CIT 1307: Introduction to Information Technology

Big Data Systems

Overview

- Big Data Computing
  - Applications of Big data
  - 3 V’s
  - Big Data Tools and Techniques
    - Database management systems
    - Data mining and machine learning
  - Big Data systems and platforms
    - MapReduce & Hadoop

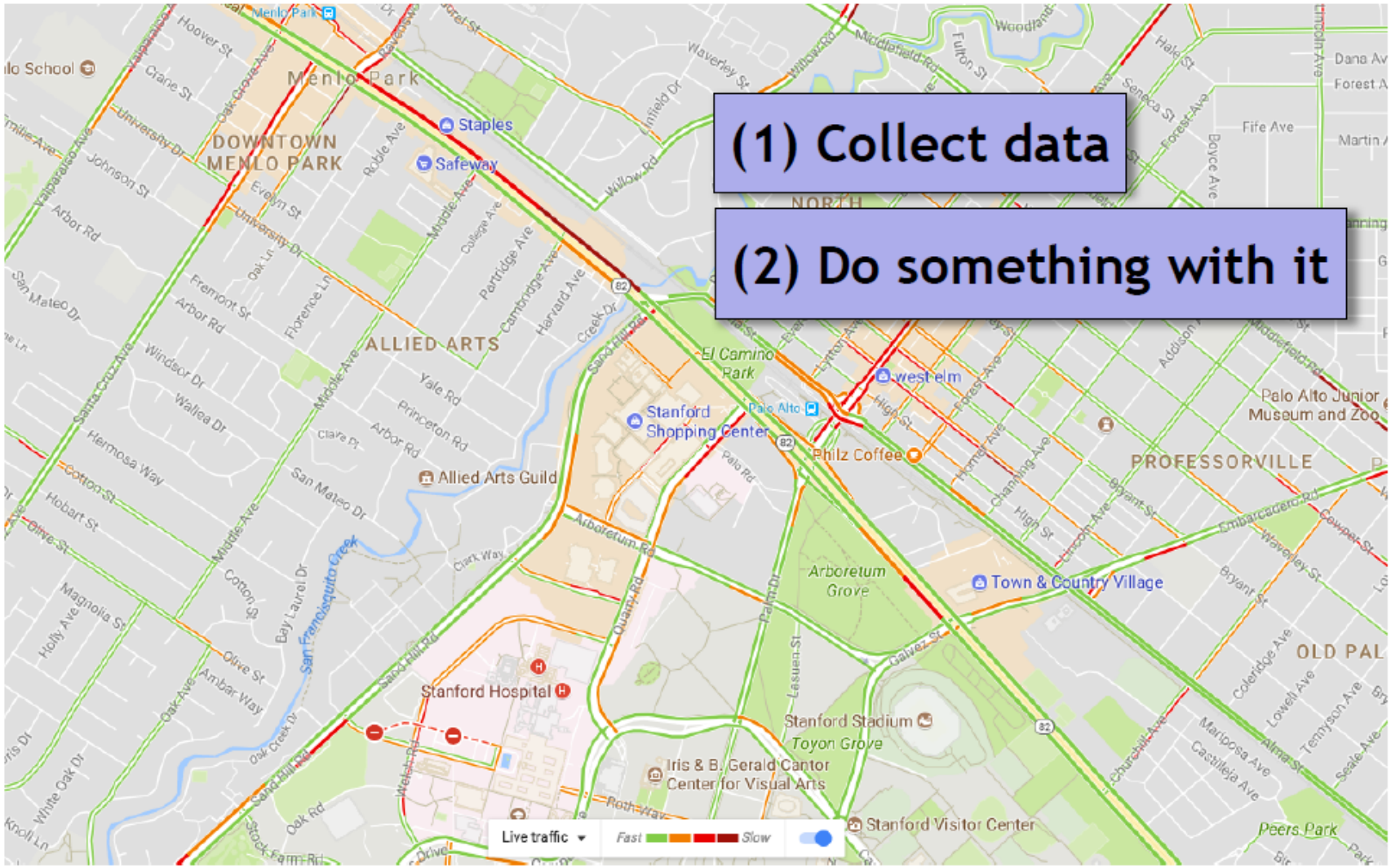
What Does "Big Data" mean?

1. Collecting large amounts of data - via computers, sensors, people, events …
2. Doing something useful with it - making decisions, confirming hypotheses, gaining insights, predicting future …

