Eliciting Model Steering Interactions From Users via Data and Visual Design Probes

Anamaria Crisan ⬤, Maddie Shang, and Eric Brochu

Abstract—**Visual and interactive machine learning systems (IML) are becoming ubiquitous as they empower individuals with varied machine learning expertise to analyze data. However, it remains complex to align interactions with visual marks to a user's intent for steering machine learning models. We explore using data and visual design probes to elicit users' desired interactions to steer ML models via visual encodings within IML interfaces. We conducted an elicitation study with 20 data analysts with varying expertise in ML. We summarize our findings as pairs of target-interaction, which we compare to prior systems to assess the utility of the probes. We additionally surfaced insights about factors influencing how and why participants chose to interact with visual encodings, including refraining from interacting. Finally, we reflect on the value of gathering such formative empirical evidence via data and visual design probes ahead of developing IML prototypes.**

Index Terms—**Design probes, interactive machine learning, model steering, semantic interactions.**

I. INTRODUCTION

VISUAL and Interactive machine learning (IML) systems are powerful interfaces for incorporating humans into the machine learning loop [1], [2], [3], [4]. However, designing visual encodings and interactions for steering machine learning models is a complex problem [5], [6]. One source of this challenge is the tension between model and data-centric priorities [7], [8], [9] that require different interaction modalities, for example using control panels or direct manipulation of marks [10]. Another is a broad user base with diverse expertise in machine learning (ML) and data science (DS), but who regularly employ these systems [11], [12]. Finally, there are challenges toward aligning human intent with ML and data interactions [6], [13].

We explore the use of data and visual design probes as a way to gather formative empirical evidence to align visual

IML systems interactions with users' anticipated outcomes and intents. Probes are an important component of pre-design formative research capable of co-creating user experiences with impacted stakeholders and surfacing problems ahead of development effort [14]. They can be useful instruments for assessing the *semantic distance* [15] between a user's actions and the meaning behind them - for example, whether changing the position of a mark suffices a user's intent of modifying its class label. While probes exist in a variety of forms and are widely used in human computer interaction (HCI) research [14], [16], including prior IML applications, [13], [17] they are under-explored in visualization research. Instead, visualization research has tended to emphasize the development and evaluation of prototypes, which can result in cementing design choices ahead of sufficient formative input from end users [14]. Prototypes can also suffer from the *attribution problem* and can lack "*observed results are loosely attributed to the system as a whole, but without accurate explanations or insights as to what component(s) played a bigger role to achieve those results*" [5]

Here, we developed a set of data and visual design probes that we used to conduct an elicitation study with twenty participants that routinely analyze data but vary in their ML and DS expertise. The goals of our study are threefold. First, we seek to elicit participants' intents coupled with direct manipulations of visual marks to update an ML model. We deliberately place a lower emphasis on other features of a visual IML system and focus solely on single visual encodings to mitigate attribution issues [5]. Second, we explore to what extent and in what ways the interactions differ in accordance with ML/DS expertise; while we can make some basic assumptions the goal is to gather data in lieu of making design decisions on instinct. Finally, we reflect on how the empirical evidence we gather can be used to prioritize the elements of visual IML systems.

Collectively, we present three core contributions:

- The design of data and visual design probes for examining direct manipulations of visual encodings
- A set of target-interaction pairs that describe both *what* data interactions act upon and *how*
- The results of an elicitation study that explores factors influencing participants' interactions

These contributions present opportunities for expanding the visualization research toolbox of formative methodologies for delineating an interaction design space that aligns with users' needs and expectations.

II. RELATED WORK

We summarize pertinent research from visualization, human computer interaction, and machine learning research.

A. Frameworks for Human-AI/ML Interaction

Theoretical frameworks for paradigms for human-AI/ML interaction inform and influence the development of techniques for visual and interactive machine learning [18], [19], [20], [21], [22], [23]. Within the visualization community, these frameworks are rooted in earlier articulations of interactions more generally. Ceneda et al. [24] presented a framework that incorporated human guidance into the visualization processes proposed by van Wikj's [25] model of interactions. Building on their processes is the work by Sperrle et al. [3], who present a framework for a co-adaptive process for people to both learn from a model and incrementally guide it toward a better understanding of the human's analytic intents over time. They describe action-reaction pairs that complement the findings of the target-interaction pairs from the elicitation study. Sacha et al. [2] and Collins et al. [1] proposed frameworks to address limitations of incorporating human guidance via visual analytics processes. They delineate various points and methods for human intervention, including semantic interactions, across stages of model building. Beyond this work, others have recently explored how an algebraic visualization design process (AVD) [26] can help surface the side-effects of ML model updates on visual representations [20], [27]. Dudely et al. [6] place visualization tools in the context of broader user interface affordances and describe how they can benefit end-users with varying degrees of ML experts

The development of our visual design probes is influenced by these prior frameworks for human-AI/ML interaction. We were especially influenced by Sacha's [2] framework of "what you see is what you can change", to enable direct manipulations with data encoding marks.

B. Techniques for Interactive Model Steering

Direct manipulations for user interface elements have long been a component in steering models of IML systems [6], [28], [29]. A common approach is to allow users to change the model's parameters by directly changing the values of input widgets (i.e., change a number in an input box or slider). Although common, this approach has limitations) [10], [30]. Input widgets are 'model-centric knobs' that can be overwhelm users and be ineffective [6], [31], [32].

Semantic Interactions are an attractive alternative to input widgets. Semantic interactions shield end-users from the complexity of the model while using a visual metaphor that interprets interactions with visual encodings to the underlying model parameters [33]. Moreover, these interactions are incorporated into model learning and provide feedback to end-users through the visual metaphor. Endert et al. demonstrated the utility of these semantic interactions with the ForceSPIRE [33] system, with further extension by Bradel et al. to develop StarSPIRE [34].

Semantic interaction techniques have also been applied to standard text modeling algorithms, as evidenced by iVisCluster [35] and Utopian [36]. While this early work focused on spatial metaphors, others have shown their limitations [37] and have explored complementary and alternative approaches.

Other research has explored direct manipulations with data via visual encodings but does not explicitly refer to themselves as semantic interactions. Saket et al. [38], [39], [40] have explored possible interactions with different encoding types; we leverage their ideas and approach in our design probes. Other systems explore visual encodings more coarsely than [38], [39], [40], but prioritize the development of techniques for model steering. Brown et al. [41] demonstrates how interactions with spatialized data can learn and refine and update a distance function. Similarly, Podium learns a ranking function through users' interactions [42]. More recent work on collaborative semantic inference [43] uses MLhooks to capture interactions and update a deep learning model. Finally, there exist techniques that more directly scrutinize the effects of data changes, and not just parameter or function modifications, on the model changes. Systems like iCluster [44] and VIAL [45] demonstrate that visual clustering and labeling of data can outperform online learning alone [46]. The Chameleon [47] prototype enables ML practitioners to visualize the effects of their data changes on the model's features and performance.

Here, we list only a subset of the many techniques that exist for incorporating human actions to steer ML models via visual and interactive interactions. We refer the reader to more recent systematic reviews for in-depth and comprehensive overview [30], [48], [49], [50]. To reflect on the utility of our approach, we compare the interactions we elicit design probes to existing and varied techniques for human-AI/ML interaction and model steering.

C. Design Probes for Interactive Machine Learning

Design probes, generative toolkits, and prototypes are all complementary approaches for designing and evaluating human-centered technological system [14], [51]. A framework by Sanders et al. [14] emphasizes the particular importance of probes for conducting pre-design work that aims to find *"inspiration in users' reactions to [the probe's] suggestions"*. They contrast probes to prototypes that aim to test a specific idea and tend to be used in evaluative research [52]. Prior research has demonstrated that design probes can be especially useful for creating a human-centered experience for machine learning [53], [54], which becomes increasingly essential in light of the increasing diversity of those using (and impacted by) AI/ML technology. While methods for developing and using design probes vary, a recent framework by Subramonyam et al. [17] has synthesized these common approaches into a Model Informed Prototyping Workflow that *"combines model exploration with UI prototyping tasks"*. Recently, they also propose the use of *data probes* to co-create Ml/AI interactive experiences for non-expert analysts [55]. The mechanisms for capturing user interactions and an ML model's response varies by according to the approach. For example, design probes can be coupled with

Wizard of Oz approach that simulates ML model behavior [56], [57], [58]. Prototypes like Gamut [59] automatically capture interactions and display model responses. While visualization researchers use a variety of approaches, we note a preference for high-fidelity prototypes assessed in summative evaluations [52]. We see opportunities to further explore approaches for co-creating visual and interactive ML systems together with users in pre-design research [60].

III. Data and Visual Design Probes

In this section, we describe the construction of a data probe and a set of visual design probes that we use in a subsequent elicitation study. Supplemental materials for our probes have been made available online: https://osf.io/8wbgf.

The primary objective of our data and design probes is to elicit interactions with a machine learning model with the intent of steering the model's training: We do not limit ourselves to uncovering only novel interactions, but also to collect the broad scope of users' desired interactions. In addition to collecting these interactions, we wanted to also explore the factors motivating those interactions including reasons why interaction is not desirable. Collecting both desired interactions and users' expressed intents allows examine and contrast what is both desirable and feasible.

