BIG DATA ANALYTICS

Introduction

Olga Klopp
klopp@essec.edu

Course Objectives

- Understand challenges and opportunities associated with Big Data
- Provide methodological principles of multidimensional data analysis methods
- Learn how to deploy effective decision-making models to a production environment

Course organization

- **Lectures:**
  - Presentation of the statistical methods
  - Some practical examples
- **Grading pattern:**
  - Case study 40%
  - Quiz: **27th of November** 60%

Tools

-  
- graphics and visualization tools: ggplot2
- documentation generation: knitr or R Markdown
- For the case study you can use: R, Python, Apache Spark ...
- We will work specific problems in R; to understand whats going on, youll need to run the examples by yourselves. **Use R** `help()`!

Ref: **- R in Action, Second Edition, by Robert Kabacoff (www. manning.com/kabacoff2/)**

Short round table to introduce yourself

- Educational background
- Career goals
- What is Big Data for you?
- Experience in data

What is Big Data?

- Big Data is a term that refers to *extremely large, very fast, highly diverse and complex* data that can not be managed with traditional data management tools.
- Big Data includes all kinds of data
- Ideally, it helps deliver the right information at the right time and make the right decisions.

Objectives for today

- We will see the key challenges and benefits of Big Data
- An overview of the essential tools and technologies:
  - Hadoop, MapReduce, NoSQL databases ...

Outline

Understanding Big Data

Why is Big Data a Big Deal?

Big Data applications

Technology Challenges for Big Data

Big Data and Privacy

Big Data Analytics

Outline

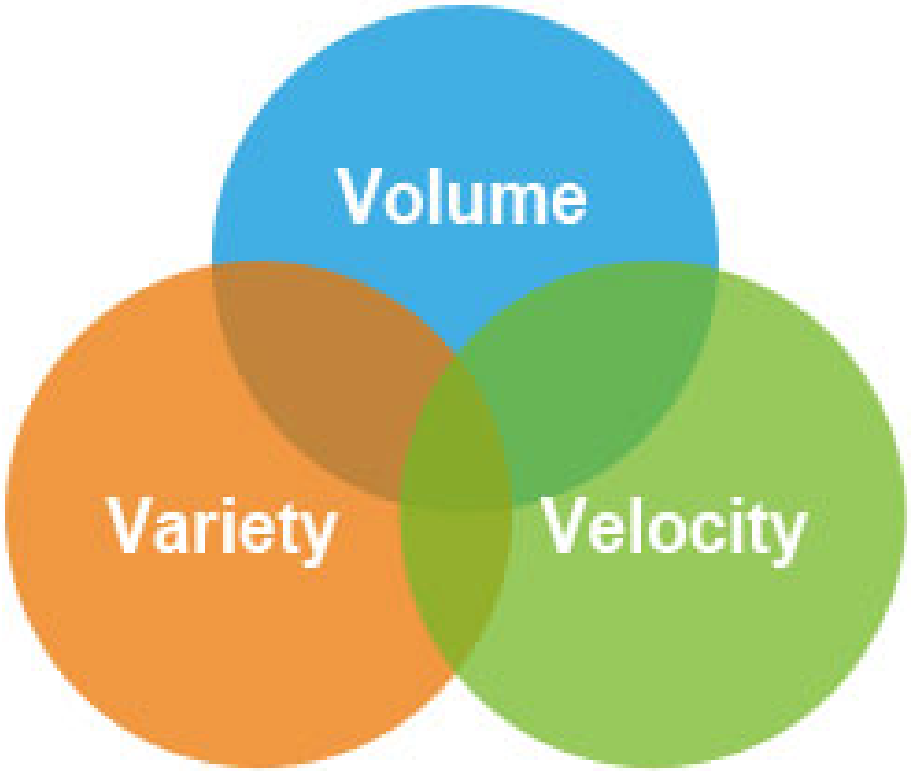
How big is Big Data?

Big Data

if the data can not be handle in "reasonable" or "useful" time by a system composed of a single node.

- If we want to analyze in a few minutes 1 teraoctet (To) of the data (size of a hard disk) one needs to use parallel processing and storage in multiple nodes
- **Big Data is not only the size of the data set but also the speed of the processing**

Three V's of BigData



3Vs of Big Data

3 V's: Volume

- Volume is the V most associated with Big Data

BigData: some numbers

- In the modern world we are producing data voraciously:
  - Every day hundreds of millions of people take photos, make videos and send texts
  - Across the globe, businesses collect data on consumer preferences, purchases and trends
  - Governments regularly collect all sorts of data from census data to incident reports in police departments
- This deluge of data is growing fast:
  - the total amount of data in the world was 4.4 zettabytes in 2013. That is set to rise steeply to 44 zettabytes by 2020

Recall

$$\text{zettabyte} = 10^{21} \text{ bytes}$$

- Go = gigaoctet = 10^9 octets
- To = teraoctet = 10^{12} octets
- Po = petaoctet = 10^{15} octets

Data or Information?

