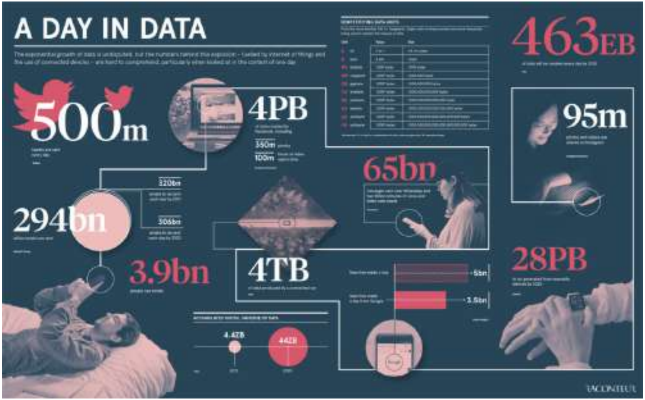
Overview: what is behind data science?

Erwan Scornet (Associate professor, Ecole Polytechnique)

Outline

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

The quantity of data produced each day per person in 2020 is 2GB which is equivalent to

6h of zoom meeting recording *or* 200 pictures *or* 6.000 Office documents

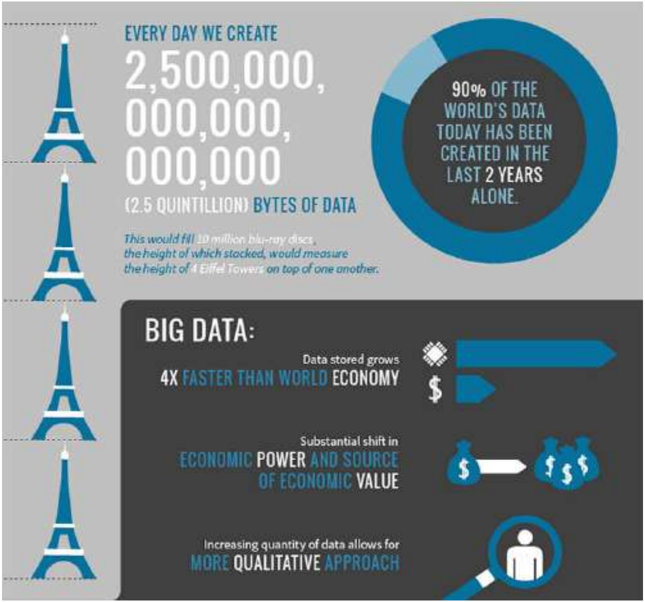
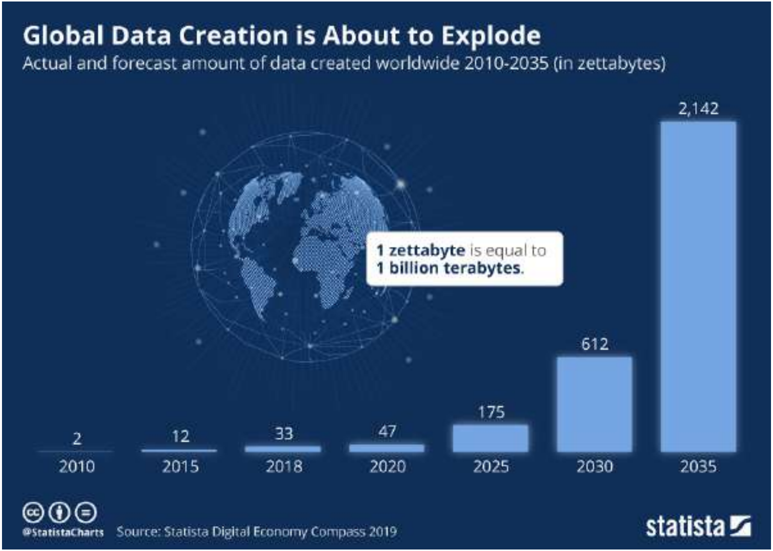
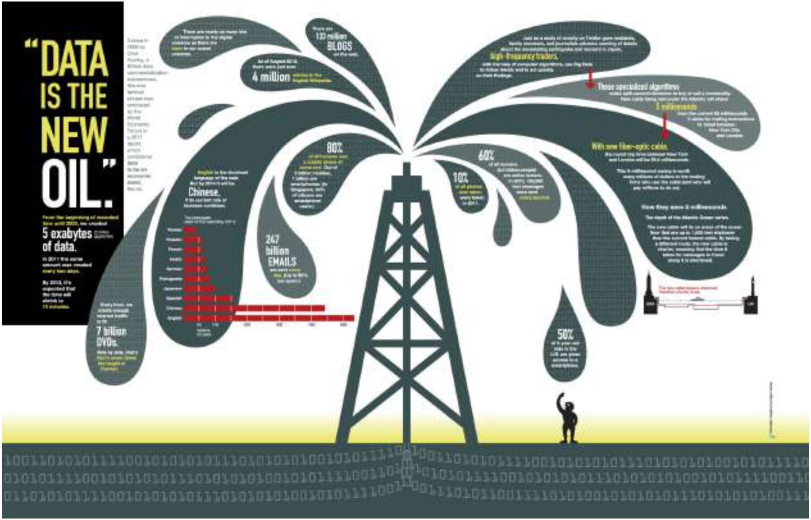


