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

Fernando Chirigati *New York University*

Harish Doraiswamy *New York University*

Theodoros Damoulas *University of Warwick*
Alan Turing Institute

Juliana Freire *New York University*

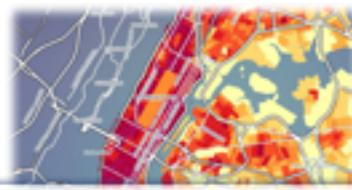
NYU

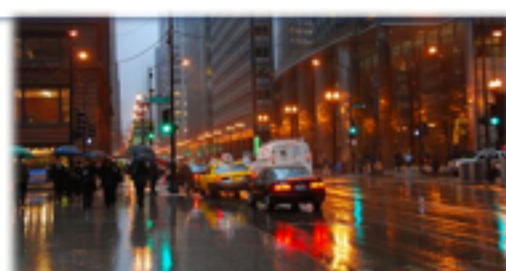
THE UNIVERSITY OF WARWICK

Data Exhaust from Cities

Infrastructure

Environment





People

Opportunity: make cities more efficient and sustainable, and improve the lives of citizens

While understanding NYC…

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

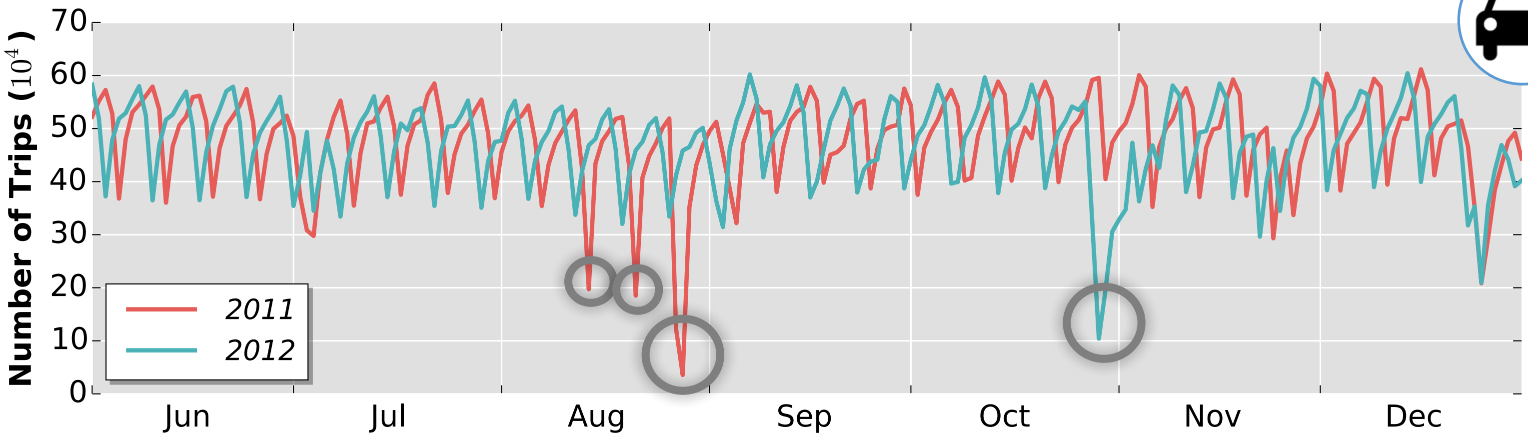
Good luck, lady. Photo: Jacobs Stock Photography/Getty Images

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

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

While understanding NYC…

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



NYC Taxi Data

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

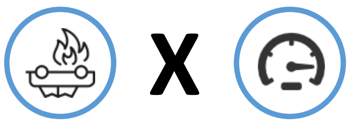
Where to start?
Which data sets to analyze?

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

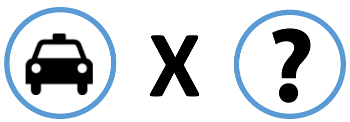


Q: Would a reduction in traffic speed reduce the number of accidents?

Find all relationships between Collisions and Traffic Speed data sets

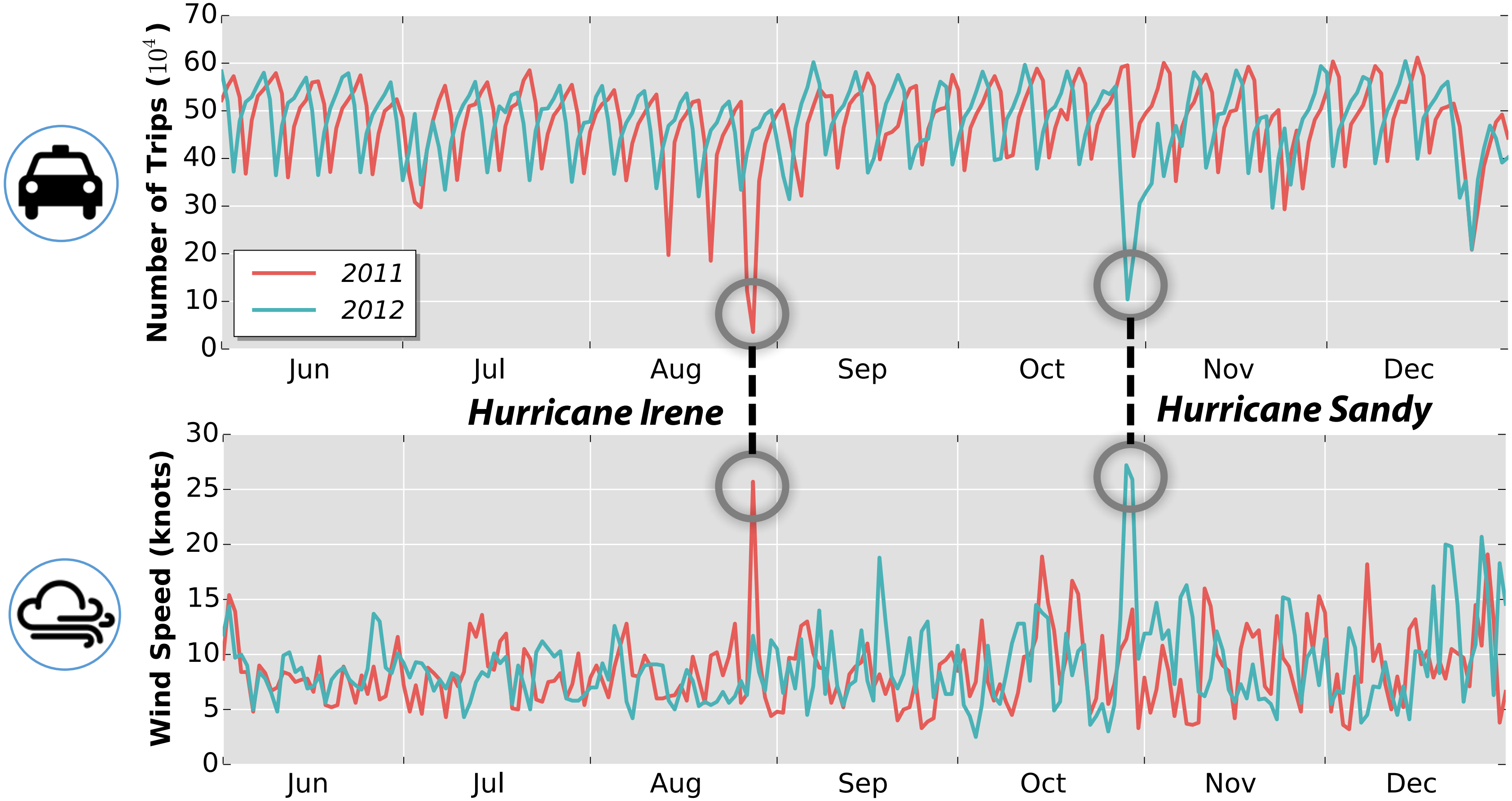
Q: Why the number of taxi trips is too low?

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, mutual information, DTW) 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**

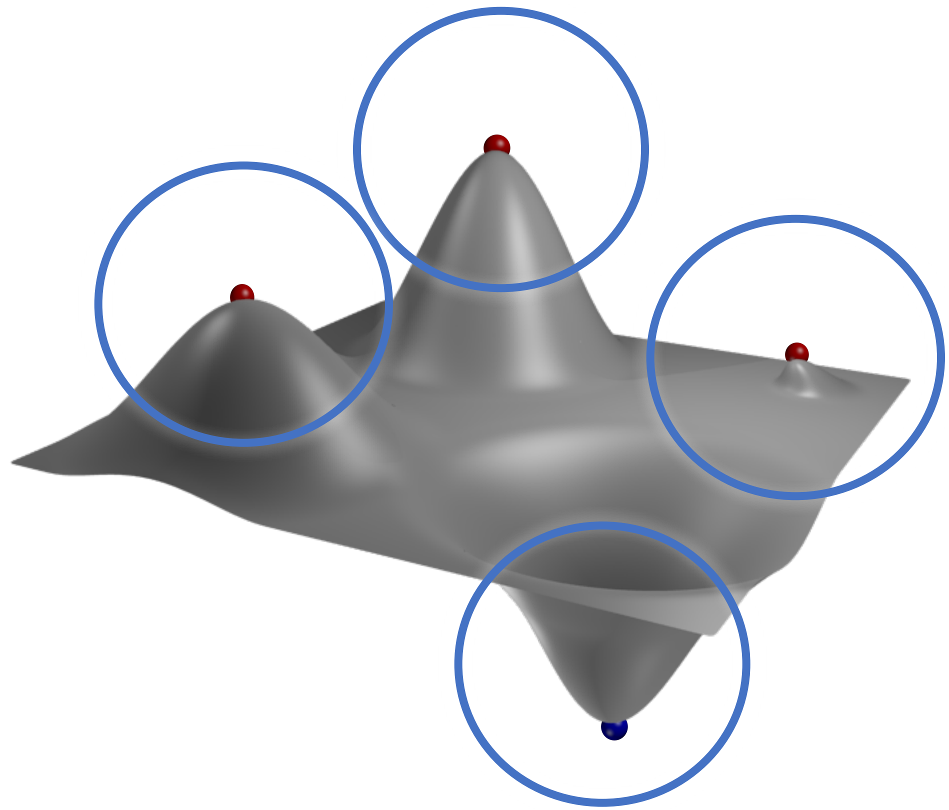
meaningful relationships ⟷ *needle in a haystack*

Our Approach:
The Data Polygamy Framework

1) How to define a relationship between data sets?
   Our solution: *Topology-based relationships*

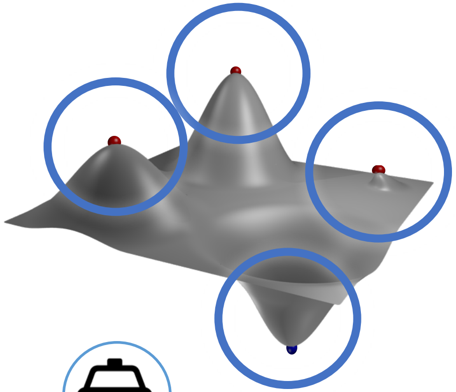
Topology? What!?

Peaks
Valleys
Critical Points



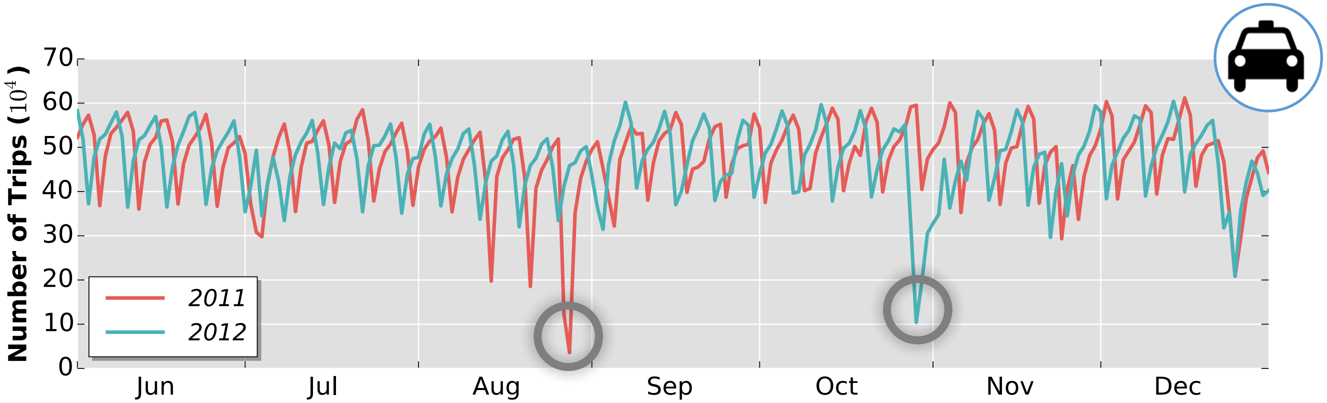
Our Approach:
The Data Polygamy Framework

1) How to define a relationship between data sets?
   Our solution: *Topology-based relationships*



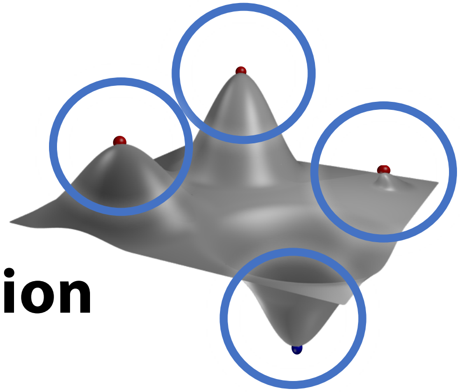
Why topology?

✓ Naturally captures the features of the data



Our Approach:
The Data Polygamy Framework

1) How to define a relationship between data sets?
   Our solution: *Topology-based relationships*



Why topology?

✓ Naturally captures the features of the data

✓ Works on data **in any resolution or dimension**

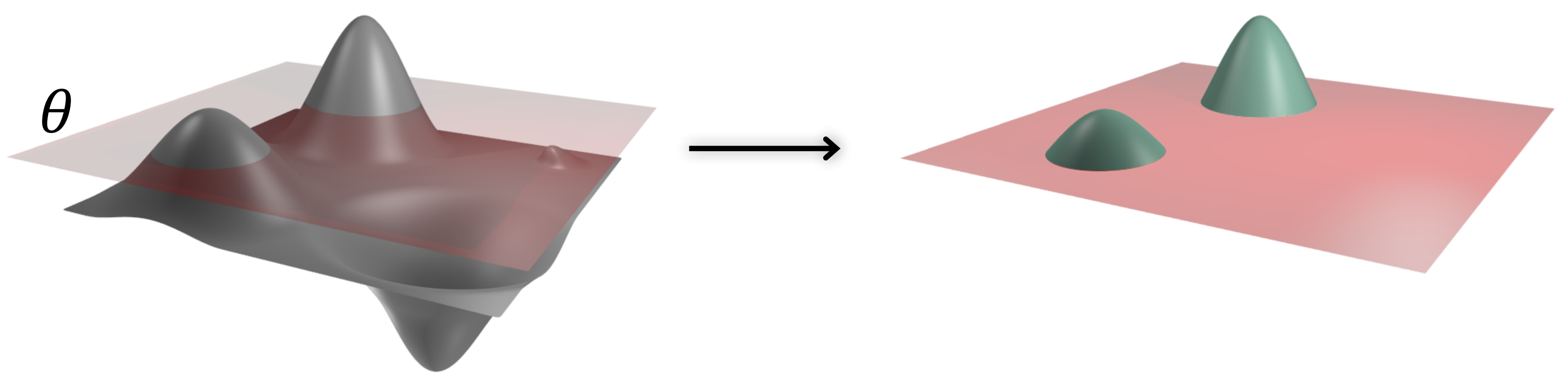
✓ Very efficient

Our Approach:
The Data Polygamy Framework

1) How to define a relationship between data sets?
   Our solution: *Topology-based relationships*

Key Techniques:

✓ Use of *topological persistence* to automatically compute thresholds (data-driven approach)



Our Approach:
The Data Polygamy Framework

1) How to define a relationship between data sets?
   Our solution: *Topology-based relationships*

Key Techniques:

✓ Use of *topological persistence* to automatically compute thresholds (data-driven approach)

✓ *Merge Tree Index* to identify topological features

Our Approach: The Data Polygamy Framework

2) Large data complexity
   Our solution: *Monte Carlo significance tests*

Why?

✓ Removes potentially coincidental relationships

↓

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

Summary of Experiments

Data Polygamy implemented using *Hadoop*

Performance

- Framework can evaluate relationships at a rate greater than 10K relationships per minute
- Using significance tests: decrease of around 99% on the number of output relationships!
- Approach is robust to noise

(Some) Interesting Relationships

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

Collisions *Traffic Speed*

 **X** 

Positive relationship between number
Positive relationship between number

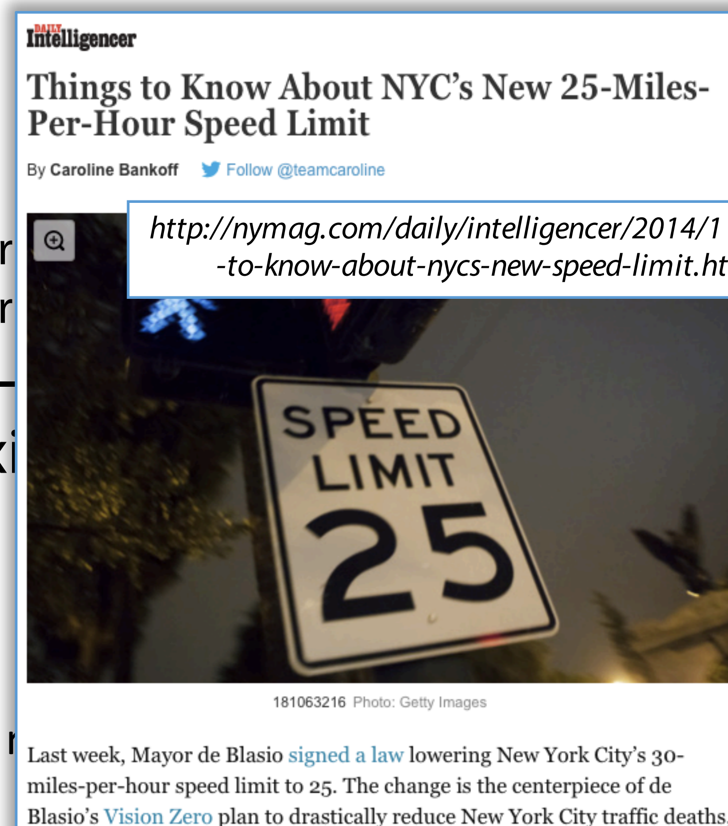
---

2. Why it is so hard to find a taxi

Taxi Fare *Precipitation*

 **X** 

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

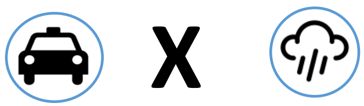
Last week, Mayor de Blasio signed a law lowering New York City's 30-miles-per-hour speed limit to 25. The change is the centerpiece of de Blasio's Vision Zero plan to drastically reduce New York City traffic deaths,

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

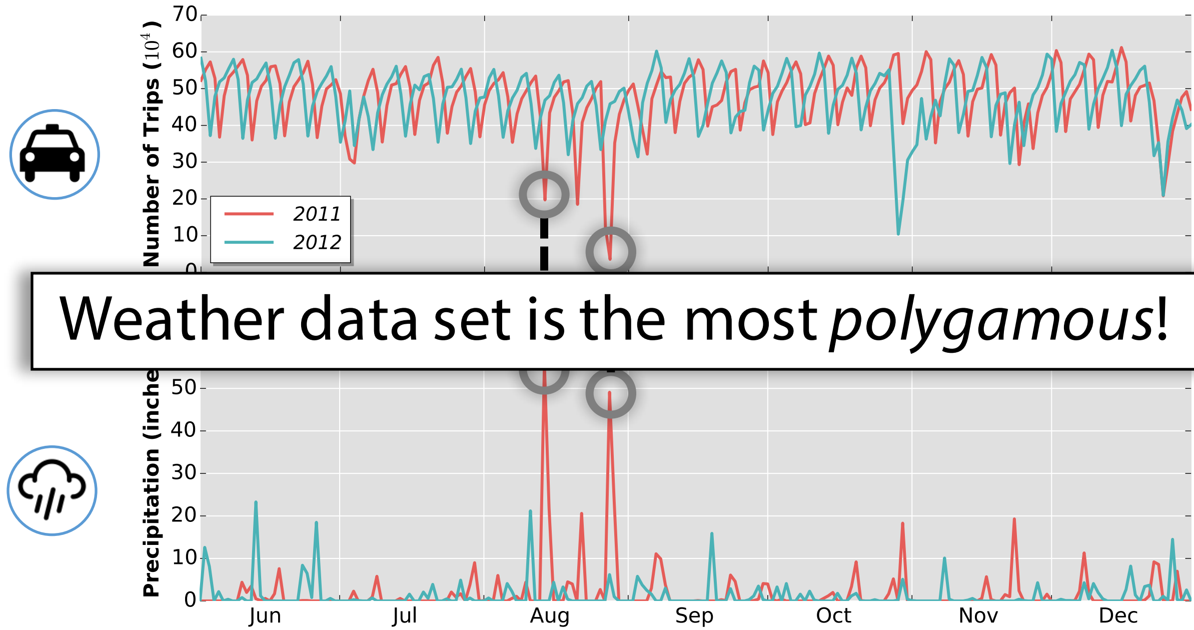
(Some) Interesting Relationships

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

 # Taxis *Precipitation*



Negative relationship between number of taxis and average precipitation



Weather data set is the most *polygamous*!

