Advanced DataBases

(Grau en Ciència I Enginyeria de les dades)

Alberto Abelló & Sergi Nadal

Facultat d'Informàtica de Barcelona

Universitat Politècnica de Catalunya

Introduction

Knowledge objectives

1. Explain the data-driven decision making framework
2. Explain what the "Business Intelligence Cycle" is
3. Identify the different data science flows
4. Give a definition of Big Data

NEEDS

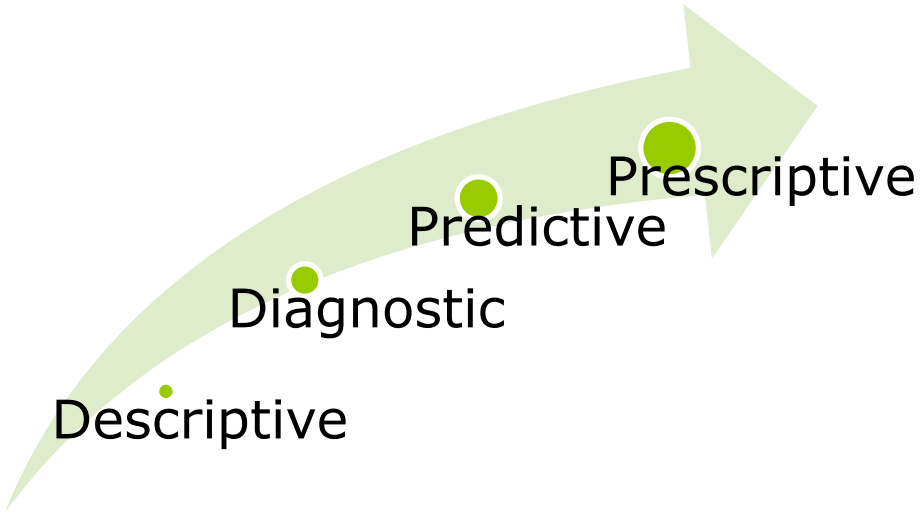
Motivation

"*Without data you are just another person with an opinion*."

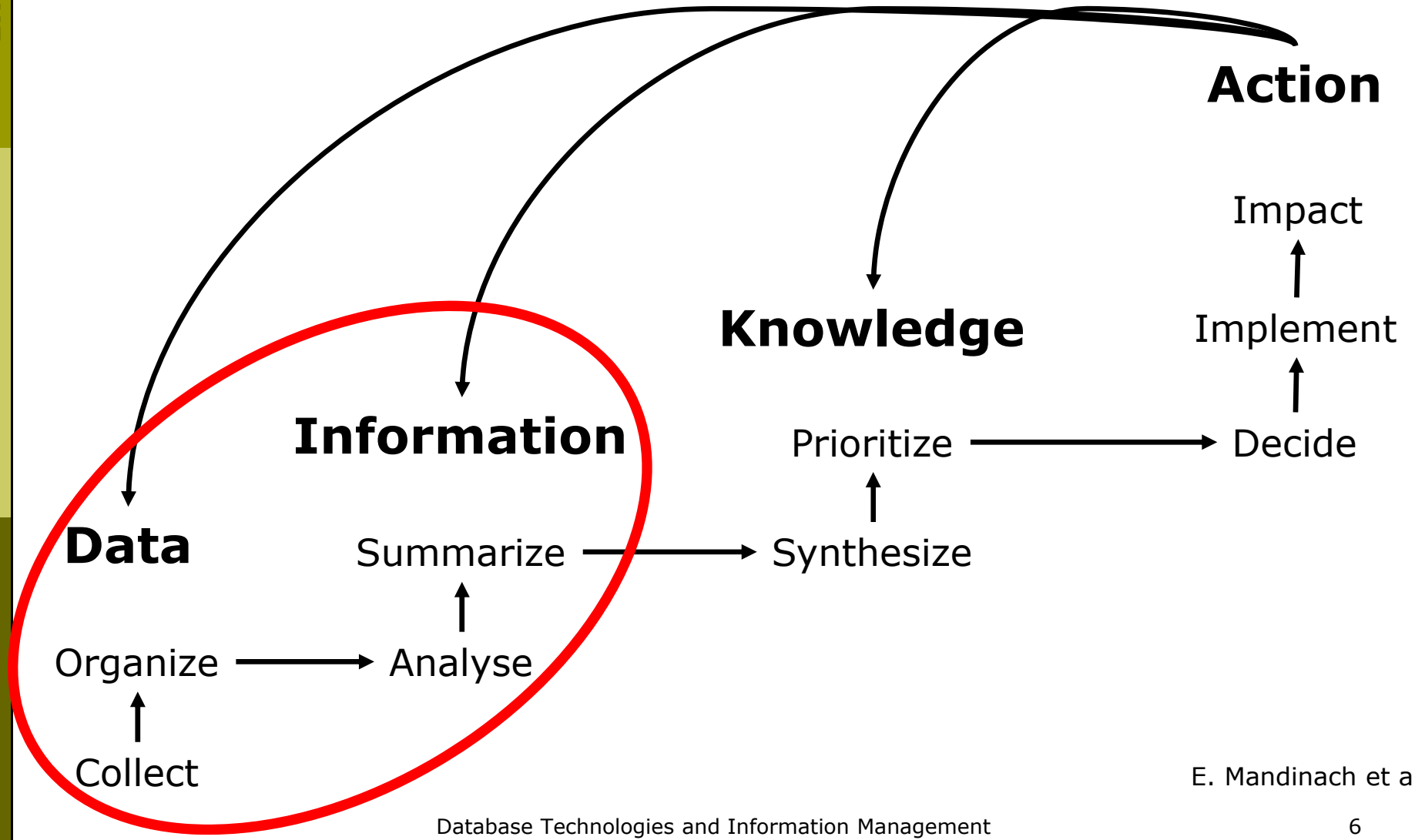
William Edwards Deming

"*It is a capital mistake to theorize before one has data*."