- Big volume of data does not represent value by itself
- **What kind of information we can extract from the data sources**
  - determine customers' purchasing patterns $\rightarrow$ personalize the offer

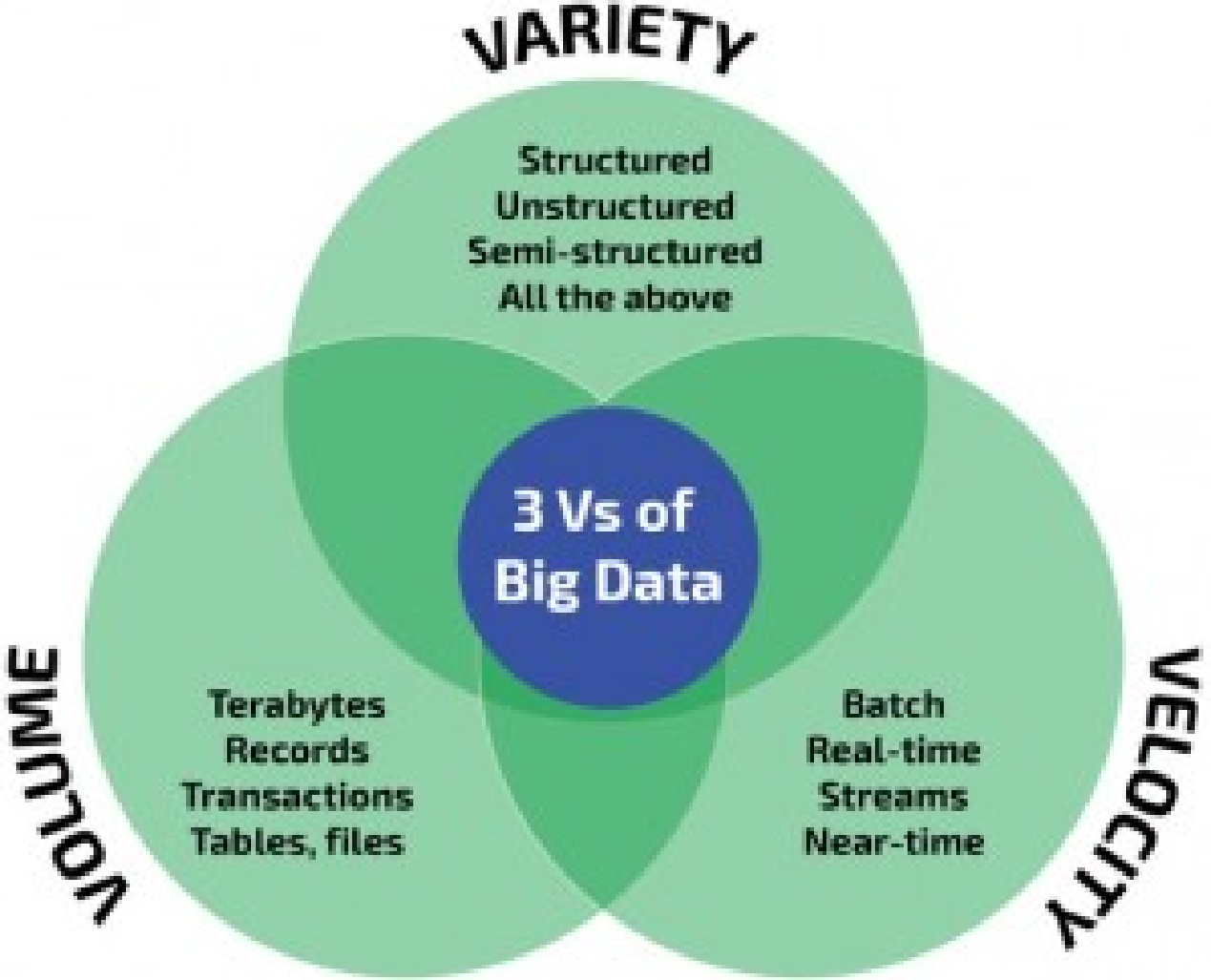
3 V's: Velocity

- How fast the data is coming in
- The speed of data flow has experimented a similar transformation as that of the volume of the data:
  - Facebook has to handle a tsunami of photographs every day. It has to ingest it all, process it, file it, and later, be able to retrieve it.
  - Trading: velocity is a real competitive advantage
  - The key factor for in real time processing of customer wishes and supply availability.

3 V's: Variety

- Tweets, photographs, sensor data ... each of these is very different from the others
- Big ambition: cross-checking information and matching data from a variety of sources
  - the main interest
  - for the moment, there is no universal solution
  - the most tricky problem with a solution that needs to be adapted to each particular situation
  - Example: the use of highly variable data in matching between the data of a CRM (Customer Relationship Management) system, geo-localization data, data from social media in the goal of expand user's profile with personal information.

Three V's of BigData



Veracity

- Accuracy and quality of data: Big Data is messy
- Misinformation and disinformation.
- The reasons for poor quality of data can range from technical error, to human error and malicious intent:
  - The source of information may not be reliable.
  - The data may not be communicated and received correctly because of human or technical failure.
  - The data provided may also be intentionally wrong. There could be disinformation and malicious information spread for strategic reasons.

