Highly Scalable Tile-Based Visualization for Exploratory Data Analysis

Strata NY: Hadoop and Beyond, 10/17/2014

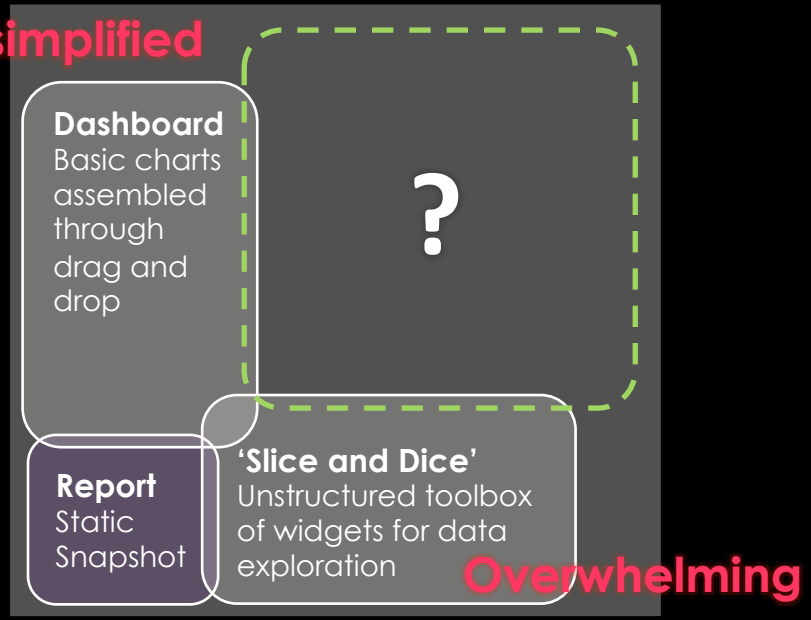
David Jonker, Rob Harper

oculus

Making Sense of Big Data

> "Big Data for us is in complexity. We have complex data. We have big complexity."
> – Analyst

> "No one wants more data, everyone wants better data."
> - Analyst



Less Time Required to acquire insights

More Information Available To make an effective decision

Challenges of Effective Visual Analytics

ESSENTIAL

- Richly **informative** and true to reality.
- Answers easily arrived at, **easily understood**.

ALSO, every business problem is unique,
but cannot afford an entirely unique solution for every problem.

- Need relatively universal, **repeatable** technical approaches.

Tile-Based Visualization for Big Data

WEB MAPPING APPS have an established model for intuitively navigating and understanding big data, based on zoomable multiresolution **tiles** and **layers**.

BUT geospatial data is relatively static. How can a tiled, layered approach be applied with **dynamic** data, and non-geospatial problems, at scale?

Aperture Tiles

TILE-BASED VISUAL ANALYTICS

- Hierarchical data tiling using cluster computing.
- Interactive on-demand image tile generation.
- Layers of raw data and derivative analytics.

OPEN SOURCE

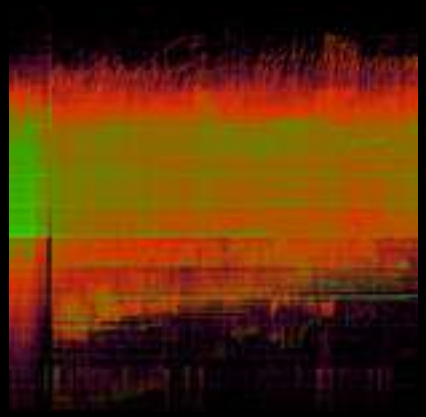
- Oculus research product.
- Built on Apache Spark and Hadoop.
- Preparing for version one product release.

Example Applications

Geospatial

Social Media

Biomedical

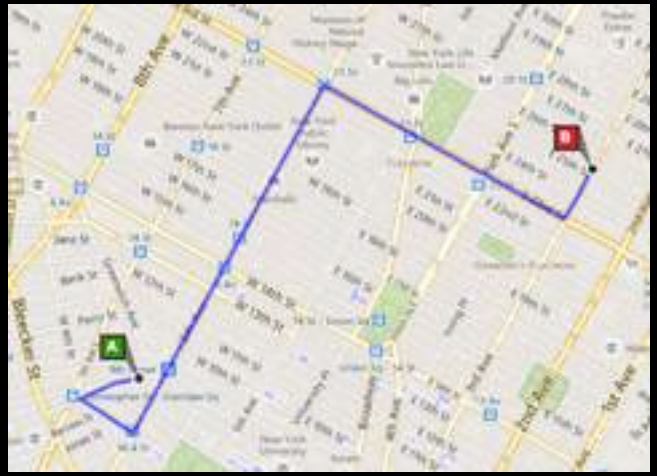


Financial

Photo by Sakeeb Sabakka, CC

NYC Taxi Data

FOIL request by Chris Whong, March 2014



Taxi ID	Hack, Medallion, Vendor
Origin	Location, Time
Dest	Location, Time
Trip	Duration, Distance, # Passengers
Fare $	Base, Tip, Tolls, Taxes, Payment method

