

TRUST BUT VERIFY

Optimistic Visualizations of Approximate Queries for Exploring Big Data

Dominik Moritz @domoritz



Paul G. Allen School of CSE
University of Washington

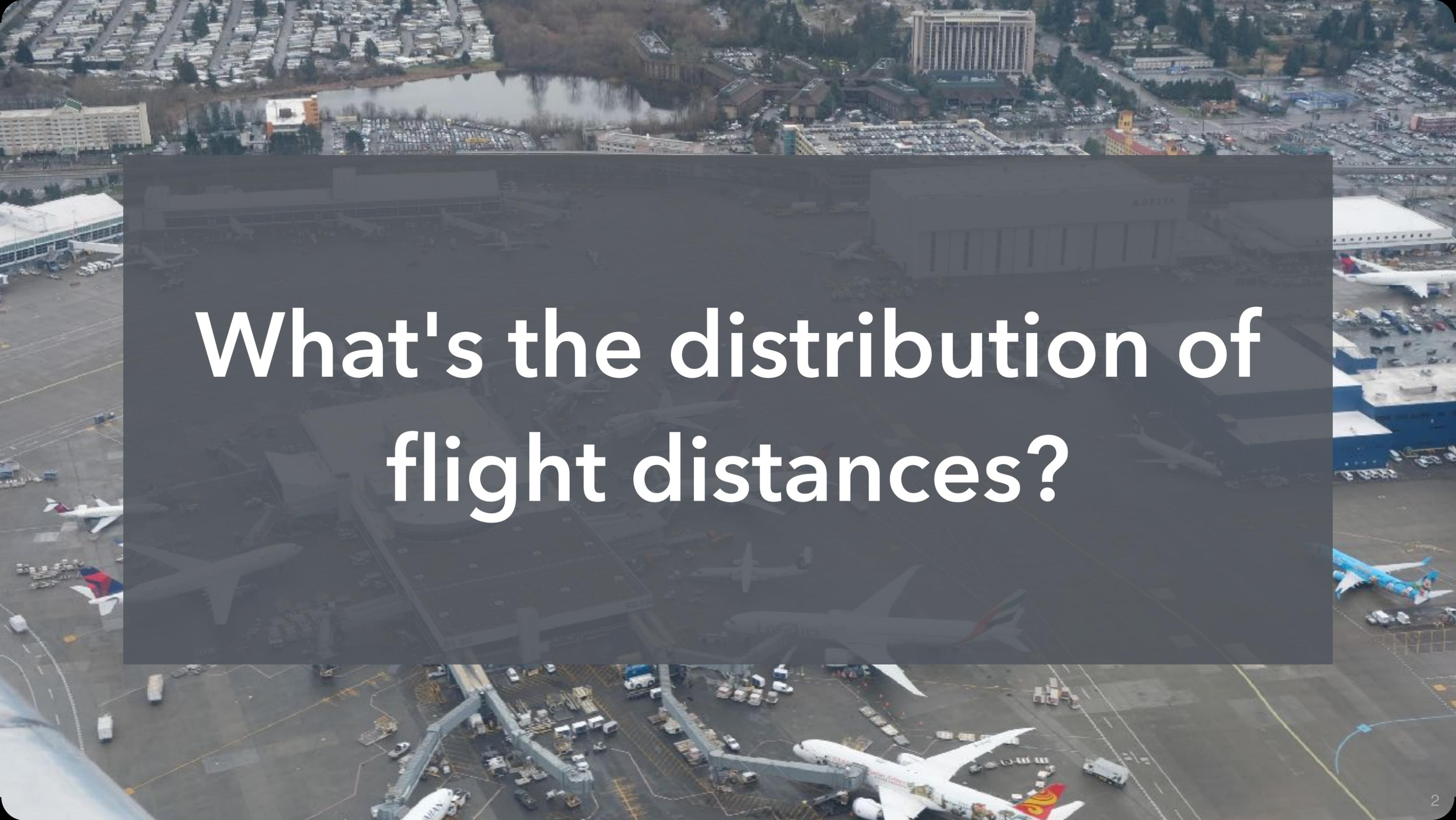
Danyel Fisher @FisherDanyel

Bolin Ding @AtlasDing

Chi Wang



HCI and DMX
Microsoft Research

An aerial photograph of an airport tarmac. In the foreground, a large white aircraft with a red and yellow tail is being serviced by ground crew and equipment. To the left, another aircraft with a red and blue tail is visible. The background shows airport buildings, including one with 'DELTA' written on it, and a parking lot. A dark grey semi-transparent box is overlaid on the center of the image, containing white text.

**What's the distribution of
flight distances?**

```
$ wget https://www.transtats.bts.gov/download.zip
```

```
=====> 70GB
```

```
-> Done
```

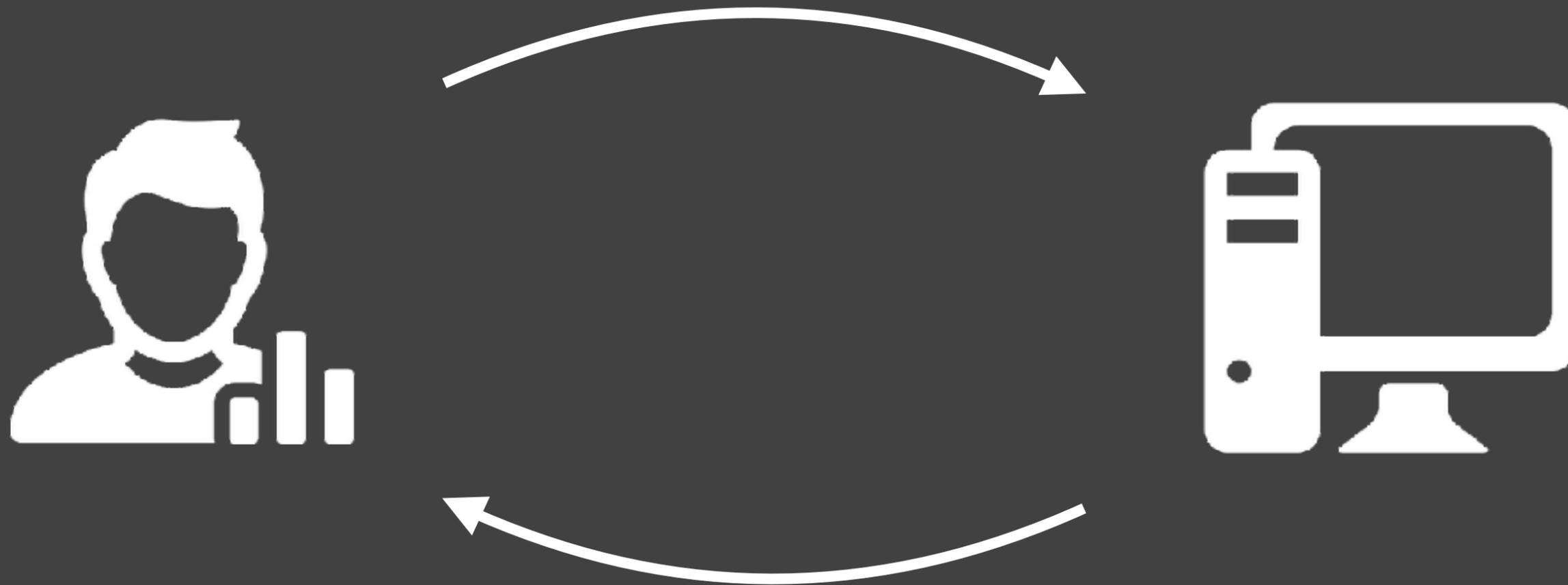
```
$ import download.zip
```

```
-> Done
```

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

```
-> Running Query. Please wait ...
```

Visual Analysis



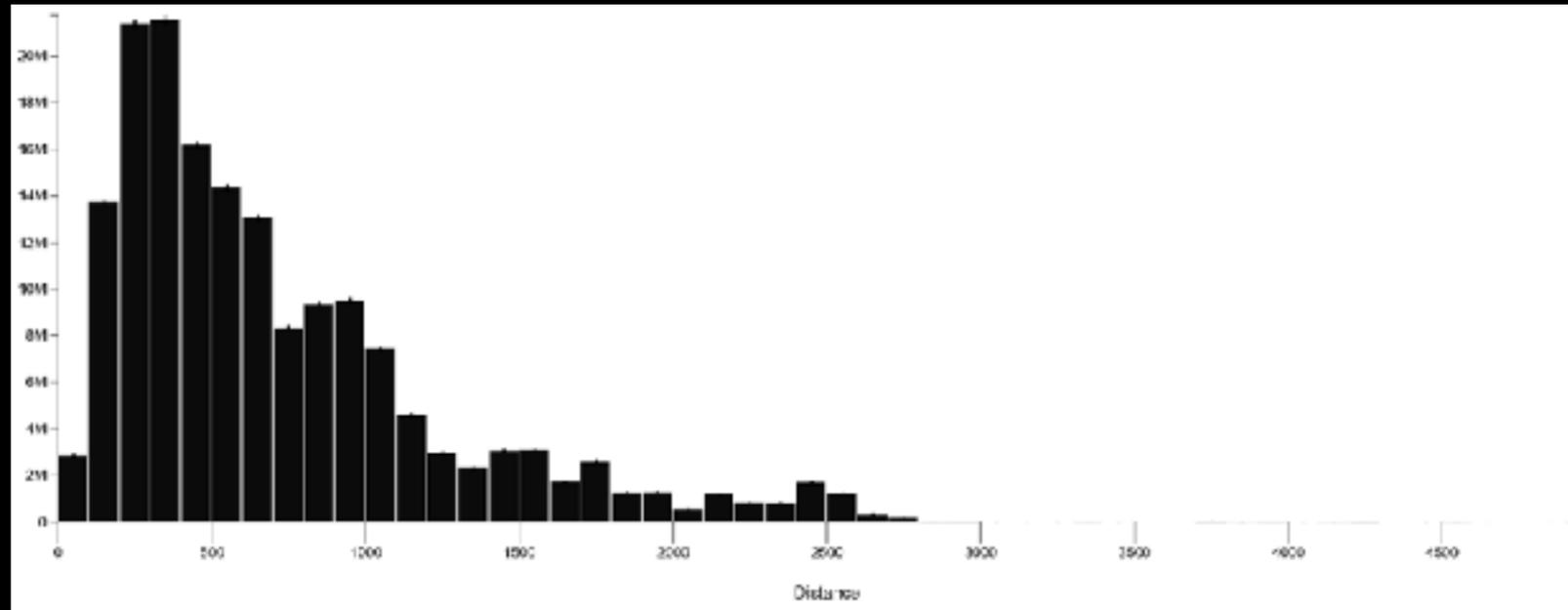
Big Data Visual Analysis



Query finished!

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

->



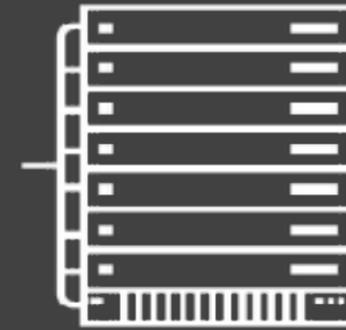
```
$ SELECT bin(distance), count(*)  
FROM flights  
WHERE airline = 'hi'
```

-> Running Query. Please wait ...

State of the Art in Big Data Exploration

Distributed Systems

Expensive and high latency.



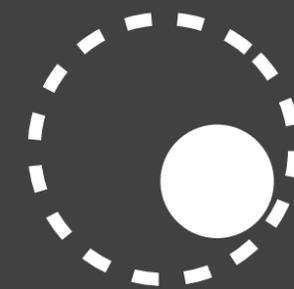
Indexes (Data Cubes)

Requires pre computation and limited queries.