Figure: OECD, 2019

Big picture on data growth



Data is power



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

Question

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.

Data terminology

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

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)

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

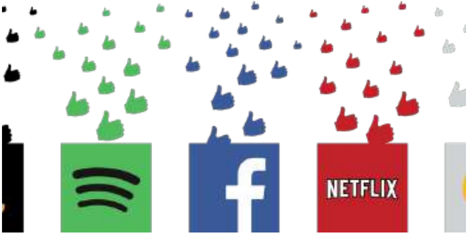
Different uses of AI

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.

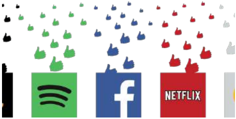
Alexa recognized 92 per cent of topics, and gave appropriate advice 19 per cent of the time.

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

Different uses of AI



Different uses of AI

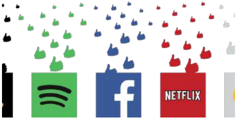


NEWS · 30 OCTOBER 2019

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

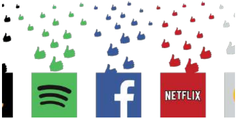
DeepMind's AlphaStar beat all but the very best humans at the fast-paced sci-fi video game.

Different uses of AI



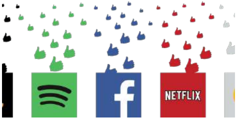
AI artwork sells for $432,500 — nearly 45 times its high estimate — as Christie's becomes the first auction house to offer a work of art created by an algorithm

Different uses of AI



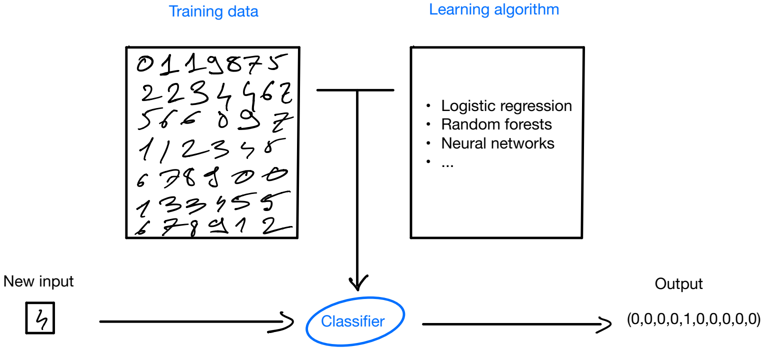
A Look Inside Augmented Analytics And Its Business Value In 2020

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.

