Origin-Destination Matrix Using Mobile Network Data with Spark

Dr. Javiera Guedes
Teralytics

Teralytics

We transform raw, human mobile activity data into valuable insights

180 Billion

events processed in *(near)* real-time every day covering location, demographics, web traffic, spending etc.

Offices in
New York, Zurich and Singapore
with clients across four continents

We Provide Insights into Human Behavior

Location Data

Human behavior data obtained from the mobile network

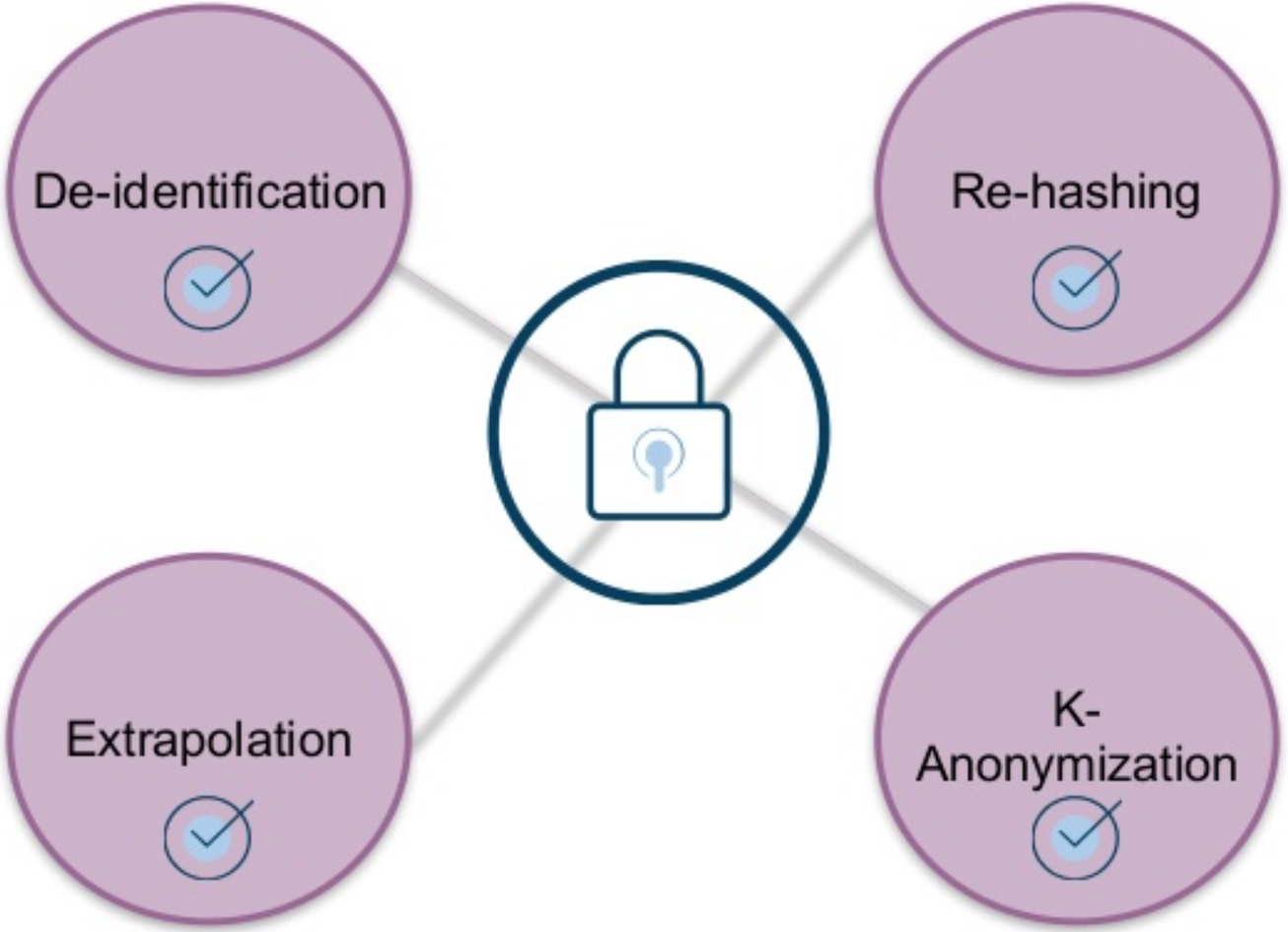
Powerful analytics

Anonymization, aggregation and analytics based on state-of-the-art algorithms in a secure environment

Unprecedented insights

Real-time and historic view of actionable insights presented on a beautiful web dashboard

Data Privacy



Problem Overview

OD Matrix

In transportation processes, it is important to understand how many people travel between zones. This interchange is organized in an origin-destination table or matrix.

One approach to *statically* computing this matrix is:

$$T_{ij} = T_i \frac{A_j f(C_{ij}) K_{ij}}{\sum_{j=1}^{n} A_j f(C_{ij}) K_{ij}}$$

where:

- T_{ij} : trips from i to j.
- T_i : trips from i, as per our generation analysis
- A_j : trips attracted to j, as per our generation analysis
- $f(C_{ij})$: travel cost friction factor, say = C_{ij}^b
- K_{ij} : Calibration parameter

These factors can be estimated from mobility and census data, surveys, ground truth collection data extrapolation, etc.

With telco data we can do this *dynamically,* over time, and including mode of transport to answer major transportation questions.

