KI - Künstliche Intelligenz (2020) 34:131–142
https://doi.org/10.1007/s13218-020-00634-1

TECHNICAL CONTRIBUTION

Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning

Mario Nadj[1] · Merlin Knaeble[1] · Maximilian Xiling Li[1] · Alexander Maedche[1]

Received: 16 September 2019 / Accepted: 2 January 2020 / Published online: 11 January 2020

Abstract
Labeling is the process of enclosing information to some object. In machine learning it is required as ground truth to leverage the potential of supervised techniques. A key challenge in labeling is that users are not necessarily eager to behave as simple oracles, that is, repeatedly answering questions whether a label is right or wrong. In this respect, scholars acknowledge designing interactivity in labeling systems as a promising area for further improvements. In recent years, a considerable number of articles focusing on interactive labeling systems have been published. However, there is a lack of consolidated principles how to design these systems. In this article, we identify and discuss five design principles for interactive labeling systems based on a literature review and offer a frame for detecting common ground in the implementation of corresponding solutions. With these guidelines, we strive to contribute design knowledge for the increasingly important class of interactive labeling systems.

Keywords Interactive labeling · Interactive machine learning · Training data

1 Introduction

Machine learning (ML) has become one of the most rapidly growing areas in computer science [1] generating massive attention in both academic and business communities [2]. Recent market studies illustrate that the growing dissemination of ML-based systems strongly impact both the societal and business context [3]. In the ML community, specifically supervised ML (SML) techniques are well-established and widely applied in diverse contexts, such as speech recognition, natural language processing (NLP) or computer vision [4, 5]. Typically, they require large quantities of labeled training data as a ground truth to learn successfully [6]. Hereby, labeling refers to the process of enclosing information to some object [6].

Creating labeled data is often a costly, error-prone and labor-intensive activity that might even frustrate users [6, 7]. In particular, studies illustrate that users are not necessarily eager to behave as simple oracles, that is, repeatedly answering questions whether a label is right or wrong [8]. For instance, Cakmak et al. [9] show that a steady stream of questions to users when teaching a task to a robot is assessed as instable and annoying. To alleviate these problems, research suggests to account for human factors, in particular, by designing interactivity in labeling systems [8]. Such interactive approaches for collecting labeled data are manifold ranging from simple approvals or rejections [e.g., 9] over label corrections or new label assignments [e.g., 10] to deeper explanations users may want to offer to learners [e.g., 8]. So far, a considerable number of articles focusing on interactive labeling systems has been published in the field of ML, despite its relatively young nature. Although some scholars reviewed and summarized common challenges for improving approaches for interactive ML (IML) [e.g., 8, 10], there is a lack of agreed upon design principles for interactive labeling systems.

However, appropriate designs of interactive labeling solutions are crucial to the success of such systems. In an ideal implementation, users would feel more important and engaged as they could build their own learned concepts by generating or collecting training data in congruence with their need [10]. In general, we believe that interactive labeling systems are in particular valuable for contexts where the

✉ Mario Nadj
mario.nadj@kit.edu

Merlin Knaeble
merlin.knaeble@kit.edu

[1] Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

given labels are complex and require a high level of domain expertise. Hereby, we see a trade-off looming. As an example one can imagine the medical field in which doctors have to classify pathological images [11]. On the one hand, these users are more involved in the whole labeling process and perceived as relevant through interactive labeling systems, as they can provide more input than just simple approvals or rejections (i.e., disease present or not). A higher perceived user importance and engagement could therefore promote not only more accurate labeling results, which would ultimately increase the performance of the resulting ML models, but also user trust and acceptance of the system. On the other hand, the higher user engagement has a cost. The demand for deeper explanations next to the label means that the user has to spend more time on the labeling process. Although such information could be used to optimize selection mechanisms when it comes to asking users for labels and reducing the total number of labels required in the long term, it seems essential to compare the effort of the labeling process with the benefits when designing interactive labeling systems.

On these grounds, this article illustrates a review and characterization of labeling principles from an interaction perspective based on a literature review to consolidate observations from user studies and establish common ground. The described principles may serve as guidance to designers in charge of developing effective interactive labeling solutions. Following the introduction, we describe the foundations of IML in Sect. 2, before related domains are illustrated in Sect. 3. In Sect. 4, we discuss the underlying research method of the literature review. Section 5 introduces design principles for building effective interactive labeling systems. Lastly, we conclude the article in Sect. 6.

2 Foundations of Interactive Machine Learning

IML represents "an interaction paradigm in which a user or user group iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review" [10, p. 4].

Refinements of the model are typically quick as they are immediately executed after the user — who typically is a domain expert for the problem at hand — has provided their input [8]. In addition, such model refinements concentrate on specific aspects and are rather incremental as the model does not change radically from the preceding iteration enabling the user to directly review the impact of their inputs while adjusting succeeding actions to achieve their goals. Hereby, practitioners with high expertise in ML—in the following, we call these experts "ML practitioners"—are in charge of developing the interface for the users making them capable of building and refining the model [10]. On these grounds, even users with little ML, however a high degree of domain expertise, can experiment in a "trial-and-error" manner to guide the behavior of the ML model [8]. Imagine the following scenario from biomedicine: Yimam et al. [12] have built an IML system to annotate entities in complex biomedical texts. Hereby, medical users with domain expertise are able to use the IML system in order to recognize new entities and their relationships. Manual corrections from the medical users are leveraged to immediately and incrementally improve the underlying ML model or train new models for unseen texts. In addition, the medical users can directly observe how their adjustments change the behavior of the ML model.

This is in contrast to traditional ML workflows as they typically involve lengthy and difficult iterations operated by a ML practitioner. Initially, data is offered by users, before ML practitioners cooperate with them to derive features

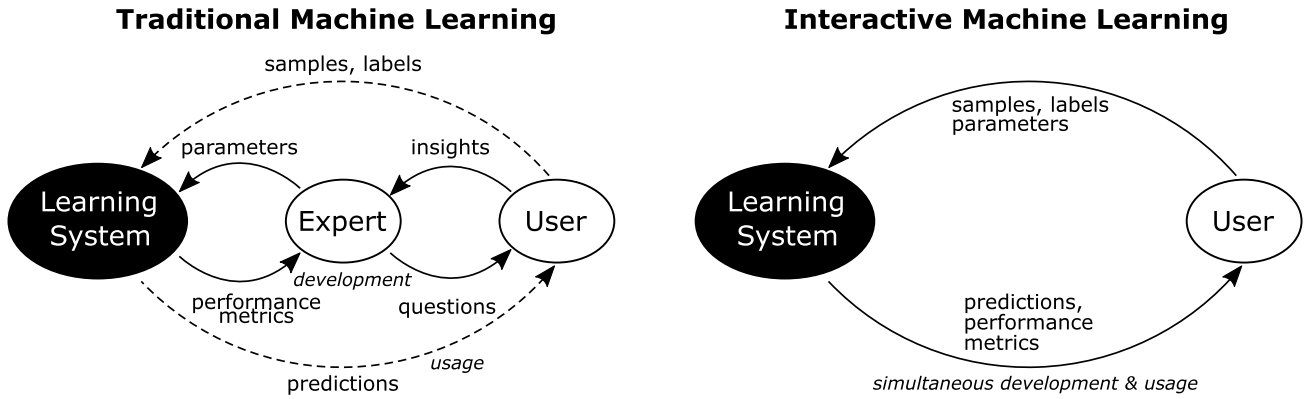

Fig. 1 Process of developing traditional machine learning systems versus interactive machine learning systems; based on Amershi et al. [8]; the "Expert" is represented by the machine learning practitioner, whereas the "User" refers to the domain expert

from the data. In the next iterations, ML practitioners apply various algorithms, adapt parameters and features to increase the performance of the ML model. On these grounds, the model updates are diverse and commonly different from the preceding iteration [8]. In such iterations, users do not directly guide the behavior of the model nor see the impact of their input on the model [13].

