Streamlining and Understanding Curation with Vizier **Pls: 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.

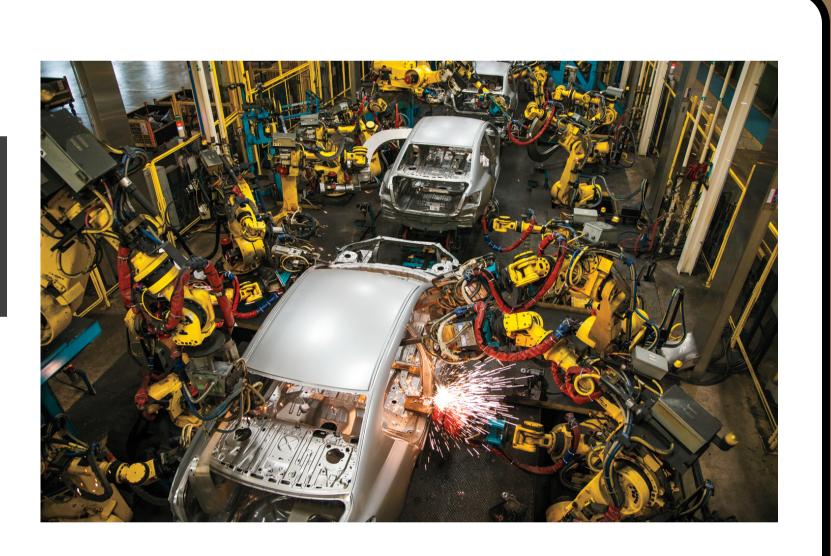
Vizier Will Make It Easier

- ... 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.

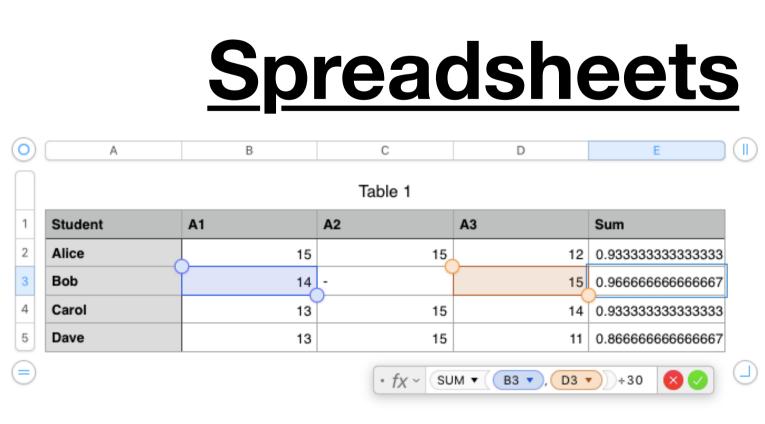
SQL

SELECT student id, SUM(assignment * weight) / 45 FROM grades;

Great at batch processing Hard to declare special cases *(Ignore Bob's assignment 3)*







Everything is a special case Not as good at batch processing

(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



Vizier Combines...

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

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

<u>GProM</u>: Open source tool for finegrained 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



Û	Spre	edsheet							
PE TDF WITH headline		candname	candid	generalpay	election	totalpay	primarypay	officedist	officecd
	0	Addabbo, Joseph	007	0	2005	0	0	32	5
OSITION 10	1	Akbar, Celestina	884	0	2005	0	0	23	5
SITION 1	2	Antoine, Royston	864	0	2005	59900.00	59900.00	41	5
	3	Arroyo, Maria C	838	0	2005	0	0	17	5
	4	Avella, Tony	51	75120.00	2005	75120.00	0	19	5
	5	Baez, Maria	591	0	2005	0	0	14	5
	6	Baldeo, Albert J	835	0	2005	82500.00	82500.00	28	5
'Addabbo, Joseph'	7	Barron, Charles		58577.00	2005	141077.00	82500.00	42	5
officeboro MOVE COLUMN canclass	8	Bernace, Victor A	560	0	2005	48900.00	48900.00	7	5
	9	Betancourt Jr, Ismael		0	2005	82500.00	82500.00	13	5
	10	Beys, Michael P	870	0	2005	82500.00	82500.00	2	5
	11	Bilal, Charles A	321	0	2005	0	0	28	5
	12	Billups, Charles B	852	44924.00	2005	44924.00	0	35	5
	13	Blackwell, Eric S		0	2005	0	0	35	5
	14	Bloch, Darren S	849	0	2005	82500.00	82500.00	2	5
MOVE COLLUNN candid	15	Bloodsaw, Daryl G	508	37700.00	2005	37700.00	0	9	5
	16	Boudouvas, Peter T	890	81960.00	2005	81960.00	0	19	5
	17	Brewer, Gale A	399	0	2005	0	0	6	5
INSERT ROW AT POSITION 1	18	Brightharp, Joan J	830	0	2005	25512.00	25512.00	2	5
	19	Brodsky, Meryl	812	0	2005	55776.00	55776.00	4	5
DELETE ROW 1	20	Brown Jr, Will	1011	0	2005	0	0	9	5
	21	Cabbagestalk, Jr., Damon L	848	0	2005	0	0		2
	22	Carlino, John	892	0	2005	0	0	2	5
	23	Carrion, Jr., Adolfo		0	2005	0	0		4
	24	Carroll, Rodney L	840	0	2005	72420.00	72420.00	9	5
	0.5	a		•	0005	•	^	•	r

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