Many more details and experiments in the paper!

Our poster was yesterday…

… but feel free to talk to us!

I ♥♥♥♥ DATA POLYGAMY

Code, data, and experiments available at:

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

Acknowledgments: Funding: National Science Foundation, Moore-Sloan Data Science Environment at NYU, DARPA, and Alan Turing Institute

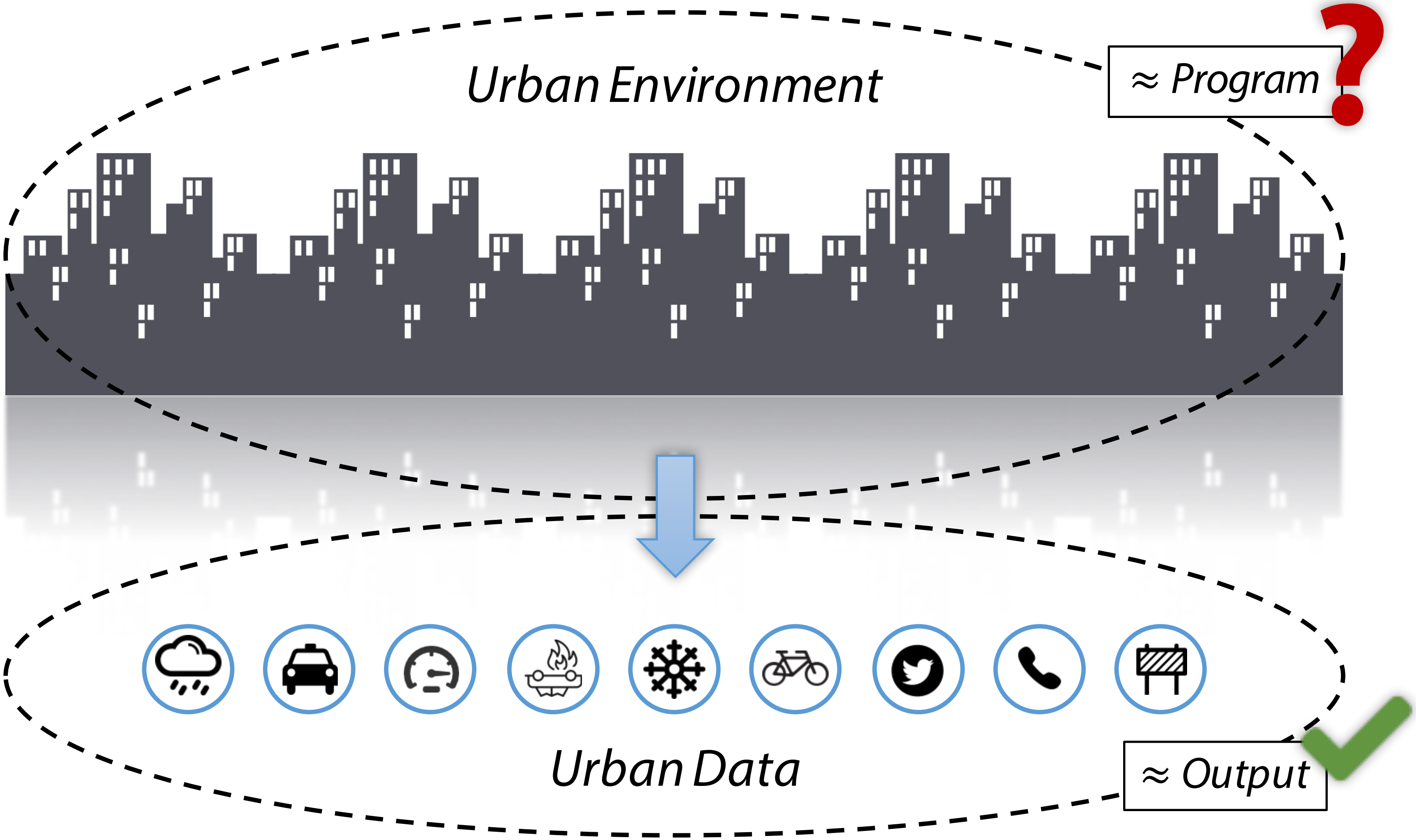
Insightful comments and suggestions from Divesh Srivastava and anonymous reviewers

Thank you!

NYU

THE UNIVERSITY OF WARWICK

Additional Material

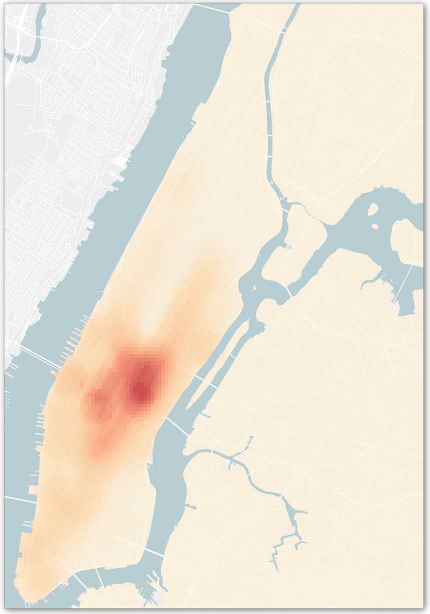


Opportunity: make cities more efficient and sustainable, and improve the lives of citizens

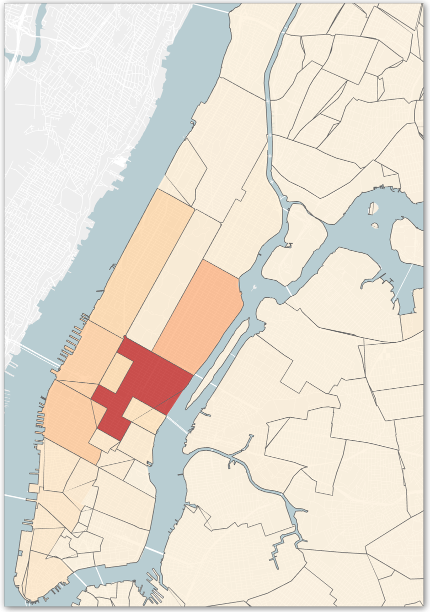
Data Set Transformation
↓
Feature Identification
↓
Relationship Evaluation

Scalar Functions

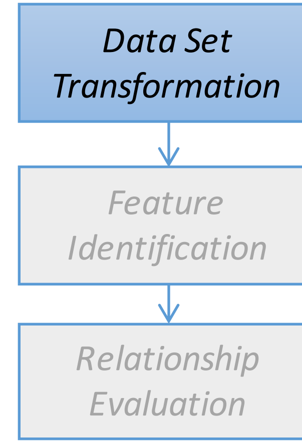
- Each data set represented as a *time-varying scalar function*
  - $f : [\mathbb{S} \times \mathbb{T}] \rightarrow \mathbb{R}$
  - Maps each point in the domain (city) over time to a scalar value



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

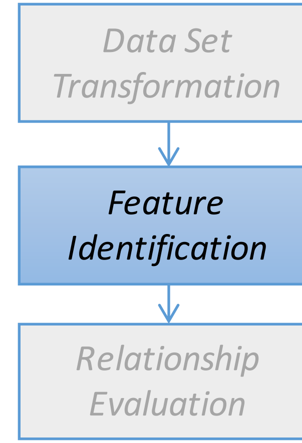


$\mathbb{S}$: *Neighborhood Resolution*



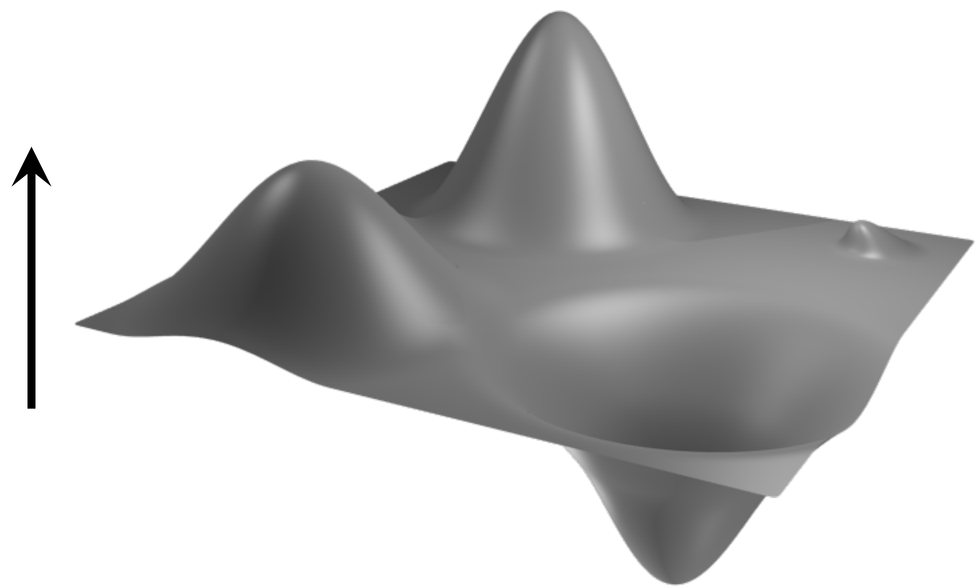
Scalar Functions

- 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



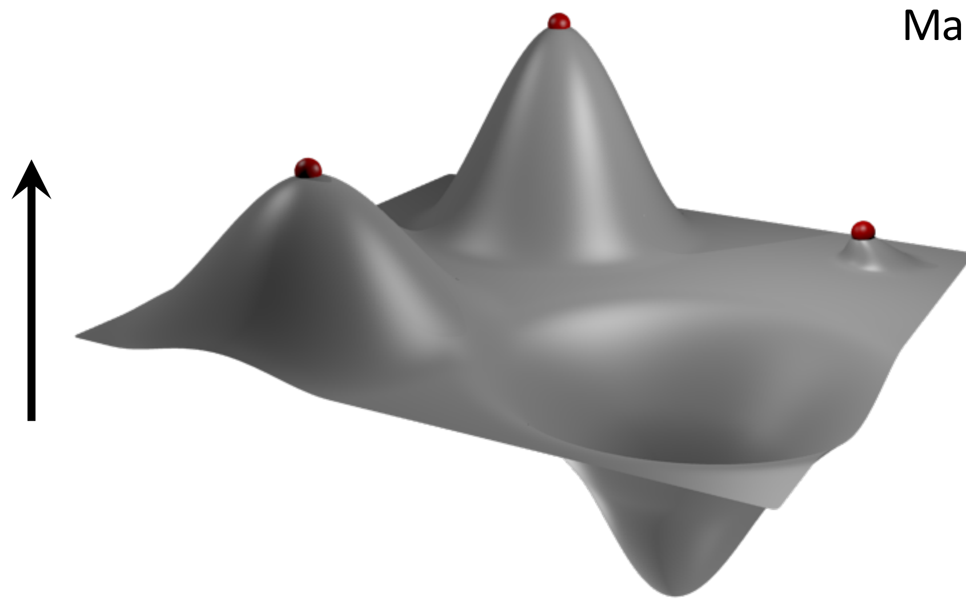
Identifying Topological Features

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



Identifying Topological Features

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



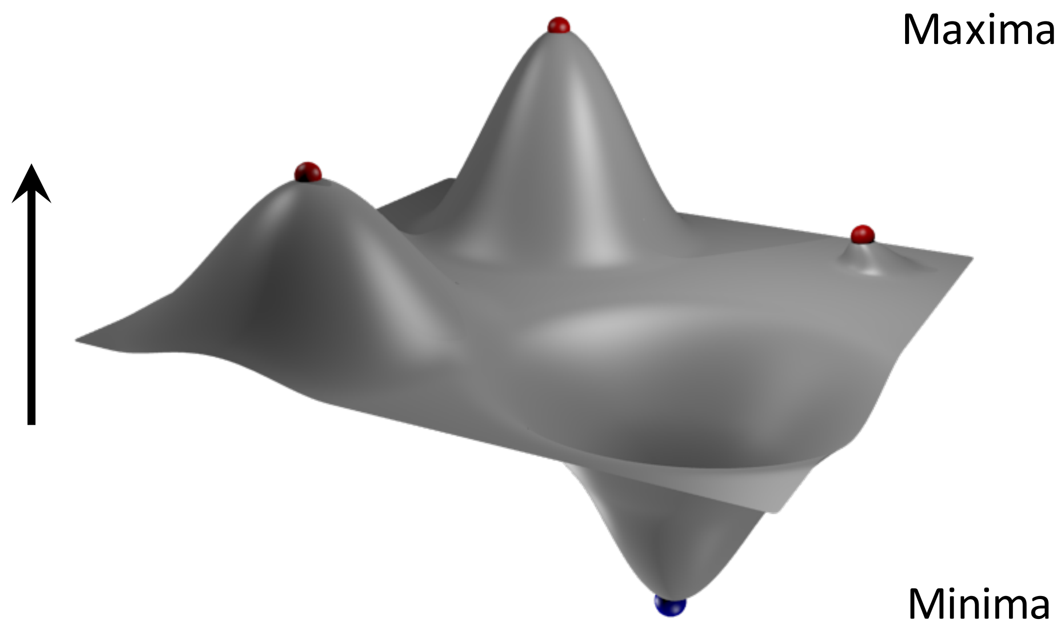
Maxima

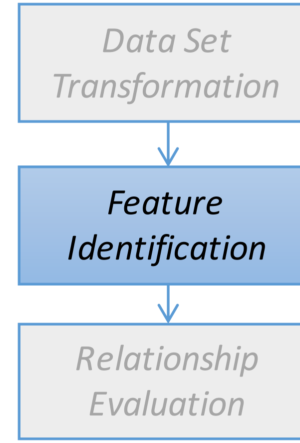
Data Set Transformation → ***Feature Identification*** → *Relationship Evaluation*

Data Set Transformation → **Feature Identification** → Relationship Evaluation

Identifying Topological Features

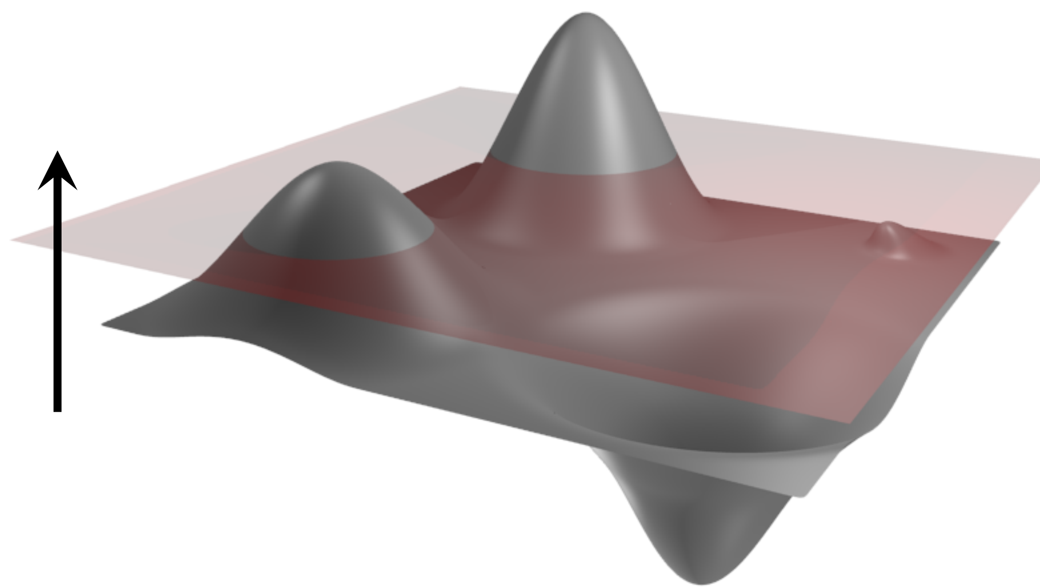
- Topological features of the scalar function
  - Neighborhoods of critical points



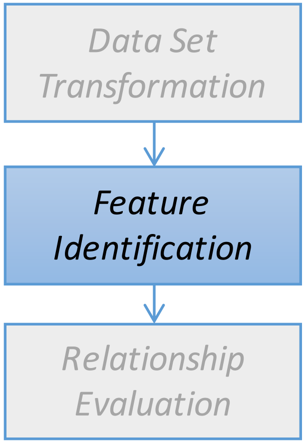


Identifying Topological Features

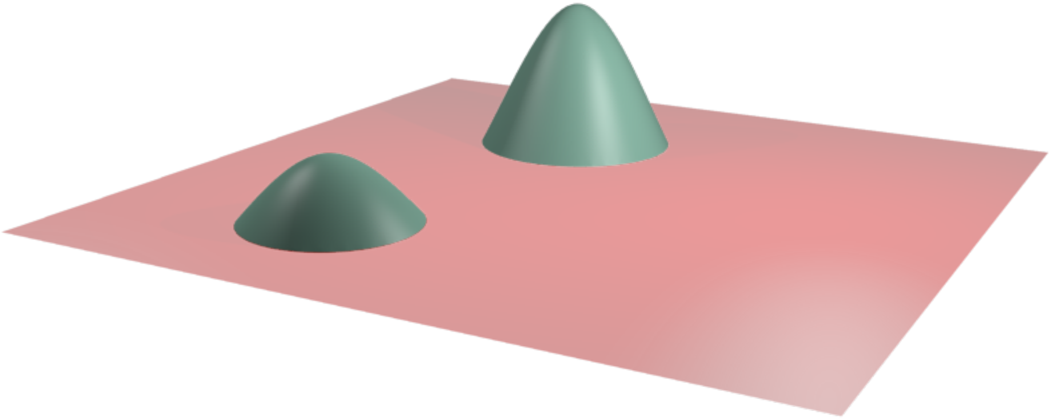
- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold



Identifying Topological Features



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



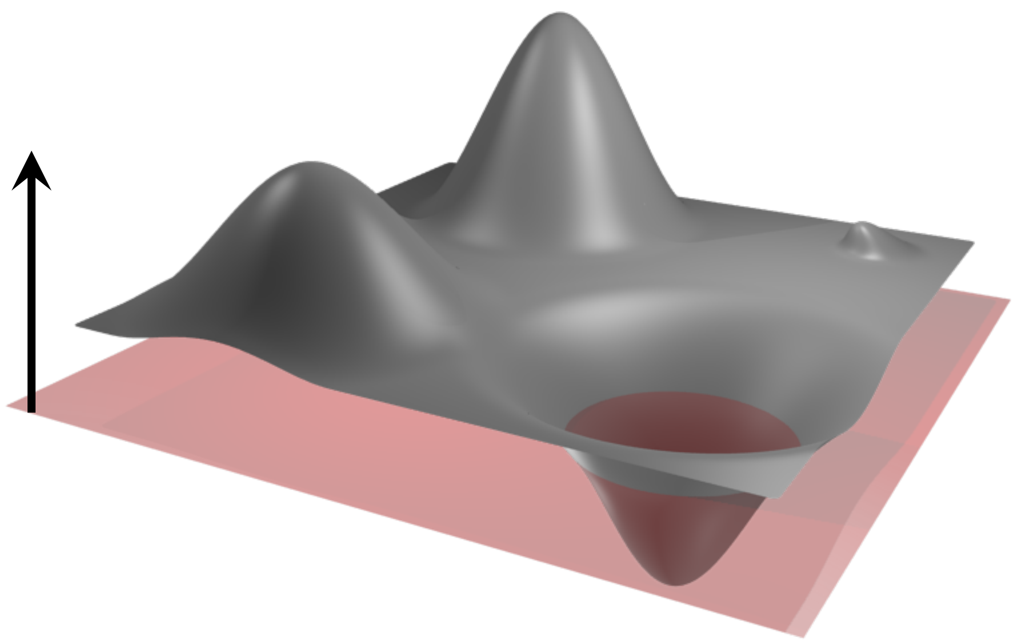
Data Set Transformation

Feature Identification

Relationship Evaluation

Identifying Topological Features

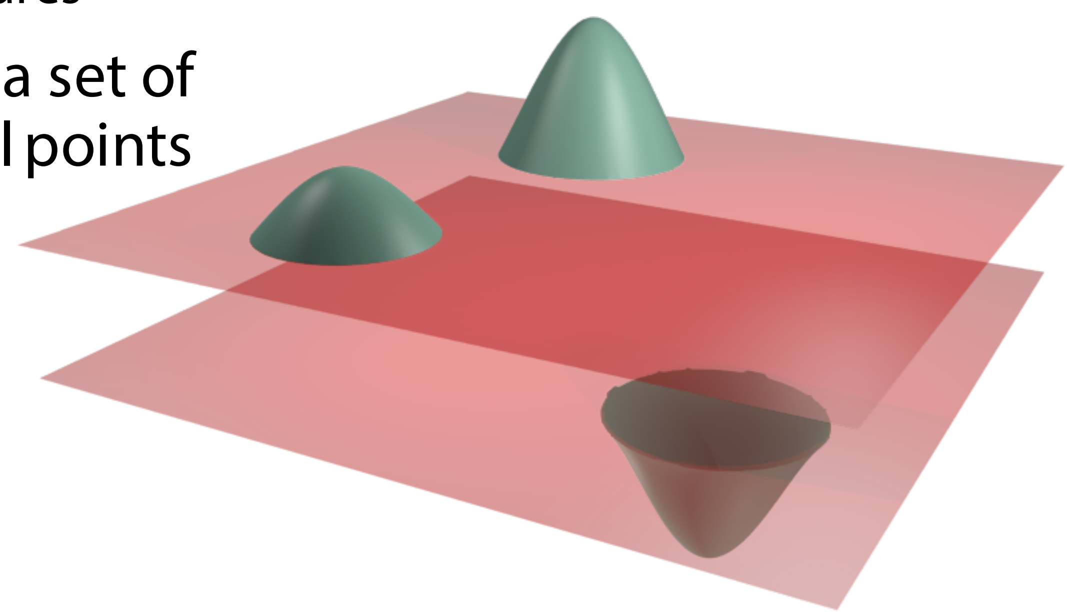
- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold
  - Positive Features



