Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop

Anamaria Crisan
Tableau Research
acrisan@tableau.com

Brittany Fiore-Gartland
Tableau Software
bfioregartland@tableau.com

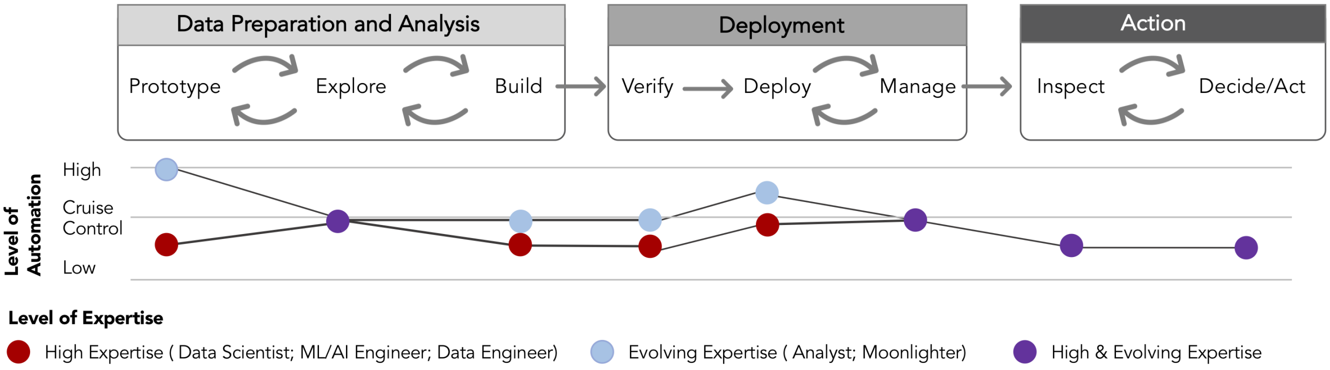


Figure 1: Levels of Automation in Data Science Work. From our interviews we illustrate the desired level of automation according to level of technical expertise in data science. We ground our findings in the levels of automation proposed by Parasuraman et al. [41] and Lee et al. [34]

ABSTRACT

AutoML systems can speed up routine data science work and make machine learning available to those without expertise in statistics and computer science. These systems have gained traction in enterprise settings where pools of skilled data workers are limited. In this study, we conduct interviews with 29 individuals from organizations of different sizes to characterize how they currently use, or intend to use, AutoML systems in their data science work. Our investigation also captures how data visualization is used in conjunction with AutoML systems. Our findings identify three usage scenarios for AutoML that resulted in a framework summarizing the level of automation desired by data workers with different levels of expertise. We surfaced the tension between speed and human oversight and found that data visualization can do a poor job balancing the two. Our findings have implications for the design and implementation of human-in-the-loop visual analytics approaches.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**.

KEYWORDS

Data Science, Automation, Machine Learning, Data Scientists

ACM Reference Format:
Anamaria Crisan and Brittany Fiore-Gartland. 2021. Fits and Starts: Enterprise Use of AutoML and the Role of Humans in the Loop. In *CHI '21: ACM Conference on Human Factors in Computing Systems, May 08–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Organizations are flush with data but bereft of individuals with the technical expertise required to transform these data into actionable insights [45]. To bridge this gap, organizations are increasingly turning toward automation in data science work beginning with the adoption of techniques that automate the creation of machine learning models [13, 46]. However, the adoption of this technology into enterprise settings has not been seamless. Currently AutoML offerings have limitations in what they can flexibly support. End-to-end systems encompassing the full spectrum of data science work, from data preparation to communication, are not yet fully realized [34, 54]. Consequently, AutoML systems still require human intervention to be practically applicable [42, 46]. This mode of human-machine collaboration presents a number of challenges [2, 35], chief among them being the importance of balancing the speed afforded by AutoML with the agency of individuals to interpret, correct, and refine automatically generated models and results [20]. Data visualization can play an important role in facilitating this human-machine collaborative process [20, 46], but there are few studies

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
CHI '21, May 08–13, 2021, Yokohama, Japan
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

that examine if and how data visualization is used in real-world settings together with AutoML. To fill this gap, we conduct interviews with 29 individuals from organizations of different sizes and that extend across different domains to capture how they currently, or plan to, use AutoML to carry out data science work. We examine specifically if and how participants use data visualization as a way to integrate the human in the automation loop.

Our investigation reveals that the practical use of AutoML technology in real world settings requires considerable human effort. This effort is complicated by the need to trade-off data work between individuals with different expertise, for example data scientists and business analysts. This trade-off is exacerbated by a data knowledge gap that participants believe AutoML technology is widening. While participants saw the value of data visualization as one way to facilitate human-in-the-loop interactions with AutoML tools, many still reported using visualization in a limited way. Participants found that creating quality visualizations for AutoML was often too difficult and time consuming and had the effect of slowing down automation often with limited benefit. Moreover, participants reported a lack of useful visualization tools to support them in some of their more pressing needs, such as collaborating on data work among their diverse teams and with AutoML technology. Altogether our study makes the following timely contributions to the existing literature on AutoML and the design of human-in-the-loop tools for data science:

- An interview study that presents real world uses of AutoML technologies in enterprise settings with a focus on the role of the human-in-the-loop facilitated by data visualization
- A summary of three use cases for AutoML according to different organizational needs
- A framework that illustrates the level of automation that is desirable for individuals with different levels of technical expertise.

As AutoML systems continue to gain traction in enterprise settings, our contributions will be a resource to the research communities developing human-in-the-loop approaches that support an appropriate balance of automation and human agency.

2 RELATED WORK

We review prior work that investigates the use of AutoML in data science, the ways that humans act within these processes, and current data visualization approaches that mediate these processes.

As we reviewed this work, we were challenged by the varied use of the term 'AutoML'. The preliminary goals of automation in machine learning began with the objective of removing the human specifically from hyper-parameter tuning and model selection steps [50]. However, it quickly became clear that other steps, such as data preparation or feature engineering, were also critical to the success of hyper-parameter tuning. The scope of the term AutoML, and more recently "AutoAI" or "driverless AI", began to encompass broader steps in the data science workflow [46, 54]. We observed that the terms AutoML, AutoAI, and the phrase 'automation in data science' are often used interchangeably in the literature. Here, we use the term AutoML to broadly encompass automation across multiple data science steps, from preparation to monitoring to deployment.

2.1 AutoML in Data Science

Data science leverages techniques from machine learning to derive new and potentially actionable insights from real-world data [3, 6, 12]. AutoML systems have been developed to automate the computational work involved in building a data analysis pipeline that enable individuals to derive these insights from data. Several commercial systems already exists and are used within different types of organizations, including AWS SageMaker AutoPilot [24], Google's Cloud AutoML [26], Microsoft's AutomatedML [29], IBM's AutoAI [28], H20 Driverless AI [27], and Data Robot [25]. There are also implementations of AutoML that build upon widely used data science packages, such as the scikit-learn [43] python library, autosklearn [15, 16] and TPOT [38, 39]. The focus of these AutoML systems are toward largely supervised tasks concerning feature engineering, hyper-parameter tuning, and model selection [14, 50, 54]. Recent innovations have proposed possible end-to-end solutions that also support data preparation [34, 50, 54] and it is likely that AutoML technologies will continue to expand toward broader end-to-end support.

The means and extent to which AutoML systems integrate with a computational data science pipeline is variable. Some AutoML systems exist as a single component within a larger pipeline, such as automated feature selection step, that the analyst or data scientist creates. At other times, AutoML systems can also *create* these pipelines with minimal user input. In their comprehensive analysis of existing AutoML tools, Zöller et al. [54] describe three common configurations for including AutoML in data science work. The two configurations are "fixed structure pipelines", where the AutoML system assumes a very specific configuration of computational pipeline. The authors differentiate between fixed pipelines that are optimized for specific AutoML methods (for example, neural networks or random forests) compared to those that are not. While these fixed systems are common, they have limitations when confronted with different data types and tasks. For example, image data or text data demand more flexibility within the structure of the computational pipeline. The second category is a "variable structure pipeline", which refers to a fairly recent approach that aims to learn the appropriate steps within a data science pipeline [54]. TPOT [38, 39] is an example of one of the first variable pipelines. Unlike fixed models that *execute* a pre-determined set of processes, variable structure approaches *learn* a network of process in response to different datasets and user objectives.

While the stated goal of many of these AutoML systems is to effectively remove humans from many aspects of data science work [50], a view that data scientists themselves express [46], today these systems still rely on considerable human labor to be of use [19]. These limitations stem from both the complexity of data science work and the brittleness of fixed structure pipelines that are in common use [54]. Our study catalogs this human labor across data science work and examines how visualization is used by individuals engaged in data work.

