Introduction to Urban Data Science
Lecture 3

Topological Data Analysis: Applications to Urban Data

Harish Doraiswamy

New York University

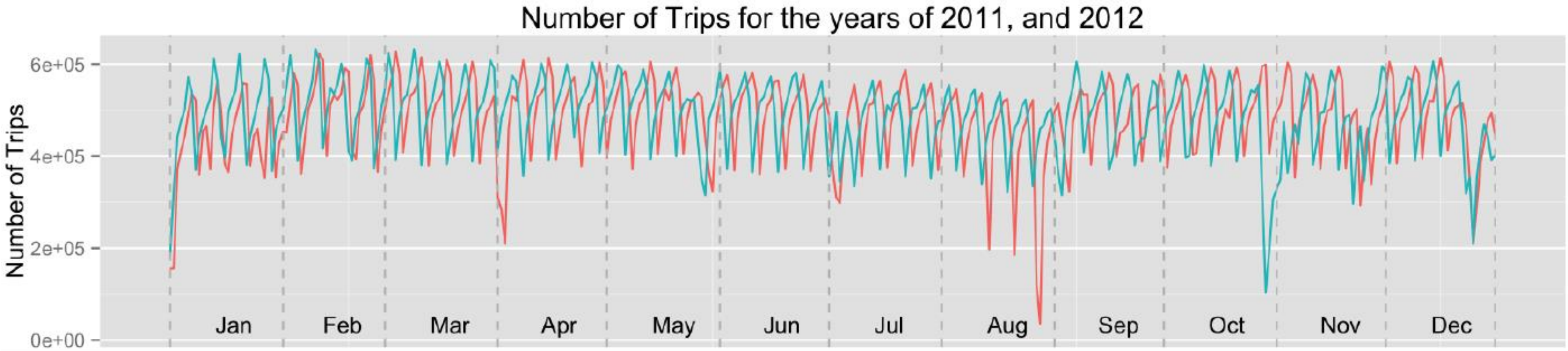
NYU

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

NYC Taxi Data

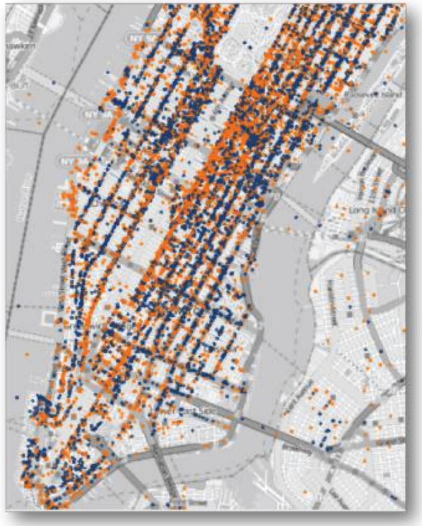
- Yellow cab trips
- ~175 million trips / year
- Spatial-Temporal
  - 2 spatial attributes
  - 2 temporal attributes
- Other attributes
  - Fare, tip
  - Distance
  - Duration
  - …

Analysis: Example



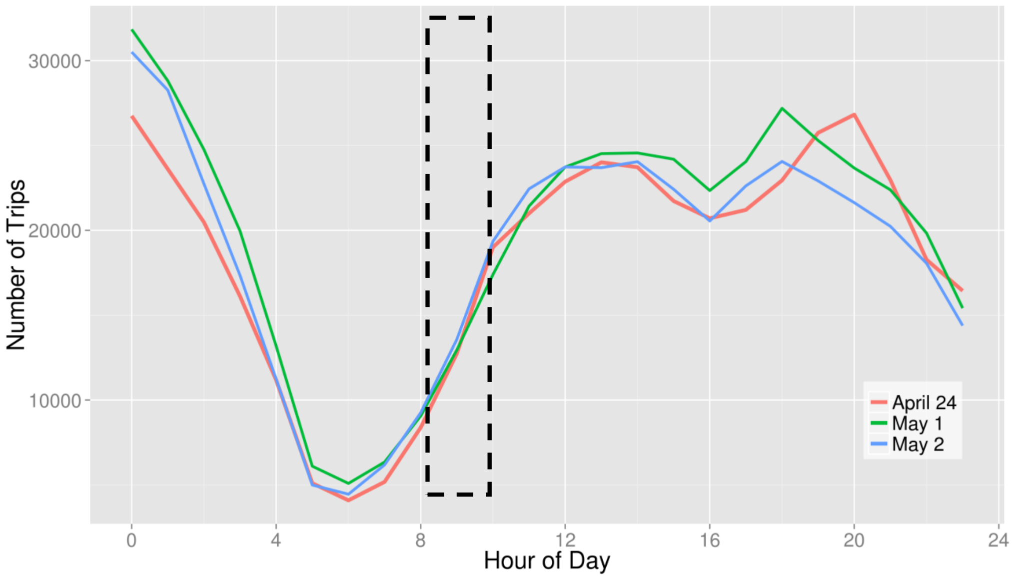

7am - 8am


8am - 9am


9am - 10am

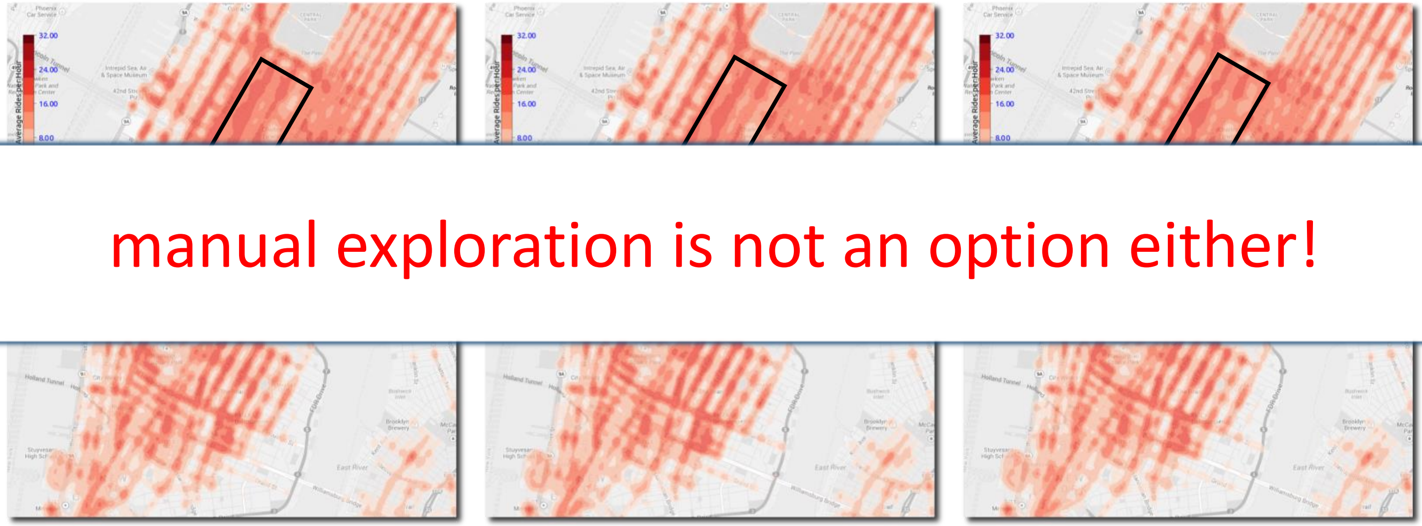

10am - 11am

Aggregate over **Space**



8am - 9am

Aggregate over **Time**



manual exploration is not an option either!

April 24 May 1 May 8

Goal

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

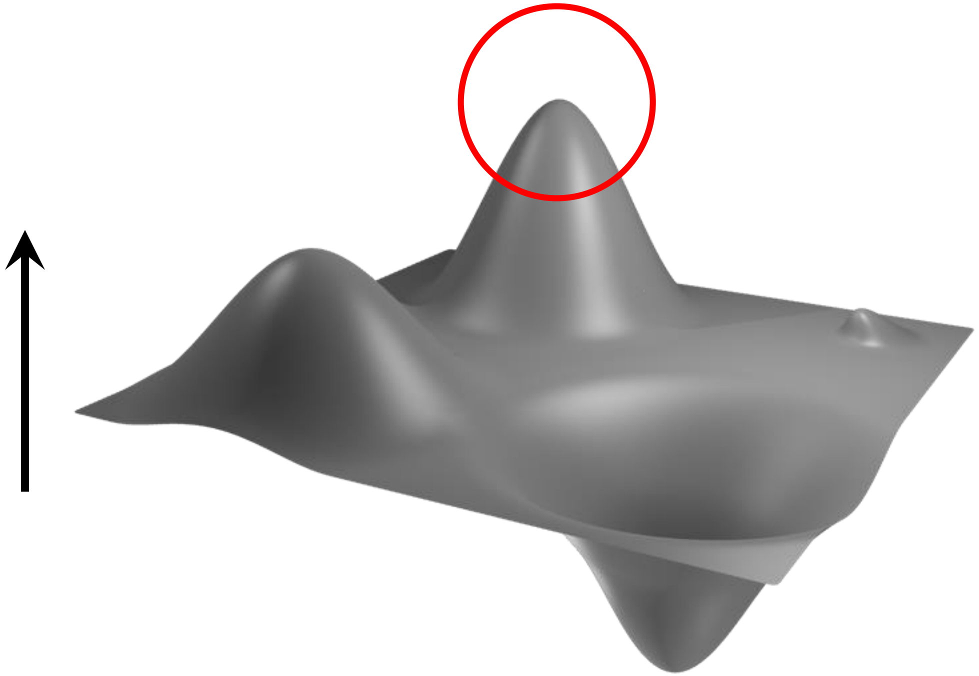
Harish Doraiswamy, Nivan Ferreira, Theodoros Damoulas, Juliana Freire, Cláudio T. Silva

IEEE TVCG 2014

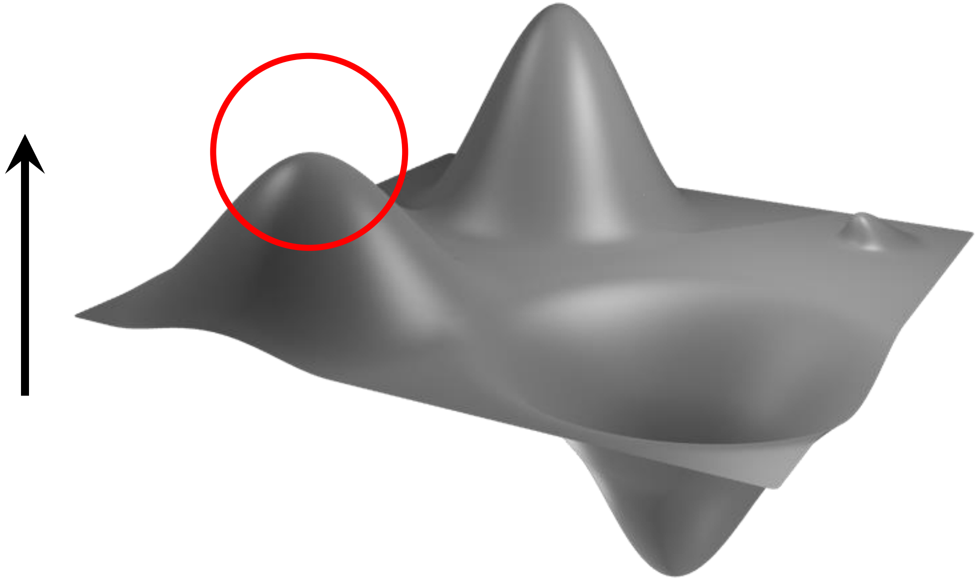
- Guide users towards potentially interesting data slices
- What is an interesting data slice?
  - Contains an "event"
- Flexible definition of events
  - Arbitrary spatial structure
  - Different types of events
  - Multiple temporal scales
- Efficient search for similar event patterns