A detailed discussion on the differences between traditional ML and IML processes, as well as case studies that emphasize the value of IML can be found in Amershi et al. [8]. In summary, the IML workflow is user-driven to accomplish the desired behavior. However, the system still has impact on the workflow and can make decisions autonomously, for instance, intelligently defining a data subset for labeling. In addition, ML practitioners can also influence model adjustments by preprocessing the user input before it is internalized by the ML model. The underlying rationale is to give the user control over the high-level system behavior that may not necessarily be executed in each interaction [10]. Fig. 1 visualizes the process of developing traditional ML systems versus IML systems.

3 Related Domains

Interactive labeling is a subdomain of IML that is part of the Human-in-the-Loop (HITL) methodology. This methodology aims to reduce the limits of fully automated systems through user interaction [14] and relate to four underlying learning approaches that are introduced in the following.

3.1 Supervised Machine Learning (SML)

SML is one of the most common learning approaches that is supplied with training data consisting of several sets of feature vectors, each with a desired label [15]. On this basis, a model is learned that predicts the correct label on previously unseen feature vectors. For instance, imagine the problem of predicting car insurance claims. The input feature vectors could include information in form of age, gender and home address of the driver as well as the car type. Experienced ML practitioners would work with users (in this case insurance professionals) to derive these sets of feature vectors. The labels could be provided by past records of similar insurance claims. In addition, the users could also correct or assign new labels [10], however their possibility to directly guide the behavior of the prediction model would be rather limited. Ultimately, a prediction model for potential claims could be developed, which in turn would help adjust the monthly insurance premiums.

Within the IML domain, the degree of user interactivity varies along context, system and task. Particularly, users would exert greater control over the ML-based system than just correcting or assigning new labels [10]. For instance, they could provide deeper explanations incorporating the proposal of alternative features, changing the weight defined to features or altering the information extracted from the text in order to fit their defined goals and needs [8]. In the case of our car insurance example, insurance professionals could be asked to assign importance values to the individual input fields when recording new claims. These user inputs would have a direct impact on future training iterations and would (potentially) further improve the performance of the prediction model.

3.2 Active Learning (AL)

AL refers to a subdomain of SML where the model identifies interesting key data points and queries the user or another source of information for its label [16]. Thus, only a subset of data must be labeled. This is particularly beneficial for scenarios where the label collection is expensive, time-consuming or complex.

For instance, in astronomy, AL could be used to identify the buildup of celestial bodies. The underlying model would be trained based on a database of existing observations. Whenever an interesting example appears, for instance, a data point that could explain many similar observations, the AL system asks for a label. In such cases, a user with domain expertise could be commissioned to provide the needed labels.

IML can build on the AL concept, however learner-driven point selection strategies are complemented with user-driven input [10]. In particular, domain experts could identify the instances within the data set for which the system should query labels. But even they may not be knowledgable enough to assess all possible data points in the set, suggesting a hybrid approach of user- and system-selected subsets. In particular, an IML system used for the astronomical task could allow domain experts to mark certain features (such as the size or color of the stars) or individually point out interesting examples. With such course of action, users would be more involved in the training process and a higher degree of their domain-specific knowledge would be transferred into the ML model.

3.3 Reinforcement Learning (RL)

In turn, RL does not have access to labeled training data but instead gets feedback about its choices and actions from a (time-delayed) reward or punishment. Such input may be trigged by a human judge, a machine or even another ML algorithm [15].

Imagine a robotic agent trained to fly a quadrocoptic drone in a stunt competition. The agent gets a positive reward score in proportion to the difficulty of a successfully completed stunt

maneuver. Punishment is applied in the event of rule violations, for example as soon as the drone crashes.

When applying RL in the IML domain, a human judge is in charge to reward or punish the action in an interactive and natural fashion. For instance, in the context of the stunt drone, rewards may be provided by hand clapping after the agent performed an action to the human judge's satisfaction instead of explicitly entering a numerical score to conform to the technical constraints of the system.

3.4 Preference Learning (PL)

Lastly, PL learns from observations that either explicitly or implicitly illustrate the preferences of a user or class of users [17]. For instance, recommendation systems create the users' preference model iteratively by evaluating their actions [10]. However, oftentimes users are unconscious that the underlying algorithm is learning implicitly from their actions. Conversely, other recommender systems make users explicitly aware that they can affect their preference model [18]. For instance, movie streaming services might use an explicit user rating to improve their recommendations or even production strategies by correlating the ratings to actors, directors, genre or even visual features of the video. Furthermore, implicit statistics can be inferred by recording the patterns when users abandon a series or the order in which movies are consumed (e.g., an action movie is followed by a romantic comedy).

When augmenting PL with IML, most of the changes occur in the explicit inputs. Implicit inputs are hidden from the user and only work if they remain so. But the explicit ratings by users can be improved by providing more detailed rating responses as well as by illustrating their impact. The rating options should not increase in quantity (i.e., the range of scores) but rather in quality. In the case of movie recommendations, the user could individually rate different aspects such as musical scores, cast or pacing rather than simply providing a binary rating for the entire movie. Moreover, increasing the awareness of the users what impacts their action could have positive effects on their underlying engagement.

In summary, IML applies these approaches to advance the interaction between the user and the system by emphasizing the co-adaptive learning process and role of the user (cf. Fig. 2). Interactive labeling falls within this field and refers to the interactive elicitation of labeled training data, which is a prerequisite for many applications involving ML.

4 Research Method

To offer a characterization of design principles for interactive labeling systems, we consolidated observations from user studies in this domain. Hereby, we conducted a literature review by following the methodological guidelines by Webster and Watson [19].

