Sao Paulo Advanced School on Smart Cities

Analysis and Visualization of Urban Data

Juliana Freire & Cláudio Silva
Computer Science & Engineering
Visualization, Imaging and Data Analysis Center (VIDA)
Center for Data Science (CDS)
Center for Urban Science and Progress (CUSP)

Joint work with Huy Vo, Harish Doraiswamy, Fernando Chirigati, Theo Damoulas, Nivan Ferreira, Masayo Ota, Jorge Poco, Yeuk Yin Chan, and many others

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

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, infrastructure
- Good news: Lots of data being collected from traditional and *unsuspecting* sensors

Data Exhaust from Cities

Infrastructure

Condition, Operations

Environment

Meteorology, pollution, noise, flora, fauna

People

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

Opportunity: Use data to make cities more efficient and sustainable, and improve the lives of their residents

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

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

Benefit residents

Urban Data: Success Stories

- NYC gets 25,000 illegal-conversion complaints a year and **only 200 inspectors** to handle them…
- Data-driven approach
  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%

Make City more efficient

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.

Enlarge This Image

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."

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

http://furmancenter.org/

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

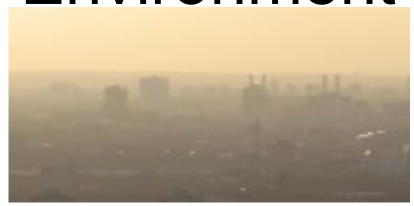
Affect policy

Urban Data: What is hard?

Infrastructure

Condition, operations

Environment



Meteorology, pollution, noise, flora, fauna

People

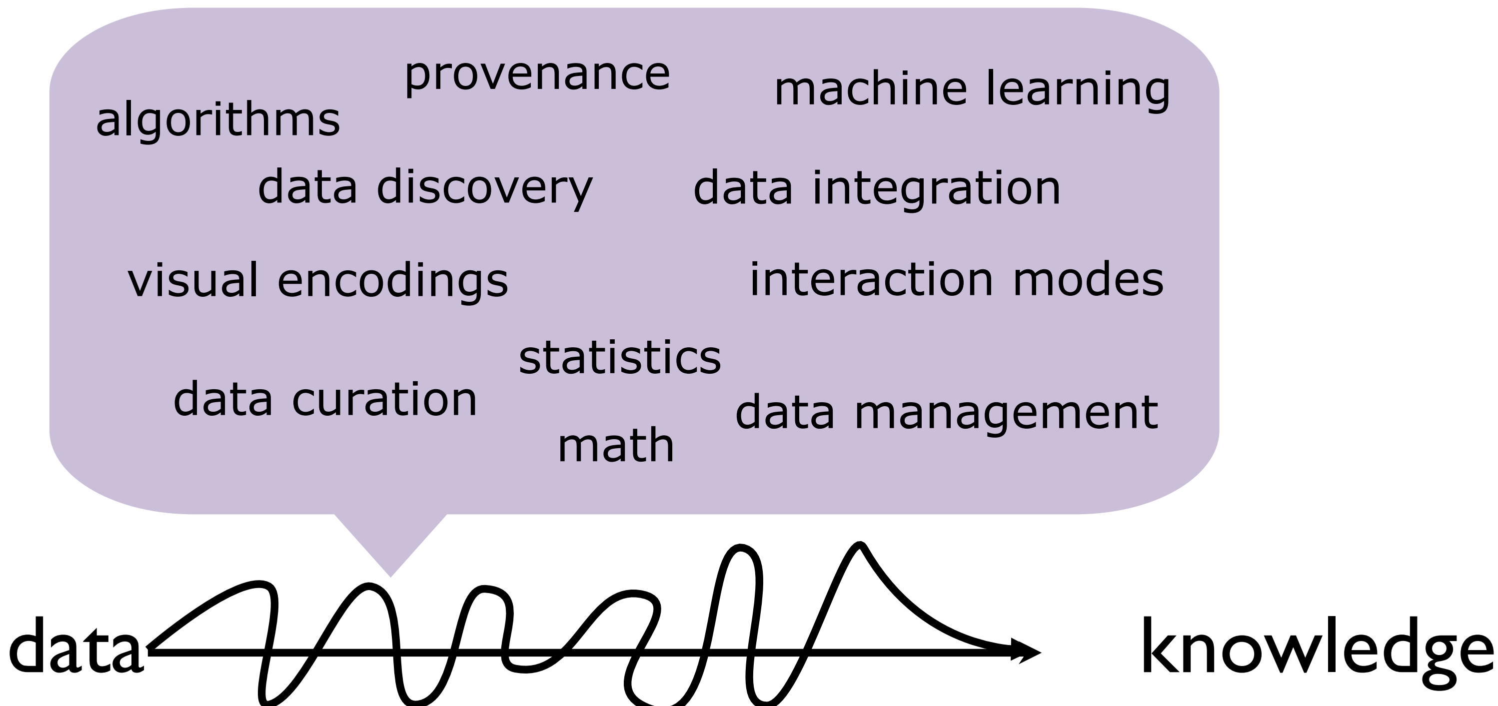
Relationships, economic activities, health, nutrition, opinions, …

- City components interact in complex ways
- Need to analyze the city *data exhaust* to understand these interactions
- Lots of **heterogeneous** and **dirty** data
- Processes occur over **time** and **space**

Urban Data: What is hard?

- Scalability for **batch** computations is not the biggest problem
  - Lots of work on distributed systems, parallel databases, cloud computing…
  - Elasticity: Add more nodes!
- Scalability for people is!

regardless of whether data are big or small



NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Urban Data Analysis: Common Practice

1. Domain experts and policy makers formulate hypotheses
2. Data scientists select data sets and slices, perform analyses, and derive plots
3. Domain experts examine the plots, goto 1.

Issues:

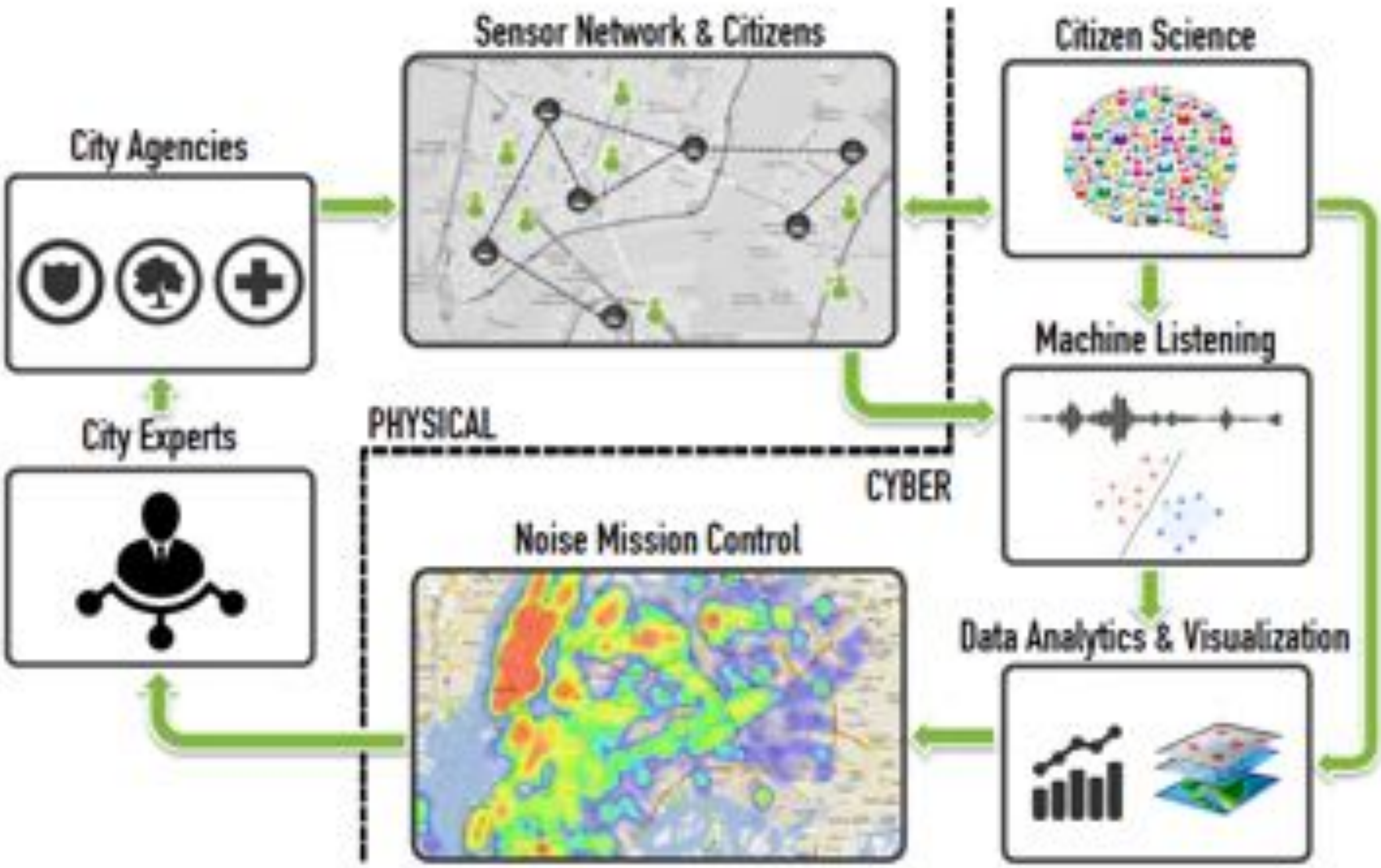
- Dependency on data scientists distances domain experts from the data
- Batch-oriented analysis pipeline hampers exploration – analyses are mostly confirmatory [Tukey, 1977]
- Data are complex – often multivariate spatio-temporal
- Analysis often limited to samples or small number of data slices
- Finding relevant data among the many data sets available

Urban Data Analysis: Desiderata

- Scalable tools and techniques that help *domain experts* find, clean, integrate, *interactively* explore and explain data
- Cater to different kinds of users with little or no CS training
- *Automate* tedious tasks as much as possible
- Guide users in the exploration process

Data analysis for all!

Sounds of New York City



NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

https://www.youtube.com/watch?v=d-JMtVLUSEg

Outline for Today

- What does the data look like?
- Big Problems
- Data Cleaning
  - Overview and Challenges
  - Cleaning the NYC Taxi Data: A Case Study
- Exploring Urban Data: Usability and Interactivity
- Finding Interesting Features
- Using Data to Discover and Explain Data

Opportunity: Lots of Open Data

As of December 2016, over 1,600 data sets are available on the NYC Open Data catalog.

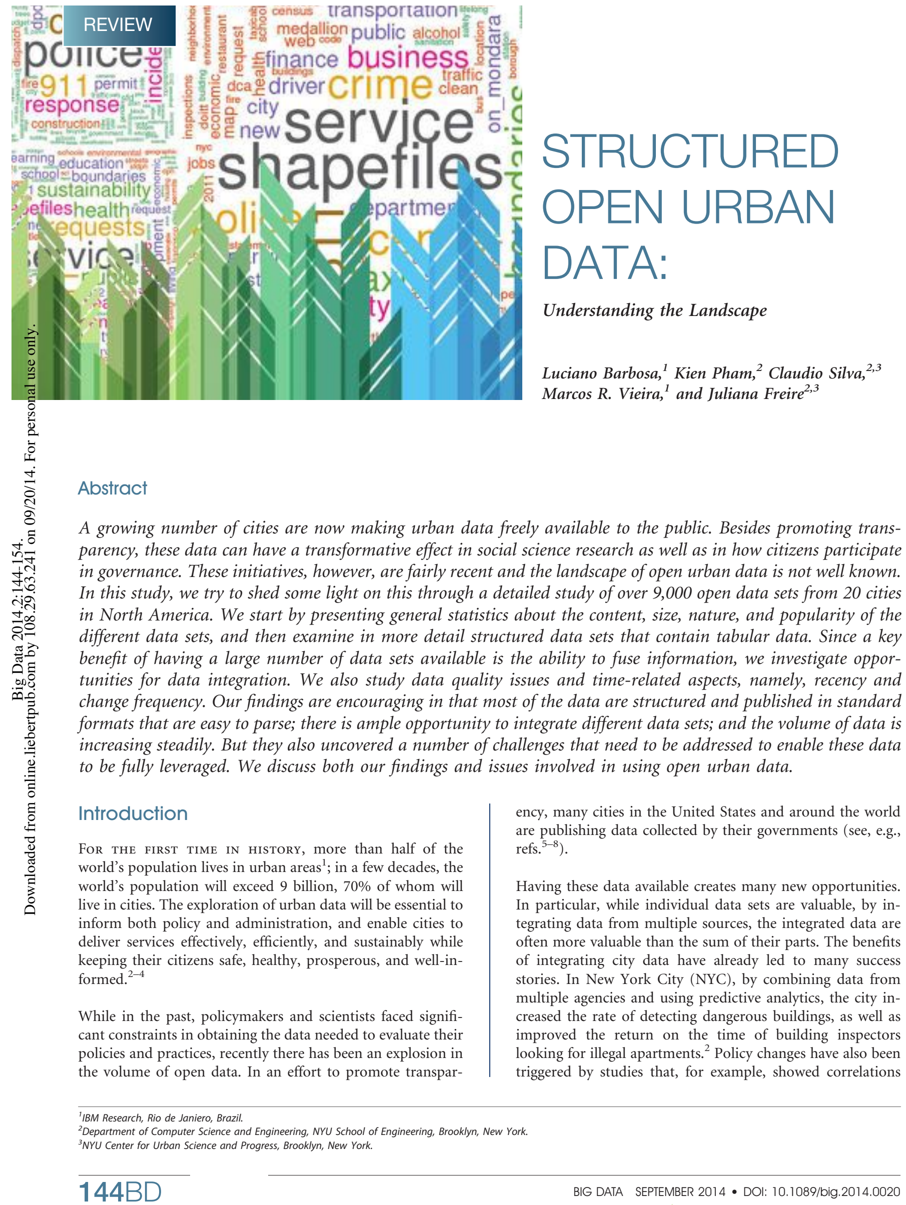
Open Urban Data (as of 2014)

- Study: 20 cities in North America, 9,000 data sets
- Investigated
  - Nature of the data
  - Opportunities for integration

“People are tribal, but data doesn’t care”

Mike Flowers

[Barbosa et al., Big Data 2014]



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

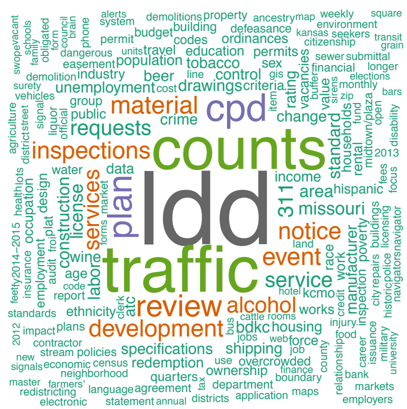
[1]IBM Research, Rio de Janiero, Brazil.
[2]Department of Computer Science and Engineering, NYU School of Engineering, Brooklyn, New York.
[3]NYU Center for Urban Science and Progress, Brooklyn, New York.

 BIG DATA SEPTEMBER 2014 • DOI: 10.1089/big.2014.0020

Some Findings

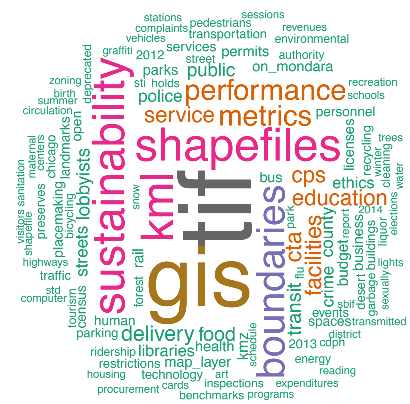
- 75% of the data sets are available in tabular formats, e.g., CSV: ability to pose ‘complex’ queries and re-use data cleaning/integration techniques
- Many topics are covered

(a) NYC



(b) Kansas City

(c) Seattle



(d) Chicago

Some 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 are big and small tables

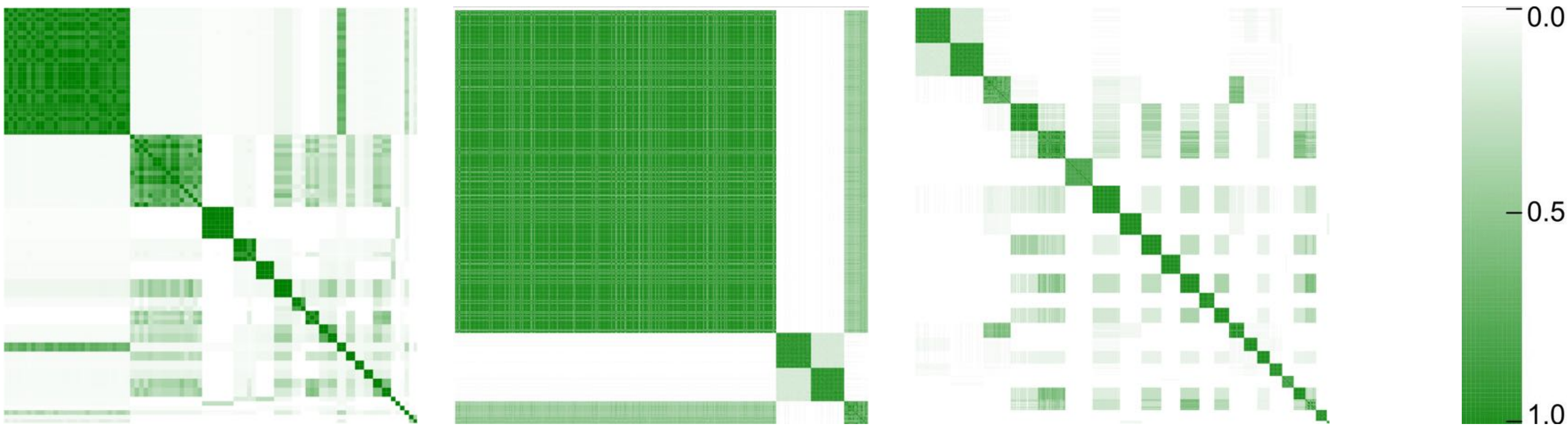
No. of records	*Percentage of total*
0–1K	65.3
1K–10K	17.0
10K–100K	11.7
100K–1M	5.5
1M–10M	0.3

>800M trips (5 years)

Some 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 are big and small tables
- Lots of spatio-temporal data:
  - Over 50% of the tables have lat+long and over 40% have date
- 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 among tables

- Potential for joining tables
- Hints about horizontally partitioned tables

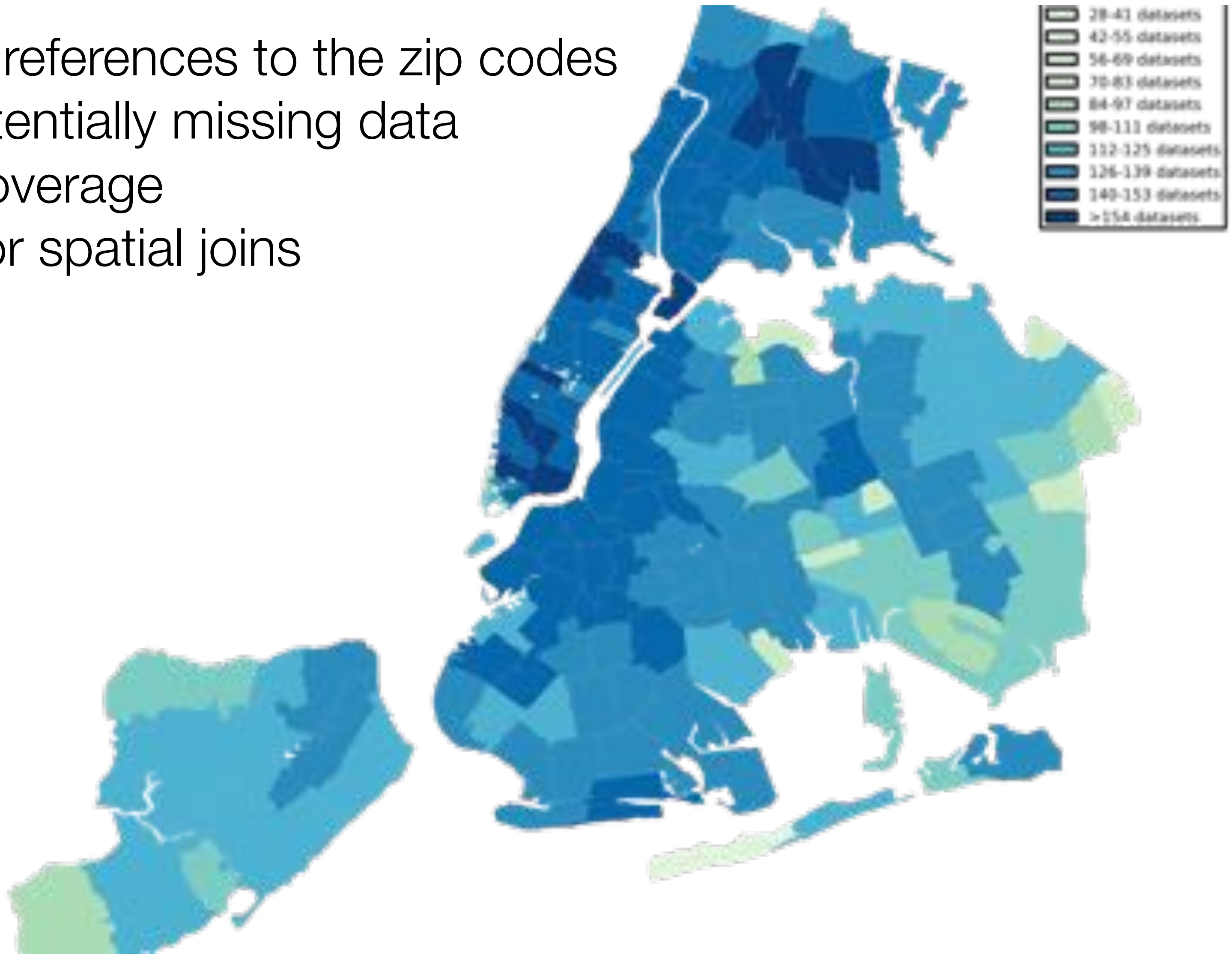
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Integration Opportunities

Frequency of references to the zip codes

- Identify potentially missing data
- Quantify coverage
- Potential for spatial joins



Geographical coverage and overlap

It’s not all roses…

Big Problems: Opportunities for Research

- Finding the Data
  - Data spread in many different repositories, e.g., NYC Open Data, Chicago Open Data, NYC MTA, …
  - Incomplete metadata

Data search engine

- Using the Data
  - Hard for domain experts without training in computing
  - Need to re-structure and integrate data
  - For Big Data, need advanced techniques, including the cloud and associated software stack

Usable tools

- Data Quality
  - Can we trust the data? No provenance is provided!
  - Lots of dirt…
  - Data cleaning and curation require substantial human intervention

Quality Issues in Urban Data

Challenge: Data Quality Issues

DOHMH New York City Restaurant Inspection Results

DBA	STREET	BUILDING
MADANGSUI	WEST 35 STREET	35
@NINE	9 AVENUE	592
TACO HUT	BROADWAY	3210

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Challenge: Data Quality Issues

DOHMH New York City Restaurant Inspection Results

DBA	STREET	BUILDING
MADANGSUI	WEST 35 STREET	35
@NINE	9 AVENUE	592
TACO HUT	BROADWAY	3210
TERROIR AT THE PORCH	**W 15th Street @ 10th Ave**	**HIGHLINE**

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Challenge: Data Quality Issues

DOHMH New York City Restaurant Inspection Results

DBA	STREET	BUILDING
MADANGSUI	WEST 35 STREET	35
@NINE	9 AVENUE	592
TACO HUT	BROADWAY	3210
TERROIR AT THE PORCH	**W 15th Street @ 10th Ave**	**HIGHLINE**

People that generate data get ‘creative’ to fit information to data models.

