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 devcies – are hard to comprehend, particularly when looked at in the context of one day

DEMYSTIFIYING 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	bit	0 or 1	1/8 of a byte
B	byte	8 bits	1 byte
KB	kilobyte	1,000 bytes	1,000 bytes
MB	megabyte	$1,000^2$ bytes	1,000,000 bytes
GB	gigabyte	$1,000^3$ bytes	1,000,000,000 bytes
TB	terabyte	$1,000^4$ bytes	1,000,000,000,000 bytes
PB	petabyte	$1,000^5$ bytes	1,000,000,000,000,000 bytes
EB	exabyte	$1,000^6$ bytes	1,000,000,000,000,000,000 bytes
ZB	zettabyte	$1,000^7$ bytes	1,000,000,000,000,000,000,000 bytes
YB	yottabyte	$1,000^8$ 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.

463EB

of data will be created every day by 2025

IDC

500m

tweets are sent every day

Twitter

4PB

of data created by Facebook, including

350m photos

100m hours of video watch time

Facebook Research

95m

photos and videos are shared on Instagram

Instagram Business

320bn

emails to be sent each day by 2021

294bn

billion emails are sent

306bn

emails to be sent each day by 2020

Radicati Group

65bn

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

Facebook

3.9bn

people use emails

4TB

of data produced by a connected car

Intel

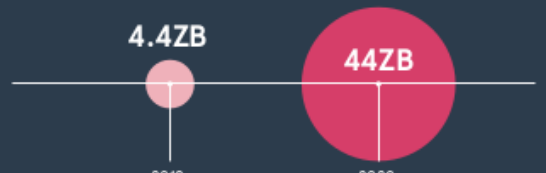
Smart Insights

28PB

to be generated from wearable devices by 2020

Statista

ACCUMULATED DIGITAL UNIVERSE OF DATA



Source: Raconteur

Data is everywhere

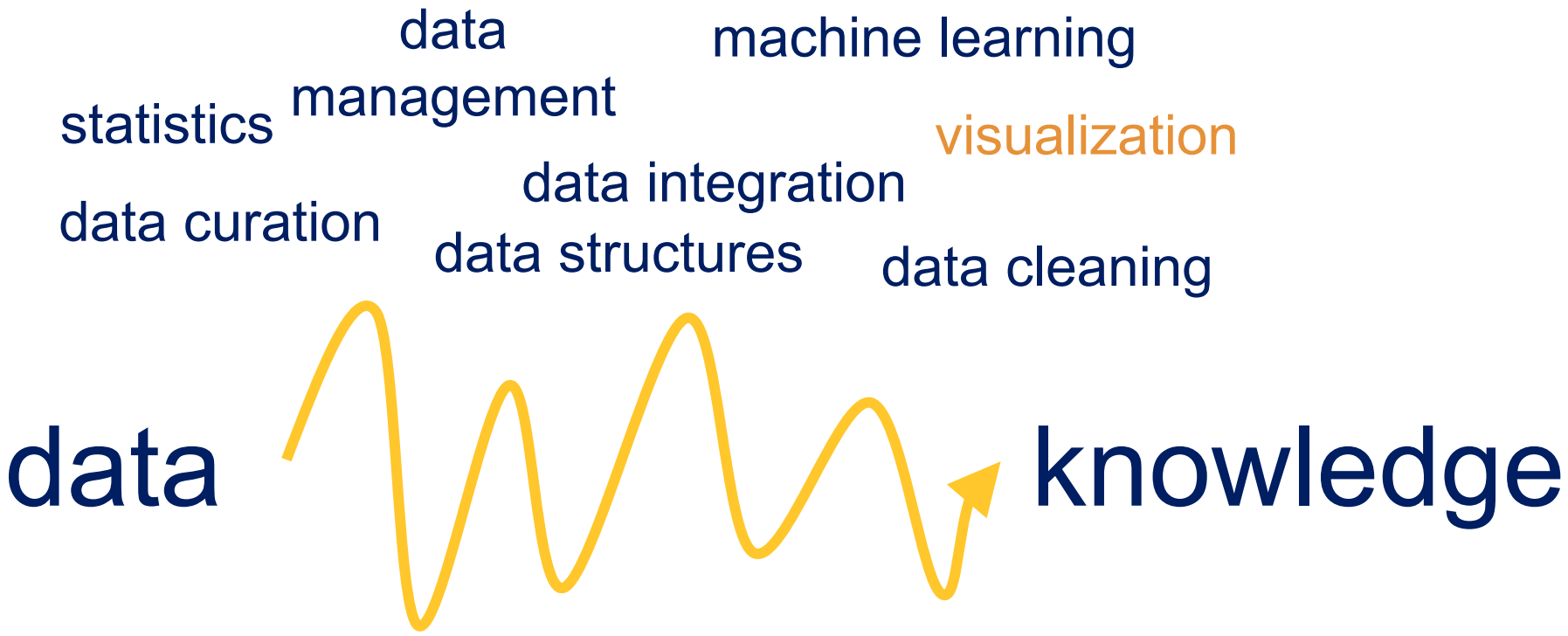
"The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades, … because now we really do have essentially free and ubiquitous data."

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

Data to knowledge

Data to knowledge

Data to knowledge

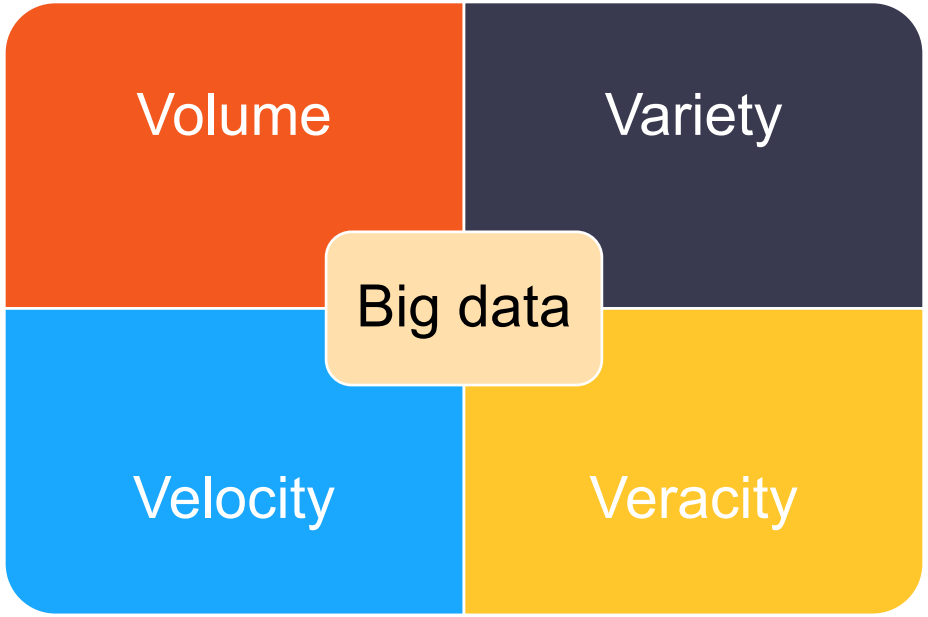


What is Big Data?