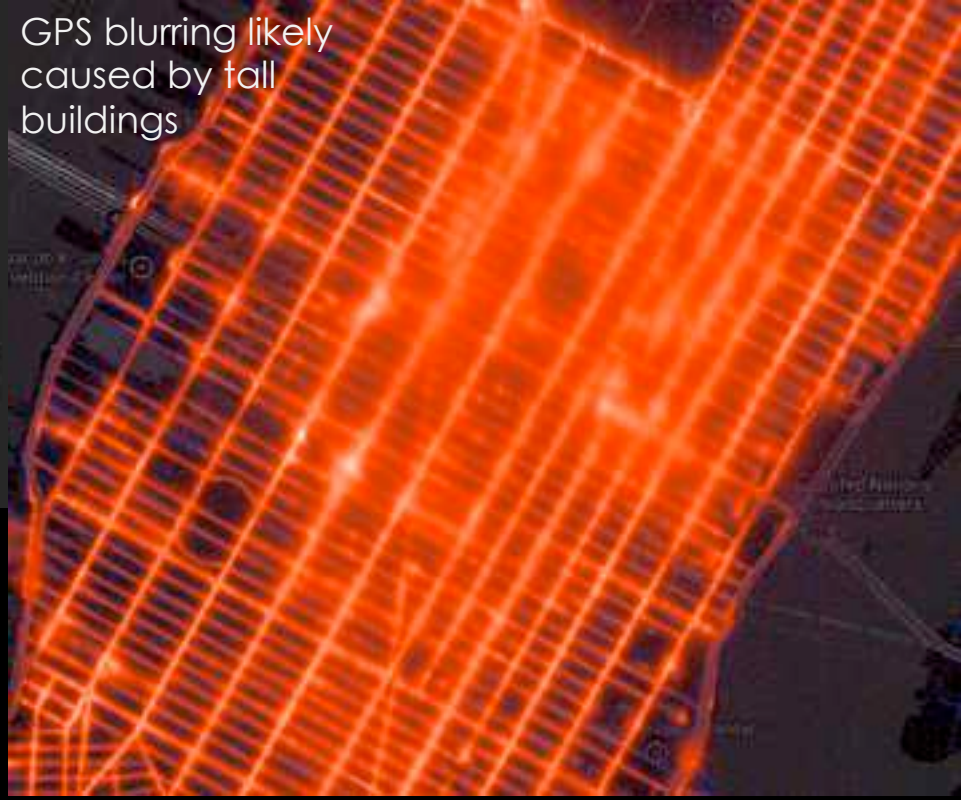
Visualizing NYC Taxi Data

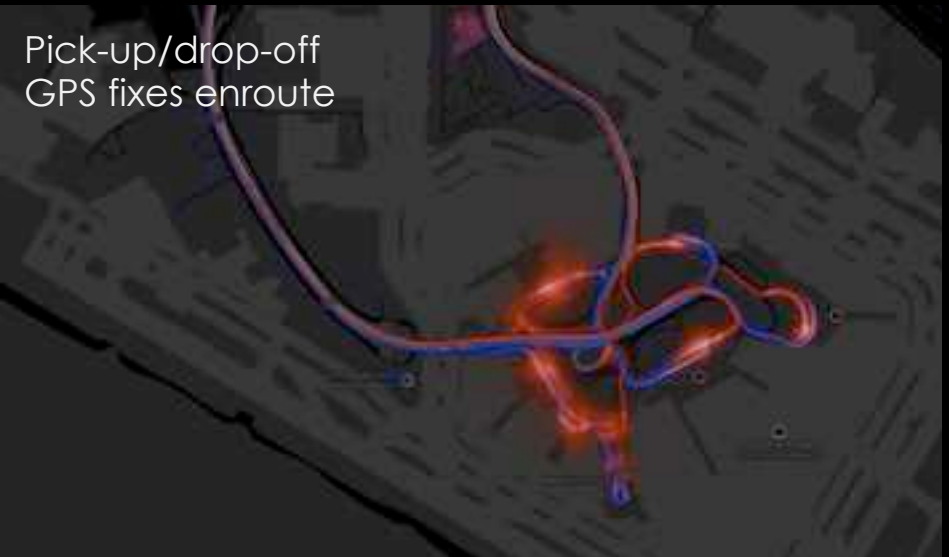
Eric Fischer, MapBox

www.mapbox.com/blog/nyc-taxi/

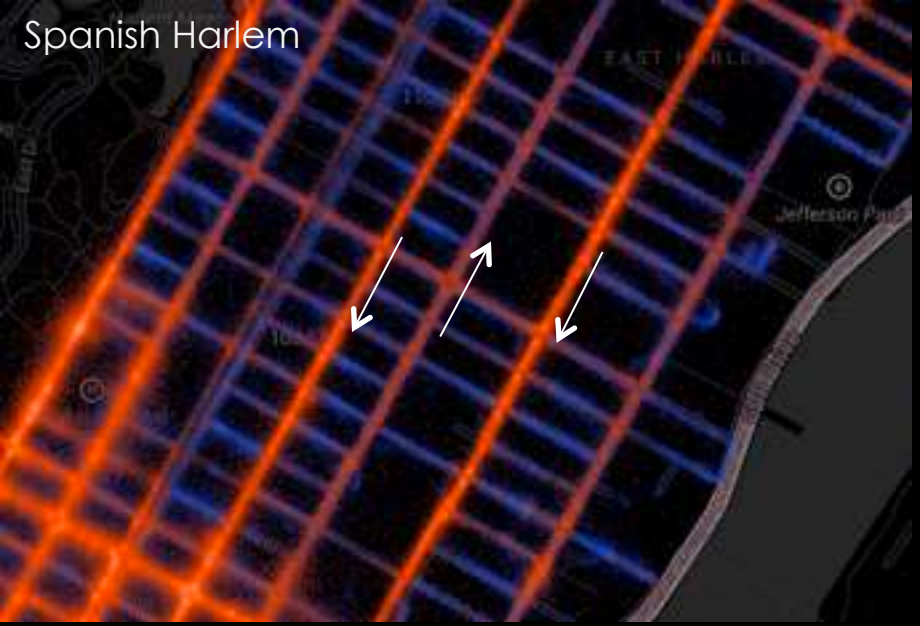
Demo

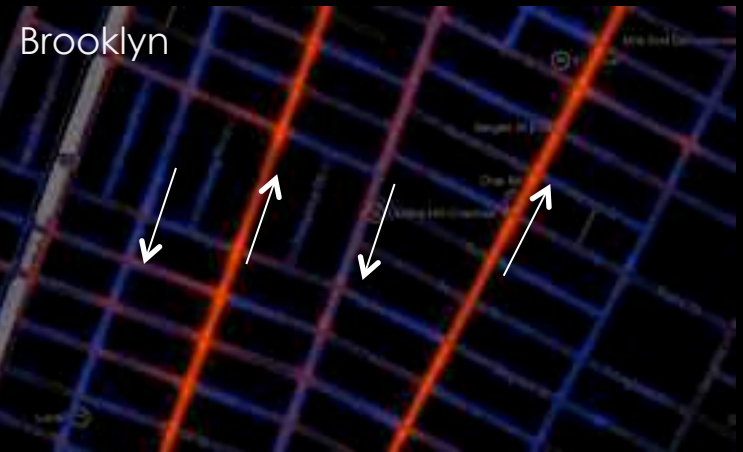
Continue presentation





Visualizing all of the data including "bad data" helps identify source of errors, understand how errors manifest, and how they may affect analysis

