Visual Exploration of Big Urban Data

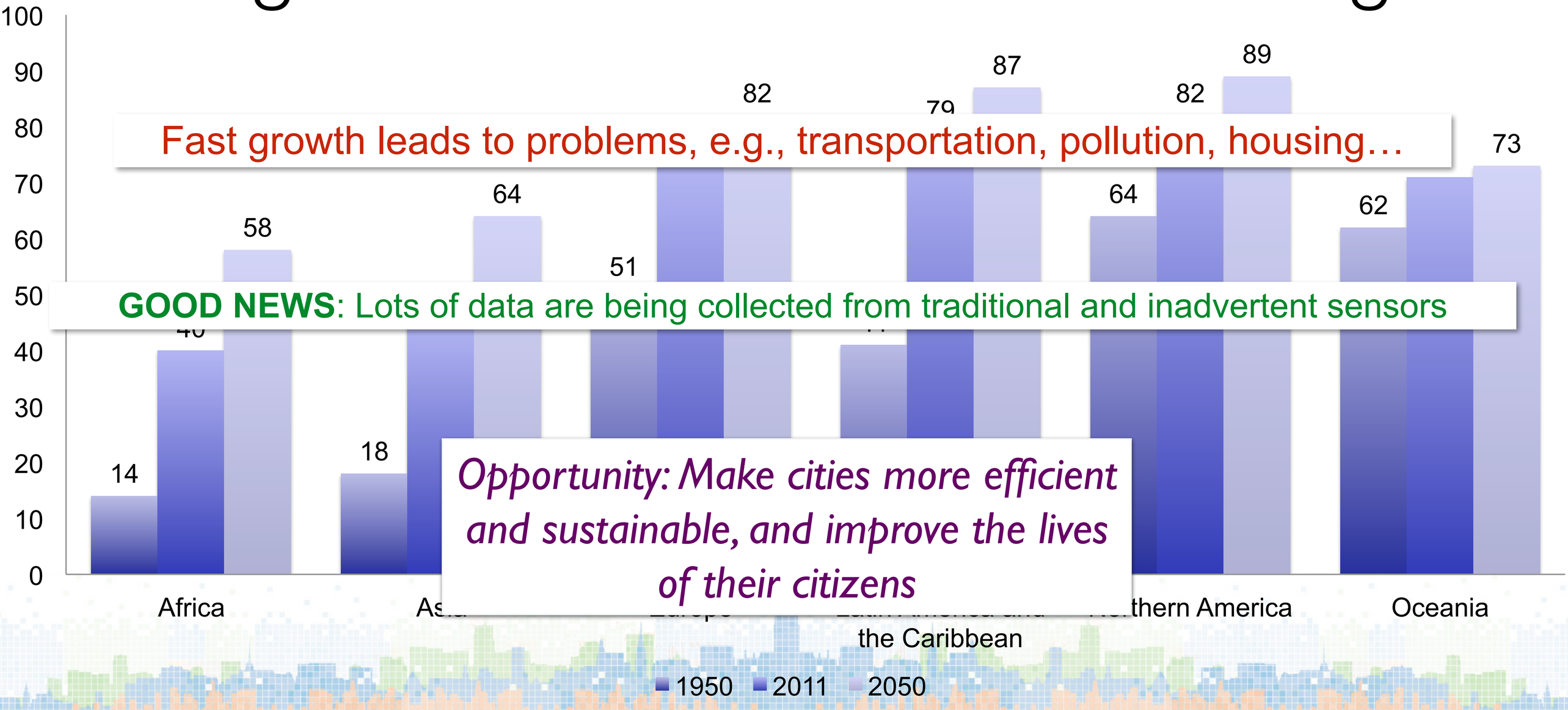
Huy T. Vo
Center for Urban Science and Progress (CUSP)
Polytechnic School of Engineering (Poly)
New York University (NYU)

NYU-CUSP and Urban Science

“Research center that uses New York City as its laboratory and classroom to help cities around the world become more productive, livable, equitable, and resilient. CUSP observes, analyzes, and models cities to optimize outcomes, prototype new solutions, formalize new tools and processes, and develop new expertise/experts”

- Multisector collaboration: universities, industry, national labs and city agencies
- Multidisciplinary collaboration
- Acquire, integrate, explore large diverse datasets while respecting privacy
- Training students who will create the new discipline http://cusp.nyu.edu

Big Cities – the world is urbanizing



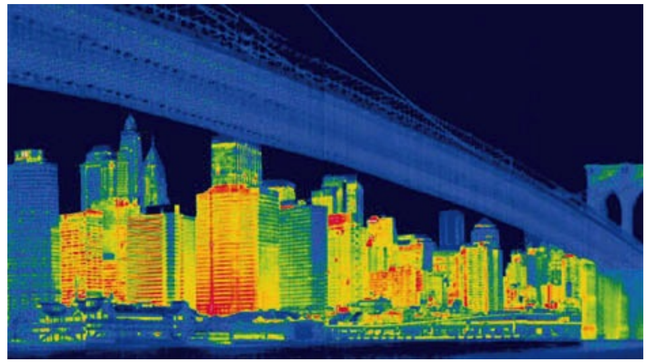
Fast growth leads to problems, e.g., transportation, pollution, housing…

GOOD NEWS: Lots of data are being collected from traditional and inadvertent sensors

Opportunity: Make cities more efficient and sustainable, and improve the lives of their citizens

Urban Data

- **Organic data flows**
  - Administrative records & transactions(census, permits, sales…)
  - Operational (traffic, transit, utilities, health system, …)
  - News and social media (Twitter feeds, blog posts, Facebook, …)
- **Sensors**
  - Personal (location, activity, physiological)
  - Fixed *in situ* sensors
  - Crowd sourcing (mobile phones, …)
- **Opportunities for "novel" sensor technologies**
  - Visible, infrared and spectral imagery, RADAR, LIDAR
  - Gravity and magnetic, seismic, acoustic
  - Ionizing radiation, biological, chemical



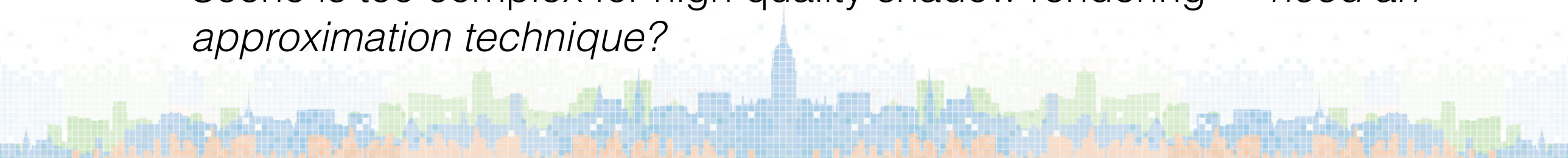
Big Data

“Big data is an all-encompassing term for any collection of data sets so **large and complex** that it becomes **difficult** to process using **traditional** data processing **applications**.”

— *Wikipedia:Big_Data*

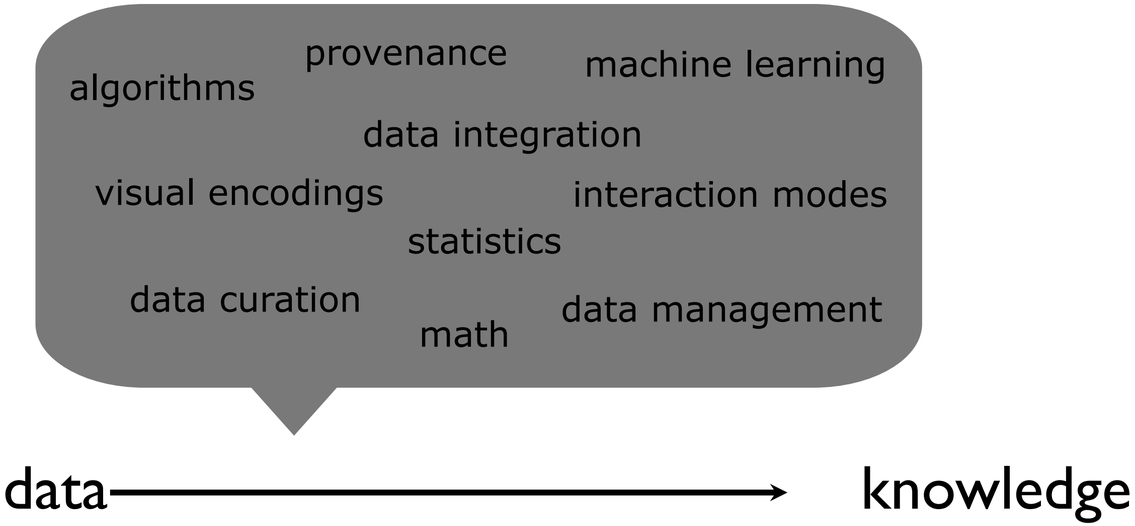
Big Urban Data

- Urban data that is too large:
  - Microsoft Excel couldn't load the data — *need a more scalable tool?*
  - ArcGIS is too slow — *need a faster database index?*
- Urban data that is too complex:
  - Searching the entire parameter space is too expensive for my simulator — *need a distributed algorithm?*
  - Scene is too complex for high quality shadow rendering — *need an approximation technique?*



Big Data is not new!

