



DataPilot: Utilizing Quality and Usage Information for Subset Selection during Visual Data Preparation

Arpit Narechania
Georgia Institute of Technology
Atlanta, USA
arpitnarechania@gatech.edu

Fan Du
Adobe Research
San Jose, USA
dufan2013@gmail.com

Atanu R Sinha
Adobe Research
Bengaluru, India
atr@adobe.com

Ryan A. Rossi
Adobe Research
San Jose, USA
ryrossi@adobe.com

Jane Hoffswell
Adobe Research
Seattle, USA
jhoffs@adobe.com

Shunan Guo
Adobe Research
San Jose, USA
sguo@adobe.com

Eunye Koh
Adobe Research
San Jose, USA
eunye@adobe.com

Shamkant B. Navathe
Georgia Institute of Technology
Atlanta, USA
sham@cc.gatech.edu

Alex Endert
Georgia Institute of Technology
Atlanta, USA
endert@gatech.edu

ABSTRACT

Selecting relevant data subsets from large, unfamiliar datasets can be difficult. We address this challenge by modeling and visualizing two kinds of auxiliary information: (1) *quality* – the validity and appropriateness of data required to perform certain analytical tasks; and (2) *usage* – the historical utilization characteristics of data across multiple users. Through a design study with 14 data workers, we integrate this information into a visual data preparation and analysis tool, DataPilot. DataPilot presents visual cues about “the good, the bad, and the ugly” aspects of data and provides graphical user interface controls as interaction affordances, guiding users to perform subset selection. Through a study with 36 participants, we investigate how DataPilot helps users navigate a large, unfamiliar tabular dataset, prepare a relevant subset, and build a visualization dashboard. We find that users selected smaller, effective subsets with higher quality and usage, and with greater success and confidence.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in visualization; Visualization systems and tools.**

KEYWORDS

data quality, data usage, subset selection, data preparation, visualization, visual data analysis, design study

ACM Reference Format:

Arpit Narechania, Fan Du, Atanu R Sinha, Ryan A. Rossi, Jane Hoffswell, Shunan Guo, Eunye Koh, Shamkant B. Navathe, and Alex Endert. 2023. DataPilot: Utilizing Quality and Usage Information for Subset Selection



This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '23, April 23–28, 2023, Hamburg, Germany
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9421-5/23/04.
<https://doi.org/10.1145/3544548.3581509>

during Visual Data Preparation. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*, April 23–28, 2023, Hamburg, Germany. ACM, New York, NY, USA, 18 pages. <https://doi.org/10.1145/3544548.3581509>

1 INTRODUCTION

Data are never truly raw [42] but still require processing through cleaning, integration, transformation, and selection before they can be utilized for their intended purposes [92]. Modern organizations often ingest all incoming data in their native form with the intent of performing analytics later [32]. The inherent information overload due to this “load-first” philosophy poses several challenges in data navigation and knowledge discovery [24, 33, 44]. For example, consider a user task, “analyze a large e-commerce dataset and build a dashboard visualizing recent geographic trends for predicting future sales.” To perform this task, users must first identify relevant data attributes pertaining to customers’ locations (e.g., “ZipCode”) and then select the desired data records by applying a temporal filter (e.g., monthly). Unfortunately, new users unfamiliar with the data may adopt “trial and error” inspection strategies [21] resulting in the selection of irrelevant, inferior attributes while missing out on important attributes, undermining the outcome of the subsequent analysis. Even experienced users may rely upon their own past usage and not explore new attributes of a new dataset, also putting the analysis outcome into question. Furthermore, users may spend more time finding relevant data than performing the analytic task at hand [33]. Thus, we ask, “How to design user interfaces that provide guidance to users to analyze large, unfamiliar datasets and select relevant and effective subsets for downstream analytics and visualization tasks such as building dashboards and customer segmentation?”

We interviewed 14 data workers from a large technology company who select data subsets (extract a smaller set of attributes and records from a larger dataset) for making dashboards (data analysts), training machine learning models (data scientists), and running digital marketing campaigns (marketers). All data workers communicated the importance of the quality of data; some of them, who relied on others for preparing these data subsets as they lacked the

necessary skill set, also reflected on the potential of surfacing other data characteristics such as their usage across users. This feedback from the data workers call for an interactive, self-service tool that facilitates data preparation with two kinds of auxiliary information: (1) quality and (2) usage. We model this auxiliary information using the data, associated meta-data, and corresponding usage logs and visually present it to users to guide them during subset selection and analysis, a task that they all perform for different purposes.

Prior art defines data quality from multiple perspectives: consumer [39], business [35, 51, 74, 87, 89], and standards-based [17, 77]. A single definition covering the different contexts is difficult [39]. Contextual to this work, we define *quality* as “the validity and appropriateness of data required to perform certain analytical tasks.” Quality is important because data are often messy, and organizations’ “load-first” philosophy often results in “big data graveyards” [105] comprising large volumes of missing, erroneous, and irrelevant information. Ideally, these data deficiencies would trigger corrective measures or even non-use; however, most organizations fail to maintain data quality standards [83] as “everyone wants to do the [ML] model work, not the data work” [99]. In this work, we model three quality dimensions [89], deemed important by the experts: (1) *completeness*: frequency of non-missing values in the data. (2) *correctness*: frequency of correct values in the data. (3) *objectivity*: extent that values conform to a target distribution.

We define *usage* as “the historical utilization characteristics of data across multiple users,” inspired by the “data utility” descriptor [101]. Users often collaborate at work [6, 26, 64, 99, 128], but much more around code than around data [65]. Understanding how data are created and shared inside an organization is underexplored [65]. We believe leveraging usage logs of current and past users, and meta-data can be one way to guide other users. Motivated by use cases from the data workers, we derive three dimensions of usage for a subset selection and dashboard building task, where data refers to attributes and records of a tabular dataset:

- (1) *in-subsets*: percentage of users that put the data in their subset.
- (2) *in-filters*: percentage of users that applied a filter on the data.
- (3) *in-visualizations*: percentage of users that visualized the data.

We integrate both quality and usage information into a visual data preparation and analysis tool, DataPilot. DataPilot facilitates preparing a subset from a large tabular dataset for building a visualization dashboard. Specifically, DataPilot computes a standardized score out of 100 for each of the quality and usage dimensions, e.g., *in-subsets* score for the “Profit” attribute is 94 out of 100. DataPilot also presents visual cues to guide users about the “good” and “bad” aspects of their data, e.g., highlighting missing and incorrect data values by coloring them in red. Lastly, DataPilot provides graphical user interface (GUI) controls as interaction affordances to assist users during subset selection, e.g., range sliders to filter out less popular data and sorting widgets to order and group data with similar characteristics together. Modern data tools [7, 25, 47, 79, 81, 107, 110, 112] provide a myriad of features such as interactive GUIs to help prepare data; however, to the best of our knowledge, no tool leverages usage information from the usage logs and associated meta-data to provide interaction affordances that facilitate interactive subset selection and analysis.

We conducted a user study with 36 participants to investigate how the DataPilot user interface guides users (nudging them one

way or another) to navigate a large and unfamiliar tabular dataset, prepare a relevant subset, and build a visualization dashboard. Our findings indicate that quality and usage information together help users to create smaller, effective subsets with greater success and confidence. We define an effective subset as one that has a higher percentage of attributes and records with high overall scores on quality and usage. Importantly, participants expressed caution about excessive reliance on usage behaviors of previous users as it can reduce exploration of quality data (pursuing novelty less), in favor of exploitation (repeating what has worked so far). Challenging convention, our findings also call for visual data analysis tools to prioritize and integrate data preparation affordances directly into analysis workflows to foster more effective use of data.

The primary contributions of this work include:

- (1) A design study with 14 data workers about tasks and challenges during data preparation and analysis that revealed the importance of data quality and the potential of surfacing additional characteristics such as usage to improve these workflows and also improve user collaboration (Section 3).
- (2) Modeling of two kinds of auxiliary information: **quality** and **usage**, by leveraging the data, associated meta-data, and usage logs of users (Section 4).
- (3) A visual data preparation and analysis tool, **DataPilot**, integrated with quality and usage information to guide users during subset selection and analysis (Section 5).
- (4) A user study with 36 participants that revealed how DataPilot helped users to select smaller, effective subsets from large, unfamiliar datasets with greater success and confidence during a subset selection and dashboard building task (Section 6).

Note that judging the true effectiveness of the selected subsets and the created dashboards depends on the end goals and other contextual circumstances, requiring expert assessment; we did not pursue this angle because our participants were not domain experts. Also, while DataPilot focuses on subset selection, additional tools and studies are needed to evaluate other downstream analytics tasks such as ranking and clustering across other applications.

2 RELATED WORK

2.1 Data Preparation

Data preparation (or pre-processing) involves analyzing the data to ensure high-quality results through collection, integration, transformation, cleaning, reduction, and discretization [129].

2.1.1 Subset Selection. Subset selection (or data reduction) involves reducing the size of the dataset [36, 56, 91]; it can be performed in two ways: feature set reduction (attributes or columns of a tabular dataset) or sample set reduction (records or rows of a tabular dataset). Feature set reduction is common when training ML models wherein users either drop irrelevant features [57] or reduce them through dimensionality reduction techniques [36]. Sample set reduction is common during market segmentation [111] wherein select groups of consumers are shortlisted to satisfy segment specific goals. These techniques have been used to combat selection bias, e.g., by visualizing how a subset compares to the original dataset [12, 44]. In this work, we support subset selection tasks by presenting data quality and usage information to users.

2.1.2 Data Quality Assessments and Tools. Real-world datasets are often “dirty” and include a variety of data quality problems [58] that speculatively cost organizations trillions of dollars [48, 95]. Data quality is crucial to ensure that systems using the data can perform the intended task in a performant, scalable, accurate, and unbiased manner [8, 19, 22, 46, 78]. Umbrich et al. [114] point out that even low meta-data quality (missing meta-data) affects both the discovery and the consumption of the datasets. However, Kandel et al. [61] revealed that practitioners consider data wrangling tedious and time-consuming. Sambasivan et al. [99] provided empirical evidence of “Data Cascades” – compounding events that cause negative, downstream effects from data quality issues. A growing body of work, thus, has been focused on understanding and improving data quality to avoid the “garbage in, garbage out” problem [49, 96].

The Data Nutrition Label [53] framework, like the Nutrition Facts label on food, highlights the “ingredients” of a dataset to help determine if the dataset is healthy for a particular statistical use case. DataPilot provides similar at-a-glance information about a dataset’s quality. Tableau Prep [107], OpenRefine [47], and Wrangler [60] are self-service data preparation tools that provide interactive affordances to explore, clean, structure, and shape the data before analysis. Most relevant to DataPilot is Profiler [62], a visual analysis tool for assessing quality issues in tabular data; Profiler applies data mining methods to automatically flag problematic data and also suggests coordinated summary visualizations for assessing the data in context, albeit without DataPilot-like usage information.

Pipino et al. [89] first presented sixteen objective and subjective dimensions for assessing data quality, that have since been extended [3, 69, 87, 101, 103, 115] as there is “no one size fits all set of metrics” [89] and also “no single dominant tool” [3]. Based on feedback from our domain expert interviewees, we model three of these dimensions (*completeness*, *free-of-error*, and *objectivity*) to guide users about potential data quality concerns. While data cleaning (e.g., imputing missing values) [18, 20, 47, 60, 72, 73, 76, 88, 94, 104, 125] is deferred to future work, DataPilot currently provides novel GUI interaction affordances to support subset selection.

2.1.3 Collaboration among Users. Prior work has examined human collaboration for information sharing and access [11, 55, 64, 65, 99, 128]. Social translucence theory [31] describes designing digital systems to support collaboration in large groups by making participants and their activities visible to one another. These collaborations have direct costs (e.g., employee salaries) and indirect costs (e.g., time delays due to user preferences and availability) [98, 108, 109], motivating efforts to mitigate inefficiencies.

In the visualization domain, collaborative systems [55, 117, 120, 123] have focused on supporting both synchronous [116] and asynchronous [50, 122] models. In the data science domain, Auto-suggest recommends data preparation steps in computational notebooks [126], albeit based on previously written code and not usage logs. Presenting users with readable, reusable code (over manual programming) for data wrangling in computational notebooks has been shown to increase user efficiency, trust, and confidence [27].

In the database domain, providing a rich set of *starter queries* from experts has been shown to empower non-experts to use SQL for data analysis with ad hoc databases [54]. In the machine learning (ML) and artificial intelligence (AI) community, Almahmoud et al. [6]

studied how different team members communicate about the quality of ML models; they found a mismatch between user-focused and model-focused notions of performance and a difficulty in understanding concerns beyond one’s role. Ehsan et al. [29] found that social transparency can potentially calibrate trust in AI, improve decision-making, facilitate organizational collective actions, and cultivate holistic explainability. This prior work motivated us to model and present usage characteristics from prior utilization of data to help users during subset selection. DataPilot’s modeling of usage information is a key novelty in addition to data quality.

2.2 Data, Analytic Provenance, and Guidance

With the proliferation of big data, more data [24] and meta-data [113] (e.g., application logs) are being stored and processed. In database contexts, “data provenance” is used to reason about the current state of a data object [127], e.g., describe its provenance characteristics (“Data Descriptors” [101]), study secure provenance schemes and associated issues (Zafar et al. [127]), and document the purpose, performance, safety, and security of data and models (“Fact-Sheets” [9]) and computational workflows (Wings/Pegasus [63]).

