Exploring Big Urban Data

and confessions of a plumber

Juliana Freire
Computer Science & Engineering
Visualization and Data Analysis (ViDA)
Center for Data Science (CDS)
Center for Urban Science and Progress (CUSP)

Joint work with Cláudio Silva, Huy Vo, Harish Doraiswami,
Fernando Chirigati, Theo Damoulas, Nivan Ferreira,
Masayo Otta, Kien Pham, Jorge Poco, Luciano Barbosa, Marcos Vieira

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

CUSP
CENTER FOR URBAN SCIENCE+PROGRESS

Our Research at the ViDA Center

- Empower a *wide range of users* to *explore* the vast repositories of urban data
  - Data-savvy analysts, domain experts, policy makers and citizens
- Address issues at the different stages of the data lifecycle
- Key ingredients (that we work on)
  - Finding information on the Web, hidden, dark and surface
  - Information integration
  - Data analysis
  - Visualization and visual analytics
  - Data and provenance management
- Focus on usability – tools must be powerful and easy to use
- From concepts and algorithms to deployed systems

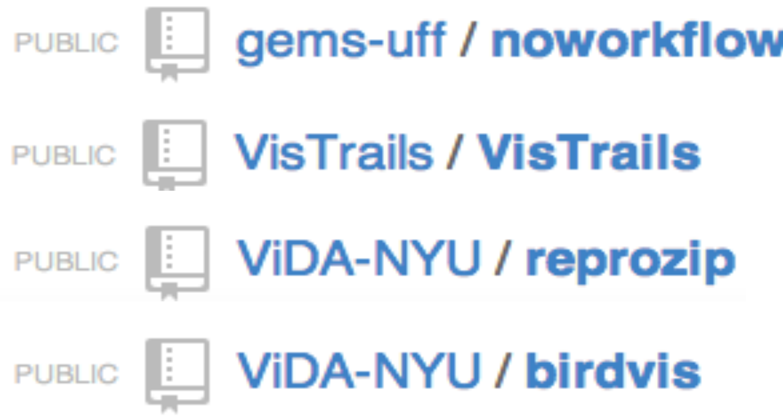
Visualization and Data Analysis @NYU

50+ papers
(since 2011)

Over 30 active members

- 3 TT faculty members
- 3 Research faculty
- 3 Postdocs
- 2 Research engineers
- 15+ PhD students
- Constant stream of visiting researchers and students

<u>Over 50 people since 2011</u>



PUBLIC gems-uff / noworkflow

PUBLIC VisTrails / VisTrails

PUBLIC ViDA-NYU / reprozip

PUBLIC ViDA-NYU / birdvis

Moore-Sloan Data Science Initiative

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

NYU

Berkeley UNIVERSITY OF CALIFORNIA

W UNIVERSITY *of* WASHINGTON

A 5-year, $37.8 million cross-institutional collaboration

Moore-Sloan Data Science Initiative

At a time when … sciences are all producing data with relentlessly increasing volume, variety and velocity, capturing the full potential of a progressively data-rich world has become a daunting hurdle for researchers. …, data (intensive) science is already contributing to scientific discovery, yet substantial systemic challenges need to be overcome to maximize its impact on academic research.

http://www.moore.org/programs/science/data-driven-discovery/data-science-environments

Moore-Sloan Data Science Initiative

- Develop meaningful and sustained interactions and collaborations between researchers from different disciplines to move science forward;
- Establish career paths that are long-term and sustainable to retain scientists whose research focuses on the multi-disciplinary research and the development of the tools that enable this research; and
- Build an ecosystem of analytical tools and research practices that is sustainable, reusable, extensible, learnable, easy to translate across research areas and enables researchers to spend more time focusing on their science.

Do breakthrough science, and enable breakthrough science

Moore-Sloan Data Science Initiative

- Do breakthrough science
  - In scientific theme areas
  - In data science methodology areas
  - Establish a research program in the science of data science

- Enable breakthrough science
  - Through new tools and methods
  - Through changing the process of discovery and driving cultural changes
  - Foster the adoption of reproducible research
  - New career paths for researchers at universities
  - Establish graduate programs in data science

- Establish a “virtuous cycle”

Join us!

- If you are passionate about data,
- Like challenges, and
- Want to have real impact

- Open positions:
  - Post-docs
  - Research engineers
  - Programmers

- And we always welcome great students!

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/

Students: Admissions Summary, MS Applied Urban Science

Cycle Dates: December 18, 2012 through June 30, 2013 (~6months)

24
Inaugural Class

21%
Selectivity

27
Years
Average Age

36%
Female

3.5
Average
Undergraduate GPA

20
Undergraduate
Disciplines

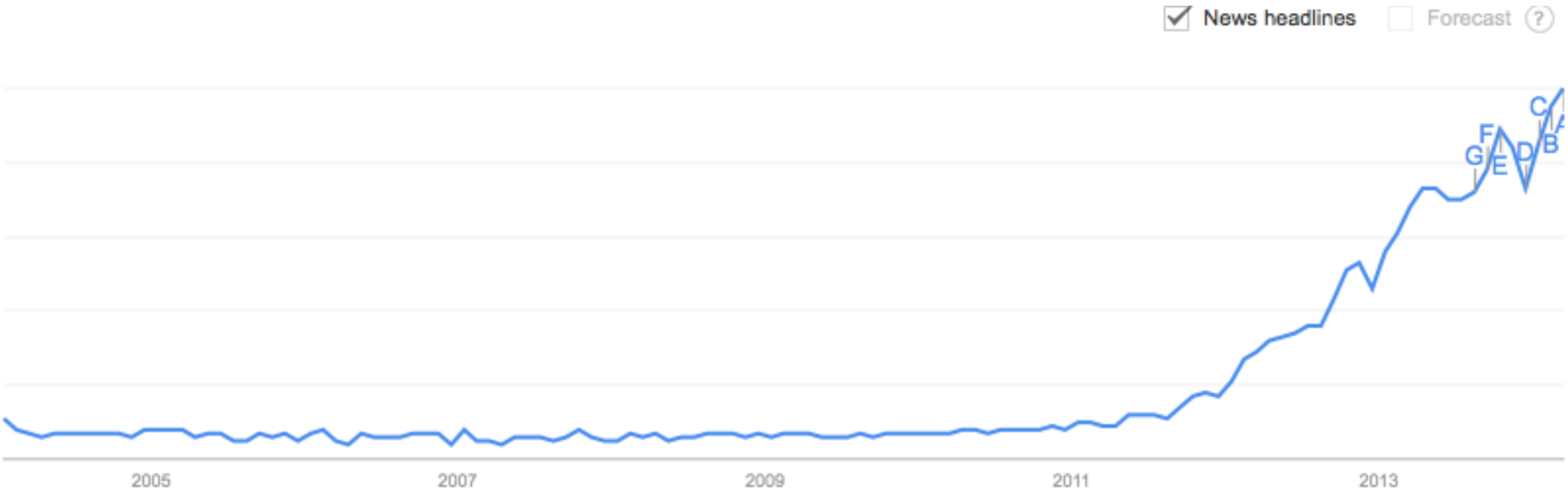
48%
International

9
Countries Represented

4
Years Average
Work Experience

28%
With Graduate Degree

Big Data: What is the **Big** deal?



http://www.google.com/trends/explore#q=%22big%20data%22

Big Data: What is the **Big** deal?

http:/ g%20data%22

Big Urban Data: What is the **Big** deal?

- Cities are the loci of economic activity
- 50% of the world population lives in cities --- by 2050 the number will grow to 70%
- Growth leads to problems, e.g., transportation, environment and pollution, housing
- Good news: Lots of data are being collected from traditional and *unsuspecting* sensors
  - Census, crime, emergency visits, taxis, public transportation, real estate, noise, energy, Twitter, …

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

Big Urban Data: Success Stories

OneBusAway

Serving up fresh real-time transit information for the ☐ region.

http://onebusaway.org

- Real-time arrival predictions
- 94% reported increased or greatly increased satisfaction with public transit
- Significant decrease in actual wait time per user, and an even greater decrease in *perceived* wait time
- 78% of riders reported increased walking --- a significant public health benefit

Big Urban Data: Success Stories

- Michael Flowers @ NYC

New York City gets 25,000 illegal-conversion complaints a year, but it has only 200 inspectors to handle them.

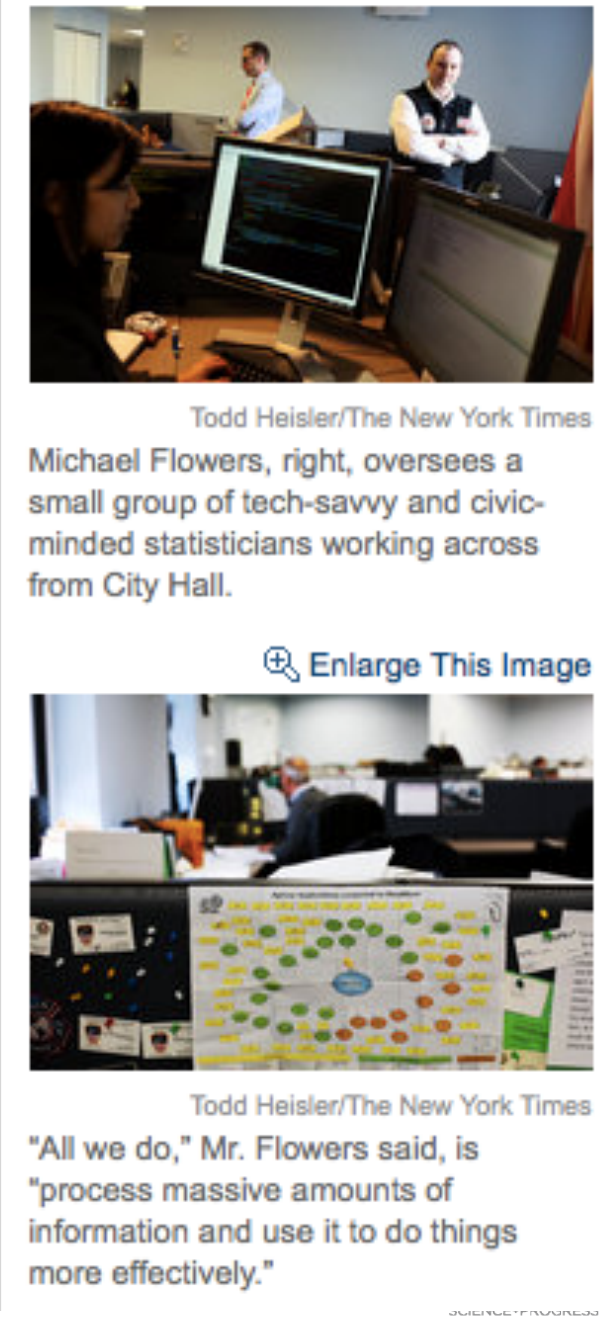
Flowers' group: (1) integrated information from 19 different agencies that provided indication of issues in buildings

- E.g., Late taxes, foreclosure proceedings, service cuts, ambulance visits, rodent infestation, crime

(2) Compared with 5 years of fire data

(3) Created a prediction system

Result: hit rate for inspections went from 13% to 70%



Todd Heisler/The New York Times

Michael Flowers, right, oversees a small group of tech-savvy and civic-minded statisticians working across from City Hall.

Todd Heisler/The New York Times

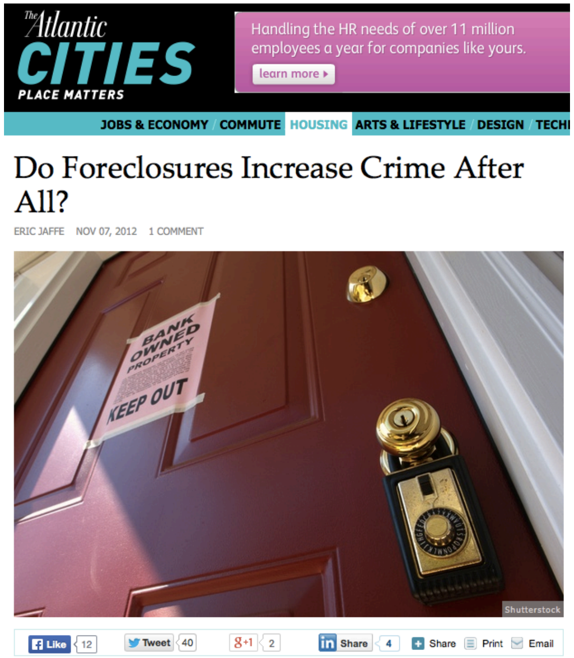
"All we do," Mr. Flowers said, is "process massive amounts of information and use it to do things more effectively."

SCIENCE+PROGRESS

Big Urban Data: Success Stories

- The NYU Furman Center
  - Analysis of the impact and benefits of subsidized housing on the surrounding neighborhoods → influenced city spending decisions
  - Assessment of crime data and property-level foreclosure data led to the finding that neighborhoods with concentrated foreclosures see an uptick in crime for each foreclosure notice issued → updates to policing strategies

FURMAN CENTER
FOR REAL ESTATE & URBAN POLICY
NEW YORK UNIVERSITY
SCHOOL OF LAW • WAGNER SCHOOL OF PUBLIC SERVICE



Do Foreclosures Increase Crime After All?

ERIC JAFFE NOV 07, 2012 1 COMMENT

http://furmancenter.org/

Big Urban Data: What is hard?

Infrastructure

Condition, operations

Environment

Meteorology, pollution, noise, flora, fauna

People

Relationships, economic activities, health, nutrition, opinions, …

- City components interact in complex ways
- Need to look at the city *data exhaust* to understand these interactions

DATA.GOV
EMPOWERING PEOPLE

NYC OpenData

twitter

facebook

YouTube
Broadcast Yourself

data.gov.in
Open Government Data Platform India

सत्यमेव जयते

BRASIL Acesso à informação

Ir para o conteúdo 1 Ir para o menu 2 Ir para a busca 3 Ir para o rodapé 4

Dados Abertos
GOVERNO FEDERAL

CUSP
CENTER FOR URBAN SCIENCE+PROGRESS

Big Urban Data: What is hard?

Environment

Meteorology, pollution, noise, flora, fauna

People

city = program

Relationships, economic activities, health, nutrition, opinions, …

Infrastructure

Condition, operations

?

DATA.GOV
EMPOWERING PEOPLE

NYC OpenData

twitter

data = output

be
urself

data.gov.in
Open Government Data Platform India
सत्यमेव जयते

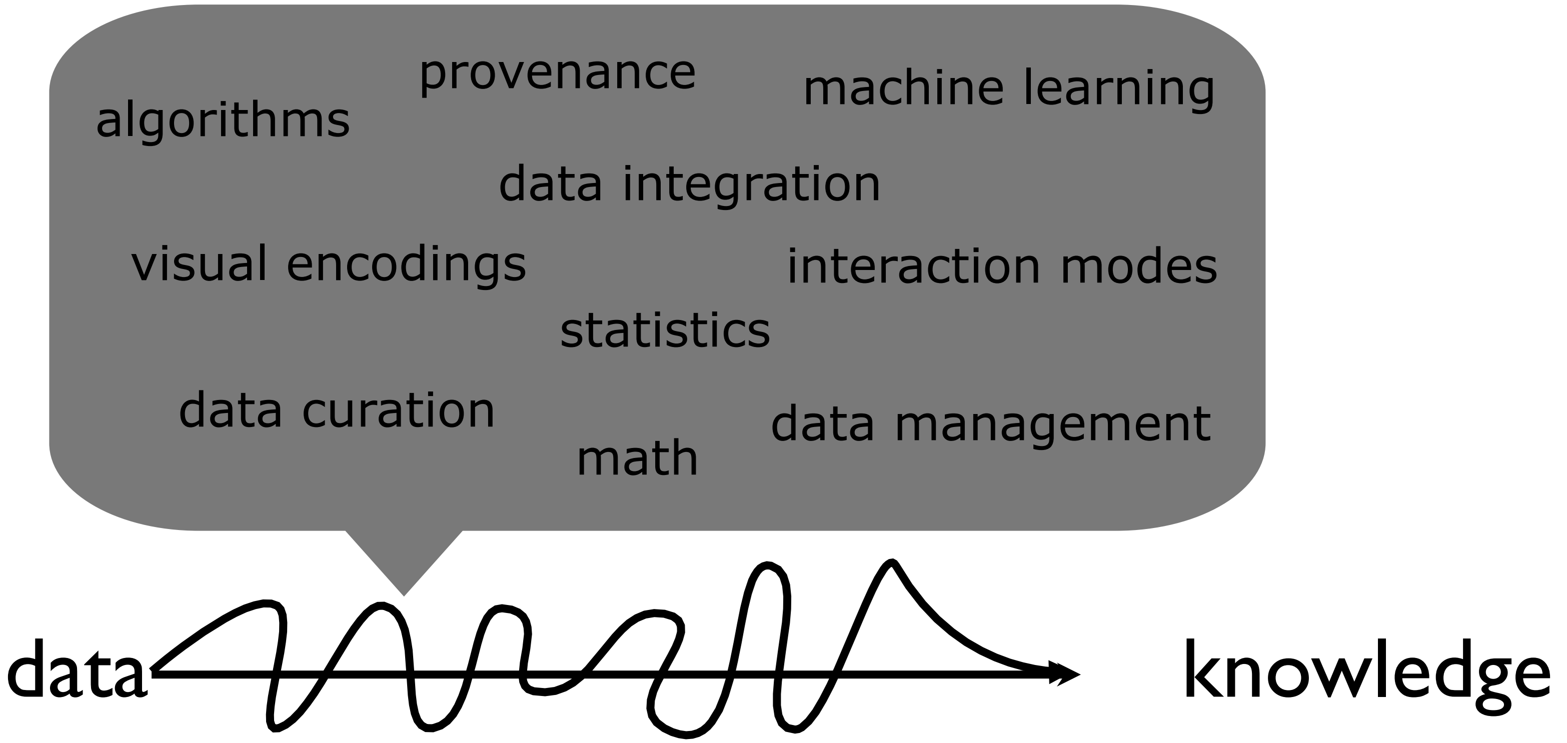
BRASIL Acesso à informação

Ir para o conteúdo 1 Ir para o menu 2 Ir para a busca 3 Ir para o rodapé 4

Dados Abertos
GOVERNO FEDERAL

Big Urban Data: What is hard?

- Scalability for batch computations is *not* the biggest problem
  - Lots of work on distributed systems, parallel databases, …
  - Elasticity: Add more nodes!
- Scalability for people is! regardless of whether data are big or small



algorithms

provenance

machine learning

data integration

visual encodings

interaction modes

statistics

data curation

math

data management

data

knowledge

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

- Ask questions about the *past, present and future*

Outline

- A short story …
- Understanding the past and present
- Predicting the future
- Confessions
- Conclusion

Getting Data about NYC Taxi Trips

http://vimeo.com/31298658

How can I get the taxi data

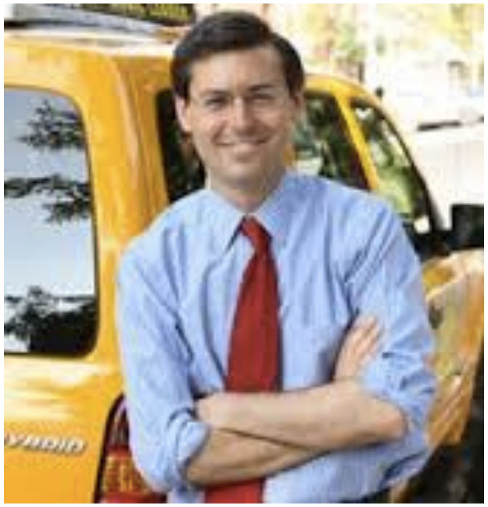
I don't know, ask Mitch

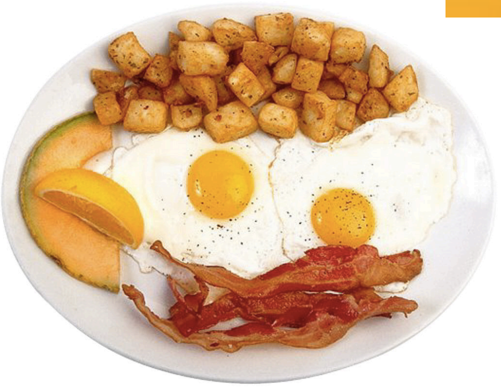
David Yassky will speak at the Rudin Center!

Getting Data about NYC Taxi Trips

Can I have the taxi data?

Sure, bring a HD!





Getting Data about NYC Taxi Trips

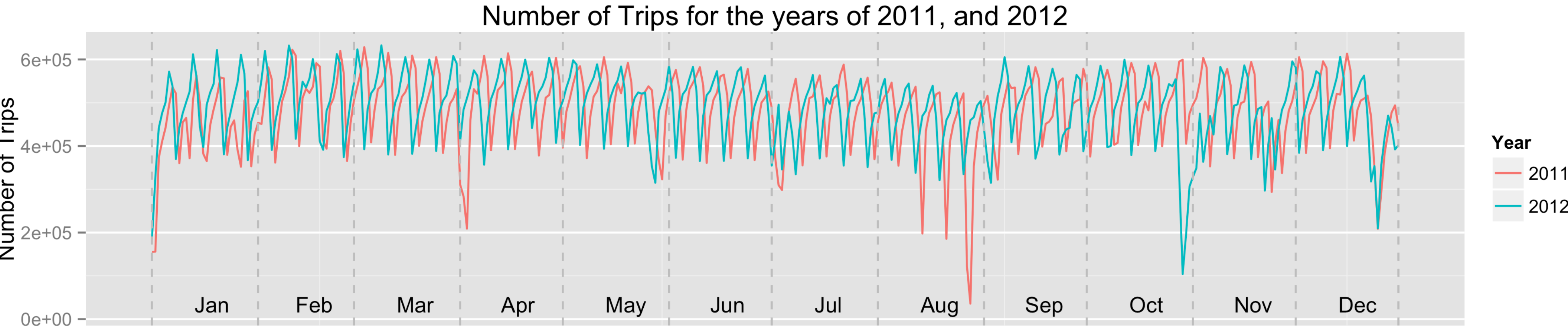
NYC
Taxi and Limousine Commission

NYC Taxi & Limousine Commission
Rector St
Wall St
Broad St
Maiden Ln
South Street Seaport
Brooklyn Bridge
Bowling Green
DTCC Deriv/Serv LLC
Whitehall St
South Ferry Loop
M 15 min every 4 min
M 16 min every 6 min
M 12 min every 10 min
M 15 min every 5 min
East River
Brooklyn Bridge
Water St
DUMBO
John St
Plymouth St
Front St
Pearl St
Brooklyn Queens Expy
York St
Middagh St
Cranberry St
Orange St
High St
Willow St
Cadman Plaza E
Hugh L. Carey Tunnel (Toll road)
Dimond Reef
Brooklyn Queens Expy
Jay St
Adams St
Grace Ct
Remsen St
Court St
Joralemon St
Borough Hall
Borough Hall
Willow Pl
Sidney Pl
Jay St - MetroTech
Jay St
2 MetroTech
Colonie
Hoyt St
Carder Rd
278
478
York St
M
F

And the ViDA Center lived happily ever after…

Understanding the Past and Present

Exploring Urban Data: NYC Taxis



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

  "How the taxi fleet activity varies during weekdays?"

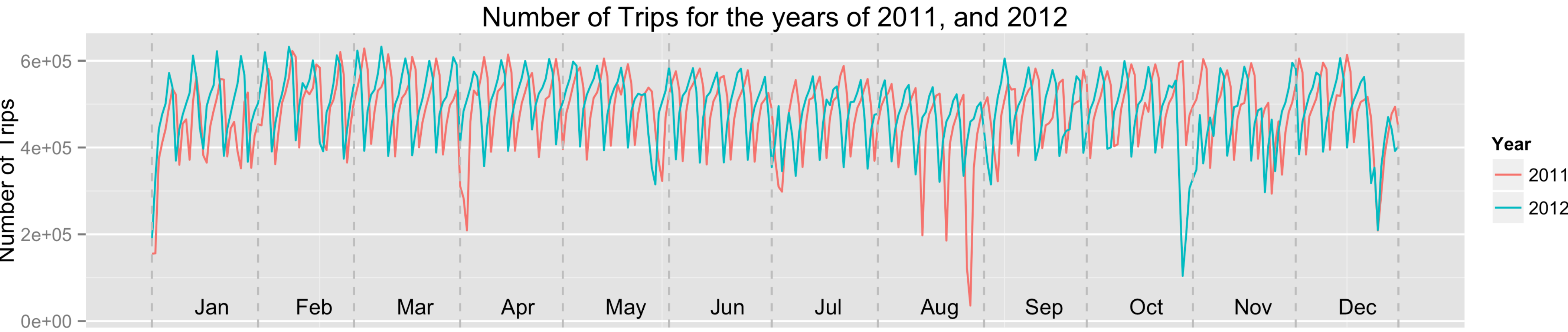
  "What is the average trip time from Midtown to the airports during weekdays?''

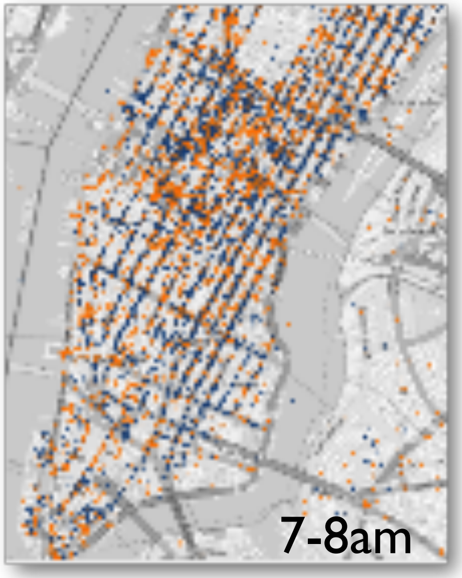
  "How was activity in Midtown affected during a presidential visit?''

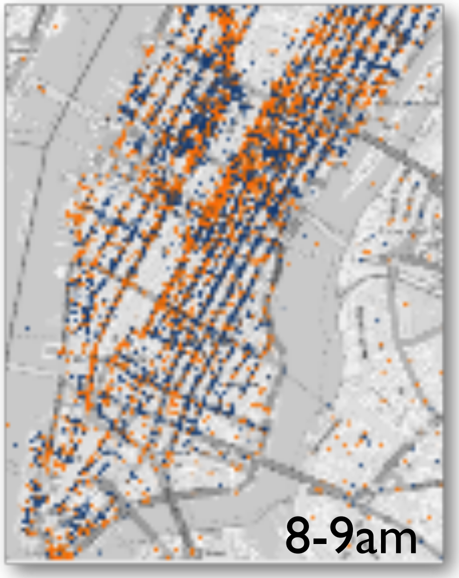
  "How did the movement patterns change during Sandy?"

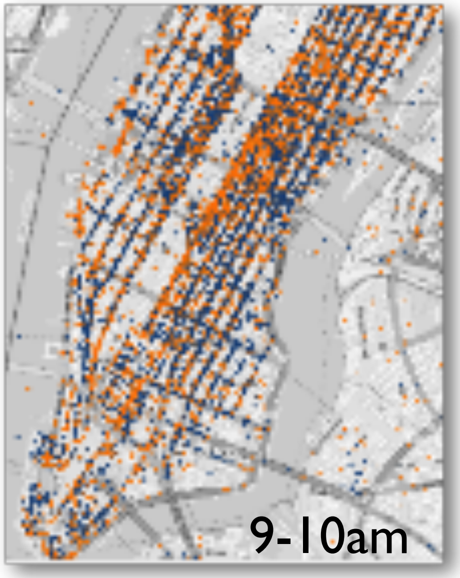
  "Where are the popular night spots?"

