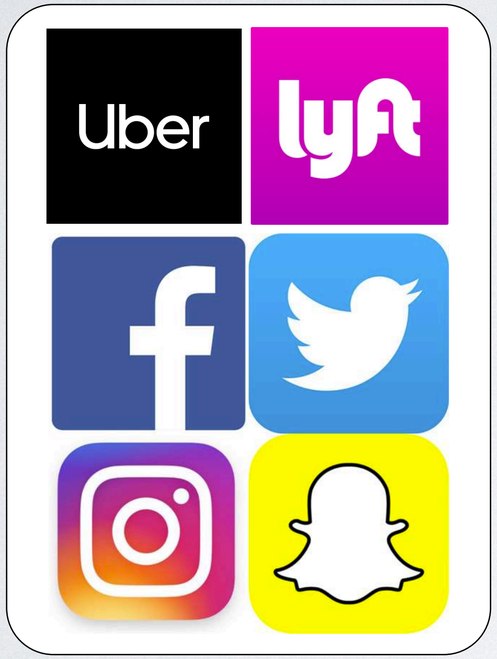
Turbocharging Geospatial Visualization Dashboards via a Materialized Sampling Cube Approach

Jia Yu Mohamed Sarwat

Arizona State University

Big geospatial data

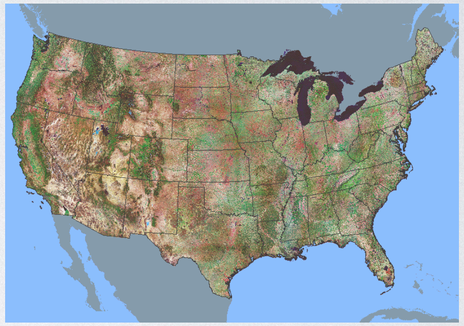
Mobile devices



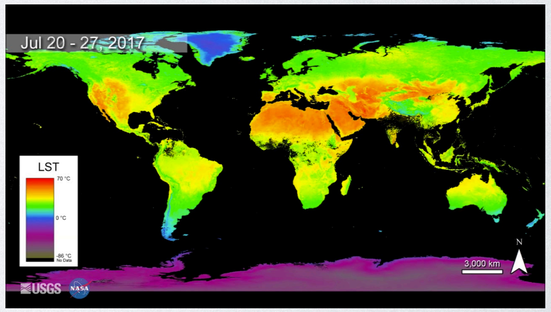
IoT sensors

Climate monitoring

USGS
science for a changing world



NASA



https://www.statista.com/statistics/274774/forecast-of-mobile-phone-users-worldwide/

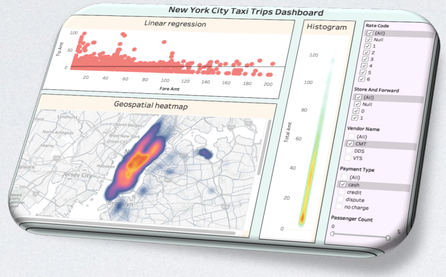
https://iot-analytics.com/state-of-the-iot-update-q1-q2-2018-number-of-iot-devices-now-7b/