Unsupervised Learning

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

Supervised Learning

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

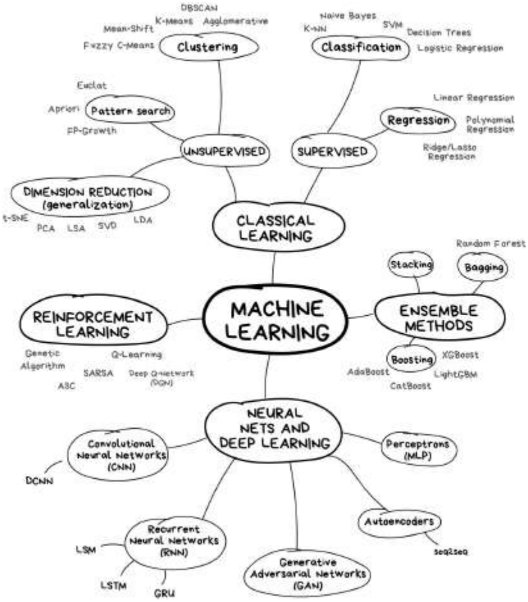
Reinforcement Learning

- **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



MACHINE LEARNING
CLASSICAL LEARNING
UNSUPERVISED
SUPERVISED
Clustering
DBSCAN
K-Means
Agglomerative
Mean-Shift
Fuzzy C-Means
Classification
Naive Bayes
K-NN
SVM
Decision Trees
Logistic Regression
Pattern search
Euclat
Apriori
FP-Growth
Regression
Linear Regression
Polynomial Regression
Ridge/Lasso Regression
DIMENSION REDUCTION (generalization)
t-SNE
PCA
LSA
SVD
LDA
REINFORCEMENT LEARNING
Genetic Algorithm
Q-Learning
SARSA
Deep Q-Network (DQN)
A3C
ENSEMBLE METHODS
Stacking
Bagging
Random Forest
Boosting
XGBoost
AdaBoost
LightGBM
CatBoost
NEURAL NETS AND DEEP LEARNING
Convolutional Neural Networks (CNN)
DCNN
Perceptrons (MLP)
Recurrent Neural Networks (RNN)
LSM
LSTM
GRU
Generative Adversarial Networks (GAN)
Autoencoders
seq2seq

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?*

Cost of storing data

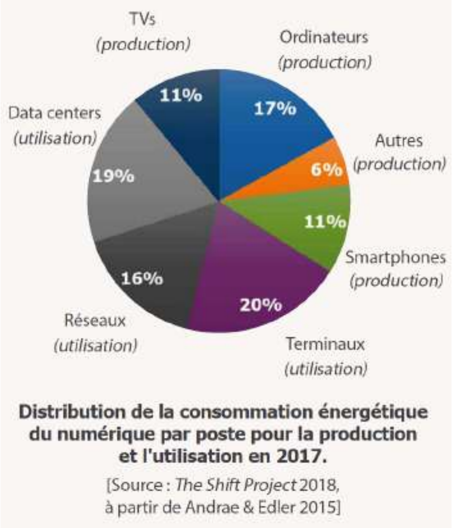
Cost of storing data

- 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.

Environmental issues



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

Car accidents

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/

COMPAS

What is COMPAS?

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

COMPAS

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.

COMPAS

(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.

Ethical issues - Bias

- Car crash tests lead to unfair vehicles.

  $\rightarrow$ **Debias the data** : collect more, better quality, better representativity.

- Correctional Offender Management Profiling for Alternative Sanctions (Compas) used in the US.

  $\rightarrow$ **Debias the algorithm** : twist predictions to annihilate one bias.

- Social Credit System / DeepNude

  $\rightarrow$ **Impact on society** : do we want these algorithms in our life?

Steps of a data project

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 Project Framework

Figure: http://www.anovaanalytics.com/data-science-consulting/

Data Science in 1 Slide

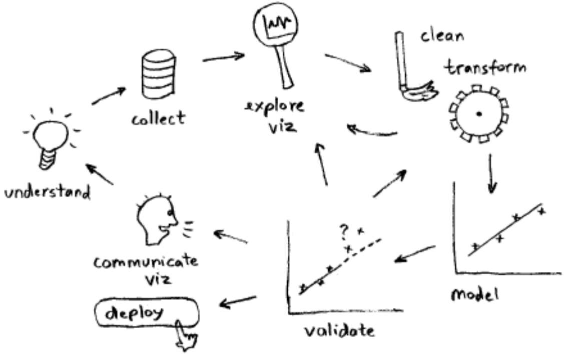
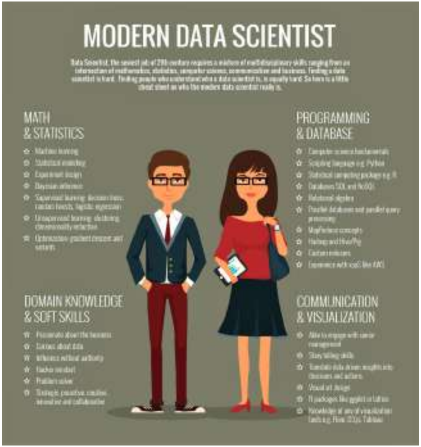


Figure: Source: Sz. Pafka

CRISP-DM

- CRoss-Industry Standard Process for Data Mining (1999)
- Adapted by Szilard Pafka.

Data Scientists and Challenges

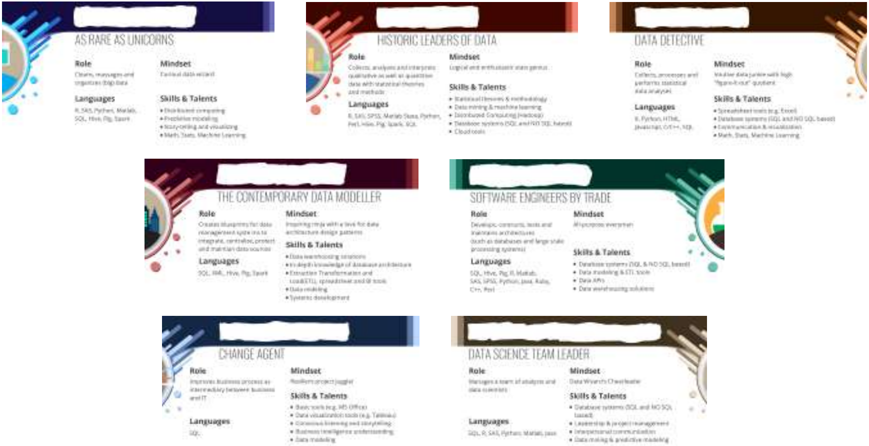


Data Scientist

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

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.