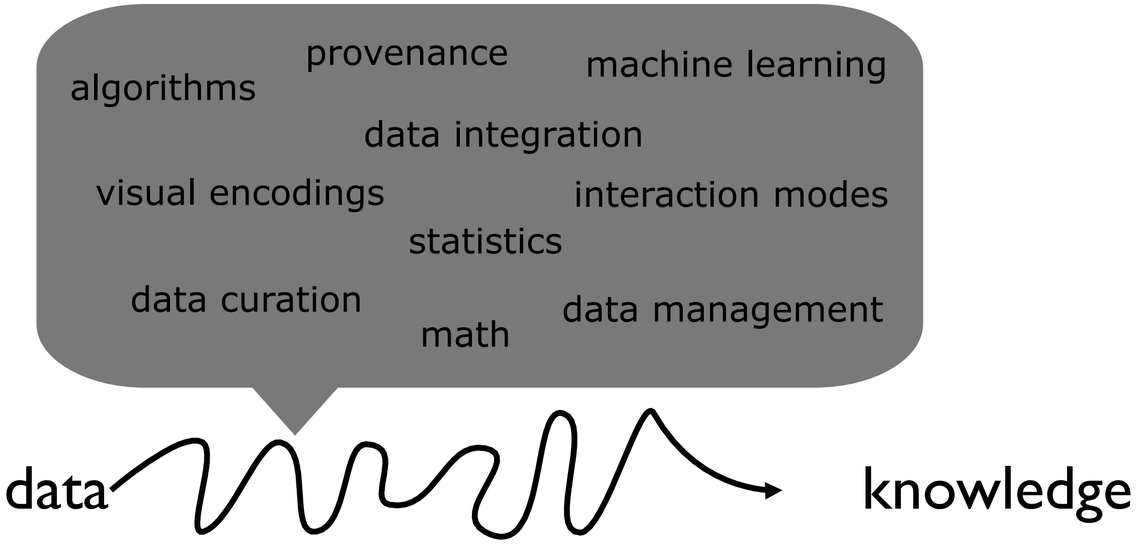
- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”
- Scalability for people is!
  - Data owners are non-expert



Big Data is not new!

- Scalability for batch computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Data were “in good hands”
- Scalability for people is!
  - Data owners are non-expert

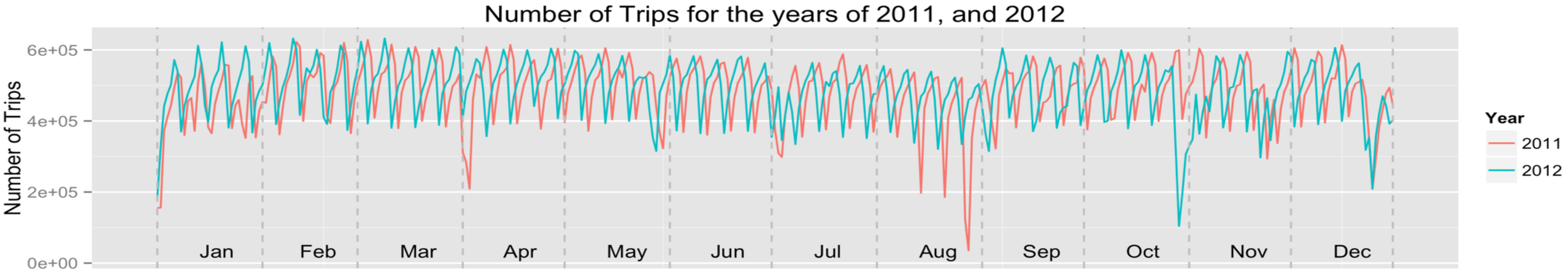
regardless of whether data are big or small



Urban Data Analysis: Desiderata

- Scalable tools and techniques that aid data enthusiasts to find, integrate, and interactively explore data
  - Automate tedious tasks as much as possible
  - Guide users in the exploration process
- Many different kinds of users with little or no CS training
  - Social scientists
  - Government employees
  - Citizens
- Needs big data management and big data visualization

Exploring Big Urban Data: NYC Taxis



Taxis are ***sensors*** that can provide unprecedented insight into city life: economic activity, human behavior, mobility patterns, …

"What is the average trip time from Midtown to the airports during weekdays?''

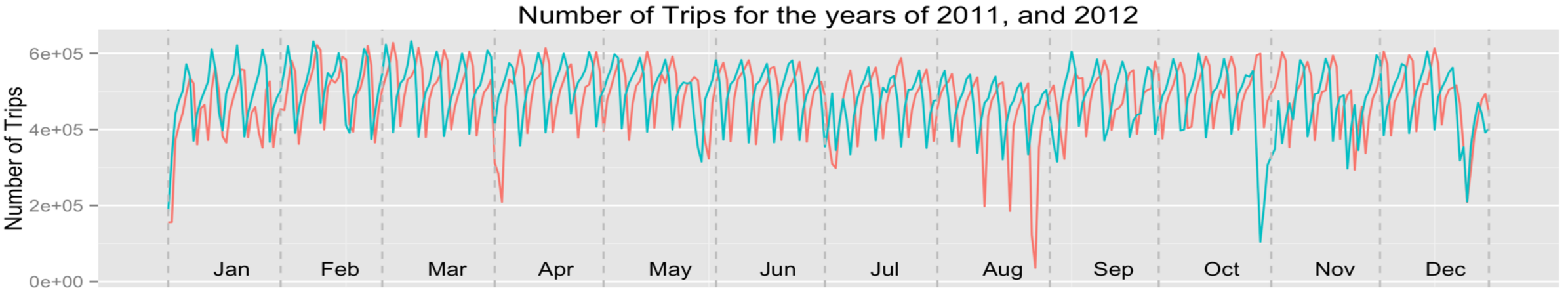
"How the taxi fleet activity varies during weekdays?"

"How was the taxi activity in Midtown affected during a presidential visit?''

"How did the movement patterns change during Sandy?"

"Where are the popular night spots?"

