FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Big Data Analytics
Lecture 1

Yilu Zhou, PhD
Associate Professor
Gabelli School of Business
Fordham University

* Some slides are adopted from Professor Kunpeng Zhang's Big Data course @UMD

Yilu Zhou 周轶璐

FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Research Interest： Big Data, Data Science, Business Intelligence, Natural Language Processing, Machine Learning, Multilingual Knowledge Management.

2013 FORDHAM UNIVERSITY — Associate Professor, Information Systems

2006 GW — Assistant Professor of Information Systems

2001 THE UNIVERSITY OF ARIZONA — Artificial Intelligence Laboratory

2001 SHANGHAI JIAO TONG UNIVERSITY 1896 — Computer Science

About this class

- **This class** is covers

  (1) **introduction to the Big Data problem**

  current challenges, trends, and applications

  (2) **algorithms for Big Data analysis**

  mining and learning algorithms that have been developed specifically to deal with large datasets

  (3) **technologies for Big Data management**

  Big Data technology and tools, special consideration made to the Map-Reduce paradigm and the Hadoop ecosystem.

About this class

- This class is NOT
  - a machine learning or data mining course
  - a programming course (however, you need to know basic programming with Python and some basic commands in Linux

- Syllabus

Class Overview

- Introduction to Hadoop and MapReduce programming
- Hadoop overview
  - Framework / architecture
- Cloud computing (Amazon AWS)
- Data management
  - Hbase, hive, pig, sqoop, kafka

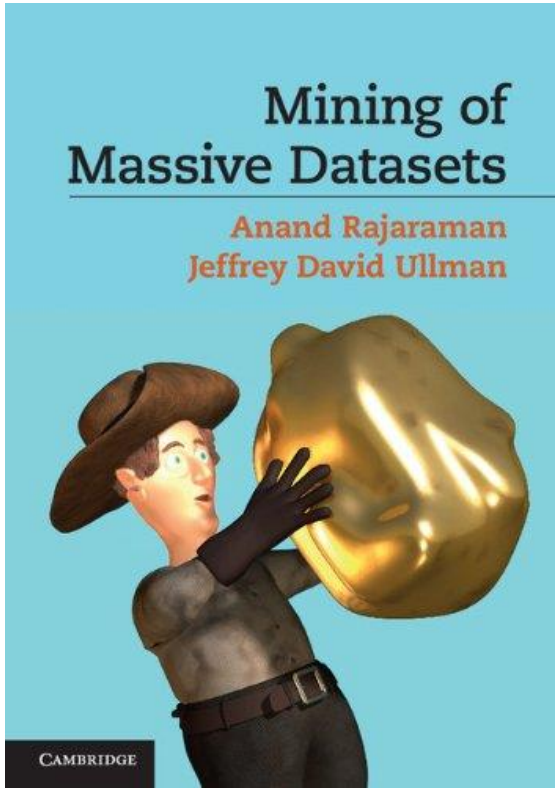
Class Overview –cont'd

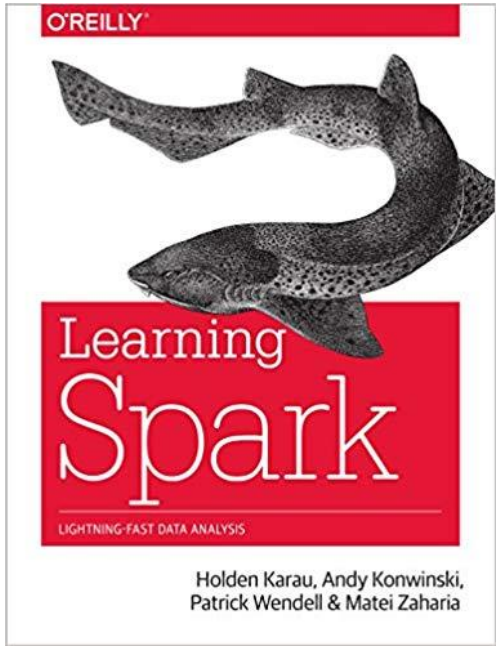
- **Data analytics**
  - Association rule mining
  - Clustering
  - Recommender system
  - Topic modeling
  - Social network analysis
  - **Spark**

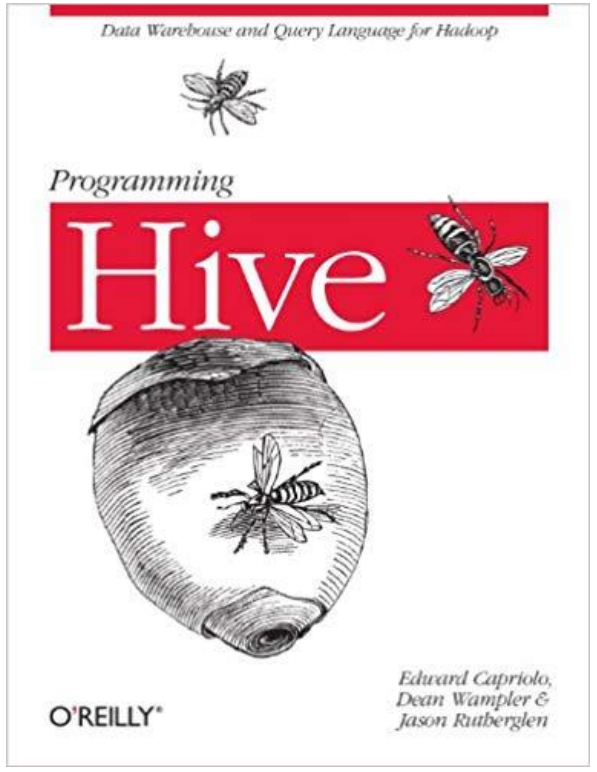
Algorithm level
(not just the terminology)

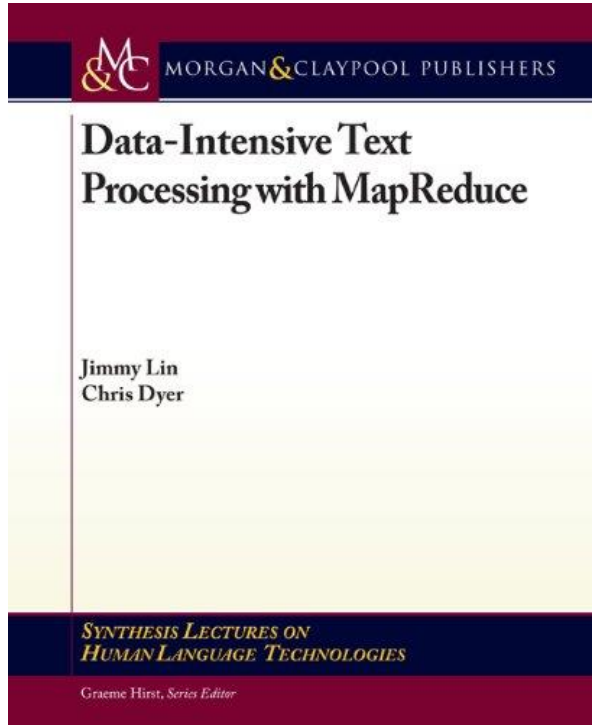
Textbooks

- Recommend you read the following books:

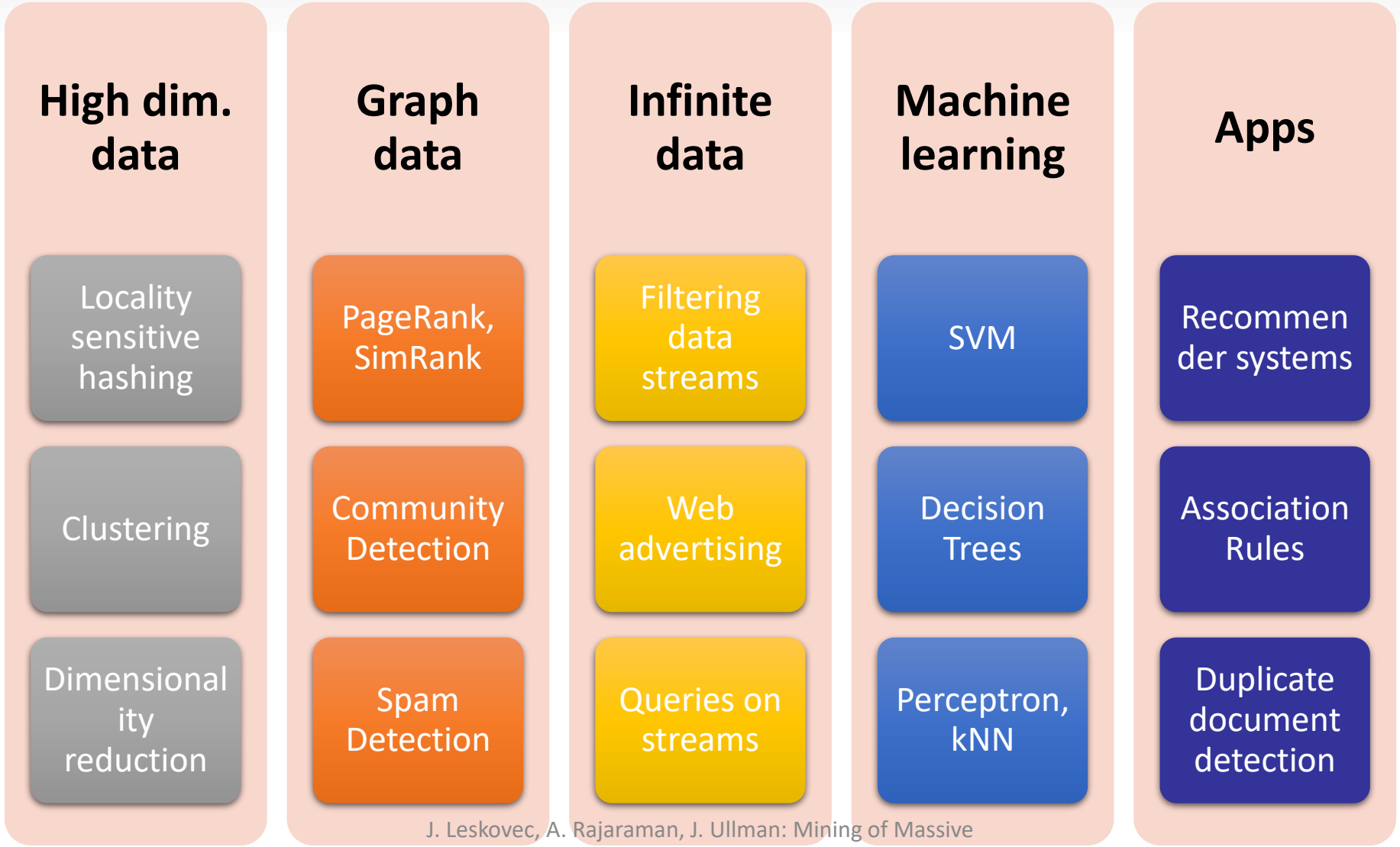








How It All Fits Together



Labs

- Configuration and installation of Hadoop
- Data management (HQL)
- Spark practice
- ML practices (multiple!)

