After each participant completed the 2 teaching tasks, we conducted a semi-structured interview to better understand their cognitive process and challenges that they had encountered.

Data Analysis

We collected 3 forms of data from each participant: (1) Labeled and annotated articles, (2) Knowledge summary, and (3) Audio and video recording of study session, transcribed into text via a third-party transcription service.

We used a grounded theory approach to qualitatively analyze all 3 forms of data from each participant, following Strauss and Corbin's process of first open coding, followed by axial coding, and finally selective coding [28]. For each research question, coding was conducted in an iterative manner: the first author developed an initial set of codes from the dataset, which the second and third authors then reviewed and adjusted based on their own interpretations of the dataset during a series of immersive group meetings. All disagreements were resolved through group discussion, and the updated set of consensus codes was ultimately applied across the whole dataset.

RESULTS

RQ 1: Types of Knowledge

We identified 3 broad categories of knowledge types that participants thought were useful for the learner to understand in order to perform the multi-label classification task - concepts, relationships, and rules - as well as sub-categories within each.

Concepts are ideas or notions related to the decision to be made. We identified 5 sub-categories of concepts that participants wanted to teach:

Semantic concepts are dependent on label meanings (i.e., In our study, the labels were "Food" and "Business"). Examples included "types of food," "food actions/verbs," "recipe," "types of business," "buying/selling," and "money." For many semantic concepts, participants listed keywords or symbols that are indicators of that concept (e.g., keywords like "cookie" and "beef" are indicators of the concept "types of food").

Structural concepts are independent of label meanings, but dependent on components of the data type (i.e., In our study, the data type was text-based news articles). Examples included "title," "sub-headers," "author," "paragraphs," "ordered lists," "sentences," and "words."

Stylistic concepts depend only on the overall data type (i.e., In our study, the overall data type was text). Examples included linguistic style concepts such as "language" (i.e., English), "tone," "informality," and "figures of speech," as well as visual style concepts such as "font type" and "font size."

Meta concepts are independent of label meanings, independent of data type, and computable as a function of other types of knowledge. Examples included *implicit* meta concepts that require some level of subjective interpretation to determine, such as "main subject," "intended audience," and "goal/intent" of a news article, as well as *explicit* concepts that can be determined directly based on the article text, such as the "presence," "frequency," and "repetition" of keywords or concepts.

Task goal concepts are externally- or user-imposed constraints that are not computable from the data itself. Examples included "my personal interests/non-interests" such as "real or fake news" and "Thai food", as well as "my objective" such as speed and accuracy of the ML model.

Relationships, or schemas, describe relations and constraints between concepts. We identified 2 sub-categories of relationships that participants wanted to teach:

Semantic relationships are based on the meanings of labels and concepts. Examples included taxonomical relationships (e.g., "types of food" is a sub-concept of "food"; "company" is an example of "types of business"), positive/negative association (e.g., "market" is associated with "business"; "business" and "politics" are sometimes related, "suicide" is rarely related to "food" or "business"), and mathematical relationships (e.g., word count is greater than five).

Structural relationships are independent of label or concept meanings, but dependent on the data type. Examples included co-occurrence (e.g., words/concepts appear together; presence of certain words/concepts in the absence of other words or concepts), and spatial relationships (e.g., "cookie" is in the title; "$" is before a number).

Rules describe how to apply/combine concepts and relationships to assign labels to documents. We identified 2 subcategories of rules that participants wanted to teach:

Procedures are sequences of steps or if-then statements for how to assign labels to documents. These included instructions on the order in which actions should be performed (e.g., First look at the title and find these keywords. Next, look at subsection headers. Then look at the body.) and criterion that need to be met in order to assign labels (e.g., "If the frequency of these keywords is greater than 5, then label it as "Food.").

Weights are degrees of strength or confidence that the user subjectively assigns to each concept, relationship, procedure, or label. Participants expressed this type of rule in many different ways. For example, some used the words "strong" vs. "weak" or a 1-3 star rating system to indicate the importance of each keyword, concept, or relationship to a label, while others assigned numerical confidence scores to each concept (e.g., If you see this set of words, then 90% sure it's this label. If you see this other set of words, then 70% sure it's this label. If you see this final set of words, then 50% sure it's this label.)

