TOWARDS AN INTEGRATIVE THEORETICAL FRAMEWORK OF INTERACTIVE MACHINE LEARNING SYSTEMS

Miguel Angel Meza Martínez  
*Karlsruhe Institute of Technology (KIT)*, miguel.martinez@kit.edu

Mario Nadj  
*Karlsruhe Institute of Technology (KIT)*, mario.nadj@kit.edu

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

Follow this and additional works at: https://aisel.aisnet.org/ecis2019_rp

Recommended Citation  
TOWARDS AN INTEGRATIVE THEORETICAL FRAMEWORK OF INTERACTIVE MACHINE LEARNING SYSTEMS

Research paper

Meza Martínez, Miguel Angel, Karlsruhe Institute of Technology, Karlsruhe, Germany, miguel.martinez@kit.edu, SAP, Walldorf, Germany, miguel.angel.meza.martinez@sap.com
Nadj, Mario, Karlsruhe Institute of Technology, Karlsruhe, Germany, mario.nadj@kit.edu
Maedche, Alexander, Karlsruhe Institute of Technology, Karlsruhe, Germany, alexander.maedche@kit.edu

Abstract

Interactive machine learning (IML) is a learning process in which a user interacts with a system to iteratively define and optimise a model. Although recent years have illustrated the proliferation of IML systems in the fields of Human-Computer Interaction (HCI), Information Systems (IS), and Computer Science (CS), current research results are scattered leading to a lack of integration of existing work on IML. Furthermore, due to diverging functionalities and purposes IML systems can refer to, an uncertainty exists regarding the underlying distinct capabilities that constitute this class of systems. By reviewing extensive IML literature, this paper suggests an integrative theoretical framework for IML systems to address these current impediments. Reviewing 2,879 studies in leading journals and conferences during the years 1966-2018, we found an extensive range of applications areas that have implemented IML systems and the necessity to standardise the evaluation of those systems. Our framework offers an essential step to provide a theoretical foundation to integrate concepts and findings across different fields of research. The main contribution of this paper is organising and structuring the body of knowledge in IML for the advancement of the field. Furthermore, we suggest three opportunities for future IML research. From a practical point of view, our integrative theoretical framework can serve as a reference guide to inform the design and implementation of IML systems.

Keywords: Interactive Machine Learning, Interaction, Information Systems, Systematic Literature Review.

1 Introduction

Recent technological advances have led to an unprecedented ability to collect and store data (Woods et al., 2002). The amount of data exceeds the human capacity to digest what is meaningful and informative (Porter et al., 2013). Machine learning (ML) can help in these situations by providing the required capabilities to build systems that can support the clean-up, filtering and identification process of the most important subsets and patterns of the data (Gillies et al., 2016; Porter et al., 2013). One of the biggest ML advantages refers to its capability of decoding complex data relationships to model behaviours without the need for explicit programming (Dudley and Kristensson, 2018). Hereby, ML systems have broad coverage and have been successfully implemented in many fields, such as computer vision, speech recognition, or natural language processing (Anthes, 2017).

While ML performs very well in many contexts, as shown by the potential of self-driving cars (Holzinger, 2016), ML methods in complex contexts are at risk of lacking domain-specific user input (Porter et al., 2013). In general, there exist two options to address this problem. On the one hand, skilled practitioners analyse, translate and define ML systems based on what they have learned from users (who are experts for their domain) (Porter et al., 2013). However, following Amershi et al. (2014), this development approach is often characterised by a limited user engagement, leading to
lengthy and complex design iterations with a great dependency on the availability of skilled practitioners. Furthermore, resulting ML systems are typically considered a “black box” that might perform poorly for the intended purposes (Jiang et al., 2018) ending up in low levels of users’ trust into the system (Porter et al., 2013). On the other hand, the concept of interactive machine learning (IML) offers a promising way to address this problem by placing the user in the centre of the interaction with the ML system (Porter et al., 2013). The aim is to engage users directly and create a ML system that fits to the users’ goals and needs by building ML models iteratively through user input (Amershi et al., 2014). This approach enables users to review model outputs, make corrections by giving feedback and observe model changes and verify them (Gaurav, 2016). For instance, a prominent case stems from biomedicine – a context characterised by high dimensional, probabilistic and incomplete data (Holzinger, 2016). Physicians can make diagnoses with great reliability without being able to explicitly specify the underlying rules of their procedure. IML systems could help to equip algorithms with such “instinctive” knowledge and learn from it (Holzinger, 2016). The importance becomes evident when the use of automated solutions gets more difficult due to the incompleteness of ontologies (Atzmüller et al., 2006). Thus, in such problem spaces it seems beneficial to combine user interaction with ML. Up to now, numerous studies have already been published demanding assessment and structuration. Furthermore, as IML relies on specific techniques for achieving distinct capabilities, IML systems include system classes that vary strongly in their functionality and purpose (Amershi et al., 2014). However, there seems to be no established form to classify IML systems and due to the diverging purposes and functionality an uncertainty remains concerning the characteristics that constitute these systems (Jiang et al., 2018). In addition, IML terminology has been adopted by different fields, such as Human-Computer Interaction (HCI), Information Systems (IS) or Computer Science (CS), leading to a lack of integration of the present work on IML.