Lack of provenance information means we have to attempt to understand their decisions and the data generation process.

https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

Challenge: Data Quality Issues

- Columns containing Telephone Numbers in NYC Open Data
- Think of a (simple) way to distinguish the ‘Good’ from the ‘Bad’ <u>and</u> to transform the bad into good.

.	(000)000-0000
0	(201)368-1000
212 NEW YORK	(201)373-9599
311	(718) 206-1088
511	(718) 206-1121
911	(718) 206-1420
0000000000	(718) 206-4420
1111111	(718) 206-4481
1111111111	(914) 681-6200
1212669311	(718) 868-2300 x206
2012162746	(718) 206-0545/(718) 298-0117
2015954606	(718) 262-9072/(718) 658-1537
2033631907	(718) 297-4708/c: (347) 806-4588
9737924762	(888) 8NYC-TRS
9737924769	(888)-VETS-NYS
Fax7189801021	1-800-CUNY-YES
Fax:7189187823	800-624-4143

Challenge: Data Quality Issues

- Columns containing Boroughs, Cities, Neighborhoods in NYC Open Data
- Cities, neighborhoods and boroughs all mixed: how to fix this?

borough (0)
BRONX
BROOKLYN
MANHATTAN
QUEENS
STATEN ISLAND

city (1)
ASTORIA
BRONX
BROOKLYN
CHELSEA
CLINTON
FLUSHING
HARLEM
JAMAICA QUEENS
MANHATTAN
NEW YORK
STATEN ISLAND

manhattan neighborhood (2)
CHELSEA
CHINATOWN
CLINTON
HARLEM
SOHO
TRIBECA

Challenge: Data Quality Issues

- Assumption about valid values in a column, i.e., the domain

  Data Type (INT, DECIMAL, TEXT, DATE)

- Semantic constraints often not explicitly documented

  ZIP Code is a 5 digit number between 10000 and 99999

  Monetary value in US$

  Date in format YYYY-MM-DD

  Name in format <first> <last>

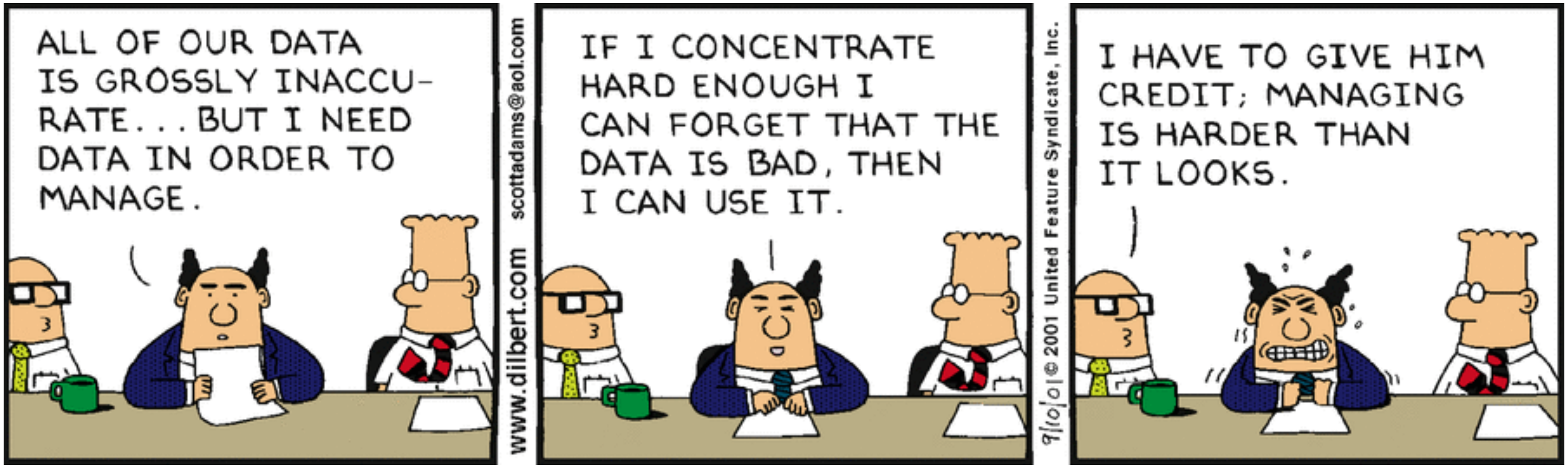
Attribute: illegal and missing values

- Pairs of records that contradict each other or violate a functional dependency ZIP → City

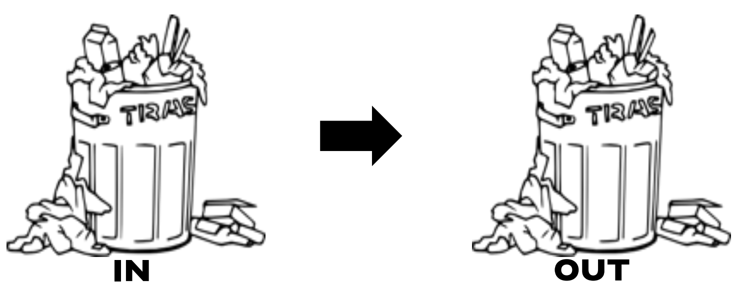
ZIP	City
10003	NYC
10003	Chicago

- Uniqueness violations, conflicting values, missing records

Data Quality



- **Data is a critical resource** that supports analytics and decision making
- As data volumes increase, so does the complexity of managing it and the **risks of poor data quality**.



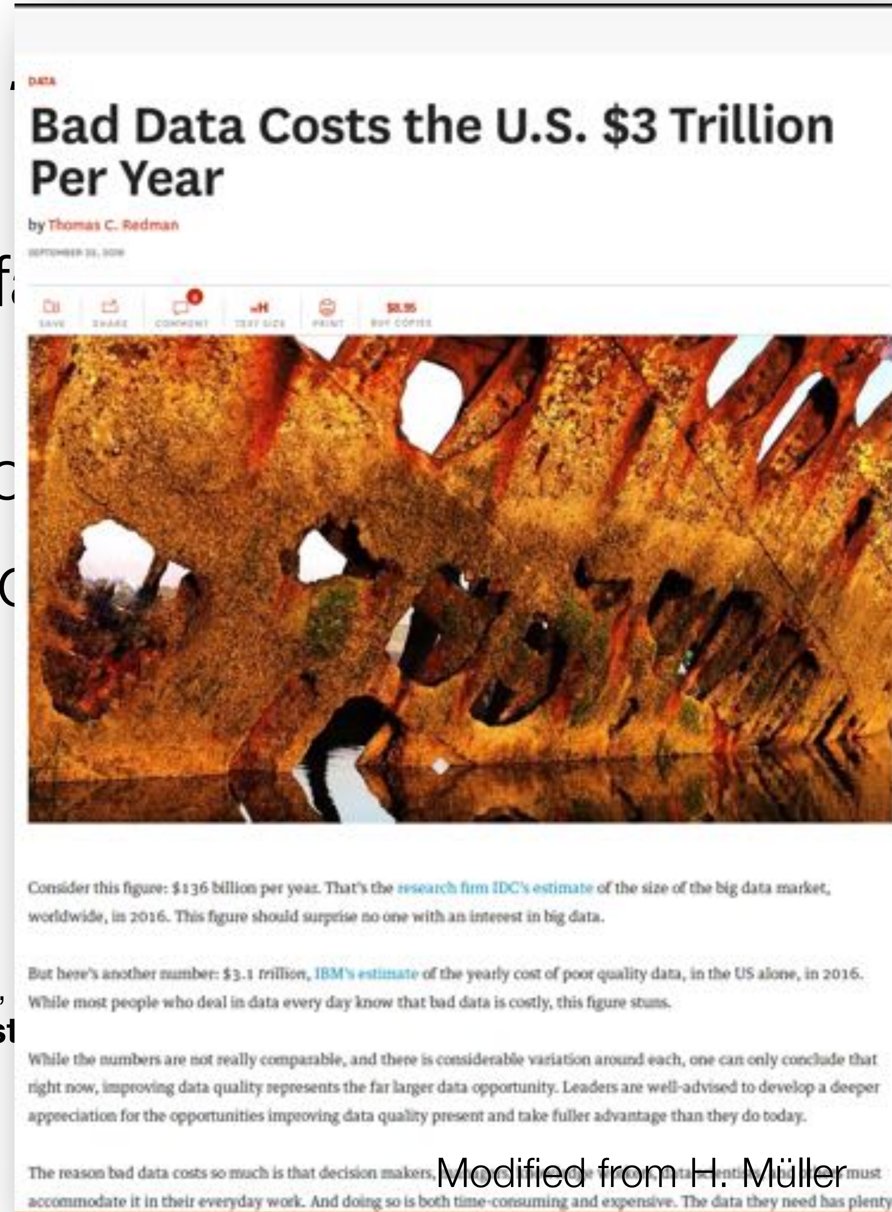
Modified from H. Müller

The Impact of Data Quality

Because of poor data quality …

- 88% of data integration projects f
  budgets
- 75% of organizations have additio
- 33% of organizations delayed or c
- $611bn per year is lost in the US

In [Marsh 2005] summarizing reports by **Gartner Group**,
Warehousing Inst


DATA

Bad Data Costs the U.S. $3 Trillion Per Year

by Thomas C. Redman

Consider this figure: $136 billion per year. That's the research firm IDC's estimate of the size of the big data market, worldwide, in 2016. This figure should surprise no one with an interest in big data.

But here's another number: $3.1 *trillion*, IBM's estimate of the yearly cost of poor quality data, in the US alone, in 2016. While most people who deal in data every day know that bad data is costly, this figure stuns.

While the numbers are not really comparable, and there is considerable variation around each, one can only conclude that right now, improving data quality represents the far larger data opportunity. Leaders are well-advised to develop a deeper appreciation for the opportunities improving data quality present and take fuller advantage than they do today.

The reason bad data costs so much is that decision makers, … must accommodate it in their everyday work. And doing so is both time-consuming and expensive. The data they need has plenty

Modified from H. Müller

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Are you excited about data cleaning?

15,078 views

Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says

Gil Press, CONTRIBUTOR

FULL BIO

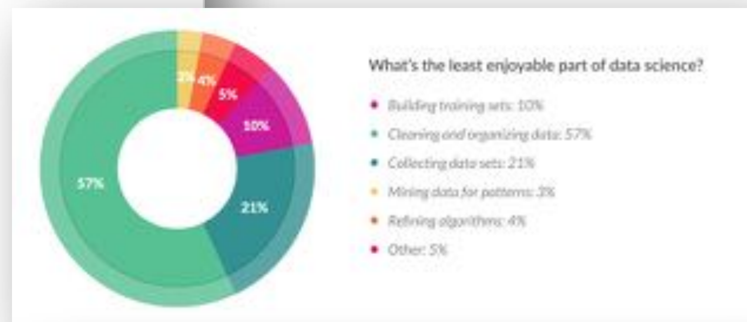
Opinions expressed by Forbes Contributors are their own.

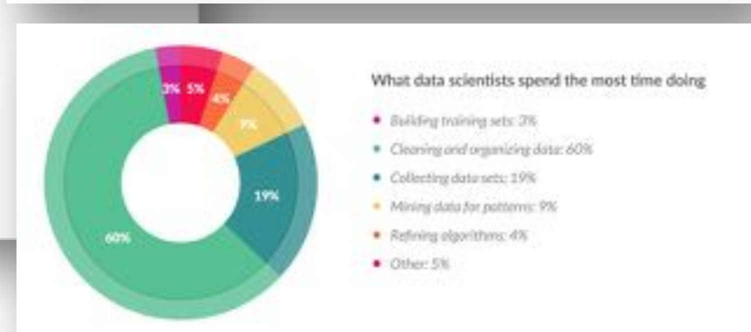
TWEET THIS

data scientists found that they spend most of their time massaging rather than mining or modeling data.

76% of data scientists view data preparation as the least enjoyable part of their work

A new survey of data scientists found that they spend most of their time massaging rather than mining or modeling data. Still, most are happy with having the sexiest job of the 21st century. The survey of about 80 data scientists was conducted for the second year in a row by CrowdFlower, provider of a "data enrichment" platform for data scientists. Here are the highlights:





- *Least enjoyable part of Data Science?*
  - Collecting data (21%)
  - Cleaning and organizing data (57%)
- *Spend most time doing*
  - Collecting data (19%)
  - Cleaning and organizing data (60%)

https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says

Modified from H. Müller

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Cleaning Small Data

- To extract value from data we must
  - Remove errors
  - Fill in missing information
  - Transform units and formats
  - Map and align columns
  - Remove duplicates records
  - Fix integrity constraint violations
- Specify all domain knowledge as integrity constraints
  - Reject updates that violate constraints
- Very rich literature and many tutorials
- Some tools are available
  - https://www.tamr.com, https://www.trifacta.com/products/wrangler, http://openrefine.org

Modified from Chu & Ilyas

Big Data + Data Quality: Challenges

- Constraints are not know a priori…
- Size: huge volume of data from multiple sources
- Complexity: large variety of data and sources
- Speed: dynamic data, collected and analyzed at high velocity
- Evolution: considerable variability of data, semantics over time
- Active area of research
  - Learn/infer models (semantics) from the data
  - Automatically identify data glitches
- Need (semi) automated methods and toolkits
  - Get ready to build your own!

Complete domain knowledge infeasible

Domain knowledge becomes obsolete

Modified from D. Srivastava

Toolbox of a Data Cleaner

- *External (High Quality) Data Sources*
  - E.g., lookup tables for city names and ZIP codes
- *Integrity Constraints*
  - Define and enforce constraints that high quality data adhere to
- *Regular Expressions*
  - Define format of values
- *String Similarity Functions*
  - Identify typos at data entry
  - Find records that represent the same entity (duplicates)
- *Conflict Resolution Functions*
  - Resolve contradicting information (in data integration)

Find Attribute Outlier Values

- *Sort attribute values in alphabetical order*
  - ‘Interesting’ values often appear at the beginning and end of list

The following examples are from the ***DOB Permit Issuance*** *dataset in* ***NYC Open Data***

owner_s_business_name
(JOANNE H. SIEGMUN 2ND OWNER)
(PERSONAL RESIDENCE)
(PRIVATE RESIDENCE)
(TENANT IN COMMON)
(TENANTS IN COMMON)
****
****
****
++++++++++++++++++++++++++
-
--

.
..

[…]

[…]

______N/A
altered state restoration
c/o Bowery Hotel
c/o Cooper Square Realty
c/o Leibovitz Studio
individual
mtp investment
n/a
na
new hempstead home for the adult
none
not applcable
owner
renaissanc
same
sierra realty corp.
wm maidmanfamily lp

Outliers in Alphabetical Order

city
(646)4396000
, FLORAL PARK
,ELMSFORD
.
1
10012
10013
10452
10462
105

A large number of quality problems are a result of 'parsing errors' or invalid file formats (e.g., too many or missing column delimiters in CSV file).

```
QUEENS|4144683|147-57 |78 AVE |421156046|01|A1||06688|00040
|408|11367|1|YES|||PL|ISSUED|RENEWAL|PL|02| | |NOT APPLICABLE
|11/06/2016|11/06/2016|11/06/2017|11/10/2015|CONSTANTINE |KOUMPAROULIS
|ARIANA CONTRACTING INC |7187215018|MASTER PLUMBER |0001101| | | | | |
| | |INDIVIDUAL ||N/A |ARTUR |KHAIMOV |147-57 |78TH AVENUE |KEW
GARDENS |NY|11367 |6464022132|11/07/2016
```

Find Attribute Outlier Values

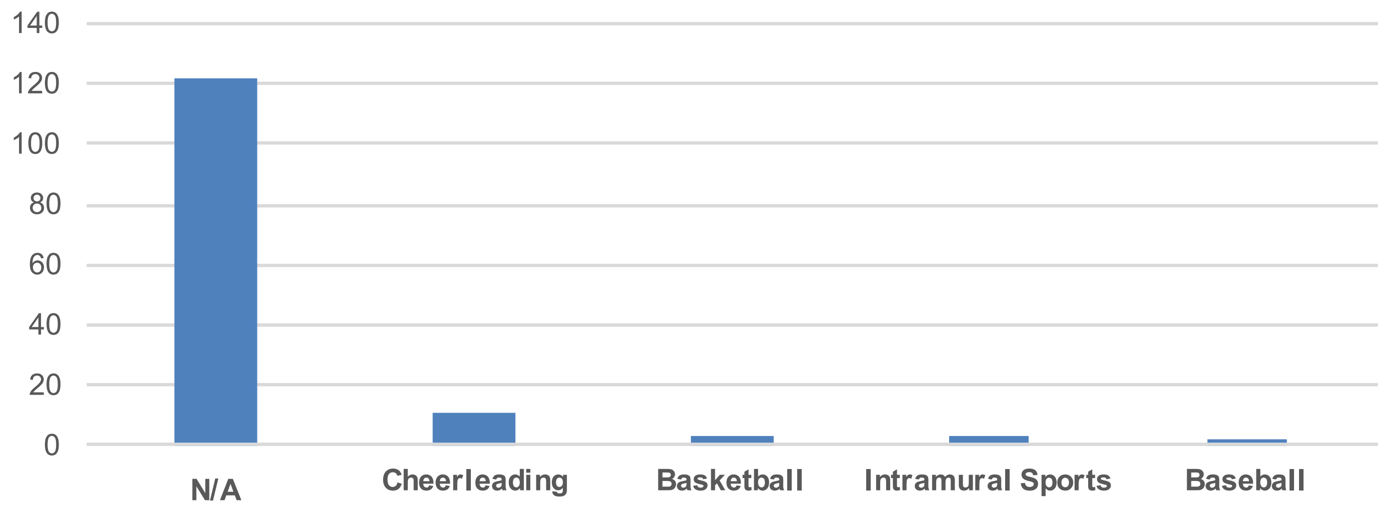
- *Sort attribute values in alphabetical order*
  - ‘Interesting’ values often appear at the beginning or end of list.
- *Frequency outliers*
  - NULL values sometimes have significantly different frequency (high or low) compared to other column values.

Frequency Outliers

DOE High School Directory 2013-2014

NYC Open Data

school_sports



Frequency Outliers (cont.)

