DOI: 10.1111/cgf.14329
EUROVIS 2021
N. Smit, K. Vrotsou, and B. Wang
(Guest Editors)

Volume 40 (2021), Number 3
STAR – State of The Art Report

A Survey of Human-Centered Evaluations in Human-Centered Machine Learning

F. Sperrle[1], M. El-Assady[1], G. Guo[2], R. Borgo[3], D. Horng Chau[2], A. Endert[2], and D. Keim[1]

[1]University of Konstanz [2]Georgia Institute of Technology [3]King's College London

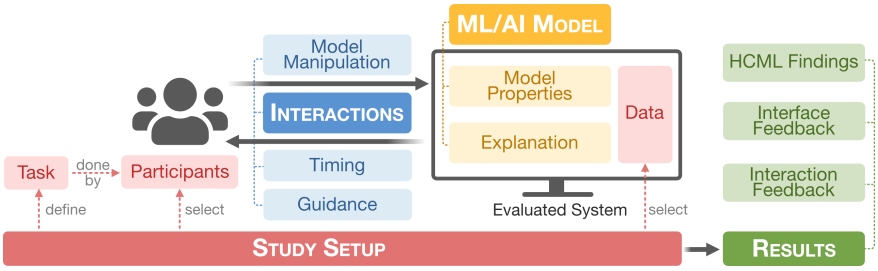


Figure 1: *The four main aspects (**study setup, ML/AI models, interactions**, and **results**) of human-centered evaluations in human-centered machine learning and their related dimensions. The assigned colors are used as a structuring visual element throughout this survey.*

Abstract

Visual analytics systems integrate interactive visualizations and machine learning to enable expert users to solve complex analysis tasks. Applications combine techniques from various fields of research and are consequently not trivial to evaluate. The result is a lack of structure and comparability between evaluations. In this survey, we provide a comprehensive overview of evaluations in the field of human-centered machine learning. We particularly focus on human-related factors that influence trust, interpretability, and explainability. We analyze the evaluations presented in papers from top conferences and journals in information visualization and human-computer interaction to provide a systematic review of their setup and findings. From this survey, we distill design dimensions for structured evaluations, identify evaluation gaps, and derive future research opportunities.

1. Introduction

Recent advances in artificial intelligence (AI) and machine learning (ML), have led to numerous breakthroughs across many application domains. Often, complex systems are developed by combining the latest innovations from ML, interactive systems, visual analytics, and many other fields. The emerging research area **human-centered machine learning** (HCML) takes a holistic view on the ML process, placing particular focus on human input, interactions, and collaboration, and the involvement of different stakeholders in the ML process [SSSE20] to enact the iterative context-sensitive checking characteristic of the human brain [Seg19]. HCML combines research in AI and ML with research in *visualization* (VIS) and *human-computer interaction* (HCI) and has become a core topic of *visual analytics* research over the last years, as indicated in Figure 2. HCML is closely linked to current research efforts in *eXplainable AI* (XAI) [Gun17], and the intelligibility of machine learning models [WB19].

A challenge of current HCML research is the ability to provide nuanced evaluations of systems, given their complexity and multifaceted nature. Most papers provided small-scale evaluations of simplified and encapsulated tasks [BBL18]. The well-established methodology for ML evaluation (e.g., accuracy, F-score, squared error) only covers some result-oriented aspects of human work, such as their impact on model quality. To holistically evaluate HCML processes, human factors like trust and effort also need to be evaluated. Due to the field's novelty, there is no established, general methodology for evaluations of HCML systems yet. Such an established methodology would benefit current HCML research efforts in making evaluations replicable and more comparable.

In this *State-of-the-Art-Report* (STAR), we present the first focused review of **human-centered evaluations** (HCE) of **human-centered machine learning** (HCML), providing a grouping of papers by HCML task. We discuss the particular challenges and evaluation designs that are frequently used in different domains and distill our findings into a checklist to provide guidance for the design of HCML evaluations, advancing towards a structured evaluation methodology for HCML. The aspects and dimensions we use to group papers and structure this STAR are shown in Figure 1. They

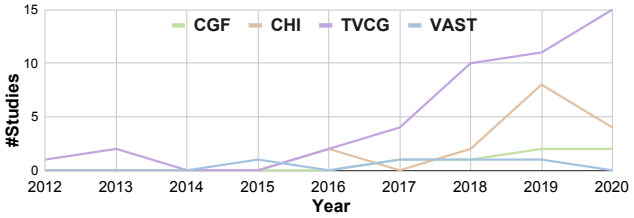


Figure 2: ***HCEs per Year & Venue****: We see a trend in more human-centered evaluations over the past years, particularly, since 2017.*

are derived from a visual analytics perspective, considering three main concepts: the system as well as its (machine learning) models, the user, and interactions between them. Rather than focusing on individual machine learning tasks, such as clustering or classification (already surveyed in [ERT*17]), we consider aspects of models and explanations that are particularly important for a successful human-machine interaction, such as trust. These aspects go beyond simple interactions like providing relevance feedback or weighting dimensions and are more difficult to tackle.

Our STAR has four major contributions: (1) a survey of human-centered evaluations of HCML systems, (2) dimensions for the structured and comparable evaluation of such systems that are derived from the results of a comprehensive survey, (3) a checklist as guidance for the design of HCML evaluations and a template for reporting HCE findings, and (4) an overview of underexplored and infrequently evaluated dimensions as a starting point for future research.

For an interactive overview of our results, see the survey browser at https://human-centered-evaluations-star.dbvis.de.

We structure our paper as follows: in Section 2 we synthesize a definition of HCML from previous work, compile the main challenges in HCML, and discuss problems of current evaluations. In Section 3, we introduce our survey's methodology, including our iterative coding process and paper selection criteria. Then, in Section 4, we introduce the dimensions of analysis, including evaluation setup, model properties, and interaction and guidance techniques. In Section 5, we discuss the evaluation of technical contributions of HCML, and in Section 6, we cover application-specific evaluations in the domains of bio-medicine, machine learning, and linguistics. In Section 7, we summarize our findings, discuss limitations of our survey, and present opportunities for future research which emerge from our survey.

2. Background

While there is no unified definition of HCML, there is a consensus that HCML considers factors pertaining to human involvement in machine learning pipelines, whether as users or as teachers [FG18]. Below, we provide several perspectives that build on each other before providing a unified definition that will be used in this survey.

2.1. Definition of HCML

In 2014, Amershi et al. [ACKK14] defined *interactive machine learning* as a form of machine learning that directly includes an end-user into the loop to enable rapid feedback and model development. They contrast this approach with *applied machine learning*, in which domain experts rely on ML practitioners to train models in slow, asynchronous loops. Consequently, they find that "*interactive machine learning can facilitate the democratization of applied machine learning, empowering end-users to create machine-learning-based systems for their own needs and purposes*." [ACKK14]

In visual analytics, Endert et al. called for a paradigm shift from *human in the loop* to what they called "*the human is the loop*" [EHR*14], making a first step from interactive machine learning towards human-centered machine learning. According to their vision, systems should facilitate sensemaking tasks by seamlessly integrating analysis capabilities into existing workflows without disrupting users. Amongst others, this includes enabling more expressive forms of user feedback and the use of spatialization to define common ground between humans and machines. Dimensionality reduction algorithms are particularly suitable to generate spatializations as they are typically unsupervised. Further, they can benefit from user interaction to monitor errors and reduce reduction losses, as surveyed by Nonato and Aupetit [NA19] and Sacha et al. [SZS*17].

More recently, Fiebrink et al. state that HCML should consider both the human work and the "*human contexts*" [FG18] in machine learning workflows. Such human work comes in many forms, such as collecting and annotating training data, deriving machine learning pipelines, interacting with intelligent systems to derive knowledge [SSS*14] and fine-tuning them. Understanding how humans interact in such situations can help to not only make the systems more usable but also discover new areas in which machine learning could be helpful [FG18]. Sacha et al. [SSZ*17] highlight how interaction offers considerable potential for improved support of ML with respect to interpretability, understandability, evaluation, and refinement. They also advocate for the integration of a multidisciplinary perspective as a contributor to bridge the gaps between automated ML methods and human reasoning.

Gilles et al. see the potential for HCML to lead to "*new ways of framing learning computationally*" [GFT*16]. According to their perspective, HCML includes "*exploring the co-adaptation of humans and systems*" [GFT*16]. In the visual analytics context, such co-adaptation can facilitate knowledge generation and is particularly applicable in the context of guidance. There, Sperrle et al. [SJB*20] have recently proposed to view co-adaptive guidance from the perspective of simultaneous learning and teaching processes.

Going beyond co-adaptive human-machine collaboration, human-centered AI "*is a perspective on AI and ML that intelligent systems must be designed with an awareness that they are part of a larger system consisting of human stakeholders, such as users, operators, clients, and other people in close proximity*" [Rie19]. This view does not only include user perspectives into ML and AI systems but aims to provide a holistic, systemic perspective. Shneiderman aims to operationalize human-centered AI by providing a framework that clarifies how to (1) design for high levels of human control and high levels of computer automation so as to increase human performance, (2) understand the situations in which full human control or full computer control are necessary, and (3) avoid the dangers of excessive human control or excessive computer control [Shn20].

More generally, recent research in explainable artificial intelligence (see [CPC19; Mil19; Rie19] for an overview) has focused on measuring and improving algorithm transparency, trustworthiness,

and intelligibility. Wortman Vaughan and Wallach postulate that depending on the different stakeholders in the AI process, rather than model intelligibility, the intelligibility of "*datasets, training algorithms or performance metrics*" [WW21] could be more critical.

Building on this previous work, we define human-centered machine learning as follows:

> *Human-centered machine learning is a field of research that considers humans and machines as equally important actors in the design, training, and evaluation of co-adaptive machine learning scenarios.*

All surveyed resources advocate the need to not only include humans in machine learning pipelines but also comprehensively consider human factors. While earlier approaches like interactive machine learning were primarily concerned with increased efficiency and agency, more recent work bridges the gap to psychology and sociology and emphasizes that humans are deeply embedded into HCML workflows. Consequently, evaluations should also be human-centered.

2.2. Surveys of HCML Methods

As a result of its interdisciplinary nature, HCML relies on a multitude of techniques and methods for designing and implementing machine learning processes that address the challenges outlined above. Here, we focus on related surveys from the visual analytics domain that summarize existing approaches. Endert et al. [ERT*17] survey the integration of machine learning techniques into visual analytics applications. They note a particular increase in the tight coupling of bespoke visualization systems and steerable machine learning algorithms. The resulting systems place equal importance on visualization, machine learning algorithms, and interaction affordances to balance human and machine effort and increase user trust and model interpretability. Chatzimparmpas et al. [CMJ*20] specifically survey methods to increase trust in machine learning through visualizations and collect a large set of techniques and methods for different domains and tasks. Hohman et al. provide a human-centered survey of visual analytics for deep learning that, amongst others, aims to identify "*types of people and users that would use and stand to benefit from visualizing deep learning*" [HKPC19], in which circumstances deep learning visualization is typically used. However, in contrast to this paper, those works do not focus on whether and how the presented approaches were evaluated. Yuan et al. [YCY*20] provide the most recent survey of visual analytics for machine learning. They distinguish techniques that are employed before, during, or after model training and that enable human involvement at the respective stage.

2.3. Challenges in HCML

Here, we outline common challenges in the research of HCML, synthesized from related work and the papers surveyed in this STAR. These challenges manifest in challenges for human-centered evaluation that will be introduced in Section 7.2.

HCML-C1: **Interdisciplinarity.** Human-centered machine learning unites various fields like machine learning, explainable artificial intelligence, human-computer interaction, and psychology. Consequently, successful work in this area must bridge the gaps between domains and encourage interdisciplinary collaboration [SSZ*17].

HCML-C2: **Complexity.** As HCML systems should be designed "*in full recognition of the agency and complexity of human users*" [FG18] they tend to be sophisticated, bespoke solutions to a given problem. Due to the systems' integrated nature, suboptimal (design) choices can hamper overall success. For instance, neglecting the underlying machine learning algorithms could lead to systems with great usability but weak performance. In contrast, well-performing machine learning models might not be used to their potential when embedded in poorly designed systems. Hence, some or all of their parts might not be intelligible to stakeholders, necessitating appropriate trust calibration methods. Furthermore, biased perceptions due to novelty effects [SCG09] and participant response biases [DVM*12] complicate effective evaluation and rapid iteration.

HCML-C3: **Co-Adaptation.** Many HCML systems observe user interactions and adapt their models to specific users, changing their characteristics over time (e.g., [CVL*18; SSKE19]). At the same time, users observe system responses and adapt their workflows [SJB*20]. In this process, systems might learn false or eventually outdated information. Consequently, they must offer interaction sequences that allow reverting previous adaptations. Further, co-adaptation can become a source of frustration for users when they are implicitly expected to participate in the system training. Moreover, the user becoming a teacher may make the system vulnerable to user biases. In addition to raising challenges during system design and use, co-adaptation also complicates the design of replicable studies.

HCML-C4: **Stakeholder Diversity.** Multiplicity of stakeholders poses a challenge to the assessment of validity and result generalizability. User segmentation is confounded by factors including cultural and educational background, age, gender, expertise, moral, and social contexts [BBL18]. In the context of HCML, all these factors coupled with subtle differences like personality traits can influence how an action is perceived, reacted to, and executed. Similar to co-adaptation, stakeholder diversity may hinder result replicability.

2.4. Evaluation of HCML

Due to the complexity and interdisciplinary challenges of HCML systems discussed above, their evaluation is typically complex given the many different factors that can be considered. Boukhelifa et al. distinguish between human-centered evaluations that focus on the interaction quality and algorithm-centered evaluations aiming to assess the robustness of the deployed algorithms [BBL18]. It is important to note that HCML systems can be successfully evaluated using both approaches. However, in this STAR, we focus on the first group, human-centered evaluations, to emphasize the role of the human in the interactive machine learning process. With increasing complexity of systems, many designs account for "*issues of fairness, accountability, interpretability, and transparency*" [Rie19]. These factors, often inspired by recent research in explainable artificial intelligence, inherently require a human perspective for evaluation.

Algorithm-Centered Evaluations – As algorithm-centered evaluations are frequently used in machine learning and artificial intelligence research, there are established methodologies that can be applied. They typically rely on quantitative analysis and report on model properties performance. Clear cut metrics (e.g., accuracy and F-score) exist to evaluate supervised ML techniques. Evaluation of

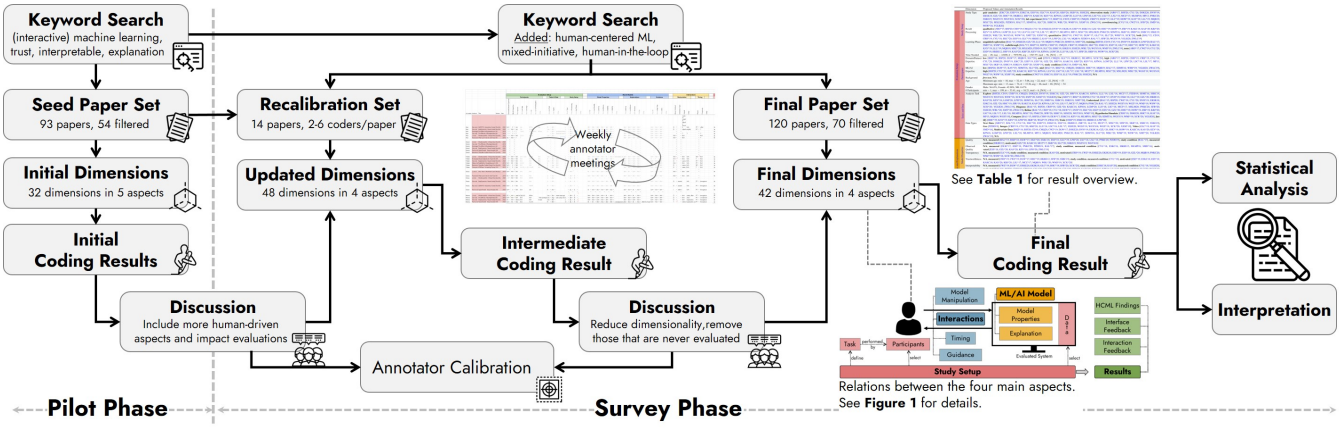


Figure 3: ***Survey Methodology****: We first compiled a set of seed papers to derive initial coding dimensions. Based on these results, we updated the dimensions and strengthened the human-centric focus of the survey. We finally code 71 papers in 42 dimensions using an iterative process.*

unsupervised learning is more complex with metrics commonly separated into the two categories internal validation and external validation [Pal19]. The lack of ground truth in unsupervised ML techniques makes user knowledge about data core to several evaluations that comparatively validate models against user expectations. In this context, interaction plays a central role [NA19] for lack of established metrics that could replace ground-truth data.

Human-Centered Evaluations – In contrast, a comprehensive overview for human-centered evaluations is currently missing. Here it is important to note that human-centered evaluations can be both qualitative and quantitative, as these two dimensions are orthogonal. In our selection, 90% of papers report qualitative evaluations, while 37% report quantitative, human-centered measurements. Human-centered evaluations of HCML systems must consider methods that adequately evaluate all aspects of the human machine partnership. For instance, testing only interface usability may miss out on the system's ability to interpret user input to create more accurate models. Similarly, only measuring model accuracy misses out on evaluating the user experience. Ideally, HCE should evaluate both cognitive and emotional elements involved in human-machine interaction aspects, such as co-adaptation and co-creation, pushing the boundary beyond the assessment of usability and user experience.

2.5. Surveys of Evaluations

Boukhelifa et al. [BBL18] discuss challenges in the evaluation of interactive machine learning systems and base their insights on previous experience in the field, as well as a survey of recent works. They consider all types of evaluations and find that "*current evaluations tend to focus on single isolated components such as the robustness of the algorithm, or the utility of the interface*" [BBL18]. In contrast to their survey, we focus on human-centered evaluations. As a result, we report on more focused coding dimensions and follow different paper selection and exclusion criteria (see Section 3.2). While the survey of evaluations in information visualization by Borgo et al. [BMB*18] is not directly related to HCML, it provides detailed dimensions for reporting study designs and participant characteristics. As these dimensions are equally important in evaluations of HCML, we report on them as well.

3. Methodology

As described above, the focus of this survey is on human-centered evaluations of human-machine interaction in the fields of visual data analysis and machine learning. Thus, we collected papers from the following high-quality journals and conferences in that domain:

- IEEE Transactions on Visualization and Computer Graphics (TVCG, including IEEE InfoVis and IEEE VAST proceedings)
- Computer Graphics Forum (CGF, including EuroVis proceedings)
- Proceedings of ACM Computer Human Interaction (CHI)

For all venues, we considered the years 2012 to 2020 to focus on recent developments. We do not consider short papers or workshop papers as they do not typically provide extensive evaluations.