In this paper, we present the results of a systematic literature review (SLR) we conducted to tackle down the impediments in current research. Hereby, we formulated the following research question (RQ): What is the state-of-the-art in IML systems and how can the existing approaches and results be conceptualised in a unified way? Using an established method for a systematic literature review (SLR) (Kitchenham and Charters, 2007; Webster and Watson, 2002) research on IML was analysed with the objective to synthesise the resulting knowledge into an integrative theoretical framework (Baumeister and Leary, 1997). Based on the framework, we aim to identify gaps and outline future research avenues. Hereby, this paper makes three main contributions to the area of IML systems. First, creating an integrative theoretical framework is a foundational step towards a route map on how to conceptualise IML systems. Second, we integrate and structure the body of knowledge across disciplines (e.g., IS, CS, HCI) for the advancement of the field, which is specifically relevant due to the diverging nature of purposes and functionality of IML systems. Third, this systematic review identifies gaps in the literature and suggest further research avenues for scholars. For practice, our integrative theoretical framework can be used as a reference guide for designing and implementing IML systems.

The structure of this paper is as follows: The foundations of IML are shown in section 2. Section 3 describes the SLR method. Section 4 presents the results of the analysis. The discussion regarding the theoretical and practical implications is illustrated in section 5. Section 6 concludes the paper.

2 Conceptual Foundations

Studies on the interaction between users and ML systems have a long history in scientific literature relying on different means and interaction forms to incorporate domain-specific user input. IML has been built on the foundations of different learning algorithms with a Human-in-the-loop approach. Human-in-the-loop defines a technique to reduce the limitations of fully automated systems by engaging the user in an interactive process (Kim and Pardo, 2018). This technique has been successfully applied in diverse areas, such as image retrieval systems (Thomee and Lew, 2012) or sound event detection (Kim and Pardo, 2018) and is based on three underlying learning approaches: (1) Supervised Learning (SL), (2) Active Learning (AL) and (3) Reinforcement Learning (RL). Figure 1 depicts the relationships between the different approaches.
Applications in which the training data consist of examples of input elements along with their corresponding target output are known as SL (Bishop, 2006). Typically, humans play a crucial role on the labelling process to generate the target outputs. In turn, RL differs from SL in the absence of comprehensive labels of training examples; instead, the learner is guided towards a desired behaviour based on a (time-delayed) reward. There exist many applications in which the reward is obtained automatically or by implicit interaction with users (Taghipour et al., 2007). Research has examined the benefits of integrating users explicitly in the reward-loop interaction to steer up the learning algorithm (Knox and Stone, 2015). Hereby, plenty of areas have successfully integrated the user in RL systems; examples range from recommender systems (Gao et al., 2015) to exploratory search systems (e.g., Glowacka et al., 2013). Lastly, AL focuses on the system selecting unlabelled points in the dataset and querying the user for their labels (Settles, 2010). IML has built on this concept, but with the main difference that the selection is driven by the user rather than the learner (Dudley and Kristensson, 2018). Typical AL examples refer to handwritten text transcription (e.g., Serrano et al., 2010) or image retrieval (e.g., Tong and Chang, 2001). Research on IML combines strategies used on all the previous mentioned approaches to improve the iterative learning process with humans, however, the manifestations differs across studies.

3 Method

We conducted a SLR in the area of IML following the well-established research methods presented by Kitchenham and Charters (2007) and Webster and Watson (2002). The SLR was organised along three distinct stages (plan, conduct, and report; see Figure 2). During the plan stage, we recognised the need for a SLR, created a review protocol and assessed it. During the conduct stage, we performed the databases search, selected relevant studies, and reviewed them. Lastly, we described our results within the report stage.
Research questions. To keep our systematic review focused and to answer the overarching research question, we defined the following subordinated RQs: (1) How can the extent research on IML systems be conceptualised into an integrative theoretical framework? (2) What are predominant research areas in the IML domain? (3) What are potential directions for future research?

Search strategy. The search string was created in several steps. An explorative search with common literature databases (e.g., Google Scholar) was conducted using a search term which consisted of four parts: First, we extracted the term (1) “learn*” as a starting term from our research question. Second, we used the terms (2) “user” and (3) “system” to emphasise the importance of building ML systems for the intended user purposes. Third, as the main mean of IML refers to increasing the interaction of users with ML systems, we extracted the term (4) “interact*” as highly relevant for the initial search term. Finally, we used Boolean operators to create the initial search string: “learn AND user AND system AND interact*”.