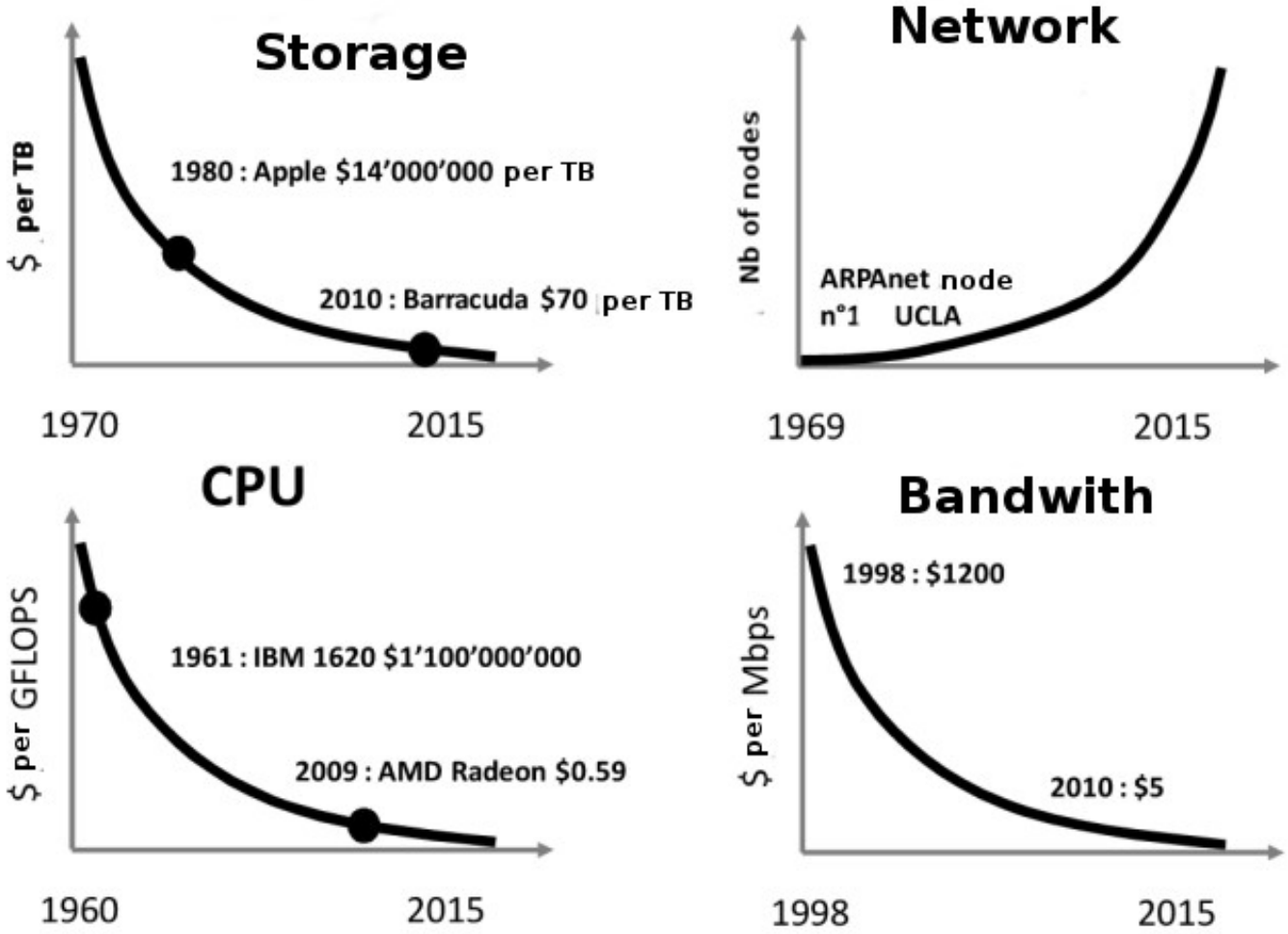
Big Data needs to be sifted and organized by quality.

Why Big Data now?

- Price of IT devices has fallen substantially

Why Big Data now?

- Price of IT devices has fallen substantially



Why Big Data now?

- Internet Giants (Google, Facebook, Yahoo, Amazon...) developed new technologies to take advantage of it:
  - Google: MapReduce (programming model for processing Big Data sets with a parallel distributed algorithm on a cluster)
  - Open source parallel processing system: Hadoop Apache
  - New non relational mechanism for storage and retrieval of data: NoSQL database
    - Key-value database: data is presenting as a collection of key-value pairs s.t. each possible key appears at most once, e.g., [digit, its base-ten logarithm]
    - Graph

Recall: Relational Data Base

A digital database whose organization is based on the relational model of data

School Table

ID	Name
S001	University of Technology
S002	University of Applied Science

Student Table

School ID	ID	Name	DOB
S001	UT-1000	Tommy	05/06/1995
S001	UT-1000	Better	16/04/1995
S002	UAS-1000	Linda	02/09/1995
S002	UAS-1000	Jonathan	22/06/1995

