

# 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.

(1) Load Vizier in your browser

(2) Start with any CSV File

(6) Trace all of your edits, and go back to earlier versions or branches

(3) Vizier creates a spreadsheet-like data view

(4) Edit anything like in a normal spreadsheet

(5) Vizier converts your edits into **VizUAL**, a SQL-like language

## 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



Student	A1	A2	A3	Sum
Alice	15	15	12	0.9333333333333333
Bob	14	15	15	0.9666666666666667
Carol	13	15	14	0.9333333333333333
Dave	13	15	11	0.8666666666666667

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