Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective

Webinar Series
Fall 2017

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective

9/12/17 – Building Data Acumen *(recording posted)*

9/19/17 – Incorporating Real-World Applications

9/26/17 – Faculty Training and Curriculum Development

10/3/17 – Communication Skills and Teamwork

10/10/17 – Inter-Departmental Collaboration and Institutional Organization

10/17/17 – Ethics

10/24/17 – Assessment and Evaluation for Data Science Programs

11/7/17 – Diversity, Inclusion, and Increasing Participation

11/14/17 – Two-Year Colleges and Institutional Partnerships

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective
Incorporating Real-World Applications

Cláudio T. Silva, New York University
Professor of computer science
and engineering and data science

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of
Data Science & Advanced Analytics at
MassMutual Financial Group

Provide input and learn more about the study at *www.nas.edu/EnvisioningDS*

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Using Urban and Sports Data in Student Projects

Cláudio T. Silva, New York University
Professor of computer science and engineering and data science

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Using Urban and Sports Data in Student Projects

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, Harish Doraiswamy, Carlos Dietrich, 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

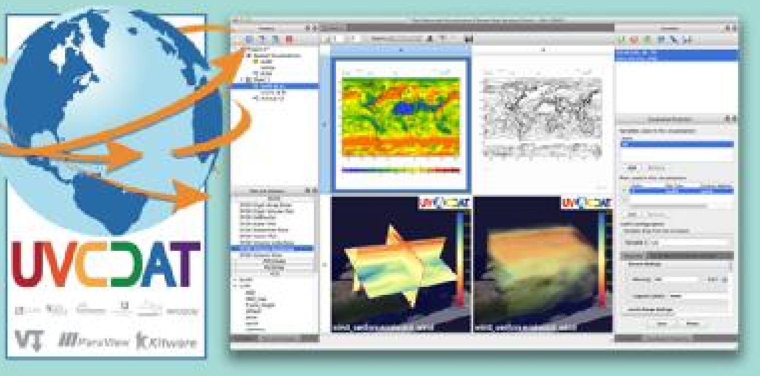
Funded by grants/gifts from NSF, NASA, DOE, Moore and Sloan Foundations, MLB.com, DARPA, AT&T, NVIDIA, and IBM

NYU TANDON SCHOOL OF ENGINEERING

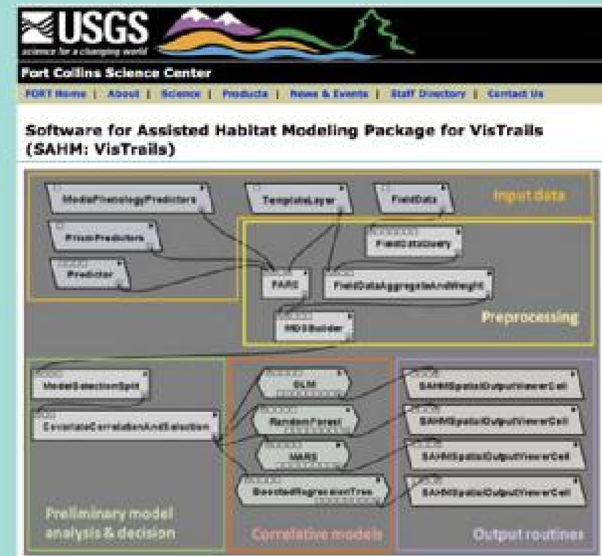
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

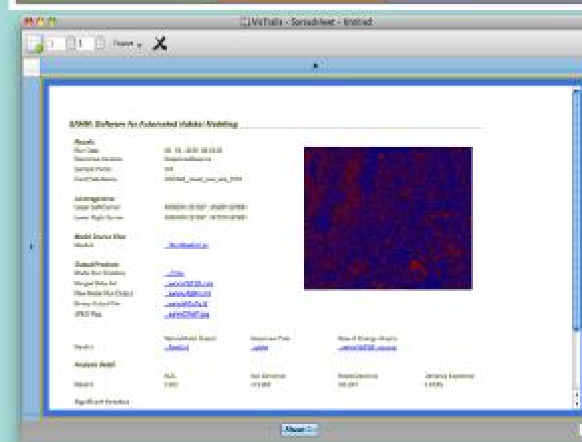
Data Science Applications - I

Climate Data Analysis

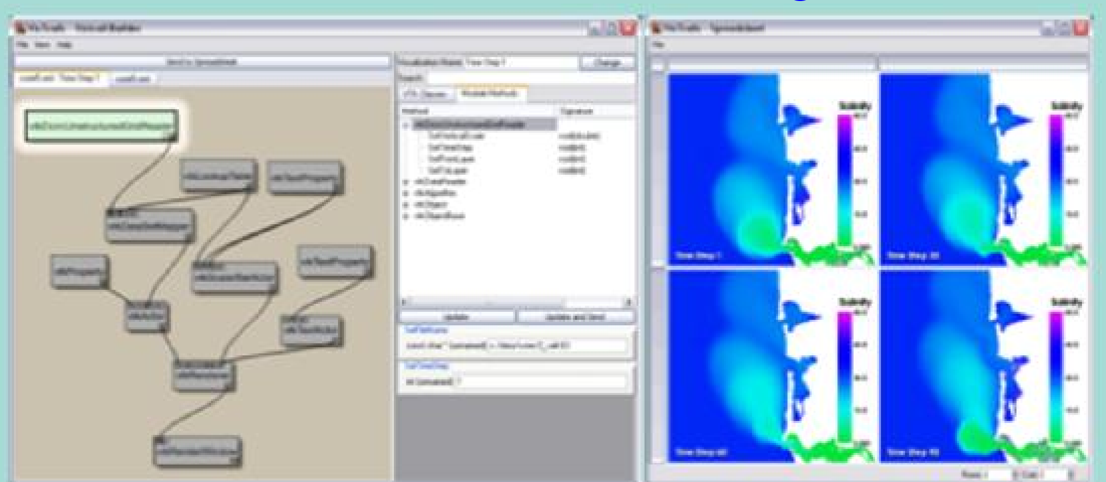


Modeling the Spread of Invasive Species





VisTrails — www.vistrails.org

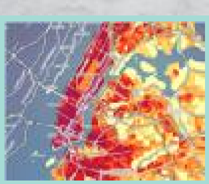


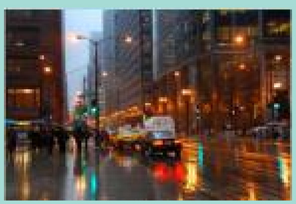
Data Science Applications - II

Urban Applications

Infrastructure | *Environment* | *People*





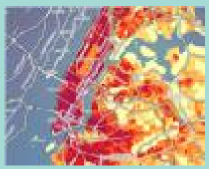

Sports Data Analytics

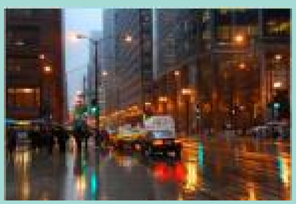
Data Science Applications - II

Urban Applications

Infrastructure *Environment* *People*



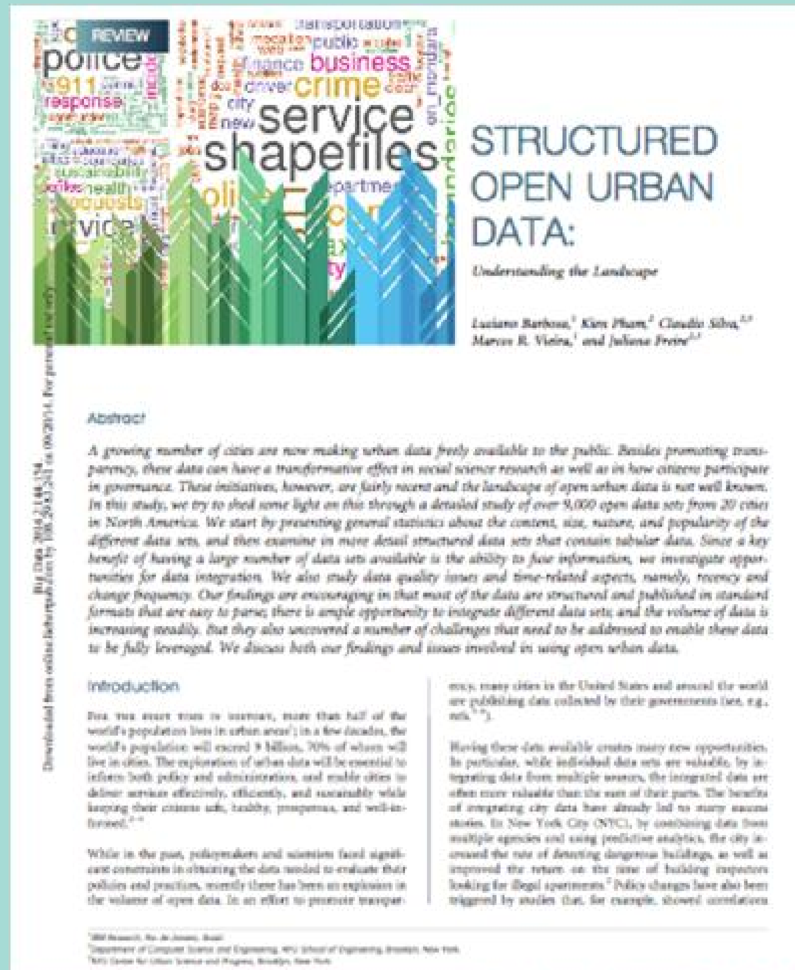



Applications in these areas are attractive to students since data is closer to their interests and they can tap into their personal experiences

Sports Data Analytics

Urban Data

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



REVIEW

STRUCTURED OPEN URBAN DATA:

Understanding the Landscape

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

Abstract

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

Introduction

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

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.).

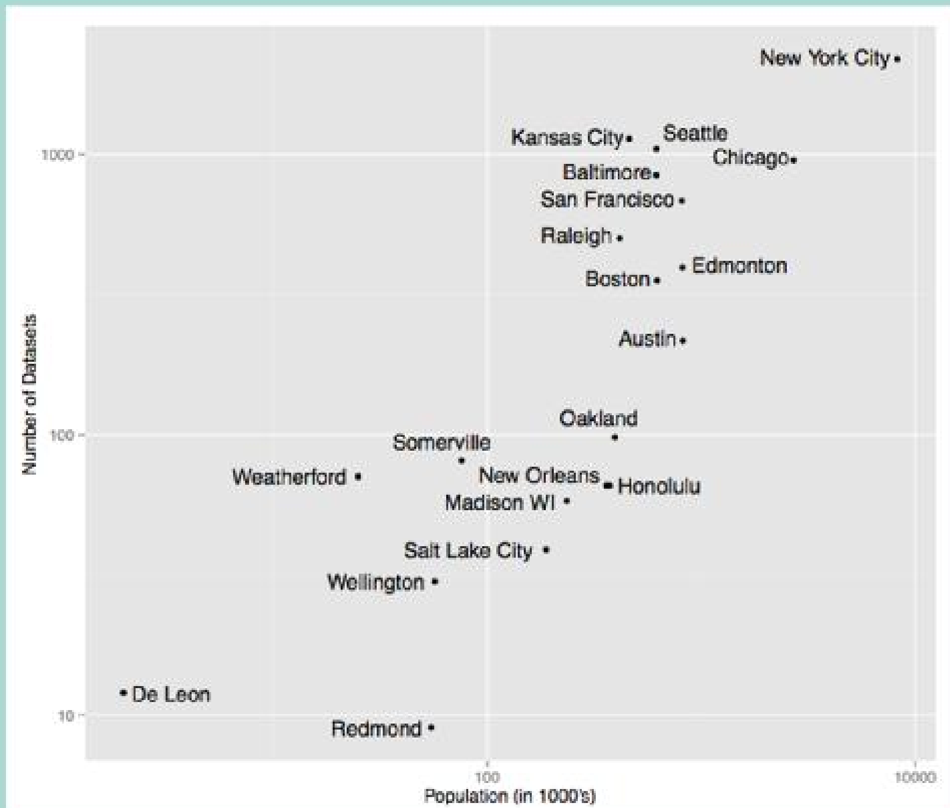
Having these data available creates many new opportunities. In particular, while individual data sets are valuable, by integrating data from multiple sources, the integrated data are often more valuable than the sum of their parts. The benefits of integrating city data have already led to many success 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. Policy changes have also been triggered by studies that, for example, showed correlations