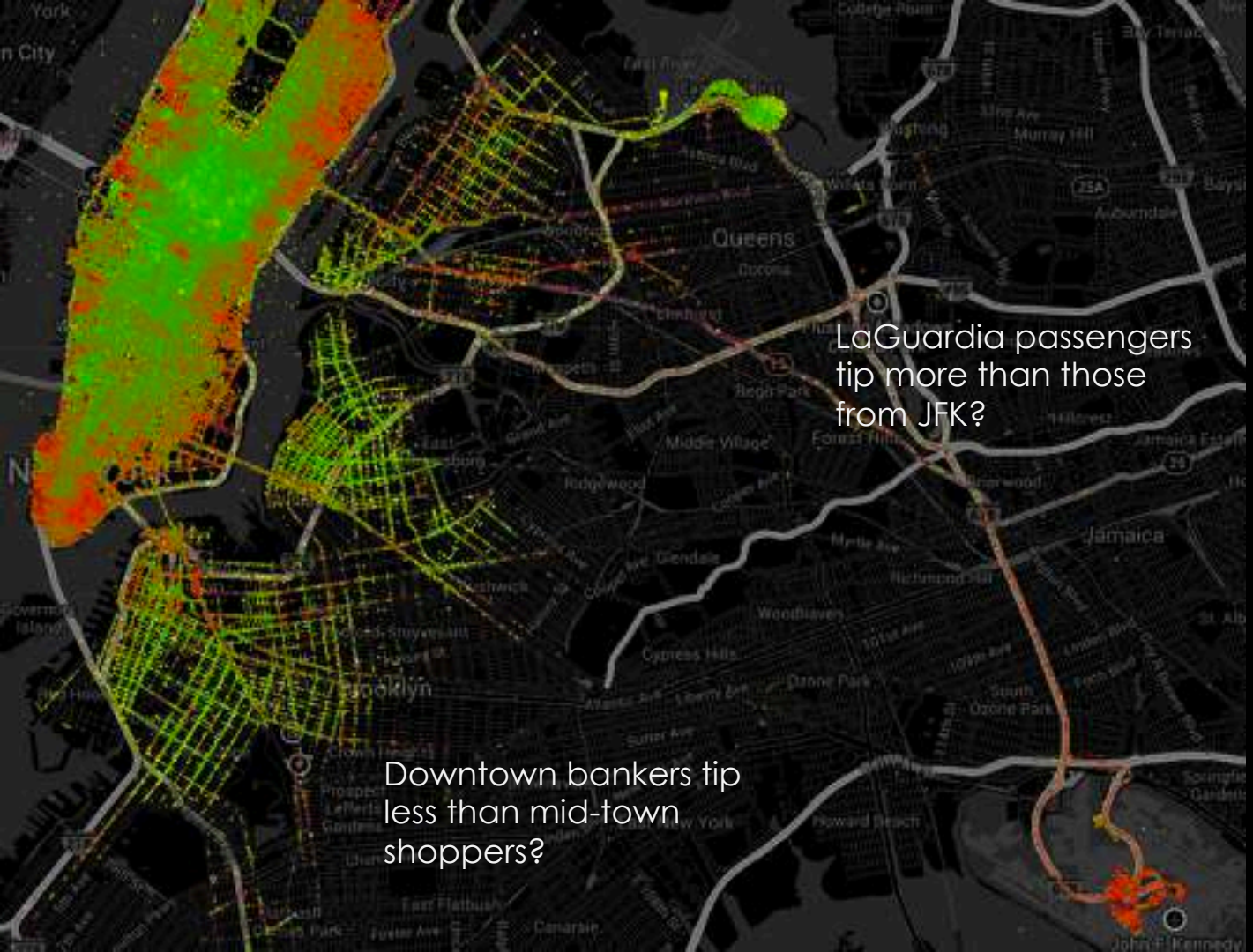




Comparing Pick-up and Drop-off Locations

Red – Pick-up Locations
Blue – Drop-off Locations

Multi-scale visualization allows exploration of macro and micro trends.



Average Tip % by Pick-up Location*

Red	~15%
Green	~20%
Yellow	~25%

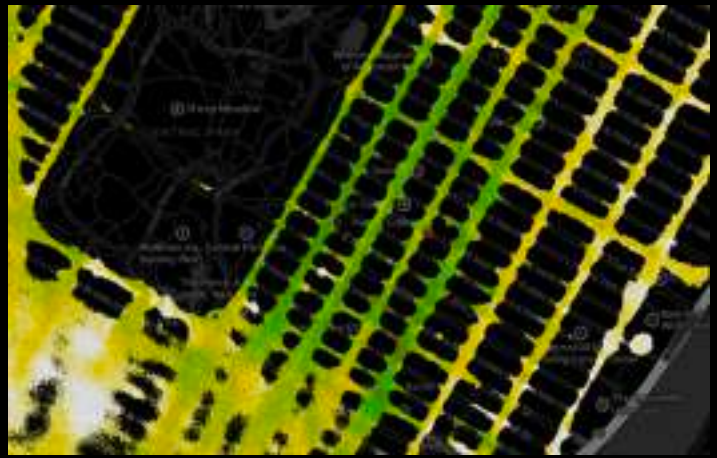
*CC payments only

Using a spectral color ramp helps us compare differences across the data (luminance better for multi-series).

Tiles uses aggregate data to generate rasters, not pre-computed images

Easily switching between fully zoomable multi-scale layers makes the task of understanding the data easier.

Very short ride anomaly in upper east side exposed by range filtering – hospital sending patients on a short ride to clinic?



Distance
Red = 1m
Yellow = 18m

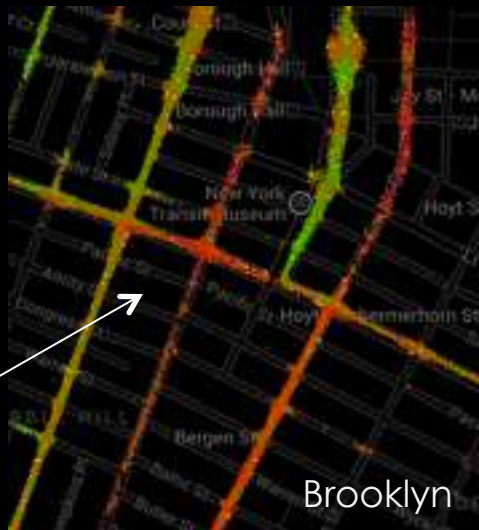
Average $/hr by Pick-up Location

Red	$30/hr
Yellow	$75/hr

Calculated using:
(fare + tip) / (duration of trip + time until next fare)

Hot zones at ends of tunnels due to GPS fix lag observed earlier?

Profitability correlates with pickup likelihood on one-way streets



Up until now have focused on pick-up location, what about drop-off?

