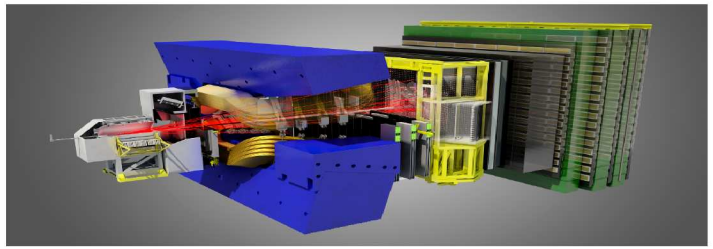
Selected topics in Data Science

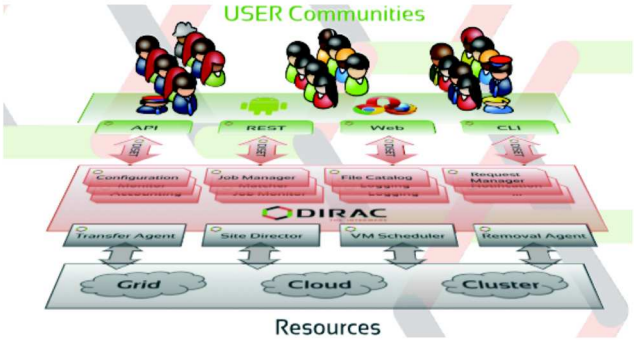
Grandjean, Martin (2014).
"La connaissance est un réseau".
Les Cahiers du Numérique

Wojciech Krzemień

16.10 2020, NCBJ

About me





- Wojciech Krzemień
- Pasteura 7, 4th floor, room 404
- contact via e-mail: wojciech.krzemien@ncbj.gov.pl
- website: http://koza.if.uj.edu.pl/~krzemien

Course scope

Selected topics in Data Science

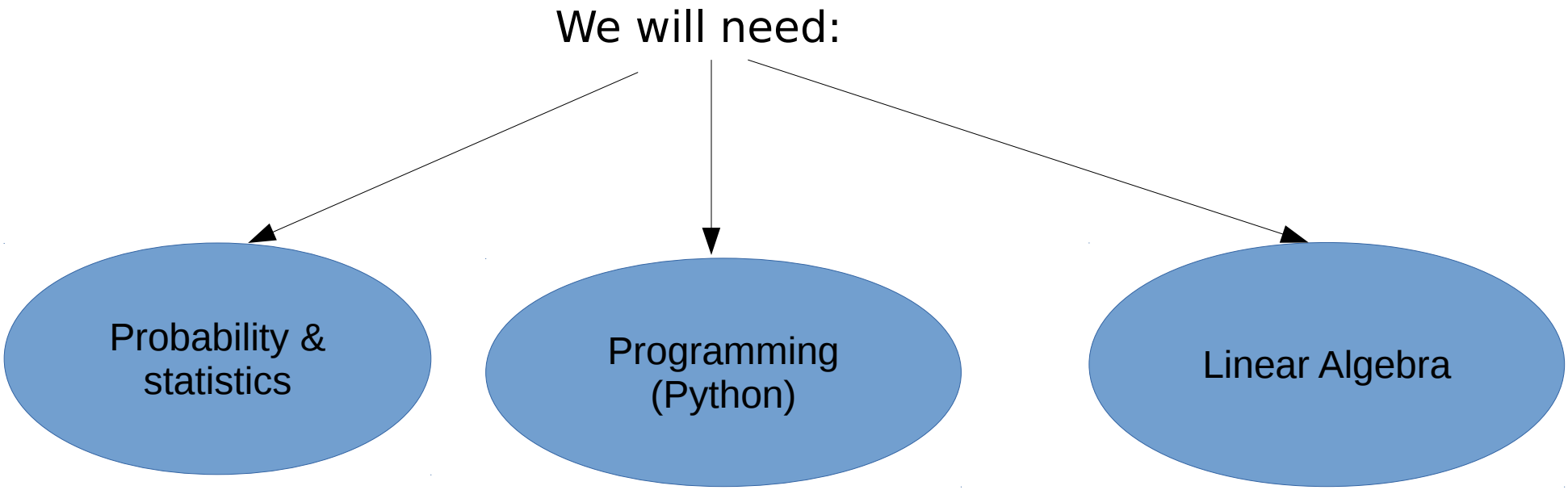
List of topics:

- Machine learning, A.I. techniques, regression vs classification
- Supervised and Unsupervised learning,
- K-nearest neighbors,
- Linear Regression,
- Logistic Regression,
- Neural Networks,
- Decision Trees and Random Forests,
- Models evaluations: bias-variance trade-off, ROC curves,
- Monte Carlo simulations,
- Clustering, K-means clustering,
- Bayesian classifiers
- Support Vector Machines,
- Kernel density estimation
- Dimensionality reduction – Principal Component Analysis,
- Optimization techniques, genetic algorithms,
- ...

Only selected topics will be covered – some of the topics could serve as a separated semester courses e.g MC simulations or Neural Networks.

Main idea of the course

- Discuss key **statistical** ideas on which studied methods are based.
- Understand how it works by **implementing** a simplified version of algorithms in **Python**.



Remark:
We will not focus on performance of our programs. In practice, always try to use the implementations from official software packages.

Course organization

- place: ~~Pasteur 7, room 404~~ (remotely with Microsoft Teams)
- time: Fridays, 14.30-17.30 (to be discussed)
- Info & material:

  http://koza.if.uj.edu.pl/~krzemien/machine_learning2021/
- password: ML2021
- format: lectures & classes

If you spot any bug or error or even typo, please let me know.

Evaluation

- Programming projects
- Problems to solve
- Extra points from quizzes during lectures
- Final exam

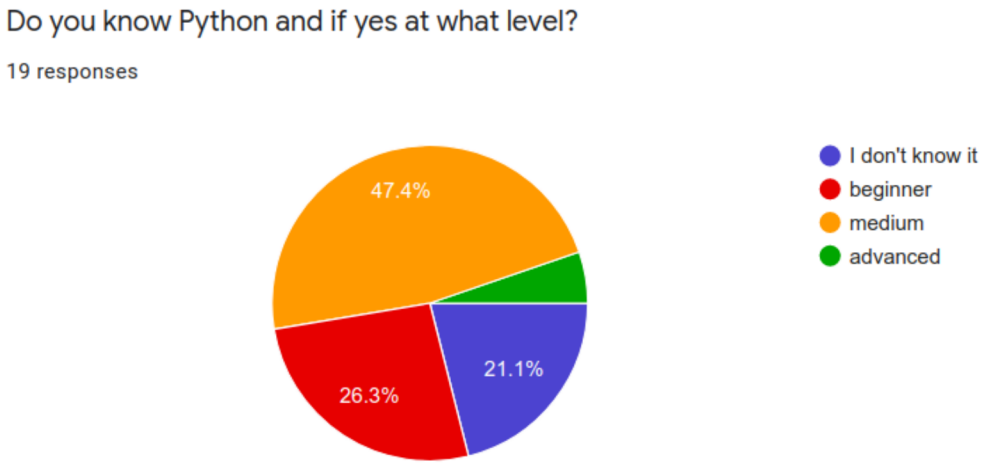
Bibliography & Materials

- „The Elements of Statistical Learning” - Trevor Hastie, Robert Tibshirani, Jerome Friedman
- „Programming Collective Intelligence” Toby Segaran
- „Statistical Analysis Techniques in Particle Physics” Ilya Narsky, Frank C. Porter
- „Neural Networks and Machine Learning” Simon Haykin
- „Python Machine Learning” Sebastian Raschka
- „Statistical data analysis” Glen Cowan
- Some additional materials will be available on the website:

  http://koza.if.uj.edu.pl/~krzemien/machine_learning2021/

Todays plan

- Introduction lecture + first ML algorithm
- Crash course in Python
- Crash course in Statistics (part I)



