

5-11-2023

## **Tools of Trade of the Next Blue-Collar Job? Antecedents, Design Features, and Outcomes of Interactive Labeling Systems**

Merlin Knaeble

*Karlsruhe Institute of Technology (KIT), merlin.knaeble@kit.edu*

Mario Nadj

*TU Dortmund University, mario.nadj@tu-dortmund.de*

Luisa Germann

*TU Berlin, l.germann@camput.tu-berlin.de*

Alexander Maedche

*Karlsruhe Institute of Technology (KIT), alexander.maedche@kit.edu*

Follow this and additional works at: [https://aisel.aisnet.org/ecis2023\\_rp](https://aisel.aisnet.org/ecis2023_rp)

---

### **Recommended Citation**

Knaeble, Merlin; Nadj, Mario; Germann, Luisa; and Maedche, Alexander, "Tools of Trade of the Next Blue-Collar Job? Antecedents, Design Features, and Outcomes of Interactive Labeling Systems" (2023). *ECIS 2023 Research Papers*. 373.

[https://aisel.aisnet.org/ecis2023\\_rp/373](https://aisel.aisnet.org/ecis2023_rp/373)

This material is brought to you by the ECIS 2023 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2023 Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact [elibrary@aisnet.org](mailto:elibrary@aisnet.org).

# TOOLS OF TRADE OF THE NEXT BLUE-COLLAR JOB? ANTECEDENTS, DESIGN FEATURES, AND OUTCOMES OF INTERACTIVE LABELING SYSTEMS

*Complete Research*

Knaeble, Merlin, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany,  
merlin.knaeble@kit.edu

Nadj, Mario, TU Dortmund University, Dortmund, Germany, mario.nadj@tu-dortmund.de

Germann, Luisa, TU Berlin, Berlin, Germany, l.germann@camput.tu-berlin.de

Maedche, Alexander, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany,  
alexander.maedche@kit.edu<sup>1</sup>

## Abstract

*Supervised machine learning is becoming increasingly popular - and so is the need for annotated training data. Such data often needs to be manually labeled by human workers, not unlikely to negatively impact the involved workforce. To alleviate this issue, a new information systems class has emerged - interactive labeling systems. However, this young, but rapidly growing field lacks guidance and structure regarding the design of such systems. Against this backdrop, this paper describes antecedents, design features, and outcomes of interactive labeling systems. We perform a systematic literature review, identifying 188 relevant articles. Our results are presented as a morphological box with 14 dimensions, which we evaluate using card sorting. By additionally offering this box as a web-based artifact, we provide actionable guidance for interactive labeling system development for scholars and practitioners. Lastly, we discuss imbalances in the article distribution of our morphological box and suggest future work directions.*

*Keywords:* Interactive Labeling, Annotation, Interactive Machine Learning, Crowdwork.

## 1 Introduction

Machine learning (ML) has become pervasive in our daily lives and work. Current approaches are often based on supervised ML, which, allows scaling beyond hundreds of thousands or even millions of training instances to achieve accurate results on complex problems (Hestness et al., 2017). This strength of scaling, however, is also one of the main drawbacks, i.e. it requires copious amounts of labeled training data. When suitable data is not available and alternatives like synthetic data do not suffice, such labels must be created manually. In this regard, labeling is often performed on crowdworking platforms like MTurk, and is repeatedly referred to as “the blue-collar job of the age of artificial intelligence (AI)” (Gahntz, 2018; Reese, 2016). However, this also has serious downsides: Labeling, being defined as adding information to existing data (Zankl et al., 2012) by human workers has been called a costly, error-prone, and labor-intensive process that often negatively impacts the workers involved (Bernard et al., 2018; Nadj et al., 2020; L. Zhang et al., 2008b). Specifically, in homes across the globe, crowdworkers perform labeling tasks in minuscule pieces (Alpar and Osterbrink, 2020) often leading to motivational problems, adverse performance effects, or boredom (Gadiraju et al., 2019).

---

<sup>1</sup> Research funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project number 447287107.

To alleviate these issues, labeling tasks are nowadays performed in cooperation with machine intelligence. Hereby, a new class of information systems has emerged - that of interactive labeling (IL) systems. IL is defined as “combining automatic steps with incremental user input” (Bernard et al., 2018). IL differentiates from plain labeling tasks by incorporating machine intelligence in an interactive fashion, as opposed to using manual work start to end. For instance, when having crowdworkers annotate documents, Ramos et al. (2020) provide guidance in form of an iterative, incremental process, leveraging the workers capabilities to teach humans a skill. With only minimal training, their novices matched the performance of professionals. Their guidance further leads to lower frustration among participants. Chang et al. (2017) have worked on designing IL systems leveraging the scalability of crowdworking platforms. When having images labeled, they used a series of synchronized stages involving multiple workers each. By allowing for ambiguity and labeling as structures instead of binary decisions, they could decrease instruction effort and allow for an ex-post setting of label boundaries. While it may be easy to decide that a photo of a house cat should be labeled ‘cat’, and that of a Caterpillar-branded (short CAT) excavator should not be, the decision for a wild leopard is not straight forward. Hereby the stages of their system allow for labeler-led creation of new subcategories, if disagreement occurs. This is especially helpful to prevent fatal inconsistencies in the dataset, if, for instance, some labelers find leopards to be cats, while others do not. The IL system asks labelers for explanations for such disagreement, extracts sub-category names, and uses these as new labels to allow for the task issuers to resolve the question after labeling has concluded. So far, a considerable number of articles have already been published despite the young nature of this new class of information systems. Amershi et al. (2014) provide an initial overview, differentiate IL from previous works, and introduce the mistreatment of users as oracles. Further, we have previously studied user archetypes addressed in IL research illustrating two competing perspectives of how users are treated (Knaeble et al., 2020): (a) oracle (e.g. users are asked if a label is correct) versus (b) teacher (e.g. users can provide deeper explanations in labeling tasks). Moreover, recently, we derived five general design principles along prominent literature findings and exemplary IL approaches illustrating that the design of such systems is diverse, and can significantly vary in scope and functionality (Nadj et al., 2020). However, this diversity encompasses a complex of dimensions and characteristics across different IL systems that is not currently well understood. Y. Zhang et al. (2022) have recently argued for more support for those developing IL systems, as they commonly need to be tailor-made. Knowing which design features IL systems offer (and which alternatives exist) is therefore of utmost importance. Hereby, designers and developers of IL systems require an overview of the available design features of IL systems, to be able to make educated decisions. Furthermore, research needs to draw connections between such design features, foregoing antecedents, and following outcomes. Recent research in crowdworking has highlighted the importance of both (Alpar and Osterbrink, 2020). Antecedents are hereby task and context specific requirements and necessities, in the context of IL systems this could be what data is supposed to be labeled, whereas outcomes are goals which are seen as desirable by designers, such as efficiency. By understanding such antecedents and outcomes, IL designers and developers get examples on how to optimize their systems regarding given antecedents or desired outcomes. As IL is a young and popular field of research the need for conducting such a study is exacerbated further. As we will show, the amount of relevant research is increasing rapidly. While we have derived knowledge about how IL systems currently treat users (Knaeble et al., 2020) and we outlined general goals to aim for when creating IL systems (Nadj et al., 2020), there is a gap of design knowledge on how to achieve such goals. Specifically there is no overview of existing design features, foregoing antecedents, and following outcomes. Hereby, research has created, implemented, and evaluated numerous design features of IL systems, but is yet to aggregate such results for a holistic picture. On this basis, we formulate the following research question (RQ): *How to describe antecedents, design features, and outcomes of interactive labeling systems in a morphological box?* Hereby, we rely on a systematic literature review (SLR), deductive and inductive coding, as well as a card sorting evaluation.

Our theoretical contributions lie in the following. First, we present a literature-grounded and user-evaluated systematic overview of a new class of information systems, along antecedents, design features,

and outcomes in form of a morphological box. Thereby, any given IL system can be described along the dimensions and respective characteristics of the box. Second, we provide guidance and a starting point for future research in IL, whereas information systems scholars may use the box to identify gaps in existing research, i.e. dimensions with little previous work. Practically, we contribute by delivering our morphological box not only in form of a static table in this research paper, but also instantiated it as a web-based living artifact. Therefore, one can filter through our 188 included papers by selecting desired configurations of the box. Practitioners thereby benefit from an overview of the possibilities in IL system design and development, as well as a clear outline of required antecedents and expectable outcomes. Further, we invite others to propose new additions to the box, as well as to use its code to serve other paper's boxes. The remainder of this article is structured as follows. Initially we explain foundational concepts in IL. Afterwards we outline our methodology to make our research process transparent. We then present and discuss our results, limitations, and future work, before coming to a conclusion.

## 2 Foundations

IL belongs to the broader research area of IML, which focuses on “building ML models iteratively” (Trivedi, 2016). Within IML, researchers have found benefits in collaborative approaches combining the complementary strengths of humans and ML models. They identify challenges, as well as advantages, ranging in application contexts from teaching robots (Cakmak et al., 2010) to organizational learning (Sturm et al., 2021). Hereby researchers repeatedly identify the need for a guiding overview to coordinate such efforts and call for further research on the topic (Amershi et al., 2014; Sturm et al., 2021) and find strong performance benefits from combining manual and automated approaches in labeling (Gu and Leroy, 2019). T. Zhang et al. (2020) report an increase in label granularity without negative effects on mental workload, while Hamidi-Haines et al. (2019) discuss avenues to increase users trust in the product of their labeling efforts. Figure 1 shows an overview of previous conceptualizations of IL research. Firstly, we outline the related domains of IL and IML, as explained by Amershi et al. (2014), based on a visualization of Trivedi (2016). This is in contrast to the user-centered perspective we take in Knaeble et al. (2020), whereas IL systems treat users either as the as teachers or oracles. Lastly, in Nadj et al. (2020) we have already offered five design principles as a guiding lens for IL systems design (c.f. Figure 1).

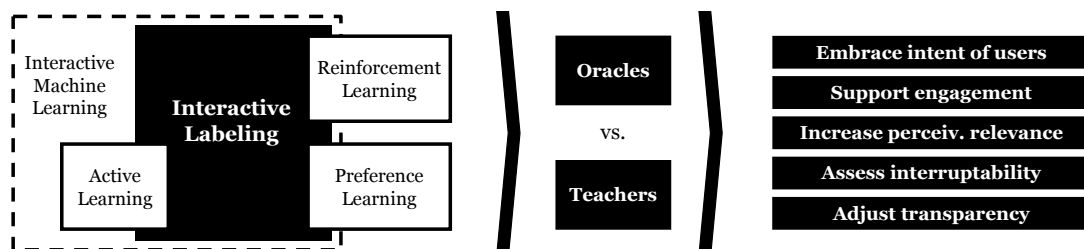


Figure 1. Previous Conceptualizations of IL Research Based on Amershi et al. (2014), Knaeble et al. (2020), Nadj et al. (2020), and Trivedi (2016).

The IML paradigm has been founded as a reaction to the shortcomings of active learning (AL) with regard to its user involvement (Amershi et al., 2014). Hereby, established AL approaches mistreat the user as an oracle (e.g. users are repeatedly asked if a label is correct). AL thereby works by querying a label for the unlabeled data instance in which the system is most uncertain, i.e. sees the most potential for learning. Such approaches may lead to frustration, a loss of trust, and ultimately a lower perceived system performance (Amershi et al., 2014; Cakmak et al., 2010). Further research implies motivational deficiencies for crowdworkers in such restricted setups (Durward et al., 2020). Organizational shortcomings like wage-theft further contribute to this issue (Shafiei Gol et al., 2019). On the technical side, AL has difficulties with modern architectures like deep learning, as they cannot deliver the early uncertainty estimations that

AL requires for sample selection, nor quickly retrain on a single instance. Additionally long training cycles could delay each user interaction. These two-fold criticisms have led to the IL paradigm as a derivative of AL (Amershi et al., 2014). IL further intersects with preference learning, as well as reinforcement learning. Preference learning hereby defines preferences towards certain errors or error types, to “refine the decision boundaries” (Amershi et al., 2014). Techniques from preference learning could indicate tendency towards certain errors. Imagine, e.g. a production scenario in which quality assurance might rather falsely discard an actually good part, in place of mistakenly keeping a faulty one. Reinforcement learning on the other hand is a paradigm in which the learning agent bases its decision on feedback in form of rewards or punishments, issued by a human, an automated judge, or another ML system (Knaeble et al., 2020; Trivedi, 2016). Cakmak et al. (2010) have shown how the IL paradigm helps to differentiate between human and machine judges.

### 3 Methodology

To address our RQ, we combine several methodological approaches in sequence. Our initial set of publications is derived from a SLR. Based on these SLR results, we employ qualitative methods combining bottom-up with top-down coding, to derive a morphological box organized along the three focal points of interest (i.e. antecedents, design features, and outcomes). Finally, we evaluate and subsequently adapt our box with a card sorting study, relying on both qualitative insights from interview workshops, and quantitative measures of fit.