OD Matrix Germany

1 Matrix per
- Time bucket (15 min, hour, day)
- Mode of transport (flight, car, train)

8199 zip code level 5 geometries

Region	1	2	3
1	120	12	58
2	12	80	95
3	56	90	34

40 days, 100 Billion events

Zip code level 5 geometries, Germany

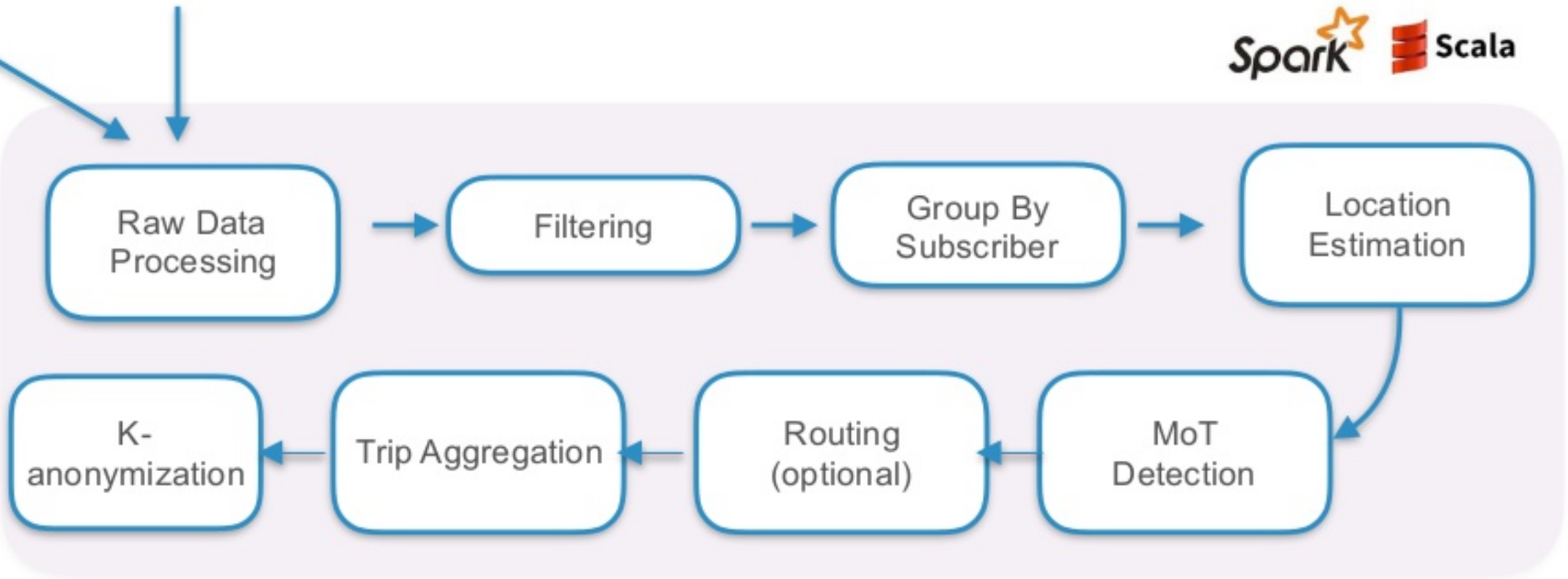
Approach

Approach

Telco Data
HDFS

Network Info
PostgreSQL

USA: 25 Billion events / day