Sherlock Holmes (A Study in Scarlett)

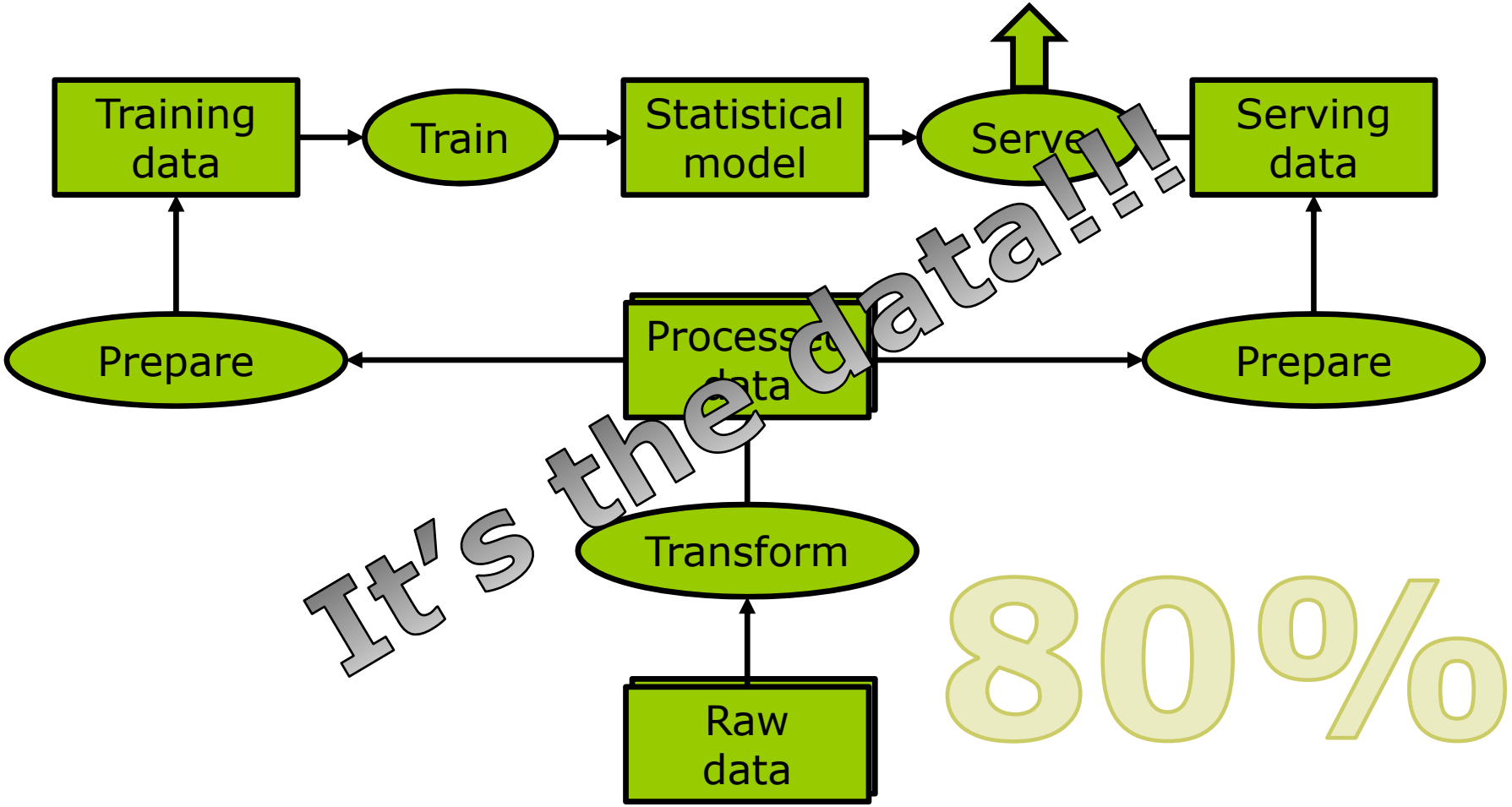


Data driven decision making



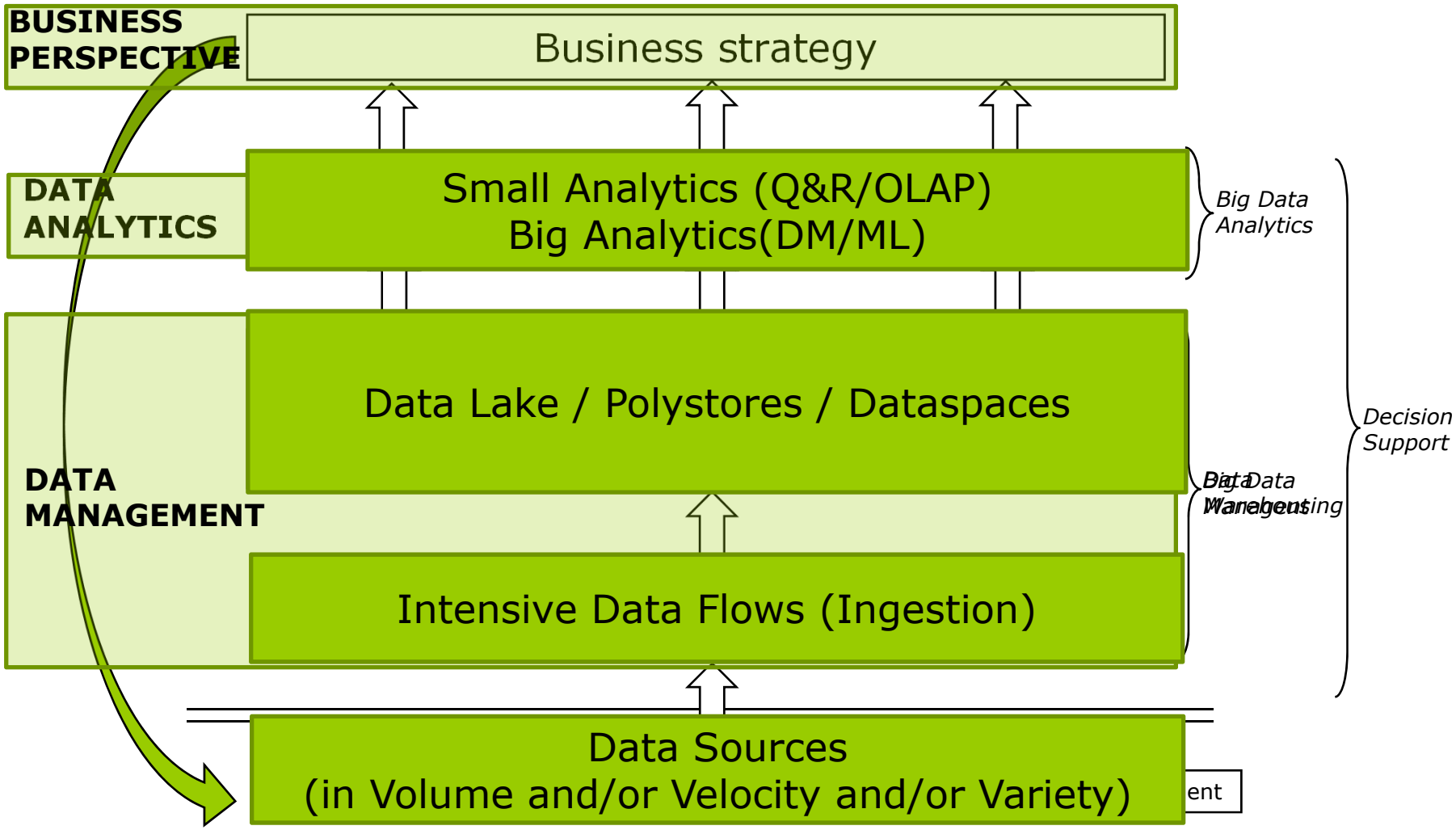
E. Mandinach et al.

