Passing the Data Baton : A Retrospective Analysis on Data Science Work and Workers

Anamaria Crisan, Brittany Fiore-Gartland, and Melanie Tory

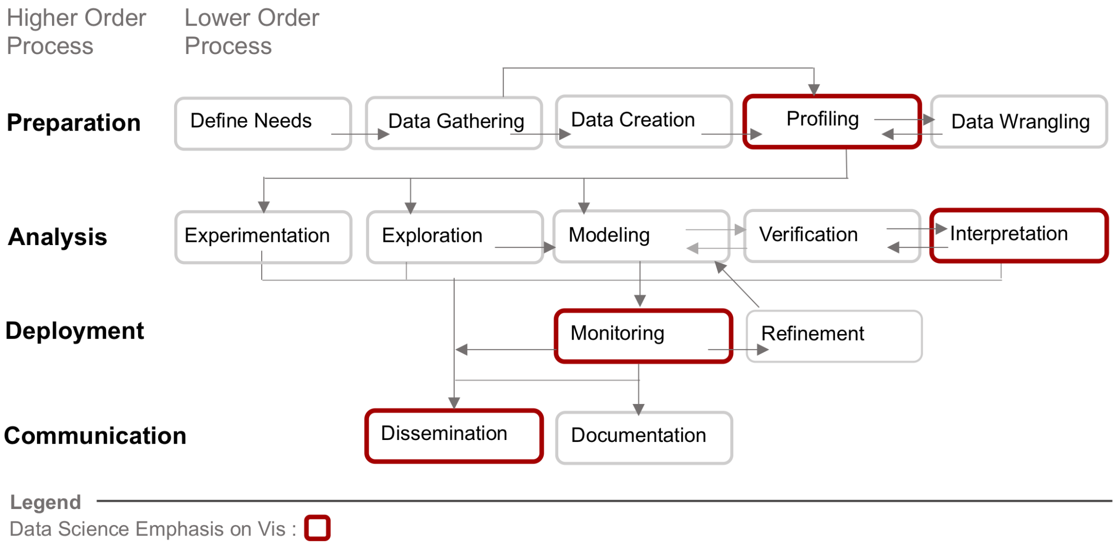

Fig. 1. Our model of data science work synthesized from an extensive and systematic literature review. We summarize data science work processes as constituting four higher order and fourteen lower order processes. Red boarders around the lower order processes highlight where we found explicit evidence in the literature for data visualization as a core component of the work being carried out, these processes were profiling, interpretation, monitoring, and dissemination. We also identified two emergent processes, collaboration and pedagogy, that we believe are of growing importance but not consistently acknowledged to be a part of data science work.

Abstract—Data science is a rapidly growing discipline and organizations increasingly depend on data science work. Yet the ambiguity around data science, what it is, and who data scientists are can make it difficult for visualization researchers to identify impactful research trajectories. We have conducted a retrospective analysis of data science work and workers as described within the data visualization, human computer interaction, and data science literature. From this analysis we synthesis a comprehensive model that describes data science work and breakdown to data scientists into nine distinct roles. We summarise and reflect on the role that visualization has throughout data science work and the varied needs of data scientists themselves for tooling support. Our findings are intended to arm visualization researchers with a more concrete framing of data science with the hope that it will help them surface innovative opportunities for impacting data science work.
Data availability: `https://osf.io/z2xpd/?view_only=87fa24be486a473884adb9ffbe8db4ec`

Index Terms—Information Visualisation, Visual Analytics, Data Science, Systematic Literature Review

1 Introduction

Data science and visualization share a common goal of helping people understand their data, offering complementary approaches toward this aim. In this spirit, both communities have created visualization platforms and libraries that support data science work. From the visualization community notable examples include libraries such as D3 [12] or Vega [71], as well as foundational research that led to systems like Trifacta [36] and Tableau [77]. From the statistical and machine learning communities libraries such as ggplot [83] and techniques such as t-SNE [79] or UMAP [52] have similarity seen wide adoption. But as data science work has become more common, the minatue of this work has grown in complexity and scope [23]. In tandem with the change is the growing diversity of individuals who engage in data work, which has provoked researchers in other disciplines to reexamine what it means to be a data scientist [23, 31, 40]. Existing studies in the visualization literature [6, 9, 36, 47, 87] have examined data scientists and analysts and similarity interrogated them about their processes and visualization needs. While these studies have generated important and actionable insights for the visualization community, we noted that their frame of reference for data science work was not consistent. Upon deeper reviewer found that these studies each captured a different aspect of data science processes and further that what was captured did not entirely overlap with processes reported in the data science literature (i.e [15,23]). While there did exist different degrees of overlap we were overall intrigued by the inconsistency and concerned it could introduce a misalignment between the efforts of the visualization and data science communities.

We believe what is lacking is a modern framework of data science, one that captures the changes in data work over time and the diversity of data workers. With such a framework in hand, researchers and

- *Anamaria Crisan is with Tableau Research. E-mail: acrisan@tableau.com.*
- *Brittany Fiore-Gartland and Melanie Tory are with Tableau Software. E-mail: {bfioregartland, mtory}@tableau.com.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

practitioners in both communities would be better able to reflect on their tools, techniques, and even elicitation methods to assess where there exist unmet needs amongst data scientists. Toward this objective we conduct a comprehensive and retrospective analysis of the data science and visualization research literature. Our contributions through this research are the following:

- A novel model of data science work, broken down into four higher order processes (Preparation, Analysis, Deployment, and Communication) composed of fourteen lower order processes.
- A novel breakdown of data scientists into nine roles and a delineation of their expertise with respect to statistics, computer science, specific domains, and human-centered design.

We believe this synthesis, derived from multiple bodies of research literature, is timely and important to both communities. Furthermore, we hope that our emphasis here on the intersection of data visualization and data science can serve as a fruitful starting point of collaboration between our two communities.

2 Related Work

We begin with a review prior research from multiple disciplines describing data science work and data scientists.

2.1 Visualization and data science

Models for data science work have their foundations in those initially proposed from the knowledge discovery and databases (KDD) community. Initially framed for data mining work, these models have influenced both academic and industrial data science practices. The KDD model was the first such model [27] and constitutes five phases: *data selection*, *preprocessing*, *transformation*, *mining*, and finally *interpretation / evaluation*. One of the best known iterations of this model was CRISP-DM, the Cross-Industry Standard Processes for Data Mining (CRISP-DM) [84]. CRISP-DM constitutes five phases: *business understanding, data understanding, modeling, evaluation, and deployment*. The CRISP-DM model also augmented these phases with descriptions of so-called 'generic tasks', 'specific tasks', and finally lower level 'processes'. A more recent variation of CRISP-DM is TDSP, the Team Data Science Process model, which acknowledges that data science work is increasingly carried out by teams that must coordinate [2]. Visualization research studies that have examined data scientist have arrived at a similar set of data science work phases [6,9,36,47,87].

We note with some consternation that neither the KDD, CRISP-DM, or TDSP models refer to data visualization, although all have explicit steps for statistical modeling. Khan *et al.* [38] in 2014 showed that visualization is a low priority concern for data scientists, but it is not clear whether this is because their needs are met by existing tools or because they do not often use visualization in their work. There are explicit references to visualization in other contemporary models, such as in Grolemund and Wickham's model for data exploration [30], Menger *et al.*'s interactive data mining model (CRISP-IDM) [54], and Moreno *et al.*'s data science life cycle [59]. However, visualization appears to occupy a limited role in these models, typically as a means to *present* the results of statistical or machine learning models.

Our research extends prior work by synthesizes a novel and contemporary model of data science work from the existing literature. In doing so, we amalgamated and reconciled many of the different phases, steps, and tasks in existing models.

2.2 Analyzing the analyzers

Harris *et al.*'s [31] seminal study took an early look at those called "data scientists"; noting the diversity in their ranks, classifying them as data business people, creatives, developers, and researchers. Kim et al. [39,40] examined the "data scienctist" role within Microsoft over a single year of explosive growth. Their findings not only corroborated Harris's but extended them by introducing new roles. Work by Wang *et al.* [81] and Zhang *et al.* [91] codified the challenges data scientists experience when working together, examining the way they share code, data, and tasks, as well as the different needs across data scientist roles.