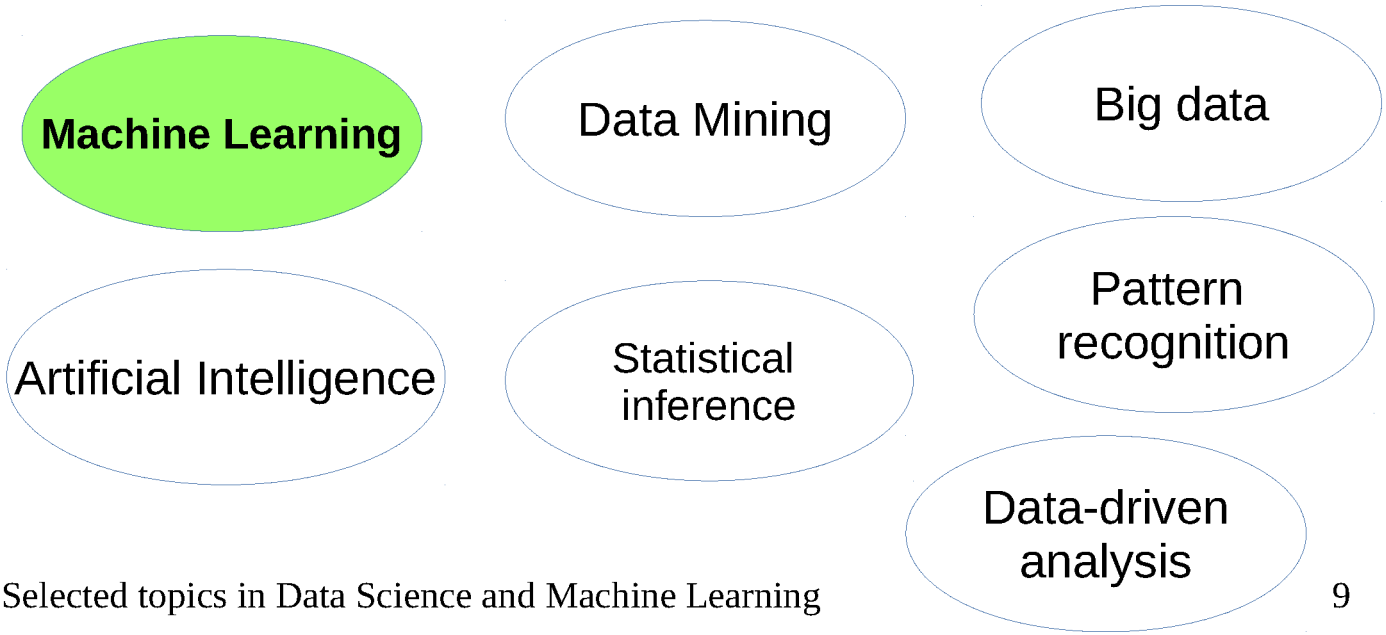
What is Data Science?

Definition from Wikipedia (https://en.wikipedia.org/wiki/Data_science):

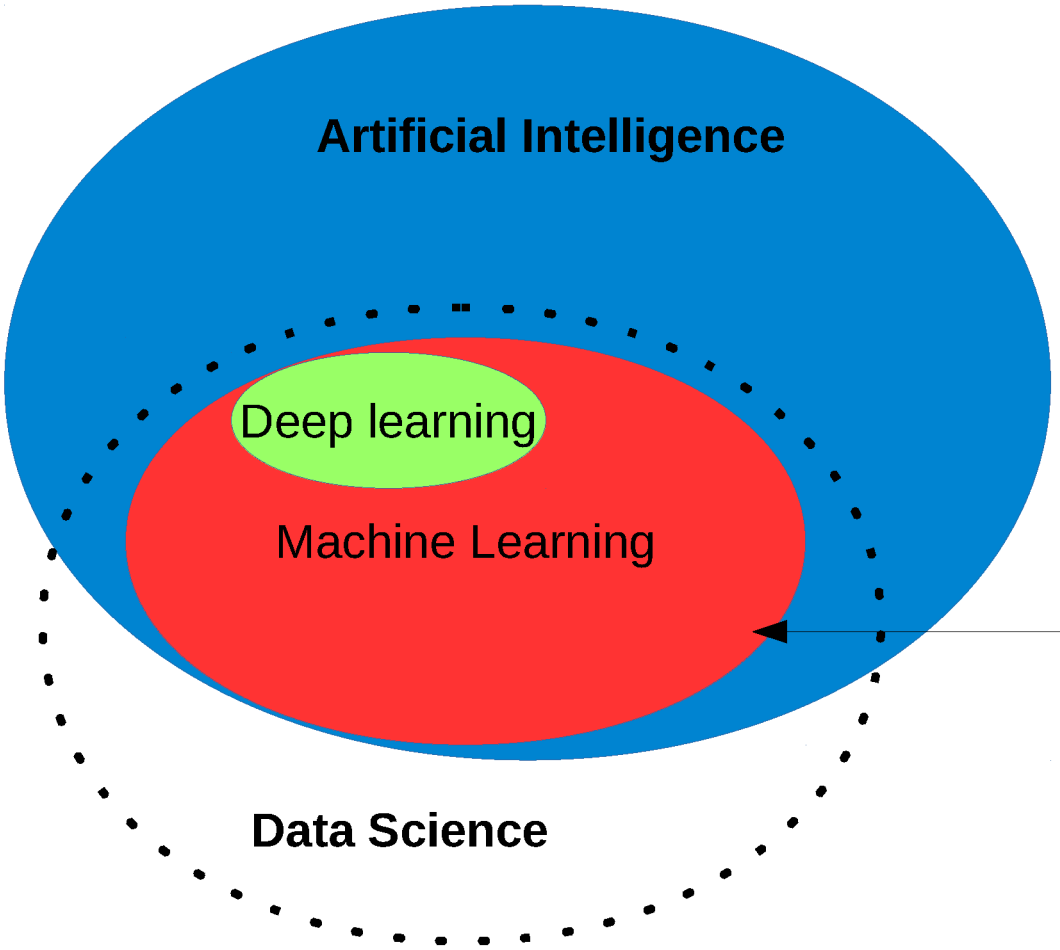
*„**Data science**, also known as **data-driven** science, is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights **from data** in various forms, either **structured** or **unstructured**, similar to data mining.(...)"*

Data Scientist: The Sexiest Job of the 21st Century

Other buzzwords:



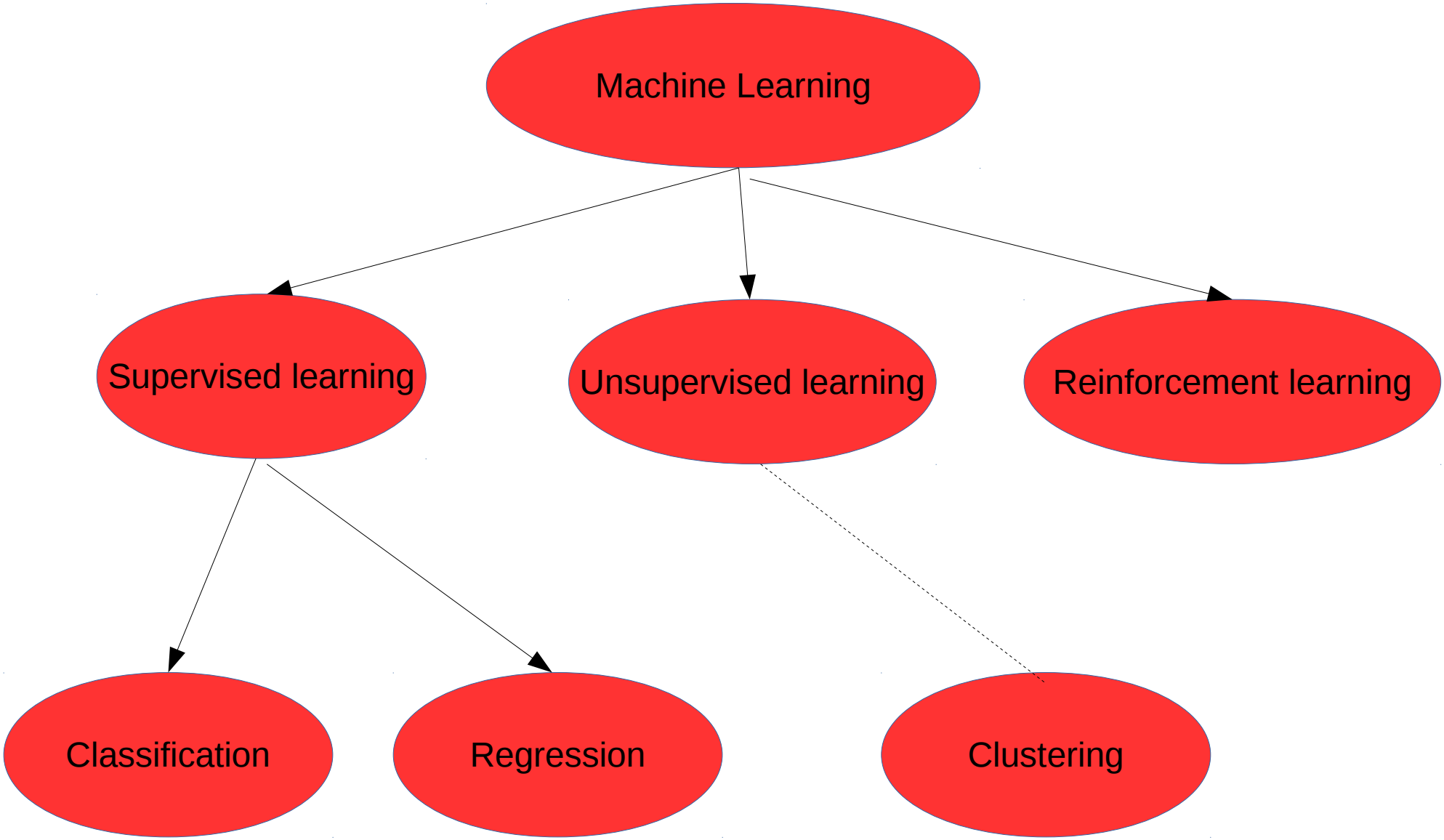
Data Science, Artificial Intelligence (AI) and Machine Learning (ML)



