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

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

NYU | TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Data-Driven Exploration

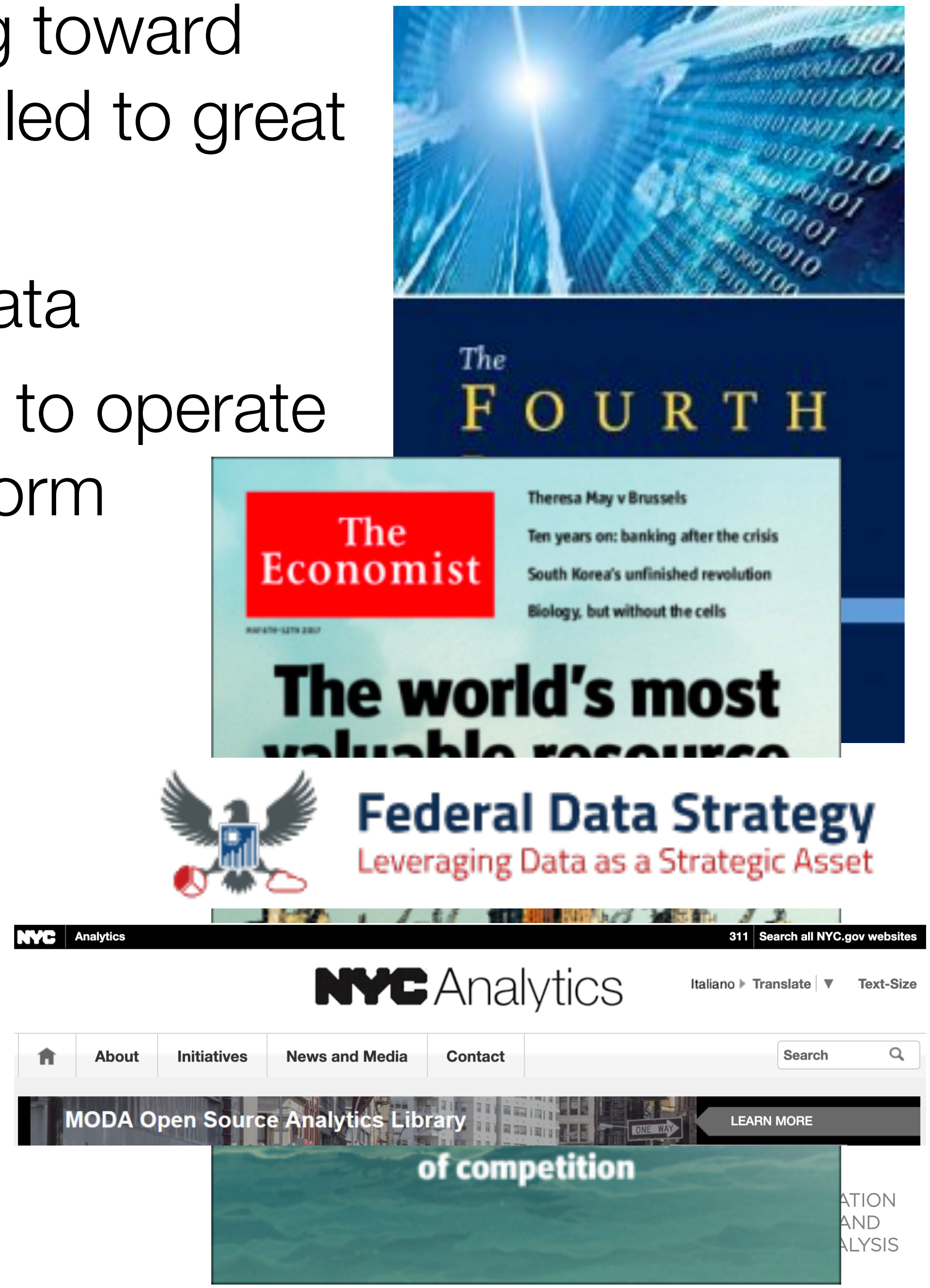
- Every scientific domain is moving toward data-driven exploration, this has led to great advances and discoveries
- Companies are capitalizing on data
- Government agencies uses data to operate efficiently, make policies, and inform decisions

Computing is free

Storage is free

Data are abundant

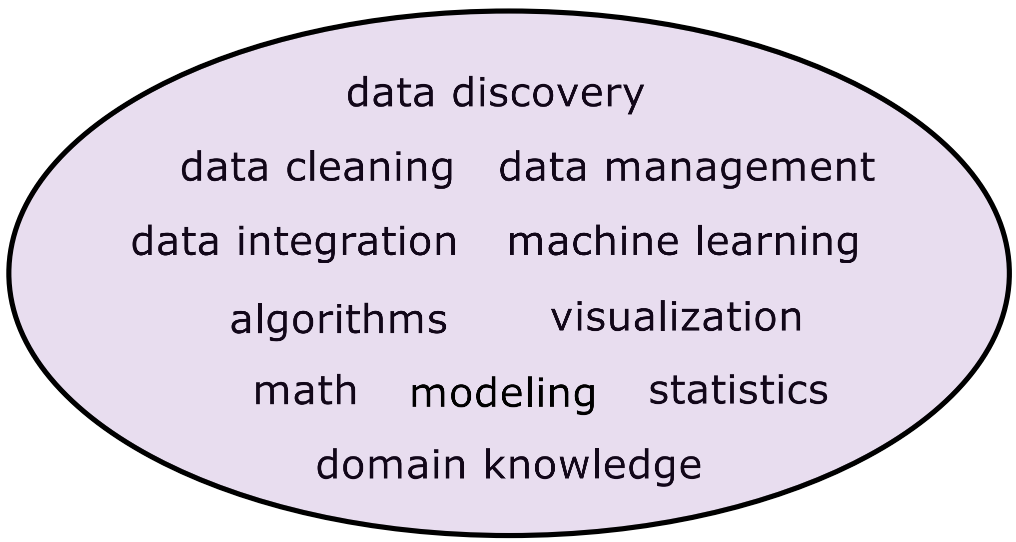
The bottlenecks lie with people



NYU TANDON SCHOOL OF ENGINEERING

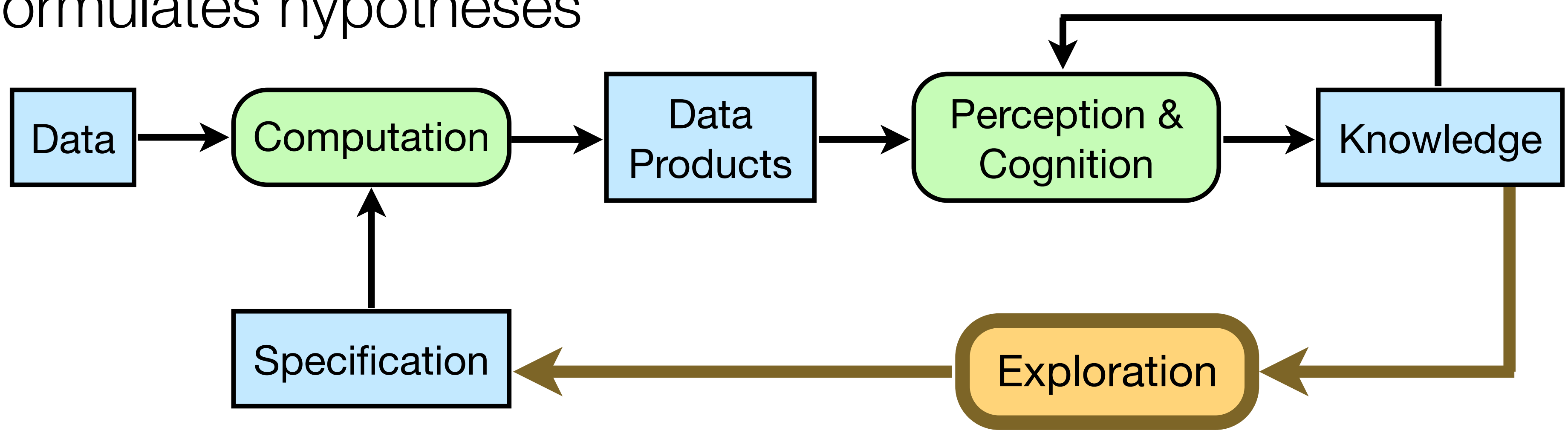
Data-Driven Exploration: Challenges

- Data are vast and produced at unprecedented rates
  - Sources are broad, varied, unreliable, and difficult to integrate
- Complex computational processes are required to extract insight -- hard to assemble and require expertise in a wide range of topics and tools



Data-Driven Exploration: Challenges

- Diverse group of experts and data enthusiasts
- It is difficult for domain experts to explore data
  - Dependence on data scientists distances domain experts from the data
  - Analyses are mostly confirmatory (Tukey, 1977) – batch-oriented analysis hampers exploration
- Exploratory analyses are inherently iterative as one tests and formulates hypotheses



[Modified from Van Wijk, Vis 2005]

Data-Driven Exploration: Challenges

- After many steps…
  - It is easy to get lost and not remember how a result was derived
  - Did I make any mistakes? Were there any problems with the data?
  - Processes can break or misbehave in unforeseen ways
  - Results can be hard to understand, interpret and trust

data → knowledge → decisions

Incorrect conclusions can lead to bad decisions!

My Research

- Empower a *wide range of users* to explore and obtain *trustworthy*, *actionable insights* from data
  - Reduce bottlenecks and make humans more effective
  - Support responsible exploration
- Develop methods that address problems at the different stages of the data lifecycle
- Combine methods from different computer science areas and solve real problems
- Build systems ☺
- NYU Visualization, Imaging and Data Analysis (VIDA) Center

http://vida.engineering.nyu.edu

https://github.com/ViDA-NYU

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

My Research (last 4 years)

data

https://auctus.vida-nyu.org
[WWW 2019;
WSDM 2018]

Data Discovery and Search

[pVLDB 2020;
CIDR 2020;
SIGMOD 2019,
IEEE DEB 2016,
HILDA 2016]

Data Curation

[SIGMOD 2020;
AutoML 2019,
SIGMOD 2018,
pVLDB 2017,
CGF 2018,
ICDE 2016]

Data Management

[ACM TDS 2020,
[IEEE TVCG 2020;
AutoML 2019,
SIGMOD 2017,
SIGMOD 2016]

Data Analysis & Visualization

[IEEE TVCG 2020;
ICML AutoML 2019,
HILDA 2019]

Modeling and Learning

knowledge

Provenance and Reproducibility

[ACM Surveys 2019;
MSR 2019,
IEEE DEB 2018,
pVLDB 2017]

Debugging and Trust

[SIGMOD 2020;
DEEM 2019]

Urban Data Science

[IEEE Big Data 2019;
IEEE CG&A 2019;
IEEE TBD 2017,
IEEE Big Data 2016]

Misinformation

[AAAS 2020;
WWW 2020;
WWW 2019]

Scientific Reproducibility

[National Academies 2019]
https://github.com/gems-uff/noworkflow
https://www.vistrails.org
https://www.reprozip.org

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Talk Outline

- Usability and performance

Interactive exploration of spatio-temporal urban data

Data Management

Data Analysis & Visualization

- Reproducibility and trust

Provenance for data exploration and debugging

Provenance and Reproducibility

Debugging and Trust

Usability and Performance

Interactive exploration of spatio-temporal urban data

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

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%
- Growth leads to problems, e.g., transportation, environment and pollution, housing, infrastructure
- Good news: Lots of data being collected about many cities in the world

NYC OpenData london.gov.uk

RIO PREFEITURA Armazém de Dados INSTITUTO PEREIRA PASSOS

Stadt Zürich

Startseite Portal der Stadt Zürich

facebook

San Francisco Data

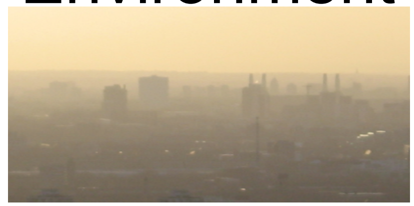
twitter

City of Chicago Data Portal

Understanding Cities

Infrastructure

Environment



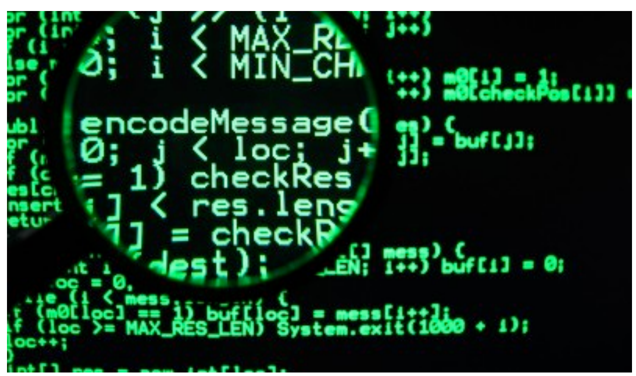
Meteorology, pollution

People

Opportunity:

Crack the codes of cities and understand how different components interact over space and time

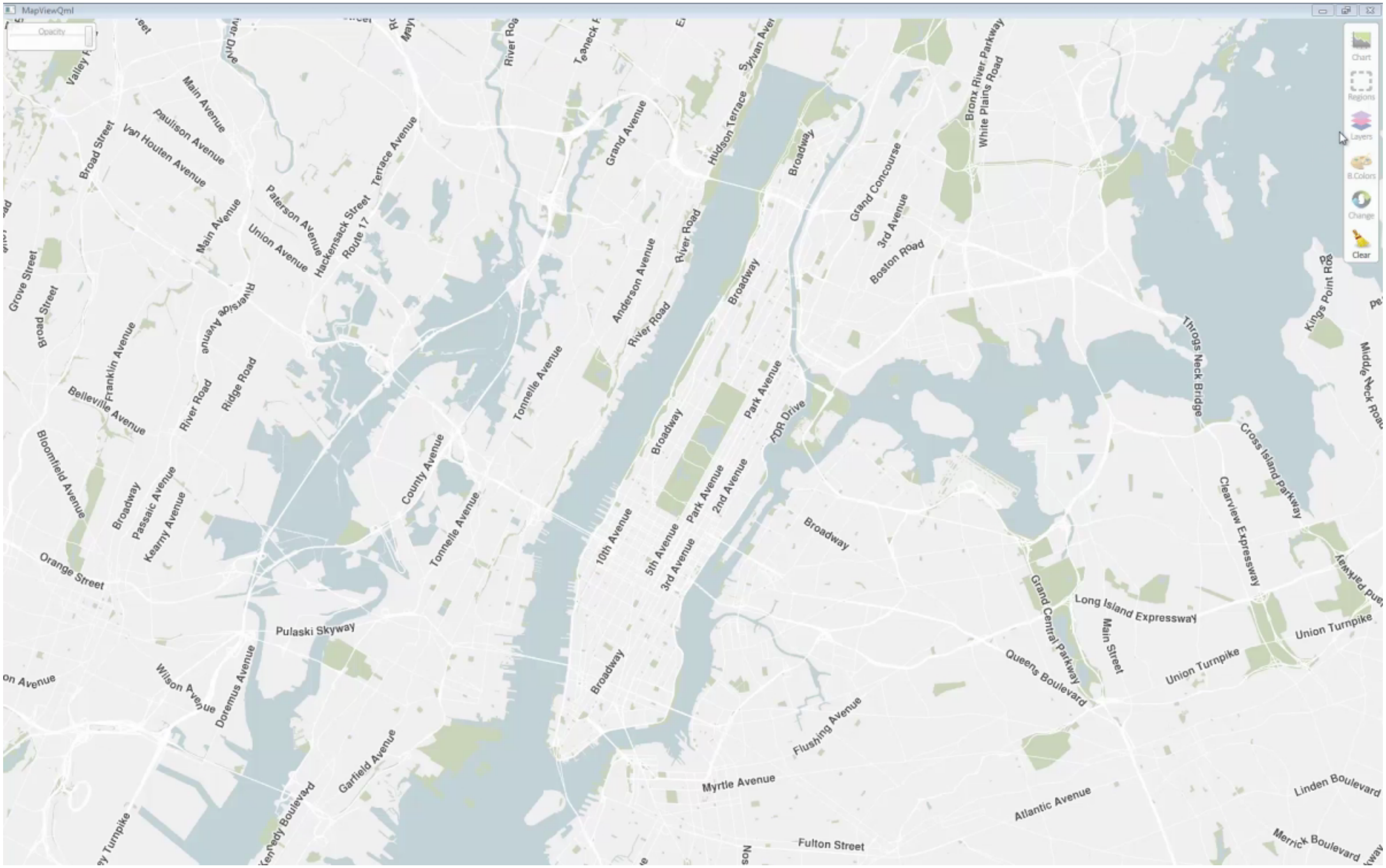
Use insights to make cities more efficient and sustainable, and improve the lives of their residents



NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

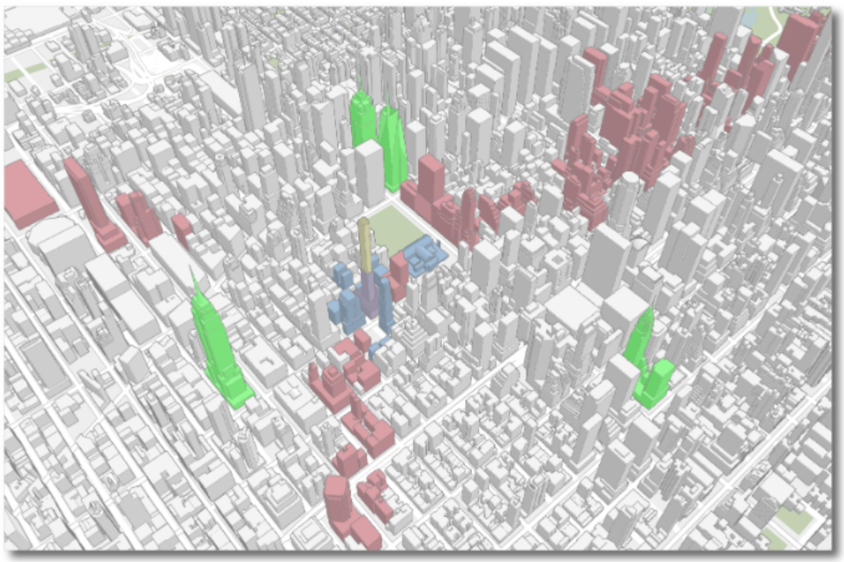
Urbane: Exploring Urban Data



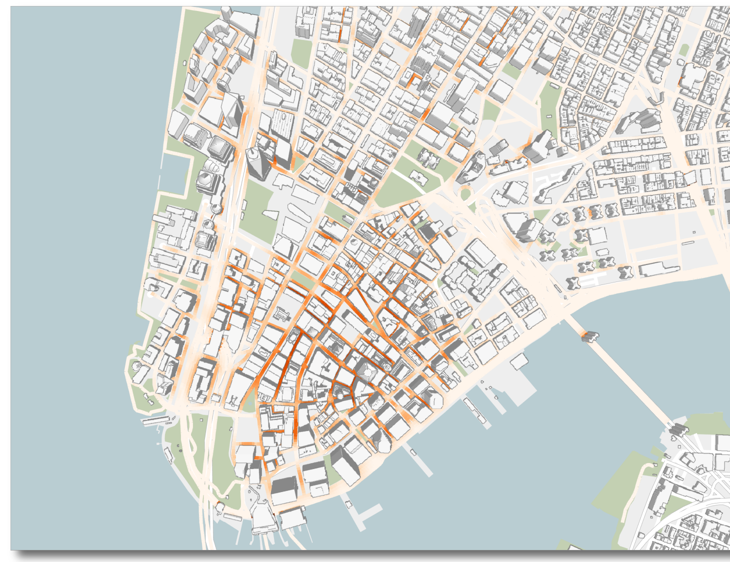
https://youtu.be/Y54pNn3BkA4

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

Urbane: Usability through Visual 3D Queries



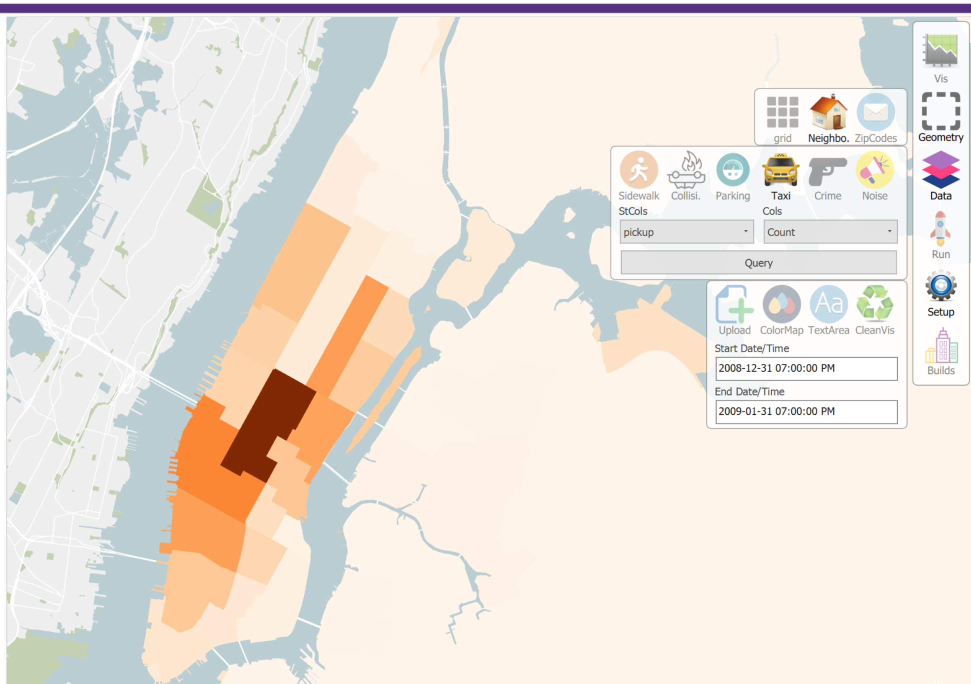
View Impact Queries



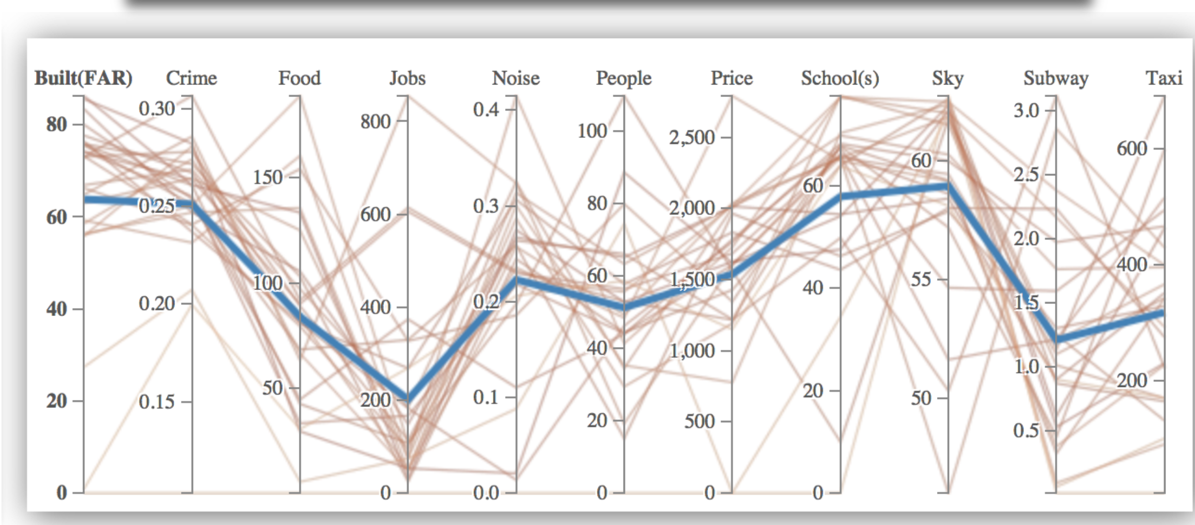
Sky Exposure Queries

[Ferreira et al., IEEE VAST 2015]

Urbane: Usability through Visual 2D Queries



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
INSIDE *N*.g
AND *C*.date in Janu
GROUP BY *N*.id

Food
Jobs
Noise
People
Price
Schools
Sky
…

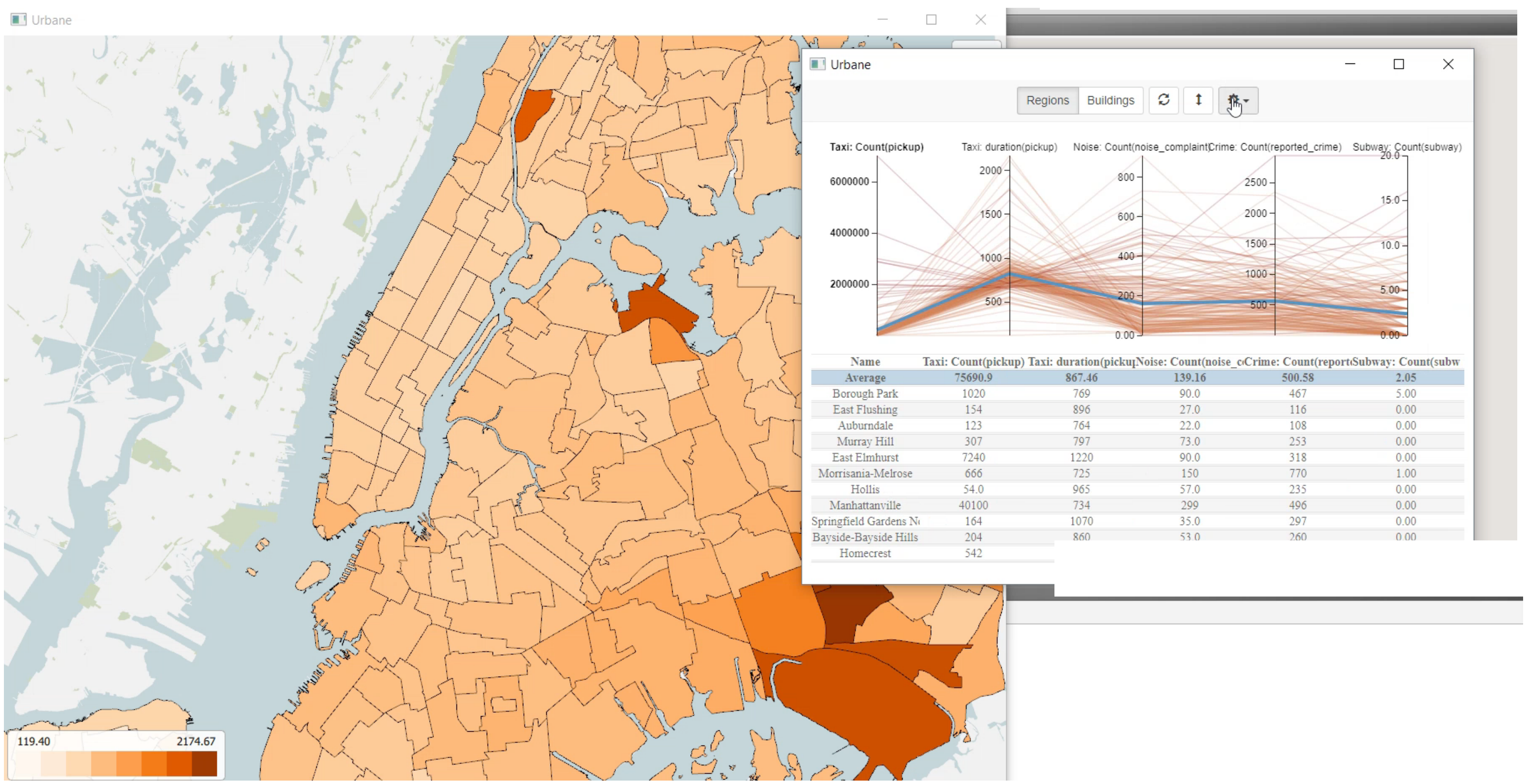
NYU TANDON SCHOOL OF ENGINEERING

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

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Challenge: Interactive Query Evaluation

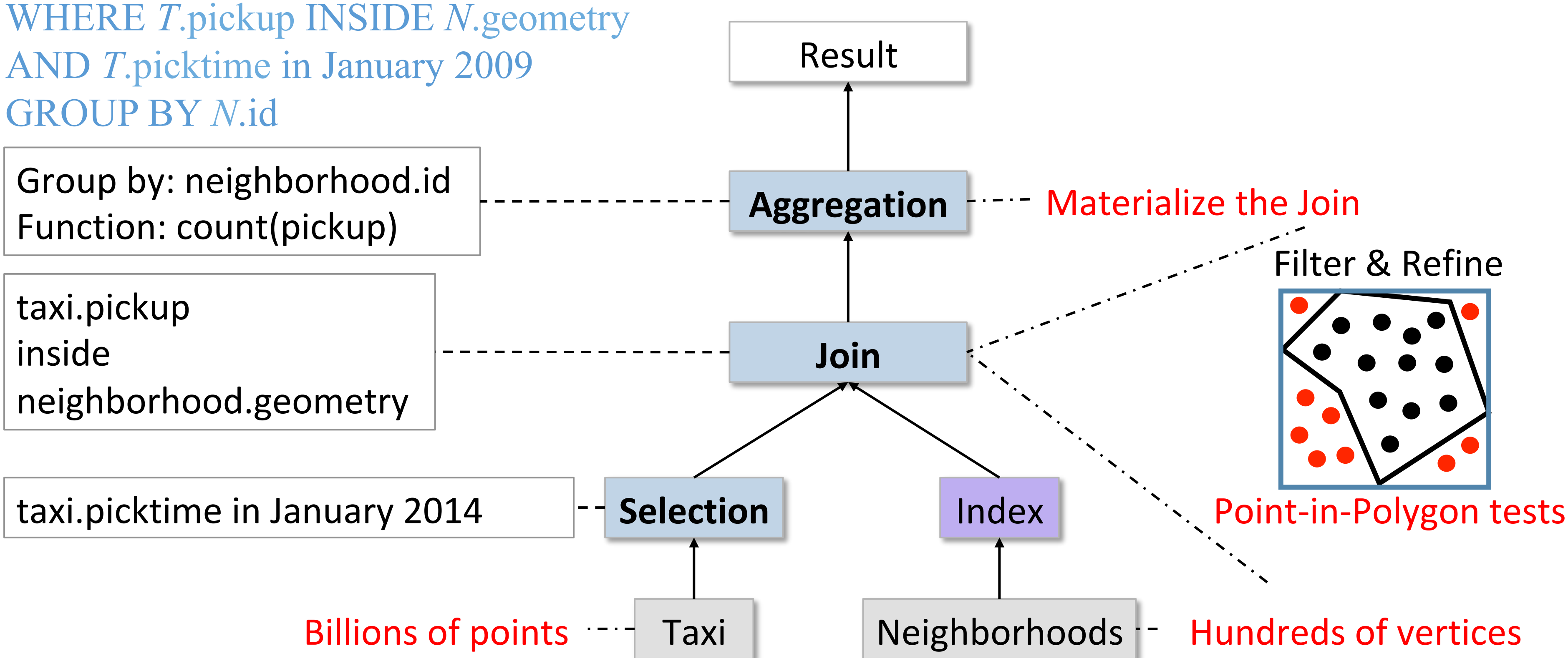
"increased latency reduces the rate at which users make observations, draw generalizations and generate hypotheses" [Liu and Heer, IEEE TVCG 2014]



High query rate

Spatial Aggregation

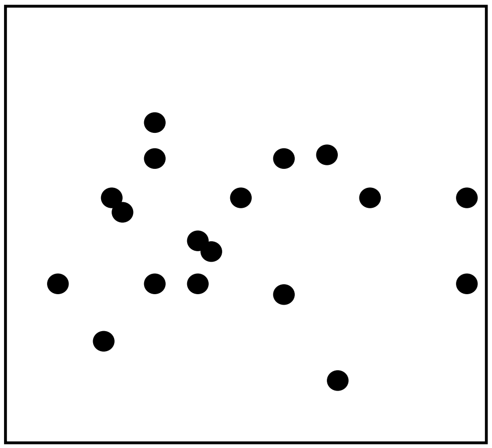
SELECT COUNT(*)
FROM taxi *T*, neighborhoods *N*
WHERE *T*.pickup INSIDE *N*.geometry
AND *T*.picktime in January 2009
GROUP BY *N*.id



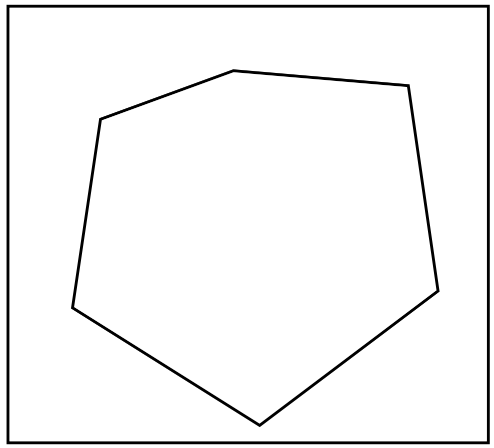
Several minutes on a commercial database

Spatial Aggregation: A Geometric Perspective

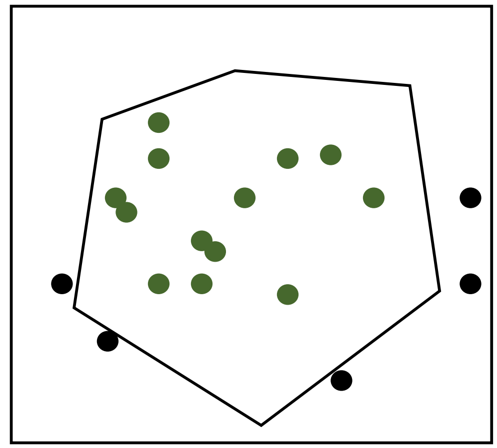
Spatial join = "Drawing" points and polygons on the same canvas