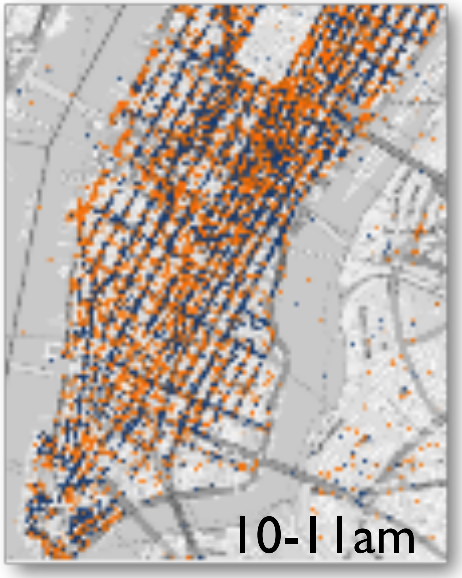
Exploring Urban Data: NYC Taxis











Urban Data Analysis

- Common practice:
  - Domain scientists and policy makers formulate hypotheses
  - Data scientists select data slices, perform analyses, and derive plots
- Issues:
  - Analyses are mostly confirmatory (Tukey, 1977) -- batch-oriented analysis pipeline hampers exploration
  - Data are complex -- often multivariate spatio-temporal
  - Queries are expensive
  - Tools are not scalable, e.g., Excel, GIS, SAS, …
  - Dependency on data specialists distances domain experts from the data

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
- Data are *complex*:
  - *spatio-temporal:* pick up + drop off
  - *trip attributes*: e.g., distance traveled, cost, tip
- Too many data slices to examine
- Our goals: Design a *usable* interface, efficiently support *interactive* + *exploratory* queries

TaxiVis: Visually Exploring NYC Taxi Data

- New model that allows users to visually query taxi trips, easily select and compare different spatial-temporal slices
  - Data selection through visual manipulations
  - Use visualization to explore selected data
- Support for origin-destination queries that enable the study of mobility across the city
- Use multiple coordinated views to allow comparisons, and brushing to support query refinements
- Use of adaptive level-of-detail rendering and heat maps to generate clutter-free visualization for large results
- Scalable system that provides interactive response times for spatio-temporal queries over large data

[Ferreira et al., IEEE TVCG 2013]