Data science lifecycle

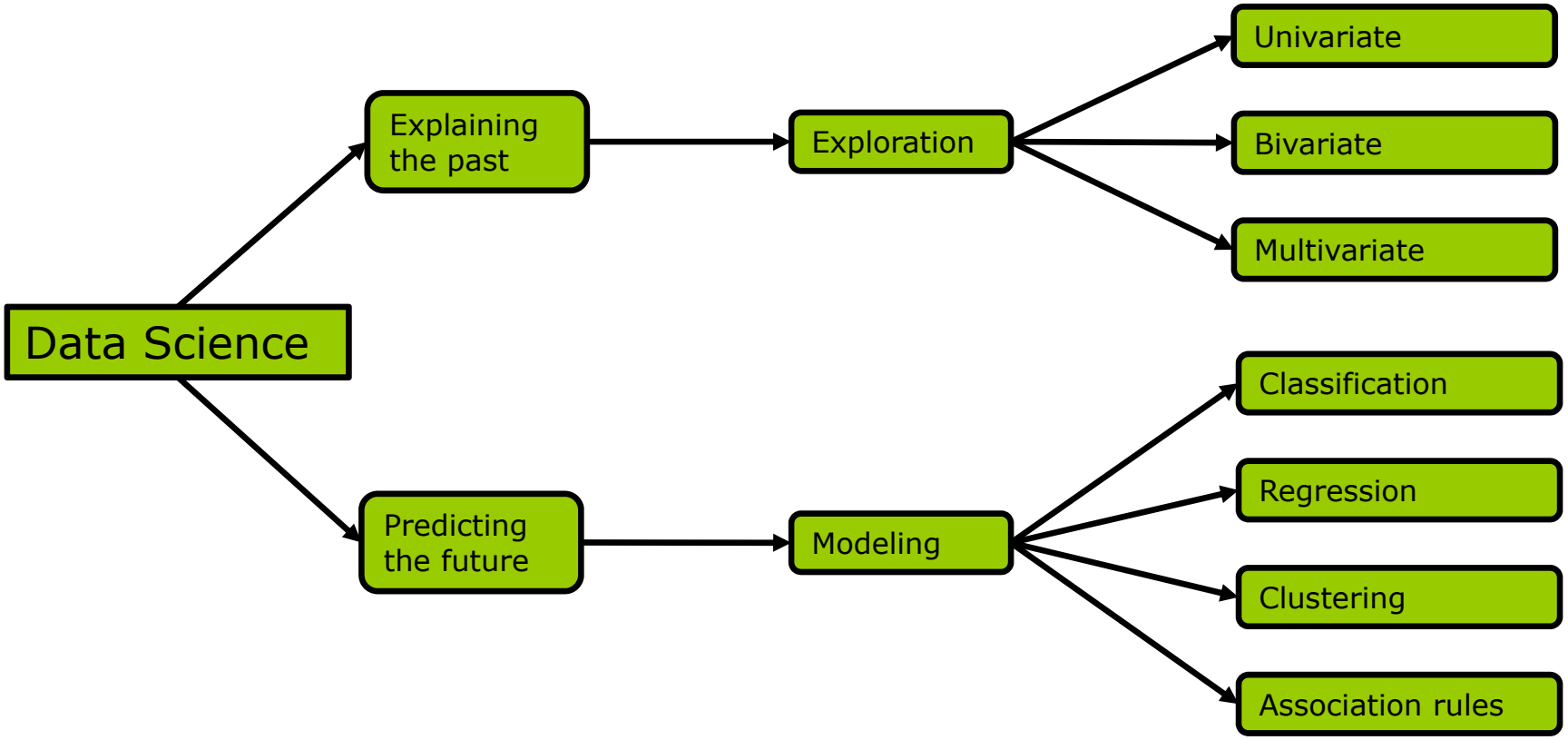


http://www.redbooks.ibm.com/redpapers/pdfs/redp5120.pdf

Business Intelligence Cycle



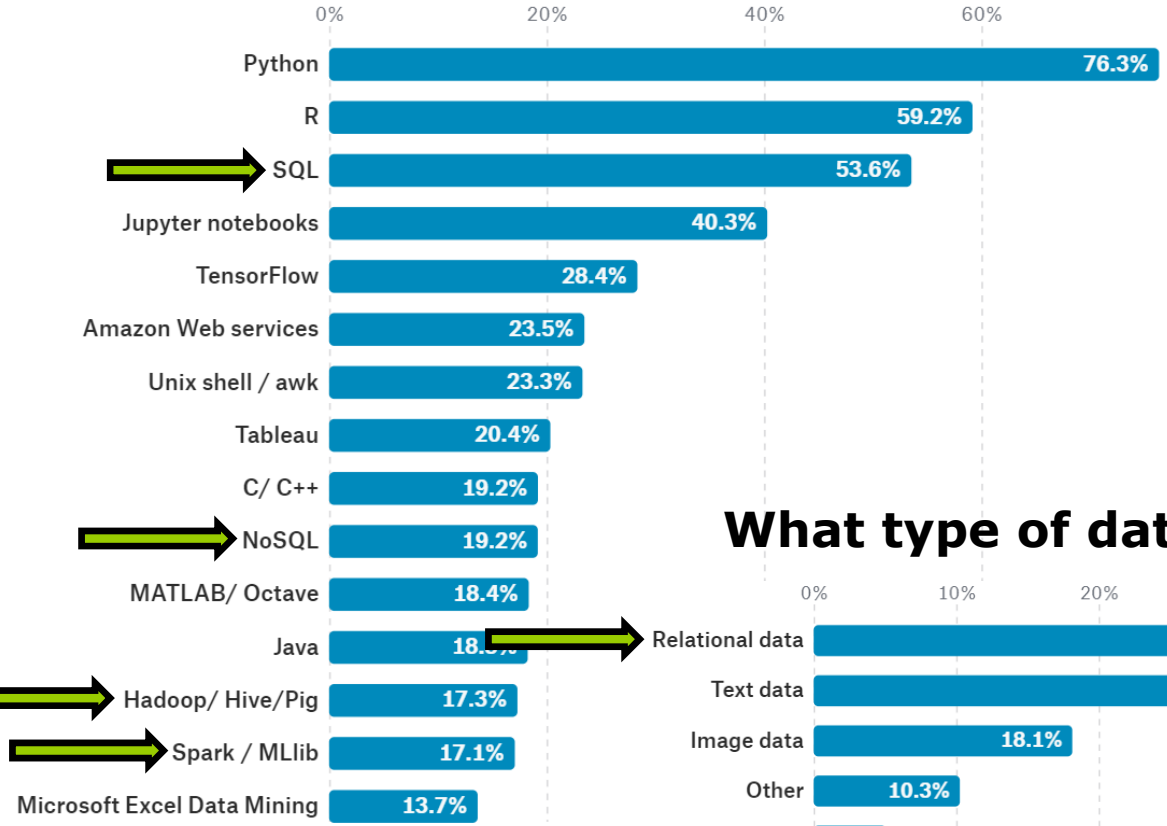
Data Science tools



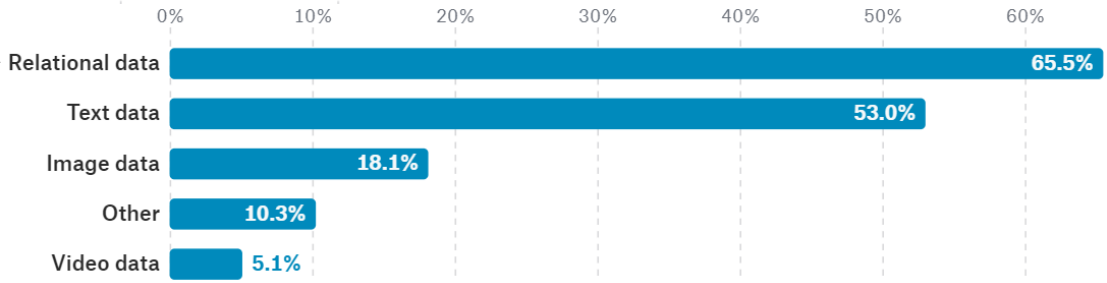
http://www.saedsayad.com/data_mining_map.htm

Kaggle report

What tools are used at work?



What type of data is used at work?

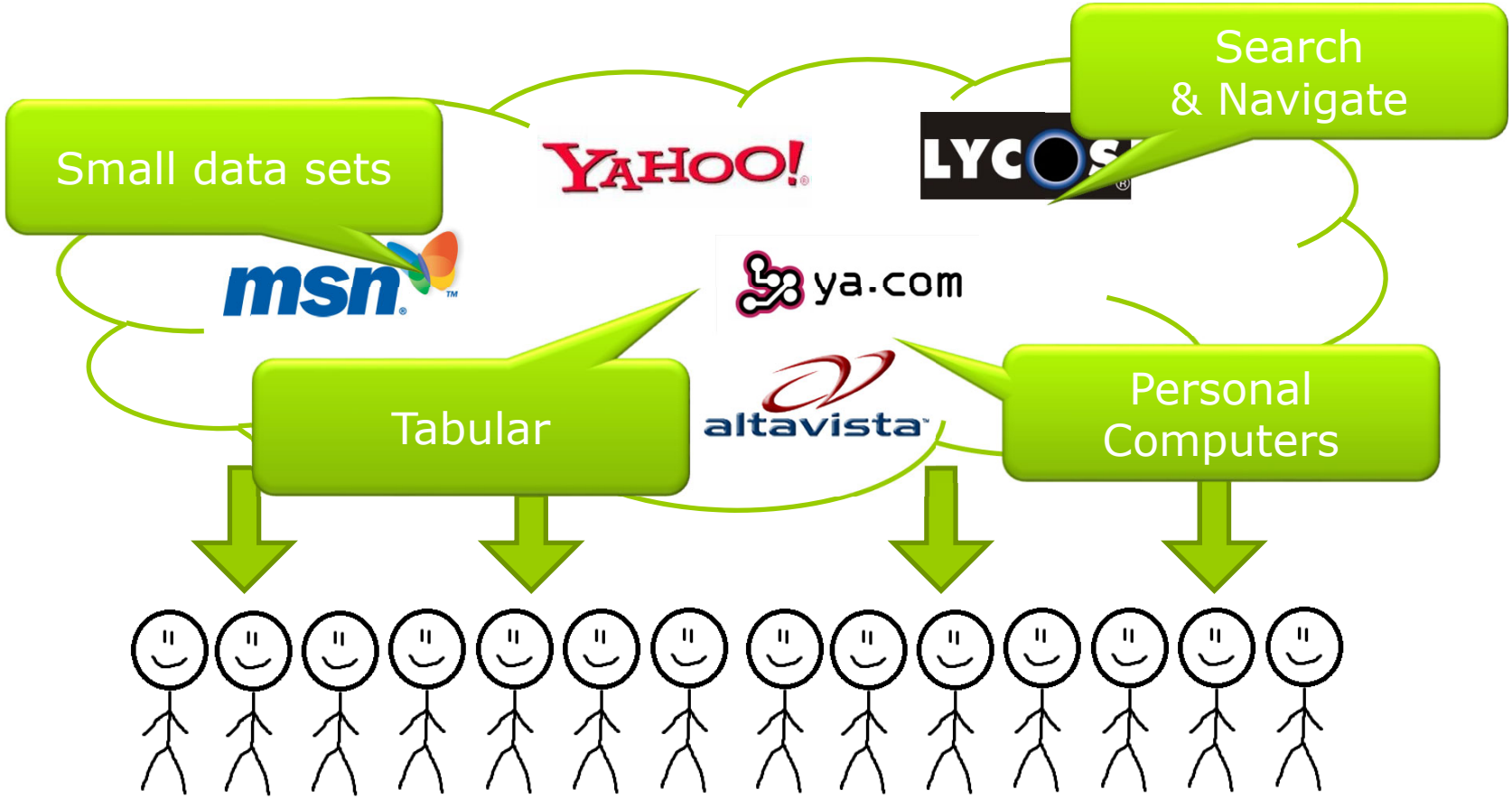


https://www.kaggle.com/surveys/2017

BIG DATA

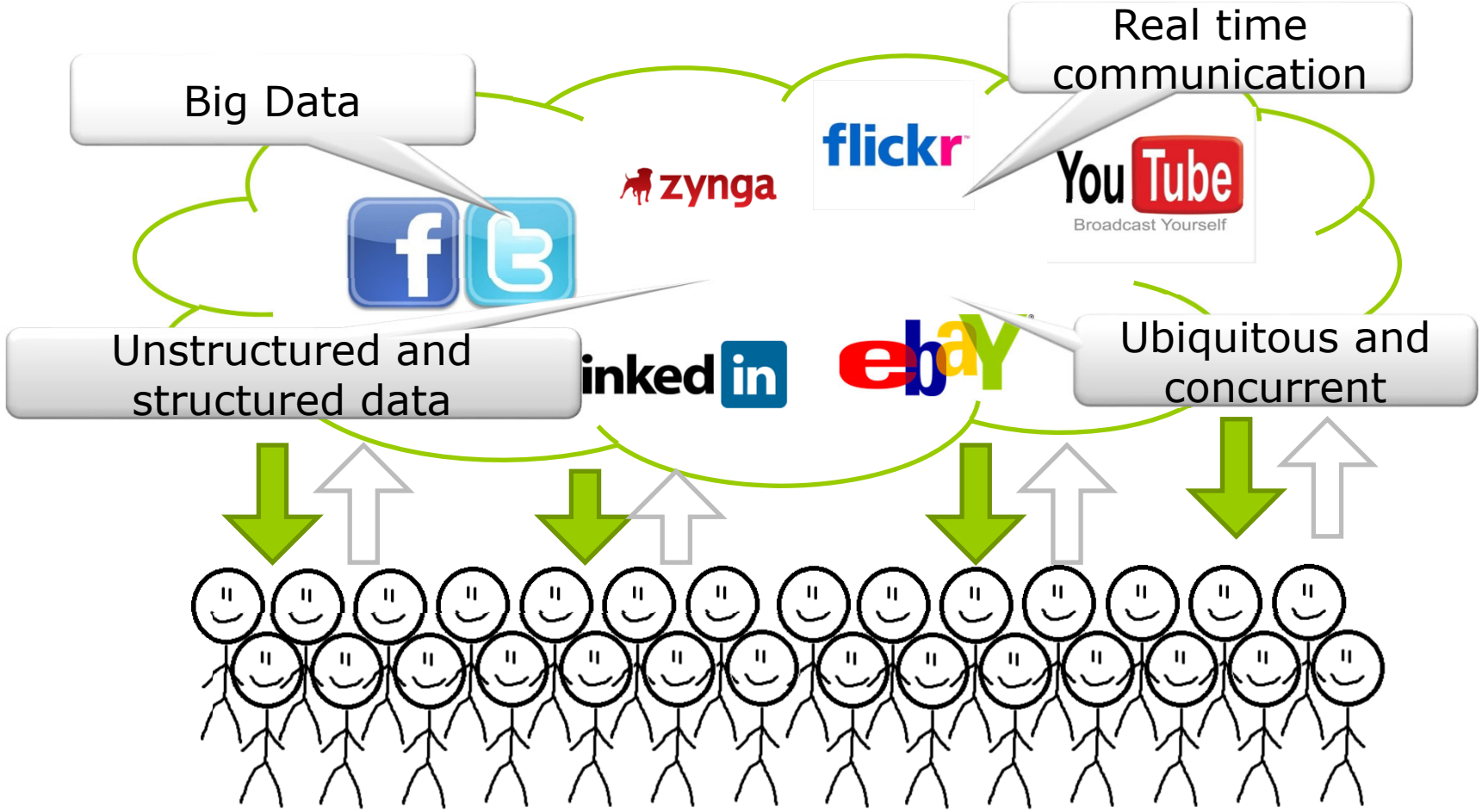
The End of an Architectural Era

WEB 1.0 – Read Era



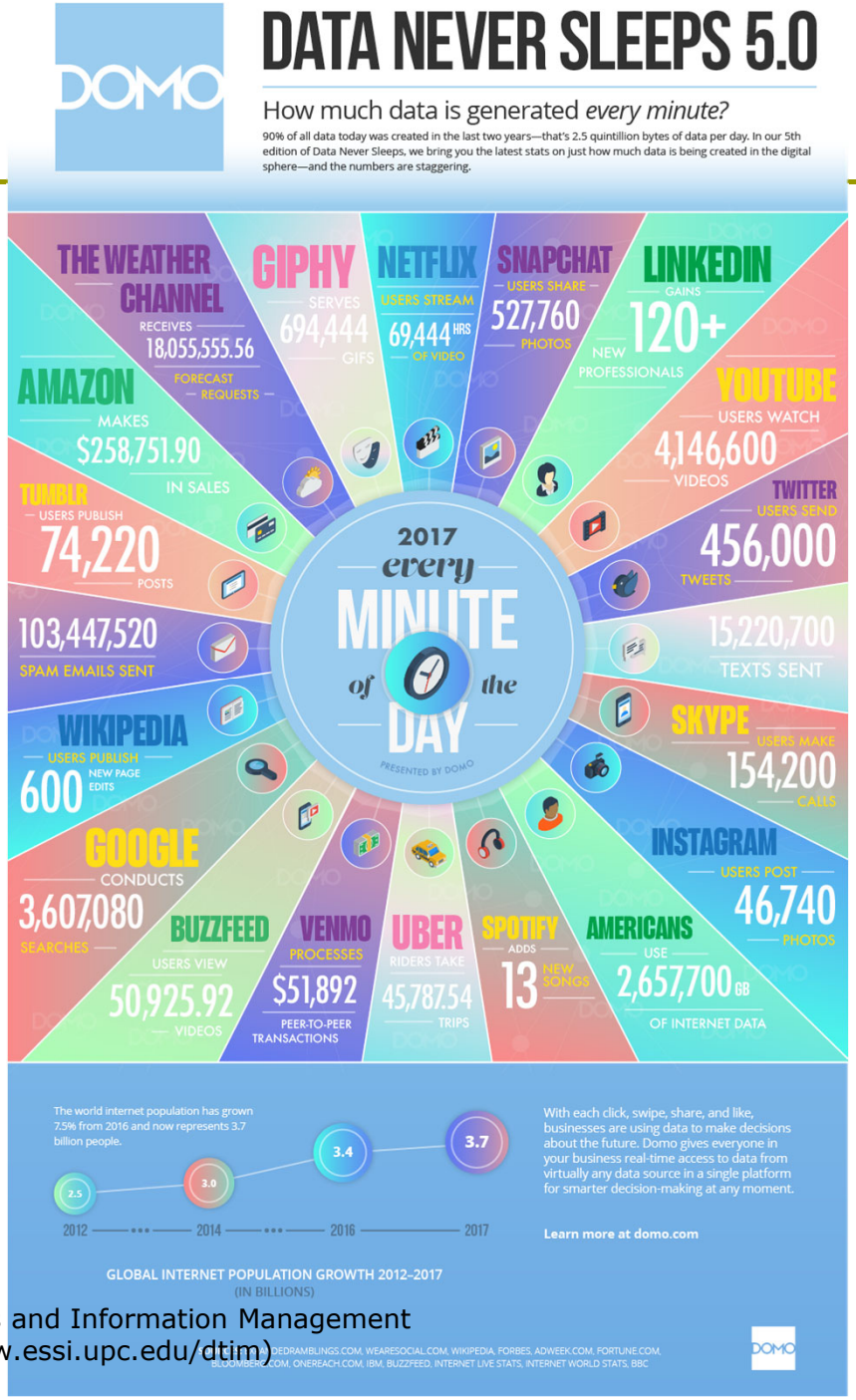
The End of an Architectural Era

WEB 2.0 – Write Era



Maria Belen Bianchi

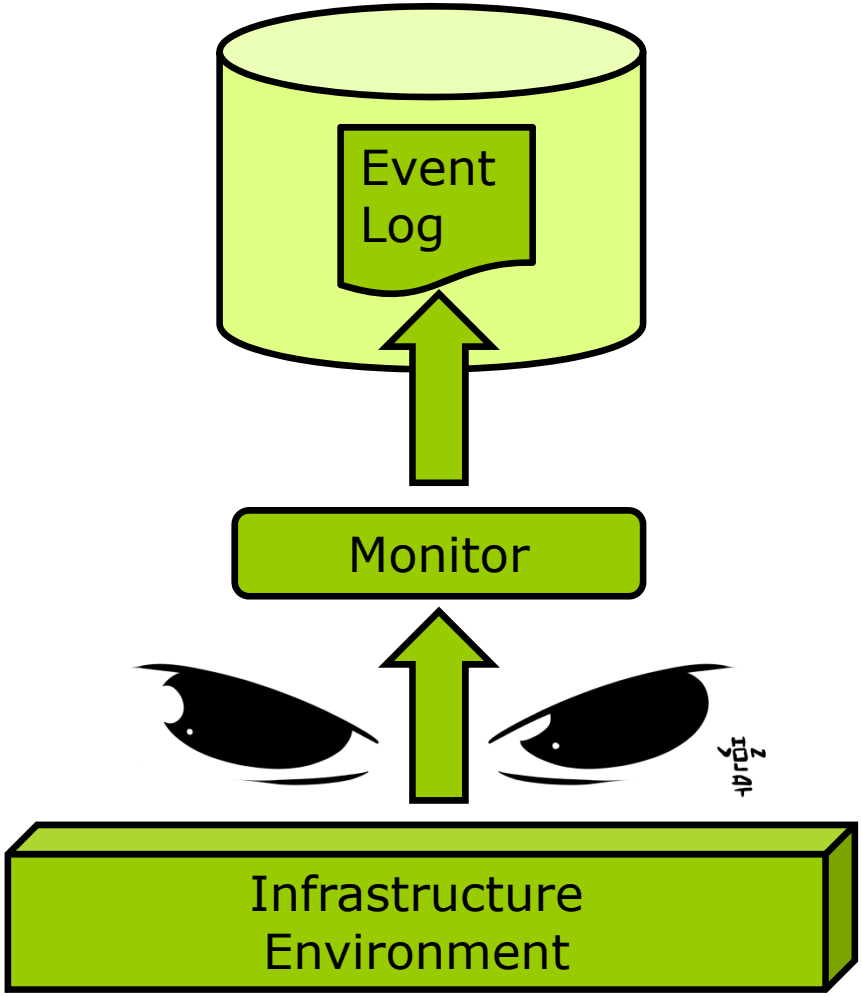
It is estimated that in 2020 there will be more data than sand grains in the world (40 Zb)



New business model

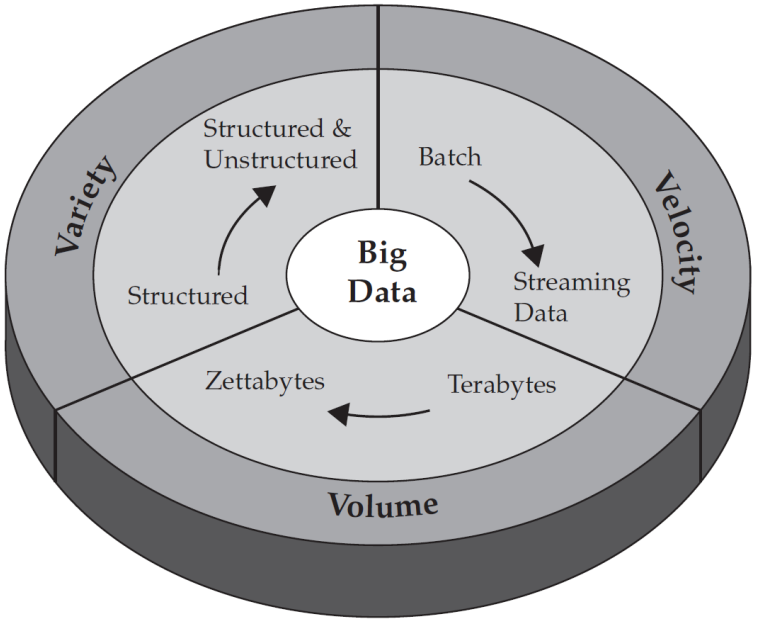
- Hello good afternoon. Renato Pizza?
- No sir, this is Google Pizza.
- Excuse me, I'll have the wrong number ...
- No sir, Google has bought and renamed it.
- Oh perfect! Well I would like to order.
- Very good, Mr. López. The usual order?