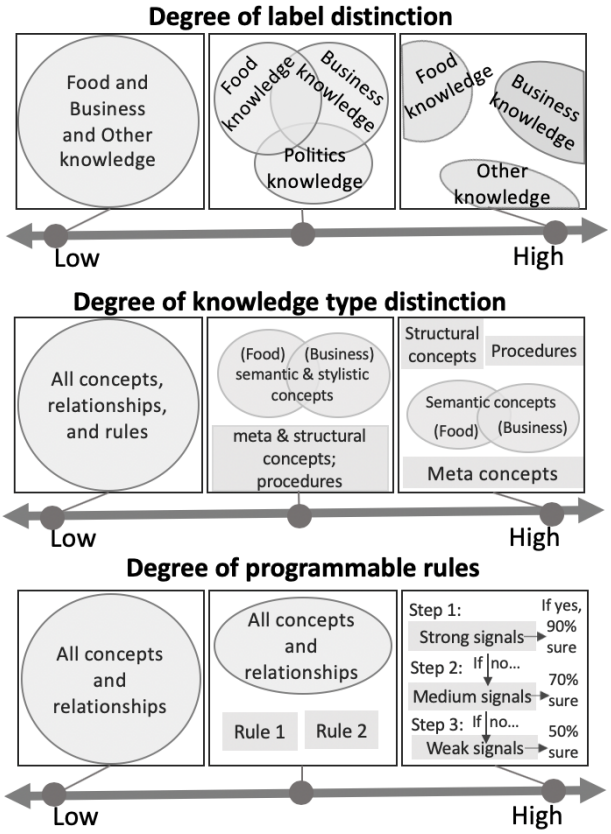

Figure 3. Schematic examples of participants' decomposition structures at low, medium and high points of the spectrum for each of the 3 key dimensions from our exploratory study.

Figure 2 shows the number of participants who wanted to teach each type of knowledge. For brevity, we refer to each concept, each relationship, and each rule that a participant identified as a "knowledge unit" from now on.

RQ 2: Types of Decomposition Structures

We observed 3 key dimensions along which the decomposition structures in participants' knowledge summaries varied: degree of label distinction, degree of knowledge type distinction, and degree of programmable rules. Figure 3 shows schematic examples of participants' decomposition structures at low, medium, and high points of each dimension's spectrum.

Degree of label distinction refers to how participants categorized each knowledge unit in relation to label categories. For example, 7 participants created decomposition structures with mutually exclusive categories of knowledge for each label (i.e., "Food" knowledge vs. "Business" knowledge). In contrast, 11 participants created decompositions with a Venn Diagram structure such that there were additional categories of knowledge for multiple labels (i.e., "Food AND Business" knowledge). The other 2 participants created decompositions that did not organize knowledge units by label category, but rather by knowledge type (see next paragraph).

Degree of knowledge type distinction refers to how participants categorized each knowledge unit in relation to knowledge type. For example, 11 participants created decomposition

structures that contained separate groups for different knowledge types (e.g., for semantic concepts, for structural concepts, for relationships, for rules), while others did not segregate knowledge units by type, but rather solely by label category.

Degree of programmable rules refers to how participants organized knowledge units together in relation to a set of step-by-step operations. For example, 2 participants created decomposition structures that were a complete set of logic statements on what concepts and relationships to look for and how to combine them together into a label decision on each document. In contrast, other participants created decomposition structures that were simply a set of concepts and relationships without any specific procedures or only a few unconnected rules.

These 3 dimensions were orthogonal to one another (i.e., some decomposition structures were high on multiple dimensions, while others were high on 1 dimension and low on the other dimensions). However, the choices that each participant made on these dimensions reflected their mental model of how the learner works. We discuss implications of this observation in more depth in the upcoming section "RQ 4: User Challenges."

RQ 3: Knowledge Decomposition Processes

We found that participants used an iterative sensemaking process to identify and express the knowledge that they wanted to teach the machine. Specifically, we identified 5 discrete steps that they performed, which we discuss through the lens of the sensemaking process's components outlined by Pirolli & Card: the Foraging Loop and the Sensemaking Loop [24]. Figure 4 illustrates these 5 steps and the iterative nature in which they were performed.

In the **Foraging Loop**, participants sifted through many examples to develop a rough idea of what types of documents and knowledge units exist and which ones could be useful vs. not useful for the classification task. No committed decisions were made in the Foraging Loop.

Search for knowledge - In this step, participants searched through examples either from their own memory or from the documents in front of them to find knowledge units that could be useful for the classification task. This step was performed in the beginning of the task and sometimes returned to during the Testing steps of the process. When performed in the beginning of the task, participants specifically searched for positive and negative examples for each label category. When performed during the Testing steps, participants searched for examples containing specific knowledge units that they were testing.

Shoebox knowledge - In this step, participants stored potentially useful documents and knowledge units, either by physically marking and putting them aside for later examination or by mentally making note of them. Documents were often shoeboxed into separate piles, according to participants' initial categorizations (i.e., Food, Business, Both, Neither, Not Sure). Knowledge units were often shoeboxed by annotating (e.g., highlighting, underlining, circling) them directly on the documents and/or scribbling them on Post-it notes that were temporarily stored for later testing.

