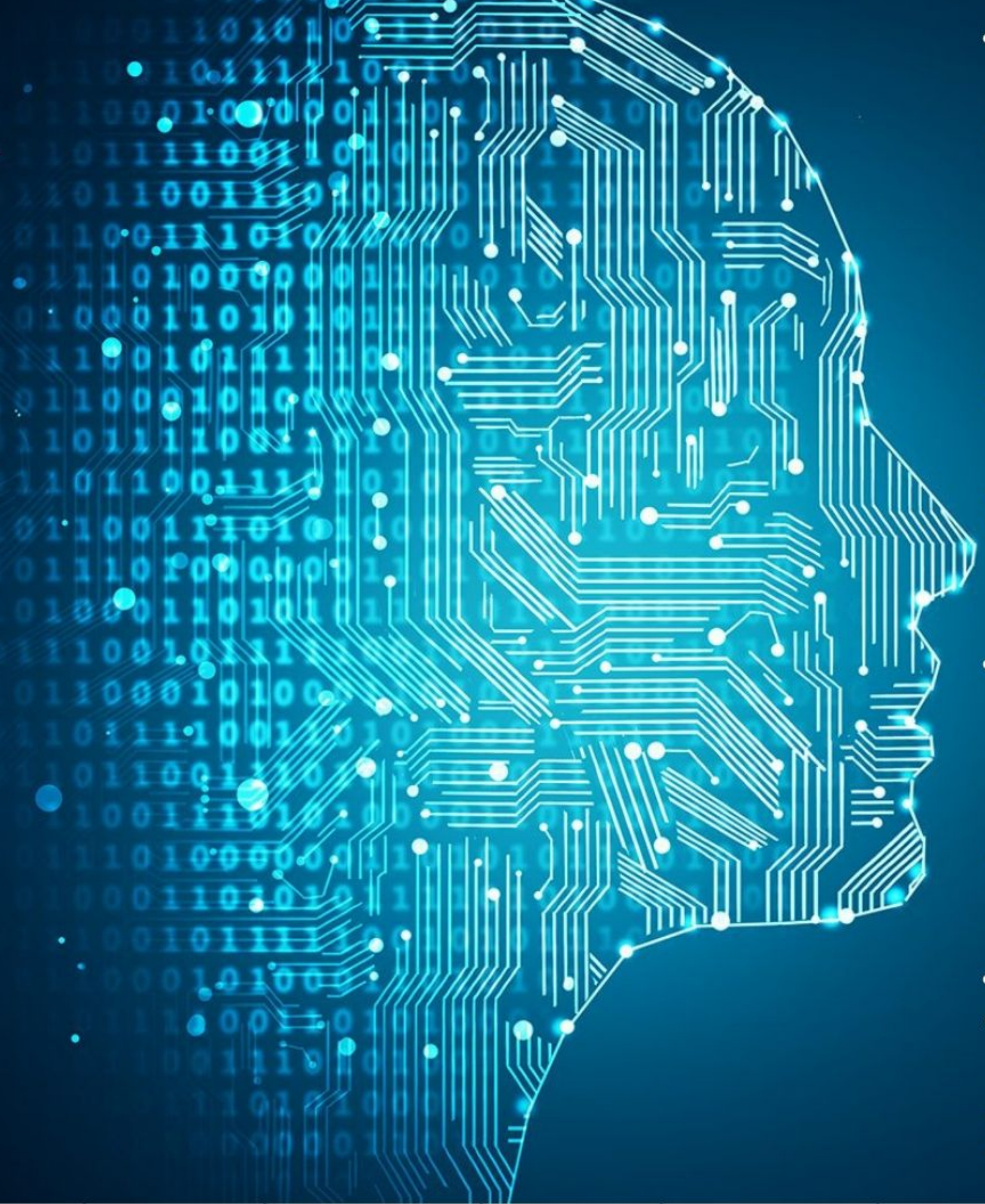


Machine Learning Vs Human Brain

Goals of today's event

Motivations:

- Scientific event with a subject transverse to the 4 departments
- Clear trend in our fields since already many years
- Already several use of ML in our researches

Goals:

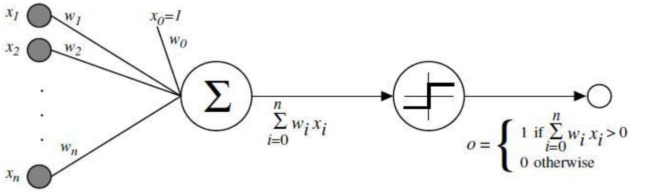
- Inform about activities done @ IPHC
- Starting point to sharing expertise?
- Network ?
- Common interests – Structuration ?