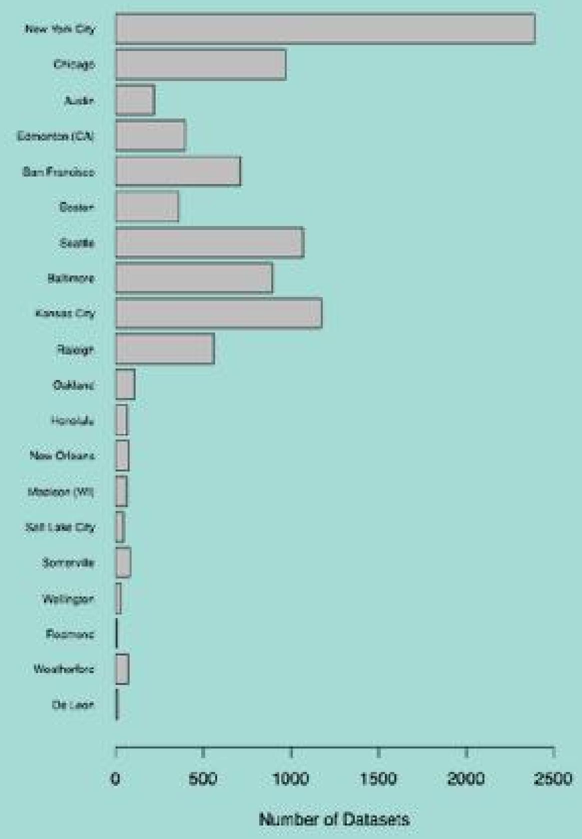
[Barbosa et al., Big Data 2014]

https://github.com/ViDA-NYU/urban-data-study

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





https://github.com/ViDA-NYU/urban-data-study

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.

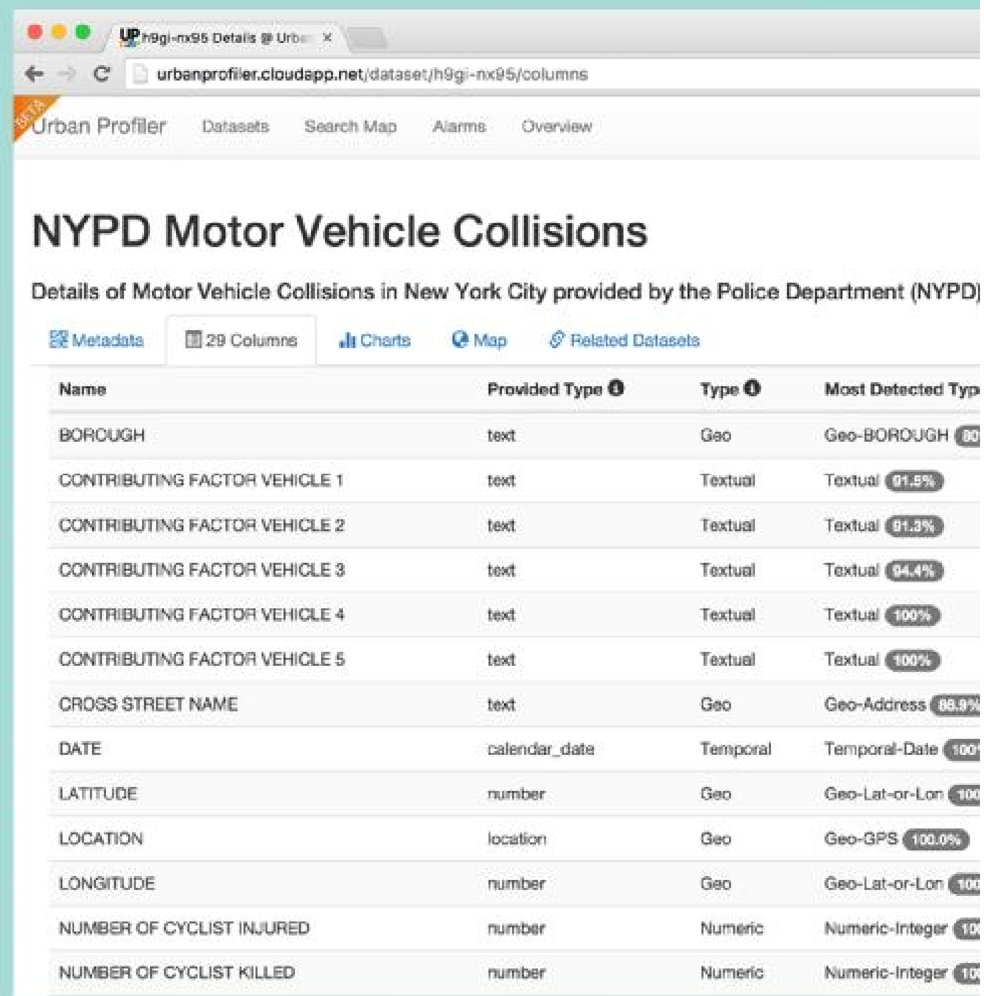
WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742135.

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, 18, 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,



https://datahub.cusp.nyu.edu/

Taxi drivers petition NYC for fare hike over soaring gas prices

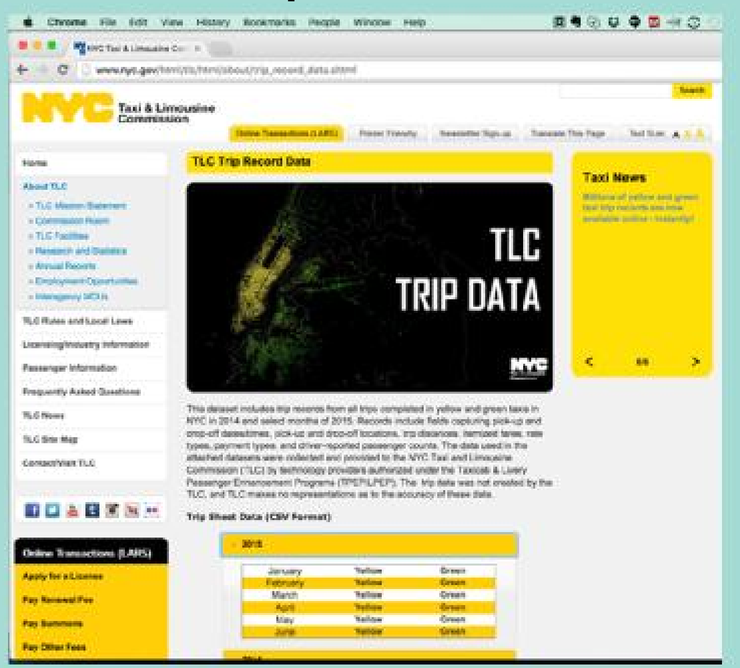
BY PETE DONOHUE / DAILY NEWS STAFF WRITER

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

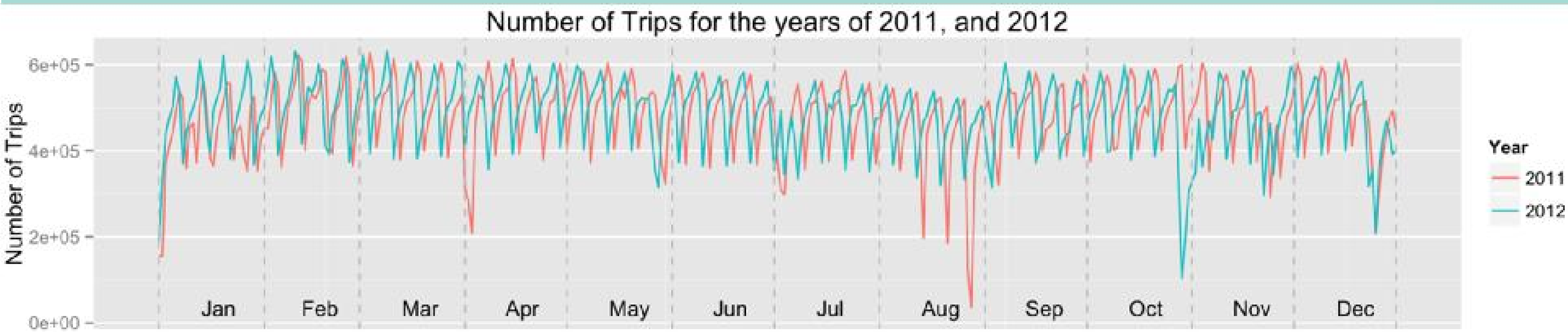
ailynews.com/new-york

NYC Taxi Data

- It is relatively *big*: ~500k trips/day - several hundred million trips in 5 years
- ... and relatively *complex*:
  - *spatio-temporal:* pick up + drop off
  - *trip attributes*: e.g., distance traveled, cost, tip
- Many data slices to examine



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?"

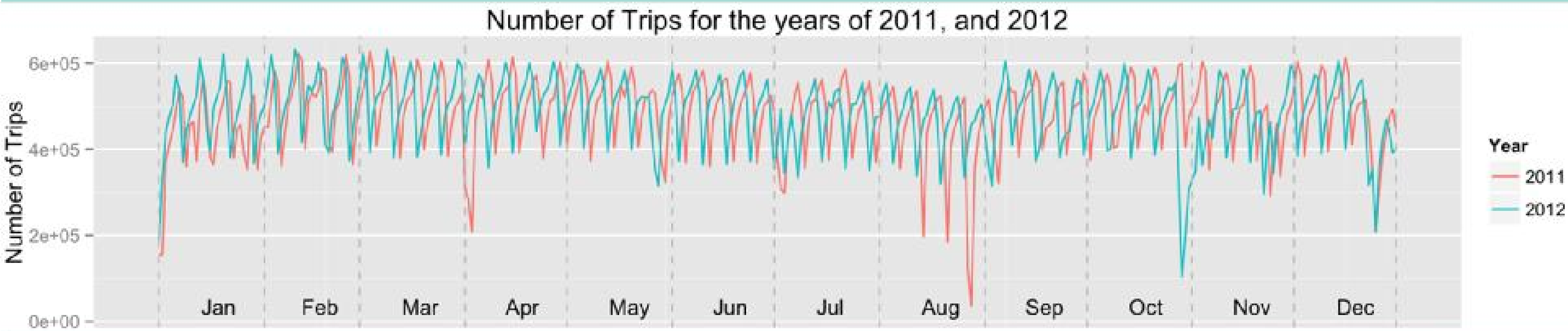
  "What is the average trip time from Midtown to the airports during weekdays?"

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

  "How did the movement patterns change during Sandy?"

