Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems

Jaimie Drozdal
drozdj3@rpi.edu
Rensselaer Polytechnic Institute

Justin Weisz
Dakuo Wang
jweisz@us.ibm.com
dakuo.wang@ibm.com
IBM Research

Gaurav Dass
Bingsheng Yao
Changruo Zhao
dassg2@rpi.edu
yaob@rpi.edu
zhaoc6@rpi.edu
Rensselaer Polytechnic Institute

Michael Muller
michael_muller@us.ibm.com
IBM Research

Lin Ju
linju@ca.ibm.com
IBM

Hui Su
huisuibmres@us.ibm.com
IBM Research
Rensselaer Polytechnic Institute

ABSTRACT

We explore trust in a relatively new area of data science: Automated Machine Learning (AutoML). In AutoML, AI methods are used to generate and optimize machine learning models by automatically engineering features, selecting models, and optimizing hyperparameters. In this paper, we seek to understand what kinds of information influence data scientists' trust in the models produced by AutoML? We operationalize trust as a willingness to deploy a model produced using automated methods. We report results from three studies – qualitative interviews, a controlled experiment, and a card-sorting task – to understand the information needs of data scientists for establishing trust in AutoML systems. We find that including transparency features in an AutoML tool increased user trust and understandability in the tool; and out of all proposed features, model performance metrics and visualizations are the most important information to data scientists when establishing their trust with an AutoML tool.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Empirical studies in HCI*; • **Computing methodologies** → *Artificial intelligence*; *Machine learning*.

KEYWORDS

AutoAI, AutoML, AutoDS, Automated Artificial Intelligence, Automated Machine Learning, Automated Data Science, Trust

1 INTRODUCTION

The practice of data science is rapidly becoming automated. New techniques developed by the artificial intelligence and machine learning communities are able to perform data science work such as selecting models, engineering features, and tuning hyperparameters [26, 30, 33, 36, 68]. Sometimes, these automated methods are able to produce better results than people [38]. Given the current shortage of data scientists in the profession [8], automated techniques hold much promise for either improving the productivity of current data scientists [61] or replacing them outright [56].

Many companies and open source communities are creating tools and technologies for conducting automated data science [4, 9, 13, 15, 21, 42, 46, 49]. However, in order for these automated data science techniques – which we collectively refer to as "AutoML" – to become more widely used in practice, multiple studies have recently suggested that a significant hurdle in *establishing trust* must first be overcome [12, 34, 61, 62]. Can AI-generated models be trusted? What factors contribute to the trust of AutoML systems?

In this paper, we discuss users' **trust** in AutoML systems as it pertains to **transparency** and **understandability**, thus we believe it is necessary to clarify these three concepts. Our transparency concept derives from [40] that transparency of the automation is "the quality of an interface pertaining to its ability to afford an operator's comprehension about an intelligent agent's intent, performance, future plans and reasoning process." Understandability is the quality of comprehensibility in an automation tool. However, high transparency of a system does not necessarily lead to high understandability.

In this paper, we use a definition of trust as "the extent to which a user is confident in, and willing to act on the basis of, the recommendations, actions, and decisions of an artificially intelligent decision aid" [37]. Furthermore, we adopt an informational perspective of

trust [17, 19, 51, 67]: having information about how an AutoML system works, as well as the artifacts it produces (i.e. machine learning models), ought to increase trust in that system due to increased levels of transparency. Conversely, not having information about an AutoML system or its artifacts ought to decrease trust in that system. Thus, we seek to understand what kinds of information are important for an AutoML system to include in its user interface in order to establish trust with its users.

Current AutoML systems offer a myriad of information about their own operation and about the artifacts they produce. We coarsely group this information into three categories: information about the *process* of how AutoML works (e.g. information about the *data* (such as how it was transformed or pre-processed), how it performs feature engineering), and information about the *models* produced by the system (such as their evaluation metrics).

In this paper, we report results from three studies designed to evaluate how the inclusion or exclusion of these types of information impacts data scientists' trust in an AutoML system. Our first study is formative, consisting of a number of semi-structured interviews designed to capture the universe of information currently present across a representative sample of commercial AutoML products, and identify information that is not commonly included in those products. We hypothesize that the inclusion of this "hidden" information will increase the transparency of AutoML systems, and hence, increase the amount of trust users' place in it. Our second study is evaluative, designed to quantitatively test the impact of the inclusion of new information – what we refer to as "transparency features" – on ratings of trust. Our third study is an open card-sorting task that aims to understand the relative importance of different kinds of information in the AutoML "information space."

We focus on addressing two research questions in our work.

- **RQ1**. To what extent does the inclusion of new transparency features affect trust in and understanding of an AutoML system?
- **RQ2**. What information is highly important for establishing trust in an AutoML system? What information is not important?

Our results make a number of significant contributions to the existing literature on data science work and to the IUI community.

- We find quantitative evidence that the inclusion of transparency features – visualizations of input data distributions and a visual depiction of the feature engineering process – increases peoples' ratings of trust and understanding of an AutoML system.
- We provide a ranking of relative importance of different kinds of informational "nuggets" in establishing trust in an AutoML system.

We expect our work to inform the design of AutoML systems by highlighting the different types of informational needs data scientists have in order to establish trust in the system.

2 RELATED WORK

We first review literature on human-in-the-loop machine learning, focused on understanding the work practices and tool use of data scientists. We then discuss recent advances in automated data science, and summarize issues of trust and transparency in machine learning.

2.1 Human-in-the-loop Machine Learning

Data science is the process of generating insights from primarily quantitative data [32]. Often, data scientists leverage techniques from machine learning to build models that make predictions or recommendations based on historical data. Studies have suggested that data science work practices are different from traditional engineering work [14, 32, 44]. For example, Muller et al. [44] decomposed the data science workflow into four sub-stages, based on interviews with professional data scientists: data acquisition, data cleaning, feature engineering, and model building and selection. They argue that a data science task is more similar to a crafting work practice than an engineering work practice, as data scientists need to be deeply involved in the curation and the design of data, similar to how artists craft their work.