3.1. Iterative Coding Methodology

Our methodology is split into two distinct phases that are outlined in Figure 3. In the initial *pilot phase*, we performed a keyword search for *machine learning, interactive machine learning, trust, interpretable, interpretability, explanation*, and *explainability* on the titles, abstracts, as well as contents of published works, retrieving an initial set of potentially relevant papers. We manually screened all papers and excluded those that did not deal with some form of machine learning or artificial intelligence or that did not perform user-based evaluations. Further, we excluded all papers describing systems that did not afford user-model interactions, leaving us with 54 papers at this stage. The final paper selection and exclusion criteria and examples of excluded papers will be presented below.

Starting with eight papers randomly selected among the collected 54, we began an exploratory coding phase in which we extracted all potentially relevant dimensions and distilled them into coding guidelines. Next, we refined the guidelines until an agreement between all coders was reached. This left us with 32 dimensions in 5 aspects, focusing on *user characteristics*, *XAI properties*, *model properties*, *tasks and environment*, and *study setup*. When discussing the initial coding results, it became clear that the survey was too focused on properties of (X)AI models and explanations and did not sufficiently cover the effects and timings of various interaction options.

Venue	#Collected	#Coded	#Experiments
IEEE Transactions on Visualization and Computer Graphics	64	41	44
IEEE Conference on Visual Analytics Science and Technology (VAST)	7	5	5
Computer Graphics Forum	9	6	6
Conference on Human Factors in Computing Systems (CHI)	41	15	16
Total	121	67	71

Table 1: ***Publications per Venue****: We collected 121 papers; out of which 67 met our selection criteria and where coded. Four papers report on two relevant evaluations, leading to 71 coded experiments.*

As a result, we added the following keywords to our search terms: *human-centered machine learning, mixed-initiative, human-in-the-loop*, and *intelligible*, and entered the *survey phase*: We started a recalibration process and derived new dimensions in four aspects: *study setup*, *ML/AI model*, *interactions*, and *results*.

We then assigned 12 selected papers to all six authors, such that each paper was coded by two annotators. Further, we selected two papers that were coded by all authors. Through a discussion, we then used the obtained results to both refine our coding guidelines and calibrate our annotations to ensure inter-annotator agreement. Further, we derived initial criteria to decide whether to code or exclude a paper. From there, we entered an iterative coding cycle in which we collected a total of 121 papers. Table 1 provides an overview of the publication venues of the 67 papers that we coded as relevant. We used weekly annotator meetings to discuss the obtained results and ensure continued inter-annotator agreement. In this phase, we removed six dimensions that were not reported in any of the coded papers and fine-tuned our paper exclusion criteria.

When coding papers, we did not attempt to resolve potential conflicts, ambiguities, or overlaps between concept definitions (e.g., transparency, intelligibility) but captured them as presented by the authors. Instead, we present short definitions of all dimensions in Section 4. Refining these concepts and converging on a common vocabulary presents an opportunity for future research.

3.2. Paper Selection and Exclusion Criteria

We manually evaluated all potential papers of interest and excluded those that did not deal with some form of interactive machine learning or artificial intelligence or that do not provide a user-based evaluation. We focus on systems that afford direct or indirect interactions with the underlying models. As a result, we exclude papers that do not include interactivity related to the analysis task. In particular, papers matching any of the following criteria are excluded:

- Papers that provide use cases or usage scenarios developed by the authors without the inclusion of expert feedback, or case studies that do not consider human factors pertaining to the expert (e.g., [GWGvW19; KTC*19; LJLH19; PLM*17; SJS*18; WPB*20]). This was the most frequent reason for exclusion.
- Papers that describe applications that do not allow user interaction with the model beyond filtering of data points (i.e., purely exploratory systems in which the user can neither influence the model behavior during the analysis session nor optimize towards a specific model output) (e.g., [JVW20; LLT*20; XXL*20]).
- Papers not describing system evaluations but research agendas (e.g., [AVW*18]), or workshops (e.g., [AW18; BCP*19]).
- Papers that provide quantitative evaluations of results not generated by participants in a study setting (e.g., [BZL*18; YDP19]).

We recognize that these criteria exclude a significant number of (HC)ML papers at the intersection between machine learning, visualization, and human-computer interaction. However, the focus of this STAR is on human-centered evaluations; several recent surveys on visual analytics and machine learning without this focus exist.

4. Dimensions of Analysis

Following our methodology, we iteratively refined the dimensions coded in our review. They are summarized in Table 2. Below, we introduce all dimensions, provide definitions where necessary, and present summary statistics. For an overview of all annotation results and definitions of all coding values, see the supplementary material.

4.1. Evaluation Setup

This aspect captures properties of the study setup, the participants, and the analysis tasks and data types used in the study. It is fundamental to assessing a study's internal validity, as the level of precision in reporting each dimension supports evaluating the strengths and truthfulness of inferences regarding cause-effect or causalities.

4.1.1. Study Setup

The first category of the evaluation setup is the study setup. This describes study protocols and methodologies for data collection, analysis of results, and forms of participant training when included. Study setup dimensions are interlinked with the *Participants* dimensions, with the method chosen in the Learning Phase dimension being correlated with the required expertise level of the participants.

Study Type

Definition: The study type defines how the study was designed and carried out.

Values: Observation Study, Pair Analytics, Lab Experiment, Crowdsourcing.

Across the 71 experiments surveyed, 37% each were lab studies (e.g., [CHH*19; LLL*19; MQB19]), 42% observational studies (e.g., [BHZ*18; BSP20; PNKC20]), and 13% pair analytics studies in which visual analytics experts support participants with the technical challenges raised by complex systems (e.g., [KAS*20; SKB*18]). Four of the analyzed papers presented results from multiple studies, and one relies on multiple study types by combining a pair analytics and a lab study [BAL*15]. Only two studies used long term analytics [KPN16; MP13] while six evaluations used crowdsourcing [CVL*18; CWZ*19; SFB*20; SMD*16; WSW*18; YGLR20] .

Result Processing

Definition: The result processing defines the type of data collected within a study, such as qualitative and/or quantitative.

Values: Qualitative, Quantitative, Both.

Qualitative research appears to be the favored approach, with 65% of the studies focusing on gathering qualitative feedback in the form of interviews, surveys, and observations, often leveraging think-aloud

		Dimension	Proposed Values and Annotation Results
Evaluation Setup	Study Setup	Study Type	**pair analytics** [EKC*20; ESD*19; ESKC18; ESS*18; GLC*19; KAS*20; KBJ*20; SKB*18; SSSE20], **observation study** [ARO*17; BSP20; CYL*20; DSKE20; DVH*19; EKSK18; GZL*20; HHC*19; HKBE12; JSR*19; KAKC18; KEV*18; KPN16; LGM*20; LLS*18; LPH*20; LSC*18; LSL*17; LXL*18; MCZ*17; MLMP18; MP13; PNKC20; SSKE19; WGSY19; WGYS18; XCK*20], **lab experiment** [BAL*15; BHZ*18; CD19; CHH*19; CMQ20; CRH*19; DLW*17; dSBD*12; GLC*19; HOW*19; KAY*19; LLL*19; MQB19; MXC*20; MXLM20; PZDD19; RAL*17; SDMT16; SLC*20; SSBC19; WBL*20; WMJ*19; XXM*19; ZWLC19], **crowdsourcing** [CVL*18; CWZ*19; SFB*20; SMD*16; WSW*18; YGLR20]
		Result Processing	**qualitative** [ARO*17; BSP20; CHH*19; CMQ20; CYL*20; DSKE20; DVH*19; EKSK18; ESD*19; ESKC18; GZL*20; HHC*19; HOW*19; JSR*19; KAKC18; KAS*20; KBJ*20; KEV*18; KPN16; LGM*20; LLL*19; LLS*18; LSC*18; LSL*17; MCZ*17; MLMP18; MP13; MXC*20; MXLM20; PNKC20; SDMT16; SKB*18; SMD*16; SSBC19; SSKE19; SSSE20; WBL*20; WGYS18; WSW*18; XMT*20; XXM*19], **quantitative** [BHZ*18; CWZ*19; DLW*17; GLC*19; SLC*20; WMJ*19; XCK*20], **both** [BAL*15; CD19; CRH*19; CVL*18; dSBD*12; EKC*20; ESS*18; GLC*19; HKBE12; KAY*19; LPH*20; LXL*18; MQB19; PZDD19; RAL*17; SFB*20; WGSY19; YGLR20; ZWLC19]
		Learning Phase	**unguided exploration** [BAL*15; DSKE20; GZL*20; LLL*19; MQB19; PNKC20; SDMT16; XMT*20], **training** [BSP20; CD19; CVL*18; DVH*19; EKSK18; LPH*20; RAL*17; SMD*16; XXM*19], **walkthrough** [BAL*15; BHZ*18; BSP20; CHH*19; CMQ20; CRH*19; DSKE20; EKC*20; ESKC18; ESS*18; GLC*19; HHC*19; HOW*19; KAKC18; KAY*19; LLL*19; MQB19; MXC*20; MXLM20; PZDD19; SLC*20; SSBC19; SSKE19; SSSE20; WBL*20; WGYS18; WMJ*19; ZWLC19], **none** [ARO*17; CWZ*19; CYL*20; dSBD*12; ESD*19; HKBE12; JSR*19; KAS*20; KBJ*20; KEV*18; KPN16; LGM*20; LLS*18; LSL*17; SFB*20; SKB*18; WSW*18; XCK*20]
		Time Needed	min = 20, max = 43200, σ = 7870.90383664717, avg = 1547.96666666667, med = 56, |N/A| = 38
	Participants	Domain/Dataset Expertise	**low** [BHZ*18; BSP20; DLW*17; MQB19; SLC*20], **mid** [CD19; CMQ20; GLC*19; HKBE12; MLMP18; XCK*20], **high** [ARO*17; BSP20; CHH*19; CRH*19; CVL*18; CYL*20; DSKE20; DVH*19; EKC*20; ESD*19; ESS*18; GZL*20; JSR*19; KAKC18; KBJ*20; KEV*18; KPN16; LGM*20; LLL*19; LPH*20; LSC*18; LSL*17; MP13; MXC*20; SKB*18; SSBC19; SSKE19; XMT*20; XXM*19], **study condition** [ESKC18; SMD*16], **N/A**
		ML/AI Expertise	**low** [BSP20; DLW*17; KAY*19; SDMT16; SLC*20], **mid** [BAL*15; BHZ*18; CMQ20; DSKE20; HHC*19; MQB19; RAL*17; SDMT16; WMJ*19; YGLR20; ZWLC19], **high** [BSP20; CYL*20; GZL*20; KAKC18; KEV*18; KPN16; LLS*18; LSC*18; LSL*17; LXL*18; MCZ*17; MLMP18; MXC*20; MXLM20; WBL*20; WGSY19; WGYS18; WGZ*19; WSW*18; XXM*19], **study condition** [CWZ*19; ESKC18; ESS*18; LLL*19; PNKC20; SSSE20], **N/A**
		Background	***free text***, **N/A**
		Age	Minimum age: min = 10, max = 32, σ = 5.06, avg = 22, med = 22, |N/A| = 54 Maximum age: min = 13, max = 74, σ = 15.36, avg = 46, med = 48, |N/A| = 55
		Gender	Male: 56.63%, Female: 42.90%, NB: 0.47%
		# Participants	min = 1, max = 199, σ = 33.21, avg = 15.99, med = 6, |N/A| = 4
	Tasks and Data	Analysis Task	**Explore** [BSP20; CD19; CHH*19; CMQ20; DSKE20; DVH*19; ESKC18; GZL*20; JSR*19; KAKC18; KPN16; LLL*19; LXL*18; MCZ*17; PZDD19; SDMT16; SSBC19; WGSY19; WGYS18; WSW*18; XCK*20; XMT*20; XXM*19; YGLR20], **Use** [ARO*17; BHZ*18; BSP20; CVL*18; DLW*17; DVH*19; ESKC18; GLC*19; GZL*20; HKBE12; KAS*20; KEV*18; LGM*20; LPH*20; SDMT16; SLC*20; SMD*16; SSBC19; SSKE19; XMT*20], **Understand** [BAL*15; BSP20; CWZ*19; CYL*20; DVH*19; EKSK18; ESKC18; GZL*20; HHC*19; JSR*19; KAKC18; KAS*20; KPN16; LSC*18; LSL*17; MCZ*17; MQB19; PNKC20; RAL*17; SSSE20; WGSY19; WGZ*19; WMJ*19; WSW*18; XCK*20; YGLR20; ZWLC19], **Diagnose** [BAL*15; BSP20; CRH*19; dSBD*12; GZL*20; KAKC18; KPN16; LGM*20; LLS*18; LSC*18; MCZ*17; MXLM20; PNKC20; SFB*20; SSSE20; WBL*20; XMT*20; ZWLC19], **Refine** [BAL*15; CRH*19; CVL*18; DLW*17; DVH*19; EKC*20; ESD*19; ESS*18; GZL*20; HHC*19; HOW*19; JSR*19; KBJ*20; LSC*18; LSL*17; LXL*18; MLMP18; MXC*20; PNKC20; SFB*20; SSBC19; SSSE20; WGYS18; WMJ*19], **Hypothesize/Simulate** [CRH*19; DSKE20; HHC*19; KAY*19; MP13; MQB19; WGSY19], **Compare** [BAL*15; BSP20; CHH*19; DLW*17; ESKC18; KEV*18; MLMP18; MXC*20; SDMT16; WGSY19; WMJ*19; XCK*20; YGLR20], **Justify** [HHC*19; KAY*19; KEV*18; LPH*20; SKB*18; WGZ*19; ZWLC19], **Train** [CHH*19; ESKC18; HKBE12; LPH*20]
		Data Types	**Text Data** [ARO*17; BAL*15; CVL*18; EKC*20; ESD*19; ESKC18; ESS*18; HKBE12; JSR*19; LLL*19; MCZ*17; MXC*20; SFB*20; SKB*18; SSBC19; SSKE19], **Geo** [PZDD19], **Images** [CRH*19; CYL*20; KBJ*20; LLS*18; LSC*18; LSL*17; SSSE20; WGSY19; WGYS18; WGZ*19; XCK*20; XXM*19], **Video** [GLC*19; KAY*19; SMD*16], **Multivariate Data** [BHZ*18; BSP20; CD19; CMQ20; CWZ*19; DLW*17; dSBD*12; DSKE20; DVH*19; EKSK18; GZL*20; HHC*19; HOW*19; KAKC18; KAS*20; KEV*18; KPN16; LGM*20; LPH*20; LXL*18; MLMP18; MP13; MQB19; MXLM20; PNKC20; RAL*17; SDMT16; SLC*20; WBL*20; WMJ*19; WSW*18; XMT*20; YGLR20; ZWLC19], **N/A**
Model Properties and Explanations	Model Properties	Quality	**N/A**, **measured** [BAL*15; BHZ*18; DLW*17; EKC*20; ESKC18; ESS*18; GLC*19; LPH*20; LSC*18; LXL*18; PNKC20; PZDD19], **study condition** [RAL*17], **measured condition** [dSBD*12; HKBE12], **motivated** [GZL*20; KAKC18; MCZ*17; SKB*18; SLC*20; SSKE19; WGSY19; WGYS18]
		Observed Quality	**N/A**, **measured** [DLW*17; ESD*19; PNKC20; PZDD19; RAL*17], **study condition** [dSBD*12], **measured condition** [CVL*18; ESKC18; HKBE12; MLMP18; SMD*16], **motivated** [ESS*18; GZL*20; KAS*20; KEV*18; LPH*20; ZWLC19]
		Transparency	**N/A**, **measured** [GLC*19], **study condition**, **measured condition** [KAS*20], **motivated** [CRH*19; CWZ*19; DSKE20; EKSK18; ESD*19; ESS*18; GZL*20; MQB19; PNKC20; WMJ*19; WSW*18; XCK*20; ZWLC19]
		Trustworthiness	**N/A**, **measured** [CRH*19; CWZ*19; DLW*17; HHC*19; HKBE12; SFB*20; SSBC19], **study condition**, **measured condition** [CVL*18], **motivated** [ESD*19; ESKC18; ESS*18; KAKC18; KAS*20; KBJ*20; LSL*17; MCZ*17; MQB19; WBL*20; WMJ*19; XCK*20]
		Interpretability	**N/A**, **measured** [CWZ*19; DLW*17; DSKE20; EKSK18; GLC*19; HHC*19; SFB*20; XCK*20], **study condition** [ESKC18; KAS*20], **measured condition** [CVL*18; YGLR20], **motivated** [BAL*15; BSP20; CHH*19; CRH*19; CYL*20; EKC*20; ESS*18; GZL*20; JSR*19; KAKC18; KEV*18; KPN16; LLL*19; MCZ*17; MQB19; SLC*20; WGSY19; WGYS18; WSW*18; XMT*20; ZWLC19]
		Controllability	**N/A**, **measured** [DSKE20; WGSY19], **study condition** [ESKC18; SFB*20], **measured condition** [SEH*18], **motivated** [BSP20; CHH*19; CRH*19; EKC*20; ESD*19; GZL*20; HKBE12; JSR*19; KEV*18; LPH*20; PNKC20; SLC*20; WBL*20; WMJ*19]
	Explanations	Transparency	**N/A**, **measured** [DLW*17; EKSK18], **study condition**, **measured condition**, **motivated** [BSP20; GZL*20]
		Trustworthiness	**N/A**, **measured** [BAL*15; DLW*17; EKSK18; HHC*19], **study condition**, **measured condition** [CVL*18], **motivated** [ESKC18; LPH*20]
		Effectiveness	**N/A**, **measured** [CWZ*19; ESKC18; GLC*19; KAY*19; SKB*18; WSW*18], **study condition**, **measured condition** [CVL*18; SFB*20], **motivated** [BSP20; CHH*19; DVH*19; EKSK18; GZL*20; PNKC20]
		Fidelity	**N/A**, **measured**, **study condition**, **measured condition** [CVL*18], **motivated** [BSP20; MQB19]
Interactions and Guidance	Model Manipulation	Direct / Indirect	**direct** [BSP20; CMQ20; CVL*18; dSBD*12; DVH*19; EKC*20; ESD*19; ESKC18; GLC*19; GZL*20; HKBE12; HOW*19; KAKC18; KAS*20; KBJ*20; KPN16; LLL*19; LPH*20; LSL*17; MCZ*17; MP13; MXC*20; MXLM20; PZDD19; SDMT16; SFB*20; SKB*18; SMD*16; SSBC19; WBL*20; XMT*20; YGLR20; ZWLC19], **indirect** [ARO*17; CHH*19; CWZ*19; DLW*17; EKSK18; ESS*18; JSR*19; KEV*18; LGM*20; LSC*18; LXL*18; MLMP18; MQB19; PNKC20; WMJ*19; WSW*18; XCK*20; XXM*19], **both** [BAL*15; BHZ*18; CRH*19; CYL*20; DSKE20; HHC*19; KAY*19; SLC*20; SSKE19; WGSY19; WGYS18], **N/A**
		Interaction Type	***free text***, **N/A**
		Impact	***free text***, **N/A**
	Guidance	Time/Phase	**data selection** [WBL*20; YGLR20], **data preprocessing** [BHZ*18; SMD*16; WBL*20], **training** [ARO*17; BAL*15; CHH*19; CRH*19; DSKE20; EKC*20; ESD*19; ESKC18; ESS*18; GLC*19; HOW*19; JSR*19; KAS*20; KBJ*20; KEV*18; LPH*20; MP13; MXC*20; PNKC20; SKB*18; SLC*20; WGSY19; WMJ*19], **post-training** [BAL*15; CHH*19; CMQ20; CVL*18; CWZ*19; CYL*20; DLW*17; dSBD*12; EKSK18; GZL*20; HHC*19; HKBE12; JSR*19; KAKC18; KAY*19; KPN16; LGM*20; LLL*19; LLS*18; LSC*18; LSL*17; LXL*18; MCZ*17; MLMP18; MQB19; MXLM20; PZDD19; RAL*17; SDMT16; SFB*20; SSBC19; SSKE19; SSSE20; WGSY19; WGYS18; WGZ*19; WSW*18; XCK*20; XMT*20; XXM*19; ZWLC19]
		Frequency	**throughout**, **on-demand**, **N/A**
	Guidance	Degree	**orienting** [BHZ*18; BSP20; CVL*18; CYL*20; DVH*19; ESD*19; GZL*20; JSR*19; KAKC18; SKB*18; SSKE19; WMJ*19; XCK*20; YGLR20], **directing** [CVL*18; DSKE20; KEV*18; LGM*20; MLMP18; SMD*16; YGLR20], **prescribing** [EKC*20; GLC*19], **N/A**
		Knowledge Gap	**data** [BHZ*18; BSP20; CVL*18; CYL*20; DSKE20; GZL*20; KAKC18; KEV*18; SKB*18; SMD*16; XCK*20], **task** [EKC*20; ESD*19; GLC*19; JSR*19; MLMP18; SKB*18; SSKE19; YGLR20], **VA method** [BSP20; DVH*19; ESKC18; SKB*18], **user**, **infrastructure**, **N/A**
		Adaptation	**content** [CVL*18; DSKE20; GLC*19; GZL*20; SMD*16; SSKE19], **context** [JSR*19; SKB*18], **both**, **N/A**
Results		Main HCML Finding	***free text*** (see Table 3), **N/A**
		UI Feedback	***free text*** (see Table 4), **N/A**
		Interaction Feedback	***free text*** (see Table 5), **N/A**

