Data Polygamy

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

Fernando Chirigati, **Harish Doraiswamy**, Theodoros Damoulas, Juliana Freire

NYU TANDON SCHOOL OF ENGINEERING
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER
THE UNIVERSITY OF WARWICK

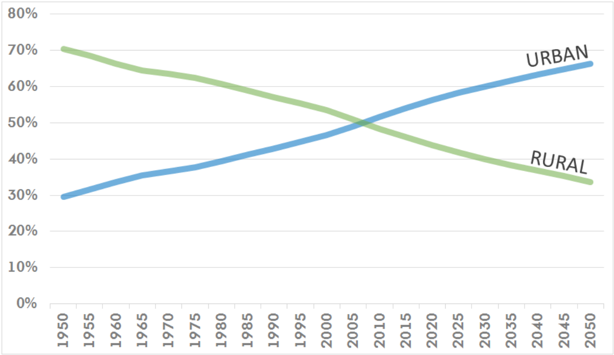
Big Urban Data: What's the Big Deal?

Cities are the loci of economic activity

50% of the world population lives in cities

By 2050, this number will grow to 70%

Growth leads to many problems



Plot by [Akantamn]

Data is the light at the end of the tunnel

Data Exhaust from Cities

Infrastructure

Environment

Photo by *MTA*

Photo by *Yinka Oyesiku*

People

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

While Exploring NYC Data...

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?

Intelligencer

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

By Annie Lowrey Follow @AnnieLowrey

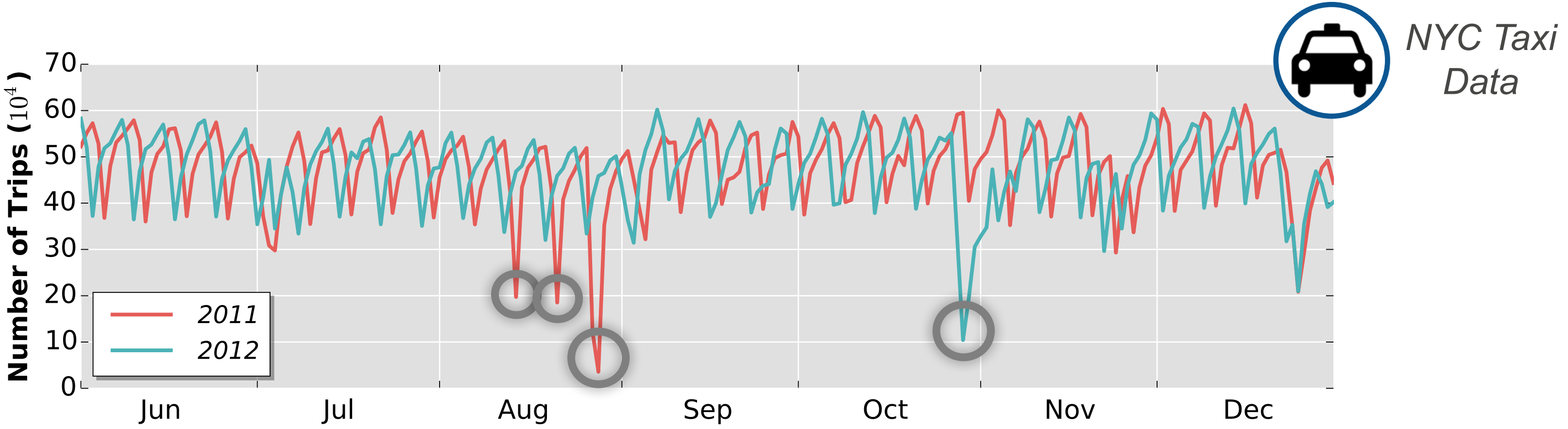
Good luck, lady. Photo: Jacobs Stock Pho

It's pouring rain. You're running late. You de
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 Exploring NYC Data...

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?



Urban Data Interactions

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

Uncovering relationships

⇩

Uncovering important attributes

*Urban data sets can be very **Polygamous**!*

Data are available...
... but we are talking about **big** data!

NYC OpenData

1,200 data sets
(and counting)

Where to start? Which data sets to analyze?

NYC OpenData

> 300 data sets
are **spatio-temporal**

Which spatio-temporal slices 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

Identify important variables

Hypothesis Testing

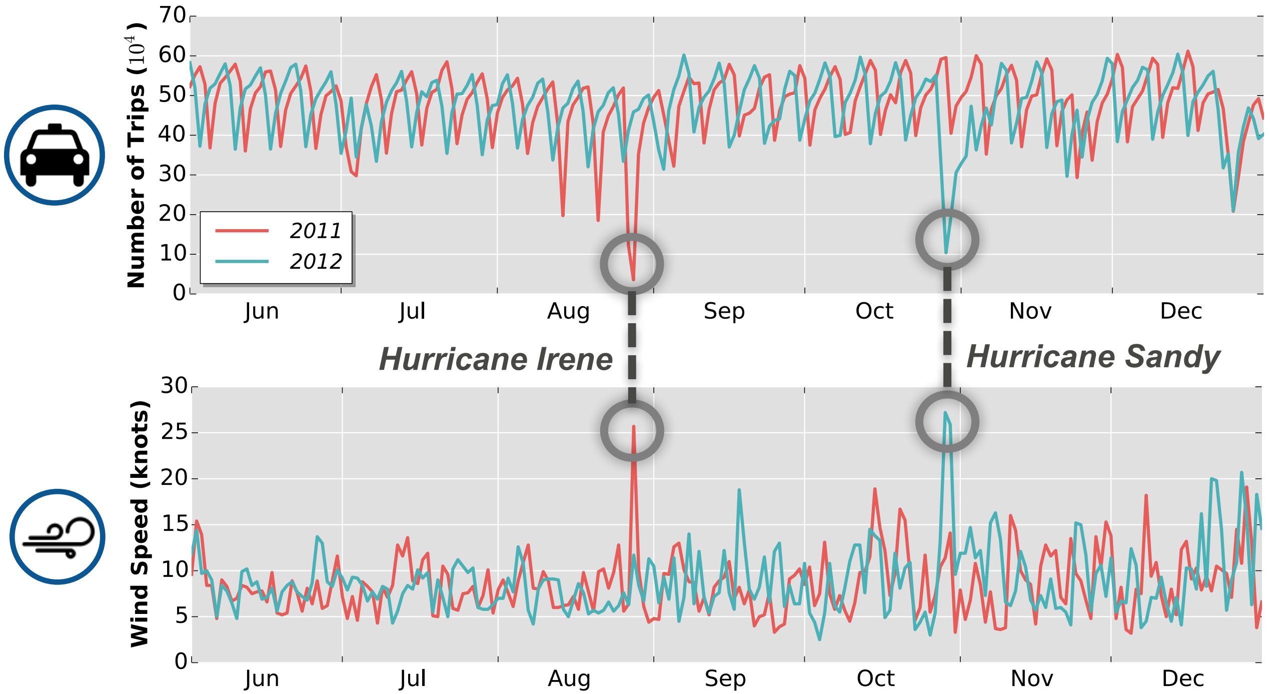
Q: Would a reduction in traffic speed reduce the number of accidents?
Find all relationships between Collisions and Traffic Speed data sets

Q: Why the number of taxi trips is too low?
Find all data sets related to the Taxi data set

Hypothesis Generation

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

NYC OpenData | 8 attributes per data set

> 200 attributes

Combinatorially large number of relationships to evaluate
≈**2.4 million** possible relationships among NYC Open Data alone for a **single spatio-temporal resolution**

meaningful relationships ⟷ *needle in a haystack*

Our Approach: *Data Polygamy*

1) How to define a relationship between data sets?

   Our solution: *Topology-based relationships*

2) Large data complexity

   Our solution: *Implementation using map-reduce*

NYU VIDA

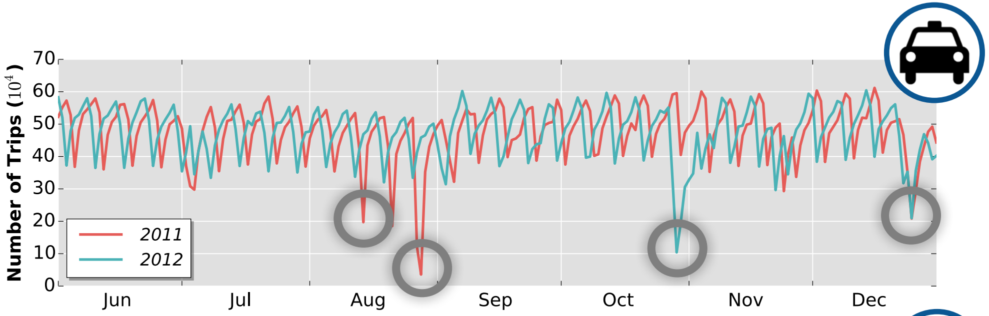
Topology-based Relationships

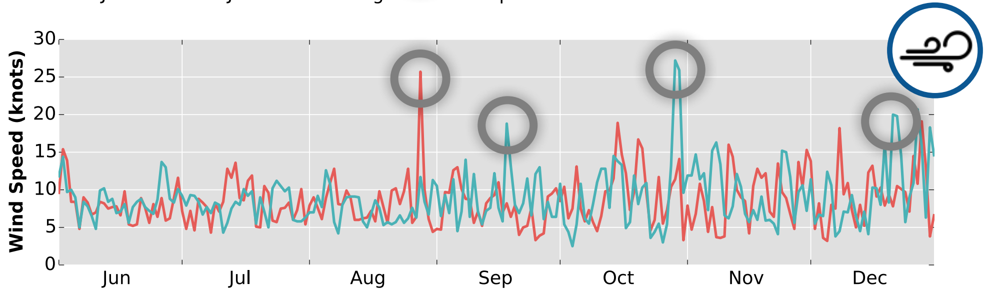
Topological Features

Valleys

Peaks

Critical Points





Advantage

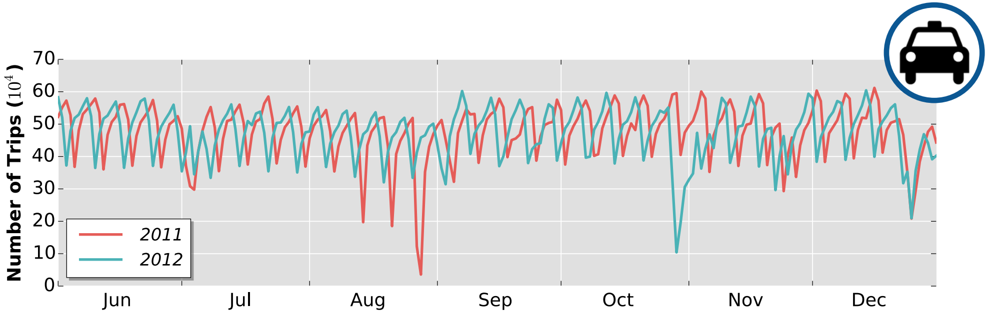
1. Naturally captures such features

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



NYU

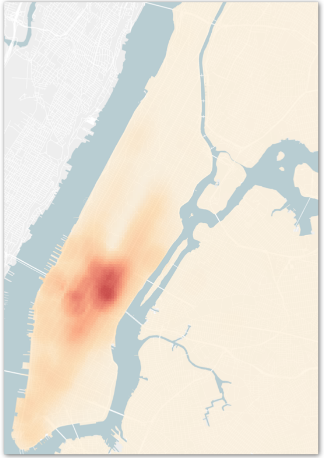
VIDA

Scalar Functions

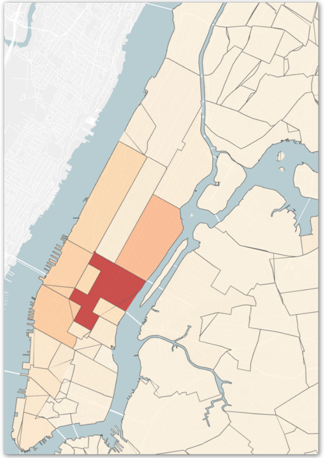
Each data set represented as a *time-varying scalar function*