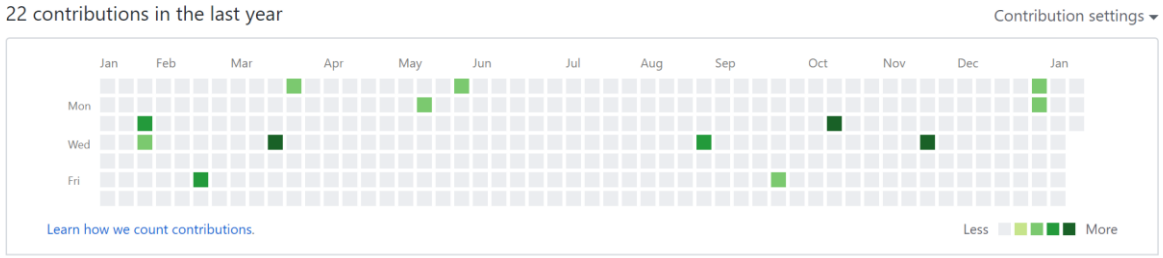
Assignments

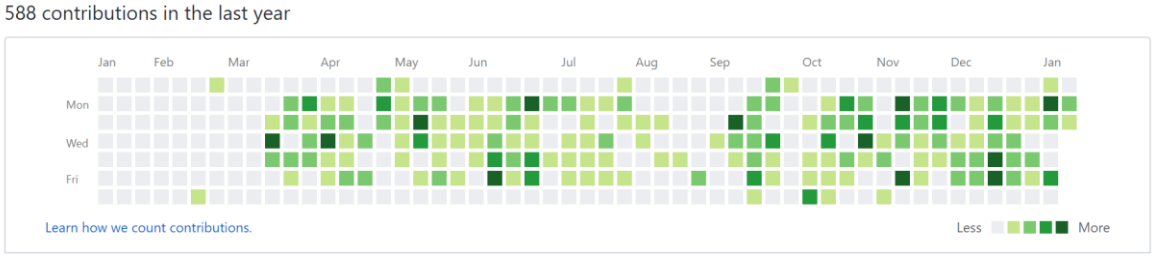
- 4 assignments (all of them require programming)
  - Basic MapReduce programming
  - Data management
  - Frequent Itemset Mining
  - K-Means
  - Recommendation System
  - Topic Models
  - Social network analysis

create github account

Two of these topics

Start to use GitHub





- Goal: to have at least 4 commits by the end of the semester
- For MSBA: use it as your extended resume

Project

1. **Proposal presentation.** Maximum 6 PPT pages of project proposal, dataset, business problem, data processing, models and (optional) expected outcomes.

2. **Final Presentation.** There will be 1 presentation during the last class in the semester, major results are expected during the presentation.

3. **Report.** Maximum 15-page report (single-space, 12-point font) highlighting consisting on the traditional sections of introduction, motivation, method, results, and conclusion.

5. **Confidential peer-evaluation form**. You will evaluate the contribution of each of your group members.

Big Data Overview

Why Big Data?

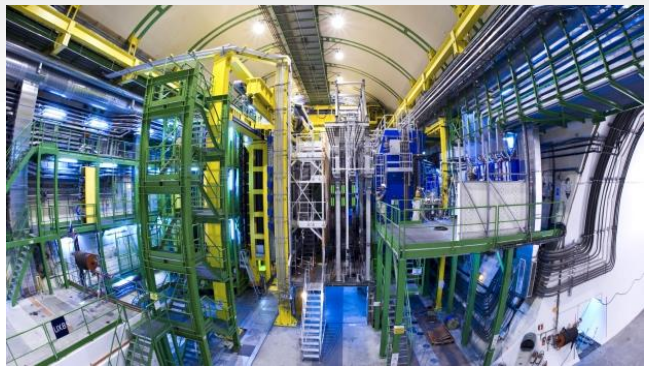
The U.S. could face a shortage by 2018 of 140,000 to 190,000 people with "deep analytical talent" and of 1.5 million people capable of analyzing data in ways that enable business decisions. (McKinsey & Co)

- Science
- Engineering
- Business
  - In 2012, the Obama administration announced the Big Data Research and Development Initiative 84 different big data programs spread across six departments.
  - Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data.
  - Facebook handles 40 billion photos from its user base.
  - Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide.
- Healthcare
- …

Big data is everywhere…

Google

processed about 24 petabytes of data per day in 2009.



By 2012, LHC collision data was being produced at approximately 25 petabytes per year.

The new at&t Your world. Delivered.

transfers about 30 petabytes of data through its networks each day.

As of January 2013, Facebook users had uploaded over 240 billion photos, with 350 million new photos every day.

facebook

Twitter now sends and receives as many as 200 million “tweets” every day.

S3: 449B objects, peak 290k request/second (7/2011)
1T objects (6/2012)

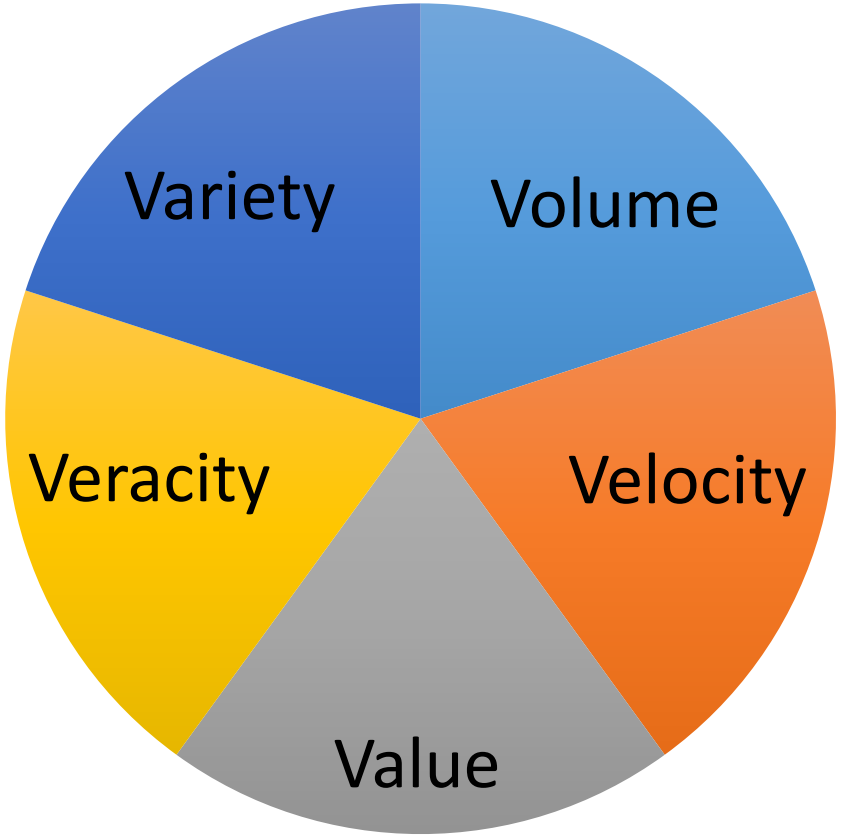
amazon

150 PB on 50k+ servers running 15k apps (6/2011)

CHASE

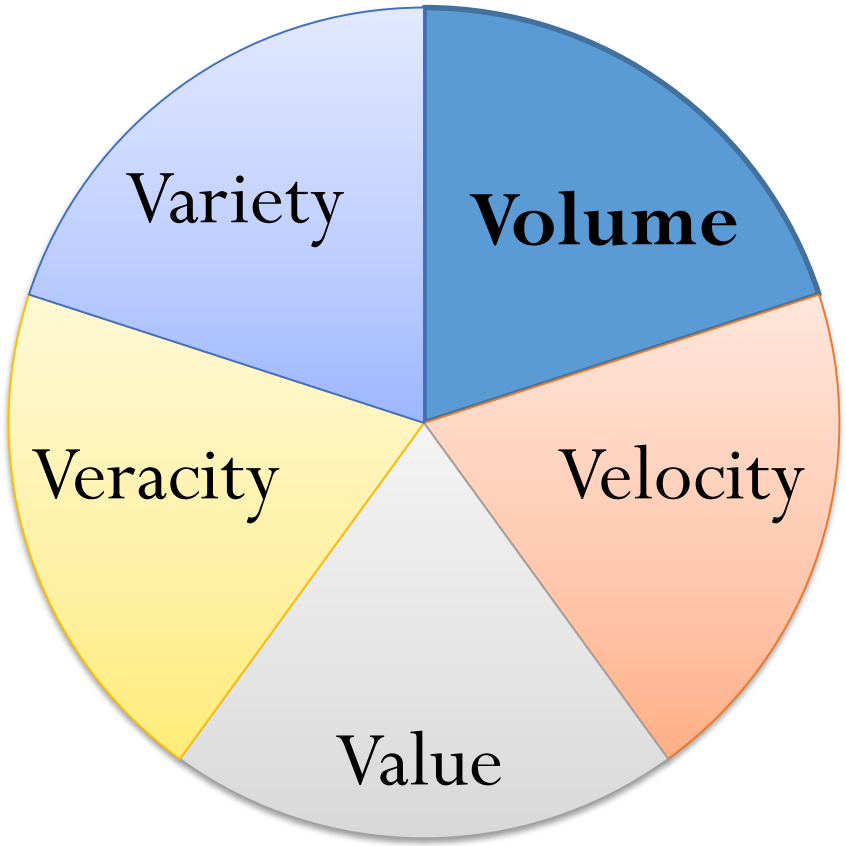
5 Vs of Big Data

To get better understanding of what big data is, it is often described using 5 Vs.



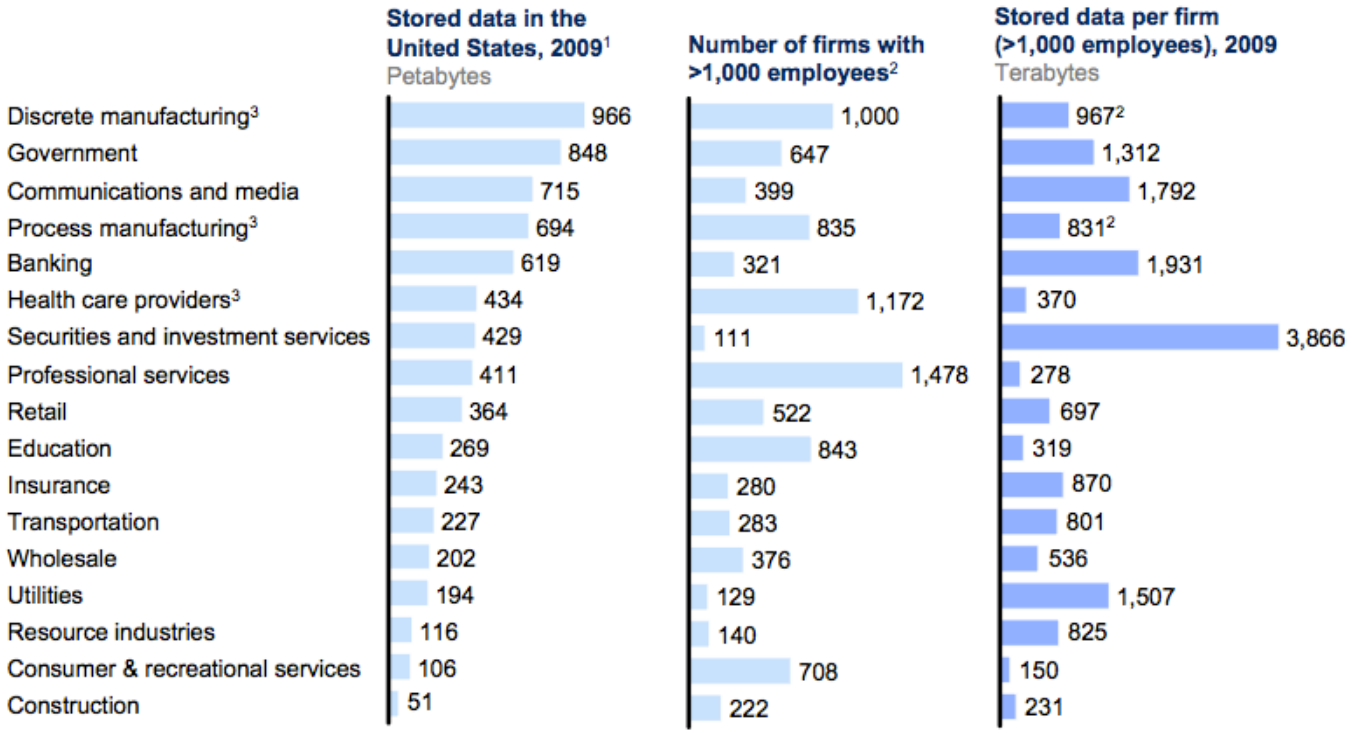
We see increasing volume of data, that grow at exponential rates

Volume refers to the vast amount of data generated every second. We are not talking about Terabytes but Zettabytes or Brontobytes. If we take all the data generated in the world between the beginning of time and 2008, the same amount of data will soon be generated every minute. This makes most data sets too large to store and analyze using traditional database technology. New big data tools use distributed systems so we can store and analyze data across databases that are dotted around everywhere in the world.