Table 2: ***The Surveyed Dimensions, their values, and coding results****: For all non-free-text dimensions we summarize our results by listing corresponding references. N/A values and free text comments are not included; any other value without reference indicates that it was not found.*

protocols (e.g. [BSP20; CD19; HHC*19; SDMT16; XXL*20]. No significant change in trend across the years has been detected with quantitative data collection representing only 9% (e.g. [BHZ*18; CWZ*19; DLW*17; SEH*18]) overall. Studies using a mix of qualitative and quantitative data gathering protocols represent the remaining 26% (e.g. [BAL*15; CD19; HKBE12; LXL*18; RAL*17]). No significant difference was found between evaluation supervised versus unsupervised methods, with both favoring qualitative approaches over quantitative ones.

Learning Phase

Definition: The learning phase identifies the type and amount of training provided to participants before they interact with the system being evaluated.

Values: Unguided Exploration, Structured Training, Walkthrough, None.

The aim of a learning phase in human-centered evaluation is to reduce the chance that participant interaction with the experimental setup might be influenced by confounding effects such as lack of clarity of task requirements and task execution, insufficient familiarity with the study interface, or other elements pertinent to study infrastructure. HCML approaches are interested in investigating understanding and interpretation. Thus training phases need to balance the amount of information provided to participants against the potential introduction of bias towards a system, technique, or model. Our survey highlighted walkthrough as the preferred training method overall (38%). When participants had prior knowledge in either domain, dataset, or ML, structured training was used (14%) [CVL*18; EKSK18; LPH*20]. Unguided exploration was employed as an alternative to a walkthrough for those cases where participants had high levels of competencies and familiarity with core aspects of the study (13%) [GZL*20; PNKC20; XMT*20]. A large number of studies did not report training information 35%.

Time Needed

Definition: Total time needed to complete a study. Time can be average study completion time, average time per task, or fixed when allocated as part of the study design.

Values: Free text, format: min, hr, etc.

Completion time clustered between two ranges with the majority of studies taking between 30 min to 1 hr (e.g., [EKC*20; SDMT16; SSBC19]), followed by studies lasting between 90 min to 3 hrs (e.g., [MXLM20; WBL*20; XXM*19]). Studies with a total completion time of less than 30 min were crowdsourced studies (e.g. [SFB*20]). Two studies lasted for 24 hrs [CHH*19; CWZ*19] and one for four months [KPN16], with the former being a lab study and the latter a long term observation study.

4.1.2. Participants

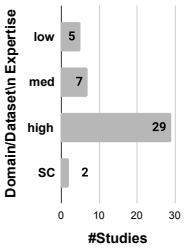
A core factor in human-centered evaluation is the clear profiling of participants. In human-centered machine learning, the depth of such profiling is even more complex. Elements belonging to the user's personal, private, and social spheres are likely to influence interaction with the model. Among others, these include the propensity to trust, differences between trust in humans and machines, prejudice built from previous experience, confidence, and self-esteem. In our survey, we did not find studies that performed any considerable evaluation of such traits. Participant profiling remains limited, focusing mainly on skills and knowledge. We have also found limited reporting of details related to reducing bias and balancing diversity in dimensions such as gender and age.

Domain and Dataset Expertise

Definition: This dimension distinguishes between participant expertise and familiarity with the problem domain and/or dataset used during evaluation.

Values: Low, Mid, High, Study Condition, N/A.

Our analysis revealed a clear distinction between Domain versus Dataset expertise. The former implies knowledge and understanding of the essential aspects of a specific field, the latter, in the context of HCML, implies knowledge and familiarity with the specific dataset(s) under investigation. The majority of studies (62%) reported values for this dimension detailing expertise levels (e.g. [KPN16; LSL*17; MLMP18; MXC*20]) and distribution across levels (e.g., [CVL*18; SMD*16]). Few papers use this dimension as controlled study condition, comparing results across participants' expertise levels (e.g., [ESKC18]).

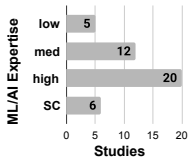


ML/AI Expertise

Definition: This dimension reports participant expertise with respect to machine learning models and/or their development.

Values: Low, Mid, High, Study Condition, N/A.

Technical expertise of the participant was reported both at the level of expertise with respect to ML models, their development, as well as with respect to ML interactive systems and framework development. Similar to the case for Domain & Dataset expertise, this dimension was reported in the majority of studies (62%). The dimension often appeared as a study condition (e.g., [CWZ*19; ESS*18; LLL*19]), or as criteria used for participant segmentation together with Domain & Dataset expertise [CYL*20; KAKC18; KEV*18].



Age

Definition: The participants' age range or average age.

Values: Tuple (min, max), single numerical value, N/A.

Only 20% of the surveyed studies reported participant age ranges, with some including standard deviation (e.g. [WMJ*19]). Age range and distribution represent important information to explore data and feedback related to the perception of Model specific categories such as Explanations and Model Properties.

Gender

Definition: This dimension reports summary statistics about the gender of participants.

Values: Free text, N/A

In the context of machine learning, bias can be introduced by lack of representation of demographic categories. In our surveys only 31% of the total studies reported gender distribution, with 3 studies reporting equal distribution of male and female participants (e.g.,

[CD19; EKSK18; GLC*19]), and 2 studies reporting non-binary or unspecified gender participants (e.g. [GLC*19; SFB*20]).

Number of Participants

Definition: Total number of participants who completed a study and who were accounted for in the study analysis.

Values: Numerical value.

All surveyed studies reported the total number of participants. Where applicable, studies also differentiated between the total number of recruited participants versus the total number of participants considered for eventual analysis. In those cases, authors reported details on exclusion criteria that were applied to filter out participants (e.g., [CVL*18; CWZ*19; SFB*20; SMD*16]).

4.1.3. Tasks and Data

The task and data dimensions represent characteristics of HCML approaches and their evaluations. Based on the analysis tasks and data types, we can directly compare different systems and paper contributions. Hence, we use these two dimensions in the next two sections to discuss techniques and application domains. Section 5 describes **technique**-focused evaluations in relation to identified tasks. Section 6 describes **application**-centered evaluations in relation to the data types considered.

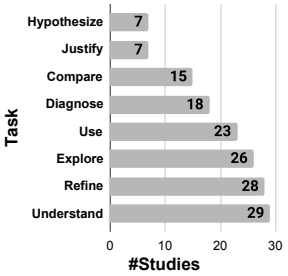
Analysis Task(s)

Definition: The main analysis task(s) participants worked on during the evaluation.

Values: Understand, Diagnose, Refine, Compare, Explore, Use, Hypothesize, Justify.

HCML approaches usually target one or more tasks from the data and visual analytics pipelines, ranging from configuring and training a machine learning model to using it, comparing it to other models, or justifying its decisions. Thus, we surveyed the reported tasks in our paper set and grouped them into the values described above. Specifically, we focused on the tasks that were performed during the evaluation described. For instance, if a tool is motivated to help users *refine* a model, but the evaluation only tested the comparison of a model, it is categorized under *compare*. If tools were evaluated for multiple tasks, each of these tasks was coded. While several task taxonomies for information visualization exist (e.g., [BM13; vLFB*14]), they do not appropriately capture several typical HCML tasks. The tasks listed here build on work by Liu et al. [LSL*17] and were iteratively compiled during paper coding.

Following the data and visual analytics pipelines, we start with tasks performed during training, where models get iteratively **refined** (e.g., [EKC*20; ESS*18; KBJ*20]); **diagnosed** (e.g., [LLS*18; MXLM20; WBL*20]); and **compared** (e.g., [CHH*19; DLW*17; MLMP18]). Followed by tasks performed post training, where model results are **explored** (e.g., [BSP20; CMQ20; KAKC18]); **understood** (e.g., [CWZ*19; MQB19; XCK*20]); and **used** (e.g., [DSKE20; HHC*19; HKBE12]). Another task after training is the **refinement** of results (e.g., [BAL*15]). In some cases, participants were asked to **hypothesize** (e.g., [DSKE20; KAY*19]) and provide **justifications** (e.g., [SKB*18]). We did not find many HCEs in our paper collection that tackled tasks in the data selection or preprocessing phases. The most prominent tasks we have found during coding are use, explore, and understand.

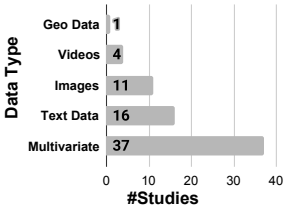


Data Type(s)

Definition: The main data type(s) the system is designed to use.

Values: Multivariate Data, Text, Images, Video, Geographic Data.

The data type(s) used in each system is another relevant dimension for comparing human-centered evaluations. The predominant data types are **multivariate** data (e.g., [PNKC20; WMJ*19; XMT*20]); **text** (e.g., [ARO*17; ESD*19; LLL*19]); and **images** (e.g., [CRH*19; LSL*17; SSSE20]). Only very few papers use other data types like **videos** (e.g., [KAY*19]) or **geo** data (e.g., [PZDD19]).



4.2. Model Properties and Explanations

Visual analytics and HCML are characterized by the integration of human intuition within automated machine learning and artificial intelligence. However, increasingly powerful models easily become infamous "*black boxes*" and novel research fields like XAI have been developed that aim to explain model decisions in support of the user's analytical and decision-making process. We were interested in studying how previous evaluations of HCML systems have dealt with different properties of models and explanations, how they correlated, and, in particular, if and how they were evaluated.

In addition to statistics, we provide short definitions for all dimensions. One of our findings is that there does not seem to be a standardized definition used systematically across different studies. The following section thus draws from existing literature on ML/XAI and aims to provide unified definitions from the perspective of HCML researchers, although agreed-upon definitions used consistently throughout the community remain challenges for future work. All dimensions are coded along four values: *measured* and *study condition* are used for dependent and controlled variables, respectively. Some studies evaluate a participant's perception of controlled variables; these cases are coded as *measured conditions*. More frequently, dimensions are *motivated* throughout the paper, but not evaluated.

4.2.1. Model Properties

This set of dimensions focuses on properties of the models themselves. The emphasis is on what aspects or properties are shown or explained to users. Additionally, the result of visual analytic tools showing these to users is often motivated by specific outcomes (e.g., trustworthiness, interpretability, etc.) described in this section.

Quality

Definition: Model quality is typically represented by *accuracy* or *F-score* and determines the correctness of the model at performing the task it was trained for.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension aims to characterize the actual quality of a model. This is also commonly referred to as correctness. Although a high accuracy is often desired of machine learning models, the focus of many HCML papers is on explaining a model or creating models that respond to user preferences instead of emphasizing model quality. We thus found that only 18% of studies measured the accuracy of the model(s) (e.g., [EKC*20; LXL*18; PZDD19]). Of these, the measure of accuracy is often studied and derived from direct comparison to ground truth data [BAL*15], or varied among study conditions [DLW*17]. This can be either benchmark datasets or datasets where experts provide data labels.

Perceived Quality

Definition: Perceived quality describes the model quality that users can observe. Notably, in the context of a study, it can be manipulated to differ from the actual model quality.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension captures the extent to which the quality of the model is exposed and observable to users. Examples include directly showing quality or allowing users to interactively explore aspects of models that allow implicit assumptions about model quality to be made (e.g., [HKBE12; KAS*20; LPH*20]). However, when interactivity is involved, this can lead to situations where the ground truth is based on domain-relevant information as opposed to verified labels (e.g., [DXG*20]). Overall, we found that only 7% of studies considered observable quality as a measured condition (e.g., [CVL*18; ESKC18; HKBE12; MLMP18; SMD*16]).

Transparency

Definition: A model is transparent when all its inner workings and decision-making processes can be observed and understood by users.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension focuses on how the transparency of models was specifically communicated and evaluated in a system or study. This is a common focus of HCML papers, and often consists of showing the mechanisms of the models themselves (e.g., [KAS*20; LPH*20]). While early work in (X)AI equated model transparency with the presence of an explanation, later work found that transparency might be overwhelming [PGH*21]. We found that while many tools were motivated to improve the transparency of the underlying model (e.g., [GZL*20; WMJ*19]), only two studies (from the same paper) measure the transparency of the proposed tools [GLC*19].

Interpretability

Definition: A system is interpretable when users can understand why it behaves in a given way under given circumstances.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

According to our definition, interpretability can be considered an inductive process, where users first create a mental model of the system and then verify whether the system is consistent with that mental model, making it interpretable. Lipton [Lip18] has previously surveyed interpretability and suggests "*that interpretability is not a monolithic concept, but in fact reflects several distinct ideas*." Improving the interpretability of models was a common motivation for papers included in this survey (e.g., [CHH*19; EKC*20; MCZ*17]). However, only 14% of papers in this survey measured whether interpretability was achieved. Studies that evaluate interpretability often test how well people can communicate their internalized understanding of how models make decisions (e.g., [HHC*19; KAS*20; SLC*20]), often through qualitative responses from participants.

Trustworthiness

Definition: A model can be considered trustworthy when users believe it is *correct*.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension captures to what extent users subjectively trust the outputs or decisions made by the models that are used in the tools. Improving user trust is another common motivation for many papers in this survey (e.g., [ESS*18; KBJ*20; WBL*20]). However, only seven studies (10%) measured whether and how the proposed tools affect user trust. Methods that have been used to capture this dimension include participant self reports Likert scales [CRH*19; CWZ*19; DLW*17; HKBE12; SFB*20], think-alouds [HHC*19], and interviews [SSBC19]. Likert scales are a particularly common method adopted across multiple papers. This suggests potential for a consistent evaluation methodology for measuring model trustworthiness in future studies and can contribute to the comparability of user trust across multiple studies.

Controllability

Definition: A system is controllable when it affords interactions that allow users to manipulate it such that they can correct decisions or modify its behavior so that it matches their expectations.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension focuses on the extent to which users are able to control and provide feedback to the models. Although 20% of studies were motivated to provide controllability for users, only two studies measured whether controllability was achieved. Both papers reported qualitative responses from participants [DSKE20; WGSY19]. The ability for domain experts to control and correct model outcomes or processes is often motivated by use cases where the decisions being made are critical and can have detrimental outcomes if not seriously considered (e.g., healthcare [CRH*19] and fraud detection [SLC*20]).

4.2.2. Explanations

The four *explanation* dimensions–*transparency*, *trustworthiness*, *effectiveness*, *fidelity*–focus on the properties of explanations gener-

ated for an ML model to describe its decision-making process to the users. Overall, these dimensions are usually included in a paper to motivate the proposed work. The evaluation of explanations in HCML has been relatively limited; encouragingly, there seems to be a rise of interest in evaluation in recent years [EKSK18; SFB*20].

Transparency

Definition: The condition of an explanation being generated such that it is easy for users to examine the process.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

In contrast to *model* transparency defined above, this dimension characterizes whether *explanations* of the model are transparently generated and easily examinable for users. Four papers considered this dimension. Improving the transparency matrix reordering algorithms was the major motivation for a system supporting users in investigating the expressiveness and usefulness of such algorithms [BSP20]. Similarly, Gou et al. [GZL*20] motivate the need for a visual analytics system to help assess deep learning-based object detectors, going beyond aggregated metrics that fail to capture important context that could affect user understanding. Only two of the surveyed papers measured the transparency of explanations [DLW*17; EKSK18]. Eslami et al. [EKSK18] investigated and measured how communicating parts of the algorithm process of selecting relevant advertisements could affect users' perception towards ads, while Dasgupta et al. [DLW*17] measured analysts' "*levels of trust*" in visualizations that could lead to more transparent analysis processes. Both papers report the perceived explanation transparency through qualitative feedback and participant's quotes. Dasgupta et al. [DLW*17] also discussed transparency as an important design criterion to increase explanation trustworthiness.

Trustworthiness

Definition: The ability for the explanation to be believed in or accepted by the user as an honest representation or correct description.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension considers to what extent a user would accept an explanation as a faithful representation of the model's decision-making process. Note that this dimension is different from the trustworthiness of the model itself: consider how a model that performs poorly may be considered untrustworthy, but an explanation of that model may still be highly accurate and considered trustworthy. Thus, we report on the trustworthiness of models and explanations separately. In our survey, we found only four studies (6%) that measured user trust in the system explanation. Dasgupta et al. [DLW*17] discussed design criteria to increase user trust of a visual analytics system and evaluated user trust on a Likert scale. In contrast, three studies evaluated user trust qualitatively through participant feedback regarding the explanations they received [BAL*15; EKSK18; HHC*19]. In particular, Hohman et al. included an insightful discussion on user trust of model explanations. They observed that "*participants were eager to rationalize explanations without first questioning the correctness of the explanation itself [...] this could be troublesome when participants trust explanations without healthy skepticism*" [HHC*19]. Only one paper included trustworthiness as a measured condition by enabling and disabling an explanatory visualization [CVL*18].

