

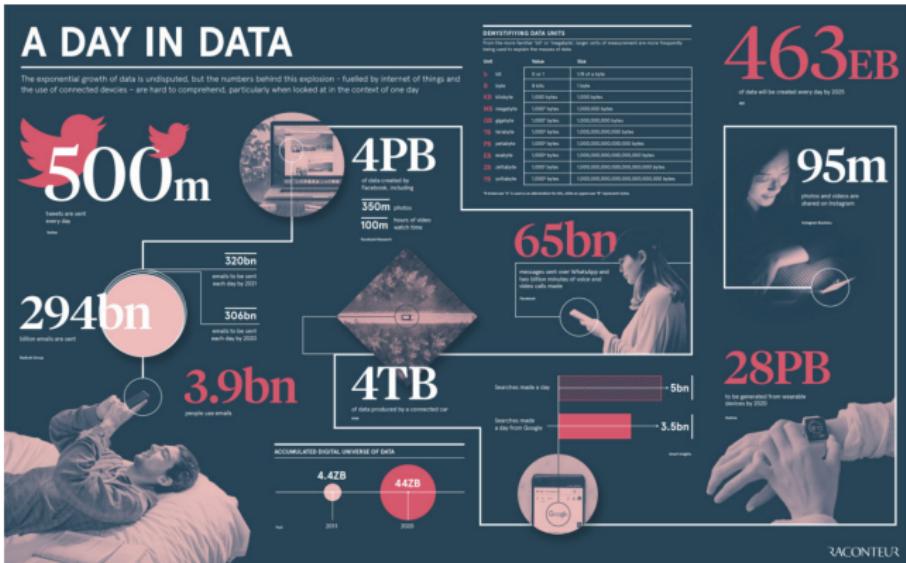
Introduction to Data Ecosystem

Erwan Scornet (Associate professor, Ecole Polytechnique)

- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

Where are the data?

Everywhere!



Good point: impossible to do a data project without data

But how much data exactly?

Wooclap: *What is the average quantity of data created per person per day in 2020?*

Some scale (on average):

- An office document (Word, PowerPoint...): 321 KB
- 1 picture with a smartphone : 10 MB
- Recording a 3 hour zoom meeting: 1 GB

But how much data exactly?

Wooclap: *What is the average quantity of data created per person per day in 2020?*

Some scale (on average):

- An office document (Word, PowerPoint...): 321 KB
- 1 picture with a smartphone : 10 MB
- Recording a 3 hour zoom meeting: 1 GB

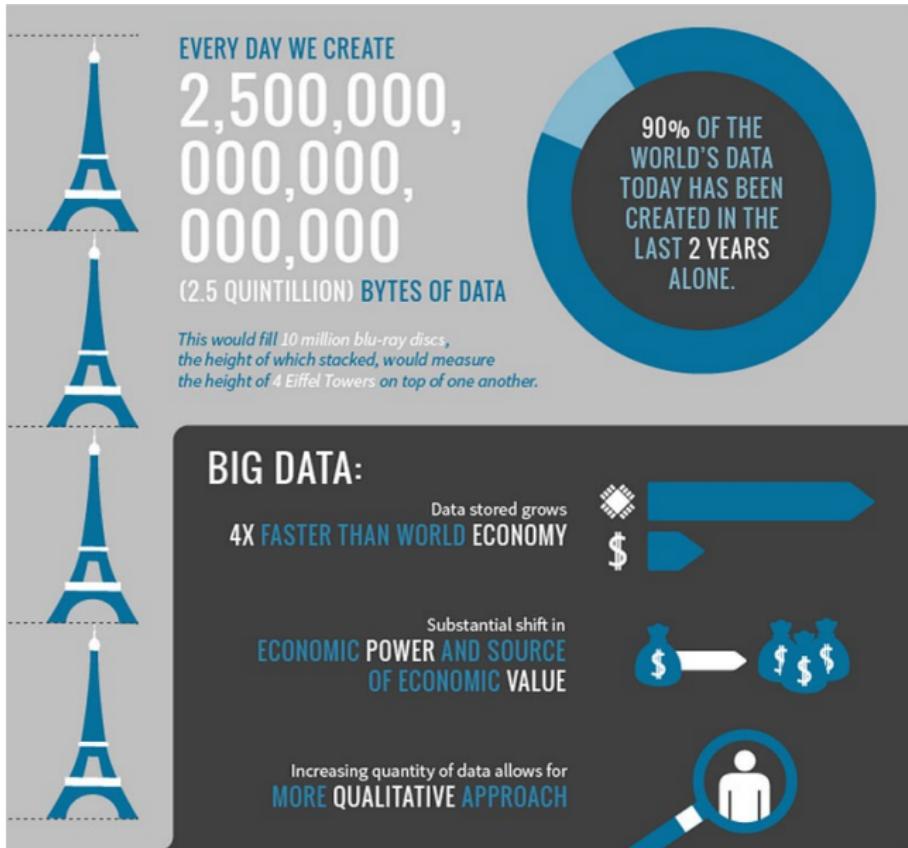


Figure: OECD, 2019

Big picture on data growth

Global Data Creation is About to Explode

Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)



@StatistaCharts

Source: Statista Digital Economy Compass 2019

statista

- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

Data Terminology



When you're fundraising, it's AI / When you're hiring, it's ML / When you're implementing, it's linear regression / When you're debugging, it's printf()

Baron Schwartz, Twitter, Nov 2017

Wooclap: Assign to each term its corresponding definition.