Morning's session

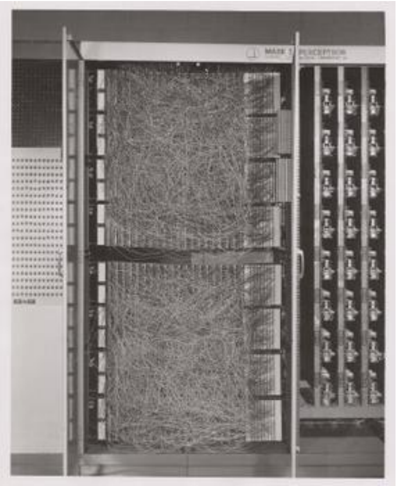
To be discussed during the round table

ML: an expanding field

- **Machine learning is not new**
  - First ideas already in 50's
  - History of Neural Network
    - First paper (1943)
    - First algo still in use (Perceptron): 1958
    - First annual meeting :NN for computing" (1985)
    - …. AlphaGo (Google DeepMind) win a professional player (2015) …

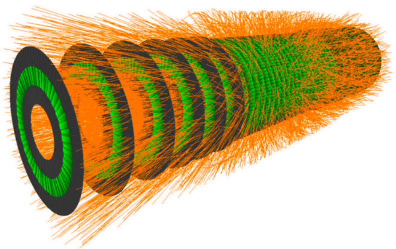


source

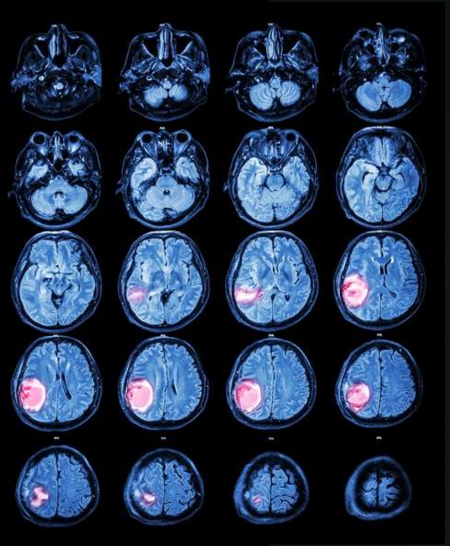


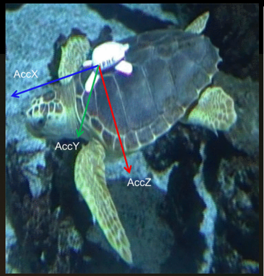
'Mark I Perceptron at the Cornell Aeronautical Laboratory', hardware implementation of the first Perceptron

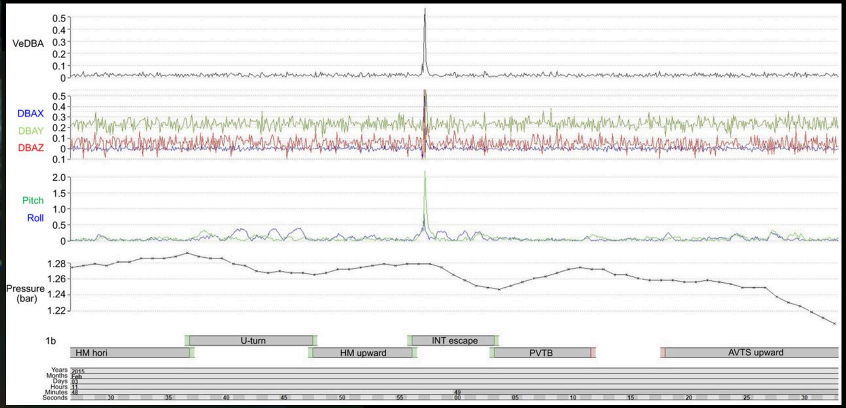
- **Expansion of the ML field**
  - Boost from big data
  - Large increase of computing resources (massive parallelisation)
  - ML implementation eased by user-friendly (open source) packages
  - Variety of applications (interests in tech compagnies and industry)
  - Dynamic field: many ML contests, …

