*The 1st ACM Summer School in Europe*
*Data Science and Big Data*

Association for Computing Machinery

# Urban Data: Opportunities, Challenges and State of the Art


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


NYU TANDON SCHOOL OF ENGINEERING

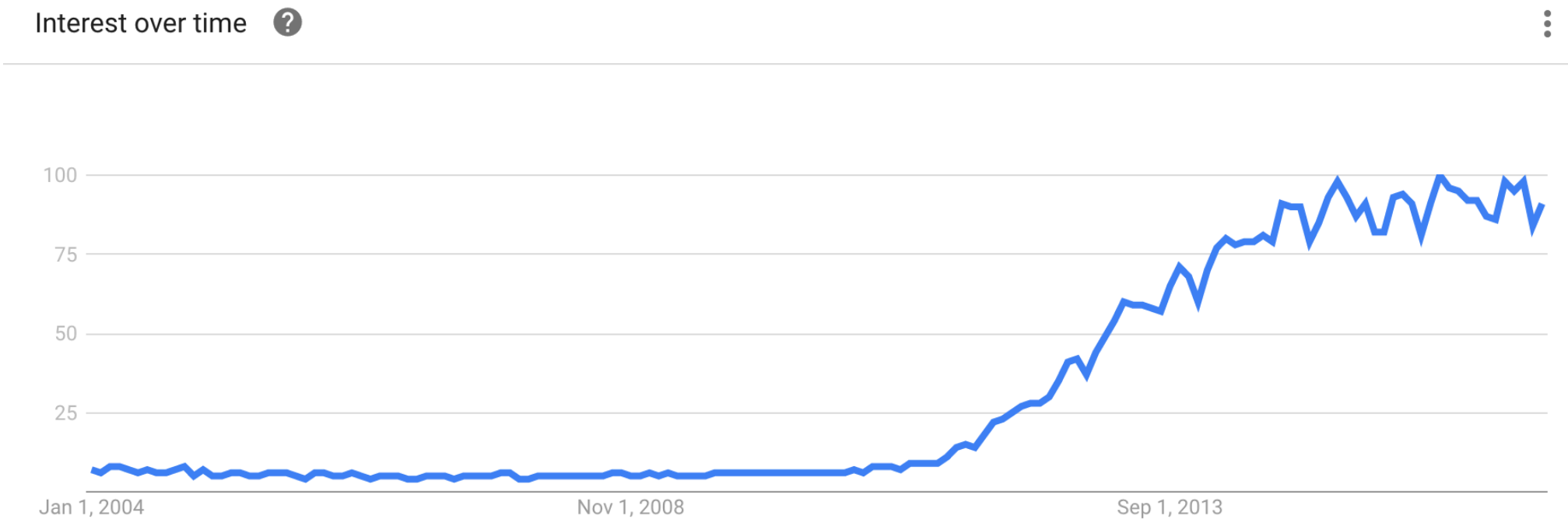
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

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


NYU TANDON SCHOOL OF ENGINEERING

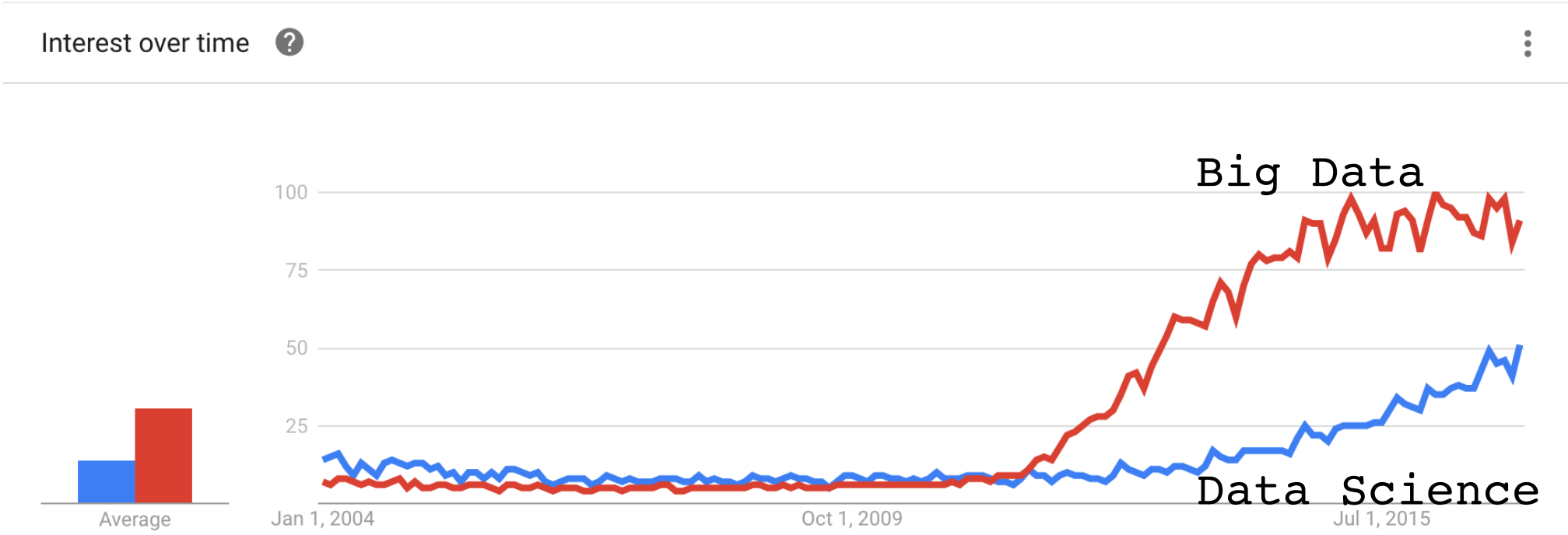
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Google Trends: Big Data




https://www.google.com/trends/explore?date=all&q=big%20data

# Google Trends: Big Data vs. Data Science

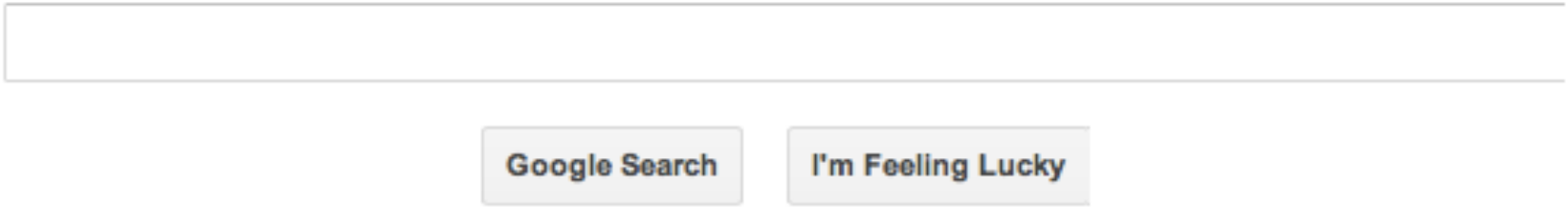



https://www.google.com/trends/explore?date=all&q=data%20science,big%20data

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

- Many success stories
  - Google: many billions of pages indexed, products, structured data
  - Facebook: 1.1 billion users using the site each month
  - Twitter: 517 million accounts, **250 million tweets/day**
- This has changed society!



# 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

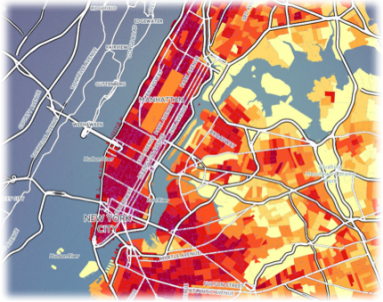
# Data Exhaust from Cities

*Infrastructure*

Condition, Operations


*Environment*

Meteorology, pollution, noise, flora, fauna



*People*

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

flickr

twitter

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

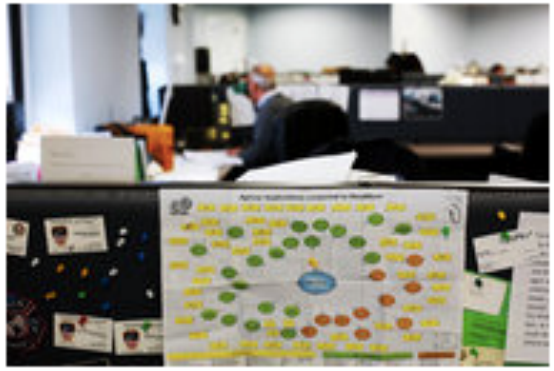
- NYC gets 25,000 illegal-conversion complaints a year and **only 200 inspectors** to handle them…
- Data-driven approach
  1. Integrated information from 19 different agencies that provided indication of issues in buildings, e.g., late taxes, foreclosure proceedings, service cuts, ambulance visits, 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

  http://furmancenter.org/

  - 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


Affect policy

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# 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**

twitter

DATA.GOV
EMPOWERING PEOPLE

YouTube
Broadcast Yourself

NYC OpenData

BRASIL Acesso à informação
Dados Abertos
GOVERNO FEDERAL

facebook

data.gov.in
Open Government Data Platform India

सत्यमेव जयते

Montréal
PORTAIL DONNÉES OUVERTES

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

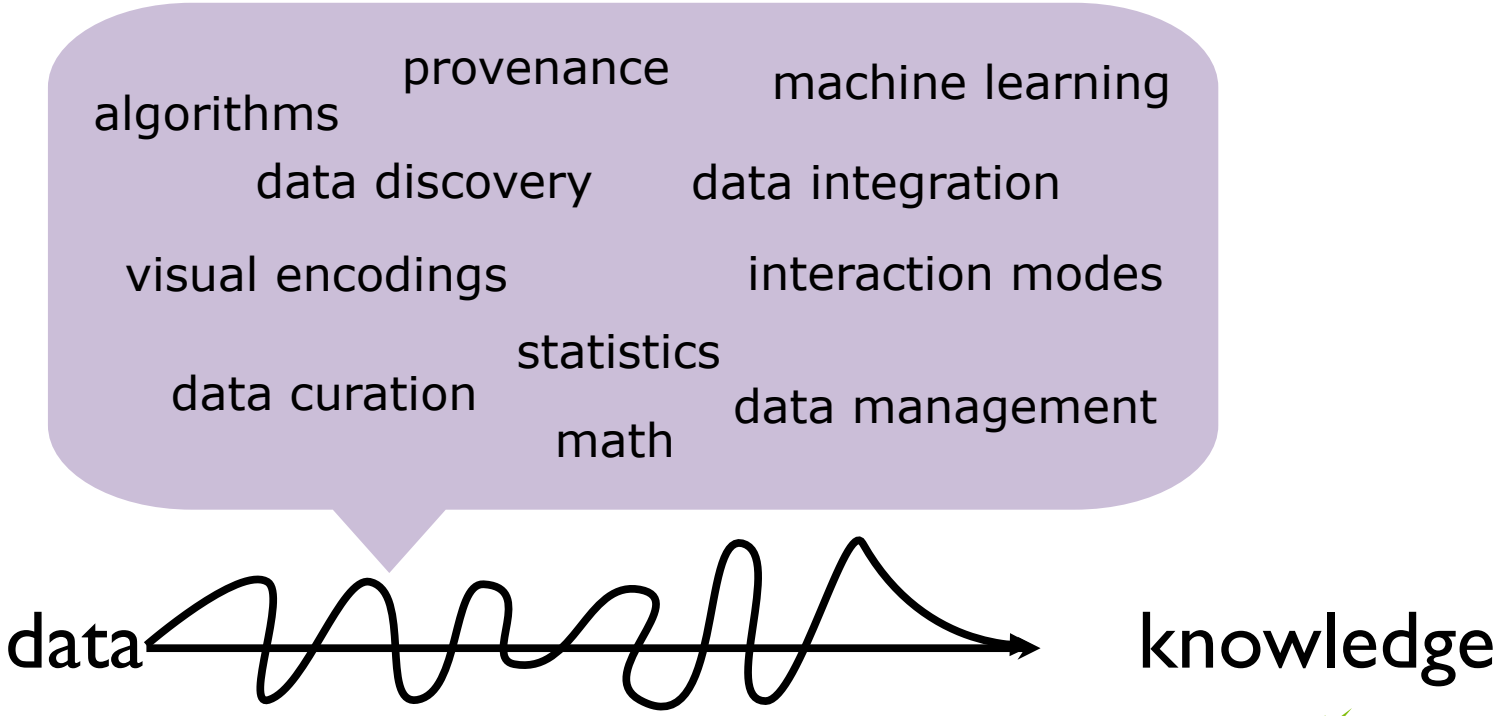
data.gouv.fr

NYU TANDON SCHOOL OF ENGINEERING

# 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



# 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!

# Outline

- 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
- Transparency and Reproducibility

# 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]




## 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

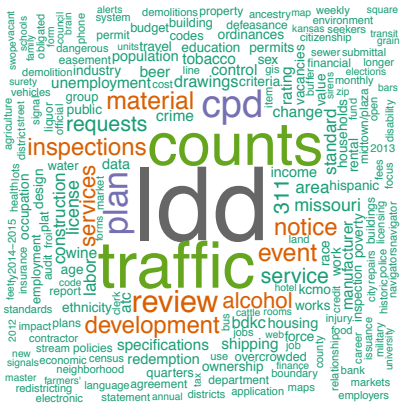
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# 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

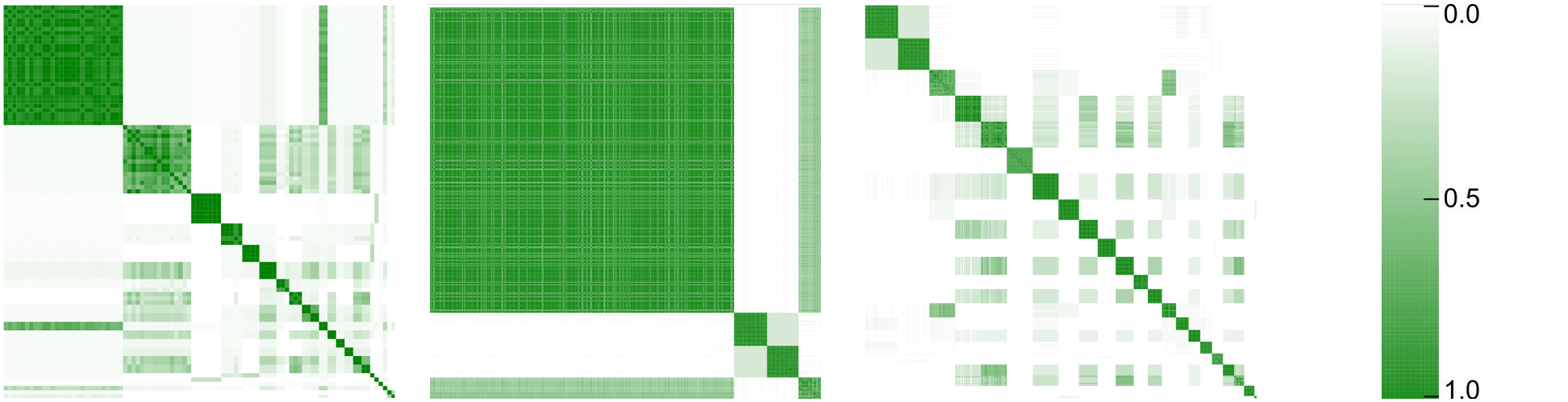
| *No. of records* | *Percentage of total* |
|---|---|
| 0–1K | 65.3 |
| 1K–10K | 17.0 |
| 10K–100K | 11.7 |
| 100K–1M | 5.5 |
| 1M–10M | 0.3 |

>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

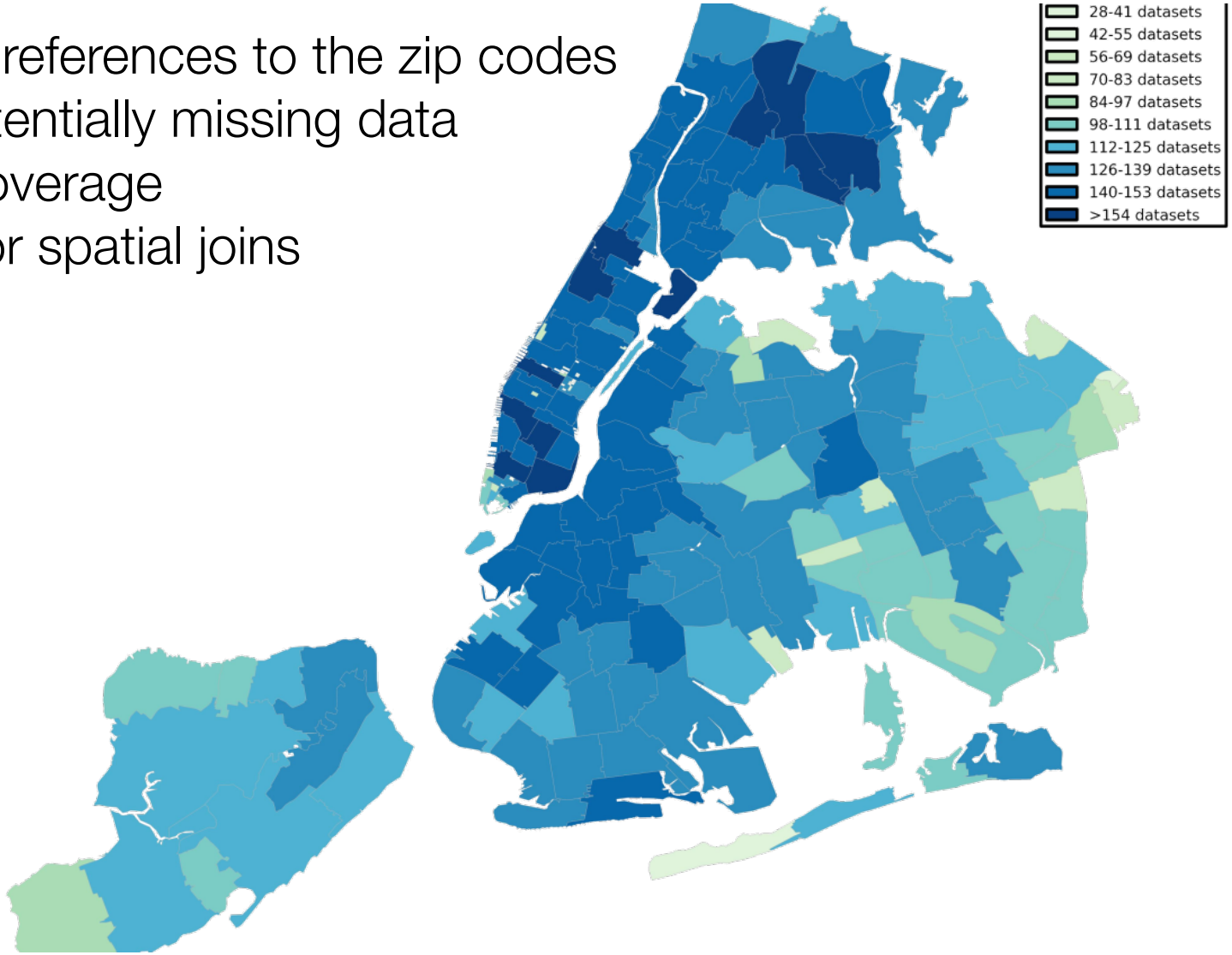
Attribute overlap among tables

- Potential for joining tables
- Hints about horizontally partitioned tables

# 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 |

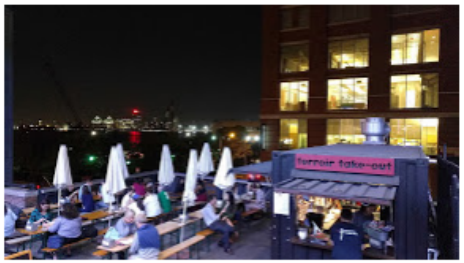
https://data.cityofnewyork.us/Health/DOHMH-New-York-City-Restaurant-Inspection-Results/43nn-pn8j

# 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’ and 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>

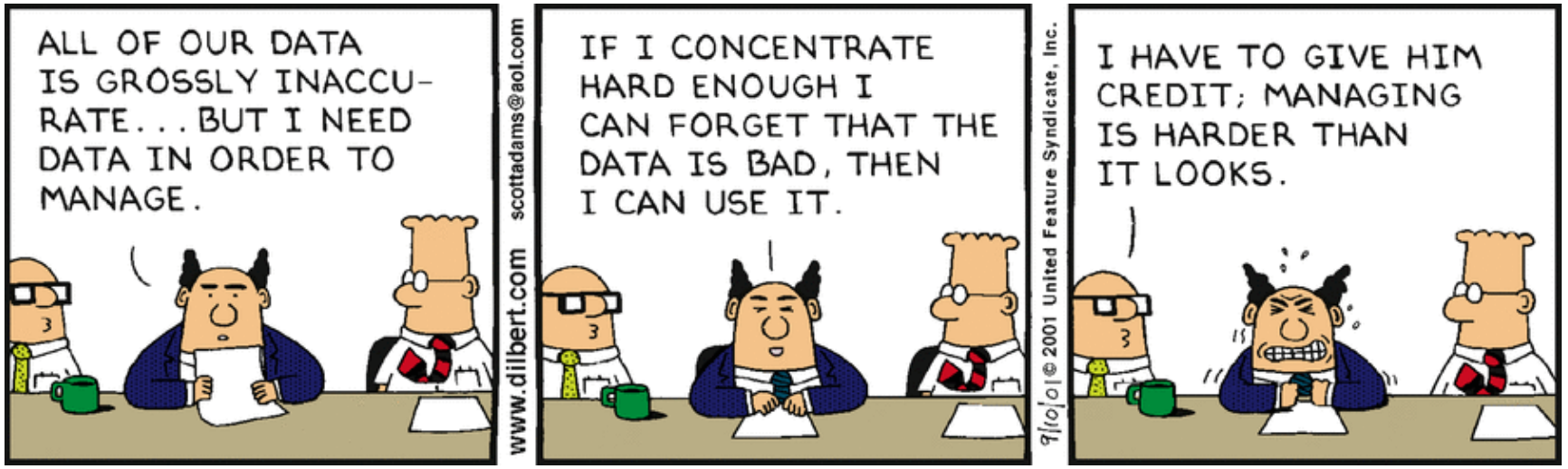
*Attribute: illegal and missing values*

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

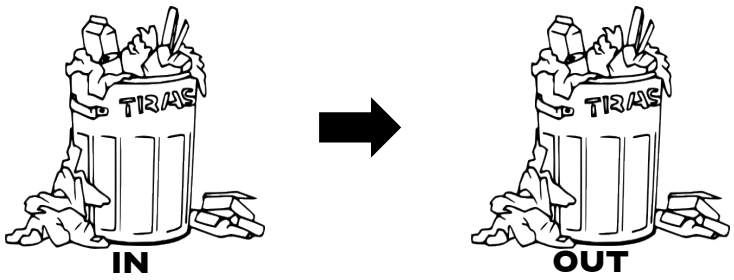
| ZIP | City |
|---|---|
| 10003 | NYC |
| 10003 | Chicago |

- 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**.



# 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
- $611bn per year is lost in the US

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

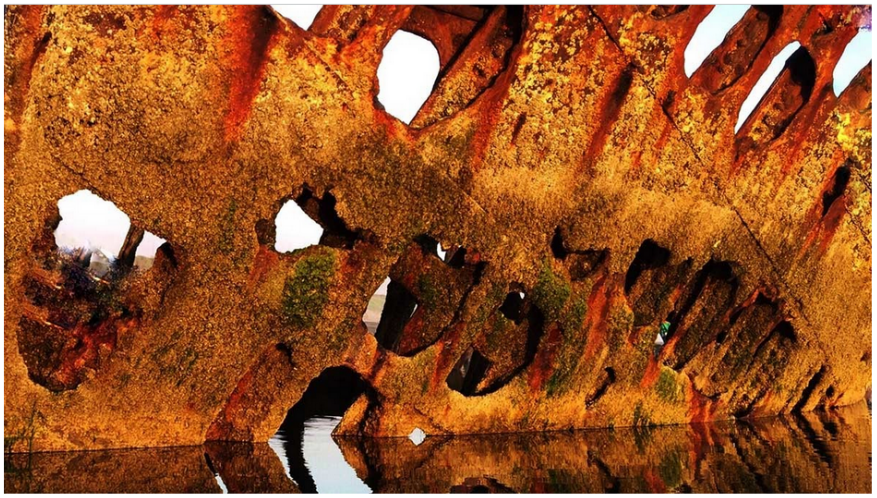
DATA

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

by Thomas C. Redman

SEPTEMBER 22, 2016

SAVE SHARE COMMENT TEXT SIZE PRINT $8.95 BUY COPIES



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, managers, knowledge workers, data scientists and others 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?

MAR 23, 2016 @ 09:33 AM 15,078 VIEWS

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

**Gil Press,** CONTRIBUTOR

*I write about technology, entrepreneurs and innovation.* FULL BIO

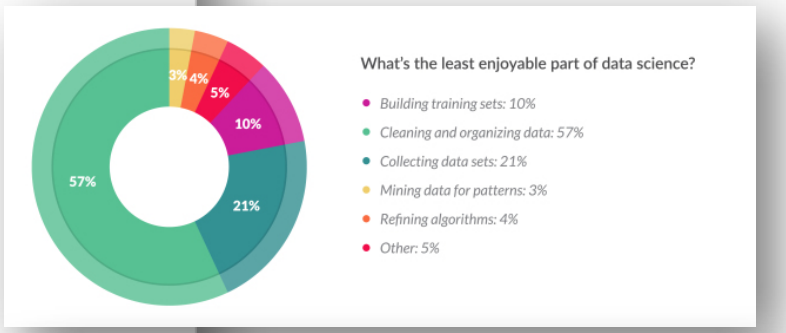
Opinions expressed by Forbes Contributors are their own.

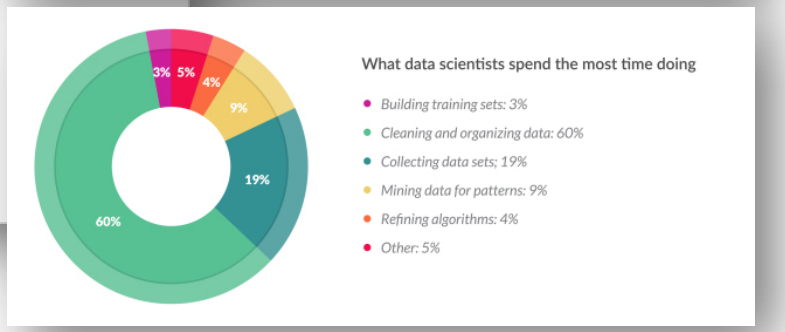
**TWEET THIS**

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

# 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)

Modified from H. Müller

# 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

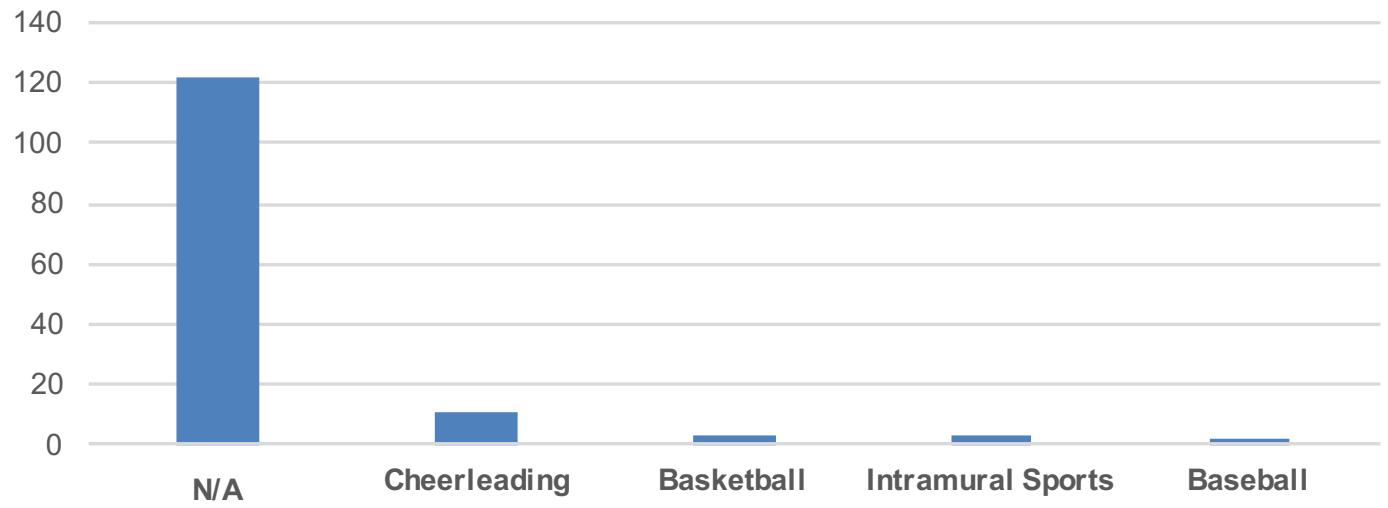
- *Sort attribute values in alphabetical order*
  - 'Interesting' values often appear at the beginning or end of list.
- *Frequency outliers*
  - NULL values sometimes have significantly different frequency (high or low) compared to other column values.

# 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

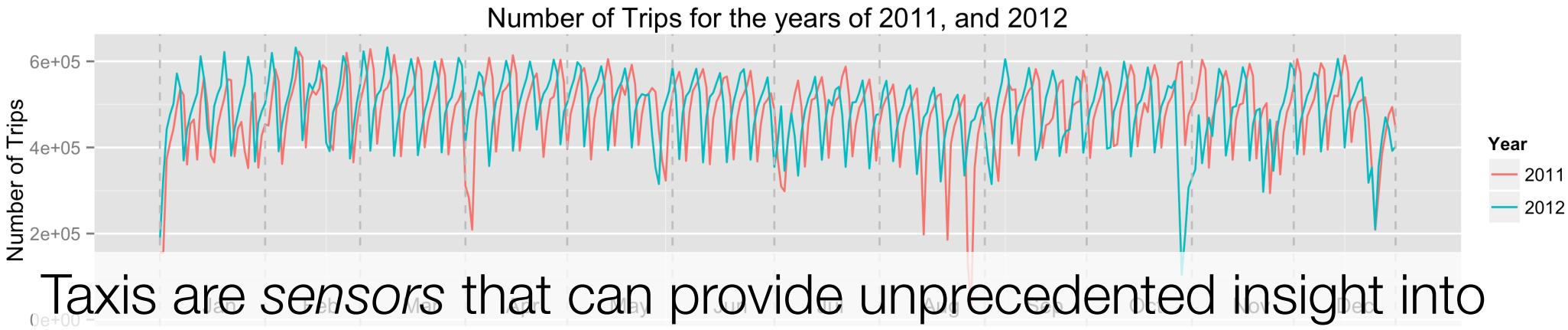
- *Sort attribute values in alphabetical order*
  - 'Interesting' values often appear at the beginning or end of list
- *Frequency outliers*
  - NULL values sometimes have significantly different frequency (high or low) compared to other column values
- *Regular expressions*
  - Find values that do not 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



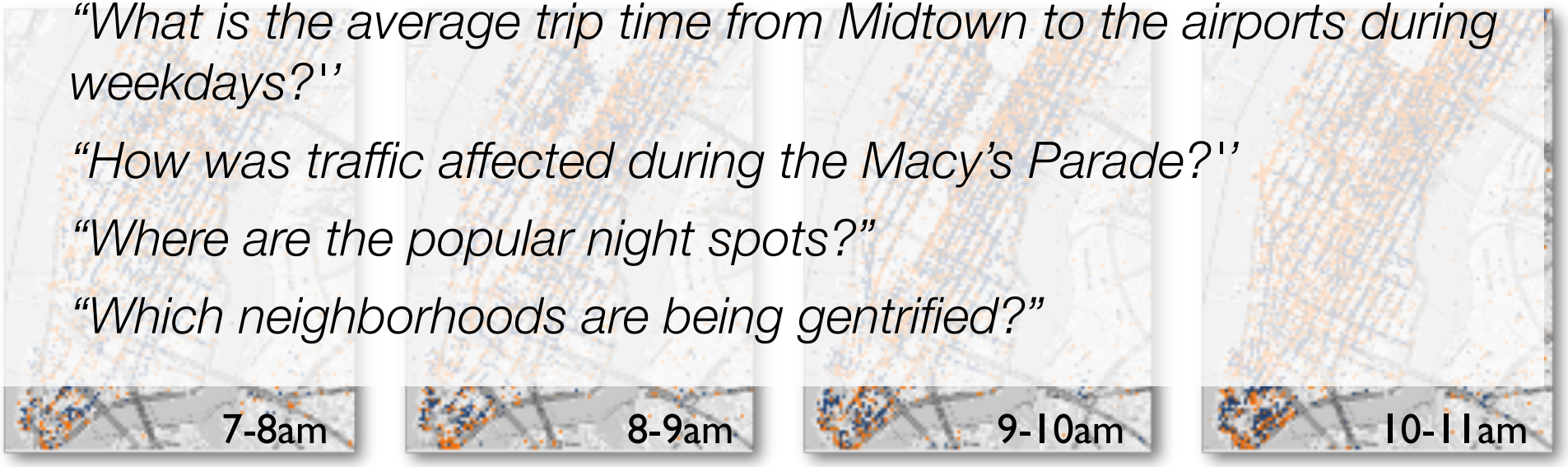

Taxis are *sensors* that can provide unprecedented insight into city life: economic activity, human behavior, mobility patterns

*"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?"*




NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Taxi Data: What to Clean and not to Clean

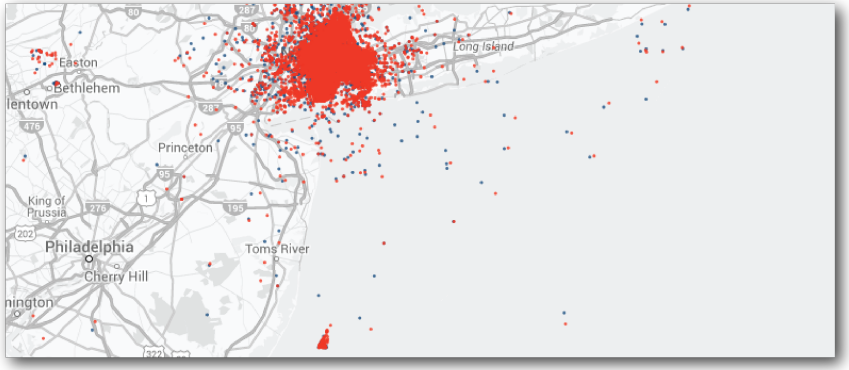
| Dataset | Statistic | Trip Duration (min) | Trip Distance (mi) | Fare Amount (US$) | Tip Amount (US$) |
|---|---|---|---|---|---|
| 2008 | Min | 0.00 | 0.00 | 0.00 | 0.00 |
| | Avg | 16.74 | 2.71 | 0.09 | 0.10 |
| | Max | 1440.00 | 50.00 | 10.00 | 8.75 |
| 2009 | Min | 0.00 | 0.00 | 2.50 | 0.00 |
| | 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

# Taxi Data: What to Clean and not to Clean


(a)

