CIF21 DIBBs: EI: Collaborative: Vizier, Streamlined Data Curation 4 April 2016
Oliver Kennedy, Boris Glavic, Juliana Freire *NSF Divisions*: CISE/III, SBS/SES
Keywords: Provenance, Curation, Ambiguity, Urban Data, Sensor Data, Wrangling

Curation as an Integral Component of Data Exploration Data curation is a critical task in data science in which raw data is structured, validated, and repaired. Data validation and repair establish trust in analytical results, while appropriate structuring streamlines analytics. Unfortunately, even with advances in automated data cleaning tools, such as Oracle's Data Guide and Trifacta's Wrangler, curation is still a major bottleneck in data exploration. Traditionally, curation has been carried out as a pre-processing task: after all data are selected for a study (or application), they are cleaned and loaded into a database or data warehouse. This is problematic because while some cleaning constraints can be easily defined (e.g., checking for valid attribute ranges), others are only discovered as one analyzes the data. As one example, consider taxis in New York City [218]. Every day, there are over 500,000 taxi trips transporting about 600,000 people from Manhattan to different parts of the city [36]. Through the meters installed in each vehicle, the Taxi & Limousine Commission (TLC) captures detailed information about trips, including: GPS readings for pick-up and drop-off locations, pick-up and drop-off times, fare, and tip amount. These data have been used in several projects to understand different aspects of the city, from creating mobility models and analyzing the benefits and drawbacks of ride sharing, to detecting gentrification. In a recent study [89], we investigated quality issues in the taxi data. We found invalid values such as negative mile and fare values, as well as trips that started or ended in rivers or outside of the US. These are clearly errors in the data. Other issues are more nuanced. An example is a fare with a tip of US$938.02 (the maximum tip value for the 2010 dataset). While this could have be an error in the data acquisition or in the credit card information, it could also be the case that a wealthy passenger overtipped her taxi driver. Issues are often detected during analytics, as different slices of the data are aggregated. Figure 1 shows the number of daily taxi trips in New York City (NYC) during 2011 and 2012. Note the large drops in the number of trips in August 2011 and October 2012. Standard cleaning techniques are likely to classify these drastic reductions as outliers that represent corrupted or incorrect data. However, by integrating the taxi trips with wind speed data (bottom plot in Figure 1), we discover that the drops occur on days with abnormally high wind speeds, suggesting a causal relation: the effect of extreme weather on the number of taxi trips in NYC. Removing such outliers would hide an important phenomenon. Conversely, detecting it upfront requires identifying a non-obvious pattern in a very high-dimensional space [69].

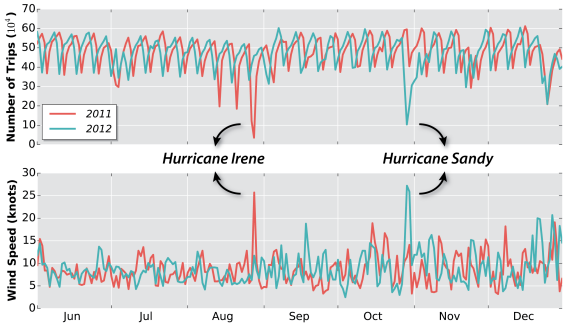

Figure 1: The plot on the top shows how the number of trips varies over 2011 and 2012. While the variation is similar for the two years, there are clear outliers, including large drops in August 2011 and in October 2012. These are not errors, but in fact correspond to hurricanes Sandy and Irene, as shown by the wind speed plot on the bottom.

Issues like these appear across all forms of analytics, making curation an integral component of data exploration. When erroneous features are identified, appropriate *cleaning operations should be applied on the fly*. Besides the need to refine a curation pipeline as the user gets more familiar with a dataset, different questions that arise during exploration may require different cleaning strategies. Thus, we need to move from the traditional cleaning function $DirtyData \rightarrow CleanData$, to a function that encapsulates the exploratory curation process: $DirtyData \times UserTask \rightarrow (CleanData, Explanation)$. The trial-and-error nature of the curation process poses several challenges. First, and foremost, the cleaned data must be accompanied by its provenance which *explains the transformations applied to the raw data*, as well as *ambiguities that arise while applying these transformations*. This is critical for subsequent analyses, as experts need to both assess the quality of the data and understand which assumptions they can rely on. If an operation is applied and later found to be incorrect (e.g., removing the outliers in Figure 1), it should be possible to *undo the operation and all of its direct and indirect effects*. It should also be possible to *modify an operation* (e.g., change the

parameters of an outlier detection operation to not consider the $938.02 tip amount as an outlier) and the effects of this modification should be propagated to derived data. Furthermore, it is often necessary to *explore and compare alternative cleaning strategies and to consolidate their results*, especially when data curation is a collaborative effort. Currently, such corrections and refinements require an error-prone, time-consuming process to track different versions of scripts and queries used, and the data feeds they were applied to.

Building Infrastructure for Data Curation In this project, we propose to build Vizier, a system that unifies curation and data exploration through provenance. Vizier tracks the provenance of the exploratory process and all the computations performed [93, 98, 159, 224] as well as the provenance of the affected data items [215]. By connecting these two forms of provenance, Vizier is able to provide a detailed audit trail that explains which operations caused which changes to the data and to reproduce these steps on new data. Vizier will integrate and extend three production/near-production quality systems that we have developed in previous work: *Mimir* [143, 167, 231, 232], a system that supports probabilistic pay-as-you-go data curation operators; *VisTrails* [42, 43, 51, 99, 101, 128, 151, 200, 202, 209], an NSF-supported open-source system designed for interactive data exploration; and *GProM* [10, 11, 106, 107, 109, 116, 175], a database middleware that efficiently tracks fine-grained data provenance, even through update operations. A core challenge of implementing Vizier will be to integrate the data and provenance models of these systems and to develop a unified interface that synergystically combines the features of each.

To define the requirements for the system design, we have taken into consideration the curation needs of previous and ongoing projects led by the PIs in different domains—from urban to industrial data, as well as the results of a survey (see Section 1). Curation is a difficult task for experts in data management and computer science, and even more so for the growing number of *data enthusiasts*, with expertise in particular domains but that often lack training in computing. Furthermore, the process demands the iterative, trial-and-error application and tuning of data cleaning operations. To support these requirements, Vizier will provide a novel "sandbox" interface for exploration and curation that is easy to use, and whose unique features are enabled by the integration of fine- and coarse-grained provenance.

Vizier's interface is a hybrid of notebook-style interfaces and spreadsheets. Notebook interfaces, as exemplified in Jupyter (`http://jupyter.org`), interleave code, data, summaries, visualizations, and documentation into a single train of thought that allows readers to retrace an analyst's exploratory process. Conversely, spreadsheets allow analysts an unprecedented level of flexibility and freedom, while still supporting a powerful programming model. Thus, in contrast to classical notebook software where all data manipulations must be handled programmatically, Vizier allows data, summaries, and visualizations to be edited directly as in a spreadsheet *and* in a classical programming environment *simultaneously*.

Vizier uses provenance to aid the user in the exploration process by enabling her to navigate the evolution of her curation pipeline, to understand the impact of cleaning operations she devises, and to understand sources and effects of uncertainty or ambiguity in her data. Vizier also uses provenance to provide recommendations to the user on how to tune curation operations, which curation operations to apply next [152], and how to generalize her curation workflow developed over a small sample dataset to deploy it over a large scale dataset (e.g., to deploy the workflow on a Big Data platform).

Finally, provenance enables the use of ambiguity-aware data transformations, including automated zero-configuration curation operators called lenses [167, 232]. Unlike classical clean-before-use approaches to data curation, Vizier allows data wranglers to defer resolving ambiguities. Ambiguities persist through transformations and appear as annotations (e.g., asterisks) on data in Vizier that indicate the cause and quantify the effects of the ambiguity. By deferring ambiguity resolution until after the user has had a chance to explore, the analyst can better decide how to resolve the ambiguity, or even whether she needs to resolve the ambiguity at all. Lenses also form the basis for Vizier's extensibility, as they allow data cleaning heuristics or tools to be just "plugged" in, even if the tools have overlapping use cases or conflicting effects.

The interface features of Vizier are enabled by tracking *provenance*. Edits, whether to code or data, are all *transparently* recorded and associated with the resulting data. Guesses made by the system on behalf of the user are recorded similarly. Provenance persists through computations and modifications, providing an audit trail and allowing the system to explain the reasoning behind results it shows to the analyst. Provenance allows potentially suspicious or ambiguous results to be flagged as such, provides context for results, and

by using well-established techniques for probabilistic and uncertain data management [214], also quantifies the impact of the ambiguity on the result. An important technical challenge we will tackle in this regard is how to integrate the workflow (evolution) provenance of VisTrails with the fine-grained data provenance of GProM and the ambiguity tracking of Mimir.

Interactive exploration also requires interactive latencies, even for large datasets. We will build Vizier to provide interactive response times through two techniques. First, through its facilities for tracking provenance, Vizier can enact a form of program slicing, a more general form of incremental view maintenance that allows data representations produced by Vizier to be updated rapidly in response to small changes to input data and parameters of curation workflows. Second, similar to Trifacta's Wrangler, Vizier allows analysts to extract and use samples of large datasets for preliminary development efforts. Unlike Wrangler, however, Vizier can measure the quality or representativeness of a sample set with respect to a given summary, representation, or view of the data. Furthermore, Vizier guides the user in tweaking a curation workflow developed over a sample dataset to ensure that it generalizes to the complete data.

Community Building. Users in a broad range of application domains from urban sciences to business intelligence, have committed to collaborate with the PIs on this proposal (see Section 1 and letters of collaboration). This diversity of collaborators will ensure that Vizier is relevant to real-world needs and help to establish a healthy user base. The latter is particularly important to sustain development beyond the initial funding period. Our strategy for building a healthy community around Vizier includes demonstrations, workshops, publicly released code, and ensuring proper documentation for users and developers, and development of project governance. Our sustainability plan is detailed in Section 4.

The Team. The PIs bring cross-cutting expertise from different areas of research on data management, provenance, and visualization. Systems developed by each of the PIs: Mimir, GProM and VisTrails, will serve as the core building blocks of our proposed system Vizier. Additionally, PIs Kennedy and Glavic already have a record of collaboration that has resulted in joint papers (under submission) related to data curation, uncertainty, and provenance. Both PIs have ongoing projects in this domain that are sponsored by Oracle (see letter of collaboration from Gawlick and Hua-Liu). PI Kennedy's expertise covers optimization and incremental computation [3, 138, 145], data structures [144], uncertain data management [140–143, 167, 168, 232], online aggregation [137, 141], and mobile systems [49, 139]. He is a member of UB's Center for Multisource Information Fusion and National Center for Geographic Information Awareness, and has collaborations [49,139] based on UB's NSF-funded PhoneLab smartphone experimental platform. PI Glavic's research is focused on database systems with a strong track record in provenance [9, 11, 105–117, 174, 175, 182, 183] and data integration [12–15, 109, 110]. He has designed and implemented several provenance-aware systems including Perm [106–108, 116], GProM [9, 11, 174, 175], Vagabond [109, 110], Ariadne [111, 112], and LDV [114, 182, 183]. PI Freire's research has focused on big-data analysis and visualization, large-scale information integration, provenance management, and computational reproducibility. She has a track record of successful interdisciplinary collaborations and her work has had impact in a number of domains beyond computer science, from physics and biology to climate research and social sciences [18, 29, 83, 125, 127, 194, 195, 199]. She has co-authored multiple open-source systems [1, 34, 189, 216, 221, 224] (see `https://github.com/ViDA-NYU`), including VisTrails which is used by high-visibility scientific applications in diverse fields. As the Executive Director of the Moore-Sloan Data Science Environment at NYU, a faculty member at the NYU Center for Data Science and at the NYU Center for Urban Science and Progress, she currently leads several projects where data curation is a key challenge.

1 Applications and Requirement Gathering

The need for data curation arises in all applications of data science. To ensure that our efforts will see use in practice, the PIs will deploy Vizier through established collaborative efforts with data scientists in academia and industry. These collaborations will serve as a platform to evaluate Vizier, as well as a source of feedback, helping us to identify and address critical pain points in data curation workflows. Additionally, we have conducted an informal survey, reaching out to several affiliated data science communities and potential consumers of Vizier for feedback and desiderata. In this section, we outline our two primary collaborative efforts, as well as the high-level feedback garnered from our informal survey.

1.1 Curating Urban Data

The NYU team is working on several projects that involve curation, analysis and integration of urban data [28, 33, 47, 70, 71, 84, 102, 125, 180, 181, 184, 216]. PI Freire is a faculty member at the NYU Center for Urban Science and Progress (CUSP). CUSP is set up as a unique public-private research center that uses New York City as its laboratory and classroom to help cities around the world become more productive, livable, equitable, and resilient. Research and development at CUSP focuses on the collection, integration, and analysis of data to improve urban systems. The social sciences play an integral role in CUSP activities– people are the customers and operators of urban systems, so that understanding them and their behavior is essential to the CUSP mission. Conversely, CUSPs large, multi-modal data sets and technology have the potential to revolutionize the social sciences. CUSP has established a data facility (CDF) to support the empirical study of cities. Freire was one of the original designers of CDF and part of the vision of this proposal was motivated by the needs of CDF users, notably: different users (and projects) need to combine and clean datasets in different ways, depending on their research questions, which change over time as the process evolves and new hypotheses are formulated; users come from widely different backgrounds, and include social scientists, physicists, computer scientists, civil engineers, policy makers, and students. As Professor Lane and Dr. Rosen state in their letter of collaboration, the CDF needs a data curation infrastructure such as the one we propose to build. Vizier will bring many benefits to CDF, including: users will be to able collaboratively curate data; curated data will include detailed provenance, allowing them to be re-used; and curation pipelines will also be shared within the facility, enabling users to benefit from the collective wisdom of their peers by re-using and building upon these pipelines.