Effectiveness

Definition: The degree to which the explanation is successfully conveying the decision-making process of the model.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

Among the four *Explanation* dimensions, *effectiveness* was included by the highest number of papers, fifteen in total. The *effectiveness* dimension characterizes how well an explanation conveys the model's decision-making process. While many papers were motivated to improve explanation effectiveness [CHH*19; DVH*19; EKSK18; GZL*20; PNKC20], only six papers went to measure whether explanation effectiveness was achieved [CWZ*19; ESKC18; GLC*19; KAY*19; SKB*18; WSW*18].

Fidelity

Definition: The faithfulness, thoroughness, and degree of exactness with which the explanation represents the model's decision-making process.

Values: N/A, Measured, Study Condition, Measured Condition, Motivated.

This dimension characterizes the explanation's faithfulness in representing the model's decision-making process. None of the papers in our survey measured or evaluated explanation fidelity in their studies. However, three papers considered this dimension in their motivations and study design. RuleMatrix [MQB19] was motivated by the benefit of using a simpler rule-based surrogate model instead of a high-fidelity (but likely unintelligible) representation of a "*black box*" model. GUIRO [BSP20] used approximate 2D projections of a matrix to help users understand matrix patterns and different matrix reordering algorithms. Coppers et al. [CVL*18] reduced their Intellingo system's fidelity to create a "*simple*" version by disabling an explanatory visualization component.

4.3. Interactions and Guidance

Interactions and guidance are key elements in HCML, as they enable and drive this co-adaptive process. However, as the results below show, the impact of interactions tends to be evaluated more qualitatively than other areas of the HCML process.

4.3.1. Model Manipulation

Interaction techniques provide a means for users to manipulate and obtain further information about the model. This section characterizes the different types of interaction techniques we came across in papers, as well as the impact of the interactions (if discussed).

Directness

Definition: This dimension characterizes whether direct manipulation or indirect manipulation of the model is available in the HCML system/interface.

Values: Direct, Indirect.

We found a good distribution of both direct and indirect manipulation in the papers we surveyed, with most systems using only of both types. Direct manipulation is often used for selecting and filtering data instances (e.g., [CYL*20; KAKC18; KBJ*20; MXLM20;

WGYS18]), while indirect manipulation techniques tend to be used for data labeling (e.g., [BHZ*18; XXM*19]) or model specification and hyperparameter tuning (e.g., [CHH*19; DLW*17; JSR*19; KEV*18; PNKC20; SLC*20; XXM*19]). These tasks are not exclusive to each interaction manipulation technique. For example, there were cases where filtering was done using indirect manipulation (e.g., [MQB19; PNKC20]). Direct and indirect manipulation techniques provide different affordances and outcomes [Shn97], and a number of systems used a combination of both direct and indirect manipulation techniques (e.g., [CRH*19; DSKE20]). The impact of using different manipulation techniques has not been studied in the context of HCML and represents an avenue of potential future work.

Interaction Type

Definition: The different ways users can interact with a system to affect the model.

Values: Free response, e.g. "*drag and drop to merge clusters*"

While papers describe interactions at varying levels of detail, a few common interaction types emerged in our analysis, including selection (e.g., [KBJ*20; WBL*20]), filtering (e.g., [LGM*20; PNKC20]), zooming (e.g., [LPH*20; LSC*18]), tuning weights/hyperparameters (e.g., [DSKE20; SLC*20]), and annotation/labeling (e.g., [SSKE19; XXM*19]). Multiple of these interaction types tend to be combined in a single system, for example when users first select a set of data instances before labeling them (e.g., [BHZ*18; ZWLC19]). Less frequently, we also observed more complex combinations of interactions to enable application-specific operations such as cropping an image to the area of interest [CRH*19], aligning video frames [KAY*19], or placing game level elements[GLC*19]. In several cases, these are observed by systems and used as implicit feedback towards an underlying model.

Impact

Definition: The extent to which the achieved results of the model can be attributed to changes made via interaction.

Values: Free response, e.g. "*improved model quality shown in independent ranking study*"

While some studies evaluate the impact of interactions on system outcomes, most do not. Studies that evaluate interaction impact tend to focus on usability, reporting qualitative user feedback about the interactions implemented (e.g., [CRH*19; KEV*18; WSW*18]). A small number of papers also evaluate interaction as a study condition, comparing interactive explanations with static explanations. One such paper by Cheng et al. [CWZ*19] found that "*Interactive interfaces increased both objective understanding and self-reported understanding of the algorithm.*" This suggests that interactions play a crucial role in the design of HCML. However, further evaluation will be necessary to understand how interactions should be designed and the magnitude of their impact on HCML systems.

4.3.2. Timing

Interactions can be designed for different phases of the machine learning process. Some systems require constant input from the users, others only provide further information on demand. In this section, we provide details about when interactions are exchanged between users and systems and how much user input is required.

Phase

Definition: This indicates the machine learning phase during which interactions are provided to the user.

Values: Data Selection, Data Preprocessing, Training, Post-Training.

Machine learning is an iterative multi-step process from data selection and data preprocessing to training and post-training. Most of the papers we surveyed were situated during the training and post-training phases of the process. This makes sense since this is when the machine learning model is introduced. Only four papers focus on the data selection and preprocessing steps (see Section 5.1). Wang et al. [WBL*20] allow users to **diagnose** when the information of groups of users are at risk of exposure from inference attacks. Bernard et al. [BHZ*18] provide a system for users to label data instances in a visually interactive way. Finally, Yan et al. [YGLR20] implement Silva, a system that helps users **explore** and **understand** biases in their data sets and machine learning models.

Frequency

Definition: This indicates how often users interact with the system.

Values: Throughout, On-Demand, N/A.

Different systems require different amounts of input from the user. In systems where the user input drives the work of the system or the machine learning model, user interaction will be required throughout the functioning of the system (e.g., [CHH*19; SKB*18]). Other systems occupy a more informational role. Thus users only need to interact with the interface on-demand when they require specific information from the system (e.g., [CYL*20; KAS*20]).

4.3.3. Guidance

Guidance is a complex, co-adaptive process that aims to optimize the collaboration between machine and human [SJB*20]. As many novel guidance approaches are adaptive and learn from the user over time, they are a core component of HCML systems [CGM19]. However, the effect of guidance in general, and that of adaptive guidance in particular, is difficult to evaluate for a multitude of reasons: guidance is typically only used in complex, non-trivial scenarios that are not easily replicated in study environments and aims to close a knowledge gap of a specific user. Consequently, between-subject studies are difficult, as independent problems of the same complexity to a given user are required [BM18]. 33% of evaluations mentioned the guidance provided by the system. However, none of the evaluations specifically aimed to evaluate guidance and typically captured participant feedback in qualitative comments instead.

Knowledge Gap

Definition: Ceneda et al. [CGM*17] describe the knowledge gap in terms of information that the user is missing in order to make progress during the analysis.

Values: Data, Task, Visual Analytics Methods, User, Infrastructure, N/A.

We observed only three of the knowledge gaps (data, task, visual analytics methods) proposed by Ceneda et al. [CGM*17] being tackled by systems. Independent of the knowledge gaps, the provided guidance is typically validated through qualitative feedback. To bridge the data knowledge gap, 55% of systems that provided guidance

(e.g., [KAKC18; KEV*18]) propose automatically identified subsets or features based on a definition of interestingness [CGM*17]. Das et al. [DSKE20] evaluate the guidance provided by their system for iterative clustering by highlighting the interpretability of recommended clusters and report that users were uncertain why they were shown certain suggestions. Task-driven guidance supports users by "*hinting at what to do next*" [CGM*17]. Evaluations of this guidance consider its "*usefulness*" [EKC*20], interestingness [SKB*18], and highlight the potential to unsettle experts whenever recommendations do not align with their mental model [SSKE19].

Degree

Definition: Ceneda et al. [CGM*17] state that the "*guidance degree specifies the extent to which guidance is required and actually provided*" and characterize it in terms of three values.

Values: Orienting, Directing, Prescribing, N/A.

69% of papers that mentioned guidance in their evaluation described systems with orienting guidance (e.g., [ESD*19; SSKE19; WMJ*19]) and 18% and 9% report insights into directing [KEV*18; LGM*20; SMD*16] and prescribing guidance [EKC*20; GLC*19], respectively. As guidance is not typically evaluated in-depth, we observed no difference between the qualitative feedback for different guidance degrees. Gou et al. [GZL*20] present a system for the iterative refinement of image classifiers, providing orienting guidance to users using semantic representations of the data space. Their qualitative domain expert feedback reveals that "*they found the tool's capability to aid them in targeting weak spots is* 'specifically useful*', and visual summary was* 'the most useful feature'" [GZL*20]. Bernard et al. [BHZ*18] provide an extensive, quantitative comparison of visualization techniques for providing orienting guidance in interactive labeling. Using a within-subjects design, each participant evaluated four design alternatives and provided feedback on a five-point Likert scale. None of the surveyed papers provide a comparative evaluation between different guidance degrees.

Adaptation

Definition: Adaptation describes how, if at all, the guidance considers implicit or explicit user feedback to improve the relevance of the provided suggestions over time.

Values: Content, Context, Both, None, N/A.

Out of the 23 studies evaluating guidance, seven systems adapt in terms of content (e.g, [CVL*18; DSKE20; GLC*19; GZL*20]) and two in terms of analysis contexts in which guidance is provided [JSR*19; SKB*18]. None of the papers provide both types of adaptation simultaneously. Guzdial el al. [GLC*19] quantitatively investigate whether participants notice adapting guidance agents in game level design, and whether they prefer the adapted or original version. Their findings suggest that participants do notice and prefer adaptation. In the context of argumentation annotation, Sperrle et al. [SSKE19] suggest fragments of text for annotation that are similar to existing annotations. Their evaluation reveals that participants were aware of this adaptation, and only start to trust the learned suggestions late in the annotation process.

4.4. Results

Many evaluations provide primarily qualitative feedback that is difficult to categorize in fixed dimensions. To capture as much information as possible, we also included three free-text dimensions for each paper to document the feedback in three areas: human-centered analytics, user interface feedback, and interaction feedback. When coding these dimensions, we aimed to include quotes from the respective papers whenever concise descriptions were available. When this was not the case, we paraphrased the paper to obtain clear summaries. Below, we report topic modeling results for all three dimensions. We use the Incremental Hierarchical Topic Model (IHTM) [ESD*19] as it does not require a pre-determined number of topics and allows the integration of domain knowledge in the form of a topic backbone (see supplementary material). Providing a backbone primes the model to expect certain topics but does not force their creation to prevent model manipulation. The IHTM hierarchically clusters documents based on their cosine similarity, directly assigning the most probable topic to each document. In our case, a document consists of one finding or comment from a paper. As a result, papers are typically represented by multiple documents that can be assigned to different topics. Hence, semantically diverse annotations can accurately be represented by assigning papers to multiple topics. Before running the topic model, we process all documents, keeping only nouns, proper nouns, adjectives, and verbs. The inclusion of adjectives and verbs enables the model to capture frequent descriptions of participant actions during evaluations. Furthermore, we remove common English stopwords, as well as the words *system, help, perform, allow, support, design* and *expert* that are ubiquitous in our dataset.

Main Findings in HCML

Definition: The main finding(s) in HCML reported in a paper, e.g., in terms of result quality, user satisfaction, or model intelligibility.

Values: Quotes where concise descriptions were present, otherwise summarized findings.

Main findings cover a broad range of topics summarized in Table 3. It is interesting to note how the emerging topics relate to core tasks in HCML, including understanding, interpreting, and explain. Topics we identified as dimensions of either model properties or explanations that were not mentioned as part of the main findings include, amongst others, verifiability, transparency, fidelity, and control. Overall, topics emerging in main findings are not straightforward to interpret. Nevertheless, it provides a useful exercise to highlight those themes of interest to HCML research that still remained unexplored.

Interface Feedback

Definition: Any participant feedback concerning the interface design or usability.

Values: Quotes where concise descriptions were present, otherwise summarized findings.

In this category, we grouped topics related to feedback provided with respect to the system/model interface. Topics are summarized in Table 4. Analysis showed overlap with main findings as well as two emerging topics: Findings and Actionability. Both topics highlight user-expected behavior from an interface. It is interesting to note that actionability in several cases implied actionable communication

Topic	Keywords	Papers
Workflow	***workflow***, explore, reflect, attack, algorithm designer, typical, framework, retrieval, personal, exploit, reflect typical workflow	[BSP20; CD19; JSR*19; MXLM20; PZDD19]
Quality	cluster, ***quality***, user, intelligibility, available, data, suggestion, statistical, useful, sampling, actionable insight, feel excited, meaningful	[BAL*15; BHZ*18; CD19; CRH*19; CVL*18; DSKE20; EKC*20; ESD*19; ESS*18; GLC*19; GZL*20; HOW*19; KAS*20; KEV*18; PZDD19; RAL*17; SDMT16; SKB*18; SSBC19; WSW*18]
Debugging	***debugging***, training process, suggestion, visual, nn training, ad, aspect, communicate, perception, curation, algorithmic, speed	[EKSK18; LSL*17; SSKE19]
Refinement	machine learn, ***refinement***, interactive, agree optimize model, automl, tune, search, practical, sense, diagnoses, give groundtruth structure	[ESKC18; SLC*20; SSSE20; WMJ*19]
Understanding	***understanding***, improve, understand, human, children, representation, capable, visual analytic, ml, diagnose, text, able, demonstrate	[ARO*17; CWZ*19; GLC*19; GZL*20; HOW*19; KAS*20; KAY*19; KPN16; LGM*20; LSC*18; MCZ*17; WBL*20; WGSY19; XMT*20]
Interpretability	interpret, achieve, predictive, ***interpretability***, performance, science, similar, evaluate, boost, nlg, scientist assess interpretability	[HKBE12; KPN16; LLL*19; LXL*18; MXC*20; WGYS18]
Explainability	explain, ***explanation***, explainable, visualization, frustration, decision, develop, get, path, role, capability	[CYL*20; HHC*19; KAKC18; MQB19; SFB*20; XCK*20; YGLR20; ZWLC19]
Trust	***trust***, prefer, feedback, increase, improvement, level, substantial, expect, perceive, compare, comfort-zone, convey, acceptance reduce explanation	[CD19; CVL*18; CWZ*19; DLW*17; LPH*20; RAL*17; SFB*20; SMD*16]

Table 3: ***Topics based on Main HCML Findings****: IHTM results on a collection of sentences from the reported finding sections in our paper collection. Each paper can be attributed to multiple topics. The descriptive topic titles where manually assigned based on the provided keywords.*

Topic	Keywords	Papers
Findings	meaningful, relation, pattern, ***find***, utility, seem, think, explore, educational, effective	[KAS*20; KAY*19; WGSY19; WSW*18]
Explanations	reassuring, ***explanation***, case, way, ecosystem, output, reassure, ai, provide, annoying	[HHC*19; KAY*19; KEV*18; XCK*20]
Interpretations	information, ***interpretation***, blast, sufficient, overwhelming, enlightening, dataset, know	[BAL*15; WGYS18]
Actionability	observation, ***action***, react, straightforward, accuracy, sanity, fix, improvement strategy aim, understand, feature, model, require, visualization, split, graph, important, improve	[CVL*18; ESKC18; GZL*20; HHC*19; KAY*19; MQB19; SKB*18; WGSY19; WSW*18; XCK*20; XMT*20; ZWLC19]

Table 4: ***Topics based on Interface Findings****: IHTM results based on the reported findings related to the interface design or usability.*

representative of active engagement in the interaction process for which the interface is the primary mediator.

Interaction Feedback

Definition: Any participant feedback concerning the interaction design or system workflow.

Values: Quotes where concise descriptions were present, otherwise summarized findings.

We intentionally separated feedback related to interaction from feedback related to the interface. Topics are summarized in Table 5. All topics emerging in this dimension have no overlap with previous dimensions. It is worth noting how controllability, a dimension of model properties, is reported as part of interaction feedback. Lack of ways to provide inputs to model training was explicitly reported as a limitation of the system by at least one paper [KAS*20].

This section has introduced the dimensions in our methodology and observations about how they have been evaluated, focusing on individual aspects of evaluations. In Section 5 and Section 6, we highlight evaluations from the perspectives of task-specific techniques and domain-specific characteristics, respectively.

5. Evaluating the Technique Contributions of HCML

Application papers focus on domain-specific challenges and provide solutions to expert users from their respective fields. As a result, there are different expectations of success, making application papers more difficult to compare across domains. Technique papers, on the other hand, focus on providing solutions to general analysis tasks, such as exploration, efficient filtering, or model refinement. Where possible, these techniques are expected to be applicable across different application domains or even data types. As a result, we expect the findings from evaluations of systems employing these techniques to be transferable to other domains. In this section, we survey evaluations that do not rely on domain-specific knowledge. Following the approach by Yuan et al. [YCY*20], we group the works by the stages of the machine learning process, and their respective tasks.

Topic	Keywords	Papers
Interactivity	***interactivity***, process, play, select, create filter, feature, learn, appreciate, element	[CHH*19; HHC*19; LGM*20; SSBC19; YGLR20]
Customization	***custom*** collapsible unit, define ***custom***, specify, manipulation, construct, compare, find, complex exploration, model, experiment	[CRH*19; DSKE20; EKC*20; HHC*19; LLL*19; SDMT16; WSW*18; YGLR20]
Controllability	focus, ***control***, feedback, important, go, click, think, menu, time, relate, input, provide	[DSKE20; EKSK18; KAS*20; SFB*20; XCK*20]
Comprehension	***comprehend***, put, understanding, reduce, remain, physician, black, visualization pipeline	[SDMT16; XCK*20]

Table 5: ***Topics based on Interaction Feedback****: Identified from the reported findings related on interaction design or system workflow.*

5.1. Pre-Training

Only a few collected papers provide evaluations that fall into this phase: two approaches for efficient data labeling and two approaches for analyzing inferences between attributes. We observe no differences between supervised and unsupervised machine learning.

Data Labeling – In a within-subjects study, Bernard et al. [BHZ*18] compared active learning (AL) with visual interactive labeling (VIL) across different support techniques, dataset complexities, and datapoint selection strategies. For their three-part study, they provide detailed descriptions of the setup and dependent and independent variables for all experiments. They conclude their study with interviews to obtain subjective feedback on the usefulness and preferences for different support techniques. They find that VIL was competitive with AL and that class coloring and convex hulls were most useful at supporting users during labeling. Sarkar et al. [SMD*16] let users rank sets of videos to obtain a relative scoring in a within-subjects design with two conditions. To minimize learning effects, they keep a minimum of three days between conditions for each participant. Upon completion of a condition, participants fill out NASA TLX forms, as well as open-ended questions.