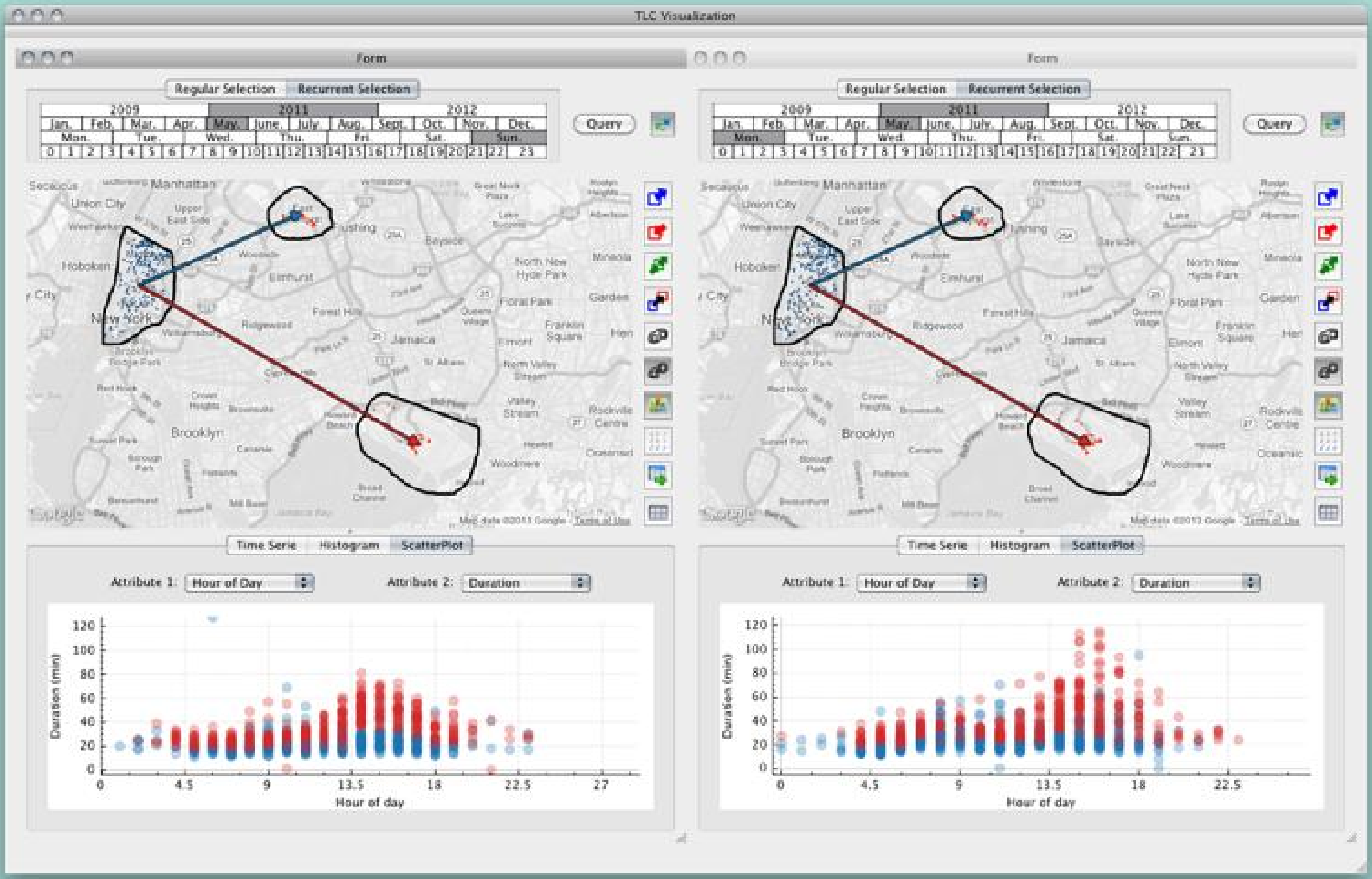
  "Where are the popular night spots?"

Exploring Urban Data: NYC Taxis



7-8am 8-9am 9-10am 10-11am

Looking at NYC Taxi Records



https://github.com/ViDA-NYU/TaxiVis

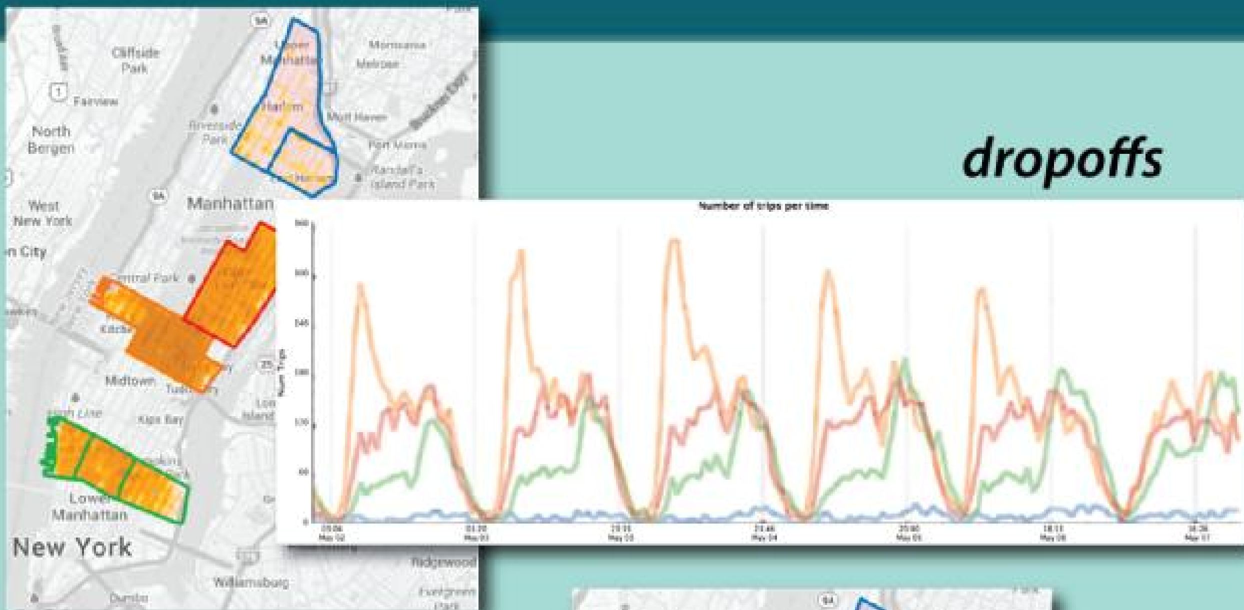
[Ferreira et al., IEEE TVCG 2013]

NYU TANDON SCHOOL OF ENGINEERING

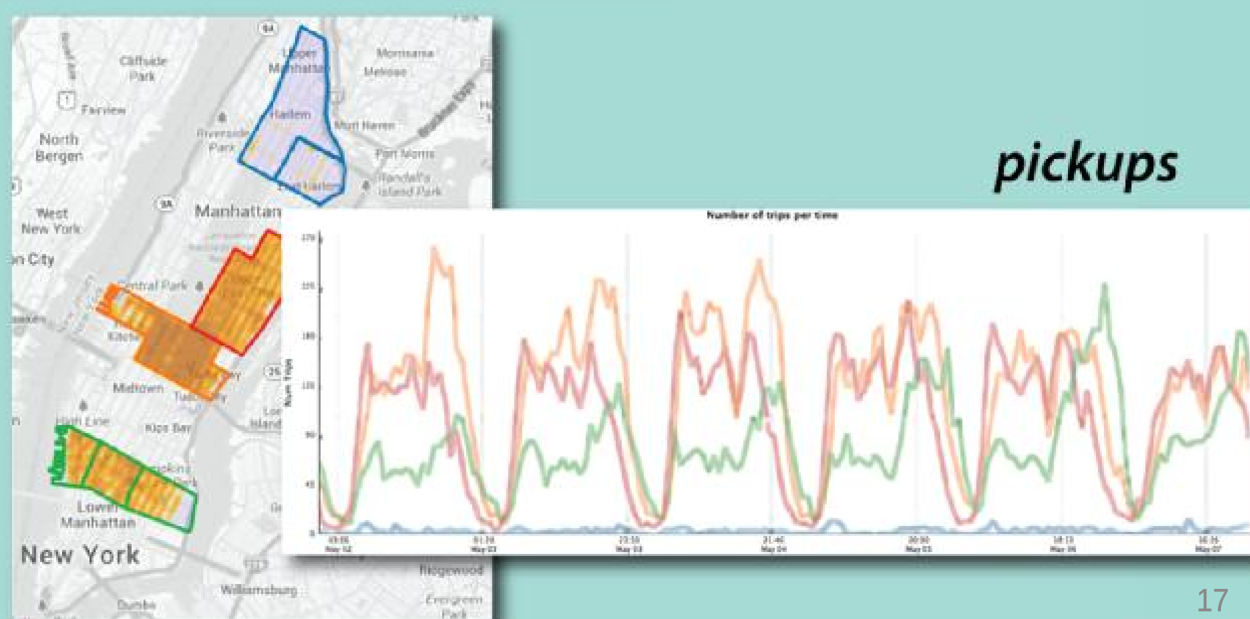
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