8am - 9am

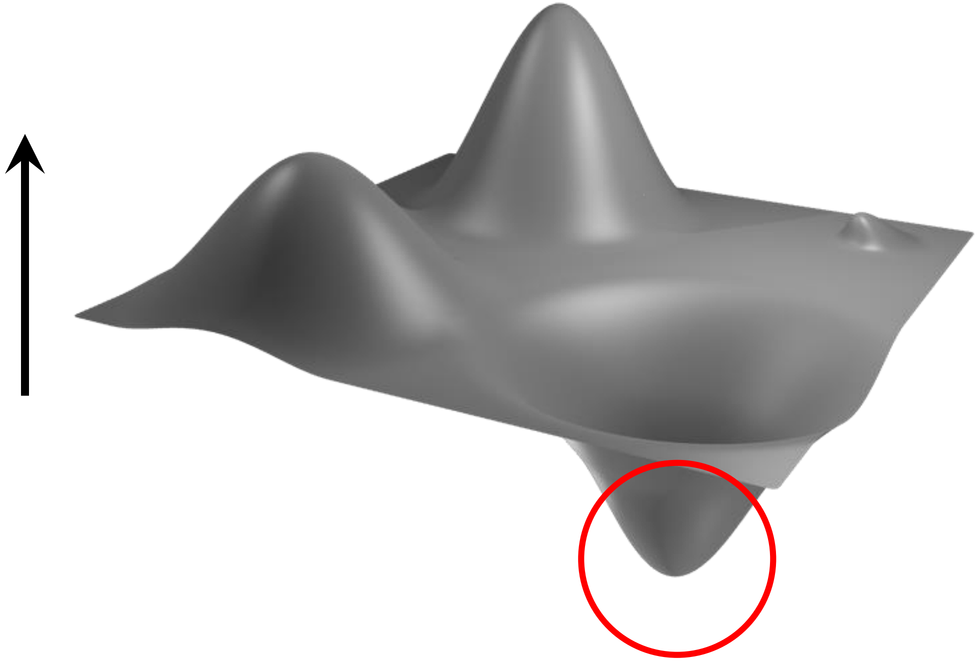
Idea: Use **Topology** of the Data

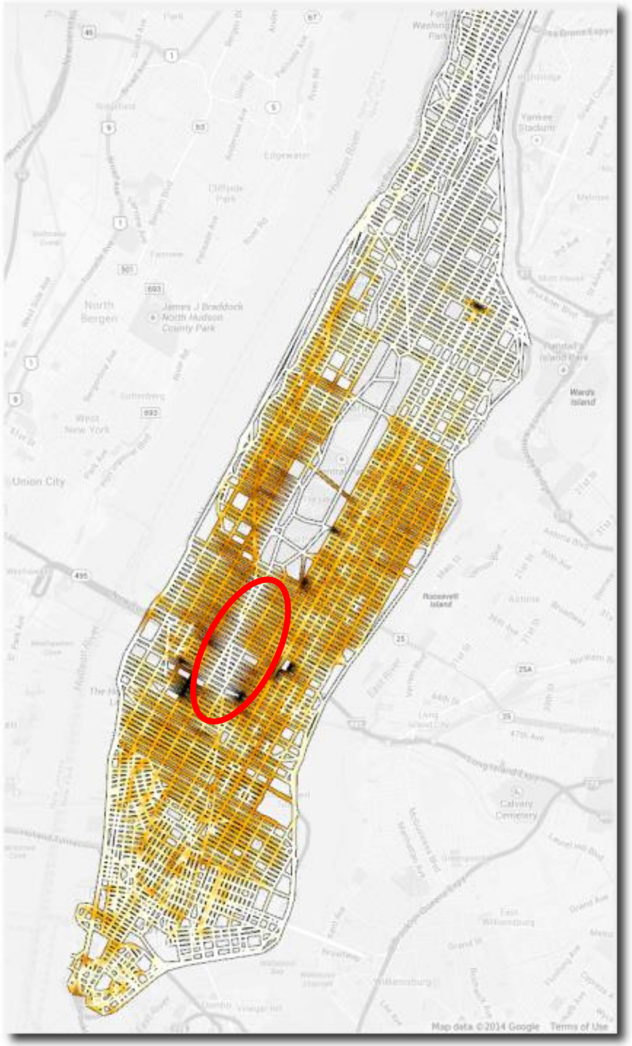


Idea: Use **Topology** of the Data



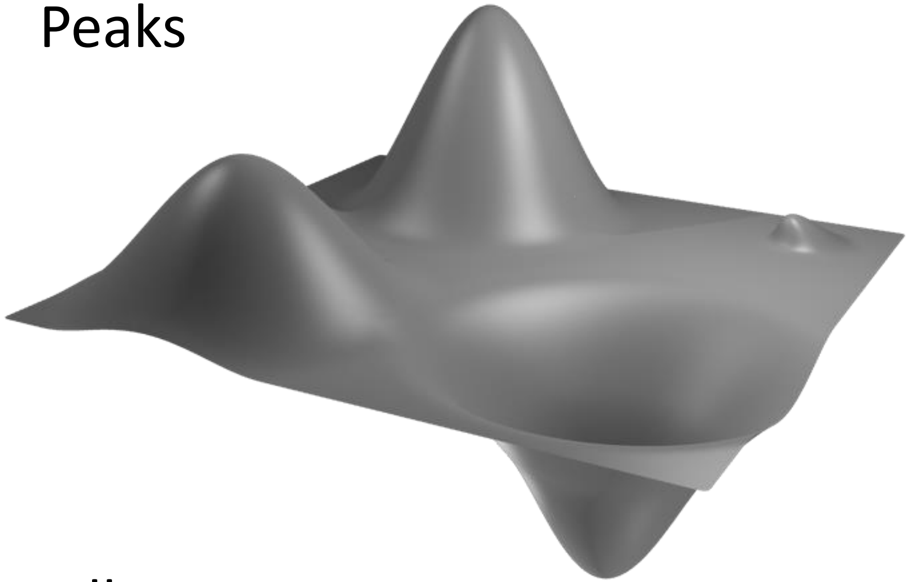
Idea: Use **Topology** of the Data



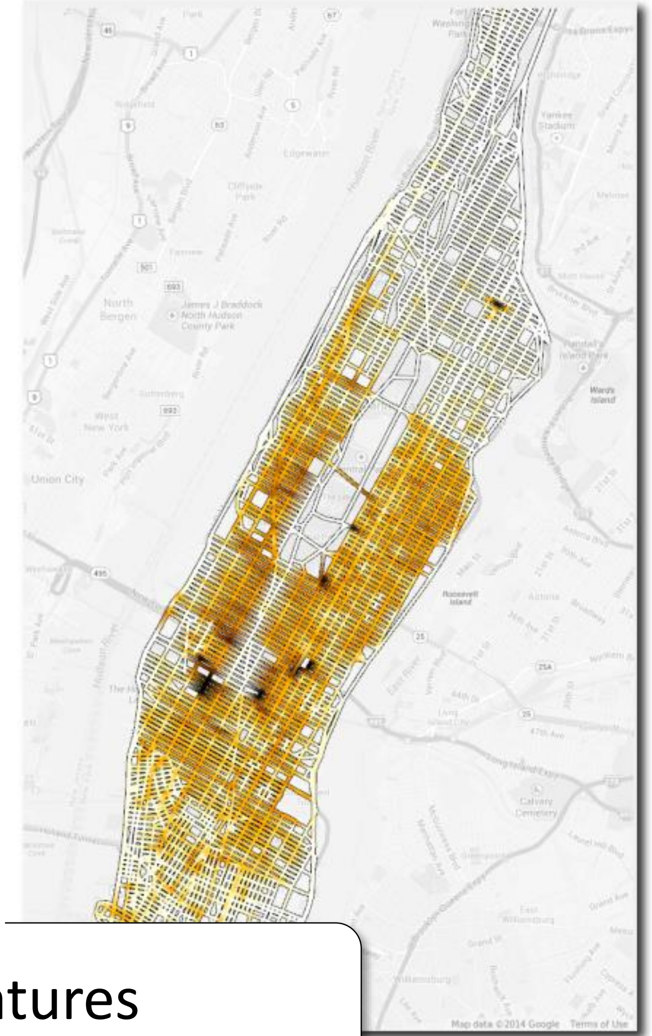


Idea: Use **Topology** of the Data

Peaks



Valleys



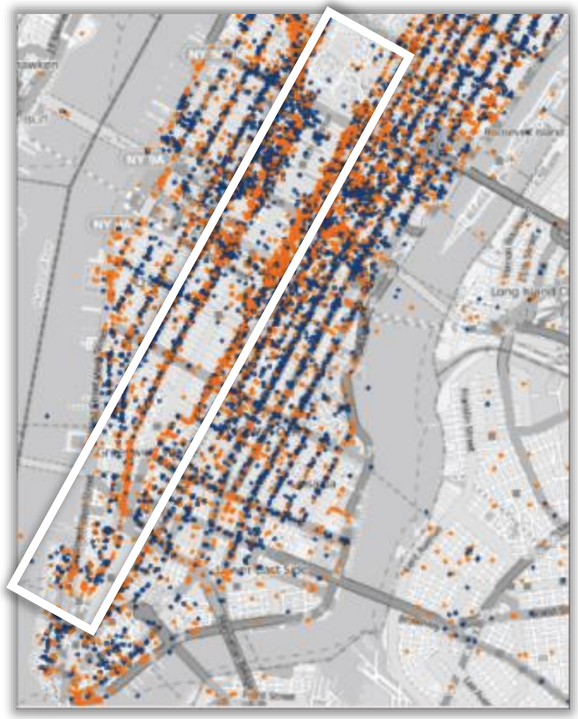
Advantage

1. Naturally captures such features

Identifying Topological Features

8am - 9am
May 1 2011

5 Boro Bike Tour



Valleys

Advantage

2. Features can have arbitrary shapes

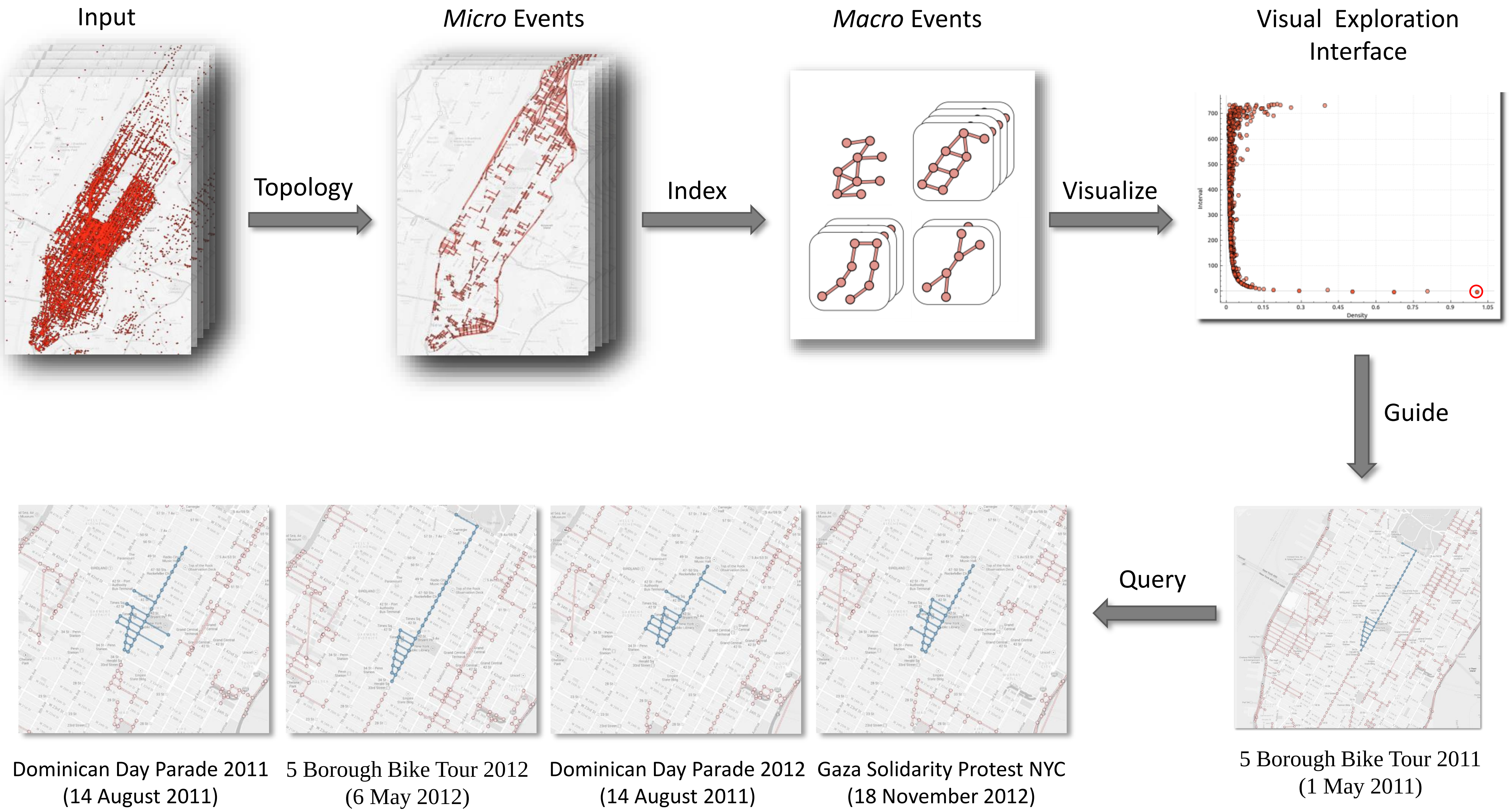
Using Topology: Advantages

1. Naturally captures such features

2. Features can have arbitrary shapes

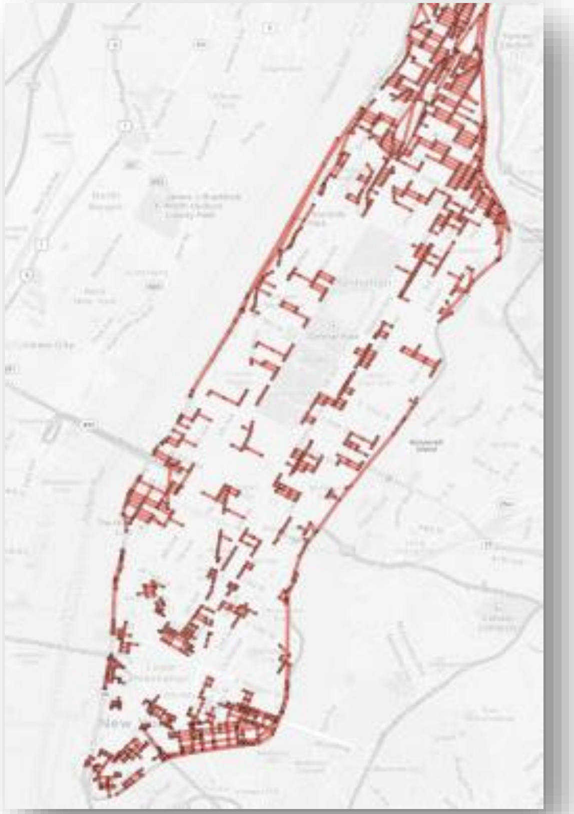
3. Very efficient

4. Effectively handle noisy data


Input
Micro Events
Macro Events
Visual Exploration Interface
Topology
Index
Visualize
Guide
Query
Interval
Density
Dominican Day Parade 2011 (14 August 2011)
5 Borough Bike Tour 2012 (6 May 2012)
Dominican Day Parade 2012 (14 August 2011)
Gaza Solidarity Protest NYC (18 November 2012)
5 Borough Bike Tour 2011 (1 May 2011)

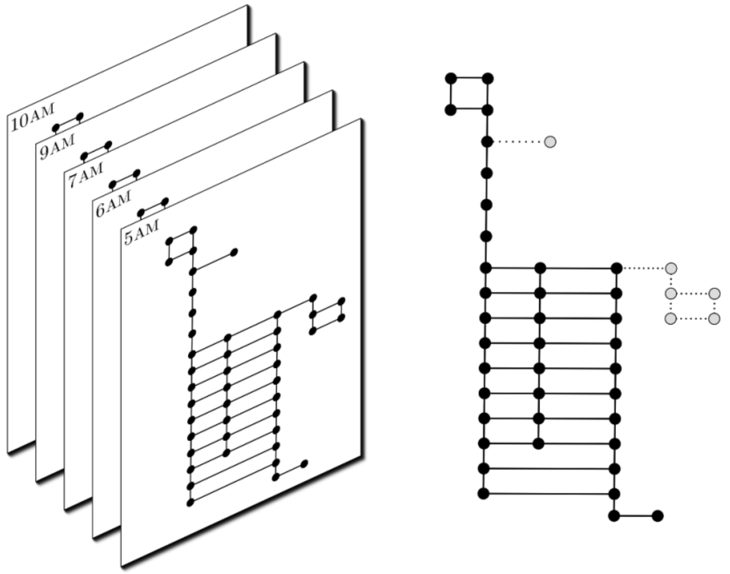
Macro Events

- Several features per time step
- Group similar features within a larger time interval
  - Represents “macro” events



Macro Events

- Several features per time step
- Group similar features within a larger time interval
  - Represents “macro” events
- Similarity
  - Geometric similarity: Shape
  - Topological similarity: Volume



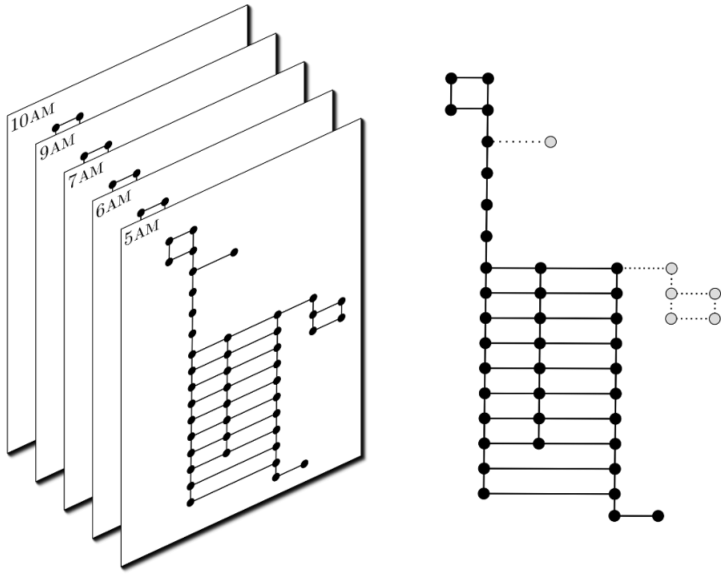
Macro Events

- Several features per time step
- Group similar features within a larger time interval
  - Represents "macro" events
- Similarity
  - Geometric similarity: Shape
    - Graph distance metric

$$\delta(E_1, E_2) = 1 - \frac{|R_1 \bigcap R_2|}{\max(|R_1|, |R_2|)}$$

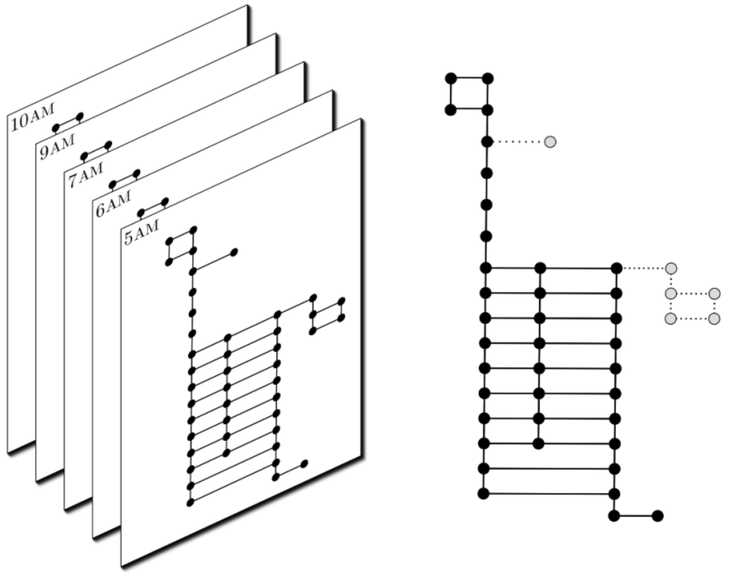
  - Topological similarity: Volume

$$T(E_1, E_2) = |\tau_1 - \tau_2|$$



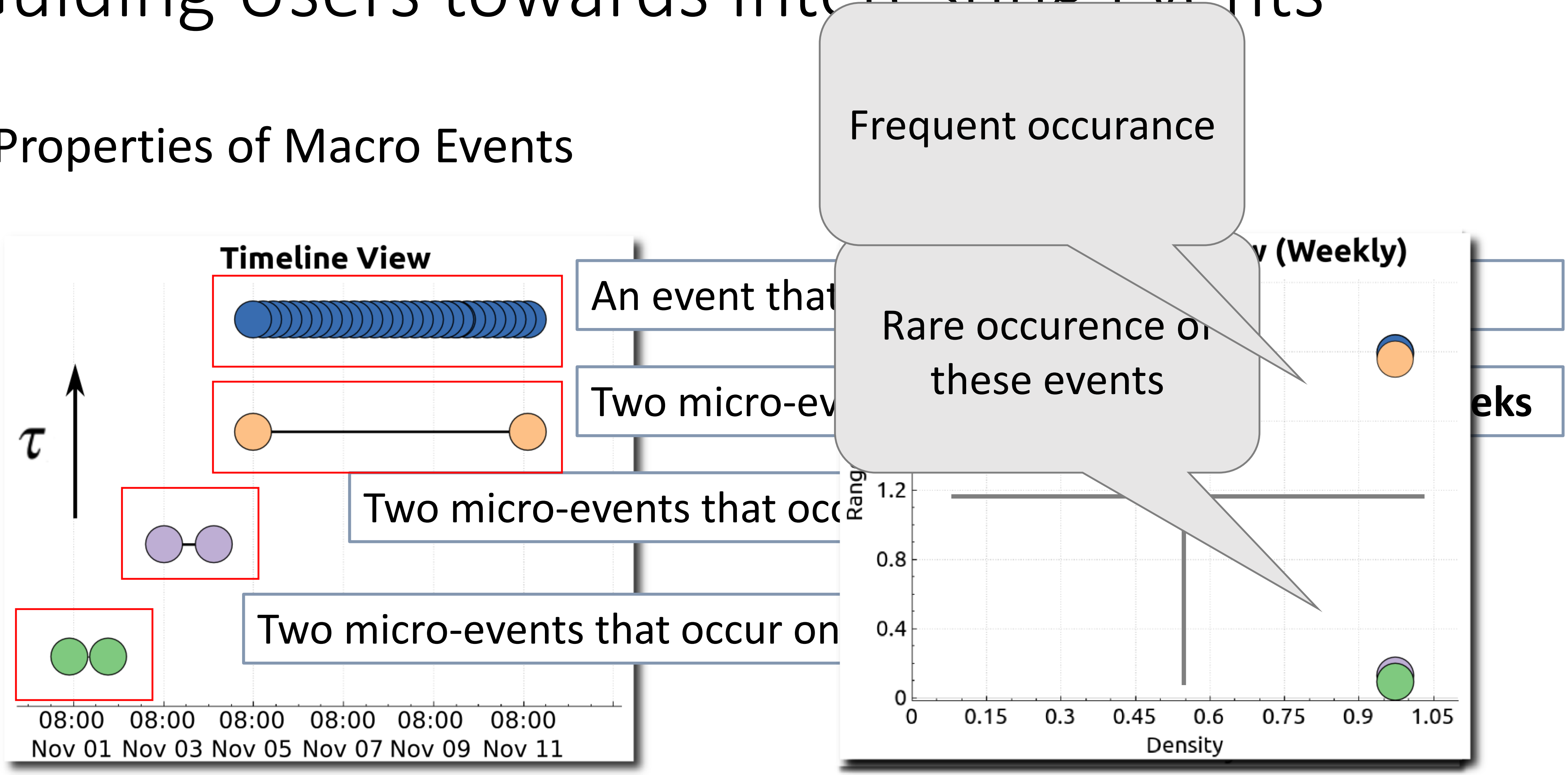
Macro Events

- Several features per time step
- Group similar features within a larger time interval
  - Represents "macro" events
- Similarity
  - Geometric similarity: Shape
  - Topological similarity: Volume
- **Key** for each group
  - *Average* shape and volume
  - Efficient search



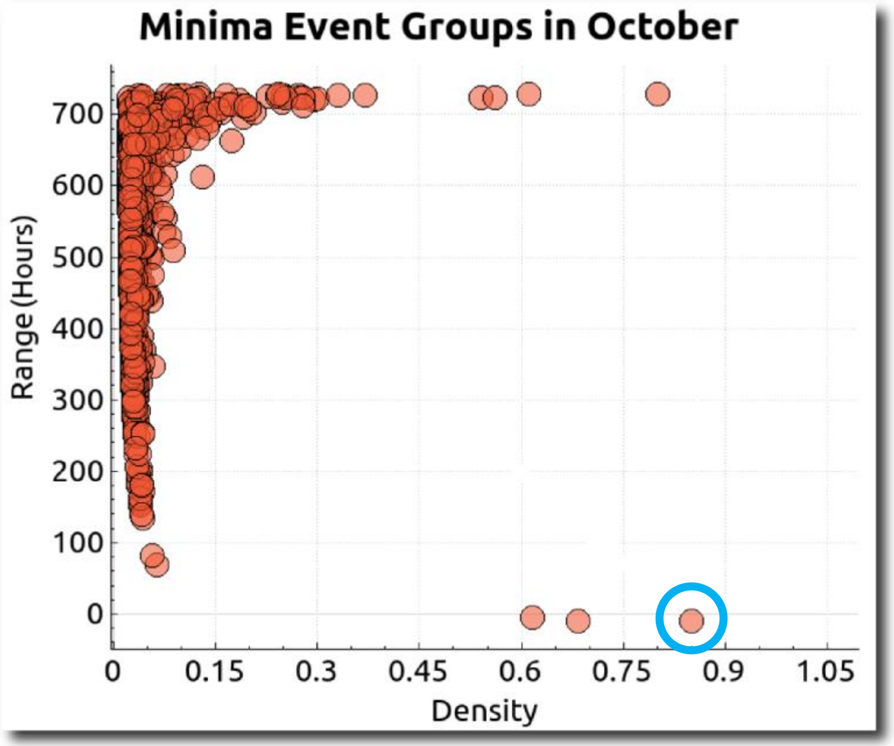
Guiding Users towards Interesting Events