Arthur Samuel (1959):

„Field of study that gives computers the ability to learn without being explicitly programmed."

Link to the original article: here

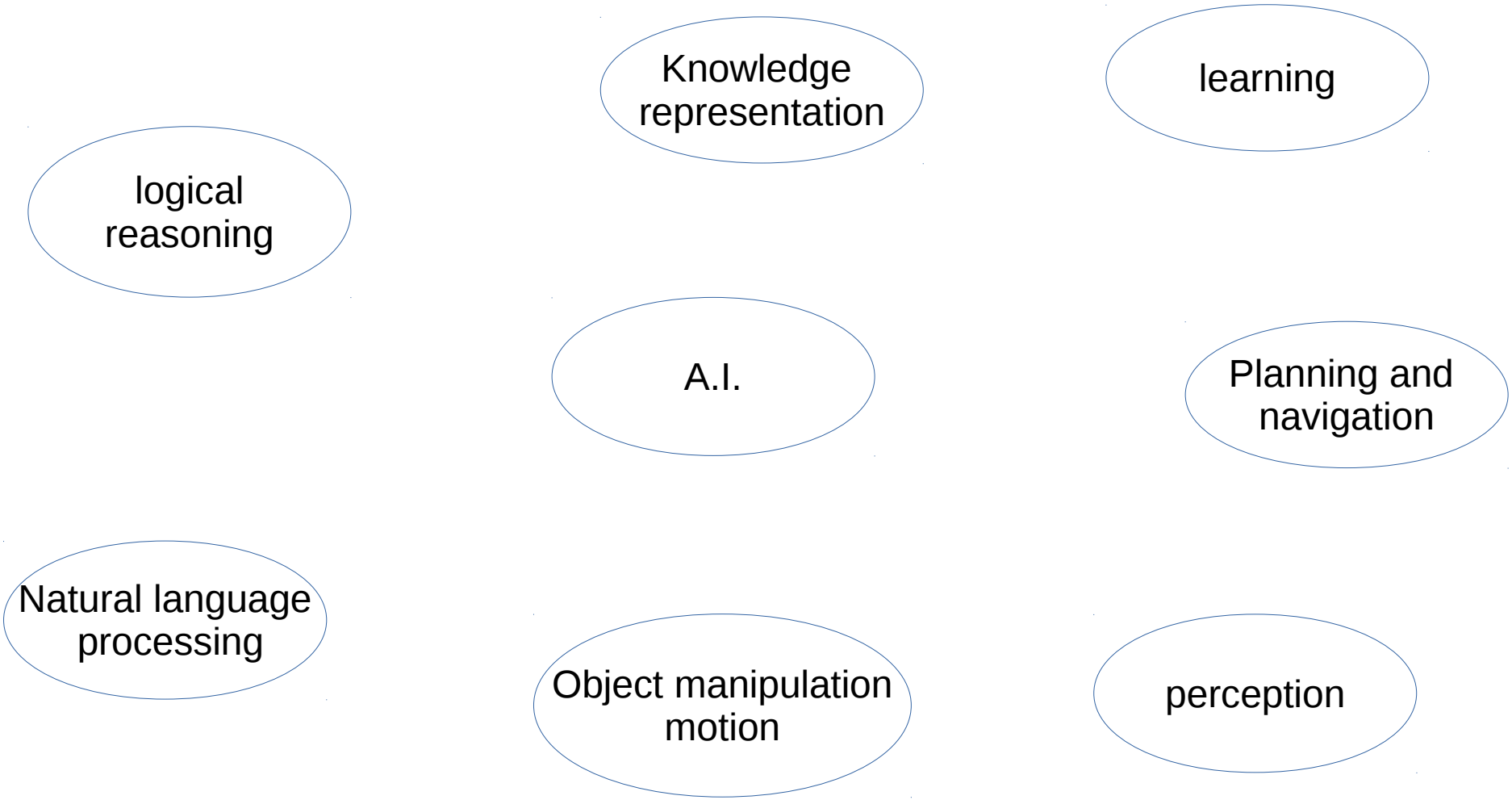

Machine Learning
Supervised learning
Unsupervised learning
Reinforcement learning
Classification
Regression
Clustering

ML is not new! Why such a hype?

- Computational Model of Neural Networks by W. S. McCulloch and W. Pitts (1943)
- Perceptron by F. Rossenblat (1958)
- Works on ML by A. Samuel (1959)
- ....

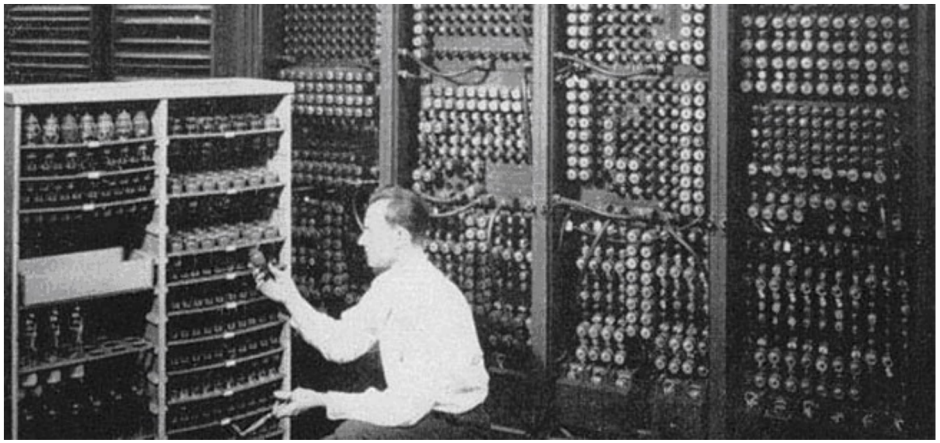
Recommended video of Frank Chen (hyperlink):
AI, Deep learning, and Machine Learning: A primer

A.I. domains


Knowledge representation
learning
logical reasoning
A.I.
Planning and navigation
Natural language processing
Object manipulation motion
perception

