

# Data Science

Fabrice Rossi

CEREMADE  
Université Paris Dauphine

2021

# Data Science?

the Science of Data!

the Science of Data!



## Informal definitions

1. Collection of measured and recorded values
2. Set of values of entities (subjects) with respect to variables (attributes)

## Data is not

1. information: structured, post-processed, organized data
2. knowledge: high level understanding

# Data?

## Informal definitions

1. Collection of measured and recorded values
2. Set of values of entities (subjects) with respect to variables (attributes)

## Data is not

1. information: structured, post-processed, organized data
2. knowledge: high level understanding

## Remarks

- ▶ *data* tends to be used as a singular mass noun (as information)
- ▶ *datum* is seldom used
- ▶ *a data set*: a collection of data/datum

# Example

	age	job	marital	education	default	balance	housing
1	30	unemployed	married	primary	no	1787	no
2	33	services	married	secondary	no	4789	yes
3	35	management	single	tertiary	no	1350	yes
4	30	management	married	tertiary	no	1476	yes
5	59	blue-collar	married	secondary	no	0	yes
6	35	management	single	tertiary	no	747	no
7	36	self-employed	married	tertiary	no	307	yes
8	39	technician	married	secondary	no	147	yes
9	41	entrepreneur	married	tertiary	no	221	yes
10	43	services	married	primary	no	-88	yes
11	39	services	married	secondary	no	9374	yes
12	43	admin.	married	secondary	no	264	yes
13	36	technician	married	tertiary	no	1109	no
14	20	student	single	secondary	no	502	no
15	31	blue-collar	married	secondary	no	360	yes
16	40	management	married	tertiary	no	194	no
17	56	technician	married	secondary	no	4073	no
18	37	admin.	single	tertiary	no	2317	yes
19	25	blue-collar	single	primary	no	-221	yes
20	31	services	married	secondary	no	132	no

# Data Science?

## A possible definition

Tools, techniques, systems, processes, etc. that extract information and possibly knowledge from data, using the scientific method



## A possible definition

Tools, techniques, systems, processes, etc. that extract information and possibly knowledge from data, using the scientific method

## Examples of standard applications

- ▶ spam filtering:

  - data** emails sorted into categories (acceptable versus spam)

  - information** a computer program that infers from the content of an email its category

- ▶ movie recommendations:

  - data** a movie database (with movie descriptions), user ratings for some movies

  - information** a computer program that infers the rating of an unseen movie for a given user

## Data $\neq$ information

Hierarchy:

1. data
2. information/insights
3. knowledge

## Data $\neq$ information

Hierarchy:

1. data
2. information/insights
3. knowledge

## SoundHound/Shazam

1. data: sound (digital recording)
2. information (low level):  
fingerprint of the recording
3. information (high level):  
metadata of the song  
associated to the fingerprint
4. knowledge: musical genre,  
band history, etc.

## Older/other names

- ▶ data mining
- ▶ knowledge discovery in databases (KDD)
- ▶ business intelligence/analytics
- ▶ statistics (!)
- ▶ big data
- ▶ artificial intelligence (AI)

## Older/other names

- ▶ data mining
- ▶ knowledge discovery in databases (KDD)
- ▶ business intelligence/analytics
- ▶ statistics (!)
- ▶ **big data**
- ▶ **artificial intelligence** (AI)

## The confusion is damaging!

- ▶ **big** data induces specific problems and *specific solutions*
- ▶ AI is **used** in data science

# Big Data?

# Big Data?



# Big Data?

## Operational definition

Data sets that are too big to be processed by a single computer or by a limited number of computers (say less than 10)



## Operational definition

Data sets that are too big to be processed by a single computer or by a limited number of computers (say less than 10)

## Remarks

- ▶ a definition by size is meaningless: computer processing power grows in parallel with storage capacity
- ▶ solutions adapted to large scale data sets are generally wasteful when applied to standard scale data sets

## Doug Laney's 3 Vs

- ▶ Doug Laney was an analyst at the META group (now Gartner)
- ▶ He proposed in 2001 the 3 V's:
  1. Volume: data size
  2. Velocity: streaming context
  3. Variety: text, image, video, etc.
- ▶ frequently used as “characteristics of big data”
- ▶ complemented by other Vs such as Veracity (data quality, confidence in the results)

# X Vs as typical corporate BS

## Volume

is the only acceptable characterization

# X Vs as typical corporate BS

## Volume

is the only acceptable characterization

## Velocity

- ▶ can be Volume when data is stored
- ▶ induces completely different challenges in a true streaming context (when data is thrown away!)
- ▶ is related to drifting and other advanced *standard* data science problems

# X Vs as typical corporate BS

## Volume

is the only acceptable characterization

## Velocity

- ▶ can be Volume when data is stored
- ▶ induces completely different challenges in a true streaming context (when data is thrown away!)
- ▶ is related to drifting and other advanced *standard* data science problems

## Variety and Veracity

have been part of data science since almost its beginning!

