Data Debugging and Exploration with Vizier

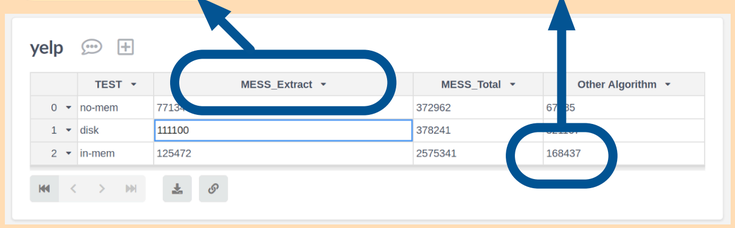
Mike Brachmann[1], Carlos Bautista[2], Sonia Castelo[2], Su Feng[3], Juliana Freire[2], Boris Glavic[3], Oliver Kennedy[1], Heiko Müller[2], Rémi Rampin[2], William Spoth[1],Ying Yang[4]

[1]University at Buffalo,[2]New York University,[3]Illinois Institute of Technology,[4]Oracle

Painless Data Ingest

Detects Headers

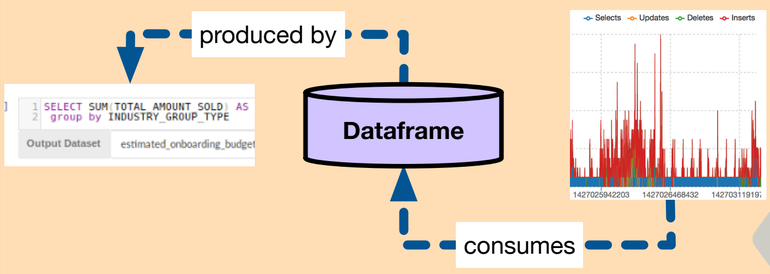
Infers Types



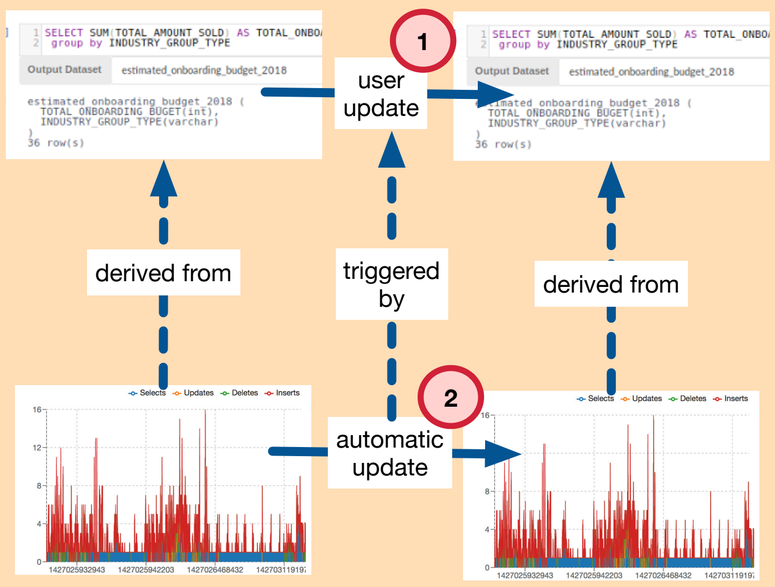
Supports your favorite data format and source

Data-Centric

Interoperability through data



Automatic refresh of dependent objects



Supporting new languages

Implement a library for Vizier's dataset API

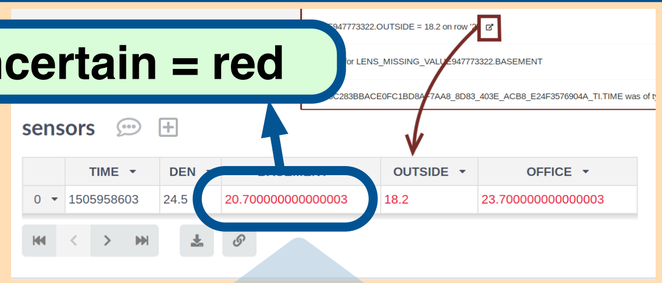
Dependency-tracking for free by monitoring dataset API calls

Languages never communicate directly! No interoperability issues!