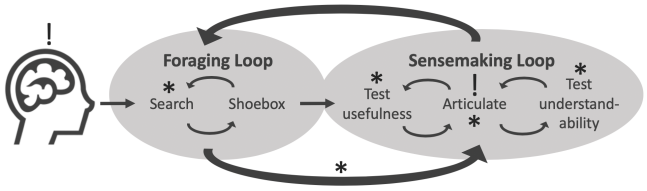

Figure 4. The knowledge decomposition process that participants used to identify and express the knowledge that they wanted to teach the machine in our exploratory study. !'s indicate points where major user challenges occurred. *'s indicate points where participants reported wanting specific supporting resources.

In the **Sensemaking Loop**, participants took each knowledge unit that they identified in the Foraging Loop and individually tested and articulated the unit to inform their decision on if/how to incorporate it into the knowledge summary that they want to teach the machine for the classification task.

Test usefulness of knowledge - In this step, participants used a variety of strategies to find supporting or counter evidence for whether a knowledge unit can be reliable used to inform what label to assign documents. These strategies included: word association (in which a knowledge unit was deemed useful if it automatically made the participant think of the label), searching for other positive examples (in which a knowledge unit was deemed useful if was consistent across other documents with the same label), and searching for counter examples (in which a knowledge unit was deemed unuseful if the participant could identify situations in which the knowledge unit would identify other documents without the same label). Participants added useful knowledge units to their knowledge summaries, and excluded unuseful ones.

Articulate knowledge - In this step, participants externalized knowledge units to the machine by verbalizing them using natural language and writing or drawing them on paper. Concepts were expressed as words or phrases written on Post-it notes or on the poster. Relationships were expressed through written descriptions on Post-it notes or on the poster, spatial arrangements of Post-it notes (e.g., physical distance indicating semantic distance between concepts), and/or annotations around Post-it notes (e.g., drawing circles or boundaries between groups of Post-it notes). Rules were also expressed through written instructions on Post-it notes or on the poster, spatial arrangements of Post-it notes (e.g., sequential steps placed in a vertical line; concepts with higher weight placed in the center and concepts with lower weight placed in the periphery), and/or annotations around Post-it notes (e.g., arrows to indicate the order in which procedures are to be performed; stars to indicate the weight of each concept).

Test understandability of knowledge - In this step, participants sought feedback on whether the way in which they articulated a knowledge unit was understandable by the machine, or whether they needed to further decompose it into simpler terms that the machine could understand. During the study, participants recognized the think aloud prompts (e.g., "How do you know if an article is about or contains [concept]?") as helpful signals for testing the understandability of their knowledge articulations, and relied on our questions to know

when to continue or stop decomposing. Some participants even proactively asked whether we had any questions after articulating knowledge units to test their understandability.

RQ 4: User Challenges

We identified 2 major challenges that participants faced during knowledge decomposition: understanding how the learner works, and articulating abstract and implicit knowledge.

Understanding how the learner works - Each participant came into the study with their own preconceived notions about how a learner works, which biased the types of documents and knowledge units that they searched for and articulated in their knowledge summary.

A common limiting assumption that participants had was that the learner is building a decision tree. This was reflected in participants' decomposition structures with high degrees of label distinction. Participants with this assumption were strongly biased towards teaching semantic concepts that are mutually exclusive to 1 label category, intentionally excluding semantic concepts that could be useful for multiple label categories. These participants found it particularly challenging to teach the machine how to handle documents with multiple labels. Some even avoided assigning multi-labels to documents all together, as P04 explained:

"*You saw me hem and haw on the multi-label thing. That was clearly challenging... I find multiple labels to be less helpful than single labels, and so I feel like it's a bit of a cop-out... 'Cause look, they all have both food and business items in it. Now you've just called it both and move on, but I don't think that's as valuable... And so, I just wanted to be, I guess, a little bit more cautious about the multi-labeling thing. But that implies a whole bunch of judgment around balance and that type of thing, which is hard to explain.*"

This assumption was also reflected in participants' decomposition structures with high degrees of programmable rules. For example, participants with this assumption thought that they needed to assign explicit weights and procedures to each knowledge unit in order for the machine to know how to combine them into a label decision, which limited the types of knowledge that they articulated to only those for which they could define specific rules (e.g., lists of keywords and if/then rules around word frequency). P16 explained:

"*I did look at if I were to code this, how would I code it, right? So that somewhere was in my head because like, I try to [identify] words which would mean something in an if-then statement.*"

Articulating abstract and implicit knowledge - 7 participants reported that the most challenging knowledge units to articulate were abstract and implicit concepts, relationships, and rules. For example, P09 explained about abstract knowledge:

"*Business is so abstract... There are a lot of foods that you could put into a list of foods - it seems like a probably never ending list, but I feel like that's a lot easier. It's just more concrete to think of than something like business, where there are probably tons and tons of sub-concepts. Food you can group in different ways, like where it's from and a lot of that sort of thing, but I feel like people have a lot, probably wider interpretation of how to group different types of business.*"

Similarly, P20 explained about implicit knowledge:

"*I think the concept of something implied, that's not in the article, was hard to explain. A lot of my prior knowledge was the reason I was able to classify these articles. There may not have been the information in the article itself, so it was hard to say what was native to the article.*

Examples of implied knowledge that participants reported having difficulty explaining how to recognize included "intended audience" of an article (P18) and "context" (P16).

