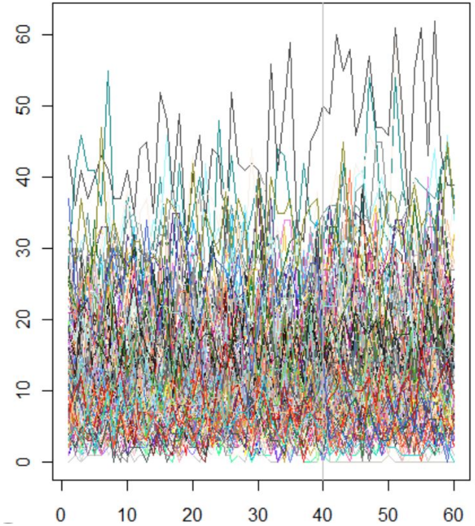
Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data

Authors: Dominik Moritz, Danyel Fisher, Bolin Ding, Chi Wang

Presenter: Alice Yeh

Obstacles in Exploratory Visualization

- Scientists want to derive insights from large datasets
- But…
  - Screen cannot render so much data and visualizations get cluttered
  - Database queries take a long time to return



Approximate Query Processing (AQP)

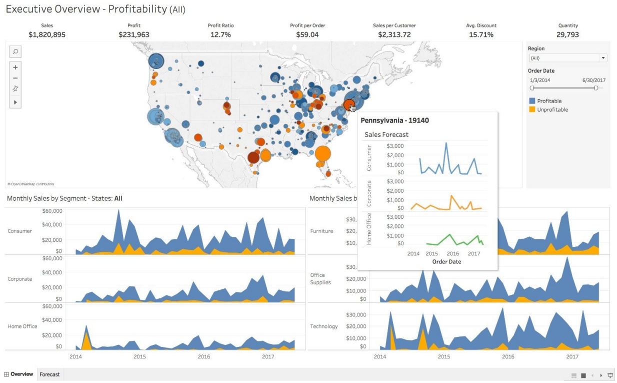
- Sample the dataset
  - Allows visualization in interactive time using approximate values
- Does this solve all our problems?
  - Approximate values can be incorrect
  - Can we trust approximate values with business-critical decisions? What if we take multiple samples–at what point are these visualizations reliable?

Optimistic Visualization

- Produces approximate results quickly and computes precise results in the background
- Best of both worlds
  - Speed of approximation and ability to check for precision

What Has Been Done in this Space?

- Exploratory visualization
  - Iterative process (broader → specific questions) that prioritizes speed
  - Enabled by visualization tools (Tableau, PowerBI, Matplotlib)
- Big data visualization
  - Data retrieval and processing are bottlenecks
    - Offline processing phase
    - Online Analytical Processing (OLAP) systems

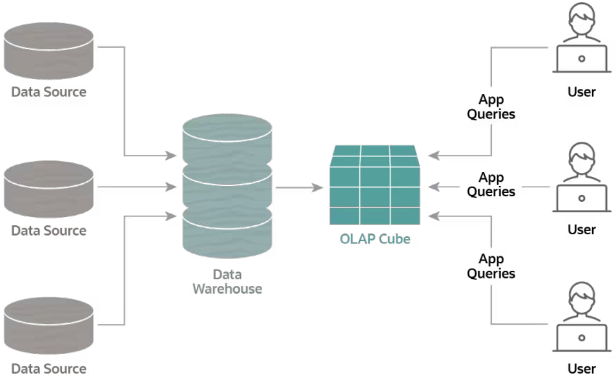


Executive Overview - Profitability (All)

Sales	Profit	Profit Ratio	Profit per Order	Sales per Customer	Avg. Discount	Quantity
$1,820,895	$231,963	12.7%	$59.04	$2,313.72	15.71%	29,793

Region: (All)

