Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets

Fernando Chirigati[1], Harish Doraiswamy[1], Theodoros Damoulas[2,3], Juliana Freire[1]

[1]New York University [2]The University of Warwick [3]Alan Turing Institute

Urban Data Sets are Polygamous!

There are multiple interactions between entities of a city.
These are captured by the **relationships** between urban data sets.

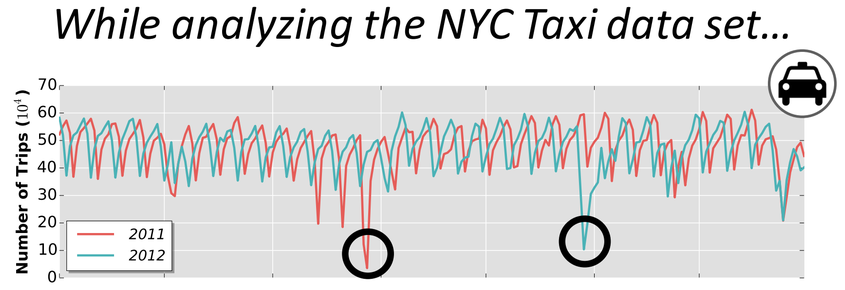
Relationship Queries

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

Enable *hypothesis generation* and *hypothesis testing*!

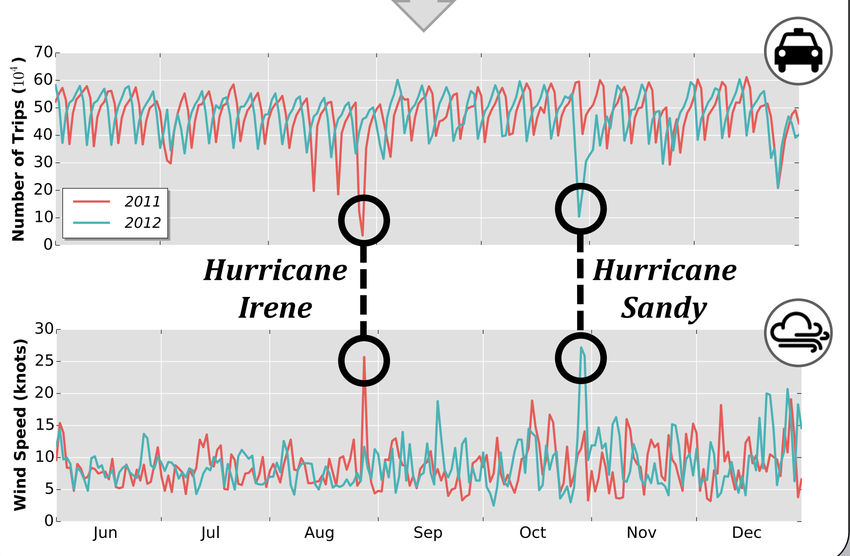
Hypothesis Generation

While analyzing the NYC Taxi data set…



*How to **explain** these features?*

*Find all data sets related to the **Taxi** data set*



Hypothesis Testing

NYC residents often struggle to get a taxi when it is raining.

Long-standing hypothesis:

- Taxi drivers set an income goal
- They reach goal faster on rainy days

*Can we test such hypothesis? **Yes!***

*Find relationships between **Taxi** and **Weather** data sets*

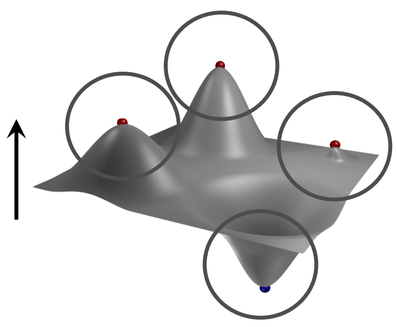
Taxi Fare **X** *Precipitation*

Challenge 1: How to define a data set relationship?

Our Approach: Computational Topology

1) Modeling the Data as a Terrain

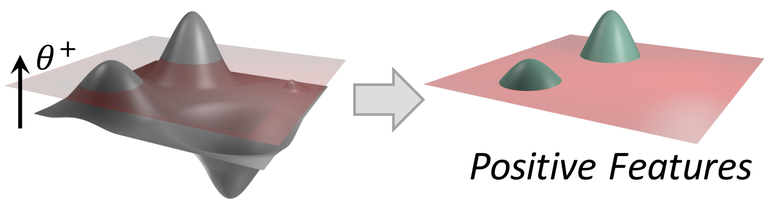
$$f : [\mathbb{S} \times \mathbb{T}] \rightarrow \mathbb{R}$$



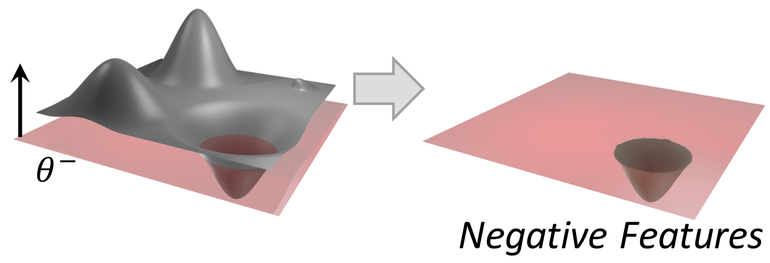
- Critical points
  - Peaks
  - Valleys

2) Identifying and Computing Topological Features

Neighborhoods of critical points

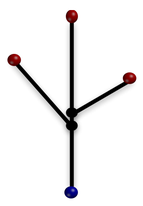


Positive Features



Negative Features

Index: Merge Tree

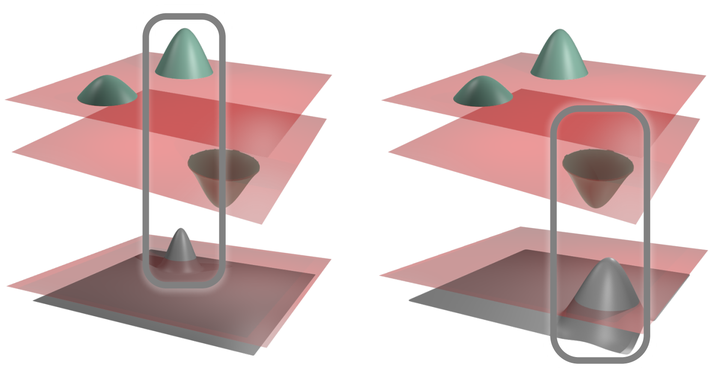


Thresholds θ^+ and θ^- computed in a data-driven approach

Computing features is **output sensitive**!

3) Identifying Topology-based Relationships

Relationship between features



Positive Relationship

Negative Relationship

Relationship between functions

- Relationship Score ($\boldsymbol{\tau}$)
  Nature of the relationship

$$\tau = \frac{\#p - \#n}{|\Sigma|}$$

- Relationship Strength ($\boldsymbol{\rho}$)
  How often they are related

$$precision = \frac{\#tp}{\#tp + \#fp} \qquad recall = \frac{\#tp}{\#tp + \#fn}$$

$$\rho = F_1(f_1, f_2) = 2 \times \frac{precision \times recall}{precision + recall}$$

Challenge 2: Data Complexity

- Multiple spatio-temporal resolutions
- Large data sets
- Relationships can be between any of the attributes

meaningful relationships
⇕
needle in a haystack

Our Approach:

- Monte Carlo tests filter potentially coincidental relationships
- Further filtering using $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$

Reduces the number of output relationships in around 99%

Interesting Relationships

Taxi and Wind Speed
– No. taxis × Wind speed

Taxi and Rainfall
– No. taxis × Precipitation
+ Taxi fare × Precipitation

Weather and Citi Bike
+ Snow precipitation × Trip duration
– Snow precipitation × Active stations

Weather is the most polygamous data set !