Germany: 3 Billion events / day



Event Generation

Cell phone tower

Coverage Areas

Location Estimation

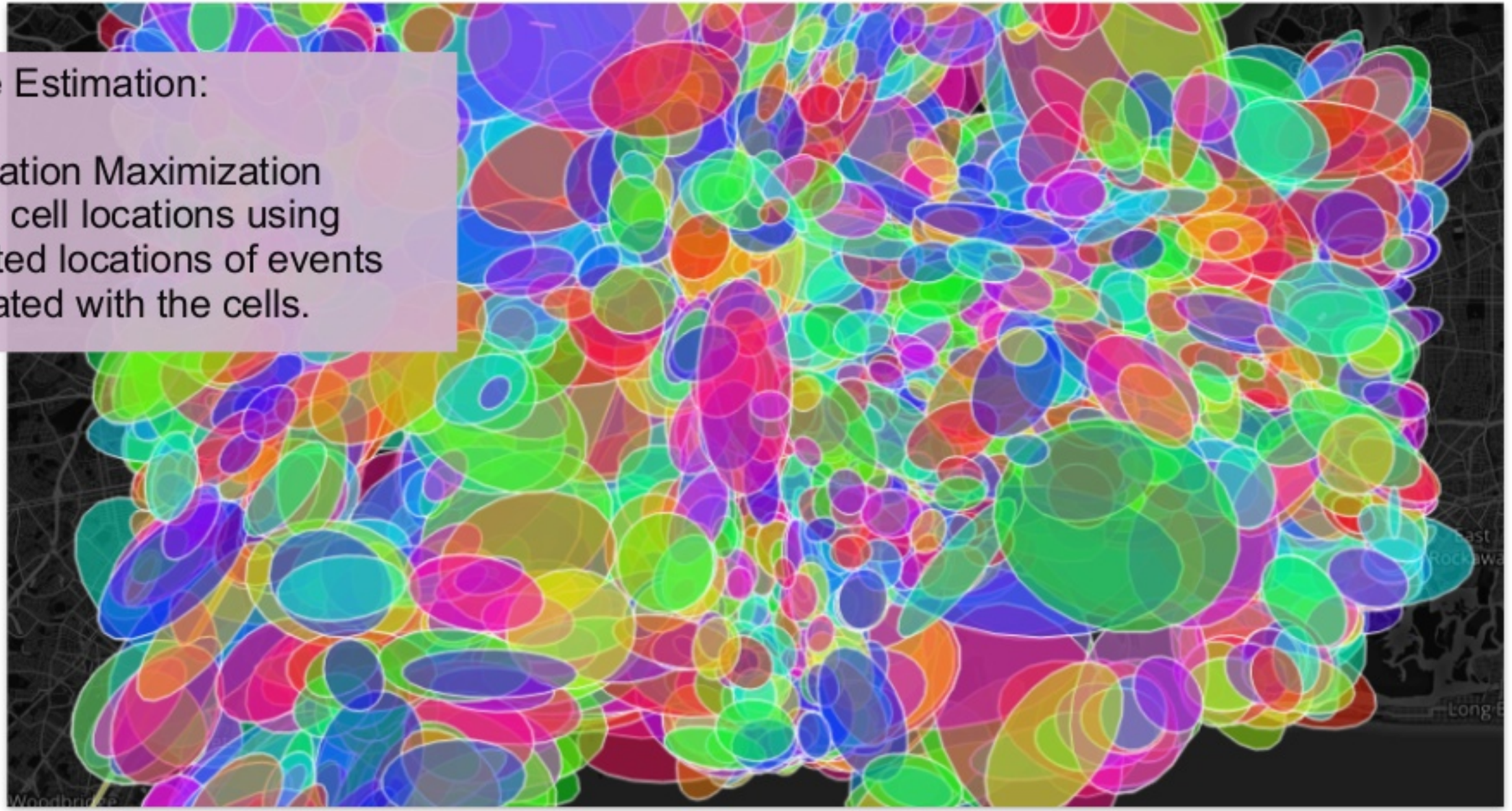
Trip Estimation

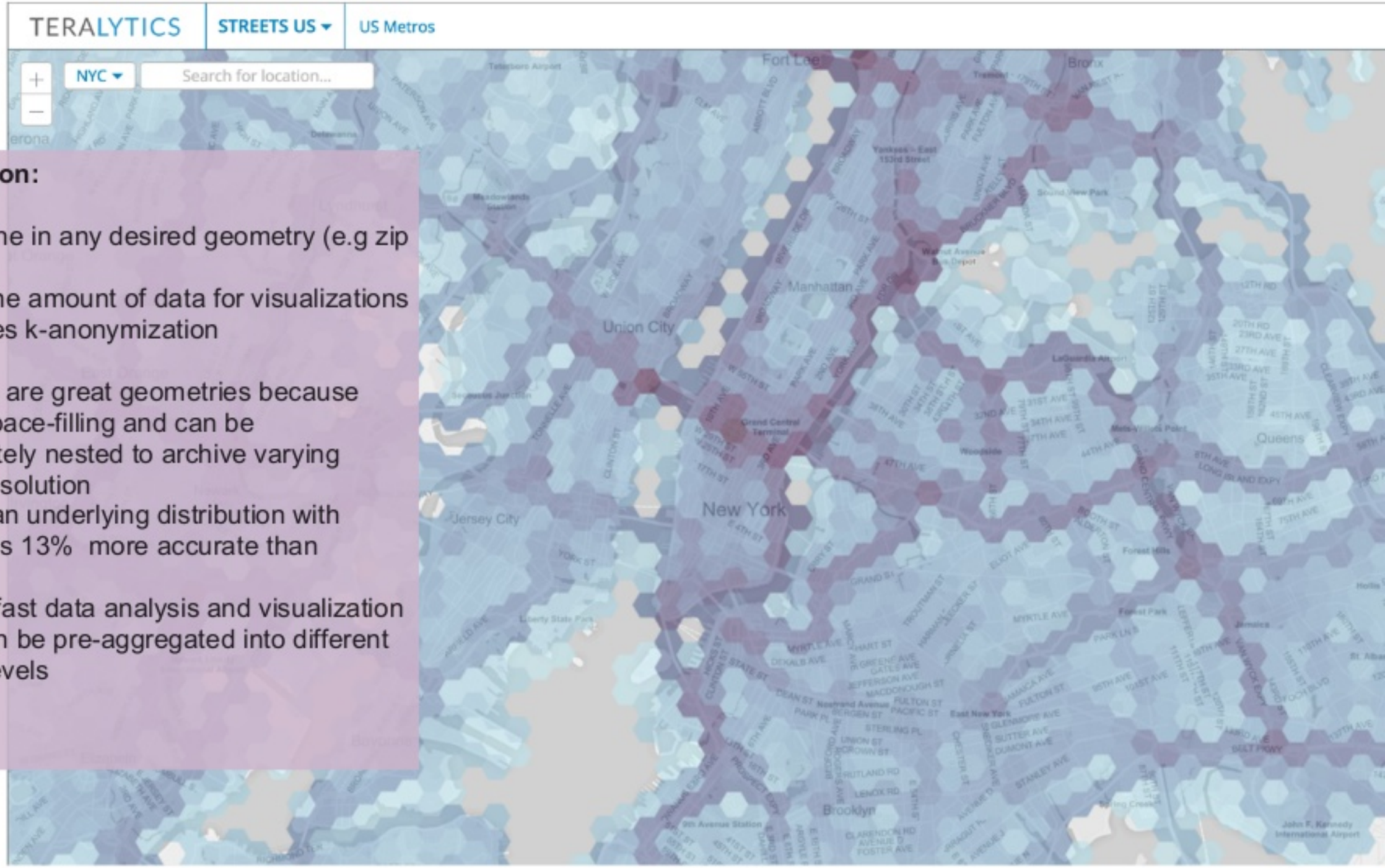
Mode of Transport Estimation

Coverage Area Estimation

Coverage Estimation:

- Expectation Maximization around cell locations using estimated locations of events associated with the cells.

