UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

BIG DATA and AI
for business

Lecture 1 (01/23, 01/28): Introduction to Big Data

Decisions, Operations & Information Technologies
Robert H. Smith School of Business
Spring, 2019


Welcome
Velkomin
Bienvenue
Bienvéni
Bonvenon
Velkommen
Dobrodošli
Selamat Datang
Swagata
Ongietorri
Yokôso
Tervetuloa
Vítejte
欢迎
Welkom
Verwelkom
Benvingut
Bonavinuta
Toivottaa
Willkommen
Herzlich
Dobredojde
добре дошъл
Sugeng Rawuh
Siyakwamukela
Swagatham
Fáilte
Acollir
Hospedar
Mirëpres
Namaste
Heten
iBiala
Qaimarutin
Akwaba
Goscic
Recoger
Kaabo
Croesawu
Sambut
Laukiamas
Swaagat
GnindiTonHap
Karibuni
Khoshumadi
Karibu
Hozta
Murakaza Neza
Sannu da zuwa
Kalosllthate
Degemer
Mai
EKomo
Maligayang Pagdating
Hwangyong Hamnida
Ontvangen
Hoan Nghênh
Uvitani
Teretulnud
Moguah
Woezor
Wëllkom
MirëSevini

Instructor

- Kunpeng Zhang (KZ)
  - Assistant professor at DOIT
- Background: Ph.D. in computer science
- Courses taught before
  - Introduction to business programming
  - Big data analytics

Syllabus

- Lecture discussion
  - Monday, Wednesday
    - 11:00 – 12:15PM (0502)
    - 12:30 – 1:45PM (0501)
  - Room
    - VMH 1333 (0501)
    - VMH 1333 (0502)
- Office hour
  - TBD
- My office
  - Room: 4316 Van Munching Hall

What cover in this course

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Deep Learning
  - ❑ Deep Neural Networks
  - ❑ CNN
  - ❑ RNN
  - ❑ Advanced topics
- Introduction to Hadoop and MapReduce programming
- Hadoop overview
  - ❑ Framework / architecture
  - ❑ Installation and configuration
  - ❑ Cloud computing (Amazon AWS)

What cover in this course

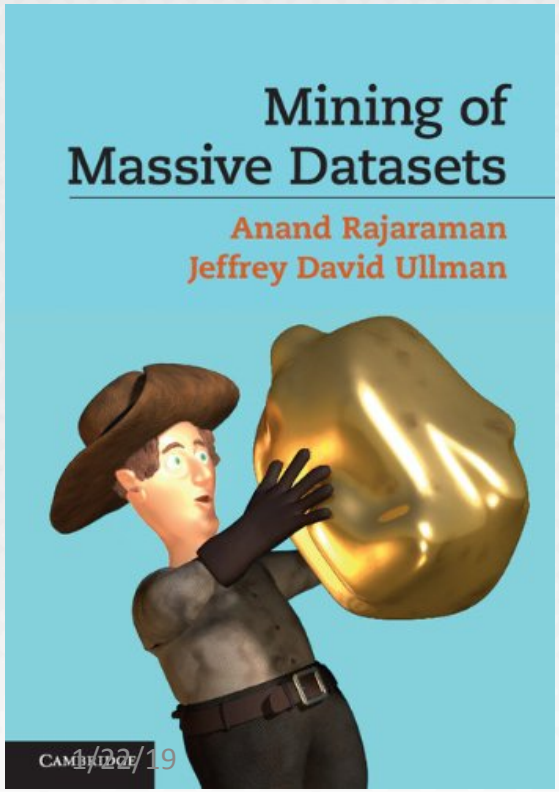
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

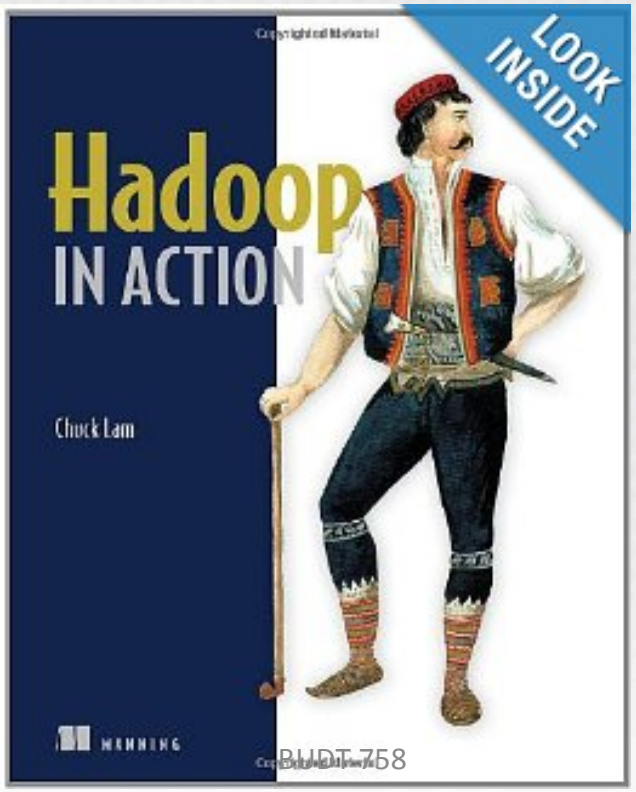
- Data management
  - Advanced SQL query
  - Hive, pig, sqoop
- Spark
  - RDD
  - Spark-SQL
  - Spark-ML/GraphX
  - Applications

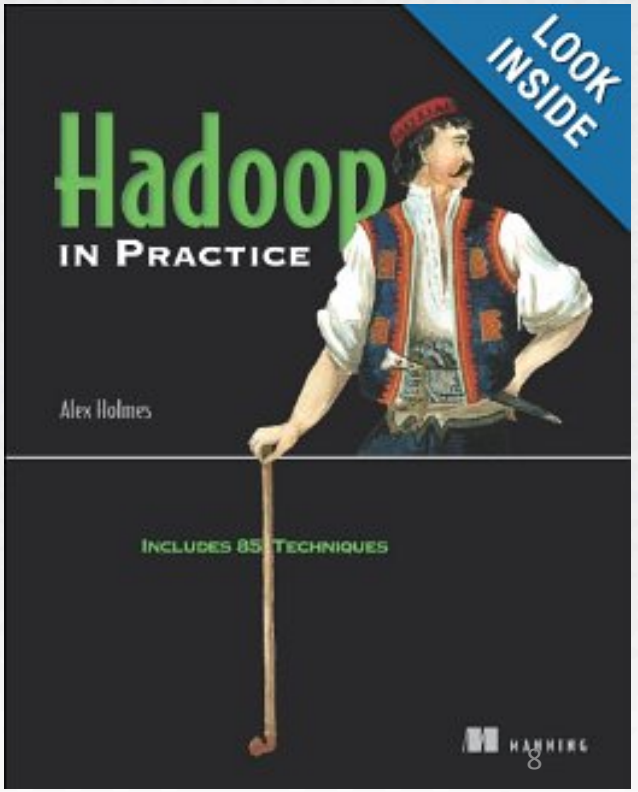
We do not cover …

Recommended textbooks

- No required textbooks, but recommend you read the following books:







Recommended textbooks

- Deep Learning
  - http://www.deeplearningbook.org/

Prerequisites

- Required
  - ❑Data Mining for Business
  - ❑Basic computer programming (Python)

- Recommended
  - ❑Database: SQL knowledge
  - ❑Math
    - Statistics, Probability, and Matrices

Lab session

- Must attend
- ~10 labs (in class)
  - ❑Deep Learning
  - ❑HDFS
  - ❑Data management: Sqoop, Hive, Pig
  - ❑Spark: RDD, SQL, ML

Quiz

- 3 quizzes:
  - ❑About Deep learning, HDFS, MapReduce
  - ❑About Pig, Hive, and Sqoop
  - ❑About Spark

Class project

- Develop a big data and AI system to solve an interesting problem with a large amount of data using what we learn in this course
  - 2/25/2019: project proposal due (1 page)
  - 5/15/2019: project report due
  - 5/13/2019: poster slam
  - Examples:
    - http://cs229.stanford.edu/proj2018/
    - http://cs230.stanford.edu/past-projects/#fall-2018
    - From kaggle competition

Group

- The instructor will assign teams
  - ❑No more than 3 people (strict rule)
  - ❑Have a name and a leader for contacting and coordinating

Grading

Class participation:	10%*1=10%
Class project:	40%*1=40%
(poster: 2 awards)	
Quizzes:	10%*3=30%
Lab assignments:	5%*4=20%

Attendance

- Encourage you to attend every lecture session and lab session
  - ❑Have several random attendance checking
  - ❑Receive ZERO if absence for 5 times
  - ❑Not receive A if absence for 2 times
- Project poster slam session is required to attend

Contact

- TA: Peiyan Yu:
  - peiyan.yu@rhsmith.umd.edu
- Non-regular Office Hour
  - Appointments by email preferred
- Email
  - kzhang@rhsmith.umd.edu
- EMLS

Plagiarism

- For lab assignment, both receive ZERO credit and the final grade will be down one level.
- For quizzes, both will be sent to the Graduate Office.
- For project, all members in all groups will be sent to the Graduate Office.

Tentative schedule

Session	Topics	Lab	Assignment Due
1/23/19	Introduction		
1/28/19	Business Value of Big data and AI, Internet of Things		
1/30/19	Deep Learning – Introduction (1)		
2/4/19	Deep Learning – Introduction (2)		
2/6/19		Deep Learning	
2/11/19	Deep Learning – CNN		
2/13/19		Deep Learning	
2/18/19	Deep Learning – RNN		
2/20/19	Deep Learning – Advanced		
2/25/19		Deep Learning	Project proposal Due
2/27/19	Overview of Hadoop Ecosystem, HDFS		
3/4/19	The MapReduce Framework	Set up Cloudera Training Virtual Machine	
3/6/19		Yarn and Hue	
3/11/19			Quiz 1
3/13/19	Sqoop		
3/18/19	**Spring break (No class)**		
3/20/19			
3/25/19	Pig		
3/27/19		Sqoop and Pig	
4/1/19	Hive		
4/3/19		Hive	
4/8/19	AWS - Introduction	AWS (Pig and Hive on Cloud)	
4/10/19			Quiz 2
4/15/19	Spark - Introduction		
4/17/19	Spark - RDD		
4/22/19	Spark - SQL	RDD	
4/24/19	Spark - ML/GraphX	Spark SQL	
4/29/19	Big Data ML Applications: Clustering	Spark ML/GraphX	
5/6/19		*K*-means on Spark	
5/8/19			Quiz 3
5/13/19	Poster slam		
5/15/19			Project report due

Introduction to big data

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

- Importance
- Definition and characteristics
- Applications
- Big data analysis pipeline
- Challenges
- Analytical techniques
- Review of data mining algorithm

Why big data?

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Science
Engineering
Business
Healthcare
…

Importance of big data

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

Government

- In 2012, the Obama administration announced the Big Data Research and Development Initiative 84 different big data programs spread across six departments.

Private Sector

- 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.

Science

- Large Synoptic Survey Telescope will generate 140 Terabyte of data every 5 days.
- Medical computation like decoding human Genome.
- Social science revolution.
- New way of science (Microscope example).

Many opportunities

- Many demands from different domains, including finance, IT, biology, physics, ...
- 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)
- Big Data industry is worth more than $100 billion growing at almost 10% a year (roughly twice as fast as the software business)

Big data analytics: data mining, statistics, computer programming, business intelligent, and others.

Usage example of big data

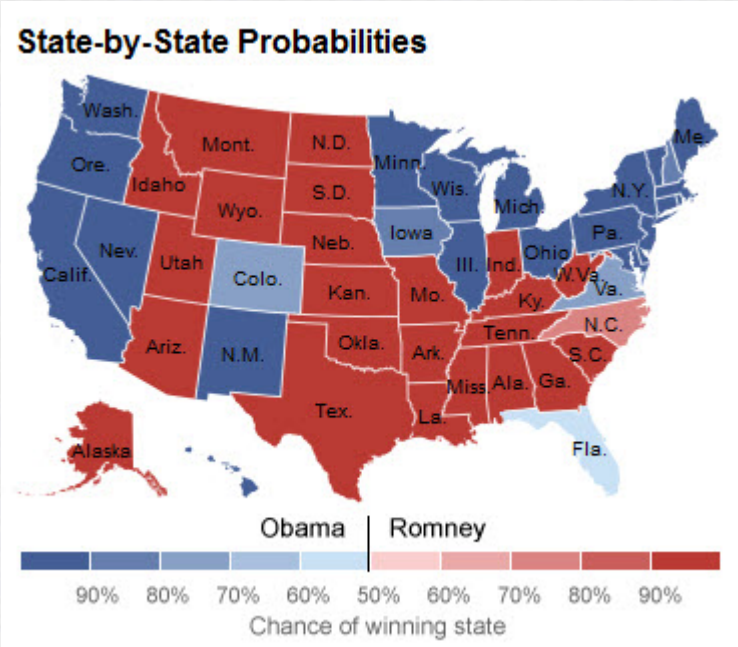
- Predictive modeling
- mybarackobama.com
- Drive traffic to other campaign sites
  - Facebook page (33 million "likes")
  - YouTube channel (240,000 subscribers and 246 million page views).
- Every single night, the team ran 66,000 computer simulations.