Big data is more prevalent than you think

Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte

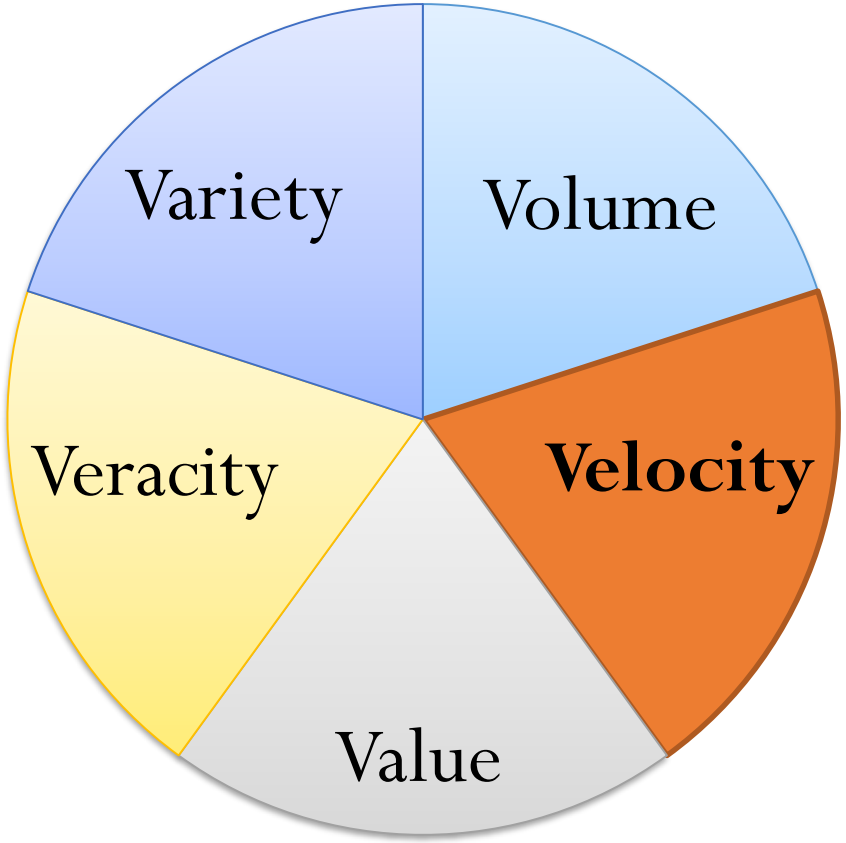


1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

SOURCE: IDC; US Bureau of Labor Statistics; McKinsey Global Institute analysis

We see increasing velocity (or speed) at which data changes, travels, or increases

Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology now allows us to analyze the data while it is being generated (sometimes referred to as it in-memory analytics), without ever putting into databases.