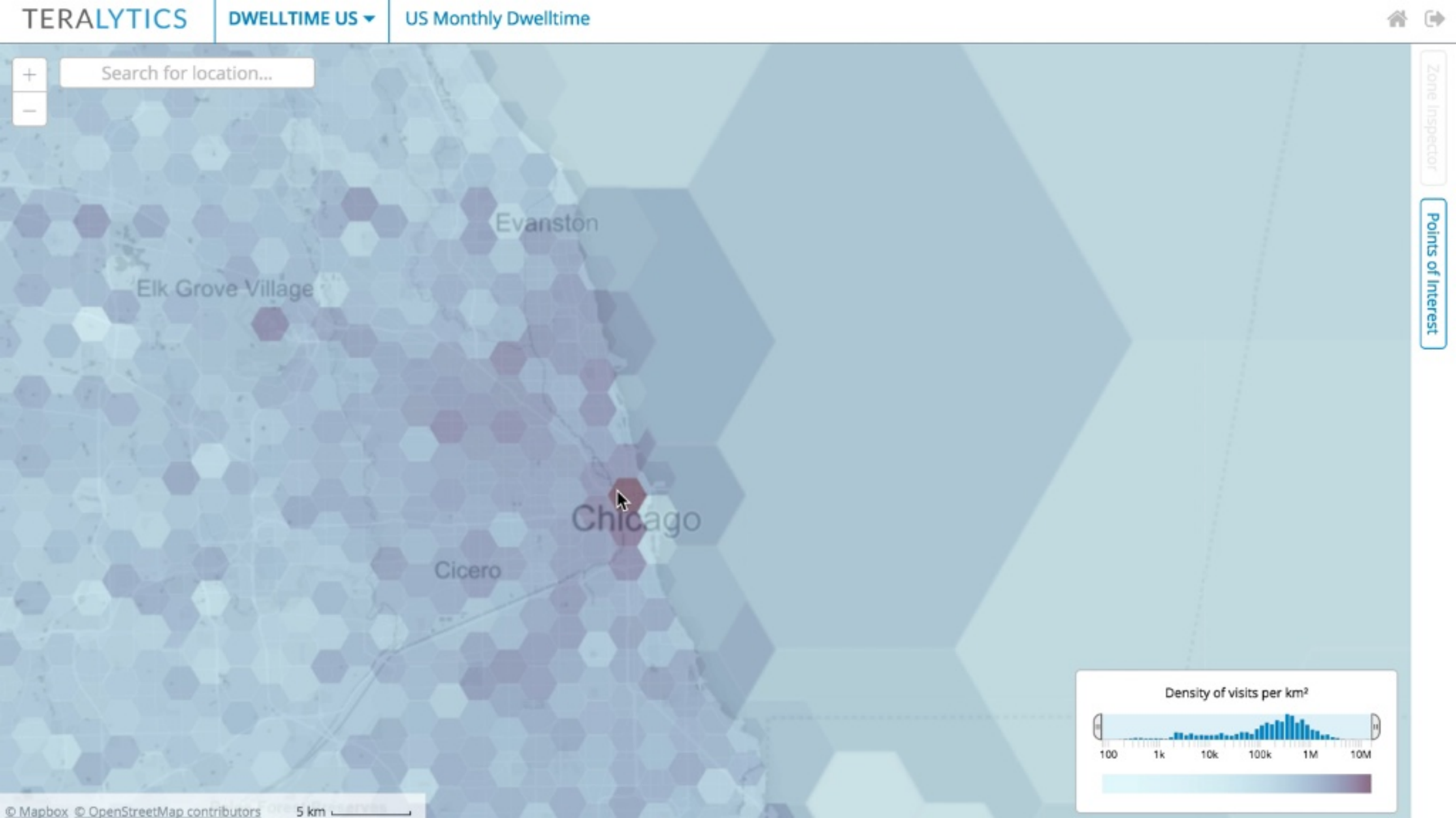




Aggregation:

- Can be done in any desired geometry (e.g zip code)
- Reduces the amount of data for visualizations and ensures k-anonymization

- **Hexagons** are great geometries because they are space-filling and can be approximately nested to archive varying levels of resolution
- Sampling an underlying distribution with hexagons is 13% more accurate than squares
- Allows for fast data analysis and visualization as data can be pre-aggregated into different hexagon levels

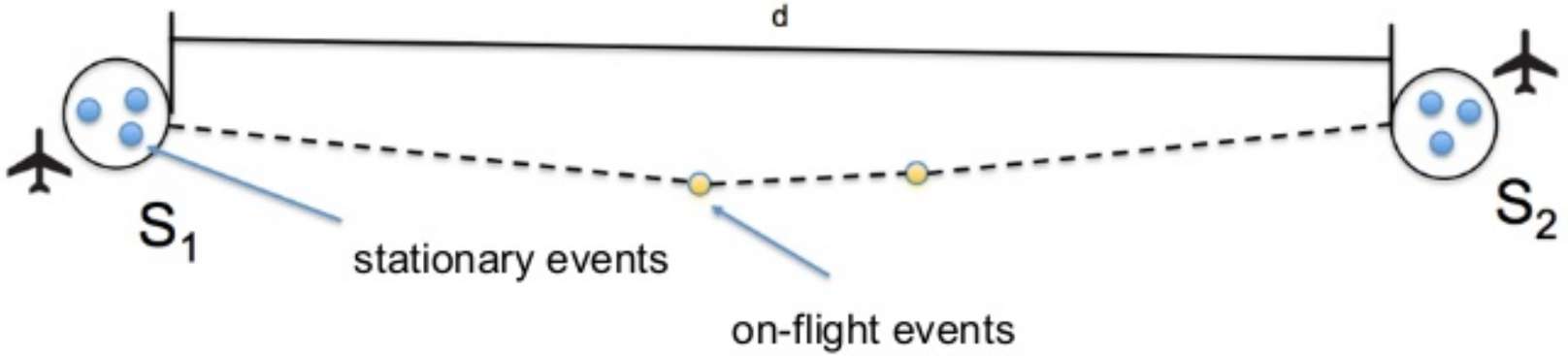

TERALYTICS
DWELLTIME US
US Monthly Dwelltime
Search for location...
Zone Inspector
Points of Interest
Evanston
Elk Grove Village
Chicago
Cicero
Density of visits per km²
100
1k
10k
100k
1M
10M
© Mapbox © OpenStreetMap contributors
5 km

Mode of Transport Estimation

Detecting the mode of transportation of a trip is a classification problem. Mostly, we want to assign trips to major mode of transportations such as flight, train, or car.

Flights:

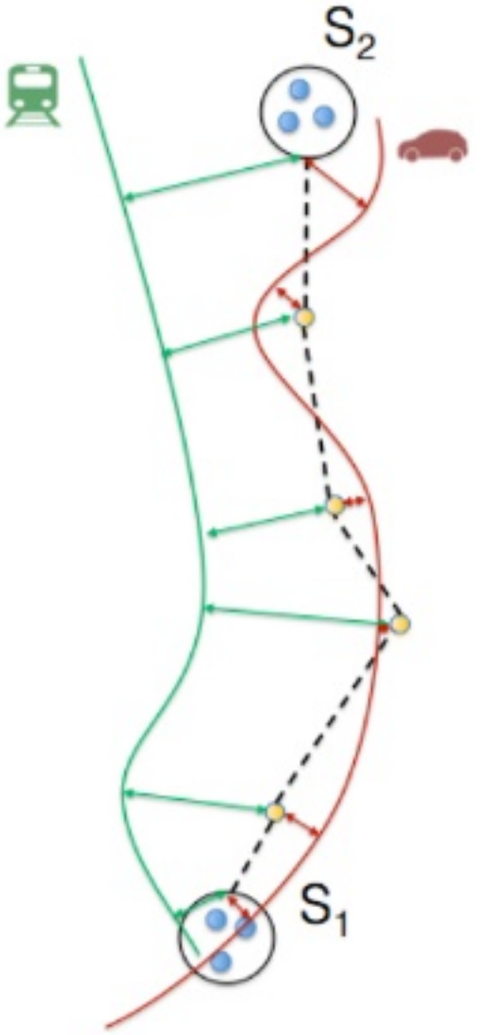
- Stationary at origin and destination
- Nearby Airports
- Speed constrains
- Feasible flight routes