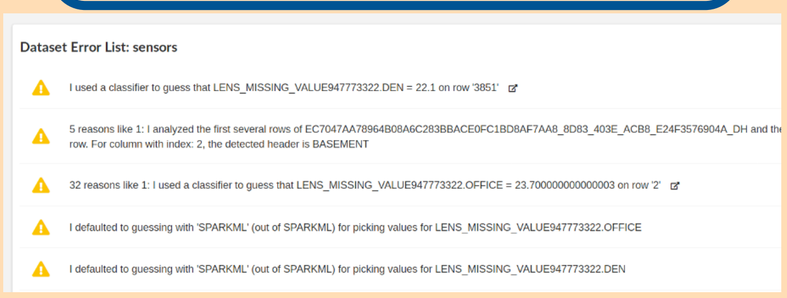
Data Testing & Debugging

Automated Curation/Cleaning Ops

uncertain = red



Error tracking & annotations

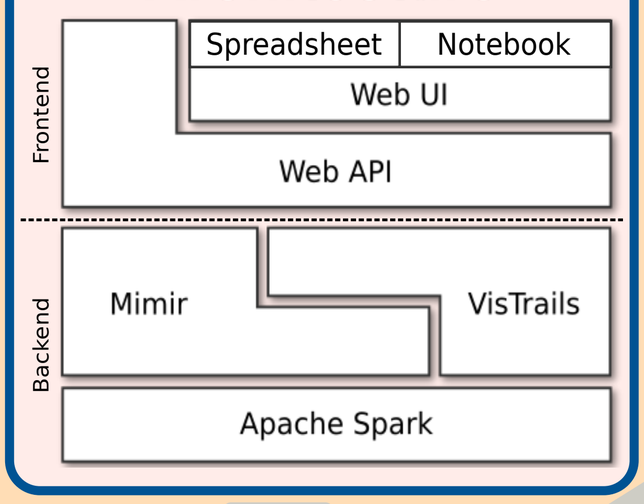


What is Vizier?

A next generation data science platform that…
- supports iterative construction of pipelines
- provides expressive interaction paradigms without sacrificing ease-of-use
- aids data debugging & testing
- scales
- provides tracebility and provenance
- aids collaboration and reuse

data-centric!

Architecture

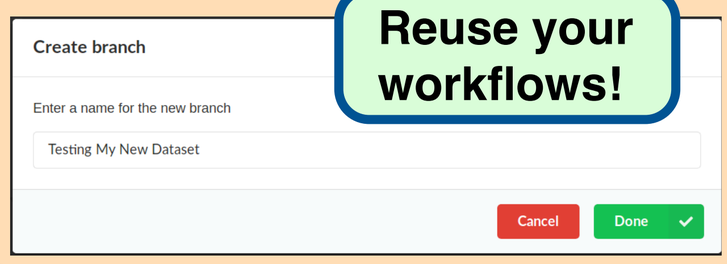


Versioning & Collaboration

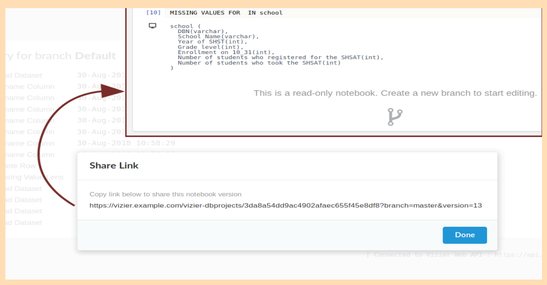
Automatic commit on edit!

Browse old versions and create branches

Reuse your workflows!