Provenance information has also been explored for dataset reuse. Koesten et al. [66] described a case study that determined how dataset provenance information in the form of GitHub-specific engagement metrics (e.g., the number of forks, watchers, stars, and committers) can predict a dataset’s likeliness of reuse. Facilitating data navigation and fostering reuse, many open data portals utilize and/or present provenance information [2] to users including “most-viewed” [1], “high-value” [1], and “trending” [59, 90] datasets and access to example projects and user discussion boards [58, 66].

In visualization contexts and most relevant to this work, “analytic provenance” tracks users’ interactions with a visualization system to provide an overview of their sensemaking process [86]. This information is then used for product and user behavior analytics purposes such as generating personalized content [121], mitigating biased analytic behaviors [84, 119], increasing user trust [9], recommending alternate design choices [28, 71], and visual data exploration [122]. In HCI contexts, traces of prior interactions have been applied in revisiting common regions of a page using scrollbar history [4], tracking user interactions with documents [52], facilitating groupware coordination [45], and tracking user focus while browsing a webpage using eye- [85] and mouse-tracking [10].

Characterizing provenance in visualization and data analysis, Ragan et al. [93] present an organizational framework comprising five types and six purposes of analytic provenance; our work most closely falls in the “Data” type (*the history of changes and movement of data, which can include subsetting, data merging, formatting, transformations, or execution of a simulation to ingest or generate new data*) and “Collaborative Communication” purpose (*communicating and sharing data, information, and ideas with others who are conducting the same analysis*). Ceneda et al. [16] characterize guidance in visual analytics along three degrees (orienting, directing, prescribing) that specify the extent to which guidance is required by users and provided by the system. DataPilot provides the least intrusive, “orienting” guidance through visual cues hinting at the good and bad aspects of data quality and usage.

3 DESIGN STUDY AND EXPERT INTERVIEWS

To learn about user tasks and challenges associated with subset selection as part of data preparation before analysis, we adopted a design study methodology [102] comprising semi-structured interviews and brainstorming and feedback sessions centered around iterative prototype design and development.

We interviewed 14 data workers (E_{1-14} ; 10 males; 4 females) from a large technology company to learn about their current tasks, challenges, and requirements. These experts comprised *data engineers* (3: E_{1-3}), *data analysts* (3: E_{4-6}), *data scientists* (4: E_{7-10}), *digital marketers* (3: E_{11-13}), and a *user interface engineer* (1: E_{14}), with relevant working experience ranging 3–34 years (median: 15.5; $\mu=16.86$). We recruited these experts from within our enterprise network using a combination of targeted emailing and snowballing strategies. We conducted these interviews remotely using Microsoft Teams over the course of twelve weeks; the earlier sessions were more frequent and spontaneous than the latter weekly sessions.

In total, there were eight sessions (including three follow-up sessions) that lasted about 45 minutes each, with 1–7 domain experts and 2–3 study administrators participating on the calls (sessions S_1 : E_2 , S_2 : $E_{7,8}$, S_3 : $E_{3,9}$, S_4 : $E_{11,12}$, S_5 : $E_{1,4-6,10,13,14}$, S_6 : E_2 , S_7 : $E_{4,5}$, S_8 : E_{14}). These experts had an established working relationship and were aware of each others' strengths and expertise, unlike a group meeting with complete strangers. Discussing problems and solutions from such a cohesive group was valuable for us. For instance, the program manager often relied on the engineers' opinion regarding the technical feasibility of an idea; similarly, engineers alluded to the program manager for questions around prioritization and timelines. During these sessions, one study administrator shared a PowerPoint presentation and another took notes while facilitating a conversation structured around the following questions:

- (1) “What (kind of) tasks related to data and analytics do you accomplish on a day-to-day basis?”
- (2) “What (types of) data do you work with? How do you prepare this data? What tools do you use?”
- (3) “(How) do you collaborate with other people within/outside your organization over your tasks?”
- (4) “What are some challenges that you face while working on your tasks? How do you overcome them?”

3.1 User Tasks and Challenges

We coded the domain experts' spoken quotes using inductive thematic analysis [13], categorized them based on their roles, tasks performed, challenges faced, and opinions about quality and usage. We make these available in supplemental material, albeit anonymized.

We found that tasks varied quite a bit based on the different user roles. Data analysts select key performance indicators (attributes) and subset relevant records to design interactive dashboards and prepare reports for business executives. Marketers subset customer behavior data and demographic data by devising strategic segmentation rules, e.g. a filter criteria to shortlist customers for running targeted digital marketing communications. Data scientists select a subset of relevant attributes (features) and records (observations) from existing datasets to build predictive models. Data engineers help other users (e.g., marketers) prepare their data for various analytical or operational uses; they also monitor the organization's data repositories to control their storage and cost footprints. User

interface (UI) engineers help design scalable interactive web applications for various user-facing use-cases.

We also learned many general as well as domain-specific challenges that these experts faced while performing their tasks. Everyone communicated to us that data quality is important, e.g., “Are the data complete? Correct? Unbiased? Having the correct datatypes?” and demanded that users be made more aware of quality issues with additional guidance during data preparation and analysis. In particular, E_1 (data engineer) acknowledged that skewness is an important problem, “What if all of the data came from Northern California (in USA)?” E_{11} (marketer) acknowledged from a marketing standpoint that segmenting a customer dataset by a skewed (and/or sparse) attribute can result in suboptimal targeting of communications; E_7 (data scientist) affirmed this concern from an ML standpoint.

With respect to sparseness, E_3 (data engineer) shared from experience that missing data can be, “an empty string (“”), ASCII-space only string (“ ”), null string (“null”), missing string (), which is different from [an] explicitly null [value]; depending on how the data comes in, these can [mess] up your segmentation logic.” E_2 (data engineer) raised the computational cost of “bad” filtering strategies (e.g., if the first filter minimizes the search-set by 95%; then users often undo that operation by running a new query which is expensive). Instead, E_2 suggested showing (quality) insights upfront as it may “instill feelings of curiosity and care in the user” and help catch the “bad” filter(s). Another challenge commonly faced was selecting “important/best/effective/relevant” attributes and records from a large dataset for preparing effective dashboards (E_4), training accurate and fair ML models (E_9), and defining segmentation rules for running successful marketing communications (E_{11}).

Many domain experts noted that collaboration during data preparation could be tedious, and thus advocated for better tools to support this process. E_{11} (marketer) mentioned that conveying their data-related requirements to the data engineers is often a tedious process, requiring multiple iterations that take time and critical information may also get lost during the exchange, advocating for an interactive self-service tool. E_1 (data engineer) suggested provisioning a visual report-card similar to LinkedIn's [80] profile completeness, but for datasets as that “can also give decision-makers outside of users who are executing these tasks an idea of how good their data is, prompting them to enforce data policies.” Some suggested surfacing insights based on the usage of data within the organization. E_3 (data engineer) first sighed that “even if the ingested data might be of low quality, users sometimes don't really care; they aren't actually using it. We just have to accept that.” However, E_3 also noted that, “the consumption [usage] of data can be really important here as it can provide a different kind of awareness” and defined two types of usage dimensions that may be beneficial from a business standpoint: “the ‘boolean logic’ [filter criteria] to perform customer segmentation (e.g., ‘Age’ > 18 is a common criteria to target adults) and ‘projections’ [referenced attributes], e.g., the ‘Name’ and ‘Email’ attributes are frequently utilized to target users during campaigns.” E_3 further noted that, “If I'm defining segments and referencing fields that might not be good [or] garbage, with more nulls than expected, I want to know them. So I love the usage aspect here.”

E_9 (data scientist) suggested that data provenance (e.g., “when, where, and by whom the datasets were last used”) can be used to assist with data housekeeping as “there are several unused, low-quality

datasets just lying around that may be archived.” E_4 (data analyst) suggested generalizing this idea by curating “a KPI (key performance indicator) catalog comprising different metrics, how they are calculated, how often they are used in dashboards and in segments and in journeys, and these scores would sit right along the side.” E_2 (data engineer) concluded that “it takes hard work to get data into a high quality form so any kind of re-use is a good thing, whether it is the output or the workflow used to obtain that output.” Inspired by these expert endorsements, we confirmed that guiding users about the quality and usage characteristics of data can be a promising way to help them better accomplish their tasks.

3.2 Design Requirements Exercises

After understanding the user tasks and challenges, we conducted two follow-up sessions, one each with E_2 (data engineer) and $E_{4,5}$ (data analysts), to discuss architecting DataPilot and handling data in terms of access (authentication and ethical considerations), processing (scalable computation strategies), and persistence (optimal storage mechanisms). Through a follow-up session with E_{14} (user interface engineer), we conducted design exercises wherein we sketched low-fidelity designs digitally as well as on paper and presented them for feedback. These sketches included visualizations, widgets, layouts, workflows, and interactions in the UI with an intention to catch errors that could surface later. We brainstormed on the pros and cons of each design resulting in multiple changes and refinements. For some designs, we developed rapid software prototypes with a dual purpose of exploring potential technologies (such as software libraries) and evaluating their feasibility, which further helped discard less-useful designs, refine the user tasks to be supported, and distill design goals, described next.

3.3 Design Goals

We derived **six** key design goals from our expert interviews that drove the design and development of DataPilot.

DG1. Facilitate data preparation and visual data analysis, in situ. Data preparation is a necessary step *before* analysis. However, users must often revisit the data preparation step even *during* analysis. We derived this core design goal to support both aspects within the same tool (in situ), minimizing unnecessary learning of and switching between multiple tools and windows. In particular, DataPilot supports building a visualization dashboard from a subset selected after navigating a large, unfamiliar tabular dataset.

DG2. Model data quality and usage information as standardized scores. Because non-technical marketers (E_{11}) often had to rely on data engineers (E_1), we derived this goal to model each dimension of quality and usage information into a standardized score (out of 100) using smart, heuristically determined rules. This scoring strategy will enable comparisons and aggregations across dimensions, making it a “self-service” experience for users, minimizing their (over)reliance on and tedious exchanges with other users. Power-users can still interactively specify the constraints by themselves, gaining some configuration control.

DG3. Provide visual guidance about data quality and usage while balancing user agency and control. This goal translates to providing guidance to users about their data’s quality and usage characteristics. To balance user agency and control as desired by

the domain experts, we provision the least intrusive “orienting” guidance [16], providing visual hints (e.g., highlight missing values; show the computed scores) without disrupting users’ analysis.

DG4. Provide interaction and specification affordances for data discovery, subset selection, and visualization dashboard creation. A key novelty, this goal involves providing “self-service” interaction affordances (e.g., sort and filter UI controls) to help users inspect quality and usage information of data attributes and records in the original datasets, as well as the selected subsets; and specification affordances to assign data attributes to visual encodings (e.g., X axis, Y axis) to build a visualization dashboard.

DG5. Enable control and context through configurability. E_3 (data engineer) had noted that quality and usage information may not always be available or applicable (e.g., there is no usage information yet for a newly uploaded dataset). Hence, this goal translates to providing options within the DataPilot UI to configure the visibility of different components pertaining to quality and usage (i.e., one, none, or both), enabling multiple levels of control and context across users and applications.

DG6. Design for scalability and performance. E_{14} (UI engineer) had reiterated the challenges associated with presenting large amounts of data on the UI (e.g., slow load times and sluggish interaction experience). We derived this goal to design a performant frontend application, offloading complex operations to a scalable backend server for an overall fluid user experience [30].

4 MODELING DATA QUALITY AND USAGE

Based on the design study described in Section 3, we now discuss how we modeled quality and usage information. Strategies to model quality and usage information depend on the types of data, users, and applications. In this work, we focus on a dashboard application in which users first upload a tabular dataset, prepare a relevant subset (by selecting relevant attributes and filtering out irrelevant records), and use it to create visualizations that constitute a dashboard. User-defined constraints and interactions with GUI elements (e.g., attribute-level selection checkboxes, range sliders for record-level filters) are used to model, quantify, and also interact with quality and usage information.

4.1 Quality

Based on existing challenges from our domain experts around data skewness (E_1, E_7, E_{11}), sparseness (E_1, E_2, E_3, E_{11}), and incorrectness (E_2) and prior work [89], we modeled three dimensions of quality at an attribute-level: *completeness*, *correctness*, *objectivity* and two dimensions at a record-level: *completeness*, *correctness* (DG2).

4.1.1 Attribute-level Quality Dimensions.

Completeness is the percentage of *non-missing values* among an attribute’s values, e.g., if 10 of 50 attribute values are *nulls* or *empty strings*, its completeness is $100 \cdot (50 - 10) / 50 = 80\%$. Completeness can help users detect sparse attributes that can, for example, alter how well ML algorithms can make accurate predictions.

Correctness is the percentage of *correct values* among an attribute’s values, e.g., if 5 out of 50 attribute values are incorrect, then its correctness is $100 \cdot (50 - 5) / 50 = 90\%$. To calculate correctness,

businesses can preconfigure SQL-like constraints in the DataPilot source code through relations ($>$, $<$, $=$), range (BETWEEN), pattern matching (LIKE), and membership (IN) operators; e.g., “`SELECT Count(*) WHERE email NOT LIKE ‘%_@_%._%’`” computes the number of records with incorrect email addresses. With correctness, users can assess the accuracy of individual attributes.

