ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

IT4142E
Introduction to Data Science

Chapter 1: overview

Lecturer:
Muriel VISANI: murielv@soict.hust.edu.vn

Acknowledgements:
Khoat Than

Department of Information Systems
School of Information and Communication Technology - HUST

Contents of the course

- Chapter 1: Overview
- Chapter 2: Data scraping
- Chapter 3: Data cleaning, pre-processing and integration
- Chapter 4: Introduction to Exploratory Data Analysis
- Chapter 5: Introduction to Data visualization
- Chapter 6: Introduction to Machine Learning
  - Performance evaluation
- Chapter 7: Introduction to Big Data Analysis
- Chapter 8: Applications to Image and Video Analysis

Contents

- **Introduction to Data Science**
  - Introduction
    - Goals of data science
    - Where is the data?
    - What can we do with the data?
  - Big data
    - What is it?
    - Challenges
  - What is a data scientist?

Goals of this chapter

Goal	Description of the goal
M1	**Understand and be able to design and manage the systems which are based on Data Science (DS)**
M1.1	Identify and understand the components of the systems based on DS
M2	**Identify and manage the opportunities from DS to boost the existing organizations, or develop new organizations**
M2.2	Identify the (possible) impacts of Data Science on their organizations

Introduction

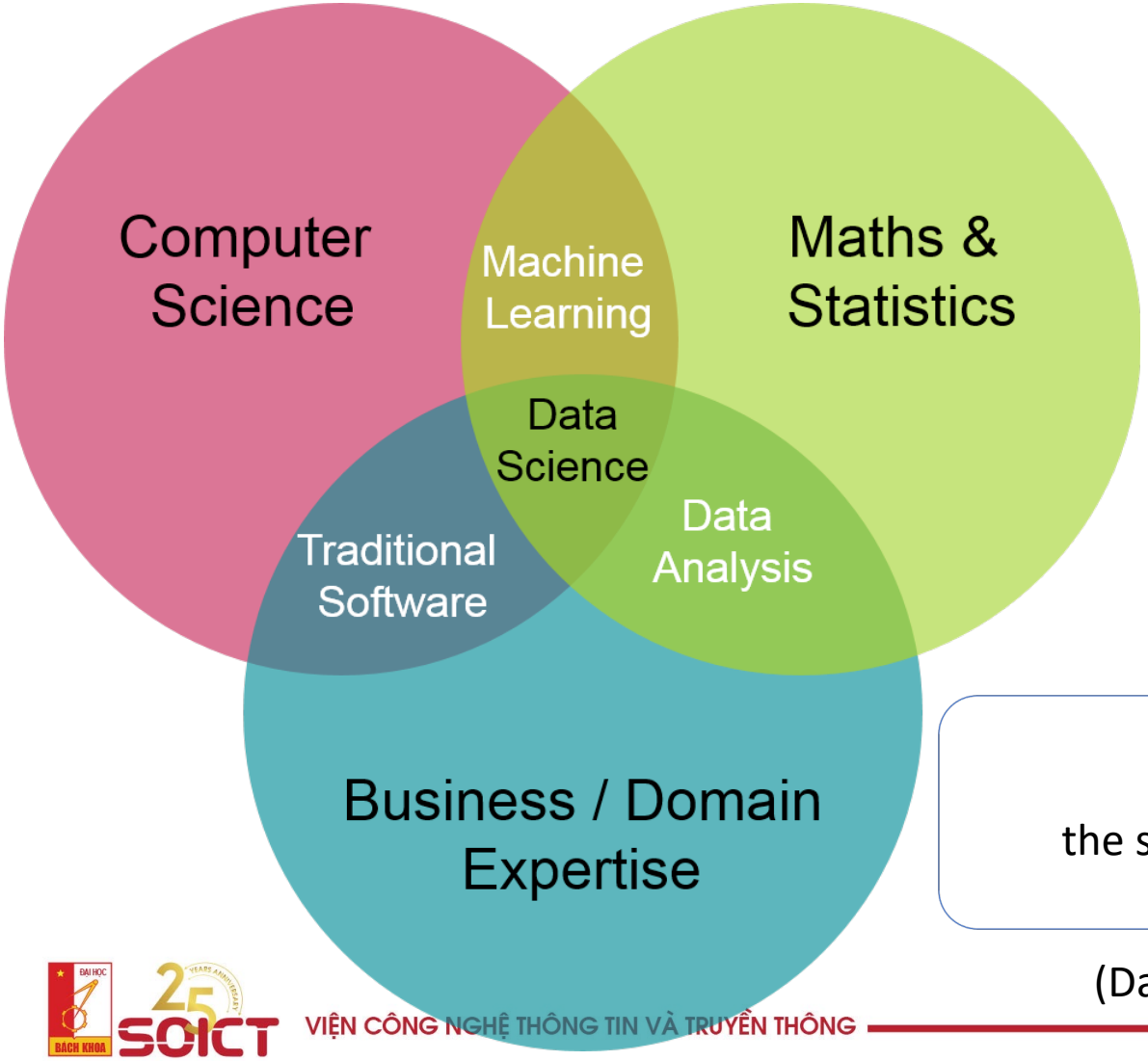
Goals of data science

Some questions

- Some questions a decision-maker might wonder:
  - Wha… by store?
  - Wha… ommend them
  - Wha… this info into acco…
  - If I a… to reimburse me (at least likely enough)?
  - Based on what client X is buying, which age range does he/she most likely belong to?
- These questions are:
  - Specific
  - Sometimes, embedded in one another
  - Unpredictable
  - → A simple planned report is not enough!

Let the data speak

What is Data Science?



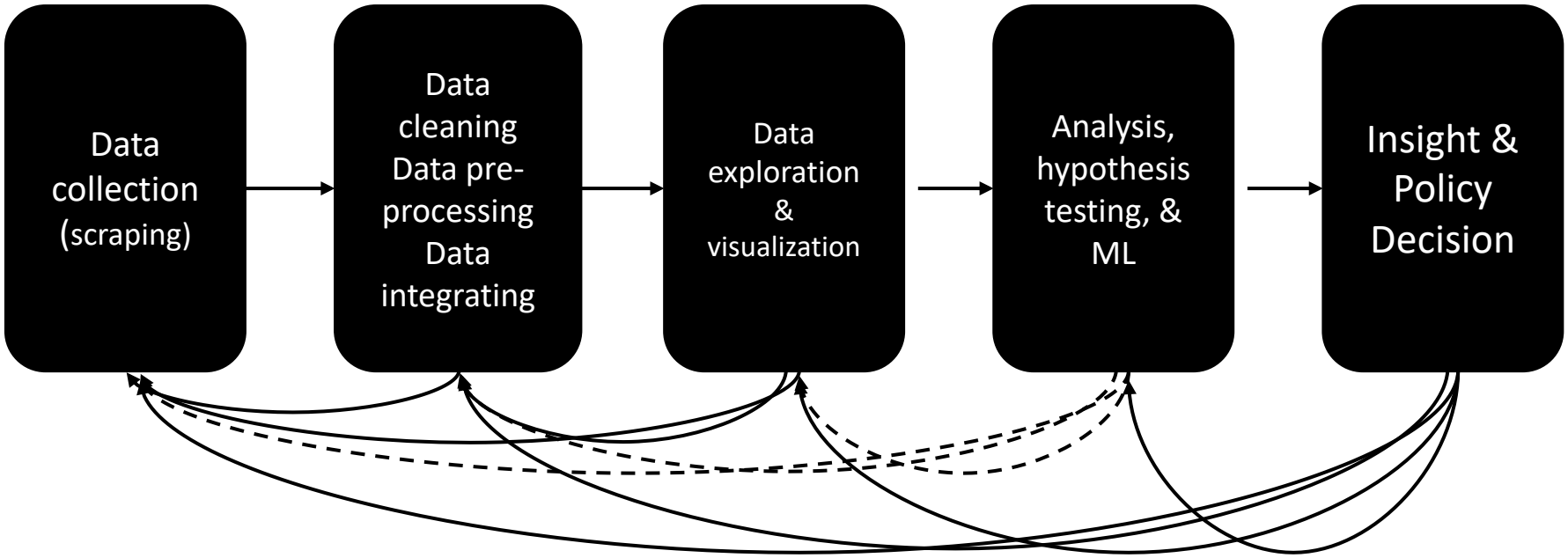
Data science is
the science of *learning from data*.