Inference Analysis – Two systems aimed to assess potential inferences in training data: one from the perspective of algorithmic fairness [YGLR20], and one to prevent privacy leaks [WBL*20]. Yan et al. [YGLR20] compared their system for assessing fairness against an existing industry solution in a controlled user study. They reported that participants found their system to be more useful (self-report on Likert scale) and that they made more true-positive discoveries. In addition, they provide qualitative comments collected

after the study. Wang et al. [WBL*20] conducted an expert review, including an interactive system demo from which they report brief, qualitative feedback.

5.2. During Training

The primary user task during model training is refinement. While our coding results show that further tasks like diagnosis or comparison (e.g., [BAL*15; MXC*20; WGSY19]) are performed during model training, they are not explicitly evaluated.

Refine – During the training phase, refinement is typically performed through iterative model optimizations. We observe a distinction between the refinement of supervised and unsupervised models in the evaluation methodologies. Working with unsupervised models a series of three related papers by El-Assady et al. [EKC*20; ESD*19; ESS*18] presented human-centered approaches for topic modeling explainability and refinement. All three papers used similar evaluation methodologies based on the same document corpus. The evaluations all consisted of two stages. First, a pair-analytics study in which three groups of experts from different domains used the respective visual analytics approaches to refine topic models. In the second stage, an annotation study, the obtained model refinements were scrutinized by independent annotators to confirm their quality. The empirical results from the first stage, such as perceived model transparency or controllability of the process, were presented as qualitative feedback, while the second stage used the intuition and understanding of independent annotators to provide quantitative results. In the context of our survey, this series of papers showcases a unique example of comparable study setups. Additionally, all three papers set the users' expertise as a study condition to understand the influence of the users' backgrounds on the topic modeling refinement results.

Krueger et al. [KBJ*20] presented a tool for semi-automatic phenotype analysis, where a human in the loop drove the analysis and steered both clustering and classification models. While they relied primarily on case studies to evaluate their approach, they highlighted expert requests for direct access to the raw data to build trust during the refinement process. Similarly, Ming et al. [MXC*20] reported that experts felt that the interactivity of the system made it easier to interpret model results.

Use – In a less common form of HCML, users interact with a system without explicit training intent, aiming to perform a domain-specific task like designing a game level [GLC*19] or annotating text fragments [SSKE19]. The respective systems gathered all interaction data and utilized it to learn the users' preferences and adapt future guidance suggestions. Both evaluations reported that participants were able to observe the adaption over time. However, from the users' perspective, this refinement was a side effect and not the primary analysis goal.

5.3. Post-Training

The majority of evaluations identified in our survey cover analysis tasks in the post-training phase.

Explore – Exploration is the task most frequently referenced in evaluation descriptions. Many papers primarily focused on case studies

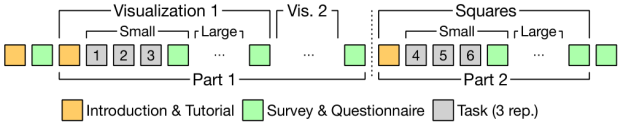


Figure 4: *Graphical representation of the study methodology used by Ren et al. [RAL*17].*

(e.g, [XXM*19]), use cases (e.g., [CMQ20; LLL*19]) and questionnaires (e.g., [WSW*18]). Stahnke et al. introduced probing as "*a general interaction approach for information visualization that is aimed at both exploring the data as well as examining its representation*" [SDMT16]. They developed a prototype application that enabled exploration of embedding spaces and the errors introduced during dimensionality reduction. They evaluated their system using a lab experiment and found that participants had difficulties in reading errors introduced by distortion and interpreting data point positions.

Understand – Ren et al. [RAL*17] run a controlled, within-subjects lab experiment to evaluate at what speed users can assess the quality of multi-class classifiers. As one of few papers, they summarize their experiment design in a figure (see Figure 4): first, they compare their system against an interactive confusion matrix and measure task completion time across three locating tasks and measure task completion times. They find that participants are not only significantly faster using their tool but also strongly prefer the tool and find it more helpful than the confusion matrix.

Hohman et al [HHC*19] run a lab study to explore how participants investigate and understand ML models. Rather than prescribing a dataset to be used during the study, they let each participant select a dataset of their preference. They find that "*participants were eager to rationalize explanations without first questioning the correctness of the explanation itself*" [HHC*19] and that system interactivity was fundamental to the participant's model understanding. Hitron et al. [HOW*19] study how children can learn the basic building blocks of machine learning and find that repeated blocks of data labeling and evaluation were essential to constructing an accurate understanding.

Ming et al. [MQB19] developed RuleMatrix, a visualization system that explains classification models through rule induction. In a user study testing how well participants understood the model, one participant expressly commented that he liked how the system supported hypothesis testing. An open question for explaining classification models with rules is the fidelity of the rules generated. The authors discuss the "*trade-off between the fidelity and complexity*", where more rules are required to explain a machine learning model with higher fidelity. However, visualizing more rules would also affect the interpretability of the system. As such, more work is required to improve the scalability of rule induction techniques for explaining classification models.

Diagnose – Several systems have been proposed to diagnose various machine learning models [LDM*18; LGM*20; LLS*18; LSC*18; MXLM20]. Zhao et al. developed iForest, an interactive visual analytics system that helps "*users interpret random forest models from various perspectives*" [ZWLC19]. In particular, visualizing the decision paths helped participants explain the predictions for

certain data instances. iForest enables users to tweak feature values to ask *what if* questions and explore "*how feature values affect predictions*" [ZWLC19]. Participants' qualitative feedback about this interaction was not reported in the paper. However, the ability for users to ask "*what-if*" questions remains an interesting feature that can be further developed in HCML systems.

Refine – Smith-Renner et al. [SFB*20] conducted an extensive crowdsourced experiment to investigate the relationship between explainability and interactivity in machine learning. They found that, for low-quality models, the availability of explanations increased frustration, while support for feedback reduced it. Further, "*trust and acceptance were reduced by explanations and increased by feedback*" [SFB*20]. Participants in conditions that allowed for user feedback expected stronger model improvements than those that did not provide feedback. However, interestingly, they found that some participants were under the impression that the model was learning from their feedback, while this was not possible in the study setup. In a second study with a higher-quality model, they found no significant effects between user feedback, explanations, and frustration. To summarize, they report that "*participants felt strongly that the opportunity to provide feedback was important*" [SFB*20] and that explanations and user feedback complement each other.

6. Application-Specific Evaluations

Most HCML papers apply their proposed techniques to tackle problems and challenges from specific application domains. Hence, in addition to evaluating a technique contribution, their evaluations often reflect the suitability of the approaches to specific users, tasks, and data. In this section, we report on the predominant clusters of application-specific evaluations in our paper collection. These highlight trends in recent research efforts, as well as open gaps for future work. This section showcases the most dominant application domains, as well as the prevailing data types.

6.1. Bio-Medical Applications

One of the largest clusters of human-centered evaluations in our paper collection are eight papers reporting on biomedical applications. These typically present approaches where human involvement and trust are critical for utilization. As a result, several papers report bespoke, complex evaluation methodologies that go beyond more traditional pair analytics [KBJ*20] and observational studies [CYL*20; DSKE20; LPH*20] that are also used.

Dasgupta et al. [DLW*17] report on a controlled lab experiment that aims to evaluate how much users trust a bespoke analysis tool versus a traditional data analysis tool of their choice (like R or Excel). To avoid learning effects commonly observed during expert evaluations, they opt for a between-subjects design. They find no significant difference in trust when participants perform retrieval tasks. For two out of four complex interpretation tasks, the bespoke tool leads to significantly higher trust levels. Further tests including participant experience revealed significant differences in trust levels only for participants with little experience, favoring the visual analytics tool.

Cancer Image Analysis – In the domain of medical image analysis, Cai et al. [CRH*19] present a tool for interactive image retrieval based on user-refined concepts. To evaluate the end-user experience while using the system, they run a counterbalanced within-subject lab study, comparing their tool to a prototypical system used in cancer image analysis. Aiming to analyze the effects of imperfect algorithms on user perception, they report results from six Likert-scale questions, as well as empirical feedback collected during participant tool usage. To ensure realistic task complexity, they select images that have received conflicting diagnosis labels from pathologists prior to the study.

Radiology – To enable physicians to understand automatic X-ray image analysis through artificial intelligence, Xie et al [XCK*20] present CheXplain. The authors follow a human-centered design process and first elicited current analysis practices from 77 medical practitioners through questionnaires before iteratively co-designing the tool with three physicians. Finally, they evaluate the tool in a remote observational study through video calls in which participants are asked to describe "*their understanding of why AI arrived at certain results of the case*" [XCK*20]. The obtained qualitative comments are transcribed and iteratively tagged before they are independently reviewed and conflicts are resolved through discussion.

Diabetes – Kwon et al. [KAS*20] describe results from a long term collaboration with clinical researchers. Over the period of one year, they held four quarterly workshops in four locations. During each workshop session, they performed pair-analytics to analyze progression trajectories of type 1 diabetes. Over this time, they elicited relevant clinician tasks and adapted the system accordingly. After all workshops were complete, they conducted unstructured interviews with nine participants and collected qualitative feedback. Similarly, Krause et al. [KPN16] performed a long-term case study with five domain experts aiming to predict diabetes from patient records. Over four months, they held bi-weekly meetings to co-design the system and ensure that it met the experts' requirements.

6.2. Machine Learning Applications

All HCML algorithms are machine learning applications by definition. In this section, we cover tools and systems that are designed to support researchers and practitioners in machine learning in understanding, diagnosing or refining their models. Across domains and data types, a number of applications are evaluated using observational studies [KAKC18; MCZ*17] and expert case studies that combine domain-specific findings with feedback interviews [LSC*18; LSL*17; WGSY19; WGYS18; XXM*19]. In their pair analytics study, Spinner et al. [SSSE20] collect qualitative feedback and provide a fine-granular coding into eight groups pertaining to user expectations, supported tasks, and interface feedback. The responses are coded according to grounded theory and encoded into compact glyph tables that are shown as

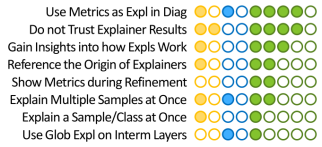


Figure 5: *Glyph table based qualitative feedback coding [SSSE20]. Each column represents one participant and color different ML experience levels. Circles are filled if a participant gave a corresponding answer and hollow otherwise, providing a detailed overview in a compact format.*

wrapfigures next to the text (see the extracted example in Figure 5).

AutoML – With ATMSeer, Wang et al. [WMJ*19] provided the first visual analytics approach for controlling AutoML processes. They state that the lack of existing systems for comparison made it difficult to establish a "*fair*" baseline model that was not drastically inferior. Consequently, they instead deem it "*more interesting and important to investigate user behavior under the characterized workflow with ATMSeer*" [WMJ*19]. In an observational study with thirteen participants, they evaluate their workflows through the system and what functionality is used or ignored. Eight small proxy tasks requiring participants to locate some information in the system are used as a proxy for system usability. From participant feedback, they also find that nine participants (69%) would consider a familiar model over a better-performing, unknown one. Park et al. deploy their AutoML system HyperTendril internally at an AI research company and collect usage data for four months [PNKC20]. They then invite the three most active users in that period to interview sessions in which they revisit their previous AutoML refinements completed with the system. The interview reveals that the integration of human domain knowledge can lead to higher-quality models in a shorter time.

Adversarial ML – Liu et al. [LSC*18] analyze the noise robustness of image classification networks and evaluate their system using an expert-driven case study, revealing mostly domain insights. Ma et al. [MXLM20] focus on explanations of vulnerabilities that are evaluated through a group expert interview following a case study. While the system receives positive feedback, they also find that its complexity causes long learning times.

6.3. Linguistics Applications

A number of approaches address questions related to linguistic application and text analysis. In the following, we describe some of the works with respect to their presented human-centered evaluations. Applications range from analyzing discourses through annotating argumentation structures [SSKE19] and untangling and reconstructing threaded conversations [ESKC18], to analyzing and comparing prosodic patterns in different languages [SKB*18]. Besides a clearly structured quantitative evaluation that tackles model intelligibility in machine translation by Coppers et al. [CVL*18], most linguistic application papers in our collection are related to the thematic analysis of text data, as described in the remainder of this section.

Information Retrieval – Heimerl et al. [HKBE12] compare three methods for classifier creation to retrieve documents. Their study comprised multiple phases in which the twelve participants were introduced to the models and then got to use two of them for fulfilling the analysis task. This evaluation is particularly interesting as it is one of few studies in our data set that directly compares more than one analysis alternative. In another information retrieval application, Ji et al. [JSR*19] introduced a system for exploring neural document embedding based on semantic features. The system is evaluated by presenting use cases that are based on expert feedback. Experts were required to configure document maps, specify clustering parameters, and analyze dimension correlations. This evaluation relies on a close collaboration with selected experts. Similarly, Abdul-Rahman et al. [ARO*17] rely on a long-term collaboration with domain experts, who were involved in providing system feedback during the development process and in the observational study they conducted for evaluation. Their approach, ViTA, is a system for visual text similarity detection based on constructive text alignment. Such evaluations provide unique insights into the specialized applicability of the presented tools and describe their benefit for expert users.

Text Classification – Brooks et al. [BAL*15] developed FeatureInsight, a system to help users ideate features for text classification. The system is evaluated using a controlled, within-subjects study with four treatments. Participants ideate four different feature sets and complete a questionnaire about enjoyment and satisfaction after each condition. The resulting features are used to train classifiers that are compared based on precision-recall curves, finding significant differences between some conditions.

Topic Model Refinement – Applying topic modeling algorithms to analyze the content of text corpora is a prevalent task in the humanities and social sciences. However, as the results of these unsupervised algorithms are typically subjective and domain-dependent, there is no single ground truth that could be used to optimize such models. Hence, refining these models has relied on humans externalizing their knowledge to adapt the results to their domain understanding. To tackle this challenge, El-Assady et al. proposed three different workflow and interface designs, each tailored to a different user group. For domain experts, they apply a machine teaching paradigm to capture the experts' knowledge [EKC*20]. Data scientists can use a progressive learning approach [ESS*18], and machine learning experts can look into the model [ESD*19] and see the impact of their interactions before applying them using *speculative execution* [SBS*18]. As described in Section 5.2, these three related papers were presented as a consecutive series using comparable study designs. To remedy the lack of ground-truth data, all papers were evaluated in two stages; first, a pair analytics study, followed by an annotation study where independent coders blindly rated their perceived quality of the refined and unrefined topic modeling results.

6.4. Others

Lastly, besides the more prominent domain clusters presented above, we also encountered other diverse application fields in our data set. This section describes a selection of papers that are particularly interesting from the point of view of human-centered evaluations.

Game Design – Guzdial et al. [GLC*19] present multiple iterations of a system for cooperative *Super Mario Bros.* level design between a human and an AI agent. The first prototype builds on simple, existing models from related work: To build a level, the agent and the user take turns adding (or removing) blocks to a level. In a comparative study, 84 participants interact with two different agents to build two levels. Afterward, they rank which agent was more fun, frustrating, challenging, helpful, or surprising and which one they would like to use again. Furthermore, they provide qualitative comments. Using the gathered interaction data from this study, the authors train a new agent that adapts through implicit feedback: it receives penalties when users delete its blocks and a bonus otherwise. In the second study, 24 participants design two levels using this agent, experiencing continuous adaptation. In addition to the five questions from the first study, the second study aims to evaluate to what extent participants felt that the agents were adapting, collaborating with them

and whether participants would have preferred to use the system without agent support. The study finds that the agents adapt to participants and that the adaptation is noticeable to participants.

Video Analysis – Kittley et al. [KAY*19] aimed to investigate the intelligibility of different processing steps throughout the computer vision pipeline. They task study participants of their controlled lab study with designing a stop-motion video, as this task can be supported through explainable keyframe matching and is "*not obvious to participants*" [KAY*19]. Each participant completes four tasks that are designed to showcase the system (1), reveal algorithm failure conditions (2), and assess user understanding (3+4). Finally, participants complete a semi-structured interview. Their answers are coded by three independent annotators and analyzed in detail according to predefined research questions. Additional quantitative analysis reveals significant differences between the perceived understandability of the four different study conditions.

Fraud Detection – In cooperation with a bank, Leite et al. [LGM*20] develop an environment to explore fraudulent behavior and state that "*trustability, reputation, security and quality are the main concerns for public and private financial institutions*" [LGM*20]. They derive four design requirements with collaborating experts and evaluate the success of the system using four tasks, each targeting one requirement. Rather than focusing on trust-aspects motivated early on in the paper, the evaluation considers comparisons to other systems (like Excel or Tableau), insights, and potential improvements. To measure insights, they annotated different interaction patterns and counted their occurrences as insights. The motivation of human-centered aspects that are not explicitly evaluated is a common pattern that appears in many papers. Sun et al. [SLC*20] evaluate their system for fraud detection with both expert and novice users. An expert interview reveals that experts have higher confidence in their diagnosis when using the system but does not elaborate on the baseline.

7. Survey Insights and the Next Frontiers

To put our findings in a broader context, we summarize all lessons learned, insights, and challenges that we uncovered by surveying human-centered evaluations in HCML. We, further, deduce a set of guidelines and best practices, and report on research gaps and opportunities. We conclude with a call to action to the research community.

Summary – This survey provides three different perspectives on the body of existing work, as it is intended to serve various readers with their diverse questions. In Section 4, we describe **evaluation dimensions** that can both serve as structuring elements to researchers planning evaluations, as well as a concise resource to identify gaps in current evaluations. In Section 5, we summarize which **analysis tasks** are evaluated throughout the machine learning pipeline, focusing on the evaluation of technique contributions. Finally, in Section 6, we provide a domain-specific view on evaluations, identifying differences and similarities between **domains and data types**.

Limitations – Our survey only considered human-centered evaluations of HCML. This particular focus on the human in both systems design and evaluations limits the number of papers considered in this survey, as many approaches to HCML that were evaluated with use cases or algorithm-centered approaches were excluded. While these works are equally relevant in the field of HCML, this survey reveals the complexity of human-centered evaluations that draw from different methodologies of various fields like machine learning, explainable artificial intelligence, human-computer interaction, and psychology. Future work should extend this survey to non-human-centered evaluations to obtain a complete picture. Further, this survey only includes papers that could be found in the respective digital libraries, as described in Section 3. In particular, we use a predefined keyword list and do not consider papers that were published before 2012. While both the nomenclature and scope of the field have evolved over time, there is a significant body of work on adaptive systems that we had to exclude. Charting the field's evolution over time represents another challenge for future work.

7.1. Findings and Observations

Overall, we found that the surveyed papers tend to focus on different aspects of HCML systems. Some papers focus on the tools' implementation, some on their design, and some on user studies and evaluations. Our observations do not suggest that authors had not considered all aspects when they developed their systems. Rather, we believe the different emphases during reporting of evaluations was a main factor that created challenges when comparing research findings or contributions across papers.