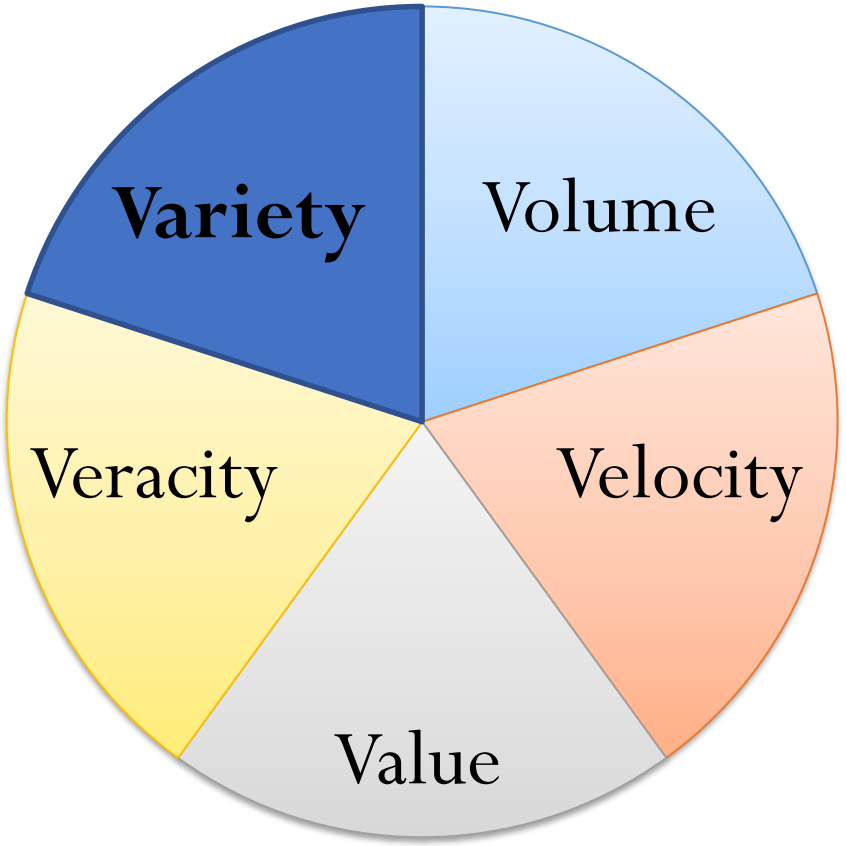


Stream Data and Real Time Data

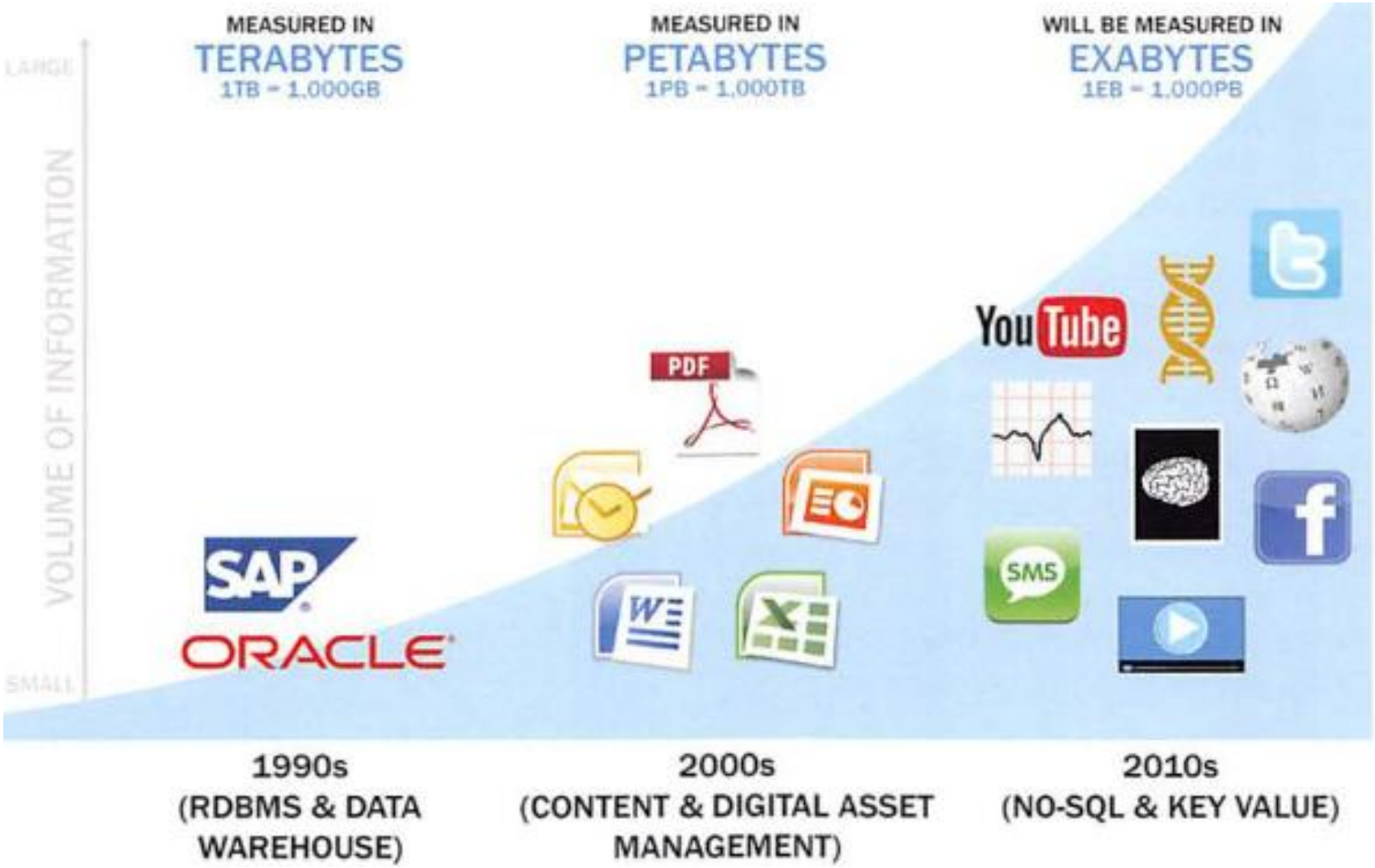
Big Data - Velocity Example

We see increasing variety of data types

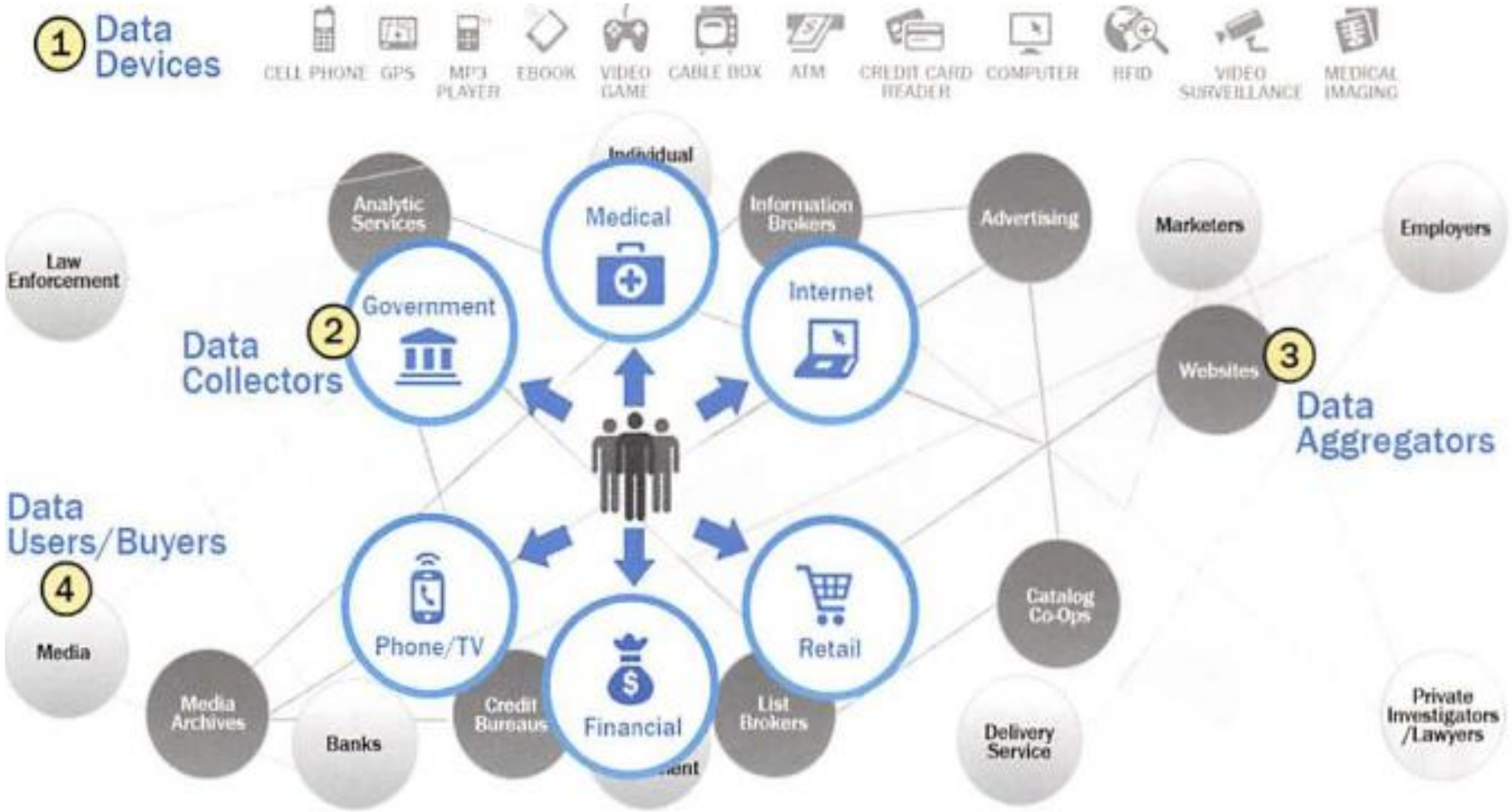
Variety refers to the different types of data we can now use. In the past we only focused on structured data that neatly fitted into tables or relational databases, such as financial data. In fact, 80% of world's data is unstructured (text, images, video, voice, etc.). With big data technology we can now analyze and bring together data of different types such as messages, social media conversations, photos, sensor data, video or voice recordings.



Data Evolution & Rise of Big Data Sources

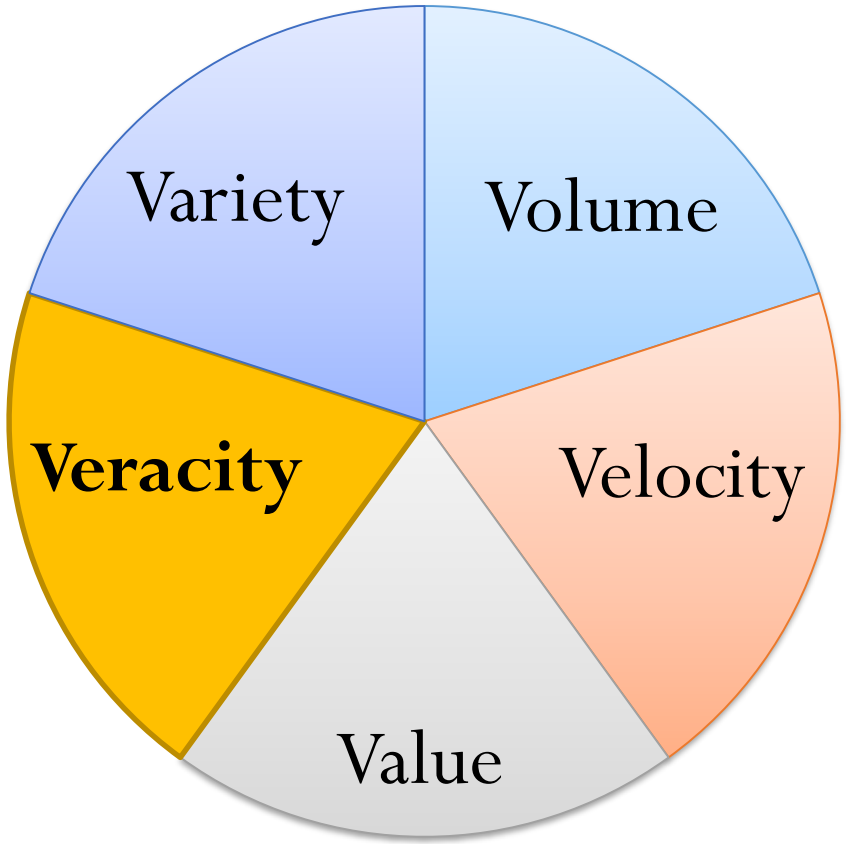


Emerging Big Data Ecosystem



We see increasing veracity (or accuracy) of data

Veracity refers to messiness or trustworthiness of data. With many forms of big data quality and accuracy are less controllable (just think Twitter posts with hash tags, abbreviations, typos and colloquial speech as well as the reliability and accuracy of content) but technology now allows us to work with this type of data.

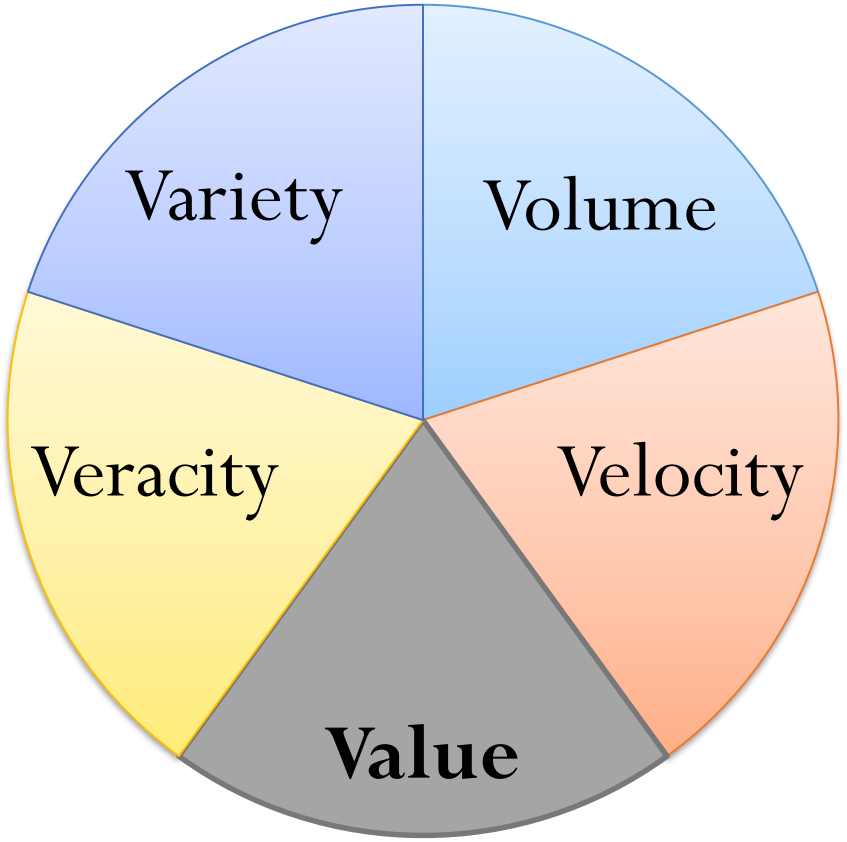


Value – the most important V of all

There is another V to take into account when looking at big data: **Value**.

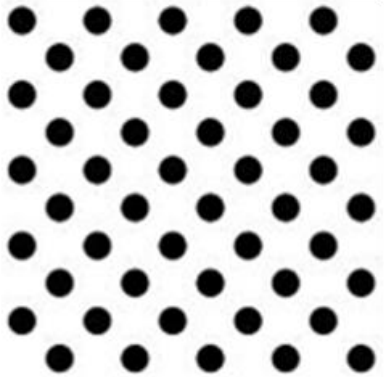
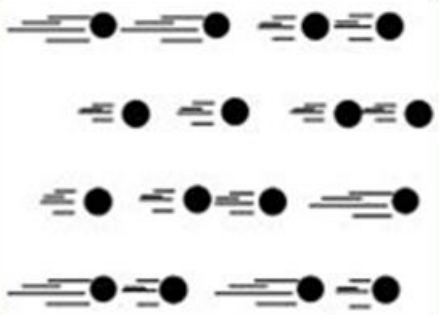
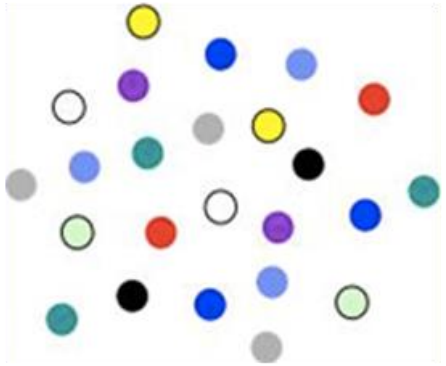
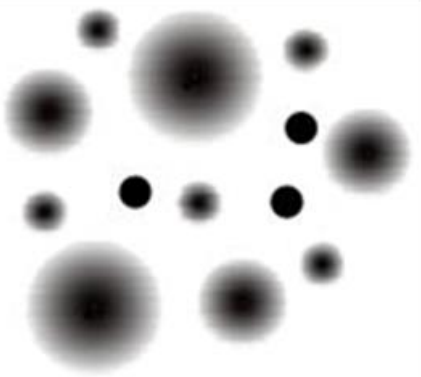
Having access to big data is no good unless we can turn it into value.

Companies are starting to generate amazing value from their big data.



Competitive advantages gained through big data

To Summarize

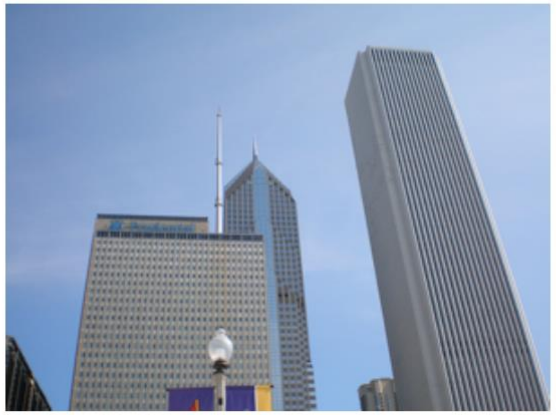
Volume	Velocity	Variety	Veracity	Value
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data into Money**
Terabytes to Exabytes of existing data to process	Streaming data, requiring milliseconds to seconds to respond	Structured, unstructured, text, multimedia,…	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations	Business models can be associated to the data

Adapted by a post of Michael Walker on 28 November 2012

Obama for America 2012

April 2011: Campaign formed in Chicago

Nov. 2012: 800 offices, 4k staff, 40k volunteers



~4000
OFA

~400
Chicago HQ

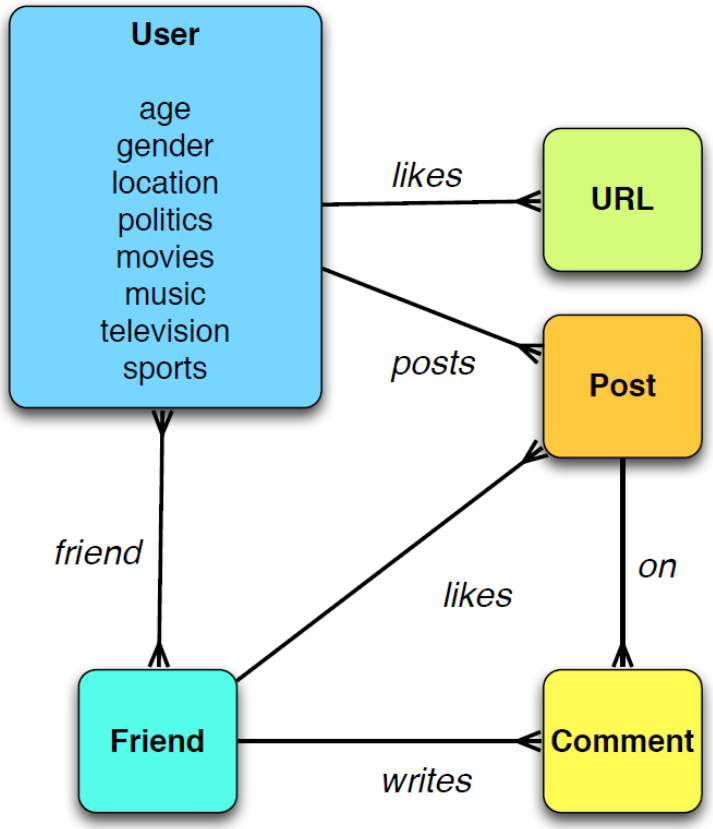
~40
Analytics

Source:

Matthew J. H. Rattigan
matt@edgeflip.com

Facebook

Facebook allows for much richer targeting, based on individual user preferences.