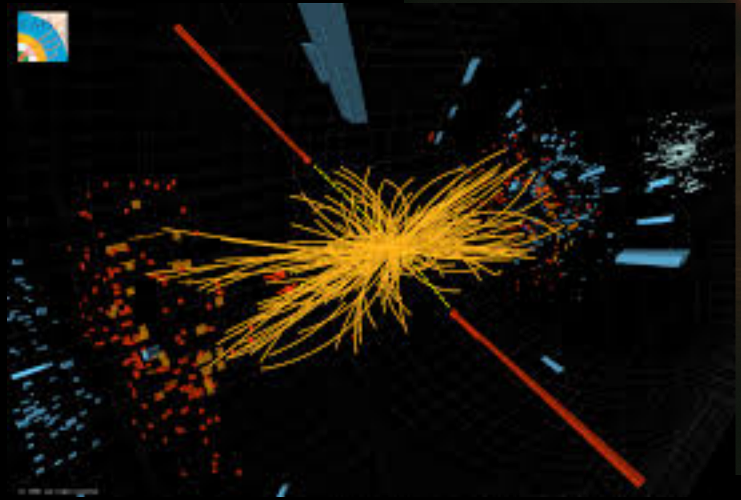


Machine Learning: *an e-crystal ball*





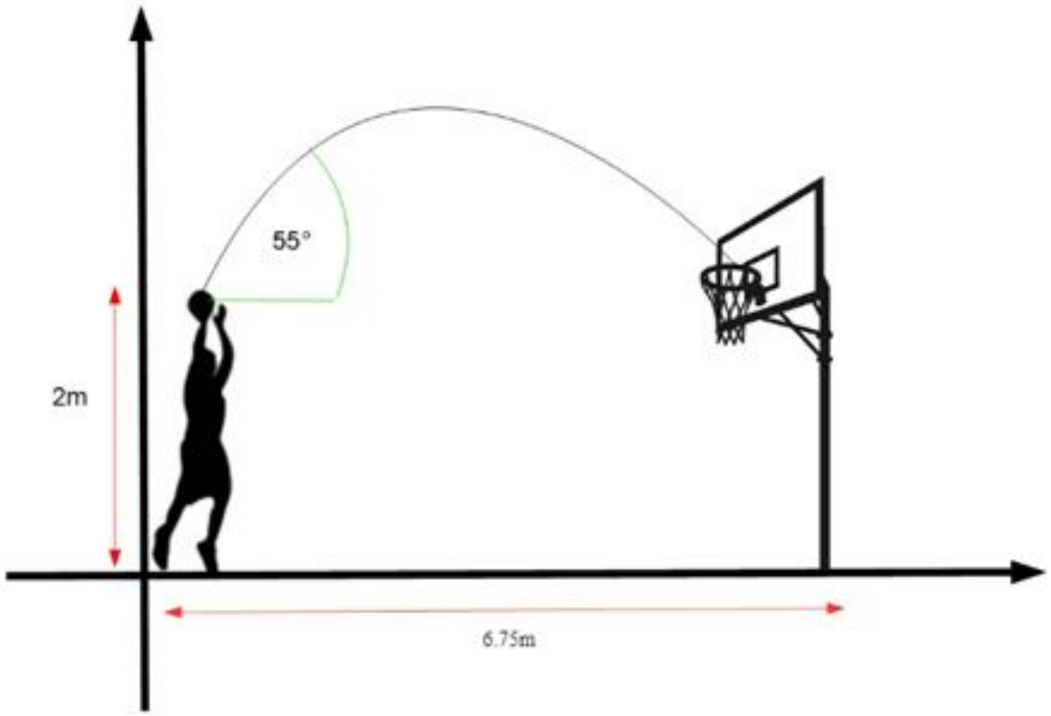




Use data to "**predict**" (estimate/infer)
a quantity / state / category /… on the studied object

Machine Learning: *an e-crystal ball*

Do we need to know gravitational low to play basketball ?



Children learn by experiment how to solve this problem without determining explicitly any physics law !

Machine Learning goals

ML goals:

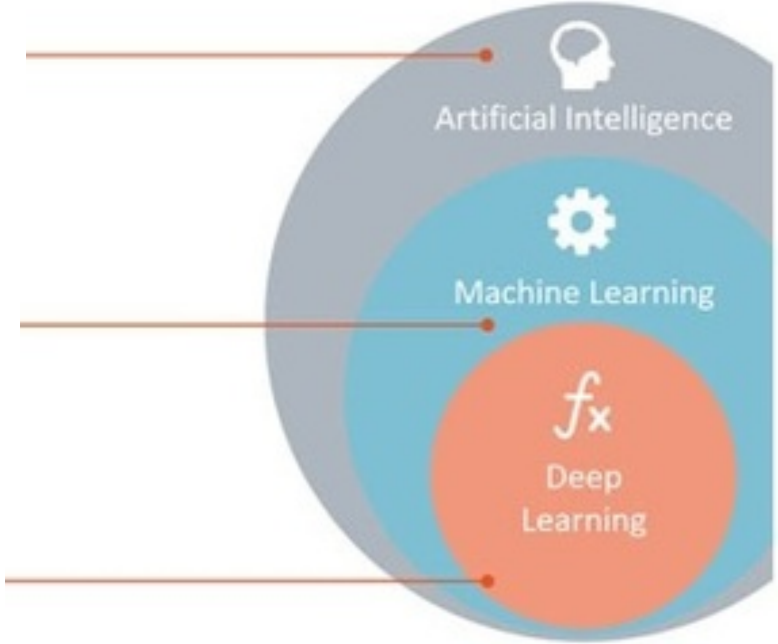
- Learning
- Finding patterns & trends
- Improving decision making
- Generating data
- Improving/optimizing “tasks”
- …

Pro(*con*)fusion of keywords …

A bit of clarification

Artificial Intelligence

Any technique which enables computers to mimic human behaviour. It uses ML but not only. Creates applications that react, adapt, ….
Ex: self-drive car, chatbot, …