In addition to these 2 major challenges, participants also reported 4 categories of supporting resources that they wanted to facilitate their knowledge decomposition process. While some of these reflect the limitations of our study setup, we still find them useful for informing the design of future MT tools:

Ability to search through more and varied samples - Most participants reported feeling their knowledge summaries were incomplete at the end of the session. To address the gaps in their knowledge summaries, 12 participants wanted the ability to search and filter through a greater number and variety of articles. They said that this would help them identify more useful knowledge units that are common across articles of each label, and to test whether knowledge units that they had identified in one article are generalizable and strong signals across other articles. For example, P16 said:

"*...read more articles on these two subjects to get more identifying words and identifying sequence of words for context... I think that would certainly help. And because it will also help identifying those core common words ... Because I think in this [Food] article, the really common words, like meat, drink, food, beverage, dining... Here [in this Business article] it's all about these words like market segmentation and trends... So I would like to identify more of those common words, which are like the intersection words, if you may, across these articles which fall in that bucket or this one.*"

Access to existing lists, dictionaries, and lexicons - During the task, many participants expressed semantic concepts by referring to real or imaginary databases of keywords, synonyms, definitions, or examples related to those concepts (e.g., "types of food," "Fortune 500 businesses"). In addition, during the post-task interview, 9 participants explicitly reported that being able to pull these from existing resources would make it easier to teach broad concepts to the learner in a more efficient way. For example, P03 said:

If I had a dictionary of global company names that restaurants, diners, bars, cafes... then I'm somehow able to automatically apply it to the article... then I'm done. Like existing vocabularies dictionaries that I can apply to my search automatically. Similarly, "food vocabulary" - like I think this category should have existing dictionaries that I can apply to and that will automatically get this done real quickly so I don't need to come up with those words."

Flexibility to revise knowledge decomposition - During the task, many participants rearranged the Post-it notes and annotations on their knowledge summary as they iterated between the Foraging Loop and the Sensemaking Loop. Also, during the post-task interview, 2 participants explicitly reported wanting the freedom and tools to revise their knowledge summary, as their mental representation of some concepts (e.g., abstract ones like "business"), relationships, and rules evolved as they read more articles and encountered new knowledge units. For example, P07 said:

"*If I can mark things a certain color, and I build that distinction, and then I have the flexibility to keep moving, so ... some container that enables me to drag concepts, and as you drag those concepts, drag elements that will be related to those concepts. ... So eventually as you were elaborating and refining your map, you might want to create super categories and sub-categories to refine it. So that flexibility, the ability to easily drag things and coloring the way... that would help me see things visually and easily.*"

Feedback on knowledge usefulness and understandability - While many participants relied on our think aloud prompts to test the understandability of their knowledge unit articulations during the study session, most participants reported during the post-task interview that they felt unsure about whether the knowledge they had articulated in their summaries are useful to the learner. In addition, 5 participants said they wanted more feedback in order to test whether the knowledge units they had identified are indeed useful and/or understandable to the machine for the classification task and to help them evaluate or debug their knowledge summaries. For example, some participants requested feedback on the ML model's label predictions and confidence levels on new articles given the knowledge that they had articulated so far in order to identify situations in which the machine's predictions are incorrect and how to improve their knowledge summary. P01 explained:

"*I would like to get feedback from the AI. Immediate feedback... You can start even with as simple as a smile, like, did you understand? 'Cause if you don't understand, I have to find another way to describe the same thing... Or it can be very specific ... the results say [this document label] now is this and that. And I can go through there and see how dirty it is, and then clean by getting better... I would imagine it will... come out like this [with my assigned labels]. And if it didn't come out like this, then I will have to say, where you went wrong, and I would try to do better.*"

DISCUSSION

Synthesizing the findings from our exploratory study, we identified 5 design opportunities for MT tools to support knowledge decomposition. We incorporated the types of knowledge that we identified from RQ1, the decomposition structures that we identified from RQ2, and the user challenges that we identified from RQ4 to generate a set of user needs, and we used the knowledge decomposition process that we identified from RQ3 as a framework to describe when each user need could be better supported by new design interventions.

Design Opportunities

For each design opportunity below, we describe the underlying user needs and when in the knowledge decomposition process to support them, we discuss how existing tools address (or do not address) the design opportunity, and we provide examples of how our findings could inspire new design ideas.