Freire also has an ongoing collaboration with the NYU Furman Center for Real Estate and Policy [125]. The Furman Center uses urban data to explore the effect that land use regulations, real estate development, and other public and private place-based investments have on the quality, affordability, and character of neighborhoods and on individual well-being (see e.g., [76, 77, 126, 205]). They take an interdisciplinary approach, applying policy and legal analyses, as well as econometric techniques to study critical current issues. Over the years, they have collected a broad array of data on demographics, neighborhood conditions, infrastructure, housing stock and other aspects of New York City's neighborhoods and real estate market [103]. They also produce a series of data products that help inform community groups, developers, policymakers and investors about trends, challenges and opportunities in particular neighborhoods [31, 32, 45, 46]. Given that their research has direct impact on policies and the data they release is widely used, for them, data quality is of utmost importance (see letter from Professor Gould Ellen).

Freire's group has multiple ongoing collaborations with NYC agencies. She is a member of the NYC Taxi & Limousine Commission (TLC) Data/Technology Advisory Committee. She has collaborated with the TLC on different projects, from the study of privacy and quality issues in data they release [89] to the deployment of TaxiVis, an open-source tool for the analysis of spatio-temporal data [84, 102, 216]. As with many other agencies that use data to improve their operations and policies, and that publicly release these data [179], the TLC faces tremendous challenges around data quality (see letter by Rodney Styles, TLC).

Our collaborators will provide us unique data and feedback that we will use in the design of Vizier, and we will collaborate with them on the deployment of the system at CUSP, Furman Center, and TLC.

1.2 Industrial Strength Data Lakes

Our second effort is part of a large, ongoing collaborative project between PIs Kennedy and Glavic and Oracle's Dieter Gawlick and Zhen Hua-Liu, as well as with Ronny Fehling, Head of Data Driven Technologies and Advanced Analytics at Airbus (see attached letters). One specific part of this effort concerns a large scale project at Airbus to merge data sources from different sectors of its business into a single, company-wide data lake. In addition to the issues of data quality that arise in our collaboration with CUSP, an operation of this scale presents several additional challenges:

Cross-Domain Data Mapping. Datasets from different sectors have distinct, non-overlapping schemas, as well as distinct attribute domains recorded at different granularities. There are also more subtle conflicts. For example, one assumption often made when using a dataset is temporal stability [143]. An anecdotal example encountered during past collaborative efforts involved an auto-generated report that computed revenue totals for sales groups at a large company. Revenue totals would fluctuate unexpectedly over time,

even decreasing for some groups. The cause was mismatched temporal assumptions between an append-only sales history dataset and a HR database linking salespeople to their groups. As salespeople moved between groups, the HR database was updated and past sales would be re-attributed to their new group.

Data Discovery. Given the number of distinct data sources and data sets available at a company of this scale, simply finding datasets applicable to a specific problem is a challenge. Forcing participants in the data lake to properly create and curate metadata for their datasets is infeasible, necessitating alternative approaches to data discovery.

Heterogeneity. The data lake draws on a heterogeneous mix of storage layers and data sources including Hadoop, client APIs, external tables, and live data feeds, a fact that makes other data quality challenges more difficult. Cross domain data mapping is more difficult, as schemas and attribute domains may not be immediately accessible. Data discovery also becomes more difficult, as complete datasets are not available and associated metadata must either be tracked independently or through purpose-built adaptors.

Versioning. As analysts progressively refine and extend a dataset, multiple revisions of the same data become available. Using an earlier revision may require an analyst to duplicate effort, while a later revision may be modified in unexpected ways that are inconsistent with the analyst's current goals.

1.3 Community Feedback

As a preliminary effort at reaching a broader audience, we have conducted an informal survey to gauge preliminary interest in Vizier, as well as to solicit feedback from different communities about the data curation challenges that they face. The survey was distributed directly to many of the PI's colleagues and collaborators at potential deployment sites like UB's National Center for Geographic Information and Analysis. Over the course of a one-week informal survey period, we received 8 detailed responses from parties expressing interest in improved tools for data cleaning and a further 3 additional off-the-record responses expressing frustration with the state of the art in data curation. The preliminary responses indicate that *spreadsheets are the tool of choice for data curation* (used by 6 out of the 8 respondents). The two respondents who do not use spreadsheets, regularly work with data at sizes to which spreadsheets do not scale. When asked informally, one of these respondents stated that a spreadsheet-style interface used to design scalable curation workflows over samples would be useful to them. When asked about the biggest challenges in their most recent analytics task, the two most common answers were: *the time taken to perform computations* and *data quality*. For the first case, standard database techniques (indexing, incremental maintenance, set-at-a-time) are often sufficient, but require too much effort to deploy, configure, and load data into. A concern with missing or garbage data was also common, and several respondents in particular noted issues of data quality when working with different schema versions simultaneously.

Throughout our project, we will continue to seek feedback from experts in different domains. This will not only ensure the proposed infrastructure will be widely available, but we also hope this will help create and sustain a strong user community for Vizier.

2 The Vizier System

The goal of Vizier is to **unify the tasks of data exploration and data curation** under a single interface (illustrated in Figure 3) that combines characteristics of a notebook and a spreadsheet. Our aim is to provide a tool that allows users to quickly and easily tabulate, visualize, clean, and summarize large datasets. As the user explores her data, Vizier provides a variety of tools for curation, from the ability to manually edit individual records, to batch operators that **automate common tasks** like schema merging, outlier detection, missing value imputation, entity resolution, constraint repair, and others. While the user is exploring and curating the data, Vizier records her curation activities to **incrementally build a workflow** [5, 51, 64, 65, 101, 105, 128, 202], as illustrated in Figure 2a. During this process, the system **recommends curation steps and tuning parameters** to the user based on successful past curation efforts on datasets with similar characteristics [200]. For example, the system may detect that the user is constructing a typical address cleaning workflow and suggest curation steps that are commonly used to clean addresses such as using an external zip-city lookup table. For large datasets, a user would typically first curate and explore a small sample and only **deploy** her curation workflow over the full dataset once she

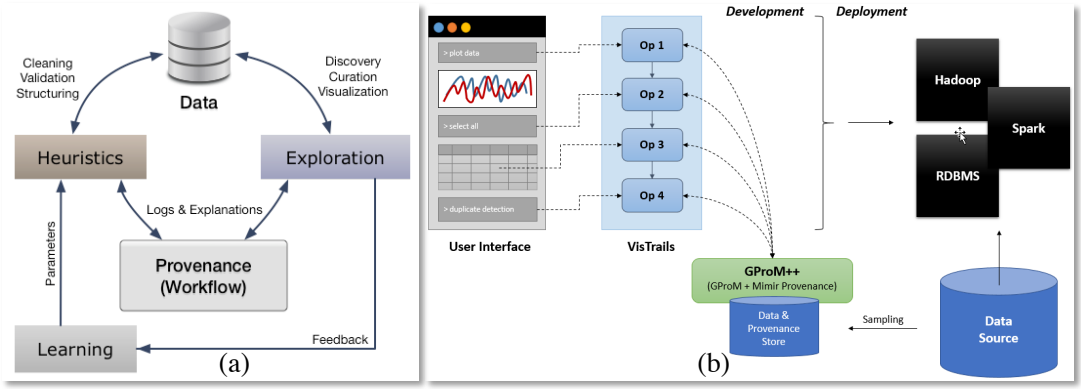

Figure 2: The Vizier System. (a) Vizier backs freeform data exploration with heuristic data curation operators and a feedback and provenance-driven learning engine that continually improves its curation heuristics. (b) Workflows are designed through Vizier's hybrid notebook-spreadsheet interface on small or sampled dataset processed locally by an engine backed by GProM and Mimir. When workflows are ready, they can also be deployed to the cloud, to a Hadoop or Spark cluster, or to a relational database system.

is sufficiently certain that the workflow resolves all data quality issues. This type of deployment at-scale is supported in Vizier, as illustrated in Figure 2b. One challenge in this setting is that the curation steps applied over the sample dataset need to be generalized to correctly apply to the full dataset. For example, the user may manually remove taxi trips with unrealistically large fare values from the sample dataset. However, the full dataset will contain many additional such trips that would not be removed when the workflow is deployed, unless the deletion of a fixed set of trips is *generalized* to a deletion of trips based on their fare value. Vizier aids the user in the deployment by providing recommendations on how to **generalize** her operations based on data characteristics and provenance. For example, the system may detect that all deleted trips have an unusually high fare in common and offer to delete all records with similarly high fares in one operation. Combining workflow provenance, data provenance, and uncertainty management, Vizier can offer non-intrusive visualizations that **fully explain** how data was produced and how ambiguities in the input data or curation steps have affected it. For example, clicking on a value in Vizier's interface will bring up an explanation view showing the value's provenance, ambiguities in its derivation, and how it compares to related values (e.g., those in the same column).

Interface. Vizier's interface (illustrated in Figure 3) combines elements of both notebooks and spreadsheets. Notebook interfaces like Jupyter use an analogy of pages in a notebook that consist of a block of code, as well as an output for the block like a table, visualization, or documentation text. Blocks are part of a continuous program, allowing a user to quickly probe intermediate state or to safely insert hypothetical, exploratory modifications by adding or disabling pages. Spreadsheets give users an infinite 2-dimensional grid of cells that can hold either constant values or computed values derived from other cells. Instead of classical programmatic specification of bulk, set-at-a-time operations, spreadsheets use the metaphor of copying code and relative, positional data dependencies to "map" operations over collections defined by contiguous regions of data. Thus, the ability to change any value anywhere in the execution process, and simple integrated visualizations combine to make spreadsheets a very viable tool for data curation and exploration. The simplicity of spreadsheets has encouraged many database-driven efforts to resolve the impedance mismatch between positional and set-at-a-time query semantics [132,160], make spreadsheets more structured [16, 17] or make databases more spreadsheet-like [131]. Vizier builds on these efforts, creating a hybrid notebook-spreadsheet interface by making a notebook's output dynamic. Vizier's users can edit tables and visualizations directly, and have those edits reflected in the notebook's code block through database-style table updates. As a result, the user's edits, however they are applied, are recorded in the notebook as a part of the workflow (see Figure 2b). Although we will not reproduce the full spreadsheet interface entirely, our goal is to replicate as many of the flexible data and schema manipulation features of spreadsheets as possible. Vizier allows users to overwrite arbitrary values, embed formulas into table cells, cut/copy/paste cells, add (or delete) columns or rows, and sort or filter the data. In addition to low-level modifications, the

user can also apply higher-level curation operations, including ones for which exact configuration parameters may not be known ahead of time (we return to this later). Unique to Vizier is its ability to present useful recommendations to the user based on provenance, to expose ambiguity through visual hints in the interface, and to explain why and how values where derived.

Incrementally building curation workflows. The workflows that are *incrementally* constructed based on the user's operations serve several purposes. By leveragering VisTrails [42, 99, 202] as a system to manage these workflows and their evolutionary construction through user actions [43], they allow seamless repeatability [9, 10]. This enables users to easily identify and revert erroneous changes and helps Vizier to learn from the user's activities. To support a user in efficiently revising a workflow, we can use provenance to propagate changes to data and workflow steps throughout dependent workflow stages. Techniques such as program slicing [2, 153, 230] and incremental view maintenance [3, 35, 38, 50, 118, 135, 138, 145] can help to improve the performance of this process. It can be further optimized in the case where update histories are available through a technique called reenactment which is readily supported in GProM [9, 10].

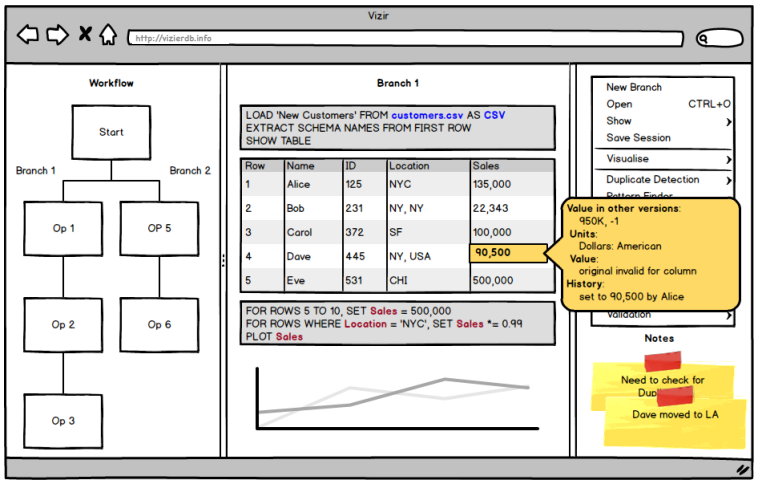

Figure 3: An example of Vizier's UI

Furthermore, a workflow generated for one dataset can be easily adapted for use with different, but similar data sets, a task for which research by PI Freire [200] provides a preliminary approach. Inputs can take the form of CSV files, relational database tables, JSON files, unstructured data, or external data sources like web APIs or HTML pages. Workflows are part of the extensive provenance tracking of Vizier [51, 65], helping the user and her collaborators to explain the system's outputs and audit their curation and exploration activities. Unique to Vizier is that workflow steps are not considered as black boxes, but correspond to programs in a declarative language described below. Vizier also supports ambiguity-tolerant operators called lenses and tracks fine-grained provenance at the data level.

Exploration, testing, and deployment. As mentioned earlier, a user would typically want to design a workflow over a small sample dataset before testing or deploying it at scale. One of the compelling benefit of workflows is the ability to adapt and re-use them across different contexts. A data curation process *designed through an exploration-friendly spreadsheet* interface can then be adapted to other datasets, replicated for recurring data curation tasks like monthly reports, or run at scale on a cluster. Like Wrangler [121, 134], Vizier allows users to easily develop data curation workflows on a small sample of the data before scaling to tera- or peta-byte scale data on a Hadoop cluster or RDBMS. In contrast to Wrangler which forces users to generalize edits upfront, Vizier instead helps users to generalize a workflow that includes manual curation steps into a workflow suitable for the complete dataset.

Curation DSL with Provenance. Vizier's flexibility is derived, in part, from a new domain-specific language for exploratory curation that we will develop as part of the proposed work. The objective of this DSL is to *integrate closely with the interface* by ensuring a 1-1 mapping between operations in the language and edits to Vizier's tabular outputs. In addition to facilitating the hybrid notebook/spreadsheet environment, the Vizier DSL serves as an intermediary between the workflows and a variety of back-end execution environments. In addition to executing locally, we will develop modules to compile the Vizier DSL down to a variety of languages for deployment on external computing platforms, such as SQL for Relational DBs or Map/Reduce or Spark programs for cloud computing clusters. To maximize compatibility with GProM and Mimir, Vizier's DSL will start as an instance of classical extended-relational algebra [123]. For

a more user-friendly imperative flavor and to make it easier to mirror manual edits from the spreadsheet onto the target program, we will add operations from SQL's DML that edit specific fields, batches of fields, and insert or delete rows; operations from SQL's DDL that add, remove, and hide columns or manipulate constraints; and automated data curation operations based on Mimir's lenses, as discussed below. In spite of the imperative flavor of the language, these operators effectively modify a table-valued query "object" modelling the whole sequence of operations [9, 10], and can be thought of as operators in a relational monad [39] that can be reduced to a single query.

Taming uncertainty and explainability. Lenses [167,232], part of the Mimir system discussed below, are data curation operators that use heuristics to minimize or eliminate configuration. In lieu of configuration, a lens is allowed to produce multiple ambiguous *possible outputs* alongside a single *best guess output*. Possible outputs remain associated with the output of queries or transformations over lenses, allowing potentially ambiguous results to be identified and explained and allowing the uncertainty stemming from this ambiguity to be quantified. For example, depending on the user's goals outliers like the $938 tip in the TLC dataset might be kept as-is, elided from the dataset, or replaced with a reasonable educated guess. In the latter case, there are also numerous heuristics that can approximate a user's best guess, including sequential interpolation along each of several candidate dimensions [59,67,136], a classifier trained on the adjacent attributes [228], a classifier that treats the data as the output of a Markov process [157,158], and numerous others. A user could be asked to select from among these options upfront, but she may not have sufficient information about the data to decide which is relevant (e.g., is the tip a data error or a legitimate outlier). Furthermore, depending on the user's goals, the choice may not even be relevant (e.g., she is generating a table of average tips by year). Workflows also provide context for data presented through the notebook interface. As illustrated in Figure 3, users can obtain statistics about output values through the Vizier UI, including the record's dependencies and formula, variations in the record's value over time, the system's certainty in the result (discussed further below), and other features like the set of of edits with the greatest impact on the value.

Automation and Recommendation. In keeping with best practices for user interface design [177], a goal of Vizier's interface is retaining the user's sense of control over the system's behavior. Accordingly, we adopt two general strategies for streamlining the user's interaction with the system that we refer to as Automation and Recommendation. Automation allows users to accomplish complex high-level tasks without concern for the low-level details of the task. Effective use of automation requires both the user's consent and awareness, as well as effective communication in the case of ambiguities. Automation in Vizier occurs through lenses. Lens behaviors are precisely defined in terms of desired effects. By requesting that a lens be applied, a user indicates both consent and an awareness of what the lens is trying to accomplish. As discussed above, ambiguities are communicated through explanations over results, minimizing upfront effort from the user, but keeping them aware of potential repercussions of using automated tools. By integrating these techniques into Vizier we enable users to combine these techniques with each other and with simpler manual curation operations (e.g., manually deleting dirty rows). Even more important, through their integration with Vizier, these operations can benefit from the uncertainty, provenance, recommendation, and deployment features of Vizier. Recommendations, instead help users to quickly reach relevant curation tools and to discover tools that they may not be aware of. Vizier uses a repository of collected provenance to provide suggestions to the user based on the current workflow and data characteristics. This is similar to Wrangler [134] which trains a Markov model on sequences of transformations to suggest transformations commonly performed together in sequence. However, our recommendations are much more advanced in that they are not just based on the current structure of the workflow, but also on the fine-grained data and workflow provenance. Four types of recommendations will be supported: (1) Recommendations on how to tune the parameters of a cleaning operation in the workflow. For example, the sensitivity of an outlier detection step could be tuned based on successful values for past outlier detection methods over data with similar characteristics. Similarly, if a past workflow containing a particular data curation operation has a similar structure as the current workflow this can also be an indication that similar tuning parameter values should be applied; (2) Recommendations on which data curation steps to apply next. These recommendations will be based both on the characteristics of the workflow as well as the data. (3) Recommendations on how to generalize specific curation steps, e.g., updating values based on a condition instead of updating a fixed set of rows. (4) Finally, Vizier offers

facilities for automated data discovery, both through simple keyword search, and offering suggestions based on datasets frequently used together.

2.1 Comparison to Existing Tools

Spreadsheets are an extremely common tool for data curation. Vizier borrows the structural and visual programming elements of spreadsheets, allowing users to freely interact with data and gracefully supporting corner cases. Although Vizier will not provide full freedom of spreadsheets, curation efforts on spreadsheets can not be easily generalized into repeatable, shareable workflows. Scripting languages like Python are another common tool for curation work, but suffer from several limitations that Vizier addresses. First, the link between output and code is unidirectional: Edits to the output are overwritten the next time the code runs, making it harder to apply one-off changes or to explore hypothetical what-if scenarios. Moreover, once a curation script is developed through interactive design, it must still be manually adapted for parallel execution via map/reduce or a tool like apache spark [212].

Recently, several new tools for data cleaning or "wrangling" have emerged from academia and industry and are gaining traction in the data science community. NADEEF [61, 73, 74, 82] uses a rule-based system to detect and repair violated constraints. The set-at-a-time interaction model of a rule-based system works well when a user knows what properties the data should satisfy upfront, but does not permit easy discovery of such properties as in Vizier. SampleClean [122, 154, 229] uses sample-based cleaning to make unbiased estimates for aggregate results over messy or dirty data; This technique is orthogonal to Vizier's workflow generalization, and could conceivably be eventually incorporated into Vizier. Habitat [130] is an extensible framework for data curation, provenance, and discovery, into which specialized modules and interfaces can be deployed; Vizier could eventually be adapted to use Habitat as a deployment target. The Data Tamer project [120, 213], commercialized as Tamr, focuses more on data integration issues like schema matching and entity resolution. These are both important tasks in data curation, and appear as operators in Vizier.

The most similar production system is the Wrangler [121, 134, 185] project, commercialized as Trifacta [220]. Here too, the goal is to develop a repeatable curation workflow. We borrow Trifacta's sample-based development model, its ability to apply global edits directly on its output, as well as the idea of predictive suggestions. However, Vizier distinguishes itself in four ways. First, Trifacta forces users into the clean-before-use model and is thus optimized for developing only generalized set-at-a-time cleaning programs. Second, Vizier's notebook metaphor allows users to explore the data simultaneously from multiple perspectives or with different hypothetical modifications. Third, Vizier tracks provenance through the workflow, making it easier to debug values and sanity check results. Finally, provenance also permits Vizier to *safely* automate common tasks, even in the presence of ambiguity; if they are later found to be incorrect, the operations can be undone.

2.2 VisTrails

The open-source VisTrails system was designed to support exploratory and creative computational tasks [41, 86, 100, 133, 204, 210, 224, 225], including data exploration, visualization, mining, machine learning, and simulations. The system allows users to create complex workflows (computational processes) that encompass important steps of data science, from data gathering and manipulation to complex analyses and visualizations. A new concept we introduced with VisTrails is the *provenance of workflow evolution* [40, 98]. In contrast to previous workflow and visualization systems which maintain provenance only for derived data products, VisTrails treats workflows as first-class data items and maintains their provenance (see *version tree* in Figure 4). Workflow evolution

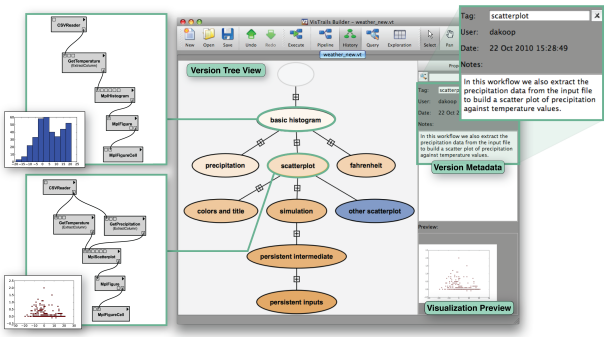

Figure 4: In VisTrails, provenance of exploration is represented as a version tree. Each node represents a workflow and an edge between two nodes encodes the set of changes applied to the parent to derive the child.

provenance supports reflective reasoning, allowing users to store temporary results, to make inferences from stored knowledge, and to follow chains of reasoning backward and forward [176]. It also works as a *version control system for computational pipelines*, thus naturally supporting collaboration [78]. Users can easily *navigate through the space of workflows* created for a given investigation, visually *compare workflows and their results*, and *explore parameter spaces* [98]. Users are encouraged to *re-use knowledge by exploring and leveraging provenance information* through specific, easy-to-use components. These include a query-by-example interface, a mechanism for refining workflows by analogy [203], and a recommendation system that aids users in the design of workflows [152]. Pipelines and their provenance can be shared through Web-based interfaces, allowing *others to validate and reproduce computational experiments* [161, 192, 197]. The system has an active community of developers and users and has been adopted in several scientific projects, both nationally and internationally, across different domains. It has enabled and supported new research in environmental science [18,48,58,127,129], psychiatry [8], astronomy [219], cosmology [7], high-energy physics [68], molecular modeling [124], quantum physics [29, 85], earth observation [60, 222] and habitat modeling [165]. Besides being a stand-alone system, VisTrails has been used as a key component of domain-specific tools including DoE's UV-CDAT [191,195,221]; USGS's SAHM [165,190]; and NASA's DV3D [227]. VisTrails was featured as an NSF Discovery [178].

2.3 Mimir

Mimir [167, 231, 232] is a system that extends existing relational databases with support for so-called on-demand, or pay-as-you-go data curation through lenses, already introduced above in the description of Vizier. Mimir's support comes through a form of an annotated cursor that identifies ambiguous values and rows who's presence in the result set is ambiguous. Furthermore, the annotated cursor can explain ambiguity, both through English statements like *"I replaced `TIP` with `NULL` on row 5239865 because you asked me to remove outliers,"* and by quantifying the effect of ambiguity on results, as in *"the `AVERAGE(TIP)` is* 10 ± 2 *with 95% confidence"*. To convey the output of this annotated cursor to users, Mimir's front-end interface (illustrated in Figure 5) provides a standard SQL interface (a). Ambiguous outputs are marked (b), and in addition to showing their lineage (c), Mimir produces explanations (d) upon request.

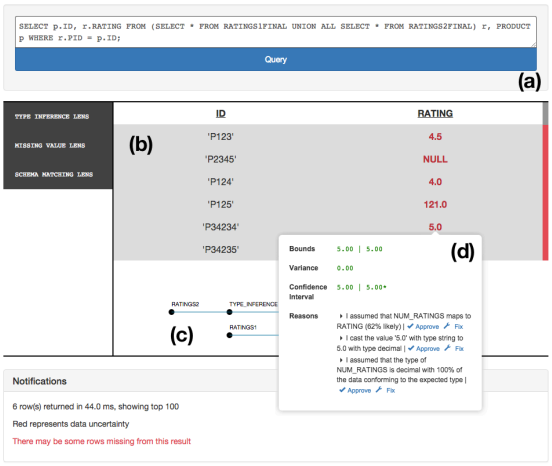

Figure 5: Mimir's User Interface shows and explains uncertain or ambiguous data

2.4 GProM

GProM (**G**eneric **Pro**venance **M**iddleware) [10, 11, 113, 174, 175] is a database-independent middleware for tracking and querying fine-grained provenance of database queries, updates, and transactions (see Figure 6). GProM supports multiple database backends and can be extended to new backends through plugins. The user interacts with the system through one of the system's declarative frontend languages (currently dialects of SQL and Datalog with constructs for requesting provenance). Importantly, GProM tracks data provenance in a non-intrusive manner, without requiring any changes to applications or the database backend. Using **reenactment** [10], a declarative replay technique which simulates the effects of past updates or transactions using queries with time travel, the provenance of an update or transaction is computed retroactively through replaying the operation instrumented to capture provenance. Time travel is supported by most commercial systems and can be implemented using, e.g., triggers in systems

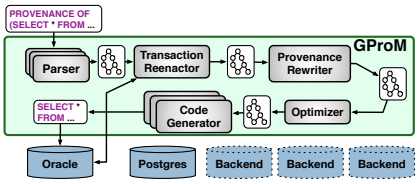

Figure 6: GProM system overview

that do not support it natively (see, e.g., [211]). GProM will be used to supply fine-grained provenance for curation operations be they queries or updates. Vizier uses this functionality to generate explanations for curated data and the curation process as well as to extract training data for the recommendations based on provenance. Furthermore, using reenactment it is possible to efficiently propagate changes to data and operations through a workflow during the exploratory phase of curation workflow construction.

3 Management Plan

Our implementation strategy for Vizier is to gradually integrate these three systems into a front-end interface that combines elements of VisTrails and Mimir, and a back-end component combining elements of Mimir and GProM. Our approach is iterative: many of our deliverables are useful stand-alone data curation tools in their own right that become progressively more powerful as they are extended and combined.

Task 1: Bulletproof GProM and Mimir. Both GProM and Mimir are stable enough for general use, but have primarily been developed as proofs of concept. The first task will be to thoroughly stress-test both systems to identify and repair any bugs, and to evaluate whether any critical core functionality necessary for Vizier is missing from either system.
Deliverable: *Production quality releases of GProM and Mimir*

Task 2: Unify GProM and Mimir. Our first integration target will be the back-end for Vizier: a data processing system with fine-grained provenance support, able to link outputs to edits applied both by users and by automated curation tasks. The back-end component will combine GProM's reenactment facilities with support for Mimir's lenses and explanations of ambiguity.
Deliverable: *A provenance and ambiguity-aware data processing system supporting queries and updates*

Task 3: Add a notebook Interface to VisTrails. VisTrails will serve as a central dispatcher for Vizier, linking the front-end interface to the fine-grained provenance and data-processing capabilities of the backend developed in Task 2. In this task, we will extend VisTrails with a UI shell based on Jupyter notebook, which will eventually serve as a front-end for Vizier. For this task, like VisTrails, the notebook shell will remain agnostic to the implementation of workflow steps.
Deliverable: *A notebook-style UI for VisTrails*

Task 4: The Vizier DSL. In parallel with the previous tasks, we will begin a research effort to design a DSL for our hybrid notebook/spreadsheet environment. As noted, our starting point will be relational algebra with additional operations based on SQL's DDL and DML, Mimir's lenses, as well as operators from existing data curation systems like Wrangler or NADEEF. Our goal is to create a language with a 1-1 mapping between edits applied to notebook's tabular output and the program generating that output.
Deliverable: *A language specification for Vizier's DSL and compliant parser*

Task 5: A Prototype of Vizier's Hybrid Notebook and Spreadsheet UI. As the DSL and the front- and back-end components mature, we will begin integrating them together using the DSL as glue. The first part of the integration process will be to create the hybrid notebook/spreadsheet UI proposed earlier by enabling spreadsheet-style editing for the notebook's tabular outputs. We then also need to create the UI elements necessary to mirror edits between the spreadsheet and code elements. Time permitting, we will explore visual cues to help users to follow the flow of code or data dependencies.
Deliverable: *A prototype front-end for Vizier*

Task 6: Implement the DSL in the Vizier Back-End. Simultaneously, we will add DSL support to the back-end. Both GProM and Mimir use an intermediate representation similar to relational algebra with updates. We anticipate that it will be possible to implement the DSL with minimal extension to this basic model (e.g., GProM already supports multiple front-end languages and, thus, it is reasonable to assume that adding support for the DSL would be possible with reasonable effort). Our initial plan will be to support local evaluation of the DSL on common formats including CSV and JSON, first to permit low-latency development, and second to create a prototype as quickly as possible. We will return to add support for partial or full deployment to external data management systems later (Task 8).
Deliverable: *A back-end with support for the Vizier DSL.*

Task 7: Link the Back- and Front-end components. The DSL developed in Task 4 serves as a common interface between the front- and back-end components of Vizier. Tasks 5 and 6 prepare these components for integration, establishing a DSL-aware front-end, and a back-end capable of evaluating DSL programs. We anticipate this task to require only minor refinements, but extensive stress-testing.
Deliverable: *A beta version of Vizier*

Task 8: External Datasource Connectors. Our next goal will be to enable access to data hosted in distributed resources such as relational databases, HDFS clusters, NoSQL data stores, generic REST APIs and other internet resources. A connector will require three components: First a way to access the complete data, second a way to sample from the available data, and finally an optional DSL translator to allow Vizier to deploy workflows to it.
Deliverable: *Connectors between Vizier and external data sources*

Task 9: Extending Mimir's Library with Automated Data Curation Operations. With the intent of getting users up and running as quickly as possible, our next goal will be to extend Mimir's existing library of Lenses with new data curation operations as well as operations that evaluate data quality and unearth interesting features of the data. Our initial goal will be to develop three modules: One for identifying potential outliers in a dataset [119], one for detecting correlations or dependencies between attributes and/or successive rows, and one for suggesting visual representations of a dataset [151]. We will revise this list based on feedback from our user community collaborators.
Deliverable: *Extended library of automated data curation operations and quality evaluation modules*

Task 10: Provenance-based Suggestions. Vizier will track the user's activities as a way to explain outputs to the user. The system will leverage the resulting repositories of both workflow provenance and fine-grained data provenance to offer suggestions to users [152,200], e.g., by identifying common relationships between user edits [134] or common properties of data elements that the user appears to be trying to fix.
Deliverable: *Automatic provenance-based recommendation module*

Task 11: Revise and Refine. We anticipate active community involvement in the development of Vizier. However, once we have assembled a beta version of the complete system, we expect to receive a substantially higher volume and quality of feedback from our collaborators and users. Our timeline includes an explicit period in which we expect to focus entirely on revising and refining the system.
Deliverable: *A full release of Vizier*

Task 12: Version Tree Operations. The change-based provenance adopted by VisTrails enables a series of operations that streamline exploration, including the ability to visually compare workflows [98] and to modify workflows by analogy [203]. We will extend VisTrails to support additional operations, including the ability to merge/reconcile curation workflows and propagate updates to multiple workflows.
Deliverable: *A comprehensive set of version tree manipulations required by exploratory curation*

3.1 Development Timeline

We have prepared a 3-year development timeline. In addition to the 3 PIs and a senior research engineer responsible for the project, we have budgeted for 4 students, 1 post-doctoral associate, and 2 developers for the full 3 years. The post-doc and developer hosted at NYU will be responsible for the implementation of Tasks 3-5, 7, 9-12. One developer hosted at UB will be implement the components in Tasks 1-2, 4, 6-8, 11. Students, one hosted at NYU and UB each, and two hosted at IIT, and the post-doc will be responsible for research & development components of the proposal. The post-doc will also serve as a bridge to our collaborators and work on outreach, giving talks and tutorials about the system. A yearly demonstration of technical capabilities, as requested by the DIBBS solicitation, will be based on the deliverables described for each task above, grouped into yearly units of work as described below.

Year 1. Preliminaries: Bulletproofing and unifying GProM and Mimir (Tasks 1-2) and preparing the user interface (Task 3). While these tasks are underway, all three sites will collaborate on a research effort to design a notebook/spreadsheet DSL (Task 4). Towards the end of the year, we expect both developers to begin prototyping efforts for their respective parts of Vizier (Tasks 5-6).

Year 2. Continued efforts to prototype the Front- and Back-end components of Vizier: We expect to see early efforts to integrate both elements (Task 7) begin early-to-mid year 2, and a beta version of the system

available late in year 2. Research efforts in year 2 will include a preliminary exploration of provenance-aware learning and heuristic quality assessment (Tasks 9 and 10), and performance-tuning, for example through techniques like reenactment and incremental maintenance (Task 11).

Year 3. Year 3 will serve as a buffer for time-overruns and as a period of refinement: improving compatibility (Task 8), making Vizier smarter (Tasks 9 and 10), and incorporating community feedback (Task 11) and adding versioning (Task 12).

3.2 Evaluation Strategy

As our goal is to simplify data curation, we will evaluate Vizier and its components primarily through community feedback and expert studies facilitated by our collaborators. We will also conduct user studies to evaluate the effectiveness of specific features or interface elements on an as-needed basis, pending IRB approval for each case. As performance is also a significant user concern, we will evaluate Vizier's back-end components using data from collaborators including Airbus and the NYC Taxi and Limousine Commission, standard benchmarks like the MayBMS probabilistic data benchmark [217], openly accessible datasets, and benchmarking tools for cleaning and integration [12–15].

With the exception of Tasks 4–6, each sub-goal results in a complete software artifact that can be evaluated in isolation. Tasks 1–3 result in stand-alone components with clear, self-contained goals. Tasks 7–12 result in versions of Vizier with progressively more features that can be evaluated through expert studies and deployment at our collaborator's sites. In each case, the task has clear, measurable deliverables. The interface design of Task 5 will be evaluated and refined primarily through expert studies, and IRB approval permitting, user studies as well. The back-end design of Task 6 will be evaluated in terms of its performance and compliance with the DSL's specifications. The DSL created in Task 4 will be evaluated and refined based on its suitability for Tasks 5 and 6.

3.3 Risk Mitigation

The primary elements of Vizier are already well-established systems and we do not anticipate any significant issues with the feasibility of the proposed system. One possible risk is time overruns due to unforeseen difficulty or unexpected events. To mitigate this risk we have budgeted a part of year 3 as a buffer. If absolutely necessary, we can also scale back the goals of Tasks 8-10 to ensure the delivery of a stable system by the end of the grant period. Another risk is that of a developer leaving. We will mitigate this potential risk to institutional knowledge by ensuring that developers actively document their efforts, holding regular code reviews, and ensuring that the student researchers are actively engaged in the design process with the developers. A third risk is the possibility that by the time a prototype of Vizier is ready, our collaborators will no longer be able to provide us with their expected contributions. To mitigate this risk, we have engaged a large, diverse group of collaborators and will continuously seek out new partnerships over the course of the grant. Finally, the risk of running out of development resources is addressed in our sustainability plan.

4 Sustainability Plan

To ensure the continuation of the project beyond the three year grant period, we will build on our existing efforts to establish a strong community of users and developers with a vested interest in Vizier. To this end, we have already engaged a diverse set of expert leaders from communities in industry, academia, and government, all interested in actively participating in the design of Vizier and helping us to evaluate and refine the results. Our preliminary set of collaborators includes representatives from NYU's Center of Urban Science and Progress (CUSP), Furman Center, The Airbus group, the NYC Taxi and Limousine Commission (TLC), and Oracle (see letters of collaboration). Oracle, in particular, has already been supporting PIs Kennedy and Glavic's research through unrestricted gifts for the past 2 and 3 years, respectively. Their collaborators at Oracle, Dieter Gawlick and Zhen Hua-Liu, actively provide feedback and guidance for the Mimir and GProM projects. We have also begun efforts to reach out to additional communities (e.g., the informal survey in Section 1.3). We will expand on these efforts as additional components of Vizier are completed through further surveys, workshops, webinars, and demonstration videos. We have also budgeted for a publicly-available demonstration release to be hosted on a cloud-computing platform.

Our primary approach to building community buy-in will be to engage prospective users in the design and development of Vizier. The first step in this process will be to release Vizier under a permissive license (e.g., Apache), and provide access via a public code repository such as GitHub. However, simply releasing code is insufficient to build an engaged community. Ensuring high-quality user and developer documentation, as well as well-structured, well-commented, and readable code will help newcomers to make use of Vizier, and eventually to extend it as well. A stable community also needs communication. We will ensure the availability of vectors for both asynchronous communication (e.g., wikis or bug/feature request trackers) and synchronous communication (IRC, Slack, or similar) with project staff. We note that the use of such tools for communication is already necessary to support collaboration between the three project sites and that making these resources publicly accessible presents a negligible overhead. Finally, to provide the community with effective leadership, we will establish a common set of guidelines for code review and community governance during our first annual on-site meeting, and will review these guidelines during subsequent annual meetings.

5 Collaboration Plan

The PIs will closely collaborate on the project and synchronize their work progress using bi-weekly virtual meetings. These meetings will be attended by the PIs, the developers funded through this project, involved Ph.D. students and PostDocs, and collaborators from the target communities (see letters of collaboration). The purpose of these meetings is to (1) plan short-term and long-term implementation efforts, (2) elicit feedback from community leaders on the current version of the Vizier system, (3) strategic planning of development direction, and (4) coordination among all involved project sites. All three PIs have extensive experience working with users from the sciences and industry and have been involved in collaborations that span multiple disciplines. The PIs and project staff will also meet in person regularly at conferences (funds for conference travel are included in our respective budgets), and at a yearly on-site meeting. We have budgeted travel funds to allow PIs, developers, students, PostDocs, and key collaborators from target communities to gather at a single (rotating) project site for a workshop event once per year. These yearly meetings will revolve around a series of lightning talks from project staff to convey detailed information about each group's activities, areas of expertise, and project requirements. The lightning talks will be followed by a strategic planning session, a review of project and community governance policies, a breakout session for free-form discussions among project participants, and a wrap-up discussion to discuss outcomes arrived at during the break-out sessions. The PIs are using web-based collaboration software called GitLab that provides document sharing, a collaborative wiki, a bug tracker, and other tools for group management, as well as other cloud collaboration tools like Dropbox and Skype. If funded, project staff will also deploy and use tools for community-management, including group chat software (e.g., Slack, IRC, or similar), forums (e.g., PHPBB or Disqus), and blogging software (e.g., Wordpress). Finally, when Vizier is ready for a preliminary release, we will begin hosting it on a public open-source code repository such as GitHub.