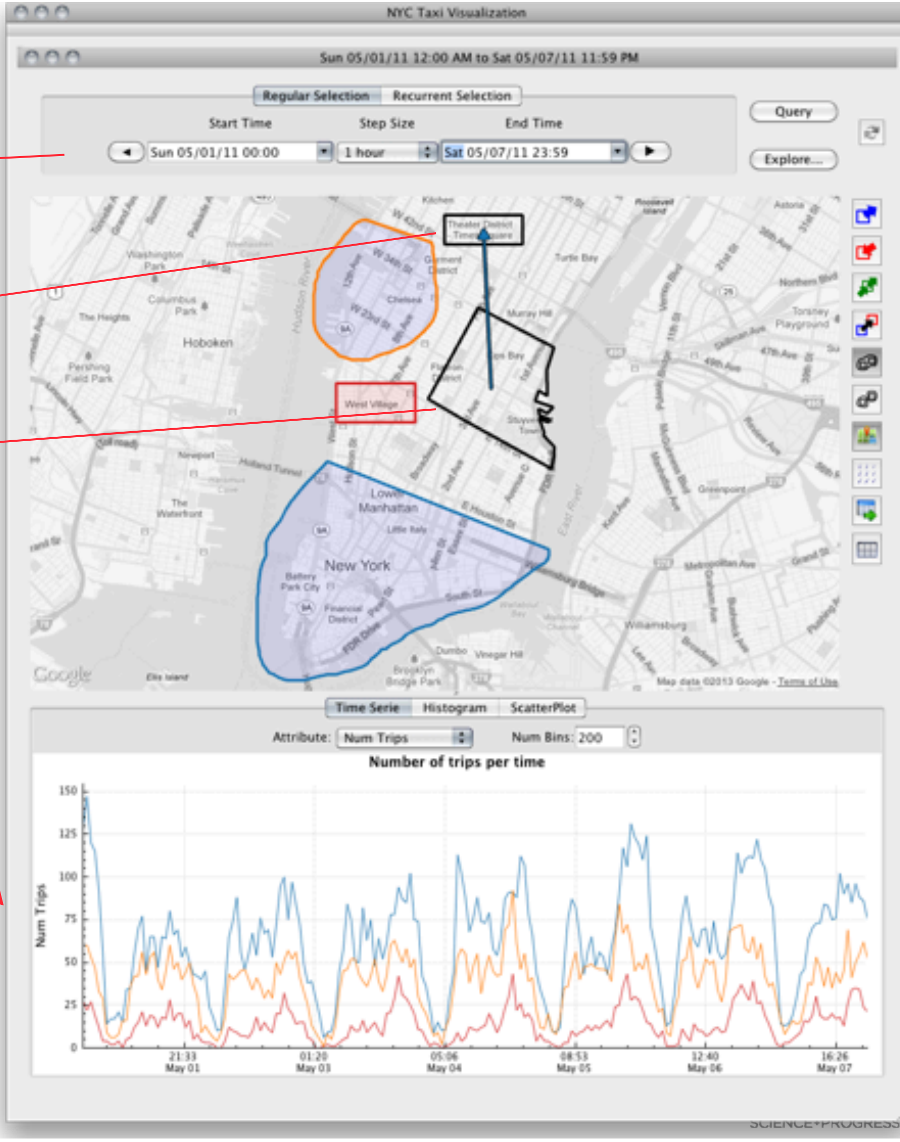
TaxiVis in Action

Usability through Visual Operations

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

```
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



NYC Taxi Visualization

Sun 05/01/11 12:00 AM to Sat 05/07/11 11:59 PM

Regular Selection | Recurrent Selection

Start Time: Sun 05/01/11 00:00

Step Size: 1 hour

End Time: Sat 05/07/11 23:59

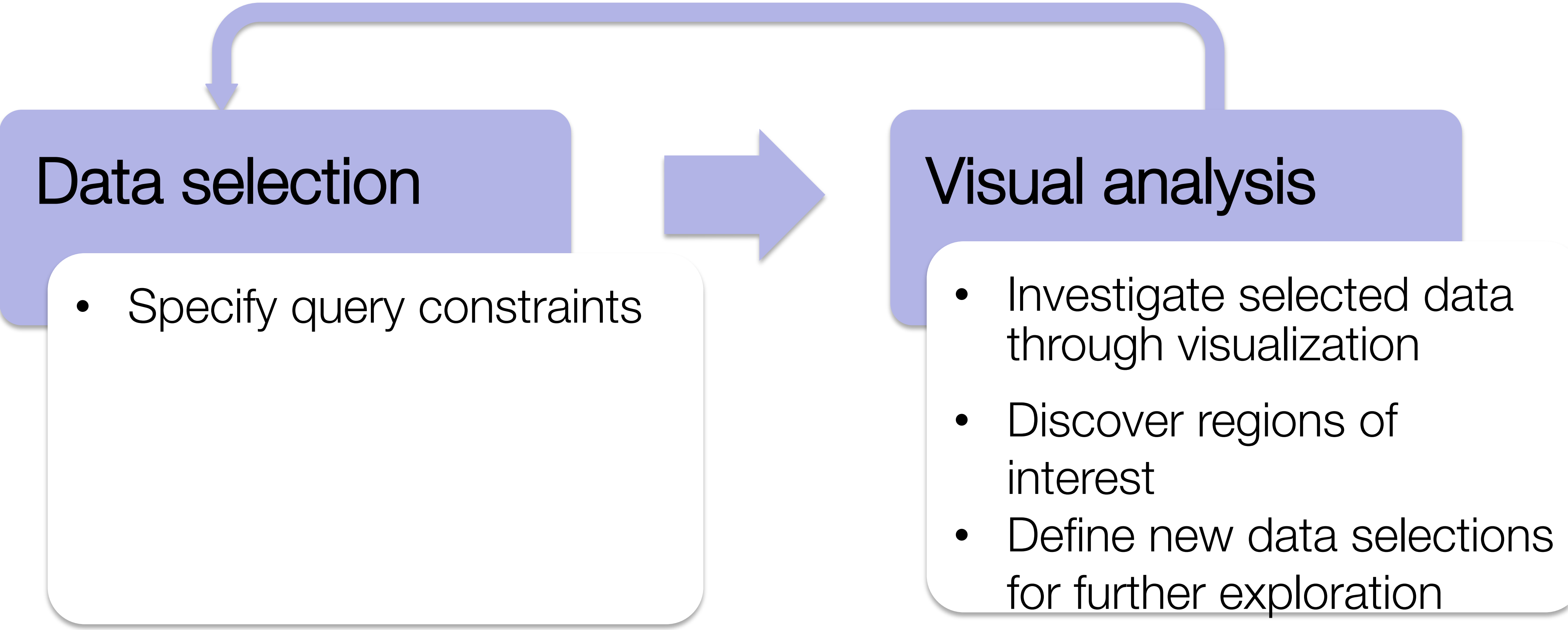
Time Serie | Histogram | ScatterPlot

Attribute: Num Trips

Num Bins: 200

Number of trips per time

Data Exploration: A Two-Phase Process



Data selection

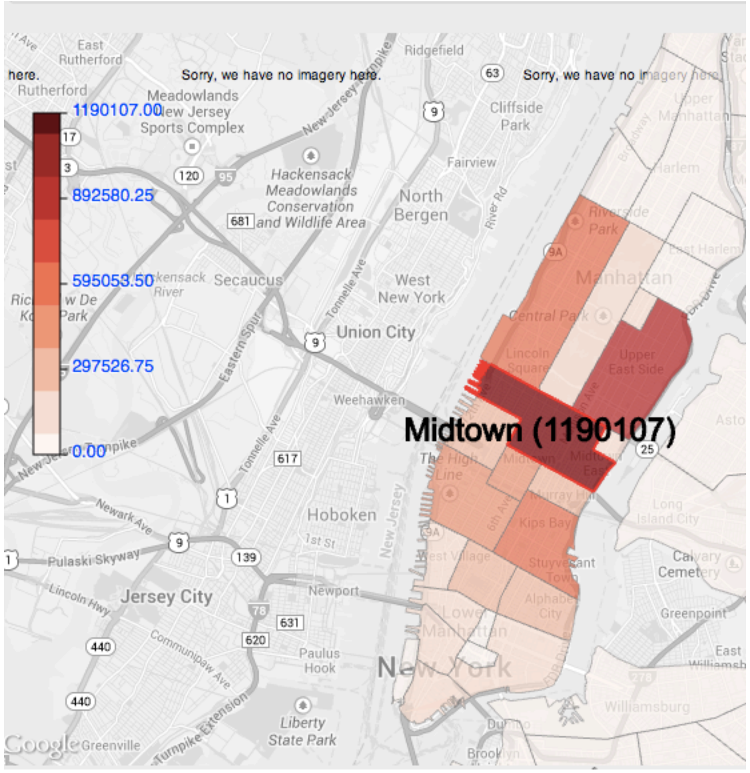
- Specify query constraints

Visual analysis

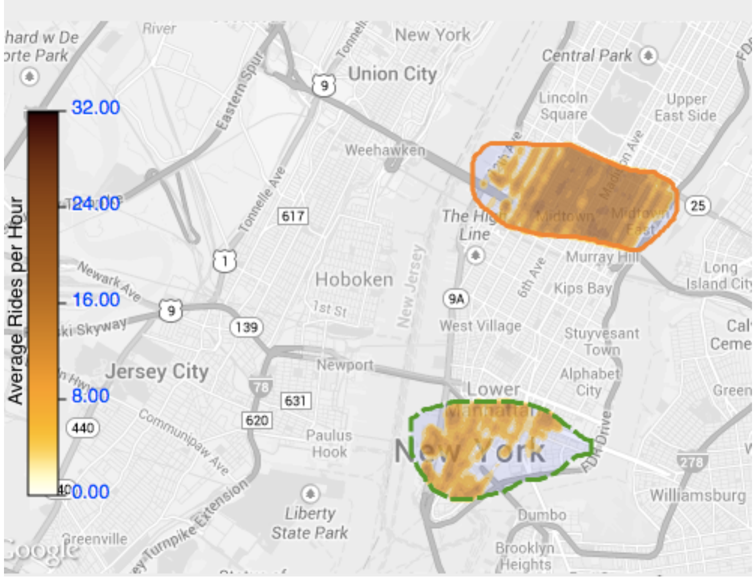
- Investigate selected data through visualization
- Discover regions of interest
- Define new data selections for further exploration

TaxiVis unifies the two phases of the process through visual operations

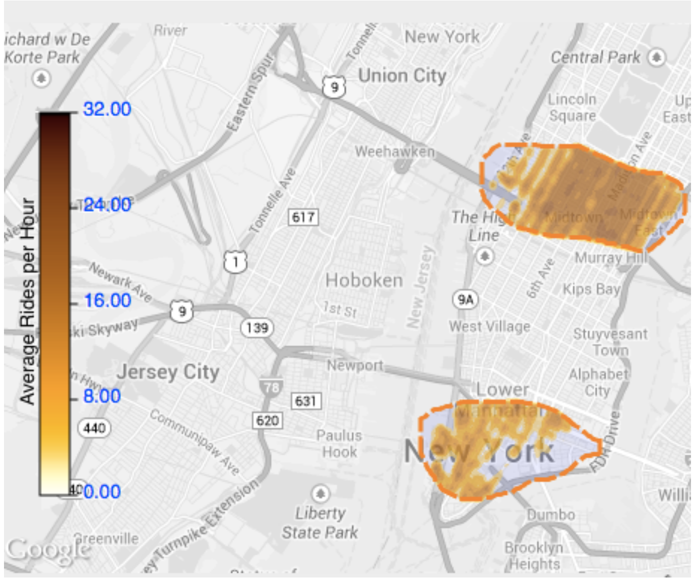
Selecting Regions – Spatial Constraints



Predefined polygons, e.g., zip, neighborhoods, etc



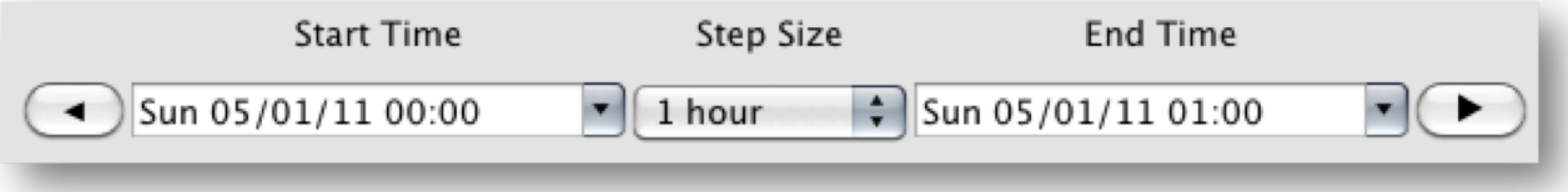
Free selection



Group regions

Selecting Time – Temporal Constraints

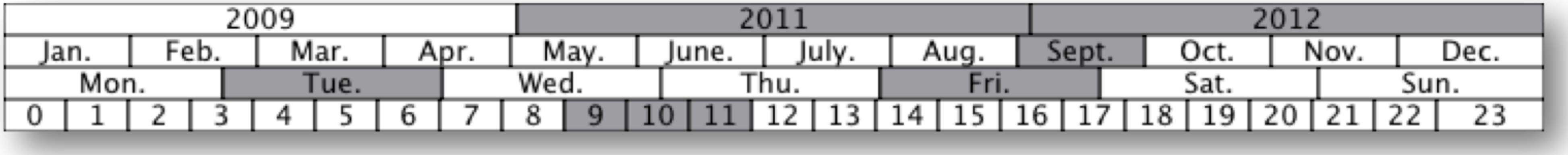
Time interval



Start Time: Sun 05/01/11 00:00

Step Size: 1 hour

End Time: Sun 05/01/11 01:00



2009 | 2011 | 2012

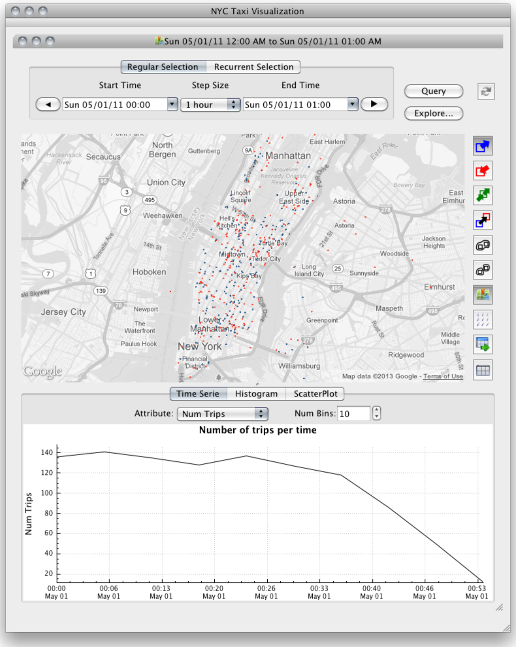
Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec.

Mon. | Tue. | Wed. | Thu. | Fri. | Sat. | Sun.

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23

Recurrent time patterns

Selecting Attributes



NYC Taxi Visualization

Sun 05/01/11 12:00 AM to Sun 05/01/11 01:00 AM

Start Time: Sun 05/01/11 00:00

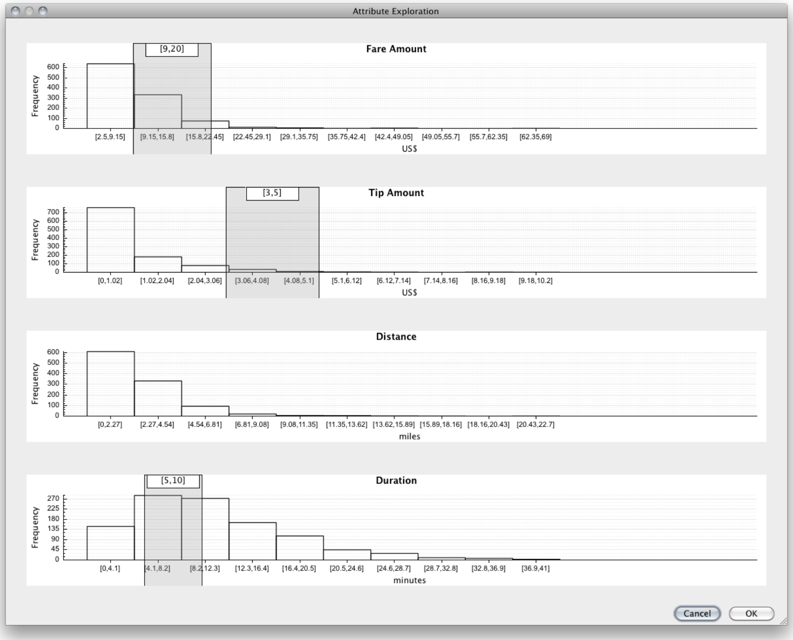
Step Size: 1 hour

End Time: Sun 05/01/11 01:00

Attribute: Num Trips

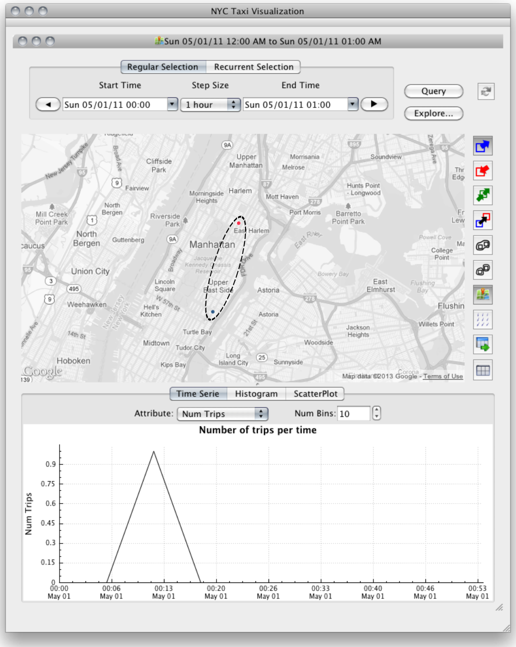
Num Bins: 10

+



Attribute Exploration

=



NYC Taxi Visualization

Sun 05/01/11 12:00 AM to Sun 05/01/11 01:00 AM

Start Time: Sun 05/01/11 00:00

Step Size: 1 hour

End Time: Sun 05/01/11 01:00

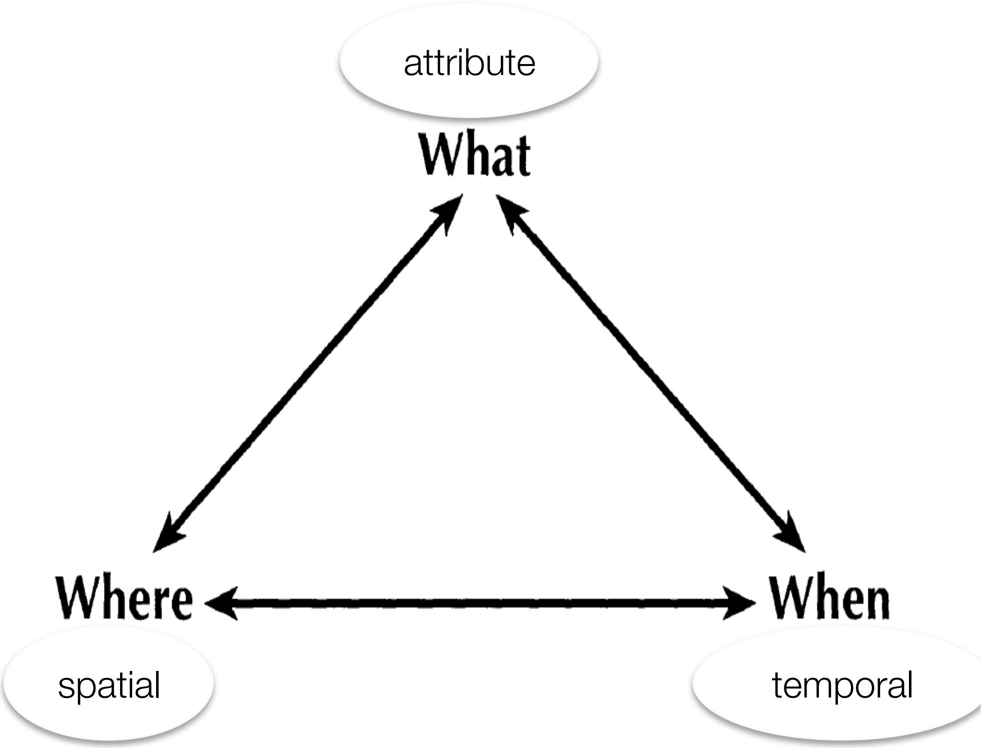
Attribute: Num Trips

Num Bins: 10

Peuquet's Triad for Spatio-Temporal Data

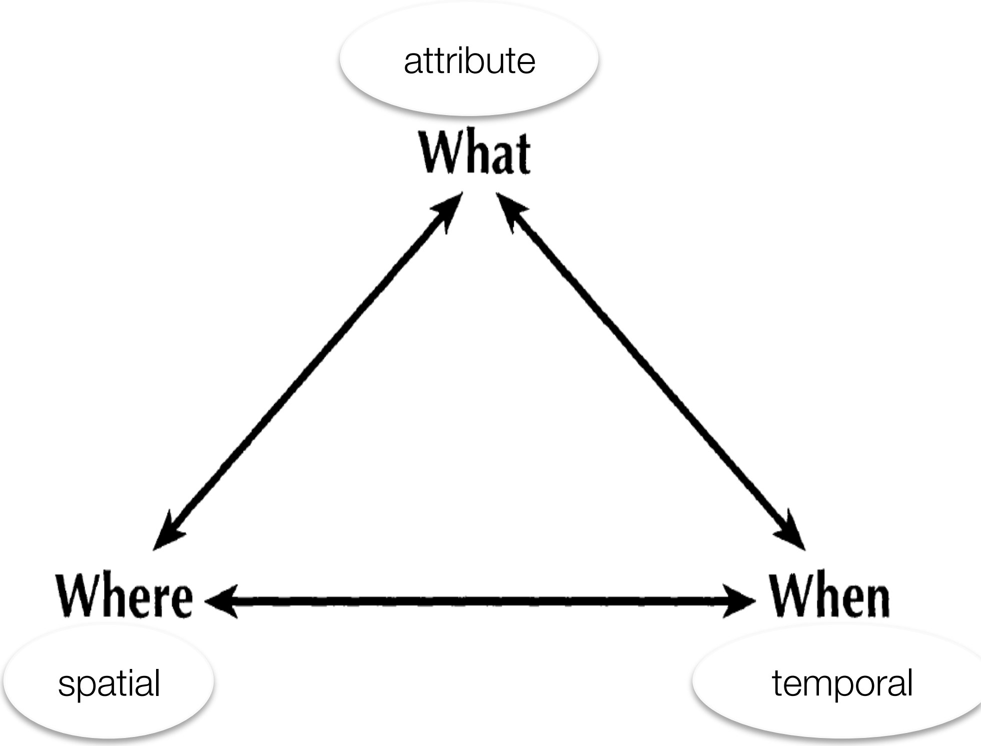
Classes of questions:

- 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?"*



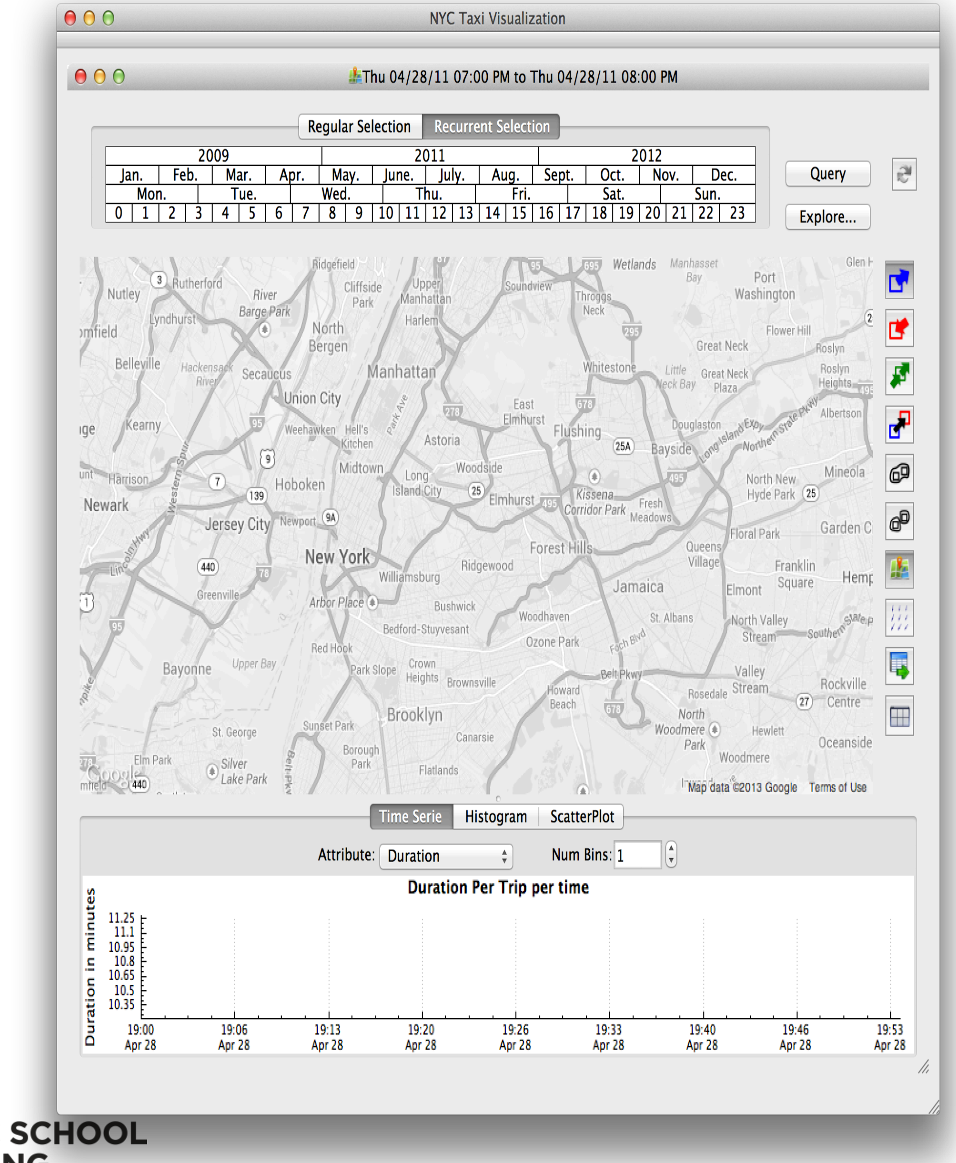
Query Expressiveness

- Our models supports these 3 classes of queries
  - when + where → what
  - when + what → where
  - where + what → when
- And more:
  - *when → what + where*
  - *where → when + what*
  - *what → where + when*



When + Where → What

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



NYC Taxi Visualization

Thu 04/28/11 07:00 PM to Thu 04/28/11 08:00 PM

Regular Selection | Recurrent Selection

2009 | 2011 | 2012

Jan. | Feb. | Mar. | Apr. | May. | June. | July. | Aug. | Sept. | Oct. | Nov. | Dec.

Mon. | Tue. | Wed. | Thu. | Fri. | Sat. | Sun.

0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23

Query

Explore...

Time Serie | Histogram | ScatterPlot

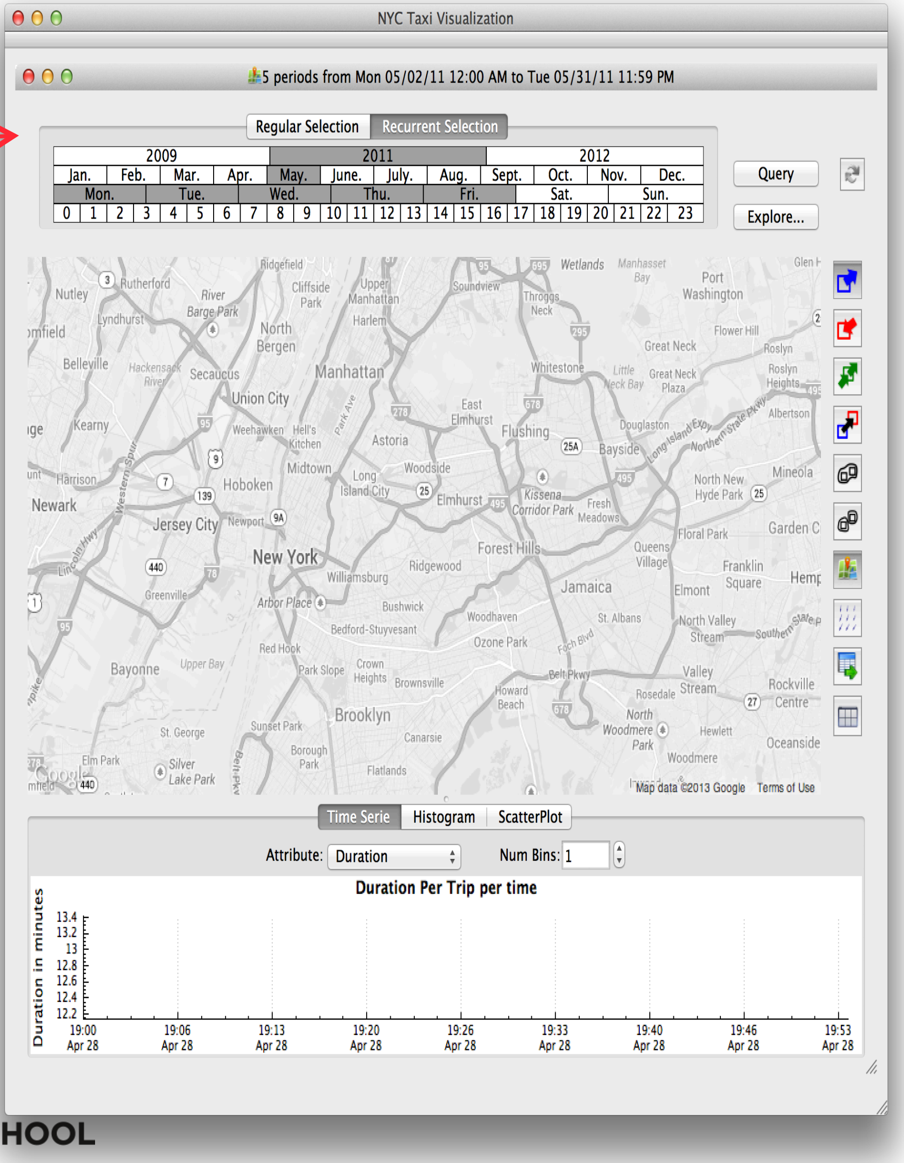
Attribute: Duration

Num Bins: 1

When + Where → What

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

When?



NYC Taxi Visualization

5 periods from Mon 05/02/11 12:00 AM to Tue 05/31/11 11:59 PM

Regular Selection | Recurrent Selection

2009	2011	2012

Jan.	Feb.	Mar.	Apr.	May.	June.	July.	Aug.	Sept.	Oct.	Nov.	Dec.

Mon.	Tue.	Wed.	Thu.	Fri.	Sat.	Sun.

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23

Time Serie | Histogram | ScatterPlot

Attribute: Duration

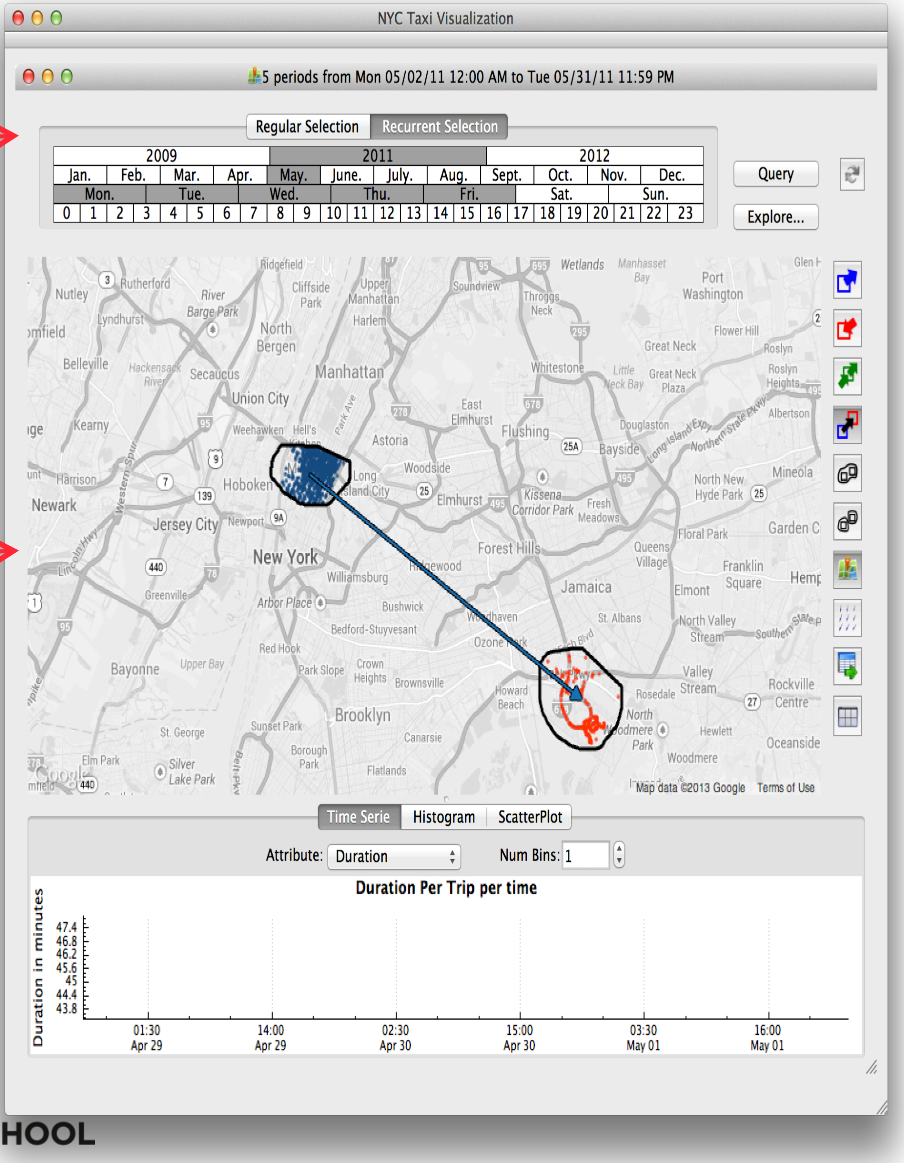
Num Bins: 1

When + Where → What

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

When?

Where?


NYC Taxi Visualization

5 periods from Mon 05/02/11 12:00 AM to Tue 05/31/11 11:59 PM

Attribute: Duration

Num Bins: 1

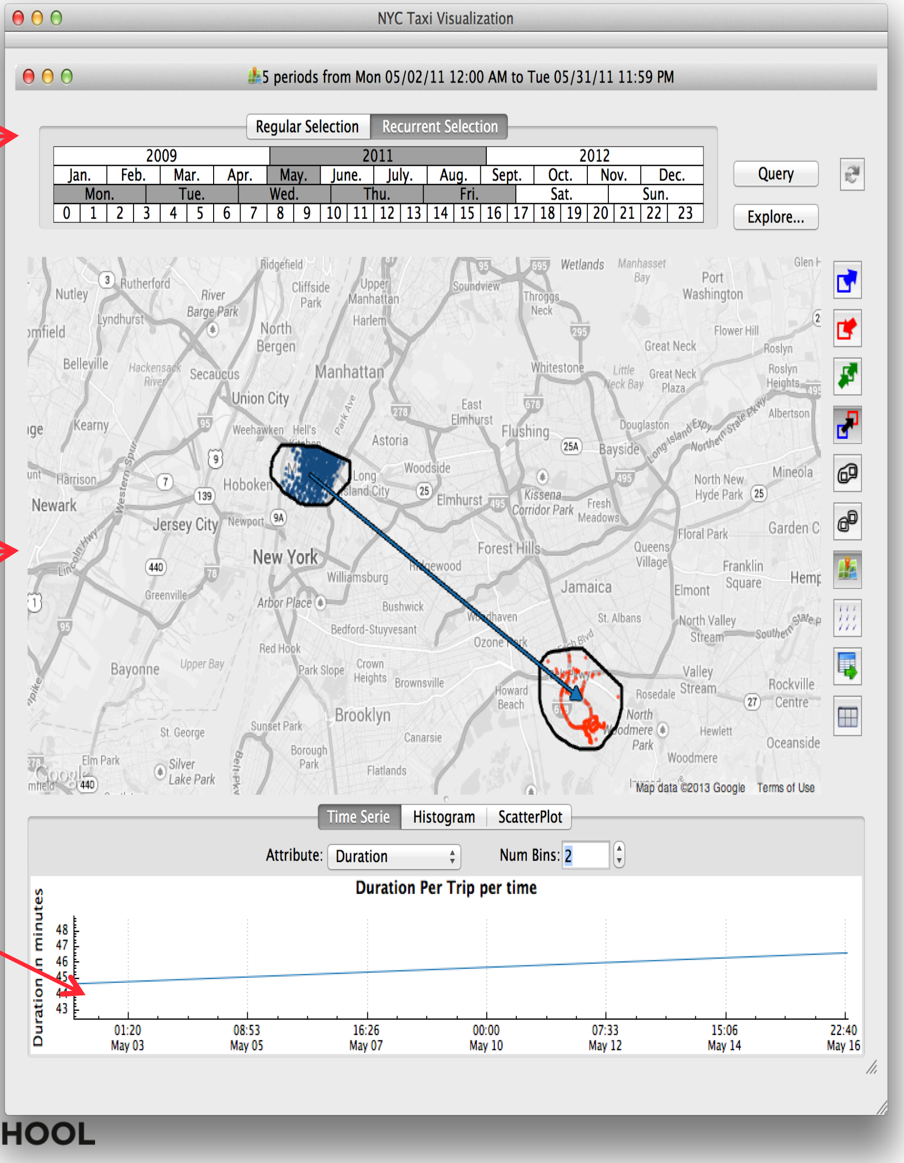
When + Where → What

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

When?

Where?

What



NYC Taxi Visualization

5 periods from Mon 05/02/11 12:00 AM to Tue 05/31/11 11:59 PM

Attribute: Duration

Num Bins: 2

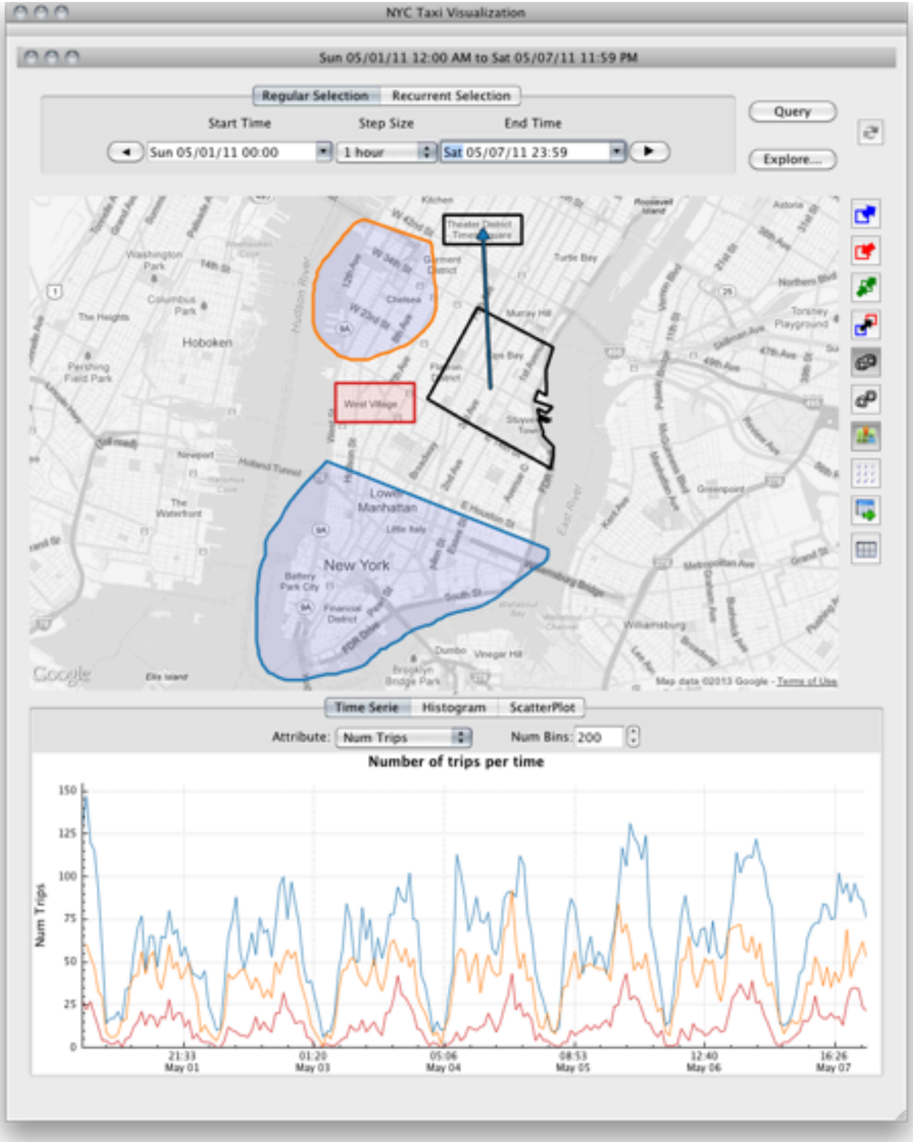
Duration Per Trip per time

Composing Queries

A query is associated with the set of trips contained in its results – queries can be composed.

Different visualizations can be applied to query results

Lines in plot are linked to the queries by their color.



NYC Taxi Visualization

Sun 05/01/11 12:00 AM to Sat 05/07/11 11:59 PM