https://earthdata.nasa.gov/about/eosdis-cloud-evolution

Spatial data science pipeline

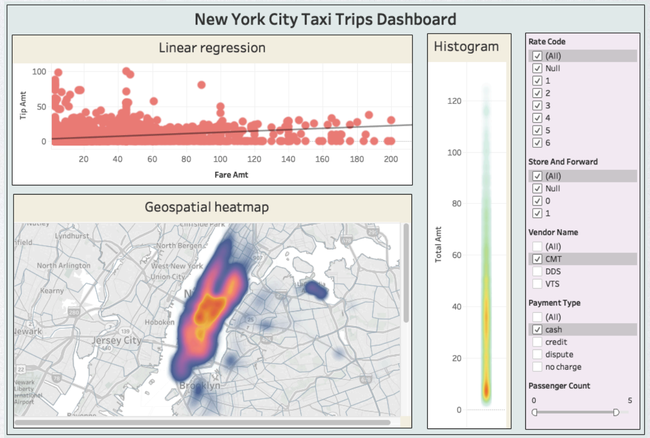
Arizona State University

Analytics tool



Spatial visualization
Spatial data mining

Interactive visualization dashboard



Spatial database

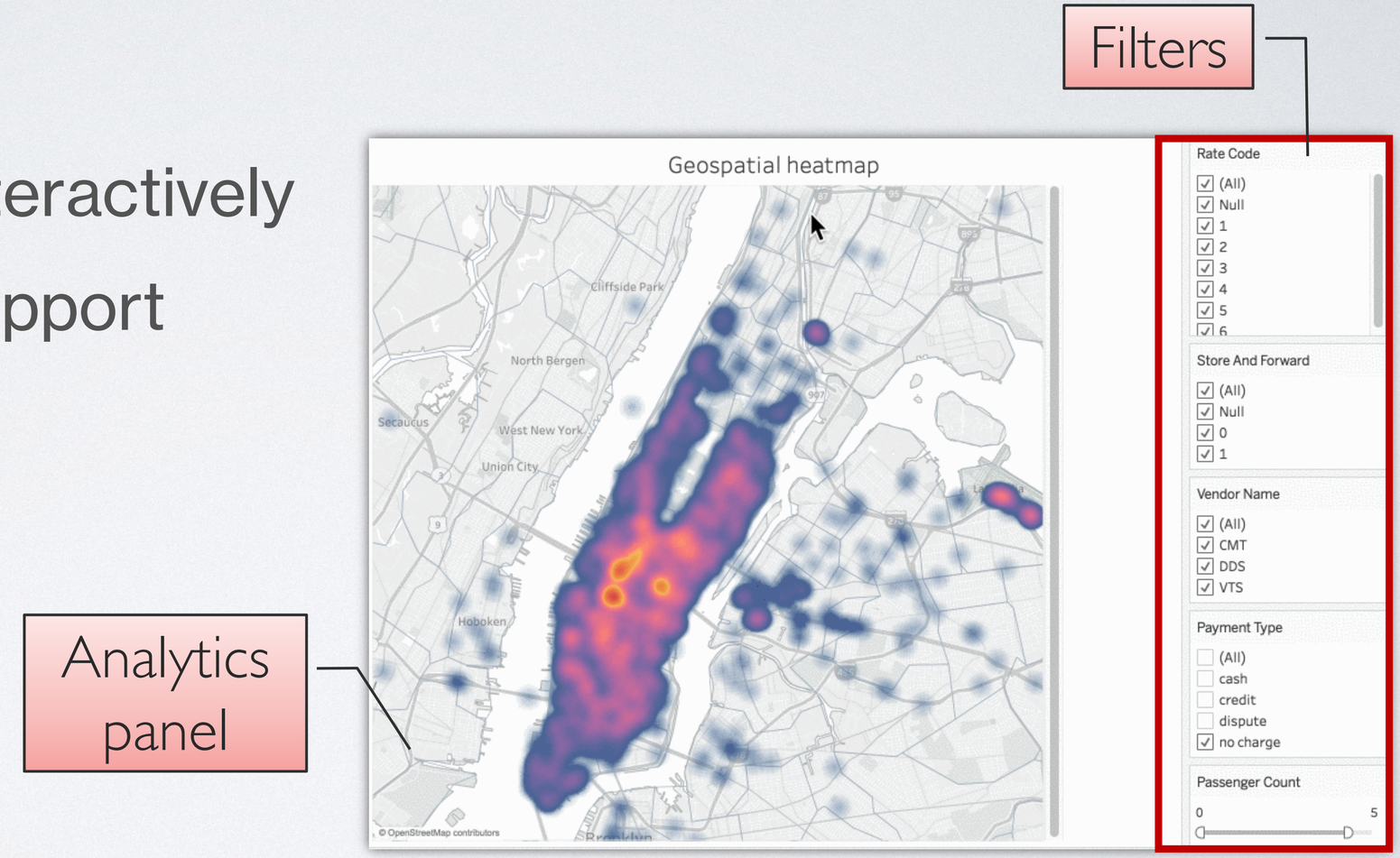
Spatial index
Spatial queries

PostgreSQL PostGIS Apache Spark Hive

tableau Apache Zeppelin

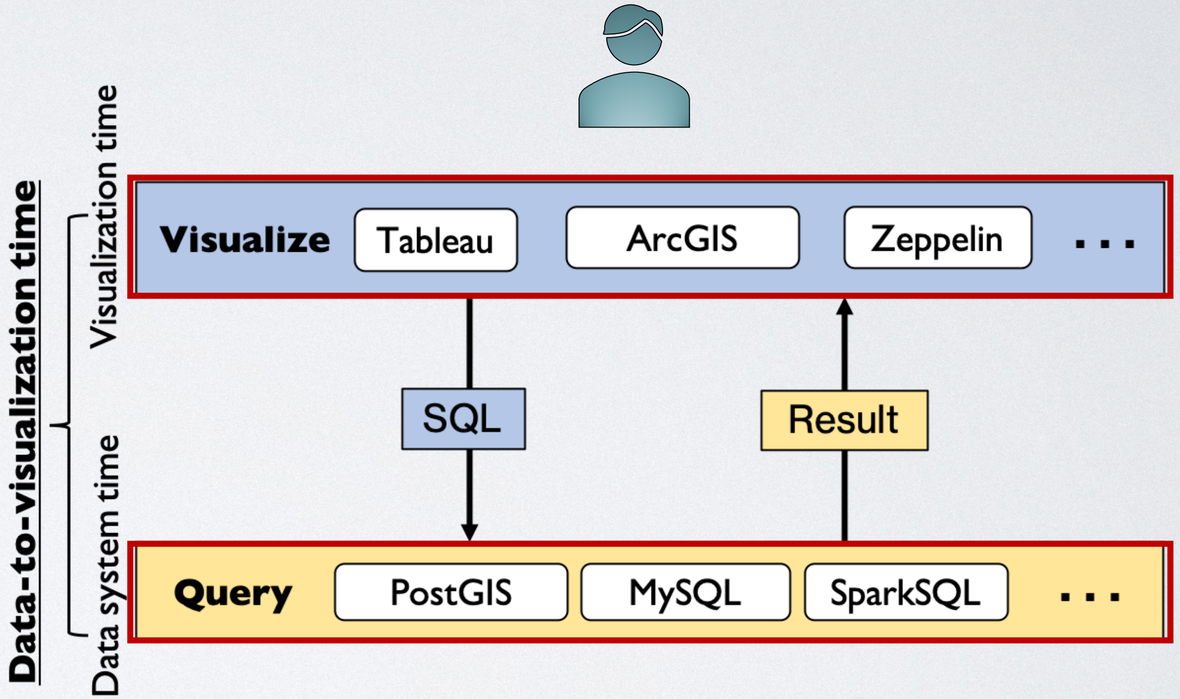
Interactive visualization dashboard

- Tableau, ArcGIS, ...
- Different population interactively
- Interactive analytics support