6 Broader Impacts

The overarching goal of this proposal, to simplify data curation and exploration, has the potential to be transformative and positively impact the state of data science in many different domains. Our initial set of collaborators (see attached letters) includes representatives from government (TLC), academia (NYU's CUSP and Furman Center), and industry (Oracle and Airbus), all heavily invested in the issues of data quality that we are poised to address. Our informal survey turned up numerous colleagues and collaborators from across multiple fields of research and sectors of industry, all of whom grapple with data quality on a regular basis. Thus, the impact of the project can be immediate. Besides advancing the state of the art in data curation, the proposed work will improve research in other fields, notably, in social sciences. It will also contribute to an emerging field: urban science. Through the deployment of our tools within CUSP and the Furman Center, our work has the potential to impact a wide range of scientists and students, as well as the NYC agencies that are partnering with CUSP and Furman on various projects. If successful, our work will positively impact government by improving data quality, which in turn will lead to better planning and policies and more efficient operations. The ability to publish provenance-rich, high-quality data will also positively impact governance and citizen engagement.

Integration of Research and Education. We see education in a broader context. In our interactions with

domain scientists, it has become clear that many of them are not up-to-date of recent developments around data curation techniques and tools. Thus, we believe that we not only need to educate our own students, but also inform the scientific community at large of the benefits and technologies related to curation. We are well positioned to reach "across our disciplines", given our on-going multi-disciplinary collaborations and the make up of our team. Four PhD students and one Postdoctoral researcher will participate in this project. The grant will also support the PIs in their existing educational outreach efforts: PI Kennedy is working with contacts at local high schools, gained through his membership in the WNY chapter of the ACM Computer Science Teacher's Association, to develop an open-data after-school program at high schools near UB. PI Glavic is actively working on integrating data cleaning and integration into IIT's CS graduate programs including the development of a new course on data integration and provenance. Vizier will be used as a tool in future installments of this course.

Minority and Undergraduate Involvement. We are committed to recruiting and mentoring minorities. All the PIs have been involved in different efforts to foster minority involvement. NYU Tandon School of Engineering ranked #1 in the nation by US News and World Report in Racial Diversity, and #3 in Economic Diversity among private universities in 2009. Typically, around half of our incoming domestic CS Ph.D. students are women or underrepresented minorities. It is well known that attracting minorities to STEM fields has been a great challenge. We believe that our research topic can contribute to attracting more minorities to computer science. The M.S. progrems at NYU CUSP and at NYU CDS (which PI Freire directs) have close to 50% female enrollments. If funded, the PIs will also apply for REUs to facilitate involvement from undergraduate researchers at their respective universities.

Technology Transfer and Software Tools. Our team has an unassailable record of translating research results into practice across the sciences, including urban science, as well as to government, industry, and the general public. The software we will develop will be released as open source. Value-added open data derived by our methods will also will be made available under creative commons licenses where permitted.

7 Results from Prior NSF Support

Dr. Oliver Kennedy. Dr. Oliver Kennedy has been tenure-track faculty since Fall of 2012 and has been a PI on one NSF award. *Intellectual Merit:* Since receiving his first award 1.5 years ago, Dr. Kennedy's funding has resulted in 2 workshop publications [155,156] and one further conference paper under submission. *Broader Impacts:* NSF: CNS-1409551 (2014-2018; $976k) has funded the development of Ettu, a tool for summarizing query logs to enable insider threat analysis, as well as the collection and anonymization of a one-week trace of *all* SQL query activity at a major US bank. The project actively supports three PhD students, and an REU supplement is actively supporting one undergraduate and supported one undergraduate student who graduated in Winter of 2015. One graduate student has completed his thesis under Dr. Kennedy's supervision. Dr Kennedy is currently advising 5 PhD students (including 2 female students), one MS student, and one BS student, and co-advising 3 PhD students (including one female student).

Dr. Juliana Freire. Since joining academia in late 2002, Dr. Freire has been PI, co-PI, or senior investigator on fourteen NSF awards. *Intellectual Merit:* Dr. Freire's NSF funding has resulted in over 87 publications [4, 6, 18–27, 29, 30, 37, 40, 41, 44, 52–56, 63, 66, 71, 72, 75, 78–81, 83, 84, 86–88, 90, 91, 93, 94, 96–98, 100, 104, 127, 146–150, 152, 152, 159, 161–164, 166, 169–172, 172, 173, 186, 187, 193–199, 201, 203, 204, 206–208, 210, 223, 226], including a IEEE Visualization 2007 best paper award and an Eurographics Educational Program best paper award. *Broader Impact:* NSF IIS-0513692 (2005-2008; $499k)) [95] and NSF CNS-1405927 (2014-16; $530k) [92] have funded the development of VisTrails [86,224], an open-source data analysis and visualization tool that provides a comprehensive provenance infrastructure. VisTrails has been adopted in several scientific projects, both nationally and internationally, including in large NSF-funded projects [57, 58, 62, 188]; and it has had impact in different scientific domains [7,8,18,48,58,68,124,127,129,219]. Thirteen graduate students have completed their degrees under Dr. Freire's supervision—these include five female and three Hispanic students. She has also supervised post-doctoral assistants and several undergraduate students. She currently advises 7 Ph.D. students, 5 of which are from under-represented minorities.

Dr. Boris Glavic. does not have NSF support yet. He is currently advising three Ph.D. students (one female), three MS students, and is co-advising one Ph.D. student.

7 References

[1] ACHE. `https://github.com/ViDA-NYU/ache`.

[2] Hiralal Agrawal and Joseph R. Horgan. Dynamic program slicing. *SIGPLAN Not.*, 25(6):246–256, June 1990.

[3] Yanif Ahmad, **Oliver Kennedy**, Christoph Koch, and Milos Nikolic. DBToaster: Higher-order delta processing for dynamic, frequently fresh views. *PVLDB*, 2012.

[4] Sihem Amer-Yahia, Fang Du, and **Juliana Freire**. A Comprehensive Solution to the XML-to-Relational Mapping Problem. In *Proceedings of ACM WIDM*, pages 31–38, 2004.

[5] Y. Amsterdamer, S.B. Davidson, D. Deutch, T. Milo, J. Stoyanovich, and V. Tannen. Putting Lipstick on Pig: Enabling Database-style Workflow Provenance. *Proceedings of the VLDB Endowment*, 5(4):346–357, 2011.

[6] Erik Anderson, Steven P. Callahan, David A. Koop, Emanuele Santos, Carlos E. Scheidegger, Huy T. Vo, **Juliana Freire**, and Cláudio T. Silva. *VisTrails: Using Provenance to Streamline Data Exploration. In *Poster Proceedings of the International Workshop on Data Integration in the Life Sciences (DILS)*, page 8, 2007.

[7] Erik W. Anderson, James P. Ahrens, Katrin Heitmann, Salman Habib, and Cláudio T. Silva. Provenance in comparative analysis: A study in cosmology. *Computing in Science and Engineering*, 10(3):30–37, 2008.

[8] Erik W. Anderson, Gil A. Preston, and Cláudio T. Silva. Towards development of a circuit based treatment for impaired memory: A multidisciplinary approach. In *IEEE EMBS Neural Engineering*, 2007.

[9] Bahareh Arab, Dieter Gawlick, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Reenacting transactions to compute their provenance. Technical Report IIT/CS-DB-2014-02, Illinois Institute of Technology, 2014.

[10] Bahareh Arab, Dieter Gawlick, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Formal foundations of reenactment and transaction provenance. Technical Report IIT/CS-DB-2016-01, Illinois Institute of Technology, 2016.

[11] Bahareh Arab, Dieter Gawlick, Venkatesh Radhakrishnan, Hao Guo, and **Boris Glavic**. A generic provenance middleware for database queries, updates, and transactions. In *Proceedings of the 6th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2014.

[12] Patricia C. Arocena, Radu Ciucanu, **Boris Glavic**, and Renée J. Miller. Gain Control over your Integration Evaluations. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 8(12):1960 – 1971, 2015.

[13] Patricia C. Arocena, **Boris Glavic**, Radu Ciucanu, and Renée J. Miller. The iBench Integration Metadata Generator. *Proceedings of the VLDB Endowment (PVLDB)*, 9(3):108–119, 2015.

[14] Patricia C. Arocena, **Boris Glavic**, Giansalvatore Mecca, Renée J. Miller, Paolo Papotti, and Donatello Santoro. Messing Up with Bart: Error Generation for Evaluating Data-Cleaning Algorithms. *Proceedings of the VLDB Endowment (PVLDB)*, 9(2):36–47, 2015.

[15] Patricia C. Arocena, **Boris Glavic**, and Renée J. Miller. Value invention for data exchange. In *Proceedings of the 39th International Conference on Management of Data (SIGMOD)*, pages 157–168, 2013.

[16] Eirik Bakke and Edward Benson. The schema-independent database ui a proposed holy grail and some suggestions. 2011.

[17] Eirik Bakke, David Karger, and Rob Miller. A spreadsheet-based user interface for managing plural relationships in structured data. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2541–2550, New York, NY, USA, 2011. ACM. `http://doi.acm.org/10.1145/1978942.1979313`

[18] António Baptista, Bill Howe, **Juliana Freire**, David Maier, and Cláudio T. Silva. *Scientific Exploration in the Era of Ocean Observatories. *Computing in Science and Engineering*, 10(3):53–58, 2008.

[19] António Baptista, Todd Leen, Y. Zhang, A. Chawla, David Maier, Wu-Chi Feng, Wu-Chang Feng, Jon Walpole, Cláudio Silva, and **Juliana Freire**. Environmental observation and forecasting systems: Vision, challenges and successes of a prototype. In *Conference on Systems Science and Information Technology for Environmental Applications (ISEIS 2003)*, 2003.

[20] Denilson Barbosa, **Juliana Freire**, and Alberto Mendelzon. Information preservation in xml-to-relational mappings. In *Proceedings of XML Database Symposium (XSym)*, pages 66–81, 2004.

[21] Luciano Barbosa and **Juliana Freire**. Siphoning hidden-web data through keyword-based interfaces. In *Proceedings of the Brazilian Symposium on Databases (SBBD)*, pages 309–321, 2004.

[22] Luciano Barbosa and **Juliana Freire**. Automatically constructing collections of online databases (poster). In *Proceedings of CIKM*, pages 796–797, 2006.

[23] Luciano Barbosa and **Juliana Freire**. *Siphoning Hidden-Web Data through Keyword-Based Interfaces. *JIDM*, 1(1):133–144, 2010.

[24] Luciano Barbosa and **Juliana Freire**. *Siphoning Hidden-Web Data through Keyword-Based Interfaces: Retrospective. *JIDM*, 1(1):145–146, 2010.

[25] Luciano Barbosa and **Juliana Freire**. *Using Latent-Structure to Detect Objects on the Web. In *Proceedings of WebDB*, 2010.

[26] Luciano Barbosa, **Juliana Freire**, and Altigran Soares da Silva. Organizing hidden-web databases by clustering visible web documents. In *IEEE International Conference on Data Engineering (ICDE)*, pages 326–335, 2007.

[27] Luciano Barbosa, Hoa Nguyen, Thanh Nguyen, Ramesh Pinnamaneni, and **Juliana Freire**. *Creating and exploring web form repositories. In *SIGMOD Conference*, pages 1175–1178, 2010.

[28] Luciano Barbosa, Kien Pham, Cláudio Silva, Marcos Vieira, and **Juliana Freire**. *Structured Open Urban Data: Understanding the Landscape. *Big Data*, 2(3), 2014.

[29] B. Bauer et al. The alps project release 2.0: open source software for strongly correlated systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(05):P05001, 2011.

[30] Louis Bavoil, Steve Callahan, Patricia Crossno, **Juliana Freire**, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *VisTrails: Enabling Interactive Multiple-View Visualizations. In *Proceedings of IEEE Visualization*, pages 135–142, 2005.

[31] Vicki Been, Sam Dastrup, Ingrid Gould Ellen, Ben Gross, Andrew Hayashi, Susan Latham, Meghan Lewit, Josiah Madar, Vincent Reina, Mary Weselcouch, and Michael Williams. State of new york city's housing and neighborhoods, 2011.

[32] Vicki Been, Sam Dastrup, Ingrid Gould Ellen, Ben Gross, Andrew Hayashi, Susan Latham, Meghan Lewit, Josiah Madar, Vincent Reina, Mary Weselcouch, and Michael Williams. State of new york city's housing and neighborhoods, 2012.

[33] Aline Bessa, Fernando de Mesentier Silva, Rodrigo Frassetto Nogueira, Enrico Bertini, and **Juliana Freire**. *RioBusData: Outlier Detection in Bus Routes of Rio de Janeiro. In *IEEE Symposium on Visualization in Data Science*, 2015.

[34] The BirdVis System. `http://www.birdvis.org`.

[35] Jose A. Blakeley, Per-Ake Larson, and Frank Wm Tompa. Efficiently updating materialized views. *SIGMOD Rec.*, 15(2):61–71, June 1986.

[36] Michael R. Bloomberg and David Yassky. 2014 taxicab fact book. `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`, 2014.

[37] Philippe Bonnet, Stefan Manegold, Matias Bjørling, Wei Cao, Javier Gonzalez, Joel Granados, Nancy Hall, Stratos Idreos, Milena Ivanova, Ryan Johnson, David Koop, Tim Kraska, René Müller, Dan Olteanu, Paolo Papotti, Christine Reilly, Dimitris Tsirogiannis, Cong Yu, **Juliana Freire**, and Dennis Shasha. Repeatability and workability evaluation of sigmod 2011. *SIGMOD Record*, 40(2):45–48, 2011.

[38] O. Peter Buneman and Eric K. Clemons. Efficiently monitoring relational databases. *ACM Trans. Database Syst.*, 4(3):368–382, September 1979.

[39] Peter Buneman, Shamim Naqvi, Val Tannen, and Limsson Wong. Principles of programming with complex objects and collection types. *Theoretical Computer Science*, 149(1):3 – 48, 1995. Fourth International Conference on Database Theory (ICDT '92).