#### 3.1 Systematic Literature Review

SLRs are an established research method in the field of information systems, and are commonly used to create an overview of emergent information system classes (e.g. Haug and Maedche, 2021). Thereby, we follow established guidelines to plan, conduct, and report upon our review and its results (Kitchenham and Charters, 2007; Webster and Watson, 2002).

Replicating the successful search strategy of Knaeble et al. (2020), who have previously investigated user roles in IL, we employ a search string along two parts. Part I is ('interactive machine learning' AND ('annot\*' OR 'label\*')) OR 'interactive annot\*' OR 'interactive label\*', with word-stemming, whereas 'annot\*' matches e.g. 'annotate', 'annotation', or 'annotates'. We created this part I as we identified early IL works, and such from different fields, to be using the terms IL and interactive annotation synonymously. Additionally, we added a criterion to include works referencing the IML paradigm in conjunction with labeling or annotation. As IL and IML intersect with several underlying learning approaches, as explained by Trivedi (2016), we include part II to capture articles that refer only to them. However, as such terms not necessarily relate to IL we employed further filtering to place an emphasis on the interactive user role in the labeling process: ('annot\*' OR 'label\*') AND 'data\*' AND 'interact\*' AND 'user\*' AND ('active learning' OR 'preference learning' OR 'reinforcement learning'). These related fields are mirrored in the overview in Figure 1. We joined the two parts with an OR operator (I OR II).

Knaeble et al. (2020) identified six highly cited luminary articles within IL (introducing the IML paradigm, Fails and Olsen, 2003; acknowledging AL drawbacks, Culotta and McCallum, 2005; introducing IL, Tian et al., 2007; using dynamic termination, Branson et al., 2011; considering labeling costs, Joshi et al., 2012; contrasting previous user mistreatment against new interactive approaches, Amershi et al., 2014). We chose the four databases that returned hits on any of these six luminary papers: the ACM Digital Library, Web of Science, Scopus, and IEEE Xplore. Knaeble et al. (2020) offer a detailed account of the luminary articles and their return on these databases. Hereby, we included only peer-reviewed articles in English. Also, we only considered articles that introduced an IL system. The aim of this had to be to interactively collect training data. Among those articles excluded, most either mentioned that as a prerequisite to their training process data needed to be labeled (without offering details), or emphasized the need for iteratively

refining their resulting ML model in an IL process. Articles that focused on algorithm development, primarily from AL, and on how to optimize training cost were excluded as well.

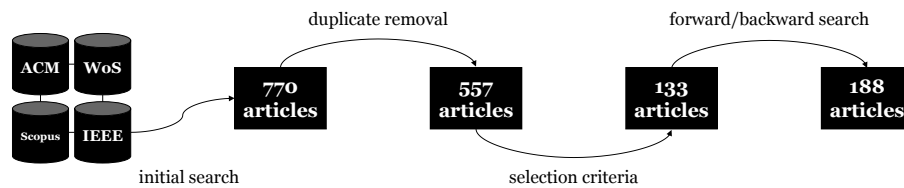


Figure 2. Search Results of the Systematic Literature Review.

We visualize our hit statistics in Figure 2. All steps in our search process were performed by two different researchers working in parallel. Upon conflicts they re-reviewed the article with a third research to reach a consensus. This was invoked for 14 articles, all of which were decided to be kept. Our literature search process therefore produced an inter-coder agreement of 97.49%. Applying our search string, we retrieved a total of 770 hits across the four databases. After removing duplicates, we were left with 557 unique articles to consider. We then applied our above mentioned selection criteria, to identify 133 articles. We followed the iterative approach outlined by Kitchenham and Charters (2007), reading titles, keywords, and abstracts to exclude clear cases. For remaining articles the authors retrieved the full text and decided upon this basis. As a last step, we followed the recommendations of Webster and Watson (2002) to perform a so-called forward/backward search. Hereby, we iterated over each of the 133 search results and applied the identical filter criteria to the papers cited in its reference list (backward search), as well as to papers citing the identified paper via Google Scholar (forward search). During this, we identified a total of 55 additional articles, for a total of 188 relevant hits for further analysis. With Fails and Olsen (2003) we find our earliest matching article from 2003, while more than half of all articles found are from 2020 or later. This further supports our argument of IL as a recently emerging class of information systems.

### 3.2 Morphological Box & Coding

We choose to present our results in form of a morphological box (Zwicky, 1967). Such boxes have been used successfully as a reliable classification scheme to structure results of literature reviews (Álvarez and Ritchey, 2015; Haug and Maedche, 2021) and are commonly defined as “a creative way of illustrating all the potential solutions to existing problems in a structured format” (Kley et al., 2011). Morphological boxes break down subjects into their fundamental dimensions, and describing each dimension by potential characteristics (Wissema, 1976). In our case, beyond an overview of predominant antecedents and outcomes, as well as available design features, it serves to compare among them and identify most feasible approaches (Kley et al., 2011).

We performed inductive coding, i.e. the characteristics were derived bottom-up. As a methodological guide for inductive coding, we relied on the grounded theory approach (Corbin and Strauss, 1990; Wolfswinkel et al., 2013) along open, axial, and selective coding. Hereby, open coding is to extract excerpts, so that concepts may start to emerge, and to capture underlying information (Wolfswinkel et al., 2013). The initial goal of our coding was to identify users, data/label properties, as well as design features of the IL systems. The latter remained in our morphological box, whereas users split into antecedents (merging with data/label properties) and outcomes. Axial coding strives to begin aforementioned aggregation by identifying main characteristics. At this point we had identified 146 characteristics in 22 dimensions. Lastly, selective coding refines the results. Key actions are revisiting the reasoning behind such emerging characteristics and dimensions, eliminating duplicate coverings, and re-organizing the results (Kley et al., 2011). All authors collaborated on this. We completed the coding process with the 14 dimensions and their characteristics as visualized in Figure 4, however at this stage differently named. To perform our coding, two researchers worked in parallel. For mismatches, a third researcher as involved to reach a

consensus. During bottom-up coding for each of the 188 articles we thus decided the characteristic for all 14 dimensions. 484 of this total of 2632 assignment decisions were discussed in the aforementioned format, leading to an inter-coder agreement of 81.61%.

To complement this bottom-up approach, we applied deductive coding (top-down) by deriving dimensions and characteristics from existing literature. Specifically, we identified two dimensions of integral components of IL systems, already covered by previous work. These were integrated into our bottom-up box (c.f. our Results chapter, on the design features of **guidance** and **instance relations**).

### 3.3 Card Sorting Evaluation

For structured classification schemes like morphological boxes or taxonomies to be beneficial to its users, the categories it is based on matter most (Bailey, 1994; Nickerson et al., 2013), their names and definitions should be meaningful, following a clear logic (Gregor, 2006). To evaluate such categorical groupings card sorting procedures are the method of choice (Moore and Benbasat, 1991). Using a web-based card sorting tool (<https://kardsort.com>), we presented participants with either all dimensions in the antecedents, in the design features, or in the outcome space as categories to sort cards (i.e. the characteristics) into. In a randomized order, the participants were presented with a list of all the characteristics, and were asked to assign them. They could reconsider and change each assignment until they confirmed their submission, but had to assign all in order to proceed. For both characteristics and dimensions, we provided a definition (as can be found in our following Section as well as Tables 1, 2, and 3). Trivial characteristics in form of “not reported” (none of the characteristics applicable) or “mixed” (combination of other characteristics) were not included in the sorting procedure.

We used a crowdworking platform (<https://www.prolific.co>) to recruit participants that had at least a Bachelor’s degree, as well as experience in software development. Further, we only recruited native English speakers. One round of sorting was conducted for each of the antecedents and outcomes. In turn, for the design features we performed two rounds of sorting and one round of interviews. For each round of sorting, we recruited 10 participants, and an additional five for the round of interviews. On average, participants were 35 years old (with a standard deviation of 10.5). Among our total of 45 participants were 26 women and 19 men. One participant held a doctorate, 11 a Master’s, and 33 a Bachelor’s degree. We briefed them on the concept of IL and the method of card sorting. The round durations varied between those sorting antecedents (average of 9 min), design features (32 min for the first round, 15 min for the second), and outcomes (11 min) as would be expected due to the difference in complexity. Interviews took 34 minutes on average. Notably, the sorting time was cut in half between the two rounds of sorting the design features. We compensated according to this duration, resulting in an effective hourly wage of over 9 USD. Participants could only take part in one round. Following the recommendations of Moore and Benbasat (1991), we iterated over our process, until two established measures of inter-rater and agreement with our results confirm a meaningful and natural organization.

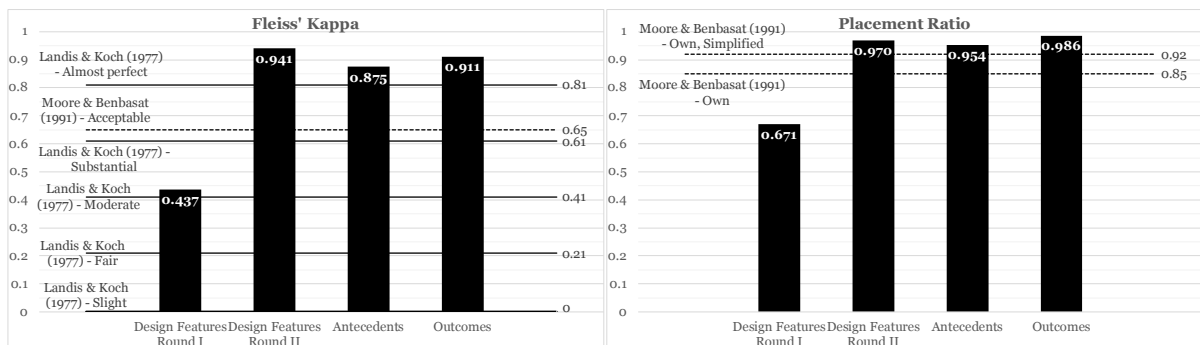


Figure 3. Results of the Card Sorting Evaluation.

To assess inter-rater agreement, we rely on Fleiss' Kappa, an extension of Cohen's Kappa for more than two raters (Fleiss, 1971). Such agreement measures indicate reliability, hence that different raters or users understand the same concepts. Moore and Benbasat (1991) define a value of 0.65 as acceptable. Landis and Koch (1977) give a range of values, whereas they define values starting from 0 as slight, above 0.21 as fair, such above 0.41 as moderate, over 0.61 as substantial, and finally values above 0.81 as almost perfect. It must be noted, however, that the thresholds from Landis and Koch (1977) were initially defined for scenarios with comparably few raters (in their case 2) and few categories (4). Increasing either will lead to a lower expected measure. Further, we measure how much the raters agree with our morphological box by calculating the placement ratio (i.e. how many characteristics are placed in the correct dimension) following Moore and Benbasat (1991). They do not define a threshold value for this metric, as for one they frame it more as an exploratory tool, and second they argue that desired values are context dependent. However, they report and accept a value of 0.85 for a complex task, and one of 0.92 for a simplified one.

Figure 3 shows an overview of our results, as well as the aforementioned thresholds. We began by evaluating the more complex sorting of the design features. In our initial round I, participants produced low scores for Fleiss' Kappa (0.437) as well as placement ratio (0.671). Following the suggestions by Rugg and McGeorge (2005), we then performed a round of five interviews to better understand thoughts and conflicts. Interviews were performed online, while the screen was shared. We recorded all interviews for later analysis. The task presented to the participants was identical to the card sorting rounds. To gain insights into their approach, the interviewer asked, and continuously reminded, participants to think aloud. Questions for clarification purposes were asked intermittently, based on the articulated thoughts. After sorting has commenced, the interviewer iterated over those design features that were sorted wrongly, asking to explain their reasoning behind their choice. Following each interview, we refined the naming and definitions in our box. After the fifth interview, we had reached a stage, where only minor changes needed to be made. Hence, we performed round II to confirm our assumption of a stable arrangement. High inter-rater agreements (0.941) and placement ratios (0.970) confirm this.

For the considerably less complex antecedents and outcomes, the initial round of sorting already produced Fleiss' Kappas (0.875 and 0.911) as well as placement ratios (0.954 and 0.986) exceeding all thresholds. Hence, here we accepted without further adjustment.