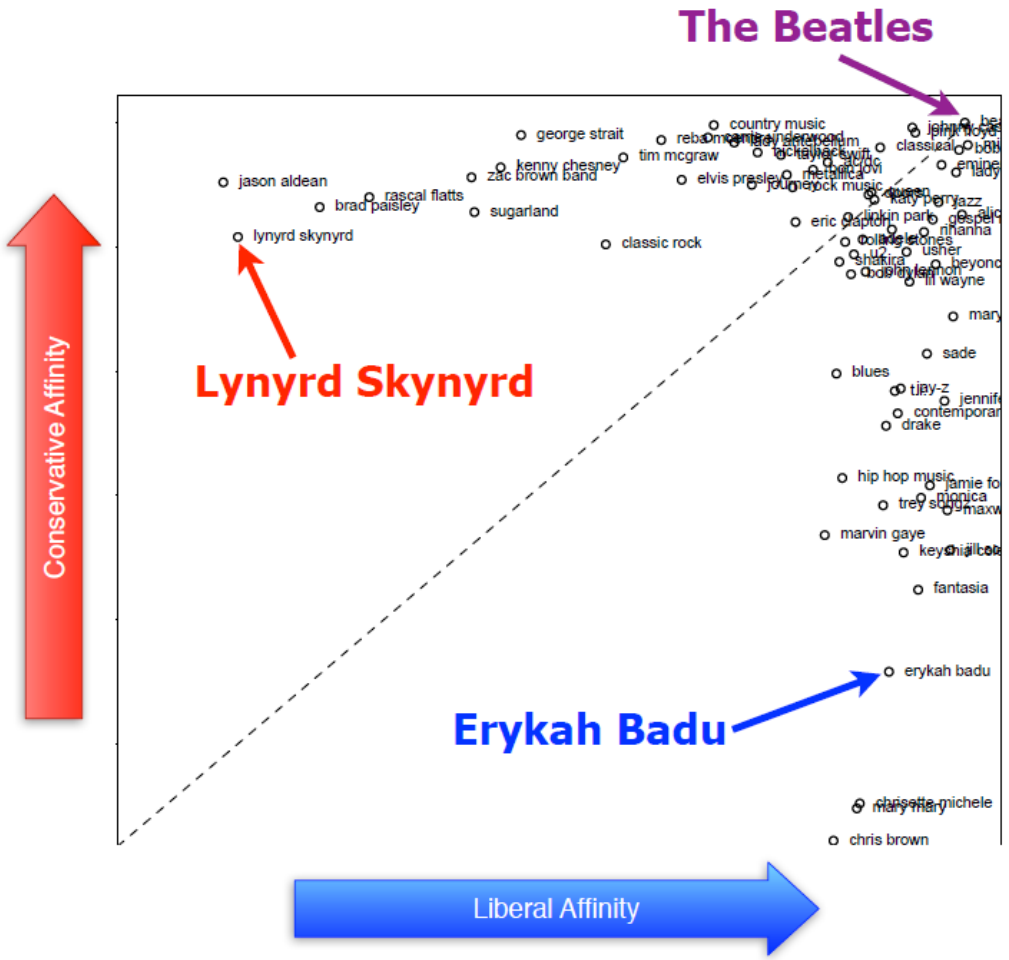


Source:

What you like tells me who you are

Matthew J. H. Rattigan
matt@edgeflip.com

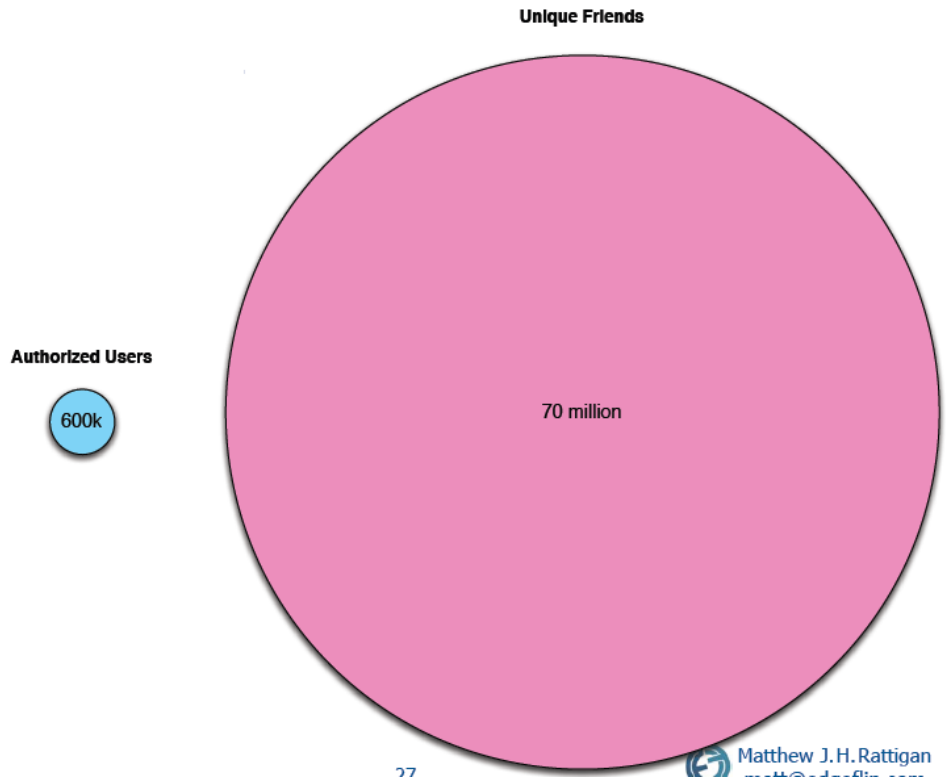
What music you like tells me who you are



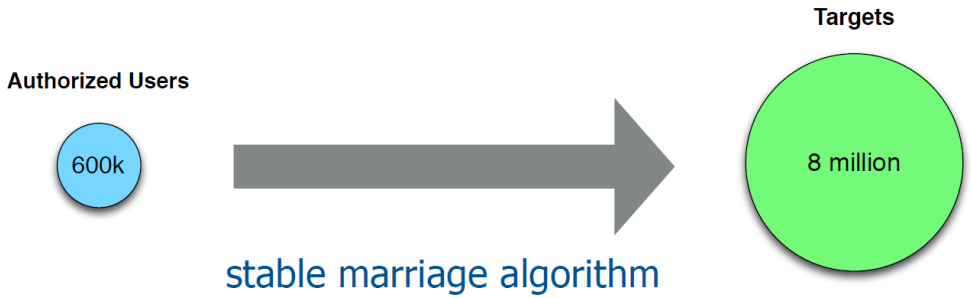
Source:

Matthew J. H. Rattigan
matt@edgeflip.com

Target sharing



Too many



Target

Source:

Matthew J.H.Rattigan
matt@edgeflip.com

Usage example of big data

- Predictive modeling
- mybarackobama.com
- Drive traffic to other campaign sites
  - Facebook page (33 million "likes")
  - YouTube channel (240,000 subscribers and 246 million page views).
- Every single night, the team ran 66,000 computer simulations.

- Data mining for individualized ad targeting
- Orca big-data app
- YouTube channel(23,700 subscribers and 26 million page views)

Applications, data, and corresponding commonly used analytical techniques

1. E-Commerce and marketing intelligence

Applications

- Recommender systems
- Social media monitoring and analysis
- Crowd-sourcing systems

Data

- Search and user logs
- Customer transaction records
- Customer generated content

Data characteristics

- Structured web-based, user-generated content, rich network information, unstructured informal customer opinions

Analytics

- Association rule mining
- Database segmentation and clustering
- Anomaly detection similar to classification but classification needs balance firs
- Graph mining
- Social network analysis
- Text and web analytics
- Sentiment and affect analysis

Impacts

- Long-tail marketing, targeted and personalized recommendation, increased sale and customer satisfaction

2. E-Government and Politics 2.0

Applications

- Ubiquitous government services
- Equal access and public services
- Citizen engagement and participation
- Political campaign and e-polling

Data

- Government information and services
- Rules and regulations
- Citizen feedback and comments

Data characteristics

- Fragmented information sources and legacy systems, rich textual content, unstructured informal citizen conversations

Analytics

- Information integration
- Content and text analytics
- Government information semantic services and ontologies
- Social media monitoring and analysis
- Social network analysis
- Sentiment and affect Analysis

Impacts

- Transforming governments, empowering citizens, improving transparency, participation, and equality

3. Science & Technology

Applications

- S&T innovation
- Hypothesis testing
- Knowledge discovery

Data

- S&T instruments and system generated data
- Sensor and network content

Data characteristics

- High-throughput instrument-based data collection, fine-grained multiple-modality and large-scale records, S&T specific data formats

Analytics

- S&T based domain-specific mathematical and analytical models

Impacts

- S&T advances, scientific impact

4. Smart Health and Wellbeing

Applications

- Human and plant genomics
- Healthcare decision support
- Patient community analysis

Data

- Genomics and sequence data
- Electronic medical records (EMR)
- Health and patient social media

app,patient like me

Data characteristics

- Disparate but highly linked content, person-specific content, and ethics issues

Analytics

- Genomics and sequence analysis and visualization
- EHR association mining and clustering
- Health social media monitoring and analysis
- Health text analytics
- Health ontologies
- Patient network analysis
- Adverse drug side-effect analysis
- Privacy-preserving data mining

Impacts

- Improved healthcare quality, improved long-term care, patient empowerment

5. Security and Public Safety

Applications

- Crime analysis
- Computational criminology
- Terrorism informatics
- Open-source intelligence
- Cyber security

Data

- Criminal records
- Crime maps
- Criminal networks
- News and web contents
- Terrorism incident databases
- Viruses, cyber attacks, and botnets

Data characteristics

- Personal identity information, incomplete and deceptive content, rich group and network information, multilingual content

Analytics

- Criminal association rule mining and clustering
- Criminal network analysis
- Spatial-temporal analysis and visualization
- Multilingual text analytics
- Sentiment and affect analysis
- Cyber attacks analysis and attribution