Sampling

Use a representative subset of the data.



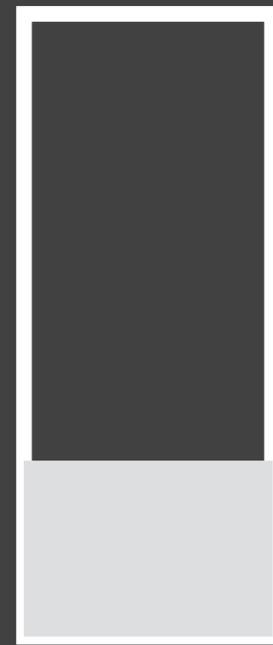
Sampling and Approximate Query Processing (AQP)

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 ~~25%~~ = ~~82~~

Sum of 100 % = 168 ±50

Query finished!

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

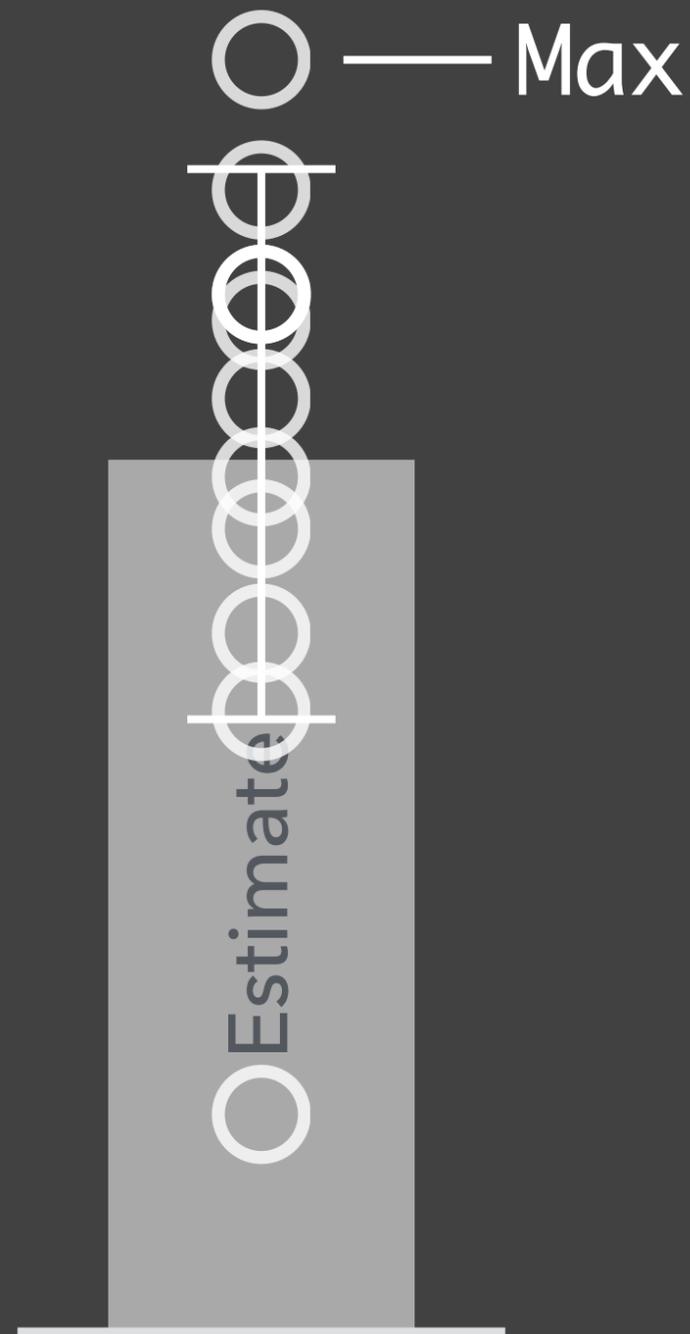
-> No Results

```
$ 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

Data: FAAData

Heatmap

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
- # DestWac

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

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

Value
Function: Count

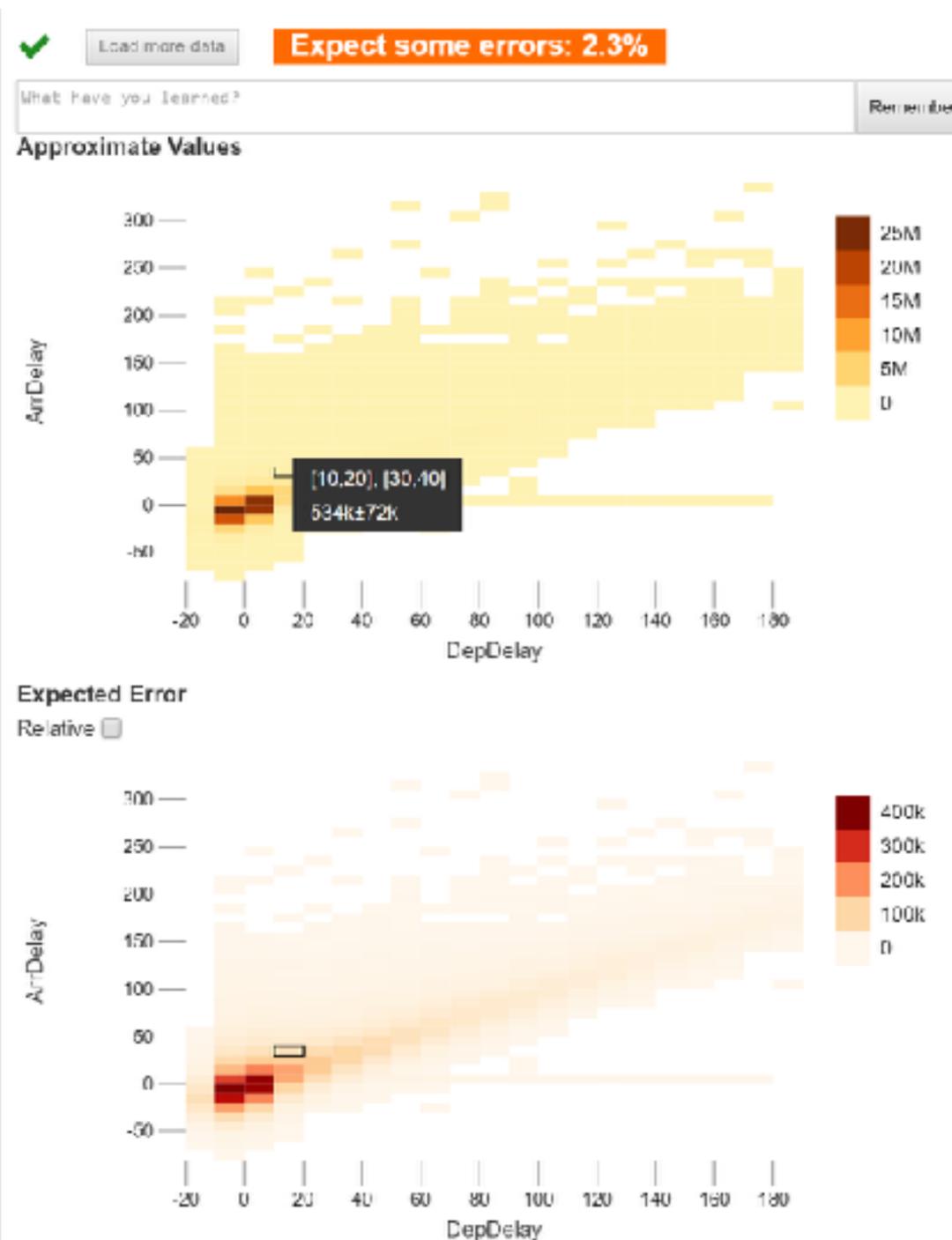
Persistent Filters
E.g. AND(Carrier = 'DL', DepDelay >= 0)

Filter set: clear

Zoom clear Capture as Filter

[ArrDelay \$RNGS
[[-140,00619517543057,390.49205043059655]]]

Query Specification



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...

Clear History Reset App

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

Heatmap

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 = 'DL', DepDelay >= 0)

Filter set: clear

Zoom clear Capture as Filter

- {ArrDelay \$RNGS [-140.00619517543057, 390.49205043059655]}
- {DepDelay \$RNGS [-19.819658218570382, 187.25649037534237]}

Load more data **Expect some errors: 2.3%**

What have you learned? Remember

Approximate Values

Expected Error Relative

Visualization View

Massive drop off after Sep 2001

Exact data loaded (18s)

3 decades of flights

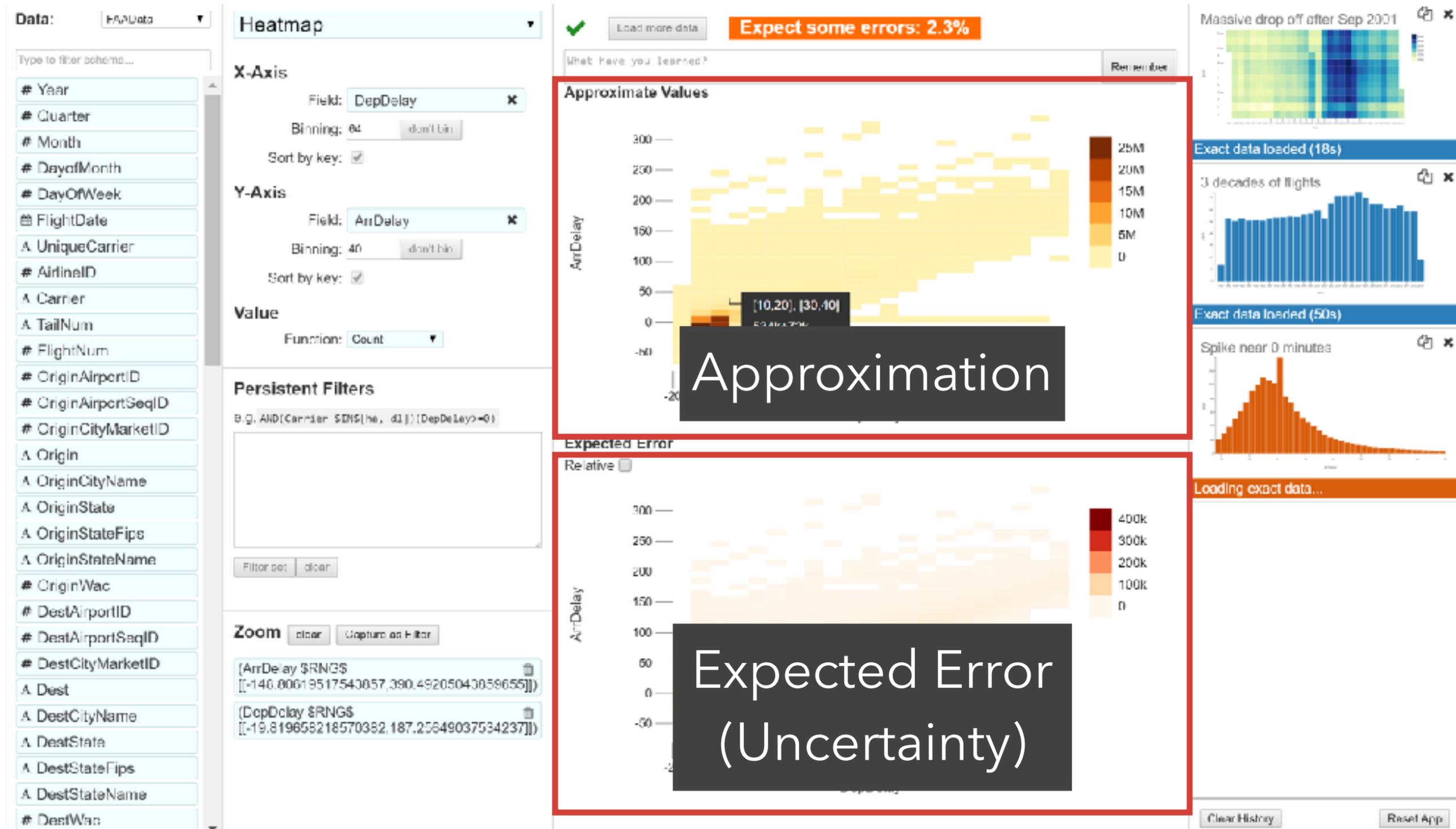
Exact data loaded (50s)

Spike near 0 minutes

Loading exact data...

Clear History Reset App

Pangloss implements Optimistic Visualization



Pangloss implements Optimistic Visualization

The screenshot displays the Pangloss data visualization interface. On the left is a sidebar with a list of fields for filtering and zooming. The main area is divided into several sections:

- Heatmap Configuration:** X-Axis is 'DepDelay' (Binning: 64), Y-Axis is 'ArrDelay' (Binning: 40), and Value is 'Count' (Function: Count).
- Persistent Filters:** A text input field containing a filter expression: `B.G. AND[Carrier $ENS[he, dl]]:(DepDelay>=0)`.
- Zoom:** Two zoomed-in regions are shown:
 - Top: `{ArrDelay $RNGS [-140.00019517543057,390.49205043059655]}`
 - Bottom: `{DepDelay $RNGS [-19.819658218570382,187.25649037534237]}`
- Annotation:** A red-bordered box highlights a green checkmark, a 'Load more data' button, and an orange box stating 'Expect some errors: 2.3%'. Below this is a text input 'What have you learned?' and a 'Remember' button.
- Approximate Values:** A dark grey box displays the coordinates `[10,20], [30,10]` and the value `534k±72k`.
- Expected Error:** A checkbox labeled 'Relative' is present.