Antecedents	 <b>Type of Label</b>	Binary Classification (32)		Multi-Class Classification (97)		Bounding Box / Rough Location (17)		Segmentation / Exact Location (20)		Meta Labels (15)		Other Types (7)	
	 <b>Type of Data</b>	Imagery (66)		3D Imagery (11)		Audio-Visual Recordings (20)		Text (31)		Compound & Structured Data (43)		Time-Series Data (17)	
	 <b>Type of Labeler</b>	Domain Experts (67)				Nondescript Users (110)				Crowdworkers (11)			
Design Features	 <b>Error Treatment</b>	Unerring Oracle (129)				Ignore User Errors (36)				Identify User Errors (23)			
	 <b>Guidance</b>	Prescribing (37)			Directing (51)			Orienting (54)			Not Reported (46)		
	 <b>Label Input Format</b>	Creation from Scratch (69)			Approval or Correction (23)			Approval or Rejection (9)			Mixed (87)		
	 <b>Proactivity</b>	System (90)						User (98)					
	 <b>Subset Sampling</b>	Choice-based (42)		Comparison-based (16)		Uncertainty-based (50)		Error-based (3)		Random (45)		Mixed (32)	
	 <b>Order of Presentation</b>	By Instance (84)				By Label (3)				Random (101)			
	 <b>Termination Criterion</b>	User Dependent (28)			Prediction Quality (11)			All Data is Labeled (18)			Budget Constraint (131)		
	 <b>Instance Relations</b>	Instance Between-Group (16)			Instance Within-Group (30)			Instance Neighbors (27)			Instance Only (115)		
	 <b>Model Training</b>	Pre-trained (14)			On-the-fly (53)			Pre- and Re-trained (79)			Ex-post (42)		
Outcomes	 <b>Performance</b>	Efficiency (54)			Effectiveness (24)			Both (79)			Not Reported (31)		
	 <b>Psychology</b>	Well-being (18)		Trust (8)	Perceived Usefulness (8)		Cognitive Load (8)		Motivation (5)		Boredom (2)	Mixed (17)	

Figure 4. Resulting Morphological Box (Number of Papers per Characteristic in Parentheses).

## 4 Results

Figure 4 illustrates the resulting morphological box. We show the number of articles representing each characteristic in parentheses. The dimensions are structured along our three points of focus: antecedents, design features, and outcomes. In this context, we define antecedents as externally given requirements directly related to the goal of the labeling process. Design features are commonly referred to as “elements that users see, hear, touch, or operate” (Han et al., 2000; Liu and Yu, 2017). Notable exclusions in the context of labeling systems would be technical features like choice of ML algorithm, data structure, or data transmission. We focus on the user side of IL systems. Lastly, we define outcomes as the measurable consequences of the labeling process, be it on the product of the work, on the time and money required to perform it, or on the people doing so. In the following subsections, we define the **dimensions** and their respective *characteristics*. In addition to the overview in Figure 4 we offer our morphological box as a web-based artifact<sup>2</sup>. Our artifact allows for filtering the list of references, also attached to this paper, by selecting characteristics. Tables 1, 2, and 3 list all dimensions, along with their definitions.

### 4.1 Antecedents




	<b>Type of Label</b>	The kind of label the system seeks to assign to each data instance.
	<b>Type of Data</b>	The data format of the unlabeled data instances.
	<b>Type of Labeler</b>	The level of expertise, involvement, and origin of users.

Table 1. Dimensions and Associated Definitions for the Antecedents.

The dimension **type of label** describes the kind of label the system seeks to assign to its data instances. If it is a *binary classification* task, the user provides one of two classes (e.g. yes/no type of decisions). In contrast, in *multi-class classification* tasks there are three or more class labels to be chosen from when assigning labels to instances. Further types of labels are the *bounding box / rough location*, where the user provides a bounding box or another rough location input (e.g. center of body), and a class label. More accurate is the *segmentation / exact location*, referring to pixel perfect segmentation or another exact location inputs (e.g. polygon trace as in Ling et al., 2019), again with a class label. For *meta-labels* the user provides labels that are not directly connected to the output the system wants to achieve, e.g. general performance feedback. In rare cases, we find diverse *other types*.

A second antecedent dimension is the **type of data**, which is the format of the unlabeled data instances present. Common are *imagery* data, i.e. two-dimensional photography. We call their three-dimensional counterpart, e.g. from LIDAR cameras as shown by Boyko and Funkhouser (2014), *3D imagery*. *Audio-visual-recordings* refer to videos as well as sound recordings. *Text* data are for instance messages, reviews, or interview transcripts. Tabular and multidimensional data-sets are referred to as *compound & structured data*. Lastly, *time-series data* are sequences of data points with timestamps, e.g. sensor data-streams.

For our third dimension of antecedents, we identify different **types of labelers**, hence a user’s level of expertise, type, and involvement with the system. On one hand, we find *domain experts*, which have experience in the target field required to execute the respective labeling, e.g. doctors labeling diseases in radiology images (X. Wu et al., 2021b). A large group is represented within the *nondescript users*, with no clearly specified expertise, but also with no distinct optimization towards crowds. These are within the *crowdworker* characteristic, specifically designed for scaling.










	<b>Error Treatment</b>	The acknowledgement of user input errors and their handling.
	<b>Guidance</b>	The support being provided by the system for the user.
	<b>Label Input Format</b>	The format of the labels, exactly as provided by the user.
	<b>Proactivity</b>	The active component that requests or provides label input.
	<b>Subset Sampling</b>	The strategy to choose which data to assign labels to.
	<b>Order of Presentation</b>	The arrangement in which instances are shown to the user.
	<b>Termination Criterion</b>	The benchmark upon which to stop the labeling process.
	<b>Instance Relations</b>	The application of label information beyond one instance.
	<b>Model Training</b>	The usage of machine learning models during labeling.

Table 2. Dimensions and Associated Definitions for the Design Features.

## 4.2 Design Features

With our first design feature **error treatment**, we identify three options for the acknowledgment of user input errors and how to handle them. For the *unerring oracle*, the user input is considered as completely error free and there is no error treatment taking place. The system designers *ignore user errors* if they recognize that input may be erroneous, but do not take up countermeasures. They trust the learning system to cope with faulty data. They may instead *identify user errors* by acknowledging erroneous input, and installing precautionary features to detect and prevent (at least some) user errors. Commonly this is done by aggregating labels across multiple coders via majority voting (e.g. T. Zhang et al., 2020).

We base our definition of **guidance**, as well as its three main characteristics on the works of Ceneda et al. (2017). It refers to whether the system provides any support for the user, hence advice or information aimed at resolving a problem, or pre-structuring of the data. Following Ceneda et al. (2017), the highest degree of guidance is *prescribing*, hence supporting the user by steering them toward a predefined goal by specifying a fixed process, giving explanations on why automated steps are undertaken, and enabling the user to intervene. S. Das et al. (2020) provide such support by depicting by color where in the feature space the user has already taken influence and what effect this had on the model, hence steering them away from regions not promising further improvement. At a medium level is *directing* guidance, which supports the user by presenting them with a potential course of action, e.g. previews to make informed decisions or assistant-like features to show options. Lowest ranks *orienting*, i.e. supporting the user by building or maintaining their mental map of the system by providing visual cues or overviews to resolve knowledge gaps, like visually structuring or presenting the data. We additionally include a characteristic for papers that do *not report* guidance features.

The **label input format** is the form of input that the user provides for the system. Here lies an important distinction to the antecedent of type of label, which refers to the types of labels that the system wants to assign, not how the user enters them. For instance, in *label creation from scratch*, the user enters exactly what is expected as the type of label, hence labels are provided and entered by the user without any further support. They could however also interact via *label approval or correction* in which case the user decides whether a predetermined label is correct and approves it if so. If not, they enter a new label manually. Hollandi et al. (2020), for instance, allow the user to accept accurate cell segmentations, or to redraw them if the suggestion turns out erroneous. The characteristic *label approval or rejection* further abstracts the input from the desired type of label, whereas the user then also decides whether a predetermined label is correct and approves it if so. If not, however, they simply reject the suggestion. There also exist *mixed* implementations of these three input formats, whereas the system may switch between them, e.g. depending on whether it can yet make a sufficiently confident suggestion.

**Proactivity** refers to which party is the leading one in the labeling process, especially with regard to requesting or giving input. As such there are two options. A proactive *system* requests specific label input

<sup>2</sup> <https://human-centered-systems-lab.github.io/ilmbbox/>

interactions from the user. Whereas alternatively, the *user* may provide input proactively and freely at their discretion by requesting where to interact and provide label input for the system. For instance, Majumder and Yao (2019) rely on the user in such a regard.

If there are more unlabeled data instances available than can be labeled, a **subset sampling** needs to be performed. It can be *choice-based* if the user selects a subset on whatever basis they find appropriate (e.g. selecting representative examples for each class). System strategies encompass a *comparison-based* selection by comparing unlabeled instances, e.g. by selecting similar or dissimilar examples. For an *uncertainty-based* subset, the system performs the selection via its prediction uncertainty on yet unseen, hence unlabeled, examples. Like Joshi et al. (2012), such IL systems are typically grounded in the AL paradigm. Whereas for the *test-based* subset the system's predictions are evaluated against an externally given ground truth, i.e. labeled examples. We define systems without an explicitly stated strategy as having a *random* selection process, as there is no conscious decision on what data to label. Even if all available data is labeled, this also constitutes a random subset of the real world data that will have to be assessed in the future. Lastly, we identify *mixed* strategies, combining the aforementioned ones.

Label interactions can be seen as tuples  $(x, y)$  of data instances  $x$  and label classes  $y$ . For instance, you might have a series of family photos as instances. Imagine, as label classes you might want to identify which photos contain your mother, your brother, daughter, best friend, or other relatives and acquaintances. To now decide how to arrange this series of labeling decisions  $(x, y)$  for the user we refer to as the **order of presentation**. There exist three fundamental options to sort such  $(x, y)$  tuples. One option is ordering the presentation *by instance* first, hence presenting all decisions on one instance (i.e. all label classes for this instance), before moving to the next instance. In our example you would label all known faces in a photo, before moving to the next. With this order of presentation you avoid having to review one instance multiple times. For complex multi-class labeling tasks, this could however be difficult. An alternative is ordering *by label* first, hence presenting all decisions on one label class ordered subsequently (i.e. all instances for this label class), before moving to the next label class. Hence, you would first search for (and consequently label) photos containing e.g. your daughter. Such an order of presentation has unique advantages, e.g. if this label class has specific importance, or if you have system-based suggestions already grouping probable occurrences together. Lastly, one could sort at *random* with no specific arrangement to the order in which the user is presented with their task (i.e. neither by label nor by instance). In our example, the system could ask you whether this vacation photo contains your spouse, and in the next image show you a selfie, asking if the person in the background is your best friend from college. While such an arrangement may seem inconsequential at first sight, from a user's perspective, random ordering is what e.g. typical AL model training involves. If the system selects the next instance and label based on an internal metric, the user has no transparent ordering by either one of the tuple's components.

To stop the labeling process, either as invoked by the user or by the system, IL systems rely on predefined **termination criteria**. A *user dependent* stop is performed manually, based on the user's assessment of the system performance or an inter-labeler agreement. Vargas-Munoz et al. (2019) argue that this allows for expert users to better decide on the trade-off between exploration and exploitation. Alternatively, one system based termination criterion is when a certain *prediction quality* has been reached. The system thus stops the labeling process when the prediction quality of the model it has been training with the labeled data exceeds a certain threshold. Further, the system may choose to stop only when *all data is labeled*, i.e. the best possible base for training a capable model has been created. Lastly, the violation of certain *budget constraints* play a role. Here, the system stops the labeling process when a certain budget is exceeded, e.g. time/money spent on labeling, or amount of labels assigned.

As with guidance, for the **instance relations** we rely on a pre-existing top-down structure comprising a range of four categories by Bernard et al. (2021) and adapted the following definitions from their work. Therefore, this dimension represents how label information is applied beyond the focus instance, suggestions the system makes regarding relationships between instances, and how the labels are spread through the instances. Highest on the spectrum proposed by Bernard et al. (2021) rank the *instance between-group* relations. Hereby, label information is applied to not only the focus instance it was

assigned to, but also to instances across multiple pre-defined groups of instances. Clustering approaches, pre-structuring the available unlabeled data into such groups, are one example of such instance relations. Next lower are the *instance within-group* relations, where information is only applied within a pre-defined group around the focus instance. Whereas for the *instance neighbors* information applies only to instances in the focus instances local spatial region. Lowest rank *instance only* approaches, that represent no further use of the label information beyond the focus instance it was initially assigned to.

As a last design feature, we identify **model training**. This dimension reflects how ML models support the labeling, when they are trained, and whether feedback is being redirected back into the learning process. For once, the IL system could rely on a *pre-trained* model. For such designs, the ML model used to support the labeling process was trained beforehand, but is not being refined during the interaction process. Data to train such a model could come from other labeling sources, synthetic generation, from a general training-dataset not offering the specific classes required, or be an old dataset in need of updating due to e.g. concept drift. A different approach is being taken in *on-the-fly* model training, where the supporting model is created only during the interaction process. The *pre- and re-trained* approach combines the previous two. A model trained beforehand is refined with user input during the interaction. Lastly, there could be no ML model being used during the labeling process. If such a model is being trained only after the interaction is finished, we categorize the IL systems under *ex-post* model training.