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

Prediction for US 2012 Election

Nate Silver's, Five thirty Eight blog predicts Obama had a 86% chance of winning predicted all 50 state correctly

Sam Wang, the Princeton Election Consortium: The probability of Obama's re-election at more than 98%

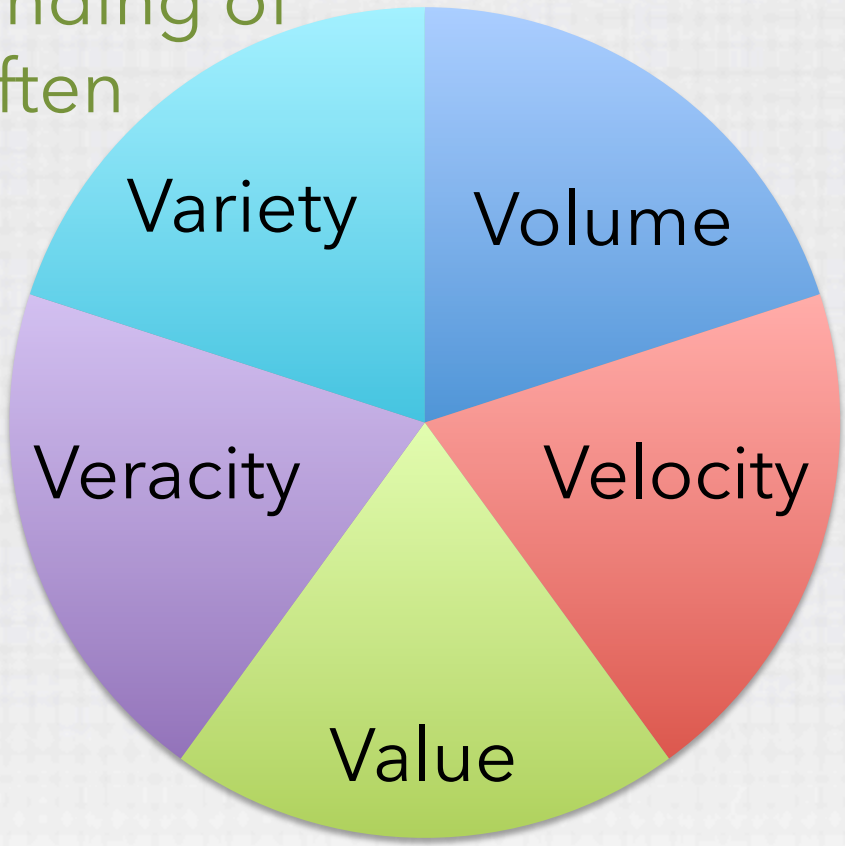


What is big data?

- Big data is a blanket term for any types of data sets so large and complex that it becomes difficult to process using on-hand data management tools or traditional data processing applications. [from Wikipedia]

5 Vs of big data

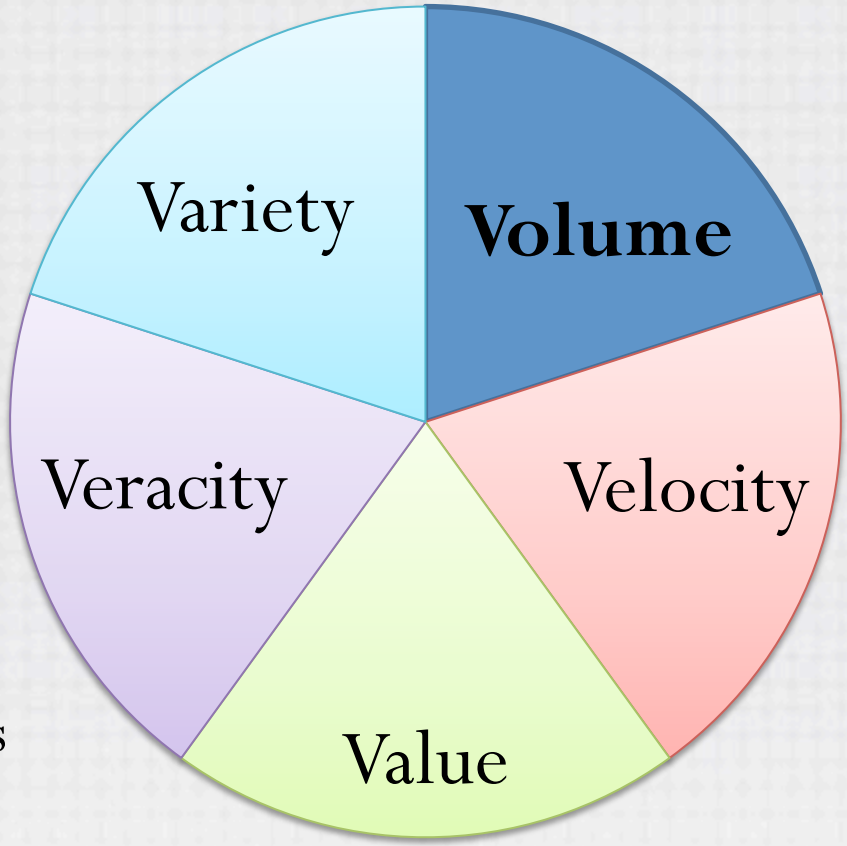
To get better understanding of what big data is, it is often described using 5 Vs.



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

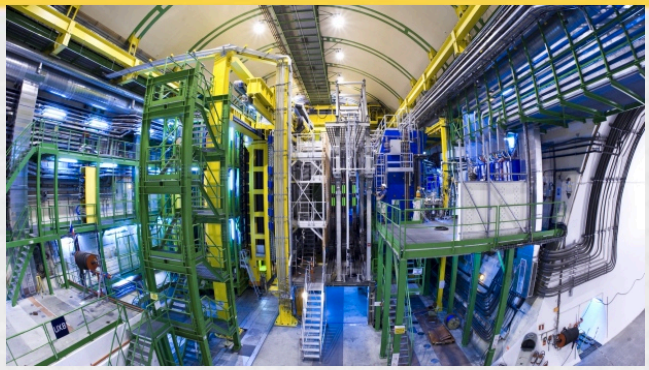
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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 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