2.2 Automation and the Human-in-the-loop

Human-in-the-loop approaches provide a way to explicitly incorporate human interaction within automated processes. Identifying

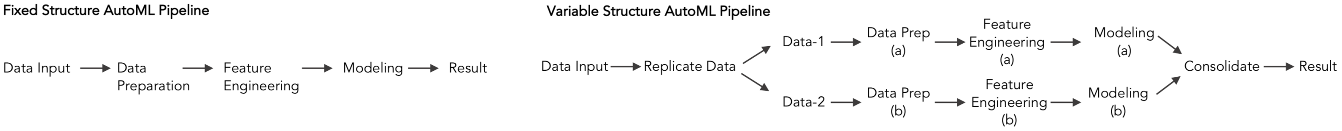


Figure 2: Example Illustrations of Fixed and Variable Structure Pipelines. Adapted with modification from Zöller et al. [54]

when and how to add the human-in-the-loop within AutoML processes is important in order to appropriately balance the speed that automation affords with the importance of human guidance. Parasuraman et al. [41] proposed a model to help designers identify the appropriate type and level of automation for information-seeking processes. They define four broad functions for how automation is used : 1) information acquisition; 2) information analysis; 3) decision and action selection; and 4) action implementation. They argue that the level of automation, from none to fully automated, should be evaluated against human performance consequences, automation reliability, and costs of actions. When the impact of automation is both significant and potentially harmful, human intervention is essential. The question of when, how, and how much to automate remains critical to the discussion of AutoML technologies today.

A number of recent studies in the HCI literature have examined this trade-off between automation and human intervention as it relates to AutoML technology.

Lee et al. [34], Gil et al. [17] and Liao et al. [35] describe a set of interaction modalities for users to engage with AutoML systems. Lee et al. [34] categorizes Parasuraman's et al. [41] levels of low to high automation into three different modes of interaction: 'user-driven', 'cruise-control', and 'autopilot'. In 'cruise control' a user directs an AutoML algorithm to a set of possible configurations to explore, as opposed to specifying a single and immediate next configuration. As an example, configuration can mean the user setting a parameter for hyper-parameter tuning during model creation. Gil et al. [17] describe a framework for human guided machine learning (HGML), which is predicated on the ability to effectively map user actions to a so-called 'AutoML planner' capable of translating and executing the action. Similar to Gil e al., Liao et al. [35] proposes a declarative way for the user to specify their objectives while allowing the system to automatically generate the underlying processes. By their descriptions, the systems proposed by Lee et al., Gil et al., and Liao et al. are akin to variable structure pipelines that were described in the previous section, in that they learn the processes in the pipeline. Studies have also examined human-ML/AI collaboration as it pertains to model authoring and interpretation specifically. While these studies are not exclusive to AutoML, they highlight key challenges for interacting with and interpreting machine learning models in enterprise settings. As an example, Hong et al. [22] interviewed 20 individuals across different domains (the majority of whom identified as data scientists), and found that collaboration among different organizational roles was of chief importance for operationalizing machine learning models into organizational practices.

Honeycutt et al. [21], Liao et al. [35], and Amershi et al. [2] describe the ways that information can be shared between humans and AutoML systems throughout a variety interactions. Honeycutt et al. [21] identifies 'relevance feedback' and 'incremental learning' as two general ways that humans can provide feedback to AutoML systems. Humans can provide relevant feedback, which informs the AutoML systems about whether its actions were effective or not. For example, humans may provide labeled data or correct errors when they arise. Humans may also provide new information in the form of incremental feedback to AutoML systems, which can be used to correct for issues like concept drift in models that have been deployed into production settings. Liao et al. and Amershi et al. focus on the flow of information in the opposite direction, which concerns the types of information humans require to interpret the results of from AutoML systems. Liao [35] conducted interviews with 20 UX design practitioners using a question bank to surface limitations in guidance targeting the development of explainable AutoML technologies. Their work demonstrates that the importance of the ML/AI results and their presentation is highly dependent on the question posed by the individual. Finally, Amershi et al. [2] proposes a comprehensive set of 18 design guidelines that outline the appropriate modes of interaction when experts, a) initially interact with an AutoML system; b) as the system is churning; c) when errors surface; and d) throughout user interactions.

Studies that examine how people use AutoML technologies and how they respond to human-in-the-loop features are also emerging. Wang et al. [46] interviewed 20 data scientists across industries to interrogate their practices and perceptions of AutoML. They found the benefits of AutoML for augmenting, but not replacing, human intuition were valued and appreciated by practitioners. Passi et al. [42] conducted an extensive six month ethnographic study that involved over 50 data scientists. Their findings surface the different organizational needs and challenges of data workers as they collaborated with each other in the context of automation in data science work. Zhang et al. [53], Drozal et al. [13], and Honeycutt et al. [21] conducted controlled experiments to evaluate decision making and trust in AutoML technologies, but their studies did not recruit current practitioners. Both Zhang et al. and Honeycutt et al. conducted their research via Mechanical Turk and Drozal et al. recruited undergraduate and graduate students in quantitative disciplines. Zhang et al. and Honeycutt et al. both found that reporting accuracy data alone was not sufficient for improving confidence and trust in the results produced by AutoML systems. Honeycutt et al. observed that the act of *interacting* with a machine learning model reduced confidence of individuals in the model's performance even when the human guidance increased accuracy. These findings by Zhang et al. and Honeycutt et al. underscore the challenges of designing useful feedback mechanisms between humans and AutoML systems.

While many human-in-the-loop approaches to support AutoML processes, and by extension data science work, exist, there are few studies aimed at understanding how they are integrated by practitioners in enterprise settings. We found two studies that concretely explore AutoML in enterprise, and we build upon these findings in our present study in order to further assesses attitudes of individuals in enterprise settings toward human-in-the-loop approaches.

2.3 Data Visualization and AutoML

Our work specifically focuses on visualization systems that support human-in-the-loop interactions for AutoML. Two prior and comprehensive state-of-the-art surveys capture the role of visualization in explaining [7] and building trust [8] in machine learning. Recent work by Yuan [51] demonstrates visual analytic approaches throughout the data science process, including prior to model building (data prep and feature engineering), during model building, and after model building (verification, deployment). These surveys show the diversity of approaches that are taken to support decision making throughout the data science pipeline. Here, we highlight five systems that collectively capture this diversity. Google Vizier [18] and ATMSeer [47] a) surface the complex latent space of models, b) search this space through interaction and visualization, and c) triage machine learning models. These systems present users with results from multiple models across their hyper-parameters through multiple coordinated views of the data. As with GoogleVizier, PipelineProfiler [40] and AutoAIViz [48] make use of parallel coordinate plots to help users navigate the model search space and to highlight possible hyper-parameter settings. AutoAIViz shows the utility of conditional parallel coordinates plots to visualize subsequent steps in an AutoML pipeline based upon the user's current selections. One limitation of visualization for AutoML in data science pipelines is the assumption of a fixed structure (see Section 2.1), making it difficult to visually compare variable AutoML pipelines. To address this limitation, PipelineProfiler was developed as a wrapper for the auto-sklearn [15] package, supporting the visualization and comparisons of different end-to-end AutoML implementations.

Taken together, we believe that these systems represent a 'cruise-control' mode of including a human-in-the-loop, balancing between the slower 'user-driven' and faster, but less transparent, 'autopilot' modes for executing and interacting with AutoML. Moreover, these systems, co-created with experts in design and data science, represent real implementations of the existing design guidance toward the use of visualization to help interrogate AutoML systems. However, it remains to be understood how such systems that are intended to build trust or transparency in AutoML actually get used, or perhaps more concerning, whether they get used at all. Our study sought to surface the visualization strategies within AutoML in enterprise settings.

2.4 Situating our Research

The current state of the art in AutoML is informed by multidisciplinary research endeavours spanning machine learning, human computer interaction, and visualization. Given this research effort, there exist a number of AutoML offerings with varying types of pipeline configurations, from fixed to variable, and that support different modes of interaction so that "intelligent services and users may collaborate efficiently to achieve the user's goals" [23]. However, there remain few studies on how this technology is applied in enterprise settings, whether users can effectively leverage the benefits of this technology, and how adding the human-in-the-loop via visualization is viewed by enterprise users. Moreover, existing studies [22, 42, 46] looking at enterprise settings focused on specific themes, namely collaboration and trust, and did not closely examine how AutoML broadly intersects with data science work. Building on these prior findings, our study conducts a broader examination of AutoML and data science work that surfaces how AutoML is situated within organizational processes.

Participant	Role	Organization Size	Sector	AutoML Use
P01	Management	50,000+	Finance	Experimenting
P02	Management	10,000 - 50,000	Management & Consulting	Experimenting
P03	Analyst	50,000+	Finances	Active
P04	Analyst	50,000+	Finance & Consulting	Active
P05	Management	50,000+	Retail	Knowledgeable
P06	Analyst	50,000+	Healthcare	Knowledgeable
P07	Analyst	1,000 - 10,000	Healthcare	Experimenting
P08	Analyst	<10	Analytics	Knowledgeable
P09	Analyst	100 - 1,000	Security	Knowledgeable
P10	Analyst	100 - 1,000	Software	Knowledgeable
P11	Analyst	100 - 1,000	Software	Knowledgeable
P12	Analyst	100 - 1,000	Education	Experimenting
P13	Management	1,000 - 10,000	Finance & Government	Experimenting
P14	Analyst	<10	Consulting	Knowledgeable
P15	Management	1,000 - 10,000	Healthcare	Experimenting
P16	Analyst	100 - 1,000	Healthcare	Knowledgeable
P17	Analyst	1,000 - 10,000	Finance	Active
P18	Management	50,000+	Telecommunications	Experimenting
P19	Analyst	100 - 1,000	Analytics	Active
P20	Analyst	100 - 1,000	Travel	Knowledgeable
P21	Management	100 – 1,000	Healthcare	Active
P22	Management	100 – 1,000	Sports Analytics	Active
P23	Management	10 - 100	Telecommunications	Knowledgeable
P24	Analyst	<10	Management & Consulting	Active
P25	Analyst	10,000 – 50,000	Manufacturing	Knowledgeable
P26	Management	50,000+	Healthcare	Experimenting
P27	Analyst	50,000+	Healthcare	Experimenting
P28	Management	50,000+	Finance	Experimenting
P29	Management	10– 1,000	Telecommunications	Active

Table 1: Summary of the Participants.