During the initial exploration on IML literature, we identified nine highly relevant studies (Amershi et al., 2012, 2014; Dudley and Kristensson, 2018; Fails and Olsen, 2003; Fiebrink et al., 2011; Kabra et al., 2013; Porter et al., 2013; Talbot et al., 2009; Ware et al., 2001). After reviewing these studies, the search string was modified to include domain-specific keywords used in these studies. After several iterations, the final definition of our search string referred to “(reinforcement learning OR active learning OR machine learning) AND (user OR scientist* OR human-in-the-loop OR expert) AND (system* OR algorithm*) AND (interact*)”.

Next, we selected appropriate digital databases. Hereby, we compared Semantic Scholar, SCOPUS, ACM DL, WebofScience, EBSCO Discover service, EBSCOhost, IEEE, Springer Link, Emerald, JSTOR, ProQuest, ScienceDirect, and Wiley on how reliable they found the initially defined relevant literature. We finally selected the ACM Digital Library for our SLR as the database covered the most of the identified highly relevant studies (66.7%). In addition, the ACM Digital Library represents a well-established database used by scholars as a reliable source for SLR in IS (e.g., Bandara et al., 1999; Hamari et al., 2014).

Study selection criteria. To incorporate relevant studies, we carefully applied the following study selection criteria: (1) Only studies in which there was an explicit interaction with the user to achieve his or her goal were included, (2) studies in which the user did not iterative built and refine the model were excluded, (3) studies where only included when the user was the principle driver of the interaction to deliver desired behaviour in the ML system (as also suggested by Dudley and Kristensson, 2018), and (4) research in progress papers, doctoral and student consortiums, workshops and demonstration papers were excluded. In the study selection process, the selection criteria were applied to abstract, keyword, and title section excluding 2591 inappropriate studies. Next, the criteria were applied to full text, again excluding 218 studies. The references of the remaining 70 relevant studies were re-
viewed and three additional studies were incorporated. Finally, 73 studies were analysed, from which 67 were classified in our theoretical framework and six were only incorporated into our analysis since they focus on key areas of IML systems.

**Descriptive Analysis.** A wide range of application areas are covered by the analysed studies. In total, 38 different application areas were identified, whereby most of them referred to: media retrieval (13.7% of the studies; e.g., Fogarty et al., 2008; Luan et al., 2007), gesture recognition (10.9% of the studies; e.g., Fiebrink et al., 2011; Lü et al., 2014) and data analysis (6.8% of the studies; e.g., Sun et al., 2017; Talbot et al., 2009) (see Table 2 Appendix). In addition, most of the studies (83.5%) were derived from conferences. The most frequently represented conferences were the following: *International Conference on Intelligent User Interfaces* (19), *SIGCHI Conference on Human Factors in Computing Systems* (10) and *CHI Conference on Human Factors in Computing Systems* (4). Meanwhile only 16.4% of the studies were published in journals. In particular, *ACM Transactions on Interactive Intelligent Systems* was the most strongly represented with three studies. Regarding research methods, most studies (40) performed laboratory experiments, whereas only eight conducted case studies. Only three studies performed a review of IML applications (e.g., Amershi et al., 2014; Dudley and Kristensson, 2018; Porter et al., 2013). Lastly, during the analysis process, multiple ML fields were identified as the foundational backbone of IML systems illustrating the broad spectrum of related domains in which research has been conducted to combine user interaction with ML systems. From the reviewed studies, 50.6% were related to IML, 17.8% to AL and 10.9% to SL. The descriptive results are summarised in Figure 3.

**Figure 3. Descriptive Results.**

In addition, we identified six studies among the results that focus on key areas of IML systems, such as foundational principles for designing and interacting with IML systems. The corresponding studies are showcased in Table 1.

<table>
<thead>
<tr>
<th>Focus Areas</th>
<th>No. of Studies</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design</td>
<td>4</td>
<td>Dudley and Kristensson, 2018; Pei et al., 2017; Porter et al., 2013; Sarkar and Advait, 2016</td>
</tr>
<tr>
<td>Interaction</td>
<td>2</td>
<td>Amershi et al., 2014; Stumpf et al., 2007</td>
</tr>
</tbody>
</table>

**Table 1. Focus Areas of IML Systems.**
4 Results

In order to address our research question, we followed the suggestions by Baumeister and Leary (1997) and Kitchenham and Charters (2007) and consolidated the results of our SLR into an integrative theoretical framework. In the following, we describe the classification process of the framework before each (sub)dimension is explained along prominent publication results.