### 4.3 Outcomes

🏆	<b>Performance</b>	The objectively measured system output in form of resulting labels.
🧠	<b>Psychology</b>	The subjective impacts the system interaction had on involved workers.

Table 3. Dimensions and Associated Definitions for the Outcomes.

We identify two **performance** measures. Herein we categorized reports regarding the objectively measured system output with a focus on results of the labeling task. *Efficiency* is improving the amount of time (measured as such, in interactions/clicks, or in monetary terms) required to assign a label or a set amount of them. *Effectiveness* on the other hand refers to improving the accuracy and completeness of the labels that are being assigned as compared against a factual ground truth. There are also articles within our search measuring *both* or *not reporting* performance measures at all.

The second dimension refers to **psychological** outcomes, i.e. reports on the impacts the interaction with the labeling system had on its users, generally measured subjectively via a user's perceptions. We find the following constructs among the identified papers. *Well-being* refers to a user's psychological health, functioning, and pleasantness in and after the labeling task. *Trust*, often addressed in explainable ML approaches, is defined as the perceived reliability and trustworthiness in the system. *Perceived usefulness*, as a well established construct in technology acceptance studies in information systems research is present for IL as well. It refers to the degree to which users believe the system enhances their labeling task and outcome. The *cognitive load* is defined as the mental effort it takes to assign labels with the system, whereas *motivation* refers to internal or external factors driving user behaviour. Additionally, we find *boredom* being investigated as a user's adaption to monotony or absence of autonomy. Furthermore, some articles report a *mixed* combination of the aforementioned outcomes, or do *not report* any.

## 5 Discussion

To answer our RQ, we have provided our morphological box (c.f. Figure 4). In the following, we will discuss noticeable insights about the distribution of articles within it. Hereby, these insights may serve as a starting point for future work on the topic of IL systems.

Regarding the types of label and data, we find no such insights, as articles are spread evenly across the characteristics. However, we find little focus on IL systems optimized for crowdworking. Such endeavours could coincide with rectifying the misalignment towards accepting input as error-free, or ignoring such potential input errors. For instance, Chang et al. (2017) introduce a crowd-centric approach with several synchronised stages involving many workers simultaneously. Hereby, they allow for inter-worker comparison administrated by workers themselves. Further, they allow workers to explicitly state their own inconfidence in certain label decisions. Such approaches, if transferred to other contexts, could aim to resolve two shortcomings (crowdworking and error treatment) of recent IL works at once.

As with the label and data types, there seems to be a wide-covered spread of work on different stages of guidance, formats of label input, options for proactivity, and subset sampling strategies. However, for the latter, there is tendency to use uncertainty-based samples (typical for AL, no ground truth required), as compared to error-based ones. A reason for this could be the issue of a cold start, whereas such a ground truth dataset to evaluate errors on would have to be labeled firstly. We find a further misbalance regarding the order of presentation. As such, we identify only three articles ordering their interaction by label. One of them is Arendt et al. (2019), who focus on labeling handwritten digits. They allow for selection and review ordered by classes, allowing users to focus on peculiarities of certain digits, e.g. ones with and without left-downward tick, or sevens, with and without a cross-bar. While their approach benefits from the limited amount of label classes, as well as a good pre-classification, this approach to structuring the fundamental sequence of labeling interactions demands further investigation.

While there is a strong focus on budget constraints, as compared to other termination criteria, it seems logical that such decisions are often made due to practical limitations. Whereas with the instance relations, we find the majority of articles working with only the current in-focus instance to disregard potential benefits of other approaches. By structuring and interconnecting unlabeled instances, label propagation can be applied to rapidly label large datasets. For instance, Lee et al. (2021) use superpixel segmentation to create disjunct groups of similar cells in medical whole-slide imaging for cancer detection. These superpixels can then be labeled with one click, instead of having to repeat the label assignment for every, almost identical cell within it. They have however yet to perform a real-world user evaluation, confirming the initially good results from simulated tests. For the last design feature of model training, we find a sensible focus on model improvement during the labeling process. Training only after the labeling has concluded seems to be mainly motivated by high training times for e.g. deep neural nets.

For our outcome dimensions of performance and psychology, we find a severe imbalance towards the former. All but 31 articles report at least either one of efficiency or effectiveness, if not both. In contrast, 122 articles do not report any user-centered psychological outcomes. The spread of the remainder over six different constructs, often with different operationalizations, underlines the recentness of the field of IL and its rapid growth across different research domains. Regardless, our results show a concerning lack of regard for the labeling workers in many of the studies.

## 5.1 Limitations & Future Research

Although we followed established methodological guidelines to ensure a high rigor of our morphological box of IL systems, this study is not without limitations to consider.

While our search strategy has been adopted from Knaeble et al. (2020), any potential bias in it, will replicate here. Following the recommendations of Webster and Watson (2002), we conducted a forward-backward search to mitigate potential biases from the search string or the selection of the databases. Furthermore, we addressed potential biases during the coding and article selection with a multi-coder approach. As IL is a young and rapidly growing discipline around an emergent class of information systems, this morphological box cannot claim exhaustive coverage of all possible design feature dimensions. Researchers could come up with entirely new ones, leading to the potential future need to update our results accordingly. The same holds true for characteristics. This becomes most apparent for the psychological outcomes, where studies must simply include a new construct in their user evaluations, to demand inclusion in an updated box.

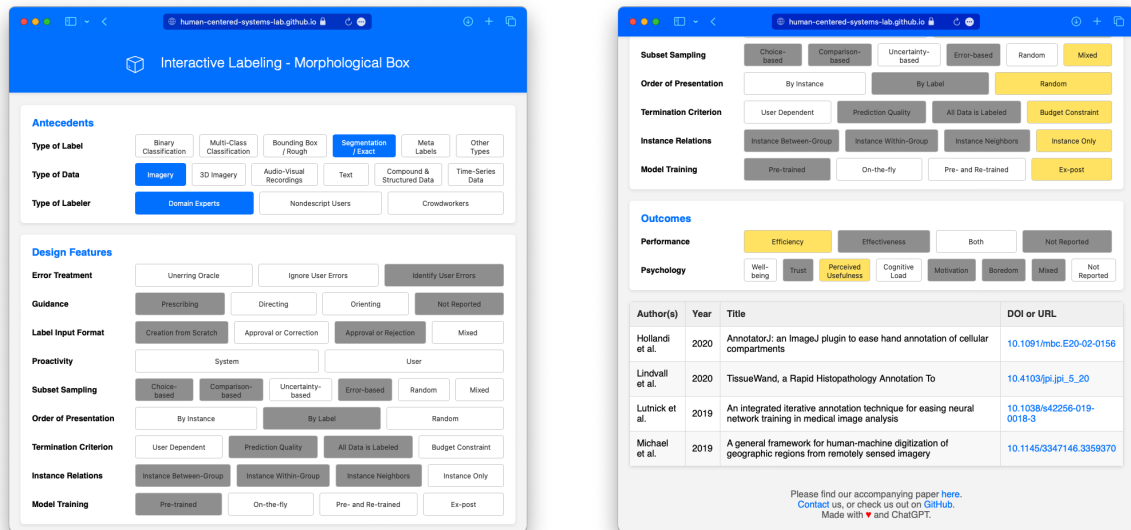


Figure 5. Screenshots from <https://human-centered-systems-lab.github.io/ilmbbox/>

Future research could further expand our box by including technical features. For instance the choice of ML algorithms or the underlying data structures could also be assessed.

Our evaluation was only performed with potential users of IL systems. While this has its strengths in supporting the comprehensibility of the box, general recommendations extend to additional expert evaluations in a real-world context (Moore and Benbasat, 1991; Nickerson et al., 2013). In the future, a study should be carried out with IL practitioners as target audience, allowing researchers to observe how such a morphological box is used for IL design and development. To facilitate such an endeavour, we have already provided our results as an easy to access web-based artifact, and invite the IL community to use it, to jointly transform this static framework into a living one, which can grow with the field of IL research. This web-based artifact can also be of use to address the aforementioned point of the design features of IL systems expanding in the future, eventually demanding an extension of the morphological box we provide here. Researchers could also expand our artifact into a recommender system, e.g. integrating empiric evidence about the effects of certain design features. Furthermore, context can determine quite powerful restrictions and requirements for the IL system. Our web-based artifact allows for filtering of the morphological box based on such restrictions. If for instance, the IL system is supposed to be used for segmenting images in the medical context, and only domain experts can be used as a labeler type, then this can be selected in our artifact. The result will then show a box specific for domain expert IL systems, omitting, e.g. design features commonly found for crowdworker application. We see further use for this artifact to be spun-off into other literature reviews resulting in morphological boxes and have consequently made it open source. Future researchers may therefore simply provide their coded list of papers, as well as the structuring of a morphological box, and the artifact will generate a matching front-end.

## 6 Conclusion

In the age of AI the creators of such models often rely on cheap labor - for instance from crowdworking platforms, home of a potential new blue-collar job - to generate labeled data. Labeling workers typically perform their tasks supported by IL systems, a new class of information systems, which form their tools of trade. Research has, until now, not offered a morphological box of antecedents, design features, and both task- and worker-centric outcomes of such systems. With this work we provide such an overview to support future development of the tools of trade of an emergent class of blue-collar workers.

## References

- Acuna, D., H. Ling, A. Kar, and S. Fidler (2018). *Efficient Interactive Annotation of Segmentation Datasets with Polygon-RNN++*. Tech. rep.
- Adhikari, B., E. Rahtu, and H. Huttunen (2021). *Sample Selection for Efficient Image Annotation*. Tech. rep.
- Alberto, B., D. L. Andrea, M. Eric, and T. Fabiano (2019). “Active Learning of Predefined Models for Information Extraction: Selecting Regular Expressions from Examples.” *Frontiers in Artificial Intelligence and Applications* 320, 645–651.
- Alpar, P. and L. Osterbrink (2018). “Antecedents of Perceived Fairness in Pay for Microtask Crowdwork.” In: *ECIS Research-in-Progress Papers*.
- (2020). “Antecedents and Consequences of Perceived Fairness in Pay for Crowdwork.” In: *ICIS*.
- Álvarez, A. and T. Ritchey (2015). “Applications of General Morphological Analysis.” 4 (1), 40.
- Amershi, S., M. Cakmak, W. B. Knox, and T. Kulesza (2014). “Power to the People: The Role of Humans in Interactive Machine Learning.” *AIMag* 35 (4), 105–120.
- Andersen, J. S., O. Zukunft, and W. Maalej (2021). “REM: Efficient Semi-Automated Real-Time Moderation of Online Forums.”
- Arendt, D., L. R. Franklin, F. Yang, B. R. Brisbois, and R. R. LaMothe (2018). “Crush Your Data with ViC 2 ES Then CHISSL Away.” In: *2018 IEEE Symposium on Visualization for Cyber Security (VizSec)*. IEEE, pp. 1–8.
- Arendt, D., C. Komurlu, and L. M. Blaha (2017). *CHISSL: A Human-Machine Collaboration Space for Unsupervised Learning*. Tech. rep. PNNL-SA-124302, pp. 429–448.
- Arendt, D., E. Saldanha, R. Wesslen, S. Volkova, and W. Dou (2019). “Towards Rapid Interactive Machine Learning.” In: *Proceedings of the 24th International Conference on Intelligent User Interfaces*. Ed. by W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, and G. Calvary. New York, NY, USA: ACM, pp. 591–602.
- El-Assady, M., R. Kehlbeck, C. Collins, D. Keim, and O. Deussen (2020). “Semantic Concept Spaces: Guided Topic Model Refinement Using Word-Embedding Projections.” *IEEE transactions on visualization and computer graphics* 26 (1), 1001–1011.
- El-Assady, M., F. Sperrle, O. Deussen, D. Keim, and C. Collins (2018). “Visual Analytics for Topic Model Optimization Based on User-Steerable Speculative Execution.” *IEEE transactions on visualization and computer graphics*.
- Bahrami, M. and W.-P. Chen (2019). “WATAPI: Composing Web API Specification from API Documentations through an Intelligent and Interactive Annotation Tool.” In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 4573–4578.
- Bailey, K. D. (1994). *Typologies and Taxonomies: An Introduction to Classification Techniques*. Quantitative Applications in the Social Sciences 07-102. Thousand Oaks, Calif: Sage Publications.
- Baur, T., A. Heimerl, F. Lingenfelder, J. Wagner, M. F. Valstar, B. Schuller, and E. André (2020). “eXplainable Cooperative Machine Learning with NOVA.” *Künstl Intell* 34 (2), 143–164.
- Beil, D. and A. Theissler (2020). “Cluster-Clean-Label: An Interactive Machine Learning Approach for Labeling High-Dimensional Data.” In: *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*. Ed. by M. Burch, M. Westenberg, Q. V. Nguyen, and Y. Zhao. New York, NY, USA: ACM, pp. 1–8.
- Benato, B. C., A. C. Telea, and A. X. Falcão (2018). “Semi-Supervised Learning with Interactive Label Propagation Guided by Feature Space Projections.” In: *2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pp. 392–399.
- Benato, B. C., J. F. Gomes, A. C. Telea, and A. X. Falcão (2021). “Semi-Automatic Data Annotation Guided by Feature Space Projection.” *Pattern Recognition* 109, 107612.
- Berg, S., D. Kutra, T. Kroeger, C. N. Straehle, B. X. Kausler, C. Haubold, M. Schiegg, J. Ales, T. Beier, M. Rudy, K. Eren, J. I. Cervantes, B. Xu, F. Beuttenmueller, A. Wolny, C. Zhang, U. Koethe, F. A.

- Hamprecht, and A. Kreshuk (2019). “Ilastik: Interactive Machine Learning for (Bio)Image Analysis.” *Nature methods* 16 (12), 1226–1232.
- Bernard, J., M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair (2018). “Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study.” *IEEE transactions on visualization and computer graphics* 24 (1), 298–308.
- Bernard, J., E. Dobermann, A. Vögele, B. Krüger, J. Kohlhammer, and D. Fellner (2017a). “Visual-Interactive Semi-Supervised Labeling of Human Motion Capture Data.” *ei* 29 (1), 34–45.
- Bernard, J., M. Hutter, M. Sedlmair, M. Zeppelzauer, and T. Munzner (2021). “A Taxonomy of Property Measures to Unify Active Learning and Human-centered Approaches to Data Labeling.” *ACM Transactions on Interactive Intelligent Systems* 11 (3-4), 1–42.
- Bernard, J., C. Ritter, D. Sessler, M. Zeppelzauer, and D. Fellner (2017b). “Visual-Interactive Similarity Search for Complex Objects by Example of Soccer Player Analysis.” In: pp. 75–87.
- Bobák, P., L. Čmolík, and M. Čadík (2019). “Video Sequence Boundary Labeling with Temporal Coherence.” In: *Advances in Computer Graphics*. Ed. by M. Gavrilova, J. Chang, N. M. Thalmann, E. Hitzer, and H. Ishikawa. Vol. 11542. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 40–52.
- Boyko, A. and T. Funkhouser (2014). *Cheaper by the Dozen*, pp. 33–42.
- Branson, S., P. Perona, and S. Belongie (2011). “Strong Supervision from Weak Annotation: Interactive Training of Deformable Part Models.” *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, 1832–1839.
- Brighi, M., A. Franco, and D. Maio (2020). “A Semi-Supervised Learning Approach for CBIR Systems with Relevance Feedback.” In: *Thirteenth International Conference on Machine Vision*. Ed. by W. Osten, J. Zhou, and D. P. Nikolaev. SPIE, p. 4.
- Burkovski, A., W. Kessler, G. Heidemann, H. Kobdani, and H. Schütze (2011). “Self Organizing Maps in NLP: Exploration of Coreference Feature Space.” *undefined*.
- Cakmak, M., C. Chao, and A. L. Thomaz (2010). “Designing Interactions for Robot Active Learners.” *IEEE Transactions on Autonomous Mental Development* 2 (2), 108–118.
- Cao, L., W. Tao, S. An, J. Jin, Y. Yan, X. Liu, W. Ge, A. Sah, L. Battle, J. Sun, R. Chang, B. Westover, S. Madden, and M. Stonebraker (2019). “Smile: A System to Support Machine Learning on EEG Data at Scale.” *Proc. VLDB Endow.* 12 (12), 2230–2241.
- Ceneda, D., T. Gschwandtner, T. May, S. Miksch, H.-J. Schulz, M. Streit, and C. Tominski (2017). “Characterizing Guidance in Visual Analytics.” *IEEE Transactions on Visualization and Computer Graphics* 23 (1), 111–120.
- Chang, J. C., S. Amershi, and E. Kamar (2017). “Revolt: Collaborative Crowdsourcing for Labeling Machine Learning Datasets.” *undefined*.
- Chatani, S., Y. Ma, H. Zhang, Y. Chen, and W. Du (2020). “Interactive Labeling System for Lung Nodules with CT Images.” In: *2020 International Symposium on Computer, Consumer and Control (IS3C)*. IEEE, pp. 529–532.
- Chegini, M., J. Bernard, P. Berger, A. Sourin, K. Andrews, and T. Schreck (2019). “Interactive Labelling of a Multivariate Dataset for Supervised Machine Learning Using Linked Visualisations, Clustering, and Active Learning.” *Visual Informatics* 3 (1), 9–17.
- Chegini, M., J. Bernard, J. Cui, F. Chegini, A. Sourin, K. Andrews, and T. Schreck (2020). “Interactive Visual Labelling versus Active Learning: An Experimental Comparison.” *Front Inform Technol Electron Eng* 21 (4), 524–535.
- Chen, B., H. Ling, X. Zeng, J. Gao, Z. Xu, and S. Fidler (2020). “ScribbleBox: Interactive Annotation Framework for Video Object Segmentation.” In: *Computer Vision - ECCV 2020: 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings*. Ed. by H. Bischof, T. Brox, and A. Vedaldi. Vol. 12346–12375. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 293–310.

- Choi, M., C. Park, S. Yang, Y. Kim, J. Choo, and S. R. Hong (2019). "AILA: Attentive Interactive Labeling Assistant for Document Classification through Attention-based Deep Neural Networks." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Ed. by S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos. New York, NY, USA: ACM, pp. 1–12.
- Christen, V., P. Christen, and E. Rahm (2020). "Informativeness-Based Active Learning for Entity Resolution." In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by P. Cellier and K. Driessens. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 125–141.
- Corbin, J. M. and A. Strauss (1990). "Grounded Theory Research: Procedures, Canons, and Evaluative Criteria." *Qualitative Sociology* 13 (1), 3–21.
- Cui, S., C. O. Dumitru, and M. Datcu (2014). "Semantic Annotation in Earth Observation Based on Active Learning." *International Journal of Image and Data Fusion* 5 (2), 152–174.
- Culotta, A. and A. McCallum (2005). "Reducing Labeling Effort for Structured Prediction Tasks." In: vol. 2, pp. 746–751.
- Dahl, V. A., M. J. Emerson, C. H. Trinderup, and A. B. Dahl (2020). "Content-Based Propagation of User Markings for Interactive Segmentation of Patterned Images." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, pp. 4280–4288.
- Danyluk, K., T. T. Ulusoy, W. Wei, and W. Willett (2020). "Touch and Beyond: Comparing Physical and Virtual Reality Visualizations." *IEEE transactions on visualization and computer graphics* PP.
- Das, K., I. Avrekh, B. Matthews, M. Sharma, and N. Oza (2017). "ASK-the-Expert: Active Learning Based Knowledge Discovery Using the Expert." In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by Y. Altun, K. Das, T. Mielikäinen, D. Malerba, J. Stefanowski, J. Read, M. Žitnik, M. Ceci, and S. Džeroski. Cham: Springer International Publishing, pp. 395–399.
- Das, S., W.-K. Wong, T. Dietterich, A. Fern, and A. Emmott (2020). "Discovering Anomalies by Incorporating Feedback from an Expert." *ACM Trans. Knowl. Discov. Data* 14 (4), 1–32.
- Datta, S. and E. Adar (2018). "Community Diff: Visualizing Community Clustering Algorithms." *ACM Transactions on Knowledge Discovery from Data* 12, 1–34.
- Deng, D., J. Wu, J. Wang, Y. Wu, X. Xie, Z. Zhou, H. Zhang, X. Zhang, and Y. Wu (2021). "EventAnchor: Reducing Human Interactions in Event Annotation of Racket Sports Videos." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Ed. by Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker. New York, NY, USA: ACM, pp. 1–13.
- Dennig, F. L., T. Polk, Z. Lin, T. Schreck, H. Pfister, and M. Behrisch (2019). "FDive: Learning Relevance Models Using Pattern-based Similarity Measures." In: *2019 IEEE Conference on Visual Analytics Science and Technology: Proceedings : Vancouver, BC, Canada, 20-25 October 2019*. Ed. by R. Chang, D. Keim, and R. Maciejewski. Piscataway, NJ: IEEE, pp. 69–80.
- Desmond, M., E. Duesterwald, K. Brimijoin, M. Brachman, and Q. Pan (2021a). "Semi-Automated Data Labeling."
- Desmond, M., M. Muller, Z. Ashktorab, C. Dugan, E. Duesterwald, K. Brimijoin, C. Finegan-Dollak, M. Brachman, A. Sharma, N. N. Joshi, and Q. Pan (2021b). "Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface." In: *26th International Conference on Intelligent User Interfaces*. Ed. by T. Hammond, K. Verbert, D. Parra, B. Knijnenburg, J. O'Donovan, and P. Teale. New York, NY, USA: ACM, pp. 392–401.
- Dias, A. G., E. E. Milios, and M. C. F. Oliveira (2019). "TRIVIR: A Visualization System to Support Document Retrieval with High Recall." In: *Proceedings of the ACM Symposium on Document Engineering 2019*. Ed. by U. Borghoff and S. Schimmler. New York, NY, USA: ACM, pp. 1–10.
- Durward, D., I. Blohm, and J. M. Leimeister (2020). "The Nature of Crowd Work and Its Effects on Individuals' Work Perception." *Journal of Management Information Systems* 37 (1), 66–95.
- Dutta, A. and A. Zisserman (2019). "The VIA Annotation Software for Images, Audio and Video." In: *Proceedings of the 27th ACM International Conference on Multimedia*. Ed. by L. Amsaleg, B.

- Huet, M. Larson, G. Gravier, H. Hung, C.-W. Ngo, and W. Tsang Ooi. New York, NY, USA: ACM, pp. 2276–2279.
- Eirich, J., D. Jackle, T. Schreck, J. Bonart, O. Posegga, and K. Fischbach (2020). “VIMA: Modeling and Visualization of High Dimensional Machine Sensor Data Leveraging Multiple Sources of Domain Knowledge.” In: *2020 Visualization in Data Science (VDS)*. IEEE, pp. 22–31.
- Eva weigl, A. Walch, U. Neissl, Pauline Meyer-Hey, and C. Eitzinger (2016). “MapView: Graphical Data Representation for Active Learning.” In.
- Fails, J. A. and D. R. Olsen (2003). *Interactive Machine Learning*, p. 39.
- Fallgren, P., Z. Malisz, and J. Edlund (2019). “How to Annotate 100 Hours in 45 Minutes.” In: *Interspeech 2019*. ISCA: ISCA, pp. 341–345.
- Fan, W., Y. Si, W. Yang, and G. Zhang (2021). “Active Broad Learning System for ECG Arrhythmia Classification.” *Measurement* 185, 110040.
- Fan, X., C. Li, X. Yuan, X. Dong, and J. Liang (2019). “An Interactive Visual Analytics Approach for Network Anomaly Detection through Smart Labeling.” *Journal of vision* 22 (5), 955–971.
- Fleiss, J. L. (1971). “Measuring Nominal Scale Agreement among Many Raters.” *Psychological Bulletin* 76 (5), 378–382.
- Françoise, J., B. Caramiaux, and T. Sanchez (2021). “Marcelle: Composing Interactive Machine Learning Workflows and Interfaces.”
- Frisoli, K., B. LeRoy, and R. Nugent (2019). “A Novel Record Linkage Interface That Incorporates Group Structure to Rapidly Collect Richer Labels.” In: *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, pp. 580–589.
- Gadiraju, U., G. Demartini, R. Kawase, and S. Dietze (2019). “Crowd Anatomy Beyond the Good and Bad: Behavioral Traces for Crowd Worker Modeling and Pre-selection.” *Computer Supported Cooperative Work (CSCW)* 28 (5), 815–841.
- Gahntz, M. (2018). *The Invisible Workers of the AI Era*. <https://towardsdatascience.com/the-invisible-workers-of-the-ai-era-c83735481ba>.
- Ghai, B., Q. V. Liao, Y. Zhang, R. Bellamy, and K. Mueller (2021). “Explainable Active Learning (XAL).” *Proc. ACM Hum.-Comput. Interact.* 4 (CSCW3), 1–28.
- Gregor, S. (2006). “The Nature of Theory in Information Systems.” *MIS Quarterly* 30 (3), 611.
- Grimmeisen, B. and A. Theissler (2020). “The Machine Learning Model as a Guide.” In: *Proceedings of the 13th International Symposium on Visual Information Communication and Interaction*. Ed. by M. Burch, M. Westenberg, Q. V. Nguyen, and Y. Zhao. New York, NY, USA: ACM, pp. 1–8.
- Gu, Y. and G. Leroy (2019). “Mechanisms for Automatic Training Data Labeling for Machine Learning.” *ICIS 2019 Proceedings*.
- Guan, S., W. Peng, J. Song, and Z. Xu (2020). “The Construction of Interactive Environment for Sentence Pattern Structure Based Treebank Annotation.” In: *Chinese Lexical Semantics*. Ed. by J.-F. Hong, Y. Zhang, and P. Liu. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 753–763.
- Hamidi-Haines, M., Z. Qi, A. Fern, F. Li, and P. Tadepalli (2019). “Interactive Naming for Explaining Deep Neural Networks: A Formative Study.” In: *Joint Proceedings Of the ACM IUI2019 Workshops, Los*.
- Han, S. H., M. Hwan Yun, K.-J. Kim, and J. Kwahk (2000). “Evaluation of Product Usability: Development and Validation of Usability Dimensions and Design Elements Based on Empirical Models.” *International Journal of Industrial Ergonomics* 26 (4), 477–488.
- Haridas, A., F. Bunyak, and K. Palaniappan (2015). “Interactive Segmentation Relabeling for Classification of Whole-Slide Histopathology Imagery.” In: *2015 IEEE 28th International Symposium on Computer-Based Medical Systems*, pp. 84–87.
- Harvey, N. and R. Porter (2013). “User-Driven Sampling Strategies in Image Exploitation.” *Visualization and Data Analysis 2014* 9017, 90170.

- Haug, S. and A. Maedche (2021). "Crowd-Feedback in Information Systems Development: A State-of-the-Art Review." In: *ICIS 2021 Proceedings*.
- Heimerl, A., T. Baur, F. Lingenfelder, J. Wagner, and E. Andre (2019). "NOVA - A Tool for eXplainable Cooperative Machine Learning." In: *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII 2019): Cambridge, United Kingdom, 3-6 September 2019*. Piscataway, NJ: IEEE, pp. 109–115.
- Heimerl, A., K. Weitz, T. Baur, and E. Andre (2020). "Unraveling ML Models of Emotion with NOVA: Multi-Level Explainable AI for Non-Experts." *IEEE Transactions on Affective Computing*, 1–1.
- Heo, J., J. Park, H. Jeong, Kwang joon Kim, J. Lee, E. Yang, and S. J. Hwang (2020). "Cost-Effective Interactive Attention Learning with Neural Attention Processes." *Proceedings of the 37th International Conference on Machine Learning, Online*.
- Hestness, J., S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. M. A. Patwary, Y. Yang, and Y. Zhou (2017). "Deep Learning Scaling Is Predictable, Empirically." *arXiv:1712.00409 [cs, stat]*. arXiv: 1712.00409 [cs, stat].
- Hollandi, R., Á. Diósdí, G. Hollandi, N. Moshkov, and P. Horváth (2020). "AnnotatorJ: An ImageJ Plugin to Ease Hand Annotation of Cellular Compartments." *Molecular biology of the cell* 31 (20), 2179–2186.
- Honeycutt, D. R., M. Nourani, and E. D. Ragan (2020). "Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy."
- Hossain, H. M. S. and N. Roy (2019). "Active Deep Learning for Activity Recognition with Context Aware Annotator Selection." In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Ed. by A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, and G. Karypis. New York, NY, USA: ACM, pp. 1862–1870.
- Huang, L., S. Matwin, Eder J. de Carvalho, and R. Minghim (2017). "Active Learning with Visualization for Text Data." In: pp. 69–74.
- Ishibashi, T., Y. Nakao, and Y. Sugano (2020). "Investigating Audio Data Visualization for Interactive Sound Recognition." In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Ed. by F. Paternò, N. Oliver, C. Conati, L. D. Spano, and N. Tintarev. New York, NY, USA: ACM, pp. 67–77.
- Islam, M. R., S. Das, J. R. Doppa, and S. Natarajan (2020). "GLAD: GLocalized Anomaly Detection via Human-in-the-Loop Learning." In: *Proceedings of the International Conference on Machine Learning 2020*.
- Jing, J., E. d'Angremont, S. Zafar, E. S. Rosenthal, M. Tabaeizadeh, S. Ebrahim, J. Dauwels, and M. B. Westover (2018). "Rapid Annotation of Seizures and Interictal-ictal Continuum EEG Patterns." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2018*, 3394–3397.
- Joshi, A. J., F. Porikli, and N. P. Papanikolopoulos (2012). "Scalable Active Learning for Multiclass Image Classification." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34 (11), 2259–73.
- Kellenberger, B., D. Marcos, S. Lobry, and D. Tuia (2019). "Half a Percent of Labels Is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning." *IEEE Transactions on Geoscience and Remote Sensing* 57 (12), 9524–9533.
- Kellenberger, B., D. Tuia, and D. Morris (2020). "AIDE: Accelerating Image-Based Ecological Surveys with Interactive Machine Learning." *Methods in Ecology and Evolution* 11 (12), 1716–1727.
- Khamesi, A. R., E. Shin, and S. Silvestri (2020). "Machine Learning in the Wild: The Case of User-Centered Learning in Cyber Physical Systems." In: *2020 International Conference on COMMunication Systems & NETworkS (COMSNETS)*. IEEE, pp. 275–281.
- Kim, B. (2018). "Leveraging User Input and Feedback for Interactive Sound Event Detection and Annotation." *undefined*.
- Kim, B. and B. Pardo (2017). "I-SED: An Interactive Sound Event Detector." In: pp. 553–557.

- Kim, B. and B. Pardo (2018). “A Human-in-the-Loop System for Sound Event Detection and Annotation.” *The ACM Transactions on Interactive Intelligent Systems* 8 (2), 1–23.
- Kim, H. and Y.-k. Lim (2021). “Teaching-Learning Interaction: A New Concept for Interaction Design to Support Reflective User Agency in Intelligent Systems.” In: *Designing Interactive Systems Conference 2021*. Ed. by W. Ju, L. Oehlberg, S. Follmer, S. Fox, and S. Kuznetsov. New York, NY, USA: ACM, pp. 1544–1553.
- Kim, J., J. Hwang, S. Chi, and J. Seo (2020). “Towards Database-Free Vision-Based Monitoring on Construction Sites: A Deep Active Learning Approach.” *Automation in Construction* 120, 103376.
- Kime, K., T. Hickey, and R. Torrey (2019). “Refining Skill Classification with Interactive Machine Learning.” In: *FIE Cincinnati 2019: 2019 Conference Proceedings*. [Piscataway, NJ]: IEEE, pp. 1–8.
- Kitchenham, B. and S. Charters (2007). *Guidelines for Performing Systematic Literature Reviews in Software Engineering*.
- Kley, F., C. Lerch, and D. Dallinger (2011). “New Business Models for Electric Cars—A Holistic Approach.” *Energy Policy* 39 (6), 3392–3403.
- Klute, F., G. Li, R. Löffler, M. Nöllenburg, and M. Schmidt (2019). “Exploring Semi-Automatic Map Labeling.” In: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Ed. by F. Banaei-Kashani, G. Trajcevski, R. H. Güting, L. Kulik, and S. Newsam. New York, NY, USA: ACM, pp. 13–22.
- Knaeble, M., M. Nadj, and A. Maedche (2020). “Oracle or Teacher? A Systematic Overview of Research on Interactive Labeling for Machine Learning.” In: *WI2020 Zentrale Tracks*. GITO Verlag, pp. 2–16.
- Kockelkorn, T., R. Ramos, J. Ramos, C. Sánchez, and B. van Ginneken (2012). “Interactive Classification of Lung Tissue in CT Scans by Combining Prior and Interactively Obtained Training Data: A Simulation Study.” *Pattern Recognition (ICPR), 2012 21st International Conference on*, 105–108.
- Kockelkorn, T. T. J. P., P. A. Jong, H. A. Gietema, J. C. Grutters, M. Prokop, and B. van Ginneken (2010). “Interactive Annotation of Textures in Thoracic CT Scans.” *Proceedings of SPIE - The International Society for Optical Engineering*, 76240.
- Kockelkorn, T. T. J. P., P. A. Jong, C. M. Schaefer-Prokop, R. Wittenberg, A. M. Tiehuis, H. A. Gietema, J. C. Grutters, M. A. Viergever, and B. van Ginneken (2016). “Semi-Automatic Classification of Textures in Thoracic CT Scans.” *Physics in medicine and biology* 61 (16), 5906–24.
- Konishi, K., T. Nonaka, S. Takei, K. Ohta, H. Nishioka, and M. Suga (2021). “Reducing Manual Operation Time to Obtain a Segmentation Learning Model for Volume Electron Microscopy Using Stepwise Deep Learning With Manual Correction.” *Microscopy : the journal of the Quekett Microscopical Club*.
- Kuebert, T., H. Puder, and H. Koepl (2019). “Daily Routine Recognition with Visual Interactive Labeling by Fusing Acceleration and Audio Signals.” In: *2019 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT 2019): Ajman, United Arab Emirates, 10-12 December 2019*. Piscataway, NJ: IEEE, pp. 1–6.
- Kutsuna, N., T. Higaki, S. Matsunaga, T. Otsuki, M. Yamaguchi, H. Fujii, and S. Hasezawa (2012). “Active Learning Framework with Iterative Clustering for Bioimage Classification.” *Nature Communications* 3 (1), 1032.
- Kuznetsova, A., A. Talati, Y. Luo, K. Simmons, and V. Ferrari (2020). “Efficient Video Annotation With Visual Interpolation and Frame Selection Guidance.”
- Lahtinen, T., H. Turtiainen, and A. Costin (2021). “Brima: Low-Overhead Browser-Only Image Annotation Tool (Preprint).” In: *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, pp. 2633–2637.
- Landis, J. R. and G. G. Koch (1977). “The Measurement of Observer Agreement for Categorical Data.” *Biometrics* 33 (1), 159.
- Laux, L., M. F. A. Cutiongco, N. Gadegaard, and B. S. Jensen (2020). “Interactive Machine Learning for Fast and Robust Cell Profiling.” *PLoS One* 15 (9), 0237972.

- Lee, S., M. Amgad, P. Mobadersany, M. McCormick, B. P. Pollack, H. Elfandy, H. Hussein, D. A. Gutman, and L. A. D. Cooper (2021). "Interactive Classification of Whole-Slide Imaging Data for Cancer Researchers." *Cancer research* 81 (4), 1171–1177.
- Legg, P., J. Smith, and A. Downing (2019). "Visual Analytics for Collaborative Human-Machine Confidence in Human-Centric Active Learning Tasks." 9 (1).
- Lekschas, F., B. Peterson, D. Haehn, E. Ma, N. Gehlenborg, and H. Pfister (2020). "Peax: Interactive Visual Pattern Search in Sequential Data Using Unsupervised Deep Representation Learning." *Computer Graphics Forum* 39 (3), 167–179.
- Li, E., S. Wang, C. Li, D. Li, X. Wu, and Q. Hao (2020). "SUSTech POINTS: A Portable 3D Point Cloud Interactive Annotation Platform System." In: *2020 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1108–1115.
- Li, H., S. Fang, S. Mukhopadhyay, A. J. Saykin, and L. Shen (2018). "Interactive Machine Learning by Visualization: A Small Data Solution." *Proc IEEE Int Conf Big Data* 2018, 3513–3521.
- Liang, J., N. Homaounfar, W.-C. Ma, Y. Xiong, R. Hu, and R. Urtasun (2020). "PolyTransform: Deep Polygon Transformer for Instance Segmentation." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 9128–9137.
- Liao, X., W. Li, Q. Xu, X. Wang, B. Jin, X. Zhang, Y. Wang, and Y. Zhang (2020). "Iteratively-Refined Interactive 3D Medical Image Segmentation With Multi-Agent Reinforcement Learning." In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 9391–9399.
- Lindvall, M., A. Sanner, F. Petré, K. Lindman, D. Treanor, C. Lundström, and J. Löwgren (2020). "TissueWand, a Rapid Histopathology Annotation Tool." *Journal of pathology informatics* 11 (1), 27.
- Ling, H., J. Gao, A. Kar, W. Chen, and S. Fidler (2019). "Fast Interactive Object Annotation With Curve-GCN." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5252–5261.
- Liu, N. and R. Yu (2017). "Identifying Design Feature Factors Critical to Acceptance and Usage Behavior of Smartphones." *Computers in Human Behavior* 70, 131–142.
- Lopresti, D. and G. Nagy (2012). "Optimal Data Partition for Semi-Automated Labeling." *Pattern Recognition (ICPR), 2012 21st International Conference on*, 286–289.
- Lorbach, M., R. Poppe, and R. C. Veltkamp (2019). "Interactive Rodent Behavior Annotation in Video Using Active Learning." *Multimedia tools and applications* 78 (14), 19787–19806.
- Lu, T., Y. Gui, and Z. Gao (2021). "Learning Document-Level Label Propagation and Instance Selection by Deep Q-Network for Interactive Named Entity Annotation." *IEEE access : practical innovations, open solutions* 9, 39568–39586.
- Lutnick, B., B. Ginley, D. Govind, S. D. McGarry, P. S. LaViolette, R. Yacoub, S. Jain, J. E. Tomaszewski, K.-Y. Jen, and P. Sarder (2019). "An Integrated Iterative Annotation Technique for Easing Neural Network Training in Medical Image Analysis." *Nature machine intelligence* 1 (2), 112–119.
- Luus, F., I. Akhalwaya, and N. Khan (2018). "Cognitive-Assisted Interactive Labeling of Skin Lesions and Blood Cells." In: *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, pp. 398–405.
- Luus, F., N. Khan, and I. Akhalwaya (2019). "Interactive Supervision with T-SNE." In: *Proceedings of the 10th International Conference on Knowledge Capture*. Ed. by M. Kejrival, P. Szekely, and R. Troncy. New York, NY, USA: ACM, pp. 85–92.
- Macadam, A., C. J. Nowell, and K. Quigley (2021). "Machine Learning for the Fast and Accurate Assessment of Fitness in Coral Early Life History." *Remote Sensing* 13 (16), 3173.
- Majumder, S. and A. Yao (2019). "Content-Aware Multi-Level Guidance for Interactive Instance Segmentation." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 11594–11603.
- Mallinar, N., A. Shah, T. K. Ho, R. Ugrani, and A. Gupta (2020). "Iterative Data Programming for Expanding Text Classification Corpora." *AAAI* 34 (08), 13332–13337.

- Marinelli, F., A. Cervone, G. Tortoreto, E. A. Stepanov, G. Di Fabbri, and G. Riccardi (2019). "Active Annotation: Bootstrapping Annotation Lexicon and Guidelines for Supervised NLU Learning." In: *Interspeech 2019*. ISCA: ISCA, pp. 574–578.
- Michael, C. J., S. M. Dennis, C. Maryan, S. Irving, and M. L. Palmsten (2019). "A General Framework for Human-Machine Digitization of Geographic Regions from Remotely Sensed Imagery." In: *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. Ed. by F. Banaei-Kashani. ACM Digital Library. New York, NY, United States: Association for Computing Machinery, pp. 259–268.
- Mishra, S. and J. M. Rzeszutski (2021). "Designing Interactive Transfer Learning Tools for ML Non-Experts." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Ed. by Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, and S. Drucker. New York, NY, USA: ACM, pp. 1–15.
- Moore, G. C. and I. Benbasat (1991). "Development of an Instrument to Measure the Perceptions of Adopting an Information Technology Innovation." *Information Systems Research* 2 (3), 192–222.
- Muthukrishnan, H. P. and D. A. Szafir (2020). "Using Machine Learning and Visualization for Qualitative Inductive Analyses of Big Data."
- Nadj, M., M. Knaeble, M. X. Li, and A. Maedche (2020). "Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning." *KI - Künstliche Intelligenz* 34 (2), 131–142.
- Nagy, G. and X. Zhang (2011). "CalliGUI: Interactive Labeling of Calligraphic Character Images." *Document Analysis and Recognition (ICDAR), 2011 International Conference on*, 977–981.
- Nalishnik, M., D. A. Gutman, J. Kong, and L. A. D. Cooper (2015). "An Interactive Learning Framework for Scalable Classification of Pathology Images." In: *2015 IEEE International Conference on Big Data (Big Data)*, pp. 928–935.
- Nashaat, M., A. Ghosh, J. Miller, and S. Quader (2020). "Asterisk: Generating Large Training Datasets with Automatic Active Supervision." *ACM/IMS Trans. Data Sci.* 1 (2), 1–25.
- Nashaat, M., A. Ghosh, J. Miller, S. Quader, C. Marston, and J.-F. Puget (2019). "Hybridization of Active Learning and Data Programming for Labeling Large Industrial Datasets." In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 46–55.
- Nickerson, R. C., U. Varshney, and J. Muntermann (2013). "A Method for Taxonomy Development and Its Application in Information Systems." *European Journal of Information Systems* 22 (3), 336–359.
- Oh, S. W., J.-Y. Lee, N. Xu, and S. J. Kim (2019). "Fast User-Guided Video Object Segmentation by Interaction-And-Propagation Networks." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 5242–5251.
- Okuma, K., E. Brochu, D. G. Lowe, and J. J. Little (2011). "An Adaptive Interface for Active Localization." In: pp. 248–258.
- Ouyang, W., T. Le, H. Xu, and E. Lundberg (2021). "Interactive Biomedical Segmentation Tool Powered by Deep Learning and ImJoy." *F1000Research* 10, 142.
- Park, S. and C. M. Yang (2019). "Interactive Video Annotation Tool for Generating Ground Truth Information." In: *2019 Eleventh International Conference on Ubiquitous and Future Networks (ICUFN)*. IEEE, pp. 552–554.
- Pedronette, D. C. G., Y. Weng, A. Baldassin, and C. Hou (2019). "Semi-Supervised and Active Learning through Manifold Reciprocal kNN Graph for Image Retrieval." *Neurocomputing* 340, 19–31.
- Piazzentin Ono, J., A. Gjoka, J. Salamon, C. Dietrich, and C. T. Silva (2019). "HistoryTracker: Minimizing Human Interactions in Baseball Game Annotation." In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. Ed. by S. Brewster, G. Fitzpatrick, A. Cox, and V. Kostakos. New York, NY, USA: ACM, pp. 1–12.
- Plummer, B. A., M. H. Kiapour, S. Zheng, and R. Piramuthu (2018). *Give Me a Hint! Navigating Image Databases Using Human-in-the-loop Feedback*. Tech. rep.
- Pohl, D., A. Bouchachia, and H. Hellwagner (2018). "Batch-Based Active Learning: Application to Social Media Data for Crisis Management." *Expert Systems with Applications* 93, 232–244.

- Radeva, P., M. Drozdal, S. Segui, L. Igual, C. Malagelada, F. Azpiroz, and J. Vitria (2012). "Active Labeling: Application to Wireless Endoscopy Analysis." In: *2012 International Conference on High Performance Computing & Simulation (HPCS)*, pp. 174–181.
- Rajadell, O., P. Garcia-Sevilla, V. C. Dinh, and R. P. W. Duin (2011). "Semi-Supervised Hyperspectral Pixel Classification Using Interactive Labeling." *Hyperspectral Image and Signal Processing: Evolution in Remote Sensing (WHISPERS), 2011 3rd Workshop on*, 1–4.
- Ramos, G., C. Meek, P. Simard, J. Suh, and S. Ghorashi (2020). "Interactive Machine Teaching: A Human-Centered Approach to Building Machine-Learned Models." *Human-Computer Interaction* 35 (5-6), 413–451.
- Reddy, S., A. D. Dragan, S. Levine, S. Legg, and J. Leike (2020). "Learning Human Objectives by Evaluating Hypothetical Behavior." *37th International Conference on Machine Learning, ICML 2020 Part F* 168147–11.
- Reese, H. (2016). *Is 'data Labeling' the New Blue-Collar Job of the AI Era?* <https://www.techrepublic.com/article/is-data-labeling-the-new-blue-collar-job-of-the-ai-era/>.
- Rietz, T. and A. Maedche (2020). *Towards the Design of an Interactive Machine Learning System for Qualitative Coding*. Tech. rep. AIS eLibrary (AISEL).
- Rietz, T., P. Toreini, and A. Maedche (2020). "Cody: An Interactive Machine Learning System for Qualitative Coding." In: *Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology*. UIST '20 Adjunct. New York, NY, USA: Association for Computing Machinery, pp. 90–92.
- Riou, M., B. Jabaian, S. Huet, and F. Lefevre (2019). "Joint On-line Learning of a Zero-shot Spoken Semantic Parser and a Reinforcement Learning Dialogue Manager." In: *2019 IEEE International Conference on Acoustics, Speech, and Signal Processing: Proceedings : May 12-17, 2019, Brighton Conference Centre, Brighton, United Kingdom*. Piscataway, NJ: IEEE, pp. 3072–3076.
- Rosenthal, S. L. and A. K. Dey (2010). "Towards Maximizing the Accuracy of Human-Labeled Sensor Data." In: *Proceedings of the 15th International Conference on Intelligent User Interfaces - IUI '10*. Hong Kong, China: ACM Press, p. 259.
- Rudovic, O., M. Zhang, B. Schuller, and R. Picard (2019). "Multi-Modal Active Learning From Human Data: A Deep Reinforcement Learning Approach." In: *2019 International Conference on Multimodal Interaction*. Ed. by W. Gao, H. M. Ling Meng, M. Turk, S. R. Fussell, B. Schuller, Y. Song, and K. Yu. New York, NY, USA: ACM, pp. 6–15.
- Rugg, G. and P. McGeorge (2005). "The Sorting Techniques: A Tutorial Paper on Card Sorts, Picture Sorts and Item Sorts." *Expert Systems* 22 (3), 94–107.
- Sarkar, A., C. Morrison, J. F. Dorn, R. Bedi, S. Steinheimer, J. Boisvert, J. Burggraaff, M. D'Souza, P. Kotschieder, S. R. Bulò, L. Walsh, C. Kamm, Y. Zaykov, A. Sellen, and S. E. Lindley (2016). "Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision." *undefined*.
- Schramowski, P., W. Stammer, S. Teso, A. Brugger, F. Herbert, X. Shao, H.-G. Luigs, A.-K. Mahlein, and K. Kersting (2020). "Making Deep Neural Networks Right for the Right Scientific Reasons by Interacting with Their Explanations." *Nature machine intelligence* 2 (8), 476–486.
- Shafiei Gol, E., M.-K. Stein, and M. Avital (2019). "Crowdwork Platform Governance toward Organizational Value Creation." *The Journal of Strategic Information Systems* 28 (2), 175–195.
- Shang, X., D. Di, J. Xiao, Y. Cao, X. Yang, and T.-S. Chua (2019). "Annotating Objects and Relations in User-Generated Videos." In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*. Ed. by A. El Saddik, A. Del Bimbo, Z. Zhang, A. Hauptmann, K. S. Candan, M. Bertini, L. Xie, and X.-Y. Wei. New York, NY, USA: ACM, pp. 279–287.
- Shen, T., J. Gao, A. Kar, and S. Fidler (2020). "Interactive Annotation of 3D Object Geometry Using 2D Scribbles." In: *Computer Vision – ECCV 2020*. Ed. by A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 751–767.
- Sheng, M., J. Dong, Y. Zhang, Y. Bu, A. Li, W. Lin, X. Li, and C. Xing (2020). "AHIAP: An Agile Medical Named Entity Recognition and Relation Extraction Framework Based on Active Learning." In: *Health*