3 METHODOLOGY

We conducted semi-structured interviews to develop an understanding of how AutoML is used to automate data science work. We were also interested in surfacing the role of the human-in-the-loop as it is mediated by data visualization tools, such as those used to explore data or support model tuning and selection.

3.1 Interviews and Data Collection

We recruited participants through a snowball sampling approach [9], with the first point of contact being individuals that had participated in prior studies, were known to the authors, or other collaborators. We recruited and conducted interviews with 29 individuals that self-identified as data scientists, or analysts engaged in data science type work, or a manager overseeing a team comprising either entirely data scientists or a mixture of data scientists with others. The semi-structured interview format prompted participants to discuss data science work in their organization; if and how they currently use AutoML systems, or plan to deploy AutoML systems; and the ways they use data visualization, using the Tableau platform or other tools.

Interviews were scheduled for approximately 60 minutes, audio-recorded, and transcribed. Our participant screening questionnaires and interview guides are provided as supplemental materials. Due to the nature of semi-structured interviews, the range of topics that participants chose to touch upon were quite broad. Moreover, due to the novelty and diversity of uses for AutoML technology the perceptions and pain points described by our participants were not always overlapping. All interviews were conducted over video conferencing software. A summary of participants, their organization size, and domain are summarized in Table 1.

3.1.1 Sensitization to Emerging Concepts.

Sensitizing concepts are an important component of qualitative research, as they ground the analysis in important emergent features and operate as a key interpretative device in data analysis [4]. At the outset of our study, we had some preliminary concepts that we were sensitized to from prior research we conducted that examined the nature of data science work and workers [10]; we used this prior research as part of selective coding processes. In addition to this prior framework, we also had our own notions of concepts that could be pertinent to AutoML, visualization, and human-in-the-loop interactions specifically and these informed our initial interview questions.

As we completed interviews we debriefed and conducted initial thematic coding of transcripts, we became sensitized to particular themes in our analysis that further refined our existing concepts of data science work and generated new ideas that we had not previously considered. Specifically, these emerging themes included the importance of different types and levels of participant expertise and participants' use and attitudes around "click" (low or no-code solutions) as opposed to "code-based" solutions. Analysis of participant pain points surfaced issues of tool switching, trust, and collaboration. Finally, we also found that predictive modeling was the primary way that these organizations applied AutoML technology. As we became sensitized to these concepts, we revised our interview guide to ask more pointed questions about these themes. We provide both the preliminary and modified interview guides in our supplemental materials.

3.1.2 Participant Characteristics and AutoML Use.

Participants self-identified as either analysts or managers. Analysts were individuals that were engaged in the day-to-day tasks of data analysis, including data scientists, business analysts, or other technical analysts engaged in data science work. Managers oversaw teams that often contained a mixture of data scientists, business analysts, or other types of organizational decision makers. In total, 17 participants in our study were analysts and 12 were managers. Overall, participants had high data science expertise, although one could be classified more as a citizen data scientist, which did not have formal training in data science but was exploring this field with the aide of AutoML. Participants also represented organizations of different sizes performing a variety of functions. Four participants were at organizations with fewer than 100 individuals, 10 with between 100 to 1,000 individuals, 4 with between 1,000 and 10,000, 2 with between 10,000 and 50,000, and 9 with more than 50,000 individuals. Participants worked in a broad range of organizations across different industries that were focused on data analytics, finance, government, healthcare, management and consulting, security, telecommunications, and travel.

We further stratified participants according to their current usage of AutoML technology. *Active* users were those who reported that they, or members of their team, used a specific AutoML technology to conduct their work. We did not stipulate some required frequency of use (daily vs not) or the number of individuals currently using this technology. *Experimenting* users were those that reported creating proof of concepts or described at least some preliminary projects specifically for the purposes of exploring AutoML technologies. Unlike active users, those that were experimenting with the technology articulated that their use of AutoML was in the early stages and exploratory in nature. Finally, those individuals that we categorized as *knowledgeable* had high context for data science work including AutoML, but were not using or planning to use this technology in their work. Among our 29 participants 8 were active users, 10 were experimenting, and 11 were knowledgeable.

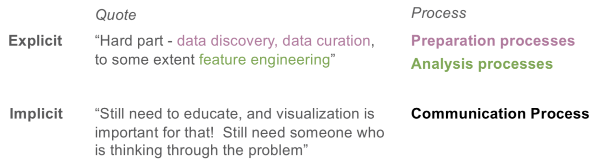
3.2 Selective Coding Process

A prior study [10] used an open coding process to define a framework of data science work that comprises four **higher-order processes** and fourteen *lower-order processes*. We use the set of codes from this prior study to carry out a selective coding of our interview transcripts. Selective coding is a stage in grounded theory research that serves to organize the analysis around a core set of variables [5], in this case the processes of data science work, rather than derive and organize a new set of codes as is done in open and axial coding. The reason we use selective, as opposed to more commonly used open and axial coding approaches [37] is due to AutoML technology being relatively new and as a result participants having varied experiences with it. While our interviews captured a rich diversity of experiences with AutoML this diversity also led to spareness in our data that made it difficult to achieve theoretical saturation in an open coding process. Using a selective coding process allowed us to scaffold our analysis around a cohesive narrative of how AutoML is used *across* data science work. The selective coding process still makes use of constant comparison that allowed us to eventually achieve theoretical saturation in our findings.

The set of four **higher-order processes** and fourteen *lower-order processes* in this framework for data science work were:

- **Preparation:** *Defining Needs, Data Gathering, Data Creation, Profiling*, and *Data Wrangling*
- **Analysis:** *Experimentation, Exploration, Modeling, Verification*, and *Interpretation*
- **Deployment:** *Monitoring* and *Refinement*
- **Communication:** *Dissemination* and *Documentation*

The authors of [10] also indicated two *lower-order* processes, *Collaboration* and *Pedagogy*, that were identified as emergent but did not have sufficient context to place within the higher order categories.



	Quote	Process
Explicit	"Hard part - data discovery, data curation, to some extent feature engineering"	Preparation processes Analysis processes
Implicit	"Still need to educate, and visualization is important for that! Still need someone who is thinking through the problem"	Communication Process

Figure 3: Example of Annotating Interviews with processes from an existing framework of data science work and workers [10]

In Figure 3, we exemplify how we performed a selective coding for these processes across our interviews. Some statements made explicit references to data science processes, for example, "Hard part - data discovery, data curation" are explicit references to the preparation higher-order process, as the terminology used can be linked directly to a higher order or lower order processes in the existing framework. By comparison, some references to processes were more implicit and were inferred by the authors with other context from the interviews. For example, "still need to educate, and visualization is important for that. Still need someone who is thinking through the problem" was determined to be an implicit reference to communication processes although there were not explicit terms specific to communication.

As with putting any model into practice, the prior framework not only deepened our analysis, but also generated natural tensions between our observations and our framing of data science. We leveraged these tensions as points of inquiry that allowed us to critique or expand upon these frameworks based upon the participants' reported experiences. We reflect on our approach and propose modifications to it that we describe in Section 5.4 (as these modifications were motivated by our analysis) and again in discussion (Section 6).

4 RESULTS

We first present our general findings regarding prevailing attitudes toward AutoML and the role of automation of data science work. Then, we examine the intersection of AutoML in data science work at a level of higher-order processes.

4.1 Attitudes Toward Automation

In this section, we describe prevailing attitudes toward adopting automation in data science as described by our participants. We identified four primary themes that encapsulate these attitudes: the role of AutoML to drive productivity, the importance of tool integration, the concerns about automating bad decisions, and, finally, the desire to limit the role of the human-in-the-loop.

4.1.1 Improving Productivity. AutoML was embraced with cautious optimism by participants at organizations of different sizes, but we found that it tended to be more widely used or explored at larger organizations. It is not clear what is driving these differences, but we believe that it may relate to different amounts and rates of data collection at larger organizations that motivate a greater need for automation. Among larger organizations that had implemented data science automation tools, the primary reason for investing in AutoML was that it "just makes data scientists more productive" (P01) by automating aspects of their work and allowing them to triage to focus on more pressing problems. As a specific example, P07 indicated that there is

> Lots of waste determining which models to work on. [I] Wonder [if we] should focus on managing the pipeline so [that] the things get through have more impact. [We] need a predictive model to figure out which predictive models are the ones to work on.

The automation of routine work, like model tuning and selection, was seen as a desirable way to shift the effort of human labor toward model verification and, if needed, correction tasks. While some participants felt strongly that technical expertise was required to safely use AutoML technology for productivity gains (a topic we return to in 4.1.3), others (P03, P12) saw the benefit of AutoML to democratize data work. Individuals without a background in statistics and computer science that occupy roles of "business analysts" or "moonlighters" [10, 33] would benefit from the lower barrier to entry that AutoML affords. This democratization effort may improve productivity in those roles, but it also opens a door to capabilities across self-service analytics that were previously inaccessible.

Overall, AutoML systems reduce the amount of code that is needed to use data within data science workflows. Individuals with high technical expertise, such as data scientists can leverage AutoML systems to improve speed and efficiency of routine tasks. For non-experts, AutoML systems can also democratize the accessibility of data science workflows and machine learning solutions.