4.1 Classification Process

For the classification process, we applied a three-step approach to create the integrative theoretical framework. In a first step, categories for our framework were derived deductively (top-down) by focusing on the focus areas of IML systems. Thereafter, we leveraged the suggestions of Sacha et al. (2017) and Jiang et al. (2018) for structuring ML analysis tasks to derive the first four dimensions of the framework, namely (1) Interactive Classification, (2) Interactive Clustering, (3) Interactive Information Retrieval and (4) Interactive Regression. Hereby, (1) Interactive Classification refers to a ML classification task, where training data comprises examples of input elements along with their corresponding target output. The goal of the system is to assign new input elements into a finite number of discrete categories (Bishop, 2006). Many ML problems can be framed as a classification problem if the number of possible outputs can be represented by classes. In turn, (2) Interactive Clustering refers to the task of interactively partitioning instances in groups based on their similarity (Bishop, 2006). Hereby, the interaction of the user plays an important role since the concept of similarity is a subjective measure, thus, it is important to integrate the user in the refinement process of the model to define the similarity function. For (3) Interactive Information Retrieval, we considered interactive systems in which the main goals of the task is to provide a recommendation list of relevant information, to offer relevant elements in a search and to interactively showcase optimal solutions for decision-making (Jiang et al., 2018). Lastly, (4) Interactive Regression refers to tasks where the training data comprises examples of input elements with their corresponding target output. However, in contrast to classification tasks, the target output is represented by a continuous variable (Bishop, 2006). Thus, the system makes a prediction of continuous values by modelling the relations between independent and dependent variables (Jiang et al., 2018).

In the second step, we applied inductive reasoning (bottom-up) to analyse, whether all studies could be classified under the four previously defined dimensions. It became evident that changes on the initial framework were required to typecast all the different IML systems within the reviewed literature. Leveraging on the work from Krening and Feigh (2018), we introduced a new dimension entitled (5) Teaching Intelligent Agents. It considers applications in which the user has been incorporated to interact with intelligent agents and robots to train them for specific tasks or to personalise their behaviour to the user’s preference. In addition, modifications were made to the definition of subdimensions that were proposed by Jian et. al (2018). In particular, we added a new subdimension entitled Interactive Model Selection for Interactive Classification to address tasks in which the user interactively creates a combination of multiple models. Furthermore, we incorporated the concept of Constraint Clustering as a subdimension for Interactive Clustering to incorporate applications in which clustering is performed through constraints on similarity instead of exploration. Lastly, we needed to derive the subdimensions for the recently added dimension of Teaching Intelligent Agents. Hereby, four different subdimensions were created: Interactive Personalisation was introduced due to systems where the agent adapts to the needs and preferences of an individual (Clabaugh, 2017); Interactive Training was selected to represent a task in which the user performs training of robots; Preference Elicitation covered the aspect of applications in which a user indicates a preference between multiple options and the system is able to learn and adapt his or her behaviour; Programming-by-demonstration was derived to address tasks where the user interactively demonstrates his or her objective by executing an action that the agent needs to learn in order to execute it later.

Finally, during the third step, we performed a revision and integration of the dimensions and subdimensions and proceeded to classify all studies. Overall, five dimensions and 15 subdimensions were defined. The representation of the integrative theoretical framework together with the statistics of the
classification distribution are illustrated in Figure 4. We moved the coding schema of the integrative theoretical framework with a detailed view of all studies into the appendix (see Table 3 Appendix).

<table>
<thead>
<tr>
<th>Interactive</th>
<th>Interactive Clustering</th>
<th>Interactive Information</th>
<th>Interactive</th>
<th>Teaching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Engineering</td>
<td>Classification</td>
<td>Retrieval</td>
<td>Regression</td>
<td>Intelligent Agents</td>
</tr>
<tr>
<td>Interactive Feature Engineering</td>
<td>Interactive Labelling</td>
<td>Interactive Cluster-based Exploratory Data Analysis</td>
<td>Comparative Cluster Analysis</td>
<td>Constraint Clustering</td>
</tr>
<tr>
<td>Interactive Model Selection</td>
<td>Parameter Spaces Analysis</td>
<td>Data Analysis</td>
<td>Interactive Information Retrieval</td>
<td>Interactive Recommendation</td>
</tr>
<tr>
<td>Parameter Tuning</td>
<td>Interactive Retrieval Analysis</td>
<td>Interactive Model Optimisation</td>
<td>Interactive Selection</td>
<td>Model Optimisation</td>
</tr>
<tr>
<td>Interactive Labelling</td>
<td>Interactive Recommendation</td>
<td>Interactive Personalisation</td>
<td>Interactive Training</td>
<td>Interactive Preference Elicitation</td>
</tr>
<tr>
<td>Interactive Model Selection</td>
<td>Interactive Personalisation</td>
<td>Interactive Programming by Demonstration</td>
<td>Interactive Training</td>
<td>Preference Elicitation</td>
</tr>
<tr>
<td>Interactive Feature Engineering</td>
<td>Interactive浪</td>
<td>Interactive Personalisation</td>
<td>Interactive Programming by Demonstration</td>
<td>Interactive Programming by Demonstration</td>
</tr>
<tr>
<td>Interactive Feature Engineering</td>
<td>Interactive Programming by Demonstration</td>
<td>Interactive Programming by Demonstration</td>
<td>Interactive Programming by Demonstration</td>
<td>Interactive Programming by Demonstration</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>44.8%</td>
<td>13.4%</td>
<td>32.8%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>77.8%</td>
<td>11.1%</td>
<td>72.7%</td>
</tr>
<tr>
<td>18.2%</td>
<td>9.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>20.0%</td>
<td>20.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>20.0%</td>
<td>73.3%</td>
<td>3.3%</td>
</tr>
<tr>
<td>3.3%</td>
<td>77.8%</td>
<td>11.1%</td>
</tr>
<tr>
<td>72.7%</td>
<td>18.2%</td>
<td>9.1%</td>
</tr>
<tr>
<td>100.0%</td>
<td>20.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>44.8%</td>
<td>13.4%</td>
<td>32.8%</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>77.8%</td>
<td>11.1%</td>
<td>72.7%</td>
</tr>
<tr>
<td>18.2%</td>
<td>9.1%</td>
<td>100.0%</td>
</tr>
<tr>
<td>20.0%</td>
<td>20.0%</td>
<td>40.0%</td>
</tr>
<tr>
<td>20.0%</td>
<td>73.3%</td>
<td>3.3%</td>
</tr>
<tr>
<td>3.3%</td>
<td>77.8%</td>
<td>11.1%</td>
</tr>
<tr>
<td>72.7%</td>
<td>18.2%</td>
<td>9.1%</td>
</tr>
<tr>
<td>100.0%</td>
<td>20.0%</td>
<td>40.0%</td>
</tr>
</tbody>
</table>