Artificial Intelligence Winters

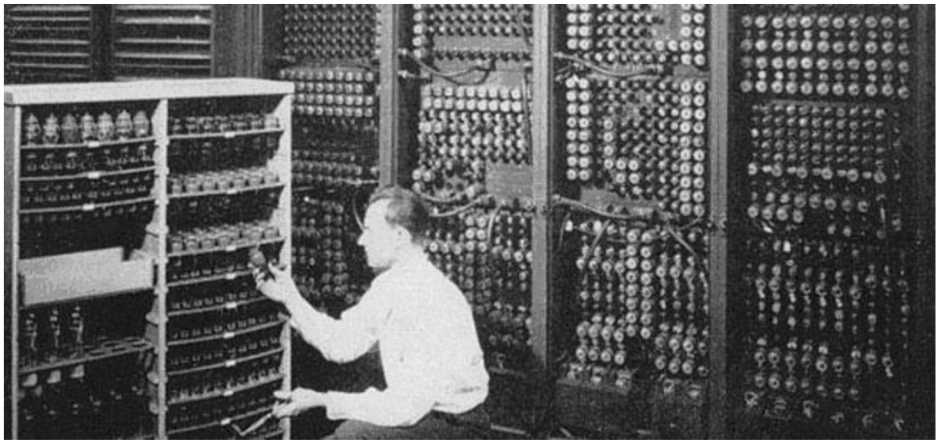
Machine translation tries during the Cold War (1954):



„The spirit was willing but the flesh was weak”

Artificial Intelligence Winters

Machine translation tries during the Cold War (1954):

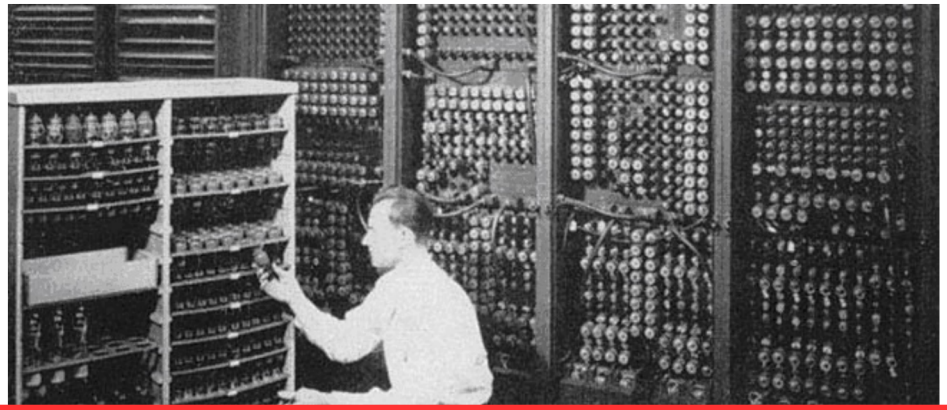


„The spirit was willing but the flesh was weak”

„The vodka was good, but the meat was rotten”

Artificial Intelligence Winters

Machine translation tries during the Cold War (1954):



Several periods of raising interests in A.I (and funds)

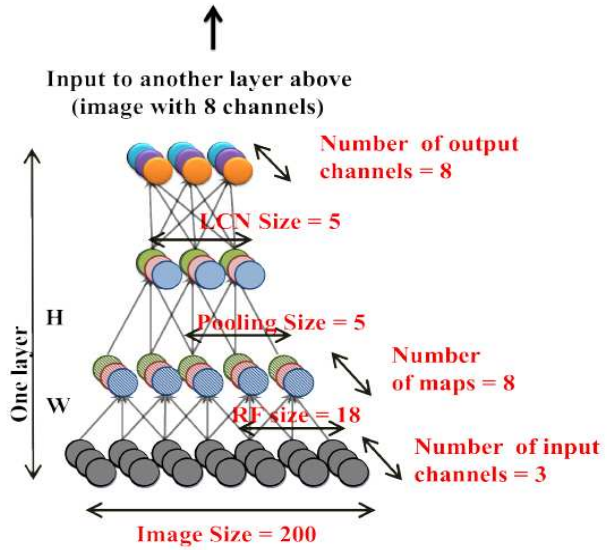
followed by the period of desillusions (and lack of funds)

„The vodka was good, but the meat was rotten”

TRANSLATOR
Russian → English

Deep learning „breakthrough”

- Google performed studies for face detector using the unsupervised learning approach.
- huge model: 9 interconnected layers, 1 billion connections (10^9)
- huge processing network: 1000 machines → 16000 cores
- huge training set: 10 mln images from YouTube (200x200 pixels)

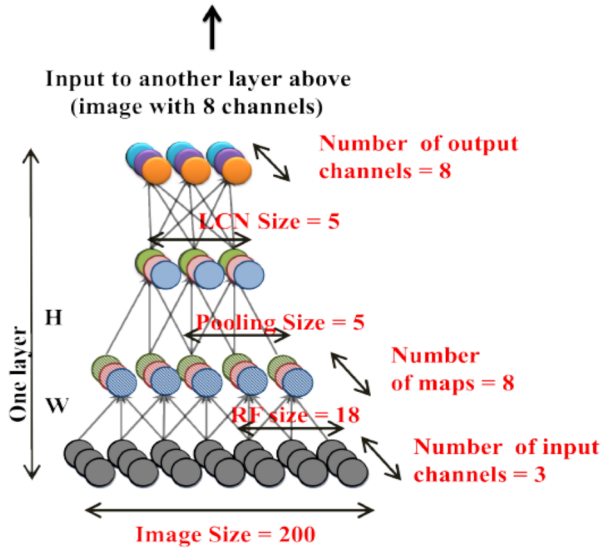


Building High-level Features Using Large Scale Unsupervised Learning (2012)

https://arxiv.org/abs/1112.6209

Deep learning „breakthrough”

- Google performed studies for face detector using the unsupervised learning approach.
- huge model: 9 interconnected layers, 1 billion connections (10^9)
- huge processing network: 1000 machines → 16000 cores
- huge training set: 10 mln images from YouTube (200x200 pixels)

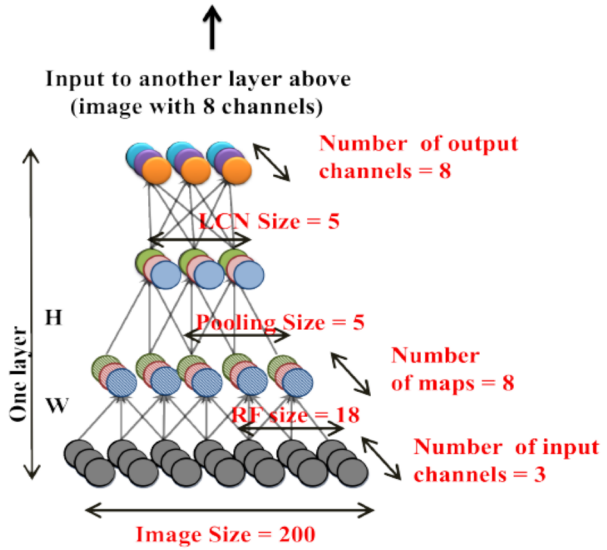


Building High-level Features Using Large Scale Unsupervised Learning (2012)

https://arxiv.org/abs/1112.6209

Deep learning „breakthrough”

- Google performed studies for face detector using the unsupervised learning approach.
- huge model: 9 interconnected layers, 1 billion connections (10^9)
- huge processing network: 1000 machines → 16000 cores
- huge training set: 10 mln images from YouTube (200x200 pixels)



Building High-level Features Using Large Scale Unsupervised Learning (2012)

https://arxiv.org/abs/1112.6209

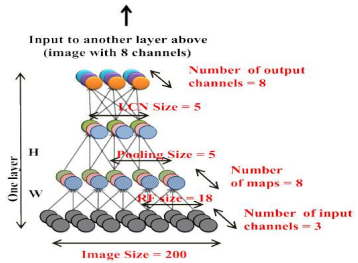
Deep learning „breakthrough”

- Recognition of objects such as: cats, human body, human face.
- Possible to recognize ~22,000 object categories from ImageNet (18,7% accuracy)
- Face detector rather robust to the distortion, e.g. rotation, scaling
- Huge improvement (recognizing simple shapes lines → recognizing complex objects)

Deep learning „breakthrough”

- Recognition of objects such as: cats, human body, human face.
- Possible to recognize ~22,000 object categories from ImageNet (18,7% accuracy)
- Face detector rather robust to the distortion, e.g. rotation, scaling
- Huge improvement (recognizing simple shapes lines → recognizing complex objects)