Support useful mental models of how the learner works

Our findings revealed that users need an accurate or at least useful mental model of how the learner works before the knowledge decomposition process even begins. We found that a user's mental model may consist of incorrect assumptions about what types of knowledge the learner can use and how the predictions work (e.g., whether it is a strictly rule-based system), and that some users have natural preferences for label distinctions (e.g., exclusivity between target classes). To support these user needs, teaching tools should be designed to reduce misguided user decisions during knowledge decomposition by preempting or correcting such misconceptions.

Existing ML, IML, and MT tools use tutorials, explanations, and system feedback to help users develop useful mental models of how the overall learning system works. For example, AuPair users received a 30-minute introductory tutorial with illustrated examples on how the system determines music recommendations, the types of features it "knows" about, and how it extracts this information from audio files [19], and novice MATE users received introductory tutorial materials on what types of knowledge the system needs (i.e., samples, labels, features) in order to make label predictions [34], while EluciDebug provides feedback on how features were used to predict the topic of email messages [18].

Our work identifies several design opportunities for new MT tools to expand on these approaches to help guide users toward accurate understandings of how to affect learning or prediction behaviors through knowledge decomposition. For example, illustrated examples of each type of concept, relationship, and rule can be provided prior to the start of a MT session to break limiting preconceptions about the types of knowledge that users can teach, and inspire them to provide more diverse forms than only keywords and if-then statements. Another example could be designing systems where teachers can directly manipulate the learner's parameters, the predictor's parameters, or both. Opportunities include expansions on interventions like the ones presented in [34], where a system can advise at certain points during a teaching session that instead of providing more labeled examples, a teacher should instead add new semantic features. Tools can also be designed to support awareness of different types of knowledge that can be taught by providing different sections of the interface or different types of interactions for inputting different types of knowledge (e.g., text entry for keywords and semantic concepts; spatial manipulation of visual elements for relationships). Additionally, misconceptions about a strictly rule-based system might be broken by designing feedback functions showing the user that the model is capable of combining all articulated knowledge units into valid label predictions without the user explicitly specifying weights and step-by-step operations on each concept or relationship.

Support user search and filter through many varied samples
During the Searching step in the Foraging Loop of the knowledge decomposition process, user needs that we identified included finding: positive and negative examples of target labels, examples containing target knowledge units, and examples with new knowledge units that have not been encountered before. To support these user goals, MT tools should be designed to enable efficient identification of each of these types of documents and knowledge units.

Prior work has evaluated machine-initiated or user-initiated techniques in searching for useful samples. For example, active learning methods in ML use a machine-initiated approach to algorithmically identify unlabeled samples in the dataset that may improve the model and query users to label those samples [26]. IML tools such as AnchorViz [31] support user-initiated approaches to exploring samples in the dataset by visualizing the similarity between semantic concepts or documents, and allowing users to decide which samples to label. MT tools such as MATE [34] have sampling functionalities that allow users to specifically search for positive examples of documents with target labels or target keywords.

Building on these existing approaches, our work identifies several design opportunities for new MT tools to support knowledge unit-based search in addition to document-based search. For example, we might design tools that help users identify new knowledge units by visually clustering documents and inspecting for commonalities or algorithmically extracting commonalities between documents to suggest new knowledge units for the user to consider. Additionally, we might design tools that identify specific knowledge units that require additional examples before being useful for learning. Expanding beyond search through examples in the sample set, we might also design tools that help users search through examples in their own memory by prompting them with suggestions or questions that break them out of fixation on the limited sample set, and make them think about more diverse knowledge units from other hypothetical examples.

Support flexible user knowledge evolution
During iterative transitions between the Foraging Loop and the Sensemaking Loop of the knowledge decomposition process, a user need that we identified is the flexibility for knowledge evolution, as users encounter new data that changes their mental representation of a target concept or their evaluation of whether certain knowledge units are still useful. To support this need, MT tools should enable users to dynamically add, remove, edit, and restructure knowledge units, and provide a temporary design space for underdeveloped knowledge units.

Existing ML, IML, and MT tools offer limited support for knowledge evolution. For example, traditional ML tools typically require users to re-label large numbers of samples to reflect updated mental representations of concepts. Some IML tools use structured labeling techniques to facilitate more efficient re-labeling during concept evolution [17], but they are not designed to address evolution of other knowledge types. MT tools like MATE [34] allow users to add, remove, and edit concepts in the model through text entry, but are also not designed to support evolution of other knowledge types.