- *Values that frequently occur as high frequency outliers*
  - Values that occur with frequency >50% in + 15,000 columns of NYC Open Data datasets

```
0                     (x 262)
N/A                   (x 71)
UNSPECIFIED           (x 67)
S                     (x 57)
-                     (x 50)
0.00                  (x 47)
NY                    (x 38)
1                     (x 25)
0.0                   (x 20)
IND                   (x 12)
CLOSED                (x 10)
100                   (x 8)
NOT AVAILABLE         (x 8)
0 UNSPECIFIED         (x 6)
NONE                  (x 5)
```

Find Attribute Outlier Values

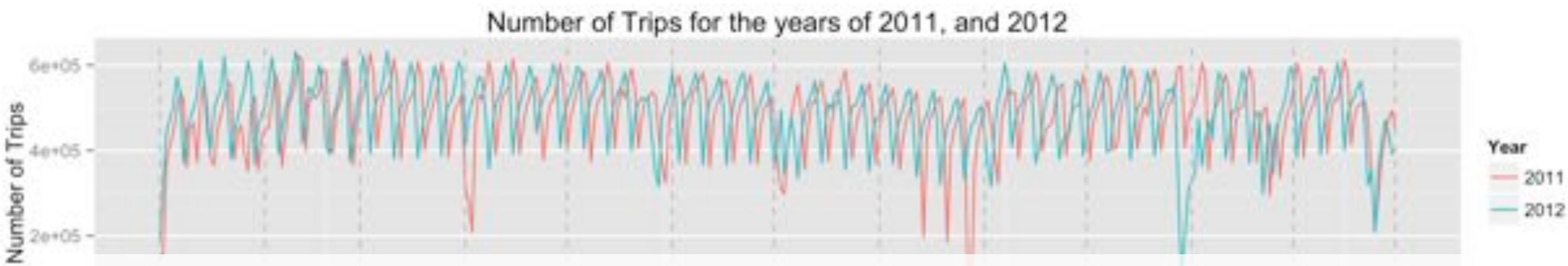
- *Sort attribute values in alphabetical order*
  - ‘Interesting’ values often appear at the beginning or end of list
- *Frequency outliers*
  - NULL values sometimes have significantly different frequency (high or low) compared to other column values
- *Regular expressions*
  - Find values that do not match the expected format of a column
- Often identify outliers and potential problems during data exploration

Exploring Urban Data: A Look into Quality issues in Taxi Trips

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

NYC Taxis



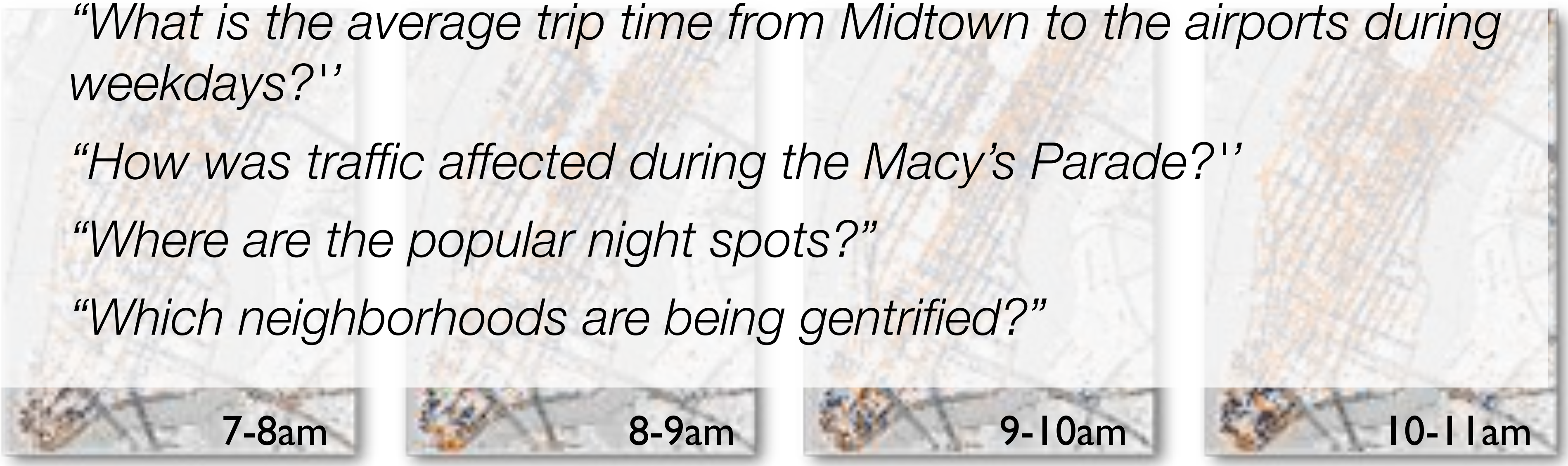
Taxis are *sensors* that can provide unprecedented insight into city life: economic activity, human behavior, mobility patterns

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

“How was traffic affected during the Macy’s Parade?''

“Where are the popular night spots?”

“Which neighborhoods are being gentrified?”



Taxi Data: What to Clean and not to Clean

Dataset	Statistic	Trip Duration (min)	Trip Distance (mi)	Fare Amount (US$)	Tip Amount (US$)
2008	Min	0.00	0.00	0.00	0.00
	Avg	16.74	2.71	0.09	0.10
	Max	1440.00	50.00	10.00	8.75
2009	Min	0.00	0.00	2.50	0.00
	Avg	7.75	6.22	6.04	0.38
	Max	180.00	180.00	200.00	200.00
2010	Min	-1,760.00	-21,474,834.00	-21,474,808.00	-1,677,720.10
	Avg	6.76	5.89	9.84	2.11
	Max	1,322.00	16,201,631.40	93,960.07	938.02
2011	Min	0.00	0.00	2.50	0.00
	Avg	12.35	2.80	10.25	2.22
	Max	180.00	100.00	500.00	200.00
2012	Min	0.00	0.00	2.50	0.00
	Avg	12.32	2.88	10.96	2.32
	Max	180.00	100.00	500.00	200.00

Negative values are clearly errors.
But high tip may not be an error…

Different processes were used to process data in different years,
but no provenance information is provided

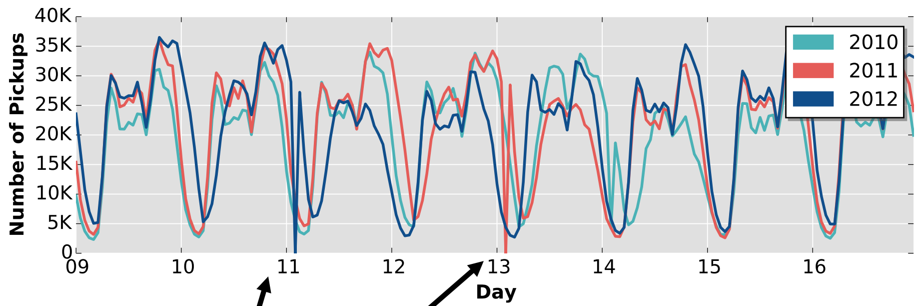
[Freire et al., IEEE DEB 2016]

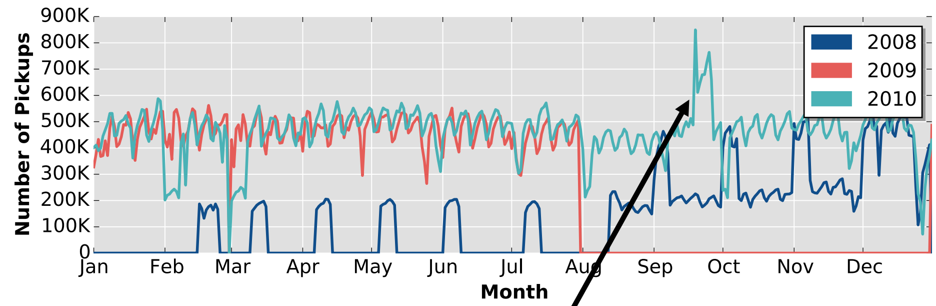
Taxi Data: What to Clean and not to Clean

(a) (b) (c)

Need to consider spatial constraints:
Trips in rivers, ocean and Central America

Taxi Data: What to Clean and not to Clean





No trips at 2am

Missing data in 2008

Missing data in 2009

Daylight savings:
March 13, 2011
March 11, 2012

Big spike on Sept 19th, 2010

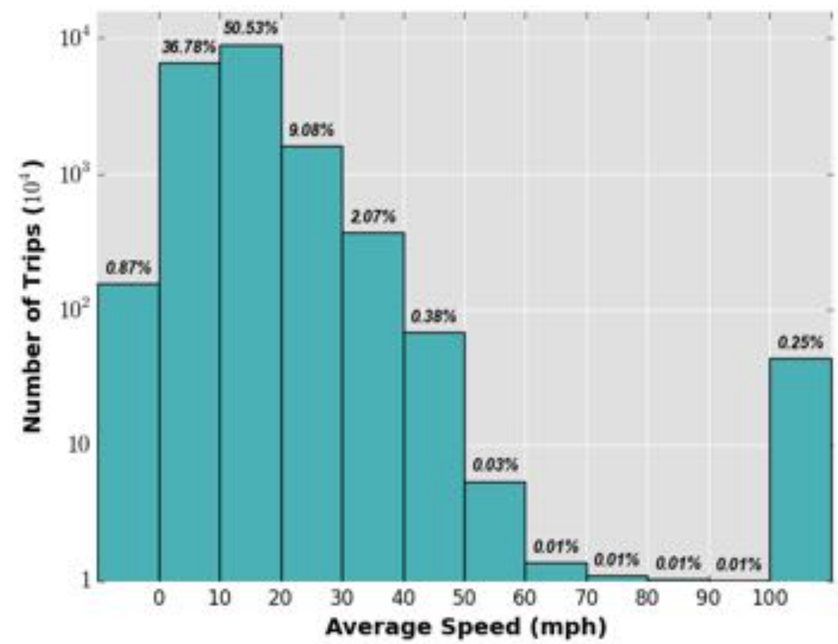
Unusually large number of consecutive and extremely short trips (lasting less than a minute)

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Taxi Data: What to Clean and not to Clean

- Ghost trips
  - Overlapping trips for the same taxi, i.e., for a given taxi, a new trip starts before the previous trip has ended
- Speed too high or too low
  - Incorrect values can negatively impact predictive models, e.g., which rely on average speeds
  - Speed = 0, easily an error
  - But what about high speeds?



Takeaway: Big Urban Data Cleaning

- Data cleaning has been performed as a pre-processing step

  Dirty Data → *Clean Data*

- Cleaning is an integral part of data exploration: constraints that should be checked in the cleaning function, and which might not be evident at first, are naturally discovered
- Different question/analyses require different cleaning strategies

$$DirtyData \times UserTask \rightarrow (CleanData, Explanation)$$

Takeaway: Big Urban Data Cleaning (cont.)

- Spatio-temporal data adds a new set of constraints and issues that need to be considered
- Visualization is essential!
- Traditional cleaning techniques are useful
- It is not always clear what is dirt and what is a feature
- Need domain knowledge
- Promising research direction: New techniques that leverage multiple data sets
  - Holistic data cleaning and integration
  - Use data to explain data (more soon!)

Data Cleaning References

- Tutorial: *Data Cleaning: Overview and Emerging Challenges* http://sigmod2016.org/sigmod_tutorial1.shtml
- Tutorial: Knowledge curation and knowledge fusion: challenges, models, and applications (SIGMOD 2015) http://lunadong.com/talks/KFTutorial_sigmod.pptx
- *Profiling relational data: a survey.* VLDB J. 24(4): 557-581 (2015)

Exploring Urban Data: Usability and Interactivity

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Exploring Taxi Data: Challenges

- Data: ~500k trips/day; 868 million trips in 5 years
  - *spatio-temporal:* pick up + drop off
  - *trip attributes*: e.g., distance traveled, fare, tip
- Government, policy makers and scientists are unable to *interactively* explore the *whole* data
  - Too many data slices to examine
- Our goal: Design a *usable* interface, efficiently support *interactive* + *exploratory* queries

Exploring Taxi Data

http://www.taxivis.org

Usability through Visual Operations

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

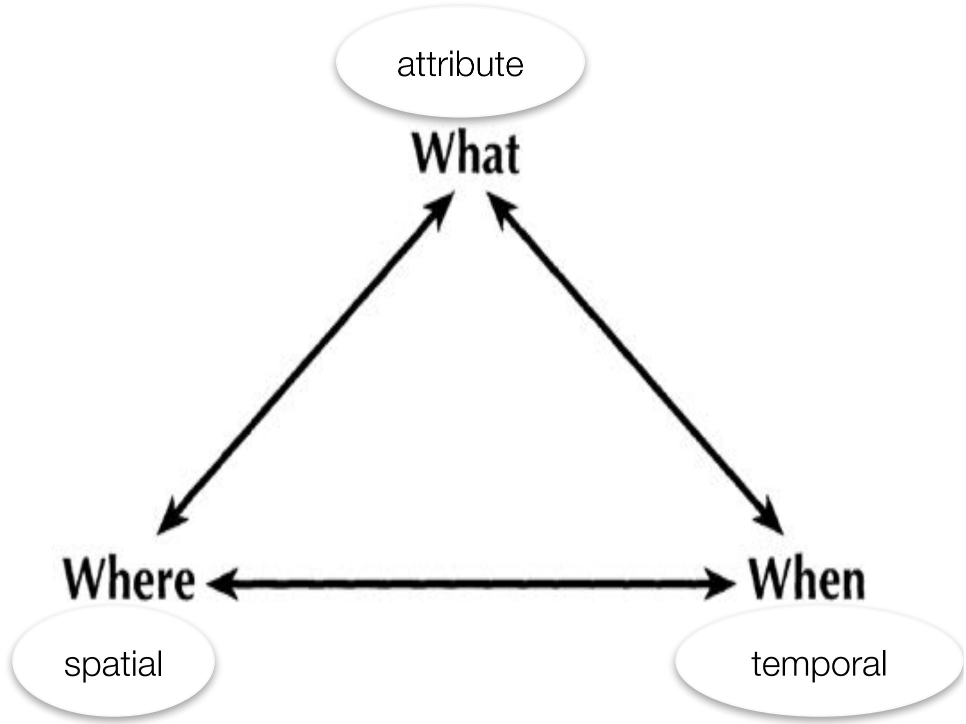
SELECT *
FROM trips
WHERE pickup_time in (5/1/11,5/7/11)
AND dropoff_loc in "Times Square"
AND pickup_loc in "Gramercy"

Data selection and result exploration are unified

Visual Query Model

Expressiveness:

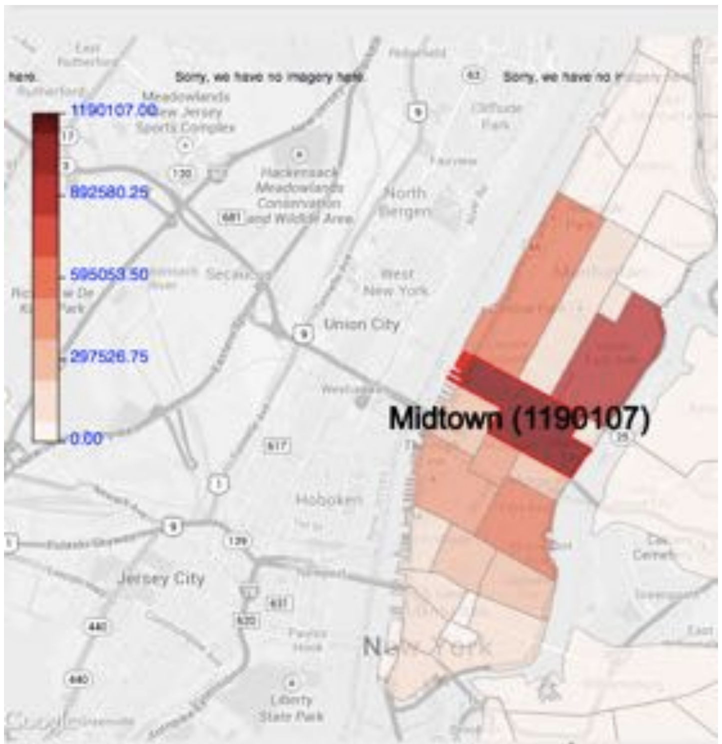
- when + where ➔ what: *"What is the average trip time from Midtown to the airports during weekdays?''*
- when + what ➔ where: *"Where are the hot spots in Manhattan in weekends?"*
- where + what ➔ when: *"When were activities restored in Lower Manhattan after the Sandy hurricane?"*



Peuquet's Triad

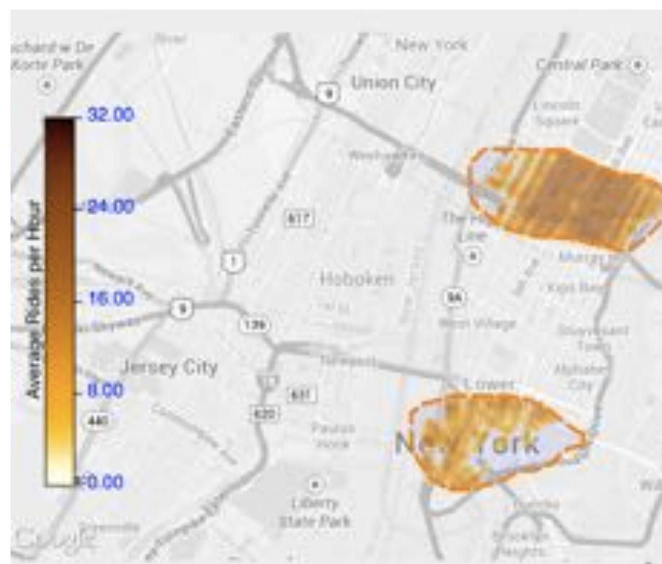
Model is also able to express other types of queries, including *when* → *what* + *where*, *where* → *when* + *what*, and *what* → *where* + *when*

Selecting Regions – Spatial Constraints



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

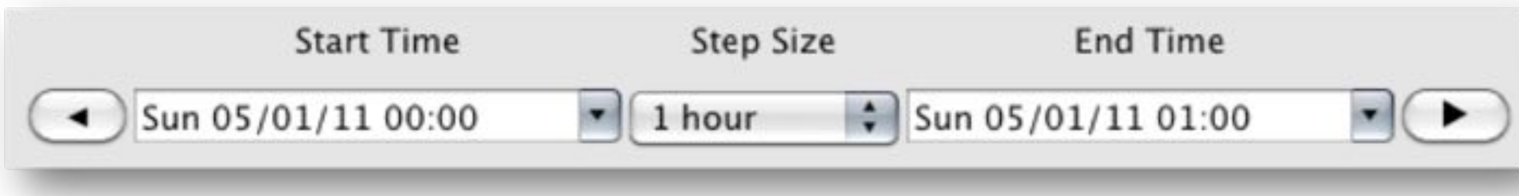
Free selection

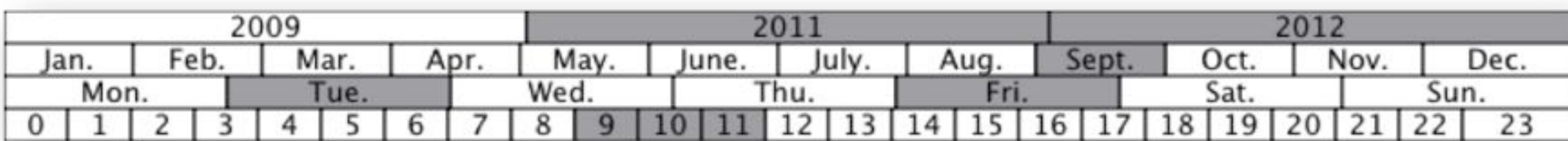


Group regions

Selecting Time – Temporal Constraints

Time interval

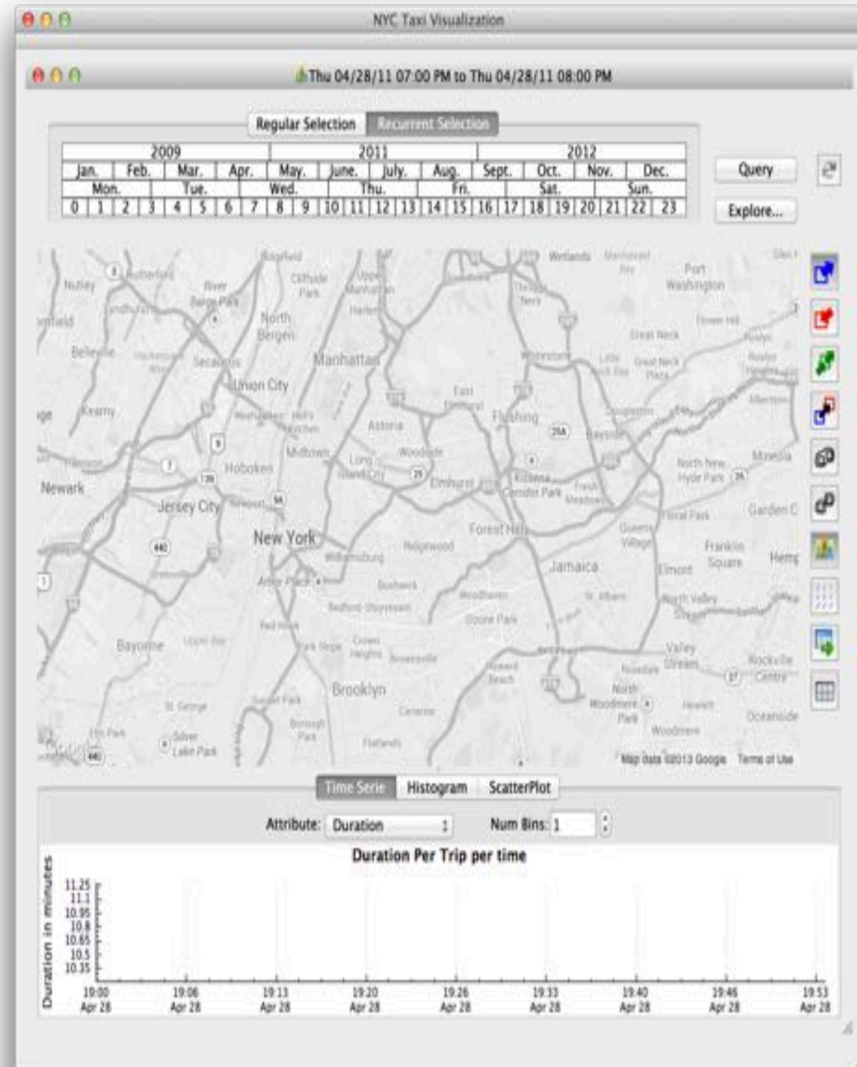




Recurrent time patterns

When + Where → What

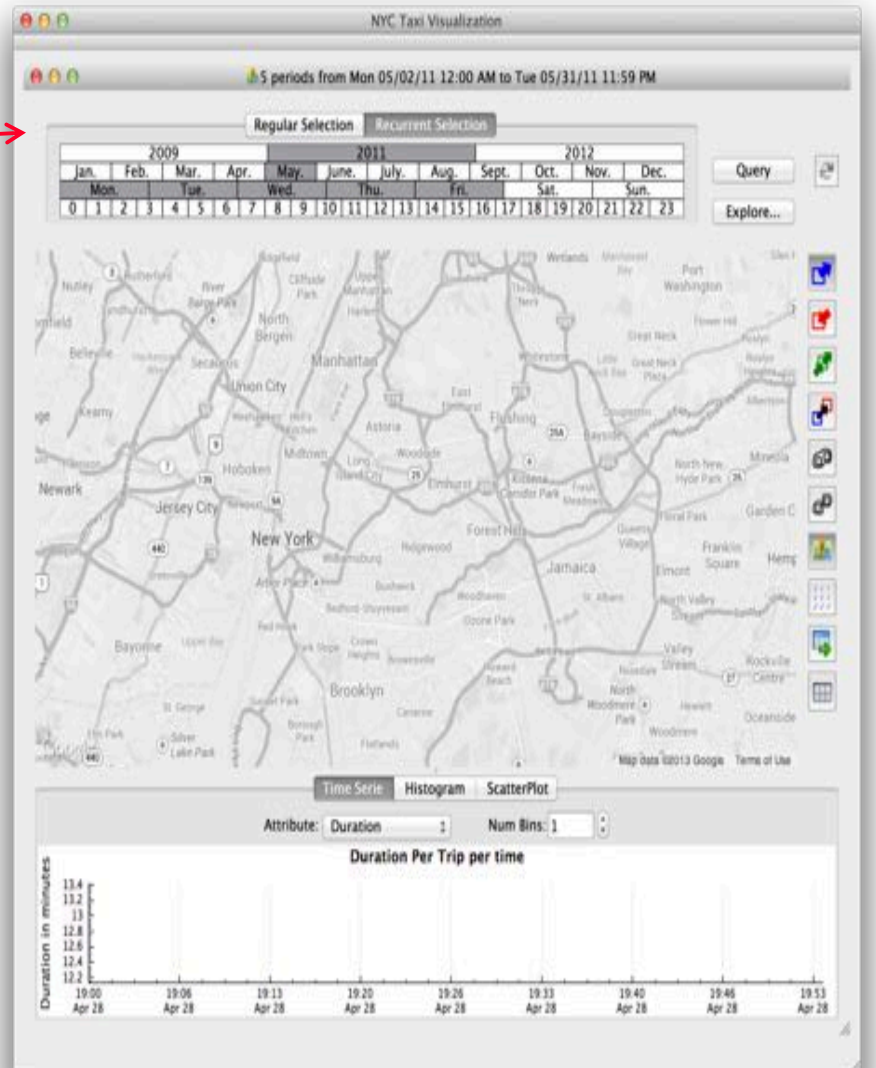
"What is the average trip time from Midtown to the airports during weekdays?"