Traces that contain no flights are further classified as land-based

Mode of Transport Estimation

Detecting the mode of transportation of a trip is a classification problem. Mostly, we want to assign trips to major mode of transportations such as flight, train, or car.



Train / Car:

- Features derived from e.g. distance to railways or highways

  $p_{car} / p_{train} > t$ MoT = Car

  $p_{train} / p_{car} > t$ MoT = Train

- Routing

PRODUCT

Departure times | Arrival times

Monday, September 7, 2015

5,814,636

Mo 07 Tu 08 We 09 Th 10 Fr 11 Sa 12 Su 13

5,814,636 people movements in total

Mode of transport

Number of movements by mode of transport.

Train 514,863

Plane 2,394

Private 5,297,377

Time of day

Number of movements by time of day.

00:00 02:00 04:00 06:00 08:00 10:00 12:00 14:00 16:00 18:00 20:00 22:00 24:00

Distance

Top Flows

Top Origins

OD Matrix Product:

- Departure time from origin
- Arrival time at destination
- Mode of Transport

- Aggregated by desired region (state, city, zip code, etc.)
- Bucketed in any desired time (15 min, 1 hr, 1 week, etc.)

Filter flows by magnitude

10 100 1,000 10,000

99% of movements shown

OD Matrix Germany

Goal:

- Compute a matrix between all zip code level 5 areas in Germany (8199 PLZ5)²
- Bucketed by 15 min, hour, day, or week
- Split by mode of transport: flight, car, train

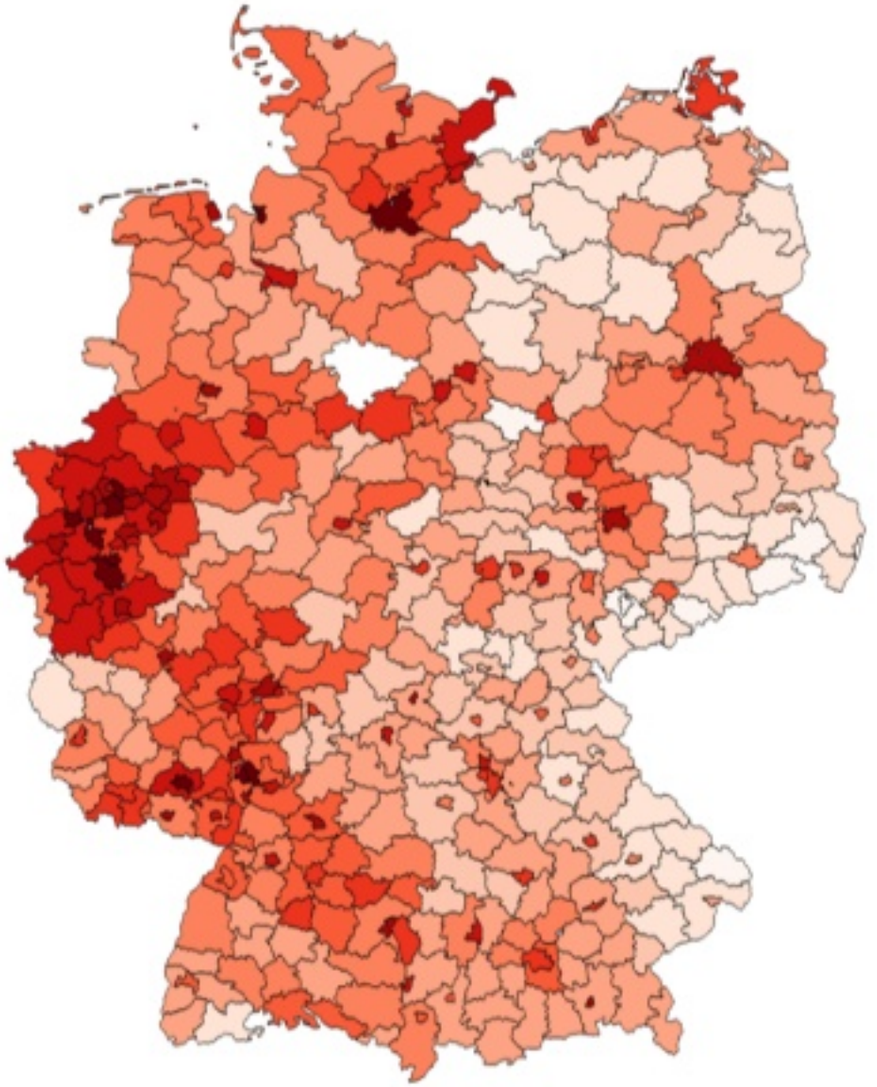
Data:

- 40 days
- 100 Billion events

Zip code level 5 geometries, Germany

Extrapolation

The extrapolation factors for each entry of the OD matrix are computed using the market share of the origin and destination regions.



Estimated local market share based on telco and census data in Germany
0 < MS < 1

TERALYTICS | DEMO - TRANSPORT GERMANY (ZIP CODES) | OD - DE - Zip Level 2 | OD Matrix Berlin - Zip Level 5

OD Matrix Germany

- Telco data is intrinsically dynamic, patterns can be seen varying over time, or even over the course of day.
- Data accompanied by visualizations that allow customers to dice and slice the data for specific regions

21,926,008 people movements in total

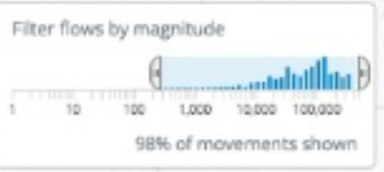
Distance

Top Flows

Top Origins

Top Destinations

Matrix is k-anonymized, i.e. all flow with less than five commuters is removed.