- Data Management
 - Business Intelligence
 - Statistics
 - Data science
 - Big data
 - Machine learning
 - Artificial Intelligence
 - Deep Learning
- is the study of the collection, analysis, interpretation, presentation and organization of data.
 - comprises the strategies and technologies used by enterprises for the data analysis of business information.
 - is part of a broader family of machine learning methods based on learning data representations, as opposed to task-specific algorithms.
 - is the study of the generalizable extraction of knowledge from data.
 - is an all-encompassing term for any collection of data sets so large and complex that it becomes difficult to process using traditional data processing applications.
 - comprises all disciplines related to handling data as a valuable resource.
 - is the subfield of computer science that gives computers the ability to learn without being explicitly programmed.
 - aims at designing and studying devices that perceive its environment and take actions that maximize its chance of success at some goal.

Artificial Intelligence - an old buzzword (Dartmouth conference)

On September 1955, a project was proposed by McCarthy, Marvin Minsky, Nathaniel Rochester and Claude Shannon introducing formally for the first time the term "[Artificial Intelligence](#)".

The study is to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves.

Proposal for Dartmouth conference on AI (1956)

Misconception of AI

AI is about electronic device able to **mimic** human thinking:

- Artificial **intelligence**
- One famous class of AI algorithms are called **neural networks**.
- **Android** are close to humans in shape so they must think like humans.

Most AI algorithms do **not** aim at **reproducing human reasoning**.

*Artificial intelligence is the science
of making machines do things that
would require intelligence if done by
men*

Marvin Minsky (1968)



Dave... I'm afraid I can't
let you do that...

2001: A Space Odyssey

Artificial Intelligence is not human intelligence

What often happens is that an engineer has an idea of how the brain works (in his opinion) and then designs a machine that behaves that way. This new machine may in fact work very well. But, I must warn you that that does not tell us anything about how the brain actually works, nor is it necessary to ever really know that, in order to make a computer very capable. It is not necessary to understand the way birds flap their wings and how the feathers are designed in order to make a flying machine [...] It is therefore not necessary to imitate the behavior of Nature in detail in order to engineer a device which can in many respects surpass Nature's abilities.

Richard Feynman (1999)

AI technology - Autonomous cars

- Originates from 1920 (NY)
- First use of neural networks to control autonomous cars (1989)
- Four US states allow self-driving cars (2013)
- First known fatal accident (May 2016)
- Singapore launched the first self-driving taxi service (Aug. 2016)
- A Arizona pedestrian was killed by an Uber self-driving car (March 2018).



AI technology - virtual assistant / chatbot

- Voice recognition tool "Harpy" masters about 1000 words (1970s, CMU, US Defense).
- System capable of analyzing entire word sequences (1980).
- Siri was the first modern digital virtual assistant installed on a smartphone (2011).
- Watson won the TV show Jeopardy! (2011)



Different uses of AI



Different uses of AI



Different uses of AI



Français	↔	Anglais
Une manière rapide, efficace et assez précise de traduire vos textes	x	A fast, efficient and fairly precise way to translate your texts

Ouvrir dans Google Traduction

Commentaires

The study found that Google Home performed the best, recognizing 98 per cent of topics accurately and providing advice that matched with Red Cross first aid guidelines 56 per cent of the time.

Alexa recognized 92 per cent of topics, and gave appropriate advice 19 per cent of the time.

The responses from Siri and Cortana were so low that researchers determined that they couldn't analyze them.

Different uses of AI

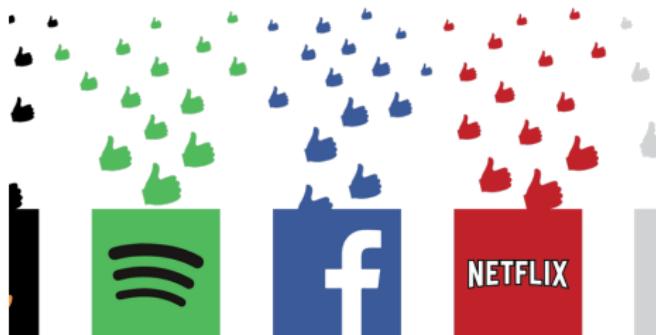


French ▾ x English ▾

Une manière rapide, efficace et assez précise de traduire vos textes

A fast, efficient and fairly precise way to translate your texts

Ouvrir dans Google Traduction Commentaires

A screenshot of a Google Translate interface. It shows a comparison between French text ("Une manière rapide, efficace et assez précise de traduire vos textes") and its English translation ("A fast, efficient and fairly precise way to translate your texts"). The interface includes language selection dropdowns and standard translation controls.