4.1.2 The Importance of Integrating Tools. Participants reported using a variety of commercial tools to facilitate AutoML work and the need to create custom solutions for preparation, deployment, and communication data science processes. Participants used or were actively investigating platforms like Alteryx (n=5) for automating their workflows and integrating with Data Robot (n=3) or H2o.ai (n=3) to facilitate automatic modeling steps. Dataiku was also used in lieu of Alteryx and was seen as a better tool for facilitating collaboration across processes. Participants also reported using Sagemaker (n=2), PowerBI (n=2) (Azure), and Data Bricks (1). Participants report leveraging libraries like TensorFlow (n=4), scikit-learn(n=2), and mxnet (n=1) for their AutoML work. Python, R, and their attendant notebook environments were described as being used by roles that had higher technical expertise. However, P02 observed that "businesses can't deal with the notebooks" because "data scientist(s) are now [needing] to build things that can run in a production environment" in order to operationalize models. Moreover, as organizations seek to spread out the data work from data scientists to others in the organization, P15 observed that they were "starting to see heavier reliance on data science products that don't require heavy coding.". For individuals without a data science or computer science background the need to write code appear to be a barrier, but our participants comments indicate that AutoML can lower, or potentially remove, this barrier. Moreover, there is an increasing appetite for a "platform [that] helps people move between tools that they have selected"(P04) and where "75% of the organization could work". Tool switching is common in data science because there are "Different tools for different analysis" and yet we found most users would "prefer to stay in one environment"(P06).

Importantly, data science processes are not linear but occur in a "big recursive" (P09) loop . Constantly changing environments within multiple cycles of iteration and refinement is time consuming and impractical. Participants did report visualizing their data via systems like Tableau, or via charting libraries in R or Python, but many described their limited use:

> As organizations scale, they're going to spend less and less time doing visualizations [...] the job is to deliver results of a model in some form[...] Data scientists aren't going to deploy with ggplot, but they may use it for static reporting or just for their information. (P17)

This participant didn't see visualization tools as scaling well alongside AutoML and other data science work, leading to abandonment, especially within results communication. Two other participants had to awkwardly move back and forth across modeling and visualization tools, and as a result the role of visualization was limited throughout the process.

4.1.3 *Concerns of Automating Bad Decisions.*

Participants also clearly understood that blind trust in AutoML could lead to potentially catastrophic failures. P20 worries that "lots of people will try to predict things without really understanding. [...] People will make horrific mistakes and not realize they've made them." P12 colorfully indicated that "having this [AutoML] tooling may just allow people to make stupid mistakes easier!" P05 observed that it's "bad to slap together models and try to make decisions from it without understanding how things work. [That's like] giving a loaded gun to a child". Participants were also concerned about regulatory constraints, for example the European GDPR legislation, that requires organizations to be able to explain decisions made by automated data science technologies. Even without legislative pressures, there were internal organizational concerns around these technologies, especially when large financial decisions were involved. A tolerance for errors and failure was an important factor in evaluating the use of AutoML technology; however, in many cases perfection was not required. For instance, P12 observed that there were "lots of business use cases where 80% accurate could be okay". P29's observation echoes P12's that it is desirable for automation to help your surface failure points in your data preparation and analysis processes:

> I think they [executives] want you to use those [automated insights] to look at a graph and say, "Oh wow, this is life changing. Let's go make this change in our business." We didn't use it like that. We used it to make sure that the results we were getting back made common sense.

A surprising finding was the general concern around the use of AutoML by "citizen data scientists" or domain experts that were not formally trained in data science, statistics or computer science. P12 stated that while they understand organizations want to democratize data science work, it still worries them because "in practice you'll still have to be pretty technical" to analyze data. P22 raised the issue of the overhead needed to ensure those "who aren't as well versed...in the data science space are able to not make silly mistakes that they shouldn't be making". Perhaps the strongest stance we heard was from a participant who stated that they would "restrict it [AutoML] just to the data scientists" and "use it to get efficiency after demonstrating they know they are doing" (P05). This attitude was a recurrent theme in our study.

Overall, from participants' responses we see that the promise of automating data science is tempered by very real concerns of how things could go wrong. However, this does not mean that organizations are pulling back from their investment in these technologies. As P02 noted: "ambition is still 'industry 4.0' with lots of automation."

4.1.4 *Limiting the Human-in-the-Loop.*

Concerns toward safety and trust of AutoML in many ways highlight the value of humans-in-the-loop approaches to balance automation with human oversight. However, participants also expressed concerns about the inclusion of humans within the AutoML loop. For example, while P02 expressed that there was still "lots of room for humans-in-the-loop" innovations in their industry, they also stated that "the manual part, where you have to visualize something, is getting cut out as much as possible". P01 stated that even as "there are lots of automated tools in place making decisions", at the same time there was a lot of "anxiety in the firm about what people do with ML," stemming from concerns of automating decisions at scale. We interpret these seemingly contradictory positions to mean that human-in-the-loop approaches are valuable when applied at the right time and in the right way.

4.1.5 *Summary.*

We briefly summarize the key takeaways for participant attitudes toward AutoML. While participants expressed concerns about the potential to automating bad decision-making, there was also a growing interest in using AutoML technology to produce a 'good enough' result that could scope out the viability and possible issues with the data or machine learning product. AutoML allows for the creation of sophisticated tools with minimal code and offers opportunities to 'fail fast', which enables data scientists, and even so-called 'citizen data scientists', to surface issues earlier. However, what is most clear is that applying AutoML technology at suitable points in data science work is very important, otherwise it is dismissed as intrusive. To further explore *when* and *where* AutoML technologies could support data science, we analyzed interviews through the lens of an existing model for data science work.

4.2 AutoML in Data Science Work

In this section we summarize our findings on the use of AutoML in the data science process. We consider places where considerable human labor is required either to support the creation of a machine learning model or to interpret, communicate, and act on its findings. Following the framework described in Section 3.2, we begin by examining AutoML in data preparation, followed by analysis, deployment, and communication.

4.2.1 *Data Preparation Continues to be a Rate Limiting Step for Automation.*

Participants understood that without robust data preparation the AutoML portion of their Data Science processes would be ineffective. P02 succinctly stated that,

> AutoML has never been the solution - it's a shiny toy. [It's] always going to be about the quality of the data - do you understand what you are modeling?

In our interviews, participants identified several challenges in data preparation work that still required a lot of human labor, from gathering data, profiling it, and wrangling it into shape for analysis. It is important to emphasize that participants rarely began by cleaning a single tabular dataset, but often needed to bring several data sources together. P12 reported that "40-50% of my team's time [was spent] on Alteryx to bring data together", while P23 reflected that "the most difficult part of my day is getting the data that I need to work with". Once the data were gathered, participants faced difficulties with data profiling, a challenge that was also surfaced in prior work by Kandel [31] and Alspaugh [1]. P19 expressed that automation in

> Data profiling would be a huge win. I spend a ton of time having to explain the shape of data, and what shapes work best, how to explore the data, and how to refactor as needed

Even when participants have data gathered and profiled, they still need to assess its utility for further downstream analysis. Rapid iteration via AutoML plays a role in speeding up this manual process. P23 described that they "get [the data] to a point where maybe it's 60% [clean], and they start to run algorithms...[in order to]...expand on the data itself". The workflow described here resonates with rapid prototyping and failing fast to discover issues or limitations in the data. This observation also echoes data reconnaissance and task wrangling processes previous described in [11], in which individuals acquire and quickly view data in order to assess its suitability for analysis and decide whether to pursue additional data sources. Despite the usefulness of AutoML-driven prototyping, challenges associated with data preparation remain. This reflects an emerging theme from our analysis that there is a growing need to more tightly couple investments in data prep and model building. For instance, the experience of these challenges led P03 and their team to make more significant investments toward "tools for democratization of data prep and (to a lesser extent) model building".

It is not surprising that data preparation is both time consuming and important to the successful application of AutoML and data science more generally. Prior studies [33, 46] with data scientists and other domain experts have routinely pointed to this bottleneck for years, and visualization tools such as Trifacta (and its academic predecessors Wrangler [32] and Profiler [30]), and Tableau Prep have been developed to address this challenge. It is disconcerting that preparation continues to be such a significant bottleneck despite existing tools. Our observations suggest that one reason for this may be that existing tools for data preparation do not easily fit within a data workers' analytics environment. By extension of these observations, AutoML technology needs to be well-integrated with existing tooling environments, while also surfacing the manual labor and lack of adequate tooling for data preparation.

4.2.2 Use of AutoML in Analysis Varies by Data Science Role.

Perhaps as a testament to the advancement of AutoML technology, participants reported that model building "is fast and easy"(P12). P25 further elaborated that

> If I can actually get through all the data stuff, then getting the predictive model is not really that hard. There's a huge bunch of code to get the data ready for the model, then a tiny bit of code for the model. And then the rest of the work is delivery to the customer.