The possible design space for these data and visual probes is vast. We prioritize understanding how users directly manipulate data-encoding marks of visual encodings to steer an ML model. The result is that our visual design probes emphasize different chart types, similar in spirit to the approach taken by Saket et. al. [38], and not all aspects of the interface. Instead, we use our probes to elicit the other aspects of the user interface (i.e., performance diagnostic plots, explainability features, etc.) they believe would be useful. Finally, we aimed to minimize potential attribution issues that impact visual IML systems [5]. We created our data and visual design probes with these objectives in mind.

A. Data Probe

Data probes are a useful tool for encouraging divergent thinking about the behaviors on IML systems, considering their boundaries and limitations, and imagining (or proposing) alternative behaviors [55]. For our purposes, data probes are useful for reflecting on the efficacy of the different visual encoding choices and the interactions they afford. We created a design probe for a synthetic movies dataset. We use synthetic data, over an actual dataset, because it allows us to control the shape and distribution of the data, including the introduction of noise and errors, to probe end-users perceptions and their desired interactions.

Overview: Our data probe explores the output of a topic classification model. We created a dataset of 50 movies that we compiled from IMBD and Rotten Tomatoes, harvesting their title, a brief synopsis, and primary genre (i.e., its classification label). We select five movies belonging primarily to one of ten genres: Action, Animation, Comedy, Drama, Fantasy, Holiday, Horror, Romance, Sci-Fi, and Thriller. To construct our synthetic dataset, we first deliberately introduce errors into the labels.

For example, *Dr. Strangelove* is a comedy movie, but we set its predicted genre as a romance; *Die Hard* is a 'Holiday' movie but is predicted to be 'Action').[1] Next, we use a Gaussian Mixture Model (GMM)s to simulate the classification probabilities as well as a set of two-dimensional XY coordinates for each data point; the 2D coordinates are a substitute for computing a dimensional reduced representation of the dataset. We explored different parameter settings for the GMM and its components, to modify the number spread of each cluster, their proximity and orientation to each other, and the extent that they overlap (see supplemental materials). We selected one set of simulated results for the final data probe.

Design Considerations: While movies can belong to multiple genres, we deliberately imposed a single classification label per movie to observe if participants would be motivated to modify this. If participants were to introduce multiple labels per movie, then they would be introducing a change to the model that would go beyond simple parameter updating. We similarly use errors and ambiguities as prompts for discussion around the data and the model. In both instances, we are interested in assessing if participants would discover these prompts via the visual design probes and how they would respond to them.

The overall size of the data probe is smaller than a typical ML workflow, which features thousands, millions, or even billions of points depending on the dataset. The smaller size of 50 points was pragmatically selected to make the design and interaction of the visual design probes feasible, whilst still having diversity and variation in the dataset. Visualization research has not proposed a singular dataset for clustering data, nor has prior research concerning clustering with visualization discussed the relationship between dataset size, total marks displayed, and interactions. Lacking such guidance, we opted for a dataset size amenable to our research objectives. However, we argue that while datasets are massive, end-users rarely interact with all data points at once. They may filter or aggregate information, leaving them with a more reasonable number of points to interact with. While the design of our data probe may not be suitable for answering all possible questions about model steering through visual encodings, it does allow us to explore the fundamental questions of how end-users interact with data encoding marks where each mark represents one data point.

Generality and Transferability: We derived the probe data from two commonly used machine learning datasets with modifications to deliberately introduce errors and ambiguity. Moreover, the movies that we selected for our probe are popular such that the majority, if not all, participants were familiar with them and their premise. The choice of dataset also served to ameliorate two potential confounding factors in this study: dataset characteristics and participant's dataset familiarity. Our dataset is not intended to serve as such a benchmark, others may find our approach transferable to other datasets and applications.

---

[1] The movie 'Die Hard' takes place on Christmas Eve and has prompted a (sometimes facetious, sometimes serious) debate of whether it is a holiday film or not. Among our participants, 60% said it was solely an action movie, 35% thought it was both an action and holiday movie, and 5% were unsure.

TABLE I
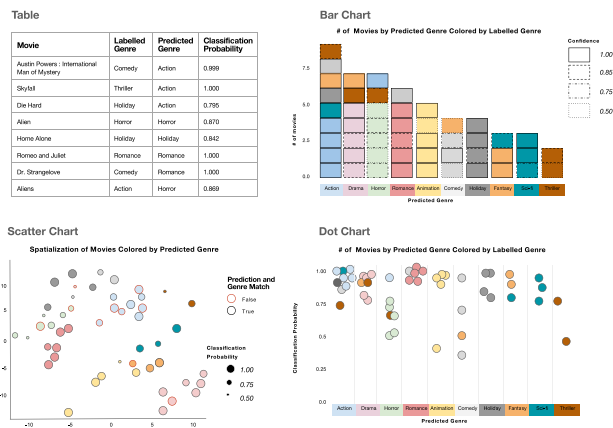MAPPING OF ELEMENTS OF THE DATA PROBE TO VISUAL ENCODING

Fig. 1. Visual design probes created for our elicitation study.

B. Visual Design Probe

We used the data probe to create a set of visual design probes. Visual and interactive ML interfaces can be complex and multi-faceted and include modalities for the user both to interact with data, through both existing and bespoke visual encodings [49], and get feedback on the effects of their interactions [6]. The complexity of these interfaces can also lead to attribution challenges [5]; it is not clear what the value of individual interface components is. We construct visual design probes that examine the relationship between the visual encoding design choices and participants' interactions with individual data marks. Our visual design probes would constitute *one component* of an overall IML system and we use these probes to also elicit from users *what the other interface components should be*.

Overview: We create visual design probes for a simple text classification problem. We mapped the elements of the data probe to marks and channels to a set of four visual encodings: a table, a bar chart, a scatter chart, and a dot chart. While the table is potentially an unusual type of visual encoding, it remains a common way to view and report on machine learning results and serves as an important baseline. Encoding choices are summarized in Table I and the four visual design probes are shown in Fig. 1.

Table: Each row displays one movie along with its labeled and prediction genres and the classification probability.

Bar Chart: Each rectangular mark in the bar chart represents a single movie. The marks are vertically stacked and positioned along the x-axis according to the predicted genre. The x-axis is also sorted by the number of movies predicted to belong to the genre. The color of the individual marks represents the labeled genre. The model's prediction probability is shown via the mark's outline; a solid outline indicates higher confidence, whereas the dashes indicate lower confidence.

Scatter Chart: Each point in the scatter chart is a single movie. The marks are positioned according to x and y coordinates that are simulated from the GMM. The marks are colored according to their predicted genre and sized according to the model's predicted probability; smaller mark sizes indicate lower prediction probabilities. A red stroke outline around the marks indicates whether the predicted genre matches the labeled genre (black outline) or not (red outline). The scatter chart is also intended to represent lower dimensionality data representations or embeddings that are routinely presented alongside ML models.

Dot Chart: Each point in the dot chart is a single movie. The points are positioned along the x-axis according to their predicted genre and along the y-axis according to the predicted probability. As with the bar chart, the genres are sorted according to class size from largest to smallest. The colors of the point marks encode the labeled genre.

Alternative Chart Types: We recognize that different individuals may expect different visual encoding types, or may not find the default encoding choices effective. In considering how we implement our probes we deliberately sought to provide participants with the tooling to challenge our design choices and even propose new visual encodings.

Design Considerations: We choose these three visual encodings (and table) because, based on our experience, they are common ways to visual ML model data and results. Others have also previously explored these visualizations in prior work of direct manipulations of data encoding marks [38]. The goal of consistently encoding data across probes is to reduce the potential for anyone visualization to have an outsized effect on the interactions. Others who leverage data and visual probes in their work may find greater variability in the probe design desirable. However, in a pilot study (Section IV-F) we found that too much variability was disorienting. We modified our probes ahead of the full study (Section V) to the versions shown in Fig. 1.

We use the same visual design probes for both ML and non-ML Experts. While these two groups have different expertise, which influences their needs, we believe that visualizations are a common and reasonable baseline for both groups. We rely on both the expressivity of our probe medium and our elicitation study structure to capture additional elements participants would need, as well as, the similarity and differences between these groups.

C. Visual Design Probe Implementation

We developed visual design probes that are situated between a paper prototype and a web-based app. Through our prototype, an end-user can interact by directly manipulating marks in the

chart (i.e., change its position, color, size, etc.). The responses to these interactions are driven by a human 'wizard' (i.e., the study administrators) and assessed by the end-users.

Overview: We considered different materials for implementing our visual design probes and found Google Slides to be a surprisingly compelling medium to meet our goals. As a medium, Slides allowed us to implement the different visual encodings, provided a way for participants to easily change them (i.e. change position, color, size, shape, etc.), and introduce new encoding elements. It also allowed us to securely share and save individual sessions with different users. The slides environment was also familiar, reducing concerns about the effects of systems novelty [5]. Finally, Slides was also collaborative allowing for an administrator to create interactive responses to end-user interactions.

Affordances of Visual Design Probes: The four visual encodings in Fig. 1 were implemented in Google Slides with individual marks and channels capable of being directly manipulated (i.e., moving a point, changing its size, or color). We chose this medium because it provided considerable encoding and interaction affordances. At a minimum, we were able to support the full spectrum of interactive operations reported in Saket et al. [38] (i.e., modifying position, size, color, height, and width encodings) in addition to others such as adding text annotations, new shapes, or comments. We also created tooltips for each mark by making use of hyperlink previews.