Different uses of AI



French ▾

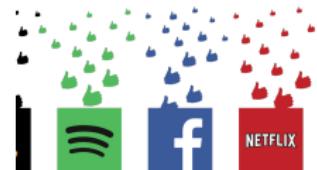
Anglais ▾

Une manière rapide, efficace et assez précise de traduire vos textes

A fast, efficient and fairly precise way to translate your texts

Google Traduction

Commentaires



NEWS - 30 OCTOBER 2019

Google AI beats top human players at strategy game *StarCraft II*

DeepMind's AlphaStar beat all but the very best humans at the fast-paced sci-fi video game.

Different uses of AI



French ▾

Anglais ▾

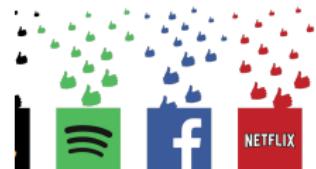
Une manière rapide, efficace et assez précise de traduire vos textes

A fast, efficient and fairly precise way to translate your texts

Google dans Google Traduction

Commentaires

This block shows a comparison between French and English text using Google Translate. The French text "Une manière rapide, efficace et assez précise de traduire vos textes" is on the left, and the English translation "A fast, efficient and fairly precise way to translate your texts" is on the right. Below the text, there are buttons for "Google dans Google Traduction" and "Commentaires".

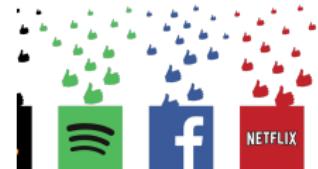


AI artwork sells for \$432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer a work of art created by an algorithm

Different uses of AI



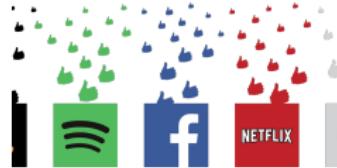
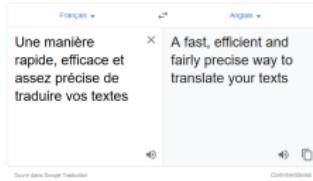
A screenshot of the Google Translate mobile application. The interface is split into two sections: the left section shows text in French ("Une manière rapide, efficace et assez précise de traduire vos textes"), and the right section shows the same text translated into English ("A fast, efficient and fairly precise way to translate your texts"). The bottom of the screen includes standard mobile navigation icons.



524 views | Jan 16, 2020, 08:00am

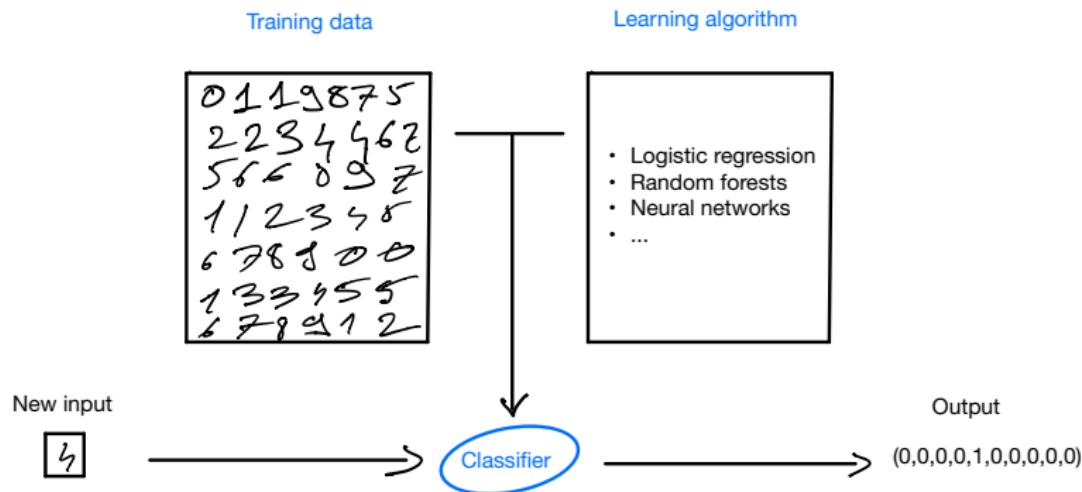
**A Look Inside
Augmented Analytics
And Its Business Value In
2020**

