Streamlining and Understanding Curation with Vizier (Award #1640864)

Data curation or wrangling, is a critical stage in data science in which raw data is validated and repaired to establish trust in the data, and structured to streamline analytics. Traditionally, data curation has been performed as a pre-processing task: only after all the data selected for a study (or application) are curated, are they ready to be loaded into an analytics system for use. 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. In short, curation and exploration are both part of an ongoing, *iterative* process.

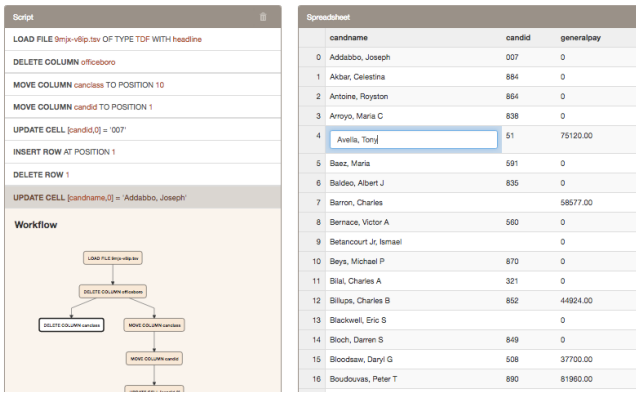

Figure 1: Prototype Vizier User Interface

Vizier will link exploration and curation and allows analysts to leverage the full power of their existing SQL-based analytics platform to explore data, even if it has not yet been fully curated. As the analyst explores and queries her data, Vizier will present her with provenance information, quality assesments, and opportunities for quality improvement relevant to her current exploration efforts. Provenance and quality information help the analyst to evaluate whether she can trust the data she's looking at. If she decides that more curation effort is required, Vizier can help her to direct her curation efforts.

Vizier's intuitive hybrid notebook-spreadsheet interface will make it easy for analysts to gracefully transition from preliminary data surveys (best done in a spreadsheet) to more rigorous, procedural data manipulations (best done in an imperative language). As shown in the prototype interface above, Vizier presents users with both a tabular spreadsheet-style interface, and a script interface. Edits to the spreadsheet are reflected as operations in the script, while bulk procedural transformations in the script are decomposed into a single, editable expressions in the spreadsheet. As a result, analysts can both transform data through quick, visual edits, as well as bulk, set-at-a-time scripting and automated data curation heuristics, e.g., missing value imputation.

Preliminary Challenges

Vizier will be composed of three existing systems: GProM — A system for generic fine-grained provenance queries, Mimir — A system for probabilistic data curation, and VizTrails — A system for workflow management and provenance. Our first year is largely dedicated to getting these systems to work together.

Integrating Different Provenance Granularities: All three systems adopt different provenance models. Over the coming year, we will need to develop a multi-level provenance model with the attribute-level detail required by Mimir, but without sacrificing the simplicity of coarse-grained provenance.

Defining Cleaning Workflows: Vizier requires a scripting language that can gracefully capture spreadsheet interactions and common curation tasks. Our first year goals include a draft of this language.

Bulletproofing: GProM and Mimir are both academic projects. Our goals for the first year include bulletproofing, testing, and extending these systems to meet the requirements of the Vizier system.

Interface Design: Over the first year, we will work with domain experts — particularly those without an extensive programming background — to understand their existing curation workflows and how Vizier's interface can be designed to best fit their needs.

Systems Integration: Finally, there is the mechanical challenge of getting systems written in different languages to talk to one another. We are making progress and do not anticipate significant roadblocks.