| A: Layer percentage - B: Total number of coverage - C: Coverage percentage |

**Figure 4. Integrative Theoretical Framework of Interactive Machine Learning Systems.**

## 4.2 Integrative Theoretical Framework

Following our classification process, this section explains the integrative theoretical framework along each dimension as well as emphasises the number of classified studies and prominent research results.

### 4.2.1 Interactive Classification

30 of the identified studies explicitly address Interactive Classification, resulting in the highest value in our analysis. Interactive Classification has been applied to a wide variety of applications, such as gesture recognition (e.g., Sarasua et al., 2016) or sentiment analysis (e.g., Huang et al., 2013). Quality of labels, features, parameter tuning, as well as model selection are key design factors for building a classifier with a high performance. Based on the classifier’s focus on one of these factors, we have extended the categories proposed by Jiang (2018) and classified the reviewed studies along those sub-dimensions. Most studies focus on interactive labelling (22 studies), followed by interactive feature engineering (6), interactive model selection (1) and parameter space analysis (1).

**Feature engineering** refers to the selection of appropriate input variables that can help to improve the model’s performance. Recent work has focused on eliciting the user’s domain knowledge on feature relevance by supporting the user to understand the correlation between features and classes (Micaleff et al., 2017). In turn, in **interactive labelling**, the system interacts with the user to generate the target labels of relevant instances to refine the underlying model. For instance, Wallace (2012) implements an interactive classification system to find relevant documents in a citation screening process. **Model selection** focuses on allowing the user to indicate his or her model preferences; thereafter those are incorporated into the classifier. Talbot et al. (2009) developed an interactive classification system that allows the user to build an ensemble model that best adapt to his or her objective model by combining a set of predefined ML algorithms. Lastly, in **parameter space analysis**, the users can navigate and explore the parameter space by providing preferences on the tolerance of misclassification errors.

### 4.2.2 Interactive Clustering

We found nine studies that focus on the area of Interactive Clustering, which represent the third ranked representation in our analysis. Interactive Clustering has been successfully implemented in
many applications as a tool to help the user to explore the elements of a dataset, as exemplified by AnchorViz (Chen et al., 2018) in an application that facilitates the identification of error discovery in data exploration. From the analysis of relevant studies, it was identified that most of applications performing an interactive clustering task focused on cluster-based exploratory data analysis (7), follow by comparative cluster analysis (1) and constraint clustering (1).

Cluster-based exploratory data analysis refers to tasks in which the user’s goal is to explore elements in order to understand the relationships that exist among them. Recent work has implemented interactive systems in which the user creates groups of similar topics (Nourashrafeddin et al., 2013), provides his or her preferences for the distance metric learning in ontology constructions (Yang and Callan, 2008) or interactively groups search results (Chang et al., 2016). Constraint clustering considers an interesting approach to find structures in the dataset by focusing on the use of constraints for the similarity between elements. An example of research in this area, focuses on creating interactive clusters by using pairwise constraints (e.g., Okabe and Yamada, 2012). Finally, cluster comparison refers to using multiple cluster methods to help the user visualise different structure possibilities. CommunityDiff (e.g., Datta and Adar, 2018) demonstrates this, by allowing users to conduct cluster comparisons of different ensembles.

4.2.3 Interactive Information Retrieval

Interactive Information Retrieval represents the second most relevant area of research with 22 studies. The concept has been successfully implemented in many applications, such as urban design (e.g., Chirkin and König, 2016) or media retrieval (e.g., Ayache et al., 2010; Fogarty et al., 2008; Glowacka et al., 2013). We subclassified the dimension of Interactive Information Retrieval system along information retrieval (16), interactive recommendation (4) and model optimisation (2).