# Specific solutions

## Big data

cannot be

- ▶ stored
- ▶ queried
- ▶ processed

on a single computer (even a powerful one!)

## Specific solutions

- ▶ e.g. Google:
  - ▶ in 2016, 2.5 millions of servers (estimation)
  - ▶ Google File System
  - ▶ Bigtable
  - ▶ Map Reduce
- ▶ open source: Apache Hadoop

# Consequences of the confusion

## A typical story

1. we want to do predictive modelling but we confuse that with big data
2. we read tutorials and books about big data finding discussions about predictive modelling, reinforcing our confusion
3. we use the dominant open source solution, hadoop
4. performances are very bad so we add more computers, reinforcing our belief that we do big data magic
5. wash, rinse, repeat

# Consequences of the confusion

## A typical story

1. we want to do predictive modelling but we confuse that with big data
2. we read tutorials and books about big data finding discussions about predictive modelling, reinforcing our confusion
3. we use the dominant open source solution, hadoop
4. performances are very bad so we add more computers, reinforcing our belief that we do big data magic
5. wash, rinse, repeat

## 2019 version

replace

1. big data by AI (and/or deep learning)
2. hadoop by TensorFlow



# Consequences of the confusion

## A typical story

1. we want to do predictive modelling but we confuse that with big data
2. we read tutorials and books about big data finding discussions about predictive modelling, reinforcing our confusion
3. we use the dominant open source solution, hadoop
4. performances are very bad so we add more computers, reinforcing our belief that we do big data magic
5. wash, rinse, repeat

## 2020 version

replace

1. big data by AI with deep learning
2. hadoop by PyTorch

General Introduction

The Data Science value chain

Data science skills and knowledge

## Value chain concept

- ▶ business oriented concept
- ▶ process point of view
- ▶ activity of an organization:
  - ▶ a process that receives inputs and produces outputs in a series of steps
  - ▶ each step increases the “value” of the objects it transforms

## Data science case

- ▶ natural representation in terms of steps
- ▶ “value”: information content, operability

## A possible breakdown of the data science process

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. Data management and querying
5. Data analysis

# Recurring examples

## Spam filtering

Overall goal: automatic detection of unsolicited emails

## Product recommendation

Overall goal: automatic recommendation of products to consumers

## Activity tracker

Overall goal: fitness reporting and health related advice

## A possible breakdown of the data science process

1. **Data generation and acquisition**
2. Data preprocessing
3. Data storage
4. Data management and querying
5. Data analysis

# Data generation and acquisition

## The input phase

- ▶ data collection is the first phase of any data science project
- ▶ generation and acquisition is the low level part of this collection phase

## Generation (design)

- ▶ measurements of entities
- ▶ what are the entities?
- ▶ what is measured?

## Acquisition (engineering)

- ▶ how to conduct measurements?
- ▶ how to retrieve the measurement results?

## Basic view

- ▶ email: text content + tag (acceptable/spam)
- ▶ acquisition by the email service

## More advanced generation

- ▶ leveraging email structure
  - ▶ body: main text
  - ▶ subject: text
  - ▶ sender, recipient: email addresses
  - ▶ more technical headers: date, server, authentication aspects, etc.
- ▶ usage analysis for web mail



## Generation

- ▶ shopping cart: association between consumers and products
- ▶ consumer profile: as many variable as possible (location, gender, age, etc.), previous shopping activity
- ▶ product profile: text description, image, reviews and comments, previous sales
- ▶ web site usage

## Acquisition

- ▶ shopping cart: as part of the buying process
- ▶ user profile: during registration, solicited afterwards
- ▶ product profile: as part of the inclusion of the product in the merchant offer

## Generation

- ▶ personal information: age, gender, weight, etc.
- ▶ heart rate
- ▶ temperature
- ▶ step counts
- ▶ position
- ▶ etc.

## Acquisition

- ▶ user profile: during registration
- ▶ dual approach: tracker + external device
- ▶ on tracker collection (and preprocessing!)
- ▶ periodic synchronization with a connected device and through it to a server

# Data revolution

Data has never been easier to acquire!