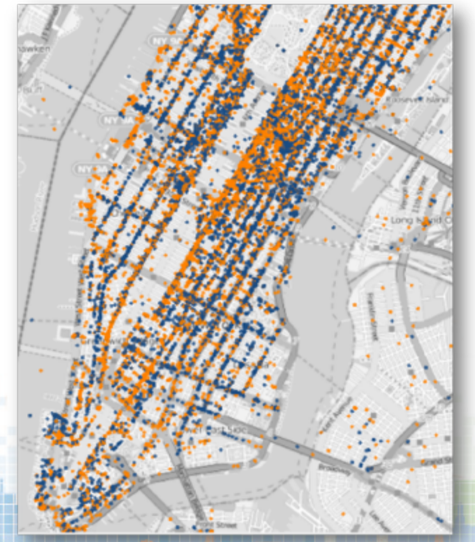
Exploring Big Urban Data: NYC Taxis



7-8am

8-9am

9-10am



10-11am

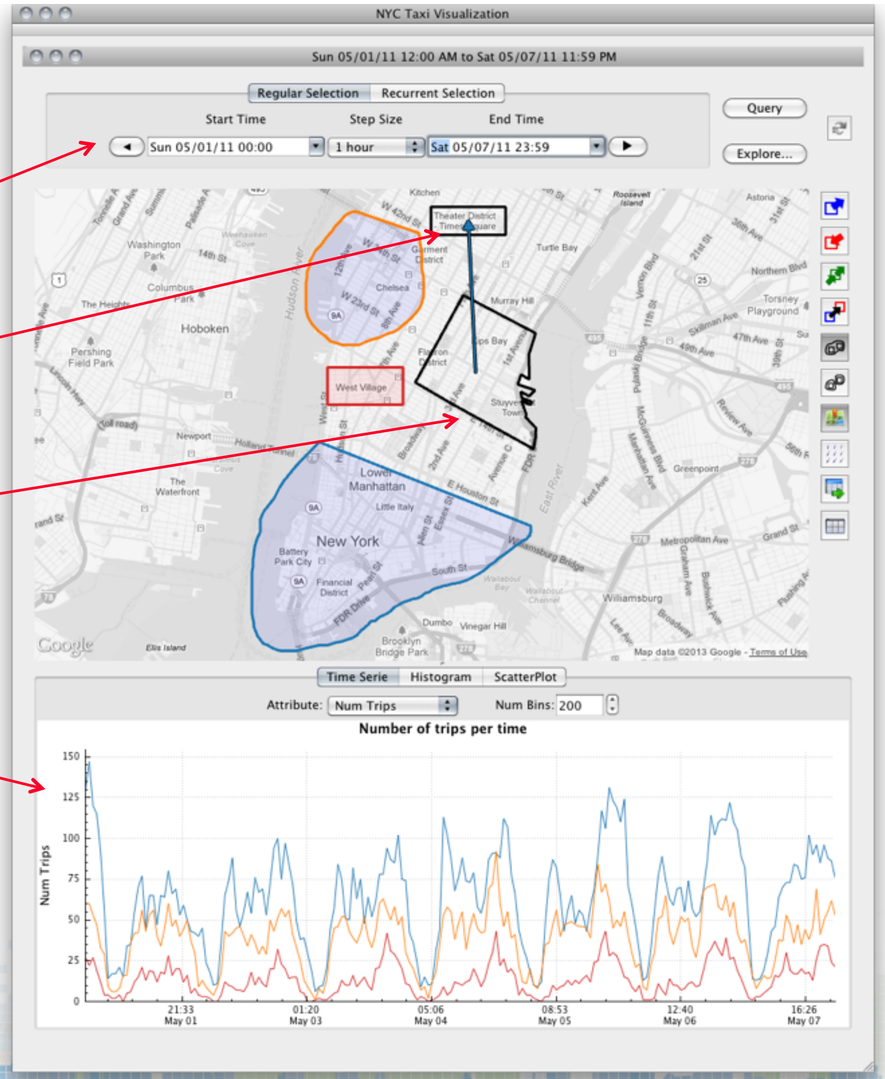
Exploring Taxi Data: Challenges

- Data are big: ~500k trips/day - 780 million trips in 5 years
- Government, policy makers and scientists are unable to explore the whole data — their tools are not scalable!
  - dependency on data specialists, limited to confirmatory tasks
- Data are complex:
  - spatio-temporal: pick up + drop off
  - trip attributes: e.g., distance traveled, cost, tip
- Too many data slices to examine — expensive queries
- Our goals: Design a usable interface, efficiently support interactive + exploratory queries

TaxiVis: Visually Exploring NYC Taxi Data

```
SELECT  *
FROM    trips
WHERE pickup_time in (5/1/11,5/7/11)
        AND
          dropoff_loc in “Times Square”
        AND
          pickup_loc   in “Gramercy”
```

Data selection and result exploration are unified



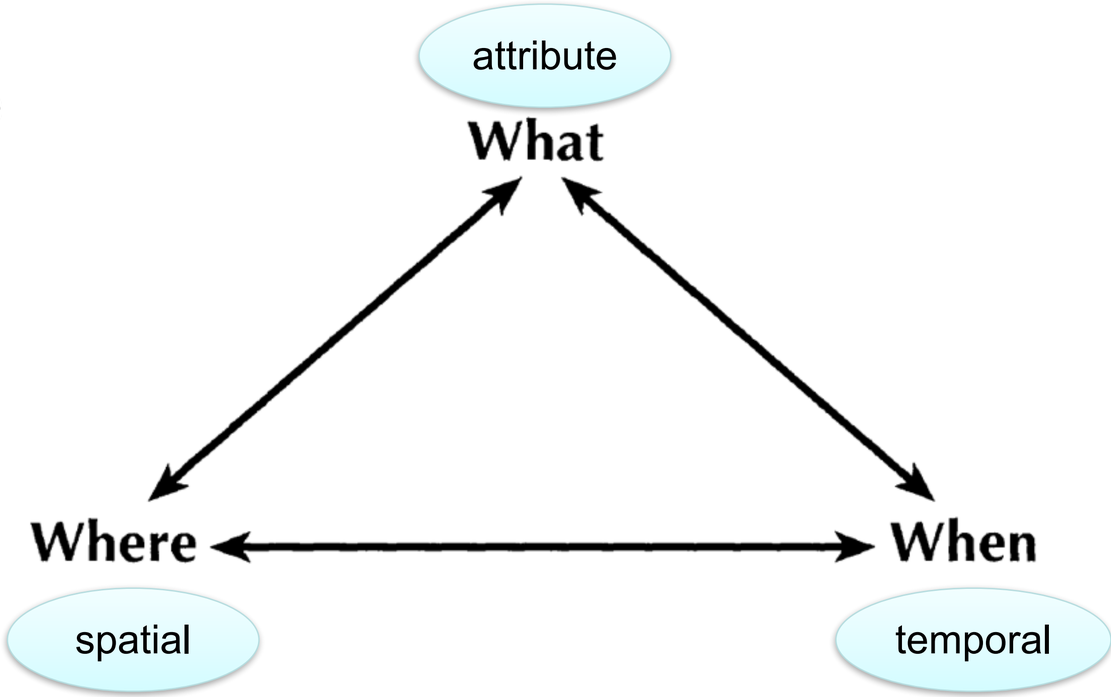
Users select a data slice by specifying spatial, temporal and attribute constraints

[Ferreira et al., IEEE TVCG 2013] http://taxivis.org

Peuquet's Triad for Spatio-Temporal Data

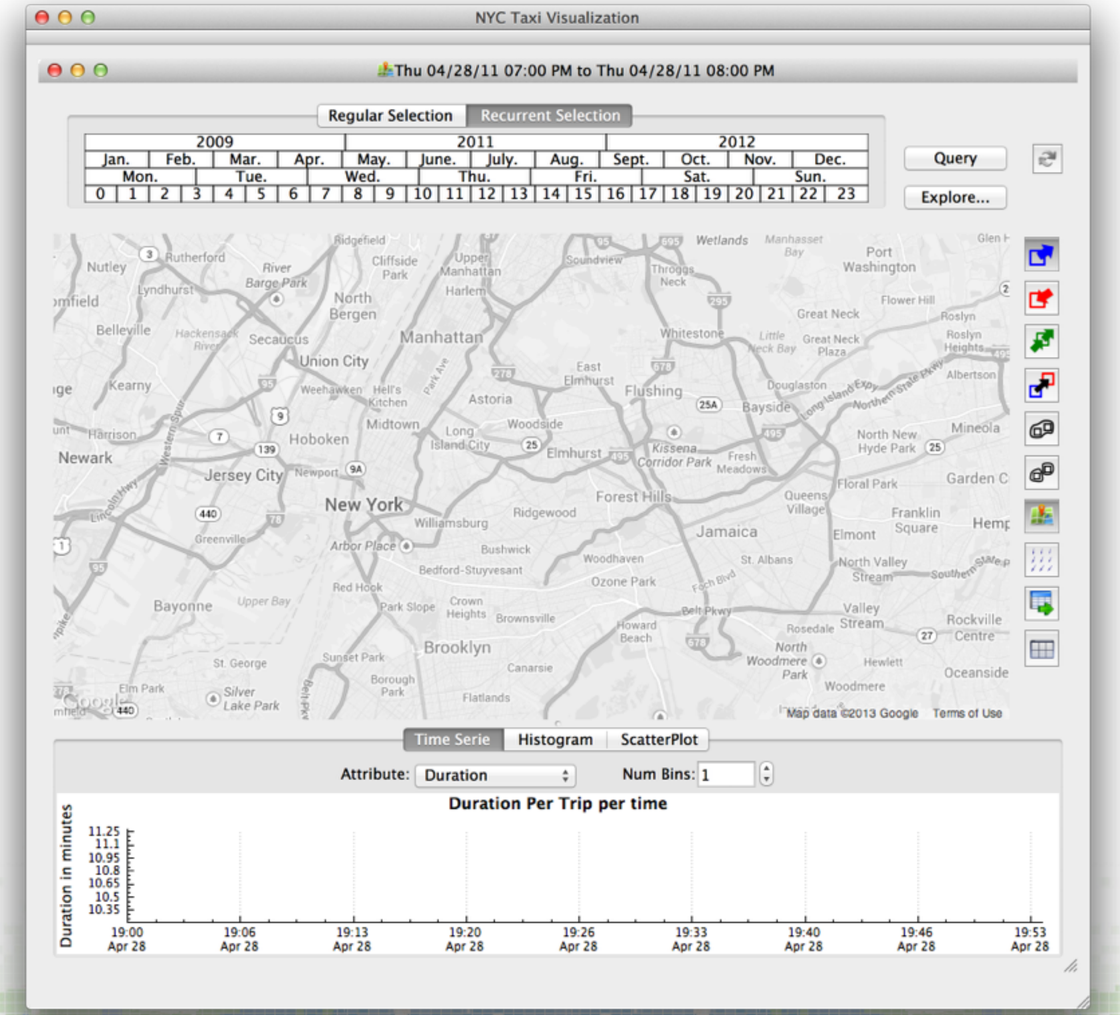
Classes of questions:

- when + where ➔ what: "What is the average trip time from Midtown to the airports during weekdays?''
- when + what ➔ where: "Where are the hot spots in Manhattan in weekends?"
- where + what ➔ when: "When were activities restored in Lower Manhattan after the Sandy hurricane?"