Objectivity is the extent that values conform to a target distribution, e.g., if the *Gender* attribute has 120 males and 45 females, then it is evidently skewed towards males and hence, from a gender equality standpoint, not objective. We utilize Wall et al.’s Attribute Distribution (AD) metric [118] for measuring the deviation between the observed and the expected objective distribution (baseline); AD scores range from [0,1] so we standardize them by multiplying by 100. With this dimension, users can detect anomalous phenomena, e.g., if the majority of applicants are of a specific gender, against expectations. Like *correctness*, businesses can preconfigure *objectivity* constraints in the DataPilot source code.

4.1.2 Record-level Quality Dimensions.

Completeness is the percentage of *non-missing* values in each dataset record, e.g., if a record has 50 values (one for each attribute), 20 of which are *nulls* or *empty strings*, then its completeness is $100 \cdot (50 - 20) / 50 = 60\%$. With this dimension, users can, e.g., discard sparse customer profiles (records) for marketing campaigns where success is determined by the profiles’ richness.

Correctness is the percentage of correct values in each record, e.g., if a record has 50 attribute values, 15 of which are incorrect (based on set constraints), its correctness is $100 \cdot (50 - 15) / 50 = 70\%$. With this dimension, marketers can discard customer profiles (records) with invalid email addresses and social media handles that are useless for running marketing campaigns.

Objectivity is inapplicable for record-level dimensions as each record comprises values from different, incomparable attributes.

4.1.3 Overall Scores: Aggregations and Customizations.

We compute a configurable heuristics-based *overall* score for each attribute and record that defaults to the arithmetic *mean* of the corresponding dimensions. Based on work by Vaziri et al. [115], users can specify different weights for different dimensions (e.g., a user might prefer an overall dimension that comprises 75% completeness and 25% correctness, and ignores objectivity) as well as different attributes and records (e.g., a digital marketer may want to weigh the “Phone” attribute more than “Email Address” for correctness).

4.2 Usage

Based on positive feedback from our domain experts, we modeled usage information (DG2) across three dimensions at an attribute-level: *in-subsets*, *in-filters*, and *in-visualizations* and one dimension at a record-level: *in-subsets*.

4.2.1 Attribute-level Usage Dimensions.

In-subsets score of an attribute is the percentage of users who selected that attribute to be in their subset for later use, e.g., if 15 out of 20 users select a feature for training an ML model, then the *in-subsets* score is $100 \cdot 15 / 20 = 75\%$. With this dimension, new users

can, e.g., perform quick and efficient analysis by selecting highly used (important?) attributes based on subsets of prior users.

In-filters score is the percentage of users who applied a filter on that attribute, e.g., by choosing a multiselect dropdown option (*Gender* = “Female”) or dragging range slider handles (*Age* $\in$ [40,50]). With this dimension, digital marketers can, e.g., determine segmentation rules (filter criteria to pick certain customer profiles) for running marketing campaigns based on previous ones. Note that *in-filters* is not a subset of *in-subsets*; users can filter (or not) by an attribute and (not) select it in their subset and vice versa.

In-visualizations score is the percentage of users who assigned that attribute to one or more visual encodings (e.g., X axis) and utilized the resultant visualization in a dashboard. With this dimension, users can refer to popular (important?) attributes from past business reports to assist with the design of present ones.

4.2.2 Record-level Usage Dimensions.

In-subsets score of a record is the percentage of users who selected that record to be in their subset (as a result of filters). With this dimension, users can, e.g., select a subset of popular (important?) records and re-run new marketing campaigns by targeting customer profiles (records) from previous successful campaigns. This dimension is in essence the same as record-level *in-filters* and *in-visualizations* usage dimensions because DataPilot treats a filtered dataset as the selected subset that is used in the visualization.

4.2.3 Overall Scores: Aggregations and Customizations.

Like overall quality, we computed a heuristics-based *overall* score for each attribute and record, but as the *maximum* of the constituent dimensions. Because attributes are seldom utilized simultaneously in subsets, filters, and visualizations, choosing *mean* would result in low scores that would be ineffective and demotivating for the user; hence, we chose *maximum*. Users can ignore one or more usage dimensions, e.g., *In-filters* usage, if it is irrelevant to their use-case.

5 DATAPILOT USER INTERFACE

To support subset selection and analysis in the same tool (DG1), we designed the DataPilot UI to have a three-step workflow with each step navigable from others via the top left corner (Figures 1 and 2). We finalized this design based on pilot studies with four users; Section 5.5 discusses some of the alternate, discarded designs.

5.1 Step 1: Review Raw Data

This step, also the landing page of DataPilot, enables users to analyze a dataset and select a relevant subset (Figure 1). It consists of the following views:

(A) **Attribute View** shows all attributes as a flattened list (☰) or as a nested list (☷), the latter being helpful for hierarchical datasets. To efficiently display a large number of attributes, we utilize the virtual scrolling principle preventing unnecessary rendering of objects not visible in the viewport (DG6). A search field allows quick attribute lookup via keyword-based queries. Users can also sort by quality and usage dimensions at the attribute-level. Each list item shows the attribute’s name (e.g., “sales.product.name”), its datatype (e.g., **A**: Categorical, **#**: Numerical), a bi-colored circular glyph (DG3),

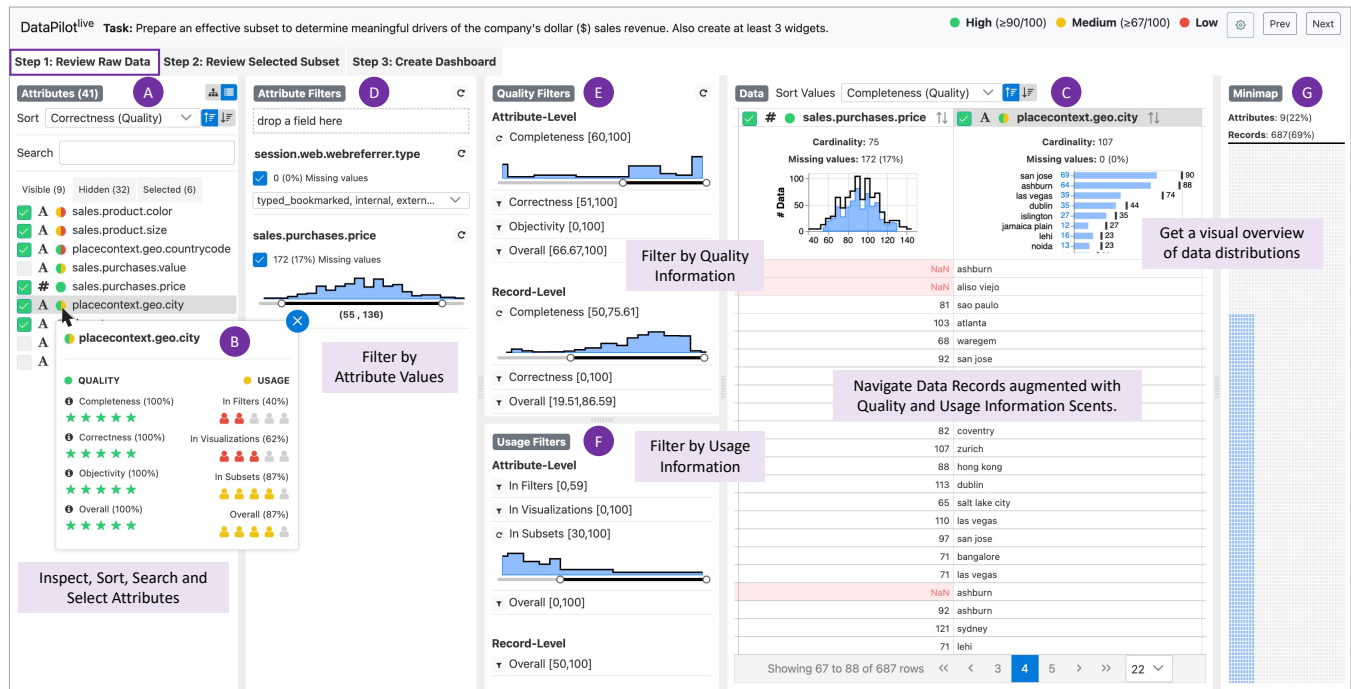


Figure 1: The DataPilot user interface showing Step 1 (Review Raw Data) of the three-step workflow. Users can inspect the list of dataset attributes (A. Attribute View), inspect quality and usage dimension scores for an attribute (B. Attribute Detail View), visualize attribute distributions and navigate dataset records (C. Data View), incrementally filter records by attribute values (D. Attribute Filter View), incrementally filter attributes and records by both quality (E. Quality Filters View) and usage dimensions (F. Usage Filters View) to reduce the search space, get a visual summary of this filtered dataset (G. Minimap View), and explicitly select attributes (A. Attribute View) and records (automatically selected based on filters) for the desired subset.

e.g., (combination of green , yellow , red colors), where the left-half shows the *overall quality* score and the right-half shows the *overall usage* score. Note that when the uploaded dataset has only either quality or usage information available, these bi-colored glyphs automatically transform into single-colored glyphs; users can also manually configure them from the settings in the top-right corner (DG5). The high (≥ 90), medium (≥ 67 but < 90), low cutoffs (that determine the three categories) and the corresponding colors (to accommodate color-related accessibility concerns), can be configured from the legend in the top-right corner. Each checkbox allows users to select ☒ or deselect ☐ attributes in the subset (DG4). Hovering on an attribute's name shows its description in a tooltip. Clicking the bi-colored glyph opens the **Attribute Detail View**.

(B) Attribute Detail View is an overlay showing details of the attribute quality and usage, like LinkedIn's [80] profile completeness (DG3). Like the bi-colored glyph, the left column shows data quality dimensions and the right shows usage dimensions along with the scores visualized on 5-point icon-array rating scales, e.g., "placecontext.geo.city" has a 100% completeness score (★★★★★) and an 87% overall usage score (👤👤👤👤👤). Hovering the info icon ⓘ shows the dimension's definition (e.g., "Completeness is the percentage (%) of non-missing values in the attribute") and any pre-configured rules for the calculation (e.g., "sales.purchases.price is considered correct if it is ≥ 0 ") to help educate the user (DG2).

(C) Data View shows the entire dataset in an interactive table. The first row shows a summary view of attribute characteristics such as cardinality (number of unique values), missing values, and distribution plots (area charts for numerical #, bar charts for categorical A attributes that show the underlying data distribution in black and the filtered data distribution in blue) (DG3). Table cells that have missing or incorrect values (e.g., "sales.purchases.price" = "NaN") are highlighted in red with details shown on hover (DG3). Standard operations such as search, pagination, and sorting are integrated within the table controls. Users can also sort by quality and usage dimensions at the record level (DG4). In Figure 1, the records are sorted by *completeness* (the "Sort Values" dropdown in the **Data View**) and the columns are sorted by *correctness* (the "Sort" dropdown in the **Attribute View**), both in the ascending order .

(D) Attribute Filter View enables users to filter the dataset by applying filters for each attribute by dragging them (from the **Attribute View** or the **Data View**) into this view's drop-zone (DG4). Multi-select dropdowns for categorical A and range-sliders for numerical # attributes along with visual scents (embedded visualizations that provide information scent cues for navigating information spaces [122]) for the distribution of attribute values in the original dataset (in black) and after applying filters (in blue) help the user determine appropriate filter criteria (DG3). Unlike selection of attributes, where one must explicitly check checkboxes



Figure 2: DataPilot Step 2 (Review Selected Subset) and Step 3 (Create Dashboard). Users review their selected attributes (H. Attribute View) and records (I. Data View), assign attributes (J. Attribute View) to encodings (K. Encodings View), inspect the resulting visualization (L. Visualization Canvas) and save it to the dashboard (M. Saved Visualizations). Users can freely navigate between the three steps.

to add to the subset, DataPilot automatically selects all remaining records after filtering into the subset.

(E) Quality Filters View enables users to filter the dataset by quality dimensions at both an attribute and a record level (DG4). For example, applying the attribute-level completeness filter $\in [60, 100]$ removes all data attributes (columns) that have a completeness score outside the range. Similarly, a record-level completeness filter $\in [50, 75.61]$ filters out all records (rows) outside that range.

(F) Usage Filters View, like the **Quality Filters View**, enables users to filter the dataset based on usage dimensions (DG4). For example, applying the attribute-level *in-subsets* usage filter $\in [30, 100]$ removes all attributes that were selected by less than 30% of users.

(G) Minimap View provides a novel, visual overview of the proportion of attributes and records originally in the dataset (gray),

currently visible after applying filters (blue), and selected in the dataset subset (green) (DG4). We disabled the green (selected) state by default as our pilot users found it to be overwhelming (Section 5.5). The width and height of the rectangular area encode the number of attributes and records, respectively. This view is discretized into small rectangles proportional to the dataset size.

5.2 Step 2: Review Selected Subset

This *review* step consists of the (H) **Attribute View** and (I) **Data View** with *just* the ☒ selected attributes and records (Figure 2). Viewing all selected attributes stacked together enables users to inspect the relative distributions of high, medium, and low quality and usage scores; this view also makes it easy to inspect the distribution of the red highlights (missing or incorrect values) in the selected table cells; both of these tasks would be difficult in

Step 1 in the presence of deselected attributes. This step makes users pause and reflect on their subset selection performance before moving onto building a dashboard (DG1).