TaxiVis: Comparing Neighborhoods

dropoffs



pickups

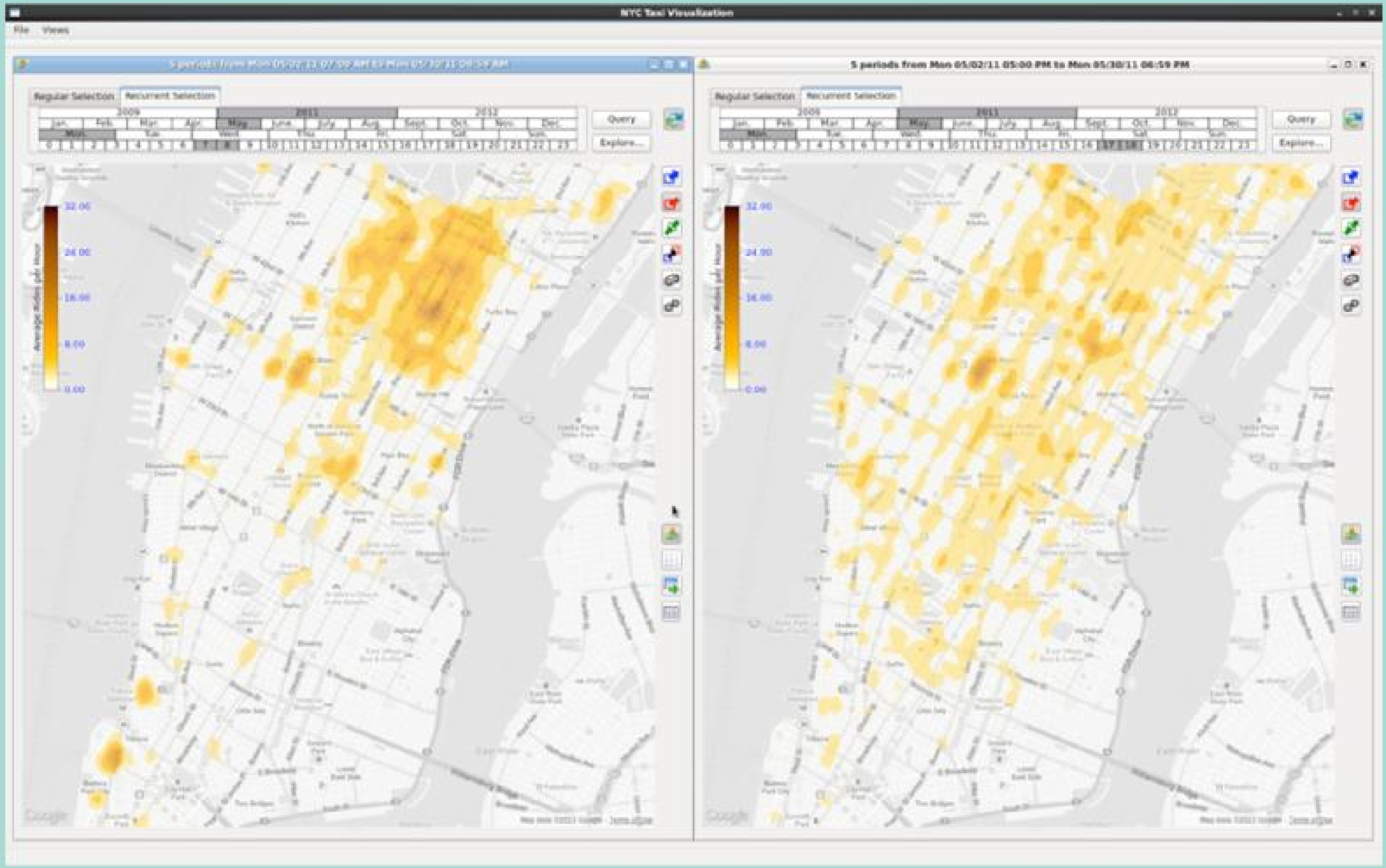


Time Exploration

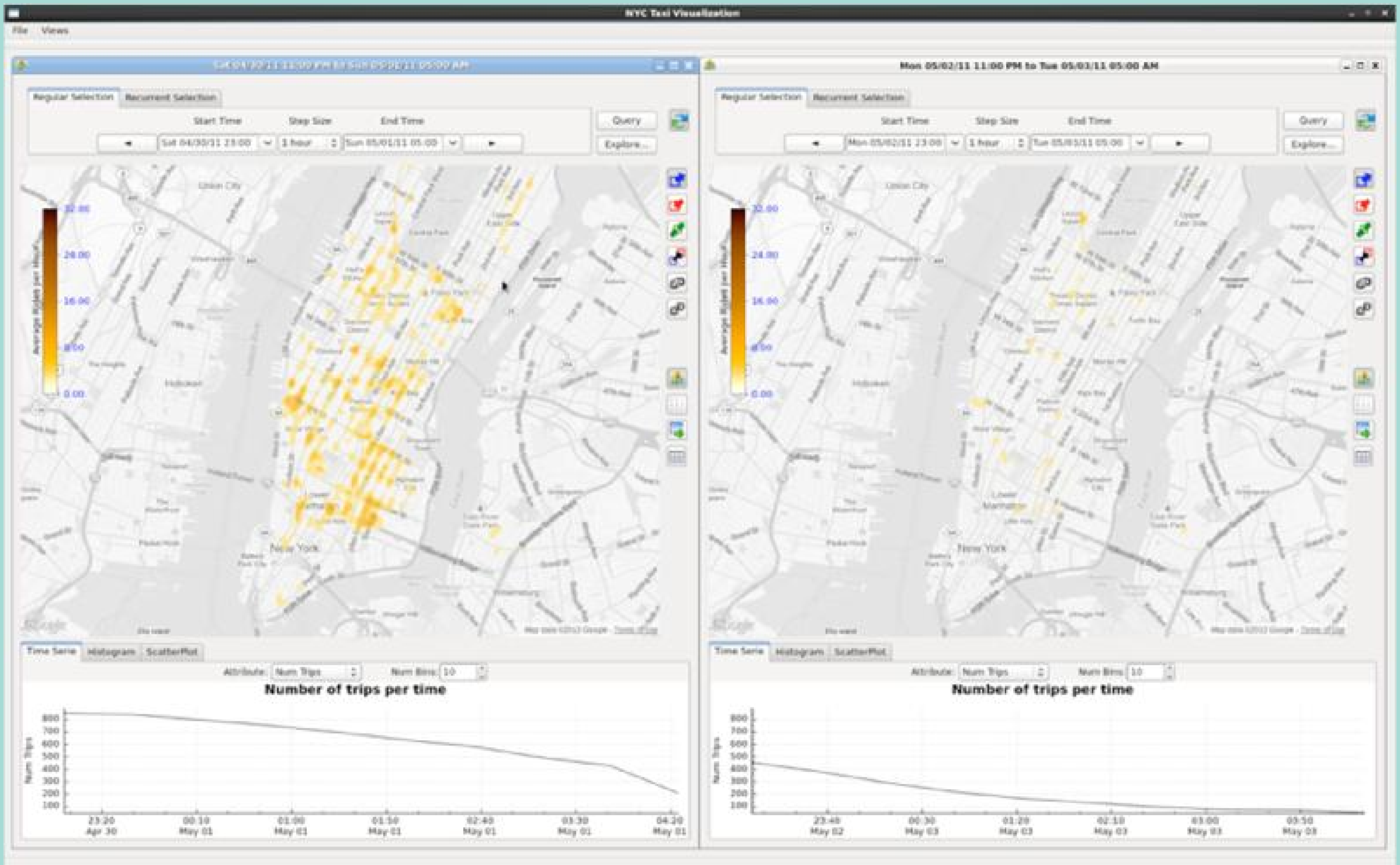
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

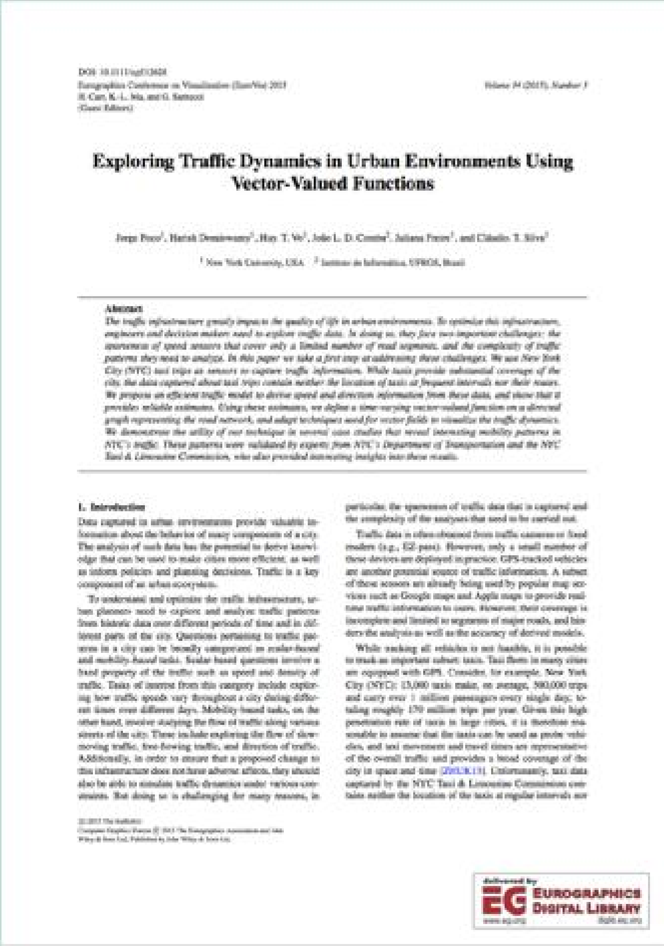
Dropoffs Before vs. After Work



Night Life Saturday vs. Monday



A Taxi over 24 hours



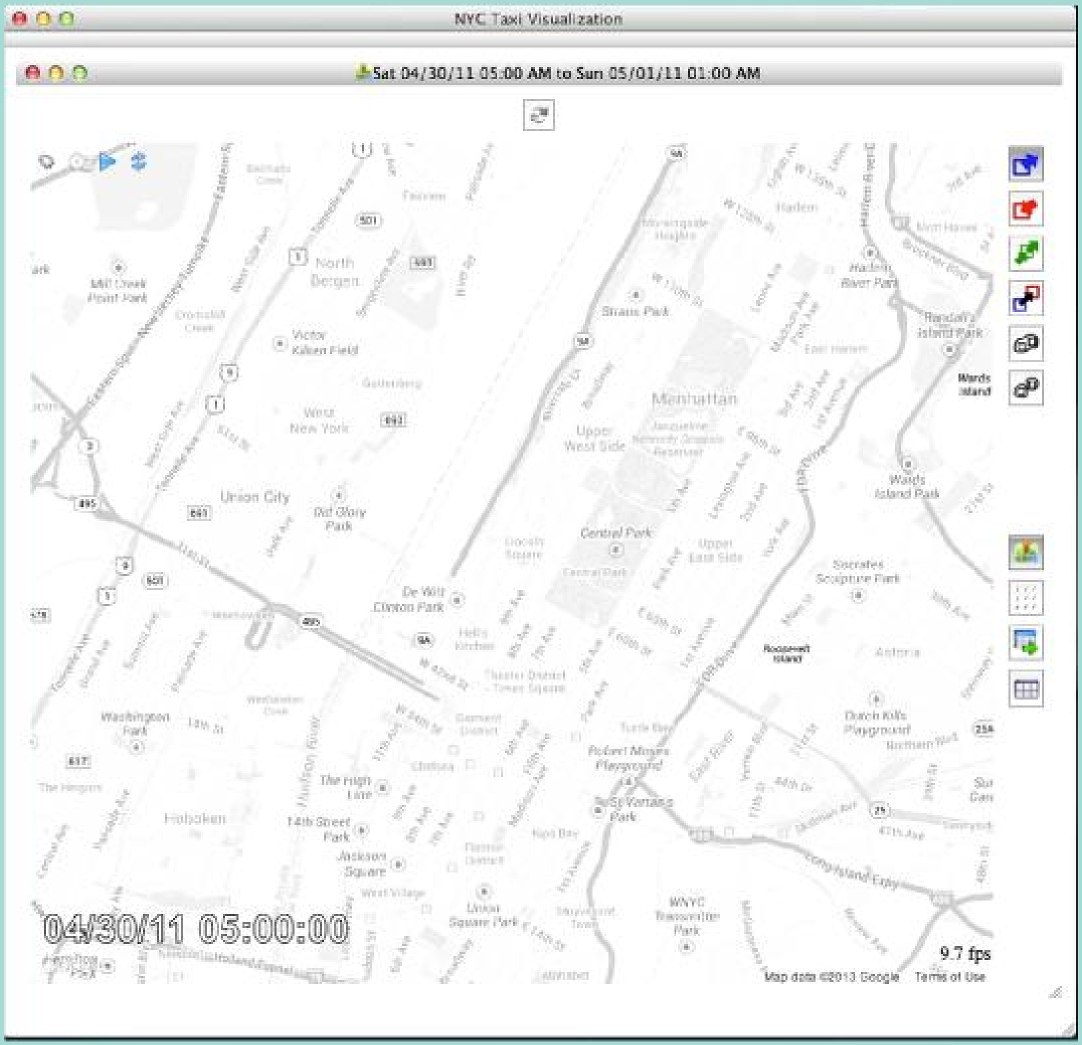
Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions

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

1. Introduction



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

Student Course Projects

The Daily Commute: An In-depth Analysis of Manhattan Traffic Patterns Between Yellow Cab, Uber, and CitiBike

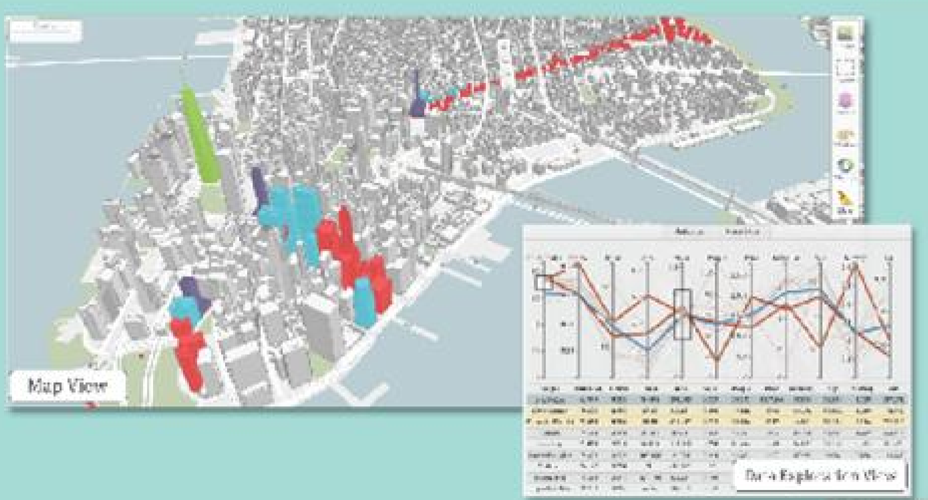
Crime Analysis in New York City 2006-2015

Detecting Gentrification with Taxi Patterns in NYC

Optimizing Walking Paths Based on Interestingness

NYC Real Estate Price Prediction

Projects with 3D data



[Ferreira et al., IEEE VAST 2015]

https://www.nytimes.com/interactive/2016/12/21/upshot/Mapping-the-Shadows-of-New-York-City.html?mcubz=0

Projects with other data modalities, e.g., sound

https://wp.nyu.edu/sonyc/

Thank you!

csilva@nyu.edu

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Cláudio T. Silva, New York University
Professor of computer science and engineering and data science

Using Urban and Sports Data in Student Projects

Q&A

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Building a talent pipeline through a strategic career development program & academic-industrial partnerships

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of
Data Science & Advanced Analytics at
MassMutual Financial Group

Building a talent pipeline

Academic Partnerships

Development Program
- Support creation of undergraduate programs
- Blend academic rigor with industry application
- Integrate academic approach to development

Research and Senior Talent
- Senior data science talent generation
- Collaborative research opportunities

Junior Data Scientist → Data Scientist → Senior Data Scientist → Lead Data Scientist

Year 1: Portfolio Assembly
- Enroll in MS program
- Participate in 2 projects

Year 2: Communicating Results
- Participate in 2 projects
- Continue course work
- Defend results

Year 3: Leading a Project
- Complete coursework
- Identify problem in business
- Scope and assess value
- Executed with junior member
- Defend project

