Towards Usability, Transparency, and Trust in Data-Driven Exploration

Juliana Freire
Visualization, Imaging and Data Analysis Center (VIDA)
Computer Science & Engineering
Center for Data Science (CDS)

Joint work with A. Ailamaki, A. Bessa, F. Chirigati, Y. Chan, T. Damoulas, H. Doraiswamy, N. Ferreira, M. Lage, R. Lourenço, F. Miranda, J. Poco, D. Shasha, C. Silva, D. Srivastava, E. Tzirita, H. Vo

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Data-Driven Exploration

- Every scientific domain is moving toward data-driven exploration, this has led to great advances and discoveries
- Companies are capitalizing on data
- Government agencies uses data to operate efficiently, make policies, and informed decisions

Computing is free

Storage is free

Data are abundant

The bottlenecks lie with people

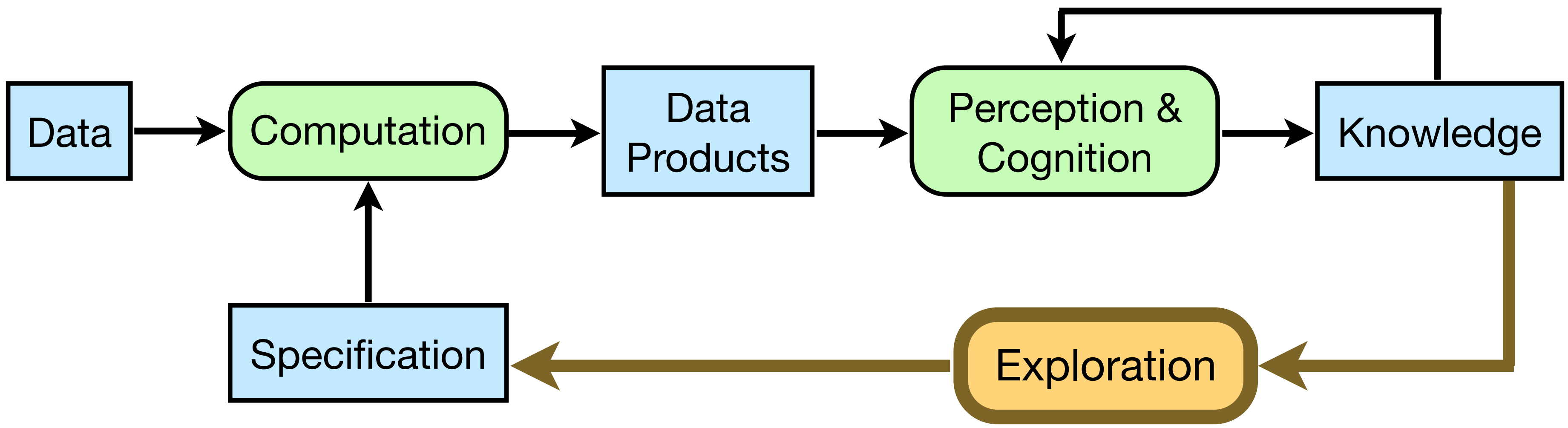
Data-Driven Exploration: Challenges

- Data are vast and produced at unprecedented rates
  - Sources are broad, varied, and unreliable
- Computational processes are required to extract insight
  - But they hard to assemble

provenance
machine learning
algorithms
data integration
data discovery
interaction modes
visual encodings
statistics
data curation
data management
math

Data-Driven Exploration: Challenges

- Data are vast and produced at unprecedented rates
  - Sources are broad, varied, and unreliable
- Computational processes are required to extract insight
  - But they hard to assemble
- Exploratory tasks are inherently iterative as one tests and formulates hypotheses



[Modified from Van Wijk, Vis 2005]

Data-Driven Exploration: Challenges

- After many steps…

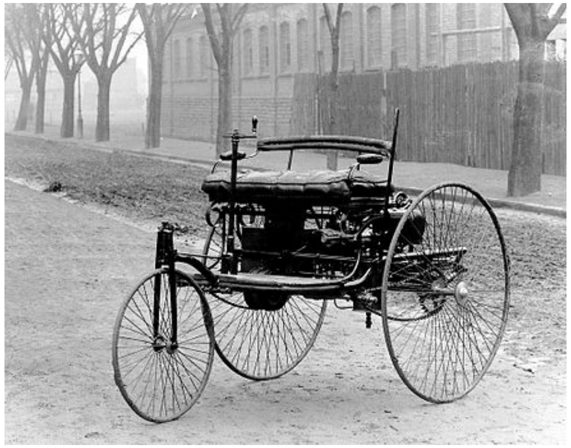
"An analysis has 30 different steps. It is tempting to just do this then that and then this. You have no idea in which ways you are wrong and what data is wrong" [Kandel et al., VAST 2012]

- It is easy to get lost and not remember how a result was derived
- Processes can break or misbehave in unforeseen ways
- Results can be hard to understand, interpret and trust

Incorrect conclusions can lead to bad decisions!

An Analogy: Cars

Data-Driven Exploration: Goal



?

Grand challenge for data science and engineering:

Empower a wide range of users to explore and obtain trustworthy, actionable insights from data.

Talk Outline

- Interactive exploration of spatio-temporal urban data
- Using data to explain and discover data
- Open problems for database research

Usability in data exploration

Guiding users and building trust

Urban Data

- Cities are the loci of economic activity
- 50% of the world population lives in cities, by 2050 the number will grow to 70%
- (… ent
  a
- (… n the
  \
Ir

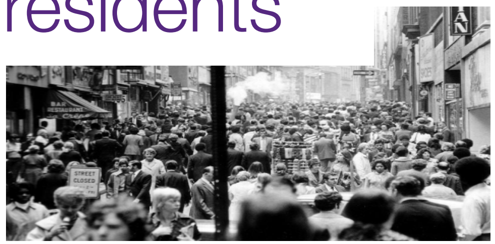
Opportunity:

Analyze the data exhaust to understand how different components interact over space and item

Use these insights to make cities more efficient and sustainable, and improve the lives of their residents

Condition, operations

Meteorology, pollution, noise, flora, fauna



Relationships, economic activities, health, nutrition, opinions, …

Urban Data: Success Stories

- Real-time bus arrival pred
  - 94% reported increa ... ly increased satisfaction with p ... ansit
- Illegal conversions in NYC
  - DS team (1) integrated data ... erent agencies that provided ... of issues in buildings; (2) cr ... ire data; (3) Created a predictio
  - Hit r ... nspections went from 13% to 70%
- Foreclosures and crime
  - Neighborhoods with concentr ... oreclosures see an uptick in crime f ... closure notice issued
  - NYPD updated its policing strategies

Benefit residents

Make cities more efficient

Impact policy

OneBusAway

Serving up fresh real-time transit information for the region.

Todd Heisler/The New York Times

Michael Flowers, right, oversees a small group of tech-savvy and civic-minded statisticians working across from City Hall.

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Urban Data: What is hard?

Infrastructure

Condition, operations

Environment

Meteorology, pollution, noise, flora, fauna

People

Relationships, economic activities, health, nutrition, opinions, …

- City components interact in complex ways
- Need to explore the city *data exhaust* to understand these interactions

DATA.GOV EMPOWERING PEOPLE

NYC OpenData

twitter

facebook

YouTube Broadcast Yourself

data.gov.in Open Government Data Platform India सत्यमेव जयते

BRASIL Acesso à informação Dados Abertos GOVERNO FEDERAL

Urban Data Analysis: Common Practice

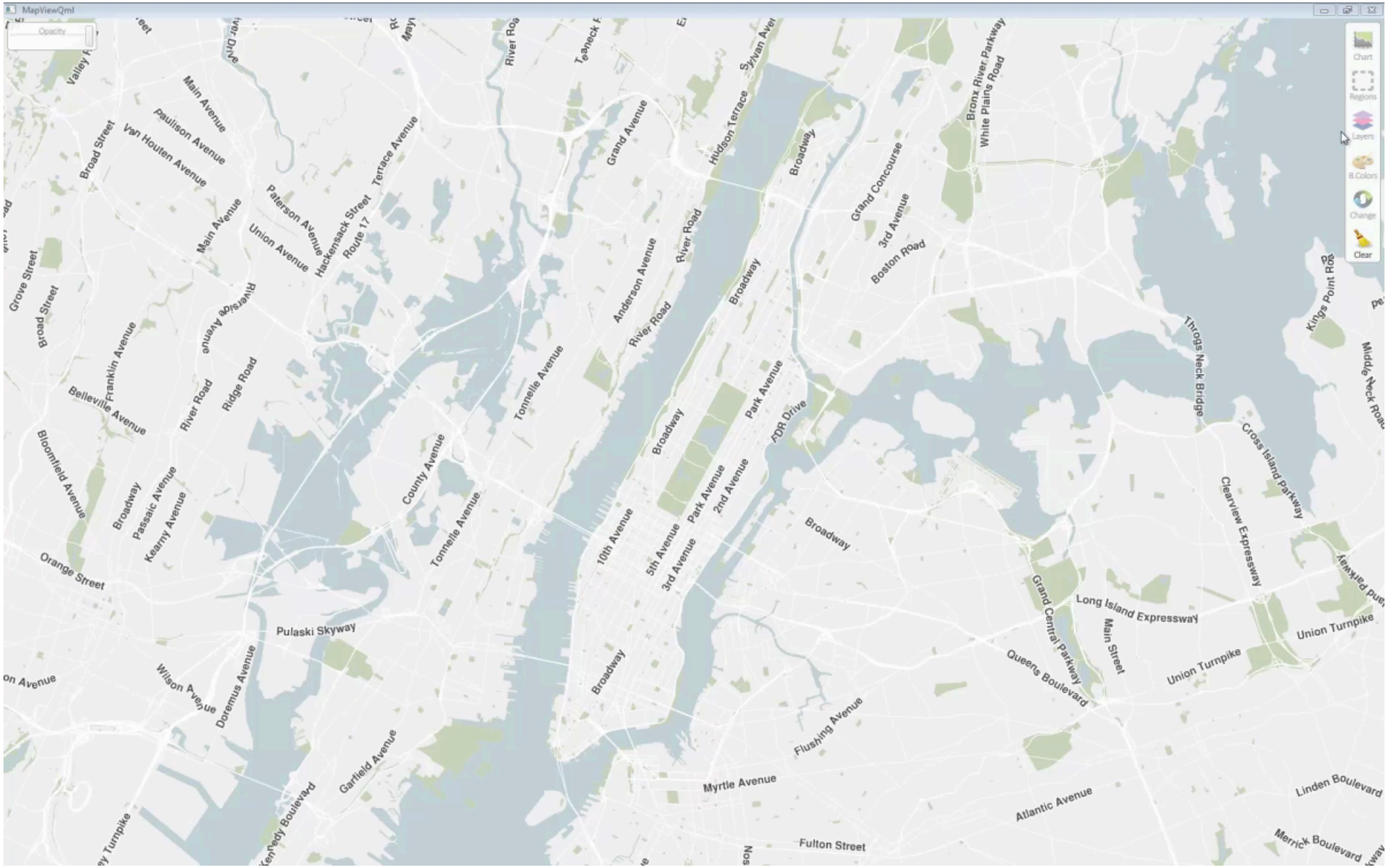
- Domain experts formulate hypotheses
- Data scientists select data sets and slices, perform analyses, and derive plots
- Domain experts examine the plots
- Issues:
  - Analyses are mostly confirmatory (Tukey, 1977) – batch-oriented analysis hampers exploration
  - Dependency on data specialists distances domain experts from the data
  - Data are noisy and complex – often multivariate spatio-temporal
  - Queries are expensive: widely-used tools are not scalable, e.g., Excel, GIS, SAS, …

Need scalable tools and techniques that help domains experts *interactively* explore data

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

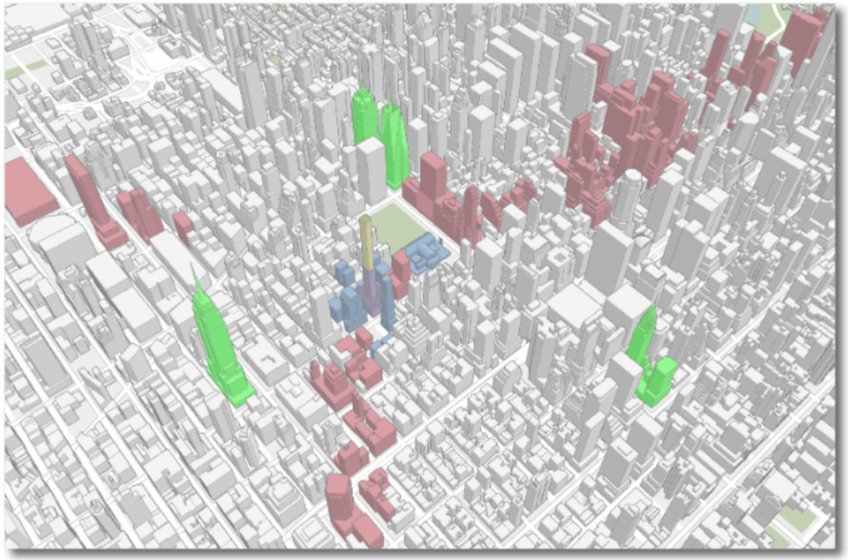
Urbane: Exploring Urban Data



https://www.youtube.com/watch?v=_B35vxCgDw4&feature=youtu.be

[Ferreira et al., IEEE VAST 2015;
Doraiswamy et al., ACM SIGMOD 2018]