5.3 Step 3: Create Dashboard

After reviewing the selected subset, this step helps users create and save univariate and bivariate visualizations, collectively forming a dashboard (Figure 2) (DG1). This step consists of:

(J) Attribute View is the same as the **Attribute View** in **Step 2**.

(K) Encodings View allows users to create visualizations by specifying a chart type (bar chart, scatter plot, line chart), dragging attributes onto visual encodings (X, Y), and determining aggregations (sum, mean, max, min) wherever applicable (DG4).

(L) Visualization Canvas renders the visualization based on the specifications configured in the **Encodings View**. Users can save a visualization by giving it a title and clicking the save icon .

(M) Saved Visualizations View shows the list of all visualizations saved from the **Visualization Canvas**. This view also allows users to delete  one or all saved visualizations as needed (DG4).

5.4 Implementation

We developed the DataPilot frontend in Angular [43], which interfaces with a Python [37] server in real-time over the HTTP REST [97] and websocket [34] protocols. The datasets, user interaction logs (collected from the frontend), and auxiliary information were all stored in PostgreSQL, and queried later using SQL (DG6).

5.5 Design Alternatives

Before finalizing the design of DataPilot, we presented an initial version of the interface to *four* pilot users to assess the feasibility of certain designs as well as the fidelity of the evaluation task planned for the user study (Section 6). Some of our design considerations that did not make it to the current version are described next.

Before fixating on the bi-colored glyphs next to the attribute names, we experimented with other visual variables such as size (e.g., a larger circle means higher score) and shape (e.g., quality is square and usage is a circle). We did not choose these alternatives in order to satisfy DG5 (configure DataPilot to support one, none, or both of quality and usage information); the bi-colored glyphs were more aesthetic as they retained a consistent circular shape while using different colors to describe different dimensions across configurations. Next, we picked a discrete three-class (high, medium, low) scale over a continuous scale to help users perceptually distinguish between (and form groups of) attributes by color hue instead of the less effective saturation [67]. For the five-class rating scales in the *Attribute Details View*, we considered a progress bar-like continuous widget that encodes the size (length), but eventually chose discrete icon arrays as they are easy to read [41]. For selecting records into the subset, we considered if they should, like attributes, be selected manually through checkboxes; however, this one-by-one selection was deemed tedious and was hence discarded. Finally, to facilitate data preparation along with analysis, we had several workflow-related considerations, e.g., How many steps should we have? Should they be linear? Is the review step necessary? Our pilot users helped us finalize the flexible, linear, three-step workflow.

5.6 DataPilot Example Scenarios

To illustrate how DataPilot can help users prepare relevant subsets from large, unfamiliar datasets, we developed two usage scenarios about two hypothetical users - Sunny (data engineer) and Kiran (data analyst); these scenarios were developed in collaboration with the domain experts to ensure domain relevance (Section 3).

5.7 Case 1: Expert User, Improved Performance

Sunny, an experienced data engineer, often prepares data subsets for analysts who then prepare business reports. They open DataPilot, upload a recent batch of customer transactions data for an e-commerce app, and begin analysis. Given their domain expertise, they quickly lookup known attributes via the search field and select *five* attributes for their subset: “sales.product.name” , “sales.purchase.price”  (in USD), “timestamp”  (of purchase), “placecontext.geo.countrycode”  (e.g., ‘IN’ for India), and “environment.operatingsystem”  (e.g., ‘iOS’).

They switch to **Step 2: Review Selected Subset** where they observe several cells in the data table (which now only shows the five selected attributes) with a red background. In particular, the “placecontext.geo.countrycode”  column is highlighting cells with the value “AA” () and the “environment.operatingsystem”  column is highlighting cells with blank (missing) values (). Realizing no country has “AA” as their code (as per DataPilot’s *correctness* constraint and from their own knowledge) and that a majority (706 out of 1000) of values for operating system are missing, they go back to **Step 1: Review Raw Data** to make amends.

They drag the “placecontext.geo.countrycode”  attribute from the **Attribute View** into the **Filter Panel** to remove all records with “AA” values () and separately alert the data collection team about this issue. To absolutely ensure that their data are correct across all attributes, they apply a record-level “Correctness” filter () to only keep 100% correct records. Finally, they deselect “environment.operatingsystem”  from the subset and instead select another attribute “environment.browserdetails.useragent”  that has similar information, e.g., ‘Mozilla/5.0 (iPhone; CPU OS 12_0 like Mac OS X; en_US)’ and although it has not been used often before (right half is red), it is of high overall quality (left half is green). In this way, throughout their working session, DataPilot helped Sunny become aware of issues with their data, guiding them to prepare a more complete and correct subset.

5.8 Case 2: New User, Effective Onboarding

Kiran recently joined a data analytics company and is tasked with becoming familiar with a client’s data for designing future dashboards. They upload a client dataset of e-commerce transactions into DataPilot and start analyzing. The dataset is large and unfamiliar. They start inspecting the attribute names and descriptions from the **Attribute View** and the corresponding values and distribution plots in the **Data View** (). Overwhelmed by the sheer size of the data and wanting to speed up their onboarding, they modify their strategy to only target *important* attributes.

They try to reduce the attribute search space by applying attribute-level filters in the **Quality Filters View** and **Usage Filters View** as proxies for *importance*. Specifically, they inspect the

distributions over the respective range sliders and filter out attributes with an *overall* quality score < 75 (📊) and an *overall* usage score < 25 (📊), reducing the number of attributes to a manageable 17. Finally, they sort these attributes by *overall* quality score in the descending order (Sort Overall Quality 📊) and start inspecting their name, description, and ● Completeness (80%) and ● Correctness (100%) scores in the **Attribute Detail View** (via the bi-colored circular glyphs 🟡🟢). In this way, DataPilot helped Kiran get onboarded to a new, unfamiliar dataset quickly and effectively.

6 EVALUATION

We conducted a user study to evaluate and understand how the quality and usage information in DataPilot guides users in preparing effective data subsets for subsequent analysis.

Task: We designed a task involving subset selection and visual analysis wherein participants are expected to:

“Explore a dataset of online customer behavior on an e-commerce website, prepare an effective subset¹ to determine meaningful drivers of \$ (dollar) sales revenue for the company, and create a dashboard of at least three visualizations to convey their findings.”

Participants: We recruited 36 participants consisting of professionals and researchers from industry and academia: *students* (23), *business consultants* (2), *senior data analysts* (2), *assistant professor*, *associate product manager*, *data science manager*, *postdoctoral scholar*, *program manager*, *quality assurance engineer*, *scientist (clinical trials)*, *software developer*, and *UX designer*. Participants were pursuing or had received *bachelors* (3), *masters* (14), or *doctoral* (19) degrees in *computer science* (21), *human-centered computing* (4), *human-computer interaction* (2), *business administration* (3), *pharmaceutical sciences*, *economics*, *electronics engineering*, *systems engineering*, *data science*, or *information studies*. Demographically, they were in the *18-24* (13), *25-34* (19), *35-44* (3), or *preferred not to say* (1) age groups (in years) and of *female* (16), *male* (19), *other* (0), or *preferred not to say* (1) genders. They self-reported their experience performing any kind of data analysis using visual analysis tools (e.g., Excel, Tableau) or programming as either *everyday or part of the job* (10), *often* (13), *occasionally* (13), *rarely* (0), or *never* (0).

Dataset: For the purpose of a thorough evaluation of all DataPilot capabilities and to ensure completion of the task within the stipulated study duration, we used a random sample of 1000 records and 42 attributes (columns) from an open-source digital marketing dataset [68] and infused certain quality issues pertaining to *correctness* and *objectivity* (by setting appropriate constraints). We marked quality and usage (and overall) scores such that ≥ 90 is marked as *high* 🟢, ≥ 67 but < 90 as *medium* 🟡, and the rest as *low* 🔴. We fixed these thresholds to realize a reasonable distribution of attributes and records across the three (high, medium, low) categories, so that participants are neither demoralized (all scores are low) nor overconfident (all scores are high).

System Configurations as User Study Conditions: To achieve **DG5**, we designed DataPilot to support four configurations: (1) neither quality nor usage, (2) only quality, (3) only usage, and (4) both quality and usage. Of these four configurations, we did not explicitly evaluate the (3) only usage configuration because our expert interviews highlighted addressing data quality concerns as most important and that usage information alone must never power “data-driven” analysis and decision-making, at least not without more important aspects such as quality. Hence, we utilized the other three DataPilot configurations as standalone study conditions in a between-subjects evaluation, described next.

[B] Baseline: With this configuration, we aim to understand user strategies *without* quality and usage information, also simulating what many current systems do (e.g., Tableau [106]). Specifically, the bi-colored glyphs next to the attribute name, filter and sort options, and visual scents (in the table) for usage and quality are all hidden.

[Q] Quality: With this configuration, we aim to understand how users utilize only quality information to perform the study task, also simulating what many current systems do (e.g., Profiler [62], Trifacta [110]). This condition would also enable us to compare against the following *D* configuration (that has both quality and usage information). Specifically, only single-colored circular glyphs next to the attribute name, sort and filter options, and visual scents (in the data table) that are relevant to quality are visible and enabled.

[D] DataPilot: This all encompassing configuration shows both data quality and usage information in the interface. Specifically, all features described in Section 5 are enabled. Usage information for the *D* condition were computed by processing the interaction logs of the participants in the *B* and *Q* conditions (24 participants). We computed each attribute’s *in-subsets* score as the percentage of participants who selected that attribute to be in their subsets, *in-filters* score as the percentage of participants who filtered by that attribute, *in-visualizations* score as the percentage of participants who assigned that attribute to a visual encoding, and an *overall* score as the maximum of the three aforementioned scores. Similarly, for each record, we computed the *in-subsets* score (also the *overall* score in this case) by computing the percentage of participants who selected that record (automatically as a result of applied filters) to be in their subsets. To disregard temporary, unplanned, and accidental selections during analysis, we compute this information only based on the final state of the interface at the end of the task (selected subset, applied filters, saved visualizations).

Study Session: We assigned participants to one of the three study conditions (*B*, *Q*, *D*) while trying to balance for their backgrounds, demographics and visual data analysis literacies. Each study session lasted between 60 and 90 minutes, with *D* taking longer than *Q* than *B* due to differences in participants’ training and practice times. We compensated each participant with a \$15 gift card for their time. We conducted the study remotely using Microsoft Teams [82]; the experimenter provided participants access to the study environment by sharing their (experimenter’s) computer screen and granting input control to the participant. After providing consent, participants saw a video tutorial (*B*:5, *Q*:7, *D*:10 minutes long) that demonstrated

¹Note that a data subset comprises attributes and records less than or equal to those in the original presented dataset.

the features of DataPilot². Participants then performed a practice task on a *dataset of houses* (adapted from [23]) to get acquainted with the UI before starting the actual task.

The actual task lasted a maximum duration of 30 minutes. Participants were not required to think aloud during the task to simulate a realistic work setting (although some participants felt comfortable doing so). During the task, participants' interactions with the system (e.g., the filters they applied, the data subsets they selected) were logged. The study ended with participants completing a questionnaire to rate the usefulness of DataPilot's features and a semi-structured debriefing interview for 10 minutes in which participants reflected on their overall experience, provided feedback, and answered other questions. At the end of the debriefing interview, the experimenter also demonstrated the *D* configuration to both *B* and *Q* participants to get their initial reactions and elicit feedback on how the new set of aids would have hypothetically helped them accomplish their task differently. Each debriefing interview was screen- and audio-recorded for subsequent qualitative analysis.

6.1 Hypotheses

We structure our study analysis according to the hypotheses below, predetermined before the study based on our expectations from the intended purpose of the tool, former perception studies, feedback from pilot studies, and our own instincts. > implies *more or greater than*; < implies *less or smaller than*.

- H1** *B (Baseline)* > *Q (Quality)* > *D (DataPilot)* in terms of the number of attributes and records in the selected subsets.
- H2** *B* > *Q* > *D* in terms of the proportion of attributes and records with *low* quality and usage in the selected subsets.
- H3** *B* < *Q* < *D* in terms of the proportion of attributes and records with *high* quality and usage in the selected subsets.
- H4** *B* < *Q* < *D* in terms of success and confidence after the task.
- H5** *B* < *Q* < *D* in terms of amount of effort, temporal demand, mental demand, and frustration while doing the task.
- H6** Participants will find quality information to have greater utility than usage information while doing the task.

6.2 Results

Below, we present findings from the user study and discuss them in the context of qualitative participant feedback. *B*_{1,...,12}, *Q*_{1,...,12}, *D*_{1,...,12} refer to the 36 participants in the *Baseline* (*B*), *Quality* (*Q*), and *DataPilot* (*D*) conditions, respectively. Participant quotes spoken during the debriefing interview and responses written in the questionnaires were both coded and categorized using affinity diagramming [40], an inductive thematic analysis [13] technique. One experimenter came up with an initial set of categories that were then refined during iterations with three other experimenters until a consensus was reached; the final codebook consisted of 6 high-level categories with 43 detailed, low-level codes. Relevant study material consisting of the users' interaction logs, questionnaires, interview transcripts, assigned qualitative codes, data analysis scripts, and relevant figures with RainCloudPlots [5] (instead of box plots) are made available in the supplemental material.