(David Donoho, Stanford University)

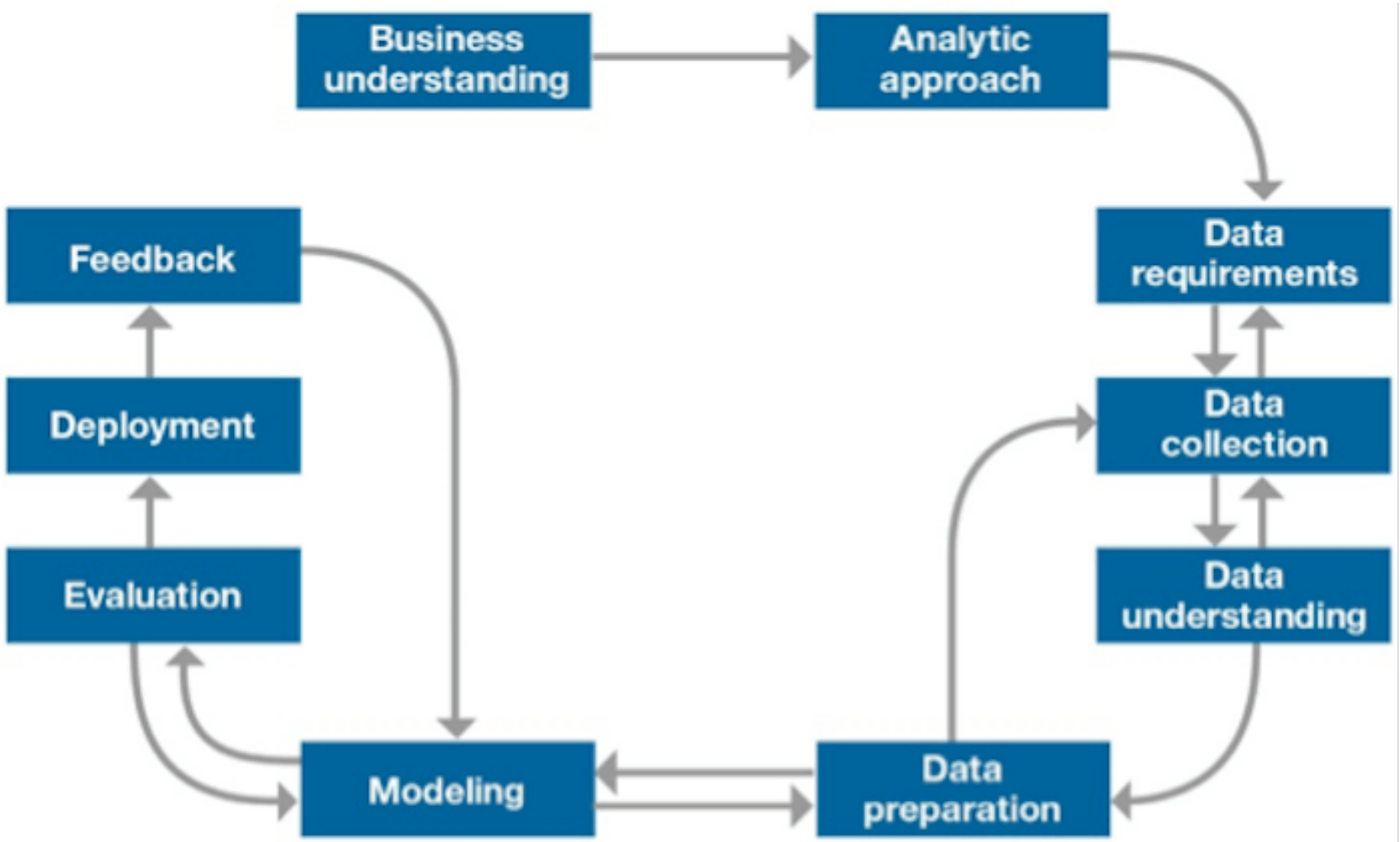
Goals of Data Science

- The final goals of data science might be classified into
  - Description
  - Segmentation
  - Association
  - Prediction
- In order o 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

DS methodology: insight-driven



DS methodology: product-driven



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

Some online platforms for DS competitions

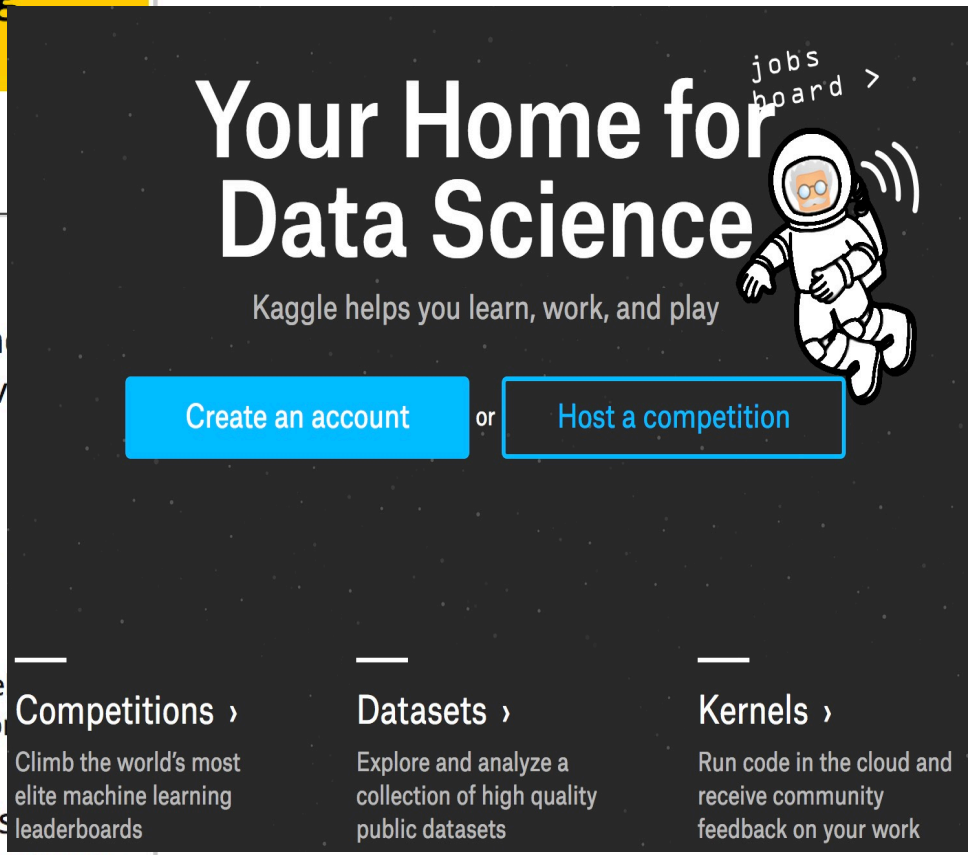
KDnuggets

Analytics, Data Science, Data Mining Competitions

Notable Recent Competitions

- GE NFL $10 Million Head Health Challenge, for more a
  diagnoses of mild brain injury and prognosis for recov
  following acute and/or repetitive injuries.
- GE Hospital Quest on Kaggle.
  Your challenge: Contribute to the design of the ultimate patient experience. Prize Pool: $100,000
- GE Flight Quest on Kaggle.
  Your Challenge: Develop a usable and scalable algorithm that de
  real-time flight profile to the pilot, helping them make flights mo
  efficient and reliably on time. Prize Pool: $250,000
- Heritage Health Data Analysis Prize ($3M), can adminis

kaggle

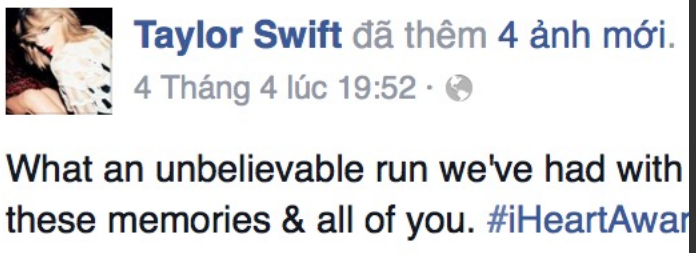


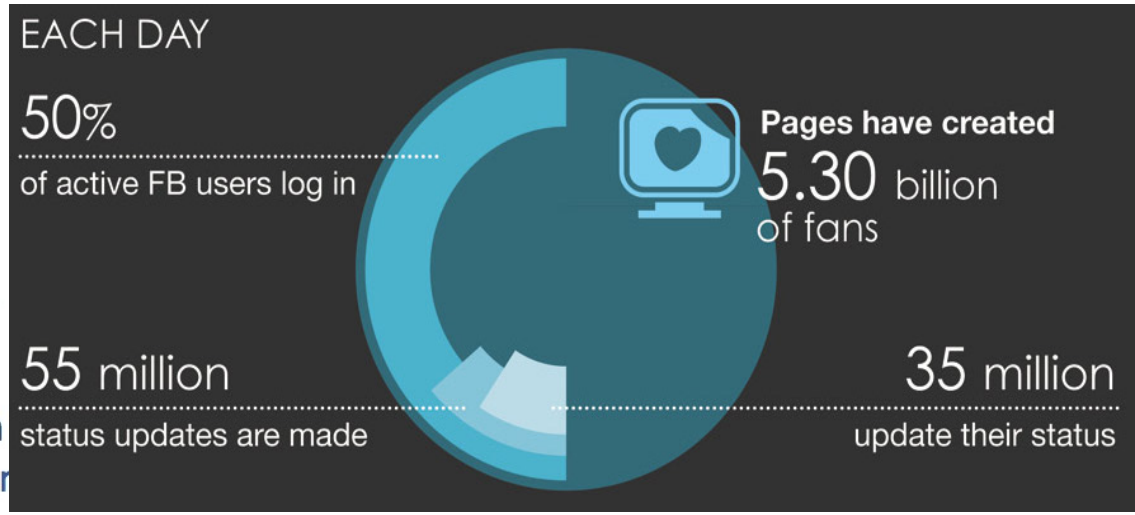
Introduction

Where is the data?

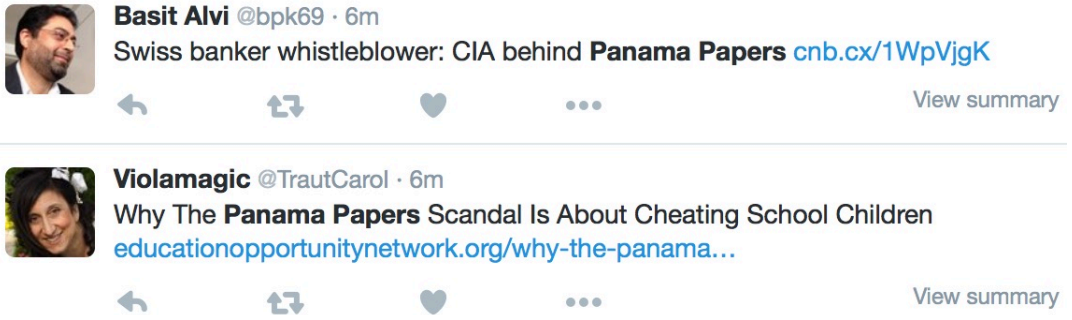
Where is the data? Social networks

facebook





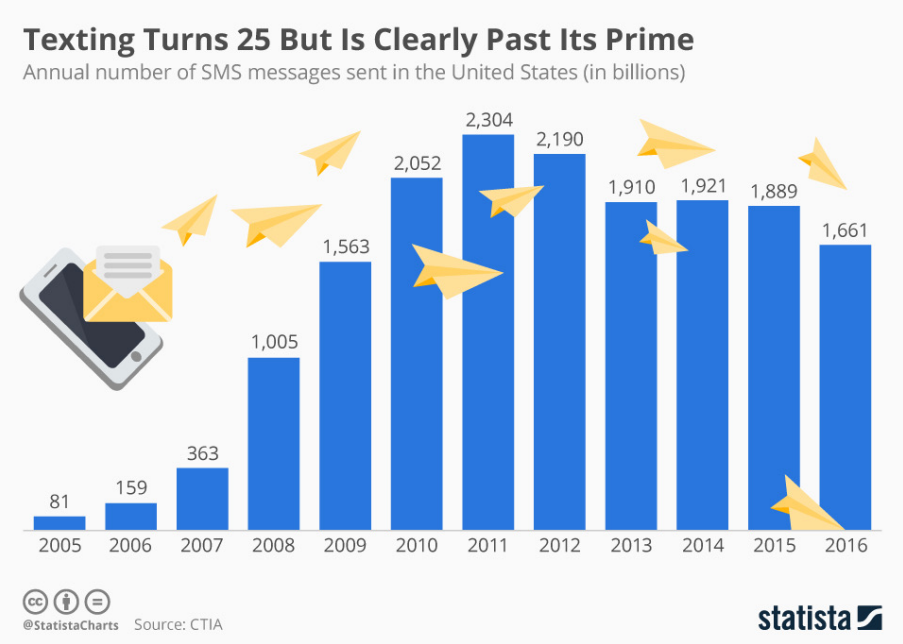
twitter



7,174 Tweets sent in 1 second

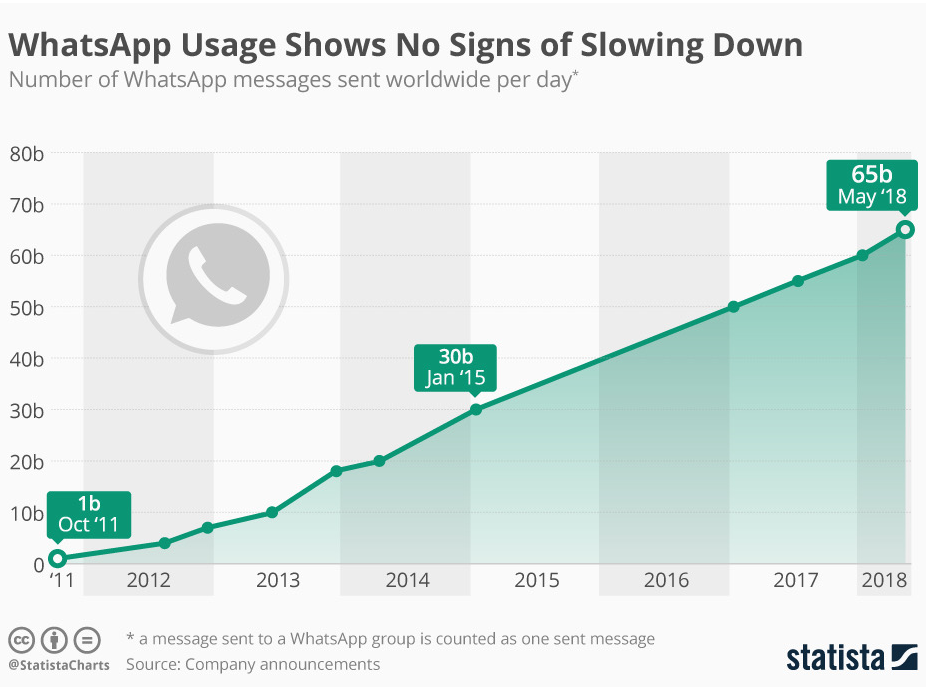
862,696 Tweets since opening this page
0:02:00 seconds ago

Where is the data? Mobile messages



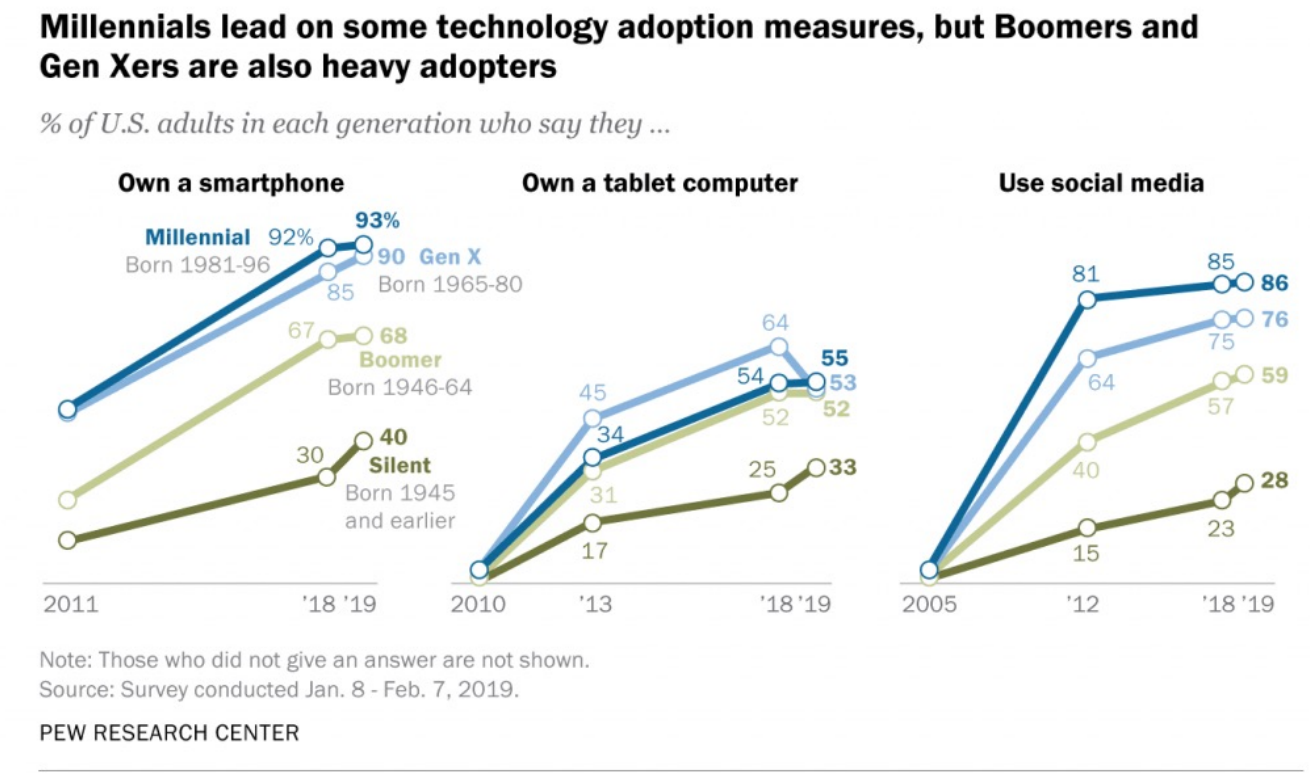
Rise and fall of SMS

Rise of messaging apps



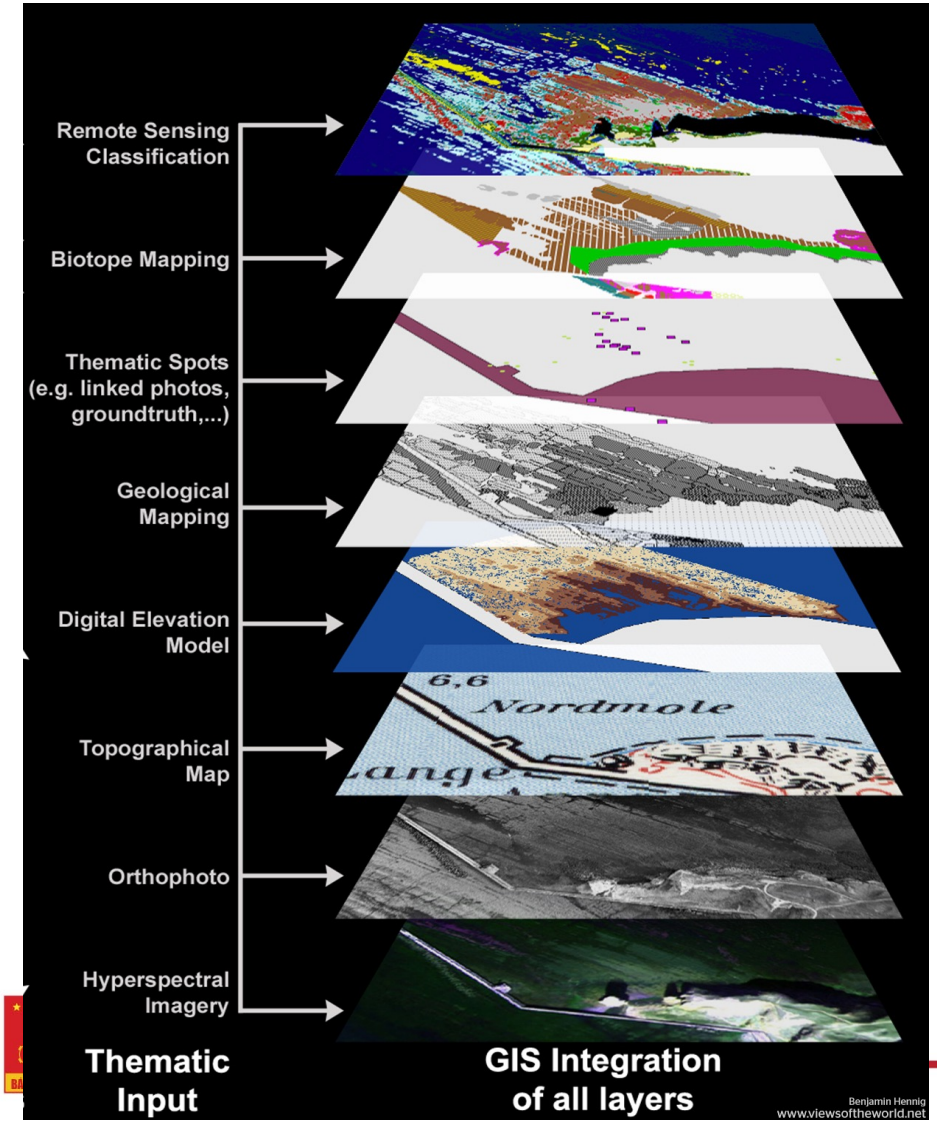
Where is the data? Internet

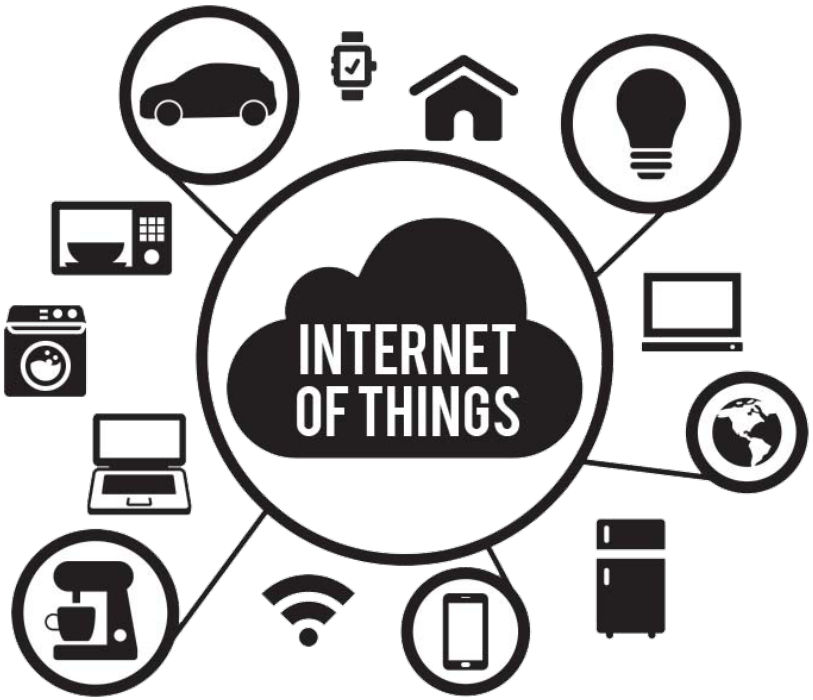
- In the US:



- 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

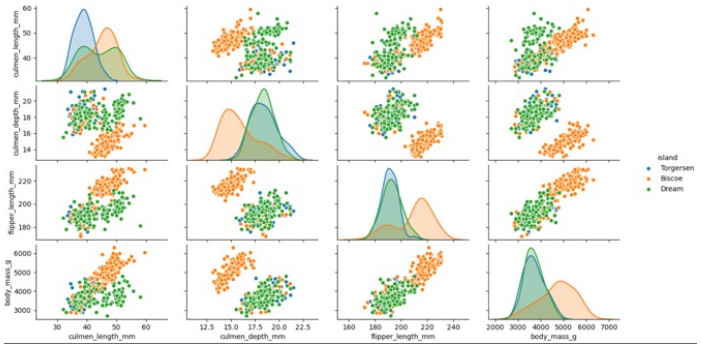
- Data description consists in summarizing the data in an “understable” way, either:
  - Through **exploratory data analysis**: see Chapter 4
    - Mostly descriptive statistics such as average, standard deviation, median, PCA…
  - Through **data visualization**: see Chapter 5

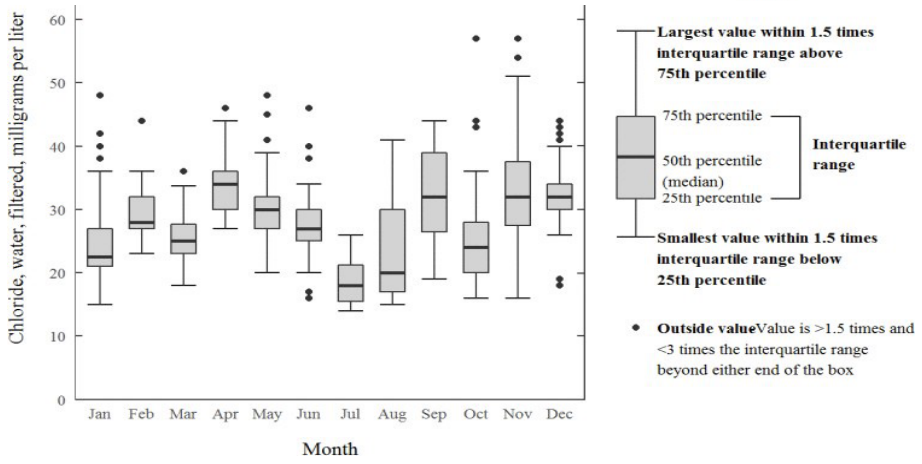
[http://graphics.wsj.com/job-market-tracker/]

What can we do with the data?
Data description through Exploratory Data Analysis

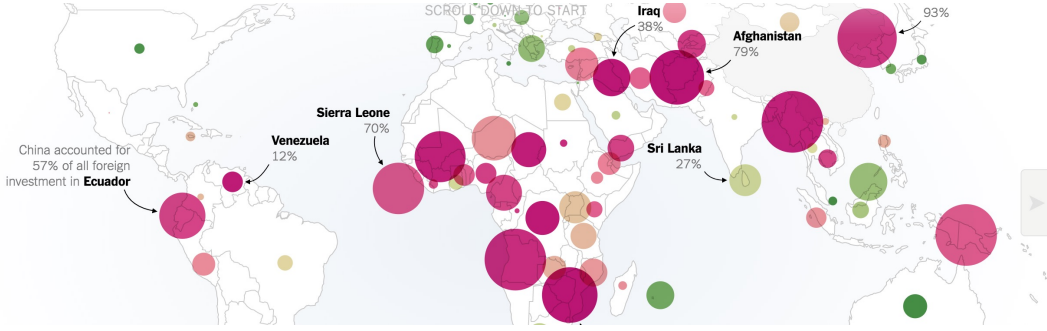
	INDIVIDUALS' AGE					
	Subadult		*Adult*		*Senior*	
Acoustic parameters	**median**	**range**	**median**	**range**	**median**	**range**
call duration [s]	10.3 (8.1, 12.8)	4.4–56.5	8.4(6.8, 10.6)	1.9–201.6	8.3 (7.2, 9.5)	1.7–29.8
number of elements	8 (6, 9)	4–50	9 (7, 12)	3–127	9 (8, 10)	1–33
element duration [s]	0.46 (0.20, 0.73)	0.01–1.73	0.27 (0.14, 0.58)	0.01–2.09	0.24 (0.13, 0.58)	0.02–1.73
interval duration [s]	0.88 (0.50, 1.24)	0.06–4.47	0.54 (0.32, 1.01)	0.04–4.34	0.46 (0.29, 0.97)	0.09–2.74
start F0 [Hz]	680 (640, 760)	530–1460	700 (620, 850)	390–1540	670 (600, 870)	10–1530
end F0 [Hz]	920 (820, 1040)	600–1700	950 (810, 1090)	480–1950	960 (820, 1100)	430–1570
max F0 [Hz]	930 (870, 1010)	640–1430	950 (840, 1040)	530–1480	930 (810, 1040)	500–1400
location of max F0 [s]	1040 (950, 1170)	710, 1900	1070 (930, 1210)	540–1950	1120 (960, 1210)	510–1680
number of elements	*N* = 2476		*N* = 5557		*N* = 2344	
number of calls	*N* = 259		*N* = 533		*N* = 257	