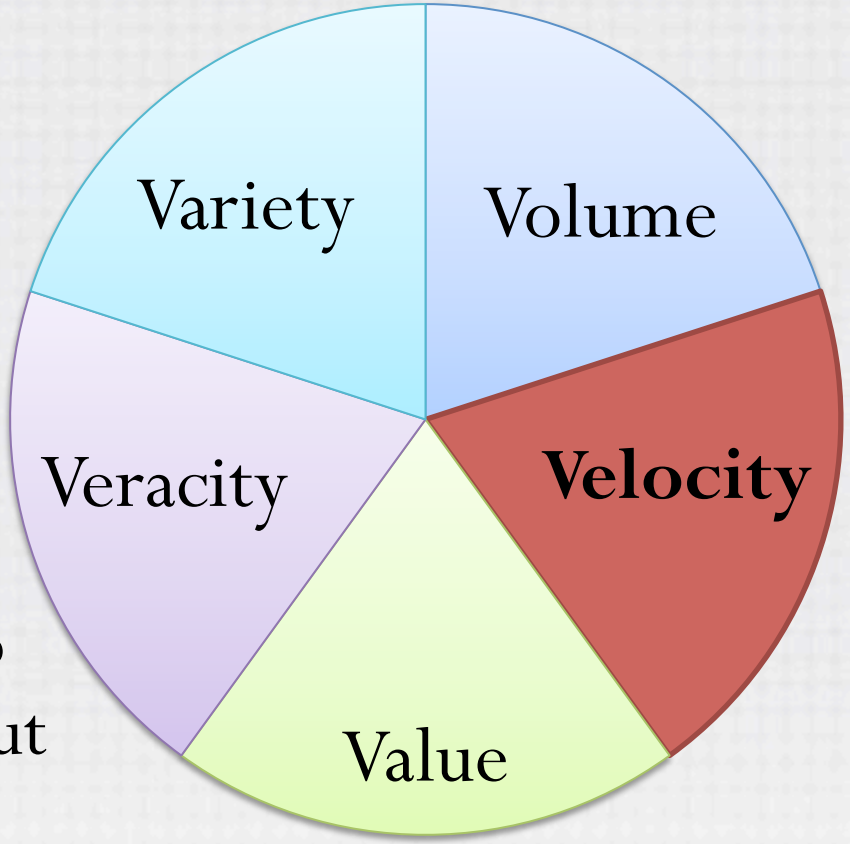
150 PB on 50k+ servers running 15k apps (6/2011)

CHASE

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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

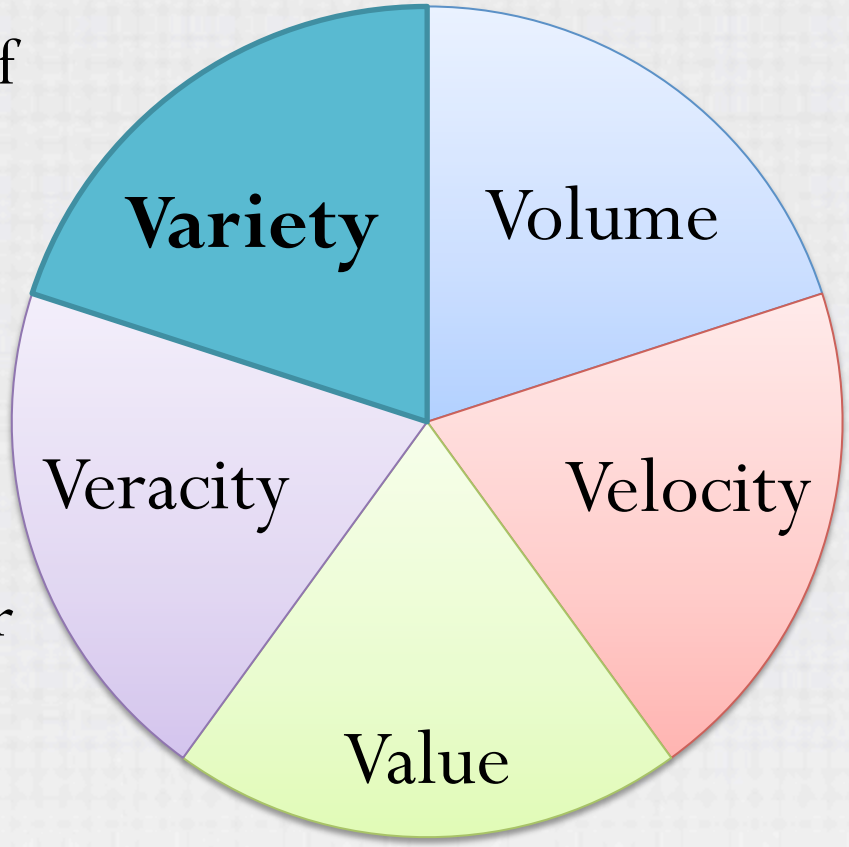
Velocity refers to the speed at which new data is generated and the speed at which data moves around. Just think of social media messages going viral in seconds. Technology now allows us to analyze the data while it is being generated (sometimes referred to as it in-memory analytics), without ever putting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



We see increasing variety of data types

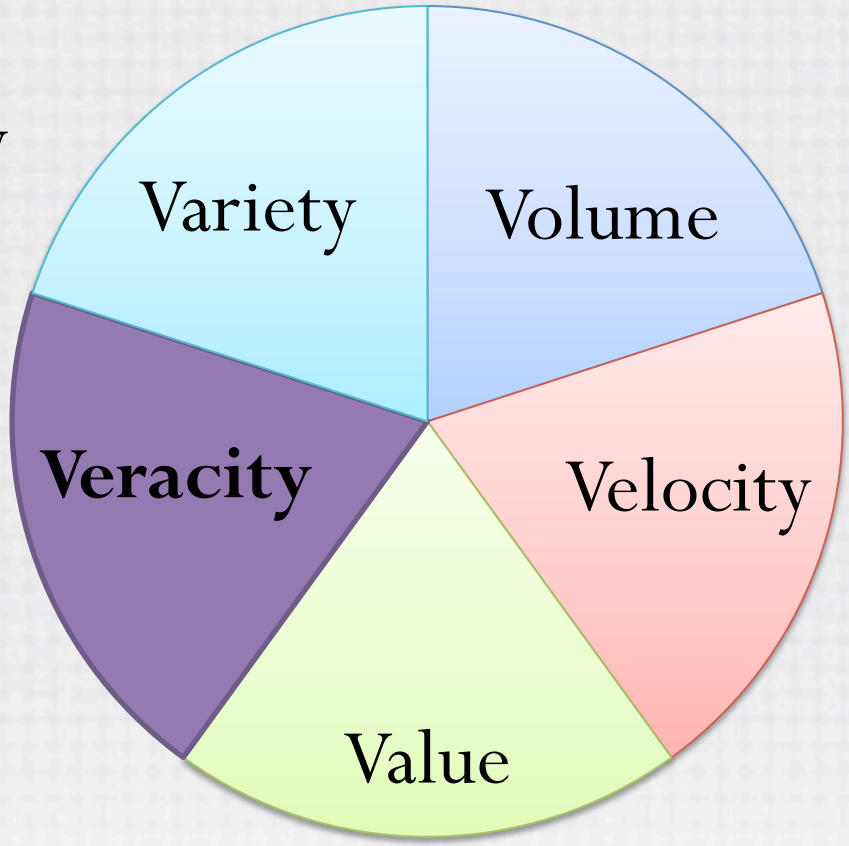
UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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.



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.