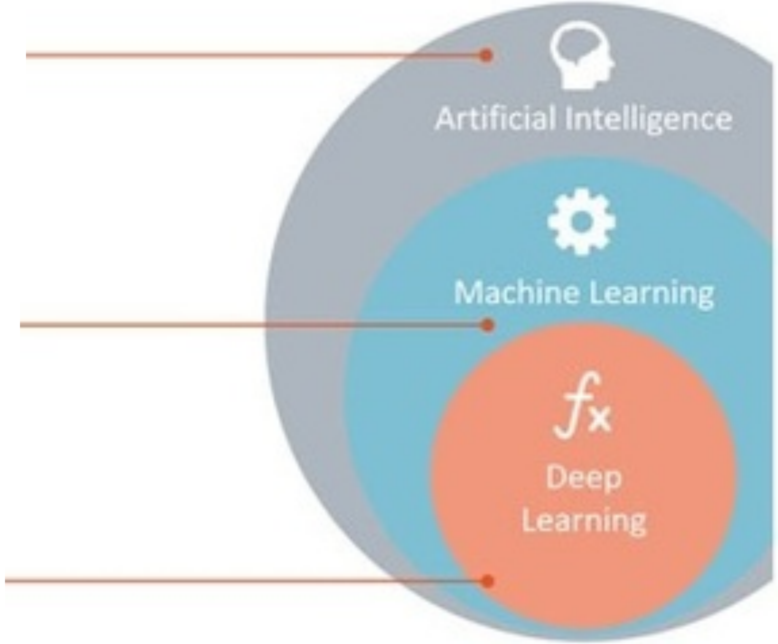
A bit of clarification

Artificial Intelligence

Any technique which enables computers to mimic human behaviour. It uses ML but not only. Creates applications that react, adapt, ….
Ex: self-drive car, chatbot, …

Machine Learning

Algorithms able to find patterns/features based on data using statistics with the aim to perform predictions



A bit of clarification

Artificial Intelligence

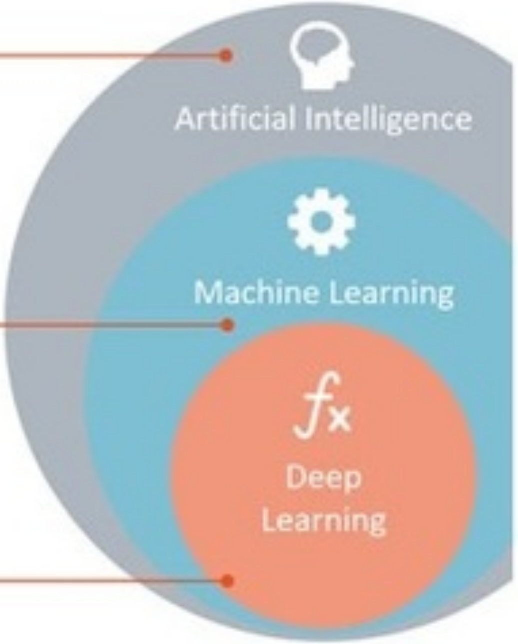
Any technique which enables computers to mimic human behaviour. It uses ML but not only. Creates applications that react, adapt, ….
Ex: self-drive car, chatbot, …

Machine Learning

Algorithms able to find patterns/features based on data using statistics with the aim to perform predictions

Deep Learning

Subset of ML which make use of multi-layer neural networks (Artificial NN, Convolutional NN, Recurrent NN, …)
"deep" → number of layers & neurons
→ many parameters to be estimated
→ need large datasets and long training period



A bit of clarification

Maths/Statistics

Programming

Biological inspired rules

Artificial Intelligence

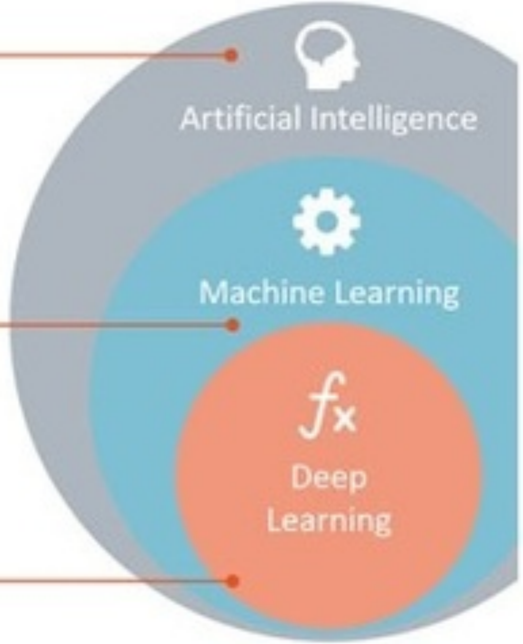
Any technique which enables computers to mimic human behaviour. It uses ML but not only. Creates applications that react, adapt, ….
Ex: self-drive car, chatbot, …

Machine Learning

Algorithms able to find patterns/features based on data using statistics with the aim to perform predictions

Deep Learning

Subset of ML which make use of multi-layer neural networks (Artificial NN, Convolutional NN, Recurrent NN, …)
"deep" → number of layers & neurons
→ many parameters to be estimated
→ need large datasets and long training period



A bit of clarification

- Maths/Statistics
- Programming
- Biological inspired rules

Big data

- Does not imply the use of ML
- ML complexity & performances depends on the size of the data used to learn

Artificial Intelligence

Any technique which enables computers to mimic human behaviour. It uses ML but not only. Creates applications that react, adapt, ….
Ex: self-drive car, chatbot, …