Our work identifies design opportunities for new MT tools to better support evolution of different knowledge types. For example, we might design interfaces that align more closely with users' knowledge decomposition processes by providing separate spaces for conducting intermediary steps (e.g., temporarily shoeboxing documents and knowledge units during the Foraging Loop vs. articulating useful knowledge units and comparing different versions of articulations for a knowledge unit during the Sensemaking Loop). Alternatively, we might design a functionality that allows users to toggle individual knowledge units "on" or "off" to easily edit the model without necessarily removing knowledge that they may later decide to add again. Additionally, to expand evolution support beyond concepts, we might design interactions that allow users to explicitly revise previously articulated relationships or rules by adding newly identified conditions or exceptions.

Support user articulation of varied knowledge types
During the Articulating step in the Sensemaking Loop of the knowledge decomposition process, user needs that we identified included the articulation of semantic, structural, stylistic, meta, and task goal concepts, as well as semantic and structural relationships, and rules in the form of procedures and weights. To support these needs, MT tools should provide interaction languages for the user to express these varied types of knowledge (including abstract and implicit ones).

Existing ML, IML, MT tools are limited in the types of knowledge articulation that they support. A typical ML process only supports human labels as a form of input while using automated feature selection methods (e.g., bag-of-words), which is successful in practice but hard for humans to interpret. Semantic dictionaries can be articulated manually by the domain experts [32] or imported from external sources such as WordNet [15], but these features are typically limited to semantic concepts. Similarly, MT tools, such as LUIS and MATE [34], support the articulation of labels and (semantic) features that are computed from raw text, as well as some forms of semantic relationships. However, these tools do not currently consider concepts relating to text structure or style.

Our work identifies design opportunities for new MT tools to support the articulation of more types of concepts, relationships, and rules. For example, to better support the articulation of semantic concepts, we might design tools that help suggest related words or phrases to users by enabling them to query existing knowledge graphs and databases such as ConceptNet [9]. To better support the articulation of structural concepts, we might design tools that enable users to indicate specific components of the document, either through a markup language like HTML or through direct manipulation like a click-and-drag interaction. To better support the articulation of relationships and rules, we might design tools that enable the user to input multiple data types, including textual data (e.g., keywords) and graphical data (e.g., spatial arrangements of words on a 2D canvas; circles, lines, arrows between words).

Support user testing and learner feedback on knowledge
During the Testing step in the Sensemaking Loop of the knowledge decomposition process, user needs that we identified included feedback on: the usefulness of each knowledge unit,

the understandability of each knowledge unit articulation, and the performance of the under-construction model.

Existing tools offer some forms of feedback, but are lacking in others. For example, ML tools provide feedback on the overall model performance (e.g., F1 score, precision, recall), but not on knowledge units. IML tools, such as [2, 6], provide feedback to help users assess the usefulness of knowledge units through increases or decreases in overall model performance scores after their addition or removal. MT tools, such as MATE [34], also provide feedback to help users assess the usefulness of knowledge units through changes in label predictions for documents in the sample set after the addition or removal of knowledge units. However, none of these tools provide feedback on the understandability of knowledge units.

Our work identifies several design opportunities for supporting more nuanced feedback on the usefulness and understandability of individual knowledge units. For example, to provide feedback that helps users assess the usefulness of knowledge units, we might design tools that help users interpret the discriminatory power of a knowledge unit across positive vs. negative samples of each label by visualizing how often it appears across different subsets of documents in the sample set. (i.e., If a knowledge unit appears exclusively in documents of 1 label and not other labels, then it is likely useful for discriminating between the labels). Similarly, to provide feedback on the understandability of knowledge units, we might design tools that help users compare their own expectations of what an articulated knowledge unit means to what the learner thinks it means by visualizing areas of documents where the machine recognizes the presence vs. absence of the knowledge unit. (i.e., If the learner indicates the presence of a knowledge unit in a place where the user does not think it is present, then this indicates an understandability issue.). In addition, to provide feedback on how confused the learner is about the understandability of a knowledge unit, we might design tools that sort such visualizations by the model's confidence level on the presence/absence of knowledge units in each document.

Limitations and Future Work

Our exploratory study findings and the design opportunities that we derived from them help build foundational frameworks for understanding and supporting knowledge decomposition, and pave the way for future MT research and design. Below, we describe 2 areas in which our work can be further expanded.

Evaluating knowledge decompositions

The pen-and-paper wizard-of-oz-styled task was appropriate for our exploratory purposes. However, a trade-off of this approach is that we were not able to implement any of the knowledge that participants wanted to teach into a real ML model or provide feedback on model performance. Future work is needed on how to implement each type of knowledge, and to measure how different types of knowledge units and decompositions affect model performance.

More broadly, we need more work on how to evaluate knowledge decompositions beyond traditional ML metrics (e.g., F1 score, precision, recall). The reusability of model components (i.e., the ability to share knowledge units across MT tasks), the transparency of model predictions and explanations, and the ease of the knowledge decomposition process for non-ML experts are all important to the success of MT and require systematic evaluation. The types of knowledge and dimensions of decomposition structure that we developed from this study can provide a foundation for evaluation rubrics, but additional work is needed to understand which types of knowledge and which types of processes are more effective in what contexts.