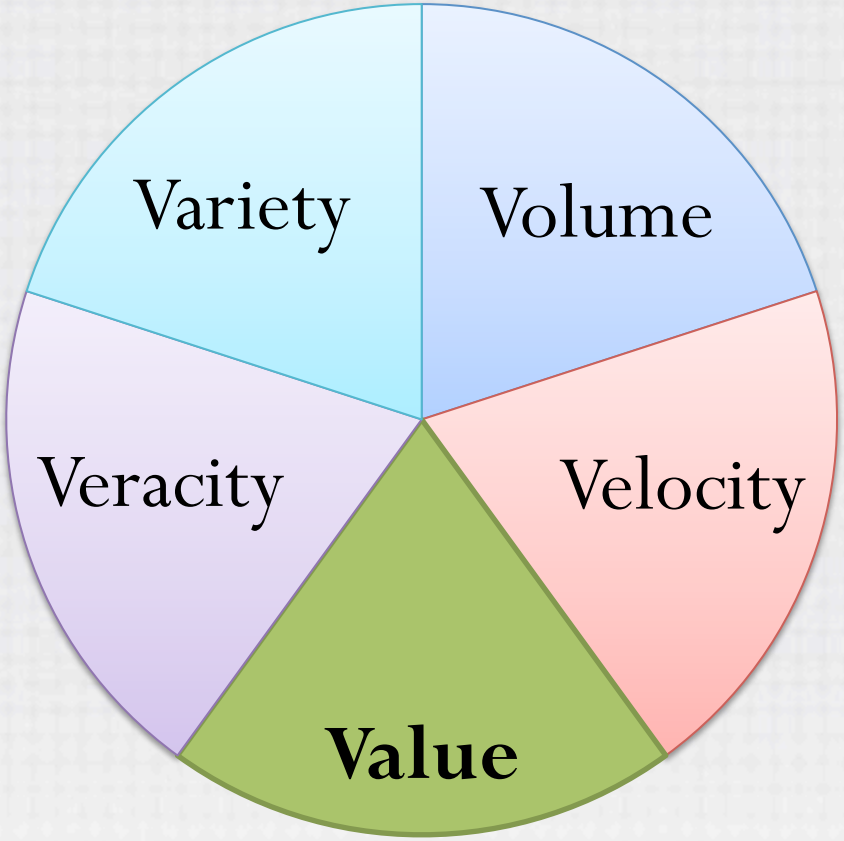


Value – the most important V of all

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.

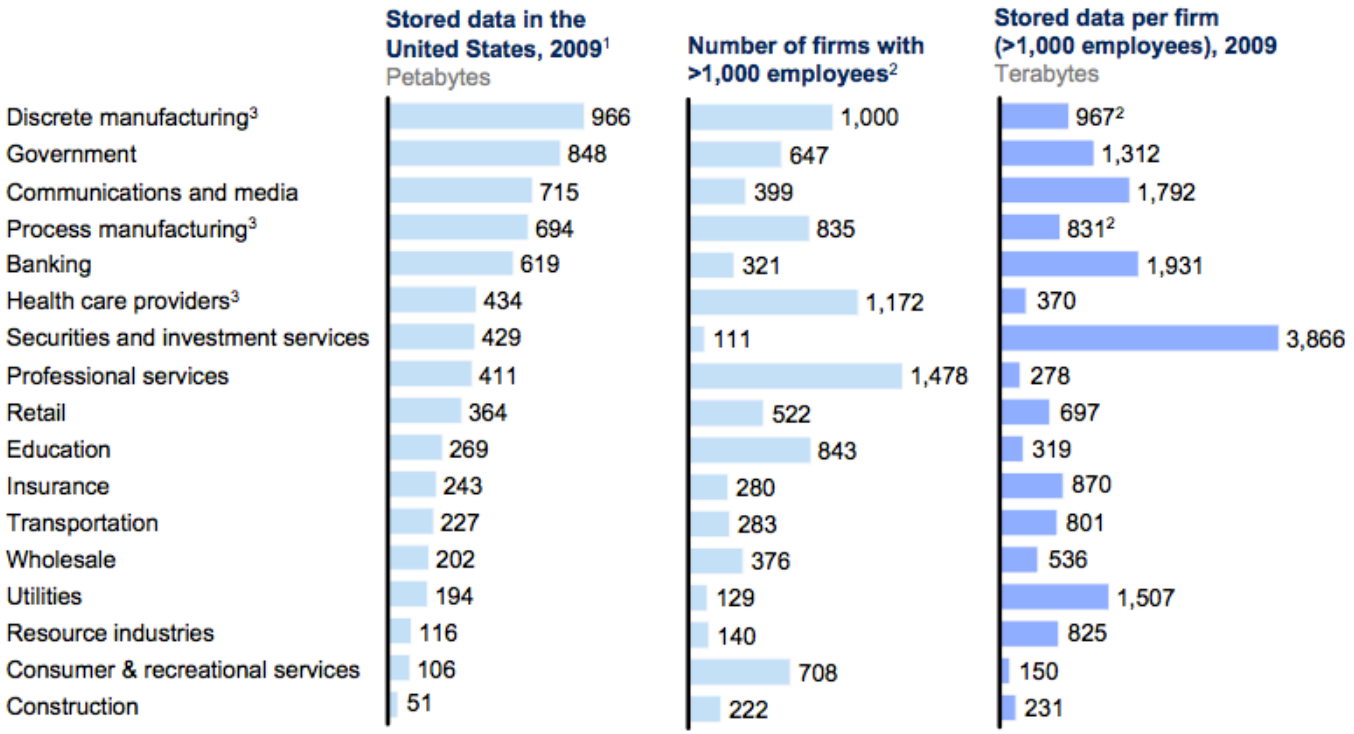


Why big data matters to us?

Big data is more prevalent than you think

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

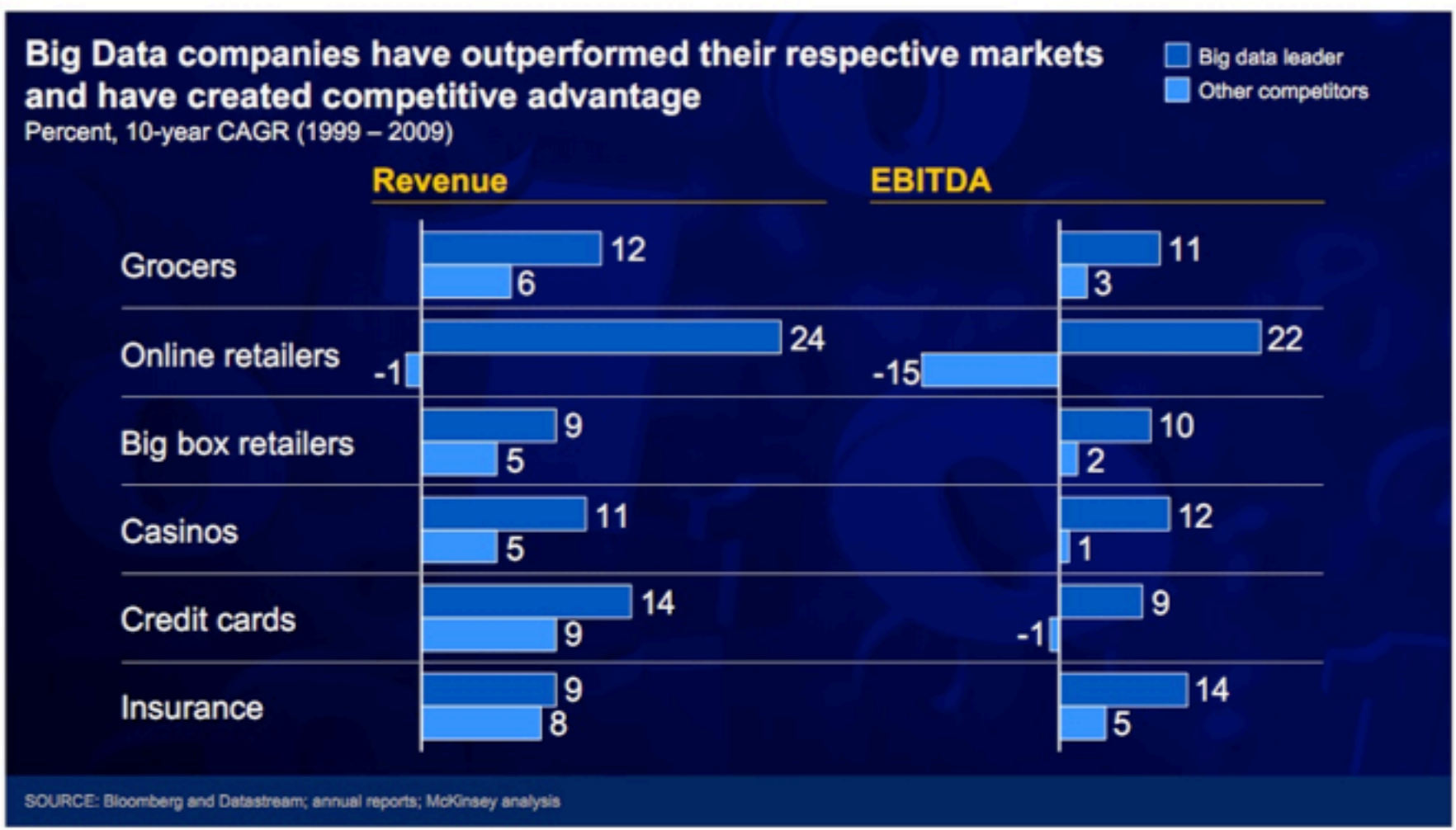
Companies in all sectors have at least 100 terabytes of stored data in the United States; many have more than 1 petabyte



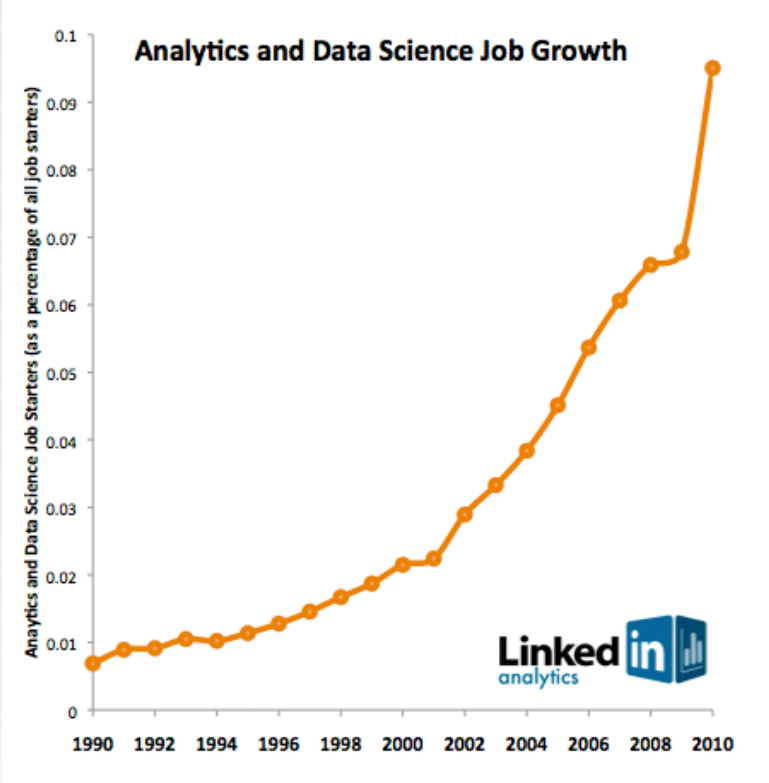
1 Storage data by sector derived from IDC.
2 Firm data split into sectors, when needed, using employment
3 The particularly large number of firms in manufacturing and health care provider sectors make the available storage per company much smaller.

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

Competitive advantages gained through big data



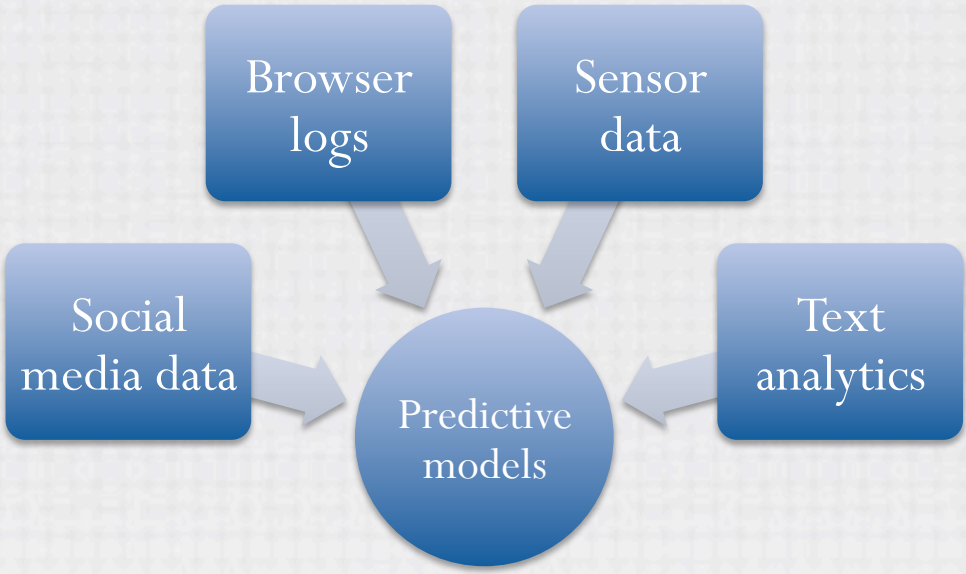
Big data jobs



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

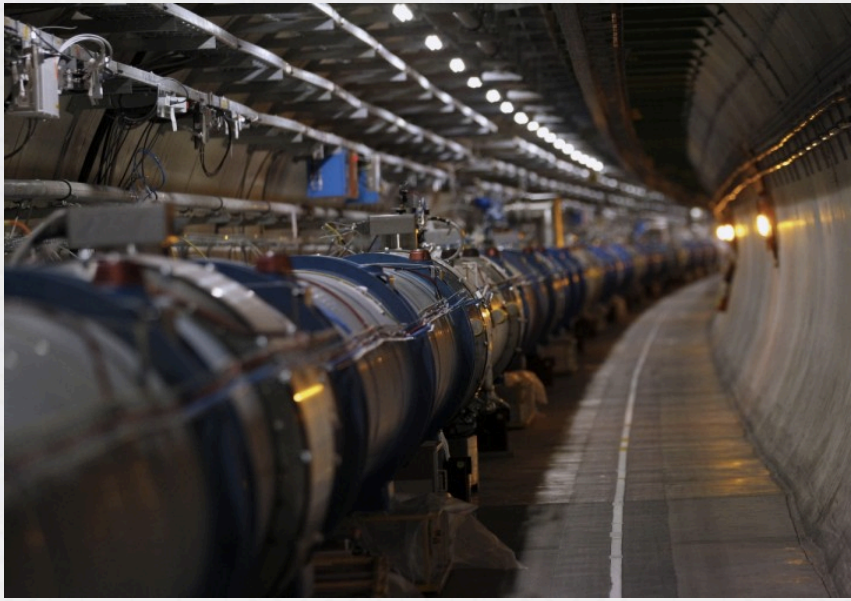
- Big data techniques are already being used to monitor babies in a specialist premature and sick baby unit;
- Big data analytics allow us to monitor and predict the developments of epidemics and disease outbreaks;
- By recording and analyzing every heart beat and breathing pattern of every baby, infections 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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

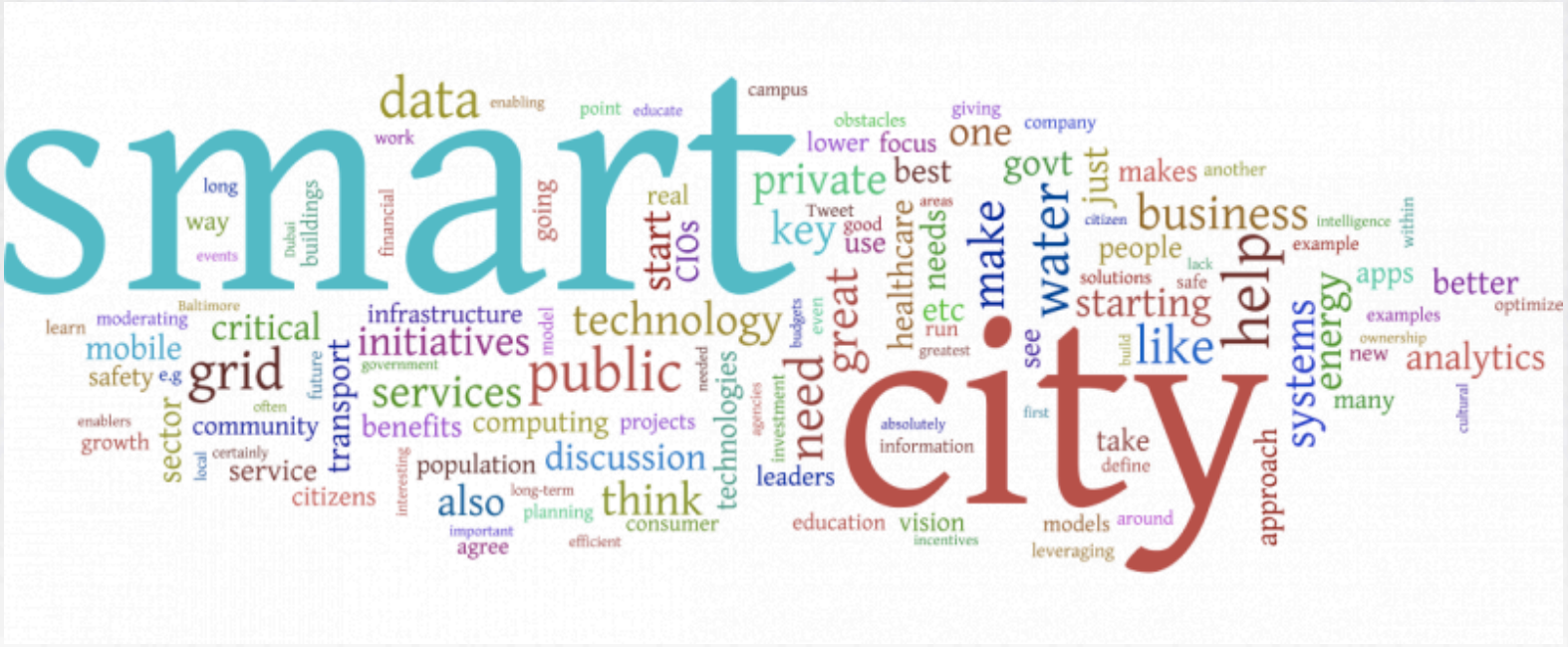
- Google self-driving car: the Toyota Prius is fitted with cameras, GPS, powerful computers and sensors to safely drive without the intervention of human beings;
- Big data tools are also used to optimize energy grids using data from smart meters.

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.

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).

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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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

Phase #5

- Interpretation
  - Report
  - Visualization

Challenges in Big data

Challenge #1

- Heterogeneity and incompleteness
  - Data from different sources/platforms
  - Data formats are different
  - Data missing due to security, privacy, or other reasons

Challenge #2

- Scaling: data volume is scaling faster than compute resources.
  - Moving towards cloud computing

Challenge #3

- Timeliness
  - Query and indexing techniques to find suitable elements/records quickly

Other challenges

- Privacy
- Human collaboration

Applications, data, and corresponding commonly used analytical techniques

1. E-Commerce and marketing intelligence

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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
- 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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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