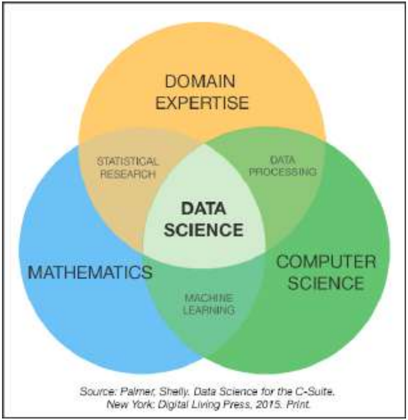


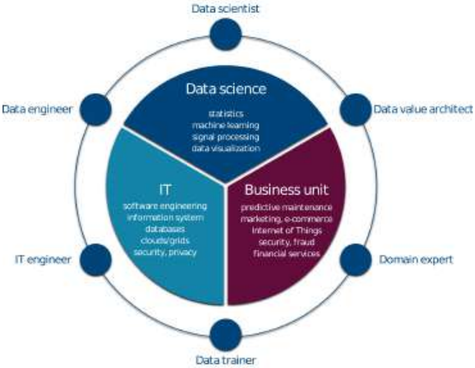
Different occupations in a data project

Several Profiles

- Several kind of problem / several kind of tools
- Much more variety than this...
- Importance of balanced teams.

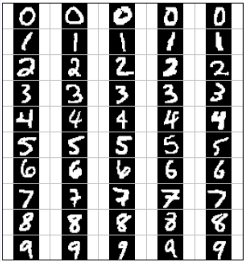
Different training and different occupations





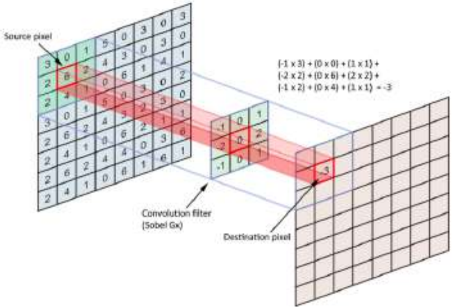
Problem

Number Recognition

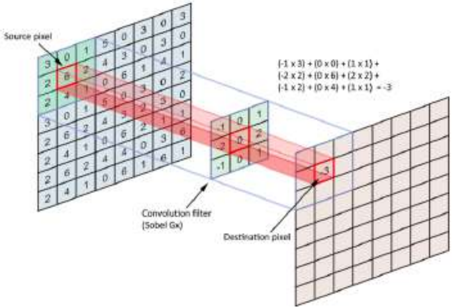


- Data: Annotated database of images (each image is represented by a vector of $28 \times 28 = 784$ pixel intensities)
- Input: Image
- Output: Corresponding number

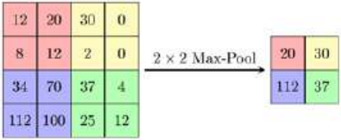
Fundamental elements



Convolution

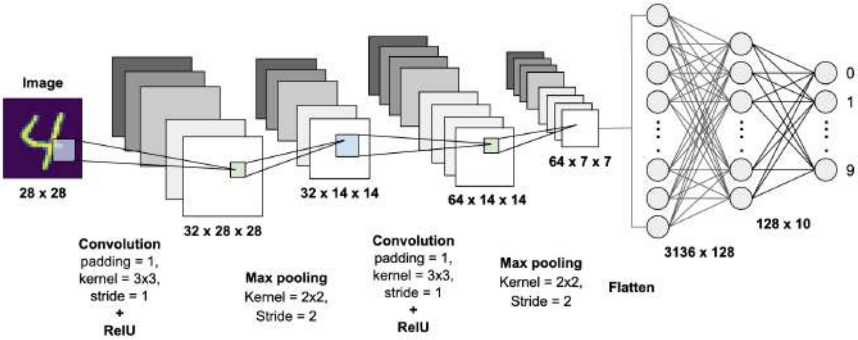


Convolution



Max-Pooling

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 under represented 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

Take-home message II

- Limitations of ML
  - Data may be biased because our wolrd 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 wolrd 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 wolrd 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!