Regular Selection | Recurrent Selection

Start Time: Sun 05/01/11 00:00

Step Size: 1 hour

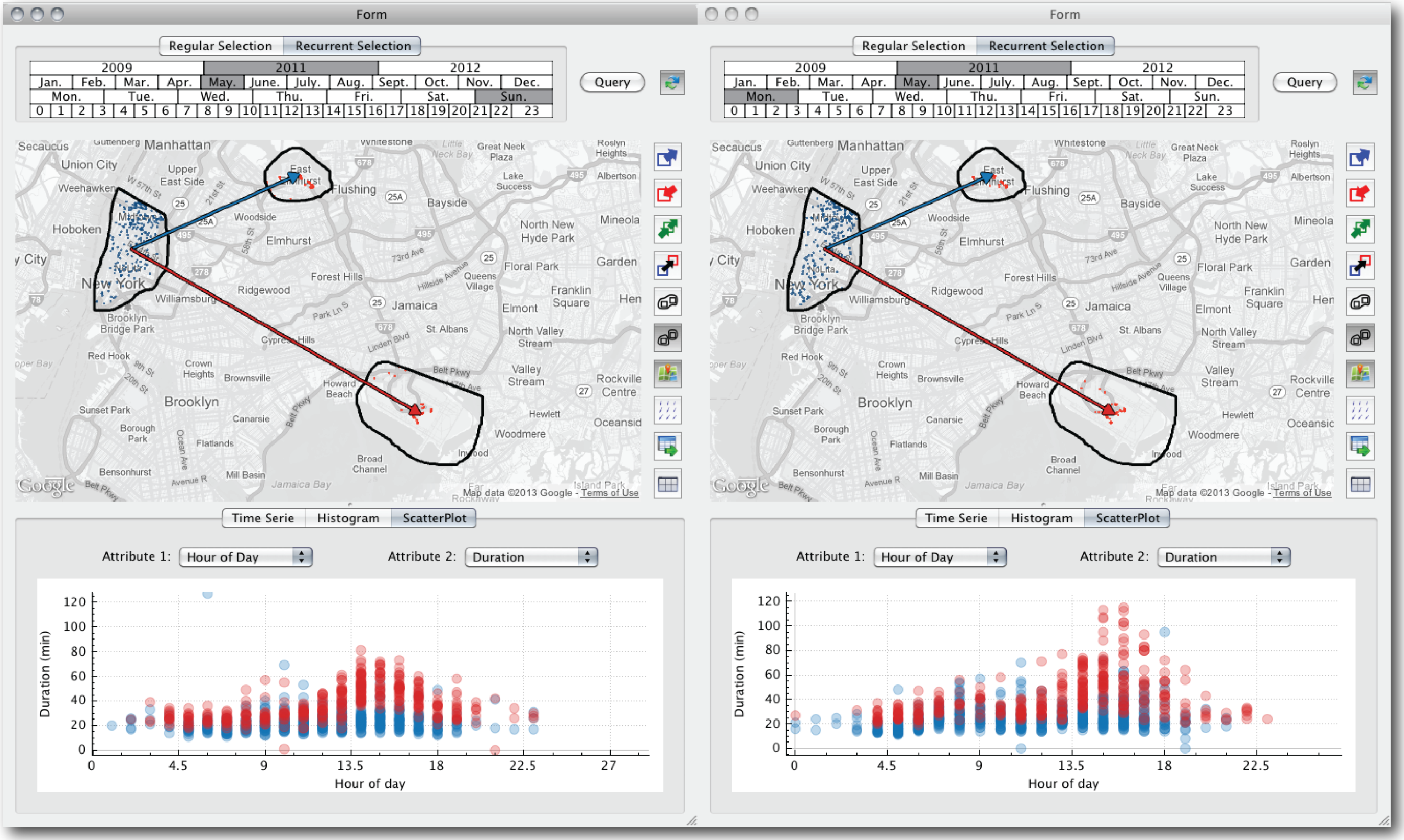
End Time: Sat 05/07/11 23:59

Time Serie | Histogram | ScatterPlot

Attribute: Num Trips

Num Bins: 200

TaxiVis: Studying Mobility



Attribute 1: Hour of Day

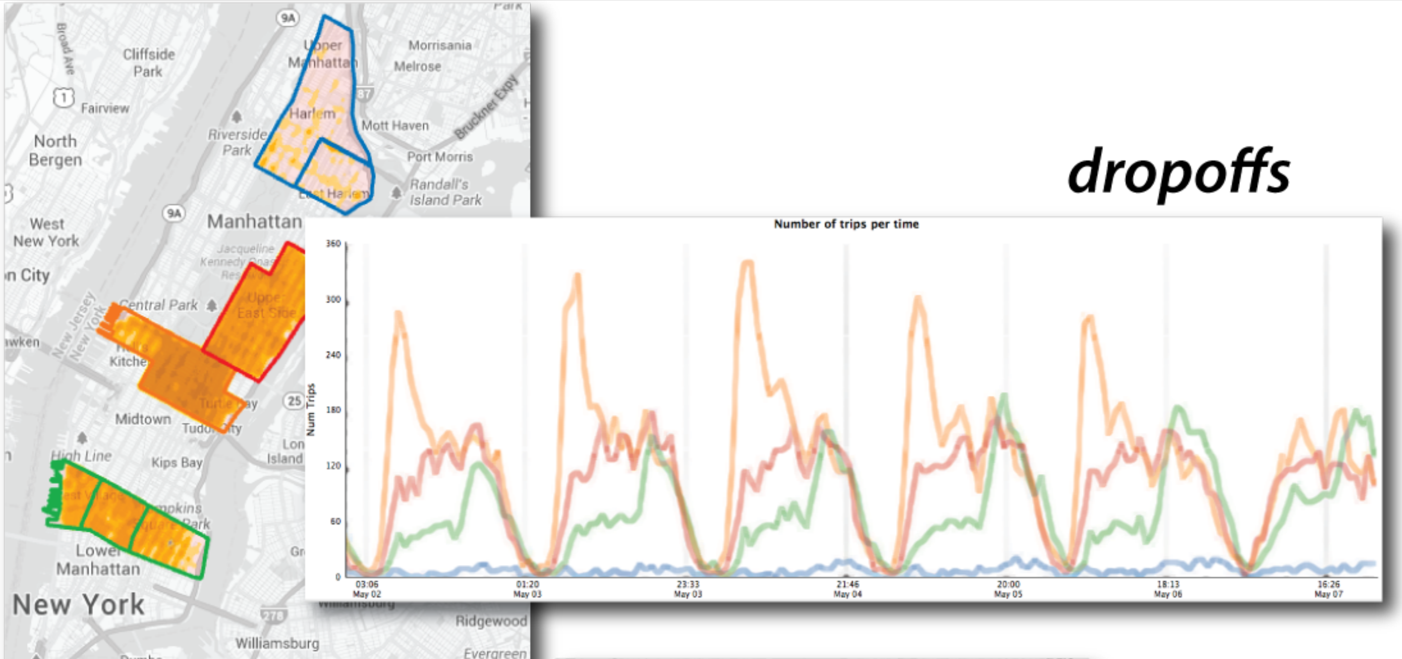
Attribute 2: Duration

NYU POLYTECHNIC SCHOOL OF ENGINEERING

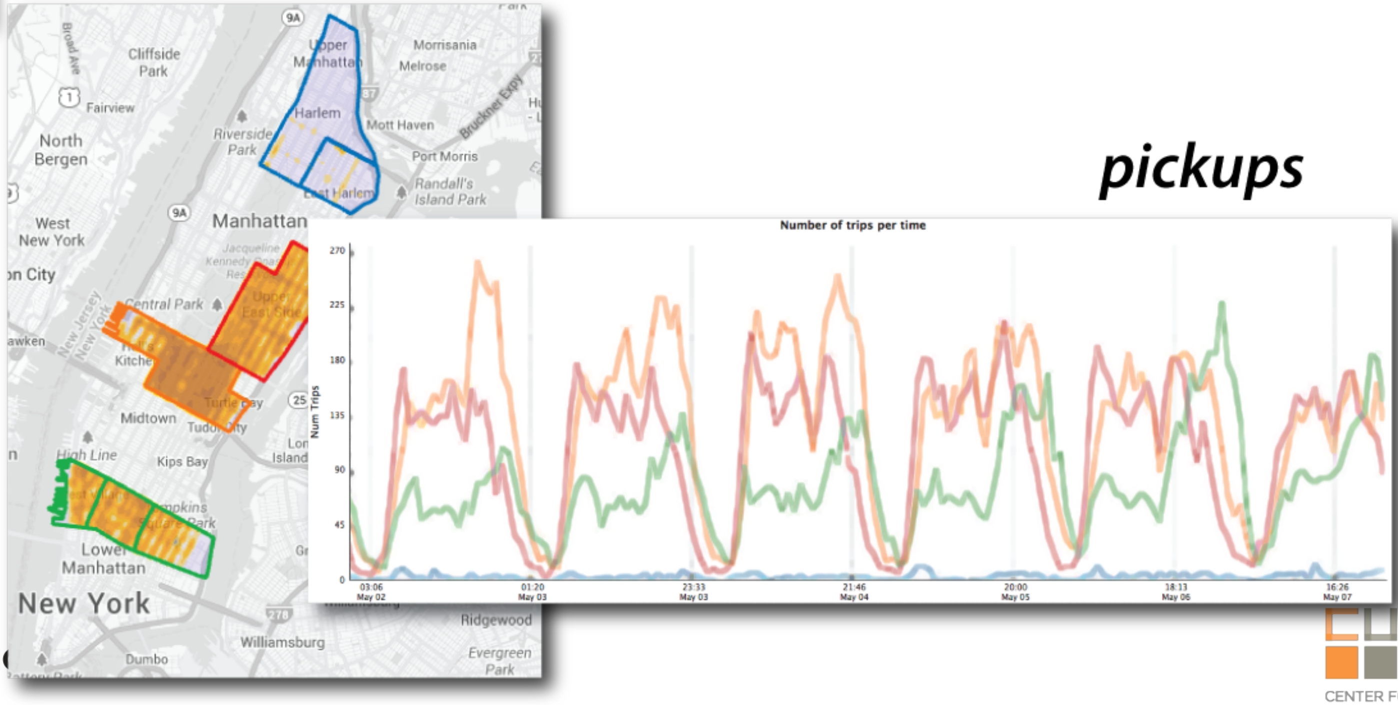
CUSP CENTER FOR URBAN SCIENCE+PROGRESS

TaxiVis: Comparing Neighborhoods

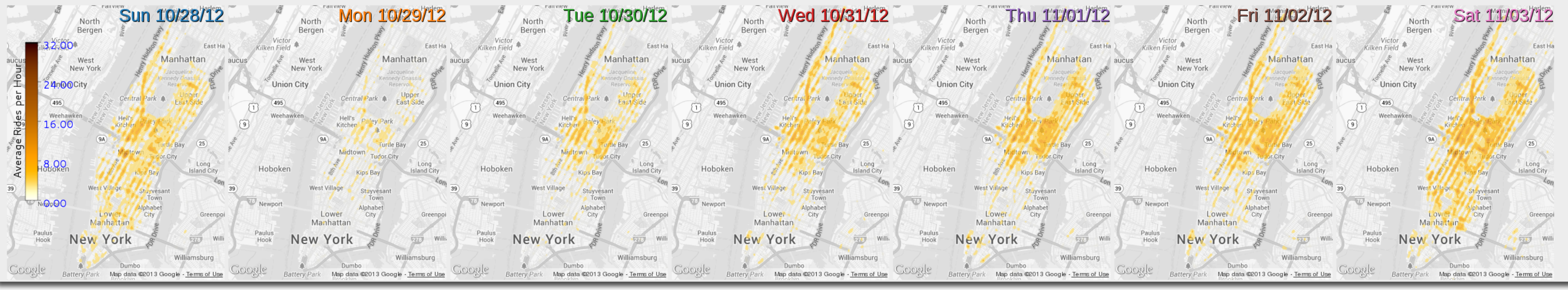
dropoffs

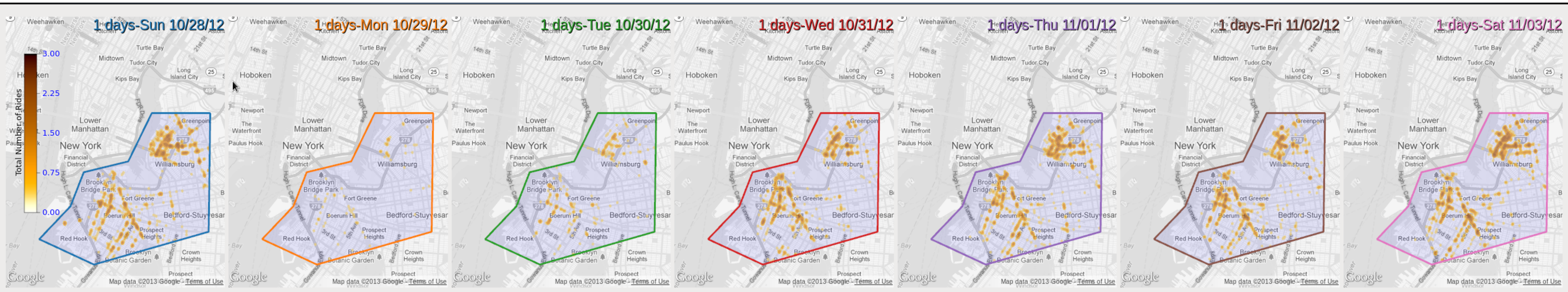


pickups

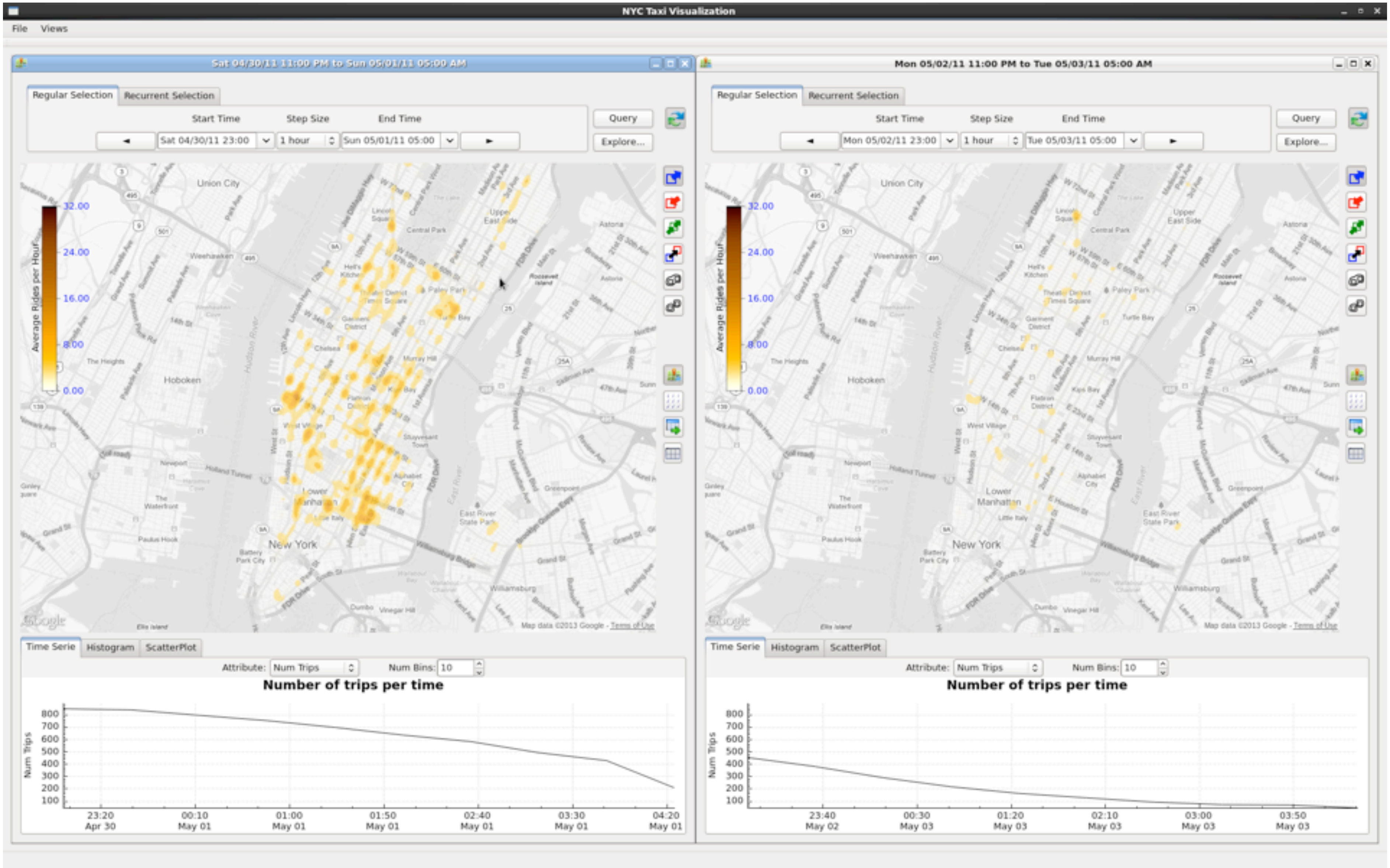


Exploring the Effect of Major Events: Sandy





Night Life in NYC: Saturday vs. Monday



NYC Taxi Visualization

File Views

Sat 04/30/11 11:00 PM to Sun 05/01/11 05:00 AM

Regular Selection | Recurrent Selection

Start Time: Sat 04/30/11 23:00 — Step Size: 1 hour — End Time: Sun 05/01/11 05:00

Query — Explore...

Time Serie | Histogram | ScatterPlot

Attribute: Num Trips — Num Bins: 10

Number of trips per time

Mon 05/02/11 11:00 PM to Tue 05/03/11 05:00 AM

Regular Selection | Recurrent Selection

Start Time: Mon 05/02/11 23:00 — Step Size: 1 hour — End Time: Tue 05/03/11 05:00

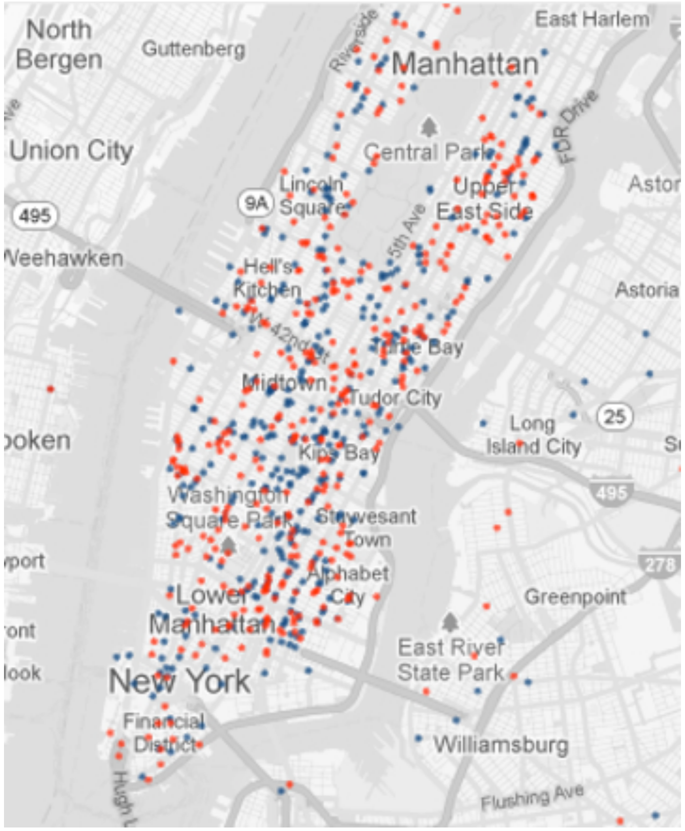
Query — Explore...

Time Serie | Histogram | ScatterPlot

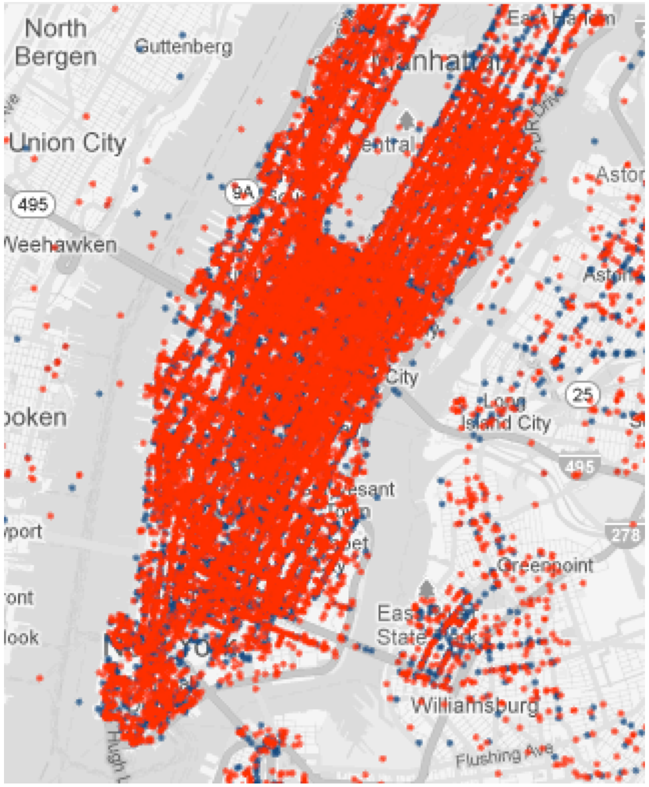
Attribute: Num Trips — Num Bins: 10

Number of trips per time

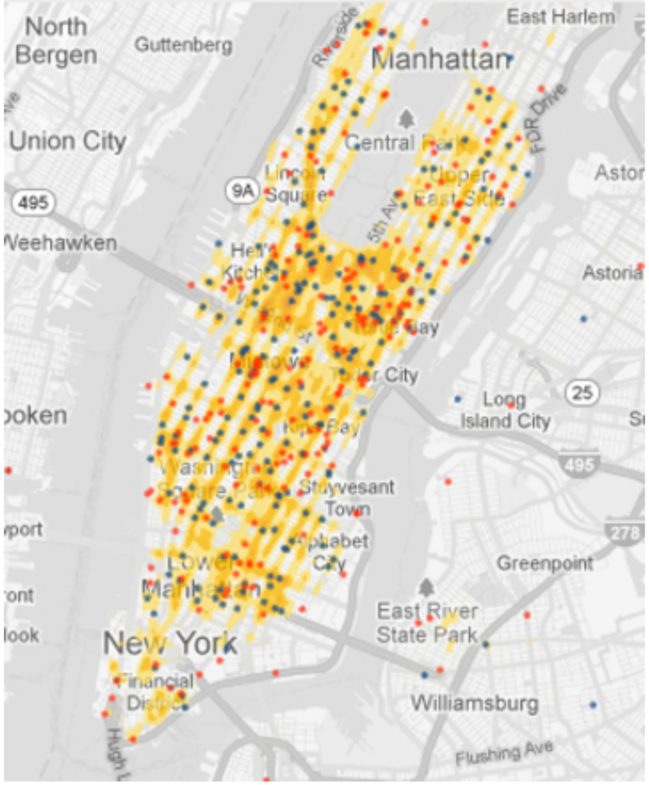
Visualizing Lots of Data



trips in an hour



trips in a day
too much information!



trips in a day
using level of detail and heat maps

NYU POLYTECHNIC SCHOOL OF ENGINEERING

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

TaxiVis: The Plumbing

- Requirement: support 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

Solution 1: In-memory spatio-temporal index based on kd-trees

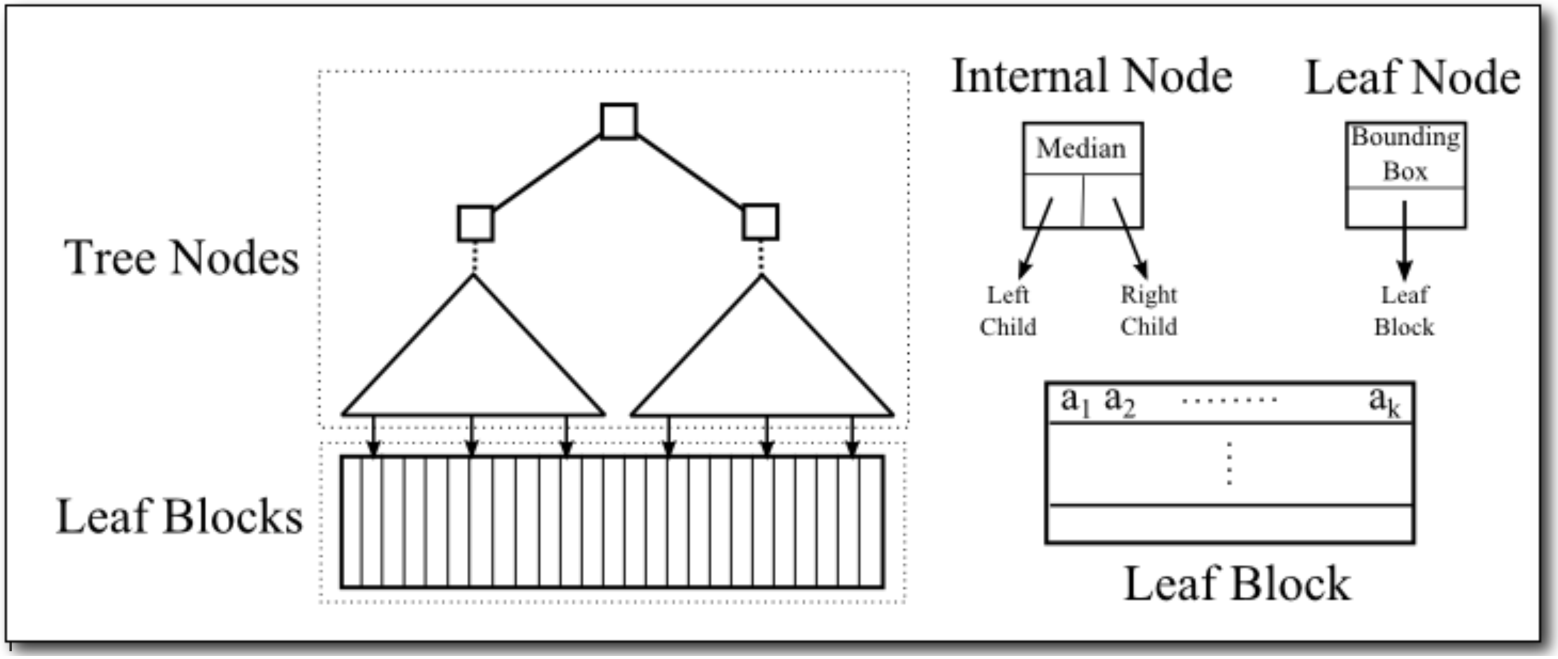
- Can index multiple attributes!
- Tree nodes store kd-tree
- A leaf node represents a k-dimensional node that satisfies the path constraints

	SQLite	Postgre SQL	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

Solution 2: Spatio-temporal index based on out-of-core kd-tree using GPUs *(work in progress)*

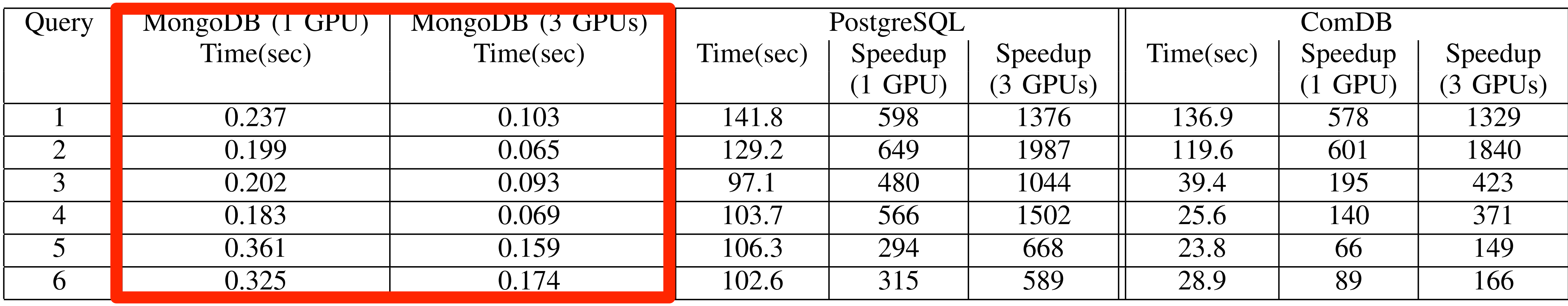
- Can index multiple attributes!
- Tree nodes store kd-tree
- Leaf nodes represent a *set of k-dimensional nodes*
  - Point to a leaf block containing records that satisfy the path constraints
  - Store the bounding box for the records



Supporting Interactive Queries

Solution 2: Spatio-temporal index based on out-of-core kd-tree using GPUs *(work in progress)*

- Can index multiple attributes!
- Tree nodes store k-d tree
- Leaf nodes represent a *set of k-dimensional nodes*
  - Point to a leaf block containing records that satisfy the path constraints
  - Store the bounding box for the records

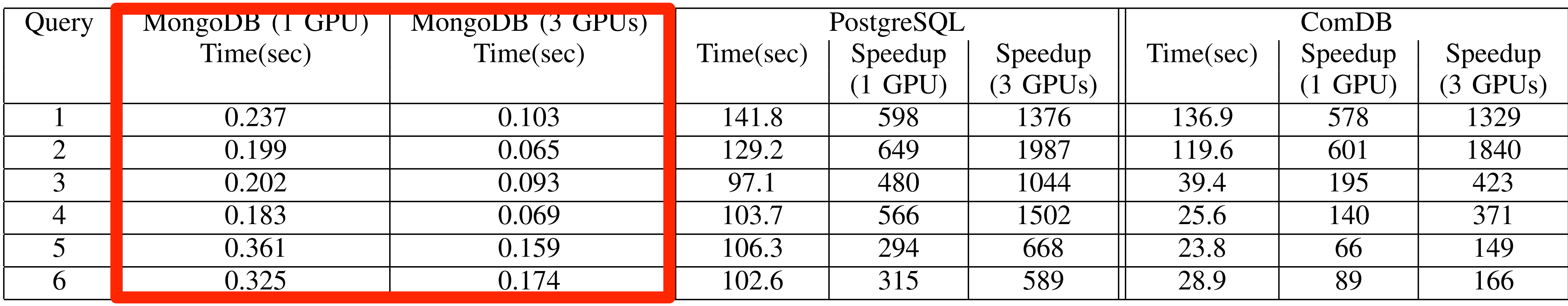


Query	MongoDB (1 GPU) Time(sec)	MongoDB (3 GPUs) Time(sec)	PostgreSQL			ComDB		
			Time(sec)	Speedup (1 GPU)	Speedup (3 GPUs)	Time(sec)	Speedup (1 GPU)	Speedup (3 GPUs)
1	0.237	0.103	141.8	598	1376	136.9	578	1329
2	0.199	0.065	129.2	649	1987	119.6	601	1840
3	0.202	0.093	97.1	480	1044	39.4	195	423
4	0.183	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

2 orders of magnitude faster than RBDMSs

Supporting Interactive Queries

Solution 2: Spatio-temporal index based on out-of-core kd-tree using GPUs *(work in progress)*

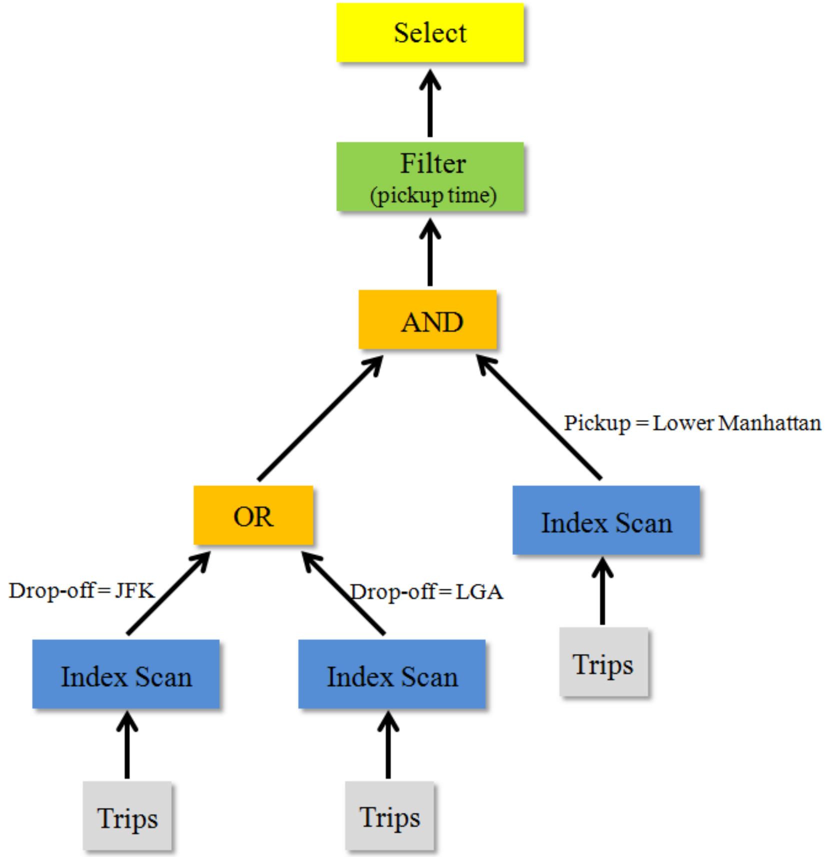


Query	MongoDB (1 GPU)	MongoDB (3 GPUs)	PostgreSQL			ComDB		
	Time(sec)	Time(sec)	Time(sec)	Speedup (1 GPU)	Speedup (3 GPUs)	Time(sec)	Speedup (1 GPU)	Speedup (3 GPUs)
1	0.237	0.103	141.8	598	1376	136.9	578	1329
2	0.199	0.065	129.2	649	1987	119.6	601	1840
3	0.202	0.093	97.1	480	1044	39.4	195	423
4	0.183	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

2 orders of magnitude faster than RBDMSs

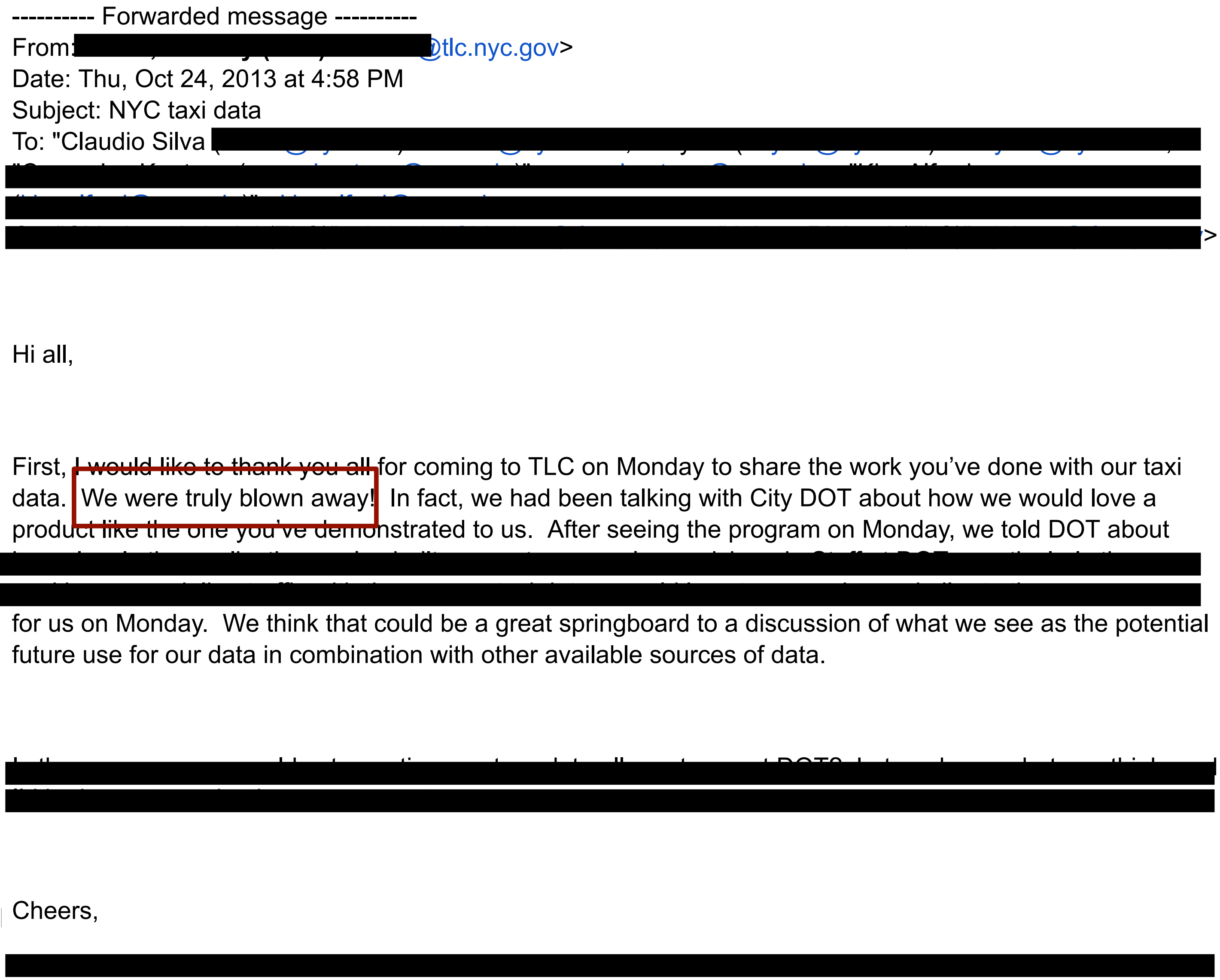
Why are RDBMS so slow?

- Designed for batch queries, not very good for interactive queries [Fekete & Silva, IEEE DEB 2012]
- R-trees and R*-trees are limited to a single spatial attribute
  - Taxi data has origin + destination
  - Needs a join!
- Lots of point-in-polygon tests
  - Filter location before other query constraints
  - Too many intermediate results



TaxiVis: Status

- Demoed to NYC DOT and TLC
  - They are currently using the prototype!



---------- Forwarded message ----------
From: [redacted]@tlc.nyc.gov>
Date: Thu, Oct 24, 2013 at 4:58 PM
Subject: NYC taxi data
To: "Claudio Silva [redacted]>

Hi all,

First, I would like to thank you all for coming to TLC on Monday to share the work you've done with our taxi data. We were truly blown away! In fact, we had been talking with City DOT about how we would love a product like the one you've demonstrated to us. After seeing the program on Monday, we told DOT about [redacted] for us on Monday. We think that could be a great springboard to a discussion of what we see as the potential future use for our data in combination with other available sources of data.

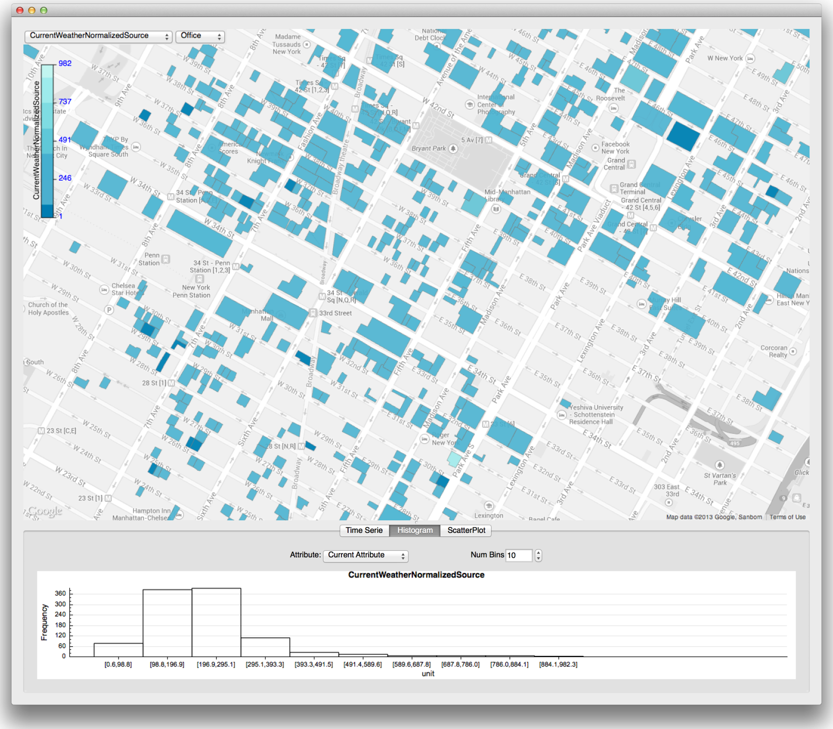
[redacted]

Cheers,

[redacted]

TaxiVis: Status

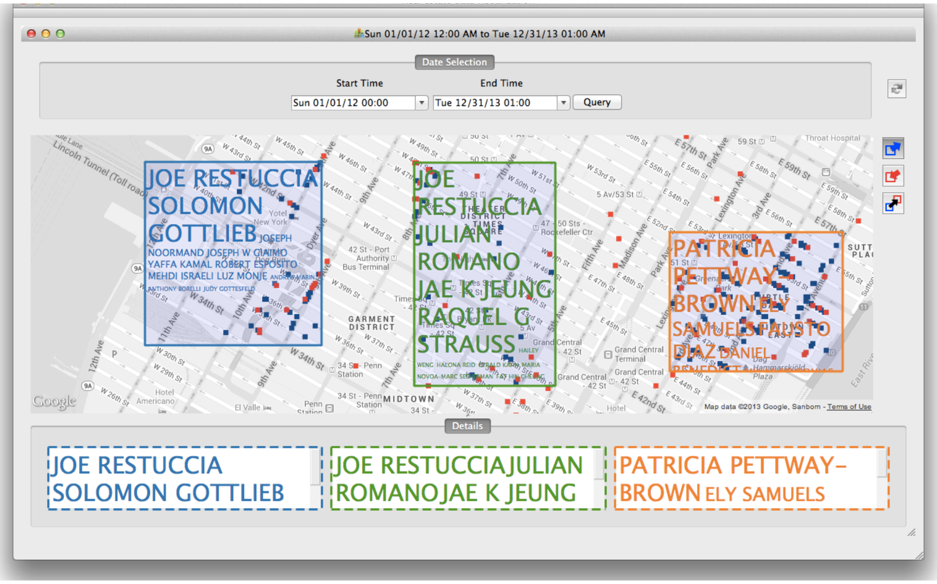
- Demoed to NYC DOT and TLC
  - They are currently using the prototype!
- Applying to different data sets
  - Bikes, energy consumption, property ownership, etc.
- Improving scalability by leveraging multiple cores
- Vision: the GIS of the future
  - Scalable
  - Powerful analytics



CurrentWeatherNormalizedSource
Office

Time Serie | Histogram | ScatterPlot

Attribute: Current Attribute
Num Bins 10



Sun 01/01/12 12:00 AM to Tue 12/31/13 01:00 AM

Date Selection

Start Time: Sun 01/01/12 00:00
End Time: Tue 12/31/13 01:00
Query

Details

JOE RESTUCCIA SOLOMON GOTTLIEB

JOE RESTUCCIAJULIAN ROMANOJAE K JEUNG

PATRICIA PETTWAY-BROWN ELY SAMUELS

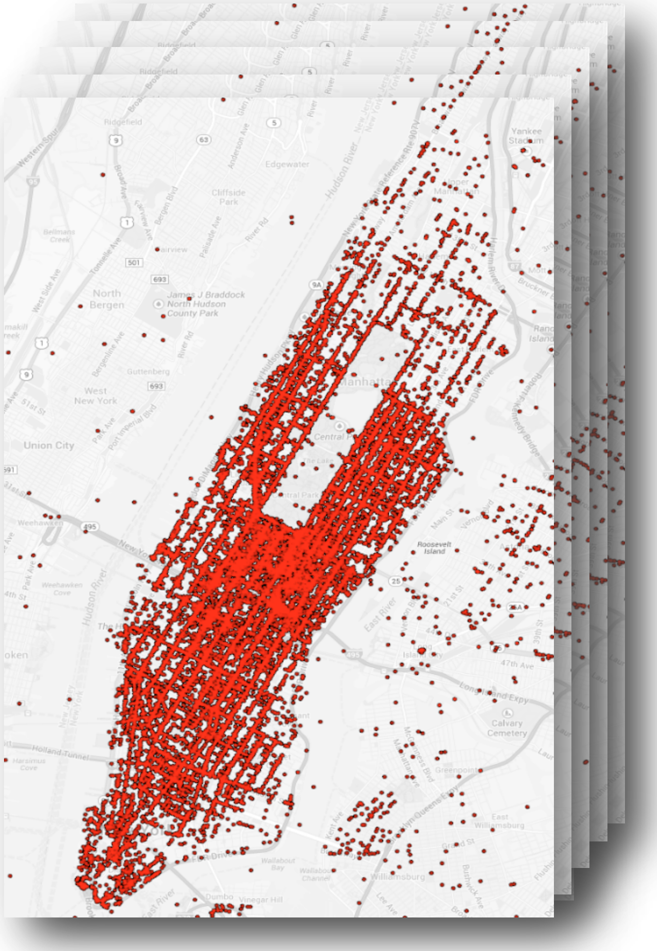
NYU POLYTECHNIC SCHOOL OF ENGINEERING

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

Taxi Data: Too Many Slices

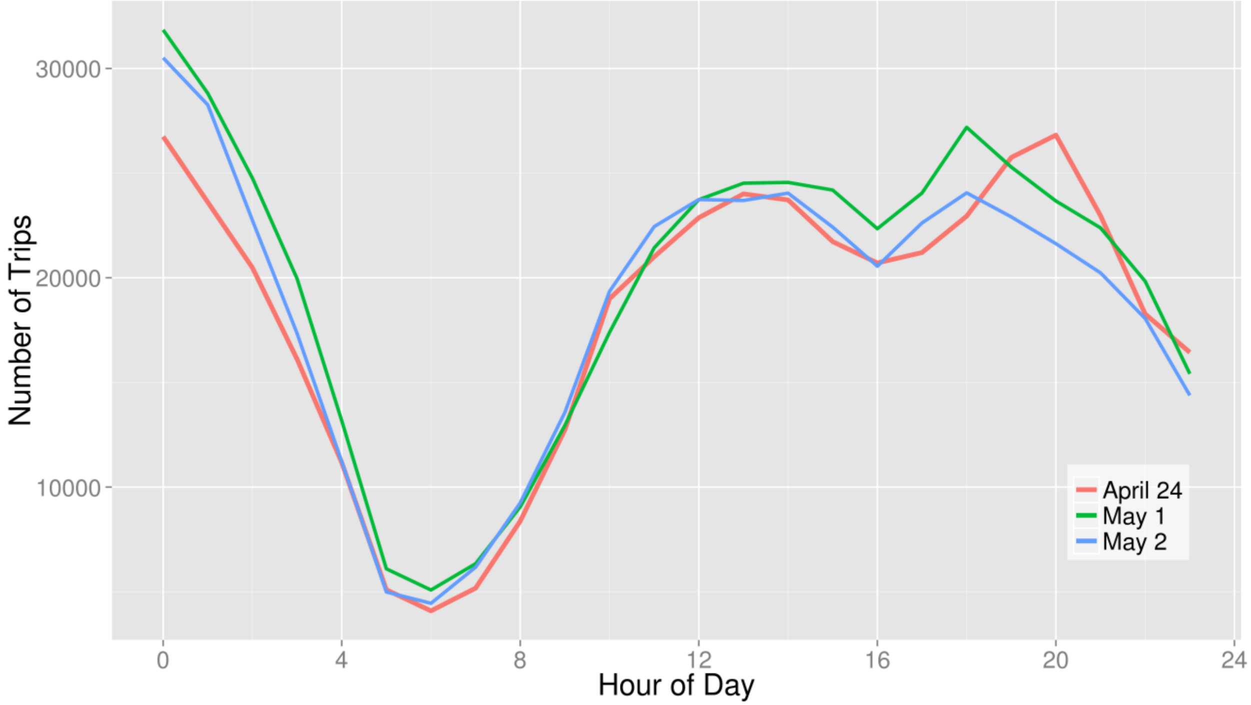
- 170 Million trips / year
- Spatial context
  - pick-up and drop-off locations
- Temporal attribute
  - pick-up and drop-off times

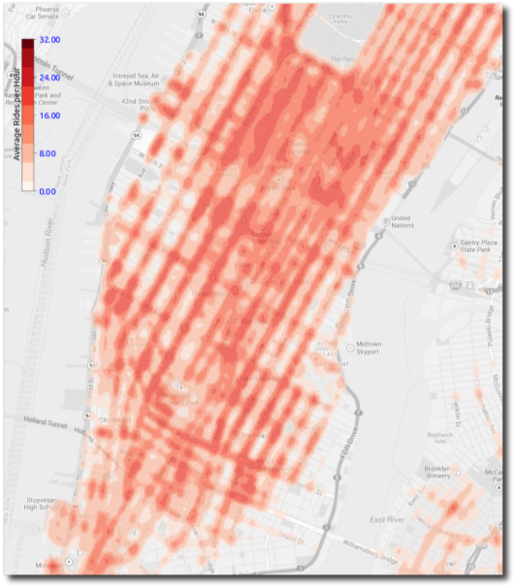
- Which slices are interesting?
- Can we guide users to interesting features in the data?



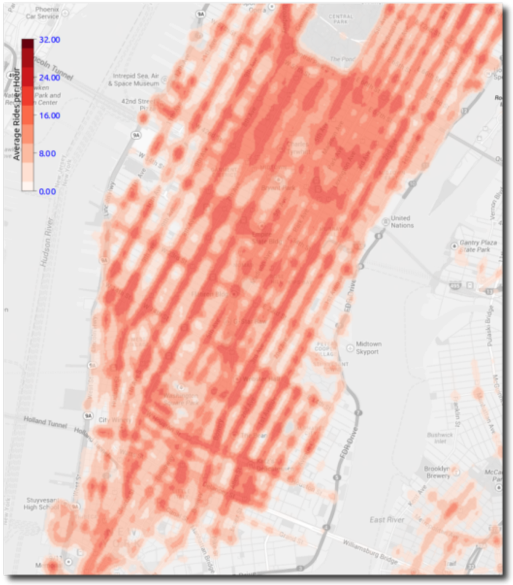
Reducing the Number of Slices

- *Aggregate* over space
- *Aggregate* over time

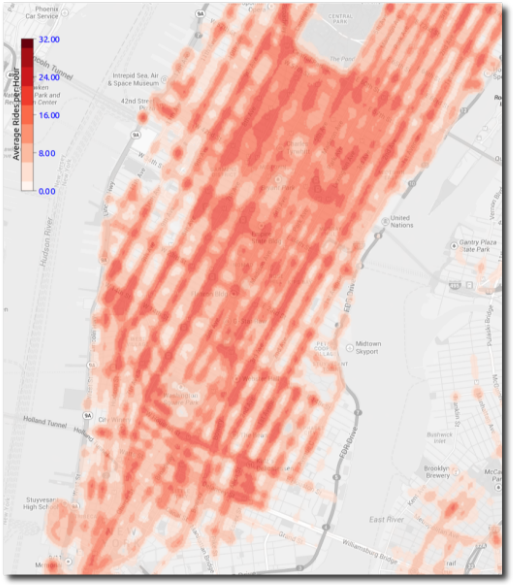




April 24

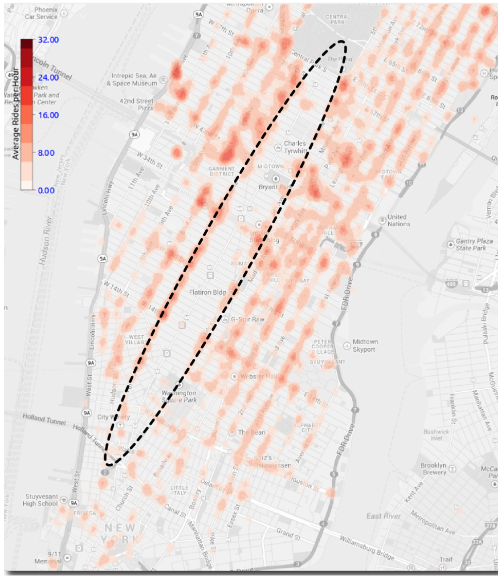


May 1

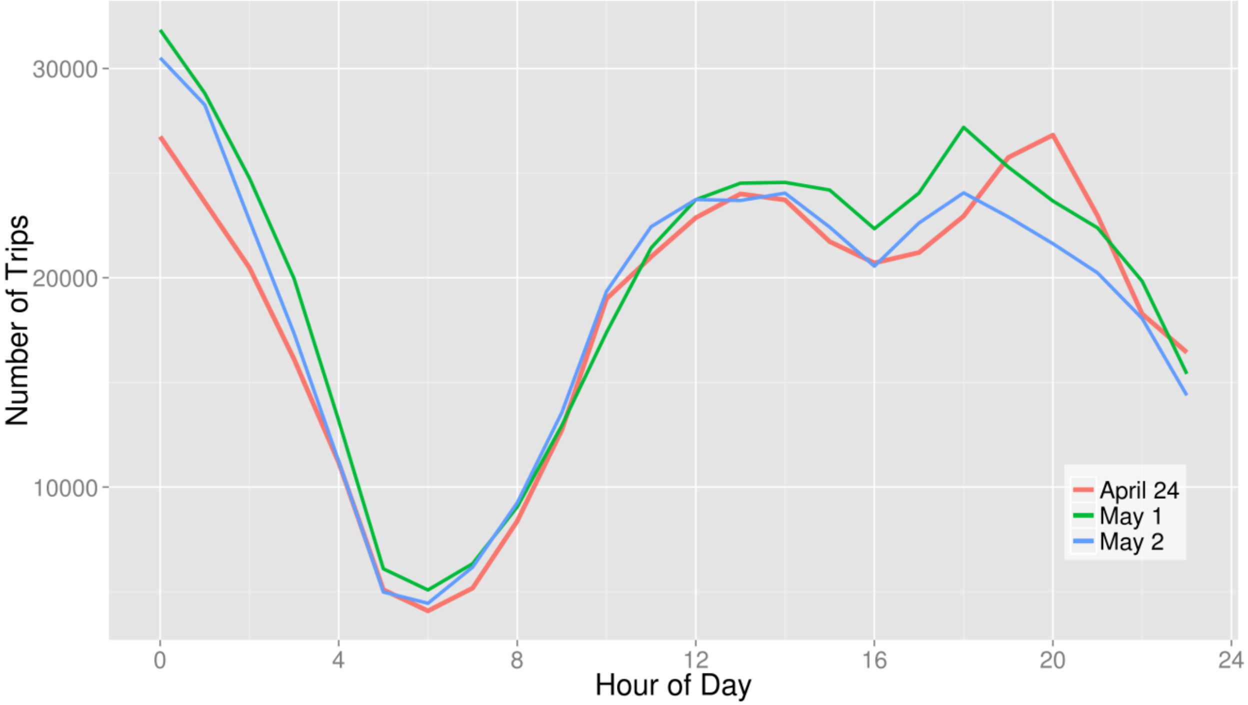


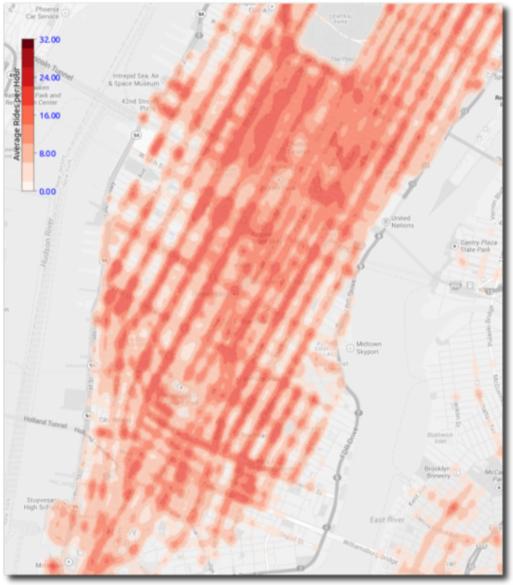
May 8

Miss Interesting Slices

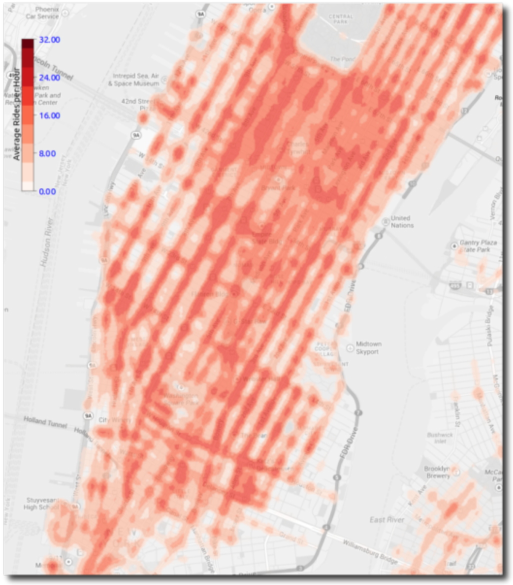


May 1 (8-9am)

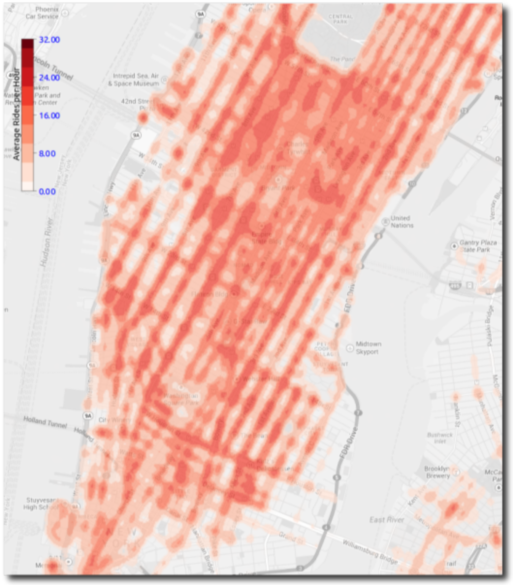




April 24



May 1



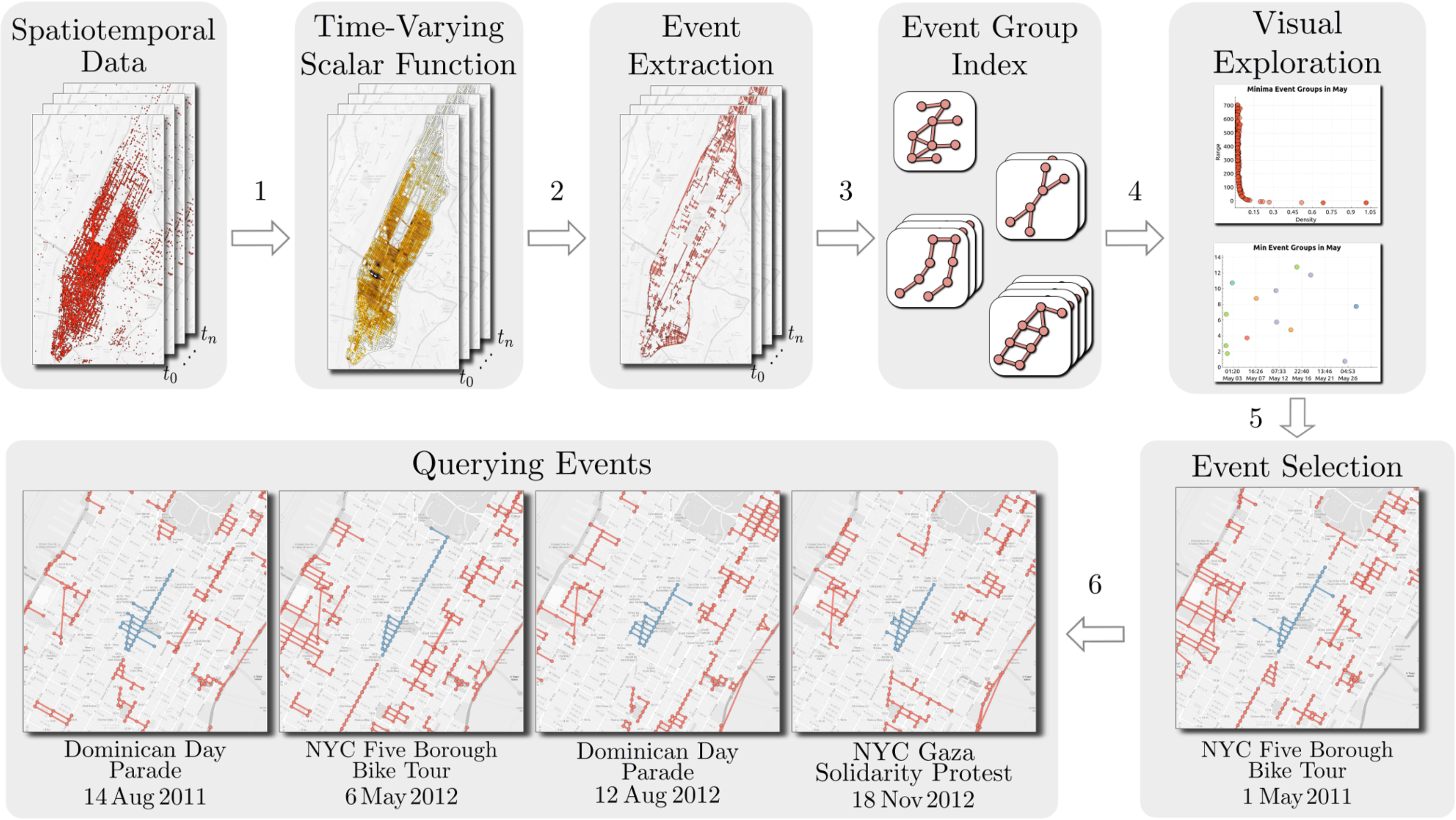
May 8

Finding Events at Multiple Granularities

- Goal: guide users towards interesting data slices
- Use topology-based techniques to efficiently identify potential events
- Use a simple visual interface to *explore* and *query* the events of interest
  - Efficient search for similar event patterns
- Flexible definition of events
  - Arbitrary spatial structure
  - Different types of events
  - Multiple temporal scales

[Doraiswamy et al., IEEE TVCG 2014]

Our Approach: Overview



Spatiotemporal Data

Time-Varying Scalar Function

Event Extraction

Event Group Index

Visual Exploration

Event Selection

Querying Events

Dominican Day Parade 14 Aug 2011

NYC Five Borough Bike Tour 6 May 2012

Dominican Day Parade 12 Aug 2012

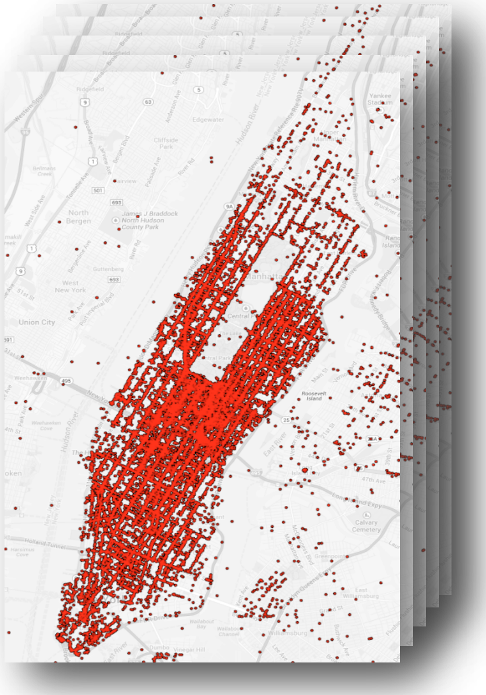
NYC Gaza Solidarity Protest 18 Nov 2012

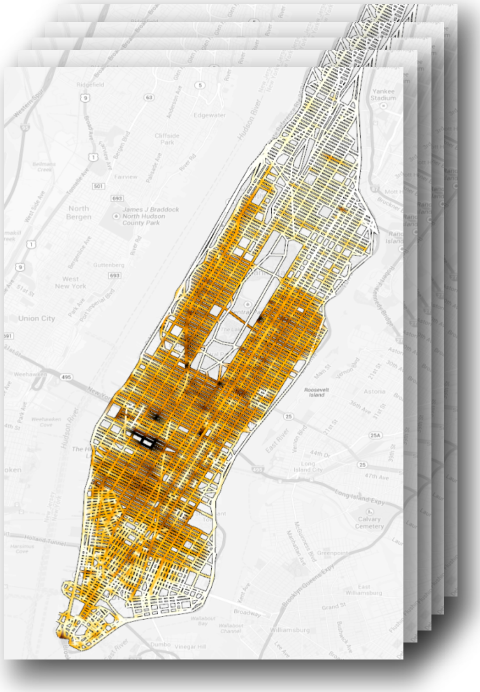
NYC Five Borough Bike Tour 1 May 2011

[Doraiswamy et al., IEEE TVCG 2014]

Identifying Potential Events

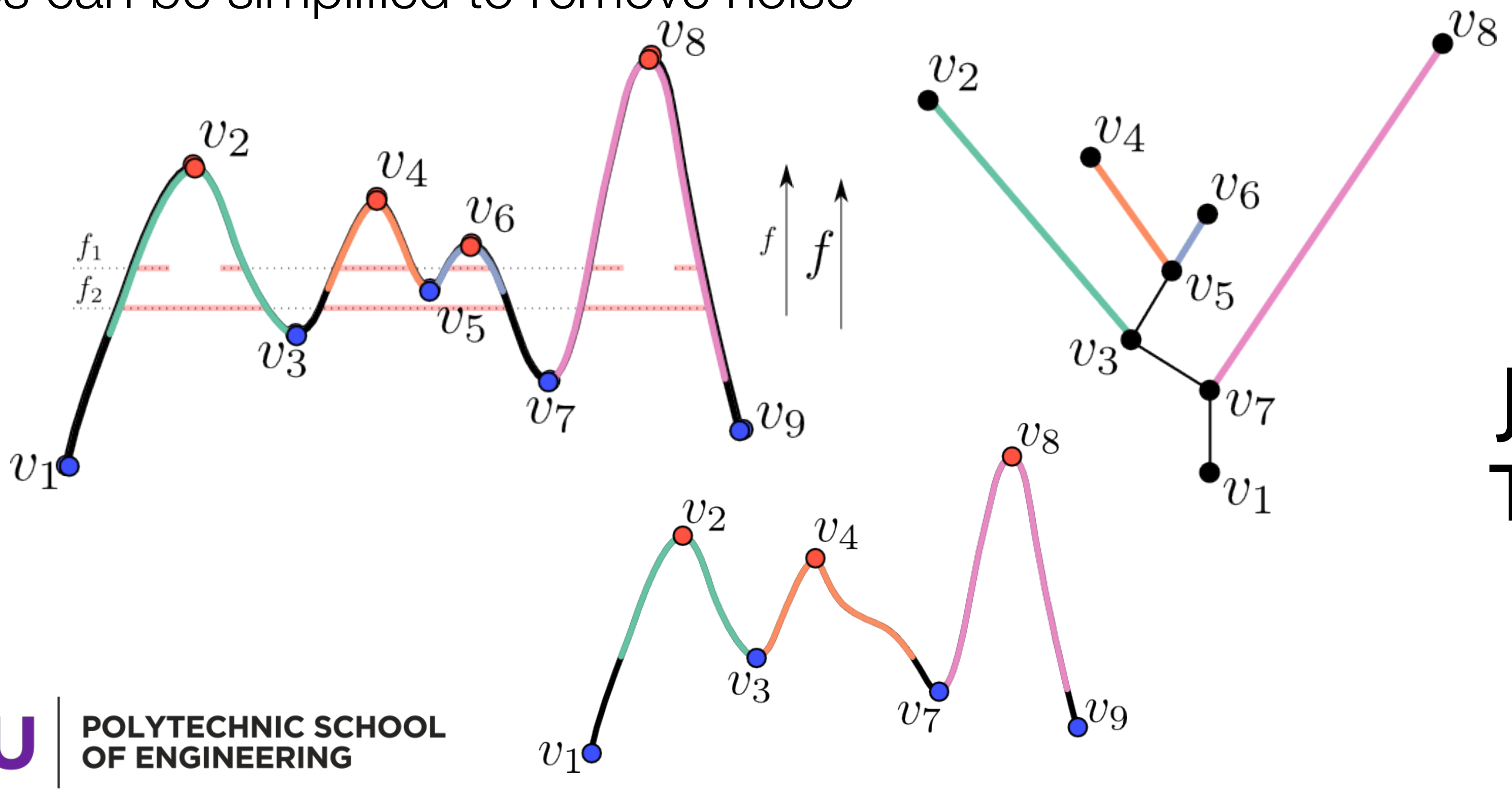
- Model data as a time-varying scalar function defined on a graph
  - Graph = road network
  - Function = density of taxis





Identifying Potential Events

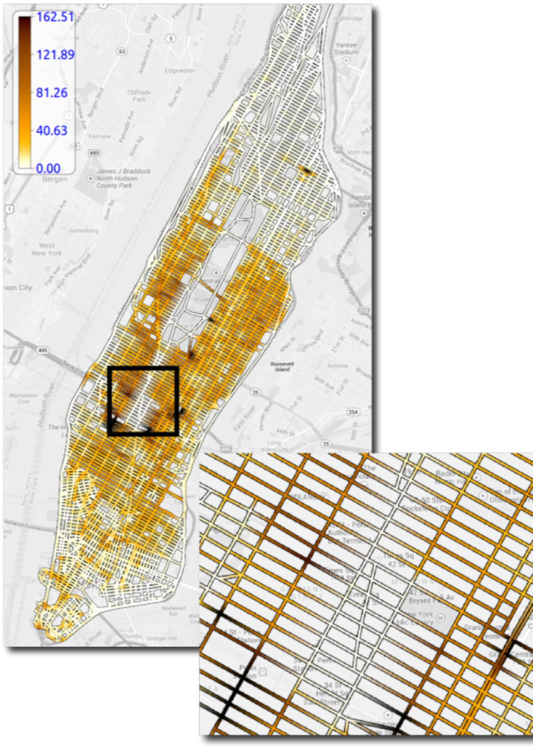
- Compute the regions corresponding to the set of *maxima* and *minima – the set of potential events*
  - Intuition: a region is interesting if its density is different from that of its neighborhood
- Join and Split tree can be used to efficiently represent regions
  - Trees can be simplified to remove noise



Join Tree

Potential Events – Taxi Data

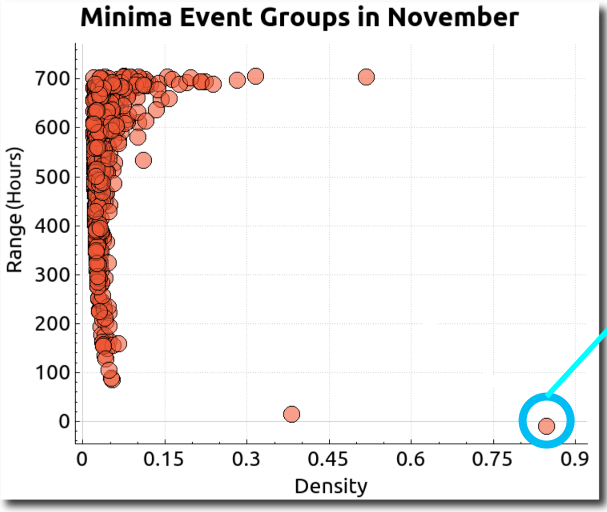
- Compute events based on the scalar function for each time step
  - Density of taxis in 1-hour intervals
- Minima: lack of taxis
  - Region where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade
- Maxima: popular taxi locations
  - Region where density is higher than local neighborhood
  - Could denote tourist locations, train stations
- Too many events: group similar events and create an index
