Streamlining and Understanding Curation with Vizier

Data curation (aka wrangling and cleaning) is a critical stage 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 tools (e.g., Oracle's Data Guide and Trifacta's Wrangler), wrangling 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. Furthermore, as domain experts integrate data sets to test their hypotheses, erroneous data points identified when a data set was analyzed in isolation may actually turn out to be important features.

Consider, for example, taxis in New York City.[1] Every day, there are over 500,000 taxi trips transporting about 600,000 people from Manhattan to different parts of the city.[2] 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 [6], 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. We observe 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.

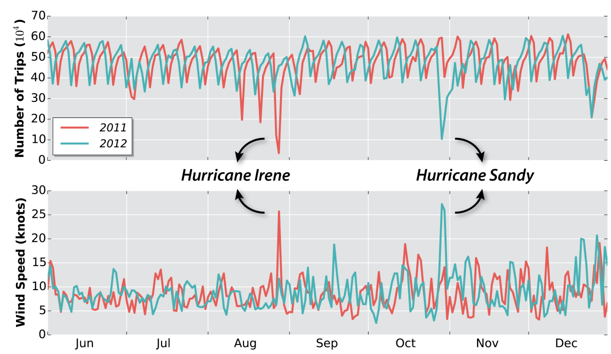

Figure 1: The plot on the top shows how the number of trips varies over 2011 and 2012. While the variation is similar for the two years, there are clear outliers, including large drops in August 2011 and in October 2012. These are not errors, but in fact correspond to hurricanes Sandy and Irene, as shown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

Issues much like these appear across all forms of analytics, making curation an integral component of data exploration. As data are analyzed and erroneous features are identified, appropriate *cleaning operations should be applied on the fly*. The trial-and-error nature of an exploratory curation process poses several challenges. First, if an operation is applied and later it is 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*. Second, it should be possible to *modify an operation* (e.g., change the parameters of an

[1] `http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml`

[2] `http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf`

outlier detection operation to no longer consider the \$938.02 tip amount as an outlier) and the effects of this modification should be propagated to derived data. Third, it is often necessary to *explore and compare alternative cleaning strategies and to consolidate their results*. Last, but not least, because analysis results are highly dependent on the curation process applied to the input data, *the ability to explain and audit the cleaning process* is crucial.

The Vizier System

To address these challenges, in this project we will build Vizier, a system that unifies curation and data exploration through provenance. We will integrate and significantly extend three different components we have developed in previous work: *Mimir* [14, 16, 21, 22], a system that supports probabilistic pay-as-you-go data curation operators; *VisTrails* [3–5, 7, 8, 13, 15, 18–20], an NSF-supported open-source system designed for interactive data exploration that provides a comprehensive provenance management infrastructure; and *GProM* [1, 2, 9–12, 17], a database middleware that efficiently supports fine-grained data provenance. Vizier leverages and extends unique features of the systems it integrates. At its core are provenance tracking mechanisms that capture both the exploratory curation process—how the cleaning workflows evolve, and how data changes over time. By connecting these different types of provenance, Vizier will not only support the auditing of curation processes, but also explain the context in which they were applied. This, in turn, will make data science easier, faster, and more broadly accessible.

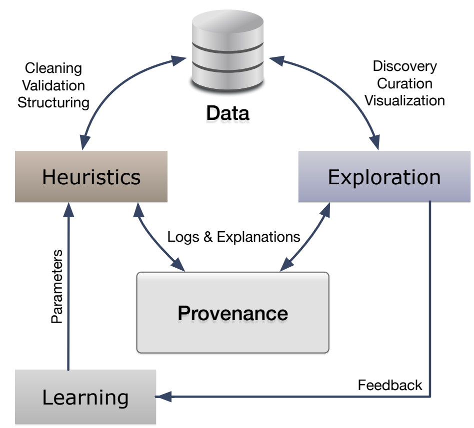

Figure 2: The Vizier System

As a user explores a data set and defines cleaning operations, a curation workflow is incrementally constructed. The change-based provenance introduced by the **VisTrails** system [8], tracks how the workflow evolves over time. Similar to a version control system, different versions are maintained. This naturally supports *collaboration*. Users can easily *navigate through the space of workflows* created for a given curation task, *visually compare workflows and their results*, undo changes without losing any results, and thus, enact *reflective reasoning* – making inferences from stored knowledge and following chains of reasoning backward and forward. This is key to supporting the trial-and-error nature of data curation. Users can also *re-use knowledge by exploring provenance information*. They can query the workflows both to reason about the process used to create them and to find examples that can help in the construction of new workflows. Errors in one stage of a curation workflow might not be detected until several stages later. As errors are repaired, the correcting operations must be propagated throughout the workflow, or even different versions of the workflow. With change-based provenance, these *modifications can be automatically applied to a workflow collection by analogy* [18].

VisTrails tracks provenance at the workflow level and views data input into and derived by workflows as a black box. For curation, it is important to understand the effects of the cleaning operations to individual data items – these are capabilities provided by GProM and Mimir. Mimir is a database-independent pay-as-you-go curation middleware that uses qualitative metrics to help users understand the quality of their data and judge how much curation effort is required. GProM is database-independent provenance middleware for computing this fine-grained provenance for queries, updates, and transactions. Our proposed system, Vizier will use VisTrails as a hub, supplementing it with the fine-grained provenance models of GProM and Mimir.

GProM tracks data provenance in a non-intrusive manner, without requiring any changes to applications or the database backend. By combining fine-grained data provenance and workflow provenance, Vizier will *support data-dependent explanations and recommendations.* For instance, the system will precisely explain how a record or value produced by a curation workflow was derived – which workflow stages modified it, which data items contributed to it. In previous work [15], we designed a recommendation system which, by mining workflow evolution provenance, suggests workflow steps to be applied. While this approach took only the workflow structure into account, in VIZIER, we will extend this approach to also consider data provenance. For example, if we know which data items were successfully cleaned by an operation, we can recommend this operation for data sets with similar characteristics. Furthermore, using the declarative replay technique called *reenactment*, the main enabler of GProM's provenance tracking mechanism for updates, it is possible to efficiently propagate changes to data and operations through a workflow. Reenactment turns updates into queries and, thus, changes to data can be virtualized, enabling any operation in VIZIER to be easily undone. There are many automated tools for entity resolution, schema matching, JSON shredding, log extraction, interpolation, or virtually any other data repair task. By tracking the effectiveness of these tools for particular data sets and tasks, it is possible to guide the domain expert in selecting the right tool for a task, in configuring these tools, and in understanding their interactions and outputs. In addition, tools and operations can also be suggested based on the actual data being explored.

Mimir provides a suite of data curation operations called Lenses that require minimal *upfront* configuration or tuning from users to perform cleaning tasks like data imputation, structural curation tasks like schema matching, or outlier detection through algorithms like MSET. When first created, a Lens makes a best-effort guess about its tuning parameters, allowing users to immediately run SQL queries over messy, semi- or un-structured data. These best effort guesses *may* initially result in of low-quality curation, and some of them will need to be refined. To help the user understand the impact of these guesses and react accordingly, each Lens tags its output with provenance markers that persist through queries. When displaying results, Mimir uses these markers to help users understand when a result is uncertain, why the result is uncertain, the magnitude of its uncertainty, and what parameters the user can tune to fix it. For example, the US$938.02 tip is an outlier, but may still be correct. However, large aggregate computations (e.g., the average tip per mile traveled) may not be significantly affected by this outlier; whether or not the tip is correct is irrelevant. Mimir provides quantitative metrics like bounds and standard deviations that help users decide whether their output is sufficiently precise. In cases where it is not (e.g., segmenting the analysis by hour and neighborhood), Mimir provides a prioritized list of curation tasks that guides users through the process of improving the quality of their data.

VIZIER will support a series of cleaning operations and tools that will be mixed and matched in the curation workflows. These range from regular expressions and user-defined functions to Mimir's Lenses. Extending VisTrails' workflow-level provenance with GProM's data-level provenance makes it possible to precisely track the effects of these operations. When combined with Mimir's facilities for measuring and communicating uncertainty, VIZIER becomes a powerful tool for putting data in context and for establishing trust in a dataset with minimal effort.

Applications. The proposed work will be used in real applications as part of two ongoing collaborative projects between the PIs, domain experts, government, and industry. PI Freire is working with social scientists and New York City agencies in the development of new methods to explore urban data. PIs Kennedy and Glavic are working with Oracle and Airbus on IoT/Situational Awareness challenges like managing and querying large sensor logs (e.g., for flight metrics or server clusters). These existing projects will naturally foster a close interaction between the computer scientists and domain experts, which in turn will lead to continuous feedback in the development of the proposed techniques and tools. *We will evaluate the efficiency and effectiveness of our techniques using these applications.* Furthermore, the use of our work in these projects affords us the opportunity to have immediate practical impact across disciplines: the proposed techniques have the potential to contribute to advances in empirical research by enabling domain experts to carry out analyses that were not possible previously.

The Team. This project brings together a team of researchers that is uniquely qualified to carry out the proposed work. They have complementary expertise in data cleaning/curation, data provenance, and workflow provenance. They also have a proven track record of building open-source, widely-used systems.

References

[1] B. Arab, D. Gawlick, V. Krishnaswamy, V. Radhakrishnan, and B. Glavic. Formal foundations of reenactment and transaction provenance. Technical Report IIT/CS-DB-2016-01, Illinois Institute of Technology, 2016.

[2] B. Arab, D. Gawlick, V. Radhakrishnan, H. Guo, and B. 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.

[3] S. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. 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.

[4] S. P. Callahan, J. Freire, E. Santos, C. E. Scheidegger, C. T. Silva, and H. T. Vo. Managing the evolution of dataflows with vistrails. In *Data Engineering Workshops, 2006. Proceedings. 22nd International Conference on*, pages 71–71. IEEE, 2006.

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

[6] J. Freire, A. Bessa, F. S. Chirigati, H. Vo, and K. Zhao. Exploring what not to clean in urban data: a study using new york city taxi tripsa computational reproducibility benchmark. *IEEE Data Eng. Bull.*, 2016.

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

[8] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo. Managing rapidly-evolving scientific workflows. In *Provenance and Annotation of Data*, pages 10–18. Springer, 2006.

[9] B. Glavic and G. 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.

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

[11] B. Glavic, G. Alonso, R. J. Miller, and L. 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.

[12] B. Glavic, R. J. Miller, and G. 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.

[13] B. Howe, P. Lawson, R. Bellinger, E. W. Anderson, E. Santos, J. Freire, C. E. Scheidegger, A. Baptista, and C. 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.

[14] O. Kennedy, Y. Yang, J. Chomicki, R. Fehling, Z. H. Liu, and D. 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, Berlin, Heidelberg, 2015.

[15] D. Koop, C. E. Scheidegger, S. P. Callahan, J. Freire, and C. T. Silva. Viscomplete: Automating suggestions for visualization pipelines. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1691–1698, 2008.

[16] A. Nandi, Y. Yang, O. Kennedy, B. Glavic, R. Fehling, Z. H. Liu, and D. Gawlick. Mimir: Bringing ctables into practice. *CoRR*, abs/1601.00073, 2016.

[17] X. Niu, R. Kapoor, and B. Glavic. Heuristic and cost-based optimization for provenance computation. In *TaPP*, 2015.

[18] C. Scheidegger, H. Vo, D. Koop, J. Freire, and C. Silva. Querying and creating visualizations by analogy. *IEEE Transactions on Visualization & Computer Graphics*, 13(6):1560–1567, 2007.

[19] C. E. Scheidegger, H. Vo, D. Koop, J. Freire, and C. 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.

[20] C. T. Silva, J. Freire, and S. Callahan. Provenance for Visualizations: Reproducibility and Beyond. *Computing in Science and Engineering*, 9(5):82–89, 2007.

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

[22] Y. Yang, N. Meneghetti, R. Fehling, Z. H. Liu, and O. Kennedy. Lenses: an on-demand approach to etl. *Proceedings of the VLDB Endowment*, 8(12):1578–1589, 2015.