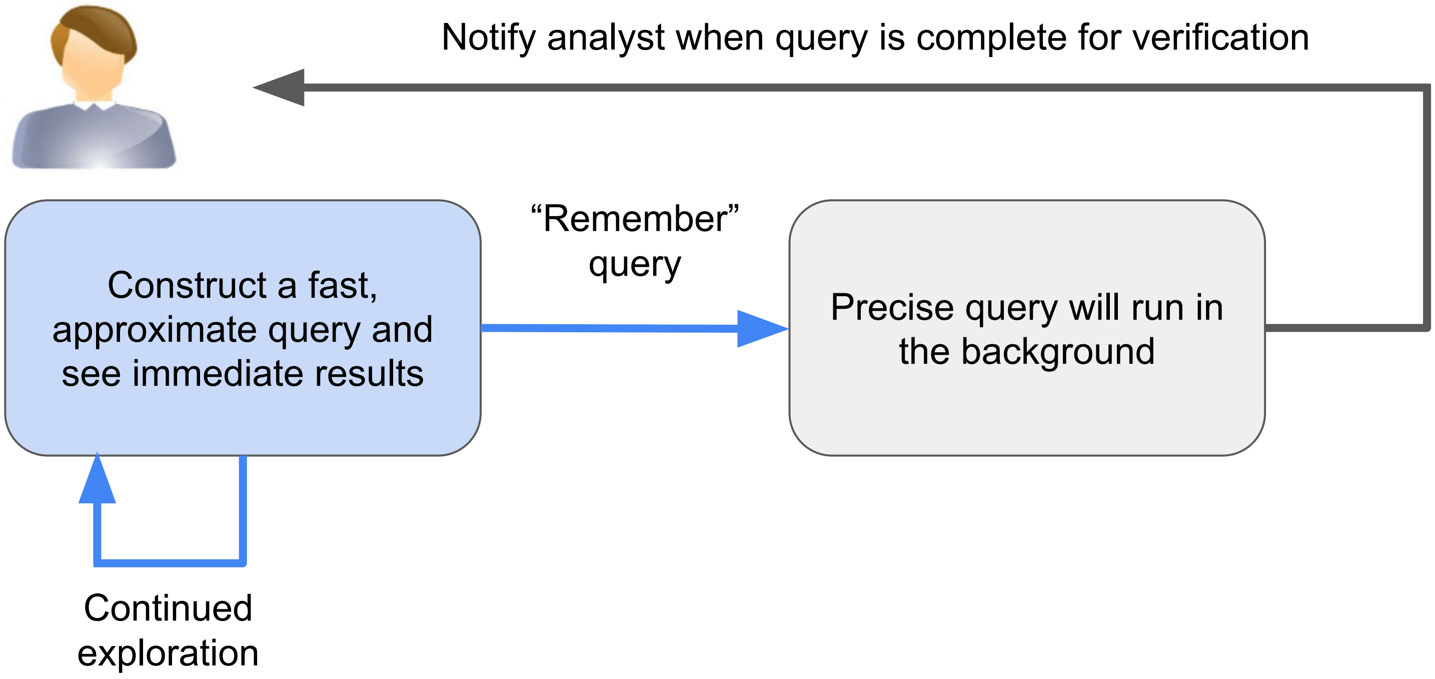
Order Date: 1/3/2014 – 6/30/2017



What Has Been Done in this Space?

- Approximate query processing
  - Look at less data more quickly
  - Available tools…
    - Create sample of data before user begins analysis but precision diminishes as analyst filters records
    - Pick a sample and compute results with estimated error bounds but it is up to the analyst to choose between max query runtime or error bound
- Progressive visualization with online aggregation (OLA)
  - Picks increasing sample sizes and displays results, user decides when to end process
  - Optimistic visualization is asynchronous form of this