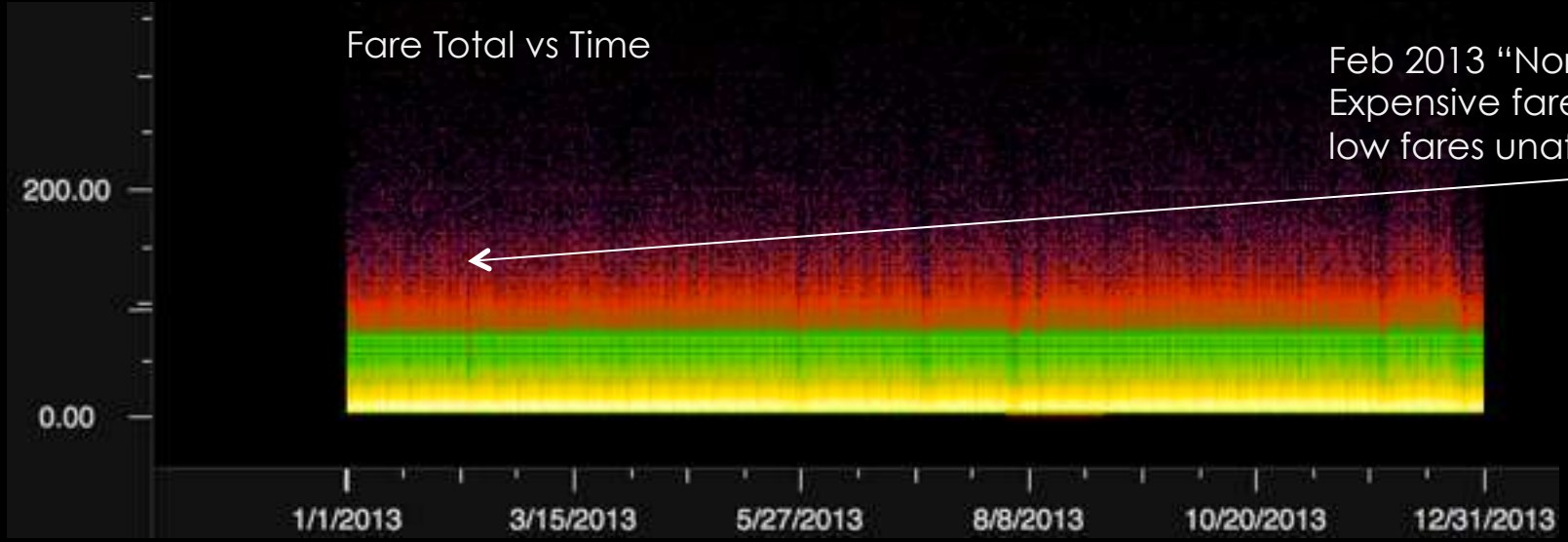
Average $/hr by Drop-off Location

Red	$30/hr
Yellow	$60/hr

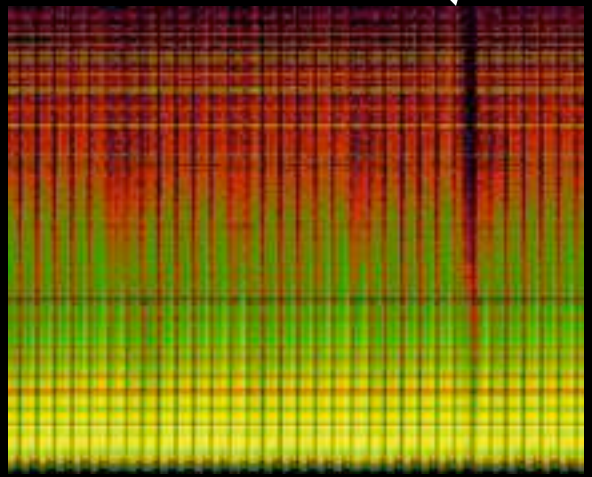
Calculated using:
(fare + tip) / (duration of trip + time until next fare)

Average annual taxi income by pick-up location

Red	$80k
Yellow	$100k



Feb 2013 "Nor`easter storm"
Expensive fares drop off by
low fares unaffected



Tiles approach isn't limited to geographic plots

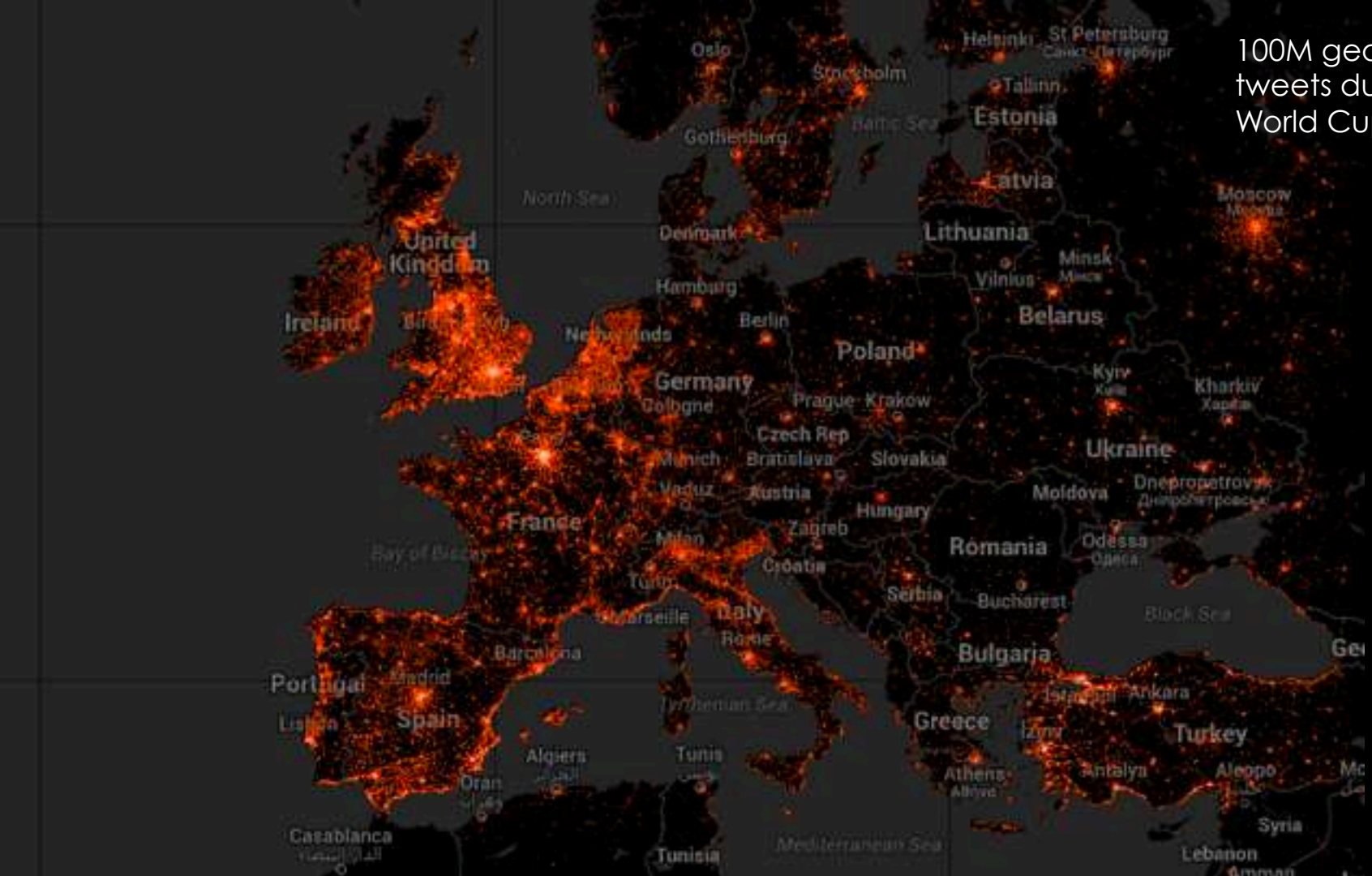
X-Y crossplots can show 3 arbitrary dimensions and expose macro and micro patterns.

e.g. Time + Fare Total + Frequency

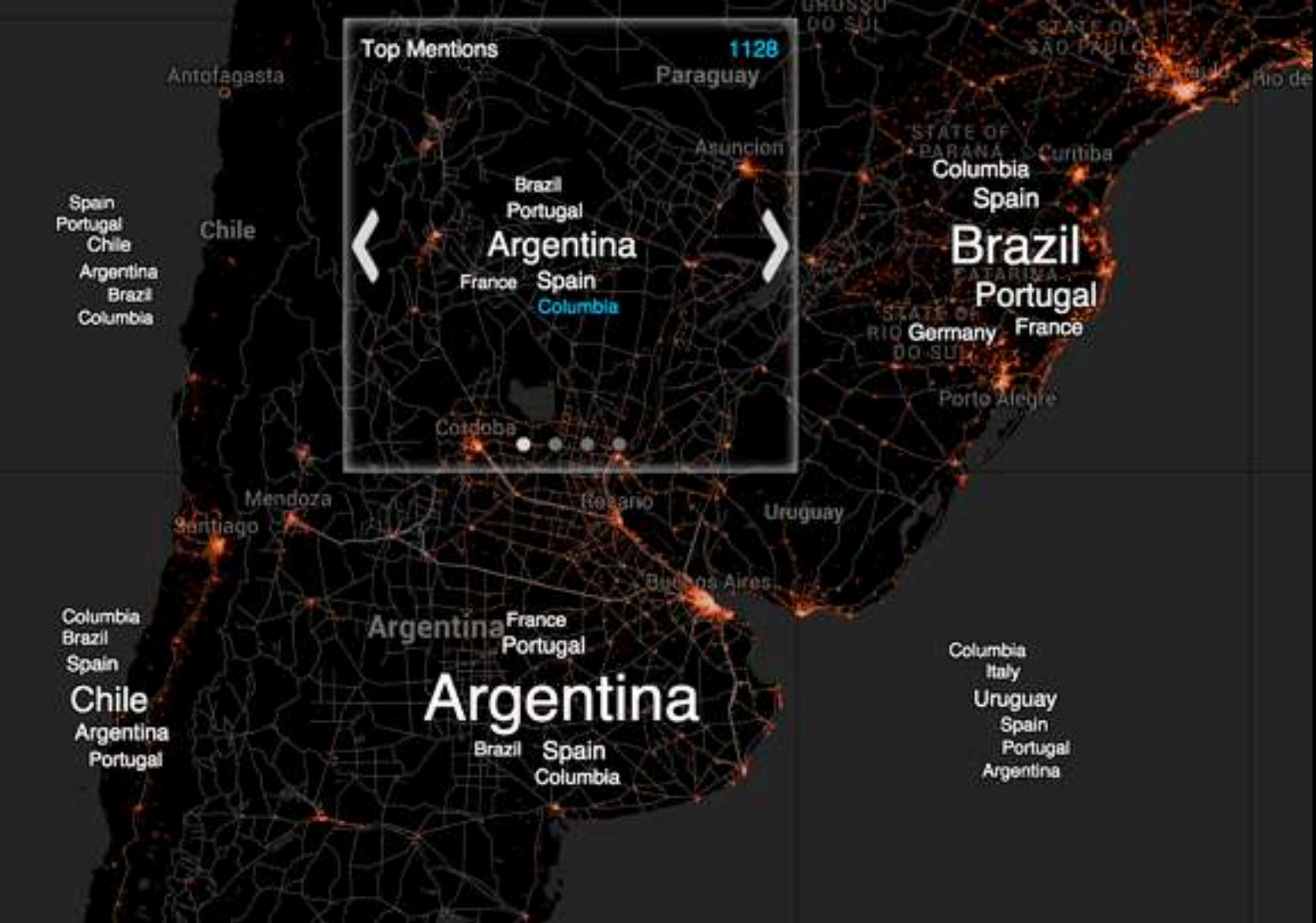
World Cup Tweets

player @mentions

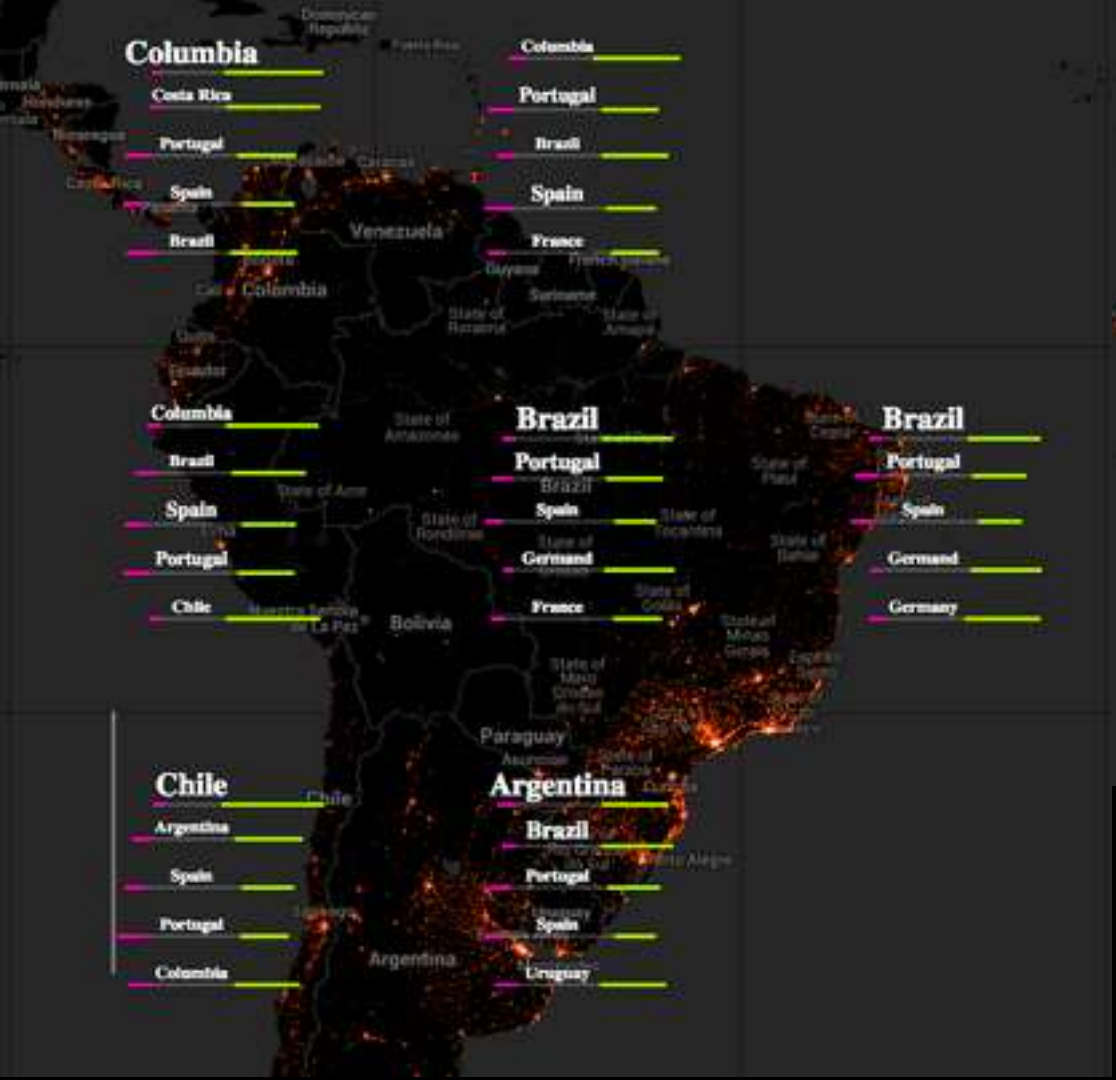
Photo by paulisson miura, CC

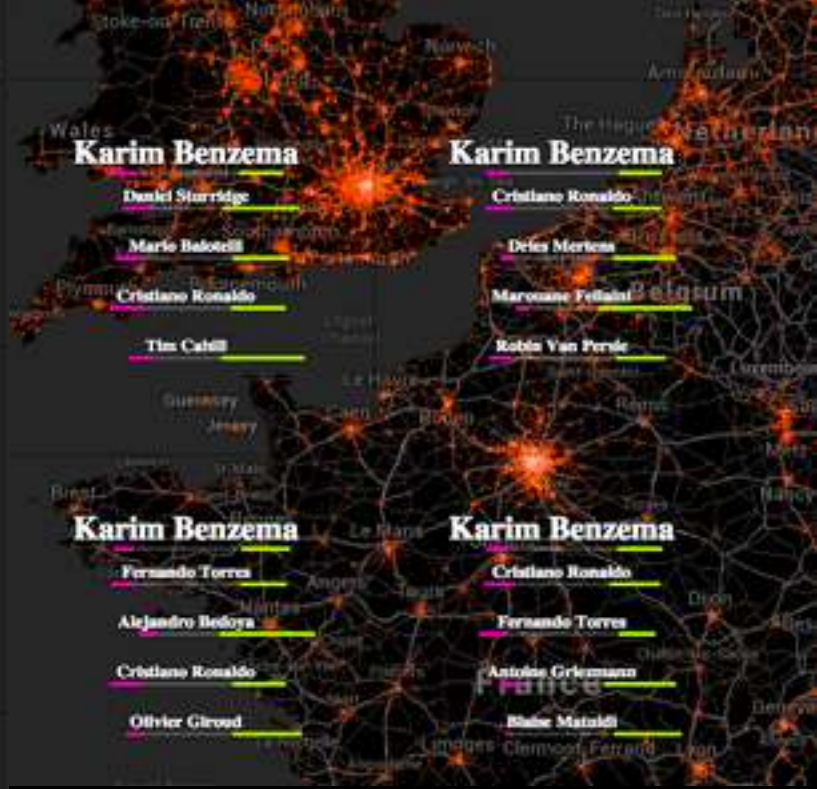


100M geolocated tweets during the World Cup



Analytic overlays showing @mention summaries