Big Data is Here to Stay

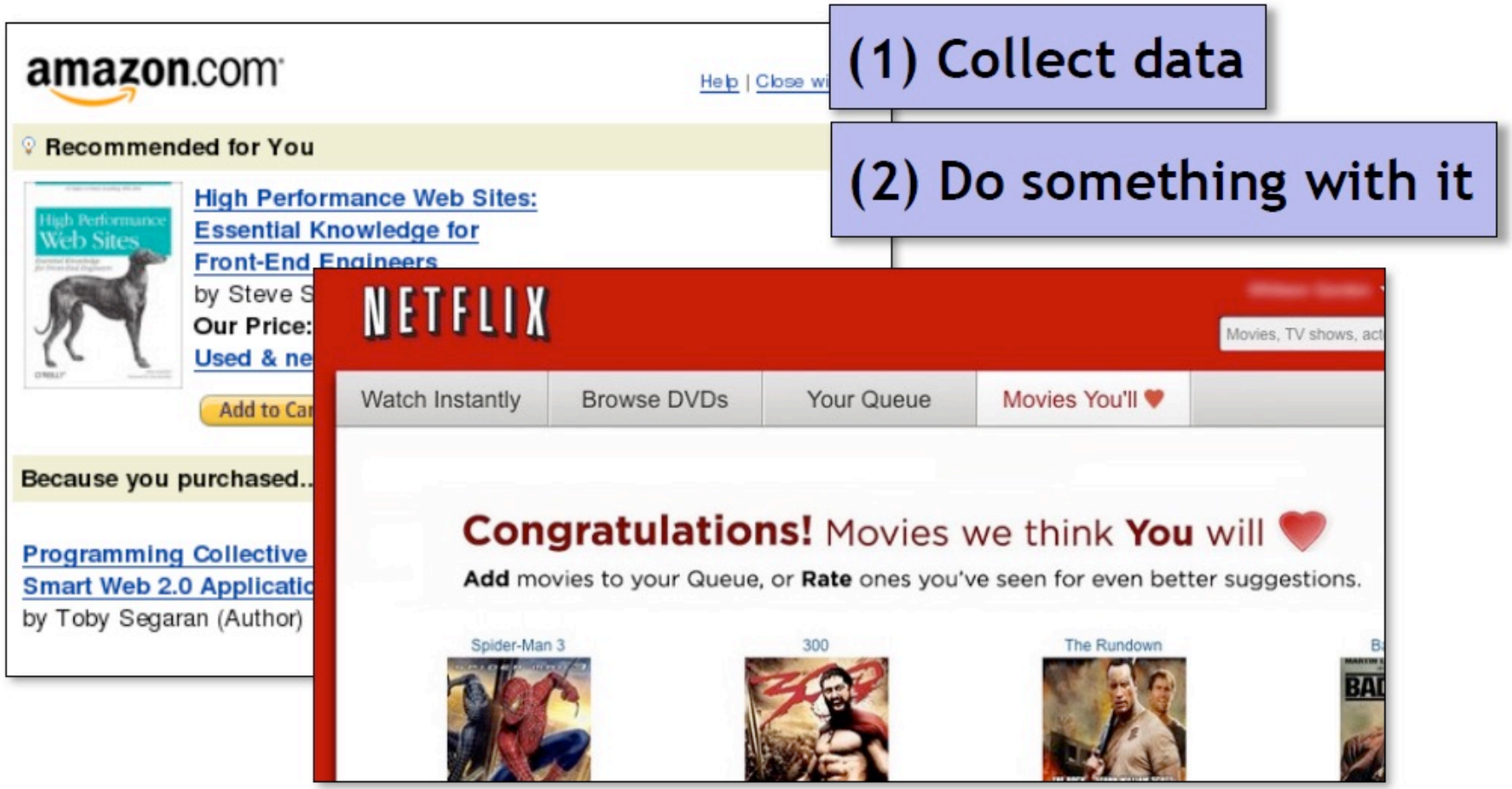
- Ability to collect data will only increase
- Ability to analyze data will only improve

Applications of "Big Data": Traffic



ACK: Jennifer Widom

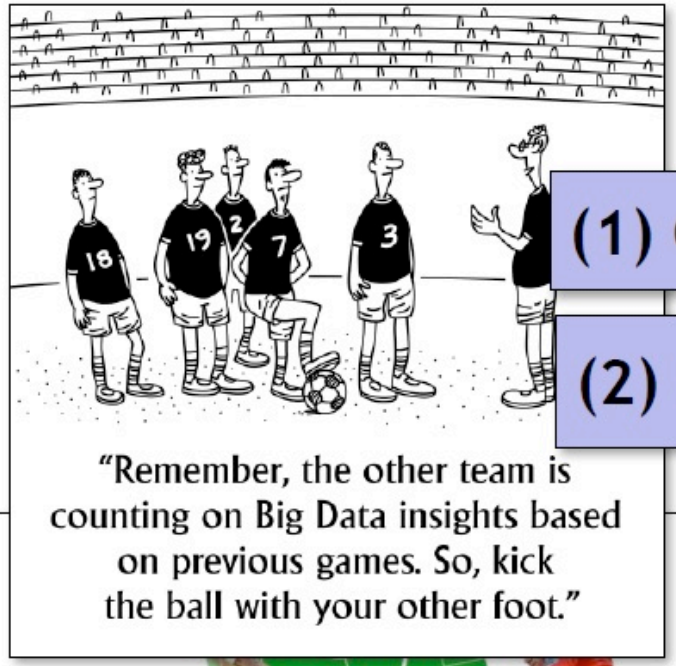
Applications of "Big Data": Recommender System



+ music, news, friends, romantic partners, and many more!

ACK: Jennifer Widom

Applications of "Big Data": Sports



(1) Collect data

(2) Do something with it

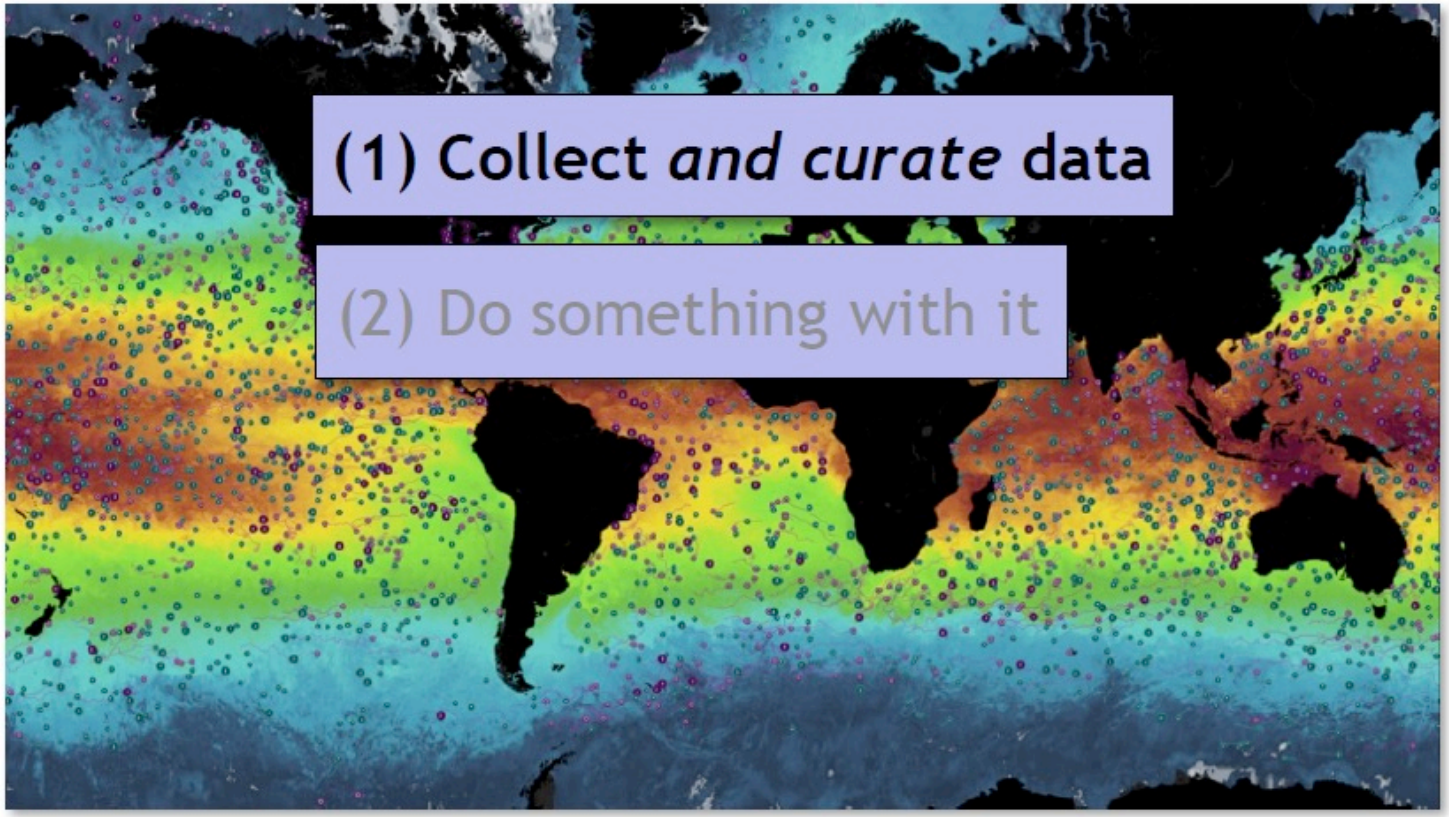
How big data gave the German football team a leg up