Recall: Relational Data Base

School Table

ID	Name
S001	University of Technology
S002	University of Applied Science

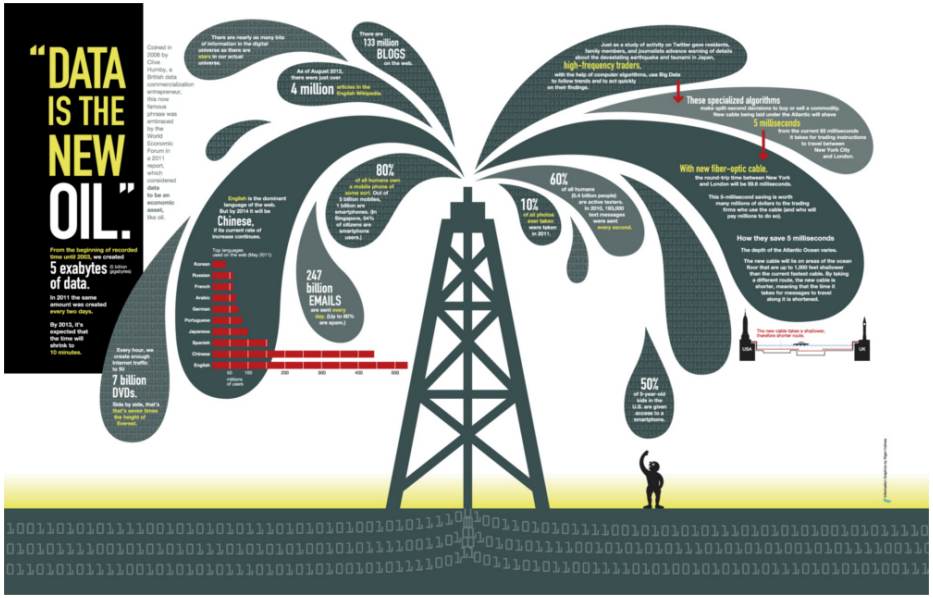
Student Table

School ID	ID	Name	DOB
S001	UT-1000	Tommy	05/06/1995
S001	UT-1000	Better	16/04/1995
S002	UAS-1000	Linda	02/09/1995
S002	UAS-1000	Jonathan	22/06/1995

- The data is organized into one or more tables of columns and rows
- Each table represents one "entity type" (e.g. customer or product)
- Rows represent instances of that type of entity (e.g. name or type of product)
- Columns represent values attributed to that instance (address or price)

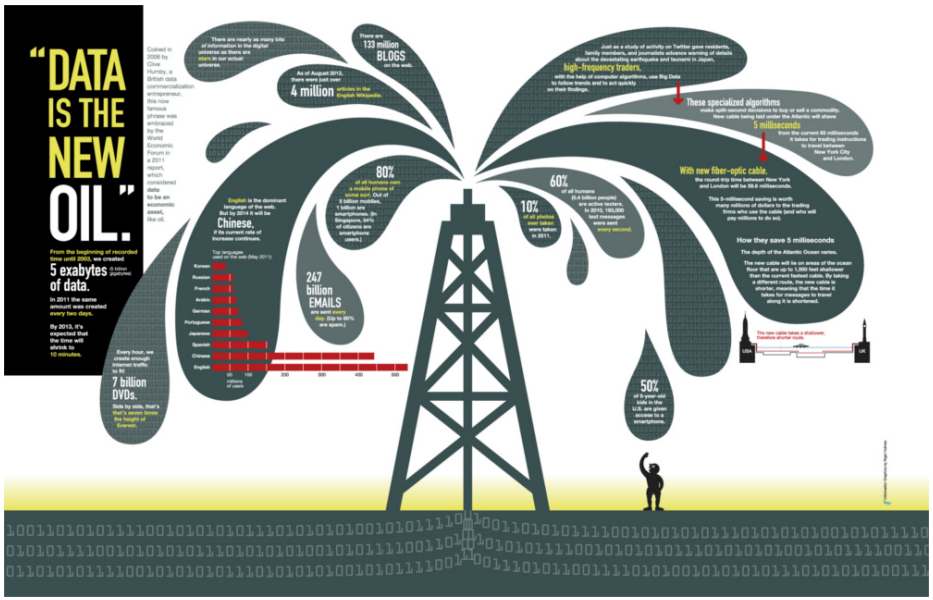
Outline

New Oil



Entire business sectors are being reshaped by Big Data

New Oil

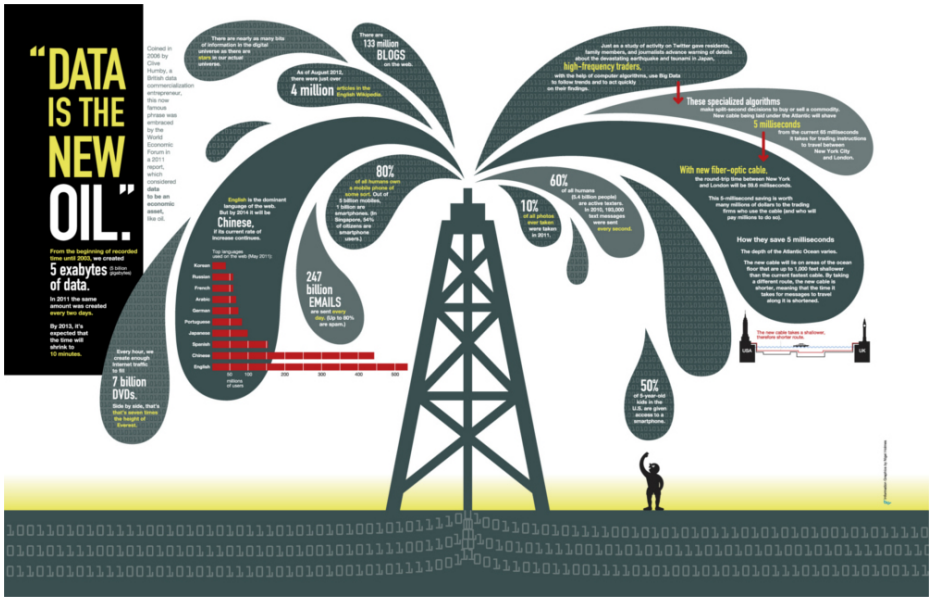


Entire business sectors are being reshaped by Big Data

- For many businesses, their critical data used to be limited to their transactional databases and data warehouses:
  - Data organized into orderly rows and columns
  - Every byte of information well understood in terms of its nature and business value.