Optimistic Visualization, Visualized



Notify analyst when query is complete for verification

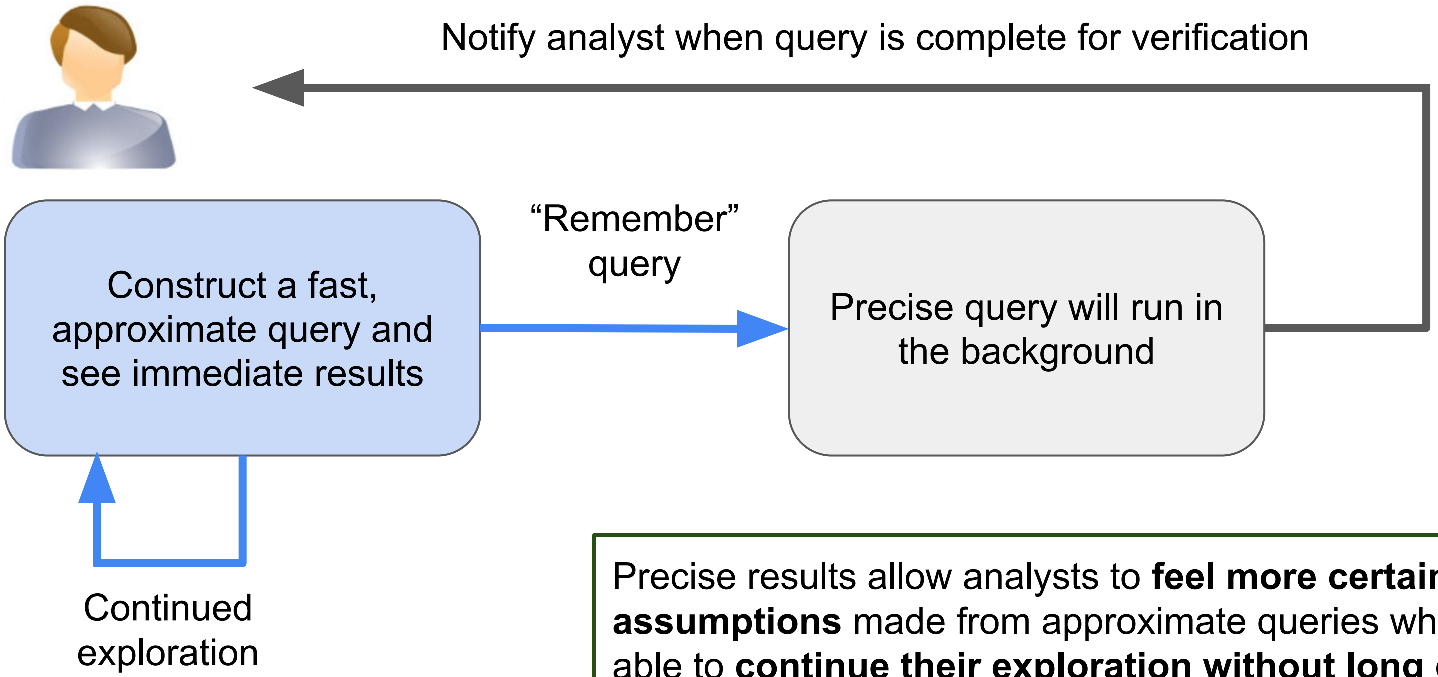
Construct a fast, approximate query and see immediate results

“Remember” query

Precise query will run in the background

Continued exploration

Optimistic Visualization, Visualized



Notify analyst when query is complete for verification

Construct a fast, approximate query and see immediate results

"Remember" query

Precise query will run in the background

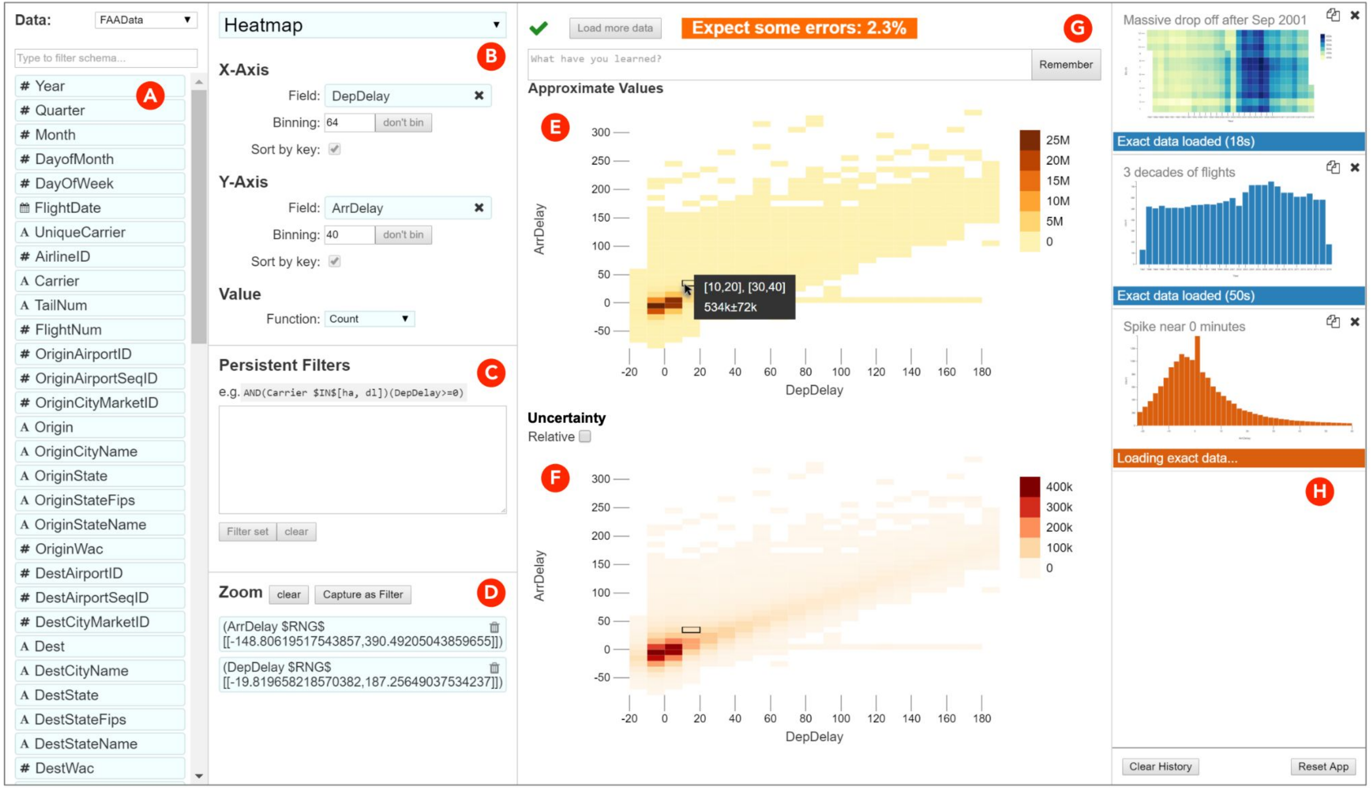
Continued exploration

Precise results allow analysts to **feel more certain about assumptions** made from approximate queries while being able to **continue their exploration without long delays**.

Pangloss: Optimistic Visualization Tool with AQP

- Enables analysts to rapidly explore large multi-dimensional datasets
    - Grouping, aggregating, filtering functionality
- Web based UI that queries Sample+Seek (AQP system)
    - Highly responsive to aggregate queries on a single table
    - Incrementally loads more records until uncertainty bound is below a threshold or timeout
    - Uses measure-biased sampling
        - Fewer samples necessary for same accuracy (vs. uniform sampling)
        - Optimizes distribution uncertainty

Pangloss UI



Data: FAAData

X-Axis
Field: DepDelay
Binning: 64
Sort by key: ☑

Y-Axis
Field: ArrDelay
Binning: 40
Sort by key: ☑

Value
Function: Count

Persistent Filters
e.g. `AND(Carrier $IN$[ha, dl])(DepDelay>=0)`

Zoom
(ArrDelay RNG [[-148.80619517543857,390.49205043859655]])
(DepDelay RNG [[-19.819658218570382,187.25649037534237]])

Expect some errors: 2.3%

What have you learned?

Approximate Values

[10,20], [30,40]
534k±72k

Uncertainty
Relative □

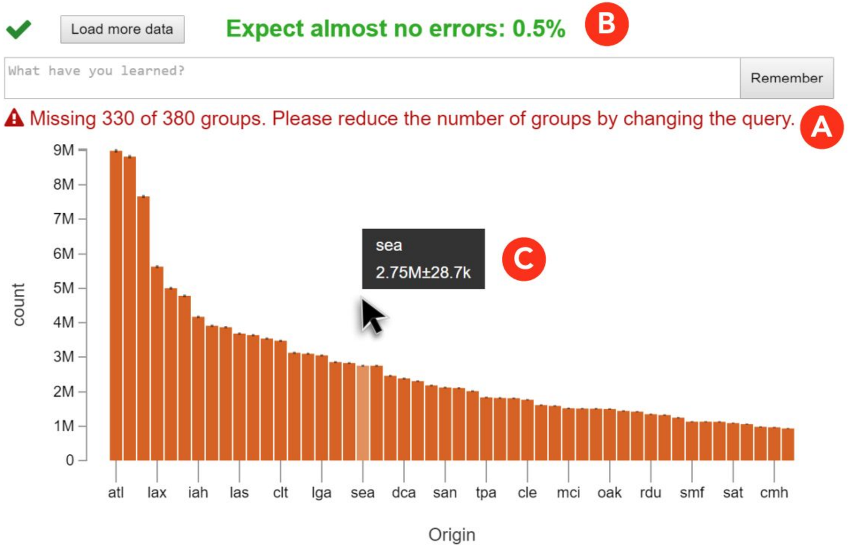
Massive drop off after Sep 2001
Exact data loaded (18s)

3 decades of flights
Exact data loaded (50s)

Spike near 0 minutes
Loading exact data...