The goal of our visual design probes was to hone in on direct manipulations with data encoding marks, not a whole system. We believe this focused approach can collect useful empirical data by providing a canvas for expressing desired interactions with visual encodings. Moreover, the affordances of our probe can provide sufficiently generative power to meaningfully extract the relationships between data, interactions, and desired model updates.

Simulating Model Modifications: We used a Wizard-of-Oz [57] approach to simulate model updates and system behaviors. Our Wizard made two assumptions about participant interactions with visual design probes. First, any interactions with data encoding marks were valid. For example, if a user wanted to change the size, position, and/or color of a mark to express the same intent, such as changing a class label of one or more data points, these were all valid actions. Second, participants could introduce *new* data marks and the wizard could learn from them. For example, adding an ellipse around a group of points to denote a class. Given these interaction assumptions, the possible responses by the Wizard were broad, but allowed us the flexibility to collect a variety of participant perspectives as part of a pre-design empirical phase. Notably, we made no assumptions towards the back-end ML model for the Wizard's responses. We assumed that some of the participants' interactions could conform to existing models, including the GMM we used to generate the data, but we also wanted to capture possible instances were they did not.

A study administrator would craft a model's possible responses in response to participant interactions and the visual encoding choice. For example, if an end-user changed the position of the mark, a study administrator could respond by changing its color or moving additional points that were proximal to the one that moved. The administrator prompted participants for their expected rate of update, for example, with each interaction or following a sequence of interactions. Lastly, before the model was updated participants were asked to indicate their expected response. The wizard used these utterances to craft a response by changing marks in the visual encoding. We did not have a prescribed procedure for how the Wizard should interpret the utterances; it was left to them to interpret based on session and the context. Participants were not informed that their utterances were used to carry out the response. After the update was carried out, participants were prompted again to see if the changes aligned with their expectation. Given the medium, participants were aware that a human was driving the updates, but they did not know how the updates were being made. This approach allowed for co-creation processes of ideating around different individual expectations of model responses.

IV. ELICITATION STUDY

We conducted an elicitation study using the visual design probes. Elicitation interviews are a qualitative technique that uses an iterative approach to reveal increasingly granular details of a participant's experience within a situated scenario, in this case steering a classification model [61]. Our goal of the elicitation study is to explore "approaches, processes, and subjective experiences" [61] of interacting with ML models. We argue that this approach allows us to explore the *semantic distance* [15] between a variety of encoding and interaction affordances and the user's intended model modifications. To that end, we seek to not only elicit novel interactions, but to examine *how* interactions are linked to possible modifications of ML models and *what* the intended effects of these interactions are.

A. Research Questions

We are motivated by the challenges that domain experts experience when needed to update machine learning models, often without the use of code [11], [20], [62]. We also wanted to understand if and how the needs of these domain experts differ from individuals that have formal ML training. Our study sought to examine the following research questions :
- *RQ1* : *How* do participants choose to interact with marks of visual encodings to steer a model?
- *RQ2* : *What* are participants' perspectives or concerns towards these interactions?
- *RQ3* : *What differences* exist (if any) between ML and non-ML experts?

We use a the same visual design probes to assess these research questions with both ML and non-ML experts. While the needs of both groups are different at face value, we were interested in understanding to what extent and in what ways expectations of these two groups overlap or differ relative to a common baseline. We posit that using a common baseline as part of pre-design empirical work [14], [60] helps to calibrate and evaluate IML interfaces.

B. Participant Recruitment and Background

ML and Non-ML Expert Definition: We use the definition of a non-ML expert articulated by Qian et. al. [11], which is that they are *"people who are not formally trained in ML and are actively building ML solutions to serve their needs in the real world"*. This definition excludes lay people that, in addition to lacking forward knowledge, do not actively build or develop ML workflows even though they may encounter ML models in their daily work [12]. We define an ML expert to be someone with formal training (i.e., undergraduate or graduate-level education) in ML or DS.

Recruitment: We recruited an even number of participants with and without ML expertise to participate in our study. We posted calls to participate on Twitter and various machine learning and data science Slack Channels, in addition to using snowball sampling [63] to recruit and screen eligible participants and determine if they were classified as ML or non-ML experts. We also sought to have some gender balance in our recruiting, specifically reaching out to individuals that identified as female to participate. We recruited a total of 20 participants; half of the participants identified as female or non-binary. The initial four recruits were used in a pilot study (Section IV-F) that served to both refine the data and design probes and also to establish an appropriate sample size. Following the initial full study data analysis, we reflected on the findings and established that we had reached analytic saturation. Based on this analysis we determined another round of recruitment was not required.

Background: A total of 8 of 20 (40%) identified as data scientists, 6 (30%) as data or machine learning engineers, 6 (30%) as software engineers, 3 analysts (15%), and finally 3 (15%) as visualization researchers. A total of 6% of individuals did not identify with any of these roles. Note that participants could select more than one category to report their role. Aligning with our definition of ML and non-ML experts, all participants had regular exposure to ML models but reported a varied focus on model work in their daily activities. In total 8 (40%) reported that training machine learning models were a daily component of their job; note that building an ML solution does not require training a model, as it is possible to use pre-trained models. An additional 10 (50%) report primarily interpreting machine learning results in their daily work.

C. Session Procedures

On boarding: All sessions took place over a video conferencing platform. We participants were provided with an overview of the study after which we obtained consent to participate to record audio and video of the session. Participants were then provided with a unique study link to access a set of training materials and the study probes. For the training materials, participants were given an overview of the data probe (Section III-A). They were also provided with a simple bar chart encoding and asked to complete a set of basic interaction tasks (i.e., move a point, change its color or position, hover to reveal a tooltip) to familiarize themselves with the affordances of Google Slides. As most participants had experience using Slides, or some other presentation tool, all found this task to be straightforward.

Interacting with Visual Design Probes: Participants were then presented with the design probes (Fig. 1) in one of four possible orders (Section IV-F). For each visual encoding, participants were given a brief overview of the data that is encoded and were provided with a set of possible interactions (i.e., move a point, change its color). *Importantly, participants were informed that it was valid to indicate that they did wish to make any modifications.* Participants were prompted to speak out loud to describe what they were doing and why. In response to both participant interactions and responses, the study administrator would simulate a response (see Section III-C). These were sometimes simple responses (filtering, sorting/re-ordering, or highlighting) or were model responses (updating the positions of other points, reclassifying data according to new labels, updating classification probabilities). Lastly, participants were encouraged to consider what additional feedback would help them understand and contextualize their actions; if they chose not to interact, they were asked to discuss why.

Facilitating Model Updates: Model updates proceeded at a cadence in accordance with participants' stated preferences (see Section III-C); in practice, nearly all participants wanted updates after completing a sequence of interactions, not with each individual interaction. For the update step, the wizard would manipulate individual data encoding marks, or groups of marks, depending upon the participants' interactions and their expressed model steering intents. Given the size of the data probe and the simplicity of the visual design probes the 'model update' could be carried out in under a minute in most instances. Participants could see individual marks being moved in real-time. At the end of the update, indicated by a prolonged pause, participants were asked to comment on the 'responses' and whether the cadence of change and magnitude of change aligned with their expectations. Once again, the specific responses were tailored to each individual and the particulars of that session. The process of model refinement would continue until participants indicated they wished to conclude.

Wrapping Up: Participants were asked to reflect on their experience overall and provide any final thoughts.

D. Data Collection and Analysis

We collected and analyzed approximately 20 hours of recordings. We retained modifications participants made to the design probes and took notes throughout the sessions. Video recordings captured participants' interactions with the design probes and audio was transcribed for analysis. We manually annotated the video recordings for participants' direct manipulation of visual encoding marks. We record single instances of an interaction per visualization, not per mark; for example, if a participant changes the color of 10 different point marks in the scatter chart, this is tallied as just one instance of semantic interaction and not 10. We use descriptive statistics to summarize the types and frequencies of interactions across sessions. We use participant utterances to add context to both the purpose of the interactions, and any hesitations.

E. Session Overview

Sessions lasted on average 50 minutes; the shortest session was 40 minutes and the longest was 1 h and 7 minutes. Participants spent on average 5 minutes and 32 seconds with each encoding; administrators checked in with participants either at the five-minute mark or when they observed a prolong paused (one minute or greater). In total participants spent 22 minutes and 40 seconds (roughly 50%) of the overall session time across all four encodings. The rest of the session duration consisted of consent, overview, practice, and wrapping up and addressing technical issues (video, connectivity, screen sharing) if any arose. On average participants spent 5m33 s with the table, 5m31 s with the bar chart, 5m56 s with the scatter chart, and 5m17 s with the dot chart. The interactions with the scatter chart were slightly slower because it took more time for participants to identify a mark they were interested in with the hover action. The longest a participant spent with any one encoding was approximately 9 minutes. We did not precisely time the responses of the Wizard, but did not note any response that took especially long to craft for any encoding. This is largely because the Wizard was reasonably familiar with correspondence between marks and certain movies and given the dataset size was able to quickly respond.