Still extremely important ...

New Oil



Entire business sectors are being reshaped by Big Data

- Business are now differentiating themselves by how they are finding value in the large volumes of other kind of data:
  - Internally: website clickstream data, typed notes from call center operators, emails and instant messaging ...
  - Externally: open data from public and private entities, such as social media data ...

New Oil

- Learn about their consumers, improve their service delivery and design new products
- Gather and store this data for later analysis, or sell it to other organizations that might benefit from it
- Discard parts of their data for privacy or legal reasons

Businesses can not afford to ignore this shift in mindset about how data can be used, that is, to create a new form of economic value.

New Oil

- *Proctor&Gamble* have inserted Big Data into all aspects of its planning and operations
- *Volkswagen* requires all its business units to identify some realistic initiative using Big Data to grow their units sales
- *Netflix* processes 400 billion user actions every day to understand their customers needs.

Outline

Big Data applications

- Big Data applications exist in many industries and aspects of life.
- There are three major types of Big Data applications:
  - **Monitoring and Tracking**
  - **Analysis and Insight**
  - **Product Development**

Monitoring and Tracking applications

- The first and basic applications of Big Data
- Help improve the efficiency of business

Consumer Sentiment Monitoring

- Social media has become more powerful than advertising
- Many consumer goods companies have moved a bulk of their advertising budget from traditional media into social media
- Companies have set up Big Data listening platforms where social media data streams (tweets, Facebook posts and blog posts) are filtered and analyzed for certain keywords, by certain demographics and regions
- Information from this analysis is delivered to marketing professionals

Introducing a New Coffee Product at Starbucks

- Starbucks was introducing a new coffee product but was concerned that customers would find its taste too strong
- The morning that the coffee was rolled out, Starbucks monitored blogs, Twitter, and coffee forum discussion groups to assess customers reactions.
- By mid-morning, Starbucks discovered that although people liked the taste of the coffee, they thought that it was too expensive.
- Starbucks lowered the price, and by the end of the day all of the negative comments had disappeared

Introducing a New Coffee Product at Starbucks

Compare this fast response with a more traditional approach:

- waiting for the sales reports to come in and noticing that sales are disappointing
- A next step might be to run a focus group to discover why
- Perhaps in several weeks Starbucks would have discovered the reason and responded by lowering the price.

Monitoring Trucks at U.S. Xpress

- U.S. Xpress is a transportation company
- Its cabs continuously stream more than 900 pieces of data related to the condition of the trucks and their locations.
- This data is stored in the cloud and analyzed in various ways
- Information delivered to various users, from drivers to senior executives

Monitoring Trucks at U.S. Xpress

- When a sensor shows that a truck is low on fuel, the driver is directed to a filling station where the price is low.
- If a truck appears to need maintenance, drivers are sent to a specific service depot.
- Routes and destinations are changed to ensure that orders are delivered on time
- ...

By monitoring its trucks, U.S. Xpress has saved millions in fuel costs and reduced emissions into the environment

Preventive Machine Maintenance

- All machines, including cars and computers, will fail sometime
- Any important equipment could be equipped with sensors
- The continuous stream of data from sensors is monitored and analyzed to forecast the status of key components, and, thus, monitor the overall machine's health

Preventive maintenance reduces the cost of downtime

Improved Risk Management

Credit card monitoring provides a good example of a streaming application:

- A common fraudulent sequence of events is
  - $5 charge for gas at a convenience store (to see whether the credit card is good)
  - the purchase of thousands of dollars of electronic equipment at a big-box store
- When this stream is detected, store personnel are alerted to possible fraud.

Analysis and Insight Applications

- Help increase the effectiveness of business and to have transformative potential.

- Big Data is structured and analyzed to produce insights and patterns that can be used to make business and public services better

Predictive Policing

- The Los Angeles Police department (LAPD) invented the concept of Predictive Policing
- They worked with UC Berkeley researchers to analyze iThosts database of 13 million crimes recorded over 80 years
- They predicted the likeliness of crimes of certain type, at certain times, and in certain locations
- Crime patterns were mathematically modeled using a model for earthquakes

By aligning the police cars patrol schedules in accordance with the models predictions, the LAPD was able to reduce crime by 12 % to 26 % for different categories of crime.

Winning Political Election

- Political organizations use Big Data to **micro-target** voters and win elections.
- Both Obama 2008 campaign and his reelection made extensive use of organized and analyzed information.

Micro-targeting

- It begins with a voter database:
  - While votes are private, voter registration and voting records (whether and when you actually voted) are public
  - Those records form the starting point of voter databases
- Obamas vast campaign database includes supplemental information gleaned from commercial and other sources:
  - information on voters' magazine subscriptions, car registrations, housing values and hunting licenses, etc
  - scores estimating how likely they were to cast ballots for his reelection

Micro-targeting

- Obama's vast campaign database includes supplemental information gleaned from commercial and other sources
  - Voter contacts in person, on the phone, via e-mail or through visits to the campaigns Web site
  - Facebook friends, along with scores measuring the intensity of those relationships and whether they lived in swing states.
  - If their last names sounded Hispanic, a key target group for the campaign, the database recorded that, too.