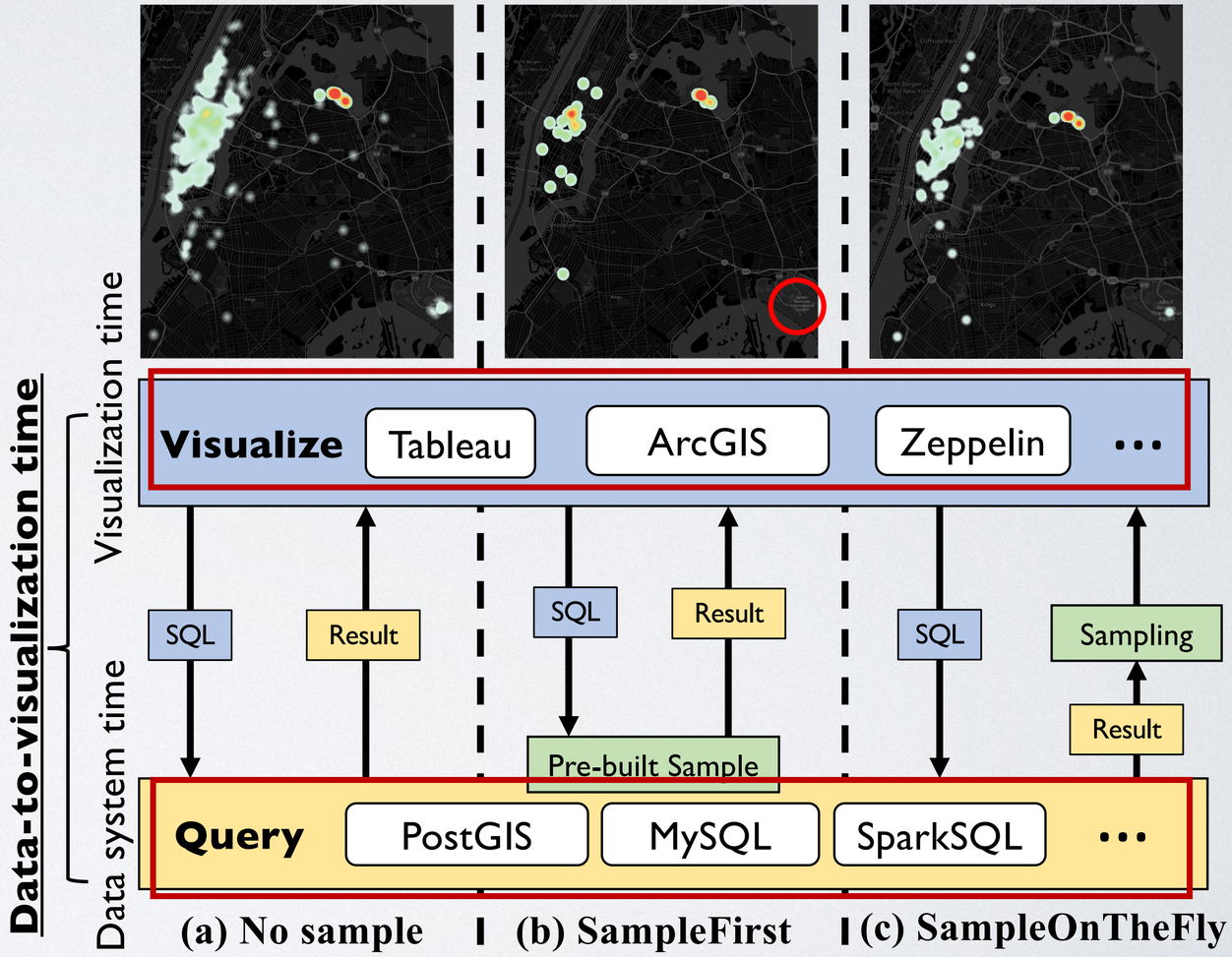
Interactive visualization dashboard

- Problem on big spatial data
- Step 1: DB query
  - Several minutes
  - Increase with data size
- Step 2: Visualize results
  - Long or crash
    - Tableau / Google Maps: stuck at 100 MB for heat map



Sampling techniques

- People may tolerate some accuracy loss for visualization
- Sample first
  - Ignore important patterns
- Online sample after every query
  - Sample on the fly
  - Viz fast and accurate
  - Query still slow
  - POIsam and VAS

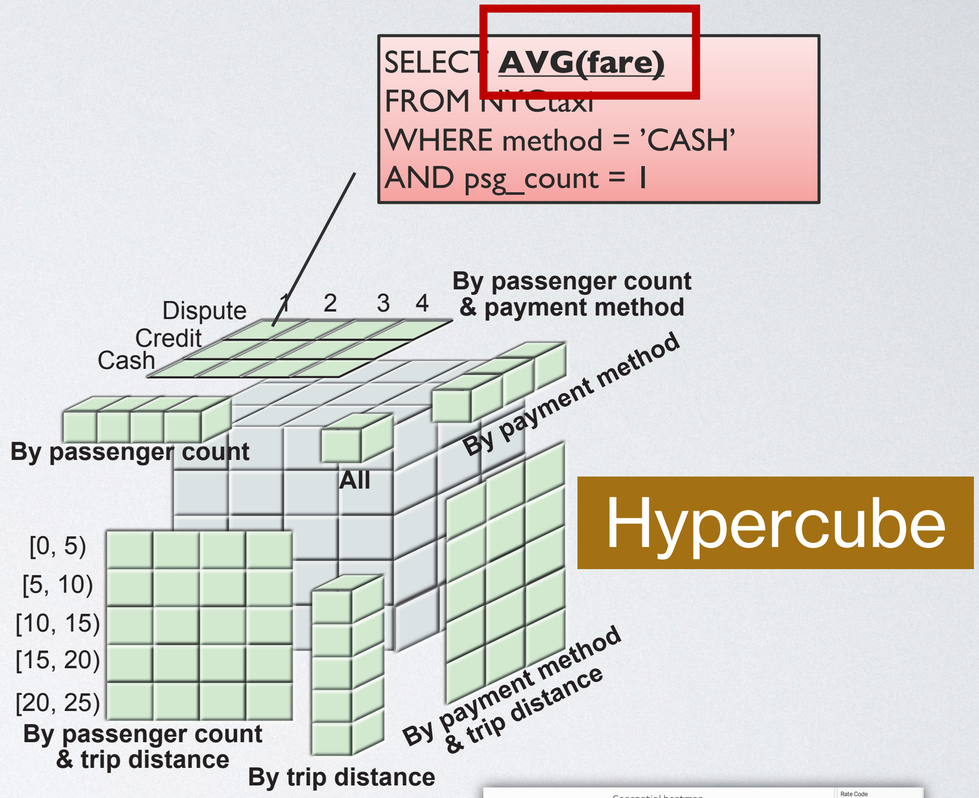


Yongjoo Park, Michael J. Cafarella, Barzan Mozafari: Visualization-aware sampling for very large databases. ICDE 2016: 755-766

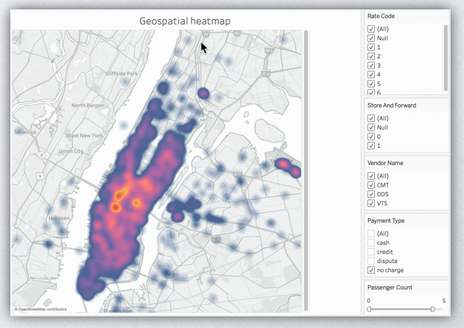
Tao Guo, Kaiyu Feng, Gao Cong, Zhifeng Bao; Efficient Selection of Geospatial Data on Maps for Interactive and Visualized Exploration. SIGMOD Conference 2018: 567-582

Existing stratified sampling tech.