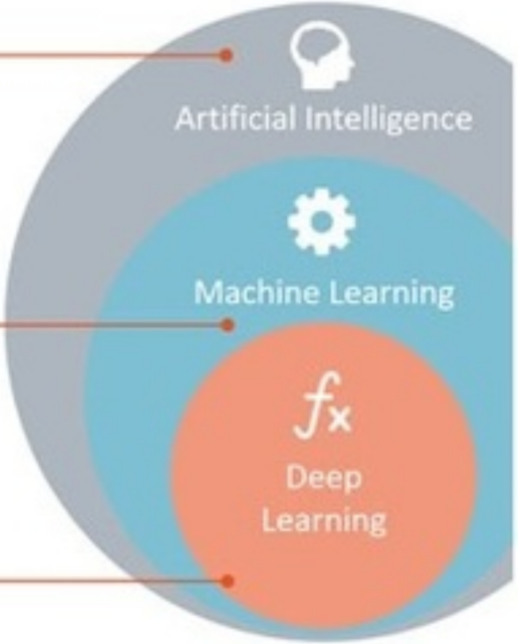
Machine Learning

Algorithms able to find patterns/features based on data using statistics with the aim to perform predictions

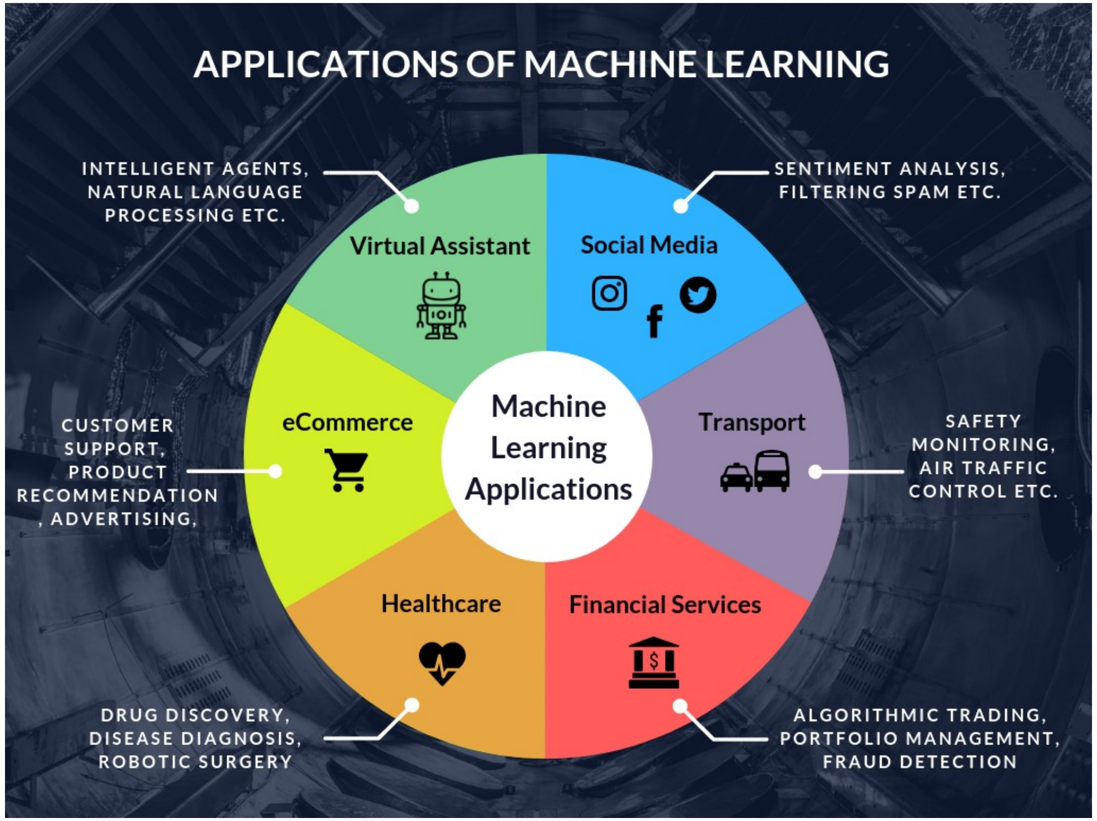
Deep Learning

Subset of ML which make use of multi-layer neural networks (Artificial NN, Convolutional NN, Recurrent NN, …)

