Data mining and machine learning

dsminingf17vm

Physics MSc course

01 - Course introduction

Pataki Bálint Ármin
ELTE, Physics of Complex Systems Department
2020.09.07.

Introduction & Course details

Webpage: https://csabaibio.github.io/physdm/

Main audience: Physics MSc students @ Scientific data analytics and modeling orientation

Prerequisites: Python programming experience (core python, numpy, matplotlib), basic linear algebra & basic probability theory

Language: English

Course time schedule:

- Lecture: Monday, 12:15-13:00, online, Microsoft TEAMS
  - Presentation, slides are available at the webpage
  - Also the recording
  - Theory
- Computer lab: Monday, 13:15-14:45, online, Microsoft TEAMS
  - Homework presentation & code examples
  - Technical help & personal feedback if needed

Contact: Pataki Bálint Ármin - patbaa@gmail.com - room 5.103 @ North building

Requirements

Attend lectures&labs is welcomed, not compulsory.

Grading:

- 80%: Weekly homeworks, 10 point for each. Will be checked offline.
  - Must be submitted until Sunday 23:59 on the same week (will be handed out on Monday).
  - Platform: https://kooplex-edu.elte.hu/hub
- 20%: Data mining project with written report @ the end of the semester.
  - Dataset will be provided OR custom datasets that fit your interest are welcome!
  - You need to analyze, come up with ideas how and what to model on the data
    - Perform meaningful supervised learning task
  - Write ~ 5 pages PDF report (wo/ code)
  - Deadline: 15th December: ~5 page written report

Discussion of homeworks is welcomed, but not at code level!

We will check code similarity.

Code should be own work (using referenced online public code is OK).

Resources

Webpage: https://csabaibio.github.io/physdm/
Kooplex: https://kooplex-edu.elte.hu/hub

Books:

- An Introduction to Statistical Learning
  - http://www-bcf.usc.edu/~gareth/ISL/
- The Elements of Statistical Learning: Data Mining, Inference, and Prediction
  - https://web.stanford.edu/~hastie/ElemStatLearn/

Online courses:

- Coursera - Machine Learning - https://www.coursera.org/learn/machine-learning

Tools:

- Python3, Jupyter - Easy install with Anaconda
- Numpy, scipy, sklearn, statmodels, matplotlib, seaborn, keras, pytorch, xgboost
- Kooplex ← educational cloud with everything pre-installed to handle homeworks

Schedule

1. Course introduction
2. Unsupervised learning and clustering
3. Overview of supervised learning
4. Linear regression
5. Linear methods for classification
6. Model selection and regularization
7. SVM
8. Decision trees, random forest
9. Neural networks
10. Convolutional neural networks
11. More neural networks
12. Natural language processing
13. Natural language processing II.

We have an other course, focused more on neural networks:

Deep learning and machine learning in science
deeplea17em
https://patbaa.github.io/physdl/
Spring semester

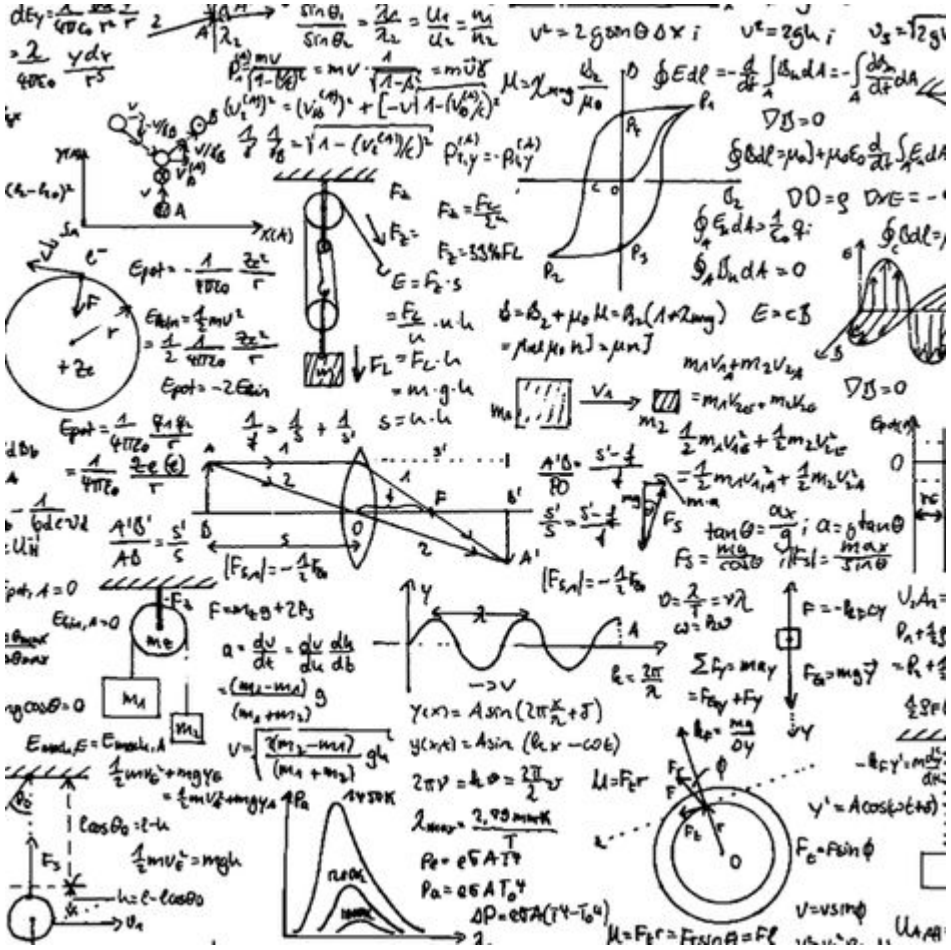
Goal of the course

- Broad overview of machine learning
  - “I’ve heard of it before, it is not that hard, let’s Google it how it was”
- Concepts
  - Most important
    - How to handle a problem
    - How not to fool yourself with a model
    - Measure & interpret results
- Hands-on experience
  - A practical field. Experience is needed.
  - Specific tools are important, but they change quickly
    - Overall experience is >> knowing API for a given tool
  - We use jupyter-python-sklearn-keras-tensorflow
    - 5-10 years ago:
      - R/matlab-Caffe-Theano

Why are we here?

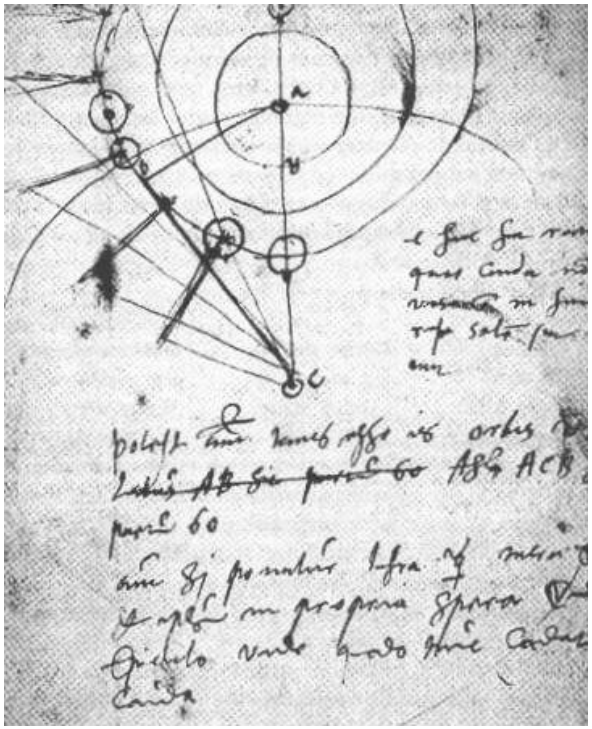
Scientific Data Analytics and Modelling?

Data mining and machine learning? But I applied for Physics MSc!

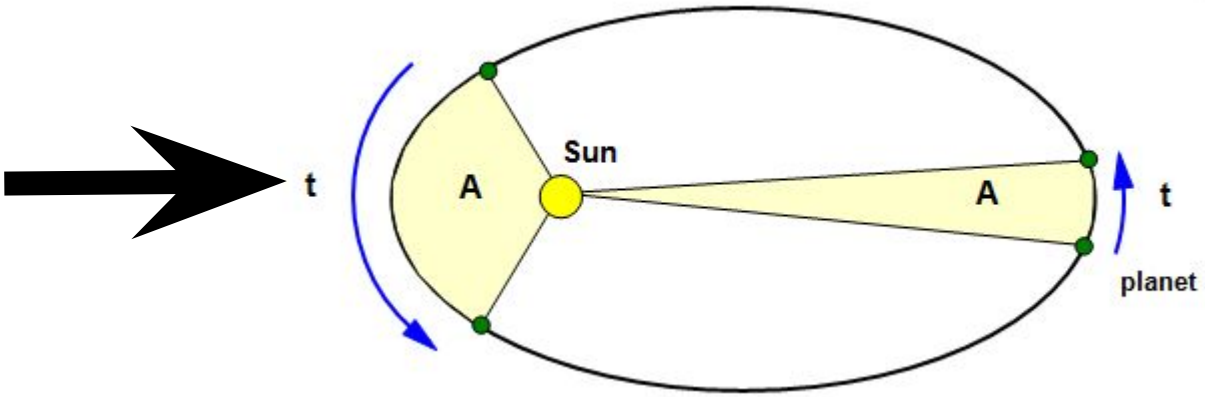


http://blog.cambridgecoaching.com/what-physics-equation-sheets-can-do-for-you-and-what-they-really-really-cant

Physics some times ago...

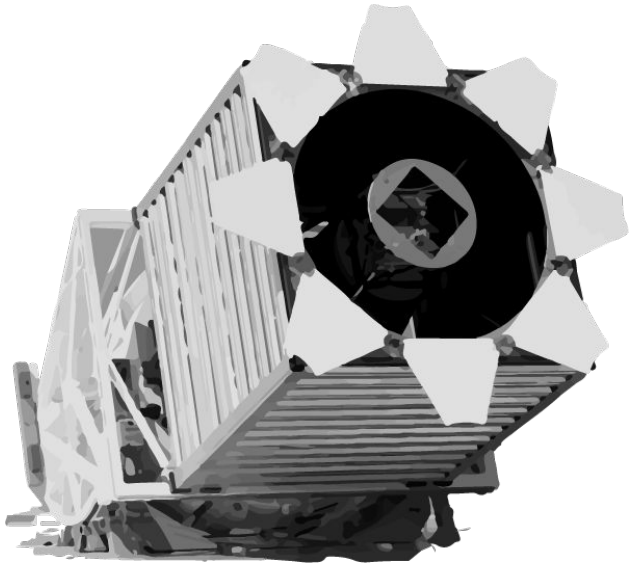


Tycho Brahe, 1500s, MB scale



Johannes Kepler, cc 1600

Physics some times ago... ...and now



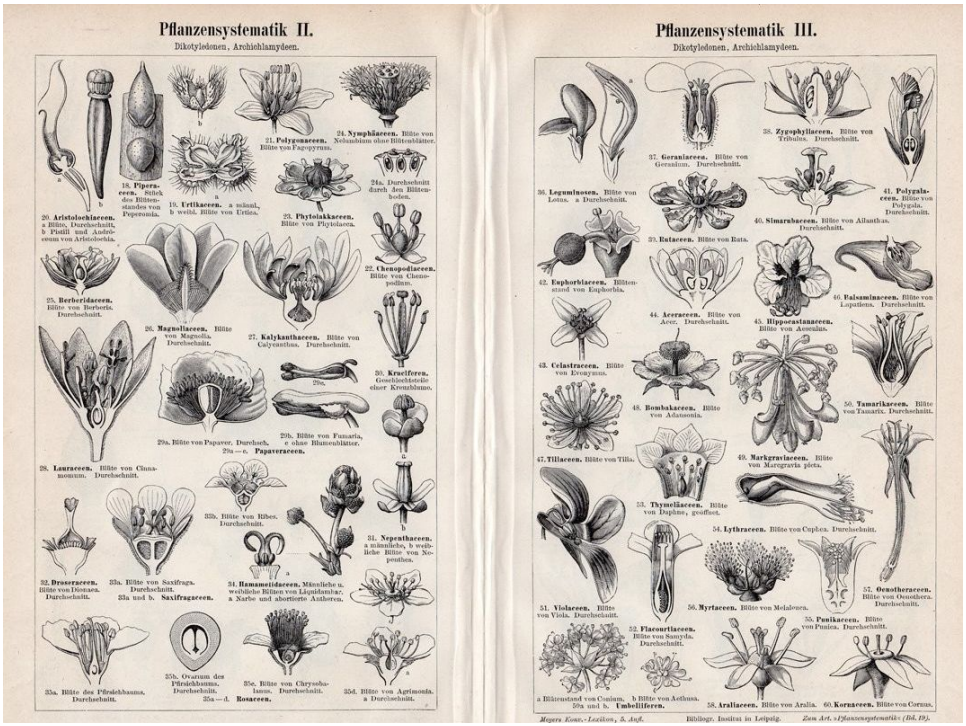
SDSS, 2000s, 116 TB

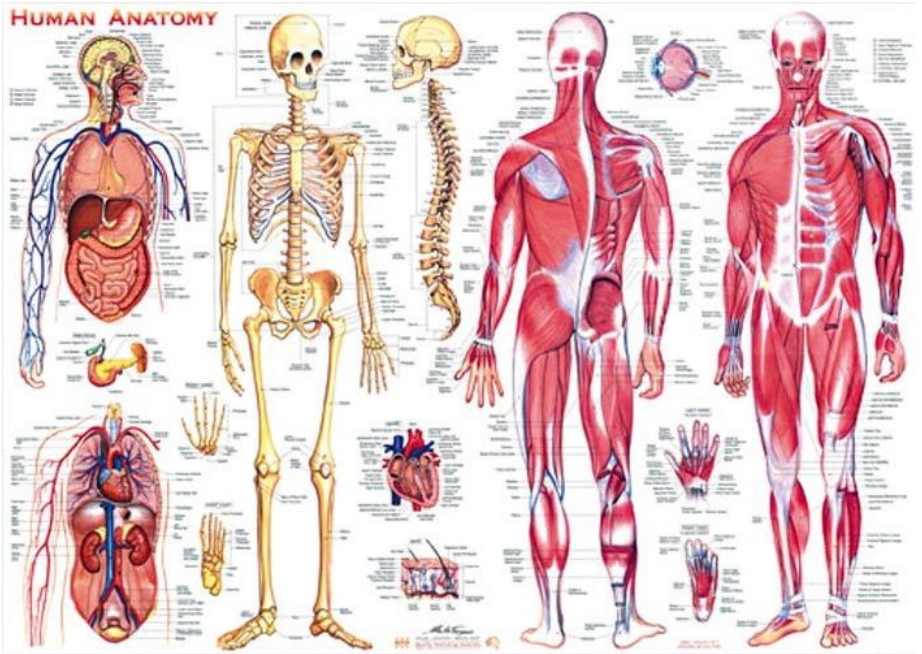
LIGO, 2000s, 1 TB per day

CERN, 2000s, 1000 TB per **day**

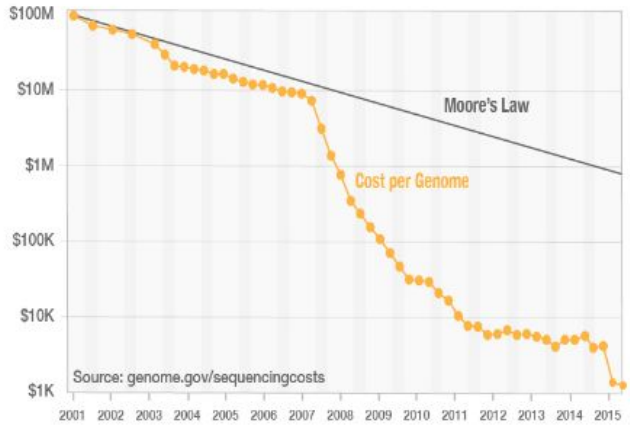
New tools are needed! Paper & pen is not enough!

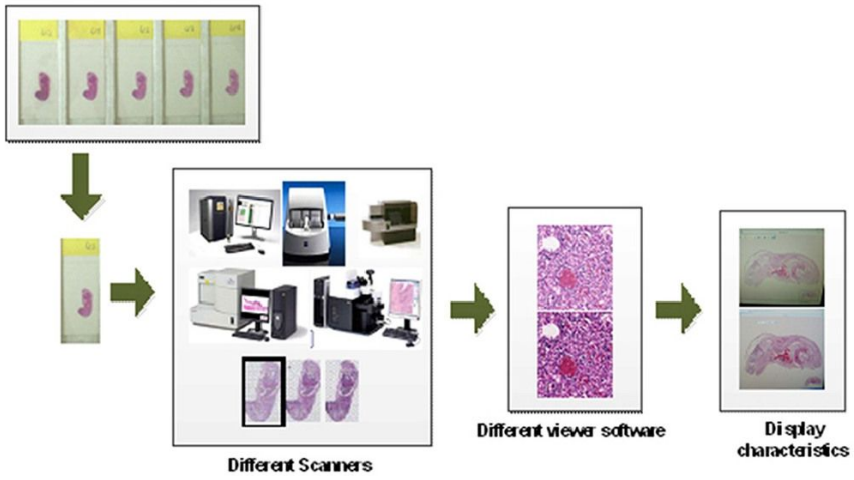
Biology some times ago...





Biology some times ago... ...and now

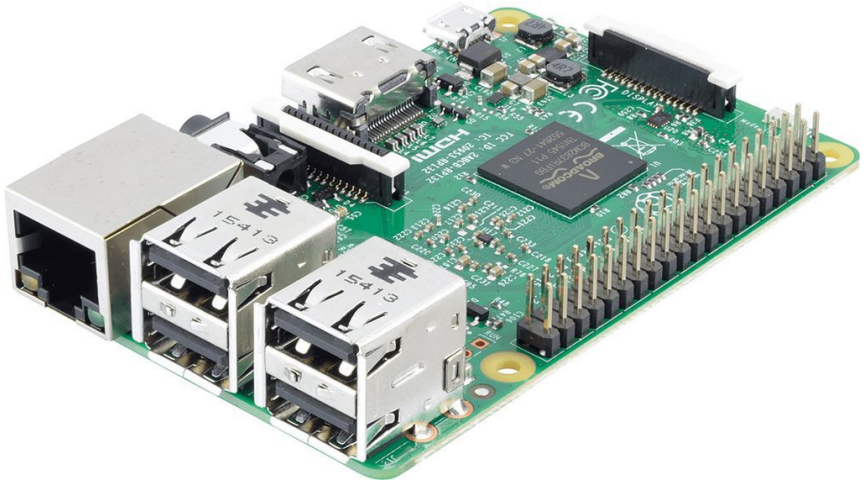




New tools are needed! A microscope & pipette is not enough!

IT some times ago... ...and now

Apollo Guidance Computer and DSKY



Raspberry Pi 3

Machine learning? AI? Why?

nature > articles > article

MENU

nature

Article | Published: 01 January 2020

International evaluation of an AI system for breast cancer screening

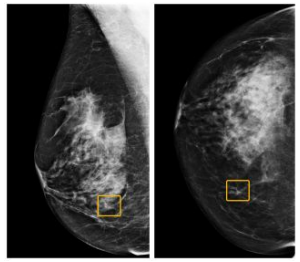
Scott Mayer McKinney, Marcin Sieniek, [...] Shravya Shetty

Nature **577**, 89–94(2020) | Cite this article

31k Accesses | **1** Citations | **3417** Altmetric | Metrics

Abstract

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To



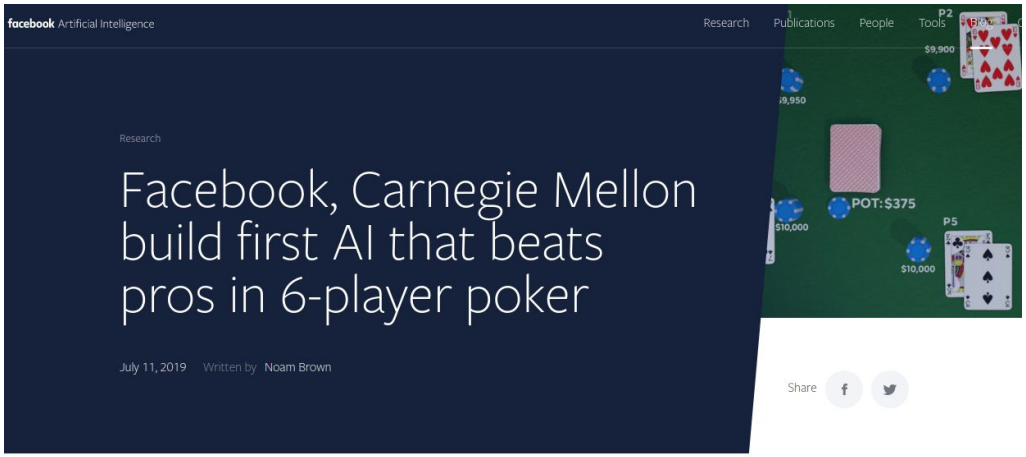
Machine learning? AI? Why?

‘On 19 November 2019, Lee announced his retirement from professional play, arguing that he could never be the top overall player of Go due to the increasing dominance of AI. Lee referred to them as being "an entity that cannot be defeated"’

Machine learning? AI? Why?

Machine learning? AI? Why?

Machine learning? AI? Why?



Review

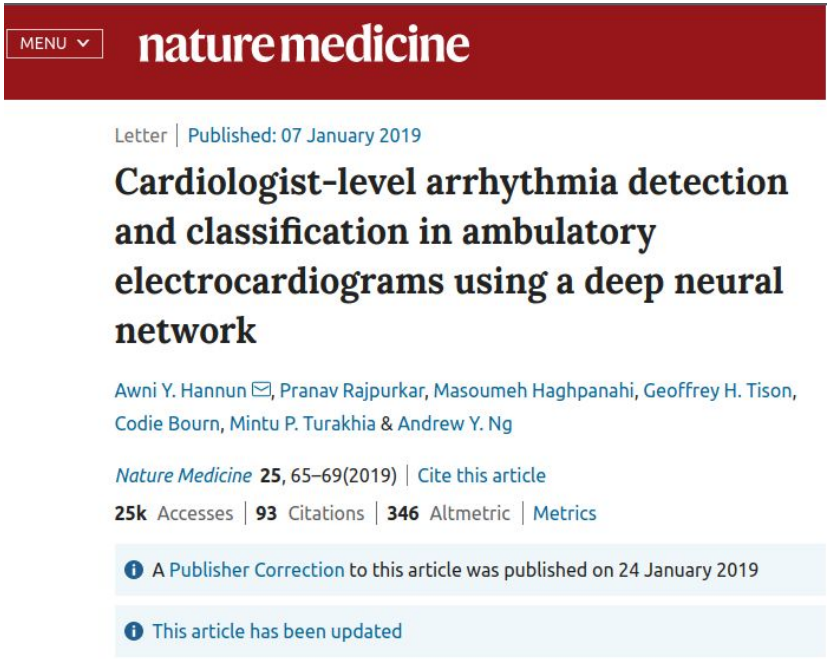
Informatics

Artificial intelligence in drug development: present status and future prospects

Kit-Kay Mak [1, 2], Mallikarjuna Rao Pichika [2, 3]

Show more

https://doi.org/10.1016/j.drudis.2018.11.014 Get rights and content



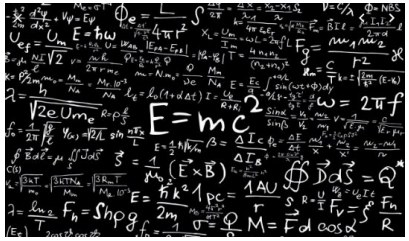
Many of them after 2018

Physics vs machine learning

Good model in physics

- Corresponds with former observations
- Accurately predict future experiments
- Based on data and human intuition
- Deep understanding of a phenomena
- Clear limitations
- Dense extraction of knowledge
- Fits to a t-shirt (3 letters are preferred)
  $E = mc^2$, $\underline{F} = m\underline{a}$,

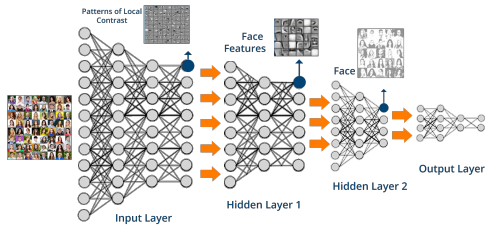
Main goal: deep & accurate understanding



Good model in ML

- Corresponds with former observations
- Accurately predict future experiments
- Based on data
- Complex (100M+ parameters)
- How and why is not the primary goal

Main goal: accuracy, usability



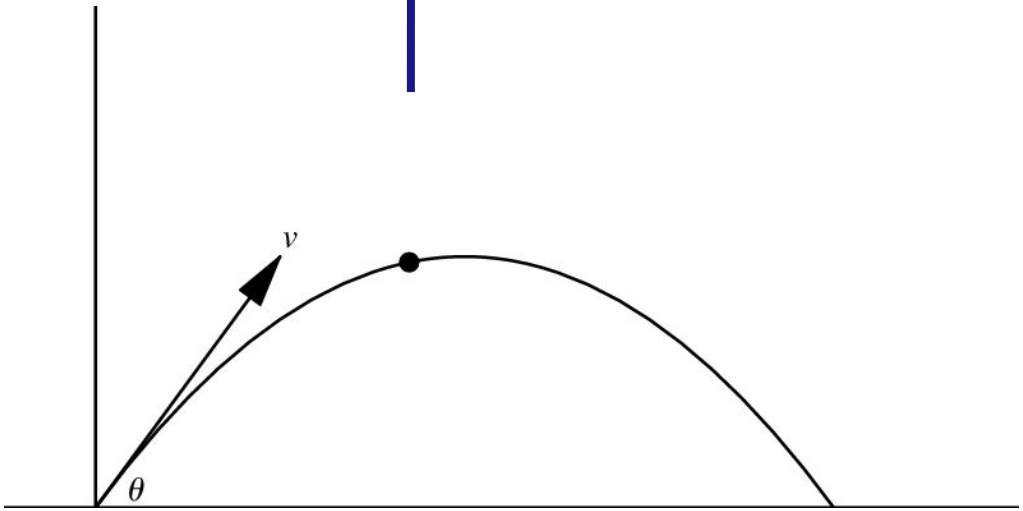
Physics vs machine learning by example

Physicist approach

- Many experiments
- Concepts
  - Gravity
  - Inertia
  - Speed
  - Acceleration
- Newton
  - Simple equations with a few parameters

ML approach

- Hundreds/thousands measurements
- Storing variables
  - Initial speed, angle
  - Mass, place of impact
  - Air temperature, wind
- Fitting a complicated model on data
  - No much intuition
  - Might work accurately



Physics vs machine learning by example

Physicist approach

- Many experiments
- Concepts
  - Gravity
  - Inertia
  - Speed
  - Acceleration
- Newton
  - Simple equations with a few parameters

ML approach

- Hundreds/thousands measurements
- Storing variables
  - Initial speed, angle
  - Mass, place of impact
  - Air temperature, wind
- Fitting a complicated model on data
  - No much intuition
  - Might work accurately

Stunningly hard to write up equation to many real word problem.

Any example idea?

Traditional programing vs machine learning by example

Traditional programming

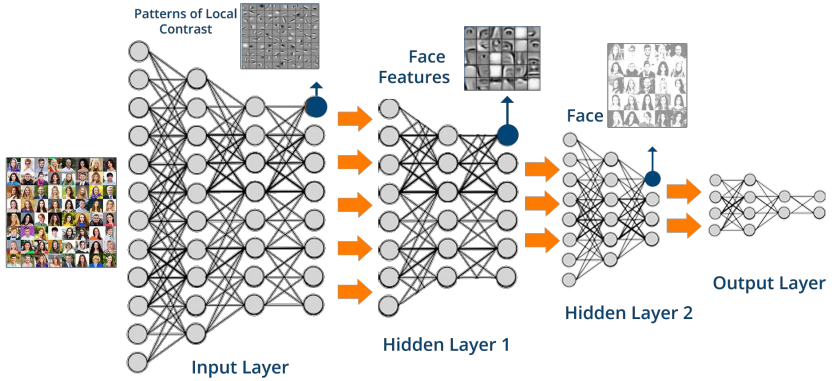
- Execution of pre-defined instructions

```
1  isPrime = True
2  for i in [2, 3, ... N-1]:
3      if divisible(N, i):
4          isPrime = False
5
6  print(isPrime)
```

- Works well in some cases, but...

Machine learning

- Model is a complex parametric function
- Parameters needs to be adjusted
  - Based on examples

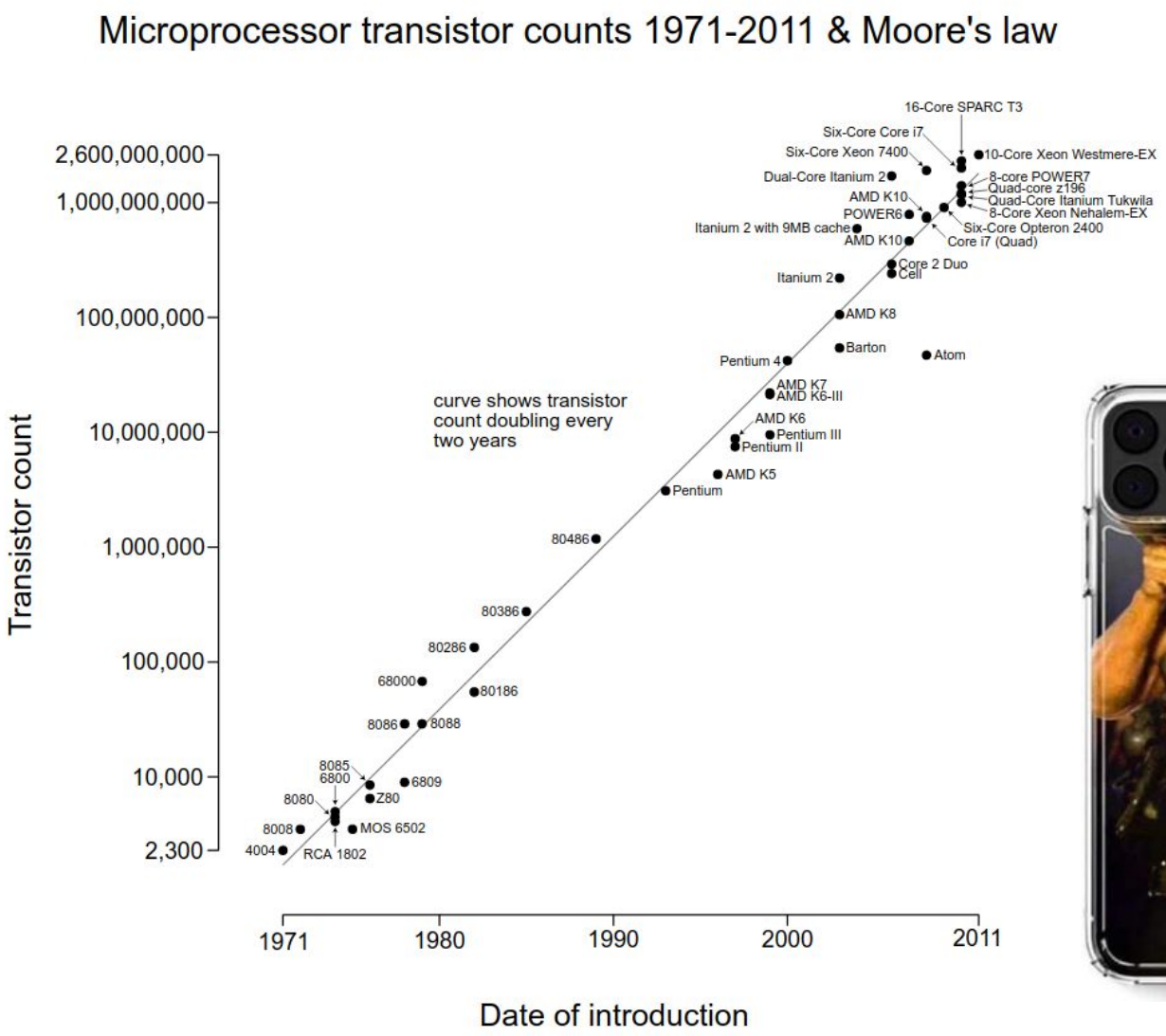


How would you write a program to recognise a car?

It is not a problem for a young kid.



Machine learning? Deep learning? Why now?



By Wgsimon - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=15193542

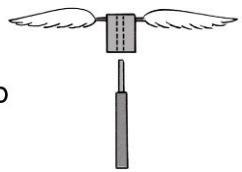
iPhone 11 Pro

Samsung Galaxy S10+

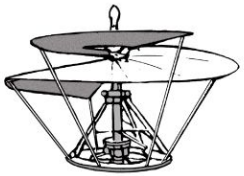
Helicopter history

History of the helicopter

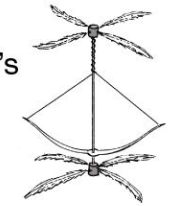
1100
Chinese flying top



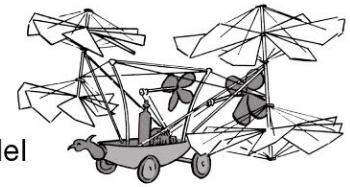
1483
Leonardo da Vinci's helical airscrew



1784
Launoy and Bienvenu's feather model

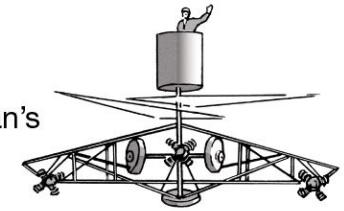


1843
Sir George Cayley's steam-powered model



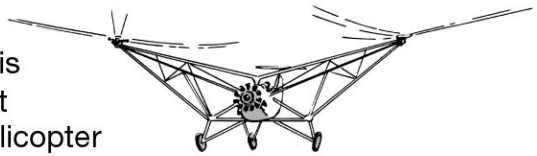
1907
Paul Cornu's first man-carrying helicopter

1916
István Petróczy and Theodore von Kármán's tethered helicopter

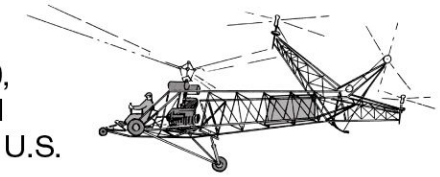


1923
Juan de la Cierva's autogiro

1936
Focke Achgelis Fa 61, the first successful helicopter



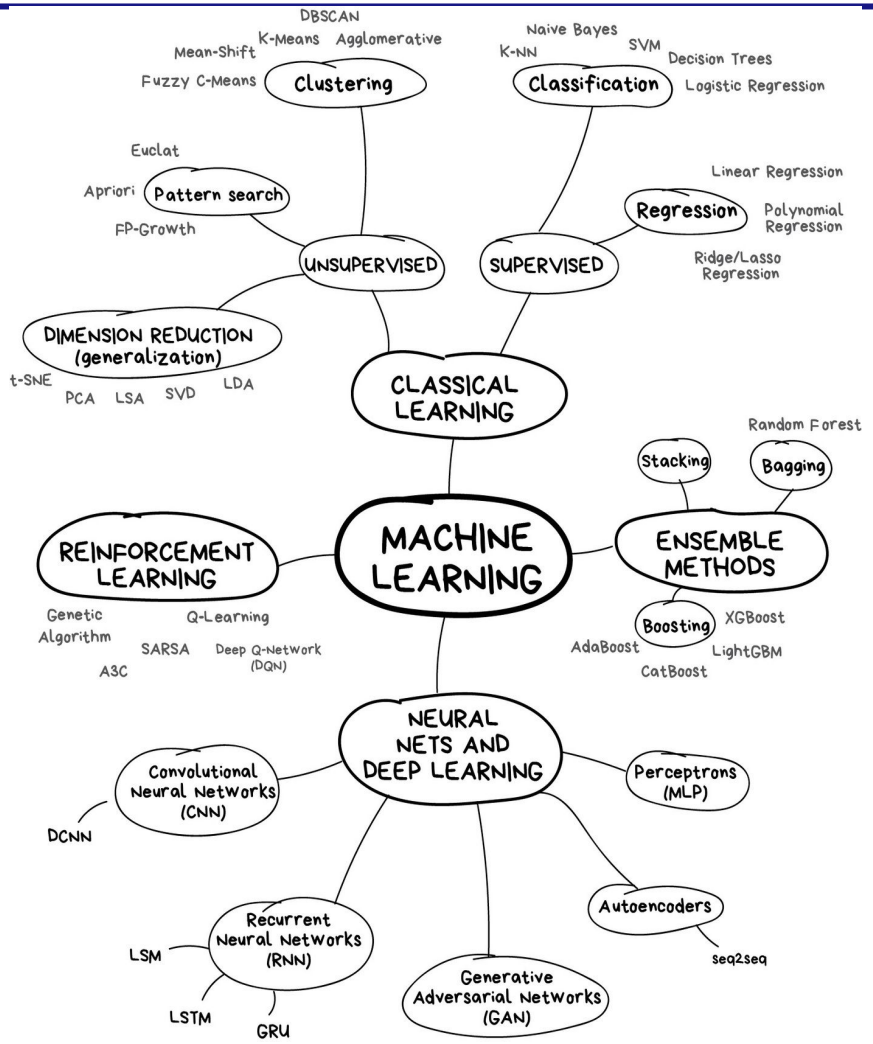
1939
Sikorsky VS-300, the first practical helicopter in the U.S.



A key component was missing for long: a powerful, reliable, not too heavy internal combustion engine

Our internal combustion engine: data & compute power, which arrived lately

Machine learning has many branches

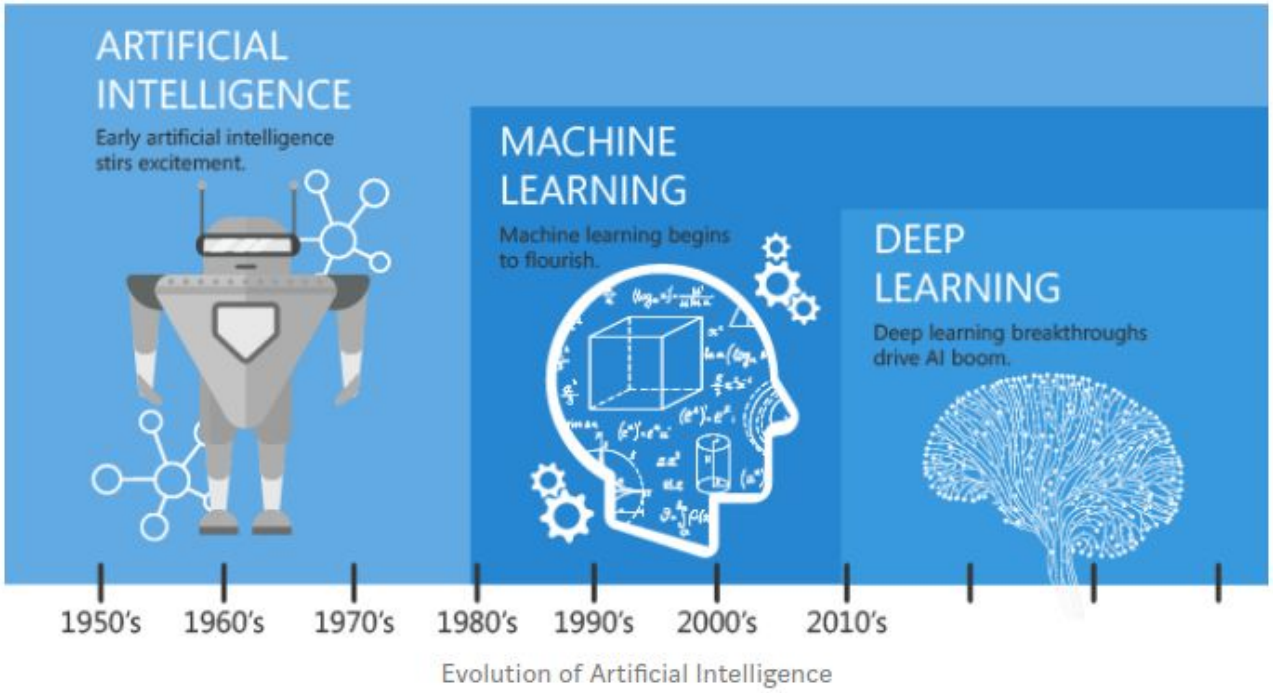


http://valyrics.vas3k.com/blog/machine_learning/

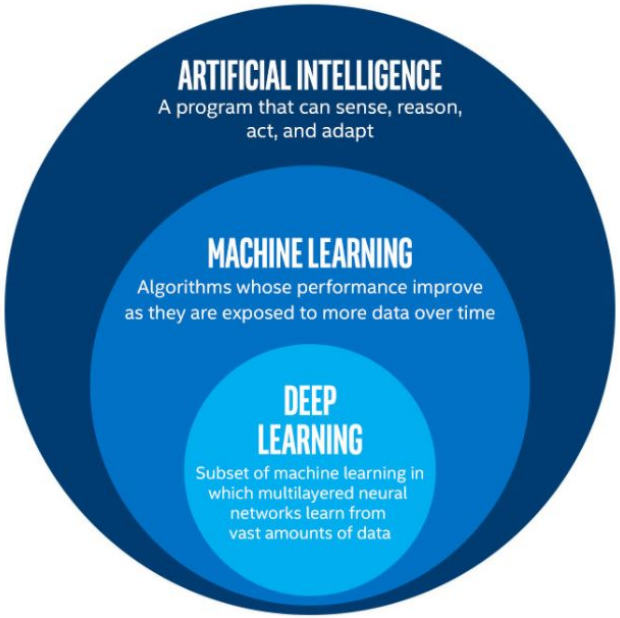
1. Course introduction
2. Unsupervised learning and clustering
3. Overview of supervised learning
4. Linear regression
5. Linear methods for classification
6. Model selection and regularization
7. SVM
8. Decision trees, random forest
9. Neural networks
10. Convolutional neural networks
11. More neural networks
12. Natural language processing
13. Natural language processing II.

Rule of thumbs exist, but there is no ultimate superior model, method!

Machine learning vs AI vs Deep Learning

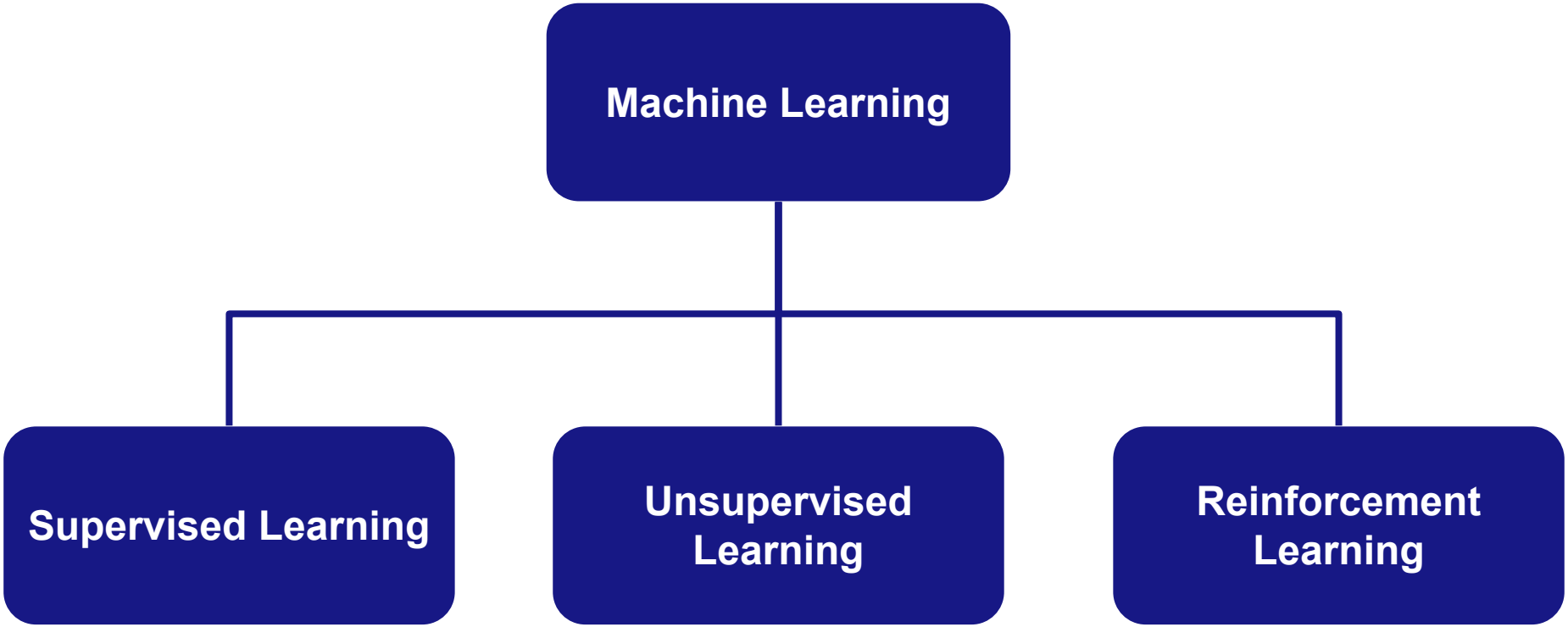


https://towardsdatascience.com/artificial-intelligence-vs-machine-learning-vs-deep-learning-2210ba8cc4ac



* in news: everything is AI

Subgroups within machine learning



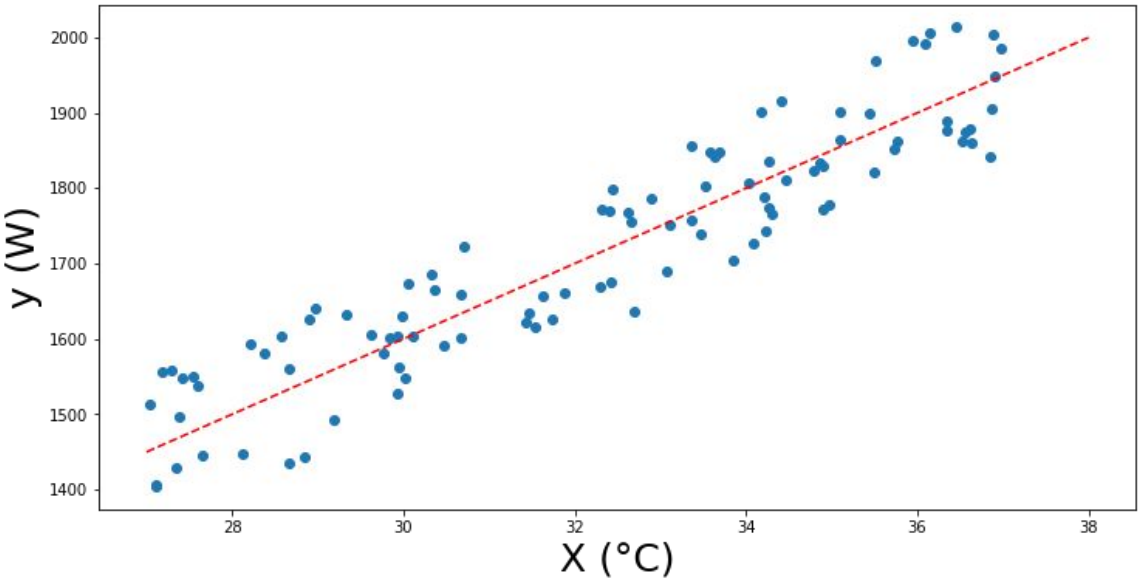
- Regression
- Classification

- Clustering
- Outlier detection
- Dimension reduction

- Interaction w environment
  Eg: playing Go

Machine learning as curve fitting

- Data: (**X**, **y**) pairs
  - **X** is the input data
  - **y** is what we want to predict
  - Eg: temperature ➔ power usage
- Goal: tell **y** for a new **X**
- Simple model:
  - Linear relationship
  - $\mathbf{y} = f(\mathbf{X}, w) = w_0 + x * w_1$
    - w_0 is the axis section
    - w_1 is the slope
  - How to set w_0 and w_1?
    - Minimize the error!
      - $(\mathbf{y} - f(\mathbf{X}, w))^2$



Machine learning as curve fitting

$$y = f(\ \text{  }\ , w)$$

- w is not a set of 2 but 10/100 million parameters
- f not a simple linear function, we need to set up f smartly
- Training a neural network: optimizing **w** based on training examples

But it is not a simple fitting ➔ validation

Supervised learning is not a usual optimization!

The goal is to build a model that will work in the future during real-world conditions.

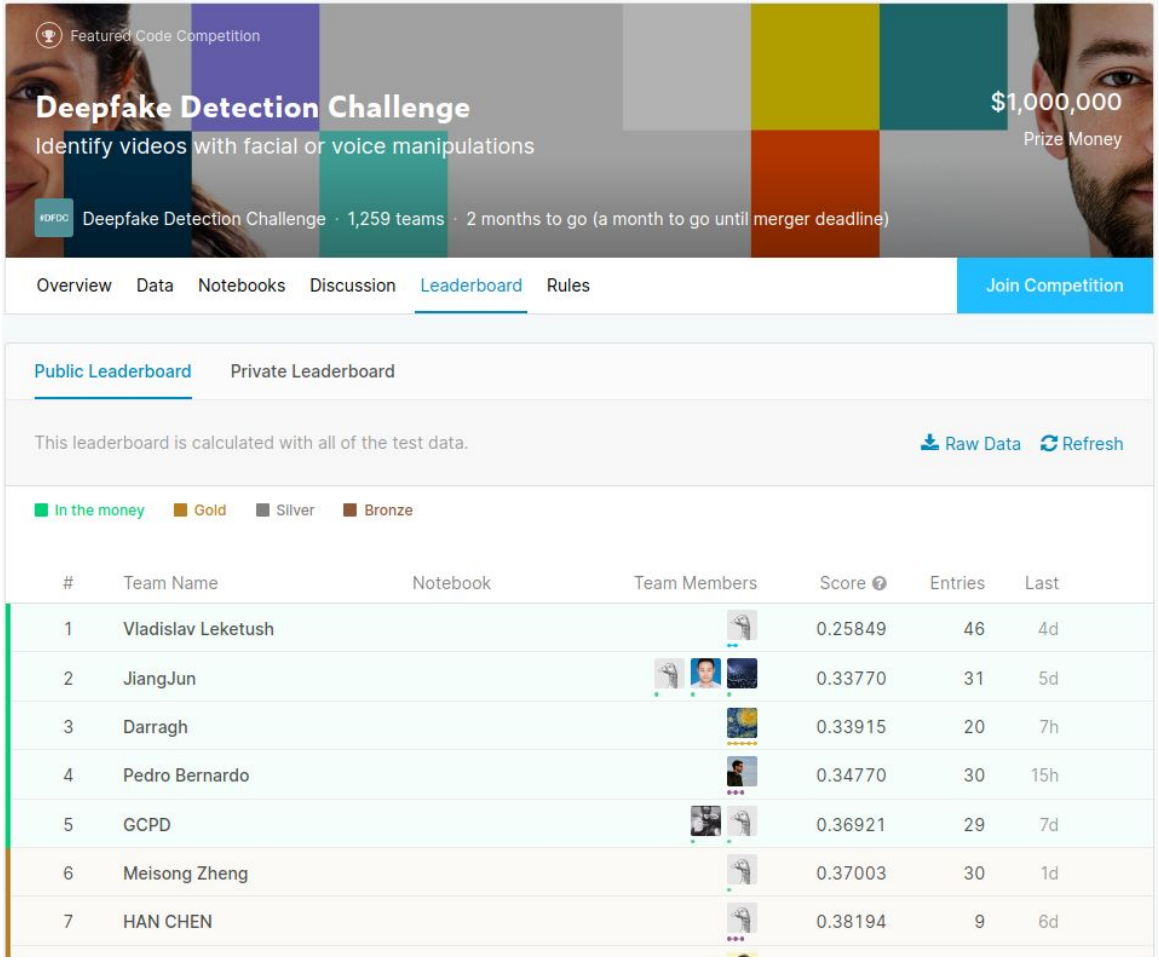
- Drives your car
- Detects cancer
- Makes you money on the stock market

You want to estimate the power of your model. How do you do that?

- Performance on the training set?
  - Complex models can memorize the training set ➔ bad idea
- Performance on an unseen dataset?
  - Much better idea, if you have enough data!
  - Try many different model, train all on test data and pick the one that is the best on test?
    - What if one was just lucky? If you have enough model one will be!
    - Would you accept the lottery winners' advice on how to play lottery?

Machine learning competitions

- Kaggle, Dream Challenges, Alcrowd, Drivendata etc.
- Train data released
  - Both X and y (labels)
- Test data is partially released
  - Labels kept private
- Local model fitting
  - Submitting predictions
  - Predictions vs true labels
    - ranking



Exploratory data analysis - where all work starts

You are given a bunch of data to:

- Predict something that is important
- ‘Make something useful out of it’

Data size usually large to ‘manually’ fully understand, what to do:

- Eyeballing
  - Excel / image viewer
  - Strange values? Systematic missing values? Is something strange?

- Visualize everything you can
  - Distributions, histograms, box-plots
  - Scatter plots
  - Bar charts, trend lines
  - Heatmaps, correlations
  - GPS coordinates ➔ map plot
  - Whatever that gives you insight about the data!

Simpler machine learning models can help understanding even during EDA!
Goal: have an overall insight what the data contains & to reduce the future surprises!
On Kaggle you can find excellent EDAs for many different data!

Learning more + Homework & lab

Hot topic, quick development ➔ well-written up-to-date books are rare

- Competitions, hand-on experience
- Online courses (Coursera, Stanford etc)
- Research papers
- Hobby projects

Homework: EDA - visualization & data handling in Python

This week during the lab we will review examples for

- Numpy, matplotlib, pandas, sklearn
- Exploratory data analysis
- Get familiar with kooplex

Any questions?

Assignment 1

Exploratory data analysis

http://patbaa.web.elte.hu/physdm/data/titanic.csv

On the link above you will find a dataset about the Titanic passengers. Your task is to explore the dataset.

Help for the columns:

- SibSp - number of sibling/spouses on the ship
- Parch - number of parent/children on the ship
- Cabin - the cabin they slept in (if they had a cabin)
- Embarked - harbour of entering the ship
- Pclass - passenger class (like on trains)

1. Load the above-linked csv file as a pandas dataframe. Check & plot if any of the columns has missing values. If they have, investigate if the missingness is random or not.

Impute the missing values in a sensible way:

- if only a very small percentage is missing, imputing with the column-wise mean makes sense, or also removing the missing rows makes sense
- if in a row almost all the entries is missing, it worth to remove that given row
- if a larger portion is missing from a column, usually it worth to encode that with a value that does not appear in the dataset (eg: -1).

The imputing method affects different machine learning models different way, but now we are interested only in EDA, so try to keep as much information as possible!

2. Create a heatmap which shows how many people survived and dies with the different Pclass variables. You need to create a table where the columns indicates if a person survived or not, the rows indicates the different Pclass and the cell values contains the number of people belonging the that given category. The table should be colored based on the value of the cells in the table.

3. Create boxplots for each different Pclass. The boxplot should show the age distribution for the given Pclass. Plot all of these next to each other in a row to make it easier to compare!

4. Calculate the correlation matrix for the numerical columns. Show it also as a heatmap described at the 1st task.

Which feature seems to play the most important role in surviving/not surviving? Explain how and why could that feature be important!

5. Create two plots which you think are meaningful. Interpret both of them. (Eg.: older people buy more expensive ticket? people buying more expensive ticket survive more? etc.)