Analysis and Visualization of Urban Data

Claudio T. Silva
Tandon School of Engineering
Center for Data Science
Center for Urban Science + Progress
Courant Institute for Mathematical Sciences
New York University

Joint work with Juliana Freire, Huy Vo, Carlos Dietrich, Harish Doraiswamy, Fernando Chirigati, Theo Damoulas, Nivan Ferreira, Masayo Otta, Kien Pham, Jorge Poco, Luciano Barbosa, Marcos Vieira, Marcos Lage, Joao Comba, Luis Gustavo Nonato, Luc Wilson, Heidi Werner, Muchan Park, Jonathas Costa, and many others

Funded by grants/gifts from Moore and Sloan Foundations, NSF, NASA, DOE, MLB.com, AT&T, and IBM

NYU TANDON SCHOOL OF ENGINEERING

VIDA@NYU (2011-)

Close to 40 members

- 5 TT faculty @ NYU / 1 TT @ CUNY
- 2 Research faculty
- 6 Postdocs
- 4 Research engineers
- ~20 PhD students
- Constant stream of visiting faculty, researchers and students

PUBLIC gems-uff / noworkflow

PUBLIC VisTrails / VisTrails

PUBLIC ViDA-NYU / reprozip

PUBLIC ViDA-NYU / birdvis

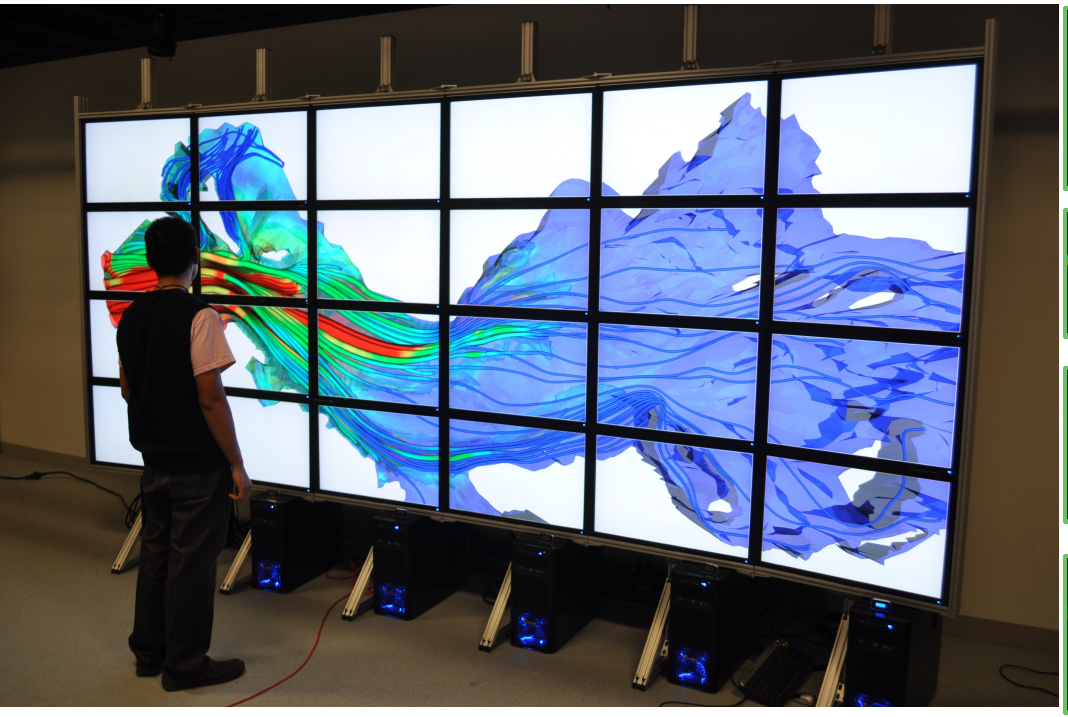
ViDA-NYU / ache

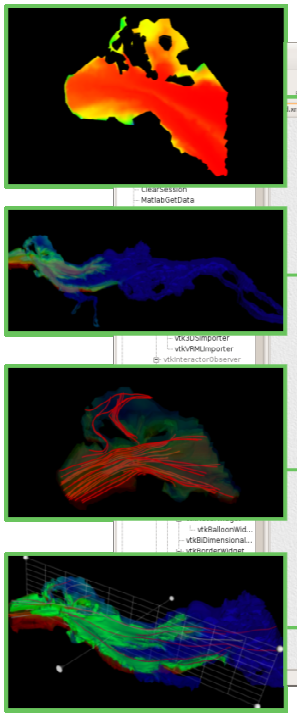
NYU TANDON SCHOOL OF ENGINEERING

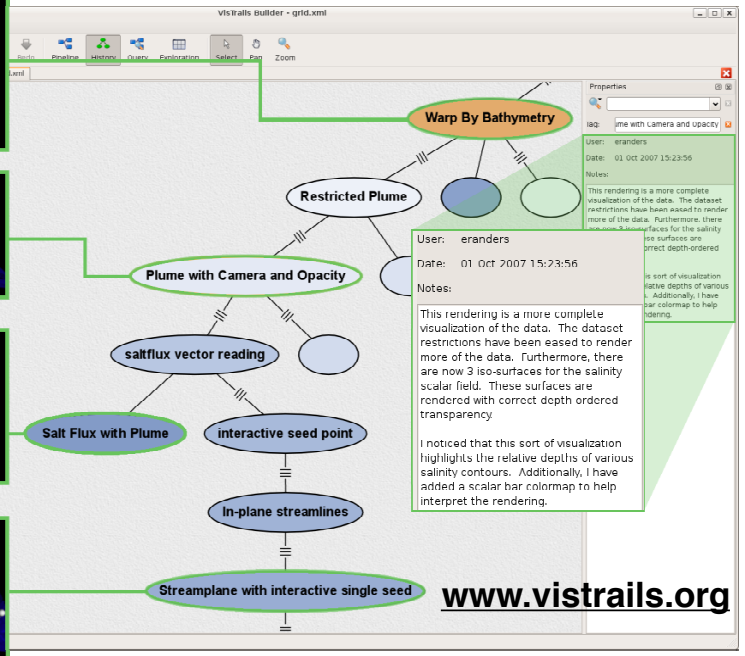
Our Research at the ViDA Center

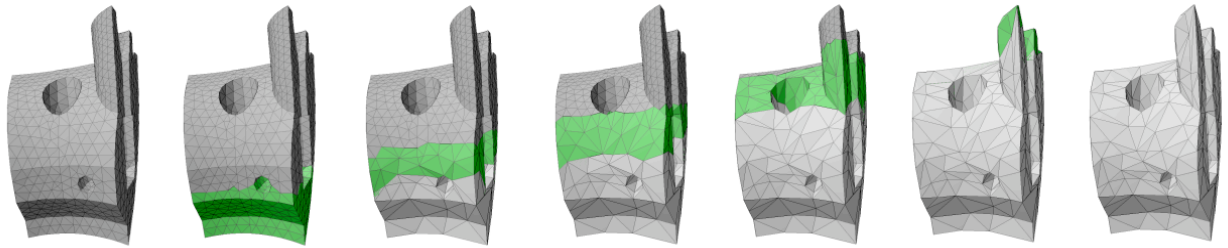
- Empower a *wide range of users* to *explore* the vast repositories of urban data
  - Data-savvy analysts, domain experts, policy makers and citizens
- Address issues at the different stages of the data lifecycle
- Key ingredients (that we work on)
  - Finding information on the Web, hidden, dark and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

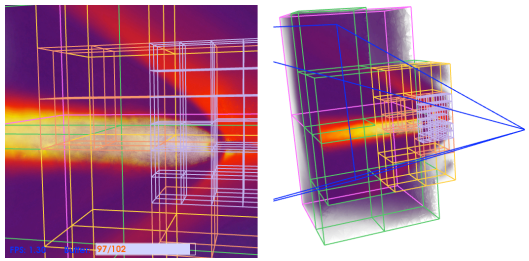
NYU TANDON SCHOOL OF ENGINEERING

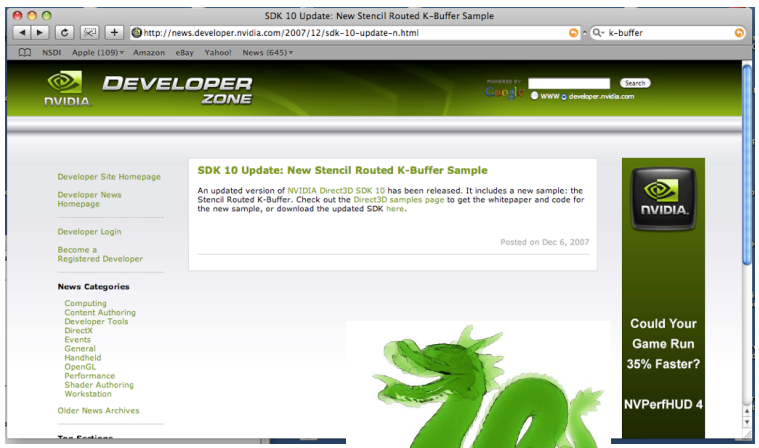


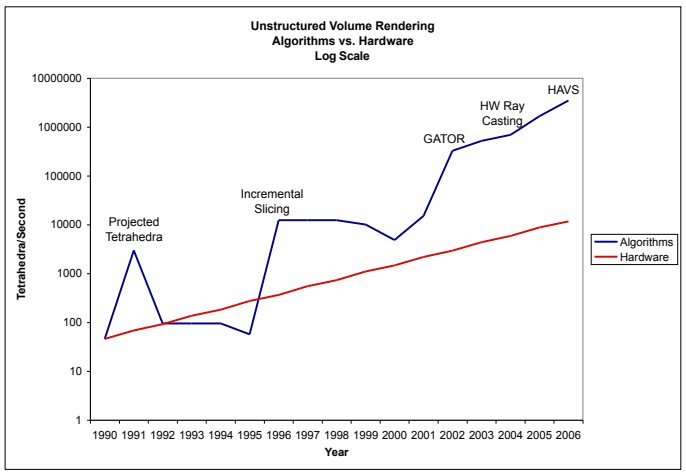


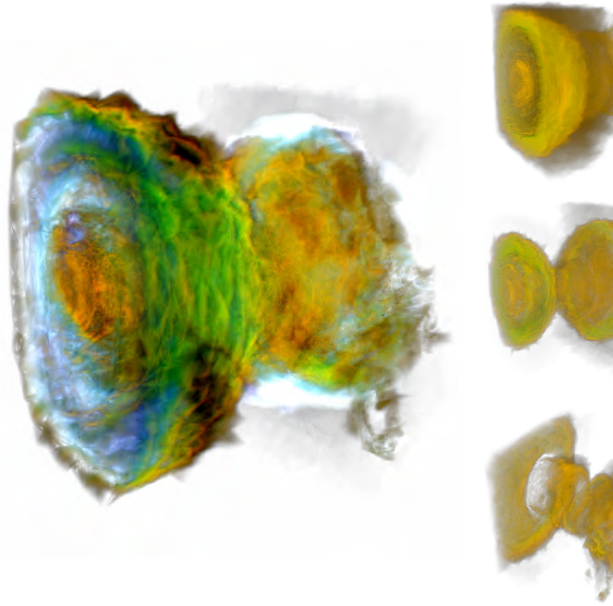

vistrails Builder - grid.xml
Warp By Bathymetry
Restricted Plume
Plume with Camera and Opacity
saltflux vector reading
Salt Flux with Plume
interactive seed point
In-plane streamlines
Streamplane with interactive single seed
User: eranders
Date: 01 Oct 2007 15:23:56
Notes:
This rendering is a more complete visualization of the data. The dataset restrictions have been eased to render more of the data. Furthermore, there are now 3 iso-surfaces for the salinity scalar field. These surfaces are rendered with correct depth ordered transparency.
I noticed that this sort of visualization highlights the relative depths of various salinity contours. Additionally, I have added a scalar bar colormap to help interpret the rendering.
www.vistrails.org






SDK 10 Update: New Stencil Routed K-Buffer Sample
DEVELOPER ZONE
SDK 10 Update: New Stencil Routed K-Buffer Sample
Could Your Game Run 35% Faster?
NVPerfHUD 4


Unstructured Volume Rendering
Algorithms vs. Hardware
Log Scale
Tetrahedra/Second
Year
Projected Tetrahedra
Incremental Slicing
GATOR
HW Ray Casting
HAVS
Algorithms
Hardware



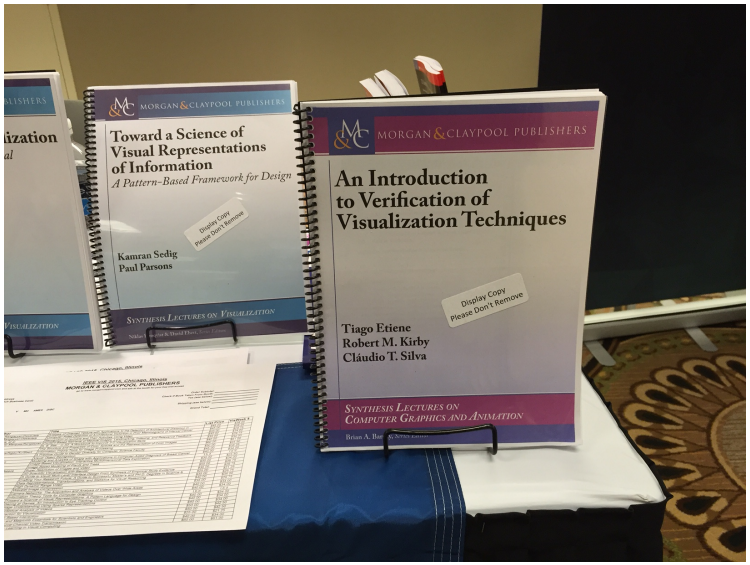
Verifiable Visualizations: Improving the Accuracy of your Visualizations

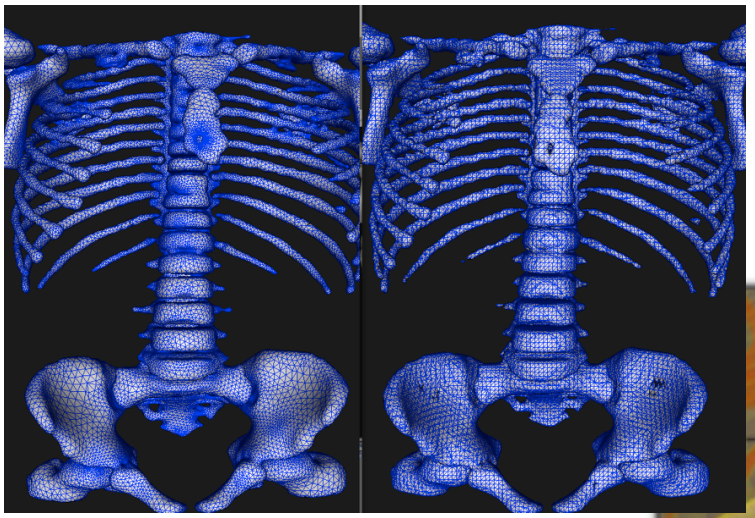
Verifying Volume Rendering Using Discretization Error Analysis

Tiago Etiene, Daniel Jönsson, Timo Ropinski, *Member*, *IEEE Computer Society*, Carlos Scheidegger, João L.D. Comba, Luis Gustavo Nonato, Robert M. Kirby, *Member*, *IEEE*, Anders Ynnerman, *Member*, *IEEE*, and Cláudio T. Silva, *Fellow*, *IEEE*

Abstract—We propose an approach for verification of volume rendering correctness based on an analysis of the volume rendering integral, the basis of most DVR algorithms. With respect to the most common discretization of this continuous model (Riemann summation), we make assumptions about the impact of parameter changes on the rendered results and derive convergence curves describing the expected behavior. Specifically, we progressively refine the number of samples along the ray, the grid size, and the pixel size, and evaluate how the errors observed during refinement compare against the expected approximation errors. We derive the theoretical foundations of our verification approach, explain how to realize it in practice, and discuss its limitations. We also report the errors identified by our approach when applied to two publicly available volume rendering packages.

Index Terms—Discretization errors, volume rendering, verifiable visualization, verification, testing





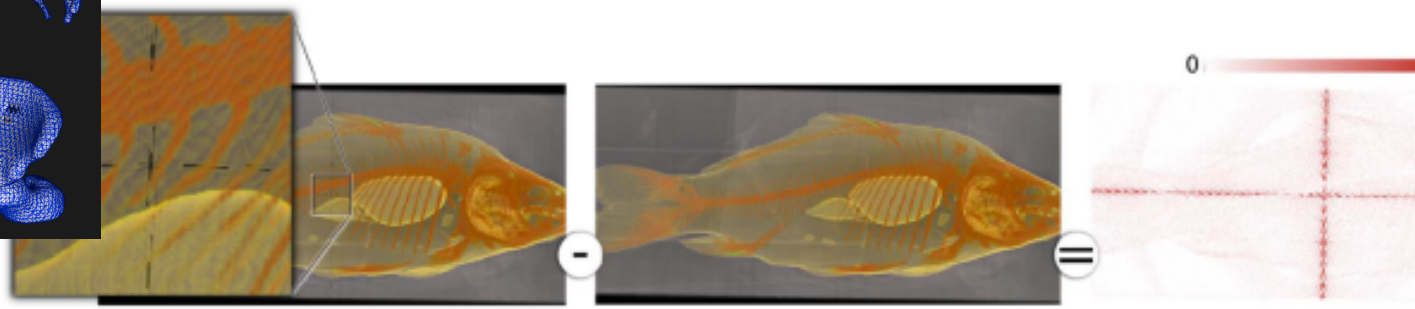


Fig. 10. A CT scan of a carp, rendered with VTK 5.6.1 and Fixed-Point Raycast Mapper (FP). On the left, we see the artifacts prevented FP convergence. In the middle, we see the results after fixing the issues that prevented convergence. The artifacts are On the right we see the difference image.