Evaluation Methodologies and Procedures – In general, we did not observe a standardized methodology throughout the surveyed papers. Most evaluations were diverse in their format, as well as their content. Hence, it was challenging to converge to a meaningful coding structure that was able to capture these diverse methodologies, as is evident from Figure 3. In addition, due to HCML systems' complexity, we observed that in several instances, papers are motivated using factors that should influence the analysis but do not consider them in the subsequent evaluation that happens at a different level on Munzner's nested model [Mun09]. Most papers evaluate the systems they represent by conducting user studies in the lab. There are few longitudinal or in-the-wild studies. In addition, most evaluations were conducted with a low number of experts, which may limit the generalizability of findings to real-world usage.

In addition to evaluating the system and ML/AI performance, in HCML, human factors are, by definition, of central importance. In our surveyed paper corpus, we did not find many evaluations with a nuanced consideration of human factors. We thus derive that there is a need for interdisciplinary collaboration with fields, such as psychology and social sciences, to derive insightful results on the diversity of human impact on such collaborative systems.

Evaluation Reporting – Overall, we have not identified a general reporting structure among the surveyed works. We have observed incomplete descriptions in some cases. For example, in some papers, it was not entirely clear who the study participants were, what instructions or tasks they were given, or what the study intended to measure. In addition, the majority of papers reported more than one evaluation, with frequent combinations including case studies and expert interviews, as well as algorithm-centered and human-centered feedback. 86% of papers report qualitative results, while only 37% report quantitative values. Lastly, there was a small number of papers that did present use cases as case studies or case studies as use cases. Based on these observations, in Section 7.3, we derive a number

of guidelines and best practices in an attempt to help future paper authors avoid common pitfalls.

Supervised and Unsupervised Models– We observed similar numbers of qualitative and quantitative studies across evaluations of both supervised and unsupervised ML, with a predominance of qualitative (63%) or mixed (26%) evaluations. Notably, several papers belonging to areas of unsupervised learning did not qualify for the present survey as they lacked human centred evaluations. In particular, our survey contains few examples of dimensionality reduction [dSBD*12], and even less with comprehensive human-centered evaluations [SDMT16]. While the inherent complexity of dimensionality reduction might be a reason, other unsupervised model classes like topic modeling are successfully evaluated using human-centered approaches. In general, we observe a tendency for evaluations of unsupervised HCML to focus on quality metrics related to the model or user interaction with the system (e.g., through parameter selection or output exploration). User interactions directly influencing the underlying machine learning models are evaluated less frequently.

7.2. Challenges in Human-Centered Evaluations

In Section 2.3, we have collected a list of challenges in HCML. This section presents how these are manifested in associated challenges for human-centered evaluations.

HCE-C1: Exploratory Analysis. A typical challenge in interactive machine learning is the lack of ground truth data [BBL18]. In HCML, many systems are explicitly designed to promote personalized results that match a user's requirements. Such personalization makes the comparison with a gold standard–if available–complicated or even irrelevant. Several papers provide multi-stage evaluations to tackle this problem, as described in Section 6.3, by collecting the results obtained during an expert study and passing them on to other participants to evaluate, rank, or validate them. This downstream, outside perspective is either obtained using the evaluated system itself or externally to mitigate several possible issues like, for example, novelty effect that can bias qualitative feedback.

HCE-C2: Participants. Expert-level systems typically cannot be evaluated at the same scale as non-expert systems due to the limited availability of participants. Non-expert systems can and should be evaluated at scale – for each approach, it has to be determined where it falls on the scale and what an appropriate evaluation is. Whenever few participants are available, combinations with case studies and quantitative evaluations are common. In some cases, it might be possible to extract parts of the system that can be evaluated without the integration of experts – however, careful assessment of the results is needed to avoid overgeneralizing any findings back to expert-level users. While expertise is the factor most commonly considered (domain expertise: 43/71 papers, ML/AI expertise: 43/71 papers), other factors like personality can also be influential. Personal characteristics of interest include, among others, the propensity to trust, differences between trust in humans and machines, prejudice built from previous experience, confidence, or self-esteem. However, we did not observe studies that performed a considerable evaluation here. This is related to **HCML-C3** (Co-Adaptation) and **HCML-C4** (Stakeholder Diversity).

HCE-C3: Evaluation Focus. To effectively evaluate a HCML system, researchers must consider methods that adequately test diverse aspects of these systems (**HCML-C1** (Interdisciplinarity) and **HCML-C2** (Complexity)). For instance, testing only the usability of the interface may miss out on the ability of the system to interpret user input to create more accurate models. Similarly, only measuring model accuracy misses out on evaluating the user experience. The challenge lies in finding the right balance between human-centered and algorithm-centered evaluation. A large number of papers surveyed in this STAR tackle this issue by providing both types of evaluations. However, as the number of excluded papers shows, there is still a bias towards algorithm-centered evaluation. We encourage researchers to be creative and investigate methodologies that can provide human-centered insight without sacrificing algorithm-centered or domain-specific insights, for example, using repeated study designs or evaluations combining several smaller, focused studies. Large, multi-faceted projects built on collaborative efforts across disciplines (e.g., HCI, visualization, ML, and humanities) can also benefit from a series of tailored publications, each evaluating a different perspective in detail, rather than providing only superficial insight in one overly complex paper. While novel HCI aspects can be evaluated using a large-scale user study, complex domain-specific workflows might require case studies with lower participant counts.

7.3. Guidelines for Human-Centered Evaluations

As HCML systems are becoming increasingly complex, it is typically infeasible or even impossible to evaluate all aspects in detail. From the papers we surveyed, it was often difficult to extract what the main findings with respect to HCML were. Rather than providing generic expert feedback, we encourage authors to select specific areas to be evaluated and clearly state which aspects are not evaluated. To provide a more actionable plan for researchers preparing and conducting human-centered evaluations, as well as reporting their results, in this section, we describe guidelines and best practices based on our observations from the current state-of-the-art for human-centered evaluations in HCML.

Guidelines for Conducting HCE Studies – To assist researchers in planning out their evaluations, in Table 6, we have collected guiding questions. These are structured in the form of a seven-point ***checklist*** that covers all aspects that can help shape the studies to be successfully conducted. We expect such a "*cheat sheet*" to be especially helpful in supporting junior researchers to cover all aspects for preparing their evaluations. In particular, for evaluations that rely on qualitative feedback, we highlight the importance to clearly define the main hypotheses and research questions. To that end, researchers also have to decide which parts of the system might be better evaluated with an algorithm-centered evaluation, as opposed to a human-centered one, or a combination of both.

Best Practices for Study Reporting – As there is currently no accepted default structure for reporting results of human-centered evaluations, study descriptions are difficult to compare and do not always include all necessary information. The machine learning community has recently suggested *model cards* for model reporting [MWZ*19] and *datasheets for datasets* [GMV*20] to summarize the respective most important aspects. For the InfoVis community, Borgo et al. have generated a template form to "*be filled in and provided as supplementary material to adequately report all the*

details of an InfoVis crowdourcing experiment" [BMB*18]. Building on this idea, we provide a ***template*** for the reporting of human-centered evaluation results in Table 7.

While it is not our goal to prescribe that every evaluation must consider specific properties, we encourage authors to use the template to help them in systematically capturing all aspects worth considering and reporting. Further, using the dimensions introduced by this STAR will ensure that new human-centered evaluations are comparable to the ones studied here. In the following, we directly address future authors, and recommend a set of aspects to help improve the clarity of the reported results:

1. Clearly **state the research questions** that you are aiming to investigate (e.g., [BHZ*18; HHC*19; LGM*20]).
2. **Structure the evaluation** section to support readability, e.g., [HHC*19; SFB*20]. Highlight key findings.
3. If applicable, **represent a study's phases** as a figure, such as provided by Ren et al. [RAL*17] (depicted in Figure 4).
4. Identify and **report what** properties or aspects **are being evaluated**, and how, e.g., usability, learning, knowledge generation, user satisfaction, interaction design, workflow, etc.
5. Provide definitions for abstract concepts like trustworthiness or intelligibility.
6. Consider providing **statistics about participants**, e.g., age and gender. They are relevant to scope potential replication studies.
7. Describe the **background of participants** (where were they born, raised?; where do they live?), as many human-centered aspects depend on cultural backgrounds.
8. Be precise in describing the **expertise levels** that you report, especially considering the two dimensions of domain/dataset and ML/AI expertise.

1) Hypotheses and Main Questions to Investigate
☐ Are the research questions defined?
☐ Human-centered vs. algorithm-centered evaluation?
2) Study Setup
☐ Which study type is most appropriate?
☐ How are results processed? Qualitative or quantitative?
☐ Which analysis or coding methods are applied?
3) Tasks and Dataset
☐ What tasks do participants perform?
☐ What can be measured through those tasks?
☐ Which dataset will you use?
☐ Do participants need training? How can it be provided?
☐ How do you verify that participants understand tasks?
4) Data Collection
☐ What data (screen, audio, interactions, . . .) will be collected?
☐ How will it be stored in a privacy-preserving way?
☐ Do you have to ensure GDPR-compliant data handling?
5) Ethical Clearance
☐ Does the study require ethical approval?
☐ How will participant consent be collected?
6) Participants
☐ How can the required expertise be ensured?
☐ Has diversity in backgrounds been considered?
☐ Screening: inclusion and exclusion criteria
☐ Are participants compensated? If yes, how?
7) Reproducibility and Open Sourcing
☐ How can reproducibility of results be ensured?
☐ Will the results of the study be publicly available?

Table 6: ***Checklist*** *of factors to consider when planning a human-centered evaluation, ensuring all relevant information is captured.*

7.4. Research Opportunities for Human-Centered Evaluations

We believe an exciting research direction is for the community to develop a *structured evaluation framework*, which will improve many HCE facets, including synergizing HCE with algorithm-centric evaluations and increasing research reproducibility and accessibility.

Structured Evaluation Framework – To advance and mature the field of HCML, a structured evaluation framework that enables comparable evaluations is needed. Our STAR can serve as a first step in this direction, providing an overview of current human-centered evaluation practices. In order to converge towards a complete evaluation framework, algorithm-centered evaluations must also be surveyed, and guidelines developed on how to balance both integral aspects of HCML. The surge of interest among major companies in understanding the best practices in human-AI interaction (e.g., from Apple [App19], Google [Goo19], Microsoft [AWV*19], and summarized and compared by Wright et al. [WWP*20]), provides fertile common ground for researchers to collaborate with industry practitioners to make advances on all fronts.

Shared Vocabulary – As a foundation for a common evaluation framework, convergence on a shared vocabulary is needed. In this paper, we have coded terms like *trustworthiness* and *transparency* as mentioned by the respective authors. However, it is not clear whether all authors have the same definition of these abstract terms. As a result, this STAR can not provide unifying definitions, instead highlighting them as challenges that remain for future work. Future HCE evaluations should consider providing definitions for evaluated terms and concepts or reference existing definitions.

Open Access and Reproducible Studies – We call on all authors to "*open-source*" their evaluations, to help advance the reproducibility of research results, and to help the community more easily access high-quality evaluation approaches, and to contribute to improving them. Indeed, open-sourcing evaluation aligns well with the ongoing practices of open-sourcing code repositories in the machine learning and AI communities (e.g., top venues typically strongly encourage researchers to do so for their submitted work), and the increasing encouragement of posting final preprint versions of accepted articles to open access repositories (e.g., at IEEE VIS). Specifically, we encourage researchers and authors to make publicly available all the evaluation materials, such as the protocol transcripts of evaluation sessions. Even better would be to also open-source such materials so that other researchers can contribute to extending and improving them (similar to how contributions are made to code repositories, say, on GitHub). We believe a structured evaluation framework will accelerate the open-sourcing efforts and vastly increase the amount of evaluation results that can be compared, because the structure provides a common ground that naturally facilitates comparison and invites collaborative development; independent research teams

Study Setup	**Study Type**: ○ Observational Study ○ Pair Analytics ○ Lab Experiment ○ Field Study ○ Crowdsourcing Study **Study Design:** ○ Between-Subjects ○ Within-Subjects ○ Mixed-Design **Study Conditions & Controls:** 1. _____ 2. _____ 3. _____ ... **Study Duration:** ○ One Session ○ Multiple Sessions ○ Long Term | **Session Duration:** ____ mins
Research Objectives & Hypothesis	**Research Questions:** 1. _____ 2. _____ 3. _____ ... **Application Domain**: _____ | **Evaluation Goals:** ☐ Interface Design ☐ Interaction Design ☐ General Usability ☐ Model Performance ☐ Domain Insight ☐ Interaction Impact ☐ Technique ☐ Other: _____ **ML Properties**: ☐ Quality ☐ Transparency ☐ Trustworthiness ☐ Interpretability ☐ Controllability ☐ Other: _____ **Explanation Properties**: ☐ Transparency ☐ Trustworthiness ☐ Effectiveness ☐ Fidelity ☐ Other: _____ **Interface** Feedback on: _____ | **Interaction** Feedback on: _____ | **Guidance** Feedback on: _____
Methodology & Procedure	**Learning Phase:** ☐ Training ☐ Walkthrough ☐ Unguided Exploration | **Learning Phase Duration:** ____ mins **Study Phases (Duration):** 1. _____ (__ mins) 2. _____ (__ mins) 3. _____ (__ mins) **Structure** (per Phase): ☐ Structured ☐ Unstructured ☐ Semi-Structured | **Study Leadership:** _____
Tasks	**Main Task:** ☐ Refine ☐ Diagnose ☐ Compare ☐ Explore ☐ Understand ☐ Use ☐ Hypothesize ☐ Justify **Domain-Specific Task Description:** 1. _____ 2. _____ 3. _____ ...
Data	**Description**: _____ | ☐ Cleaned ☐ Processed ☐ Labeled ☐ Contains Ground Truth **Data Type**: ☐ Multivariate Data ☐ Text ☐ Images ☐ Videos ☐ Geographic Data ☐ Other: _____ **Availability**: ○ Open Source ○ Upon Request ○ Restricted | **Source**: _____
Participants	**Number of Participants**: ____ | **Demographics:** Age (min:__ max:__ mean:__ SD:__) Gender (#f:__ #m:__ #o:__) **Background**: _____ | **Education**: _____ | **Culture**: _____ | Other: _____ ☐ **Domain Expertise**: _____ | Distribution: (___% Low, ___% Mid, ___% High) ☐ **Dataset Expertise**: _____ | Distribution: (___% Low, ___% Mid, ___% High) ☐ **ML/AI Expertise**: _____ | Distribution: (___% Low, ___% Mid, ___% High)
Analysis & Reporting	**Data Collection**: Recordings (☐ Screen, ☐ Video, ☐ Audio) ☐ Tracking ☐ Questionnaire ☐ Protocol ☐ Sketches & Notes **Analytical Methods:** ☐ Summarizing Observations ☐ Statistical Measures ☐ Grounded Theory ☐ Others: _____ **Result Processing:** ☐ Qualitative ☐ Quantitative | **Result Presentation**: ☐ Tables ☐ Text Description ☐ Figures **Study Data Availability:** ☐ Raw Data ☐ Processed/Aggregated Data | ○ Open Source ○ Upon Request ○ Restricted
Findings	☐ **HCML** Findings: 1. _____ 2. _____ 3. _____ ... ☐ **Interaction** Findings: 1. _____ 2. _____ 3. _____ ... ☐ **Interface** Findings: 1. _____ 2. _____ 3. _____ ...

Table 7: ***Template*** *for reporting the results of human-centered evaluations, based on Borgo et al. [BMB*18]. Enhancing the reproducibility of study designs, this form can serve both, as a guide to structure evaluation reporting within papers, as well as a paper's supplementary material.*

could selectively adopt and extend specific parts that are relevant to them and contribute their extensions back to the "*source*" repository and to the community at large. For instance, open-sourcing the evaluation materials that focus on recruiting experts as participants could help researchers more easily extend it to also recruit novices by leveraging much of the structure of the existing study protocol, and focus mainly on modifying the parts that are critical to change (e.g., how participants are recruited and trained, develop simpler tasks).

To further increase the likelihood for evaluation results to be reproduced and compared, we encourage researchers to adopt repeatable methods like grounded theory [Dey99; IZCC08] to code transcribed qualitative feedback. We also believe that it will also be beneficial for evaluations to consider factors like personality characteristics or traits of the participants (which have not been considerably evaluated by the surveyed papers), such as their propensity to trust, confidence in interacting with technologies, and prior belief about machine learning; all of which could have an impact on how participants interact with a machine learning system.

7.5. Calls for Action

Lastly, we conclude our reflection with four calls for action (CFA) aimed at both authors and the scientific community.

CFA for Authors – (1) Try to make the aspects that were evaluated clear as HCML papers are becoming very complex, and it is infeasible to evaluate all of the dynamics at hand. Also, clearly state which aspects you did not evaluate. (2) Be clear on limitations of study results, including factors limiting generalization of findings and clearly distinguishing between statements supported by evidence and statements seeding future research directions.

CFA for the Community – (1) Consider splitting papers; do not expect complex HCML systems to be evaluated within the remit of a single paper; create a venue for HCML evaluations; value replication studies. (2) Value candid, realistic reporting of study results, especially with respect to limitations. The complexity of the human factors is exacerbated in HCML; we need to learn to accept small and perhaps constrained but sound steps pushing research forward.

8. Conclusion

This paper presents a comprehensive survey of evaluations of HCML techniques in visual analytics. For an interactive overview of our results, see the survey browser at https://human-centered-evaluations-star.dbvis.de. Evaluating these tools is complex, as there are human-focused aspects such as usability and utility in terms of task performance, as well as model performance metrics to measure model quality. This survey focuses on factors of the evaluation that may influence trust, interpretability, and explainability. Through systematic analysis of the evaluations performed in papers from relevant conferences and journals, we came up with a series of design dimensions to describe structured evaluations. Finally, we discuss gaps in evaluation methodologies and future opportunities to advance the research.

Acknowledgements – We would like to thank Thilo Spinner, Udo Schlegel, Mariem Mahmoud, Hassan Elassady, and Rita Sevastjanova for their help in preparing the manuscript. This project is supported in part by NSF IIS-1750474, EPSRC EP/R033722/1, and the DFG within grant number 455910360 (SPP-1999). Open access funding enabled and organized by Projekt DEAL. [Correction added on 08 November 2021, after first online publication: Projekt Deal funding statement has been added.]

References

[ACKK14] AMERSHI, S., CAKMAK, M., KNOX, W. B., and KULESZA, T. "Power to the People: The Role of Humans in Interactive Machine Learning". *AI Magazine* 35.4 (2014), 105–120. DOI: 10.1609/aimag.v35i4.2513.

[App19] APPLE. *Human Interface Guidelines for Machine Learning*. https://developer.apple.com/design/human-interface-guidelines/machine-learning/overview/introduction/. 2019.

[ARO*17] ABDUL-RAHMAN, A., ROE, G., OLSEN, M., GLADSTONE, C., WHALING, R., CRONK, N., MORRISSEY, R., and CHEN, M. "Constructive Visual Analytics for Text Similarity Detection". *Computer Graphics Forum* 36.1 (2017), 237–248. DOI: 10.1111/cgf.12798.

[AVW*18] ABDUL, A., VERMEULEN, J., WANG, D., LIM, B. Y., and KANKANHALLI, M. "Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2018, 1–18. DOI: 10.1145/3173574.3174156.