- Carefully create a stratified sample
  - Consider different filter selections
  - Sample+Seek, BlinkDB, SnappyData
- Make the returned aggregates accurate (SUM, AVG, COUNT)
- Example: filter selections -> hypercube (selection space)



Hypercube



Requires a sample per query

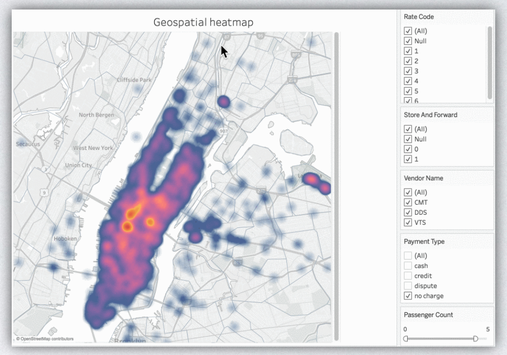
Bolin Ding, Silu Huang, Surajit Chaudhuri, Kaushik Chakrabarti, Chi Wang: Sample + Seek: Approximating Aggregates with Distribution Precision Guarantee. SIGMOD Conference 2016: 679-694

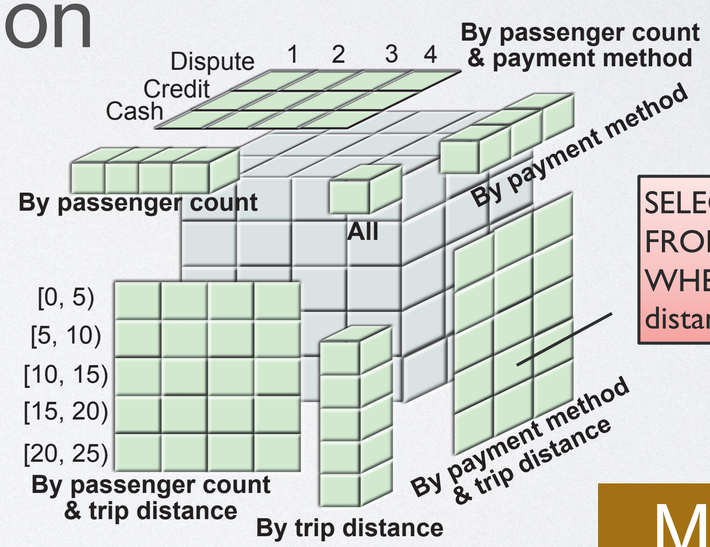
Sameer Agarwal, Barzan Mozafari, Aurojit Panda, Henry Milner, Samuel Madden, Ion Stoica: BlinkDB: queries with bounded errors and bounded response times on very large data. EuroSys 2013: 29-42

Barzan Mozafari, Jags Ramnarayan, Sudhir Menon, Yogesh Mahajan, Soubhik Chakraborty, Hemant Bhanawat, Kishor Bachhav: SnappyData: A Unified Cluster for Streaming, Transactions and Interactice Analytics. CIDR 2017

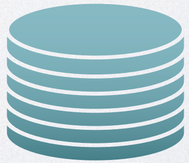
Tabula: sampling middleware system

- Interactive analytics on dashboard
- Local samples for all future queries
  - All cells in the cube
- Return a sample for every interaction
- Never go back to the raw data
- Materialized sampling cube
  - Huge storage overhead
  - Long construction time





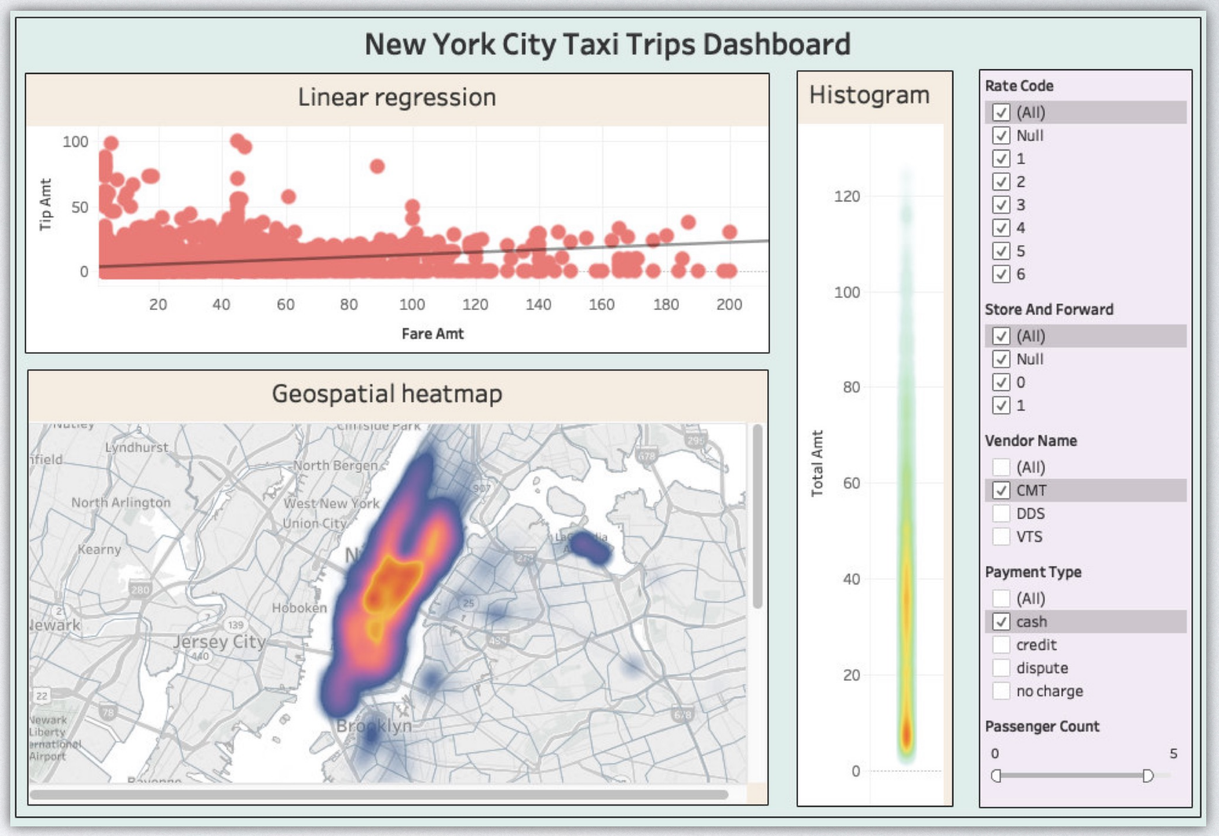

```
SELECT sample
FROM NYCtaxi
WHERE method = 'CASH' AND
distance = [0, 5)
```