When + Where → What

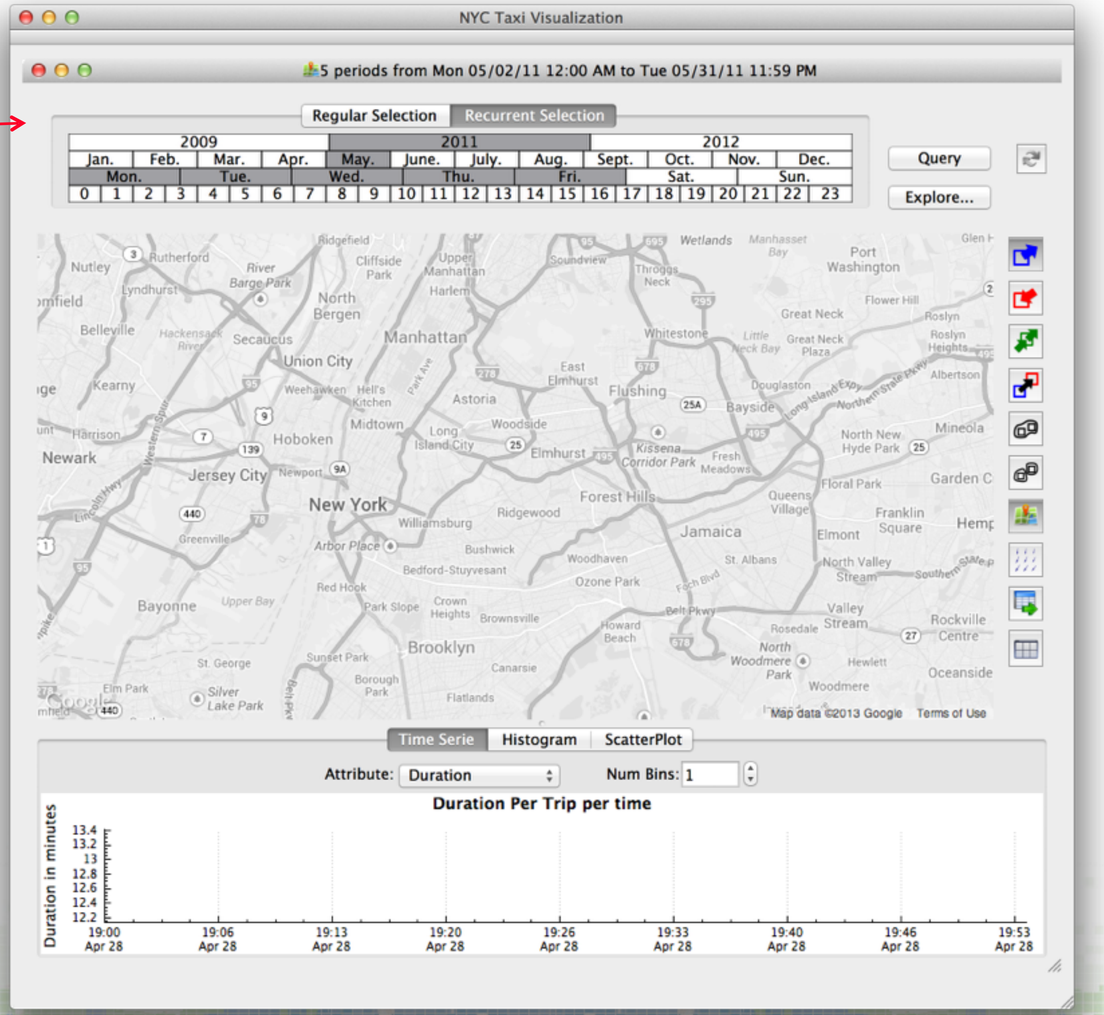
"What is the average trip time from Midtown to the airports during weekdays?



When + Where → What

"What is the average trip time from Midtown to the airports during weekdays?

When?

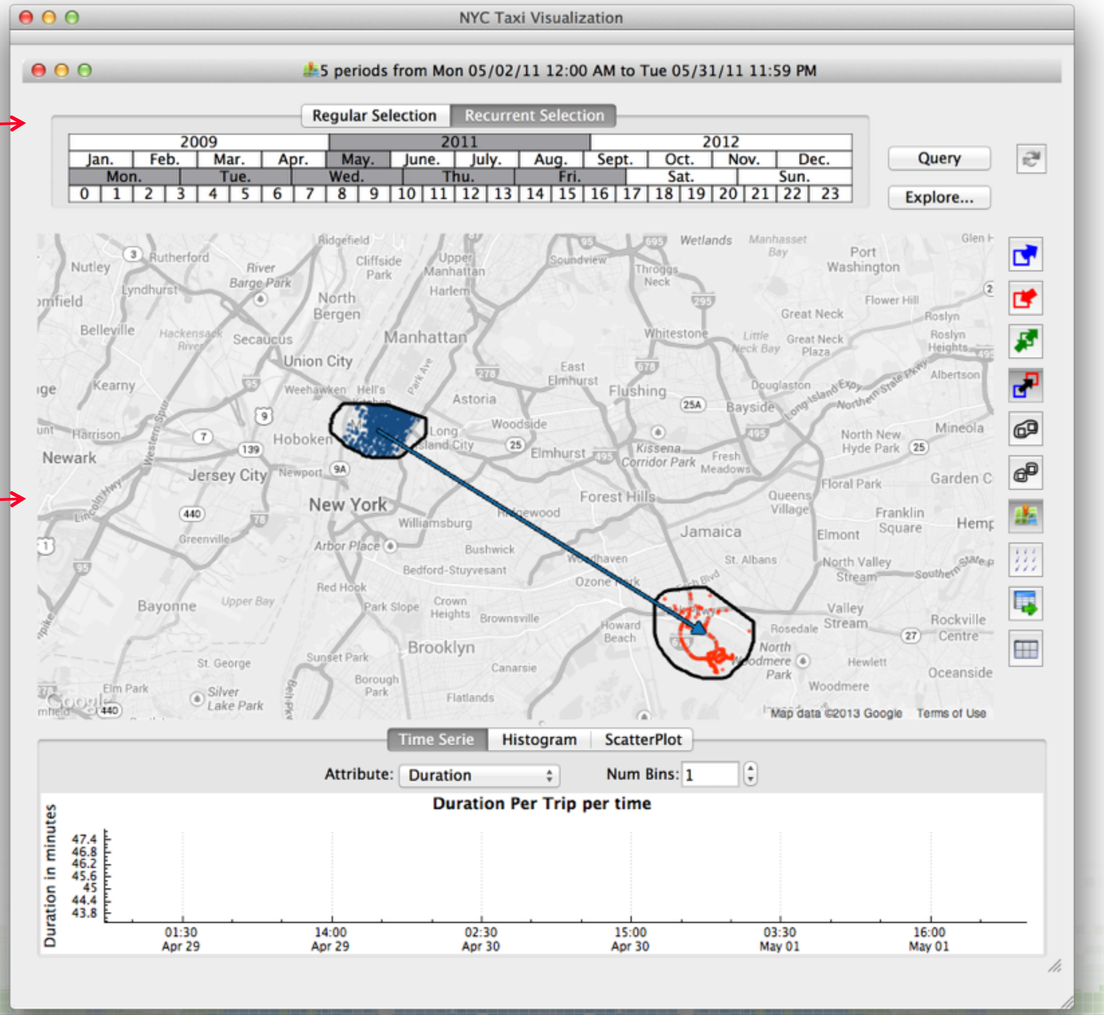


When + Where → What

"What is the average trip time from Midtown to the airports during weekdays?

When?

Where?