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

Maps each point in the domain (city) over time to a scalar value



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

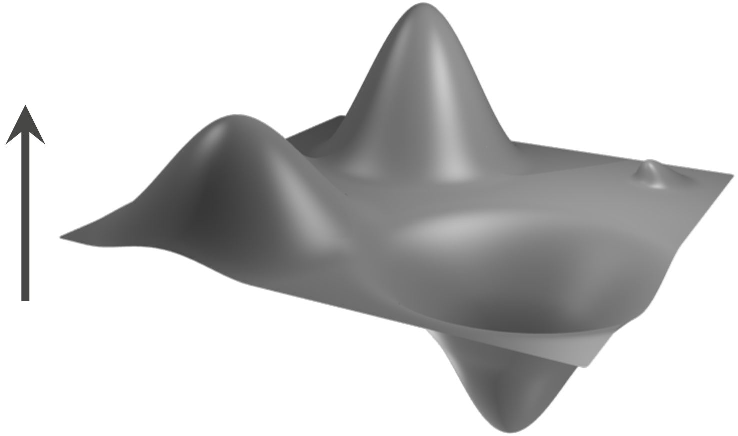


$\mathbb{S}$: *Neighborhood Resolution*

Identifying Topological Features

Topological features of the scalar function

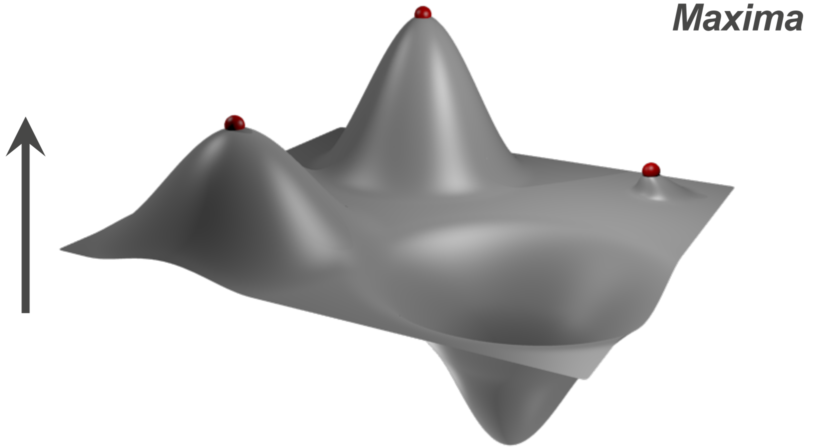
- Neighborhoods of critical points



Identifying Topological Features

Topological features of the scalar function

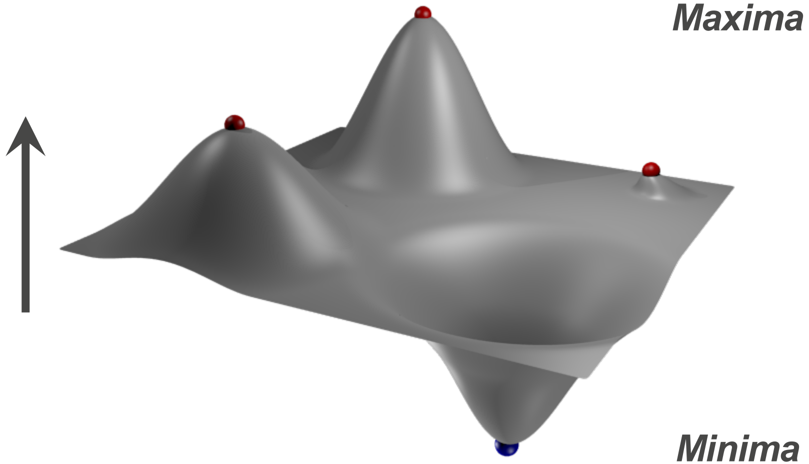
- Neighborhoods of critical points



Identifying Topological Features

Topological features of the scalar function

- Neighborhoods of critical points

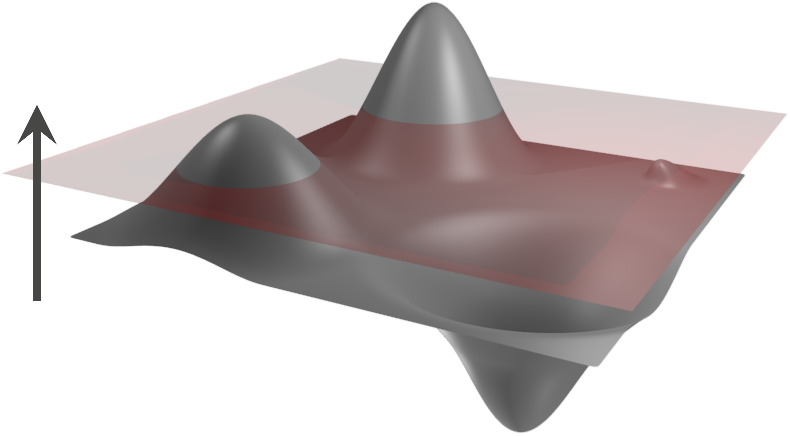


Identifying Topological Features

Topological features of the scalar function

Neighborhoods of critical points

Neighborhood defined by a threshold



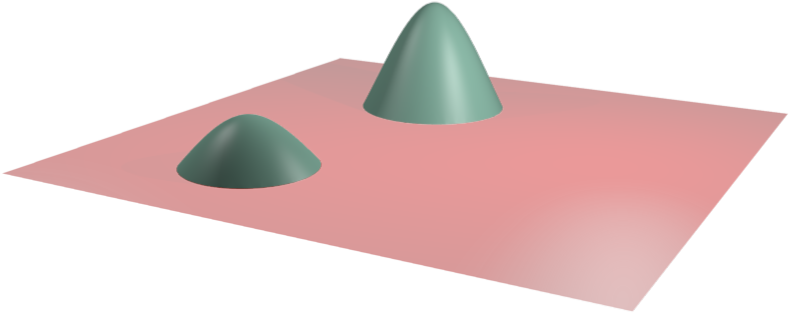
Identifying Topological Features

Topological features of the scalar function

- Neighborhoods of critical points

Neighborhood defined by a threshold

- Positive Features



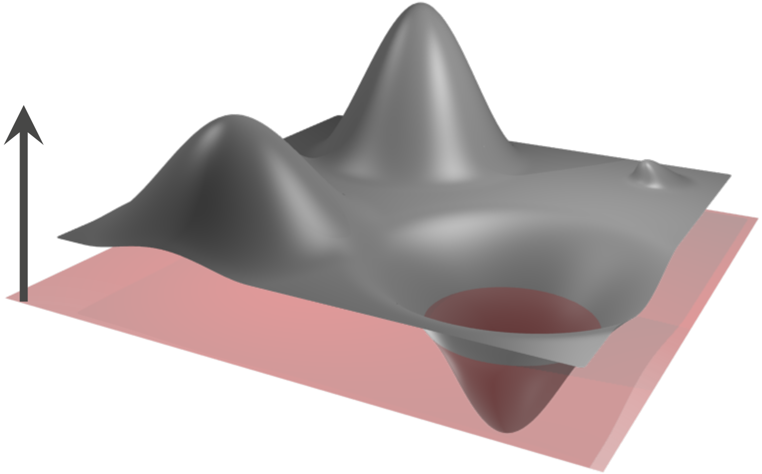
Identifying Topological Features

Topological features of the scalar function

- Neighborhoods of critical points

Neighborhood defined by a threshold

- Positive Features



Identifying Topological Features

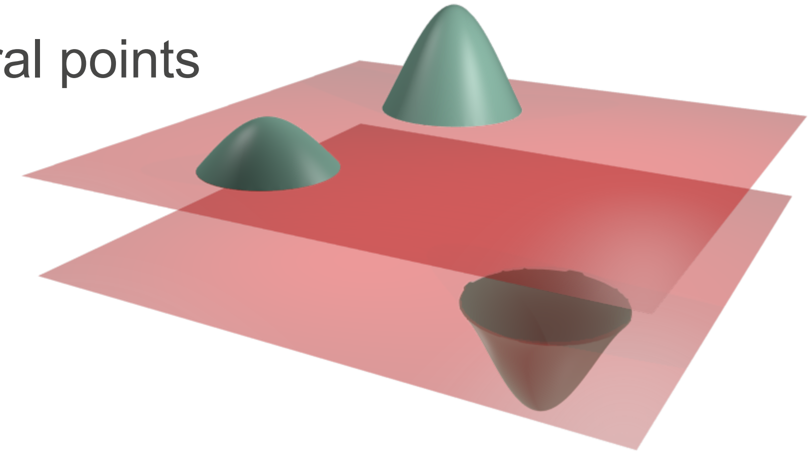
Topological features of the scalar function

- Neighborhoods of critical points

Neighborhood defined by a threshold

- Positive Features
- Negative Features

Represented as a set of spatio-temporal points



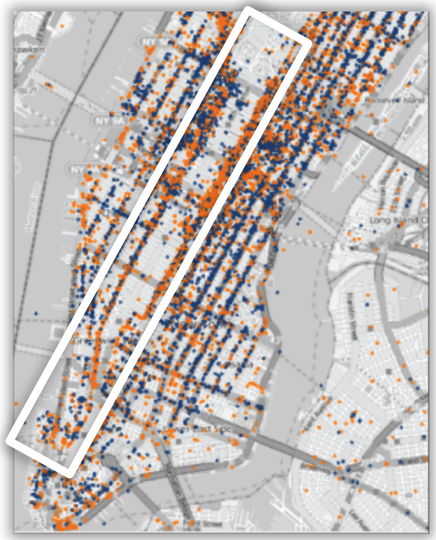
NYU

VIDA

Identifying Topological Features

8am - 9am
May 1 2011

5 Boro Bike Tour



Negative Feature

Advantage

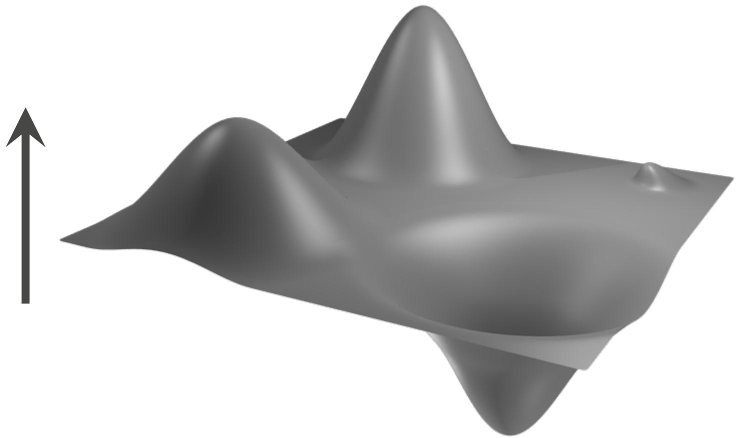
2. Features can have arbitrary shapes

NYU

VIDA

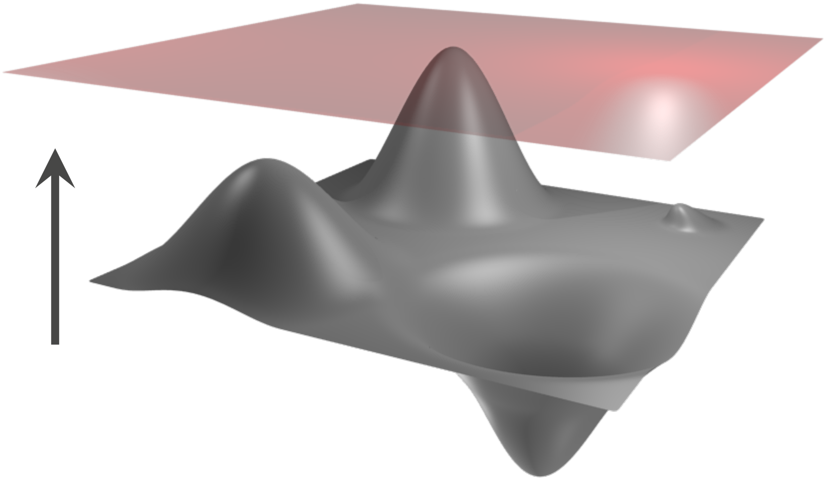
Computing Topological Features

Index: Merge Tree



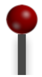
Computing Topological Features

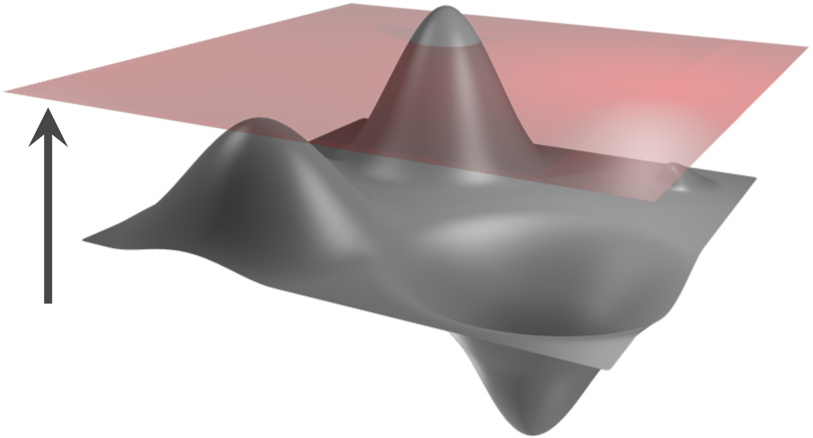
Index: Merge Tree



Computing Topological Features

Index: Merge Tree



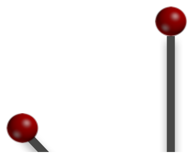


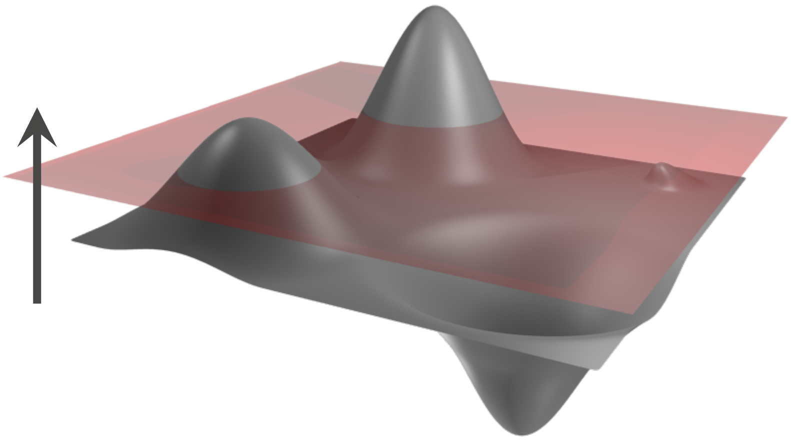
NYU

VIDA

Computing Topological Features

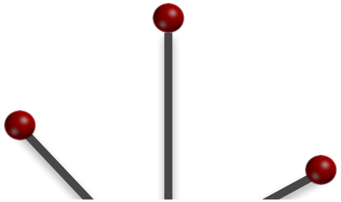
Index: Merge Tree

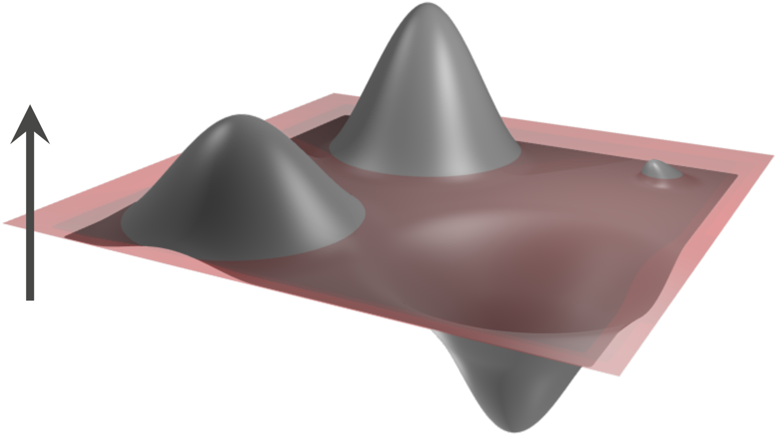




Computing Topological Features

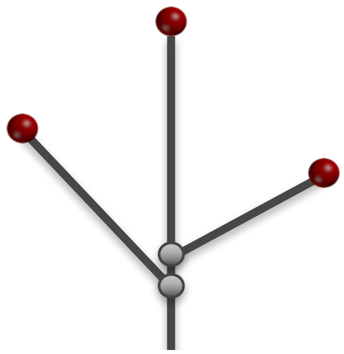
Index: Merge Tree

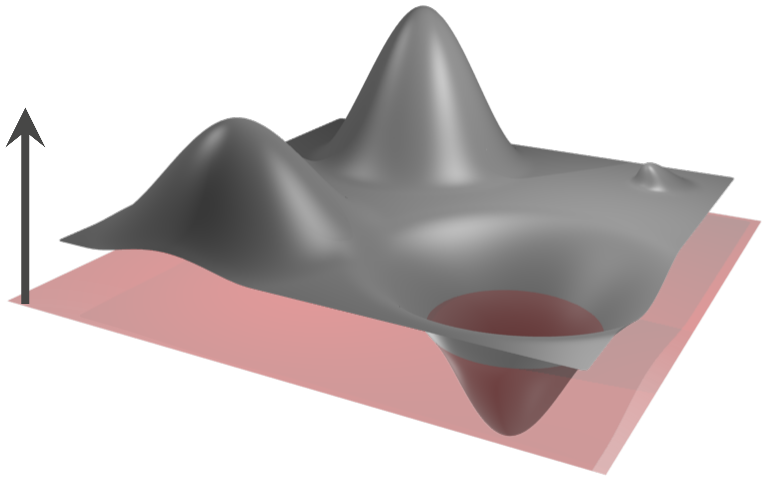




Computing Topological Features

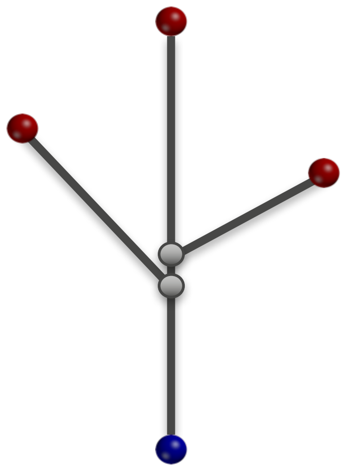
Index: Merge Tree

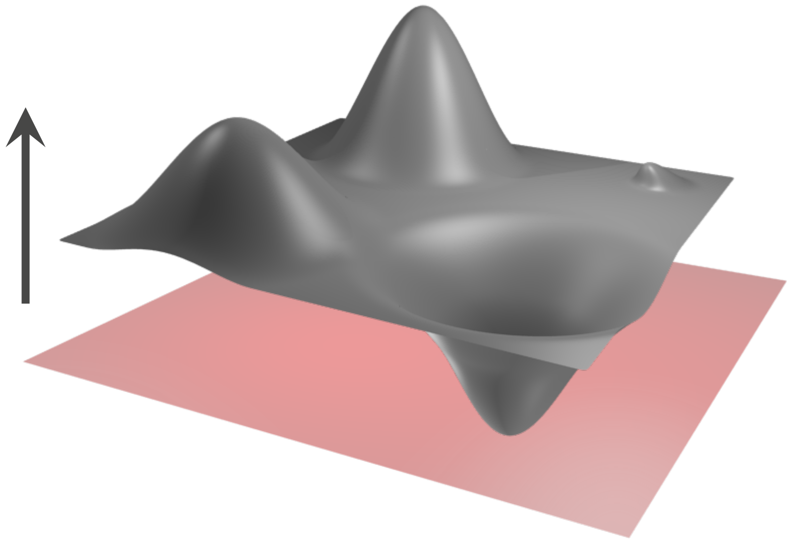




Computing Topological Features

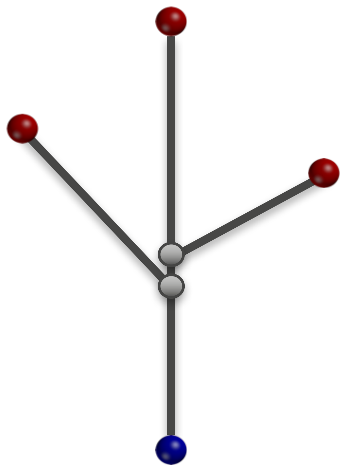
Index: Merge Tree

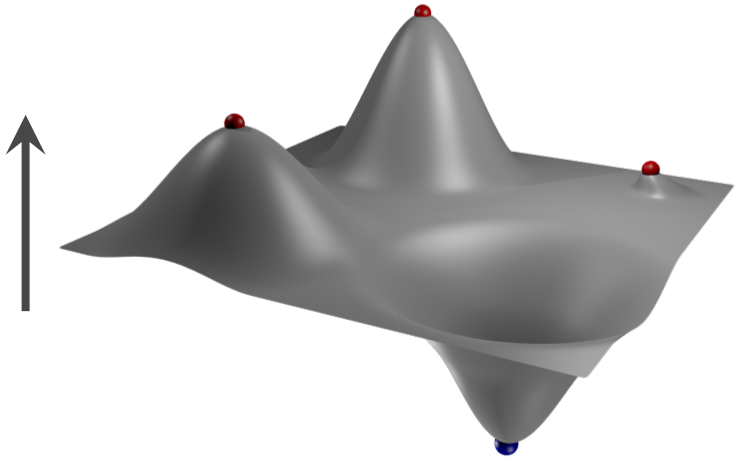




Computing Topological Features

Index: Merge Tree





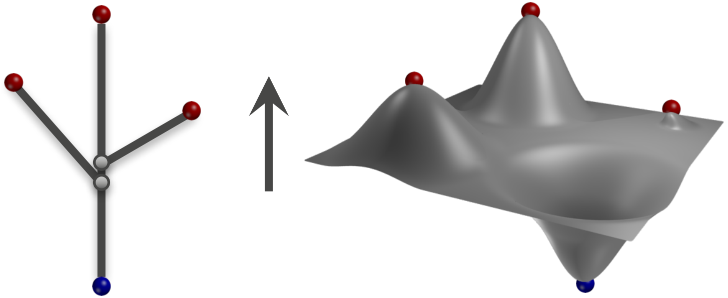
Computing Topological Features

Index: Merge Tree

Computing Merge Tree

$O(n \log n)$

Computing Features



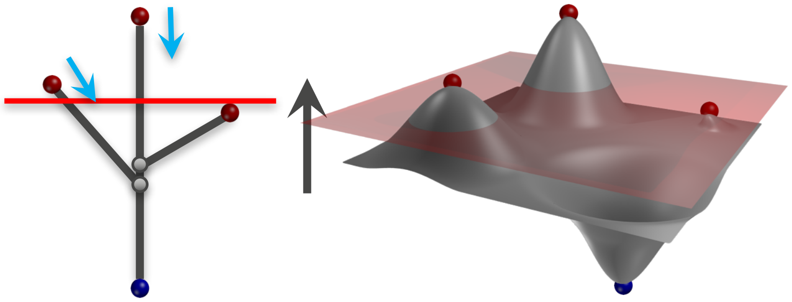
Computing Topological Features

Index: Merge Tree

Computing Merge Tree

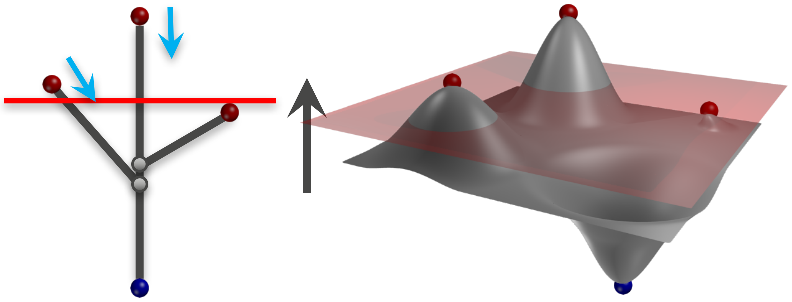
$O(n \log n)$

Computing Features



Computing Topological Features

Index: Merge Tree



Computing Merge Tree

O(n log n)

Computing Features

Output-sensitive time complexity

Advantage

3. Very efficient

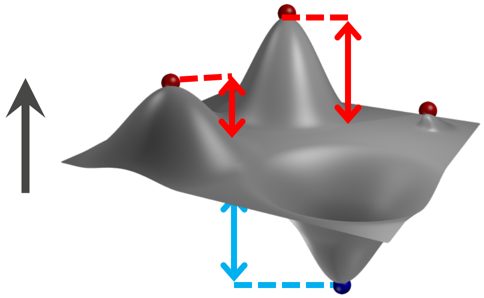
NYU

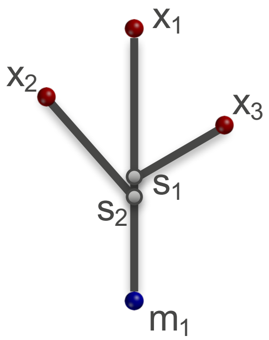
VIDA

Computing Feature Threshold

Feature thresholds are computed in a data-driven approach

- Uses topological persistence of the features
- "Life time" of the topological features
- Persistence can be efficiently computed using the merge tree





Computing Feature Threshold

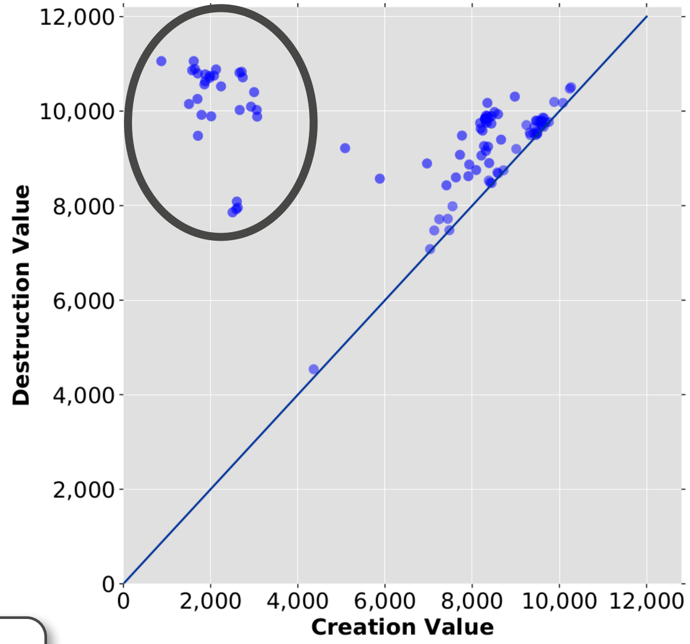
Use persistence diagram

Plots “birth” vs “death”

High persistent features form a separate cluster

2-means cluster

Use the high persistent cluster to compute the threshold

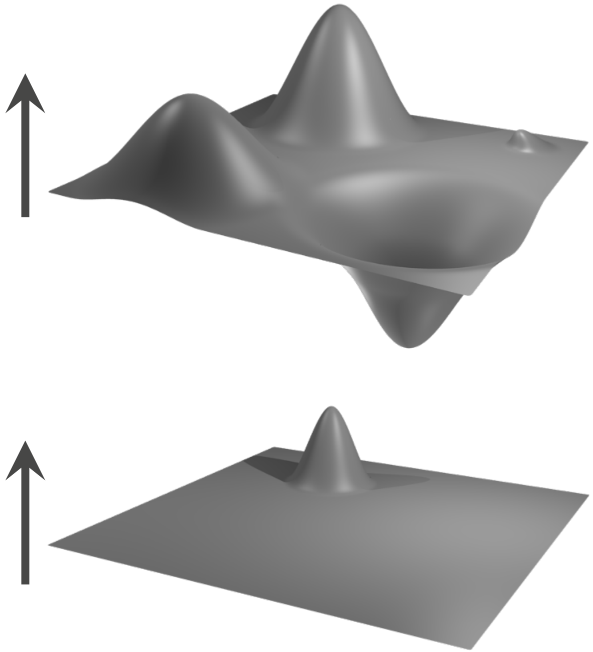


Advantage

4. Robust to noise

Relationship Evaluation

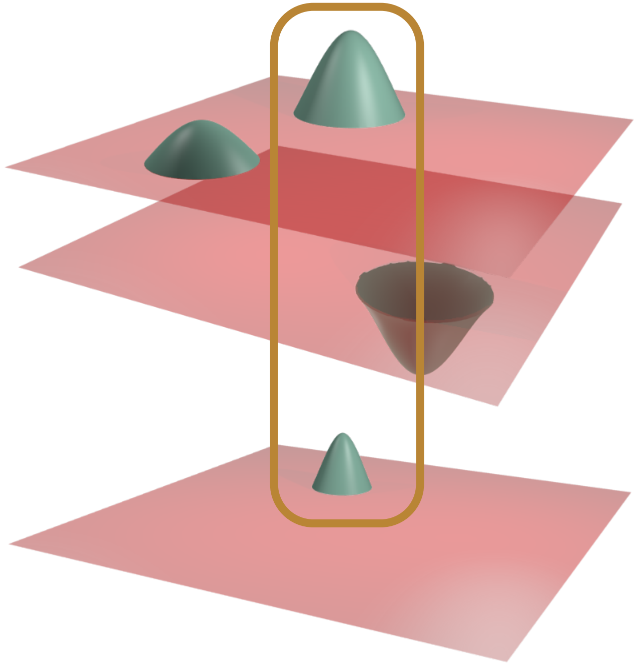
Relationship between features



Relationship Evaluation

Relationship between features

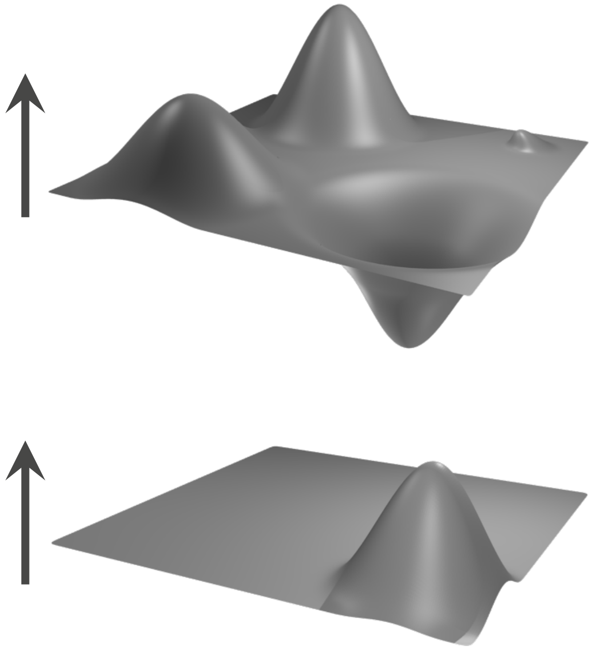
- Related features
- **Positive** Relationship



Relationship Evaluation

Relationship between features

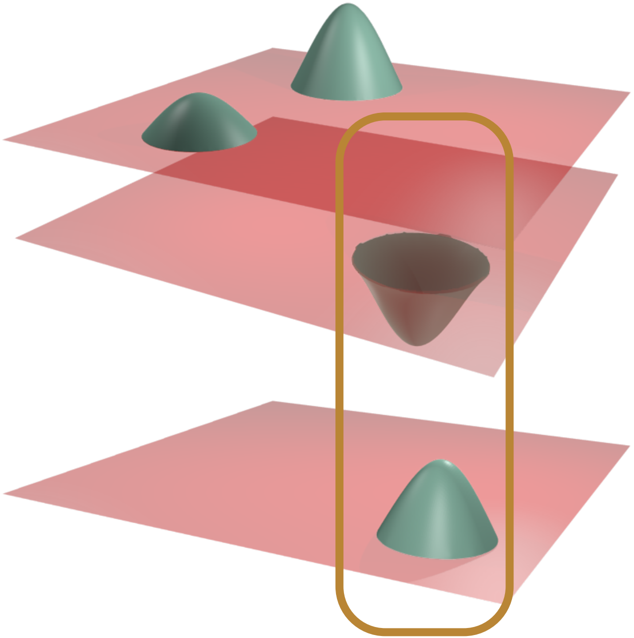
- Related features
- **Positive** Relationship



Relationship Evaluation

Relationship between features

- Related features
- **Positive** Relationship
- **Negative** Relationship



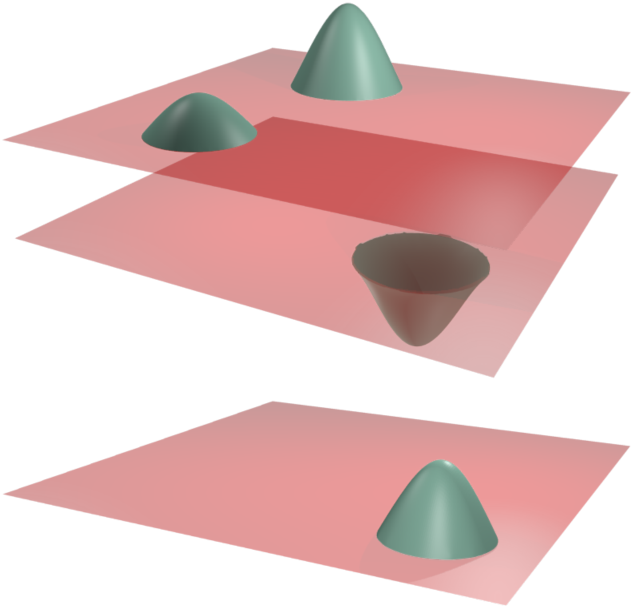
Relationship Evaluation

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

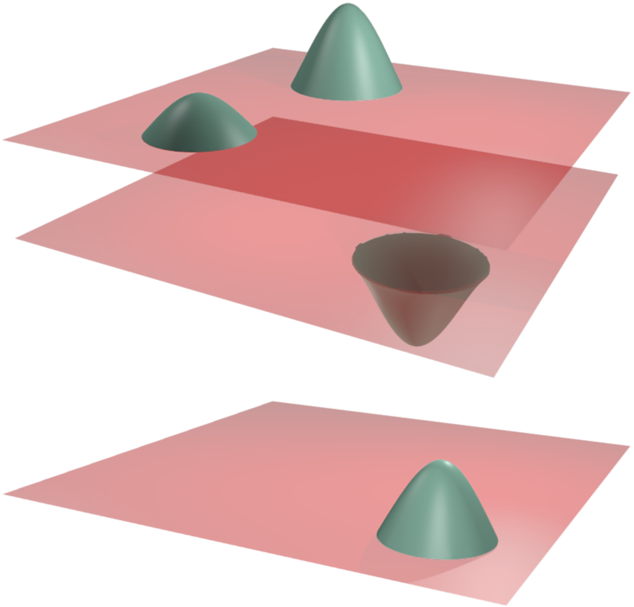
Relationship Score: τ

- 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

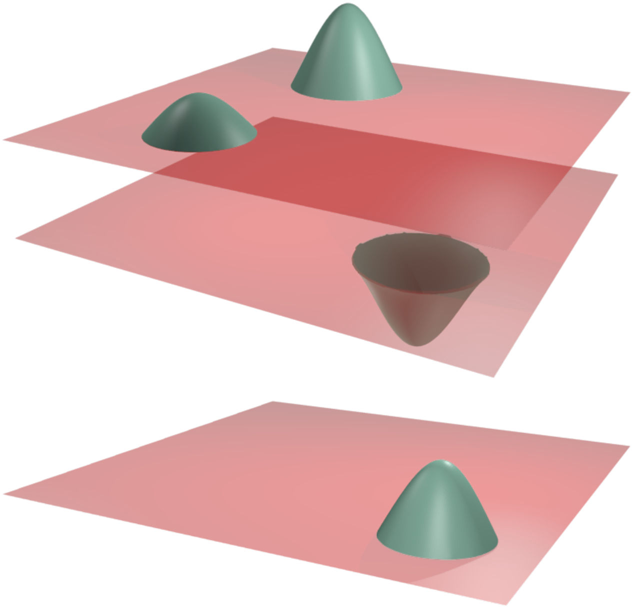
- How related the two functions are
- Captures the nature of the relationship

Relationship Strength: ρ

- How often the functions are related

Significant relationships

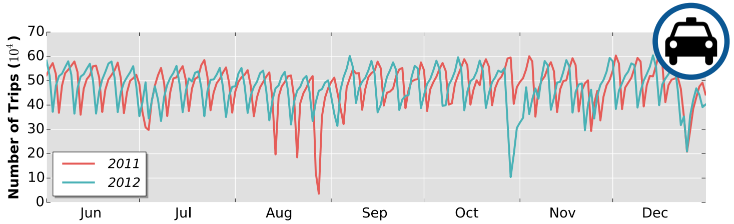
- Monte Carlo tests filter potentially coincidental relationships