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

Large-scale collections or lots of (not necessarily big) data sets

Needs to be processed quickly
- Streaming
- Interactivity



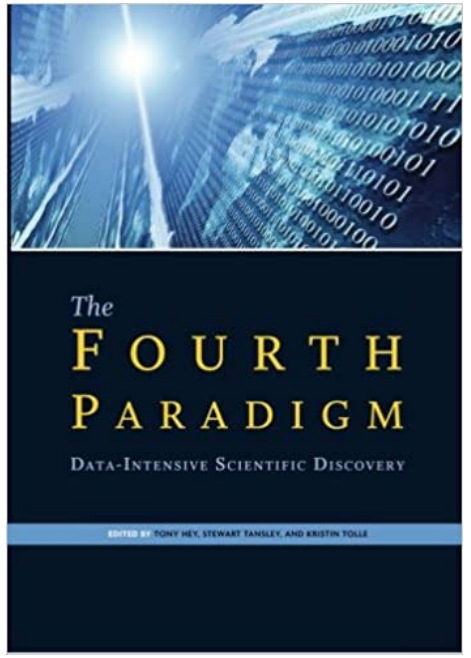
Does not fit neatly in an existing tool
- Data sets that need to be cleaned, processed and integrated
- E.g., Twitter, news, geometries, news, images, videos

Different levels of accuracy and quality
- Uncertainty

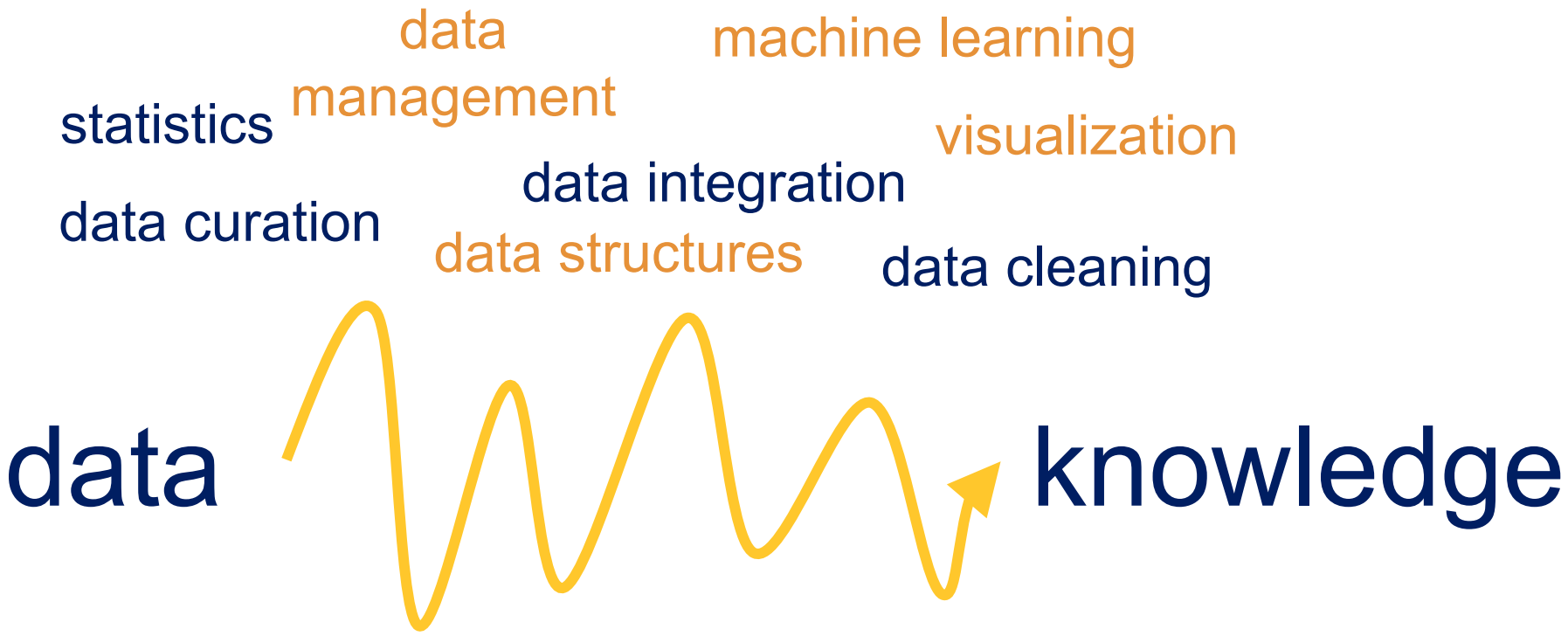
The big deal of big data

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.

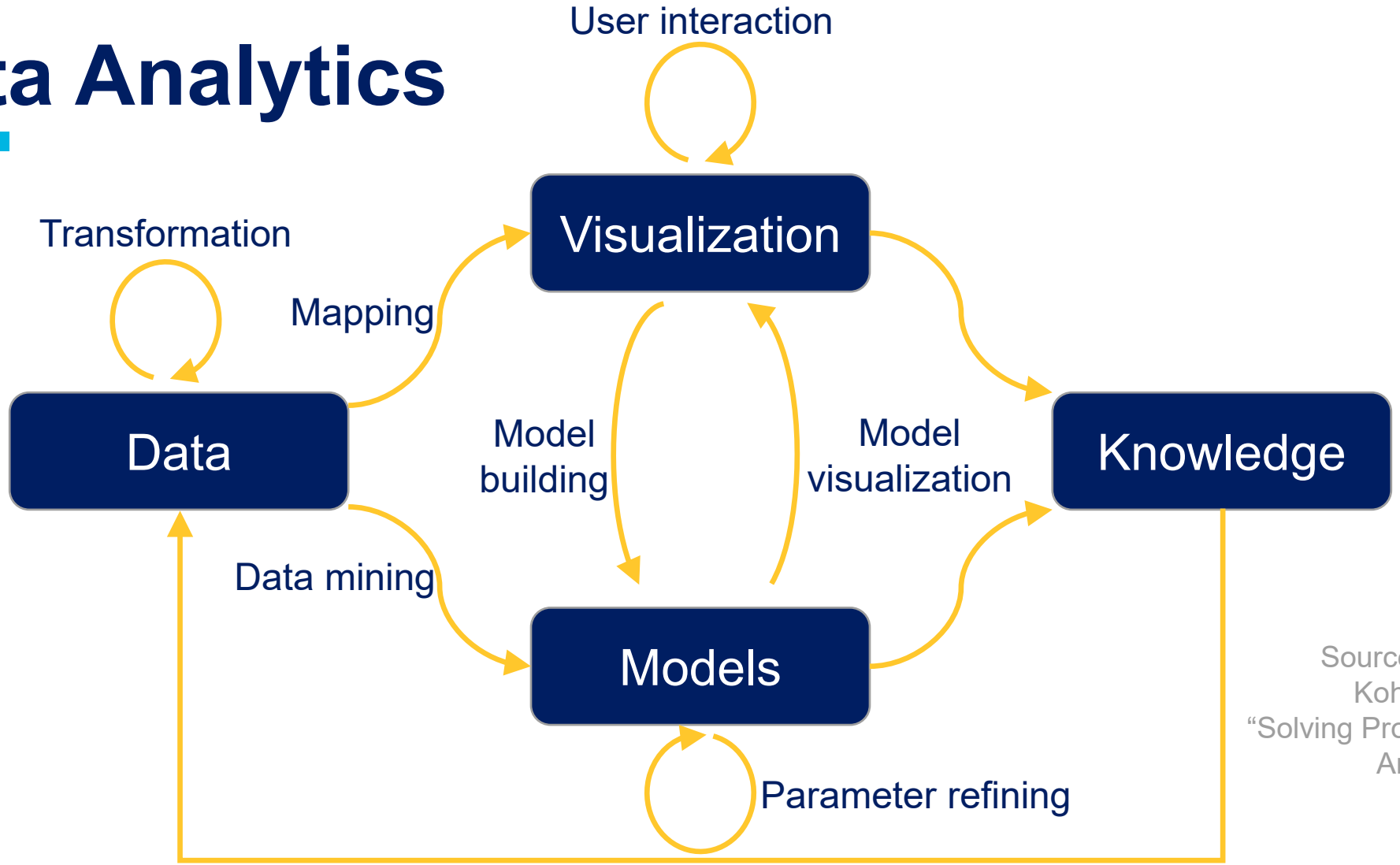
Big data analysis: What we need

- 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*".

    [Liu and Heer, IEEE TVCG 2014]
- 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*".

    [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***."

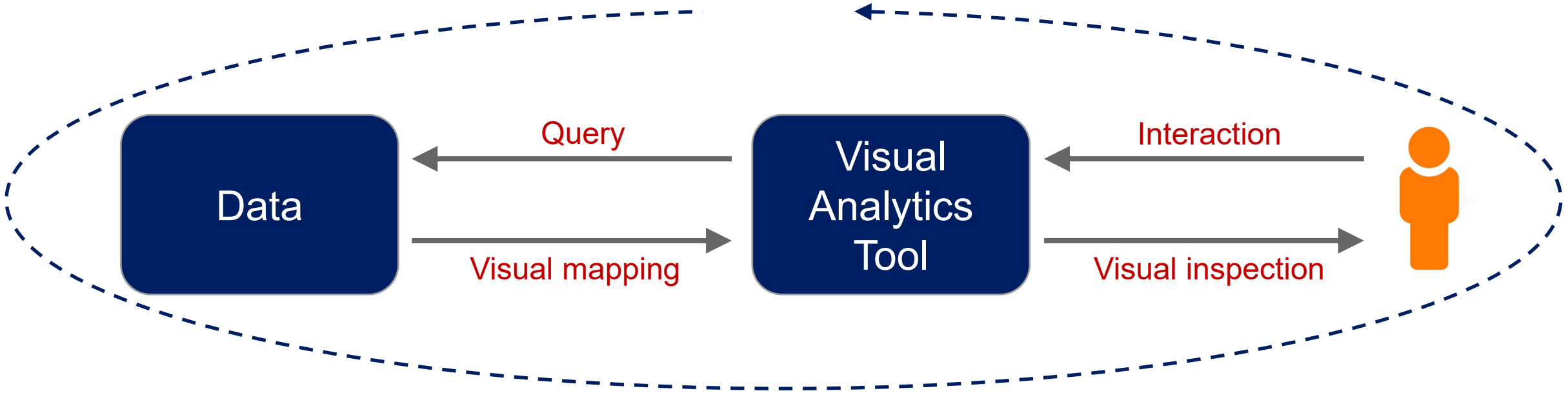
[Battle, Chang, Stonebraker SIGMOD 2016]

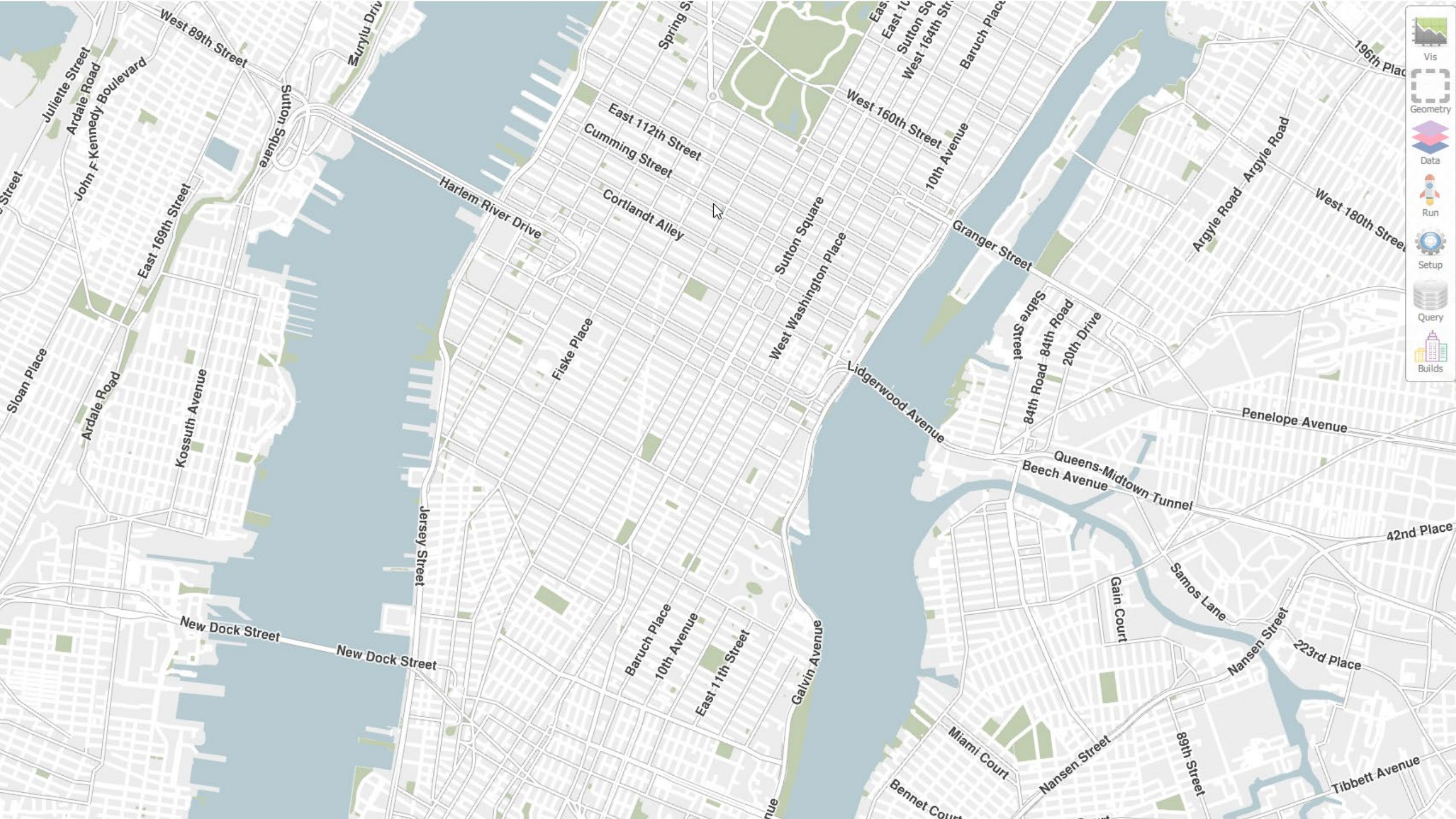
Big data challenges

- 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 formulates 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




West 89th Street
Juliette Street
Ardale Road
John F Kennedy Boulevard
Sutton Square
East 169th Street
Harlem River Drive
East 112th Street
Cumming Street
Cortlandt Alley
West 160th Street
10th Avenue
Granger Street
Sutton Square
West Washington Place
Fiske Place
Sloan Place
Ardale Road
Kossuth Avenue
Jersey Street
Lidgerwood Avenue
Sabre Street
84th Road
20th Drive
Penelope Avenue
Queens-Midtown Tunnel
Beech Avenue
42nd Place
Argyle Road
West 180th Street
New Dock Street
New Dock Street
Baruch Place
10th Avenue
East 11th Street
Galvin Avenue
Gain Court
Samos Lane
Nansen Street
223rd Place
Miami Court
Nansen Street
89th Street
Tibbett Avenue
Bennet Court
Vis
Geometry
Data
Run
Setup
Query
Builds