Knowledge decomposition for other ML tasks

While the findings, frameworks, and design opportunities presented in this paper stem from a study on classification of text documents, we hypothesize that they can generalize across different data types (e.g., images, videos) and across other MT tasks (e.g., for regression, clustering, etc.) as well. For example, the types of knowledge that we identified in our study can guide the ideation of tools for knowledge articulation with images and videos: To support the articulation of *structural concepts* and *structural relationships* in images, we might design tools that enable users to indicate components of the image as "foreground," "background," "on top of," "bottom left corner," etc.; in videos, components such as "first 5 seconds", "center of frame", "end credit". Similarly, to support the articulation of *stylistic concepts* in images, we might design tools that enable users to indicate the art style (e.g., abstract, minimalist, realist, etc.) or art medium (e.g., pencil, acrylic, digital, etc.); in videos, film styles (e.g., animated, live-action, black-and-white, etc.) or film genre (e.g., action, comedy, horror, etc.). Future work should apply our frameworks across a variety of additional data types and ML tasks to see whether there are other knowledge or processes used for them that are not yet captured by the current study.

Furthermore, although we recruited participants from a diverse sample of professional experiences within the constraints of our resources, future work should also validate our frameworks with a larger, more varied population outside of large technology company employees. Together, these extensions of our work will help build a more comprehensive framework to inform the design of future tools that support knowledge decomposition in a broader range of contexts.

CONCLUSION

MT is a promising approach to building ML models in a more accessible, efficient, and semantic way than traditional ML approaches by allowing users to teach richer forms of knowledge to the machine learner than simply labels. To guide the design of effective MT tools, we built foundational frameworks for understanding and supporting knowledge decomposition by conducting an exploratory study on the knowledge and processes that people want to use when teaching a machine how to perform an ML task. From our findings, we derived a set of design opportunities to better support user needs during key points in the knowledge decomposition process. In the future, we encourage researchers and designers to build on our work by further exploring these design opportunities and expanding upon them in the context of different types of ML tasks.

REFERENCES

[1] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. 2014. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine* 35, 4 (2014), 105–120. http://www.aaai.org/ojs/index.php/aimagazine/article/view/2513

[2] Saleema Amershi, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 337–346. DOI: http://dx.doi.org/10.1145/2702123.2702509

[3] Josh Attenberg and Foster Provost. 2010. Why Label when You Can Search?: Alternatives to Active Learning for Applying Human Resources to Build Classification Models Under Extreme Class Imbalance. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '10)*. ACM, New York, NY, USA, 423–432. DOI: http://dx.doi.org/10.1145/1835804.1835859

[4] David Badre and Anthony D Wagner. 2002. Semantic retrieval, mnemonic control, and prefrontal cortex. *Behavioral and cognitive neuroscience reviews* 1, 3 (2002), 206–218. DOI: http://dx.doi.org/10.1177/1534582302001003002

[5] Bonsai. 2019. Bons.ai: BRAINs for Autonomous Systems. (2019). https://www.bons.ai Accessed June 2019.

[6] Michael Brooks, Saleema Amershi, Bongshin Lee, Steven M Drucker, Ashish Kapoor, and Patrice Simard. 2015. FeatureInsight: Visual support for error-driven feature ideation in text classification. In *2015 IEEE Conference on Visual Analytics Science and Technology (VAST)*. IEEE, 105–112. DOI: http://dx.doi.org/10.1109/VAST.2015.7347637

[7] Maya Cakmak, Crystal Chao, and Andrea Lockerd Thomaz. 2010. Designing Interactions for Robot Active Learners. *IEEE Trans. Autonomous Mental Development* 2, 2 (2010), 108–118. DOI: http://dx.doi.org/10.1109/TAMD.2010.2051030

[8] Justin Cheng and Michael S. Bernstein. 2015. Flock: Hybrid Crowd-Machine Learning Classifiers. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing (CSCW '15)*. ACM, New York, NY, USA, 600–611. DOI:http://dx.doi.org/10.1145/2675133.2675214

[9] ConceptNet. 2019. ConceptNet: An open, multilingual knowledge graph. (2019). https://conceptnet.io Accessed June 2019.

[10] Albert Corbett and John Anderson. 1995. Knowledge Decomposition and Subgoal Reification in the ACT Programming Tutor. In *Artificial Intelligence in Education, 1995: Proceedings of the 7th World Conference on Artificial Intelligence in Education*. AACE, 469–476.