- Other Visualizations:** On the right, there are three smaller charts: a heatmap titled 'Massive drop off after Sep 2001', a bar chart titled 'Loaded (18s)', and a histogram titled 'Spike near 0 minutes'. A status bar at the bottom right shows 'Loading exact data...'.

Annotation + Remember Button

Pangloss implements Optimistic Visualization

The screenshot displays the Pangloss data visualization interface, which implements Optimistic Visualization. The interface is divided into several sections:

- Data:** A dropdown menu showing 'FAADATA' and a search bar for filtering schemas.
- Heatmap:** The main visualization area, currently showing a heatmap of flight delays. The X-axis is 'DepDelay' and the Y-axis is 'ArrDelay'. The value function is set to 'Count'. The heatmap shows a dense cluster of yellow and orange pixels, indicating a high frequency of flights with low delays. A tooltip shows a value of 534k ± 72k for a specific bin.
- Approximate Values:** A section showing the approximate values of the data, with a color scale ranging from 0 to 25M. A status bar indicates 'Expect some errors: 2.3%'.
- Expected Error:** A section showing the expected error, with a color scale ranging from 0 to 400k.
- History:** A panel on the right showing a list of data loading events, including 'Massive drop off after Sep 2001', 'Exact data loaded (18s)', '3 decades of flights', 'Exact data loaded (50s)', and 'Spike near 0 minutes'. A 'History' button is visible at the bottom of this panel.

The interface also includes a sidebar with a list of fields for filtering, such as Year, Quarter, Month, DayofMonth, DayOfWeek, FlightDate, UniqueCarrier, AirlinerID, Carrier, TailNum, FlightNum, OriginAirportID, OriginAirportSeqID, OriginCityMarketID, Origin, OriginCityName, OriginState, OriginStateFips, OriginStateName, OriginWac, DestAirportID, DestAirportSeqID, DestCityMarketID, Dest, DestCityName, DestState, DestStateFips, and DestStateName.

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

Heatmap

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

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

Value: Function: Count

Persistent Filters: B.G. AND(Carrier = 'DL', DepDelay >= 0)

Zoom: clear, Capture as Filter

- {ArrDelay \$RNGS [-140.00619517543057, 390.49205043059655]}
- {DepDelay \$RNGS [-19.819658218570382, 187.25649037534237]}

Load more data **Expect some errors: 2.3%**

What have you learned? Remember

Approximate Values

ArrDelay

DepDelay

Expected Error Relative

ArrDelay

DepDelay

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...

Clear History Reset App

Data: FAAData

Barchart

Load more data

Expect almost no errors: 0.2%

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

X-Axis

Field: Carrier

Binning: 0 bin

Secondary Field:

Sort by key:

Value

Function: Count

