

Introduction and Overview

CS524: Big Data Visualization & Analytics

Fabio Miranda

<https://fmiranda.me>

A DAY IN DATA

The exponential growth of data is undisputed, but the numbers behind this explosion – fuelled by internet of things and the use of connected devices – are hard to comprehend, particularly when looked at in the context of one day

 500m
tweets are sent every day

Twitter

294bn
billion emails are sent

Radicati Group



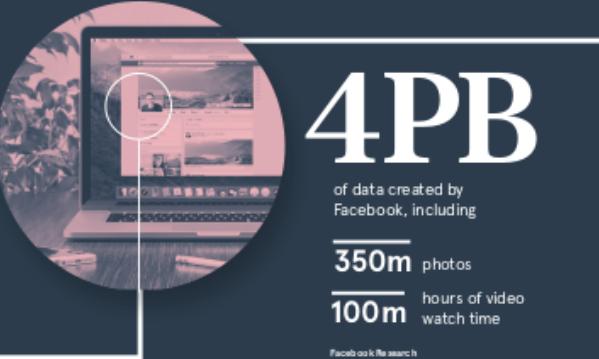
3.9bn
people use emails

Radicati Group

320bn
emails to be sent each day by 2021

306bn
emails to be sent each day by 2020

Radicati Group



Facebook Research

4TB
of data produced by a connected car

Intel

ACCUMULATED DIGITAL UNIVERSE OF DATA

4.4ZB

2010

44ZB

2020

DEMYSTIFYING DATA UNITS

From the more familiar ‘bit’ or ‘megabyte’, larger units of measurement are more frequently being used to explain the masses of data

Unit	Value	Size
b	0 or 1	1/8 of a byte
B	8 bits	1 byte
KB	1,000 bytes	1,000 bytes
MB	1,000 ³ bytes	1,000,000 bytes
GB	1,000 ⁶ bytes	1,000,000,000 bytes
TB	1,000 ¹² bytes	1,000,000,000,000 bytes
PB	1,000 ¹⁵ bytes	1,000,000,000,000,000 bytes
EB	1,000 ¹⁸ bytes	1,000,000,000,000,000,000 bytes
ZB	1,000 ²¹ bytes	1,000,000,000,000,000,000,000 bytes
YB	1,000 ²⁴ bytes	1,000,000,000,000,000,000,000,000 bytes

A lowercase ‘b’ is used as an abbreviation for bits, while an uppercase ‘B’ represents bytes.

65bn

messages sent over WhatsApp and two billion minutes of voice and video calls made

Facebook



Searches made a day → 5bn

Searches made a day from Google → 3.5bn

Smart Insights



Source: Raconteur

463EB

of data will be created every day by 2025

idc

95m

photos and videos are shared on Instagram

Instagram Business

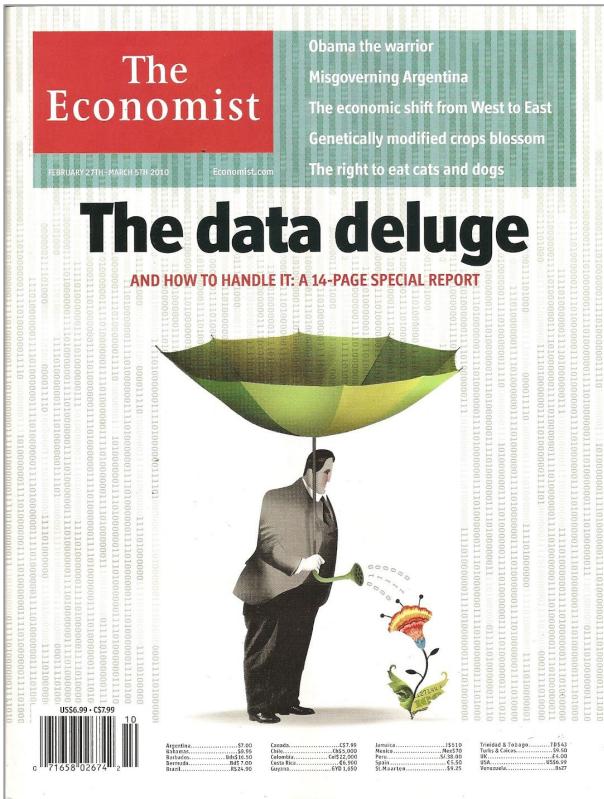
28PB

to be generated from wearable devices by 2020

Statista



Data is everywhere



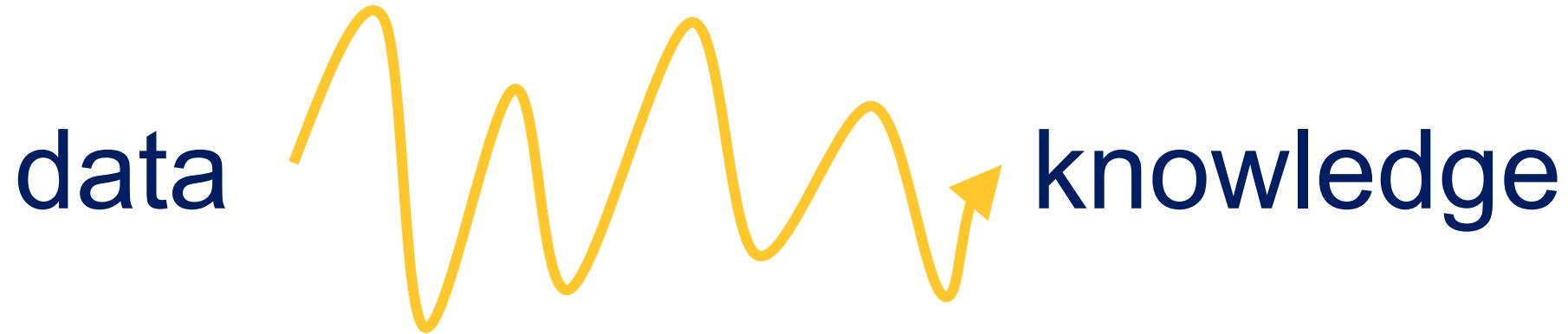
“The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that’s going to be a hugely important skill in the next decades, ... because now we really do have essentially free and ubiquitous data.”

Hal Varian, Google's Chief Economist
The McKinsey Quarterly, Jan 2009

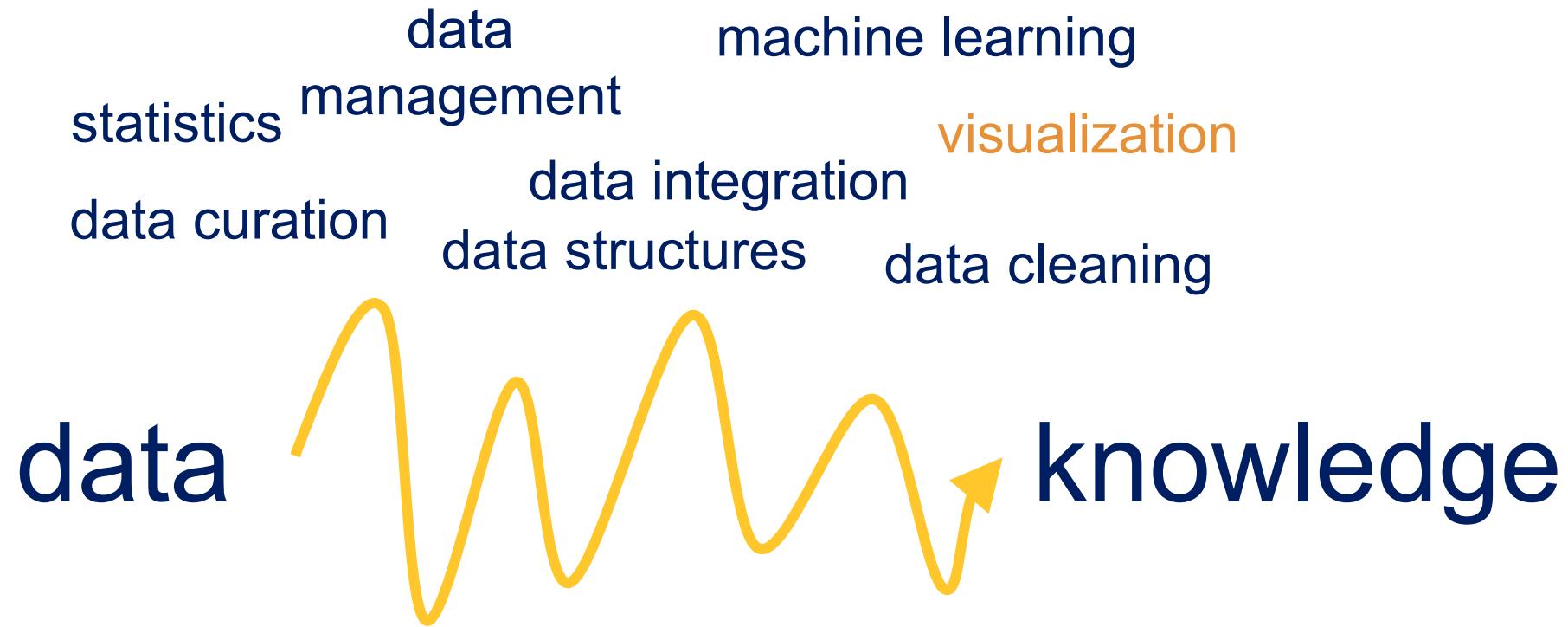
Data to knowledge

data → knowledge

Data to knowledge

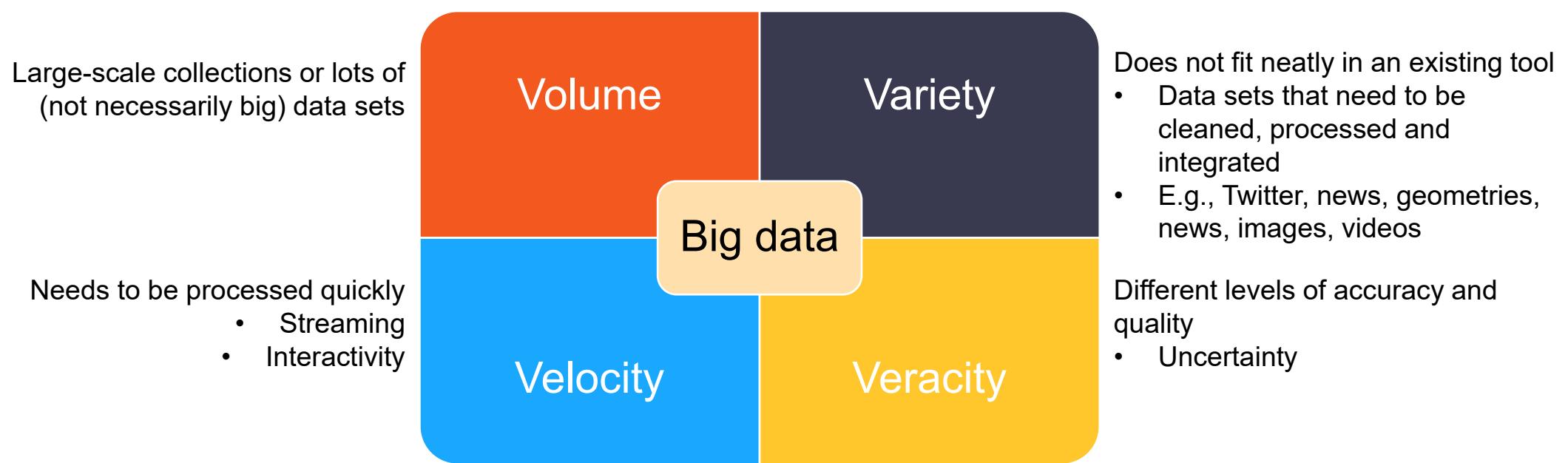


Data to knowledge

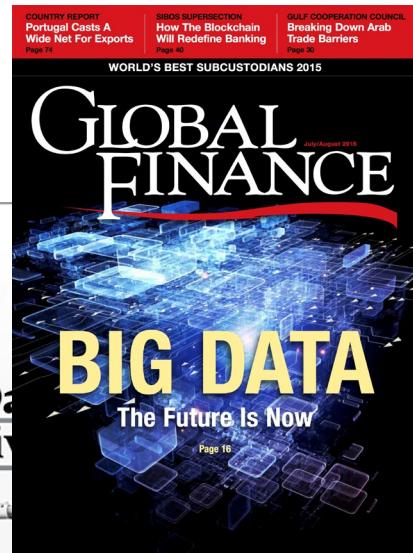


What is Big Data?

- Broad term for data so large and complex that traditional data processing applications are inadequate.

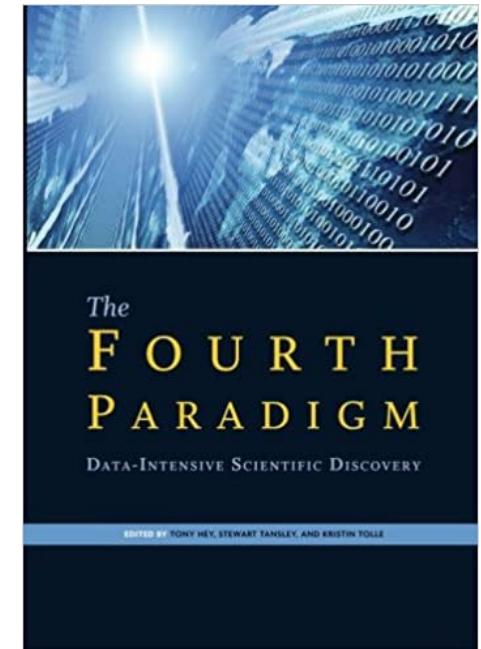


The big deal of big data

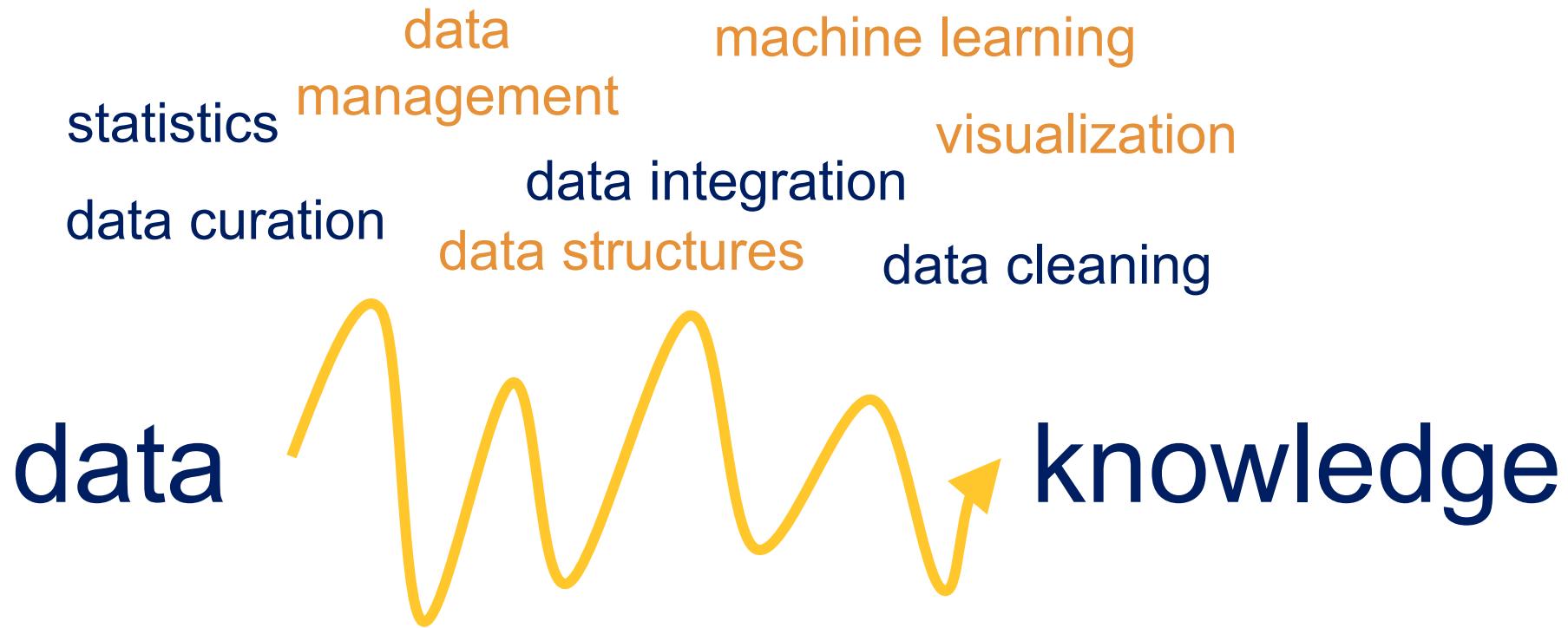


The big deal of big data

- **Science:** several domains are moving toward data-driven exploration: *increasing use of data is bringing a paradigm shift to the nature of science.*
 - Climate data, urban data, physics data, health data.
- **Industry:** companies are capitalizing on data – users are consuming and producing data.
 - Social media data, crowdsourced data, sensor data.
- **Government:** agencies use data to operate efficiently, make policies, make informed decisions.
 - Data.gov: 224,669+ datasets; NYC Open Data: 1,400 datasets.



Big data to knowledge



Big data analysis: Common practices

1. Domain experts and policy makers formulate hypotheses.
 2. Computer scientists or data scientists select data sets and slices, perform analyses, and derive plots.
 3. Domain experts examine the plots, go to step 1.
- Issues:
 - Dependency on computer scientists or data scientists distances domain experts from the data.
 - Batch-oriented analysis pipeline hampers exploration – mostly confirmatory analyses.
 - Data are complex – often multivariate spatiotemporal.
 - Analysis limited to samples.

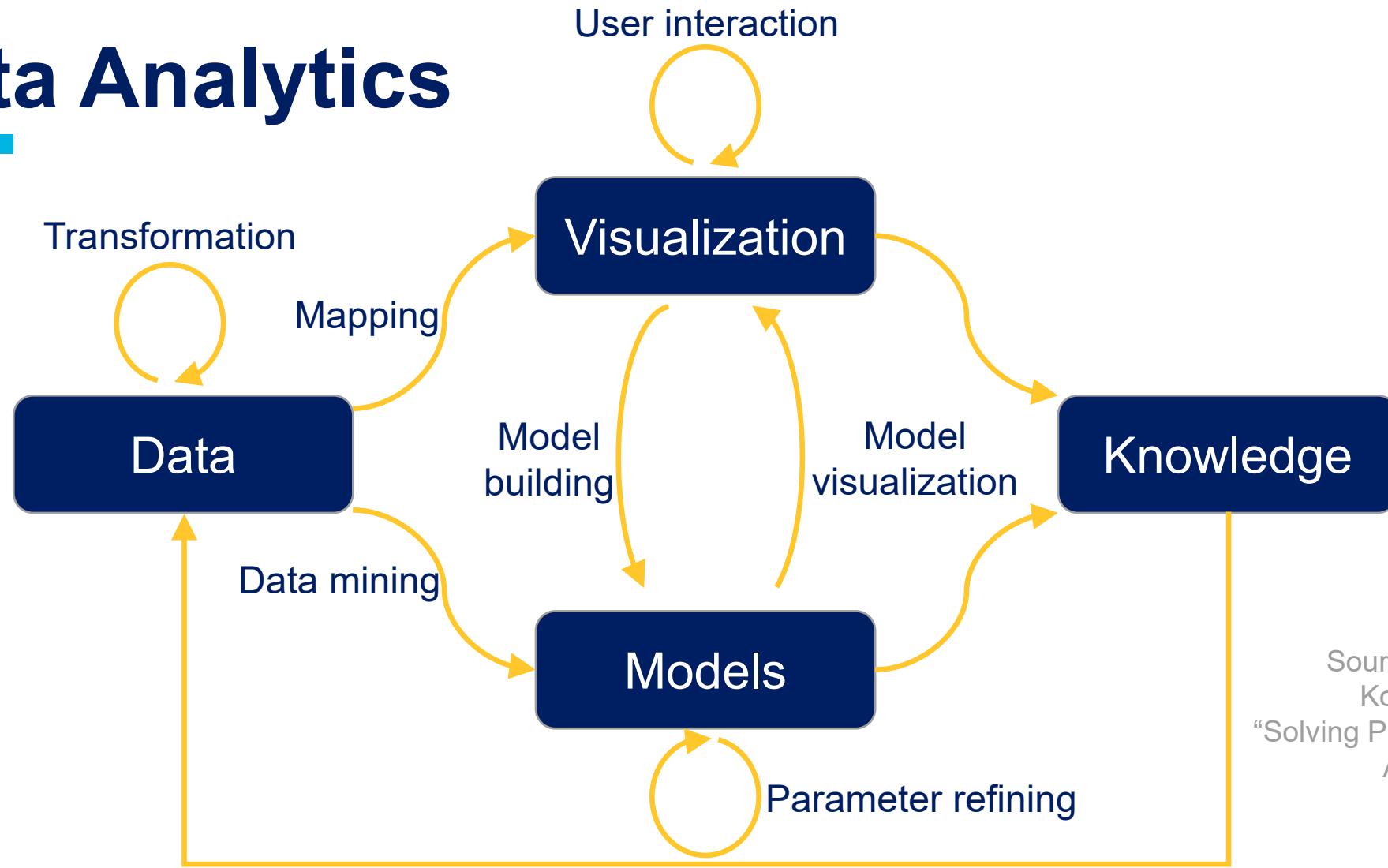
Big data analysis: What we need

- Scalable tools and techniques that help domain experts find, clean, integrate, ***interactively*** explore, and explain data.
- Guide users in the exploration process.
- Support interactive queries:
 - “*increased latency reduces the rate at which users make observations, draw generalizations, and generate hypotheses*”.
- Interdisciplinary:
 - “*as data scale and complexity increases, the novel solutions that will ultimately enable interactive, large-scale exploratory data analysis will have to come from truly interdisciplinary work*”.

[Liu and Heer, IEEE TVCG 2014]

[Chang, Fekete, Freire and Scheidegger, Dagstuhl Reports 2017]

Data Analytics



Source: Kleim and Kohlhammer,
“Solving Problems with Visual
Analytics”

Big data challenges

- “*Although modern database management systems (DBMS’s) allow users to perform complex scientific analyses over large datasets, DBMS’s are not designed to respond to queries at interactive speeds.*”

[Battle, Chang, Stonebraker SIGMOD 2016]

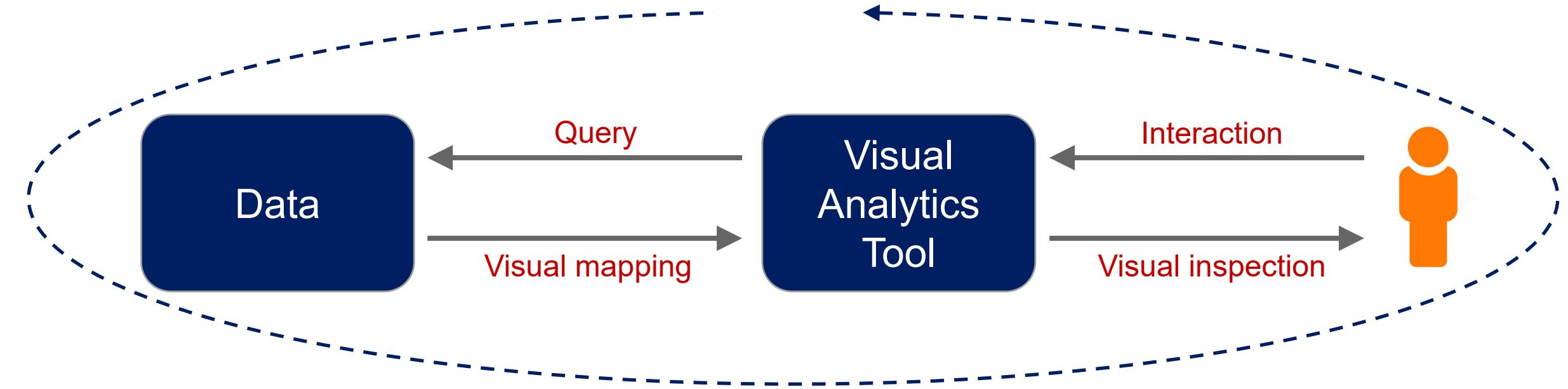


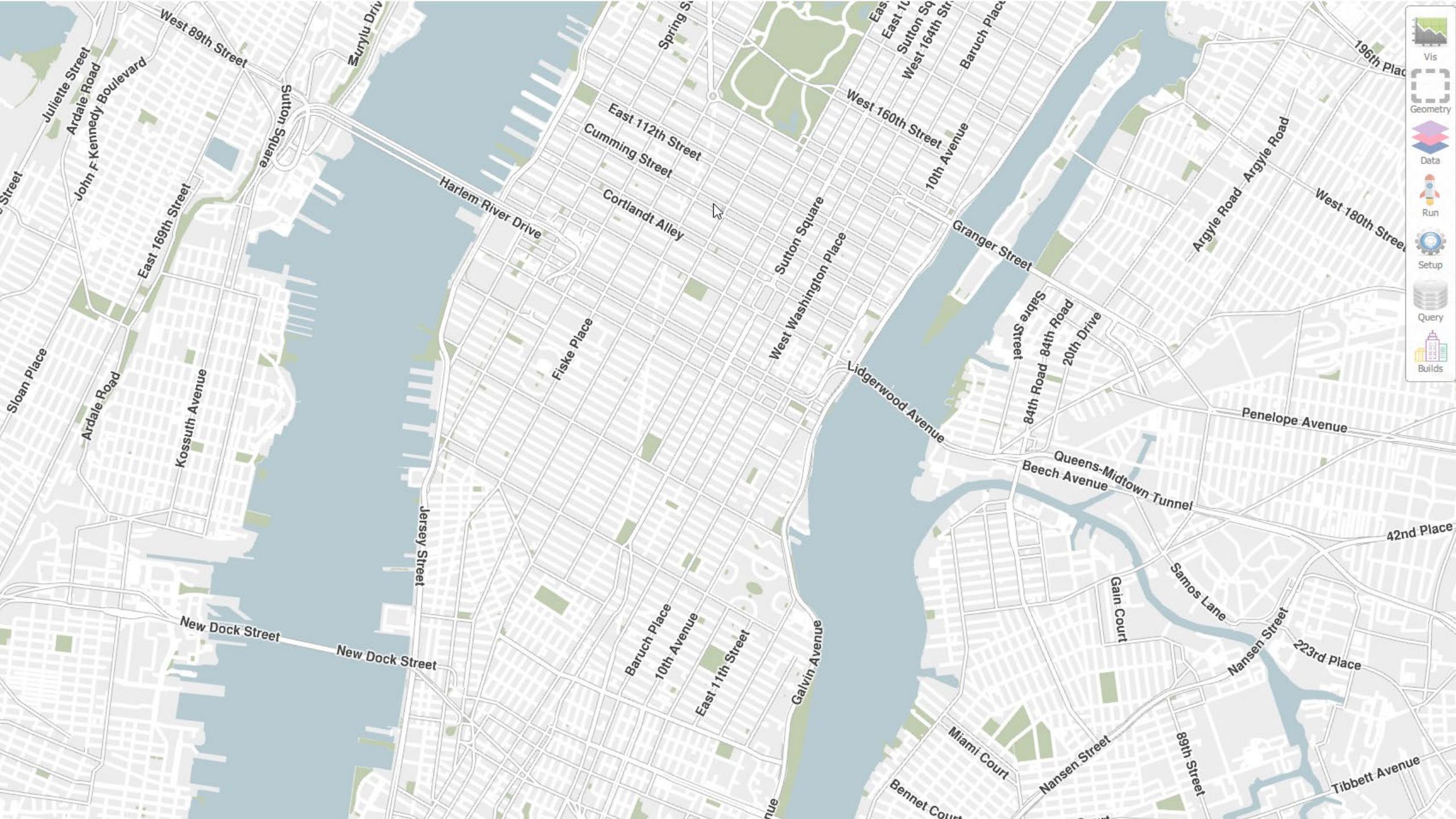
Big data challenges

- Data are vast and produced at unprecedented rates.
 - Sources are broad, varied, and unreliable.
- Computational processes are required to extract insight.
 - Hard to assemble and require expertise in a wide range of topics.
- Exploratory task are inherently iterative as one tests and formulate hypotheses:
 - *“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]

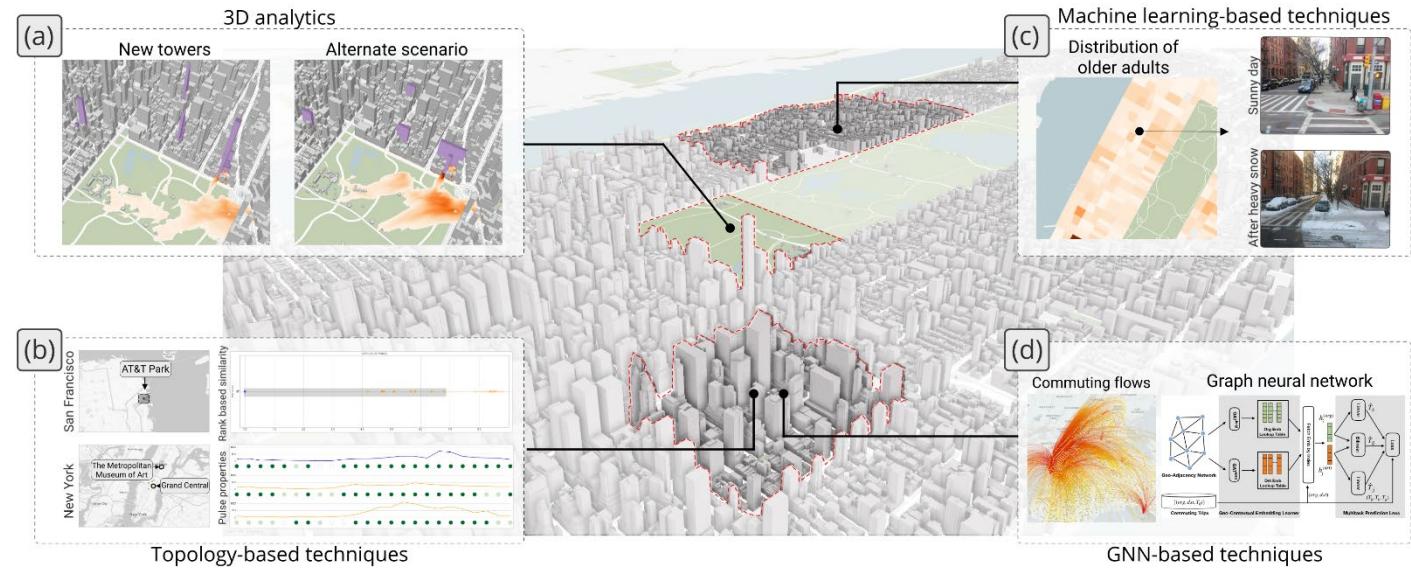
Interactive visual analysis





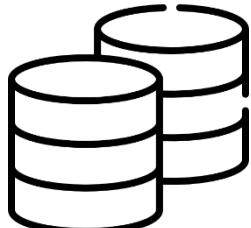
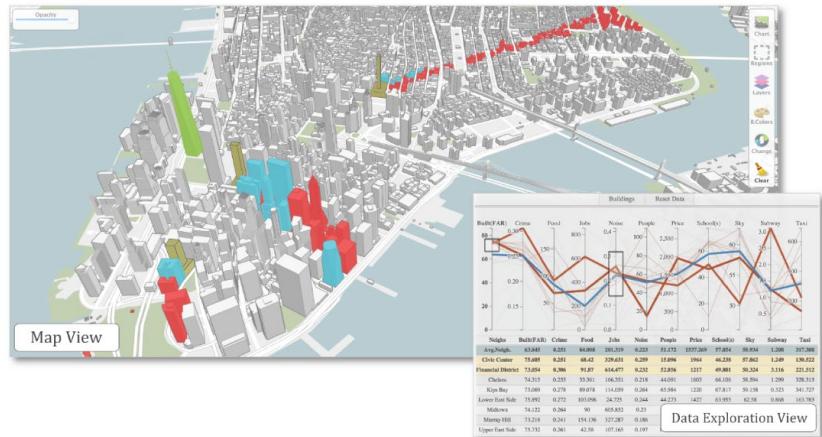
Research

Assistant Professor, CS (UIC)
PhD, CS 2018 (NYU)
MSc, CS 2012 (PUC-Rio)
BSc, CS 2009 (UFMG)

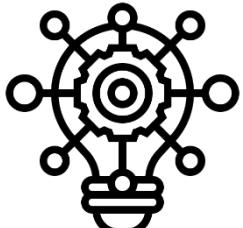


- Methods and techniques that follow a **human-centered approach to data science**, fostering the involvement of domain experts in the analysis process of big data.
- **Interactive tools and frameworks** that combine visualization, data management, human-computer interaction, and machine learning to support data-driven decision making by domain experts.

Big data vis



Data storage

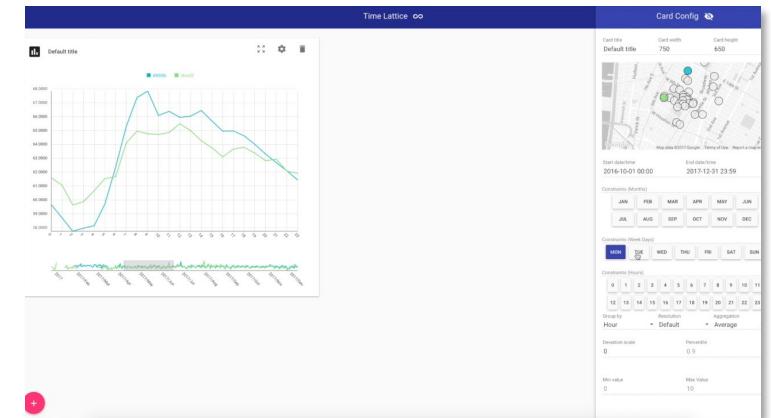
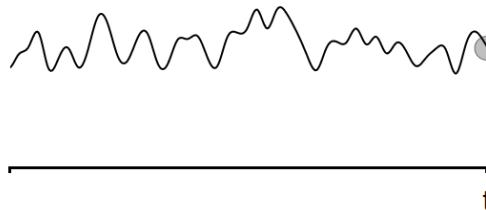


Data analytics



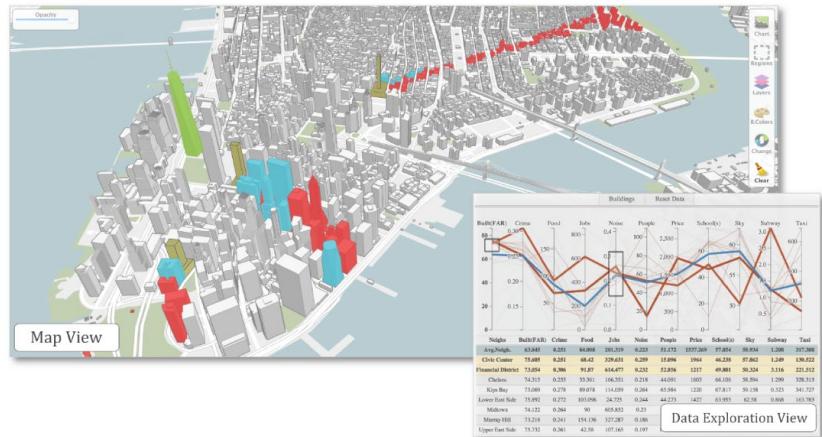
Data visualization

[Best SIGMOD'18 paper]

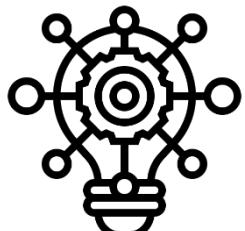


Time Lattice [CGF 2017]

Big data vis



Data storage

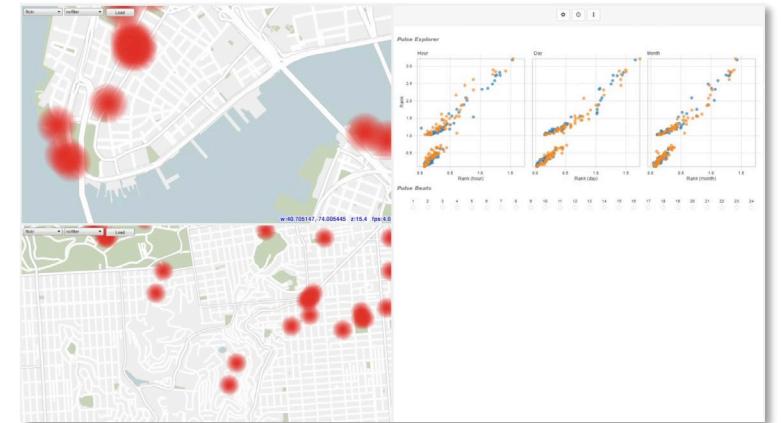
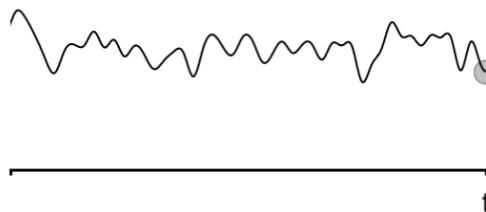


Data analytics



Data visualization

[Best SIGMOD'18 paper]

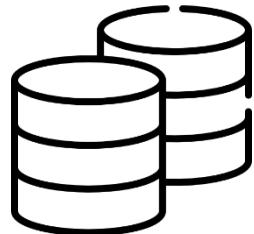
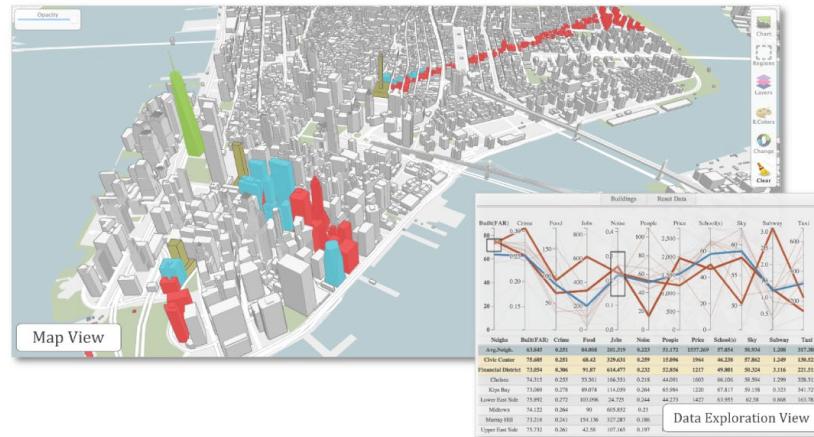


Urban Pulse [TVCG 2016]

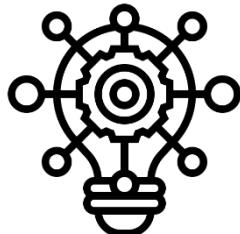


COMPUTER SCIENCE

Big data vis



Data storage

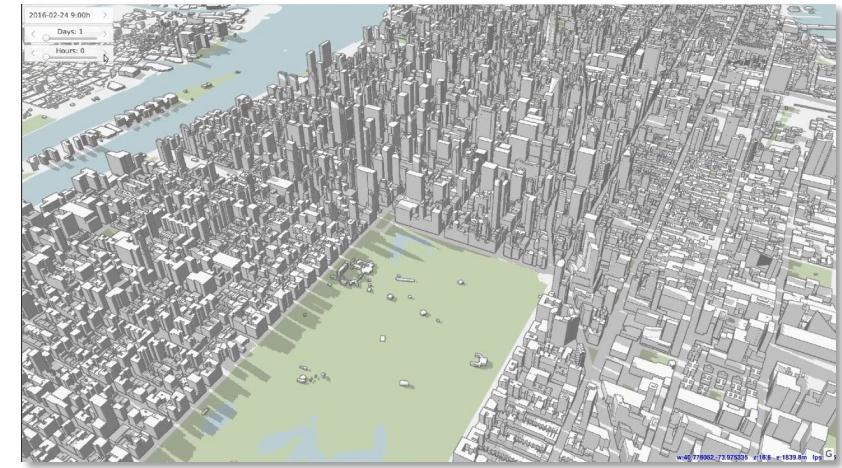
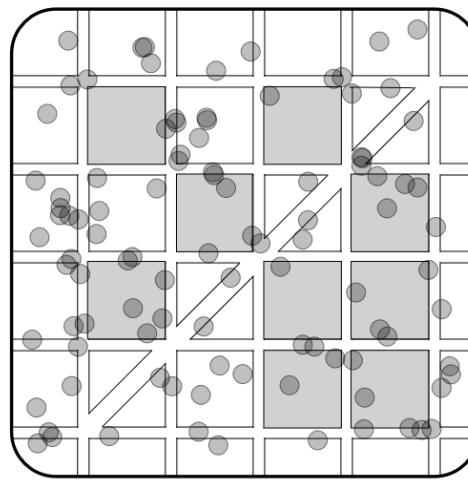


Data analytics



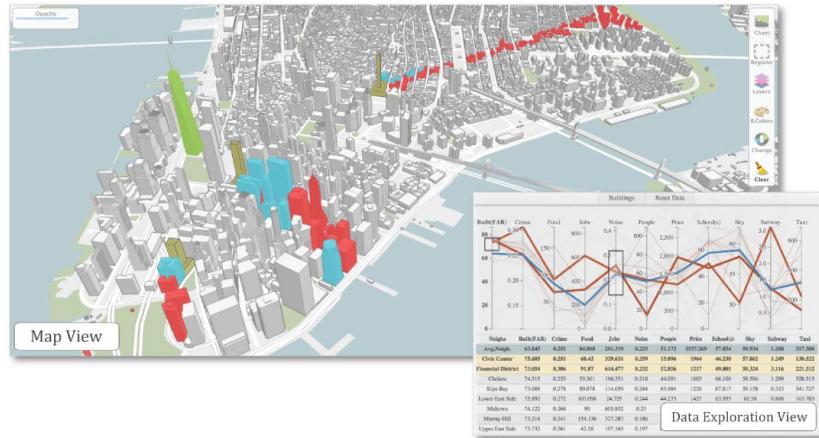
Data visualization

[Best SIGMOD'18 paper]



Shadow Accrual Maps [TVCG 2019, New York Times 2018]

Big data vis



Data storage

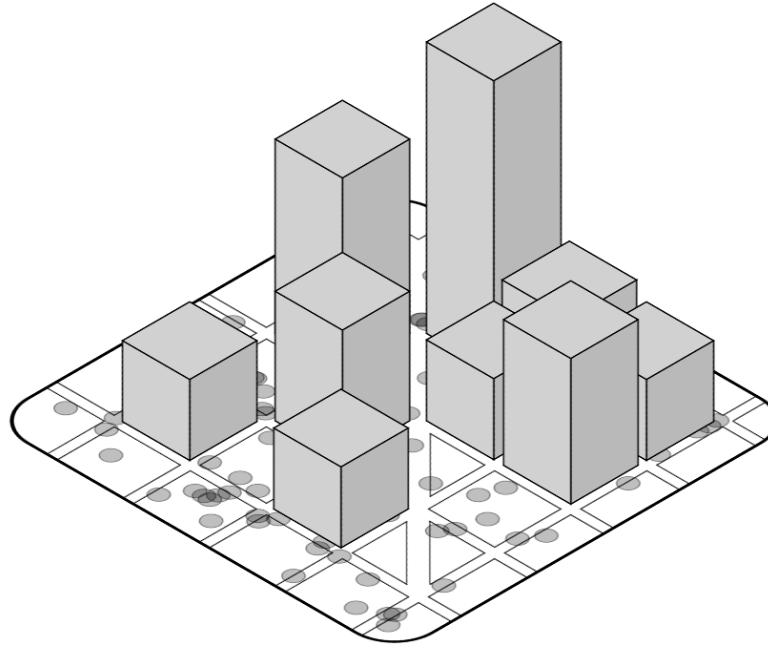


Data analytics



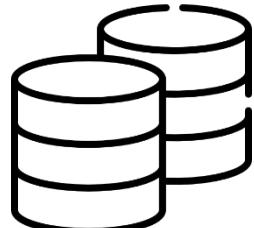
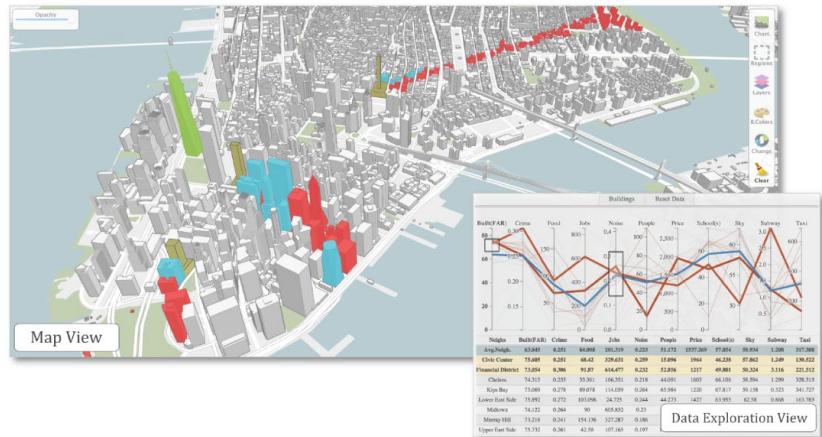
Data visualization

[Best SIGMOD'18 paper]

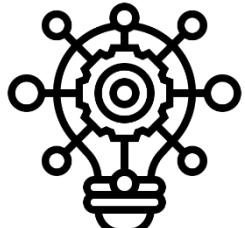


Urban Mosaic [CHI 2020]

Big data vis



Data storage

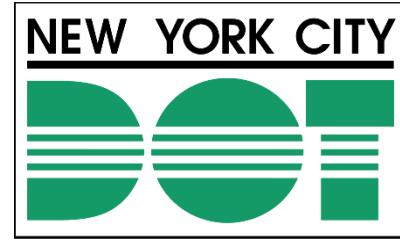


Data analytics



Data visualization

[Best SIGMOD'18 paper]



KPF

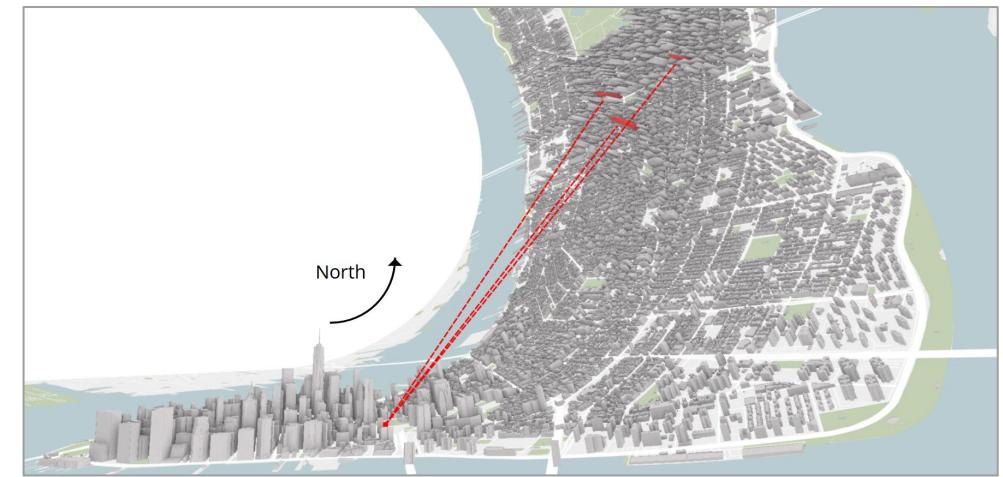
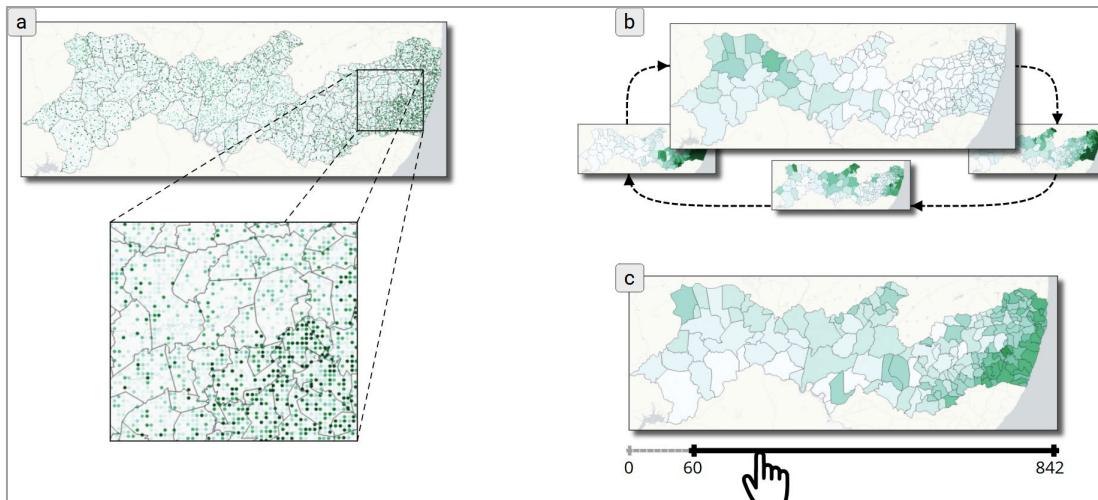
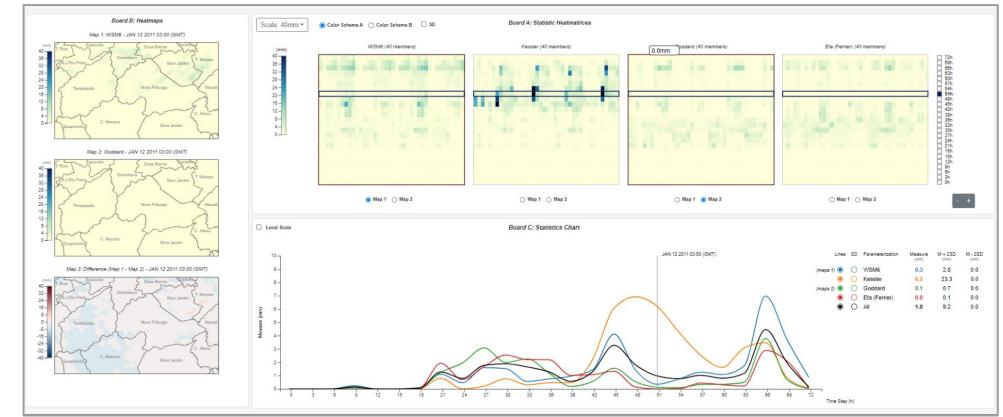
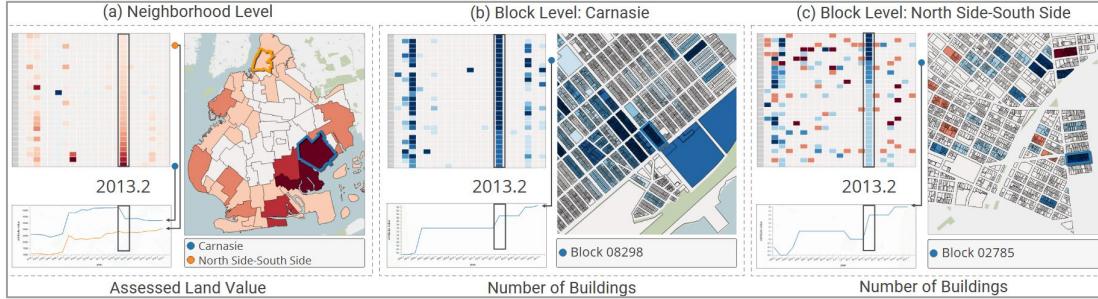
The New York Times



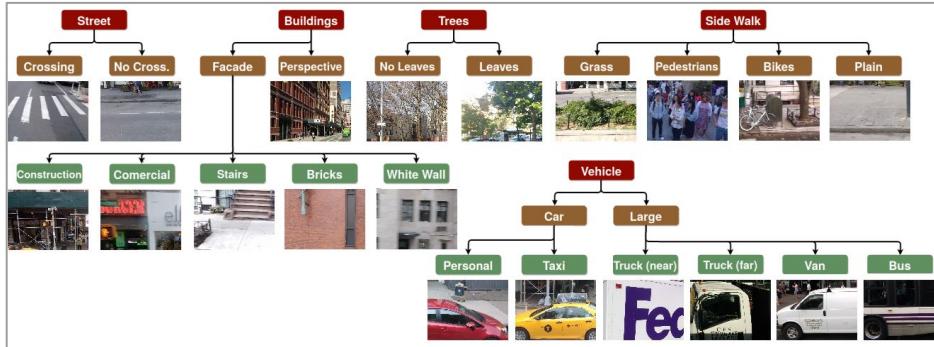
DRAW
BROOKLYN



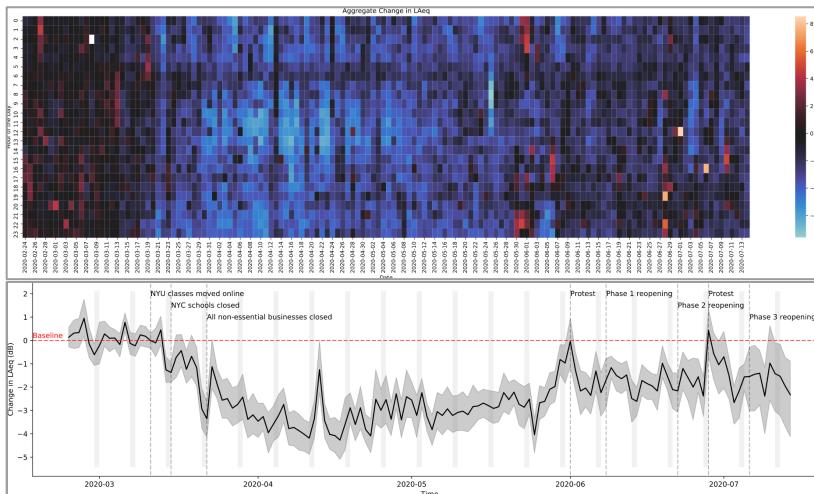
Big data vis



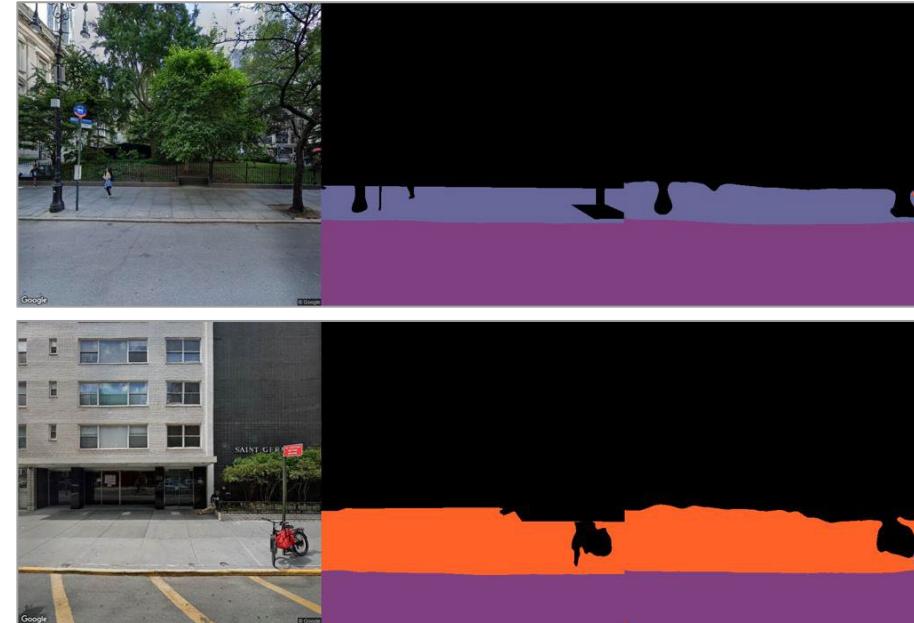
Big data vis



Creation of Urban Taxonomies from Street-level Image Collections



Impact of COVID-19 lockdowns on urban noise and perceived urban noise



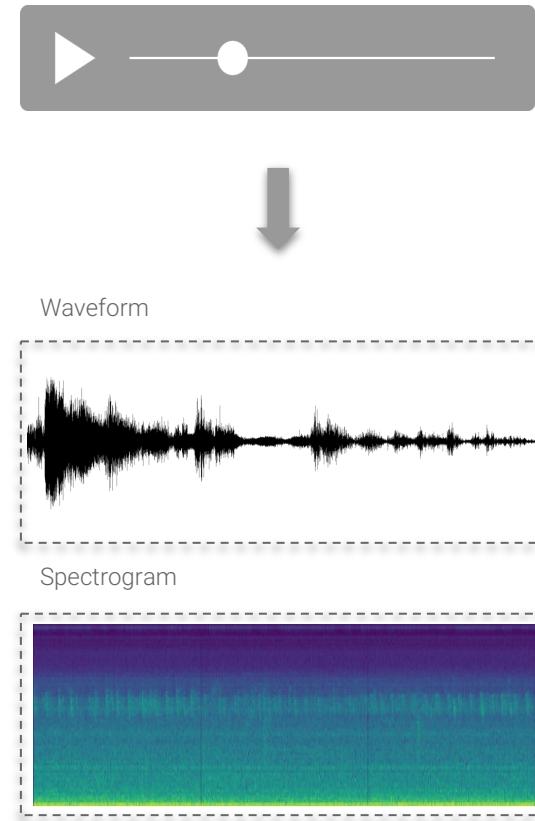
CitySurfaces: NN model for the segmentation and assessment of sidewalk material



Python toolkit for urban data visualization and analysis

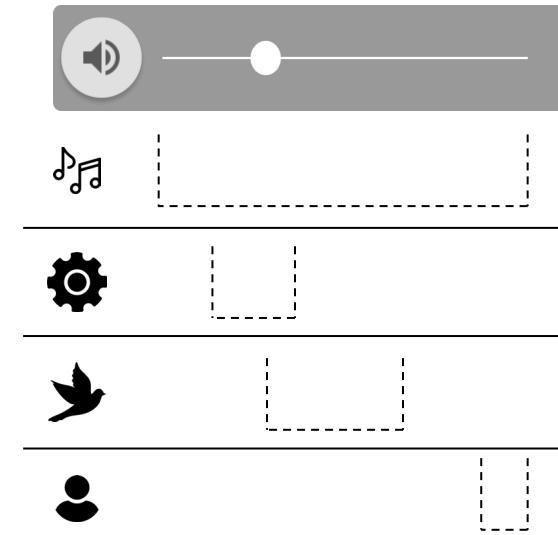
Urban audio data exploration

- Audio recordings are consumed in a **serial** way by us.
- To understand events happening in a 10-second audio snippet, users **must listen through the entire recording**.
- Although the **visualization of specific frequencies or loudness** can help identify interest periods of the recording, it is still difficult to build a **semantic understanding** of the recording.



Urban audio data exploration

- Unlike images, where visual objects are opaque, sound objects are conceptually ***transparent***, meaning that multiple objects (sound sources) can have energy at the same frequency.
- At any given instant in time, a sound recording might have a **mixture of background** (birds, dog barks) and **foreground sounds** (party, sirens).

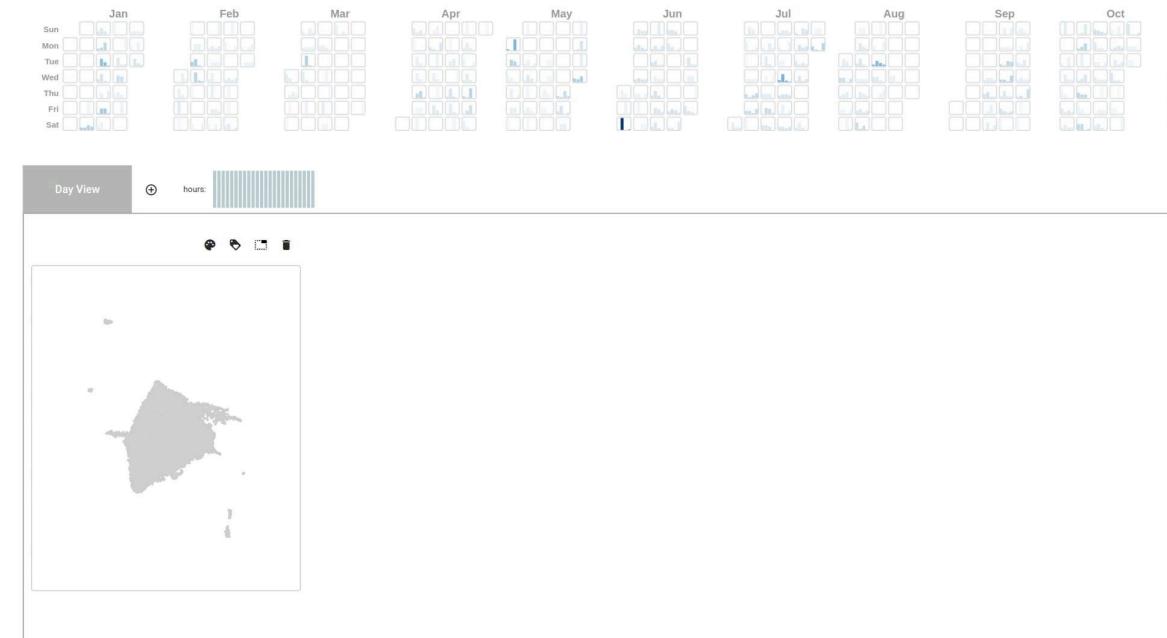


Urban Rhapsody: Creating prototypes

Binary Classification Model



Representatives



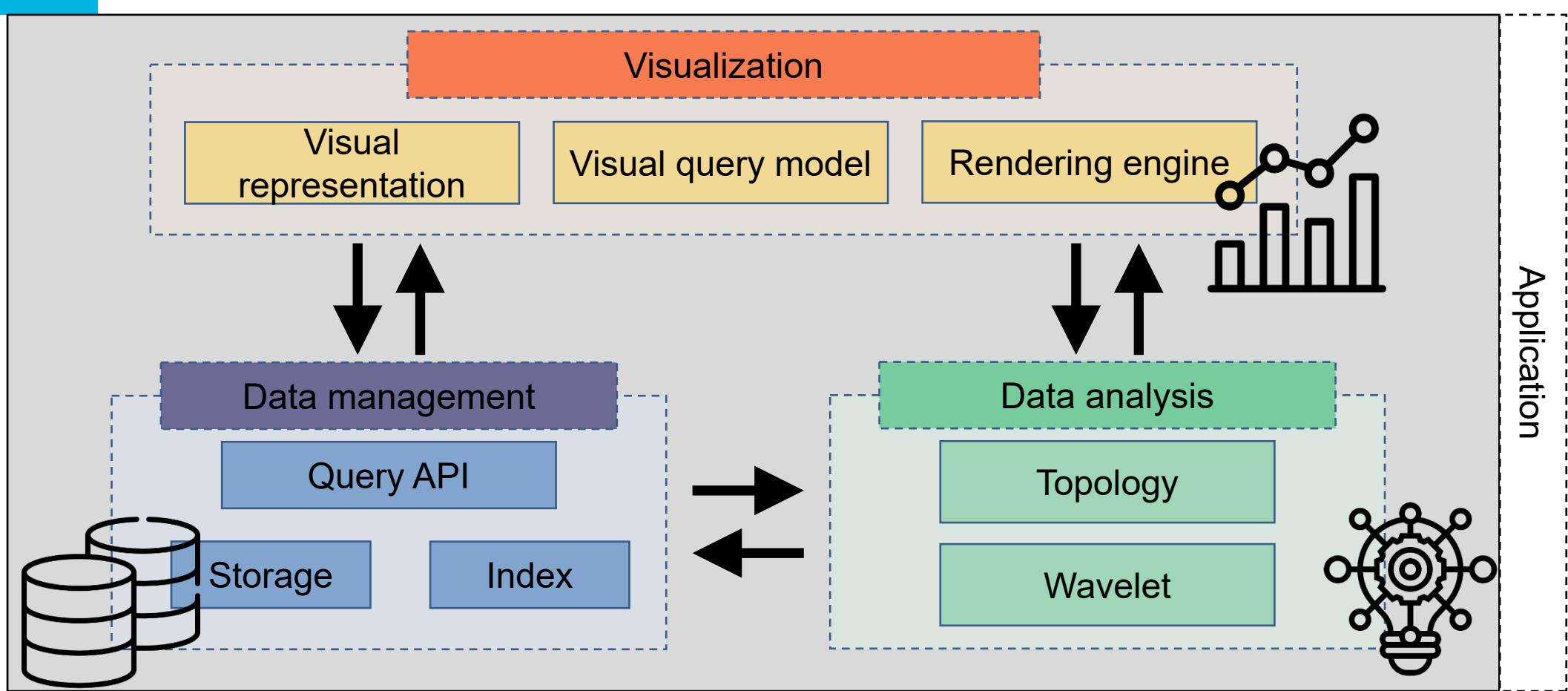
CS524: Big Data Visualization & Analytics

- Intersection between visualization, data management and data mining, covering the necessary topics to build visual analytics tools to handle big data.
- Broad definition of big data: any dataset with size (or complexity) that goes beyond the ability of standard tools and techniques to **interactively** manage and process it.
- At the end of the course, you will be able to:
 - Design and implement visual analytics systems capable of handling large data (combining visualization and data mining techniques, data structures and algorithms)

CS524: Big Data Visualization & Analytics

- Necessary techniques to build visual analytics tools to handle big data:
 1. Building blocks: current technologies and libraries to build visual analytics systems for big data.
 2. Visualization: visual implications of handling big data.
 3. Data management: techniques to handle big data.
 4. Analytics: data mining and technical frameworks to extract patterns or features that can drive visual analytics systems.

CS524: Big Data Visualization & Analytics



Requirements

- No specific pre-requisite courses, but students are expected to be comfortable with programming and be able to learn new programming languages as required by the assignments and projects, as well as write technical documents.

Logistics

- Syllabus: <https://fmiranda.me/courses/cs524-spring-2023/>
- We will meet twice a week:
 - Tue 9:30am – 10:45pm (Central)
 - Thur 9:30am – 10:45pm (Central)
- In-person: EVL's Continuum.
- Office hours:
 - Friday 10:00am – 12:00pm (Central).
 - 2032 ERF or Zoom (link on Discord).

Assignment 1

- Front-end technologies and libraries:
 - Javascript
 - Angular
 - D3
- Simple assignment to make sure everyone is on the same page.
- Week 2, due week 3.



Assignment 2

- Back-end technologies and libraries:
 - Boost (C++)
 - Qt (C++)
 - Flask
- Simple assignment to make sure everyone is on the same page.
- Week 3, due week 4.



Research paper presentation

- Presentation from one research paper selected by you.
- ~10-minute presentation.
- Week 7.
- You should cover:
 - Overview of related work
 - Major contributions
 - Comparison with previous works
 - Evaluation

Final project

- Cumulative final project, you will need to demonstrate your research skills by combining visualization and data mining techniques, data structures and algorithms that work in tandem to enable interactive data exploration.
- Three milestones (3 months):
 - Detailed instructions for each milestone will be made available following the evaluation schedule.
 - M1: Project proposal (due week 5)
 - Two-page document describing proposal.
 - List of potential projects will be posted on week 3, but you are free to propose your own project.
 - M2: Midterm review (due week 10)
 - ~10-minute presentation.
 - M3: Final delivery (due week 15)
 - ~10-minute presentation.
 - Final report.

Final project

- Detailed instructions regarding proposals and presentations will be posted following the course schedule.
- Teams of 2 students.

Grading policy

- Assignments: 10%
- Research paper presentation: 25%
- Final project midterm review: 25%
- Final project: 30%
- Participation: 10%

Overview

Week	Topic	Resources and readings	Slides
1	Intro to course and visualization	Required: [1]; Others: [2], [3]	
2	Building blocks: technologies and libraries	[1], [2], [3]	
3	Vis: Big data visual analytics systems	Required: [1]; Others: [1], [2], [3]	
4	Vis: Progressive & Uncertainty	Required: [1]; Others: [2], [3], [4], [5], [6], [7], [8]	
5	Proposal presentations		
6	Vis: Vis for machine learning	Required: [1]	
7	Paper presentations		
8	Analytics: Machine learning for vis	Required: [1], [2]	
9	Analytics: Computational topology	Required: [1], [2]; Others: [1]	
10	Final project midterm review		
11	Data: Visualization data structures	Required: [1]; Others: [2], [3], [4]	
12	Data: Spatial queries	Required: [1]; Others: [2]	
13	Data: MapReduce, Hadoop, Spark, Dask	[1], [2], [3]	
14	Data: Approximate queries, hashing, learned indices	Required: [1]; Others: [2]	
15	Project presentations		

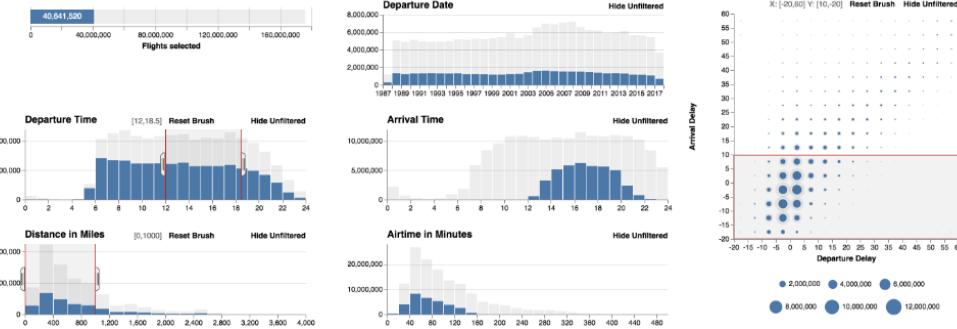
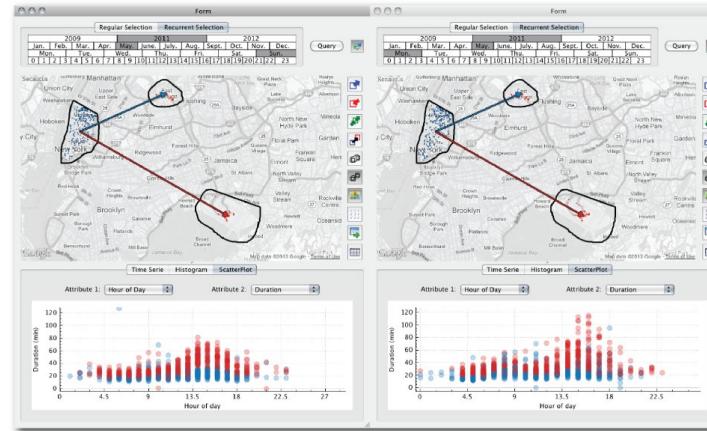
Building blocks: Weeks 2-3

- Necessary tools and frameworks for the development of big data analytic systems:
 - Angular (Javascript)
 - Boost (C++)
 - Qt (C++)
 - Flask (Python)



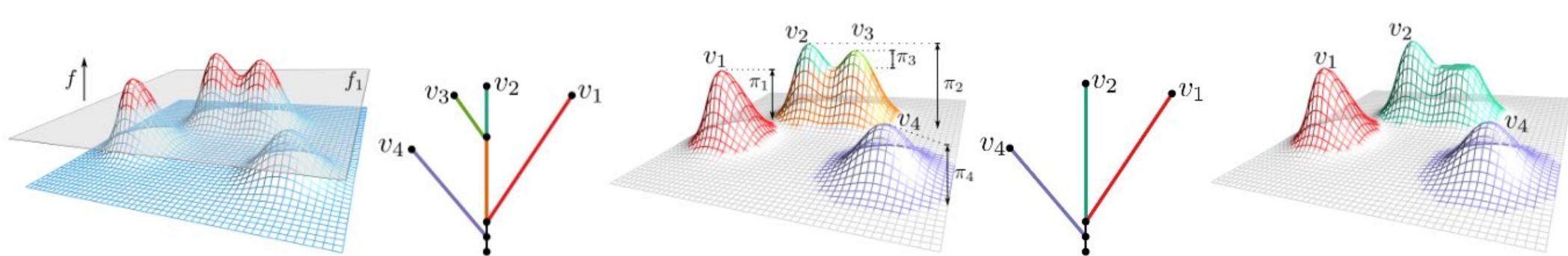
Visualization: Weeks 3-6

- Visual analytics systems
 - Interactivity requirements.
 - Components of the system.
- Progressive visualization
 - Progressively build visualizations, maintaining interactivity.
- Uncertainty visualization
 - Defining uncertainty.
 - Displaying uncertainty.
- Visualization for machine learning



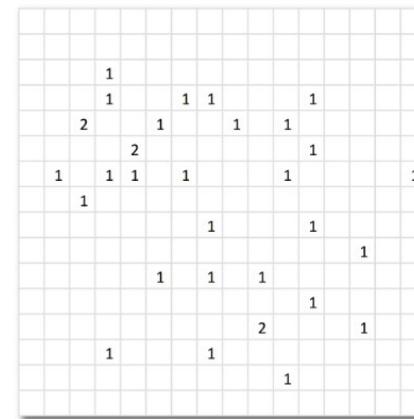
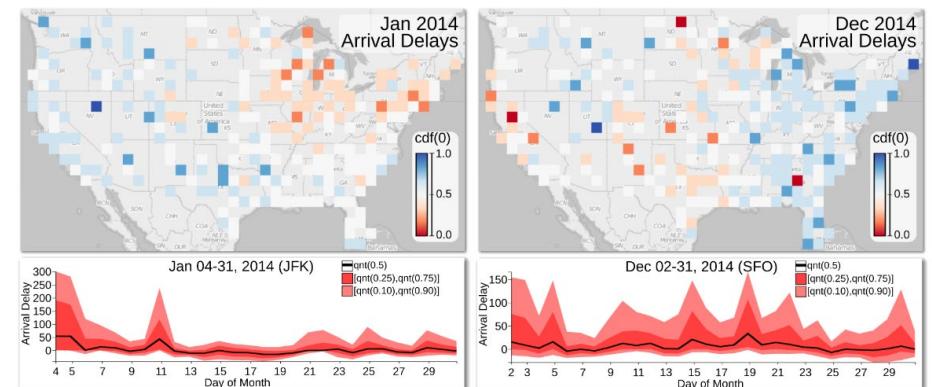
Analytics: Weeks 8-9

- Data mining and analytics techniques to extract features and patterns from big data:
 - Computational topology
 - Wavelet
 - Techniques for streaming data
- Machine learning for visualization

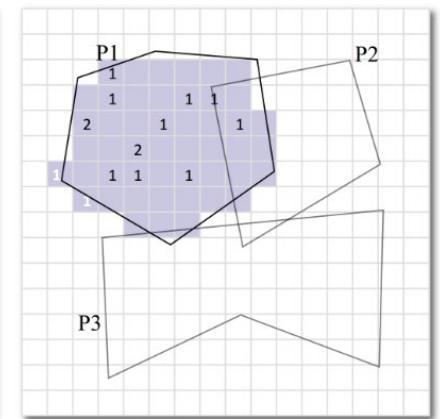


Data management: Weeks 11-14

- Approximate queries, hashing, learned indices
- Spatial structures
 - Nanocubes
- Spatial queries
 - GPU-based indices
- MapReduce
 - Hadoop, Spark



(a)



(b)

Other considerations

- Piazza will be the main platform for communication, please check Blackboard for the Piazza invitation link.
- Github Classroom will be used for assignments and final project.
- Required readings.
- Late submissions / presentations will be penalized at a deduction rate of 20%.
- Participation grade will take into consideration your contribution to a productive environment, either in the classroom, discord or office hours.