[11] E.N. Corlett, J.R. Wilson, and N. CORLETT. 1995. *Evaluation of Human Work, 2nd Edition*. Taylor & Francis. https://books.google.com/books?id=_Uq--hONFDUC

[12] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (June 2018), 37 pages. DOI: http://dx.doi.org/10.1145/3185517

[13] Jerry Alan Fails and Dan R. Olsen, Jr. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 39–45. DOI: http://dx.doi.org/10.1145/604045.604056

[14] E.A. Feigenbaum. 1982. *Knowledge Engineering for the 1980s*. Stanford University Computer Science Department. https://books.google.com/books?id=wlZBtwAACAAJ

[15] Christiane Fellbaum. 1998. A Semantic Network of English: The Mother of All WordNets. *Language Resources and Evaluation* 32, 2-3 (1998), 209–220. DOI:http://dx.doi.org/10.1023/A:1001181927857

[16] Anna Hart. 1986. *Knowledge acquisition for expert systems*. Technical Report. School of Computing, Lancashire Polytechnic, Preston.

[17] Todd Kulesza, Saleema Amershi, Rich Caruana, Danyel Fisher, and Denis Charles. 2014. Structured Labeling for Facilitating Concept Evolution in Machine Learning. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems (CHI '14)*. ACM, New York, NY, USA, 3075–3084. DOI: http://dx.doi.org/10.1145/2556288.2557238

[18] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. DOI: http://dx.doi.org/10.1145/2678025.2701399

[19] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 1–10. DOI: http://dx.doi.org/10.1145/2207676.2207678

[20] Alex Martin and Linda L Chao. 2001. Semantic memory and the brain: structure and processes. *Current opinion in neurobiology* 11, 2 (2001), 194–201. DOI: http://dx.doi.org/10.1016/S0959-4388(00)00196-3

[21] Microsoft. 2019. Language Understanding (LUIS). (2019). https://luis.ai/ Accessed June 2019.

[22] Gregory Murphy. 2004. *The big book of concepts*. MIT press.

[23] Randall C O'Reilly and Jerry W Rudy. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. *Psychological review* 108, 2 (2001), 311. DOI: http://dx.doi.org/10.1037/0033-295X.108.2.311

[24] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.

[25] Ava Santos, Sergio E Chaigneau, W Kyle Simmons, and Lawrence W Barsalou. 2011. Property generation reflects word association and situated simulation. *Language and Cognition* 3, 1 (2011), 83–119. DOI: http://dx.doi.org/10.1515/langcog.2011.004

[26] Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers. DOI: http://dx.doi.org/10.2200/S00429ED1V01Y201207AIM018

[27] Patrice Y. Simard, Saleema Amershi, David Maxwell Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, and Mo Wang an John Wernsing. 2017. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *CoRR* abs/1707.06742 (2017). http://arxiv.org/abs/1707.06742

[28] Anselm Strauss and Juliet Corbin. 1990. *Basics of qualitative research*. Sage publications.

[29] Rudi Studer, V.Richard Benjamins, and Dieter Fensel. 1998. Knowledge engineering: Principles and methods. *Data & Knowledge Engineering* 25, 1 (1998), 161 – 197. DOI:http://dx.doi.org/https://doi.org/10.1016/S0169-023X(97)00056-6

[30] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward Harnessing User Feedback for Machine Learning. In *Proceedings of the 12th International Conference on Intelligent User Interfaces (IUI '07)*. ACM, New York, NY, USA, 82–91. DOI: http://dx.doi.org/10.1145/1216295.1216316

[31] Jina Suh, Soroush Ghorashi, Gonzalo Ramos, Nan-Chen Chen, Steven Drucker, Johan Verwey, and Patrice Simard. 2019. AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 10, 1, Article 7 (Aug. 2019), 38 pages. DOI: http://dx.doi.org/10.1145/3241379

[32] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based Methods for Sentiment Analysis. *Comput. Linguist.* 37, 2 (June 2011), 267–307. DOI: http://dx.doi.org/10.1162/COLI_a_00049

[33] Lorraine K Tyler and Helen E Moss. 2001. Towards a distributed account of conceptual knowledge. *Trends in cognitive sciences* 5, 6 (2001), 244–252. DOI: http://dx.doi.org/10.1016/S1364-6613(00)01651-X

[34] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching. In *Human-Computer Interaction - INTERACT 2019 - 17th IFIP TC 13 International Conference, Paphos, Cyprus, September 2-6, 2019, Proceedings, Part III*. 578–599. DOI:http://dx.doi.org/10.1007/978-3-030-29387-1_34

[35] Webhose. 2019. Webhose.io: Tap Into Web Content at Scale. (2019). http://webhose.io Accessed June 2019.

[36] Xiaojin Zhu. 2015. Machine Teaching: An Inverse Problem to Machine Learning and an Approach Toward Optimal Education. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*. 4083–4087. http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9487