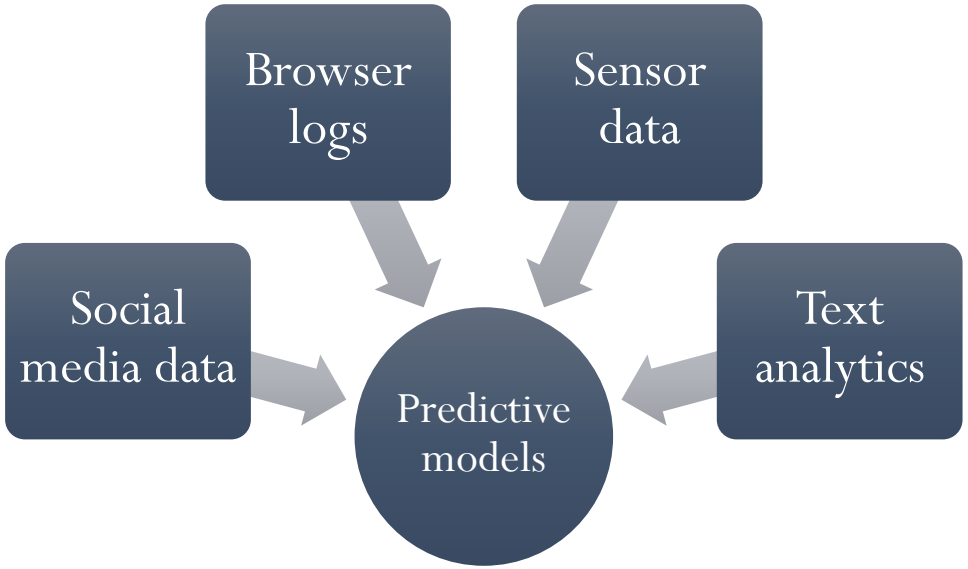
Impacts

- Improved public safety and security

Typical applications in big data

1. Understanding and targeting customers

- Big data is used to better understand customers and their behaviors and preferences.
  - Target: very accurately predict when one of their customers will expect a baby
  - Wal-Mart can predict what products will sell
  - Car insurance companies understand how well their customers actually drive
  - Obama use big data analytics to win 2012 presidential election campaign



2. Understanding and optimizing business processes

- Retailers are able to optimize their stock based on predictions generated from social media data, web search trends, and weather forecasts;
- Geographic positioning and radio frequency identification sensors are used to track goods or delivery vehicles and optimize routes by integrating live traffic data, etc.

3. Personal quantification and performance optimization

- The Jawbone armband collects data on our calorie consumption, activity levels, and our sleep patterns and analyze such volumes of data to bring entirely new insights that it can feed back to individual users;
- Most online dating sites apply big data tools and algorithms to find us the most appropriate matches.

4. Improving healthcare and public health

- Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit;
- Big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks;
- By recording and analyzing every heart beat and breathing pattern of every baby, infections can be predicted 24 hours before any physical symptoms appear.

5. Improving sports performance

- Use video analytics to track the performance of every player;
- Use sensor technology in sports equipment to allow us to get feedback on games;
- Use smart technology to track athletes outside of the sporting environment: nutrition, sleep, and social media conversation.

6. Improving science and research

- CERN, the Swiss nuclear physics lab with its Large Hadron Collider, the world's largest and most powerful particle accelerator is using thousands of computers distributed across 150 data centers worldwide to unlock the secrets of our universe by analyzing its 30 petabytes of data.

7. Optimizing machine and device performance

- Google self-driving car: the Toyota Prius is fitted with cameras, GPS, powerful computers and sensors to safely drive without the intervention of human beings;
- Big data tools are also used to optimize energy grids using data from smart meters.

8. Improving security and law enforcement

- National Security Agency (NSA) in the U.S. uses big data analytics to foil terrorist plots (and maybe spy on us);
- Police forces use big data tools to catch criminals and even predict criminal activity;
- Credit card companies use big data to detect fraudulent transactions.

fraud detect is important

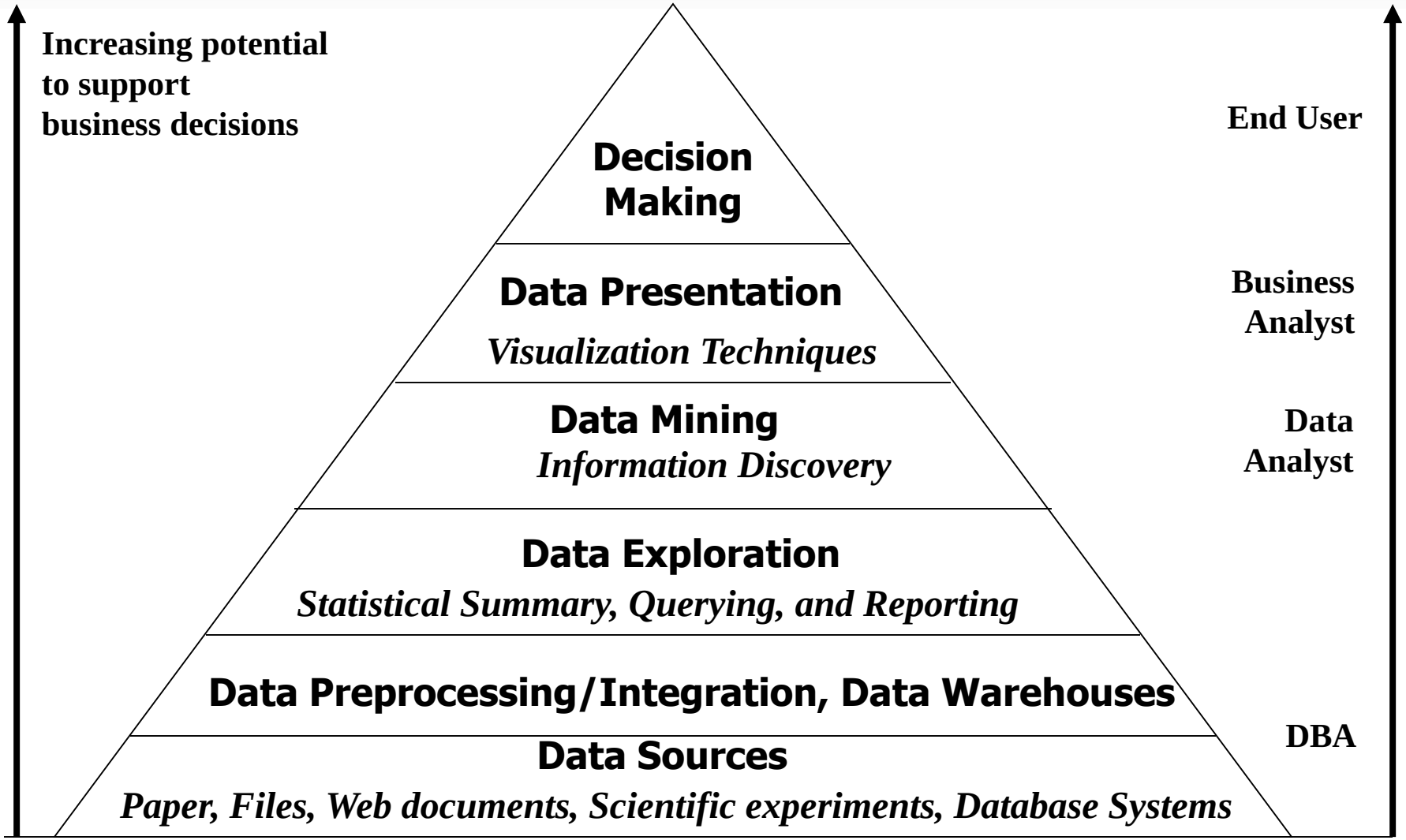
9. Improving and optimizing cities and countries

- Smart cities optimize traffic flows based on real time traffic information as well as social media and weather data.

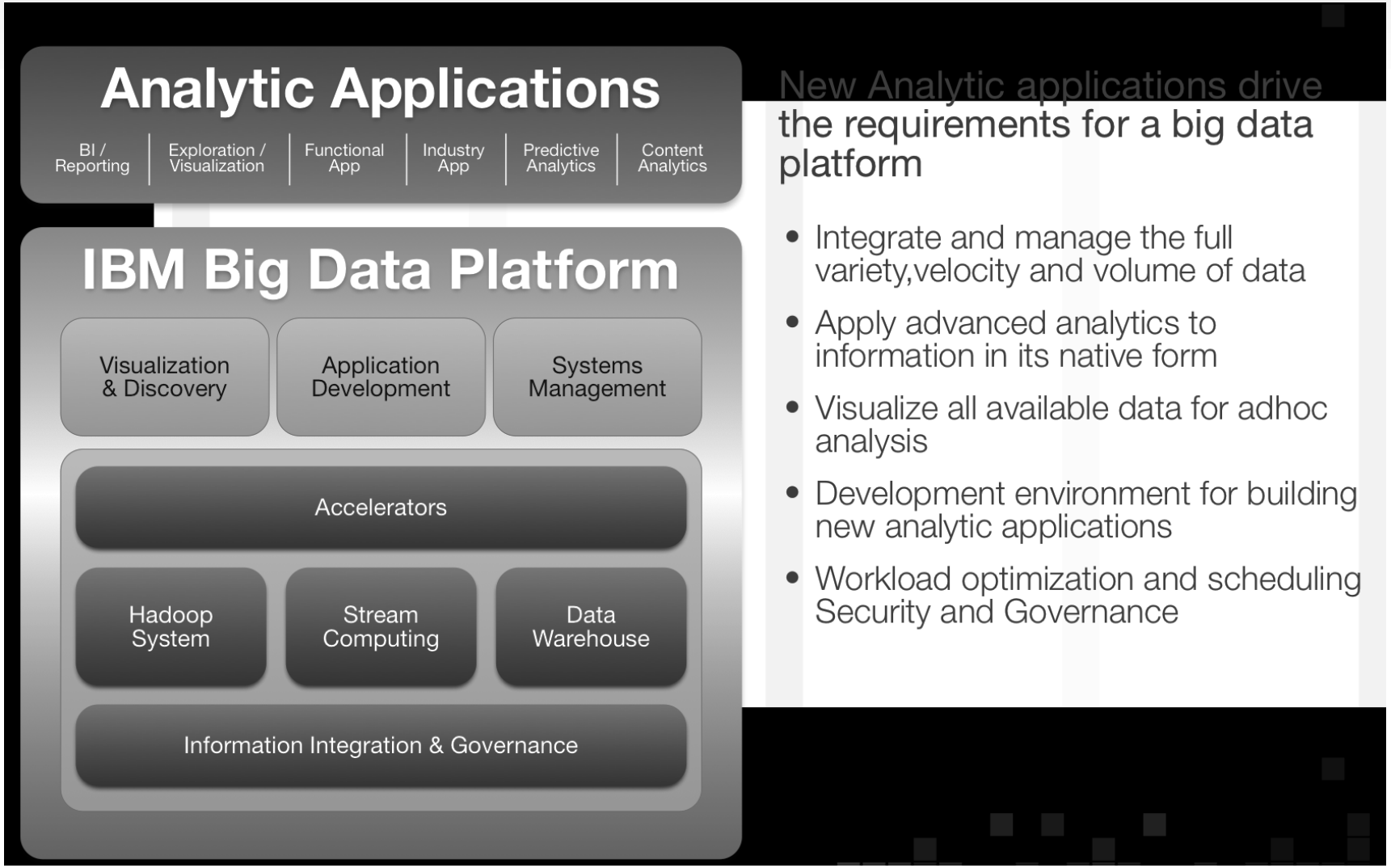
10. Financial trading

- The majority of equity trading now takes place via data algorithms that increasingly take into account signals from social media networks and news websites to make, buy and sell decisions in split seconds (High-Frequency Trading, HFT).