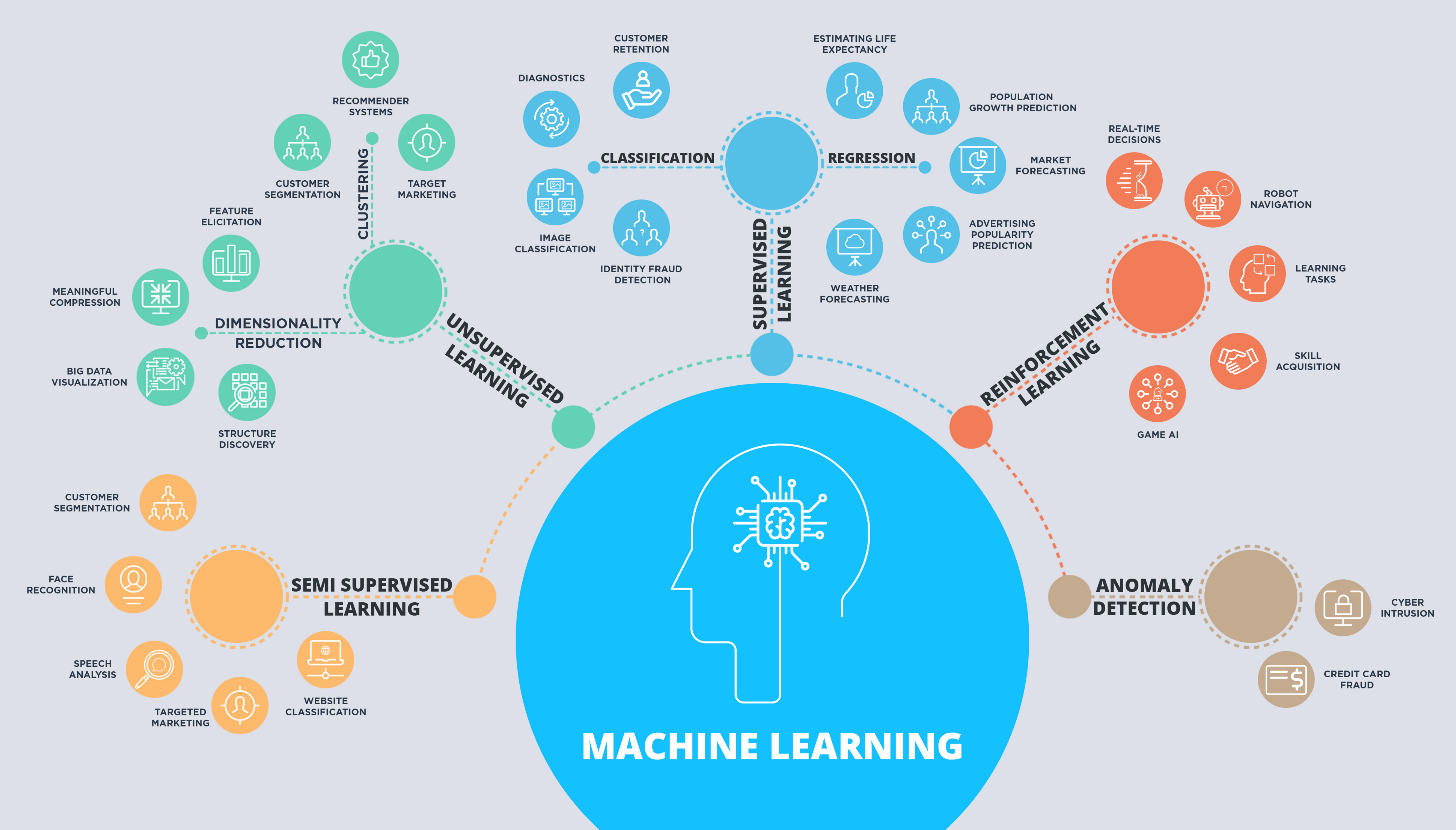
"deep" → number of layers & neurons
→ many parameters to be estimated
→ need large datasets and long training period



ML is everywhere in everyday's life



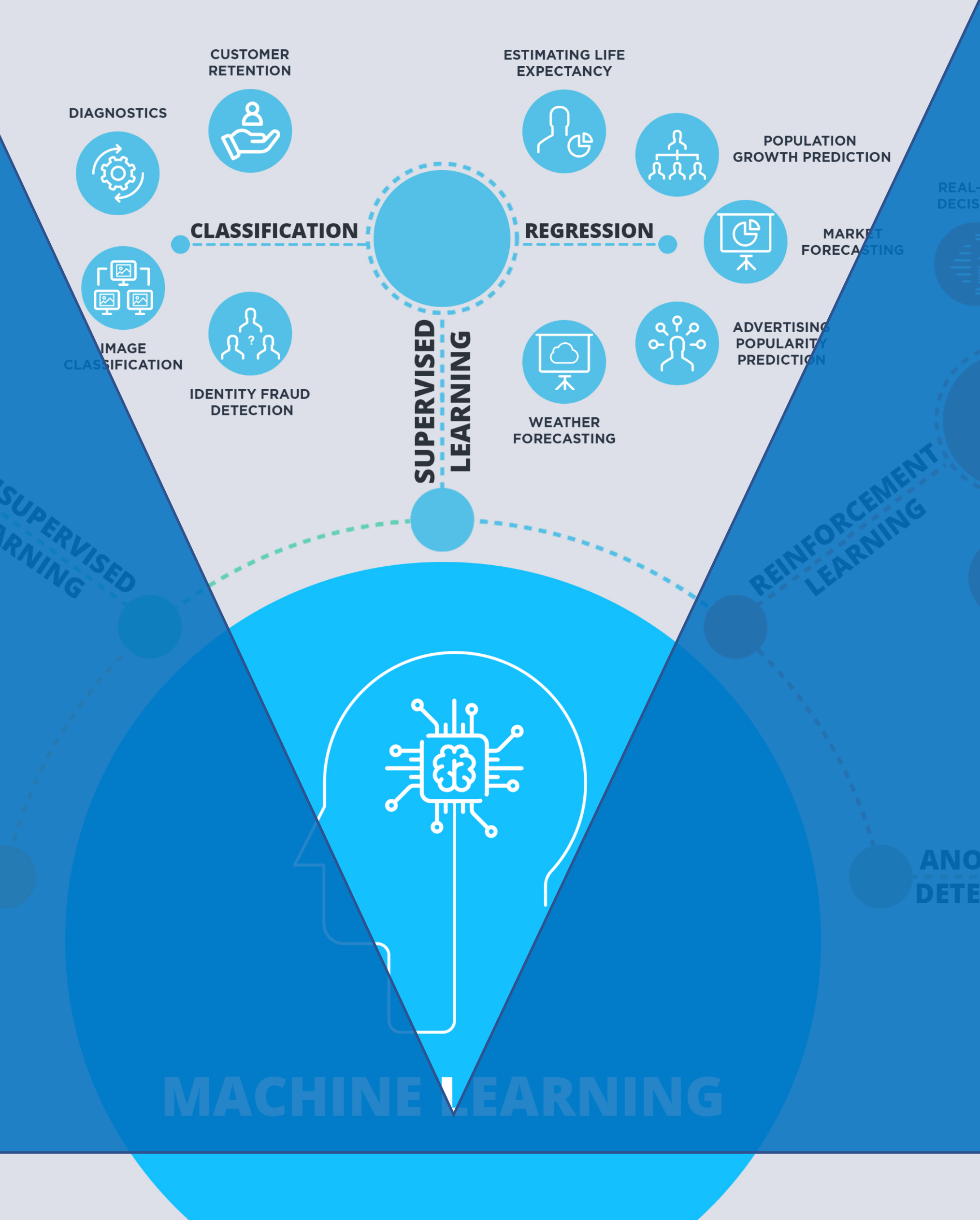
Class of problems solved by ML algorithms