The desired amount of oversight and control over AutoML in the analysis appeared to vary by level of technical expertise. P26 felt that individuals with high expertise in statistics and/or computer science were less likely to use automation because "they want everything to be customizable", whereas those with less technical expertise, whom they refer to as *"citizen data scientits"* were "focused on integrating intelligence to their app" and tended to prefer higher levels of automation. This latter group is growing as P03 observed : "the vast majority of data science type work is done by non-data-scientist[s]". From participants' responses we also noted that this group with a low or evolving technical expertise often needed heavy guidance, low or no-code AutoML implementations, and visualization. P29 also observed how individuals with higher technical expertise could act as a rate limiting step to analysts engaged in data work, but that automation could serve as a catalyst:

> [Analysts] maybe can do their own little, their own little predictions, off to the side. And they can do them fast [...] PhDs could always do it better, but are they there? Do they have time? The answer is almost always no. You can either do good or you can do nothing. Better is there, but you're not getting better. They're [PhDs] busy, working on the big problems.

Moreover, summarize that individual with without high technical expertise benefited from proper guidance, no-code solutions, and data visualization to help situate themselves within the analysis process given many "steps in (data science) workflow contain lots of details that are hidden" (P03).

It was surprising that visualization was not more widely mentioned to steer the model authoring processes even though there exists a number of visualizations systems to help individuals do so [7, 8, 53]. One possibility is that individuals with high technical expertise are constructing novel and bespoke models that existing data visualization tools cannot easily support. Instead, these technical experts might benefit from highly customized visualizations for certain classes of models, such as those generated by tensorboard [49]. In contrast, individuals with lower technical expertise lack the background knowledge to orient themselves and effectively interact with visualizations exposing the mathematical underpinnings of machine learning models. They rely on the AutoML systems to make model decisions and it may not be easy for them to correct and refine these models.

While participants did raise concerns that AutoML-derived models and results were a 'black box', it also appears that technical acumen in statistics and computer science was perceived as necessary and possibly sufficient to 'open up the black box'. Echoing concerns we summarized in Section 4.1.3 participants saw AutoML as potentially contributing an existing "knowledge gap [that is] becoming wider" (P19) because the availability of AutoML meant that individuals with lower technical expertise could harness the power of machine learning without having to seriously engage with its technical and nuanced underpinnings. Our interpretation of these concerns was that trust in the individual conducting the data analysis was as important as trust in the AutoML process itself. Moreover, collaboratively sharing knowledge or including human oversight may mitigate some of these concerns and could be achieved through visualization of data science processes. The differences in data science roles and the extension of data science work to individuals without formal training in statistics and computer science has several implications for the design of AutoML and visualization tools. While AutoML tools reduce or even eliminate the need to write code it becomes important to consider what kinds of guard rails might need to be put in place. We believe that data visualization tools are an important component of such guardrails, but that we require a finer-grained understanding of data workers to design such tools effectively.

4.2.3 *'Good Enough' Rapid Prototyping to Bootstrap Analysis.*

Participants used AutoML technologies to rapidly prototype viable solutions in both data preparation and analysis. The need for rapid prototyping stems from the challenges of generalizing AutoML to a variety of problems, which requires manual effort. P21 acknowledged that "AutoML is really hard, and I think we have so many operations with such nuance that we actually most of the time really... just want to be doing simple stuff correctly, rather than adding additional layers of complication." P24 was much more explicit in stating that "every customer is different [...] [but] AutoML is supposed to be a generalized framework. So, that is a problem". The challenges of AutoML to generalize to a variety of problems are known, especially when it concerns fixed structure computational pipelines (which is the most common implementation of AutoML) [54]. Despite this lack of generalizability, participants found that they could leverage AutoML to rapidly prototype data science solutions. P23 offered description of such a use case:

> I talk to clients daily. If I could get ML done just real quick, as a prototype into what we could build (in depth), that would be super helpful. [...] Can you get me 50% of the way to answering some question quickly, that would benefit me

P21 reported using automatically generated results to start a conversation with others ahead of making serious personnel or infrastructure investments. They shared that "when starting out with a client, we'll run the default model and we'll say, 'Hey, here are some of the topics, some of the interesting trends that are coming out' " and then use the clients reaction to further refine upon default model or craft a different solution altogether. As we previously reported, P12 and P29 also make the case for 'failing-fast' to discover issues in the data or analysis without expensive upfront investment in fully developing an automated data science pipeline. This rapid and iterative use of AutoML to drive different data conversations evokes a complex picture of a data worker operating within multiple loops of data science and organizational processes. This prototyping scenario offers an evocative example of how the limitations of AutoML technology can be beneficial leveraged through human-in-the-loop interactions. We argue that with adequate guardrails in place AutoML systems may also be able to further support the process of surfacing potentially more complex issues of bias.

4.2.4 *Inadequate Support for Governed and Managed Deployment Processes.*

The majority of machine learning models often do not advance to production environments where they are applied to real data. Those that do are often required to go through a set of governed processes before they are deployed and are constantly monitored once they are out in the wild. These governed processes vary as P01 described :

> [A] Governed workflow [is important]. Looking at what all teams are doing - are there divergences around governance, etc. e.g. models with a financial impact have a very stringent governance process.

The volume and variety of both data and models makes it challenging to monitor and govern AutoML models deployed in production. P03 reported that their practice was to "err on the side of letting people use tools [they preferred]" and to "monitor what tools are being used". They emphasized that vigilance was important to ensure mistakes do not "clog up the server with bad content", which might happen when pushing a model generated by AutoML into production without adequately vetting it. These problems exist for all software code in general, but may be further exacerbated by the novelty and complexity of AutoML. Moreover, the amount of data produced by automation can make it overwhelming to effectively govern models that need to be continually validated, and have a process that employs someone to "look for drift, look to increase accuracy and effectiveness of these models over time"(P07). Larger organizations working under enforceable regulatory constraints struggle to find the right balance between integrating a potentially valuable new technology like AutoML while conforming to these constraints:

> [There are] areas where critical models are developed that will likely have very strict controls, often imposed by a regulator [...]if you have no clue what that [model result] means, you are on pretty thin ice (P03)

As with preparation, the processes of governing deployed models still requires considerable human effort. Moreover, we believe that adding some sort of automation to these processes is desirable in order to reap the efficiency benefits of AutoML. Dashboards are often used to monitor changes in data [44], but participants did not report using dashboards for AutoML work even though they may already use dashboards for other types of work. We hypothesize that this relates to the tooling environment and that monitoring and the governance of AutoML systems require more specialized dashboards that are not well supported by existing tools. There may be fruitful work here for visualization and HCI research to improve governance processes through better monitoring and, at least, help them triage governance violations. Improved awareness and consideration for governance throughout the visualization design process can also inform the implementation of guardrails for AutoML throughout data science work.

4.2.5 *Correction and Repair of Deployed Models.*

Human oversight of automation is critical to detecting when something new or unexpected has happened, identifying the source of what has happened, and implementing appropriate corrective actions as needed. While there may be some ability to automatically detect anomalies, and thus make governed workflows easier to monitor, participants expressed doubt that such an approach would work in practice. P12 stated that if "the forecast is clearly wrong a human can detect this" whereas it is harder for the AutoML tool to do so. Still, other participants articulated the limited abilities of analysts to intervene appropriately, suggesting that some analysts "wouldn't necessarily know what to do next...[whereas]...a data scientist might know what to do next - for improving forecasts, for analyzing how good it might be." P26 succinctly summarized that as "you're only as good as what you debug". These observations align with a recurrent theme in our findings that trust between the underlying technology and the data science teams is critical for the wider adoption of AutoML. Moreover, these observations expose the brittleness of AutoML technology and its reliance on iterative loops of correction and refinement with humans. In the next section, we also emphasize that AutoML loops are not closed systems; rather these are loops that interact with many other loops of business processes and pre-existing modes of human collaboration. Thus, 'debugging' not only requires technical expertise that spans the preparation to deployment processes, but includes sufficient domain expertise to recognize and account for other 'loops' that interact with and beyond data science.

4.2.6 Communication and Collaboration Build Trust in AutoML.

Communication and collaboration are essential data science processes, including but not limited to model automation [33, 46, 52]. AutoML systems require communication between humans and the technical system. P22 expresses one such mode of communication in which the AutoML system helps guide users in analysis by communicating "this is what it is that you're about to do, and this is the impact it will have" and should also prompt users with respect to certain actions with "are you sure you want to do this?". P17 also noted that more could be done to "walk the user through that [a data analysis, for example] given a kind of data to predict, here are the kinds of models and visualizations to use". AutoML introduces a new mode of collaboration between humans as well. This new frontier can also be challenging to navigate as P29 observed:

> So, there's human interaction along the whole life cycle. And interpreting that human interaction is what we're trying to get machine learning to do.

However, participants indicated that these diverse individuals must still work together to deliver actionable and safe results from automating technology. A common theme emerged around the desire to broaden data engagement across the organization, bringing more people into the data sensemaking loops. In P21's words,

> [We] need our workforce to be more data savvy across the board. An engineer needs to be able to play with data as much as the MBA does [...] [and] giving them better tools will help with ramp up.

A big part of having teams work more effectively together is to provide more situational awareness of data science workflows and who has done which task. For instance, P06 described what would be needed to support teams working together across workflows. This support includes surfacing "notification[s] that people [are] working on the same step ...[and]... underlying metadata about how people were using the platform " (P06). Integral to this collaboration was the ability to hand-off different aspects of the data or analysis processes to different team members, likely with a different data science role:

> They can create a workflow, share it with other people, people can build off of that workflow, grab a table from that workflow and then build their own. That collaboration aspect of it was important to us. (P29)

In order to improve collaboration some participants defined a viable solution around making workflows visual, interrogable, and extensible.