[40] Steve Callahan, **Juliana Freire**, Emanuele Santos, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *Managing the Evolution of Dataflows with VisTrails *(Extended Abstract)*. In *IEEE Workshop on Workflow and Data Flow for Scientific Applications (SciFlow)*, 2006.

[41] Steve Callahan, **Juliana Freire**, Emanuele Santos, Carlos Scheidegger, Cláudio Silva, and Huy Vo. *VisTrails: Visualization meets Data Management. In *Proceedings of ACM SIGMOD*, pages 745–747, 2006.

[42] Steven Callahan, **Juliana Freire**, Emanuele Santos, Carlos Eduardo Scheidegger, Claudio T. Silva, and Huy Vo. VisTrails: Visualization meets Data Management. In *SIGMOD '06: Proceedings of the 32th SIGMOD International Conference on Management of Data (demonstration)*, pages 745–747, 2006.

[43] Steven P Callahan, **Juliana Freire**, Emanuele Santos, Carlos E Scheidegger, Claudio T Silva, and Huy T Vo. Managing the evolution of dataflows with vistrails. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 71–71. IEEE, 2006.

[44] Steven P. Callahan, **Juliana Freire**, Carlos Eduardo Scheidegger, Cláudio T. Silva, and Huy T. Vo. *Towards Provenance-Enabling ParaView. In *IPAW*, pages 120–127, 2008.

[45] Sean Capperis, Jorge De la Roca, Ingrid Gould Ellen, , Brian Karfunkel, Yiwen (Xavier) Kuai, Shannon Moriarty, Justin Steil, Eric Stern Michael Suher, Max Weselcouch, Mark Willis, and Jessica Yager. State of new york city's housing and neighborhoods, 2014.

[46] Sean Capperis, Jorge De la Roca, Kevin Findlan, Ingrid Gould Ellen, Josiah Madar, Shannon Moriarti, Justin Steil, Mary Weselcouch, and Mark Williams. State of new york city's housing and neighborhoods, 2013.

[47] Daniel Castellani Ribeiro, Huy T. Vo, **Juliana Freire**, and Cláudio T. Silva. *An Urban Data Profiler. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 1389–1394. ACM, 2015.

[48] Cdat newsletter: Cdat v5.0 - highlights. http://www-pcmdi.llnl.gov/software-portal/Newsletter/Vol3/news.html, June 2007.

[49] Geoffrey Challen, Jerry Antony Ajay, Nick DiRienzo, **Oliver Kennedy**, Anudipa Maiti, Anandatirtha Nandugudi, Sriram Shantharam, Jinghao Shi, Guru Prasad Srinivasa, and Lukasz Ziarek. maybe we should enable more uncertain mobile app programming. In *HotMobile*, pages 105–110, 2015. `http://doi.acm.org/10.1145/2699343.2699361`

[50] Surajit Chaudhuri, Ravi Krishnamurthy, Spyros Potamianos, and Kyuseok Shim. Optimizing queries with materialized views. In *Proceedings of the Eleventh International Conference on Data Engineering*, ICDE '95, pages 190–200, Washington, DC, USA, 1995. IEEE Computer Society. `http://dl.acm.org/citation.cfm?id=645480.655434`

[51] Fernando Chirigati and **Juliana Freire**. Towards integrating workflow and database provenance. In *Provenance and Annotation of Data and Processes*, pages 11–23. Springer, 2012.

[52] Fernando Seabra Chirigati and **Juliana Freire**. Towards integrating workflow and database provenance. In *IPAW*, pages 11–23, 2012.

[53] Fernando Seabra Chirigati, **Juliana Freire**, David Koop, and Cláudio T. Silva. *VisTrails provenance traces for benchmarking. In *EDBT/ICDT Workshops*, pages 323–324, 2013.

[54] Fernando Seabra Chirigati, Dennis Shasha, and **Juliana Freire**. Packing experiments for sharing and publication. In *SIGMOD Conference*, pages 977–980, 2013.

[55] Fernando Seabra Chirigati, Dennis Shasha, and **Juliana Freire**. *ReproZip: Using Provenance to Support Computational Reproducibility. In *USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2013.

[56] Fernando Seabra Chirigati, Matthias Troyer, Dennis Shasha, and **Juliana Freire**. *A Computational Reproducibility Benchmark. *IEEE Data Eng. Bull.*, 36(4):54–59, 2013.

[57] CLEO Experiment.

[58] NSF Center for Coastal Margin Observation and Prediction (CMOP).

[59] Daniel Crankshaw, Peter Bailis, Joseph E. Gonzalez, Haoyuan Li, Zhao Zhang, Michael J. Franklin, Ali Ghodsi, and Michael I. Jordan. The missing piece in complex analytics: Low latency, scalable model management and serving with velox. 09 2014.

[60] Council for Scientific and Industrial Research (CSIR) in South Africa.

[61] Michele Dallachiesa, Amr Ebaid, Ahmed Eldawy, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, and Nan Tang. Nadeef: A commodity data cleaning system. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, SIGMOD '13, pages 541–552, New York, NY, USA, 2013. ACM.
`http://doi.acm.org/10.1145/2463676.2465327`

[62] The Data Observation Network for Earth (DataONE).

[63] Susan B. Davidson, Sarah Cohen Boulakia, Anat Eyal, Bertram Ludäscher, Timothy M. McPhillips, Shawn Bowers, Manish Kumar Anand, and **Juliana Freire**. Provenance in scientific workflow systems. *IEEE Data Eng. Bull.*, 30(4):44–50, 2007.

[64] Susan B. Davidson, Sarah Cohen-Boulakia, Anat Eyal, Bertram Ludäscher, Timothy McPhillips, Shawn Bowers, and **Juliana Freire**. Provenance in Scientific Workflow Systems. *IEEE Data Engineering Bulletin*, 32(4):44–50, 2007.

[65] Susan B Davidson and **Juliana Freire**. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1345–1350. ACM, 2008.

[66] Susan B. Davidson and **Juliana Freire**. *Provenance and scientific workflows: challenges and opportunities. In *SIGMOD*, pages 1345–1350, 2008.

[67] Amol Deshpande and Samuel Madden. Mauvedb: Supporting model-based user views in database systems. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 73–84, New York, NY, USA, 2006. ACM.
`http://doi.acm.org/10.1145/1142473.1142483`

[68] Andrew Dolgert, Lawrence Gibbons, Christopher D. Jones, Valentin Kuznetsov, Mirek Riedewald, Daniel Riley, Gregory J. Sharp, and Peter Wittich. Provenance in high-energy physics workflows. *Computing in Science and Engineering*, 10(3):22–29, 2008.

[69] H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C.T. Silva. Using topological analysis to support event-guided exploration in urban data. *IEEE TVCG*, 20(12):2634–2643, 2014.

[70] H. Doraiswamy, H. Vo, C.T. Silva, and J. Freire. *A GPU-Based Index to Support Interactive Spatio-Temporal Queries over Historical Data. In *ICDE*, 2016.

[71] Harish Doraiswamy, Nivan Ferreira, Theodoros Damoulas, **Juliana Freire**, and Cláudio T. Silva. *Using Topological Analysis to Support Event-Guided Exploration in Urban Data. *IEEE Trans. Vis. Comput. Graph.TVCG*, 20(12):2634–2643, 2014.

[72] Fang Du, Sihem Amer-Yahia, and **Juliana Freire**. Shrex: Managing xml documents in relational databases. In *Proceedings of VLDB*, pages 1297–1300, 2004.

[73] Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge Arnulfo Quiane-Ruiz, Nan Tang, and Si Yin. *NADEEF: A generalized data cleaning system*, volume 6, pages 1218–1221. 12 edition, 8 2013.

[74] Amr Ebaid, Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quiane-Ruiz, Nan Tang, and Si Yin. Nadeef: A generalized data cleaning system. *Proc. VLDB Endow.*, 6(12):1218–1221, August 2013.

[75] Eric Eide, Tim Stack, Leigh Stoller, **Juliana Freire**, and Jay Lepreau. *Integrated scientific workflow management for the Emulab network testbed. In *Proceedings of USENIX*, pages 363–368, 2006.

[76] I.G. Ellen, J. Lacoe, and C.A. Sharygin. Do foreclosures cause crime? *Journal of Urban Economics*, 74:59–70, 2013.

[77] Ingrid Gould Ellen and Johanna Ruth Lacoe. Do foreclosures cause crime? Technical report, New York University, 2013.

[78] Tommy Ellkvist, David Koop, Erik W. Anderson, **Juliana Freire**, and Cláudio T. Silva. Using provenance to support real-time collaborative design of workflows. In *IPAW*, pages 266–279, 2008.

[79] Tommy Ellkvist, David Koop, **Juliana Freire**, Cláudio Silva, and Lena Strömbck. *Using Mediation to Achieve Provenance Interoperability (Extended Abstract). In *IEEE International Conference on eScience*, pages 398–399, 2008.

[80] Tommy Ellkvist, Lena Strömbäck, Lauro Didier Lins, and **Juliana Freire**. A first study on strategies for generating workflow snippets. In *Proceedings of the ACM SIGMOD Intenational Workshop on Keyword Search on Structured Data (KEYS)*, pages 15–20, 2009.

[81] Tommy Ellqvist, David Koop, **Juliana Freire**, Claudio Silva, and Lena Stromback. Using mediation to achieve provenance interoperability. *IEEE Congress on Services*, pages 291–298, 2009.

[82] Ahmed Elmagarmid, Ihab F. Ilyas, Mourad Ouzzani, Jorge-Arnulfo Quiané-Ruiz, Nan Tang, and Si Yin. Nadeef/er: Generic and interactive entity resolution. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 1071–1074, New York, NY, USA, 2014. ACM.
`http://doi.acm.org/10.1145/2588555.2594511`

[83] Nivan Ferreira, Lauro Lins, Daniel Fink, Steve Kelling, Chris Wood, **Juliana Freire**, and Cláudio Silva. *BirdVis: Visualizing and Understanding Bird Populations. *IEEE Transactions on Visualization and Computer Graphics*, 17:2374–2383, 2011.

[84] Nivan Ferreira, Jorge Poco, Huy T. Vo, **Juliana Freire**, and Claudio T. Silva. *Visual Exploration of Big Spatio-Temporal Urban Data: A Study of New York City Taxi Trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013.

[85] M. H. Freedman, J. Gukelberger, M. B. Hastings, S. Trebst, M. Troyer, and Z. Wang. Galois conjugates of topological phases. *Phys. Rev. B*, 85:045414, Jan 2012.

[86] J. Freire, D. Koop, E. Santos, C. Scheidegger, C. Silva, and H. T. Vo. **The Architecture of Open Source Applications*, chapter VisTrails. Lulu.com, 2011.

[87] **Juliana Freire**. Provenance management: Challenges and opportunities. In *Datenbanksysteme in Business, Technologie und Web (BTW)*, page 4, 2009.

[88] **Juliana Freire** and Michael Benedikt. Managing xml data: An abridged overview. *IEEE Computing in Science & Engineering*, 6(4):12–19, 2004.

[89] **Juliana Freire**, Aline Bessa, Fernando Seabra Chirigati, Huy Vo, and Kai Zhao. Exploring what not to clean in urban data:
a study using new york city taxi trips. *IEEE Data Eng. Bull.*, 2016.

[90] **Juliana Freire**, Philippe Bonnet, and Dennis Shasha. *Exploring the Coming Repositories of Reproducible Experiments: Challenges and Opportunities. *PVLDB*, 4(12):1494–1497, 2011.

[91] **Juliana Freire**, Philippe Bonnet, and Dennis Shasha. *Computational reproducibility: state-of-the-art, challenges, and database research opportunities. In *SIGMOD Conference*, pages 593–596, 2012.

[92] **Juliana Freire** and David Koop. Ci-en: Enhancing and supporting a community-based data analysis, visualization, and provenance platform, 2014.

[93] **Juliana Freire**, David Koop, Emanuele Santos, and Cláudio T. Silva. *Provenance for Computational Tasks: A Survey. *Computing in Science and Engineering*, 10(3):11–21, 2008.

[94] **Juliana Freire**, M. Ramanath, and L. Zhang. A flexible infrastructure for gathering XML statistics and estimating query cardinality. In *IEEE International Conference on Data Engineering (ICDE)*, 2004.

[95] **Juliana Freire** and Cláudio Silva. Managing complex visualizations, July 2005.

[96] **Juliana Freire** and Cláudio Silva. *Simplifying the Design of Workflows for Large-Scale Data Exploration and Visualization. In *Proceedings of the Microsoft eScience Workshop*, 2008.

[97] **Juliana Freire** and Claudio Silva. *Towards Enabling Social Analysis of Scientific Data. In *ACM CHI Social Data Analysis Workshop*, 2008.

[98] **Juliana Freire**, Cláudio Silva, Steve Callahan, Emanuele Santos, Carlos Scheidegger, and Huy Vo. *Managing Rapidly-Evolving Scientific Workflows. In *International Provenance and Annotation Workshop (IPAW)*, LNCS 4145, pages 10–18. Springer Verlag, 2006.

[99] **Juliana Freire** and Cláudio T. Silva. Making computations and publications reproducible with vistrails. *Computing in Science and Engineering*, 14(4):18–25, 2012.

[100] **Juliana Freire** and Cláudio T. Silva. *Making Computations and Publications Reproducible with VisTrails. *Computing in Science and Engineering*, 14(4):18–25, 2012.

[101] **Juliana Freire**, Cláudio T Silva, Steven P Callahan, Emanuele Santos, Carlos E Scheidegger, and Huy T Vo. Managing rapidly-evolving scientific workflows. In *Provenance and Annotation of Data*, pages 10–18. Springer, 2006.