  - Geometric and topological similarity



The scalar function corresponding to the time step 10 am-11 am on 24 November 2011

Visual Exploration Interface

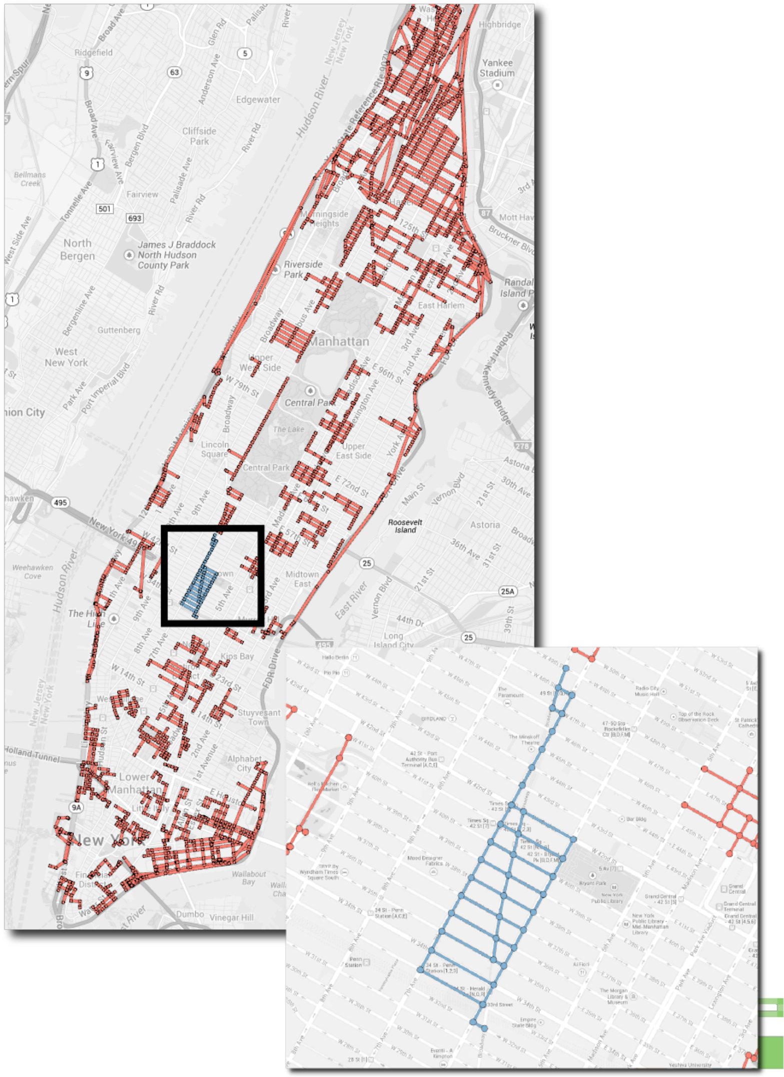
- Too many event groups
  - Many not interesting
- Visual interface to guide users
- Filter based on group size, event size, event time, spatial region



long → short time span

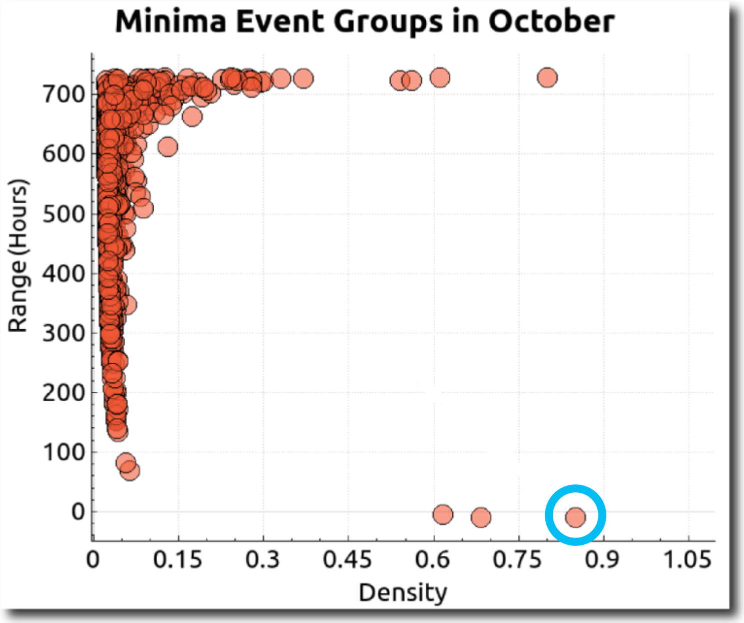
small → large groups

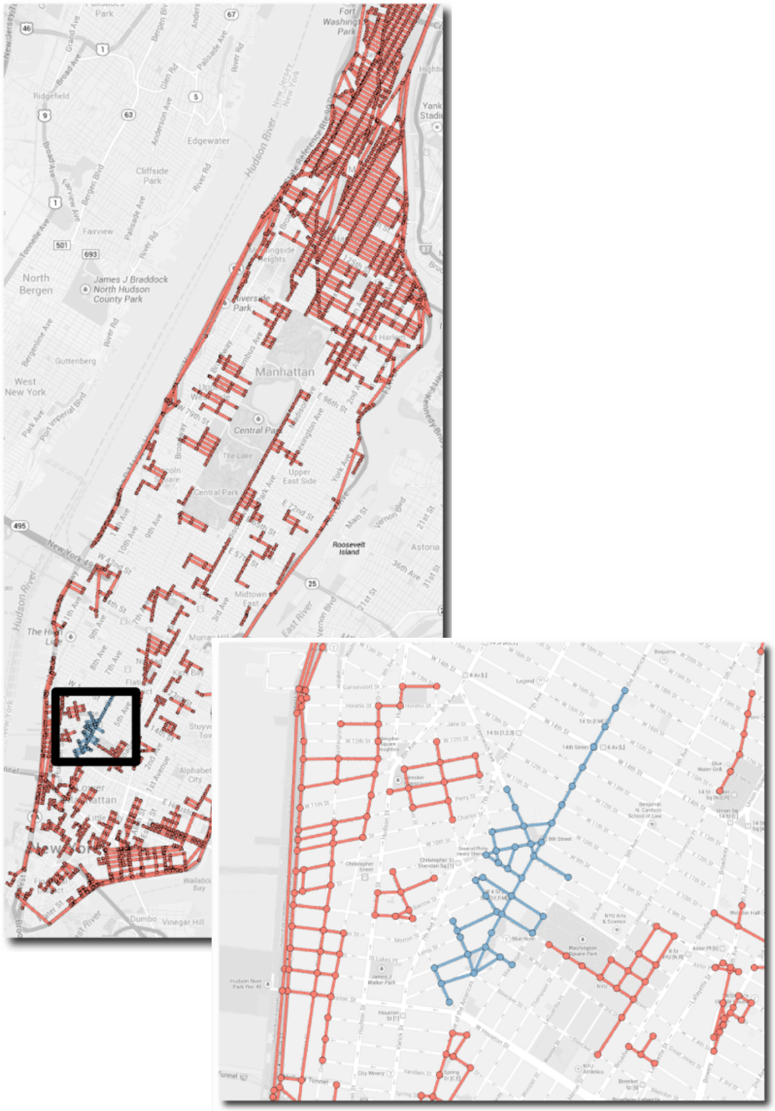
Macy's parade



Minima Events - Hourly

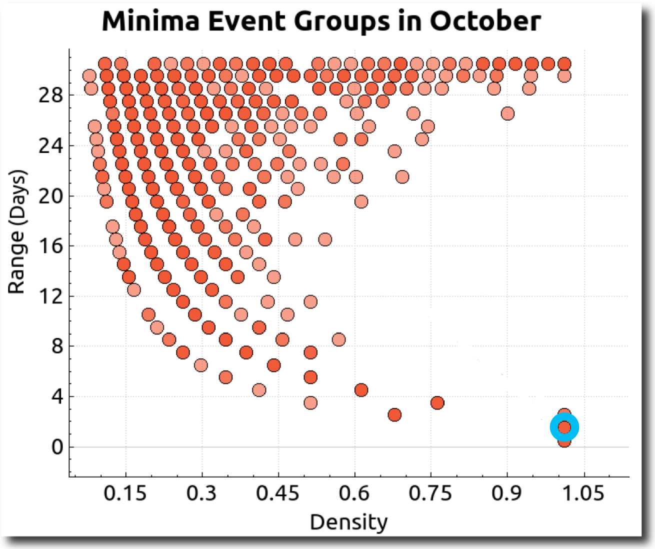
- October
  - Halloween Parade

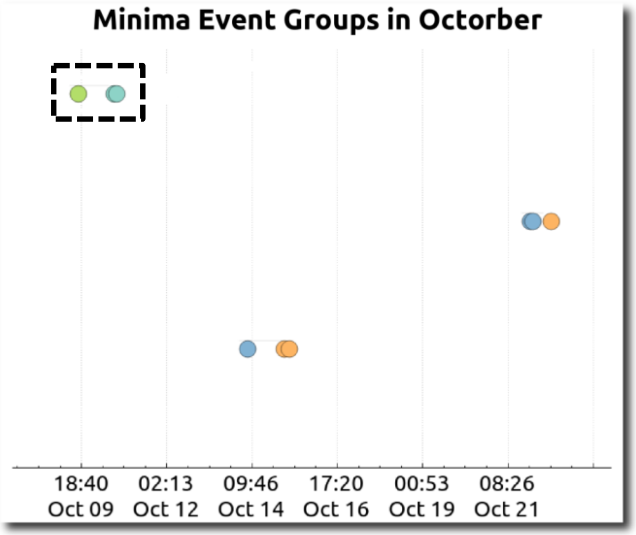


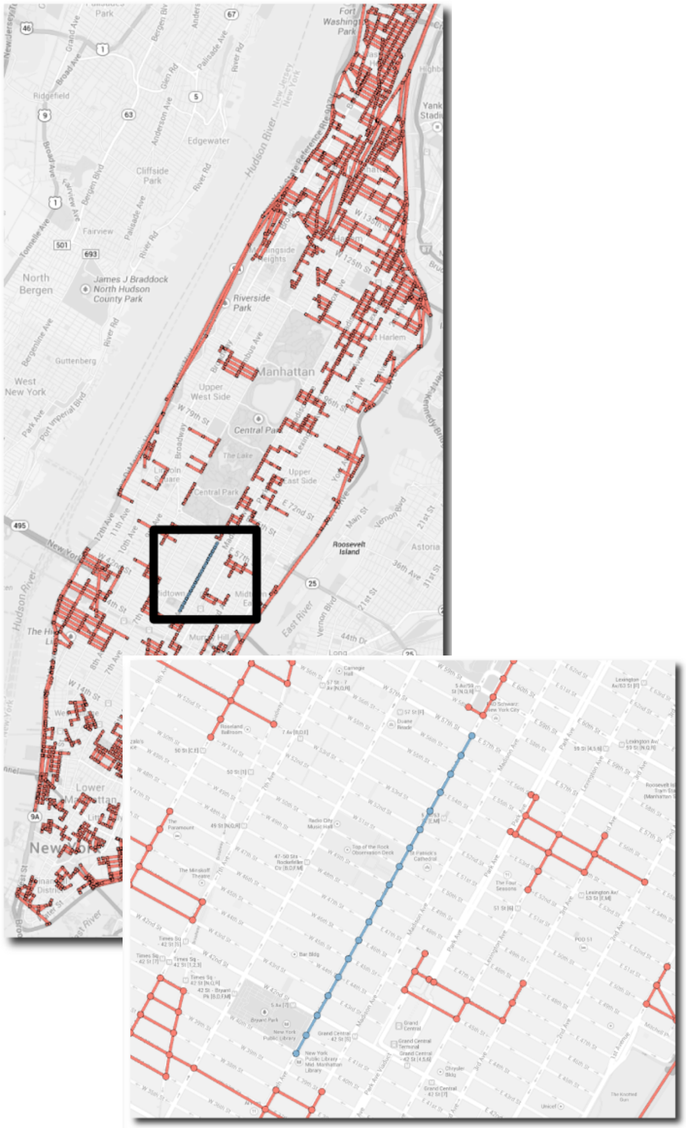


Minima Events - Daily

- October
  - No. of Days = 2
  - Hispanic Day Parade – Oct 9th
  - Columbus Day Parade – Oct 10th





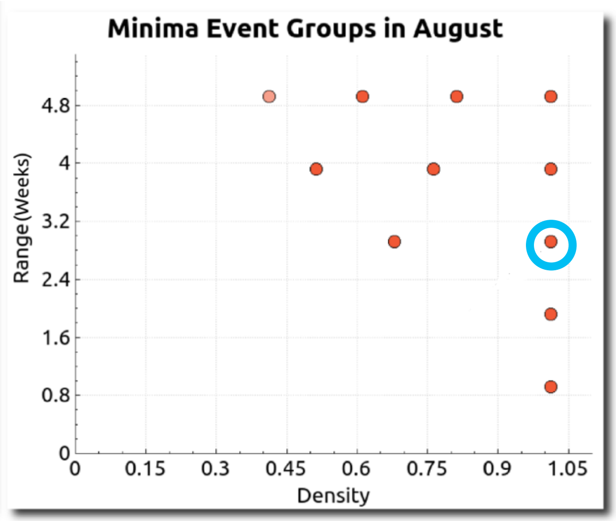


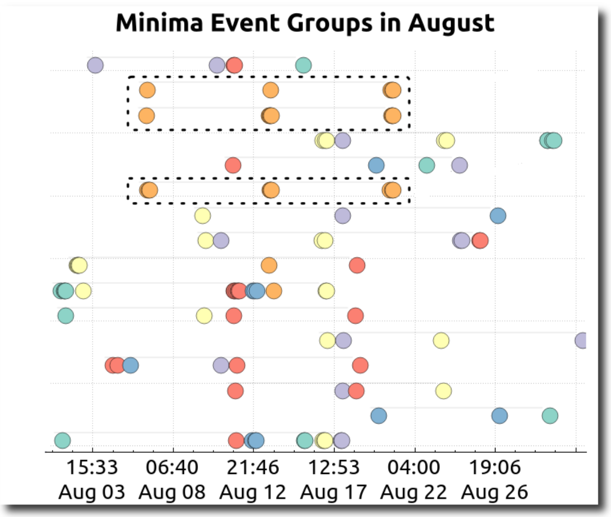
NYU POLYTECHNIC SCHOOL OF ENGINEERING

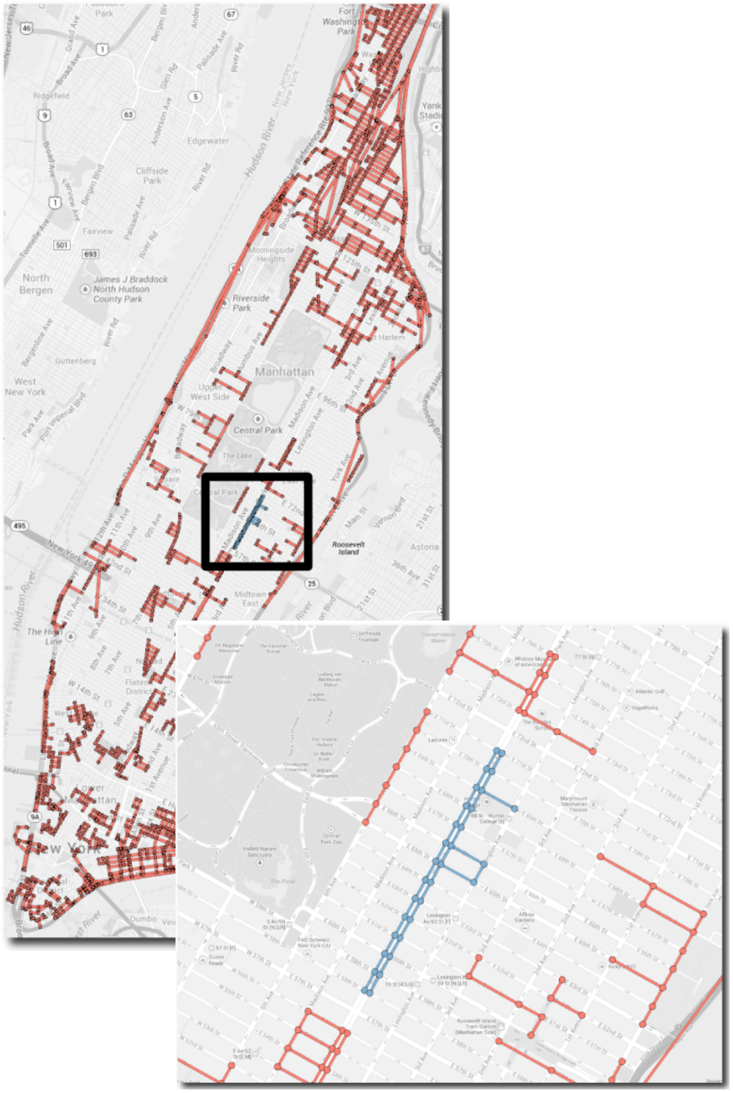
CENTER FOR URBAN SCIENCE+PROGRESS

Minima Events - Weekly

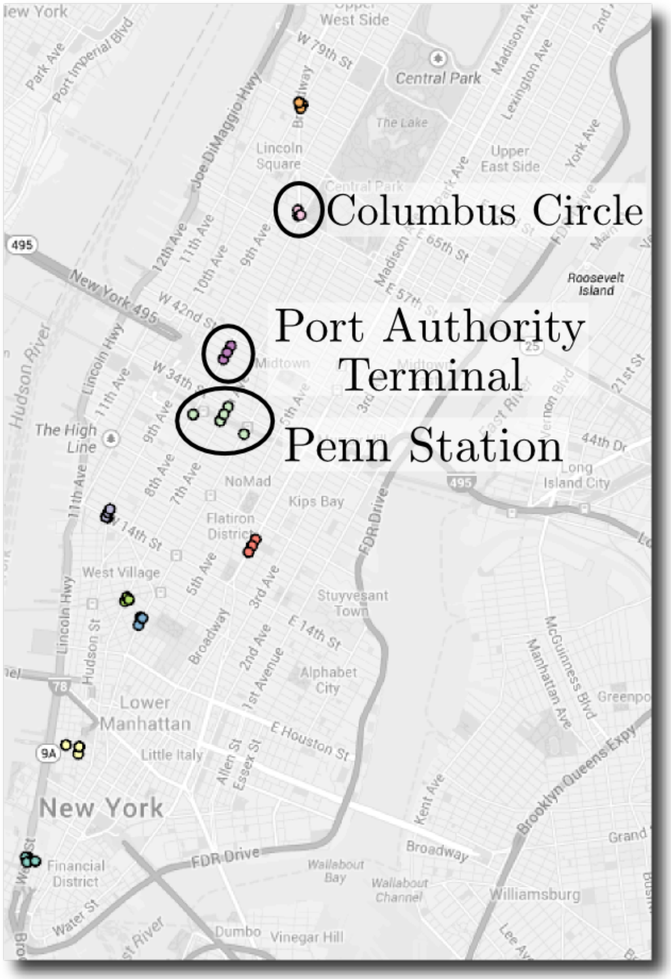
- August
  - No. of weeks = 3
  - NYC Summer Streets: 3 consecutive Saturdays


Minima Event Groups in August

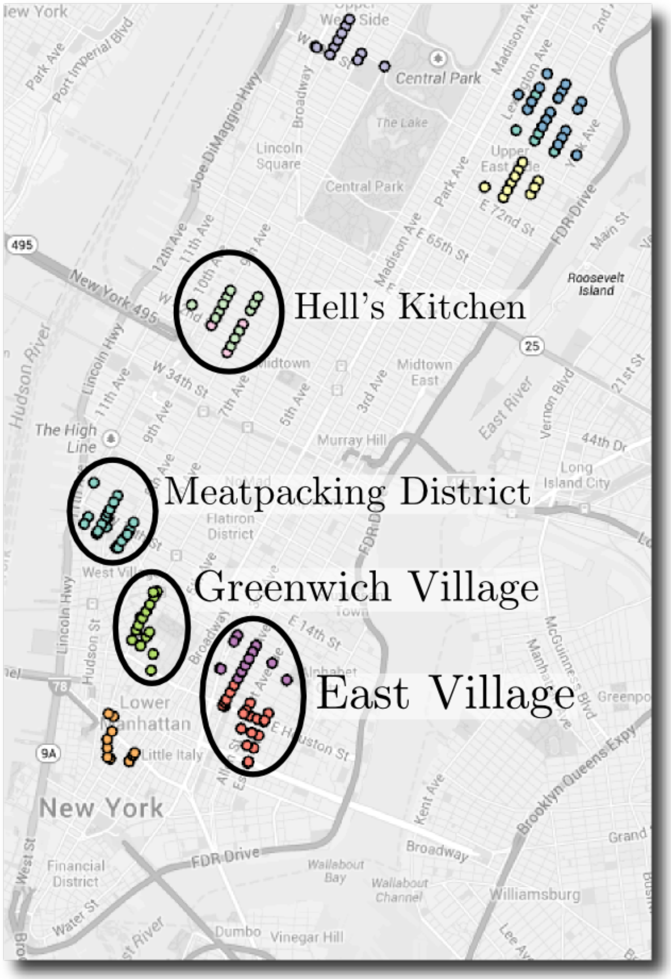

Minima Event Groups in August



Maxima Events

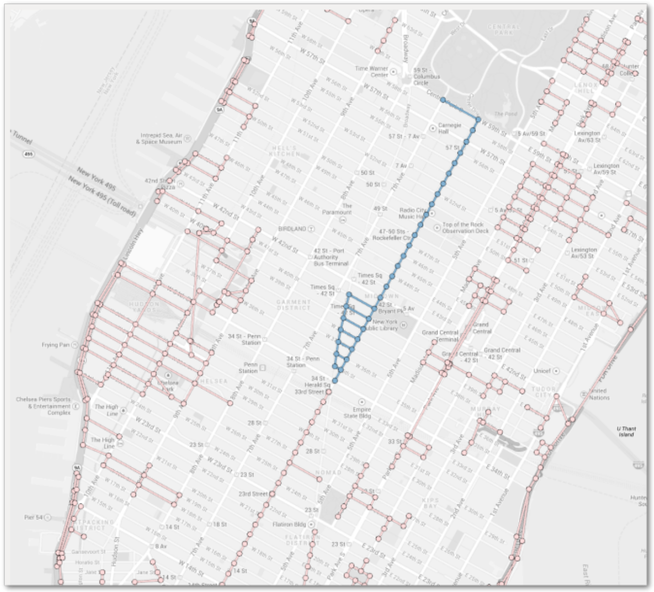


General trends



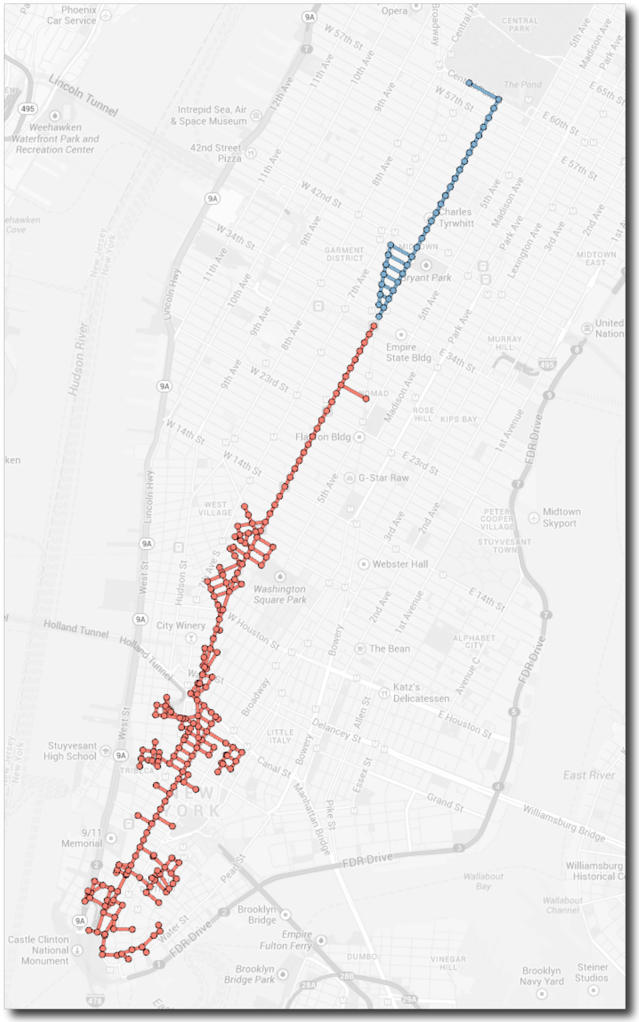
Night time trends

Event Guided Exploration

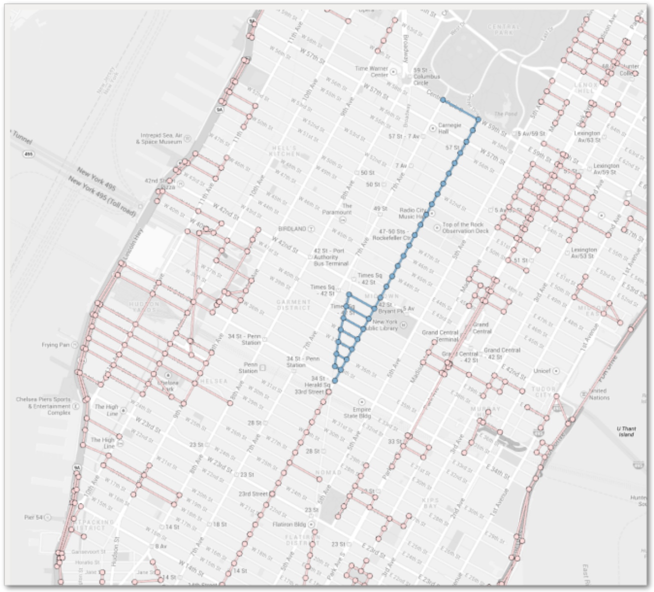


5 Borough Bike Tour 2011
(1 May 2011)

Go to Time slice

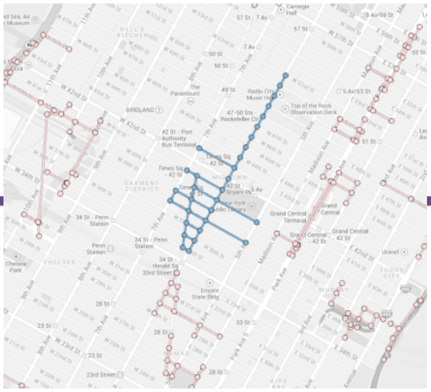


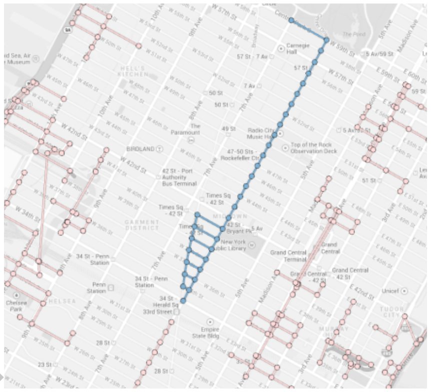
Querying Events

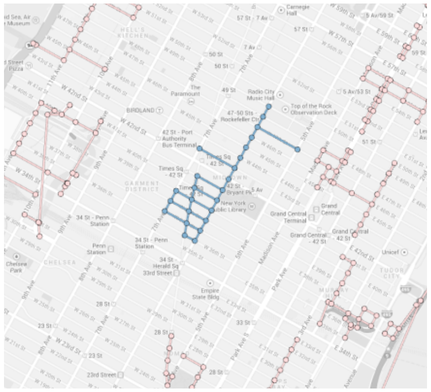


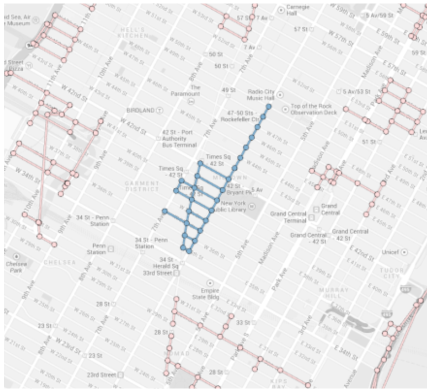
5 Borough Bike Tour 2011
(1 May 2011)

Query









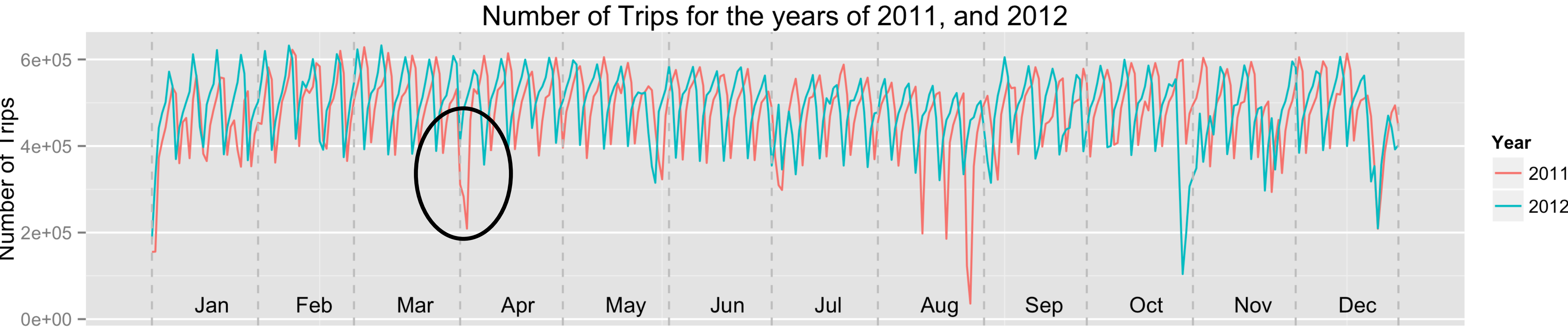
Dominican Day Parade 2011
(14 August 2011)

5 Borough Bike Tour 2012
(6 May 2012)

Dominican Day Parade 2012
(12 August 2012)

Gaza Solidarity Protest NYC
(18 November 2012)

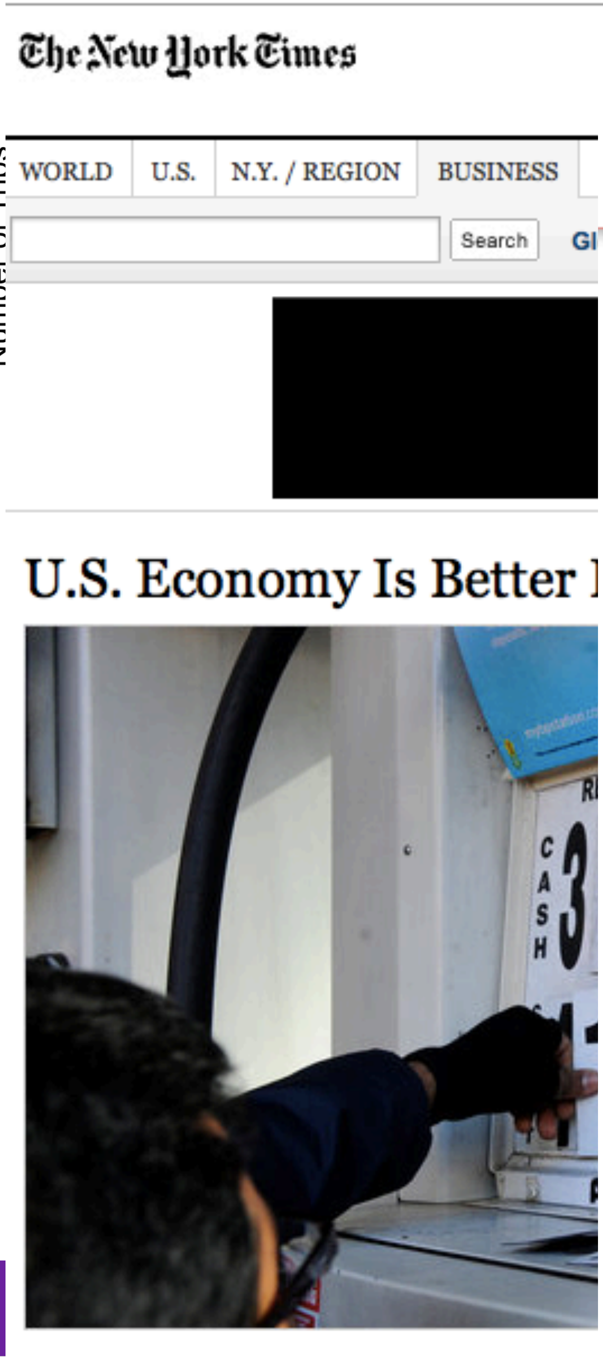
Understanding Events



- What happened around March/April 2011?

Work in progress

Understanding Events

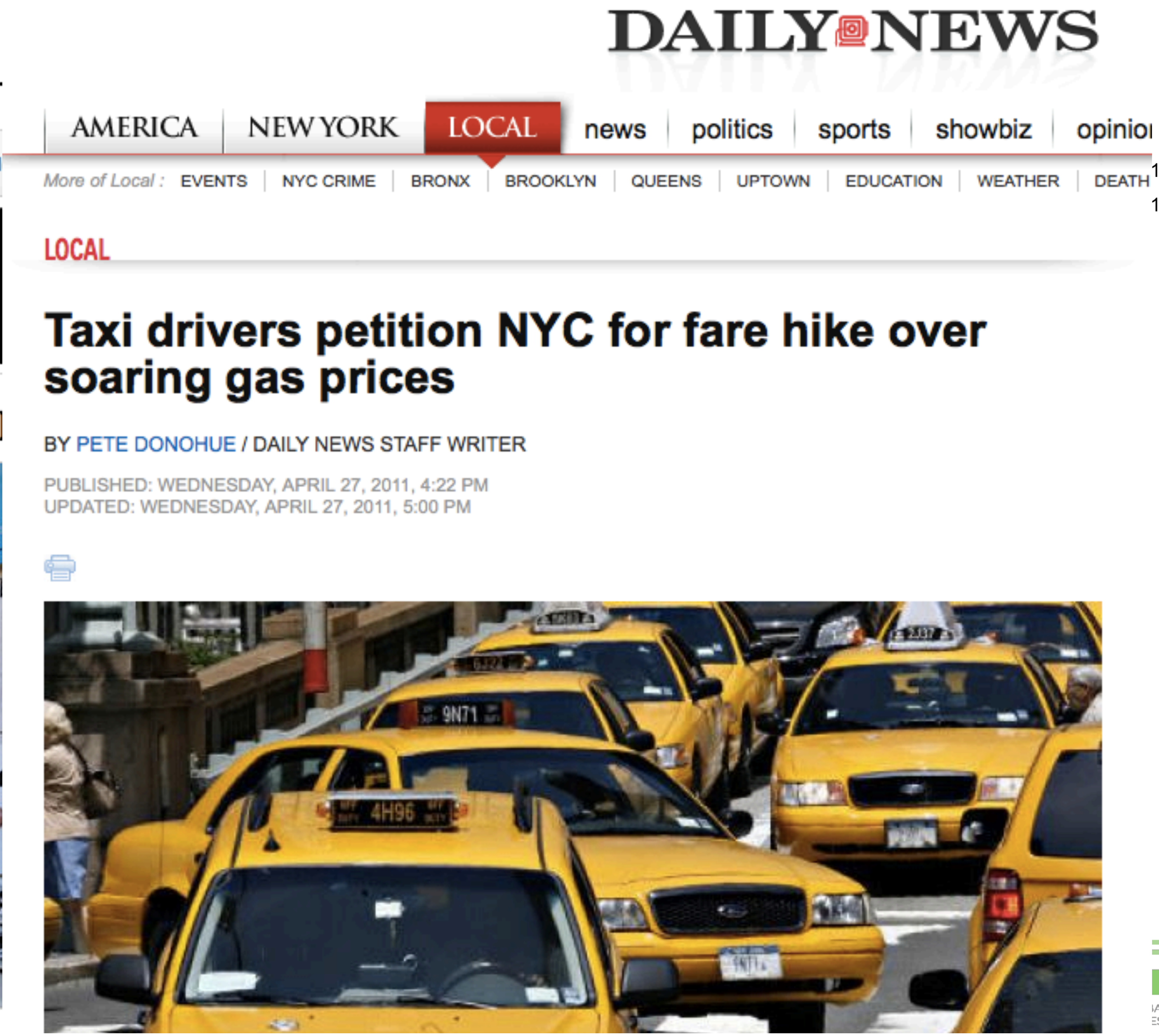


The New York Times

WORLD | U.S. | N.Y. / REGION | BUSINESS

[redacted]

U.S. Economy Is Better I



DAILY NEWS

AMERICA | NEW YORK | LOCAL | news | politics | sports | showbiz | opinio
More of Local : EVENTS | NYC CRIME | BRONX | BROOKLYN | QUEENS | UPTOWN | EDUCATION | WEATHER | DEATH

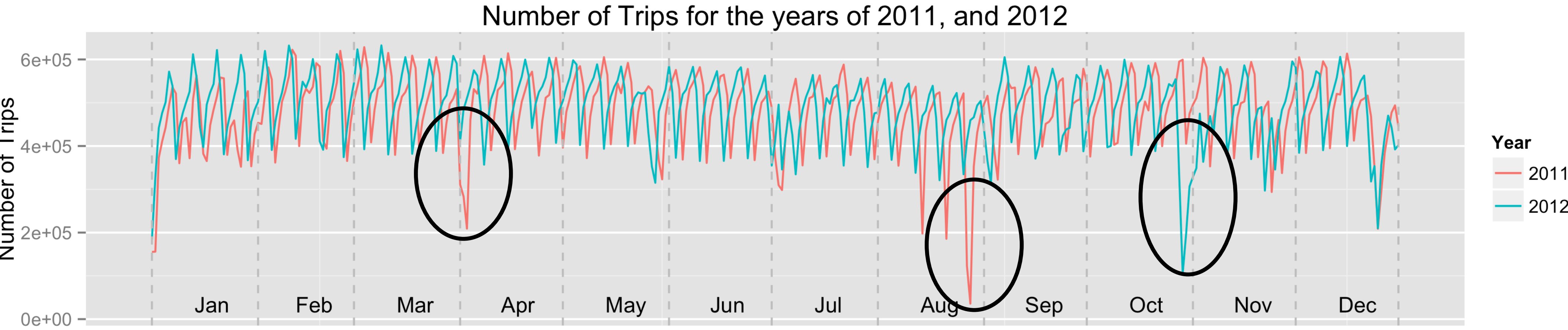
LOCAL

Taxi drivers petition NYC for fare hike over soaring gas prices

BY PETE DONOHUE / DAILY NEWS STAFF WRITER

PUBLISHED: WEDNESDAY, APRIL 27, 2011, 4:22 PM
UPDATED: WEDNESDAY, APRIL 27, 2011, 5:00 PM

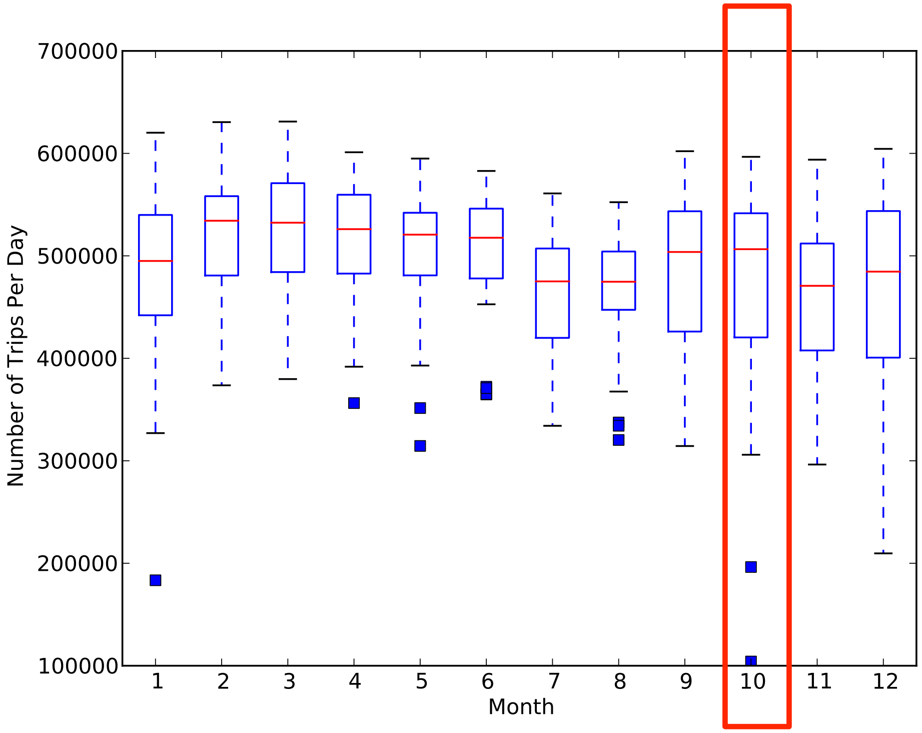
Understanding Events

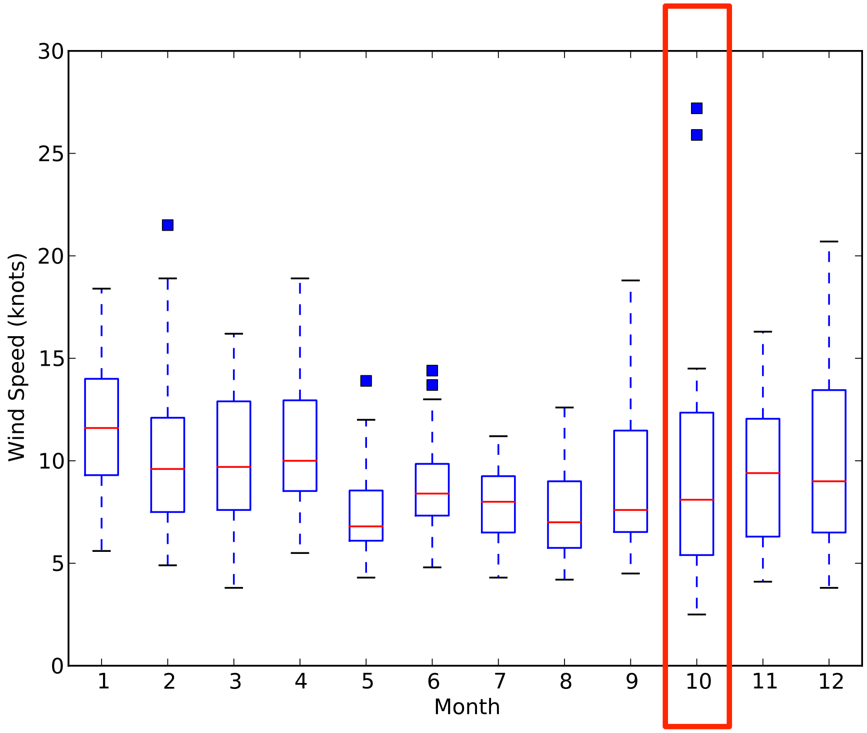


- What happened around March/April 2011?
  - Gas prices went up
  - How do gas prices affect taxi availability?
- What happened in Aug 2011 and Oct 2012?
  - Hurricanes Irene and Sandy
  - How does weather affect taxi service and mobility in a city?

Need to combine different data sets!

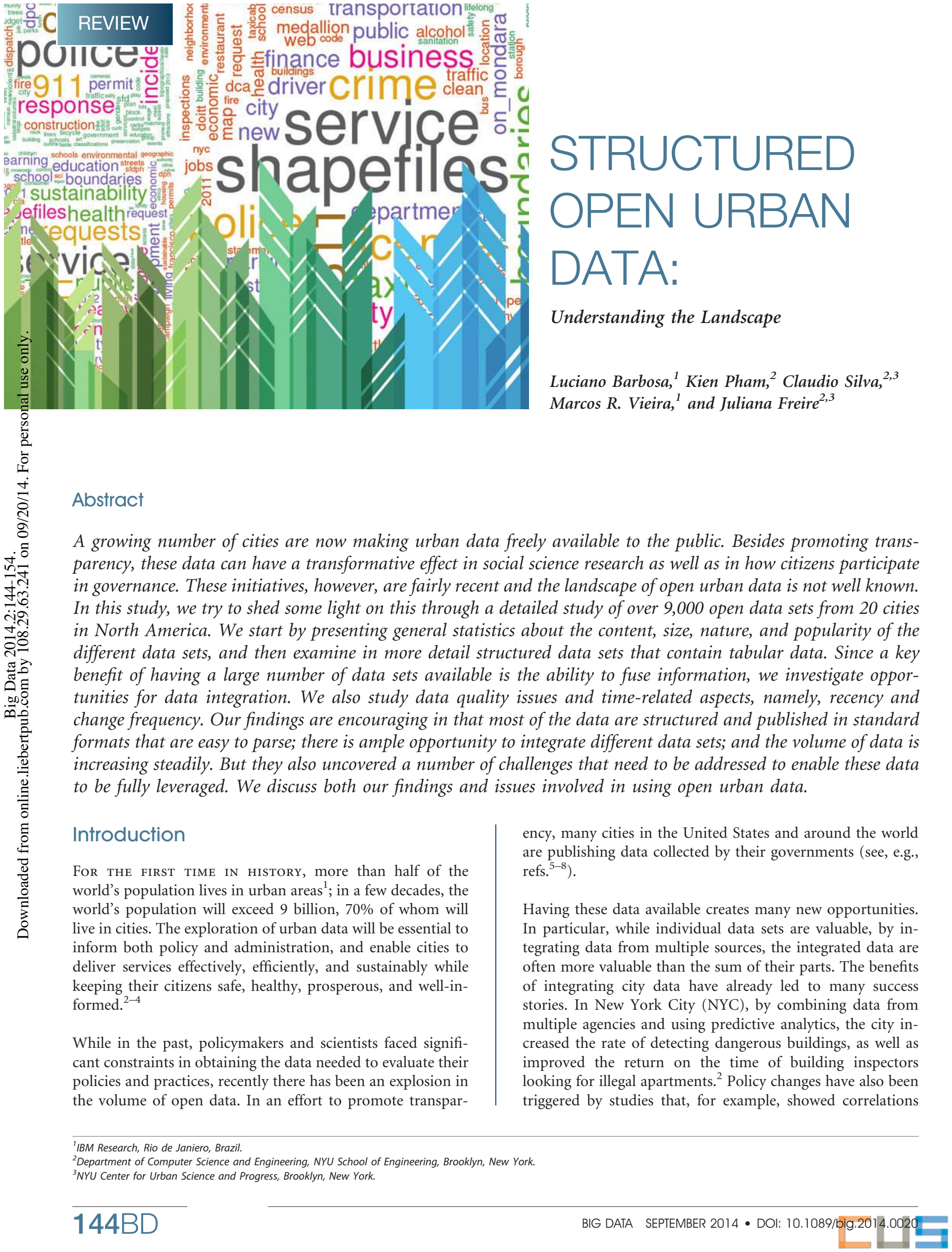
Combining Multiple Data Sets





Integrating Urban Data

- Many data sets available
- Trend: cities are opening their data
- Study: 20 cities in North America, 9,000 data sets
- Investigated
  - Nature of the data
  - Opportunities for integration



REVIEW

STRUCTURED OPEN URBAN DATA:

Understanding the Landscape

Luciano Barbosa,[1] Kien Pham,[2] Claudio Silva,[2,3] Marcos R. Vieira,[1] and Juliana Freire[2,3]

Abstract

A growing number of cities are now making urban data freely available to the public. Besides promoting transparency, these data can have a transformative effect in social science research as well as in how citizens participate in governance. These initiatives, however, are fairly recent and the landscape of open urban data is not well known. In this study, we try to shed some light on this through a detailed study of over 9,000 open data sets from 20 cities in North America. We start by presenting general statistics about the content, size, nature, and popularity of the different data sets, and then examine in more detail structured data sets that contain tabular data. Since a key benefit of having a large number of data sets available is the ability to fuse information, we investigate opportunities for data integration. We also study data quality issues and time-related aspects, namely, recency and change frequency. Our findings are encouraging in that most of the data are structured and published in standard formats that are easy to parse; there is ample opportunity to integrate different data sets; and the volume of data is increasing steadily. But they also uncovered a number of challenges that need to be addressed to enable these data to be fully leveraged. We discuss both our findings and issues involved in using open urban data.

Introduction

For the first time in history, more than half of the world's population lives in urban areas[1]; in a few decades, the world's population will exceed 9 billion, 70% of whom will live in cities. The exploration of urban data will be essential to inform both policy and administration, and enable cities to deliver services effectively, efficiently, and sustainably while keeping their citizens safe, healthy, prosperous, and well-informed.[2–4]

While in the past, policymakers and scientists faced significant constraints in obtaining the data needed to evaluate their policies and practices, recently there has been an explosion in the volume of open data. In an effort to promote transparency, many cities in the United States and around the world are publishing data collected by their governments (see, e.g., refs.[5–8]).

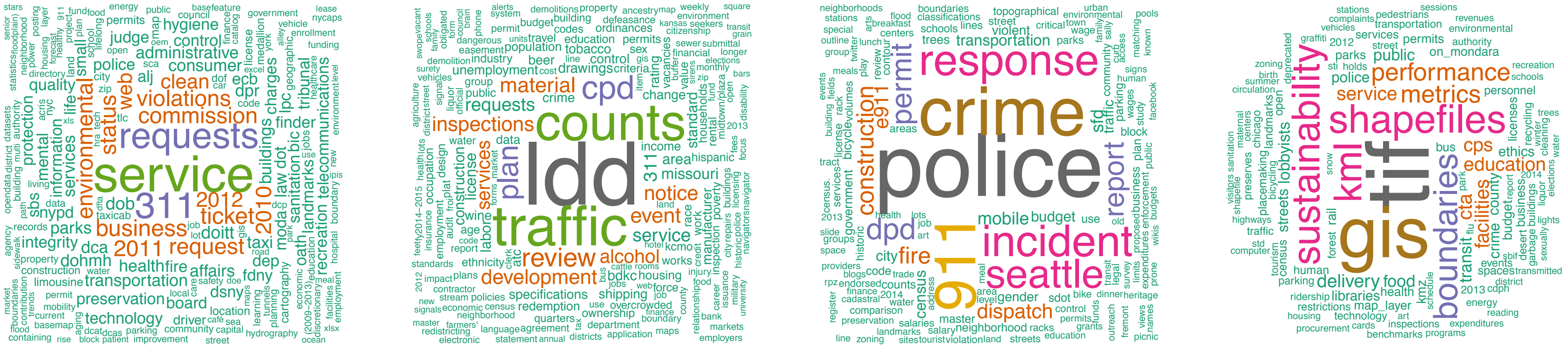
Having these data available creates many new opportunities. In particular, while individual data sets are valuable, by integrating data from multiple sources, the integrated data are often more valuable than the sum of their parts. The benefits of integrating city data have already led to many success stories. In New York City (NYC), by combining data from multiple agencies and using predictive analytics, the city increased the rate of detecting dangerous buildings, as well as improved the return on the time of building inspectors looking for illegal apartments.[2] Policy changes have also been triggered by studies that, for example, showed correlations

[1] *IBM Research, Rio de Janiero, Brazil.*
[2] *Department of Computer Science and Engineering, NYU School of Engineering, Brooklyn, New York.*
[3] *NYU Center for Urban Science and Progress, Brooklyn, New York.*

[Barbosa et al., Big Data 2014]

Key Findings

- 75% of the data sets are available in tabular formats, e.g., CSV
- Many topics are covered



(a) NYC

(b) Kansas City

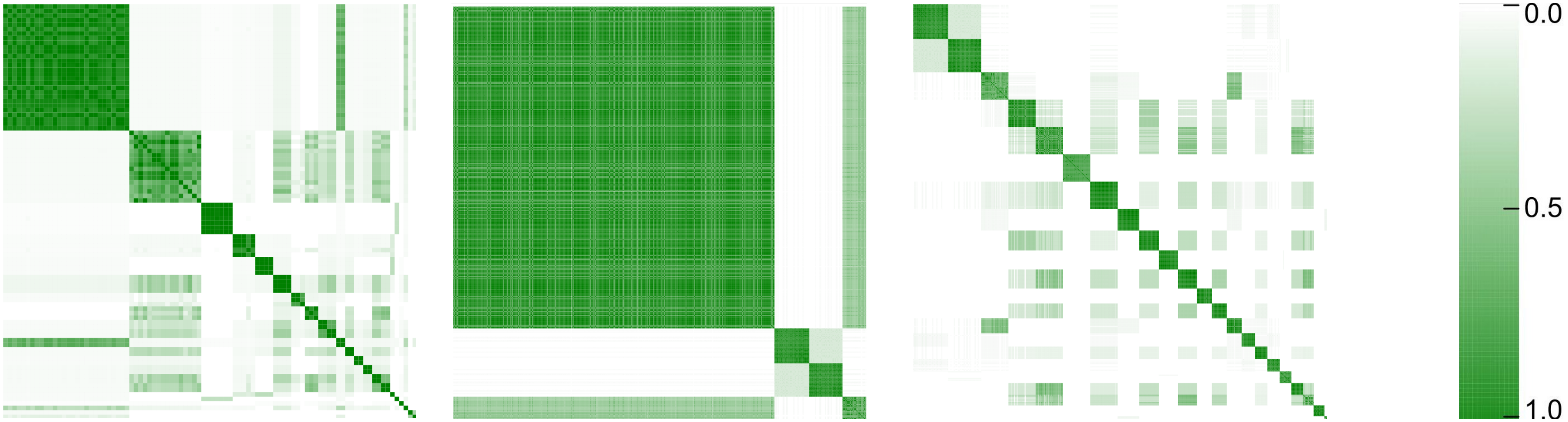
(c) Seattle

(d) Chicago

Key Findings

- Most data are available in tabular formats, e.g., CSV
- Many topics are covered
- Number of data sets is growing
  - In 2013, more data sets were added than in the 3 previous years combined
- Data is small: 70GB for all cities
  - Compare against 1 year of taxi data: 50GB/year
- There is ample opportunity for integration – significant overlap across tables: schema and spatial!

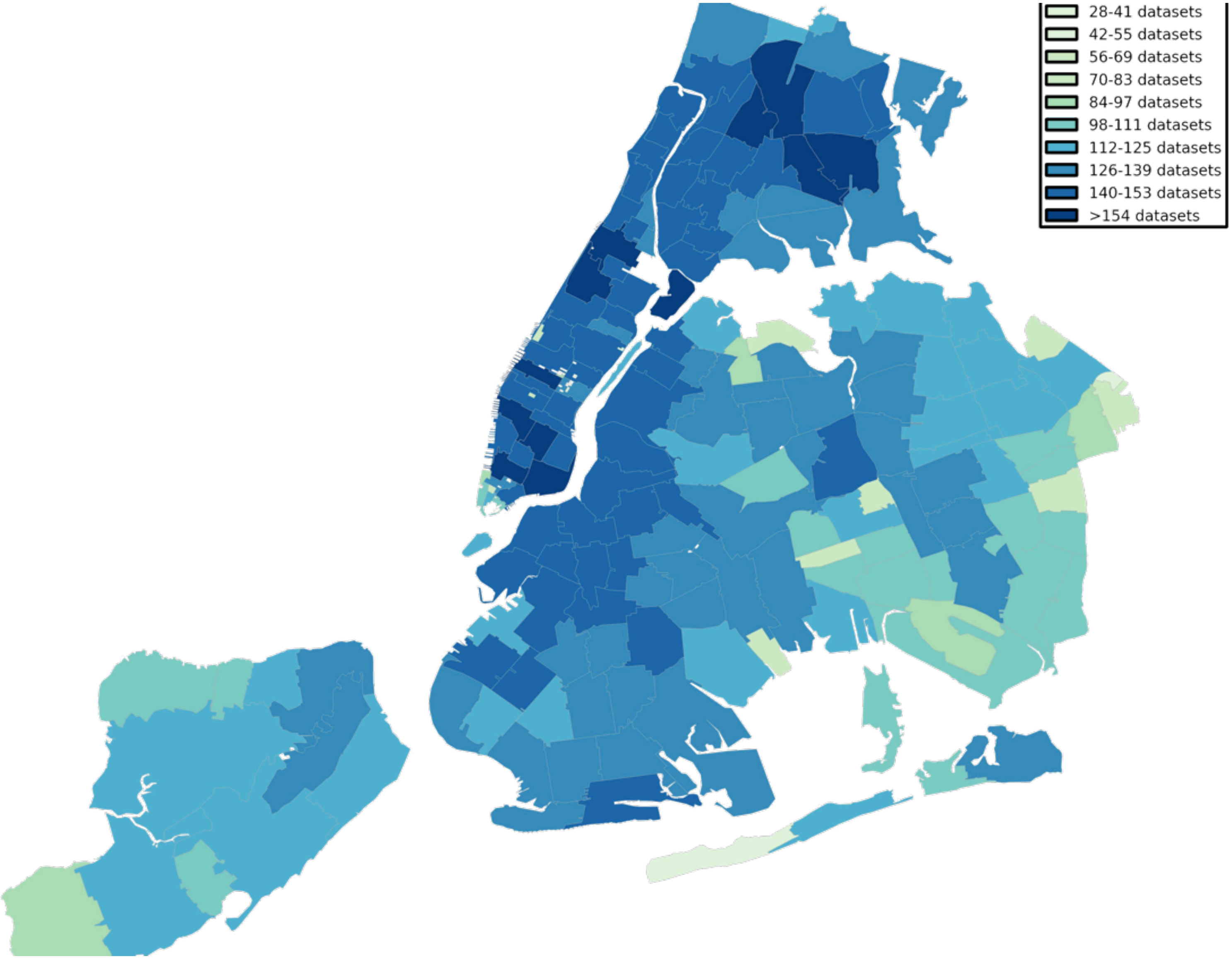
Integration Opportunities



(a) Boston (b) 4 largest NYC clusters (c) NYC without 311 data set (d) Similarity Scale

Attribute overlap

Integration Opportunities



Geographical coverage and overlap

Challenges

- The old data integration problem: heterogeneity
- But also: dirty data, lack of schema and type information
- Data are available, but can be hard to find – they are spread over many different sites
  - How to find data sets?
- Search interfaces are primitive, e.g., keyword-based search over meta-data
  - Need to support more complex queries, e.g., find all data sets that cover Lower Manhattan from Jan – March 2013.
- Too many data sets
  - How to find *related* sets?
  - How to integrate them?

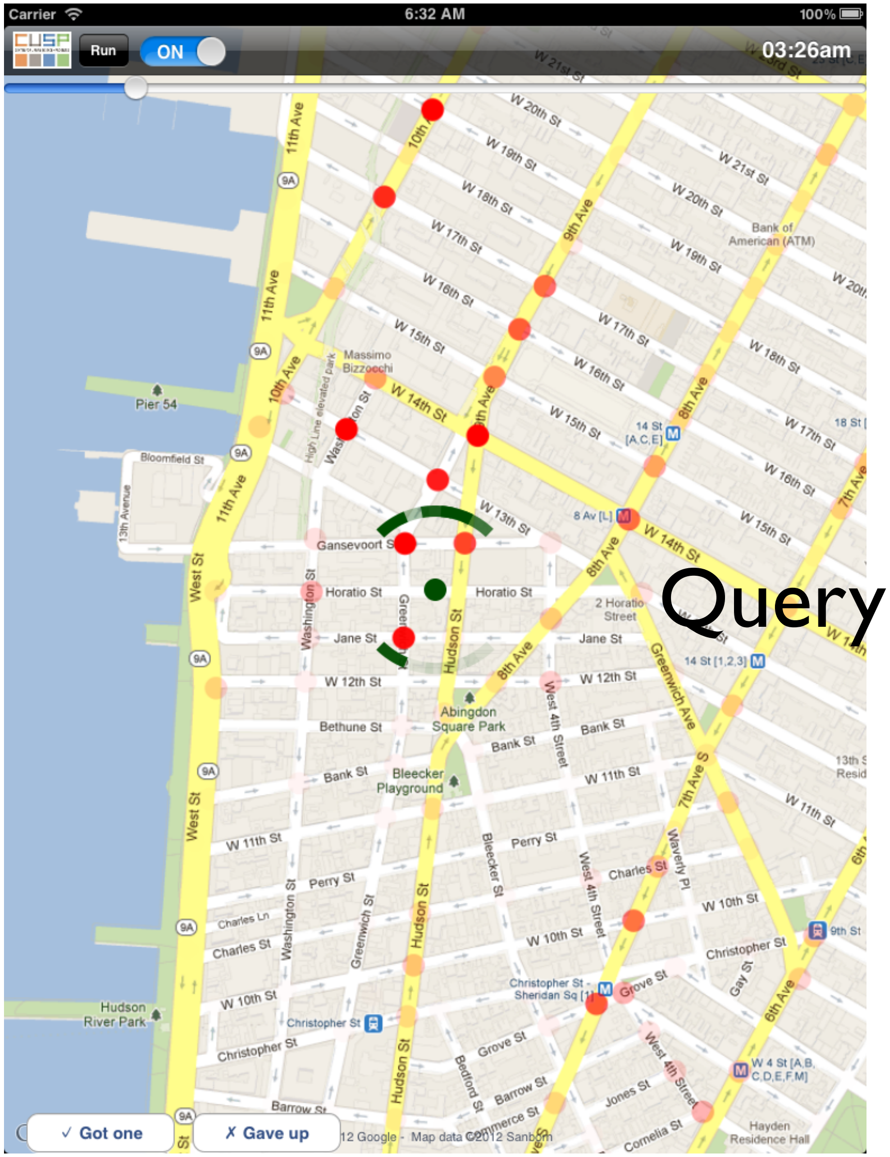
Need support for task-guided data integration

Planning for the Future *and Inferring Unknown Information*

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

CENTER FOR URBAN SCIENCE+PROGRESS

Find-a-Cab App



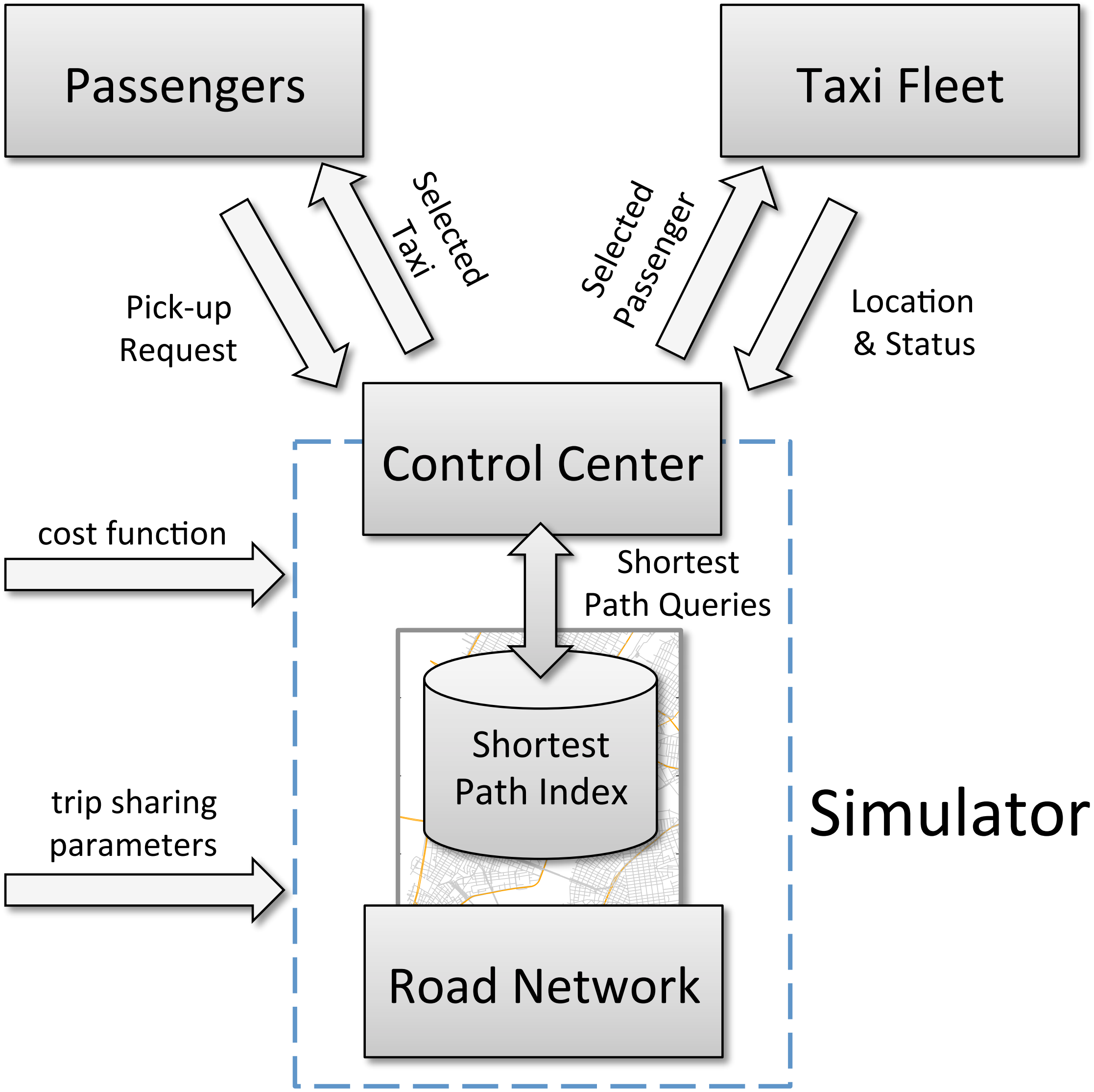
Querying the future!

Cab Ride Sharing

- Traffic and pollution are major problems in big cities
- NYC: 13,000 cabs, 500k+ trips/day
- Ride sharing can attenuate these problems – but it has never been widely adopted
  - Very controversial issue
- Challenge: Different and conflicting interests
  - Government: reduce traffic and pollution
  - Cab companies: maximize profits
  - Passengers: reach their destination quickly and cheaply
- Solution: use *historical data* to study and better understand the trade-offs

Our Approach: Data-Driven Simulation