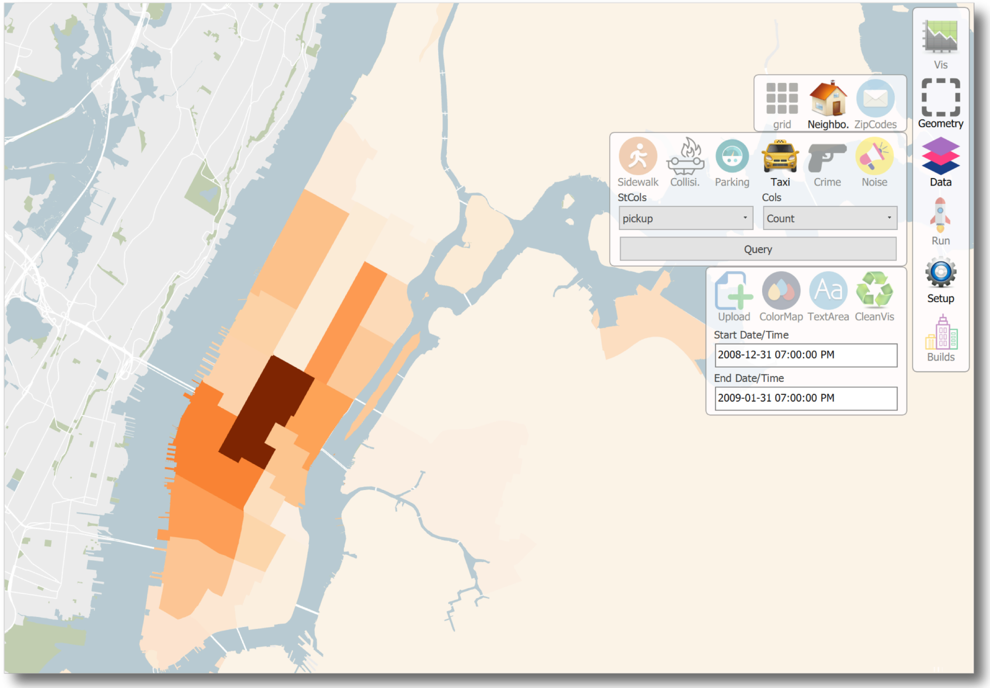
Usability through Visual 3D Queries



View Impact Queries

Sky Exposure Queries

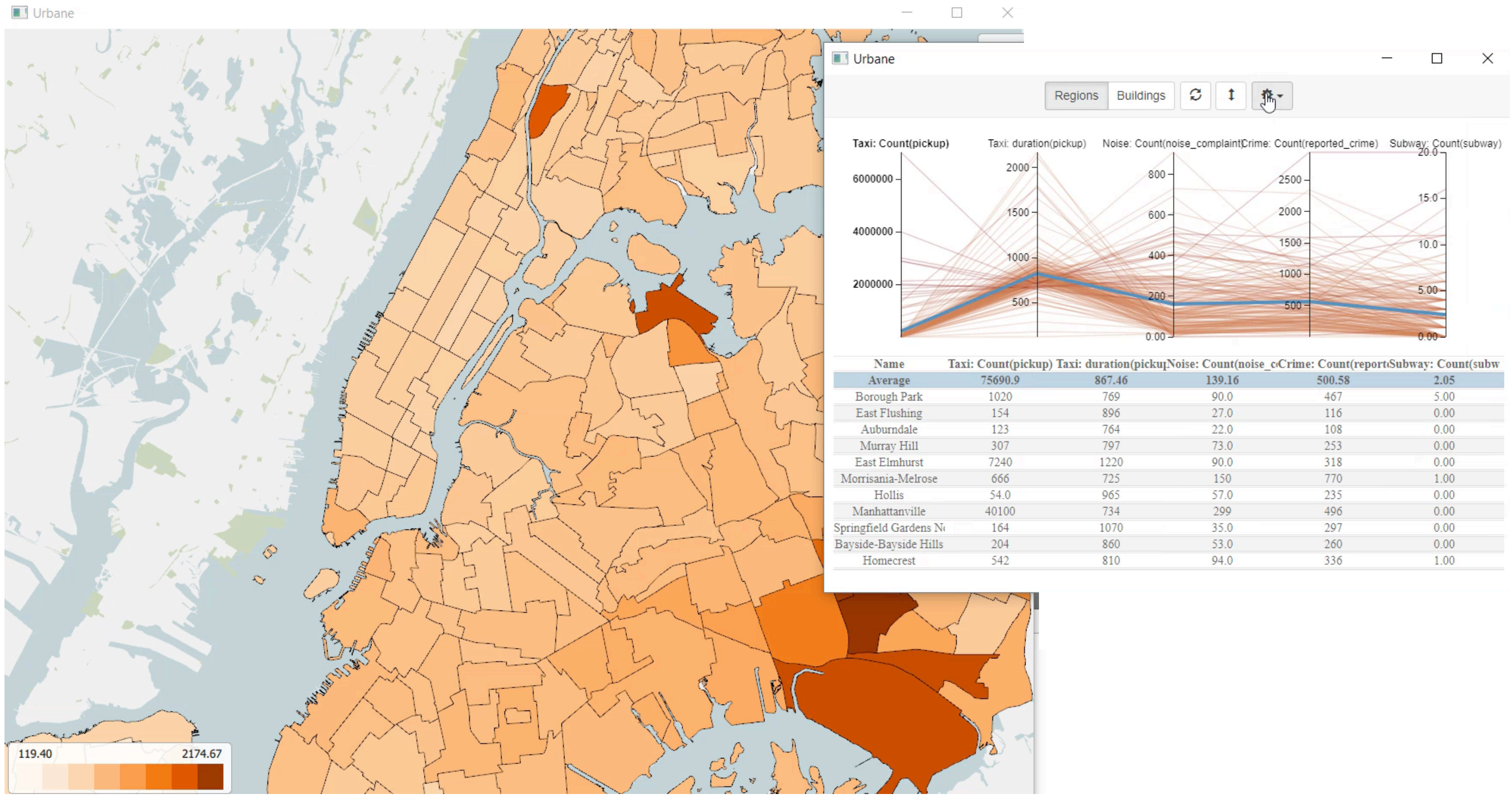
Usability through Visual 2D Queries



SELECT COUNT(*)
FROM taxi *T*, neighborhoods *N*
WHERE *T*.pickup INSIDE *N*.geometry
AND *T*.picktime > 2008-12-31
AND *T*.picktime < 2009-01-31
GROUP BY *N*.id

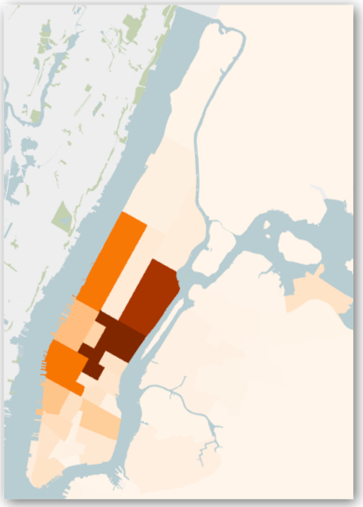
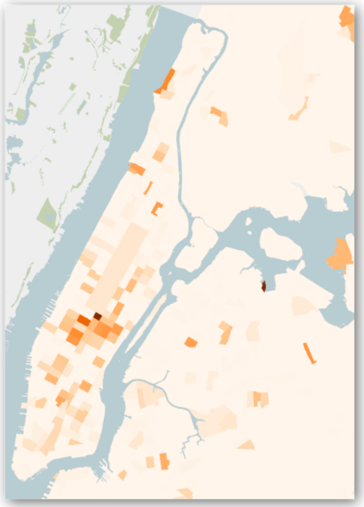
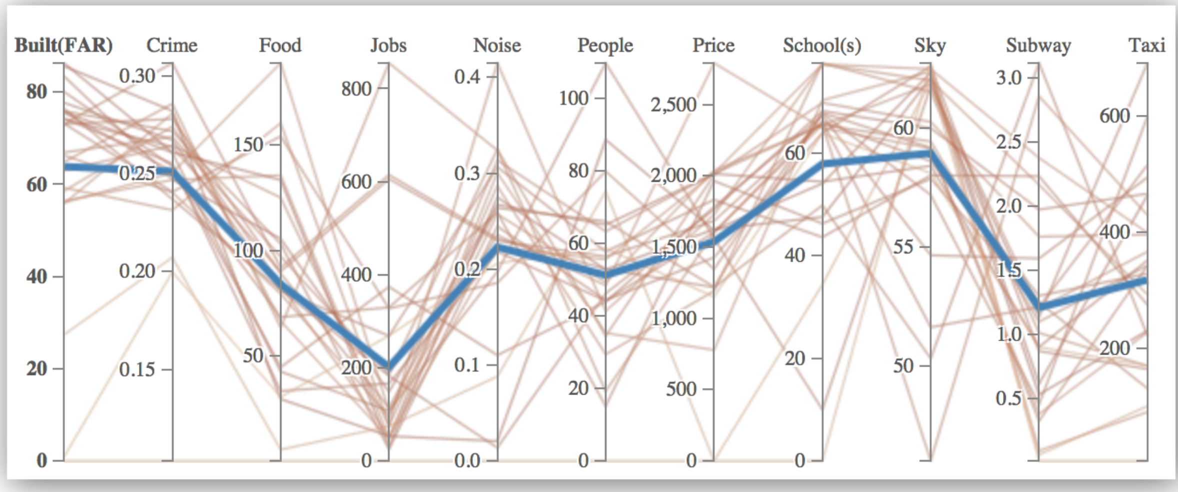
Challenge: Interactive Query Evaluation

"increased latency reduces the rate at which users make observations, draw generalizations and generate hypotheses" [Liu and Heer, IEEE TVCG 2014]



High query rate

Challenge: Spatial Aggregation

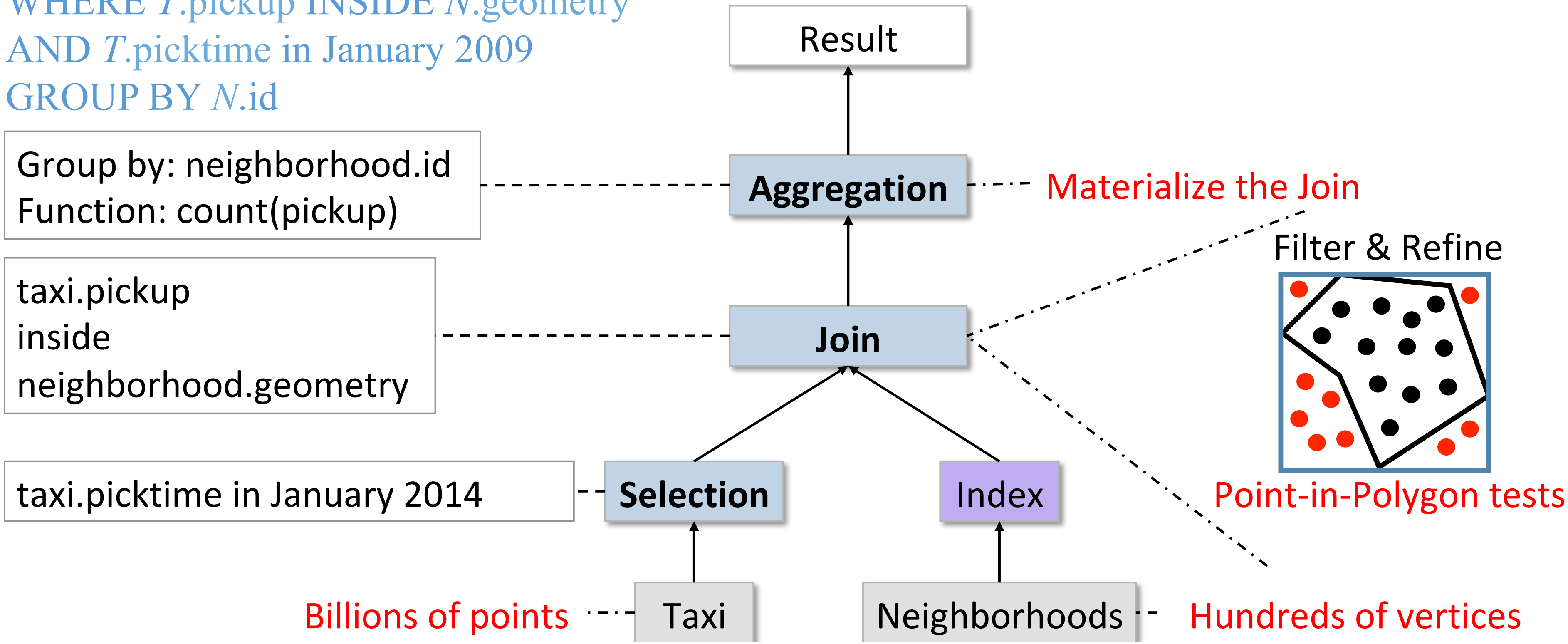
SELECT COUNT(*)
FROM taxi *T*, neighborhoods *N*
WHERE *T*.pickup INSIDE *N*.geometry
AND *T*.picktime in January 2009
GROUP BY *N*.id

SELECT COUNT(*)
FROM crime *C*, neig
WHERE *C*.location I
N.geometry
AND *C*.date in Janua
GROUP BY *N*.id

Food
Jobs
Noise
People
Price
Schools
Sky
…

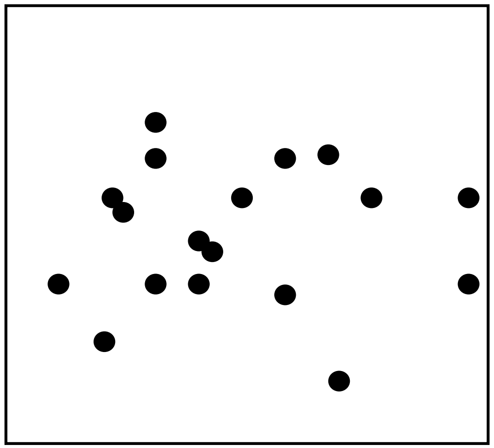
Spatial Aggregation

SELECT COUNT(*)
FROM taxi *T*, neighborhoods *N*
WHERE *T*.pickup INSIDE *N*.geometry
AND *T*.picktime in January 2009
GROUP BY *N*.id

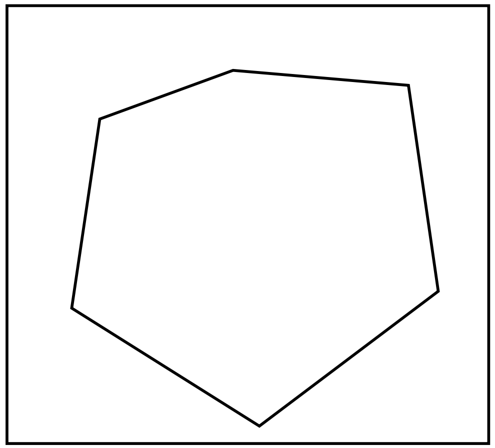


Spatial Aggregation: A Geometric Perspective

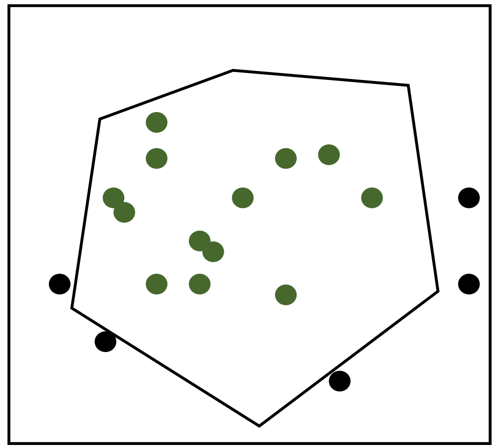
Spatial join = “Drawing” points and polygons on the same canvas



Input points



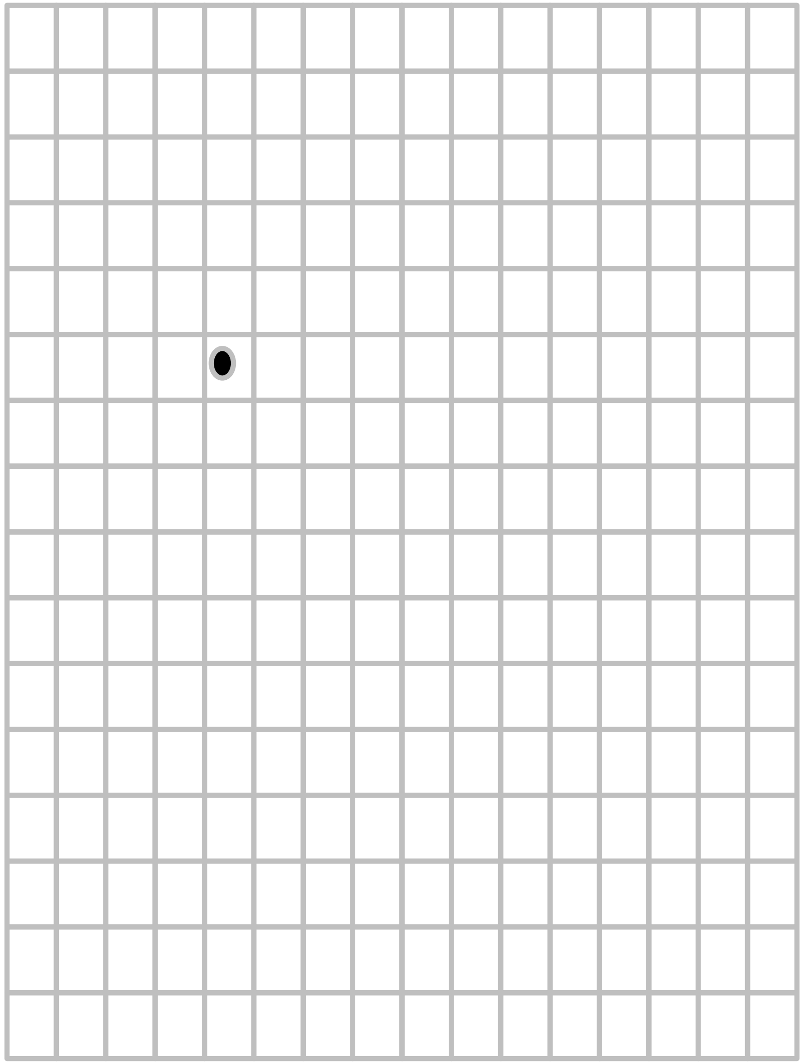
Input polygon



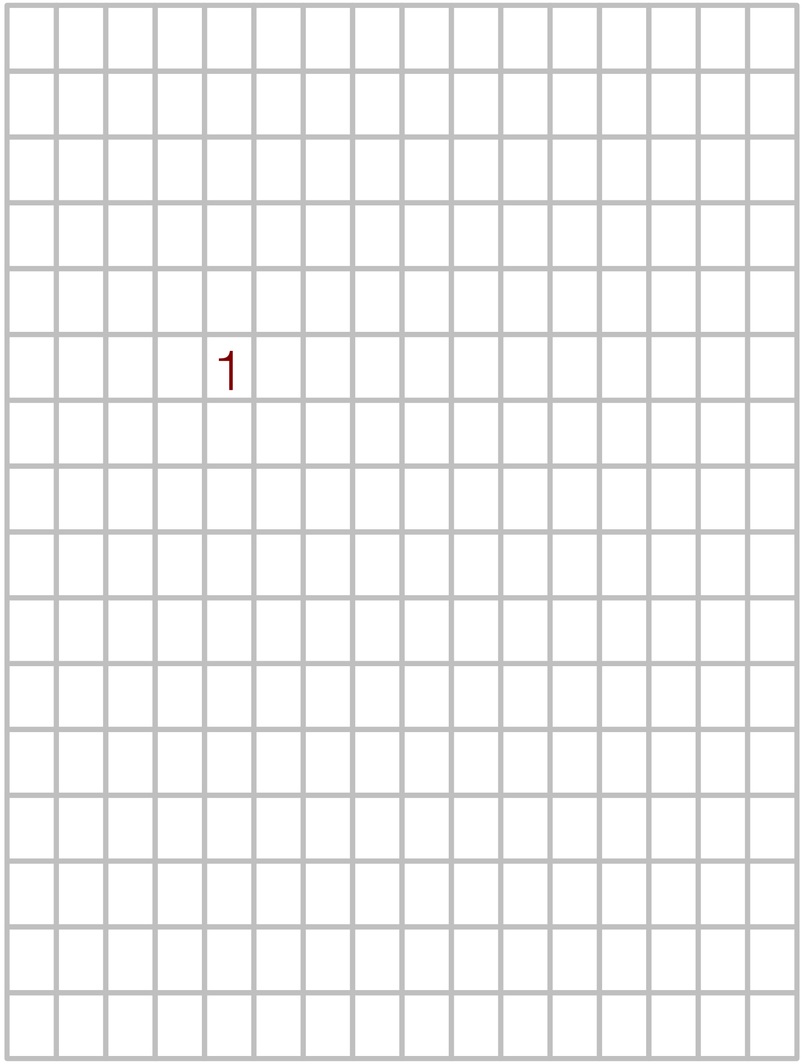
Spatial join

Leverage the graphics pipeline of the GPU

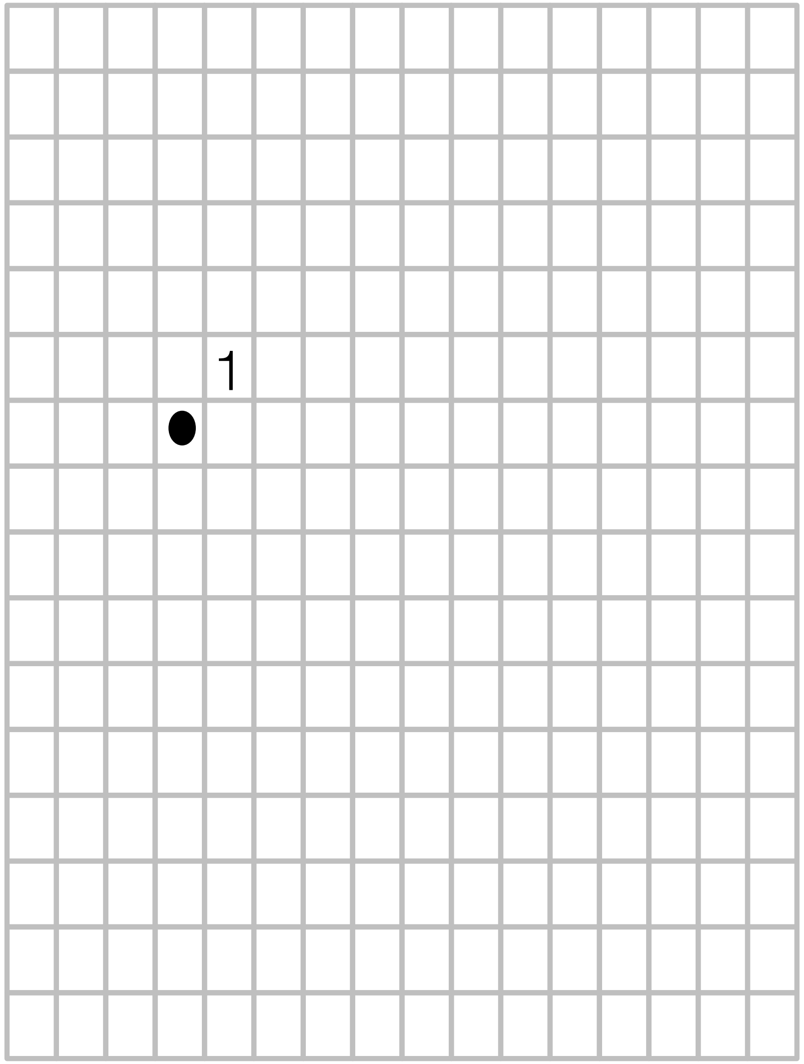
Raster Join: I. Draw the Points



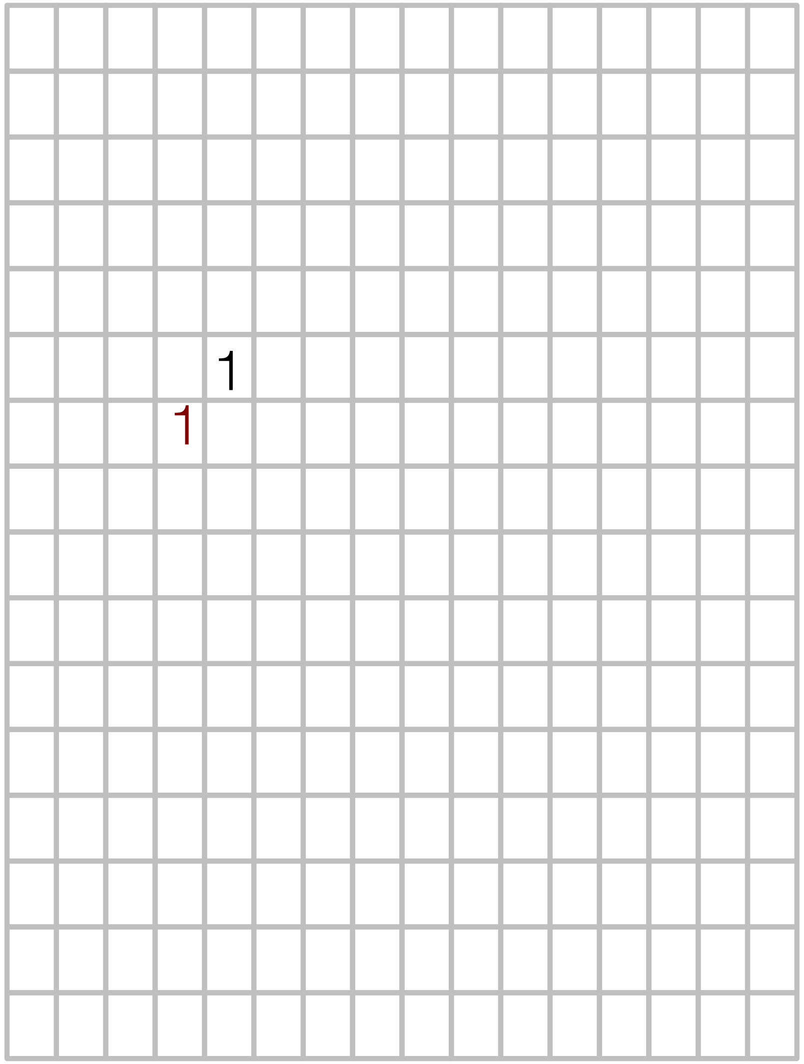
Raster Join: I. Draw the Points



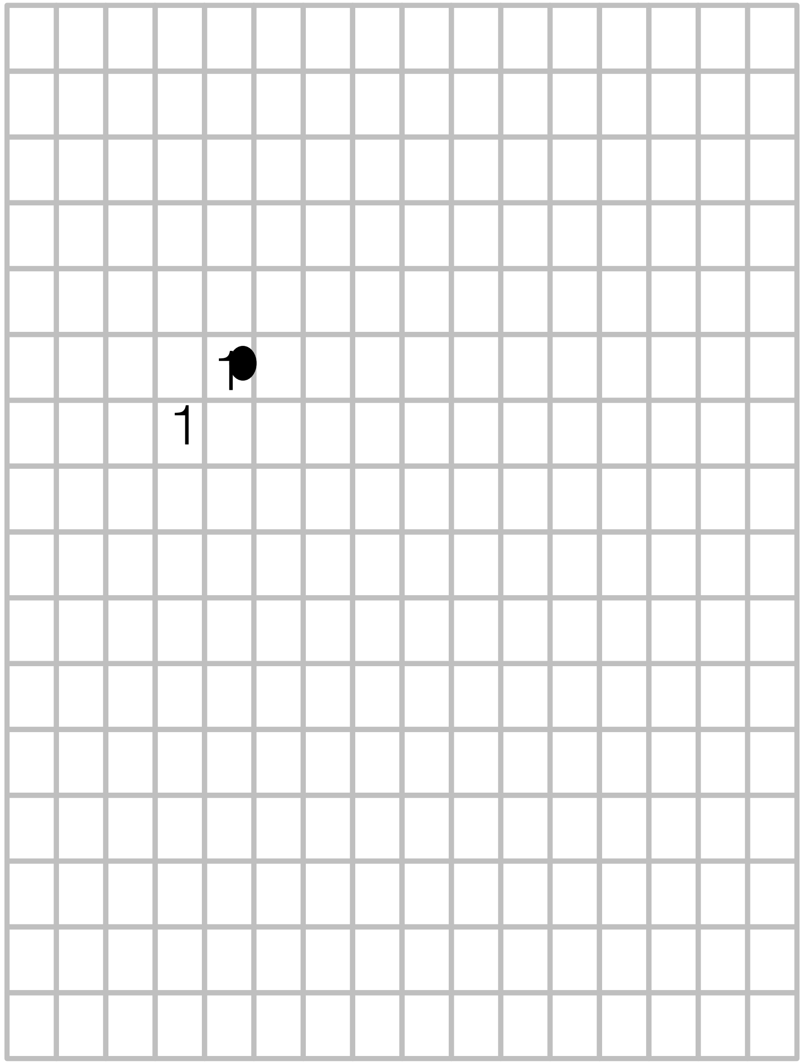
Raster Join: I. Draw the Points



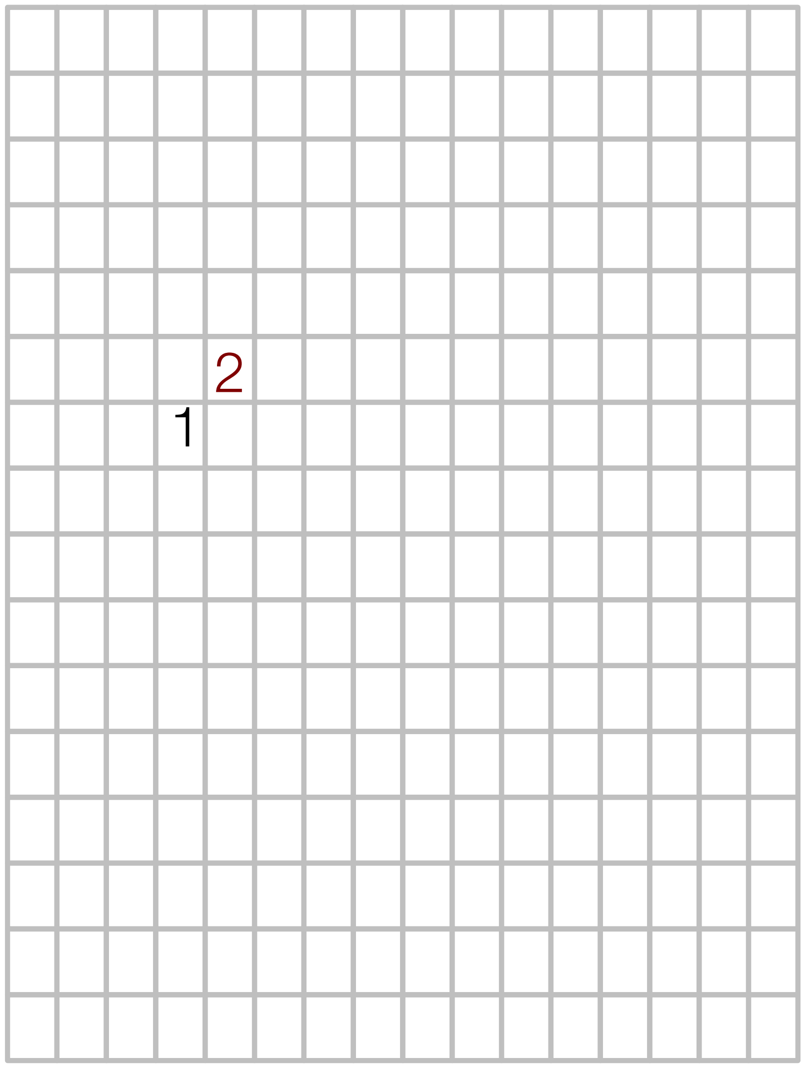
Raster Join: I. Draw the Points



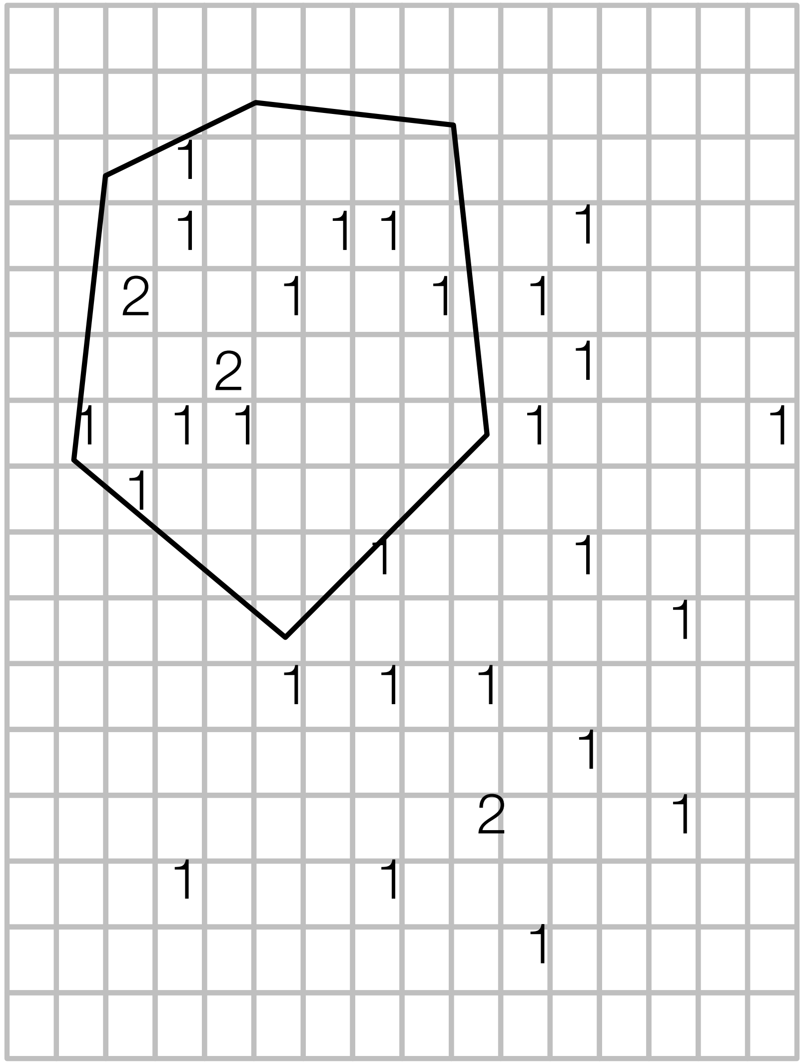
Raster Join: I. Draw the Points



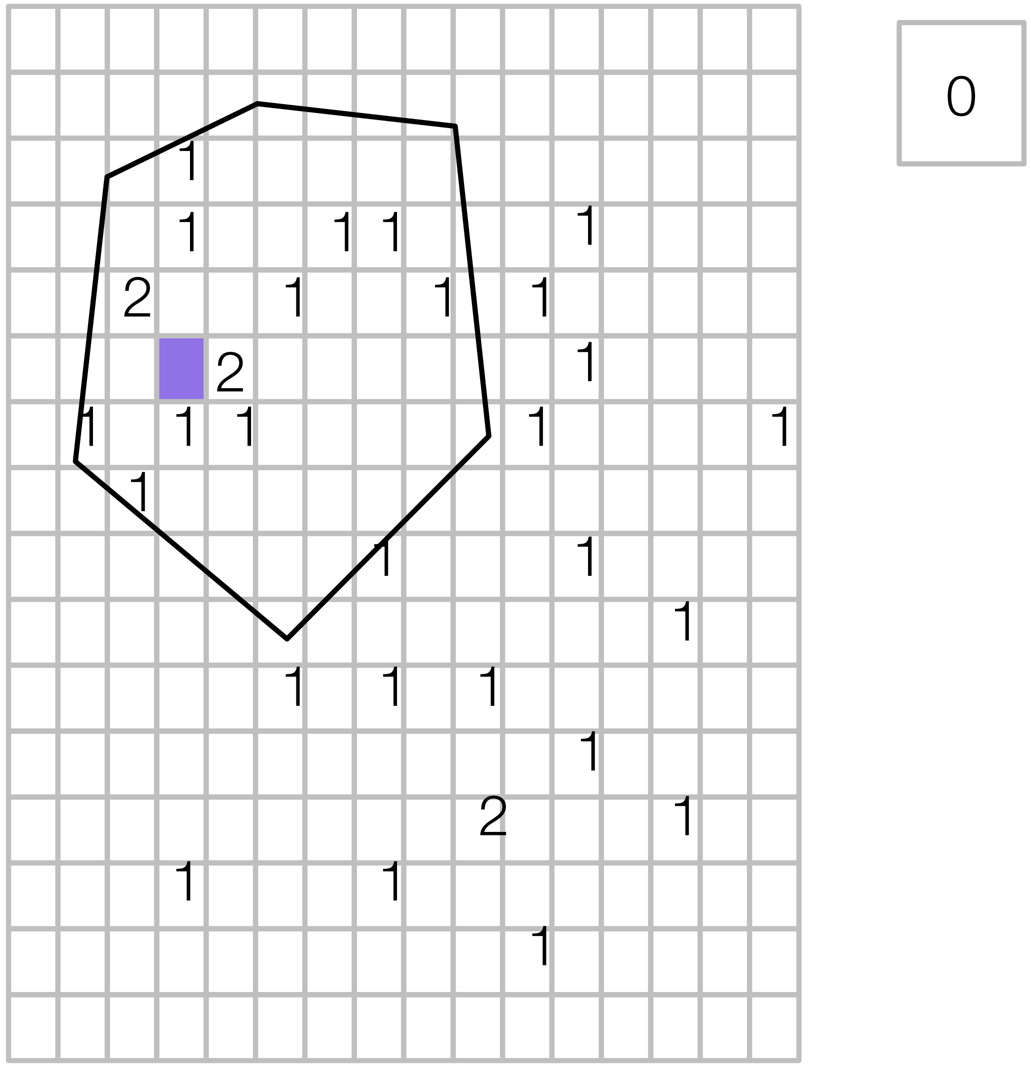
Raster Join: I. Draw the Points



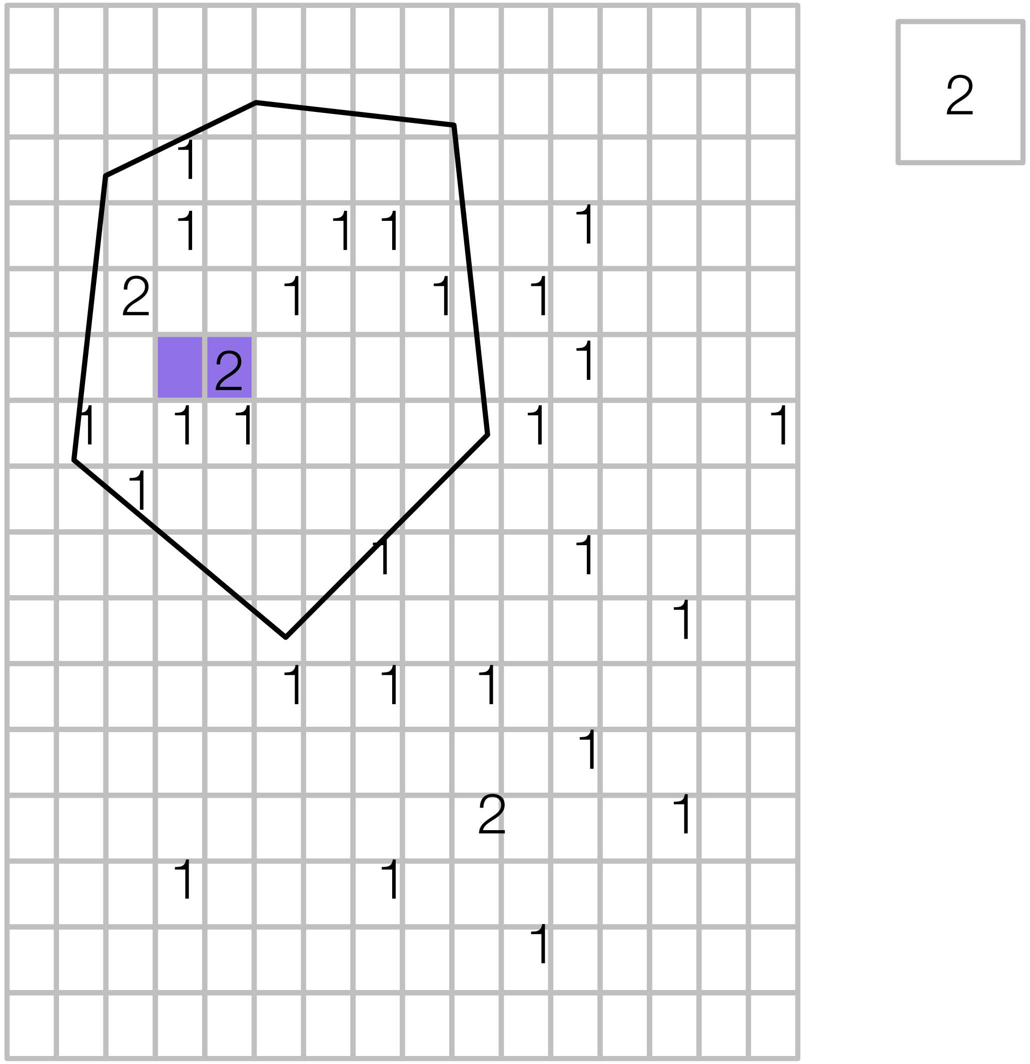
Raster Join: I. Draw the Points



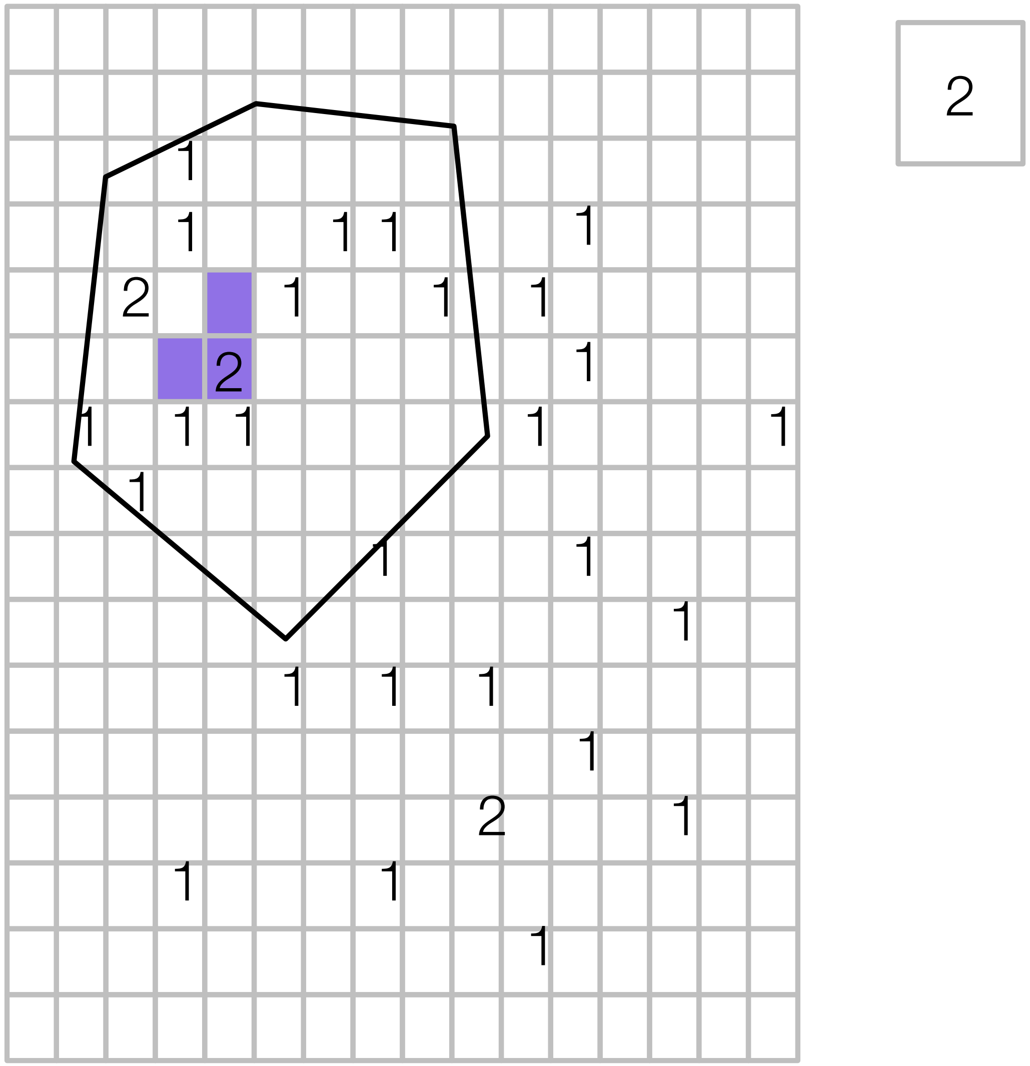
Raster Join: II. Draw the Polygons



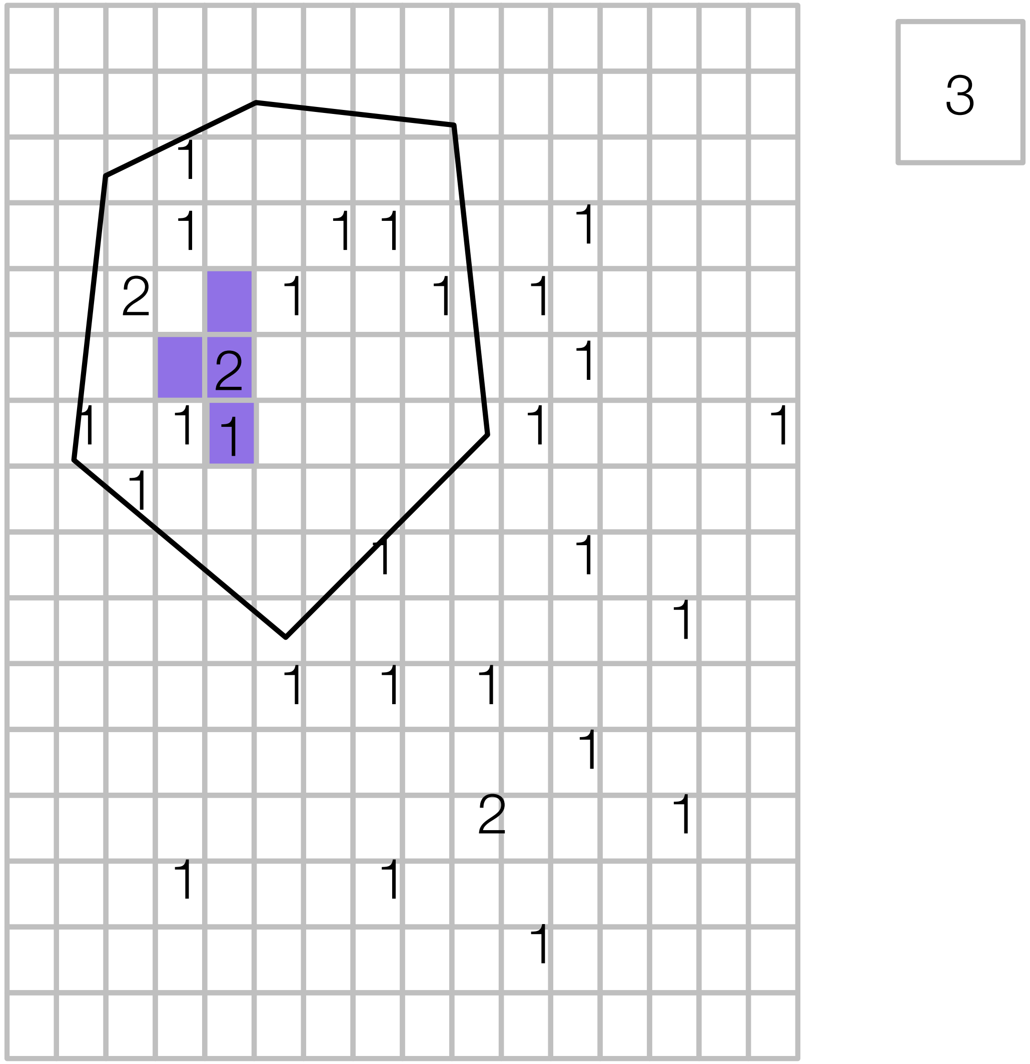
Raster Join: II. Draw the Polygons



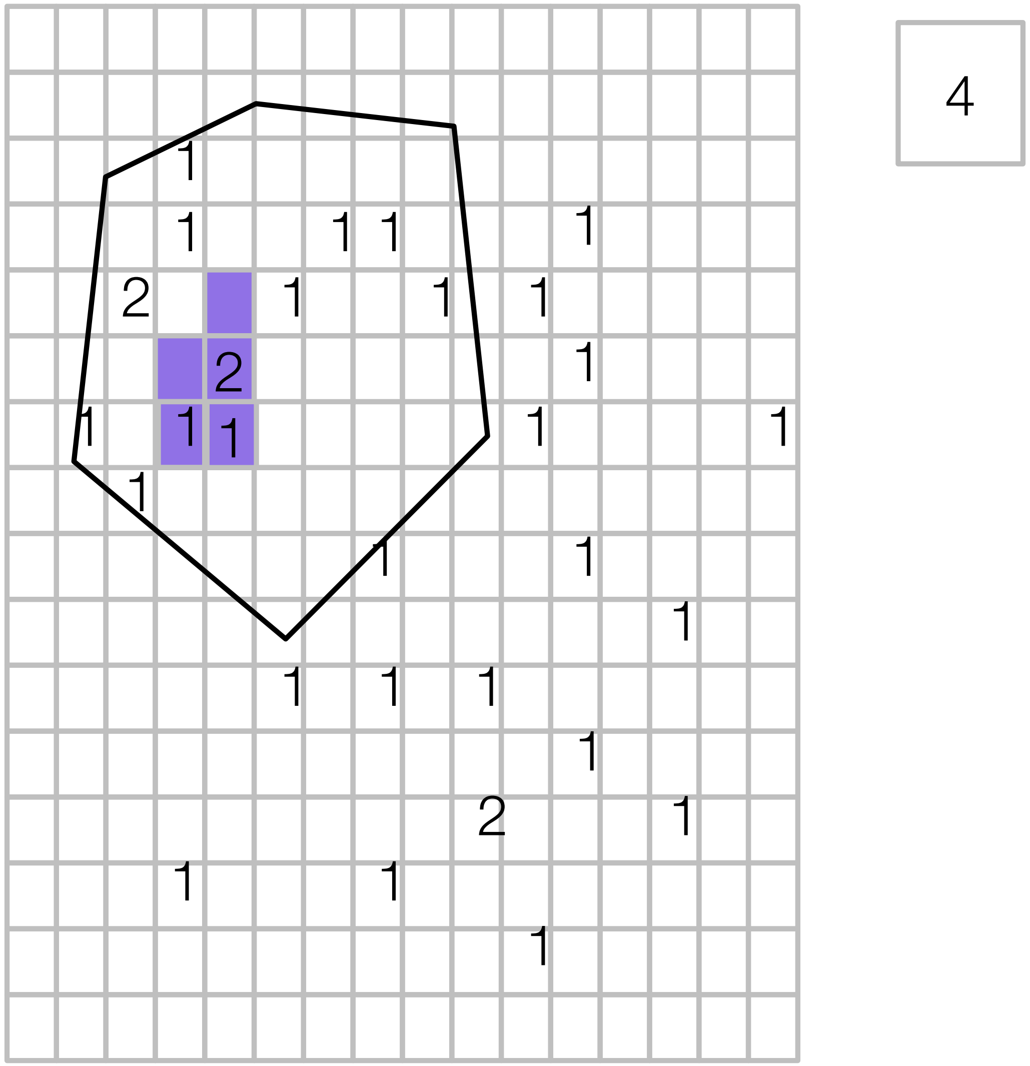
Raster Join: II. Draw the Polygons



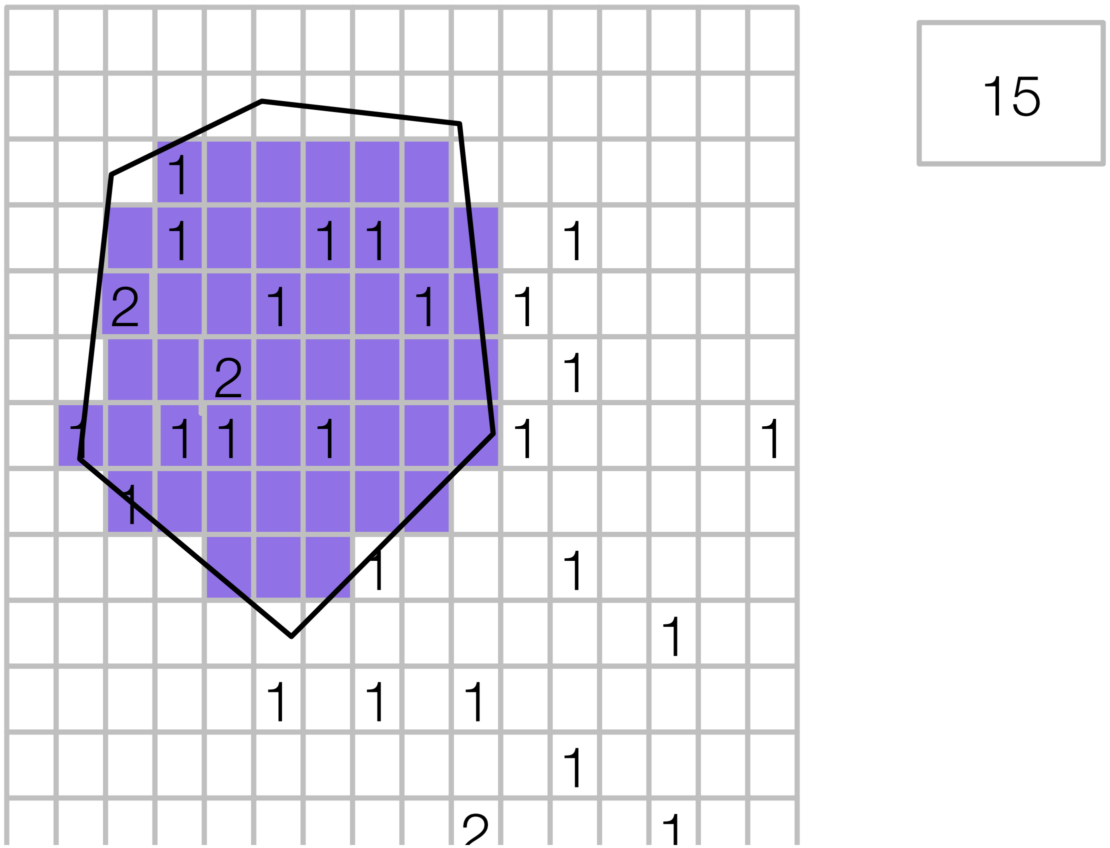
Raster Join: II. Draw the Polygons



Raster Join: II. Draw the Polygons



Raster Join: II. Draw the Polygons



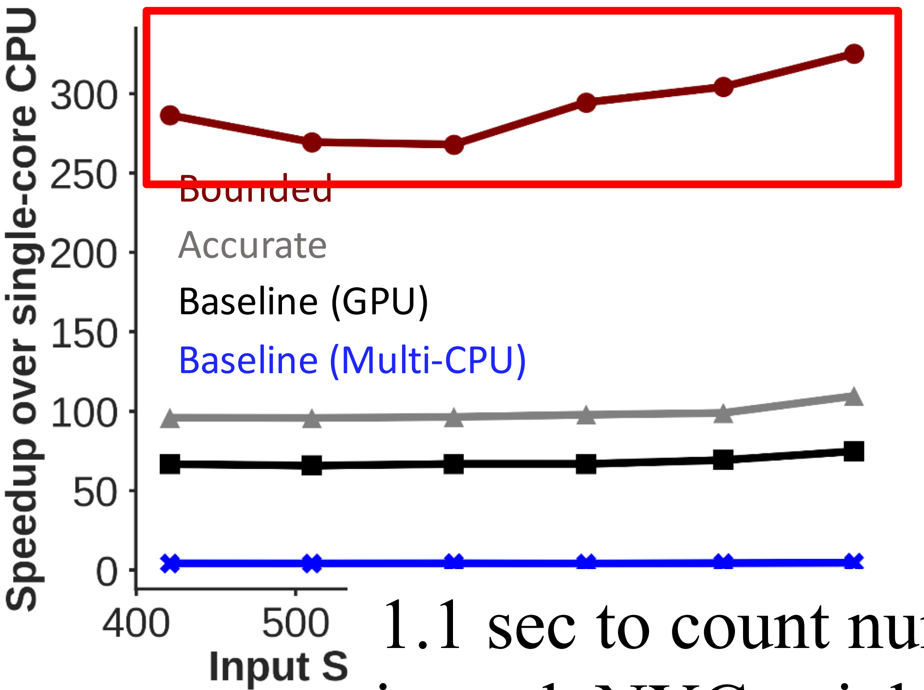
Exploits the native support for drawing in GPUs

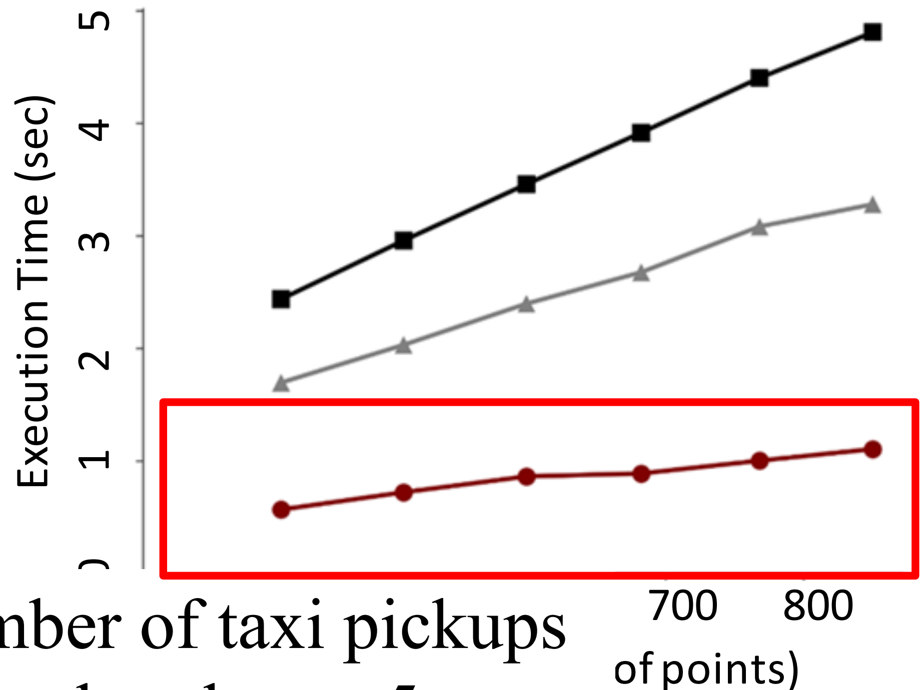
Combines the aggregation with the join operation

No Point-in-Polygon tests

Performance Evaluation

Hardware: Laptop with Intel Core i7 Quad-Core @2.8 GHz, 16GB RAM.
NVIDIA GTX 1060 GPU, 6GB VRAM (usage limited to 3GB)
Data Sets: NYC Taxi data (over 868 million points), 260 NYC neighborhood polygons





1.1 sec to count number of taxi pickups in each NYC neighborhood over 5 years

https://github.com/ViDA-NYU/raster-join

[Tzirita et al., PVLDB 2017]

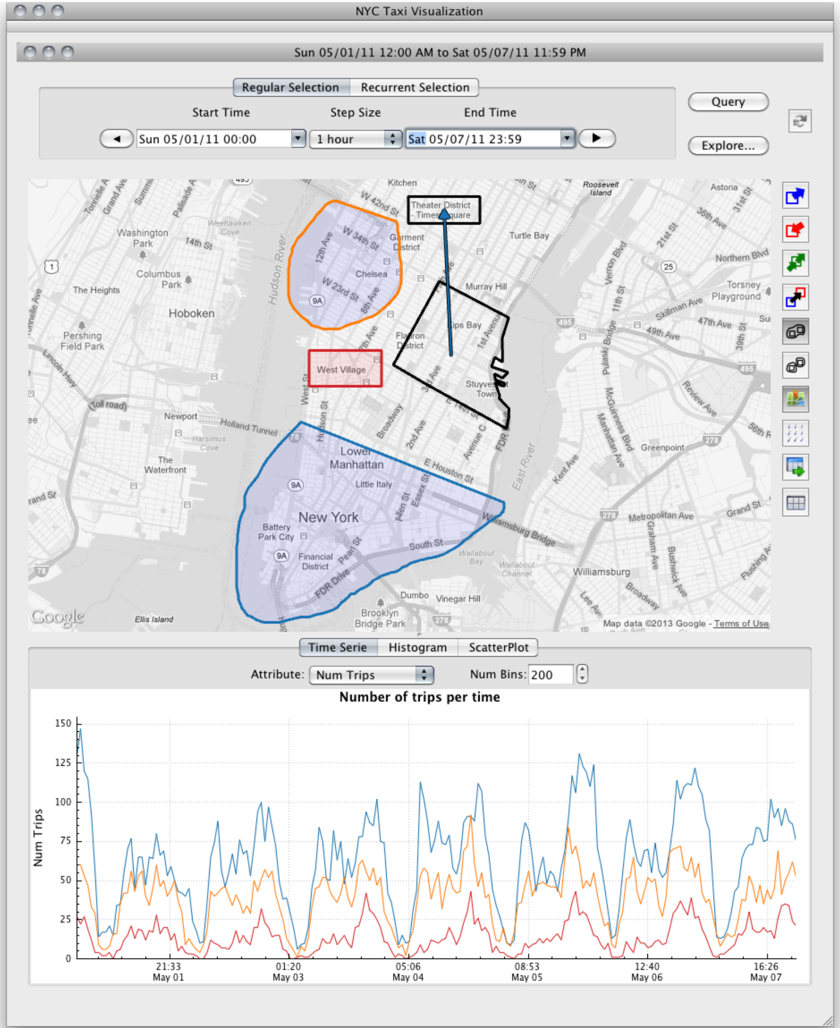
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Interactive Spatio-Temporal Selection

- Spatio-temporal index based on out-of-core kd-tree using GPUs
- Can index and simultaneously filter multiple attributes: avoid joins and reduce the number of point-in-polygon (PIP) tests
- Block-based kd-tree
  - Tree nodes store kd-tree, leaf nodes represent a *set of k-dimensional nodes* that point to a leaf block
  - Create *big* blocks – tree is small and fits in memory
  - Use GPU to search the blocks in parallel – speeds up PIP tests

http://www.taxivis.org



https://github.com/harishd10/mongodb

[Doraiswamy et al., ICDE 2016]

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Performance Evaluation

Find all trips between Lower Manhattan and the two airports, JFK and LGA, during all Sundays in May 2011.

Query	MongoDB	PostgreSQL		ComDB	
	Time	Time	Speed up	Time	Speed up
1	0.075	503.9	6718	20.6	274
2	0.080	501.9	6273	23.3	291
3	0.067	437.8	6534	21.6	322
4	0.070	437.1	6244	32.6	465

Time in Seconds

868 million trips; ~13k results/query

[Doraiswamy et al., ICDE 2016]

Take Away

- You don't need big iron to analyze big data, you can do it on your laptop!
- Usability requires combining techniques from Visualization, Computer Graphics, HCI, and data management [Doraiswamy et al., CG&A 2018]
- Connecting Visualization and Data Management Research [Chang et al., Dagstuhl 2018]
- Great potential for impact: democratizing large-scale data analysis

Impact: TaxiVis

---------- Forwarded message ----------
From: [redacted]@tlc.nyc.gov>
Date: Thu, Oct 24, 2013 at 4:58 PM
Subject: NYC taxi data
To: "Claudio Silva (csilva@nyu.edu)" <csilva@nyu.edu>, "Huy Vo (huy.vo@ny
"Caryn Joy Knutsen (caryn.knutsen@nyu.edu)" <caryn.knutsen@nyu.edu>, "
(kim.alfred@nyu.edu)" <kim.alfred@nyu.edu>
[redacted]

Hi all,

First, I would like to thank you all for coming to TLC
data. We were truly blown away! In fact, we had be
product like the one you've demonstrated to us. Aft
[redacted]
for us on Monday. We think that could be a great s
future use for our data in combination with other av

[redacted]

Cheers,

[redacted]

"The speed at which the tool permits us to work has saved multiple hours of staff time and has dramatically improved the unit's output and capabilities."

Assistant Commissioner, DoT

http://www.taxivis.org

[Ferreira et al., IEEE TVCG 2013]

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Impact: Urbane

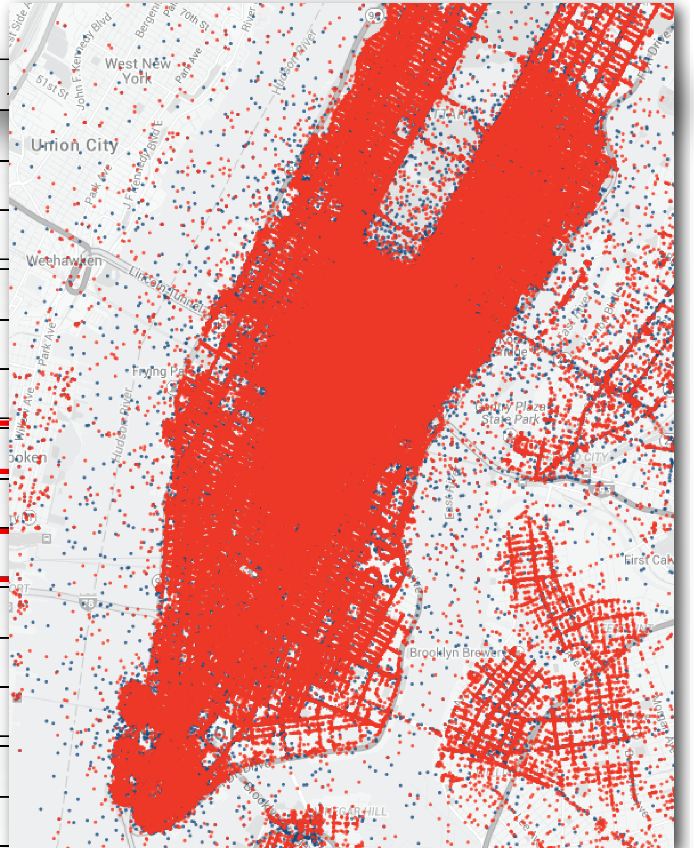
Neighs	Built(FAR)	Crime	Food	Jobs	Noise	People	Price	School(s)	Sky	Subway	Taxi
Avg.Neigh.	63.845	0.251	84.008	201.319	0.223	51.172	1537.269	57.854	58.934	1.208	317.388
Financial District	73.054	0.306	91.87	614.477	0.232	52.856	1217	49.801	50.324	3.116	221.512
Civic Center	75.605	0.251	68.42	329.631	0.259	15.096	1964	46.238	57.862	1.249	130.522
Chelsea	74.315	0.255	33.361	166.351	0.218	44.091	1603	66.106	58.594	1.299	328.313
Kips Bay	73.069	0.278	89.078	114.059	0.264	65.984	1220	67.817	59.158	0.323	341.727
Lower East Side	75.892	0.272	103.096	24.725	0.244	44.273	1427	63.955	62.58	0.868	163.783
Midtown	74.122	0.264	90	605.852	0.23	56.309	1683	66.046	51.632	1.213	515.279
Murray Hill	73.216	0.241	154.136	327.287	0.186	79.791	1345	67.687	57.185	0.453	598.477
Tribeca	77.706	0.267	80.932	268.418	0.306	57.652	1832	43.578	58.006	2.233	274.208
Upper East Side	75.732	0.261	42.58	107.165	0.197	109.632	1562	9.975	61.748	1.136	690.693

[Ferreira et al., IEEE VAST 2015; Doraiswamy et al., ACM SIGMOD 2018]

Urban Data Quality

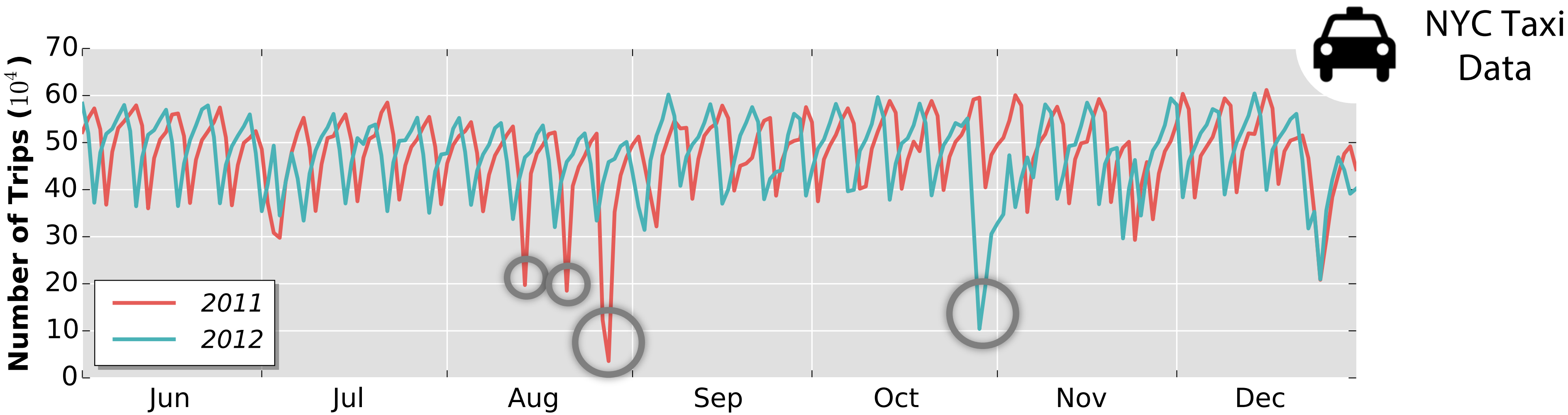
- NYC Taxi Data: ~500k trips/day; 868 million trips in 5 years

Dataset	Statistic	Trip Duration (min)	Trip Distance (mi)	Fare	
2008	Min	0.00	0.00		
	Avg	16.74	2.71		
	Max	1440.00	50.00		
2009	Min	0.00	0.00		
	Avg	7.75	6.22		
	Max	180.00	180.00		
2010	Min	-1,760.00	-21,474,834.00		
	Avg	6.76	5.89		
	Max	1,322.00	16,201,631.40		
2011	Min	0.00	0.00		
	Avg	12.35	2.80		
	Max	180.00	100.00		
2012	Min	0.00	0.00		
	Avg	12.32	2.88		
	Max	180.00	100.00	500.00	200.00



Data quality issues [Freire et al., IEEE DEB 2016]

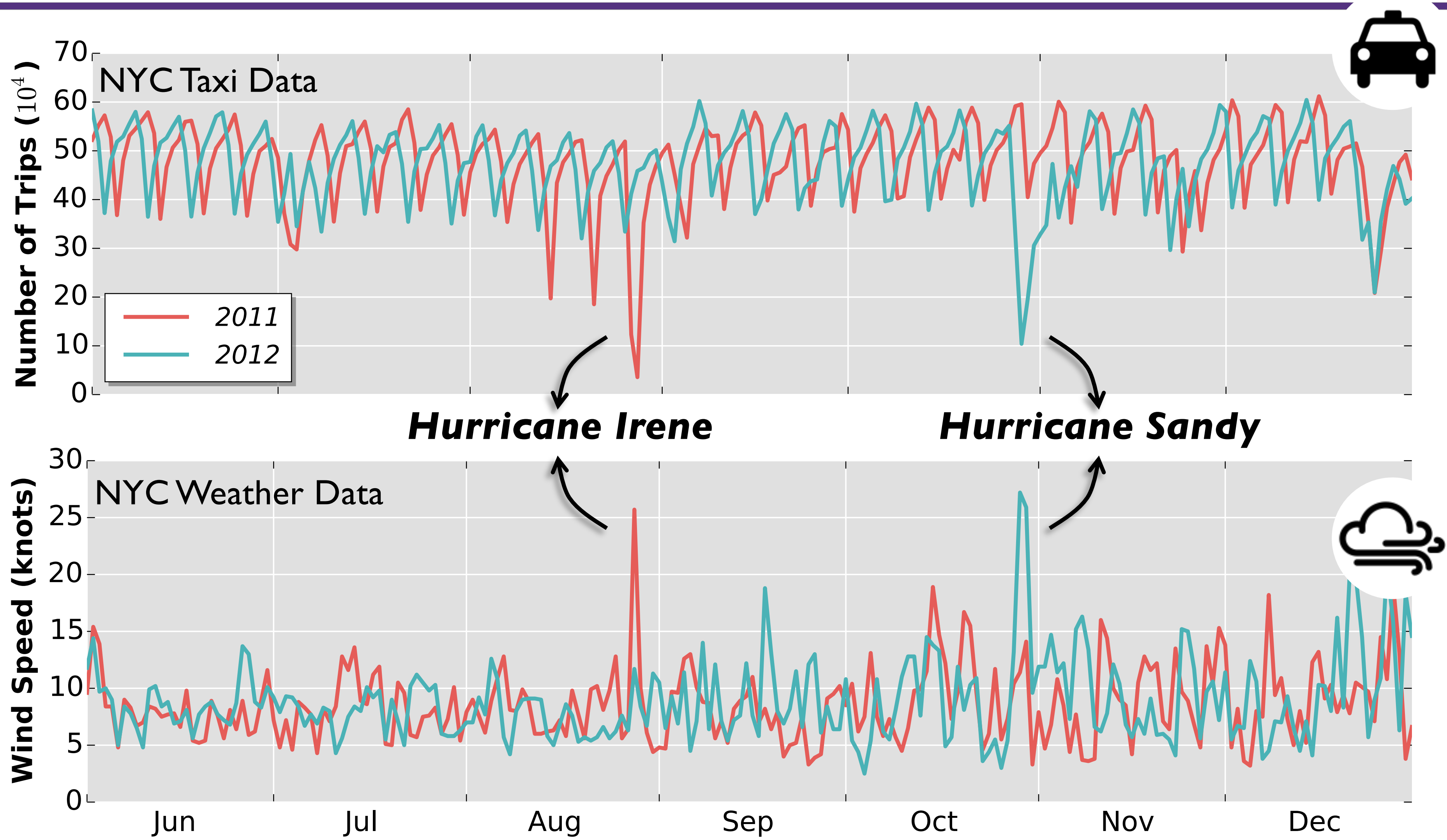
Understanding Data



Are these big drops data quality issues in the data?

Or do they correspond to *real* events?

Understanding Data



Can we use data to explain data?

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

The Data Polygamy Framework

- **Automatically discovers relationships** between data sets
- Each data set can be related to **zero or more** data sets through several attributes: *Data sets are polygamous*
- Guide users in **data discovery and analysis** by allowing them to pose ***relationship queries***

Find all data sets related to a given data set $\mathbb{D}$

Identify potential data quality issues

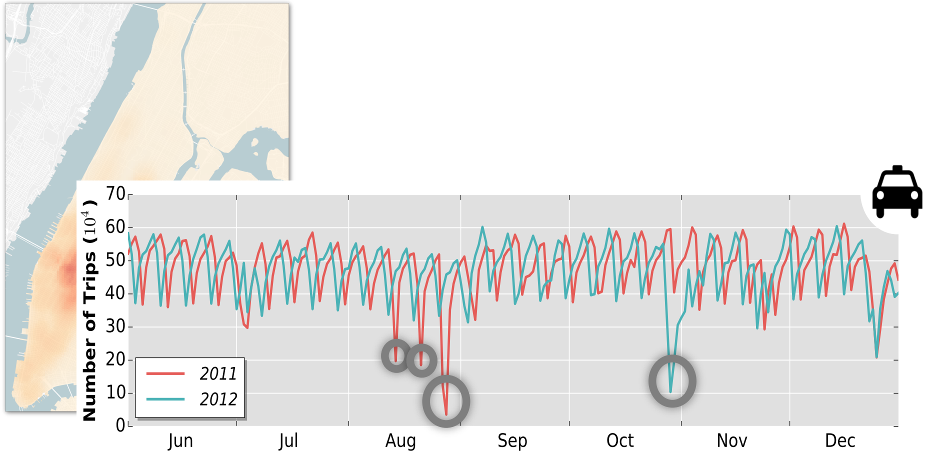
Discover attributes for predictive models

Explain *interesting* features

[Chirigati et al., ACM SIGMOD 2016; Chan et al., ACM SIGMOD 2017]

Relationship Discovery

- Desiderata:
  - Take both space and time into account
  - Capture *atypical* behavior



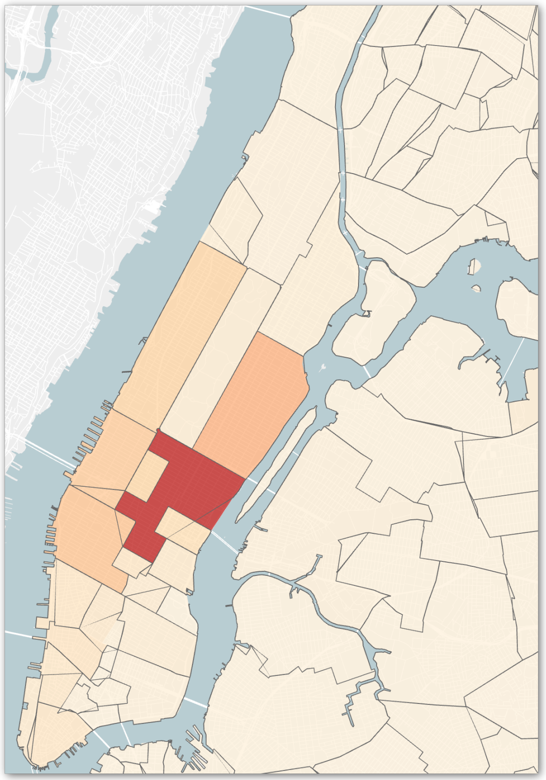
- Challenges
  - Many data sets, each consisting of many attributes, e.g., Weather data: >200 attributes; NYC Open data: 8 attributes per data set on average
  - Data sets can be large, e.g., 180M trips per year
  - Data at multiple spatio-temporal different resolutions
  - Combinatorially large number of relationships to evaluate
    - ~2.4 million possible relationships among NYC Open Data alone for a single spatio-temporal resolution

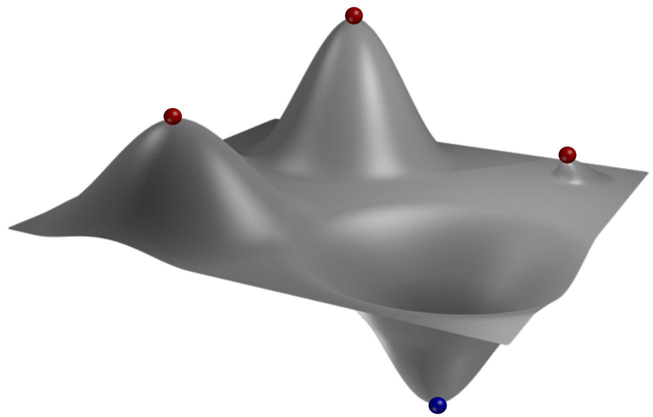
NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Topology-Based Relationships

- Use topological representation for the data
- Each attribute is represented as a set of time-varying scalar functions: $f : [\mathbb{S} \times \mathbb{T}] \rightarrow \mathbb{R}$

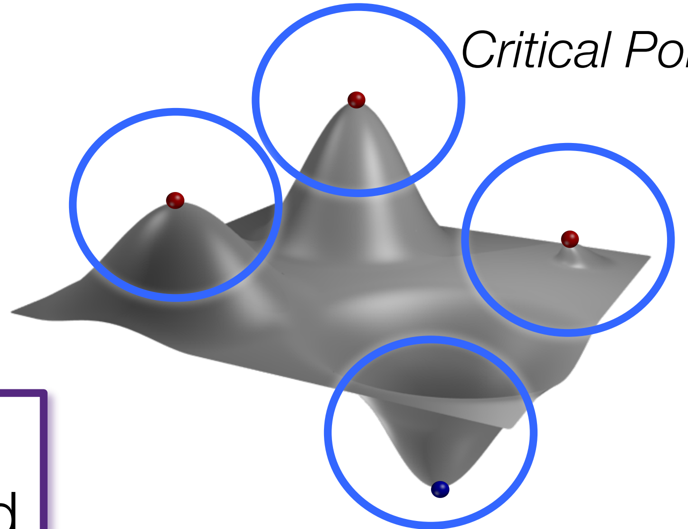




Topology-Based Relationships

- Use topological representation for the data
- Each attribute is represented as a set of time-varying scalar functions: $f : [\mathbb{S} \times \mathbb{T}] \rightarrow \mathbb{R}$
- Uniform representation for all data
- Naturally captures atypical behavior –

salient features



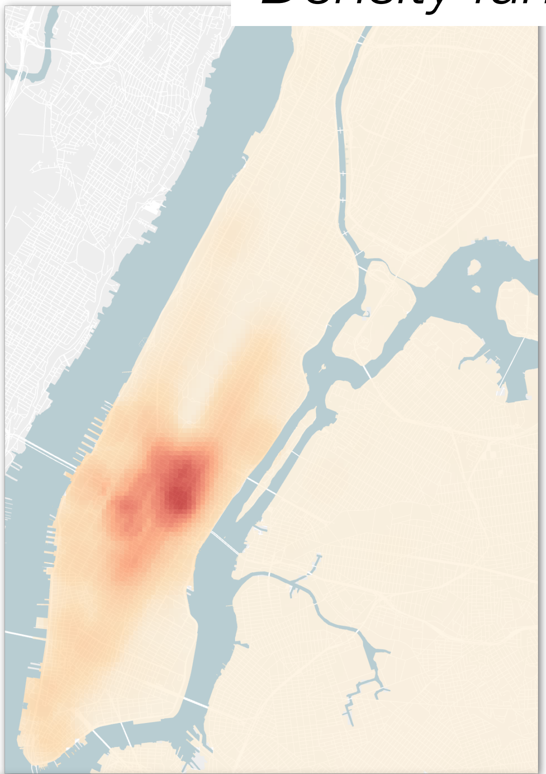
A salient feature is a spatio-temporal region whose behavior differs from its neighborhood

Two attributes are *related* if their **salient features** overlap in space and time

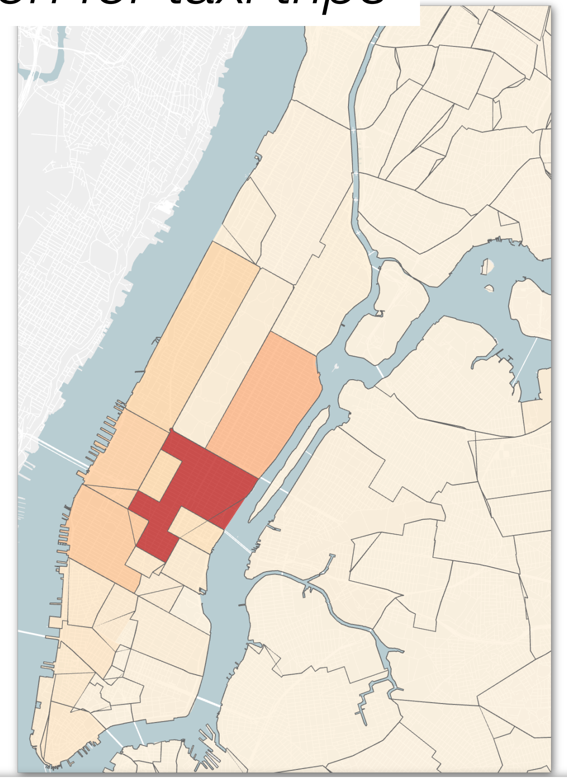
Data Set to Scalar Functions

- Each attribute in a data set represented as a set of time-varying scalar functions
- Functions computed at all possible resolutions

Density function for taxi trips



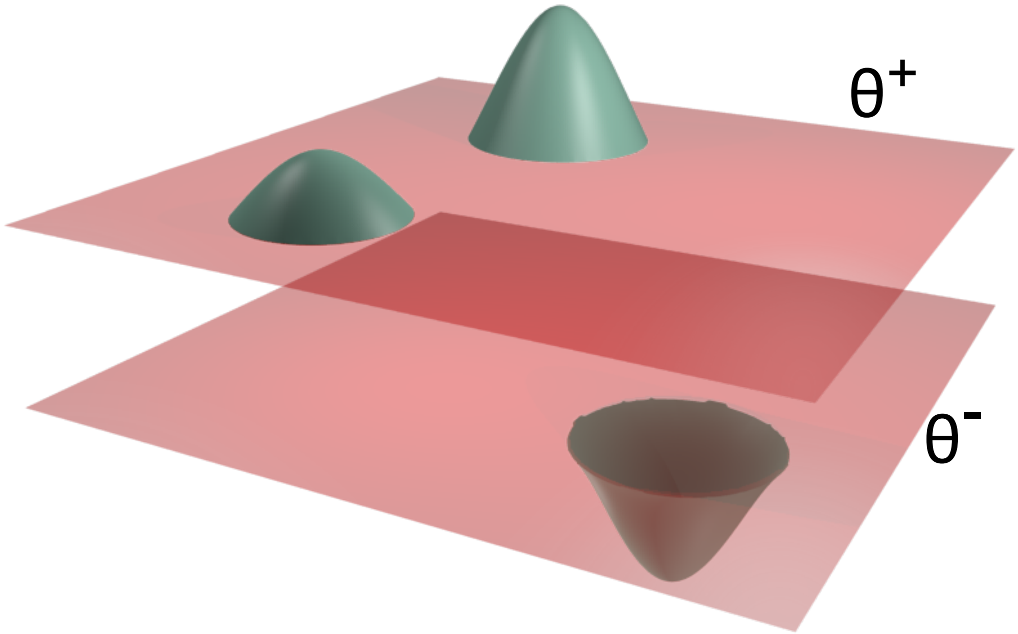
$\mathbb{S}$ *High Resolution Grid*



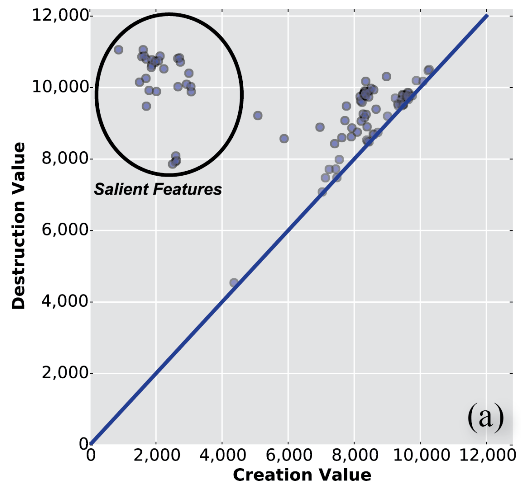
$\mathbb{S}$ *Neighborhood Resolution*

Identify Salient Features

- Topological features of *a scalar function:* salient features correspond to peaks and valleys
- Neighborhood defined by a threshold
  - Use *topological persistence* to automatically compute thresholds in a data-driven fashion



minima of the taxi-density function

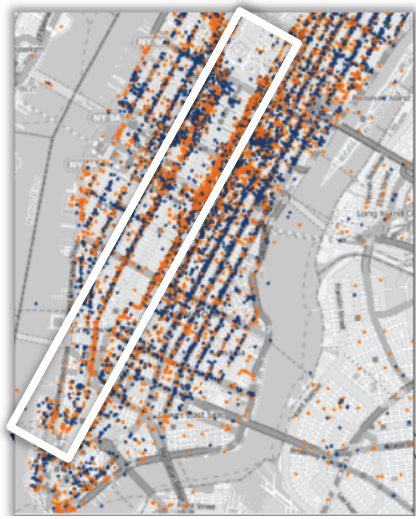


Identify Salient Features

- Topological features of *a scalar function;* salient features correspond to peaks and valleys
- Neighborhood defined by a threshold
  - Use *topological persistence* to automatically compute thresholds in a data-driven fashion
- *Merge Tree Index* efficiently identifies features at all resolutions
  - *O(n log n)* to construct
  - Computing features is output sensitive
- Benefit: features can have arbitrary shapes

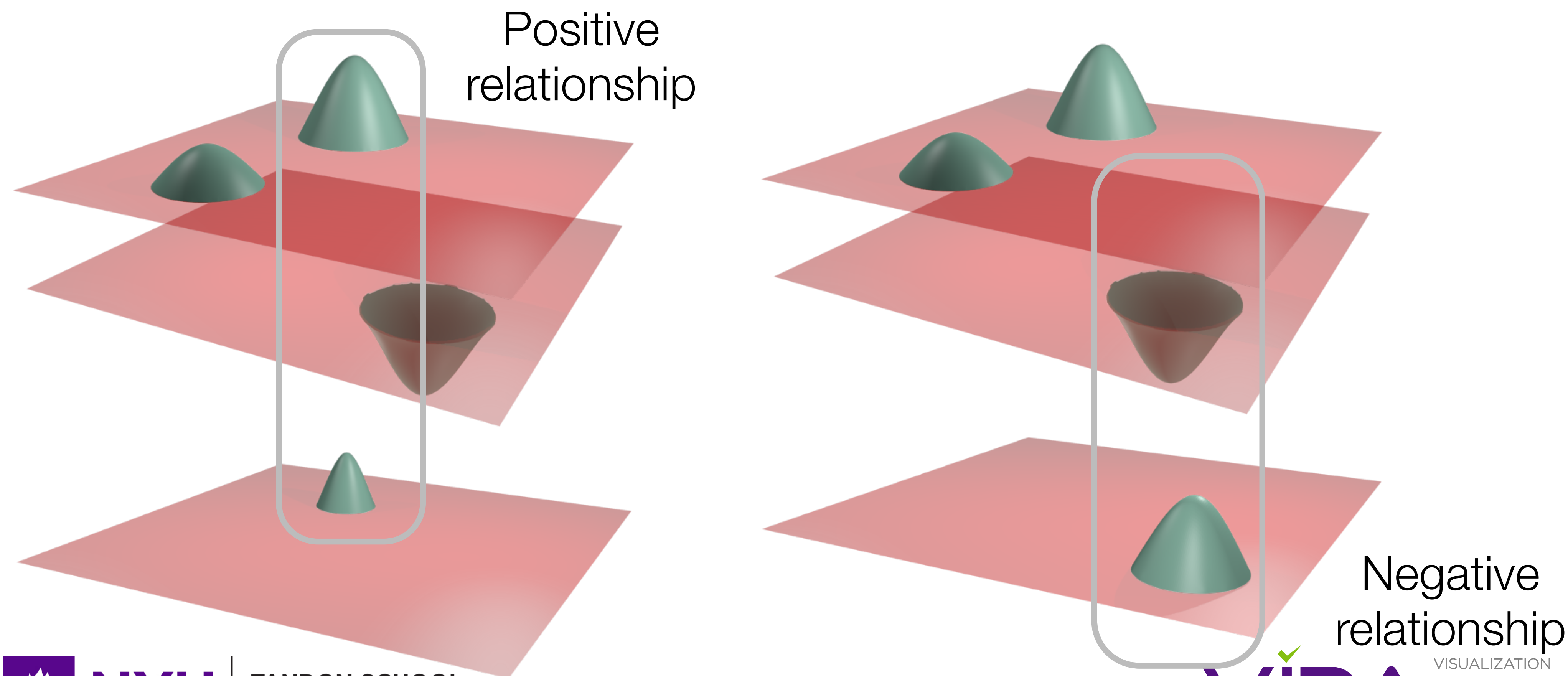
8am - 9am
May 1 2011

5 Boro Bike Tour



Find Candidate Relationships

- Relationship between functions *f* and *g* consists of the set of spatio-temporal points that are features in both functions
  - E.g., for Hurricane Sandy, there is a negative feature in the taxi density function and a positive feature in the wind speed function



Evaluating Relationships

- *Relationship Score:* Captures the nature of the relationship – how positively or negatively related

$$\tau = \frac{\#p - \#n}{\#p + \#n}$$

p – no. of positive features
n – *no.* of negative features

- *Relationship Strength:* How often the functions are related – strong or weak

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

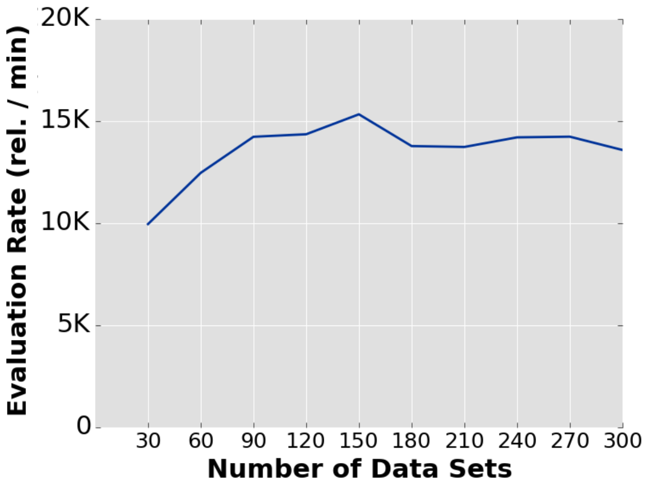
- Restricted Monte Carlo procedure to test the *statistical significance* accounting for the spatial and temporal proximity
  - Prune potentially coincidental relationships