²*D* participants saw both quality and usage; *Q* only saw quality; and *C* saw neither; hence the difference in the duration of the respective video tutorials.

6.2.1 Feedback on DataPilot's Quality and Usage Information.

DataPilot, the system. Overall, participants found DataPilot to be useful, reporting above average system usability (SUS [15]) scores across the three conditions as {*B*: 80.21, *Q*: 74.17, *D*: 71.67}. *D*₄ commented that “*Providing detailed auxiliary information such as the quality and usage of each data attribute is very important and missing in current tools like Tableau and PowerBI.*” *Q*₈ also explained why quality and usage information are important noting, “*80-90% of true data analysis, data science, machine learning is [the data preparation] step. These [quality and usage] measurements that you're creating to allow users to start [working on their tasks] and make them explore some of the unintended consequences is very powerful. It has ample opportunity for future discovery to continuously make this a better product, so very very fascinating stuff.*”

Quality information. Participants had overall positive feedback for the quality information. *Q*₁₀ commented that “*There are invisible problems with your data and you don't necessarily find out until you start playing around with the visualizations. [Furthermore,] in aggregate visualizations, you either have limited or no ability to identify quality problems so I appreciate that DataPilot is just very explicit about these quality issues.*” *Q*₇ noted that “*It is important for systems to provide such out-of-the-box insights so that users like me who don't write code don't completely ignore these aspects and can rely on the green attributes and just get started with analysis.*” *Q*₈ saw “*a lot of value to enable users to more quickly filter [attributes and records] through the signal of these measurements of quality as opposed to learning [them] on their own.*” However, *Q*₈ also expressed caution about “*confounding factors, especially missing data, because many times data is not missing at random it is actually missing and telling a story,*” suggesting quality information can provide a good starting place but additional analysis by users may still be required.

Usage information. There was mixed feedback regarding the usage information. Participants with positive feedback suggested using usage information to perform fast and efficient analysis (*Q*₆), to seek validation “*by performing little investigations*” (*Q*₁), “*to check if they have a similar opinion as others*” (*D*₄), “*to identify new things where other people are not looking*” (*D*₃), to seek guidance from predecessors (e.g., *Q*_{2,11}), to avoid repeating past mistakes (*D*₃), and to choose between conflicting choices (e.g., “*for some attributes it's not easy to decide...but usage can help choose*” - *D*₈). Participants with mixed and negative feedback said they would not care (*D*₃) or rely on what other people did as they do not know anything about the other users and would have to assume they did a great job with their analysis (*Q*₁, *D*₁₀). Participants also raised concerns around bias and following the crowd as “*one might miss out on an uncommon attribute that is also useful*” (*B*₇).

6.2.2 Comparing Prepared Subsets. Table 1 and Figure 3 show the sizes of subsets (total number of attributes out of 42 and records out of 1000) selected by the participants (Figure 3a) and the distribution of *high*, *medium*, *low* values of attribute- and record-level quality and usage information (Figure 3b). Validating **H1**, *D* chose the fewest attributes and records followed by *Q* followed by *B*.

Furthermore, *D* chose a higher percentage of *high* overall quality attributes than *Q* than *B*. Because the dataset was sparse (a majority of values in each record were empty), no record had a *high* overall

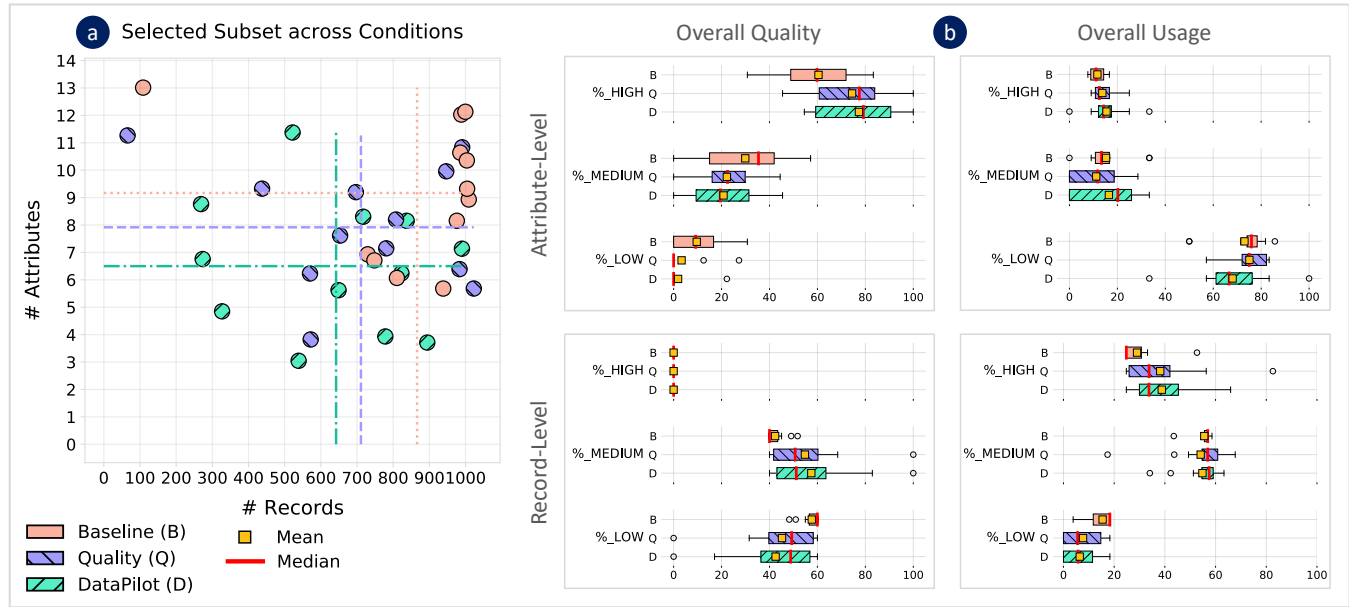


Figure 3: (a) Number of attributes and records in the participants’ selected subsets and (b) attribute-level and record-level distributions of high, medium, low overall scores for both quality and usage across the three study conditions (B, Q, D).

Table 1: Statistics associated with the prepared dataset subsets in terms of their “Size” and distribution of *high* (“% H”), *medium* (“% M”), *low* (“% L”) values for attribute- (“A”) and record-level (“R”) quality and usage scores across the three study conditions (B, Q, D). The bolded and highlighted values in each row support our hypothesis, specifically H1, H2, H3, e.g., **6.5 (D)** has the smallest μ of number (“Size”) of attributes (“A”) selected in the subset, supporting H1. No record (“R”) had a high (“% H”) overall quality score because the chosen dataset was sparse. In addition, medium (“% M”) values were not part of our hypotheses; thus, the table cells corresponding to these values are neither highlighted nor formatted.

		Baseline (B)		Quality (Q)		DataPilot (D)	
		μ	σ	μ	σ	μ	σ
Size of Prepared (Selected Subsets)							
Size	A	9.17	2.44	7.92	2.19	6.5	2.32
	R	866.17	253.08	710.83	282.85	642.17	249.18
Distribution of Overall Quality Scores							
% H	A	60.45	16.63	74.47	18	77.32	17.73
	R	0	0	0	0	0	0
% M	A	29.90	20.43	22.22	13.94	20.83	16.36
	R	42.36	4.12	54.78	17.09	57.45	18.7
% L	A	9.65	10.32	3.31	8.36	1.85	6.42
	R	57.64	4.12	45.22	17.09	42.55	18.70
Distribution of Overall Usage Scores							
% H	A	11.67	33.20	13.72	4.54	15.45	8.27
	R	29.03	8.14	38.16	16.89	38.78	12.97
% M	A	15.29	9.53	11.18	9.81	16.46	12.95
	R	55.54	3.90	54.11	13.16	54.82	8.59
% L	A	73.04	11.37	75.09	7.87	68.08	16.45
	R	57.64	4.12	45.22	17.09	42.55	18.70

quality score, hence the corresponding μ_R , σ_R values for B, Q, D were all 0. D also chose a higher percentage of *high overall usage* attributes and records than Q than B. These results validate H3.

Similarly, D chose a lower percentage of *low overall quality* attributes and records than Q than B. Furthermore, D chose a lower percentage of *low overall usage* attributes and records than Q and B, validating H2. These findings suggest that quality and usage information nudged users to prepare smaller, more effective subsets.

6.2.3 Task Fidelity Scores. Figure 4 shows participant feedback on the fidelity of the task on a seven-point *Disagree* (1) to *Agree* (7) scale. D reported higher or comparable mental demand ($M_D=5$; $M_Q=5$; $M_B=4.5$; M=median), hard work ($M_D=5$; $M_Q=4$; $M_B=4$), and frustration ($M_D=2.5$; $M_Q=2.5$; $M_B=2$) than Q than B, finding some evidence in support of H5. We attribute this result to the increased complexity due to additional user interface elements in D, that may have affected users’ cognitive load. However, D reported greater success ($M_D=6$; $M_Q=5.5$; $M_B=5$) and confidence ($M_D=5.5$; $M_Q=5$; $M_B=4.5$) in the end, validating H4 and suggesting that the auxiliary information helped participants perform the task more effectively.

6.2.4 Importance of General, Quality, and Usage Information. We asked participants about the importance of different kinds of general, quality, and usage information in the interface on a *Not at all important* (1) to *Very important* (7) scale. Except attribute *datatypes*, other general information such as attribute *names*, *values*, *distributions*, *cardinalities*, and *descriptions* were mostly useful (Figure 5a).

Figures 5b, 5c show that overall, both Q and D participants found quality information to be useful ($M_D=5$; $M_Q=5$; M=median). At the attribute-level, *completeness* ($M_D=6$; $M_Q=6$) was more important than *correctness* ($M_D=5$; $M_Q=6$) and overall ($M_D=5$; $M_Q=5$), while *objectivity* ($M_D=3.5$; $M_Q=4.5$) received mixed scores. Many participants felt completeness was the most important (Q_5 , $D_{3,6,9,10}$).

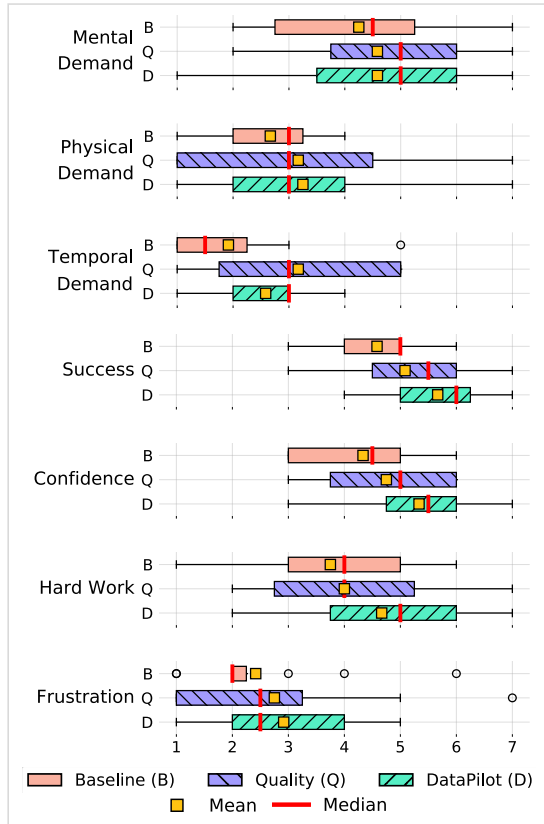


Figure 4: Assessment of the fidelity of the study task as reported by participants on a seven-point Disagree (1) to Agree (7) scale. D participants reported higher or comparable mental demand, hard work, and frustration but greater success and confidence at the end of the task than Q than B.

because “[they were] not the one who set the rules for correctness and objectivity” (D_6). Scores were mixed for the record-level dimensions: overall ($M_D=4$; $M_Q=4$), correctness ($M_D=4$; $M_Q=3.5$), and completeness ($M_D=4$; $M_Q=4$). Q_4 tried to make their subset as authentic as possible with mostly complete records but Q_7 did not as they felt it would be counterproductive after applying attribute-level filters. B participants, when presented with quality information during the debriefing, stated that they either assumed there were no missing values ($B_{2,10}$), forgot to look for them and vowed to be more alert next time (B_9), or thought of but ignored them ($B_{4,7}$).

Figures 5b, 5d show that overall, D participants had mixed feedback about the usage information ($M_D=5$; $M=$ median). $D_{2,7,8}$ found them useful, D_1 not so much, and $D_{3,4,5,10}$ raised concerns about bias and loss of originality, suggesting usage be provided with care in specific situations. At the attribute-level, overall ($M_D=5$) was more important than in-subsets ($M_D=4$), in-visualizations ($M_D=3.5$), and in-filters ($M_D=3$). Most participants also stated overall to be the most important dimension except D_6 who “went for the highest [usage] in filters.” Participants found the record-level dimensions less useful (in-subsets: $M_D=3$). Q and B participants, when they were presented simulated usage information during the debriefing interview reflected that usage can “give [them] more confidence in

selecting attributes” (Q_4), help verify their work (Q_1), and be guided by others’ work (B_8 , Q_2). Overall, participants found quality to be more important than usage, as noted by D_4 , “Data quality is way more important in our daily life and only if there are several people working on the same dataset or tool, then data usage may be helpful” and Q_{12} , “If an attribute is of high quality but low usage, I would still pick that attribute.” Collectively, these results validate H6.

7 DISCUSSION

7.1 Participant strategies to select subsets.

Only quality. Ten Q and two D participants relied only on quality: Q_4 discarded incomplete records by applying a completeness filter, $D_{1,5}$ filtered out attributes based on completeness, and Q_3 looked for high overall quality attributes via the colored glyphs.

Only usage. No D participant relied only on usage, vindicating our domain experts’ judgment that quality is still the most critical information during data-driven preparation and analysis.

Both quality and usage. Seven out of twelve D participants used both quality and usage. For example, D_9 applied quality filters and then focused on the bi-colored glyphs to avoid the low usage attributes. D_8 sorted attributes by overall usage scores before applying quality filters, D_{11} inspected the in-subsets usage dimension after applying quality filters, and $D_{4,6}$ used quality to make initial selections and then usage to verify and validate.

Neither quality nor usage. All B (as they did not see any auxiliary information), two Q ($Q_{1,2}$), and three D participants ($D_{2,3,10}$) primarily relied on general attribute information (e.g., attribute names and descriptions) and correlation and trend analysis (e.g., by creating visualizations) to select their subsets.

Other non data-driven strategies. Participants also relied on their preconceptions (Q_3 , D_4), common sense (D_1), intuition ($D_{2,3,5,7}$), and trial and error practices ($D_{3,6}$) as secondary strategies, highlighting the role of human-intelligence in data-driven analysis. Modeling auxiliary information such as quality, usage can minimize uncertainties and inconsistencies associated with such strategies.

7.2 Reflections on the three-step workflow.

We designed DataPilot to facilitate a three-step workflow: (1) **Review Raw Data**, (2) **Review Selected Subset**, and (3) **Create Dashboard**, that forces the user to first select attributes and records of interest before creating visualizations. This approach deviates from many visual data analysis workflows wherein either there are no steps and no means to (de)select attributes (e.g., Voyager [124], Lumos [84]) or all attributes are selected by default and users can only hide irrelevant ones (e.g., Tableau [106]). In Power BI [81], users are first presented with a separate “Query Editor” to transform data before analysis; however, because data preparation is an iterative process, users can utilize the “Transform Data” feature to open the “Query Editor” window at any time during analysis. Tableau Prep [107] on the other hand, is a separate tool that provides data preparation affordances before use in Tableau [106]; Tableau Prep does, however, have the “Open sample in Tableau Desktop” feature for users to test how a sample of the data currently under preparation would appear during the eventual analysis in Tableau.

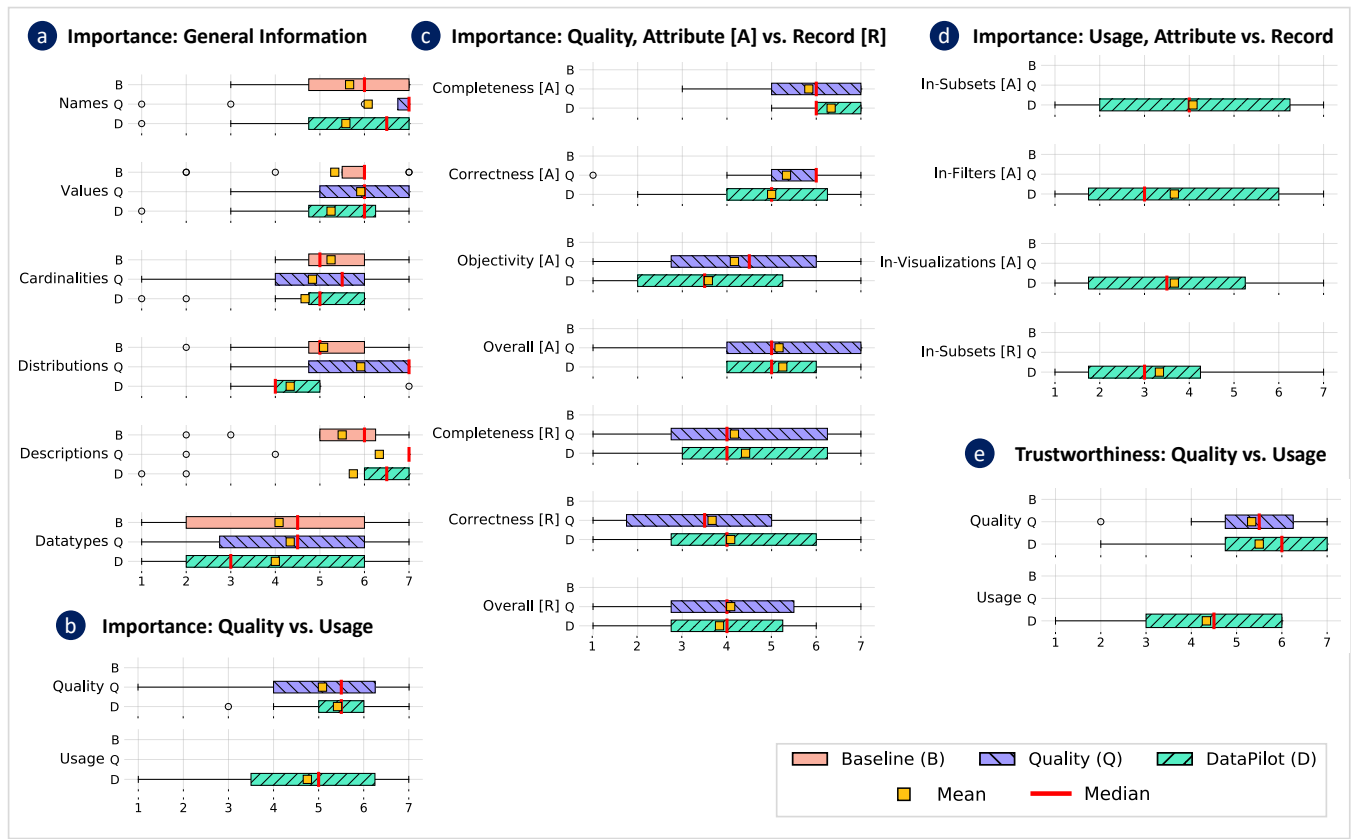


Figure 5: Importance and trustworthiness scores of general, quality and usage information for attributes and records across the three study conditions. There are no box plots for some study conditions, e.g., Baseline (B) in (b)-(e), as they were not applicable.

Regarding DataPilot’s flexible, three-step workflow, participants found it useful as “it made [them] think about what is important, whereas in Tableau, one imports the dataset and then immediately goes on to the chart making step, dragging and dropping attributes hoping to find something interesting” (B₁). D₁₁ commented, “I think **Step 1** is the most important step for me in creation of the dataset. I know that charts are very important but they are appropriately put at the third step otherwise it would get overwhelming while having all the attributes.” For some participants, the workflow helped them focus on individual aspects of data (D₁₀), was time saving because they could quickly identify if the attributes that sounded important and interesting were not worth looking at (Q₁₀), and prevented junk data from reaching the chart creation step (D₇). B₃ used the *Review Raw Data* step as more exploratory and found it convenient “to move back and forth between the steps to remove certain attributes that [they] don’t need” and liked the *Review Selected Subset* step as “they get to see just their smaller, cleaner subset of data.” However some others requested support for “creating charts using all attributes” (Q₁₀) so that “[they] don’t have to tab back and forth” (D₁₀).

7.3 Trust, bias, convergence, ethics concerns.

We asked Q and D participants to rate the trustworthiness of the auxiliary information they interacted with. As shown in Figure 5e, both Q and D participants found quality ($M_Q=5.5$; $M_D=6$) to be more

trustworthy than usage ($M_D=4.5$). Whereas D₁₂ simply “trusted the overall [quality] score,” some others exhibited hesitation in trusting the quality scores, referencing the preconfigured constraints for the correctness and objectivity dimensions and the lack of clarity around how these were defined; some participants stated that “[they] don’t trust [their] manager or the settings they’ve made” (D₃) as “they may not be doing it in a reliable or an unbiased way” (Q₅).

Participants similarly expressed a lack of trust for usage, particularly about the behavior or decisions of other people since they do not know them (D₁₀), their experience (Q₉), their expertise (Q_{1,9}, D₁), or their tasks (Q₅, D_{4,10,12}). D₁₀ noted, “I’ll trust [usage information] if I know the ten people working with me and everyone is doing this [same] task.” D₁ commented, “I don’t want to depend on previous people’s understanding of the system.” D₁₂ wondered if the people who created these data had different objectives, hence “[they] might be using data to get totally different insights as compared to another team, so [they weren’t] comfortable trusting usage.”

In addition to discussing trust, participants reported usage insights as a potential source of bias. Q₇ commented that “[usage information] would definitely bias the perspective of people who are doing data exploration themselves...it is not necessarily expert exploration but leaning towards exploitation.” Q₅ explained that “If I have to see how other people are using the dataset, then that would bias me.” Q₇ referenced convergence theory [75], “If I select an attribute

and notice that not many people have selected it in the past, then I will be compelled to deselect it. On the other hand, if previous users did not select an otherwise relevant and important attribute, then it will likely stay lowly consumed due to the convergence effect.”

Privacy was another concern reported by both participants and our domain experts. Applicability and availability aside, participants questioned if it is ethical to extract usage data and share it publicly, even if it is anonymous and aggregated. They also raised concerns around users unwilling to share their usage history. In fact, Thom-Santelli et al. [108] found that tension may arise between users (experts and novice contributors) especially when they perceive a threat (e.g., when a workplace bonus or promotion is at stake).

With the above considerations, we believe providing quality and usage information during subset selection and analysis was an effective way to alleviate many of our domain experts’ concerns; however, providing additional context and advanced configuration capabilities is the next important step, especially for power users.

8 LESSONS LEARNED

Give importance to data preparation, not just visual data analysis. Motivated by the “garbage in, garbage out” principle, applications need more data work [99], as was also echoed by our users. Achieving this balance in data analysis tools is desirable and raises the need for functionality to inspect and interact with auxiliary information such as DataPilot’s quality and usage. Furthermore, the workflow change to prioritize and integrate data preparation (e.g., by first selecting relevant attributes and records) into current visual data analysis tool workflows should be considered.

Present quality information for accurate and objective analysis. Evidenced by positive user feedback in terms of both importance and trust, quality can nudge users to pause and reflect upon the state of their data and take suitable corrective actions (e.g., clean the data) before performing analysis, or not use the data at all.

Present usage information, albeit with care and caution. Users had mixed feedback about usage in terms of both importance and trust. While usage information can help nudge users to draw inspiration from previous judgments, it can also be counterproductive, leading to the propagation of negative practices (e.g., biased analytic behaviors) or hampering creativity and originality (e.g., preventing fresh, new ideas to flourish) within the organization. One way to achieve a good balance is to present usage information on demand, e.g., “*when I get to know the dataset, I want to hide the second half [usage] of the [bi-colored glyph]*” (Q_5).

Different tasks call for flexible information. Data preparation and subsequent analysis are contextualized by specific task requirements and user preferences. Tools must provide the desired flexibility to, e.g., assign different weights to constituent quality and usage dimensions or determine different aggregation functions (max, mean) for calculating the *overall* scores, override preset constraints to assess *correctness* and *objectivity*, and modify the thresholds to determine the *high*, *medium*, *low* score cutoffs.

Additional degrees of guidance towards quality and usage characteristics may be pursued. Participants found that the visual interactive affordances for quality and usage information were useful for “orienting” [16] them with the dataset. Future tools could

also explore higher guidance degrees (e.g., “directing” or “prescribing” [16]) to more actively steer users rather than just passively increasing awareness hoping they react in good conscience.

Organizations should start building and utilizing collective intelligence. Organizations should capture usage logs across their databases and applications to model usage information to increase general user awareness within and across teams. Moreover, as new data is regularly ingested (and old data archived), persistence and subsequent monitoring of the auxiliary information can help detect shifting trends, flag anomalous events, and generally track data provenance, ensuring accurate and efficient data management.

9 LIMITATIONS AND FUTURE WORK

We noted five key limitations related to our studies and tool.

Study limitations. One, during the design study, we structured our interviews in a group setting; while these interviewees had a strong working relationship, this mode of discussion may result in complexities around gender and organizational hierarchies and must be accounted for (e.g., through 1-1 interviews). Two, during the user study, we made a fair assumption that our participants were unfamiliar with the dataset and hence exhibited similar expertise, supporting internal validity; however, this assumption may not hold true for real-world cases from an external validity standpoint [100]. Future work may incorporate weighting mechanisms to more accurately approximate usage based on recency of use (e.g., give more importance to recent data), user expertise (favor experts), or the criticality of the application that utilized the data. Three, because our participants were not domain experts, we did not have experts assess the selected subsets or final dashboards; future user studies with domain experts should further evaluate the quality of these results. Four, we focused on the particular task of exporting visualizations for a dashboard, which may have impacted how the attributes and records were chosen; future work should consider developing additional tools to study downstream analytics tasks other than subset selection such as ranking and clustering. Five, although data quality dimensions are not easily transferable across domains [70], study participants suggested utilizing DataPilot for searching datasets on Kaggle [59] (Q_{10}), generating tutorials for software use (Q_2 , D_5), and preparing fair and accurate datasets for machine learning (D_4), which are left for future work.