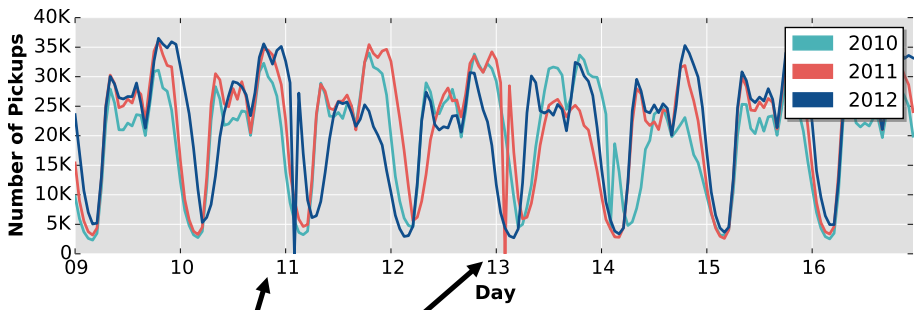


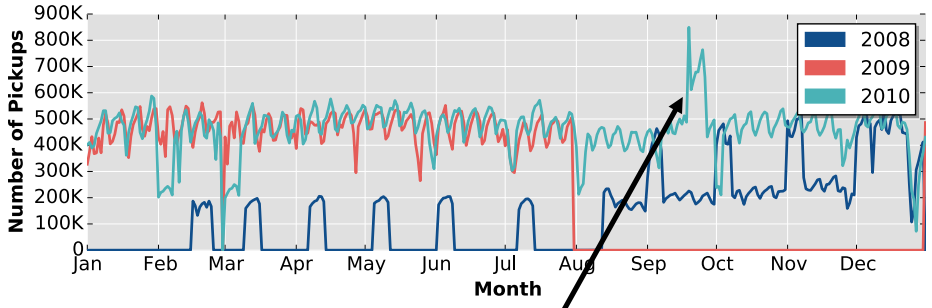

(b)


(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

Daylight savings:
March 13, 2011
March 11, 2012

Missing data in 2008

Big spike on Sept 19th, 2010

Missing data in 2009

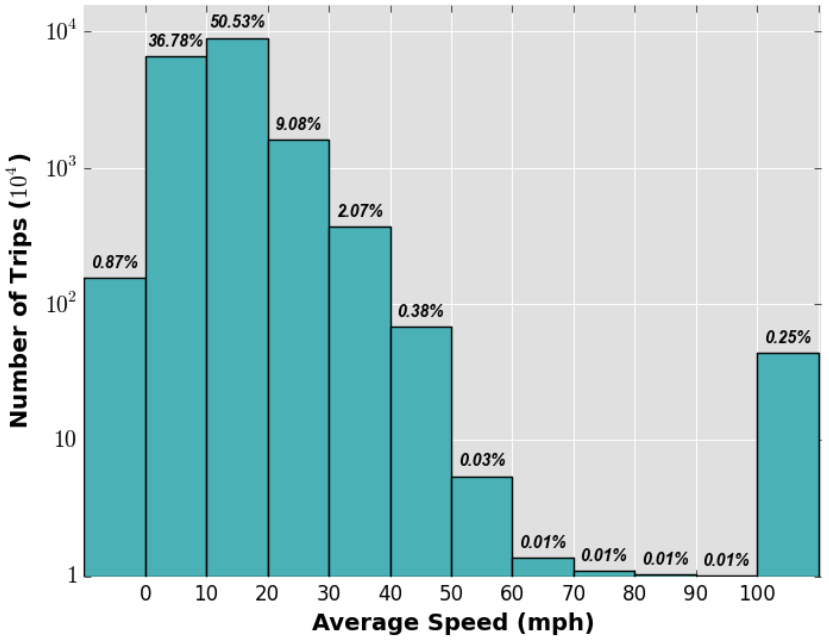
Unusually large number of consecutive and extremely short trips (lasting less than a minute)

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* × *UserTask* → (*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