Class of problems solved by ML algorithms

Supervised
vs
Unsupervised

Supervision means that label(s) are attached to data, *ie* the targeted value (discrete or continuous) is known in the training sample.



Class of problems solved by ML algorithms

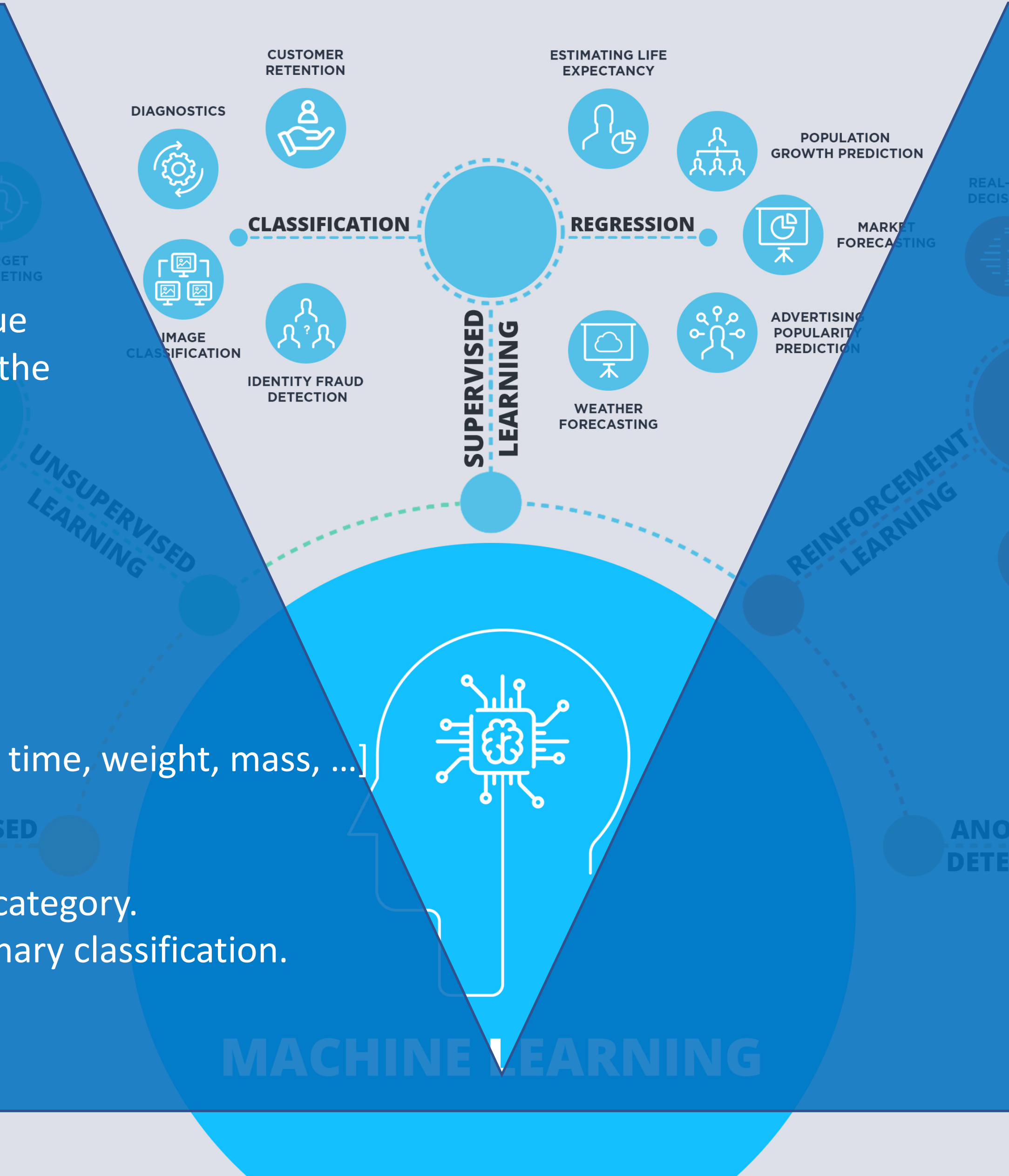
Supervised
vs
Unsupervised

Supervision means that label(s) are attached to data, *ie* the targeted value (discrete or continuous) is known in the training sample.

