ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Introduction to
Data Science
(IT4142E)

Contents

Introduction

Goals of data science

Some questions

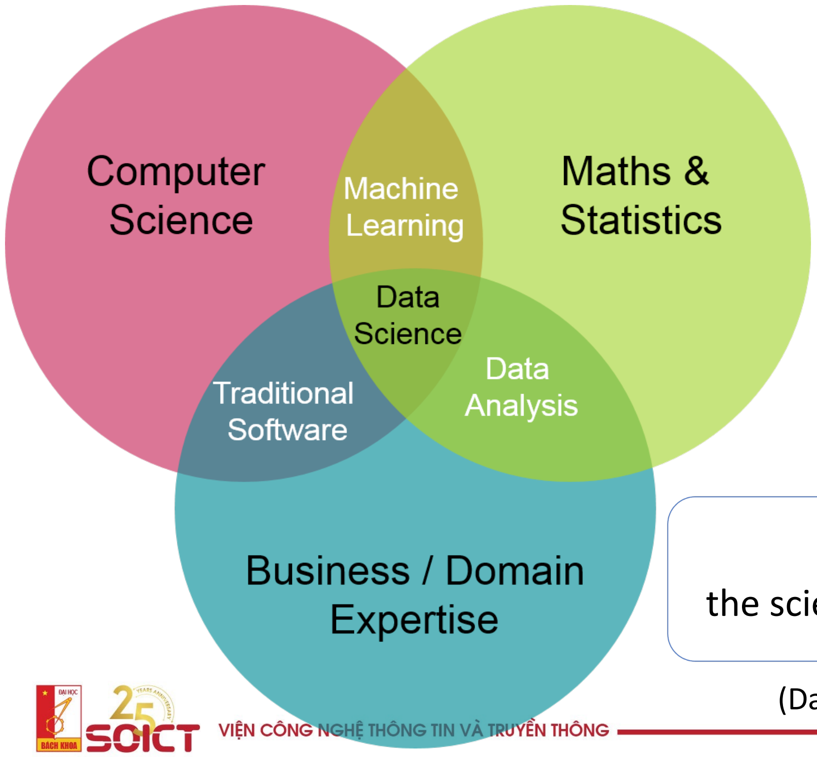
- Some questions a decision-maker might wonder:
  - What is the evolution of my stores' turnover, by month and by store?
  - Based on what client X is buying, which age range does he/she most likely …
  - What … mmend them, …
  - What … his info into accou…
  - If I ac… o reimburse me (at …

Let the data speak

- These questions are:
  - Specific
  - Sometimes, embedded in one another
  - Unpredictable

→**A simple planned report is not enough!**

What is Data Science?



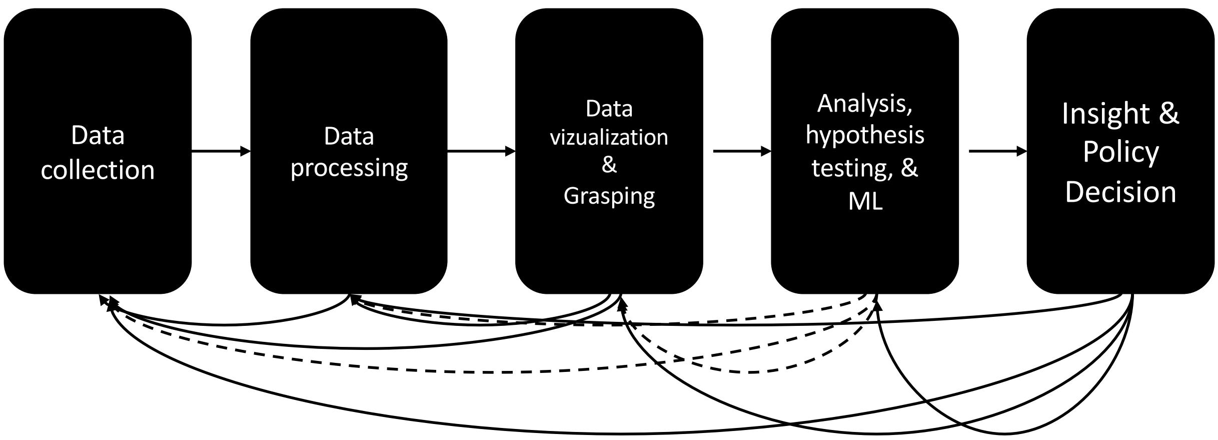
Data science is the science of *learning from data*.

(David Donoho, Stanford University)

Goals of Data Science

- The final goals of data science might be classified into
  - Description
  - Prediction
- In order to achieve these goals, several tasks are required:
  - Data scraping
  - Data cleaning, pre-processing and integration
  - Machine learning
  - Visualization
- Data science may apply to any kind of data
  - Raw data (numbers)
  - Text analysis
  - Image and video analysis
  - Graph analysis