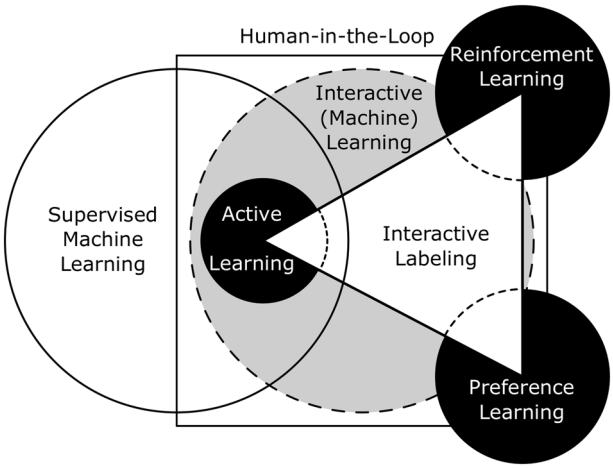

Fig. 2 Related domains based on Trivedi [15]

We queried three well-established databases, namely ACM, Web of Science (WoS) and IEEE, to identify relevant user studies on interactive labeling systems. Our search string consists of two parts. Besides the general concept of IML, we observed that researchers specifically rely on two different terms for the process of training data interactively, namely interactive annotation and interactive labeling. On these grounds, the first part of the search term was created (Part I):

"interactive machine learning" OR
"interactive annot*" OR
"interactive label*"

The second part of the search string relies on established terms from related domains that have been already introduced in Sect. 3 (i.e., AL, PL, RL, HITL). However, as these terms do not necessarily have to be associated with interactive labeling, we connected them with terms, such as "label", "annotation", "data", "user" and "interact" to ensure that only articles were in the corpus which are relevant for the aim of our literature review. In particular, only articles that emphasized the user role, its interaction, the labeling (annotation) process and the context of (training) data were taken into consideration for further analysis (Part II):

("active learning" OR
"preference learning" OR
"reinforcement learning" OR
"human-in-the-loop") AND
"user*" AND "interact*" AND
("annot*" OR "label*") AND
("training data" OR "data*")

Finally, an OR operator combined both parts (Part I OR Part II) for the final search string. Our query returned 1599 peer-reviewed publications for ACM, 110 for WoS and

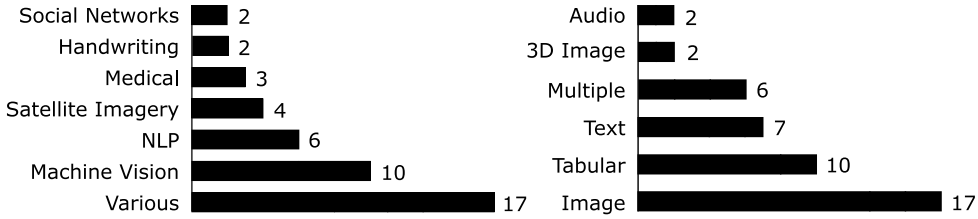

Fig. 3 Context (left) and data type (right) of articles identified in the literature review

2662 for IEEE (4371 in total). Hereby, 1312 duplicates were removed. Next, we filtered by scanning the title, abstract and keywords of each article and applying the following exclusion criteria: Articles without a focus on designing interactive labeling systems were excluded. We would only have considered articles in English. On this basis, 302 articles were left. Following the same criteria for a full text review, 25 relevant articles remained. Finally, we employed a forward and backward search and included eight more articles to our corpus leading to a sample of 33 relevant articles. Furthermore, due to our general research focus on ML, we included two flagship conferences in the area of ML, namely the "Conference on Neural Information Processing Systems" (NIPS[1]) and "International Conference on Machine Learning" (ICML[2]). Lastly, we also covered two ICML workshops (i.e., Human In the Loop Learning—HILL and Human Interpretability in Machine Learning—WHI) as both are of specific relevance for IML. We applied the same exclusion criteria as above and were able to identify another eleven articles for a grand total of 44 relevant articles for our review.

The identified articles are spread across different contexts. Ten articles concentrate on machine vision. Six articles address NLP and four satellite imagery, whereas two articles each deal with handwriting and social networks. Three articles investigate the medical field. The remaining 17 articles study diverse contexts. Examples refer to event detection for audio surveillance [14], movie recommender systems [21] and teaching robots task orders and concepts [22]. Regarding the data types, most articles (17) deal with images, followed by tabular data (10; mostly numerical or Boolean typed), text (7), combinations of multiple data types (6), as well as 3D images (2) and audio (2). Fig. 3 summarizes these results.

For the derivation of the design principles, we relied on a qualitative research strategy as described by Zikmund et al. [23]. A qualitative research strategy seems specifically supportive for generating new concepts, such as design principles, and when the aim is to create an understanding of some phenomenon in much depth and in great detail [23]. The role of theory was inductive and we applied three steps to identify our design principles. First, we documented the aim, method, context, data type and findings for each article. Second, based on this documentation, we have relied on open coding to create a set of concepts that relate to excerpts supported in the 44 articles. During the analysis, it became apparent that the described interactive labeling approaches are grounded on five underlying core themes: (1) user intent, (2) user engagement, (3) perceived relevance of the user, (4) interruptability and (5) transparency. Third, we typecasted all articles along the themes and derived five design principles for interactive labeling systems.

5 Design Principles for Interactive Labeling Systems

In the following, we describe the design principles, which are summarized in Table 1, along prominent publication results and exemplary interactive labeling approaches to offer a useful frame for detecting common ground for implementing interactive labeling solutions.

5.1 Embrace the Intent of Users

Scholars have shown that there is an ambiguous relationship between user input, which can be captured but is prone to error, and the user intent, which is typically hidden [10]. For example, if users can only provide input in form of yes or no answers, they might be too restricted by the functionality of the labeling system to express their intent. Thus, minimizing this ambiguity is of specific relevance to the training process of a model. However, algorithms face the problem of distinguishing between

Table 1 Design principles for interactive labeling systems

Design principle	References
Embrace the intent of users	[10, 20, 22, 24–37]
Support the engagement of users	[8, 9, 38–42]
Increase the perceived relevance of users	[10, 11, 14, 21, 43–46]
Assess the degree of interruptability	[6–10, 46–52]
Adjust the transparency to users	[8, 12, 14, 20, 53–56]

[1] We queried the NIPS back to the year of 2003 as this coincides with our oldest previously identified publication [cf. 20]

[2] The regular proceedings of ICML were already covered by our database search up until 2009. From 2010 onwards we relied on icml.cc and proceedings.mlr.press to retrieve the corresponding articles.

error-prone user input and examples that are considered important by users [10, 24].

Fogarty et al. [24] tries to circumvent this problem by showing only the best and worst results directing the user towards either "good" or "bad" training examples. Another technique is inspired by domains where comparisons are easier to collect than precise assessments [25]. In this regard, Xu et al. [25] propose a general algorithmic learning framework based on both: labels for comparisons and assessments. Imagine a clinical setting where a "precise assessment of each individual patient's health status can be difficult, expensive and/or risky (e.g. it may require application of invasive sensors or diagnostic surgeries), but comparing relative statuses of two patients at a time may be relatively easy and accurate" [25, p. 1]. Similarly, Shivaswamy and Joachims [26] enable users to offer implicitly preferences from which the system infers a ranking, instead of requiring explicitly ranked lists as training data. In turn, Borovikov et al. [27] rely on interactive user input to train virtual agents in the context of video gaming and report improved performance results compared to modeling the virtual agent behavior. In the navigation tool of Plummer et al. [28], users evaluate image attributes so that the search process of the system can be aligned with the user intent. For instance, when searching for a "sandal", users may be presented with images of a flip flop shoe and more traditional sandals, accompanied by an explicit question regarding which of those images represents more likely what they are looking for. Amershi et al. [29], on the other hand, rely on both explicit and implicit user input for creating a system for custom social networks. For instance, in case of implicit user input, the contacts will be labeled as negative examples if past contacts are omitted by the user.