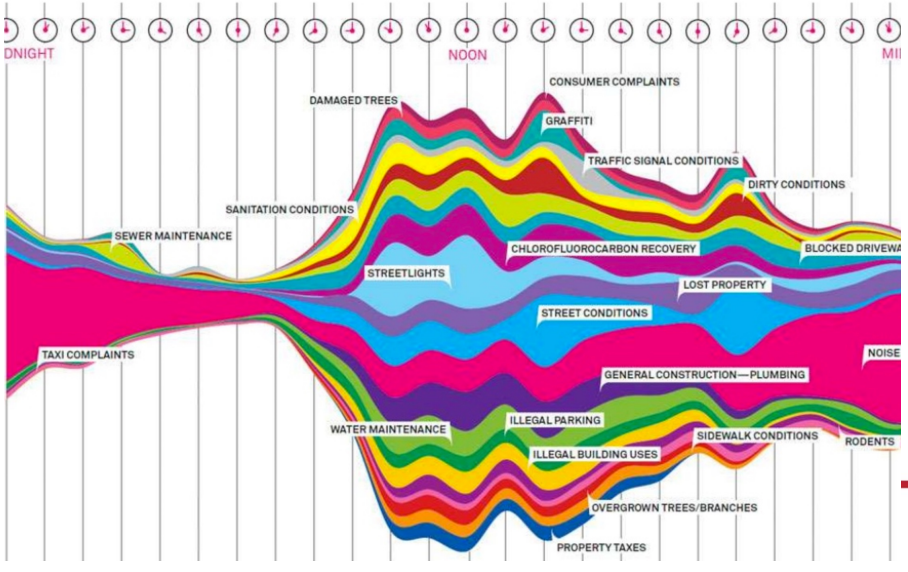
doi:10.1371/journal.pone.0082748.t004

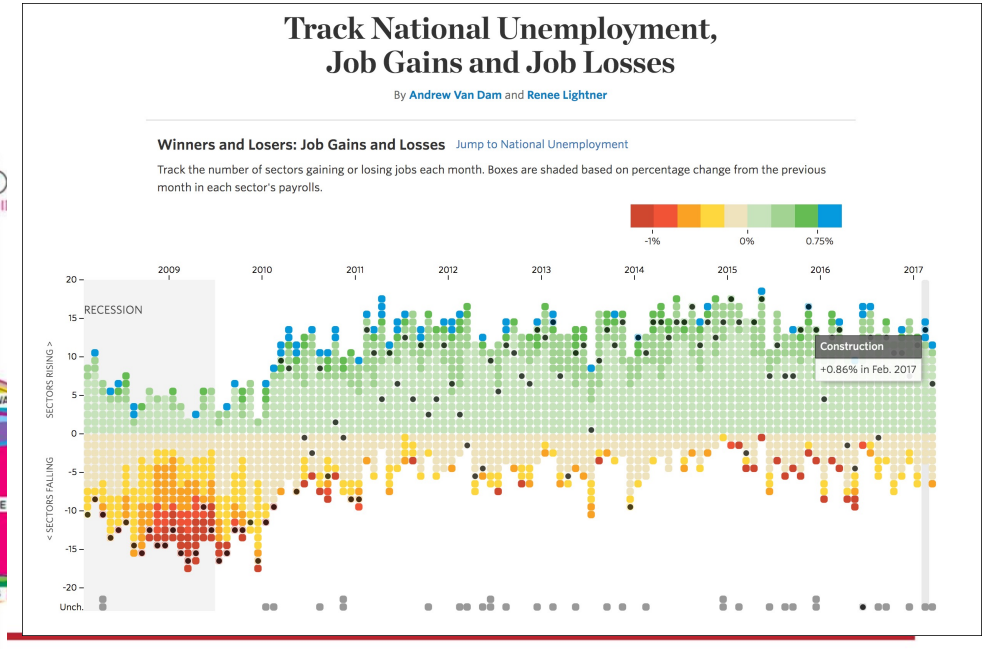




What can we do with the data?
Data description through visualization





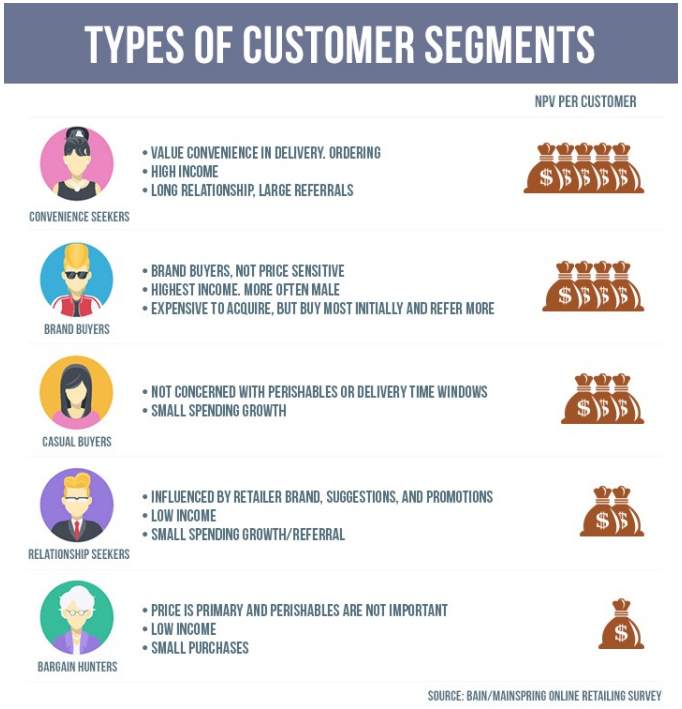


[http://graphics.wsj.com/job-market-tracker/]

What can we do with the data?
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** machine learning methods can be used: see Chapter 6

What can we do with the data?
Data 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.

What can we do with the data?
Data association

- 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** machine learning methods can be used: see Chapter 6

What can we do with the data?
Data association

amazon

"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

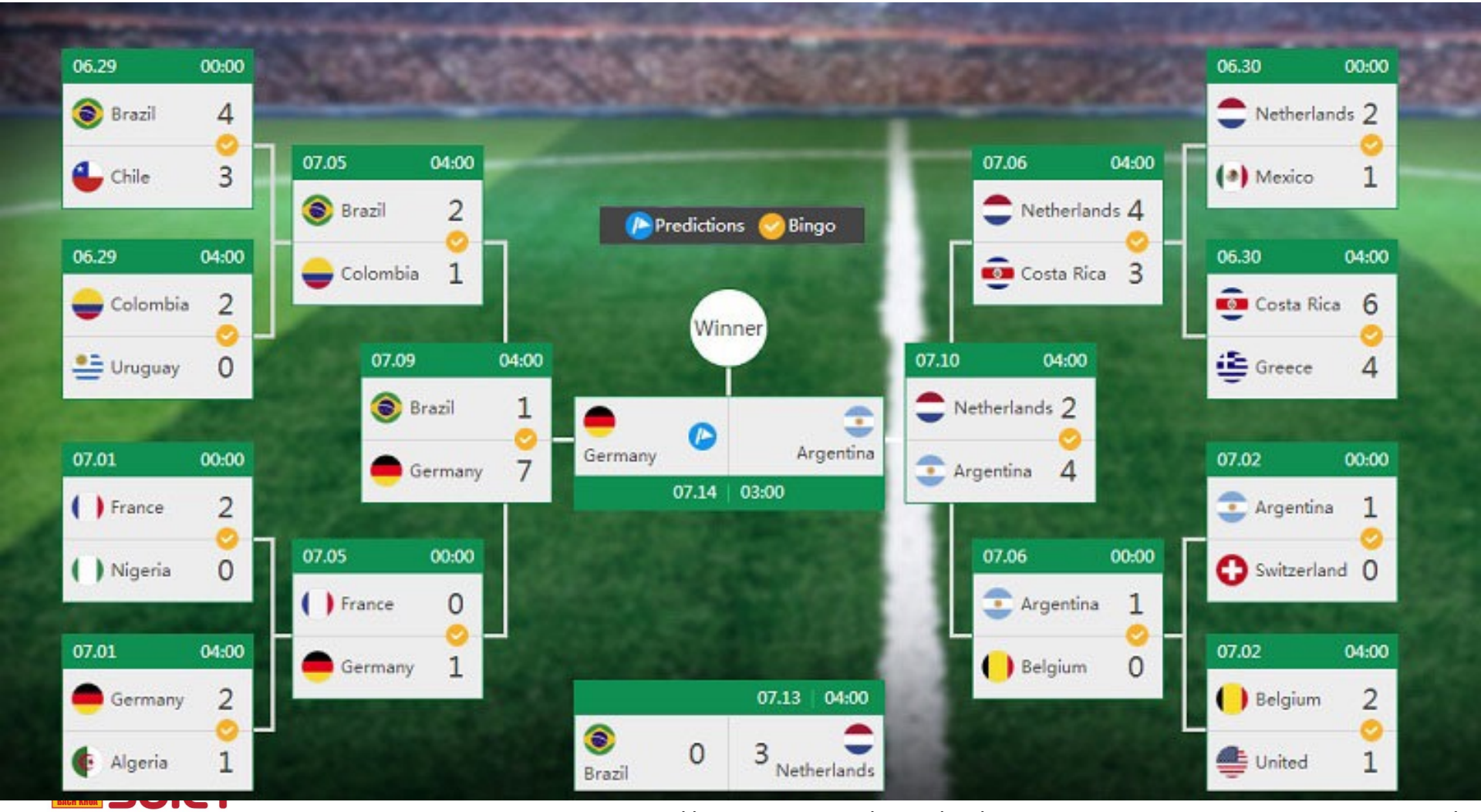
What can we do with the data?
Data prediction

- Data prediction consists in either:
  - predicting (in the future) or estimating (in the present) 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** machine learning methods can be used: see Chapter 6

Accuracy ~93%.

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

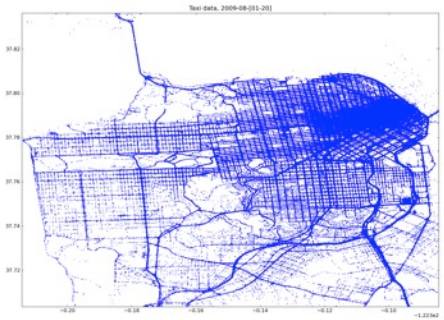
What can we do with the data?
Data prediction

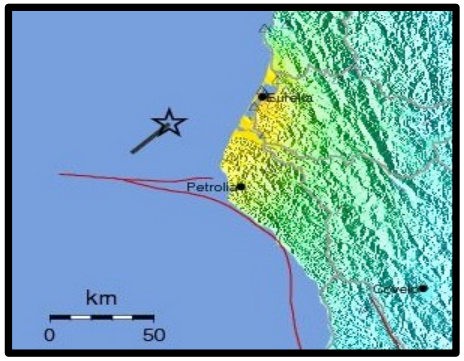


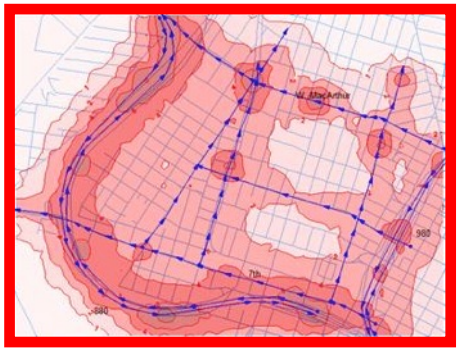
Accuracy ~93%.

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

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***.

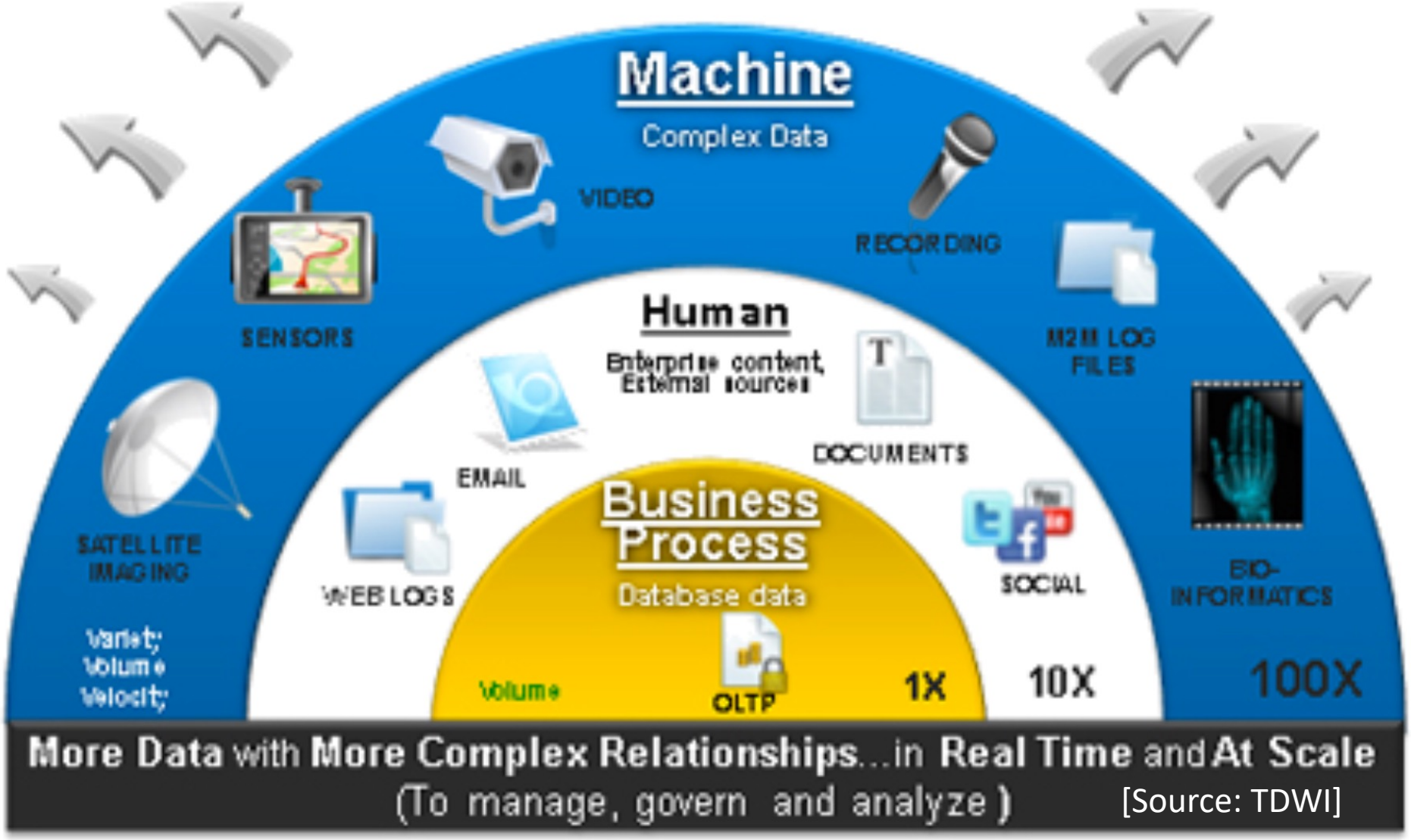
The image was created by blending more than 1,400 separate photos taken over the course of 15 hours—a meticulous process that took him nearly three months.

