# Oracle or Teacher? A Systematic Overview of Research on Interactive Labeling for Machine Learning

Merlin Knaeble, Mario Nadj, and Alexander Maedche

Karlsruhe Institute of Technology, Institute of Information Systems and Marketing (IISM), Karlsruhe, Germany
{merlin.knaeble, mario.nadj, alexander.maedche}@kit.edu

**Abstract.** Machine learning is steadily growing in popularity – as is its demand for labeled training data. However, these datasets often need to be labeled by human domain experts in a labor-intensive process. Recently, a new area of research has formed around this process, called interactive labeling. While much research exists in this young and rapidly growing area, it lacks a systematic overview. In this paper, we strive to provide such overview, along with a cluster analysis and an outlook on five avenues for future research. Hereby, we identified 57 relevant articles, most of them investigating approaches for labeling images or text. Further, our findings indicate that there exist two competing views how the user could be treated: (a) oracle, where users are queried whether a label is right or wrong versus (b) teacher, where users can offer deeper explanations in the interactive labeling process.

**Keywords:** Interactive Labeling, Interactive Machine Learning, Training Data

## 1 Introduction

Both academia and business have been increasingly attentive towards machine learning (ML) [1], making it one of the fastest growing disciplines [2]. Supervised ML (SL) is a method of choice for many scholars and practitioners. It has been applied to various contexts like natural language processing (NLP) or computer vision [3, 4]. However, SL needs large quantities of labeled training data to successfully accomplish its learning task [5]. Hereby, labeling is the process of information enclosure to objects [5]. For example, one could think of drawing bounding-boxes around pedestrians for self-driving cars.

Labeling is a costly, labor-intensive, and error-prone process, not unlikely to frustrate involved users [5, 6]. Especially repeated queries whether a label is rightly assigned, thus, being treated like a so-called “oracle”, is perceived as undesirable [7], instable, or annoying [8]. Active Learning (AL) research has defined oracles as users without system control who are faced with a steady stream of questions regarding label correctness [7]. To extenuate these negative effects, research suggest accounting for human factors in ML, emphasizing interactivity in particular [7]. Hereby, users could also be treated as teachers who offer deeper explanations in the labeling process.

15th International Conference on Wirtschaftsinformatik,
March 08-11, 2020, Potsdam, Germany

https://doi.org/10.30844/wi_2020_a1-knaeble

Approaches range from basic approval/rejection [8], over label corrections [9] to deeper insights users may be able to give [7]. Despite the relatively young nature of interactive labeling (IL), it has already produced a remarkable number of articles. Analyzing IL in a systematic manner is required due to multiple reasons: First, the numerous articles published need to be structured. Second, although some scholars summarized challenges for improving interactive ML (IML) approaches [7, 9], there is a lack of such guidance in IL for future work to be well directed. Third, to the best of our knowledge, no systematic literature review (SLR) on IL research has been published to date. Forth, IL has been applied by various fields, such as Human-Computer Interaction (HCI), Information Systems (IS), or computer science (CS), leading to a lack of integration of the present work.

In this paper, we illustrate the SLR results in the field of IL. We formulate the following research question (RQ): *What is the state-of-the-art of IL research for ML?* Following acknowledged SLR guidelines [10, 11], we identified 57 relevant articles and employed an established clustering approach [12] to identify patterns.

The remainder of the paper is structured as follows. We lay out the foundations of IL in Section 2, before the SLR method (Section 3) and results (Section 4) are described. Future work is suggested in Section 5. We conclude our paper in Section 6.

## 2 Conceptual Foundations of Interactive Labeling

IL belongs to the field of IML, which is defined as "the process of building ML models iteratively through end-user input" [13]. IML is part of the Human-in-the-Loop (HITL) methodology that tries to decrease the boundaries of fully-automated systems by means of user interaction [14]. HITL and IML intersect with four underlying ML approaches (1) SL, (2) AL, (3) Reinforcement Learning (RL), and (4) Preference Learning (PL).