Participants did not continuously interact with an encoding during a session, instead they would perform as a set of interactions after which they would be prompted by the study administrators to explain the meaning behind them. Participants varied in how much or little they chose to discuss their interactions; again while we lack precise timing, anecdotally some participants were more verbose than others, which may have contributed to longer time spent with an encoding.

F. Pilot Study

We conducted a preliminary pilot study with four participants to refine the visual design probes (i.e., spelling errors, complex flow, and design inconsistencies). The pilot study also revealed the ordering effect of our visual probes, which introduced an anchoring effect [64]. To mitigate this effect, we grouped the table (T), bar chart (B), scatter chart (S), and dot chart (D) into two sets and varied the order in which they would be presented to the participants as follows: [(TS)(DB)], [(TD)(BS)], [(ST)(DB)], [(DB)(ST)]. Given these presentation orders, our study would need to recruit at least eight participants (4 ML and 4 non-ML). We wanted at least each order to be viewed twice within each group, which resulted in a minimum study size of 16 participants.

V. ELICITATION STUDY FINDINGS

We summarize the results of our elicitation study according to our research questions (Section IV-A).

A. RQ1: A Spectrum of Interaction

We summarize participants' interactions with visual encoding marks as target-interaction pairs. **Targets** are the data attributes that participants sought to modify and **interactions** are the mechanism they used to do so. Participants demonstrated a variety of strategies and consistency when manipulating marks of our visual design probes. The full set is enumerated in Fig. 2 and a subset of the most common target and interaction pairs are shown in Fig. 3. Across all of our study sessions, we recorded a total of 114 examples of interactions that we summarize as a set of 43 unique target-interaction pairs. These interactions were not gathered uniformly across participants; two (one ML and one non-ML) provided no interactions at all (per the session procedures (Section IV-C) it was valid to indicate no desire to interact). Conversely, one participant demonstrated 16 (of 43) unique interactions; the median was 8. Participants targeted their interactions both at individual data encoding marks, and sets of marks. Critically, we found the interactions were deliberate and that participants could articulate their interaction intents and expectations.

1) Data Targets: We identified a total of 9 data targets; this number exceeds the total number of raw attributes in the dataset because participants also introduced new attributes through the encodings. As an example, one participant added a column to the table probe to introduce a weight to each movie in the dataset. The participant wanted the weight column to inform the model of data points to prioritize in retraining: *"I want to tell the model to worry about these points and don't worry as much about getting these [others ones] right [...] which would allow me to say, I really care about these and I don't so much care about those"* [ML:07]. One participant indicated wanted to augment the data through the interactions, a reflection of the expressivity of our design probes: *"when I think about what I am feeding the machine learning model there, it's essentially creating another dimension for it to run through[..] I am saying 'Here's a user dimension that I want you to consider [..] with the other dimensions that you have'*[ML:04]. Lastly, nearly all participants also introduced multiclass labels for each movie, which for analysis we treated as separate from the original single class label (the primary genre).

Considering both initial datasets and additional participant-derived attributes collectively, we found that the most common targets of interactions are primary and predicted labels (genres) and model probabilities (Fig. 2). However, we also observed instances when participants also wanted to indicate when they agreed with the predicted label instead of the initial label: *"I also have this urge to say that I like how this is being classified [over the primary]"* [non-ML:06]. These findings show that participants are not just seeking to update the ML model parameters. They are operating on a meta-modeling level by saying that the way the data is modeled might not be true to the meaning of the data itself; they are challenging the "ground truth" of the given data. Moreover, although our data probe is constructed for a simple classification model, our findings show that the scope of modifications participants wish to make would push the boundaries of such a model by challenging the data assumptions it was trained on.

2) Mechanisms of Interaction: Participants demonstrated a variety of strategies for interacting with data targets (Fig. 2). The most common types of interactions were changing the position, size (or height), and color of the mark. These types of interactions were common to the bar, scatter, and dot plots. With
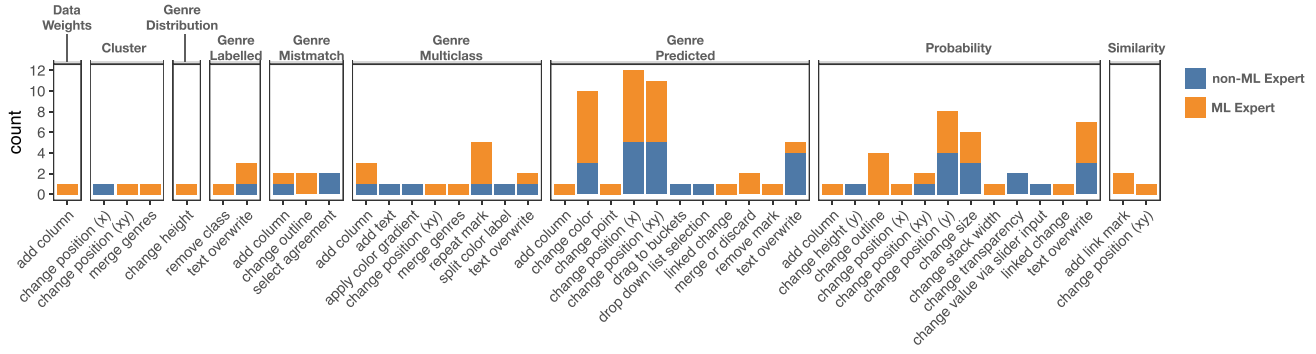
Fig. 2.   Frequency of Interactions (x-axis) and their Data Targets (facets) elicited from a study using the visual design probes.

Fig. 3.   Subset of common target-interaction pairs elicited from participants in our elicitation study. The difference between the initial (V_0) and resultant visualization (V_1) can be used to refine an ML model without directly exposing the person to its parameters.

the table data, participants would frequently overwrite existing text, for example, to modify a movie's genre or its prediction probability. Interestingly, participants also introduced graphical elements to the table. The most common was coloring the rows to reflect the classification probability. Some participants added marks to new columns (see Fig. 4) that they wanted to use to modify the data weight of the row.

Participants explored encoding designs beyond these defaults, a set of examples from four participants is shown in Fig. 4. One participant [non-ML:03] used the design prob to prototype a "accordion view" with the bar chart; surprisingly, and independently, another participant [non-ML:05] had a similar request but described it instead of implementing it. Participant [non-ML:03] proposed that clicking on a column within the stacked bar chart would trigger the individual marks to reposition themselves next to each other. The participant demonstrated how they would then inspect the individual marks and modify

Fig. 4.   Examples of larger changes participants made, which include changing the visual encoding, adding new elements, and adding new components to the visual interface like comments.

their heights (prediction probability) and even introduce new genres by recoloring segments of the mark. This was the most complex strategy we observed and indicated to us that the visual design probes provided affordances for participants to creatively express themselves. Such expressivity is useful for challenging existing visual design ideas, as was the case for our probes, and iterating on encoding and interaction design. Other new marks participants proposed included adding links between marks (ML-04) to indicate similarity, and adding columns to the tables to add more data, such as data weights (see Section V-A1). Participants introduced new marks by repeating an existing one. For example, when wanting to express that a movie belongs to multiple classes, several participants choose the strategy of copying a mark and pasting it into multiple genres. Lastly, participants also explored introducing new, or alternative, channel encodings such as color gradients and transparency to encode multiple classes and prediction probability, respectively. Overall, both ML and non-ML experts used similar types of interactions and targeted similar data attributes for model steering.

Reasons not to Interact: Only two (of 20) participants declined to interact with any of the visual design probes. One of these participants was a non-ML expert and expressed concerns about injecting bias (other participants also expressed these concerns and we elaborate further in Section V-B1). The other participant was an ML expert and felt that visualizations were primarily for communication and not for model steering - they preferred to code a solution, but this is not an option for many non-ML experts. Other ML expert participants saw the value of interacting through visualizations to explore model steering for certain tasks. These ML participants also indicated instances when they did not want to refine the model but completely retrain it or select a new model. Three participants indicated that they would want to retrain the model on only the misclassified results. Two said that they would use the visualizations to decide whether to choose a different model. In these instances, ML experts felt it was better to drop into code instead of interacting through an interface.

Additional Properties of Interactions: We observed that individual interactions could also be both compounded and hierarchical. Compounded actions affect multiple data targets simultaneously, for example, moving a mark from one position to another, with the intent to change its class, should be accompanied by an automatic change to its probability (higher for the new class, lower for the new class), and vice versa. Actions were hierarchical when the participants wanted to make changes that should cascade down to individual points. Two participants described hierarchy actions via anchor points; these were points participants chose to represent a cluster and any modifications to those anchor points should be applied to others near it.

3) Feasibility and Utility of Interactions: We identified two factors impacting the feasibility of implementing the interactions participants demonstrated via our probes: whether it was reasonable to implement the interaction and whether it was possible to use the data generated by the interaction to update an ML model. First, some interactions, although possible to implement, might not be reasonable to scale. An example is using gradients to indicate multiple classes (i.e, a movie belongs to two categories). The design space we collect offers multiple

alternative strategies for encoding multiple classes, elicited by participants, that may be more favorable according to multiple participants and reasonable to implement, such as repeating a mark (see Fig. 2). Another factor of feasibility is whether the data generated by the interaction could actually be used in an ML training regimen. For example, participants expressed introducing new class labels, having multi-class assignments, as well as introducing hierarchy to data points. Not all classification models would be able to accommodate such interactions and prototype systems that failed to anticipate them would not be able to support them at all. In this case, the probes can suggest further directions for expanding ML models to incorporate user feedback. Importantly, it may also be used to help refine an ML model backed in a visual IML system. For example, if participants consistently introduce multiple classes or data hierarchy, as participants did in our study, then we could choose a model that supports the incorporation of such data. It may be possible, but challenging, to revise a more complex prototype to align with user expectations.