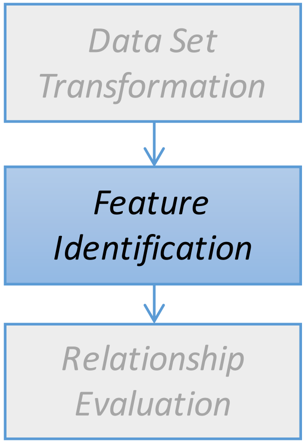
Data Set Transformation → **Feature Identification** → Relationship Evaluation

Identifying Topological Features

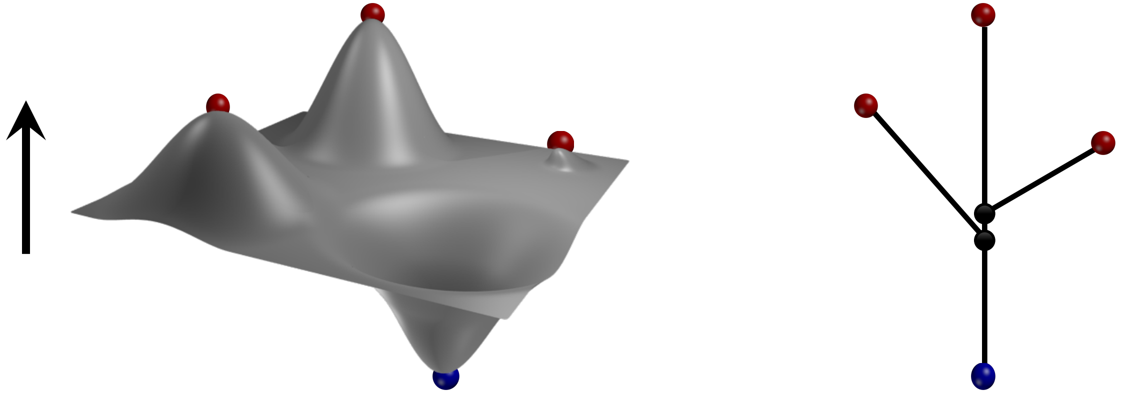
- Topological features of the scalar function
  - Neighborhoods of critical points
- Neighborhood defined by a threshold
  - Positive Features
  - Negative Features
- Represented as a set of spatio-temporal points



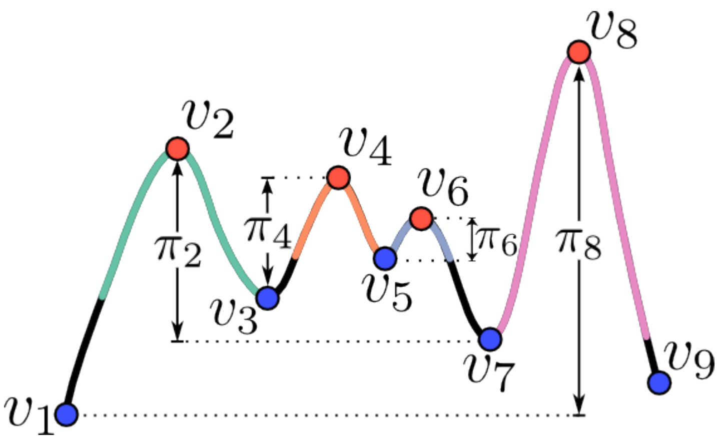
Computing Topological Features



- Index: Merge Tree
  - Topological data structure
  - Tracks evolution of the topology of level sets
  - Data can be of any dimension
- Output-sensitive time complexity
- Feature thresholds are computed in a data-driven approach
- Features are computed at all possible resolutions



Topological Thresholds



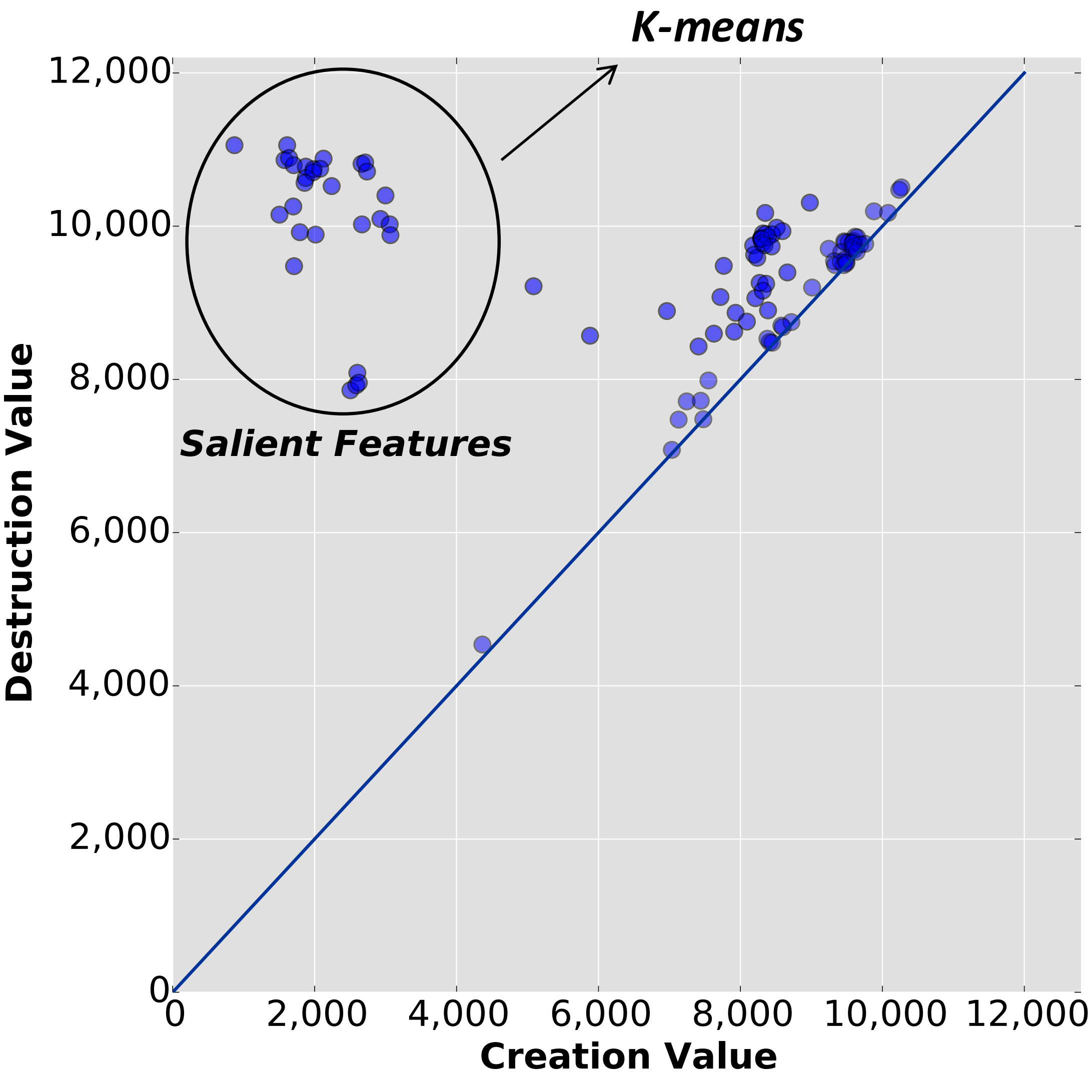
- π is the persistence value
  - E.g.: π_8 has v_8 as creator and v_1 as destroyer

Salient Features

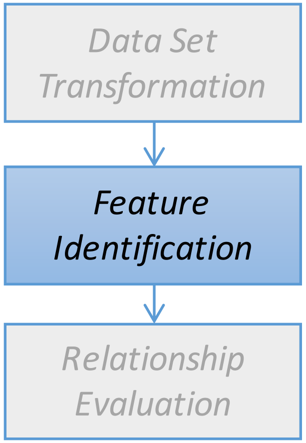
Data Set Transformation

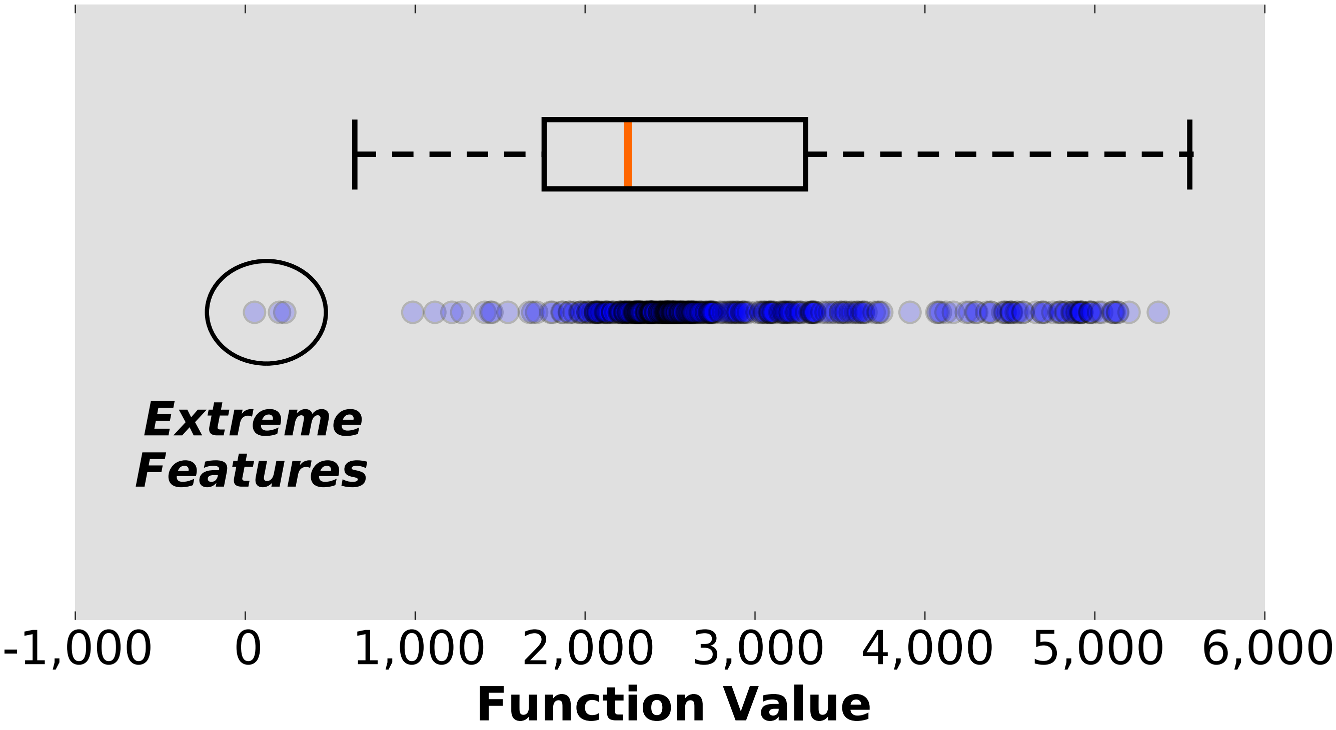
Feature Identification

Relationship Evaluation



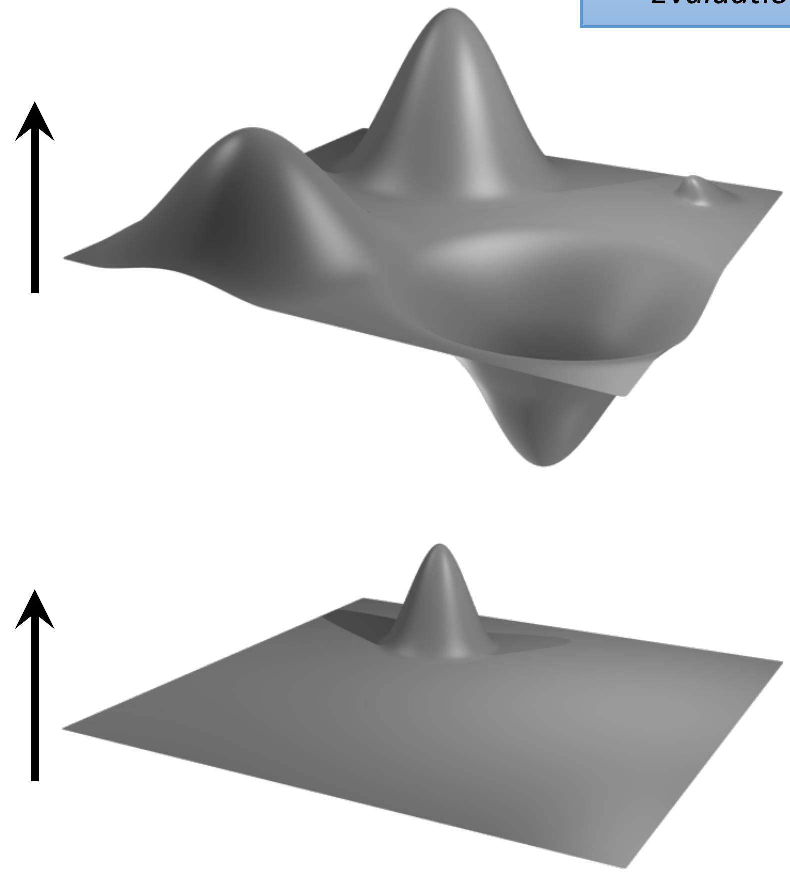
Extreme Features

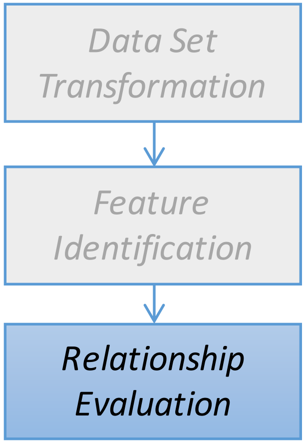




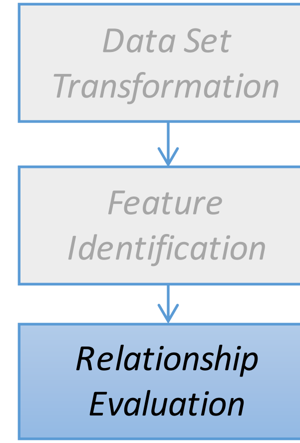
Relationship Evaluation

- Relationship between features

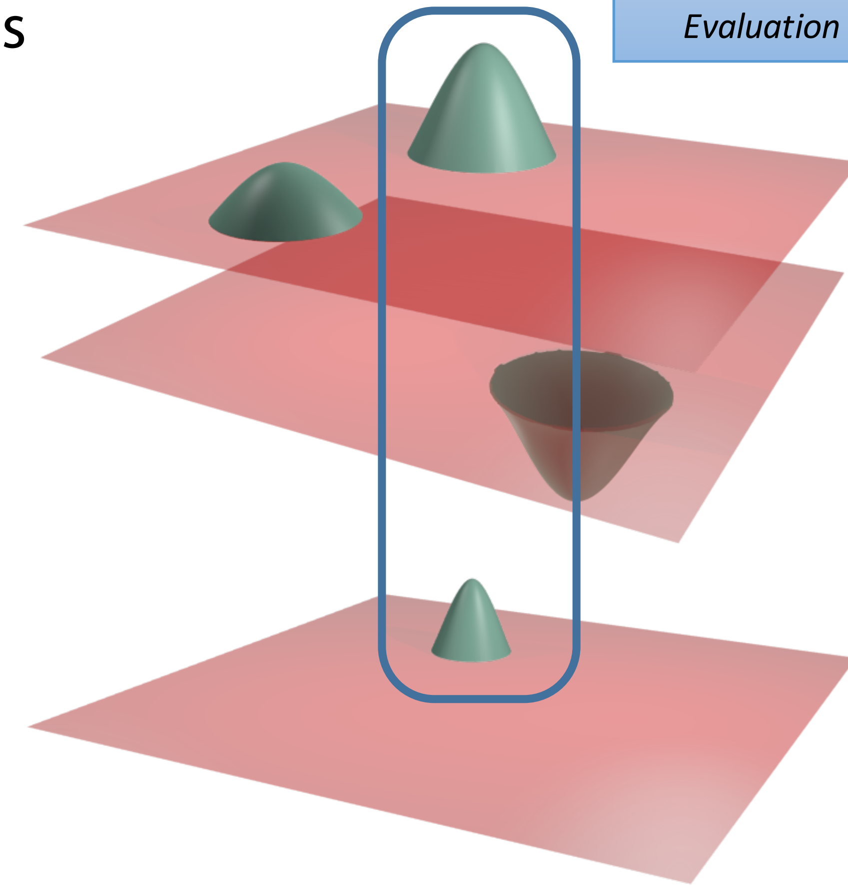




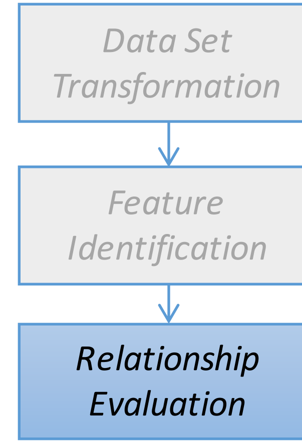
Relationship Evaluation



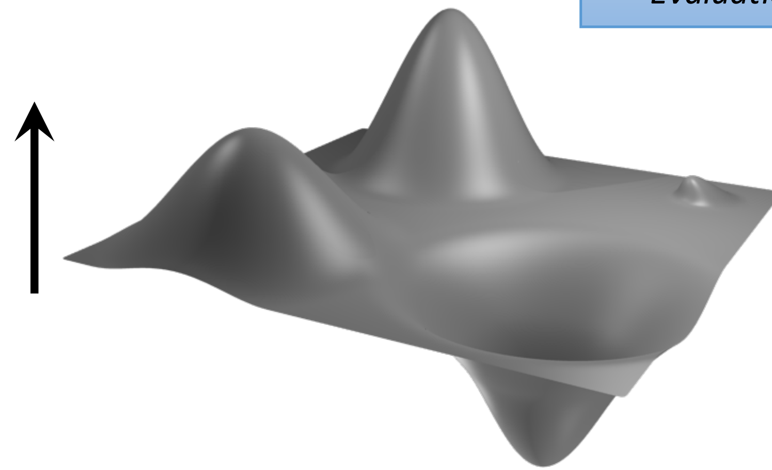
- Relationship between features
  - Related features
  - **Positive** Relationship

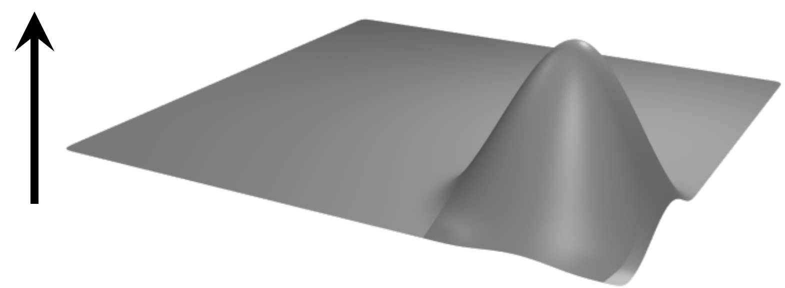


Relationship Evaluation

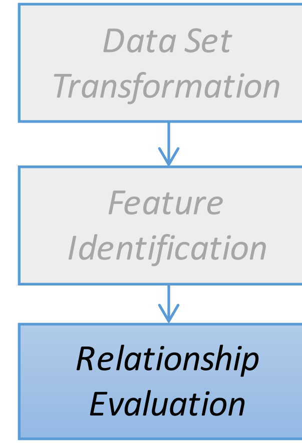


- Relationship between features
  - Related features
  - **Positive** Relationship

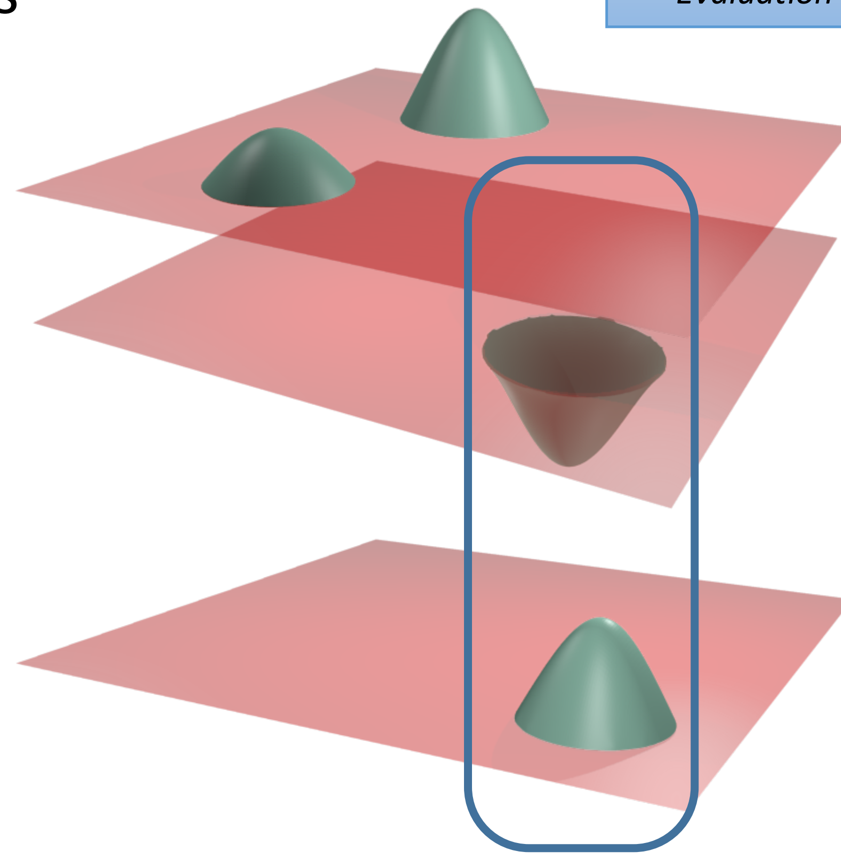


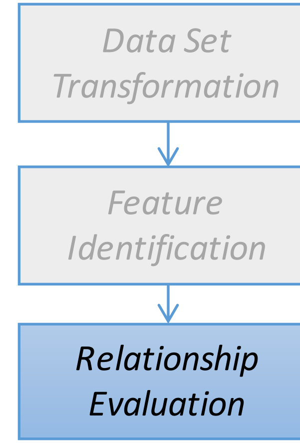


Relationship Evaluation



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

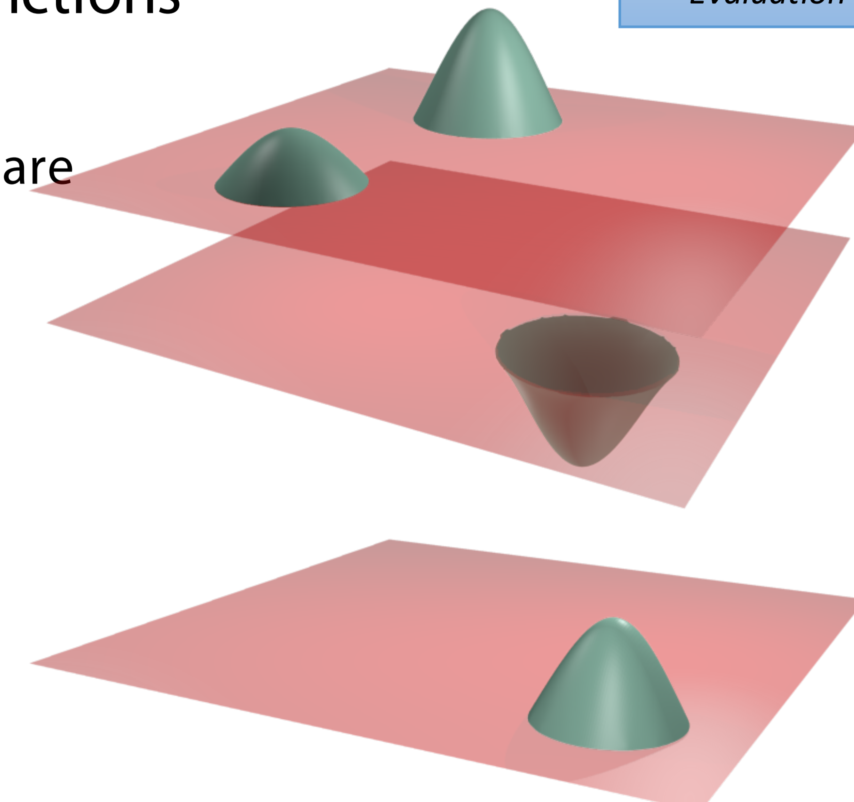




Relationship Evaluation

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

Negative Relationship

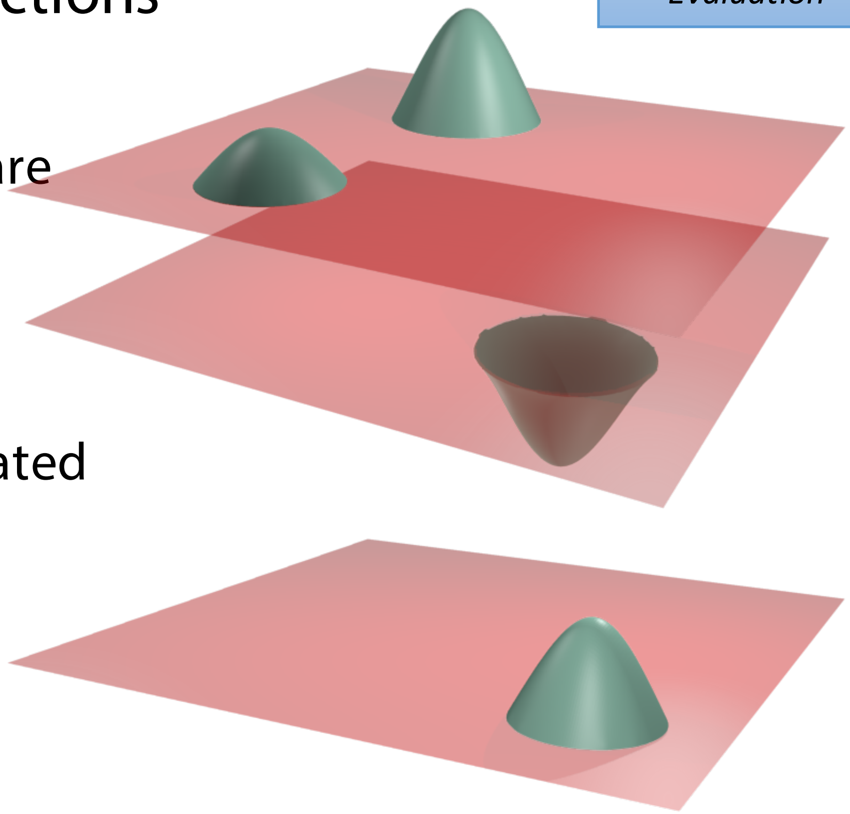


Data Set Transformation → Feature Identification → ***Relationship Evaluation***

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



Data Set Transformation

↓

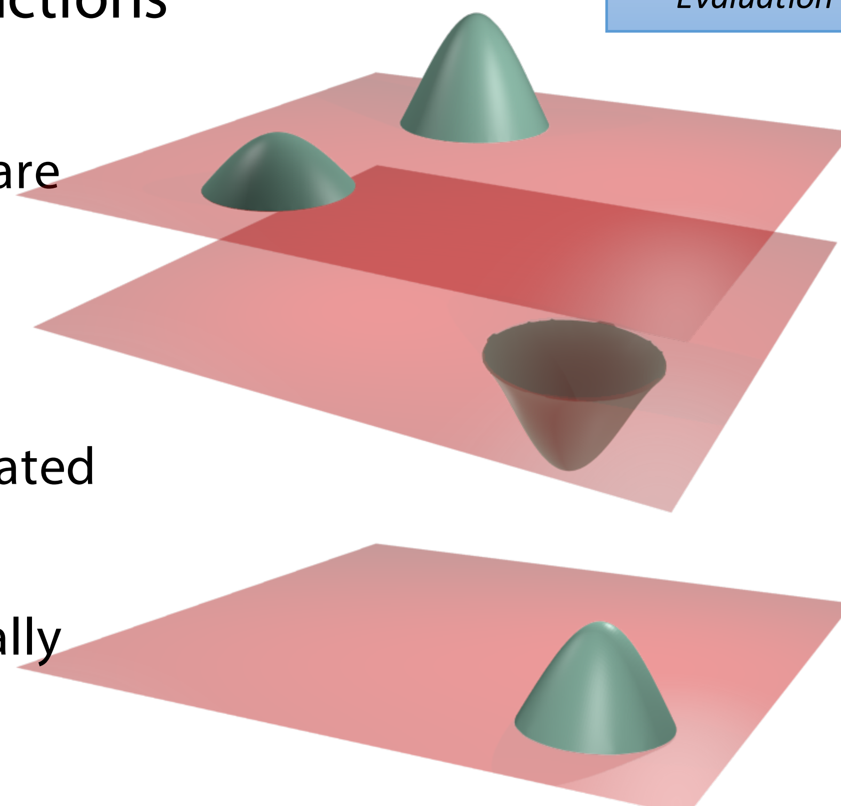
Feature Identification

↓

Relationship Evaluation

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



Data Set Transformation

↓

Feature Identification

↓

Relationship Evaluation

Relationship Evaluation

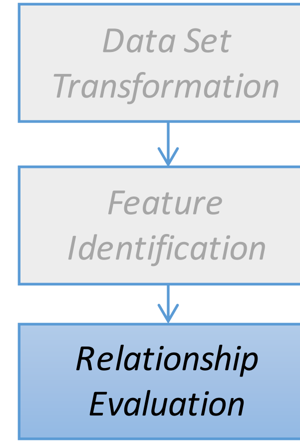
- Relationship Score

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

  - *#p* – no. of positively feature-related points
  - *#n* – no. of negatively feature-related points
  - $\boldsymbol{\Sigma}$ – set of feature-related points

  - $\boldsymbol{\tau}$ close to 1 $\Rightarrow$ positive relationship
  - $\boldsymbol{\tau}$ close to -1 $\Rightarrow$ negative relationship

Relationship Evaluation



- Relationship Strength

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

- $\mathbf{\Sigma}_1$ – set of points that are features in f_1
- $\mathbf{\Sigma}_2$ – set of points that are features in f_2

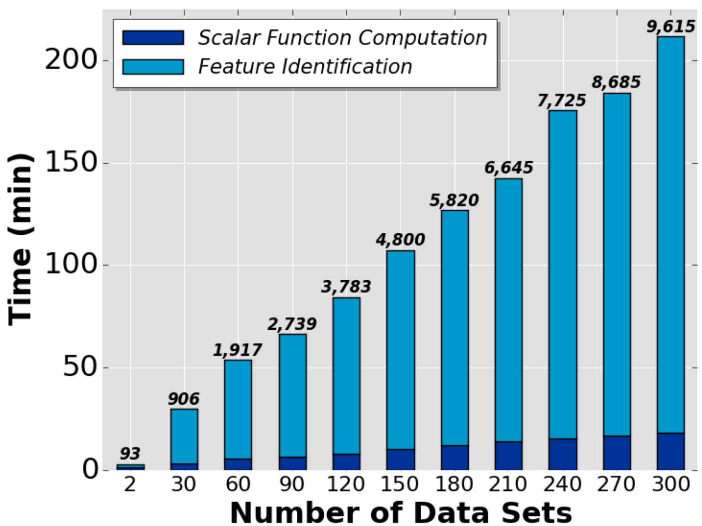
Let x be a spatio-temporal point:

- $x \in \mathbf{\Sigma}_1$ and $x \in \mathbf{\Sigma}_2 \Rightarrow$ true positive
- $x \in \mathbf{\Sigma}_1$ and $x \notin \mathbf{\Sigma}_2 \Rightarrow$ false positive
- $x \notin \mathbf{\Sigma}_1$ and $x \in \mathbf{\Sigma}_2 \Rightarrow$ false negative

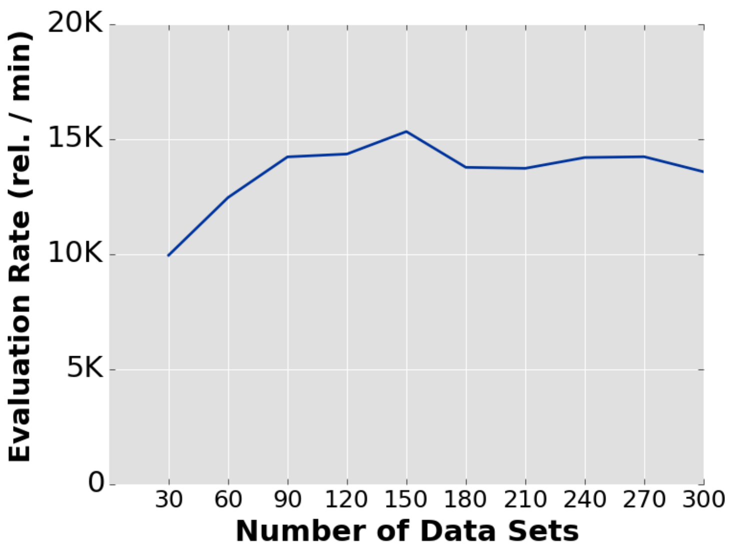
- $\boldsymbol{\rho}$ close to 1 $\Rightarrow$ strong relationship: a feature in one function almost always indicate a feature in the other function

Efficiency

- *NYC Open Data*
  - Index created and features computed for all possible resolutions
  - Query: Find all relationships among the given subset of data sets



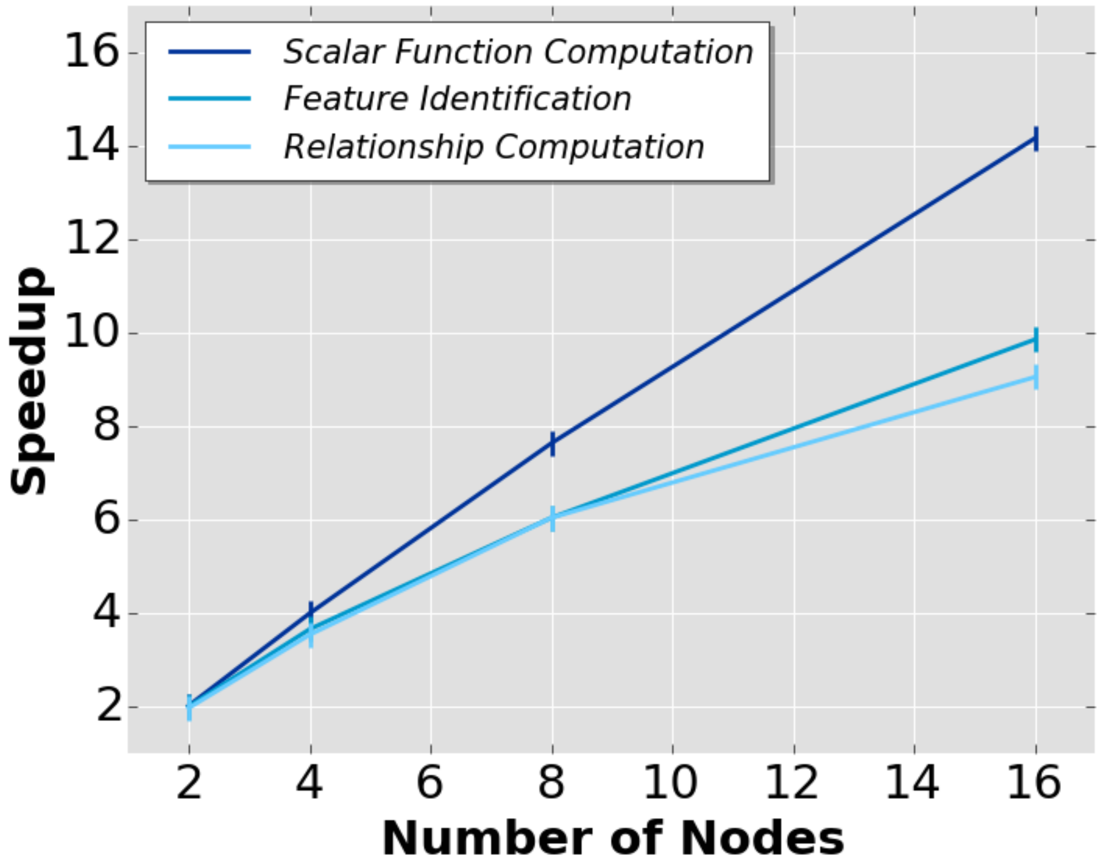
Index Creation



Query Rate

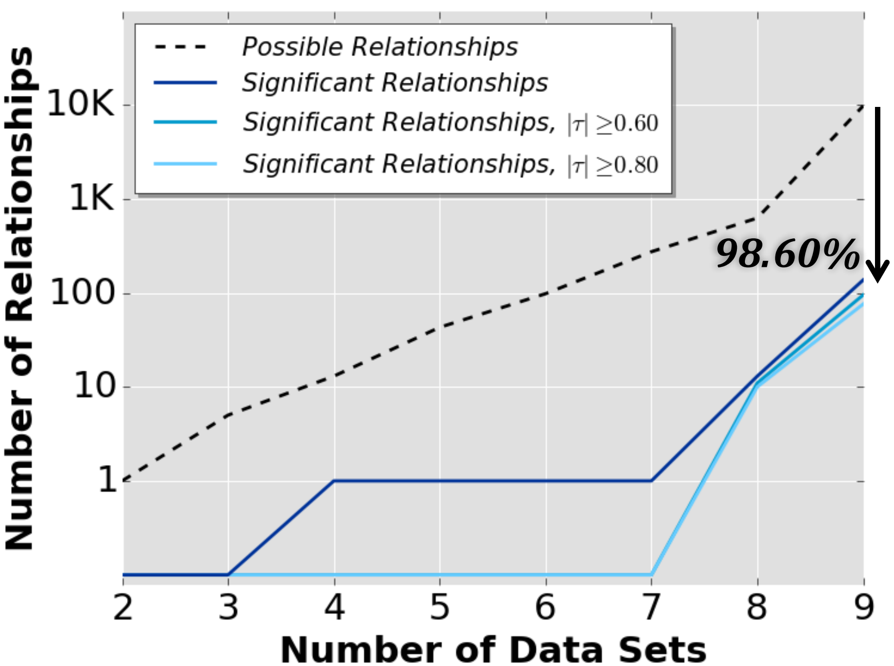
Scalability

- *NYC Urban* on AWS

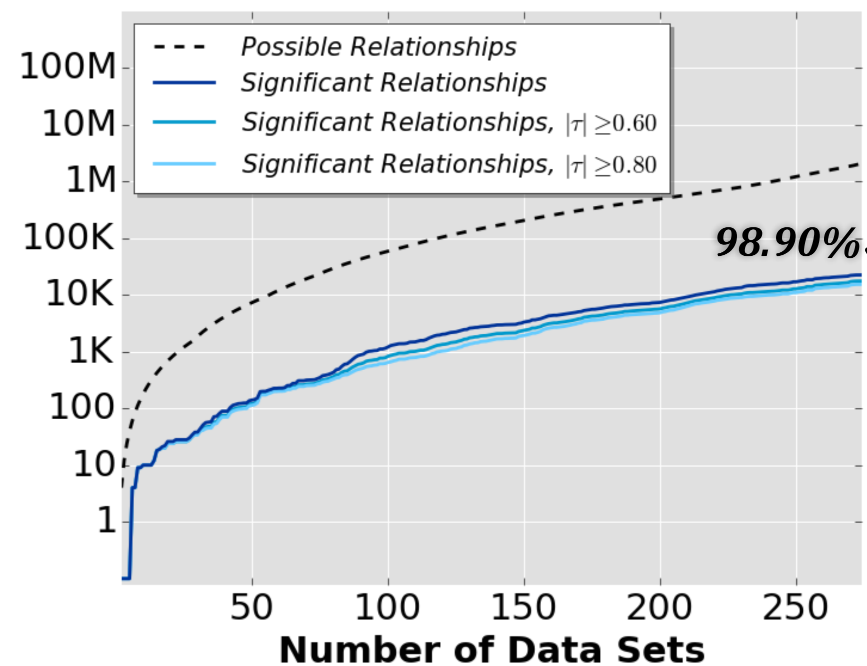


Relationship Pruning

- Query
  - Find all relationships among the given subset of data sets
  - (week, city) resolution



NYC Urban



NYC Open Data

Robustness

- Evaluation using the density of taxi trips
  - Random Gaussian noise; noise amount bounded by a fraction of the inter-quartile range of the function