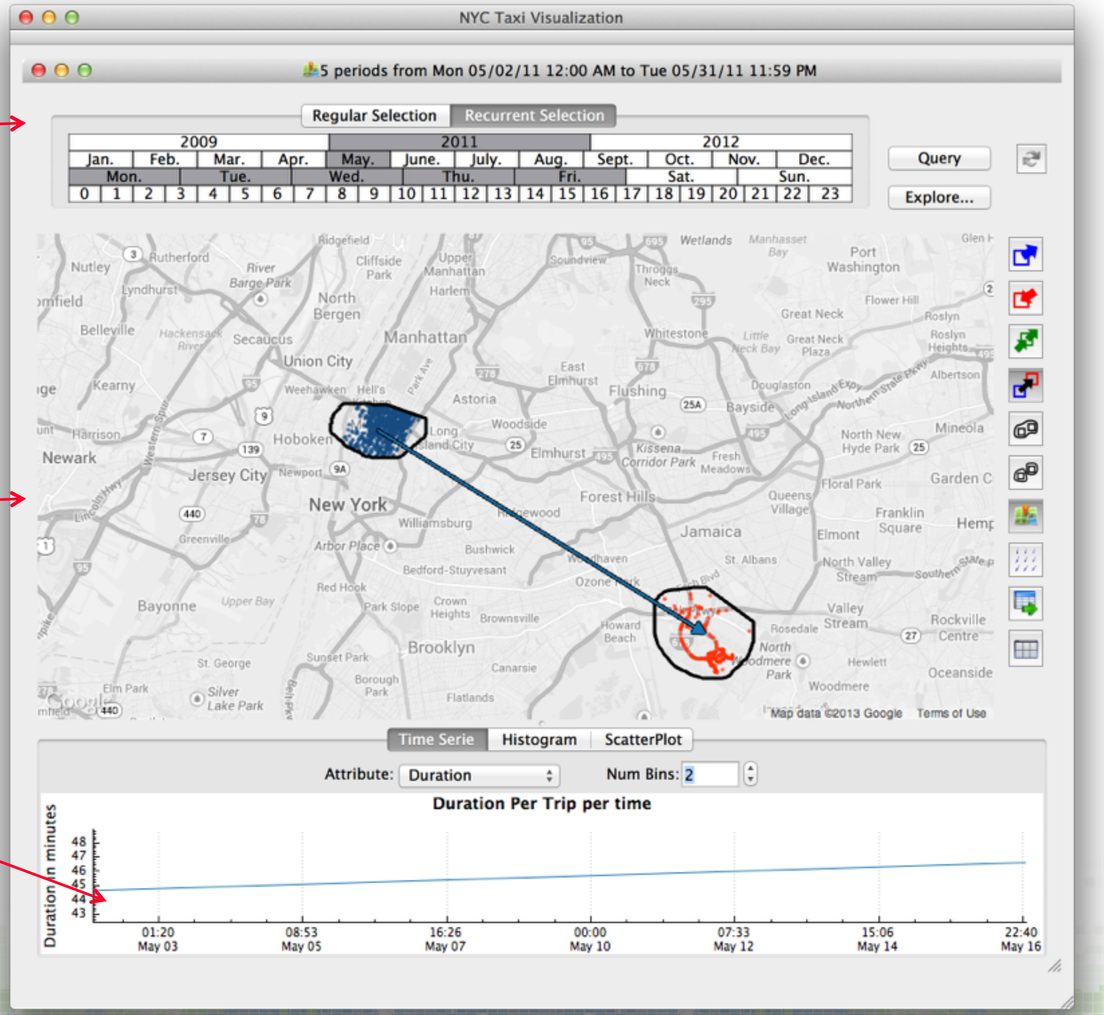
When + Where → What

"What is the average trip time from Midtown to the airports during weekdays?

When?

Where?

What



TaxiVis in Action

Supporting Interactive Queries

- Raw data:
  - 3 years
  - 150 GB in 48 CSV files
  - 520M trips
  - 12 fields, 2 spatial-temporal attributes
- After ETL: 50 GB in binary format

	SQLite	PostgreSQL
Storage Space in GB	100	200
Building Indices in Minutes (One Year of Data)	3,120	780
1K Items Query in Seconds	8	3
100K Items Query in Seconds	85	24

Supporting Interactive Queries

Spatio-temporal index based on kd-trees:

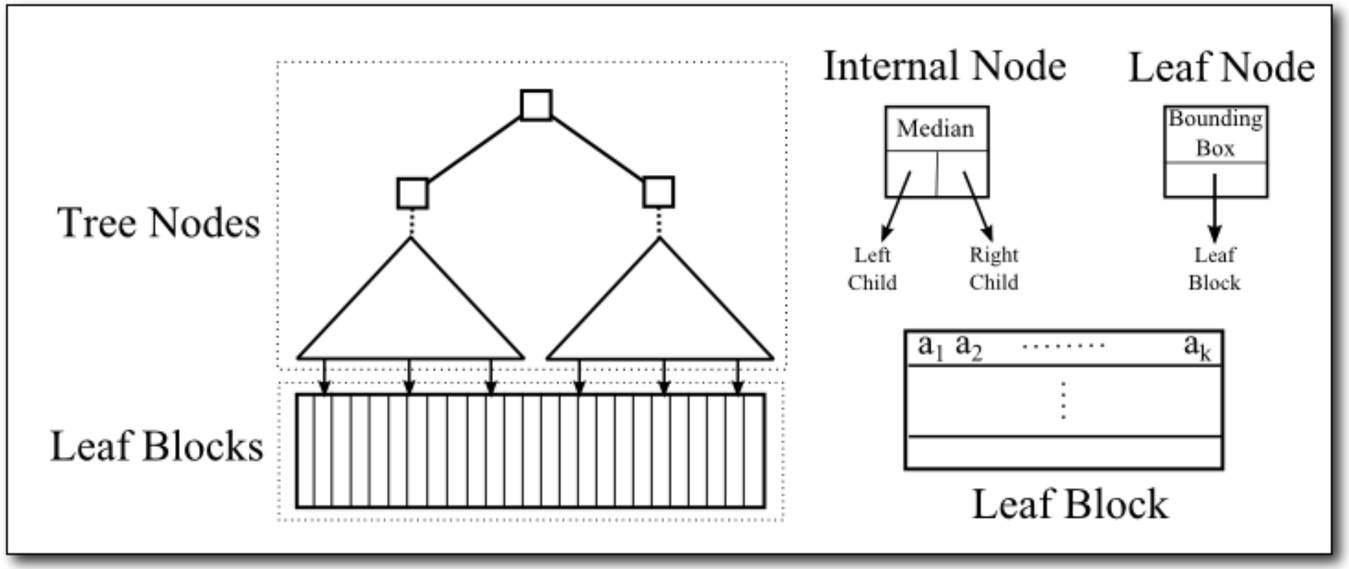
- Faster queries (>10x faster)
- Faster indexing time (~30x faster)
- Smaller footprint (3-5x smaller)

	SQLite	PostgreSQL	Our Solution
Storage Space in GB	100	200	30
Building Indices in Minutes (One Year of Data)	3,120	780	28
1K Items Query in Seconds	8	3	0.2
100K Items Query in Seconds	85	24	2

Supporting Interactive Queries

Spatio-temporal index based on kd-trees:

- Can index multiple attributes!
- Compact representation
  - store only 2 elements per internal node
- Support disk-based access
  - post-ordered traversal memory mapped files



Query	MongoDB (1 GPU) Time(sec)	MongoDB (3 GPUs) Time(sec)	PostgreSQL Time(sec)	PostgreSQL Speedup (1 GPU)	PostgreSQL Speedup (3 GPUs)	ComDB Time(sec)	ComDB Speedup (1 GPU)	ComDB Speedup (3 GPUs)
1	0.237	0.103	141.8	598	1376	136.9	578	1329
2	0.199	0.065	129.2	649	1987	119.6	601	1840
3	0.202	0.093	97.1	480	1044	39.4	195	423
4	0.183	0.069	103.7	566	1502	25.6	140	371
5	0.361	0.159	106.3	294	668	23.8	66	149
6	0.325	0.174	102.6	315	589	28.9	89	166