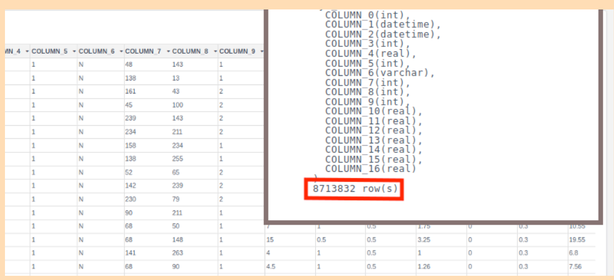
Share links to any version of
… a dataset
… a notebook
… a plot



Scalable

Backed by Spark

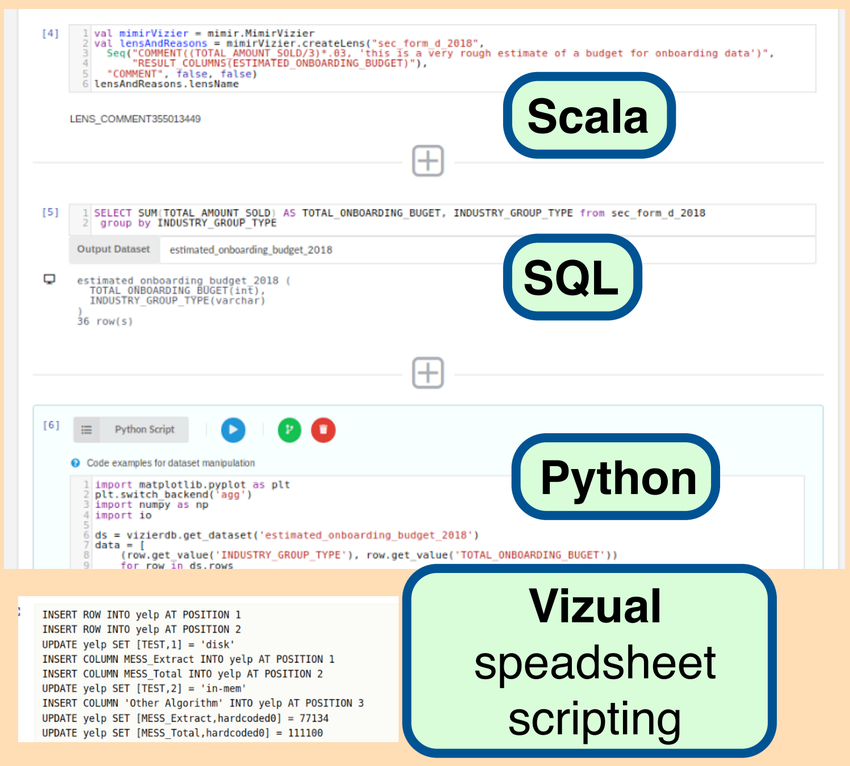
Vizier dataset = Spark dataframe



Interested in how Vizier tracks uncertainty?

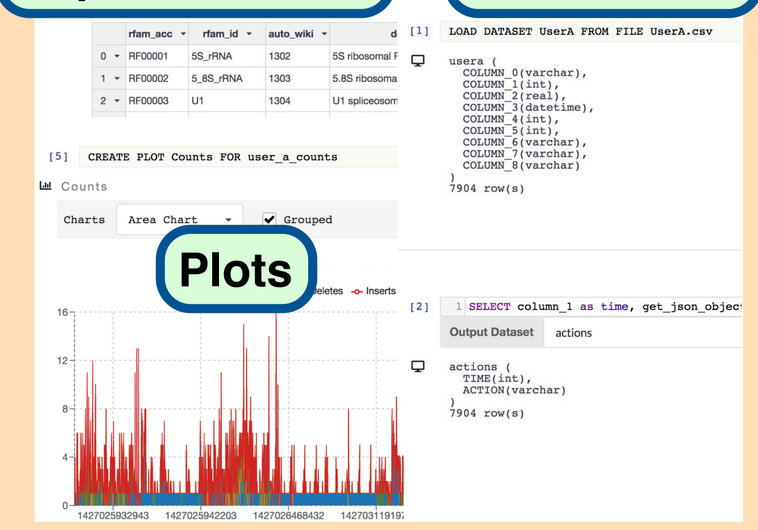
Come to our research talk [1]:
Thu, 11:30-12:50 SIGMOD Research 13, Administratiezaal

Multi-lingual + Multi-modal



Spreadsheets

Notebooks



Project Page

Demo Video

Github Repo

[1] S. Feng, A. Huber, B. Glavic, O. Kennedy. **Uncertainty Annotated Databases - A Lightweight Approach for Approximating Certain Answers.** SIGMOD 2019.