Data Mining in Business Intelligence



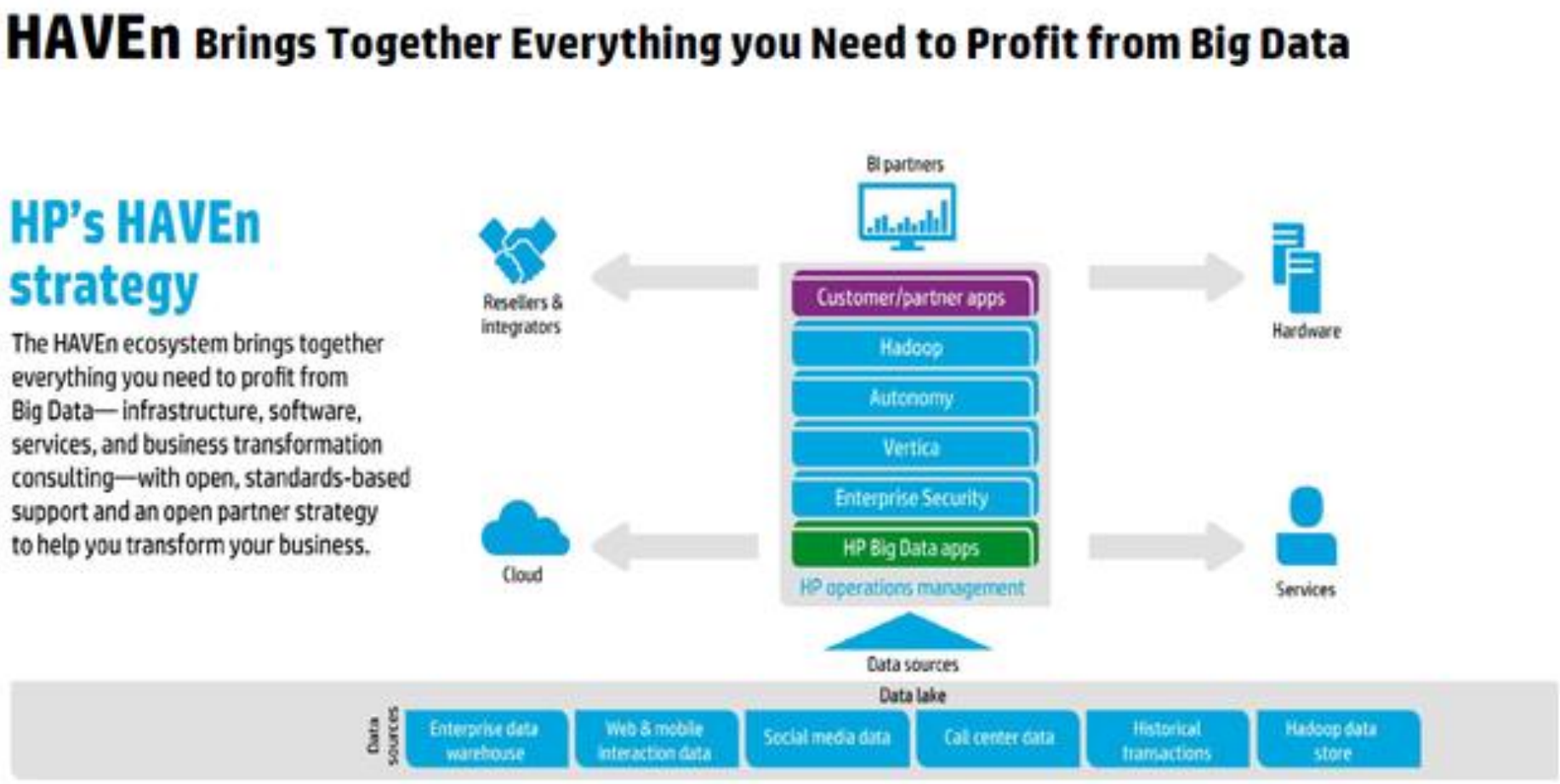
Big Data Platforms



Amazon EC2

- Elastic MapReduce
- DynamoDB

HP HAVEn


HAVEn Brings Together Everything you Need to Profit from Big Data
HP's HAVEn strategy
The HAVEn ecosystem brings together everything you need to profit from Big Data— infrastructure, software, services, and business transformation consulting—with open, standards-based support and an open partner strategy to help you transform your business.
BI partners
Resellers & integrators
Customer/partner apps
Hadoop
Autonomy
Vertica
Enterprise Security
HP Big Data apps
HP operations management
Hardware
Cloud
Services
Data sources
Data lake
Data sources
Enterprise data warehouse
Web & mobile interaction data
Social media data
Call center data
Historical transactions
Hadoop data store

Using Hadoop

- Java language
- High-level languages on top of Hadoop
  - Hive (Facebook)
    - A data warehouse system for Hadoop that facilitates easy data summarization, ad-hoc queries, and the analysis of large datasets stored in Hadoop compatible file systems
    - Provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL
    - It also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL

- Pig (Yahoo)
  - A platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs
- Jaql (IBM)
  - Primarily a query language for JavaScript Object Notation (JSON), but supports more than just JSON. It allows you to process both structured and nontraditional data

Big data analysis pipelines

Phase #1

- Data acquisition and recording
  - Filters: not discard useful data and not store irrelevant data
  - Metadata: describe what data is recorded and how it is recorded and measured
  - Data provenance: data quality

Phase #2

- Information extraction and cleaning
  - Raw data in different formats
  - Inaccurate data due to many reasons

Phase #3

- Data integration, aggregation, and representation
  - Database techniques: NoSQL DB

Phase #4

- Query processing, data modeling, and analysis
  - Data mining techniques
  - Statistical modeling
  - Query, indexing, searching techniques

Phase #5

- Interpretation
  - Report
  - Visualization

Review of Data Mining Concepts

Terminologies

- Dataset
  - Training set
  - Testing set
  - Validating set
- Data representation
  - Feature vector
- TF/IDF

Document 1

Term	Term Count
this	1
is	1
a	2
sample	1

Document 2

Term	Term Count
this	1
is	1
another	2
example	3

$$\text{idf}(\text{this}, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

$$\text{tf}(\text{example}, d_2) = 3$$

$$\text{idf}(\text{example}, D) = \log \frac{2}{1} \approx 0.3010$$

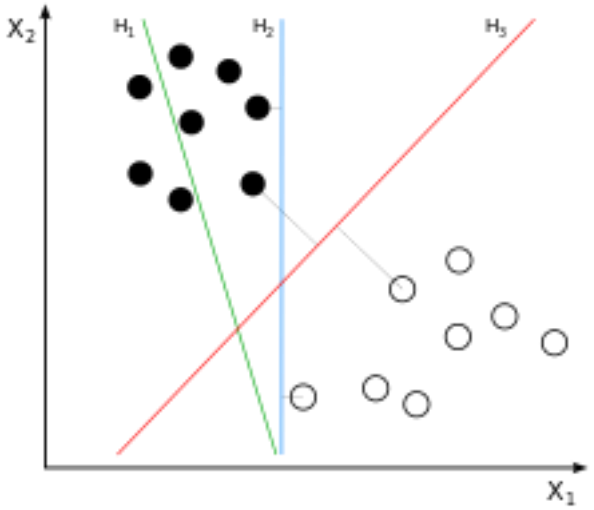
$$\text{tfidf}(\text{example}, d_2) = \text{tf}(\text{example}, d_2) \times \text{idf}(\text{example}, D) = 3 \times 0.3010 \approx 0.9030$$

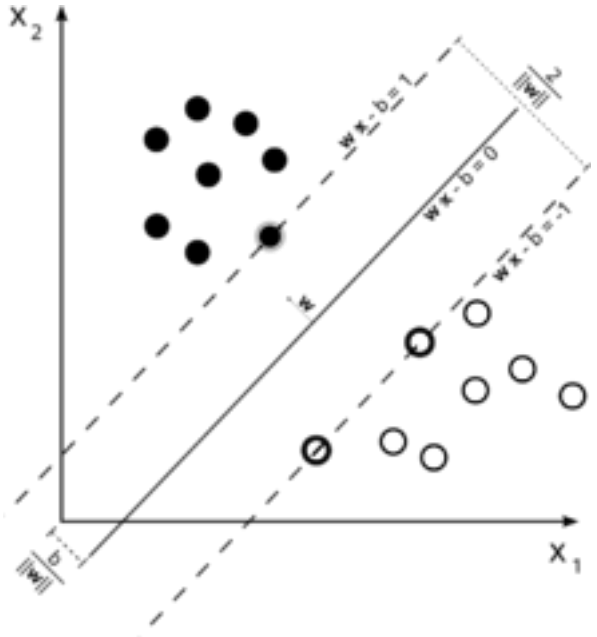
Supervised Learning

- Regression
  - Linear regression
  - Logistic regression
- Naïve Bayes
  - Strong independence assumption
- K-nearest neighboring (KNN)
- Decision Tree
  - C4.5
  - Can handle both numerical and categorical features
  - Missing values

Support Vector Machine

- Find a hyper-plane to maximize the functional margin.





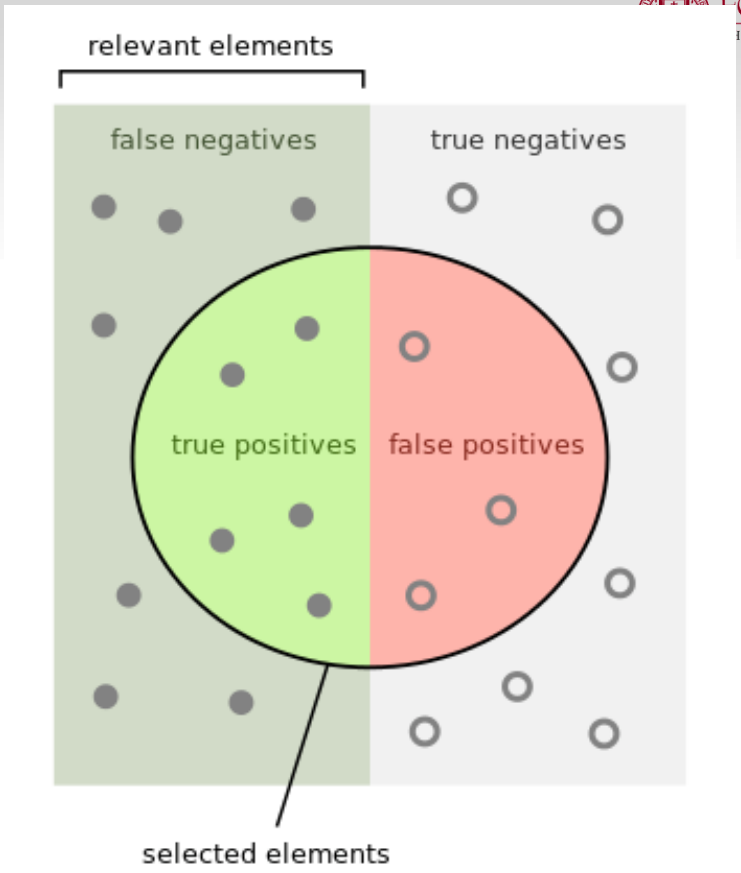
- Evaluation
  - Accuracy
  - Precision-recall

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

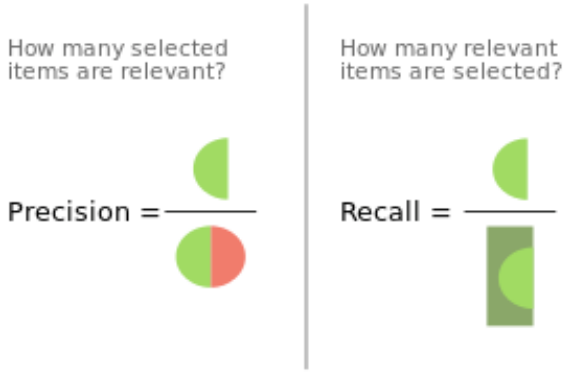
$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