Micro-targeting

- The campaign learns what messages work best with whom:
  - perhaps a candidate will learn that Californian Latinas with older children respond most strongly to messages about work opportunities for young people
  - those with very young children are more concerned about public schools
  - while gun owners in general might react negatively to the candidates gun control stance, they react positively to the same candidates economic plan

This information tells candidates and their campaign staff what messages appeal most to voter groups and even individual voters

Micro-targeting

- Another benefit of having the right data and using it well is greater impact from the paid advertising budget:
  - Obama for America campaign used TV viewership data available from TV ratings firms to reach desired audiences at lower than normal cost.

The right data collection and analytics can enable campaigns to match the specific issues and positions from the candidates portfolio that are most appealing to specific voter groups and even individual voters.

Targeted Marketing Campaigns

- Advertisement agencies use quite similar strategies to design more targeted marketing campaigns more quickly
- Marketers are interested in micro-campaigns that are designed specifically for a micro-segment or, in some cases, for specific customers.
  - One can collect channelsurfing data from the set-top box (STBs)
  - The grocery stores offer frequent shopper cards that can be used by the grocer to track purchase habits
  - We can correlate the channel surfing data with retail purchases by the household and insert appropriate commercials to run micro campaigns based on household purchases.

Product Development

- Incoming data could be used to design new products:
  - We read books electronically online or on our favorite handheld devices, giving publishers an opportunity to understand what we read, how many times we read it, and which parts we look at
  - We watch television using a two-way set-top box that can record each channel click
  - We make all of our ordering transactions electronically, giving third parties opportunities to analyze our spending habits by month, by season, by ZIP code, and by tens of thousands of micro-segments

Analytics plays a major role in customizing, personalizing, and changing products based on customer feedback.

Big Data catches the Flu

- In the United States, the Centers for Disease Control and Prevention (CDC) requested that doctors inform them of new flu cases
- The picture is always a week or two out of date:
  - People might feel sick for days but wait before consulting a doctor
  - Relaying the information back to the central organizations took time
  - CDC only tabulated the the numbers once a week
- With a rapidly spreading disease, a two-week delay is way too large!

Big Data catches the Flu

- Google published a remarkable paper in Nature: the authors explained how Google could predict the spread of the winter flu in the United States.
- They achieved this by looking at what people were searching for on the Internet:
  - Google receives more than three billion search queries every day and saves them all
  - It had plenty of data to work with
  - Google took the 50 million most common search terms that Americans type and compared the list with the CDC data on the spread of seasonal flu between 2003 and 2008.

The idea is to identify areas infected by flu virus by what people searched for on the Internet

Big Data catches the Flu

- Others had tried to do this, but no one else had as much data, processing power, and statistical know-how as Google
- Googles software found a combination of 45 search terms
- Like the CDC, Google could tell where the flu had spread
- Unlike the CDC they could tell it in near real time, not with a week or two delay

Google's system proved to be a more useful and timely indicator than government traditional statistics

Big Data catches the Flu

- Google Flu Trends was an enormously successful influenza forecasting service build on Big Data
- However, it failed spectacularly to predict the 2013 flu outbreak
- Data used to predict Ebolas spread in 2014 - 2015 yield also widely inaccurate results

Big Data catches the Flu

Google Flu Trends failed for two reasons:

- Google Flu Trends predictions were based on a commercial search algorithm that frequently changes based on Googles business goals
- The quantity of data does not mean that one can ignore fundamental issues of measurement, construct validity and reliability and dependencies among data

There are some powerful and exciting tools for making predictions from data but they are not magic! You should be skeptical. They require good data and proper internal validation

Outline

Technology Challenges for Big Data

- Traditional approach (more and more complex data bases and powerful computers) has shown its limitation in front of this buzz
- Progress goes through simplification and delegation of processing.
- Various directions
  - Computing equipment to allow faster running
  - Parallelization simplifies nested processing
  - New database queries systems
  - Use of statistical learning algorithms to explore the data

Technology Challenges for Big Data

There are four major technological challenges in managing Big Data:

- Storing Huge Volumes
- Ingesting streams at an extremely fast pace
- Handling a variety of forms and functions of data
- Processing data at huge speeds

Storing Huge Volumes

- No storage machine is big enough to store the relentlessly growing quantity of data
- Data needs to be stored in a large number of smaller inexpensive machines
- There is the inevitable challenge of machine failure which could entail a loss of data stored on it

Storing Huge Volumes

- The first layer of Big Data technology helps store huge volumes of data at an affordable cost, while avoiding the risk of data loss:
  - It distributes data across a large cluster of inexpensive commodity machines
  - It ensures that every piece of data is systematically replicated on multiple machines to guarantee that at least one copy is always available
  - **Apache Hadoop** is the most well-known clustering technology for Big Data:
    - Its data storage system is called Hadoop Distributed File System (HDFS)
    - This system is built on the patterns of Googles Big File system