Another stream of literature focuses on interactive clustering to embrace the intent of users. In particular, Self et al. [30] introduce a selection tool when pulling a data point to let users think about the clustering of other data points in order to support the congruence of user intent with user input. Similarly, Dasgupta et al. [31] show that better results can be achieved if word groups are clustered interactively by users via anchor words. In turn, Cheng et al. [32] showcase that user-driven clustering is sensitive to skewness as each user typically observes only a subset of the complete dataset. The authors propose an algorithm for using feature embeddings to normalize multiple user inputs.

Furthermore, research discusses the use of RL techniques to address the intent of the user. For instance, MacGlashan et al. [33] develop a RL algorithm that is capable of handling scenarios with a human judge. In particular, this algorithm can cope with changing reward scores depending on past actions. Typically, users tend to give great rewards as soon as a mistake is not repeated for the first time, whereas traditional RL frameworks expect the reward to scale only by the action taken and not by the overall behavior [33].

According to Porter et al. [34], users are able to cope with deficient tools to creatively accomplish their goals even without clearly revealing their intent. Still their performance is negatively affected, creating a need of systems to take advantage of these abilities. For instance, an image retrieval system built by Guo et al. [35] applies NLP to refine the search results. Furthermore, Thomaz and Breazeal [22] showcase that by allowing users a higher degree of freedom in time—that is, giving input to both future and previous actions of an agent—the learning performance of the agent can be drastically increased. Similarly, Fails and Olsen [20] allow the user a higher degree of freedom in space, by using imprecise, hand drawn markings on an image to separate foreground from background. These markings do not have to be placed pixel perfect on the border between the image segments, but the system interprets the location and form of the user input. Hebbalaguppe et al. [36] illustrate that users prefer to apply such a technique with the additional input possibilities of rectangular bounding boxes to make a broad, first selection (cf. Fig. 4). However, for precise segmentation, Acuna et al. [37] rely on polygons instead of pixel-based approaches. Polygons are able to represent any shape that humans can perceive as object contours and more similar to what the user sees and intends. Their sparse representation allows for quick and easy user input while significantly reducing the amount of clicks.

In summary, to embrace the intent of the user, the designer should beware of excessively restricted capabilities and offer higher degrees of freedom in terms of user input when creating interactive labeling solutions.

5.2 Support the Engagement of Users

Although context-dependent, users oftentimes share a negative attitude towards being (mis)used as simple oracles, that is, being forced to answer repeated yes or no questions or executing ordinary labeling tasks [9]. In turn, users have the tendency to offer deeper insights [38]. Recent research confirms this perspective by showcasing that in some cases users may rather exert greater control over systems than just providing a label [8]. In particular, Bryan and Mysore [39] develop a system that visualizes an audio spectogram enabling users to highlight problematic noise artifacts via a broad paintbrush. Furthermore, Stumpf et al. [40] report that users naturally offer a wide range of input types incorporating the suggestion of weight or importance changes, feature alternatives or information altering. For instance, some interactive labeling approaches enable users to label features (e.g., choosing characteristic words from a document) and not only instances (e.g., matching documents to classes of groups) in order to support user engagement.

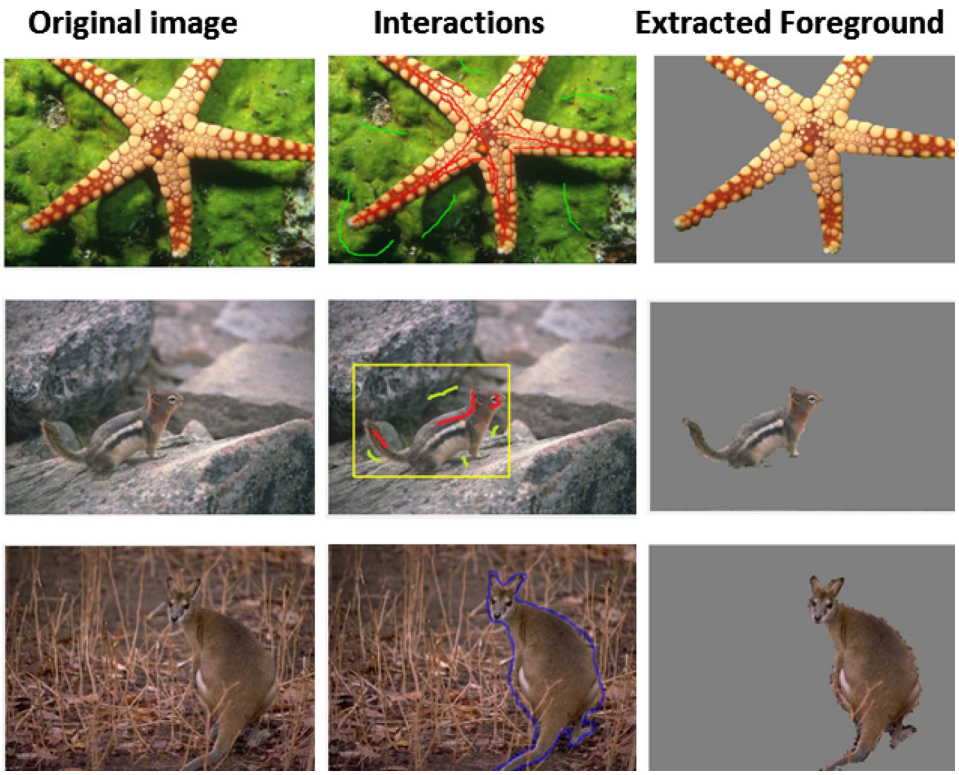

Fig. 4 Image segmentation methods evaluated by Hebbalaguppe et al. [36]