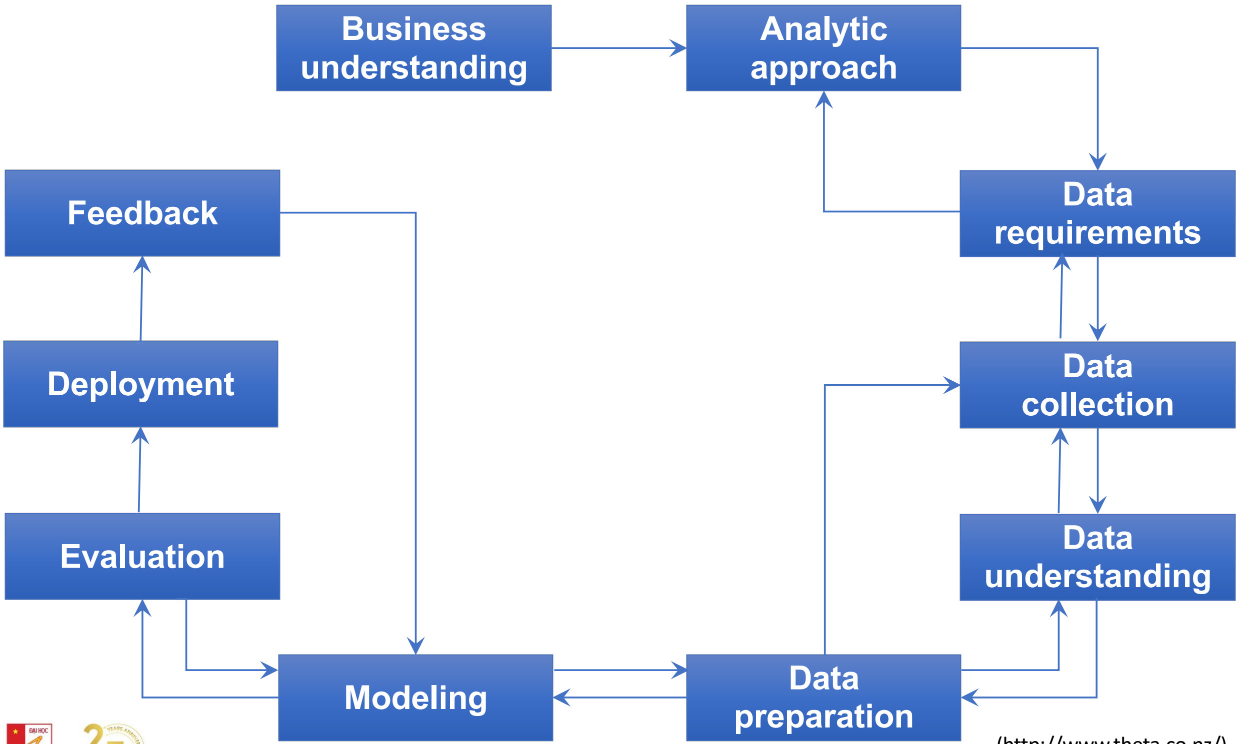
Methodology: insight-driven



70-90% of the whole process

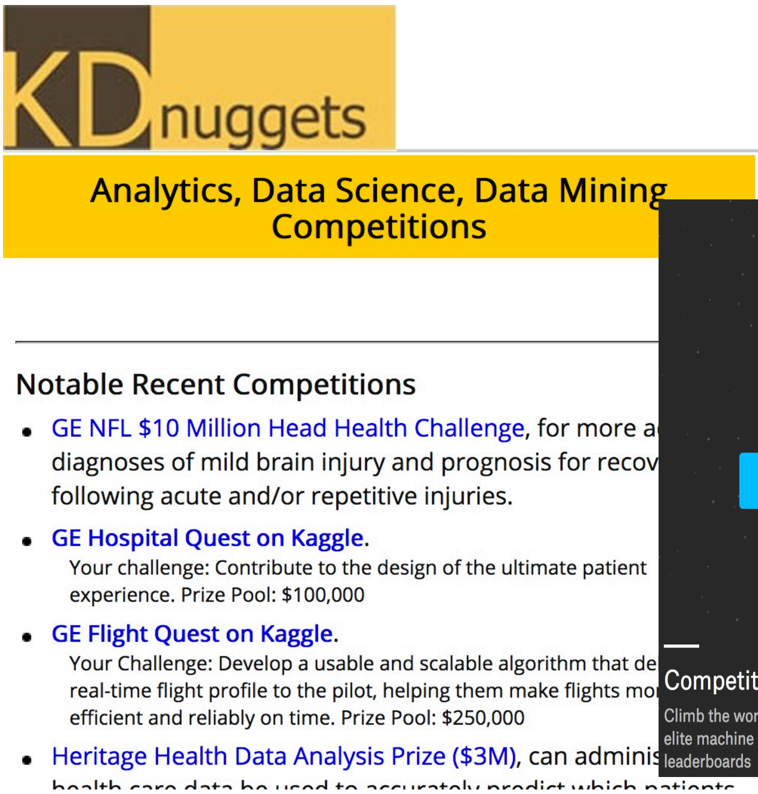
(John Dickerson, University of Maryland)

Methodology: product-driven

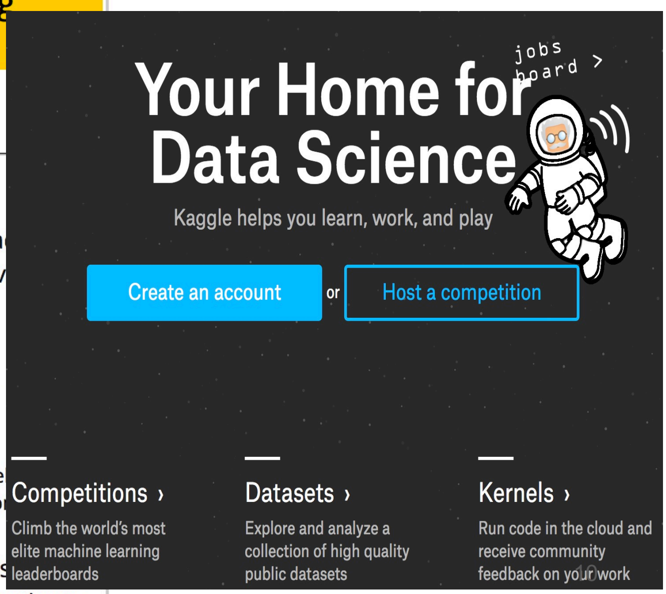


(http://www.theta.co.nz/)

Some online platforms for DS competitions



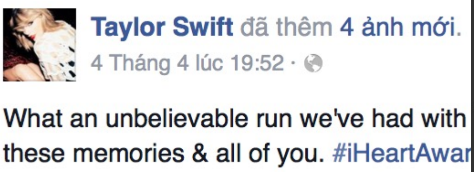
kaggle

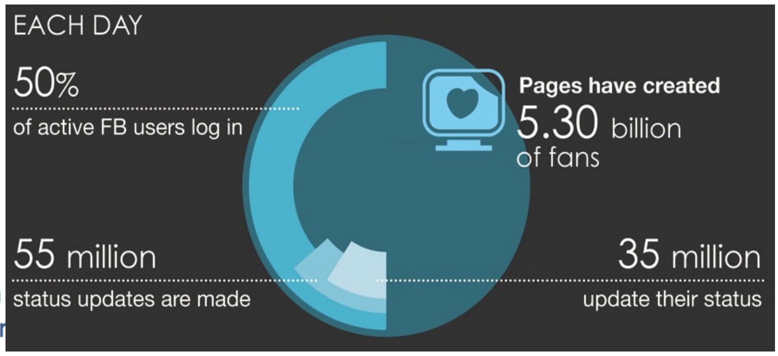


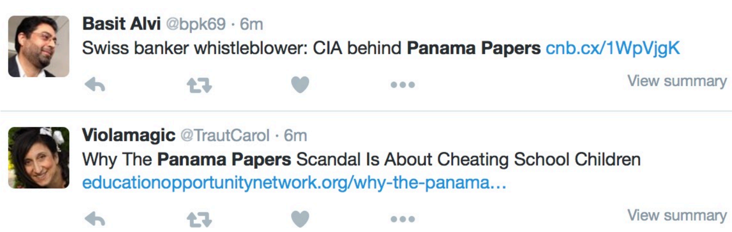
Introduction

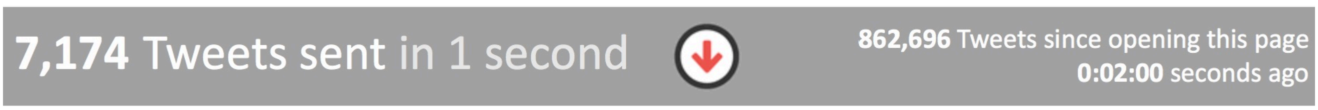
Where is the data?

Where is the data? Social networks

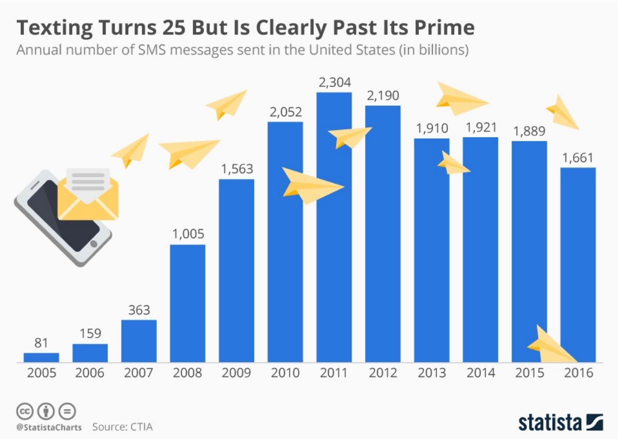






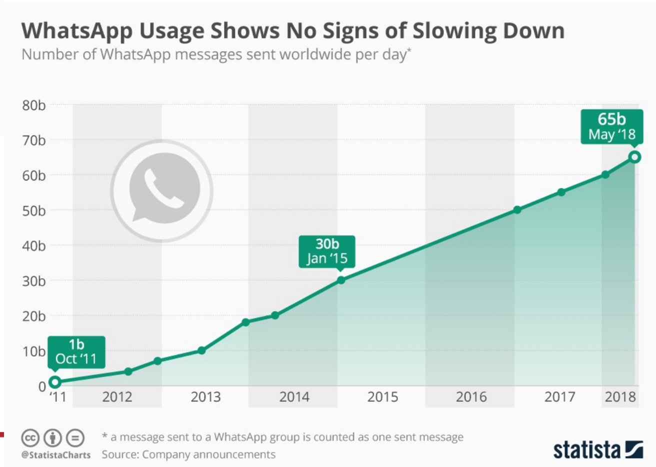


Where is the data? Mobile messages



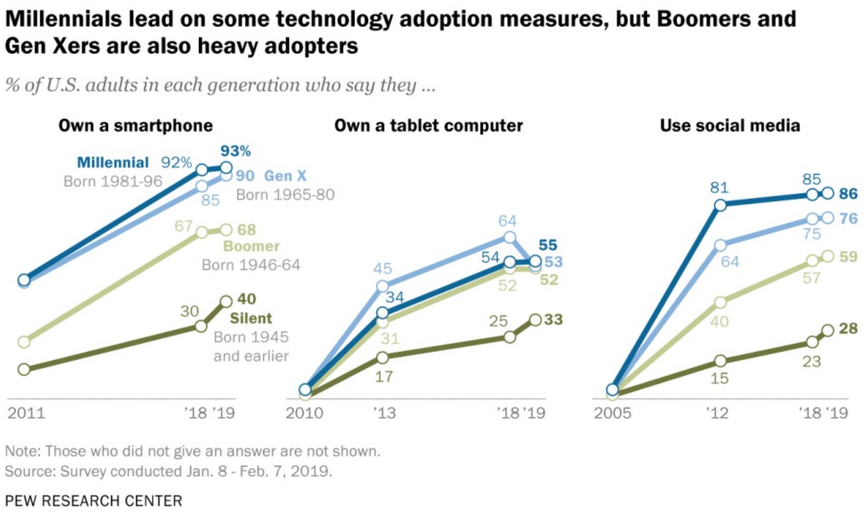
Rise and fall of SMS

Rise of messaging apps



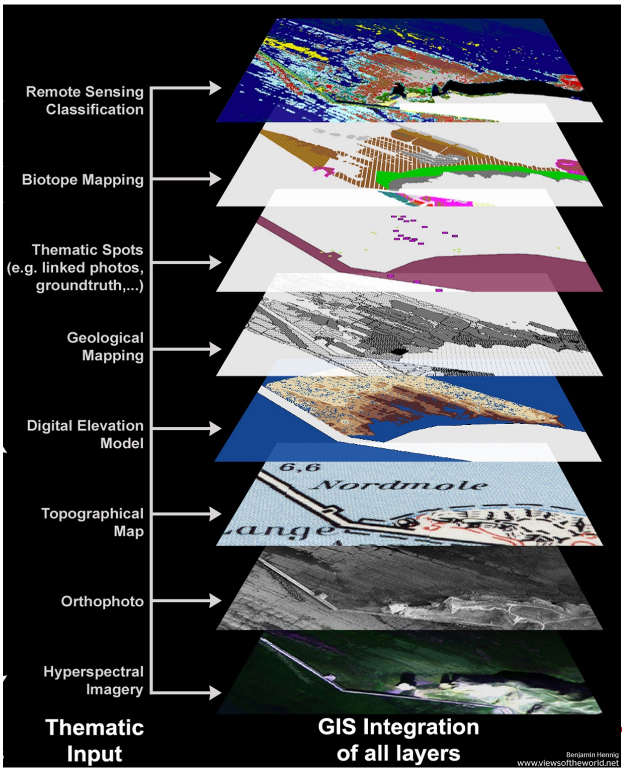
Where is the data? Internet

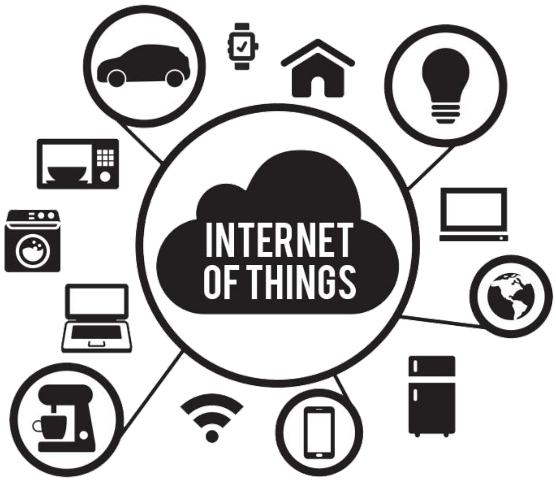
- In the US:



- https://www.internetlivestats.com

Where is the data? And more

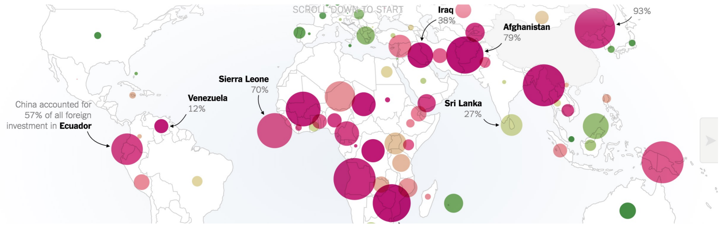


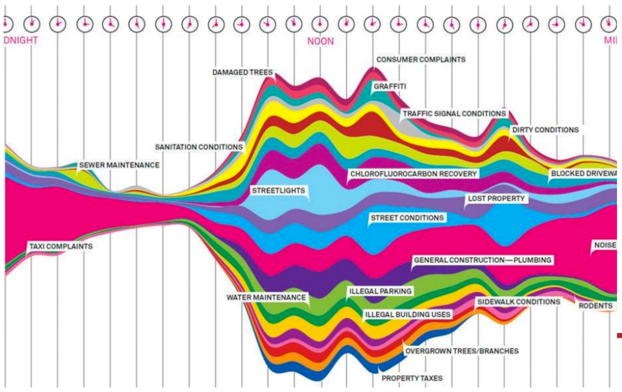


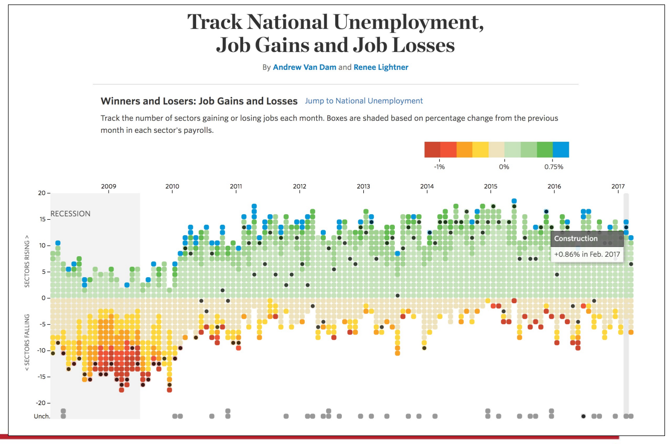
Introduction

What can we do with the data?

What can we do with the data?
Data description through visualization



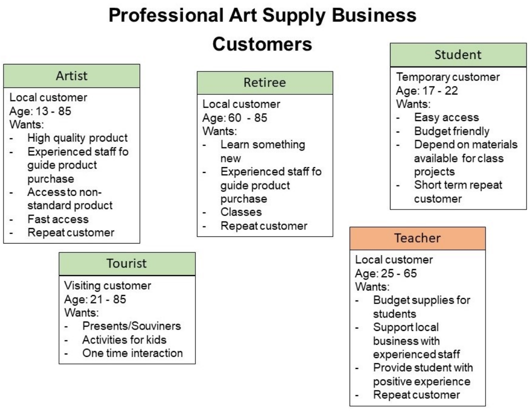


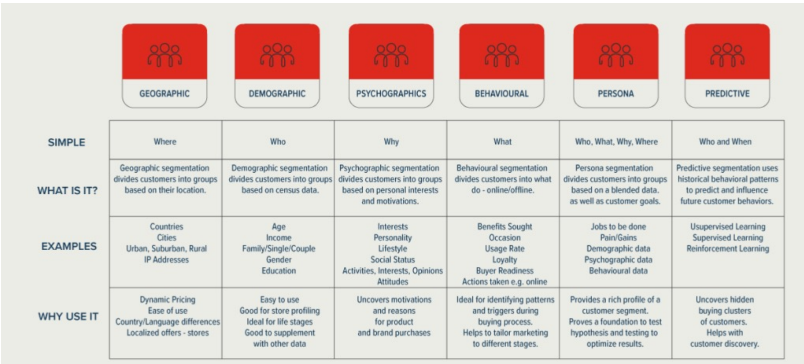


Data description

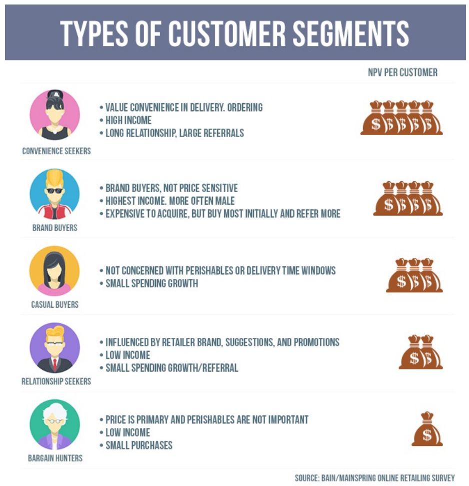
- Data description consists in summarizing the data in an “understable” way, either:
  - Through **exploratory data analysis**
    - Mostly descriptive statistics such as average, standard deviation, median, mode,…
  - Through **data visualization**