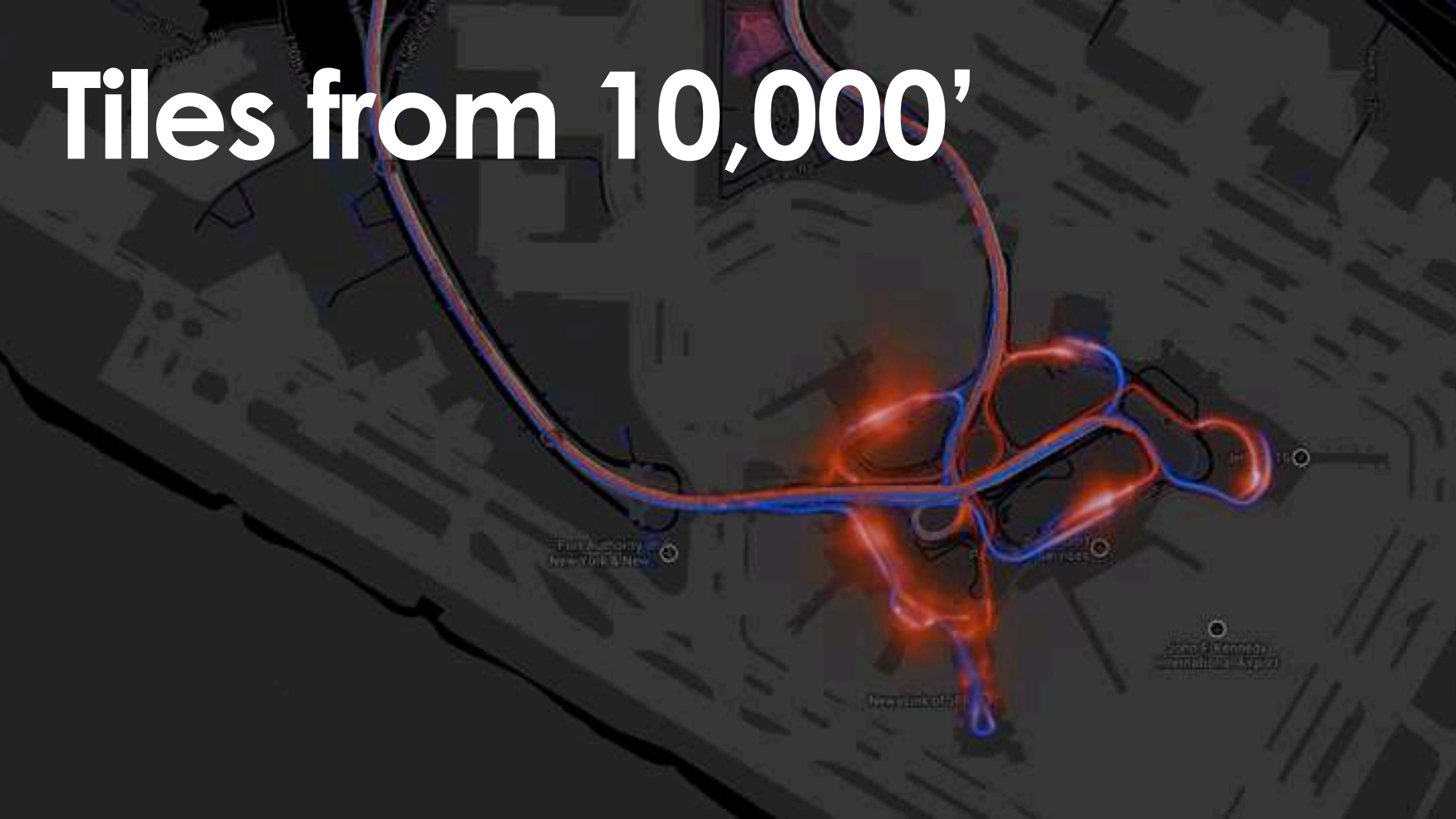




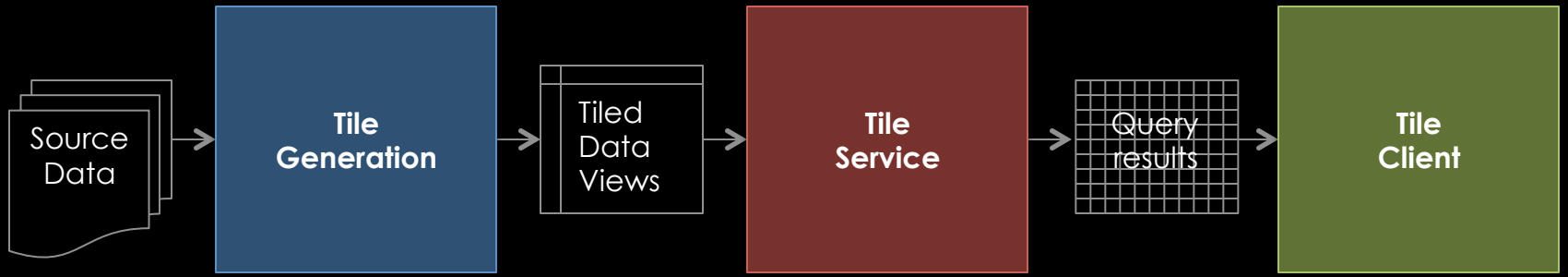
Team and Player @mention sentiment

Interactive overlays

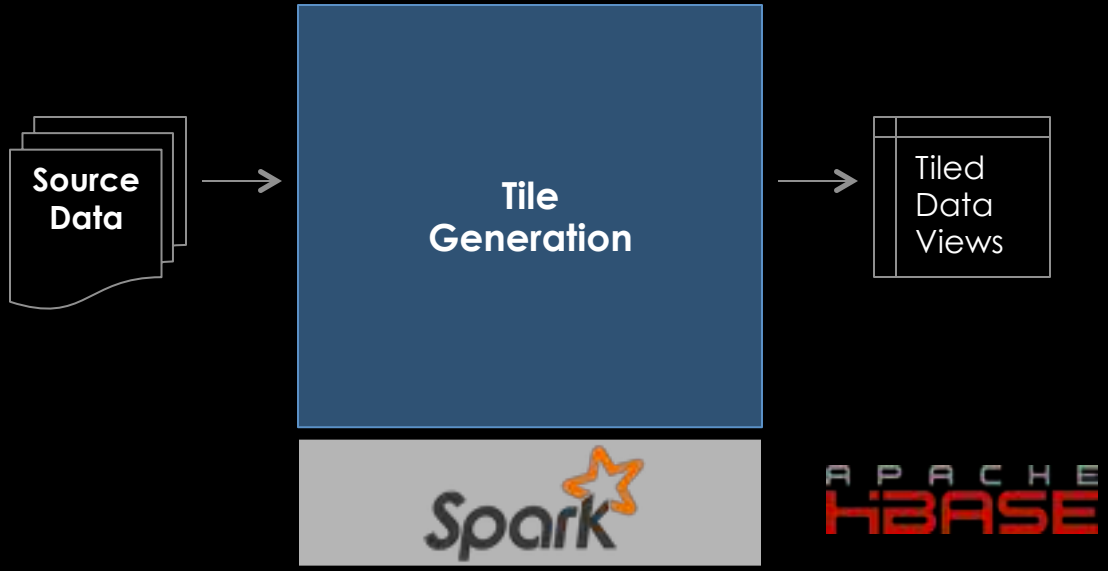
Tiles from 10,000’



Three Tiers



Tile Generation



Tiled Data Views

Aggregate views of source data designed for answering **analytic questions**