Tool limitations. One, DataPilot currently supports quality information for tabular datasets; future work may explore other structured (e.g., relational databases) and unstructured (e.g., text, documents) datasets. Two, there are also other data-dependent (e.g., consistent representation, ease of manipulation, and timeliness [89]) and process-dependent (e.g., data collection [38]) aspects of quality, and similarly, other aspects of usage beyond a subset selection and dashboard building task (e.g., co-usage frequencies of multiple attributes in a visualization, frequency of visualization interactions such as zooming and panning [14]) that may be operationalized in the future. Three, DataPilot’s dashboard view currently supports creation of disconnected visualizations; future work may explore the effects of interactive affordances such as brushing and linking. Four, to ensure scalability, DataPilot computes quality and usage scores using SQL queries (objectivity is computed using both SQL

and Python); some of these new dimensions, however, may be difficult to operationalize using SQL and hence challenging to scale. Five, the completeness, correctness, and objectivity quality constraints are currently hard-coded in the DataPilot source code in a SQL-like syntax. Future work can provide interactive affordances for the user to configure these constraints and also clean the data (e.g., handle missing values) directly via the user interface.

10 CONCLUSION

DataPilot is a visual data preparation and analysis tool that models two kinds of auxiliary information, *quality* and *usage*, to assist users in analyzing a large and unfamiliar tabular dataset, selecting a relevant subset, and building a visualization dashboard. DataPilot is an outcome of a design study with 14 data workers over a period of two months who communicated the importance of data quality and also suggested surfacing data usage characteristics to guide users during data preparation. A user study with 36 participants suggested that quality and usage information together help users select smaller, effective data subsets with greater success and confidence. We posit that through quality *and* usage information, organizations can build collective intelligence, increasing transparency and accuracy to foster closer collaboration and cooperation among teams.

ACKNOWLEDGMENTS

We thank our data worker interviewees, study participants, members of the Georgia Tech Visualization Lab, and anonymous reviewers for providing feedback at different stages of this work.

REFERENCES

- [1] 2012. Open Government Data (OGD) Platform India. Retrieved 21-Nov-2022 from <https://data.gov.in/>
- [2] 2021. CERN Open Data Portal. Retrieved 21-Nov-2022 from <https://opendata.cern.ch>
- [3] Ziawasch Abedjan, Xu Chu, Dong Deng, Raul Castro Fernandez, Ihab F. Ilyas, Mourad Ouzzani, Paolo Papotti, Michael Stonebraker, and Nan Tang. 2016. Detecting Data Errors: Where Are We and What Needs to Be Done? *Proc. VLDB Endow.* 9, 12 (aug 2016), 993–1004. <https://doi.org/10.14778/2994509.2994518>
- [4] Jason Alexander, Andy Cockburn, Stephen Fitchett, Carl Gutwin, and Saul Greenberg. 2009. Revisiting read wear: analysis, design, and evaluation of a footprints scrollbar. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1665–1674.
- [5] Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A Kievit. 2019. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome open research* 4 (2019).
- [6] Jumana Almahmoud, Robert DeLine, and Steven M Drucker. 2021. How Teams Communicate about the Quality of ML Models: A Case Study at an International Technology Company. *Proceedings of the ACM on Human-Computer Interaction* 5, GROUP (2021), 1–24.
- [7] Amazon. 2022. Deequ. Retrieved July 29, 2022 from <https://aws.amazon.com/blogs/big-data/test-data-quality-at-scale-with-deequ/>
- [8] Clemens Arbesser, Florian Spechtenhauser, Thomas Mühlbacher, and Harald Piring. 2017. Visplause: Visual Data Quality Assessment of Many Time Series Using Plausibility Checks. *IEEE Transactions on Visualization and Computer Graphics* 23, 1 (2017), 641–650. <https://doi.org/10.1109/TVCG.2016.2598592>
- [9] Matthew Arnold, Rachel KE Bellamy, Michael Hind, Stephanie Houde, Sameep Mehta, Aleksandra Mojsilović, Ravi Nair, K Natesan Ramamurthy, Alexandra Olteanu, David Piorkowski, et al. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6–1.
- [10] Ernesto Arroyo, Ted Selker, and Willy Wei. 2006. Usability tool for analysis of web designs using mouse tracks. In *CHI'06 extended abstracts on Human factors in computing systems*. 484–489.
- [11] Anant Bhardwaj, Souvik Bhattacharjee, Amit Chavan, Amol Deshpande, Aaron J Elmore, Samuel Madden, and Aditya G Parameswaran. 2014. Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798* (2014).
- [12] David Borland, Wenyan Wang, Jonathan Zhang, Joshua Shrestha, and David Gotz. 2019. Selection bias tracking and detailed subset comparison for high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 429–439.
- [13] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development*. sage.
- [14] Matthew Brehmer and Tamara Munzner. 2013. A multi-level typology of abstract visualization tasks. *IEEE transactions on visualization and computer graphics* 19, 12 (2013), 2376–2385.
- [15] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.
- [16] David Ceneda, Theresia Gschwandtner, Thorsten May, Silvia Miksch, Hans-Jörg Schulz, Marc Streit, and Christian Tominski. 2017. Characterizing Guidance in Visual Analytics. *IEEE Transactions on Visualization and Computer Graphics* (2017). <https://doi.org/10.1109/TVCG.2016.2598468>
- [17] Wo L Chang, Arnab Roy, Mark Underwood, et al. 2019. NIST Big Data Interoperability Framework: Volume 4, Security and Privacy. (2019).
- [18] Xu Chu and Ihab F. Ilyas. 2016. Qualitative Data Cleaning. *Proc. VLDB Endow.* 9, 13 (sep 2016), 1605–1608. <https://doi.org/10.14778/3007263.3007320>
- [19] Xu Chu, Ihab F. Ilyas, Sanjay Krishnan, and Jiannan Wang. 2016. Data Cleaning: Overview and Emerging Challenges. In *Proceedings of the 2016 International Conference on Management of Data (San Francisco, California, USA) (SIGMOD '16)*. Association for Computing Machinery, New York, NY, USA, 2201–2206. <https://doi.org/10.1145/2882903.2912574>
- [20] Xu Chu, Ihab F Ilyas, and Paolo Papotti. 2013. Holistic data cleaning: Putting violations into context. In *2013 IEEE 29th International Conference on Data Engineering (ICDE)*. IEEE, 458–469.
- [21] National Research Council et al. 2000. *How people learn: Brain, mind, experience, and school: Expanded edition*. National Academies Press.
- [22] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. <https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G>
- [23] Dean De Cock. 2011. Ames, Iowa: Alternative to the Boston housing data as an end of semester regression project. *Journal of Statistics Education* 19, 3 (2011). <https://doi.org/10.1080/10691898.2011.11889627>
- [24] Dong Deng, Raul Castro Fernandez, Ziawasch Abedjan, Sibao Wang, Michael Stonebraker, Ahmed K Elmagarmid, Ihab F Ilyas, Samuel Madden, Mourad Ouzzani, and Nan Tang. 2017. The Data Civilizer System.. In *Cidr*.
- [25] Dremio. 2022. Dremio. Retrieved July 29, 2022 from <https://www.dremio.com/>
- [26] Ian Drosos, Titus Barik, Philip J. Guo, Robert DeLine, and Sumit Gulwani. 2020. Wrex: A Unified Programming-by-Example Interaction for Synthesizing Readable Code for Data Scientists. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376442>
- [27] Ian Drosos, Titus Barik, Philip J. Guo, Robert DeLine, and Sumit Gulwani. 2020. Wrex: A Unified Programming-by-Example Interaction for Synthesizing Readable Code for Data Scientists (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3313831.3376442>
- [28] Susan Dumais, Robin Jeffries, Daniel M Russell, Diane Tang, and Jaime Teevan. 2014. Understanding user behavior through log data and analysis. In *Ways of Knowing in HCI*. Springer, 349–372.
- [29] Upol Ehsan, Q Vera Liao, Michael Muller, Mark O Riedl, and Justin D Weisz. 2021. Expanding explainability: towards social transparency in AI systems. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.
- [30] Niklas Elmqvist, Andrew Vande Moere, Hans-Christian Jetter, Daniel Cernea, Harald Reiterer, and TJ Jankun-Kelly. 2011. Fluid interaction for information visualization. *Information Visualization* 10, 4 (2011), 327–340.
- [31] Thomas Erickson and Wendy A. Kellogg. 2000. Social Translucence: An Approach to Designing Systems That Support Social Processes. *ACM Trans. Comput.-Hum. Interact.* 7, 1 (mar 2000). <https://doi.org/10.1145/344949.345004>
- [32] Mina Farid, Alexandra Roatis, Ihab F. Ilyas, Hella-Franziska Hoffmann, and Xu Chu. 2016. CLAMS: Bringing Quality to Data Lakes (SIGMOD '16). Association for Computing Machinery, New York, NY, USA, 2089–2092. <https://doi.org/10.1145/2882903.2899391>
- [33] Raul Castro Fernandez, Ziawasch Abedjan, Famen Koko, Gina Yuan, Samuel Madden, and Michael Stonebraker. 2018. AURUM: A data discovery system. In *2018 IEEE 34th International Conference on Data Engineering (ICDE)*. IEEE.
- [34] Ian Fette and Alexey Melnikov. 2011. *The websocket protocol*. Technical Report.
- [35] Mike Fleckenstein, Lorraine Fellows, and Krista Ferrante. 2018. Data Quality. In *Modern data strategy*. Springer.
- [36] Imola K Fodor. 2002. *A survey of dimension reduction techniques*. Technical Report. Lawrence Livermore National Lab., CA (US).
- [37] Python Software Foundation. 2022. Python. Retrieved July 29, 2022 from <https://www.python.org/>
- [38] Lawrence M Friedman, Curt D Furberg, and David L DeMets. 2010. Data collection and quality control. In *Fundamentals of clinical trials*. Springer.