Visualizations in Pangloss

- Approximate visualizations
  - Displays top *k* bars or cells to deal with queries with long tails
  - Distribution uncertainty is displayed
- Zooming and filtering
  - Operations will force new query to run → aggregate values and uncertainty can change
  - Negative filtering semantics
- Data transformations

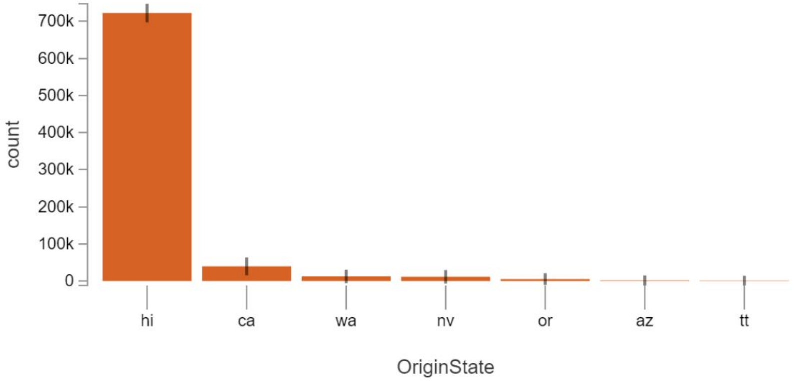


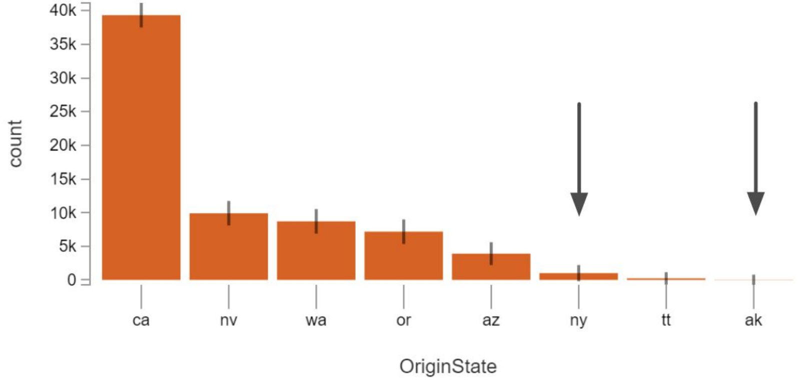
Expect almost no errors: 0.5%

Missing 330 of 380 groups. Please reduce the number of groups by changing the query.

Visualizations in Pangloss

- Approximate visualizations
  - Displays top *k* bars or cells to deal with queries with long tails
  - Distribution uncertainty is displayed
- Zooming and filtering
  - Operations will force new query to run → aggregate values and uncertainty can change
  - Negative filtering semantics
- Data transformations