Urban Data Visualization

Joint work with **Juliana Freire, Huy Vo, Harish Doraiswamy,**
Fernando Chirigati, Theo Damoulas, Nivan Ferreira, Masayo Otta, Kien Pham, Jorge Poco, Luciano Barbosa, Marcos Vieira, Marcos Lage, Joao Comba, Luis Gustavo Nonato, Luc Wilson, Heidi Werner, Muchan Park, Jonathas Costa, and many others

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

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

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

Big Urban Data: Success Stories

OneBusAway

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

http://onebusaway.org

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

Big Urban Data: Success Stories

- Michael Flowers @ NYC

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

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

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

(2) Compared with 5 years of fire data

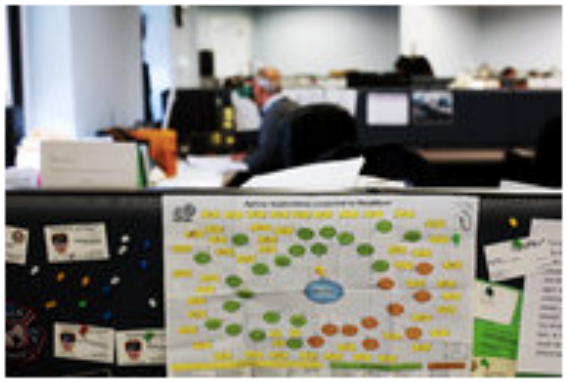
(3) Created a prediction system

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

Todd Heisler/The New York Times

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

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."

Big Urban Data: What is hard?

Infrastructure

Condition, operations

Environment

Meteorology, pollution, noise, flora, fauna

People

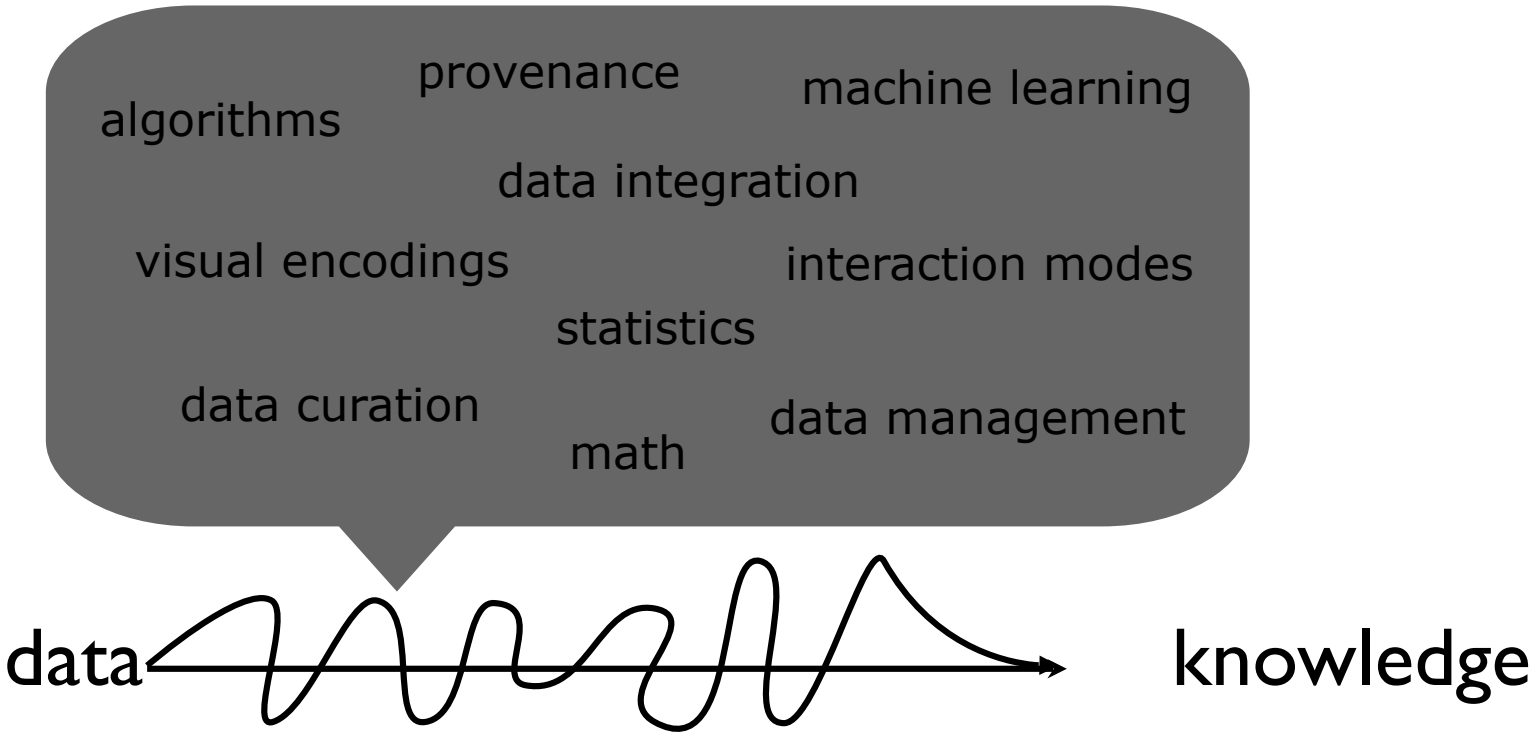
Relationships, economic activities, health, nutrition, opinions, …

- City components interact in complex ways
- Need to look at the city *data exhaust* to understand these interactions
- Processes occur over time and space

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

What dataset should we start with?

What dataset should we start with?

Firefox File Edit View History Bookmarks Tools Window Help — 68% Wed 11:42 PM Claudio Silva

Utah teapot - Wikipedia, th...

https://en.wikipedia.org/wiki/Utah_teapot

Search

Create account Not logged in Talk Contributions Log in

Article Talk | Read Edit View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
العربية
Català
Čeština

Utah teapot

From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(November 2014)*

The **Utah teapot** or **Newell teapot** is a 3D computer model that has become a standard reference object (and something of an in-joke) in the computer graphics community. It is a mathematical model of an ordinary teapot of fairly simple shape, that appears solid, cylindrical and partially convex. A teapot primitive is considered the equivalent of a "hello, world" program, as a way to create an easy 3D scene with a somewhat complex model acting as a basic geometry reference for scene and light setup. Many programming libraries even have functions dedicated to drawing teapots.[1]

The teapot model was created in 1975 by early computer graphics researcher Martin Newell, a member of the pioneering graphics program at the University of Utah.[2]

A modern rendering of the Utah teapot model.

Contents [hide]

History [edit]

Newell needed a moderately simple mathematical model of a familiar object for his work. His wife Sandra Newell suggested modelling their tea service since they were sitting down for tea at the time. He got some graph paper and a pencil, and sketched the entire teapot by eye.[*citation needed*] Then he went back to the lab and edited bézier control points on a Tektronix storage tube, again by hand.[*citation needed*]

The teapot shape contains a number of elements that made it ideal for the graphics experiments of the time: it is round, contains saddle points, has a genus greater than zero because of the hole in the handle, can project a shadow on itself, and looks reasonable when displayed without a complex surface texture.

NYU TANDON SCHOOL OF ENGINEERING

What dataset should we start with?

Chrome File Edit View History Bookmarks People Window Help — 85% — Wed 11:12 PM — Claudio Silva

Stanford bunny - Wikipedia

https://en.wikipedia.org/wiki/Stanford_bunny

Create account | Not logged in | Talk | Contributions | Log in

Article | Talk | Read | Edit | View history | Search

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Create a book
Download as PDF
Printable version

Languages
Deutsch
فارسی
Français
Latviešu

Stanford bunny

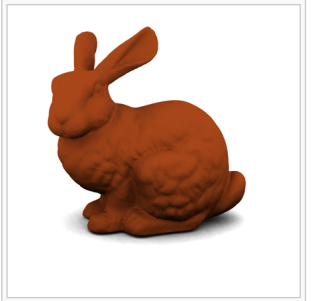
From Wikipedia, the free encyclopedia

This article **needs additional citations for verification**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(August 2013)*

The **Stanford bunny** is a computer graphics 3D test model developed by Greg Turk and Marc Levoy in 1994 at Stanford University. It is available for free download in various formats.[1]

The *bunny* consists of data describing 69,451 triangles determined by 3D scanning a ceramic figurine of a rabbit. The data can be used to test various graphics algorithms; including polygonal simplification, compression, and surface smoothing. By today's standards in terms of geometric complexity and triangle count, it is considered a simple model. There are a few problems with this dataset that can occur in any 3D scan data. The problems are that it is manifold connected, and that it has holes in the data (some due to scanning limits and some due to the object being hollow). Though being "problems", they provide a more realistic input for any algorithm that is benchmarked with this bunny.

The model was originally available in .ply (polygons) file format with 4 different resolutions, 69,451 polygons being the highest.



The Stanford bunny rendered in YafRay

Rabbits and hares portal

See also [edit]

- 3D modeling
- Utah teapot
- Suzanne (3D model)
- Cornell box
- List of common 3D test models

References [edit]

1. ^ Stanford bunny model downloads

External links [edit]

- Original article on the technique
- The Stanford 3D Scanning Repository provides the S

Zippered polygon meshes from range images
G Turk, M Levoy - Proceedings of the 21st annual conference on …, 1994 - dl.acm.org
Abstract Range imaging offers an inexpensive and accurate means for digitizing the shape of three-dimensional objects. Because most objects self occlude, no single range image suffices to describe the entire object. We present a method for combining a collection of ...
Cited by 1394 Related articles All 41 versions Cite Saved More

NYU TANDON SCHOOL OF ENGINEERING

Taxi drivers petition NYC for fare hike over soaring gas prices

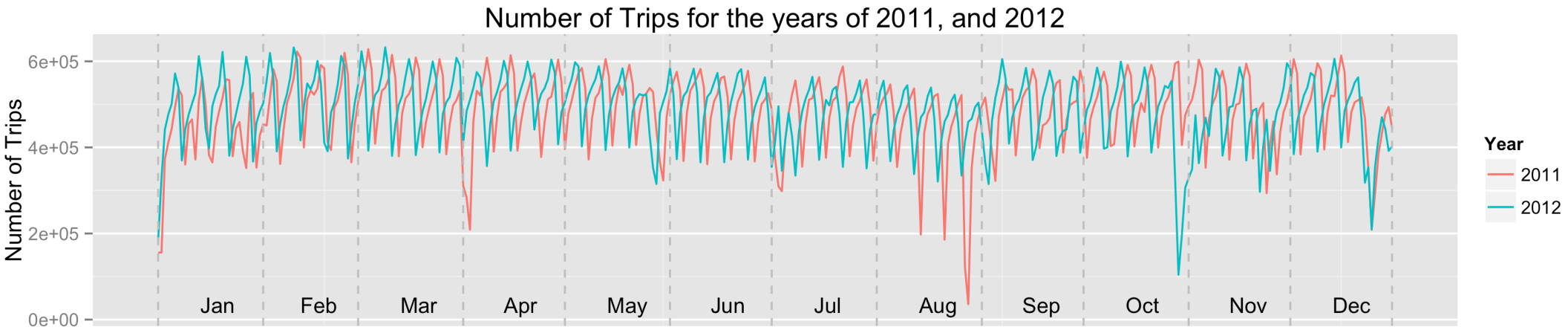
BY PETE DONOHUE / DAILY NEWS STAFF WRITER

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

What dataset should we start with?

ailynews.com/new-york

Exploring Urban Data: NYC Taxis



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

  "How the taxi fleet activity varies during weekdays?"

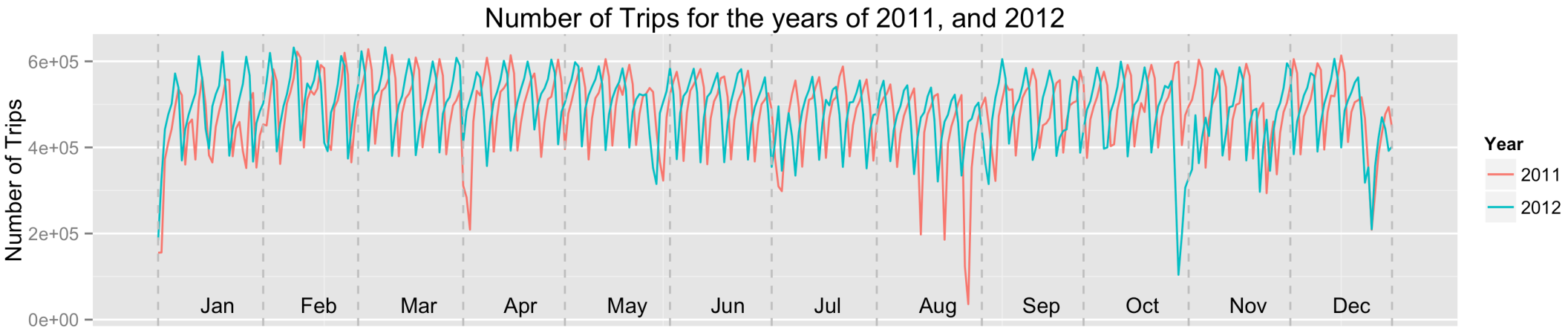
  "What is the average trip time from Midtown to the airports during weekdays?''

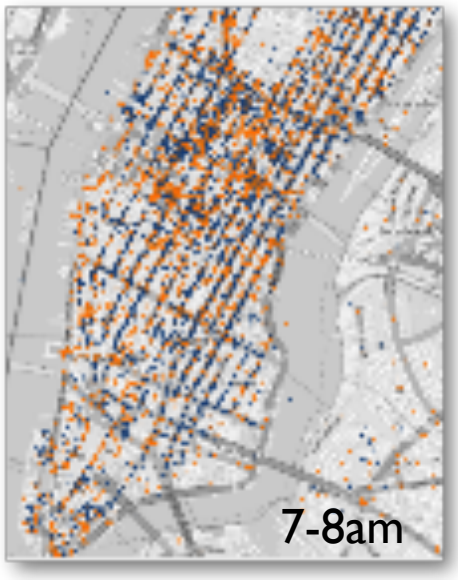
  "How was activity in Midtown affected during a presidential visit?''

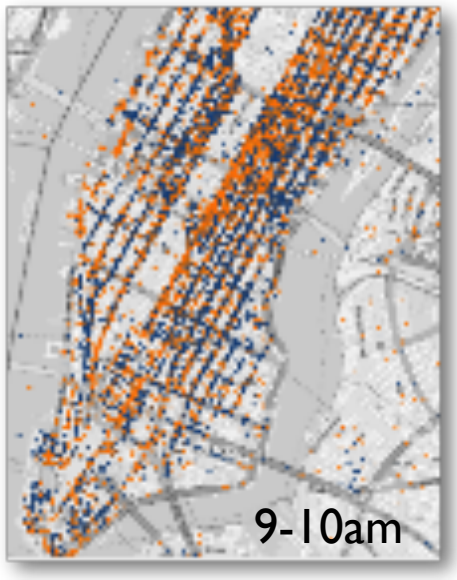
  "How did the movement patterns change during Sandy?"

  "Where are the popular night spots?"

Exploring Urban Data: NYC Taxis



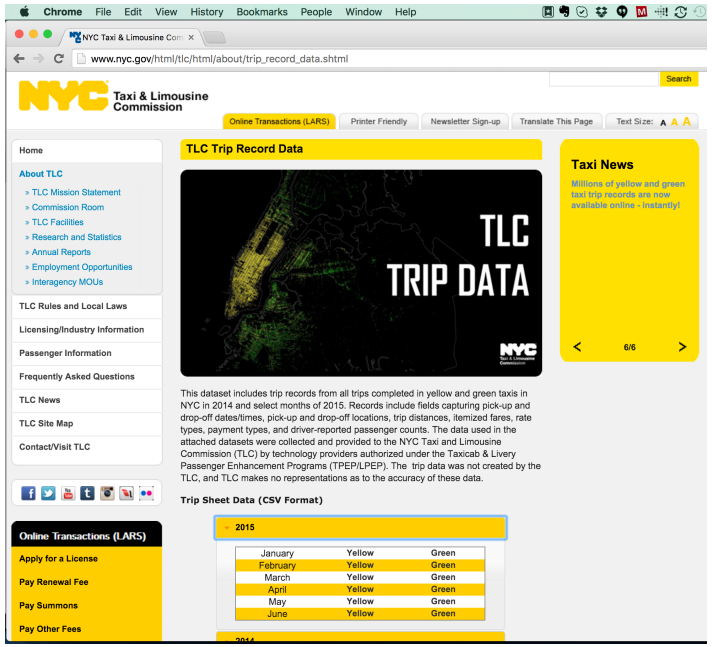




NYU TANDON SCHOOL OF ENGINEERING

Exploring Taxi Data: Challenges

- Data are *big*: ~500k trips/day - 780 million trips in 5 years
- Government, policy makers and scientists are unable to explore the *whole* data
- Data are *complex*:
  - *spatio-temporal:* pick up + drop off
  - *trip attributes*: e.g., distance traveled, cost, tip
- Too many data slices to examine
- Our goals: Design a *usable* interface, efficiently support *interactive + exploratory* queries

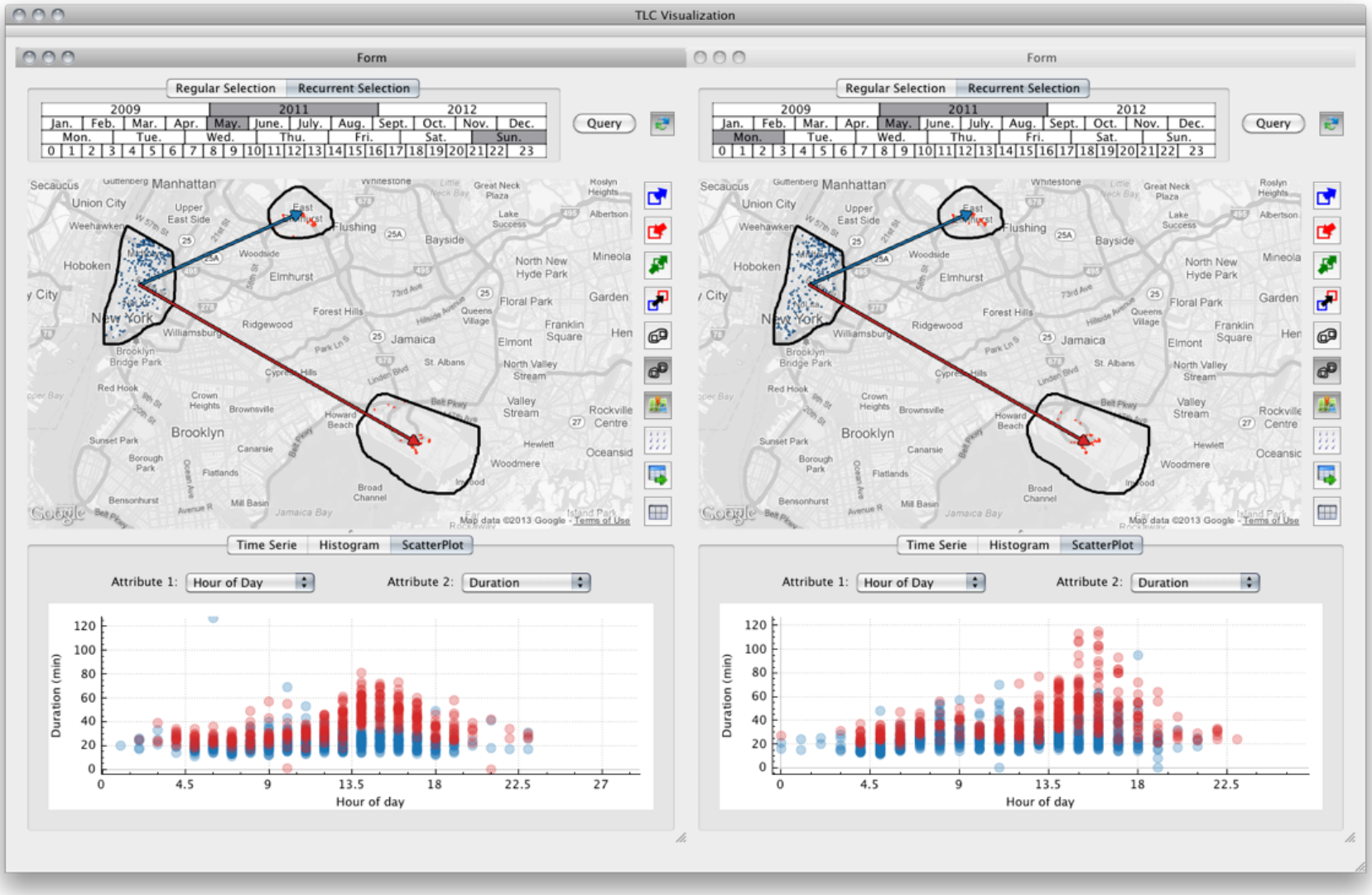


TaxiVis: Visually Exploring NYC Taxi Data

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

[Ferreira et al., IEEE TVCG 2013]