Materialized Sampling cube

System design philosophy

- A sampling middleware system
  - Plug and play
  - No change to front-end dashboard
  - No change to underlying data infra.
- Pluggable function for sample quality
  - Domain experts know their needs
  - Support various analytics apps

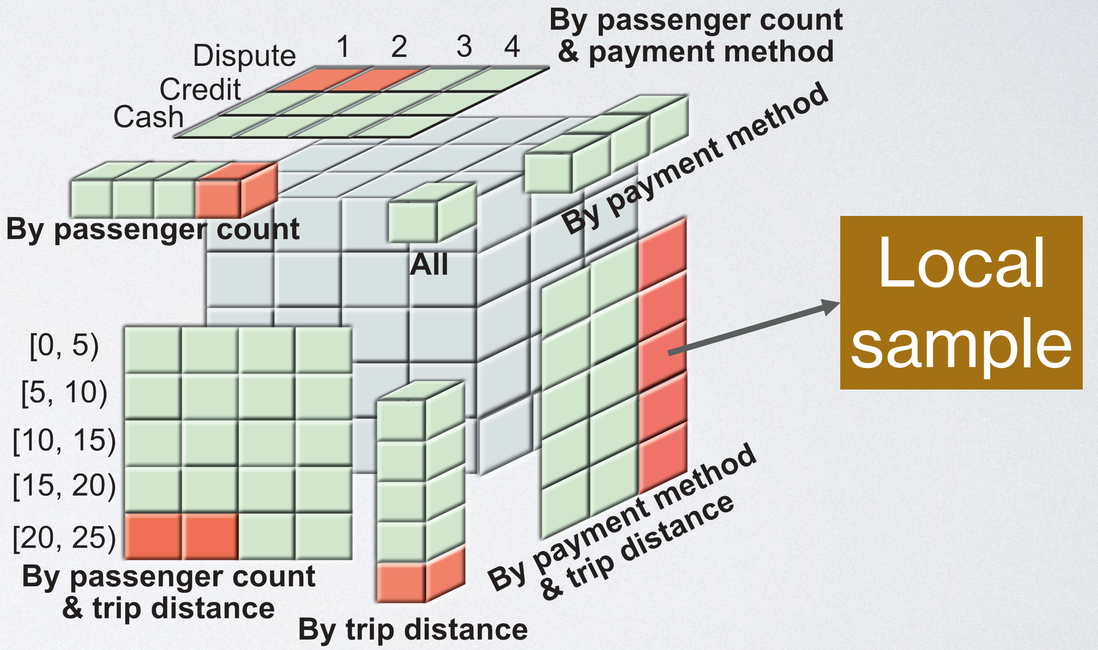


Reduce the storage overhead

Partially materialized sampling cube

- Draw a global sample first
- Use it whenever it is possible
- Only draw local samples for low-accuracy queries (cells)

Global sample



Local sample

Partially materialized sampling cube

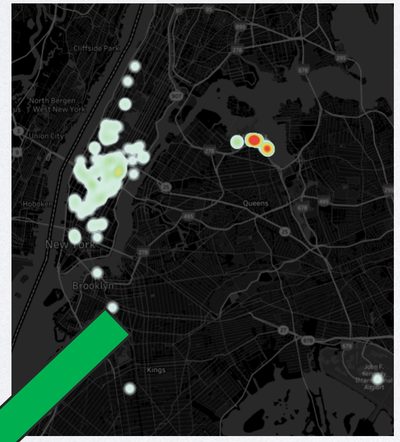
```
CREATE TABLE SamplingCube AS
SELECT D, C, M, SAMPLING(*,θ) AS sample
FROM nyctaxi
GROUPBY CUBE(D, C, M)
HAVING loss(pickup, Sam_global ) > θ
```

Reduce the storage overhead

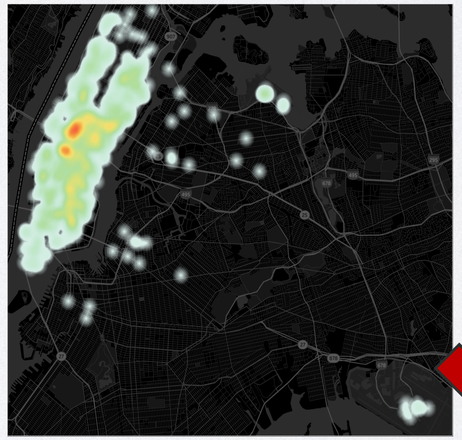
- Low accuracy query
  - If use global sample as the query result, the produced viz will exceed accuracy loss threshold