Participants also highlighted the importance communicating to individuals that were one step removed from the data analysis processes, most often communication to executives or other business leaders

In this process, visualization plays a clear role as a communication tool. A theme emerging from the analysis was that participants often framed this part of the process as more difficult than the modeling itself. This was due in part to the extra work required. For example, P19 describes the additional labor it takes, especially when the visualization tools are not well integrated into existing processes :

> There is a gap once your analysis is done on presenting the results. Nobody wants to spend more hours in another tool to build charts for explanation.

For instance, P05 described the challenges of authoring compelling visualizations and how "nice visualizations feel like a hack that the average user can't build themselves".

The other challenge often cited was around the efficacy of visualizations as a medium of communication to drive business decisions or processes. Participants described the challenge of translating their work to business users that were not versed in modeling vernacular. In one participant's words, "we don't want people to actually understand model jargon, we want to help them understand what the model is saying in business terms." This participant relies on interactive visualizations to support the dialogues that they anticipate will happen when showing a snapshot of results to bridge the gap. Still, despite efforts to translate for business users this participant's team experienced a range of challenges in operationalizing their models. P26 describes how "the challenges that we saw as the data science team is...we give this [model or results] to them, but then actually, the action of implementation of this in the market sometimes doesn't always pull through. So it's like we did all this work, you said it was good, but now you have to take it to the last mile, actually get to marketing, creative, and content, and get it out to market." They point to communication difficulties between data scientists and others at the organizations as exacerbating this 'last mile' problem, which results from "either lack of funding or sense of disbelief in prediction models and ML techniques" (P26). Validation measures may also be required for regulated industries that can slow this process down.Taken together, the collaborative nature of data science work imposes constraints on the design of visualization tools, which must be usable across the organization and interoperable by individuals embedded within a variety of analytic, business, and governance processes.

4.2.7 Summary.

Our examination of AutoML technology along the data science pipeline, both where it exists and where it does not, helps us to understand the current capabilities of this technology and how the technology and its surrounding ecosystem can be further developed to support data scientists and others. We see gaps for AutoML technology outside of data analysis processes and that translate to unmet tooling needs in data preparation, governance, and deployment processes. These are also processes where considerable human labor is still required to make AutoML technology in data analysis viable. Automation that extends to these other processes, ideally with appropriate guardrails, could improve both the quality and speed of data science work. Moreover, the relationship between automation and data science expertise emerged as a critical consideration for what future tools should support, including the types of guardrails that should be built in. We were surprised to surface some of the tensions that existed between data science experts, and data workers with different training. Emerging from this tension was one heavy-handed guardrail strategy to restrict access to AutoML technology that many sought to implement. We believe this view has surfaced from a lack of adequate tools to support the safe creation, deployment, and governance of these models and that there are many fertile opportunities for visualization research in this space. However, our analysis of participants' comments reveals that existing visualization tools are falling short of their needs. Moreover, that data visualizations tools can have a steep learning curve and their is little motivation to use following intensive analysis. We underscore that it is critical to understand the diversity of teams that carry out data science work and the ways they intersect with many organizational processes. In other words, visualization tools need to work for many humans engaged in many loops.

5 INTERPRETATION OF FINDINGS

We now reflect on our findings and summarize the central themes that emerged from our analysis.

5.1 Three Usage Scenarios for AutoML Emerge

The general attitudes toward AutoML suggest three usage scenarios for this technology that are conditioned on the technical expertise (statistics and computer science) of the individual analyzing the data and the magnitude of consequences associated with errors. The first usage scenario is automating routine tasks, thereby reducing the coding efforts of data science teams and improving the speed of the analysis processes. A second usage scenario is the rapid exploration of potential data science solutions through low-effort prototyping. Such prototyping approaches can be used by individuals with varying degrees of technical expertise. Its possible that for individuals with high technical expertise (such as, data scientists, generalists, research scientists, ML/AI engineers, and data shapers) prototyping allows them to quickly create a base framework that they further develop into novel solutions for arising technical challenges. For other individuals, prototyping enables them to have a conversation around the data with customers and other members of their organization. Prototyping also enables individuals and data science teams to fail fast and discover issues with their data and analysis before investing in considerable engineering effort. A third and final usage scenario is the use of AutoML toward democratizing the ability to create a machine learning model, empowering individuals that would not be able to build a model otherwise. In this third scenario, we argue that individuals require heavy guidance and guardrails from an AutoML systems and may have very limited ability to identify errors or correct them.

The delineation of these usage scenarios is intended to guide visualization researchers as they explore opportunities to develop techniques or systems for AutoML.

5.2 Varying the Level of Automation Across Data Science Processes

Considerable human labor is still expended to prepare data, govern and deploy a model, and to communicate the results to impacted individuals and other decision-makers. An end-to-end AutoML solution capable of addressing the full scope of such data science work does not currently exist, and as a result data workers, which includes individuals that are and are not data scientists, are finding *ad hoc* ways of bootstrapping AutoML technology into their work. In Figure 1, we outline a common set of eight steps synthesized from participants' responses describing AutoML use in enterprise settings. We further align these steps within higher order data science processes. For data preparation and analysis, these tasks were prototyping, exploring the results, and settling on a solution to implement. Should this solution reach a certain level of maturity it is deployed into production following a verification of the solution (including compliance of regulations), where it is consistently monitored while in production. Finally, these deployed models can be used to take action through an inspection of the results that surface new insights for decision making. We illustrate the levels of automation [41] that we believe are desirable for future AutoML systems to support, considering the range of participant challenges and concerns this study surfaced. Importantly, the level of automation is not consistent across all data science processes. Human oversight is still required throughout data science work and is dictated by both regulatory requirements and organizational practices. Most automation likely needs to adopt a 'cruise control' mode of interaction [34], where humans can oversee and steer AutoML systems without needing to guide the systems at each step. Even this would be an improvement over current AutoML systems that appear to oscillate between 'autopilot' and 'user-driven' modes. We further illustrate the level of automation required by individuals with high expertise in computer science and/or statistics (Data Scientists, ML/AI or Data Engineers), and low or an evolving technical expertise in these areas (Business Analysts, Moonlighters). Individuals with high expertise can benefit from full automation, for example when speeding up routine work (Usage Scenario One) or to rapidly prototype and explore new solutions (Usage Scenario Two). Even in these two usage scenarios individuals with high technical expertise still rely on considerable manual effort, but this in fact might be an appropriate use of their expertise and focus on "bigger problems", especially if other trivially automated tasks are reliably handled by an AutoML system. Individuals with lower or and evolving technical expertise require much more support and guidance and would rely much more on full automation to rapidly prototype solutions (Usage Scenario Two) or even to begin to engage in data science work more generally (Usage Scenario Three). However, while these individuals rely on AutoML systems to guide them, their domain expertise still needs to be incorporated in downstream steps.

While Figure 1 is a useful illustrative summary of our findings, it needs to be further validated in future studies that assess its generalizability. We suggest how to do so in our Discussion section.

5.3 Eliciting Tasks for Visualization Design

Taken together, these usage scenarios an levels of automation impose a set of constraints for the design of visualization tools that operate together with AutoML technologies. Visualization researchers need to carefully consider where and how automation is currently deployed, the diversity and expertise of the data science teams, and the full breadthof data science processes. We have illustrated a set of steps and proposed the levels of automation in Figure 1 that data workers with different levels of expertise desire. Importantly, by illustrating an end-to-end pipeline, we encourage visualization researchers to consider how changes *across a workflow* influence the kinds of data to be visualized and the fundamental tasks that these workflow steps support. For example, 'prototyping' may have different tasks associated to it depending on whether the analysts want to develop a new model, fail fast, or prototype some solution for a customer. The 'monitor' process in deployment could reasonably rely on high automation until the system requires human action, much like auto pilot in aircraft. Alternatively, 'exploration' may require less automation if the user is expected to steer the algorithm. Without a concrete understanding of usage scenarios, data science steps, and level of automation, researchers risk eliciting inappropriate tasks and creating visualization tools that will be dismissed because they are not well integrated into end-to-end data science workflows. Visualization researchers can reference our findings and the summary in Figure 1 as a guide to support their own task elicitation for the design and evaluation of visualization tools.

5.4 Modifications to our Analytic Framework

Lastly, we briefly reflect on our findings and propose modifications to the framework of data science work and workers reported in [10]. We remind the reader that this framework is described in Section 3.2 and delineates a set of **higher** and associated *lower* order data science processes that we used as part of selective coding analysis. First, we propose that *Collaboration* be added as a lower-order processes of **Communication**. While collaboration was part of the original framework there was not enough evidence to determine how it should be incorporated. This analysis suggests it belongs as a component of **Communication** alongside *documentation* and *dissemination* lower order processes. Moreover, *collaboration* emphasizes the ways that individuals engage in multidirectional exchanges of knowledge and data products (data, code, models, documents), whereas *dissemination* refers to a more unidirectional exchange of knowledge from an individual to others. Second, we propose that *governance* be included as a lower-order process of *deployment.* While governance processes can technically encompass all of data science work, our findings point to its specific importance in managing the process of launching, monitoring, and refining machine learning models deployed into production settings. Finally, we propose a new higher order process, **Guidance**, which follows communication. We assign three lower order guidance processes based upon our analysis : *human-machine guidance*, *human-human guidance* (or pedagogy), and *organizational guidance. Human-machine guidance* describes the interplay between AutoML tools surfacing new data insights to humans and humans making corrections and refinements of AutoML models and results. *Human-human* guidance describes the collective work in building a data savvy organization and other efforts to bridge the data science "knowledge gap". Alternatively this could be referred to as pedagogical process. Finally, *organization guidance* refers to regulations and other organizational processes that impose constraints on the use of data, models, and the level of automation.