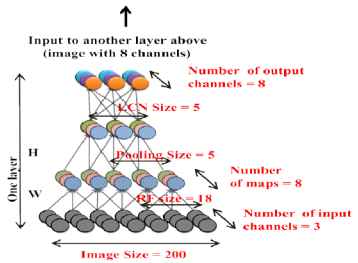
What has really changed ? **scale, model complexity!!!**

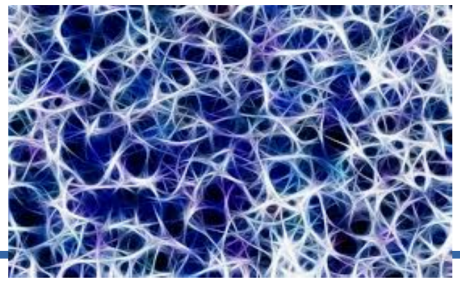


Deep learning „breakthrough”

- Recognition of objects such as: cats, human body, human face.
- Possible to recognize ~22,000 object categories from ImageNet (18,7% accuracy)
- Face detector rather robust to the distortion, e.g. rotation, scaling
- Huge improvement (recognizing simple shapes lines → recognizing complex objects)

What has really changed ? **scale, model complexity!!!**





Still only tiny fraction of biological neural connections

(~10^9 vs ~10^{14} connections in the human brain)

DeepFace (2015)

- Face recognition system developed by Facebook
- Neural Network with 9 layers and 120 millions of connections ($1.2 * 10^8$)
- ~97.35 % of accuracy (FBI system ~85 %, humans: ~97.5 %)
- Huge improvement (recognizing simple shapes lines → recognizing complex objects)

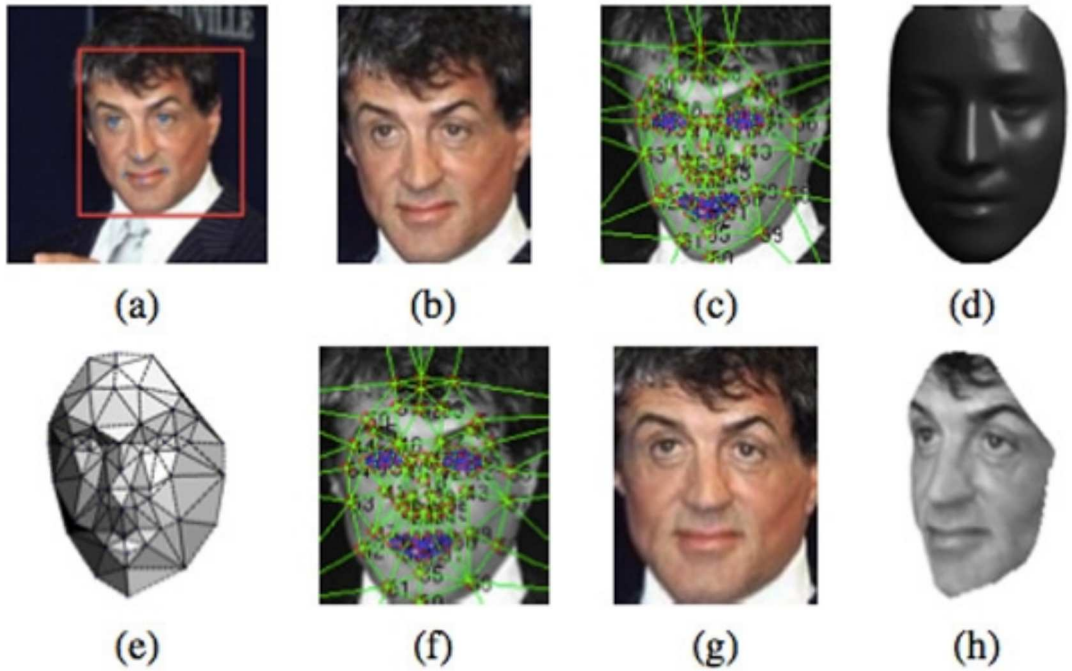


Image via Facebook

DeepBlue and AlphaGo

DeepBlue vs Gary Kasparov (1997)

- Evaluation of 200 million positions per second

AlphaGo vs Lee Sedol (2016)

- MC tree search
- Deep neural networks

AlphaStar and Real Time Strategy (Starcraft II)

AlphaStar won 10-1 against two pro players.

More details: https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/

Video transmission: https://www.youtube.com/watch?v=cUTMhmVh1qs&feature=youtu.be

MENU

nature

Article | Published: 30 October 2019

Grandmaster level in StarCraft II using multi-agent reinforcement learning

Oriol Vinyals, Igor Babuschkin, [...] David Silver

Nature (2019) | Cite this article

Abstract

Many real-world applications require artificial agents to compete and coordinate with other agents in complex environments. As a stepping stone to this goal, the domain of StarCraft has emerged as an important challenge for artificial intelligence research, owing to its iconic and enduring status among the most difficult professional esports and its relevance to the real world in terms of its raw complexity and multi-agent challenges. Over the course of a decade and numerous competitions[1,2,3], the strongest agents have simplified important aspects of the game, utilized superhuman capabilities, or employed hand-crafted sub-systems[4]. Despite these advantages, no previous agent has come close to matching the overall skill of top StarCraft players. We chose to address the challenge of StarCraft using general-purpose learning methods that are in principle applicable to other complex domains: a multi-agent reinforcement learning algorithm that uses data from both human and agent games within a diverse league of continually adapting strategies and counter-strategies, each represented by deep neural networks[5,6]. We evaluated our agent, AlphaStar, in the full game of StarCraft II, through a series of online games against human players. AlphaStar was rated at Grandmaster level for all three StarCraft races and above 99.8% of officially ranked human players.

https://deepmind.com/blog/article/AlphaStar-Grandmaster-level-in-StarCraft-II-using-multi-agent-reinforcement-learning

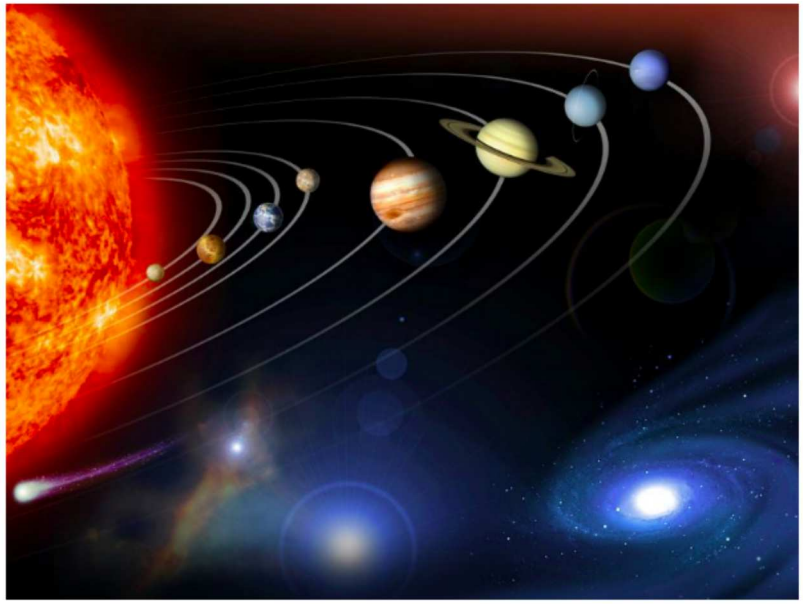
MENU nature Subscribe

NEWS · 07 NOVEMBER 2019

AI Copernicus 'discovers' that Earth orbits the Sun

A neural network that teaches itself the laws of physics could help to solve quantum-mechanics mysteries.

Davide Castelvecchi



Physicists have designed artificial intelligence that thinks like the astronomer Nicolaus Copernicus by realizing the Sun must be at the centre of the Solar System. Credit: NASA/JPL/SPL

RELATED ARTICLES

Reimagining of Schrödinger's cat breaks quantum mechanics – and stumps physicists

AI peer reviewers unleashed to ease publishing grind

Quantum machine goes in search of the Higgs boson

How machine learning could help to improve climate forecasts

Machine learning spots treasure trove of elusive viruses

SUBJECTS

Computer science Physics

https://www.nature.com/articles/d41586-019-03332-7

Discovering physical concepts with neural networks

https://arxiv.org/pdf/1807.10300.pdf

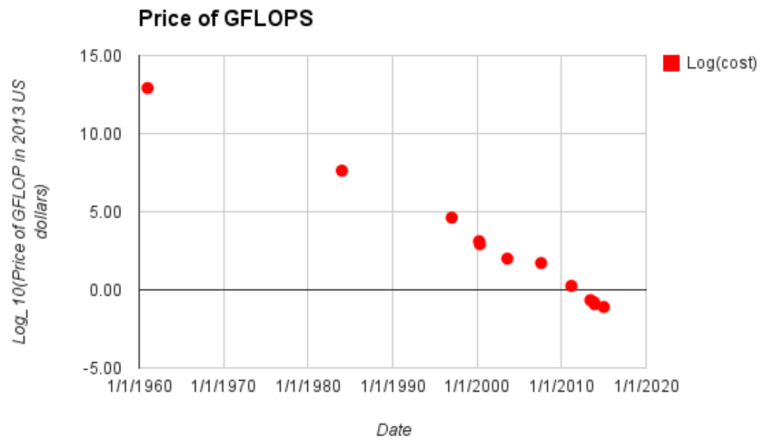
2018 ACM A.M. Turing awards

link: https://awards.acm.org/about/2018-turing

New wave of A.I.

Lower cost of computing power:

- Grid & Cloud computing, Chips dev.



https://aiimpacts.org/trends-in-the-cost-of-computing/

Huge amount of data gathered:

- Big data

Google facebook NETFLIX

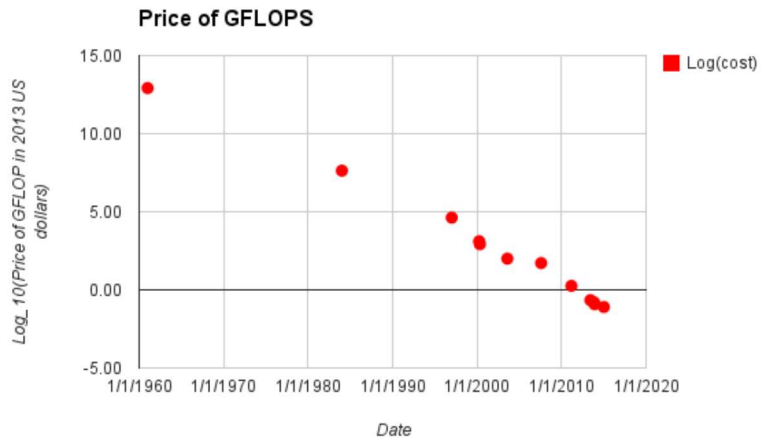
Better algorithms:

- Deep Learning
- ....

New wave of A.I.

Lower cost of computing power:

- Grid & Cloud computing, Chips dev.



https://aiimpacts.org/trends-in-the-cost-of-computing/

Huge amount of data gathered:

- Big data

Google facebook NETFLIX

Better algorithms:

- Deep Learning
- ....

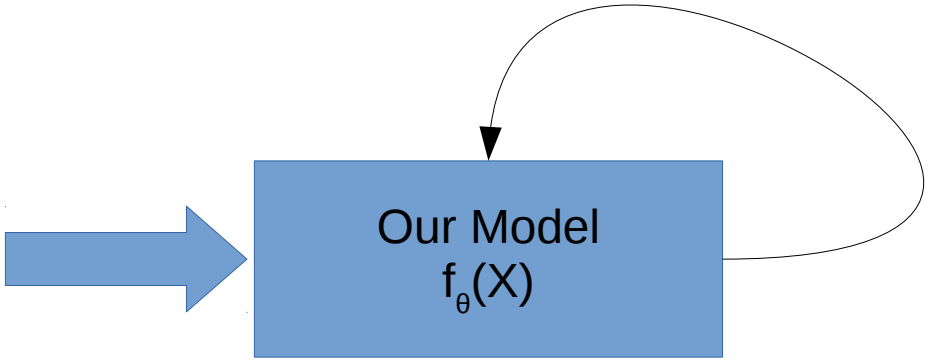
Don’t forget about A.I. winters

Supervised learning

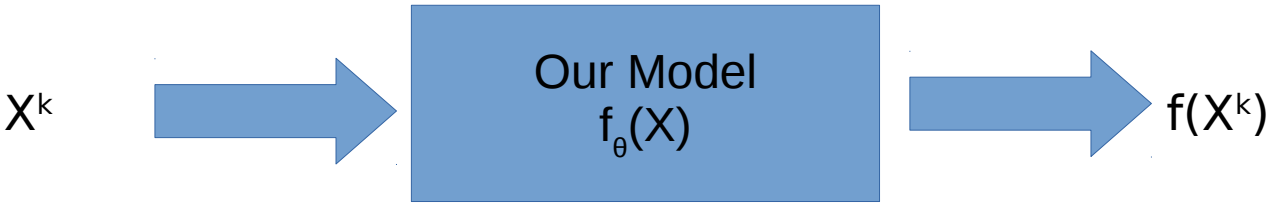
1.Training

$\{X^1,Y^1\}$, $\{X^2,Y^2\}$, ...

training data



2.Prediction:



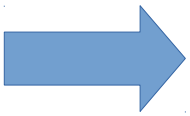
X – input, set of features (attributes)
Y – output
Training samples – pairs of $\{X^i, Y^i\}$ i=1,...N
θ – some parameters of the model

Supervised learning - spam filter

I am Mohammed Abacha,the son of the late Nigerian Head of State who died on the 8th of June 1998.If you are conversant with world news,you would understand better,while I got your contacts through my personal research.Please,I need your assistance to make this happen and please; do not undermine it because it will also be a source of upliftment to you also.You have absolutely nothing to loose in assisting us instead, you have so much to gain.

Please my dear,I repose great confidence in you and I hope you will not betray my confidence in you.I have secretly deposited the sum of **$30,000,000.00** with a security firm abroad whose name is withheld for now until we open communications.The money is contained in a metal box consignment with Security Deposit Number 009GM.

$\{X_1$, "non-spam"$\}$, $\{X_2$,"spam" $\}$

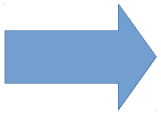


Our Model
$f_\theta(X)$

X – set of words indicative for spam
Y – „spam"/"non-spam"

Supervised learning handwriting recognition

{ 5 ,5}, { 2,2}, ...

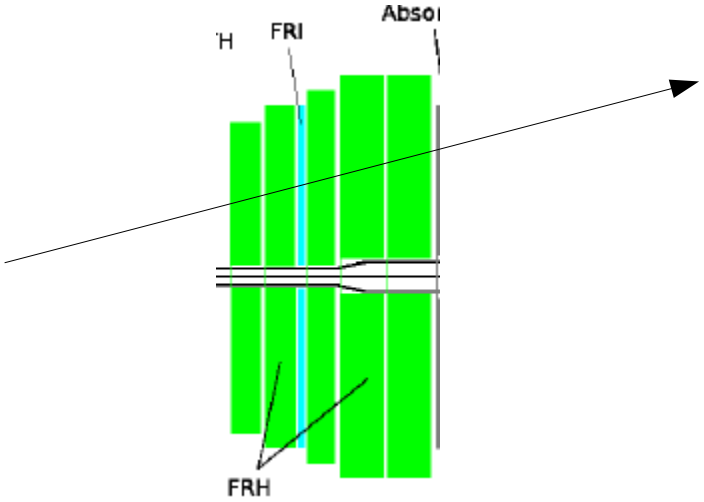


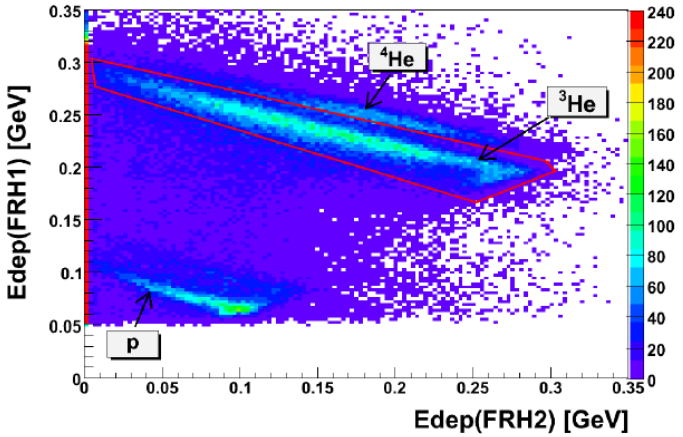
Our Model
$f_\theta(X)$

X – image of a digit
Y – numerical value

Yann LeCun – CNN pioneer work

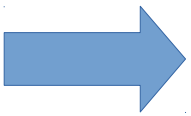
Supervised learning
Particle identification





{E(FRH1), E(FRH2) , “proton”}, {E(FRH1), E(FRH2),”pion”, },

{{E(FRH1), E(FRH2),”3He”}, ...