In return retrieval, the user can refine the system results to a query by providing feedback on the return elements’ relevance. Thus, the system can interactively learn the user’s objectives and improve the corresponding search results. Multiple information retrieval applications have been conducted within a wide range of elements that can be processed by the system. Some examples include retrieval of images (e.g., Fogarty et al., 2008; Sato et al., 2018), retrieval of documents (e.g., Glowacka et al., 2013), or retrieval of videos (e.g., Haas et al., 2004). Interactive recommendation systems can learn the user’s preference on a specific domain to provide the appropriate suggestions. Recently, research has focused on providing recommendations for stock transactions (e.g., Yoo et al., 2003), scheduling of meeting (e.g., Kozierek and Maes, 1993; Weber and Pollack, 2007) and members for group membership (e.g., Amershi et al., 2012). Model optimisation refers to tasks in which the user collaborates with an interactive system to provide his or her preference in an optimisation problem context and where the system can return suggestion of optimised models that might be relevant considering the user’s objective. For example, Brochu et al. (2010) developed an interactive system, where the user provides examples of animations through manipulation of parameters and the system performs an optimisation to return a list of animations that might be relevant for the user.

4.2.4 Interactive Regression

Only one of the studies explicitly address working with Interactive Regression tasks, which represents the lowest value in our analysis. Interactive Regression considers task in which the objective of the user is to perform a prediction on a continuous variable, instead of finite set of classes as it is done in classification. Working with continuous variables extends the possibilities of application areas for IML because continuous variables are present in many natural sciences and engineering applications.

Interactive numerical prediction was defined as the only subdimension for Interactive Regression which represents all tasks in which the user has the goal to obtain a numerical prediction from the system. An example of an Interactive Regression is presented by Daee et al. (2018) in a system that performs predictions in a sentiment analysis task.
4.2.5 Teaching Intelligent Agents

As for Teaching Intelligent Agents, five of the studies address this area, representing the second-lowest value in our analysis. As mentioned previously, Teaching Intelligent Agents refers to task in which the user interacts with agents or robots to train them for specific tasks or to personalise their behaviour. The number of studies classified for each subdimension are: interactive personalisation (1), interactive training (2), preference elicitation (1) and programming-by-demonstration (1).

**Interactive personalisation** is defined by Clabaugh (2017) as the process by which an intelligent agent adapts to the user’s needs and preferences through eliciting information directly from the user about his or her states. In this light, Clabaugh (2017) analyses a set of social and computational trade-offs for the elicitation process. For **preference elicitation**, an agent assists the user by adaptively and interactively learning about the underlying preference model on basis of user feedback. Lee et al. (2004) successfully implemented an interactive system that learns from the user’s preferences in the selection of wireless services. In turn, the objective of **interactive training** systems refers to enabling users to personalise the behaviour of an intelligent agent. For **programming-by-demonstration**, the user generates training demonstrations for activities he or she is interested in to teach the intelligent agent. The agent is capable to observe and learn from such demonstrations, while the user may refine the execution by providing feedback or correcting errors (Berthouzoz et al., 2011). For instance, Berthouzoz (2011) implemented a programming-by-demonstration system to train the system to replicate complex photo manipulations.

5 Future Work

In this section, we offer an overview on our insights, highlight emerging research gaps and propose suggestions on suitable research methods to answer them. Our integrative theoretical framework represents a valuable baseline for future studies as it offers a reference of the state-of-the-art and trends within the literature. In reverse, this also indicates what has not yet been addressed and accordingly facilitates the disclosure of future research avenues. Using our framework, we were able to identify several research gaps within the existing body of knowledge. Thus, we would like to call the scholars’ attention to the following three research avenues as a promising starting point:

**Understand the Impact of Interaction on User’s Level of Trust.** So far, extensive research has been conducted on exploring new application scenarios that could take advantage of the interaction between users and ML systems to personalise the applications and enable users to achieve their particular goals (e.g., Chirkin and König, 2016; Kabra et al., 2013; Katan et al., 2015). Research has also proposed to improve the interaction with IML systems by providing new functionality or visualisation techniques that support users to easily refine their model and better understand the system’s behaviour (e.g., Amershi, 2011; Kulesza et al., 2012; Stumpf et al., 2007). However, studies have predominantly addressed the facilitation of user’s interaction, while typically neglecting the impact that such IML interaction approaches have on outcome variables such as the user’s level of trust. As ML systems are typically considered a “black box” that might perform poorly for the intended purposes (Jiang et al., 2018), users’ trust into an IML system seems essential (Porter et al., 2013). In our analysis, the notion to evaluate the impact on user’s experience, such as trust, has also been confirmed by several studies (e.g., Ankerst et al., 2000; Dudley and Kristensson, 2018; Krening and Feigh, 2018; Kulesza et al., 2015; Yang et al., 2018). Thus, future research should evaluate the different interaction forms of IML systems along the user’s level of trust. Conducting comprehensive lab experiments, where different interaction forms could serve as treatments, may represent an adequate starting point to address this research gap.