Saheli Roy Choudhury | @sahelirc
Thursday, 7 Jul 2016 | 12:39 AM ET
CNBC

How Big Data is Changing the World of Football

ACK: Jennifer Widom

Applications of "Big Data":Ocean Health



44,000 sensors, over 2 billion measurements

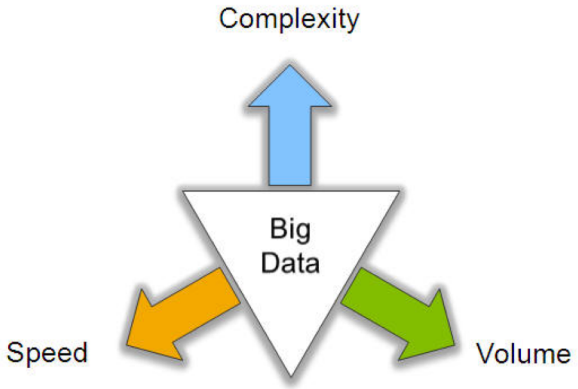
Physical, chemical, biological …

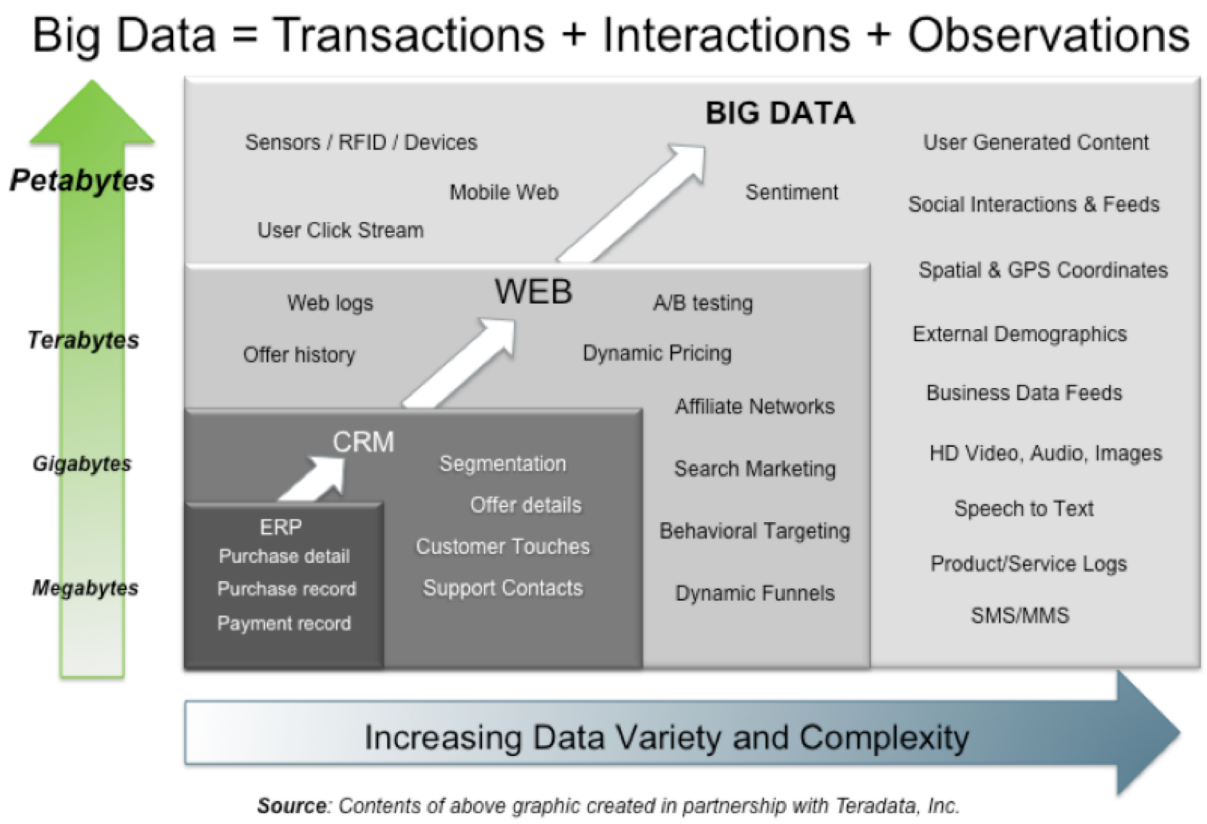
ACK: Jennifer Widom

And Many More

- Weather prediction
- Medical diagnosis
- Financial markets
- Resource management
- Computational social science
- Smart buildings and cities
- The list goes on and on and it’s still early days.