- Real-time ride sharing
- Enables the study of different scenarios
  - Passenger preferences, e.g., max number of addl. stops, wait time
  - Vendor constraints: cab capacity, max number of shared trips
- Challenge: assigning trips to taxis is computationally expensive



[Ota and Vo, work in progress]

Our Approach: Data-Driven Simulation

- Optimization algorithm that is linear wrt the number of trips, and uses an efficient index to support shortest-path computations
- Compute sharing costs in parallel
- Simulate different days in parallel
- One simulation using over 150 million trips can be run in under 10 minutes using a 1200-core cluster
- Some results:
    - If each taxi is allowed to share up to 2 and 3 trips, the total travel distance is saved by 30% and 37% with less than 4 and 5 minute delay on average
    - Sharing with at most 1 trip leads to 19% saving with the average delay 2 minutes.

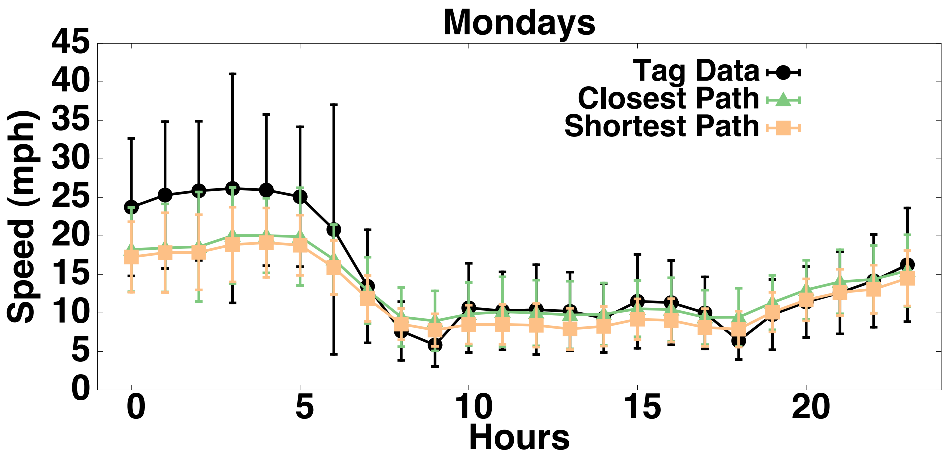
[Ota and Vo, work in progress]

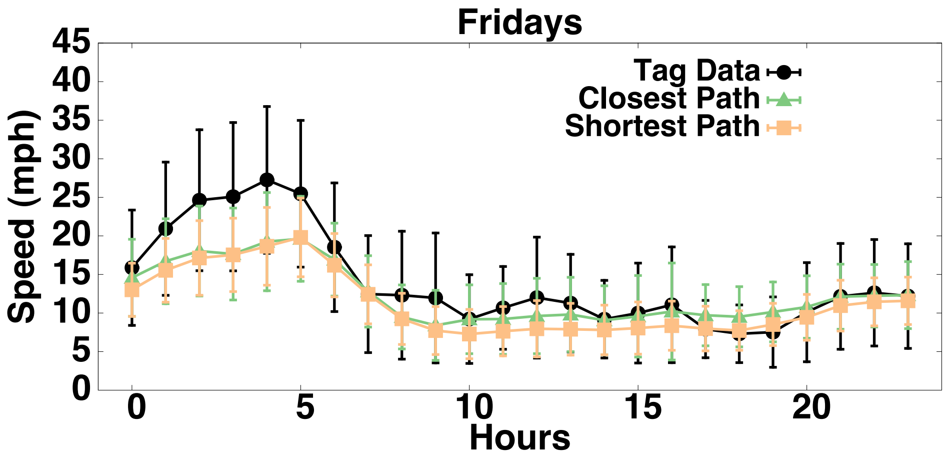
How does Traffic Move in Manhattan?

- There are sensors spread over the city, e.g., speed cameras, E-Zpass readers
- But coverage is very sparse
- Can we use the taxi data to infer the missing information?
- Challenge: taxi data contains only start and end positions of a trip

How does Traffic Move in Manhattan?

- Our approach:
  - Identify plausible routes for every trip
  - Use routes to infer traffic speed along the various roads – leverage the power of big data
- Validated results against E-Zpass data

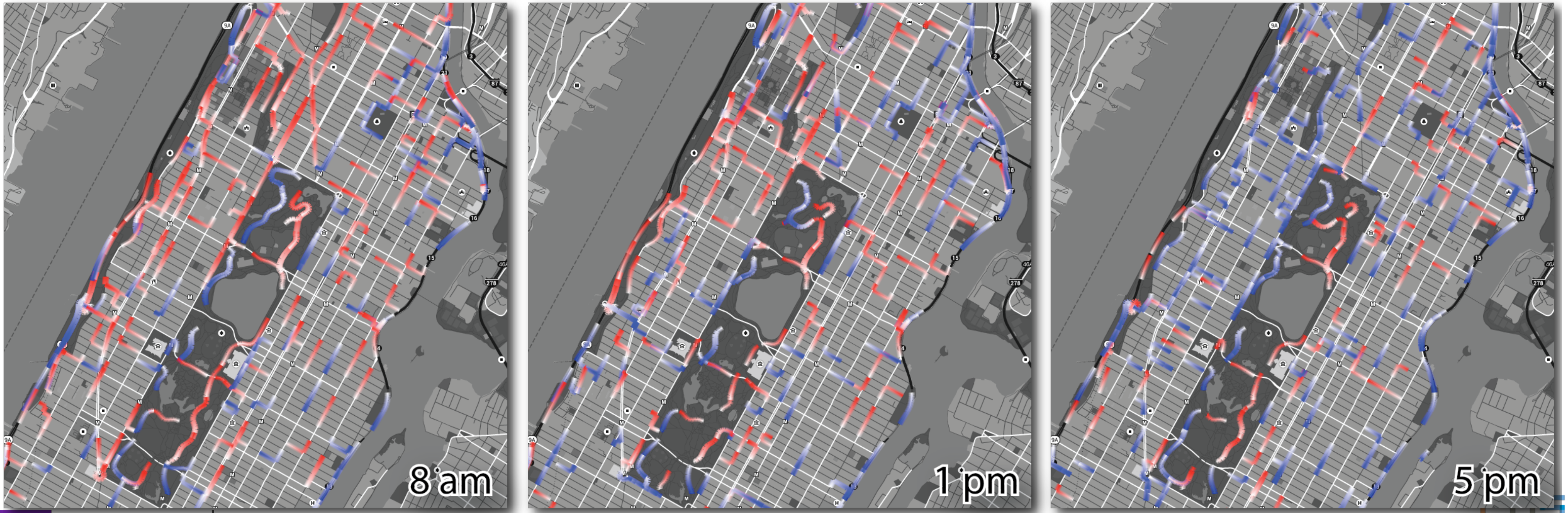




[Poco et al., work in progress]

How does Traffic Move in Manhattan?

- Our approach:
  - Identify plausible routes for every trip
  - Use routes to infer traffic speed along the various roads – leverage the power of big data
- Validated results against E-Zpass data
- Use visualization to explore traffic patterns



[Poco et al., work in progress]

Confession 1

There are lots of data, but some are not easy to get…

Need connections!

Confession 2

Why is TaxiVis so fast?

Not everybody knows what an index is…or can appreciate the intricacies of DB plumbing

My analogy: If the Yellow pages were not sorted alphabetically, how long would it take you to find a business phone number?

They got it!

Confession 3

Solving a real problem takes time

An end-to-end solution often requires expertise in different areas: data management, visualization, machine learning…

Need to collaborate!

If you collaborate with Vis folks, your papers will be beautiful and your plumbing will be more attractive

Confession 4

Building systems takes time

Building systems that others will use takes even more time

The impact of your work can be greatly magnified, but you can also waste a lot of time

Select carefully

NYU POLYTECHNIC SCHOOL OF ENGINEERING

CUSP
CENTER FOR URBAN SCIENCE+PROGRESS

Confession 5

I talked about many problems and cool solutions…
That others came up with!

I am just a supporting cast member

Building systems and solving real problems requires a great team

Select your people carefully

Conclusions

- Data exploration is challenging for both small and big data – need tools that are easy to use
  - The issue is not just size, but complexity
  - Need to enable domain experts to analyze data
- Visualization is a powerful tool for data exploration
  - Pictures help us think – substitute perception for cognition
  - To support interactive visualizations, need more synergy between Vis and Data Management
    - Vis community is building their own database solutions [Fekete & Silva, IEEE DEB 2012]
    - Can we do better?
  - A way to get more visibility for our plumbing work ;-)

Conclusions (cont.)

- Lots of open data available – opportunity to better understand how cities work
  - Need the ability to freely weave together multiple data sets
- Need to guide and support users more effectively explore data
  - Event-guided exploration
- Great potential for data-driven simulation
  - Need scalable techniques to handle large and complex data
- Many open problems around spatio-temporal data
  - Integration, querying, modeling
  - Querying the present: streaming data
- *DB community is well positioned to contribute and have tremendous practical impact – both in technology and education*

धन्यवाद

Thank you

Obrigada

Danke

Merci

Ευχαριστω

NYU | POLYTECHNIC SCHOOL OF ENGINEERING

CUSP
CENTER FOR URBAN SCIENCE+PROGRESS