Validation

Criteria for Validation

We validate out results using several criteria and data sets:

- MoT validation using limited calibration data labeled to measure performance.

	precision	recall
flights	0.96	0.81
car	0.96	0.88
train	0.95	0.91

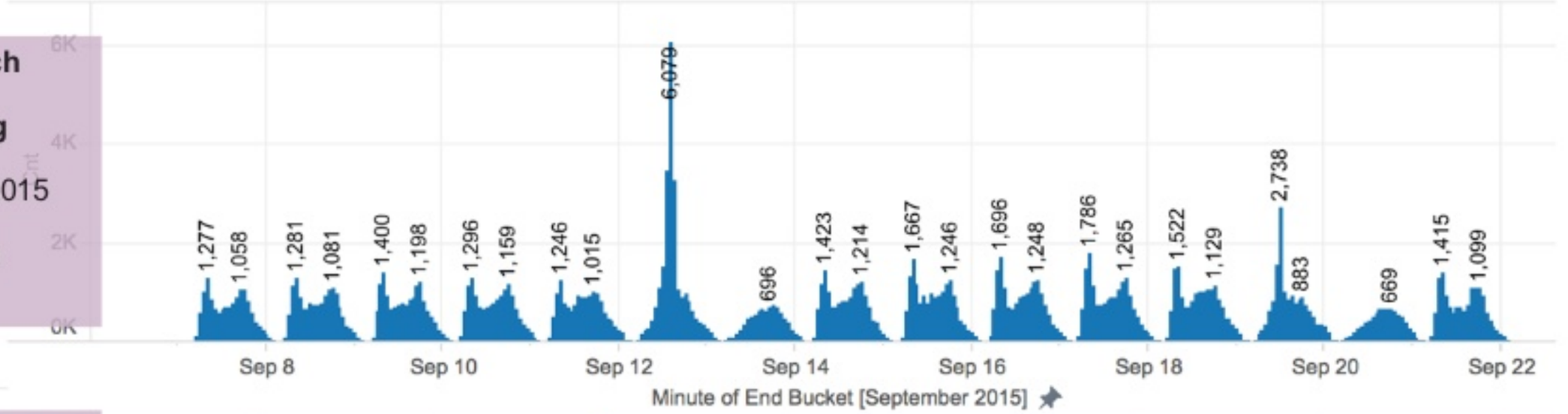
- Limited ground truth for train and flight trips from our providers and customers
- Common sense checks, e.g. whether counts are balanced:

$$OD[i,j] \sim OD[j,i]$$

Bundesliga Games

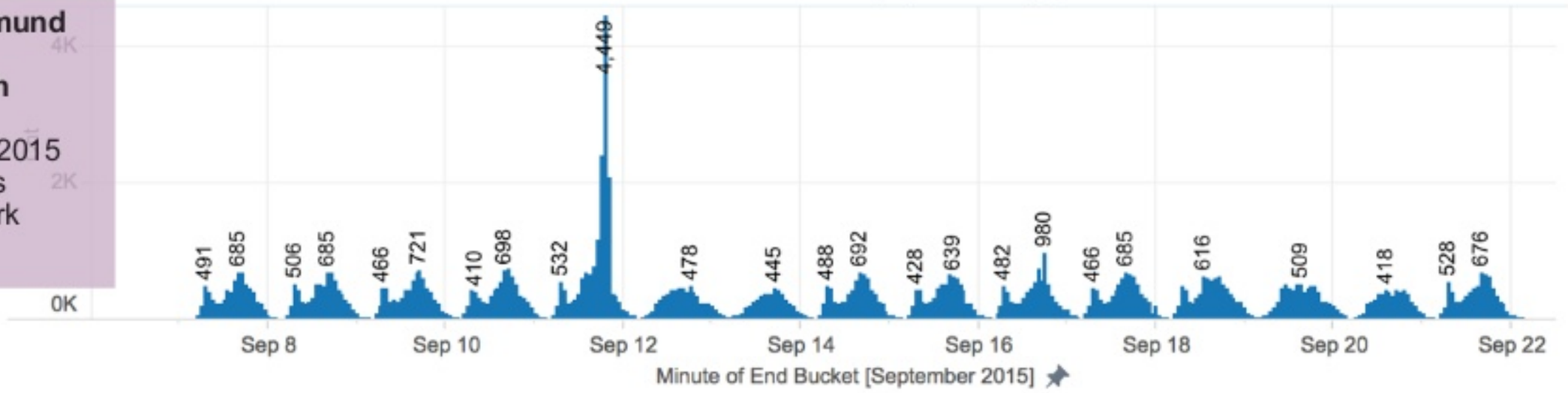
Bayern Munich
vs
FC Augsburg

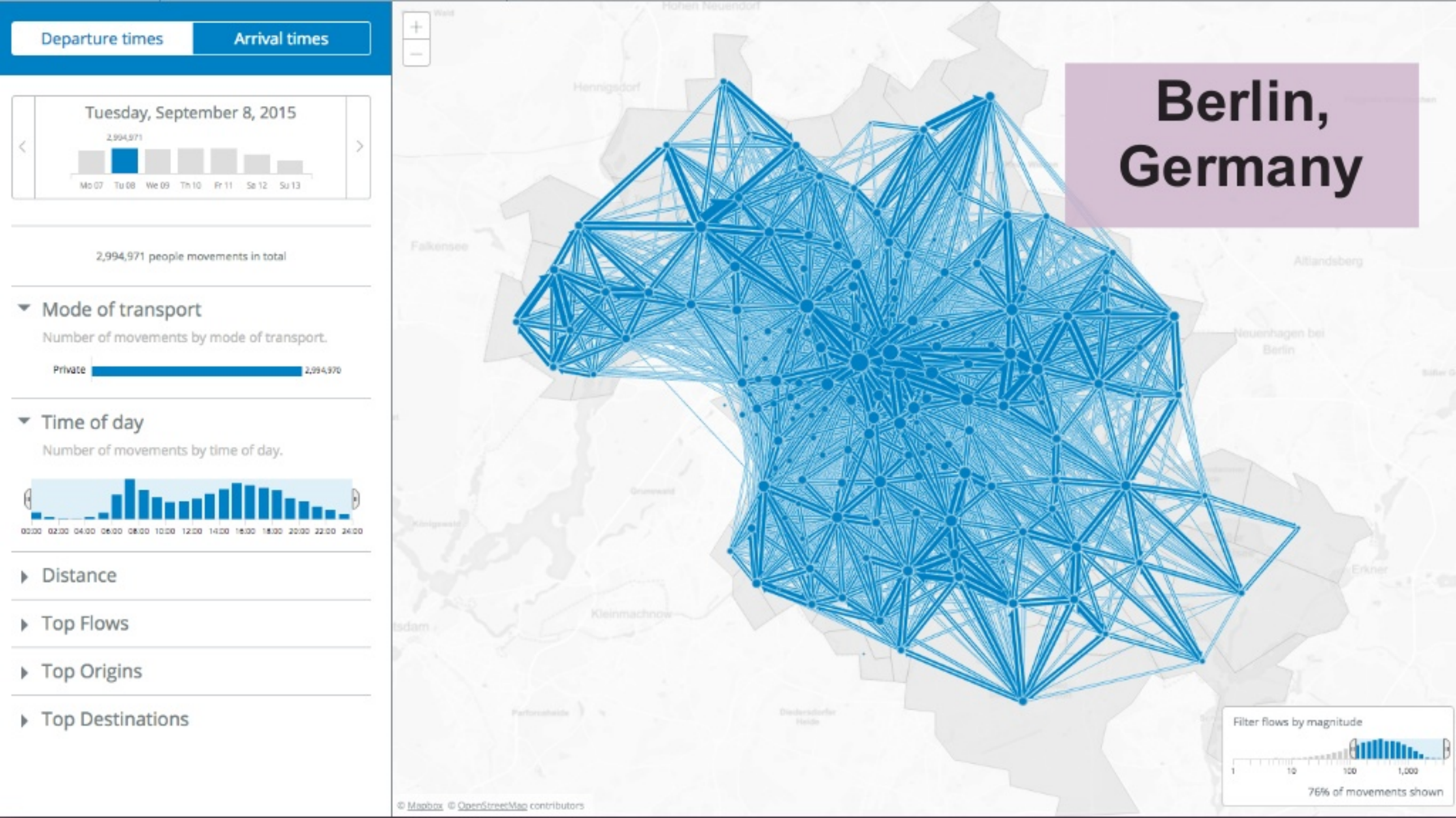
September 12, 2015
75,000 fans
Allianz Arena



Borussia Dortmund
vs
M'gladbach

September 11, 2015
54,000 fans
Borussia Park




Departure times
Arrival times
Tuesday, September 8, 2015
2,994,971
Mo 07
Tu 08
We 09
Th 10
Fr 11
Sa 12
Su 13
2,994,971 people movements in total
Mode of transport
Number of movements by mode of transport.
Private
2,994,970
Time of day
Number of movements by time of day.
02:00
04:00
06:00
08:00
10:00
12:00
14:00
16:00
18:00
20:00
22:00
24:00
Distance
Top Flows
Top Origins
Top Destinations
Berlin, Germany
Filter flows by magnitude
1
10
100
1,000
76% of movements shown
© Mapbox © OpenStreetMap contributors

Guangzhou, China

Departure times | Arrival times

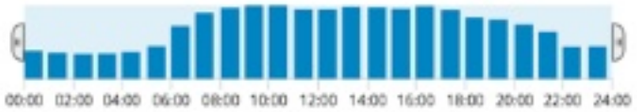
Tuesday, January 26, 2016

47,063,431 people movements in total

Mode of transport

Time of day

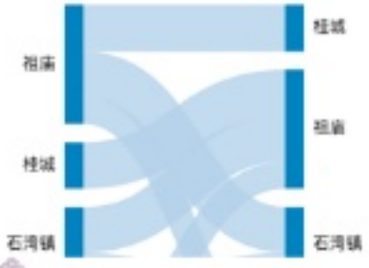
Number of movements by time of day.