- [39] Christian Fürber. 2016. Semantic Technologies. In *Data Quality Management with Semantic Technologies*. Springer, 56–68.
- [40] Gerry Gaffney. 1999. Affinity diagramming. Retrieved January 3 (1999), 2013.
- [41] Mirta Galesic, Rocio Garcia-Retamero, and Gerd Gigerenzer. 2009. Using icon arrays to communicate medical risks: overcoming low numeracy. *Health Psychology* 28, 2 (2009), 210. <https://doi.org/10.1037/a0014474>
- [42] Lisa Gitelman. 2013. *Raw data is an oxymoron*. MIT press.
- [43] Google. 2022. Angular. Retrieved July 29, 2022 from <https://angular.io/>
- [44] David Gotz, Shun Sun, and Nan Cao. 2016. Adaptive contextualization: Combating bias during high-dimensional visualization and data selection. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*. 85–95.
- [45] Carl Gutwin. 2002. Traces: Visualizing the immediate past to support group interaction. In *Graphics interface*. Citeseer, 43–50.
- [46] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.
- [47] Kelli Ham. 2013. OpenRefine (version 2.5). <http://openrefine.org>. Free, open-source tool for cleaning and transforming data. *Journal of the Medical Library Association: JMLA* 101, 3 (2013), 233. <https://doi.org/10.3163/1536-5050.101.3.020>
- [48] Anders Haug, Frederik Zachariassen, and Dennis Van Liempd. 2011. The costs of poor data quality. *Journal of Industrial Engineering and Management (JIEM)* 4, 2 (2011), 168–193.
- [49] Kim Hazelwood, Sarah Bird, David Brooks, Soumith Chintala, Utku Diril, Dmytro Dzhulgakov, Mohamed Fawzy, Bill Jia, Yangqing Jia, Aditya Kalro, et al. 2018. Applied machine learning at facebook: A datacenter infrastructure perspective. In *2018 IEEE International Symposium on High Performance Computer Architecture (HPCA)*. IEEE, 620–629.
- [50] Jeffrey Heer, Fernanda B Viégas, and Martin Wattenberg. 2007. Voyagers and voyeurs: supporting asynchronous collaborative information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 1029–1038.
- [51] Thomas N Herzog, Fritz J Scheuren, and William E Winkler. 2007. What is Data Quality and Why Should We Care? In *Data quality and record linkage techniques*. Springer, 7–15.
- [52] William C Hill, James D Hollan, Dave Wroblewski, and Tim McCandless. 1992. Edit wear and read wear. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 3–9.
- [53] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielewski. 2020. The dataset nutrition label. *Data Protection and Privacy, Volume 12: Data Protection and Democracy* 12 (2020), 1.
- [54] Bill Howe, Garret Cole, Nodira Khousainova, and Leilani Battle. 2011. Automatic Example Queries for Ad Hoc Databases. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data (Athens, Greece) (SIGMOD '11)*. Association for Computing Machinery, New York, NY, USA. <https://doi.org/10.1145/1989323.1989487>
- [55] Petra Isenberger, Niklas Elmqvist, Jean Scholtz, Daniel Cernea, Kwan-Liu Ma, and Hans Hagen. 2011. Collaborative visualization: Definition, challenges, and research agenda. *Information Visualization* 10, 4 (2011), 310–326.
- [56] George H. John, Ron Kohavi, and Karl Pflieger. 1994. Irrelevant Features and the Subset Selection Problem. In *Machine Learning Proceedings 1994*, William W. Cohen and Haym Hirsh (Eds.). Morgan Kaufmann, San Francisco (CA), 121–129. <https://doi.org/10.1016/B978-1-55860-335-6.50023-4>
- [57] A. Jović, K. Brkić, and N. Bogunović. 2015. A review of feature selection methods with applications. In *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. 1200–1205. <https://doi.org/10.1109/MIPRO.2015.7160458>
- [58] Kaggle. 2019. Kaggle ML & DS Survey. Retrieved 02/25/2022 from <https://www.kaggle.com/c/kaggle-survey-2019>
- [59] Kaggle. 2022. Kaggle. Retrieved July 29, 2022 from <https://www.kaggle.com/>
- [60] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. 2011. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3363–3372.
- [61] 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.
- [62] Sean Kandel, Ravi Parikh, Andreas Paepcke, Joseph M. Hellerstein, and Jeffrey Heer. 2012. Profiler: Integrated Statistical Analysis and Visualization for Data Quality Assessment. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (Capri Island, Italy) (AVI '12)*. Association for Computing Machinery, New York, NY, USA, 547–554. <https://doi.org/10.1145/2254556.2254659>
- [63] Jihie Kim, Ewa Deelman, Yolanda Gil, Gaurang Mehta, and Varun Ratnakar. 2008. Provenance trails in the wings/pegasus system. *Concurrency and Computation: Practice and Experience* 20, 5 (2008), 587–597.
- [64] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2017. Data scientists in software teams: Jean of the art and challenges. *IEEE Transactions on Software Engineering* 44, 11 (2017), 1024–1038.
- [65] Laura Koesten, Emilia Kacprzak, Jeni Tennison, and Elena Simperl. 2019. Collaborative practices with structured data: Do tools support what users need?. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [66] Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020. Dataset reuse: Toward translating principles to practice. *Patterns* 1, 8 (2020), 100136.
- [67] Kurt Koffka. 2013. *Principles of Gestalt psychology*. Routledge.
- [68] Jonathan Lancar. 2022. Luma. https://github.com/adobe/experience-platform-dsw-reference/blob/master/datasets/luma/luma_post_extended.csv.
- [69] Nuno Laranjeiro, Seyma Nur Soydemir, and Jorge Bernardino. 2015. A survey on data quality: classifying poor data. In *2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC)*. IEEE, 179–188.
- [70] Sabina Leonelli. 2019. Data governance is key to interpretation: Reconceptualizing data in data science. *Harvard Data Science Review* 1, 1 (2019), 10–1162.
- [71] Zipeng Liu, Zhicheng Liu, and Tamara Munzner. 2020. Data-driven Multi-level Segmentation of Image Editing Logs. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [72] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. Interactive cleaning for progressive visualization through composite questions. In *2020 IEEE 36th International Conference on Data Engineering (ICDE)*. IEEE.
- [73] Yuyu Luo, Chengliang Chai, Xuedi Qin, Nan Tang, and Guoliang Li. 2020. VisClean: Interactive Cleaning for Progressive Visualization. *Proc. VLDB Endow.* (aug 2020). <https://doi.org/10.14778/3415478.3415484>
- [74] Rupa Mahanti. 2019. Data, Data Quality, and Cost of Poor Data Quality. In *Data quality: dimensions, measurement, strategy, management, and governance*. Quality Press.
- [75] Antony SR Manstead, Miles Ed Hewstone, Susan T Fiske, Michael A Hogg, Harry T Reis, and Gün R Semin. 1995. *The Blackwell encyclopedia of social psychology*. Blackwell Reference/Blackwell Publishers.
- [76] Chris Mayfield, Jennifer Neville, and Sunil Prabhakar. 2010. ERACER: a database approach for statistical inference and data cleaning. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*. 75–86.
- [77] Srdan Medić, Biljana Karlović, and Zrinko Cindrić. 2016. New standard ISO 9001: 2015 and its effect on organisations. *Interdisciplinary Description of Complex Systems: INDECS* 14, 2 (2016), 188–193.
- [78] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.
- [79] Metaplane. 2022. Metaplane. Retrieved July 29, 2022 from <https://www.metaplane.dev/>
- [80] Microsoft. 2022. LinkedIn. Retrieved July 29, 2022 from <https://linkedin.com/>
- [81] Microsoft. 2022. Power BI Desktop. Retrieved May 25, 2022 from <https://powerbi.microsoft.com/en-us/>
- [82] Microsoft. 2022. Teams. Retrieved July 29, 2022 from <https://www.microsoft.com/en-us/microsoft-teams/group-chat-software>
- [83] Tadhg Nagle, Thomas C Redman, and David Sammon. 2017. Only 3% of companies' data meets basic quality standards. *Harvard Business Review* (2017).
- [84] Arpit Narechania, Adam Coscia, Emily Wall, and Alex Endert. 2021. Lumos: Increasing awareness of analytic behavior during visual data analysis. *IEEE Transactions on Visualization and Computer Graphics* (2021).
- [85] Jakob Nielsen and Kara Pernice. 2010. *Eyetracking web usability*. New Riders.
- [86] Chris North, Remco Chang, Alex Endert, Wenwen Dou, Richard May, Bill Pike, and Glenn Fink. 2011. Analytic provenance: process+ interaction+ insight. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. 33–36.
- [87] Boris Otto, Kai M Hüner, and Hubert Osterle. 2009. Identification of Business Oriented Data Quality Metrics. In *ICIQ*. 122–134.
- [88] Jinglin Peng, Weiyuan Wu, Brandon Lockhart, Song Bian, Jing Nathan Yan, Linghao Xu, Zhixuan Chi, Jeffrey M Rzeszotarski, and Jiannan Wang. 2021. DataPrep.EDA: Task-Centric Exploratory Data Analysis for Statistical Modeling in Python. In *Proceedings of the 2021 International Conference on Management of Data*. 2271–2280.
- [89] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. *Commun. ACM* 45, 4 (apr 2002), 211–218. <https://doi.org/10.1145/505248.506010>
- [90] EU Open Data Portal. 2016. EU open data. (2016).
- [91] Pavel Pudil and Jana Novotná. 1998. Novel methods for feature subset selection with respect to problem knowledge. In *Feature extraction, construction and selection*. Springer, 101–116.
- [92] Dorian Pyle. 1999. *Data preparation for data mining*. morgan kaufmann.
- [93] Eric D Ragan, Alex Endert, Jibonananda Sanyal, and Jian Chen. 2015. Characterizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 31–40.
- [94] Vijayshankar Raman and Joseph M Hellerstein. 2001. Potter's wheel: An interactive data cleaning system. In *VLDB*, Vol. 1. 381–390.
- [95] Thomas C. Redman. 2016. Bad Data Costs the U.S. \$3 Trillion Per Year. <https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year>
- [96] Thomas C Redman. 2018. If your data is bad, your machine learning tools are useless. *Harvard Business Review* 2 (2018).
- [97] Robert Richards. 2006. Representational state transfer (rest). In *Pro PHP XML and web services*. Springer, 633–672.

- [98] Steven G Rogelberg, Linda Rhoades Shanock, and Cliff W Scott. 2012. Wasted time and money in meetings: Increasing return on investment. *Small Group Research* 43, 2 (2012), 236–245.
- [99] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. “Everyone wants to do the model work, not the data work”: Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [100] Mark A Schmuckler. 2001. What is ecological validity? A dimensional analysis. *Infancy* 2, 4 (2001), 419–436.
- [101] Hans-Jörg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. 2017. A systematic view on data descriptors for the visual analysis of tabular data. *Information Visualization* 16, 3 (2017), 232–256.
- [102] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2431–2440. <https://doi.org/10.1109/TVCG.2012.213>
- [103] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. 2012. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval & Knowledge Management*. 300–304. <https://doi.org/10.1109/InfRKM.2012.6204995>
- [104] Qinqiao Song and Martin Shepperd. 2007. Missing Data Imputation Techniques. *Int. J. Bus. Intell. Data Min.* 2, 3 (oct 2007), 261–291. <https://doi.org/10.1504/IJBIDM.2007.015485>
- [105] Brian Stein and Alan Morrison. 2014. Data lakes and the promise of unsiloed data. *PricewaterhouseCooper, Technology Forecast: Rethinking integration* (2014). <https://www.pwc.com/us/en/technology-forecast/2014/cloud-computing/assets/pdf/pwc-technology-forecast-data-lakes.pdf>
- [106] Tableau. 2022. Tableau. Retrieved July 29, 2022 from <https://www.tableau.com/>
- [107] Tableau. 2022. Tableau Prep. Retrieved May 25, 2022 from <https://www.tableau.com/products/prep>
- [108] Jennifer Thom-Santelli, Dan Cosley, and Geri Gay. 2010. What do you know? Experts, novices and territoriality in collaborative systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1685–1694.
- [109] Jennifer Thom-Santelli, Dan R Cosley, and Geri Gay. 2009. What’s mine is mine: territoriality in collaborative authoring. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1481–1484.
- [110] Trifacta. 2022. Wrangler. Retrieved July 29, 2022 from <https://www.trifacta.com/>
- [111] A Caroline Tynan and Jennifer Drayton. 1987. Market segmentation. *Journal of marketing management* 2, 3 (1987), 301–335.
- [112] Uber. 2022. Uber Databook. Retrieved July 29, 2022 from <https://eng.uber.com/databook/>
- [113] Hannes Ulrich, Ann-Kristin Kock-Schoppenhauer, Noemi Deppenwiese, Robert Gött, Jori Kern, Martin Lablans, Raphael W Majeed, Mark R Stöhr, Jürgen Stausberg, Julian Varghese, Martin Dugas, and Josef Ingenerf. 2022. Understanding the Nature of Metadata: Systematic Review. *J Med Internet Res* 24, 1 (11 Jan 2022), e25440. <https://doi.org/10.2196/25440>
- [114] Jürgen Umbrich, Sebastian Neumaier, and Axel Polleres. 2015. Quality assessment and evolution of open data portals. In *2015 3rd international conference on future internet of things and cloud*. IEEE, 404–411.
- [115] Reza Vaziri, Mehran Mohsenzadeh, and Jafar Habibi. 2019. Measuring data quality with weighted metrics. *Total Quality Management & Business Excellence* 30, 5-6 (2019), 708–720.
- [116] Fernanda B Viegas and Martin Wattenberg. 2006. Communication-minded visualization: A call to action. *IBM Systems Journal* 45, 4 (2006), 801.
- [117] Fernanda B. Viegas, Martin Wattenberg, Frank van Ham, Jesse Kriss, and Matt McKeon. 2007. ManyEyes: A Site for Visualization at Internet Scale. 13, 6 (nov 2007), 1121–1128. <https://doi.org/10.1109/TVCG.2007.70577>
- [118] Emily Wall, Leslie M Blaha, Lyndsey Franklin, and Alex Endert. 2017. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *2017 IEEE conference on visual analytics science and technology (vast)*. IEEE, 104–115.
- [119] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2021. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE Transactions on Visualization and Computer Graphics* (2021).
- [120] Dakuo Wang, Judith S Olson, Jingwen Zhang, Trung Nguyen, and Gary M Olson. 2015. DocuViz: visualizing collaborative writing. In *Proceedings of the 33rd Annual ACM conference on human factors in computing systems*. 1865–1874.
- [121] Steve Wedig and Omid Madani. 2006. A large-scale analysis of query logs for assessing personalization opportunities. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. 742–747.
- [122] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. 2007. Scented widgets: Improving navigation cues with embedded visualizations. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1129–1136.
- [123] Wesley Willett, Jeffrey Heer, Joseph Hellerstein, and Maneesh Agrawala. 2011. CommentSpace: structured support for collaborative visual analysis. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*.
- [124] Kanit Wongsuphasawat, Dominik Moritz, Anushka Anand, Jock Mackinlay, Bill Howe, and Jeffrey Heer. 2015. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 649–658.
- [125] Mohamed Yakout, Laure Berti-Équille, and Ahmed K Elmagarmid. 2013. Don’t be scared: use scalable automatic repairing with maximal likelihood and bounded changes. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. 553–564.
- [126] Cong Yan and Yeye He. 2020. Auto-Suggest: Learning-to-Recommend Data Preparation Steps Using Data Science Notebooks. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data (Portland, OR, USA) (SIGMOD ’20)*. Association for Computing Machinery, New York, NY, USA, 1539–1554. <https://doi.org/10.1145/3318464.3389738>
- [127] Faheem Zafar, Abid Khan, Saba Suhail, Idrees Ahmed, Khizar Hameed, Hayat Mohammad Khan, Farhana Jabeen, and Adeel Anjum. 2017. Trustworthy data: A survey, taxonomy and future trends of secure provenance schemes. *Journal of network and computer applications* 94 (2017), 50–68.
- [128] Amy X Zhang, Michael Muller, and Dakuo Wang. 2020. How do data science workers collaborate? roles, workflows, and tools. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–23.
- [129] Shichao Zhang, Chengqi Zhang, and Qiang Yang. 2003. Data preparation for data mining. *Applied Artificial Intelligence* 17, 5-6 (2003), 375–381. <https://doi.org/10.1080/713827180> arXiv:<https://doi.org/10.1080/713827180>