Big Data: 3V's

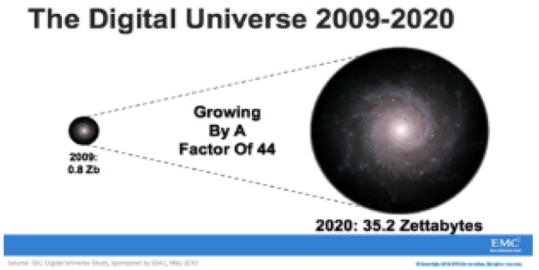


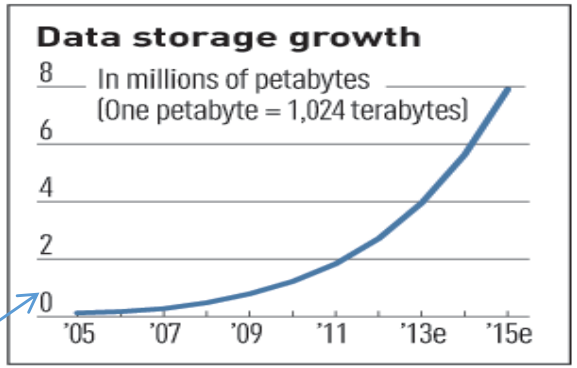


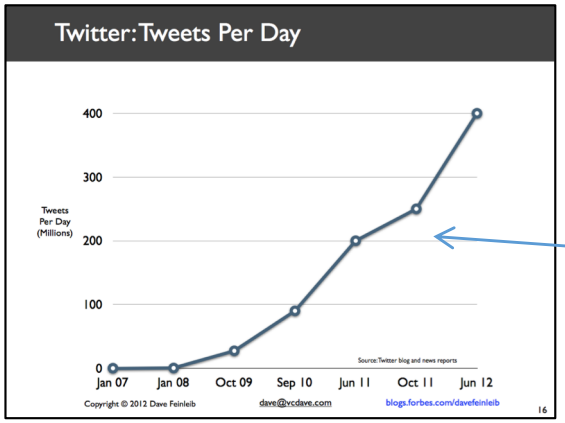
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 to 2020
  - From 0.8 zettabytes to 35ZB
- Data volume is increasing exponentially
- 90% of world data was created in last two years.



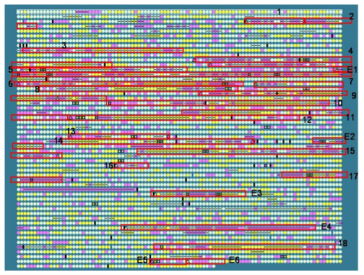


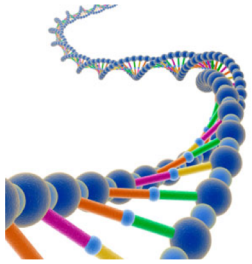


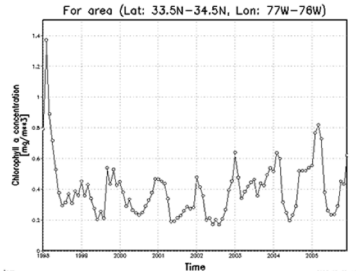
Exponential increase in collected/generated data

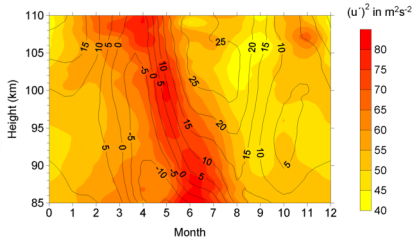
Variety (Complexity)

- **Structured Data**: Data that has a pre-set format, e.g., Address Books, product catalogs, banking transactions
- **Unstructured Data**: Data that has no pre-set format. Movies, Audio, text files, web pages, computer programs, social media
- **Semi-Structured Data**: Unstructured data that can be put into a structure by available format descriptions, i.g. XML.
- 80% of data is unstructured.
- Batch vs. Streaming Data
- **Real-Time Data**: Streaming data that needs to analyzed as it comes in. E.g., Intrusion detection. Aka "*Data in Motion*"
- **Data at Rest**: Non-real time. E.g., Sales analysis.
- **Metadata**: Definitions, mappings, scheme
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)









To extract knowledge➔ all these types of data need to be linked together

Velocity (Speed)

- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

Real-time/Fast Data

Social media and networks

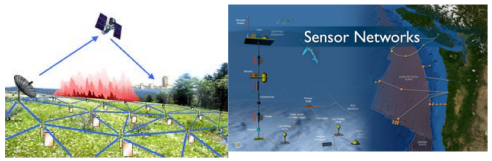
(all of us are generating data)

Scientific instruments

(collecting all sorts of data)

Mobile devices

(tracking all objects all the time)