[102] **Juliana Freire**, Cláudio T. Silva, Huy T. Vo, Harish Doraiswamy, Nivan Ferreira, and Jorge Poco. *Riding from Urban Data to Insight Using New York City Taxis. *IEEE Data Eng. Bull.*, 37(4):43–55, 2014.

[103] Furman center: Data services. `http://furmancenter.org/data`.

[104] Robert B. Gilbert, Fulvio Tonon, **Juliana Freire**, Cláudio Silva, and David R. Maidment. Visualizing uncertainty with uncertainty multiples. In *GeoCongress 2006: Geotechnical Engineering in the Information Technology Age*, pages 1–6. ASCE, 2006.

[105] **Boris Glavic**. Big data provenance: Challenges and implications for benchmarking. In *2nd Workshop on Big Data Benchmarking (WBDB)*, pages 72–80, 2012.

[106] **Boris Glavic** and Gustavo Alonso. Perm: Processing Provenance and Data on the same Data Model through Query Rewriting. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE)*, pages 174–185, 2009.

[107] **Boris Glavic** and Gustavo Alonso. Provenance for Nested Subqueries. In *Proceedings of the 12th International Conference on Extending Database Technology (EDBT)*, pages 982–993, 2009.

[108] **Boris Glavic** and Gustavo Alonso. The Perm Provenance Management System in Action. In *Proceedings of the 35th ACM SIGMOD International Conference on Management of Data (SIGMOD) (Demonstration Track)*, pages 1055–1058, 2009.

[109] **Boris Glavic**, Gustavo Alonso, Renée J. Miller, and Laura M. Haas. TRAMP: Understanding the Behavior of Schema Mappings through Provenance. *Proceedings of the Very Large Data Bases Endowment (PVLDB)*, 3(1):1314–1325, 2010.

[110] **Boris Glavic**, Jiang Du, Renée J. Miller, Gustavo Alonso, and Laura M. Haas. Debugging Data Exchange with Vagabond. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 4(12):1383–1386, 2011.

[111] **Boris Glavic**, Kyumars Sheykh Esmaili, Peter M. Fischer, and Nesime Tatbul. Ariadne: Managing fine-grained provenance on data streams. In *Proceedings of the 7th ACM International Conference on Distributed Event-Based Systems (DEBS)*, pages 291–320, 2013.

[112] **Boris Glavic**, Kyumars Sheykh Esmaili, Peter M. Fischer, and Nesime Tatbul. Efficient stream provenance via operator instrumentation. *Transactions on Internet Technology (TOIT)*, 13(1):7:1–7:26, 2014.

[113] **Boris Glavic**, Sven Köhler, Sean Riddle, and Bertram Ludäscher. *Towards Constraint-based Explanations for Answers and Non-Answers. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2015.

[114] **Boris Glavic**, Tanu Malik, and Quan Pham. *Making Database Applications Shareable. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP) (Poster)*, 2015.

[115] **Boris Glavic** and Renée J. Miller. Reexamining Some Holy Grails of Data Provenance. In *Proceedings of the 3rd USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[116] **Boris Glavic**, Renée J Miller, and Gustavo Alonso. Using sql for efficient generation and querying of provenance information. In *In search of elegance in the theory and practice of computation: a Festschrift in honour of Peter Buneman*, pages 291–320. Springer, 2013.

[117] **Boris Glavic**, Javed Siddique, Periklis Andritsos, and Renée J. Miller. Provenance for data mining. In *Proceedings of the 5th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2013.

[118] Timothy Griffin and Leonid Libkin. Incremental maintenance of views with duplicates. *SIGMOD Rec.*, 24(2):328–339, May 1995.

[119] K.C. Gross, R.M. Singer, S.W. Wegerich, J.P. Herzog, R. VanAlstine, and F. Bockhorst. *Application of a model-based fault detection system to nuclear plant signals.* May 1997.

[120] M. Gubanov, M. Stonebraker, and D. Bruckner. Text and structured data fusion in data tamer at scale. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, pages 1258–1261, March 2014.

[121] Philip J. Guo, Sean Kandel, Joseph M. Hellerstein, and Jeffrey Heer. Proactive wrangling: Mixed-initiative end-user programming of data transformation scripts. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 65–74, New York, NY, USA, 2011. ACM.
`http://doi.acm.org/10.1145/2047196.2047205`

[122] Daniel Haas, Sanjay Krishnan, Jiannan Wang, Michael J. Franklin, and Eugene Wu. Wisteria: Nurturing scalable data cleaning infrastructure. *Proc. VLDB Endow.*, 8(12):2004–2007, August 2015.

[123] Garcia-Molina Hector, Jeffrey D Ullman, and Jennifer Widom. *Database systems: The complete book*. Prentice-Hall, 2002.

[124] Randy Heiland, Maciek Swat, Benjamin Zaitlen, James Glazier, and Andrew Lumsdale. Workflows for parameter studies of multi-cell modeling (hpc). In *Proceedings of the ACM High Performance Computing Symposium*, 2010.

[125] Tuan-Anh Hoang-Vu, Vicki Been, Ingrid Gould Ellen, Max Weselcouch, and **Juliana Freire**. Towards understanding real-estate ownership in new york city: Opportunities and challenges. In *Proceedings of the Workshop on Data Science for Macro-Modeling*, 2014.

[126] Keren Mertens Horn, Ingrid Gould Ellen, and Amy Ellen Schwartz. Do housing choice voucher holders live near good schools? *Journal of Housing Economics*, 24(0):109–121, 2014.

[127] Bill Howe, Peter Lawson, Renee Bellinger, Erik Anderson, Emanuele Santos, **Juliana Freire**, Carlos Scheidegger, António Baptista, and Cláudio Silva. *End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. In *IEEE International Conference on eScience*, pages 127–134, 2008.

[128] Bill Howe, Peter Lawson, Renee Bellinger, Erik W. Anderson, Emanuele Santos, **Juliana Freire**, Carlos Eduardo Scheidegger, Antonio Baptista, and Claudio T. Silva. End-to-End eScience: Integrating Workflow, Query, Visualization, and Provenance at an Ocean Observatory. In *eScience '08: Proceedings of the 4th IEEE International Conference on eScience*, pages 127–134, 2008.

[129] Bill Howe, Claudio Silva, and **Juliana Freire**. A science cloud on your desktop: Vistrails + gridfields, 2009.

[130] Zachary G Ives, Zhepeng Yan, Nan Zheng, Brian Litt, and Joost B Wagenaar. Looking at everything in context.

[131] H. V. Jagadish, Li Qian, and Arnab Nandi. Organic databases. *IJCSE*, 11(3):270–283, 2015.

[132] HV Jagadish, A. Chapman, A. Elkiss, M. Jayapandian, Y. Li, A. Nandi, and C. Yu. Making database systems usable. *Proceedings of the 2007 ACM SIGMOD international conference on Management of data*, pages 13–24, 2007.

[133] Emanuele Santos Juliana Freire and Cláudio Silva. Provenance-enabled data exploration and visualization with vistrails. In *SciDAC*, volume 125, 2009.

[134] Sean Kandel, Andreas Paepcke, Joseph Hellerstein, and Jeffrey Heer. Wrangler: Interactive visual specification of data transformation scripts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 3363–3372, New York, NY, USA, 2011. ACM.
`http://doi.acm.org/10.1145/1978942.1979444`

[135] Manos Karpathiotakis, Ioannis Alagiannis, Thomas Heinis, Miguel Branco, and Anastasia Ailamaki. Just-In-Time Data Virtualization: Lightweight Data Management with ViDa. In *Proceedings of the 7th Biennial Conference on Innovative Data Systems Research (CIDR)*, 2015.

[136] Yannis Katsis, Yoav Freund, and Yannis Papakonstantinou. Combining databases and signal processing in plato. In *CIDR*, 2015.

[137] O. Kennedy, C. Koch, and A. Demers. Dynamic approaches to in-network aggregation. In *ICDE*, pages 1331–1334, 2009.

[138] **Oliver Kennedy**, Yanif Ahmad, and Christoph Koch. DBToaster: Agile views for a dynamic data management system. In *CIDR*, pages 284–295, 2011.

[139] **Oliver Kennedy**, Jerry Ajay, Geoffrey Challen, and Lukasz Ziarek. Pocket Data: The need for TPC-MOBILE. In *TPC Technology Conference on Performance Evaluation & Benchmarking*, 2015.

[140] **Oliver Kennedy** and Christoph Koch. Pip: A database system for great and small expectations. In *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, pages 157–168. IEEE, 2010.

[141] **Oliver Kennedy**, Steve Lee, Charles Loboz, Slawek Smyl, and Suman Nath. Fuzzy prophet: Parameter exploration in uncertain enterprise scenarios. In *SIGMOD*, pages 1303–1306, 2011. `http://doi.acm.org/10.1145/1989323.1989482`

[142] **Oliver Kennedy** and Suman Nath. Jigsaw: Efficient optimization over uncertain enterprise data. In *SIGMOD*, pages 829–840, 2011. `http://doi.acm.org/10.1145/1989323.1989410`

[143] **Oliver Kennedy**, Ying Yang, Jan Chomicki, Ronny Fehling, Zhen Hua Liu, and Dieter Gawlick. *Enabling Real-Time Business Intelligence: International Workshops, BIRTE 2013, Riva del Garda, Italy, August 26, 2013, and BIRTE 2014, Hangzhou, China, September 1, 2014, Revised Selected Papers*, chapter Detecting the Temporal Context of Queries, pages 97–113. Springer Berlin Heidelberg, 2015.

[144] **Oliver Kennedy** and Lukasz Ziarek. Just-in-time data structures. In *CIDR*, 2015.

[145] Christoph Koch, Yanif Ahmad, **Oliver Andrzej Kennedy**, Milos Nikolic, Andres Nötzli, Daniel Lupei, and Amir Shaikhha. DBToaster: Higher-order delta processing for dynamic, frequently fresh views. *VLDBJ*, 2013.

[146] D. Koop, J. Freire, and C.T. Silva. Visual summaries for graph collections. In *Visualization Symposium (PacificVis), 2013 IEEE Pacific*, pages 57–64, 2013.

[147] David Koop and **Juliana Freire**. *Reorganizing Workflow Evolution Provenance. In *USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2014.

[148] David Koop, Emanuele Santos, Bela Bauer, Matthias Troyer, **Juliana Freire**, and Cláudio T. Silva. Bridging workflow and data provenance using strong links. In *SSDBM*, pages 397–415, 2010.

[149] David Koop, Emanuele Santos, Phillip Mates, Huy T. Vo, Philippe Bonnet, Bela Bauer, Brigitte Surer, Matthias Troyer, Dean N. Williams, Joel E. Tohline, **Juliana Freire**, and Cludio T. Silva. *A Provenance-Based Infrastructure to Support the Life Cycle of Executable Papers. *Procedia Computer Science*, 4:648–657, 2011. Proceedings of the International Conference on Computational Science, ICCS 2011.

[150] David Koop, Carlos Scheidegger, **Juliana Freire**, and Cláudio T. Silva. *The Provenance of Workflow Upgrades. In *IPAW*, pages 2–16, 2010.

[151] David Koop, Carlos E Scheidegger, Steven P Callahan, **Juliana Freire**, and Cláudio T Silva. Viscomplete: Automating suggestions for visualization pipelines. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1691–1698, 2008.

[152] David Koop, Carlos Eduardo Scheidegger, Steven P. Callahan, **Juliana Freire**, and Cláudio T. Silva. *VisComplete: Automating Suggestions for Visualization Pipelines. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1691–1698, 2008.

[153] Bogdan Korel and Janusz Laski. Dynamic program slicing. *Information Processing Letters*, 29(3):155 – 163, 1988.

[154] Sanjay Krishnan, Jiannan Wang, Eugene Wu, Michael J. Franklin, and Ken Goldberg. Activeclean: Interactive data cleaning while learning convex loss models. 01 2016.

[155] Gokhan Kul, Duc Thanh Luong, Ting Xie, Patrick Coonan, Varun Chandola, **Oliver Kennedy**, and Shambhu Upadhaya. * Ettu: Analyzing query intents in corporate databases. In *ERMIS*, 2016.

[156] Gokhan Kul and Shambhu Upadhyaya. * A preliminary cyber ontology for insider threats in the financial sector. In *Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats*, MIST '15, pages 75–78, New York, NY, USA, 2015. ACM. `http://doi.acm.org/10.1145/2808783.2808793`

[157] Julia Maureen Letchner. *Lahar: warehousing markovian streams*. PhD thesis, University of Washington, 2010.

[158] Julie Letchner, Christopher Ré, Magdalena Balazinska, and Matthai Philipose. Lahar demonstration: warehousing markovian streams. *Proceedings of the VLDB Endowment*, 2(2):1610–1613, 2009.
[159] Lauro Lins, David Koop, Erik W. Anderson, Steven P. Callahan, Emanuele Santos, Carlos Eduardo Scheidegger, **Juliana Freire**, and Cláudio T. Silva. Examining statistics of workflow evolution provenance: A first study. In *SSDBM*, pages 573–579, 2008.
[160] Bin Liu and HV Jagadish. A spreadsheet algebra for a direct data manipulation query interface. In *Data Engineering, 2009. ICDE'09. IEEE 25th International Conference on*, pages 417–428. IEEE, 2009.
[161] Phillip Mates, Emanuele Santos, **Juliana Freire**, and Cláudio T. Silva. *CrowdLabs: Social Analysis and Visualization for the Sciences. In *SSDBM*, pages 555–564, 2011.
[162] Sergio L. S. Mergen, **Juliana Freire**, and Carlos A. Heuser. Querying structured information sources on the web. In *iiWAS*, pages 470–476, 2008.
[163] Luc Moreau, **Juliana Freire**, Joe Futrelle, Robert McGrath, Jim Myers, and Patrick Paulson. The open provenance model (v1.00), December 2007.
[164] Luc Moreau, **Juliana Freire**, Joe Futrelle, Robert E. McGrath, Jim Myers, and Patrick Paulson. The open provenance model: An overview. In *IPAW*, pages 323–326, 2008.
[165] J. Morisette, C. Jarnevich, T. Holcombe, C. Talbert, D. Ignizio, M. Talbert, C. T. Silva, D. Koop, A. Swanson, and N. Young. VisTrails SAHM: Visualization and workflow management for ecological niche modeling. *Ecography*, 2012. To appear.
[166] Leonardo Murta, Vanessa Braganholo, Fernando Seabra Chirigati, David Koop, and **Juliana Freire**. noworkflow: Capturing and analyzing provenance of scripts. In *IPAW*, 2014.
[167] Arindam Nandi, Ying Yang, **Oliver Kennedy**, **Boris Glavic**, Ronny Fehling, Zhen Hua Liu, and Dieter Gawlick. Mimir: Bringing ctables into practice. *CoRR*, abs/1601.00073, 2016.
[168] Suman K Nath, Seung Ho Lee, Slawomir Smyl, Charles Z Loboz, and **Oliver Andrzej Kennedy**. Efficient optimization over uncertain data, December 20 2012.
[169] Hoa Nguyen, Eun Yong Kang, and **Juliana Freire**. Automatically extracting form labels. In *ICDE*, pages 1498–1500, 2008.
[170] Hoa Nguyen, Thanh Nguyen, and **Juliana Freire**. Learning to extract form labels. *PVLDB*, 1(1):684–694, 2008.
[171] Huong Nguyen, Thanh Nguyen, Hoa Nguyen, and **Juliana Freire**. *Querying Wikipedia Documents and Relationships. In *Proceedings of WebDB*, 2010.
[172] Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, and **Juliana Freire**. *Multilingual Schema Matching for Wikipedia Infoboxes. *PVLDB*, 2012. Conditionally accepted.
[173] Thanh Hoang Nguyen, Hoa Nguyen, and **Juliana Freire**. Prusm: a prudent schema matching approach for web forms. In *CIKM*, pages 1385–1388, 2010.
[174] Xing Niu, Raghav Kapoor, Dieter Gawlick, Zhen Hua Liu, Vasudha Krishnaswamy, Venkatesh Radhakrishnan, and **Boris Glavic**. Interoperability for provenance-aware databases using prov and json. In *Proceedings of the 7th USENIX Workshop on the Theory and Practice of Provenance (TaPP)*, 2015.
[175] Xing Niu, Raghav Kapoor, and **Boris Glavic**. Heuristic and cost-based optimization for provenance computation. In *TaPP*, 2015.
[176] Donald A. Norman. *Things That Make Us Smart: Defending Human Attributes in the Age of the Machine*. Addison Wesley, 1994.
[177] Donald A Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.
[178] Nsf discovery: A new vision for scientific visualizations. http://www.nsf.gov/discoveries/disc_summ.jsp?cntn_id=114322 March 2009.
[179] NYC OpenData. `https://nycopendata.socrata.com`.
[180] Masayo Ota, Huy T. Vo, Cláudio T. Silva, and **Juliana Freire**. *A scalable approach for data-driven taxi ride-sharing simulation. In *IEEE International Conference on Big Data*, pages 888–897, 2015.
[181] Cesar Palomo, Zhan Guo, Cláudio T. Silva, and **Juliana Freire**. *Visually Exploring Transportation Schedules. *IEEE Trans. Vis. Comput. Graph.*, 22(1):170–179, 2016.

[182] Quan Pham, Tanu Malik, **Boris Glavic**, and Ian Foster. *LDV: Light-weight Database Virtualization. In *Proceedings of the 25th IEEE International Conference on Data Engineering (ICDE)*, pages 1179–1190, 2015.

[183] Quan Pham, Richard Whaling, **Boris Glavic**, and Tanu Malik. *Sharing and Reproducing Database Applications. *Proceedings of the VLDB Endowment (PVLDB) (Demonstration Track)*, 8(12):1988 – 1999, 2015.

[184] Jorge Poco, Harish Doraiswamy, Huy Vo, João LD Comba, **Juliana Freire**, Cláudio Silva, et al. *Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions. *Computer Graphics Forum*, 34(3):161–170, 2015.

[185] Vijayshankar Raman and Joseph M. Hellerstein. Potter's wheel: An interactive data cleaning system. In *Proceedings of the 27th International Conference on Very Large Data Bases*, VLDB '01, pages 381–390, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. `http://dl.acm.org/citation.cfm?id=645927.672045`

[186] Maya Ramanath, **Juliana Freire**, Jayant Haritsa, and Prasan Roy. Searching for efficient xml-to-relational mappings. In *Proceedings of XML Database Symposium (XSym)*, pages 19–36, 2003.

[187] Maya Ramanath, Lingzhi Zhang, **Juliana Freire**, and Jayant Haritsa. Imax: Incremental maintenance of schema-based xml statistics. In *IEEE International Conference on Data Engineering (ICDE)*, pages 273–284, 2005.

[188] Remote data analysis and visualization (rdav), 2009.

[189] ReproZip. https://github.com/ViDA-NYU/reprozip.

[190] Software for Assisted Habitat Modeling Package for VisTrails (SAHM: VisTrails).

[191] E. Santos, J. Poco, Yaxing Wei, Shishi Liu, B. Cook, D.N. Williams, and C.T. Silva. Uv-cdat: Analyzing climate datasets from a user's perspective. *Computing in Science and Engineering*, 15(1):94–103, 2013.

[192] Emanuele Santos. *Simplifying the Creation and Deployment of Collaborative Data Analysis and Visualization Tools*. PhD thesis, University of Utah, 2010.

[193] Emanuele Santos, **Juliana Freire**, and Claudio Silva. Information sharing in science 2.0: Challenges and opportunities. In *ACM CHI Workshop on The Changing Face of Digital Science: New Practices in Scientific Collaborations*, 2009.

[194] Emanuele Santos, **Juliana Freire**, Claudio Silva, Ayla Khan, Julien Tierny, Brad Grimm, Lauro Lins, Valerio Pascucci, Scott A. Klasky", Roselyne D. Barreto, and Norbert Podhorszki. Enabling advanced visualization tools in a simulation monitoring system. In *Proceedings of the 5th IEEE International Conference on e-Science*, pages 358–365. IEEE, December 2009.

[195] Emanuele Santos, David Koop, Thomas Maxwell, Charles Doutriaux, Tommy Ellqvist, Gerald Potter, **Juliana Freire**, Dean Williams, and Claudio Silva. Designing a provenance-based climate data analysis application. In *IPAW*, pages 214–219, 2012.

[196] Emanuele Santos, David Koop, Huy T. Vo, Erik W. Anderson, **Juliana Freire**, and Cláudio T. Silva. *Using Workflow Medleys to Streamline Exploratory Tasks. In *SSDBM*, pages 292–301, 2009.

[197] Emanuele Santos, Lauro Lins, James Ahrens, **Juliana Freire**, and Cláudio T. Silva. *VisMashup: Streamlining the Creation of Custom Visualization Applications. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1539–1546, 2009.

[198] Emanuele Santos, Lauro Lins, James P. Ahrens, **Juliana Freire**, and Cláudio T. Silva. *A First Study on Clustering Collections of Workflow Graphs. In *IPAW*, pages 160–173, 2008.

[199] Emanuele Santos, Phillip Mates, Erik Anderson, Brad Grimm, **Juliana Freire**, and Claudio Silva. Towards supporting collaborative data analysis and visualization in a coastal margin observatories. ACM CSCW Workshop on The Changing Dynamics of Scientific Collaborations, 2010.

[200] Carlos Scheidegger, Huy Vo, David Koop, **Juliana Freire**, and Claudio Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization & Computer Graphics*, 13(6):1560–1567, 2007.

[201] Carlos Eduardo Scheidegger, David Koop, Emanuele Santos, Huy T. Vo, Steven P. Callahan, **Juliana Freire**, and Cláudio T. Silva. *Tackling the Provenance Challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008.

[202] Carlos Eduardo Scheidegger, Huy Vo, David Koop, **Juliana Freire**, and Claudio T. Silva. Querying and Re-using Workflows with VisTrails. In *SIGMOD '08: Proceedings of the 34th SIGMOD International Conference on Management of Data*, pages 1251–1254. ACM, 2008.

[203] Carlos Eduardo Scheidegger, Huy T. Vo, David Koop, **Juliana Freire**, and Cláudio T. Silva. *Querying and Creating Visualizations by Analogy. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1560–1567, 2007.

[204] Carlos Eduardo Scheidegger, Huy T. Vo, David Koop, **Juliana Freire**, and Cláudio T. Silva. *Querying and re-using workflows with VisTrails. In *SIGMOD*, pages 1251–1254, 2008.

[205] Amy Ellen Schwartz, Ingrid Gould Ellen, Ioan Voicu, and Michael H. Schill. The external effects of place-based subsidized housing. *Regional Science and Urban Economics*, 36(6):679 – 707, 2006.

[206] Cláudio Silva, **Juliana Freire**, and Steven P. Callahan. *Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science & Engineering*, 9(5):82–89, 2007.

[207] Cláudio T. Silva, Erik Anderson, Emanuele Santos, and **Juliana Freire**. *Using VisTrails and Provenance for Teaching Scientific Visualization. In *Proceedings of the Eurographics Education Program*, 2010.

[208] Cláudio T. Silva and **Juliana Freire**. *Software Infrastructure for exploratory visualization and data analysis: past, present, and future. *Journal of Physics: Conference Series*, 25:012100 (15pp), 2008. SciDAC 2008 Conference.

[209] Claudio T. Silva, **Juliana Freire**, and Steven Callahan. Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science and Engineering*, 9(5):82–89, 2007.

[210] Cludio T. Silva, Erik Anderson, Emanuele Santos, and **Juliana Freire**. *Using VisTrails and Provenance for Teaching Scientific Visualization. *Computer Graphics Forum*, 30(1):75–84, 2011.

[211] Richard T Snodgrass, Jim Gray, and Jim Melton. *Developing time-oriented database applications in SQL*, volume 42. Morgan Kaufmann Publishers San Francisco, 2000.

[212] Apache Spark. Apache spark–lightning-fast cluster computing. 2014.

[213] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cherniack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. Data curation at scale: The data tamer system. In *CIDR*, 2013.

[214] Dan Suciu, Dan Olteanu, Christopher Ré, and Christoph Koch. *Probabilistic databases, synthesis lectures on data management*. Morgan & Claypool, 2011.

[215] Wang Chiew Tan. Provenance in databases: Past, current, and future. *IEEE Data Eng. Bull.*, 30(4):3–12, 2007.

[216] TaxiVis. `https://github.com/ViDA-NYU/TaxiVis`.

[217] The MayBMS project. Pdbench. http://pdbench.sourceforge.net.

[218] TLC Trip Record Data. `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`, 2015.

[219] Joel E. Tohline, Jinghya Ge, Wesley Even, and Erik Anderson. A customized python module for cfd flow analysis within vistrails. *Computing in Science and Engineering*, 11(3):68–73, 2009.

[220] Trifacta. Trifacta wrangler. https://www.trifacta.com.

[221] Ultrascale Visualization - Climate Data Analysis Tools (UV-CDAT).

[222] TL Van Zyl, G McFerren, and A Vahed. Earth observation scientific workflows in a distributed computing environment. Technical Report 7727, CSIR, 2011.

[223] Karane Vieira, André Luiz Costa Carvalho, Klessius Berlt, Edleno S. Moura, Altigran S. Silva, and **Juliana Freire**. On finding templates on web collections. *World Wide Web*, 12(2):171–211, 2009.

[224] VisTrails. http://www.vistrails.org.

[225] VisTrails Users Guide.

[226] H. T. Vo, J. Bronson, B. Summa, J. Comba, J. Freire, B. Howe, V. Pascucci, , and C. Silva. *Parallel Visualization on Large Clusters using MapReduce. In *Proceedings of IEEE Symposium on Large-Scale Data Analysis and Visualization (LDAV)*, pages 81–88, 2011.

[227] vtDV3D VisTrails Package.

[228] Daisy Zhe Wang, Eirinaios Michelakis, Minos Garofalakis, and Joseph M. Hellerstein. Bayesstore: Managing large, uncertain data repositories with probabilistic graphical models. *Proc. VLDB Endow.*, 1(1):340–351, August 2008.

[229] Jiannan Wang, Sanjay Krishnan, Michael J. Franklin, Ken Goldberg, Tim Kraska, and Tova Milo. A sample-and-clean framework for fast and accurate query processing on dirty data. In *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*, SIGMOD '14, pages 469–480, New York, NY, USA, 2014. ACM.
`http://doi.acm.org/10.1145/2588555.2610505`

[230] Mark Weiser. Program slicing. In *Proceedings of the 5th International Conference on Software Engineering*, ICSE '81, pages 439–449, Piscataway, NJ, USA, 1981. IEEE Press.
`http://dl.acm.org/citation.cfm?id=800078.802557`

[231] Ying Yang. On-demand query result cleaning. In *VLDB PhD Workshop*, 2014.

[232] Ying Yang, Niccolo Meneghetti, Ronny Fehling, Zhen Hua Liu, and **Oliver Kennedy**. Lenses: an on-demand approach to etl. *Proceedings of the VLDB Endowment*, 8(12):1578–1589, 2015.