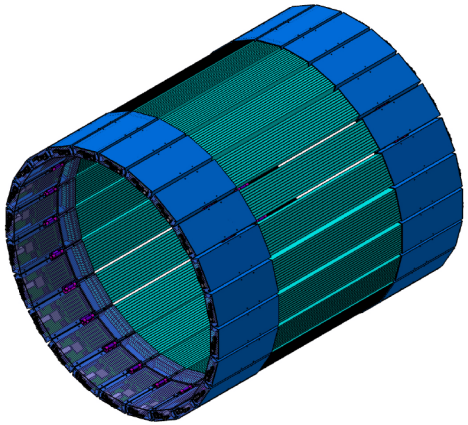


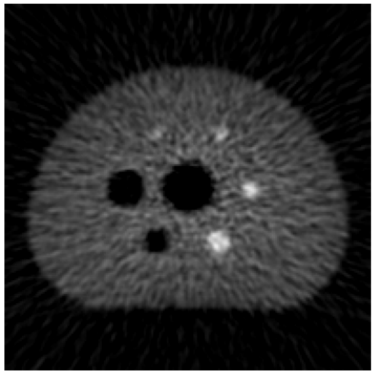
Our Model
$f_\theta(X)$

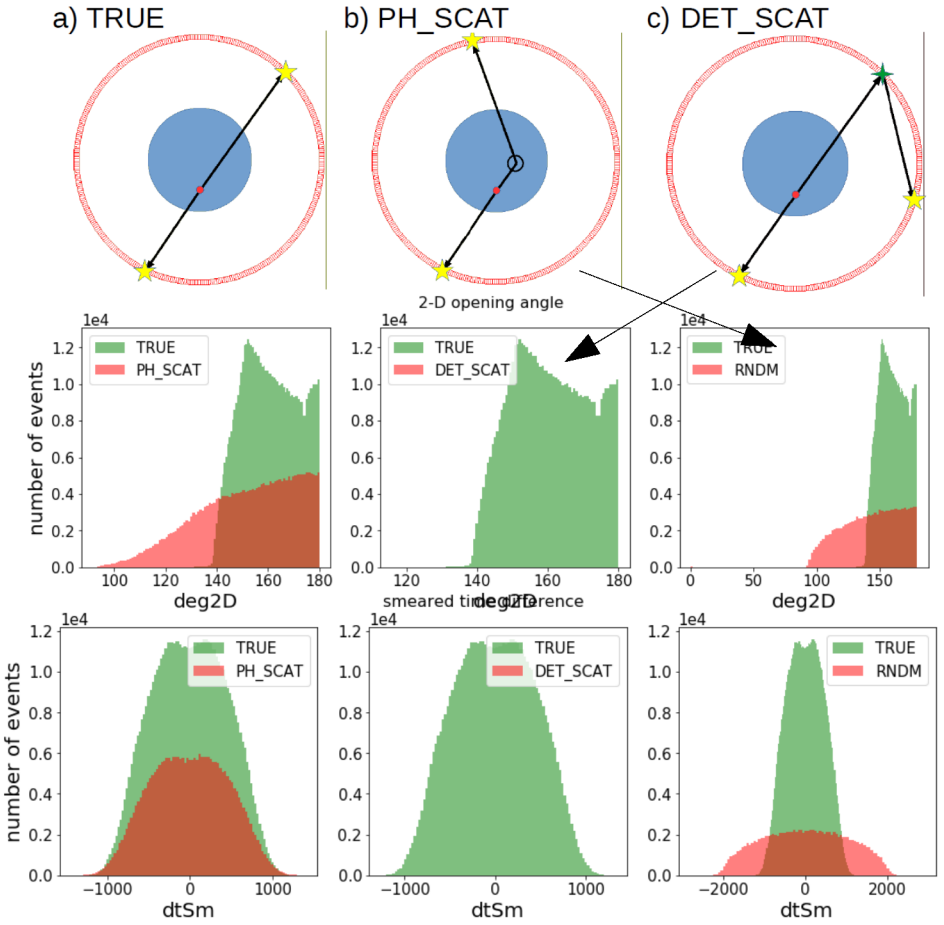
X – values of deposited energy
Y – particle type

We can use Monte Carlo simulations to train our model

Supervised learning
Positron Emission Tomography – background discrimination



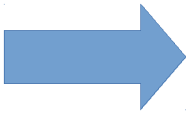




X – combined positions, deposited energies, times ..
Y – event type

{Δt ,Δ2Dangle, LOR length, $|E_1+E_2|$,$|E_1-E_2|$ "TRUE"}, {Δt

,Δ2Dangle, LOR length, $|E_1+E_2|$,$|E_1-E_2|$,"DET_SCATTER", }, ..



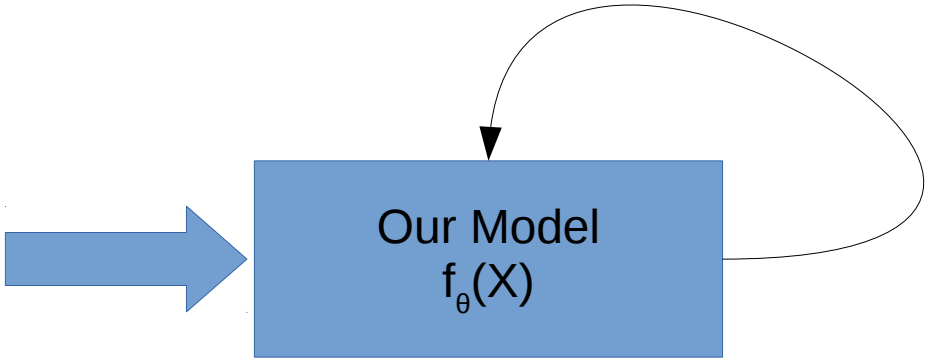
Our Model
$f_\theta(X)$

Unsupervised learning

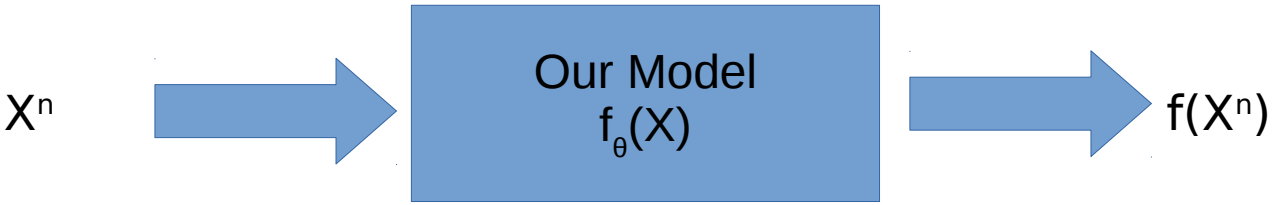
1.Training

$\{X^1\}$, $\{X^2\}$, ...

training data



2.Prediction:



X – input, set of features
Training samples – $\{X^i\}$ i=1,...N
θ – some parameters of the model