Persistent Filters

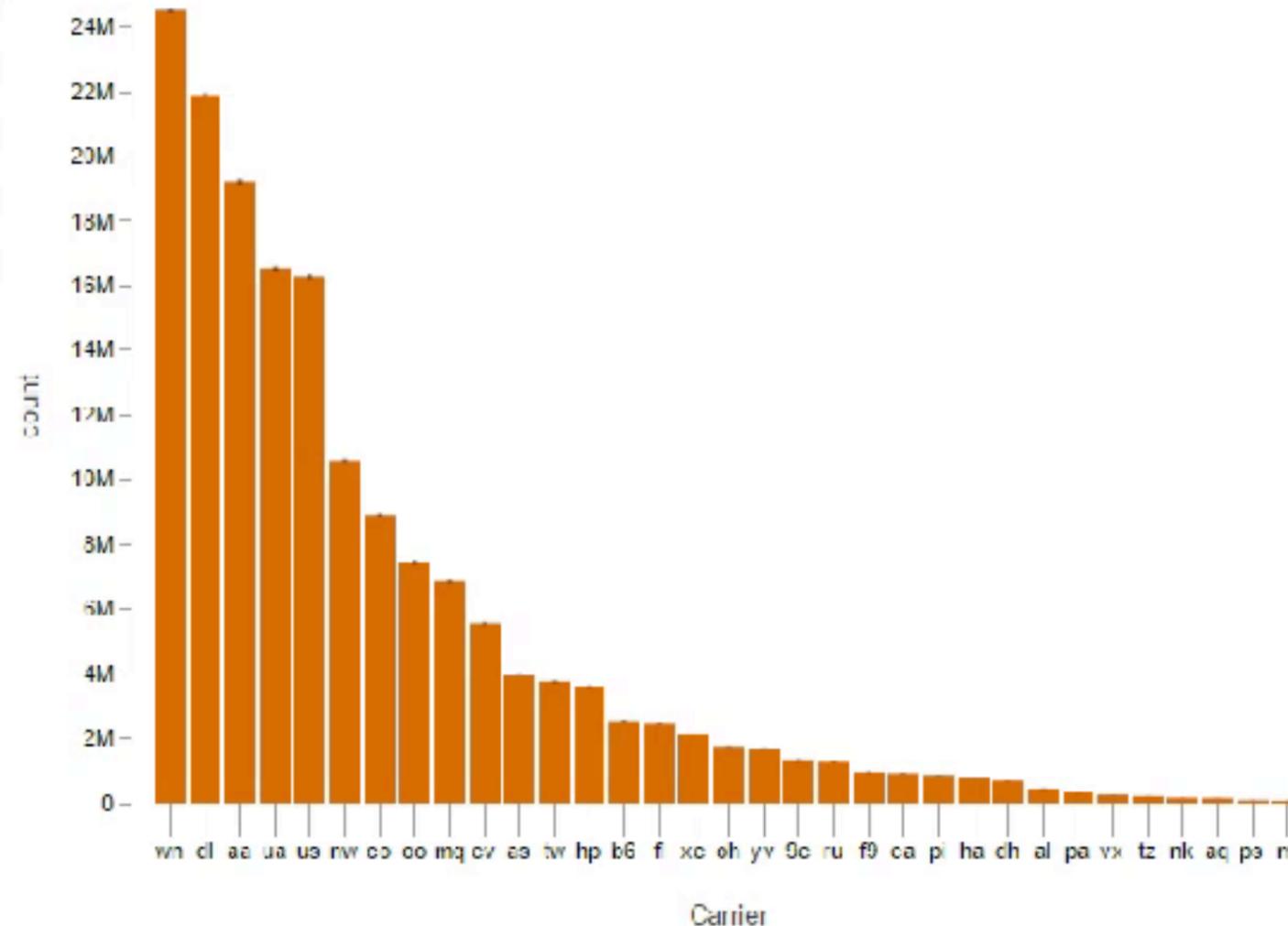
e.g. AND(Carrier IN ['ha', 'dl'])(DepDelay>=0)

Filter set clear

Zoom clear Capture as Filter

What have you learned?

Remember



170M ~100ms query time (30 years).

Clear History

Reset App

Data: FAAData

Barchart

Load more data

Expect almost no errors: 0.3%

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

X-Axis

Field: Year

Binning: 0 bin

Secondary Field:

Sort by key:

Value

Function: Count

Persistent Filters

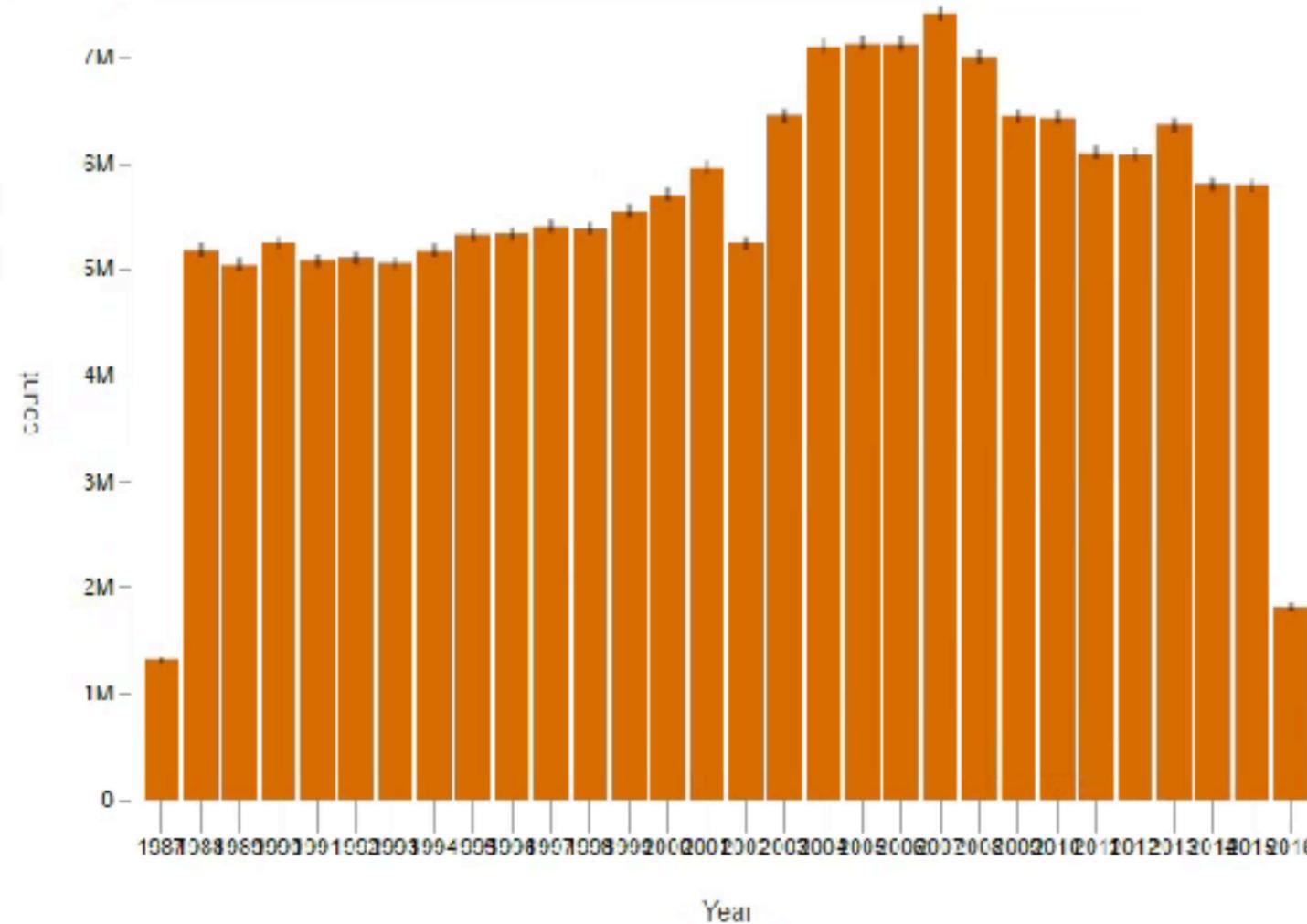
e.g. AND(Carrier IN ['ha', 'dl'])(DepDelay>=0)

Filter set clear

Zoom clear Capture as Filter

What have you learned?

Remember



Data: FAAData

Barchart

Load more data

Expect almost no errors: 0.3%

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

X-Axis

Field: Year

Binning: 0 bin

Secondary Field:

Sort by key:

Value

Function: Count

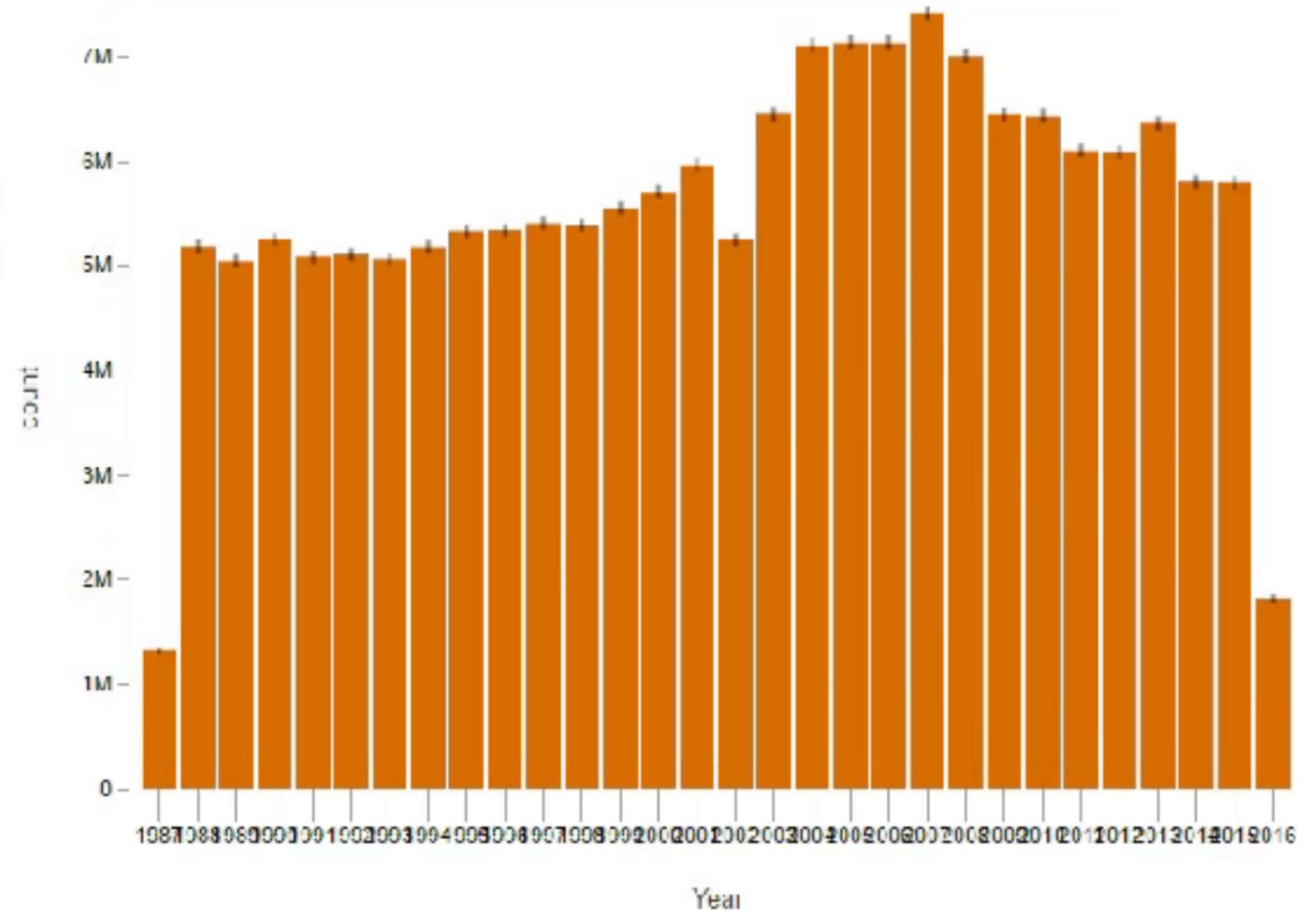
Persistent Filters

e.g. AND(CARRIER IN ['ha', 'dl'])(DEPDELAY >= 0)

Filter set clear

Zoom clear Capture as Filter

3 decades of flights Remember



"Remember" button moves query into the background

Clear History

Reset App

Data: FAAData

Barchart



Load more data

Expect almost no errors: 0.3%

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

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

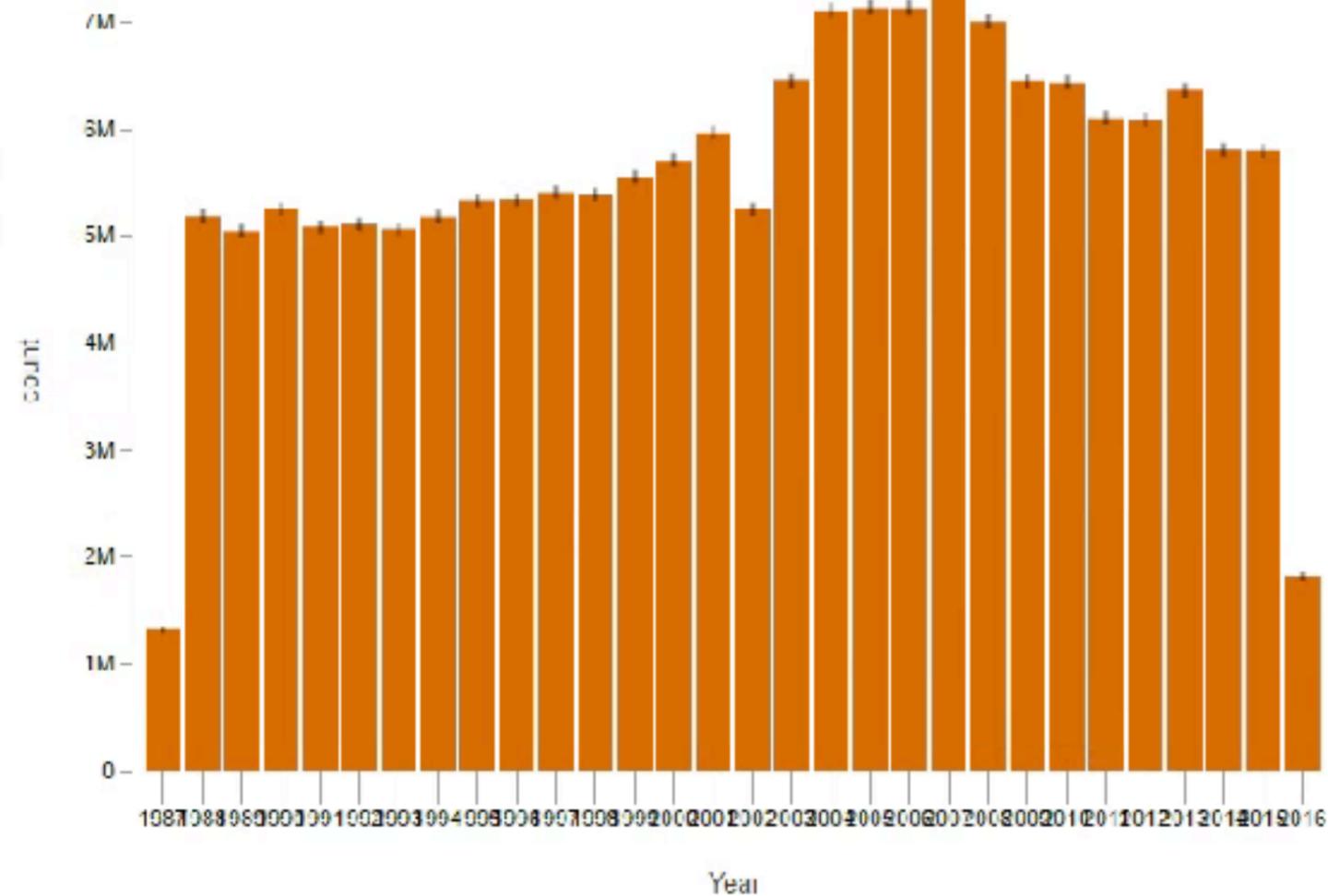
Zoom

clear

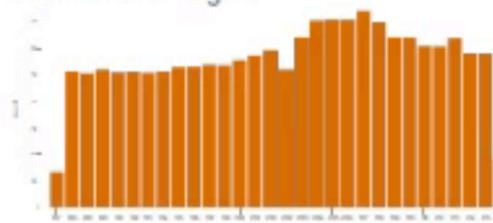
Capture as Filter

What have you learned?

Remember



3 decades of flights



Loading exact data...

Continue exploration without waiting

Clear History

Reset App

Data: FAAData

Barchart

- orig
- # OriginAirportID
- # OriginAirportSeqID
- # OriginCityMarketID
- A Origin
- A OriginCityName
- A OriginState
- A OriginStateFips
- A OriginStateName
- # OriginWac

X-Axis

Field: OriginCityName

Binning: 0 bin

Secondary Field:

Sort by key:

Value

Function: Count

Persistent Filters

e.g. AND(CARRIER \$INS[ha, d1])(DepDelay>=0)

(OriginState=tt)

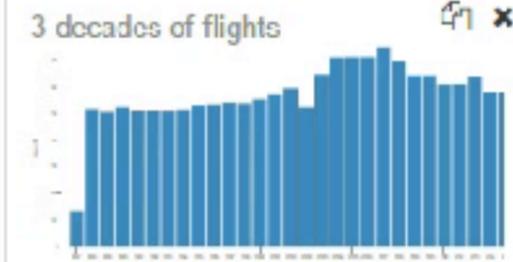
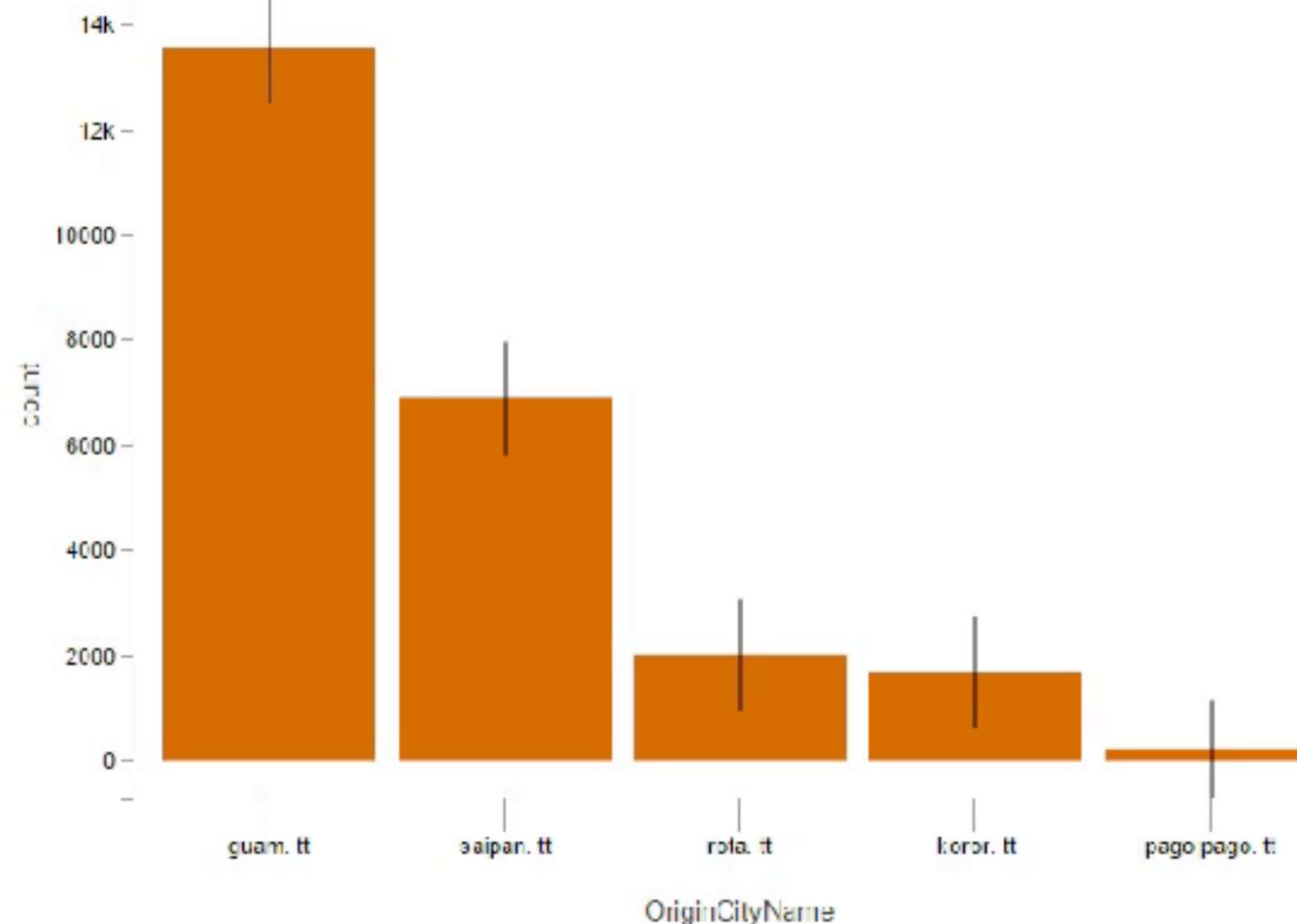
Filter set clear

Zoom clear Capture as Filter

Load more data Expect some errors: 7.5%

What have you learned?

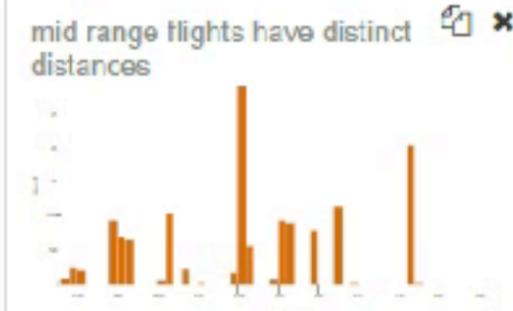
Remember



Exact data loaded (61.156s)



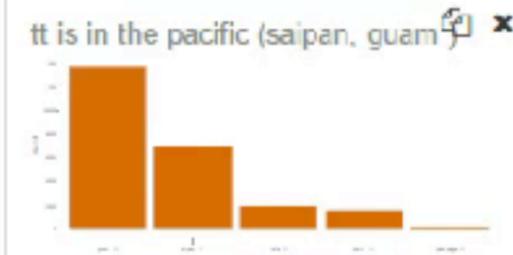
Exact data loaded (61.153s)



Loading exact data...



Loading exact data...



Clear History

Reset App

Orange → Approximate Blue → Precise

Data: FAAData

Heatmap

mostly ca to ha

- # OriginAirportID
- # OriginAirportSeqID
- # OriginCityMarketID
- A Origin
- A OriginCityName
- A OriginState
- A OriginStateFips
- A OriginStateName
- # OriginWac

X-Axis

Field: OriginState

Binning: 0

Sort by key:

Y-Axis

Field: DestState

Binning: 0

Sort by key:

Value

Function: Count

Persistent Filters

