Trust But Verify

Optimistic Visualizations of Approximate Queries for Exploring Big Data

Dominik Moritz @domoritz

Paul G. Allen School of CSE
University of Washington

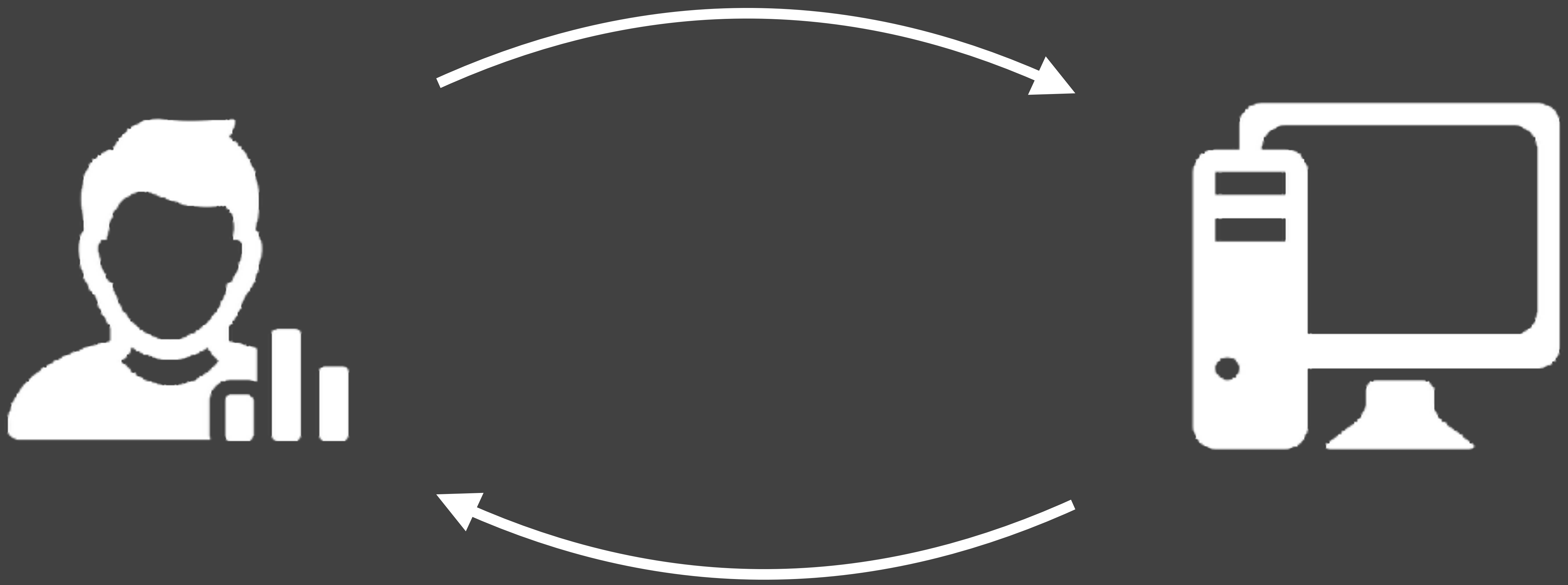
Danyel Fisher @FisherDanyel
Bolin Ding @AtlasDing
Chi Wang

HCI and DMX
Microsoft Research

What's the distribution of flight distances?

```
$ wget https://www.transtats.bts.gov/download.zip
==========================================> 70GB
-> Done
$ import download.zip
-> Done
$ SELECT bin(distance), count(*)
  FROM flights
-> Running Query. Please wait ...
```

Visual Analysis



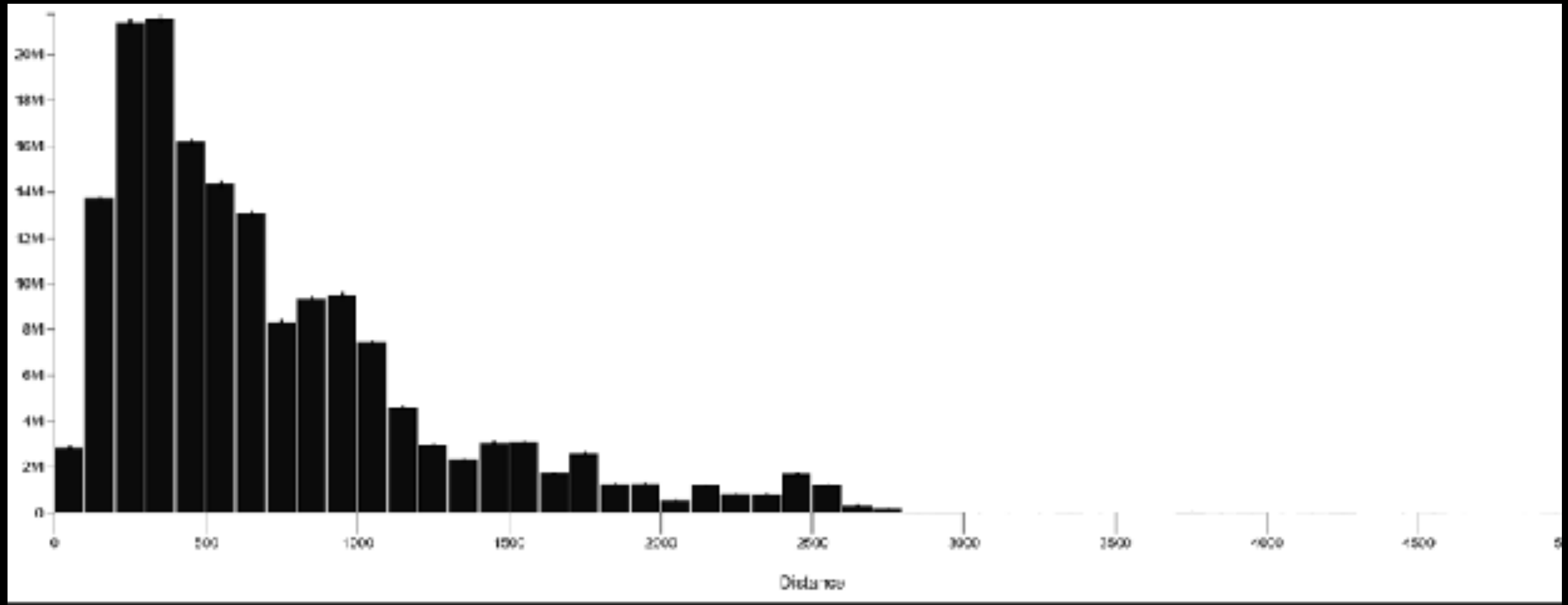
Computer by Simple Icons from the Noun Project
analyst by Gregor Cresnar from the Noun Project

Big Data Visual Analysis

`Query finished!`

```
$ SELECT bin(distance), count(*)
  FROM flights
```

->

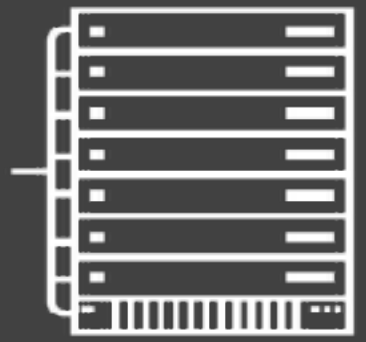


```
$ SELECT bin(distance), count(*)
  FROM flights
  WHERE airline = 'hi'
```

-> Running Query. Please wait ...

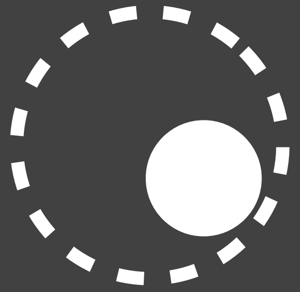
State of the Art in Big Data Exploration

Distributed Systems
Expensive and high latency.



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

Sampling
Use a representative subset of the data.



Rubik's Cube by Aleks from the Noun Project

Cluster servers by Branis Panos from the Noun Project

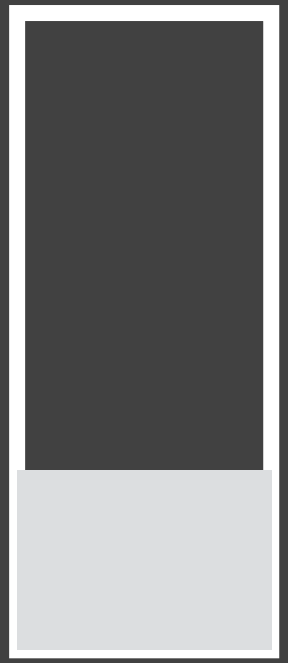
Sampling and Approximate Query Processing (AQP)

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

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

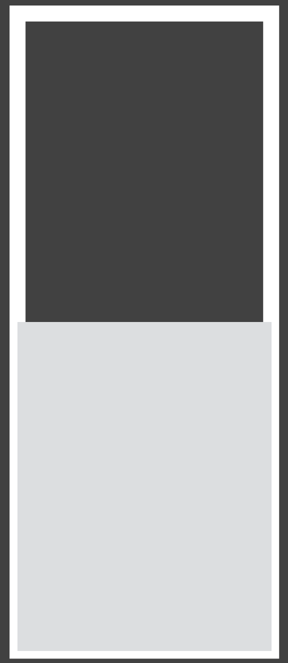
Estimate

Uncertainty

Progressive Visualization with Online Aggregation

Growing sample → continuously improving results

Analysts watch updates until bounds errors are low enough



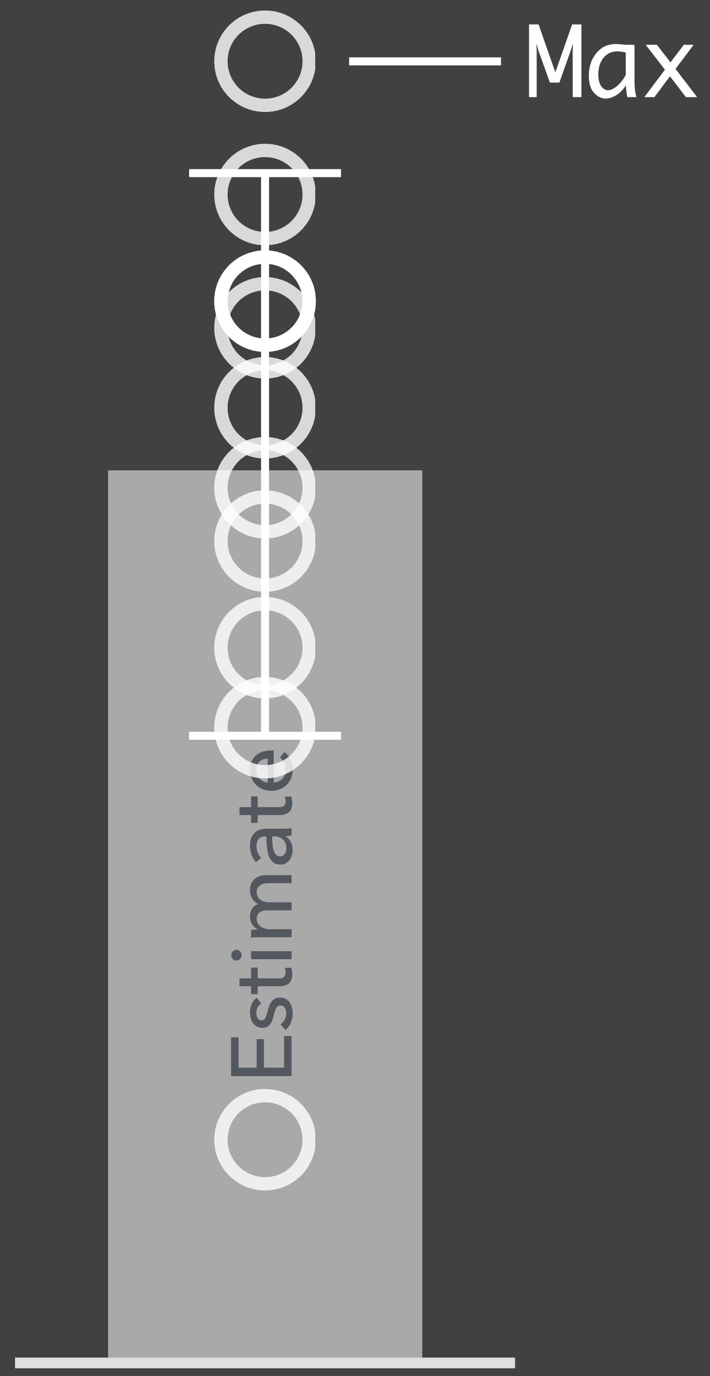
Sum of 50% = 84

Sum of 100 % = 168 ±50

Query finished!

```
$ SELECT bin(distance), count(*)
  FROM flights
  WHERE airline = 'hi'
-> No Results
$ SELECT bin(distance), count(*)
  FROM flights
  WHERE airline = 'ha'
-> Running Query. Please wait ...
```

Challenges with AQP



Approximate results

→ Convey uncertainty

Probabilistic guarantees

Unbounded errors

Arbitrary aggregation or joins

Optimistic Visualization

A UX approach to challenges with AQP traditionally treated as database problems.