When + Where → What

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

When?



NYU TANDON SCHOOL OF ENGINEERING

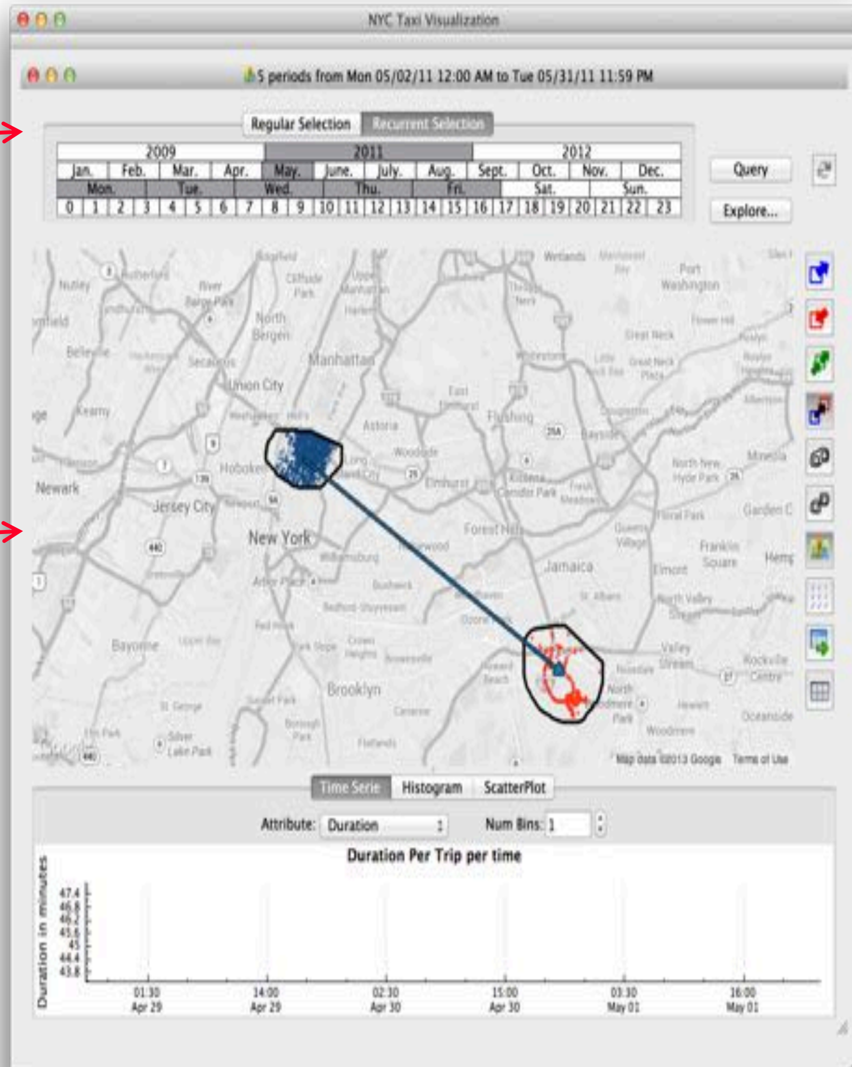
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

When + Where → What

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

When?

Where?



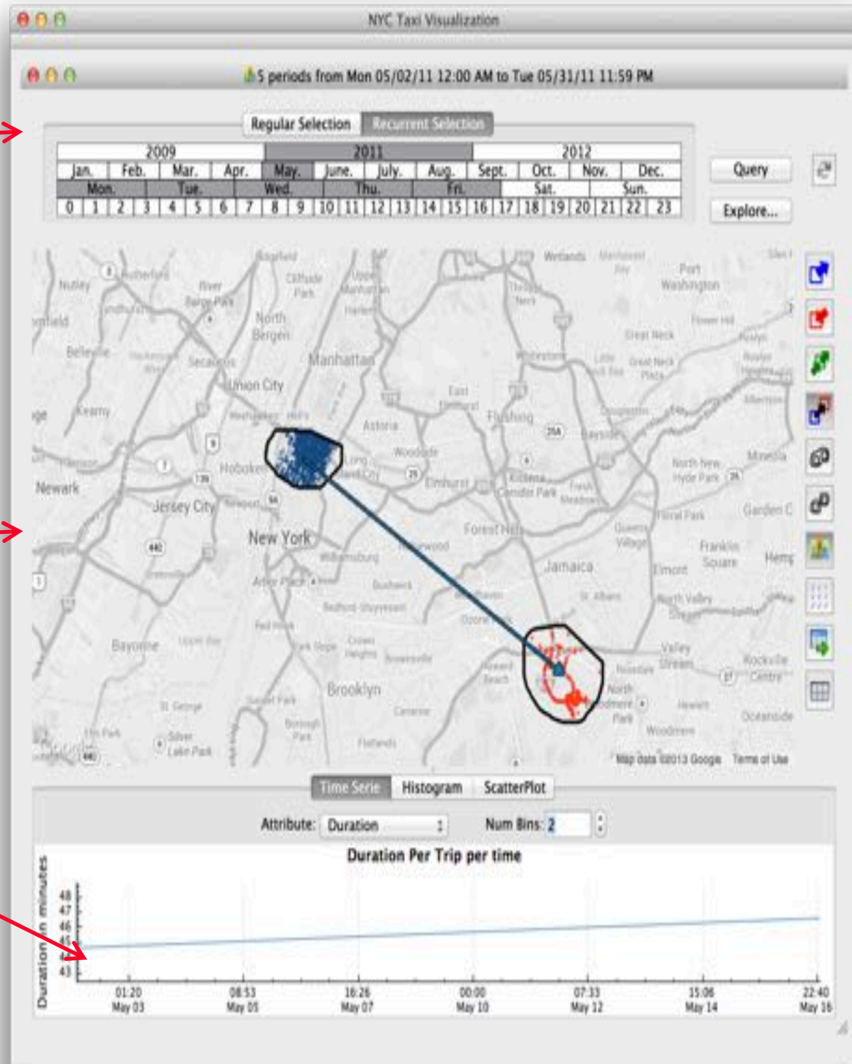
When + Where → What

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

When?

Where?

What



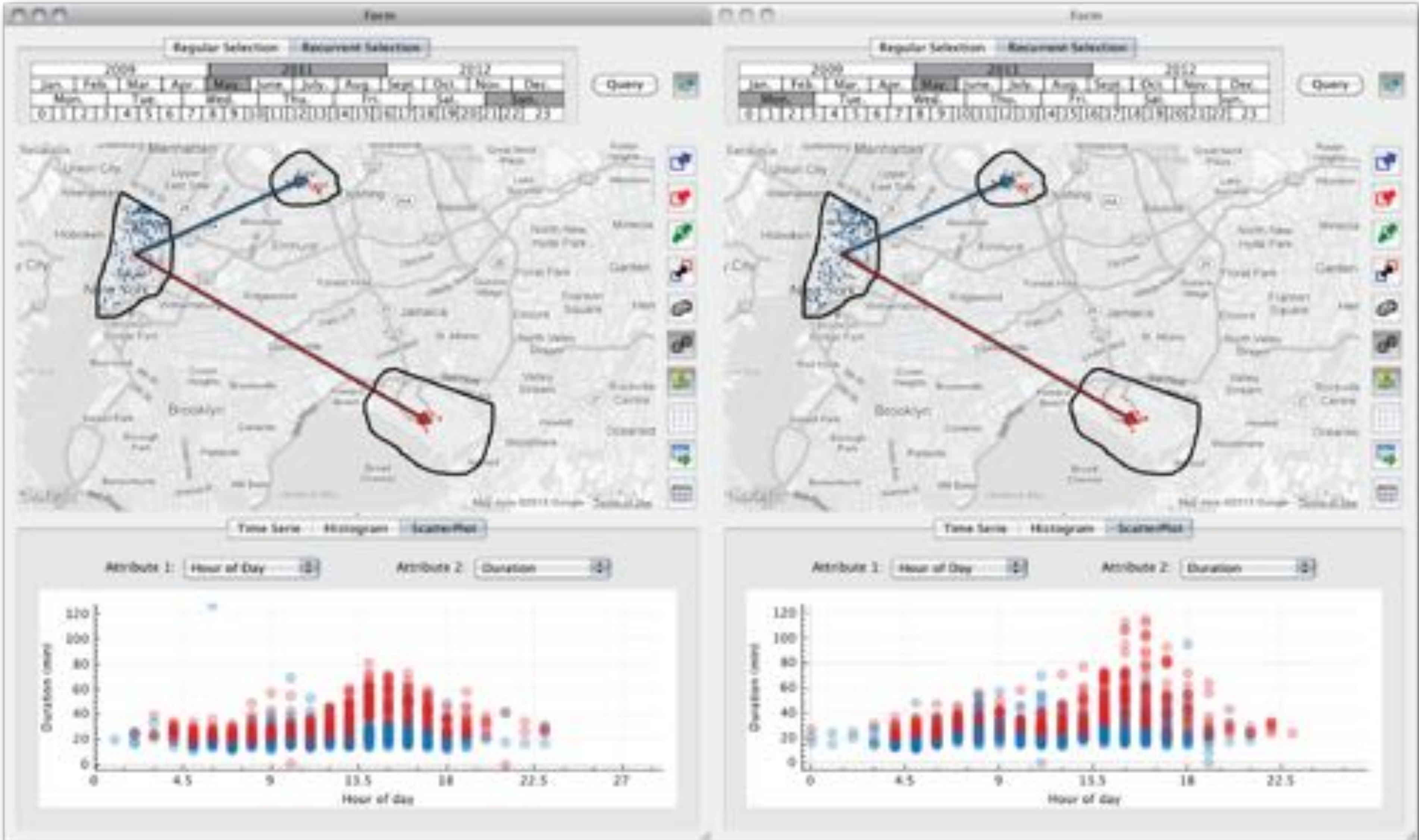
Composing Queries

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.

TaxiVis: Studying Mobility



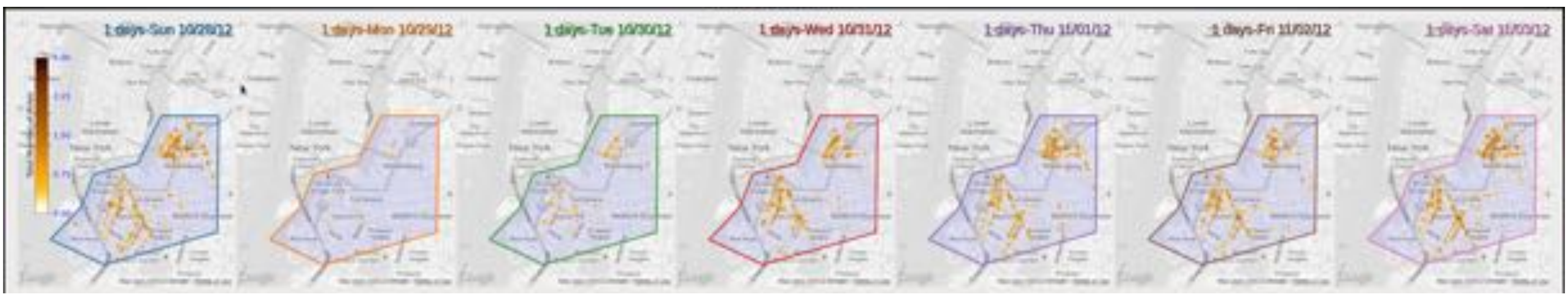
[Ferreira et al., IEEE TVCG 2013]

TaxiVis: Comparing Neighborhoods

dropoffs

pickups

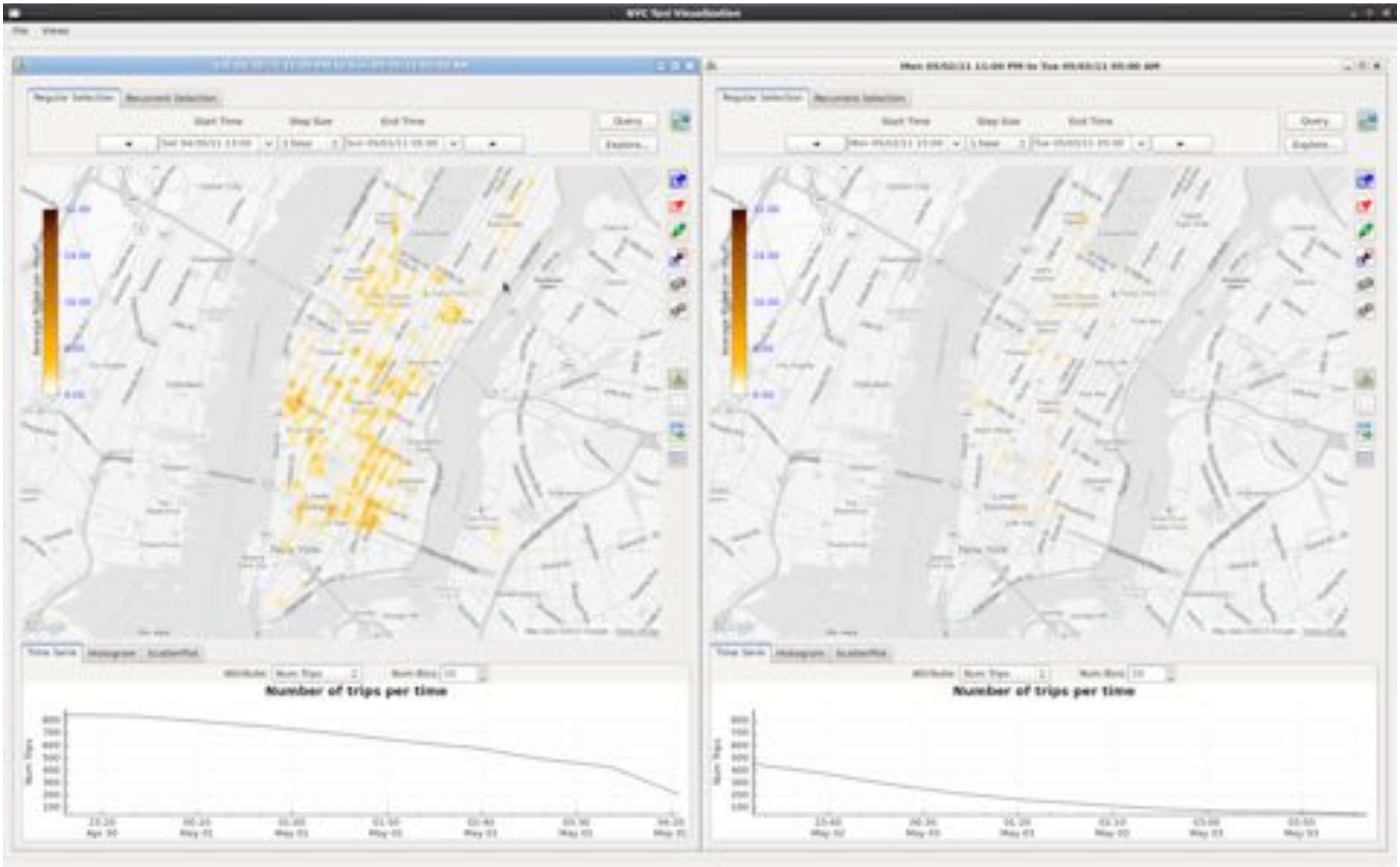
Exploring the Effect of Major Events: Sandy



NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Night Life in NYC: Saturday vs. Monday



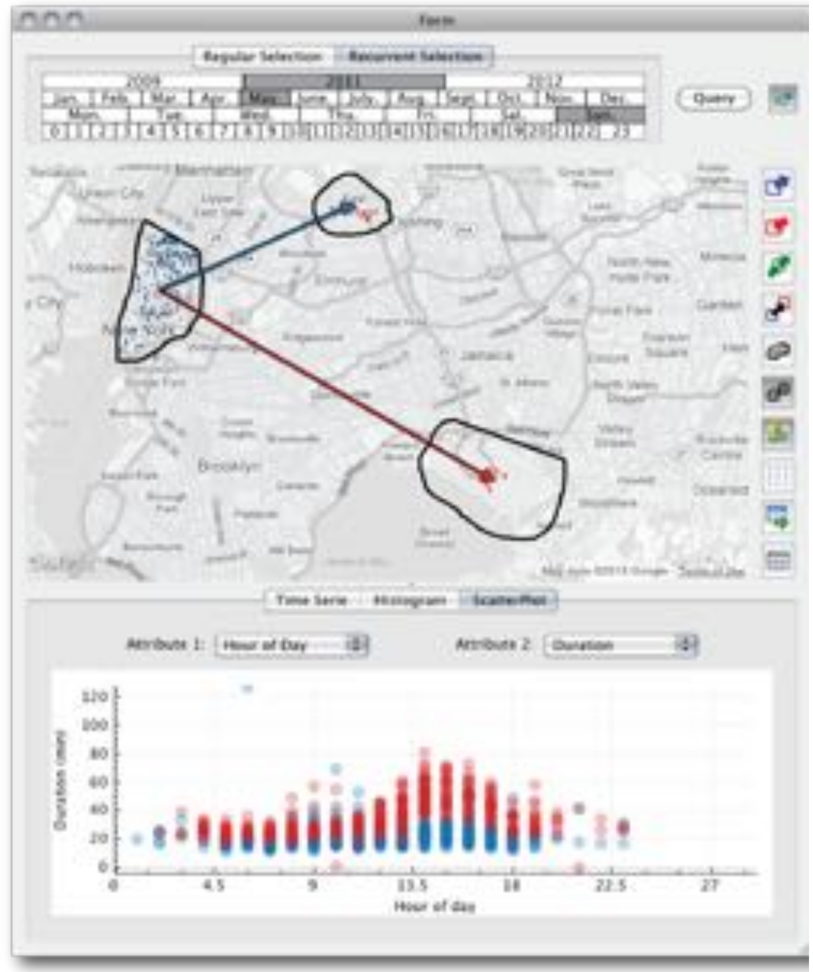
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Challenge: *Interactive* Query Evaluation

- Typical query:

Find all trips that occurred between lower Manhattan and the two airports, JFK and LGA, during all Sundays in May 2011



Query time (sec)	**PostgreSQL**	**ComDB**
	503.9	20.6

"increased latency reduces the rate at which users make observations, draw generalizations and generate hypotheses"

[Liu and Heer, IEEE TVCG 2014]

Challenge: *Interactive* Query Evaluation

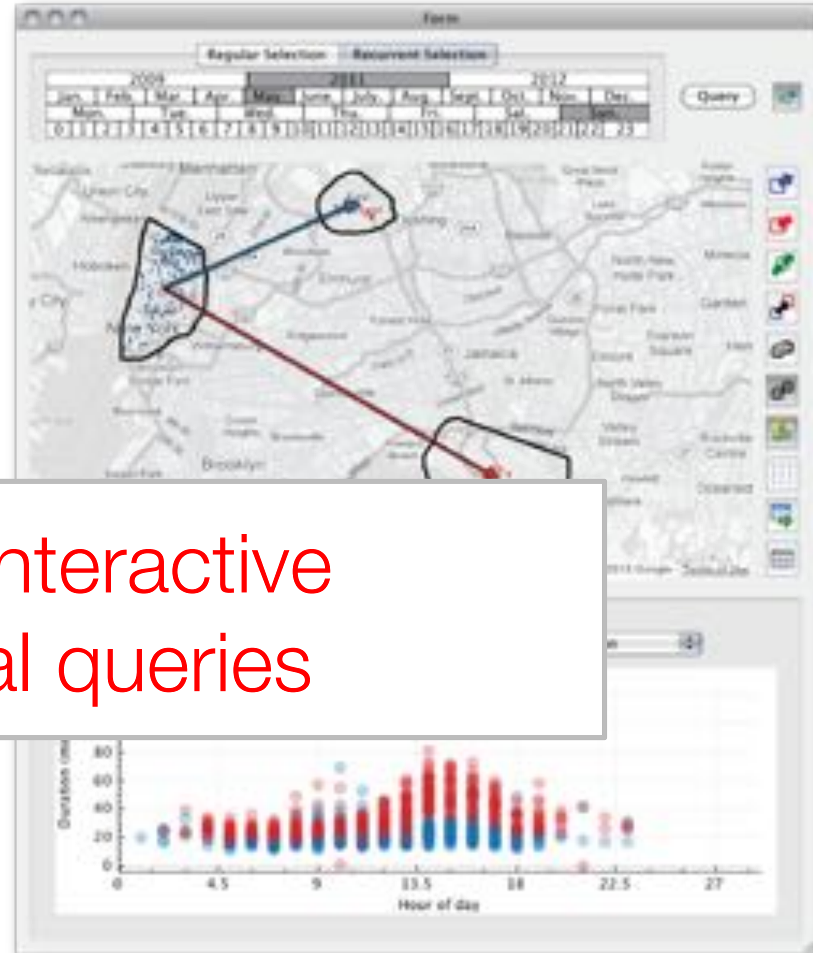
- Typical query:

Find all trips that occurred between

lower Manhattan and the two airports,

JFK and LGA,

during all Sundays in May 2011



Query time (sec)	503.9	20.6

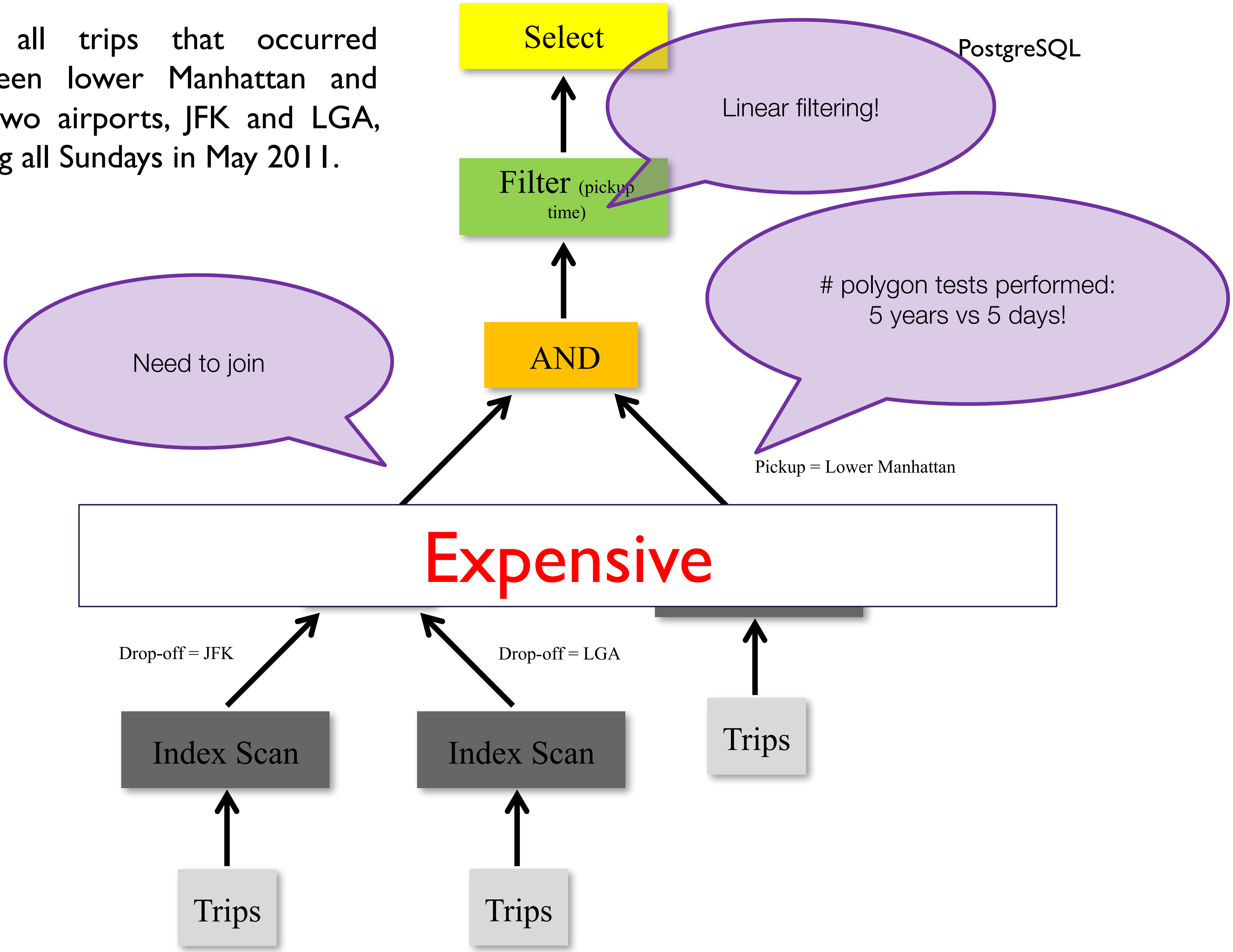
Goal: Support interactive spatio-temporal queries

"increased latency reduces the rate at which users make observations, draw generalizations and generate hypotheses"

[Liu and Heer, TVCG 2014]

Find all trips that occurred between lower Manhattan and the two airports, JFK and LGA, during all Sundays in May 2011.

PostgreSQL



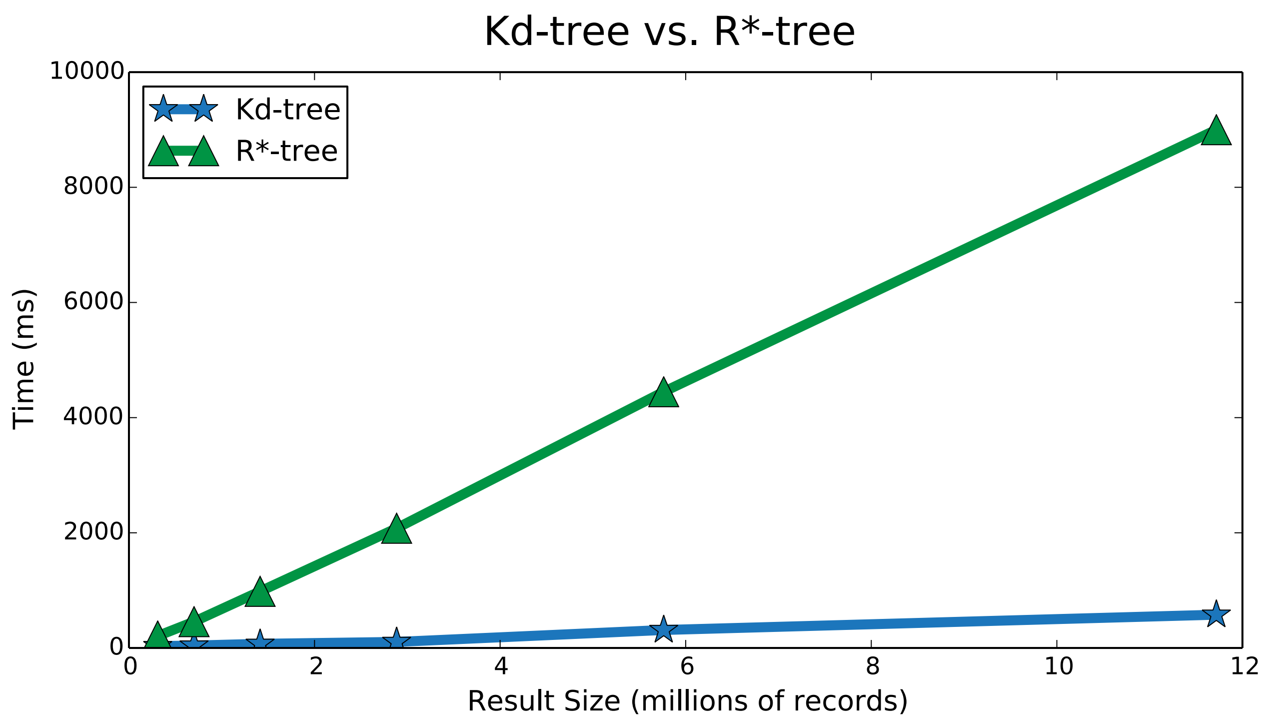
Design Goals

- Avoid joins
  - Filter simultaneously over multiple attributes
  - Need a multi-dimensional data structure
- Speed-up polygon containment tests
  - Each test is independent of another
  - GPUs are optimized for such operations
  - Make use of GPUs
- Index structure should be GPU-compatible
  - Minimize data transfer
  - Maximize occupancy

Choice of Data Structure

R*-Tree	KD-Tree
Balanced	Balanced
Allows update	Update does not maintain balance
Sibling nodes intersect	Sibling nodes do not intersect

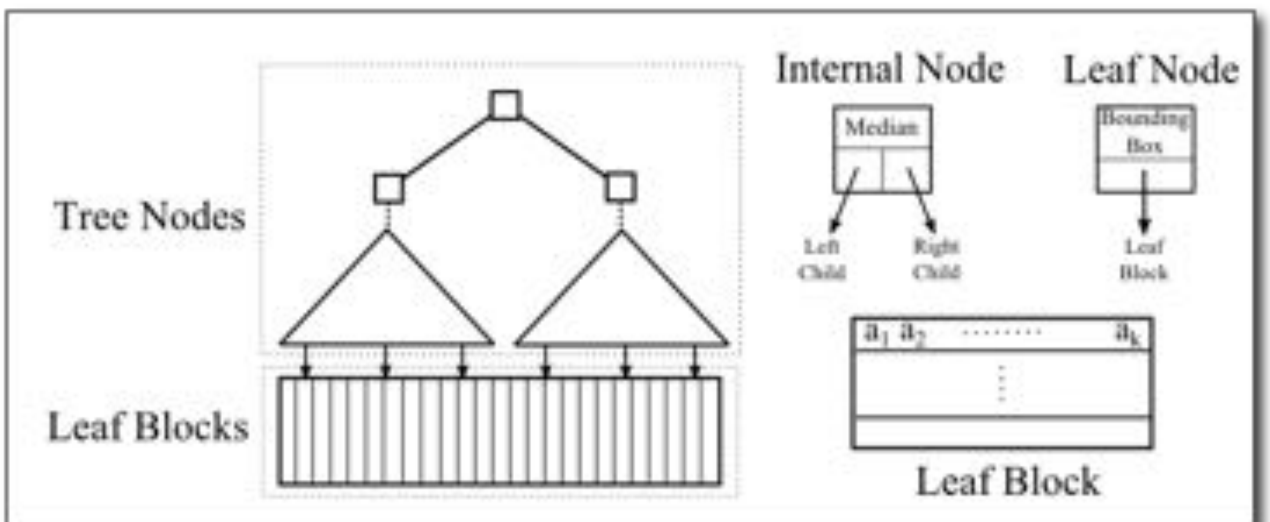
Choice of Data Structure



Supporting Interactive Queries

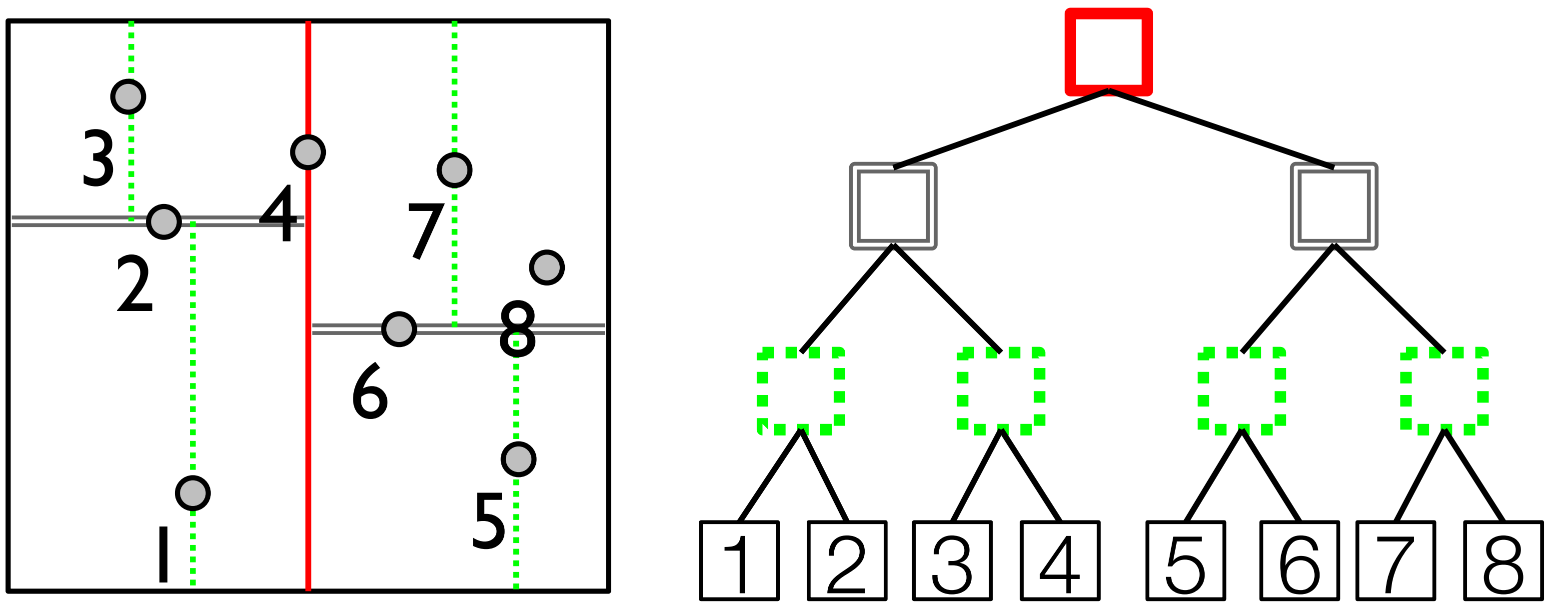
Solution: Spatio-temporal index based on out-of-core kd-tree using GPUs (STIG)

- Can index and simultaneously filter multiple attributes: avoid joins and reduce the number of point-in-polygon (PIP) tests
- 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



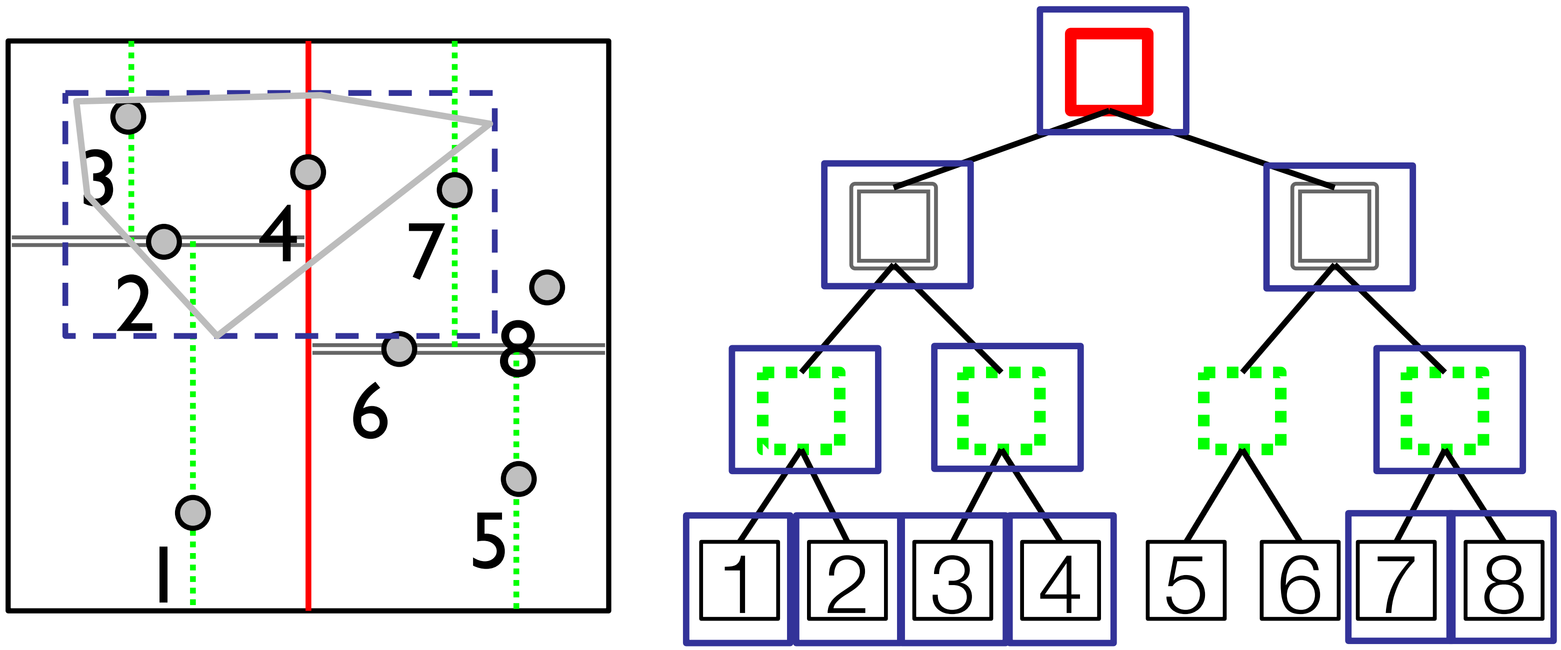
[Doraiswamy et al., ICDE 2016]

KD-Tree



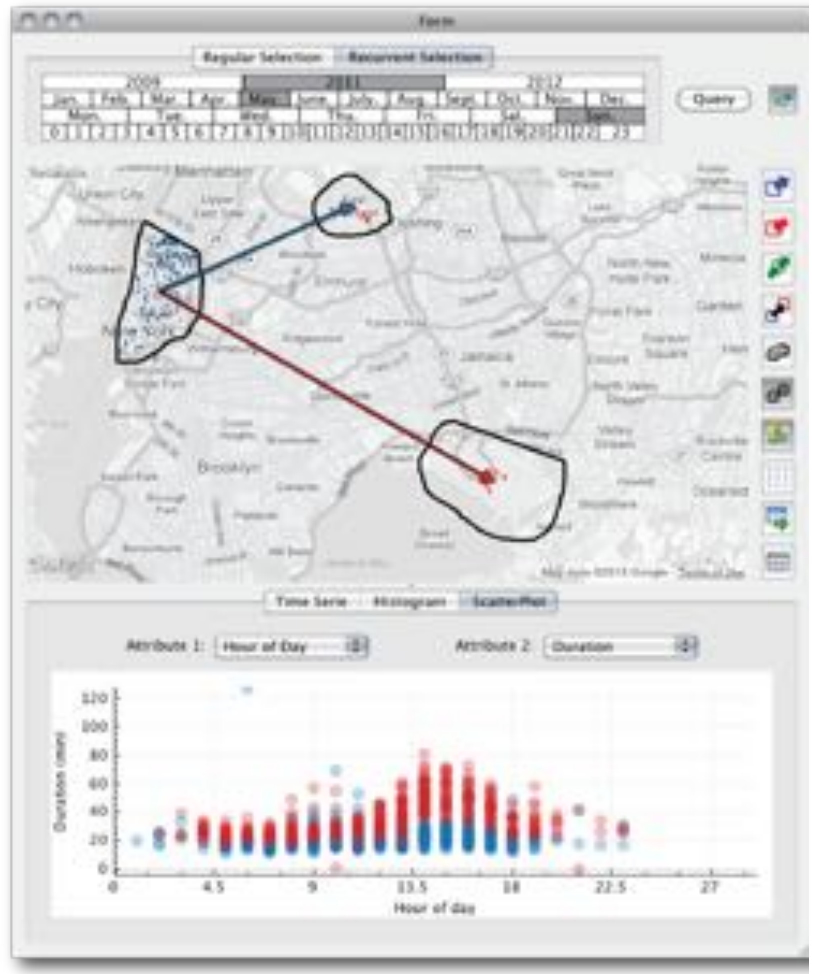
KD-Tree

- Polygon containment query
  - Search based on Bounding Box
  - Test with query polygon

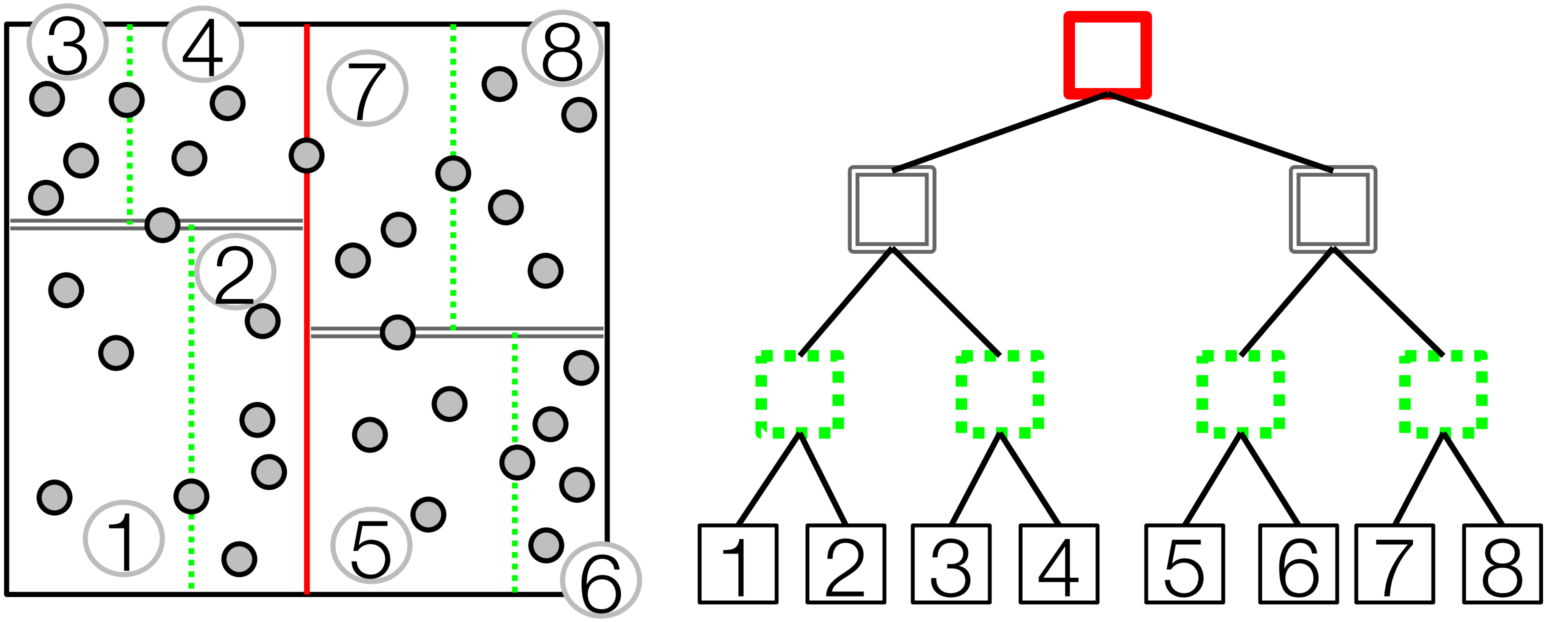


PIP Tests are Expensive

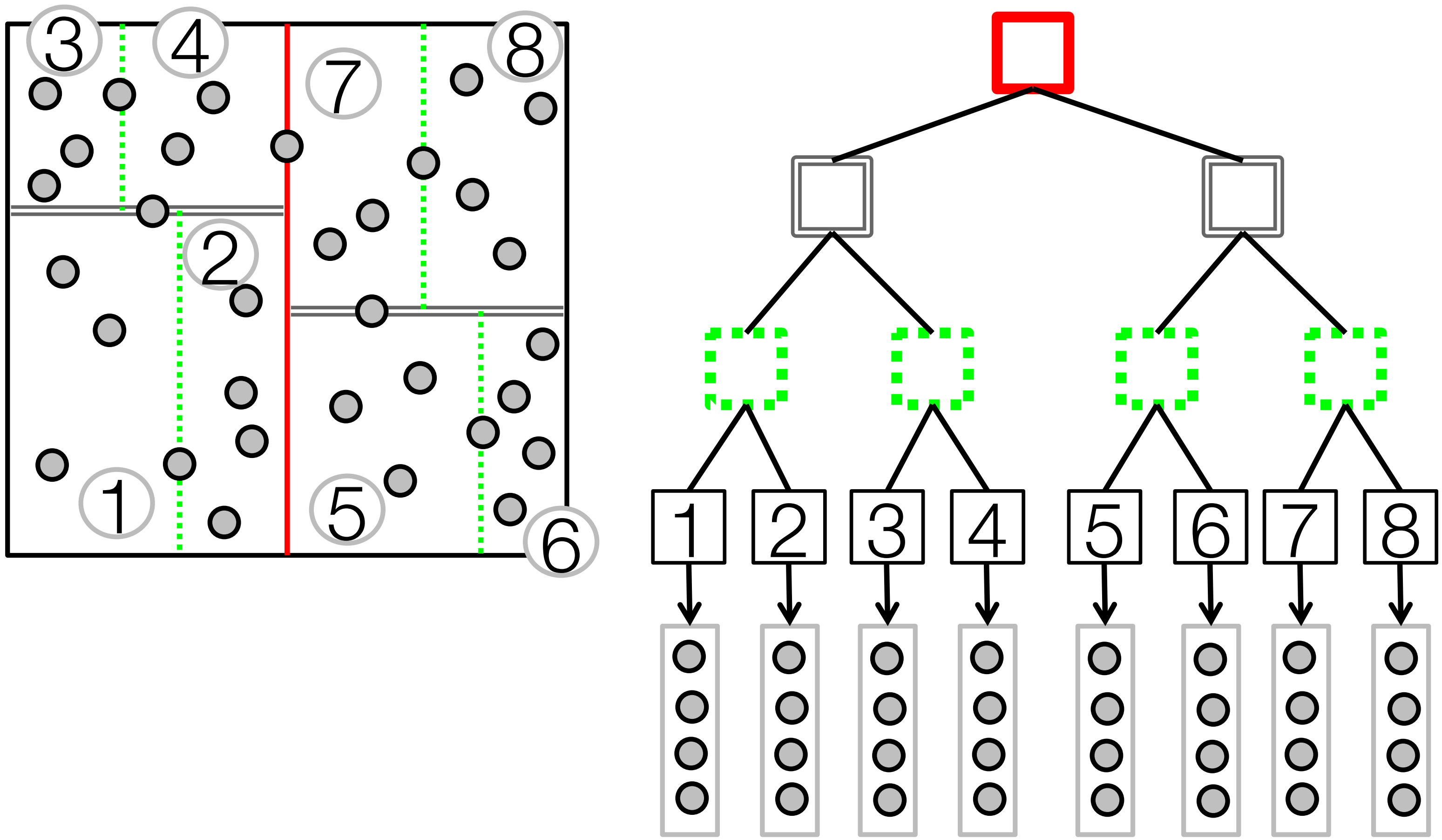
6.5 million such tests have to be performed even though the query returns only around 13,000 records



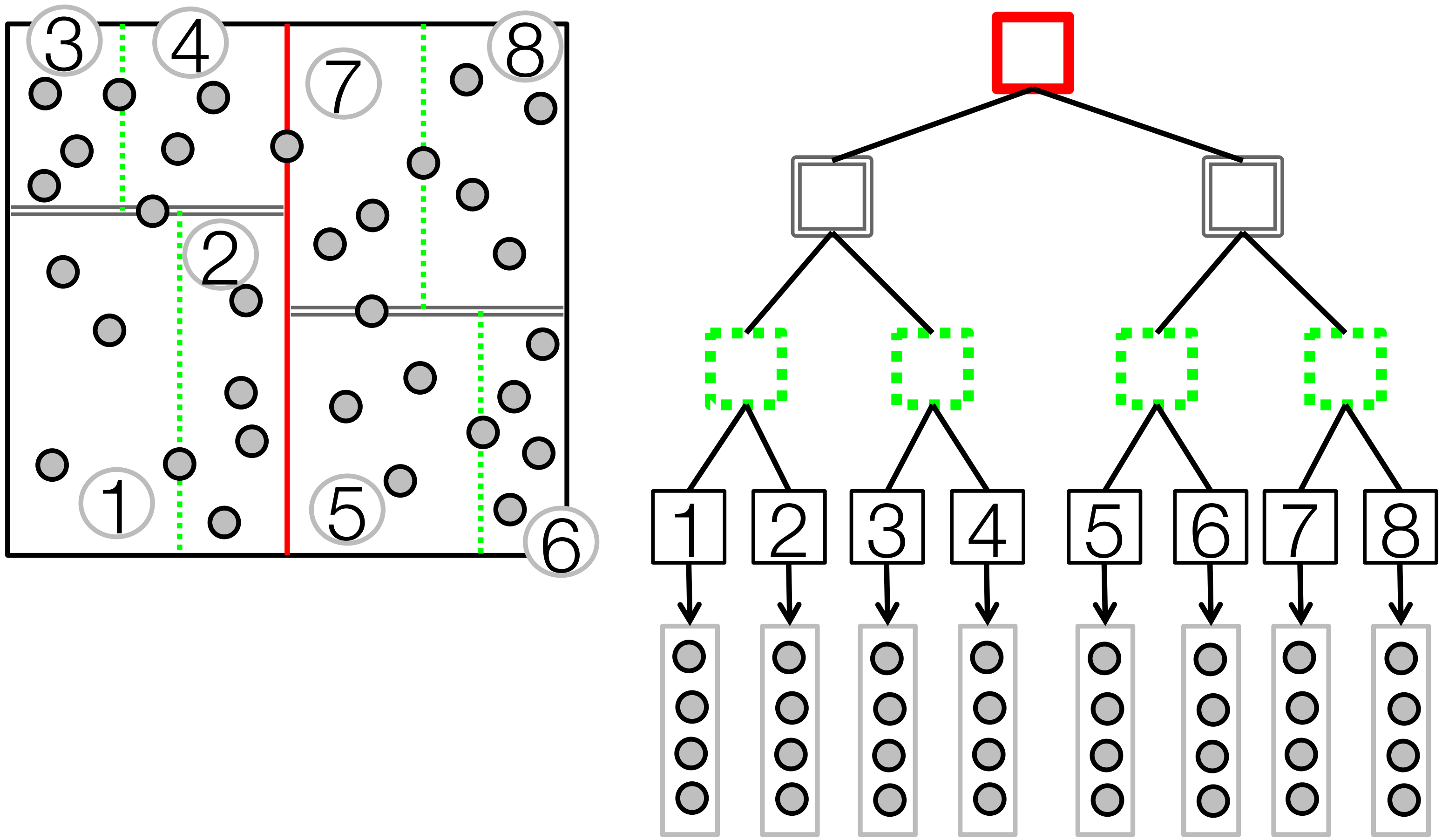
The STG Tree



Stg Tree

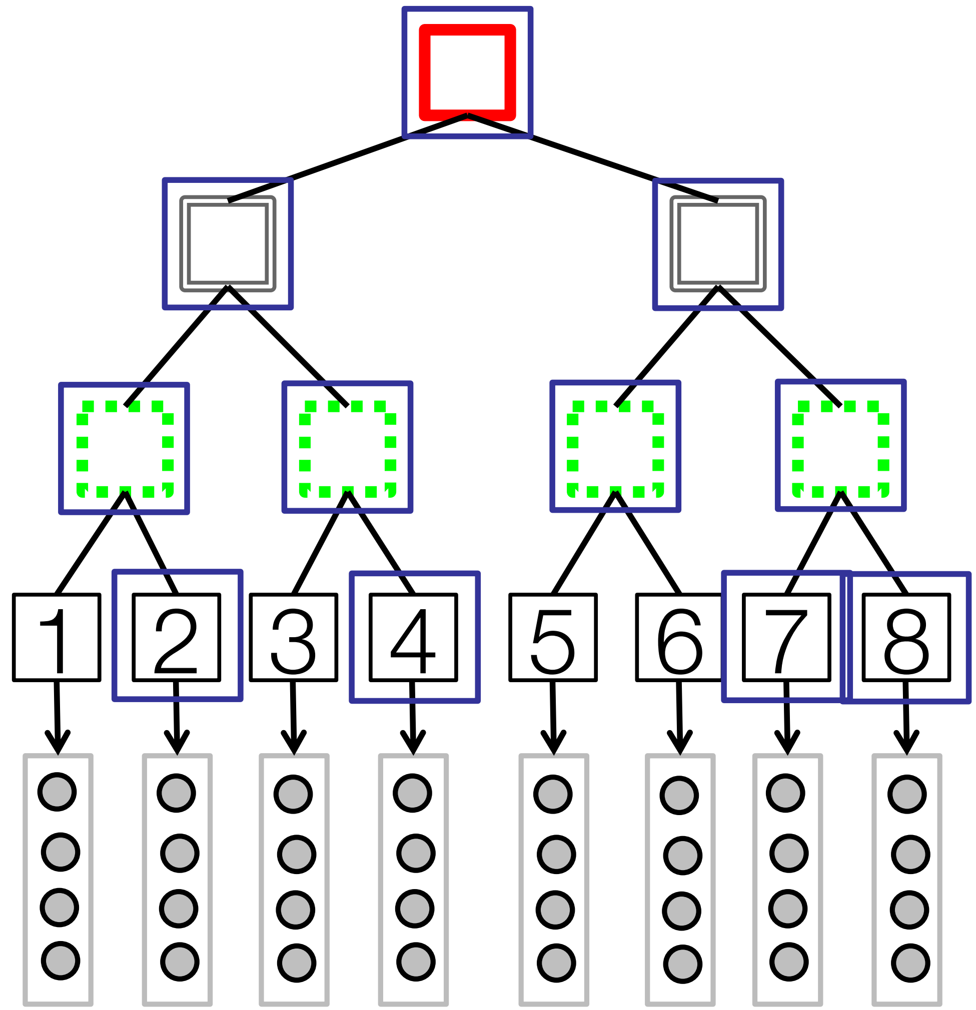


Stg Tree



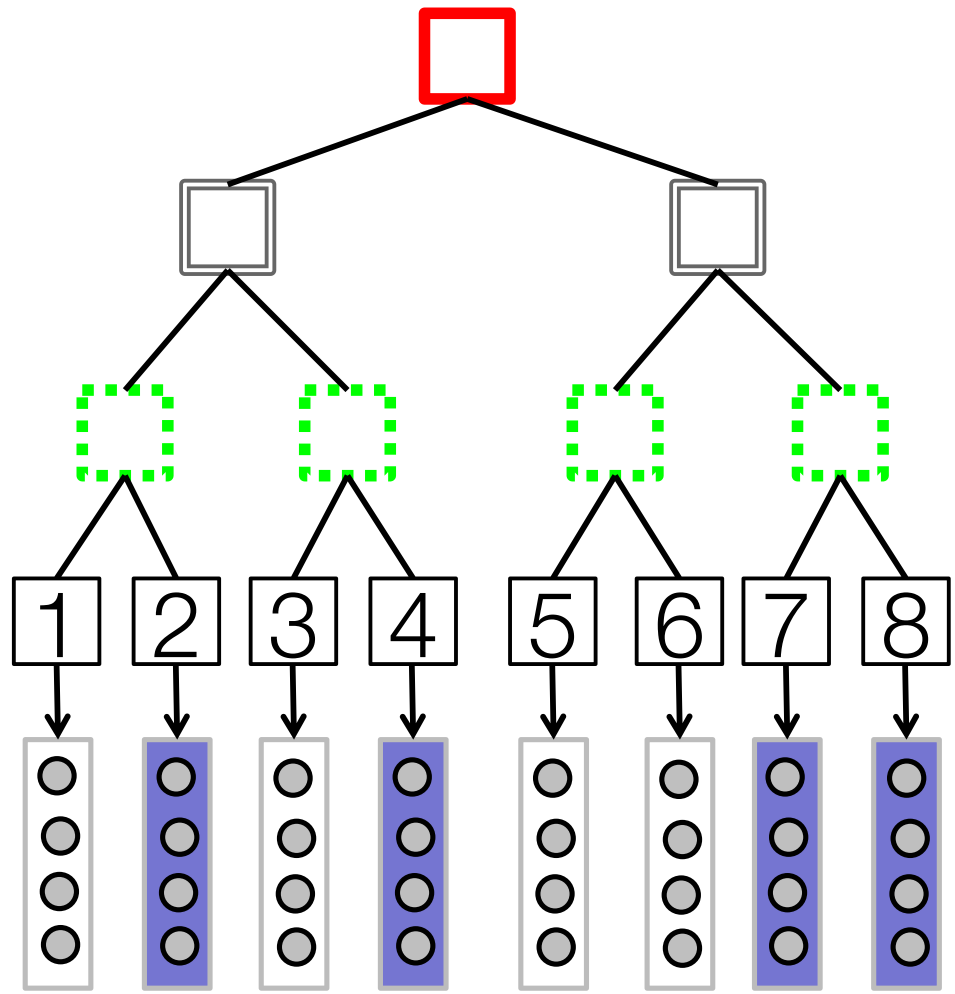
STIG Query

- Two steps
  - Search tree nodes



STIG Query

- Two steps
  - Search tree nodes – in memory
  - Search leaf blocks – in GPU



Supporting Interactive Queries

Solution: Spatio-temporal index based on out-of-core kd-tree using GPUs

- Can index and simultaneously filter multiple attributes: avoid joins and reduce the number of point-in-polygon (PIP) tests
- 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
- Create *big* blocks – tree is small and fits in memory
- Use GPU to search the blocks in parallel – speeds up PIP tests

- Source code available at

https://github.com/harishd10/mongodb

[Doraiswamy et al., ICDE 2016]

Performance Evaluation

Setup:

- 12-code Xeon processor @2.4 GHz
- 8 TB storage
- 256 GB memory
- 3 x NVIDIA GeForce TITAN
  - 6 GB memory

Performance: Taxi Data

Find all trips between Lower Manhattan and the two airports, JFK and LGA, during all Sundays in May 2011.

Query	MongoDB	PostgreSQL		ComDB	
	Time	Time	Speed up	Time	Speed up
1		503.9		20.6	
2		501.9		23.3	
3		437.8		21.6	
4		437.1		32.6	

Time in Seconds

868 million trips; ~13k results/query

Performance: Taxi Data

Find all trips between Lower Manhattan and the two airports, JFK and LGA, during all Sundays in May 2011.

Query	MongoDB	PostgreSQL		ComDB	
	Time	Time	Speed up	Time	Speed up
1	0.075	503.9	6718	20.6	274
2	0.080	501.9	6273	23.3	291
3	0.067	437.8	6534	21.6	322
4	0.070	437.1	6244	32.6	465

Time in Seconds

868 million trips; ~13k results/query

[Doraiswamy et al., ICDE 2016]

Performance: Twitter Data

Query	MongoDB	PostgreSQL		ComDB	
	Time	Time	Speed up	Time	Speed up
1	0.246	161.2	655	109.6	445
2	0.288	151.2	525	157.7	547
3	0.558	286.0	512	216.8	388

Time in Seconds

1.1 billion tweets; 130k-370k results/query

What Next: Urbane

https://www.youtube.com/watch?v=_B35vxCgDw4&feature=youtu.be

[Ferreira et al., IEEE VAST 2015]

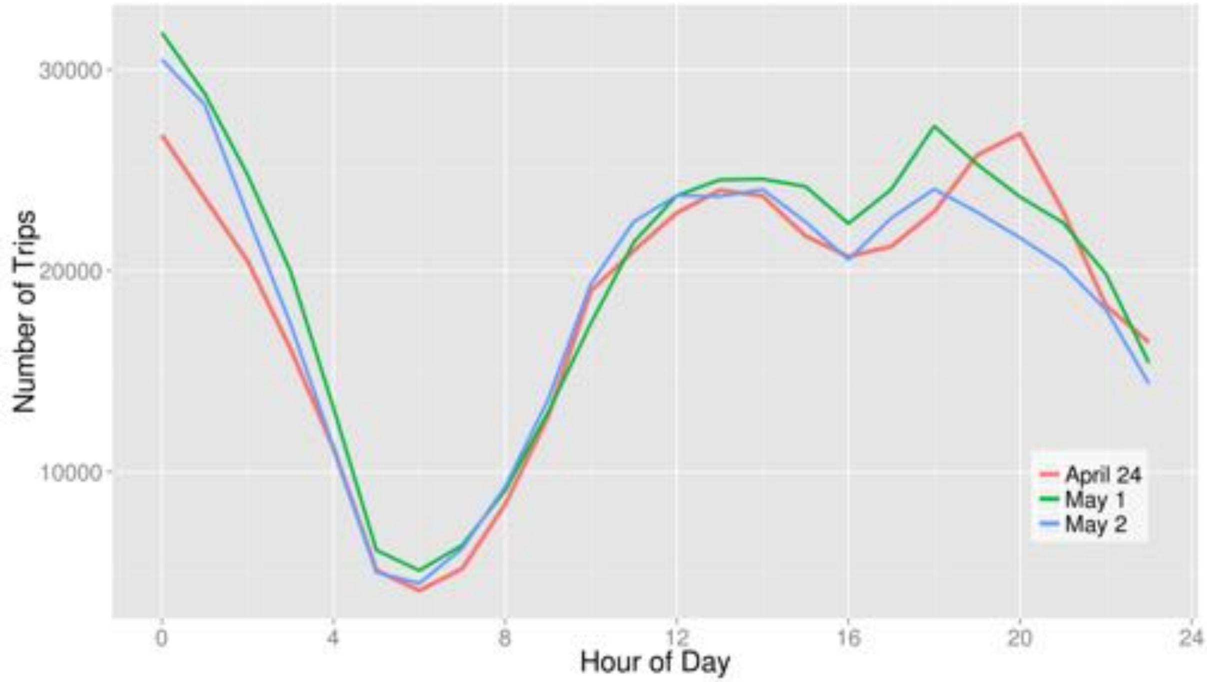
Finding Interesting Features

Taxi Data: Too Many Slices

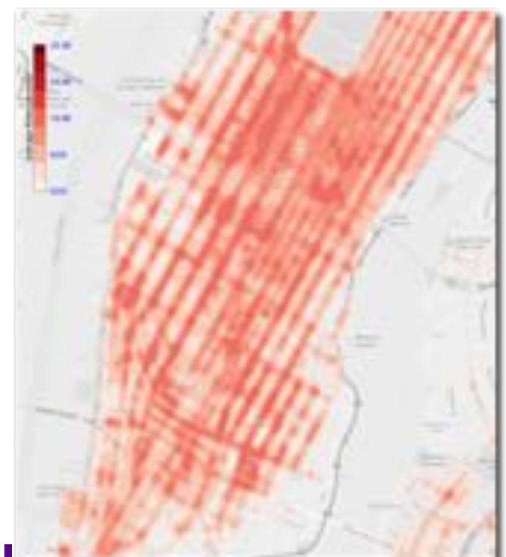
- 365*24 1-hour slices in one year
- Which slices are interesting?

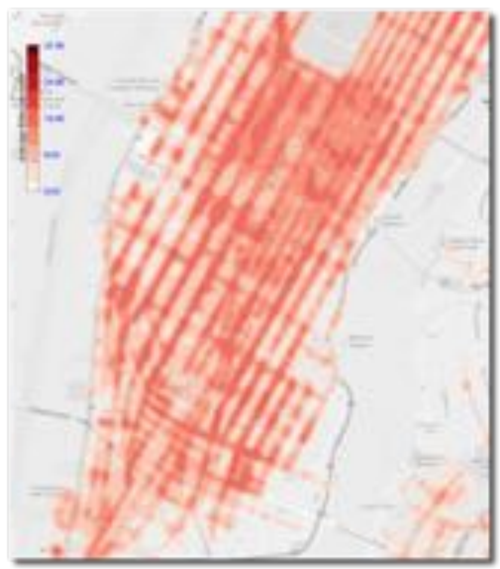
Reducing the Number of Slices

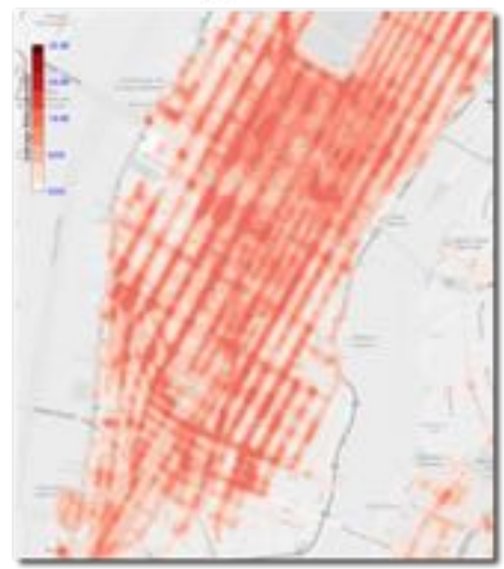
Aggregate over space



Aggregate over time

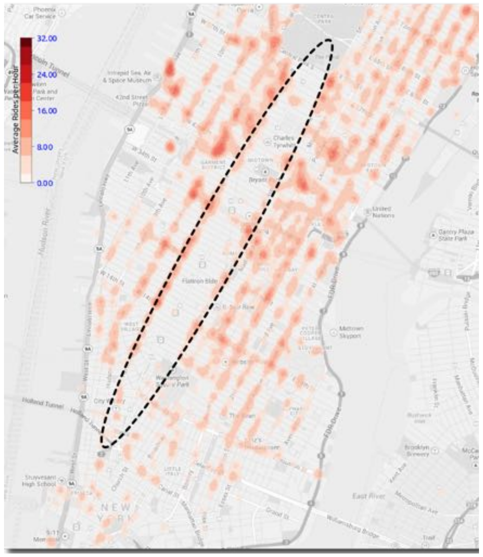

April 24


May 1

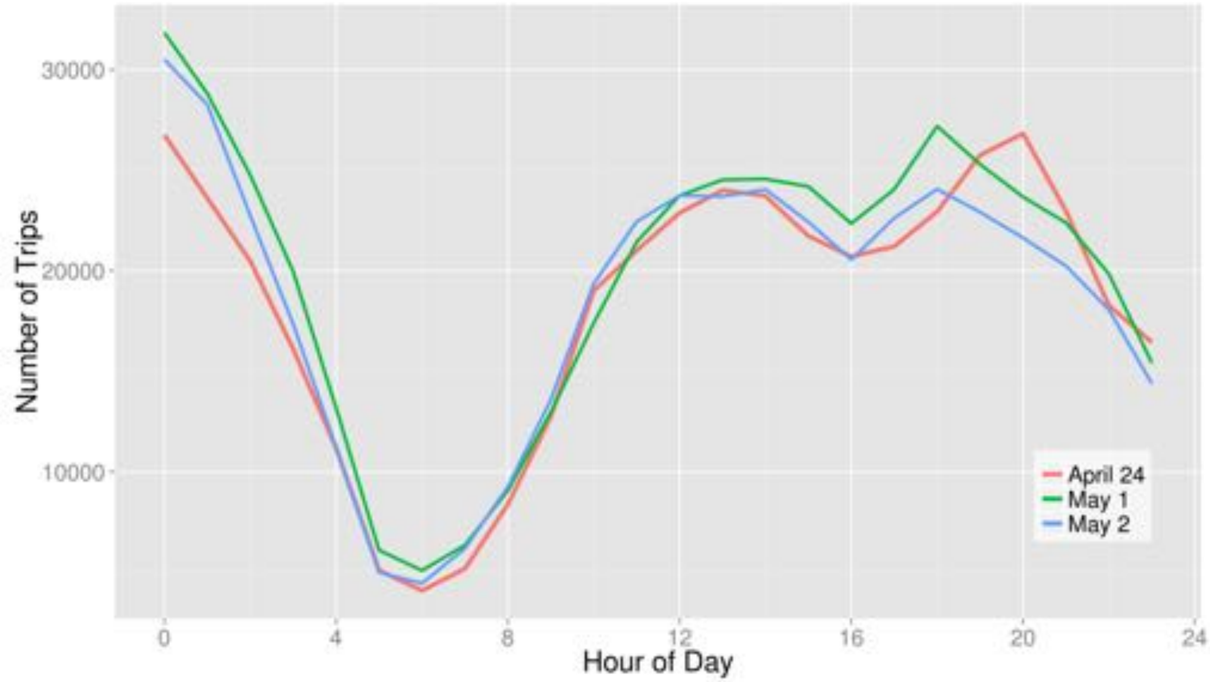

May 8

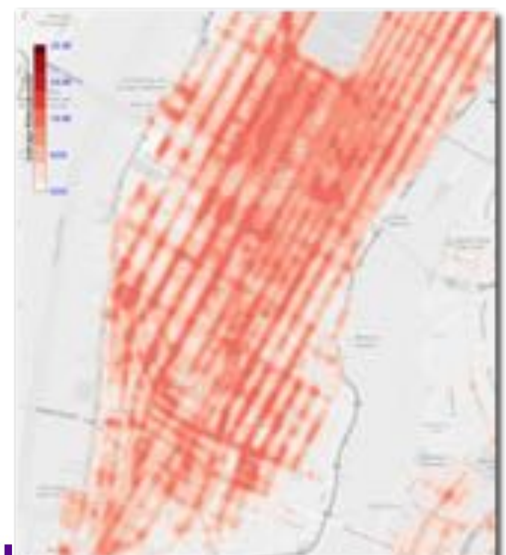
VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Miss Interesting Slices

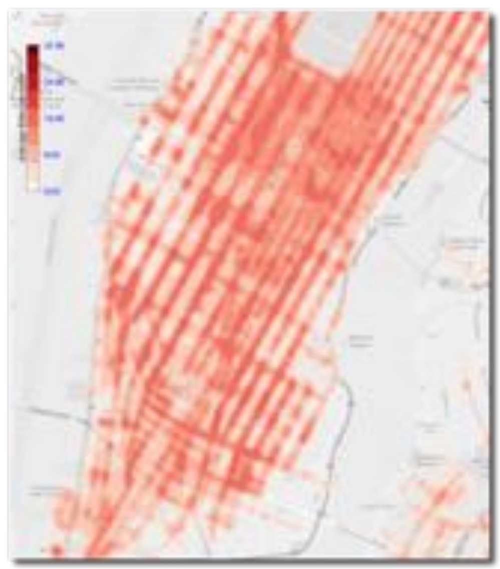


May 1 (8-9am)

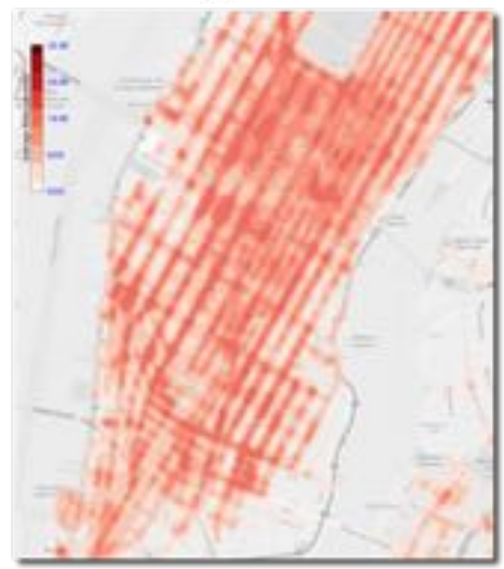




April 24



May 1



May 8

Finding Interesting Slices

Goal: guide users towards *interesting* data slices

- Desiderata: automatically identify *events* with arbitrary spatial structure and at multiple temporal scales
- Our solution:
  - Use computational topology techniques to efficiently discover events
  - Simple visual interface to *explore* and *query* the events of interest

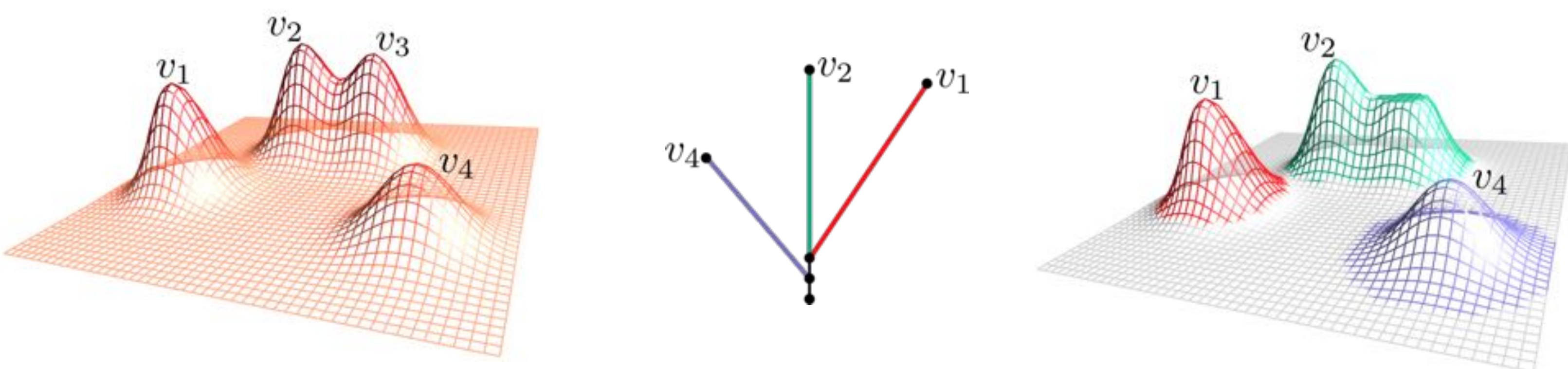
[Doraiswamy et al., IEEE TVCG 2014]

Identifying Potential Events

- Model data as a time-varying scalar function defined on a graph
  - $f : G \rightarrow R$
  - Taxi data: Graph = road network; Function = density of taxis
  - Subway data: Graph = track network; Function = delay of trains

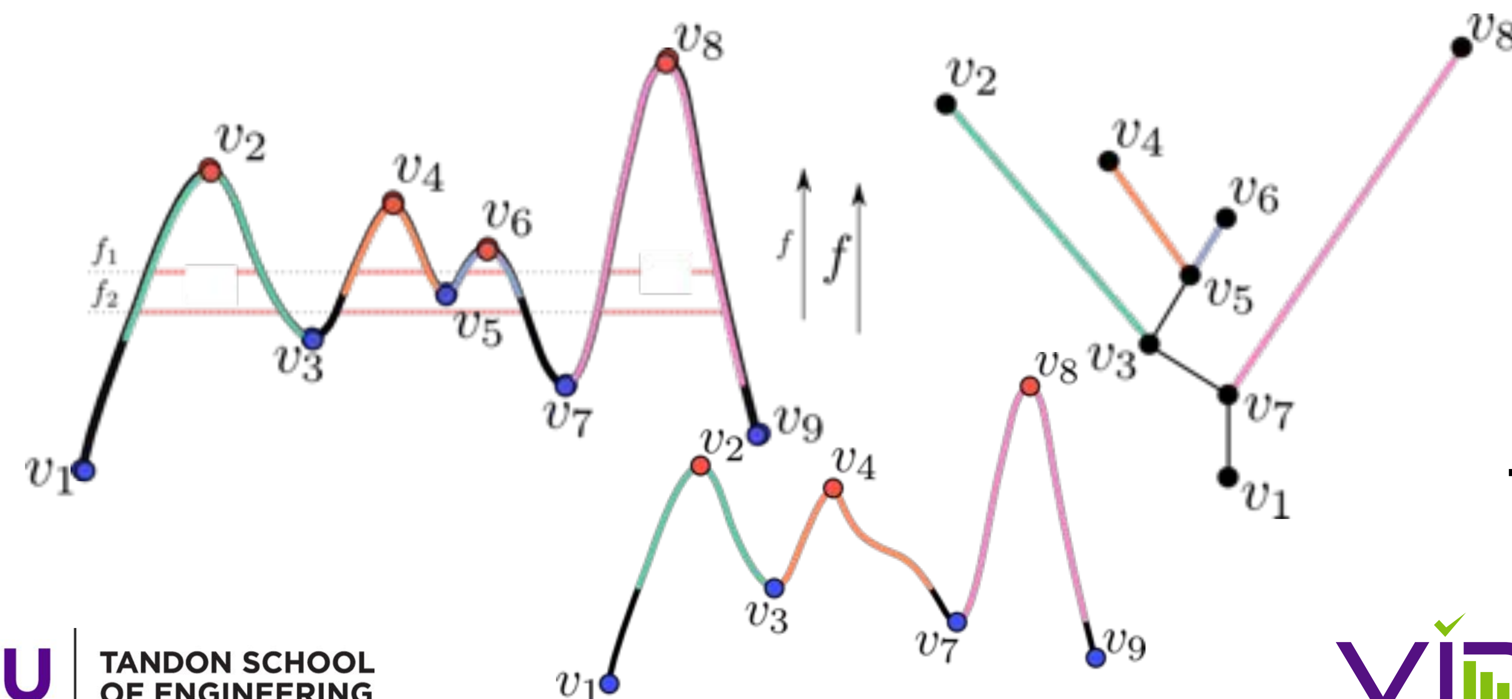
Identifying Potential Events

- Use Merge Trees to efficiently identify events in each time step
- Compute the regions corresponding to the set of *maxima* and *minima – the set of potential events*
  - Intuition: a region is interesting if its behavior differs from that of its neighborhood
  - Unimportant events can be simplified



Identifying Potential Events

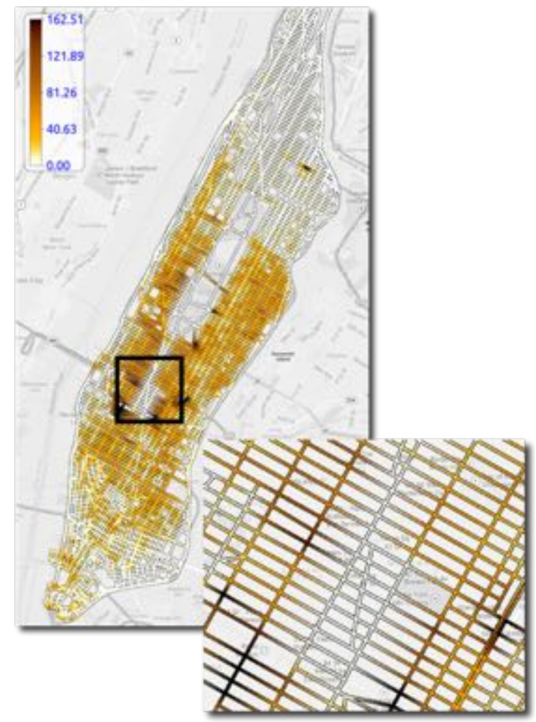
- Join (and Split tree) can be used to efficiently represent regions
  - Topological changes occur at critical points
  - Trees can be simplified to remove noise



Join Tree

Taxi Data: Potential Events

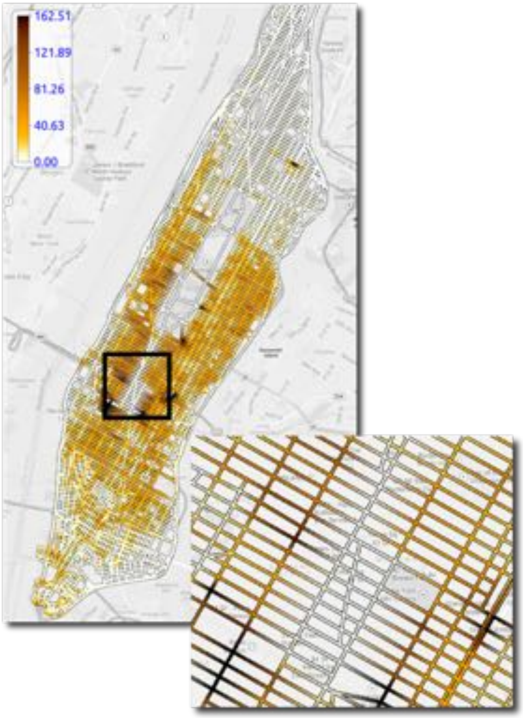
- Minima: lack of taxis
  - Regions where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade

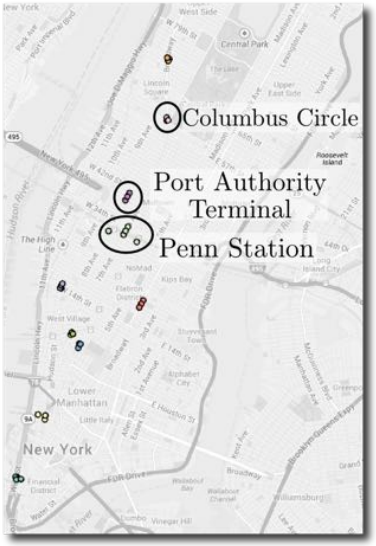


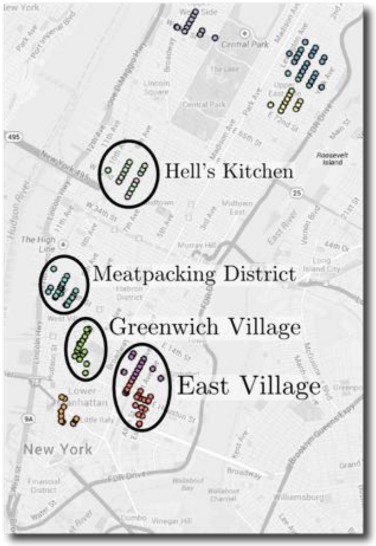
Scalar function corresponding to the time step 10 am-11 am on 24 November 2011

Taxi Data: Potential Events

- Minima: lack of taxis
  - Regions where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade
- Maxima: popular taxi locations
  - Regions where density is higher than local neighborhood
  - Could denote tourist locations, train stations





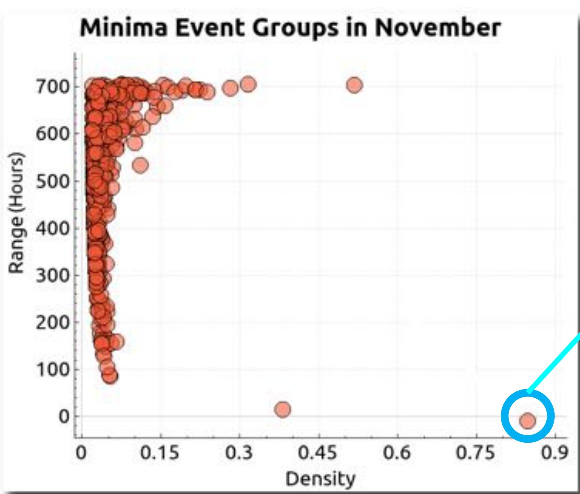


Grouping and Exploring Events

- Too many events!
- Group similar events and create an index
  - Geometric and topological similarity
- Visual interface to guide users
- Filter based on group size, event size,

event time, spatial region

short → long time span



small → large groups

Macy's parade

Querying Events

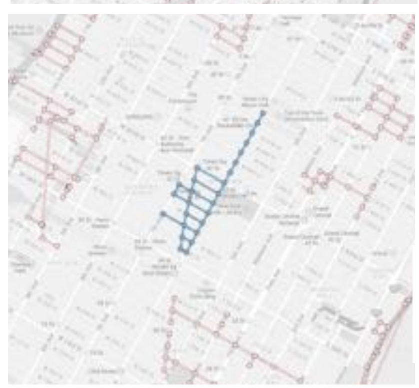
5 Borough Bike Tour 2011
(1 May 2011)

Query

Dominican Day Parade 2011
(14 August 2011)

5 Borough Bike Tour 2012
(6 May 2012)

Dominican Day Parade 2012
(12 August 2012)



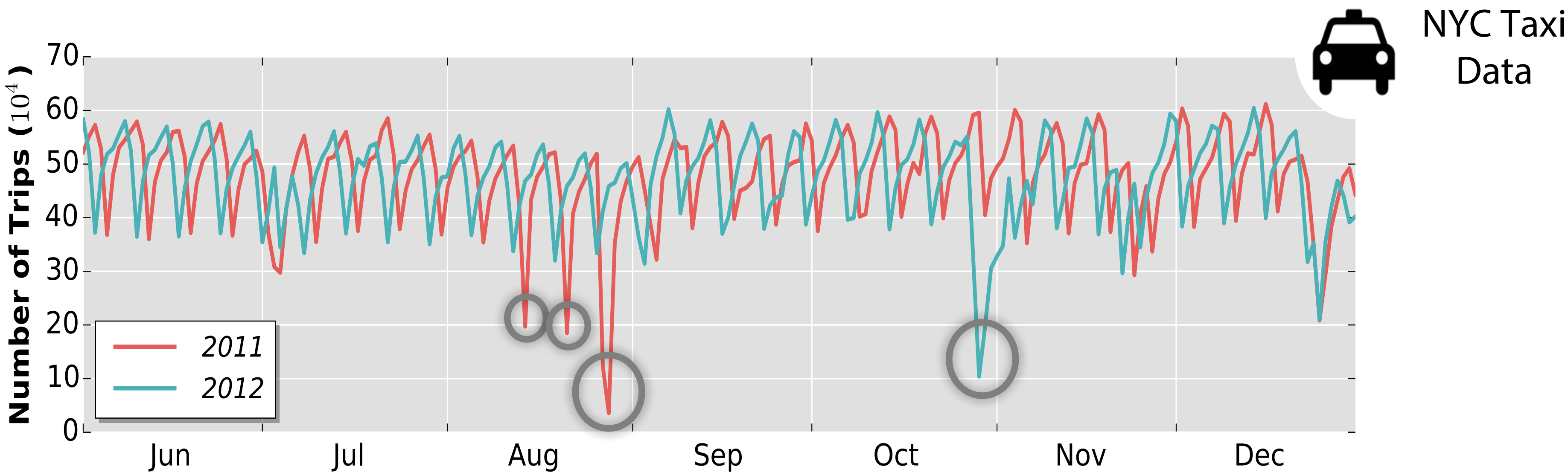
Gaza Solidarity Protest NYC
(18 November 2012)

Using Data to Explain Data

NYU TANDON SCHOOL OF ENGINEERING

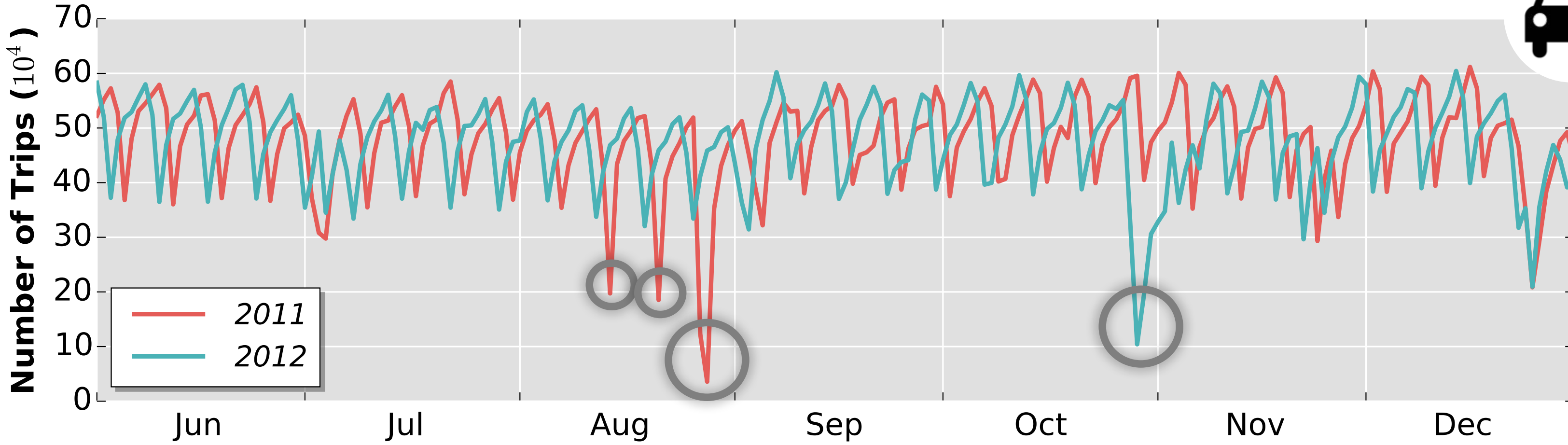
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Explaining Events



- Are these big drops data quality issues in the data?
- Or do they correspond to *real* events?

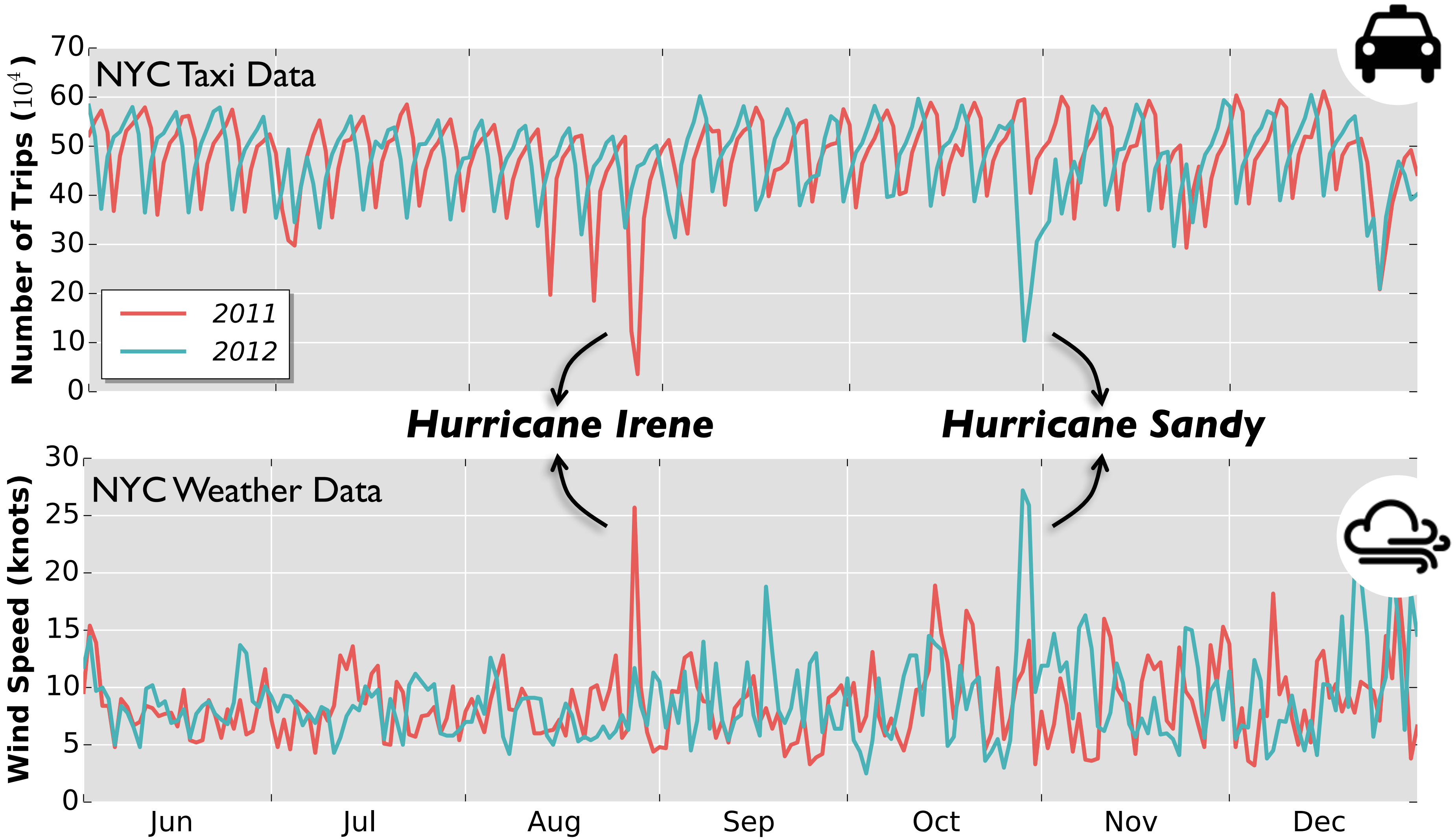
Explaining Events



- Are these big drops data quality issues in the data?
- Or do they correspond to *real* events?

Find all data sets related to the Taxi data set

Using Data to Explain Events



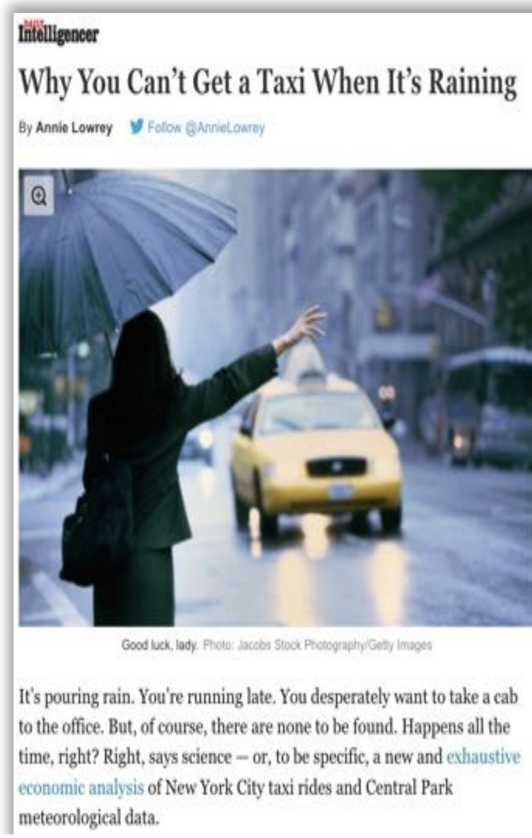
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Using Data to Explain and Predict NYC

1. Would a reduction in traffic speed reduce the number of accidents? What other factors contribute to accidents?
2. Why it is so hard to find a taxi when it is raining?

http://nymag.com/daily/intelligencer/2014/11/why-you-cant-get-a-taxi-when-its-raining.html



Daily Intelligencer

Why You Can't Get a Taxi When It's Raining

By Annie Lowrey Follow @AnnieLowrey

Good luck, lady. Photo: Jacobs Stock Photography/Getty Images

It's pouring rain. You're running late. You desperately want to take a cab to the office. But, of course, there are none to be found. Happens all the time, right? Right, says science — or, to be specific, a new and exhaustive economic analysis of New York City taxi rides and Central Park meteorological data.

Urban Data Interactions

By uncovering **relationships** between data sets, we can

- Better understand a city and how its different components interact
- Discover important attributes that can inform the construction of predictive models

Where to start?

- Data are available!
- Answers are likely in the data
- But there are too many data sets, and even more attributes to consider

NYC OpenData	1,200 data sets (and counting)	8 attributes per data set
weather	> 200 attributes	

Which data sets to analyze?

The Data Polygamy Framework

- **Discover relationships** between data sets to better understand urban data and how the different components of city interact
- Each data set can be related to **zero or more** data sets through several attributes

Data sets are polygamous!

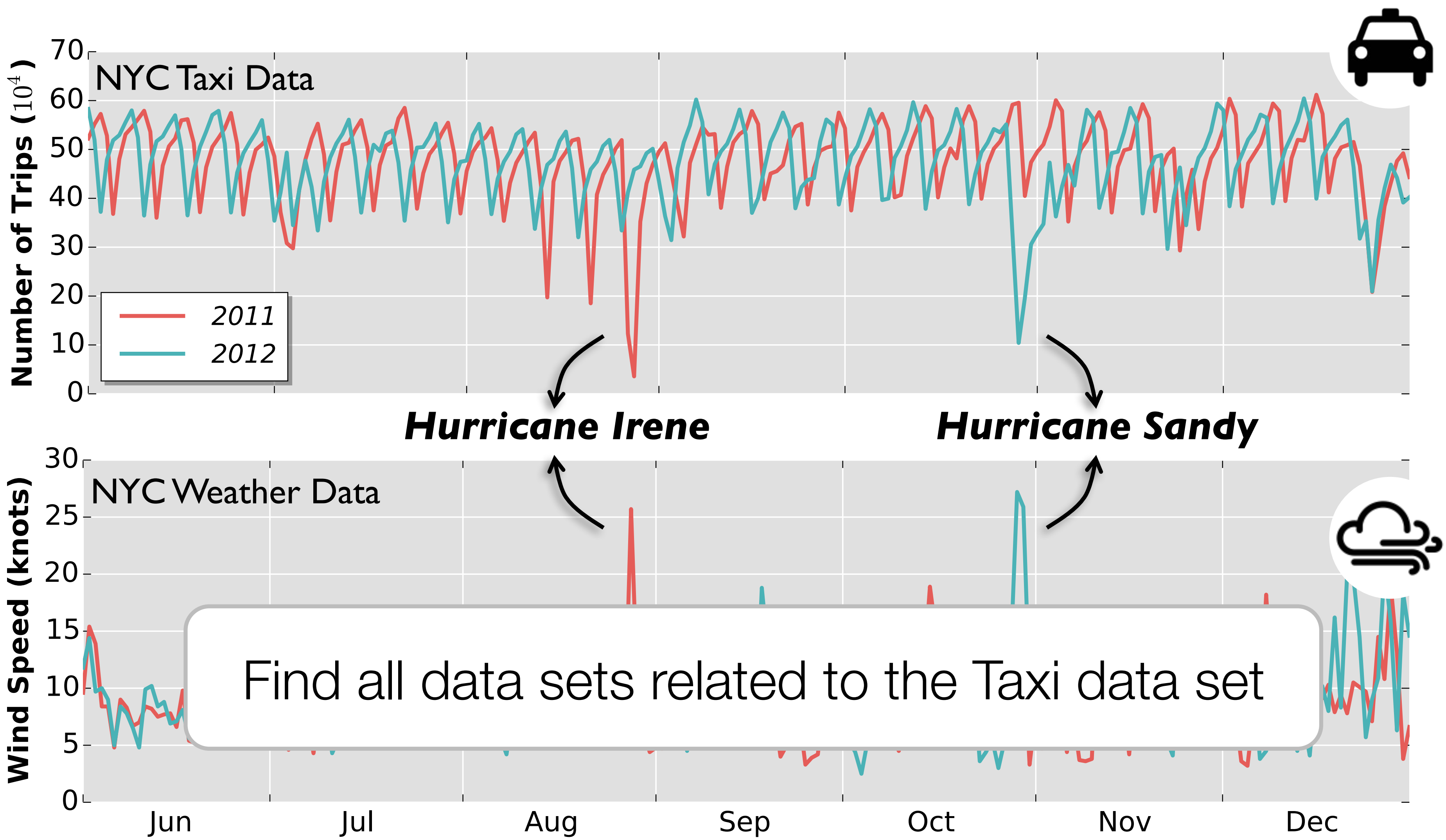
- Guide users in **data discovery and analysis** by allowing them to pose ***relationship queries***

Find all data sets related to a given data set $\mathbb{D}$

- **Support both hypothesis generation and testing**

[Chirigati et al., ACM SIGMOD 2016]

Hypothesis Generation

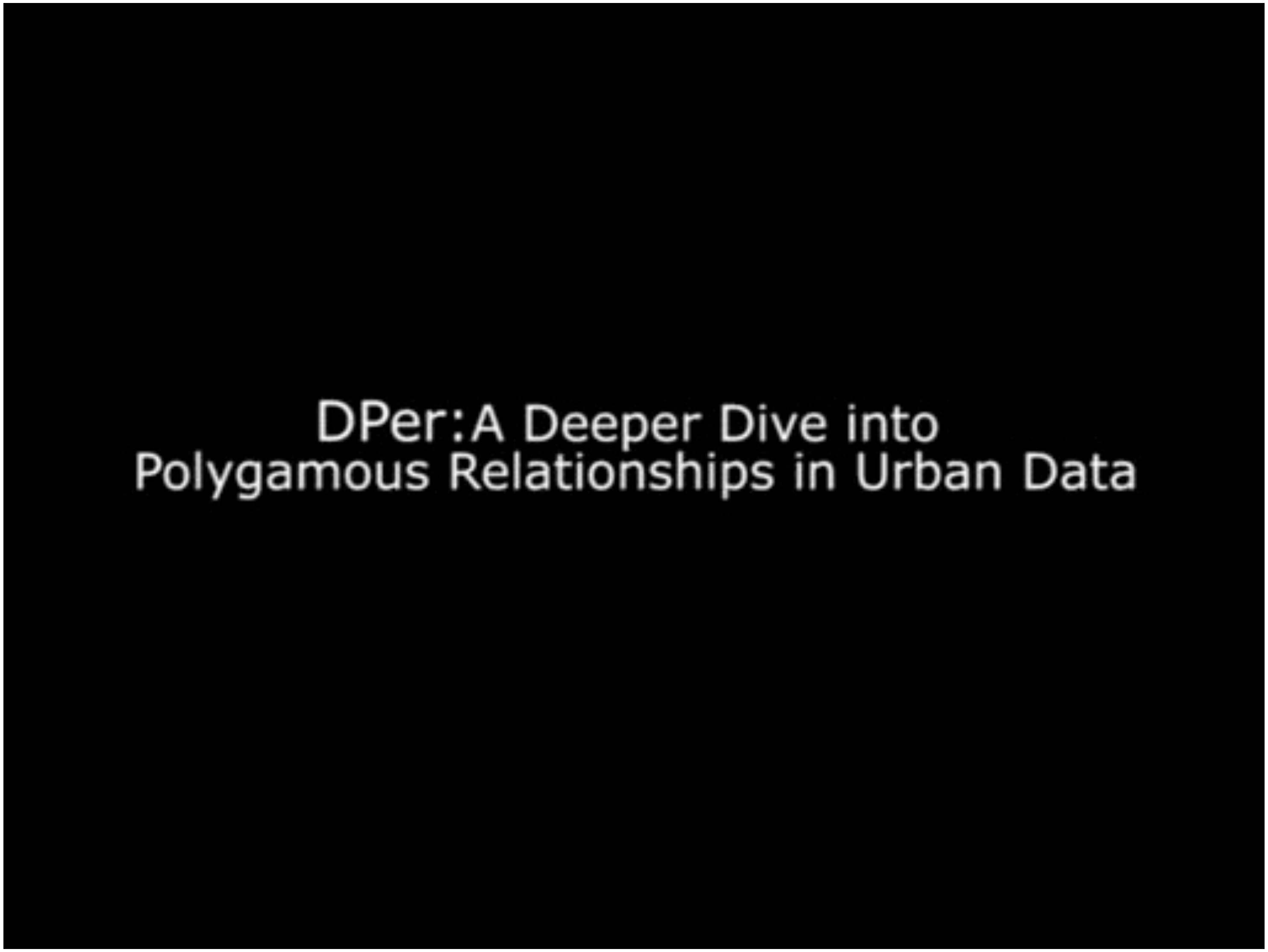


Find all data sets related to the Taxi data set

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Visually Exploring Relationships



https://vgc.poly.edu/~juliana/videos/dper2.mov

Takeaway: Urban Data Exploration

- Usability is of paramount importance
  - Need to empower domain experts to explore their data
- Exploration requires interactivity – improve the rate at which *users make observations, draw generalizations and generate hypotheses*
- Visualization must meet data management!
  - It already is at HILDA (Workshop on Human-In-the-Loop Data Analytics) http://hilda.io/2017
  - Growing number of papers in DB and Vis conferences
- By talking to and collaborating with domain experts, we can
  - Find many interesting research problems, and
  - Have practical impact

Conclusions

- New opportunities to better understand how cities work by analyzing their data exhaust
- Data has been democratized, now we need **tools that empower domain experts** to explore and extract knowledge from data
- Some steps towards **democratizing data exploration:**
  - Visual and interactive analysis of spatio-temporal data
  - Automatic event detection: point users to interesting features
  - Data Polygamy: discover relationships in data by leveraging a large collection of data sets
- Data Polygamy is also useful for data discovery, model construction, and explaining features

Conclusions

- Need interdisciplinary teams
  - Visualization, data management, computational topology
  - Collaboration with domain experts
- Many open problems around urban spatio-temporal data
  - *Cleaning, integration, querying, modeling, streaming (ongoing work)*
- Database community is well positioned to have tremendous practical impact
- Let's collaborate and build open-source tools!

Acknowledgments

- NYC Taxi & Limousine Commission for providing the data used in this paper and feedback on our results.
- Funding: Google, National Science Foundation, Moore-Sloan Data Science Environment at NYU, and DARPA.

NSF Google DARPA

GORDON AND BETTY MOORE FOUNDATION

ALFRED P. SLOAN FOUNDATION

고맙습니다
Merci
Thank you
Obrigada
благодаря
Kiitos
धन्यवाद
Tack
Danke
Ευχαριστω
Bedankt

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER