Principles for Working with Big Data

Juliana Freire
Visualization and Data Analysis (ViDA) Lab
Computer Science & Engineering
Center for Urban Science & Progress (CUSP)
Center for Data Science
New York University

NYU POLYTECHNIC SCHOOL OF ENGINEERING

CENTER FOR URBAN SCIENCE+PROGRESS

The Big Data Analysis Pipeline

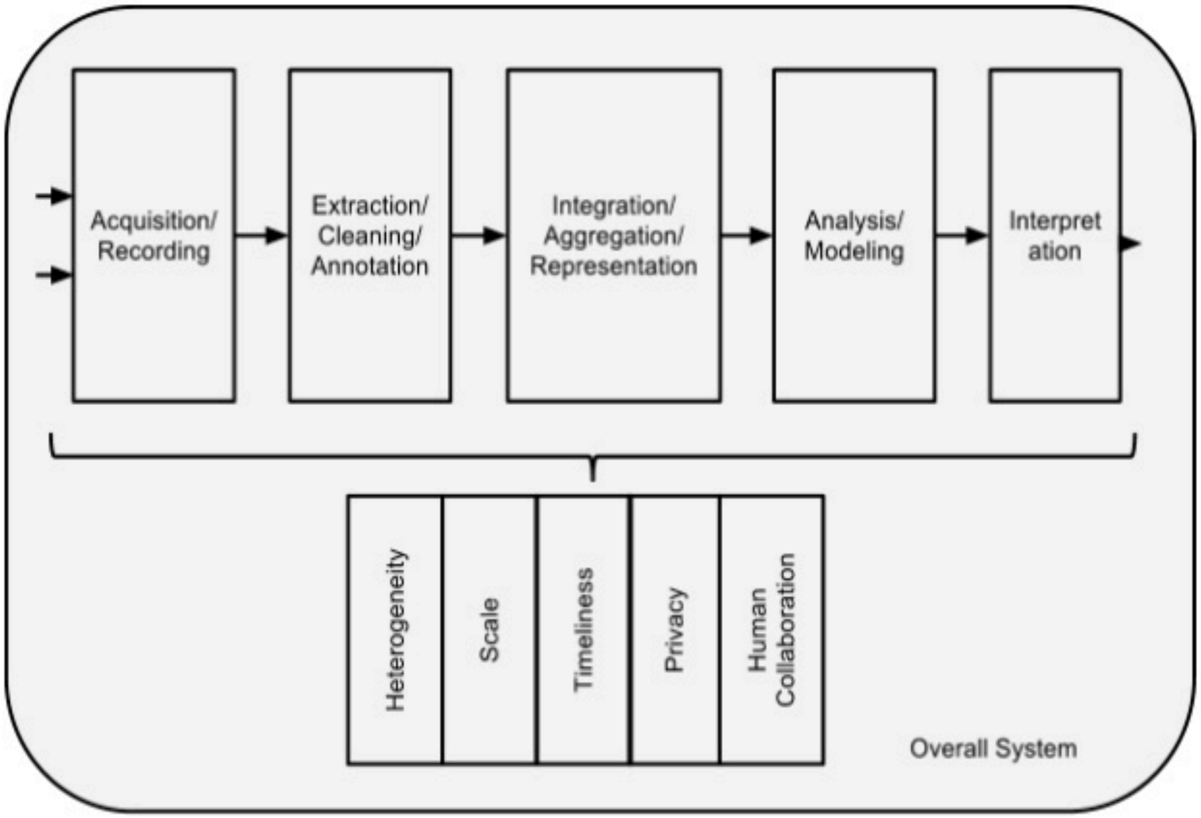
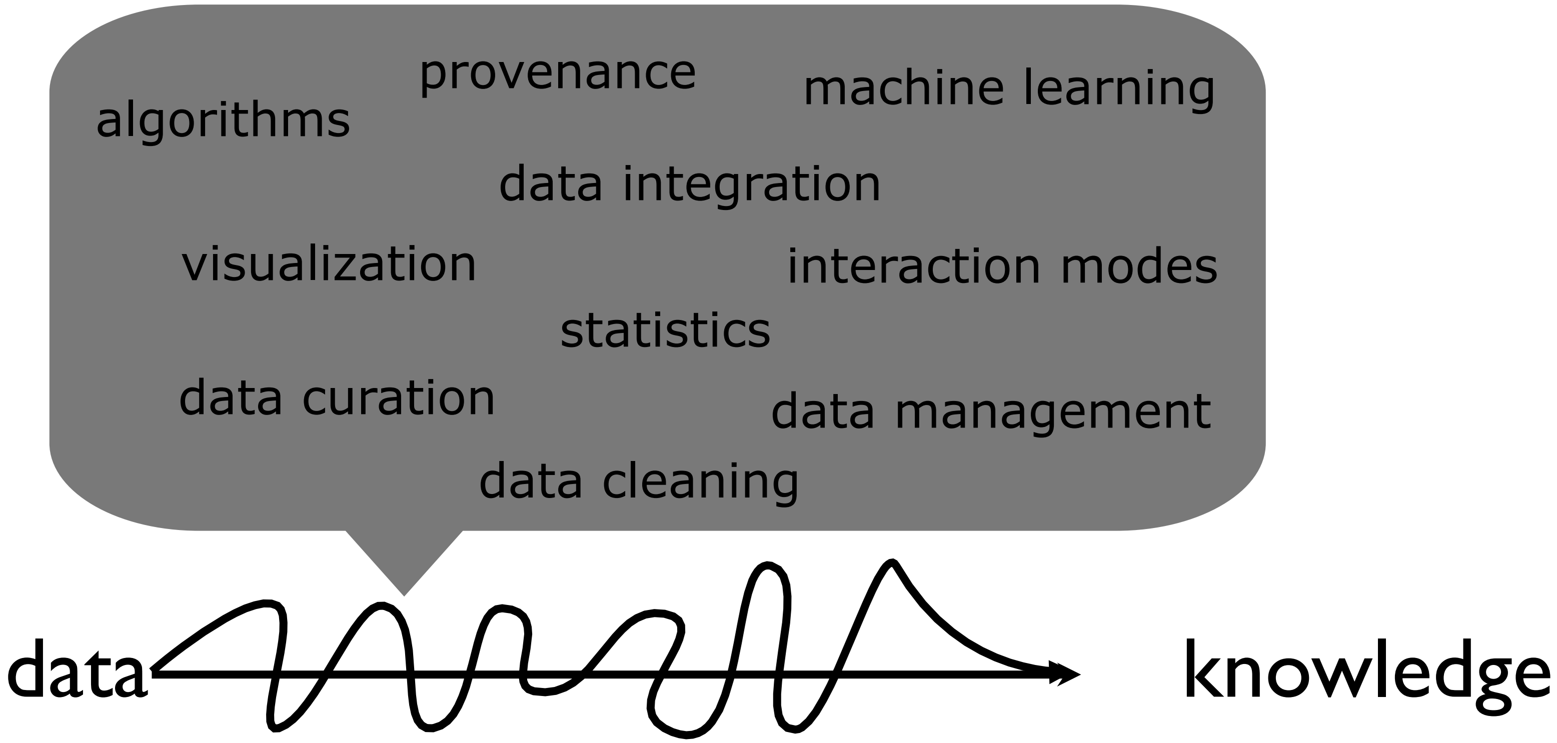


Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

Challenges and Opportunities with Big Data, CRA 2012
`http://cra.org/ccc/docs/init/bigdatawhitepaper.pdf`

Big Data: What is hard?

- Scalability for batch computations is *not* hard
  - Lots of work on distributed systems, parallel databases, …
  - Elasticity: Add more nodes!
- Scalability for people is!
  - Data exploration is hard regardless of whether data are big or small



algorithms

provenance

machine learning

data integration

visualization

interaction modes

statistics

data curation

data management

data cleaning

data

knowledge

Principles for Working with Data Big Data

Information Integration

Statistics

Programming

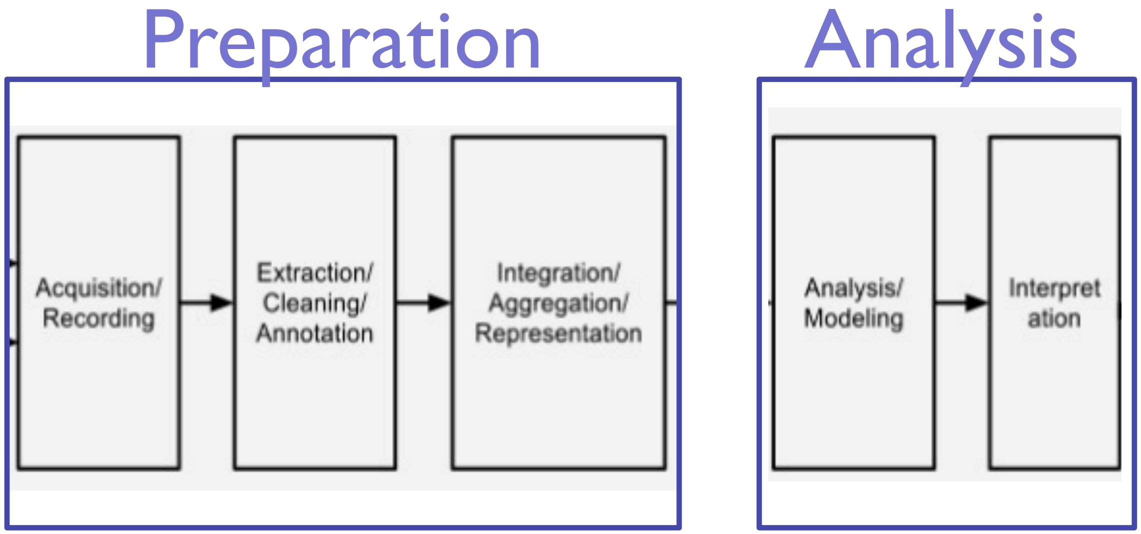
Data Management

Machine Learning

MapReduce + Hadoop ecosystem

Data Mining

Visualization



Preparation

Acquisition/ Recording

Extraction/ Cleaning/ Annotation

Integration/ Aggregation/ Representation

Analysis

Analysis/ Modeling

Interpretation

(Big and Small) Data Exploration

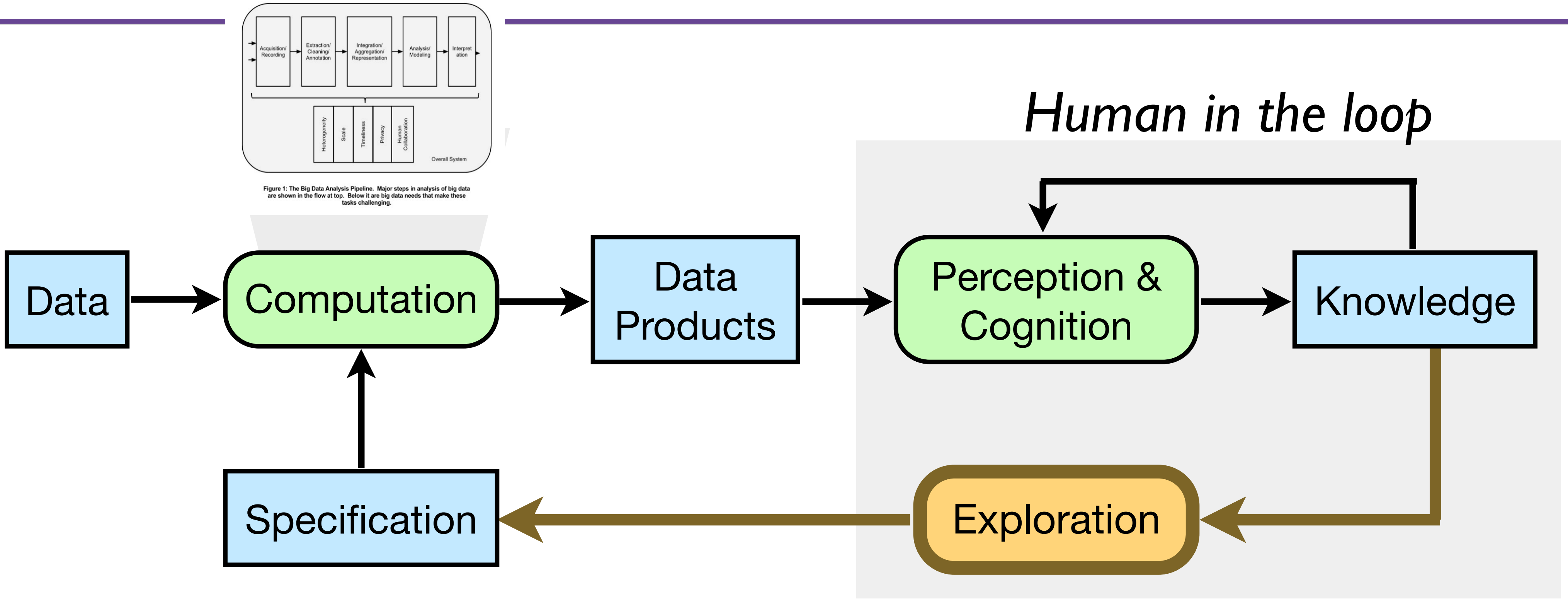


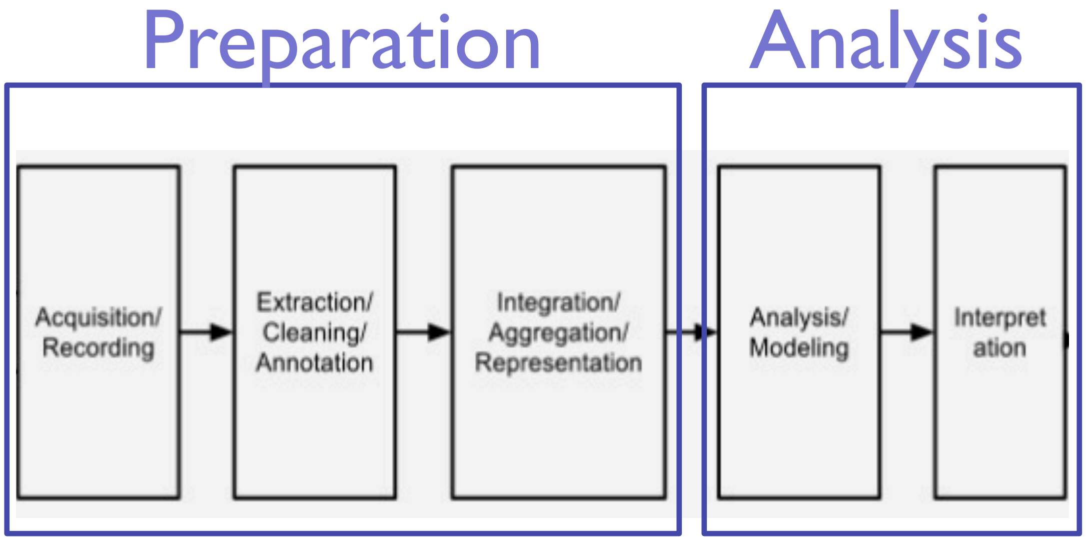
Figure 1: The Big Data Analysis Pipeline. Major steps in analysis of big data are shown in the flow at top. Below it are big data needs that make these tasks challenging.

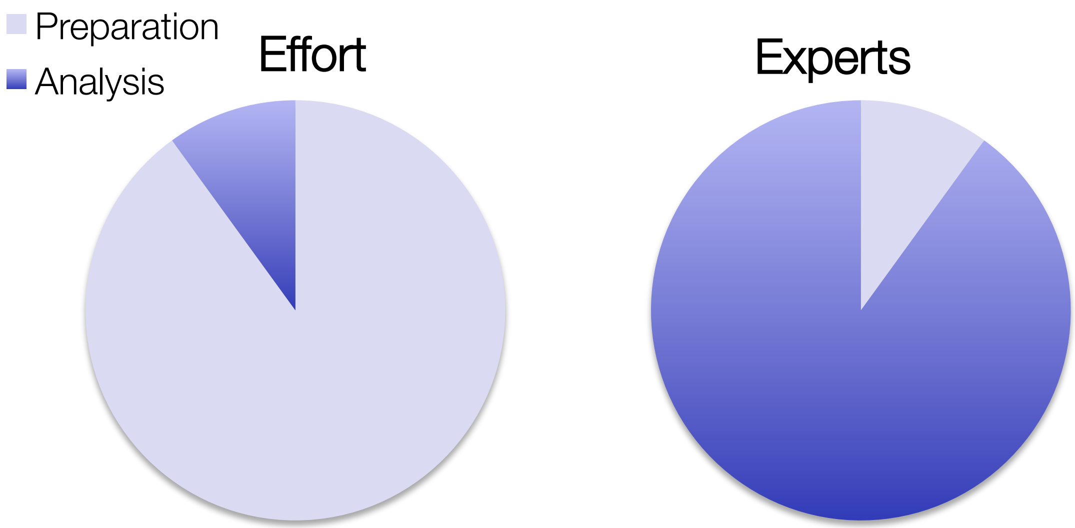
[Modified from Van Wijk, Vis 2005]

- Iterative process to generate and test hypotheses
- *Easy to get lost*---derive a result and not remember how you got there
- Need to capture ***provenance*** of the exploration process – for transparency, reproducibility and knowledge re-use

Provenance Management

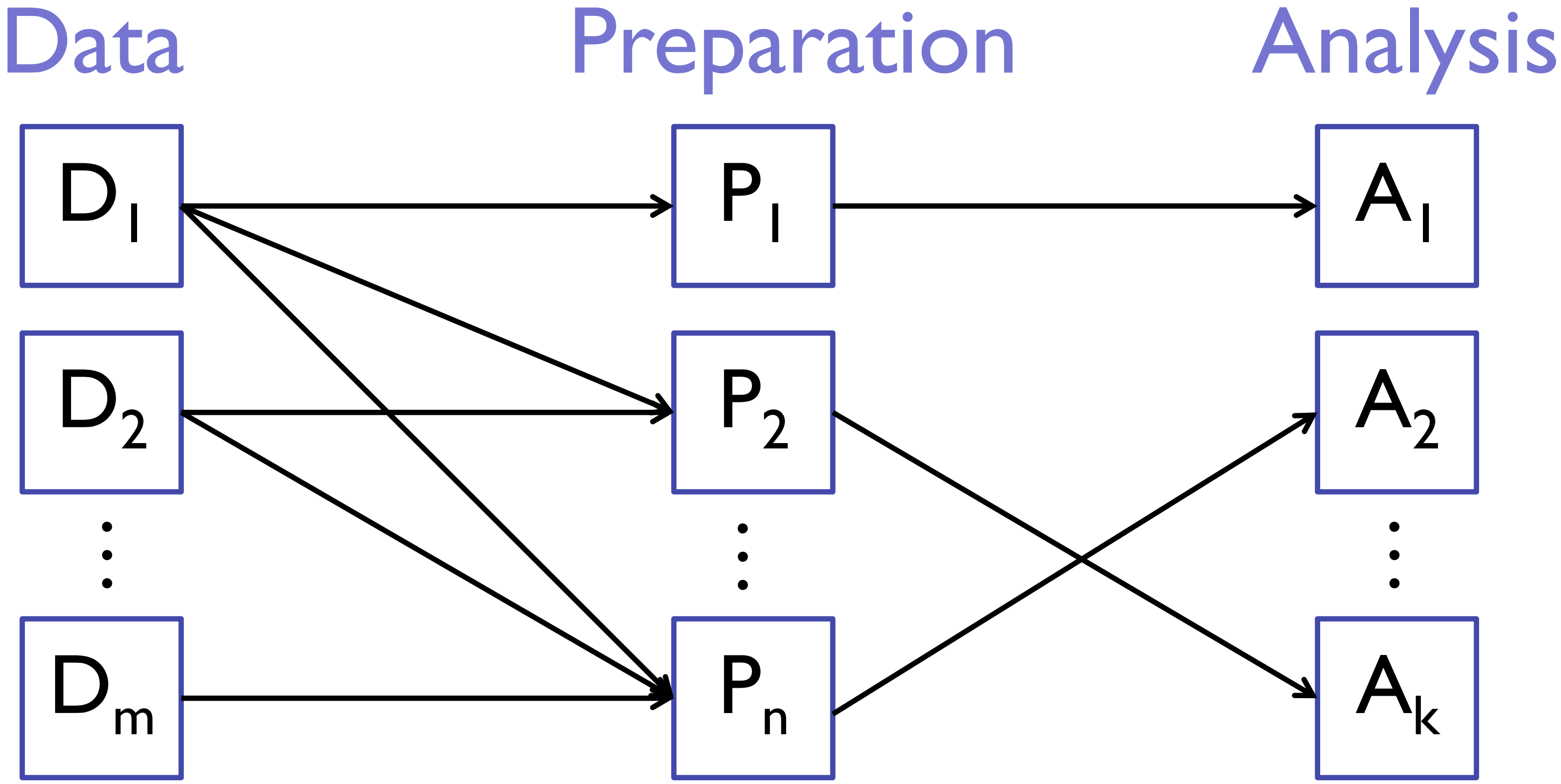
The Big Data Conundrum





The Big Data Conundrum

- Data preparation is a bottleneck
- Limits analyses

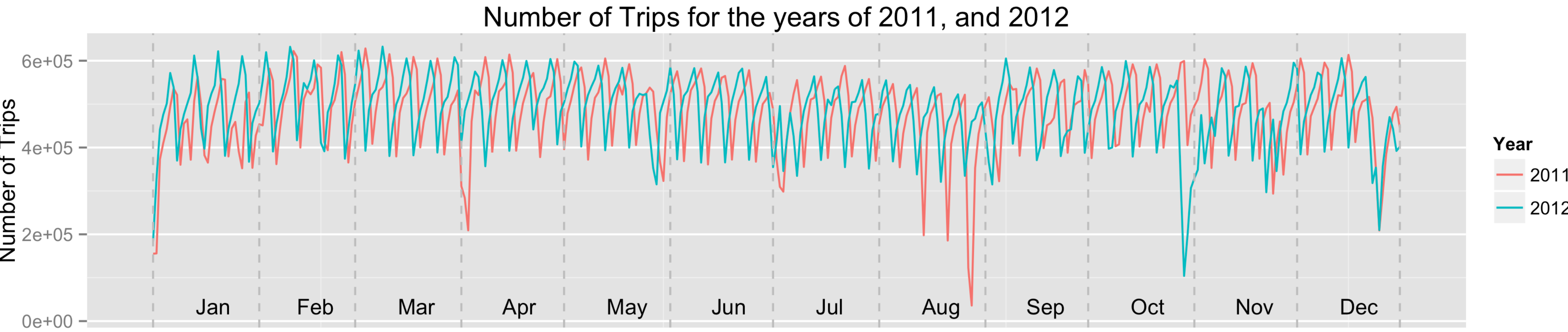


Big Data: A Moving Target

- New data (and analyses) bring new challenges
- Many tools, but many more needs…
- Knowing the principles is key to build effective solutions

Big Data: Experience from the Trenches

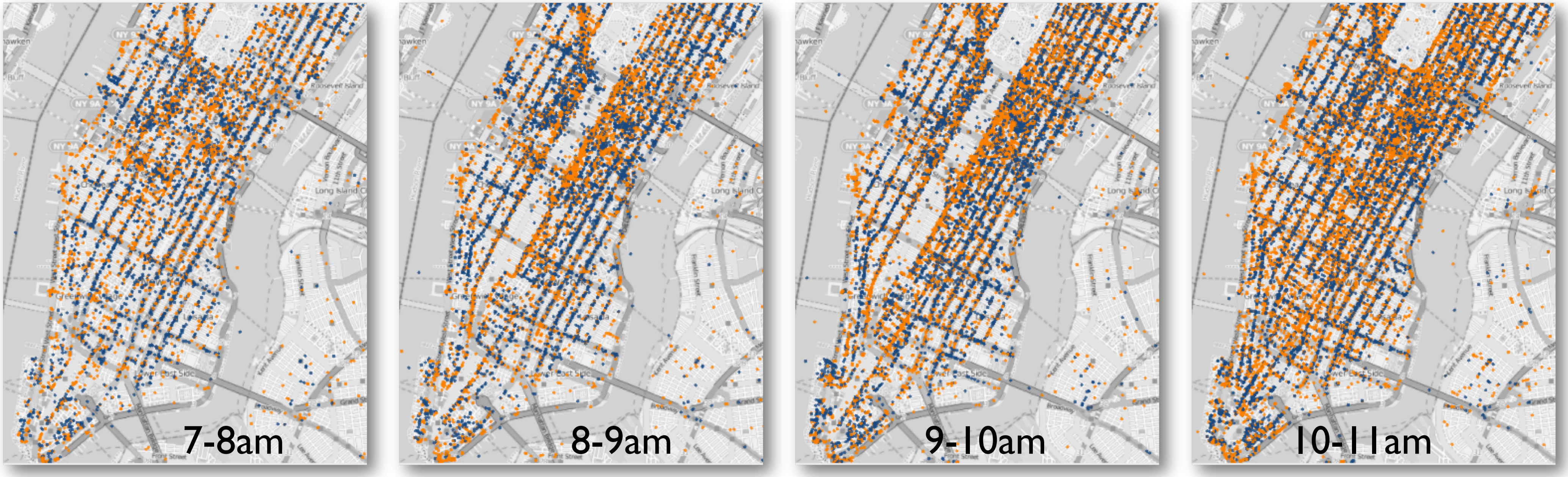
- NYC taxis as sensors for can city life: economic activity, human behavior, mobility patterns, …
- Taxi data are "big", complex and dirty
  - ~500k trips/day
  - Multiple variables: *spatial temporal* + *trip attributes*
- Domain scientists and decision makers are unable to explore the *whole* data



NYU POLYTECHNIC SCHOOL OF ENGINEERING

CUSP CENTER FOR URBAN SCIENCE+PROGRESS

A Study of NYC Taxis

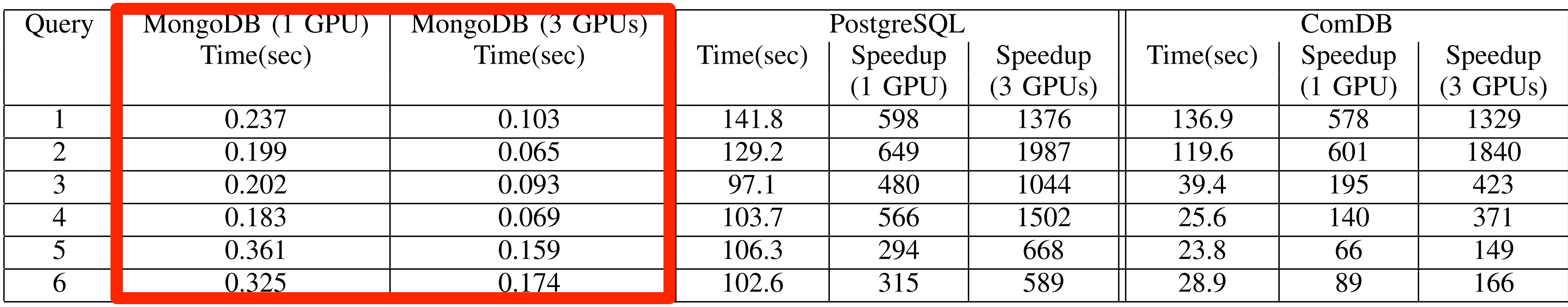


- Requirement: support interactive queries
- Raw data: 520M trips (3 years) -- 150 GB in 48 CSV files
    - 12 fields, 2 spatial-temporal attributes

A Study of NYC Taxis: Preparation

	SQLite	PostgreSQL
Storage Space in GB	100	200
Building Indices in Minutes (One Year	3,120	780
Seconds		

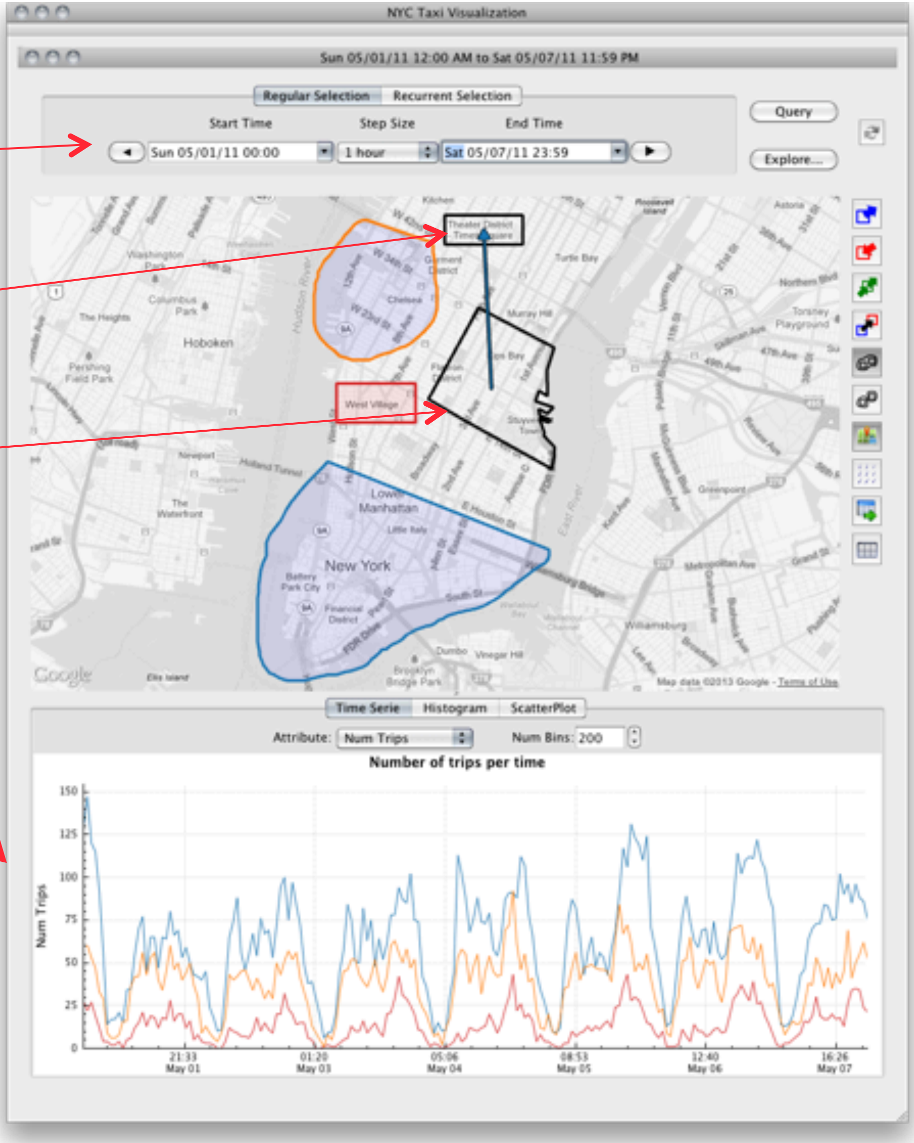
- Spatio-temporal index based on out-of-core kd-tree [Ferreira et al., TVCG 2013]
  - Deployed at TLC and DoT!
- New index that leverages GPU – 2 orders of magnitude speedup [Vo and Doraiswami, in progress]



Query	MongoDB (1 GPU) Time(sec)	MongoDB (3 GPUs) Time(sec)	PostgreSQL Time(sec)	PostgreSQL Speedup (1 GPU)	PostgreSQL Speedup (3 GPUs)	ComDB Time(sec)	ComDB Speedup (1 GPU)	ComDB Speedup (3 GPUs)
1	0.237	0.103	141.8	598	1376	136.9	578	1329
2	0.199	0.065	129.2	649	1987	119.6	601	1840
3	0.202	0.093	97.1	480	1044	39.4	195	423
4	0.183	0.069	103.7	566	1502	25.6	140	371
5	0.361	0.159	106.3	294	668	23.8	66	149
6	0.325	0.174	102.6	315	589	28.9	89	166

A Study of NYC Taxis: Analysis

```
SELECT   *
FROM     trips
WHERE pickup_time in (5/1/11,5/7/11)
                 AND
            dropoff_loc in “Times Square”
                 AND
            pickup_loc   in “Gramercy”
```



NYC Taxi Visualization

Sun 05/01/11 12:00 AM to Sat 05/07/11 11:59 PM

Start Time: Sun 05/01/11 00:00

Step Size: 1 hour

End Time: Sat 05/07/11 23:59

Attribute: Num Trips

Num Bins: 200

Interactively explore data through the map view and plot widgets

New, scalable, map rendering infrastructure

[Ferreira et al., IEEE TVCG 2013]

Food for Thought

- The expertise gap
  - Domain experts do not know what is possible
  - Techies do not understand the domain
  - π-shaped scientists [A. Szalay]

- Data scientists will solve this problem! – or not…
  - You need at least 3 experts to *make* data scientist: DB, ML/Stat, Vis

- Has computer science (and data management) research failed?

  Yes – we don't have a good track record of developing *usable tools*

- Or is this problem just too hard?

  Yes – the complexity is often underestimated

Thanks