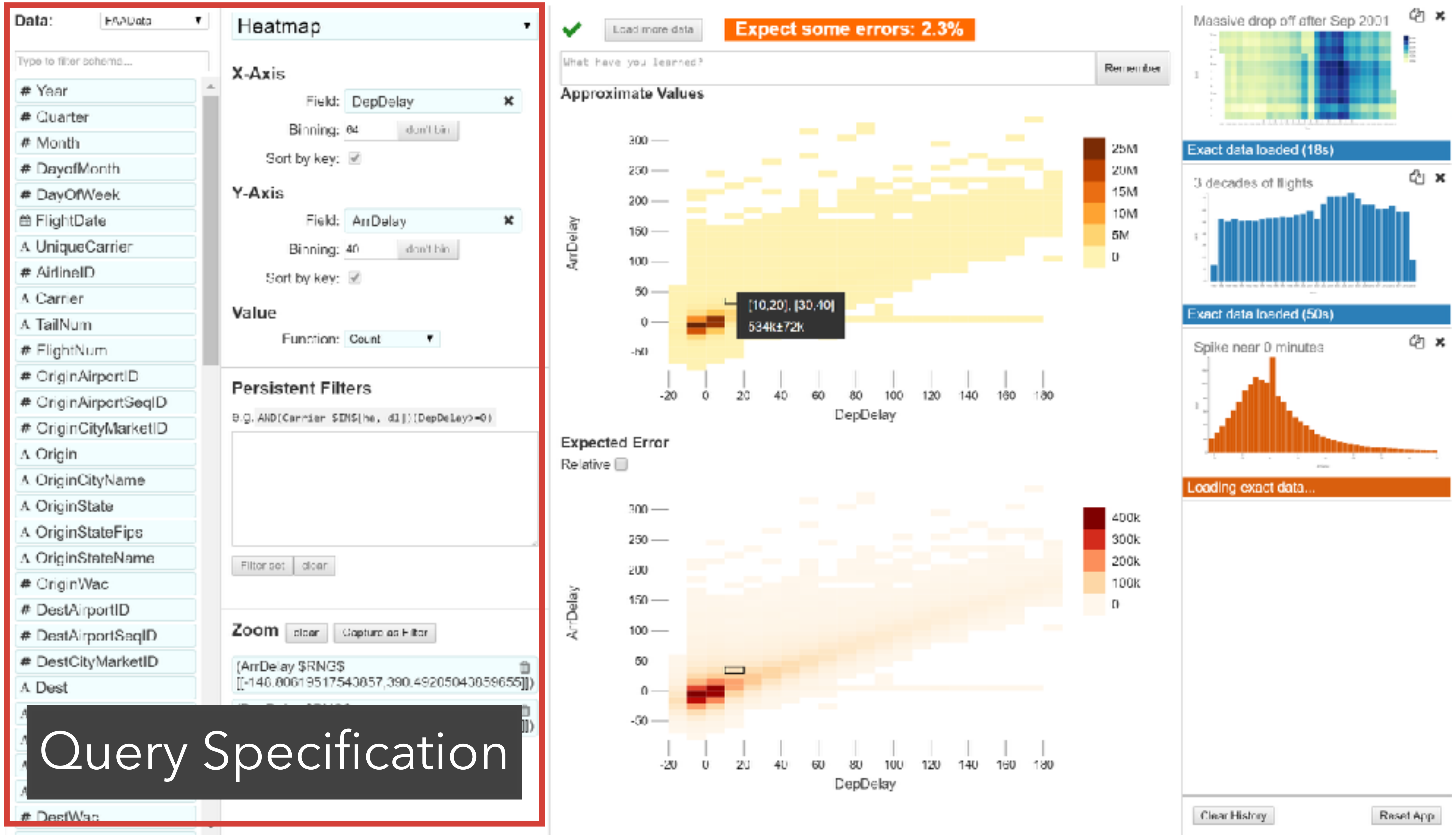
Optimistic Visualization

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

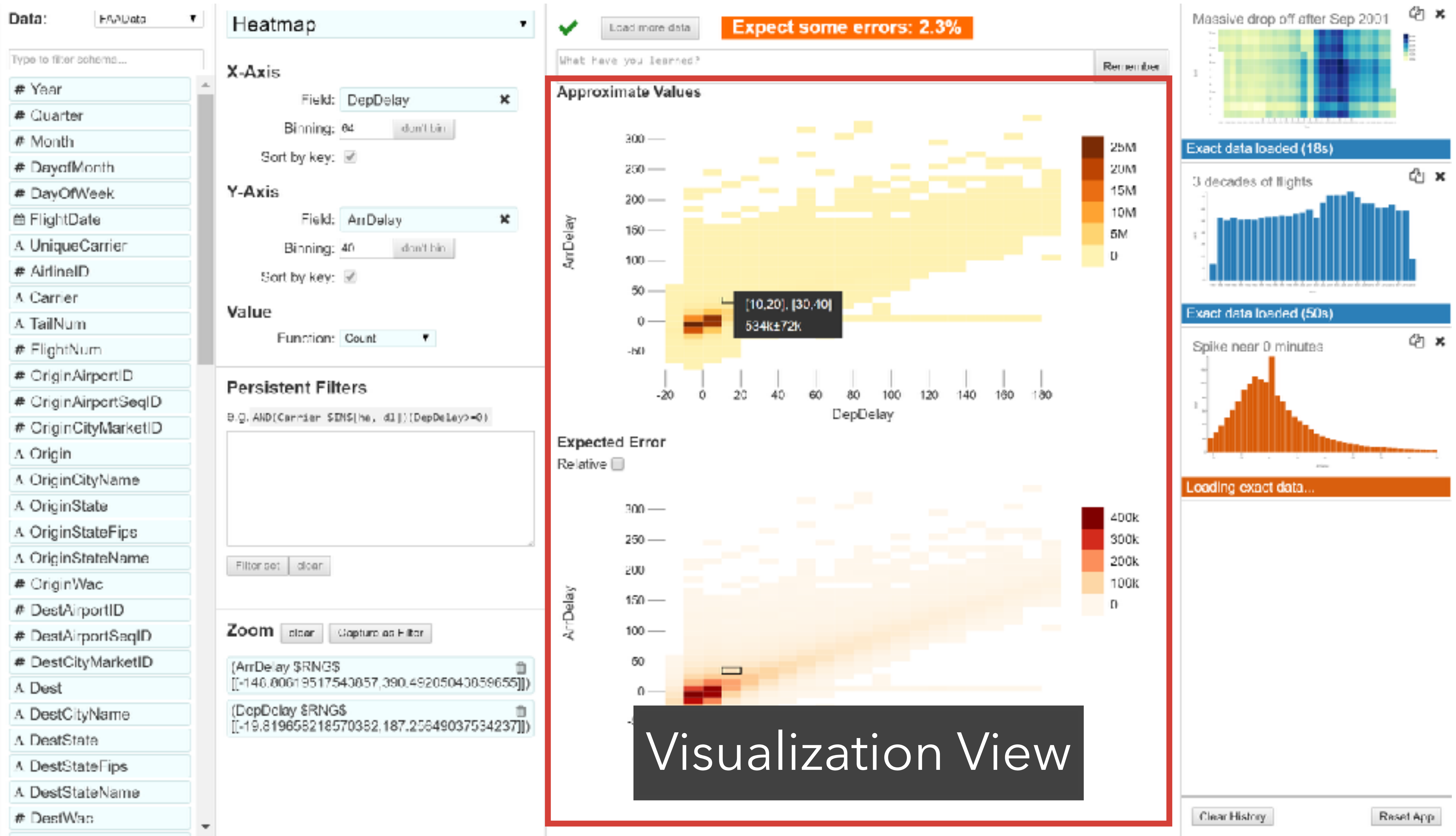
Analysts use initial estimates, run precise query in background, and confirm results later.

Gives users confidence in using AQP.

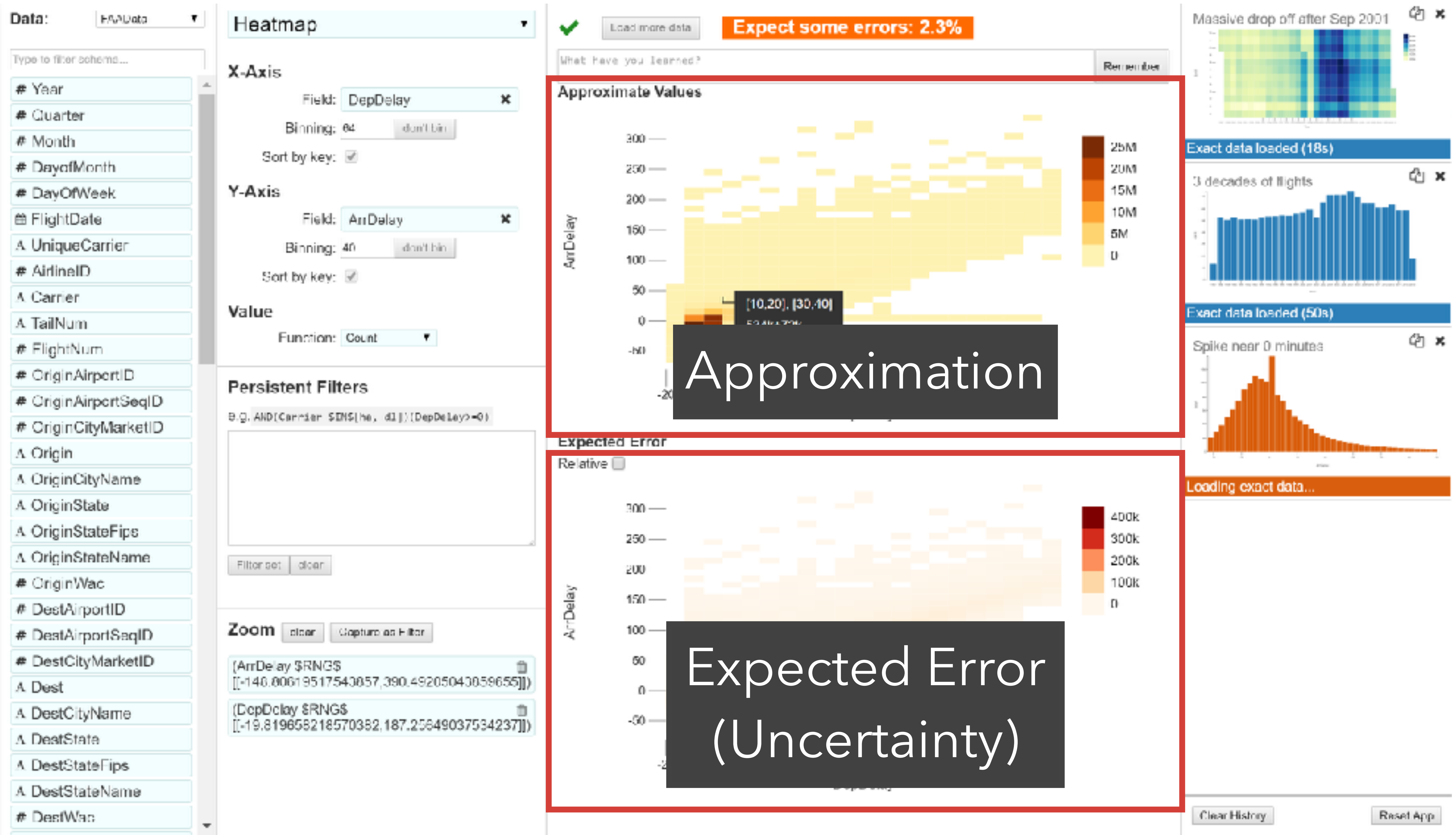
Pangloss implements Optimistic Visualization



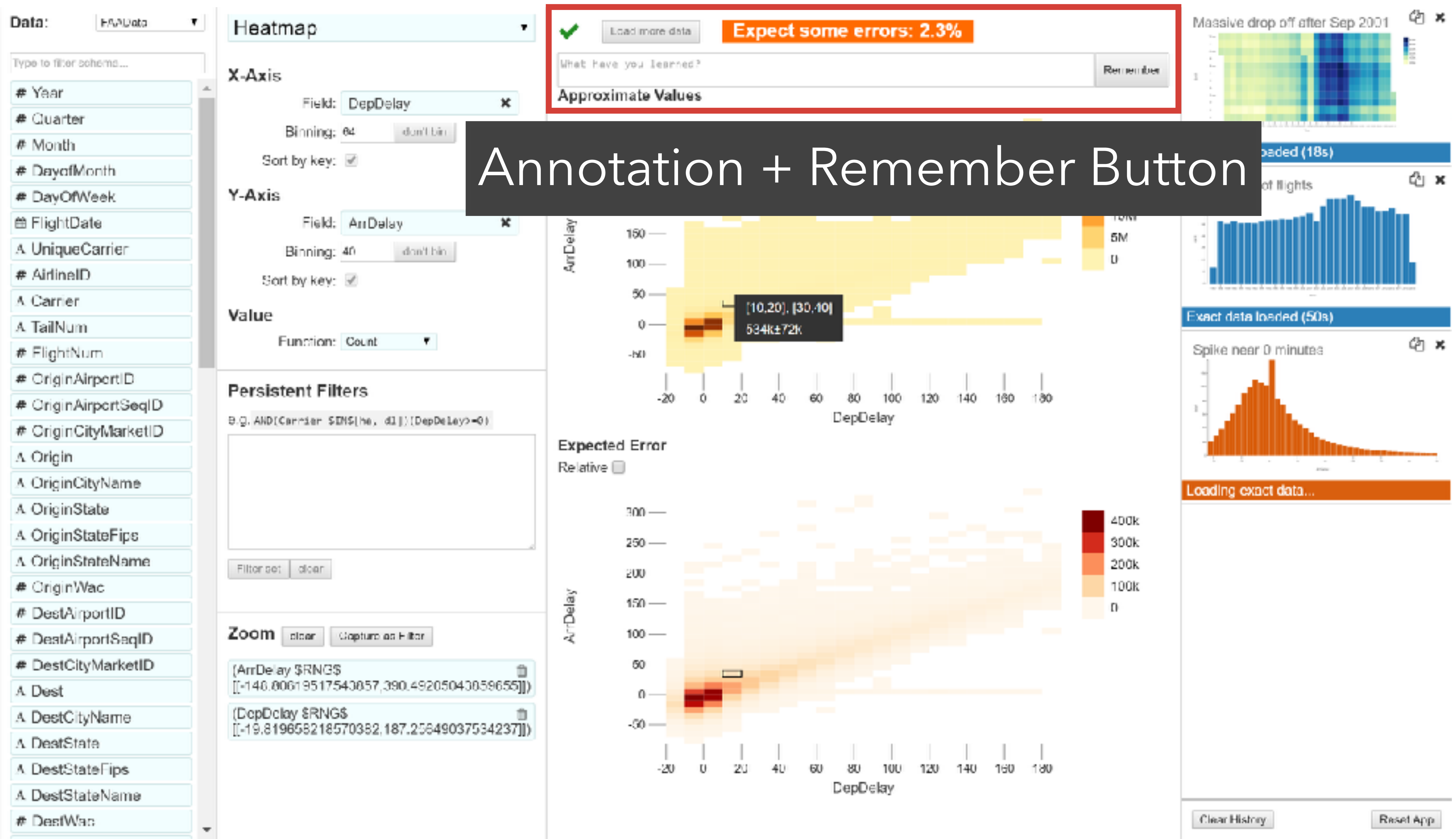
Pangloss implements Optimistic Visualization



Pangloss implements Optimistic Visualization

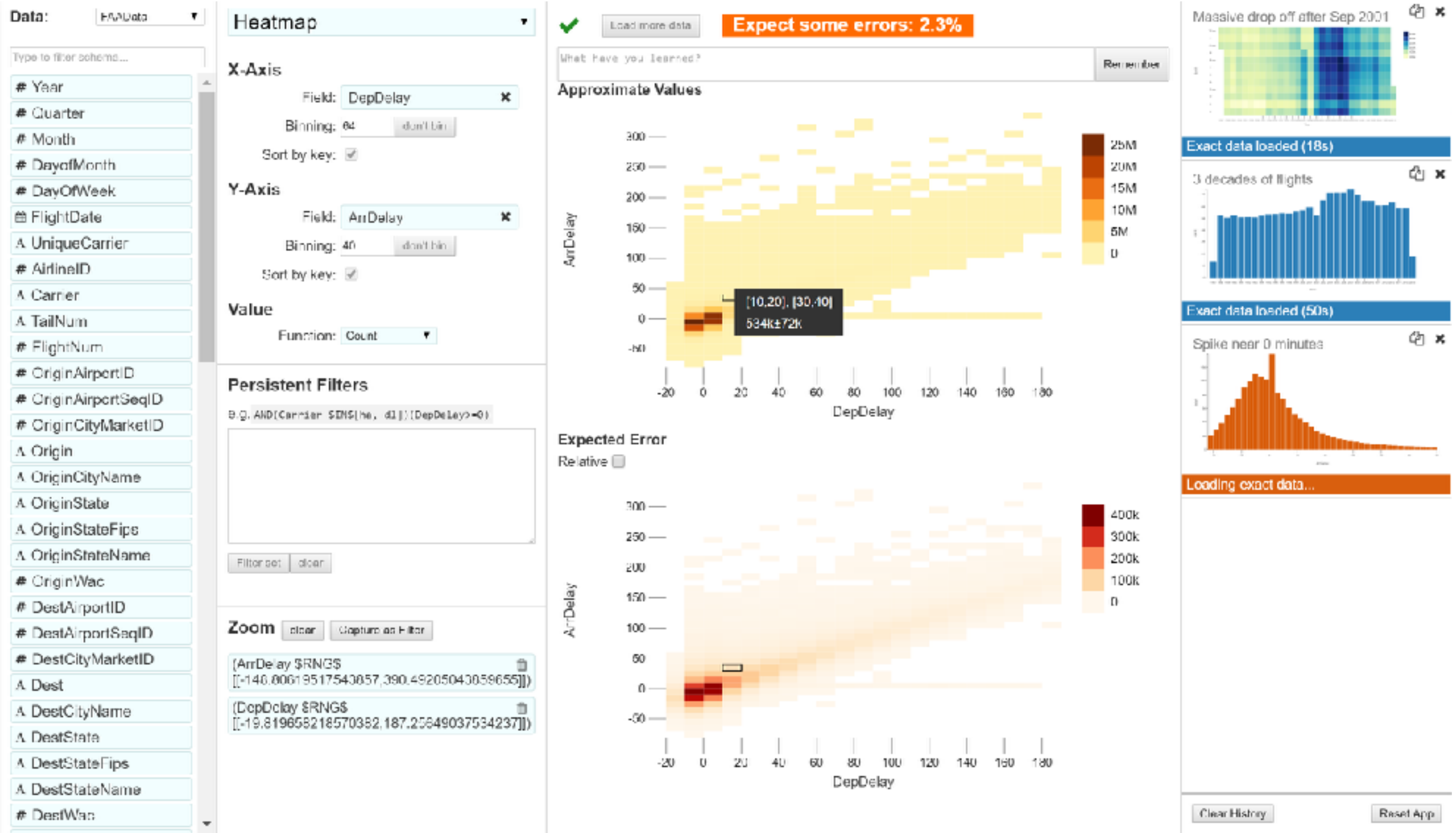


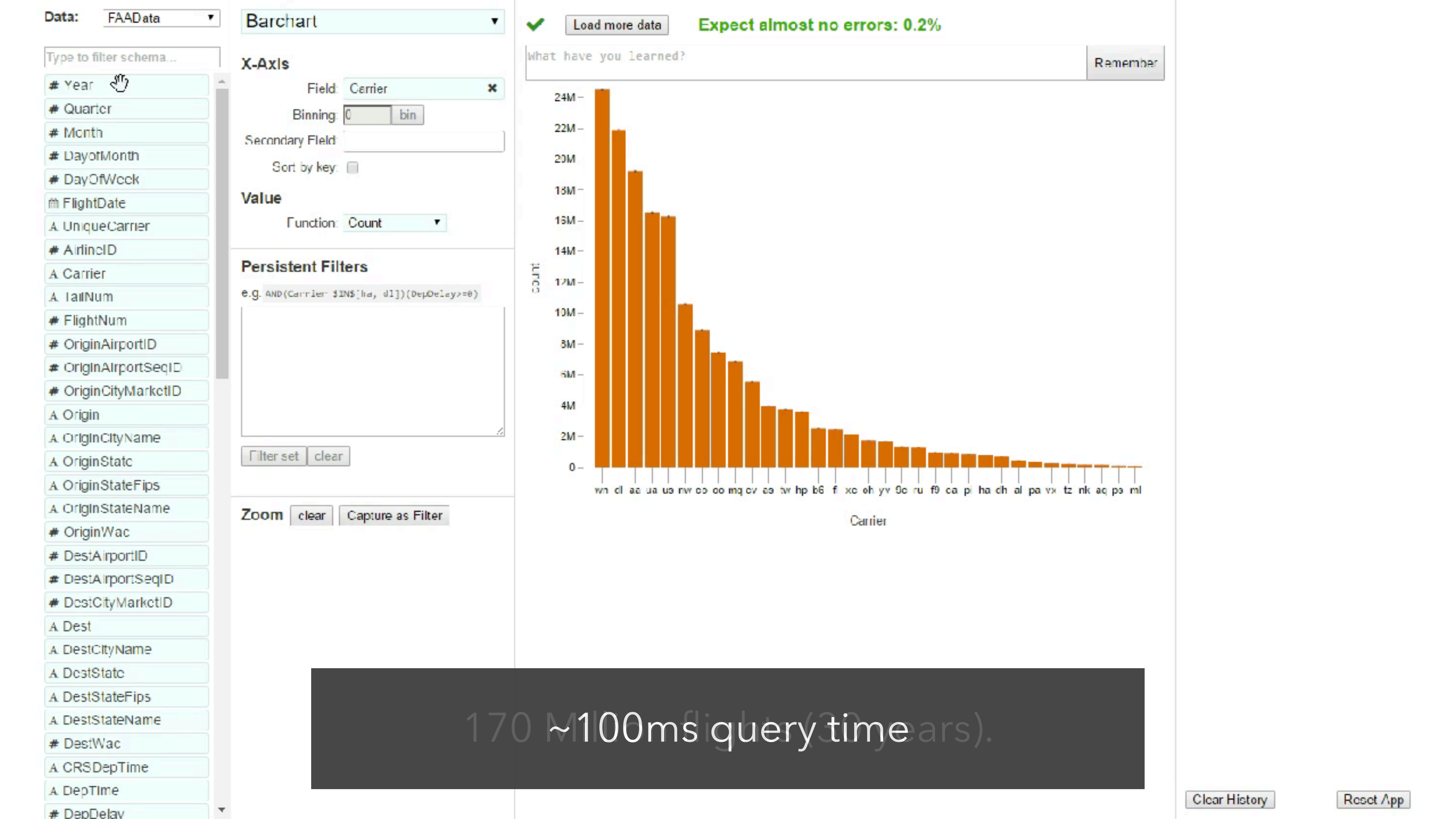
Pangloss implements Optimistic Visualization



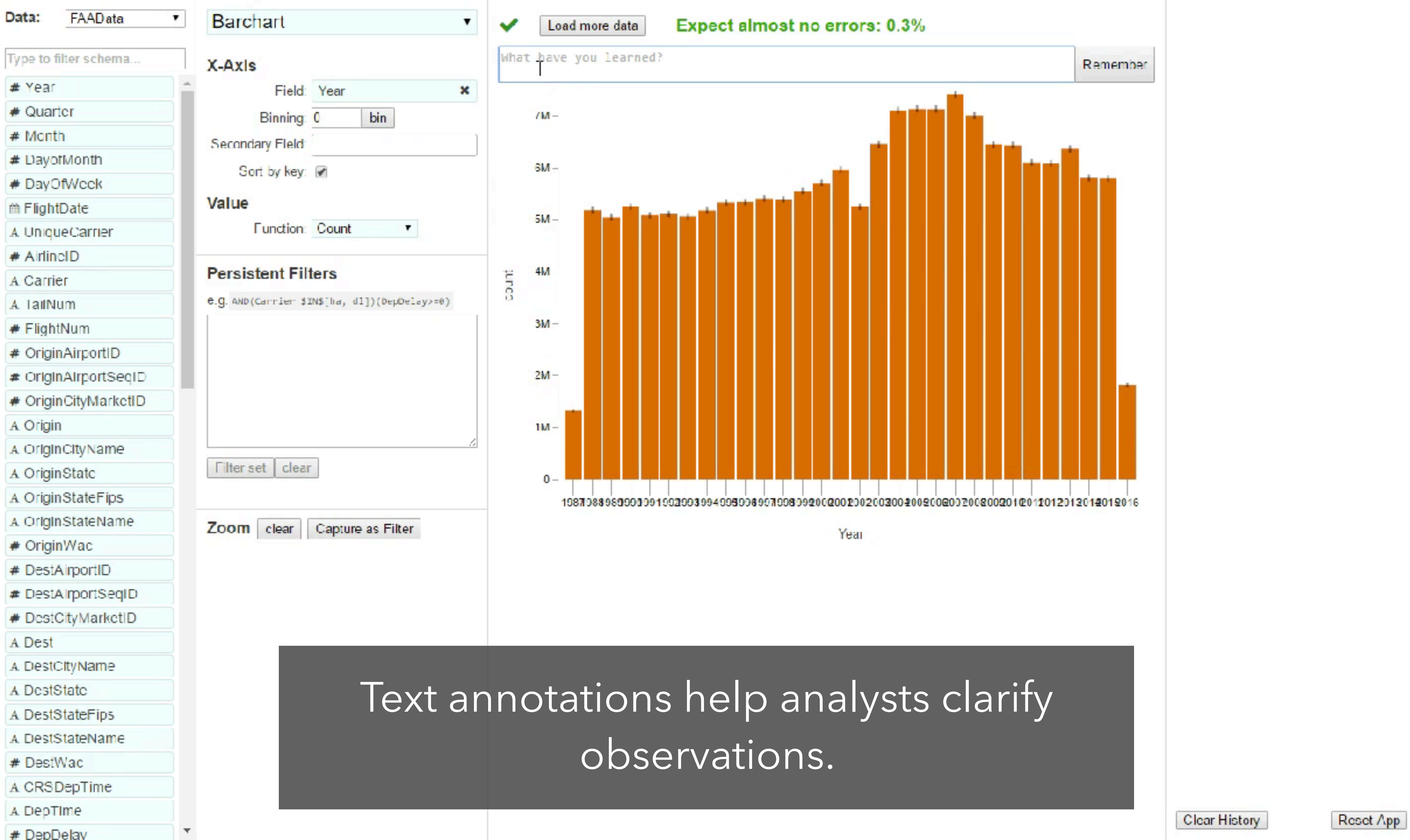
Pangloss implements Optimistic Visualization

Pangloss implements Optimistic Visualization

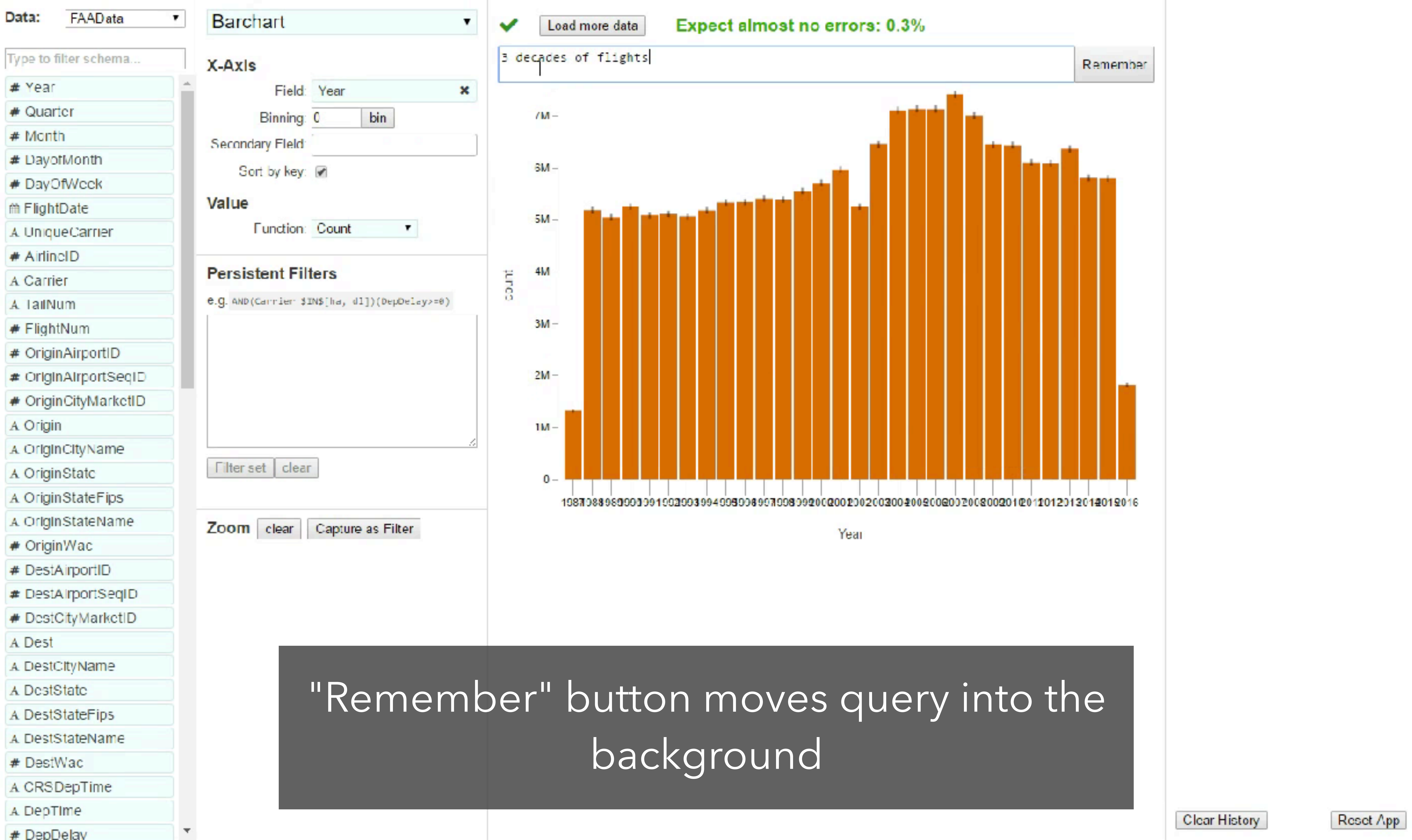




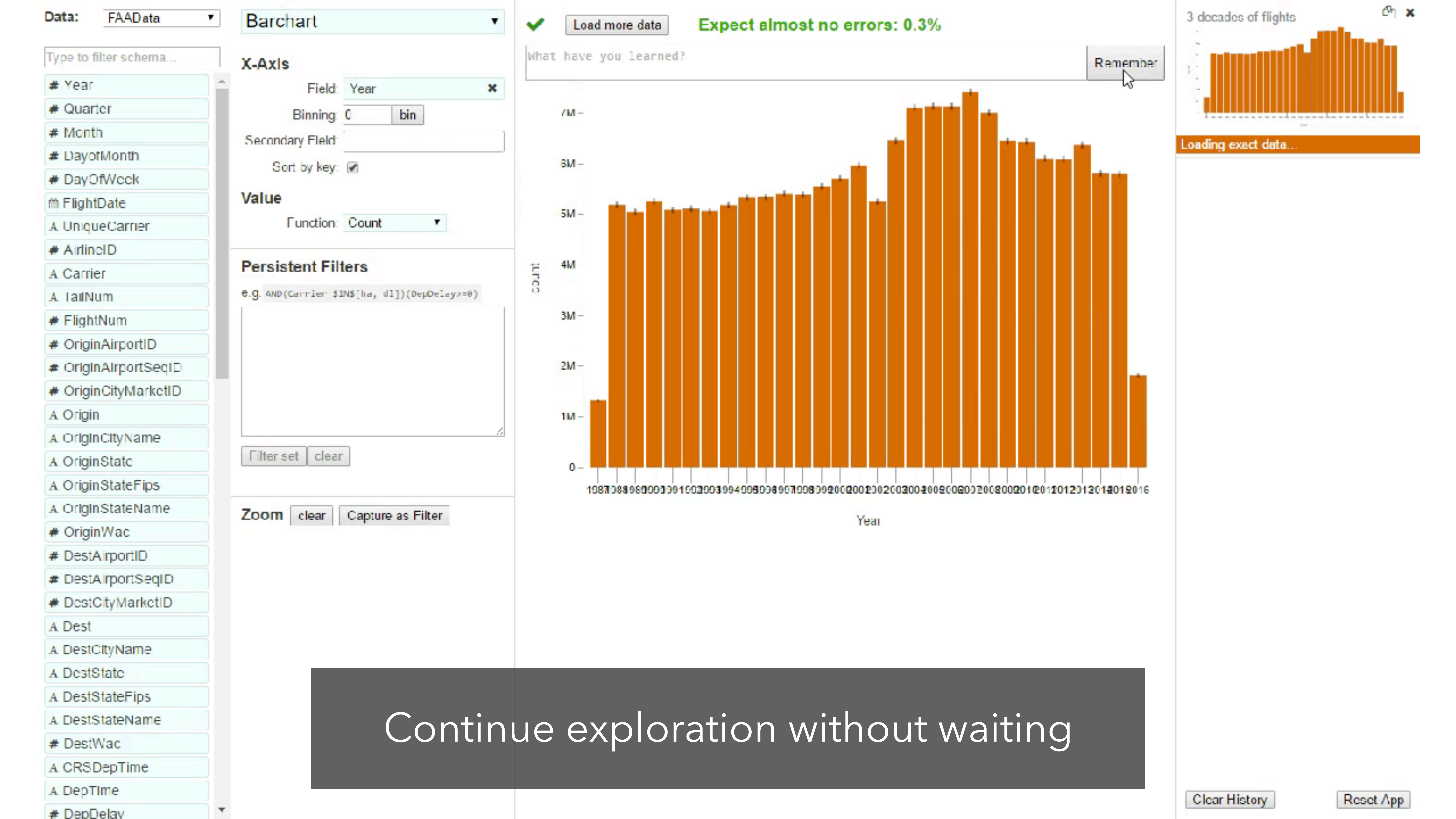
Data: FAAData
Type to filter schema...
Year
Quarter
Month
DayofMonth
DayOfWeek
FlightDate
A UniqueCarrier
AirlineID
A Carrier
A TailNum
FlightNum
OriginAirportID
OriginAirportSeqID
OriginCityMarketID
A Origin
A OriginCityName
A OriginState
A OriginStateFips
A OriginStateName
OriginWac
DestAirportID
DestAirportSeqID
DestCityMarketID
A Dest
A DestCityName
A DestState
A DestStateFips
A DestStateName
DestWac
A CRSDepTime
A DepTime
DepDelay
Barchart
X-Axis
Field: Carrier
Binning: 0 bin
Secondary Field
Sort by key:
Value
Function: Count
Persistent Filters
e.g. AND(Carrier IN[ha, dl])(DepDelay>=0)
Filter set
clear
Zoom
clear
Capture as Filter
Load more data
Expect almost no errors: 0.2%
What have you learned?
Remember
count
Carrier
Clear History
Reset App
~100ms query time

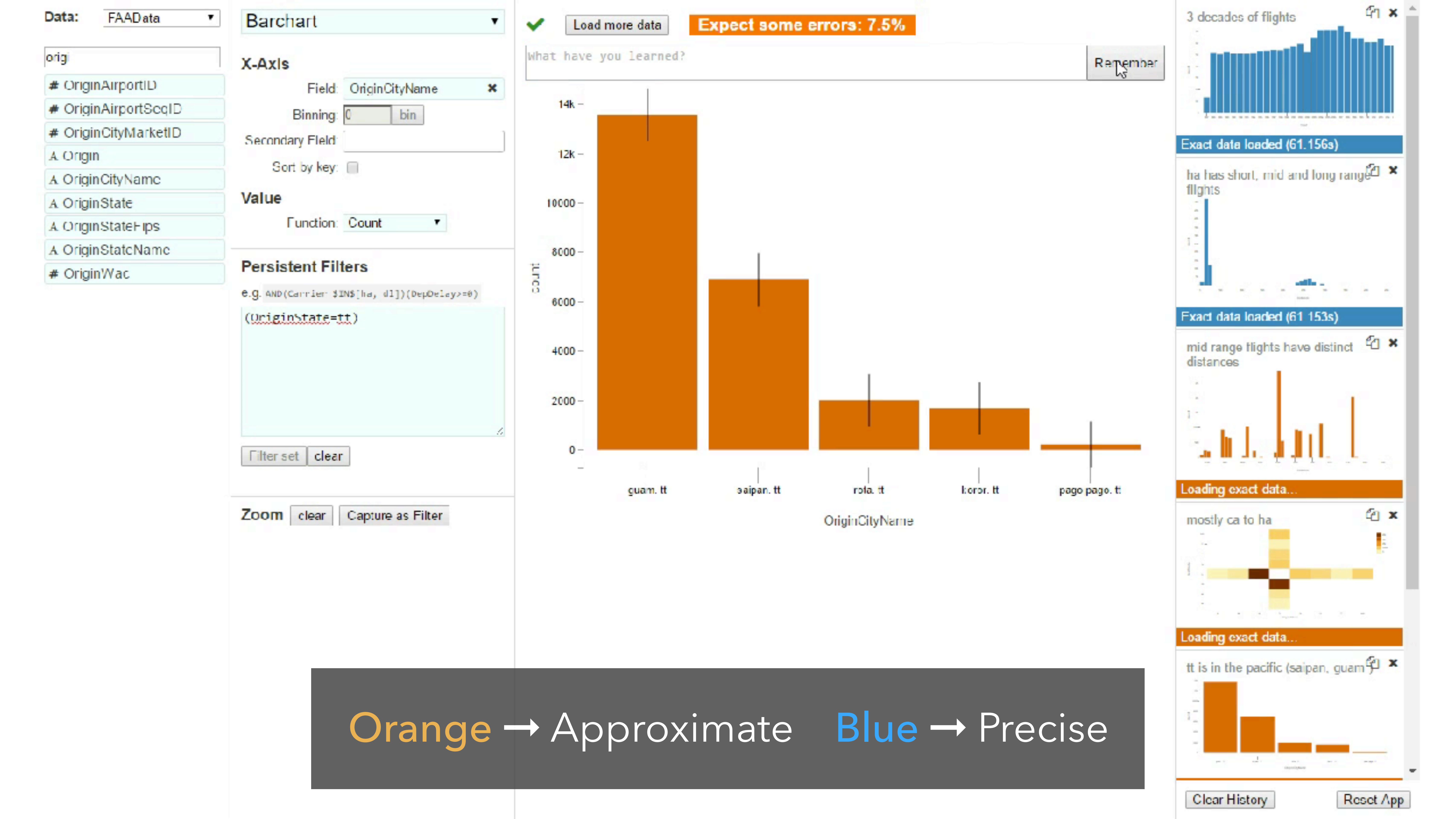


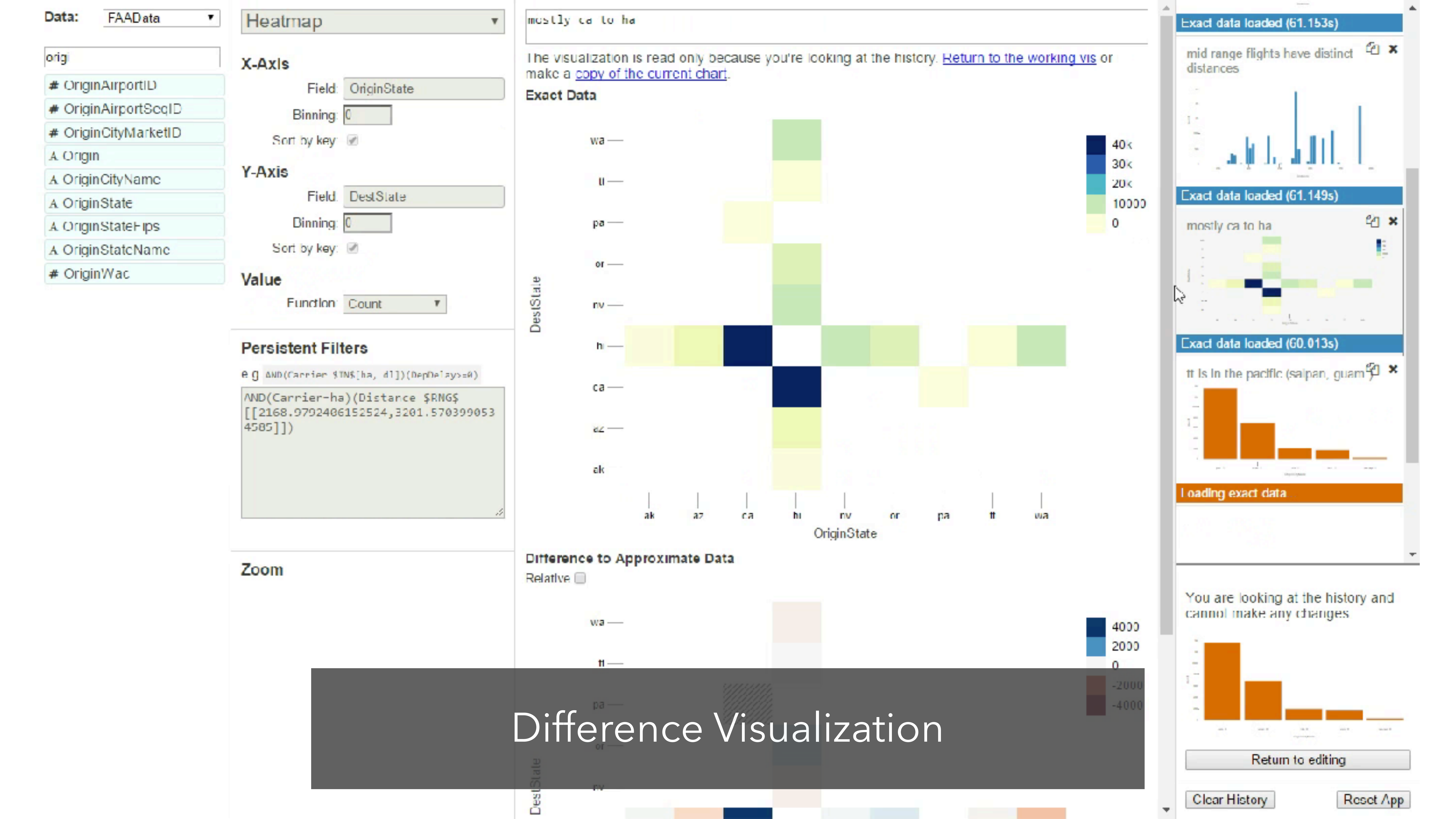
Text annotations help analysts clarify observations.



"Remember" button moves query into the background


Data: FAAData
Type to filter schema...
Year
Quarter
Month
DayofMonth
DayOfWeek
FlightDate
UniqueCarrier
AirlineID
Carrier
TailNum
FlightNum
OriginAirportID
OriginAirportSeqID
OriginCityMarketID
Origin
OriginCityName
OriginState
OriginStateFips
OriginStateName
OriginWac
DestAirportID
DestAirportSeqID
DestCityMarketID
Dest
DestCityName
DestState
DestStateFips
DestStateName
DestWac
CRSDepTime
DepTime
DepDelay
Barchart
X-Axis
Field: Year
Binning: 0 bin
Secondary Field:
Sort by key:
Value
Function: Count
Persistent Filters
e.g. AND(Carrier IN[ha, dl])(DepDelay>=0)
Filter set
clear
Zoom
clear
Capture as Filter
Load more data
Expect almost no errors: 0.3%
What have you learned?
Remember
count
Year
3 decades of flights
Loading exact data...
Clear History
Reset App
Continue exploration without waiting


Data: FAAData
orig
OriginAirportID
OriginAirportSeqID
OriginCityMarketID
A Origin
A OriginCityName
A OriginState
A OriginStateFips
A OriginStateName
OriginWac
Barchart
X-Axis
Field: OriginCityName
Binning: bin
Secondary Field:
Sort by key:
Value
Function: Count
Persistent Filters
e.g. AND(Carrier IN[ha, dl])(DepDelay>=0)
(OriginState=tt)
Filter set
clear
Zoom
clear
Capture as Filter
Load more data
Expect some errors: 7.5%
What have you learned?
Remember
14k
12k
10000
8000
6000
4000
2000
0
count
guam. tt
saipan. tt
rota. tt
koror. tt
pago pago. tt
OriginCityName
3 decades of flights
Exact data loaded (61.156s)
ha has short, mid and long range flights
Exact data loaded (61.153s)
mid range flights have distinct distances
Loading exact data...
mostly ca to ha
Loading exact data...
tt is in the pacific (saipan, guam
Clear History
Reset App
Orange → Approximate
Blue → Precise



Data: FAAData
orig
OriginAirportID
OriginAirportSeqID
OriginCityMarketID
A Origin
A OriginCityName
A OriginState
A OriginStateFips
A OriginStateName
OriginWac
Heatmap
X-Axis
Field: OriginState
Binning: 0
Sort by key
Y-Axis
Field: DestState
Binning: 0
Sort by key
Value
Function: Count
Persistent Filters
e.g. AND(Carrier IN[ha, dl])(DepDelay>=0)
AND(Carrier=ha)(Distance RNG [[2168.9792406152524,3201.5703990534505]])
Zoom
mostly ca to ha
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.
Exact Data
DestState
OriginState
Difference to Approximate Data
Relative
Exact data loaded (61.153s)
mid range flights have distinct distances
Exact data loaded (61.149s)
mostly ca to ha
Exact data loaded (60.013s)
tt is in the pacific (saipan, guam)
Loading exact data
You are looking at the history and cannot make any changes
Return to editing
Clear History
Reset App
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.”

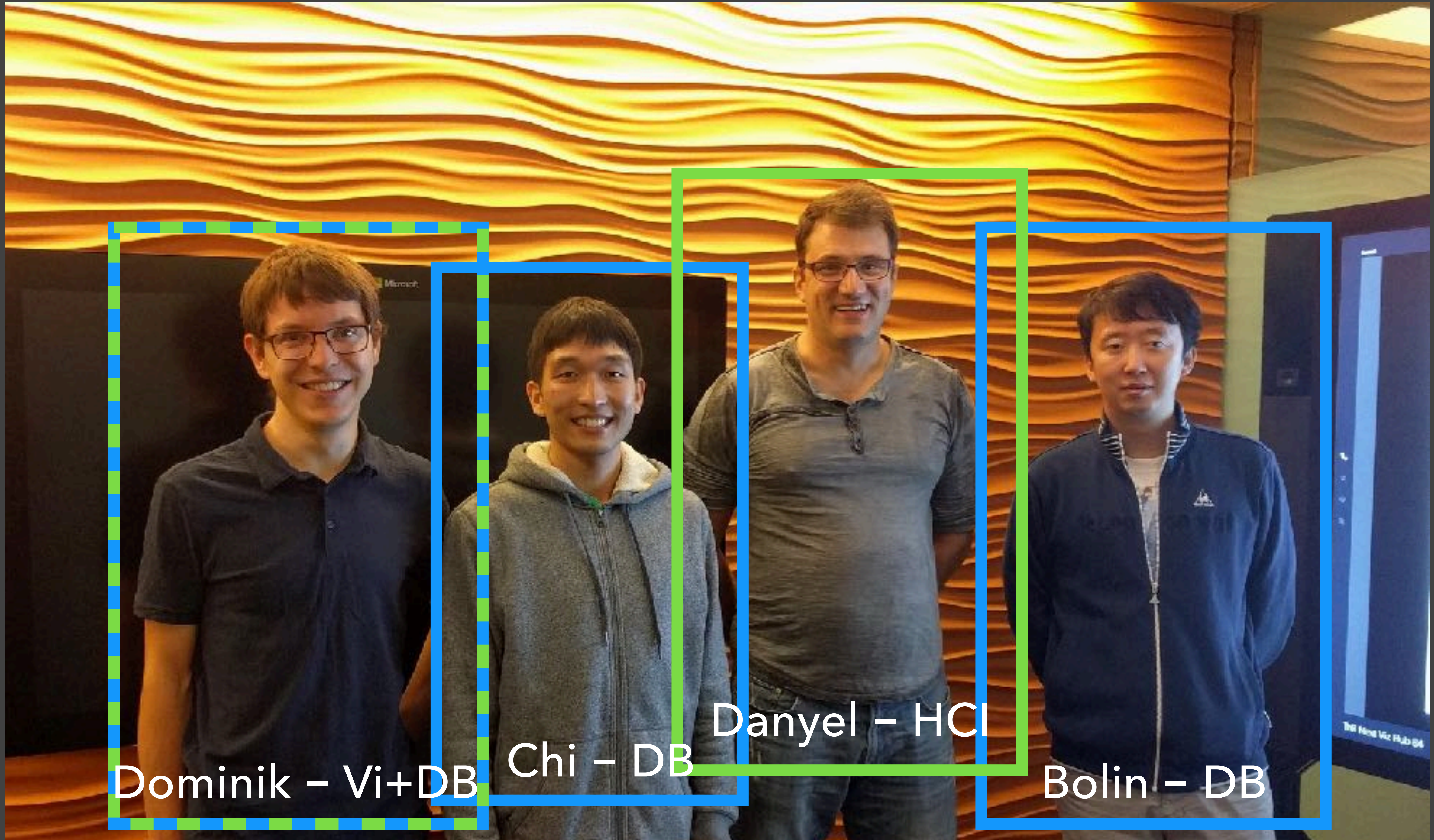
Conclusions

Fundamental problems with AQP addressed as **UX problem**

Gives analysts confidence in AQP

Future: Alerting, Remembering, Progressive + Optimistic

AQP needs Multi-Disciplinary Solutions



Implications for the Database Community

HILDA at SIGMOD 2017

What Users Don't Expect about Exploratory Data Analysis on Approximate Query Processing Systems

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

Danyel Fisher
Microsoft Research
danyelf@microsoft.com

ABSTRACT

Pangloss implements "Optimistic Visualization", a method that gives analysts confidence to use approximate results for exploratory data analysis. In this paper, we outline how analysts' experience with an approximate visualization system did not match their intuitions. These observations have implications for the design of future data exploration systems that expose uncertainty. We also describe requirements for approximate query engines to enable the next generation of exploratory visualization systems.

CCS CONCEPTS

data system: it allows users to explore their data through fast, approximate queries; users can then request precise responses with slow queries over the full data. In the first phase, the engine that drives Pangloss, called "Sample+Seek" [3], returns approximate results in interactive time with an overall uncertainty level.

This is a new experience for users. Pangloss requires users to work with a new uncertainty model, with two-round queries, and to directly face the implications of uncertainty. We see all of these as important and valuable changes in a world that increasingly embraces AQP; however, they can be surprising for users. These user stories will allow us to begin to design for interacting with

Trust But Verify: Optimistic Visualizations for AQP

Fundamental problems with AQP addressed as UX problem

Optimistic Visualization gives analysts confidence in AQP

Integrates well into existing Visual Analysis tools

Future: Alerting, Remembering, Progressive

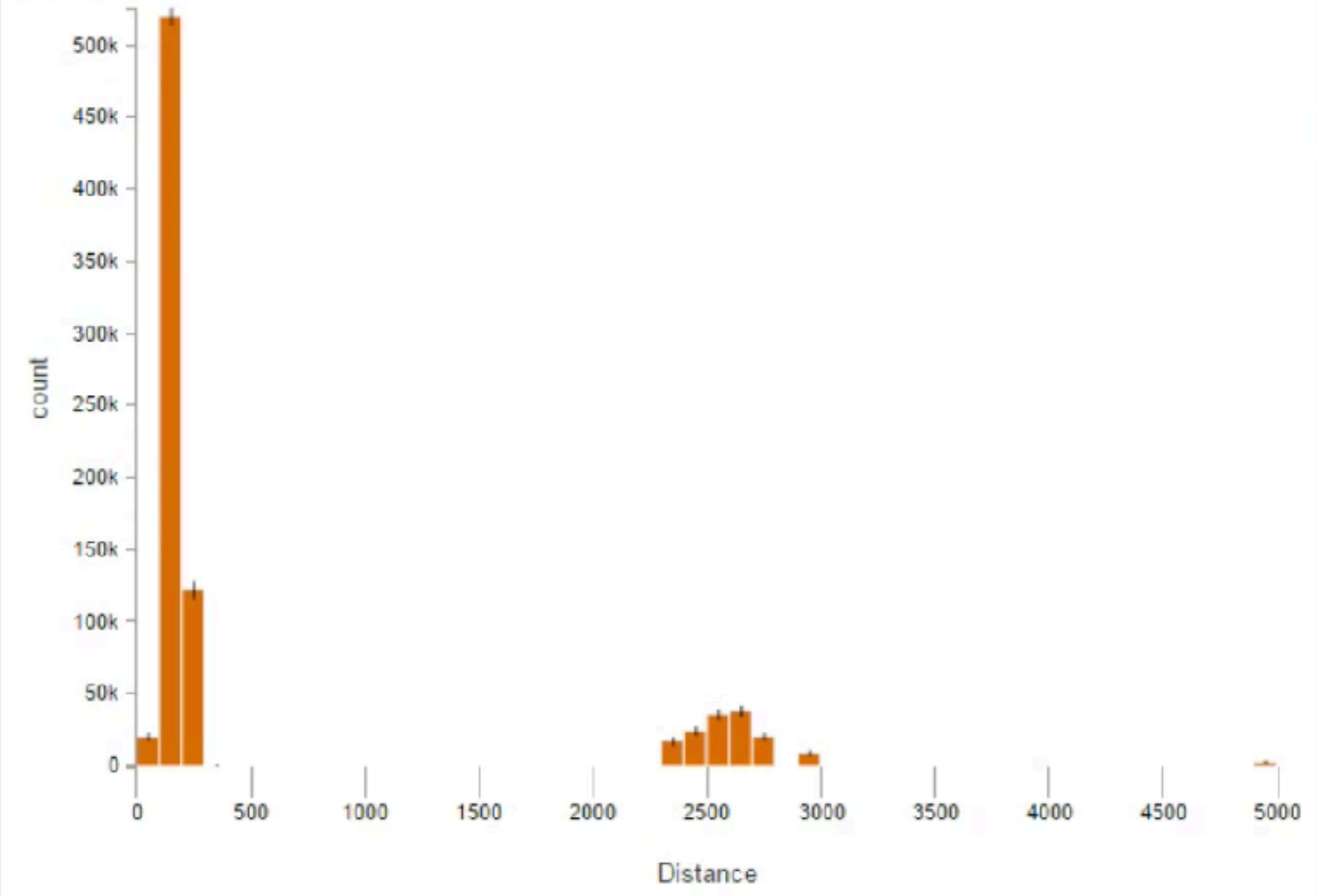
Details: bit.ly/2pwQQg7

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

Query finished!

Backup Slides

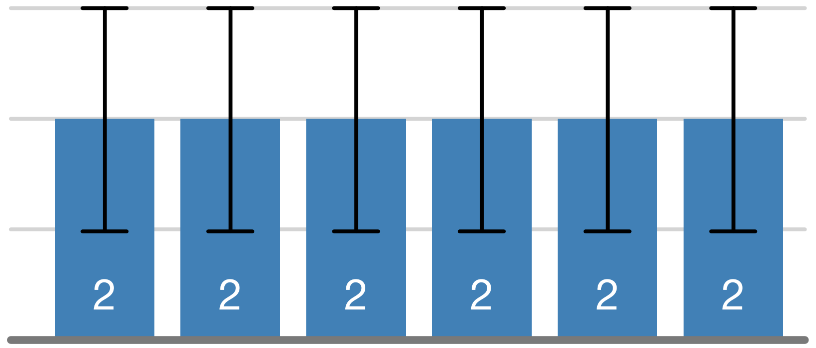
Histogram of Distances for Hawaiian Airlines



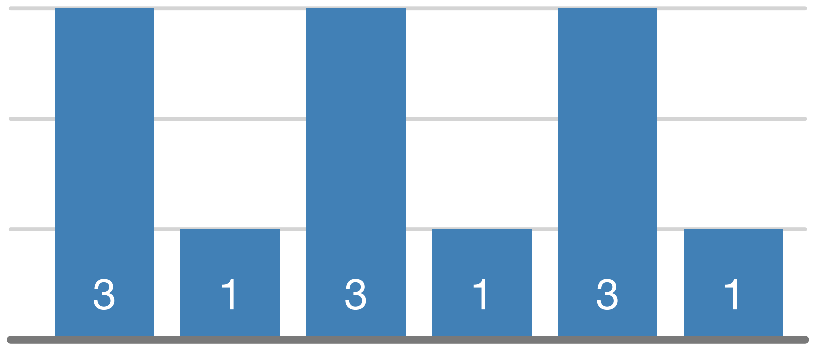
Distribution Uncertainty

Approximation

Distribution Uncertainty: 4

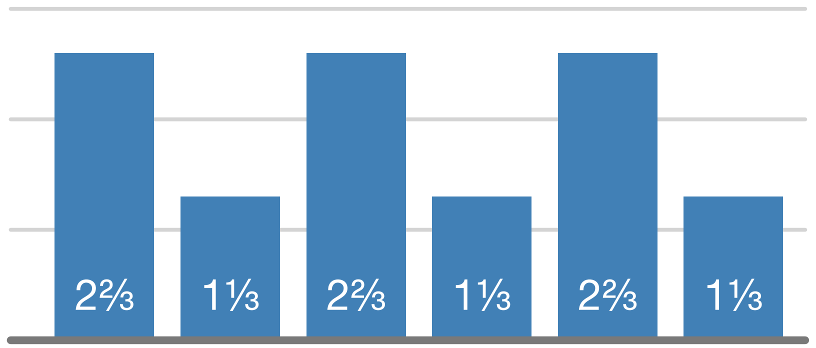


Outside Distribution Uncertainty

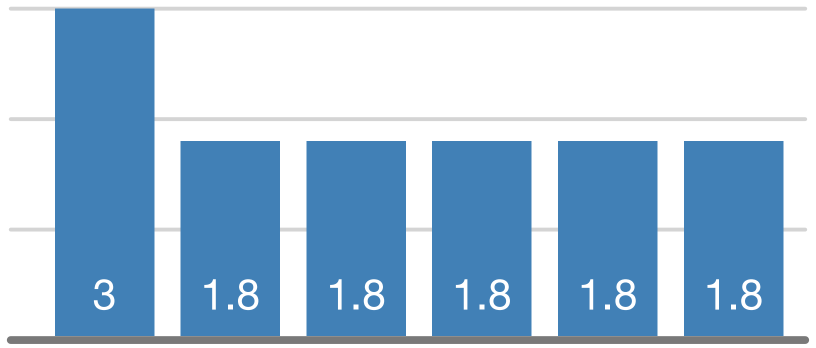


Error: **6**
Sum: 12

Within Distribution Uncertainty

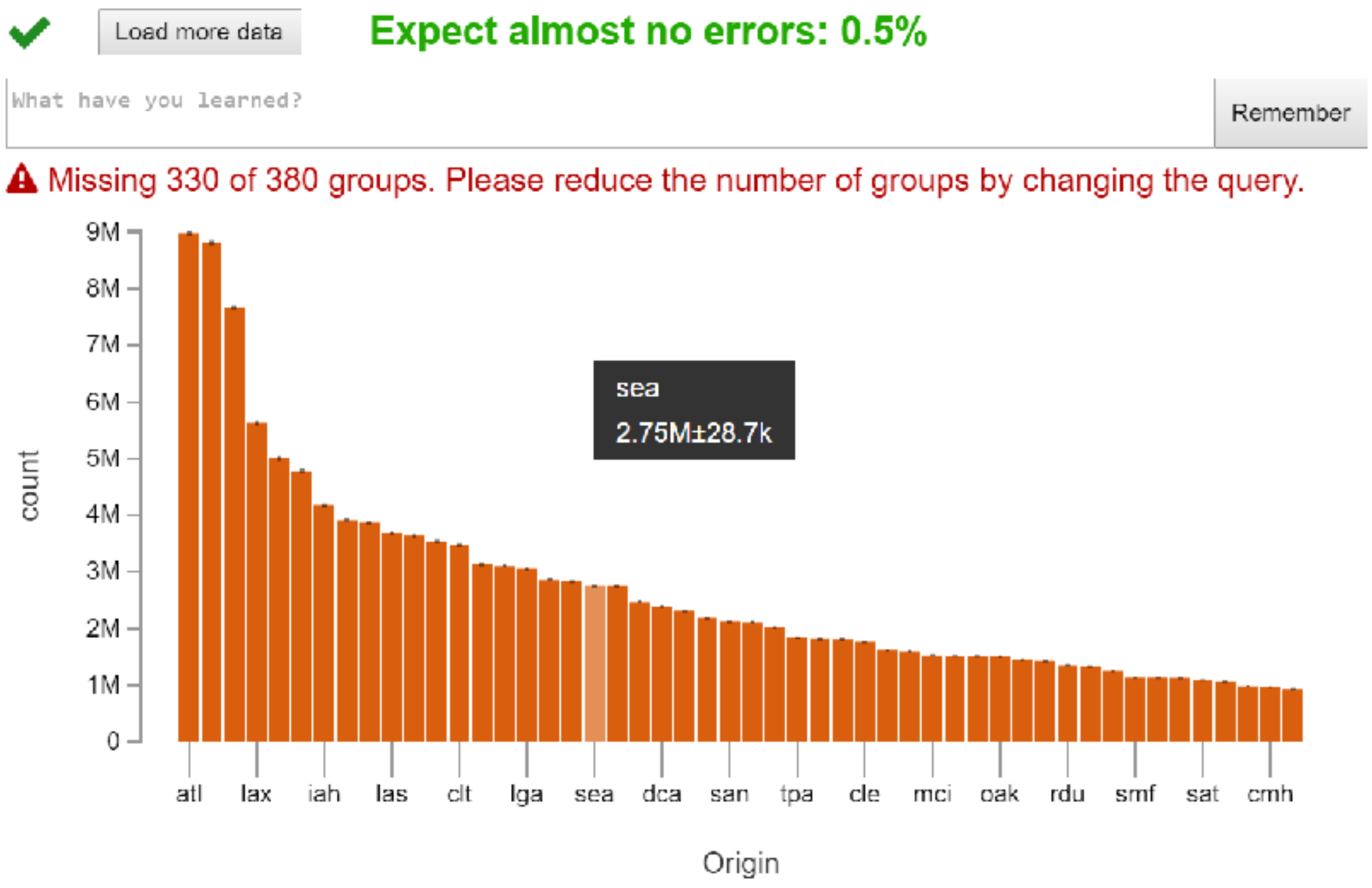


Error: 4
Sum: 12



Error: 4
Sum: 12

Distribution Uncertainty



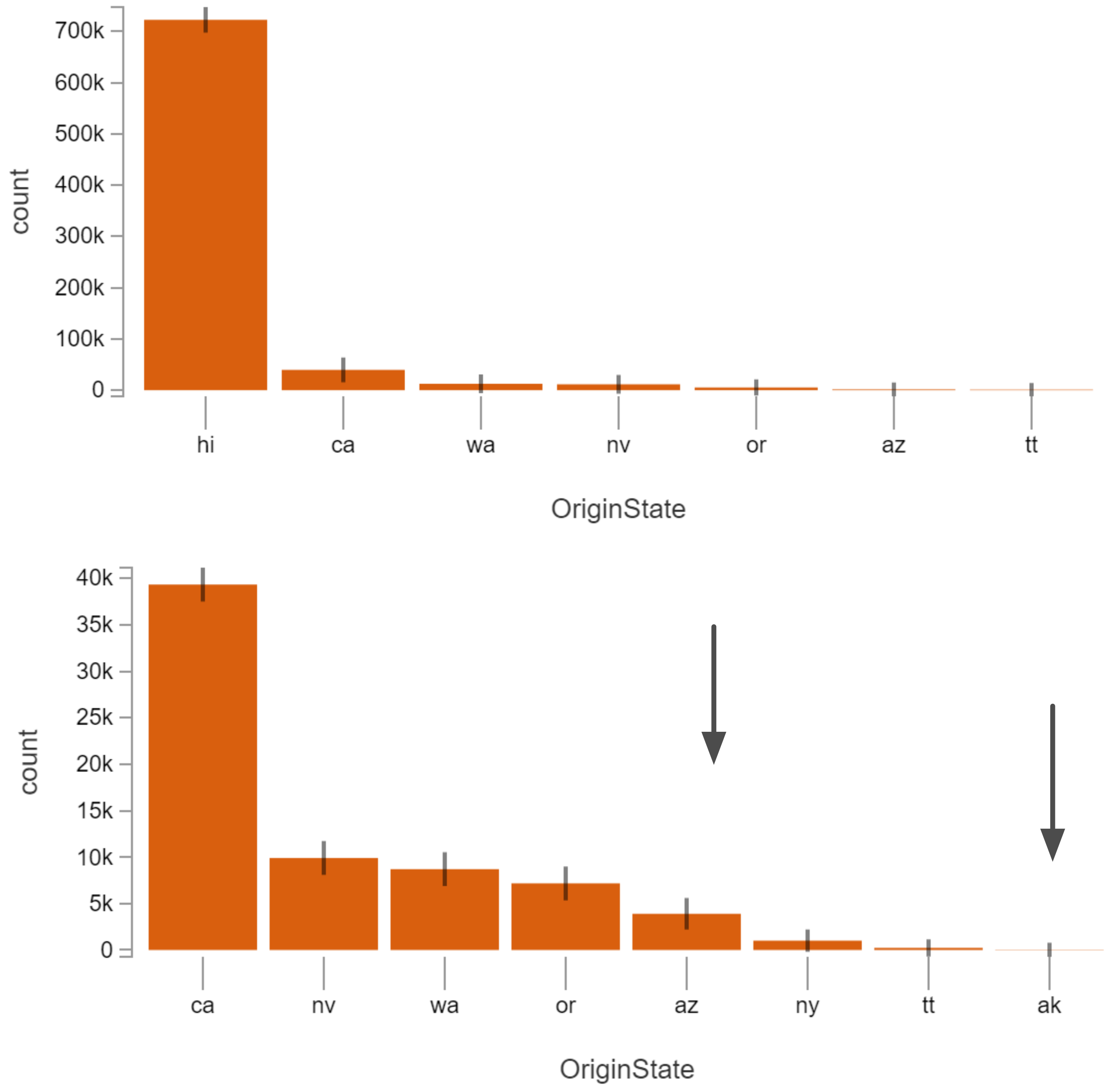
Filtering can show new groups

new predicate →

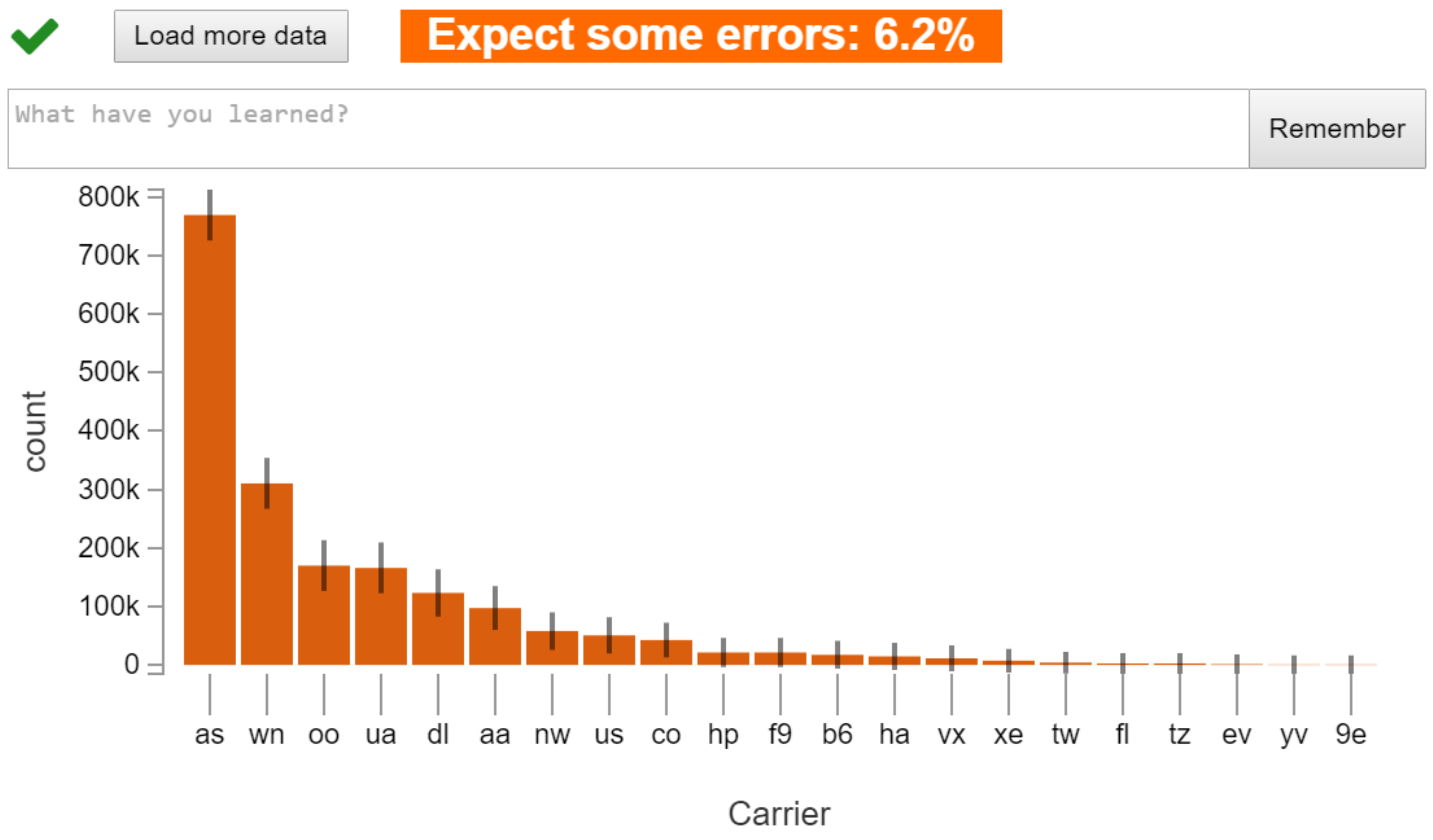
new query →

different sample →

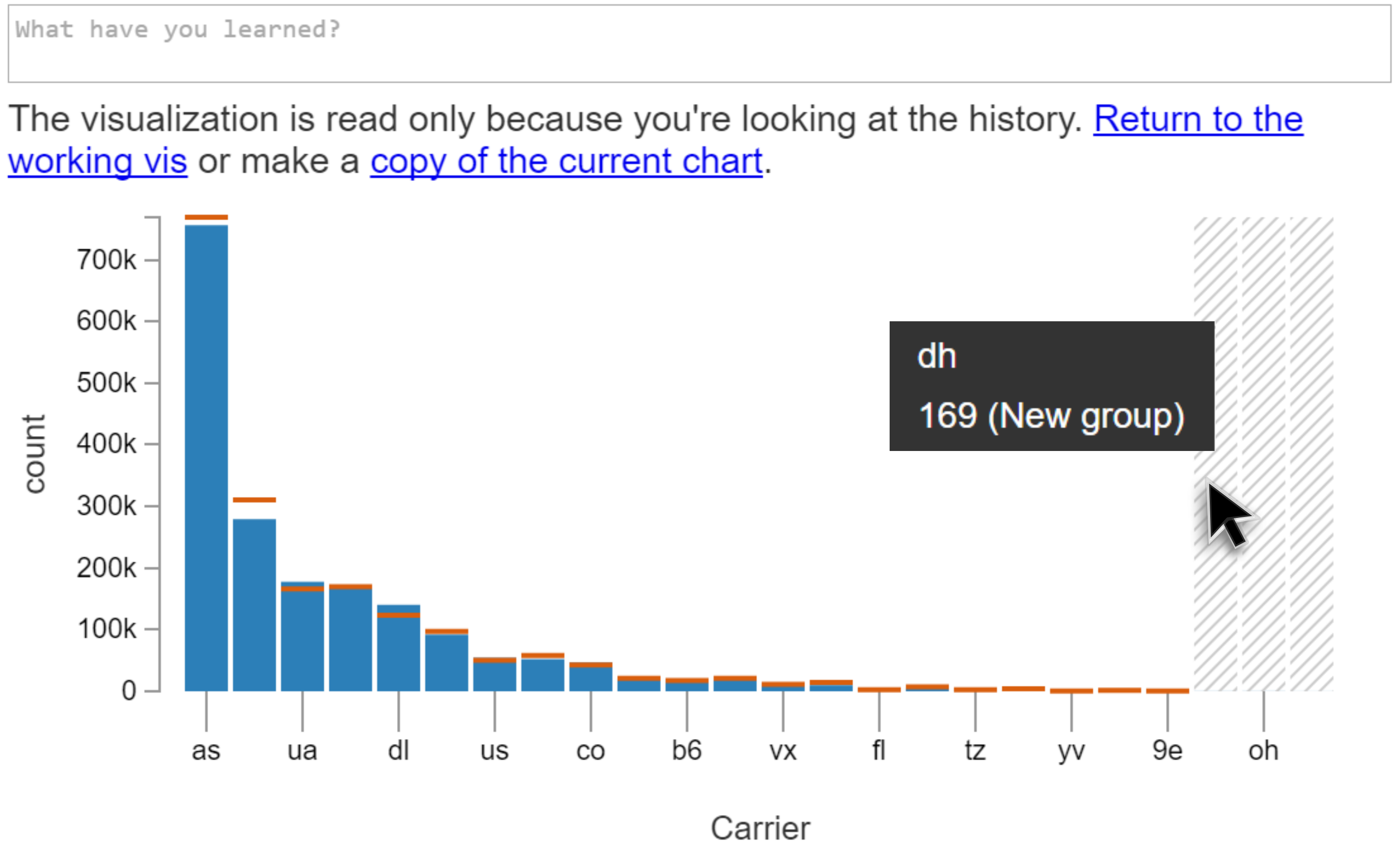
different groups



Precise results can show new groups

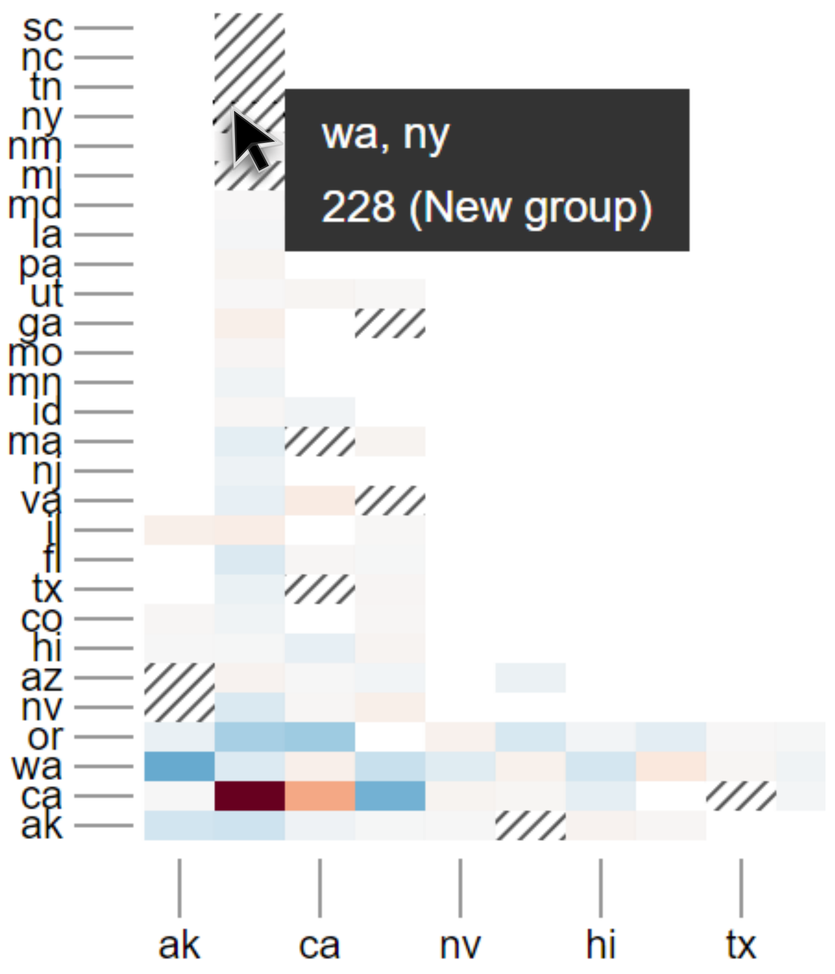


Approximate

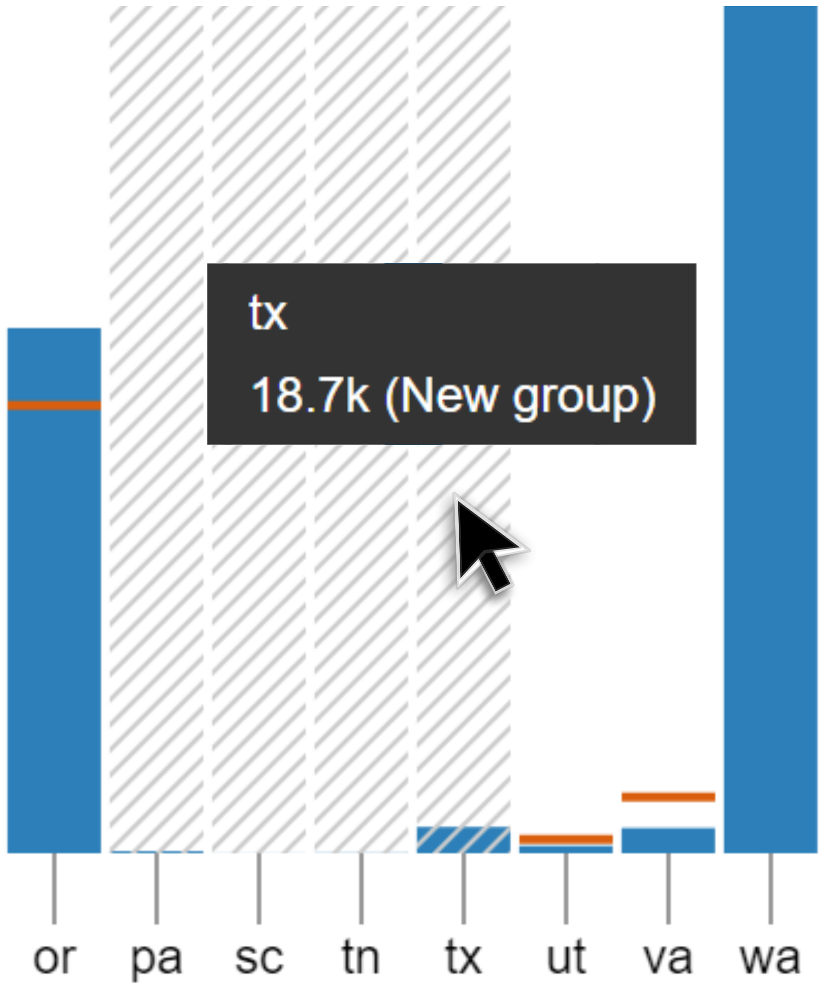


Precise

Vocabulary of visual cues



Heatmap



Barchart