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications – Q&A

Cláudio T. Silva, New York University
Professor of computer science and engineering and data science

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of Data Science & Advanced Analytics at *MassMutual Financial Group*

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

9/12/17 – Building Data Acumen
(recording posted)

9/19/17 – Incorporating Real-World Applications

9/26/17 – Faculty Training and Curriculum Development

10/3/17 – Communication Skills and Teamwork

10/10/17 – Inter-Departmental Collaboration and Institutional Organization

10/17/17 – Ethics

10/24/17 – Assessment and Evaluation for Data Science Programs

11/7/17 – Diversity, Inclusion, and Increasing Participation

11/14/17 – Two-Year Colleges and Institutional Partnerships

Provide input and learn more about the study at
www.nas.edu/EnvisioningDS

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Webinar Series
Fall 2017

The National Academies of SCIENCES ENGINEERING MEDICINE

nas.edu/EnvisioningDS

Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective

9/12/17 – Building Data Acumen *(recording posted)*

9/19/17 – Incorporating Real-World Applications

9/26/17 – Faculty Training and Curriculum Development

10/3/17 – Communication Skills and Teamwork

10/10/17 – Inter-Departmental Collaboration and Institutional Organization

10/17/17 – Ethics

10/24/17 – Assessment and Evaluation for Data Science Programs

11/7/17 – Diversity, Inclusion, and Increasing Participation

11/14/17 – Two-Year Colleges and Institutional Partnerships

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Cláudio T. Silva, New York University
Professor of computer science
and engineering and data science

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of
Data Science & Advanced Analytics at
MassMutual Financial Group

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Cláudio T. Silva, New York University

Professor of computer science and engineering and data science

Using Urban and Sports Data in Student Projects

Using Urban and Sports Data in Student Projects

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, Harish Doraiswamy, Carlos Dietrich, 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

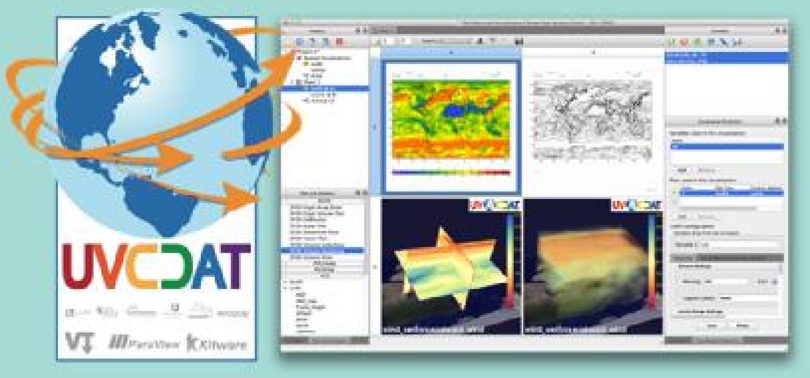
Funded by grants/gifts from NSF, NASA, DOE, Moore and Sloan Foundations, MLB.com, DARPA, AT&T, NVIDIA, and IBM

NYU TANDON SCHOOL OF ENGINEERING

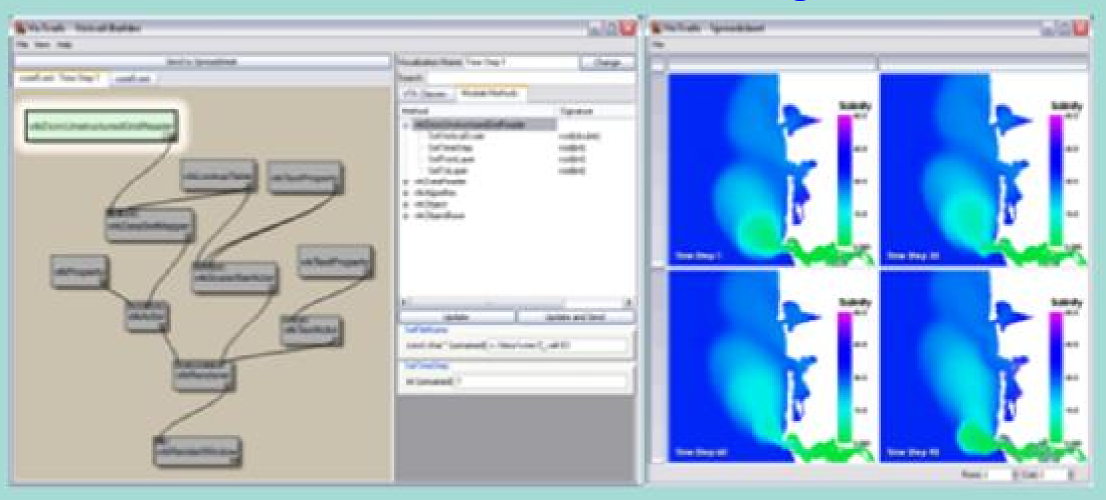
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Data Science Applications - I

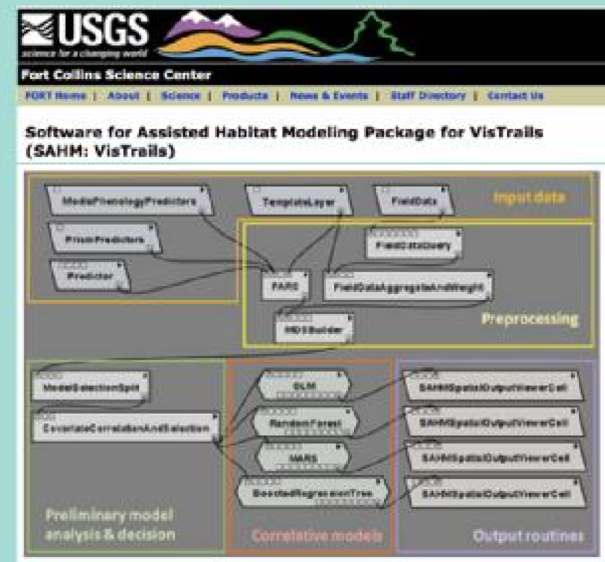
Climate Data Analysis

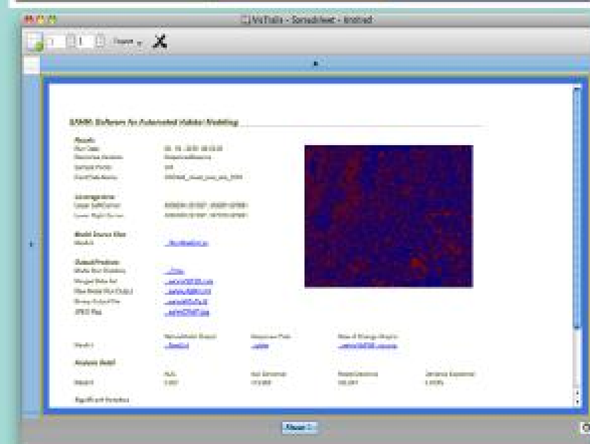


VisTrails — www.vistrails.org



Modeling the Spread of Invasive Species





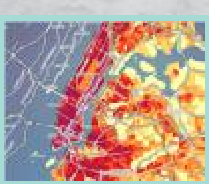
Data Science Applications - II

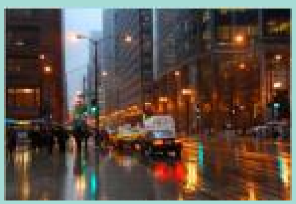
Urban Applications

Infrastructure | *Environment* | *People*



flickr


twitter

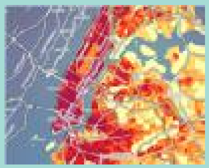


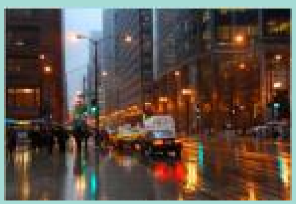

Sports Data Analytics

Data Science Applications - II

Urban Applications

Infrastructure | *Environment* | *People*



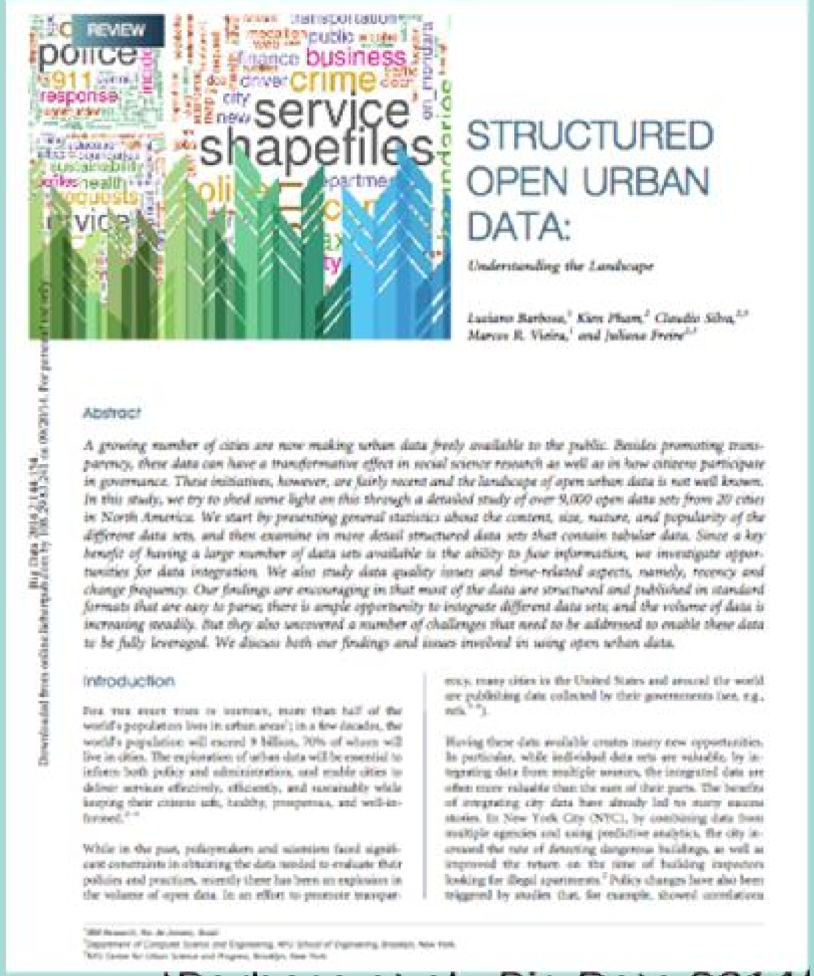


Applications in these areas are attractive to students since data is closer to their interests and they can tap into their personal experiences

Sports Data Analytics

Urban Data

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



REVIEW

STRUCTURED OPEN URBAN DATA:

Understanding the Landscape

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

Abstract

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

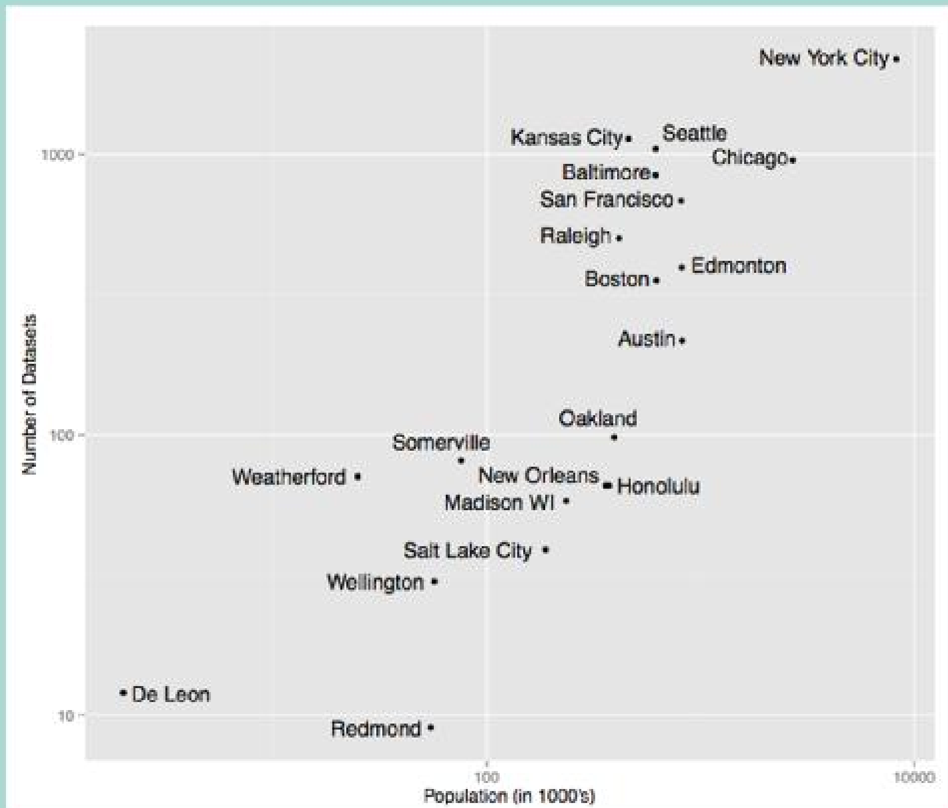
Introduction

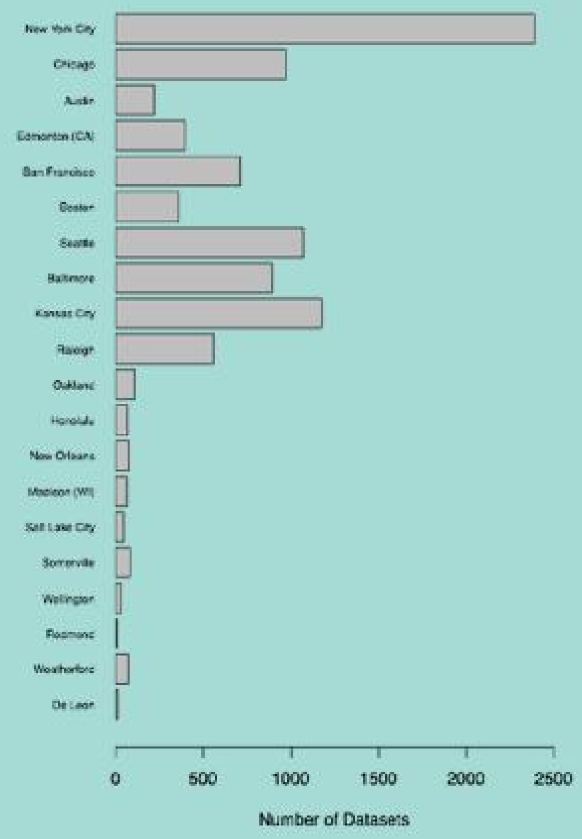
[Barbosa et al., Big Data 2014]

https://github.com/ViDA-NYU/urban-data-study

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





https://github.com/ViDA-NYU/urban-data-study

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.

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.

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,

h9gi-nx95 Details @ Urba ×
urbanprofiler.cloudapp.net/dataset/h9gi-nx95/columns

BETA Urban Profiler | Datasets | Search Map | Alarms | Overview

NYPD Motor Vehicle Collisions

Details of Motor Vehicle Collisions in New York City provided by the Police Department (NYPD)

Metadata | 29 Columns | Charts | Map | Related Datasets

Name	Provided Type	Type	Most Detected Typ
BOROUGH	text	Geo	Geo-BOROUGH 80
CONTRIBUTING FACTOR VEHICLE 1	text	Textual	Textual 91.5%
CONTRIBUTING FACTOR VEHICLE 2	text	Textual	Textual 91.3%
CONTRIBUTING FACTOR VEHICLE 3	text	Textual	Textual 94.4%
CONTRIBUTING FACTOR VEHICLE 4	text	Textual	Textual 100%
CONTRIBUTING FACTOR VEHICLE 5	text	Textual	Textual 100%
CROSS STREET NAME	text	Geo	Geo-Address 88.9%
DATE	calendar_date	Temporal	Temporal-Date 100
LATITUDE	number	Geo	Geo-Lat-or-Lon 100
LOCATION	location	Geo	Geo-GPS 100.0%
LONGITUDE	number	Geo	Geo-Lat-or-Lon 100
NUMBER OF CYCLIST INJURED	number	Numeric	Numeric-Integer 10
NUMBER OF CYCLIST KILLED	number	Numeric	Numeric-Integer 10

https://datahub.cusp.nyu.edu/

Taxi drivers petition NYC for fare hike over soaring gas prices

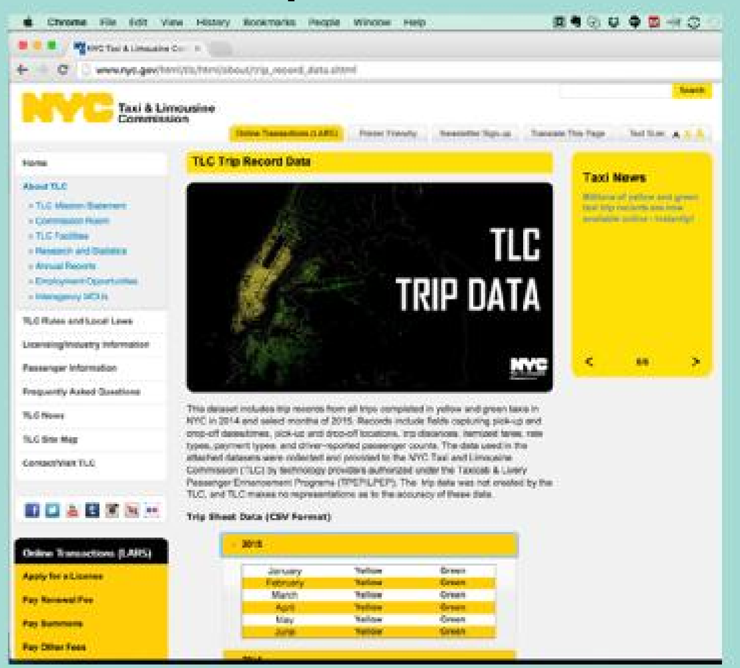
BY PETE DONOHUE / DAILY NEWS STAFF WRITER

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

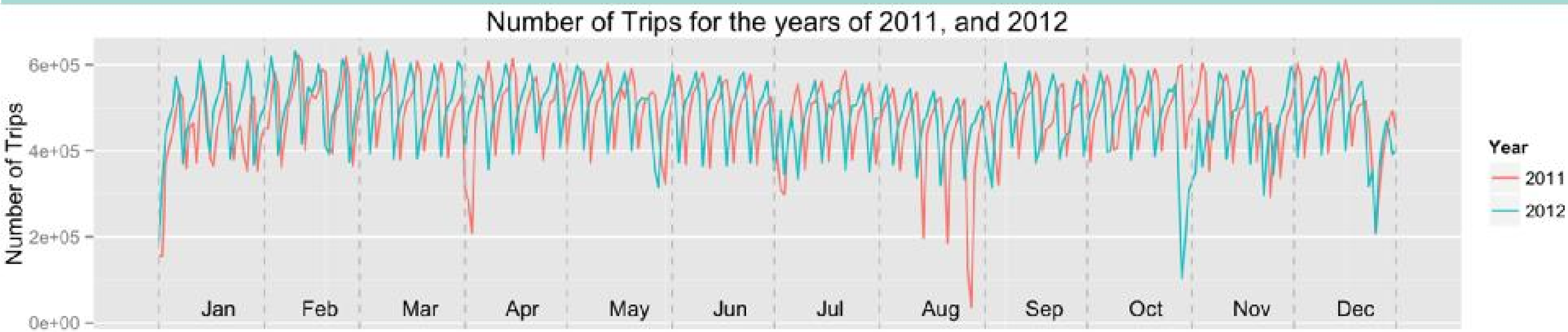
ilynews.com/new-york

NYC Taxi Data

- It is relatively *big*: ~500k trips/day - several hundred million trips in 5 years
- ... and relatively *complex*:
  – *spatio-temporal:* pick up + drop off
  – *trip attributes*: e.g., distance traveled, cost, tip
- Many data slices to examine



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?"

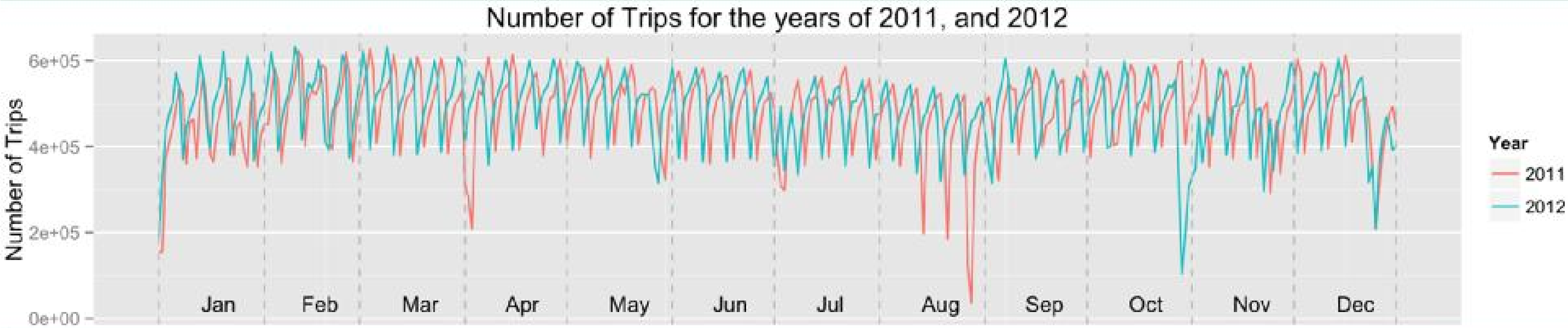
  "What is the average trip time from Midtown to the airports during weekdays?"

