Cognitive Dimensions of Between-Table Context Support in Direct Manipulation Wrangling Interfaces

Steve Kasica

Nov. 4, 2019

Intro

On Nov. 19, 2015, *BuzzFeed News* published an article visually exploring the previous ten years of refugee immigration data in the United States (Singer-Vine, 2015). In this visual analysis is a bar chart comparing the US states receiving refugees per capita. Less than 24 hours after the article was published, *BuzzFeed News* issued a correction regarding this chart. The chart unintentionally omitted Wyoming. Because the author of the article also published the raw data and analysis code to GitHub, we know this omission was caused by a mismatch in keys in an outer-join operation on a table of received refugees by state with a table of state population. Because Wyoming did not accept any refugees, this state was completely absent from the refugee table. This example shows that issues commonly associated with *data wrangling* can have editorial consequences for data journalists.

Data wrangling includes all the operations data analysts, including journalists, have to do on their data in order to make it usable (Kandel, Heer, et al., 2011) for analysis. Under the umbrella of wrangling are the well-known area of data working, such as cleaning, integration, and transforming. Common data wrangling activities include: combining multiple data sources, fixing spelling mistakes, resolving duplicates, removing outliers, and reformatting table schemas.

This project builds upon work I conducted over the summer performing thematic analysis on the workflows of data journalists, with a focus on the wrangling activities of this user group. From a systematic process of open-coding 50 repositories of computational notebooks and scripts used in the reporting of published data journalism articles at newspapers including *The New York Times, The Washington Post*, and *The Guardian.* An analysis of the wrangling code written by data journalists can inform a broader discussion of data wrangling because the chief product of data wrangling isn't clean data, but "an editable and auditable description of the data transformations applied" (Kandel, Paepcke, Hellerstein, & Heer, 2011). These Jupyter notebooks are an editable, auditable, and reproducable description of data as they are being transformed in the wild.

When it comes to wrangling, I found that journalists operate in two different contexts. When operations consist entirely of manipulating the rows and columns of a table, then the user is working in a within-table context. However, there is also a between-table context that is characterized by using multiple tables in the course of making data usable. While many applications commonly used in data wrangling, such as Microsoft Excel, support a within-table context, there appears to be little support for users to operation in a between-table among commercial and open-sourced wrangling applications.

Therefore, the aim of this project is to compare the between-table context support in direct manipulation wrangling interfaces. To accomplish this, I will recreate the wrangling workflows originally conducted with programming language with two freely-available, GUI-based wrangling applications. The set of workflows used in this project include high-level wrangling tasks typically addressed in the between-table context.

Data and Tasks

There are two senses of data in this project: datasets and workflows. First, datasets refer to the actual data used by a journalist in the course of the reporting process. The types of datasets that will be used in this project come from federal governments, state governments, non-governmental organizations. These datasets are either archived in a public repository with analysis code or still publically available on the Internet.

The second kind of data are workflows captured in the computational notebooks and programming scripts produced by these journalists. The open coding process I conducted this summer partially abstract the API calls that describe how a table is transformed into higher-level descriptions of what the journalist is aiming to achieve. It will take a small amount of time to revise the sequence of these higher-level codes into a tool-independent sequences of wrangling activities. These tool-independent sequences will be the "script" for reproducing these workflows with either of the two GUI-based tool.

These datasets were identified through analysis that successfully wrangled the raw data into its final forms through manipulation via programming languages. It is understood that such data transformations are possible. What this project aims to understand is how well are these data transformations originally implemented in code are supported by direct manipulation interfaces, if at all.

Task Descriptions

Reproducing workflows on the same datasets with different tools will allow me to compare how well they address abstract tasks that data journalists perform. My previous work analyzing data journalism workflows identified many tasks; however, I will focus on a subset of tasks that I are easier to address in a between-table than a within-table context. The hyphenated strings of the analysis refer to the individual repository from my summer research.

Supplement a table with additional data

Description: When the variables from one table are supplemented with the variables from another, such as when two tables are merge with an outer join.
Analysis: 2015-11-refugees-in-the-united-states from *BuzzFeed News*
Datasets: 2014 US Census State Populations, Worldwide Refugee Admissions Processing System (WRAPS) arrivals by destination and region

Identify the occurrence of a lossy join

Description: When some observations from merging two tables is lost, such as when there is a mismatch between joining variables in two tables.
Analysis: 2015-11-refugees-in-the-united-states from *BuzzFeed News*
Datasets: 2014 US Census State Populations, Worldwide Refugee Admissions Processing System (WRAPS) arrivals by destination and region

Align joining keys

Description: When the entities of a variable in the table are modified to match the entities of another variable from a different table before joining the two tables on both variables.
Analysis: school-star-ratings-2018 from *The Baltimore Sun*
Datasets: Maryland State Department of Education Report Card

Quarantine and treat dirty data

Description: If one a subset of a table has a data quality issue, then that section of the table can be partitioned, addressed, and recombined with the rest of the data.
Analysis: federal_employees_trump_2017 from *The Washington Post*
Datasets: FedScope Employment Cube 2008 and 2016

Split, Compute, and Merge

Description: Splitting a table into multiple tables, computing upon those tables, and merging results

back together.
Analysis: california-cc-score from *The Los Angeles Times*
Datasets: California's State Water Resources Control Board water usage

Proposed Methods and Tools

Since no one data analysis incorporates all the tasks previous mentioned, the analysis of this project analysis will include several datasets corresponding to different tasks. It is understood that the dataset has certain characteristic that make it applicable to a specific task. For example, concerning the task of identifying a lossy join, it is known that the dataset contains a mismatch between the joining keys of the two tables. Each dataset in this project will begin in the raw form it took when it the journalist began the wrangling and analysis process.

There are many open source and commercial GUI-based tools that supporting data wrangling, including: Microsoft Excel, Google Sheets, OpenRefine, Workbench, Trifacta Wrangler, and Cloud Dataprep. The two direct manipulation interfaces that this project will compare are OpenRefine (Huynh, 2012) and Cloud Dataprep by Trifacta (Trifacta, n.d.). Both products were acknowledged as interchangeable data wrangling tools for data journalism (*Preparing data*, n.d.). OpenRefine, previously Google Refine, is an open source application for data wrangling that runs in the web browser. Cloud Dataprep is a serverless application for data wrangling that is based on Trifacta Wrangler. Dataprep is free to use under a 12-month trial period.

These data wrangling tools are visualization tools as they incorporate visualization idioms into their workflows. Both OpenRefine and Dataprep incorporate univariate visualizations, such as histograms and color bars, into its interface in order to aid the user in the wrangling process. Dataprep also includes a directed, acyclic node-link diagram in the form of a control flow graph to show provide a high-level view of how data is transformed.

In order to evaluate how well these tools satisfy the requirements of these tasks, this project will employ aspects the Cognitive Dimensions of Notation framework (Blackwell & Green, 2003), a common domain-agnostic vocabulary often used in evaluating visual interfaces. This project will substitute the six generic types of notation use activity for the tasks enumerated in Supporting Materials because this project is explicitly focused on characterizing the work of a specific user group, data journalists.

Related Work

To the best of my knowledge, there is no existing literature specifically on the evaluation of wrangling tools. Much of the existing literature evaluating wrangling tools comes for usage scenarios and user studies evaluating the utility or usability of a novel technique or design.

Origraph (Bigelow, Nobre, Meyer, & Lex, 2019), a graphical interface for wrangling network data, evaluated the utility of its design through two usage scenarios. This analysis is similar to usage scenarios than user studies. but more resilient to threats of external validity than a typical usage scenario as the scenarios in this analysis come from actual data journalism workflows.

Like Origraph, Wrangler (Kandel, Paepcke, et al., 2011) conducted a usage scenario. However, the usage scenario in Wrangler is more of a motivating example and description of the problems the interface addresses rather than an evaluation of its utility. Wrangler also conducted a user study that compared this early precursor to the commercial Trifacta Wrangler with Excel. While combining multiple datasets was an acknowledge activity central to the concept of wrangling in the paper, user study subjects only conducted tasks that could be entirely performed in one table.

While this project is thematically related to wrangling, methodologically it is closer related to work applying theoretical concepts for evaluating interactive systems to visualization tools. Critical reflection is a process by which the authors of similar visualization tools collaboratively compare and contrast the underlying frameworks, system architectures, and interface design of completed systems (Satyanarayan et al., 2019). Both the critical reflection approach and this project draw upon evaluate similar completed systems using the Cognitive Dimensions of Notation framework. However, this analysis is limited to the interface design since I am not a contributor to either OpenRefine nor Dataprep.

Milestones and Schedule

As I am working solo on this project, all this work will be done by myself. My current estimate is that it'll I currently estimate that it will require at the most two hours per analysis per tool. I should have reproduced at least five analyses in the course of this project.

Oct. 28 – Nov. 3: Draft proposal, 5 hours; review related work, 5 hours

Nov. 4 – Nov. 10: Finalize proposal, 3 hours; extrapolate workflows, 5 hours; further familiarization with tools, 3 hours

Nov 4, 10 p.m. Turn in project proposal

Nov. 11 – Nov. 16: Test OpenRefine vs. Data Prep in analyses with integration-like tasks, 10 hours

Nov. 18 – Nov. 23: Incorporating feedback, 10 hours

Nov 19: Peer Project Review 1

Nov. 25 – Nov. 3: Test OpenRefine vs. Data Prep in analysis aggregation-like tasks, 10 hours

Dec. 2 – Dec. 7: Incorporating feedback, 10 hours

Dec 4: Peer Project Review 2

Dec. 9 – Dec. 13: Final polish, presentation, and revisions.

Dec. 10 Final presentation

Dec. 13 Final Paper Due at 11:59 p.m.

Bibliography

Bigelow, A., Nobre, C., Meyer, M., & Lex, A. (2019). Origraph: Interactive Network Wrangling. *Proc. IEEE VAST 2019*, 12.

Blackwell, A., & Green, T. (2003). Notational Systems—The Cognitive Dimensions of Notations Framework. In *HCI Models, Theories, and Frameworks* (pp. 103–133). https://doi.org/10.1016/B978-155860808-5/50005-8

Huynh, D. (2012). *Open Refine*. Retrieved from openrefine.org

Kandel, S., Heer, J., Plaisant, C., Kennedy, J., van Ham, F., Riche, N. H., … Buono, P. (2011). Research Directions in Data Wrangling: Visuatizations and Transformations for Usable and Credible Data. *Information Visualization*, *10*(4), 271–288. https://doi.org/10.1177/1473871611415994

Kandel, S., Paepcke, A., Hellerstein, J., & Heer, J. (2011). Wrangler: Interactive Visual Specification of Data Transformation Scripts. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3363–3372. https://doi.org/10.1145/1978942.1979444

Preparing data. (n.d.). Retrieved from https://journalismcourses.org/course/view.php?id=44&section=3

Satyanarayan, A., Lee, B., Ren, D., Heer, J., Stasko, J., Thompson, J., … Liu, Z. (2019). Critical Reflections on Visualization Authoring Systems. *IEEE Transactions on Visualization and Computer Graphics*. https://doi.org/10.1109/TVCG.2019.2934281

Singer-Vine, J. (2015, November 19). Where U.S. Refugees Come From—And Go—In Charts. Retrieved November 4, 2019, from BuzzFeed News website: https://www.buzzfeednews.com/article/jsvine/where-us-refugees-come-from-and-go-in-charts

Trifacta. (n.d.). *Cloud Dataprep*.