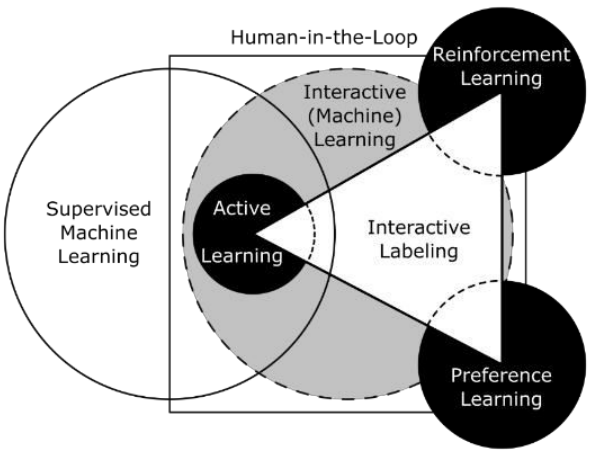


**Figure 1.** Conceptual Foundations (based on [13])

https://doi.org/10.30844/wi_2020_a1-knaeble

In SL, the learner is provided with feature vectors with labels and the task to predict this label on unseen vectors [13]. AL is a subdomain of SL that focuses on unlabeled data and querying for their labels [15]. IML builds on this concept, however learner-driven selection is combined with user-driven input [9]. RL learners do not have access to labeled training data but instead get feedback about their choices from a reward, which may be triggered by a human, a machine, or even another ML algorithm [13]. PL includes user preferences towards errors, with the aim to "refine the decision boundaries" [7].

## 3 Research Method

We followed the established guidelines for SLRs [10, 11] along three distinct stages: *plan*, *conduct* and *report*. In the *plan* stage, we identified the need for an SLR and defined a research protocol. During the *conduct* stage, we followed through with our predefined protocol, applied our search strategy, and analyzed the results. Lastly, we *reported* our findings. To address the overall RQ, we formulated the following sub-RQs: (1) *What is the role of humans in IL?* (2) *What are predominant research areas in IL?* (3) *What are potential directions for future research?*

**Table 1.** Luminary Articles and their Coverage across the Databases

| *Reference* | *Cited by* | *A* | *B* | *C* | *D* | *E-J* |
|---|---|---|---|---|---|---|
| Tian et al. (2007) [17] | 88 | | | ● | ● | |
| Culotta & McCallum (2005) [18] | 211 | | | ● | | |
| Joshi et al. (2012) [19] | 99 | | ● | ● | ● | |
| Branson et al. (2011) [20] | 106 | | | ● | ● | |
| Amershi et al. (2014) [7] | 226 | | | ● | | |
| Fails & Olsen (2003) [16] | 323 | ● | | | | |

**Search strategy.** Based on the pre-eminent work on IML [7, 16], we performed a forward/backward search to gain an initial overview about IL research. We identified four additional relevant and highly cited articles (Table 1) [17–20]. We noticed that scholars used two terms synonymously: interactive annotation and IL. On this basis, we created the following search term (Part I): ('interactive machine learning' AND ('annot*' OR 'label*')) OR 'interactive annot*' OR 'interactive label*'. Part II was focused on retrieving articles that relied on the underlying learning approaches, that is AL, RL, and PL (Section 2). As these terms do not necessarily relate to IL, we filtered with terms emphasizing the user role: ('annot*' OR 'label*') AND 'data*' AND 'interact*' AND 'user*' AND ('active learning' OR 'preference learning' OR 'reinforcement learning'). Lastly, both parts were joined with an OR (I OR II). Hereby, we selected the following 10 databases (DB): ACM DL (A), Web of Science (B), Scopus (C), IEEE Xplore (D), Science Direct (E), Emerald (F), JSTOR (G), Wiley Online (H), ProQuest (I) and AISeLibrary (J). These DBs are established in research and cover the related areas of IS, HCI, and CS. We checked which of our six luminary papers were covered in the hits of

https://doi.org/10.30844/wi_2020_a1-knaeble

each DB (Table 1). We selected those four DBs that returned hits on any of our luminary papers (A, B, C, D).

**Selection criteria.** (1) Only peer-reviewed articles were included. (2) Articles without a focus on supplying training data interactively were excluded. (3) Only articles in the English language are included. Our search returned 549 articles, spread across the DBs. After eliminating duplicates, we were left with 426. Next, we employed the above-mentioned selection criteria to title, abstract, and keywords; hereby 135 articles were left. Following the same criteria for a full text review, 52 articles remained. Lastly, we employed a forward/backward search and included 5 more articles (57 in total).

**Distribution of articles.** Although no time boundary was applied for the SLR, no article was older than 2003. In fact, more than half of the results were from 2016 or newer and more than another third was published between 2010 to 2016.