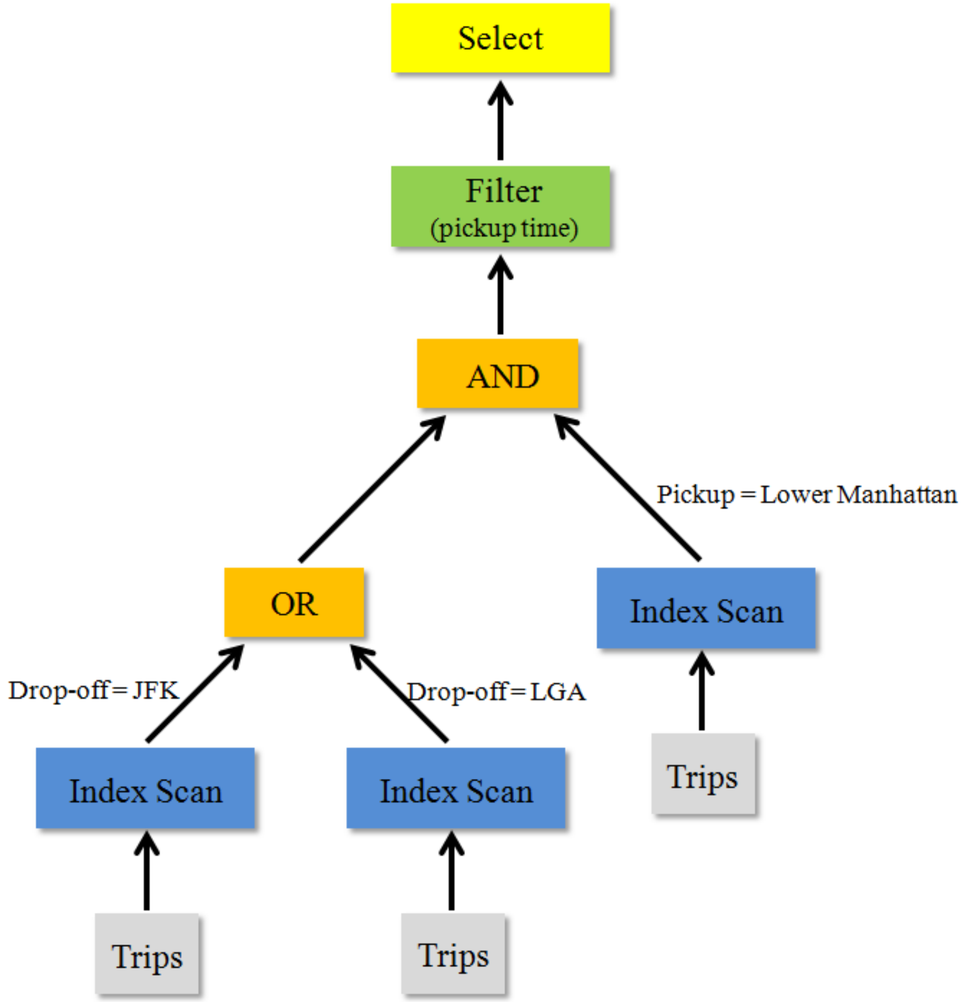
Works on GPUs!

2 orders of magnitude faster than RBDMSs

[under submission]

Why are RDBMS not as “efficient”?

- They **are** efficient **BUT**
  - designed for batch queries, not very good for interactive queries [Fekete & Silva, IEEE DEB 2012]
  - amortized cost of millions of queries
- R(*)-trees are limited to a single spatial attribute
  - Taxi data has origin + destination — needs a join!
  - Expensive in high-dimensional data [Kriegel, VLDB 96]
- Lots of expensive point-in-polygon tests
  - Filtering by other query constraints helps!
  - Using GPUs helps even more!



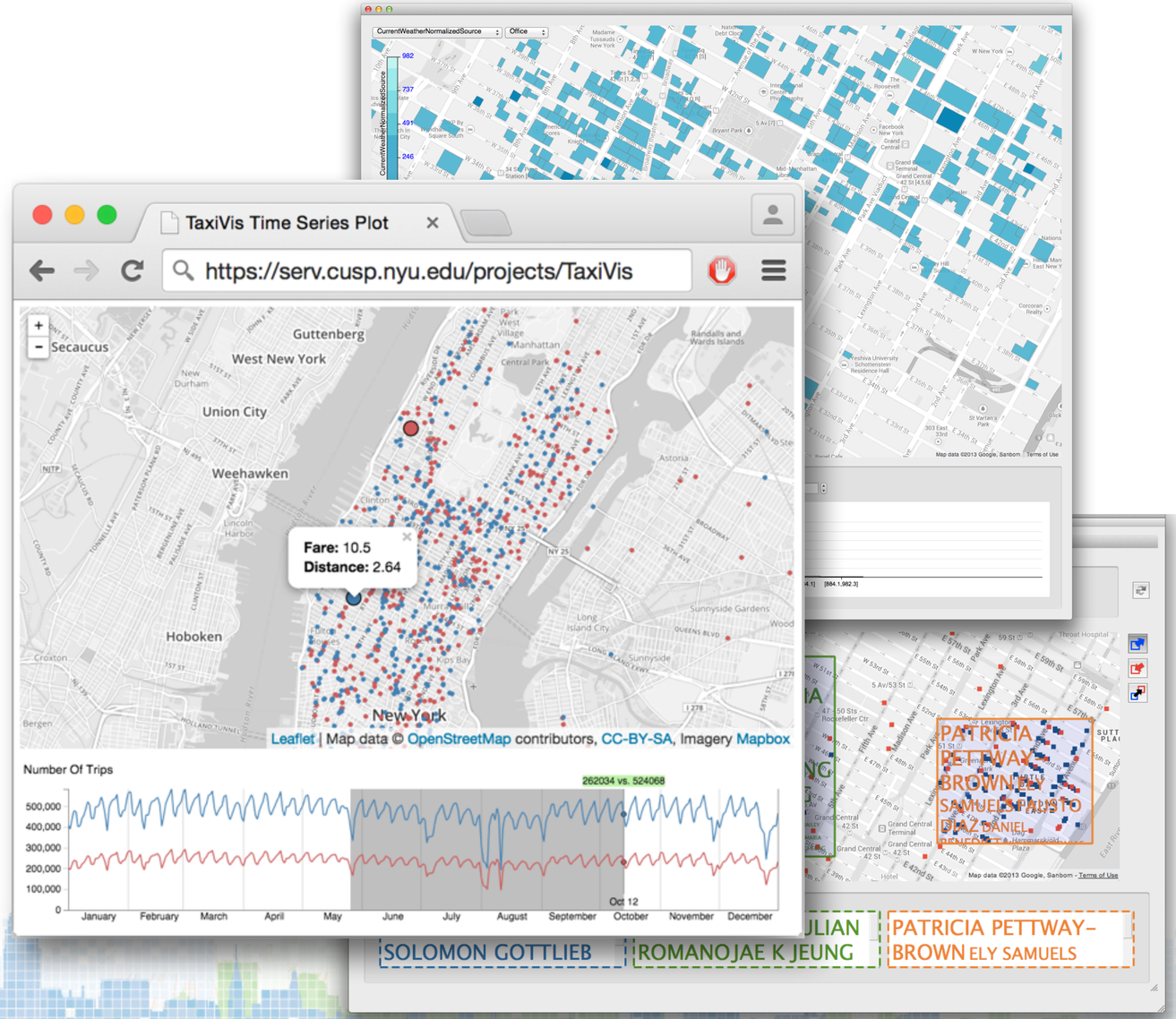
R-tree or kd-tree?

- Depends on the community: Database or Computer Graphics?
- R-tree:
  - Fast update time
  - Support disk-based access
- Kd-Tree:
  - Flexibility in high-dimensional data
  - Tuned for in-memory read-optimized access

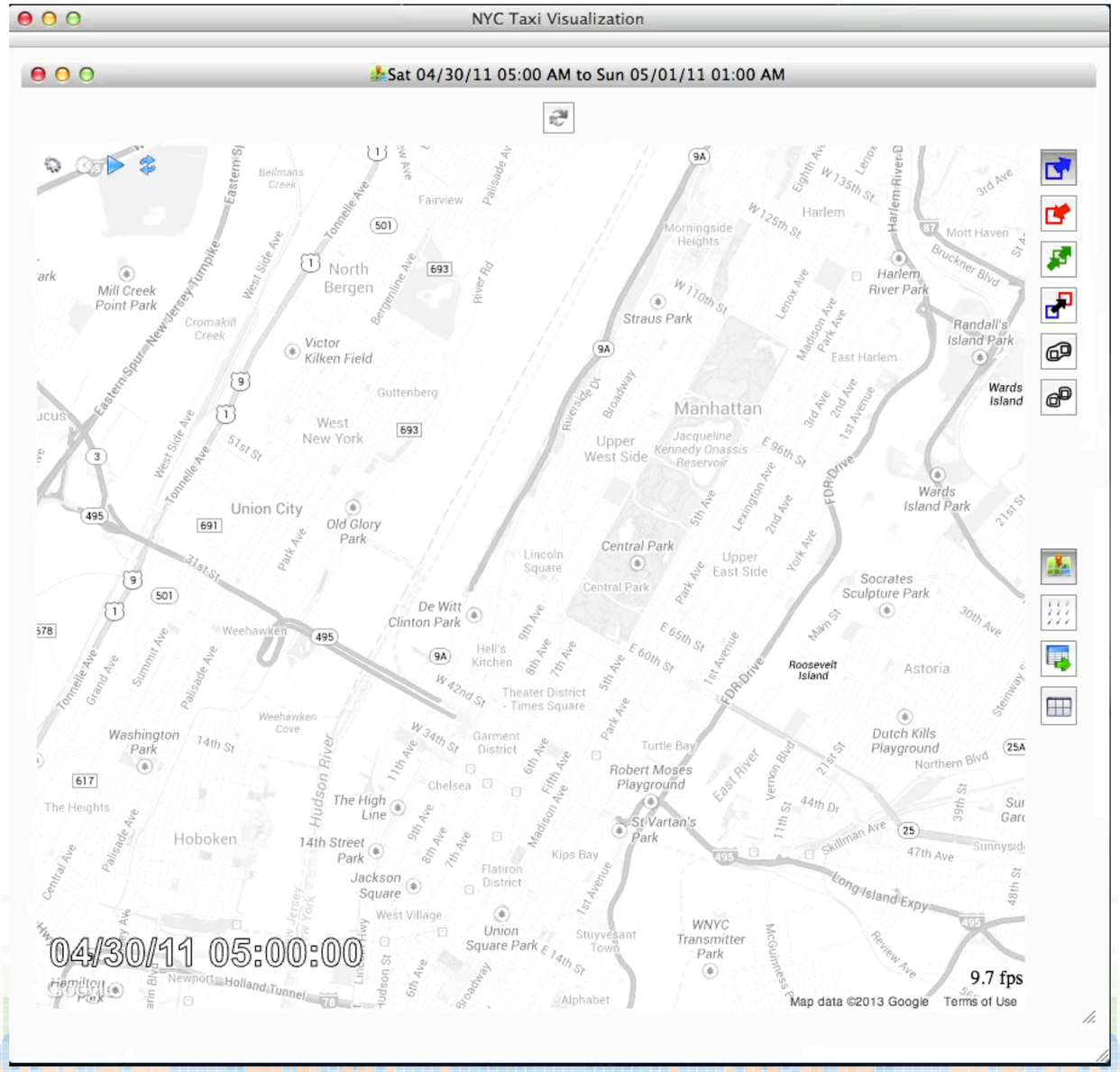
our implementation is a kd-tree with disk-based access!