Different uses of AI



- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

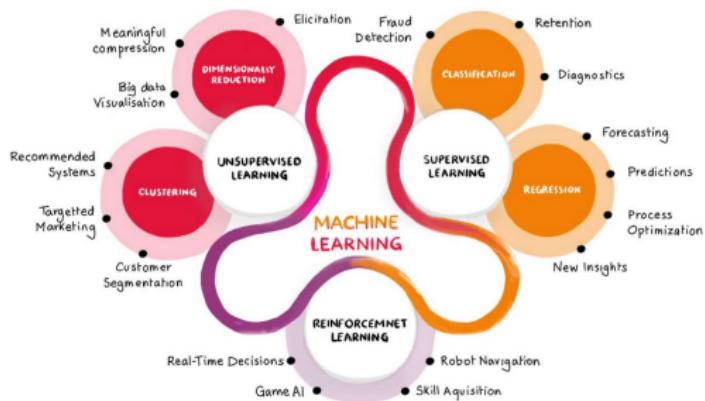
Supervised learning



A definition by Tom Mitchell (<http://www.cs.cmu.edu/~tom/>)

A computer program is said to learn from **experience E** with respect to some **class of tasks T** and **performance measure P**, if its performance at tasks in T , as measured by P, improves with experience E.

Three Kinds of Learning



Unsupervised Learning

- Task:** Clustering/DR
- Performance:** Quality
- Experience:** Raw dataset
(No Ground Truth)

Supervised Learning

- Task:** Prediction
- Performance:** Average error
- Experience:** Predictions
(Ground Truth)

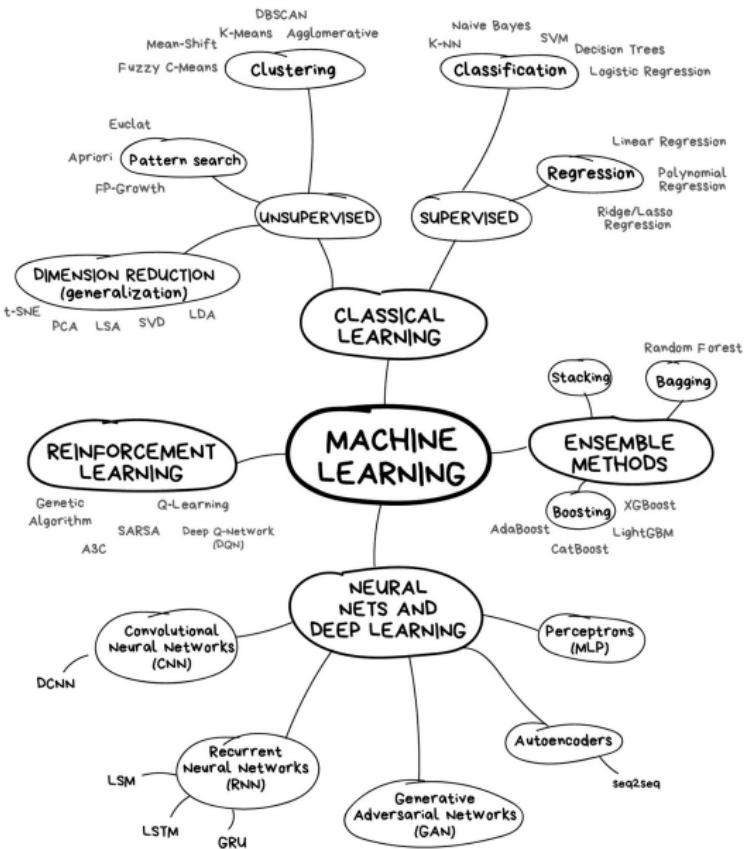
Reinforcement Learning

- Task:** Action
- Performance:** Total reward
- Experience:** Reward from env.
(Interact. with env.)

- Timing:** Offline/Batch (learning from past data) vs Online (continuous learning)

Figure Source: BCG

Algorithms



Difficulties related to (Big) data

- The prediction must be **accurate**: difficult for some tasks like image classification, video captioning...
- Predictions must be **fast**: online recommendation should not take minutes.
- Data must be **stored** and **easily accessible**.
- It may be difficult to **access all data simultaneously**. Data may come sequentially.
- Data must be **clean**.
- Data should be **relevant**.



Wooclap: *How would you evaluate the performance of an ML algorithm aiming at diagnosing a patient?*

- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

Cost of storing data

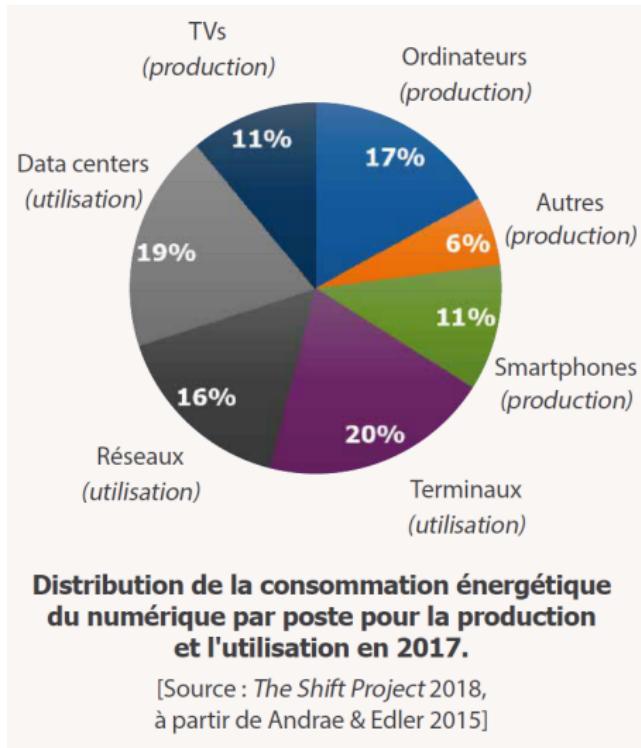
Cost of storing data

- Money: 300.000 US dollars in Google Cloud to store 1 petabyte during one year.

Cost of storing data

- Money: 300.000 US dollars in Google Cloud to store 1 petabyte during one year.
- Data centers and environment
 - 2% of the total electricity consumption in the US.
 - 626 billion liters of water.
 - 2% of total global greenhouse emissions.





The Shift Project <https://theshiftproject.org/lean-ict/>

Female drivers and right front passengers are approximately

**17 percent more likely
to be killed**

in a car crash than a male occupant of the same age.

Any seatbelt-wearing female vehicle occupant has

**73 percent greater odds of being
seriously injured**

in a frontal car crash than the odds of a seatbelt-wearing male occupant being injured in the same kind and severity of crash.

Sources: NHTSA and the journal Traffic Injury Prevention

Analysis of crash and injury data compiled from the National Automotive Sampling System Crashworthiness Data System for the years 1998 to 2015.

Bias in crash test



<https://www.consumerreports.org/car-safety/crash-test-bias-how-male-focused-testing-puts-female-drivers-at-risk/>

What is COMPAS?

Correctional Offender Management Profiling for Alternative Sanction used in US justice courts to predict the reoffending probability.



What is COMPAS?

Correctional Offender Management Profiling for Alternative Sanction used in US justice courts to predict the reoffending probability.

Assessing the fairness of COMPAS

(A) Calibration

Given a score, the percentage of black people who reoffend is the same as the percentage of white people who reoffend.

(B) Parity - False Positive rate

The false positive rates (probability of being classified at risk while being not at risk) are the same for the group of black people and white people.

(C) Parity - False Negative rate

The false negative rates (probability of being classified not at risk while being at risk) are the same for the group of black people and white people.

- (A) According to Northpoint, **COMPAS is calibrated**.

Among defendants who scored a seven on the COMPAS scale, 60 percent of white defendants reoffended, which is nearly identical to the 61 percent of black defendants who reoffended.

- (B) According to ProPublica, **COMPAS does not satisfy parity** for false Positive rate.

Among defendants who ultimately did not reoffend, blacks were more than twice as likely as whites to be classified as medium or high risk (42 percent vs. 22 percent).

- (C) Parity - False Negative rate

Theorem: assume that reoffending cannot be **exactly** predicted via the input features (life is always a bit random), then there is no algorithm that satisfies (A), (B), (C).

Washington Post : A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear.

- Car crash tests lead to unfair vehicles.
 - **Debias the data** : collect more, better quality, better representativity.
- Correctional Offender Management Profiling for Alternative Sanctions (Compas) used in the US.
 - **Debias the algorithm** : twist predictions to annihilate one bias.
- Social Credit System / DeepNude
 - **Impact on society** : do we want these algorithms in our life?

- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

Wooclap: *Order the following different steps of a data project.*

- Data collection
- Model evaluation
- Predictive modeling
- Continuous Optimization
- Solution Deployment
- Data Wrangling (gathering data in a usable format)
- Business understanding
- Testing / validation

You can also mention how each step interacts with the others.

Data Scientists and Challenges

MODERN DATA SCIENTIST

Data Scientist, the sexiest job of 21st century requires a mixture of multidisciplinary skills ranging from an intersection of mathematics, statistics, computer science, communication and business. Finding a data scientist is hard. Finding people who understand who a data scientist is, is equally hard. So here is a little cheat sheet on who the modern data scientist really is.