Interactive Visual Exploration of NYC Taxi Records



Source code at: https://github.com/ViDA-NYU/TaxiVis

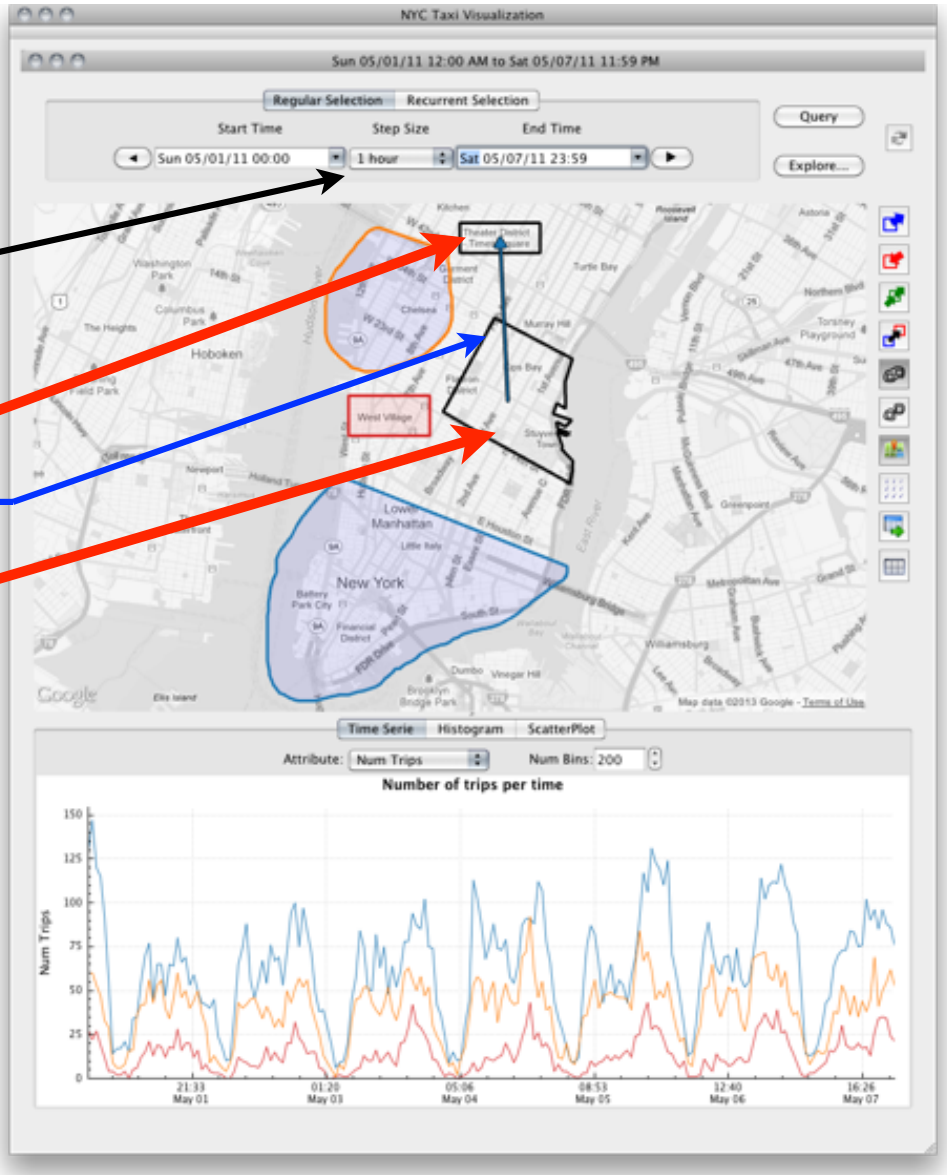
NYU TANDON SCHOOL OF ENGINEERING

Visual Data Selection

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

Visual Data Selection

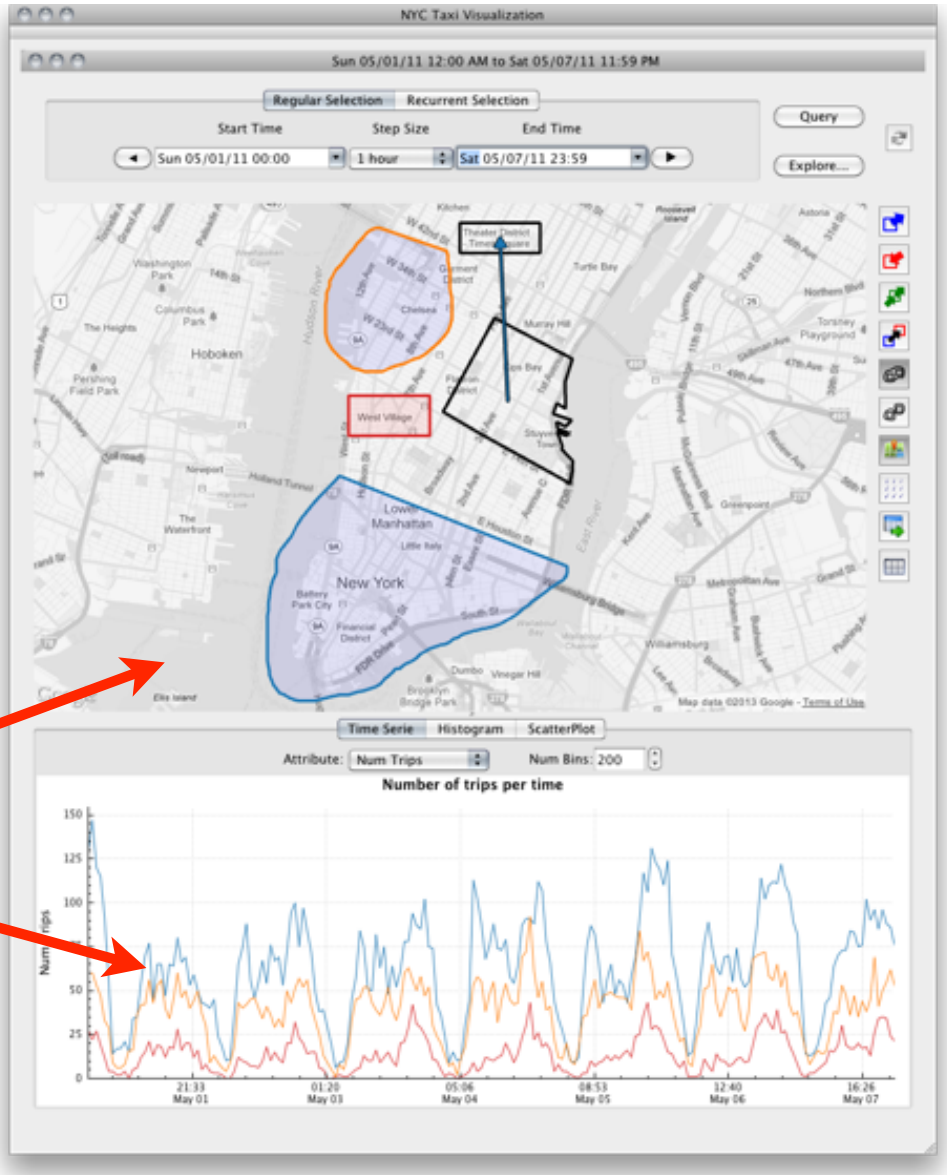
```
SELECT  *
FROM    trips
WHERE   pickup_time   in (5/1/11,5/7/11)
                AND
        dropoff_loc   in "Times Square"
                AND
        pickup_loc    in "Gramercy"
```



Visual Data Exploration

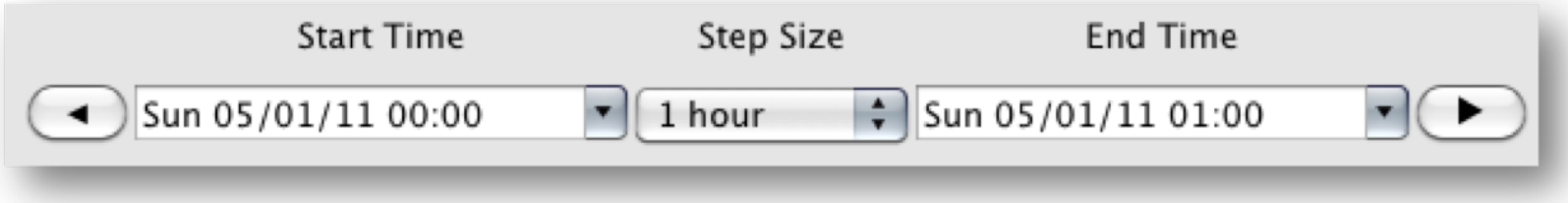
```
SELECT  *
FROM    trips
WHERE   pickup_time   in (5/1/11,5/7/11)
                AND
        dropoff_loc   in "Times Square"
                AND
        pickup_loc    in "Gramercy"
```

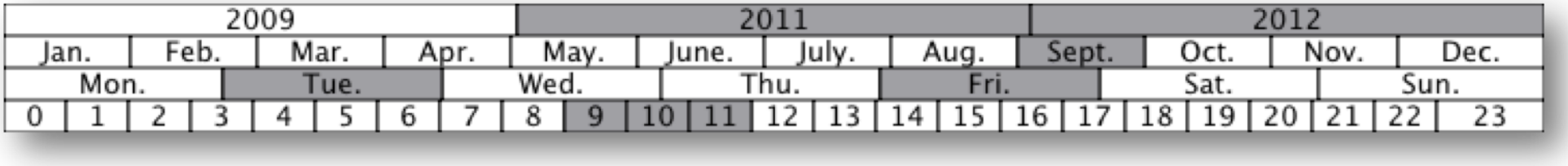
Interactively explore data through the map view and plot widgets



Selecting Time – Temporal Constraints

Time interval





Recurrent time patterns

Video: http://vgc.poly.edu/projects/taxivis/resources/vast2013_submission_270.mp4

TaxiVis: The Plumbing

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

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

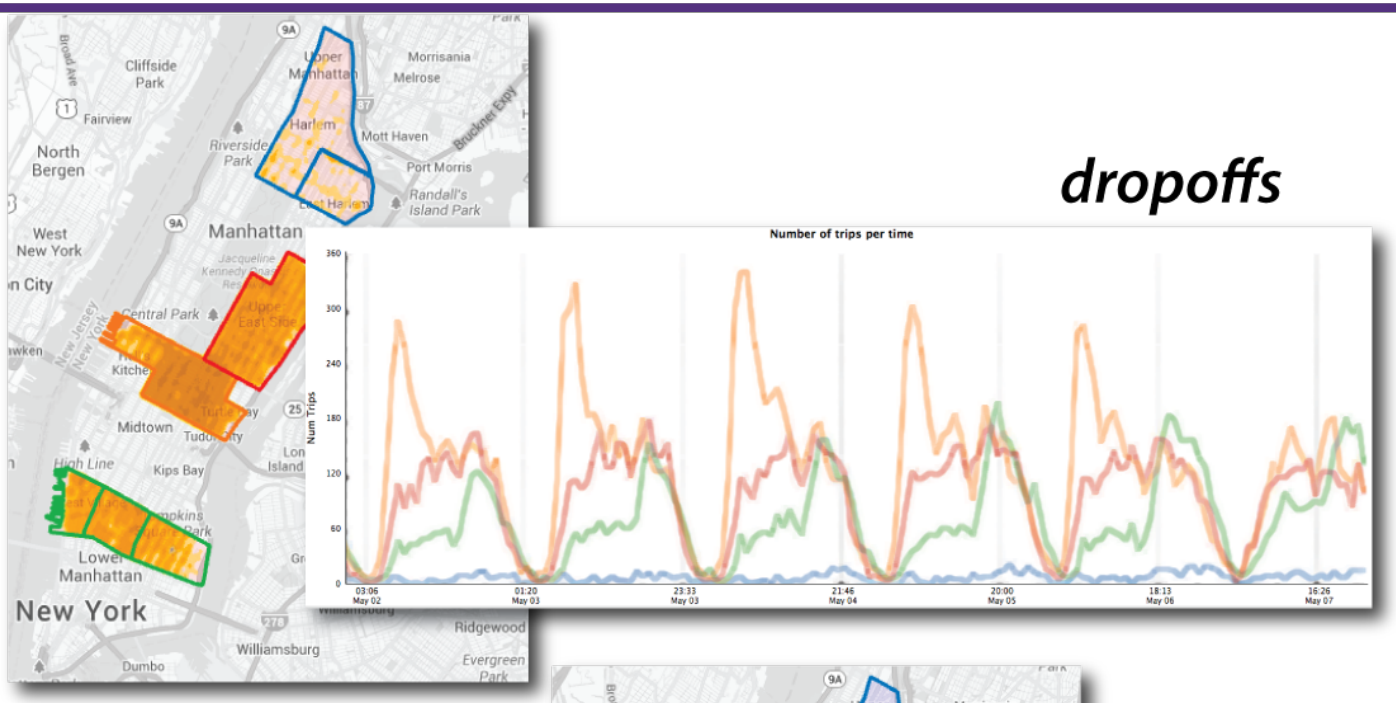
Supporting Interactive Queries

Solution 1: Spatio-temporal index based on kd-trees

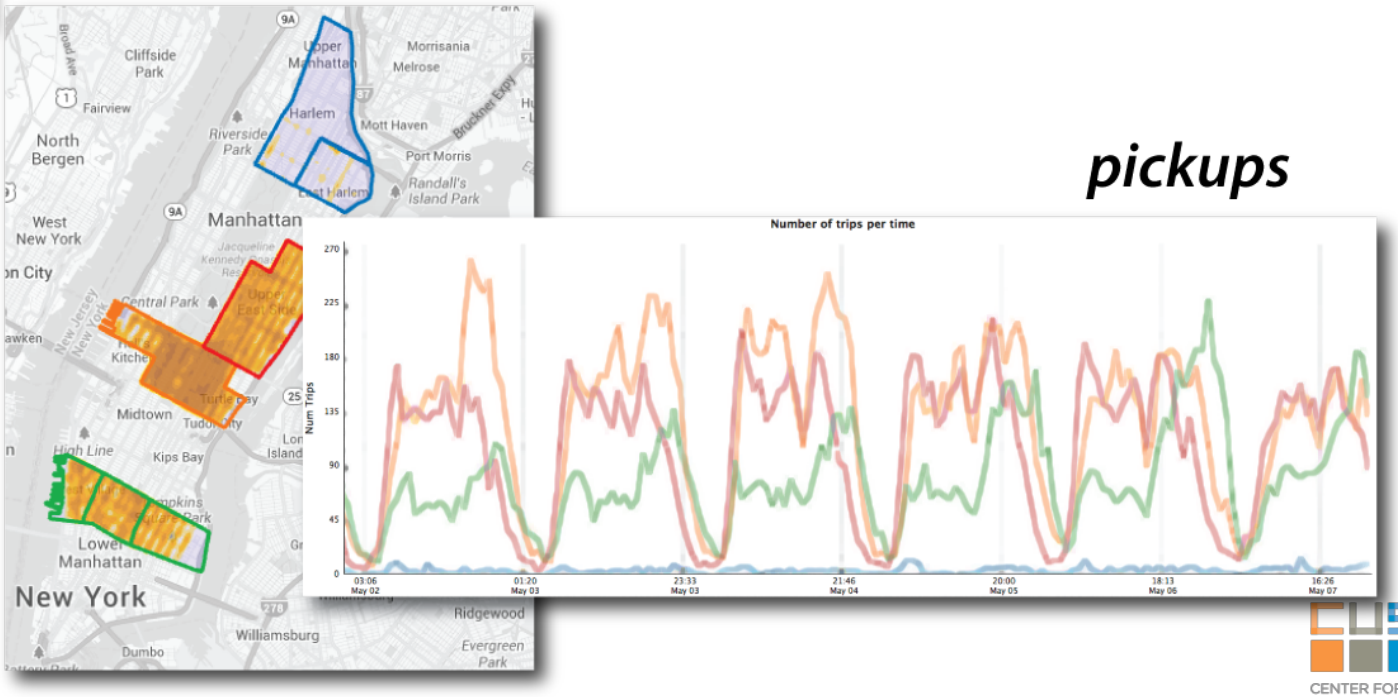
- Can index multiple attributes!
- Tree nodes store kd-tree
- A leaf node represents a k-dimensional node that satisfies the path constraints

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

TaxiVis: Comparing Neighborhoods



dropoffs

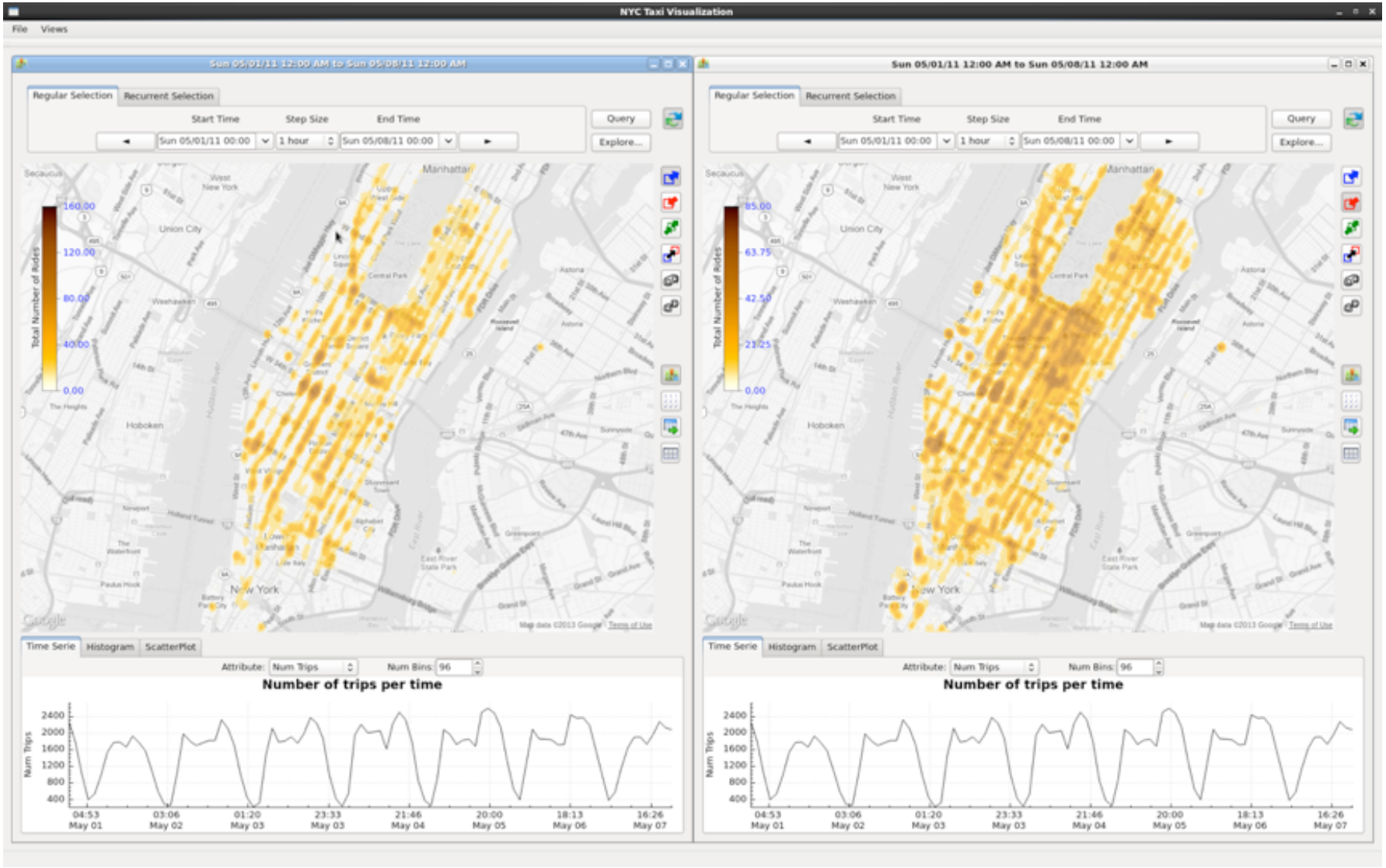


pickups

NYU TANDON SCHOOL OF ENGINEERING

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

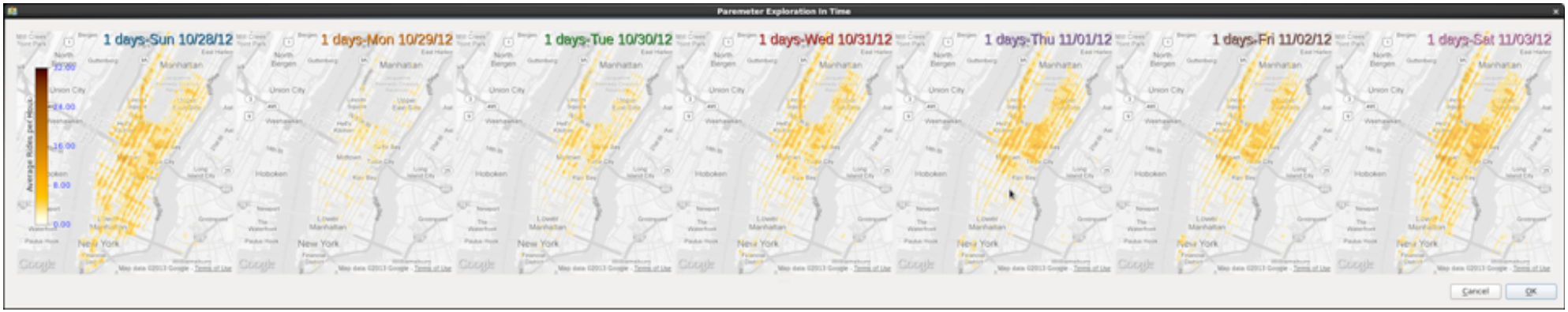
Pickups vs Dropoffs

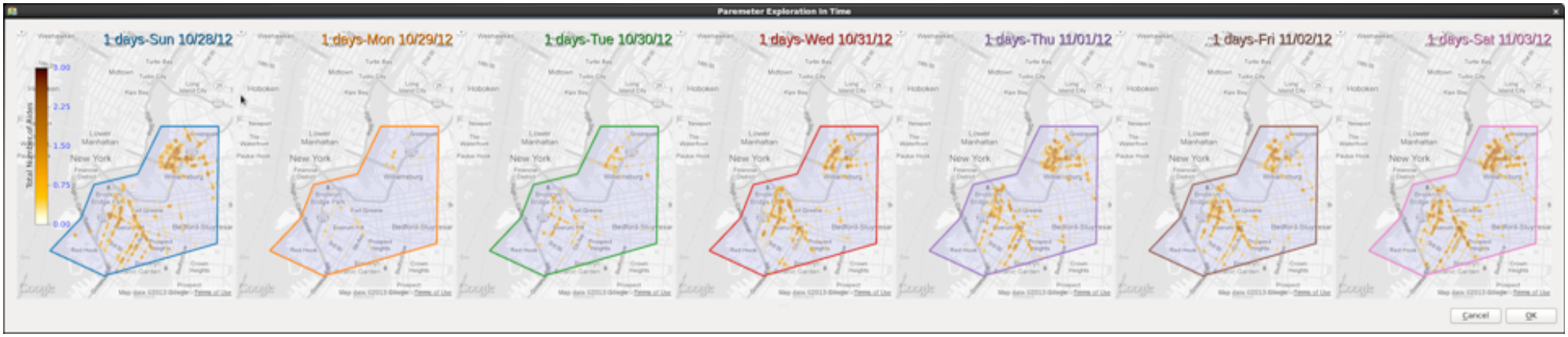


Multiple Coordinated Views

NYU TANDON SCHOOL OF ENGINEERING

The Week of Sandy

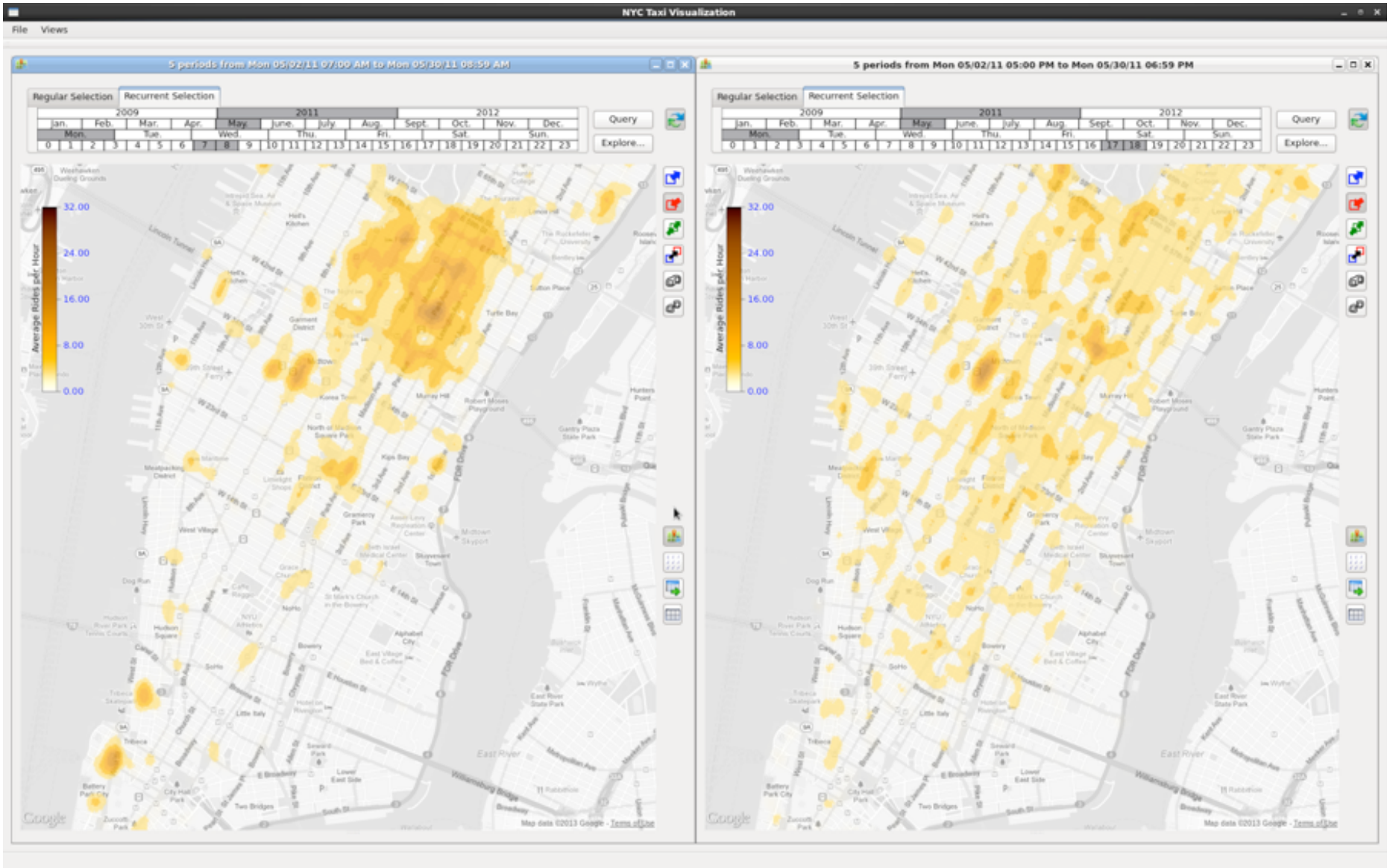




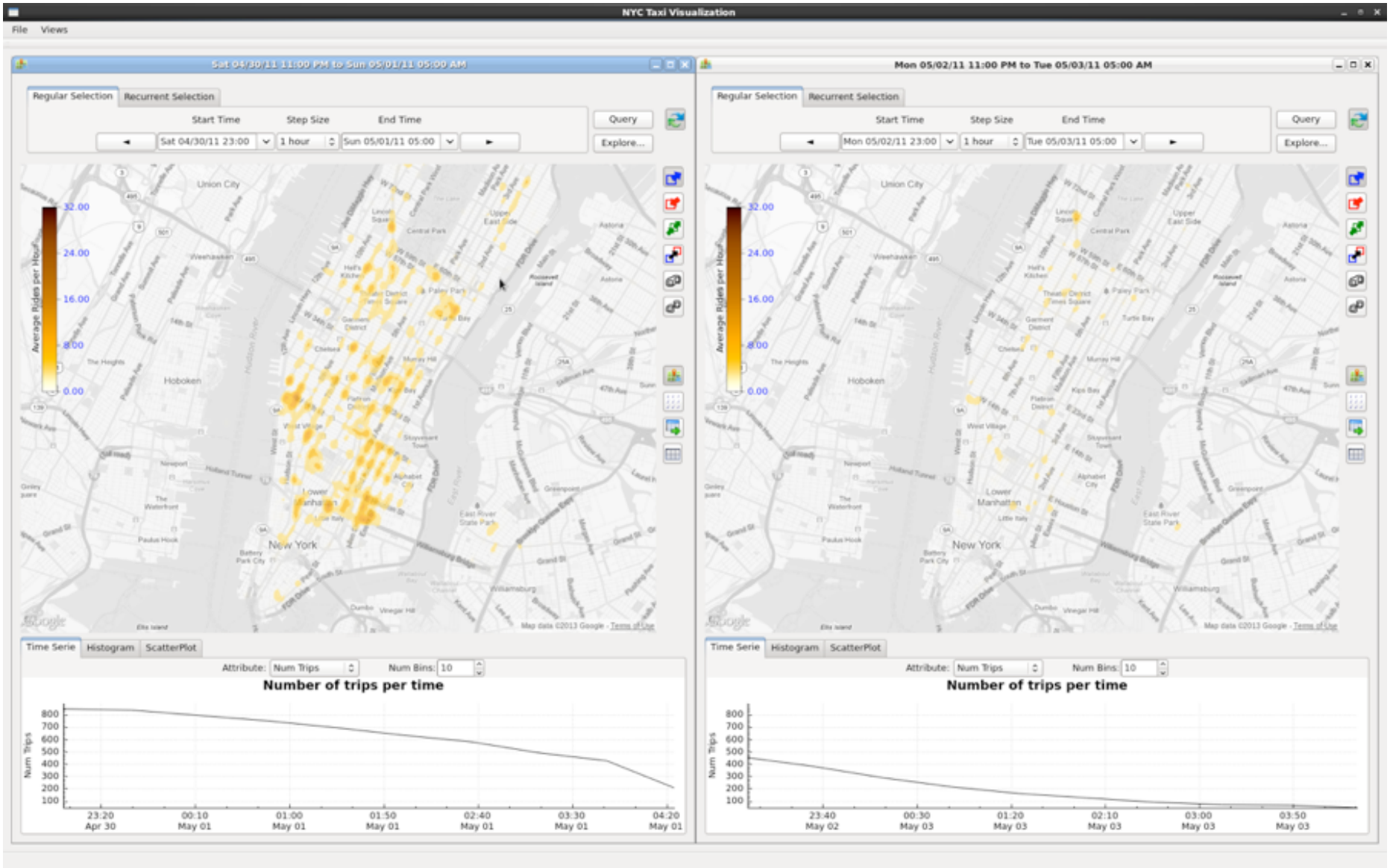
Time Exploration

NYU TANDON SCHOOL OF ENGINEERING

Dropoffs Before vs. After Work

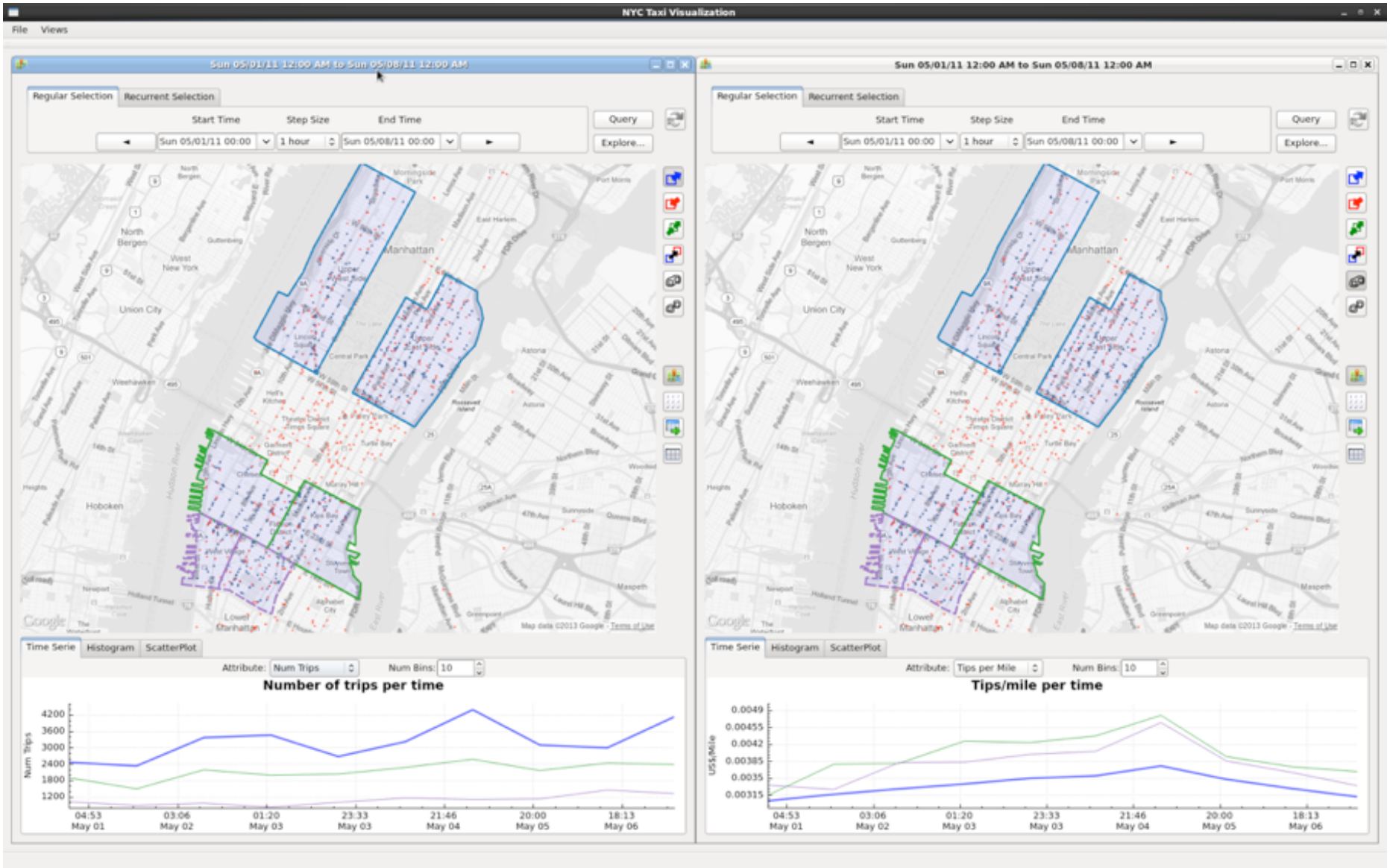


Night Life Saturday vs. Monday



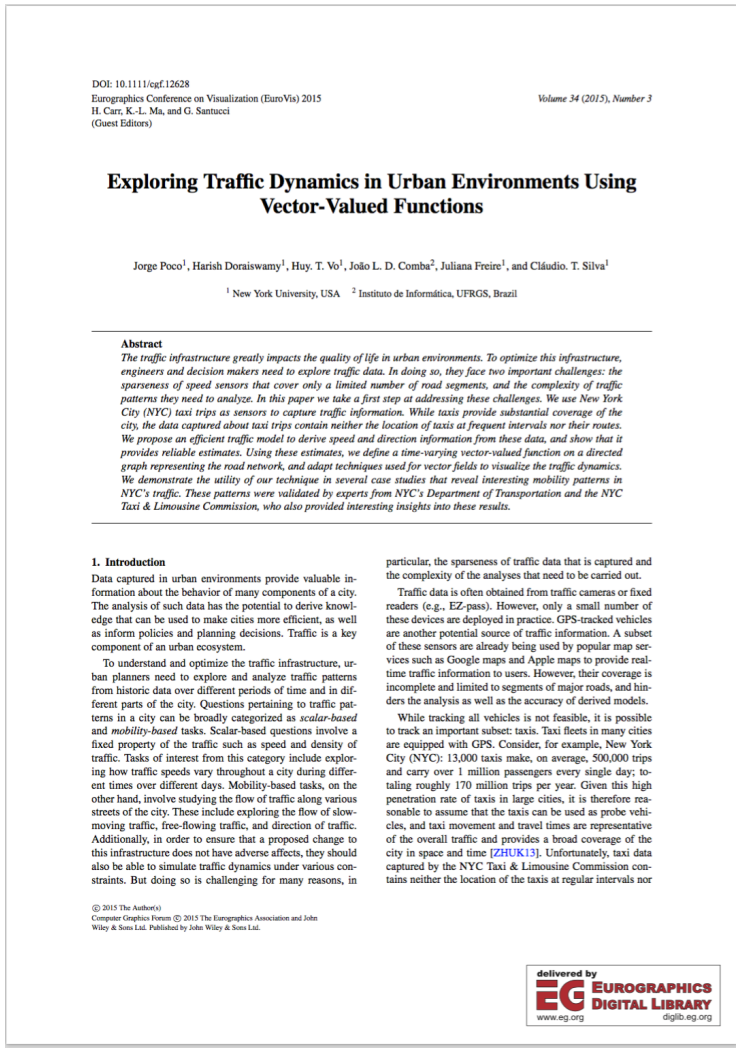
NYU TANDON SCHOOL OF ENGINEERING

Tips Vary by Neighborhoods



NYU TANDON SCHOOL OF ENGINEERING

A Taxi over 24 hours



DOI: 10.1111/cgf.12628
Eurographics Conference on Visualization (EuroVis) 2015
H. Carr, K.-L. Ma, and G. Santucci
(Guest Editors)
Volume 34 (2015), Number 3

Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions

Jorge Poco[1], Harish Doraiswamy[1], Huy. T. Vo[1], João L. D. Comba[2], Juliana Freire[1], and Cláudio. T. Silva[1]

[1] New York University, USA [2] Instituto de Informática, UFRGS, Brazil

Abstract

The traffic infrastructure greatly impacts the quality of life in urban environments. To optimize this infrastructure, engineers and decision makers need to explore traffic data. In doing so, they face two important challenges: the sparseness of speed sensors that cover only a limited number of road segments, and the complexity of traffic patterns they need to analyze. In this paper we take a first step at addressing these challenges. We use New York City (NYC) taxi trips as sensors to capture traffic information. While taxis provide substantial coverage of the city, the data captured about taxi trips contain neither the location of taxis at frequent intervals nor their routes. We propose an efficient traffic model to derive speed and direction information from these data, and show that it provides reliable estimates. Using these estimates, we define a time-varying vector-valued function on a directed graph representing the road network, and adapt techniques used for vector fields to visualize the traffic dynamics. We demonstrate the utility of our technique in several case studies that reveal interesting mobility patterns in NYC's traffic. These patterns were validated by experts from NYC's Department of Transportation and the NYC Taxi & Limousine Commission, who also provided interesting insights into these results.

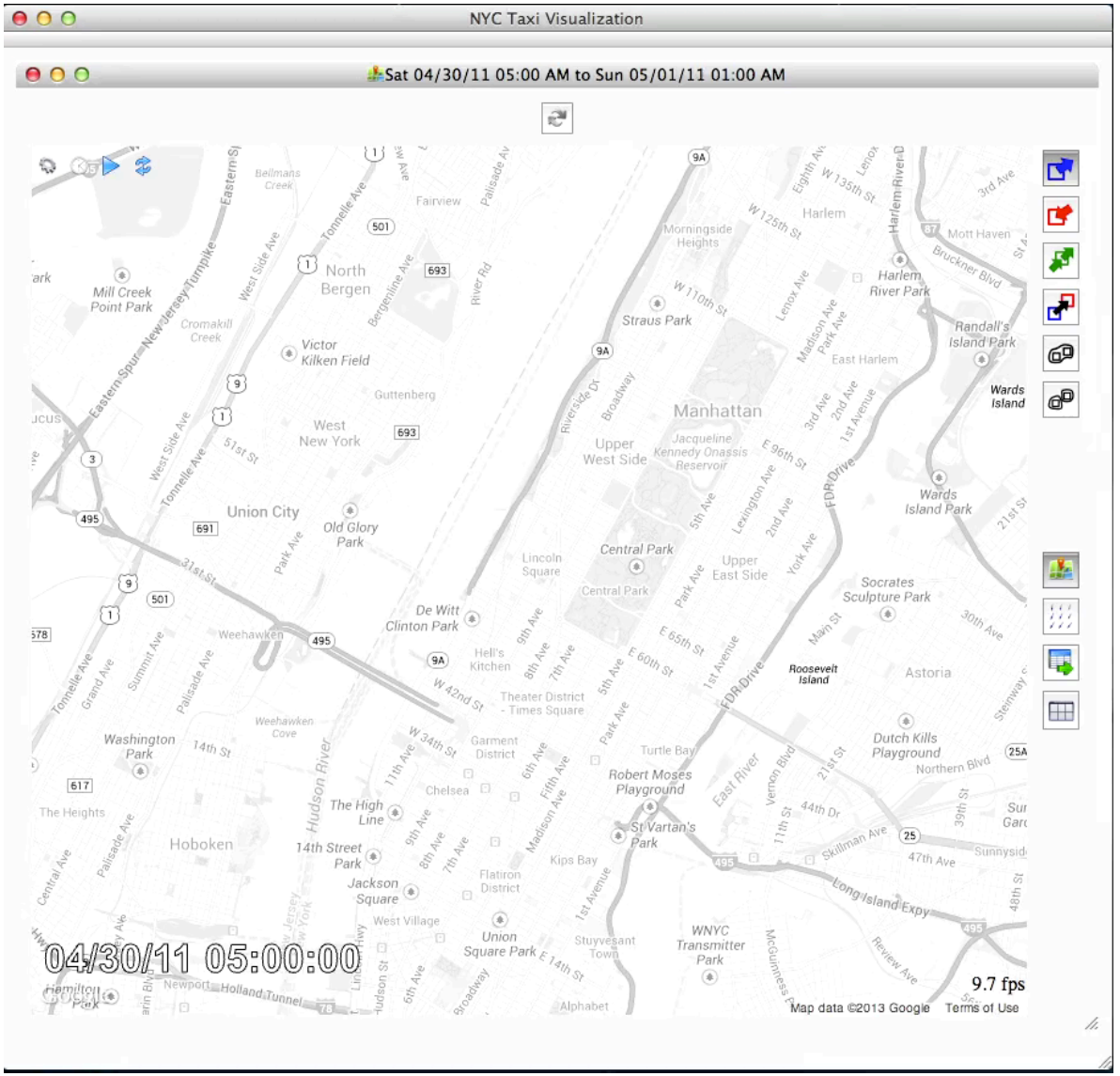
1. Introduction

Data captured in urban environments provide valuable information about the behavior of many components of a city. The analysis of such data has the potential to derive knowledge that can be used to make cities more efficient, as well as inform policies and planning decisions. Traffic is a key component of an urban ecosystem.

To understand and optimize the traffic infrastructure, urban planners need to explore and analyze traffic patterns from historic data over different periods of time and in different parts of the city. Questions pertaining to traffic patterns in a city can be broadly categorized as *scalar-based* and *mobility-based* tasks. Scalar-based questions involve a fixed property of the traffic such as speed and density of traffic. Tasks of interest from this category include exploring how traffic speeds vary throughout a city during different times over different days. Mobility-based tasks, on the other hand, involve studying the flow of traffic along various streets of the city. These include exploring the flow of slow-moving traffic, free-flowing traffic, and direction of traffic. Additionally, in order to ensure that a proposed change to this infrastructure does not have adverse affects, they should also be able to simulate traffic dynamics under various constraints. But doing so is challenging for many reasons, in particular, the sparseness of traffic data that is captured and the complexity of the analyses that need to be carried out.

Traffic data is often obtained from traffic cameras or fixed readers (e.g., EZ-pass). However, only a small number of these devices are deployed in practice. GPS-tracked vehicles are another potential source of traffic information. A subset of these sensors are already being used by popular map services such as Google maps and Apple maps to provide real-time traffic information to users. However, their coverage is incomplete and limited to segments of major roads, and hinders the analysis as well as the accuracy of derived models.

While tracking all vehicles is not feasible, it is possible to track an important subset: taxis. Taxi fleets in many cities are equipped with GPS. Consider, for example, New York City (NYC): 13,000 taxis make, on average, 500,000 trips and carry over 1 million passengers every single day; totaling roughly 170 million trips per year. Given this high penetration rate of taxis in large cities, it is therefore reasonable to assume that the taxis can be used as probe vehicles, and taxi movement and travel times are representative of the overall traffic and provides a broad coverage of the city in space and time [ZHUK13]. Unfortunately, taxi data captured by the NYC Taxi & Limousine Commission contains neither the location of the taxis at regular intervals nor



https://serv.cusp.nyu.edu/files/hvo/cab_hired_empty.mp4

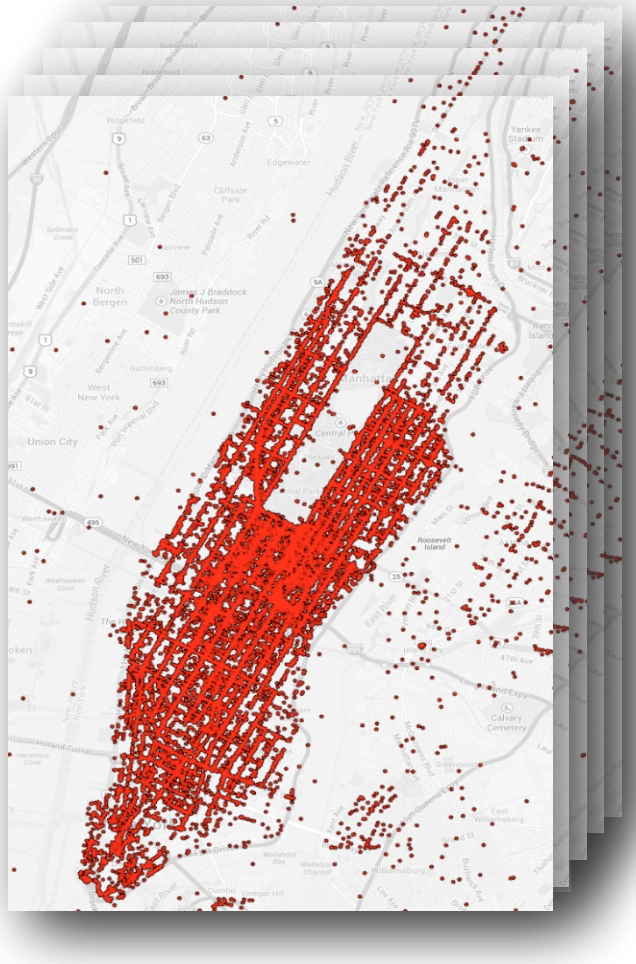
NYU TANDON SCHOOL OF ENGINEERING

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

Taxi Data: Too Many Slices

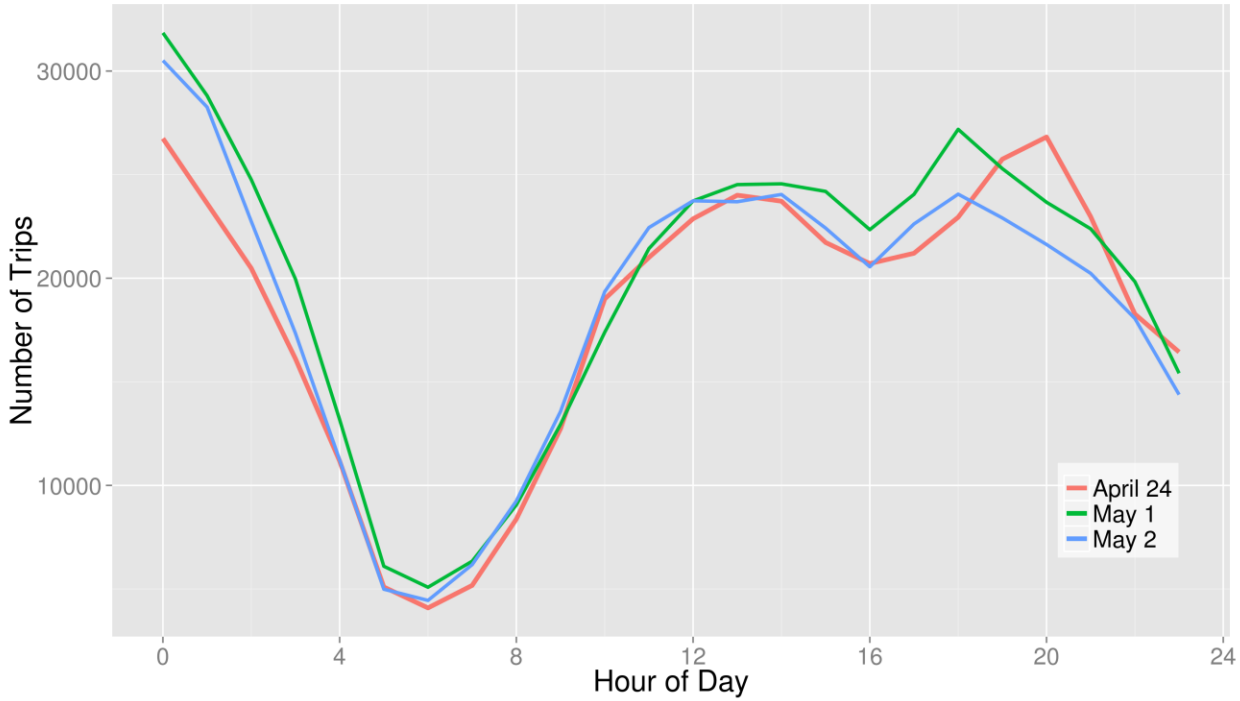
- 170 Million trips / year
- Spatial context
  - pick-up and drop-off locations
- Temporal attribute
  - pick-up and drop-off times

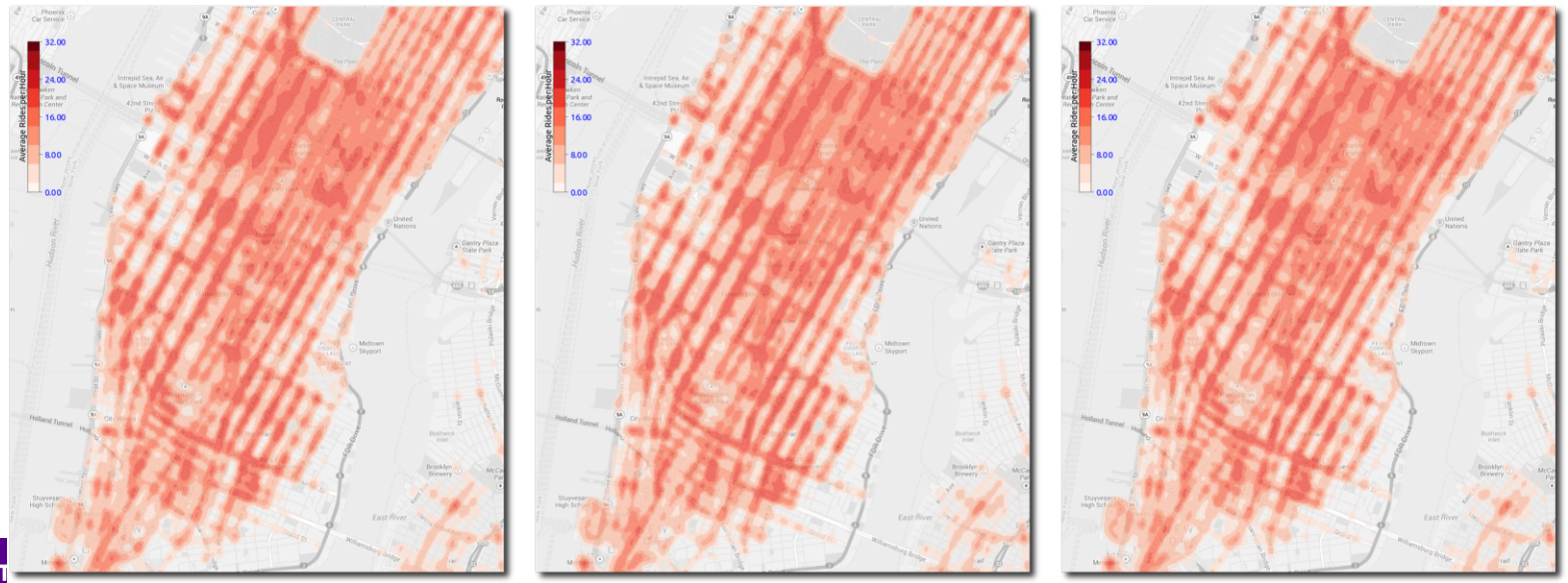
- Which slices are interesting?
- Can we guide users to interesting features in the data?