TaxiVis: Status

- Demoed to NYC DOT and TLC
  - They are currently using the prototype!
- Applying to different data sets
  - Bikes, energy consumption, property ownership, etc.
- BusVis — in progress
  - Web-based Visualization



Life of a Taxi in a Day

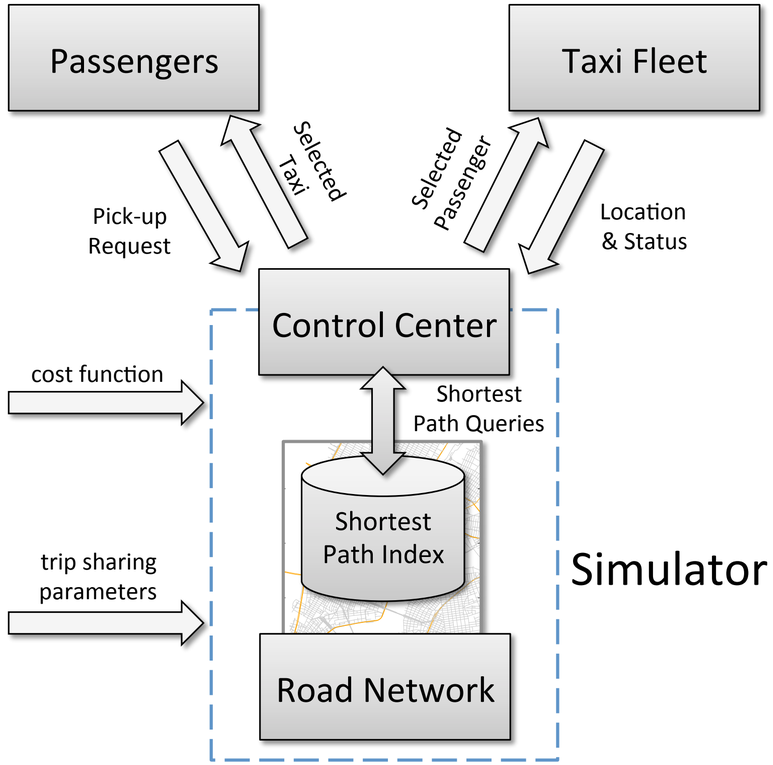


[IEEE Big Data 2013]

Beyond TaxiVis: “TaxiMining”

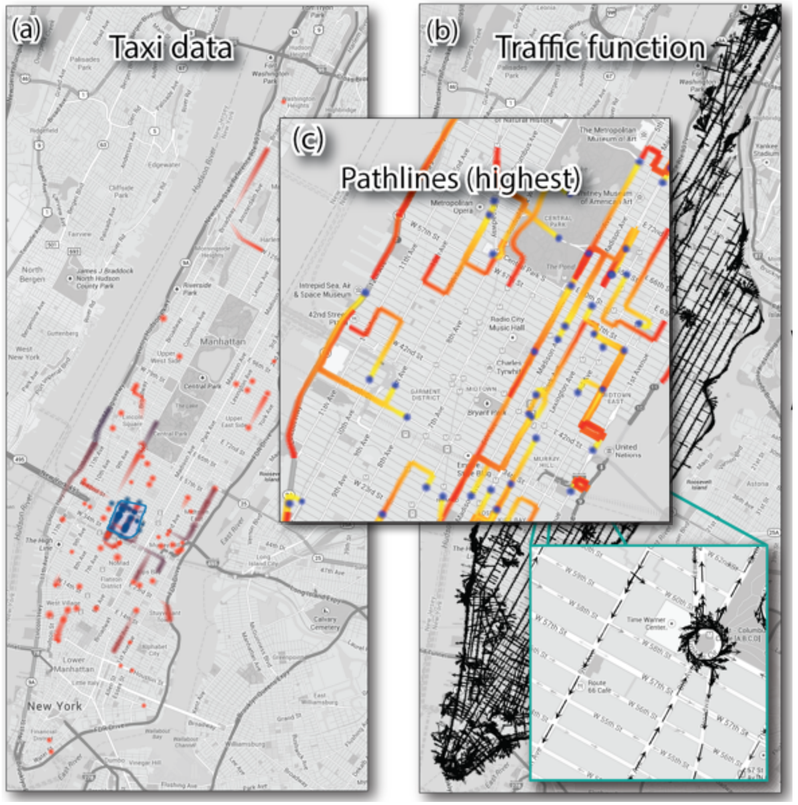
Find-a-Cab App
mobile

[iOS prototype]



Ride-sharing Simulation
at scale

[in submission]



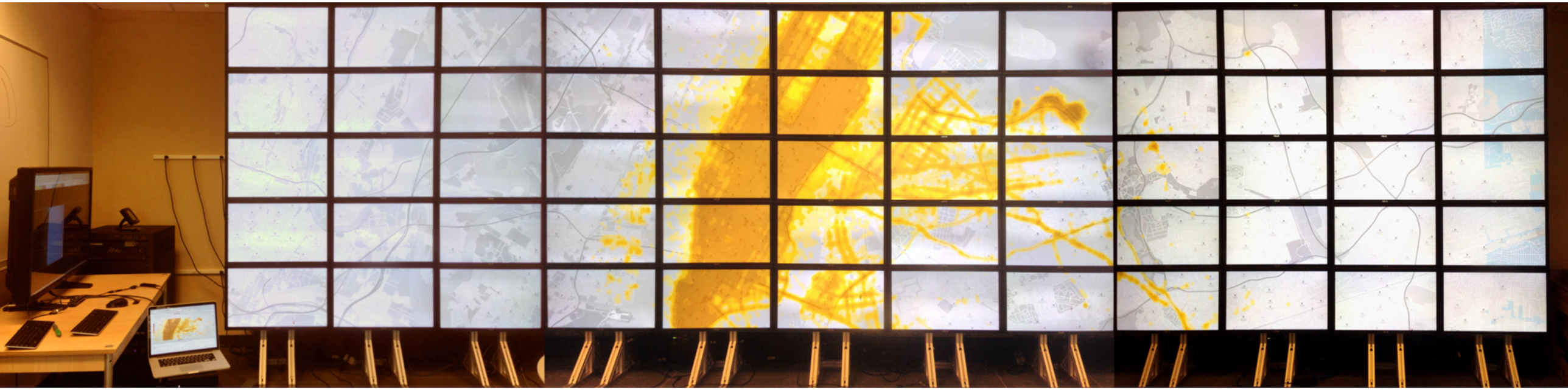
Exploring Traffic Dynamics
vector field visualization

[to appear EuroVis 2015]

Future Direction: A Scalable 3D Urban GIS Platform

- Urban science needs 3D visualization: building informatics, sky view analysis, urban heat island
  - InfoVis < “UrbanVis” < SciVis
- Bottlenecks in current urban visualization engines (aka. “mapping”):
  - Limited 3D support — do not provide 3D API, mostly eye candy
  - Scalable mapping frameworks are largely tailored for rendering purposes — no data integration and selection
  - Not scalable for large displays, including 4k monitors or display walls