Wang et al. proposed a three stage framework of data science workflow with ten sub-steps [61]. It expands [44] workflow, which mostly focuses on the model training steps and takes into account the model deployment steps after the model is trained. In this paper, we adopt the framework of [61], as shown in Figure 1, and we focus mostly on data scientists' trust in the *model validation* sub-step of the *modeling* phase.
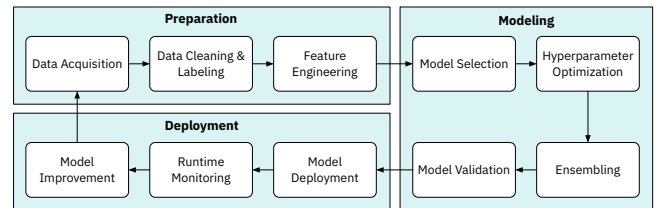
Figure 1: A data science workflow, consisting of three high-level phases: data preparation, model building, and model deployment [61]

CSCW researchers have also looked at collaborative aspects of data science work. Hou and Wang [20] conducted an ethnography study to explore collaboration in a civic data hackathon event where data science workers help Non-profit organizations develop insights from their data. Mao et al. [39] interviewed biomedical domain experts and data scientists who worked on the same data science projects. Their findings partially echo previous literature [14, 32, 44] results that data science workflow has multiple steps. In addition, they also suggest that data science is a highly collaborative effort where domain experts and data science workers need to work closely together to advance the workflow. Often times, the two parties are not "speaking the same language" [20] or do not have the common ground related to their goal [39]. Thus, a "broker role" in the team, who can bridge the two backgrounds and constantly re-calibrate the goal, may help to ease these tensions and support the success of this cross-discipline data science work.

These empirical findings guided the designers and system builders to propose human-in-the-loop data science design principles [1, 2,

11, 27, 28, 59]. Gil et al. surveyed papers about building machine learning systems and developed a set of design guidelines for building human-centered machine learning systems [11]. Amershi et al. in parallel reviewed a broader spectrum of AI applications and proposed a set of design suggestions for AI system in general [2], some of which are overlapping with the ones in [11].

With these design suggestions, more and more machine learning tools are built to support data scientists. For example, Jupyter Notebook is one of the successful examples [22]. It incorporates data scientists' three needs of coding, documenting narrative, and observing execution results [32] into a cohesive user interface. Thus, many data science workers adopt Jupyter notebooks as their primary working environment. Researchers have studied how data scientists use Notebooks [48, 50], how to build version control components into Notebooks [27], how to enable multi-users synchronous editing in Notebook [59], and other innovative designs.

In this paper, we build upon the understandings of how data science teams work and how to build systems to support data scientists work practice, but focus specifically on the scenario where data scientists interact with automated machine learning techniques in a data science work practice.

2.2 Automated Machine Learning (AutoML)

Automated Data Science or Machine Learning (AutoML for short) refers to systems that automatically select and optimize the machine learning model in each of the data science steps [62]. For example, AI researchers have developed various algorithms to automatically *clean data* [31]. The *feature engineering* step is tedious for human data scientists thus lots of solutions have been proposed to automatically generate new features and select the best subset of the features while balancing good model performance [25, 26, 33]. As for the *model selection and hyperparameter optimization* steps, data scientists are already relying on publicly available libraries such as Auto-sklearn [4] and TPOT [9], instead of writing code from scratch.

However, all of these automation technologies focus only on a single step and the data scientists still need to put various pieces of the puzzle together to assemble a model generation pipeline. End-to-end AutoML solutions have only recently become a reality. These technologies arguably can complete the entire data science workflow (as in Figure 1) from data acquisition to model selection, and then to model deployment and improvement. Large technology companies have released automated data science products, such as Google's AutoML [13], IBM's AutoAI [21], and Microsoft's Azure Studio [42]. Small startups such as H2O [15] and Data Robot [49] are also capturing significant market share.

With these latest developments, Automated Machine Learning becomes increasingly promising, and more and more researchers have begun to explore its user experience [11, 34, 61]. For example, Gil et al. reviewed existing literature on how data scientists use machine learning applications, from which they proposed design guidelines for the development of future AutoML systems [11]. Wang et al. interviewed 20 professional data scientists and asked their perceptions on AutoML technology [61]. They found that data scientists in general hold a positive position towards the collaborative future where "human and AI work together to build models".

Lee et al. referenced back to the "mixed-initiative" literature and argued that AutoML and human users can "collaborate efficiently to achieve [the] user's goals" [34], thus it is a human-in-the-loop perspective of AutoML.

This emerging group of empirical studies enrich our understanding of how data scientists think, interact, and collaborate with AutoML tools and generate useful design guidelines to improve an AutoML system's usability. However, as recent works suggested [12, 34, 47, 62], transparency and trust in this new AutoML technology is another major hindrance for large scale adoption. Thus, in the next subsection, we will focus on the transparency and trust issues of AutoML.

2.3 Trust in Machine Learning and in AutoML

AutoML is a relatively new topic, and therefore not many works have investigated trust issues of AutoML systems in particular - e.g., [11, 34, 61]. All of these works made design suggestions that, in order to make AutoML systems accessible and effortless for the data scientist users in the future, designers and system builders should present not only the final model results coming out of the system, but also the pipeline steps and decisions made in each of those steps along with the model generation process. This argument implies users demand higher transparency from AutoML system.

To accommodate the user needs of transparency and trust, a few recent works propose various design prototypes for increasing AutoML system transparency [12, 34, 47, 62]. For example, Wang et al. developed a first of its kind visualization systems, ATMSeer, that aims to open up the blackbox of an end-to-end AutoML system [62]. The authors argue that their tool can provide a multi-granularity visualization (both for the model selection as well as for hyperparameter selection) to the users so that they can monitor the AutoML process and adjust the search space in real time [62].

This approach of using visualizations to increase transparency of a system has been common in the traditional machine learning system designs [45, 60, 63]. For example, Google Vizer is a visualization tool that can reveal the optimization details of the hyperparameter tuning step [12]. In addition, it provides a visualization that shows the range of each hyperparameter in a model and the relationship between performance and hyperparameters. The authors hope that in this way, the users can understand how a final choice of the hyperparameter value is decided among alternative options (i.e., it leads to better model performance). VisualHyperTuner is a similar visualization-based system that focuses only on the hyperparameter tuning step [47].

All of these systems fall short in presenting an overview of the AutoML process and how each model pipeline was created. Dignum et al. suggests that transparency resides in not only the result of the model but also the data and the processes where the model has been generated [6]. This task is challenging for a visualization, as some of the steps in the AutoML pipeline have a categorical search space (e.g., various algorithms in the model selection step) but others have a continuous search dimension (e.g., hyperparameters values in the optimization step). Furthermore, among these AutoML steps, some have a sub-step or even a sub-sub-step. Thus, using one visualization to convey both the overview view of the pipeline

as well as the necessary details of each step in the workflow is challenging.

