Evaluation of Interactive Machine Learning Systems

Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton

Abstract The evaluation of interactive machine learning systems remains a difficult task. These systems learn from and adapt to the human, but at the same time, the human receives feedback and adapts to the system. Getting a clear understanding of these subtle mechanisms of co-operation and co-adaptation is challenging. In this chapter, we report on our experience in designing and evaluating various interactive machine learning applications from different domains. We argue for coupling two types of validation: *algorithm-centered* analysis, to study the computational behaviour of the system; and *human-centered* evaluation, to observe the utility and effectiveness of the application for end-users. We use a visual analytics application for guided search, built using an interactive evolutionary approach, as an exemplar of our work. We argue that human-centered design and evaluation complement algorithmic analysis, and can play an important role in addressing the "black-box" effect of machine learning. Finally, we discuss research opportunities that require human-computer interaction methodologies, in order to support both the visible and hidden roles that humans play in interactive machine learning.

1 Introduction

In interactive Machine Learning (iML), a human operator and a machine collaborate to achieve a task, whether this is to classify or cluster a set of data points [1, 11], to find interesting data projections [5, 6, 47], or to design creative art works [36, 43]. The underlying assumption is that the human-machine co-operation yields better results than a fully automated or manual system. An interactive machine learning

N. Boukhelifa, E. Lutton
INRA, Université Paris-Saclay, 1 av. Brétignières, 78850, Thiverval-Grignon, France, e-mail: {nadia.boukhelifa, evelyne.lutton}@inra.fr
A. Bezerianos
Univ Paris-Sud &CNRS (LRI), INRIA, Université Paris-Saclay, France

system comprises an automated service, a user interface, and a learning component. The human interacts with the automated component via the user interface, and provides iterative feedback to a learning algorithm. This feedback may be explicit, or inferred from human behaviour and interactions. Likewise, the system may provide implicit or explicit feedback to communicate its status and the knowledge it has learnt.

The interactive approach to machine learning is appealing for many reasons including:

- to integrate valuable experts knowledge that may be hard to encode directly into mathematical or computational models.
- to help resolve existing uncertainties as a result of, for example, bias and error that may arise from automatic machine learning.
- to build trust by making humans involved in the modelling or learning processes.
- to cater for individual human differences and subjective assessments such as in art and creative applications.

Recent work in interactive machine learning has focused on developing working prototypes, but less on methods to evaluate iML systems and their various components. The question of how to effectively evaluate such systems is challenging. Indeed, human-in-the-loop approaches to machine learning bring forth not only numerous intelligibility and usability issues, but also open questions with respect to the evaluation of the various facets of the iML system, both as separate components and as a holistic entity [40]. Holzinger [27] argued that conducting methodically correct experiments and evaluations is difficult, time-consuming, and hard to replicate due to the subjective nature of the "human agents" involved. Cortellessa and Cesta [18] found that the quantitative evaluation of mixed-initiative systems tend to focus either on problem solving performance of the human and what they call the artificial solver, or the quality of interaction looking at user requirements and judgment of the system. This statement also applies to iML systems, where current evaluations tend to be either *algorithm-centered* to study the computational behaviour of the system, or *human-centered* focusing on the utility and effectiveness of the application for end-users [6, 8, 7].

The aim of this chapter is to review existing evaluation methods for iML systems, and to reflect upon our own experience in designing and evaluating such applications over a number of years [3, 10, 32, 34, 36, 48, 49]. The chapter is organised as follows: First we provide a review of recent work on the evaluation of iML systems focusing on types of human and system feedback, and the evaluation methods and metrics deployed in these studies. We then illustrate our evaluation method through a case study on an interactive machine learning system for guided visual search, covering both algorithm-centered and human-centered evaluations. Finally, we discuss research opportunities requiring human-computer interaction methodologies in order to support both the visible and hidden roles that humans play in machine learning.

2 Related Work

In this section, we review recent work that evaluates interactive machine learning systems. We consider both qualitative and quantitative evaluations. Our aim is not to provide an exhaustive survey, but rather to illustrate the broad range of existing methods and evaluation metrics.

Paper	Implicit User Feedback	Explicit User Feedback	System Feedback	Case Study	User Study	Observational Study	Survey	Objective metrics	Subjective metrics
3D Model Repository Explorator [22]		✓		✓	✓			✓	
Co-integration [2]	✓				✓			✓	
EvoGraphDice [6]	✓	✓	✓		✓			✓	
EvoGraphDice [10]	✓	✓	✓			✓	✓	✓	✓
DDLite [20]		✓	✓	✓				✓	✓
Dis-Function [11]		✓	✓		✓			✓	✓
ForceSPIRE [21]	✓			✓				✓	
ForceSPIRE [38]	✓					✓		✓	✓
Interest Driven Navigation [25]	✓	✓	✓	✓					✓
ISSE [13]		✓			✓		✓	✓	✓
OLI [50]	✓			✓				✓	
RCLens [35]		✓	✓	✓			✓	✓	✓
ReGroup [1]	✓				✓		✓	✓	✓
RugbyVAST [33]		✓	✓	✓	✓			✓	✓
SelPh [31]	✓	✓	✓		✓		✓	✓	✓
User Interaction Model [19]	✓		✓		✓		✓	✓	✓
UTOPIAN [17]	✓		✓	✓				✓	
View Space Explorer [5]		✓	✓	✓				✓	
Visual Classifier [26]		✓	✓		✓		✓	✓	✓

Table 1 iML system evaluations of reviewed papers. We studied these systems in terms of the type of human feedback they support (implicit, explicit, or mixed), the evaluation methods used (case study, user study, observational study, survey), and whether they used objective or subjective evaluation metrics. Note that mixed user feedback combines implicit and explicit mechanisms.¡UPDATE CAPTION¿

2.1 Method

We systematically reviewed papers published between 2012-2017 from the following venues: IEEE VIS, ACM CHI, EG EuroVis, HILDA workshop, and CHI HCML workshop. We downloaded then filtered the proceedings to include papers having

the following keywords: "learn AND algorithm AND interact AND (user OR human OR expert) AND (evaluation OR study OR experiment)". We then drilled down to find papers that describe an actual iML system (as defined in the introduction) with an evaluation section. In this chapter, we focus on studies from the fields of visualization and human-computer interaction. Our hypothesis was that papers from these domains are likely to go beyond algorithm-centered evaluations. In total, we reviewed 19 recent papers (see Table 1), from various application domains including multidimensional data exploration [5, 6, 19, 25, 50], data integration [2], knowledge base construction [20], text document retrieval [26], photo enhancement [31], audio source separation [13], social network access control [1], and category exploration and refinement [35]. We examined these evaluations in terms of the types of user feedback, the nature of system feedback, and their evaluation methods and metrics.

