WHAT USERS DON’T EXPECT ABOUT EXPLORATORY DATA ANALYSIS ON APPROXIMATE QUERY PROCESSING SYSTEMS

Dominik Moritz @domoritz
Paul G. Allen School of CSE
University of Washington

Danyel Fisher @FisherDanyel
HCI
Microsoft Research
Big Data Visual Analysis
State of the Art in Big Data Exploration

Distributed Systems
Expensive and high latency.

Indexes (Data Cubes)
Requires pre computation and limited queries.

Sampling
Use a representative subset of the data.
Sampling and Approximate Query Processing (AQP)

Use a representative subset of the data and estimate the true values of aggregate results.

 Decide on **acceptable uncertainty** or **timeout**

Sum of 25% = 42

Sum of 100% = 168 ±10

Estimate

Uncertainty
Progressive Visualization with Online Aggregation

Growing sample ➔ continuously improving results

Analysts watch updates until bounds errors are low enough

Sum of 25% = 82
Sum of 100% = 168 ± 50
Challenges with AQP

- Approximate results
  - Convey uncertainty
- Probabilistic guarantees
- Unbounded errors
- Arbitrary aggregation or joins
Optimistic Visualization

A UX approach to challenges with AQP for visual data analysis traditionally treated as database problems.
Optimistic Visualization

Assume that approximation is mostly right but offer a way to detect and recover from mistakes.

Analysts use initial estimates, run precise query in background, and confirm results later.

Gives users confidence in using AQP.
Pangloss implements Optimistic Visualization

Query Specification
Pangloss implements Optimistic Visualization
Pangloss implements Optimistic Visualization

Approximation

Expected Error (Uncertainty)
Pangloss implements Optimistic Visualization

Annotation + Remember Button
Pangloss implements Optimistic Visualization
Pangloss implements Optimistic Visualization
170 Million flights (30 years).

~100ms query time
Text annotations help analysts clarify observations.
"Remember" button moves query into the background
Continue exploration without waiting
The visualization is read only because you're looking at the history. Return to the working vs or make a copy of the current chart.

Difference Visualization
Evaluation

Lab Study
5 users
Flight delay data
(170 Million records)
1 hour each

Case Study
3 teams
Product insights,
Social media,
Bing
~1+ hour exploration
Findings from the study

AQP works: “seeing something right away at first glimpse is really great”

Optimism works: “I was thinking what to do next— and I saw that it had loaded, so I went back and checked it . . . [the passive update is] very nice for not interrupting your workflow.”

Need for guarantees: “[with a competitor] I was willing to wait 70-80 seconds. It wasn’t ideally interactive, but it meant I was looking at all the data.”
Findings from the study (cont)

“When I’m using your system, there is a path that I need to follow.”

“Now that I’ve been sitting here for an hour, after I go back, it makes a lot of sense [to have these annotations], but as I was doing it, I was thinking, ‘I want to move on, I want to move on.’”
Adopt Optimistic Visualization

Uncertainty Visualization is not strictly required

Precise query can benefit from highly optimized Databases

Optimistic Visualization can help with adoption of AQP
Understanding Approximation Error

Approximation Error
The true error of the approximation. Only known after we run the full query!

Uncertainty
Expected approximation error.
2D Uncertainty

No best practices

Currently: juxtaposed heatmaps
2D Uncertainty

Percentage different? vs Value different?
Distribution Uncertainty

**Approximation**
- Distribution Uncertainty: 4

**Outside Distribution Uncertainty**
- Error: 6
- Sum: 12

**Within Distribution Uncertainty**
- Error: 4
- Sum: 12

Error:
- Sum: 4
- Sum: 12
Distribution Uncertainty

Expect almost no errors: 0.5%

⚠ Missing 330 of 380 groups. Please reduce the number of groups by changing the query.
What are some samples that created this value?
Optimizing the Language for Data Exploration

Tweaking SQL for high-level operations & sessions

```
SELECT HISTOGRAM(DISTANCE) WITH ALGORITHM="nice"
SELECT HISTOGRAM(DISTANCE) WITH BUCKETS=(0,10,20,30)
```

Knowing what queries are related in an exploration session enables new optimizations, e.g. ForeCache.
More complex filters
= more samples
= slower performance
Filtering can show new groups

New Predicate

New Query

Different Sample

Different Groups
Precise results can show new groups
Vocabulary of visual cues

Heatmap

Barchart
Conclusions

Optimistic Visualization addresses fundamental problems with AQP as **UX problem**

UI tools make invalid assumptions, AQP tools are not designed for visual analytics

Need to continue exploring the UX issues with AQP
AQP needs Multi-Disciplinary Solutions

Dominik – Viz+DB  Chi – DB  Danyel – HCI  Bolin – DB
Trust, but Verify: Optimistic Visualizations of Approximate Queries for Exploring Big Data

Dominik Moritz  
University of Washington  
domoritz@cs.uw.edu

Danyel Fisher  
Microsoft Research  
danyelf@microsoft.com

Bolin Ding, Chi Wang  
DMX, Microsoft Research  
bolind@microsoft.com, chiw@microsoft.com

ABSTRACT
Analysts need interactive speed for exploratory analysis, but big data systems are often slow. With sampling, data systems can produce approximate answers fast enough for exploratory visualization, at the cost of accuracy and trust. We propose optimistic visualization, which approaches these issues from a user experience perspective. This method lets analysts explore approximate results interactively, and provides a way to detect and recover from errors later. Pangolin implements these ideas. We discuss design issues raised by optimistic visualization.

In this paper, rather than addressing the problems with AQP from an algorithmic or systems perspective, we formulate them as user experience problems. What user experience would enable analysts to gain the benefits of approximate queries, while still being able to trust the results?

We propose an approach which we call optimistic visualization. Optimistic visualization produces approximate results quickly, and computes precise results in the background. The analyst can make observations on the approximation, and later check the results. Pangolin allows analysts to explore approximate results quickly, and provides a way to detect and recover from errors later.
What Users Don’t Expect about Exploratory Data Analysis on Approximate Query Processing Systems

Optimistic Visualization addresses fundamental problems with AQP as UX problem

UI tools make invalid assumptions, AQP tools are not designed for visual analytics

Need to continue exploring the UX issues with AQP

Dominik Moritz @domoritz
Danyel Fisher @FisherDanyel
Bolin Ding @AtlasDing
Chi Wang