Unsupervised learning – DNA microarrays

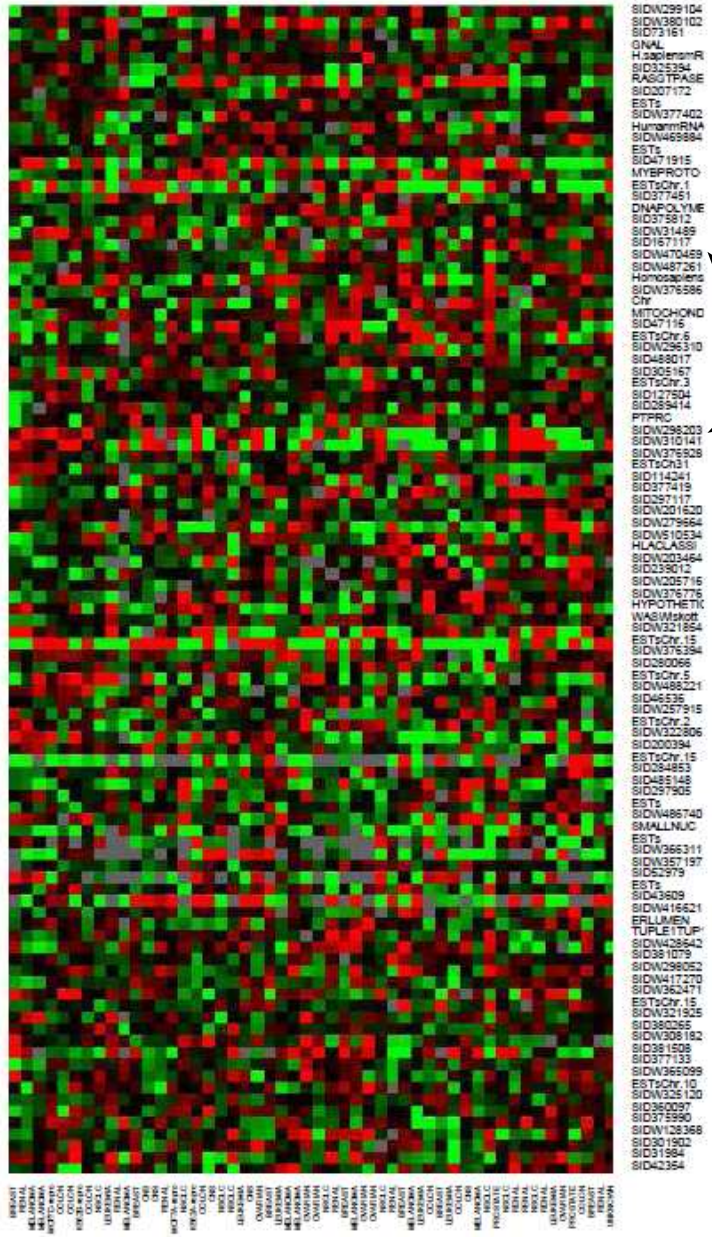
Target and reference tumor cancer samples

Colors:

- **Red** – higher intensity of RNA in target s.
- **Green**– higher intensity of RNA in reference s.

X – genes (several k features !)

Y – how to cluster the samples together ?



Taken from „The Elements of Statistical Learning" - Trevor Hastie, Robert Tibshirani, Jerome Friedman

A.I. dream

The final goal of A.I. ?

- Design of a “truly” intelligent machine
- Creativity
- Emotions
- Consciousnesses
- **New form of life**

"I fear that AI may replace humans altogether. If people design computer viruses, someone will design AI that improves and replicates itself. This will be a new form of life that outperforms humans."

Stephen Hawking
(interview for Wire Magazine)

A.I. dream

The final goal of A.I. ?

- Design of a “truly” intelligent machine
- Creativity
- Emotions
- Consciousnesses
- **New form of life**

"I fear that AI may replace humans altogether. If people design computer viruses, someone will design AI that improves and replicates itself. This will be a new form of life that outperforms humans."

Stephen Hawking
(interview for Wire Magazine)

A.I. dream

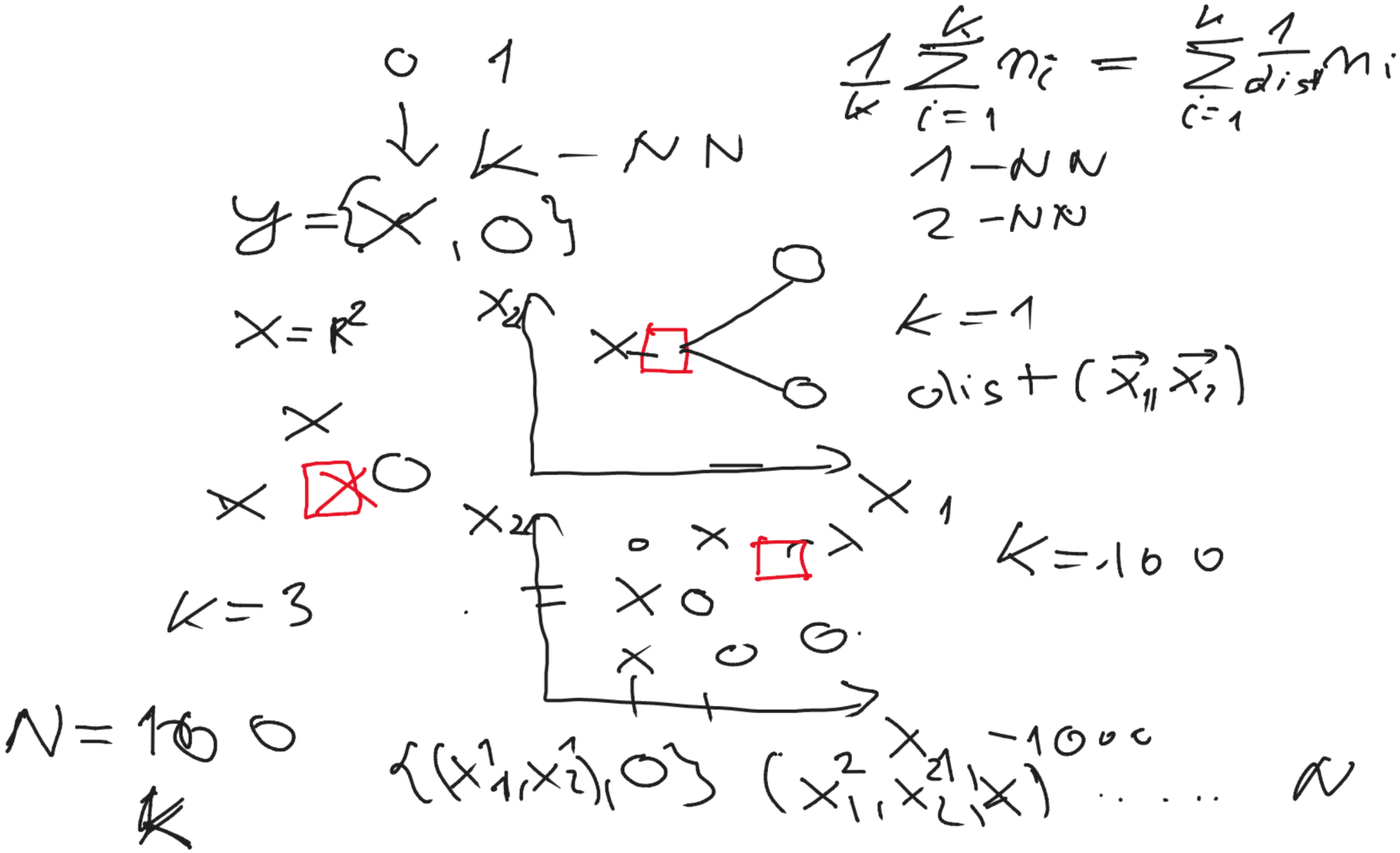
The final goal of A.I. ?

- Design of a “truly” intelligent machine
- Creativity
- Emotions
- Consciousnesses
- **New form of life**

"I fear that AI may replace humans altogether. If people design computer viruses, someone will design AI that improves and replicates itself. This will be a new form of life that outperforms humans."

Stephen Hawking
(interview for Wire Magazine)

Quiz


0 1
K – NN
y = {X, O}
X = R²
K = 3
N = 100
K
1/K Σ_{i=1}^{K} n_i = Σ_{i=1}^{K} 1/dist n_i
1 – NN
2 – NN
K = 1
dist (x_11, x_1)
K = 100
{(x¹₁, x²₁), O} (x²₁, x²₂, X) ..... N
-100c

Crash course in Python

https://github.com/wkrzemien/dataScienceAndML2020/tree/master/notebooks/intro