2.2 Human Feedback

Broadly speaking, human feedback to machine learning algorithms can be either *explicit* or *implicit*. The difference between these two mechanisms stems from the field of Information Retrieval (IR). In the case of implicit feedback, humans do not assess relevance for the benefit of the IR system, but rather to fulfill their own task. Besides, they are not necessarily aware that their assessment is being used for relevance feedback [30]. In contrast, for explicit feedback, humans indicate their assessment via a suitable interface, and are aware that their feedback is interpreted for relevance judgment. Whereas implicit feedback is *inferred* from human interactions with the system, explicit feedback is directly provided by humans.

The systems we reviewed either use implicit (7 papers), explicit (8 papers), or mixed (4 papers) human feedback. In the case of mixed feedback, the system tries to infer information from user interactions to complement the explicit feedback.

Implicit Human Feedback

Endert et al. [21, 38] developed *semantic interaction* for visual analytics where the analytical reasoning of the user is inferred from their interactions, which in turn helps steer a dimension reduction model. Their system ForceSpire learns from human input (e.g. moving objects) to improve an underlying model and to produce an improved layout for text documents. Similarly, UTOPIAN [17] supports what the authors describe as a "semantically meaningful set of user interactions" to improve topic modelling. These interactions include keyword refinement, and topic splitting and merging. Implicit feedback may also be gathered from user interactions with raw data. For example, Azuan et al. [2] developed a tool where manual data corrections, such as adding or removing tuples from a data table, are leveraged to improve data integration and cleaning.

Interactive machine learning systems may infer other types of information such as attribute salience or class membership. In [50], the iML system infers the importance of data attributes from user manipulations of nodes and clusters in order

to improve a layout algorithm. The ReGroup tool [1] learns from user interactions and faceted search on online social networks to create custom on-demand groups of actors in the network.

In the previous examples, the system learns from single users. In contrast, Dabek and Caban [19] developed an iML system that learns from crowd interactions with data to generate a user model capable of assisting analysts during data exploration.

Explicit Human Feedback

Often explicit human feedback is provided through annotations and labels. This feedback can be either binary or graduated. The View Space Explorer [5] for instance, allows users to choose and annotate relevant or irrelevant example scatter plots. Gao et al. [22] proposed an interactive approach to 3D model repository exploration where a human assigns "like" or "dislike" labels to parts of a model or its entirety. RCLens [35] supports user guided exploration of rare categories through labels provided by a human. In a text document retrieval application [26], humans decide to accept, reject or label search query results. Similarly but for a video search system [33], users can either accept or reject sketched query results.

A richer and more nuanced approach to human feedback is proposed by Brown et al. in their Dis-function system [11], where selections of scatterplot points can be dragged and dropped to reflect human understanding of the structure of a text document collection. In this case, the closer the data points in the projected 2D space, the more similar they are. Ehrenberg et al. [20] proposed the "data programming" paradigm, where humans encode their domain expertise using simple rules, as opposed to the traditional method of hand-labelling training data. This allows to generate a large amount of noisy training labels, which the machine learning algorithm then tries to de-noise and model. Bryan et al. [13] implemented an audio source separatio system where humans annotate data and errors, or directly paint on a time-frequency or spectrogram display. In each of these cases, human feedback and choices are taken into consideration to update a machine learning model.

Mixed Human Feedback

To guide user exploration of large search spaces, EvoGraphDice [10, 6] combines explicit human feedback regarding the pertinence of evolved 2D data projections, and an implicit method based on past human interactions with a scatterplot matrix. For the explicit feedback, the user ranks scatterplots from one to five using a slider. The system also infers view relevance by looking at the visual motifs [51] in the ranked scatterplots. For example, if the user tends to rank linear point distributions highly, then this motif will be favored to produce the next generation of scatterplots. Importantly, the weights of these feedback channels are set to equal by default, but the human can choose to change the importance of each at any time during the exploration.

Healey and Dennis [25] developed interest-driven navigation in visualization, based on both implicit and explicit human feedback. The implicit feedback is gathered from human interactions with the visualization system, and from tracking to infer preferences based on where the human is looking. Their argument is that data

gathered through implicit feedback is noisy. To overcome this, they built a preference statement interface, where humans provide a subject, a classification, and a certainty. This preference interface allows the human to define rules to identify known elements of interest.

Another example is the SelPH system [31], which learns implicitly from a photo editing history, and explicitly from the direct interaction of a human with an optimisation slider. Together, these two feedback channels help to exclude what the authors call the "uninteresting" or meaningless" design spaces.

2.3 System Feedback

System feedback goes beyond showing the results of the co-operation between the human and the machine. It seeks to inform humans about the state of the machine learning algorithm, and the provenance of system suggestions, especially in the case of implicit user feedback.

System feedback can be *visual*: Boukhelifa et al. [10] used color intensity and a designated flag to visualise the system's interpretation of the mixed user feedback regarding the pertinence of 2D projections. Heimerl et al. [26] implemented a visual method and text labels to show the classifier's state, and the relevance of the selected documents to a search query. Legg et al. [33] visualised the similarity metrics they used to compute a visual search.

System feedback can be *uncertain*: Koyama at al. [31] indicated the system's confidence in the estimation of humans' preferences with respect to color enhancement. Behrisch et al. [5] provided a feature histogram and an incremental decision tree. These meta visualizations also communicate the classifier's uncertainty. Lin et al. [35] showed visualization of rare categories using their "category view", and a glyph-based visualization to show classification features as well as confidence.

System feedback can be *progressive*: Dabek and Caban [19] discussed the importance of choosing when to propose something to the human. Their approach consisted in providing feedback when the human is in need of guidance. They established a number of rules to detect when this occurs. UTOPIA [17] visualises intermediate output even before algorithmic convergence. Ehrenberg et al. [20] showed "on-the-spot" performance feedback using plots and tables. They claimed that this allows the user to iterate more quickly on system design, and helps navigate the key decision points in their data programming workflow.

For the majority of the iML systems we reviewed, system feedback was provided. It appears that this feedback is an important feature, perhaps because it helps humans better interpret the results, and allows them to correct any mistakes or areas of uncertainty in the inferred user model. The challenge, however, is to find the right level of feedback without having to fully expose the inner workings of the underlying models and their parameters.

2.4 Evaluation Methods and Metrics

In total, for the systems we reviewed, there were nine papers with case studies and usage scenarios [5, 17, 20, 21, 22, 25, 33, 35, 50], ten user studies [1, 2, 6, 11, 13, 19, 22, 26, 31, 33] and two observational studies [10, 38], in addition to surveys, questionnaires and interviews (7 papers). Although papers included some form of a controlled user study of an iML system, it was however acknowledged that this type of evaluation is generally difficult to conduct due to the various potential confounding factors such as previous knowledge [33]. Indeed, evaluating accuracy of an iML system is not always possible as ground truth does not always exist [1].

Objective Performance Evaluations

One way to evaluate how well the human-machine co-operation performs to achieve a task is to compare the iML system with its non-interactive counterpart (i.e. no human feedback), or to an established baseline system. Legg at al. [33] conducted a small scale empirical evaluation with three participants using three metrics inspired from content-based information retrieval: time, precision and recall. The idea was to manually identify five video clips as the ground truth, then to compare an iML video search system with a baseline system (a standard video tool with fast-forward) for a video search task. They found that participants performed better in the iML condition for this task. In a user study with twelve participants, Amerish et al. [1] compared traditional manual search to add people to groups on online social networks (using an alphabetical list or searching by name), to an interactive machine learning approach called ReGroup. They looked at the overall time it took participants to create groups, final group sizes, and speed of selecting group members. Their results show that the traditional method works well for small groups, whereas the iML method works best for larger and more varied groups.

Another way to objectively evaluate the success of the human-machine co-operation is to look at insights. In the context of exploratory data visualization, Endert et al. [21] and Boukhelifa et al. [10] found that with the help of user feedback, their respective iML systems were able to confirm known knowledge and to discover new insights.

Other evaluations in this category compared the iML application with and without system feedback. Dabek et al. [19] proposed a grammar-based approach to model user interactions with data, which is then used to assist other users during data analysis. They conducted a crowdsourced formal evaluation with 300 participants to assess how well their grammar-based model captures user interactions. The task was to explore a census dataset and answer twelve open ended questions that require looking for combinations of variables and axis ranges using a parallel coordinates visualization. When comparing their tool with and without system feedback, they found that system suggestions significantly improved user performance for all their data analysis tasks, although questions remain with regards to the optimal number of suggestions to display to the user.

A number of studies looked at algorithmic performance when user feedback was implicit versus explicit. Azuan et al. [2] who used a “pay-as-you-go” approach to

solicit user feedback during data integration and cleaning, compared the two human feedback methods for a data integration task. They found that user performance under the implicit condition was better than for the explicit feedback in terms of number of errors. However, the authors noted some difficulties in separating usability issues related to the explicit feedback interface from the performance results.

Finally, some authors focused on algorithm-centered evaluations, where two or more machine learning methods are compared. For instance, in the context of topic modelling, Choo et al. [17] compared latent dirichlet allocation and non-negative matrix factorisation algorithms, from the practical viewpoints of consistency of multiple runs and empirical convergence. Another example is by Bryan et al. [13] who chose objective separation quality metrics defined by industry standards, as objective measures of algorithmic performance for audio source separation.

Subjective Performance Evaluations

The subjective evaluations described in Table 1 were carried out using surveys, questionnaires, interviews, and informal user feedback. They included evaluation metrics related to these aspects of user experience: happiness, easiness, quickness, favorite, best helped, satisfaction, task load, trust, confidence in user and system feedback, and distractedness. Moreover, the observational studies [10, 38] that we reviewed provided rich subjective user feedback on iML system performance. Endert et al. [38] looked at semantic interaction usage, in order to assess whether the latter aids the sensemaking process. They state that one sign of success of iML systems is when humans forget that they are feeding information to an algorithm, and rather focus on "synthesising information relevant to their task".

Other evaluations looked at human behavioural variations with regards to different iML interfaces. Amerish et al. [1] compared two interfaces for adding people to online social networks, with and without the interactive component of iML. They looked at behavioural discrepancies in terms of how people used the different interfaces and how they felt. They found that participants were frustrated when model learning was not accurate. Koyama et al. [31] compared their adaptive photo enhancement system with the same tool stripped of advanced capabilities, namely the visual system feedback, the optimisation slider functions, and the ordering of search results in terms of similarity. Because photo enhancement quality can be subjective, performance of the iML system was rated by the study participants. In this case, they were satisfied with the iML system and preferred it over more traditional workflows.

In summary, There are many aspects of interactive machine learning systems that are being evaluated. Sometimes authors focus on the quality of the user interaction with the iML system (*human-centered evaluations*), or the robustness of the algorithms that are deployed (*algorithm-centered evaluations*), and only in a few cases detailed attention is drawn to the quality of human-machine co-operation and learning. These studies use a variety of evaluation methods, as well as objective and subjective metrics. Perhaps our main observation from this literature review, is that for the majority of the reviewed papers, only a single aspect of the iML system is evaluated. We need more evaluation studies that examine the different aspects of

iML sytems, not only as separate components but also from an integrative point of view.

In the next section, we introduce an interactive machine learning system for guided exploratory visualization, and describe our *multi-faceted* evaluation approach to study the effectiveness and usefulness of this tool for end users.

3 Case Study: Interactive Machine Learning For Guided Visual Exploration

Exploratory visualization is a dynamic process of discovery that is relatively unpredictable due to the absence of a-priori knowledge of what the user is searching for [24]. The focus in this case is on the organisation, testing, developing concepts, finding patterns and definition of assumptions [24]. When the search space is large, as is often the case for multi-dimensional datasets, the task of exploring and finding interesting patterns in data becomes tedious. Automatic dimension reduction techniques, such as principle component analysis and multidimensional scaling, reduce the search space, but often are difficult to understand [42], or require the specification of objective criteria to filter views before exploration. Other techniques guide the exploration towards the most interesting areas of the search space based on information learned during the exploration, which appears to be more adapted to the free nature of exploration [6, 12].

In our previous work on guided exploratory visualization [6, 7, 10, 46, 47], we tried to address the problem of how to efficiently explore multidimensional datasets characterised by a large number of projections. We proposed a framework for Evolutionary Visual Exploration (EVE, Figure 1) that combines visual analytics with stochastic optimisation by means of an Interactive Evolutionary Algorithm (IEA). Our goal was to guide users to interesting projections, where the notion of "interestingness" is defined *implicitly* by automatic indicators such as the amount of visual pattern in the two-dimensional views visited by the user, and *explicitly* via subjective human assessment.

In this section, we report on our experience in building and evaluating an interactive machine learning system called EvoGraphDice (Figure 3) using the EVE framework. We note that existing evaluations of interactive evolutionary systems tend to be algorithm-centered. Through this case study, we argue for a *multi-faceted* evaluation approach that takes into account all components of an iML system. Similar recommendations can be found for evaluating interactive visualization systems. For example, Carpendale [14] advocates for adopting a variety of evaluative methodologies that together may start to approach the kind of answers sought.

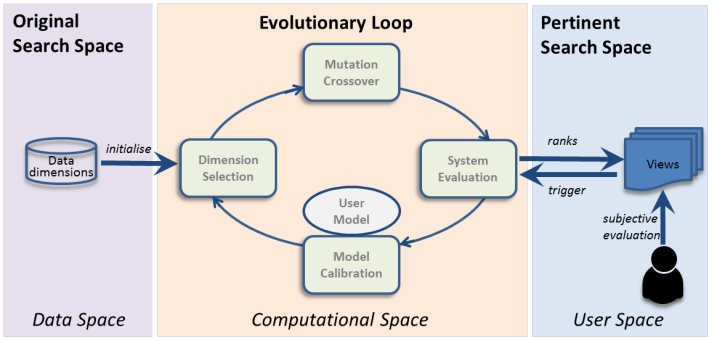

Fig. 1 The Evolutionary Visual Exploration Framework (EVE). Raw data dimensions (from the data space) are fed into an evolutionary loop in order to progressively evolve new interesting views to the user. The criteria for deciding on the pertinence of the new views is specified through a combination of automatically calculated metrics (from the computational space) and user interactions (at the user space).

3.1 Background on Interactive Evolutionary Computation IEC

There are many machine learning approaches, including artificial neural networks, support vector machines and bayesian networks. Moreover, many machine learning problems can be modelled as optimisation problems where the aim is to find a trade-off between an adequate representation of the training set and a generalisation capability on unknown samples. In contrast to traditional local optimisation methods, Evolutionary Algorithms (EAs) have been widely used as a successful stochastic optimisation tool in the field of machine learning in the recent years [44]. In this sense, machine learning and the field of Evolutionary Computation (EC), that encompasses EAs, are tightly coupled.

Evolutionary Algorithms (EAs) are stochastic optimisation heuristics that copy, in a very abstract manner, the principles of natural evolution that let a population of individuals be adapted to its environment [23]. They have the major advantage over other optimisation techniques of making only few assumptions on the function to be optimised. An EA considers populations of potential solutions exactly like a natural population of individuals that live, fight, and reproduce, but the natural environment pressure is replaced by an "optimisation" pressure. In this way, individuals that reproduce are the best ones with respect to the problem to be solved. Reproduction (see Figure 2) consists of generating new solutions via variation schemes (the genetic operators), that, by analogy with nature, are called mutation if they involve one individual, or crossover if they involve two parent solutions. A *fitness function*, computed for each individual, is used to drive the selection process, and is thus optimised by the EA. Evolutionary optimisation techniques are particularly efficient to address complex problems (irregular, discontinuous) where classical deterministic methods fail [4, 39], but they can also deal with varying environments [29], or non computable quantities [45].

Interactive Evolutionary Computation (IEC) describes evolutionary computational models where humans, via suitable user interfaces, play an active role, *implicitly* or *explicitly*, in evaluating the outputs evolved by the evolutionary computation (Figure 2). IEC lends itself very well to art applications such as for melody or graphic art generation where creativity is essential, due to the subjective nature of the fitness evaluation function. For scientific and engineering applications, IEC is interesting when the exact form of a more generalised fitness function is not known or is difficult to compute, say for producing a visual pattern that would interest a human observer. Here, the human visual system, together with their emotional and psychological responses are far superior than any automatic pattern detection or learning algorithm.

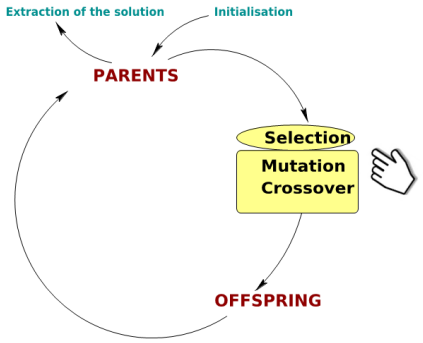

Fig. 2 The evolutionary loop: user interactions can occur at any stage including the selection and evaluation of individuals and the genetic operators.

Whereas current IEC research has focused on improving the robustness of the underlying algorithms, much work is still needed to tackle human-factors in systems where adaptation between users and systems is likely to occur [37].

3.2 The Visible and Hidden Roles of Humans in IEC

The role of humans in IEC can be characterised by the evolutionary component at which they operate, namely: initialisation, evolution, selection, genetic operators, constraints, local optimisation, genome structure variation, and parameters tuning. This may or may not be desirable from a usability perspective, especially for non-technical users. The general approach when humans are involved, especially for parameter tuning, is mostly by trial-and-error and by reducing the number of parameters. Such tasks are often visible, in that they are facilitated by the user interface. However, there exists a hidden role of humans in IEC that has often been neglected. Algorithm and system designers play a central role in deciding the details of the

fitness function to be optimised and in setting the default values of system parameters, and thus contributing to the "black-box" effect of IEC systems. Such tasks are influenced by the designer's previous experience and end-user task requirements.

Besides this hidden role in the design stage, there is a major impact of the "human-in-the-loop" on the IEC. This problem is known as the "user bottleneck", i.e. human fatigue due to the fact that the user and machine do not live and react at the same rate. Various solutions have been considered in order to avoid systematic and repetitive or tedious interactions, and the authors themselves have considered several of them, such as: (i) reducing the size of the population and the number of generations; (ii) choosing specific models to constrain the exploration in a-priori "interesting" areas of the search space; and (iii) performing an automatic learning (based on a limited number of characteristic quantities) in order to assist the user and only present interesting individuals of the population, with respect to previous votes or feedback from the user. These solutions require considerable computational effort. A different approach and new ideas to tackle the same issue could come from HCI and usability research, as discussed later on in this chapter.

3.3 EvoGraphDice Prototype

EvoGraphDice [6, 7, 10, 47] was designed to aid the exploration of multidimensional datasets characterised by a large space of 2D projections (Figure 3). Starting from dimensions whose values are automatically calculated by a Principle Component Analysis (PCA), an IEA progressively builds non-trivial viewpoints in the form of linear and non-linear dimension combinations, to help users discover new interesting views and relationships in their data. The criteria for evolving new dimensions is not known a-priori and is partially specified by the user via an interactive interface. Pertinence of views is modelled using a fitness function that plays the role of a predictor: (i) users select views with meaningful or interesting visual patterns and provide a satisfaction score; (ii) the system calibrates the fitness function optimised by the evolutionary algorithm to incorporate user's input, and then calculates new views. A learning algorithm was implemented to provide pertinent projections to the user based on their past interactions.

3.4 Multi-Faceted Evaluation of EvoGraphDice

We evaluated EvoGraphDice quantitatively and qualitatively following a mixed-approach, where on the one hand we analysed the computational behaviour of the system (algorithm-centered approach), and on the other hand we observed the utility and effectiveness of the system for the end-user (human-centered approach).

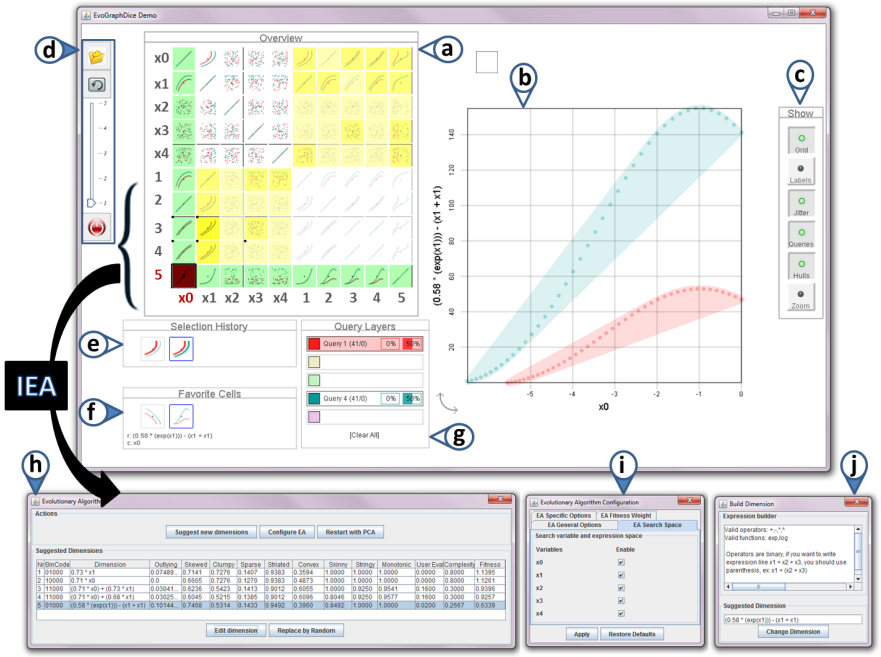

Fig. 3 EvoGraphDice prototype showing an exploration session of a synthetic dataset. Widgets: (a) an overview scatterplot matrix showing the original data set of 5 dimensions (x0..x4) and the new dimensions (1..5) as suggested by the evolutionary algorithm. (b) main plot view. (c) tool bar for main plot view. (d) a tool bar with (top to bottom) favorite toggle button, evolve button , a slider to evaluate cells and a restart (PCA) button. (e) the selection history tool. (f) the favorite cells window. (g) the selection query window. (h) IEA main control window. (i) window to limit the search space. (j) dimension editor operators.

3.4.1 Quantitative Evaluation

For this study [6], we synthesised a 5D dataset with an embedded curvilinear relationship between two dimensions and noise for the rest of the dimensions. The task was to find a data projection that shows a derived visual pattern. We logged user interactions with the tool and the state of the system at each algorithm iteration. For log data analysis, we used both statistical and exploratory visualization techniques.

Algorithm-Centered Evaluation

This evaluation focused on two aspects of our iML system: the *robustness* of the underlying algorithm, and the *quality of machine learning*. To study robustness, we conducted two types of analyses: (a) *convergence analysis* to assess the algorithms ability to steer the exploration toward a focused area of the search space, and (b) *diversity analysis* to assess the richness and variability of solutions provided by the algorithm. These two analyses are relevant because they relate to two important

mechanisms in evolutionary algorithms, *exploitation* and *exploration* [4], where on the one hand users want to visit new regions of the search space, and on the other hand they also want to explore solutions (combined dimensions) close to one region of the search space. In terms of objective metrics, we used the number of generations and task outcome to measure algorithmic performance, and mean visual pattern differences (using scagnostics [51]) to assess diversity. To evaluate the quality of learning, we used the rate of concordance between user evaluation scores, and the "predicted" values as calculated by the algorithm.

Our analysis showed that on average the interactive evolutionary algorithm followed the order of user ranking of scatterplots fairly consistently, even though users seemed to take different search and evaluation strategies. For example, some participants tended to lump evaluation scores to fewer levels, others used the five provided score levels, whereas the rest alternated between the two strategies at different stages of the exploration. Moreover, the results indicated a possible link between user evaluation strategy, and outcome of exploration and speed of convergence, where users taking a more consistent approach converged more quickly. The diversity analysis showed that, in terms of the visual pattern, the IEA provided more diverse solutions at the beginning of the exploration session before slowly converging to a more focused search space.

Human-Centered Evaluation

The user-centered evaluation of EvoGraphDice focused on two different aspects related to human interactions with the iML system. First we performed a *user strategy analysis* to understand the different approaches users took to solve the task. The evaluation metrics we used here were the type of searched visual pattern, ad stability of the exploration strategy. Second, we looked at *user focus* to highlight hot spots in the user interface and assess user evaluation strategies. In this case, our evaluation metrics were related to the user visitation and evaluation patterns.

In terms of results, the user strategies analysis showed that EvoGraphDice allows for different types of exploration strategies centered around three dominant types of scagnostics (skinny, convex and sparse) that appear to be relevant for the game task. We also found that the stability of the exploration strategy may be an important factor for determining the outcome of the exploration task and the speed of convergence, since successful exploration sessions had a more consistent strategy when compared to the unsuccessful ones, and they converged more quickly on average. From the user visitation and evaluation analyses, we found that users were more likely to visit scatterplots showing dimensions relevant to their task. Moreover, these plots were on average ranked highly by the user. Since for this game task, the main dimensions relevant to the task appeared on the top left side of the proposed cells, users intuitively started navigating that way. What we saw in these results was probably a mixture of task-relevance and intuitive-navigation, as the relevant original dimensions are placed in a prominent position in the matrix.

3.4.2 Qualitative Evaluation

To assess the usability and utility of EVE, we conducted another user study [10] where we tried to answer these three questions: is our tool understandable and can it be learnt; are experts able to confirm known insight in their data; and are they able to discover new insight and generate new hypotheses. We designed three tasks: (a) a game-task (similar to the task in the quantitative evaluation above) with varying levels of difficulty to assess participants abilities to operate the tool; (b) we asked participants to show in the tool what they already know about their data; and (c) to explore their data in light of a hypothesis or research question that they prepared. This sequence of tasks assured that experts became familiar with the tool, and understood how to concretely leverage it by looking for known facts, before looking for new insights. This evaluation approach sits between an observational study and an insight-based evaluation such as the one proposed by Saraiya et al. [41].

The study led to interesting findings such as the ability of our tool to support experts in better formulating their research questions and building new hypotheses. For insight evaluation studies such as ours, reproducing the actual findings across subjects is not possible as each participant provided their own dataset and research questions. However, reproducing testing methodologies and coding for the analysis is. Although we run multiple field studies with domain experts from different domains, with sessions that were internally very different, the high level tasks, their order and the insight based coding were common. Training expert users on simple specific tasks that are not necessarily "theirs" also seemed to help experts become confident with the system, but of course comes at a time cost.

4 Discussion

We conducted qualitative and quantitative studies to evaluate EVE which helped us validate our framework of guided visual exploration. While the observational study showed that using EVE, domain experts were able to formulate interesting hypothesis and reach new insights when exploring freely, the quantitative evaluation indicated that users, guided by the interactive evolutionary algorithm, are able to converge quickly to an interesting view of their data when a clear task is specified. Importantly, the quantitative study allowed us to accurately describe the relationship between user behaviour and algorithms response.

Besides interactive machine learning, guided visualization systems such as EVE fall under the wider arena of knowledge-assisted visualization and mixed-initiative systems [28]. In such cases, where the system is learning, it is crucial that users understand what the system is proposing or why changes are happening. Thus, when evaluating iML systems with users, we need to specifically test if the automatic state changes and their provenance are understood. It would be interesting, for example, to also consider evolving or progressive revealing of the provenance of system suggestions. This way, as the user becomes more expert, more aspects of the underlying

mechanics are revealed. When creativity and serendipity are important aspects, as it is the case in artistic domains and data exploration, new evaluation methodologies are required.

Research from the field of mixed initiative systems describes a set of design principles that try to address systematic problems with the use of automatic services within direct manipulation interfaces. These principles include considering uncertainty about a user's goal, transparency, and considering the status of users' attention [28]. We can be inspired by the extensive experience and past work from HCI, to also consider how user behaviour can in turn adapt to fit our systems [37].

During the design, development and evaluation of EVE, we worked with domain experts at different levels. For the observational study, we worked with data experts from various disciplines which allowed us to asses the usefulness, usability and effectiveness of our system in different contexts. In particular, we largely benefited from having one domain expert as part of the design and evaluation team. This expert explored multidimensional datasets as part of her daily work, using both algorithmic and visual tools. Involving end-users in the design team is a long-time tradition in the field of HCI as part of the user-centered design methodology. This is a recommendation we should consider more, both as a design and as a system validation approach. While HCI researchers acknowledge the challenges of forming partnerships with domain experts, their past experience (e.g. [16]) can inform us on how to proceed with the evaluation of iML systems.

5 Research Prospects

We report on observations and lessons learnt from working with application users both for the design and the evaluation of our interactive machine learning system, as well as the results of experimental analyses. We discuss these below as research opportunities aiming to facilitate and support the different roles humans play in iML, i.e. in the design, interaction and evaluation of these systems.

Human-Centered Design: during the design, development and evaluation of many of our tools, we worked with domain experts at different levels. For EvoGraphDice, for instance, we largely benefited from having a domain expert as part of the design and evaluation team. However, this was carried out in an informal way. Involving end-users in the design team is a long-time tradition in the field of HCI as part of the user-centered design methodology. Participatory design, for instance, could be conducted with iML end-users to incorporate their expertise in the design of the learning algorithm or user models. This is a recommendation we should consider in a more systematic way, both as a design and as a system validation approach.

Interaction and Visualization: often the solutions proposed by the iML systems are puzzling to end-users. This is because the inner workings of machine learning algorithms, and the user exploration and feedback strategies that led to system sug-

gestions are often not available to the user. This "black-box" effect is challenging to address as there is a fine balance to find between the richness of a transparent interface and the simplicity of a more obscure one. Finding the tipping point requires an understanding of evolving user expertise in manipulating the system, and the task requirements. Whereas HCI and user-centered design can help elicit these requirements and tailor tools to user needs over time, visualization techniques can make the provenance of views and the system status more accessible.

At the interaction level, HCI can contribute techniques to capture rich user feedback without straining the user, that are either implicit: e.g., using eye-tracking; or explicit such as using simple gestures or interactions mediated by tangible objects to indicate user subjective assessment of a given solution. Here, our recommendation is to investigate rich and varied interaction techniques to facilitate user feedback, and to develop robust user models that try to learn from the provided input.

Multifaceted Evaluation: the evaluation of iML systems remains a difficult task as often the system adapts to user preferences but also the user interprets and adapts to system feedback. Getting a clear understanding of the subtle mechanisms of this co-adaptation [37], especially in the presence of different types and sources of uncertainty [9], is challenging and requires to consider evaluation criteria other than speed of algorithm convergence and the usability of the interface.

In the context of exploration, both for scientific and artistic applications, creativity is sought and can be characterised by lateral thinking, surprising findings, and the way users learn how to operate the interactive system and construct their own way to use it. For IEC, Our observation is that augmented creativity can be achieved with the right balance between randomness and user-guided search. What is important to consider for evaluating any iML system, in the context of creativity, are the exploration components. Our recommendation with this respect is two-fold: first, to work towards creating tools that support creativity (something that the HCI community is already looking into [15]); and second, to investigate objective and subjective metrics to study creativity within iML (e.g. to identify impacting factors such as the optimisation constraints, user engagement and the presence or absence of direct manipulation). Some of these measures may only be identifiable through longitudinal observations of this co-adaptation process.

6 Conclusion

User-driven machine learning processes such as the ones described in this chapter, rely on systems that adapt their behaviour based on user feedback, while users themselves adapt their goals and strategies based on the solutions proposed by the system. In this chapter, we focused on the evaluation of interactive machine learning systems, drawing from related work, and our own experience in developing and evaluating interactive machine learning systems. We showed through a detailed literature review that despite the multifaceted nature of iML systems, current evaluations

tend to focus on single isolated components such as the robustness of the algorithm or the utility of the interface. Through a visual analytics case study, we show how coupling algorithm-centered and user-centered evaluation methods can bring forth insights on the underlying cooperation and adaptation mechanisms between the algorithm and the human. Interactive machine learning presents interesting challenges and prospects to conduct future research not only in terms of designing robust algorithms and interaction techniques, but also in terms of coherent evaluation methodologies.

References

1. Saleema Amershi, James Fogarty, and Daniel Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 21–30, New York, NY, USA, 2012. ACM.
2. Nurzety A. Azuan, Suzanne M. Embury, and Norman W. Paton. Observing the data scientist: Using manual corrections as implicit feedback. In *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pages 13:1–13:6, New York, NY, USA, 2017. ACM.
3. Benjamin Bach, André Spritzer, Evelyne Lutton, and Jean-Daniel Fekete. Interactive Random Graph Generation with Evolutionary Algorithms. In Springer, editor, *Graph Drawing*, Lecture Notes in Computer Science. Springer, 2012.
4. Wolfgang Banzhaf. *Handbook of Evolutionary Computation*. Oxford University Press, 1997.
5. M. Behrisch, F. Korkmaz, L. Shao, and T. Schreck. Feedback-driven interactive exploration of large multidimensional data supported by visual classifier. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 43–52, Oct 2014.
6. N. Boukhelifa, A. Bezerianos, W. Cancino, and E. Lutton. Evolutionary visual exploration: Evaluation of an iec framework for guided visual search. *Evol. Comput.*, 25(1):55–86, March 2017.
7. Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. A Mixed Approach for the Evaluation of a Guided Exploratory Visualization System. In W. Aigner, P. Rosenthal, and C. Scheidegger, editors, *EuroVis Workshop on Reproducibility, Verification, and Validation in Visualization (EuroRV3)*. The Eurographics Association, 2015.
8. Nadia Boukhelifa, Anastasia Bezerianos, Alberto Tonda, and Evelyne Lutton. Research prospects in the design and evaluation of interactive evolutionary systems for art and science. In *CHI workshop on Human Centred Machine Learning*, San Jose, United States, 2016.
9. Nadia Boukhelifa, Marc-Emmanuel Perrin, Samuel Huron, and James Eagan. How data workers cope with uncertainty: A task characterisation study. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3645–3656, New York, NY, USA, 2017. ACM.
10. Nadia Boukhelifa, Waldo Cancino Ticona, Anastasia Bezerianos, and Evelyne Lutton. Evolutionary visual exploration: Evaluation with expert users. *Comput. Graph. Forum*, 32(3):31–40, 2013.
11. E. T. Brown, J. Liu, C. E. Brodley, and R. Chang. Dis-function: Learning distance functions interactively. In *2012 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 83–92, Oct 2012.
12. Eli T. Brown, Jingjing Liu, Carla E. Brodley, and Remco Chang. Dis-function: Learning distance functions interactively. In *IEEE VAST*, pages 83–92. IEEE Computer Society, 2012.
13. Nicholas J. Bryan, Gautham J. Mysore, and Ge Wang. Isse: An interactive source separation editor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 257–266, New York, NY, USA, 2014. ACM.