Distance

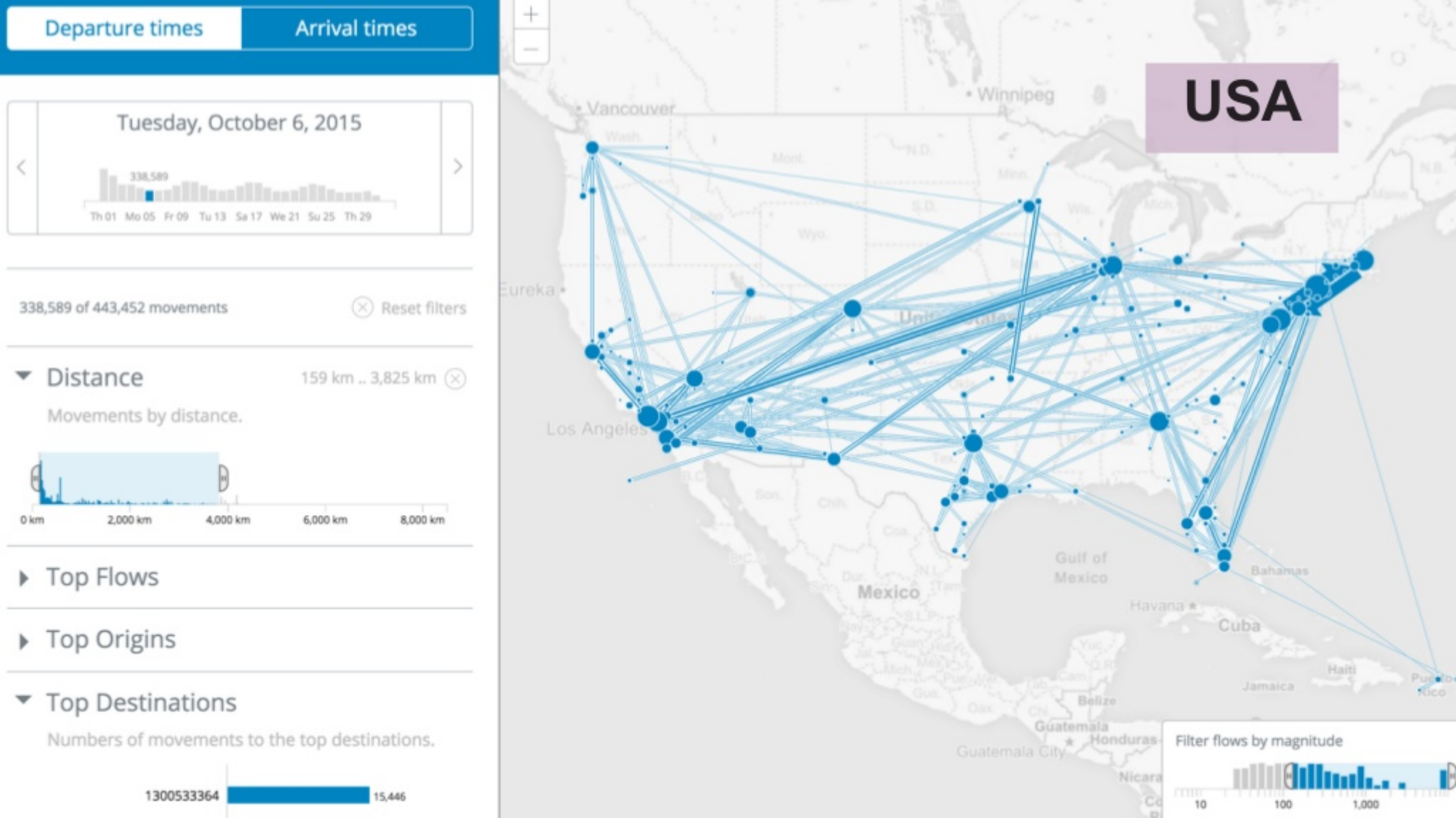
Top Flows

The flows with the most movements.



Filter flows by magnitude

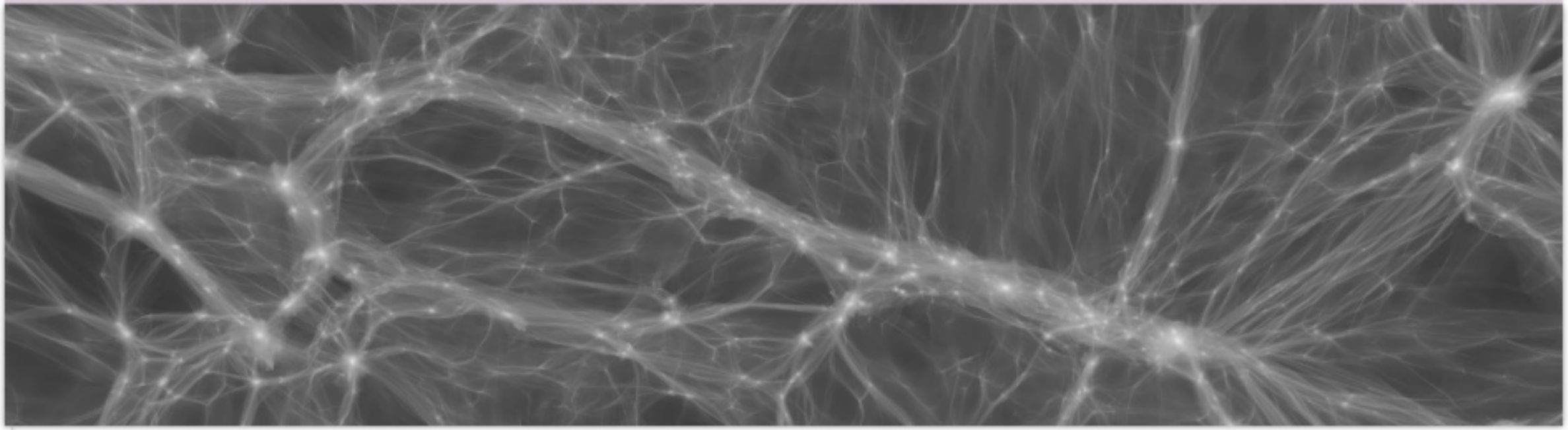
100% of movements shown


Departure times
Arrival times
Tuesday, October 6, 2015
338,589
Th 01 Mo 05 Fr 09 Tu 13 Sa 17 We 21 Su 25 Th 29
338,589 of 443,452 movements
Reset filters
Distance
159 km .. 3,825 km
Movements by distance.
0 km
2,000 km
4,000 km
6,000 km
8,000 km
Top Flows
Top Origins
Top Destinations
Numbers of movements to the top destinations.
1300533364
15,446
USA
Vancouver
Winnipeg
Eureka
Los Angeles
Mexico
Gulf of Mexico
Havana
Cuba
Bahamas
Haiti
Jamaica
Belize
Guatemala
Honduras
Guatemala City
Filter flows by magnitude
10
100
1,000

Why Spark

Our production code is written entirely in Scala / Spark:

- Terabyte data sets processed on a single job
- Fast and compatible with Hadoop HDFS and Accumulo
- The domain is easy to parallelize since every trip is independent



Dark Matter Streams | Kavli Institute Stanford University

Why Spark

Our production code is written entirely in Scala / Spark:

- Terabyte data sets processed on a single job
- Fast and compatible with Hadoop HDFS and Accumulo
- The domain is easy to parallelize since every trip is independent
- Natural choice of the map / reduce nature of our algorithmic approach

Spark is used for slicing and dicing data in dashboards

Spark was not our top choice for data visualization dashboard, since only small batches of data are queried

THANK YOU.

javiera.guedes@teralytics.ch
http://www.teralytics.net