- The usual? Mr. López? Do you know me?
- According to our caller ID, the last 12 times, you has ordered an individual *Quattro Formaggio*.
- Exactly, that's what I want.
- Can I suggest you try this time our Vegetable with ricotta, arugula, eggplant, zucchini and dried tomato?
- No thanks. I hate vegetables.
- Yeah, but it would be better for your cholesterol whose level is not very good.
- Excuse me? How do you know that?
- Through your subscription to the Online Medical Guide, we see your blood tests of the last 5 years.
- But I do not like that pizza, I hate the vegetable. Also, I'm being treated and taking the right medication.
- Mr. López, you know that you do not take medication regularly, 5 months ago you bought a box of 30 pills at Otero García Pharmacy, and you didn't buy more ...
- That's not true, I bought more at another pharmacy.
- Well, it does not appear on your credit card statement ...
- Because I paid in cash.
- Well, according to your balance, you have hardly any cash in your pocket ...
- I have cash at home.
- Seriously? Well, you have not declared it in your last income declaration ... recognizing that you declare less than you earn? That is a crime, Mr. López.
- But, WHAT DO YOU HAVE ...?! Enough! I'm sick of Google, Facebook, Twitter, WhatsApp, Instagram ... I'm going to a deserted island without Internet, where there are no phones, and nobody can spy me!
- I understand, gentleman. But remember that you must renew your passport, expired three months ago ...