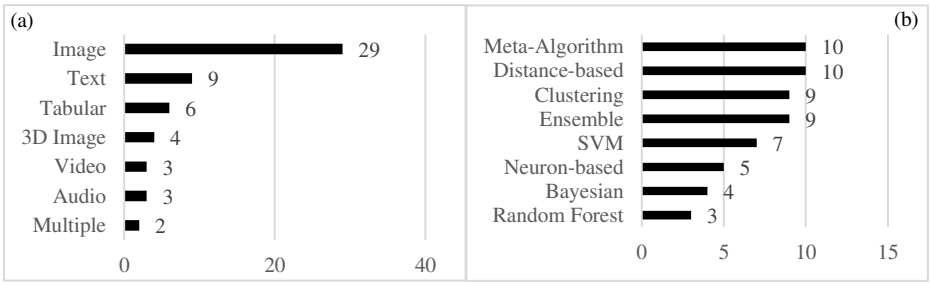


**Figure 2.** Articles per (a) data type, and (b) ML type

This supports our notion of a young and rapidly growing research field. 39 articles were derived from conferences; the remaining 18 were published in journals. For our descriptive overview of the related work, we applied inductive reasoning to analyze the types of ML and data used in the articles (Figure 2) from the bottom up. Inductive reasoning lets "salient concepts arise from the literature" [22, p. 2], which allows for our SLR to be open for whatever the investigated literature entails. The most prominent data type referred to images (29), followed by text (9), tables (6), and 3D images (4). Although IML comprises AL, PL, and RL, we did not find any application of PL for the IL field in our corpus. Just one article applied RL [21] and 30 applied AL. Regarding ML algorithms, distance-based methods (10), such as kNN, were used frequently, as well as meta-algorithms (10), followed by clustering approaches (9) and ensembles (9). Support Vector Machines (SVM) were another frequent contender, with 7 mentions.

**Classification process.** Following Wolfswinkel [22], we used a three-step strategy to develop a concept matrix and address the first sub-RQ. As a first step, we relied on open coding to create a set of concepts and insights based on excerpts supported across the 57 articles. During the analysis, it became apparent that IL initiatives include four underlying categories: (1) *select*, (2) *present*, (3) *act*, and (4) *evaluate*. *Select* refers to the process of how much data the user must label, that is, the complete data set, a subset (pre-defined by either the user or the system), or hybrid forms. All solutions related to the ordering (one-by-one, multiple at once, or various forms) and visualization (item-focused or inter-item comparisons) of the labels belong to the *present* category. *Act* incorporates the degree of user interactivity for collecting labeled data, such as

https://doi.org/10.30844/wi_2020_a1-knaeble

approval/rejections, labeling instances, or having the opportunity to offer deeper explanations. Lastly, the IL approach needs to be *evaluated*. As typical forms refer to model feedback to the user and reports of effort improvements, we created one subcategory for each alternative. In addition, we incorporated a subcategory for whether the user was treated as an oracle, as recent research has shown that this may lead to negative outcomes [7]. During the second step (axial coding), we developed hierarchical relationships between the (sub) categories. For the selection category, it became apparent that a flexible number of instances could be labeled. Within the action category, a subcategory for system-supplied data was introduced referring to either a label suggestion or providing a structure of the underlying dataset. For conceptualizing the user supplied labels, we used three levels: approval, label and explanations. The lower levels are implicitly contained in the uppers. In the last step (selective coding), we refined our codes for clarity. Table 2 (Appendix) shows the concept matrix. Similar to the procedure for the types of ML and data used, we applied inductive reasoning to track additional meta-characteristics. We tracked the context, user type, and, whether the users were simulated during the evaluation (i.e., the queries are not answered by an actual user but are rather sent to a database containing the ground truth dataset).

**Clustering.** Using a two-step cluster analysis method [12] already established in literature [23, 24], we identified clusters of research based on our concept matrix. We used the concept matrix as data to find homogenous groups across the categories. Next, the algorithm scanned the entire data to form pre-clusters. This was performed by using sequential clustering with log-likelihood distances (as advisable for binary data). The resulting cluster feature tree of the sequential clustering was used as an input for agglomerative hierarchical clustering. With Akaike´s information criterion the appropriate number of clusters can be determined – in our case four.

## 4 Results

In the following, we discuss the clusters identified to address the second sub-RQ. The four clusters we identified directly relate to the eponymous question of teacher or oracle, with each role being fulfilled by two clusters. We present the clusters in order of interactivity, from low to high. In general, we see the following meta characteristics across all clusters. The four types of users referred to across the 57 articles are domain experts (51%), ML experts (11%), crowd workers (3%) and not further specified "users" (35%). This shows how a focus of IL lies on domain experts as labelers. Further, we analyzed the contexts in which the systems were analyzed. Systems often were applied to or intended for multiple contexts. General machine vision, segmentation, and object detection makes up for more than half of the articles. The medical field was the most frequent specific context, with 15 articles. Thus, fields with highly expensive and scarce domain experts seem to be of special interest for IL.