Ingesting streams at an extremely fast pace

- Handling unpredictable and torrential streams of data
- Some of the data streams may be too large to be stored, but must still be monitored
- The solution lies in creating scalable ingesting systems that can open an unlimited number of channels for receiving data
- These systems can hold data in queues, from which business applications can read and process data at their own pace and convenience

Ingesting streams at an extremely fast pace

The second layer of Big Data technology manages this velocity challenge:

- It uses a special stream-processing engine, where all incoming data is fed into a central queueing system
- From there, a fork-shaped system sends data to batch storage as well as to stream processing directions
- **Apache Spark** is the most popular system for streaming applications

Handling a variety of forms and functions of data

Structuring and access of all varieties of data that comprise Big Data:

- Storing them in traditional flat or relational structures would be too impractical, wasteful and slow
- Accessing and analyzing them requires different capabilities

Handling a variety of forms and functions of data

The third layer of Big Data technology solves this problem by storing the data in non-relational systems:

- **NoSQL** (Not Only SQL) databases
- These databases are optimized for certain tasks such as query processing, or graph processing, document processing, etc
- **HBase and Cassandra** are two of the better known NoSQL databases systems
  - HBase stores each data element separately along with its key identifying information (key-value pair format)
  - Cassandra stores data in a columnar format
- NoSQL languages, such as Pig and Hive, are used to access this data

Processing data at huge speeds

The fourth challenge relates to moving large amounts of data from storage to the processor:

- This would consume enormous network capacity and choke the network
- The alternative is to do just the opposite: to move the processing to where the data is stored.

Processing data at huge speeds

The fourth layer of Big Data technology avoids the choking of the network:

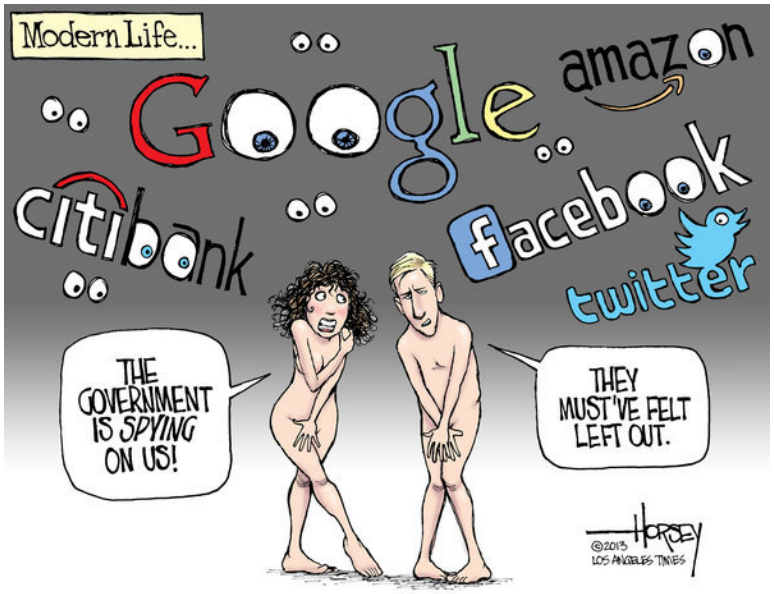
- It distributes the task logic throughout the cluster of machines where the data is stored
- Those machines work, in parallel, on the data assigned to them, respectively
- A follow-up process consolidates the outputs of all the small tasks and delivers the final results
- **MapReduce**, invented by Google, is the best-known technology for parallel processing of distributed Big Data

Technology Challenges for Big Data

- There are many additional technologies to make the task of managing Big Data easier:
  - e.g. technologies to monitor the resource usage and load balancing of the machines in the cluster
- More: "Big Data Algorithms, Techniques and Platforms"
- Once these major technological challenges arising from the 3 Vs of data are met, all traditional analytical and presentation tools, such as machine learning and statistics, can be reliably applied to Big Data
- This course: some of **statistical learning tools**. Additional machine learning tools in the machine learning course

Outline

Big Data and Privacy



A big data issue that is gaining attention and will become more important is individual privacy

What data should the government and organizations be allowed to collect and what safeguards should be in place about how it is used?

Big Data and Privacy

- Data security breaches:
  - Yahoo! security breach that exposed 453,000 unencrypted user names and passwords
  - Predictive models may also uncover private facts not yet shared openly and could also lead to privacy loss:
    - Statisticians at Target created a customer segmentation model that analyzed customer purchase behavior to predict customer life-cycle stages and related micro-segments
    - One of the predictive models was a pregnancy-prediction model that could predict with reasonable accuracy whether the customer making the purchase was expecting a baby
    - The resulting Target campaign reached a house with a girl in high school, whose parents were not informed of her pregnancy...

Big Data and Privacy

- One can identify three different ways to characterize online invasions of privacy:
  1. Uninvited intrusion into a users personal space
     - Online marketing
     - Spam advertising
     - Pop-ups
     - Sponsored sites around the edges of a Web page
  2. The most serious threats are fraudulent e-commerce transactions and identity theft

Big Data and Privacy

- One can identify three different ways to characterize online invasions of privacy:
  3. Personal profiling for commercial advantage
     - Google, Facebook, and Yahoo! combine hundreds or thousands of pieces of data from different sources to understand who you are, where you live, where you go, who your friends are, what you buy, and like
     - This blended information may be used simply to make offers that are likely to appeal to you
     - or for less benign purposes such as knowing whether you engage in risky hobbies and should be charged higher insurance rates

Big Data and Privacy

- Some laws limit the activities of telecommunications companies (e.g., they cant listen to phone calls)
- Few regulations and laws are applicable for the new digital age and are largely non-existent for Internet firms
- These companies privacy policies largely serve their commercial interests rather than protecting individuals privacy

We need laws about individual privacy!