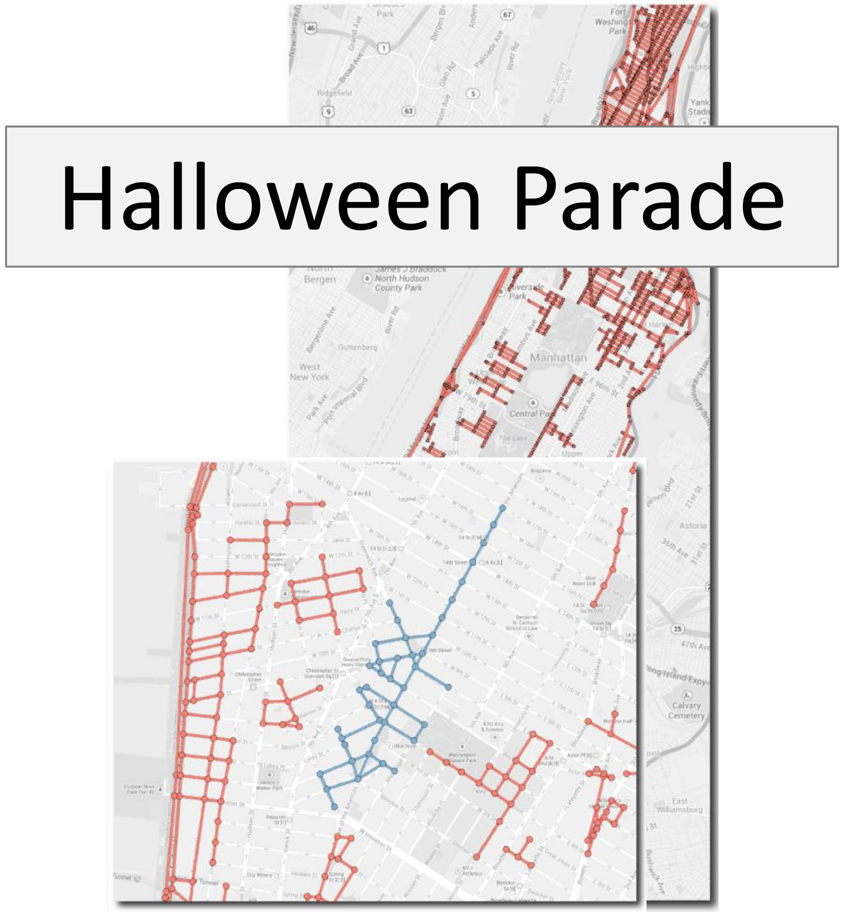
- Properties of Macro Events



Rare Events - Hourly

- October

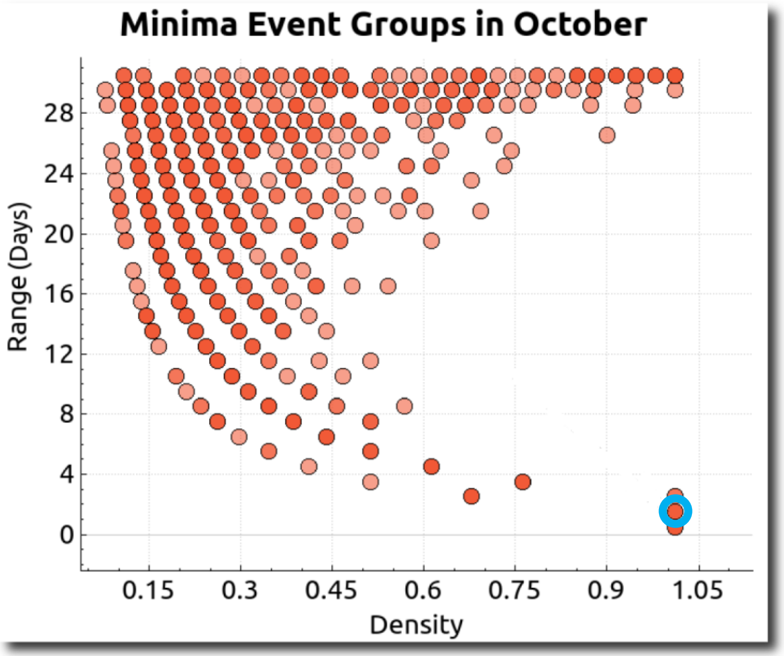


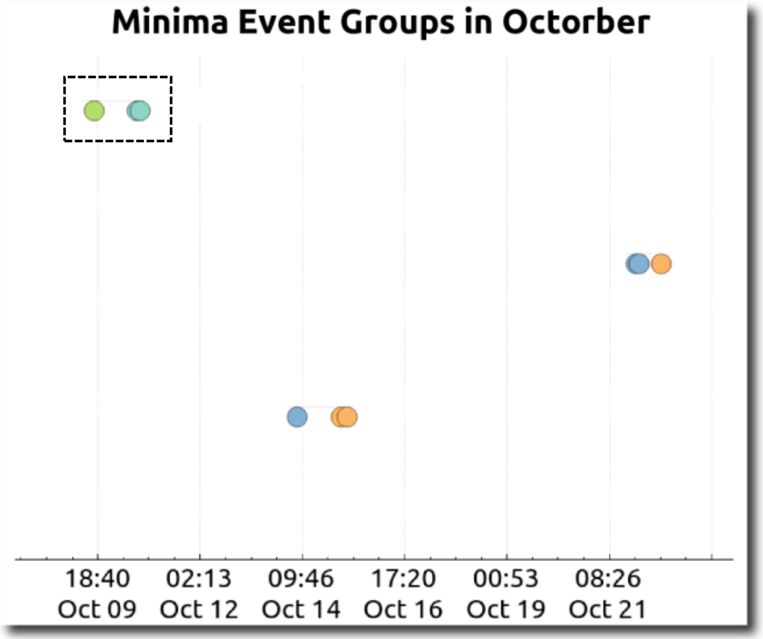


Rare Events - Daily

- October

1. Hispanic Day Parade (Oct 9 2011)
2. Columbus Day Parade (Oct 10 2011)





Frequent Events

- Maxima: Taxi hotspots
- Filter over time

General trends

Night time trends

Event-Guided Exploration

5 Borough Bike Tour 2011
(1 May 2011)

Go to Time slice

Similarity Search

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)

Similarity Search

Hispanic Day Parade 2011
(9 Oct 2011)

Query

St. Patrick's Day Parade 2011

Pulaski Day Parade 2011

Labor Day Parade 2011

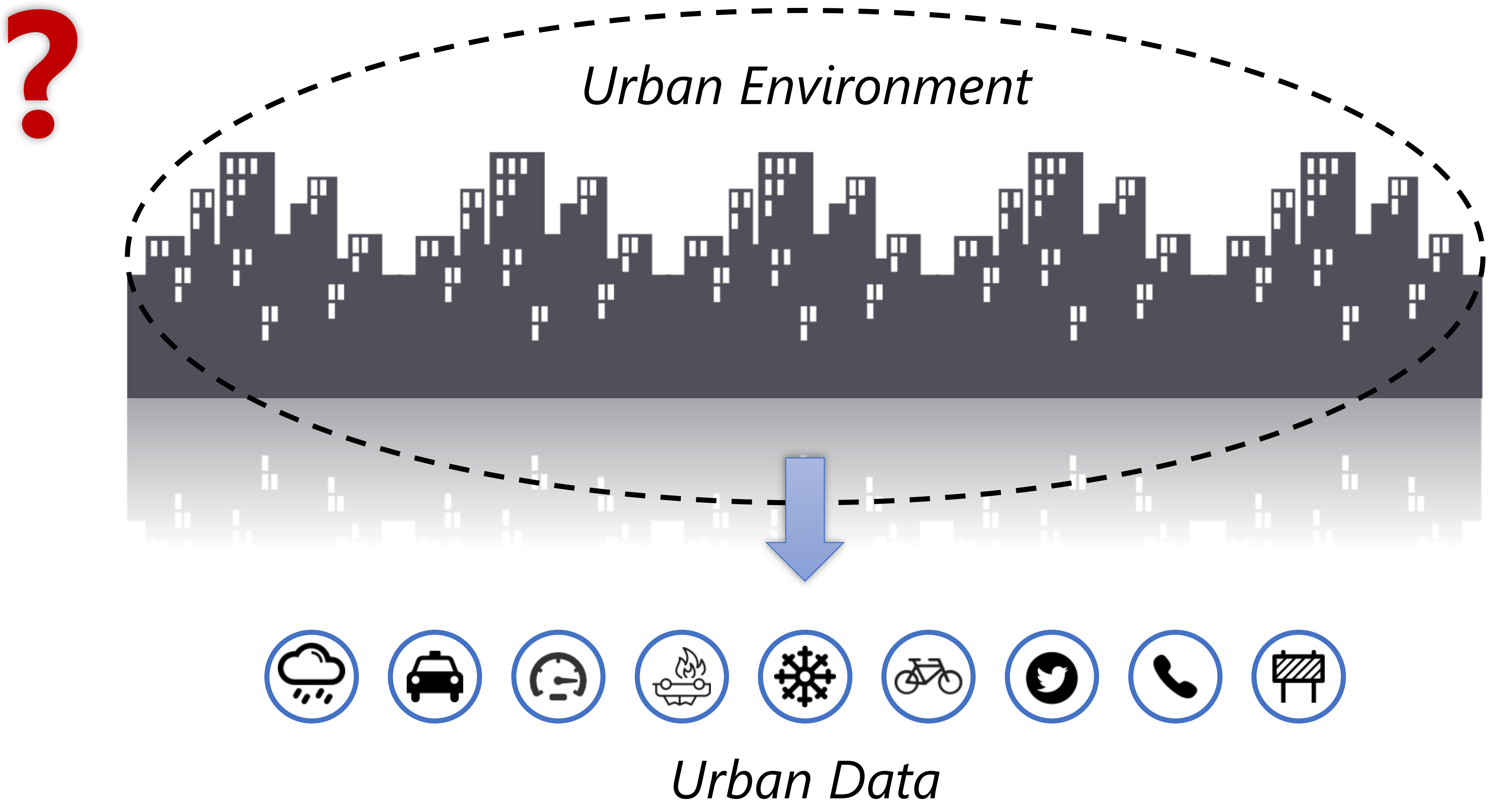
Labor Day Parade 2012

Columbus Day Parade 2012

Hispanic day parade 2012

Veterans Day Parade 2012

Event Guided Exploration
Hourly Events



how can we use multiple data sets to understand the city

Objective

How to compare cities?

- Design of public spaces
  - Understand what works / doesn't work in one city
  - Use this to improve design in another city

Union Square

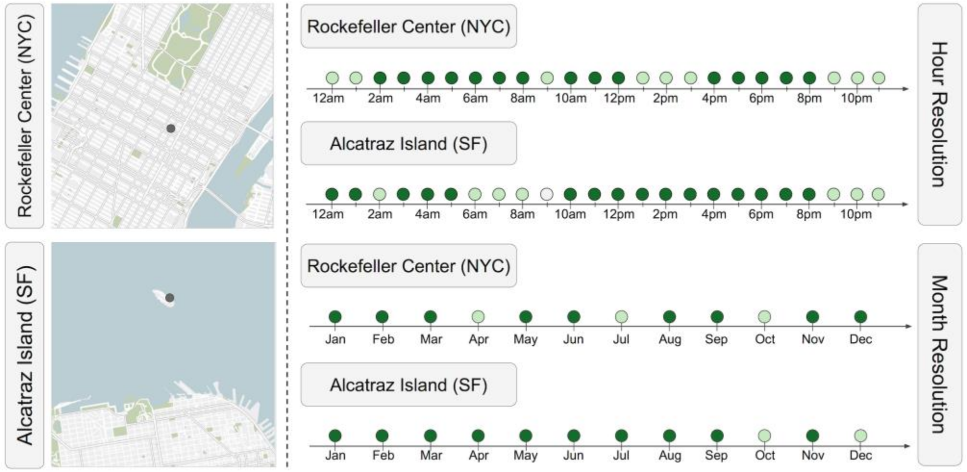
Objective

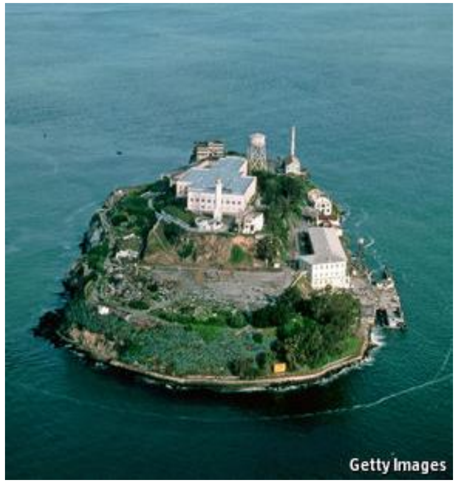
How to analyze / compare different properties of a city?

- How do cities behave during different times?
  - Summer vs. Winter
  - Weekdays vs. Weekends

Greenwich Village

Urban Pulse





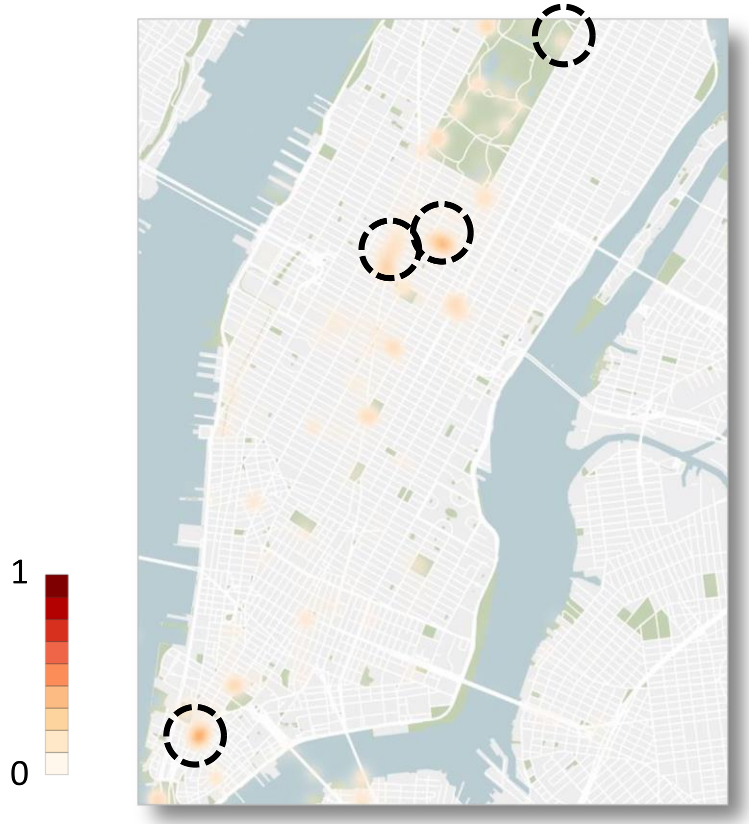
Urban Pulse: Capturing the Rhythm of Cities

Fabio Miranda, Harish Doraiswamy, Marcos Lage, Kai Zhao, Bruno Gonçalves, Luc Wilson, Mondrian Hsieh, Cláudio Silva