We can also consider the utility of interactions surfaced. We observed that the interactions surfaced by our visual design probes have also been identified in prior studies, which lends validity to the types of interactions our probes are able to elicit. For example, others have reported [44], we also observed behaviors of grouping points and annotating clusters, in one instance to create completely new clusters with labels that did not exist in the initial (training) dataset. When we compare our results to the comprehensive summary of cluster interactions in Bae et al. [50], we also find overlapping examples of participants adding or merging, modifications to the cluster's structure (i.e., moving marks in and out of clusters), adding constraints by anchoring marks to clusters, as well as correcting and validating perceived errors (i.e., Dr. Strangelove is not a Romance film, but a comedy). Actions to overwrite text in tables are similar to the behaviors of systems like StarSPIRE [34]. Moreover, some interactions that users undertook evoked model steering mechanisms that have also been proposed by prior works. Participants' desire to re-weight points is an updating approach that appears to be similar to iteratively refining a weak learner via boosting [65], but with manual intervention from a human. Re-weighting data in accordance with user interactions has also been used in other IML systems [33], [66].

While there are limitations on the full spectrum of interactions our visual design probes could capture, overall, the alignment between our findings and prior work is an encouraging demonstration of the probes' capacity for capturing useful interactions. Moreover, we can identify novel interactions with respect to the combination of targets and actions that prior work has not identified. For example, prior work does not identify probability as an interaction target for clustering and classification, yet the use of visual design probes surfaced it and identified several consistent interaction strategies (Figs. 2 and 3).

B. RQ2: Participants Perceptions and Concerns

In addition to capturing participants' interactions, we also dive more deeply into the factors that influence their interaction choices, including their hesitations. What is noteworthy about

participants' responses is the sensitivity of design choices on participants' willingness to interact with a system. For example, building a system with diagnostic plots to contextualize the effects of interactions would have been not useful to others, especially non-ML experts. Prior work has also shown that the magnitude of model changes is not always easy to capture [67]. Overall, these findings show the value of design probes to anticipate challenges proactively and influence the final system design.

1) Factors Impacting Interaction: Effects of Encoding Design on Interactions The choice of visual encoding impacted how 'inviting' it was to engage in an interaction. For one participant, the table was more inviting because when they see visualizations they are *"not sure how moving things around [in the scatter plot] change anything [and I] see a table as a natural place to provide row-level feedback."* [non-ML:06]. For other chart types, certain interactions were less desirable. For example, some participants would repeat a mark to indicate a movie belonged to two or more genres. Most participants performed this mark repetition interaction with the dot and scatter chart encodings but did not do so with the bar chart even though it was possible. When prompted, one participant offered that *"it just feels wrong"* (ML:04). There was also a general consensus among both ML and non-ML experts that the scatter chart was the most difficult to reason with. For many participants, the axes of scatter charts are tied to single concrete attributes (i.e., age, height, eye color, etc.) and it was difficult to reason about axes that represented embeddings or dimensionally reduced data. Participant non-ML:05 described the dissonance as 'L1-interference', not to be confused with L1-regularization, which occurs when concepts are translated from one's native language (L1) to a new language (L2). Their rationale was that using visualizations to steering models was like learning a new language that was similar, but different, from one that they were accustomed to.

Anticipating the Effects of Interactions: The choice of visual encoding was also tied to participants' ability to understand and contextualize the effects of their interactions. Participants expressed preferring simpler visual encodings because it was easier to anticipate the effects of their actions: *" I feel like engaging with the other charts I feel more confident [..] because[..] when I change position [on the scatter chart] I will change the distance to all the other points."* [ML:06]. Visual encodings also evoked 'degrees of freedom' that a participant could use to interact, with too many being overwhelming: *"I think I prefer this vis [bar chart]. There's just too much freedom in the scatter plot"* [ML:05].

One way to ameliorate this challenge is to include additional visuals for contextual information. Our visual design probes were intended to give participants the opportunity to raise these contextual needs - and they did so. Model explanations were commonly requested, especially amongst the non-ML experts. Participants with an ML background also wanted to view additional diagnostic plots to assess model performance, which they felt would help them better understand and contextualize the effects of their actions: *"I would love to be able to do this [action] and have the model spit back out, what's the AUC looks like"* [ML:04]. Non-ML experts did not express interest in seeing diagnostic plots. Interestingly, there was disagreement amongst participants towards the value of model diagnostic and performance metrics, with one participant summarizing this tension succinctly: *"I am little confused if what we're trying to do is see how good the model trains vs how well the model represents reality [...] you can have an accurate model or a representative model"* [ML:06]. Prior work in interactive ML systems also emphasizes the importance of building representative, over strictly accurate, models [68].

Concerns about Injecting Bias: Three participants (two non-ML experts and one ML expert) articulated strong concerns about injecting bias into the ML model as a primary reason they would hesitate to engage with a visualization to refine an ML model. For one non-ML expert, it was a reason not to engage at all – even if they expressed disagreement with the model results, they would still refrain from interacting due to concerns of bias: *"maybe I really disagreed with it[the classification] but it is my own view imposed on something cultural [a movie]"* [non-ML:08]. One non-ML expert also doubts their expertise to introduce model refining changes but would feel more confident if they knew safeguards existed: *"I really think that human safeguard with the expertise is a necessary step"* [non-ML:03]. Participants also indicated that the model refinements should be considered experimental and were not to be applied to production settings. Overall, these findings show that participants are aware of their potential to introduce bias and this impacts their willingness to steer the model. While additional components to the IML systems (i.e., model explanations) could help address these concerns, participants' comments also point to the importance of socio-technical structures that include peer safeguards as being important to them.

2) Learning From Interactions: As participants interacted with visual encodings, the study administrators initiated responses to their actions. We asked participants what their expectations were toward what the model was learning, how often it should update, and how they interpreted the responses.

Frequency of Model Updates: Participants wanted model feedback in real-time, but they wanted to control the frequency of updates and to only 'commit' to certain changes only once they were satisfied. This form of visual speculative execution [69] aligns with the refine-forecast model update modality described by Strobelt et al. [70].

Learn from Minor Examples, not Major Refinements: Participants expected the ML model to learn from a small number of interactions. Participants consistently interacted with the marks of movies they were more familiar with than others; they did not express the desire to make changes to those they were unfamiliar with. Additionally, all participants articulated prioritizing their interactions and placing different emphasis on different types of errors: 1) instances where the model was outright wrong (i.e., predicted that *Dr. Strangelove* is a romance); 2) instances where the model was not completely wrong but should be nudged in a different direction (i.e., whether *Die Hard* is an action or holiday movie); 3) instances where the participant disagreed with both the predicted and labeled genre and wanted to provide their own. Personal notions of how incorrect the model was, and the 'extent of error', also dictated how much weight the model should give

these modifications, with error correction given a higher weight than nudges; we discuss this observation in the next subsection (Section V-B2).

It is unlikely that even using their own dataset a participant is intimately familiar with each data point. We argue that participants will always have some prioritization strategy based on familiarity with some subset of the data over others. It is also important to consider that datasets are frequently reused both in research and in practice [71]. It may be both interesting and relevant to capture *which* subsets are explored and by whom. Interestingly, two participants astutely observed that many other people view and interact with these visualizations, and they wanted a way to see the changes others had made.

Incorporating Uncertainty to Interaction Inputs: Many participant concerns and hesitations toward interacting with data encoding marks of visualizations can be summarized by their desire to add uncertainty to their interactions with the model. Nearly all participants articulated that different types of interactions allowed them to convey a different amount of uncertainty. For example, changing the size of the mark, its color gradient, or transparency were seen as imprecise interactions meant to convey a larger sense of uncertainty. Conversely, overwriting a value or changing a mark's position was meant to convey less uncertainty. Participants had different preferences for conveying uncertainty. Some prefer less precision: *"feel that the confidence is more easily visualized here [...] a circle is much easier to say 'Oh I want to make it bigger or smaller compared to a number where I have to guess some random number [to change it to]"* [non-ML:06]. While others wanted more precision and control: *"I would want to move the points individually, there could be more precision like that [...] because then I can have more control over how confident I am reclassifying something.''* [non-ML:04]. The choice of visual encoding did impact a participant's ability and willingness to convey a notion of uncertainty. Data targets that participants acted on in other visual encodings were not targeted in others:

> Admin: In the previous vis you changed the confidence via the [mark] size, do you want to do that in the table?
> non-ML:07: No (emphatic) and I don't know why. I mean I guess because it's numbers, don't ever change the numbers.