Visualization researchers have studied data scientists as well. Kandel *et al.*'s [36] interview study established data preparation challenges that were used to guide the development of a visualization system [37]. Wongsuphasawat *et al.* [87] used a similar interview approach to examine the processes and pain points for data exploration. More recent work by Liu *et al.* [47] demonstrated how it was possible to construct graphs describing the paths of data science work unique to individual data scientists. These studies observed data science processes and how they were unique to individual data scientists and roles. Other research on data scientists examined the utility of specific visualization interventions. For example, Alspaugh *et al.* [6] and Batch *et al.* [9] both examined the use of interactive visualization in exploration and showed that it was underutilized. As we noted earlier, these studies capture and emphasize different aspects of data science work. These differences were driven by the authors' focusing on a specific aspect of data science work, for example Alspaugh *et al.* [6] focuses on exploration, or by the experiences of their participant pools, such as Kandel [36] or Wongsuphasawat [87]. We believe it is valuable to conduct an analysis across these different studies in conjunction with the findings of the data science community to capture and summarize this diversity of experiences. Finally, these visualization studies do not appear to address the diversity amongst the role of a data scientist, which has been observed by Harris [31] and others, and thus may have missed surface unmet visualization needs in data work.

As with data science work, there is considerable diversity in the description of data scientists across the literature. Our work extends existing research by conducting a synthesis of these prior studies and derives a contemporary classification of data science roles.

3 Literature Search and Analysis Methodology

In this section we describe our methodology for searching and analyzing the data science and data visualization research literature. We use established approaches from prior studies in visualization research (i.e., [35,43]), with some modifications that we describe here. An online repository contains the final list of reviewed literature and the analysis we conducted (see Abstract for link).

We used a "snowball approach" [85] to conduct our literature review. We began by assembling an initial set of documents, removing those that we considered to be outside the scope of our study, and using backward (references) and foreword (citations) to augment this initial set. To generate an initial set of data science literature, we used Semantic Scholar (`https://www.semanticscholar.org/`), which aggregates published research from peer-reviewed (i.e., PubMed, Springer, Nature, etc.) and pre-print (i.e., ArXiv) databases.

We searched for articles on (`''data science'' OR ''big data''`) published between January 2010 and January 2020. We included the term 'big data' as existing reviews on data science [14,23] indicated that data science is often synonymous with big data. Using these search criteria, Semantic Scholar returned approximately 19,600 works. We further restricted the analysis to 'journal article' types (n = 9,230 publications) that, for the purposes of reproducibility, have a PDF available (n = 3,008). In a first pass review, we noticed low specificity in the search results and further rejected articles that did not contain 'data science', 'big data', or 'data analysis' in their titles or abstracts, resulting in a final dataset of 218 articles. We randomly checked 50 of the rejected articles to be sure that we did not miss important content. We conducted a second pass review by considering all 218 articles in detail (title, abstract, full text) and accepted only papers that we determined to be relevant to data science. Two primary grounds for exclusion were: 1) the article reported data analysis specific to a domain *without* generic steps relevant beyond its context, or 2) the article presented a highly specific technique without context of how it pertains to data science as a whole. Of 218 articles, 70 were used for our analysis of the data science literature. We provide an exhaustive list of exclusion criteria in our online materials. Finally, via snowball sampling of the references, we identified an additional 9 articles. Finally, we added an additional 7 articles that were known to us. In total we conducted a deep review of 86 articles.

3.1 Qualitative Coding Process

We collected text, figures, and tables that described data science work and workers from this set of 86 articles. We performed an initial open and axial coding pass that produced a set of codes for the definition of data science, the description of data science work as processes, and finally a classification of data science roles. We further arranged codes for data science work into a set of four higher-order processes (preparation, analysis, deployment, and communication) that comprised fourteen lower-order processes. We also identified two emergent processes (collaboration, and pedagogy) that we deemed were important to data science work but did not appear to be recognized as such; due to this lack of strong evidence these processes were not assigned to a higher-order process but we still included them for consideration. As we developed these processes codes we also paid close attention to the specific mention of visualization use. Moreover, through this round of coding the importance of the roles became evident, something we had not anticipated. We found that in the literature outside of the visualization research community linked data science work to the expertise of the analyst; this was especially true of studies like Harris *et al.* [31] and Kim *et al.* [40]. We observed that data scientist expertise was measured consistently along axes of computer science, statistics, and domain expertise. Focusing on a subset of studies that reported on data scientists, we developed a set of codes for different roles and also classified them according to higher-order processes.

These codes and the evidence that supports them were shared with a larger group of individuals that engage in various facets of data work. We recorded their feedback in written notes and used these, along with the artifacts we collected from the research articles, to perform another coding iteration. We met frequently to discuss the codes and we continued to refine them until we reached a point of saturation. The final results of our coding processes is reflected of the structure of our paper, beginning with our working definition of data science, our description of data science processes, and finally our classification of data science roles. The online materials contains our coding artifacts.

4 What is Data Science?

There exists ambiguity and even some contention as to what data science actually is. It has been debated whether data science is a field of study in its own right, or whether it is a descriptor for a set of methods and techniques [3, 23, 66]. A number of experts from across several disciplines have argued that data science deserves a distinct status from other disciplines because of the unique emphasis it places on the *integration* of computation, statistics, and domain-specific knowledge [11, 14, 19, 23]. That is, domain experts may use data science techniques to achieve their means, but are generally less inclined to "scientifically study data analysis", which is regarded as the purview of data science [23]. That is, the field of data science adds critical structure to the ***way*** that statistical and computational techniques are applied to answer data questions. However, others argue that data science augments applied research, which has developed a similar set of methods and techniques because of a need to analyze increasingly large and heterogeneous data [11, 13, 23, 24, 33, 54, 72, 76, 86]. As a middle ground, some disciplines have crafted the nomenclature of 'data science for X' [14], where X is a substitute for a discipline such as biology, economics, or nursing. Where experts generally agree that the need to integrate across multiple disciplines coupled with the challenges of creating a robust analytic infrastructure have introduced a unique set of challenges that are best tackled by so-called "data scientists" [11,23–25, 65, 66]. It is also clear that the rise of the "data scientist" has impacted academic curricula in statistics, computer science, and other domains, which are being actively redeveloped to prepare new researchers and practitioners with the set of skills that data science work demands [11, 19, 23, 63, 65].

To define the scope of our investigation, we use the following working definition : **Data science is a multidisciplinary field that aims to learn new insights from real-world data through the structured application of primarily statistical and computational techniques.** We acknowledge that the validity of data science as a field in its own right is complex and entangled in the,history of data analysis and science in general, as succinctly described by Gray's 'Fourth Paradigm' [33]. We, as do others [11, 23], attribute the technological changes that enables the collection of larger datasets to have played an important role in the evolution of data science. Domain experts in specific discipline have used computational and statistical techniques to conduct analytic research on these large datasets, however, the ***means*** for applying those techniques did not emerge from domain experts alone. Nor did those means emerge solely from statistics or computer science. It is subtle, but it is that intermediary space between disciplines that data scientists occupy and whose efforts to integrate techniques from across disciplines are essential. We consider this to be a worthwhile and interesting group and so focus on them in our investigation.

5 Data Science Processes

We model data science work as a set of interconnected higher and lower order processes. Prior research has used the terminology of 'generic' and 'specific' tasks to describe data science work [84]. We have instead chosen to use 'processes' because it is a term that encapsulates the variety of tasks and procedures a data scientist may carry out. From our analysis of the data science literature we have synthesized **four higher order** and *fourteen lower order* processes that model data science work:

- **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*

Altogether these processes form our model of data science work, as shown in Fig. 1. We also illustrated the relationships between these processes by including arrows that suggest the flow of knowledge and artifacts (data, code, and models). *While we present these processes as discrete ordered chunks, in reality, there is considerable overlap between them as well as a non-linearity in terms of when and how they take place. Thus, this model should be consider more of baseline to contextualize data science work rather than as an absolute ground truth.* Two additional lower order processes that we identified, collaboration and pedagogy, are not routinely considered to be data science work. While we exclude these processes from Fig. 1, we believe that they are important and merit discussion.

We now present definitions for the higher and lower order processes that we synthesized from the data science literature. We also identify visualization research that aligns with these processes.

5.1 Preparation

Preparing the data ahead of analysis is an essential component of data science work [22, 23, 27, 36, 39, 81, 84, 87]. Given the importance of preparation work, it is not surprising that prior research has found these processes to be the most time-consuming [1, 23, 40, 56, 60]. Preparation processes requires considerable human intervention and are difficult to automate [40]. ***We synthesized five preparation processes : defining needs, data gathering, data creation, profiling, and data wrangling.***

Defining needs is the process of translating analytic objectives and goals, usually defined by an external stakeholder, as a viable set of requirements for data science work [84]. These requirements can include data, statistical analysis plans, and expected deliverables. Deliverables are typically static reports, but they can also include statistical or machine learning models, dashboards, or computational infrastructure [23]. Correctly framing a viable "data question" is a critical component of defining needs, but a challenge for many data scientists [51]. Curiously, not all models of data science work include this step, but we do not know whether that is because it is implied or because it is overlooked.

Data gathering is the process of identifying suitable datasets from a landscape of many candidates. Gathering data can include developing a set of criteria to gather data, if it does not currently exist, or generating new data by integrating and transforming two or more existing datasets [56]. Identifying suitable data is challenging, especially when there are many potential datasets to navigate [22, 78]. Even when data exist, gaining access to data and being permitted to integrate and transform data can be challenging [21, 60].

While the majority of data science work will use real data, there are situations where, as part of data preparation, new data must be generated. **Data creation** is the process of creating new data that cannot be collected or directly observed [26, 56, 60]. For example, Abt *et. al.* [4] highlights the challenges of obtaining the real world data for cyber security experiment and the way that synthetic data could be used to *inform* decision making and not simply to validate methods. Akios *et. al.* [41] highlighted the different ways to create synthetic dataset and the importance of domain expertise in the data creation processes. Finally Crisan et. al. [21] highlight a case study of synthetic data generation to enable visualization design and evaluation [21]. Synthetic data has been more commonly used as a way to validate downstream modeling and techniques [23], but we argue that the role of synthetic data to inform decision making is also a valid use of such data and one that is often overlooked. Finally, we believe it important to distinguish between our use of data creation here and the derivation of new data via data transformations. When there *does not* exist a dataset and one must be generated, then we refer to this as a process of data creation. When a data set exists, but transformation methods are applied to it to derive new dataset or features then we consider this to be a data wrangling process.

Profiling is the process of assessing attributes (individually and jointly) to understand the distribution of their values, identify missing values, and examine their associations with other attributes. Profiling is tightly coupled with assessing data quality and undertaking steps to understand data content [36, 84]. Prior data science models use a myriad of terms to describe profiling, such as as 'pre-processing', 'tidying', 'cleaning', 'wrangling', 'exploratory data analysis (EDA)', or 'data understanding' [30, 38, 84]. We found evidence that data visualization is often used in profiling to create "very simple plots of EDA — histograms, scatterplots, time series plots" [23], implying that profiling involves simple plots of univariate and bivariate of data distributions.

Data Wrangling is the process of shaping and transforming data so that it is suitable for analysis [36]. It is perhaps the best known of the preparation processes. Data wrangling has also been called 'tidying data' [30]. We separate profiling from wrangling because we consider profiling to be the process of understanding and assessing the data, while wrangling is the process of acting on profiling knowledge by shaping or transforming data.

5.1.1 Data visualization in preparation

The data science literature identifies a role for visualization in the profiling and wrangling of data. In this area, Kandel *et. al.* [36] produced a highly influential study that others within and beyond the visualization research community have built upon including Kandel *et. al.*'s follow-on work, the Profiler system [37]. Visualization research has also touched upon other preparation processes. As for defining needs, there exists a body of methods, primarily centered around design studies [73]. However, these methods are intended for visualization researchers and have not been adopted by data scientists. Given that data scientists often develop their own visualizations, they may benefit from these methods. Visualization researchers have also explored data gathering and creation, but to a lesser degree. Crisan *et. al.* [22] highlight the challenges that analysts face when confronted with a complex data landscape and showed the present gaps in visualization research toward helping data scientists with these challenges. Finally, while there is some visualization research on data creation, it is our assessment that most of this research is geared toward volume rendering rather than the generative modeling processes used by data scientists [23].

We also noted that the use of visualization for EDA is more narrowly defined toward profiling, compared to visualization research that applies EDA more generally [10]. We argue here that it is important to differentiate between an initial probing of the data to assess its quality and suitability, which we call profiling, and a more surreptitious exploration, which we consider to be an analysis.

5.2 Analysis

Analysis processes take (ideally prepared) data and apply statistical and computational techniques to derive insights from the data. These processes are arguably the heart of data science. Modeling tends to receive the most attention of all analysis processes, but it is in reality a relatively small part of data science work overall [19, 23]. ***We synthesized five analysis processes: experimentation, exploration, modeling, verification, and interpretation.***

Experimentation is a processes evaluating a cause-and-effect hypothesis, and one where the analysts exerts considerable control over the administration of some intervention (i.e. website layout, a visual encoding). In it's simplest form this kind of hypothesis verification can involve the application of standard methods of null hypothesis testing. Experimentation can have a closely coupled relationship with data gathering when experts collect data specifically to test a hypothesis, such as is often done in A/B tests and other variations of randomized trials. It is important here to distinguish between the colloquial use of the term 'experiment', which can mean to tinker with or try out something, compared to the more constrained definition we use here that relates to the process to setting up and executing a specific experiment.

Exploration seeks to uncover new insights from data that, unlike in experimentation, were not predetermined from the outset. This does not mean that data scientists do not have some specific question in mind, but rather that the data itself is not gathered for one specific purpose. Our definition of exploration is actually quite narrow here and excludes the initial 'getting to know the data' type exploration, which we attribute to the profiling process. In several data science models, the process of exploration is also more general and encompasses modeling (Grolemund and Wickham's model [30] for example). Model driven exploration can result in new hypotheses being generated from the data. Visual exploration of the data can also drive hypothesis generation [6, 10, 10], but we did not find much evidence that exploratory visual analysis is used much beyond profiling. While exploration is valuable, it is also often discouraged, primarily to avoid p-hacking [23].

Modeling is the process of applying statistical and computational techniques to derive an actionable insight from data.Although we present the modeling here as a discretized process, we acknowledge that it overlaps with experimentation and exploration. We differentiate modeling from these processes in two ways. First, we consider modeling to apply to data that has not be gathered under experimental conditions. We argue it is worth making this distinction as experimental data has different properties compared to routinely gathered data. Second, we also wanted to capture a common set of tasks associated with creating, tuning, and selecting models that are separate from exploring data.

The data science literature has identified four primary models: descriptive, diagnostic, predictive, and prescriptive [15, 23]. *Descriptive modeling*, also sometimes called explanatory modeling, summarizes the present state of the data. For example, descriptive modeling can inform a regional sales manager that her sales are down this quarter. *Diagnostic modeling* attempts to assign some root cause to an observed outcome, for example, identifying *why* sales are down this quarter. *Predictive modeling*, as the name suggests, attempts to predict some future outcome; for example, that sales will also be down the next quarter. Forecasting is a subset of predictive modeling that is specific to time-series data. Finally, *prescriptive modeling* identifies a specific intervention that can be taken to modify future outcomes; for example, if the sales manager hires more people, her sales will increase. Prescriptive modeling is extremely useful, but it is difficult to execute as it relies on well-established gold standard datasets (sometimes, but not exclusively, derived from experimental studies); the necessity for such a gold standard datasets is that, in order for prescriptive model to be useful it must have an established set of cause and effect relationships to learn from. While these four types of modeling can all involve different underlying statistical and computation techniques, they all share some common model development tasks; feature engineering, tuning, and optimization are some examples where the particulars may differ based upon on the type of modeling [81]. Visualization can also be used to verify that the data satisfy the modeling assumptions and that the outputs make sense [30, 54]; this could be called "model profiling".

Verification is an evaluative process to confirm the robustness of the results, code, and models [19, 38, 81]. Importantly, this verification process must make assessments toward external and ecological validity, whereas assessments of internal validity should be carried out in the modeling process. For example, a withheld or ideally newly generated dataset enables assessments of the external validity while the traditional train/test split of the data can verify only the internal validity. It is also emerging that verification is increasingly important to assess the fairness and transparency of data science processes [50, 81].

Interpretation is the process of understanding the results of an experiment, exploration, or model in terms of real world applications. The interpretability of analysis processes can be limited by the 'black box nature' of both complex models and datasets, which impacts how data science results can be safely deployed and used by others [50]. Few data science models that we investigated contained an explicit interpretation step and some combined both verification and interpretation with modeling [8, 14, 23, 38]. Visualization has an emergent role in the area of interpretation, especially toward explainable machine learning / artificial intelligence systems [34].

5.2.1 Data visualization in analysis

The data science literature is inconsistent on the use of visualization in analysis. Visualization appears often as a "sidekick" to modeling processes and is often limited to simple static charts. In contrast, visualization research appears to be highly engaged in data science analysis processes, with interactive visual analytics tools for exploration, modeling, and increasingly interpretation. Battle *et al.* [10] succinctly describe the prior art in exploratory visual analytics, including applications relevant to what we consider to be profiling. To address concerns of p-hacking in these exploratory visual analyses, Zgraggen *et al.* [90] and Lee *et al.* [44] propose mechanisms for penalizing serendipitous exploration. However, it is not clear from the data science literature whether strategies for exploratory visual analytics are addressing pressing problems for data scientists. Visualization research into modeling is likewise vast and spans different types of data (text, spatial, network, etc.) and different applications. One area where visualization research appears to play an active role is in visualization tools that support model development; we refer the reader to several in depth state of the art reports by Hohman *et. al.* [34], Chatzimparmpas *et al.* [16, 17], and Yaun *et al.* [89] for an overview of current techniques and strategies. One noteworthy tool for visual model development that has bridged between data science and visualization research is TensorBoard [88]. TensorBoard is a great example of what our two communities can achieve together. Finally, visualization research is taking tentative steps towards improving understanding and interpretability of machine learning models. This is best exemplified by work presented in the VDS symposia and VisXAI workshops and in the contributions of VIS researchers to publications like Distill (`https://distill.pub/`).

5.3 Deployment

Within many organizations, data science work extends to a phase where data preparation and analysis processes are routinely put into production and deployed to tackle real-world problems [39, 54, 59, 81]. ***We synthesized two deployment processes: monitoring and refinement.***

Monitoring is the passive observation of productionized data science processes to ensure that they behave in consistent ways [14, 81]. This could also be considered a surveillance process [59]. Monitoring is not restricted just to data and models, but includes the computing infrastructure that automatically processes incoming data (and increasingly, the costs associated with that infrastructure).

Refinement is the processes of updating and optimizing analytic models and data science processes once they have been deployed to operate on real-world data. Some refinements may happen automatically, such as updating a model with new data, while other refinements are the direct result of feedback from end users or through monitoring [38, 81, 84].

5.3.1 Data visualization in deployment

The data science literature often ascribes the use of dashboards to monitoring processes, in order to surface and conduct initial investigations into anomalous events [59]. Visualization researchers have broadly examined the use of dashboards, including for operational purposes [69]. However, we could not distill whether operational uses of dashboards encompass monitoring and refinement of complex data science artifacts like data, code, analytic models, and computing infrastructure. Moreover, these artifacts change over time and the specific challenges this introduces are not fully described in prior visualization research [69]. There do exist systems that, although not described as dashboards, could approximate the needs of data scientists in monitoring processes. LiveRAC [53] and EventFlow [58] are two examples of systems that provide high level overviews of temporal trends and, in different ways, allow a user to drill down and examine anomalies.

5.4 Communication

Communication is a cross-cutting set of processes within data science work that is essential to the circulation of artifacts and knowledge. ***We synthesized two communication processes: documentation and dissemination.***

Dissemination is the process of creating and circulating insights surfaced throughout the prior data science processes. In other models, dissemination has been called 'communication' [30], 'reporting' [36, 87], or 'presentation' [23]. We consider communication to be much more encompassing than the dissemination of findings and have deemed it to be a higher-order process. In the data science literature, reporting is about creating documents, often static, that summarize data science work and its findings. However, dashboards are also a widely-used reporting method [23]. While reporting can be about presenting findings, it also refers to the act of delivering a presentation, again often accompanied by some visual aids. We consider dissemination to encompass both reporting and presentation. There is a broad consensus that visualization plays an important role in dissemination [23].

Documentation is the processes of generating a record that describes and synthesizes data science work and its artifacts. There is a long history in computer science and engineering work of documenting code, which informs modern data science practices. However, since data science work and its artifacts include data, code for workflows or explorations, and models, the nature of documentation in data science is more complex than in other engineering practices. Computational notebooks are a relatively new form of documentation that enable a wide range of interactions [1, 64, 67, 80, 91]. Their uptake in the data science community, in particular Jupyter notebooks [1, 64], demonstrates their value across data science work. Prior studies advocate for such notebooks to support reproducible workflows, not only amongst groups of data scientists but for the individual data scientist as well.

5.4.1 Data visualization in communication

Amongst all the data science literature we reviewed, the greatest consensus on the importance of visualization was for communication. The data science literature seems to agree with observations by other visualization researchers [69] on the importance of using dashboards to provide some interactive engagement, but it does not go much further toward evaluating whether the dashboards produced by data scientists serve their intended purpose. In contrast, there has been less focus, in both visualization and data science communities, on documentation. Code notebooks are essential to data science work [1, 91] and within them are rich visualizations that document analytic processes and their results. Finally, the modeling phase grows in complexity, visualization also serves as an important medium to communicate between humans and machines [32]. Along with the development of additional metrics and heuristics visualization is used to support the exposition of decision making mechanisms behind machine learning, and more recently so called artificial intelligence, algorithms [5, 57, 75]..

5.5 Emerging types of data science work

We also identified collaboration and pedagogy as two lower-order processes that are not typically acknowledged to be a part of data science work, but we strongly believe that they should be.

Collaboration within data science offers unique challenges due to the diversity of artifacts and individuals involved in data science work.

Study	Year	Size	Context	Participants
Kandel [34]	2012	35	Industry Wide	Data Analysts
Kandogan [36]	2014	34	Industry Wide	Data Analysts
Harris [27]	2015	250	Industry Wide	Data Scientists
Kim [38]	2016	16	Microsoft Only	Data Scientists
Kim [39]	2017	793	Microsoft Only	Data Scientists
Batch [7]	2017	9	Government	Data Scientist & Economists
Alspaugh [4]	2018	33	Academia & Industry Wide	Data Analysts & Data Scientists
Muller [57]	2019	21	IBM Only	Data Scientists
Wongsuphasawat [82]	2019	18	Academia & Industry Wide	Data Analysts
Kaggle [1]	2019	4140	Industry Wide	Data Scientist
Wang [77]	2019	195	Industry Wide	Data Scientists & Others
Zhang [87]	2020	183	IBM Only	Data Scientists

Table 1. Studies of data scientists, or data analysts engaged in stated data science processes, that we analyzed.

We note that these challenges are also mirrored in the growing open science movement more generally [29, 62]. Documentation serves an important role in generating artifacts that support collaborative processes. Although notebooks are the primary means through which data scientists collaborate, they are not the only means of collaboration. Project management techniques, borrowed from software engineering, are commonly used to help data scientists break down "data questions" into viable tasks [51, 68]. In aggregate, data scientists developed a Common Task Framework that is shared amongst themselves to accomplish data science work [46]. The importance of developing a shared understanding of tasks is emphasized by Donoho [23], who credits it as an essential, but underrated, component of data science's success.

Pedagogy is a part of data science work because of the relative newness of the field, which necessitates not only active curricula development but also surfaces the need for 'on the job' training and mentorship [19, 23, 63]. Increasingly, data science learning takes place via massive online open courses (MOOCs) with thousands of students in applied online learning environments [42, 74].

5.5.1 Data visualization in collaboration and pedagogy

The collaboration and pedagogical challenges that arise from data science work are often not acknowledged, but we argue that they should be. Moreover, visualization research could play a larger role in alleviating challenges stemming from these. One area where data scientists struggle is around the breakdown and sharing of tasks [51, 68]. Here visualization research already offers some potentially interesting solutions. Loorak *et al.* [48] describe the problem of asynchronous hand-off in dashboard creation, designing a system to mitigate these problems. Zhao *et al.* [92] similarly tackled this hand-off problem with knowledge graphs; these graphs could be integrated with ideas from Liu *et al.* [47] to address collaborative pain points for data scientists. Furthermore, Sarvghad *et al.* showed how analytic provenance provides a mechanism for knowledge sharing that ultimately improves the analysis of data [70]. Toward pedagogy, visualization researchers could play a larger role in developing a rich curricula for visualization in data science. Visualization researchers could also advocate for better pedagogical approaches for data visualization; recent workshops on visualization pedagogy have showcased several promising approaches. While often overlooked, it is also conceivable for visualization to play a larger role in the very process of teaching data science. Keeping track of students' progress and pain points in MOOC environments can be challenging, but this could be better supported with data visualization tools (e.g. [18, 28]).

6 WHO ARE DATA SCIENTISTS?

A surprising finding from our analysis was the diversity among data scientists and how this related to the nature of the work they carried out. The role of "data scientist" did not exist more than a decade ago, yet today is a highly sought after title [11, 23]. There is a general consensus that data scientists have a multidisciplinary background, but when Harris *et al.* [31] conducted a study of data scientists, they found an impressive array of diversity. Their findings led them to designate four data science roles: developers, researchers, creatives, and business people. Moreover, Harris demonstrated that these "data scientists" varied in terms of the data science processes they carried out and their technical and domain expertise. They astutely observed that this diversity can "lead to poor communication between data scientists and those who seek their help". Follow-on studies ([24, 25, 39, 40, 91]) and commentaries on data science work [11, 15, 23] have confirmed Harris' findings and identified additional data scientist roles.

In this section, we analyze twelve studies of data scientists (Table 1) and from these synthesize a total of nine data science roles (Table 2). Commensurate with previous studies, we also illustrate the expertise of these roles in statistics, computer science, and specific application domains. Additionally, we include human-centered design (HCD) in this analysis. An HCD skill set is increasingly importantly to help data scientists translate their findings in communication with stakeholders and end users [7, 11], but this is an area of expertise where they receive the least training. We argue that the relationship between roles and expertise to influences the types of visualization tools that data scientists may need. For example, an ML/AI engineer and moonlighter could both be engaged in a tasks of model tuning and selection, but the engineer (who has more expertise in computer science and statistics) will need a different kind of visualization tool compared to a moonlighter who does not have this expertise. Critically, these different roles that we define here may all be classified as 'data scientist' organizations; titles such as ML/AI engineer or Data engineer are more recent and born our of growing organizational maturity toward data science. Thus, if visualization researchers do not recognize the diversity within data science roles, there is the risk of forming inaccurate or incomplete understanding of data science work in our studies. However, we also acknowledge that classifying individuals solely on their expertise can create artificial divides, when in reality there exists more fluidity within these roles. Thus, we urge readers to use our summaries in Table 2 as a guide for their own investigations as opposed to a ground truth.

To further situate these data science roles within the visualization research literature, we also cast them according to the collaborator terminology developed by Sedlmair *et al.* [73]. *Importantly, these data science roles help provide context to the nature of data science work.* As we will show, the different data science roles carry out different types of data science work and are likely to benefit from different tooling support. Thus, understanding these roles is essential to grounding visualization research.

6.1 A breakdown of data science roles

We define nine data science roles that operate within the higher order processes we defined in Section 5. Individuals who carry out work across all data science processes tend to be generalists, but it is currently more common for data scientists to specialize and work collaboratively as a team.

Data Stewards and Shapers primarily conduct preparation processes to govern the access and use of data, as well as to set up data for use by other analysts [39, 87, 91]. In terms of the visualization literature, these roles serve as *gatekeepers* [73] because of their oversight of data. Data stewards have deeper domain knowledge that enables them to delineate the use of data and also attend to organizational and regulatory constraints that dictate its use and reuse [21]. Data shapers also have domain knowledge but will have a more computational background that is required to capture, create, shape, and transform data for analysis [40]. It is worth noting that there are other interesting data preparation roles that are not typically associated with the role of data scientist. One study on data preparation [49] identified the roles of Data Librarians and Archivists; we consider these two roles to be a part of the larger data science community of practice. Both data stewards and shapers need to understand the provenance or lineage of data, as well

				Level of Expertise			
Process	Role	Role Description	In prior Studies	Statistics	Computer Science	Domain Knowledge	Human Centered Design
Preparation	Data Steward	Domain expert responsible for governing access and use of data	Data Broker [87]; Data Owners [84]	□	□	■	□
	Data Shaper	Developer responsible for supporting the curation and preparing data for analysis	Data Shaper [38]; Data Preper [38];	░	▓	▓	□
Deployment & Engineering	Data Engineer	Engineer proficient in developing Data Science technologies, including data preparation and analysis pipelines	Platform Builder [38]; Data Developer [27]; Hacker [35]; Scripter [34]; Engineer [57]; BI engineer [84]	░	■	▓	□
	ML / AI Engineer	Engineer proficient in developing and deploying machine learning / artificial intelligence methods to support data science processes	Hacker [33]; Modeling Specialists [38]	■	▓	░	□
Analysis	Generalist	Multidisciplinary individual focused solely on data science	Polymath [39]; Data Creative [27]	▓	▓	▓	░
	Research Scientist	A domain expert involved in research typically with technical expertise in 'Data Science' technologies	Data Researcher [27, 57]; X-informatician [12]	▓	▓	■	░
	Technical Analyst	A technical individual from whom data science is not core to their job but occurs only at the margins of other work	Data Analyst [4, 38, 39, 57, 84]; Application User [35]; Business Analyst [36, 57]; Data Business Person [27]; Analysis Team Members [84]	░	░	■	□
	Moonlighter	Non-technical individual tasked to perform data science duties, either voluntarily or through necessity	Moonlighter [39]	□	□	■	□
Communication	Evangelist	Manager, team leader, or analyst tasked with disseminating findings from data science work	Data Evangelist [39]; Communicator [87]; Insight Actor [39]	░	░	▓	□

Table 2. Summary of data science roles and an illustration of their skill sets. Data science skills were classified along four axes: statistics, computer science, domain knowledge, and human centred design. We use a color gradient to illustrate a level of expertise: proficient ■; knowledgeable ▓; working ░; and little to none.

as how it is being used and by whom. Data visualization may be useful to help them understand the freshness and quality of their existing data. Shapers are more likely to need tools to manipulate data, assess data quality, and facilitate data transformation and shaping. Finally, both groups would benefit from tools that help them understand the landscape of available data and whether what is available satisfies the needs of analysts making data requests.

Data Engineers and Machine Learning / Artificial Intelligence (ML/AI) Engineers primarily work within deployment processes but also engineer systems that facilitate data science work in general. In the terminology of the visualization literature, these roles represent *fellow tool builders* [73] who, among other things, are capable of developing their own standalone data visualization tools or effectively using charting libraries in various programming languages. Data engineers [23, 31, 40] ensure that artifacts can move seamlessly across data science processes and can scale to changing demands on the system, especially in situations where organizations have massive volumes of data. Data engineers tend to have a stronger background in computer science. ML/AI engineers tend to focus much more on the modeling phase of the data science process and tend to have a stronger background in statistics. Engineers have highly specialized needs of data science and visualization tools. There is space for visualization researchers to collaborate with engineers in well defined projects; a good example is TensorBoard [88]. There are also opportunities for visualization research to support improved monitoring and refinement of data science processes in the development and production life cycles. As we indicated in Section 5, there are interesting opportunities for visualization researchers to think more deeply about how complex artifacts like code, data, and models can be effectively summarized over time.

Generalists and to some extent **Researcher Scientists** are what might be considered the prototypical data scientist. These individuals have a broad multidisciplinary skill set [14, 31, 40] that includes domain knowledge and increasingly, skills in human-centered design. While these individuals are primarily situated within analysis processes, they can be dropped into others as needed. Individuals may be generalists by necessity, especially since many organizations do not know how to support data scientists or even what is involved in data science work. However, as organizations evolve, generalists are likely to be supplanted by specialists. Individuals who hold these data science roles can be classified by visualization researchers in various ways, such as *front-line analysts* or *fellow tool builders* [73]. Generalists can also play a vital role as translators, because they typically have advanced knowledge and expertise across domains. The visualization needs for generalists and research scientists can mirror those of data and ML/AI engineers. However, these roles may be more tolerant to early-stage or less developed prototypes, whereas engineers may require more mature solutions.

Technical Analysts and **Moonlighters** are classified as data scientists by their organization because they use data tools to accomplish their domain-specific tasks [31,40,87]. In the terminology of the visualization literature, these roles are *front-line analysts* [73] or individuals who might be classified simply as domain experts. Although these individuals occupy a data science role, they are less likely to have received technical training. **Moonlighters** [39, 40] are an especially interesting group who only engage in data science work intermittently. These individuals may spend anywhere between 25% and 50% of their time engaged in analysis [40]. Technical analysts and moonlighters represent data science roles that could most benefit from analytic assistance. As such, they may make the greatest use of visual analytics tools that assist them in their analysis and help them to contextualize the results of machine learning algorithms.

Evangelists represent a growing and important data science role in communication processes. In the terminology of the visualization research literature, they are *connectors* and *translators* [73]. These individuals work closely with data science teams and other stakeholders to evangelize and contextualize the results of data science work. Because of their role, evangelists must have at least high-level knowledge of statistics, computation, and the application domain. These individuals may be managers as well as individual members of a team. "Insight actors" [40], whose primary objective is identifying and disseminating actionable data insights, is another interesting way to define this group. What sets evangelists apart from other roles is that they are more likely to be translators of data science artifacts (data, results, models) and not active developers of these tools. Evangelists need tools to help them interpret these artifacts, often generated by others, and to circulate their findings to stakeholders. Some existing visualization tools could extend to the needs of evangelists. However, the literature is also much more primed toward analysts, such that we have incomplete knowledge of evangelists and whether their needs are met by visualization tools.

6.2 On data science tools

Two of the studies that described data scientists made detailed assessments of the tools used for analysis and collaboration [1, 91]. They found a clear preference among data scientists for notebook environments and identified the importance of seamless transitions between data science processes and tools. In order for visualization research to have greater influence in data science, our contributions will likely need to integrate with these existing tools. An example of one such successful integration is UpSet, which was redeveloped as an R package a few years after the original publication and experienced wider adoption [20,45].

7 Discussion

A better understanding of data science work and the diversity of data science workers will support visualization researchers to identify opportunities for collaboration and innovation. Our modeling of data science work and workers is intended to arm visualization researchers with the means to educate, converse, and collaborate with data scientists and others in the larger data science community of practice. Visualization research has a lot to offer, including best practices, techniques, methodologies, and systems. In turn, there is much that visualization researchers can learn from data scientists.

We encourage visualization researchers to use our model of data science work and breakdown of data science roles in their own investigations. The prior art in visualization research literature has relied primarily on interview studies to elicit the nature of data science work (i.e. [6, 9, 36, 87]), which while rich capture a specific segment of data science work. We build on these results and others to develop our model of data science work. Using our findings as a working base, researchers can, for example, explore data that is being handed off between lower order processes, or characterize who data scientists are and how they use visualization, with a greater degree of specificity than was afforded by previous work. We ourselves intend to use the results of this research to inform our future study development and analysis processes. We anticipate that this model will continue to evolve over time, and we encourage others to elaborate (or contradict) our work so that all may learn. As we conducted our investigations, we also surfaced specific research avenues for visualization research in data science that we highlight in this discussion.

7.1 Visualization to support data science work

Visualization research should seek to support underserved data science processes. We suspect that the visualization research literature emphasizes exploration and modeling processes, much like the data science literature does [23], and places less emphasis on other data science processes.There do exist important and interesting ways that visualization research can support exploration and modeling. But looking beyond these processes can help us work together with data scientists to surface other, potentially more pressing, pain points in their work and either amplify existing visualization solutions or develop new ones. Moreover, visualization researchers bring knowledge and methodologies of human-centered design practices, which is a skill of increasing relevance to data science. Another challenge that the data science community is grappling with is developing techniques to learn from large, diverse real-world data that are reproducible, responsible, and ethical [50]. As an example of this conversation, consider the multidisciplinary ACM FAccT (Fairness, Accountability, and Transparency) Conference. This conversation is one among many that recognizes the importance and contributions of human interpretation and perspective throughout the data science lifecycle [61]. Human-centered design becomes critical in this new frontier of data science work and visualization researchers can play a role as partners and knowledge translators.

7.2 Visualization to innovate data science work

Data science encompasses a diverse set of roles and processes. Rather than build for a data scientist, as if all of data science exists within an individual, drive innovation by building for a data science community of practice. Viewing data science as a community of practice highlights the collaborative and multidisciplinary nature of the work and recognizes that a broader set of people might engage in aspects of data science work without identifying as data scientists [82]. Our findings allude to this community of practice through the technical analyst, moonlighter, and evangelist roles. However, the data science community of practice encompasses a broader range of roles that we have not articulated in our research. As data science work is increasingly understood as sensemaking about our world, it becomes imperative to include diverse voices and domain knowledge engaged in doing data science work [61]. In many ways, visualization has an opportunity to catalyze more robust human-centered approaches throughout the data science lifecycle [55]. However, these solutions will only be effective if they can integrate with the myriad of tools used and perspectives across the data science community of practice.

One particular challenge that our investigation surfaced was the effective sharing of knowledge and artifacts throughout data science processes and amongst data scientists and their larger community of practice. *While it was evident that data and code need to be shared across data science processes, it is less obvious that tasks also needed to be shared.* Effective task sharing is a common problem amongst data scientists that is critical to the success of data science (Section 5) and one that lacks effective solutions [23, 51, 68, 91]. We conclude that data scientists not only struggle to define a concrete taxonomy for describing tasks, but also lack a consistent way to break them down [51, 68]. Handing off tasks to other individuals within and across data science processes is also complex, and increasingly there is the need to hand off tasks between people and autonomous systems. The problem space of task sharing provides an interesting opportunity for mixed-initiative visualization systems [32] to serve as an intermediary within the data science community.

7.3 A virtuous cycle of collaboration

A virtuous collaborative cycle is one where both communities learn from each other. While we emphasize here the ways that visualization research can impact data science, we would be remiss if we did not also acknowledge there is also much that *we can learn from data science*. We have indicated throughout that data scientists do routinely visualize data, but the visualization research community knows relatively little about the visualization artifacts they create, how they create them, and how these are used. Our understanding of how visualization is used with notebook environments is also relatively sparse. Moreover, data scientists in all roles have a deep understanding of data that could augment our own understanding of different data types, how they are integrated together, and how they are analyzed. Finally, data science processes produce varied and complex artifacts that, if examined closely in partnership with data scientists, could launch new and exciting joint research trajectories.

8 Where do we go from here?

Data science is a complex multidisciplinary field carried out by teams of people with varied backgrounds and roles. The discipline of data science has experienced rapid growth in the last decade and the myriad of people and processes involved in this work are not well supported by existing tools. While we have shown that visualization research touches all of the data science processes, the actual use of visualization is, from our assessment, limited. We believe it important that visualization researchers engage more directly with data scientists and seek out opportunities for mutual collaboration that advance both of our communities. Yet without a more concrete idea of what data science is, it can be complex to initiate such partnerships. With this work, we sought to provide some more concreteness to ambiguous nature of data science. We now invite you, our reader, to take the next step. Befriend a data scientist. Share your knowledge and learn from them in return. Embark on a joint project and then expand it into a long and fruitful collaboration.

Acknowledgments

The authors wish to thank the research team at Tableau, as well as others in the organization, for their feedback and suggestions.

References

[1] State of data science and machine learning 2019. https://www.kaggle.com/kaggle-survey-2019. Accessed: 2020-03-03.

[2] Team data science process. Technical report, Microsoft, 2017.

[3] C. L. Aasheim, S. Williams, P. Rutner, and A. Gardiner. Data analytics vs. data science: A study of similarities and differences in undergraduate programs based on course descriptions. 26:14, 2015.

[4] S. Abt and H. Baier. A plea for utilising synthetic data when performing machine learning based cyber-security experiments. In *Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop*, AISec '14, p. 37–45. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2666652.2666663

[5] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[6] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):22–31, Jan 2019. doi: 10.1109/TVCG.2018.2865040

[7] C. Aragon, C. Hutto, A. Echenique, B. Fiore-Gartland, Y. Huang, J. Kim, G. Neff, W. Xing, and J. Bayer. Developing a research agenda for human-centered data science. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, CSCW '16 Companion, p. 529–535. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2818052.2855518

[8] T. K. Attwood, S. Blackford, M. D. Brazas, A. Davies, and M. V. Schneider. A global perspective on evolving bioinformatics and data science training needs. *Briefings in Bioinformatics*, 20(2):398–404, Mar 2019. doi: 10.1093/bib/bbx100

[9] A. Batch and N. Elmqvist. The interactive visualization gap in initial exploratory data analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):278–287, Jan 2018. doi: 10.1109/TVCG.2017.2743990

[10] L. Battle and J. Heer. Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in Tableau. In *Computer Graphics Forum*, vol. 38, pp. 145–159. Wiley Online Library, 2019.

[11] D. M. Blei and P. Smyth. Science and data science. *Proceedings of the National Academy of Sciences*, 114(33):8689–8692, 2017. doi: 10.1073/pnas.1702076114

[12] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011.

[13] N. L. Bragazzi, O. Guglielmi, and S. Garbarino. Sleepomics: How big data can revolutionize sleep science. *International Journal of Environmental Research and Public Health*, 16(2):291, Jan 2019. doi: 10.3390/ijerph16020291

[14] L. Cao. Data science: A comprehensive overview. *ACM Comput. Surv.*, 50(3):1–42, 2017. doi: 10.1145/3076253

[15] L. Cao. Data science: Challenges and directions. *Communications of the ACM*, 60(8):59–68, 2017. doi: 10.1145/3015456

[16] A. Chatzimparmpas, R. M. Martins, I. Jusufi, and A. Kerren. A survey of surveys on the use of visualization for interpreting machine learning models. *Information Visualization*, 19(3):207–233, 2020. doi: 10.1177/1473871620904671

[17] A. Chatzimparmpas, R. M. Martins, I. Jusufi, K. Kucher, F. Rossi, and A. Kerren. The state of the art in enhancing trust in machine learning models with the use of visualizations. *Computer Graphics Forum*, 39(3):713–756, 2020. doi: 10.1111/cgf.14034

[18] Q. Chen, X. Yue, X. Plantaz, Y. Chen, C. Shi, T.-C. Pong, and H. Qu. Viseq: Visual analytics of learning sequence in massive open online courses. *IEEE Transactions on Visualization and Computer Graphics*, 26(3):1622–1636, Mar 2020. doi: 10.1109/TVCG.2018.2872961

[19] W. S. Cleveland. Data science: An action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, 69(1):21–26, 2001.

[20] J. R. Conway, A. Lex, and N. Gehlenborg. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics*, 33(18):2938–2940, 06 2017. doi: 10.1093/bioinformatics/btx364

[21] A. Crisan, J. L. Gardy, and T. Munzner. On regulatory and organizational constraints in visualization design and evaluation. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, BELIV '16, p. 1–9. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2993901.2993911

[22] A. Crisan and T. Munzner. Uncovering data landscapes through data reconnaissance and task wrangling. In *2019 IEEE Visualization Conference (VIS)*, pp. 46–50, Oct 2019. doi: 10.1109/VISUAL.2019.8933542

[23] D. Donoho. 50 years of data science. *Journal of Computational and Graphical Statistics*, 26(4):745–766, 2017. doi: 10.1080/10618600.2017.1384734

[24] F. Emmert-Streib and M. Dehmer. Defining Data Science by a Data-Driven Quantification of the Community. *Machine Learning and Knowledge Extraction*, 1(1):235–251, Dec. 2018. doi: 10.3390/make1010015

[25] F. Emmert-Streib, S. Moutar, and M. Dehmer. The process of analyzing data is the emergent feature of data science. *Frontiers in Genetics*, 7:1–12, 2016. doi: 10.3389/fgene.2016.00012

[26] M. Feinberg. A design perspective on data. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, p. 2952–2963. ACM, May 2017. doi: 10.1145/3025453.3025837

[27] U. M. Feyyad. Data mining and knowledge discovery: making sense out of data. *IEEE Expert*, 11(5):20–25, 1996. doi: 10.1109/64.539013

[28] S. Fu, J. Zhao, W. Cui, and H. Qu. Visual analysis of mooc forums with iforum. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):201–210, Jan 2017. doi: 10.1109/TVCG.2016.2598444

[29] A. Grand, C. Wilkinson, K. Bultitude, and A. F. T. Winfield. Mapping the hinterland: Data issues in open science. *Public Understanding of Science*, 25(1):88–103, Jan 2016. doi: 10.1177/0963662514530374

[30] G. Grolemund and H. Wickham. A cognitive interpretation of data analysis: A cognitive interpretation of data analysis. *International Statistical Review*, 82(2):184–204, Aug 2014. doi: 10.1111/insr.12028

[31] H. Harris, S. P. Murphy, and M. Vaisman. *Analyzing the Analyzers: An Introspective Survey of Data Scientist and Their Work*. O'Reilley Media, Inc., Sebastopol, CA, 2nd ed., 2015.

[32] J. Heer. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6):1844–1850, Feb 2019. doi: 10.1073/pnas.1807184115

[33] T. Hey, S. Tansley, and K. Tolle. *The Fourth Paradigm: Data-Intensive Scientific Discovery*. Microsoft Research, October 2009.

[34] F. Hohman, M. Kahng, R. Pienta, and D. H. Chau. Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 2018.

[35] P. Isenberg, F. Heimerl, S. Koch, T. Isenberg, P. Xu, C. Stolper, M. Sedlmair, J. Chen, T. Möller, and J. Stasko. vispubdata.org: A Metadata Collection about IEEE Visualization (VIS) Publications. *IEEE Transactions on Visualization and Computer Graphics*, 23, 2017. doi: 10.1109/TVCG.2016.2615308

[36] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer. Enterprise data analysis and visualization: An interview study. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2917–2926, Dec 2012. doi: 10.1109/TVCG.2012.219

[37] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, AVI '12, p. 547–554. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2254556.2254659

[38] N. Khan, I. Yaqoob, I. Hashem, Z. Inayat, W. Kamaleldin, M. Alam, M. Shiraz, and A. Gani. Big data: Survey, technologies, opportunities, and challenges. *The Scientific World Journal*, 2014:18, 07 2014. doi: 10.1155/2014/712826

[39] M. Kim, T. Zimmermann, R. DeLine, and A. Begel. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, p. 96–107. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2884781.2884783

[40] M. Kim, T. Zimmermann, R. DeLine, and A. Begel. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering*, 44(11):1024–1038, Nov 2018. doi: 10.1109/TSE.2017.2754374

[41] A. Kiss and T. Szirányi. Evaluation of manually created ground truth for multi-view people localization. In *Proceedings of the International Workshop on Video and Image Ground Truth in Computer Vision Applications*, VIGTA '13. Association for Computing Machinery, New York, NY, USA,

2013. doi: 10.1145/2501105.2501106

[42] S. Kross, R. D. Peng, B. S. Caffo, I. Gooding, and J. T. Leek. The democratization of data science education. *The American Statistician*, 74(1):1–7, Jan 2020. doi: 10.1080/00031305.2019.1668849

[43] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, Jul 2012. doi: 10.1109/TVCG.2011.279

[44] D. J.-L. Lee, H. Dev, H. Hu, H. Elmeleegy, and A. Parameswaran. Avoiding drill-down fallacies with vispilot: Assisted exploration of data subsets. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, p. 186–196. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3301275.3302307

[45] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister. Upset: Visualization of intersecting sets. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1983–1992, Dec 2014. doi: 10.1109/TVCG.2014.2346248

[46] M. Liberman. Fred jelinek. *Computational Linguistics*, 36(4):595–599, Dec 2010. doi: 10.1162/coli_a_00032

[47] Y. Liu, T. Althoff, and J. Heer. Paths explored, paths omitted, paths obscured: Decision points and selective reporting in end-to-end data analysis. *ArXiv*, abs/1910.13602, 2019.

[48] M. Loorak, M. Tory, and S. Carpendale. Changecatcher: Increasing inter-author awareness for visualization development. *Computer Graphics Forum*, 37(3):51–62, Jun 2018. doi: 10.1111/cgf.13400

[49] L. Lyon, E. Mattern, A. Acker, and A. Langmead. Applying translational principles to data science curriculum development. In *iPRES*, 2015.

[50] M. A. Madaio, L. Stark, J. W. Vaughan, and H. Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20. Association for Computing Machinery, New York, NY, USA, 2020.

[51] Y. Mao, D. Wang, M. Muller, K. R. Varshney, I. Baldini, C. Dugan, and A. Mojsiloviundefined. How data scientistswork together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proc. ACM Hum.-Comput. Interact.*, 3(GROUP), Dec. 2019. doi: 10.1145/3361118

[52] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv:1802.03426 [cs, stat]*, Dec 2018. arXiv: 1802.03426.

[53] P. Mclachlan, T. Munzner, E. Koutsofios, and S. North. LiveRAC - interactive visual exploration of system management time-series data. In *In Proc. ACM Conf. Human Factors in Computing Systems (CHI*, pp. 1483–1492, 2008.

[54] V. Menger, M. Spruit, K. Hagoort, and F. Scheepers. Transitioning to a data driven mental health practice: Collaborative expert sessions for knowledge and hypothesis finding. *Computational and Mathematical Methods in Medicine*, 2016:1–11, 2016. doi: 10.1155/2016/9089321

[55] M. Meyer and J. Dykes. Criteria for rigor in visualization design study. *IEEE Transactions on Visualization and Computer Graphics*, p. 1–1, 2019. doi: 10.1109/TVCG.2019.2934539

[56] A. Milani, F. Paulovich, and I. Manssour. Visualization in the preprocessing phase: an interview study with enterprise professionals, 2019.

[57] C. Molnar. *Interpretable Machine Learning*. 2019. `https://christophm.github.io/interpretable-ml-book/`.

[58] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2227–2236, 2013.

[59] C. Moreno, R. A. Carrasco, and E. Herrera-Viedma. Data and artificial intelligence strategy: A conceptual enterprise big data cloud architecture to enable market-oriented organisations. *International Journal of Interactive Multimedia and Artificial Intelligence*, 5(6):7, 2019. doi: 10.9781/ijimai.2019.06.003

[60] M. Muller, I. Lange, D. Wang, D. Piorkowski, J. Tsay, Q. V. Liao, C. Dugan, and T. Erickson. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19. Association for Computing Machinery, New York, NY, USA, 2019. doi: 10.1145/3290605.3300356

[61] G. Neff, A. Tanweer, B. Fiore-Gartland, and L. Osburn. ç. *Big Data*, 5(2):85–97, Jun 2017. doi: 10.1089/big.2016.0050

[62] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, and et al. Promoting an open research culture. *Science*, 348(6242):1422–1425, Jun 2015. doi: 10.1126/science.aab2374

[63] V. Ortiz-Repiso, J. Greenberg, and J. Calzada-Prado. A cross-institutional analysis of data-related curricula in information science programmes: A focused look at the ischools. *Journal of Information Science*, 44(6):768–784, Dec 2018. doi: 10.1177/0165551517748149

[64] J. M. Perkel. Why jupyter is data scientists' computational notebook of choice. *Nature*, 563(7729):145–146, Nov 2018. doi: 10.1038/d41586-018-07196-1

[65] E. Pournaras. Cross-disciplinary higher education of data science – beyond the computer science student. *Data Science*, 1(1-2):101–117, Dec. 2017. doi: 10.3233/DS-170005

[66] F. Provost and T. Fawcett. Data science and its relationship to big data and data-driven decision making. *Big Data*, 1(1):51–59, Mar 2013. doi: 10.1089/big.2013.1508

[67] A. Rule, A. Tabard, and J. D. Hollan. Exploration and explanation in computational notebooks. In *ACM CHI Conference on Human Factors in Computing Systems*, number 32, p. 1–12, Apr 2018. doi: 10.1145/3173574.3173606

[68] J. saltz, R. Heckman, K. Crowston, S. You, and Y. Hegde. Helping data science students develop task modularity. 2019. doi: 10.24251/HICSS.2019.134

[69] A. Sarikaya, M. Correll, L. Bartram, M. Tory, and D. Fisher. What do we talk about when we talk about dashboards? *IEEE Transactions on Visualization and Computer Graphics*, 25(1):682–692, 2019.

[70] A. Sarvghad and M. Tory. Exploiting analysis history to support collaborative data analysis. In *Proceedings of the 41st Graphics Interface Conference*, pp. 123–130, 2015.

[71] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer. Reactive vega: A streaming dataflow architecture for declarative interactive visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):659–668, Jan 2016. doi: 10.1109/TVCG.2015.2467091

[72] M. C. Schatz. Biological data sciences in genome research. *Genome Research*, 25(10):1417–1422, Oct 2015. doi: 10.1101/gr.191684.115

[73] M. Sedlmair, M. Meyer, and T. Munzner. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Trans. Visualization and Computer Graphics (Proc. InfoVis)*, 18(12):2431–2440, 2012.

[74] E. Serrano, M. Molina, D. Manrique, and L. Baumela. Experiential learning in data science: From the dataset repository to the platform of experiences. In *Intelligent Environments*, 2017.

[75] T. Spinner, U. Schlegel, H. Schäfer, and M. El-Assady. explainer: A visual analytics framework for interactive and explainable machine learning. *IEEE Transactions on Visualization and Computer Graphics*, PP:1–1, 08 2019. doi: 10.1109/TVCG.2019.2934629

[76] Z. D. Stephens, S. Y. Lee, F. Faghri, R. H. Campbell, C. Zhai, M. J. Efron, R. Iyer, M. C. Schatz, S. Sinha, and G. E. Robinson. Big data: Astronomical or genomical? *PLOS Biology*, 13(7):e1002195, Jul 2015. doi: 10.1371/journal.pbio.1002195

[77] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relational databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):52–65, Mar 2002. doi: 10.1109/2945.981851

[78] M. Stonebraker and I. F. Ilyas. Data integration: The current status and the way forward. *IEEE Data Eng. Bull.*, 41:3–9, 2018.

[79] L. van der Maaten. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 15:3221–3245, 2014.

[80] A. Y. Wang, A. Mittal, C. Brooks, and S. Oney. How data scientists use computational notebooks for real-time collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–30, Nov 2019. doi: 10.1145/3359141

[81] D. Wang, J. D. Weisz, M. Muller, P. Ram, W. Geyer, C. Dugan, Y. Tausczik, H. Samulowitz, and A. Gray. Human-ai collaboration in data science: Exploring data scientists' perceptions of automated ai. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW), Nov. 2019. doi: 10.1145/3359313

[82] E. Wenger. *Communities of practice: learning, meaning, and identity*. Cambridge Univ. Press, 2008.

[83] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag, New York, New York, 2016.

[84] R. Wirth and J. Hipp. Crisp-dm: towards a standard process model for data mining. 2000.

[85] C. Wohlin. Guidelines for snowballing in systematic literature studies and a replication in software engineering. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*,

EASE '14. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2601248.2601268

[86] H. Wong, V. Chiang, K. Choi, and A. Loke. The need for a definition of big data for nursing science: A case study of disaster preparedness. *International Journal of Environmental Research and Public Health*, 13(10):1015, Oct 2016. doi: 10.3390/ijerph13101015

[87] K. Wongsuphasawat, Y. Liu, and J. Heer. Goals, process, and challenges of exploratory data analysis: An interview study, 2019.

[88] K. Wongsuphasawat, D. Smilkov, J. Wexler, J. Wilson, D. Mané, D. Fritz, D. Krishnan, F. B. Viégas, and M. Wattenberg. Visualizing dataflow graphs of deep learning models in tensorflow. *IEEE Trans. Visualization & Comp. Graphics (Proc. VAST)*, 2018.

[89] J. Yuan, C. Chen, W. Yang, M. Liu, J. Xia, and S. Liu. A survey of visual analytics techniques for machine learning, 2020.

[90] E. Zgraggen, Z. Zhao, R. Zeleznik, and T. Kraska. Investigating the effect of the multiple comparisons problem in visual analysis. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3173574.3174053

[91] A. X. Zhang, M. Muller, and D. Wang. How do data science workers collaborate? roles, workflows, and tools, 2020.

[92] J. Zhao, M. Glueck, P. Isenberg, F. Chevalier, and A. Khan. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):340–350, Jan 2018. doi: 10.1109/TVCG.2017.2745279