Raw query result

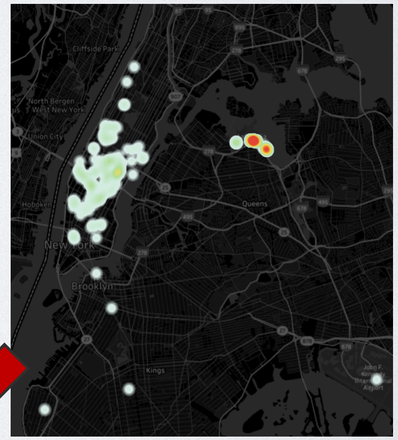
Global sample



Raw query result

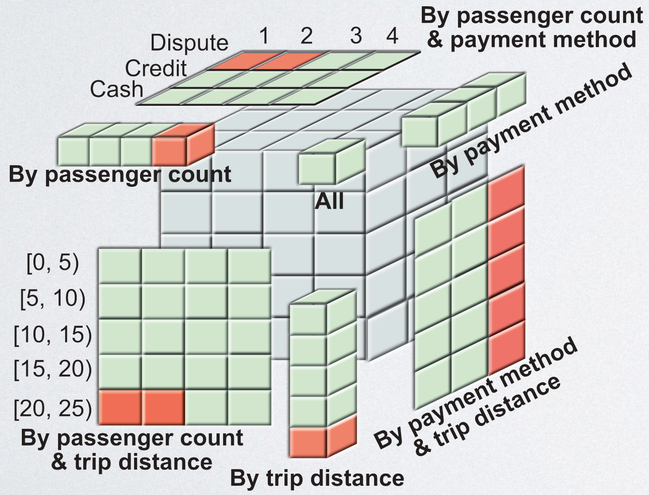


Global sample

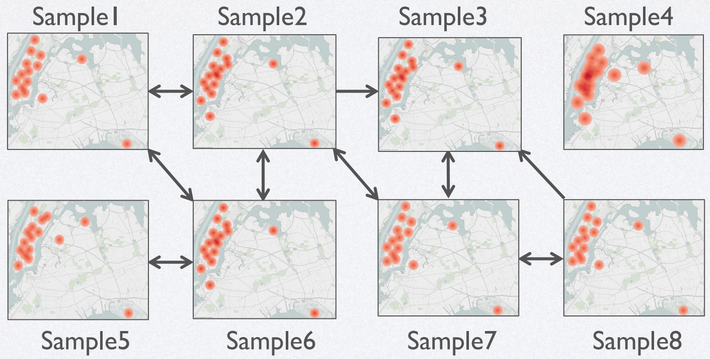


Reduce the storage overhead

- Can we reduce even more?
- Sample selection technique
- Some samples look like each other

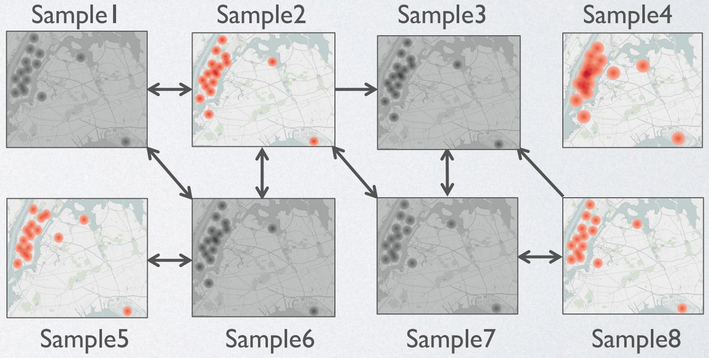


Partially materialized sampling cube



Sample representation graph

2, 8, 5, 4 selected

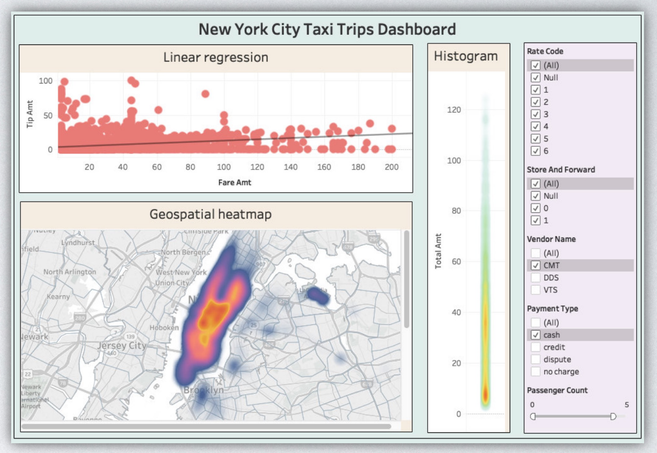


Representatives

NP-Hard

Accuracy loss function

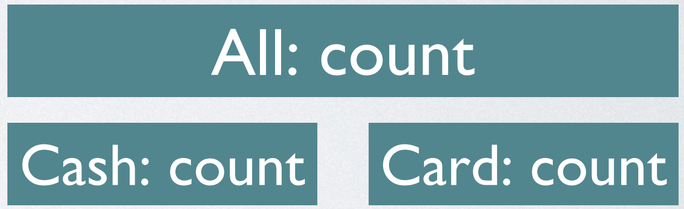
```
CREATE TABLE SamplingCube AS
SELECT D, C, M, SAMPLING(*,θ) AS sample
FROM nyctaxi
GROUPBY CUBE(D, C, M)
HAVING loss(pickup, Sam_global ) > θ
```



- User Defined accuracy loss threshold θ
  - The sample received by the dashboard never exceeds θ
- User Defined accuracy loss function
  - Domain experts know their own needs
  - Fit in different scenarios, heat map, linear regression...

Accuracy loss function

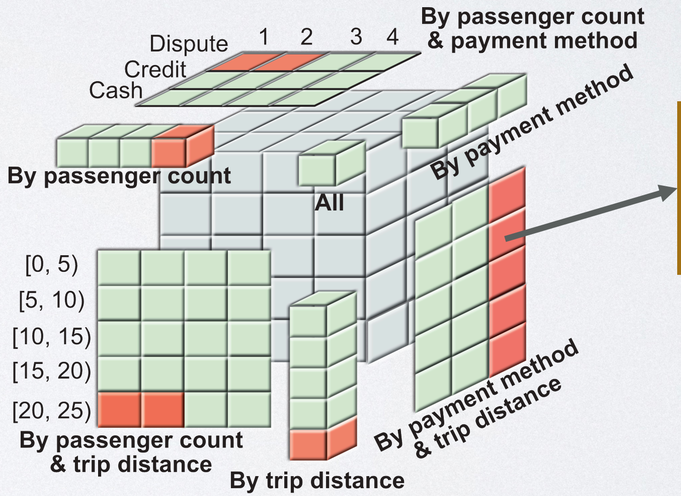
- Algebraic aggregate function
  - The function can be computed based on several functions in its sub-domains
  - Common: Count, Sum, AVG, Min, Max, ...
- Beneficial to the cube initialization



Sampling function

- The function draws the local sample for low-accuracy query
- Generic for diff loss functions
- Produce a sample which has loss < θ

```
CREATE TABLE SamplingCube AS
SELECT D, C, M, SAMPLING(*,θ) AS sample
FROM nyctaxi
GROUPBY CUBE(D, C, M)
HAVING loss(pickup, Sam_global ) > θ
```



Partially materialized sampling cube

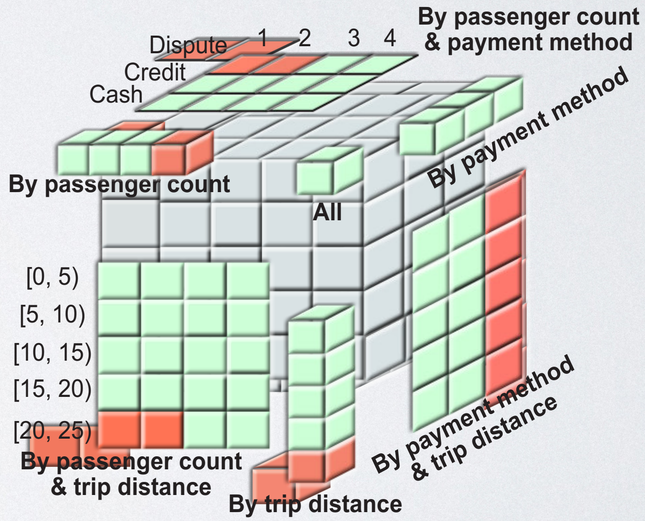
Arizona State University

Concepts are clear and Storage overhead is reduced, but...