- Information Science. Ed. by Z. Huang, S. Siuly, H. Wang, R. Zhou, and Y. Zhang. Vol. 12435. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 68–75.
- Shin, E., A. R. Khamesi, Z. Bahr, S. Silvestri, and D. A. Baker (2020). “A User-Centered Active Learning Approach for Appliance Recognition.” In: *2020 IEEE International Conference on Smart Computing (SMARTCOMP)*. IEEE, pp. 208–213.
- Shrivastava, A. and J. Heer (2020). “iSeqL: Interactive Sequence Learning.” In: *Proceedings of the 25th International Conference on Intelligent User Interfaces*. Ed. by F. Paternò, N. Oliver, C. Conati, L. D. Spano, and N. Tintarev. New York, NY, USA: ACM, pp. 43–54.
- Sivaraman, A., T. Zhang, G. van den Broeck, and M. Kim (2019). “Active Inductive Logic Programming for Code Search.” In: *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, pp. 292–303.
- Smith-Renner, A., R. Fan, M. Birchfield, T. Wu, J. Boyd-Graber, D. S. Weld, and L. Findlater (2020). “No Explainability without Accountability.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Ed. by R. Bernhaupt, F. ’. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguy, P. Bjørn, S. Zhao, B. P. Samson, and R. Kocielnik. New York, NY, USA: ACM, pp. 1–13.
- Soares Junior, A., C. Renso, and S. Matwin (2017). “ANALYTIC: An Active Learning System for Trajectory Classification.” *IEEE computer graphics and applications* 37 (5), 28–39.
- Song, M. (2020). “Personalized Image Classification by Semantic Embedding and Active Learning.” *Entropy (Basel, Switzerland)* 22 (11).
- (2021). “A Personalized Active Method for 3D Shape Classification.” *The Visual computer* 37 (3), 497–514.
- Song, M. and Z. Sun (2018). *Iterative Active Classification of Large Image Collection*.
- South, B. R., D. Mowery, Y. Suo, J. Leng, Ó. Ferrández, S. M. Meystre, and W. W. Chapman (2014). “Evaluating the Effects of Machine Pre-Annotation and an Interactive Annotation Interface on Manual de-Identification of Clinical Text.” *Journal of Biomedical Informatics* 50, 162–72.
- Souza, I. E. and A. X. Falcao (2020). “Learning CNN Filters From User-Drawn Image Markers for Coconut-Tree Image Classification.” *IEEE Geosci. Remote Sensing Lett.*, 1–5.
- Sperrle, F., R. Sevastjanova, R. Kehlbeck, and M. El-Assady (2019). “VIANA: Visual Interactive Annotation of Argumentation.” In: *2019 IEEE Conference on Visual Analytics Science and Technology: Proceedings : Vancouver, BC, Canada, 20-25 October 2019*. Ed. by R. Chang, D. Keim, and R. Maciejewski. Piscataway, NJ: IEEE, pp. 11–22.
- Sturm, T., J. Gerlach, L. Pumplun, N. Mesbah, F. Peters, C. Tauchert, N. Nan, and P. Buxmann (2021). “Coordinating Human and Machine Learning for Effective Organization Learning.” *Management Information Systems Quarterly* 45 (3), 1581–1602.
- Su, H., Z. Yin, S. Huh, T. Kanade, and J. Zhu (2016). “Interactive Cell Segmentation Based on Active and Semi-Supervised Learning.” *IEEE Transactions on Medical Imaging* 35 (3), 762–777.
- Suh, J., S. Ghorashi, G. Ramos, N.-C. Chen, S. Drucker, J. Verwey, and P. Simard (2019). “AnchorViz: Facilitating Semantic Data Exploration and Concept Discovery for Interactive Machine Learning.” *ACM Transactions on Interactive Intelligent Systems* 10 (1), 1–38.
- Sultanum, N., S. Ghorashi, C. Meek, and G. Ramos (2020). “A Teaching Language for Building Object Detection Models.” In: *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. Ed. by R. Wakkary, K. Andersen, W. Odom, A. Desjardins, and M. G. Petersen. New York, NY, USA: ACM, pp. 1223–1234.
- Tegen, A., P. Davidsson, R.-C. Mihailescu, and J. A. Persson (2019a). “Collaborative Sensing with Interactive Learning Using Dynamic Intelligent Virtual Sensors.” *Sensors (Basel, Switzerland)* 19 (3).
- Tegen, A., P. Davidsson, and J. A. Persson (2019b). “Interactive Machine Learning for the Internet of Things.” In: *Proceedings of the 9th International Conference on the Internet of Things*. New York, NY, USA: ACM, pp. 1–8.

- Tegen, A., P. Davidsson, and J. A. Persson (2020). "Activity Recognition through Interactive Machine Learning in a Dynamic Sensor Setting." *Personal and Ubiquitous Computing*.
- Theissler, A., A.-L. Kraft, M. Rudeck, and F. Erlenbusch (2020). "VIAL-AD: Visual Interactive Labelling for Anomaly Detection - An Approach and Open Research Questions." *CEUR Workshop Proceedings* 2660.
- Tian, Y., W. Liu, R. Xiao, F. Wen, and X. Tang (2007). "A Face Annotation Framework with Partial Clustering and Interactive Labeling." *2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2007), 18-23 June 2007, Minneapolis, Minnesota, USA*, 1–8.
- Trittenbach, H., A. Englhardt, and K. Böhm (2019). *Validating One-Class Active Learning with User Studies – A Prototype and Open Challenges*. Tech. rep. Karlsruhe.
- Trivedi, G. (2016). *On Interactive Machine Learning – Gaurav Trivedi*.
- van der Stappen, A. and M. Funk (2021). "Towards Guidelines for Designing Human-in-the-Loop Machine Training Interfaces." In: *26th International Conference on Intelligent User Interfaces*. Ed. by T. Hammond, K. Verbert, D. Parra, B. Knijnenburg, J. O'Donovan, and P. Teale. New York, NY, USA: ACM, pp. 514–519.
- Varga, V. and A. Lőrincz (2020). "Reducing Human Efforts in Video Segmentation Annotation with Reinforcement Learning." *Neurocomputing* 405, 247–258.
- Vargas Muñoz, J. E., D. Tuia, and A. X. Falcão (2020). "Deploying Machine Learning to Assist Digital Humanitarians: Making Image Annotation in OpenStreetMap More Efficient." *International Journal of Geographical Information Science*, 1–21.
- Vargas-Munoz, J. E., P. Zhou, A. X. Falcão, and D. Tuia (2019). "Interactive Coconut Tree Annotation Using Feature Space Projections." In: *IGARSS 2019 - 2019 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, pp. 5718–5721.
- Wall, E., S. Ghorashi, and G. Ramos (2019). "Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching." In: *Human-Computer Interaction – INTERACT 2019*. Ed. by D. Lamas, F. Loizides, L. Nacke, H. Petrie, M. Winckler, and P. Zaphiris. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 578–599.
- Wang, B., V. Wu, B. Wu, and K. Keutzer (2019). "LATTE: Accelerating LiDAR Point Cloud Annotation via Sensor Fusion, One-Click Annotation, and Tracking." In: *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*. IEEE, pp. 265–272.
- Wang, Y., P.-C. Liao, C. Zhang, Y. Ren, X. Sun, and P. Tang (2019). "Crowdsourced Reliable Labeling of Safety-Rule Violations on Images of Complex Construction Scenes for Advanced Vision-Based Workplace Safety." *Advanced Engineering Informatics* 42, 101001.
- Wang, Z., D. Acuna, H. Ling, A. Kar, and S. Fidler (2019). "Object Instance Annotation With Deep Extreme Level Set Evolution." In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 7492–7500.
- Webster, J. and R. T. Watson (2002). "Analyzing the Past to Prepare for the Future: Writing a Literature Review." *MIS Quarterly* 26 (2), xiii–xxiii.
- Westermann, H., J. Šavelka, V. R. Walker, K. D. Ashley, and K. Benyekhlef (2020). "Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents." In: *Legal Knowledge and Information Systems: JURIX 2020 : The Thirty-Third Annual Conference, BRNO, Czech Republic, December 9-11, 2020*. Ed. by S. Villata, J. Harašta, and P. Křemen. Vol. volume 334. Frontiers in Artificial Intelligence and Applications. Amsterdam: IOS Press.
- Wirth, F., J. Quehl, J. Ota, and C. Stiller (2019). "PointAtMe: Efficient 3D Point Cloud Labeling in Virtual Reality." In: *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1693–1698.
- Wissema, J. G. (1976). "Morphological Analysis: Its Application to a Company TF Investigation." *Futures* 8 (2), 146–153.
- Wolf, M., D. Ruiter, A. G. D'Sa, L. Reiners, J. Alexandersson, and D. Klakow (2020). "HUMAN: Hierarchical Universal Modular ANnotator." In: *Proceedings of the 2020 Conference on Empirical*

- Methods in Natural Language Processing: System Demonstrations*. Ed. by Q. Liu and D. Schlangen. Stroudsburg, PA, USA: Association for Computational Linguistics, pp. 55–61.
- Wolfswinkel, J. F., E. Furtmueller, and C. P. M. Wilderom (2013). “Using Grounded Theory as a Method for Rigorously Reviewing Literature.” *European Journal of Information Systems* 22 (1), 45–55.
- Wong, V. W. H., M. Ferguson, K. H. Law, and Y.-T. T. Lee (2019). “An Assistive Learning Workflow on Annotating Images for Object Detection.” In: *2019 IEEE International Conference on Big Data (Big Data)*. IEEE, pp. 1962–1970.
- Wottawa, J., M. Tahon, A. Marin, and N. Audibert (2020). “Towards Interactive Annotation for Hesitation in Conversational Speech.” *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings*.
- Wu, X., C. Chen, M. Zhong, and J. Wang (2021a). “HAL: Hybrid Active Learning for Efficient Labeling in Medical Domain.” *Neurocomputing* 456, 563–572.
- Wu, X., C. Chen, M. Zhong, J. Wang, and J. Shi (2021b). “COVID-AL: The Diagnosis of COVID-19 with Deep Active Learning.” *Medical Image Analysis* 68, 101913.
- Wu, Y., Y. Fang, S. Shang, J. Jin, L. Wei, and H. Wang (2021). “A Novel Framework for Detecting Social Bots with Deep Neural Networks and Active Learning.” *Knowledge-Based Systems* 211, 106525.
- Xu, X., D. Charatan, S. Raychaudhuri, H. Jiang, M. Heitmann, V. Kim, S. Chaudhuri, M. Savva, A. X. Chang, and D. Ritchie (2020). “Motion Annotation Programs: A Scalable Approach to Annotating Kinematic Articulations in Large 3D Shape Collections.” In: *2020 International Conference on 3D Vision (3DV)*. IEEE, pp. 613–622.
- Yang, Y., E. Kandogan, Y. Li, P. Sen, and and Walter S. Lasecki (2019). “A Study on Interaction in Human-In-The-Loop Machine Learning for Text Analytics.” In: *Joint Proceedings Of the ACM IUI2019Workshops, Los*.
- Ye, W., Y. Dong, and P. Peers (2019). “Interactive Curation of Datasets for Training and Refining Generative Models.” *Computer Graphics Forum* 38 (7), 369–380.
- Yimam, S. M., C. Biemann, L. Majnarić, Š. Šabanović, and A. Holzinger (2016). “An Adaptive Annotation Approach for Biomedical Entity and Relation Recognition.” *Brain Informatics* 3 (3), 157–168.
- Yimam, S. M., C. Biemann, L. Majnarić, S. Sabanovic, and A. Holzinger (2015). “Interactive and Iterative Annotation for Biomedical Entity Recognition.” *undefined*.
- Yimam, S. M., C. Biemann, Richard Eckart de Castilho, and I. Gurevych (2014). “Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno.” In: pp. 91–96.
- Yuan, J., X. Hou, Y. Xiao, D. Cao, W. Guan, and L. Nie (2019). “Multi-Criteria Active Deep Learning for Image Classification.” *Knowledge-Based Systems* 172, 86–94.
- Zankl, G., Y. Haxhimusa, and A. Ion (2012). “Interactive Labeling of Image Segmentation Hierarchies.” In: *Pattern Recognition*. Ed. by A. Pinz, T. Pock, H. Bischof, and F. Leberl. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 11–20.
- Zhang, J., H. Wang, S. Meng, and V. S. Sheng (2020). “Interactive Learning with Proactive Cognition Enhancement for Crowd Workers.” *AAAI* 34 (01), 540–547.
- Zhang, L., Y. Tong, and Q. Ji (2008a). “Interactive Labeling of Facial Action Units.” In: *2008 19th International Conference on Pattern Recognition*, pp. 1–4.
- Zhang, L., Y. Tong, and Q. Ji (2008b). “Active Image Labeling and Its Application to Facial Action Labeling.” In: vol. 5303, pp. 706–719.
- Zhang, T., A. El Ali, C. Wang, A. Hanjalic, and P. Cesar (2020). “RCEA: Real-time, Continuous Emotion Annotation for Collecting Precise Mobile Video Ground Truth Labels.” In: *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Ed. by R. Bernhaupt, F. ’. Mueller, D. Verweij, J. Andres, J. McGrenere, A. Cockburn, I. Avellino, A. Goguy, P. Bjørn, S. Zhao, B. P. Samson, and R. Kocielnik. New York, NY, USA: ACM, pp. 1–15.
- Zhang, X. and G. Nagy (2011). *The CADAL Calligraphic Database*, p. 37.

- Zhang, Y., Y. Wang, H. Zhang, B. Zhu, S. Chen, and D. Zhang (2022). "OneLabeler: A Flexible System for Building Data Labeling Tools." In: *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. CHI '22. New York, NY, USA: Association for Computing Machinery, pp. 1–22.
- Zhang, Y., A. Michi, J. Wagner, E. Andre, B. Schuller, and F. Weninger (2020). "A Generic Human-Machine Annotation Framework Based on Dynamic Cooperative Learning." *IEEE Transactions on Cybernetics* 50 (3), 1230–1239.
- Zhu, Y. and K. Yang (2019). "Tripartite Active Learning for Interactive Anomaly Discovery." *IEEE access : practical innovations, open solutions* 7, 63195–63203.
- Zimmer, W., A. Rangesh, and M. Trivedi (2019). "3D BAT: A Semi-Automatic, Web-based 3D Annotation Toolbox for Full-Surround, Multi-Modal Data Streams." In: *2019 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, pp. 1816–1821.
- Zwicky, F. (1967). "The Morphological Approach to Discovery, Invention, Research and Construction." In: *New Methods of Thought and Procedure*. Ed. by F. Zwicky and A. G. Wilson. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 273–297.