Monitoring the infrastructure

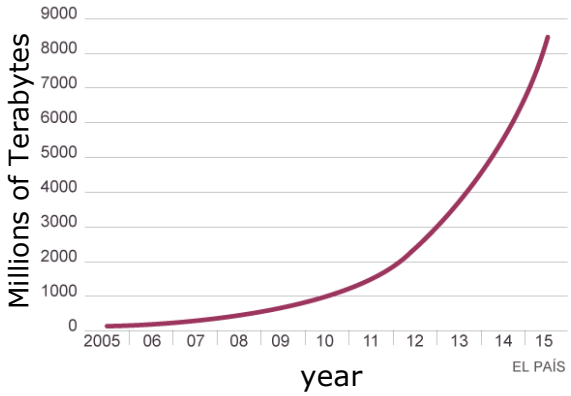


Big Data definition

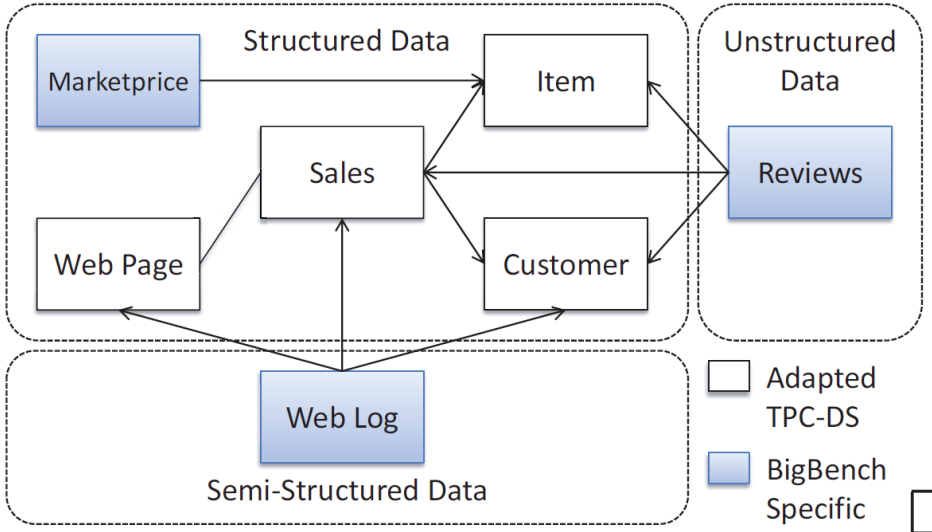
- Volume
- Velocity
- Variety
- …
- Variability
- Validity/Veracity
- Value



From IBM "Understanding Big Data"



Bigbench



Query processing type	Total	Percentage(%)
Declarative	10	33.3
Procedural	7	23.3
Mix of Declarative and Procedural	13	43.3

Data sources	Total	Percentage(%)
Structured	18	60.0
Semi-structured	7	23.3
Un-structured	5	16.7

Analytic techniques	Total	Percentage(%)
Statistics analysis	6	20.0
Data mining	17	56.7
Reporting	8	26.7

Types of Big Data Analyzed in Industry

	Manufacturing and Natural Resources	Media/ Communications	Services	Government	Education	Retail	Banking	Insurance	Healthcare	Transportation	Utilities
Transactions	73%	62%	67%	67%	54%	93%	83%	81%	75%	79%	80%
Log data	44%	57%	58%	59%	54%	40%	66%	61%	33%	71%	60%
Machine or sensor data	53%	38%	35%	33%	31%	27%	27%	48%	42%	50%	40%
Emails /documents	27%	43%	43%	41%	46%	27%	34%	39%	17%	29%	20%
Social media data	32%	52%	39%	26%	54%	73%	27%	13%	-	50%	-
Free-form text	17%	24%	28%	30%	31%	20%	34%	35%	67%	21%	40%
Geospatial data	27%	14%	19%	19%	38%	27%	27%	26%	8%	29%	40%
Images	19%	24%	17%	11%	38%	13%	5%	16%	25%	7%	-
Video	8%	29%	12%	7%	31%	13%	-	6%	8%	7%	-
Audio	10%	19%	8%	4%	8%	-	-	6%	-	-	-
Other	8%	14%	13%	15%	8%	7%	10%	16%	42%	14%	-
n =	*59*	*21**	*127*	*27**	*13**	*15**	*41*	*31*	*12**	*14**	*5**

Note: Highlighted cells indicate the top three data types by industry.
Multiple responses allowed

Source: Gartner (September 2013)

CLOSING

Summary

- Data Science flows
- Business Intelligence lifecycle
- Big Data definition

Bibliography

- E. Mandinach et al. *A Theoretical Framework for Data-Driven Decision Making*. AERA annual meeting, 2006
- D. Donoho. *50 years of Data Science*. http://courses.csail.mit.edu/18.337/2015/docs/50YearsDataScience.pdf
- D. Abadi. *Data management in the cloud: Limitations and opportunities*. IEEE Data Engineering Bulletin 32(1), 2009
- M. Madsen. *Cloud Computing Models for Data Warehousing*. Third Nature Technology White Paper, 2012
- A. Ghazal et al. *BigBench: towards an industry standard benchmark for big data analytics*. SIGMOD'13
- Gartner Reports. G00232650, G00175593, and G00219131