Big Data and Privacy

- U.S. Federal Trade Commission settled a case with Facebook that now requires Facebook to conduct regular audits
- Facebook agreed to submit to the government audits of its privacy practices.
- The company also agreed to obtain explicit approval from users before changing the type of content it makes public
- Similar processes have been put in place at MySpace and Google

Big Data and Privacy

- Research shows that most people have very little understanding and concern about how organizations are using big data [Clemons et al., 2014]
- As individuals understand the potential uses better, their concerns increase quickly
- As companies increasingly use big data analytics on customer data, the public is likely to become more concerned

In practice

- Many of the companies nowadays are doing business cross countries and continents and the differences in privacy laws are considerable and have to be taken into consideration when starting the Big Data initiative

- Big Data consists in a large amount of complex data; it is very difficult for a company to sort this data on privacy levels and apply the according security

Data obfuscation process

- Marketers are interested in customer characteristics that can be provided without Privately Identifiable Information (PII) - that is, the information about the individual that can be used to identify, locate, and contact an individual
- One possibility is to destroy all PII information, which may still provide useful information to a marketer about a group of individuals
- While PII data is destroyed, it may still leave related information that, if joined with obfuscated data, might lead to the individual:
  - if we destroyed the address and phone number but left the location information, someone could use the location information to establish the residential address.

Outline

Big Data Analytics

- Collecting and storing big data creates little value; it is only data infrastructure at this point
- It must be analyzed and the results used by decision makers in order to generate value
- The key to deriving value from big data is the use of analytics
- Big data and analytics are intertwined, but analytics is not new

Big Data Analytics

- Many analytic techniques, such as regression analysis and machine learning, have been available for many years
- The value in analyzing unstructured data such as e-mail and documents has been well understood also
- What is new is the coming together of advances in computer technology and software, new sources of data (e.g., social media), and business opportunity

Data Scientists

- Big data is also creating a high demand for people who can analyze and use big data and it is creating new jobs and changing existing ones: **data scientists**
- The job of data scientists is to discover patterns and relationships that no one else has seen or wondered about, and turn these discoveries into actionable information that creates value for the organization

Data Scientists

To do this requires a rich mixture of skills:

- understand the different types of big data and how they can be stored (e.g., RDBMS, Hadoop)
- analyze it (e.g., regression analysis, social networks)
- write code (e.g., Java, Python, R)
- access data (e.g., SQL, Hive)
- communicate findings to management in business terms (e.g., briefings, reports)

Different Kinds of Analytics

- Three kinds of analytics:
  - Descriptive analytics
  - Predictive analytics
  - Prescriptive analytics
- These differences have implications for the technologies and architectures used for big data analytics
- Some types of analytics are better performed on some platforms than on others

Descriptive analytics

- reporting/OLAP (online analytical processing)
- dashboards/scorecards
- data visualization

- Have been widely used for some time
- Are the core applications of traditional BI (business intelligence)
- Descriptive analytics are backward looking and reveal what has occurred
- New trend: include the findings from predictive analytics, such as forecasts of future sales, on dashboards/scorecards

Predictive analytics

- Suggest what will occur in the future
- Methods and algorithms for predictive analytics such as regression analysis, machine learning, and neural networks have existed for some time
- Marketing is the target for many predictive analytics applications; here, the goal is to better understand customers and their needs and preferences
- Some people also refer to exploratory or discovery analytics.

Prescriptive analytics

- Suggests what to do
- Prescriptive analytics can identify optimal solutions, often for the allocation of scarce resources
- Has been researched in academia for a long time
- Now finding wider use in practice:
  - e.g. the use of mathematical programming for revenue management is increasingly common for organizations that have perishable goods such as rental cars, hotel rooms, and airline seats

Different Kinds of Analytics

Organizations typically move from descriptive to predictive to prescriptive analytics:

What happened? $\rightarrow$ Why did it happen? $\rightarrow$ What will happen? $\rightarrow$ How can we make it happen?

Benefits of Data Analytics

Research shows the benefits of using data and analytics in decision making:

- One study of 179 large publicly traded firms found that companies that have adopted data-driven decision making have output and productivity that is 5% to 6% higher than that of other firms
- In 2010, the MIT Sloan Management Review, in collaboration with the IBM Institute for Business Value, surveyed a global sample of nearly 3,000 executives
  - top-performing organizations use analytics five times more than do lower performers
  - 37% of the respondents believe that analytics creates a competitive advantage

Benefits of Data Analytics

Research shows the benefits of using data and analytics in decision making:

- A follow-up study in 2011 found that the percentage of respondents who reported that the use of analytics was creating a competitive advantage rose to 58%
- Although these studies do not focus exclusively on big data, they do show the positive relationships between data-driven decision making, organizational performance, and competitive position

References

- Anil K. Maheshwari. *Big Data: Made Accessible.* (2017)
- Hugh J. Watson. *Tutorial: Big Data Analytics: Concepts, Technologies, and Applications.* Communications of the Association for Information Systems (2014)