MATH & STATISTICS

- ★ Machine learning
- ★ Statistical modeling
- ★ Experiment design
- ★ Bayesian inference
- ★ Supervised learning: decision trees, random forests, logistic regression
- ★ Unsupervised learning: clustering, dimensionality reduction
- ★ Optimization: gradient descent and variants

PROGRAMMING & DATABASE

- ★ Computer science fundamentals
- ★ Scripting language e.g. Python
- ★ Statistical computing package e.g. R
- ★ Databases SQL and NoSQL
- ★ Relational algebra
- ★ Parallel databases and parallel query processing
- ★ MapReduce concepts
- ★ Hadoop and Hive/Pig
- ★ Custom reducers
- ★ Experience with clouds like AWS

DOMAIN KNOWLEDGE & SOFT SKILLS

- ★ Passionate about the business
- ★ Curious about data
- ★ Influence without authority
- ★ Hacker mindset
- ★ Problem solver
- ★ Strategic, proactive, creative, innovative and collaborative

COMMUNICATION & VISUALIZATION

- ★ Able to engage with senior management
- ★ Story telling skills
- ★ Translate data driven insights into decisions and actions
- ★ Visual art design
- ★ R packages like ggplot or lattice
- ★ Knowledge of any visualization tools e.g. Flare, D3.js, Tableau

Data Scientist

- Mix of various skills.
- Hard to be an expert of everything!

Different occupations in a data project

Wooclap: Assign the following job names to the job descriptions below.

Data engineer, business analyst, statistician, data and analytics manager, data scientist, data architect, data analyst.

AS RARE AS UNICORNS

Role
Cleans, massages and organizes [big] data

Mindset
Curious data wizard

Skills & Talents

- Distributed computing
- Predictive modeling
- Story-telling and visualizing
- Math, Stats, Machine Learning

Languages
R, SAS, Python, Matlab, SQL, Hive, Pig, Spark

HISTORIC LEADERS OF DATA

Role
Collects, analyzes and interprets qualitative as well as quantitative data with statistical theories and methods

Mindset
Logical and enthusiastic stats genius

Skills & Talents

- Statistical theories & methodology
- Data mining & machine learning
- Distributed Computing (Hadoop)
- Database systems (SQL and NoSQL based)
- Cloud tools

Languages
R, SAS, SPSS, Matlab, Stata, Python, Perl, Hive, Pig, Spark, SQL

DATA DETECTIVE

Role
Collects, processes and performs statistical data analyses

Mindset
Intuitive data junkie with high "figure-it-out" quotient

Skills & Talents

- Spreadsheet tools (e.g. Excel)
- Database systems (SQL and NoSQL based)
- Communication & visualization
- Math, Stats, Machine Learning

Languages
R, Python, HTML, JavaScript, C/C++, SQL

THE CONTEMPORARY DATA MODELLER

Role
Creates blueprints for data management systems to integrate, centralize, protect and maintain data sources

Mindset
Inquiring ninja with a love for data architecture/design patterns

Skills & Talents

- Data warehousing solutions
- In-depth knowledge of database architecture
- Extraction Transformation and Load(ETL) spreadsheet and BI tools
- Data modeling
- Systems development

Languages
SQL, XML, Hive, Pig, Spark

SOFTWARE ENGINEERS BY TRADE

Role
Develops, constructs, tests and maintains architectures (such as databases and large scale processing systems)

Mindset
All-purpose everyman

Skills & Talents

- Database systems (SQL & NO SQL based)
- Data modeling & ETL tools
- Data APIs
- Data warehousing solutions

Languages
SQL, Hive, Pig, R, Matlab, SAS, SPSS, Python, Java, Ruby, C++, Perl

CHANGE AGENT

Role
Improves business process as intermediary between business and IT

Mindset
Resilient project juggler

Skills & Talents

- Basic tools (e.g. MS Office)
- Data visualization tools (e.g. Tableau)
- Conscious learning and storyelling
- Business intelligence understanding
- Data modeling