6 DISCUSSION

Visualization and HCI researchers have used enterprise studies to discover unmet needs of practitioners that have inspired new research trajectories that have ultimately led to new techniques and tools. As we consider the future of AutoML in enterprise, we believe a "cruise control" level of interaction [34] (Figure 1) is more likely to be adopted. However, we see significant barriers to implementing such a level of automation that stem from the diversity among data workers with different types expertise, a complex tooling environment that needs to be integrated, and brittle workflows that still rely on considerable human effort. Although visualization can play a role in supporting 'cruise-control' type automation, it was not being widely used to that effect and, in some cases, getting actively removed from automated data science workflows.We believe this lack of uptake is that visualization tools are potentially misspecified for the tasks they need to support and that this stems from poor understand of how automation is used in data science work and where there are opportunities for human-in-the-loop interaction. Our study fills this gap by surfacing usage scenarios and illustration of automation throughout data science work, which informs the goals and tasks feeding into visualization design and evaluation.

6.1 Implications for Automating Data Science Work

Throughout our analysis, we found both AutoML and human-in-the-loop to be misnomers for the processes that participants were describing. First, we noted in Section 2 that AutoML is used to refer to an ever-expanding set of data science processes from preparation to deployment and as such is being used interchangeably with 'automating data science' (among other phrases). We argue this is limiting as not all automation of data science needs to be in service of machine learning systems. Moreover, the notion of end-to-end AutoML obscures the human labor required for these systems to work, now and in the future, leaving inadequate support for human-machine collaboration. Echoing Wang's [46] language, we encourage researchers to **augmenting data science** with AutoML rather than automating it. It is more than a matter of semantics – the idea of augmenting data work explicitly makes space for human engagement and brings humans needs to the forefront of consideration.

Second, when we make explicit space for human engagement we are encouraged to consider the diversity among data workers. As we summarize in our three usage scenarios, this type of engagement will vary depend on the goals of data workers and their level of technical expertise. Along with prior studies [42, 52] we found collaboration among data workers to be of critical importance to the success of data work. Commensurate with findings from Hong [22] we also show that **trust amongst individuals engaged in data work was as important, or more so, than trust**

in AutoML. Surprisingly, AutoML technology appeared to erode trust among collaborators of different technical expertise by enabling so called "citizen data scientists" to potentially automate bad decision making. In theory, a human-in-the-loop paradigm for augmenting data science work can also be useful to understand the types of engagement between humans and machines that could ameliorate some of these trust concerns. However, here, too we find that human-in-the-loop is a limiting term. An AutoML correction and refinement loop not only exists within a wider scope of data science processes but also within organizational processes. While the nomenclature of human-in-the-loop is not exclusive to a single individual interacting with AutoML, we argue that the notion of "humans-in-the-loops" more accurately captures how this technology is used within enterprise settings. We note that a limitation of our findings was that study participants were primarily, although with some exceptions, experts in data science. While several were managers who oversaw mixed teams, we none-the-less believe it is useful to follow-up our findings by soliciting the views of those individuals that are not data scientists, but work closely with them.

As Visualization and HCI researchers continue to explore applications of technology like AutoML in data science work, we encourage them to consider the diversity of humans involved in data science work, their different needs and varying degrees to which they benefit from AutoML technology as well as the myriad organizational loops that are entangled within AutoML and data science.

6.2 Implications for Data Visualization Systems

Overall, we see that there are opportunities for visualization tools in data science work, especially in areas where there already exist considerable human labor. We especially see that participants struggle to get an overview of data work and that this complicates their ability to effectively handoff data, models, and results within their organizations. A visual overview of data science workflows emerged as an organic solution and is a promising area of future research. But beyond this specific example, we hope that the usage scenarios we present will help researchers identify new unmet visualization needs toward the use of AutoML that we did not surface here. However, the most troubling findings from our study concern the ecological validity of data visualization systems. We hypothesize that one reason visualization tools were not more widely used by participants was because they did not integrate well into existing data science tooling environments. This may be because existing visualization tools are developed as stand-alone systems where it is difficult to import data and export results, or because existing systems do not scale well to the volumes and varieties of data that organizations collect, or even because these visualization systems are themselves too brittle to flexibly adapt to variable data science or AutoML workflows. Moreover, visualization tools may not cater well to individuals across the gradient of technical expertise, and thus may be too rudimentary for those with high technical expertise and too complex for those with lower expertise. We encourage researchers to use our findings as a guide for surfacing these threats of ecological validity early.

Another fruitful area for visualization researchers is the creation of guardrails that surface and alert individuals of potential issues with their data, models, or results. The development of guardrails can help to examine concerns toward automating bad decisions. Our research indicates that their design is contingent upon individual expertise, the context in which individuals are using AutoML, and the level of automation that is expected. Some areas, like data preparation, will require more human labor alongside tools that automate their processes. Others, like monitoring a deployment model would rely on human labor primarily to respond to events, like the detection of model drift. Guardrails in both scenarios can help analysts contextualize and triage problems as they arise, but the design of these guardrails will differ between these two scenarios. Well designed guardrails may also increase trust and collaboration not only between data workers and automated processes, but also among data science teams. While prior research has suggested design considerations [2] and potential analytic pitfalls across visual analytics processes [36], research is needed to bring these together to explore dynamic and adaptive visualization guardrails that are appropriate for an individual's current analytic context.

6.3 Limitations and Future Work

The lack of existing studies on AutoML use in enterprise settings was the motivating factor for carrying out this research. Our findings support prior research and shed new light on the challenges and uses of AutoML in enterprise settings. However, we also found that participants had quite different experiences in their use and expectations of AutoML. As a result, our findings were simultaneously rich in capturing the diversity of experiences and sparse in that some of our findings relied on a handful of observations. To produce a cohesive analysis of these experiences we used an existing framework for data science work and workers as a scaffold. This sparseness of data and reliance on a scaffold is the primary limitation of our findings. Further work is needed to validate the generalizability of our findings, but this may be difficult due to the novelty of AutoML technology itself. One fruitful area of future work is to take the key insights from our research as constructs around which to develop a survey instrument that probes into AutoML uses more specifically than our current interview study. We did not take this approach here because we felt we needed additional information on AutoML use in the enterprise settings and beyond. A future survey instrument could also be used within a large mixed-methods approach, such as sequential explanatory design, which uses the survey results, in lieu of the framework we use here, as a more data-driven approach to inform a subsequent qualitative analysis.

7 CONCLUSION

Automating data science work through AutoML technology will continue to be commonplace in enterprise settings, especially at large organizations that work with large volumes of data. We identified three usage scenarios for AutoML that we argue are routine in current enterprise environments. These are automation routine work, rapid prototyping for a potential solution, and democratizing access to machine learning technology and data science work more generally. Moreover, we surface the complex handoff of data work between AutoML systems and data workers, as well as between data workers having different levels of technical expertise. Indeed, AutoML systems still rely on considerable human effort to be effective and even as this technology improves, human oversight will still be required to be sure it is safe and effective. While data visualization can play an important role together with AutoML, we find that it is used infrequently and is actively being minimized in data science work. We see our findings as having important implications for recasting the role of visualization in conjunction with AutoML and data science more generally.

ACKNOWLEDGMENTS

The authors wish to acknowledge and thank to study participants for sharing their insights with us. We also wish to acknowledge members of the Tableau Research, User Research, and Tableau CRM for their feedback on our study and findings.

REFERENCES

[1] Sarah Alspaugh, Nava Zokaei, Andrew Liu, Cindy Jin, and Marti A. Hearst. 2019. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 22–31. https://doi.org/10.1109/TVCG.2018.2865040

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proc CHI'19*. 1–13. https://doi.org/10.1145/3290605.3300233

[3] David M. Blei and Padhraic Smyth. 2017. Science and Data Science. *Proceedings of the National Academy of Sciences* 114, 33 (2017), 8689–8692. https://doi.org/10.1073/pnas.1702076114

[4] Glenn A. Bowen. 2006. Grounded Theory and Sensitizing Concepts. *International Journal of Qualitative Methods* 5, 3 (2006), 12–23. https://doi.org/10.1177/160940690600500304

[5] Anthony Bryant and Kathy Charmaz. 2007. *The SAGE Handbook of Grounded Theory*. Sage Publications, Los Angeles, Calif.

[6] Longbing Cao. 2017. Data Science: A Comprehensive Overview. *Comput. Surveys* 50, 3 (2017), 1–42. https://doi.org/10.1145/3076253

[7] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, and Andreas Kerren. 2020. A Survey of Surveys on the Use of Visualization for Interpreting Machine Learning Models. *Information Visualization* 19, 3 (2020), 207–233. https://doi.org/10.1177/1473871620904671

[8] Angelos Chatzimparmpas, Rafael M. Martins, Ilir Jusufi, Kucher Kostiantyn, Rossi Fabrice, and Andreas Kerren. 2020. The State of the Art in Enhancing Trust in Machine Learning Models with the Use of Visualizations. *Computer Graphics Forum* 39, 3 (2020), 713–756. https://doi.org/10.1111/cgf.14034

[9] John W. Creswell and Cheryl N. Poth. 2018. *Qualitative inquiry & Research Design: Choosing Among Five Approaches* (fourth edition ed.). Sage Publications, Los Angeles, Calif.