ATMSeer [62] presents a nice framework of proposing a visualization prototype on top of an existing AutoML system, and evaluating it with a number of users in a user study. Our paper adopts this research methodology and we propose low-fidelity design prototypes (based on the feedback from a Think-Aloud study) and evaluate these features with users on the perceived transparency level of AutoML. We hope these features can increase AutoML transparency and further promote better collaboration and trust between the data scientist and AutoML during the *Preparation and Modeling* phases (Figure 1). We leave the transparency topic in the *Deployment* phase for future work.

In particular, we join the group of researchers who argue that higher transparency of an AI system leads to higher trust by a user of that system [17, 35, 51, 66, 67]. Because transparency and trust in AutoML systems is a relatively new research topic, we want to build our baseline understanding about what information users need and how they need those information from an AutoML user interface. We designed a think-aloud pilot study as our first of a series of three studies. Based on the findings from study 1 with four data science students, we operationalize the transparency of an AutoML system into various design features such as "showing the distribution of data column (data-oriented transparency)" or "showing how AutoML performs feature engineering (process-oriented transparency)". We hope a comparison user study (study 2) between users who are and those who are not exposed to these transparency features can reveal the quantifiable differences in their trust in an AutoML system. Throwing more information to users can always increase the transparency of AutoML, but we are in danger of users' cognitive overload. To prioritize users information needs, we designed a study 3 that asks users to do a card-sorting task to prioritize those needs.

In what follows, we will start reporting our study designs and findings of the three user studies in order.

3 STUDY 1: THINK-ALOUD EVALUATION

We conducted a small pilot study with four computer science Graduate students to understand the current "information landscape" of AutoML systems. Participants were asked to use four popular commercial AutoML products while thinking aloud about the tool's information design and how it affected their feelings of trust in the tool.

Participants were presented with four tools, in random order: Google's AutoML Tables [13], Microsoft's Azure Machine Learning Studio [42], IBM's Watson Studio AutoAI [21], and H2O's Driverless AI [15]. We selected these tools based on their popularity, the fact that they require no coding from the user, and the fact that they automate the entire machine learning workflow from data preparation to model evaluation. Participants were given 30 minutes with each tool, with the task of generating and evaluating models for the Titanic dataset [24].

Participants were asked to think aloud as they interacted with each tool. Participants were also prompted with questions such as, "is this tool showing you everything you need at this moment?" and "how is this [feature/visualization/information] affecting your trust

in the tool?" At the end of reviewing all four tools, participants were asked which of the four tools they preferred.

Our procedures were approved by the Institutional Review Board (IRB) at our university, and participants provided written informed consent before participating. Participants were not compensated for their participation in this study.

3.1 Results

Overall, transparency was a highly-desired feature from all participants. Specifically, transparency of both *data* and *models* were mentioned while interacting with all four of the tools.

Participants' feedback was also varied, with many individual differences in preferences. Each of the four participants preferred a different tool. Two participants were hyper-focused on information about data and consistently commented on the lack of support for having conversations with the data [44], either in raw or processed form. Another participant cared most about the model selection process and noted the lack of transparency around the set of candidate models that the tools considered when performing model selection.

All participants expressed a lack of understanding about the different processes used by the AutoML tools.

> *"I have no idea what was done to the data." (P2)*
>
> *"I would not use this software because it's not clear what is happening when the experiment is running." (P3)*
>
> *"Feature engineering is complicated and important, but I don't know how it's doing it." (P4)*

Based on feedback from participants in this study, we focused on understanding how two transparency features might affect trust in an AutoML tool: visualizations of data distributions, and a visual depiction of how the feature engineering process works. We explored the effect of these transparency features in Study 2, discussed in the next section. Additional feedback from this study regarding the different kinds of information present in each of the AutoML interfaces was used in the design of Study 3, discussed in Section 5.

4 STUDY 2: INCREASING TRUST VIA TRANSPARENCY FEATURES

Study 1 suggested that commercial AutoML systems at the time of the study were lacking in transparency. In this study, we conducted a more detailed examination of how increasing transparency would affect trust. We recognize that transparency may be provided at different levels:

- data-oriented transparency (e.g. showing data distributions for the columns in the training set)
- process-oriented transparency (e.g. how AutoML performs feature engineering or hyperparameter optimization)
- model-oriented transparency (e.g. showing various accuracy metrics on a validation set)

In this study, we compared a baseline AutoML user interface with one that included additional *transparency features* that provided additional insight into the distributions of input features (data-oriented transparency) and the process by which the AutoML engineers new features (process-oriented transparency). These transparency features were chosen due to the overwhelming amount of

feedback from participants in Study 1, who felt that both of these features were important, yet missing from the AutoML systems. They may appear simple, but at the time of the study these features did not exist in a majority of AutoML tools. Not to mention, while including distributions of input features as a transparency feature might be argued as overly simplified the future of AutoML is one in which the entire data science process is captured in a single environment. To evaluate these transparency features, we used screenshots from a commercially-available AutoML system, removing references to the company's name and logo, in order to provide both a realistic AutoML experience, and one in which we could easily incorporate the transparency features. Specifically, we used **IBM Watson Studio AutoAI** [21]. Figure 2 shows examples of how these transparency features were shown to participants.

4.1 Participants

We recruited 21 participants who had prior experience with machine learning to complete our study. One participant was dropped from our study due to a lack of knowledge about machine learning that was uncovered during the course of the study; thus, our final sample consists of $N = 20$ participants.

Of the 20 participants, 5 (25%) were Undergraduate students and 15 (75%) were Graduate students. Five participants (25%) were female and 15 were male (75%), which is slightly higher than the proportion of women in data science (16.8%) reported in the 2018 Kaggle data science survey [25]. Students' areas of study included information technology (35%), computer science (20%), business analytics (15%), mathematics (10%), quantitative finance (10%), and other disciplines (10%).

4.2 Procedure

Prior to participating in our study, participants listened to the nature of the study and its procedures. Participants provided written informed consent. Our study was reviewed and approved by the Institutional Review Board at our institution.

Participation in the study took approximately one hour, and participants received a $10 USD gift card for their time.[1] We began by giving participants a packet of information about the dataset, which included column meanings, and all of the documentation on the tool available online from the provider. These handouts were to be used as reference materials while completing the study. Next, the experiment proceeded in two phases – reviewing AutoML-produced models (described in the next section) and a card-sorting task (described in Section 5).

4.2.1 Reviewing AutoML-produced models. In this phase, participants completed a sequence of two tasks. Each task consisted of reviewing a packet [2] of information about an AutoAI "run" for the Census-Income dataset [7]. This dataset is used to train a model that predicts whether a loan application should be approved from a set of over 35 different factors.

In this run, AutoAI produced a set of four pipelines. Each pipeline consisted of a series of steps, such as model selection, feature engineering, and hyperparameter optimization.[3] Participants reviewed a number of details about these pipelines, such as performance metrics and confusion matrices. Participants were asked by the researcher whether they trusted any of the four models enough that they would use them in a real deployment. After this, participants filled out a questionnaire containing demographic questions and different Likert scale questions measuring trust and understandability described in the following section.

To compare different user experiences of AutoAI, we developed three separate packets of information. The "V1" packet represents the base user interface provided by the commercial AutoML system. From this, we developed two variants, "V2A" and "V2B," by adding in additional visualizations of the input data and feature engineering process, respectively. We outline the informational content of each packet in Table 1. In the study, all participants saw the V1 packet, but each participant saw only one of the V2A or V2B packets. Packets were shown in a random order to control for order effects, and equal numbers of participants saw the V1 and V2 packets first.

Information	V1	V2A	V2B
Data			
Raw data table.	✓	✓	✓
Charts of input feature distributions		✓	
Process			
Visualization of pipeline creation process	✓	✓	✓
Feature engineering process diagram			✓
Model			
Metrics (ROC AUC, accuracy, F_1, etc.)	✓	✓	✓
ROC curve	✓	✓	✓
Precision Recall curve	✓	✓	✓
Confusion matrix	✓	✓	✓
Feature importance chart	✓	✓	✓
Feature transformation table (for pipelines that included feature engineering)	✓	✓	✓

Table 1: Information included in each of the V1, V2A, and V2B packets. All participants saw the V1 packet, but only one of the V2A and V2B packets. The packets were presented in random order.

4.3 Measures

For each of the phase 1 packets, we asked participants whether they trusted any of the pipelines produced by AutoAI enough to use them in a real deployment of an AI system. In addition, we asked participants to rate their overall trust in the AutoML system, based on the Merritt scale [41] reported in Hoffman et al. [19]. This scale treats trust as an attitudinal judgement of the degree to which a person can rely on an automated system to achieve their goals. We included 4 items from this scale related to confidence and dependability, and adapted them to our particular situation: "I believe the tool is a competent performer," "I trust the tool," "I have

---

[1]The local minimum wage at the time of study was $9.70.

[2]We opted not to run the actual AutoAI system during the course of the study as it would take too much time. Therefore, we ran the system on the data on ourselves and compiled a series of screenshots from the tool's user interface for participants to review.

[3]We use the terms "pipeline" and "model" interchangeably to refer to the resulting output of the AutoAI process: a binary classifier for making loan approval decisions. Some, but not all pipelines generated by AutoAI include feature engineering and hyperparameter optimization.
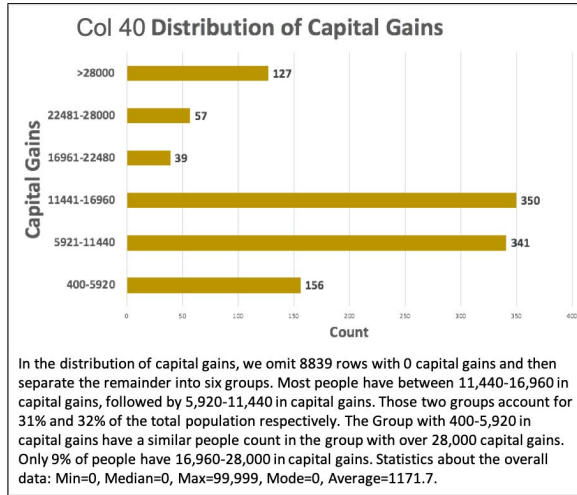
(a) Example of input data distribution for one of the columns in the data set. Written explanations accompanied each distribution to provide additional insight in interpreting the chart, as a proxy for the subject matter expertise typically available when conducting data science work.
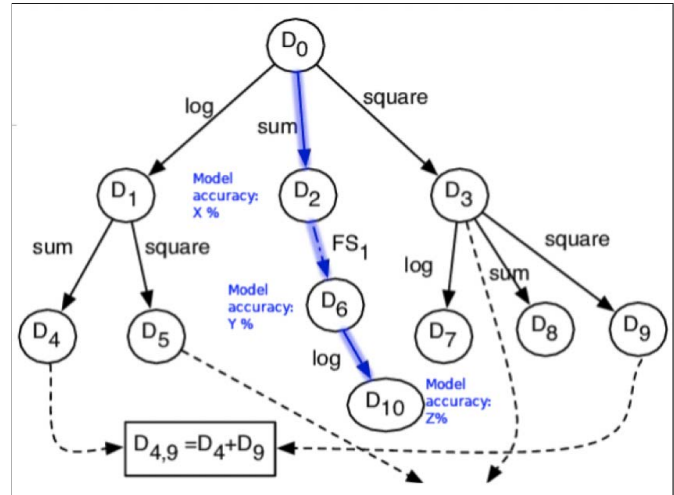
(b) Diagram illustrating the feature engineering process. Nodes represent data set versions and lines represent transformations applied to one of the data set's features, producing a new "view" of the data set. D_0 is the base data set with no transformations applied. At each step, accuracy is evaluated to determine if the transformed features provide an increase in model performance. The blue highlight represents how a sequence of feature transformations results in a model with a higher accuracy [30][29].

Figure 2: (a) Input data distribution transparency feature. (b) Feature engineering transparency feature.

confidence in the advice given by the tool," and "I can depend on the tool." These items were rated on 5-point Likert scales (Strongly Disagree to Strongly Agree).

Another aspect of trust has to do with understandability [19, 37]. We developed a 9-item scale to assess the degree to which participants understood the AutoML tool presented in each packet. Whereas the Madsen-Gregor scale [37] for evaluating perceived understandably contains high-level statements applied to a whole system (e.g. "I understand how the system will assist me with decisions I have to make"), we desired a scale that included items specifically focused on aspects of AutoML systems. Our scale included the following items: "I understand the tool," "I understood the tool's overall process," "I understand the data," "I understand how the tool performs data preprocessing," "I understand how estimators are selected," "I understand how new features are generated by the tool," "I understand the differences between the generated models," "I understand the model evaluation metrics," and "I understand the model evaluation visualizations." These items were also rated on 5-point Likert scales.

Finally, we included one more way to assess trust, by asking participants whether they felt they trusted any of the four models enough that they would use them in a real deployment. Given the high-stakes nature of the task (loan approval), we felt that participants would only answer in the affirmative if they truly understood and trusted the pipelines produced by the AutoML tool.

4.4 Results

We begin our analysis by evaluating the reliability of our trust and understandability scales. Then, we examine the effect of the transparency features on trust and understanding of the AutoAI system [21].

When conducting analyses of variance, we controlled for the effects of gender, education level (undergraduate or graduate), prior experience with AutoML systems (used previously or not), and the first packet seen by including these terms in the model. In addition, when making comparisons between the V1 and V2 packets (a within-subjects factor), we include participant ID in the model as a random effect. We report effect sizes from our ANOVA models using partial η^2, which corresponds to the proportion of variance accounted for by each main effect, controlling for all other effects [4].

4.4.1 Reliablity. Factor analysis [55] indicated a high degree of reliability[5] for the trust scale (Cronbach's $\alpha = 0.89$). Thus, despite our minor modifications to the wording of scale items, we see that its reliability falls within the $[0.87 − 0.92]$ range previously reported in Hoffman et al. [19].

Analysis of the understandability scale also indicates a high level of reliability (Cronbach's $\alpha = 0.80$), without the need for dropping any items. Therefore, we construct two outcome measures for trust

---

[4]Miles & Shevlin [43] advise that a partial η^2 of $\geq .01$ corresponds to a small effect, $\geq .06$ to a medium effect, and $\geq .14$ to a large effect.
[5]Reliability indicates the extent to which items in the scale measure the same underlying conceptual construct. Common convention holds that α values greater than 0.70 are considered reliable, and we refer to Tavakol and Dennick [54] for a more detailed discussion.

and understandability based on the averaged responses for each set of questions.

4.4.2   Effect of Transparency Features on Trust and Understandability. Overall ratings of trust for the V1 packet fell in the middle of the scale (M (SD) = 3.2 (.81) of 5), indicating that participants had neutral feelings about their trust of the base AutoML system. Ratings of understandability were higher (M (SD) = 3.6 (.52) of 5), indicating that they although they generally understood the information presented, there was also room for improvement.

We first compare the effect of having *either* of the transparency features on ratings of trust and understandability. Participants had more trust in the V2 packets (M (SD) = 4.1 (.62)) than the V1 packets, and this difference was significant and large, $F[1, 19] = 19.5, p < .001$, partial $\eta^2 = .36$. Prior experience with AutoML was also a marginally significant predictor of trust, $F[1, 15] = 3.2, p = .09$, partial $\eta^2 = .06$. Participants with prior AutoML experience had more trust in both the V1 and V2 packets than participants without prior AutoML experience.

Participants had a greater understanding of the V2 packets (M (SD) = 4.0 (.54)) than the V1 packets, and this difference was also significant and large, $F[1, 19] = 21.9, p < .001$, partial $\eta^2 = .49$. Unlike trust, participants with prior AutoML experience did not differ in their ratings of understandabilty, $F[1, 15] = .17, p = n.s.$

We observe that the inclusion of either transparency feature caused a significant increase in ratings of trust and understandability over the base AutoML interface. We next seek to understand the extent to which each individual transparency feature – input data distributions and feature engineering process – affected ratings of trust and understandability. As these are between-subjects comparisons (participants only experienced one of these features), we no longer include participant ID as a random effect in our ANOVA model.

Ratings of trust for the input data distribution variant were higher (V2A M (SD) = 3.9 (.65)) than ratings of trust for the feature engineering process variant (V2B M (SD) = 4.2 (.60)), although this difference was not significant, $F[1, 14] = .02, p = n.s.$ Ratings of understandability were equivalent for both variants (V2A M (SD) = 4.0 (.51), V2B M (SD) = 4.0 (.59), $F[1, 14] = .12, p = n.s.$). Therefore, we conclude that the inclusion of *any* transparency feature improved both trust and understandability, but the relative importance of either feature on improving trust and understandability remains unclear.

4.4.3   Would Participants Deploy AutoML Models? The decision to deploy an AutoAI [21] model was significantly correlated with both trust (Pearson's $r = .67, p < .001$) and understandability (Pearson's $r = .38, p = .02$). In general, participants did *not* trust the V1 models enough for deployment: 17 participants (85%) said they would not deploy any of the pipelines produced in V1. In contrast, participants trusted the V2 models more, with 16 participants (80%) saying that they would deploy one of the pipelines produced in V2.

As with the previous results, we do not see clear differences between the V2A and V2B variants: 7 of 10 participants having V2A would deploy one of its pipelines, and 9 of 10 participants having V2B would deploy one of its pipelines. This difference was not significant, $\chi^2 = .31, p = n.s.$

We again find evidence that the inclusion of a transparency feature improved trust, but the relative importance of the two features we examined remains unclear.

5   STUDY 3: ELICITING INFORMATION NEEDS

In order to understand peoples' information needs in an AutoML user interface, we conducted a card-sorting exercise in which participants rank-ordered individual "nuggets" of information that might be included in an AutoML UI. This information was based on our examination of the kinds of information present across the AutoML interfaces of multiple vendors (discussed in Section 3).

Each card took the form of a verb (i.e. "see," "visualize," "know how") followed by a piece of information related to the AutoML tool. For example, one nugget was "view pre-processed data," and another was "know how features are engineered."

5.1   Participants

This study ran concurrently with Study 2. First, participants completed the task described in Section 4. Next, they they completed the card-sorting task described below. The same set of participants completed both Study 2 and Study 3 in the same session.

5.2   Procedure

In this study, we used an open card-sorting method [52, 53] to gain insight into what information is important for data scientists in order to trust the models produced by an AutoML tool. Participants were provided with 27 cards, each with a different nugget of information pertaining to an AutoML run based on our results from Study 1. Participants were also provided with blank cards to fill in additional information needs they identified. Participants were asked to sort the cards from "most important for trust" to "least important for trust."

5.3   Results

The card sorting exercise provided us with the opportunity to better understand the relative importance of different pieces of information that might be reported by AutoML. Although we performed an open card sorting task, in which participants were able to write in new informational requirements on blank cards, only four participants opted to do this. Therefore, we give an overall accounting of how cards were sorted by first ignoring these new cards, and then we provide detail on the content of these new cards.

We analyzed the card sorting results by computing the mean rank assigned to a card across participants. These ranks are shown in Table 2. Each card is categorized in two ways: which aspect of the process it represents (process, data, or model), and whether the information is rendered as a visualization.

5.3.1   Information Needs for Establishing Trust. Perhaps not surprisingly, information pertaining to the performance of the generated pipelines was rated as the most important for trust, either as a raw metric (#26) or in visual form (#27). Also important were several aspects related to process: knowing how data were pre-processed before training (#10) and being able to view the process by which a pipeline was created (#23). Other process-related information

#	Aspect	Transp.	Description
Most important for trust			
26	E	Model	View evaluation metrics
27	E	Model	View visualizations of model performance
10	PPD	Process	Know how raw data was pre-processed
2	RD	Data	View the meanings of each column in the raw data
4	RD	Data	Visualize each column's distribution in the raw data
5	RD	Data	Visualize the raw data - view overall distributions
23	P	Process	View process of how a pipeline is created
7	RD	Data	View the raw data - statistics of individual distributions
3	RD	Data	Visualize outliers in the raw data
8	RD	Data	View statistics of missing values in the raw data
11	PPD	Data	View statistics of the pre-processed data
17	FE	Model	Effect of engineered features
12	PPD	Data	Visualize data after pre-processing
19	P	Model	Show adopted models in output pipelines
6	RD	Data	View statistics of outliers in raw data
15	FE	Data	View how existing features were engineered into new features
24	P	Model	Ability to edit a pipeline
25	E	Model	Compare differences between pipelines
1	RD	Data	View the raw data table
18	P	Process	Show which types of models considered for model selection
9	PPD	Data	View the pre-processed data table
20	E	Model	Compare one model against other models
14	FE	Model	View new engineered features
13	RD	Data	See how data was split (test vs. train/holdout)
16	FE	Process	Know how features were engineered
21	HP	Model	See model's hyperparameters
22	HP	Process	Know how hyperparameter optimization was performed
Least important for trust			

Table 2: Ranking of importance of different kinds of information in an AutoML user interface. Items at the top of the list were rated as being more important for establishing trust in an AutoML system. Informational aspects are: *RD*: raw data, *PPD*: pre-processed data, *HP*: hyperparameters, *P*: pipeline, *E*: model evaluation, *FE*: feature engineering.

was deemed less important, such as knowing which types of models were considered for model selection (#18) and knowing how hyperparameter optimization was performed (#22).

In our analysis of the Study 2 results, we did not see a significant difference in ratings of trust between participants who were given visualizations of input data distributions (#5) and participants who were given information on how feature engineering worked (#16). However, from the card-sorting exercise, we see a clear difference in how participants ranked the importance of these two features: participants felt that being able to see input data distributions was more important.

We were struck by participants' relative lack of interest in feature engineering (FE in Table 2, mean rank of 19/27). Participants were more interested in the raw data (mean rank of 11/27) and the model evaluation metrics (mean rank of 11/27). However, in the course of data science work, raw data are transformed and engineered into features. Therefore, data scientists' relative lack of interest in the design of these features [10, 44] should be examined in future research.

6 DISCUSSION

The goal of our research is to explore trust in the relationship between human data scientists and AutoML systems. Based on previous literature [17, 51, 62], we were interested in how the inclusion of transparency features in an AutoML system affects user trust and understanding of the tool (RQ1) and identifying the most important information data scientists need to establish trust in an AutoML system (RQ2).

We find that including certain transparency features such as visualizations of input data distributions and a graphic depicting the feature engineering process does improve user trust and understandability of AutoML. The inclusion of either transparency feature had a significant and sizable effect on both trust and understandability. Although we were unable to uncover a statistical difference between the two transparency features in Study 2, results from the card-sorting exercise (Study 3) showed users clearly ranked the "input data distributions" feature more important. Our findings support the hypothesis that the increase of transparency through adding additional information significantly lead to the increased trust and understandability. Thus, we provide support for RQ1, and we propose that future work should use a more sensitive, full repeated-measures experimental design, to clarify RQ2.

6.1 Implications for Design and Future Research

6.1.1 Visualizations. Our studies illustrate that conversations with the input data [10, 44] and information on the processes of feature engineering are valuable. Future work should explore the effectiveness of different types of data visualizations for both input data and also post-feature-engineered data. How do users want to examine these data? Are there mini-visualizations that can be applied "in place" in a pipeline diagram? What are effective ways to "sample" the data at different points in a pipeline. Will it be possible to recover transformations or feature-engineering algorithms from past analyses [27] for comparison with current data and features?

If users will be sampling data along a pipeline, then we may also want to explore effective *comparative* visualizations. Similarly to the ManyEyes project [58] with data analysts, we may also want to explore data scientists' needs for annotation and communication regarding data visualizations along a pipeline. What individual problems do data scientists want to solve in this way? What collaborative problems do they want to address collectively, and what types of messages would be helpful?

6.1.2 Individual Differences and Personalization. We recognize that there are many individual differences across those who practice data science [3, 32]: difference in backgrounds including background knowledge, skills, work practices, and experience levels make it difficult to claim that AutoML tools ought to be designed as "one size fits all" [20]. We found a wide range of individual difference

and preferences, even within our (relatively) small sample of 24 participants. For example, while thinking aloud during Study 3, six participants explicitly stated that information pertaining to the raw data was most important for them. In contrast, three participants said that the processed data was more important than the raw data, and two participants commented that the raw data was the least important and they had no desire to see this information in the tool. These differences in individual preferences can create more complications when we consider the fact that domain experts and data science workers need to work closely together, as suggested in [20, 39].

Our findings suggest that AutoML tools may need to allow for a degree of personalization to accommodate individual preferences or different domains of use. Recent research by Arya et al. [3] addresses these concerns by defining explanation methods for different audiences and domains.

6.1.3 Context of Use. In all three of our studies, participants identified a dichotomy between using AutoML for research purposes and using it in their day-to-day work practice. From these discussions, we recognized that the importance of different kinds of information depends on the intended use of the tool as well. For example, in Study 3, the two cards relating to hyperparameter optimization (#21, #22) had the lowest mean rank. But we should not conclude that this information is therefore unimportant. Further qualitative interview confirmed our speculation: two participants explicitly mentioned that information about hyperparameter optimization would be more important if the scenario of use was focused on conducting research rather than building models to approve loans.

In addition, we discovered through discussions with our participants that there is a difference between trust in a model produced by AutoML and trust in an AutoML tool itself. One participant in Study 2 commented that even though they may produce a model with low accuracy and not wish to deploy it, they would still maintain their trust in the AutoML tool itself. This disparity of trusting AutoML's artifacts versus trusting the AutoML itself is one research topic that ought to be explored further.

6.2 Limitations

There are several limitations of our work that may limit our ability to draw broad conclusions from our findings. First, our participants were drawn from a pool of undergraduate and graduate students having prior experience in data science work. However, their experience does not necessarily generalize to that of professional data scientists, whose information needs for establishing trust in AutoML may differ. Given the preponderance of AutoML systems being developed by and for enterprise users (e.g. [13, 15, 21, 42, 49]), additional work is needed to examine the viewpoints of professional data scientists. As a future work, we plan to validate our results with different user groups.

The dataset used in this study is a widely-used loan application in data science training [23]. This application domain, together with some other domains (e.g., healthcare and jurisdiction), are crucially important to the involved individuals, families, and businesses. Recent literature [16, 57] have suggested some of the datasets or AI algorithms may inherit the discrimination against people of color, or women, or women of color, in approving loans. Future

work can further investigate transparency features of presenting bias-detection and bias-mitigation information as part of the initial trust establishment.

It should also be noted that the card sorting task performed by the participants who are first-time users of AutoML system generated information needed for establishing trust and may not represent the needs of data scientists who have an established relationship with an AutoML system. The literature suggests that trust formation and trust retention are different and therefore they require different considerations in AutoML systems [18, 51].

In addition, during Study 3, participants were asked to sort the cards on a scale of most important to least important for establishing trust. As participants thought aloud about their sorting decisions, some noted that the ranks of the cards could change over time as they continued to interact with the tool. Therefore, the results of the card rankings may only represent the information needed to *establish initial trust*, which may be different from information needed to maintain trust.

7 CONCLUSION

As the AI industry is expected to grow significantly in the next few years [5], research on the relationship between user trust and automated data science systems is critical to ensure these tools can be trustworthy enough to be adopted responsibly by the public. Transparency is known to be a significant factor in trusting automated data science systems. However, much of the literature states a lack of transparency in AI [5, 64, 65]. We believe exploring the information needs and individual differences of data scientists can inform us of possible ways to increase trust in AutoML tools. Therefore, in this work we showed that increasing transparency via providing a user with more information about an AutoML tool significantly increased user trust as well as user understandability in the tool. By gathering the information requirements of data scientists to establish trust in these tools we have provided a pool of transparency features that can each be further researched to see how they impact user trust. As our work suggests it may be unreasonable to design an AutoML tool suitable for all users across all domains, we encourage the data science, HCI, and human studies research communities to continue exploring how to accommodate different AutoML users and enhance their trust in the tools based on their domains, knowledge, and common practices.

REFERENCES

[1] Saleema Amershi, Bongshin Lee, Ashish Kapoor, Ratul Mahajan, and Blaine Christian. 2011. Human-guided machine learning for fast and accurate network alarm triage. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 3.

[3] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques. *arXiv preprint arXiv:1909.03012* (2019).

[4] auto sklearn. [n. d.]. auto-sklearn. Retrieved 06-Oct-2019 from https://automl.github.io/auto-sklearn/master/

[5] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. 2019. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* 8, 8 (2019), 832.

[6] Virginia Dignum. 2017. Responsible autonomy. *arXiv preprint arXiv:1706.02513* (2017).

[7] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[8] Jen DuBois. 2019. Is There a Data Scientist Shortage in 2019? Retrieved 06-Oct-19 from https://blog.quanthub.com/is-there-a-data-scientist-shortage-in-2019

[9] EpistasisLab. [n. d.]. tpot. Retrieved 06-Oct-2019 from https://github.com/EpistasisLab/tpot

[10] Melanie Feinberg. 2017. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 2952–2963.

[11] Yolanda Gil, James Honaker, Shikhar Gupta, Yibo Ma, Vito D'Orazio, Daniel Garijo, Shruti Gadewar, Qifan Yang, and Neda Jahanshad. 2019. Towards human-guided machine learning. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM, 614–624.

[12] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D Sculley. 2017. Google Vizier: A Service for Black-Box Optimization. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1487–1495.

[13] Google. [n. d.]. Cloud AutoML. Retrieved 06-Oct-2019 from https://cloud.google.com/automl/

[14] Philip J Guo, Sean Kandel, Joseph M Hellerstein, and Jeffrey Heer. 2011. Proactive wrangling: mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. ACM, 65–74.

[15] H2O. [n. d.]. H2O. Retrieved 06-Oct-2019 from https://h2o.ai

[16] Sara Hajian, Francesco Bonchi, and Carlos Castillo. 2016. Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. ACM, 2125–2126.

[17] Peter A Hancock, Deborah R Billings, Kristin E Schaefer, Jessie YC Chen, Ewart J De Visser, and Raja Parasuraman. 2011. A meta-analysis of factors affecting trust in human-robot interaction. *Human factors* 53, 5 (2011), 517–527.

[18] R. R. Hoffman, M. Johnson, J. M. Bradshaw, and A. Underbrink. 2013. Trust in Automation. *IEEE Intelligent Systems* 28, 1 (Jan 2013), 84–88. https://doi.org/10.1109/MIS.2013.24

[19] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint arXiv:1812.04608* (2018).

[20] Youyang Hou and Dakuo Wang. 2017. Hacking with NPOs: collaborative analytics and broker roles in civic data hackathons. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 53.

[21] IBM. [n. d.]. AutoAI. Retrieved 06-Oct-2019 from https://www.ibm.com/cloud/watson-studio/autoai

[22] Project Jupyter. [n. d.]. Jupyter Notebook. Retrieved 3-April-2019 from https://jupyter.org

[23] Kaggle. [n. d.]. Kaggle: Your Home for Data Science. Retrieved 3-April-2019 from https://www.kaggle.com

[24] Kaggle. [n. d.]. Titanic: Machine Learning from Disaster. Retrieved 05-Jul-2019 from https://kaggle.com/c/titanic/data

[25] Kaggle. 2018. Kaggle Data Science Survey 2018. Retrieved 17-September-2019 from https://www.kaggle.com/sudhirnl7/data-science-survey-2018/

[26] James Max Kanter and Kalyan Veeramachaneni. 2015. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 1–10.

[27] Mary Beth Kery, Bonnie E John, Patrick O'Flaherty, Amber Horvath, and Brad A Myers. 2019. Towards Effective Foraging by Data Scientists to Find Past Analysis Choices. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 92.

[28] Mary Beth Kery, Marissa Radensky, Mahima Arya, Bonnie E John, and Brad A Myers. 2018. The story in the notebook: Exploratory data science using a literate programming tool. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 174.

[29] Udayan Khurana, Horst Samulowitz, and Deepak Turaga. 2018. Feature engineering for predictive modeling using reinforcement learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[30] Udayan Khurana, Deepak Turaga, Horst Samulowitz, and Srinivasan Parthasrathy. 2016. Cognito: Automated feature engineering for supervised learning. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. IEEE, 1304–1307.

[31] Georgia Kougka, Anastasios Gounaris, and Alkis Simitsis. 2018. The many faces of data-centric workflow optimization: a survey. *International Journal of Data Science and Analytics* 6, 2 (2018), 81–107.