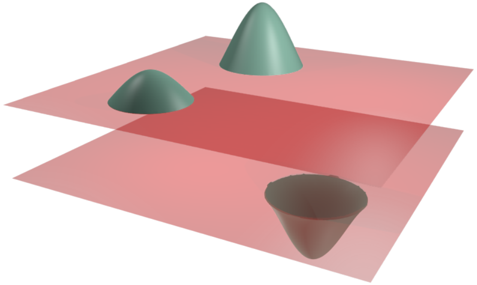


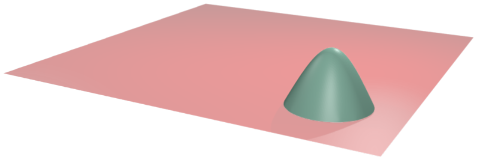
NYU

VIDA

Topology-based Relationships







Advantage

5. Works on data in any dimension

Topology: Advantages

1. Naturally captures interesting features

2. Features can have arbitrary shapes

3. Very efficient

4. Robust to noise

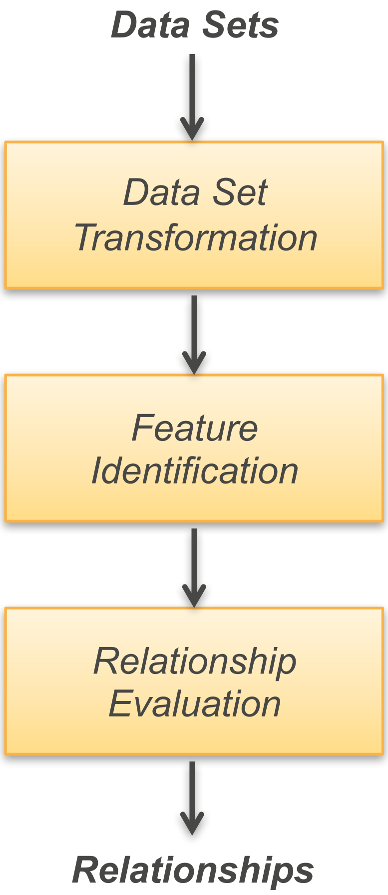
5. Works on data in any dimension

The Data Polygamy Framework

Implementation

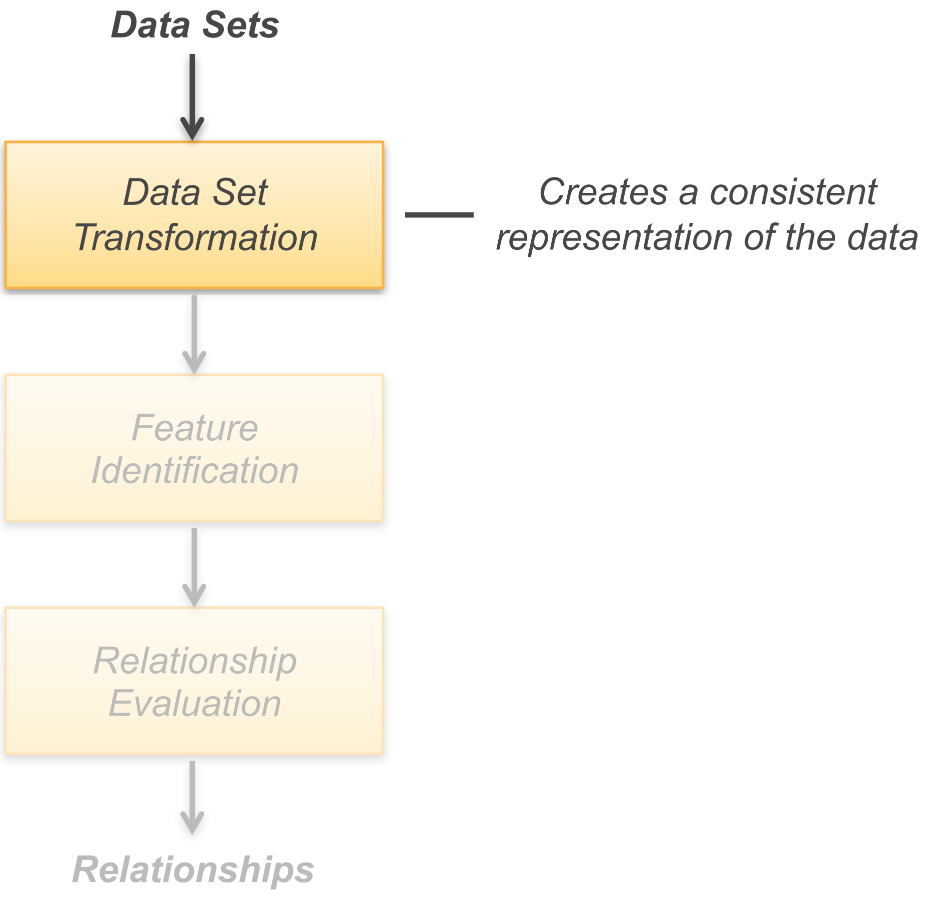
All the steps are embarrassingly parallel

Framework implemented using *map-reduce*


Data Sets
Data Set Transformation
Feature Identification
Relationship Evaluation
Relationships

NYU

VIDA


Data Sets
Data Set Transformation
Creates a consistent representation of the data
Feature Identification
Relationship Evaluation
Relationships

Scalar Functions

Two types of scalar functions: *count* and *attribute*

Count functions

E.g.: no. of taxi trips over space and time
no. of *unique* taxis over space and time

Attribute functions

E.g.: average taxi fare over space and time

Other functions can also be added!

E.g.: gradient function

Functions are computed at all possible resolutions

E.g.: data available in (GPS, seconds) can be translated to
[grid, neighborhood, city] x [hour, day, week, month]

Data Sets

Data Set Transformation

Feature Identification

Relationship Evaluation

Relationships

Map: data points mapped to different resolutions

Reduce: data is aggregated for each resolution

Data Sets

Data Set Transformation

Feature Identification

Relationship Evaluation

Relationships

Computes the set of prominent topological features

NYU

VIDA

Data Sets

↓

Data Set Transformation

↓

Feature Identification

↓

Relationship Evaluation

↓

Relationships

$<K, [A_1, A_2, A_3, A_4, \dots]>$

↓

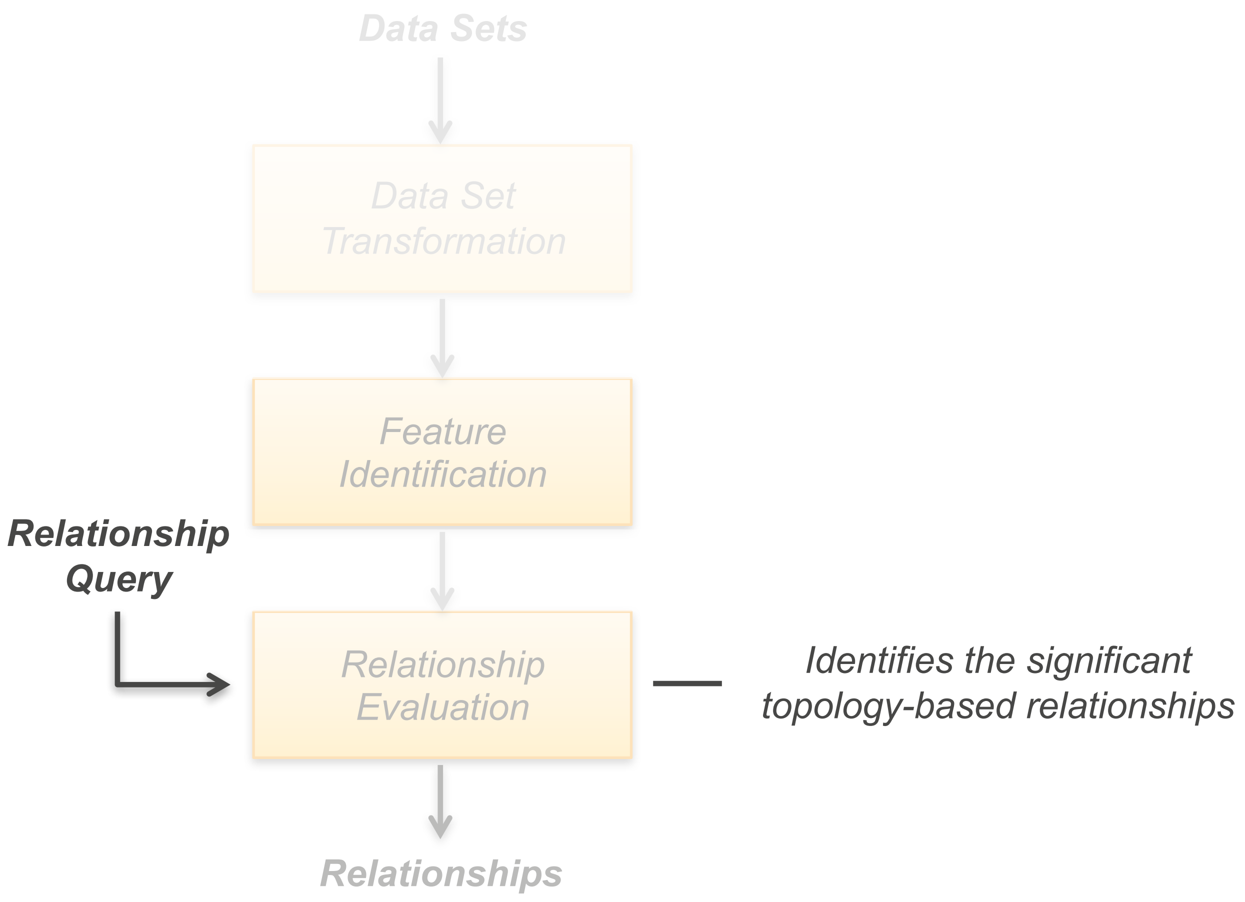
$<K, A_1> \quad <K, A_2> \quad <K, A_3> \quad <K, A_4> \quad \dots$

Map: for each data set, scalar functions are taken individually

Reduce: for each function, the merge tree index is created and the features are identified for all resolutions

↓

Merge trees are saved!


Data Sets
Data Set Transformation
Feature Identification
Relationship Query
Relationship Evaluation
Identifies the significant topology-based relationships
Relationships

NYU

VIDA

Relationship Querying

Querying for relationships

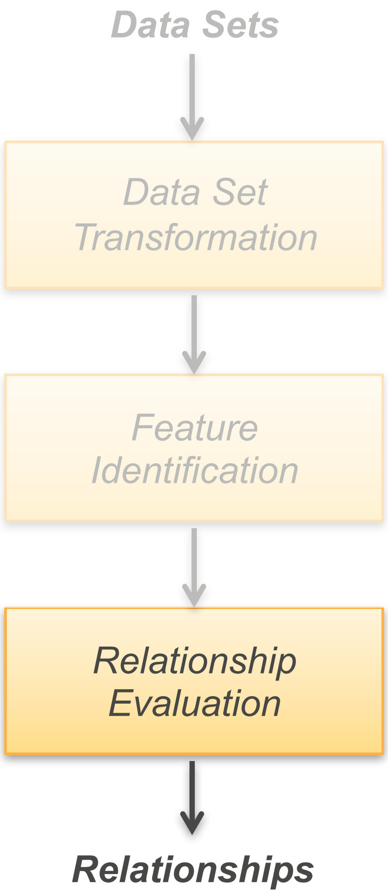
Find all data sets related to data set $\boldsymbol{D}$ satisfying **CLAUSE**

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*



Map: all possible combinations are considered, based on the query

Reduce: each relationship is evaluated

Additional Information

Software Framework: Apache Hadoop 2.2.0

Distributed File System: HDFS

Compression (BZip2 or Snappy Codec) can be used for map outputs

Framework runs on AWS

Performance Evaluation

Goal

Efficiency, scalability, robustness

Data

NYC Open Data: 300 spatio-temporal data sets

Hardware

20 compute nodes, AMD Opteron(TM) Processor 6272 (4x16 cores) running at 2.1GHz, 256GB of RAM – *for most experiments*

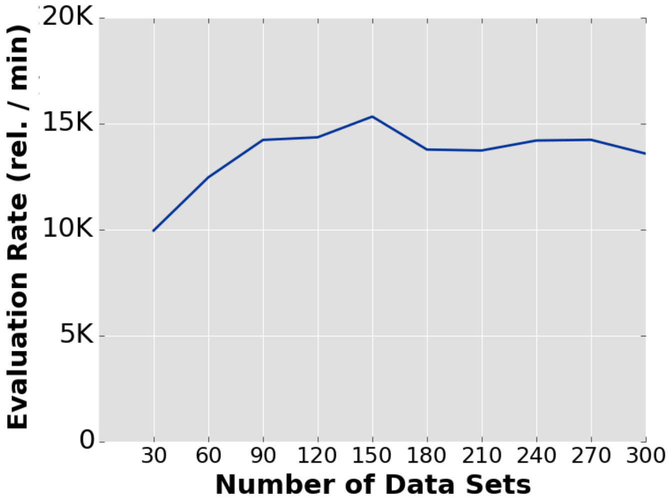
Amazon EMR: m1.medium (for master) and r3.2xlarge (for slaves) – *for scalability tests*

Performance Evaluation: Results

200 mins to compute scalar functions and features for NYC Open Data

Using significance tests: decrease of around **99%** on the number of output relationships!

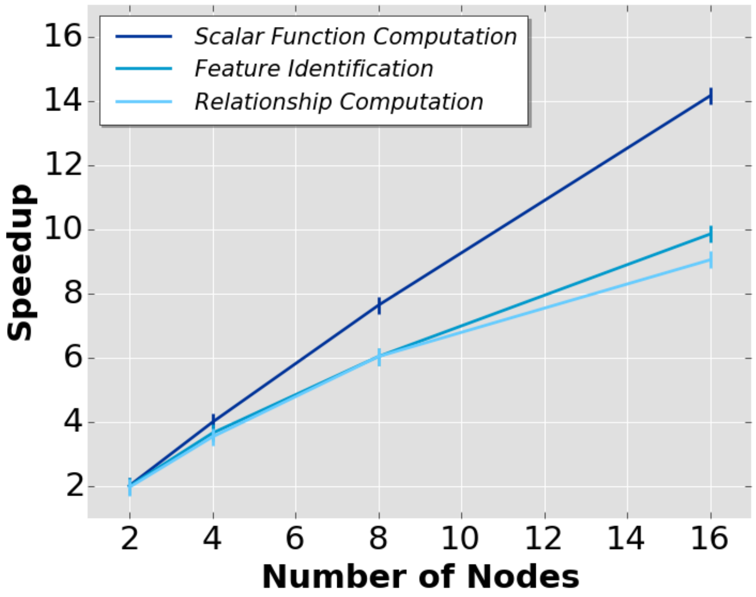
Can evaluate **> 10K relationships/min**



Performance Evaluation: Results

Approach is **robust** to noise

Approach is **scalable**



Qualitative Evaluation

Goal

Does the approach uncover *interesting*, *non-trivial* relationships?

Data

NYC Urban: 9 data sets from NYC agencies

(Some) Interesting Relationships

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

Find all relationships between Collisions and Traffic Speed data sets

 # Collisions **X** *Traffic Speed*

Positive relationship between number of collisions and spee

 # Deaths **X** *Traffic Speed*

Positive relationship between number of persons killed and

Daily Intelligencer

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

By Caroline Bankoff   Follow @teamcaroline

181063216 Photo: Getty Images

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

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,

(Some) Interesting Relationships

2. Why it is so hard to find a taxi when it is raining?

Find all relationships between Taxi and Weather data sets

 # Taxi X *Precipitation*

Negative relationship between number of taxis and average precipitation

Hypothesis: Taxi drivers are target earners

Taxi Fare X *Precipitation* → *Indicates that hypothesis is true*

Strong positive relationship between precipitation and average fare

A recent study[1] refuted this hypothesis

[1] H. S. Farber. Why You Can' t Find a Taxi in the Rain and Other Labor Supply Lessons from Cab Drivers. Technical Report 20604, National Bureau of Economic Research, 2014.

(Some) Interesting Relationships

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

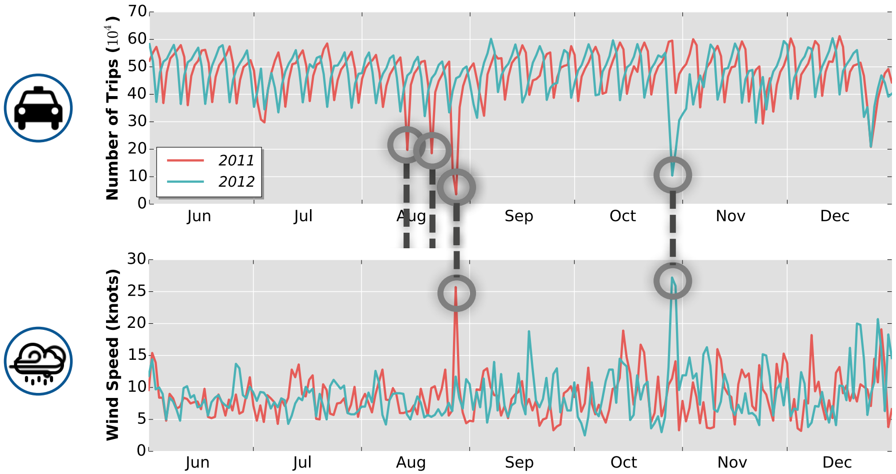
   Find all data sets related to the Taxi data set

 # Taxi *Wind Speed*

Negative relationship between number of taxis and wind speed

 # Taxi *Precipitation*

Negative relationship between number of taxis and average precipitation



(Some) Interesting Relationships

Citi Bike and Weather

 # Citi Bike stations **X** *Snow*

Negative relationship between snow precipitation and active Citi Bike stations

(day, city) ✓

(hour, city) ⃠

NYU VIDA

Many other relationships…

~ 100 significant relationships per resolution

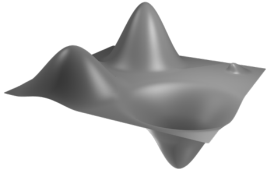
Over 35 interesting relationships

More relationships (and their implications) can be understood by having domain experts

Weather data set is the most *polygamous*!

Conclusion

Data Polygamy – discover and explore relationships in large collections of data sets



Relationships are based on the topology of the data

- Relationships between salient features
- Take into account both time and space

Framework implemented using *map-reduce*

- Efficient and scalable
- Interesting (and surprising!) relationships could be found

Querying for relationships is just the beginning…

Lessons Learned

It's hard to evaluate!

- No ground truth available
- Need benchmark
- Need real use cases from domain experts

Too many relationships!

- How to explore and analyze them?

I ❤❤❤❤ DATA POLYGAMY

Code, data, and experiments available at:

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

"Data Polygamy: The Many-Many Relationships among Urban Spatio-Temporal Data Sets", F. Chirigati, H. Doraiswamy, T. Damoulas, and J. Freire. In Proceedings of the 2016 ACM SIGMOD International Conference on Management of Data (**SIGMOD**), pp. 1011-1025, 2016

NYU

VIDA

AWS re:Invent

Thank you!

Questions?

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

amazon web services