## Online services

- ▶ e-commerce
- ▶ software as a service: email, document editions, etc.
- ▶ log generation

## Social networks

- ▶ massive sharing
- ▶ very rich and complex data

## Open data

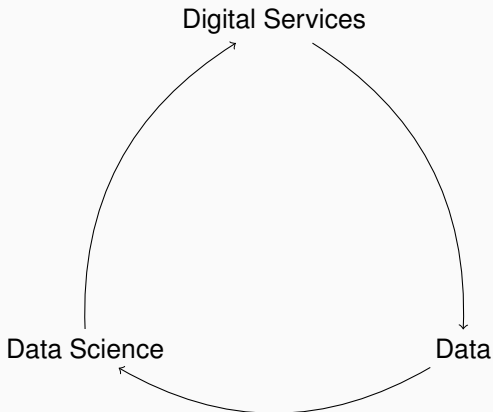
- ▶ large sources of data mainly from governmental agencies and scientific projects
- ▶ already organized, documented, etc.

Digital Services

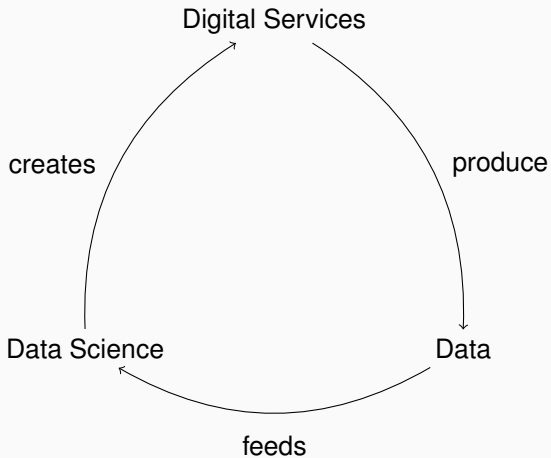
Data Science

Data

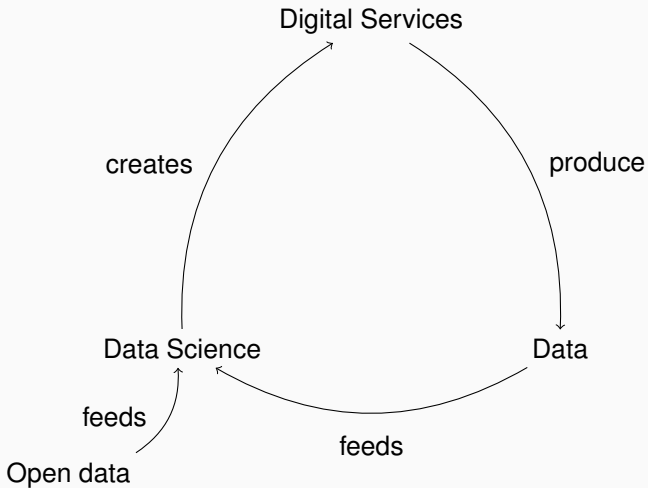
# Data Science Revolution



# Data Science Revolution



# Data Science Revolution



## Smart phones

- ▶ permanent connection
- ▶ permanent tracking
- ▶ enormous computational resources

## Internet of things

- ▶ home appliance
- ▶ smart home



# Virtuous circle of data science

## With accurate data come accurate predictions

- ▶ data quality as part of the collection
- ▶ to benefit from a service based on personal data, your personal data must be accurate!

## Examples

- ▶ spam filtering
  - ▶ manual tagging of badly classified spams
  - ▶ manual inspection of the junk folder
- ▶ product recommendation
  - ▶ actual buying/viewing act
  - ▶ reviews and ratings
- ▶ activity tracking: usage=collection!

# Generation issues

- ▶ conceptual level
- ▶ practical trade off
  - ▶ large scale collection: large scale data!
  - ▶ small scale collection: missing important data!
- ▶ conceptual trade off
  - ▶ low level data (e.g. raw data)
    - ▶ easier to measure
    - ▶ potential high acquisition costs
    - ▶ minimal risk of information loss
  - ▶ high level data (e.g. aggregated data)
    - ▶ more complex measurement strategies
    - ▶ can reduce acquisition costs and noise
    - ▶ risk of missing details