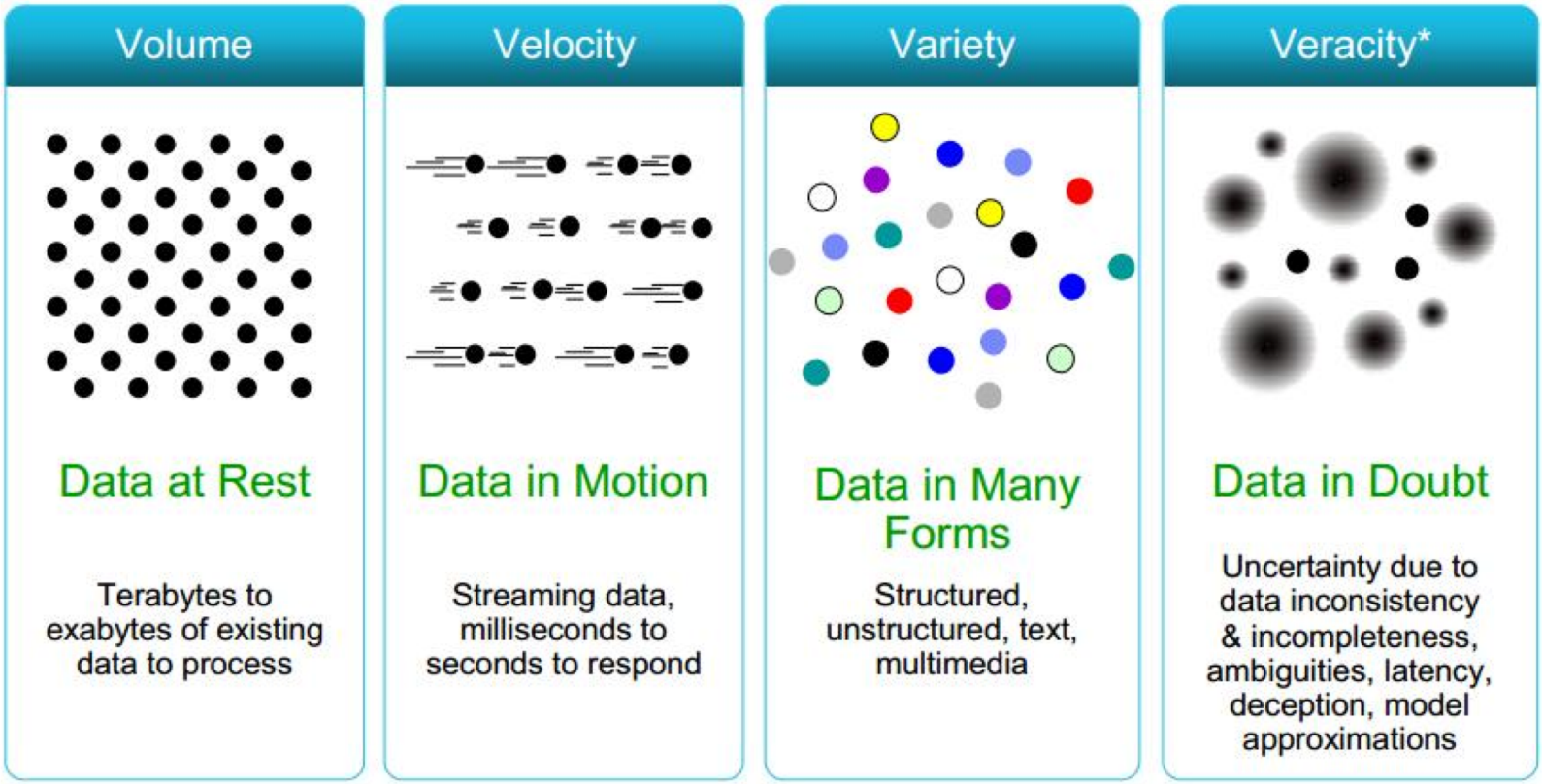


Sensor technology and networks

(measuring all kinds of data)

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Some Make it 4V's



Big Data Tools and Techniques

- Basic Data Manipulation and Analysis
  - Performing well-defined computations or
  - asking well-defined questions ("queries")
- Data Mining
  - Looking for patterns in data
- Machine Learning
  - Using data to make inferences or predictions
- Data Visualization
  - Graphical depiction of data
- Data Collection and Preparation

ACK: Jennifer Widom

Basic Data Manipulation and Analysis

Performing well-defined computations or asking well-defined questions ("queries")

- Average January low temperature for each country over last 20 years
- Number of items over $100 bought by females between ages 20 and 30
- Frequency of specific medicine relieving specific symptoms
- The ten stocks whose price varied the most over the past year

ACK: Jennifer Widom

Data Mining

Looking for patterns in data

- Items X,Y,Z are bought together frequently
- People who like movie X also like movie Y
- Patients who respond well to medicines X and Y also respond well to medicine Z
- Students going to the same university are frequently online friends
- Wealthier people are moving from cities to suburbs

ACK: Jennifer Widom

Machine Learning

Using data to make inferences or predictions

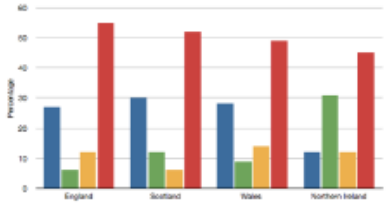
- Customers who are women over age 20 are likely to respond to an advertisement
- Students with good grades are predicted to do well on the SAT
- The temperature of a city can be estimated as the average of its nearby cities, unless some of the cities are on the coast or in the mountains.

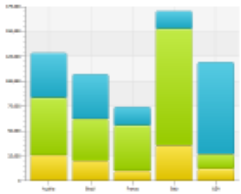
ACK: Jennifer Widom

Basic Data Visualizations

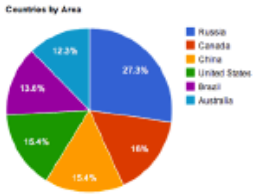
Don't underestimate the power of basic visualizations

- Bar charts

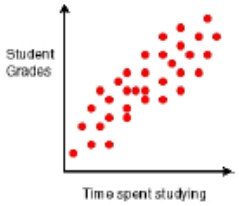


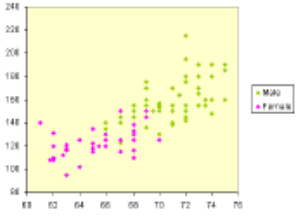


- Pie charts

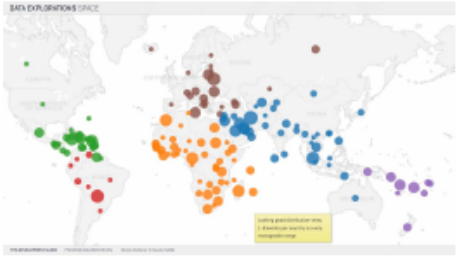


- Scatterplots



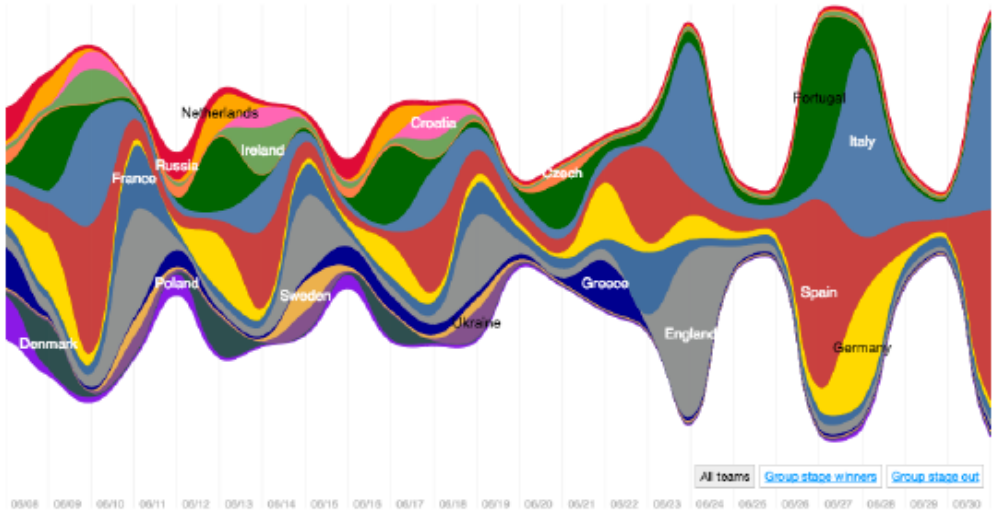


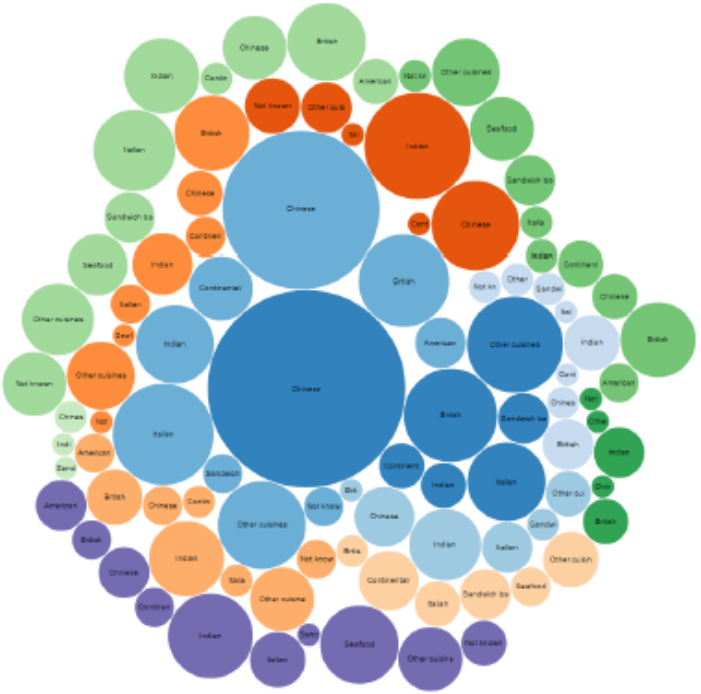
- Maps

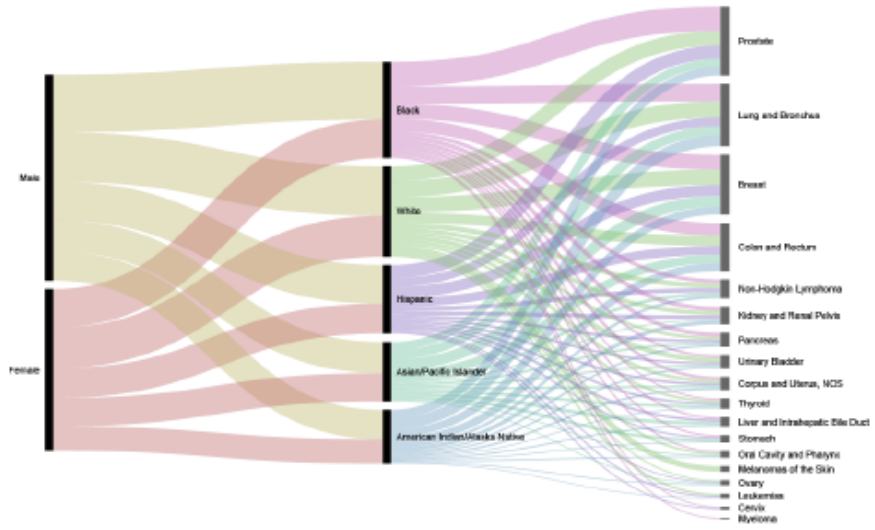


ACK: Jennifer Widom

Advanced Data Visualizations







ACK: Jennifer Widom

Big Data Systems and Platforms

- Spreadsheets
  - Surprisingly versatile and powerful for data analysis tasks, but not truly big data
- Programming languages with big-data support
  - R Language – powerful statistical features
  - Python – general-purpose language with R-like add-ons (Pandas, SciPy, scikit-learn)

Big Data Systems and Platforms

- Relational Database Management Systems
  - Also called RDBMS, SQL Systems
  - Long-standing solution for reliability, efficiency, powerful query processing
  - Works for all but truly extreme data sizes, or highly unstructured data
- "NoSQL" Systems
  - Distributed/scalable processing, unstructured data
  - Key-value row stores (e.g., Cassandra, Dynamo)
  - Document databases (e.g., MongoDB, CouchDB)
  - Graph databases (e.g., Neo4J, Giraph)

Big Data Systems and Platforms

- Specialized languages on scalable systems
  - MapReduce / Hadoop
  - Spark
- Systems for data preparation
- Systems for data visualization

Big Data Systems and Platforms: MapReduce

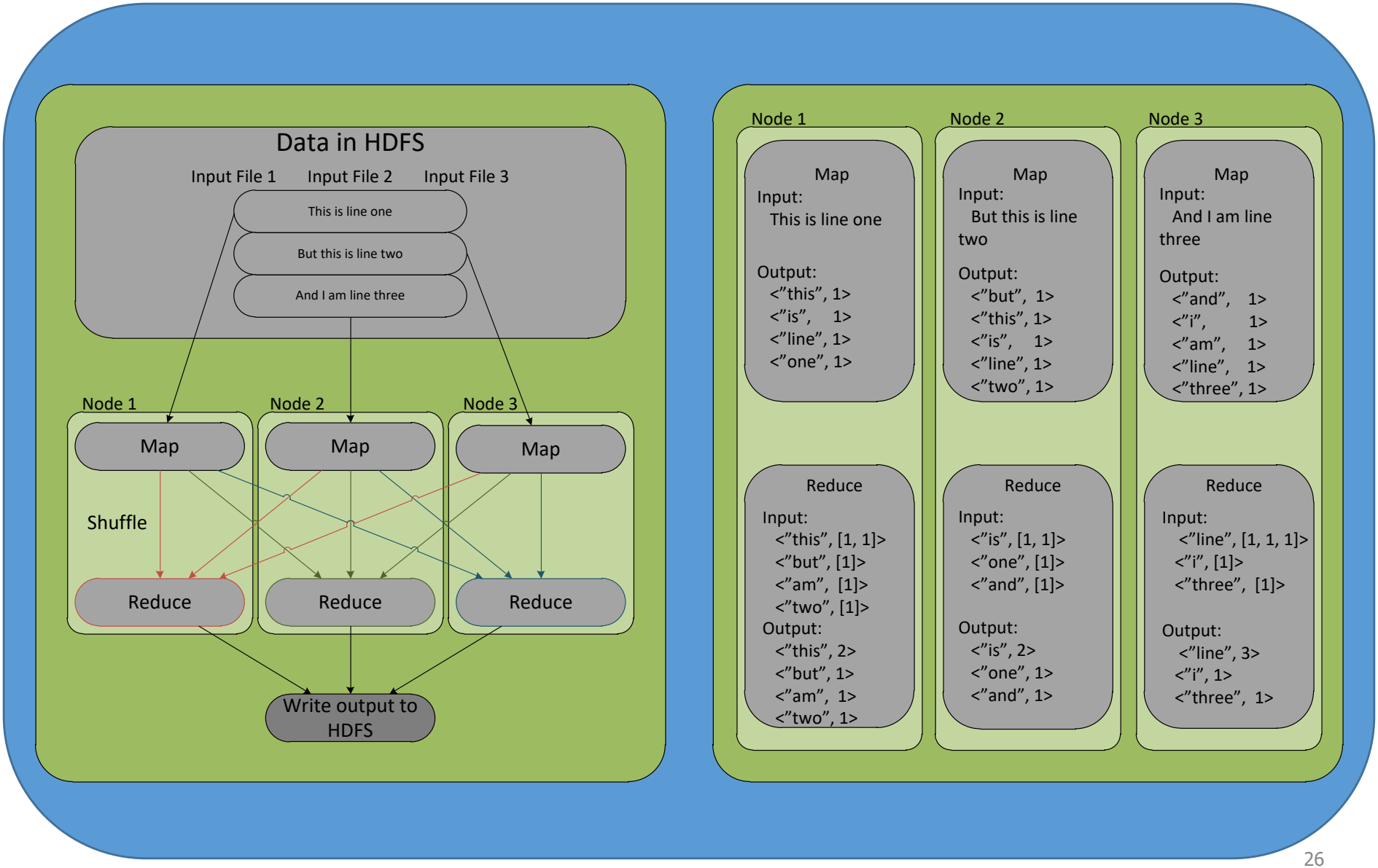
- Challenges in big data processing/Analytics
  - Extremely large data sets (petabytes of data)
  - Data is distributed across many computing nodes (potentially thousands)
  - Individual nodes are not aware of all data
  - Data is too large to be held or processed by any singular system

- How to distribute and process a task that works over petabytes of data?

Big Data Systems and Platforms: MapReduce

- MapReduce was introduced as a programming model by Google in 2004
- Spreads the task of processing data out over many computing resources
- Key-Value based system
  - Elements of divide and conquer
- Used for extremely parallel data processing/analysis
- Highly scalable
- Failure tolerant

MapReduce: Word Count



Big Data Systems and Platforms: Hadoop

- What is Hadoop?
  - Hadoop Distributed File System (HDFS)
    - The file system is dynamically distributed across multiple computers
    - Allows for nodes to be added or removed easily
    - Highly scalable
  - Hadoop Development Platform
    - Uses a MapReduce model for working with data
    - Users can program in Java, C++, and other languages

Big Data Systems and Platforms: Hadoop

- Some of the Key Characteristics of Hadoop:
  - On-demand Services
  - Rapid Elasticity
    - Need more capacity, just assign some more nodes
  - Scalable
    - Can add or remove nodes with little effort or reconfiguration
  - Resistant to Failure
    - Individual node failure does not disrupt the system
  - Uses off the shelf hardware
- **Apache Hadoop**: Open source Hadoop framework in Java. Consists of Hadoop Common Package (filesystem and OS abstractions), a MapReduce engine (MapReduce or YARN), and Hadoop Distributed File System (HDFS)

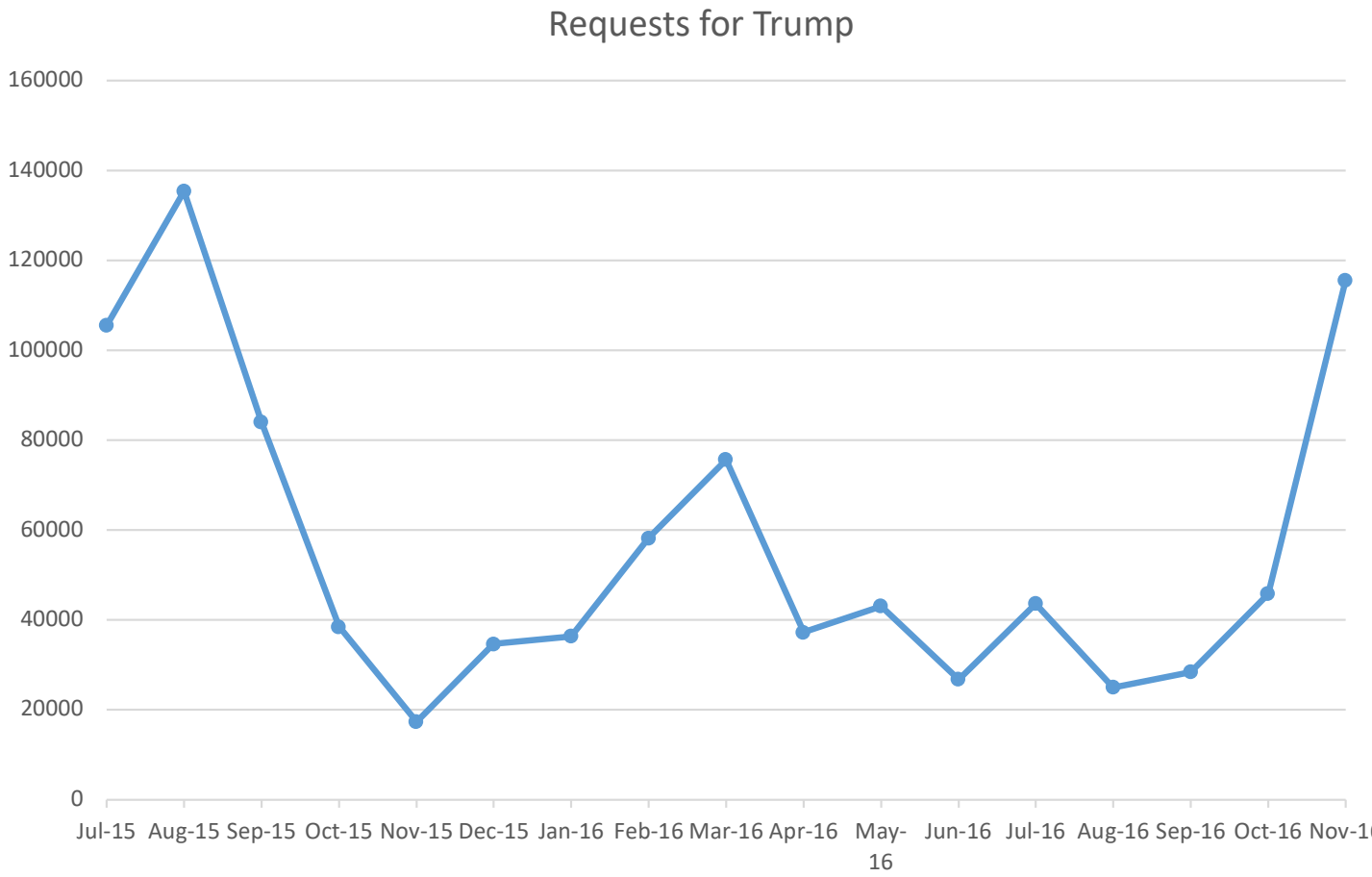
Big Data Systems and Platforms: Cloud

- Data processing in the cloud
  - Amazon Web Services, Google Cloud, Microsoft Azure
  - Data storage
  - Data processing: SQL, Hadoop, Spark
  - Machine learning libraries
  - Integration with visualization systems

Our Analysis on Wikipedia data

Page view Statistics available at

https://dumps.wikimedia.org/other/pagecounts-raw/

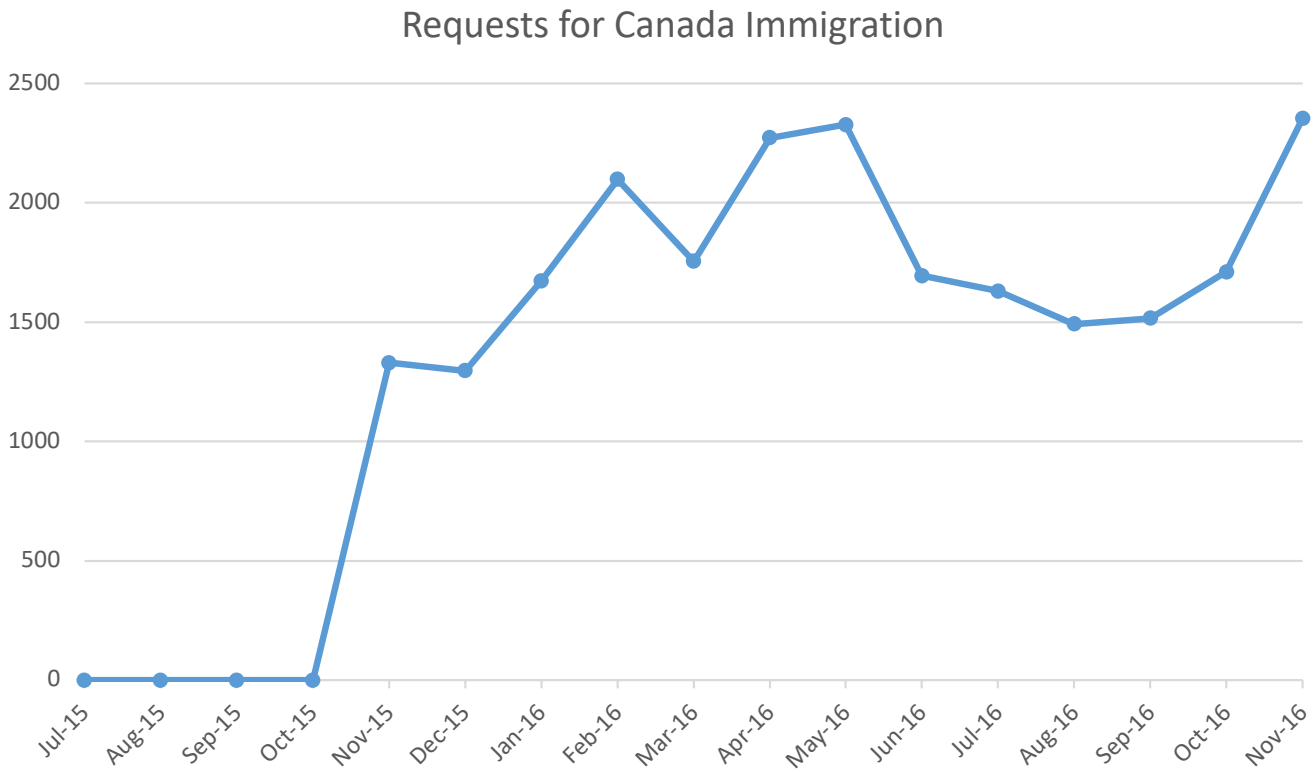


Month	Requests
Jul-15	105540
Aug-15	135381
Sep-15	84002
Oct-15	38464
Nov-15	17362
Dec-15	34680
Jan-16	36341
Feb-16	58233
Mar-16	75657
Apr-16	37245
May-16	43068
Jun-16	26760
Jul-16	43645
Aug-16	24998
Sep-16	28394
Oct-16	45815
Nov-16	115559

230 GB data/month

About 4TB data was analyzed.

Our Analysis on Wikipedia data



Month	Requests
Jul-15	0
Aug-15	0
Sep-15	0
Oct-15	0
Nov-15	1330
Dec-15	1296
Jan-16	1673
Feb-16	2098
Mar-16	1754
Apr-16	2272
May-16	2327
Jun-16	1694
Jul-16	1630
Aug-16	1492
Sep-16	1516
Oct-16	1711
Nov-16	2353