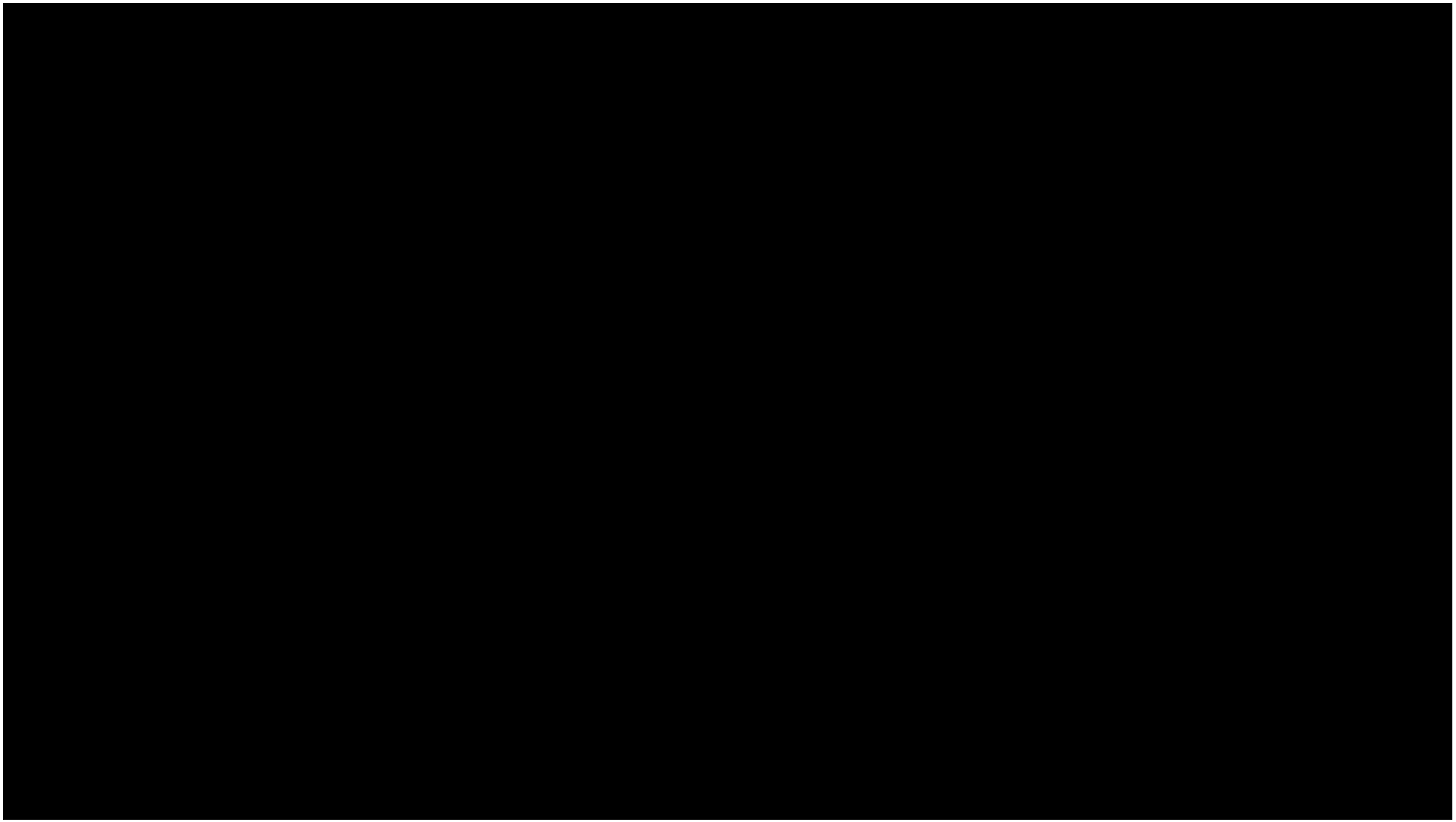
NYU TANDON SCHOOL OF ENGINEERING

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



NYU TANDON SCHOOL OF ENGINEERING

http://www.taxivis.org

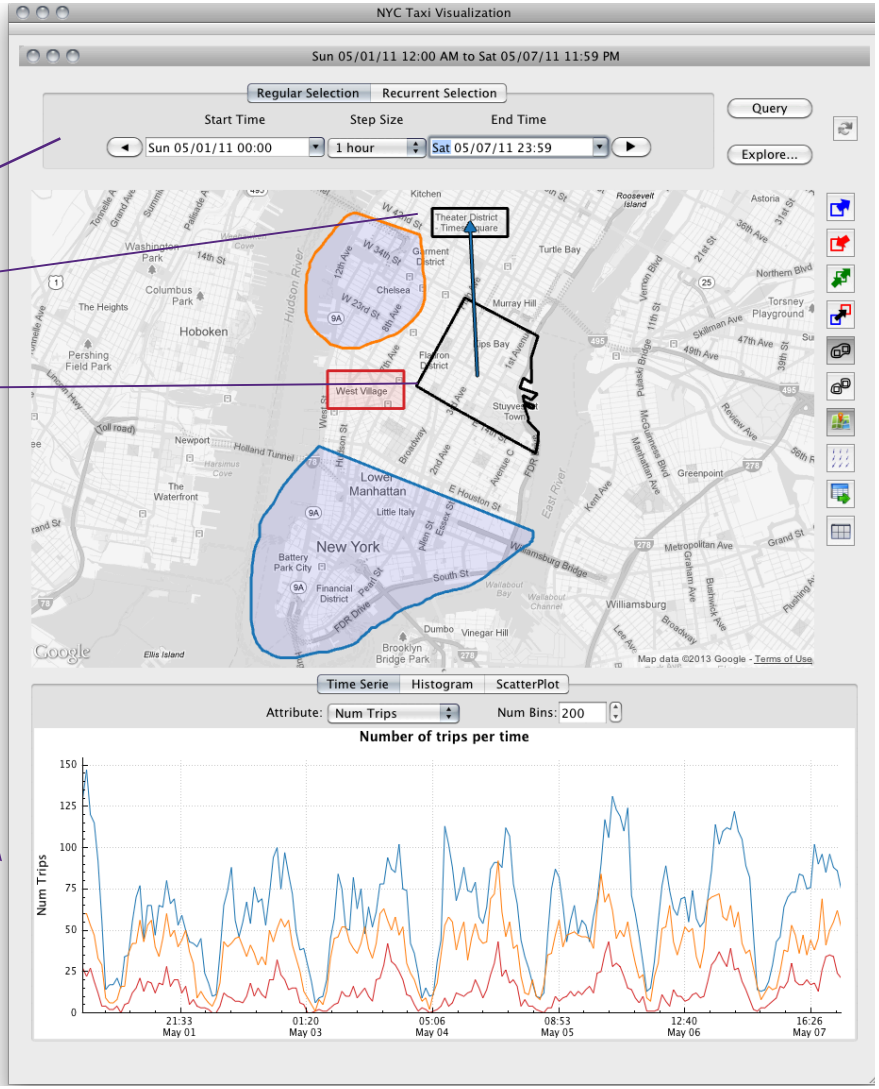
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# 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”
```

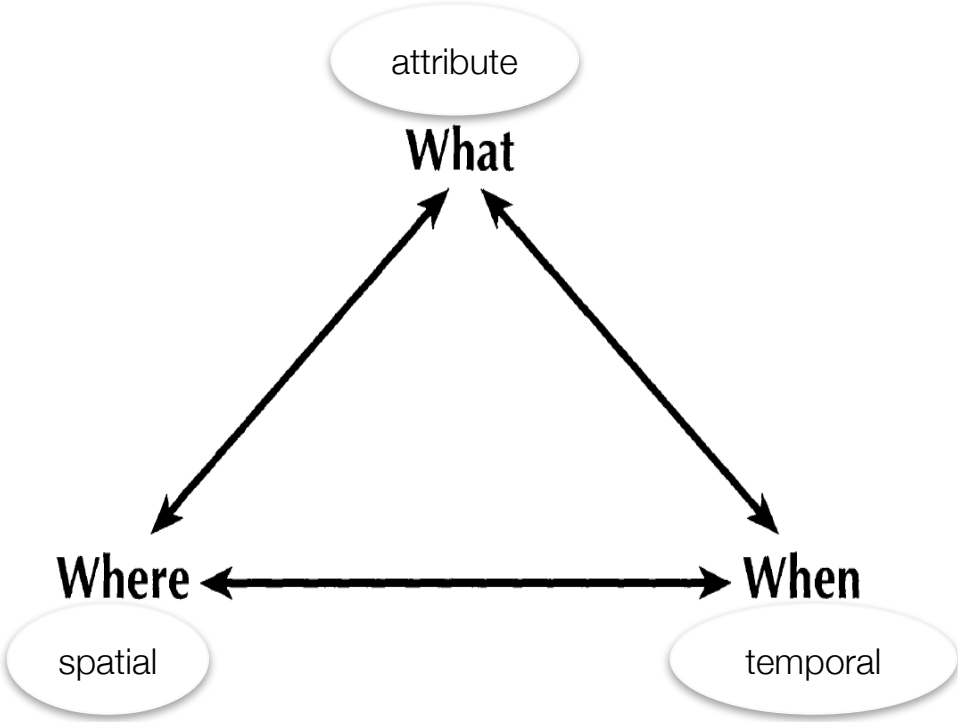
Data selection and result exploration are unified



# Visual Query Model

Expressiveness:

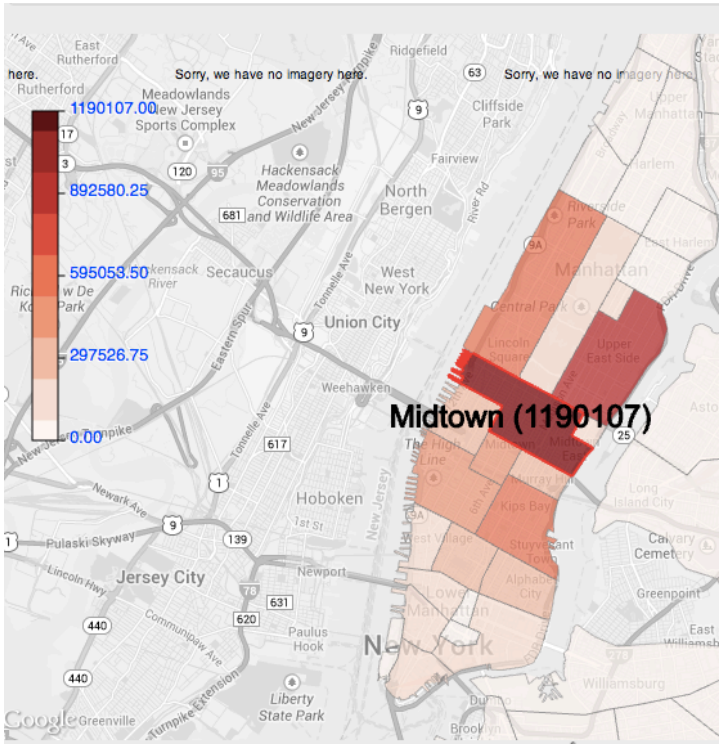
- when + where ➔ what: *"What is the average trip time from Midtown to the airports during weekdays?''*
- when + what ➔ where: *"Where are the hot spots in Manhattan in weekends?"*
- where + what ➔ when: *"When were activities restored in Lower Manhattan after the Sandy hurricane?"*




Peuquet's Triad

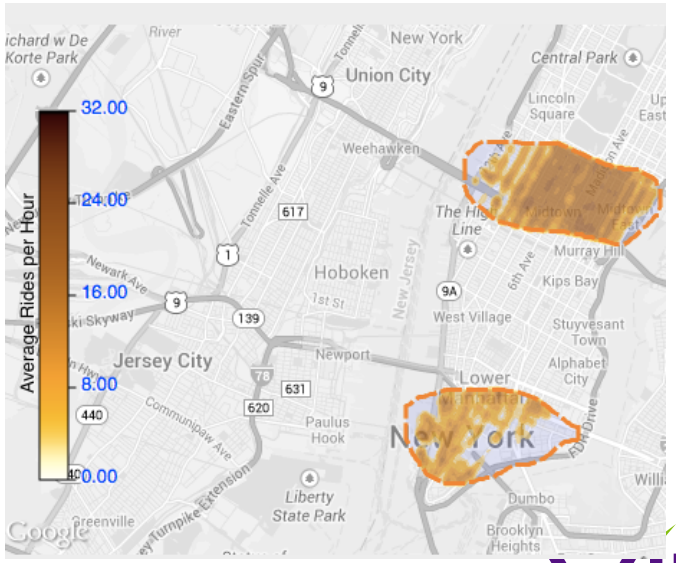
Model is also able to express other types of queries, including *when → what + where*, *where → when + what*, and *what → where + when*

# Selecting Regions – Spatial Constraints




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

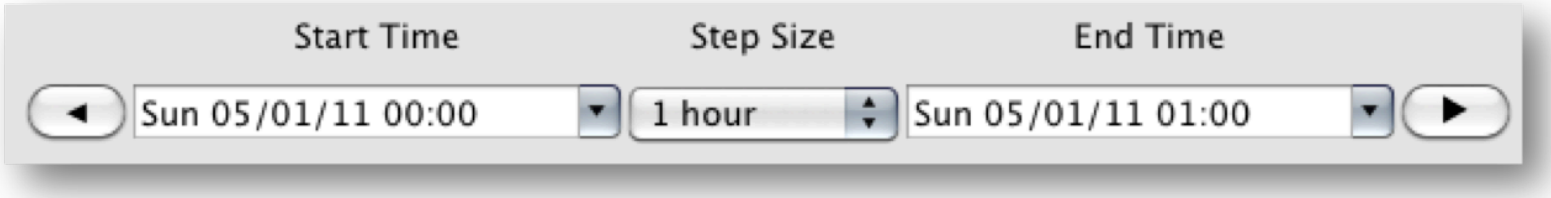

Free selection

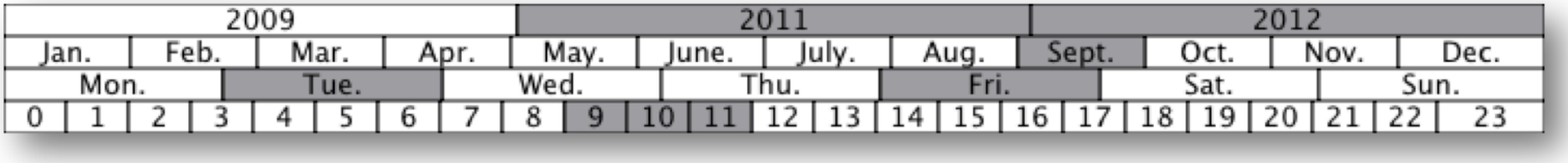



Group regions

# Selecting Time – Temporal Constraints

*Time interval*

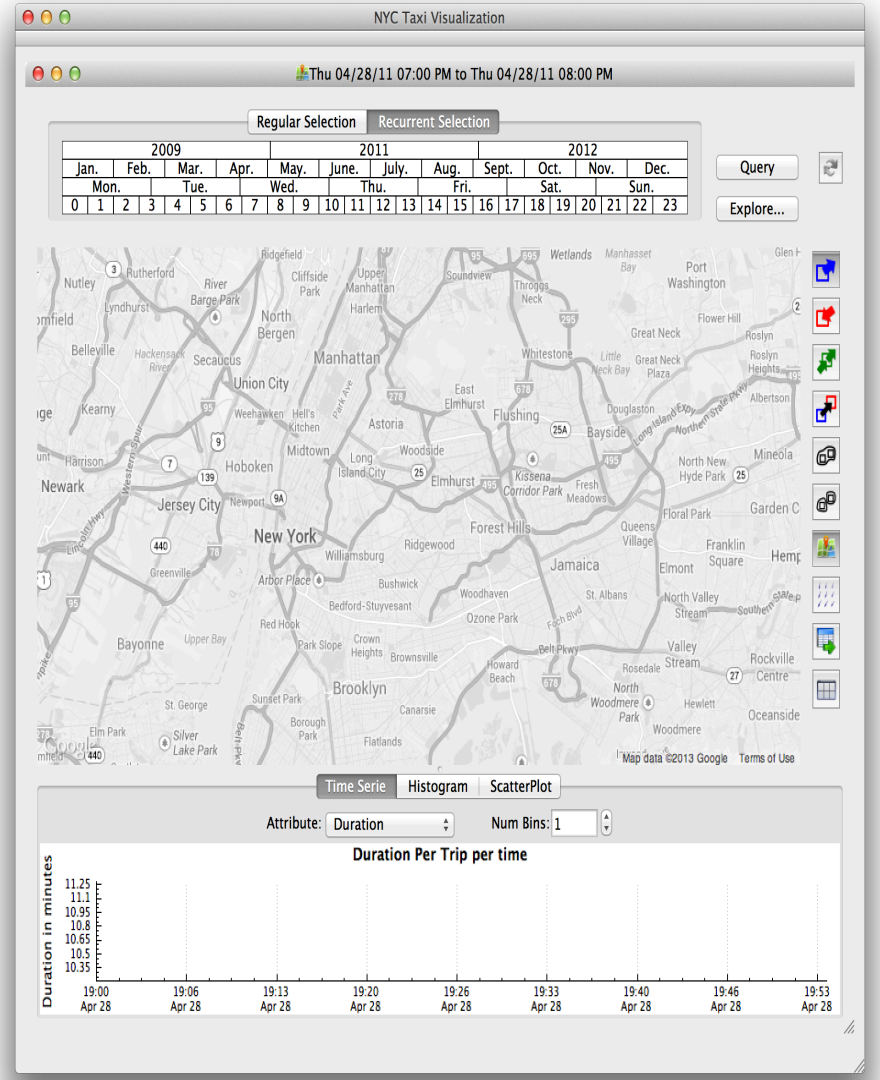



*Recurrent time patterns*

# When + Where → What

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



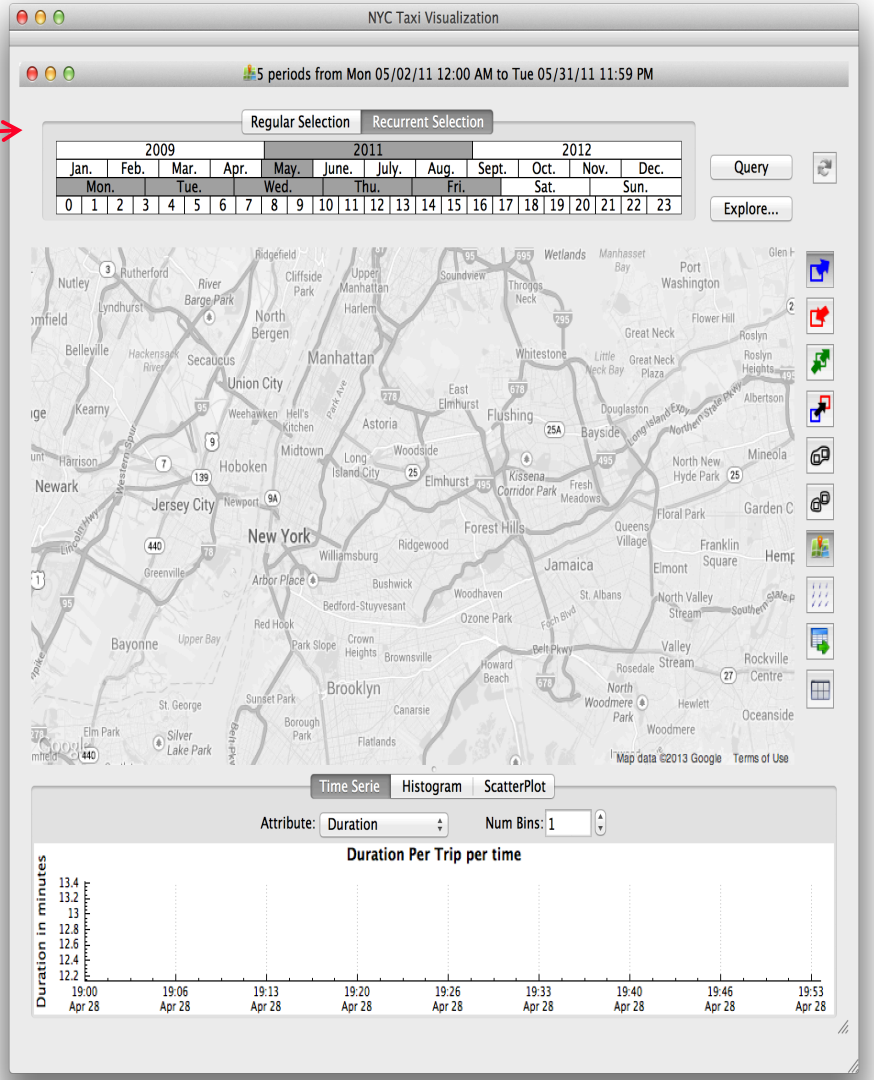

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# When + Where → What

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

When?




NYU TANDON SCHOOL OF ENGINEERING

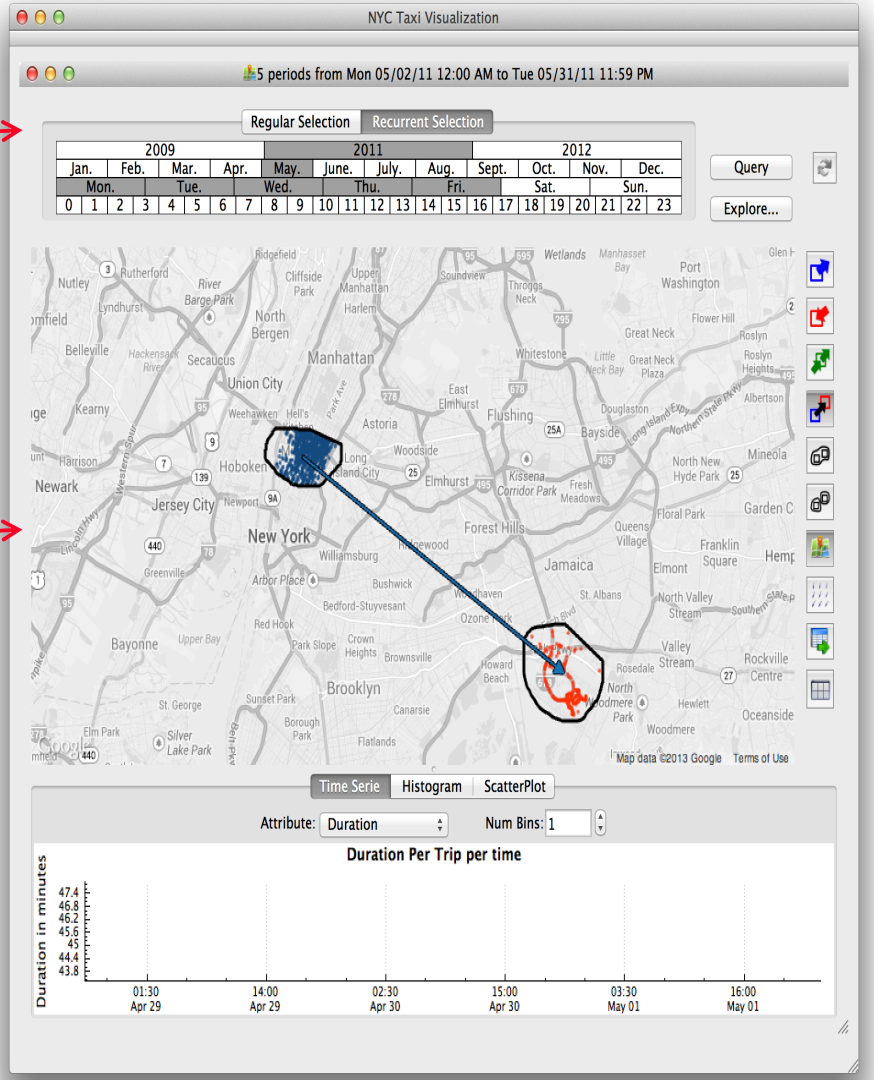
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# When + Where → What

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

When?

Where?




NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

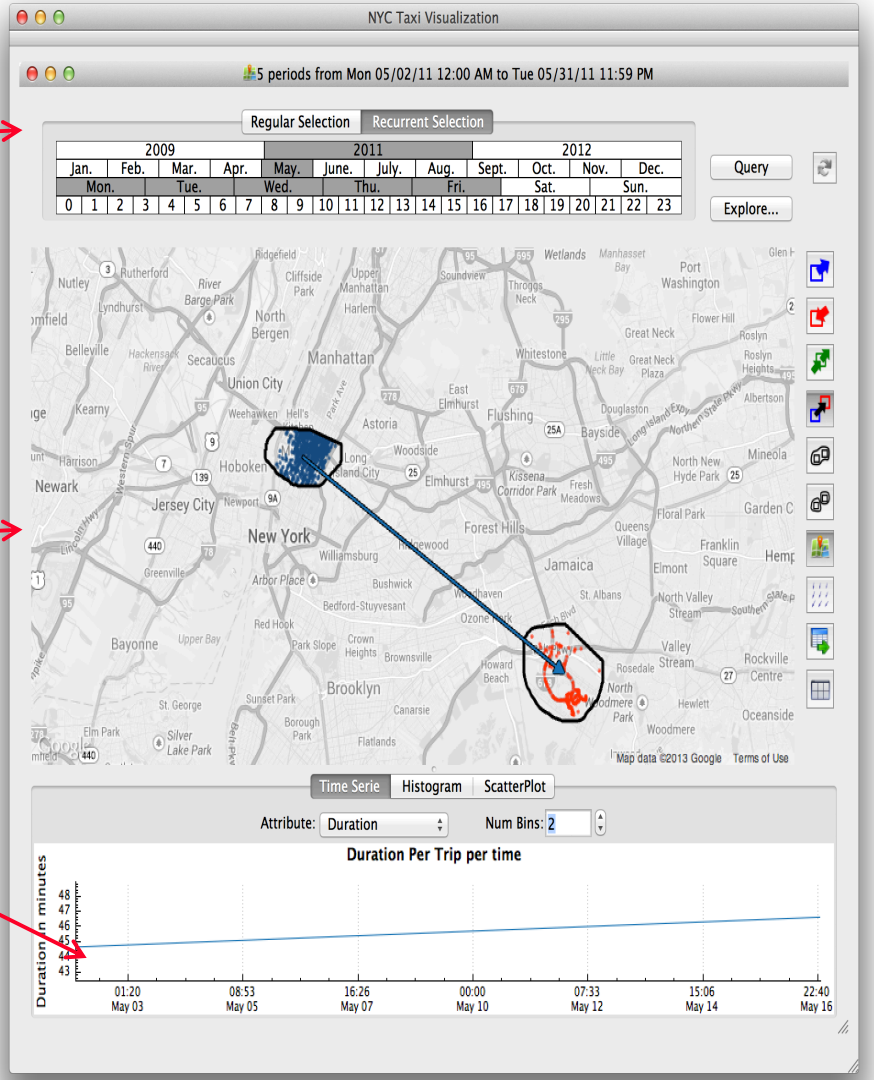
# When + Where → What

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

When?

Where?

What




NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

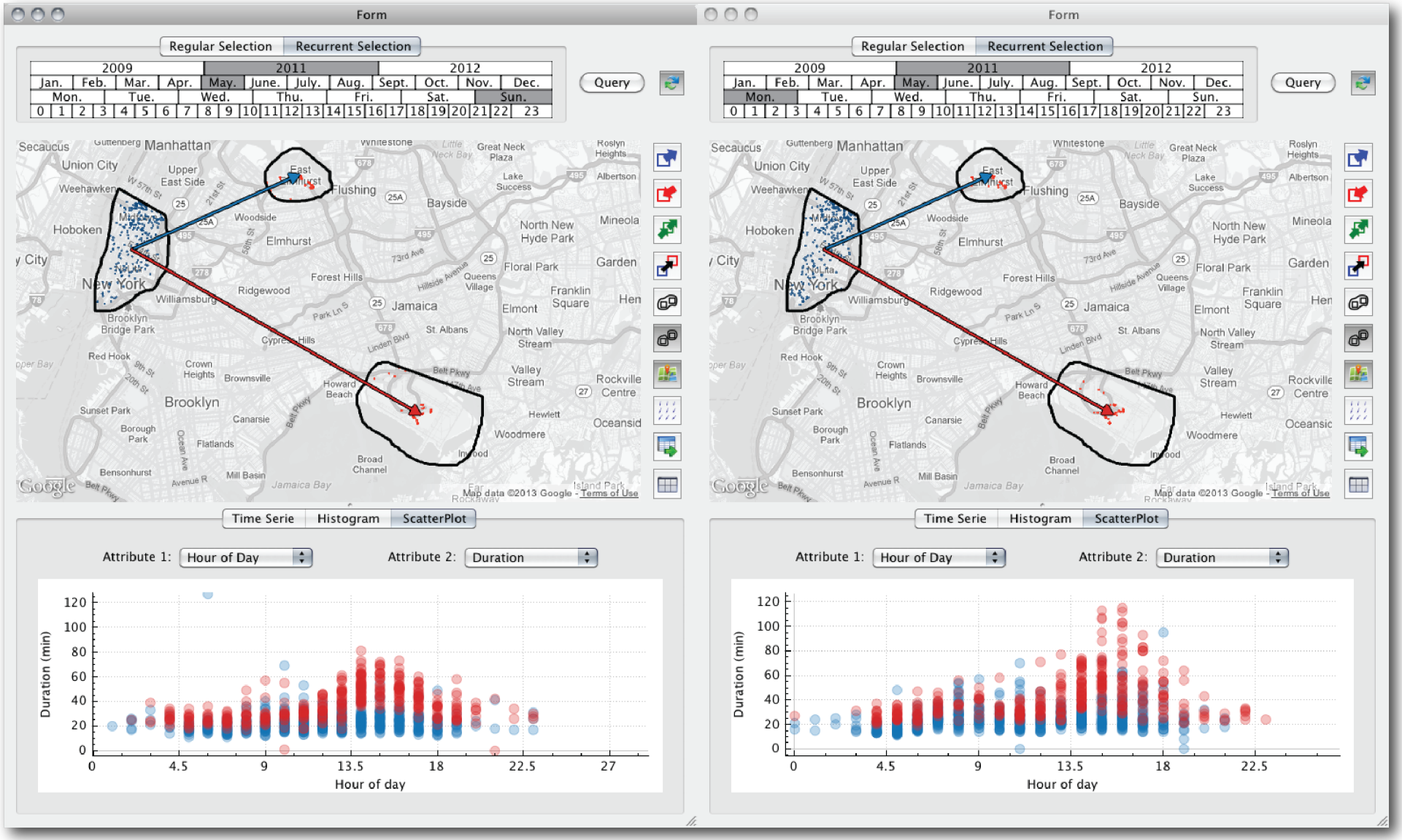
# Composing Queries

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

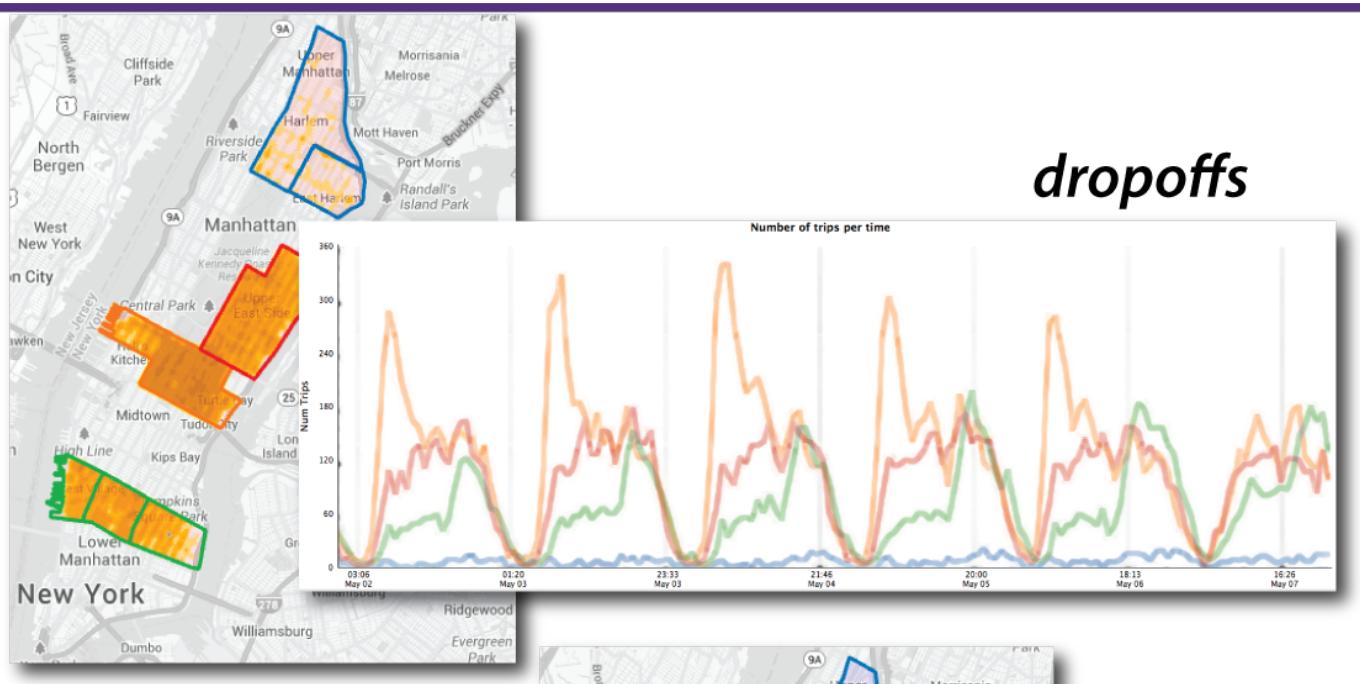


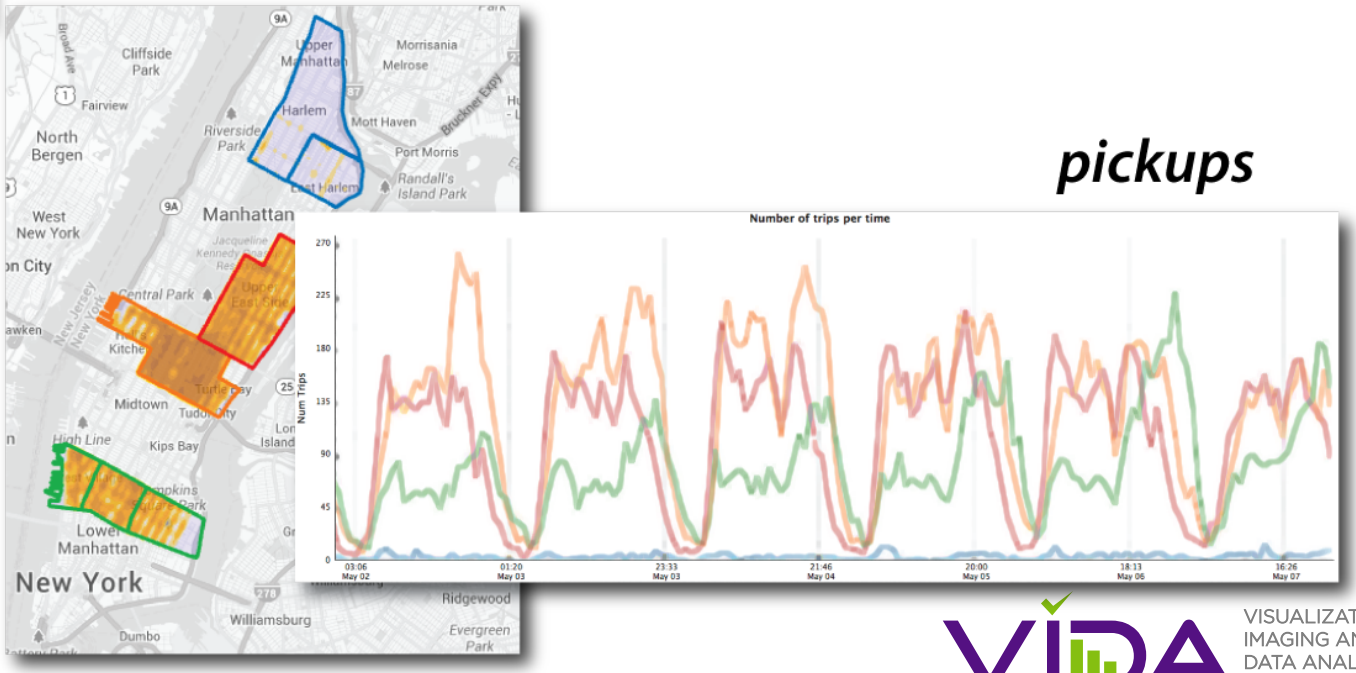

[Ferreira et al., IEEE TVCG 2013]

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# TaxiVis: Comparing Neighborhoods

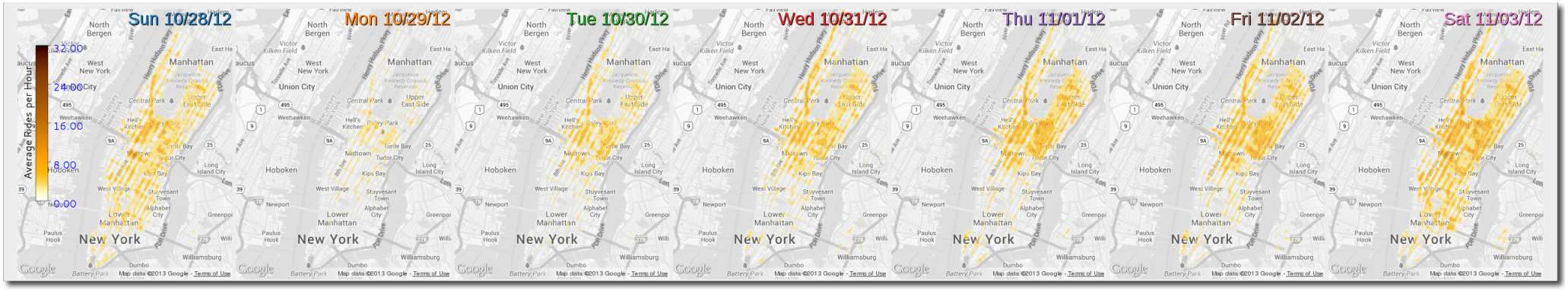


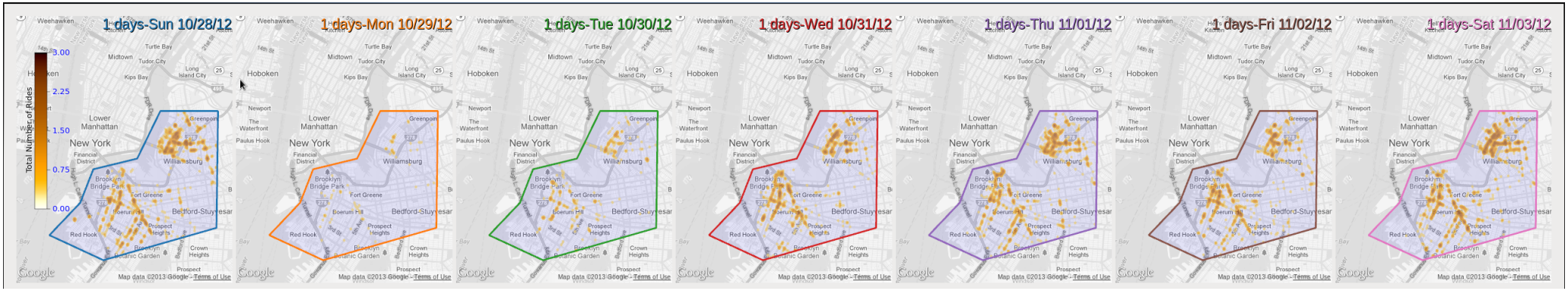




NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Exploring the Effect of Major Events: Sandy

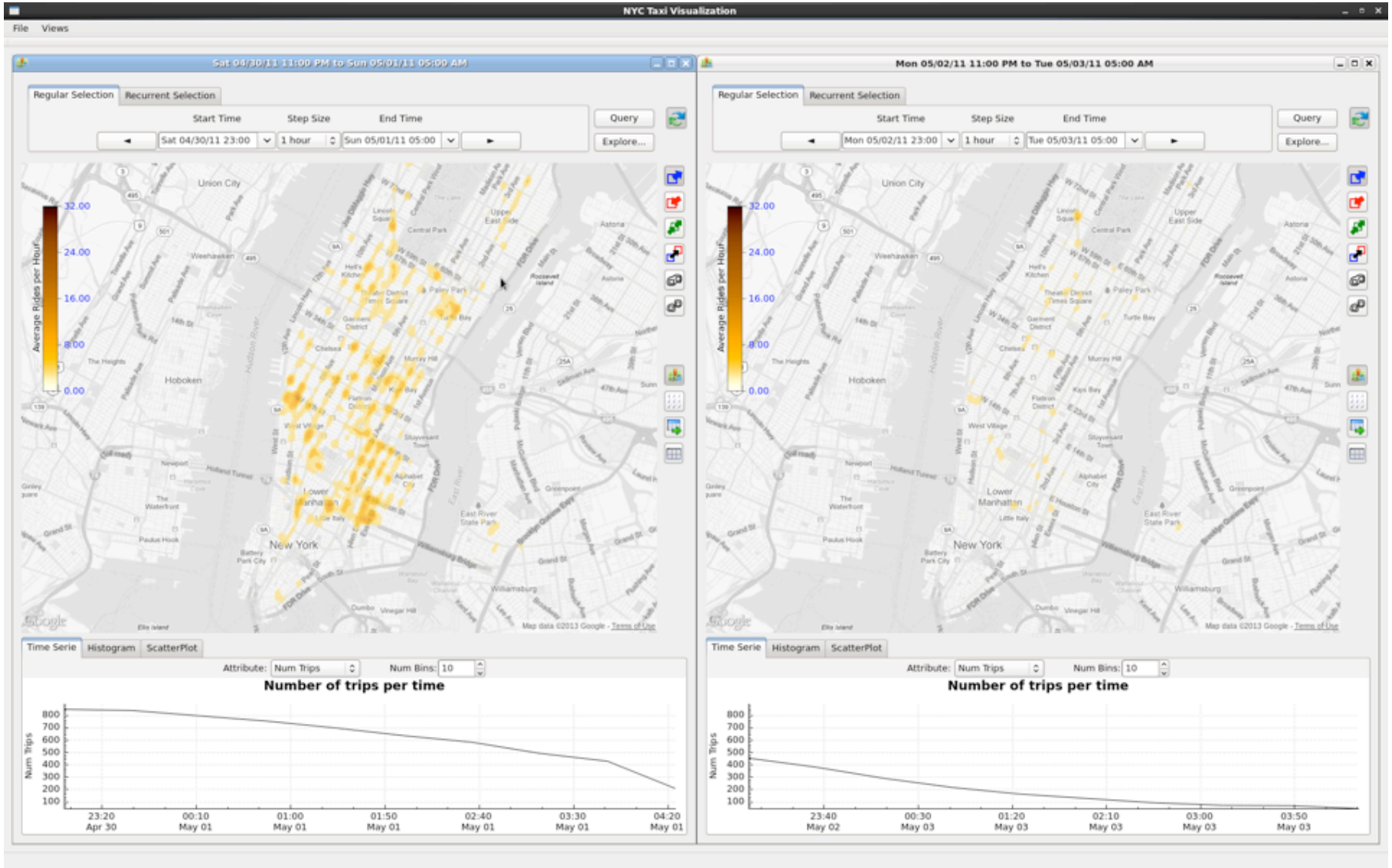



NYU TANDON SCHOOL OF ENGINEERING

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

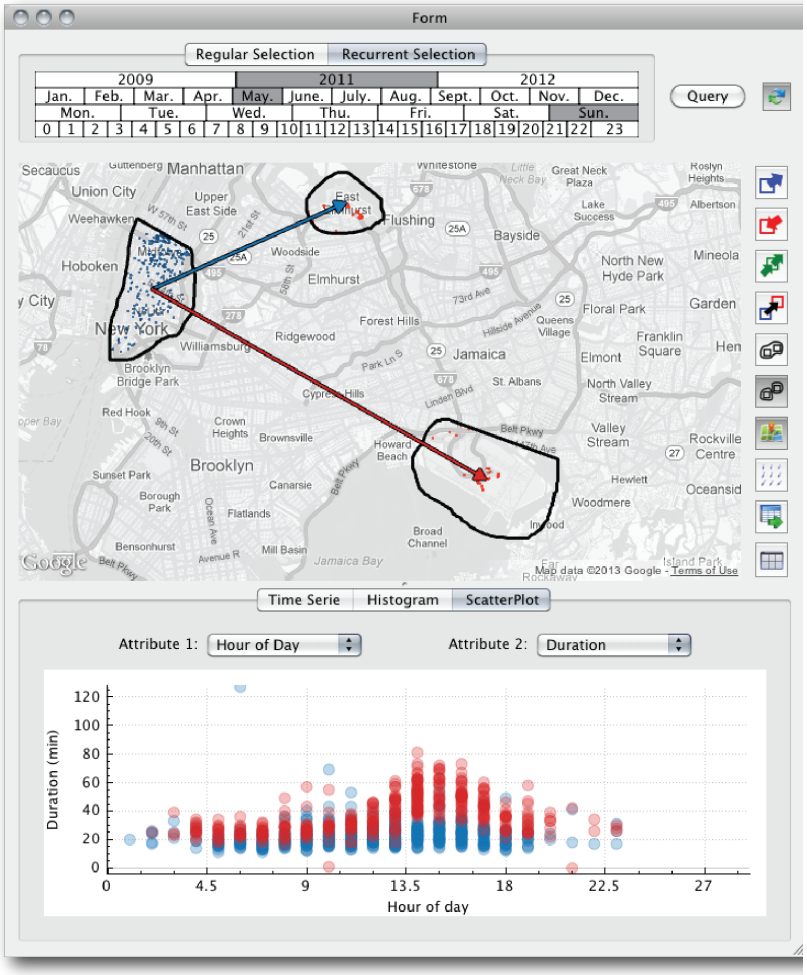
# Challenge: *Interactive* Query Evaluation

- 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]

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Challenge: *Interactive* Query Evaluation

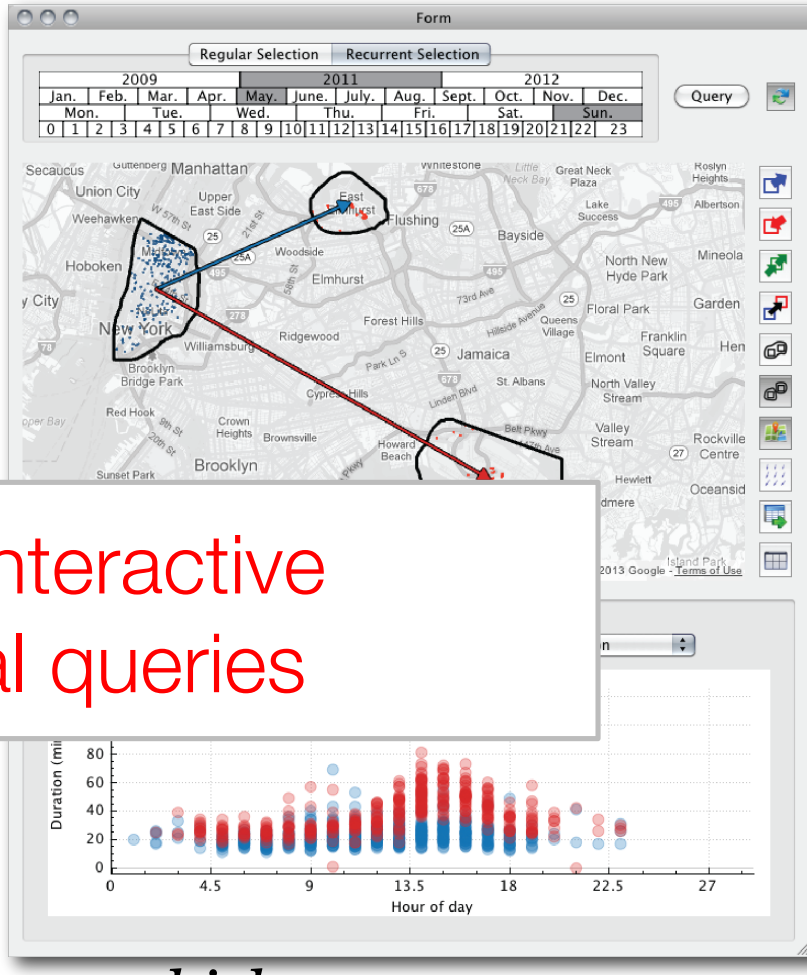
- Typical query:

*Find all trips that occurred between*

*lower Manhattan and the two airports,*

*JFK and LGA,*

*during all Sundays in May 2011*

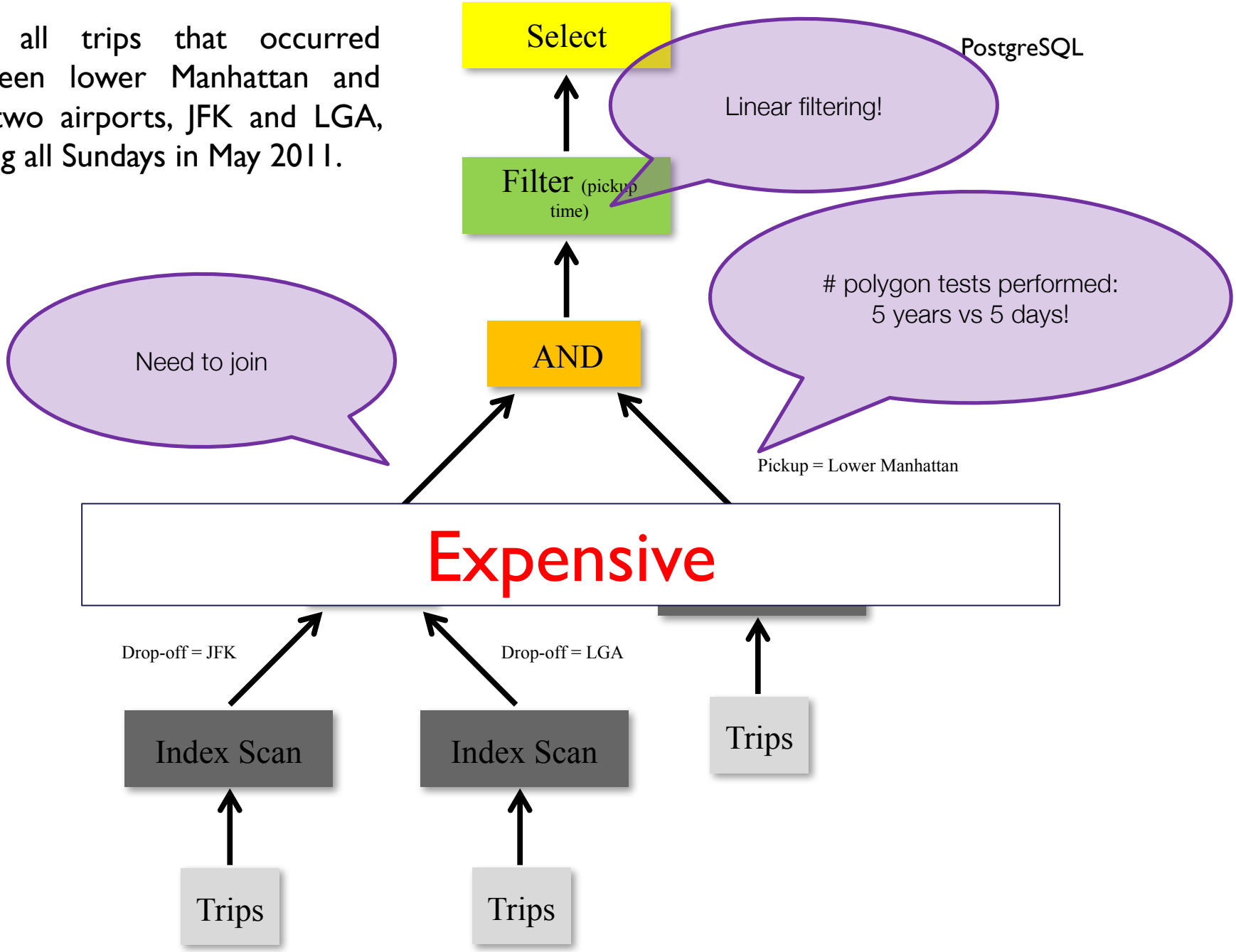



| Query time (sec) | 503.9 | 20.6 |
|---|---|---|

**Goal: Support interactive spatio-temporal queries**

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

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



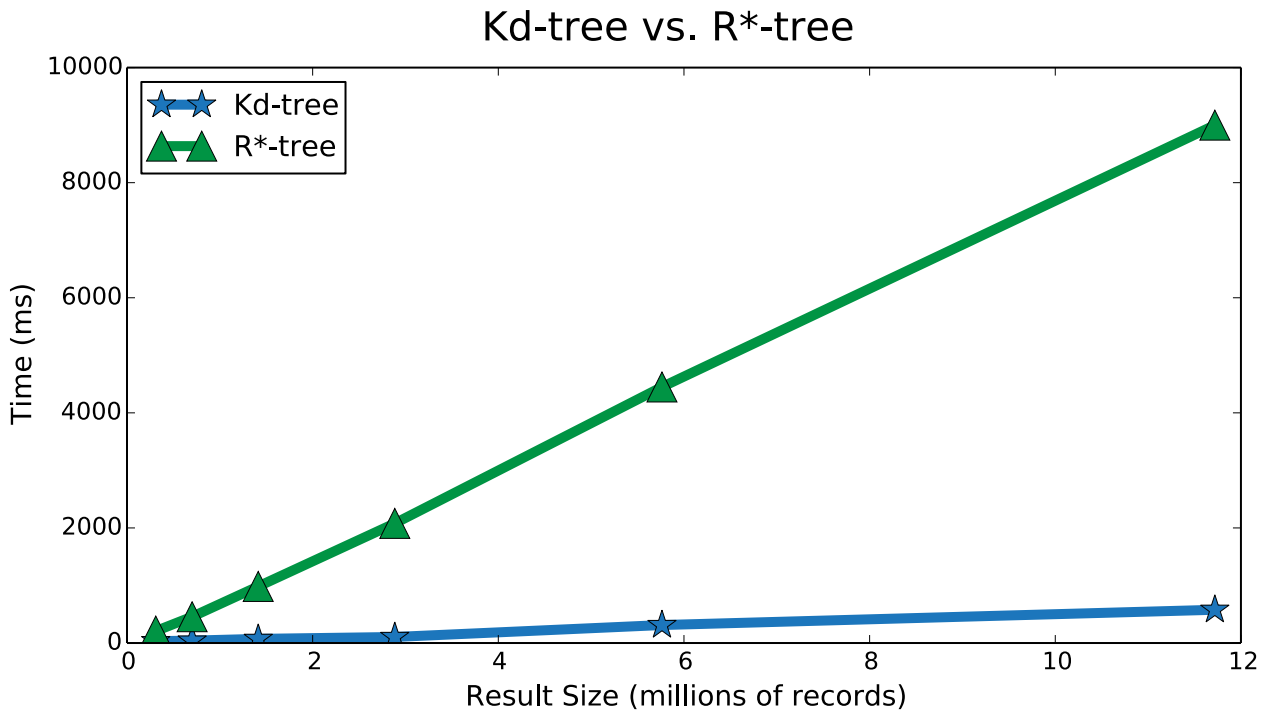
# Design Goals

- 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 |

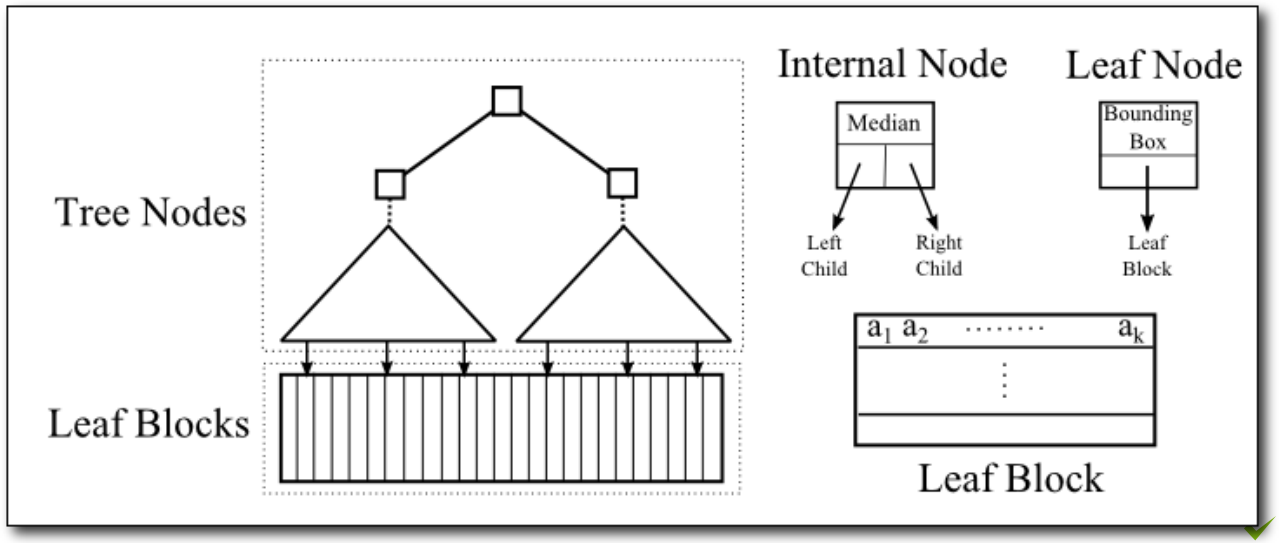
# Choice of Data Structure



# Supporting Interactive Queries

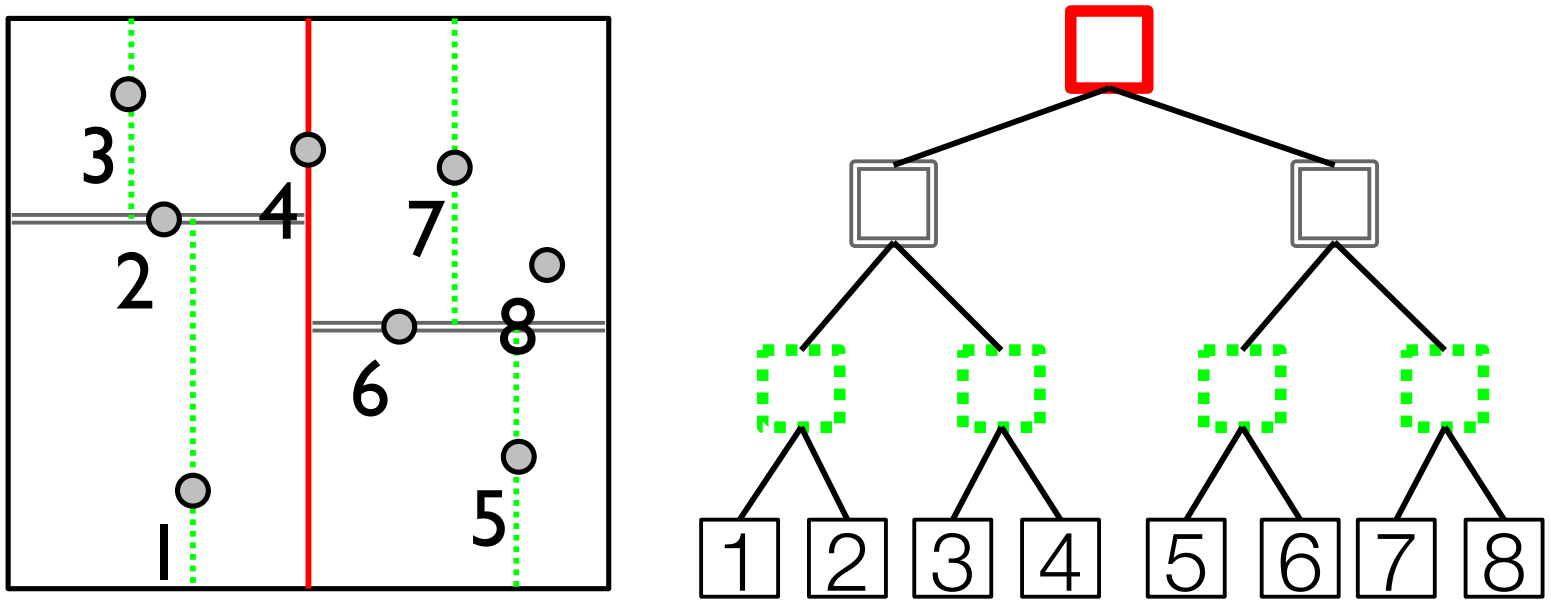
**Solution:** Spatio-temporal index based on out-of-core kd-tree using GPUs (STIG)

- 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



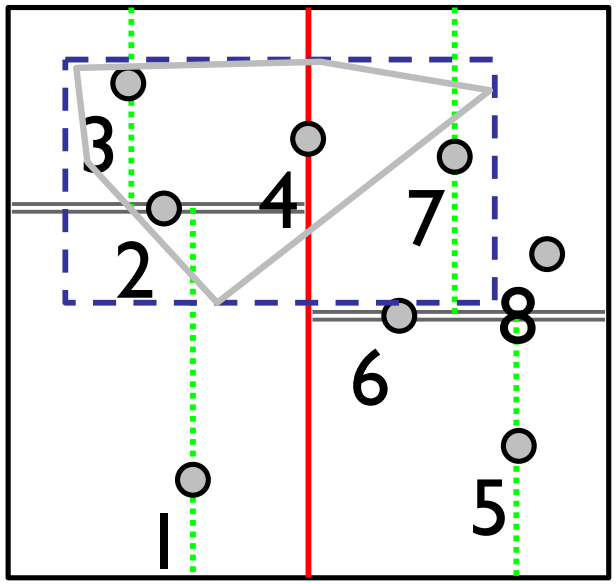

[Doraiswamy et al., ICDE 2016]

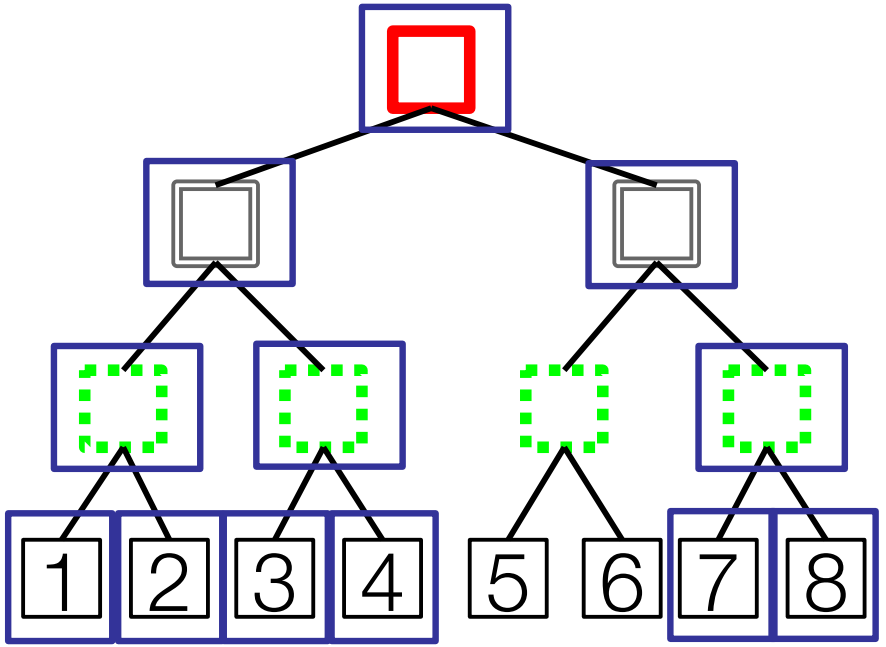
# KD-Tree



# KD-Tree

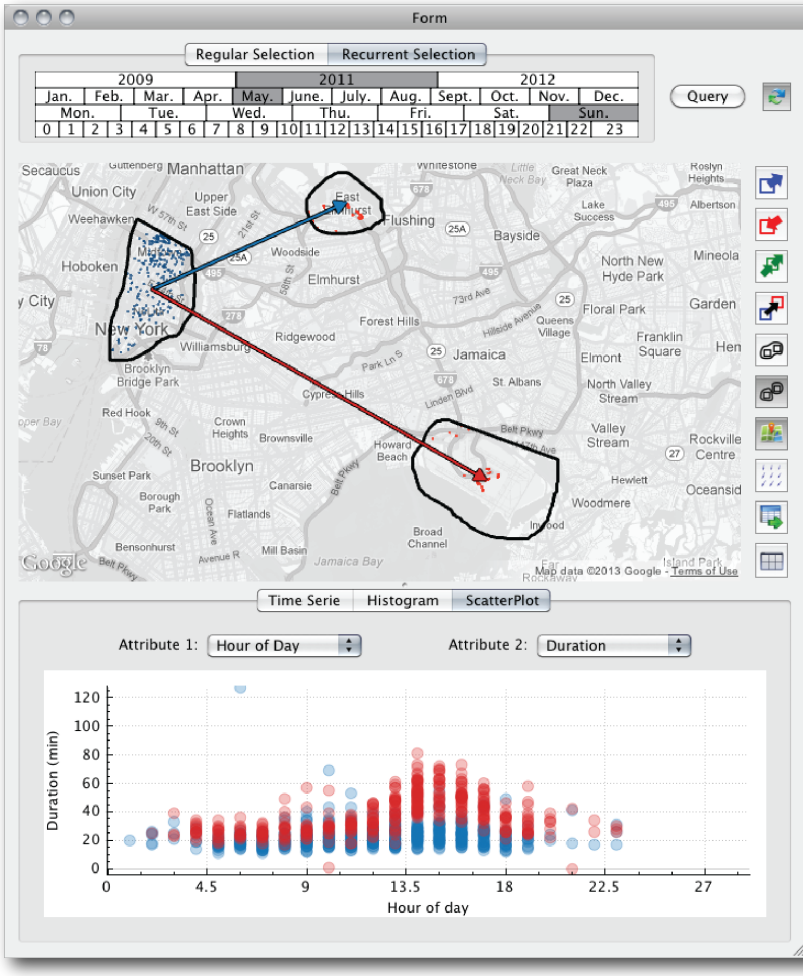
- Polygon containment query
  - Search based on Bounding Box
  - Test with query polygon



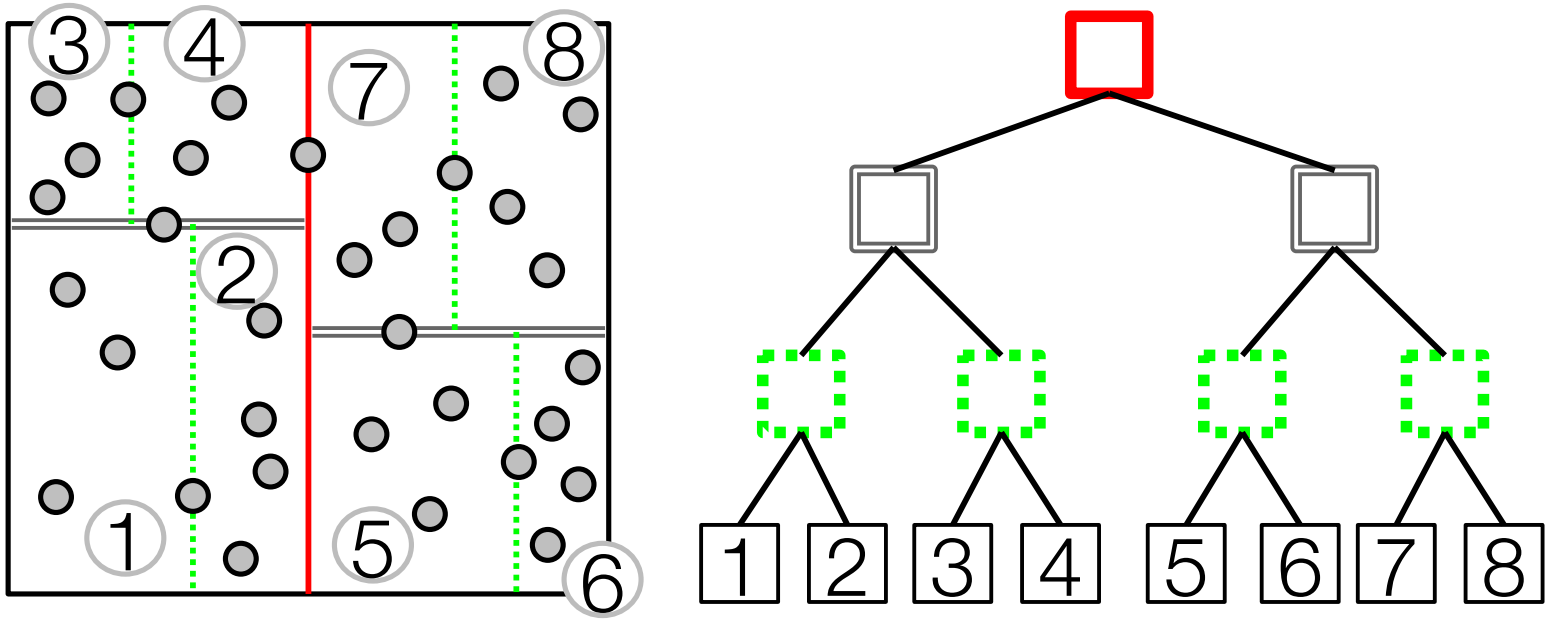



# PIP Tests are Expensive

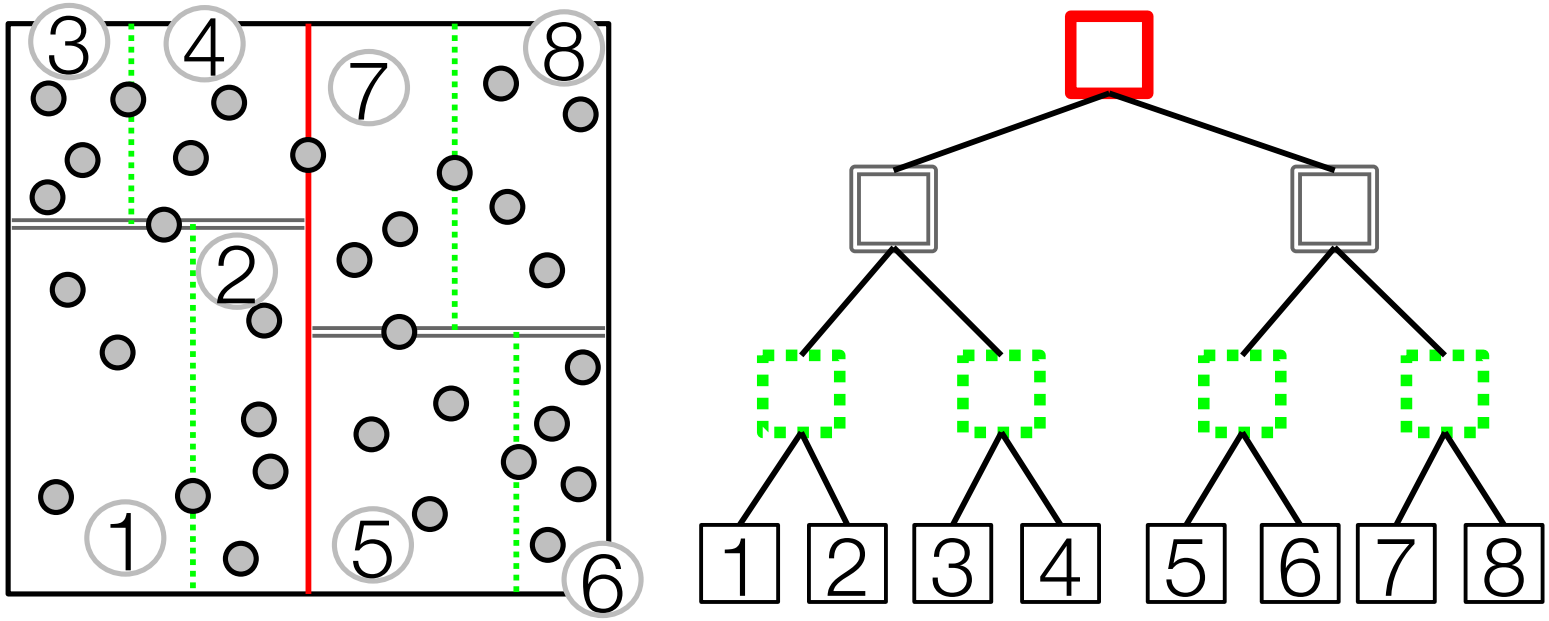
6.5 million such tests have to be performed even though the query returns only around 13,000 records



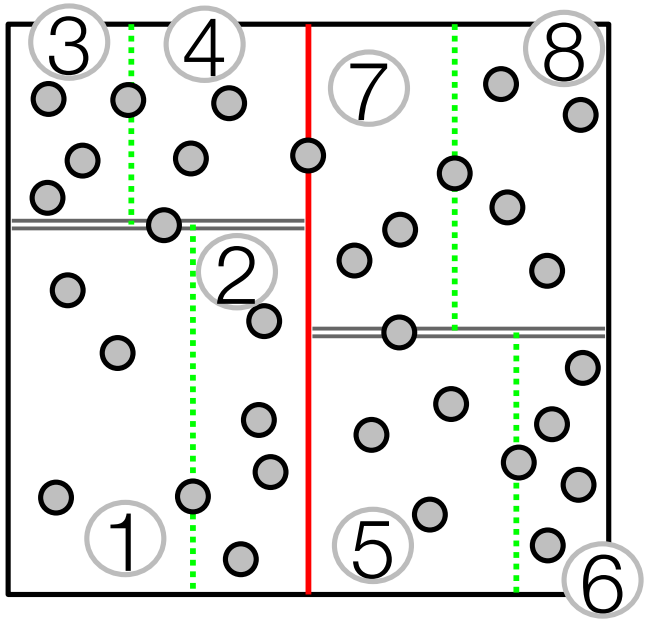
# The STG Tree

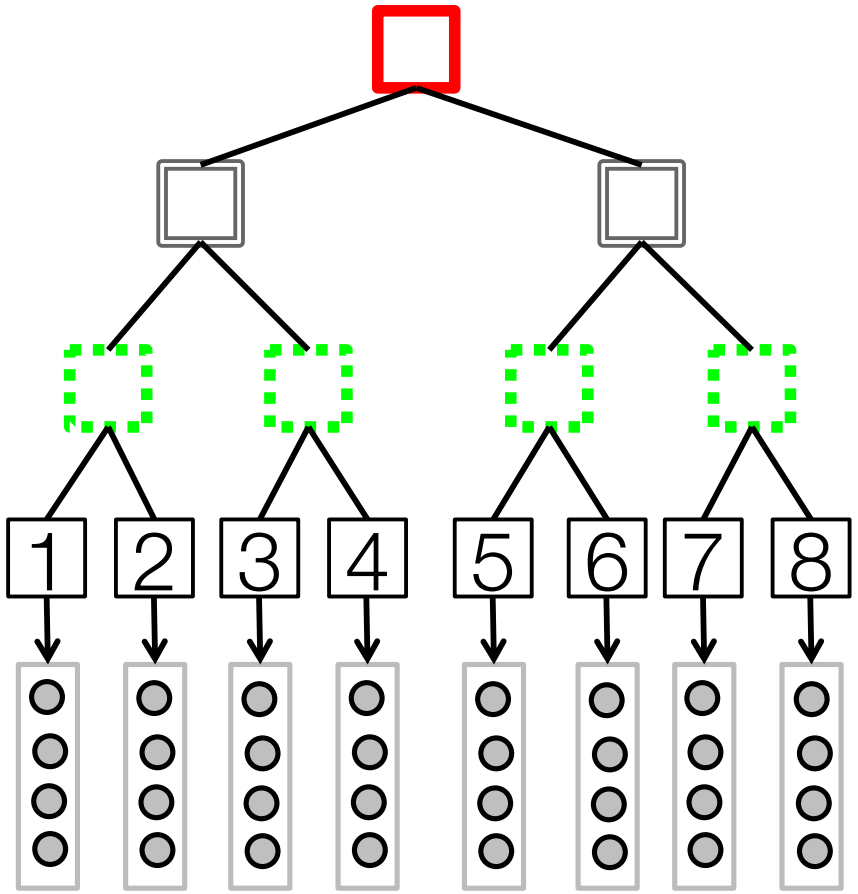


# The STG Tree

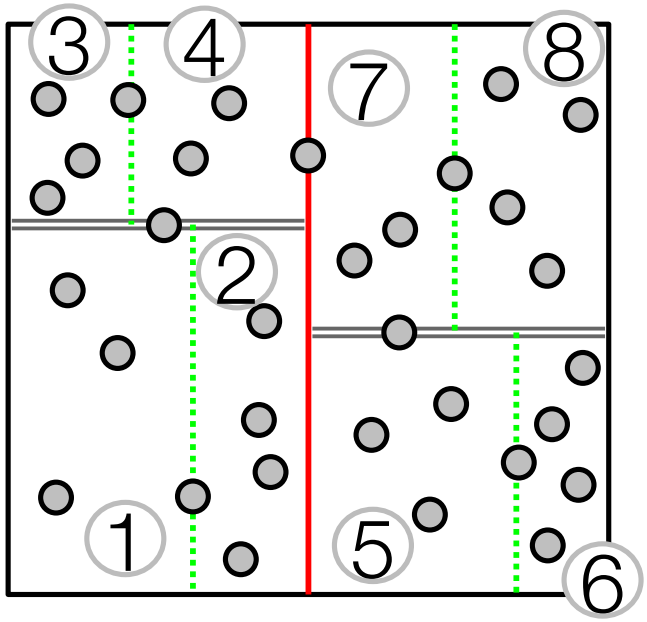


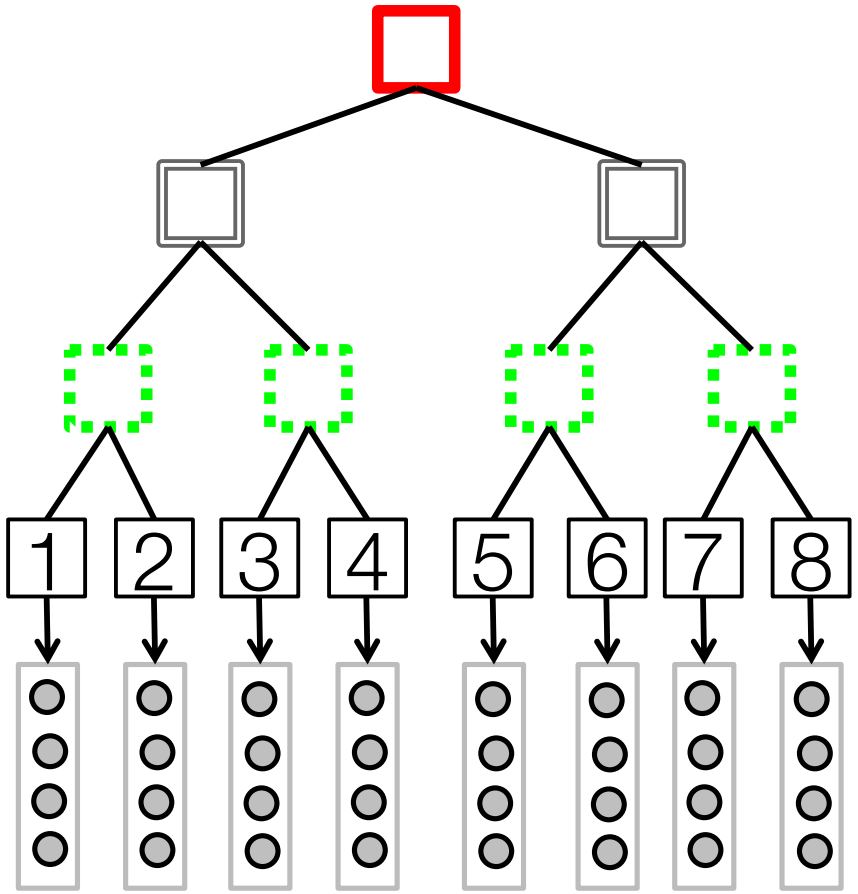
# Stg Tree



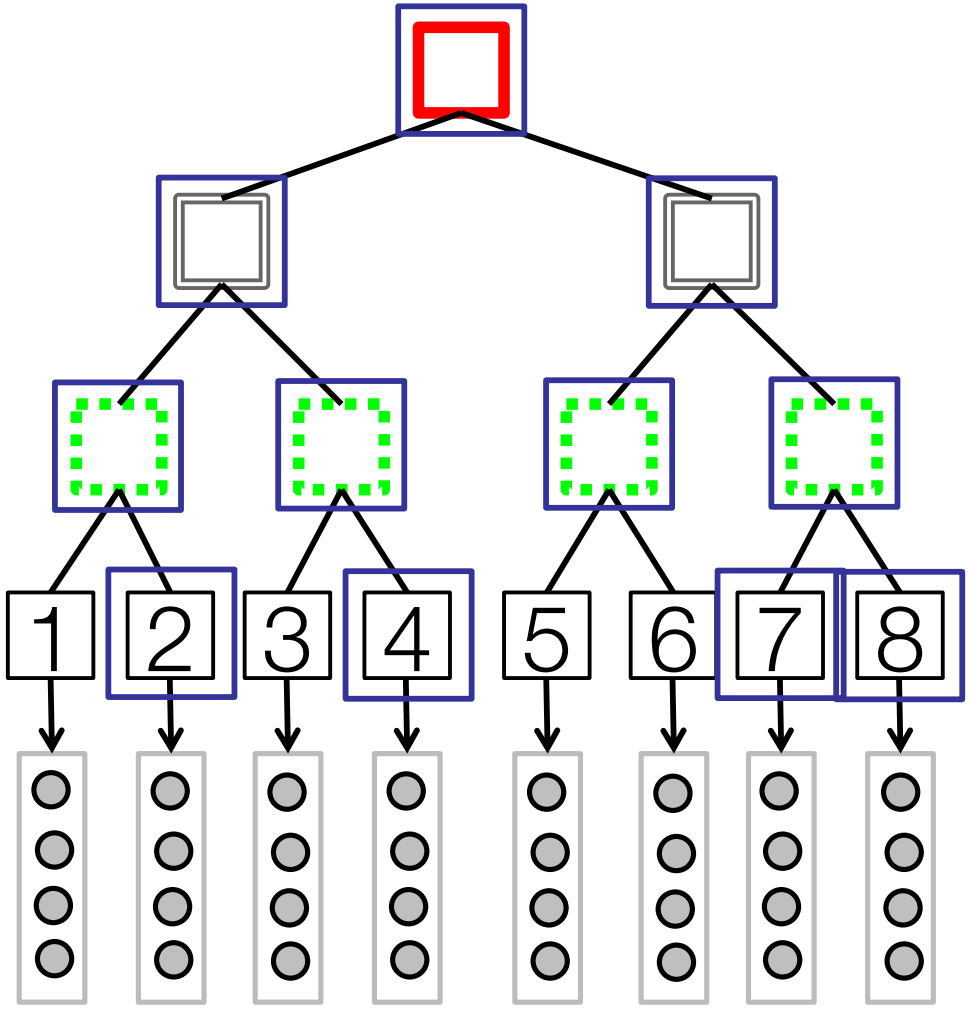
# Stg Tree



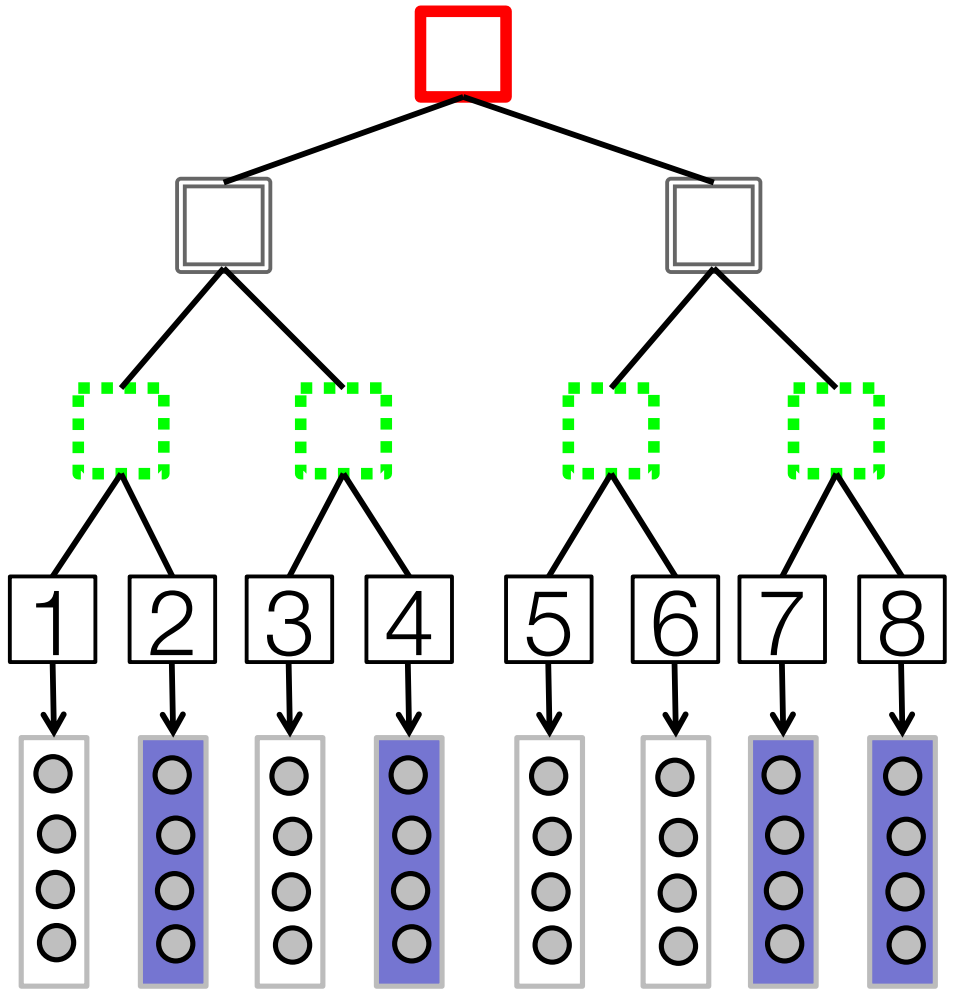
# STIG Query

- 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

# 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

# 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

# TaxiVis: Status

- Demoed to NYC DOT and TLC

---------- Forwarded message ----------
From: [redacted]@tlc.nyc.gov>
Date: Thu, Oct 24, 2013 at 4:58 PM
Subject: NYC taxi data
To: "Claudio Silva (csilva@nyu.edu)" <csilva@nyu.edu>, "Huy Vo (huy.vo@nyu.edu)" <huy.vo@nyu.edu>, "Caryn Joy Knutsen (caryn.knutsen@nyu.edu)" <caryn.knutsen@nyu.edu>, "Kim Alfred (kim.alfred@nyu.edu)" <kim.alfred@nyu.edu>
[redacted]>

Hi all,

First, I would like to thank you all for coming to data. We were truly blown away! In fact, we ha product like the one you've demonstrated to us [redacted] for us on Monday. We think that could be a gre future use for our data in combination with othe

[redacted]

Cheers,

[redacted]

*"The speed at which the tool permits us to work has saved multiple hours of staff time and has dramatically improved the unit's output and capabilities."*

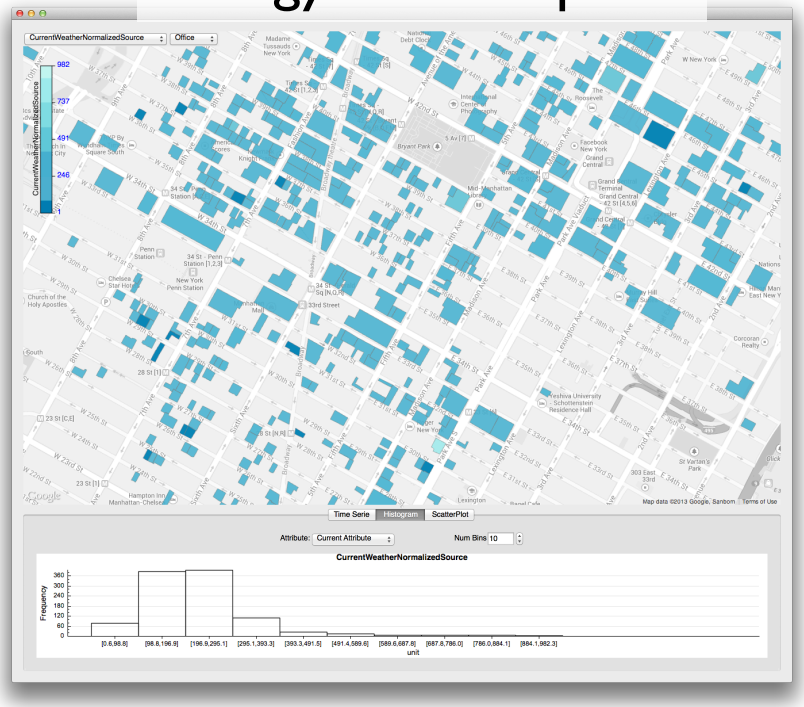
Assistant Commissioner, DoT

# TaxiVis: Status

- Demoed to NYC DOT and TLC
- System is open source
  - https://github.com/ViDA-NYU/TaxiVis
- Used to explore different data sets

## Property Ownership


## Energy Consumption




NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# What Next: Urbane

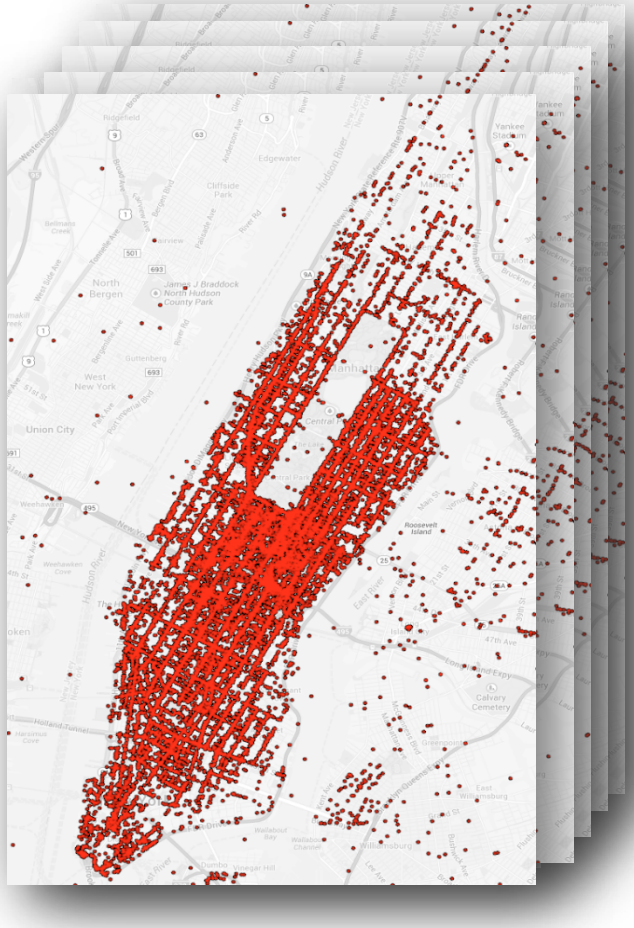

https://www.youtube.com/watch?v=_B35vxCgDw4&feature=youtu.be

[Ferreira et al., IEEE VAST 2015]

# Exploring Urban Data: Usability and Interactivity

# Finding Interesting Features

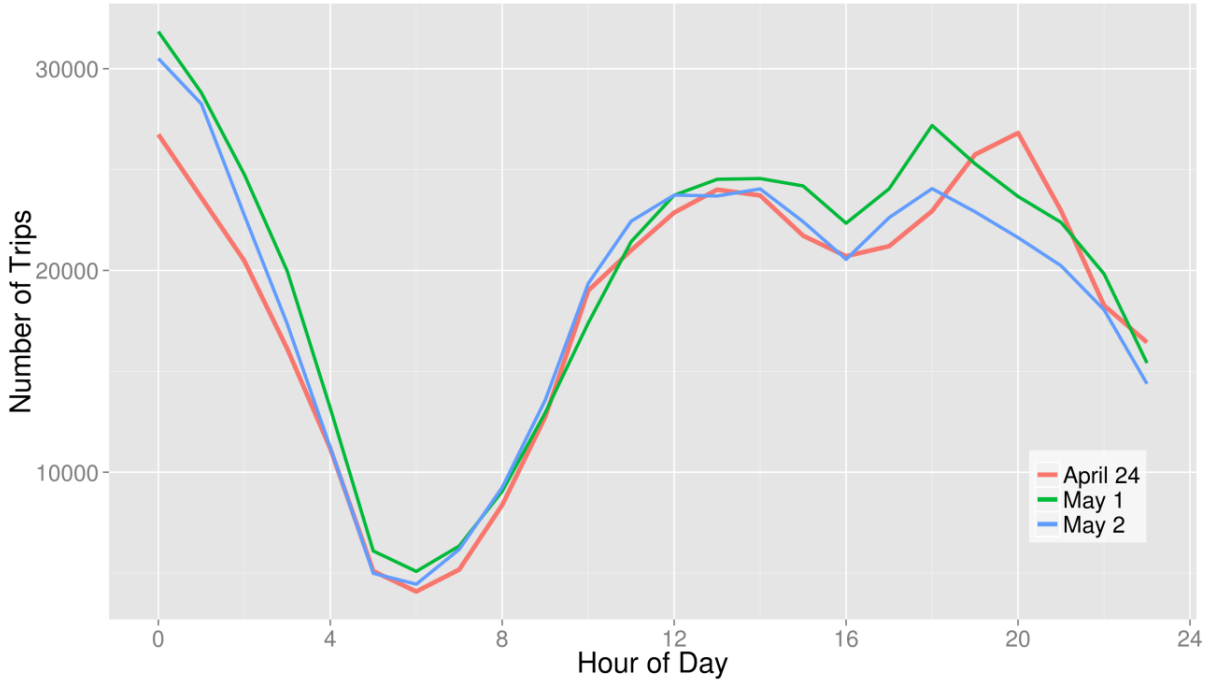
# Taxi Data: Too Many Slices



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

# Reducing the Number of Slices

*Aggregate* over space



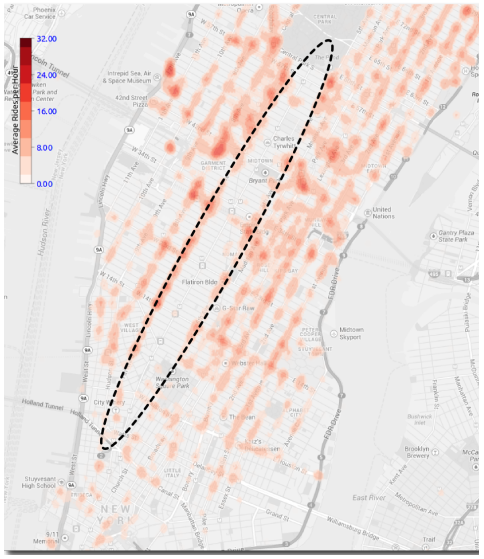

*Aggregate* over time

April 24 May 1 May 8

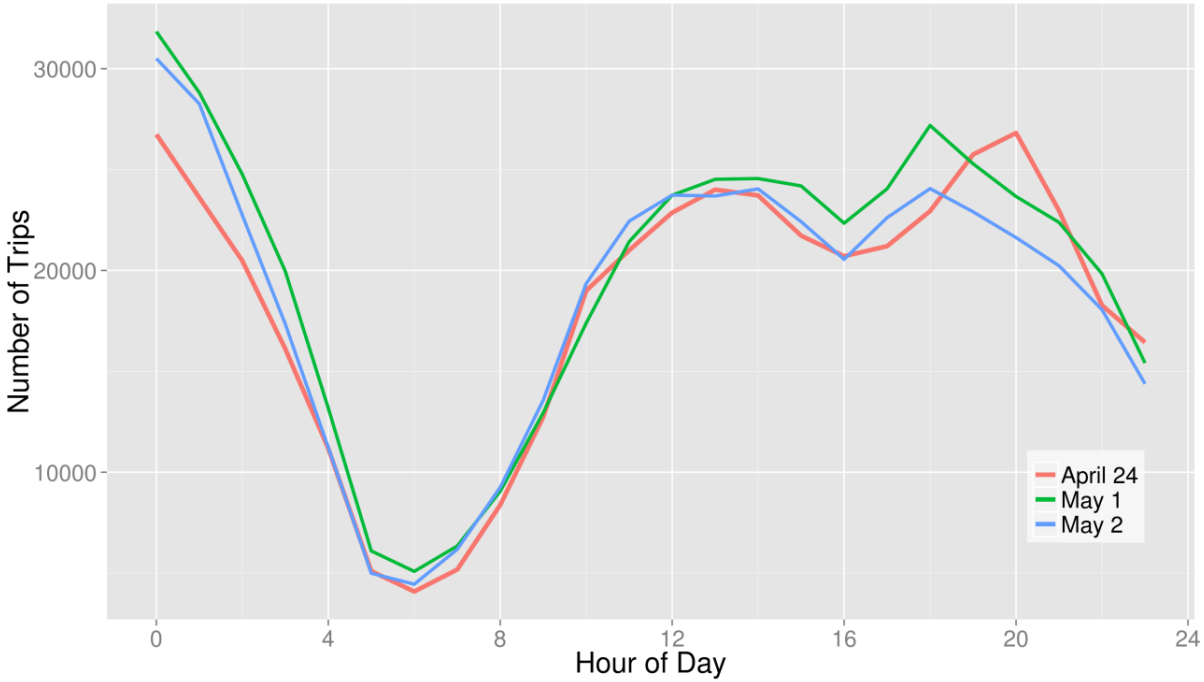
NYU

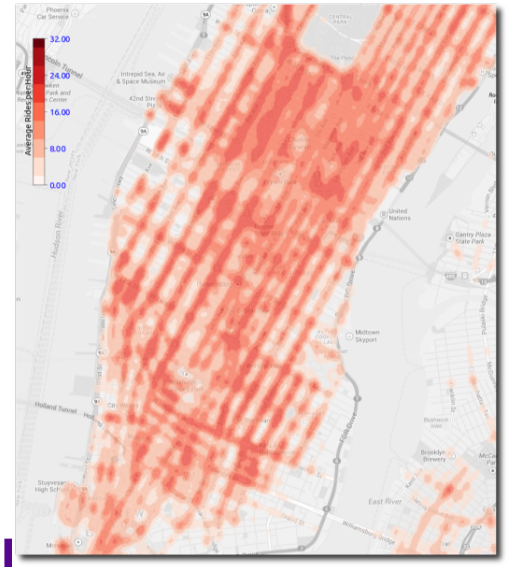
VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Miss Interesting Slices

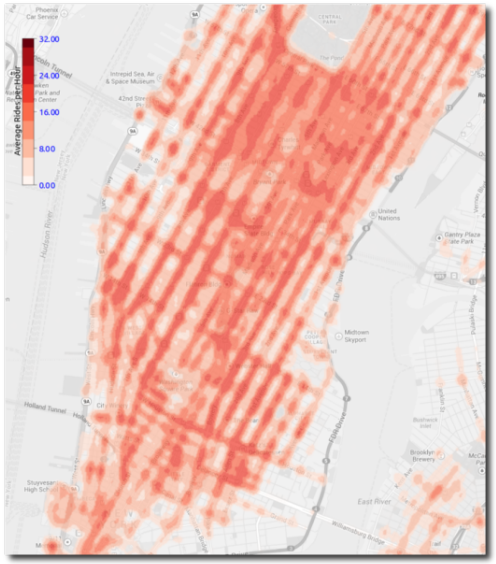


May 1 (8-9am)

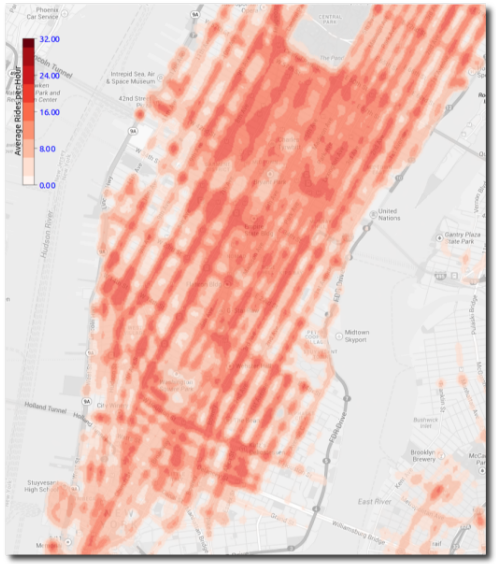


April 24



May 1



May 8

NYU

VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Finding Interesting Slices

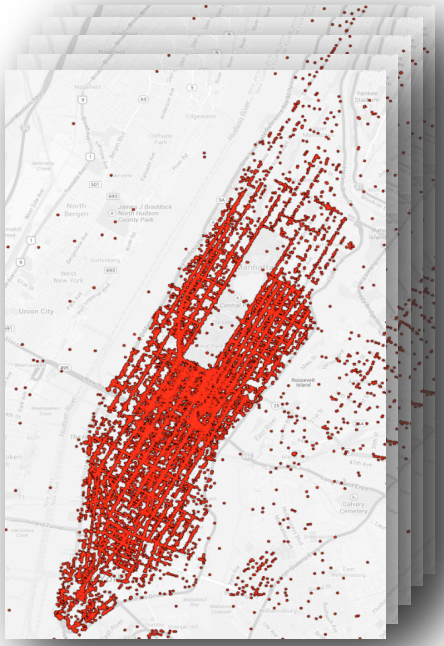
**Goal:** guide users towards *interesting* data slices

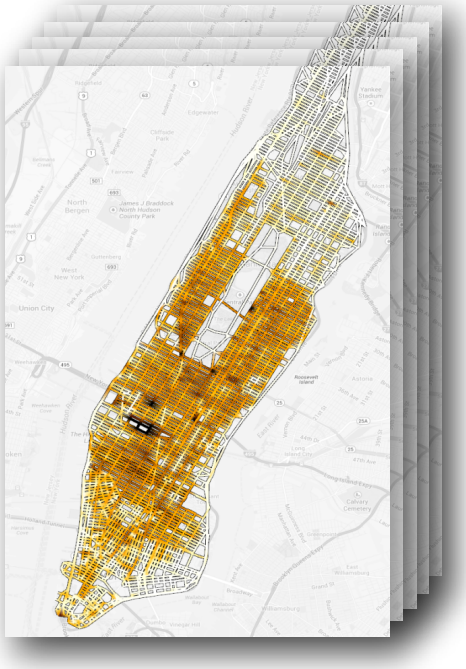
- Desiderata: automatically identify *events* with arbitrary spatial structure and at multiple temporal scales
- Our solution:
  - Use computational topology techniques to efficiently discover events
  - Simple visual interface to *explore* and *query* the events of interest

[Doraiswamy et al., IEEE TVCG 2014]

# Identifying Potential Events

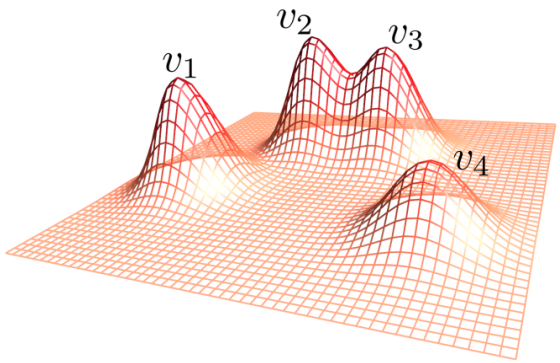
- Model data as a time-varying scalar function defined on a graph
  - $f : G \rightarrow R$
  - Taxi data: Graph = road network; Function = density of taxis
  - Subway data: Graph = track network; Function = delay of trains

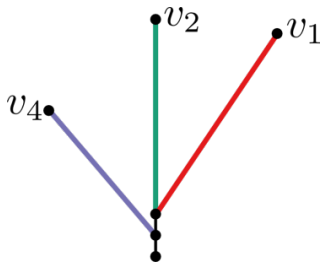


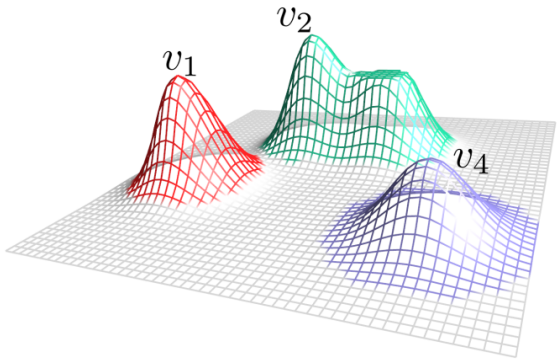


# Identifying Potential Events

- 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

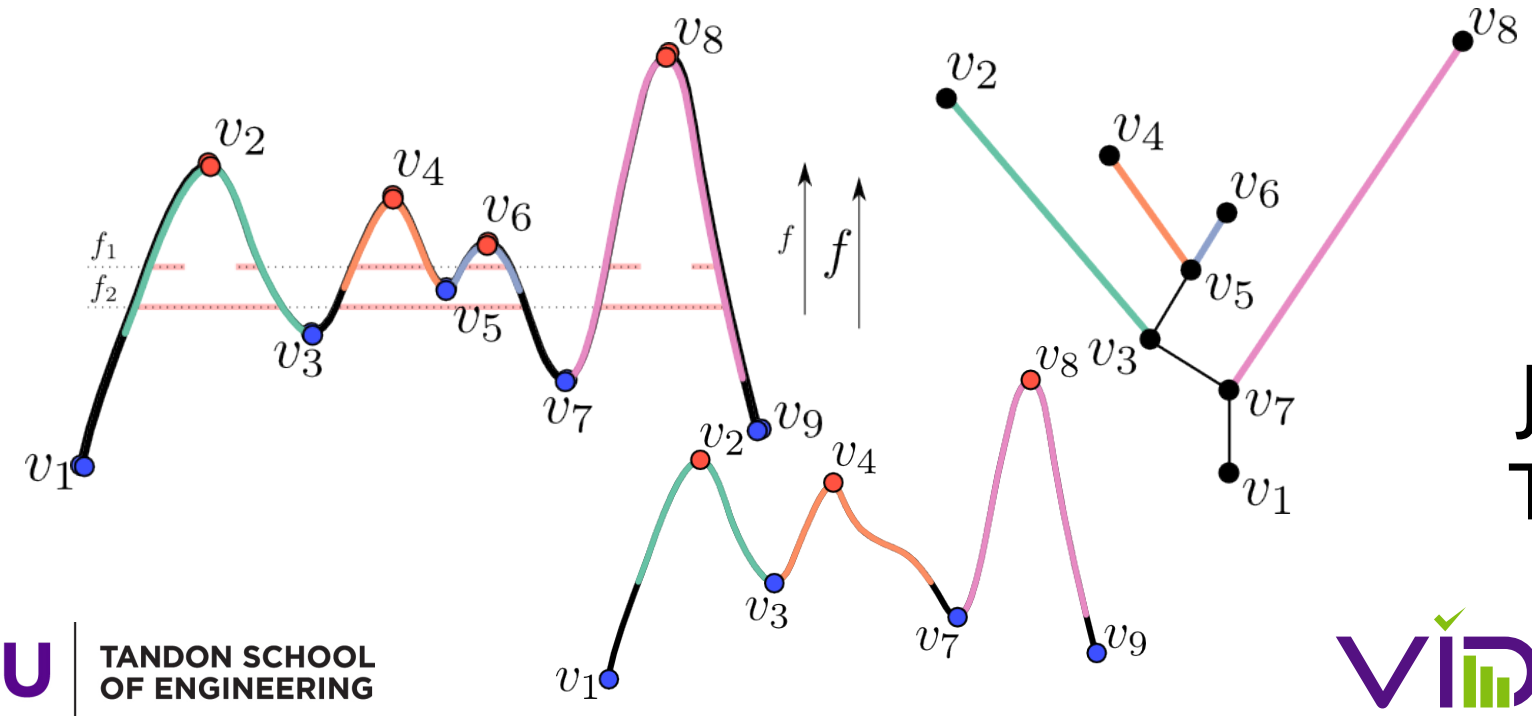


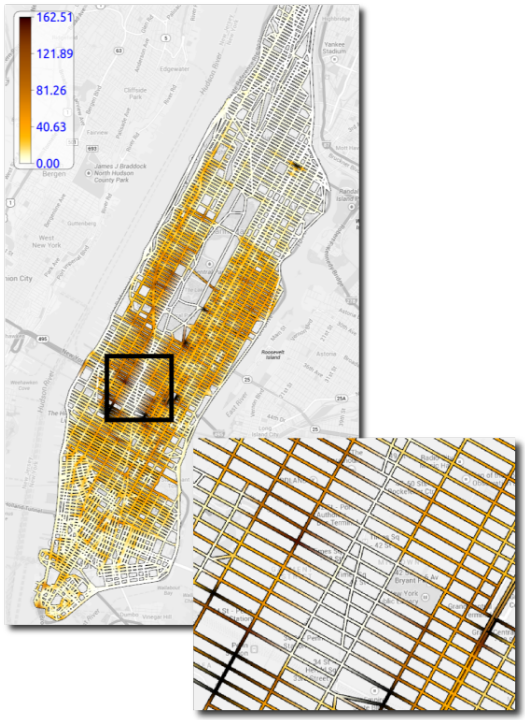
# Identifying Potential Events

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



# Taxi Data: Potential Events

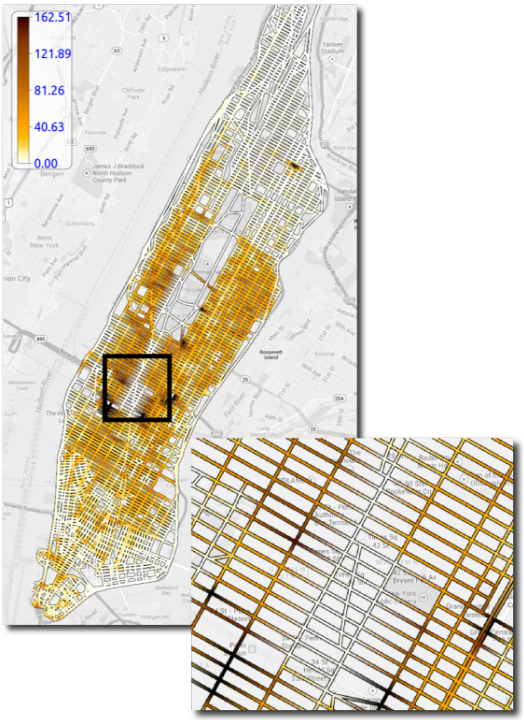
- Minima: lack of taxis
  - Regions where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade

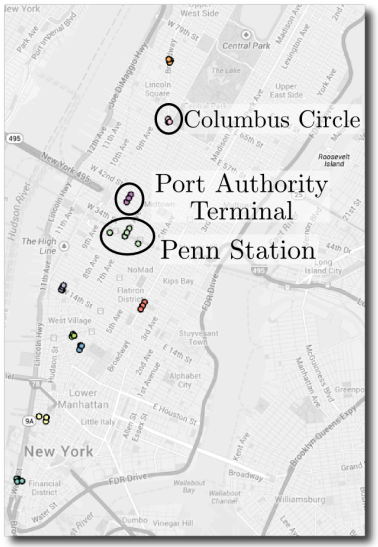



*Scalar function corresponding to the time step 10 am-11 am on 24 November 2011*

# Taxi Data: Potential Events

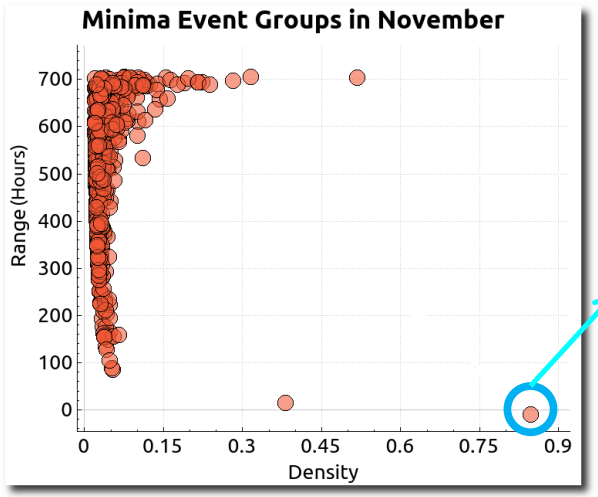
- Minima: lack of taxis
  - Regions where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade
- Maxima: popular taxi locations
  - Regions where density is higher than local neighborhood
  - Could denote tourist locations, train stations




# 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




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)

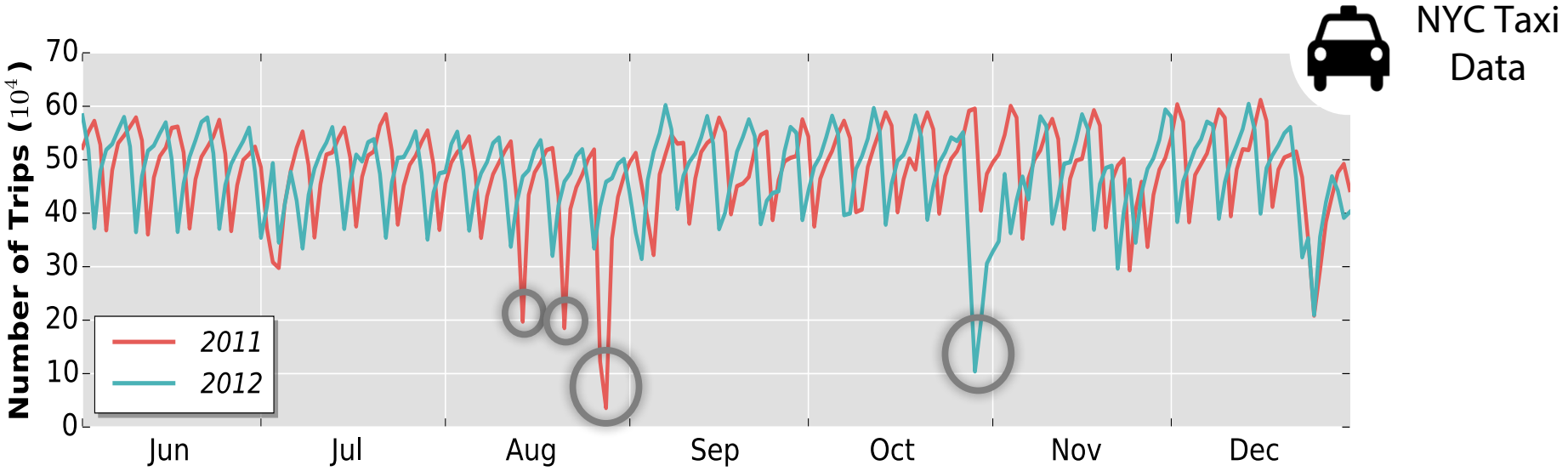
Gaza Solidarity Protest NYC (18 November 2012)

NYU TANDON SCHOOL OF ENGINEERING

CENTER

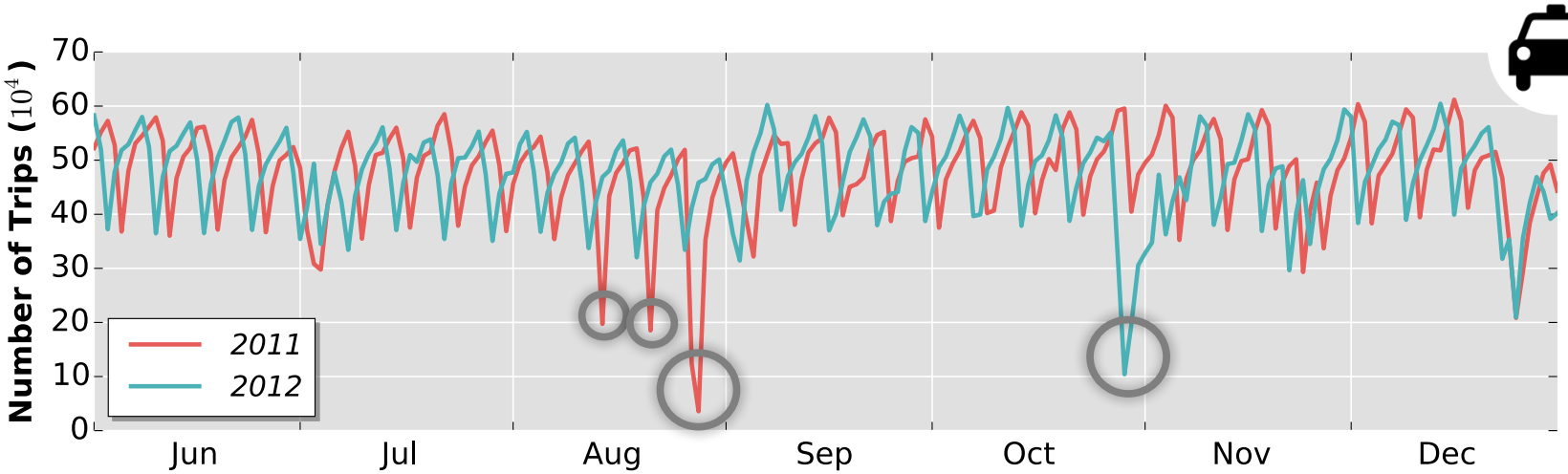
# Using Data to Explain Data

# Explaining Events




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

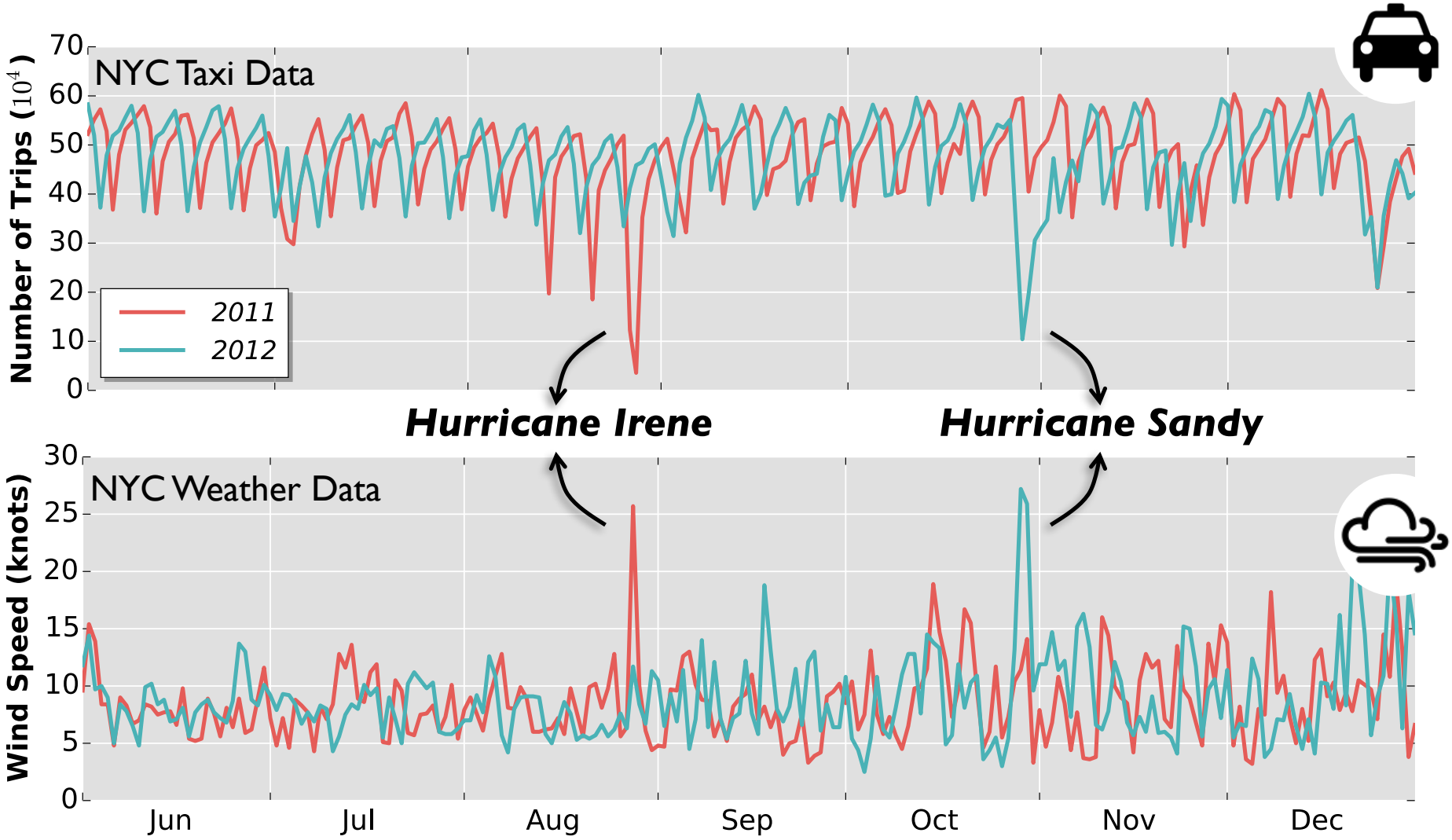
# Explaining Events




- 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

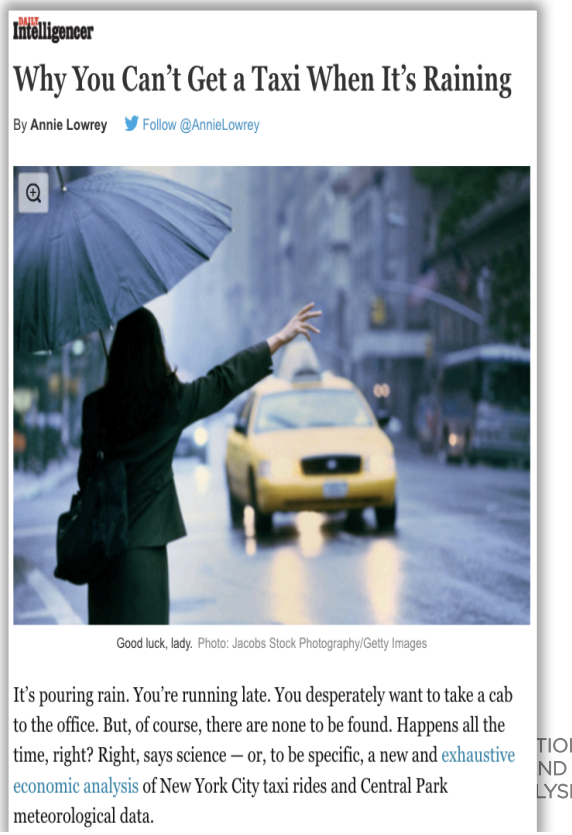
# Using Data to Explain Events



# 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*



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!*

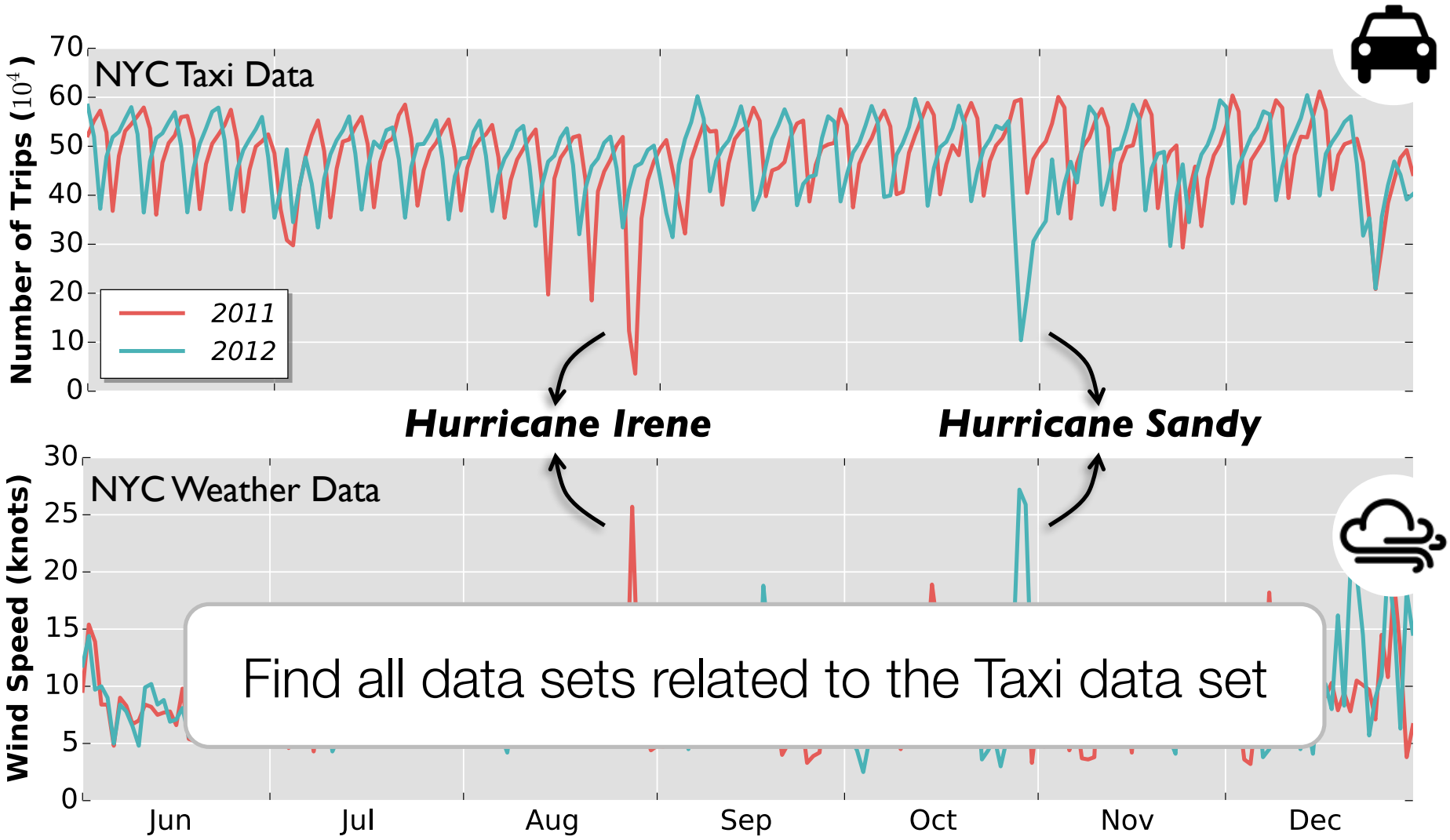
- Guide users in **data discovery and analysis** by allowing them to pose ***relationship queries***

*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



# Hypothesis Testing

- Hypothesis: Taxi drivers set an income goal, and on rainy days they reach the goal faster

*Taxi Fare*

x

*Precipitation*

Find relationships between Weather and Taxi data sets

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.

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

# Challenges

- Must take both space and time into account
- How and when are two data sets related?
  - Some relationships become visible through *atypical* behavior
- Conventional techniques (e.g., Pearson's correlation, mutual information, DTW, etc.) take into account the entire data and miss relationships that occur only at certain times/locations
  - E.g., most of the time, taxi trips and wind speed are not related

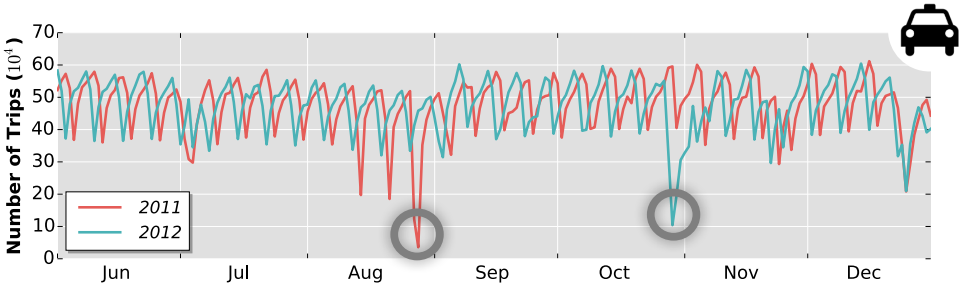
# Challenges

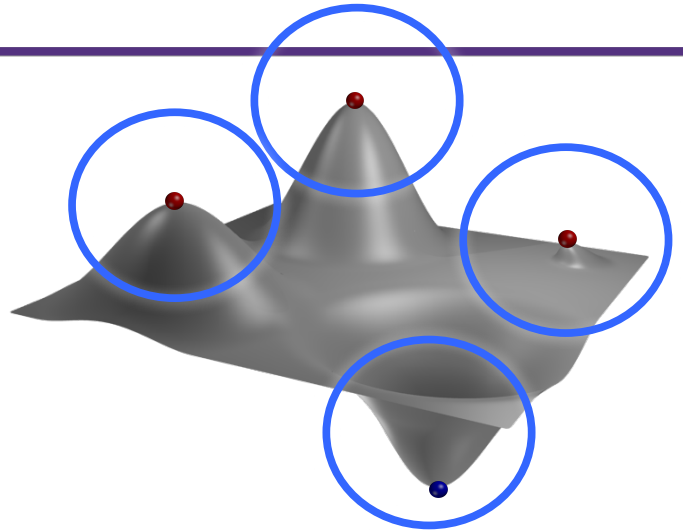
- Many data sets, each consisting of many attributes
  - Relationships can be between any of the attributes
  - Weather data: >200 attributes; NYC Open data: 8 attributes per data set on an average
- Data sets can be large
  - Taxi data: 180M trips per year
- Data at multiple spatio-temporal different resolutions
- Combinatorially large number of relationships to evaluate
  - ~2.4 million possible relationships among NYC Open Data alone for a single spatio-temporal resolution

meaningful relationship ⟷ needle in a haystack

# How and when are two data sets related?

- Topology-based relationships



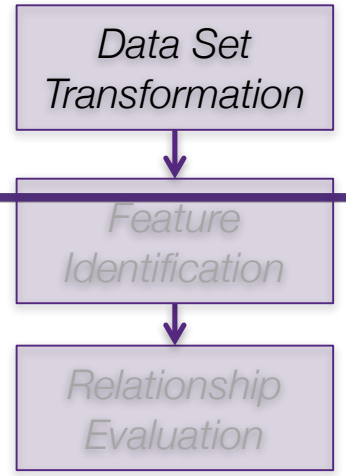
Peaks
Valleys
*Critical Poin*

- Topological representation captures salient features of the data

> A salient feature is a spatio-temporal region whose behavior differs from its neighborhood

- Supports arbitrary spatial structures and time intervals
- Efficient feature identification through Merge Tree Index
- ***Definition*:** Two data sets are related if their **salient features** are related

# Data Set to Scalar Functions

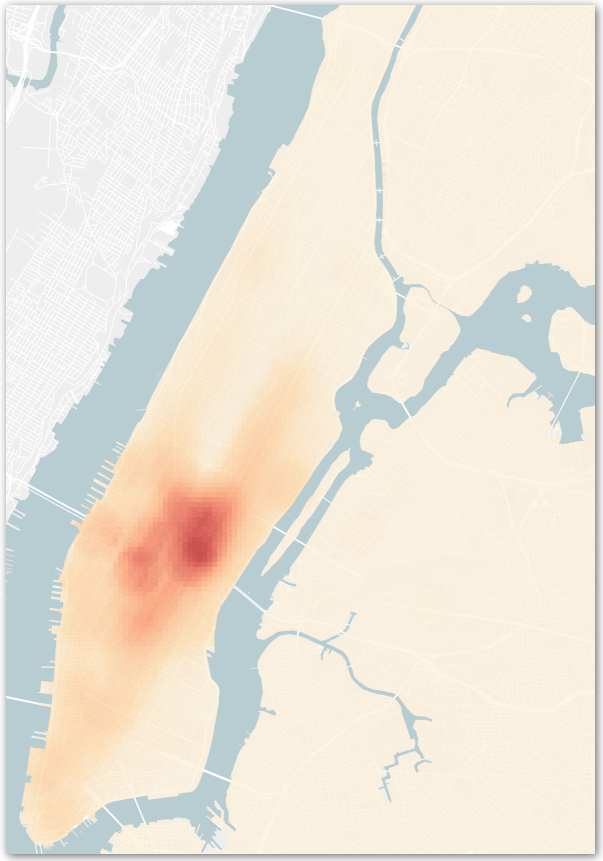



- Each data set represented as a set of time-varying scalar functions
  - $f : [\mathbb{S} \times \mathbb{T}] \rightarrow \mathbb{R}$
  - Maps each point in space and time to a real value
- Different functions can be used
  - Count: Captures the activity of an entity corresponding to the data, e.g., number of trips, number of unique taxis, etc.
  - Attribute: Captures attribute value variation, e.g., average taxi fare
- Functions computed at all possible resolutions

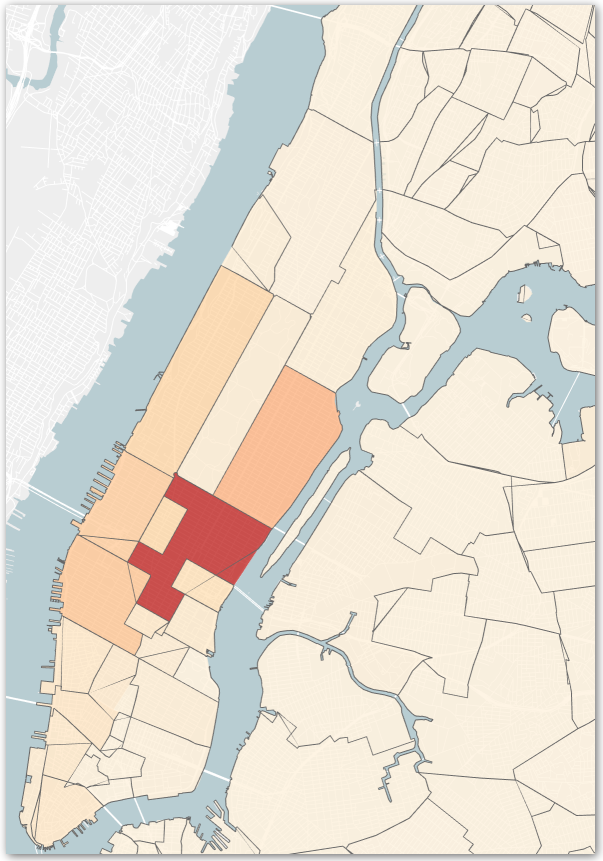
# Data Set to Scalar Functions



- Example: Density function of Taxi data

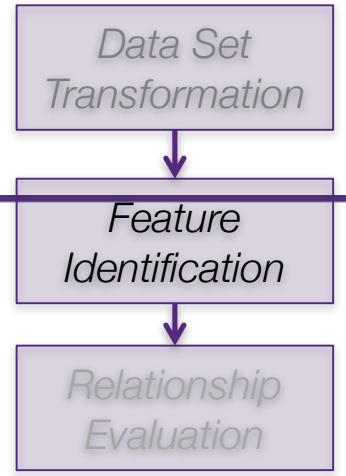


$\mathbb{S}$ : *High Resolution Grid*



$\mathbb{S}$ : *Neighborhood Resolution*

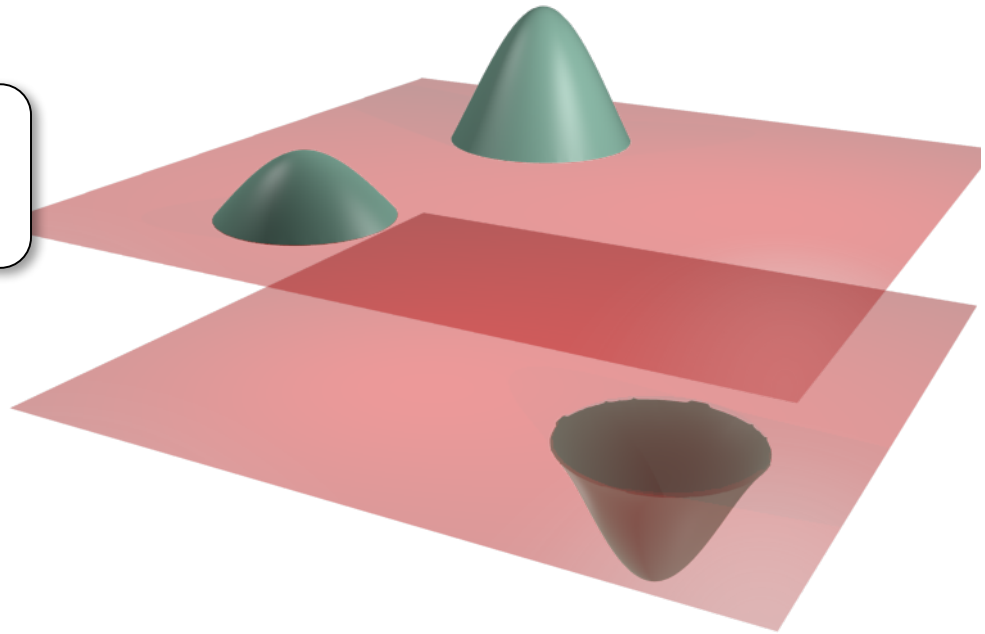
# Identify Salient Features



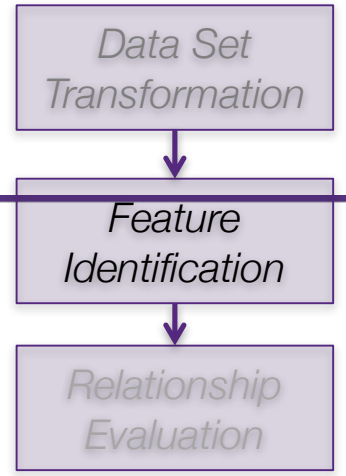

- Topological features of *a scalar function*
  - Neighborhood of critical points
  - Features represented as  spatio-temporal points
  - Positive and negative features

*Advantage*

1. Naturally captures *salient* features



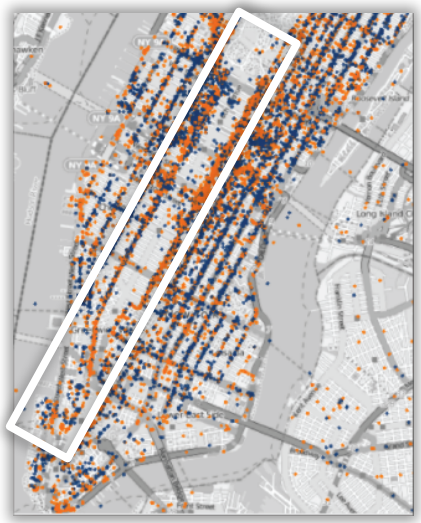
# Identify Salient Features




- Topological features of *a scalar function*
  - Neighborhood of critical points
  - Features represented as  spatio-temporal points
  - Positive and negative features

8am - 9am
May 1 2011

5 Boro Bike Tour

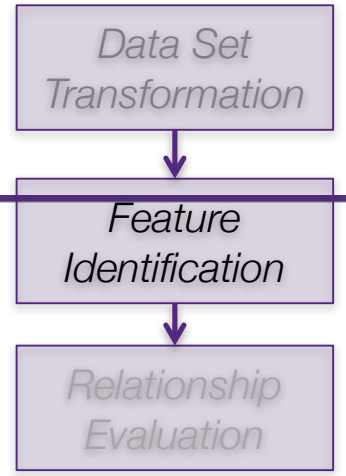


*Negative Feature*

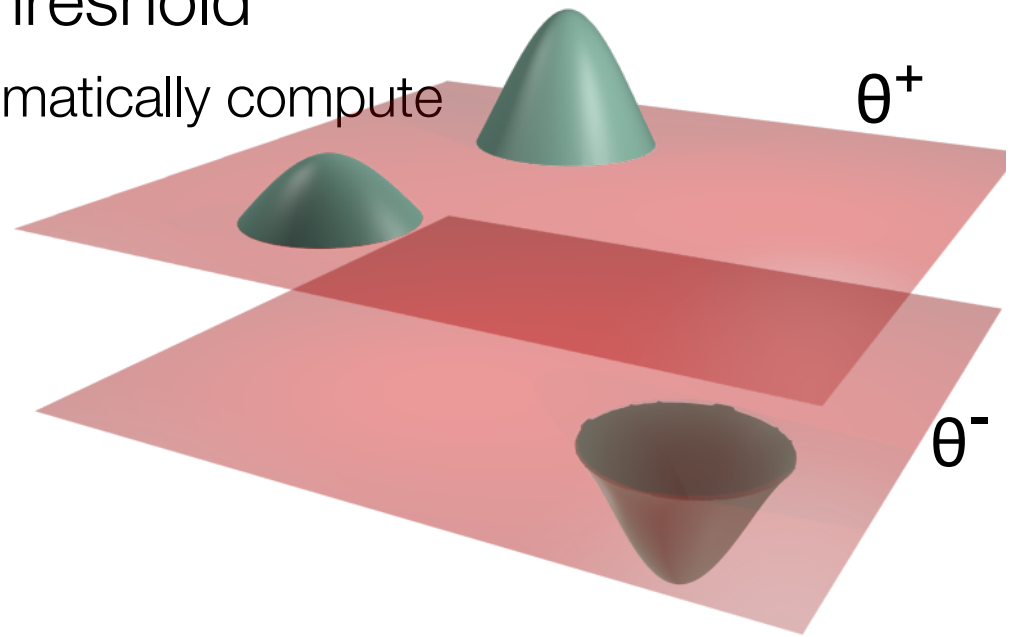
*Advantage*

2. Features can have arbitrary shapes

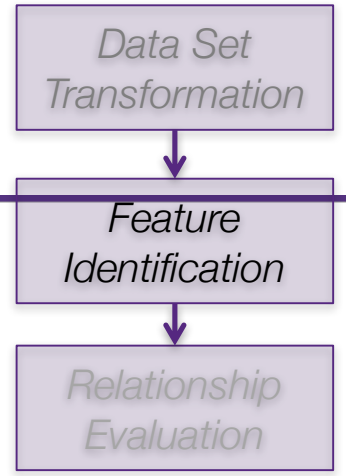
# Identify Salient Features




- Topological features of *a scalar function*
  - Neighborhood of critical points
  - Features represented as spatio-temporal points
  - Positive and negative features
- Neighborhood defined by a threshold
  - Use *topological persistence* to automatically compute thresholds in a data-driven fashion



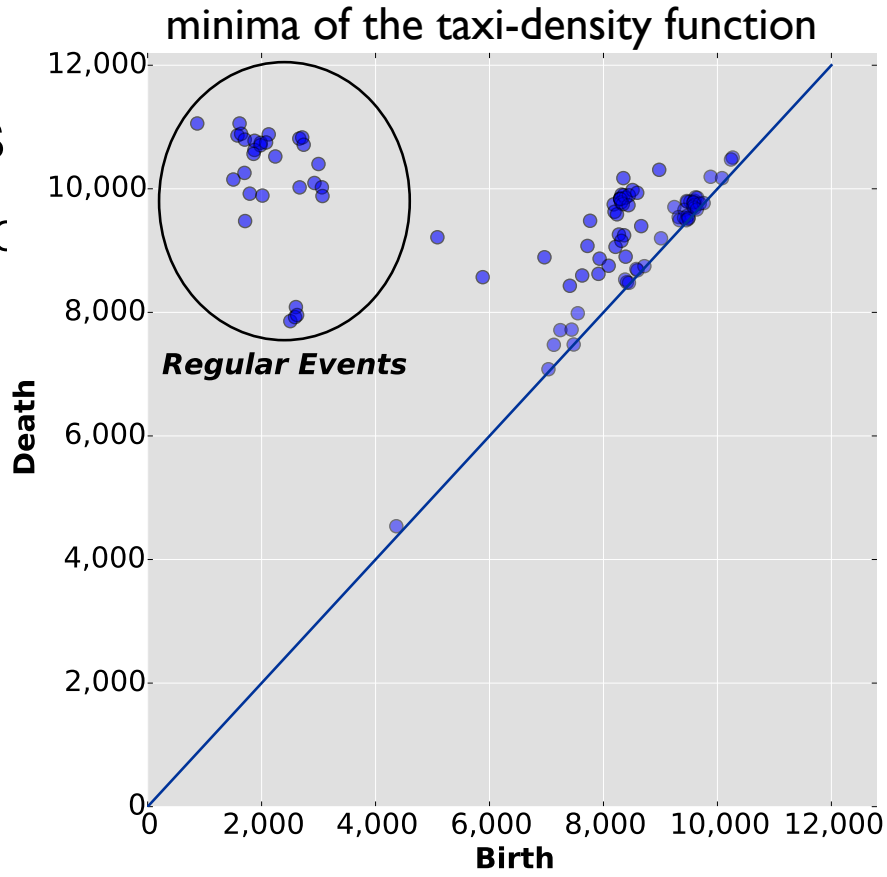
# Identify Salient Features




- Topological features of *a scalar function*
  - Neighborhood of critical points
  - Features represented as spatio-temporal points
  - Positive and negative features
- Neighborhood defined by a thres
  - Use *topological persistence* to automatic thresholds in a data-driven fashion

*Advantage*

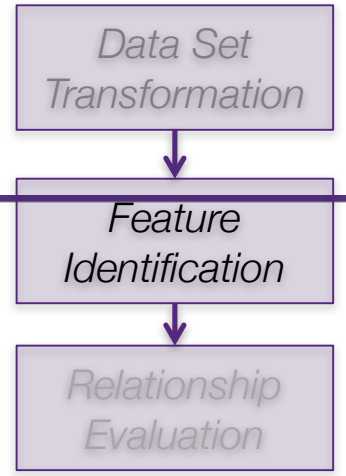
3. Data driven and robust



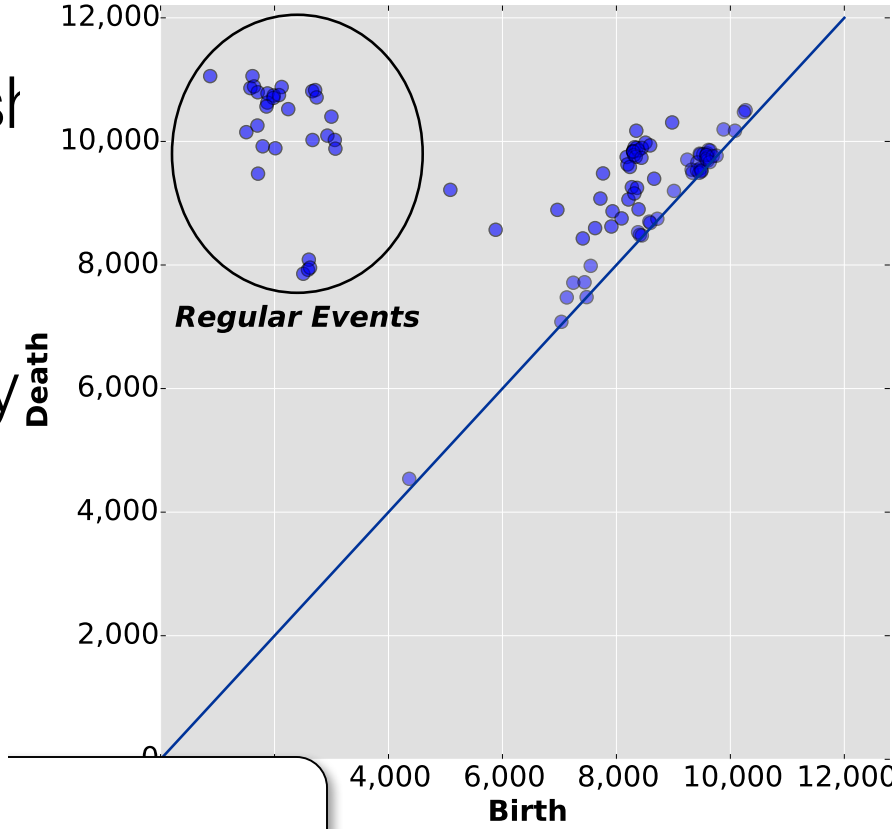

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# Identify Salient Features



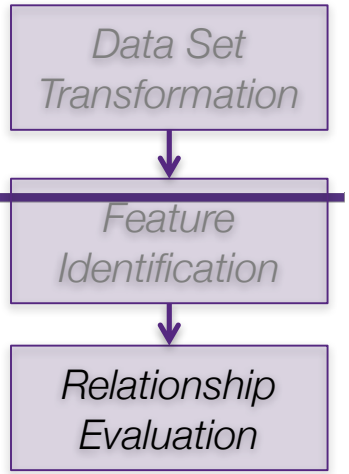

- Topological features of *a scalar function*
  - Neighborhood of critical points
  - Features represented as spatio-temporal points
  - Positive and negative features
- Neighborhood defined by a thresh
  - *Topological persistence used* to automatically compute thresholds
- *Merge Tree Index* used to identify features at all resolution
  - *O(n log n)* to construct
  - Computing features is output sensitive




*Advantage*

4. Very efficient

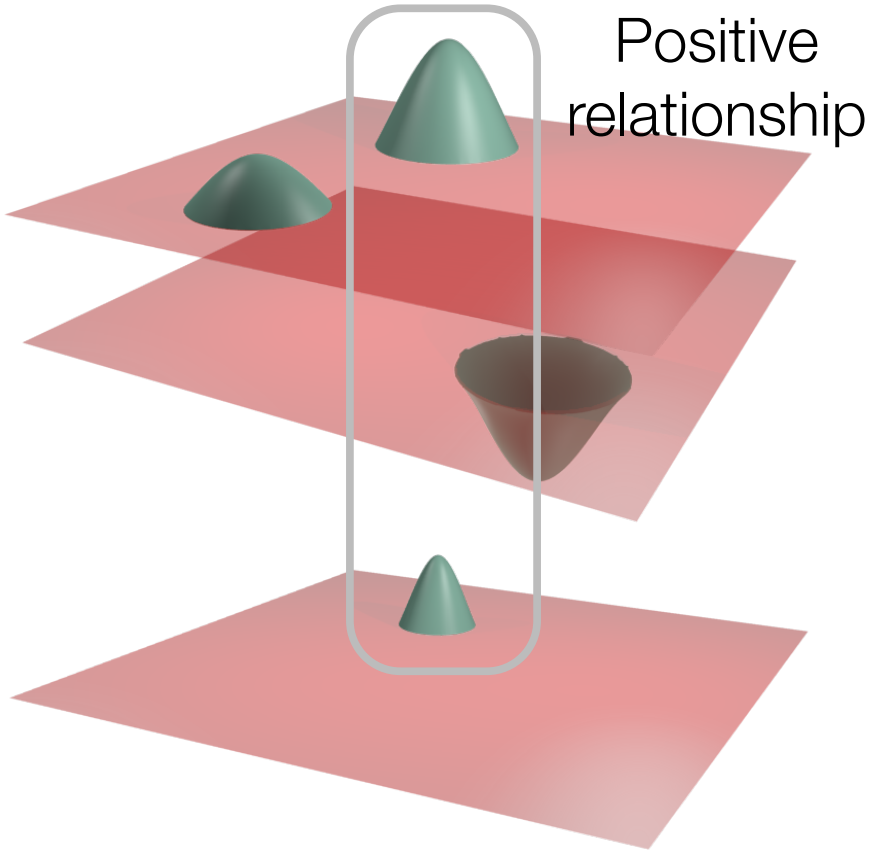
# Evaluating Relationships




- Relationship between functions *f* and *g* consists of the set of spatio-temporal points that are features in both functions
- Let $\Sigma_1$ and $\Sigma_2$ be the set of features of $f_1$ and $f_2$. $f_1$ and $f_2$ are *feature-related* at a spatio-temporal point $x = (s,t)$ if $x \in \Sigma_1 \wedge \Sigma_2$
  - E.g., for Hurricane Sandy, there is a negative feature in the taxi density function and a positive feature in the wind speed function
- *Relationship Score:* Captures the nature of the relationship – positive or negative

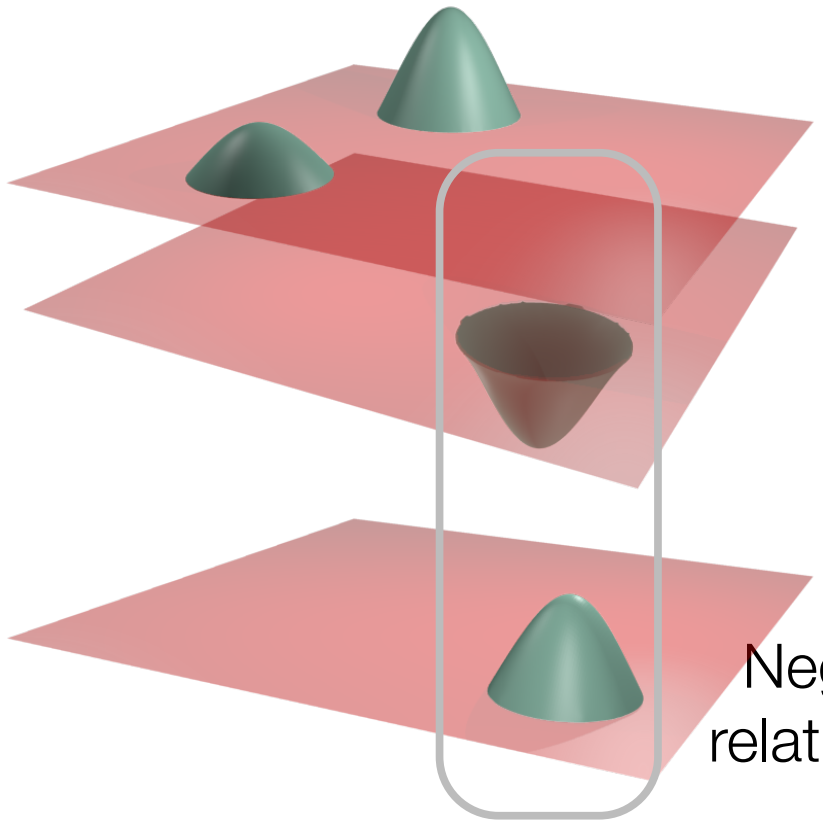
# Identifying Relationships





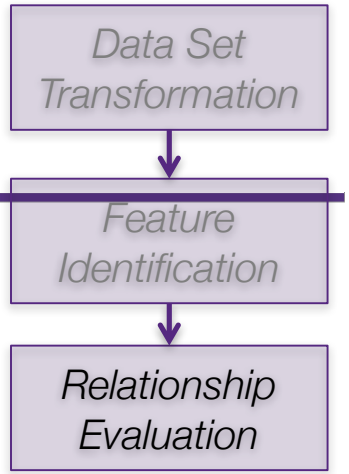
# Identifying Relationships





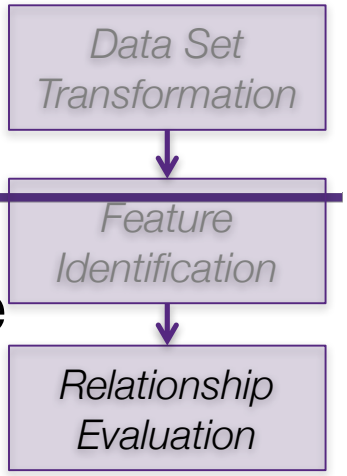
Negative relationship

# Evaluating Relationships




- Relationship between functions *f* and *g* consists of the set of spatio-temporal points that are features in both functions
- Let $\Sigma_1$ and $\Sigma_2$ be the set of features of $f_1$ and $f_2$. $f_1$ and $f_2$ are *feature-related* at a spatio-temporal point $x = (s,t)$ if $x \in \Sigma_1 \wedge \Sigma_2$
  - E.g., for Hurricane Sandy, there is a negative feature in the taxi density function and a positive feature in the wind speed function
- *Relationship Score:* Captures the nature of the relationship – positive or negative
- *Relationship Strength:* How often the functions are related – strong or weak
- Monte Carlo procedure to test the statistical significance accounting for the spatial and temporal proximity
  - Prune potentially coincidental relationships

# Evaluating Relationships




- Monte Carlo procedure to test the statistical significance accounting for the spatial and temporal proximity
  - Prune potentially coincidental relationships

Relationship between functions *f* and *g* , with score $\boldsymbol{\tau}^*$

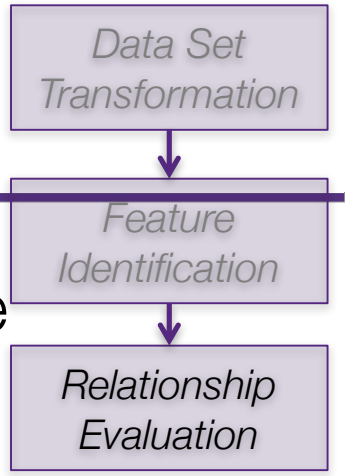
*H0*: The two functions are not related
*H1*: The two functions are related

$$p = \frac{\sum_i^N \mathrm{I}(\tau_i \geq \tau^*)}{N}$$

*N* permutations

Reject the null hypothesis if $p \leq \alpha$

# Evaluating Relationships

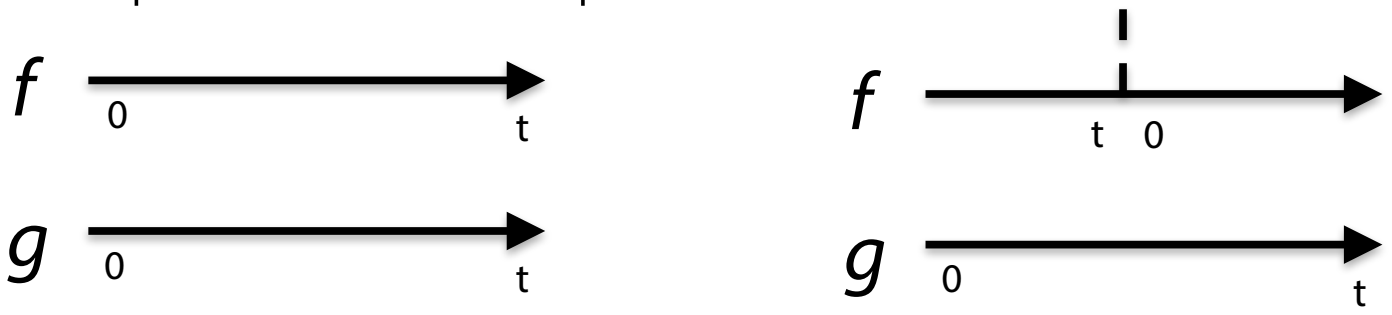



- Monte Carlo procedure to test the statistical significance accounting for the spatial and temporal proximity
  - Prune potentially coincidental relationships

**Statistical Significance**

Permutations: need to respect spatio-temporal correlations of the data!

Temporal case → Temporal shifts



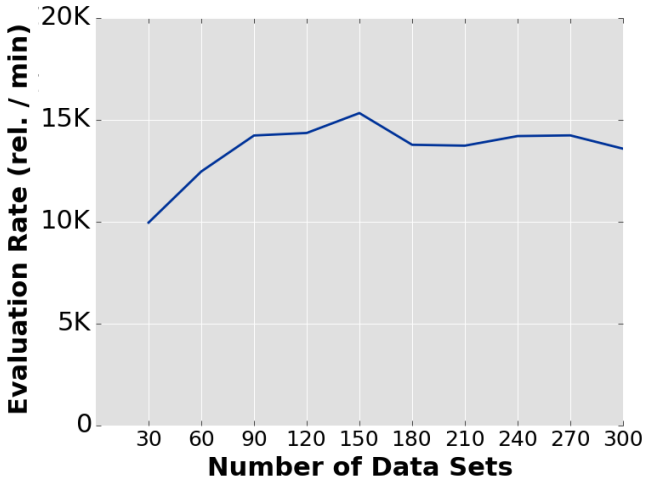

Spatial case → Spatial shifts

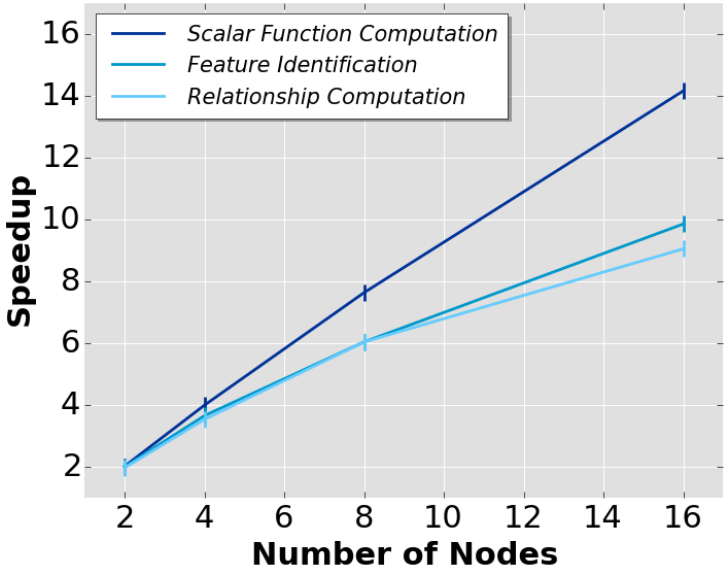
# Experimental Evaluation

- Implemented using map-reduce
  - Feature identification and relationship evaluation are independent operations
- Two collections of data sets used for experiments
  - NYC Urban: *9 data sets* from NYC agencies
  - NYC Open Data: 300 spatio-temporal data sets
- Setup
  - 20 compute nodes, AMD Opteron(TM) Processor 6272 (4x16 cores) running at 2.1GHz, 256GB of RAM – *for most experiment*
  - *Amazon EMR:* m1.medium (for master) and r3.2xlarge (for slaves) – *for scalability tests*

# Quantitative Evaluation

- Approach is efficient: 200 min to compute scalar functions and features for NYC Open Data; and 60 min for NYC Urban
- Query rate: evaluate $10^4$ relationships per minute
- Scales linearly with number of nodes
- Assessed correctness and robustness

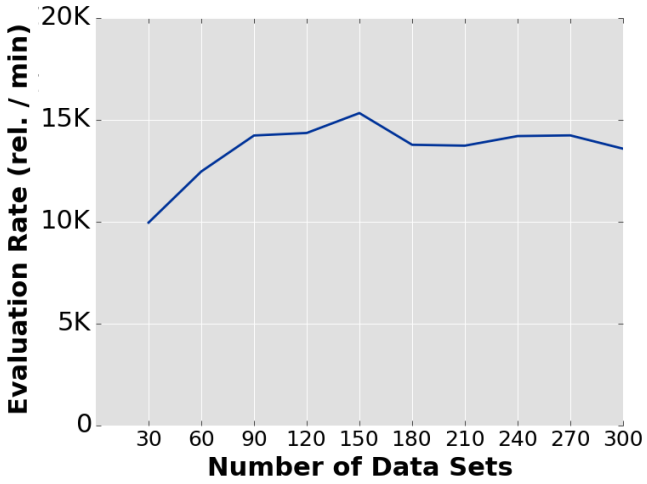


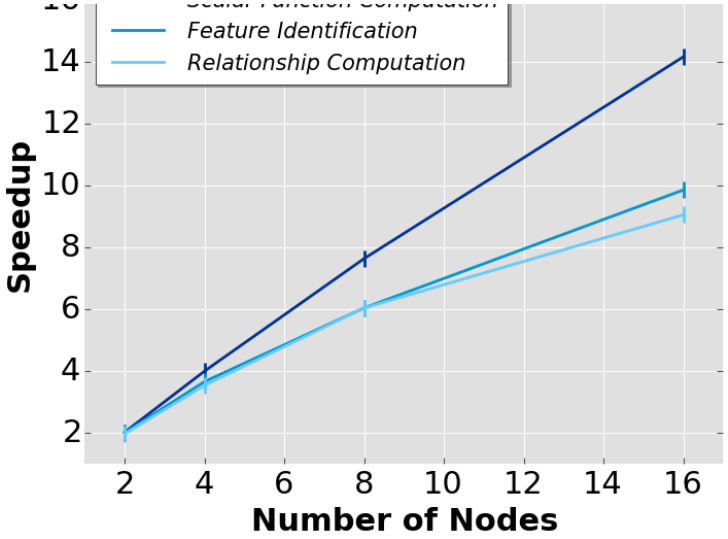




Details in [Chirigati et al., ACM SIGMOD 2016]

# Quantitative Evaluation

- Approach is efficient: 200 min to compute scalar functions and features for NYC Open Data; and 60 min for NYC Urban
- Scales linearly with number of nodes
- Code, data, and reproducible experiments available at:
- https://github.com/ViDA-NYU/data-polygamy




Details in [Chirigati et al., ACM SIGMOD 2016]

# Qualitative Evaluation

- Does the approach uncover *interesting*, *non-trivial* relationships?

Details in [Chirigati et al., ACM SIGMOD 2016]

# (Some) Interesting Relationships

1. Would a reduction in traffic speed reduce the number of accidents?

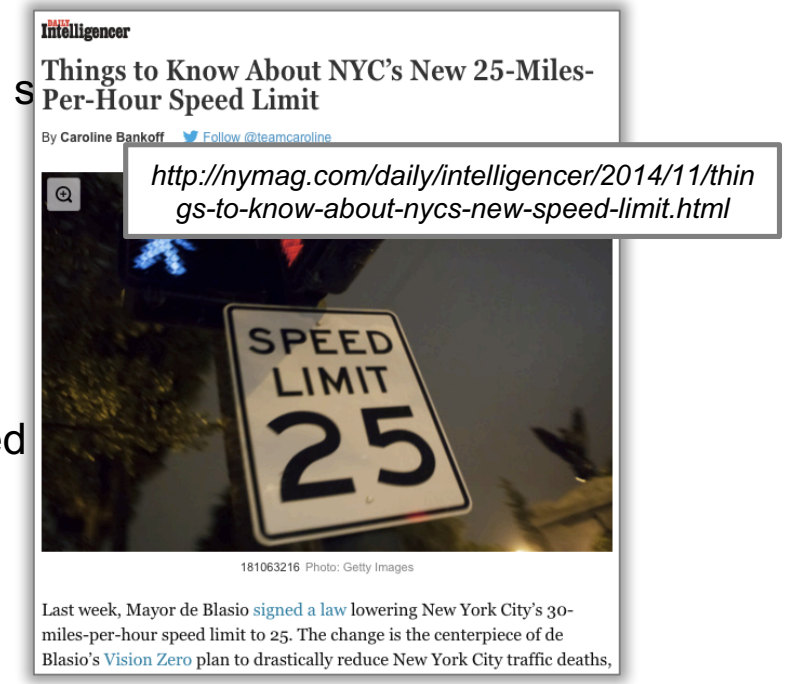
Find all relationships between Collisions and Traffic Speed data sets

*# Collisions* x *Traffic Speed*

*Positive relationship* between number of collisions and s

*# Deaths* x *Traffic Speed*

*Positive relationship* between number of persons killed speed


Intelligencer

Things to Know About NYC's New 25-Miles-Per-Hour Speed Limit

By Caroline Bankoff Follow @teamcaroline

*http://nymag.com/daily/intelligencer/2014/11/things-to-know-about-nycs-new-speed-limit.html*



181063216 Photo: Getty Images

Last week, Mayor de Blasio signed a law lowering New York City's 30-miles-per-hour speed limit to 25. The change is the centerpiece of de Blasio's Vision Zero plan to drastically reduce New York City traffic deaths,

# (Some) Interesting Relationships

2. Why it is so hard to find a taxi when it is raining?

*Find all relationships between Taxi and Weather data sets*

*# Taxi*  **X** *Precipitation* 

*Negative relationship* between number of taxis and average precipitation

***Hypothesis:*** Taxi drivers are target earners

*Taxi Fare*  **X** *Precipitation*  ⟶ *Suggests that hypothesis is true*

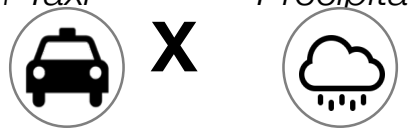
*Strong positive relationship* between precipitation and average fare

# (Some) Interesting Relationships

2. Why it is so hard to find a taxi when it is raining?

*Find all relationships between Taxi and Weather data sets*

*# Taxi* X *Precipitation*



This hypothesis had been refuted by [Farber 2014]

- Farber did not find a correlation (using OLS regression) between drivers' earnings and rainfall.
- But (i) he did not take into account the amount of rainfall—instead, he used a binary value indicating whether it rained or not; and (ii) he considered the entire time period—periods with very sparse rainfall are considered equivalent to those having higher rainfall.

It is important to consider salient features

# (Some) Interesting Relationships

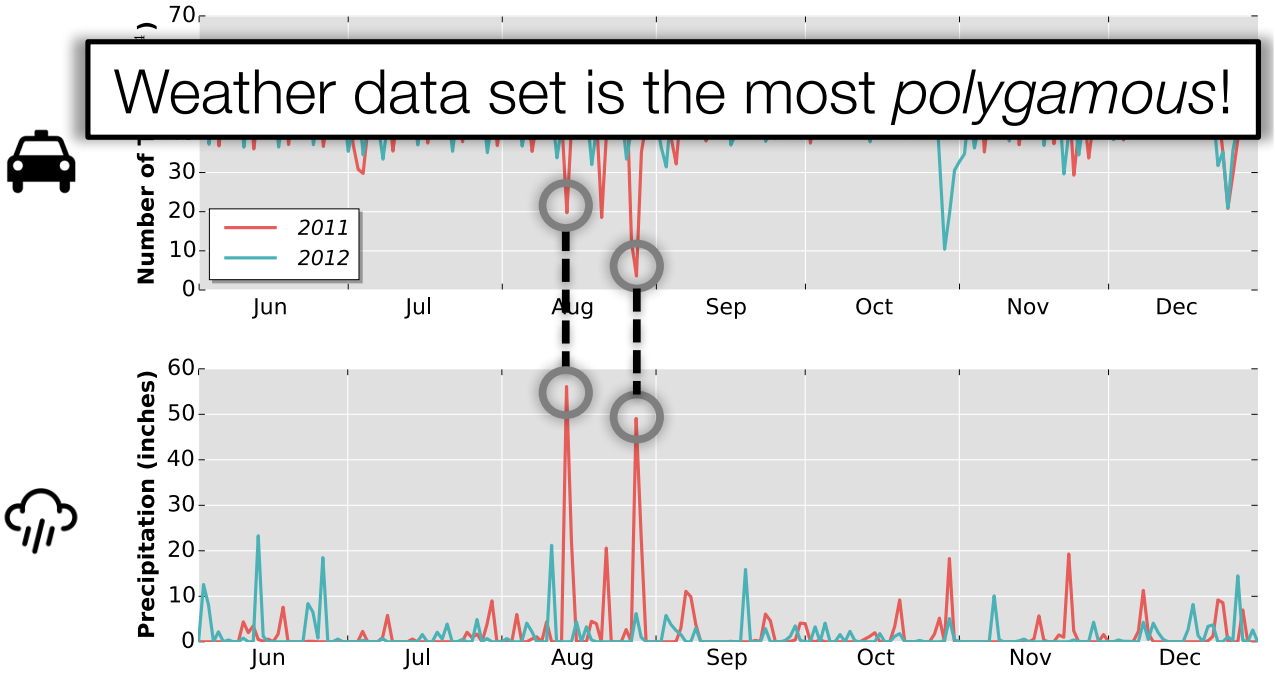
3. Why is the number of taxi trips too low?

*# Taxi* X *Precipitation*

*Negative relationship* between number of taxis and average precipitation

*# Taxi* X *Wind speed*

*Negative relationship* between number of taxis and wind speed



# (Some) Interesting Relationships

- Citi Bike and Weather

*# Citi Bike stations*  X *Snow* 

*Negative relationship* between snow precipitation and active Citi Bike stations

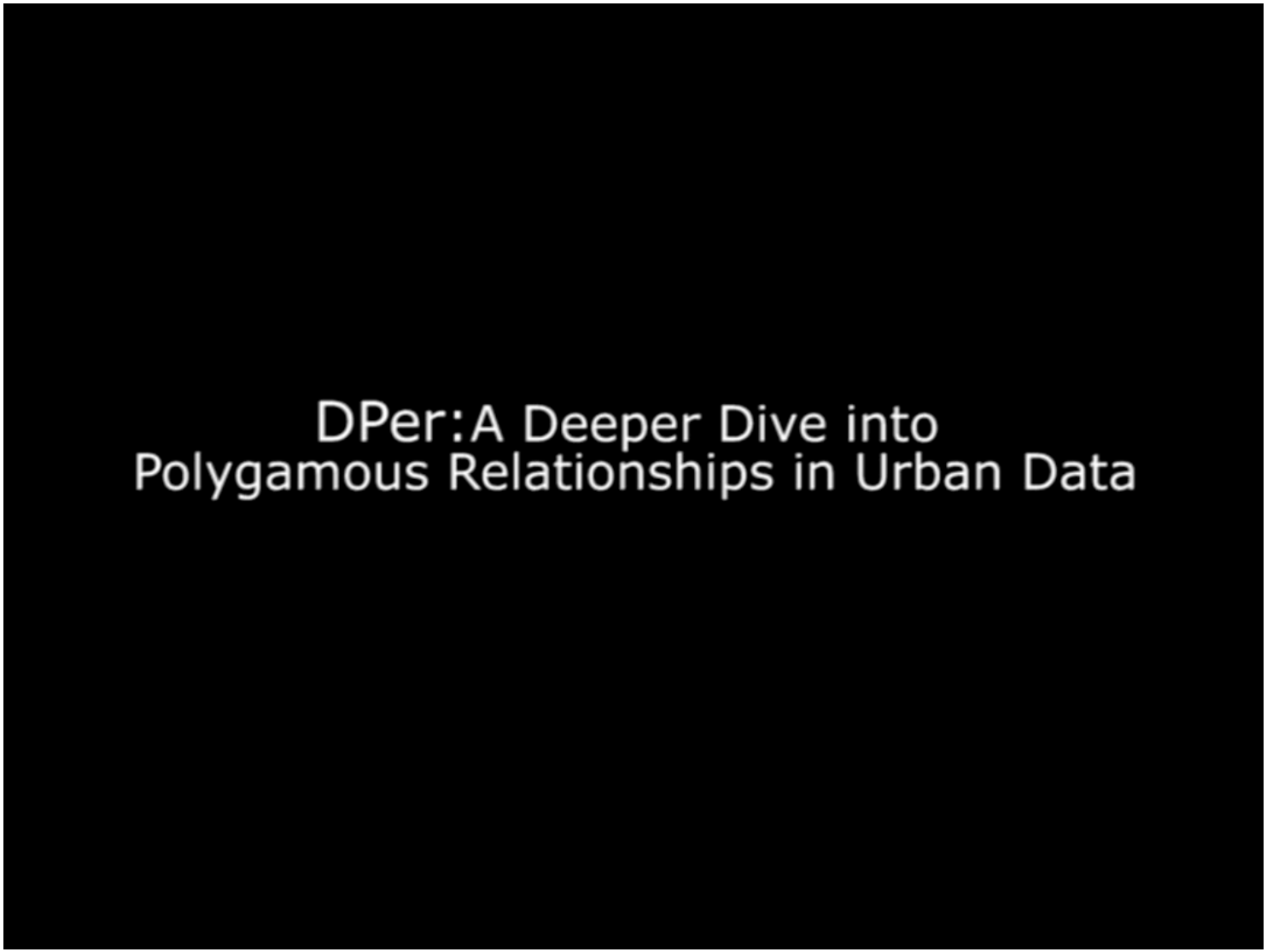
| | |
|---|---|
| *(day, city)* | ✓ |
| *(hour, city)* | ⊘ |

# Many other Relationships and Challenges

- It's hard to evaluate!
  - No ground truth available
  - Need benchmark
  - Need real use cases from domain experts
- ~ 100 *significant* relationships per resolution
  - More relationships (and their implications) can be understood by having domain experts

How to explore and analyze the relationships?

# 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

# Transparency and Reproducibility

# Science and Reproducibility

- Reproducibility is the cornerstone of science
- *If I have seen further it is by standing on the shoulders of giants.*

  *Isaac Newton*
- If we can't trust previous results, we have to start over from scratch
  - Science is incremental and self-correcting
- To increase impact, visibility [Vandewalle et al. 2009] and research quality [Begley and Ellis 2012]
- *Without reproducibility, people die!*

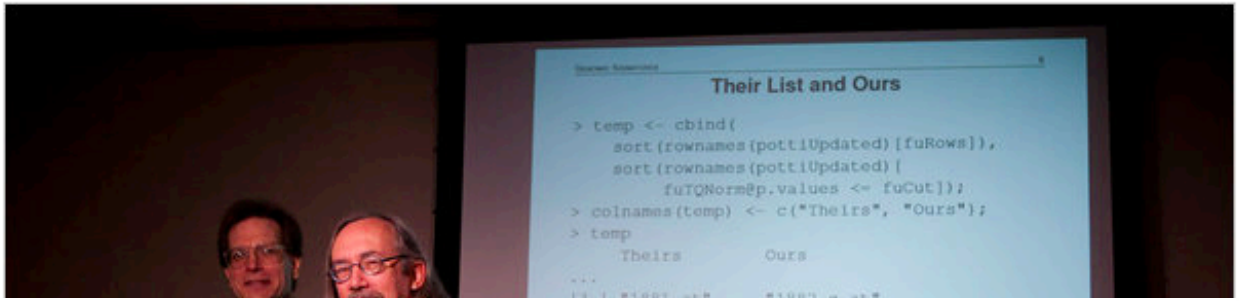
John Wilbanks, AMPS Workshop on Reproducibility 2011

# Science and Reproducibility

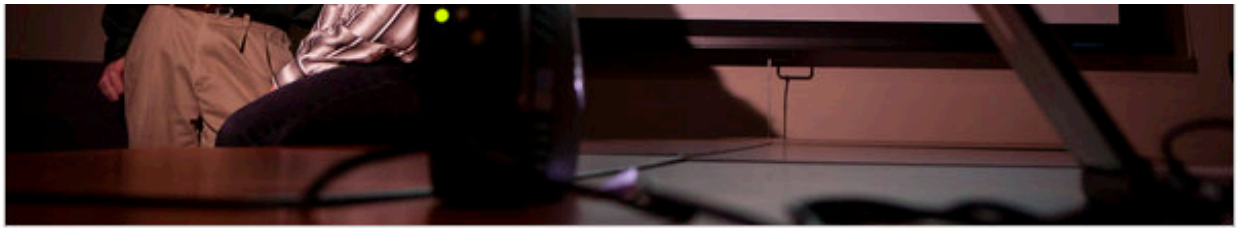
The New York Times

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

## How Bright Promise in Cancer Testing Fell Apart




But the research at Duke turned out to be wrong. Its gene-based tests proved worthless, and the research behind them was discredited. Ms. Jacobs died a few months after treatment, and her husband and other patients' relatives have retained lawyers.



Michael Stravato for The New York Times

Keith Baggerly, left, and Kevin Coombes, statisticians at M. D. Anderson Cancer Center, found flaws in research on tumors.

By GINA KOLATA
Published: July 7, 2011

When Juliet Jacobs found out she had lung cancer, she was terrified, but realized that her hope lay in getting the best treatment medicine could offer. So she got a second opinion, then a third. In February of 2010, she ended up at Duke University, where she entered a research study whose promise seemed stunning.

RECOMMEND
TWITTER
COMMENTS (76)
E-MAIL

## Nobel Laureate Retract Prize

By KENNETH CHANG
Published: September 23, 2010

Linda B. Buck, who shared the 2004 Medicine for deciphering the working retracted two scientific papers after sh unable to repeat the findings.

NYU TANDON SCHOOL OF ENGINEERING

VIDA IMAGING AND DATA ANALYSIS CENTER

# Science and Reproducibility

| | | | Real GDP growth Debt/GDP | | | | |
|---|---|---|---|---|---|---|---|
| 4 | Country | Coverage | 30 or less | 30 to 60 | 60 to 90 | 90 or above | 30 or less |
| 26 | | | 3.7 | 3.0 | 3.5 | 1.7 | 5.5 |
| 27 | Minimum | | 1.6 | 0.3 | 1.3 | -1.8 | 0.8 |
| 28 | Maximum | | 5.4 | 4.9 | 10.2 | 3.6 | 13.3 |
| 29 | | | | | | | |
| 30 | US | 1946-2009 | n.a. | 3.4 | 3.3 | -2.0 | n.a. |
| 31 | UK | 1946-2009 | n.a. | 2.4 | 2.5 | 2.4 | n.a. |
| 32 | Sweden | 1946-2009 | 3.6 | 2.9 | 2.7 | n.a. | 6.3 |
| 33 | Spain | 1946-2009 | 1.5 | 3.4 | 4.2 | n.a. | 9.9 |
| 34 | Portugal | 1952-2009 | 4.8 | 2.5 | 0.3 | n.a. | 7.9 |
| 35 | New Zealand | 1948-2009 | 2.5 | 2.9 | 3.9 | -7.9 | 2.6 |
| 36 | Netherlands | 1956-2009 | 4.1 | 2.7 | 1.1 | n.a. | 6.4 |
| 37 | Norway | 1947-2009 | 3.4 | 5.1 | n.a. | n.a. | 5.4 |
| 38 | Japan | 1946-2009 | 7.0 | 4.0 | 1.0 | 0.7 | 7.0 |
| 39 | Italy | 1951-2009 | 5.4 | 2.1 | 1.8 | 1.0 | 5.6 |
| 40 | Ireland | 1948-2009 | 4.4 | 4.5 | 4.0 | 2.4 | 2.9 |
| 41 | Greece | 1970-2009 | 4.0 | 0.3 | 2.7 | 2.9 | 13.3 |
| 42 | Germany | 1946-2009 | 3.9 | 0.9 | n.a. | n.a. | 3.2 |
| 43 | France | 1949-2009 | 4.9 | 2.7 | 3.0 | n.a. | 5.2 |
| 44 | Finland | 1946-2009 | 3.8 | 2.4 | 5.5 | n.a. | 7.0 |
| 45 | Denmark | 1950-2009 | 3.5 | 1.7 | 2.4 | n.a. | 5.6 |
| 46 | Canada | 1951-2009 | 1.9 | 3.6 | 4.1 | n.a. | 2.2 |
| 47 | Belgium | 1947-2009 | n.a. | 4.2 | 3.1 | 2.6 | n.a. |
| 48 | Austria | 1948-2009 | 5.2 | 3.3 | -3.8 | n.a. | 5.7 |
| 49 | Australia | 1951-2009 | 3.2 | 4.9 | 4.0 | n.a. | 5.9 |
| 50 | | | | | | | |
| 51 | | | 4.1 | 2.8 | 2.8 | =AVERAGE(L30:L44) | |

In this p aying pop-
historical c costs? Are
search for ely a man-
public deb
main resu empirical,
growth an historical
mal" debt ar, central
tries with Carmen M.
of GDP ar 08, 2009b).
wise; aver ly difficult
percent lo es of pub-
between p ntries, and
similar acr g markets.
economies countries
find no sy gether, the
debt levels servations
mies as a ems, insti-
exceptions y arrange-
trast, in em
debt levels
Our top on external
Public debt has been soaring in the wake of the and by private entities. For emerging markets, overnments
recent global financial maelstrom, especially in the epicenter countries. This should not be surprising, given the experience of earlier severe we find that there exists a significantly more severe threshold for total gross external debt (public and private)—which is almost exclu-

ALIZATION
ING AND
ANALYSIS
ER

# Reproducibility and Data Science

- Reproducibility is important not just for science
- Many decisions are made based on results of computational analyses
- Need transparency!

Subsidized housing

Computational hedge funds

Policing and crime prevention

Facebook ranking

Taxi fare increases

…

# How?

***Provenance***: a key ingredient for reproducibility

*"The source or origin of an object; its history and pedigree; a record of the ultimate derivation and passage of an item through its various owners."*

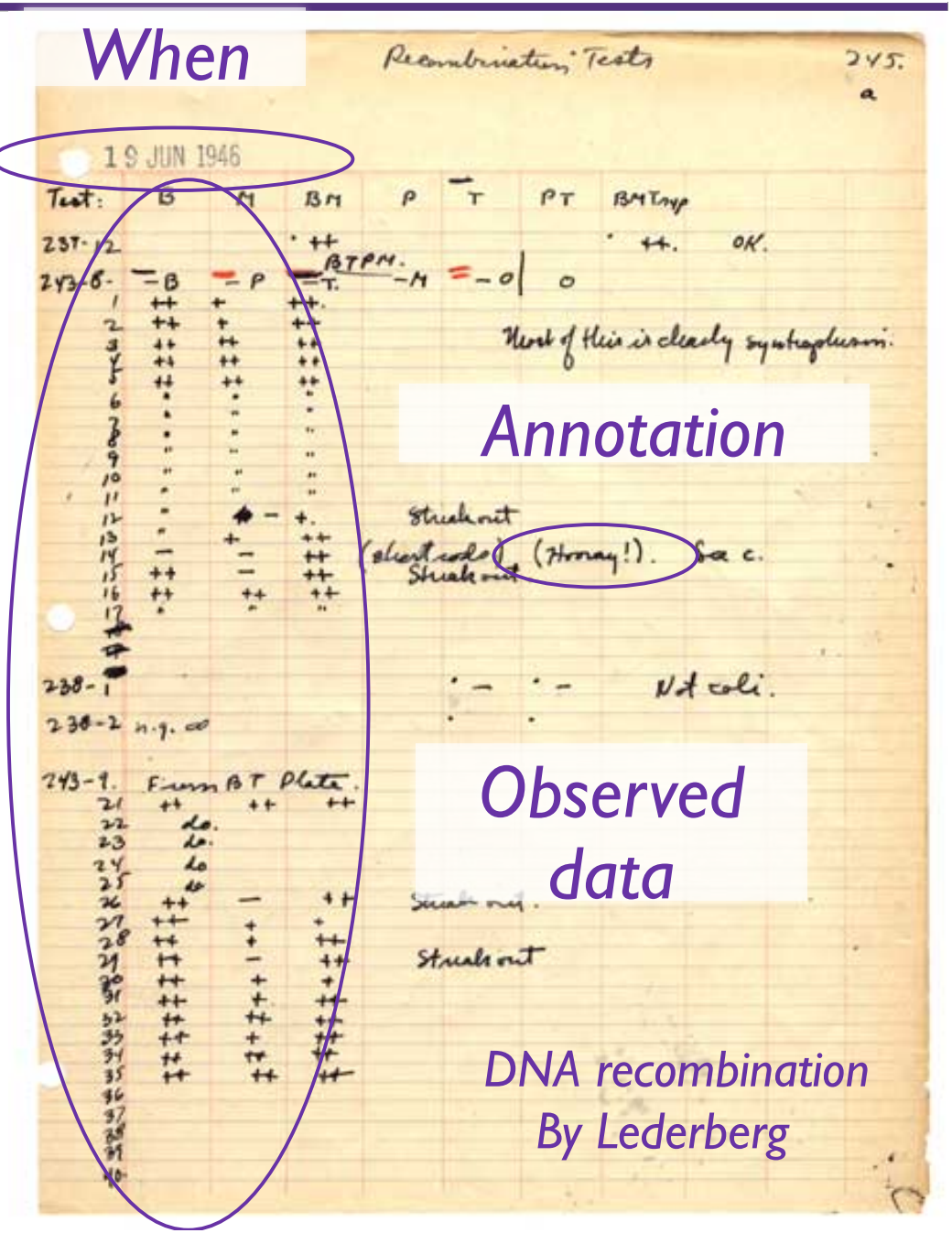
*The Oxford English Dictionary*

Helps determine the value, accuracy and authorship of an object

Used in many fields: works of art and antiques, archives and books, ***science***, …

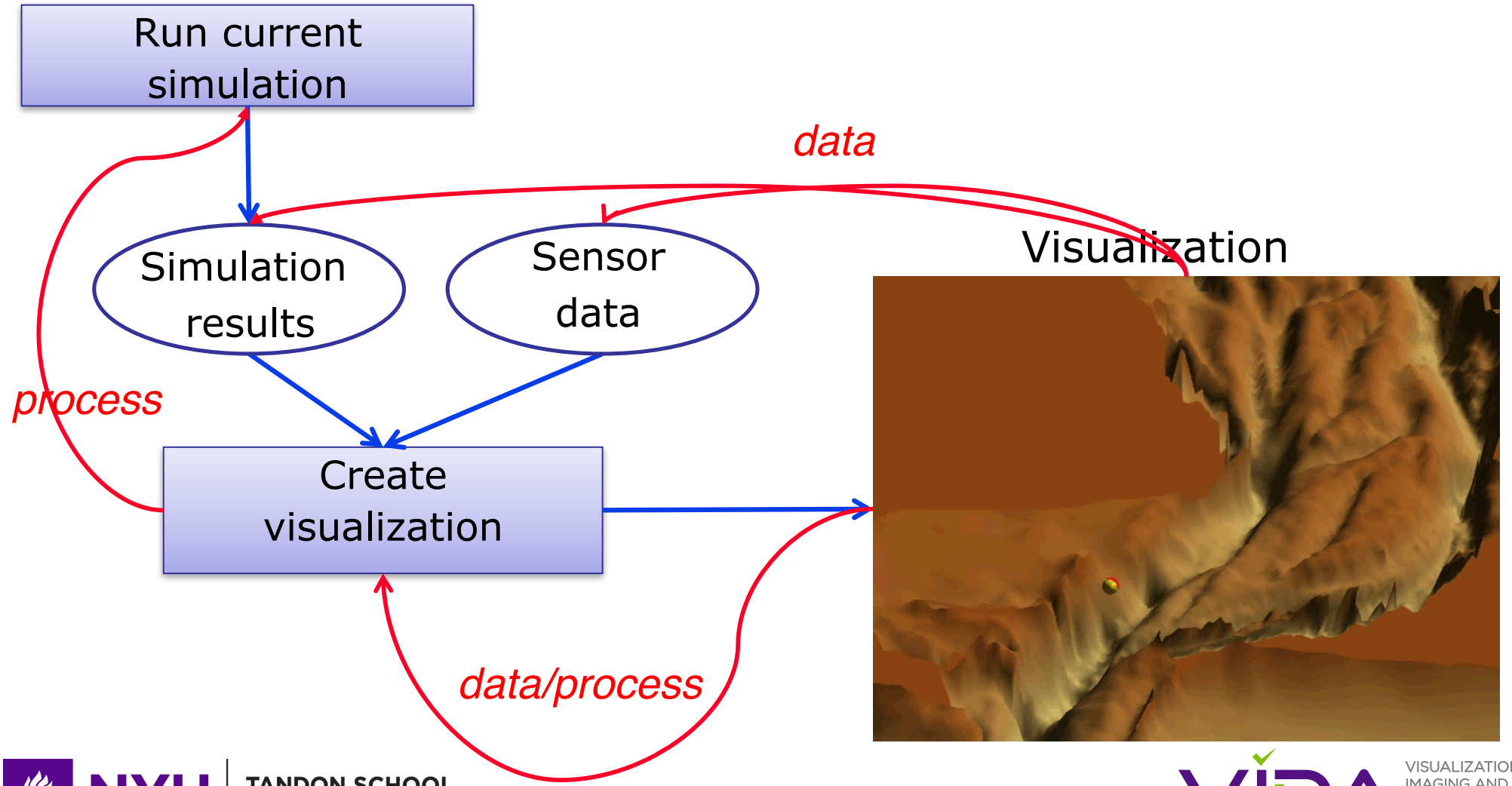
# Provenance in Science

- Not a new issue!
- Lab notebooks have been used for a long time
  - Reproduce results
  - Evidence in patent disputes
- What is new?
  - Large volumes of data
  - Complex analyses
- Writing notes is no longer an option
  - Need systematic means to capture provenance



# What is *Computational* Provenance?

- A causality graph: process and data dependencies



# Why is Provenance Important?

anon4876_base_20060331.jpg

anon4877_lesion_20060401.jpg


How were these images created?

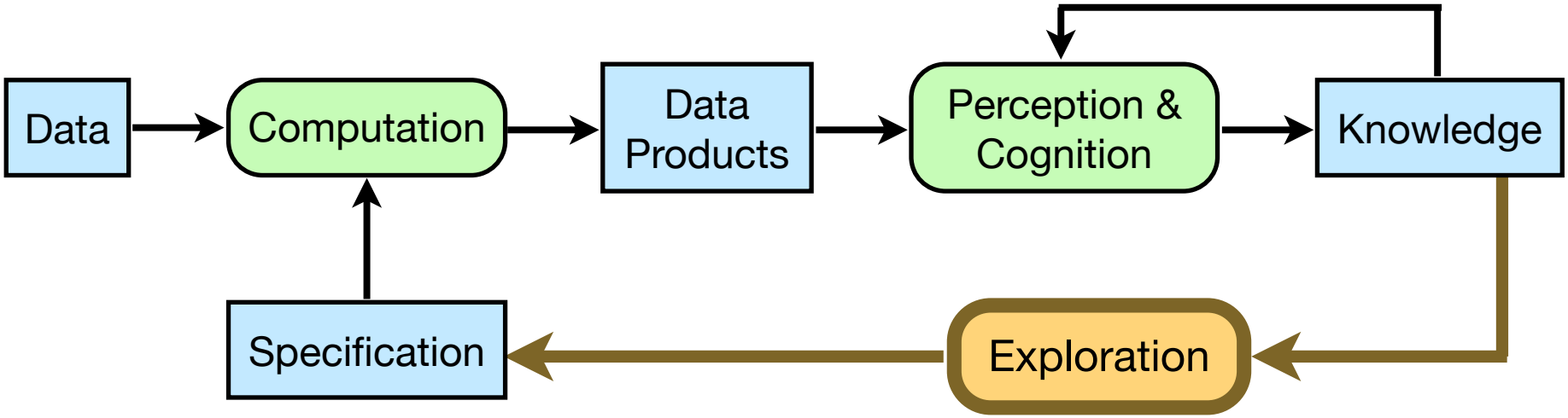
Was any pre-processing applied to the raw data?

Who created them?

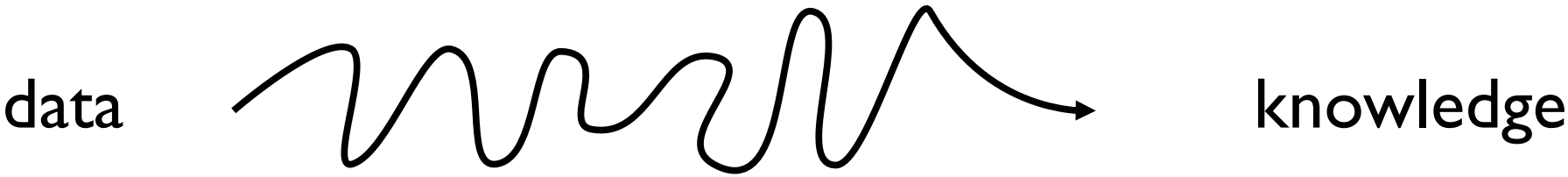
What's the difference?

Are they really from the same patient?

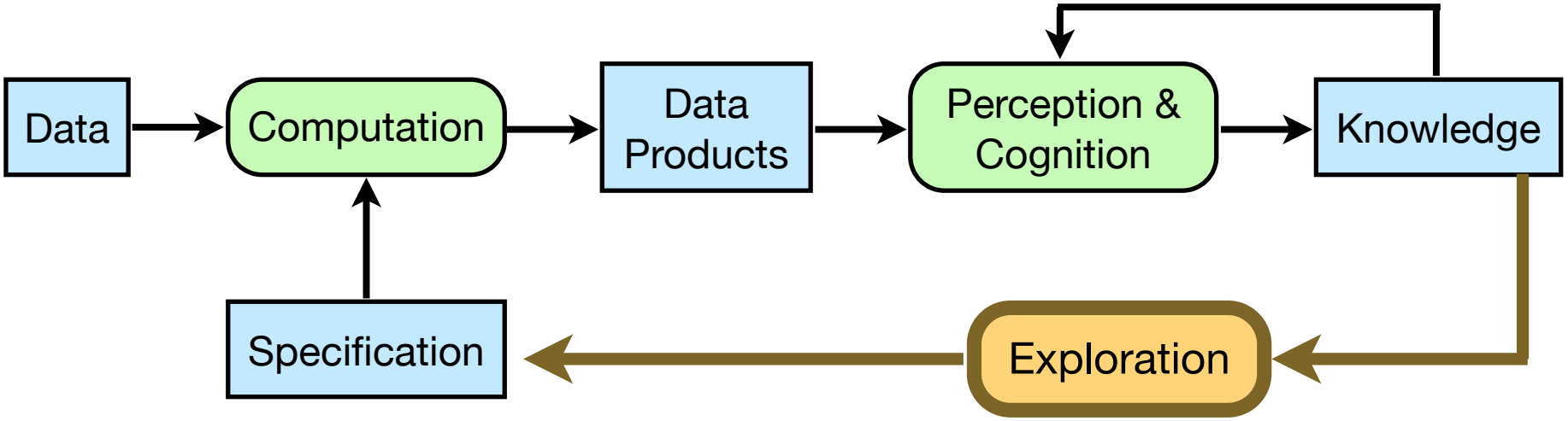
# Data Lifecycle: The Exploration Pipeline




[Modified from Van Wijk, Vis 2005]



# Data Exploration Today

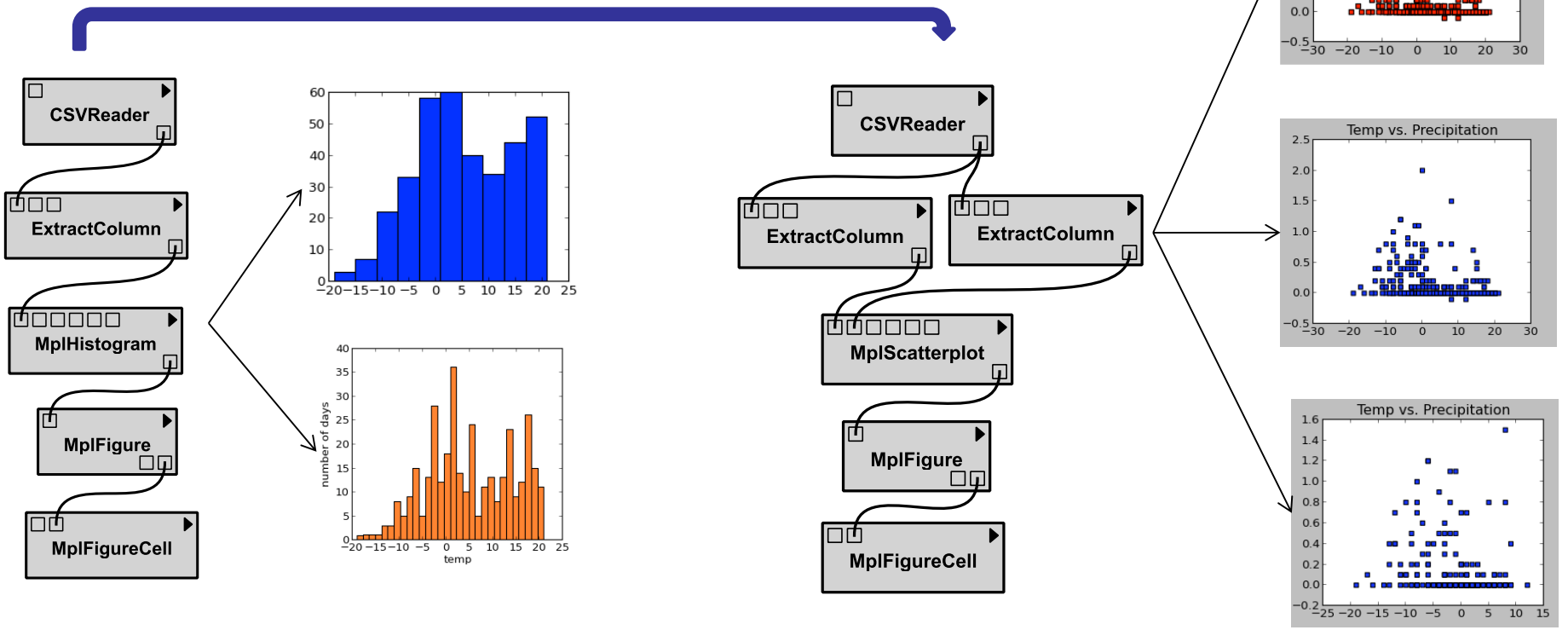



[Modified from Van Wijk, Vis 2005]

- Iterative process to generate and test hypotheses
- Easy to get lost---derive a result and not remember how you got there
- Lots of data: both input and derived
- Complex processes that encompass multiple tools

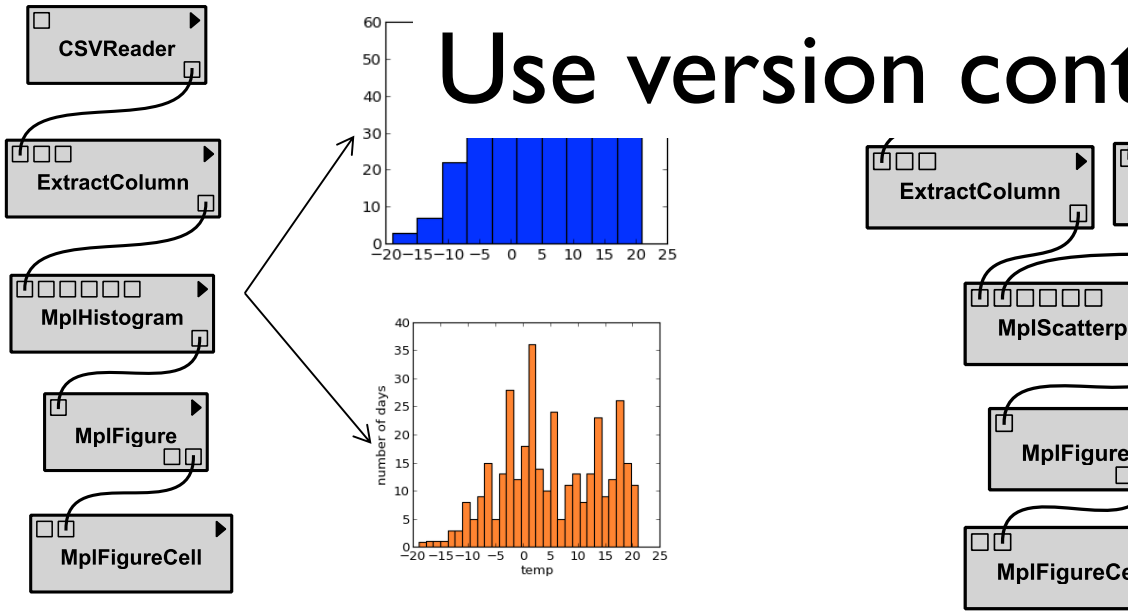
# Data Exploration and Workflows

- Workflows naturally capture *data provenance*
  - The workflow specification mirrors the provenance graph
- But in data exploration, change is the norm
  - Workflows specifications are iteratively refined



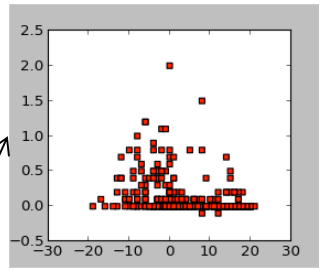
# Data Exploration and Workflows

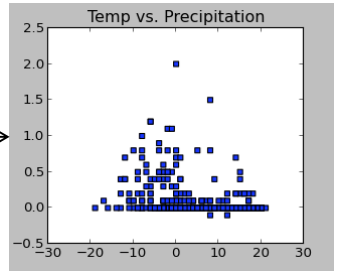
- Workflows/scripts/code naturally capture *data provenance*
  - The workflow specification mirrors the provenance graph
- But in data exploration, change is the norm
  - Specifications are iteratively refined

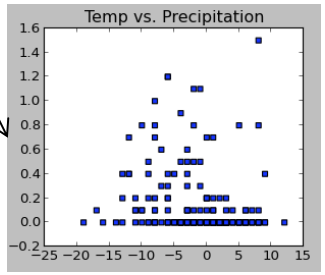



## Use version control!



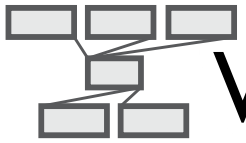





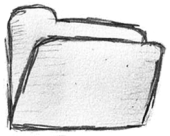
# Provenance: Reproducibility and Sharing

- Workflow/script/code provenance is not enough



WORKFLOW / GRAPH

- Need to manage data
  - Data accessed may change
  - File names are insufficient for identification purposes

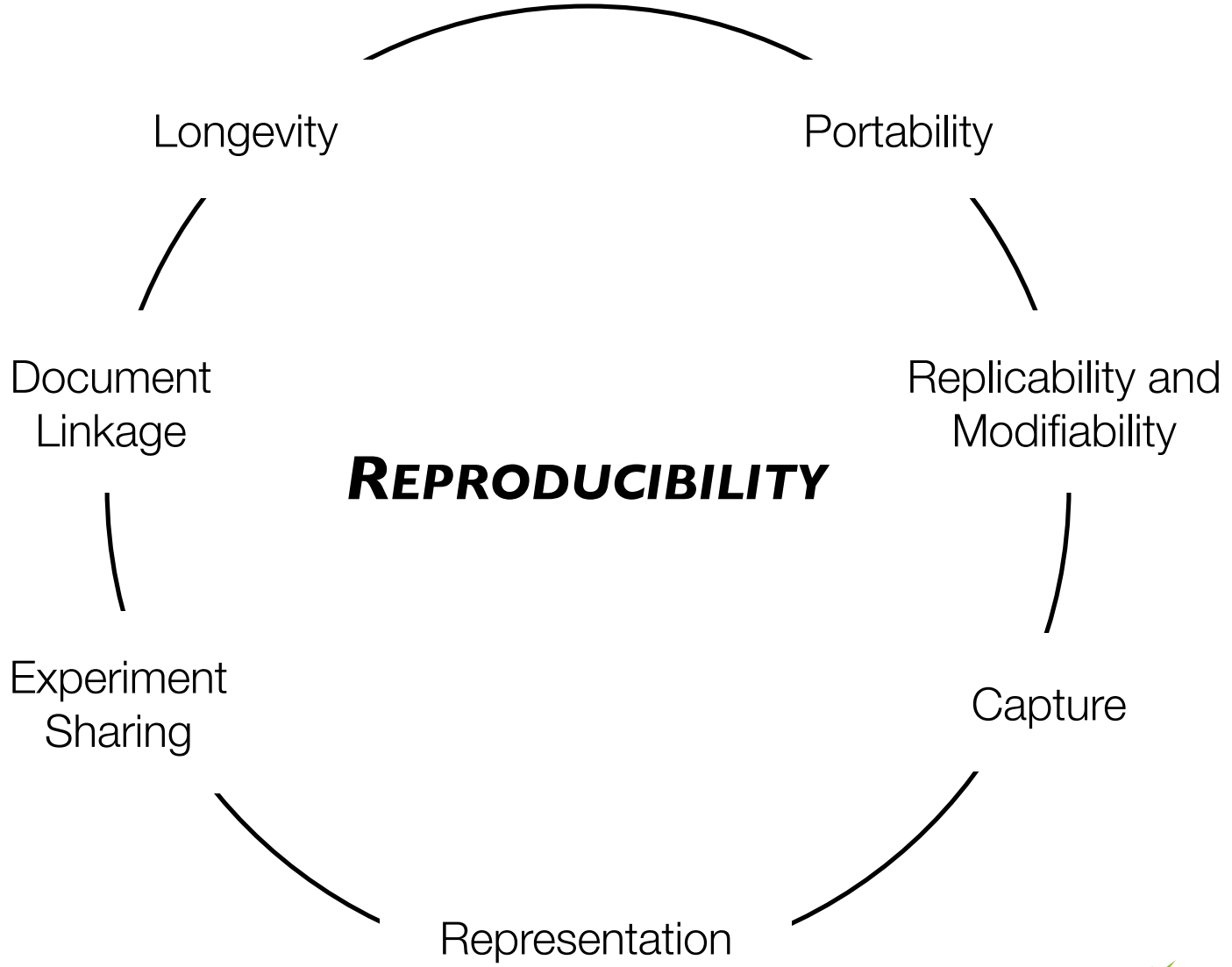


DATA

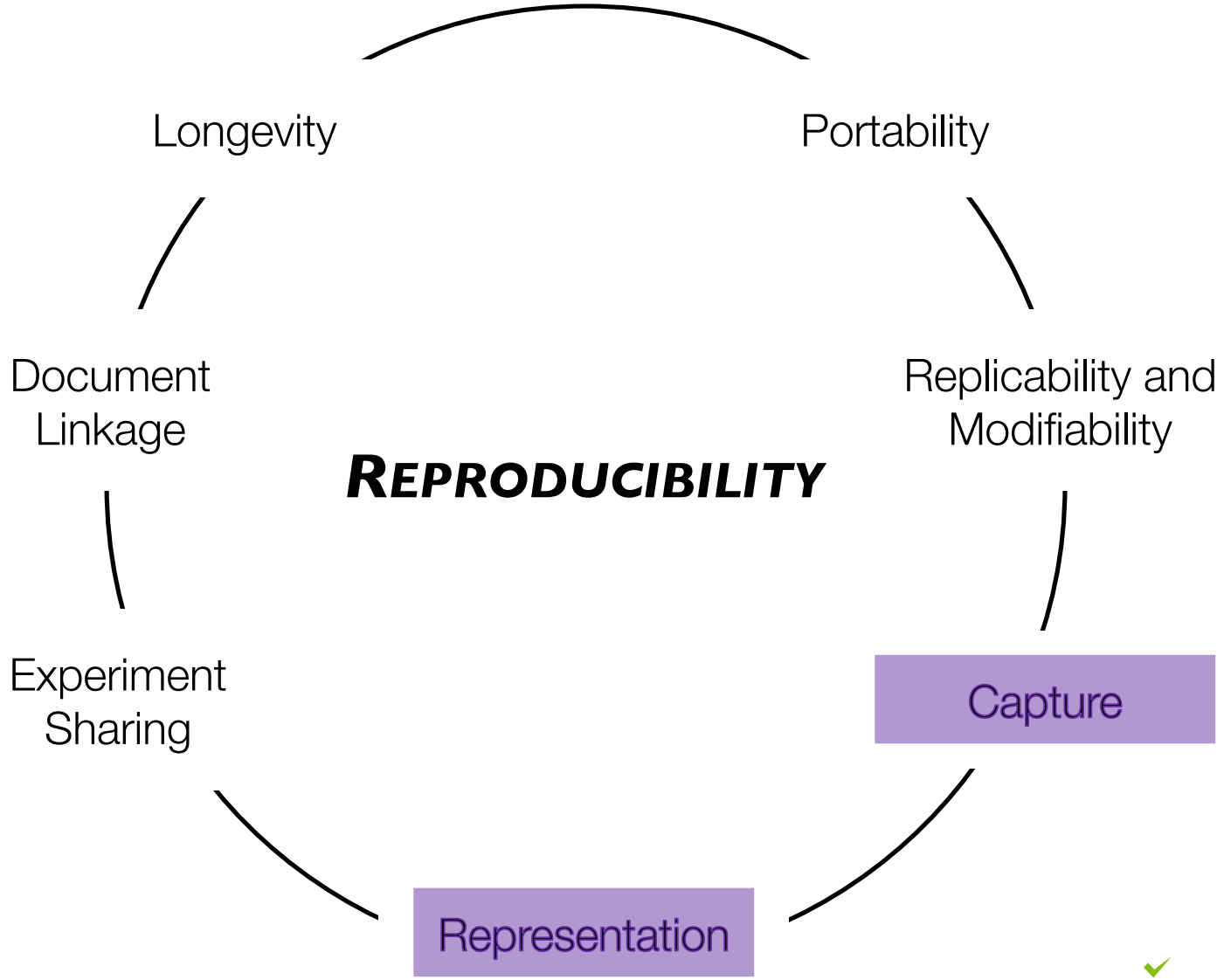
- Need portability
  - Since environment may change, provenance for the computational environment is necessary (e.g., library versions)
  - Ability to run pipelines in different environments

ENVIRONMENT

# Components for Reproducibility



# Components for Reproducibility



# Capture and Representation

*OS-Based:*

- *ES3 [Frew and Slaughter, 2008]*
- *PASS: http://www.eecs.harvard.edu/syrah/pass/*
- *ReproZip: https://vida-nyu.github.io/reprozip/*
- *CDE: http://www.pgbovine.net/cde.html*

*Source code based:*

- *Git / GitHub: https://github.com/*
- *Mercurial / Bitbucket: https://bitbucket.org/*

# Capture and Representation (cont.)

*Code-Instrumented capture and representation:*

- *noWorkflow (Python scripts): https://github.com/gems-uff/noworkflow*
- *Python CSL: https://sourceforge.net/projects/provenance-csl/*
- *Sumatra: http://neuralensemble.org/sumatra/*
- *VCR: http://vcr.stanford.edu/*

# noWorkflow

- **n**ot **o**nly **Workflow**
- Transparently captures provenance of Python scripts
- It does not require a version control system or an instrumented environment
- Usage: Instead of running

```
python simulation.py data1.dat data2.dat
```

Run:

```
now simulation.py data1.dat data2.dat
```

- Planning for reproducibility: replace Python with noWorkflow
  - Use of noWorkflow for the entire experiment's lifetime
- Reproducibility after the fact: capture of a run after the experiment is ready for publication
- Provenance analysis: Visualize and query provenance information, compare provenance traces
- Available at *https://github.com/gems-uff/noworkflow*

[Murta et al., IPAW 2014]

# Capture and Representation (cont.)

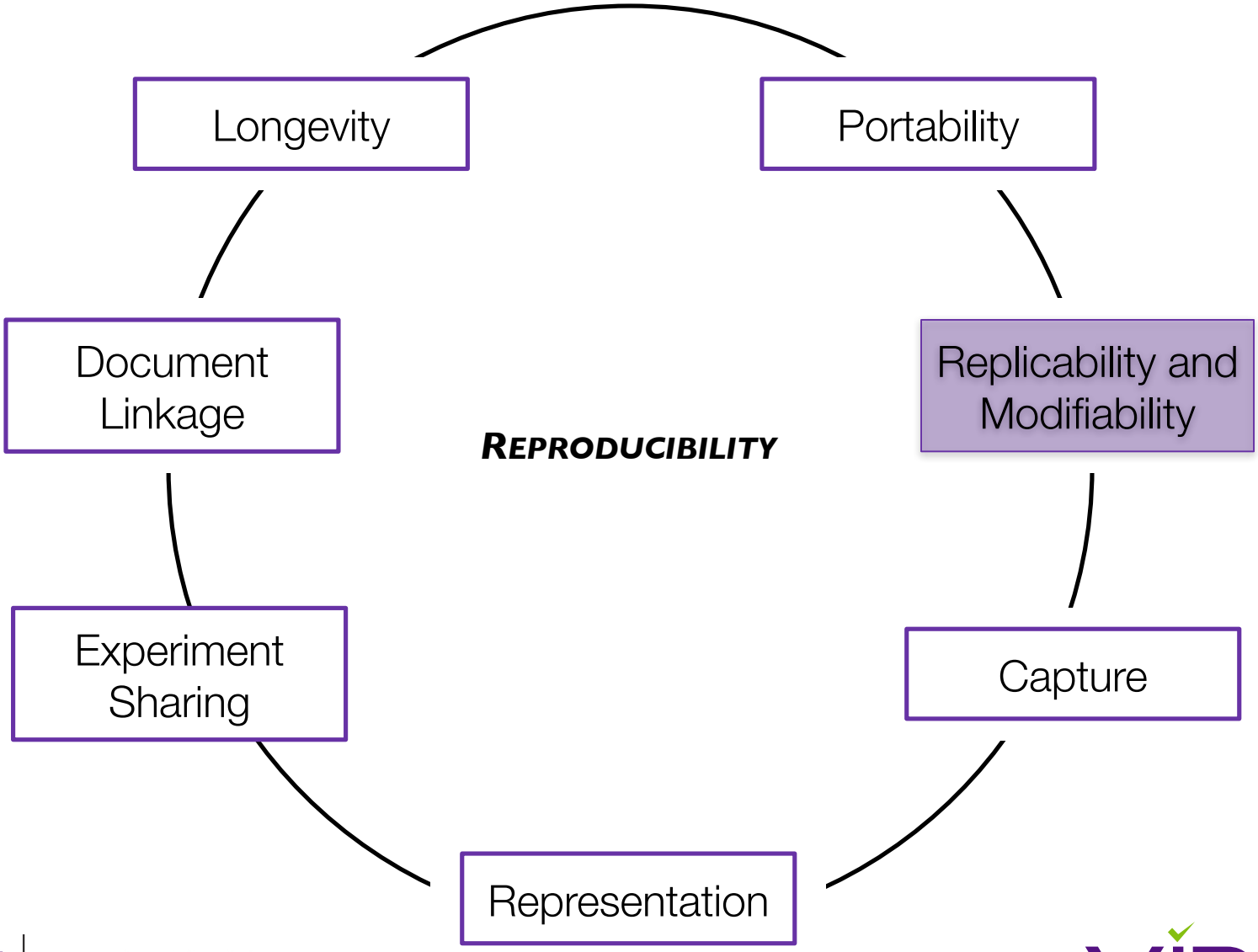
*Workflow-Based capture and representation (workflow systems):*

- *VisTrails: http://www.vistrails.org/*
- *Kepler: https://kepler-project.org/*
- *Taverna: http://www.taverna.org.uk/*

# VisTrails: Managing Exploration

- Combines features of visualization, data analysis, and scientific workflow systems
  - Integrate multiple libraries (e.g., VTK, R, matplotlib, *)
- Tracks provenance as users generate and test hypotheses
  - Data + workflow provenance---*treat workflow as a 1st-class data product*
- *Leverage provenance to streamline exploration*
  - Support for reflective reasoning and collaboration
- Open-source: www.vistrails.org
- Multi-platform: Mac, Linux, and Windows
- Focus on usability—build tools for end users

# Components for Reproducibility



# Replicability x Modifiability

***Replicability***: Ability to repeat the experiment execution with the *same* parameters and data originally used

***Modifiability***: Ability to *vary* parameters, data, and even the structure of the experiment

# *Best Practices*

***Some tools do not provide an executable representation***

*E.g.: ES3, noWorkflow*

***Some tools support only replicability***

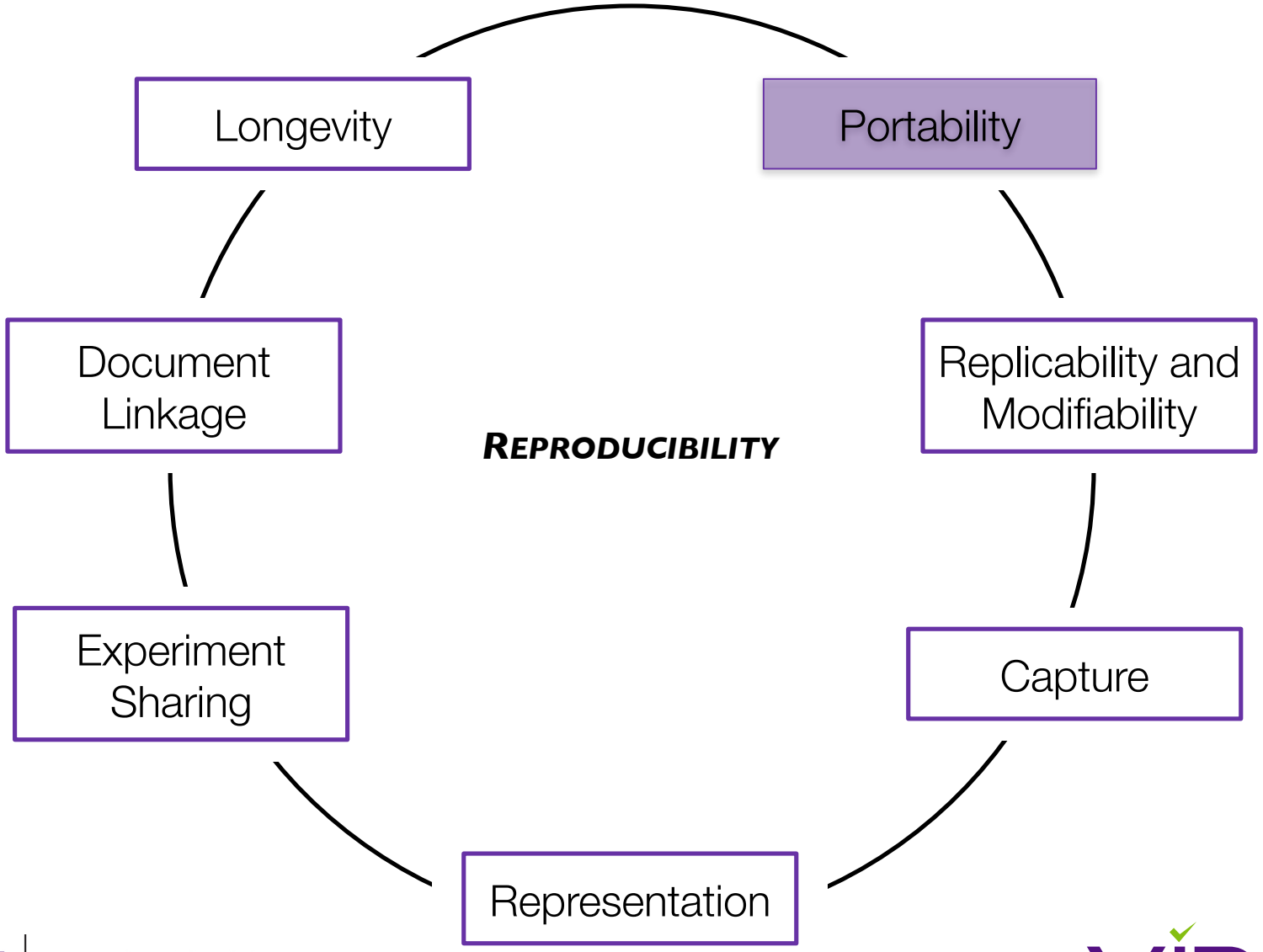
*E.g.: PASS, Sumatra*

***Some tools also support modifiability***

E.g.: ReproZip, CDE, workflow systems

*Always expose input parameters in your code/program !*

# Components for Reproducibility



# Portability

Ability for the experiment to be reproduced in different environments

Different levels:

- *Low Portability – same original environment*
- *Medium Portability – compatible environments*
- *High Portability – different environments*

# Best Practices

***How to attain high portability?***

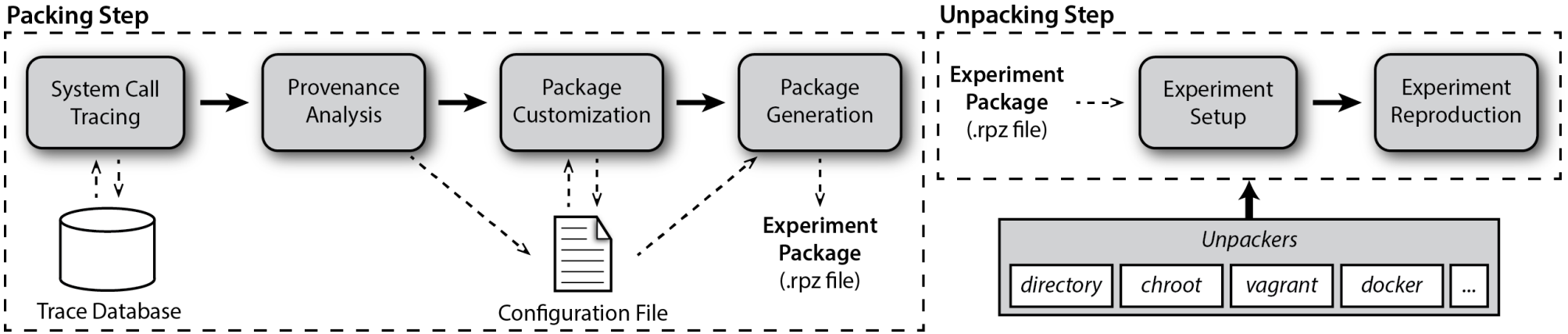
- *Virtual Machines (e.g.: Vagrant: https://www.vagrantup.com/)*
- *Docker: https://www.docker.com/*
- *ReproZip*
- *Hosted Execution (e.g.: crowdLabs)*

*Do not use hard-coded absolute paths in your program*

*Document dependencies*

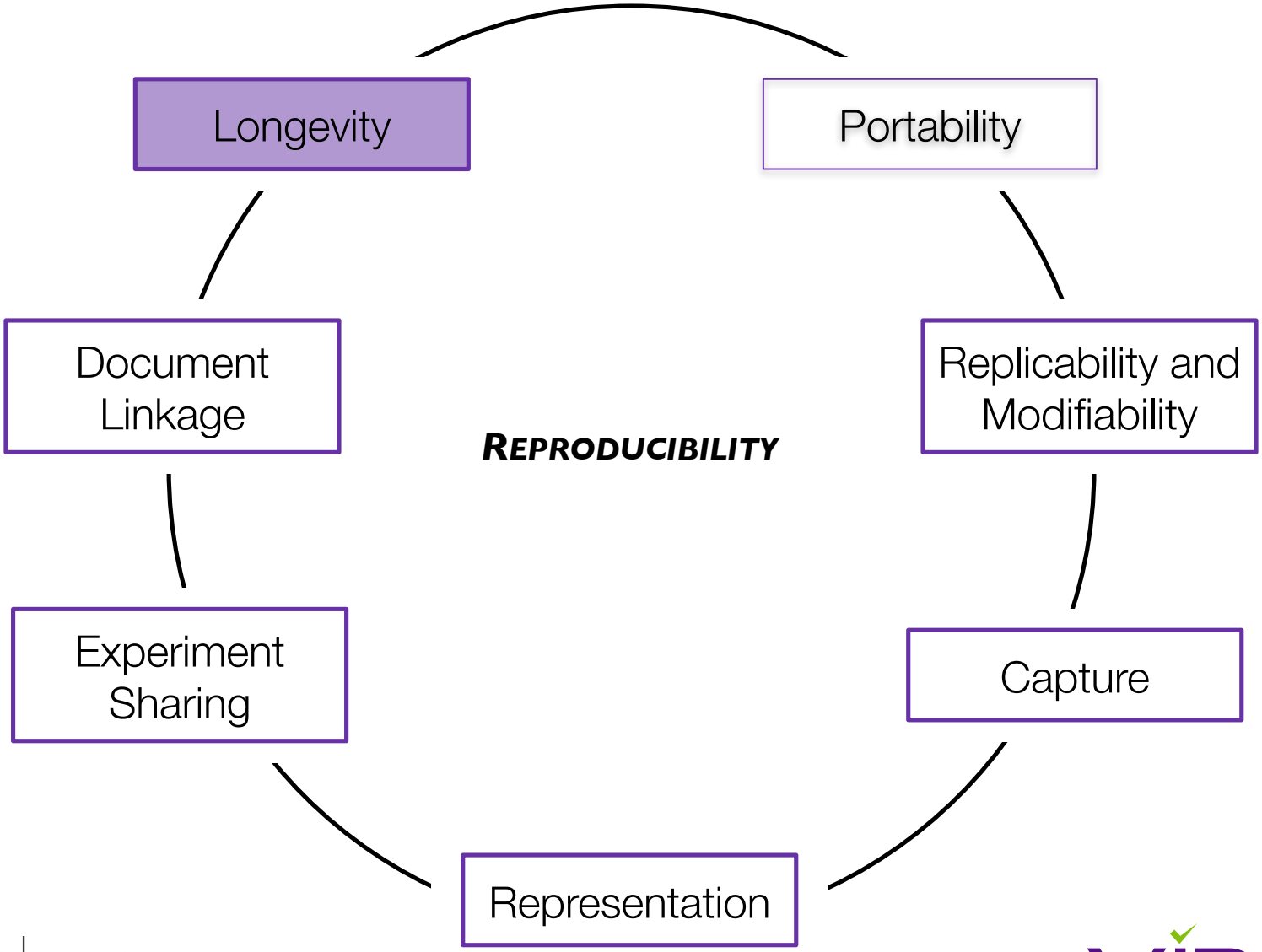
*Use open-source tools whenever possible*

# ReproZip




https://www.youtube.com/watch?v=-zLPuwCHXo0

# Components for Reproducibility



# Longevity

Ability to reproduce the experiment long after it was created

Two main mechanisms

- Longevity *by archiving*
- Longevity *by upgrading*

# *Best Practices*

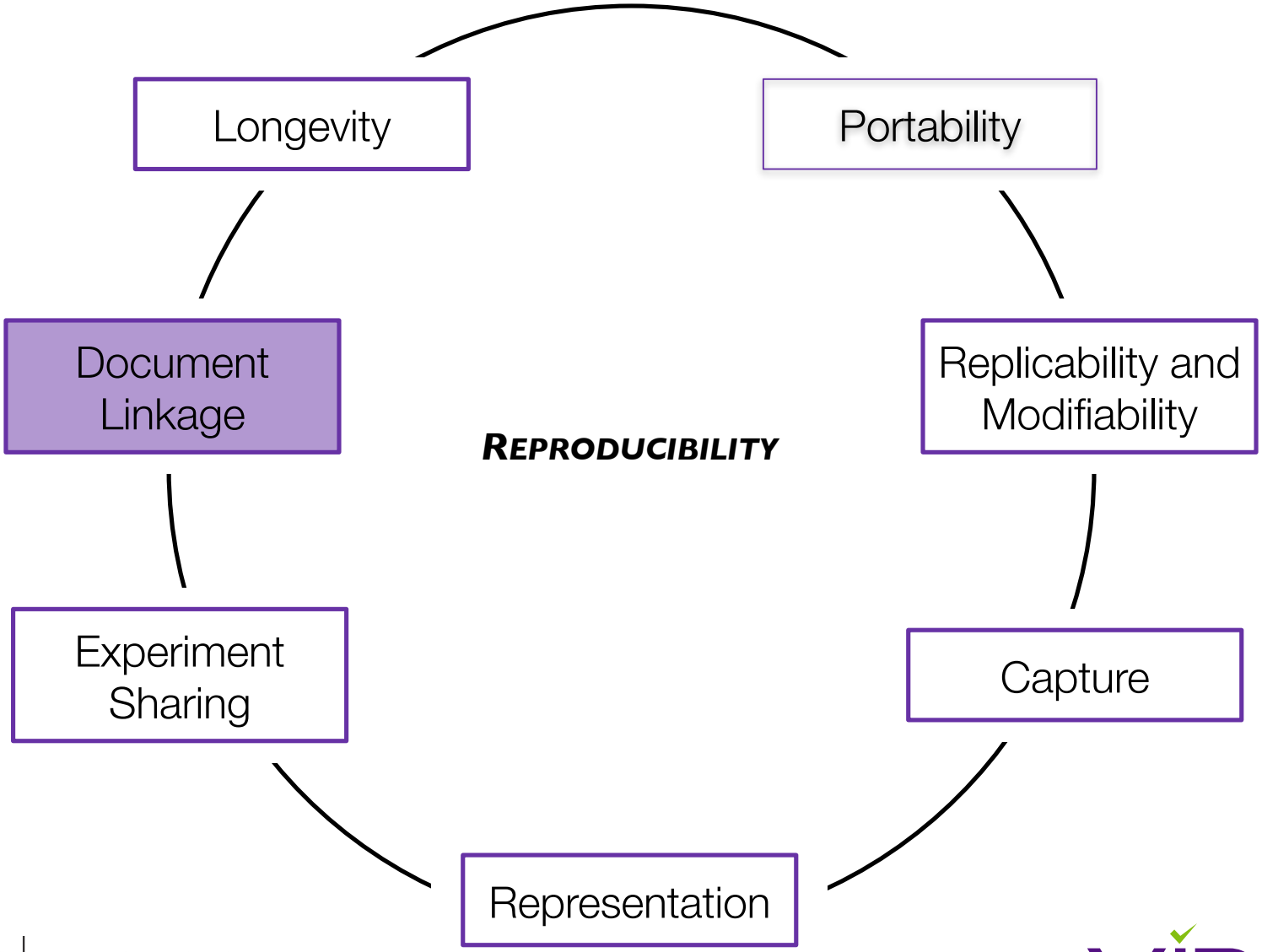
***Which tools support longevity by archiving?***

- *Virtual Machines*
- *Docker*
- *ReproZip*
- *CDE*

***Which tools support longevity by upgrading?***

- *VisTrails [Koop et al., 2010]*
- *Taverna [Belhajjame et al. 2011] – replacement of Web services*
- *Useful to integrate with new tools*

# Components for Reproducibility



# Document Linkage

[Freedman et al., Phys. Rev. 2012]



**Galois Conjugates of Topological Phases**

M. H. Freedman,[1] J. Gukelberger,[2] M. B. Hastings,[1] S. Trebst,[1] M. Troyer,[2] and Z. Wang[1]
[1]*Microsoft Research, Station Q, University of California, Santa Barbara, CA 93106, USA*
[2]*Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland*
(Dated: July 6, 2011)

Galois conjugation relates unitary conformal field theories (CFTs) and topological quantum field theories (TQFTs) to their non-unitary counterparts. Here we investigate Galois conjugates of quantum double models, such as the Levin-Wen model. While these Galois conjugated Hamiltonians are typically non-Hermitian, we find that their ground state wave functions still obey a generalized version of the usual code property (local operators do not act on the ground state manifold) and hence enjoy a generalized topological protection. The key question addressed in this paper is whether such non-unitary topological phases can also appear as the ground states of Hermitian Hamiltonians. Specific attempts at constructing Hermitian Hamiltonians with these ground states lead to a loss of the code property and topological protection of the degenerate ground states. Beyond this we rigorously prove that no local change of basis (IV.5) can transform the ground states of the Galois conjugated doubled Fibonacci theory into the ground states of a topological model whose Hermitian Hamiltonian satisfies Lieb-Robinson bounds. These include all gapped local or quasi-local Hamiltonians. A similar statement holds for many other non-unitary TQFTs. One consequence is that the "Gaffnian" wave function cannot be the ground state of a gapped fractional quantum Hall state.

PACS numbers: 05.30.Pr, 73.43.-f

## I. INTRODUCTION

*Galois conjugation*, by definition, replaces a root of a polynomial by another one with identical algebraic properties. For example, $i$ and $-i$ are Galois conjugate (consider $z^2+1=0$) as are $\phi=\frac{1+\sqrt{5}}{2}$ and $-\frac{1}{\phi}=\frac{1-\sqrt{5}}{2}$ (consider $z^2-z-1=0$), as well as $\sqrt[3]{2}$, $\sqrt[3]{2}e^{2\pi i/3}$, and $\sqrt[3]{2}e^{-2\pi i/3}$ (consider $z^3-2=0$). In physics Galois conjugation can be used to convert non-unitary conformal field theories (CFTs) to unitary ones, and vice versa. One famous example is the non-unitary Yang-Lee CFT, which is Galois conjugate to the Fibonacci CFT $(G_2)_1$, the even (or integer-spin) subset of $su(2)_3$.

In statistical mechanics non-unitary conformal field theories have a venerable history.[1,2] However, it has remained less clear if there exist physical situati… models can provide a useful descr… physics of a quantum mechanical … conjugation typically destroys the … Hamiltonian. Some non-Hermitian … prisingly have totally real spectrum… in the study of $PT$-invariant one-… some Galois conjugate many-body … seen to open the door a crack to … models. Another situation, which h… interest, is the question whether no… scribe 1D edge states of certain 2D … graphic for the bulk). In particular, … sion on whether or not the "Gaffnian… the ground state for a *gapped* fracti… state albeit with a non-unitary "Yan… edge.[5–7] We conclude that this is no… ing the possible scope of non-unitar… chanics.

We reach this conclusion quite i… is the investigation of Galois conjug…

Abelian Levin-Wen model.[8] This model, which is also called "DFib", is a topological quantum field theory (TQFT) whose states are string-nets on a surface labeled by either a trivial or "Fibonacci" anyon. From this starting point, we give a rigorous argument that the "Gaffnian" ground state cannot be locally conjugated to the ground state of any topological phase, within a Hermitian model satisfying Lieb-Robinson (LR) bounds[9] (which includes but is not limited to gapped local and quasi-local Hamiltonians).

Lieb-Robinson bounds are a technical tool for local lattice models. In relativistically invariant field theories, the speed of light is a strict upper bound to the velocity of propagation. In lattice theories, the LR bounds provide a similar upper bound by a velocity called the LR velocity, but in contrast to the relativistic case there can be some exponentially small "leakage" outside the light-cone in the lattice case. The Lieb-Robinson

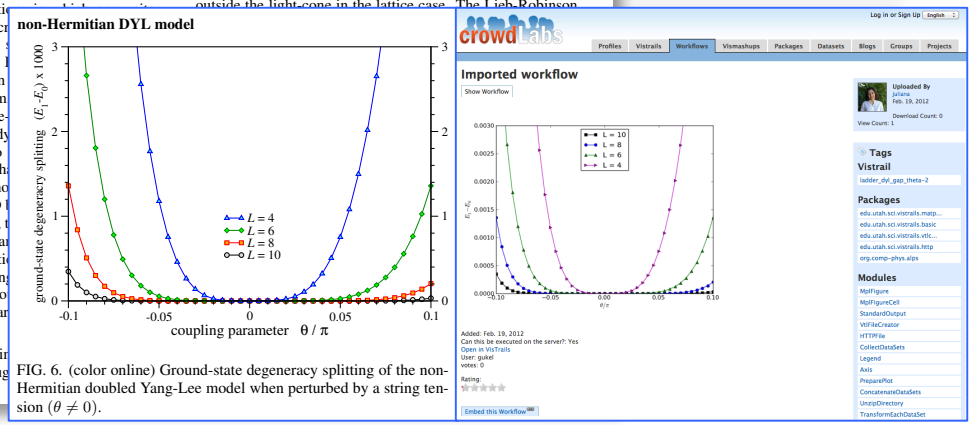
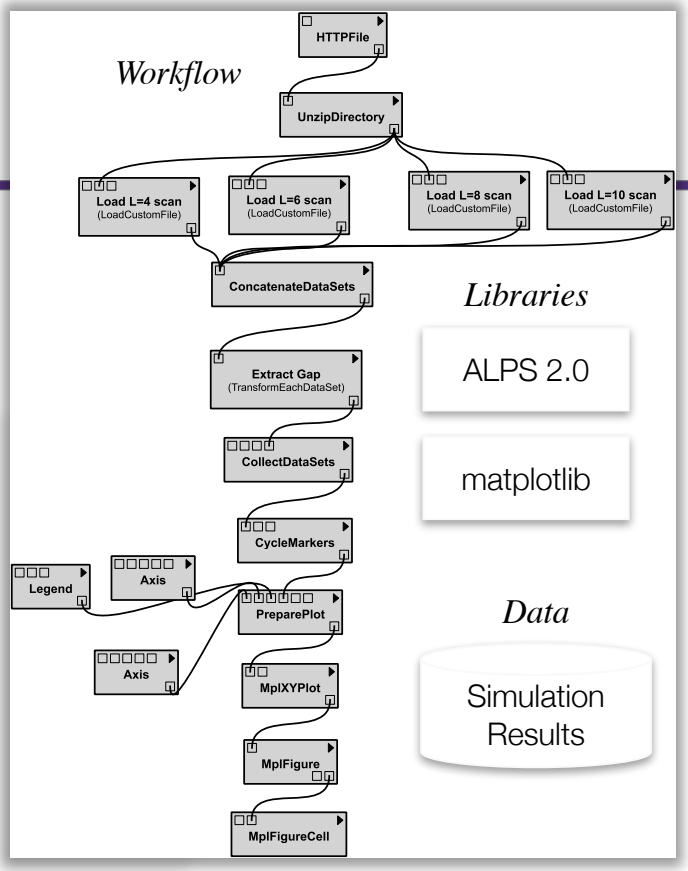



FIG. 6. (color online) Ground-state degeneracy splitting of the non-Hermitian doubled Yang-Lee model when perturbed by a string tension ($\theta \neq 0$).




*Reproducible result in paper and on the Web*




*Provenance*

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

# *Best Practices*

***Which tools support document linkage?***

- *VisTrails*
- *Jupyter notebooks: http://jupyter.org/*
  - *Literate programming*
  - *E.g.: typical signal processing tasks on strain time-series data associated with the LIGO GW150914 data release from the LIGO Open Science Center (LOSC) - by Kyle Cranmer*

    *http://mybinder.org/repo/cranmer/ligo-binder;*

    *http://app.mybinder.org/2902494521/notebooks/GW150914_tutorial.ipynb*

# Sharing

Provide infrastructure to *upload*, *archive*, and *share* data related to the experiment

Related to *data preservation: citing* data is also important!

Two types of sharing

- *Archival*
- *Hosted Execution*

# *Best Practices*

***Which tools support archival?***

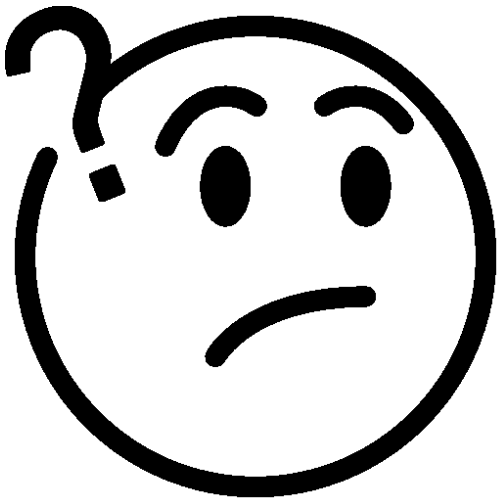
- *GitHub and Bitbucket (code)*
- *myExperiment (workflows)*
- *Dataverse: http://dataverse.org/*

***Which tools support hosted execution?***

- *crowdLabs*

# So… which tool should I use?

http://repromatch.engineering.nyu.edu



# Summary: Best Practices

1. Code Sharing and Version Control
   **Always** use a version control system!
   For code: *git* is highly recommended
   For scientific workflows: VisTrails captures workflow evolution provenance

2. Replicable Computational Environment
   Try making your code and experiment portable!
   For Python virtualization: virtualenv
   For Python builds: pip
   For general virtualization: Vagrant, Docker, ReproZip

3. Shareable Data
   Remember: science is incremental! We learn from others!
   For code: GitHub, Bitbucket
   For data: Dataverse, Mendeley Data, figshare

# 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.



Google


GORDON AND BETTY MOORE FOUNDATION


ALFRED P. SLOAN FOUNDATION

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

고맙습니다

Merci

Thank you

Obrigada

благодаря

Kiitos

**धन्यवाद**

Tack

Danke

*Ευχαριστω*

Bedankt