Input points



Input polygon

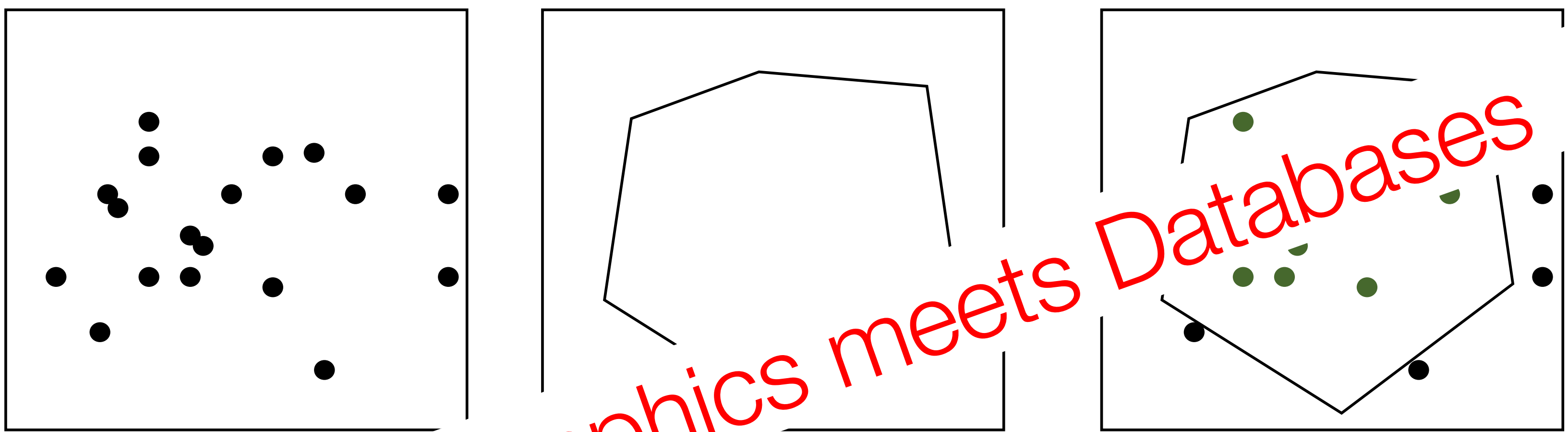


Spatial join

Leverage the graphics pipeline of the GPU

Spatial Aggregation: A Geometric Perspective

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



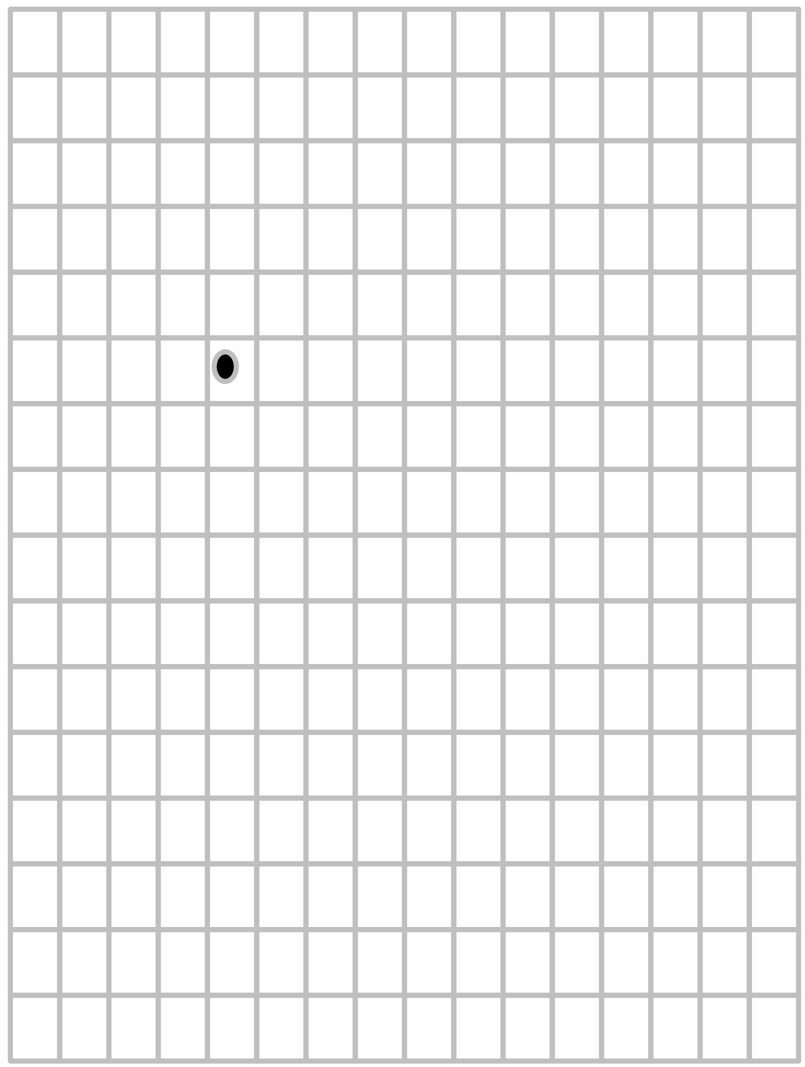
Input p   Input polygon   Spatial join

Computer Graphics meets Databases

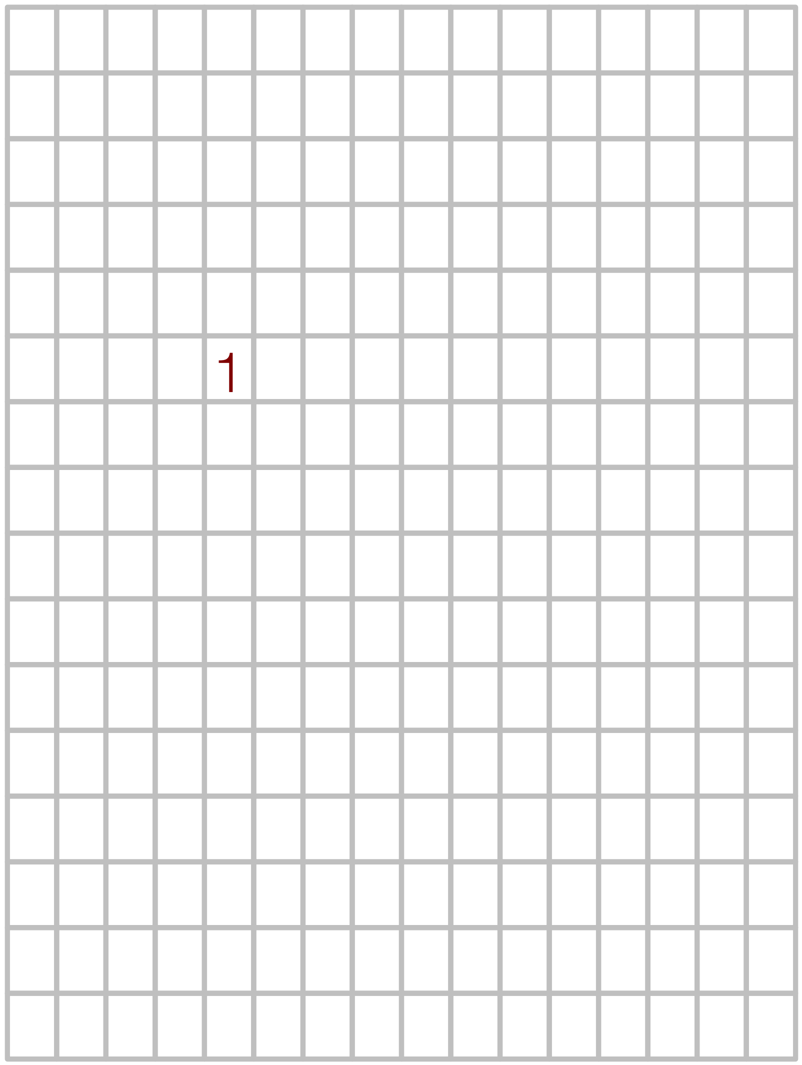
Leverage the graphics pipeline of the GPU

[Tzirita et al., PVLDB 2017]