  "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



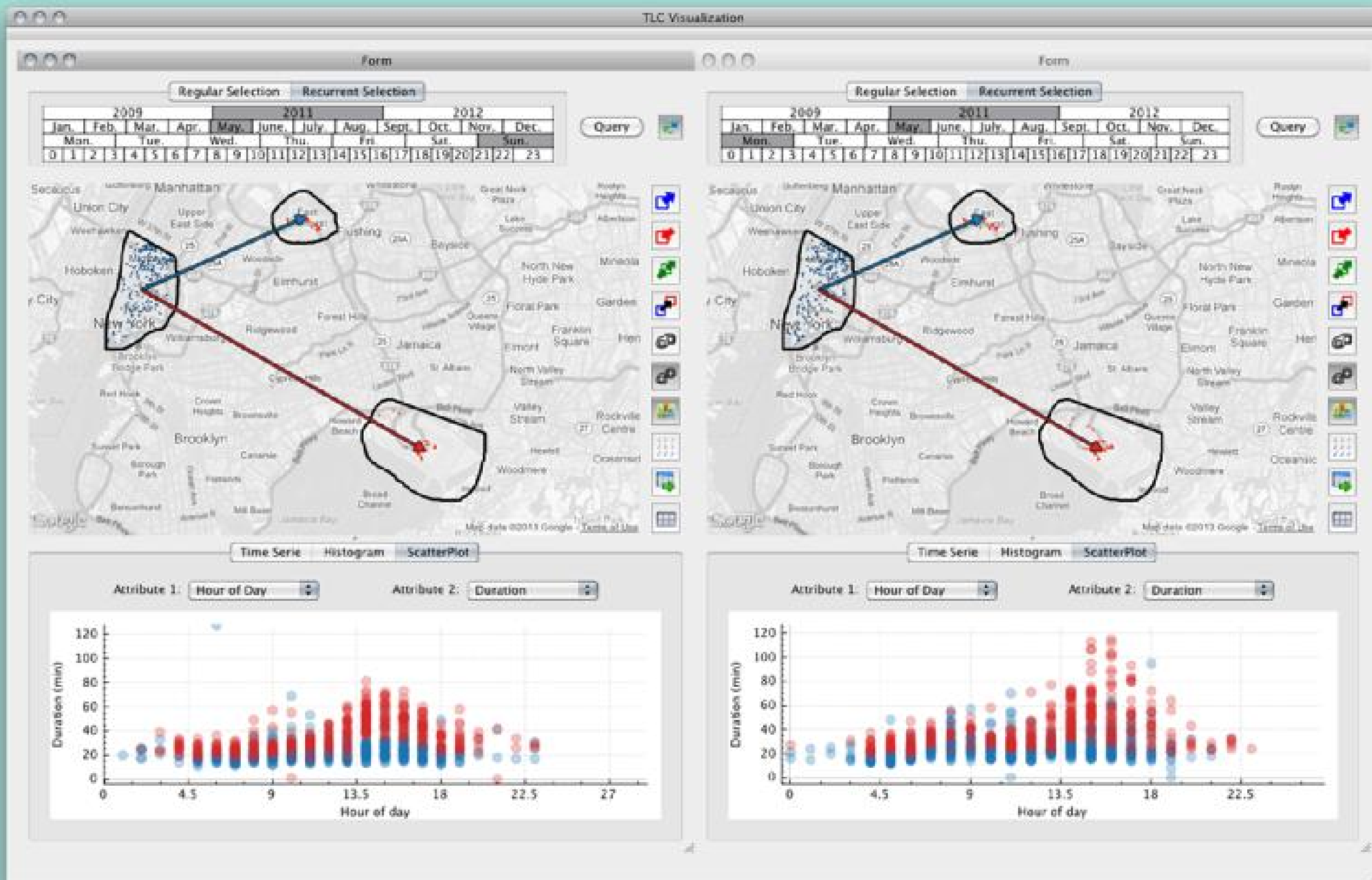
7-8am

8-9am

9-10am

10-11am

Looking at NYC Taxi Records



https://github.com/ViDA-NYU/TaxiVis

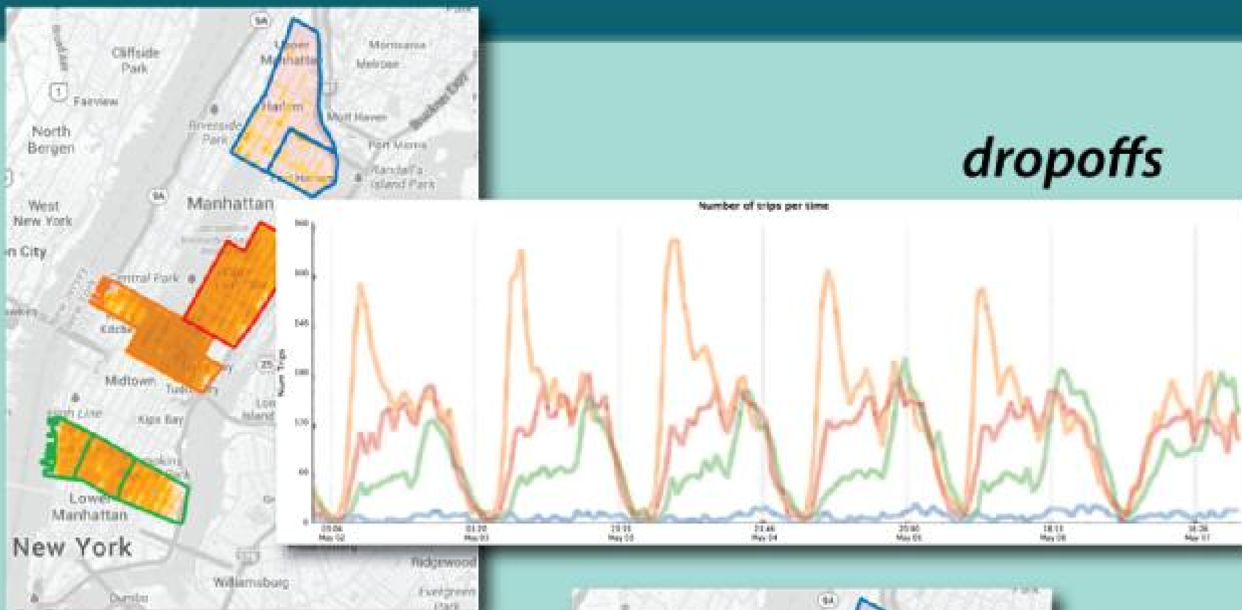
NYU TANDON SCHOOL OF ENGINEERING

[Ferreira et al., IEEE TVCG 2013]

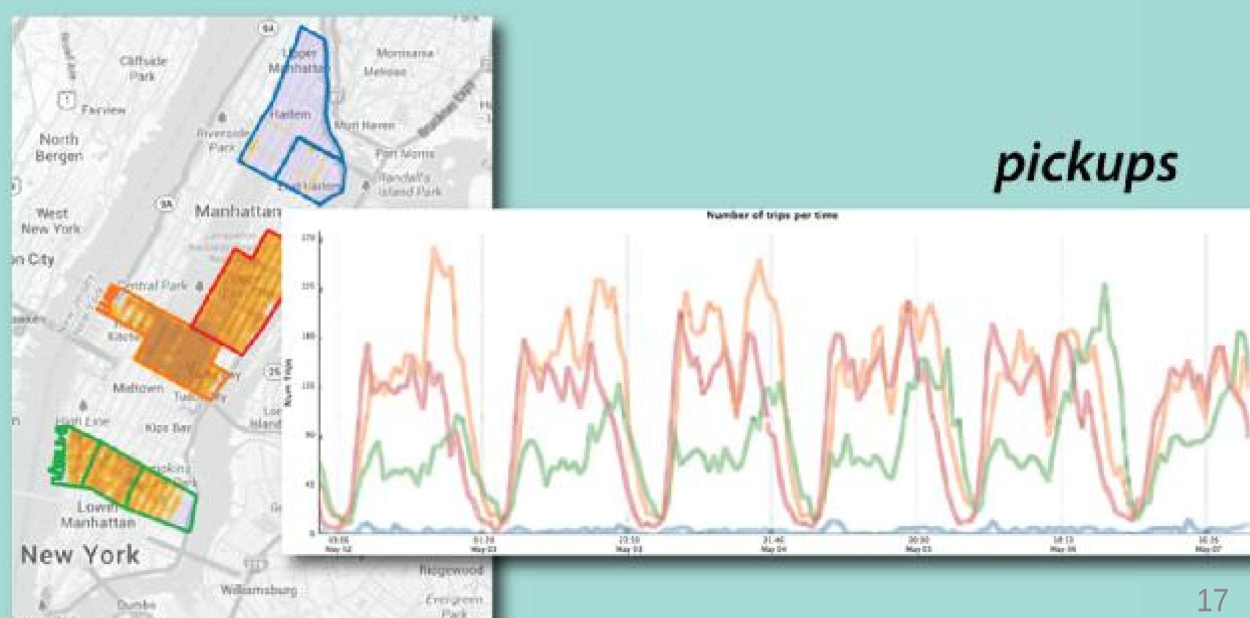
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

TaxiVis: Comparing Neighborhoods

dropoffs



pickups

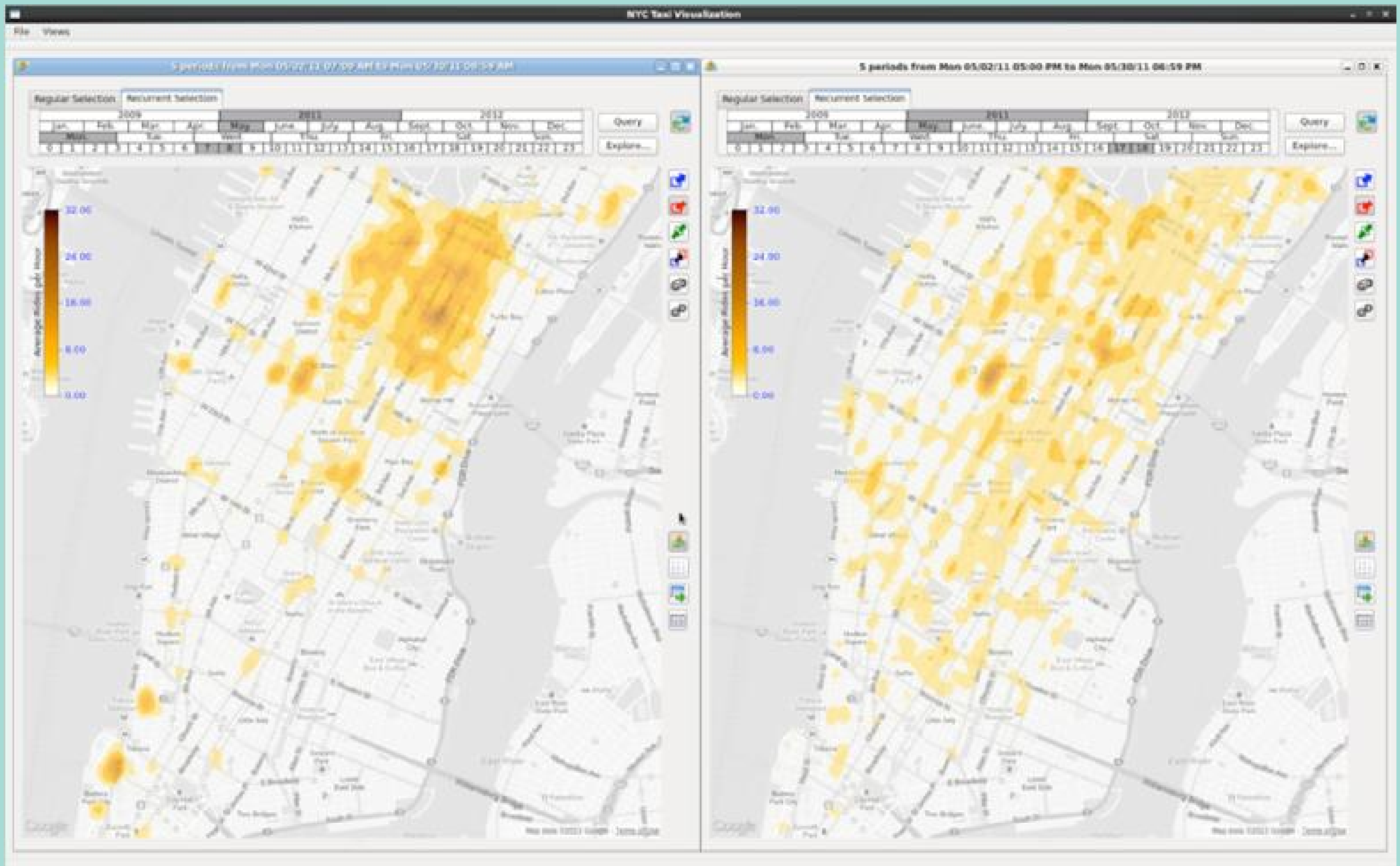


Time Exploration

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

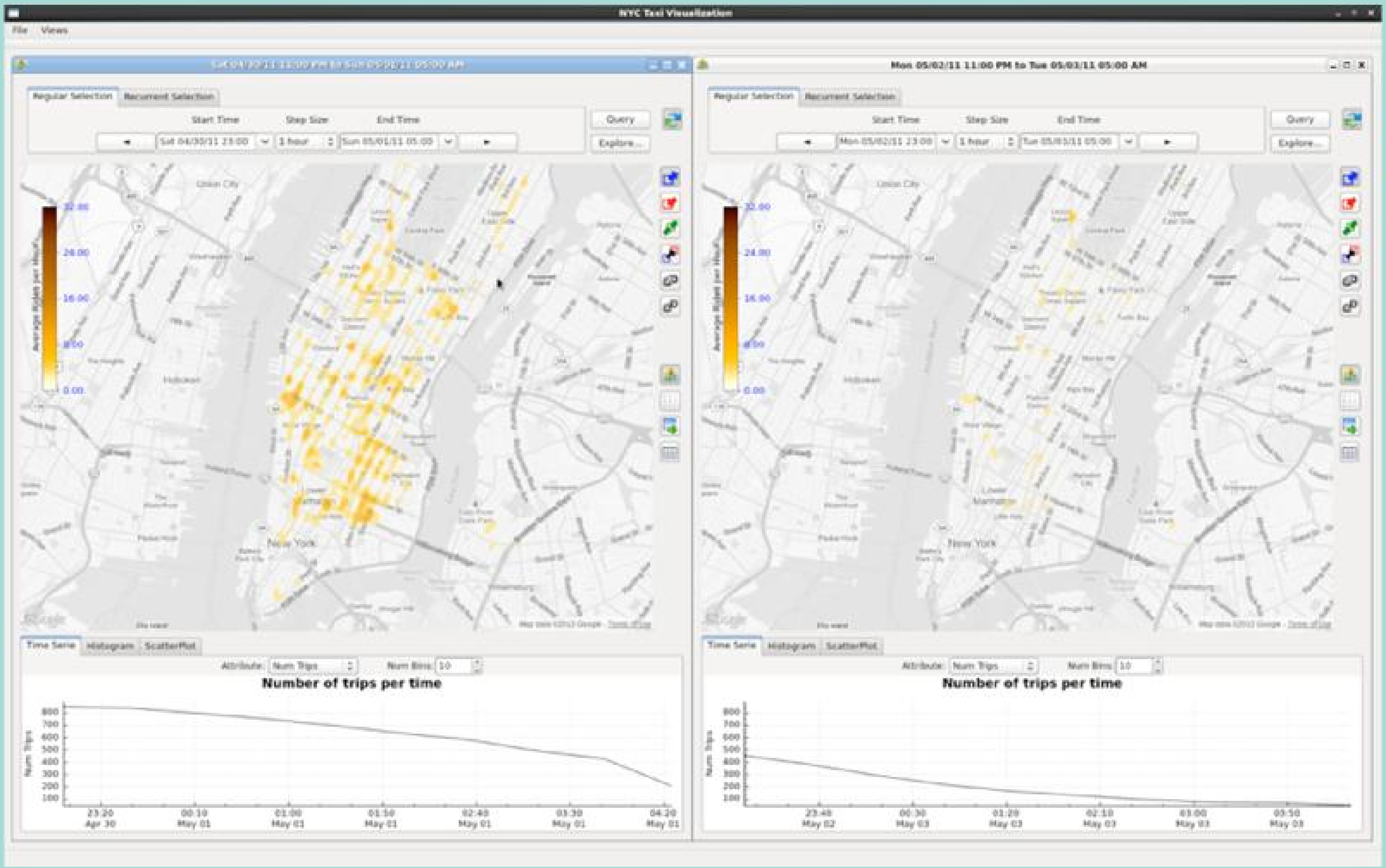
Dropoffs Before vs. After Work



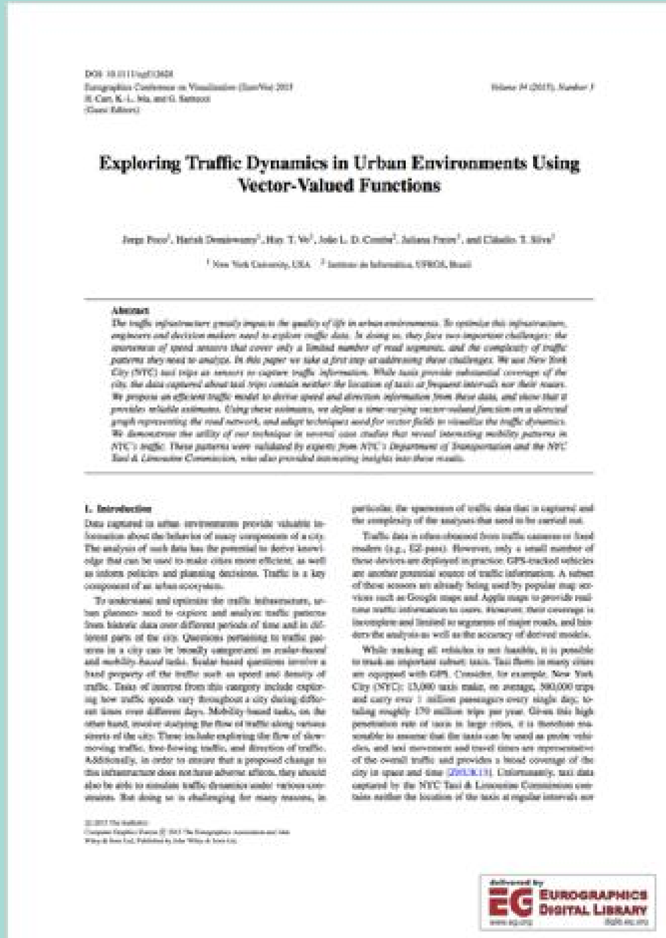
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Night Life Saturday vs. Monday



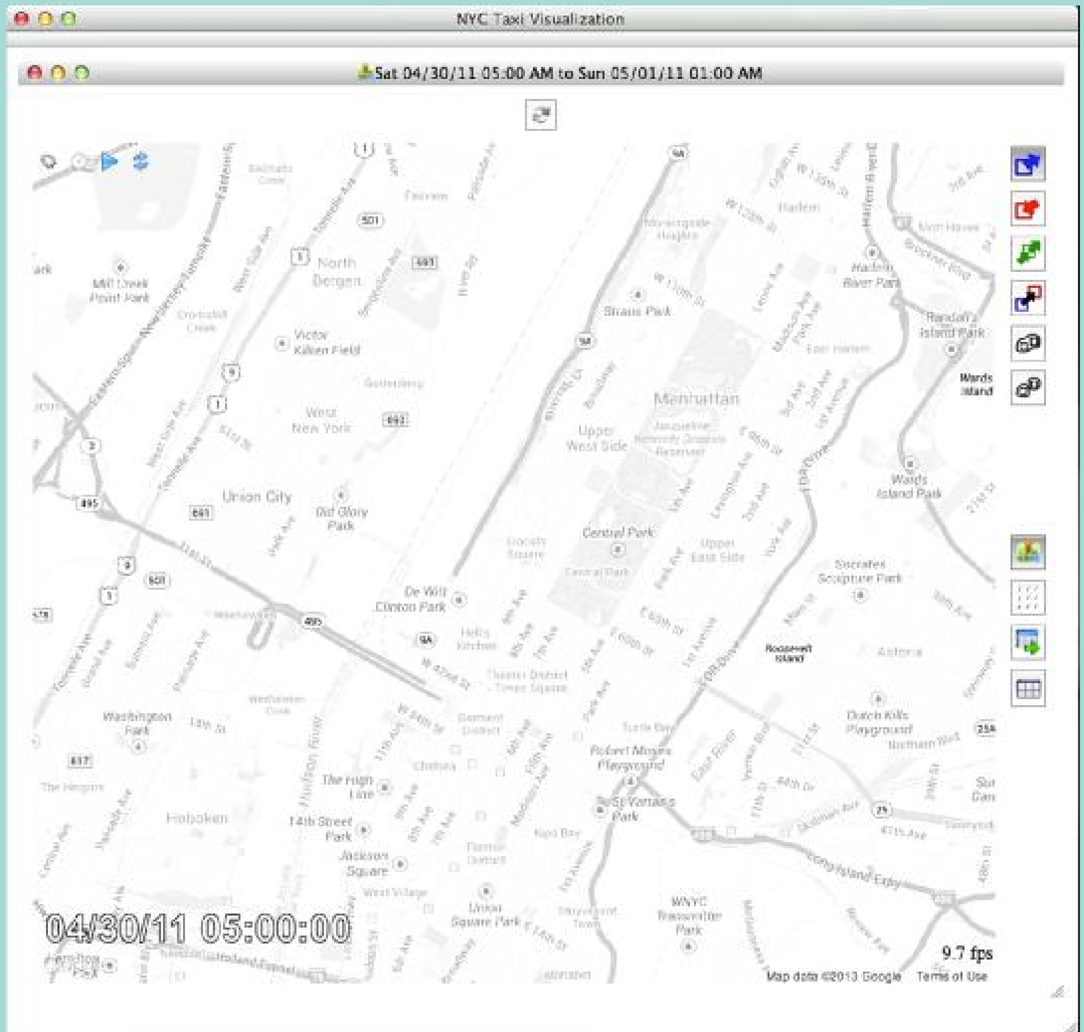
A Taxi over 24 hours



Exploring Traffic Dynamics in Urban Environments Using Vector-Valued Functions

Abstract

1. Introduction



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

Student Course Projects

The Daily Commute: An In-depth Analysis of Manhattan Traffic Patterns Between Yellow Cab, Uber, and CitiBike

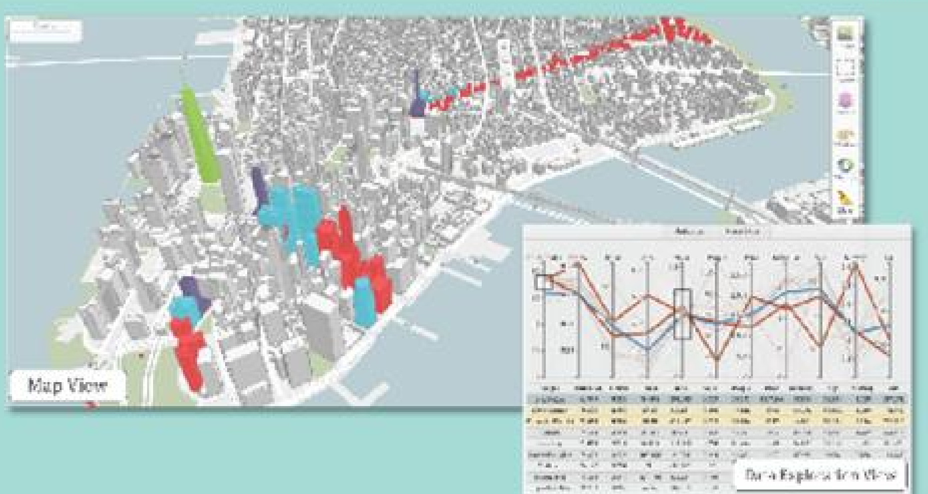
Crime Analysis in New York City 2006-2015

Detecting Gentrification with Taxi Patterns in NYC

Optimizing Walking Paths Based on Interestingness

NYC Real Estate Price Prediction

Projects with 3D data



[Ferreira et al., IEEE VAST 2015]

https://www.nytimes.com/interactive/2016/12/21/upshot/Mapping-the-Shadows-of-New-York-City.html?mcubz=0

Projects with other data modalities, e.g., sound

https://wp.nyu.edu/sonyc/

Thank you!

csilva@nyu.edu

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications

Cláudio T. Silva, New York University
Professor of computer science and engineering and data science

Using Urban and Sports Data in Student Projects

Q&A

Envisioning the
DATA SCIENCE DISCIPLINE
The Undergraduate Perspective

Incorporating Real-World Applications

Building a talent pipeline through a strategic career development program & academic-industrial partnerships

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of
Data Science & Advanced Analytics at
MassMutual Financial Group

Provide input and learn more about the study at *www.nas.edu/EnvisioningDS*

Building a talent pipeline

Academic Partnerships

Development Program
- Support creation of undergraduate programs
- Blend academic rigor with industry application
- Integrate academic approach to development

Research and Senior Talent
- Senior data science talent generation
- Collaborative research opportunities

Junior Data Scientist → Data Scientist → Senior Data Scientist → Lead Data Scientist

Year 1: Portfolio Assembly
- Enroll in MS program
- Participate in 2 projects

Year 2: Communicating Results
- Participate in 2 projects
- Continue course work
- Defend results

Year 3: Leading a Project
- Complete coursework
- Identify problem in business
- Scope and assess value
- Executed with junior member
- Defend project

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

Incorporating Real-World Applications – Q&A

Cláudio T. Silva, New York University
Professor of computer science and engineering and data science

Sears Merritt, MassMutual Financial Group
Chief Data Scientist and head of Data Science & Advanced Analytics at *MassMutual Financial Group*

Provide input and learn more about the study at www.nas.edu/EnvisioningDS

Envisioning the

DATA SCIENCE DISCIPLINE

The Undergraduate Perspective

9/12/17 – Building Data Acumen
(recording posted)

9/19/17 – Incorporating Real-World Applications

9/26/17 – Faculty Training and Curriculum Development

10/3/17 – Communication Skills and Teamwork

10/10/17 – Inter-Departmental Collaboration and Institutional Organization

10/17/17 – Ethics

10/24/17 – Assessment and Evaluation for Data Science Programs

11/7/17 – Diversity, Inclusion, and Increasing Participation

11/14/17 – Two-Year Colleges and Institutional Partnerships

Provide input and learn more about the study at www.nas.edu/EnvisioningDS