```

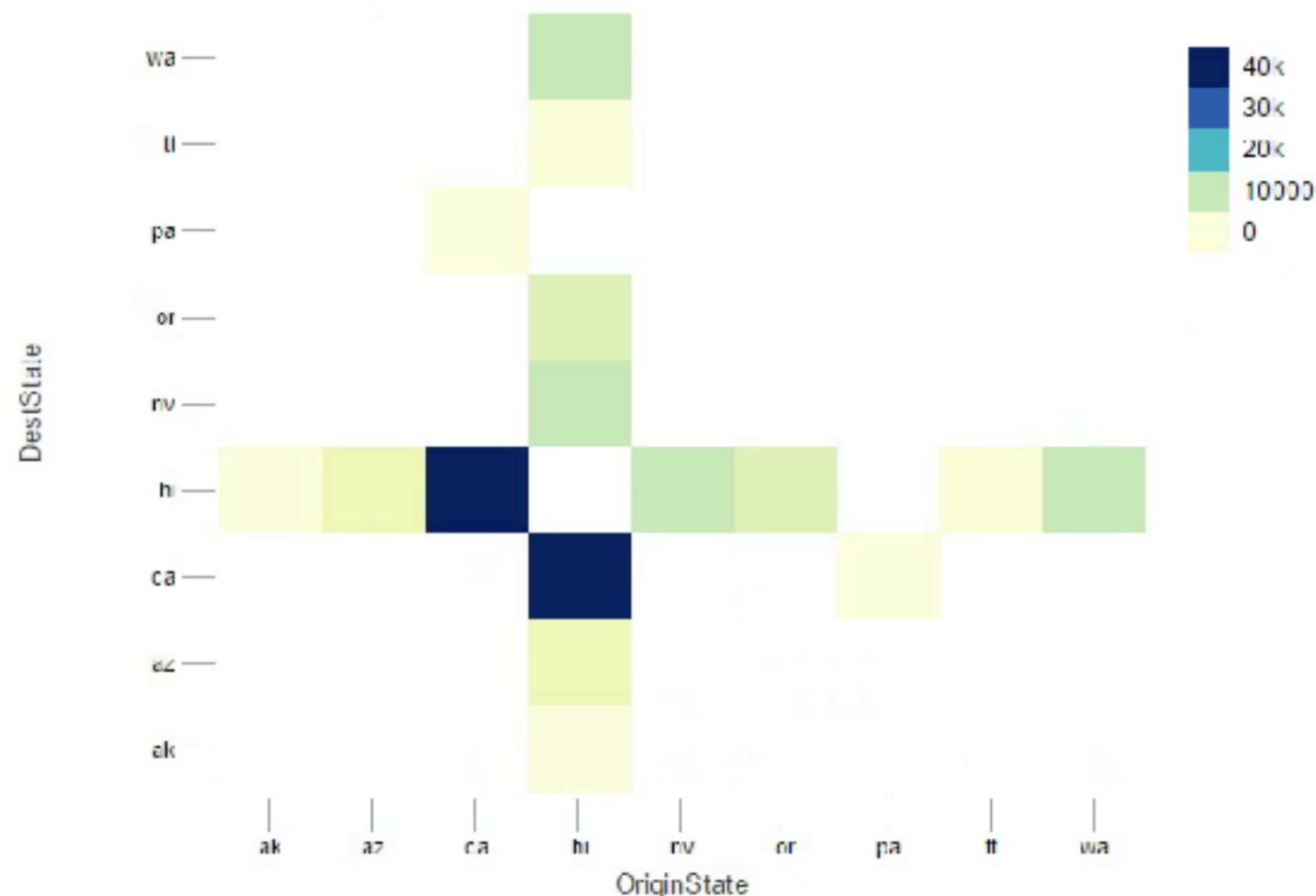
AND(CARRIER $TNS[ha, d1])(DEPARTSYS=0)
AND(CARRIER=ha)(DISTANCE $RNG$
[[2168.9792406152524,3201.570399053
4585]])

```

Zoom

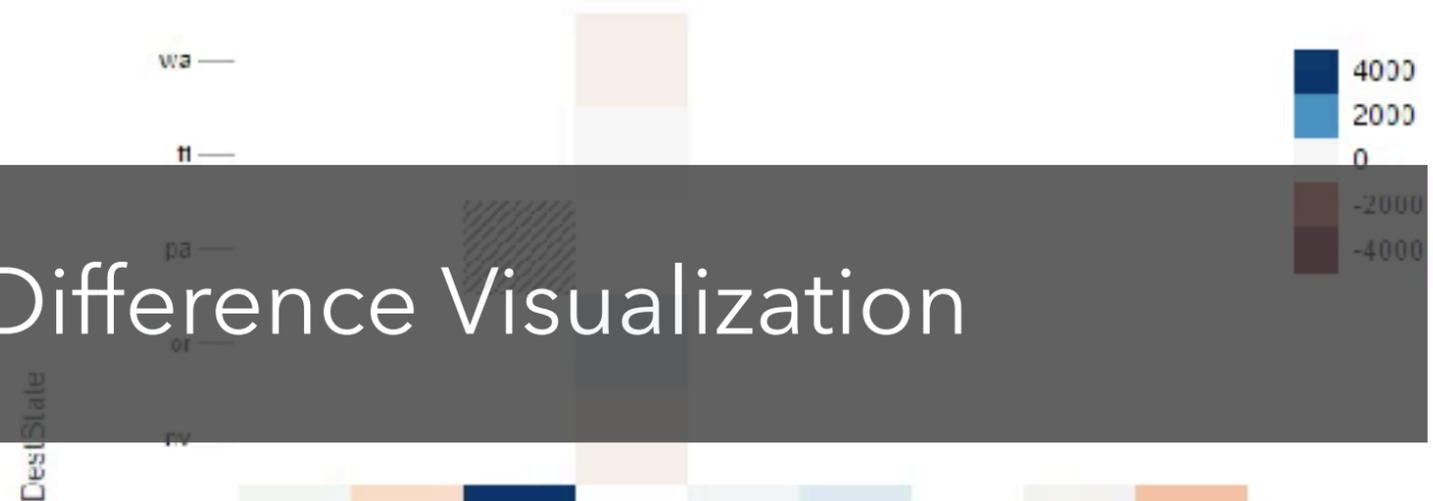
The visualization is read only because you're looking at the history. [Return to the working vis](#) or make a [copy of the current chart](#).

Exact Data



Difference to Approximate Data

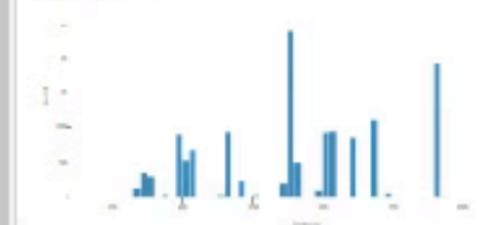
Relative



Difference Visualization

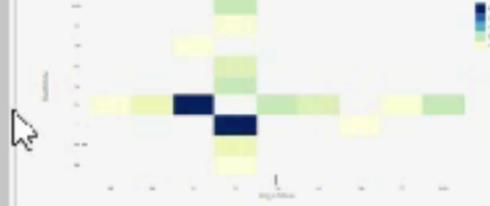
Exact data loaded (61.153s)

mid range flights have distinct distances



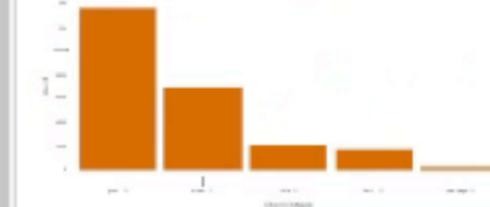
Exact data loaded (61.149s)

mostly ca to ha



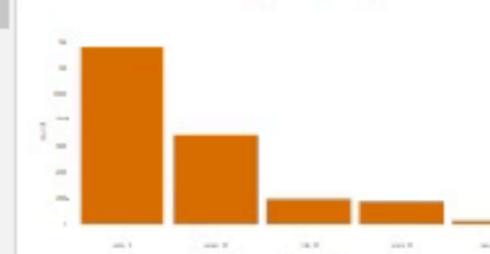
Exact data loaded (60.013s)

it 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

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.’”

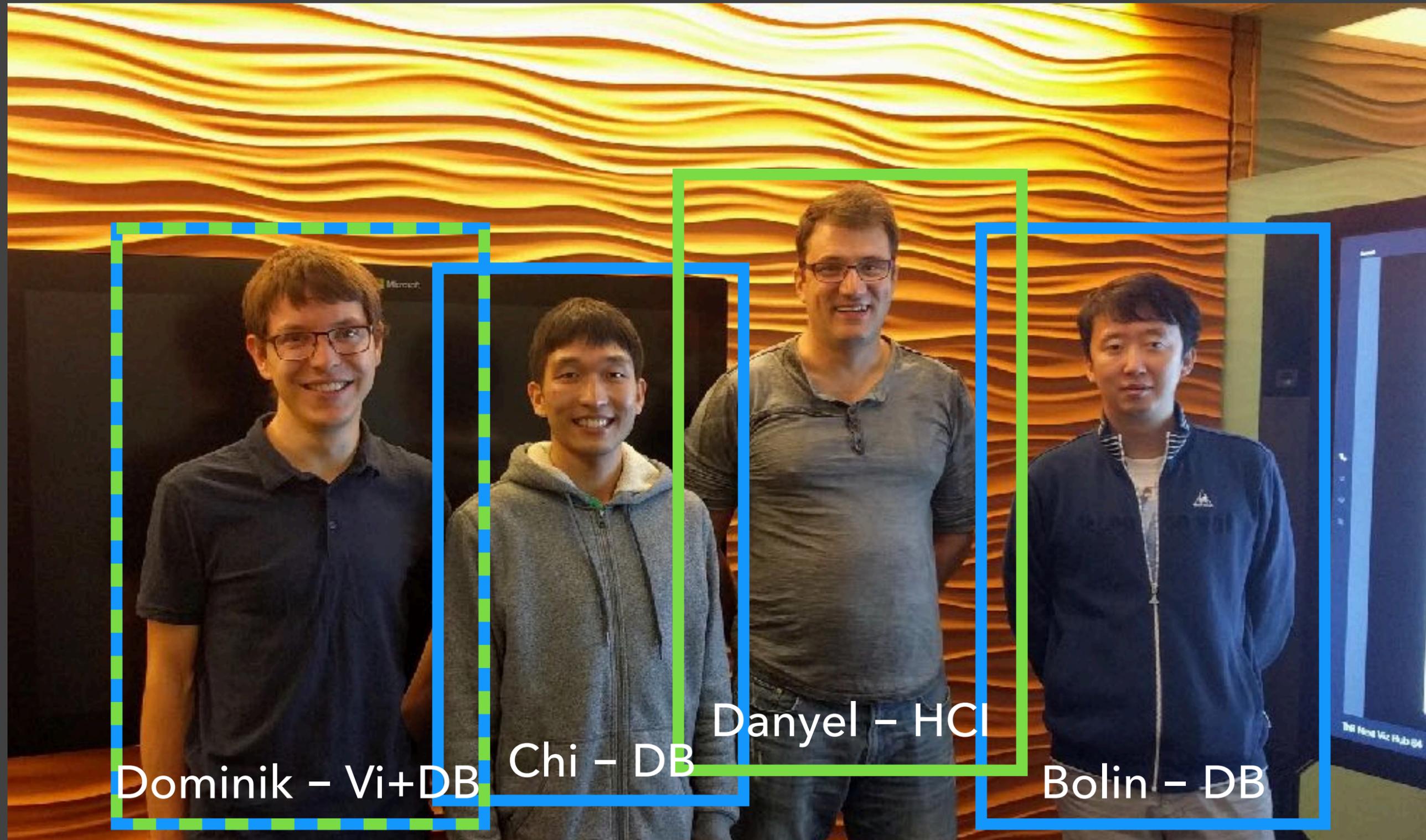
Conclusions

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

Gives analysts confidence in AQP

Future: Alerting, Remembering, Progressive + Optimistic

AQP needs Multi-Disciplinary Solutions



Dominik - Vi+DB

Chi - DB

Danyel - HCI

Bolin - DB

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

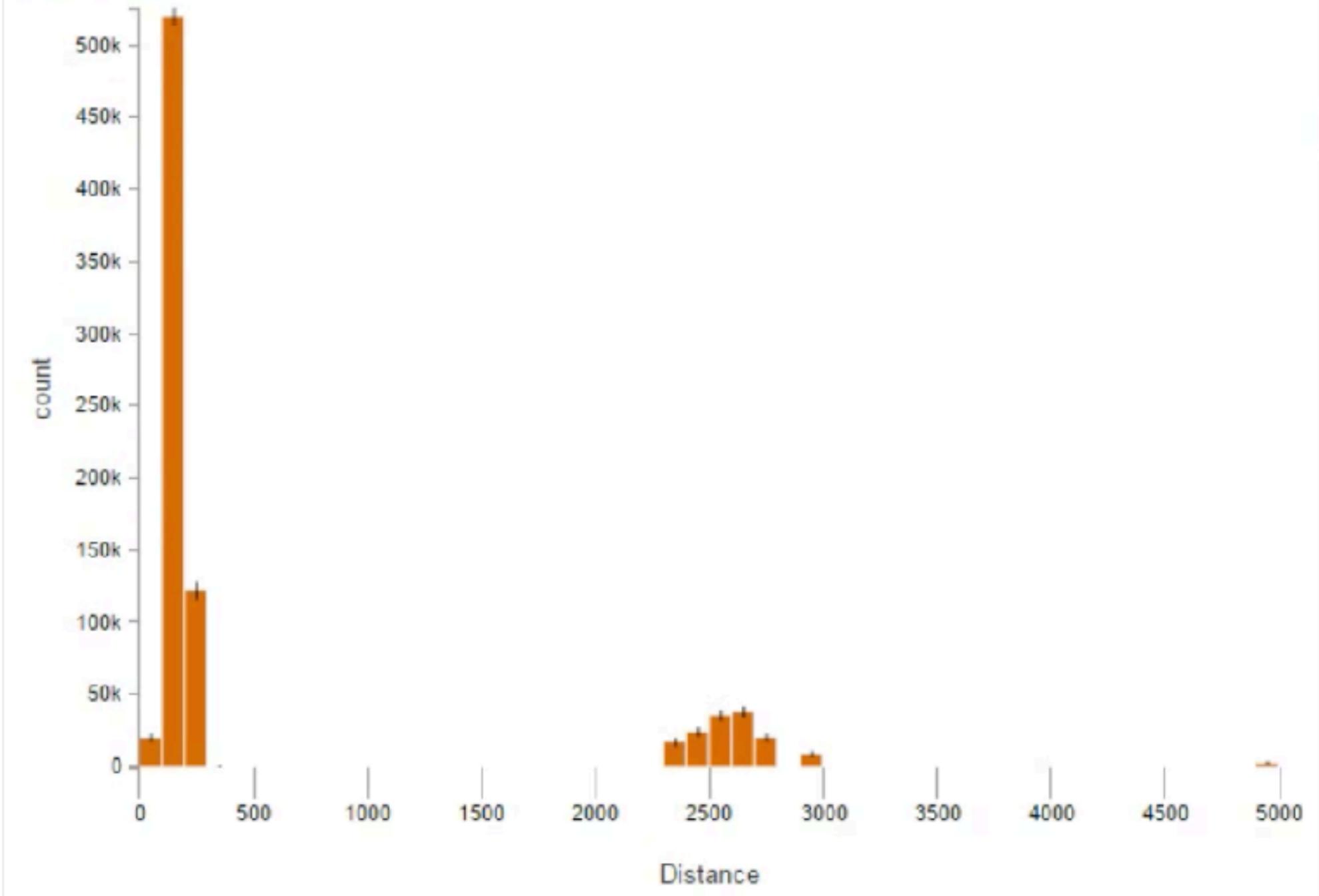
Dominik Moritz @domoritz
Danyel Fisher @FisherDanyel
Bolin Ding @AtlasDing
Chi Wang



Query finished!

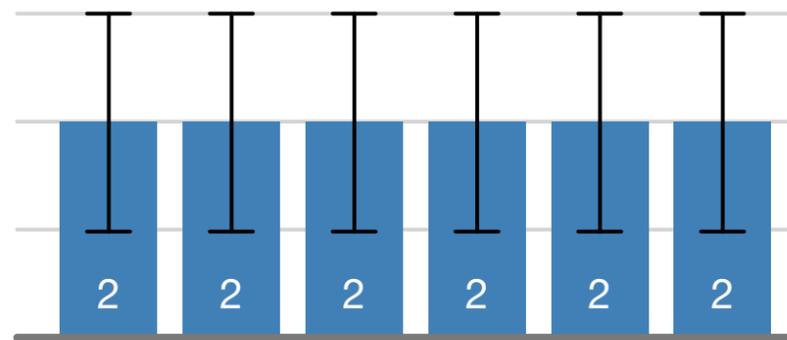
Backup Slides

Histogram of Distances for Hawaiian Airlines

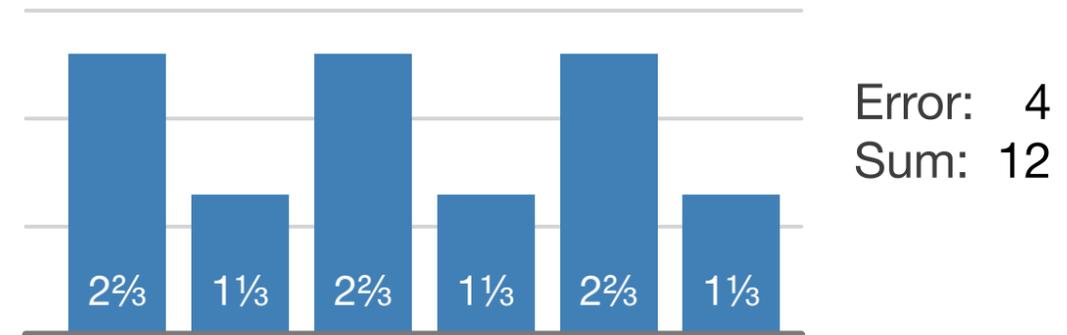


Distribution Uncertainty

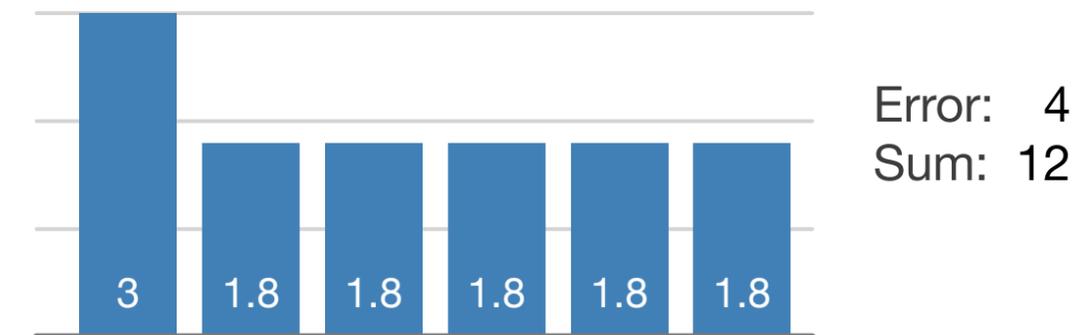
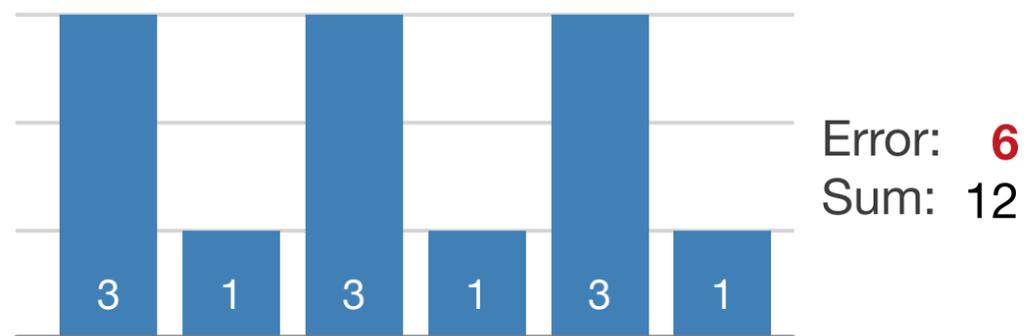
Approximation
Distribution Uncertainty: 4



Within Distribution Uncertainty



Outside Distribution Uncertainty



Distribution Uncertainty



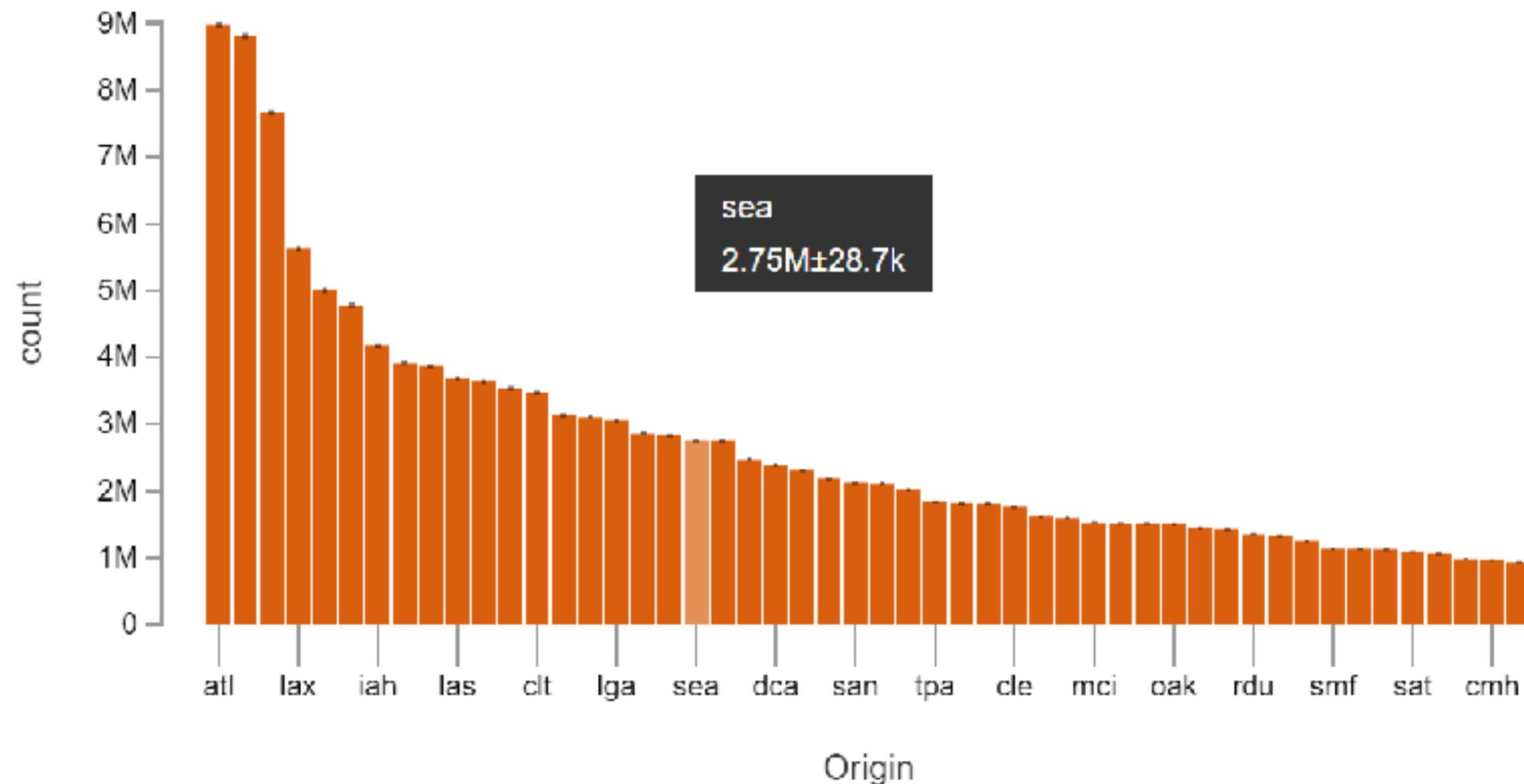
Load more data

Expect almost no errors: 0.5%

What have you learned?

Remember

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



