Streamlining and Understanding Curation with Vizier

PIs: Juliana Freire, Oliver Kennedy, Boris Glavic

Award #1640864

Data Curation is Hard

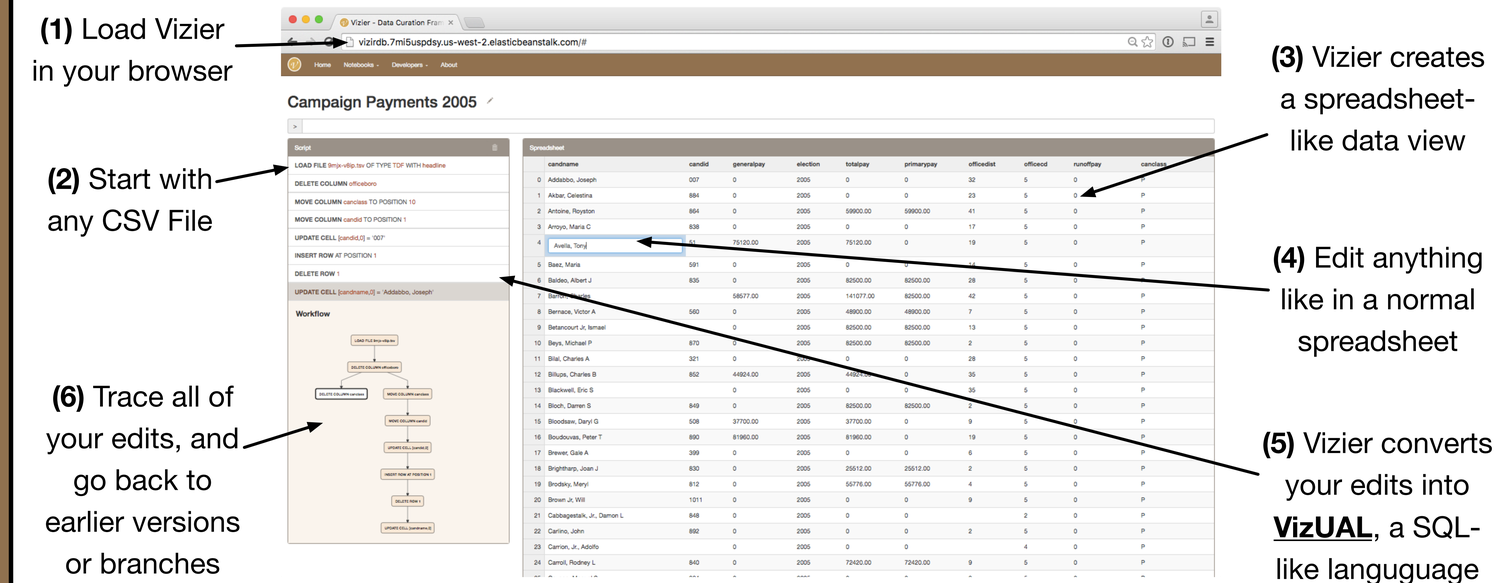
- It's hard to tell what's wrong until you see or play with a dataset.
- You might not know there are errors until your analysis is under way.

Vizier Will Make It Easier

… lets analysts leverage their existing data management infrastructures.

… tracks provenance to help you find errors *after* you start asking questions.

… uses an innovative interface to make data exploration faster and easier.



SQL

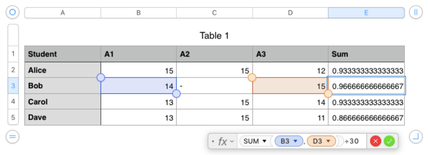
```
SELECT student_id,
       SUM(assignment * weight) / 45
FROM grades;
```

Great at batch processing

Hard to declare special cases

(Ignore Bob's assignment 3)

Spreadsheets



Everything is a special case

Not as good at batch processing

Vizier Combines…

VisTrails: Open source tool for managing workflows and coarse-grained provenance for data visualizations

Mimir: Open source tool for querying and incremental curation of messy data.

GProM: Open source tool for fine-grained provenance and SQL query introspection.

Vizier Adds…

- A hybrid spreadsheet/ notebook UI
- Integrated workflow and query-level provenance
- A seamless environment for exploring and incrementally cleaning large, messy data

First-Year Efforts

- Integrating three provenance models that operate at different granularities
- Define and formalize VizUAL, a DSL for data curation and exploration
- Bulletproofing Mimir and GProM
- Expert studies to better understand workflows and interface requirements.
- Systems Integration Engineering

University at Buffalo | NYU TANDON SCHOOL OF ENGINEERING | ILLINOIS INSTITUTE OF TECHNOLOGY