Prior work on semantic interactions had considered human inputs to be a kind of 'soft data' [33] that the model must calibrate against what is learned from the training data. However, we find that participants also want to convey uncertainty more explicitly through their interactions than the notion of 'soft data' accommodates. Kim [72] introduced the concept of an *interacted latent variable* that more closely approximates participants' expectations. Interacted latent variables explicitly include the input uncertainty from the human to calibrate human inputs against what is learned from the data. In applications of a Bayesian Case Model, Kim [72] asks participants to explicitly provide their level of certainty via an input widget, but future work could explore inferring uncertainty directly from interactions.

C. RQ3: Impacts of ML Expertise

Finally, we consider how ML and non-ML experts differed in their interactions and perceptions. Individual differences aside, non-ML experts, in particular, saw the benefit of using a visual and interactive ML system to help them collaborate with their ML expert peers. This perspective from non-ML experts impacted what they wanted out of an IML system.

ML Experts and not Substitutes for non-ML Expertise: The preceding results have already alluded to some of the commonalities and differences between ML and non-ML experts. Here we summarize the key points. If we consider only the set of target-interaction pairs (Fig. 2) we see an overlap between the interactions ML and non-ML participants would undertake. We also see that both groups would interact with similar data targets (predicted genre, and probability). While participants from both groups would also introduce new data targets (i.e., Multi-class Genres), in our study, ML experts introduced more (data weights, similarity) compared to non-ML experts. The most evident difference was toward additional information participants expressed they needed to reduce cognitive load. Non-ML experts preferred explanations of the model's behavior, whereas ML experts wanted more performance metrics (Section V-B1). Both groups felt this additional information would inform their interactions and participants assess their effects. While both ML and non-ML experts at times declined to engage with the visual design probes, their reasoning was different (Section V-A2). However, personal preferences were more indicative of whether, and how, participants interacted with the visual encodings. In summary, our findings support prior research [11]. that, despite similarities, ML experts are not a sufficient proxy for non-ML experts when developing IML systems.

Working Collaboratively With Others Several non-ML experts also expressed the desire to comment on specific changes, either as a reference for themselves or with ML experts they collaborated with. One participant indicated they would use interactive model refinement to provide feedback to the ML experts on their team: *"I would love to prototype models because that would let me refine my own notes and recommendations to the person that is potentially designing the model"* [non-ML:03]. Prior research [20], [73], [74] has shown that such collaborative modes of interaction were desired but largely absent from existing tools. Moreover, recent studies show both the necessity and value of challenging the so-called ground-truth of ML data [75] by enabling wider collaboration between data workers [76], [77]; our study suggests that interactions could facilitate this collaboration.

D. Summary of Findings

Our study elicited 43 unique interactions that define a set of data target and interaction pairs (Fig. 2). Through these interactions, participants not only modified existing data, but felt that they could introduce new data (class genres, similarity, and data weights) to steer a classification model. We were also able to capture a myriad of considerations that guided participants' interaction - including hesitating or refining from interacting. We contextualize our findings between ML and non-ML experts,

showcasing similarities and differences between these groups. Overall, our findings both align with and extend interactions discovered in prior work (i.e., [38], [44], [50]), which demonstrates the utility and value of data and design probes for developing visual and interactive ML systems.

VI. DISCUSSION

Data and design probes provide a canvas for participants and researchers to co-create elements of an interactive system [14]. Gathering empirical data prior to the development and implementation of IML systems, including those used for visual analysis, is critical to aligning the design of these systems with the needs and expectations of end-users. Our research and its findings have explored the use of data and visual design probes for conducting pre-design empirical research toward the creation of visual and interactive machine learning systems. Collectively, our probes and elicitation study revealed that users consistently target specific types of data with a varied, yet consistent, set of actions (Fig. 2). We demonstrate that probes are capable of producing tangible artifacts, from participant utterances to their design choices, and we suggest that these can be used to stimulate a more informed discussion around the design and development of visualizations for IML systems. Furthermore, as ML and AI systems grow in complexity and implicate a wider group of stakeholders, it becomes even more important to align these systems with human needs and expectations. Our findings also show that probes are useful tools for teasing apart both differences and commonalities across the continuum of ML/DS expertise. We now reflect on the implications of our findings for developing visual and interactive machine learning systems.

A. Designing Visual & Interactive ML Systems

Our initial goal was to use data and design probes as a precursor to developing an IML system. However, the complexity of the interaction design space and context for these interactions gave us pause. Our findings suggested challenges and limitations of ML models and the incorporation of human refinement that impacts IML system design. While data and design probes do not resolve these issues, they appear to surface them and provide an empirical basis for discussing trade-offs of different approaches without prematurely committing to design choices

Human Goals and IML Brittleness: What surprised us about our findings was the tensions they surfaced between what people wanted to do and what present ML models are capable of. As an example, participants wanted to exercise control over both the class label and classification probability; Die Hard should be both a holiday and action movie but with a higher prediction probability on the action genre. Accommodating this goal is not straightforward. Should an ML model be refined to reflect an end-users expressed uncertainty toward specific data points and classes? Should the initial classification problem be flipped to be treated like a regression problem? Participants also added bespoke classes of movies, which introduce out-of-distribution issues should they wish to use their interactively refined model for further prediction. As one participant indicated (see Section V-B2), the desired features for a system tend to default to a

model-centric perspective, focusing on accuracy performance, and do not accommodate human-centric desires like the representativeness of data.

Learning From Human Refinement: Refining machine learning models through human feedback via visual encodings have been explored using both reinforcement [66] and active learning [45]. However, these studies present users with a fixed set of visual encodings and do not ask the critical question: do users want to interact? The answer from our study appears to be that it depends. Our findings suggest limitations in the amount of feedback participants wish to provide (Section V-B2 and V-A1), are impacted by the choice of visual encoding and their ability to understand them (Section V-B1), their comfort initiating such interactions (Section V-B1), and finally, their preferences to refine ML models via code not by interacting with visual encodings (Section V-A2.) While reinforcement learning from human feedback (RLHF) is showing promise for large language models [78], the efficacy of such an approach for other types of data is under-explored as is the medium of interaction (e.g., text prompts vs direct manipulation).

Our study may also have unintentionally captured refinement by voice because study administrators prompted participants to discuss the rationale, which the Wizard could also use to craft a response. Our study focused participants' attention on direct manipulations of visual encodings, thus the effect of voice was not directly under study and likely incidental. However, exploring such multi-modal interactions could be an interesting trajectory of future work.

Providing Feedback to Humans: Finally, an often overlooked dimension is communicating to people what the effects of their actions were. We left this consideration entirely open to participants to dictate, both with respect to the frequency (Section V-B2) and type (Section V-C) of feedback. Here we found that expertise dictated what participants expected to see, with those receiving more formal training expecting to see typical performance metrics while non-experts preferred explanations. However, what was more surprising was that the people also wanted *feedback from other humans* as part of the refinement process. Our results suggest two modes for an IML system. An individual mode that allows analysts of varying expertise to prototype and refine ML models and a complementary feedback mode that allows participants to collaborate with others in the refinement process. Per our results, the feedback mode may extend further to also show how different individuals interact with or annotate data encoding marks.

B. Limitations

The design choices and implementation of our probes as well as our study population and size are limitations of our research. While we made these deliberate choices with the intent to minimize the number of assumptions our approach made toward user needs and expectations, we also acknowledge the effects these choices have on our findings. We chose to develop a minimal set of probes, that made the least assumptions about the components of an IML interface, instead eliciting these aspects from participants and the ML model. For purposes of feasibility, we also

limited the number of data points in our study to fifty. While we provide a rationale for these and other choices, there are aspects of the interactive experience that our probes maybe not be able to capture. Scaling design probes, without compromising the value of free expression and creativity that these probes are intended to provide, is a fruitful topic for future work. These design choices may limit some of the generative power of the probes to elicit novel encodings and interactions. The probes are best suited to identify the alignment between targets and actions; the brittleness around this combination may stimulate discussion of the need for novel encodings and interactions, but are not guaranteed to derive them. To evaluate the trade-off between scalability and free expression of our research approach, we would need to implement a system. However, for the reasons described in Section VI-A, this too would have limitations for a proper assessment. We speculate that it may require several different systems that experiment with approaches to human refinement to fully evaluate these trade-offs. Lastly, the study size of 20 participants, while commensurate with visualization research on interactive machine learning and elicitation studies [38], [42], [44], [59], [61], [66], may not to capture the total spectrum of non-ML and ML expert perspectives. While these limitations are unlikely to hamper the value and utility of data and design probes for visualization research, they do provide an outline for the reasonable scope of our approach.

C. Future Work

There are several fruitful avenues for future work that expand on our research methodology and its findings. An immediate next step is to use the findings of our elicitation study and implement a visual and interactive machine learning system. However, our probes as point to many interesting challenges that cannot be solved by the implementation of a single system (Section VI-A) and that remain open research problems for future work. One interesting area of future work is to expand the definition of so-called "soft data" [33], by allowing people to add their own perceptions of uncertainty *into* model feedback. Prior work by Kim [72] provides an interesting avenue via inter-acted latent variables, which could be explored concomitantly with visualization research in Bayesian modeling [79], [80] and uncertainty [81], [82]. Another interesting area of future work is to leverage semantic interactions to investigate how people construct and label datasets. For non-ML experts, or perhaps even the lay public more generally, interacting through visual-izations may be a more accessible way to communicate their own personal notion of a "ground truth" [75] and even elicit richer contextual information than current approaches afford [45], [76], [77], [83], [84]. Finally, we see promise in further exploring the ways that semantic interactions can facilitate collaboration amongst teams of data workers and especially between ML and non-ML experts. This team collaboration appears to add safeguards to the processes of updating ML models as a greater number of non-ML experts begin to undertake this task. Of in-terest would be to examine whether semantic interactions could enable an asynchronous and multi-user collaboration around model refinement.