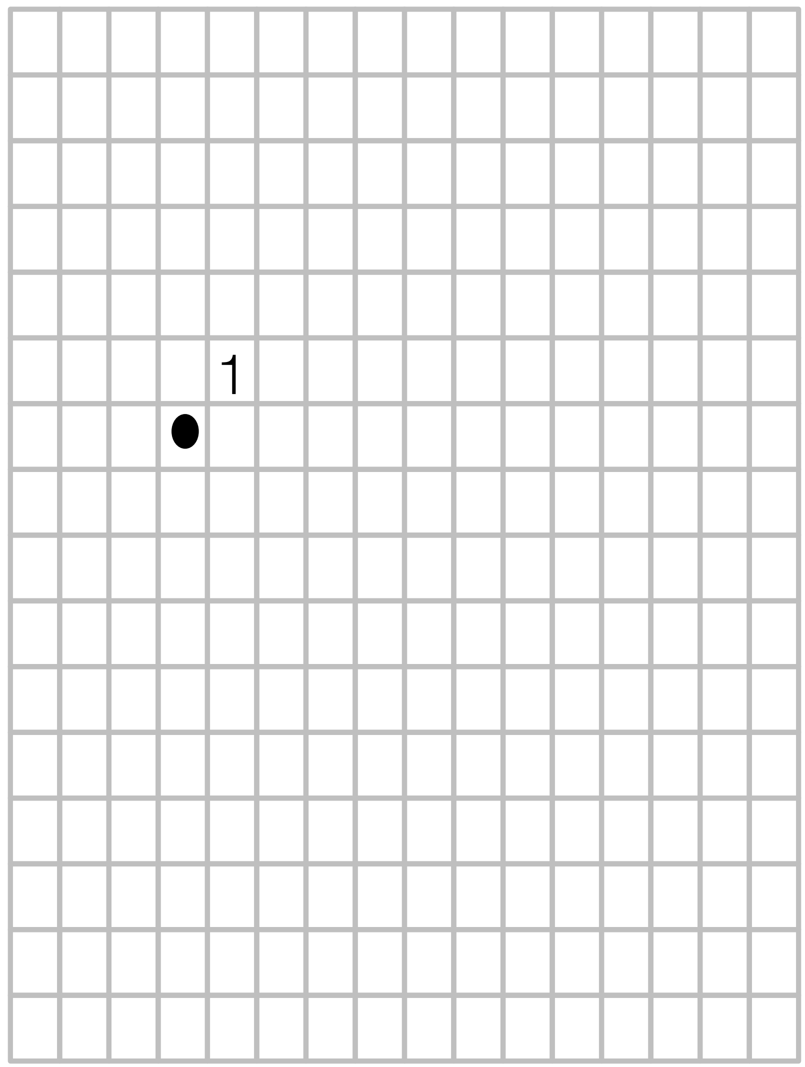
Raster Join: I. Draw the Points



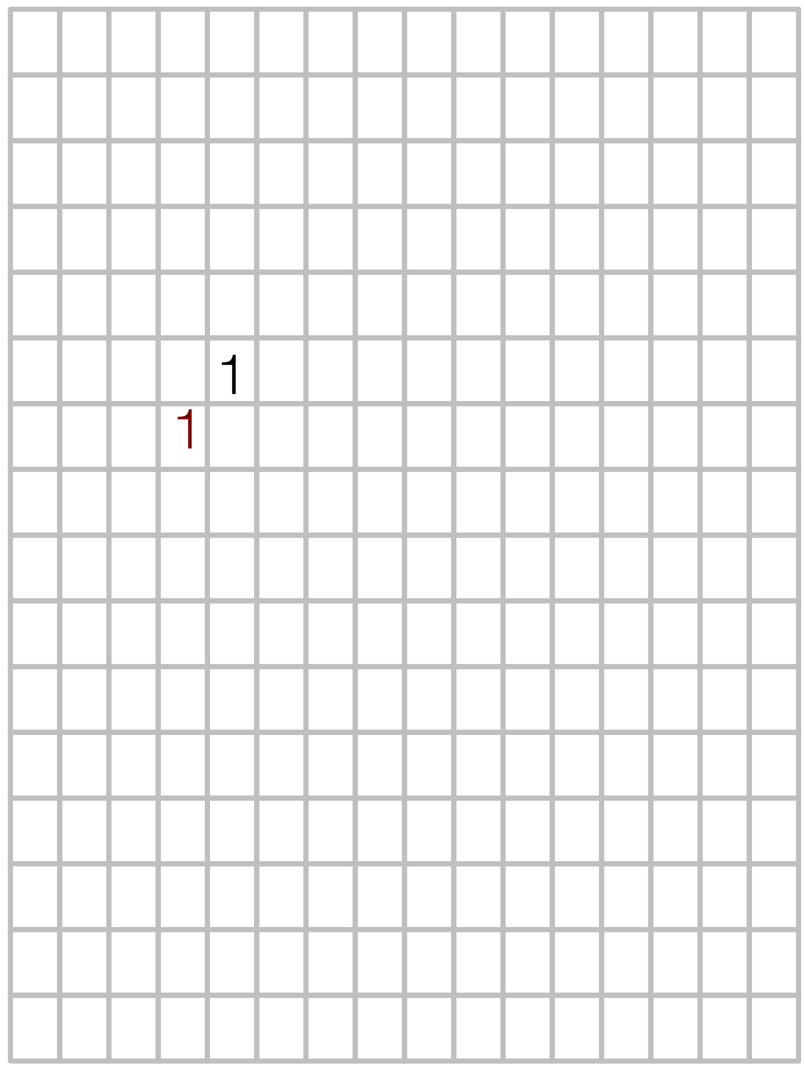
Raster Join: I. Draw the Points



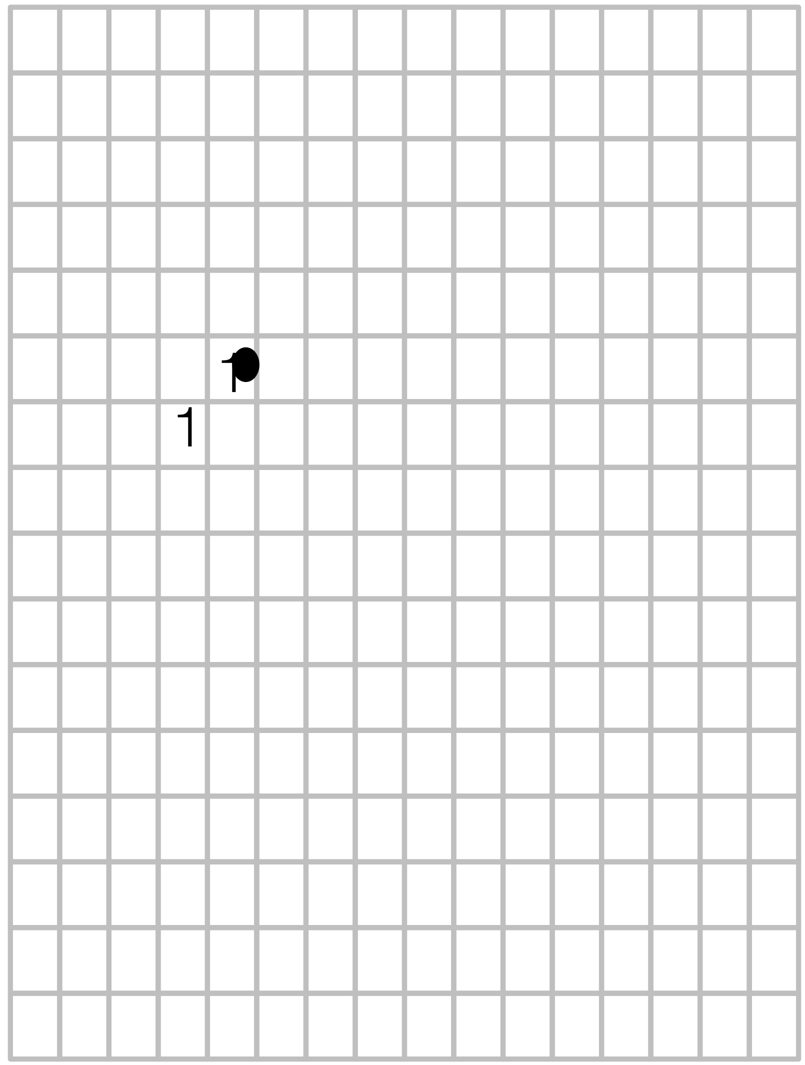
Raster Join: I. Draw the Points



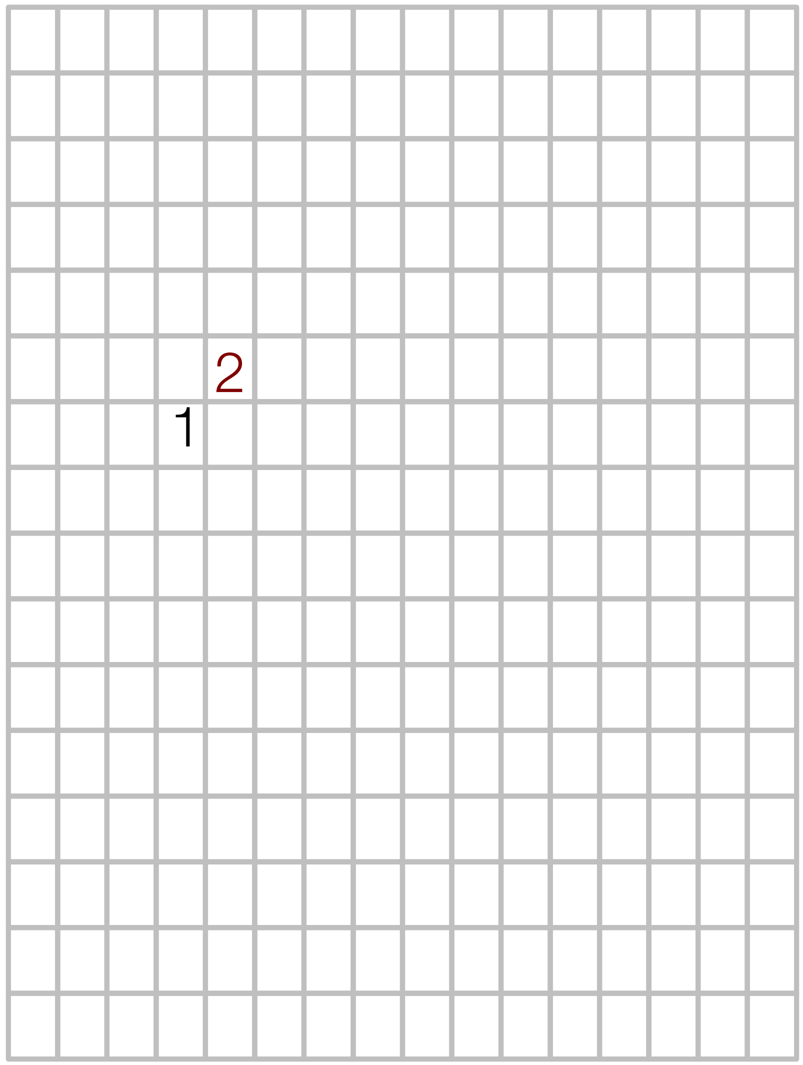
Raster Join: I. Draw the Points



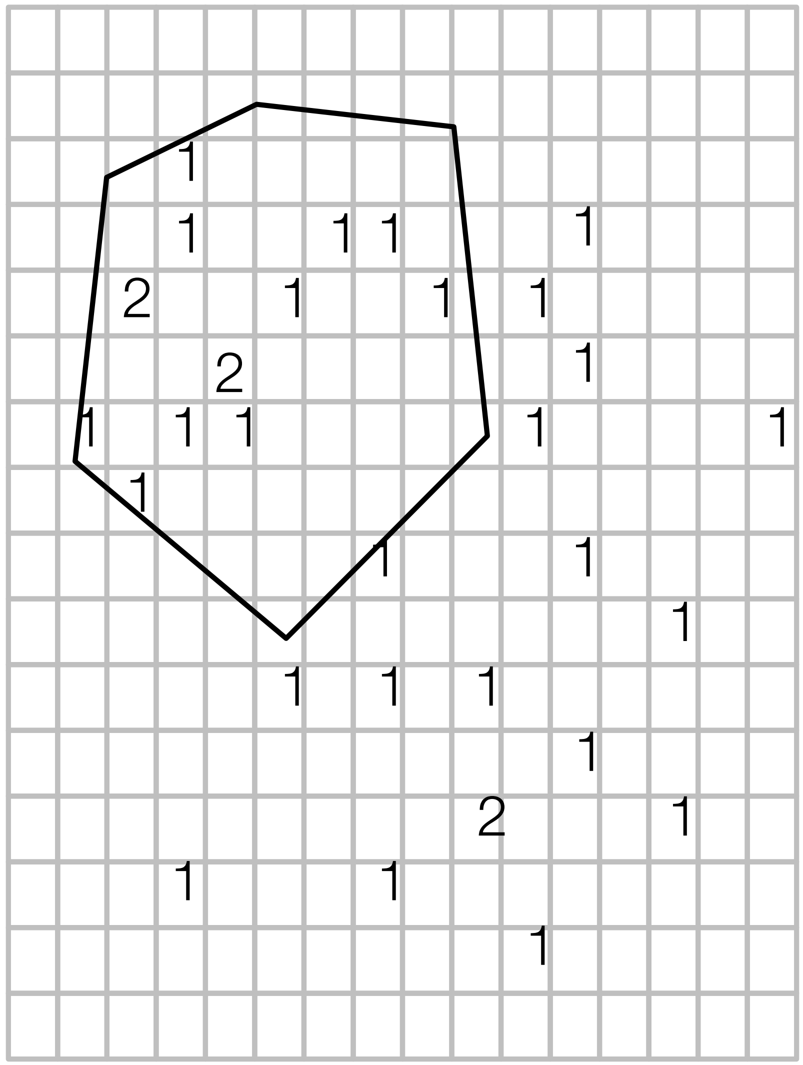
Raster Join: I. Draw the Points



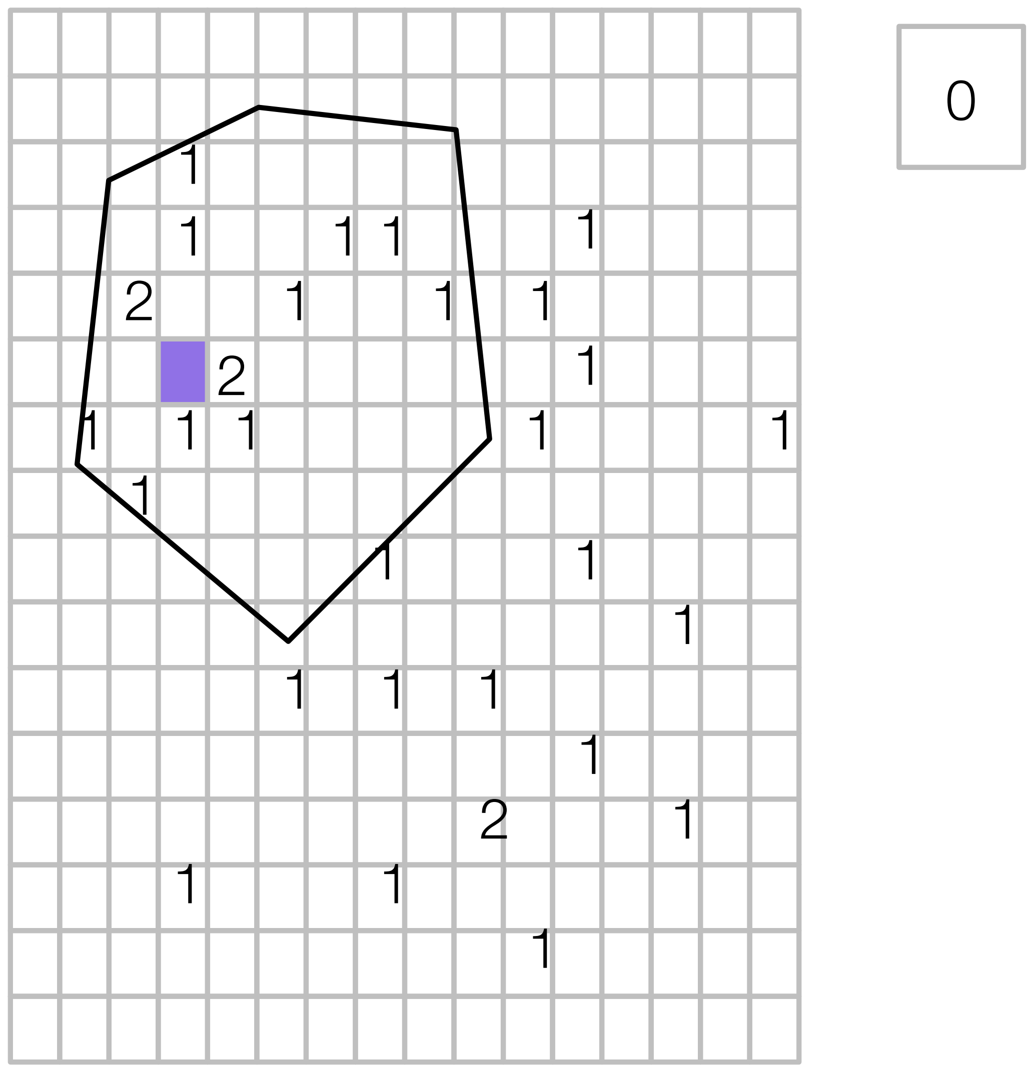
Raster Join: I. Draw the Points



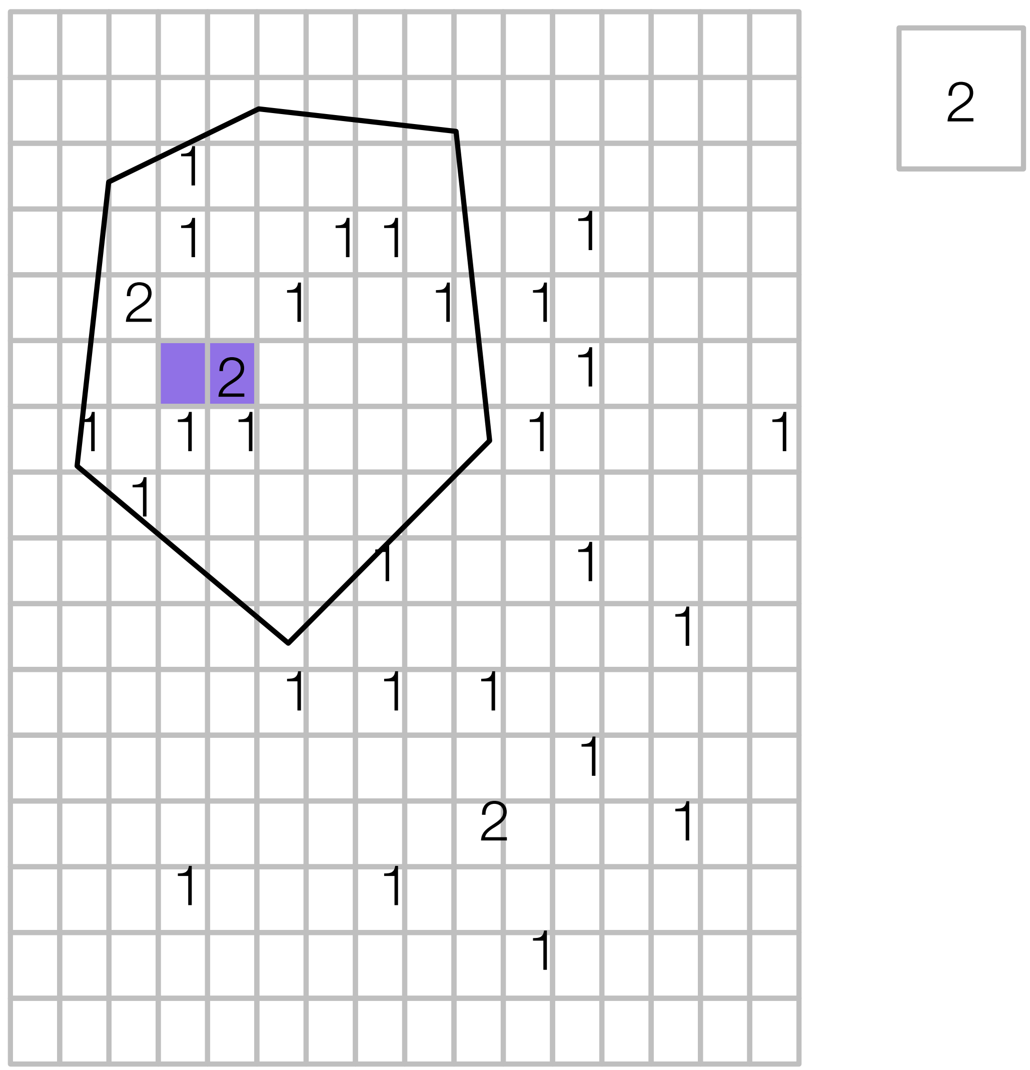
Raster Join: I. Draw the Points



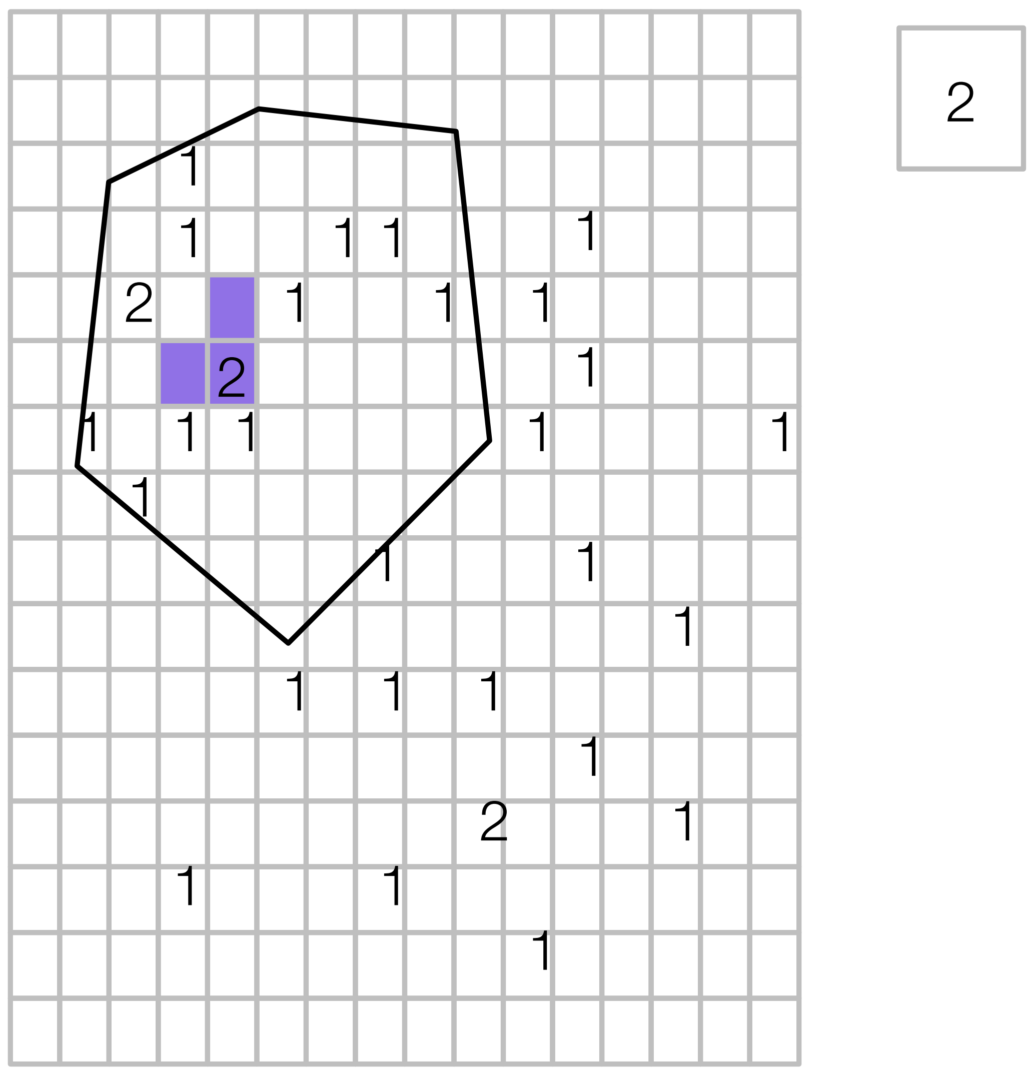
Raster Join: II. Draw the Polygons



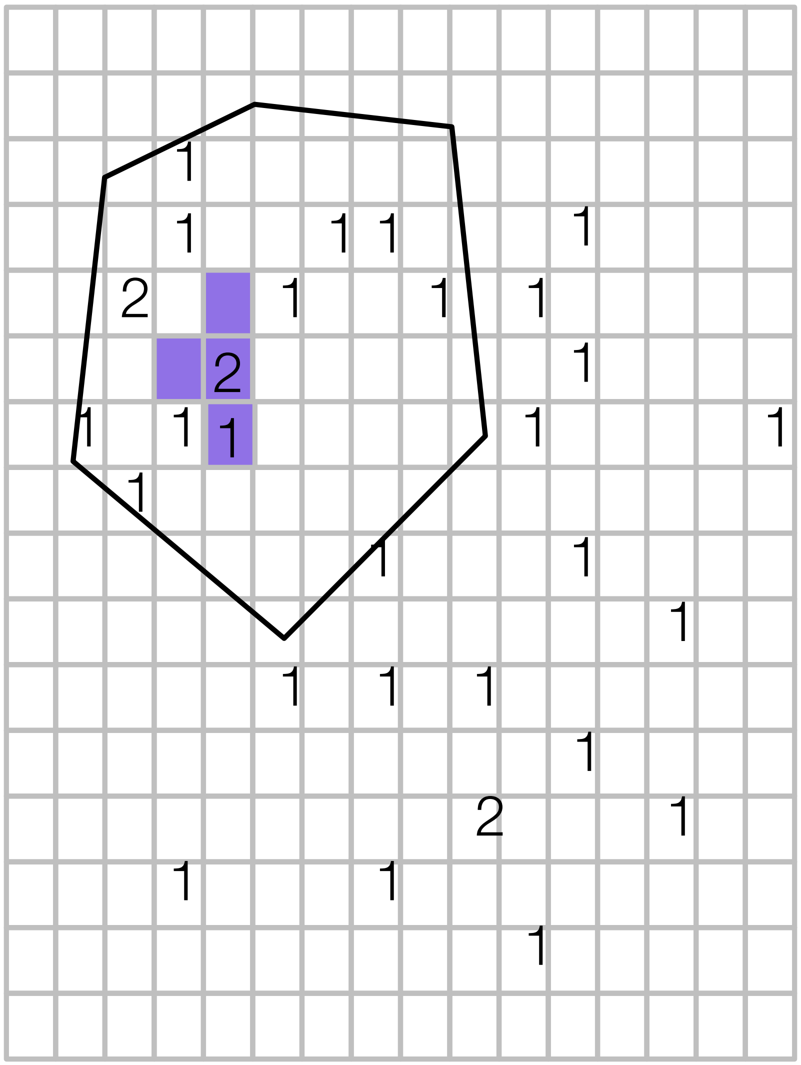
Raster Join: II. Draw the Polygons

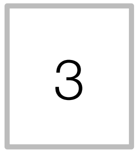


Raster Join: II. Draw the Polygons

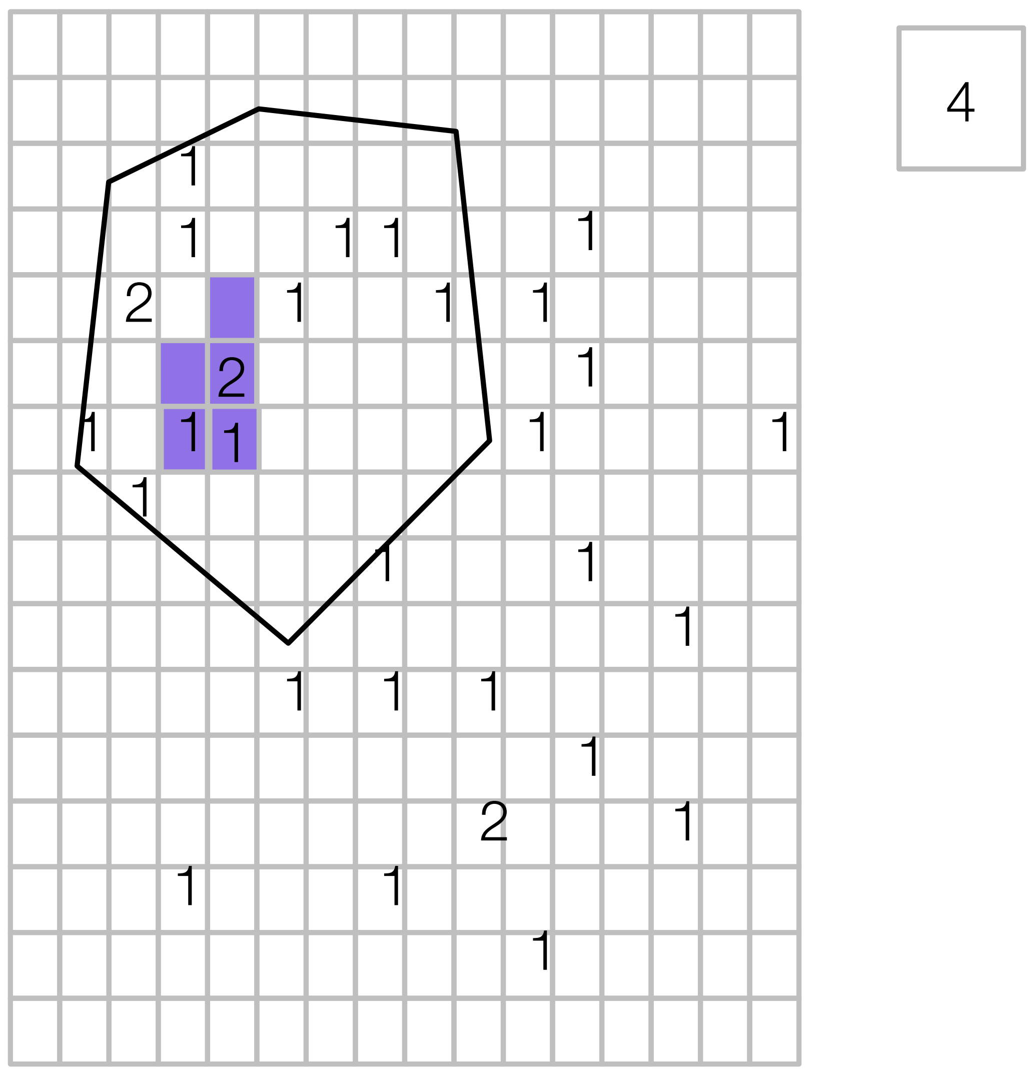


Raster Join: II. Draw the Polygons

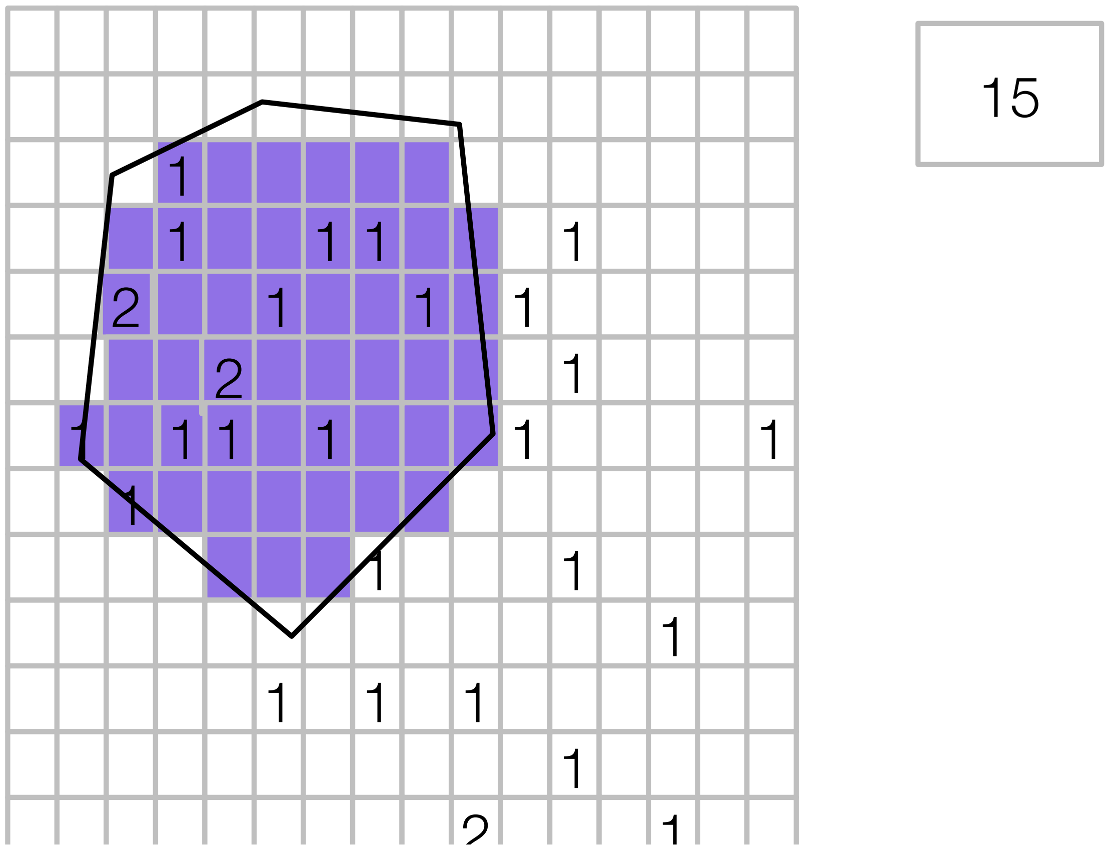




Raster Join: II. Draw the Polygons



Raster Join: II. Draw the Polygons



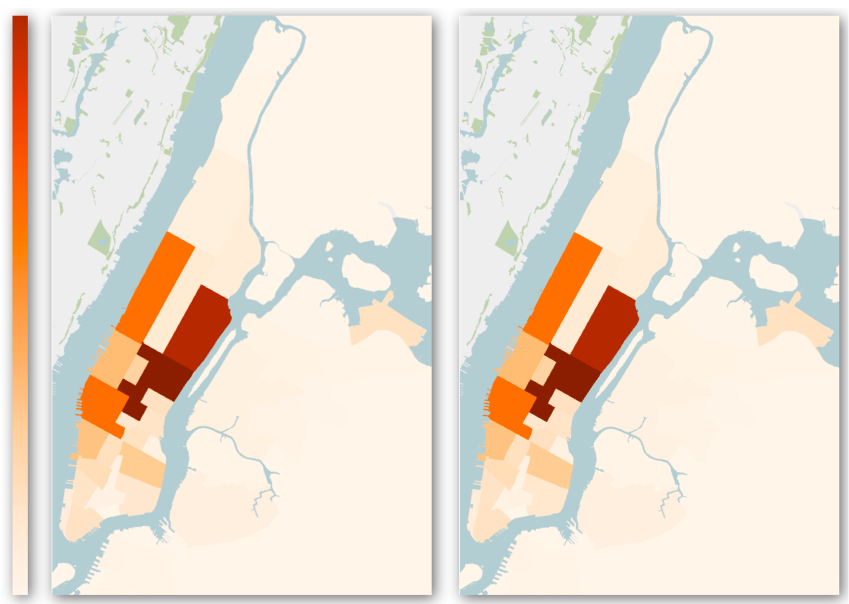
Exploits the native support for drawing in GPUs

Combines the aggregation with the join operation

No Point-in-Polygon tests

Raster Join: Some Notes

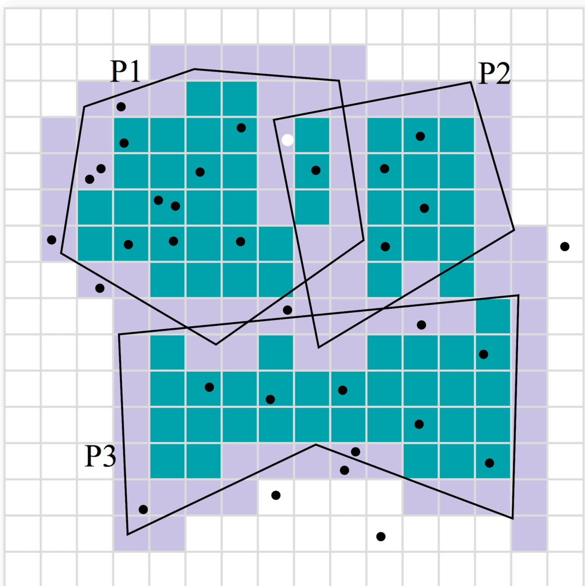
- Rasterization can introduce false positives and false negatives
- Errors can be reduced: better approximate the polygon outline by increasing the screen resolution (reducing pixel size)
- Bounded Raster Join: specify acceptable error
  - Hausdorff distance between actual and approximate polygon



approximate *accurate*

Raster Join: Some Notes

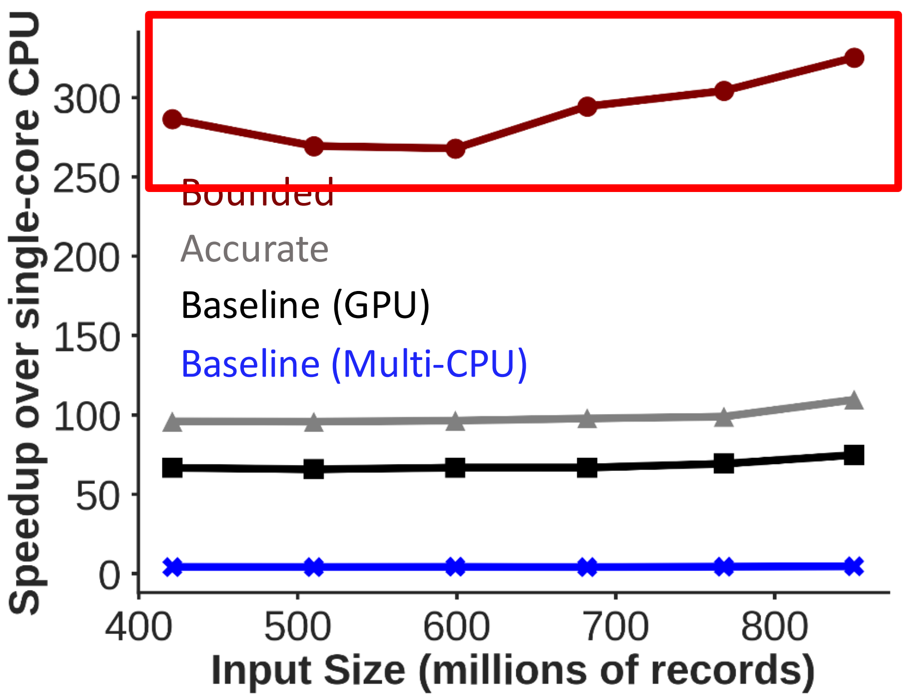
- Rasterization can introduce false positives and false negatives
- Errors can be reduced: better approximate the polygon outline by increasing the screen resolution (reducing pixel size)
- Bounded Raster Join: specify acceptable error
  - Hausdorff distance between actual and approximate polygon
- Accurate Raster Join: perform point-in-polygon tests for points in the boundary



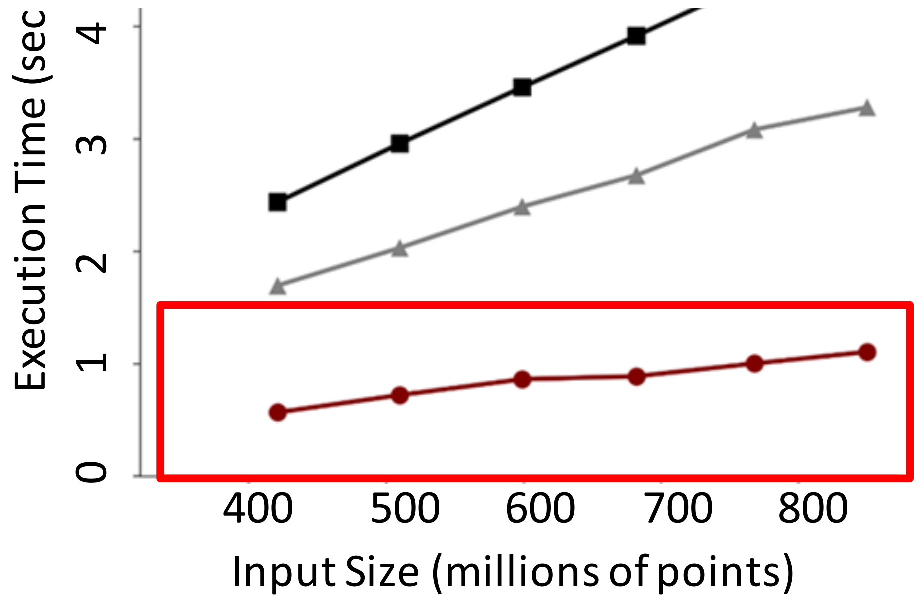
Performance Evaluation

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



300x speedup

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

[Tzirita et al., PVLDB 2017]

Take Away

- Need usability and performance for interactive exploration
  - Must combine techniques from Visualization, Computer Graphics, HCI, and data management [Doraiswamy et al., CG&A 2018; Chang et al., Dagstuhl 2018]
- GPUs can give you high-performance computing power on your laptop: You don't need big iron to analyze big data, you can do it on your **laptop**!
  - Spatial aggregation [Tzirita et al., pVLDB 2017], spatio-temporal keyword queries [Hoang-Vu et al., CIKM 2016], spatio-temporal selection queries [Doraiswamy et al., ICDE 2016]
  - And many others, e.g., [Bustos et al., ICCS 2006], [Zhang & Yu, BigSpatial 2012]

Take Away

- Need usability and performance for interactive exploration
  - Must combine techniques from Visualization, Computer Graphics, HCI, and data management [Doraiswamy et al., CG&A 2018; Chang et al., Dagstuhl 2018]
- GPUs can give you high-performance computing power on your laptop: You don’t need big iron to analyze big data, you can do it on your **laptop**!
  - Spatial aggregation [Tzirita et al., pVLDB 2017], spatio-temporal keyword queries [Hoang-Vu et al., CIKM 2016], spatio-temporal selection queries [Doraiswamy et al., ICDE 2016]
  - And many others, e.g., [Bustos et al., ICCS 2006], [Zhang & Yu, BigSpatial 2012]
- Developing code for GPUs is difficult
  - Many one-off solutions – difficult to re-use across queries and geometric objects (e.g., points, polygons)
  - Hardware-specific implementations

Source: NVIDIA

Source: AMD

Can we abstract away this complexity?

Can we build a *unified* database engine
that leverages the GPU
to support a rich set of spatial queries?

GPU-Friendly Data Model and Algebra

Key Idea: Use the computer graphics pipeline and translate spatial queries into drawing operations

- Canvas as uniform data representation for geometric objects
  - Points, lines, and polygons
  - Easy to represent on the GPU
- Algebra consists of set of composable operators
  - Optimized for the GPU
  - Operators work on canvas objects
  - Allows use across different spatial query types
  - Support common spatial queries, e.g., spatial queries such as selections, joins, aggregations, and knn queries.
- GPU hardware agnostic: graphics pipeline supported on all GPUs

GPU-Friendly Data Model and Algebra

Key Idea: Use the computer graphics pipeline and translate spatial queries into drawing operations

-

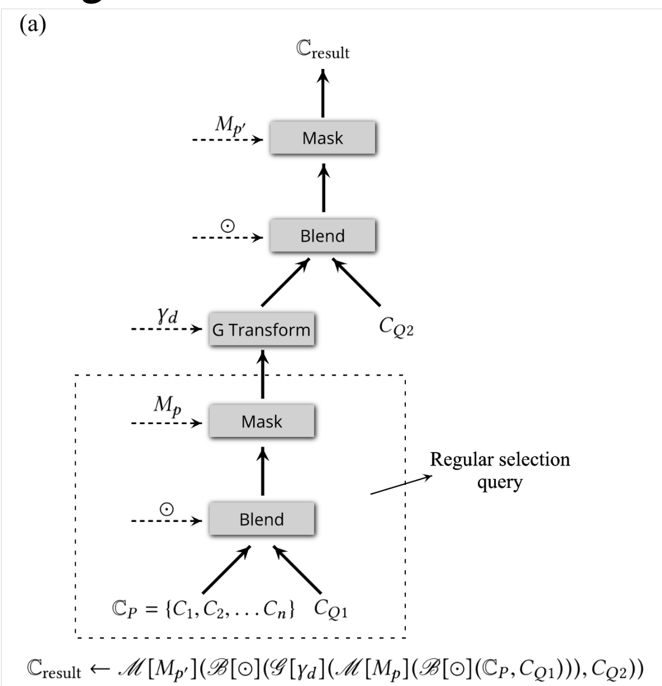
Algebra for developers: use operators to implement spatial queries

-
  - Allows use across different spatial query types
  - Support common spatial queries, e.g., spatial queries such as selections, joins, aggregations, and knn queries.
- GPU hardware agnostic: graphics pipeline supported on all GPUs

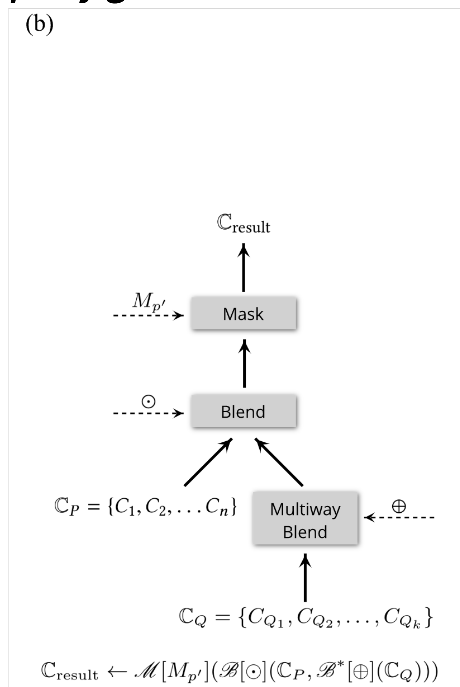
[Doraiswamy & Freire, SIGMOD 2020]

One algebra, many queries

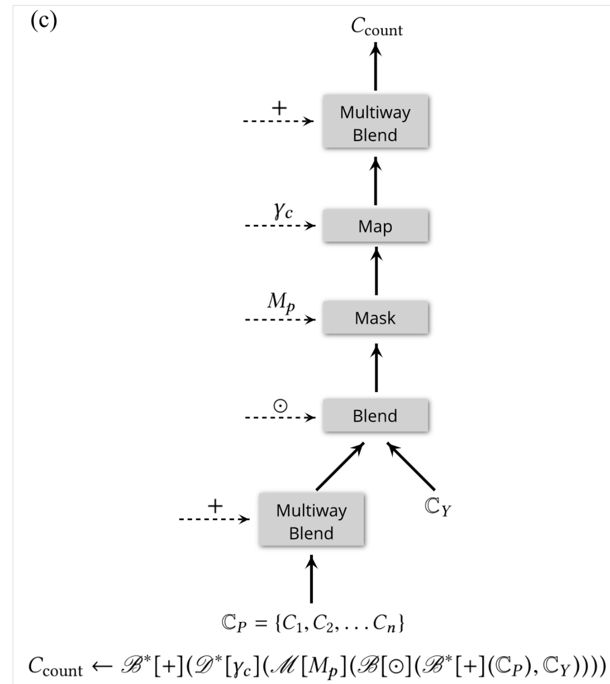
Selection over origin-destination data



Selection with multiple polygonal constraints



RasterJoin



Canvas Representation

Algebraic Operators

Blend, Mask, Transform, ...

Source: NVIDIA

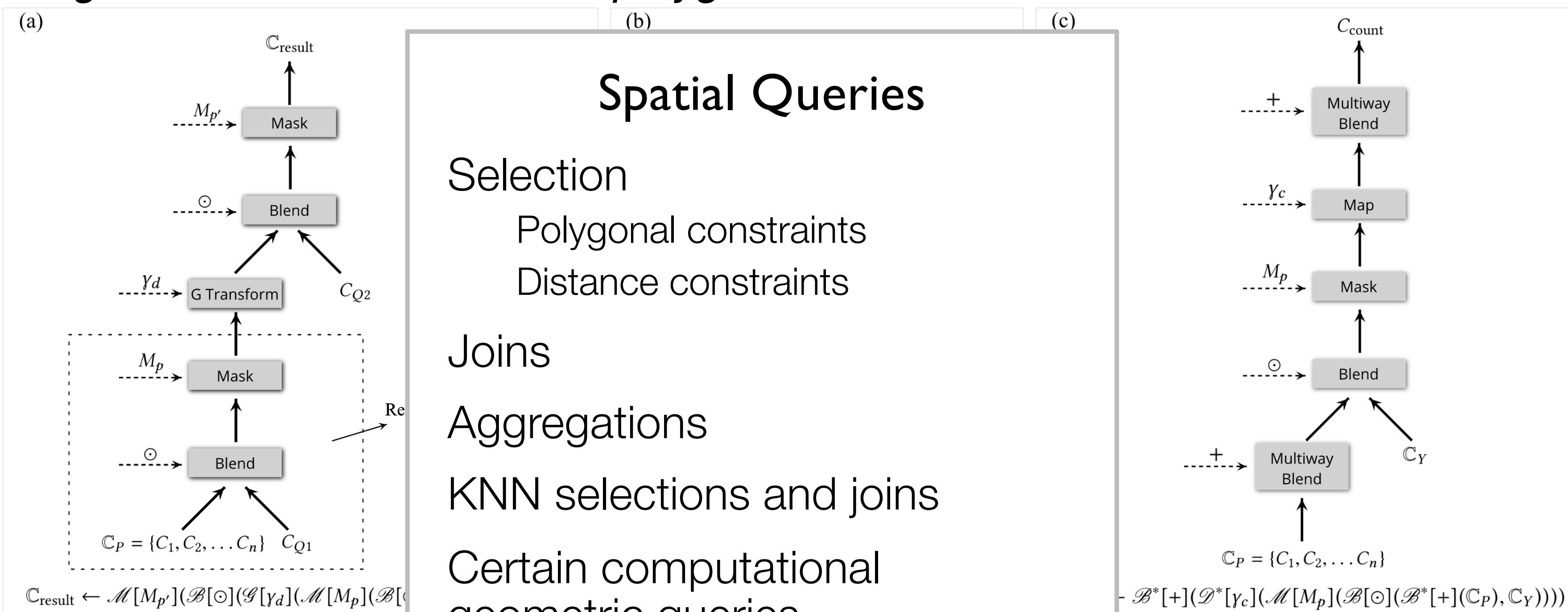
Source: AMD

One algebra, many queries

Selection over origin-destination data

Selection with multiple polygonal constraints

RasterJoin



Spatial Queries

Selection
- Polygonal constraints
- Distance constraints

Joins

Aggregations

KNN selections and joins

Certain computational geometric queries

No need to write GPU code!


Source: NVIDIA


Source: AMD

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

GPU-Powered Spatial Database Engine

- Preliminary evaluation

Spatial-algebra running on a laptop

Intel Core i7-8750H processor, 16 GB memory, 512 GB SSD, and a 2 TB external SSD connected via a Thunderbolt 3 interface

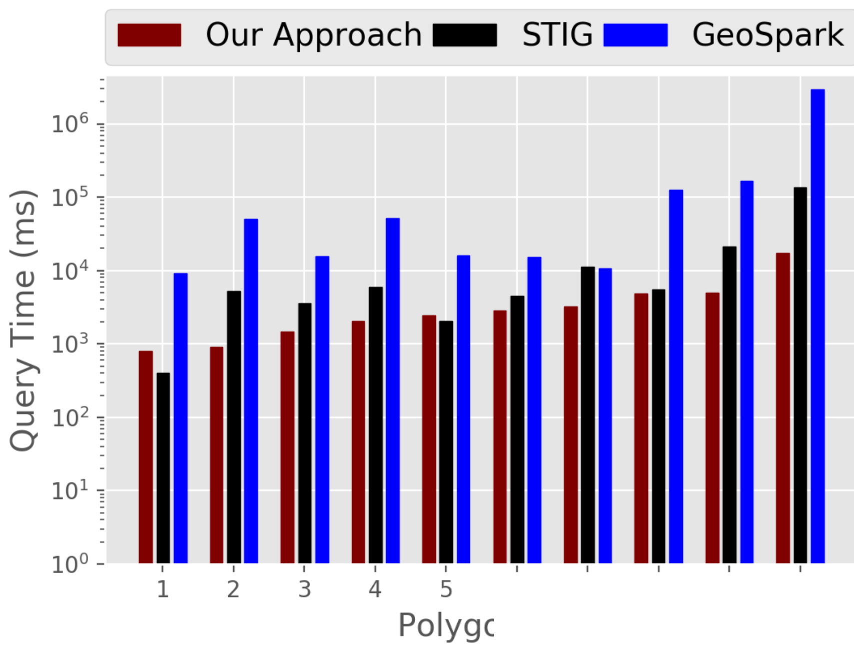
vs.

GeoSpark [Yu et al., 2015] running on a 17-node cluster

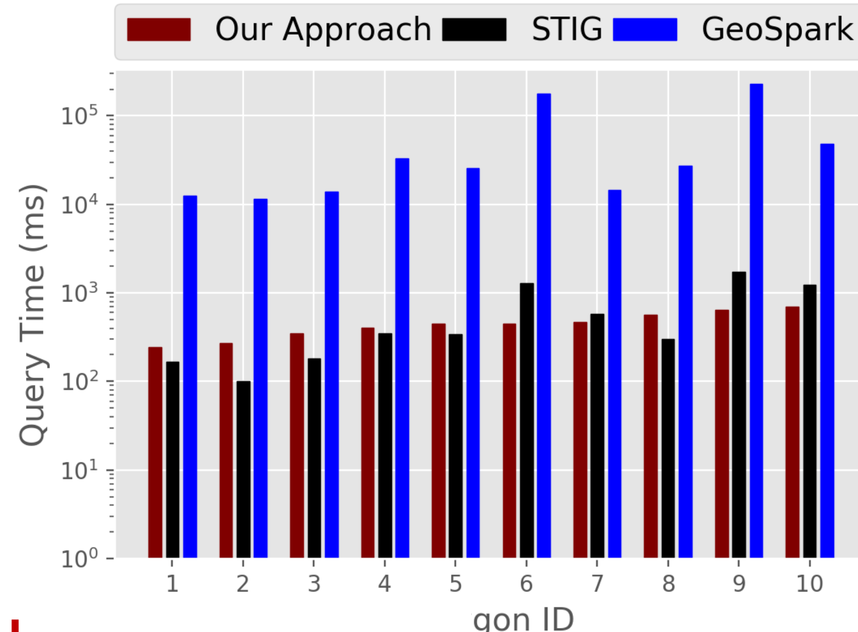
Cluster with 17 compute nodes, each node having 256 GB of RAM and 4 AMDOpteron(TM) 6276 processors running at 2.3GHz

[Doraiswamy & Freire, in preparation]

Performance Evaluation: Spatial Selection

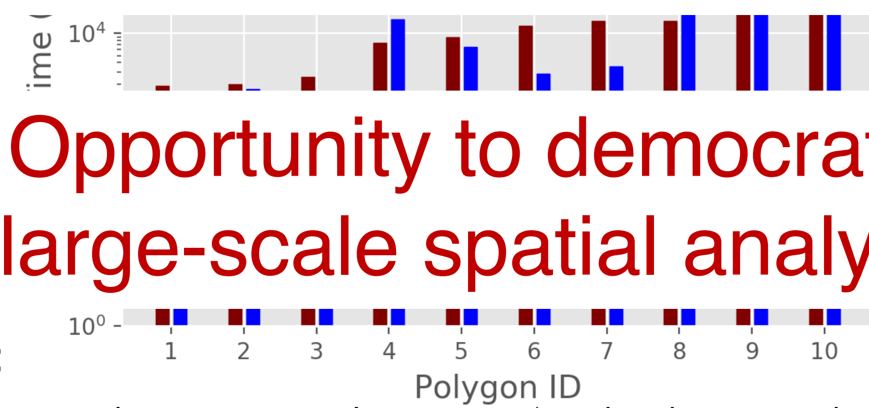


1.22 B Taxi trips (points) ir
195 NYC neighborhoods



.28B Tweets (points) in
,109 US counties

speedup ranging from 2X to as high as 108X



Opportunity to democratize large-scale spatial analytics

ildings (polygons)
le (OpenStreetMap)

NYU TANDON SCHO OF ENGINEER

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

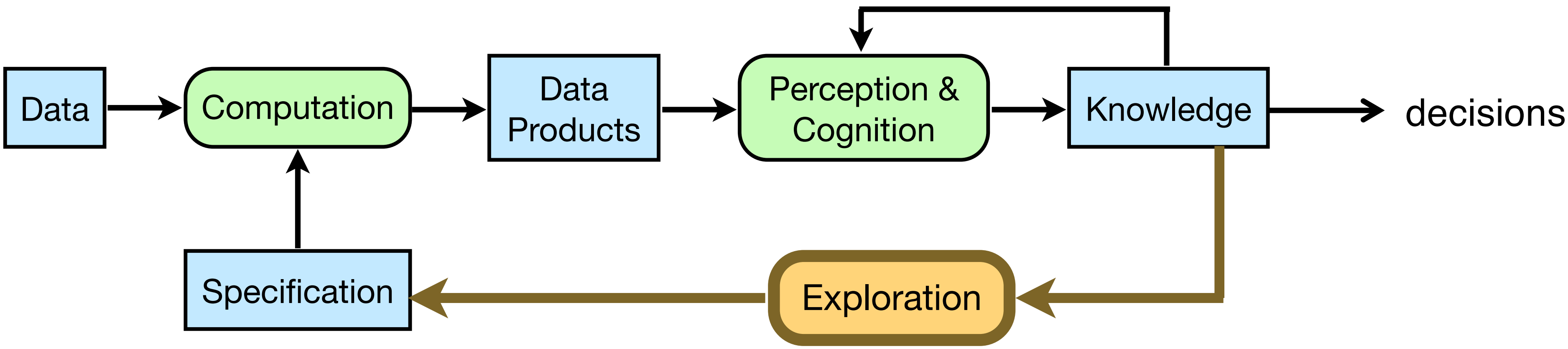
Reproducibility and Trust

Provenance for data exploration and debugging

NYU TANDON SCHOOL OF ENGINEERING

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Provenance for Data-Driven Exploration



[Modified from Van Wijk, Vis 2005]

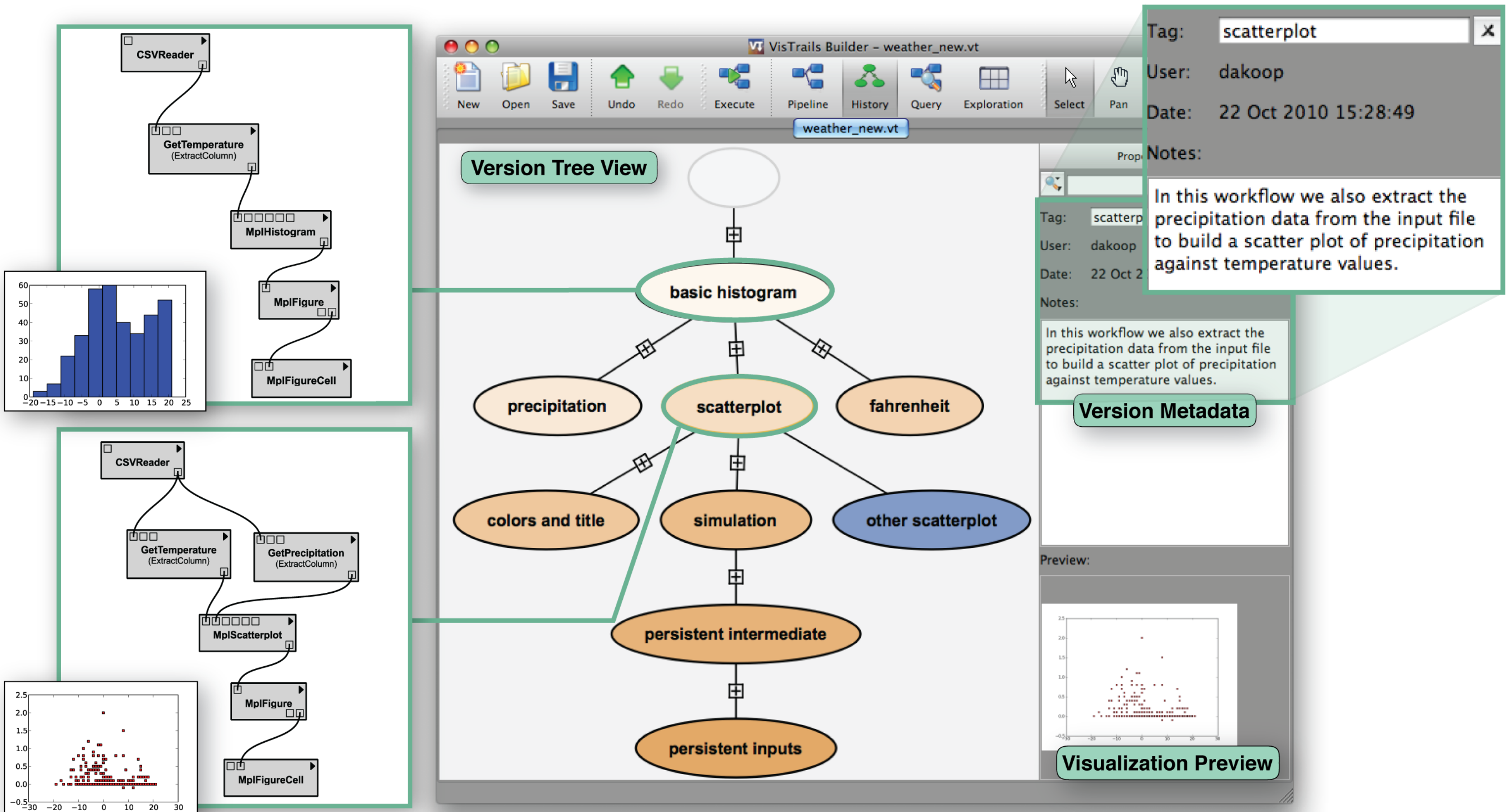
- Many things can go wrong…
- To assess validity and build trust in the results, we need detailed provenance of the **exploration process**
  - Reason about the steps followed
  - Reproduce and debug these steps

What to capture?

How to capture provenance?

Provenance Research and Systems

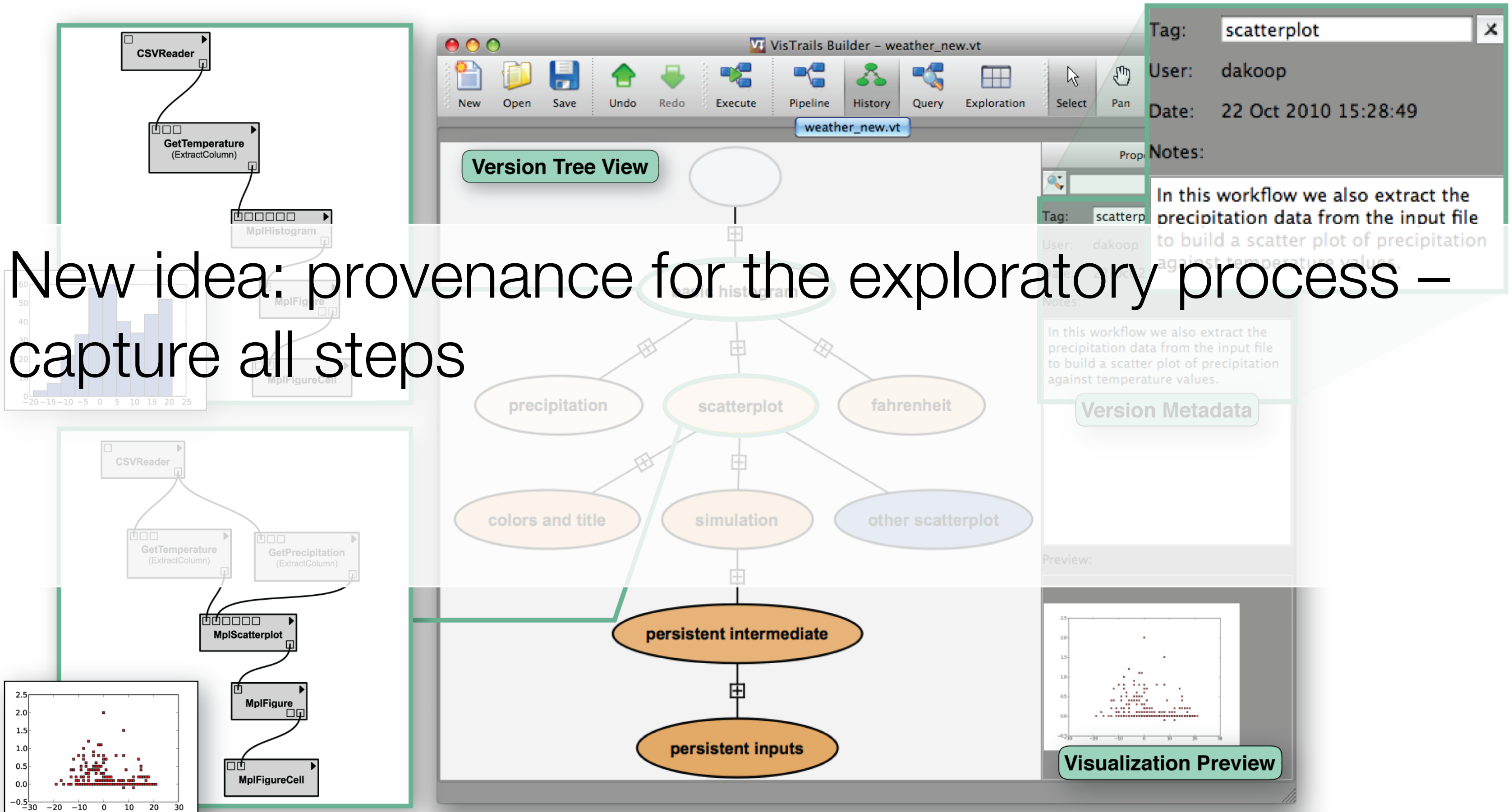
VisTrails: open-source system that supports data exploration through workflows: https://github.com/VisTrails



[Bavoil et al., IEEE Vis 2005, Freire et al., IPAW 2006; Koop et al., AOSA 2011]

Provenance Research and Systems

VisTrails: open-source system that supports data exploration through workflows: https://github.com/VisTrails



- New idea: provenance for the exploratory process – capture all steps

[Bavoil et al., IEEE Vis 2005, Freire et al., IPAW 2006; Koop et al., AOSA 2011]

Change-Based Provenance: Summary

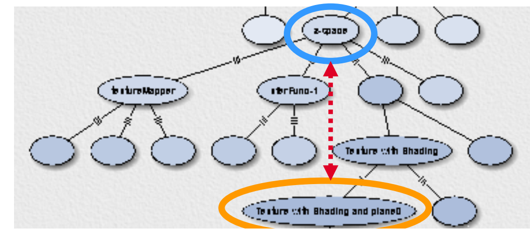
- General: Works with any system that has undo/redo
- Uniformly captures data and workflow provenance
  - Data provenance: where does a specific data product come from?
  - Workflow evolution: how has workflow structure changed over time?
- Results can be reproduced
- Provenance can be (re-)used to streamline exploration
  - Support for reflective reasoning, collaboration, recommendations, and many more [Freire & Silva, CISE 2012; Santos et al., IPAW 2012; Koop et al., ICCS, 2011; Silva et al., CGF 2011; Bauer et al., JSTAT 2011; Koop et al., SSDBM 2010; Santos et al., IEEE Vis 2009; Koop et al., IEEE TVCG 2008]

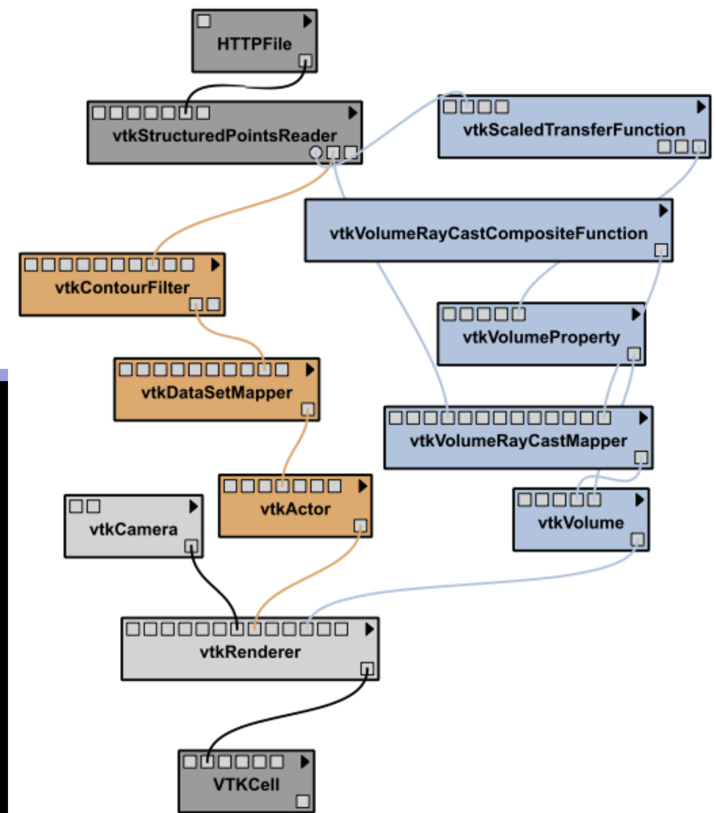
[Freire et al, IPAW 2006]

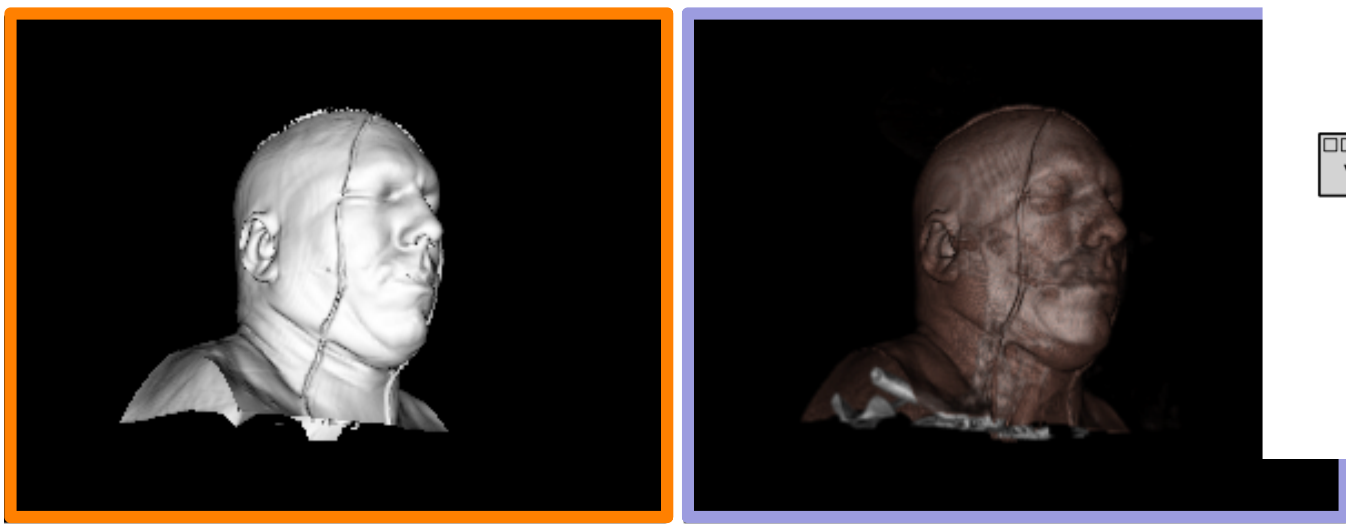
Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products

$$
\begin{aligned}
vt_1 &= x_i \circ x_{i-1} \circ \ \ldots\ \circ x_1 \circ \emptyset \\
vt_2 &= x_j \circ x_{j-1} \circ \ \ldots\ \circ x_1 \circ \emptyset \\
vt_1 - vt_2 &= \{x_i, x_{i-1}, \ldots, x_1, \emptyset\} \\
&\quad - \{x_j, x_{j-1}, \ldots, x_1, \emptyset\}
\end{aligned}
$$





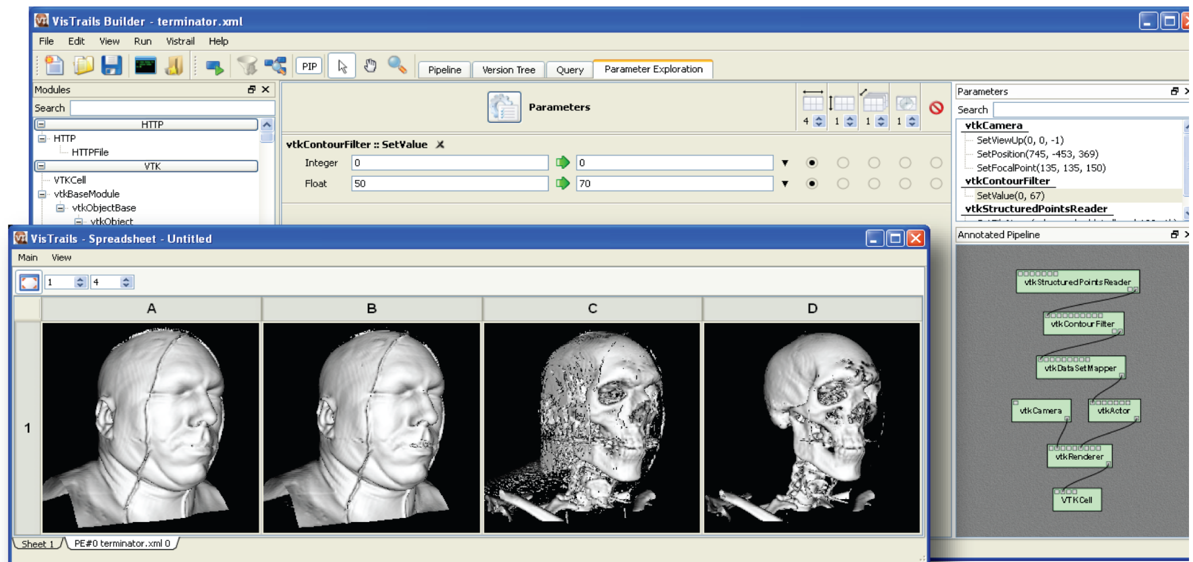


[Freire et al., IPAW 2006]

Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results
  - Also explore alternative computations

$$(setParameter(id_n, value_n) \circ \ldots \circ (setParameter(id_1, value_1) \circ v_t)$$



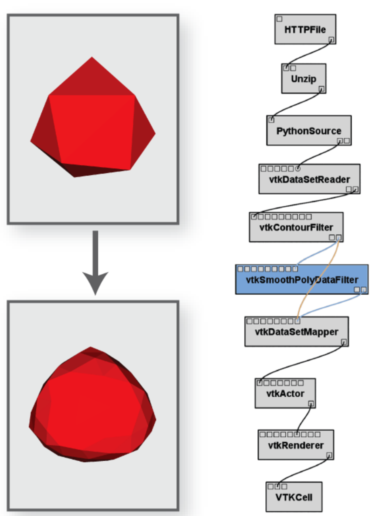
$$(addModule(id_i, \ldots) \circ (deleteModule(id_j) \circ v_1)$$
$$\ldots$$
$$(addModule(id_i, \ldots) \circ (deleteModule(id_j) \circ v_n)$$

NYU TANDON SCHOOL OF ENGINEERING

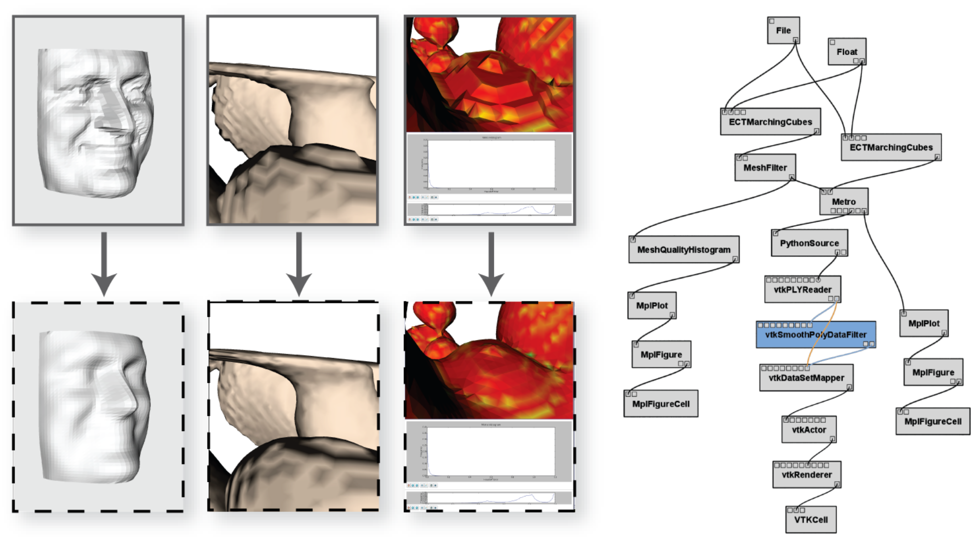
VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results
  - Also explore alternative computations
- Knowledge re-use: refine analyses by analogy



Analogy Template



Automatically constructed visualizations

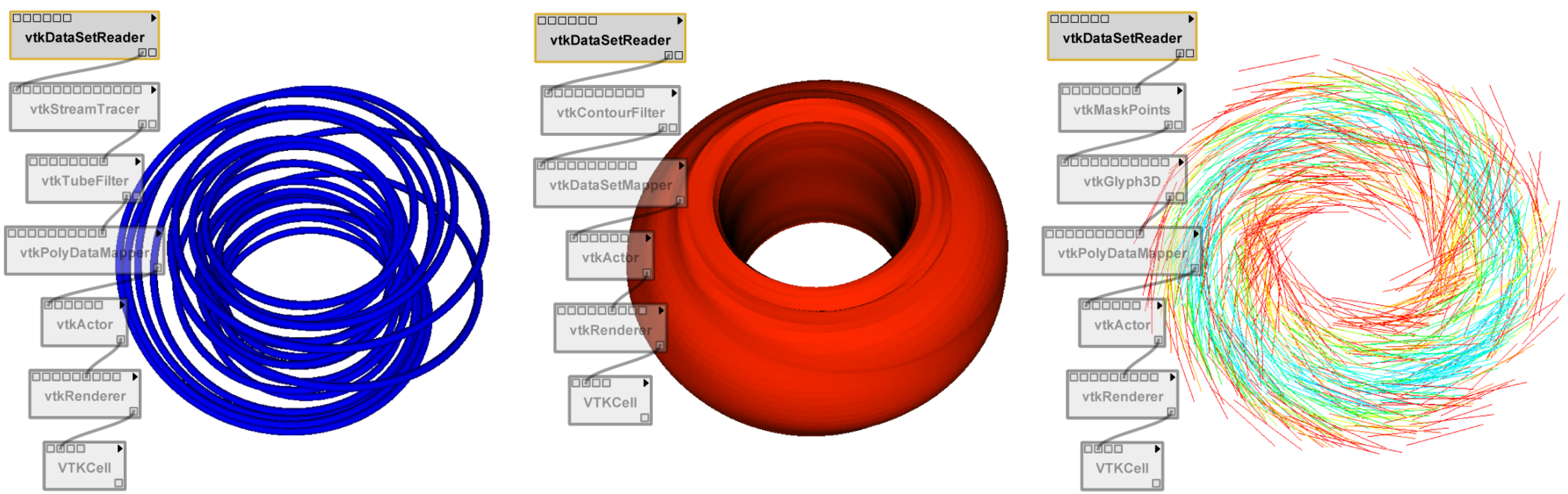
NYU TANDON SCHOOL OF ENGINEERING

[Scheidegger et al, IEEE TVCG 2007]

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

VisComplete: Workflow Recommendations

- Mine provenance collection: Identify graph fragments that co-occur in a collection of workflows
- Predict sets of likely workflow additions to a given partial workflow
- Similar to a Web browser suggesting URL completions

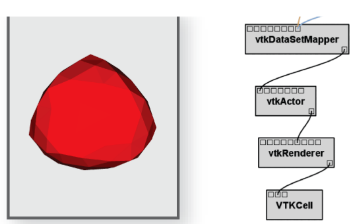


[Koop et al, IEEE TVCG 2008]

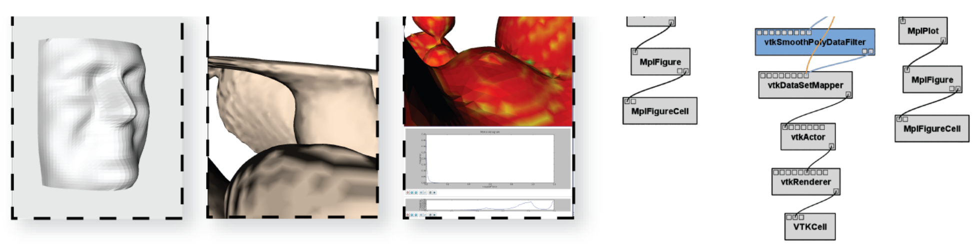
Provenance Beyond Reproducibility

- Support for reflective reasoning
- Ability to compare data products
- Explore parameter spaces and compare results
-

Same ideas can be applied to data science pipelines



Analogy Template



Automatically constructed visualizations

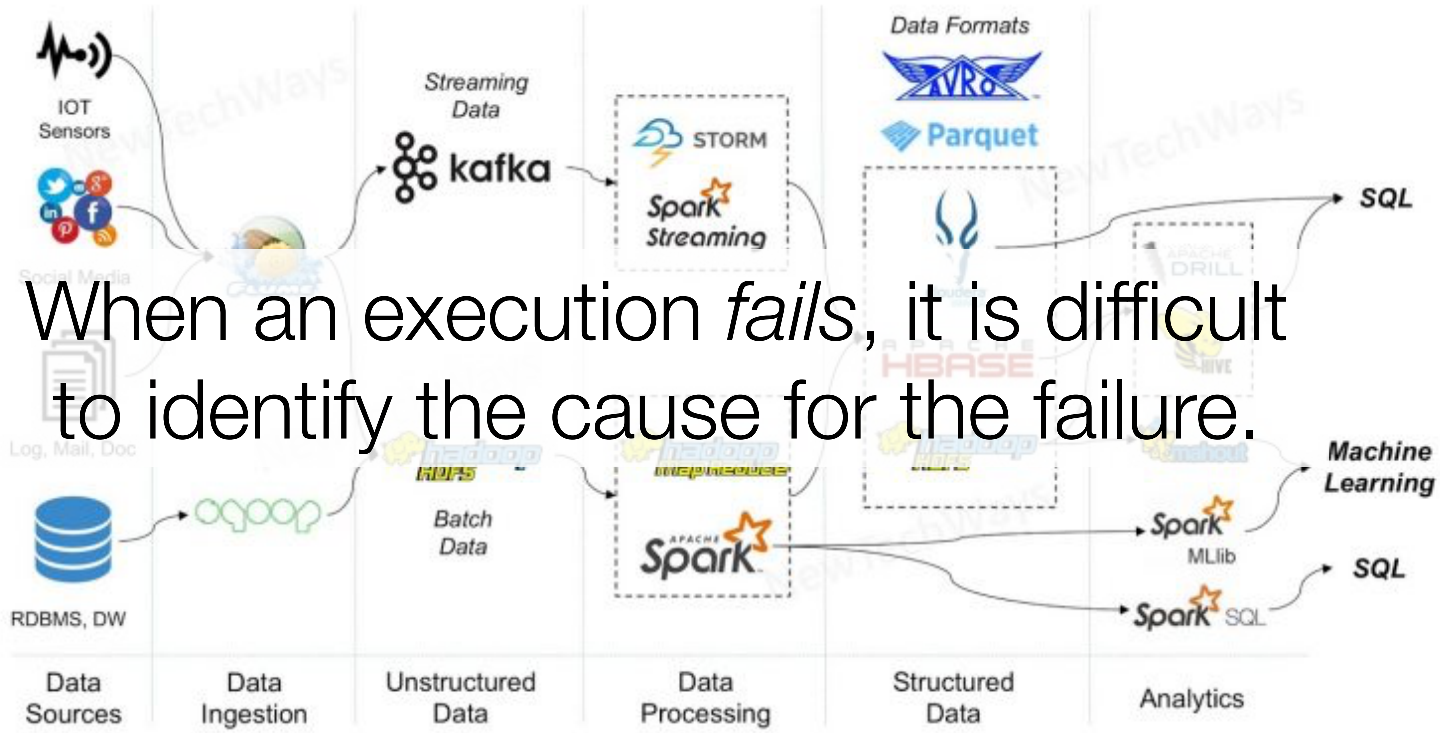
NYU TANDON SCHOOL OF ENGINEERING

[Scheidegger et al, IEEE TVCG 2007]

VIDA VISUALIZATION IMAGING AND DATA ANALYSIS CENTER

Reproducibility & Trust

- Reproducibility is a means to an end: necessary to verify and build trust in results
- And to debug computational pipelines!

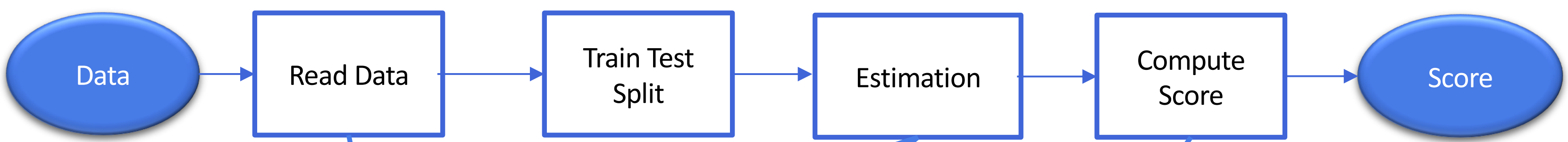


When an execution *fails*, it is difficult to identify the cause for the failure.

https://dingyuliang.me/big-data-pipeline-open-source-technologies/

Debugging Pipelines

- Problem: Given a computational pipeline, the provenance of previously-run instances, and an evaluation function that determines the success/failure of an execution, identify the root cause of failures



Execution Provenance

Instance	Data	Library	Estimator	Score	Evaluation
CP_1	Iris	1.0	Logistic regression	0.9	Success
CP_2	Digits	1.0	Decision tree	0.8	Success
CP_3	Iris	2.0	Gradient boosting	0.2	Failure
CP_4	Digits	2.0	Gradient boosting	0.3	Failure
CP_5	Iris	1.0	Decision tree	0.7	Success
CP_6	Images	1.0	Gradient boosting	0.9	Success

But are these the actual causes of failure?

Score < 0.6 → Failure

Finding Root Causes for Failures

- By executing new configurations with different parameter-value combinations, we formulate and test hypotheses
- Challenge: How to select the new configurations?

Instance	Data	Library	Estimator	Score	Evaluation
CP_1	Iris	1.0	Logistic regression	0.9	Success
CP_2	Digits	1.0	Decision tree	0.8	Success
CP_3	Iris	2.0	Gradient boosting	0.2	Failure
CP_4	Digits	2.0	Gradient boosting	0.3	Failure
CP_5	Iris	1.0	Decision tree	0.7	Success
CP_6	Images	1.0	Gradient boosting	0.9	Success
CP_7	Iris	1.0	Gradient boosting	0.7	Success

BugDoc: Debugging Pipelines

- Black-box approach to find root causes for problems in computational pipelines
- Leverages the ability to re-execute a given pipeline and iteratively create instances to test parameter-value combinations unseen in provenance
- “Shortcut” algorithm finds root causes in time **linear** wrt the number of parameters
- Debugging Decision Trees capture complex combinations of partial configurations as root causes
- Attains high precision and recall

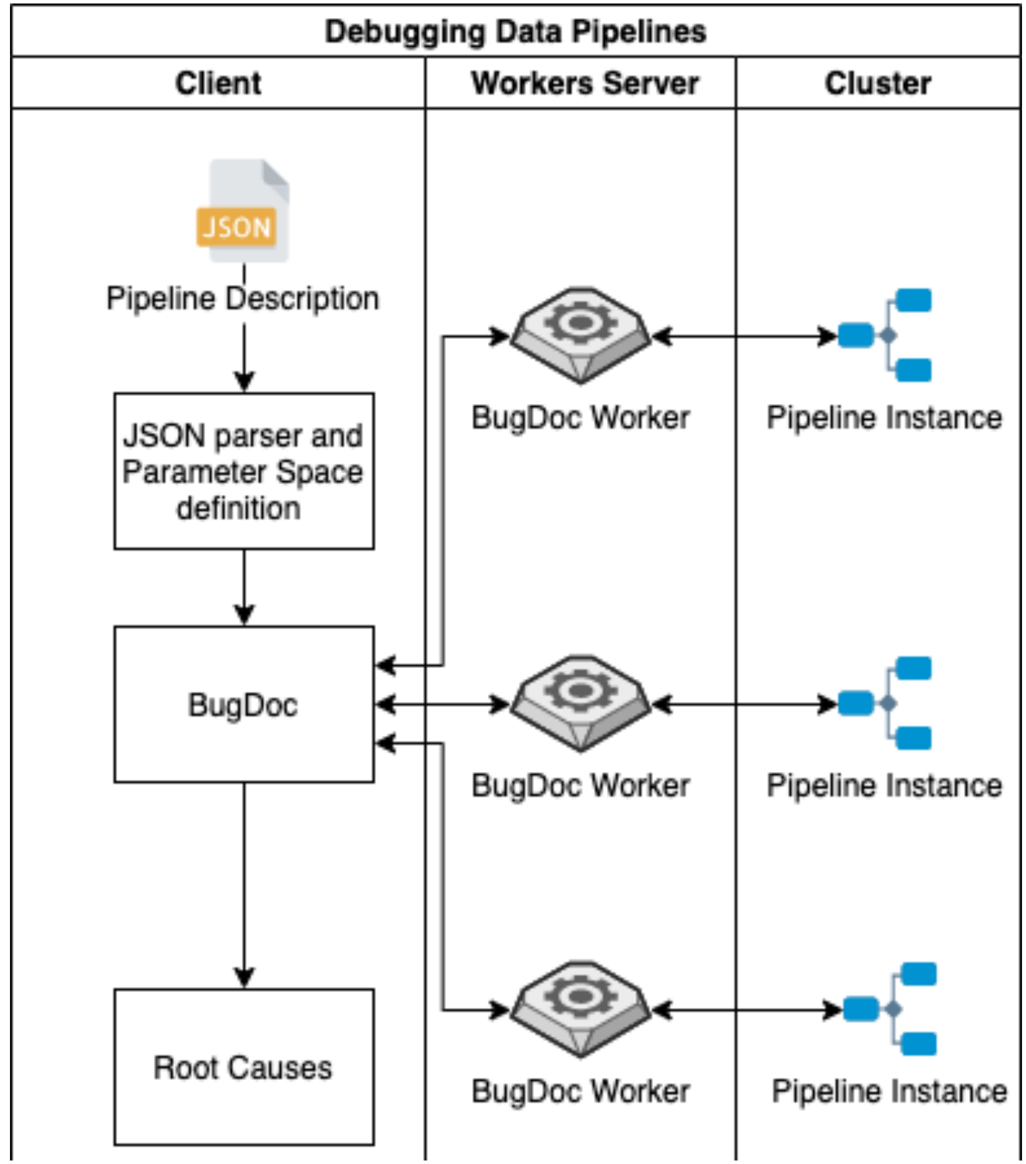
https://github.com/VIDA-NYU/BugDoc

[Lourenço et al., ACM SIGMOD DEEM 2019]

[Lourenço et al., ACM SIGMOD 2020]

BugDoc in Action at a Major Software Company

- BugDoc integrated to a proprietary Python API for a workflow system – workflows can be uploaded and executed
- Developed a parser to extract parameters and values from workflow
- BugDoc identified problematic configurations in computation scripts and input data files



Take Away

- Provenance and reproducibility are essential to build trust in and *explain* results produced by computational processes
- Not everybody produces reproducible results and follows reproducibility best practices…

- But reproducibility is not just important for science, it is also good for you!
  - Making your experiment reproducible forces you to document the execution pathways, this, in turn, helps newcomers to your group;
  - There exists preliminary evidence that reproducibility increases *impact*, *visibility* [Vandewalle et al. 2009] and *research quality* [Begley and Ellis 2012],
  - Helps you find problems in your computations, e.g., bugs, assess fairness

https://sites.nationalacademies.org/sites/reproducibility-in-science

Conclusions

- To empower domain experts, need new methods and usable tools that
  - Enable users to interactively explore *large* data
  - Help assess the quality and debug results
- Need interdisciplinary teams and collaboration with domain experts
  - Visualization, data management, computational topology, computer graphics, statistics
  - Virtuous cycle: interdisciplinary research that derives new problems and solutions for multiple areas
- Computer science and data science communities are well positioned to have tremendous practical impact

Acknowledgments

- Collaborators: Cláudio Silva, Huy Vo, Harish Doraiswamy, Dennis Shasha, Leonardo Murta, Vanessa Braganholo, Marcos Lage, Aline Bessa, Aecio Santos, Sonia Castello, Fernando Chirigati, Theo Damoulas, Nivan Ferreira, Heiko Mueller, Masayo Ota, Jorge Poco, Yeuk Yin Chan, Raoni Lourenço, Joao Felipe Pimentel, Divesh Srivastava, Remi Rampin, Fabio Miranda, Anastasia Ailamaki, Eleni Tzirita, David Koop, Carlos Scheidegger, Emanuele Santos, Lauro Lins, Steven Callahan, Matthias Troyer, Erik Anderson, Louis Bavoil
- 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