IEEE TVCG 2017

Urban Pulse

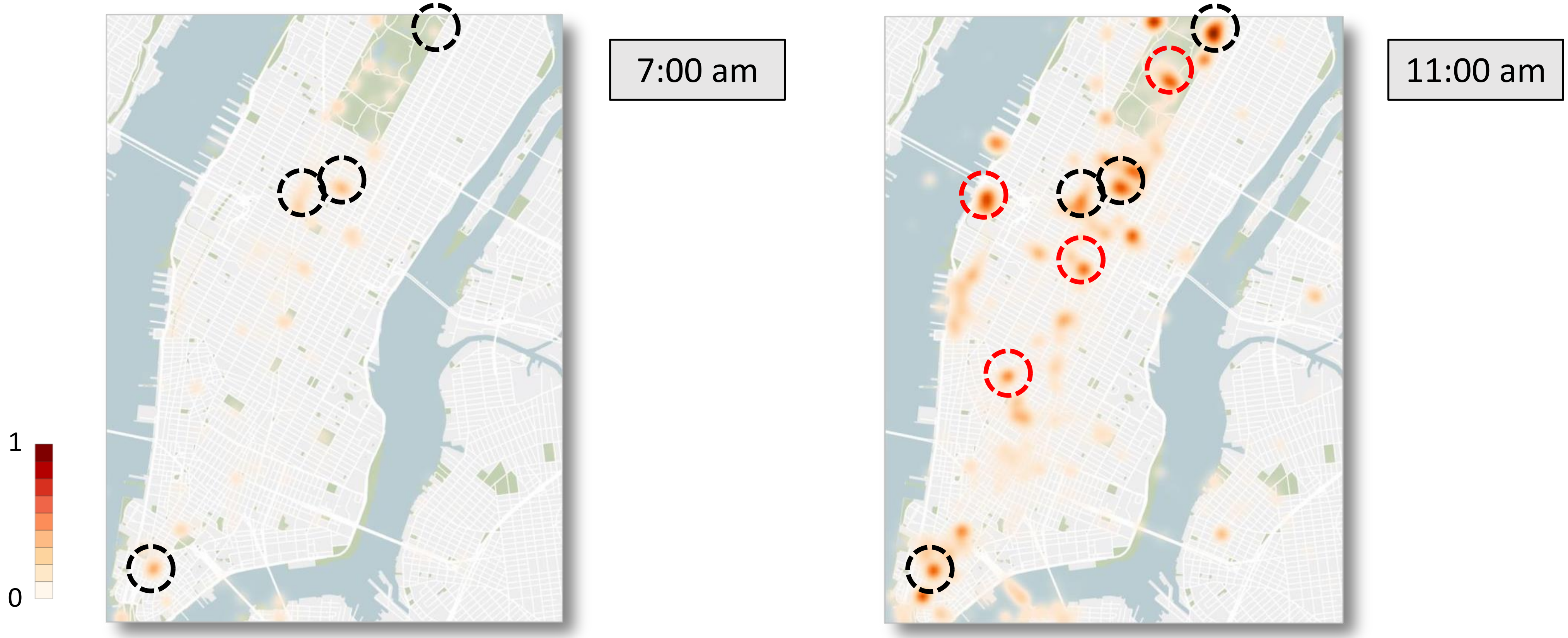
- Flickr activity in New York City



7:00 am

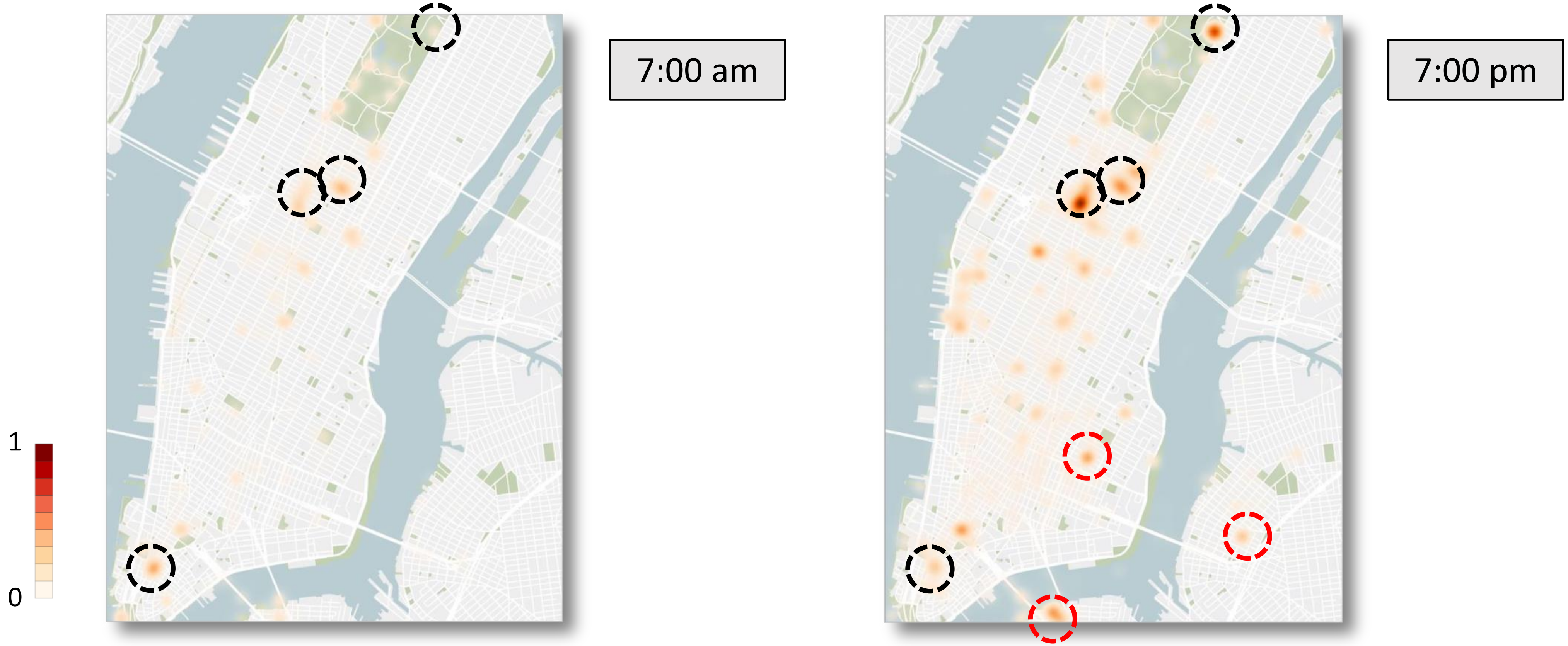
Urban Pulse

- Flickr activity in New York City



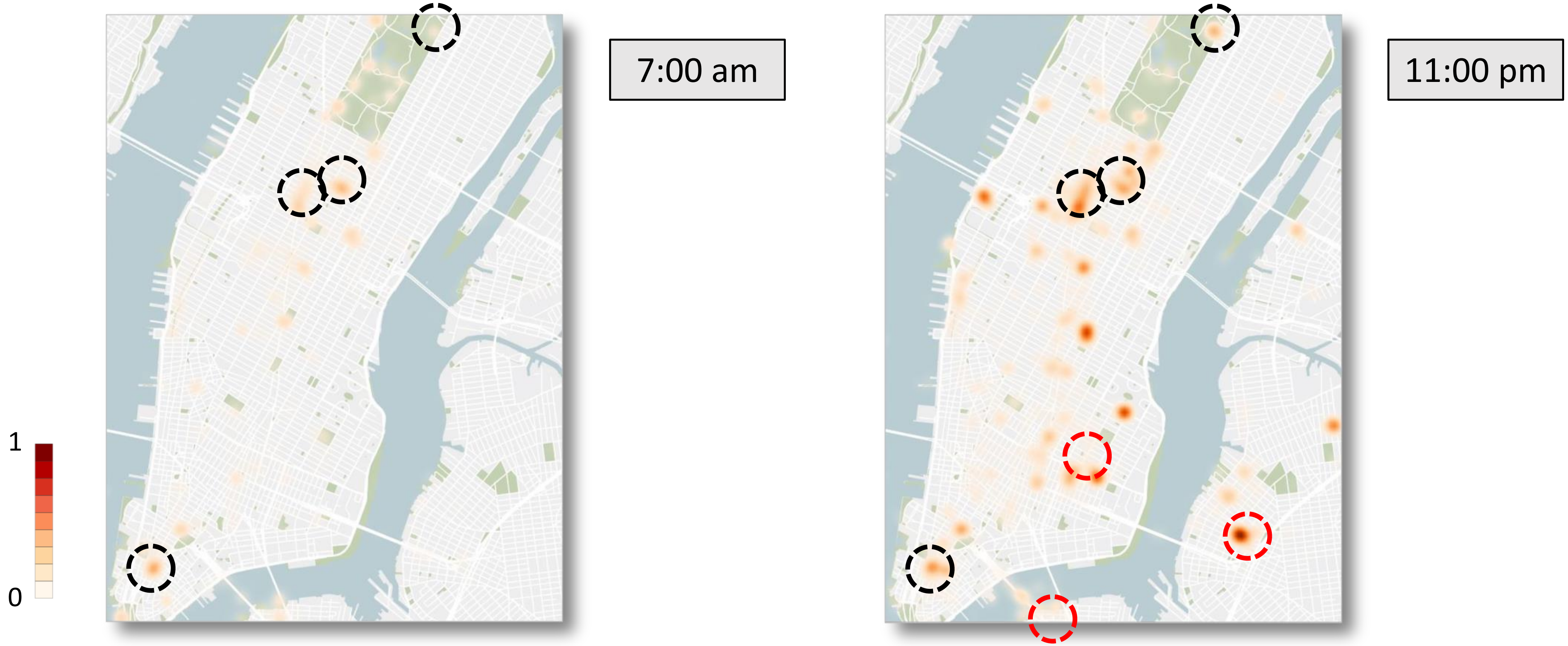
Urban Pulse

- Flickr activity in New York City



Urban Pulse

- Flickr activity in New York City



Urban Pulse: Desiderata

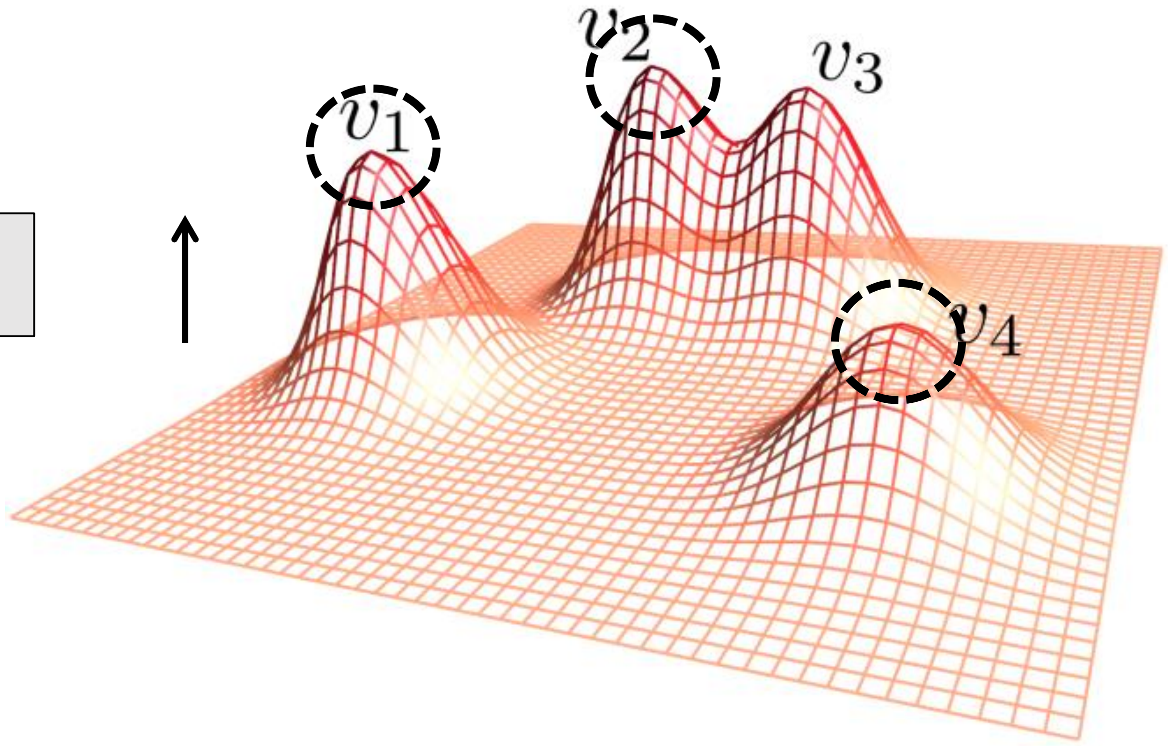
- Capture locations where the pulse is “interesting”
- Quantify the pulse
  - Track “activity”
- Temporal resolutions

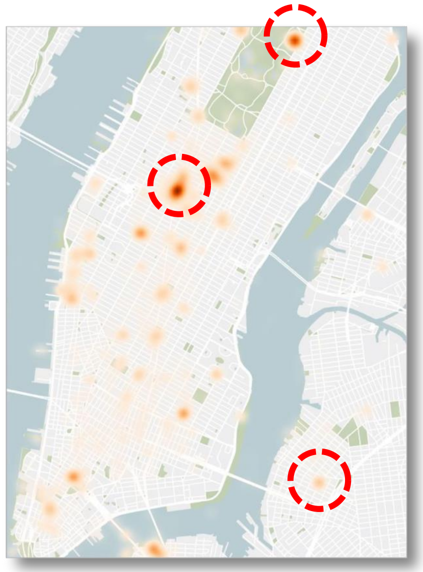
1. Identify Locations

2. Quantify Pulse

Step 1: Identify Pulse Locations

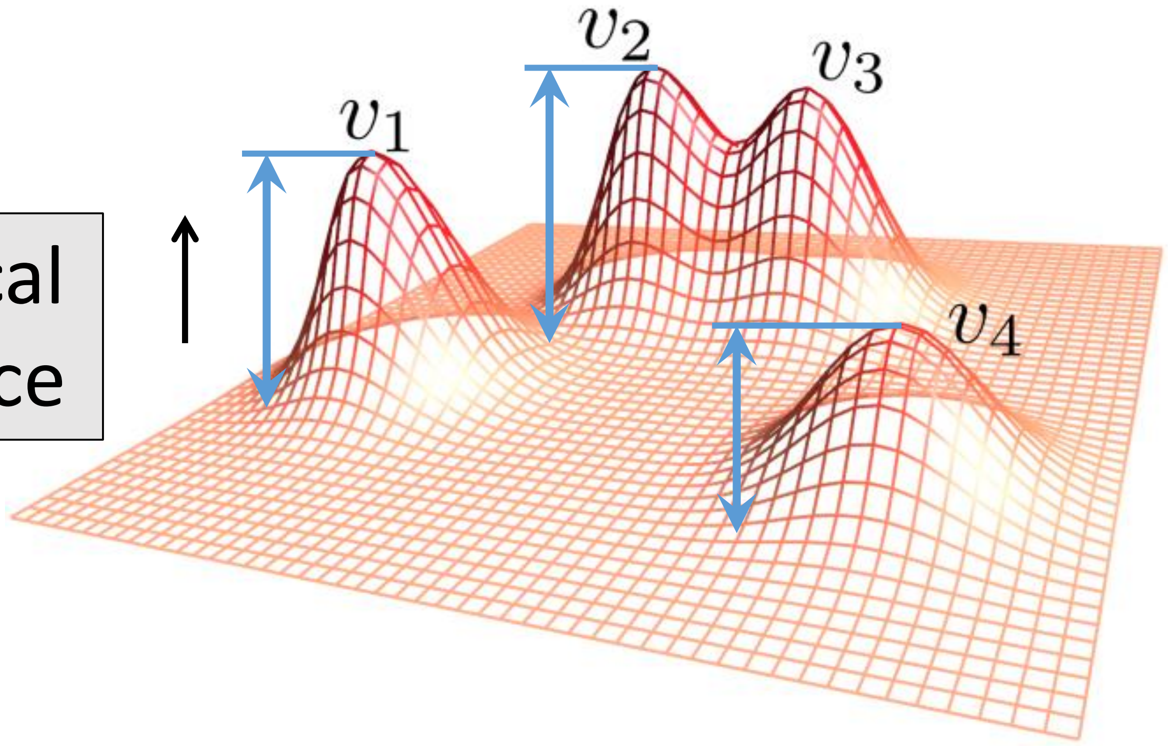
Maxima ↑

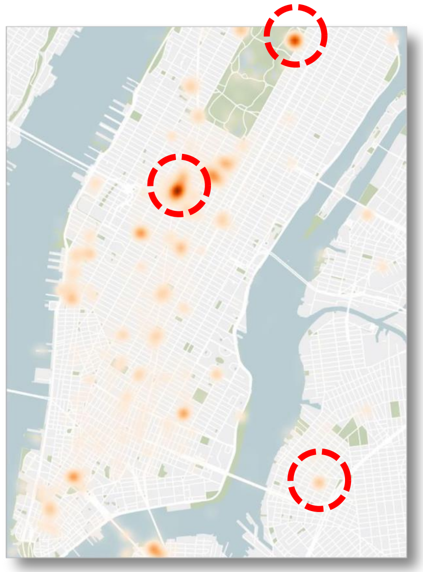




Step 1: Identify Pulse Locations

Topological Persistence



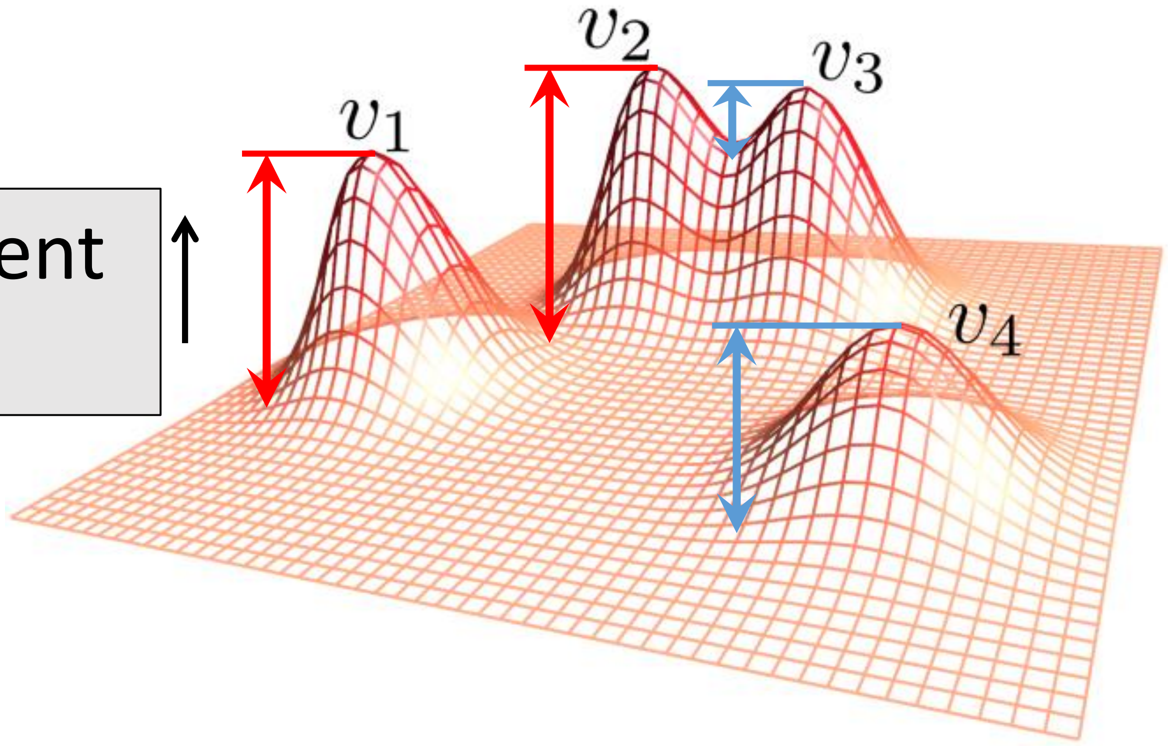


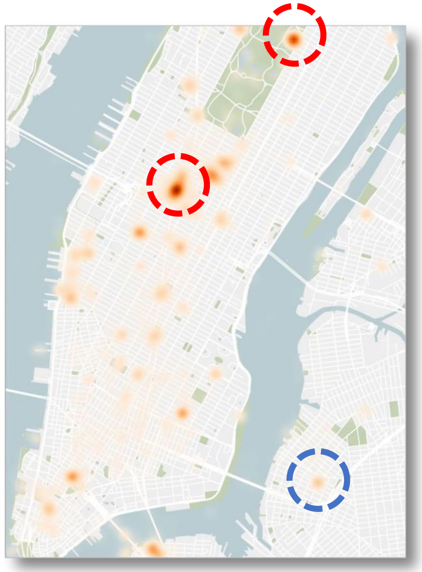
1. Identify Locations

2. Quantify Pulse

Step 1: Identify Pulse Locations

High Persistent Maxima ↑



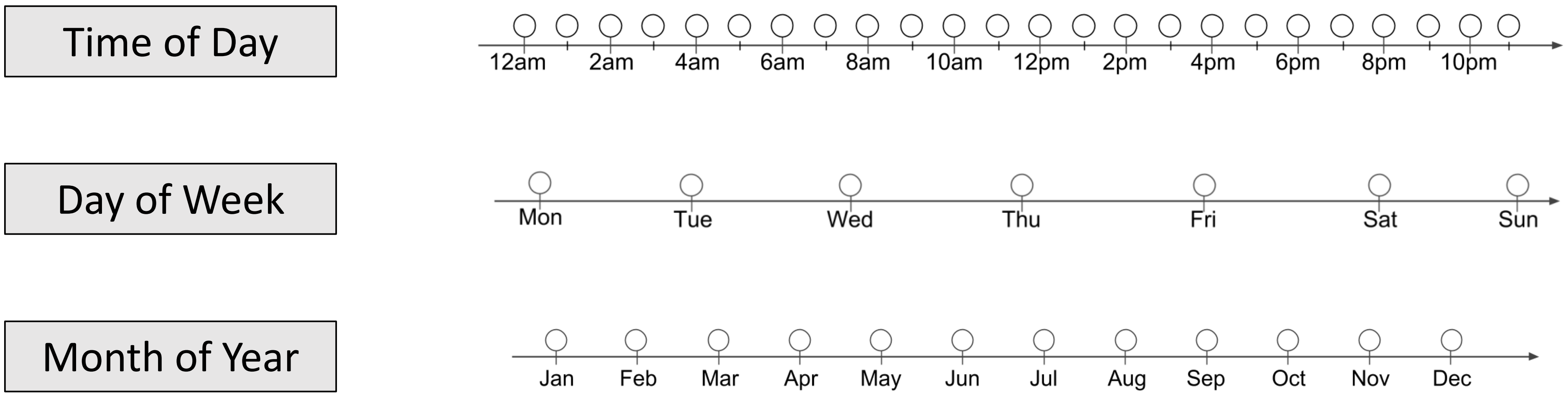


Step 1: Identify Pulse Locations

- Set of scalar functions over time
    - Density functions

Handling Temporal Resolutions

- Assume functions are defined along 3 resolutions



Group By

Step 1: Identify Pulse Locations

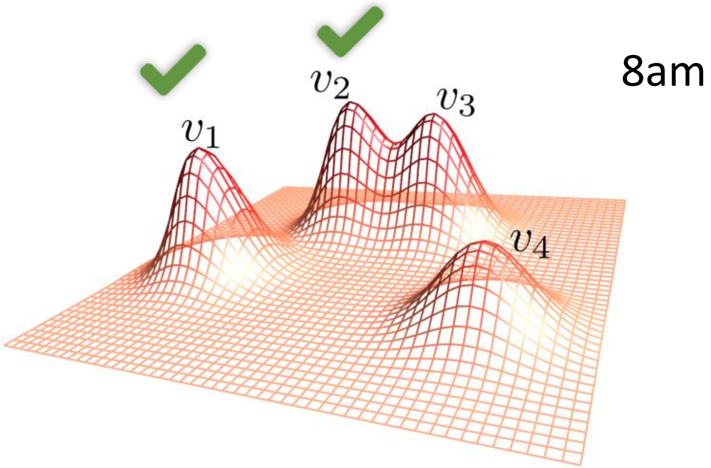
- Set of scalar functions over time
  - Density functions
- Identify all maxima
- Location of **prominent** pulses
  - is a high persistent maxima in at least 1 time step

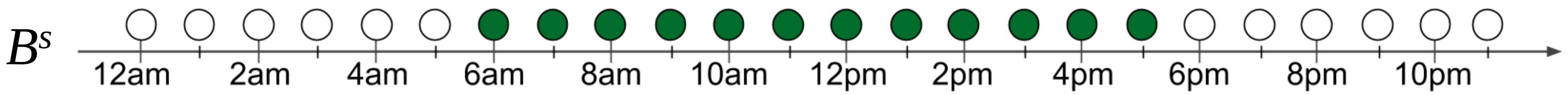
Step 1: Identify Pulse Locations

- Set of scalar functions over time
  - Density functions
- Identify all maxima
- Location of **prominent** pulses
  - is a high persistent maxima in at least 1 time step
  - is a high persistent maxima in at least 1 resolution

Step 2: Quantifying Pulse

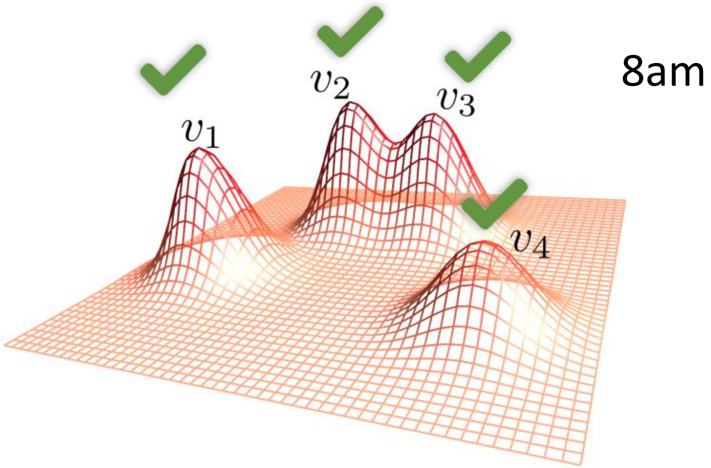
- 3 **Beats** to quantify the pulse at each location
- Significant Beats
  - Is the location a high persistent maximum?

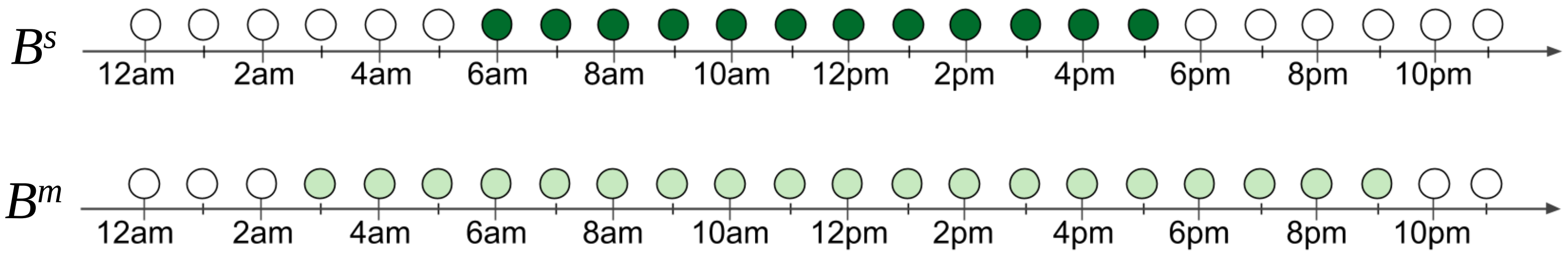




Step 2: Quantifying Pulse

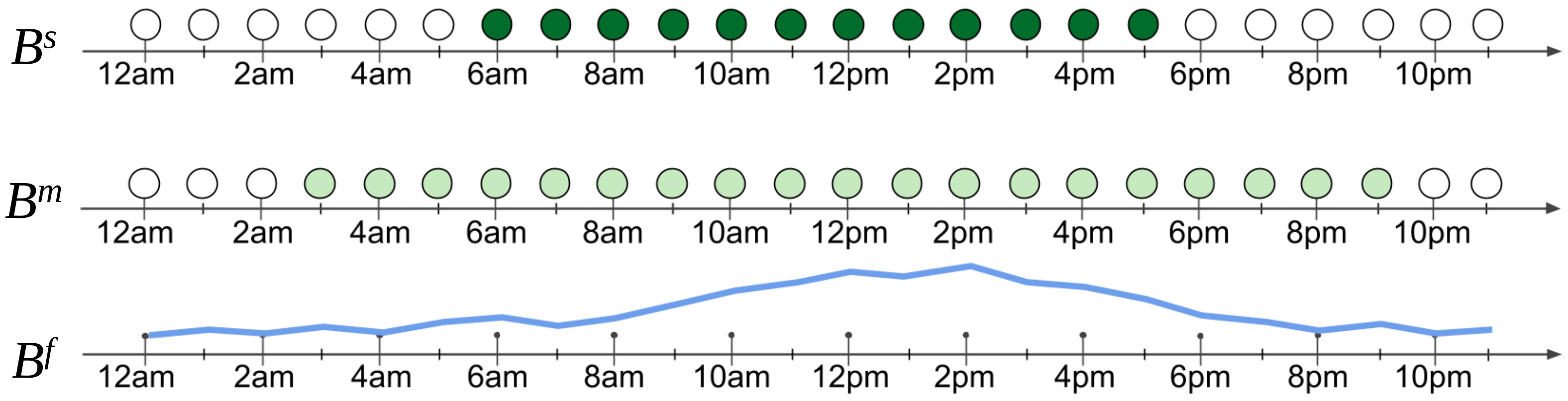
- 3 **Beats** to quantify the pulse at each location
- Maxima Beats
  - Is the location a maximum?





Step 2: Quantifying Pulse

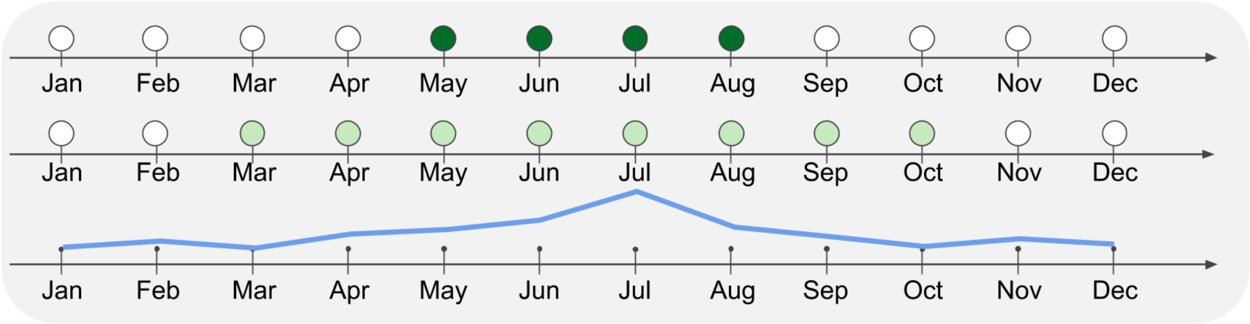
- 3 **Beats** to quantify the pulse at each location
- Function Beats B^f
  - Variation of the function values



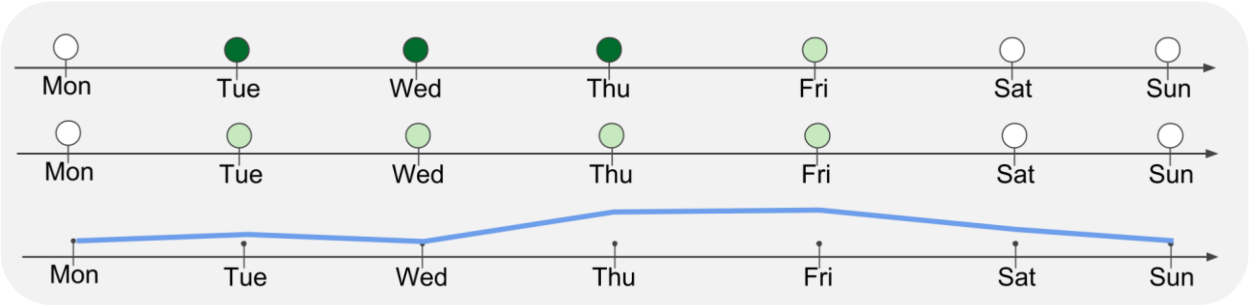
1. Identify Locations

2. Quantify Pulse

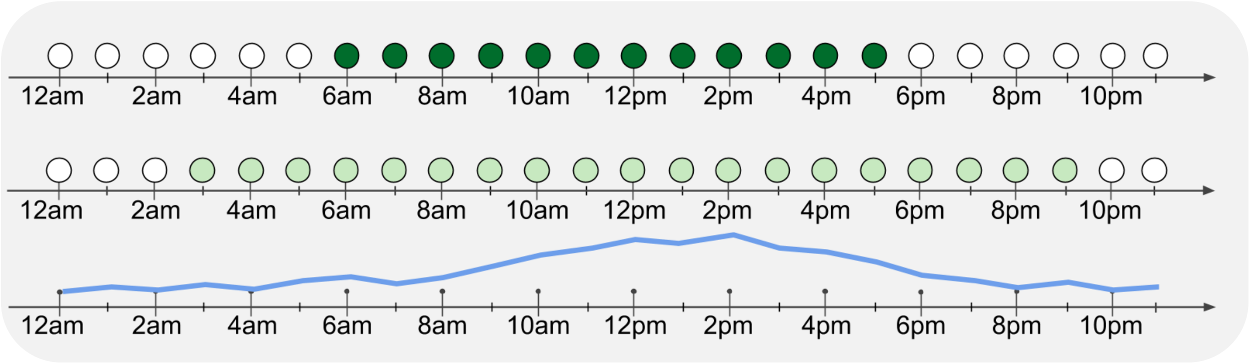
Step 2: Quantifying Pulse



Month of Year



Day of Week



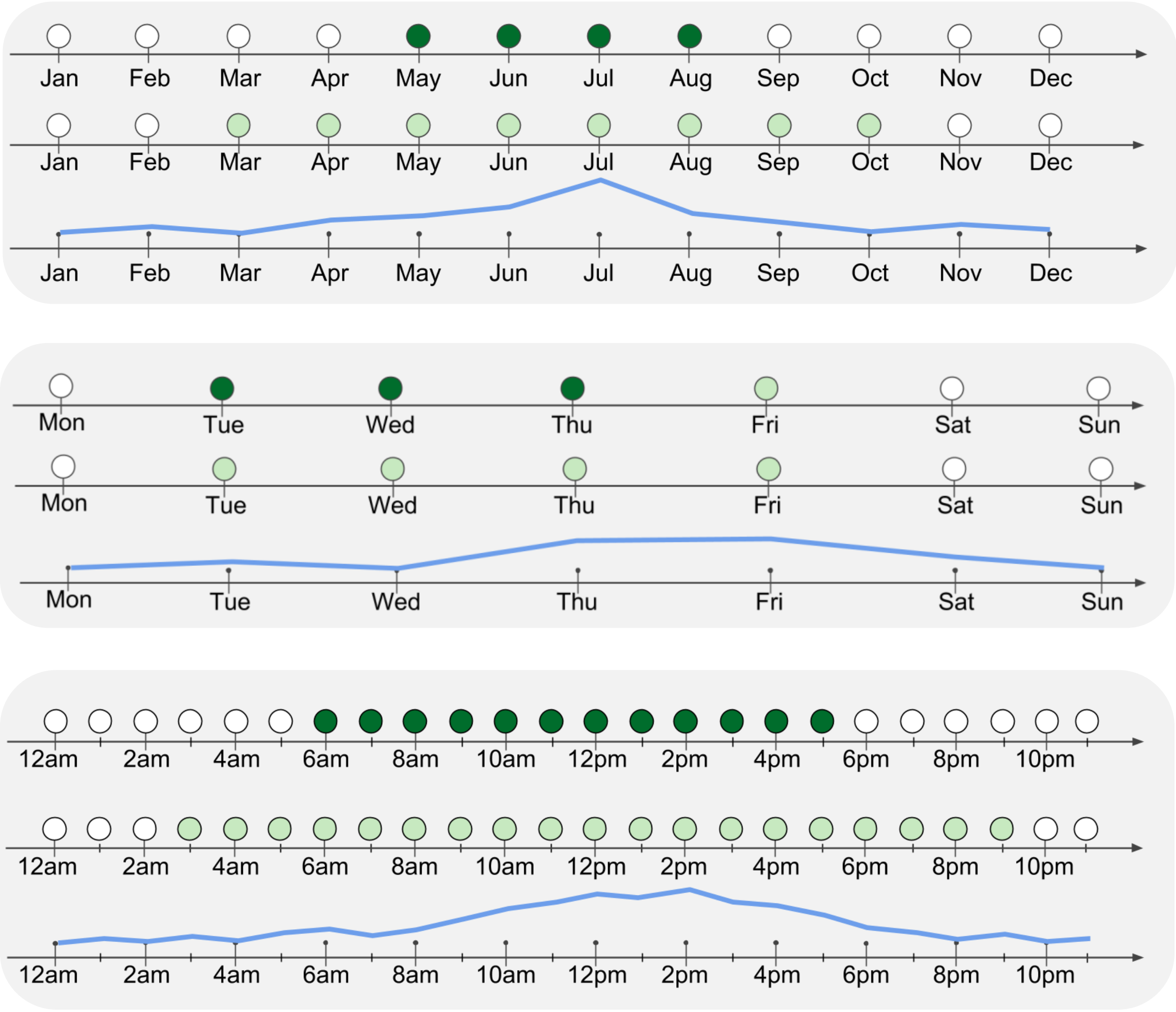
Time of Day

1. Identify Locations

2. Quantify Pulse

Step 2: Quantifying Pulse

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \\ B_7 \\ B_8 \\ B_9 \end{pmatrix}$$



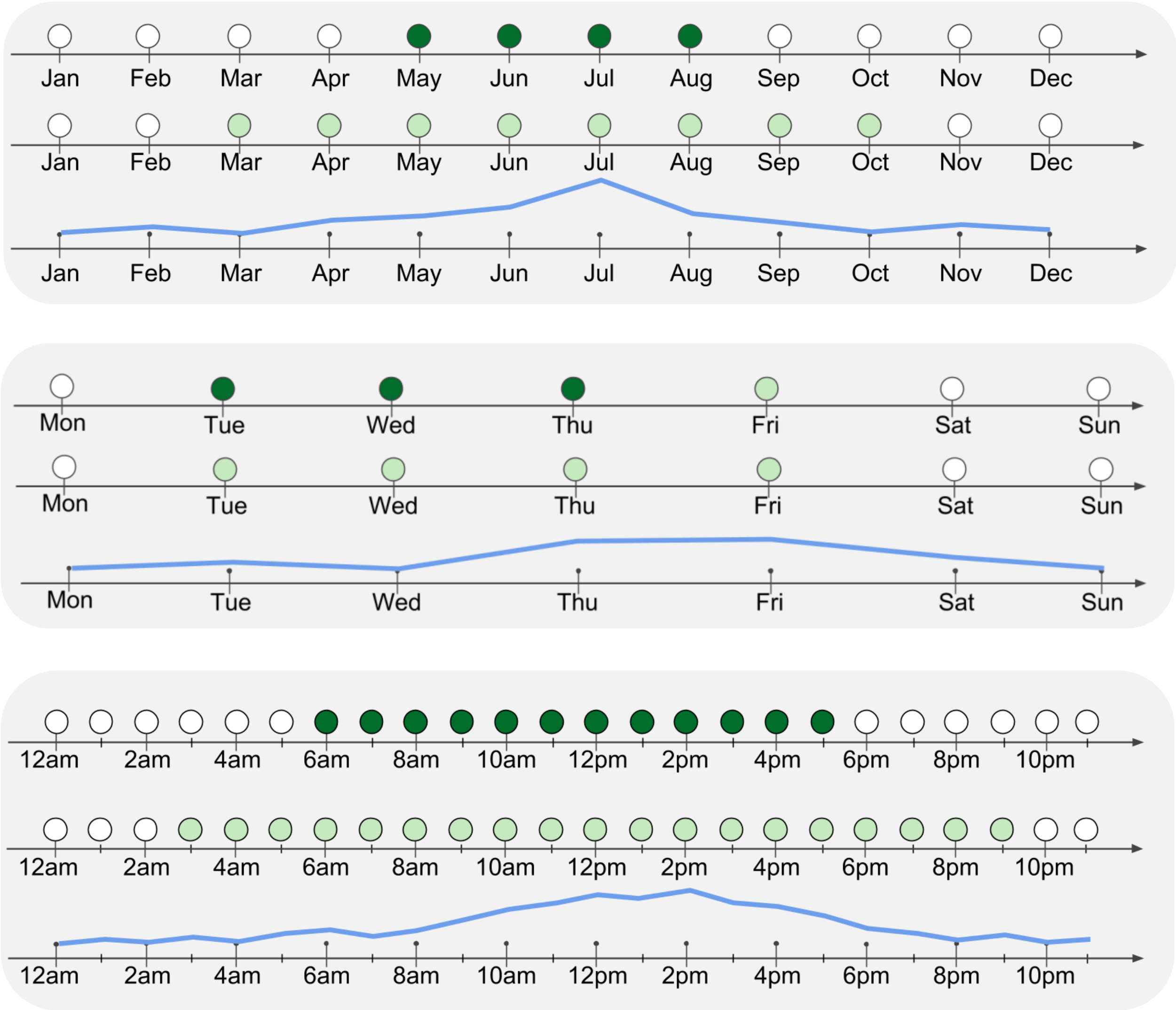
Signature

Data Oblivious

Rank

Step 2: Quantifying Pulse

$$\begin{pmatrix} B_1 \\ B_2 \\ B_3 \\ B_4 \\ B_5 \\ B_6 \\ B_7 \\ B_8 \\ B_9 \end{pmatrix}$$

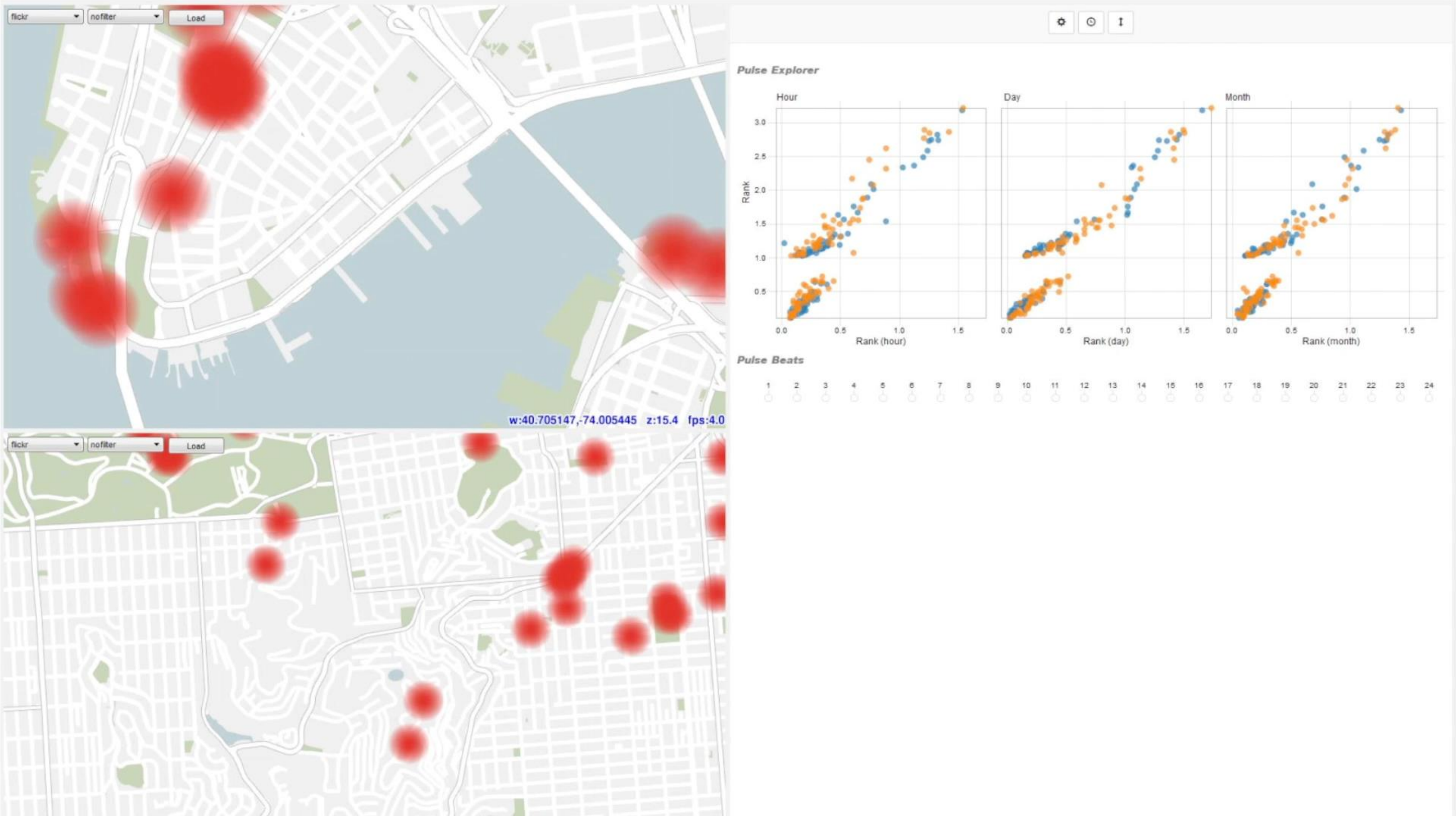


Signature

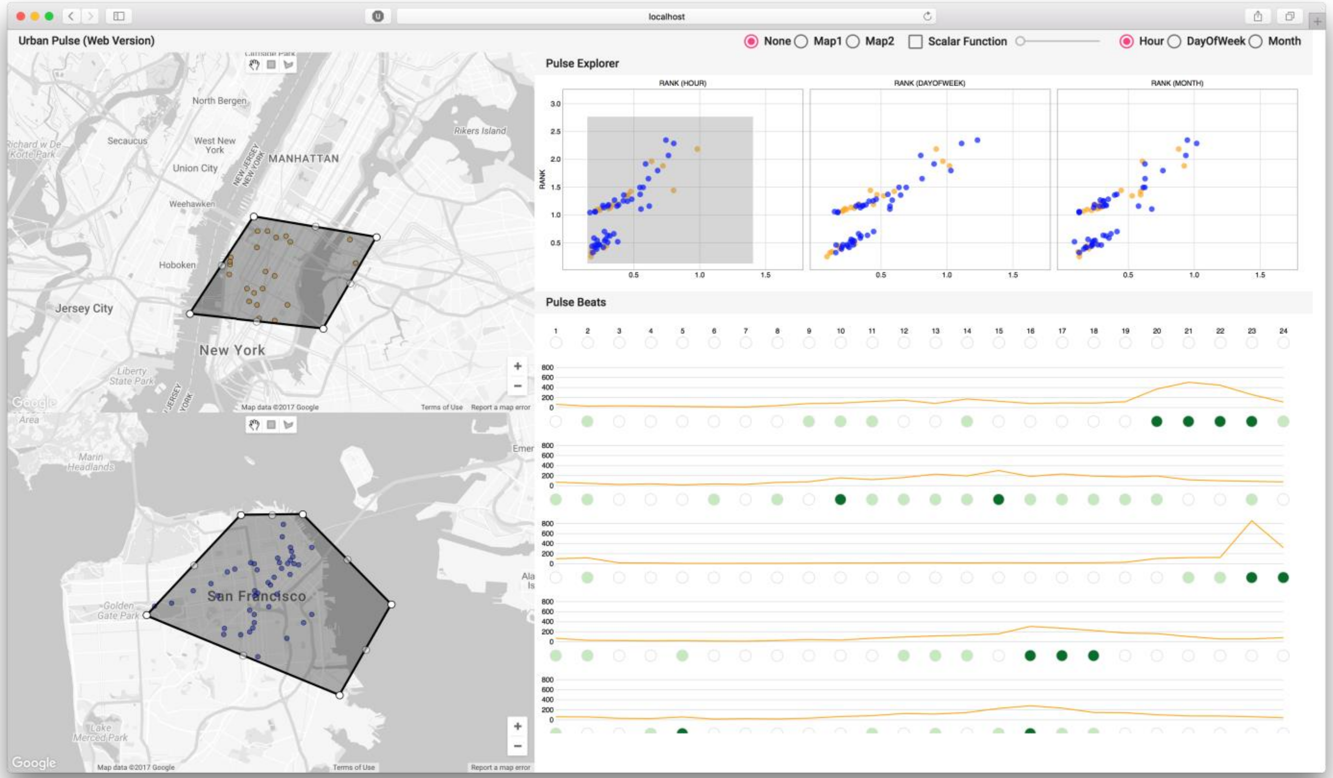
Data Oblivious

Compare

Urban Pulse Interface



Open Source (BSD License)



Code: https://github.com/ViDA-NYU/urban-pulse/

Demo: http://vgc.poly.edu/projects/urban-pulse/

Use Case

- Provided the interface to domain experts
- Architects from Kohn Pedersen Fox
  - Urban planning
- Human behavioral expert
  - Try to understand the cohabitation between cultural communities
  - Twitter as proxy for cultural communities

Use Case: Understanding Public Spaces

Rockefeller Center

Union Square

Bryant Park

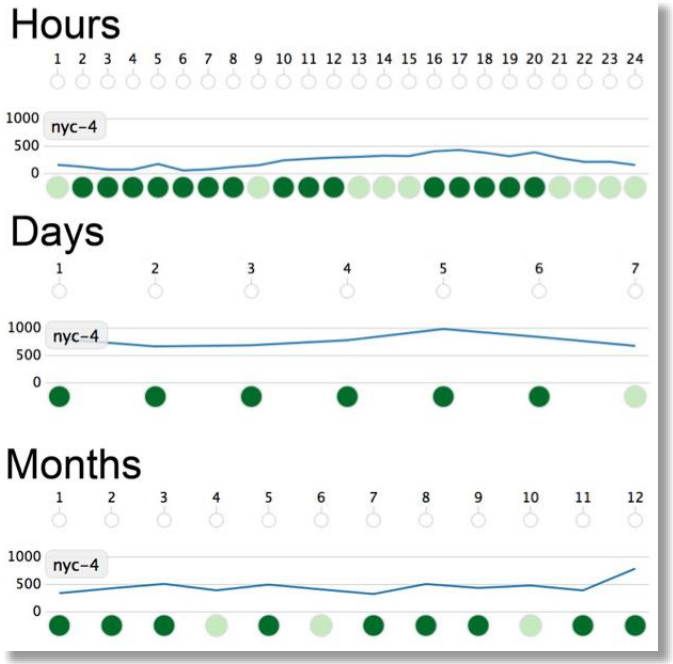
- Typically classified together as being similar

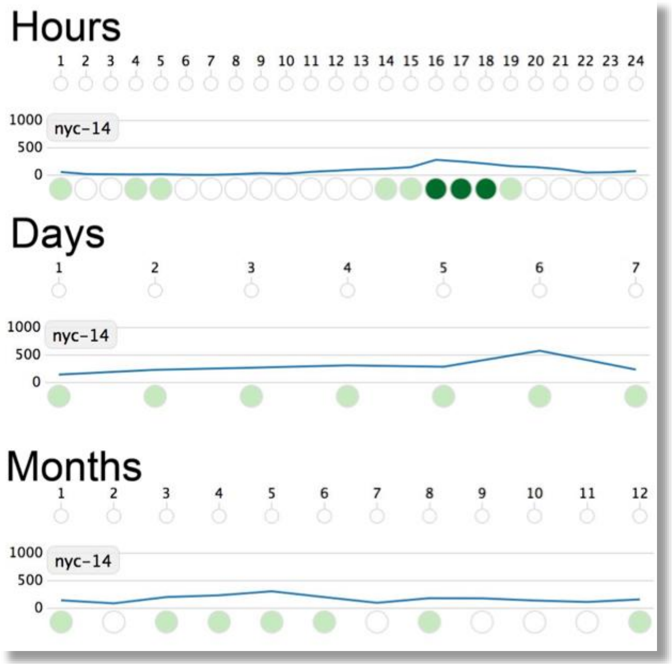
Use Case: Understanding Public Spaces

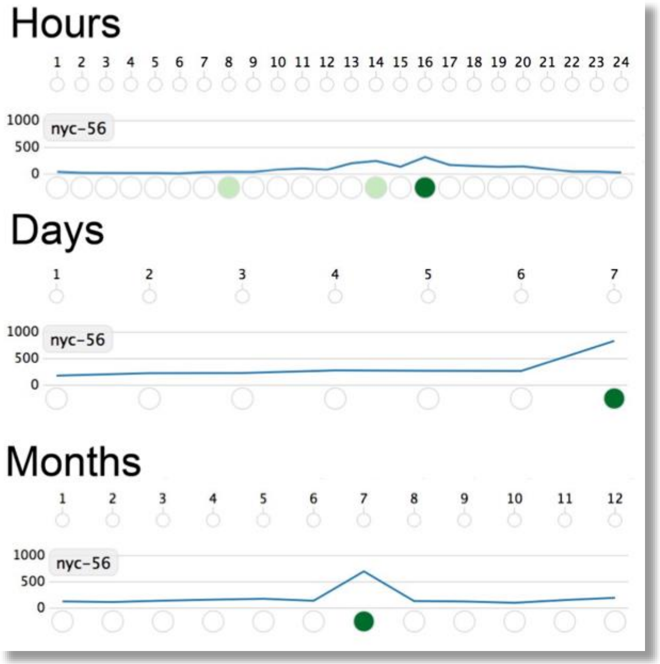
Rockefeller Center

Union Square

Bryant Park



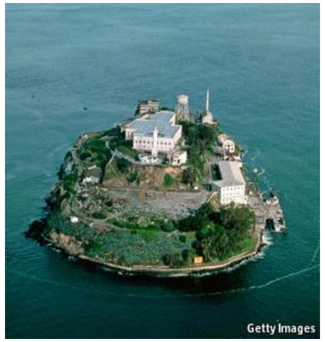


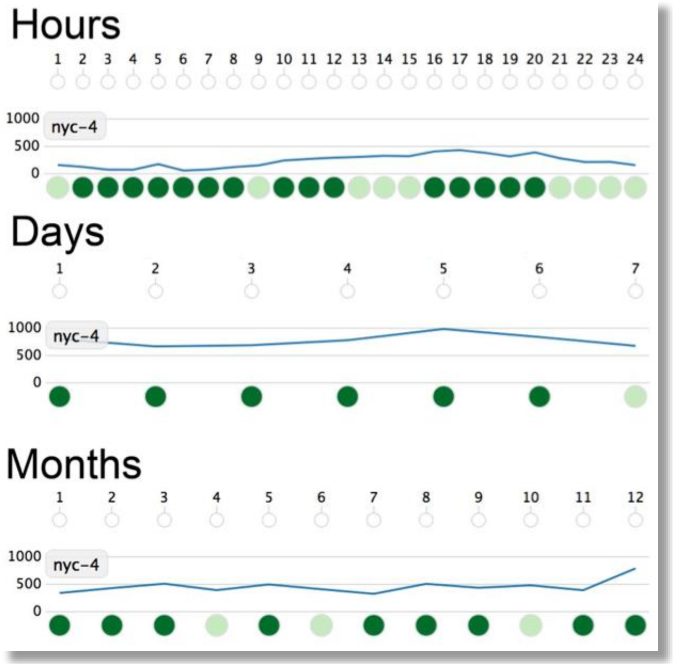


Use Case: Understanding Public Spaces

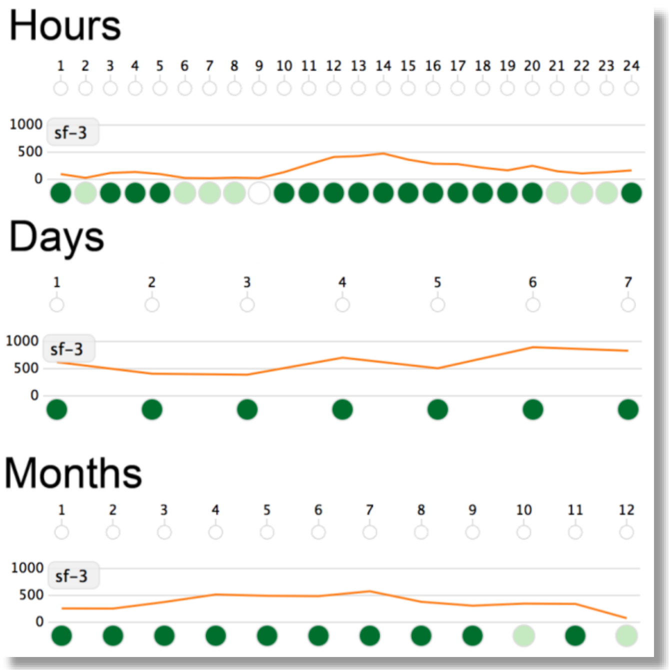
Rockefeller Center

Alcatraz



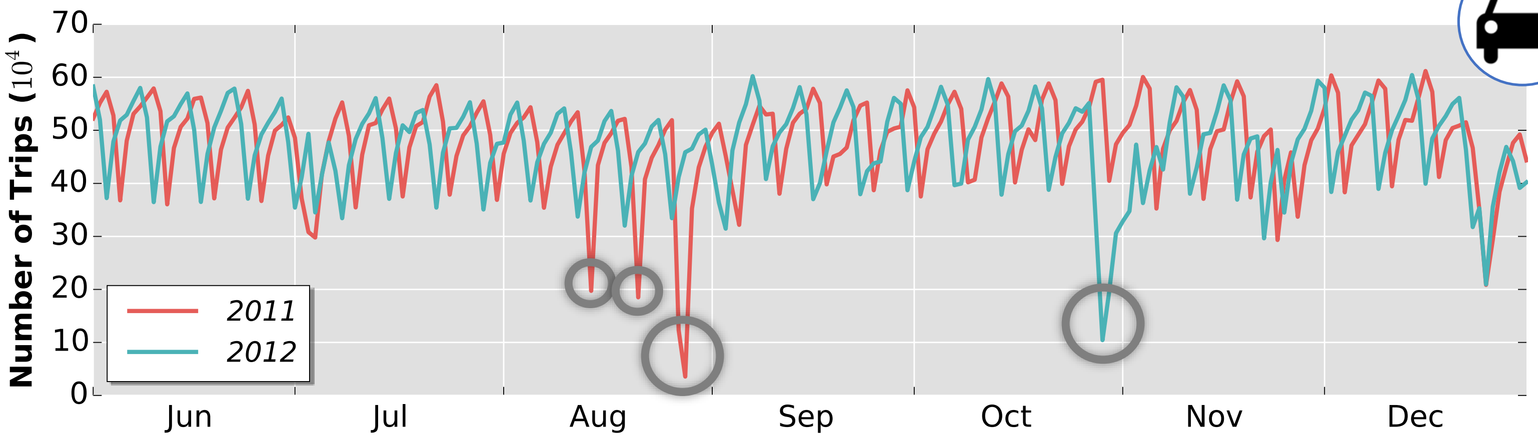


San Francisco



How to understand features?

1. Why the number of taxi trips is too low? Is this a data quality problem?



How to understand features?

1. Why the number of taxi trips is too low? Is this a data quality problem?
2. Why it is so hard to find a taxi when it is raining?

DAILY Intelligencer

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

By **Annie Lowrey** Follow @AnnieLowrey

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

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

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

How to understand features?

1. Why the number of taxi trips is too low? Is this a data quality problem?
2. Why it is so hard to find a taxi when it is raining?
3. Would a reduction in traffic speed reduce the number of accidents?

Urban Data Interactions

Uncovering **relationships** between data sets helps us better understand cities!

*Urban Data Sets are very **Polygamous**!*

Data is available…

… but it’s too much work!
Big urban data!

NYC OpenData

1,200 data sets (and counting)	> 300 data sets are **spatio-temporal**	8 attributes per data set

> 200 attributes

Where to start?
Which data sets to analyze?

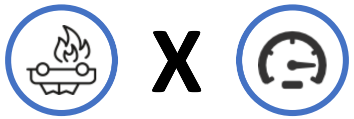
Data Polygamy Framework

Goal: Relationship Queries

*Find all data sets **related** to a given data set D*

Guide users in the data exploration process

Help identify connections amongst disparate data

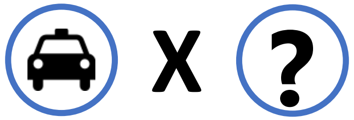


Q: Would a reduction in traffic speed reduce the number of accidents?

Find all relationships between Collisions and Traffic Speed data sets

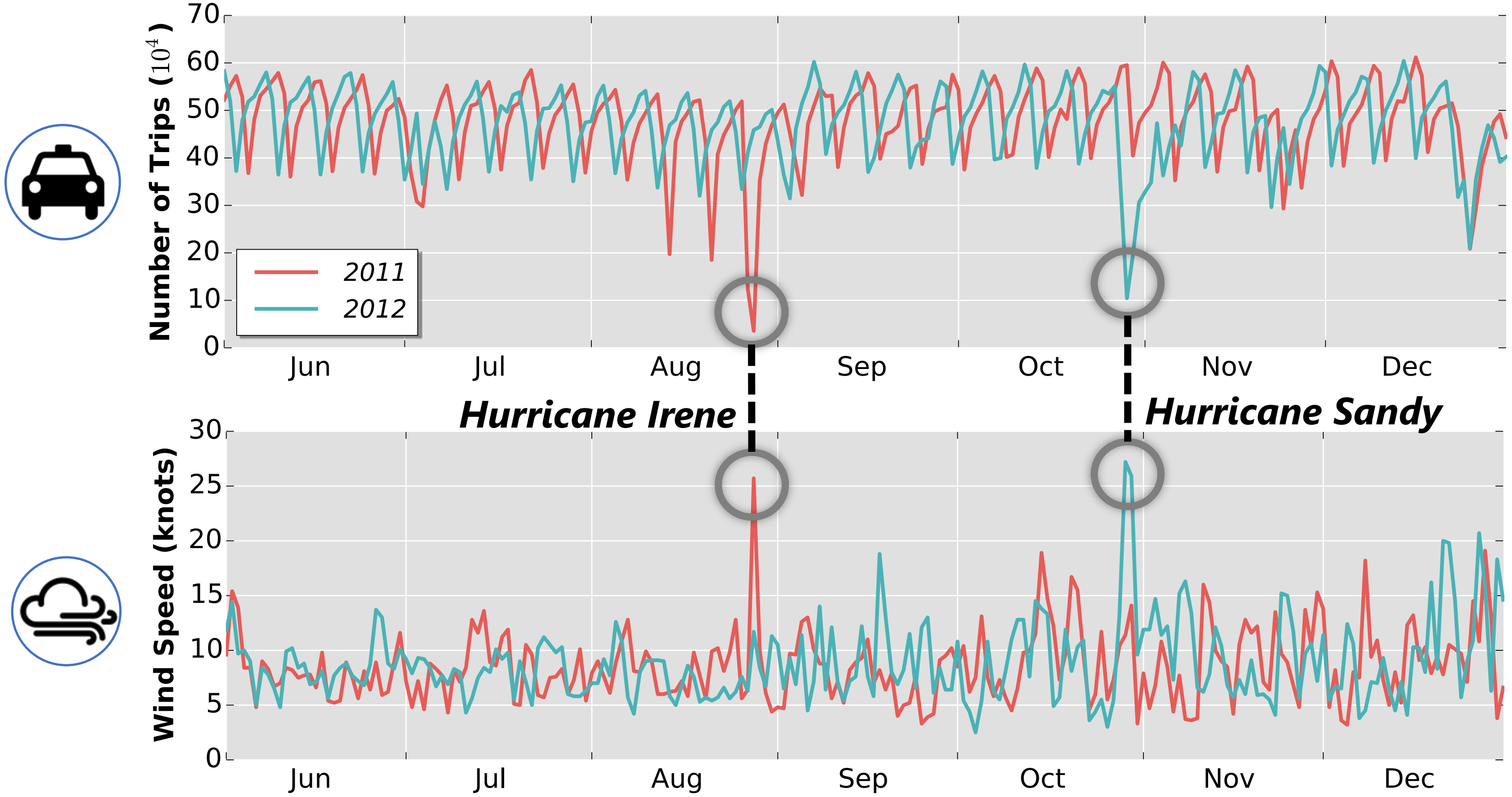
Q: Why the number of taxi trips is too low?

Find all data sets related to the Taxi data set



Challenges

1) How to define a ***relationship*** between data sets?



Challenges

1) How to define a ***relationship*** between data sets?

Relationships between interesting *features* of the data sets

Relationships must take into account both *time* and *space*

Conventional techniques (Pearson's correlation, DTW, mutual information) cannot find these relationships!

Challenges

2) Large data complexity: **Big** urban data

Many, many data sets !
Data at multiple spatio-temporal resolutions

Relationships can be between any of the attributes

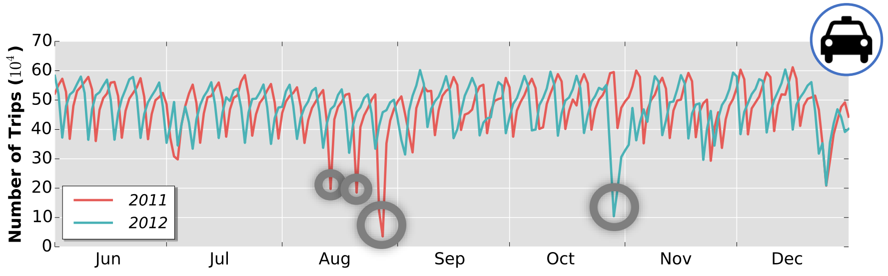
Many attributes!
≈**2.4 million** possible relationships among NYC Open Data alone for a **single spatio-temporal resolution**

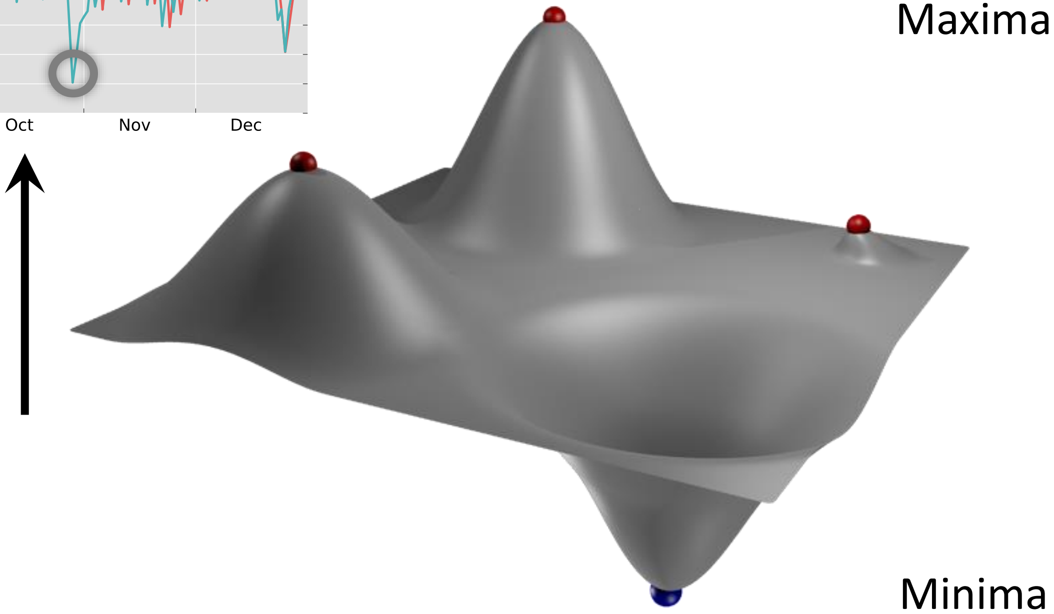
meaningful relationships ⟷ *needle in a haystack*

Key Idea:
Topology-based Relationships

Identifying Topological Features

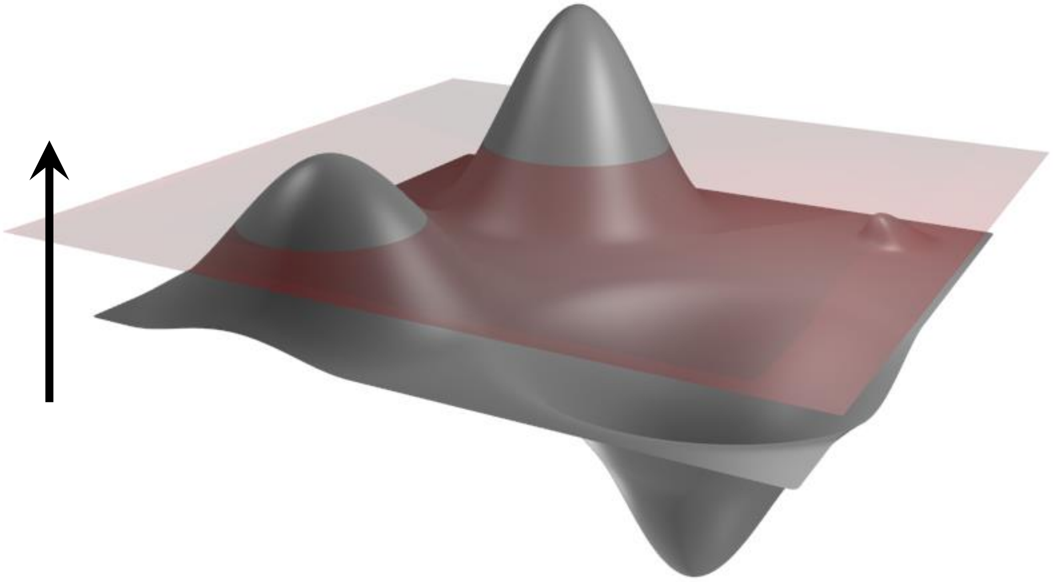
- Topological features of the scalar function
  - Neighborhoods of critical points





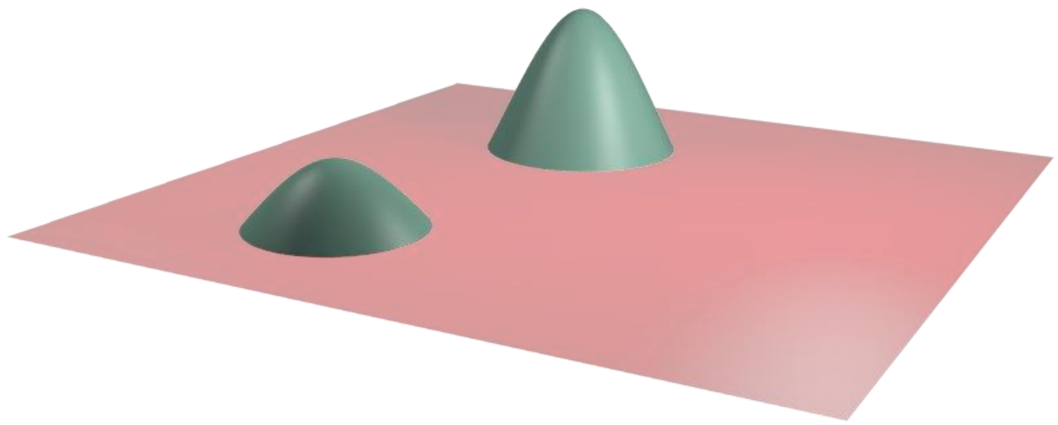
Identifying Topological Features

- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold



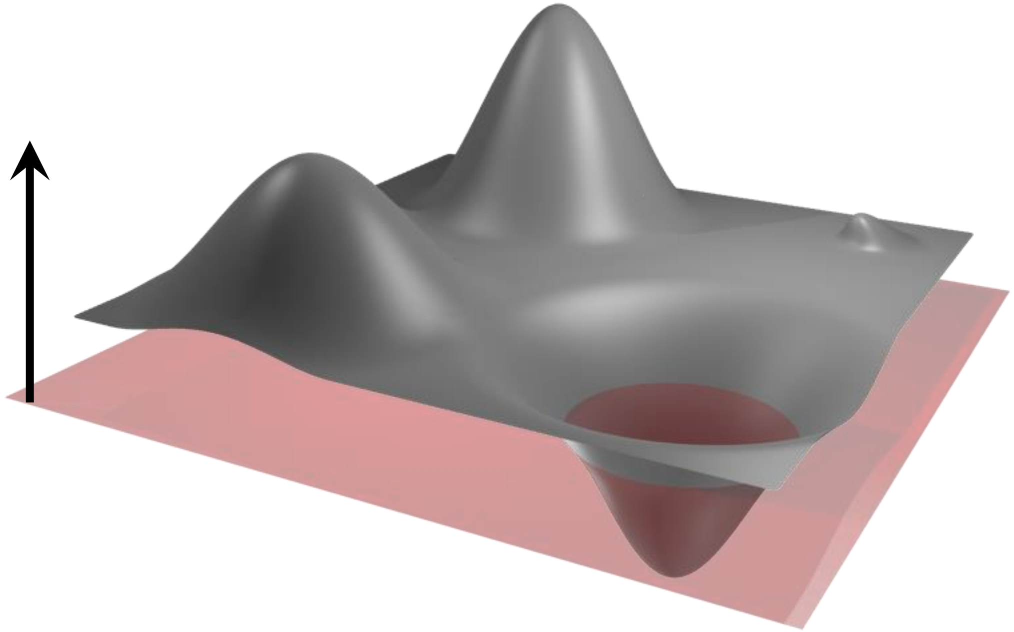
Identifying Topological Features

- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold
  - Positive Features



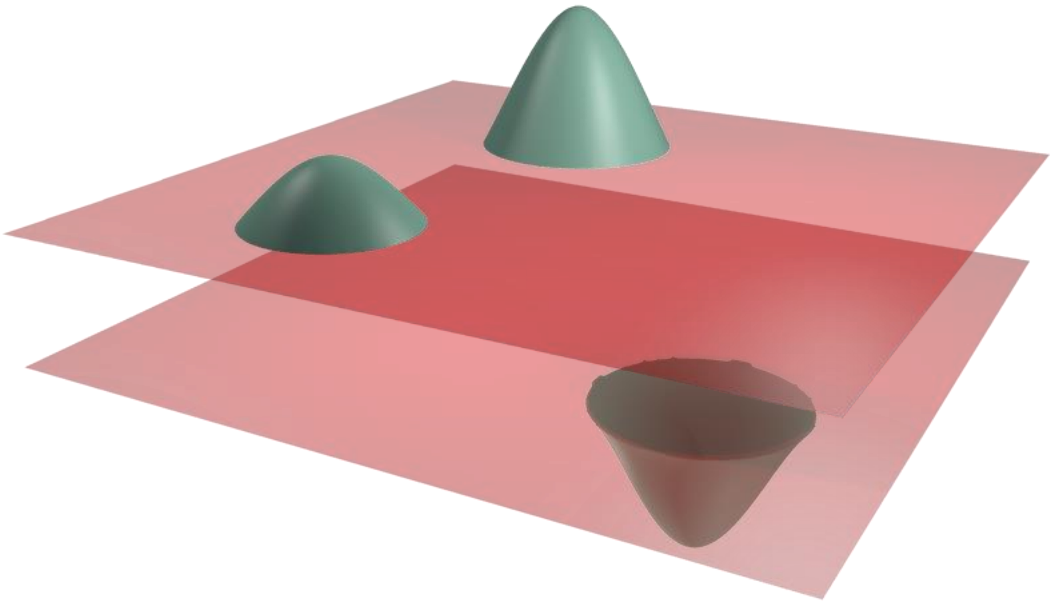
Identifying Topological Features

- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold
  - Positive Features



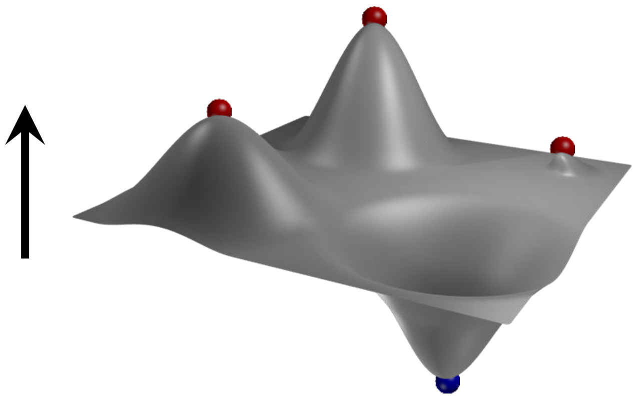
Identifying Topological Features

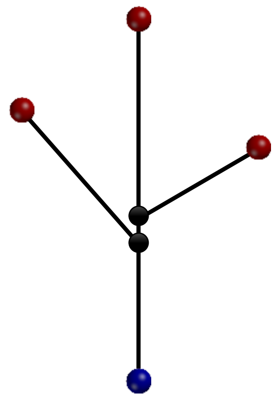
- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold
  - Positive Features
  - Negative Features
- Represented as a set of spatio-temporal points



Computing Topological Features

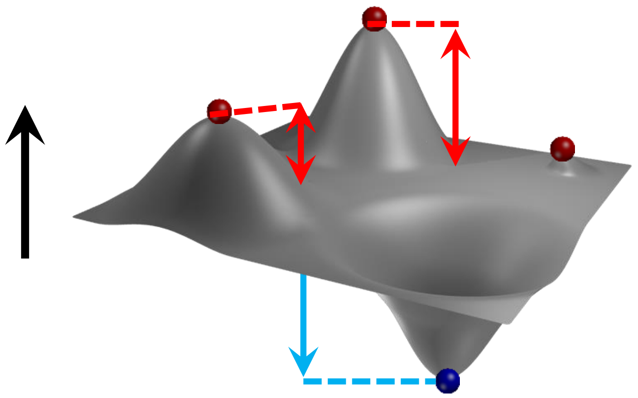
- Index: Merge Tree
  - Topological data structure
  - Tracks evolution of the topology of level sets
  - Data can be of any dimension
- Output-sensitive time complexity

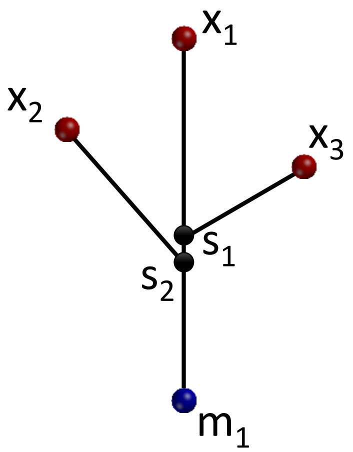




Computing Feature Threshold

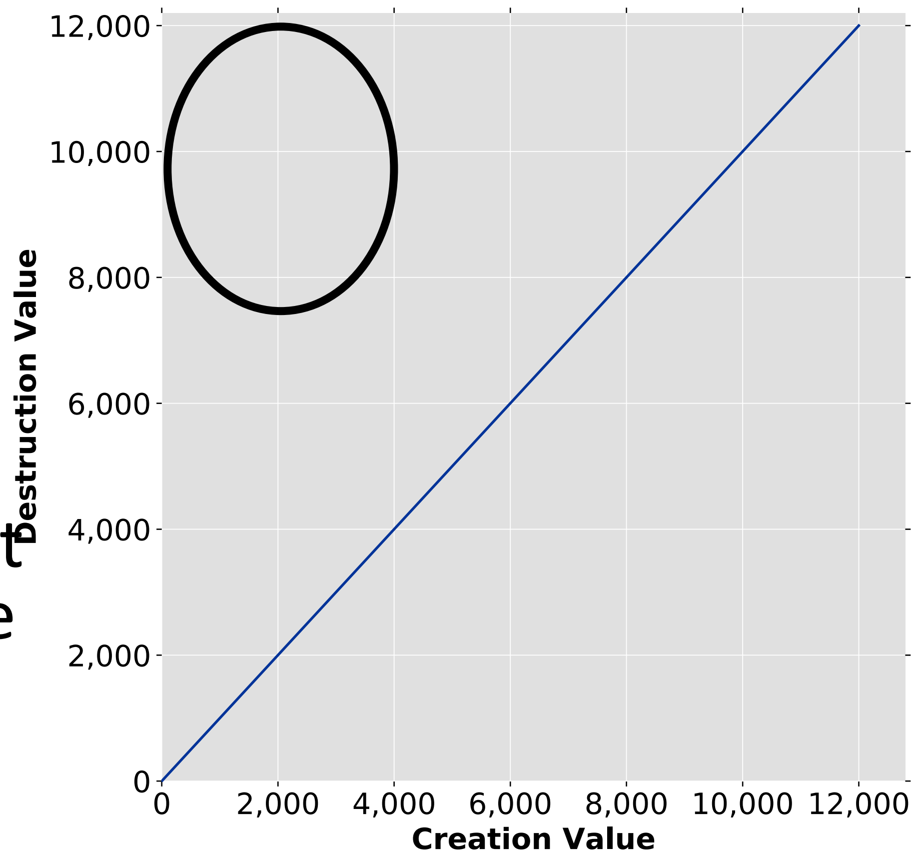
- Feature thresholds are computed in a data-driven approach
  - Uses topological persistence of the features
  - Persistence can be efficiently computed using the merge tree





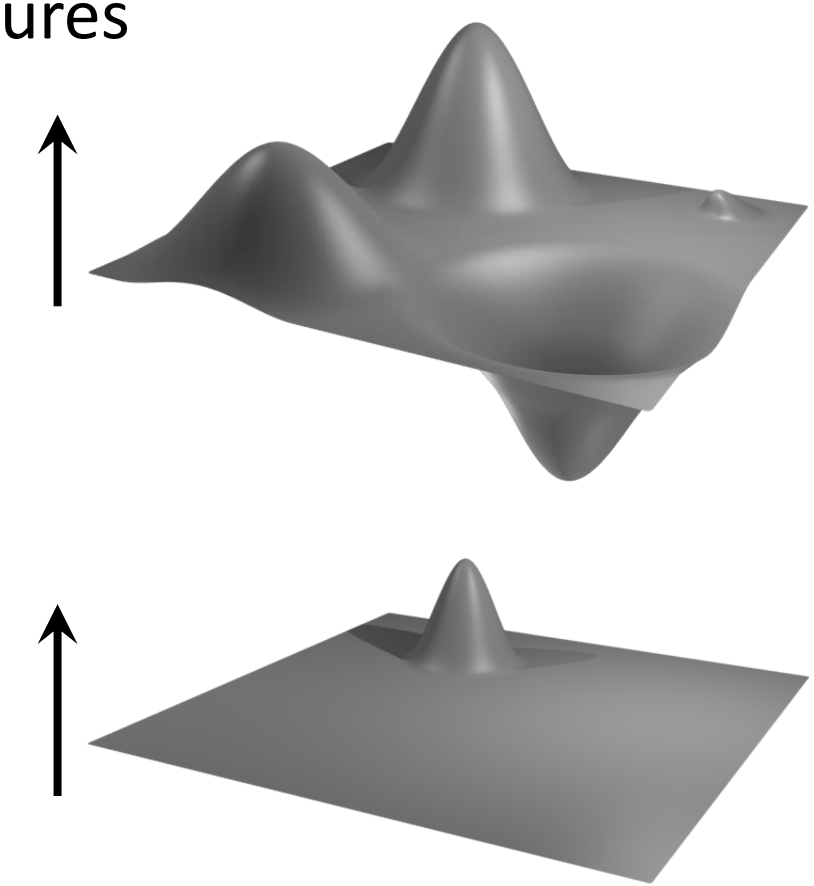
Computing Feature Threshold

- Use persistence diagram
  - Plots “birth” vs “death”
- High persistent features form a separate cluster
- 2-means cluster
- Use the high persistent cluster to compute the threshold