VII. Conclusion

Visual and interactive machine learning systems can provide avenues for data analysts with varied ML expertise to steer ML models with human guidance. However, designing such systems is uniquely hard and may make many assumptions about the needs and expectations of end-users. We examined the use of data and design probes to elicit how participants would interact with visual encodings, via direct manipulation, to update an ML model. Our findings show that the use of probes can not only capture a design space of interactions but also participants' perceptions and hesitations toward interacting with an ML model. Overall, our study demonstrates the value of expanding the visualization research toolbox to include data and design probes as a way of gathering formative empirical evidence for developing IML systems.

References

[1] C. Collins et al., "Guidance in the human–machine analytics process," *Vis. Inform.*, vol. 2, no. 3, pp. 166–180, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468502X1830041X

[2] D. Sacha et al., "What you see is what you can change: Human-centered machine learning by interactive visualization," *Neurocomputing*, vol. 268, pp. 164–175, 2017.

[3] F. Sperrle, A. Jeitler, J. Bernard, D. Keim, and M. El-Assady, "Co-adaptive visual data analysis and guidance processes," *Comput. Graph.*, vol. 100, pp. 93–105, 2021.

[4] F. Sperrle et al., "A survey of human-centered evaluations in human-centered machine learning," *Comput. Graph. Forum*, vol. 40, no. 3, pp. 543–567, 2021.

[5] N. Boukhelifa et al., "Challenges in evaluating interactive visual machine learning systems," *IEEE Comput. Graph. Appl.*, vol. 40, no. 6, pp. 88–96, Nov./Dec. 2020.

[6] J. J. Dudley and P. O. Kristensson, "A review of user interface design for interactive machine learning," *ACM Trans. Interact. Intell. Syst.*, vol. 8, no. 2, pp. 1–37, Jun. 2018. [Online]. Available: https://doi.org/10.1145/3185517

[7] B. Combemale et al., "A hitchhiker's guide to model-driven engineer-ing for data-centric systems," *IEEE Softw.*, vol. 38, no. 4, pp. 71–84, Jul./Aug. 2021.

[8] H. Subramonyam, C. Seifert, and E. Adar, "How can human-centered design shape data-centric AI?," 2021. [Online]. Available: https://sites.google.com/view/hcai-human-centered-ai-neurips/home

[9] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. Paritosh, and L. M. Aroyo, ""Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2021, pp. 1–15.

[10] A. Endert, L. Bradel, and C. North, "Beyond control panels: Direct manipulation for visual analytics," *IEEE Comput. Graph. Appl.*, vol. 33, no. 4, pp. 6–13, Jul./Aug. 2013.

[11] Q. Yang, J. Suh, N.-C. Chen, and G. Ramos, "Grounding interactive machine learning tool design in how non-experts actually build models," in *Proc. Designing Interactive Syst. Conf.*, New York, NY, USA, 2018, pp. 573–584.

[12] S. Mishra and J. M. Rzeszotarski, "Designing interactive transfer learning tools for ML non-experts," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–15.

[13] Q. Yang, A. Steinfeld, C. Rosé, and J. Zimmerman, "Re-examining whether, why, and how human-AI interaction is uniquely difficult to design," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13.

[14] E. B.-N. Sanders and P. J. Stappers, "Probes, toolkits and prototypes: Three approaches to making in codesigning," *CoDesign*, vol. 10, no. 1, pp. 5–14, 2014. [Online]. Available: https://doi.org/10.1080/15710882.2014.888183

[15] E. L. Hutchins, J. D. Hollan, and D. A. Norman, "Direct manipulation interfaces," *Hum. Comput. Interact.*, vol. 1, pp. 311–338, 1985.

[16] C. Graham and M. Rouncefield, "Probes and participation," in *Proc. 10th Anniversary Conf. Participatory Des.*, 2008, pp. 194–197.

[17] H. Subramonyam, C. Seifert, and E. Adar, "ProtoAI: Model-informed prototyping for AI-powered interfaces," in *Proc. Intell. User Interfaces*, 2021, pp. 48–58, doi: 10.1145/3397481.3450640.

[18] D. Lee, S. Macke, D. Xin, A. Lee, S. Huang, and A. G. Parameswaran, "A human-in-the-loop perspective on AutoML: Milestones and the road ahead," *IEEE Data Eng. Bull.*, vol. 42, no. 2, pp. 59–70, 2019. [Online]. Available: http://sites.computer.org/debull/A19june/p59.pdf

[19] F. M. Zanzotto, "Viewpoint: Human-in-the-loop artificial intelligence," *J. Artif. Intell. Res.*, vol. 64, pp. 243–252, Feb. 2019.

[20] A. Crisan and B. Fiore-Gartland, "Fits and starts: Enterprise use of AutoML and the role of humans in the loop," in *Proc. CHI*, 2021, pp. 1–15, doi: 10.1145/3411764.3445775.

[21] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, "The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems," 2021. [Online]. Available: https://arxiv.org/abs/2105.03354

[22] B. Shneiderman, "Human-centered artificial intelligence: Reliable, safe & trustworthy," *Int. J. Hum.–Comput. Interaction*, vol. 36, no. 6, pp. 495–504, 2020.

[23] Y. Gil et al., "Towards human-guided machine learning," in *Proc. 24th Int. Conf. Intell. User Interfaces*, 2019, pp. 614–624.

[24] D. Ceneda et al., "Characterizing guidance in visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 111–120, Jan. 2017.

[25] J. J. van Wijk, "The value of visualization," in *Proc. IEEE Vis.*, 2005, pp. 79–86.

[26] G. Kindlmann and C. Scheidegger, "An algebraic process for visualization design," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 2181–2190, Dec. 2014.

[27] A. McNutt, G. Kindlmann, and M. Correll, "Surfacing visualization mirages," in *Proc CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2020, pp. 1–16. [Online]. Available: https://doi.org/10.1145/3313831.3376420

[28] B. Shneiderman, "Direct manipulation: A step beyond programming languages," *Computer*, vol. 16, no. 8, pp. 57–69, 1983.

[29] L. Jiang, S. Liu, and C. Chen, "Recent research advances on interactive machine learning," *J. Vis.*, vol. 22, pp. 401–417, 2019.

[30] A. Endert et al., "The state of the art in integrating machine learning into visual analytics," *Comput. Graph. Forum*, vol. 36, no. 8, pp. 458–486, 2017.

[31] G. Ramos, C. Meek, P. Simard, J. Suh, and S. Ghorashi, "Interactive machine teaching: A human-centered approach to building machine-learned models," *Hum.–Comput. Interaction*, vol. 35, no. 5–6, pp. 413–451, Apr. 2020.

[32] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, "Improving fairness in machine learning systems: What do industry practitioners need?," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–16. [Online]. Available: https://doi.org/10.1145/3290605.3300830

[33] A. Endert, P. Fiaux, and C. North, "Semantic interaction for sensemaking: Inferring analytical reasoning for model steering," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 12, pp. 2879–2888, Dec. 2012.

[34] L. Bradel, C. North, L. House, and S. Leman, "Multi-model semantic interaction for text analytics," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2014, pp. 163–172.

[35] H. Lee, J. Kihm, J. Choo, J. T. Stasko, and H. Park, "iVisClustering: An interactive visual document clustering via topic modeling," *Comput. Graph. Forum*, vol. 31, pp. 1155–1164, 2012.

[36] J. Choo, C. Lee, C. K. Reddy, and H. Park, "UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.

[37] Ç. Demiralp, M. S. Bernstein, and J. Heer, "Learning perceptual kernels for visualization design," *IEEE Trans. Vis. Comput. Graph.*, vol. 20, no. 12, pp. 1933–1942, Dec. 2014.

[38] B. Saket, S. Huron, C. Perin, and A. Endert, "Investigating direct manipulation of graphical encodings as a method for user interaction," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 482–491, Jan. 2020.

[39] B. Saket, H. Kim, E. T. Brown, and A. Endert, "Visualization by demonstration: An interaction paradigm for visual data exploration," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 331–340, Jan. 2017.

[40] B. Hartmann, L. Abdulla, M. Mittal, and S. R. Klemmer, "Authoring sensor-based interactions by demonstration with direct manipulation and pattern recognition," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2007, pp. 145–154. [Online]. Available: https://doi.org/10.1145/1240624.1240646

[41] E. T. Brown, J. Lie, C. E. Brodely, and R. Chang, "Dis-function: Learning distance functions interactively," in *Proc. IEEE Conf. Vis. Analytics Sci. Technol.*, 2012, pp. 83–92.

[42] E. Wall, S. Das, R. Chawla, B. Kalidindi, E. T. Brown, and A. Endert, "Podium: Ranking data using mixed-initiative visual analytics," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 288–297, Jan. 2018.