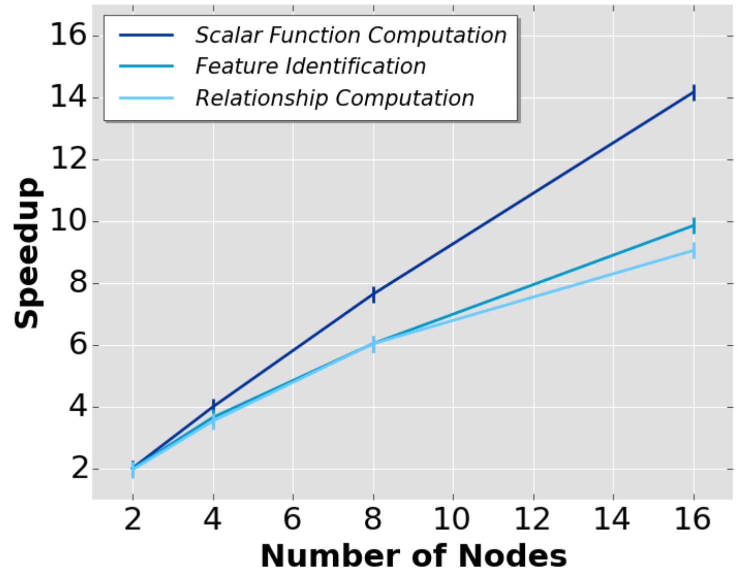
Experimental Evaluation

- Implemented using map-reduce
  - Feature identification and relationship evaluation are independent operations
- Two collections of data sets used for experiments
  - NYC Urban: *9 data sets* from NYC agencies
  - NYC Open Data: 300 spatio-temporal data sets

Quantitative Evaluation

- Approach is efficient: 200 min to compute scalar functions and features for NYC Open Data; and 60 min for NYC Urban
- Scales linearly with number of compute nodes
- Query rate: evaluate 10^4 relationships per minute
- Assessed correctness and robustness





https://github.com/ViDA-NYU/data-polygamy
[Chirigati et al., ACM SIGMOD 2016]

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Qualitative Evaluation

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

Taxis and Rainy Days

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

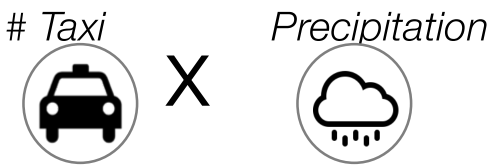
Hypothesis: Taxi drivers are target earners

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

Strong positive relationship between precipitation and average fare

Taxis and Rainy Days

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



Find all relationships between Taxi and Weather data sets

This hypothesis had been refuted by [Farber 2014]

- Farber did not find a correlation (using OLS regression) between drivers' earnings and rainfall.
- But (i) he did not take into account the amount of rainfall—instead, he used a binary value indicating whether it rained or not; and (ii) he considered the entire time period—periods with very sparse rainfall are considered equivalent to those having higher rainfall.

It is important to consider salient features

Take Away

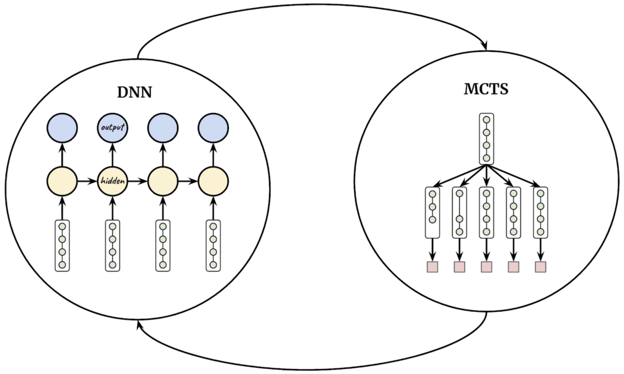
- Guide users in data exploration: use data to explain data and gain trust
- Caution: Helps generate hypotheses, further validation is needed to ascertain that a relationship really holds
- Variations of the approach are possible
  - Use different data models, event detection methods, alignment strategies, and relationship types [Bessa et al., work in progress]
- Useful for data discovery – to find *related* data sets

Vision: use as on operation in search engines for structured data [DARPA D3M]

https://www.darpa.mil/program/data-driven-discovery-of-models

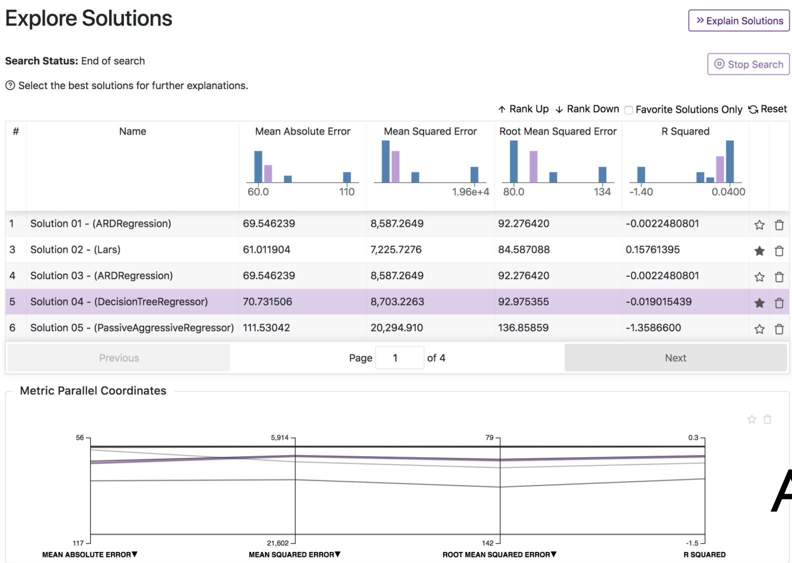
AlphaD3M + Visus + Auctus

Automatic synthesis pipelines using reinforcement learning with self-play



[Drori et al., ICML AutoML 2018]

User-guided exploration and curation of pipelines



[Santos et al., ACM SIGMOD HILDA 2019]

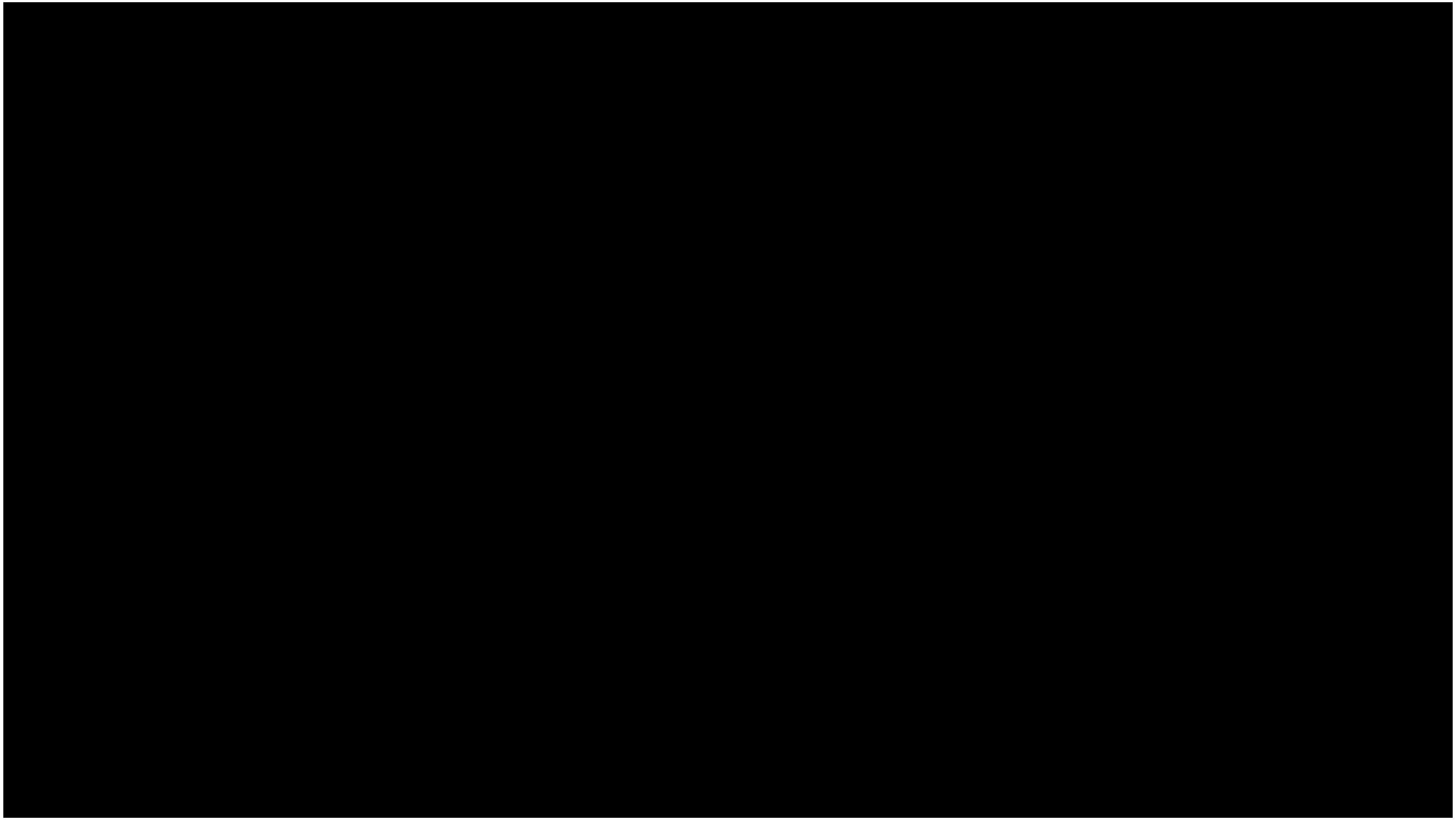
Data augmentation

[Chirigati et al., AIDR 2019]

NYU TANDON SCHOOL OF ENGINEERING

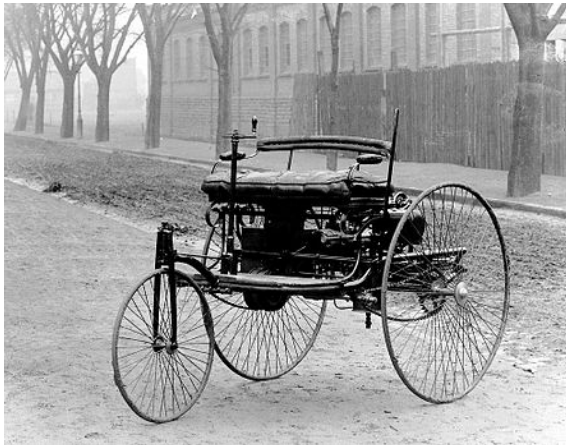
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Augmenting Data with Auctus



Conclusions & Open Problems

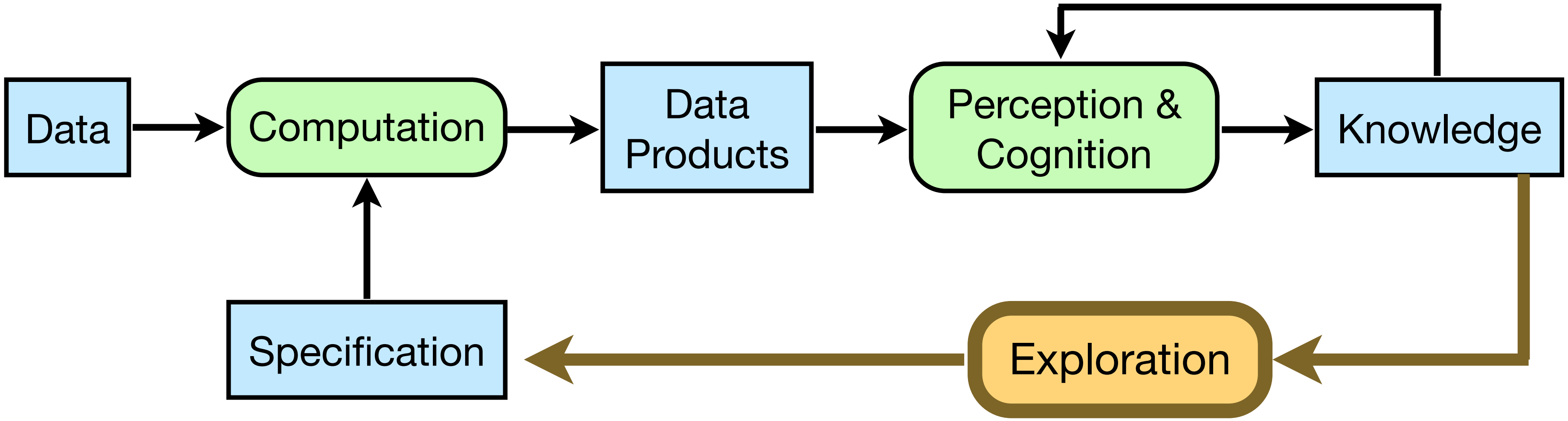
- Data-driven exploration has transformed science, government and industry
- Grand challenge: empower domain experts to effectively explore data and extract actionable knowledge they can trust



Conclusions & Open Problems

- Data-driven exploration has transformed science, government and industry
- Grand challenge: empower domain experts to effectively explore data and extract actionable knowledge they can trust
- Need new techniques and usable tools that
  - Guide users as they generate and test hypotheses
  - Help them assess the quality and debug their results

Provenance for Data-Driven Exploration

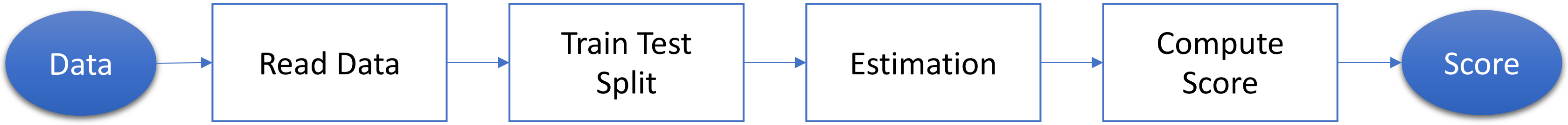


[Modified from Van Wijk, Vis 2005]

- Need to systematically capture the provenance of the exploration process [VisTrails, ReproZip, noWorkflow]
- Benefits: transparency + reproducibility

Identify root causes of problems – both in data and computational processes

Debugging Data Science Pipelines



P = {Data, Library, Estimator}
U_{data} = {Iris, Digits, Images}
$U_{library}$ = {1.0, 2.0}
$U_{estimator}$ = {Logistic regression, Decision tree, Gradient boosting}

E = score > 0.6

Instance	Data	Library	Estimator	Score	Evaluation
CP_1	Iris	1.0	Logistic regression	0.9	Succeed
CP_2	Digits	1.0	Decision tree	0.8	Succeed
CP_3	Iris	2.0	Gradient boosting	0.2	Fail
CP_4	Digits	2.0	Gradient boosting	0.3	Fail
CP_5	Iris	1.0	Decision tree	0.7	Succeed
CP_5	Images	1.0	Gradient boosting	0.9	Succeed

- Analyze provenance and explore parameter space to identify root causes

[Lourenço et al., ACM SIGMOD DEEM 2019]

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Conclusions & Open Problems

- Data discovery, cleaning, and integration
  - Handle data in the wild: no key-foreign key, incomplete metadata, dirty data
  - Advanced profiling – including relationship discovery
  - Assist users in cleaning: usability + provenance [Vizier, SIGMOD2019]
- Need interdisciplinary teams to solve real problems
  - Visualization, data management, computational topology, computer graphics, statistics
  - Collaboration with domain experts
  - Virtuous cycle: interdisciplinary research that derives new problems and solutions for multiple areas
- Data management community is well positioned to have tremendous practical impact

Acknowledgments

- Funding: Google, National Science Foundation, Moore-Sloan Data Science Environment at NYU, and DARPA.

謝謝
고맙습니다
Merci
Thank you
Obrigada
благодаря
Kiitos
धन्यवाद
Tack
Danke
Ευχαριστω
Bedankt

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER