Key Insights from a Feature Discovery User Study

Ionescu, Andra; Mouw, Zeger; Aivaloglou, Efthimia; Katsifodimos, Asterios

DOI
10.1145/3665939.3665961

Publication date
2024

Document Version
Final published version

Published in
HILDA 2024 - Workshop on Human-In-the-Loop Data Analytics Co-located with SIGMOD 2024

Citation (APA)
Ionescu, A., Mouw, Z., Aivaloglou, E., & Katsifodimos, A. (2024). Key Insights from a Feature Discovery User Study. In *HILDA 2024 - Workshop on Human-In-the-Loop Data Analytics Co-located with SIGMOD 2024* (HILDA 2024 - Workshop on Human-In-the-Loop Data Analytics Co-located with SIGMOD 2024). Association for Computing Machinery (ACM). https://doi.org/10.1145/3665939.3665961

Key Insights from a Feature Discovery User Study

Andra Ionescu, Zeger Mouw, Efthimia Aivaloglou, Asterios Katsifodimos
Delft University of Technology
The Netherlands
{a.ionescu-3,e.aivaloglou,a.katsifodimos}@tudelft.nl,z.f.mouw@student.tudelft.nl

ABSTRACT

Multiple works in data management research focus on automating the processes of data augmentation and feature discovery to save users from having to perform these tasks manually. Yet, this automation often leads to a disconnect with the users, as it fails to consider the specific needs and preferences of the actual end-users of data management systems for machine learning. To explore this issue further, we conducted 19 semi-structured, think-aloud use-case studies based on a scenario in which data specialists were tasked with augmenting a base table with additional features to train a machine learning model. In this paper, we share key insights into the practices of feature discovery on tabular data performed by real-world data specialists derived from our user study. Our research uncovered differences between the user assumptions reported in the literature and the actual practices, as well as some areas where literature and real-world practices align.

ACM Reference Format:
Andra Ionescu, Zeger Mouw, Efthimia Aivaloglou, Asterios Katsifodimos. 2024. Key Insights from a Feature Discovery User Study. In *Workshop on Human-In-the-Loop Data Analytics (HILDA 24), June 14, 2024, Santiago, AA, Chile.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3665939.3665961

1 INTRODUCTION

A significant and ongoing research effort is dedicated to developing automated methods to create training datasets for machine learning (ML) applications. This process, named feature discovery, builds upon the exploration and integration steps from dataset discovery and relies on feature selection approaches to select only the most relevant features for an ML task [5, 8, 18, 29]. This process typically starts with a query table that contains a target variable. Then, through an exploratory process, relevant candidates from a data repository are augmented to improve ML model effectiveness.

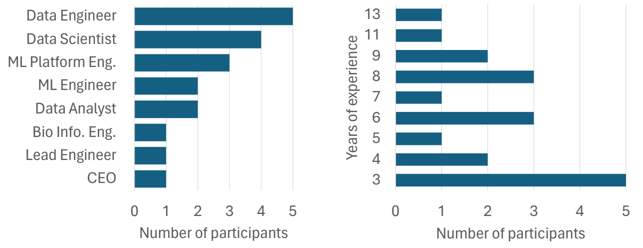
Existing research continues to operate under various assumptions regarding user workflows within the feature discovery pipeline. Some state-of-the-art works [9, 17] have incorporated user studies in their evaluation scenarios. Yet, the methods and approaches they offer are not always grounded in empirical evidence from actual user workflows. Despite this progress, a noticeable gap remains: the user perspective is lost as more automated feature discovery and augmentation approaches are developed. By understanding how users interact with and perceive the feature discovery process, we can develop more intuitive and effective methods for identifying and integrating relevant features.

We conducted a user study to understand how data professionals with hands-on experience perform feature discovery and augmentation in real life. Does the real-life process align with the theoretical one reported in the literature? To answer this question, we engaged 19 participants from various organizations and presented them with a small-scale feature discovery task. The user study allowed us to capture a nuanced understanding of how these professionals solve feature discovery challenges in real-world scenarios and compare them with the literature. This paper presents the key insights from our study in contrast with the assumptions and practises reported in state-of-the-art works as follows:

1. **The feature discovery process starts by formulating a clear hypothesis or goal.** A step often excluded from data and ML pipelines, the users spend time formulating their goal and hypothesis before any other subsequent steps.
2. **Data exploration is a collaborative and intuitive process.** It is common knowledge that users spend a significant portion of time exploring the datasets. However, this step is collaborative: users will always rely on the domain and business knowledge of data owners or use their own knowledge and intuition.
3. **Data integration process aligns with literature.** Our findings regarding the integration step are perfectly aligned with state-of-the-art literature on data integration, proving that this step in the data pipeline captures the practices and habits of users in real life.
4. **Feature selection is rarely decoupled from the ML model.** Feature selection and ML modelling are rarely treated separately. After thoroughly manipulating and engineering the dataset, users often rely on the ML model to select the best features.
5. **Data preparation is an iterative process and not a single step in the pipeline.** Data preparation is not a single step performed in the pipeline; it is rather an independent process on its own. Our participants used data preparation (e.g., data manipulation, data engineering) at every step of their pipeline.
6. **Documentation is the source of truth.** Throughout the pipeline, the subject of documenting the datasets and the process itself came recurrently. Users rely on the documentation to properly understand and manipulate the datasets.

Following, we summarise the user study design in Section 2, elaborate on each key finding in Section 3, present related work in Section 4, and conclude with Section 5.

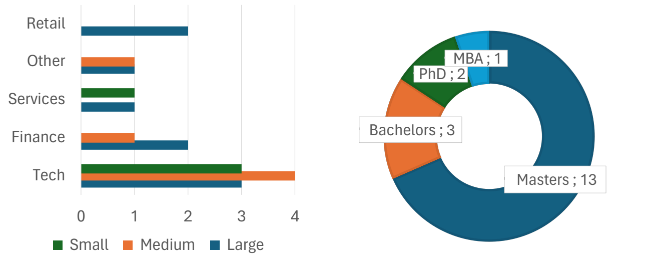
HILDA 24 , June 14, 2024, Santiago, AA, Chile
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0693-6/24/06
https://doi.org/10.1145/3665939.3665961



(a) Distribution of the roles in the company.

(b) Distribution of the years of experience with data.

Figure 1: Statistics about a) the roles of participants in the moment of the study, and b) the years of experience as reported by participants.



(a) Distribution of the industry sector and company size.

(b) Distribution of the last completed degree.

Figure 2: Statistics about a) the industry sector where participants work and the corresponding company size, and b) education as reported by the participants.

2 STUDY DESIGN

In this section, we provide preliminary information about the feature discovery process and the use case scenario in Section 2.1, and we summarise the methodology of our study in Section 2.2.

2.1 Preliminaries

Feature Discovery Process. Feature discovery is the process which discovers and augments relevant features to improve ML models [22]. Given a query table with a target variable, feature discovery retrieves candidates from a data repository focusing on the features that can improve the performance of an ML model [22]. The process contains several key steps, such as data exploration, data integration, feature selection and ML modelling [5, 10, 18].

The Use Case Scenario. The use case scenario contains 17 tables representing a small dataset of schools used for state-of-the-art feature discovery and augmentation evaluation [5, 10, 18]. The primary focus for feature discovery is the base table. The base table models a binary classification problem, where the target prediction represents the performance of each school on a standardized test based on student attributes. When applied to this base table, a decision tree ML model yields a baseline accuracy of 0.69. The objective is to enrich the base table with additional relevant features extracted from the other 16 tables so that the accuracy of the decision tree model improves.

2.2 Methodology

Participants. We recruited individuals who work with data and perform data integration tasks, such as joining tables, augmenting tables and using them for analysis. We recruited participants with diverse roles in the organisation (Figure 1a), such as data engineers, data analysts, data scientists, machine learning engineers and others with different titles at the moment of the interview but with previous experience in data engineering or analysis. We aimed to interview a balanced number of data experts per role so that potential role-specific workflow differences could be represented in the results. The participants varied in their years of experience working with data as illustrated in Figure 1b, industry sector and company size as shown in Figure 2a, and education (Figure 2b).

Interview Process. The study was structured as a think-aloud use-case scenario. The goal was to discover the real-life workflow of data practitioners. Therefore, the participants worked on their own machines, using their own set of tools, and were asked to share their screens so that we could capture information on their workflow. Participants were encouraged to think aloud during the study, explaining the motivations and expectations related to their actions. The study was conducted online and lasted between 45 and 60 minutes.

Data Processing. We analysed the data from the interviews using the thematic analysis methodological framework. In thematic analysis, determining themes (i.e., patterns) can be both theory-driven and data-driven [3]. We derived four *a priori* themes from the feature discovery pipeline: exploration, integration, feature selection, and evaluation. We finalised our set of codes *a posteriori*, extracted inductively from the interview data using the qualitative analysis tool ATLAS.ti[1]. In total, we used 60 codes to label different aspects of each step in the pipeline. We extracted a total of 1088 quotes, each labelled with one or multiple codes.

3 FINDINGS

In this section, we elaborate on the six key findings from our feature discovery user study and compare the information currently present in the literature with the information reported by the participants during the study.

3.1 Goal Setting

Our findings reveal that, contrary to the literature, the first step of the feature discovery process is not data exploration but rather defining a goal, specifying the target prediction, or consulting with clients to understand the scope of the problem. This initial goal-setting phase is commonly seen in pipelines focused on dataset search, as exemplified in Aurum [9], or in typical data science workflows [7]. However, beyond this context, goal-setting as an initial step of the pipeline is not typically observed in feature discovery or data augmentation stages. The reason for this might be either due to unintentional oversight or because of the assumption that goal-setting is an inherent part of the process [7].

[1] https://atlasti.com/

On the other hand, our participants indicated that having the goal already defined affects the subsequent steps in the process, such as data preparation. We observed that just describing the task was insufficient for the participants, who required much more information, such as information about the column names (i.e., what each column represents), in-depth details about the target variable and its meaning, and baseline accuracy.

> *"So you can, of course, do arbitrary augmentations and integrations, but at the end of the day, the real thing that matters is if the features that you have are in some way related to the thing you're trying to predict."* (P2)

3.2 Data Exploration

The literature highlights that finding relevant tables within a vast collection of datasets is an arduous and time-intensive task, as users are unaware of the relationships between the tables [5, 6, 9, 17].

In reality, users typically have a clear understanding of the location of the data and of the overall problem they need to solve with data. This is frequently achieved by examining available documentation, which acts as a road map through the data. Moreover, users actively seek the expertise of colleagues with business insight or firsthand experience with data collection, providing invaluable context. In instances where the data originates from clients, the business problems tend to be well-articulated, further guiding the users in their quest to find meaningful information. These resources empower the users and enable them to navigate through the potential overwhelm of data.

Collaboration. The collaboration between data workers and other members of organizations has been a significant focus of research. For instance, studies have shown that the scientific collaboration between biomedical scientists and data scientists can be successful when efforts are made to establish common ground and shared processing methodologies [19]. In large organizations, where data is spread across various sources, strong collaboration is often required between data workers and other organizational members, such as IT staff who assist in locating and delivering datasets [11], and business personnel who help define goals and requirements [12].

Although our study did not primarily focus on the collaboration between organizational members, we observed that data exploration is inherently a collaborative process. Our findings indicate that when faced with unclear data, data workers commonly resort to asking knowledgeable colleagues and consulting with other team members or even different departments for clarification, as noted in 6 out of 19 cases.

In our setting, we did not offer any documentation or information about the datasets. Therefore, the participants had to rely on their knowledge and intuition to understand the data.

Knowledge. Domain knowledge is key in determining which aspects of a problem are relevant and which data types can predict certain behaviours. A business context often deepens the understanding of what needs to be predicted and identifies the factors that might influence it. The team size also influences the expected knowledge someone has about their datasets (4/19).

Intuition. The participants discuss the approach of exploring a dataset they are not familiar with, emphasizing the use of common sense to navigate unknown data. They mention the importance of intuitively reasoning about the problem to identify correlations between tables or data, especially when aiming to predict specific outcomes. They resort to experimentation and intuition to form hypotheses about the significance of data (4/19). One participant even considered that relying on intuition to solve the task without any other context is irresponsible – *"It feels a little irresponsible even to push forward without having a lot more context."* (P16).

During the exploration of the datasets, automatic processes to compute the relatedness of tables were mentioned a few times (3/19). More specifically, one approach was to find schema matches by comparing the column names in the CSV files (P18). Additionally, examining the correlation between the features could help determine the relationship between the features (P2, P7).

3.3 Data Integration

We observed a blended transition between the exploration and integration steps. While exploring the data, the participants instinctively searched for columns to relate to and join the tables.

The insights from our study, particularly regarding the data integration step, resonate with the established findings in the field. Mirroring the methods outlined in COCOA [8], a few participants used automatic techniques such as computing feature correlations to determine the relevance of different tables. The participants' approach to joining all available tables or just the *"most helpful"* ones echoes the findings of [14, 17]. Furthermore, consistent with the practices documented in the literature, our participants typically work with primary key-foreign key relationships [9, 17]. Before joining tables, our participants also used data aggregation, a step that aligns with the workflow proposed in ARDA [5].

The literature on feature discovery and data augmentation techniques takes different approaches to joining plans or pipelines. Some works choose to join all tables up to a budget, then apply feature selection [5], while others have an iterative process of joining one table at a time and testing its usefulness [18]. We have observed a similar pattern among our participants.

Join All Tables. The strategy of joining all tables at once is driven by the desire for comprehensive data analysis (4/19), simplifying the initial data processing stages (6/19), and leveraging machine learning algorithms' feature selection capabilities (3/19). The methodology involves using a common key to merge tables, followed by data cleaning, aggregation, and iterative model training and enhancement (4/19). This approach allows for a thorough exploration of the data, uncovering potential insights that might not be apparent when analyzing tables in isolation (3/19) - *"If you join everything together first, then you always have the freedom to look at all of the columns in the context of each other"* (P10).

Join Tables One by One. This step-by-step process allows for careful examination of the impact of each table on the overall dataset (5/19). Most participants (4 out of the 5 discussed this process) indicated that it is particularly beneficial in complex or large datasets where an incremental approach can provide clearer insights than a bulk join. By joining tables sequentially, they can pinpoint which datasets enhance the accuracy and predictive power of the ML model (3/5), and they can also observe how each new data source contributes positively to the analysis (3/5).

3.4 Feature Selection and ML Modelling

The literature suggests that data professionals struggle with the feature selection process, as the data volume is too high [17]. In our study, we observed that while exploring the dataset, or after the feature engineering step, a few participants (6/19) discussed manually selecting specific columns relevant for the base table, and thus creating the augmented table for evaluation - *"So for this table, I'd only keep* `borough` *and* `enrollment` *[columns]."* (P18). Moreover, data professionals often have an iterative process, ensuring that each feature included contributes positively to the model's performance and overall accuracy - *"For each table I do that [i.e., cost-benefit analysis] and seeing how much it improves."* (P9).

The approach to automatic feature selection and model training is iterative and data-driven. Some participants rely on using the ML models to assess feature importance - *"Initially I would just throw a random forest or a boosted tree at it because it comes for free with feature selection."* (P1), followed by careful analysis and pruning of features. Overall, the process involves balancing automated feature selection methods, manual analysis, and continuous testing to achieve an optimal set of features for the ML model.

3.5 Data Preparation

Our observations reveal that data preparation is an iterative and integral part of the entire workflow rather than a one-time step as presented in literature [2, 4]. Data preparation is consistently revisited and initiated at any point in the pipeline throughout the various stages of the participants' workflow.

We observed participants engaging in data processing early during the exploration phase, which aids in a deeper understanding of the datasets. Data preparation is also present during the integration phase, where it serves the dual purpose of preparing the data for effective joining (e.g., data cleaning and aggregating) and tackling any issues that emerge as a result of the integration. Moreover, data preparation is linked with the feature selection process. Here, it is part of feature engineering, where the quality and relevance of features are enhanced and tailored to meet the specific requirements of the analysis – *"Ideally I also want to know what values we're dealing with and what's the appropriate way to encode the columns."* (P2). The process extends beyond feature selection to ensure that the final dataset has a high quality for the machine learning algorithms. This multi-faceted approach to data preparation highlights its significance as a dynamic, adaptive process in a data pipeline.

3.6 Documentation

The theme of documentation, or the lack of documentation, has been a recurring and significant topic throughout our interviews. Participants frequently highlighted their reliance on documentation for various tasks, such as exploration and understanding. They viewed documentation as a foundational source of truth, essential for grasping notations, definitions, data formats, and relationships between tables. The literature, however, focuses primarily on creating automated frameworks for documenting the code and computational data science notebooks [26, 27], on automating documentation to improve the reproducibility of experiments [24] without considering the importance of documenting datasets.

Our study suggests that documentation and data catalogues are crucial in understanding and effectively working with datasets. A data catalogue with detailed descriptions of tables and a dictionary or index for clarification is highly beneficial, aiding in the interpretation and utilization of data. The value of having clear, detailed documentation is underscored for both understanding the context of the data, and meeting specific requirements in data processing and presentation to the final customer (9/19). However, our participants acknowledged that lacking comprehensive and up-to-date documentation of datasets is a common issue in many organizations. This gap often hinders efficient data management and understanding, leading to data usage and interpretation challenges.

This reliance on documentation stresses its vital role in any data pipeline. Well-maintained documentation can enhance the efficiency and accuracy of data-related tasks, serving as a guide and reference for the entire data management lifecycle. Steps towards making documentation more accessible have emerged, such as automatic approaches for data versioning with explanations [25].

4 RELATED WORK

An extensive range of interview-based user studies has explored the daily workflows of data scientists [7, 11, 12, 16, 19, 21, 23, 28], specific pipeline steps such as data preparation activities [13, 20], and data exploration processes [1]. Additionally, these studies have reported on the tools and datasets used, as well as characterizations of data workers and their roles [11, 12, 15]. While these interview studies are comprehensive and provide valuable insights, they typically employ open-ended questions or retrospectively analyze existing projects within an organization. Our study, however, adopts a hands-on, practical use-case scenario approach, placing participants directly in front of a real task. This design aims to immerse the participants in their typical workflows, allowing us to capture data from their actual, hands-on work processes.

5 CONCLUSION

In this paper, we present six key findings from our user study on feature discovery on tabular data. First, we found that data professionals typically start the feature discovery and augmentation process by setting a goal, a step often overlooked in other academic studies. Secondly, data exploration emerged as a collaborative effort, whereas feature selection follows a hybrid approach as users form a mental map of key features during exploration and then rely on the ML model for subsequent feature selection.

Moreover, our findings concerning the data integration phase align perfectly with existing literature. Conversely, we observed that data preparation is not a singular step but an iterative and essential part of the entire feature discovery workflow. Additionally, documentation proved to be a critical resource for data professionals, as they heavily relied upon it at every stage of the process.

Finally, our study lays the groundwork for future research emphasising a user-centric data management approach. With this work, we want to encourage researchers to develop more intuitive and effective data management strategies that better address the real-world challenges and workflows data professionals face.

REFERENCES

[1] Sara Alspaugh, Nava Zokaei, Andrea Liu, Cindy Jin, and Marti A Hearst. 2018. Futzing and moseying: Interviews with professional data analysts on exploration practices. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 22–31.

[2] Sumon Biswas, Mohammad Wardat, and Hridesh Rajan. 2022. The art and practice of data science pipelines: A comprehensive study of data science pipelines in theory, in-the-small, and in-the-large. In *Proceedings of the 44th International Conference on Software Engineering*. 2091–2103.

[3] Erik Blair. 2015. A reflexive exploration of two qualitative data coding techniques. *Journal of Methods and Measurement in the Social Sciences* 6, 1 (2015), 14–29.

[4] Chengliang Chai, Jiayi Wang, Yuyu Luo, Zeping Niu, and Guoliang Li. 2022. Data management for machine learning: A survey. *IEEE Transactions on Knowledge and Data Engineering* 35, 5 (2022), 4646–4667.

[5] Nadiia Chepurko, Ryan Marcus, Emanuel Zgraggen, Raul Castro Fernandez, Tim Kraska, and David Karger. 2020. ARDA: automatic relational data augmentation for machine learning. *PVLDB* (2020), 1373–1387.

[6] Tianji Cong, James Gale, Jason Frantz, HV Jagadish, and Çağatay Demiralp. 2022. WarpGate: A Semantic Join Discovery System for Cloud Data Warehouse. *arXiv preprint arXiv:2212.14155* (2022).

[7] 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* 27, 2 (2020), 1860–1870.

[8] Mahdi Esmailoghli, Jorge-Arnulfo Quiané-Ruiz, and Ziawasch Abedjan. 2021. COCOA: COrrelation COefficient-Aware Data Augmentation.. In *EDBT*. 331–336.

[9] Raul Castro Fernandez, Ziawasch Abedjan, Famien Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. Aurum: A data discovery system. In *ICDE*. 1001–1012.

[10] Andra Ionescu, Kiril Vailev, Florena Buse, Rihan Hai, and Asterios Katsifodimos. 2024. AutoFeat: Transitive Feature Discovery over Join Paths. In *ICDE*. IEEE, 1861–1873.

[11] Sean Kandel, Andreas Paepcke, Joseph M Hellerstein, and Jeffrey Heer. 2012. Enterprise data analysis and visualization: An interview study. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2917–2926.

[12] Eser Kandogan, Aruna Balakrishnan, Eben M Haber, and Jeffrey S Pierce. 2014. From data to insight: work practices of analysts in the enterprise. *IEEE computer graphics and applications* 34, 5 (2014), 42–50.

[13] Stephen Kasica, Charles Berret, and Tamara Munzner. 2023. Dirty Data in the Newsroom: Comparing Data Preparation in Journalism and Data Science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[14] Aamod Khatiwada, Roee Shraga, Wolfgang Gatterbauer, and Renée J Miller. 2022. Integrating Data Lake Tables. *Proceedings of the VLDB Endowment* 16, 4 (2022), 932–945.

[15] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *Proceedings of the 38th International Conference on Software Engineering*. 96–107.

[16] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data scientists in software teams: State of the art and challenges. *IEEE Transactions on Software Engineering* 44, 11 (2017), 1024–1038.

[17] Arun Kumar, Jeffrey Naughton, Jignesh M Patel, and Xiaojin Zhu. 2016. To join or not to join? Thinking twice about joins before feature selection. In *SIGMOD*. 19–34.

[18] Jiabin Liu, Chengliang Chai, Yuyu Luo, Yin Lou, Jianhua Feng, and Nan Tang. 2022. Feature augmentation with reinforcement learning. In *ICDE*. IEEE, 3360–3372.

[19] Yaoli Mao, Dakuo Wang, Michael Muller, Kush R Varshney, Ioana Baldini, Casey Dugan, and Aleksandra Mojsilović. 2019. How data scientists work together with domain experts in scientific collaborations: To find the right answer or to ask the right question? *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–23.

[20] Alessandra Maciel Paz Milani, Fernando V Paulovich, and Isabel Harb Manssour. 2020. Visualization in the preprocessing phase: Getting insights from enterprise professionals. *Information Visualization* 19, 4 (2020), 273–287.

[21] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How data science workers work with data: Discovery, capture, curation, design, creation. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[22] Fatemeh Nargesian, Abolfazl Asudeh, and HV Jagadish. 2022. Responsible Data Integration: Next-generation Challenges. In *SIGMOD*. 2458–2464.

[23] Fahad Pervaiz, Aditya Vashistha, and Richard Anderson. 2019. Examining the challenges in development data pipeline. In *Proceedings of the 2nd ACM SIGCAS Conference on Computing and Sustainable Societies*. 13–21.

[24] Sergey Redyuk. 2019. Automated documentation of end-to-end experiments in data science. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE, 2076–2080.

[25] Roee Shraga and Renée J Miller. 2023. Explaining Dataset Changes for Semantic Data Versioning with Explain-Da-V. *Proceedings of the VLDB Endowment* 16, 6 (2023), 1587–1600.

[26] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Xuye Liu, Soya Park, Steve Oney, and Christopher Brooks. 2021. What makes a well-documented notebook? a case study of data scientists' documentation practices in kaggle. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[27] April Yi Wang, Dakuo Wang, Jaimie Drozdal, Michael Muller, Soya Park, Justin D Weisz, Xuye Liu, Lingfei Wu, and Casey Dugan. 2022. Documentation matters: Human-centered ai system to assist data science code documentation in computational notebooks. *ACM Transactions on Computer-Human Interaction* 29, 2 (2022), 1–33.

[28] Kanit Wongsuphasawat, Yang Liu, and Jeffrey Heer. 2019. Goals, process, and challenges of exploratory data analysis: An interview study. *arXiv preprint arXiv:1911.00568* (2019).

[29] Zixuan Zhao and Raul Castro Fernandez. 2022. Leva: Boosting machine learning performance with relational embedding data augmentation. In *SIGMOD*. 1504–1517.