# Acquisition issues

- ▶ engineering level (mostly)
- ▶ seldom under the responsibility of a data scientist
- ▶ sensors
  - ▶ e.g. accelerometer, GPS chip
  - ▶ typical difficulties
    - ▶ noise and calibration
    - ▶ data volume
    - ▶ reliability
- ▶ distributed acquisition (e.g. via smartphones)

## Value creation

- ▶ uncollected data is lost!
- ▶ from zero to possibly something

## Value creation

- ▶ uncollected data is lost!
- ▶ from zero to possibly something

## Value creation compromise

- ▶ however data acquisition is not (always) free!
- ▶ another aspect of the trade off between large scale and focused collection
- ▶ standard strategy: transfer part of the acquisition cost to the clients, e.g.
  - ▶ email tagging, product reviews
  - ▶ activity tracker (you have to buy it!)
  - ▶ smart phone resources (local processing and connectivity)

## General Data Protection Regulation

- ▶ EU law implemented in may 2018
- ▶ restricts personal data collection and processing
  - ▶ explicit collection and redistribution
  - ▶ collect only what is needed
  - ▶ data protection (anonymization)
  - ▶ explicit consent
  - ▶ right to withdraw consent, right of access, right of portability, right to be forgotten

## Impacts

- ▶ limits personal data collection and processing
- ▶ but clarifies and simplifies certain aspects
- ▶ long term consequences are unclear
- ▶ ongoing very active research on privacy preserving data science

## A possible breakdown of the data science process

1. Data generation and acquisition
2. **Data preprocessing**
3. Data storage
4. Data management and querying
5. Data analysis

## Transformations before storage

- ▶ data science projects leverage multiple data sources
- ▶ data should be *integrated*
- ▶ but also *cleaned* or *compressed*

## Data integration

- ▶ reference formats (for instance UTC for time)
- ▶ links between data sources (e.g. mapping between user ids)



## Compressed formats

- ▶ obvious (and almost mandatory) for media applications
- ▶ audio, image and video
- ▶ especially relevant for smart phones and smart objects with limited storage and processing capabilities
- ▶ extremely important for distributed data collection

## Advanced semantic compression

- ▶ music signature (e.g. Shazam/SoundHound)
- ▶ activity detection (activity trackers)

## Sensors

- ▶ inconsistent value removal (e.g. very noisy GPS positions)
- ▶ averaging and other low level processing

## Declarative data

- ▶ bogus value detection (such as fake phone numbers)
- ▶ auto correction

## Data integration

- ▶ linking emails to web mail usage
- ▶ integration with external verification tools (e.g DKIM)
- ▶ emails shared between users

## Compression

- ▶ deduplication for shared emails
- ▶ reduced representation for deleted emails

## Data integration

- ▶ linking web usage with products and users description
- ▶ “equivalent” product detection
  - ▶ softcover/hardcover
  - ▶ collector editions
  - ▶ multiple vendors (e.g. amazon marketplace)

## Cleaning

- ▶ fake reviews
- ▶ spam in comments or reviews
- ▶ external vendors badly written product presentation

## Semantic compression

- ▶ activity detection
  - ▶ elementary (standing, walking, etc.)
  - ▶ advanced sport related
- ▶ event detection
- ▶ multiple resolution compression

## Cleaning

- ▶ GPS positions
- ▶ heart beat detection

## A broad range of methods

- ▶ from well known methods
  - ▶ standard format (UTC, WGS 84, etc.)
  - ▶ standard compression tools (mp3, jpg, h.265, etc.)
  - ▶ classical signal processing techniques (moving average)
- ▶ to sophisticated ones
  - ▶ music signature
  - ▶ activity detection
  - ▶ etc.

## Blurry distinction

- ▶ advanced preprocessing methods are data science methods!
- ▶ recursive structure: complex data collection can be seen as a full data science pipeline

## Value creation

- ▶ obvious for data cleaning
- ▶ obvious for some specific schemes
  - ▶ activity detection
    - ▶ compressed data
    - ▶ higher information content
    - ▶ limited disclosure (vendor lock-in!)
  - ▶ outlier removal
    - ▶ compressed data
    - ▶ cleaner data

## Compromise

- ▶ similar to the acquisition compromise
- ▶ user related for e.g. smart phones (e.g. battery versus bandwidth)

## A possible breakdown of the data science process

1. Data generation and acquisition
2. Data preprocessing
3. **Data storage**
4. Data management and querying
5. Data analysis



## The easy part!

- ▶ the data revolution is a consequence of the data storage revolution
- ▶ storage cost evolution (constant dollar analysis, reference 2017)
  - 1956 first HDD 5 MB for 400 K \$
  - 1995 typical HDD 1 GB for 1350 \$
  - 2005 typical HDD 200 GB for 250 \$
  - 2010 typical HDD 1 TB for 90 \$
  - 2018 typical HDD 4 TB for 110 \$
- ▶ transparent HDD aggregation is readily available
  - ▶ e.g. Synology FlashStation FS3017: 11 K\$
  - ▶ 24 HDD: 96 TB seen as a single hard drive
  - ▶ expandable to 200 TB

None!