[32] Sean Kross and Philip J Guo. 2019. Practitioners Teaching Data Science in Industry and Academia: Expectations, Workflows, and Challenges. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 263.

[33] Hoang Thanh Lam, Johann-Michael Thiebaut, Mathieu Sinn, Bei Chen, Tiep Mai, and Oznur Alkan. 2017. One button machine for automating feature engineering in relational databases. *arXiv preprint arXiv:1706.00327* (2017).

[34] Doris Jung-Lin Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *Data Engineering* (2019), 58.

[35] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

[36] Sijia Liu, Parikshit Ram, Deepak Vijaykeerthy, Djallel Bouneffouf, Gregory Bramble, Horst Samulowitz, Dakuo Wang, Andrew Conn, and Alexander Gray. 2019. An ADMM Based Framework for AutoML Pipeline Configuration. arXiv:cs.LG/1905.00424

[37] Maria Madsen and Shirley Gregor. 2000. Measuring human-computer trust. In *11th australasian conference on information systems*, Vol. 53. Citeseer, 6–8.

[38] Susan Malaika and Dakuo Wang. 2019. AutoAI: Humans and machines better together. https://developer.ibm.com/articles/autoai-humans-and-machines-better-together/

[39] Yaoli Mao, Dakuo Wang, Michael Muller, Kush Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilovic. 2020. How Data Scientists Work Together With Domain Experts in Scientific Collaborations. In *Proceedings of the 2020 ACM conference on GROUP*. ACM.

[40] Joseph E Mercado, Michael A Rupp, Jessie YC Chen, Michael J Barnes, Daniel Barber, and Katelyn Procci. 2016. Intelligent agent transparency in human–agent teaming for Multi-UxV management. *Human factors* 58, 3 (2016), 401–415.

[41] Stephanie M Merritt. 2011. Affective processes in human–automation interactions. *Human Factors* 53, 4 (2011), 356–370.

[42] Microsoft. [n. d.]. Azure Machine Learning Studio. Retrieved 06-Oct-2019 from https://azure.microsoft.com/en-us/services/machine-learning-studio/

[43] Jeremy Miles and Mark Shevlin. 2001. *Applying regression and correlation: A guide for students and researchers*. Sage.

[44] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM, 126.

[45] Judith S Olson, Dakuo Wang, Gary M Olson, and Jingwen Zhang. 2017. How people write together now: Beginning the investigation with advanced undergraduates in a project course. *ACM Transactions on Computer-Human Interaction (TOCHI)* 24, 1 (2017), 4.

[46] Randal S Olson and Jason H Moore. 2016. TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Workshop on Automatic Machine Learning*. 66–74.

[47] Heungseok Park, Jinwoong Kim, Minkyu Kim, Ji-Hoon Kim, Jaegul Choo, Jung-Woo Ha, and Nako Sung. 2019. VisualHyperTuner: Visual analytics for user-driven hyperparameter tuning of deep neural networks. In *Demo at SysML Conference*.

[48] Samir Passi and Steven J Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 136.

[49] Data Robot. [n. d.]. Data Robot: Automated Machine Learning for Predictive Modeling. Retrieved 06-Oct-2019 from https://datarobot.com

[50] Adam Rule, Aurélien Tabard, and James D Hollan. 2018. Exploration and explanation in computational notebooks. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 32.

[51] Keng Siau and Weiyu Wang. 2018. Building trust in artificial intelligence, machine learning, and robotics. *Cutter Business Technology Journal* 31, 2 (2018), 47–53.

[52] Donna Spencer. 2009. *Card sorting: Designing usable categories*. Rosenfeld Media.

[53] Donna Spencer and Todd Warfel. 2004. Card sorting: a definitive guide. *Boxes and Arrows* 2 (2004).

[54] Mohsen Tavakol and Reg Dennick. 2011. Making sense of Cronbach's alpha. *International journal of medical education* 2 (2011), 53.

[55] Bruce Thompson. 2004. *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

[56] Pedro Uria-Recio. 2018. Can Artificial Intelligence replace Data Scientists? Retrieved 06-Oct-19 from https://towardsdatascience.com/can-artificial-intelligence-replace-data-scientists-e4d4d828e31e

[57] Wil MP van der Aalst, Martin Bichler, and Armin Heinzl. 2017. Responsible data science.

[58] Fernanda B Viegas, Martin Wattenberg, Frank Van Ham, Jesse Kriss, and Matt McKeon. 2007. Manyeyes: a site for visualization at internet scale. *IEEE transactions on visualization and computer graphics* 13, 6 (2007), 1121–1128.

[59] April Yi Wang, Anant Mittal, Christopher Brooks, and Steve Oney. 2019. How Data Scientists Use Computational Notebooks for Real-Time Collaboration. (2019).

[60] Dakuo Wang, Judith S. Olson, Jingwen Zhang, Trung Nguyen, and Gary M. Olson. 2015. DocuViz: Visualizing Collaborative Writing. In *Proceedings of CHI'15*. ACM, New York, NY, USA, 1865–1874.

[61] Dakuo Wang, Justin D. Weisz, Michael Muller, Parikshit Ram, Werner Geyer, Casey Dugan, Yla Tausczik, Horst Samulowitz, and Alexander Gray. 2019. Human-AI Collaboration in Data Science: Exploring Data Scientists' Perceptions of Automated AI. *To appear in Computer Supported Cooperative Work (CSCW)* (2019).

[62] Qianwen Wang, Yao Ming, Zhihua Jin, Qiaomu Shen, Dongyu Liu, Micah J Smith, Kalyan Veeramachaneni, and Huamin Qu. 2019. Atmseer: Increasing transparency and controllability in automated machine learning. In *Proceedings*

of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, 681.

[63] Daniel Karl I. Weidele. 2019. Conditional Parallel Coordinates. *IEEE transactions on visualization and computer graphics* 26, 1 (2019).

[64] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do You Trust Me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA '19).* ACM, New York, NY, USA, 7–9. https://doi.org/10.1145/3308532.3329441

[65] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19).* ACM, New

York, NY, USA, Article 279, 12 pages. https://doi.org/10.1145/3290605.3300509

[66] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* ACM.

[67] Ruijing Zhao, Izak Benbasat, and Hasan Cavusoglu. 2019. Do users always want to know more? Investigating the relationship between system transparency and users' trust in advice-giving systems. In *Proceedings of ECIS 2019.*

[68] Marc-André Zöller and Marco F Huber. 2019. Survey on Automated Machine Learning. *arXiv preprint arXiv:1904.12054* (2019).