  - F1 score

$$F = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

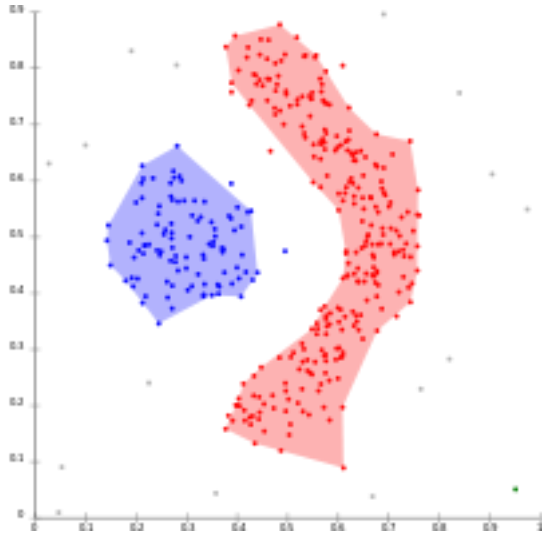


- Over fitting
  - Cross-validation
  - Regularization: L1/L2-norm
  - Early stopping
  - Pruning

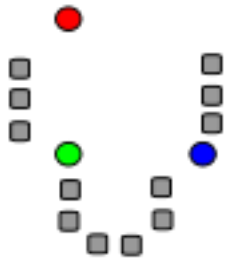


Unsupervised Learning

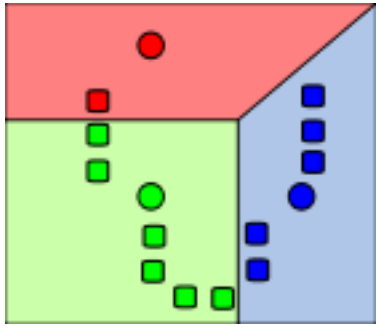
- Clustering
  - K-means
  - Spectral clustering
  - Hierarchical clustering
  - Density-based clustering (DBSCAN)
- Distance metric
  - Euclidean
  - Manhanttan
  - Cosine
  - …



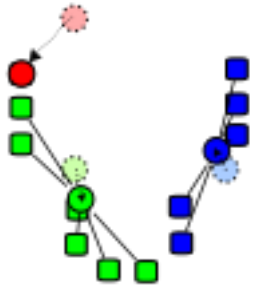
K-Means



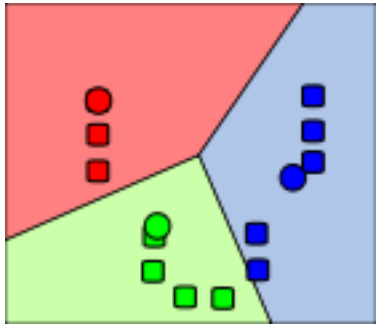
1) k initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).



2) k clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

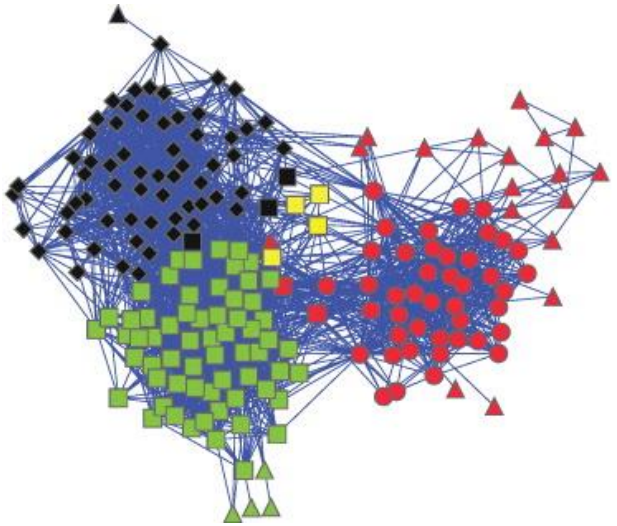


3) The centroid of each of the k clusters becomes the new mean.



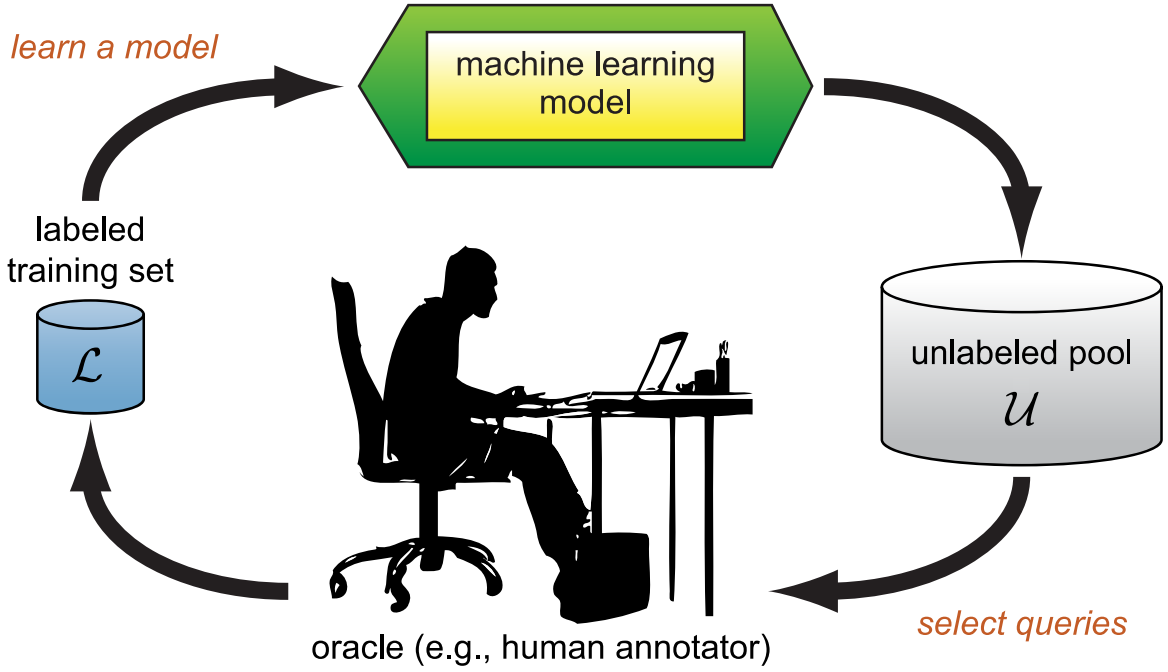
4) Steps 2 and 3 are repeated until convergence has been reached.

- Association rule mining (market basket analysis)
  - {x,y} =>{z}
  - {x,y,z}=>{u,v}
  - …
- Graph-based community detection
  - Modularity maximization-based
    - Networks with high modularity have dense connections between the nodes within modules but sparse connections between nodes in different modules.



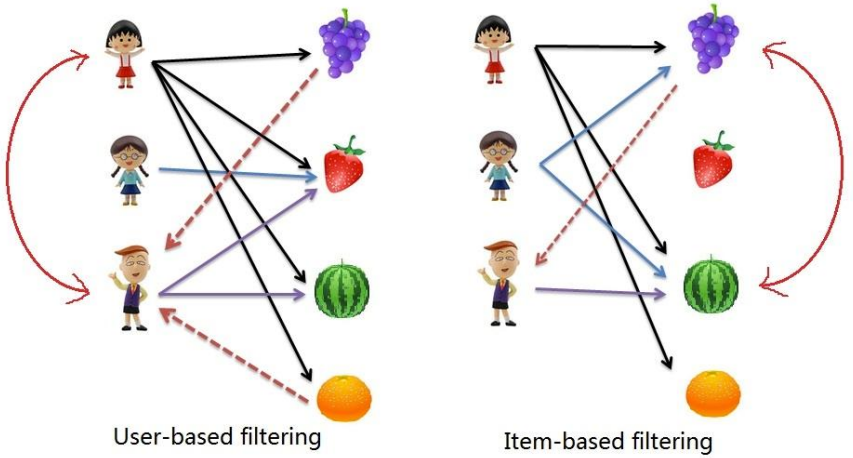
Semi-supervised learning

- Active learning



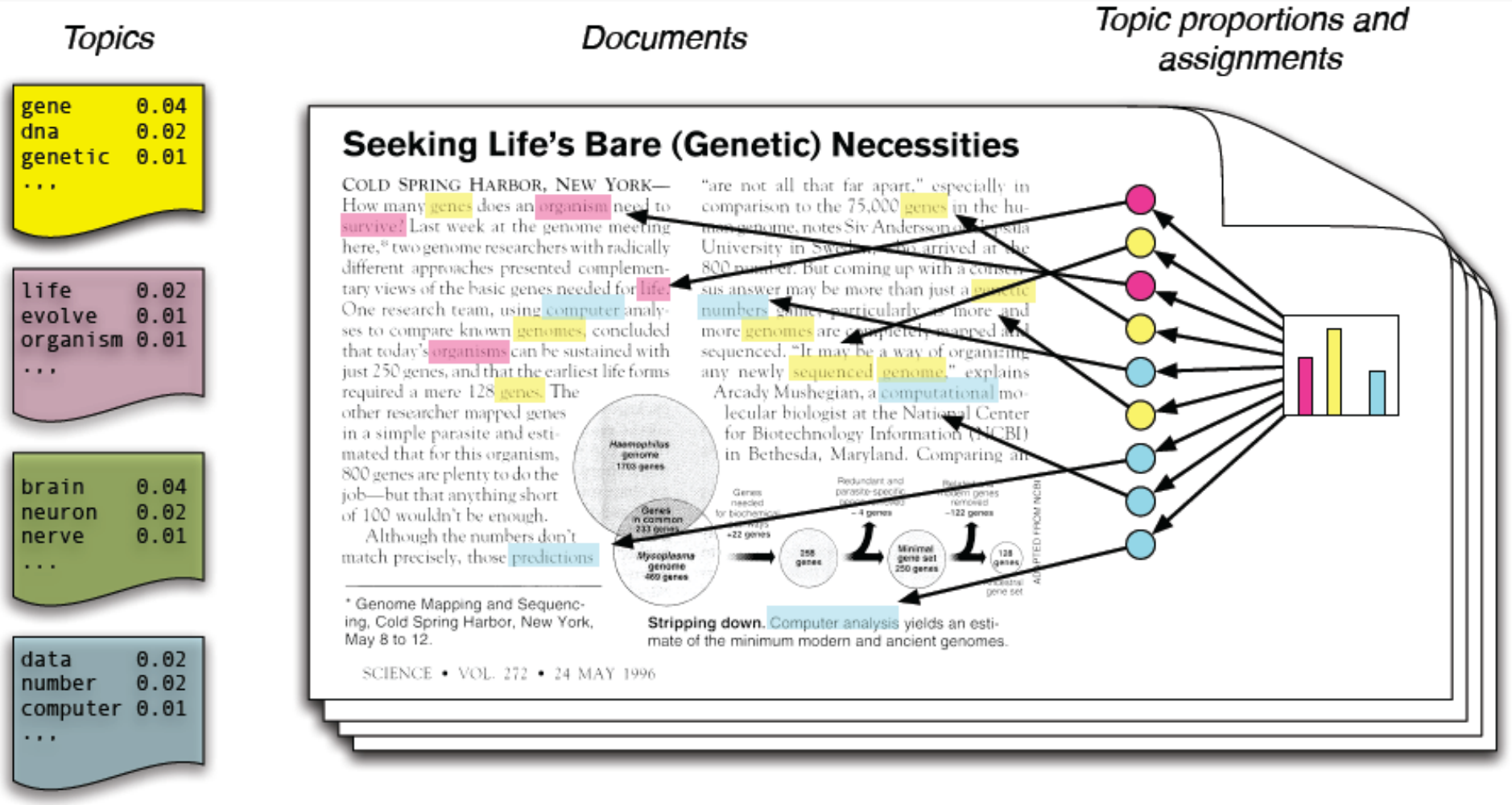
Recommender Systems

- User-based collaborative filtering
- Item-based collaborative filtering



- Sparse matrix completion
  - ❑Netflix problem

Graphical Models: Topic Modeling



FORDHAM UNIVERSITY
THE JESUIT UNIVERSITY OF NEW YORK

Questions?