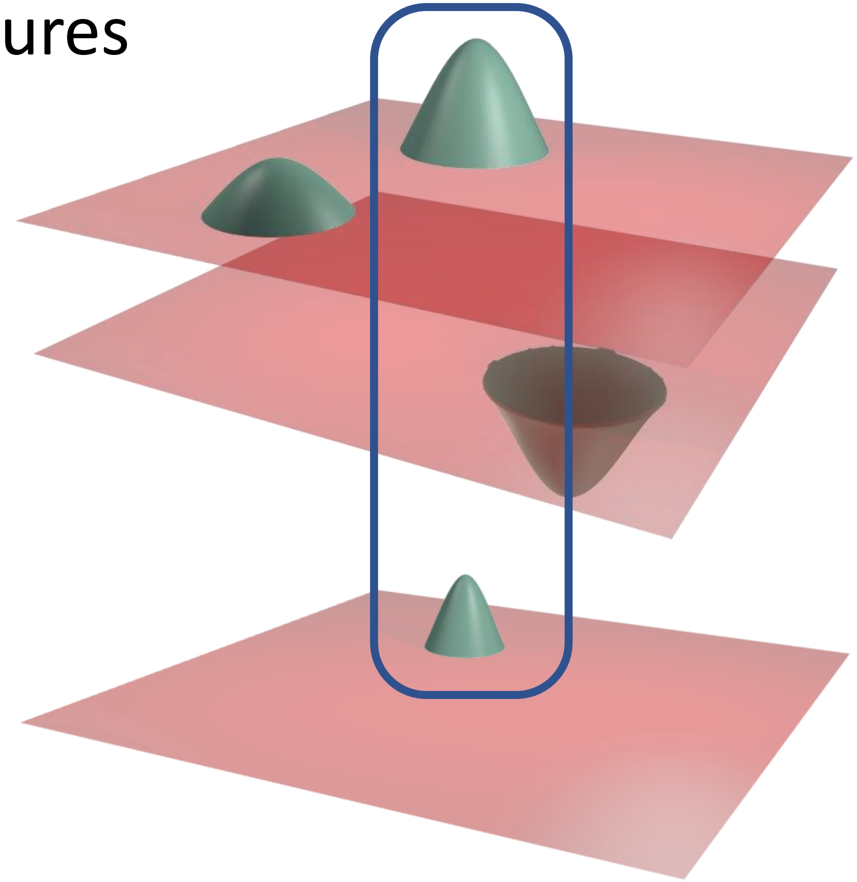
Relationship Evaluation

- Relationship between features



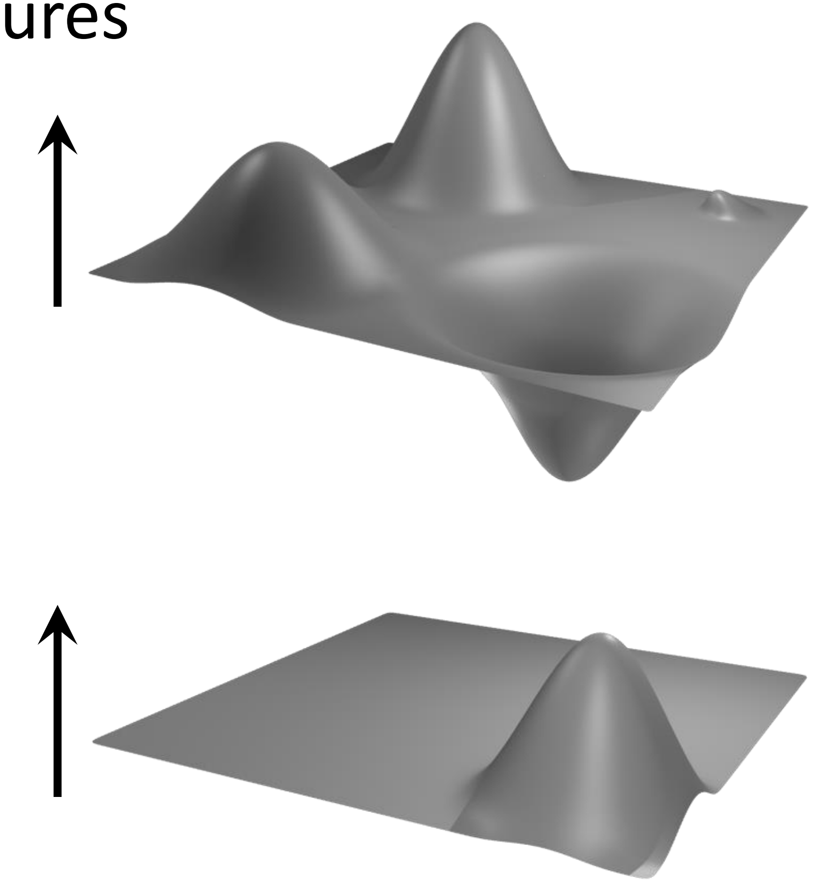
Relationship Evaluation

- Relationship between features
  - Related features
  - **Positive** Relationship



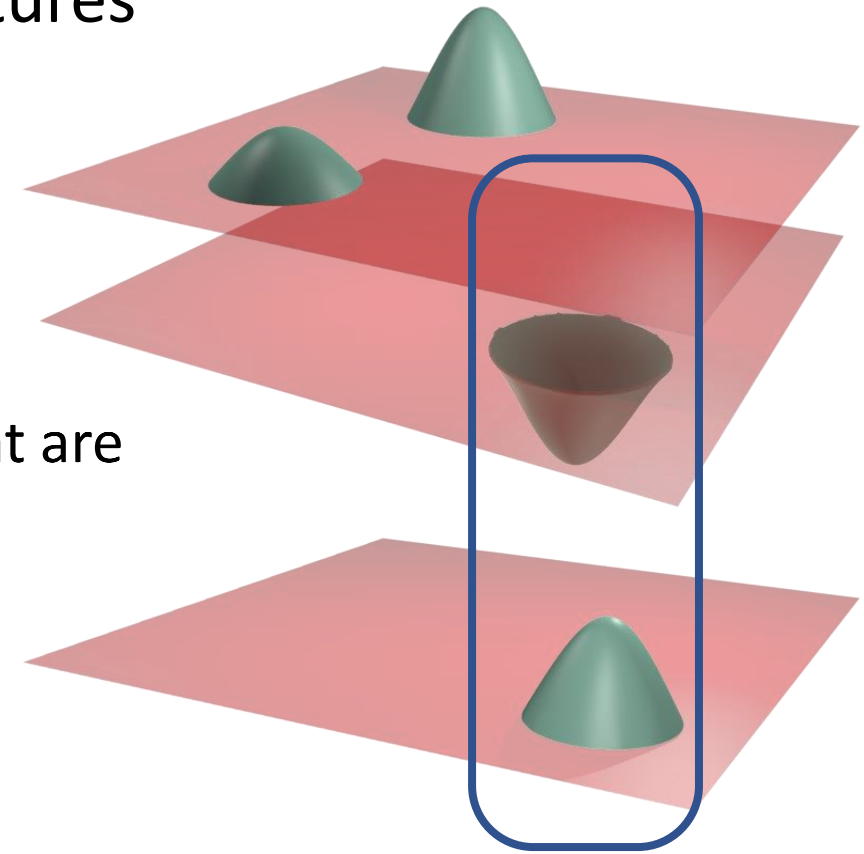
Relationship Evaluation

- Relationship between features
  - Related features
  - **Positive** Relationship



Relationship Evaluation

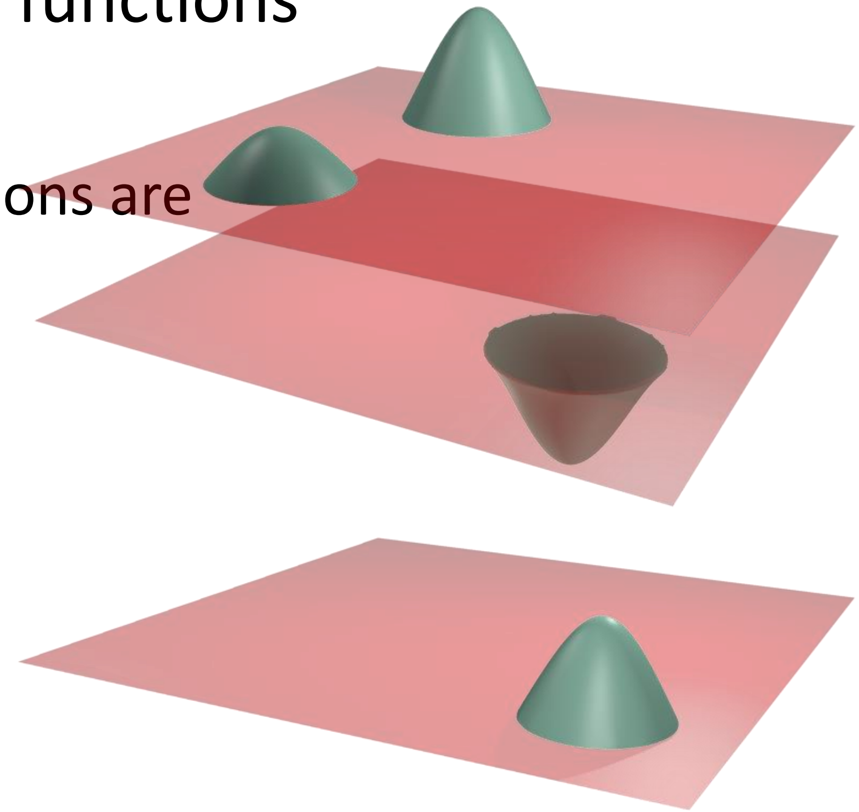
- Relationship between features
  - Related features
  - **Positive** Relationship
  - **Negative** Relationship
- Defined w.r.t. features
  - Spatio-temporal points that are features in both functions



Relationship Evaluation

- Relationship between two functions
- ***Relationship Score*** **(*τ*)**
  - How related the two functions are
  - Captures the nature of the relationship

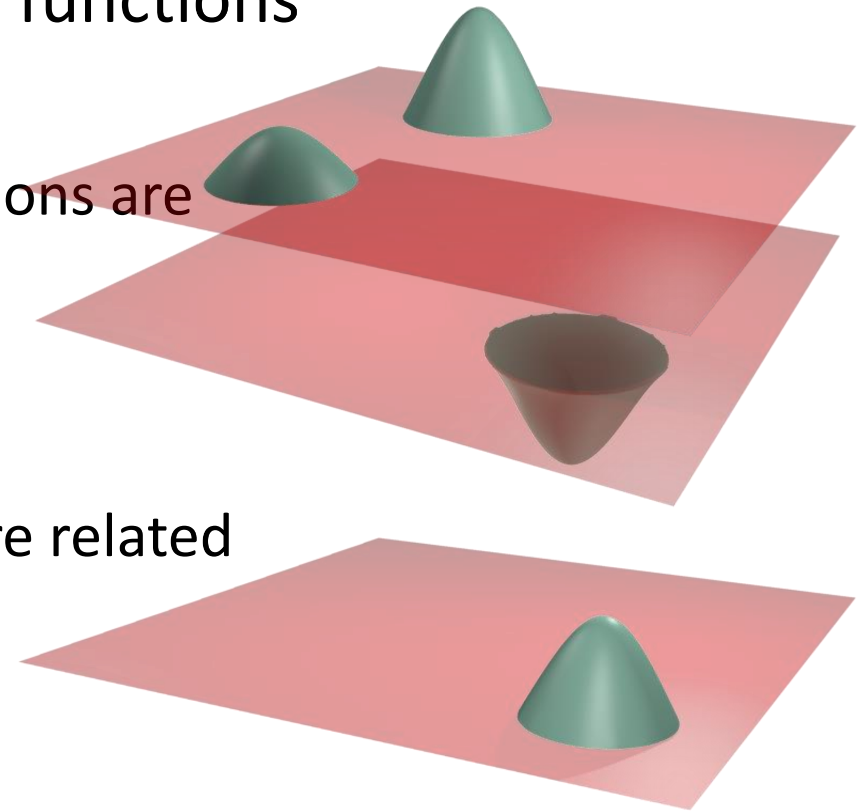
Negative Relationship



Relationship Evaluation

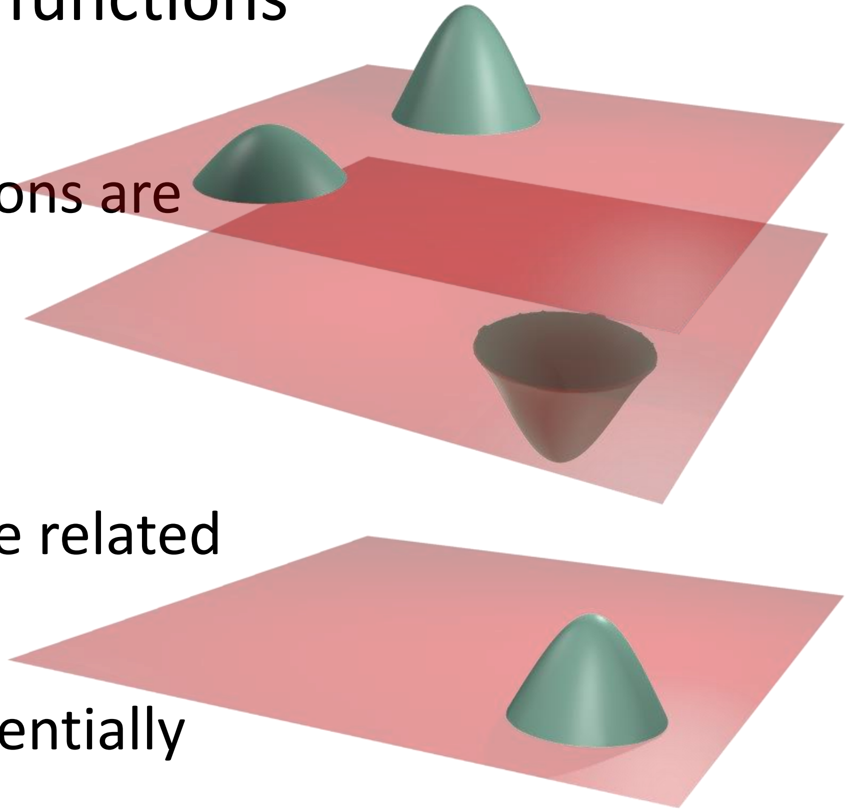
- Relationship between two functions
- ***Relationship Score*** **(τ)**
  - How related the two functions are
  - Captures the nature of the relationship
- ***Relationship Strength*** **(ρ)**
  - How often the functions are related

Weak Relationship



Relationship Evaluation

- Relationship between two functions
- ***Relationship Score*** **(*τ*)**
  - How related the two functions are
  - Captures the nature of the relationship
- ***Relationship Strength*** **(*ρ*)**
  - How often the functions are related
- *Significant* relationships
  - Monte Carlo tests filter potentially coincidental relationships



Scalar Functions

- Two types of scalar functions: *count* and *attribute*
- *Count functions*
  - Capture the activity of an entity corresponding to the data
  - Density function
    - E.g.: no. of taxi trips over space and time
  - Unique function
    - E.g.: no. of distinct taxis over space and time
- *Attribute functions*
  - Capture variation of the attribute
  - E.g.: average taxi fare over space and time
- Functions are computed at all possible resolutions

Relationship Querying

- Querying for meaningful relationships

Find relationships between $\boldsymbol{D_1}$ and $\boldsymbol{D_2}$ satisfying **CLAUSE**

- Only statistically significant relationships are returned
- **CLAUSE** can be used to filter relationships w.r.t. $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$.

↓

Significantly reduces the number of relationships the user needs to analyze !

*Goal: **guide** users in the data exploration process !*

(Some) Interesting Relationships

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

Collisions X *Traffic Speed*

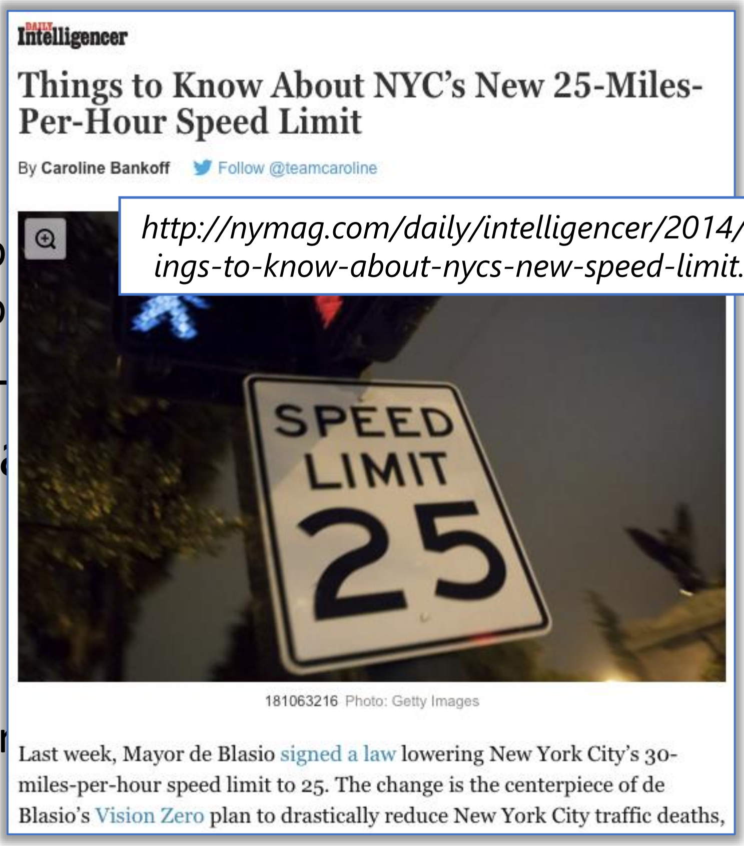
Positive relationship between numb
Positive relationship between numb

---

2. Why it is so hard to find a ta

Taxi Fare X *Precipitation*

Strong positive relationship between r
Taxi drivers are target earners!


Daily Intelligencer

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

By Caroline Bankoff Follow @teamcaroline

181063216 Photo: Getty Images

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

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

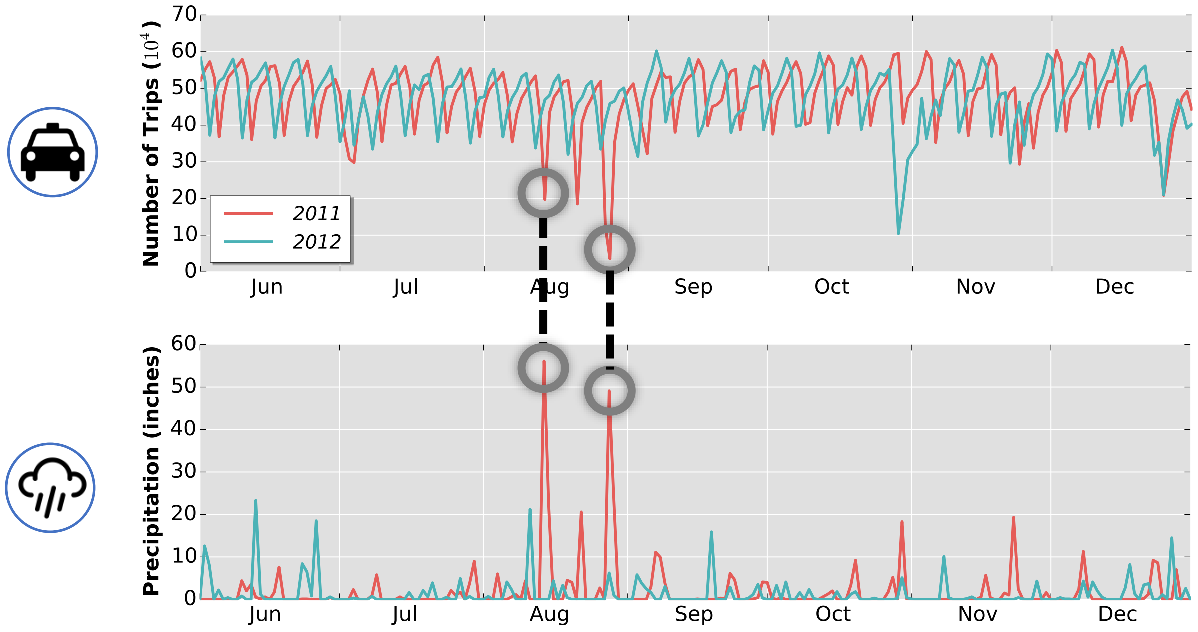
(Some) Interesting Relationships

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

 # Taxis *Precipitation*

X

Negative relationship between number of taxis and average precipitation



Many more details and experiments in the paper!

Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets, SIGMOD 2016.

Code, data, and experiments available at:

https://github.com/ViDA-NYU/data-polygamy

Weather data set is the most *polygamous*!