None!

## Big data

- ▶ a bit more complicated in the big data context
- ▶ distributed storage can be more efficient than centralized one
- ▶ specific solutions
  - ▶ mixed “small” computers
  - ▶ provide processing power and storage
  - ▶ reduce the bottleneck associated to centralized storage
  - ▶ see [Apache Hadoop](#) (HDFS)

## A possible breakdown of the data science process

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. **Data management and querying**
5. Data analysis

# Data management and querying

## Using data

- ▶ raw storage of data is arguably useless
- ▶ data science is based on data manipulation
  - ▶ selections: looking at specific items
  - ▶ links: finding related objects
  - ▶ aggregations: group analysis

## RDBMS

- ▶ relational database management system
- ▶ standard and obvious solution for such needs
- ▶ minimal ease with SQL and RDBMS is **mandatory** for a data scientist
- ▶ SQL like queries are available in many systems (e.g. in Python and R)

## Definition

- ▶ non relational databases
- ▶ and also, in some cases, non reliable database systems!

## Non relational data

- ▶ some data types are not adapted to the relational model
  - ▶ graph data (social network): too much relational!
  - ▶ text data: not enough relational!
- ▶ specific database management systems have been design to handle those data type
- ▶ important complement to RDBMS

## The (pseudo) CAP theorem

a distributed system cannot be simultaneously (Brewer, circa 2000)

- ▶ consistent (a read receive the most recent write as answer)
- ▶ available (every read receive an answer)
- ▶ partition tolerant (the system operates even when messages between its parts are lost)

## NoSQL

- ▶ RDBMS are generally ACID: strong form of consistency
- ▶ some NoSQL systems achieve high performances by dropping consistency guarantees
- ▶ this induces strong risks of e.g. losing data

## Spam filtering

- ▶ raw acquisition:
  - ▶ emails in e.g. Maildir
  - ▶ log files
- ▶ post storage: RDBMS, text oriented database

## Product recommendation

- ▶ base storage in a RDBMS (strong consistency requirements!)
- ▶ complementary storage in dedicated databases (text, images, reviews)

## Activity tracker

- ▶ RDBMS for general data
- ▶ possibly specialized databases for e.g. trajectories



## Big data

- ▶ very active research field
- ▶ back to SQL with the NewSQL movement (e.g. [Google Spanner](#))

## In practice

- ▶ RDBMS as an obvious basic solution
- ▶ large scale data should be moved to dedicated databases if needed
- ▶ database design and optimization is a complex subject!

## A possible breakdown of the data science process

1. Data generation and acquisition
2. Data preprocessing
3. Data storage
4. Data management and querying
5. **Data analysis**

## Information and knowledge extraction

- ▶ data mining, analytics, machine learning, AI (!)
- ▶ from exploratory aspects
  - ▶ dashboard
  - ▶ visualization
  - ▶ clustering
- ▶ to predictive modelling
  - ▶ recommendation
  - ▶ targeted advertisement
  - ▶ filtering

## Spam filtering

- ▶ automatic tagging of incoming emails
- ▶ supervised learning

## Product recommendation

- ▶ automatic rating of products (ranking)
- ▶ frequent association (unsupervised learning)

## Activity tracker

- ▶ personal coaching (supervised learning)
- ▶ global user base analysis (unsupervised learning)

## Computer engineering

- ▶ programming
- ▶ parallel programming
- ▶ distributed programming
- ▶ system programming
- ▶ data base manipulation

## Computer science

- ▶ computational complexity
- ▶ algorithmic thinking
- ▶ relational models
- ▶ distributed systems

## Mathematics

- ▶ probability
- ▶ statistics
- ▶ optimization

- ▶ Captain Obvious image:

<https://imgur.com/gallery/PazzF>



This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>



Last git commit: 2021-01-19

By: Fabrice Rossi (Fabrice.Rossi@apiacoa.org)

Git hash: a623238c82efeb5372d8b821e0e946cfd8c918cc

- ▶ September 2019: initial version