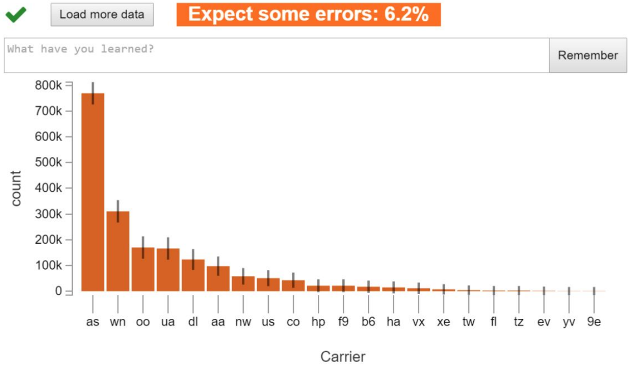


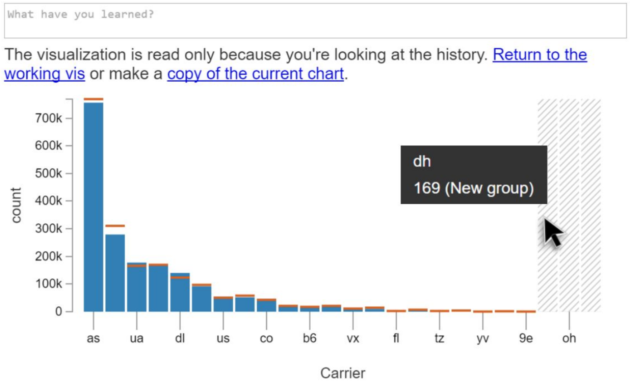


“Remembering” Views

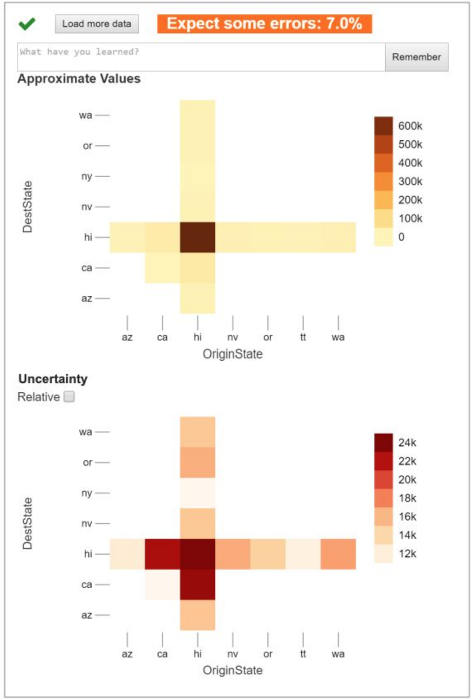
- “Remembering” view = re-running query to get precise result
- Which views should we remember?
    - Strawman: remember all past views
        - Overwhelming for user to review all views
        - Computationally expensive
    - Make it an explicit process → have users decide
- Design choices
    - “Remember” button for users to specify views to store and re-run
    - Render approximate views in orange and precise ones in blue

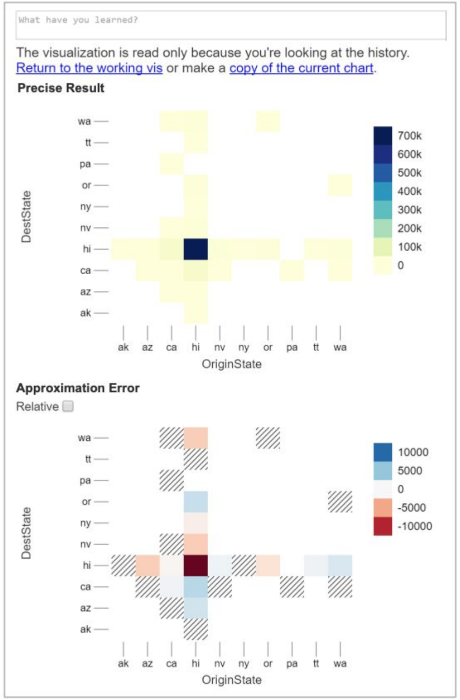
“Remembering” Views





The visualization is read only because you're looking at the history. Return to the working vis or make a copy of the current chart.





The visualization is read only because you're looking at the history. Return to the working vis or make a copy of the current chart.

User Studies

- Motivating questions
  - Comfort with incomplete or inaccurate results and usage towards exploring approximate data
  - Proceed with exploration without knowing precise results
  - Checking precise results interrupting flow
- User studies
  - Pangloss as a data analytics toolkit for usable insights
  - Pangloss applied to real-world systems

Flight Delay Study

- BTS Flight Delays dataset (70 GB), 5 data scientists (familiar w/ visualization tools)
- Participant sessions (1 hour)
    - Tutorial w/ training questions
    - Exploratory analysis with tool, prompted by introductory questions
    - Encouragement to review precise results
- Results
    - 4 users regularly “remembering” visualizations (4-7 views)
        - Most usage of “remember” functionality for uncertain results
    - Appreciation for speed of Pangloss
    - Limitation: cannot see and interact with lower-level data

Real-World Case Studies

- 3 users with >10 GB datasets
- Search terms case study
  - Analyst working on a search engine advertising platform, interested in predicting trends in searches and keywords
  - Dataset pre-aggregated with 994M rows of data
  - Usage: bulk of time spent on heatmap with a dozen or more keywords at a time
  - Trends discovered: weekly pattern, spike over a month
  - Follow up with request to use Pangloss again with a less aggregated version of the dataset

User Study Findings

- Users see precision broadly
  - Want rapid interaction in exploratory phase and precise results for presenting to decision makers
- Recording observations and “remembering” views is a useful feature
- More features desired
  - Ability to see underlying data
  - More transformations, aggregations, and projections options

Summary

- Optimistic visualization enables the benefits of both speed and precision
- Pangloss is an optimistic visualization tool that serves quick visualizations on approximate data and runs user-selected queries on precise data in the background
- User studies have shown value in Pangloss's workflow, allowing for rapid interaction during exploration and precise results for critical decision making