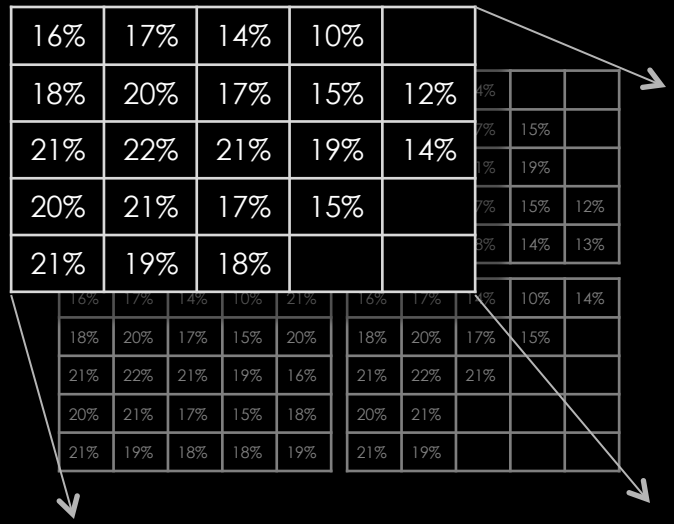
Optimized for query speed

Indexed by tile key - one view/row per tile

AVRO / Thrift

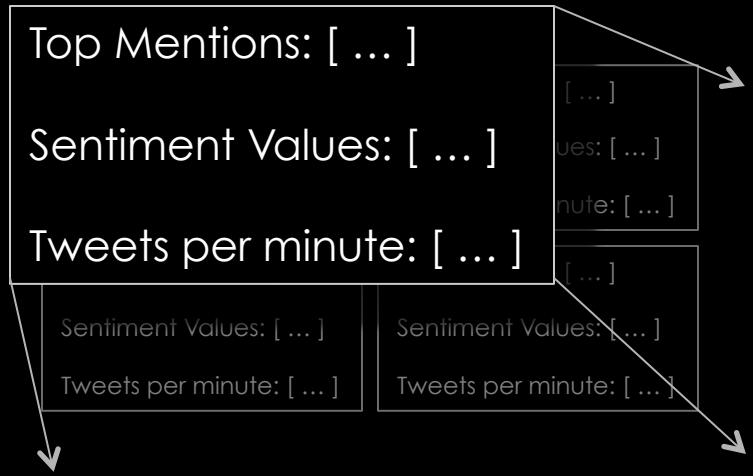
Tiled Data Views

Taxi Tip %

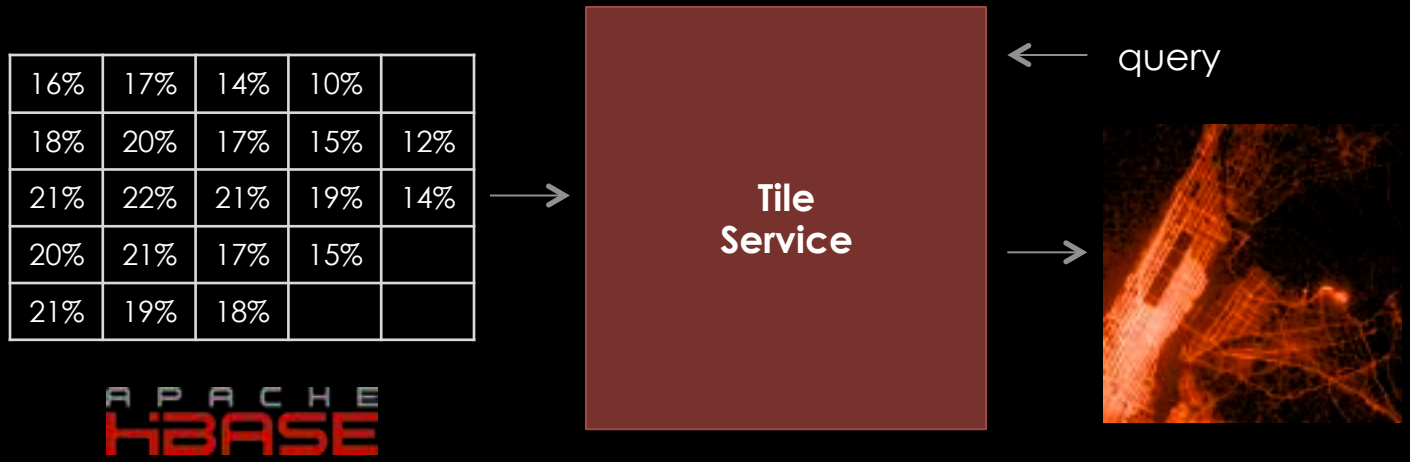


Tiled Data Views

World Cup Sentiment



Tile Generation



Tile Client

Tile Map Service (TMS)