What can we do with the data?
Customer segmentation





	GEOGRAPHIC	DEMOGRAPHIC	PSYCHOGRAPHICS	BEHAVIOURAL	PERSONA	PREDICTIVE
SIMPLE	Where	Who	Why	What	Who, What, Why, Where	Who and When
WHAT IS IT?	Geographic segmentation divides customers into groups based on their location.	Demographic segmentation divides customers into groups based on census data.	Psychographic segmentation divides customers into groups based on personal interests and motivations.	Behavioural segmentation divides customers into what do - online/offline.	Persona segmentation divides customers into groups based on a blended data. as well as customer goals.	Predictive segmentation uses historical behavioral patterns to predict and influence future customer behaviors.
EXAMPLES	Countries Cities Urban, Suburban, Rural IP Addresses	Age Income Family/Single/Couple Gender Education	Interests Personality Lifestyle Social Status Activities, Interests, Opinions Attitudes	Benefits Sought Occasion Usage Rate Loyalty Buyer Readiness Actions taken e.g. online	Jobs to be done Pain/Gains Demographic data Psychographic data Behavioural data	Usupervised Learning Supervised Learning Reinforcement Learning
WHY USE IT	Dynamic Pricing Ease of use Country/Language differences Localized offers - stores	Easy to use Good for store profiling Ideal for life stages Good to supplement with other data	Uncovers motivations and reasons for product and brand purchases	Ideal for identifying patterns and triggers during buying process. Helps to tailor marketing to different stages.	Provides a rich profile of a customer segment. Proves a foundation to test hypothesis and testing to optimize results.	Uncovers hidden buying clusters of customers. Helps with customer discovery.

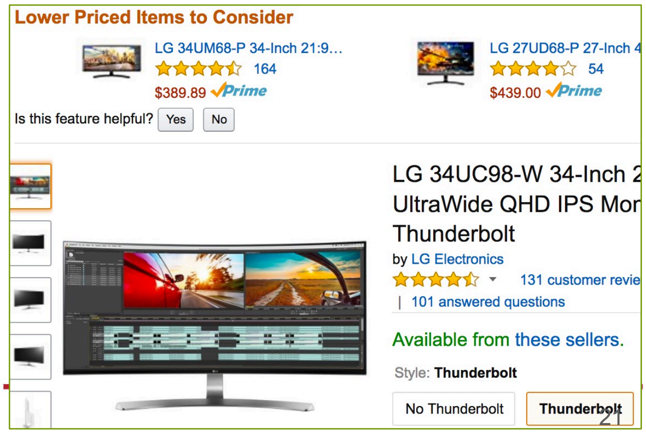


Data segmentation

- Data segmentation consists in grouping the similar records into homogeneous groups (called **clusters**)
  - Records in a group have similar attribute values
  - Technically, the goal is to lean a “new” attribute (group#) from the record’s attributes
  - **Unsupervised learning** methods can be used: see Chapter 7

What can we do with the data?
Amazon's recommendation (association)

"The company reported a **29% sales increase** to $12.83 billion during its second fiscal quarter, up from $9.9 billion during the same time last year."
– Fortune, July 30, 2012

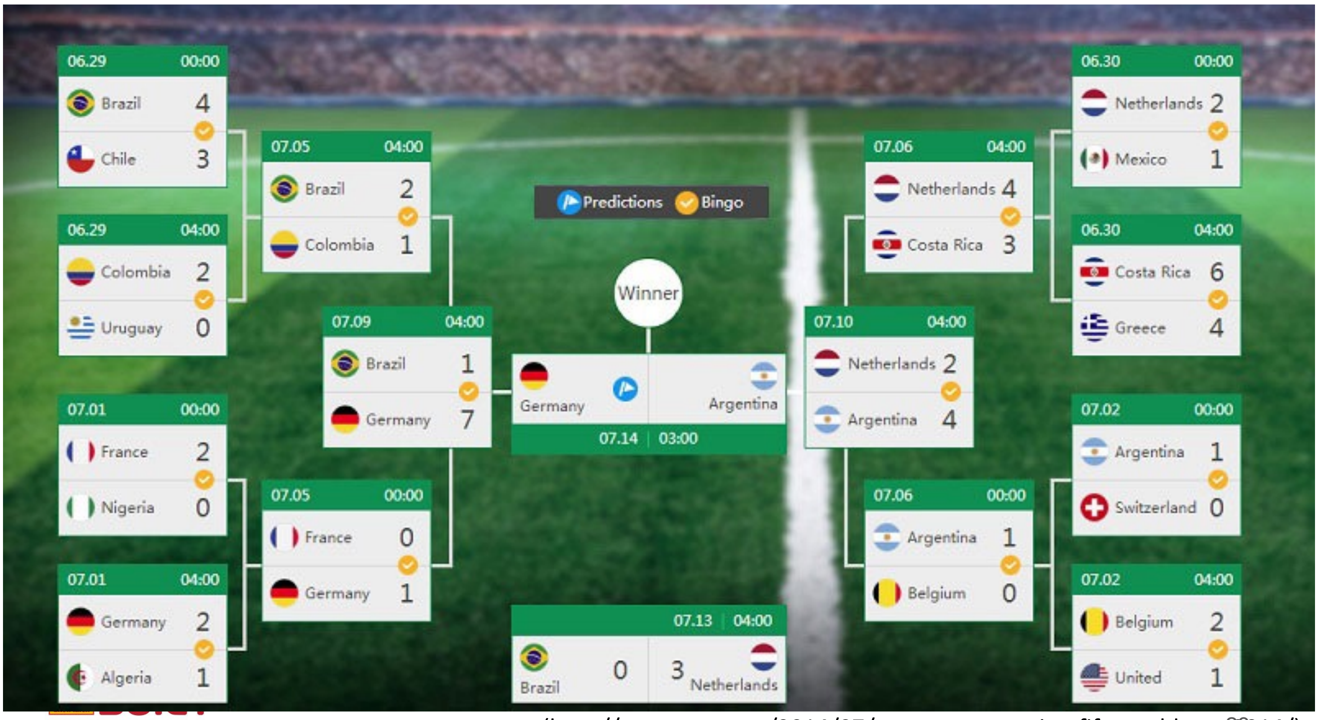


Association rules

- Association consists in discovering association rules between records, according to pre-defined criteria
  - *E.g.* the items that are often bought during one single transaction
  - Technically, the goal is to lean a "new" information (association rules) from the record's attributes
  - **Unsupervised** learning methods can be used: see Chapter 7

What can we do with the data?

FIFA predictions (2014)



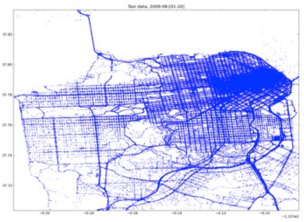
Accuracy ~93%.

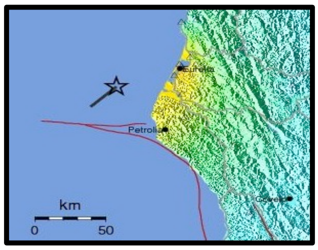
(http://yourstory.com/2014/07/germany-argentina-fifa-world-cup-2014/)

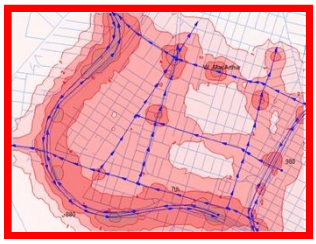
Prediction

- Prediction consists in either:
  - predicting or estimating the values of an attribute for a set or records
    - This attribute is known for other records
    - This knowledge is used to predict this attribute’s values on our set of records
    - **Supervised** learning methods can be used

What can we do with the data?
Much more!!!

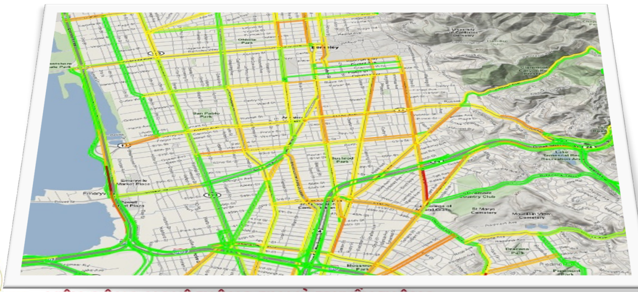






Crowdsourcing + physical modeling + sensing + data assimilation

to produce:



(Alex Bayen, UC Berkeley)

Big data

What is it?

Big data – in 2008

http://www.wired.com/wired/issue/16-07

September 2008

Big data – in 2014

THE AVERAGE PERSON TODAY PROCESSES MORE DATA IN A SINGLE DAY THAN A PERSON IN THE 1500'S DID IN AN ENTIRE LIFETIME

LOOK TO THE LEFT, and you see Times Square at dusk. Look to the right, and you see the same location at midmorning. Internationally acclaimed photographer Stephen Wilkes's time-altering image of New York's Times Square is part of his body of work titled *Day to Night*.

The image was created by blending more than 1,400 separate photos taken over the course of 15 hours—a meticulous process that took him nearly three months.

PHOTO: STEPHEN WILKES

Big data – today

The amount of information generated during the first day of a baby's life today is equivalent to 70 times the information contained in the Library of Congress

Big data – today: some numbers

Big data

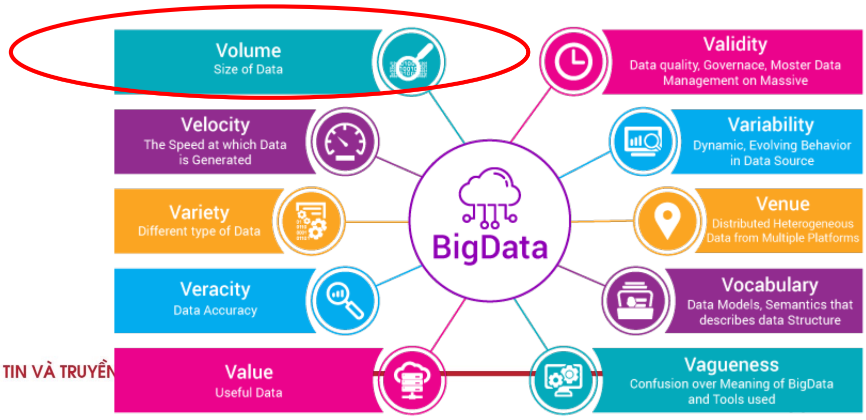
Challenges

The 10 Vs of Big data

[Source: houseofbots.com]

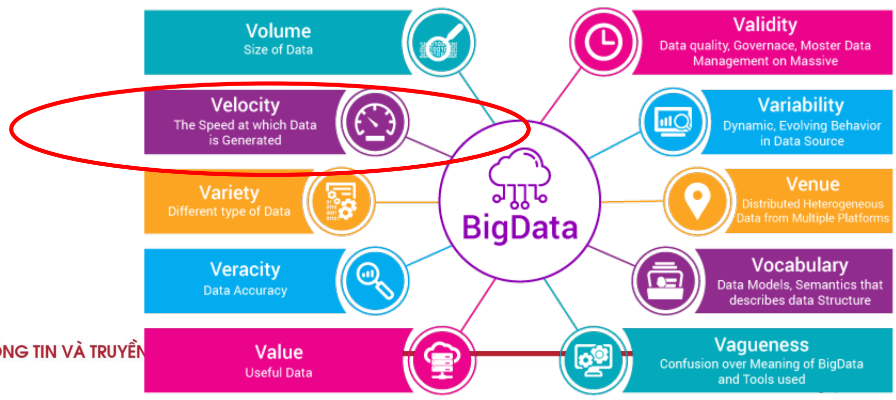
The 10 Vs of Big data: Volume

- Volume is probably the best known characteristic of big data
- More than 90% of all today's data was created in the past 2 years
- Poses challenges in terms of:
  - Exploratory Data Analysis (see Chapter 4)
  - Data visualization (see Chapter 5)