PHOTO: STEPHEN WILKES

Big data – today

Big data – today: some numbers

Big data – today: sources



[Source: TDWI]

Big data

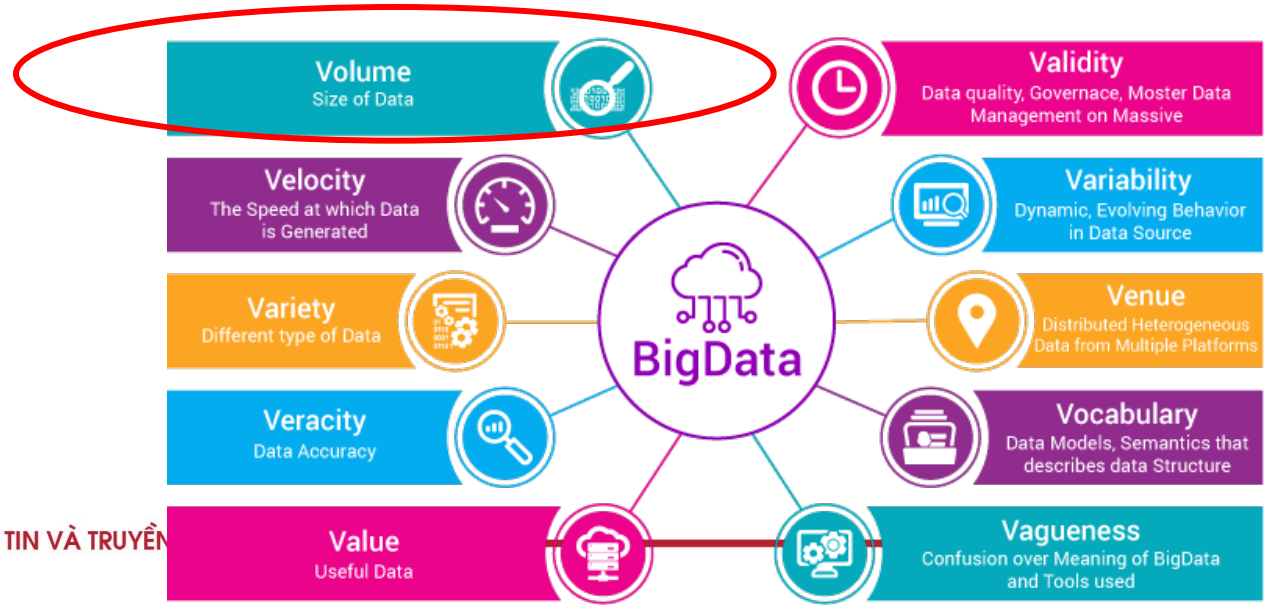
Challenges

The 10 Vs of Big data

[Source: houseofbots.com]

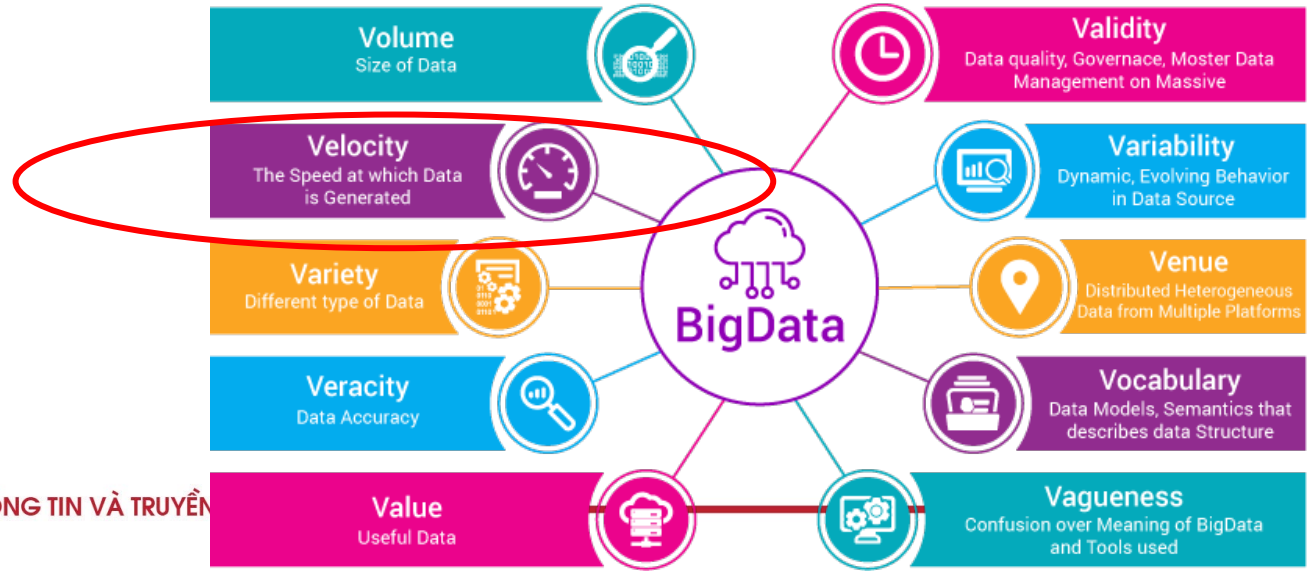
The 10 Vs of Big data: Volume

- 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 scraping (see Chapter 2)
    - data integration (see Chapter 3)



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 /scrape 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 (see Chapter 3)

The 10 Vs of Big data: Venue

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

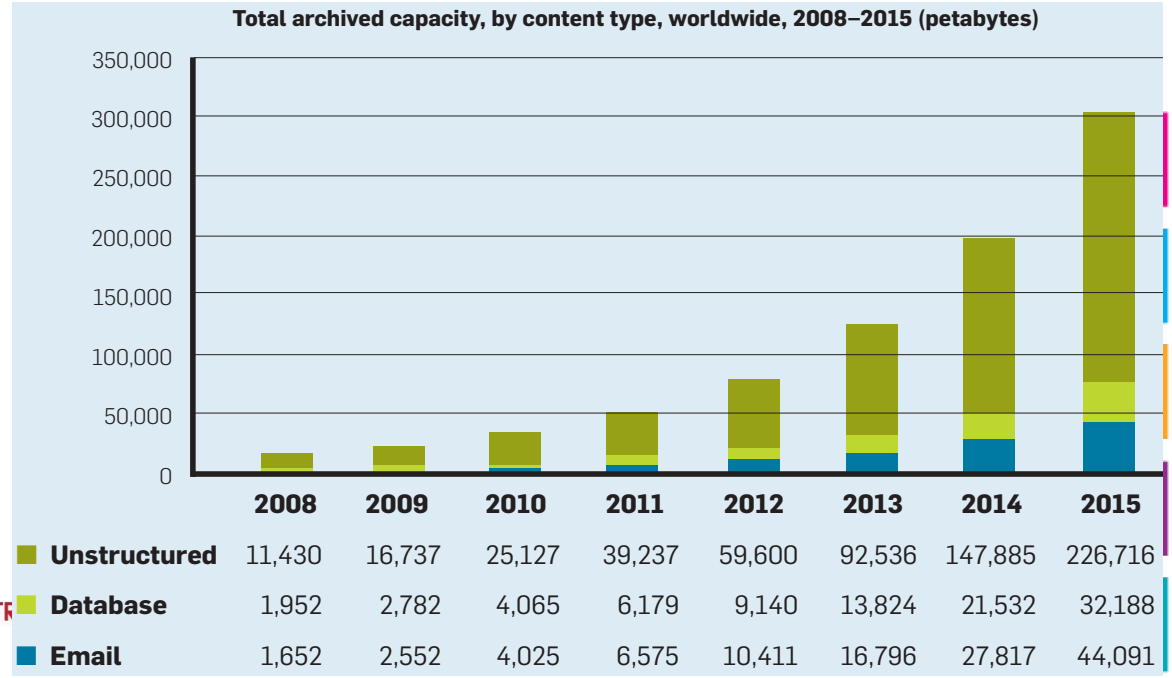
The 10 Vs of Big data: Variability

- 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
- Variability poses serious issues for data integration (see Chapter 3)

The 10 Vs of Big data: Variety

- Variety refers to the different kinds of data one has to hande:
  - **Structured** data: from OLTP datasets or Excel files, for instance
  - **Unstructured** data increases extremely fast: texts, images, tags, links, likes, emotions, …
    - Hence the need / possibility of machine learning (Chapter 6)

(Vasant Dhar, CACM, 2013)

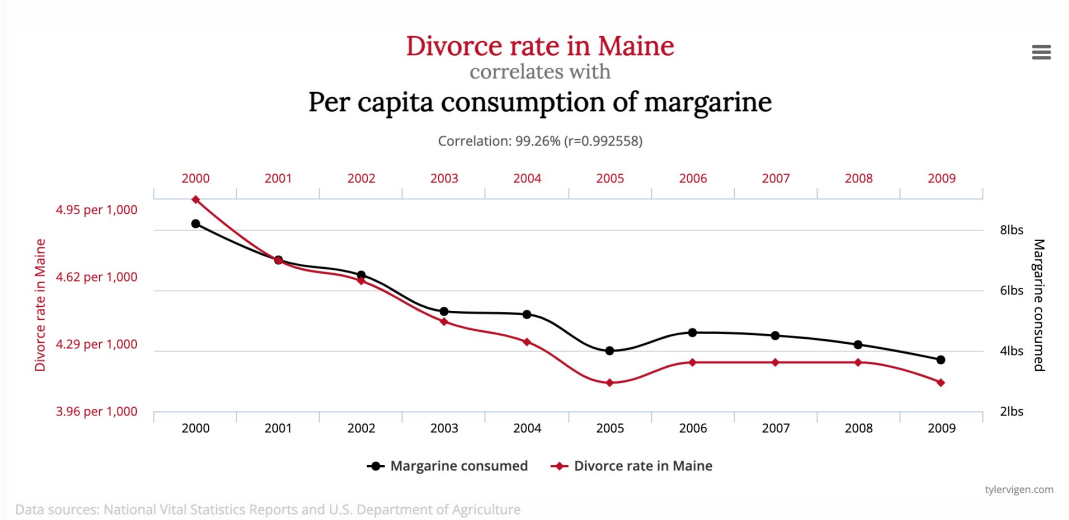


The 10 Vs of Big data: Vocabulary

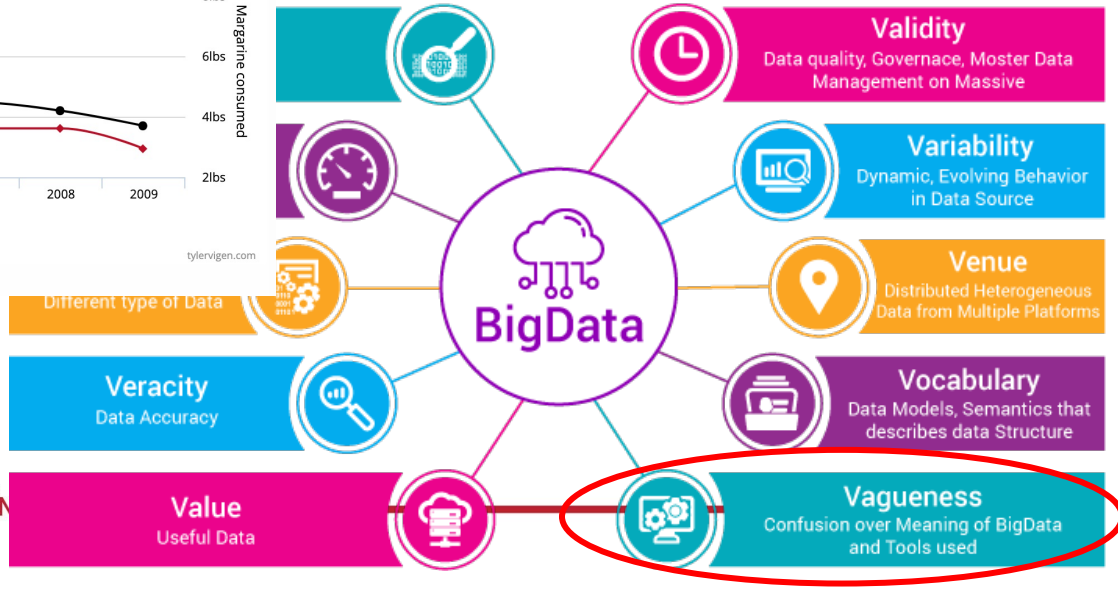
- Vocabulary refers to bringing data models / semantics (knowledge, e.g. ontologies) into the data to structure / explain it
  - See the courses « Intro to AI » and « semantic web »

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



[http://www.tylervigen.com/spurious-correlations]



The 10 Vs of Big data: Veracity

- Veracity: does the data reflect the realty?
  - **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

Some additional issues

- The interactions or **correlations** hidden in data might be really huge
  - Can affect the Machine Learning methods used
- 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

- Poses several serious challenges for Machine Learning techniques

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?

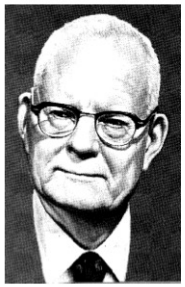
Data Science - early days

1935: "The Design of Experiments"

R.A. Fisher

1939: "Quality Control"

W.E. Demming

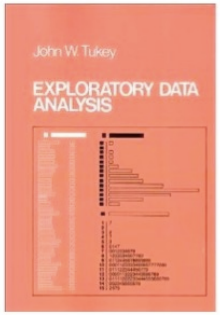


1958: "A Business Intelligence System"

Peter Luhn

1977: "Exploratory Data Analysis"

J.W. Tukey



1989: "Business Intelligence"

Howard Dresner

1997: "Machine Learning"

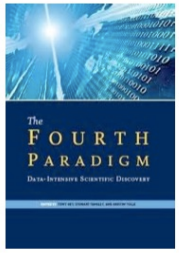
Tom Mitchell

2010:"The Data Deluge"

2009:"The Unreasonable Effectiveness of Data"

Peter Norvig

2007:"The Fourth Paradigm"



1996: Google

Google™

(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

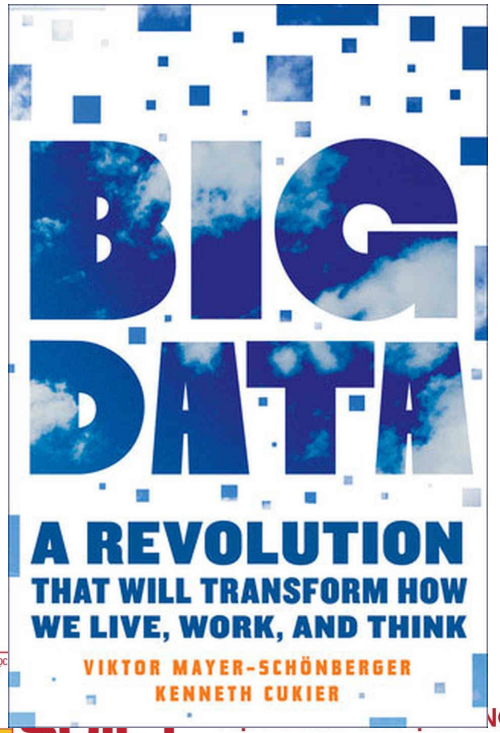
"The ability to take **data** – to be able to **understand** it, to **process** it, to **extract value** from it, to **visualize** it, to **communicate** it's going to be a hugely important skill in the next decades, not only at the professional level but even at the educational level for elementary 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

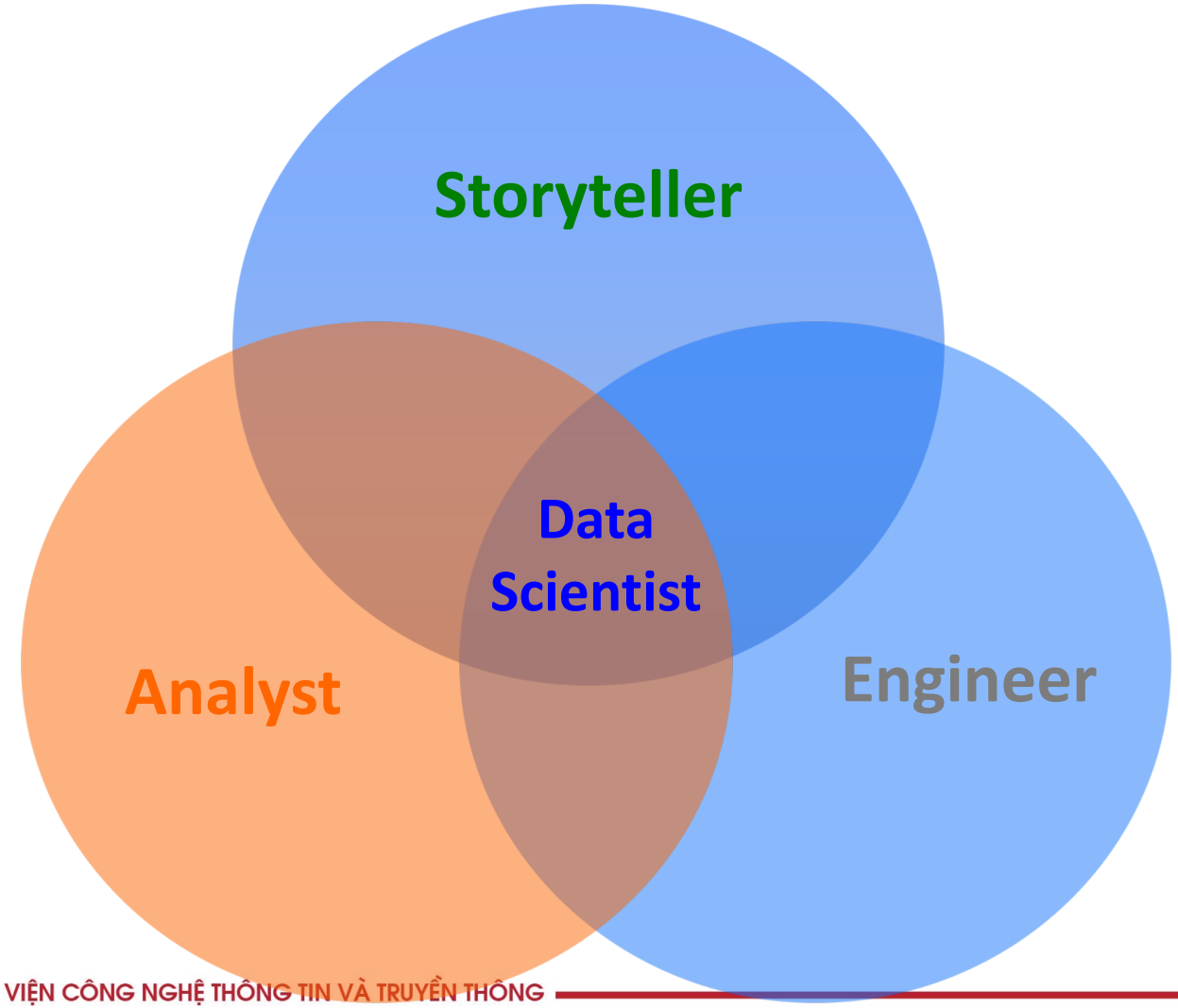
by Thomas H. Davenport and D.J. Patil



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.

Questions

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you for your attention!

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