[43] S. Gehrmann, H. Strobelt, R. Krüger, H. Pfister, and A. M. Rush, "Visual interaction with deep learning models through collaborative semantic inference," *IEEE Trans. Vis. Comput. Graph.*, vol. 26, no. 1, pp. 884–894, Jan. 2020.

[44] S. M. Drucker, D. Fisher, and S. Basu, "Helping users sort faster with adaptive machine learning recommendations," in *Proc. Int. Conf. Hum.-Comput. Interact.*, 2011, pp. 187–203.

[45] J. Bernard, M. Zeppelzauer, M. Sedlmair, and W. Aigner, "VIAL: A unified process for visual interactive labeling," *Vis. Comput.*, vol. 34, no. 9, pp. 1189–1207, Sep. 2018.

[46] J. Bernard, M. Hutter, M. Zeppelzauer, D. Fellner, and M. Sedlmair, "Comparing visual-interactive labeling with active learning: An experimental study," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 298–308, Jan. 2018.

[47] F. Hohman, K. Wongsuphasawat, M. B. Kery, and K. Patel, "Understanding and visualizing data iteration in machine learning," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2020, pp. 1–13. [Online]. Available: https://doi.org/10.1145/3313831.3376177

[48] D. Sacha et al., "Visual interaction with dimensionality reduction: A structured literature analysis," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 241–250, Jan. 2017.

[49] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu, "A survey of visual analytics techniques for machine learning," *Comput. Vis. Media*, vol. 7, no. 1, pp. 3–36, Mar. 2021.

[50] J. Bae, T. Helldin, M. Riveiro, S. Nowaczyk, M.-R. Bouguelia, and G. Falkman, "Interactive clustering: A comprehensive review," *ACM Comput. Surv.*, vol. 53, no. 1, pp. 1–39, Feb. 2020.

[51] H. Hutchinson et al., "Technology probes: Inspiring design for and with families," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2003, pp. 17–24.

[52] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale, "Empirical studies in information visualization: Seven scenarios," *IEEE Trans. Vis. Comput. Graph.*, vol. 18, no. 9, pp. 1520–1536, Sep. 2012.

[53] C. Broadley, "Visualising human-centred design relationships: A toolkit for participation," Ph.D. dissertation, The Glasgow School of Art, 2013. [Online]. Available: https://radar.gsa.ac.uk/4283/1/Cara_Broadley_Visualising_human-centred_design_relationships_thesis.pdf.

[54] S. Stumpf et al., "Interacting meaningfully with machine learning systems: Three experiments," *Int. J. Hum. Comput. Stud.*, vol. 67, no. 8, pp. 639–662, 2009.

[55] H. Subramonyam, C. Seifert, and E. Adar, "Towards a process model for co-creating AI experiences," in *Proc. Designing Interactive Syst. Conf.*, 2021, pp. 1529–1543.

[56] J. T. Browne, "Wizard of OZ prototyping for machine learning experiences," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2019, pp. 1–6. [Online]. Available: https://doi.org/10.1145/3290607.3312877

[57] D. Maulsby, S. Greenberg, and R. Mander, "Prototyping an intelligent agent through wizard of OZ," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 1993, pp. 277–284.

[58] G. Dove, K. Halskov, J. Forlizzi, and J. Zimmerman, "UX design innovation: Challenges for working with machine learning as a design material," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 278–288.

[59] F. Hohman, A. Head, R. Caruana, R. DeLine, and S. M. Drucker, "Gamut: A design probe to understand how data scientists understand machine learning models," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–13.

[60] M. Brehmer, S. Carpendale, B. Lee, and M. Tory, "Pre-design empiricism for information visualization: Scenarios, methods, and challenges," in *Proc. 5th Workshop Beyond Time Errors Novel Eval. Methods Vis.*, 2014, pp. 147–151.

[61] T. Hogan, U. Hinrichs, and E. Hornecker, "The elicitation interview technique: Capturing people's experiences of data representations," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 12, pp. 2579–2593, Dec. 2016.

[62] D. Xin, E. Y. Wu, D. Lee, N. Salehi, and A. Parameswaran, "Whither AutoML? Understanding the role of automation in machine learning workflows," 2021, *arXiv:2101.04834*.

[63] J. W. Creswell and C. N. Poth, *Qualitative Inquiry & Research Design: Choosing Among Five Approaches*, 4th ed. Newbury Park, CA, USA: Sage Publications, 2018.

[64] A. C. Valdez, M. Ziefle, and M. Sedlmair, "Priming and anchoring effects in visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 584–594, Jan. 2018.

[65] R. E. Schapire, "The strength of weak learnability," *Mach. Learn.*, vol. 5, pp. 197–227, 2005.

[66] M. El-Assady, R. Sevastjanova, F. Sperrle, D. Keim, and C. Collins, "Progressive learning of topic modeling parameters: A visual analytics framework," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 1, pp. 382–391, Jan. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8019825/

[67] A. Crisan and M. Correll, "User ex machina: Simulation as a design probe in human-in-the-loop text analytics," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2021, pp. 1–16.

[68] G. Bansal, B. Nushi, E. Kamar, W. S. Lasecki, D. S. Weld, and E. Horvitz, "Beyond accuracy: The role of mental models in human-AI team performance," in *Proc. AAAI Conf. Hum. Comput. Crowdsourcing*, 2019, pp. 2–11. [Online]. Available: https://ojs.aaai.org/index.php/HCOMP/article/view/5285

[69] F. Sperrle, J. Bernard, M. Sedlmair, D. Keim, and M. El-Assady, "Speculative execution for guided visual analytics," in *Proc. IEEE VIS Work. Mach. Learn. User Interact. Vis. Anal.*, 2018, pp. 1–8. [Online]. Available: https://learningfromusersworkshop.github.io/papers/SpecEx.pdf

[70] H. Strobelt, J. Kinley, R. Krueger, J. Beyer, H. Pfister, and A. M. Rush, "GenNI: Human-AI collaboration for data-backed text generation," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 1106–1116, Jan. 2022.

[71] B. Koch, E. Denton, A. Hanna, and J. G. Foster, "Reduced, reused and recycled: The life of a dataset in machine learning research," in *Proc. NeuroIPS*, 2021. [Online]. Available: https://arxiv.org/abs/2112.01716

[72] B. Kim, "Interactive and interpretable machine learning models for human machine collaboration," 2015. [Online]. Available: https://beenkim.github.io/papers/BKimPhDThesis.pdf

[73] D. Wang et al., "Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI," in *Proc. CSCW*, Nov. 2019, pp. 1–24, doi: 10.1145/3359313.

[74] S. K. Karmaker, M. M. Hassan, M. J. Smith, L. Xu, C. Zhai, and K. Veeramachaneni, "AutoML to date and beyond: Challenges and opportunities," 2021. [Online]. Available: https://arxiv.org/abs/2010.10777

[75] E. Denton, M. Díaz, I. Kivlichan, V. Prabhakaran, and R. Rosen, "Whose ground truth? Accounting for individual and collective identities underlying dataset annotation," 2021. [Online]. Available: https://arxiv.org/abs/2112.04554

[76] M. L. Gordon et al., "Jury learning: Integrating dissenting voices into machine learning models," in *Proc. CHI*, 2022, pp. 1–19, doi: 10.1145/3491102.3502004.

[77] A. M. Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," 2021. [Online]. Available: https://arxiv.org/abs/2110.05719

[78] N. Stiennon et al., "Learning to summarize from human feedback," in *Proc. NeuroIPS*, 2022. [Online]. Available: https://arxiv.org/abs/2009.01325

[79] Y.-S. Kim, L. A. Walls, P. Krafft, and J. Hullman, "A Bayesian cognition approach to improve data visualization," in *Proc CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2019, pp. 1–14. [Online]. Available: https://doi.org/10.1145/3290605.3300912

[80] Y.-S. Kim, P. Kayongo, M. Grunde-McLaughlin, and J. Hullman, "Bayesian-assisted inference from visualized data," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 02, pp. 989–999, Feb. 2021.

[81] L. Padilla, M. Kay, and J. Hullman, "Uncertainty visualization," Apr. 2020. [Online]. Available: psyarxiv.com/ebd6r

[82] X. Pu and M. Kay, "A probabilistic grammar of graphics," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, New York, NY, USA, 2020, pp. 1–13.

[83] O. Russakovsky, L.-J. Li, and L. Fei-Fei, "Best of both worlds: Human-machine collaboration for object annotation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2121–2131.

[84] J. C. Chang, S. Amershi, and E. Kamar, "Revolt: Collaborative crowdsourcing for labeling machine learning datasets," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2017, pp. 2334–2346.

Anamaria Crisan received the PhD degree in computer science from the University of British Columbia. She is a lead research scientist with Tableau. Her research explores the intersection of data science and data visualization. Her prior work has emphasized biomedical applications.

Maddie Shang received the degree in math from UWaterloo. She is a Sr. research engineer and team lead of OpenMined RecSys. Her open-source work in Federated Learning and Privacy was funded by PyTorch. Her interests include HCI via NLP and visual encodings, AI Ethics and reinforcement learning. She was named as one of 100 Brilliant Women in AI Ethics in 2022.

Eric Brochu received the PhD degree from the University of British Columbia in 2010, working on human-in-the-loop Bayesian optimization. He is a software engineering architect with Tableau. He currently leads Tableau's Machine Learning Engineering team.