[10] Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory. 2020. Passing the Data Baton: A Retrospective Analysis on Data Science Work and Workers. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030340

[11] Anamaria Crisan and Tamara Munzner. 2019. Uncovering Data Landscapes through Data Reconnaissance and Task Wrangling. *2019 IEEE Visualization Conference (VIS)*, 46–50. https://doi.org/10.1109/VISUAL.2019.8933542

[12] David Donoho. 2017. 50 Years of Data Science. *Journal of Computational and Graphical Statistics* 26, 4 (2017), 745–766. https://doi.org/10.1080/10618600.2017.1384734

[13] Jaimie Drozdal, Justin Weisz, Dakuo Wang, Gaurav Dass, Bingsheng Yao, Changruo Zhao, Michael Muller, Lin Ju, and Hui Su. 2020. Trust in AutoML: Exploring Information Needs for Establishing Trust in Automated Machine Learning Systems. In *Proc IUI'20*. 297–307. https://doi.org/10.1145/3377325.3377501

[14] Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. 2014. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research* 15, 90 (2014), 3133–3181. http://jmlr.org/papers/v15/delgado14a.html

[15] Matthias Feurer, Katharina Eggensperger, Stefan Falkner, Marius Lindauer, and Frank Hutter. 2020. Auto-Sklearn 2.0: The Next Generation. https://arxiv.org/abs/2007.04074

[16] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Tobias Springenberg, Manuel Blum, and Frank Hutter. 2015. Efficient and Robust Automated Machine Learning. In *Proc NeurIPs'15*. 2755–2763. https://doi.org/10.5555/2969442.2969547

[17] 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 *Proc IUI'19*. 614–624. https://doi.org/10.1145/3301275.3302324

[18] Daniel Golovin, Benjamin Solnik, Subhodeep Moitra, Greg Kochanski, John Karro, and D. Sculley. 2017. Google Vizier: A Service for Black-Box Optimization. In *Proc KDD'17*. 1487–1495. https://doi.org/10.1145/3097983.3098043

[19] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt.

[20] Jeffrey Heer. 2019. Agency Plus Automation: Designing Artificial Intelligence into Interactive Systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850. https://doi.org/10.1073/pnas.1807184115

[21] Donald Honeycutt, Mahsan Nourani, and Eric Ragan. 2020. Soliciting Human-in-the-Loop User Feedback for Interactive Machine Learning Reduces User Trust and Impressions of Model Accuracy. *Proc AAAI HCOMP'2020* 8, 1, 63–72. https://ojs.aaai.org/index.php/HCOMP/article/view/7464

[22] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proc CSCW'2020*, Article 068 (2020), 26 pages. https://doi.org/10.1145/3392878

[23] Eric Horvitz. 1999. Principles of Mixed-Initiative User Interfaces. In *Proc CHI'99*. 159–166. https://www.microsoft.com/en-us/research/publication/principles-mixed-initiative-user-interfaces-2/

[24] Amazon Inc. 2020. Amazon SageMaker Autopilot. https://aws.amazon.com/sagemaker/autopilot/. Accessed: 2020-09-01.

[25] DataRobot Inc. 2020. DataRobot: Empowering the Human Heroes of the Intelligence Revolution. https://www.datarobot.com/. Accessed: 2020-09-01.

[26] Google Inc. 2020. Cloud AutoML. https://cloud.google.com/automl. Accessed: 2020-09-01.

[27] H20.ai Inc. 2020. H20 Driverless AI. https://www.h2o.ai/products/h2o-driverless-ai/. Accessed: 2020-09-01.

[28] IBM Inc. 2020. AutoAI with IBM Watson Studio. https://www.ibm.com/cloud/watson-studio/autoai. Accessed: 2020-09-01.

[29] Microsoft Inc. 2020. Azure Machine Learning Studio. https://azure.microsoft.com/en-us/services/machine-learning/automatedml/. Accessed: 2020-09-01.

[30] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive Visual Specification of Data Transformation Scripts. In *Proc CHI'11*. 3363–3372. https://doi.org/10.1145/1978942.1979444

[31] Sean Kandel, Andreas Paepcke, Joseph M. Hellerstein, and Jeffery Heer. 2012. Enterprise Data Analysis and Visualization: An Interview Study. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2917–2926. https://doi.org/10.1109/TVCG.2012.219

[32] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proc AVI'12*. 547–554. https://doi.org/10.1145/2254556.2254659

[33] Minyung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2018. Data Scientists in Software Teams: State of the Art and Challenges. *IEEE Transactions on Software Engineering* 44, 11 (2018), 1024–1038. https://doi.org/10.1109/TSE.2017.2754374

[34] D. Lee, Stephen Macke, Doris Xin, Angela Lee, Silu Huang, and Aditya G. Parameswaran. 2019. A Human-in-the-loop Perspective on AutoML: Milestones and the Road Ahead. *IEEE Data Eng. Bull.* 42, 2 (2019), 59–70. http://sites.computer.org/debull/A19june/p59.pdf

[35] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. *Proc CHI'20* (2020), 1–15. https://doi.org/10.1145/3313831.3376590

[36] Andrew McNutt, Gordon Kindlmann, and Michael Correll. 2020. Surfacing Visualization Mirages. *Proc CHI'20*, 1–16. https://doi.org/10.1145/3313831.3376420

[37] Judith S. Olson and Wendy A. Kellogg. 2014. . Springer New York. https://doi.org/10.1007/978-1-4939-0378-8

[38] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. 2016. Evaluation of a Tree-Based Pipeline Optimization Tool for Automating Data Science. In *Proc GECCO '16*. 485–492. https://doi.org/10.1145/2908812.2908918

[39] Randal S. Olson, Nathan Bartley, Ryan J. Urbanowicz, and Jason H. Moore. 2016. Evaluation of a Tree-Based Pipeline Optimization Tool for Automating Data Science. In *Proc GECCO'16*. 485–492. https://doi.org/10.1145/2908812.2908918

[40] Jorge Piazentin Ono, Sonia Castelo, Roque Lopez, Enrico Bertini, Juliana Freire, and Claudio Silva. 2020. PipelineProfiler: A Visual Analytics Tool for the Exploration of AutoML Pipelines. *IEEE Transactions on Visualization and Computer Graphics* (2020). https://doi.org/10.1109/TVCG.2020.3030361

[41] Raja Parasuraman, Thomas B. Sheridan, and Christopher D. Wickens. 2000. A Model for Types and Levels of Human Interaction with Automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans* 30, 3 (2000), 286–297. https://doi.org/10.1109/3468.844354

[42] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc CSCW'2018* CSCW, Article 136 (2018), 28 pages. https://doi.org/10.1145/3274405

[43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 85 (2011), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

[44] Alper Sarikaya, Micharl Correll, Lyn Bartram, Melanie Tory, and Danyel Fisher. 2019. What Do We Talk About When We Talk About Dashboards? *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 682–692. https://doi.org/10.1109/TVCG.2018.2864903

[45] Bledi Taska, Steven M. Miller, Debbie Hughes, Will Markow, and Soumya Braganza. 2017. The Quant Crunch: How the Demand for Data Science Skills is Disrupting the Job Market. https://www.ibm.com/downloads/cas/3RL3VXGA

[46] 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. *Proc. CSCW'19*, 24. https://doi.org/10.1145/3359313

[47] 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 *Proc CHI'19*. 1–12. https://doi.org/10.1145/3290605.3300911

[48] Daniel Karl I. Weidele, Justin D. Weisz, Erick Oduor, Michael Muller, Josh Andres, Alexander Gray, and Dakuo Wang. 2020. AutoAIViz: Opening the Blackbox of Automated Artificial Intelligence with Conditional Parallel Coordinates. In *Proc IUI'20*. 308–312. https://doi.org/10.1145/3377325.3377538

[49] Kanit Wongsuphasawat, Daniel Smilkov, James Wexler, Jimbo Wilson, Dandelion Mané, Doug Fritz, Dilip Krishnan, Fernanda B. Viégas, and Martin Wattenberg. 2018. Visualizing Dataflow Graphs of Deep Learning Models in TensorFlow. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (2018), 1–12. https://doi.org/10.1109/TVCG.2017.2744878

[50] Quanming Yao, Mengshuo Wang, Hugo Jair Escalante, Isabelle Guyon, Yi-Qi Hu, Yu-Feng Li, Wei-Wei Tu, Qiang Yang, and Yang Yu. 2018. Taking Human out of Learning Applications: A Survey on Automated Machine Learning. http://arxiv.org/abs/1810.13306

[51] Jun Yuan, Changjian Chen, Weikai Yang, Mengchen Liu, Jiazhi Xia, and Shixia Liu. 2020. A Survey of Visual Analytics Techniques for Machine Learning. *Compunational Visual Media* (2020). https://doi.org/10.1007/s41095-020-0191-7

[52] Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc CSCW'2020*, Article 022 (May 2020), 23 pages. https://doi.org/10.1145/3392826

[53] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making. In *Proc FAccT*'20*. 295–305. https://doi.org/10.1145/3351095.3372852

[54] Marc-André Zöller and Marco F. Huber. 2019. Benchmark and Survey of Automated Machine Learning Frameworks. https://arxiv.org/abs/1904.12054