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

UNIVERSITY OF MARYLAND
ROBERT H. SMITH SCHOOL OF BUSINESS

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

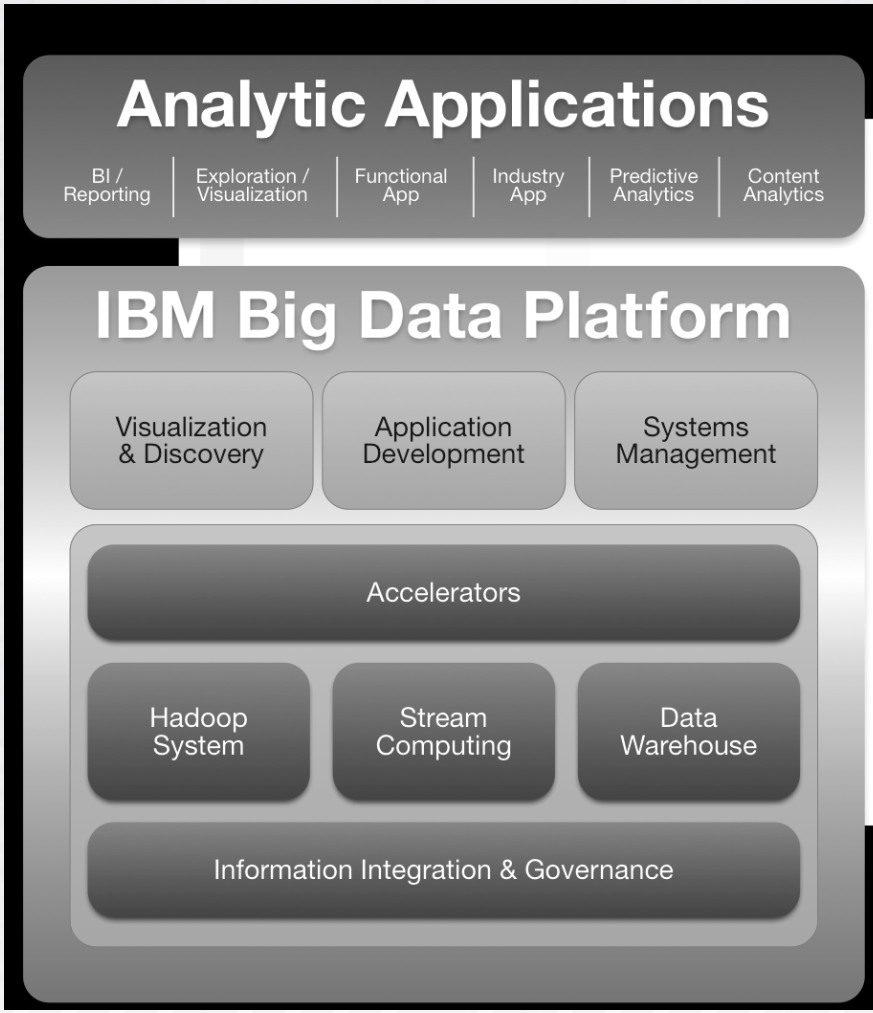
Analytics

- 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

Big Data Platforms



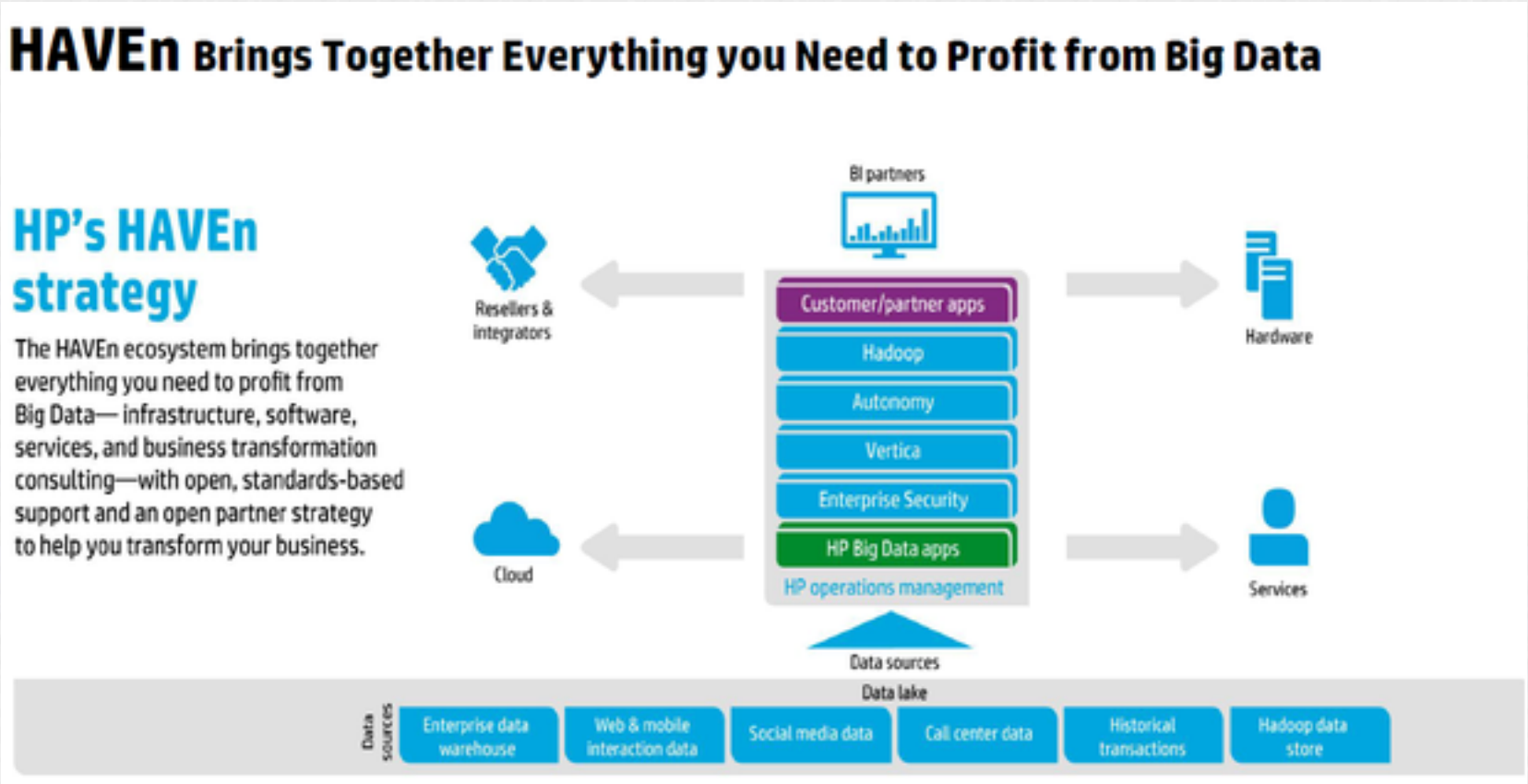
New Analytic applications drive the requirements for a big data platform

- Integrate and manage the full variety,velocity and volume of data
- Apply advanced analytics to information in its native form
- Visualize all available data for adhoc analysis
- Development environment for building new analytic applications
- Workload optimization and scheduling Security and Governance

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
Hardware
Customer/partner apps
Hadoop
Autonomy
Vertica
Enterprise Security
HP Big Data apps
HP operations management
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