**Develop an Approach for Standardising Evaluation Methodologies.** The development of a holistic approach for a standardised evaluation methodology for IML systems is a second relevant venue of future work. Such standard would also need to integrate the evaluation of user’s affective-cognitive states and behaviours based on previously well-defined constructs. Hereby, the large body of knowledge from HCI and behavioural science could be made accessible for IML research. Some of the
factors that could be integrated in this holistic evaluation approach could include constructs, such as trust, frustration, satisfaction, flow or users’ mental models. Regarding the evaluation of the system characteristics, some factors that could be considered refer to immediacy, complexity, robustness, flexibility and deterministic interaction. Thus, future research could take these elements into account in meeting the challenge of designing a standardised evaluation methodology for IML systems. Conducting design science research might help to address the identified issues (Peffers et al., 2007).

**Investigate the Problem of Overfitting within IML Systems.** The concept of overfitting relates to a ML algorithm, which is learning a model, that corresponds too closely to the training data. Such models do not generalise well with new data and typically more-effective alternative models are available (Mitchell, 1997). In IML systems, the problem of overfitting has particular relevance as the user is the main driver of the model refinement process; however, for many applications the user might not be aware of the overfitting problem. In this line, Kapoor et al. (2010) identify the risk of overfitting and emphasise the possibility to rely on traditional ML techniques. In addition, Dudley and Kristensson (2018) highlight the importance of helping the users in selecting appropriate strategies to avoid overfitting. However, only the study by Daee et al. (2018) analysed the problem of overfitting in detail. They successfully identified scenarios that can lead to overfitting; but merely within the context of knowledge elicitation. Nevertheless, the interaction between the user and the ML system varies strongly among different IML applications; thus, it is of vital importance to investigate how the problem of overfitting affects the different interaction paradigms to derive strategies to minimise this risk.

6 Conclusion

In our paper, we offered a comprehensive overview of existing IML research by developing an integrative theoretical framework to unify multiple IML concepts and to classify studies in a coherent way. Our integrative theoretical framework serves as a foundational step towards establishing a road map on how to conceptualise IML systems. Furthermore, we reviewed literature on IML across disciplines (e.g., IS, CS, HCI) for the advancement of the field enabling us to recognise unfolding research opportunities for IML scholars. In particular, we identified gaps in the literature and suggested further research avenues. For practice, our proposed framework could help to gain a state-of-the-art overview of IML systems. Such overview could be used as a guide for practitioners to identify appropriate focus areas for designing and implementing IML systems. In addition, our framework enables a benchmark comparison of practical IML solutions against other systems that have been classified in our framework according to their ML task. For example, a practitioner working in the development of interactive clustering interfaces could use our integrative theoretical framework to find IML systems that incorporate personal preferences for the distance metric used in the clustering algorithm and then perform a comparative analysis with the results that have been published for that system. Still, we are aware that our study has some limitations. Any bias within the search string selection might result in a bias of the reviewed literature. However, our SLR followed the methodological recommendations defined in the literature to reduce this probability (Webster and Watson, 2002; Kitchenham and Charters, 2007). Furthermore, we relied only on one database (ACM Digital Library) for our SLR as it covered most of the previously identified highly relevant literature (see also Search Strategy). We invite future research to expand our analysis with other established databases by using our derived integrative theoretical framework as a basis for the classification of IML systems. We hope that this framework can serve as a reference for scholars and practitioners in the broader field of IML systems and the aspects that should be addressed when investigating or designing related solutions.
# Appendix

## Table 2. Distribution of Application Areas.

<table>
<thead>
<tr>
<th>Application Area</th>
<th>#</th>
<th>%</th>
<th>Application Area</th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media Retrieval</td>
<td>10</td>
<td>13.70%</td>
<td>Communication Networks</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Gesture Recognition</td>
<td>8</td>
<td>10.96%</td>
<td>Communities Visualisation</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Conceptual Analysis IML</td>
<td>5</td>
<td>6.85%</td>
<td>Error Discovery</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Data Analysis</td>
<td>5</td>
<td>6.85%</td>
<td>Exploratory Search</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Image Processing</td>
<td>3</td>
<td>4.11%</td>
<td>Human Activity Recognition</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Text Classification</td>
<td>3</td>
<td>4.11%</td>
<td>Human Agent Interaction</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Audio Edition</td>
<td>2</td>
<td>2.74%</td>
<td>Insurance Claims</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Human Robot Interaction</td>
<td>2</td>
<td>2.74%</td>
<td>Interactive Decision Tree Classifier</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Ontology Construction</td>
<td>2</td>
<td>2.74%</td>
<td>Microarray Analysis</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Scheduling Application</td>
<td>2</td>
<td>2.74%</td>
<td>Multimedia Retrieval</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Sentiment Analysis</td>
<td>2</td>
<td>2.74%</td>
<td>Prediction of Citation Counts</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Sound Event Detector</td>
<td>2</td>
<td>2.74%</td>
<td>Search Result Clustering</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Video Retrieval</td>
<td>2</td>
<td>2.74%</td>
<td>Social Network Group Creation</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Accessible Interfaces</td>
<td>1</td>
<td>1.37%</td>
<td>Star Coordinates</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Activity Recognition</td>
<td>1</td>
<td>1.37%</td>
<td>Stock Transaction Recommender</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Alarm Triage Classification</td>
<td>1</td>
<td>1.37%</td>
<td>Topic Modelling</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Animation Design</td>
<td>1</td>
<td>1.37%</td>
<td>Urban Design</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Annotation of Animal Behaviour</td>
<td>1</td>
<td>1.37%</td>
<td>User Experience</td>
<td>1</td>
<td>1.37%</td>
</tr>
<tr>
<td>Citation Screening</td>
<td>1</td>
<td>1.37%</td>
<td>User Feedback</td>
<td>1</td>
<td>1.37%</td>
</tr>
</tbody>
</table>