Reduce the init construction time

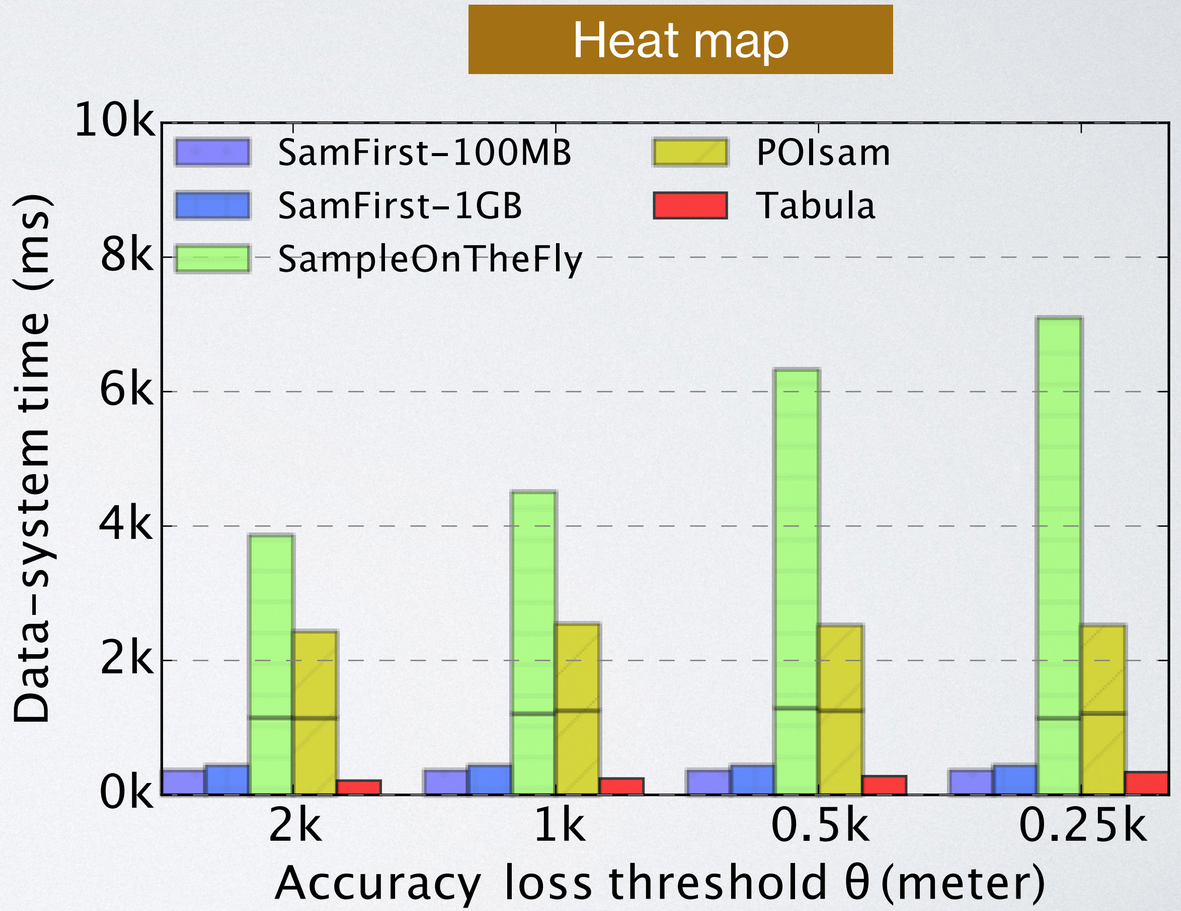
- Naïve construction
  - Exponentially with num attributes
  - 2^n GroupBy, n = num attributes
- Dry-run algorithm
  - Dry run stage: detect the low-accuracy queries
  - Real run stage: only run a few GroupBy if necessary

```
CREATE TABLE SamplingCube AS
SELECT D, C, M, SAMPLING(*,θ) AS sample
FROM nyctaxi
GROUPBY CUBE(D, C, M)
HAVING loss(pickup, Sam_global ) > θ
```



Performance: Execution time

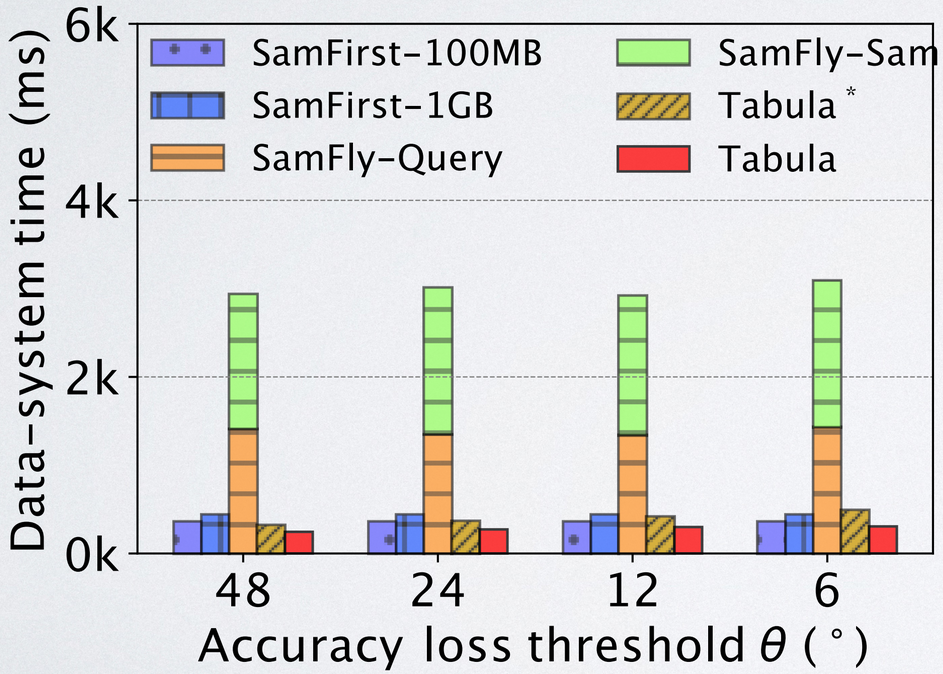
- Heat map: dashboard on Spark
- 200GB NYCtaxi, 5 columns, 17 K queries (cells)
- Sample first, sample on the fly, POIsam
- Tabula: query time = 300 ms, viz time = 400 ms



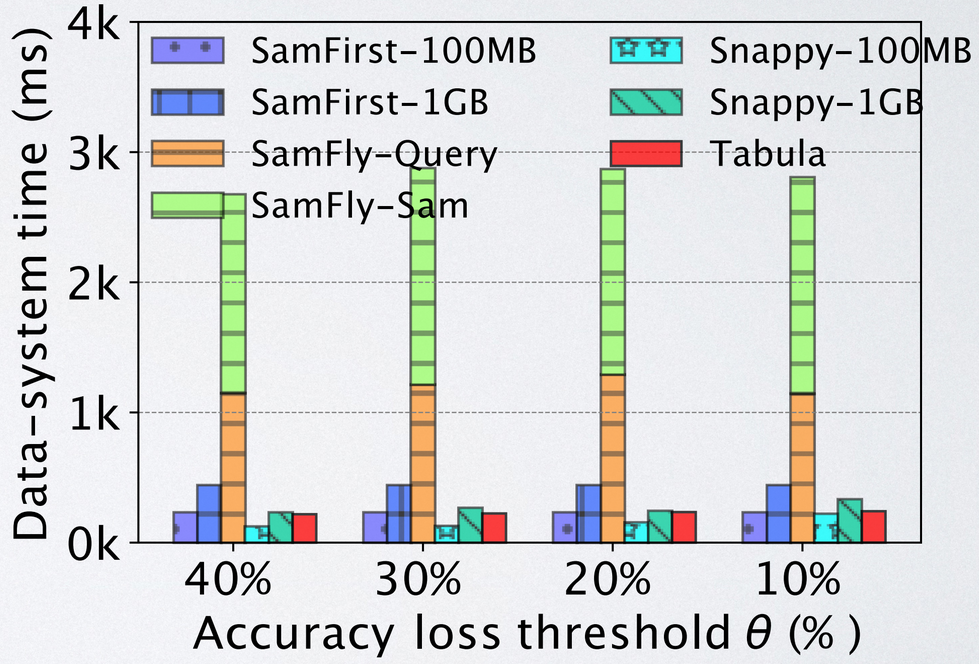
Tabula performance

- Execution time

Linear regression



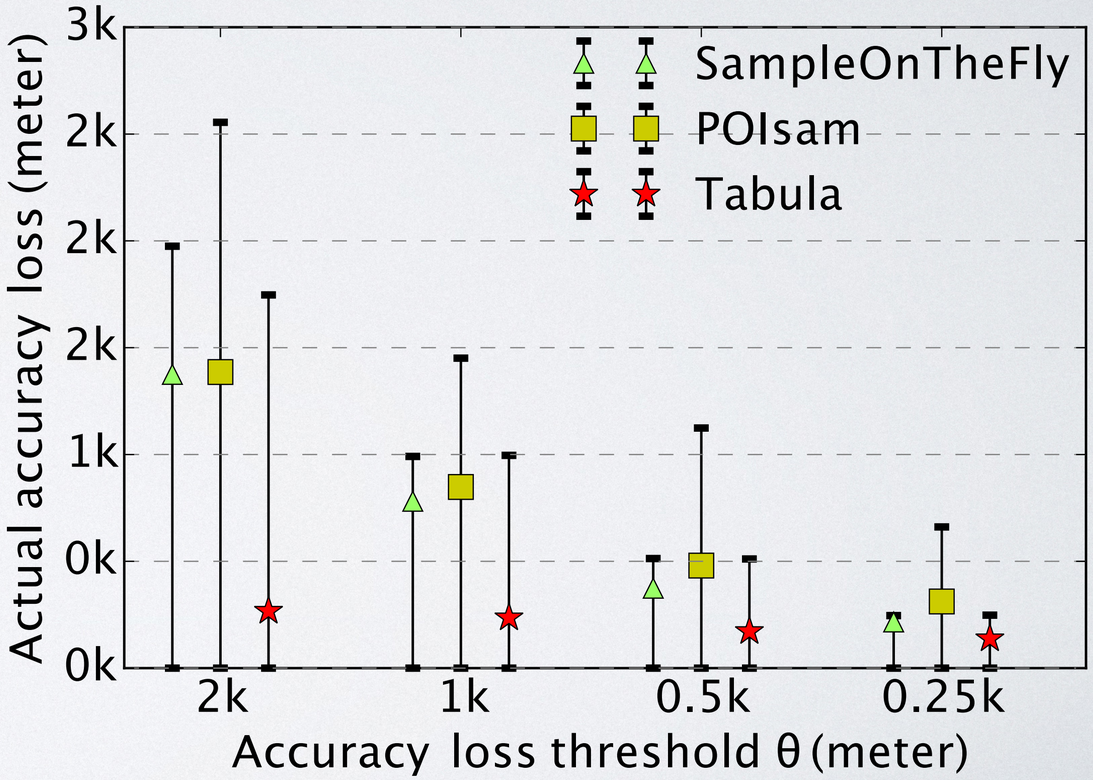
AVG



Performance: Accuracy loss

- SampleFirst: extremely bad, omitted
- Tabula and Sample on the fly guarantee the accuracy loss

Heat map



Take-away

- Tabula: sampling middleware for visualization dashboard
- Interactive performance
- Plug and play solution
- Deterministic accuracy loss guarantee
- User-defined accuracy loss with algebraic property
- Low storage overhead and quick construction