The 10 Vs of Big data: Velocity

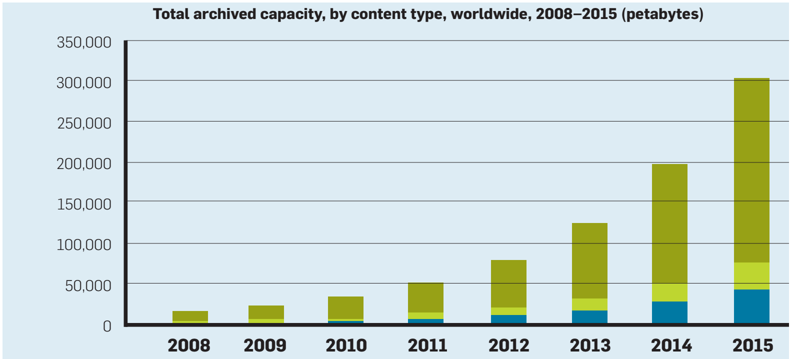
- Velocity refers to the speed at which data is being generated, produced, created, or refreshed
  - It is ever-increasing, contributing to exponentional growth in the data volume!
  - It poses several challenges in terms of data integration (see Chapter 3)



The 10 Vs of Big data: Variety

- Variety refers to the different kinds of data one has to handle:
  - **Structured** data: from OLTP datasets of Excel files for instance
  - **Unstructured** data increases extremely fast: texts, images, tags, links, likes, emotions, …

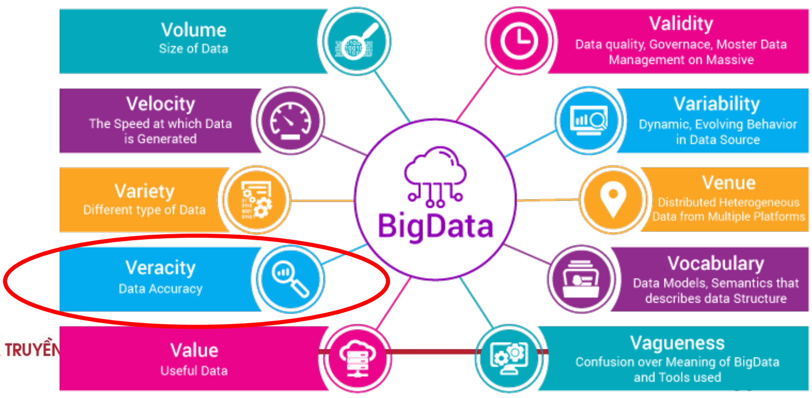
(Vasant Dhar, CACM, 2013)



	2008	2009	2010	2011	2012	2013	2014	2015
Unstructured	11,430	16,737	25,127	39,237	59,600	92,536	147,885	226,716
Database	1,952	2,782	4,065	6,179	9,140	13,824	21,532	32,188
Email	1,652	2,552	4,025	6,575	10,411	16,796	27,817	44,091

The 10 Vs of Big data: Veracity

- Veracity: does the data reflect the reality? how accurate or truthful is it?
  - **Not everything that is written on the internet is TRUE!!!**
  - Hence, the need to check the data sources' quality (see Chapter 2)
    - Almost an ethical issue
  - Noises, missing values, mistakes, biases,…
  - →Challenging for analysis

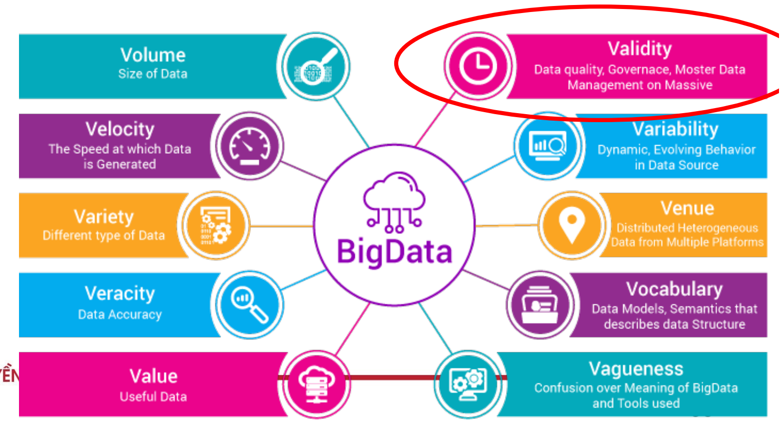


The 10 Vs of Big data: Value

- When there is so much data, it obviously poses the question of data value
  - And hence, one has to select / pre-process / integrate only the relevant data (see Chapter 2)

The 10 Vs of Big data: Validity

- When there is so much data, it of course poses the question of data validity
  - And hence, one has to check the quality of the data
    - Check its coherence with other sources of data
    - Remove outliers
  - This is pre-processing, led before integrating it for data analysis



The 10 Vs of Big data: Venue

- Venue in big data refers to the multiplicity of data sources (*e.g.* Excel files, OLTP databases, …)
  - Hence the need for data integration (see Chapter 3)

The 10 Vs of Big data: Variability

- Variability in big data refers to two things
  - The possible evolutions in the structure of the data sources
  - The different velocities at which these data sources are refreshed
  - Poses serious issues for data integration

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN

The 10 Vs of Big data: Vocabulary

- Vocabulary refers to bringing data models / semantics (knowledge, e.g. ontologies) into the data to structure / explain it
  - See the course on AI

The 10 Vs of Big data: Vagueness

- Vagueness might refer to:
  - Communication issue between provider and customer
  - Difficulty for a non-specialist to interpret the analysis output
    - *E.g.* difference between correlation and causality

More additional challenges

- The interactions or **correlations** hidden in data might be really huge
- Real problems often have extremely **high dimensions** (large number of variables)
  - Bicycle runs: 2 dimensions (a road)
  - We live in 4 dimensions
  - But an image 1024x1024: ~1 million dimensions
  - Text collections: million dimensions
  - Recommenders' system: billion dimensions (items/products)

→ The curse of dimensionality

Dữ liệu dù thu thập được lớn đến đâu thì cũng là **quá nhỏ** so với không gian của chúng

Ethical issues

- Privacy
  - Breach of privacy, collection of data without informed consent
- Security
  - The ease of stealing, including identity theft, the stealing of national security information
- Commercial exploitation
  - Commercial mining of information; targeting for commercial gain

- Issue of Power and politics
  - The use of data to perpetuate particular views, ideologies, propaganda
- Issue of Truth
  - Rumors, hoaxes, fake news
  - Bias introduced by social networks' recommender systems
- Issue of social justice
  - Information is overwhelmingly skewed towards certain groups and leaves others out of the 'digital revolution'

What is a data scientist?

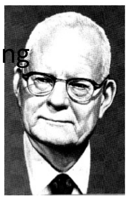
Data Science - early days

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

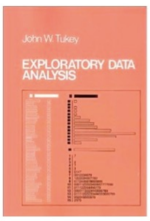
W.E. Demming



1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"

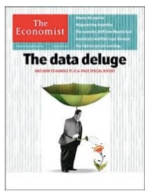


1989: "Business Intelligence"

Howard Dresner

1997: "Machine Learning"

2010: "The Data Deluge"

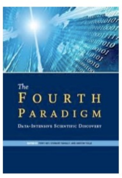


2009: "The Unreasonable Effectiveness of Data"

1996: Google

Google

2007: "The Fourth Paradigm"



(John Canny, UC Berkeley)

The rise of Data Science - 2009

I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?

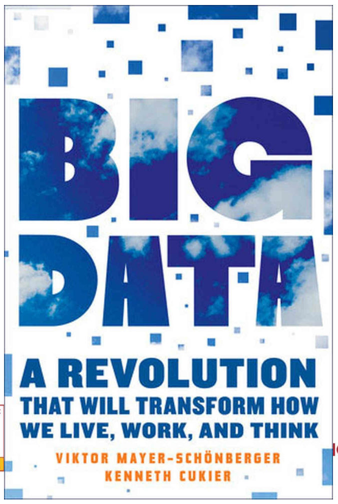
- Hal Varian, Google's Chief Economist, 2009

"The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary school kids, for high school kids, for college kids. **Because now we really do have essentially free and ubiquitous data.**"
- Hal Varian, Google's Chief Economist, 2009

Data scientist - nowadays

Data Scientist: The Sexiest Job of the 21st Century

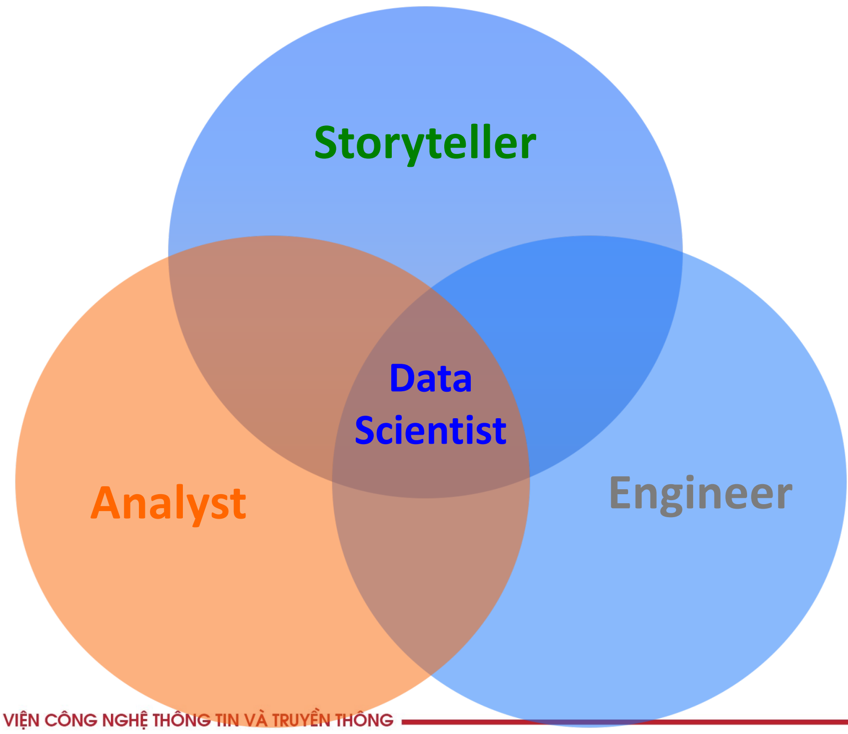
by Thomas H. Davenport and D.J. Patil



Skillset

(source: http://datasciencedojo.com/)

Roles / talents of a data scientist



Further reading

- "Job Comparison – Data Scientist vs Data Engineer vs Statistician" https://www.analyticsvidhya.com/blog/2015/10/job-comparison-data-scientist-data-engineer-statistician/
- Big Data Landscape 3.0 http://mattturck.com/big-data-landscape-2016-v18-final/
- Ten Lessons Learned from Building (real-life impactful) Machine Learning Systems http://technocalifornia.blogspot.com/2014/12/ten-lessons-learned-from-building-real.html

References

- John Dickerson. *Lectures on Introduction to Data Science*. University of Maryland, 2017.
- Longbing Cao. Data science: a comprehensive overview. *ACM Computing Surveys (CSUR)*, *50*(3), 43, 2017.
- Longbing Cao. Data science: nature and pitfalls. *IEEE Intelligent Systems*, *31*(5), 66-75, 2016.
- David Donoho. "50 years of Data Science." In *Princeton NJ, Tukey Centennial Workshop*. 2015.
- L. Duan, Y. Xiong. Big data analytics and business analytics. Journal of Management Analytics, vol 2 (2), pp 1-21, 2015.
- X. Wu, X. Zhu, G. Wu, W. Ding. Data mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, vol 26 (1), pp 97-107, 2014.
- Rafael Irizarry & Verena Kaynig-Fittau. *Lectures on Data Science*. Harvard Univ., 2014.
- John Canny. *Lectures on Introduction to Data Science*. University of California, Berkeley, 2014.
- Vasant Dhar. Data Science and Prediction. *Communication of the ACM*, vol 56 (12), pp 64-73, 2013.
- Michael Perrone. *What is Watson – an overview*. 2011.

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attentions!

soict.hust.edu.vn/ fb.com/groups/soict