[AW18] ALVARADO, O. and WAERN, A. "Towards Algorithmic Experience: Initial Efforts for Social Media Contexts". *Proc. Conference on Human Factors in Computing Systems*. 2018, 1–12. DOI: 10.1145/3173574.3173860.

[AWV*19] AMERSHI, S., WELD, D., VORVOREANU, M., FOURNEY, A., NUSHI, B., COLLISSON, P., SUH, J., IQBAL, S., BENNETT, P. N., INKPEN, K., TEEVAN, J., KIKIN-GIL, R., and HORVITZ, E. "Guidelines for Human-AI Interaction". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 1–13. DOI: 10.1145/3290605.3300233.

[BAL*15] BROOKS, M., AMERSHI, S., LEE, B., DRUCKER, S. M., KAPOOR, A., and SIMARD, P. "FeatureInsight: Visual support for error-driven feature ideation in text classification". *IEEE Conference on Visual Analytics Science and Technology*. 2015, 105–112. DOI: 10.1109/VAST.2015.7347637.

[BBL18] BOUKHELIFA, N., BEZERIANOS, A., and LUTTON, E. "Evaluation of Interactive Machine Learning Systems". *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Ed. by ZHOU, J. and CHEN, F. Springer International Publishing, 2018, 341–360. DOI: 10.1007/978-3-319-90403-0_17.

[BCP*19] BROWN, A., CHOULDECHOVA, A., PUTNAM-HORNSTEIN, E., TOBIN, A., and VAITHIANATHAN, R. "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 41:1–41:12. DOI: 10.1145/3290605.3300271.

[BHZ*18] BERNARD, J., HUTTER, M., ZEPPELZAUER, M., FELLNER, D., and SEDLMAIR, M. "Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study". *IEEE Trans. Visualization and Computer Graphics* 24.1 (2018), 298–308. DOI: 10.1109/TVCG.2017.2744818.

[BM13] BREHMER, M. and MUNZNER, T. "A Multi-Level Typology of Abstract Visualization Tasks". *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), 2376–2385. DOI: 10.1109/TVCG.2013.124.

[BM18] BEINS, B. C. and MCCARTHY, M. A. *Research Methods and Statistics in Psychology*. 2nd ed. Cambridge University Press, 2018. DOI: 10.1017/9781108399555.

[BMB*18] BORGO, R., MICALLEF, L., BACH, B., MCGEE, F., and LEE, B. "Information Visualization Evaluation Using Crowdsourcing". *Computer Graphics Forum* 37.3 (2018), 573–595. DOI: 10.1111/cgf.13444.

[BSP20] BEHRISCH, M., SCHRECK, T., and PFISTER, H. "GUIRO: User-Guided Matrix Reordering". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 184–194. DOI: 10.1109/TVCG.2019.2934300.

[BZL*18] BERNARD, J., ZEPPELZAUER, M., LEHMANN, M., MÜLLER, M., and SEDLMAIR, M. "Towards User-Centered Active Learning Algorithms". *Computer Graphics Forum* 37.3 (2018), 121–132. DOI: 10.1111/cgf.13406.

[CD19] CAVALLO, M. and DEMIRALP, Ç. "Clustrophile 2: Guided Visual Clustering Analysis". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 267–276. DOI: 10.1109/TVCG.2018.2864477.

[CGM*17] CENEDA, D., GSCHWANDTNER, T., MAY, T., MIKSCH, S., SCHULZ, H. J., STREIT, M., and TOMINSKI, C. "Characterizing Guidance in Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 111–120. DOI: 10.1109/TVCG.2016.2598468.

[CGM19] CENEDA, D., GSCHWANDTNER, T., and MIKSCH, S. "A Review of Guidance Approaches in Visual Data Analysis: A Multifocal Perspective". *Computer Graphics Forum* 38.3 (2019), 861–879. DOI: 10.1111/cgf.13730.

[CHH*19] CASHMAN, D., HUMAYOUN, S. R., HEIMERL, F., PARK, K., DAS, S., THOMPSON, J., SAKET, B., MOSCA, A., STASKO, J., ENDERT, A., GLEICHER, M., and CHANG, R. "A User-based Visual Analytics Workflow for Exploratory Model Analysis". *Computer Graphics Forum* 38.3 (2019), 185–199. DOI: 10.1111/cgf.13681.

[CMJ*20] CHATZIMPARMPAS, A., MARTINS, R. M., JUSUFI, I., KUCHER, K., ROSSI, F., and KERREN, A. "The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations". *Computer Graphics Forum* 39.3 (2020), 713–756. DOI: 10.1111/cgf.14034.

[CMQ20] CHENG, F., MING, Y., and QU, H. "DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3030342.

[CPC19] CARVALHO, D. V., PEREIRA, E. M., and CARDOSO, J. S. "Machine Learning Interpretability: A Survey on Methods and Metrics". *Electronics* 8.8 (2019). DOI: 10.3390/electronics8080832.

[CRH*19] CAI, C. J., REIF, E., HEGDE, N., HIPP, J., KIM, B., SMILKOV, D., WATTENBERG, M., VIEGAS, F., CORRADO, G. S., STUMPE, M. C., and TERRY, M. "Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 4–14. DOI: 10.1145/3290605.3300234.

[CVL*18] COPPERS, S., VAN DEN BERGH, J., LUYTEN, K., CONINX, K., van der LEK-CIUDIN, I., VANALLEMEERSCH, T., and VANDEGHINSTE, V. "Intellingo: An Intelligible Translation Environment". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2018, 1–13. DOI: 10.1145/3173574.3174098.

[CWZ*19] CHENG, H.-F., WANG, R., ZHANG, Z., O'CONNELL, F., GRAY, T., HARPER, F. M., and ZHU, H. "Explaining Decision-Making Algorithms Through UI: Strategies to Help Non-Expert Stakeholders". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 559:1–559:12. DOI: 10.1145/3290605.3300789.

[CYL*20] CHEN, C., YUAN, J., LU, Y., LIU, Y., SU, H., YUAN, S., and LIU, S. "OoDAnalyzer: Interactive Analysis of Out-of-Distribution Samples". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.2973258.

[Dey99] DEY, I. *Grounding Grounded Theory: Guidelines for Qualitative Inquiry*. Academic Press, 1999.

[DLW*17] DASGUPTA, A., LEE, J., WILSON, R., LAFRANCE, R. A., CRAMER, N., COOK, K., and PAYNE, S. "Familiarity Vs Trust: A Comparative Study of Domain Scientists' Trust in Visual Analytics and Conventional Analysis Methods". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 271–280. DOI: 10.1109/TVCG.2016.2598544.

[dSBD*12] Dos SANTOS AMORIM, E. P., BRAZIL, E. V., DANIELS, J., JOIA, P., NONATO, L. G., and SOUSA, M. C. "iLAMP: Exploring high-dimensional spacing through backward multidimensional projection". *IEEE Conference on Visual Analytics Science and Technology*. 2012, 53–62. DOI: 10.1109/VAST.2012.6400489.

[DSKE20] DAS, S., SAKET, B., KWON, B. C., and ENDERT, A. "GeoNoCluster: Interactive Visual Cluster Analysis for Biologists". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3002166.

[DVH*19] DINGEN, D., VEER, M. V., HOUTHUIZEN, P., MESTROM, E. H. J., KORSTEN, E. H. H. M., BOUWMAN, A. R. A., and WIJK, J. V. "RegressionExplorer: Interactive Exploration of Logistic Regression Models with Subgroup Analysis". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 246–255. DOI: 10.1109/TVCG.2018.2865043.

[DVM*12] DELL, N., VAIDYANATHAN, V., MEDHI, I., CUTRELL, E., and THIES, W. ""Yours is better!": Participant Response Bias in HCI". *Proc. Conference on Human Factors in Computing Systems*. 2012, 1321–1330. DOI: 10.1145/2207676.2208589.

[DXG*20] DAS, S., XU, S., GLEICHER, M., CHANG, R., and ENDERT, A. "QUESTO: Interactive Construction of Objective Functions for Classification Tasks". *Computer Graphics Forum* 39.3 (2020), 153–165. DOI: 10.1111/cgf.13970.

[EHR*14] ENDERT, A., HOSSAIN, M. S., RAMAKRISHNAN, N., NORTH, C., FIAUX, P., and ANDREWS, C. "The human is the loop: new directions for visual analytics". *Journal of Intelligent Information Systems* 43.3 (2014), 411–435. DOI: 10.1007/s10844-014-0304-9.

[EKC*20] EL-ASSADY, M., KEHLBECK, R., COLLINS, C., KEIM, D., and DEUSSEN, O. "Semantic Concept Spaces: Guided Topic Model Refinement using Word-Embedding Projections". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 1001–1011. DOI: 10.1109/TVCG.2019.2934654.

[EKSK18] ESLAMI, M., KRISHNA KUMARAN, S. R., SANDVIG, C., and KARAHALIOS, K. "Communicating Algorithmic Process in Online Behavioral Advertising". *Proc. Conference on Human Factors in Computing Systems*. 2018, 432:1–432:13. DOI: 10.1145/3173574.3174006.

[ERT*17] ENDERT, A., RIBARSKY, W., TURKAY, C., WONG, B. W., NABNEY, I., BLANCO, I. D., and ROSSI, F. "The State of the Art in Integrating Machine Learning into Visual Analytics". *Computer Graphics Forum* 36.8 (2017), 458–486. DOI: 10.1111/cgf.13092.

[ESD*19] EL-ASSADY, M., SPERRLE, F., DEUSSEN, O., KEIM, D., and COLLINS, C. "Visual Analytics for Topic Model Optimization based on User-Steerable Speculative Execution". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 374–384. DOI: 10.1109/TVCG.2018.2864769.

[ESKC18] EL-ASSADY, M., SEVASTJANOVA, R., KEIM, D., and COLLINS, C. "ThreadReconstructor : Modeling Reply-Chains to Untangle Conversational Text through Visual Analytics". *Computer Graphics Forum* 37.3 (2018), 351–365. DOI: 10.1111/cgf.13425.

[ESS*18] EL-ASSADY, M., SEVASTJANOVA, R., SPERRLE, F., KEIM, D., and COLLINS, C. "Progressive Learning of Topic Modeling Parameters: A Visual Analytics Framework". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 382–391. DOI: 10.1109/TVCG.2017.2745080.

[FG18] FIEBRINK, R. and GILLIES, M. "Introduction to the Special Issue on Human-Centered Machine Learning". *ACM Transactions on Interactive Intelligent Systems* 8.2 (2018), 7:1–7:7. DOI: 10.1145/3205942.

[GFT*16] GILLIES, M., FIEBRINK, R., TANAKA, A., GARCIA, J., BEVILACQUA, F., HELOIR, A., NUNNARI, F., MACKAY, W., AMERSHI, S., LEE, B., D'ALESSANDRO, N., TILMANNE, J., KULESZA, T., and CARAMIAUX, B. "Human-Centred Machine Learning". *Proc. Extended Abstracts Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2016, 3558–3565. DOI: 10.1145/2851581.2856492.

[GLC*19] GUZDIAL, M., LIAO, N., CHEN, J., CHEN, S.-Y., SHAH, S., SHAH, V., RENO, J., SMITH, G., and RIEDL, M. O. "Friend, Collaborator, Student, Manager: How Design of an AI-Driven Game Level Editor Affects Creators". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 1–13. DOI: 10.1145/3290605.3300854.

[GMV*20] GEBRU, T., MORGENSTERN, J., VECCHIONE, B., VAUGHAN, J. W., WALLACH, H., DAUMÉ III, H., and CRAWFORD, K. "Datasheets for Datasets". *arXiv:1803.09010 [cs]* (2020).

[Goo19] GOOGLE. *People + AI Guidebook: Designing human-centered AI products*. https://pair.withgoogle.com/. 2019.

[Gun17] GUNNING, D. "Explainable Artificial Intelligence (XAI)". *Defense Advanced Research Projects Agency (DARPA)* 2.2 (2017).

[GWGvW19] GARCIA CABALLERO, H., WESTENBERG, M., GEBRE, B., and van WIJK, J. J. "V-Awake: A Visual Analytics Approach for Correcting Sleep Predictions from Deep Learning Models". (2019). DOI: 10.1111/cgf.13667.

[GZL*20] GOU, L., ZOU, L., LI, N., HOFMANN, M., SHEKAR, A. K., WENDT, A., and REN, L. "VATLD: A Visual Analytics System to Assess, Understand and Improve Traffic Light Detection". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3030350.

[HHC*19] HOHMAN, F., HEAD, A., CARUANA, R., DELINE, R., and DRUCKER, S. M. "Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 579:1–579:13. DOI: 10.1145/3290605.3300809.

[HKBE12] HEIMERL, F., KOCH, S., BOSCH, H., and ERTL, T. "Visual Classifier Training for Text Document Retrieval". *IEEE Transactions on Visualization and Computer Graphics* 18.12 (2012), 2839–2848. DOI: 10.1109/TVCG.2012.277.

[HKPC19] HOHMAN, F., KAHNG, M., PIENTA, R., and CHAU, D. H. "Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers". *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2019), 2674–2693. DOI: 10.1109/TVCG.2018.2843369.

[HOW*19] HITRON, T., ORLEV, Y., WALD, I., SHAMIR, A., EREL, H., and ZUCKERMAN, O. "Can Children Understand Machine Learning Concepts?: The Effect of Uncovering Black Boxes". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 1–11. DOI: 10.1145/3290605.3300645.

[IZCC08] ISENBERG, P., ZUK, T., COLLINS, C., and CARPENDALE, S. "Grounded Evaluation of Information Visualizations". *Proc. Workshop BEyond time and errors: novel evaLuation methods for Information Visualization*. ACM, 2008, 1–8. DOI: 10.1145/1377966.1377974.

[JSR*19] JI, X., SHEN, H.-W., RITTER, A., MACHIRAJU, R., and YEN, P.-Y. "Visual Exploration of Neural Document Embedding in Information Retrieval: Semantics and Feature Selection". *IEEE Transactions on Visualization and Computer Graphics* 25.6 (2019), 2181–2192. DOI: 10.1109/TVCG.2019.2903946.

[JVW20] JAUNET, T., VUILLEMOT, R., and WOLF, C. "DRLViz: Understanding Decisions and Memory in Deep Reinforcement Learning". *Computer Graphics Forum* 39.3 (2020), 49–61. DOI: 10.1111/cgf.13962.

[KAKC18] KAHNG, M., ANDREWS, P. Y., KALRO, A., and CHAU, D. H. "ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 88–97. DOI: 10.1109/TVCG.2017.2744718.

[KAS*20] KWON, B. C., ANAND, V., SEVERSON, K. A., GHOSH, S., SUN, Z., FROHNERT, B. I., LUNDGREN, M., and NG, K. "DPVis: Visual Analytics with Hidden Markov Models for Disease Progression Pathways". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.2985689.

[KAY*19] KITTLEY-DAVIES, J., ALQARAAWI, A., YANG, R., COSTANZA, E., ROGERS, A., and STEIN, S. "Evaluating the Effect of Feedback from Different Computer Vision Processing Stages: A Comparative Lab Study". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 1–12. DOI: 10.1145/3290605.3300273.

[KBJ*20] KRUEGER, R., BEYER, J., JANG, W.-D., KIM, N. W., SOKOLOV, A., SORGER, P. K., and PFISTER, H. "Facetto: Combining Unsupervised and Supervised Learning for Hierarchical Phenotype Analysis in Multi-Channel Image Data". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 227–237. DOI: 10.1109/TVCG.2019.2934547.

[KEV*18] KWON, B. C., EYSENBACH, B., VERMA, J., NG, K., FILIPPI, C. D., STEWART, W. F., and PERER, A. "Clustervision: Visual Supervision of Unsupervised Clustering". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 142–151. DOI: 10.1109/TVCG.2017.2745085.

[KPN16] KRAUSE, J., PERER, A., and NG, K. "Interacting with Predictions: Visual Inspection of Black-box Machine Learning Models". *Proc. Conference on Human Factors in Computing Systems*. 2016, 5686–5697. DOI: 10.1145/2858036.2858529.

[KTC*19] KAHNG, M., THORAT, N., CHAU, D. H. P., VIEGAS, F. B., and WATTENBERG, M. "GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 310–320. DOI: 10.1109/TVCG.2018.2864500.

[LDM*18] LIU, J., DWYER, T., MARRIOTT, K., MILLAR, J., and HAWORTH, A. "Understanding the Relationship Between Interactive Optimisation and Visual Analytics in the Context of Prostate Brachytherapy". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 319–329. DOI: 10.1109/TVCG.2017.2744418.

[LGM*20] LEITE, R. A., GSCHWANDTNER, T., MIKSCH, S., GSTREIN, E., and KUNTNER, J. "NEVA: Visual Analytics to Identify Fraudulent Networks". *Computer Graphics Forum* 39.6 (2020), 344–359. DOI: 10.1111/cgf.14042.

[Lip18] LIPTON, Z. C. "The Mythos of Model Interpretability". *Queue* 16.3 (2018), 30:31–30:57. DOI: 10.1145/3236386.3241340.

[LJLH19] LIU, Y., JUN, E., LI, Q., and HEER, J. "Latent Space Cartography: Visual Analysis of Vector Space Embeddings". *Computer Graphics Forum* 38.3 (2019), 67–78. DOI: 10.1111/cgf.13672.

[LLL*19] LIU, S., LI, Z., LI, T., SRIKUMAR, V., PASCUCCI, V., and BREMER, P. "NLIZE: A Perturbation-Driven Visual Interrogation Tool for Analyzing and Interpreting Natural Language Inference Models". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 651–660. DOI: 10.1109/TVCG.2018.2865230.

[LLS*18] LIU, M., LIU, S., SU, H., CAO, K., and ZHU, J. "Analyzing the Noise Robustness of Deep Neural Networks". *IEEE Conference on Visual Analytics Science and Technology*. 2018, 60–71. DOI: 10.1109/VAST.2018.8802509.

[LLT*20] LI, Q., LIU, Q. Q., TANG, C. F., LI, Z. W., WEI, S. C., PENG, X. R., ZHENG, M. H., CHEN, T. J., and YANG, Q. "Warehouse Vis: A Visual Analytics Approach to Facilitating Warehouse Location Selection for Business Districts". *Computer Graphics Forum* 39.3 (2020), 483–495. DOI: 10.1111/cgf.13996.

[LPH*20] LEKSCHAS, F., PETERSON, B., HAEHN, D., MA, E., GEHLENBORG, N., and PFISTER, H. "PEAX: Interactive Visual Pattern Search in Sequential Data Using Unsupervised Deep Representation Learning". *Computer Graphics Forum* 39.3 (2020), 167–179. DOI: 10.1111/cgf.13971.

[LSC*18] LIU, M., SHI, J., CAO, K., ZHU, J., and LIU, S. "Analyzing the Training Processes of Deep Generative Models". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 77–87. DOI: 10.1109/TVCG.2017.2744938.