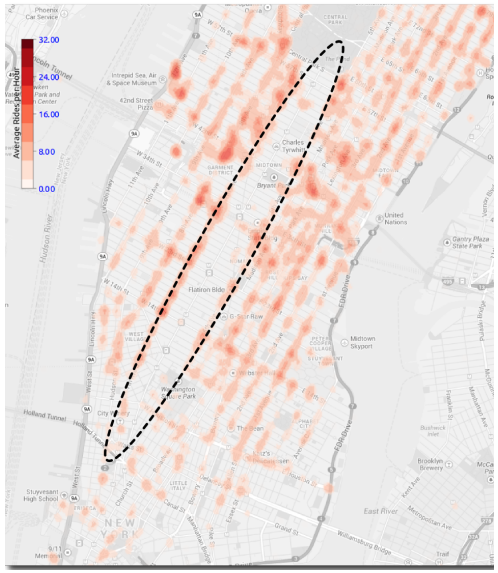
Reducing the Number of Slices

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

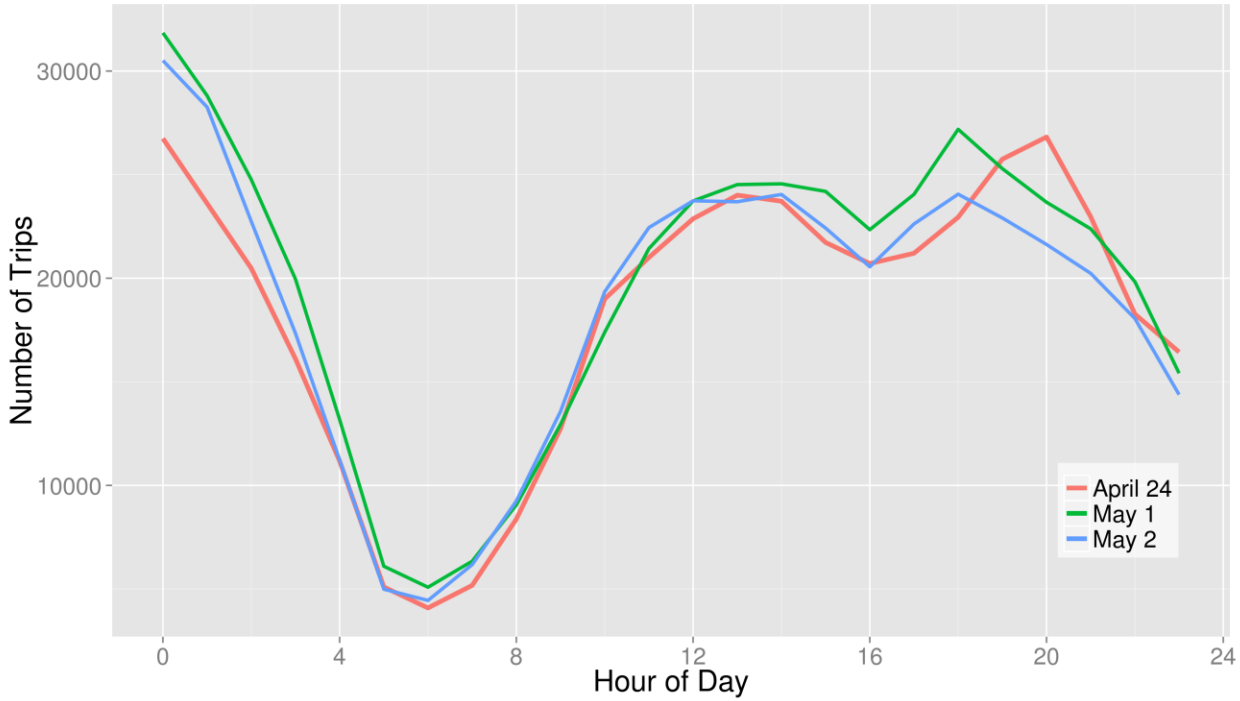


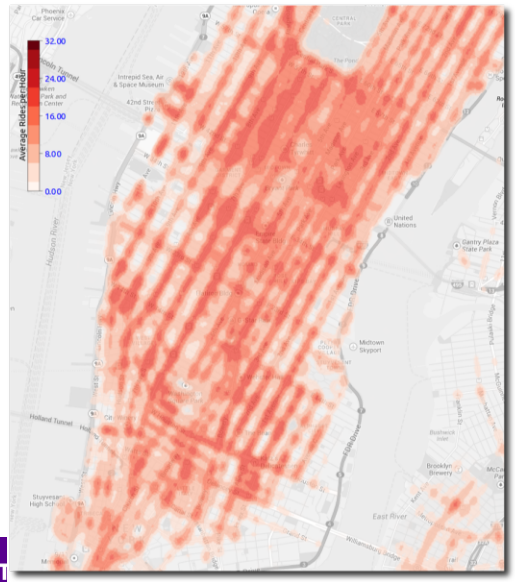


Miss Interesting Slices

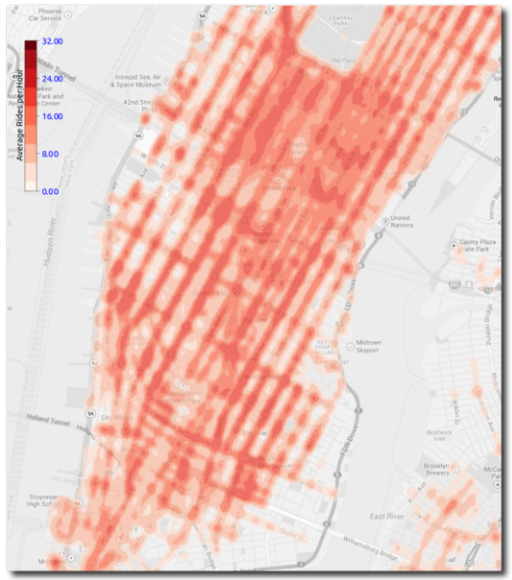


May 1 (8-9am)

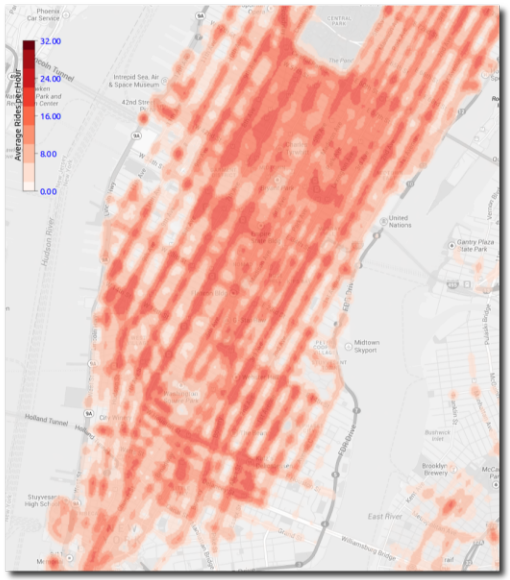




April 24



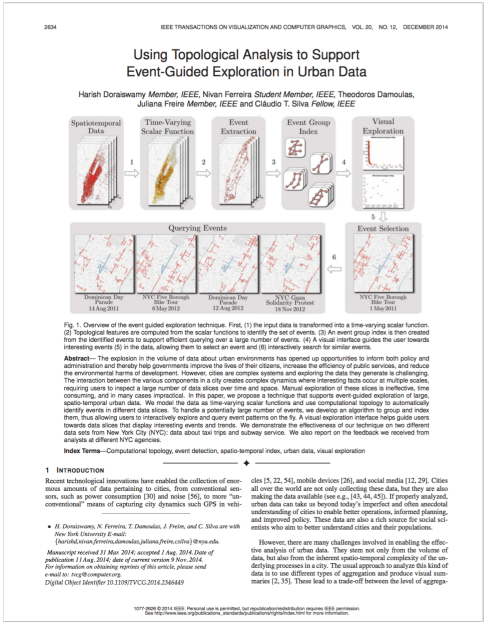
May 1



May 8

Finding Events at Multiple Granularities

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



2634 IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 20, NO. 12, DECEMBER 2014

Using Topological Analysis to Support Event-Guided Exploration in Urban Data

Harish Doraiswamy *Member, IEEE*, Nivan Ferreira *Student Member, IEEE*, Theodoros Damoulas, Juliana Freire *Member, IEEE* and Cláudio T. Silva *Fellow, IEEE*

Spatiotemporal Data | Time-Varying Scalar Function | Event Extraction | Event Group Index | Visual Exploration

Querying Events | Event Selection

Fig. 1. Overview of the event guided exploration technique. First, (1) the input data is transformed into a time-varying scalar function. (2) Topological features are computed from the scalar functions to identify the set of events. (3) An event group index is then created from the identified events to support efficient querying over a large number of events. (4) A visual interface guides the user towards interesting events (5) in the data, allowing them to select an event and (6) interactively search for similar events.

Abstract— The explosion in the volume of data about urban environments has opened up opportunities to inform both policy and administration and thereby help governments improve the lives of their citizens, increase the efficiency of public services, and reduce the environmental harms of development. However, cities are complex systems and exploring the data they generate is challenging. The interaction between the various components in a city creates complex dynamics where interesting facts occur at multiple scales, requiring users to inspect a large number of data slices over time and space. Manual exploration of these slices is ineffective, time consuming, and in many cases impractical. In this paper, we propose a technique that supports event-guided exploration of large, spatio-temporal urban data. We model the data as time-varying scalar functions and use computational topology to automatically identify events in different data slices. To handle a potentially large number of events, we develop an algorithm to group and index them, thus allowing users to interactively explore and query event patterns on the fly. A visual exploration interface helps guide users towards data slices that display interesting events and trends. We demonstrate the effectiveness of our technique on two different data sets from New York City (NYC): data about taxi trips and subway service. We also report on the feedback we received from analysts at different NYC agencies.

Index Terms—Computational topology, event detection, spatio-temporal index, urban data, visual exploration

1 INTRODUCTION

Recent technological innovations have enabled the collection of enormous amounts of data pertaining to cities, from conventional sensors, such as power consumption [30] and noise [56], to more "unconventional" means of capturing city dynamics such GPS in vehicles [5, 22, 54], mobile devices [26], and social media [12, 29]. Cities all over the world are not only collecting these data, but they are also making the data available (see e.g., [43, 44, 45]). If properly analyzed, urban data can take us beyond today's imperfect and often anecdotal understanding of cities to enable better operations, informed planning, and improved policy. These data are also a rich source for social scientists who aim to better understand cities and their populations.

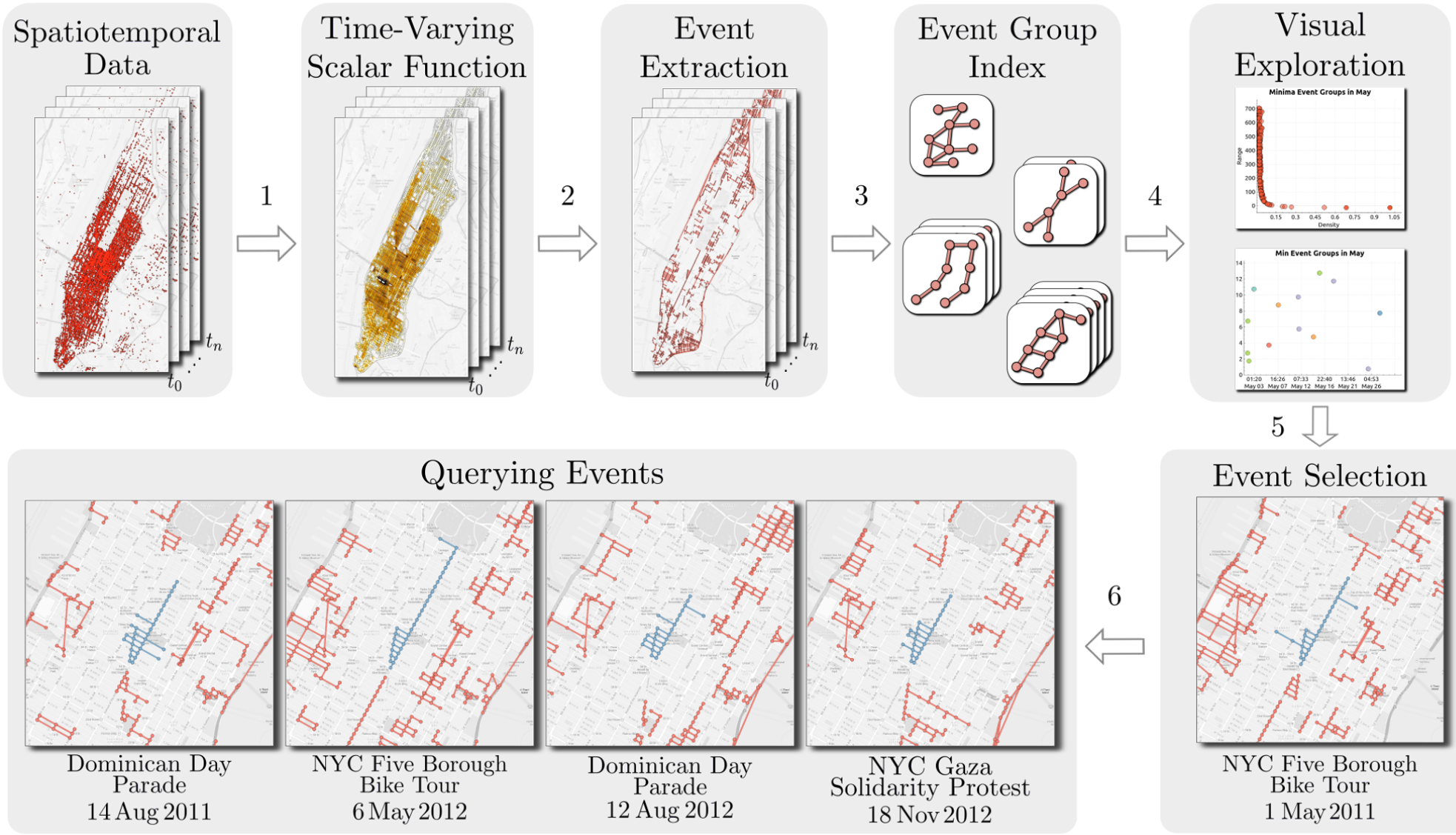
However, there are many challenges involved in enabling the effective analysis of urban data. They stem not only from the volume of data, but also from the inherent spatio-temporal complexity of the underlying processes in a city. The usual approach to analyze this kind of data is to use different types of aggregation and produce visual summaries [2, 35]. These lead to a trade-off between the level of aggrega-

• H. Doraiswamy, N. Ferreira, T. Damoulas, J. Freire, and C. Silva are with New York University. E-mail: {harish,nivan,ferreira,damoulas,juliana.freire,csilva}@nyu.edu.

Manuscript received 31 May 2014; accepted 1 Aug. 2014. Date of publication 11 Aug. 2014; date of current version 9 Nov. 2014. For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.
Digital Object Identifier 10.1109/TVCG.2014.2346449

[Doraiswamy et al., IEEE TVCG 2014]

Automatic Anomaly Detection

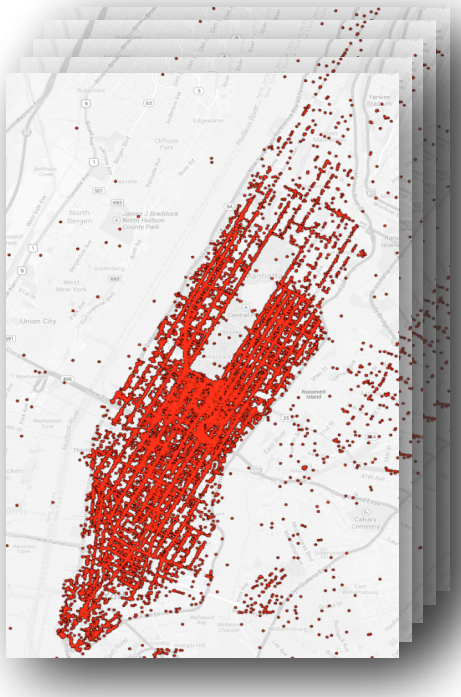


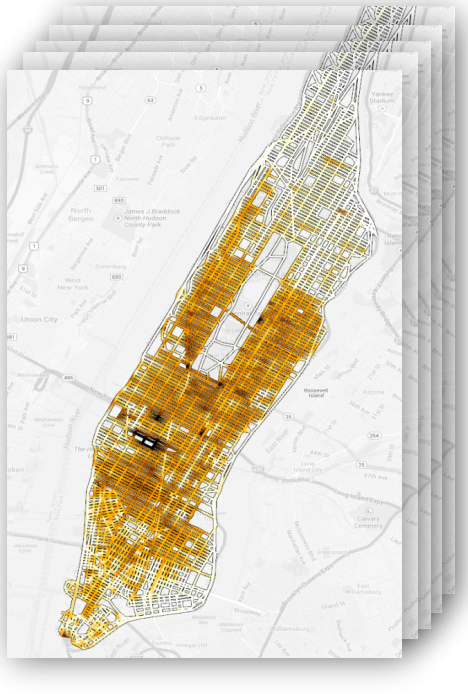
NYU TANDON SCHOOL OF ENGINEERING

[Doraiswamy et al., IEEE TVCG 2014]

Identifying Potential Events

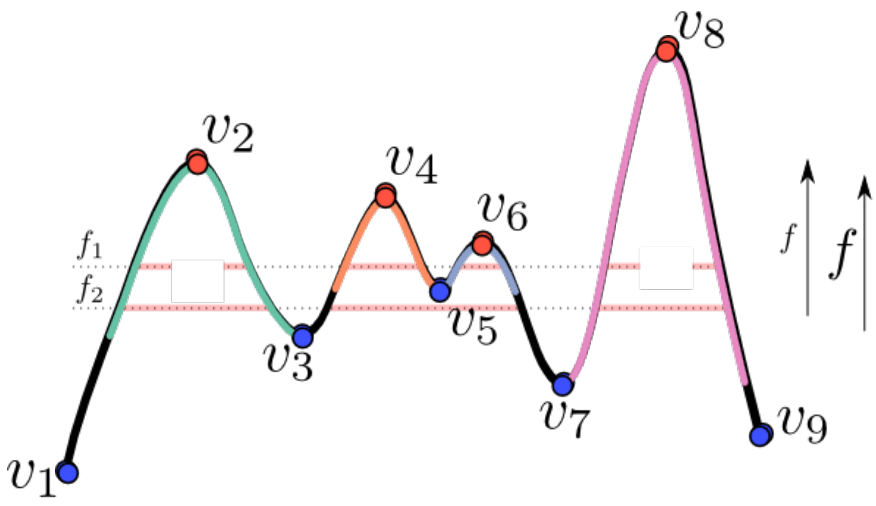
- Model data as a time-varying scalar function defined on a graph
  - Graph = road network
  - Function = density of taxis

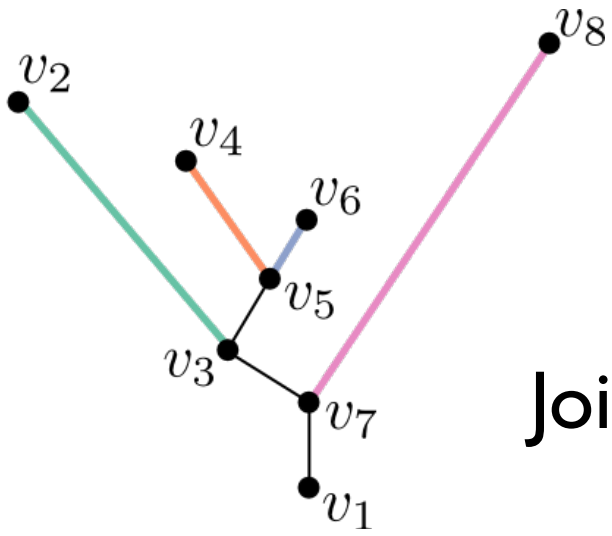




Identifying Potential Events

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

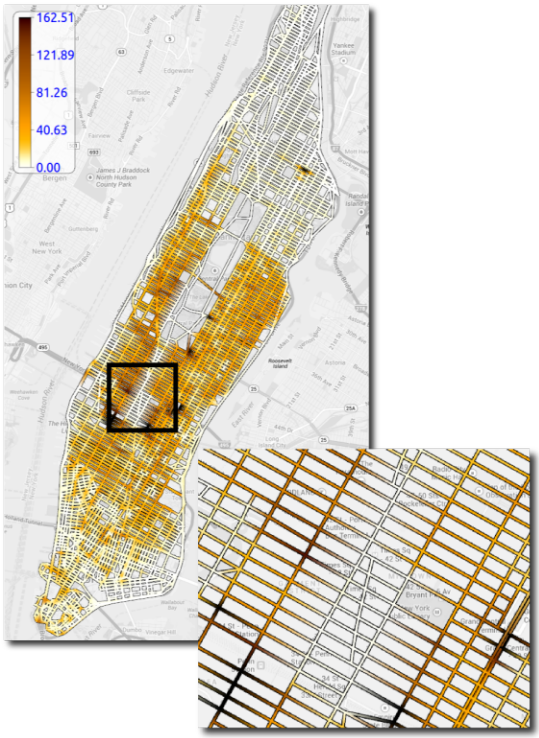




Join Tree

Taxi Data: Potential Events

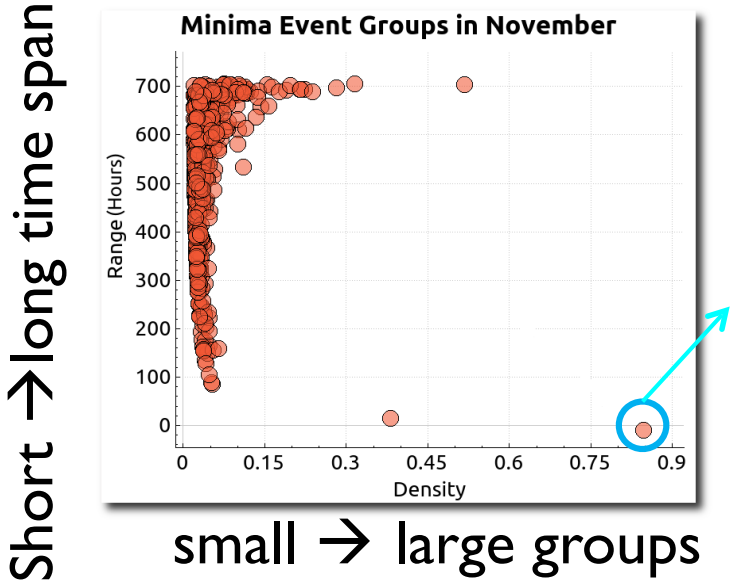
- Minima: lack of taxis
  - Region where density is lower than local neighborhood
  - Could denote road blocks, e.g., Macy's parade
- Maxima: popular taxi locations
  - Region where density is higher than local neighborhood
  - Could denote tourist locations, train stations
- Too many events: group similar events and create an index
  - Geometric and topological similarity



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

Visual Exploration Interface

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



Macy's parade

Visual Exploration Interface

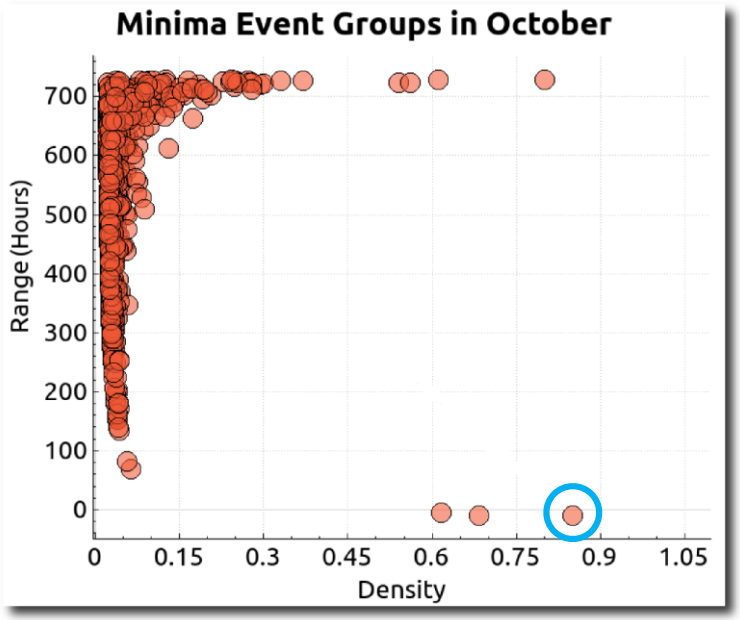
- Too many event groups
  - Many not interesting
- Visual interface to guide users
- Filter based on group size, event size,

event time, spatial region

- Hourly events: Halloween parades, Macy's parade, Gaza Solidarity, etc.
- Daily events: No. of days = 2
  - Hispanic Day Parade – Oct 9th
  - Columbus Day Parade – Oct 10th
- Weekly events: No. of weeks = 3
  - NYC Summer Streets: 3 consecutive Saturdays

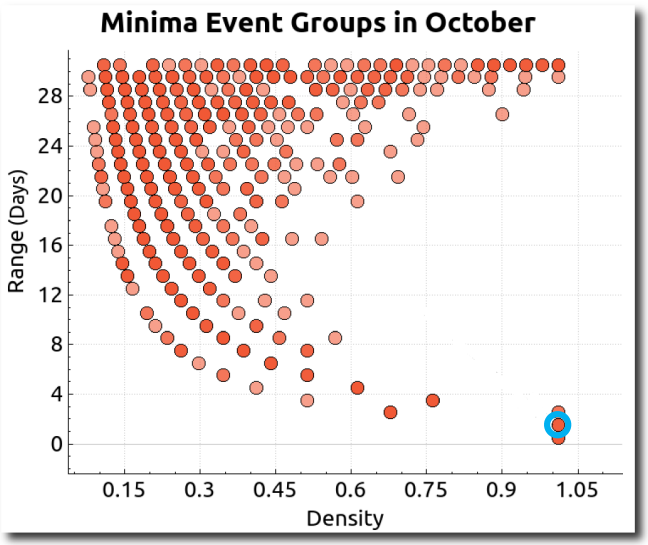
Minima Events - Hourly

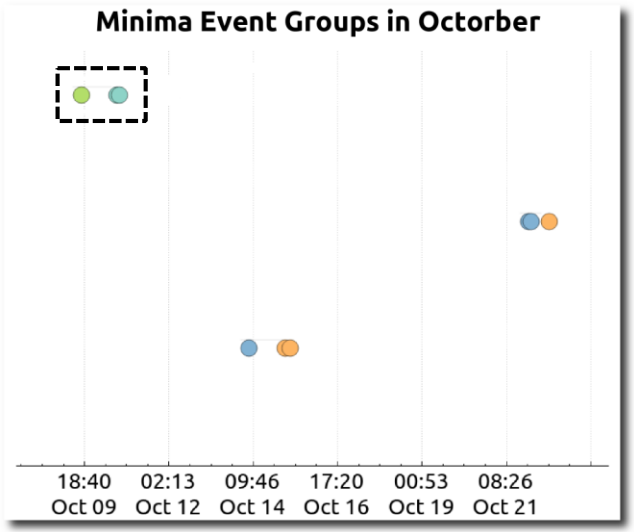
- October
  - Halloween Parade



Minima Events - Daily

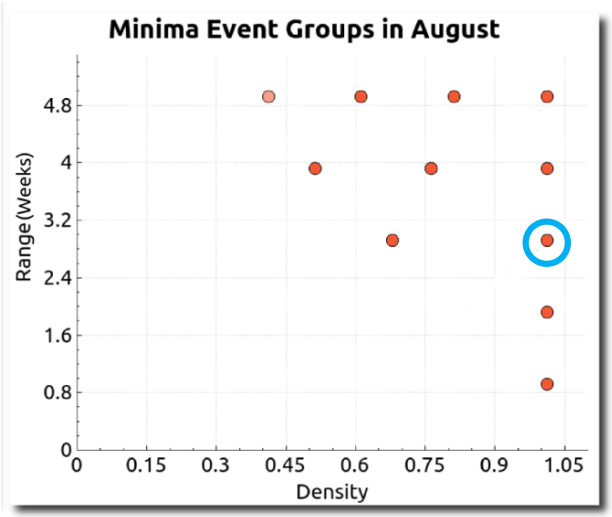
- October
  - No. of Days = 2
  - Hispanic Day Parade – Oct 9th
  - Columbus Day Parade – Oct 10th

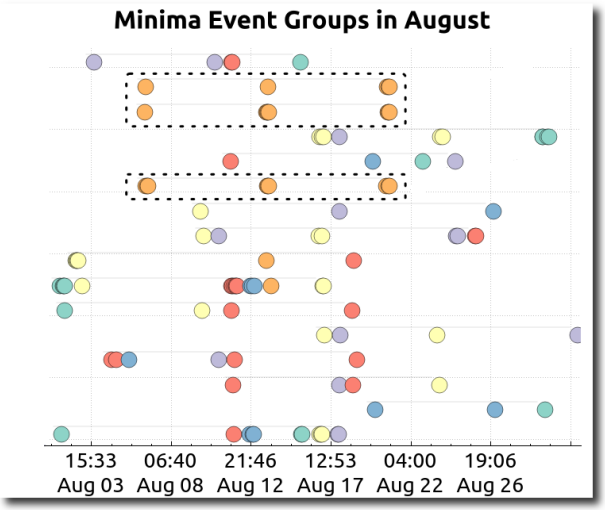




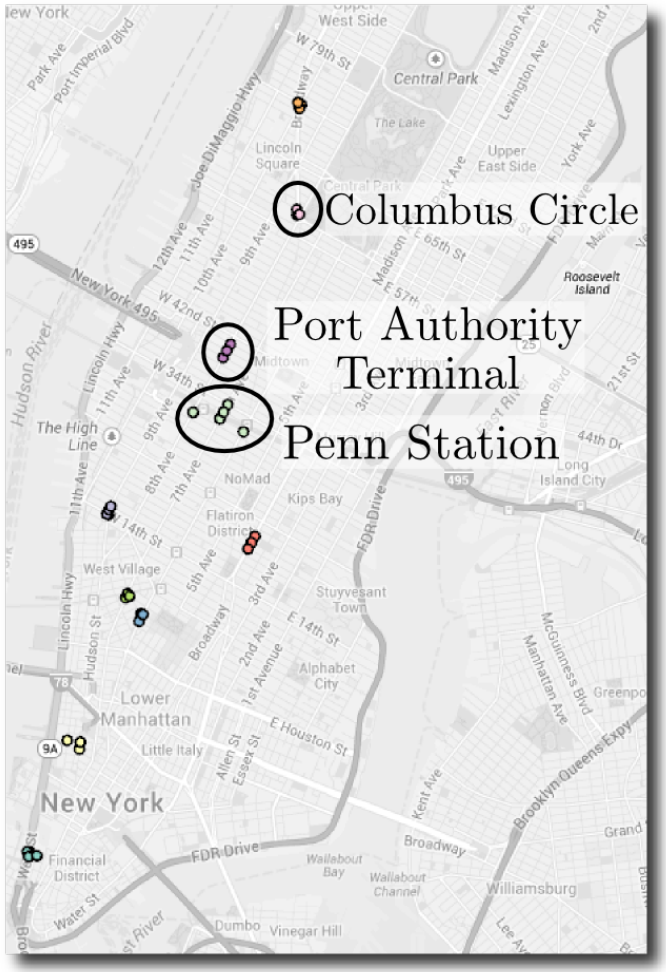
Minima Events - Weekly

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





Maxima Events



General trends

Night time trends

Event Guided Exploration

5 Borough Bike Tour 2011 (1 May 2011)

Go to Time slice

Querying Events

5 Borough Bike Tour
2011
(1 May 2011)

Query

Dominican Day Parade 2011
(14 August 2011)

5 Borough Bike Tour 2012
(6 May 2012)

Dominican Day Parade 2012
(12 August 2012)

Gaza Solidarity Protest NYC
(18 November 2012)

NYU TANDON SCHOOL OF ENGINEERING

Integrating Urban Data

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

Big Data 2014.2:144-154.

STRUCTURED OPEN URBAN DATA:

Understanding the Landscape

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

Abstract

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

Introduction

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

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

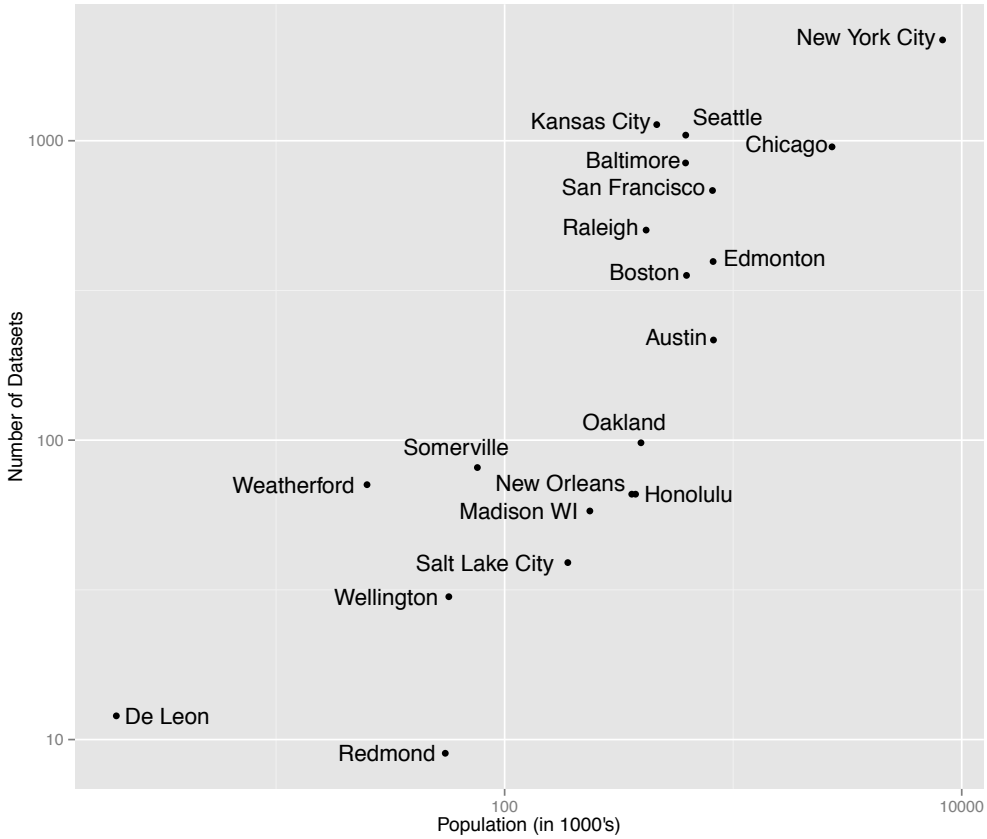
Having these data available creates many new opportunities. In particular, while individual data sets are valuable, by integrating data from multiple sources, the integrated data are often more valuable than the sum of their parts. The benefits of integrating city data have already led to many success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

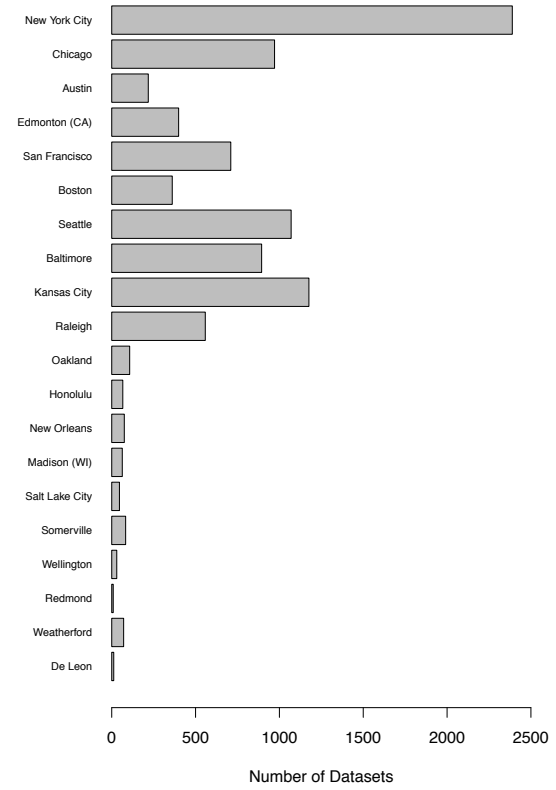
[1]IBM Research, Rio de Janiero, Brazil.
[2]Department of Computer Science and Engineering, NYU School of Engineering, Brooklyn, New York.
[3]NYU Center for Urban Science and Progress, Brooklyn, New York.

[Barbosa et al., Big Data 2014]

Structured Open Urban Data: Understanding the Landscape

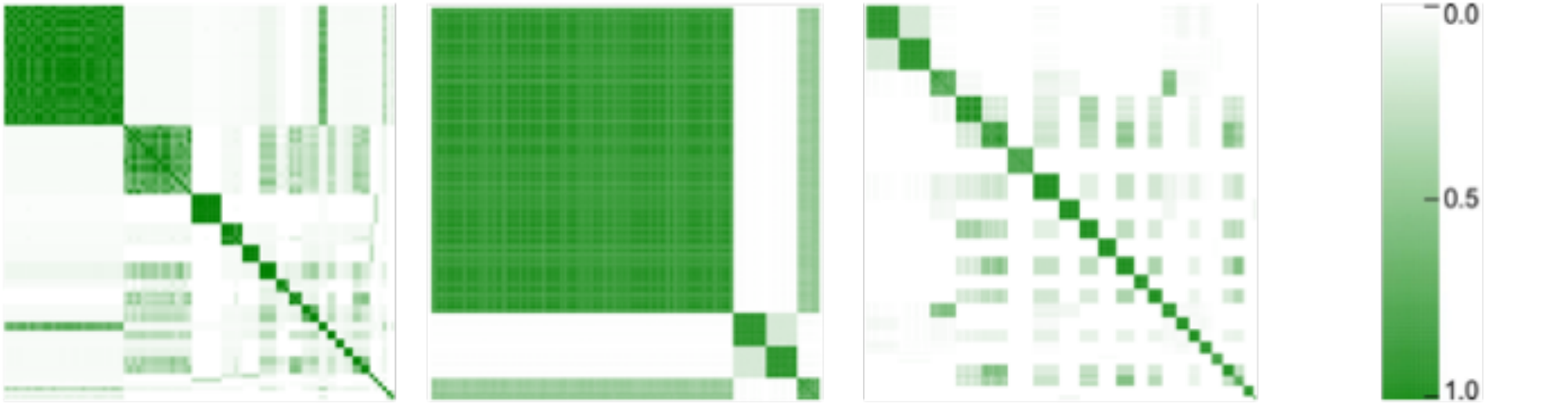
Luciano Barbosa[1] Kien Pham[2] Claudio Silva[2]
Marcos R. Vieira[1] Juliana Freire[2]
[1]IBM Research – Brazil [2]New York University





NYU TANDON SCHOOL OF ENGINEERING

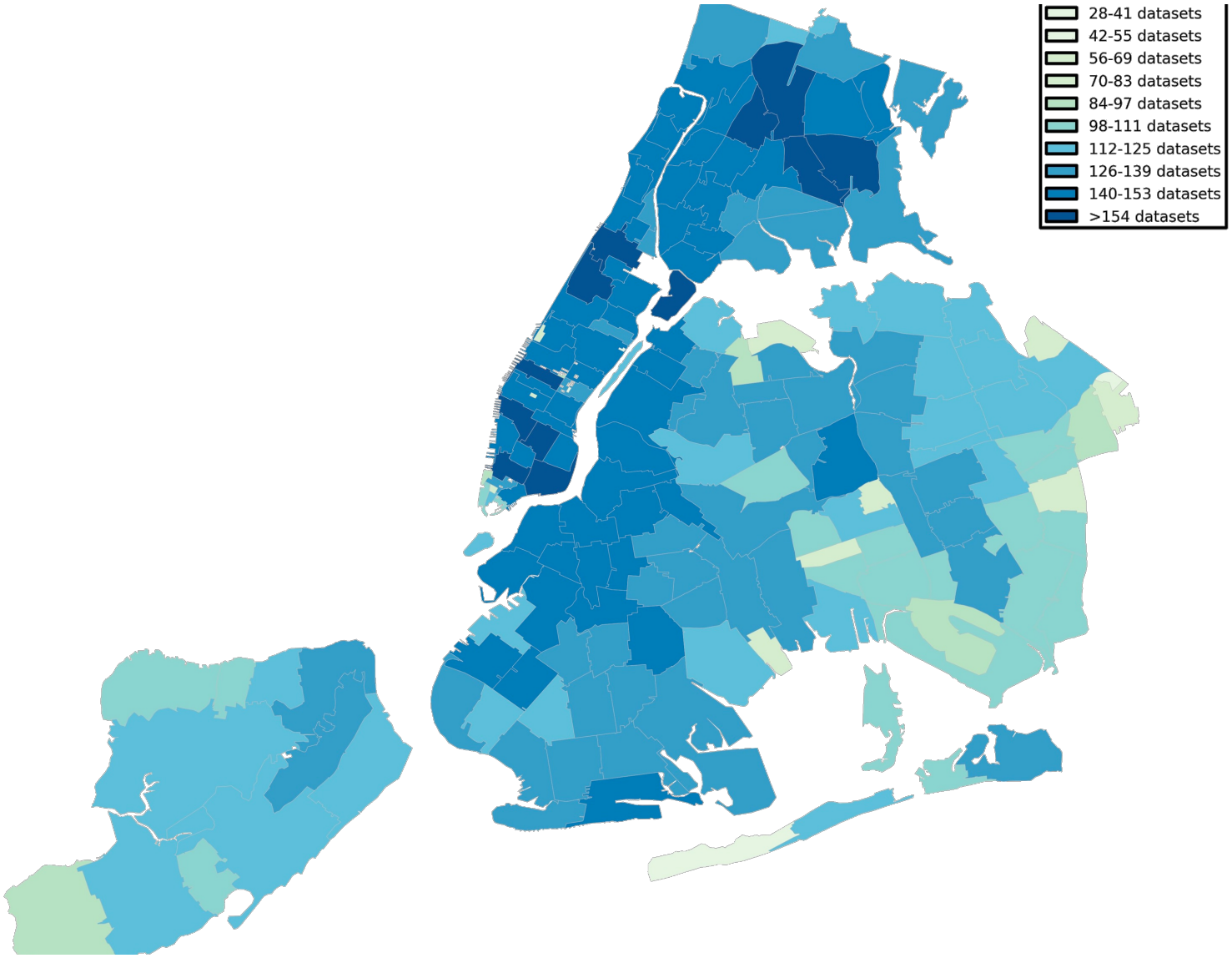
Integration Opportunities



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

Attribute overlap

Integration Opportunities



Geographical coverage and overlap

An Urban Data Profiler

Daniel Castellani Ribeiro
NYU Center for Urban Science+Progress
New York, USA
daniel.castellani@nyu.edu

Huy T. Vo
NYU Center for Urban Science+Progress
New York, USA
huy.vo@nyu.edu

Juliana Freire
NYU School of Engineering
NYU Center for Urban Science+Progress
New York, USA
juliana.freire@nyu.edu

Cláudio T. Silva
NYU School of Engineering
NYU Center for Urban Science+Progress
New York, USA
csilva@nyu.edu

ABSTRACT

Large volumes of urban data are being made available through a variety of open portals. Besides promoting transparency, these data can bring benefits to government, science, citizens and industry. It is no longer a fantasy to ask "if you could know anything about a city, what do you want to know" and to ponder what could be done with that information. However, the great number and variety of datasets creates a new challenge: how to find *relevant* datasets. While existing portals provide search interfaces, these are often limited to keyword searches over the limited metadata associated each dataset, for example, attribute names and textual description. In this paper, we present a new tool, UrbanProfiler, that automatically extracts detailed information from datasets. This information includes attribute types, value distributions, and geographical information, which can be used to support complex search queries as well as visualizations that help users explore and obtain insight into the contents of a data collection. Besides describing the tool and its implementation, we present case studies that illustrate how the tool was used to explore a large open urban data repository.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Data sharing, Web-based services*

Keywords

Metadata Extractionl; Automatic Type Detection; Dataset Analysis

1. INTRODUCTION

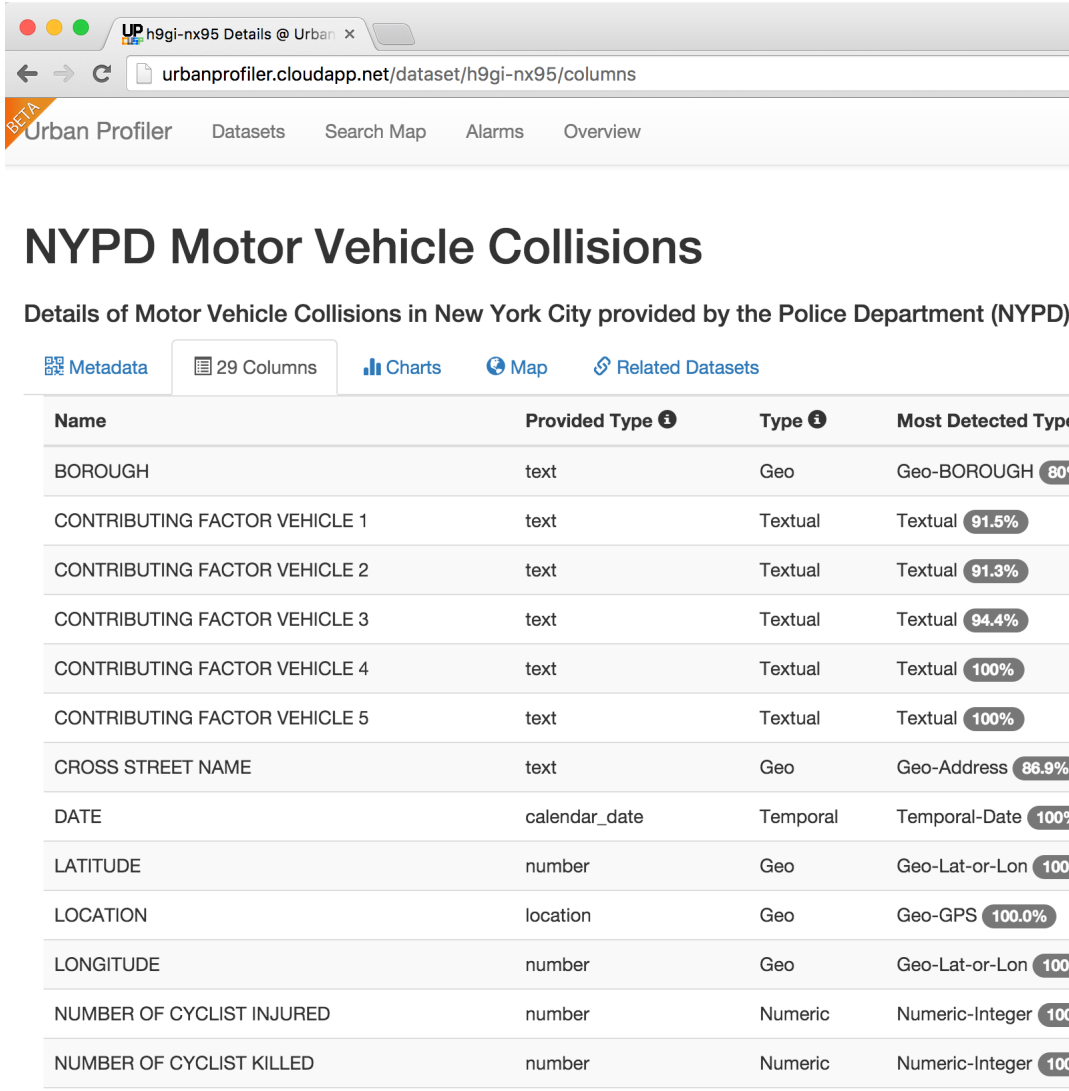
About half of humanity lives in urban environments today and that number will grow to 80% by the middle of this century; North America is already 80% in cities, and will rise to 90% by 2050. Cities are thus the loci of resource consumption, of economic activity, and of innovation; they are the cause of our looming sustainability problems but also where those problems must be solved. Our increasing ability to collect, transmit, and store data, coupled with the growing trend towards openness [1, 7, 9, 19, 6, 16, 14], creates a unique opportunity that can benefit government, science, citizens and industry. By integrating and analyzing multiple data sets, city governments can go beyond today's imperfect and often anecdotal understanding of cities to enable better operations and informed planning (see e.g., [5, 7]). Domain scientists can engage in data-driven science and explore longitudinal processes to understand people's behavior [8]; identify causal relationships across datasets, which can in turn, influence policy decisions [3, 18]; or create models and derive predictions that benefit citizens (see e.g., [4]). Putting urban data in the hands of citizens has the potential to improve governance and participation, and in the hands of entrepreneurs and corporations it will lead to new products and services. In short, it is no longer a fantasy to ask "if you could know anything about a city, what do you want to know" and to ponder what could be done with that information.

While in the past, government, policymakers and scientists faced significant constraints in obtaining the data needed for planning and evaluating their policies and practices, currently they are faced with an information overload. The number of open data portals and the volume of data they hold are growing at a fast pace around the world [14, 15, 16, 17] A big challenge, now, is how to discover datasets that are relevant for a given task or information need.

Publishing platforms such as CKAN [2] and Socrata [20], which are widely used for open urban data, provide a simple search interface over the metadata, thus, users are not able to identify datasets based on their content. Besides, there are no standards for attribute names and, often, attributes lack even basic type information [1]. This makes it hard for users to formulate discovery queries.

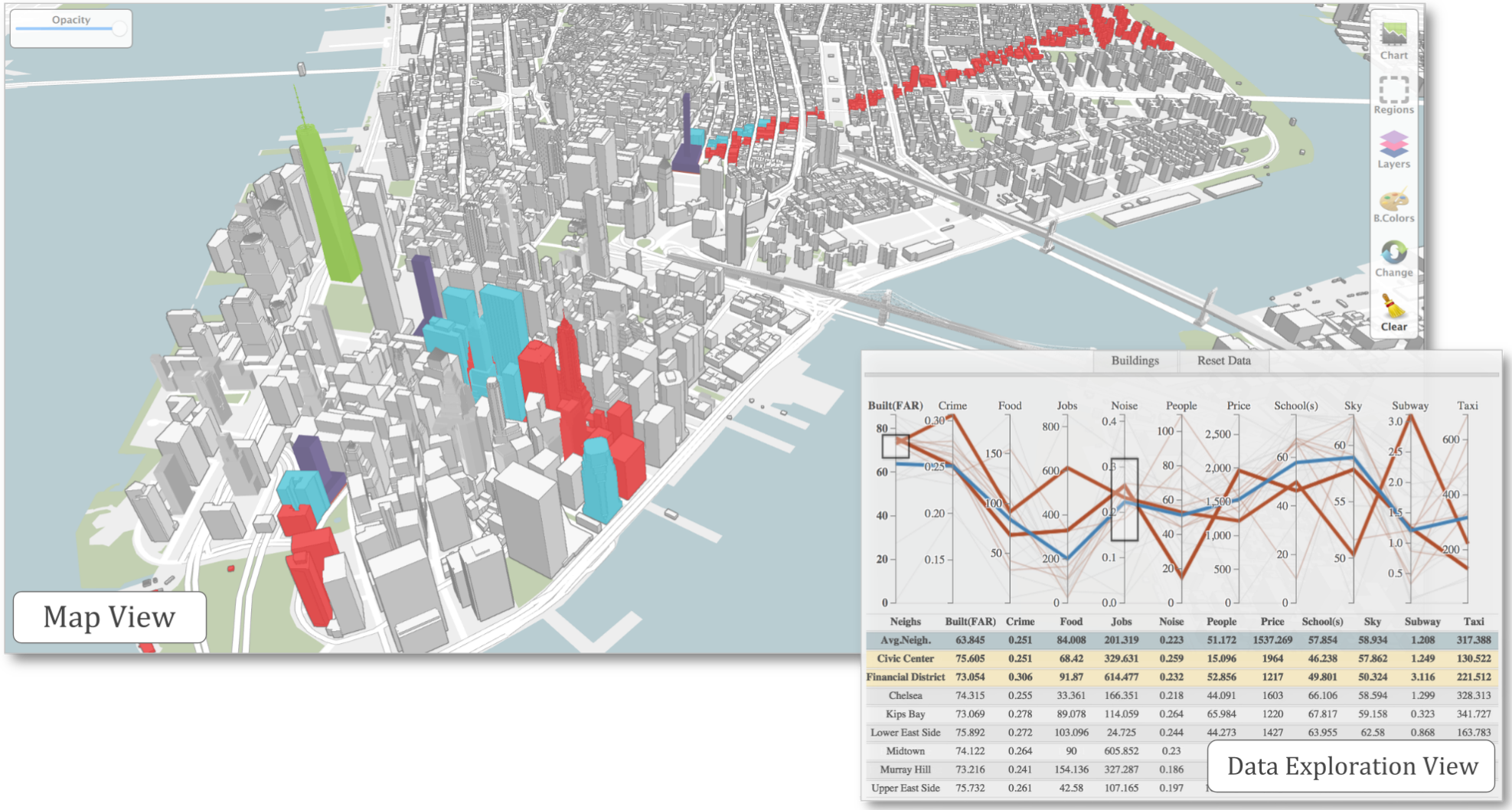
As a step towards enabling richer queries and helping users identify the datasets they need, we propose a new tool, UrbanProfiler, which automatically extracts detailed information about the contents of the datasets. The goal is to use this information to enable users explore urban data by asking queries over attributes, content, and to filter datasets based on a given time period or a region. The latter is crucial given that a large percentage of urban data contains spatial and temporal information [1]. Furthermore, longitudinal analyses often require multiple datasets that overlap in space and time. Consider, for example, a social scientist, who tries to understand the effects of adding a bike lane to a city neighborhood,

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742135.



NYU TANDON SCHOOL OF ENGINEERING

Urbane



Neighs	Built(FAR)	Crime	Food	Jobs	Noise	People	Price	School(s)	Sky	Subway	Taxi
Avg.Neigh.	63.845	0.251	84.008	201.319	0.223	51.172	1537.269	57.854	58.934	1.208	317.388
Civic Center	75.605	0.251	68.42	329.631	0.259	15.096	1964	46.238	57.862	1.249	130.522
Financial District	73.054	0.306	91.87	614.477	0.232	52.856	1217	49.801	50.324	3.116	221.512
Chelsea	74.315	0.255	33.361	166.351	0.218	44.091	1603	66.106	58.594	1.299	328.313
Kips Bay	73.069	0.278	89.078	114.059	0.264	65.984	1220	67.817	59.158	0.323	341.727
Lower East Side	75.892	0.272	103.096	24.725	0.244	44.273	1427	63.955	62.58	0.868	163.783
Midtown	74.122	0.264	90	605.852	0.23						
Murray Hill	73.216	0.241	154.136	327.287	0.186						
Upper East Side	75.732	0.261	42.58	107.165	0.197						

URBANE:

A 3D Framework to Support Data Driven Decision Making in Urban Developments

IEEE VAST 2015
Submission ID: 268

Harish Doraiswamy[1], Nivan Ferreira[1], Huy Vo[1], Claudio Silva[1], Marcos Lange[2], Muchan Park[3], Heidi Werner[3], Luc Wilson[3]

New York University[1], Universidade Federal Fluminense[2], Kohn Pederson Fox Associates PC[3]

Video: goo.gl/aZnhr7

Thank you!

csilva@nyu.edu

http://engineering.nyu.edu/