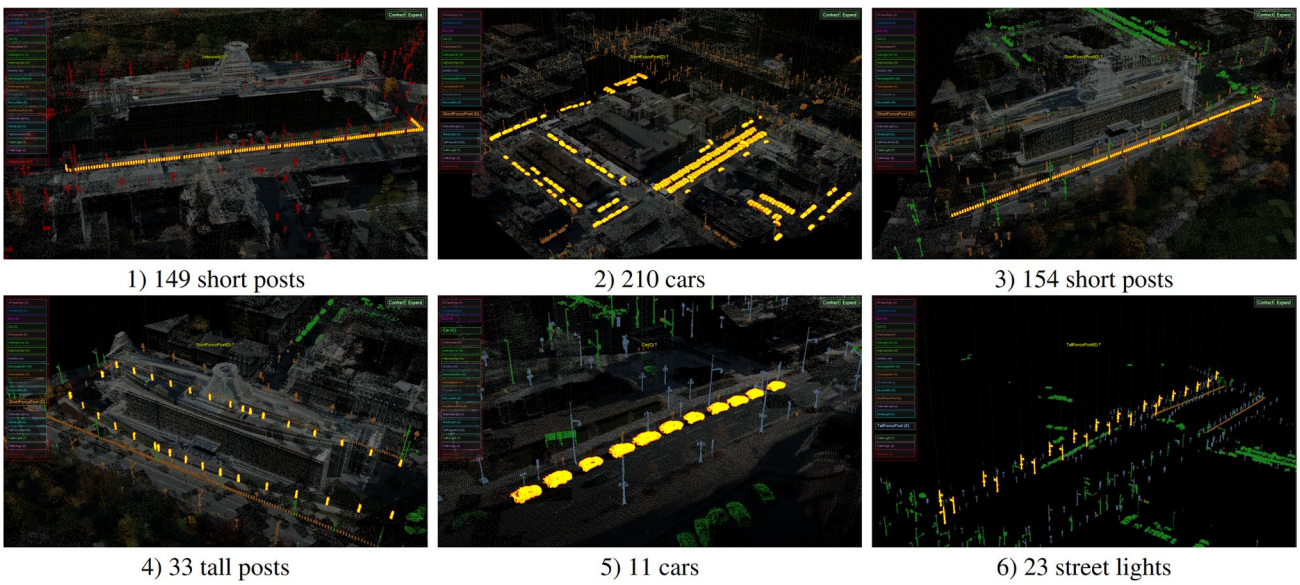

Fig. 5 Sequence of LiDAR 3D point clouds from the system built by Boyko and Funkhouser [41]; the highlighted objects are labeled by the user, with the number of objects in the class written underneath each screenshot

Fig. 5 shows several screenshots from a system built by Boyko and Funkhouser [41] for labeling 3D point clouds acquired by LiDAR scans. LiDAR represents a surveying method similar to radar. It relies on distance measurements via laser light to create digital 3D representations of the surroundings. Each measurement is a single point in 3D space, which can be clustered into groups (clouds) to indicate solid objects. In autonomous driving, point clouds from LiDAR scans are used to identify objects and their classes. Static objects, such as street lights, have to be differentiated from moving pedestrians or cars. The authors provide users a group of point clouds with a corresponding

label prediction which they can approve if appropriate [41]. But instead of only offering simple approvals or rejections, users may choose an entirely different label or ask the system to expand or contract the group if instances are missing or added incorrectly. Each possible action has a keyboard shortcut. Examples refer to a space bar for label approval or pressing a numerical key to select a different label suggestion from a numbered list.

Such mechanisms enable users to quickly label multiple objects from the same class, which reduces the time needed for labeling, while supporting their engagement. Still, users rely on bounded rationality and will contrast the effort of the labeling process with the corresponding benefits [42]. For instance, research showcases that the approval or rejection of a given label can be executed five to six times faster than selecting a label from a list [38].

In summary, supporting the user engagement may foster their willingness to spend more time in the labeling process while improving the performance of the model.

5.3 Increase the Perceived Relevance of Users

The perception by users that they offer value with their inputs, and that these inputs have an impact on the system, refers to another important design principle [10].

For instance, Sun and DeJong [43] enclose domain knowledge to a Support Vector Machine (SVM) in order to advance learning. The underlying complexity resides in the question of how domain knowledge can be converted into a format effectively accessed by the learner. Another approach by Early et al. [44] promotes the display of partial predictions facilitating users to actively advance the quality of the prediction. Commonly, users are supplied with information on the labeling coverage they have already reached [14]. Visualization plays a key part in this, as it provides the user with easy to grasp information and feedback of the current system status. For instance, research relies on opaqueness [45] and heat maps [11, 46] to show model uncertainties and class densities. Rashid et al. [21] display a short message emphasizing the importance of user input on the top of their movie recommender system MovieLens (cf. Fig. 6). In addition, a rating system of smileys illustrates the impact of the latest user rating on specific movies for different user groups. On this ground, the authors show that users are 7.4% more likely to rate movies when provided with feedback on how important their ratings are to other people who prefer to watch the same movie genre as they do [21].

In summary, designers need to illustrate the concrete impact a user's input has on the system or other users. Such mechanisms may showcase users that they represent a relevant part of the interactive labeling process.

5.4 Assess the Degree of Interruptability

User engagement and positively influencing relevance perception, however, should not come at the expense of user

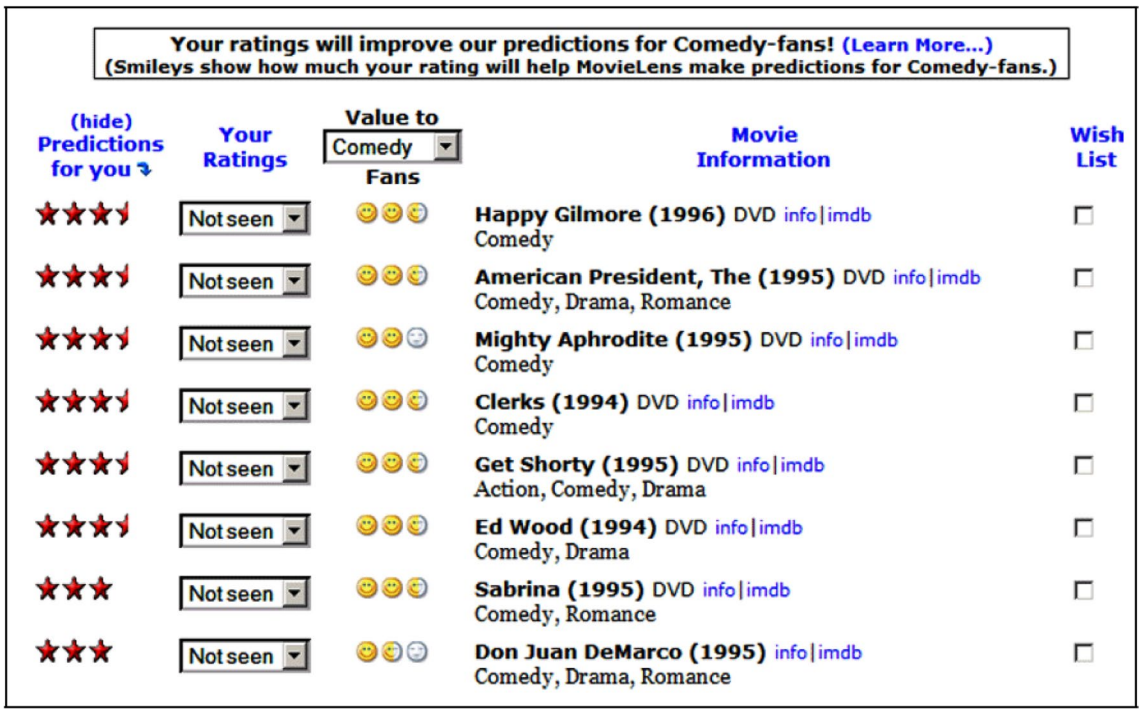

Fig. 6 MovieLens, a movie recommender system by Rashid et al. [21]

annoyance or frustration [9]. A way to address this problem resides in the careful assessment of the degree of interruptability in terms of (1) the interruption frequency and (2) the influence of individual interruptions [10].

In particular, one strategy to decrease the interruption frequency might be to offer only questions of high importance to users. For instance, Jain et al. [47] achieve accurate labels with less frequent interruptions by conglomerating together three different cues (i.e., human labels, learnt semantic similarity and geometric consistencies). In addition, based on the learner's confidence, Wallace et al. [48] assign classification labels to diverse user groups in order to take advantage of their knowledge. Hereby, the probabilistic measure of a specific instance may represent an effective indicator whether to illustrate the instance to an experienced or inexperienced user [10]. Moreover, Zhu and Yang [49] propose an algorithmic solution to assess and improve the label quality of inexperienced users by referring to "gold standard labels" provided by domain experts. Similarly, Yan et al. [50] suggest to select the best fitting crowd worker according to their expertise for the next labeling sample.

To reduce the influence of individual interruptions one may look to the selection of the subset of data that is initially labeled by the user. The AL paradigm had the system select this subset, with the user falling back to the role of a simple oracle [8]. Recent research suggests a different approach, where the user selects data points that should be labeled [7, 51, 52]. Herein the users' expertise on the current domain may be leveraged by allowing them to select prominent or difficult examples, the system alone may have overlooked. In addition, hybrid user-system approaches [6, 46] in which both the user and the system provide data points for the labeling interaction seem to represent a promising research direction. Such efforts allow for both domain knowledge and the system querying for support in low-confidence instances.

Cakmak et al. [9] conduct a study, where humans are asked to teach a robot abstract concepts. Using cutout shapes, the participants provide positive and negative examples for their chosen concept. For example, a "snowman" consists of two circles stacked on top of each other, whereas an "alien" can consist of any configuration of shapes, as long as they are green (cf. Fig. 7). Specifically, the authors compare three different modes of teaching:

- Basic strategy (BS): The robot asks a question after each training example to influence the next lesson
- Mixed interaction (MI): The robot only asks questions if the example was not informative
- Any questions (AQ): The robot only asks questions after being prompted to do so by the teacher with a key phrase

The BS strategy is found to be the most efficient, but also the least engaging one. Furthermore, participants feel annoyed and frustrated by the high amount of questions being asked [9]. AQ is described as the most natural interaction mode. It represents the preferred solution by the participants, even though it is the least efficient one. The MI mode refers to a balanced strategy. Participants feel in control during the interaction, while the learner can still ask questions (if necessary) [9].

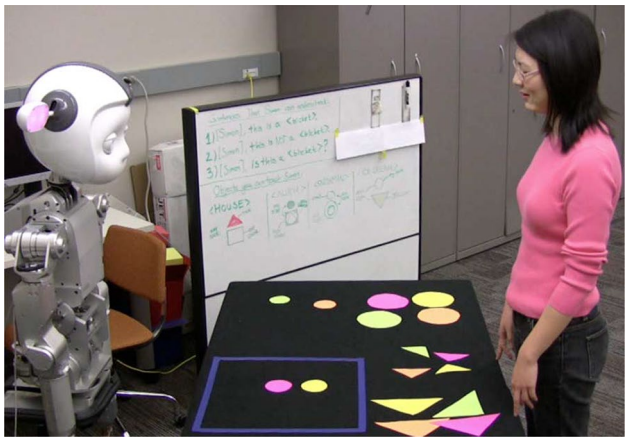

Fig. 7 Experimental setup for the study by Cakmak et al. [9]; a participant is teaching the robot a concept using the assortment of available shapes

In summary, adjustments of the interruptability degree within interactive labeling approaches can reduce the imposition on user interactivity in such cases.

5.5 Adjust the Transparency to Users

Finally, the degree of transparency within the interactive labeling approach can have a great influence on the quality of the input provided by users [8].

From time to time, users may get a label wrong and offer incorrect information. Against this backdrop, an AL approach by Rosenthal and Dey [53] suggests to offer specific information when a label is required, such as contextual features, feature explanations, prediction results or the underlying prediction uncertainty of the label [8, 54]. Such mechanisms do not only improve transparency but are also tightly interconnected with the user's perception of relevance.

Furthermore, the transparency of the labeling process can be increased by relying on information visualization techniques and visual analytics, for instance, for facilitating classification system engineering [55]. Researchers highlight areas of interest in an audio-track waveform display [14], in segmented images [20, 56] or texts [12] to support users in their understanding which parts of the data seem more relevant for the predicted labels.

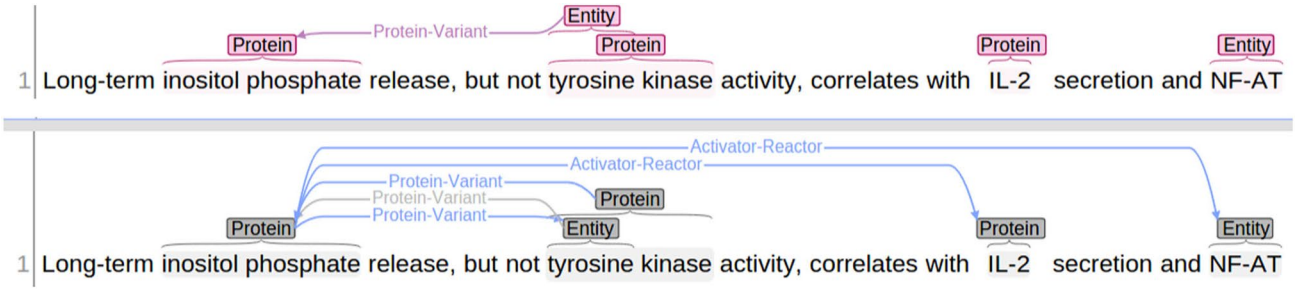

Fig. 8 Example from the system built by Yimam et al. [12]; the upper pane shows the manually annotated labels, the lower pane the suggestions by the system

Yimam et al. [12] introduce a system for automatic entity recognition in biomedical texts. In the context of biomedicine, annotations are dependent on expert knowledge, however fully manual annotations are typically regarded as labor-intesive and time-consuming. By introducing their system, the authors aim to produce high-quality labels in a short period of time with an easily extensible ML model. After medical users with domain expertise have annotated a small number of texts by hand, the system is able to provide annotations by itself for future texts. It highlights and classifies identified entities and additionally shows the relationship to other entities in the surrounding text. The user can confirm these annotations or further improve the model by providing correct annotations by themselves [12]. Such an example can be seen in Fig. 8.

In summary, research results showcase that if the system offers a satisfactory amount of contextual features and prediction results with a low degree of uncertainty, the highest labeling accuracy can be achieved. Still, not all transparency types improve the system's performance. Thus, users need to be involved in the assessment of what information seems most useful to adjust the transparency of corresponding interactive labeling approaches.

6 Summary

The exciting research area of interactive labeling offers new potentials for supporting users in labeling tasks facilitating a more effective and pleasant process. However, generating labeled data is still a challenging activity that might annoy or even frustrate users.

By concentrating more on the interactive aspects of labeling systems, this article provides a review and discussion of design principles as common ground for implementing corresponding solutions. Methodologically, we followed the structured guidelines by Webster and Watson [19] for our literature review and made all choices during our search process explicit. In particular, the article presented offers support for five design principles: The first encourages designers to embrace the user intent by providing higher degrees of freedom with regard to user input when creating interactive labeling systems, while the second emphasizes the engagement of users in the labeling process. On these grounds, users seem more willing to spend more time on the labeling tasks, which in turn improves the performance of the ML model. Third, the perception by users that they offer valuable inputs to the system or other users needs to be increased. However, designers need also to look at the expenses of interactivity to create better labeling systems. To this end, we introduce a forth design principle that addresses the degree of interruptability with regard to the interruption frequency and the influence of individual interruptions. Lastly, users need to support the assessment process which information seems most useful in order to adjust the transparency of corresponding interactive labeling approaches.

Future work should study and assess novel interaction techniques with real users to comprehend whether they promote more effective labeling solutions. We believe that the proposed design principles represent a cornerstone of further efforts to refine and expand the design of interactive labeling systems.

References

1. Chen NC, Drouhard M, Kocielnik R, Suh J, Aragon CR (2018) Using machine learning to support qualitative coding in social science. ACM Trans Interact Intell Syst 8(2):1–20. https://doi.org/10.1145/3185515
2. Watson H (2017) Preparing for the cognitive generation of decision support. MIS Q Exec 16(3):153–169. https://aisel.aisnet.org/misqe/vol16/iss3/3/
3. Baccala M, Curran C, Garrett D, Likens S, Rao A, Ruggles A, Shehab M (2018) 2018 AI predictions—8 insights to shape business strategy. https://doi.org/10.1007/s12193-015-0195-2. https://www.pwc.pl/pl/publikacje/ai-predictions-2018-report-pwc.pdf
4. Anthes G (2017) Artificial intelligence poised to ride a new wave. Commun ACM 60(7):19–21. https://doi.org/10.1145/3088342
5. Liu S, Liu X, Liu Y, Feng L, Qiao H, Zhou J, Wang Y (2018) Perceptual visual interactive learning. CoRR abs/1810.10789:1–11. arXiv:1810.10789

6. Bernard J, Hutter M, Zeppelzauer M, Fellner D, Sedlmair M (2018) Comparing visual-interactive labeling with active learning: an experimental study. IEEE Trans Vis Comput Graph 24(1):298–308. https://doi.org/10.1109/TVCG.2017.2744818
7. Zhang L, Tong Y, Ji Q (2008) Active image abeling and its application to facial action labeling. In: Forsyth D, Torr P, Zisserman A (eds) Lecture notes in computer science (including subseries Lecture notes in artificial intelligence and lecture notes in bioinformatics). vol 5303 LNCS, Springer, Berlin, Heidelberg, pp 706–719. https://doi.org/10.1007/978-3-540-88688-4_52
8. Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: the role of humans in interactive machine learning. AI Mag 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513
9. Cakmak M, Chao C, Thomaz AL (2010) Designing interactions for robot active learners. IEEE Trans Auton Ment Dev 2(2):108–118. https://doi.org/10.1109/TAMD.2010.2051030
10. Dudley JJ, Kristensson PO (2018) A review of user interface design for interactive machine learning. ACM Trans Interact Intell Syst 8(2):1–37. https://doi.org/10.1145/3185517
11. Nalisnik M, Gutman DA, Kong J, Cooper LAD (2015) An interactive learning framework for scalable classification of pathology images. In: International conference on big data, IEEE, pp 928–935. https://doi.org/10.1109/BigData.2015.7363841
12. Yimam SM, Biemann C, Majnaric L, Šabanović Š, Holzinger A (2016) An adaptive annotation approach for biomedical entity and relation recognition. Brain Inf 3(3):157–168. https://doi.org/10.1007/s40708-016-0036-4
13. Gligic L, Kormilitzin A, Goldberg P, Nevado-Holgado AJ (2019) Named entity recognition in electronic health records using transfer learning bootstrapped neural networks. CoRR abs/1901.01592:1–11, arXiv:1901.01592
14. Kim B, Pardo B (2018) A human-in-the-loop system for sound event detection and annotation. ACM Trans Interact Intell Syst 8(2):1–23. https://doi.org/10.1145/3214366
15. Trivedi G (2016) On interactive machine learning. Available at: https://www.trivedigaurav.com/blog/on-interactive-machine-learning/. Accessed 9 Jan 2020
16. Settles B (2010) Active learning literature survey. Tech. rep. University of Wisconsin, Madison
17. Fürnkranz J, Hüllermeier E (2011) Preference learning: an introduction, Springer, Berlin, pp 1–17. https://doi.org/10.1007/978-3-642-14125-6_1
18. Sen S, Vig J, Riedl J (2009) Tagommenders: connecting users to items through tags. In: Proceedings of the 18th international conference on World Wide Web, ACM, New York, NY, USA, WWW '09, pp 671–680. https://doi.org/10.1145/1526709.1526800
19. Webster J, Watson RT (2002) Analyzing the past to prepare for the future: writing a literature review. MIS Q 26(2):13–23. http://www.jstor.org/stable/4132319
20. Fails JA, Olsen DR (2003) Interactive machine learning. In: Proceedings of the international conference on Intelligent user interfaces, ACM Press, New York, NY, USA, IUI '03, pp 39–45. https://doi.org/10.1145/604045.604056
21. Rashid AM, Ling K, Tassone RD, Resnick P, Kraut R, Riedl J (2006) Motivating participation by displaying the value of contribution. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, CHI '06, pp 955–958. https://doi.org/10.1145/1124772.1124915
22. Thomaz AL, Breazeal C (2008) Teachable robots: understanding human teaching behavior to build more effective robot learners. Artif Intell 172(6–7):716–737. https://doi.org/10.1016/j.artint.2007.09.009
23. Zikmund WG (2010) Business research methods. South-Western Cengage Learning, UK
24. Fogarty J, Tan D, Kapoor A, Winder S (2008) CueFlik: interactive concept learning in image search. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, New York, NY, USA, CHI '08. pp 29–38, https://doi.org/10.1145/1357054.1357061
25. Xu Y, Zhang H, Miller K, Singh A, Dubrawski A (2017) Noise-tolerant interactive learning using pairwise comparisons. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 2431–2440. http://papers.nips.cc/paper/6837-noise-tolerant-interactive-learning-using-pairwise-comparisons.pdf
26. Shivaswamy P, Joachims T (2012) Online structured prediction via coactive learning. CoRR abs/1205.4213:1–8. arXiv:1205.4213
27. Borovikov I, Harder J, Sadovsky M, Beirami A (2019) Towards interactive training of non-player characters in video games. CoRR abs/1906.00535:1–6. arXiv:1906.00535
28. Plummer BA, Kiapour MH, Zheng S, Piramuthu R (2018) Give me a hint! navigating image databases using human-in-the-loop feedback. CoRR abs/1809.08714:1–10. arXiv:1809.08714
29. Amershi S, Fogarty J, Weld D (2012) Regroup: interactive machine learning for on-demand group creation in social networks. Proceedings of the ACM annual conference on human factors in computing systems, pp 21–30. https://doi.org/10.1145/2207676.2207680
30. Self JZ, Vinayagam RK, Fry JT, North C (2016) Bridging the gap between user intention and model parameters for human-in-the-loop data analytics. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics, ACM, New York, NY, USA, HILDA '16, pp 1–6. https://doi.org/10.1145/2939502.2939505
31. Dasgupta S, Poulis S, Tosh C (2019) Interactive topic modeling with anchor words. CoRR abs/1907.04919:1–7. arXiv:1907.04919
32. Cheng TY, Lin G, Gong X, Liu KJ, Wu SH (2016) Learning user perceived clusters with feature-level supervision. In: Lee DD, Sugiyama M, Luxburg UV, Guyon I, Garnett R (eds) Advances in neural information processing systems 29, Curran Associates, Inc., pp 532–540. http://papers.nips.cc/paper/6260-learning-user-perceived-clusters-with-feature-level-supervision.pdf
33. MacGlashan J, Ho MK, Loftin R, Peng B, Wang G, Roberts DL, Taylor ME, Littman ML (2017) Interactive learning from policy-dependent human feedback. In: Precup D, Teh YW (eds) Proceedings of the international conference on machine learning, PMLR, International Convention Centre, Sydney, Australia, Proceedings of Machine Learning Research, vol 70, pp 2285–2294. http://proceedings.mlr.press/v70/macglashan17a.html
34. Porter R, Theiler J, Hush D (2013) Interactive machine learning in data exploitation. Comput Sci Eng 15(5):12–20. https://doi.org/10.1109/MCSE.2013.74
35. Guo X, Wu H, Cheng Y, Rennie S, Tesauro G, Feris R (2018) Dialog-based interactive image retrieval. In: Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R (eds) Advances in neural information processing systems 31, Curran Associates, Inc., pp 678–688. http://papers.nips.cc/paper/7348-dialog-based-interactive-image-retrieval.pdf
36. Hebbalaguppe R, McGuinness K, Kuklyte J, Healy G, O'Connor N, Smeaton A (2013) How interaction methods affect image segmentation: User experience in the task. In: Workshop on user-centered computer vision, IEEE, pp 19–24. https://doi.org/10.1109/UCCV.2013.6530803
37. Acuna D, Ling H, Kar A, Fidler S (2018) Efficient interactive annotation of segmentation datasets with polygon-rnn++. CoRR abs/1803.09693:1–21. arXiv:1803.09693
38. Lopresti D, Nagy G (2012) Optimal data partition for semi-automated labeling. In: Proceedings of the international conference on pattern recognition, IEEE, pp 286–289

39. Bryan N, Mysore G (2013) An efficient posterior regularized latent variable model for interactive sound source separation. In: Dasgupta S, McAllester D (eds) Proceedings of the international conference on machine learning, PMLR, Atlanta, Georgia, USA, Proceedings of Machine Learning Research, vol 28, pp 208–216. http://proceedings.mlr.press/v28/bryan13.html
40. Stumpf S, Rajaram V, Li L, Burnett M, Dietterich T, Sullivan E, Drummond R, Herlocker J (2007) Toward harnessing user feedback for machine learning. In: Proceedings of the international conference on intelligent user interfaces, ACM, New York, NY, USA, IUI '07, pp 82–91. https://doi.org/10.1145/1216295.1216316
41. Boyko A, Funkhouser T (2014) Cheaper by the dozen: group annotation of 3D Data. In: Proceedings of the annual ACM symposium on user interface software and technology, ACM, New York, NY, USA, pp 33–42. https://doi.org/10.1145/2642918.2647418
42. Kim B, Glassman E, Johnson B, Shah J (2015) iBCM: Interactive Bayesian case model empowering humans via intuitive interaction. https://dspace.mit.edu/handle/1721.1/96315
43. Sun Q, DeJong G (2005) Explanation-augmented SVM. In: Proceedings of the international conference on machine learning, ACM, New York, NY, USA, ICML '05, pp 864–871. https://doi.org/10.1145/1102351.1102460
44. Early K, Fienberg SE, Mankoff J (2016) Test time feature ordering with FOCUS. In: Proceedings of the ACM international joint conference on pervasive and ubiquitous computing, ACM, New York, NY, USA, UbiComp '16, pp 992–1003. https://doi.org/10.1145/2971648.2971748
45. Weigl E, Walch A, Neissl U, Meyer-Heye P, Heidl W, Radauer T, Lughofer E, Eitzinger C (2016) MapView: graphical data representation for active learning. CEUR Workshop Proceedings, Sun SITE, Aachen, 1707:3–8
46. Datta S, Adar E (2018) CommunityDiff: visualizing community clustering algorithms. ACM Trans Knowl Discov Data 12(1):1–34. https://doi.org/10.1145/3047009
47. Jain S, Munukutla S, Held D (2019) Few-shot point cloud region annotation with human in the loop. CoRR abs/1906.04409:1–6. arXiv:1906.04409
48. Wallace BC, Small K, Brodley CE, Lau J, Trikalinos TA (2012) Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In: Proceedings of the SIGHIT international health informatics symposium, ACM, New York, NY, USA, pp 819–824. https://doi.org/10.1145/2110363.2110464
49. Zhu Y, Yang K (2019) Tripartite active learning for interactive anomaly discovery. IEEE Access 7:63195–63203. https://doi.org/10.1109/ACCESS.2019.2915388
50. Yan Y, Rosales R, Fung G, Dy JG (2011) Active learning from crowds. In: Proceedings of the international conference on machine learning, Omnipress, USA, ICML'11, pp 1161–1168. http://dl.acm.org/citation.cfm?id=3104482.3104628
51. Cui S, Dumitru CO, Datcu M (2014) Semantic annotation in earth observation based on active learning. Int J Image Data Fusion 5(2):152–174. https://doi.org/10.1080/19479832.2013.858778
52. Burkovski A, Kessler W, Heidemann G, Kobdani H, Schütze H (2011) Self organizing maps in NLP: exploration of coreference feature space. In: Proceedings of the international conference on advances in self-organizing maps, Springer, Berlin, Heidelberg, pp 228–237. https://doi.org/10.1007/978-3-642-21566-7_23
53. Rosenthal SL, Dey AK (2010) Towards maximizing the accuracy of human-labeled sensor data. In: Proceedings of the international conference on intelligent user interfaces, ACM, New York, NY, USA, pp 259–268. https://doi.org/10.1145/1719970.1720006
54. Kagy J, Kayadelen T, Ma J, Rostamizadeh A, Strnadová J (2019) The practical challenges of active learning: Lessons learned from live experimentation. CoRR abs/1907.00038:1–7. arXiv:1907.00038
55. Benato BC, Telea AC, Falcão AX (2018) Semi-supervised learning with interactive label propagation guided by feature space projections. In: SIBGRAPI conference on graphics, patterns and images, pp 392–399. https://doi.org/10.1109/SIBGRAPI.2018.00057
56. Harvey N, Porter R (2016) User-driven sampling strategies in image exploitation. Inf Vis 15(1):64–74. https://doi.org/10.1177/1473871614557659