14. Sheelagh Carpendale. Information visualization. chapter Evaluating Information Visualizations, pages 19–45. Springer-Verlag, Berlin, Heidelberg, 2008.
15. Erin Cherry and Celine Latulipe. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction.*, 21(4):21:1–21:25, 2014.
16. Parmit K. Chilana, Jacob O. Wobbrock, and Andrew J. Ko. Understanding usability practices in complex domains. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2337–2346, New York, NY, USA, 2010. ACM.
17. J. Choo, C. Lee, C. K. Reddy, and H. Park. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001, Dec 2013.
18. Gabriella Cortellessa and Amedeo Cesta. Evaluating mixed-initiative systems: An experimental approach. In *ICAPS*, volume 6, pages 172–181, 2006.
19. F. Dabek and J. J. Caban. A grammar-based approach for modeling user interactions and generating suggestions during the data exploration process. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):41–50, Jan 2017.
20. Henry R. Ehrenberg, Jaeho Shin, Alexander J. Ratner, Jason A. Fries, and Christopher Ré. Data programming with ddlite: Putting humans in a different part of the loop. In *Proceedings of the Workshop on Human-In-the-Loop Data Analytics*, HILDA '16, pages 13:1–13:6, New York, NY, USA, 2016. ACM.
21. Alex Endert, Patrick Fiaux, and Chris North. Semantic interaction for visual text analytics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 473–482, New York, NY, USA, 2012. ACM.
22. L. Gao, Y. P. Cao, Y. K. Lai, H. Z. Huang, L. Kobbelt, and S. M. Hu. Active exploration of large 3d model repositories. *IEEE Transactions on Visualization and Computer Graphics*, 21(12):1390–1402, Dec 2015.
23. David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.
24. Georges G. Grinstein. Harnessing the human in knowledge discovery. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *KDD*, pages 384–385. AAAI Press, 1996.
25. C. G. Healey and B. M. Dennis. Interest driven navigation in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(10):1744–1756, Oct 2012.
26. F. Heimerl, S. Koch, H. Bosch, and T. Ertl. Visual classifier training for text document retrieval. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2839–2848, Dec 2012.
27. Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.
28. Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, pages 159–166, New York, NY, USA, 1999. ACM.
29. Yaochu Jin and Jürgen Branke. Evolutionary optimization in uncertain environments-a survey. *IEEE Trans. Evolutionary Computation*, 9(3):303–317, 2005.
30. Diane Kelly and Jaime Teevan. Implicit feedback for inferring user preference: A bibliography. *SIGIR Forum*, 37(2):18–28, September 2003.
31. Yuki Koyama, Daisuke Sakamoto, and Takeo Igarashi. Selph: Progressive learning and support of manual photo color enhancement. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2520–2532, New York, NY, USA, 2016. ACM.
32. Yann Landrin-Schweitzer, Pierre Collet, and Evelyne Lutton. Introducing lateral thinking in search engines. *Genetic Programming an Evolvable Hardware Journal.*, 1(7):9–31, 2006.
33. P. A. Legg, D. H. S. Chung, M. L. Parry, R. Bown, M. W. Jones, I. W. Griffiths, and M. Chen. Transformation of an uncertain video search pipeline to a sketch-based visual analytics loop. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2109–2118, Dec 2013.

34. P. Legrand, C. Bourgeois-Republique, V. Pean, E. Harboun-Cohen, J. Lévy Véhel, B. Frachet, E. Lutton, and P. Collet. Interactive evolution for cochlear implants fitting. *GPEM*, 8(4):319–354, 2007.
35. H. Lin, S. Gao, D. Gotz, F. Du, J. He, and N. Cao. Rclens: Interactive rare category exploration and identification. *IEEE Transactions on Visualization and Computer Graphics*, PP(99):1–1, 2017.
36. Evelyne Lutton. Evolution of fractal shapes for artists and designers. *IJAIT, International Journal of Artificial Intelligence Tools*, 15(4):651–672, 2006. Special Issue on AI in Music and Art.
37. Wendy Mackay. Responding to cognitive overhead: co-adaptation between users and technology. *Intellectica*, 30(1):177–193, 2000.
38. C. North, A. Endert, and P. Fiaux. Semantic interaction for sensemaking: Inferring analytical reasoning for model steering. *IEEE Transactions on Visualization & Computer Graphics*, 18:2879–2888, 2012.
39. Riccardo Poli and Stefano Cagnoni. Genetic programming with user-driven selection: Experiments on the evolution of algorithms for image enhancement. In *Genetic Programming Conference*, pages 269–277. Morgan Kaufmann, 1997.
40. Dominik Sacha, Michael Sedlmair, Leishi Zhang, John Aldo Lee, Daniel Weiskopf, Stephen North, and Daniel Keim. Human-centered machine learning through interactive visualization. ESANN, 2016.
41. P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 11(4):443–456, July 2005.
42. M Sedlmair, M Brehmer, S Ingram, and T Munzner. Dimensionality reduction in the wild: Gaps and guidance. *Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep. TR-2012-03*, 2012.
43. Y. Z. Song, D. Pickup, C. Li, P. Rosin, and P. Hall. Abstract art by shape classification. *IEEE Transactions on Visualization and Computer Graphics*, 19(8):1252–1263, Aug 2013.
44. Kenneth Stanley and Risto Miikkulainen. Evolving Neural Networks Through Augmenting Topologies. *Evolutionary Computation*, 10(2):99–127, 2002.
45. Hideyuki Takagi. Interactive evolutionary computation : System optimisation based on human subjective evaluation. In *Proceedings of Intelligent Engineering Systems (INES'98)*. IEEE, 1998.
46. Waldo Cancino Ticona, Nadia Boukhelifa, Anastasia Bezerianos, and Evelyne Lutton. Evolutionary visual exploration: experimental analysis of algorithm behaviour. In Christian Blum and Enrique Alba, editors, *GECCO (Companion)*, pages 1373–1380. ACM, 2013.
47. Waldo Cancino Ticona, Nadia Boukhelifa, and Evelyne Lutton. Evographdice: Interactive evolution for visual analytics. In *IEEE Congress on Evolutionary Computation*, pages 1–8. IEEE, 2012.
48. Alberto Tonda, Andre Spritzer, and Evelyne Lutton. Balancing user interaction and control in bayesian network structure learning. In *Artificial Evolution Conference*, LNCS 8752. Springer, 2013.
49. Gregory Valigiani, Evelyne Lutton, Yannick Jamont, Raphael Biojout, and Pierre Collet. Automatic rating process to audit a man-hill. *WSEAS Transactions on Advances in Engineering Education*, 3(1):1–7, 2006.
50. John Wenskovitch and Chris North. Observation-level interaction with clustering and dimension reduction algorithms. In *Proceedings of the 2Nd Workshop on Human-In-the-Loop Data Analytics*, HILDA'17, pages 14:1–14:6, New York, NY, USA, 2017. ACM.
51. Leland Wilkinson, Anushka Anand, and Robert Grossman. Graph-theoretic scagnostics. 2005.