Languages
SQL

DATA SCIENCE TEAM LEADER

Role
Manages a team of analysts and data scientists

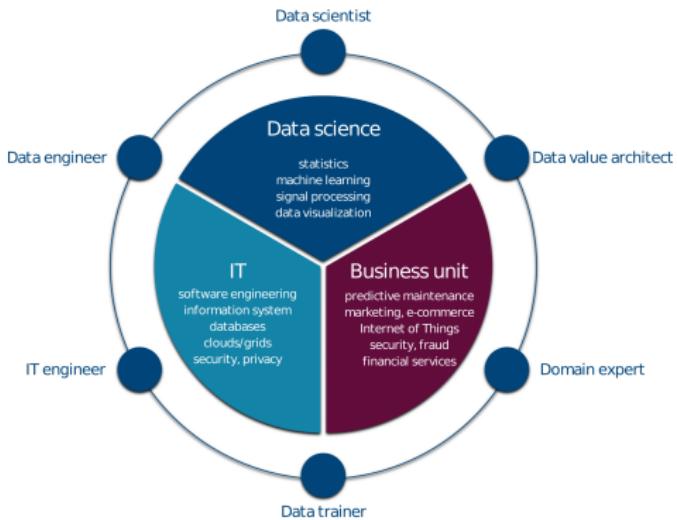
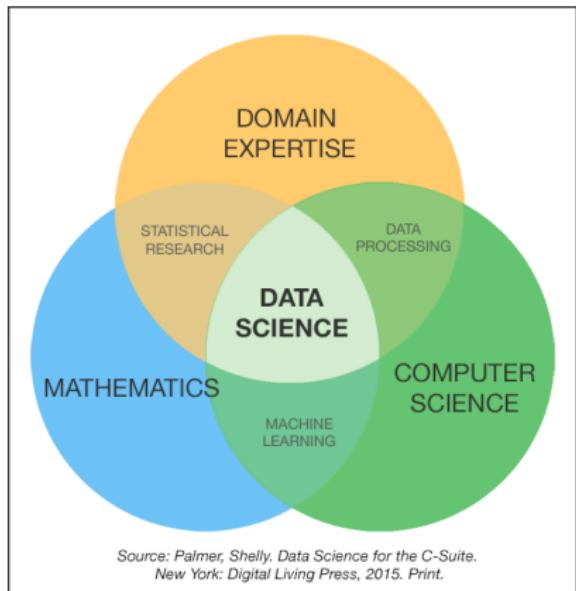
Mindset
Data Wizard's Cheerleader

Skills & Talents

- Database systems (SQL and ND SQL based)
- Leadership & project management
- Interpersonal communication
- Data mining & predictive modeling

Languages
SQL, R, SAS, Python, Matlab, Java

Different training and different occupations



- 1 No data project without data
- 2 Data science, Machine Learning, Artificial Intelligence... Which term should we use?
- 3 Different applications of AI
- 4 How does Machine Learning work?
- 5 Limitations of data projects
- 6 Data Project Organization
- 7 Unveiling the mystery of Deep Learning

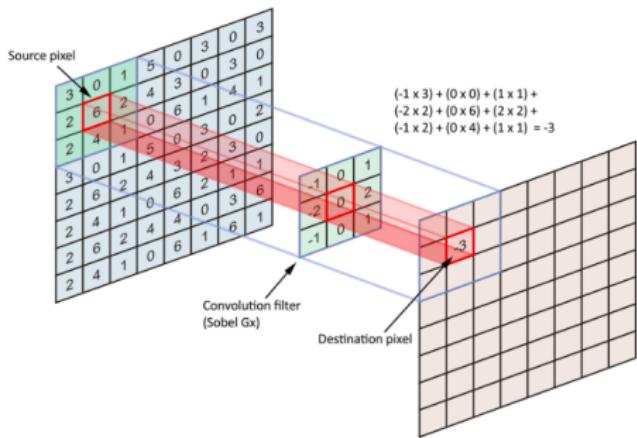
Problem

Number Recognition

0	0	0	0	0
1	1	1	1	1
2	2	2	2	2
3	3	3	3	3
4	4	4	4	4
5	5	5	5	5
6	6	6	6	6
7	7	7	7	7
8	8	8	8	8
9	9	9	9	9

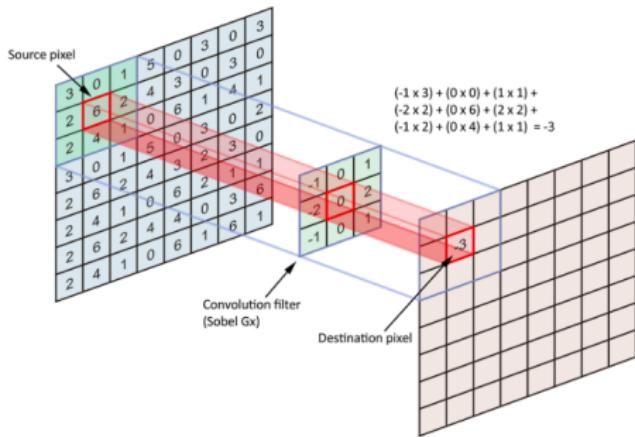
- Data: Annotated database of images (each image is represented by a vector of $28 \times 28 = 784$ pixel intensities)
- Input: Image
- Output: Corresponding number

Fundamental elements

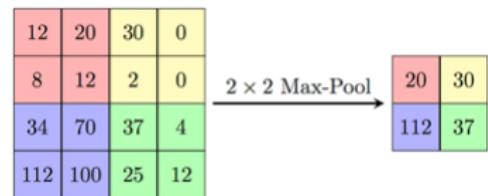


Convolution

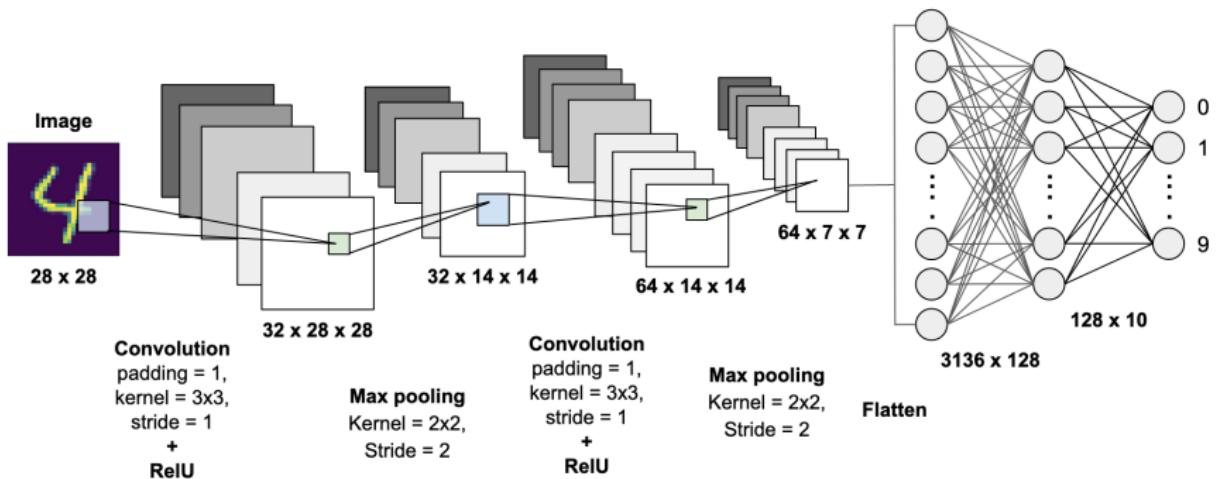
Fundamental elements



Convolution



Convolutional neural network



Results



The 82 patterns misclassified by LeNet5. Below each image is displayed the correct answer (left) and the prediction (right). These errors are mostly caused by genuinely ambiguous patterns, or by digits written in a style that are underrepresented in the training set.

Other generic applications of CNN



[Krizhevsky 2012]



[Ciresan et al. 2013]



[Faster R-CNN - Ren 2015]



[NVIDIA dev blog]

Far from terminator

- Stephen Hawking BBC, Dec 2 2014

The development of full artificial intelligence could spell the end of the human race. We cannot quite know what will happen if a machine exceeds our own intelligence, so we can't know if we'll be infinitely helped by it, or ignored by it and sidelined, or conceivably destroyed by it.



Take-home messages

- No data projects without data
 - A lot of data are available in the world
 - Difficulty of gathering the relevant ones and cleaning them (70% of the data project)
 - Environmental/Technical point of view: collecting/creating data is expensive for the planet (and the company)
- Different terms used in a data project
 - Keep in mind that AI has nothing to do with intelligence.
 - AI does not mimic human reasoning

Take-home messages

- No data projects without data
 - A lot of data are available in the world
 - Difficulty of gathering the relevant ones and cleaning them (70% of the data project)
 - Environmental/Technical point of view: collecting/creating data is expensive for the planet (and the company)
- Different terms used in a data project
 - Keep in mind that AI has nothing to do with intelligence.
 - AI does not mimic human reasoning
- How does Machine learning works?
 - Machine learning requires data to detect and learn patterns in the data.
 - Different tasks can be solved depending on the data (supervised, unsupervised, images, texts...)
 - Different tasks cannot be solved with ML notably if relevant information are not inside the collected data
 - Specific questions require specific data

- Limitations of ML
 - Data may be biased because our world is, and data are nothing but a reflection of it.
 - Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
 - ML may have trouble to adapt in temporal changes. ML has a tendency to reproduce the past.

Take-home message II

- Limitations of ML
 - Data may be biased because our world is, and data are nothing but a reflection of it.
 - Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
 - ML may have trouble to adapt in temporal changes. ML has a tendency to reproduce the past.
- Data cycle and Data jobs
 - A data cycle is composed of iterations, nothing is ever over.
 - Business analysis is very important through the cycle
 - Many different actors are involved in a data project
 - Good communication is required!

Take-home message II

- Limitations of ML
 - Data may be biased because our world is, and data are nothing but a reflection of it.
 - Detecting and removing these biases is tricky but very important if individuals are impacted by the ML solution.
 - ML may have trouble to adapt in temporal changes. ML has a tendency to reproduce the past.
- Data cycle and Data jobs
 - A data cycle is composed of iterations, nothing is ever over.
 - Business analysis is very important through the cycle
 - Many different actors are involved in a data project
 - Good communication is required!
- Data Science is evolving constantly.
 - New opportunities appear
 - New challenges are detected
 - Need for adaptability



Thank you!