# = Number of studies for a specific application area  
% = Number of studies for a specific application relative to the total amount of studies in percent
<table>
<thead>
<tr>
<th>Dimension</th>
<th>Subdimension</th>
<th>Studies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Interactive Classification (30)</td>
<td>Interactive Feature Engineering (6)</td>
<td>Ankerst et al., 2000; Bauer and Baldes, 2005; Huang et al., 2013; Kulesza et al., 2015; Micallef et al., 2017; Di Nunzio and Maria, 2016</td>
</tr>
<tr>
<td></td>
<td>Interactive Labelling (22)</td>
<td>Amershi et al., 2011; Billewicz and Agnieszka, 2018; Brenton et al., 2014; Bryan et al., 2014; Dey et al., 2014; Fails and Olsen, 2003; Fiebrink et al., 2011; Flutura et al., 2018; Françoise et al., 2016; Françoise and Bevilaqua, 2018; Ghani and Kumar, 2011; Gillies, Brenton, et al., 2015; Gillies, Kleinsmith, et al., 2015; Hipke et al., 2014; Kabra et al., 2013; Katan et al., 2015; Kim and Pardo, 2017; Lü et al., 2014; Sarasua et al., 2016; Sun et al., 2017; Wallace et al., 2012; Wu and Yang, 2006</td>
</tr>
<tr>
<td></td>
<td>Interactive Model Selection (1)</td>
<td>Talbot et al., 2009</td>
</tr>
<tr>
<td></td>
<td>Parameter Spaces Analysis (1)</td>
<td>Kapoor et al., 2010</td>
</tr>
<tr>
<td>Interactive Clustering (9)</td>
<td>Cluster-based Exploratory Data Analysis (7)</td>
<td>Awasthi et al., 2017; Chang et al., 2016; Chen et al., 2018; Chidlovskii and Lecerf, 2008; Nourashrafeddin et al., 2013; Smith et al., 2018; Yang and Callan, 2008</td>
</tr>
<tr>
<td></td>
<td>Comparative Cluster Analysis (1)</td>
<td>Datta and Adar, 2018</td>
</tr>
<tr>
<td></td>
<td>Constraint Clustering (1)</td>
<td>Okabe and Yamada, 2009</td>
</tr>
<tr>
<td>Interactive Information Retrieval (22)</td>
<td>Information Retrieval (16)</td>
<td>Amershi et al., 2009, 2010; Ayache et al., 2010; Fogarty et al., 2008; Gao et al., 2015; Glowacka et al., 2013; Gony et al., 2007; Haas et al., 2004; Keyvanpour and Asbaghi, 2008; Kim and Pardo, 2018; Lu et al., 2007; Luan, Neo, Chua, et al., 2007; Luan, Neo, Goh, et al., 2007; Nguyen et al., 2008; Sato et al., 2018; Shearin and Lieberman, 2001</td>
</tr>
<tr>
<td></td>
<td>Interactive Recommendation (4)</td>
<td>Amershi et al., 2012; Kozierek and Maes, 1993; Weber and Pollack, 2007; Yoo et al., 2003</td>
</tr>
<tr>
<td></td>
<td>Model Optimisation (2)</td>
<td>Brochu et al., 2010; Chirkin and König, 2016</td>
</tr>
<tr>
<td>Interactive Regression (1)</td>
<td>Interactive Numerical Prediction (1)</td>
<td>Daee et al., 2018</td>
</tr>
<tr>
<td>Teaching Intelligent Agents (5)</td>
<td>Interactive Personalisation (1)</td>
<td>Clabaugh, 2017</td>
</tr>
<tr>
<td></td>
<td>Interactive Training (2)</td>
<td>Krening and Feigh, 2018; Senft et al., 2017</td>
</tr>
<tr>
<td></td>
<td>Preference Elicitation (1)</td>
<td>Lee et al., 2004</td>
</tr>
<tr>
<td></td>
<td>Programming by Demonstration (1)</td>
<td>Berthouzoz et al., 2011</td>
</tr>
</tbody>
</table>

Table 3. Classification of Studies within the Integrative Theoretical Framework of Interactive Machine Learning Systems.
References


Settles, B. (2010), *Active Learning Literature Survey*. 


Advice”, Proceedings of the 8th International Conference on Intelligent User Interfaces - IUI ’03, ACM, New York, New York, USA, p. 197.