http://{rootURL}/{layer}/{zoom}/{x}/{y}.png

OpenLayers

ESRI

Tile Client

Ready-to-go Tiled Data Browser

http://tiles.oculusinfo.com

https://github.com/oculusinfo/aperture-tiles

Key Points of Value

PLOTTING ALL of the data for richer, truer insights.

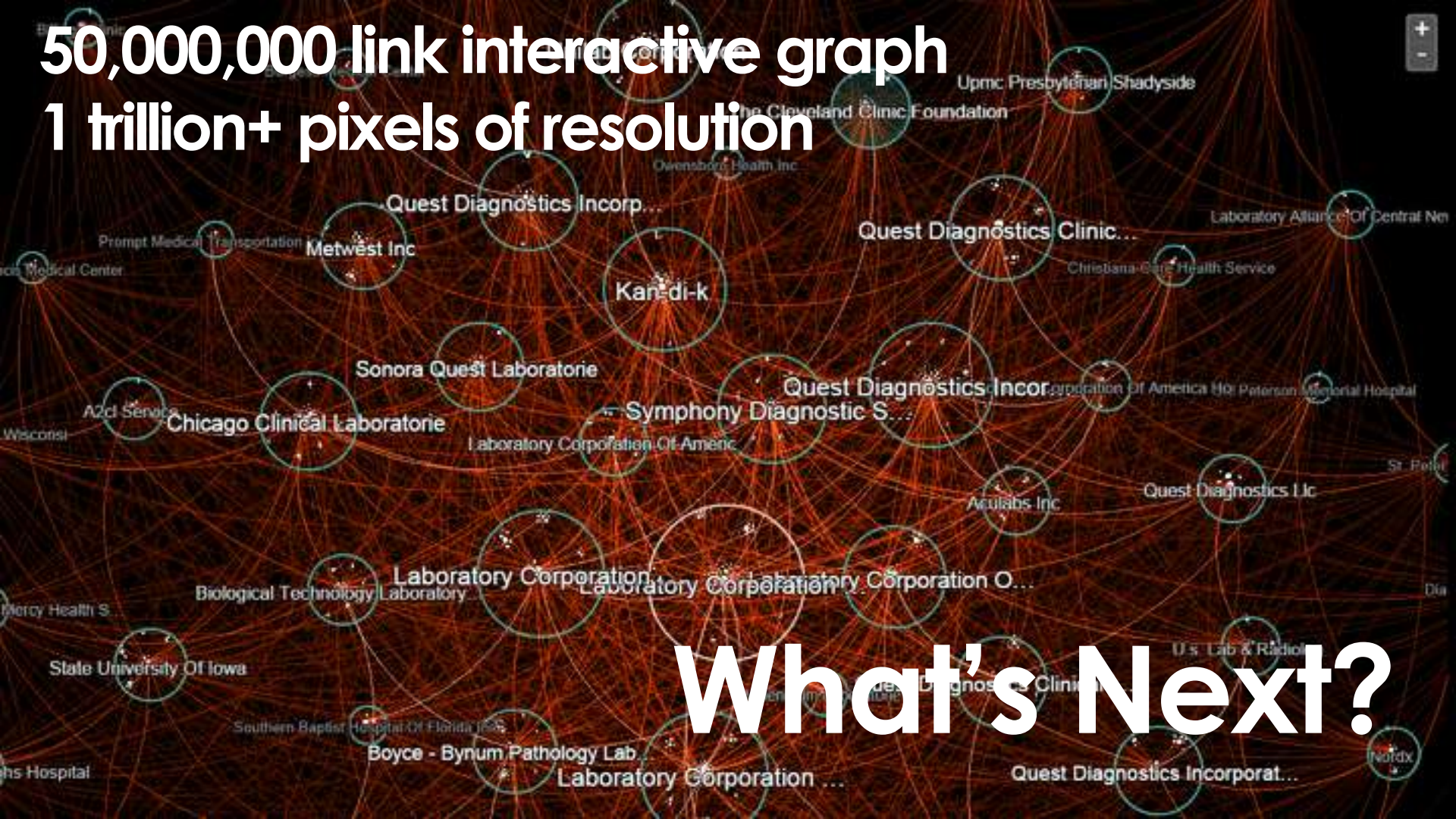
- Reveals **truths** in the data that summaries cannot.
- Provides important **context** and **evidence** for analytics.

TILED, LAYERED user interface for natural, scalable exploration.

- **Intuitive** navigation and understanding.
- Analytic layers at **levels of detail**, down to raw provenance.

INTERACTIVE analysis with data tiling and analytics.

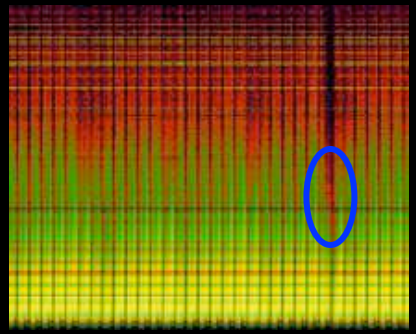
- Rapidly repeatable **tailored** solutions, with filtering + selection.


50,000,000 link interactive graph
1 trillion+ pixels of resolution
Upmc Presbyterian Shadyside
Quest Diagnostics Incorp...
Metwest Inc
Quest Diagnostics Clinic...
Kan-di-k
Sonora Quest Laboratorie
Quest Diagnostics Incor
Symphony Diagnostic S...
Chicago Clinical Laboratorie
Laboratory Corporation Of Americ
Acculabs Inc
Quest Diagnostics Llc
Laboratory Corporation
Laboratory Corporation O...
Biological Technology Laboratory
State University Of Iowa
Boyce - Bynum Pathology Lab...
Laboratory Corporation ...
Quest Diagnostics Incorporat...
What's Next?

What's Next?

Interactive
Tiled Graphs

Streaming Data
and live query



More advanced
filtering and analytics

Questions?

Oculus gratefully acknowledges the support of DARPA in research and development of this work.