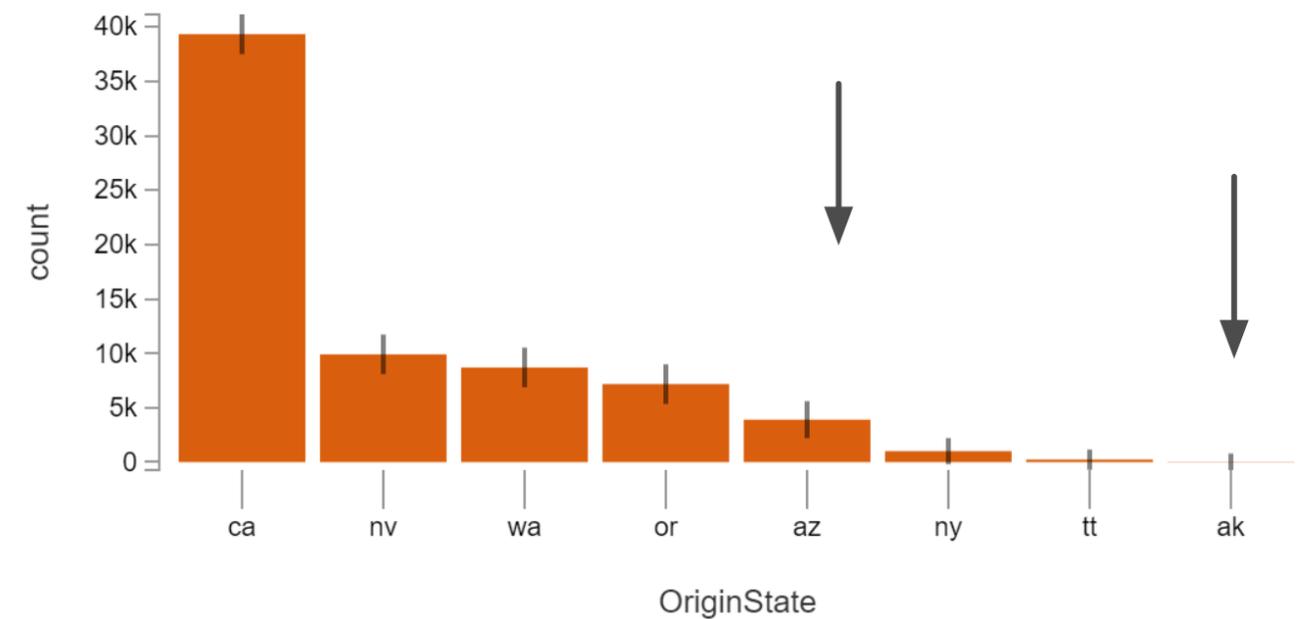
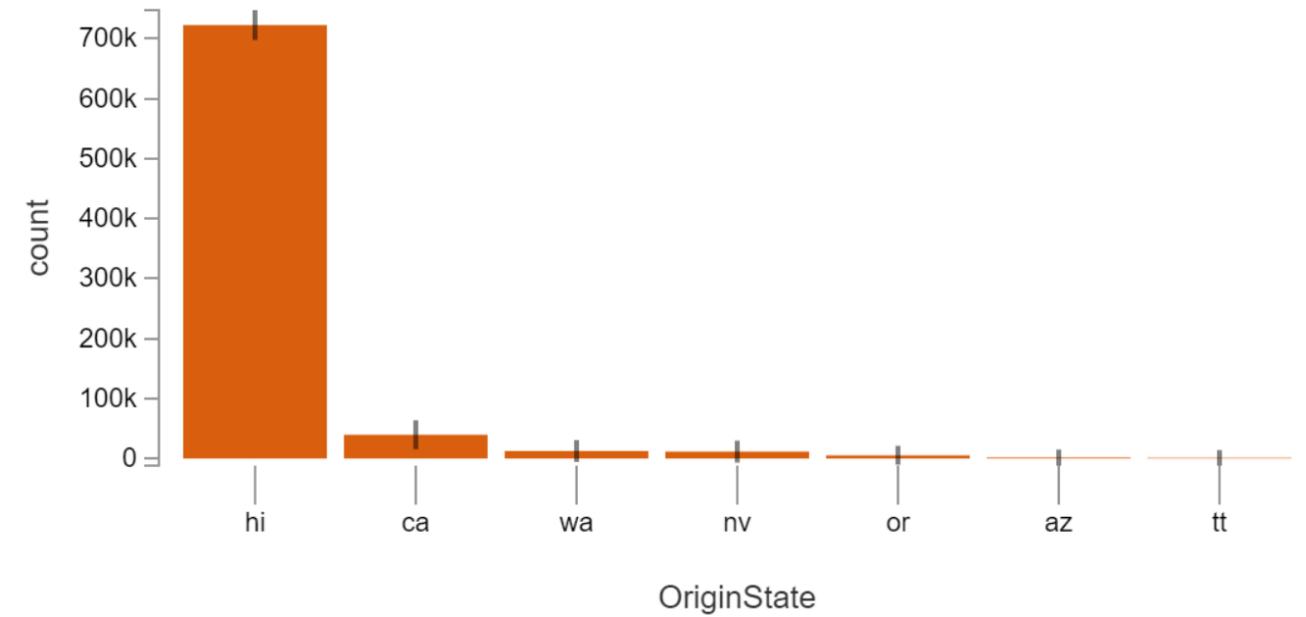
Filtering can show new groups

new predicate →

new query →

different sample →

different groups



Precise results can show new groups

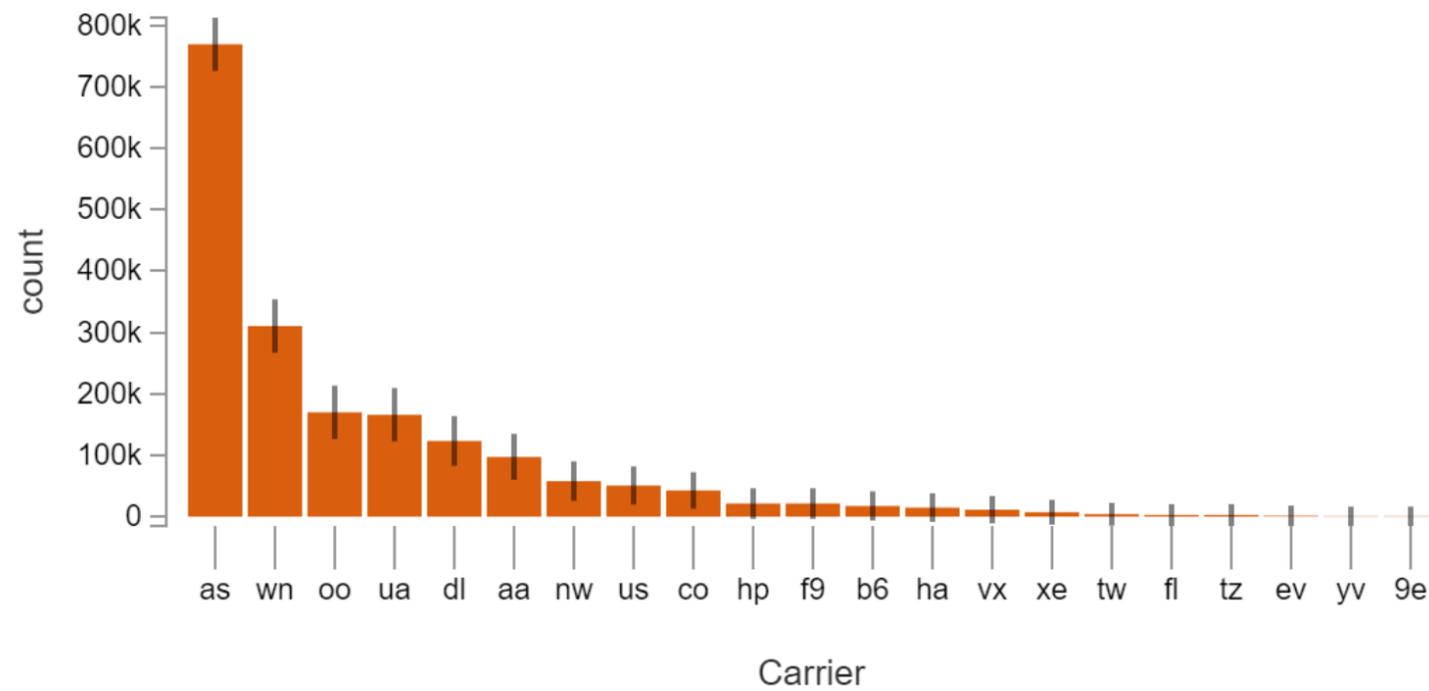


Load more data

Expect some errors: 6.2%

What have you learned?

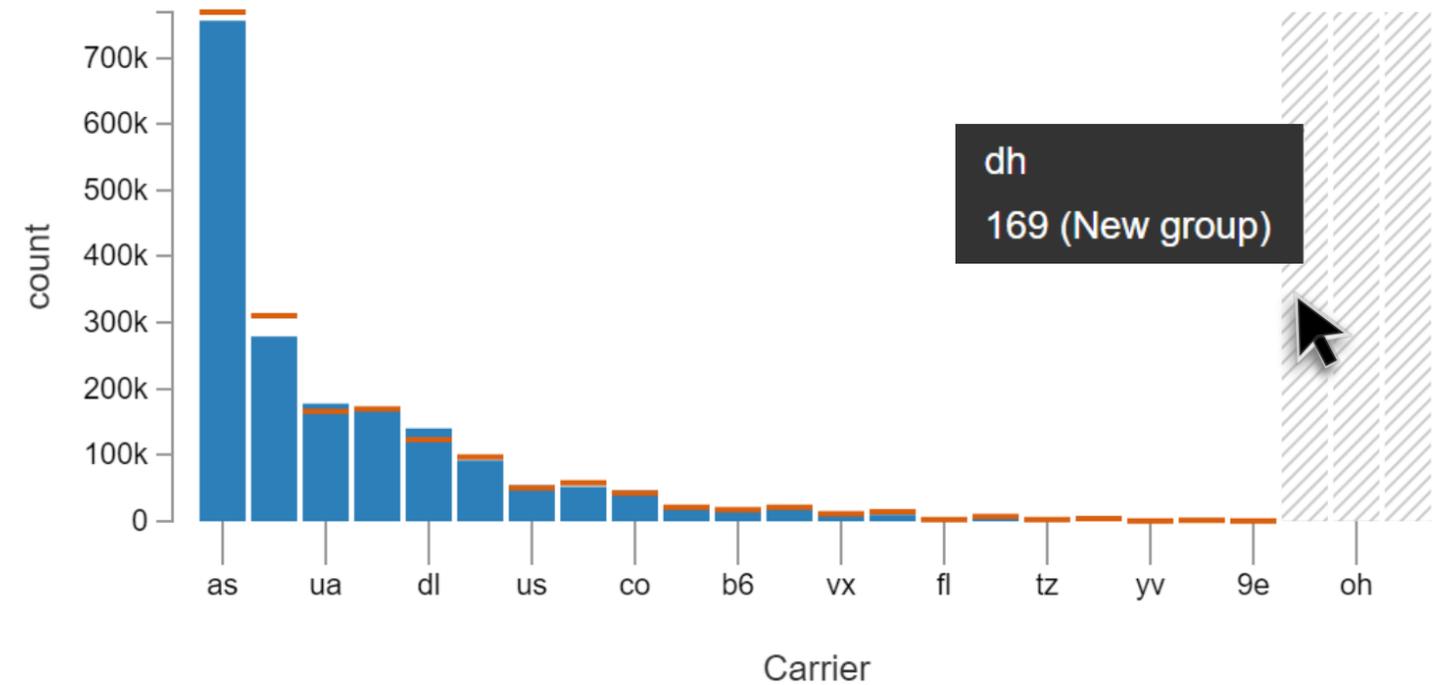
Remember



Approximate

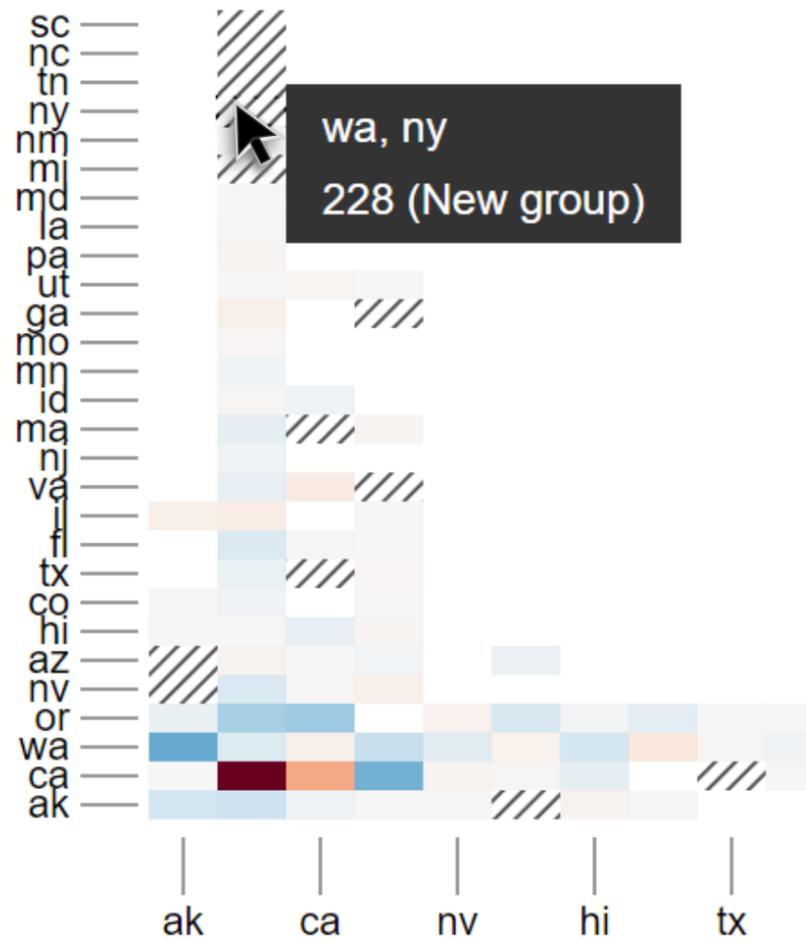
What have you learned?

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](#).

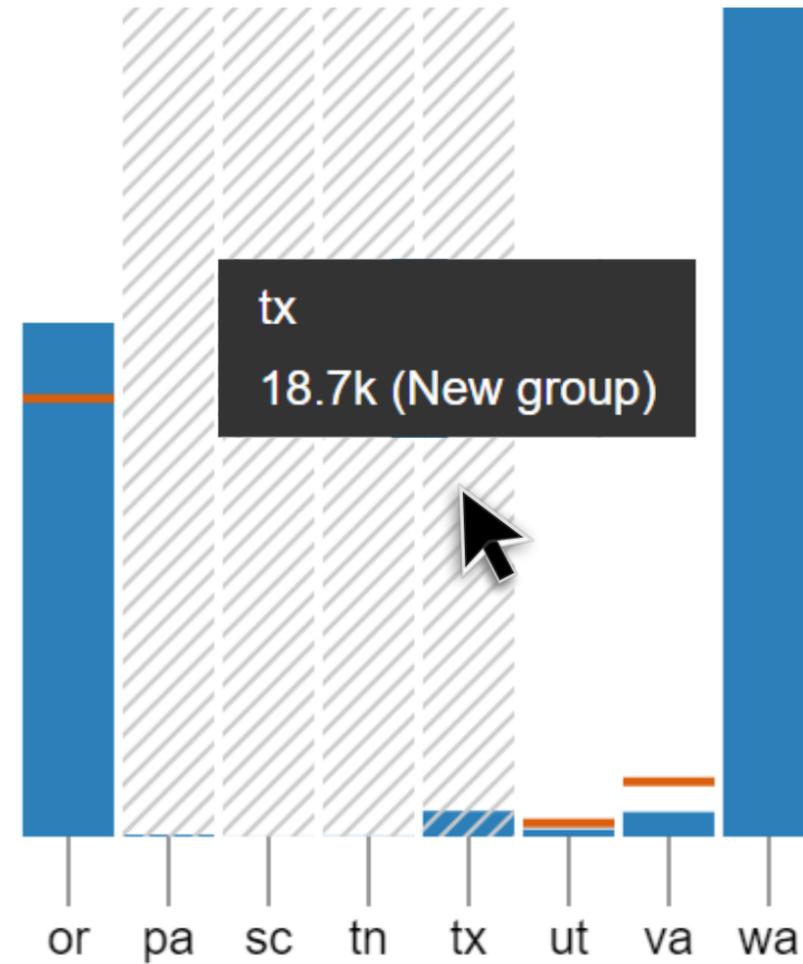


Precise

Vocabulary of visual cues



Heatmap



Bar chart