[LSL*17] LIU, M., SHI, J., LI, Z., LI, C., ZHU, J., and LIU, S. "Towards Better Analysis of Deep Convolutional Neural Networks". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 91–100. DOI: 10.1109/TVCG.2016.2598831.

[LXL*18] LIU, S., XIAO, J., LIU, J., WANG, X., WU, J., and ZHU, J. "Visual Diagnosis of Tree Boosting Methods". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 163–173. DOI: 10.1109/TVCG.2017.2744378.

[MCZ*17] MING, Y., CAO, S., ZHANG, R., LI, Z., CHEN, Y., SONG, Y., and QU, H. "Understanding Hidden Memories of Recurrent Neural Networks". *IEEE Conference on Visual Analytics Science and Technology*. 2017, 13–24. DOI: 10.1109/VAST.2017.8585721.

[Mil19] MILLER, T. "Explanation in artificial intelligence: Insights from the social sciences". *Artificial Intelligence* 267 (2019), 1–38. DOI: 10.1016/j.artint.2018.07.007.

[MLMP18] MÜHLBACHER, T., LINHARDT, L., MÖLLER, T., and PIRINGER, H. "TreePOD: Sensitivity-Aware Selection of Pareto-Optimal Decision Trees". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 174–183. DOI: 10.1109/TVCG.2017.2745158.

[MP13] MÜHLBACHER, T. and PIRINGER, H. "A Partition-Based Framework for Building and Validating Regression Models". *IEEE Transactions on Visualization and Computer Graphics* 19.12 (2013), 1962–1971. DOI: 10.1109/TVCG.2013.125.

[MQB19] MING, Y., QU, H., and BERTINI, E. "RuleMatrix: Visualizing and Understanding Classifiers with Rules". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 342–352. DOI: 10.1109/TVCG.2018.2864812.

[Mun09] MUNZNER, T. "A Nested Model for Visualization Design and Validation". *IEEE Transactions on Visualization and Computer Graphics* 15.6 (2009), 921–928. DOI: 10.1109/TVCG.2009.111.

[MWZ*19] MITCHELL, M., WU, S., ZALDIVAR, A., BARNES, P., VASSERMAN, L., HUTCHINSON, B., SPITZER, E., RAJI, I. D., and GEBRU, T. "Model Cards for Model Reporting". *Proc. Conference on Fairness, Accountability, and Transparency*. ACM, 2019, 220–229. DOI: 10.1145/3287560.3287596.

[MXC*20] MING, Y., XU, P., CHENG, F., QU, H., and REN, L. "ProtoSteer: Steering Deep Sequence Model with Prototypes". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 238–248. DOI: 10.1109/TVCG.2019.2934267.

[MXLM20] MA, Y., XIE, T., LI, J., and MACIEJEWSKI, R. "Explaining Vulnerabilities to Adversarial Machine Learning through Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 1075–1085. DOI: 10.1109/TVCG.2019.2934631.

[NA19] NONATO, L. G. and AUPETIT, M. "Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment". *IEEE Transactions on Visualization and Computer Graphics* 25.8 (2019), 2650–2673. DOI: 10.1109/TVCG.2018.2846735.

[Pal19] PALACIO NIÑO, J. "Evaluation Metrics for Unsupervised Learning Algorithms". *arXiv:1905.05667 [cs]* (2019).

[PGH*21] POURSABZI-SANGDEH, F., GOLDSTEIN, D. G., HOFMAN, J. M., WORTMAN VAUGHAN, J. W., and WALLACH, H. "Manipulating and Measuring Model Interpretability". *Proc. Conference on Human Factors in Computing Systems*. Association for Computing Machinery, 2021, 1–52. DOI: 10.1145/3411764.3445315.

[PLM*17] PEZZOTTI, N., LELIEVELDT, B. P. F., MAATEN, L. V. D., HÖLLT, T., EISEMANN, E., and VILANOVA, A. "Approximated and User Steerable tSNE for Progressive Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* 23.7 (2017), 1739–1752. DOI: 10.1109/TVCG.2016.2570755.

[PNKC20] PARK, H., NAM, Y., KIM, J.-H., and CHOO, J. "HyperTendril: Visual Analytics for User-Driven Hyperparameter Optimization of Deep Neural Networks". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3030380.

[PZDD19] PARVINZAMIR, F., ZHAO, Y., DENG, Z., and DONG, F. "MyEvents: A Personal Visual Analytics Approach for Mining Key Events and Knowledge Discovery in Support of Personal Reminiscence". *Computer Graphics Forum* 38.1 (2019), 647–662. DOI: 10.1111/cgf.13596.

[RAL*17] REN, D., AMERSHI, S., LEE, B., SUH, J., and WILLIAMS, J. D. "Squares: Supporting Interactive Performance Analysis for Multiclass Classifiers". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 61–70. DOI: 10.1109/TVCG.2016.2598828.

[Rie19] RIEDL, M. O. "Human-centered artificial intelligence and machine learning". *Human Behavior and Emerging Technologies* 1.1 (2019), 33–36. DOI: https://doi.org/10.1002/hbe2.117.

[SBS*18] SPERRLE, F., BERNARD, J., SEDLMAIR, M., KEIM, D. A., and EL-ASSADY, M. "Speculative Execution for Guided Visual Analytics". *Workshop for Machine Learning from User Interaction for Visualization and Analytics at IEEE VIS*. 2018.

[SCG09] SUNG, J., CHRISTENSEN, H. I., and GRINTER, R. E. "Robots in the Wild: Understanding Long-Term Use". *ACM/IEEE Int. Conf. on Human-Robot Interaction*. 2009, 45–52.

[SDMT16] STAHNKE, J., DÖRK, M., MÜLLER, B., and THOM, A. "Probing Projections: Interaction Techniques for Interpreting Arrangements and Errors of Dimensionality Reductions". *IEEE Transactions on Visualization and Computer Graphics* 22.1 (2016), 629–638. DOI: 10.1109/TVCG.2015.2467717.

[Seg19] SEGAL, M. "A more human approach to artificial intelligence". *Nature* 571.7766 (2019), 1–1. DOI: 10.1038/d41586-019-02213-3.

[SEH*18] SEVASTJANOVA, R., EL-ASSADY, M., HAUTLI-JANISZ, A., KALOULI, A.-L., KEHLBECK, R., DEUSSEN, O., KEIM, D. A., and BUTT, M. "Mixed-initiative active learning for generating linguistic insights in question classification". *3rd Workshop on Data Systems for Interactive Analysis (DSIA) at IEEE VIS*. 2018.

[SFB*20] SMITH-RENNER, A., FAN, R., BIRCHFIELD, M., WU, T., BOYD-GRABER, J., WELD, D. S., and FINDLATER, L. "No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2020, 1–13. DOI: 10.1145/3313831.3376624.

[Shn20] SHNEIDERMAN, B. "Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy". *International Journal of Human–Computer Interaction* 36.6 (2020), 495–504. DOI: 10.1080/10447318.2020.1741118.

[Shn97] SHNEIDERMAN, B. "Direct manipulation for comprehensible, predictable and controllable user interfaces". *Proceedings of the 2nd international conference on Intelligent user interfaces*. 1997, 33–39.

[SJB*20] SPERRLE, F., JEITLER, A., BERNARD, J., KEIM, D. A., and EL-ASSADY, M. "Learning and Teaching in Co-Adaptive Guidance for Mixed-Initiative Visual Analytics". *EuroVis Workshop on Visual Analytics (EuroVA)*. Ed. by TURKAY, C. and VROTSOU, K. The Eurographics Association, 2020. DOI: 10.2312/eurova.20201088.

[SJS*18] SCHNEIDER, B., JÄCKLE, D., STOFFEL, F., DIEHL, A., FUCHS, J., and KEIM, D. "Integrating Data and Model Space in Ensemble Learning by Visual Analytics". *IEEE Transactions on Big Data* (2018), 1–1. DOI: 10.1109/TBDATA.2018.2877350.

[SKB*18] SACHA, D., KRAUS, M., BERNARD, J., BEHRISCH, M., SCHRECK, T., ASANO, Y., and KEIM, D. A. "SOMFlow: Guided Exploratory Cluster Analysis with Self-Organizing Maps and Analytic Provenance". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 120–130. DOI: 10.1109/TVCG.2017.2744805.

[SLC*20] SUN, J., LI, Y., CHEN, C., LEE, J., LIU, X., ZHANG, Z., HUANG, L., SHI, L., and XU, W. "FDHelper: Assist Unsupervised Fraud Detection Experts with Interactive Feature Selection and Evaluation". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2020, 1–12. DOI: 10.1145/3313831.3376140.

[SMD*16] SARKAR, A., MORRISON, C., DORN, J. F., BEDI, R., STEINHEIMER, S., BOISVERT, J., BURGGRAAFF, J., D'SOUZA, M., KONTSCHIEDER, P., ROTA BULÒ, S., WALSH, L., KAMM, C. P., ZAYKOV, Y., SELLEN, A., and LINDLEY, S. "Setwise Comparison: Consistent, Scalable, Continuum Labels for Computer Vision". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2016, 261–271. DOI: 10.1145/2858036.2858199.

[SSBC19] SULTANUM, N., SINGH, D., BRUDNO, M., and CHEVALIER, F. "Doccurate: A Curation-Based Approach for Clinical Text Visualization". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 142–151. DOI: 10.1109/TVCG.2018.2864905.

[SSKE19] SPERRLE, F., SEVASTJANOVA, R., KEHLBECK, R., and EL-ASSADY, M. "VIANA: Visual Interactive Annotation of Argumentation". *IEEE Conference on Visual Analytics Science and Technology*. 2019, 11–22. DOI: 10.1109/VAST47406.2019.8986917.

[SSS*14] SACHA, D., STOFFEL, A., STOFFEL, F., KWON, B. C., ELLIS, G., and KEIM, D. A. "Knowledge Generation Model for Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), 1604–1613. DOI: 10.1109/TVCG.2014.2346481.

[SSSE20] SPINNER, T., SCHLEGEL, U., SCHÄFER, H., and EL-ASSADY, M. "explAIner: A Visual Analytics Framework for Interactive and Explainable Machine Learning". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 1064–1074. DOI: 10.1109/TVCG.2019.2934629.

[SSZ*17] SACHA, D., SEDLMAIR, M., ZHANG, L., LEE, J. A., PELTONEN, J., WEISKOPF, D., NORTH, S. C., and KEIM, D. A. "What you see is what you can change: Human-centered machine learning by interactive visualization". *Neurocomputing* 268 (2017), 164–175. DOI: https://doi.org/10.1016/j.neucom.2017.01.105.

[SZS*17] SACHA, D., ZHANG, L., SEDLMAIR, M., LEE, J. A., PELTONEN, J., WEISKOPF, D., NORTH, S. C., and KEIM, D. A. "Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis". *IEEE Transactions on Visualization and Computer Graphics* 23.1 (2017), 241–250. DOI: 10.1109/TVCG.2016.2598495.

[vLFB*14] VON LANDESBERGER, T., FIEBIG, S., BREMM, S., KUIJPER, A., and FELLNER, D. W. "Interaction Taxonomy for Tracking of User Actions in Visual Analytics Applications". *Handbook of Human Centric Visualization*. Ed. by HUANG, W. 2014, 653–670.

[WB19] WELD, D. S. and BANSAL, G. "The Challenge of Crafting Intelligible Intelligence". *Commun. ACM* 62.6 (2019), 70–79. DOI: 10.1145/3282486.

[WBL*20] WANG, X., BRYAN, C. J., LI, Y., PAN, R., LIU, Y., CHEN, W., and MA, K.-L. "Umbra: A Visual Analysis Approach for Defense Construction Against Inference Attacks on Sensitive Information". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3037670.

[WGSY19] WANG, J., GOU, L., SHEN, H., and YANG, H. "DQNViz: A Visual Analytics Approach to Understand Deep Q-Networks". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 288–298. DOI: 10.1109/TVCG.2018.2864504.

[WGYS18] WANG, J., GOU, L., YANG, H., and SHEN, H. "GANViz: A Visual Analytics Approach to Understand the Adversarial Game". *IEEE Transactions on Visualization and Computer Graphics* 24.6 (2018), 1905–1917. DOI: 10.1109/TVCG.2018.2816223.

[WGZ*19] WANG, J., GOU, L., ZHANG, W., YANG, H., and SHEN, H. "DeepVID: Deep Visual Interpretation and Diagnosis for Image Classifiers via Knowledge Distillation". *IEEE Transactions on Visualization and Computer Graphics* 25.6 (2019), 2168–2180. DOI: 10.1109/TVCG.2019.2903943.

[WMJ*19] WANG, Q., MING, Y., JIN, Z., SHEN, Q., LIU, D., SMITH, M. J., VEERAMACHANENI, K., and QU, H. "ATMSeer: Increasing Transparency and Controllability in Automated Machine Learning". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2019, 681:1–681:12. DOI: 10.1145/3290605.3300911.

[WPB*20] WEXLER, J., PUSHKARNA, M., BOLUKBASI, T., WATTENBERG, M., VIÉGAS, F., and WILSON, J. "The What-If Tool: Interactive Probing of Machine Learning Models". *IEEE Transactions on Visualization and Computer Graphics* 26.1 (2020), 56–65. DOI: 10.1109/TVCG.2019.2934619.

[WSW*18] WONGSUPHASAWAT, K., SMILKOV, D., WEXLER, J., WILSON, J., MANÉ, D., FRITZ, D., KRISHNAN, D., VIÉGAS, F. B., and WATTENBERG, M. "Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow". *IEEE Transactions on Visualization and Computer Graphics* 24.1 (2018), 1–12. DOI: 10.1109/TVCG.2017.2744878.

[WW21] WORTMAN VAUGHAN, J. and WALLACH, H. "A Human-Centered Agenda for Intelligible Machine Learning". *Machines We Trust: Getting Along with Artificial Intelligence*. Ed. by PELILLO, M. and SCANTAMBURLO, T. 2021, 224.

[WWP*20] WRIGHT, A. P., WANG, Z. J., PARK, H., GUO, G., SPERRLE, F., EL-ASSADY, M., ENDERT, A., KEIM, D., and CHAU, D. H. "A Comparative Analysis of Industry Human-AI Interaction Guidelines". *arXiv preprint arXiv:2010.11761* (2020).

[XCK*20] XIE, Y., CHEN, M., KAO, D., GAO, G., and CHEN, X. '. "CheXplain: Enabling Physicians to Explore and Understand Data-Driven, AI-Enabled Medical Imaging Analysis". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2020, 1–13. DOI: 10.1145/3313831.3376807.

[XMT*20] XIE, T., MA, Y., TONG, H., THAI, M. T., and MACIEJEWSKI, R. "Auditing the Sensitivity of Graph-based Ranking with Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* (2020), 1–1. DOI: 10.1109/TVCG.2020.3028958.

[XXL*20] XIA, M., XU, M., LIN, C., CHENG, T. Y., QU, H., and MA, X. "SeqDynamics: Visual Analytics for Evaluating Online Problem-solving Dynamics". (2020). DOI: 10.1111/cgf.13998.

[XXM*19] XU, K., XIA, M., MU, X., WANG, Y., and CAO, N. "EnsembleLens: Ensemble-based Visual Exploration of Anomaly Detection Algorithms with Multidimensional Data". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 109–119. DOI: 10.1109/TVCG.2018.2864825.

[YCY*20] YUAN, J., CHEN, C., YANG, W., LIU, M., XIA, J., and LIU, S. "A Survey of Visual Analytics Techniques for Machine Learning". *Computational Visual Media* (2020). DOI: 10.1007/s41095-020-0191-7.

[YDP19] YE, W., DONG, Y., and PEERS, P. "Interactive Curation of Datasets for Training and Refining Generative Models". *Computer Graphics Forum* 38.7 (2019), 369–380. DOI: https://doi.org/10.1111/cgf.13844.

[YGLR20] YAN, J. N., GU, Z., LIN, H., and RZESZOTARSKI, J. M. "Silva: Interactively Assessing Machine Learning Fairness Using Causality". *Proc. Conference on Human Factors in Computing Systems*. ACM, 2020, 1–13. DOI: 10.1145/3313831.3376447.

[ZWLC19] ZHAO, X., WU, Y., LEE, D. L., and CUI, W. "iForest: Interpreting Random Forests via Visual Analytics". *IEEE Transactions on Visualization and Computer Graphics* 25.1 (2019), 407–416. DOI: 10.1109/TVCG.2018.2864475.

Biographies

Fabian Sperrle is a a third-year PhD student in the Data Analysis and Visualization group at the University of Konstanz. He received his M.Sc. in Computer and Information Science from the University of Konstanz in 2018. His research interests include co-adaptive analysis processes in general and guidance in particular. He further works on interactive text analysis and visualization.

Mennatallah El-Assady is a research associate in the Data Analysis and Visualization Research Group at the University of Konstanz, Germany, and in the Visualization for Information Analysis lab at the University of Ontario Institute of Technology, Canada. Her research interests include visual text analytics, and explainable machine learning, in particular, user-steerable topic modeling.

Grace Guo is a second year PhD student in the School of Interactive Computing at Georgia Tech. She is part of the Visual Data Analytics Lab, and her current research interests are in XAI, visualization recommendation and human-computer interaction.

Duen Horng (Polo) Chau is an Associate Professor of Computing at Georgia Tech. He co-directs Georgia Tech's MS Analytics program. He is the Director of Industry Relations of The Institute for Data Engineering and Science (IDEaS), and the Associate Director of Corporate Relations of The Center for Machine Learning. His research group bridges machine learning and visualization to synthesize scalable interactive tools for making sense of massive datasets, interpreting complex AI models, and solving real-world problems in cybersecurity, human-centered AI, graph visualization and mining, and social good. His Ph.D. in Machine Learning from Carnegie Mellon University won CMU's Computer Science Dissertation Award, Honorable Mention.

Rita Borgo is Senior Lecturer in Data Visualization at the Informatics Department at King's College London (KCL), Head of the Human Centred Computing research group and Deputy Director of the Centre for Urban Science and Progress (CUSP) - London. Her research focus is on Information Visualization and Visual Analytics with particular focus on the role of Human Factors in Visualization. Her research has followed an ambitious program of developing new data visualization techniques for interactive rendering and manipulation of large multi-dimensional and multivariate datasets.

Alex Endert is an Associate Professor in the School of Interactive Computing at Georgia Tech. He directs the Visual Analytics Lab, where him and his students explore novel user interaction techniques for visual analytics. His lab often applies these fundamental advances to domains including text analysis, intelligence analysis, cyber security, decision-making, and others.

Daniel Keim is professor and head of the Data Analysis and Visualization Research Group in the Computer Science Department at the University of Konstanz, Germany. He has been actively involved in data analysis and information visualization research for more than 30 years and has developed novel visual analysis techniques for large data sets. Dr. Keim got his Ph.D. degree in Computer Science from the University of Munich, Germany. Before joining the University of Konstanz, Dr. Keim was associate professor at the University of Halle, Germany, and Technology Consultant at AT&T Shannon Research Labs, NJ, USA.