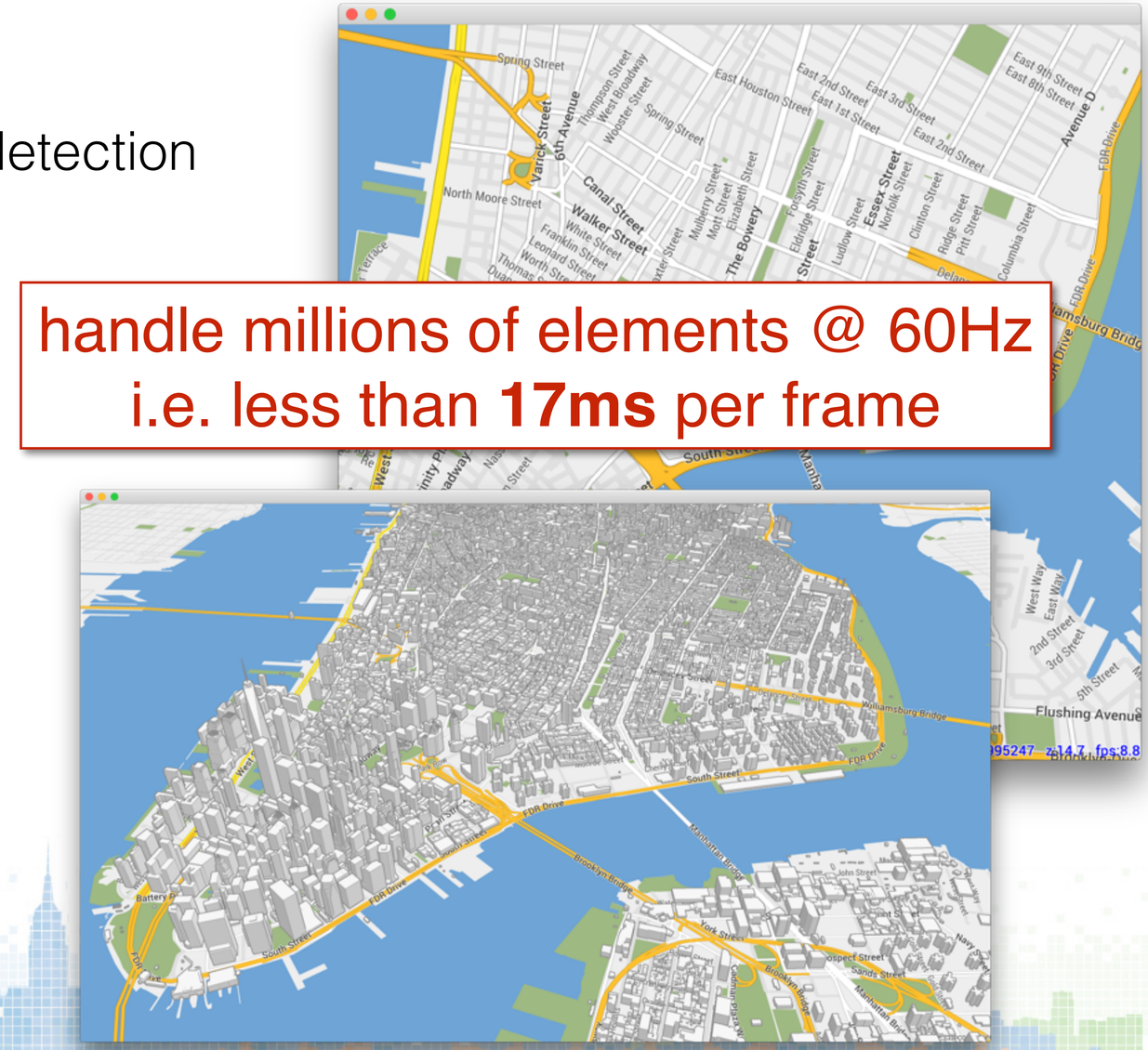
Future Direction: A Scalable 3D Urban GIS Platform



60 screens at 4K resolution each (4 times 1080p)

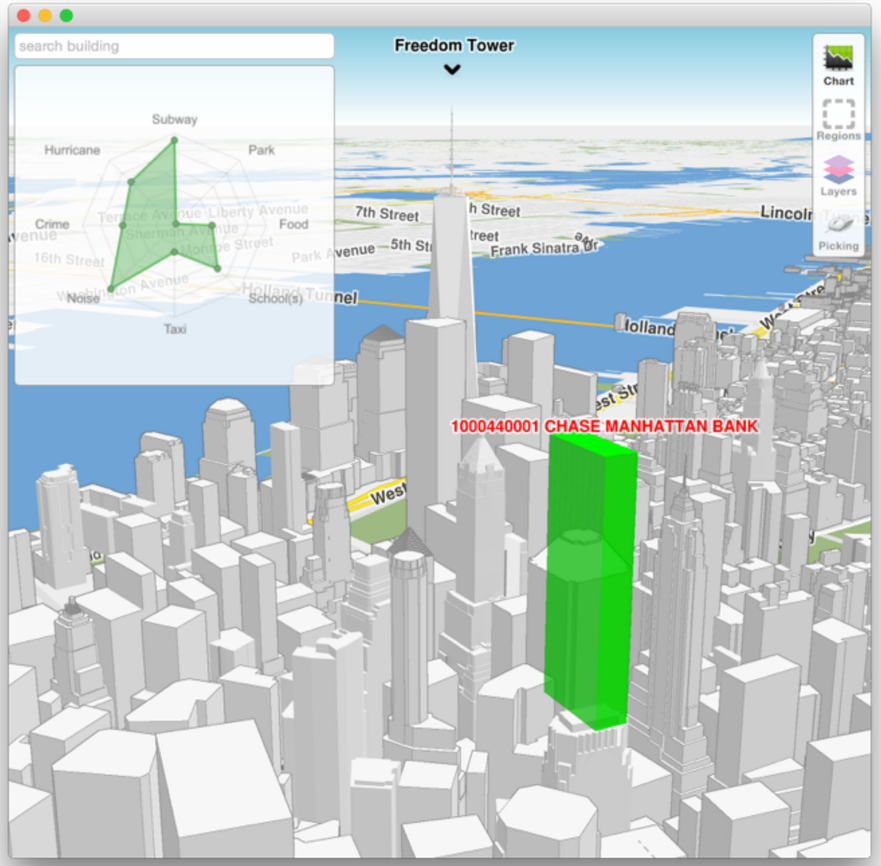
The **scalable rendering techniques** are exciting…

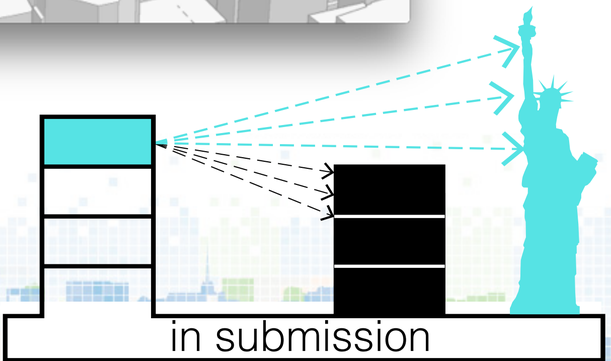
- Text Rendering
  - Dynamic label placement — collision detection
  - Compact glyphs representation
- Implicit Modeling: roads, buildings
  - Tessellation shader
- Screen-space Rendering Effects
  - Edge detection
- View-dependent Level-of-Details
  - Visibility test — occlusion culling



…but the real power is **data integration**!

- GIS data comes from multiple sources
  - Street networks — *OpenStreetMap (OSM)*
  - Crowd-source building models — *OSM*
  - High-detail building models — *proprietary*
- Flexibility in adding data layers
  - Building information — *NYC PLUTO*
  - Neighborhood profiles — *NYC Open Data*
- Linking data across layers: towards **a full integrated information system**
  - Building selection — *matching PLUTO against others*





UrbanGIS

Conclusions

- Analyzing big data requires exploration — not just confirmatory tasks
  - Needs to reach non-expert users
  - The ability to integrate multiple data sources
- (Interactive) Visualization is a powerful tool for data exploration
  - requires synergy with database/big data management
- Interdisciplinary collaborations is critical!

Acknowledgments

- Joint work with many wonderful colleagues at NYU:
  - Prof. Juliana Freire and Prof. Claudio Silva
  - Dr. Harish Doraiswamy, Dr. Marcos Lage, Dr. Tim Savage
  - Fernando Chirigati, Nivan Ferreira, Tuan-Anh Hoang-Vu, Fabio Miranda, Masayo Otta, Kien Pham, Jorge Poco, Wendel Silva
- … and researchers at KPF:
  - Mu Chan Park, Heidi Werner, Luc Wilson

CENTER FOR URBAN
SCIENCE+PROGRESS

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

Thank You

cusp.nyu.edu

NYUCUSP

@NYU_CUSP