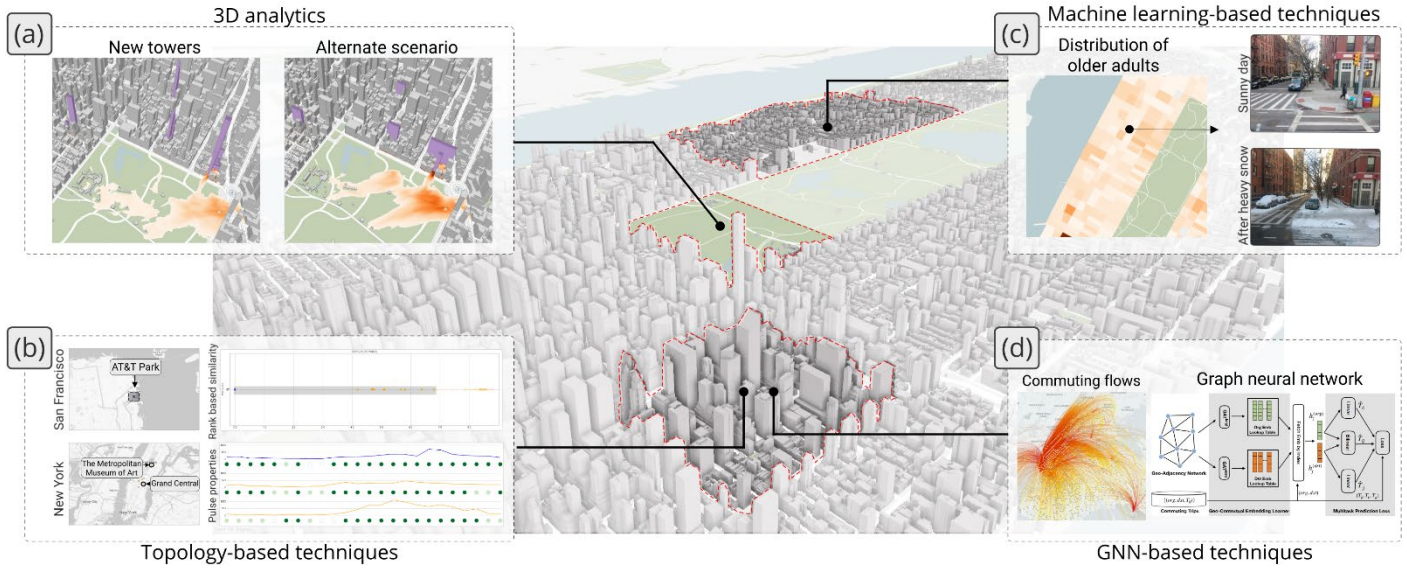
Research

Assistant Professor, CS (UIC)

PhD, CS 2018 (NYU)

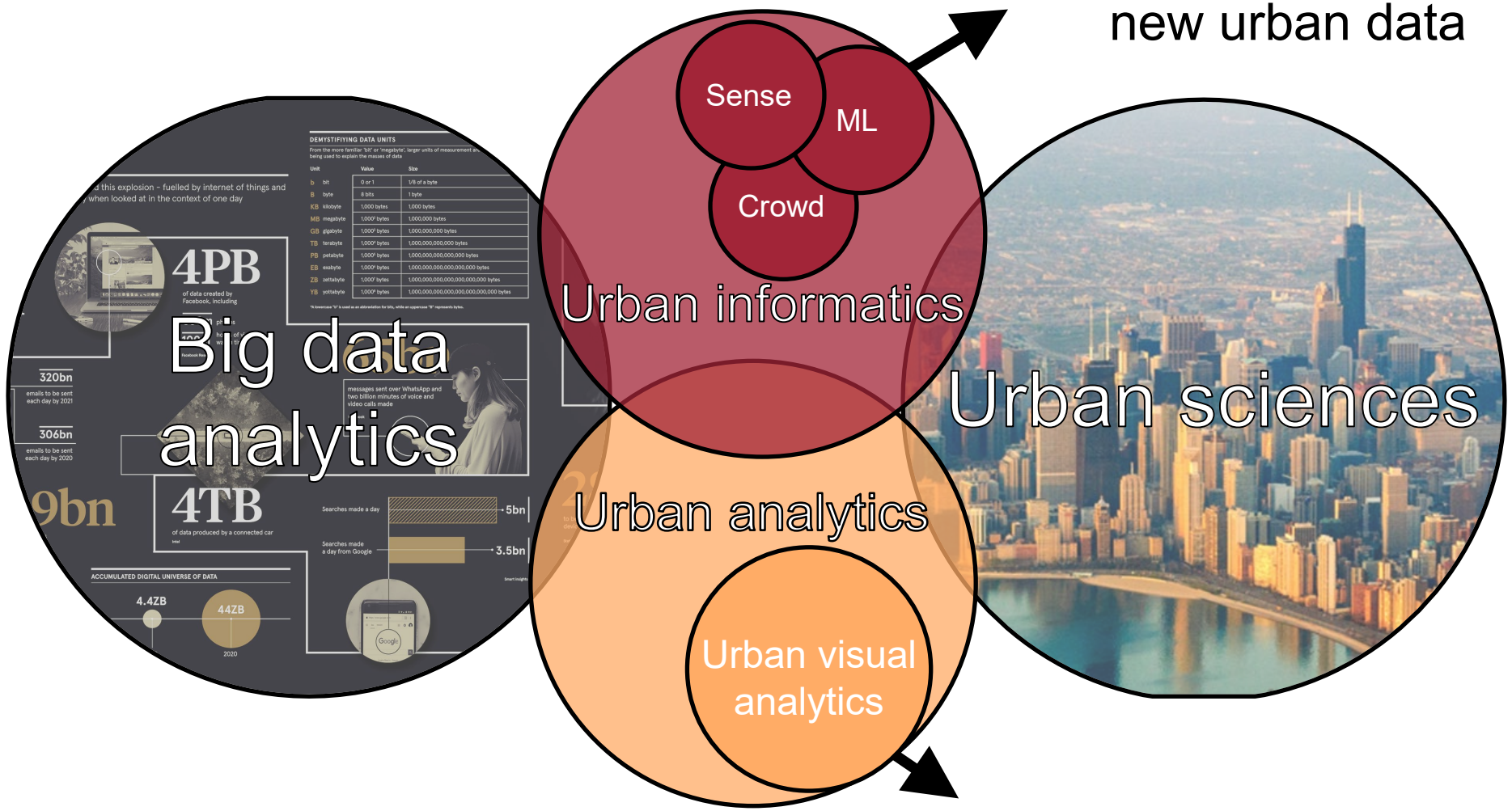
MSc, CS 2012 (PUC-Rio)

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.

Techniques to extract and generate new urban data



Visual analytics tools and systems to analyze large urban data

Visual Analysis of Uncertainty in Trajectories [Lu et al., 2014]

TrajGraph [Huang et al., 2015]

SemanticTraj [Al-Dohuki et al., 2016]

Spatiotemporal Statistical Data [Kim et al., 2017]

Forecasting Road Traffic Congestion [Lee et al., 2019]

A study of new york city taxi trips [Ferreira et al., 2013]

MaraVis [Li et al., 2020]

SEEVis [Li et al., 2020]

Air Pollution Problem in Hong Kong [Qu et al., 2007]

Air Temperature [Gautier et al., 2020]

Urban Mosaic [Miranda et al., 2020]

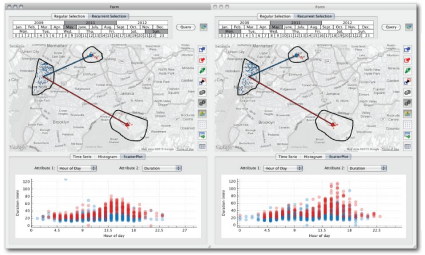
Location2vec [Zhu et al., 2019]

A correlative analysis process in a visual analytics environment [Malik et al., 2012]

HydroQual [Accorsi et al., 2014]

Shadow accrual maps [Miranda et al., 2018]

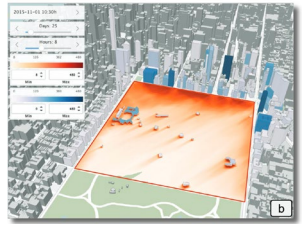
Urban Pulse [Miranda et al., 2016]



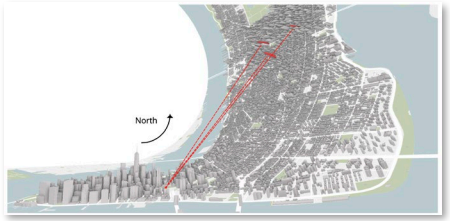
TaxiVis
(Ferreira et al., 2013)

Urbane
(Ferreira et al., 2015)

Catalogue
(Doraiswamy et al., 2015)



Shadow Profiler
(Miranda et al., 2019)

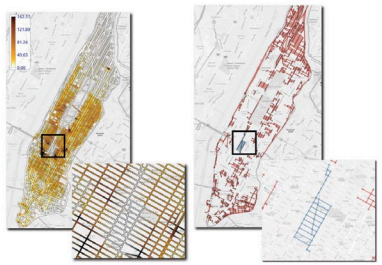


UrbanRama
(Chen et al., 2020)

UTK
(Moreira et al., 2023)

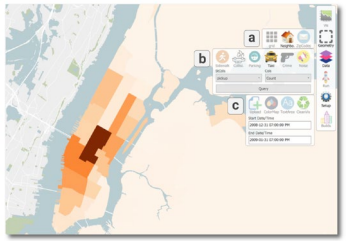
2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023

Taxi Patterns
(Doraiswamy et al., 2016)



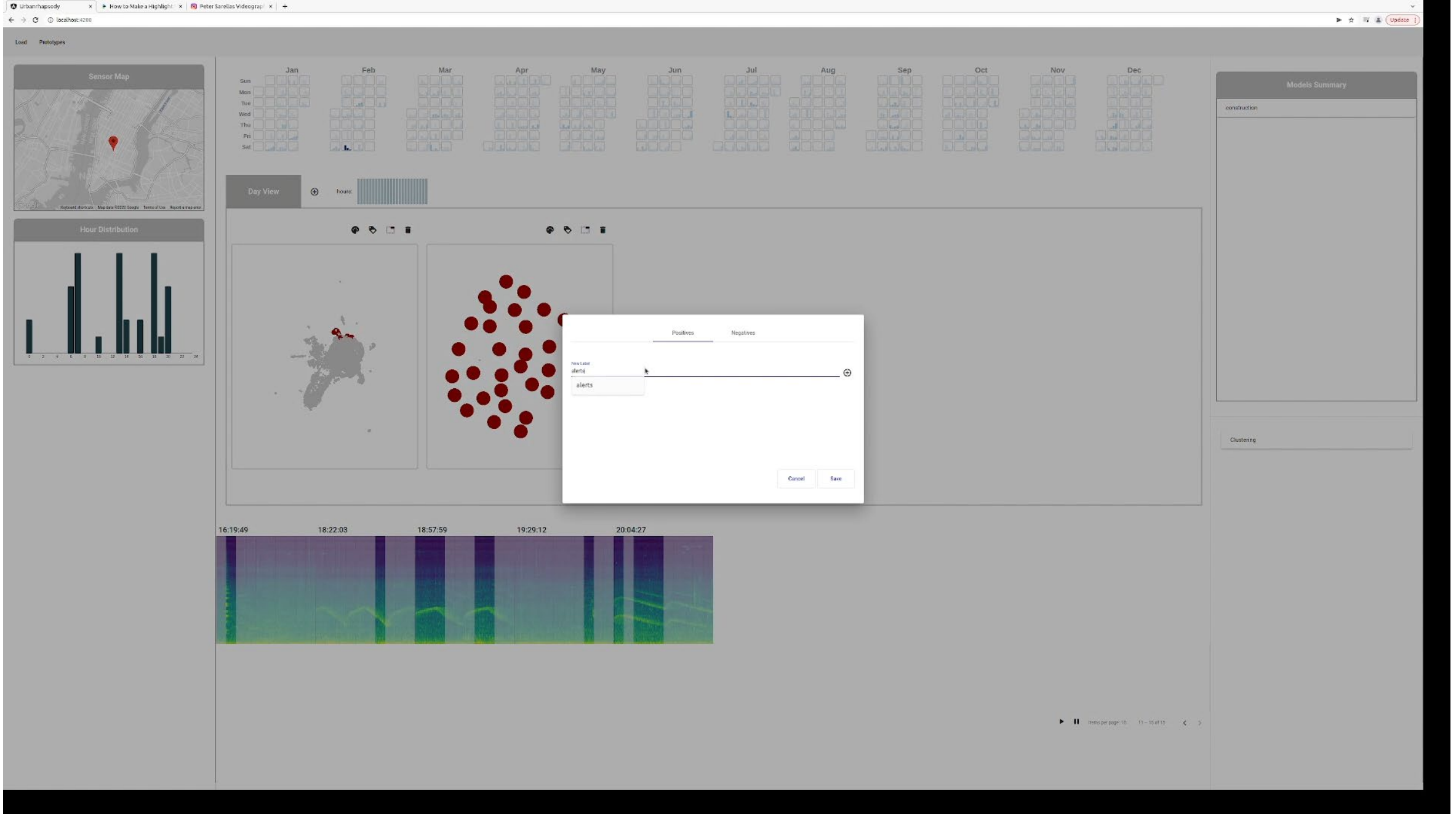
Urban Pulse
(Miranda et al., 2016)

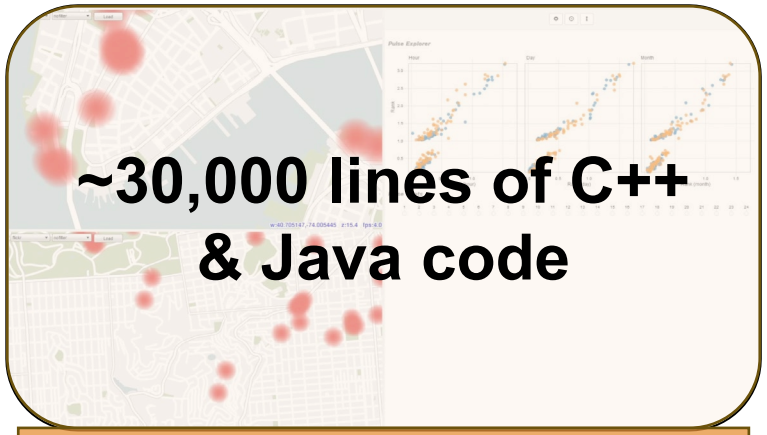
Raster-Join
(Doraiswamy et al., 2018)



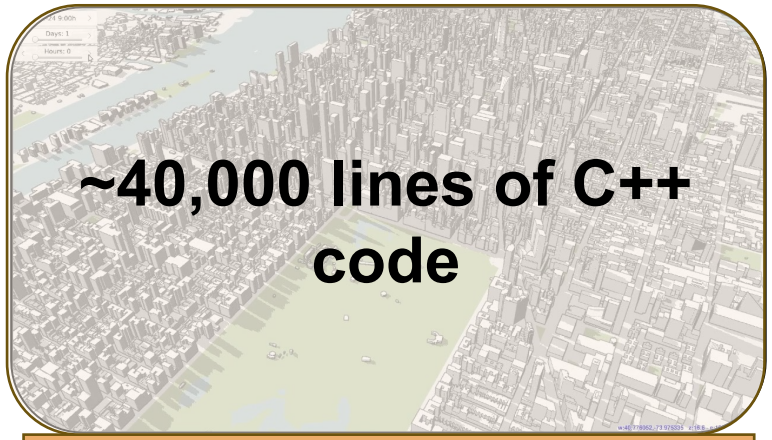
Urban Mosaic
(Miranda et al., 2020)

Urban Rhapsody
(Rulff et al., 2022)

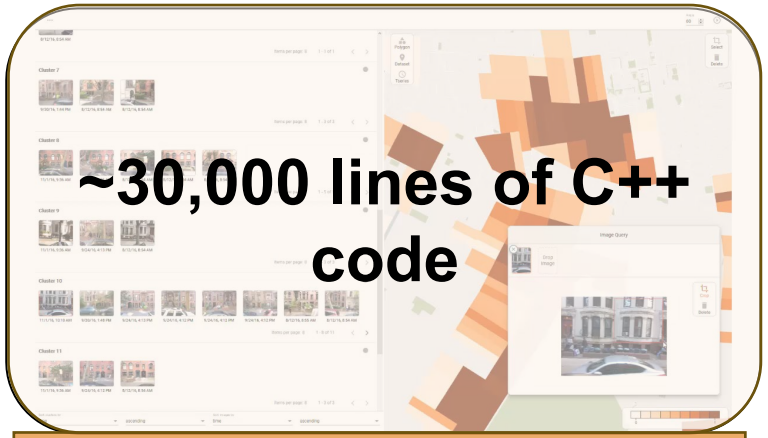

Load
Prototypes
Sensor Map
Hour Distribution
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Sun
Mon
Tue
Wed
Thu
Fri
Sat
Day View
hours
Models Summary
construction
Clustering
Positives
Negatives
alerts
Cancel
Save
16:19:49
18:22:03
18:57:59
19:29:12
20:04:27



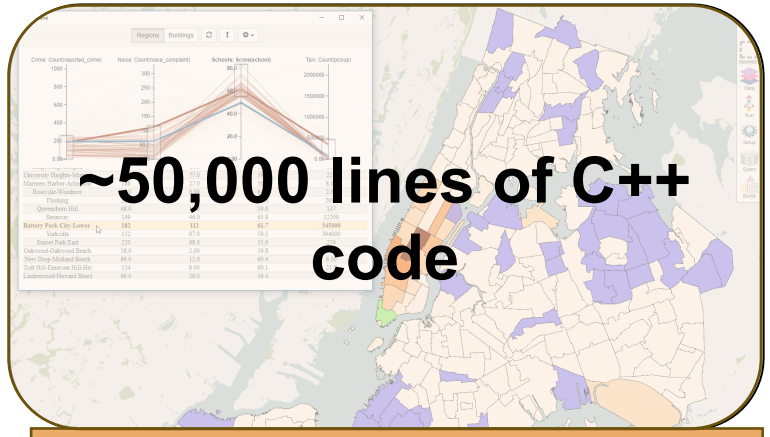
Urban Pulse:
Large-scale data mining of social media data



Shadow Profiler:
City-scale assessment of sunlight access

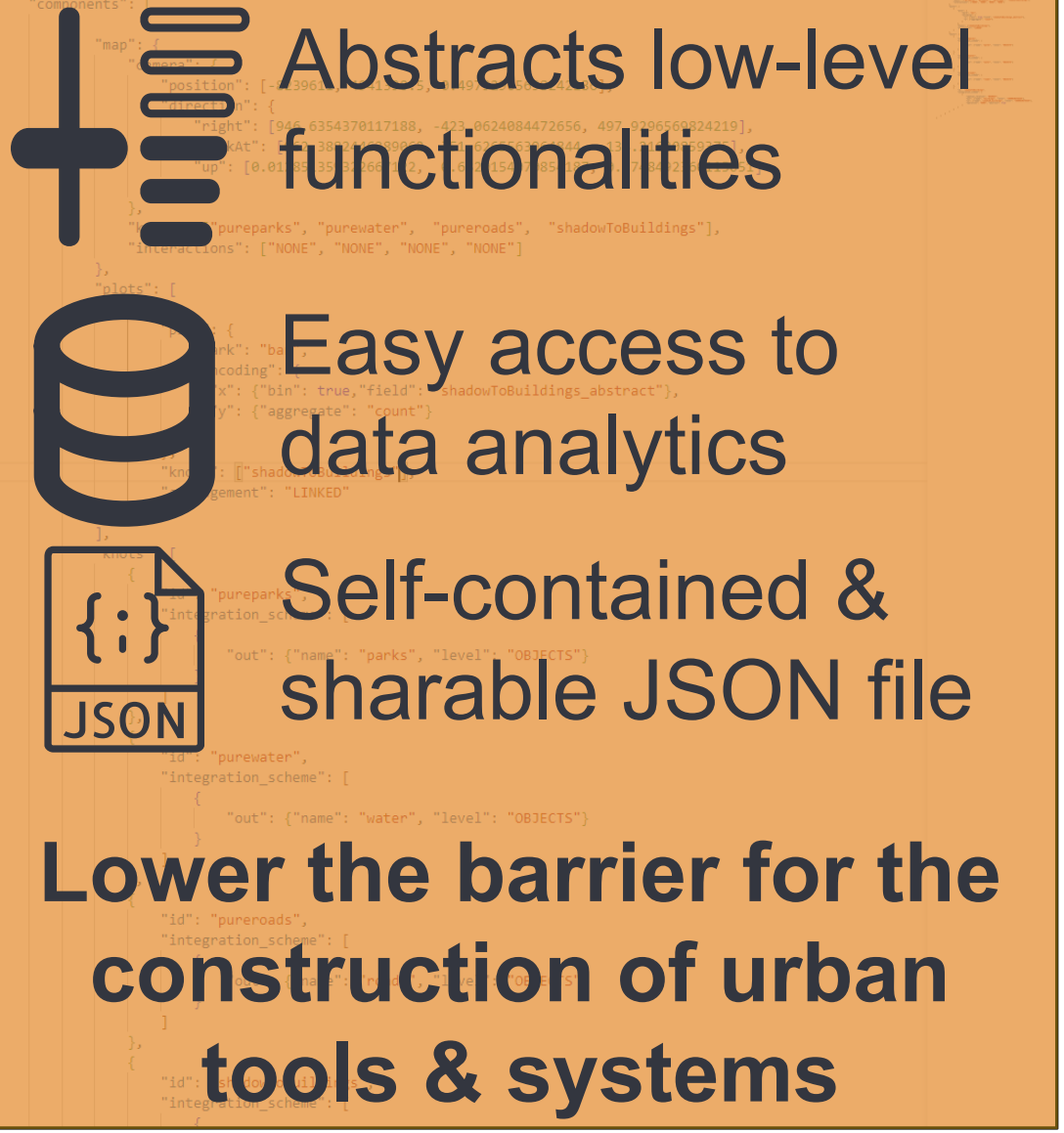


Urban Mosaic:
Interactive exploration of large imagery data



Urbane:
Interactive exploration of large data

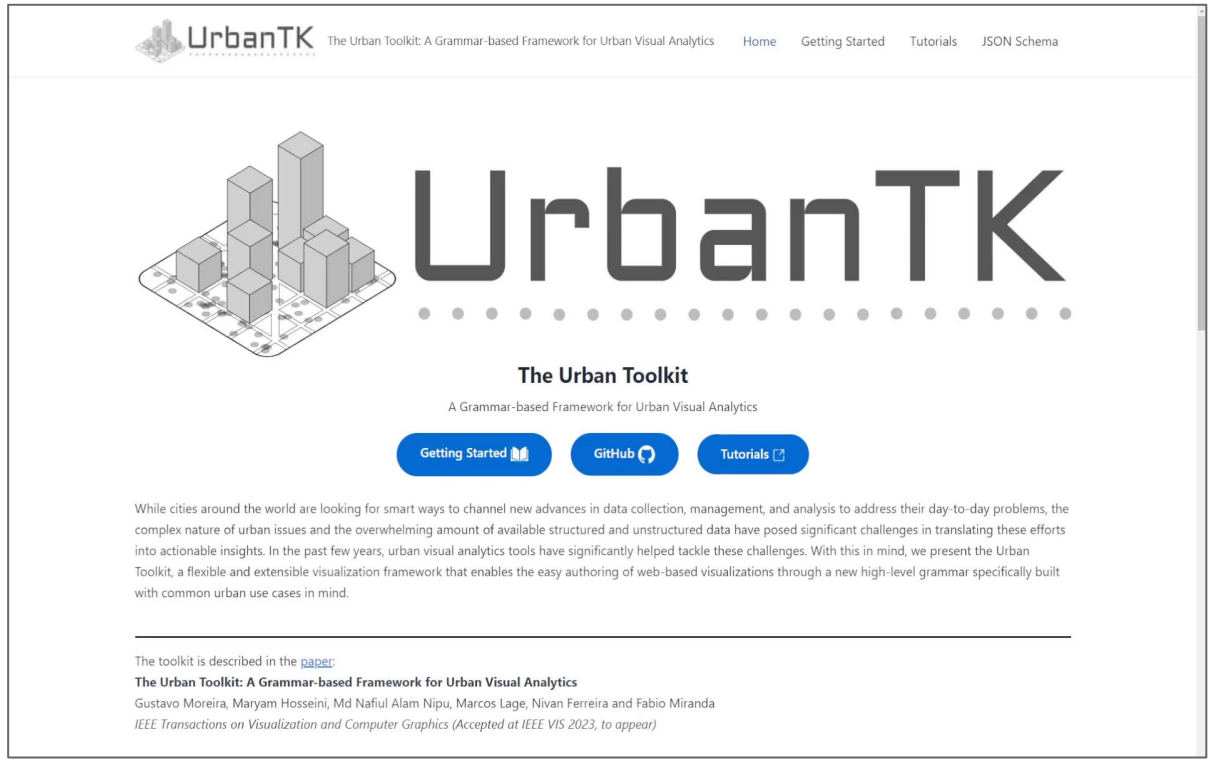
```cpp
#include "UrbaneMapView.hpp"

#include <QApplication>
#include "../MapView/BuildingRenderingLayer.hpp"

#include "../MassingGeneration/massinggeneration.h"
#include "../Util/ColorMapDivergent.hpp"
#include "UrbaneManager.hpp"

#include <QElapsedTimer>
#include <QThread>
#include <QDir>

#include <vector>

UrbaneMapView::UrbaneMapView(const QString &filename, const QRectF &vp, QWidget *parent)
    : MapView(filename,vp,parent), graphLayer(NULL)
{
    initialized = false;
    skyExposureData = false;
    this->centerIndex = GridIndex(1024,1024);
    this->currentLayer = NULL;
    this->lotUpdate = true;
}

UrbaneMapView::~UrbaneMapView() {}

void UrbaneMapView::initializeGL() {
    if(!initialized) {
        MapView::initializeGL();
        this->buildingScore.initComputeShader();
        this->skyScore.initComputeShader();
    }
    initialized = true;
}

void UrbaneMapView::paintGL()
{
    this->showOsd(false);

    // Lot data initialization in manager
    // TODO Don't know of a better place to do this
    if(lotUpdate && this->parcelLayer->isDataReady()) {
        updateLotDataDB();
        lotUpdate = false;
    }

    UrbaneManager *manager = UrbaneManager::getInstance();
    QPair<RenderingOperation, UIOperation> state = manager->getState();

    RenderingOperation operation = state.first;
    UIOperation what = state.second;
    switch(operation) {
    case RenderingOperation::UpdateVis:
    {
        bool updateFunction = false;
```



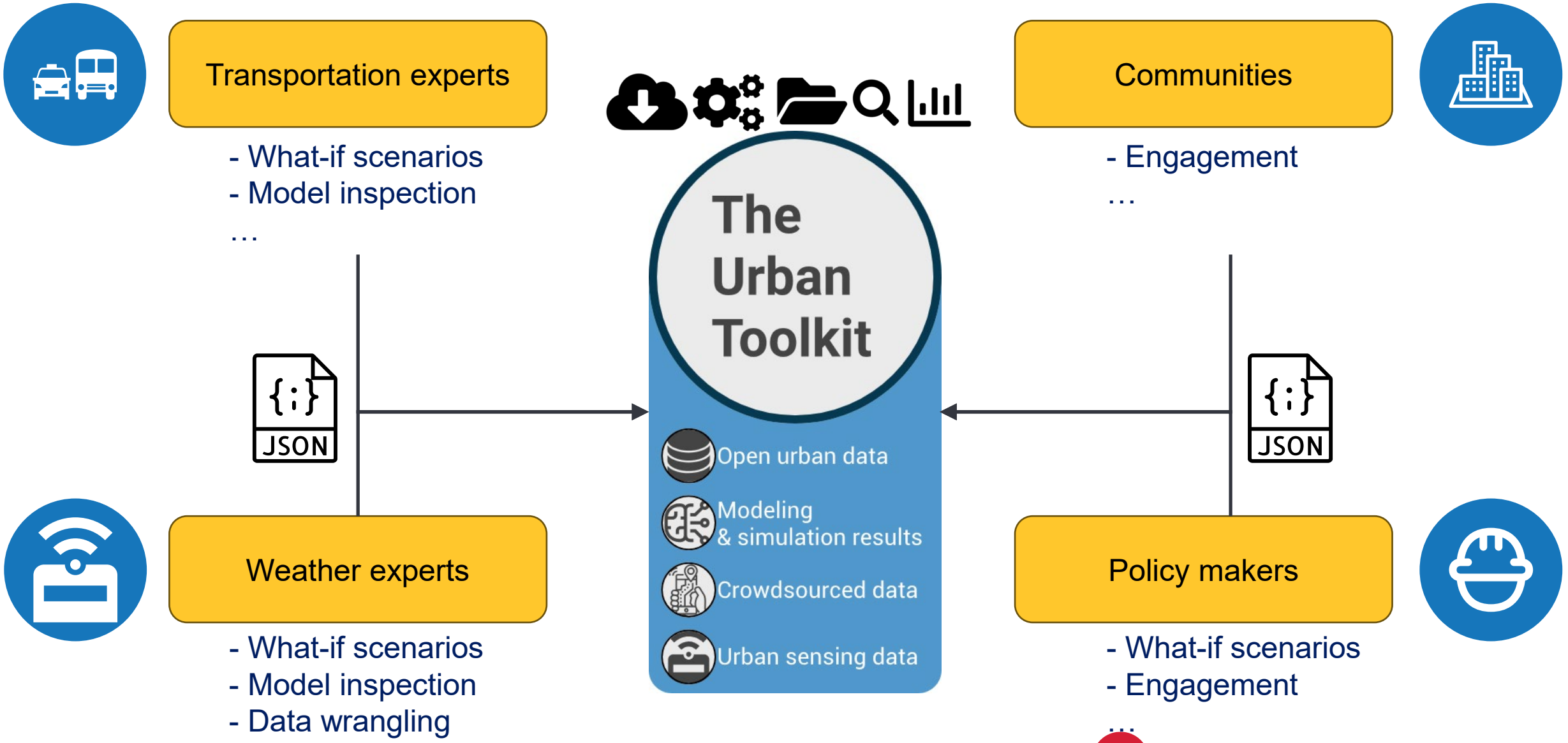
UrbanTK

Gustavo Moreira, University of Illinois Chicago
Maryam Hosseini, Massachusetts Institute of Technology
Md Nafiul Alam Nipu, University of Illinois Chicago
Marcos Lage, Universidade Federal Fluminense
Nivan Ferreira, Universidade Federal de Pernambuco
Fabio Miranda, University of Illinois Chicago

Code & tutorials:
urbantk.org



[Moreira, VIS 2023]


Transportation experts
- What-if scenarios
- Model inspection
…
Communities
- Engagement
…
The Urban Toolkit
Open urban data
Modeling & simulation results
Crowdsourced data
Urban sensing data
JSON
JSON
Weather experts
- What-if scenarios
- Model inspection
- Data wrangling
…
Policy makers
- What-if scenarios
- Engagement
…

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

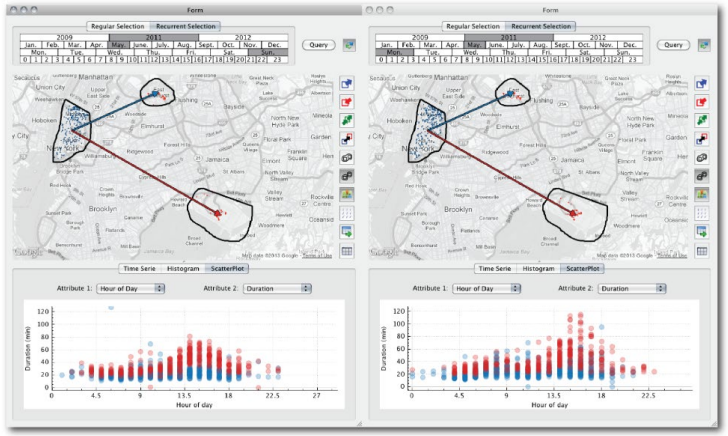
- Necessary techniques to build visual analytics tools to handle big data:
  1. Building blocks: current technologies and libraries to build visual analytics systems for big data.
  2. Visualization: visual implications of handling big data.
  3. Data management: techniques to handle big data.
  4. Analytics: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

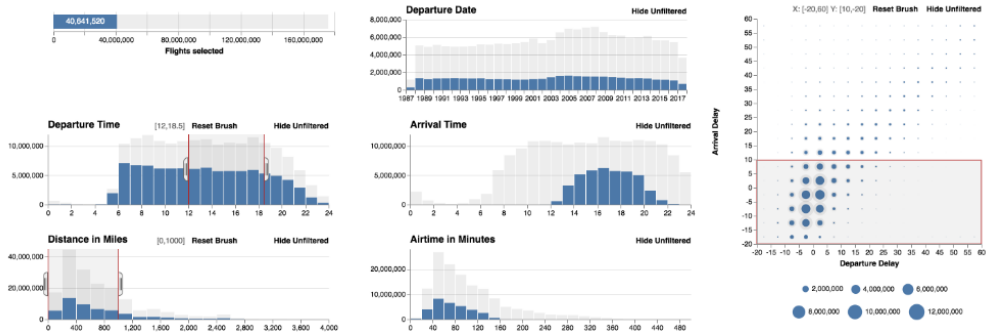
Building blocks:

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

Visualization:

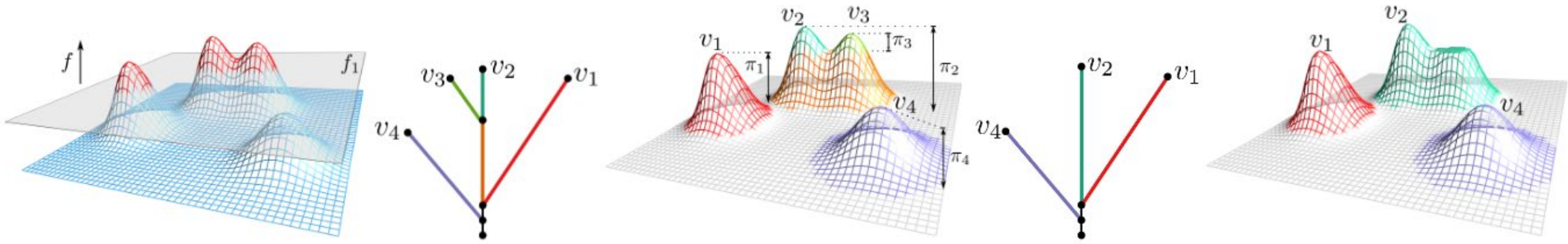
- Visual analytics systems
  - Interactivity requirements.
  - Components of the system.
- Progressive visualization
  - Progressively build visualizations, maintaining interactivity.
- Uncertainty visualization
  - Defining uncertainty.
  - Displaying uncertainty.
- Visualization for machine learning





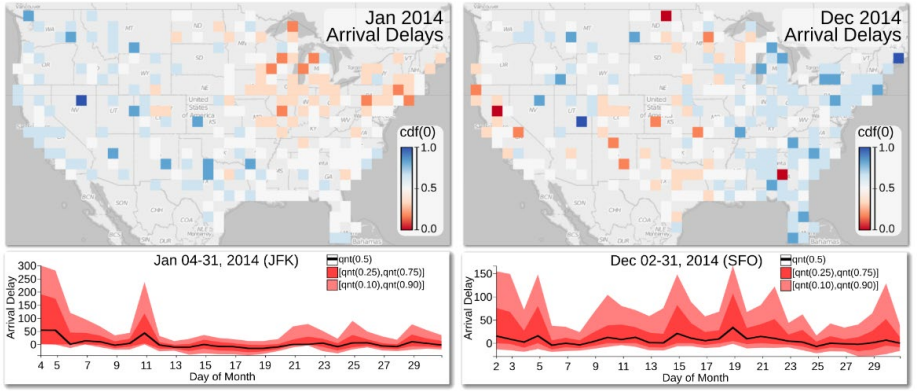
Analytics:

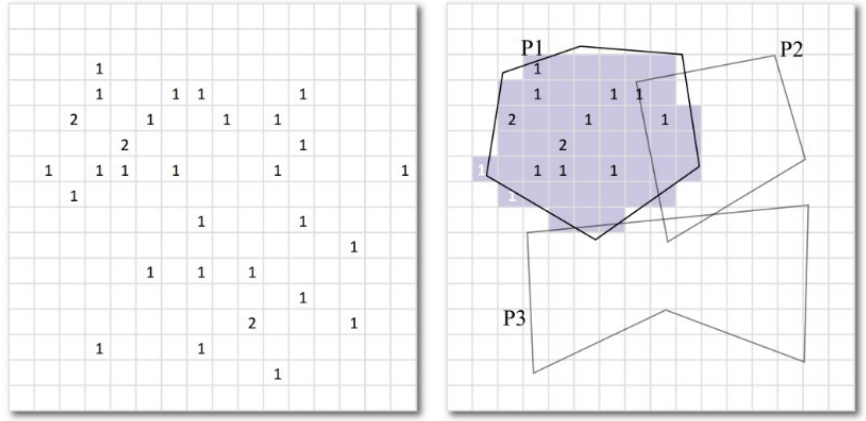
- Data mining and analytics techniques to extract features and patterns from big data:
  - Computational topology
  - Wavelet
  - Techniques for streaming data
- Machine learning for visualization



Data management:

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





(a) (b)

CS524: Big Data Visualization & Analytics