**Cluster 1 – typical oracles.** In general, this cluster comes closest to the classical AL paradigm where a system-selected sampling is used. We found this cluster to address the requirement of accuracy (R2) by limiting the user control over the IL system. For instance, by relying on the IL system to select the data to be labeled, biases (but also

https://doi.org/10.30844/wi_2020_a1-knaeble

benefits) from user selection could be prevented. Furthermore, in this cluster, the labeling task is rather simple and clearly defined as users typically only need to approve or reject a suggested label. Although a certain degree of bias can never be ruled out, such course of action can minimize the risk of systematic error caused by user misunderstanding. Regarding the meta-characteristics, for this cluster an accumulation of users being simulated in the evaluation can be observed, due to most articles in this cluster relating to AL.

All but one article already suggest a potential label for the current instance to the labeler. Still, deviations are noticeable. Seven contributions in the cluster have the user select a subset to be labeled, use all data, or employ a hybrid approach. Hybrid approaches [25], combine user- and system-selected sampling. In particular, the user may provide self-chosen examples to augment the system-selected ones or be given the opportunity to correct both suggested labels by the system and introduce new labels or examples for existing labels. For example, [25] work within the medical context to detect objects within microscopy images from pathological tissue samples. With their visualization approach, the authors enable the user to pick samples to label from a selection initially provided by the system. This hybrid approach, in which the system gives the user feedback on which samples it is uncertain, and the user selects those where they find the most learning potential, enabled the authors to accurately classify highly complex microanatomic structures. However, most of the approaches have the system provide labels for the users, who in turn often must approve or reject, and partly relabel. Still, this cluster shows promise by applying IL to a broad spectrum of contexts. Thus, it falls within the section of oracles. We coin it the *typical oracles*, as it is closest to AL, where the concept of having an oracle originated from. It ranks lowest on interactivity.

**Cluster 2 – oracles by necessity.** In this cluster, all the data needs to be labeled by the user. This is necessary mainly due to the importance of the context, such as the medical field (7 out of 11 articles refer to the medical field), where the given labels are complex and require a high level of domain expertise. To this regard, users would feel more accountable (R3) and try to produce more accurate (R2) results as the IL system largely relies on their domain expertise.

Herby, most of the articles in this cluster provide item-focused visualizations to further support the user in the underlying labeling task, by, for instance, coloring image segments [16] or highlighting words of semantic interest [26]. Another example of item focused visualization can be seen in [21], where the authors reduce the effort in creating polygonal image bounding boxes. While this cluster employs visualization and enables the users to input labels instead of approvals/rejections, it continues treating the user as an oracle. In all cases the system already provides a suggestion for the label. Many approaches rely on the AL paradigm of a steady stream of simple questions to speed up the process. In sum we coin Cluster 2 the *oracles by necessity*, as the users have to label all the data.

**Cluster 3 – fast teachers**. Whereas Cluster 2 shows all the data instead of a subset and shows visualizations of the items, the major pattern within this Cluster is as follows: present the user with a system-selected subset, show them multiple items at once, and structural information about the data determined by the system. Thus, this cluster

https://doi.org/10.30844/wi_2020_a1-knaeble

focuses on speed (R4). By combining multiple label steps into one, speed improvements were illustrated by research [28]. Often holding accuracy (R2) at a status quo is a prerequisite for this, as without it, the speed-up would be meaningless. Aspects like accountability (R3) and motivation (R1) are not particularly addressed in this cluster, although we see that the latter may be indirectly influenced by improving the speed itself.

All articles of the "fast teachers" provide structural information and all but one [27] provide a multiple item at once ordering. The articles in cluster 3 have the user now provide actual labels instead of a simple approval/rejection of what is presented. This seems a promising approach for improvements in labeling effort (8 out of 10 papers). For instance, [28] focus on labeling EEG data of intensive care patients, who had their brain activity recorded continuously. Experts, doctors in this case, needed more than three hours to manually review each 24-hour interval of data – an amount of work simply not feasible. To extenuate these negative effects, an IL approach was applied which cut the labeling effort down to about three minutes. Doctors were supplied with information about the structure of the data. Hereby, a bag-of-words based clustering was used to group similar data enabling the doctors to label many hours of data at once.

In this cluster, the user is treated as a teacher. While the data to be looked at is sampled by the algorithm, the structural information and the multiple-at-once approach enables the user to not answer a steady stream of simple questions. This combination of efficient sampling and leveraging the user's knowledge is a driver for effort reductions. For this cluster we coin the term *fast teachers,* as their focus is on speed. Still, they manage to not treat the users as an oracle, by offering more information and multiple-at-once labeling.

**Cluster 4 – explaining & comparing teachers.** Here, all systems accept explanations from the users to further improve their performance and use of the cognitive capabilities of the human labeler, which is able to grasp complex patterns with little information. By allowing for explanations as user input, the degree of interactivity by the IL system is increased, which promotes the user's motivation (R1) in the end [7]. However, we see a trade-off looming as speed (R4) is generally negatively influenced by additional inputs.

A prominent example of providing explanations refers to labeling objects in LiDAR 3D point clouds with clustering and similarity matching [29]. The user is provided with a group of similar looking objects and a suggested label for it (e.g., 30 instances of fire hydrants along a street). Next, the user may choose along the following actions: (1) approve the system's suggestion, (2) provide a different label, that is indicating that the grouping was right, but the label was wrong (e.g., the "fire hydrants" are actually lampposts), or (3) ask the system to contract or expand the group (e.g., if instances were missed or included wrongly). Such course of action cuts the time needed for labeling the data set into half. Through enabling the highest level of user input (explanations) to leverage the user's capabilities and supplying feedback on the underlying model, this cluster treats the user as a teacher. Also, inter-item comparison visualization is used in almost all articles in this cluster. This further leverages the user´s capabilities, allowing for better label input and forming the basis to provide such explanations. For this highest level of interactivity, we use the term *explaining & comparing teachers*, as here

https://doi.org/10.30844/wi_2020_a1-knaeble

actual explanations are passed from the user to the system, as well as visual aids for comparison of items are provided.

## 5 Future Research Directions

To address the third sub-RQ, we present future research directions on IL. In particular, relying on the derived concept matrix and descriptive analysis, five research gaps within the existing body or research of IL could be identified. Our concept matrix represents a valuable baseline for further investigations as it provides a systematic overview, thus indicating what has not yet been researched, facilitating avenues for future work.

**(1) Understand the Influence of Gamification on IL.** Classical labeling approaches already employed gamification successfully. For instance, in one gamification article [30] the authors created an online game people would use voluntarily, labeling images as a side product. In this game, users are matched up with a partner and describe images with words not in a taboo list. They receive points if both players mention the same word. This is then added to the taboo list for the next pair and as a label. However, none of the 57 articles reviewed employed gamification. Design science research projects might help to address this issue. For example, applying gamification to an existing IL tool and evaluating the effects along different context and tasks could represent a promising starting point. This could be done in two different ways: creating an actual game, like [30], which is played for sake of entertainment. Or introducing gamification elements like scores and achievements into a labeling environment, where the labeler is still working (in contrast to playing a game) [31]. In both cases the requirement of user motivation (R1) stands to benefit.

**(2) Study the Effectiveness of Evaluation Feedback and Visualization Forms.** Evaluation feedback on the ML model, or the labeled data, is important for the users [7] as they strive to know what they are working for. In particular, [32] apply visualization with colors and opacity to show where in the feature space classes are distributed and where there is a potential lack of labeled data. Visualization is beneficial here, as it provides the user with easy to grasp information. This refers to two IL requirements. Firstly, the user's accountability (R3) could be improved by a more transparent system, allowing for more control while maintaining a clear goal for the user. Secondly, the motivation (R1) may also be positively affected, especially for progress and ML model feedback. However, less than a fifth of the articles reviewed offered evaluation feedback on the ML model during the IL process and only three reported a visualization of such feedback. Hence, future work should investigate deeper which types of model evaluation feedback and which forms of visualization help to improve the effectiveness of the IL process. Context-independent visualizations, such as the 2D plots in [32], could be used. A laboratory study could investigate these effects. Hereby, the users' proficiency with ML, their domain knowledge, as well as their motivation to label could be considered as moderating factors.

**(3) Understand the Influences of User-Selected and Hybrid Subsets for IL.** User-selected subsets may significantly impact user accountability (R3) and motivation (R1) positively. However, so far, there exists a scarcity of work on having the user select the

https://doi.org/10.30844/wi_2020_a1-knaeble