Regression
vs
Classification

<ins>Regression</ins>:
The target variable is continuous [ex: time, weight, mass, …]

<ins>Classification:</ins>
The target variable is a discrete, *ie* a category.
The simplest case correspond to a binary classification.
[ex: signal vs bkg, species, ..]



Class of problems solved by ML algorithms

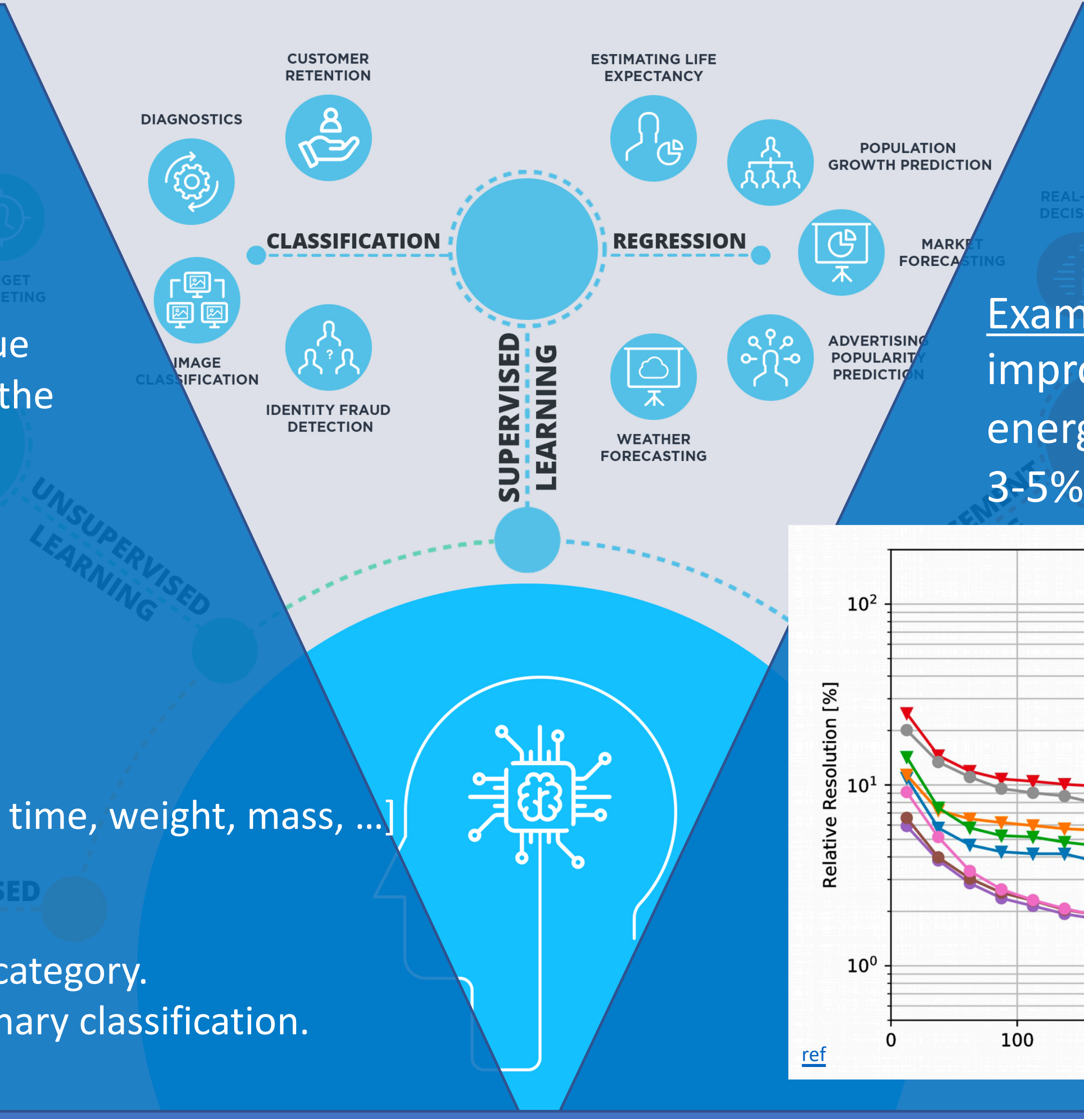
Supervised
vs
Unsupervised

Supervision means that label(s) are attached to data, *ie* the targeted value (discrete or continuous) is known in the training sample.

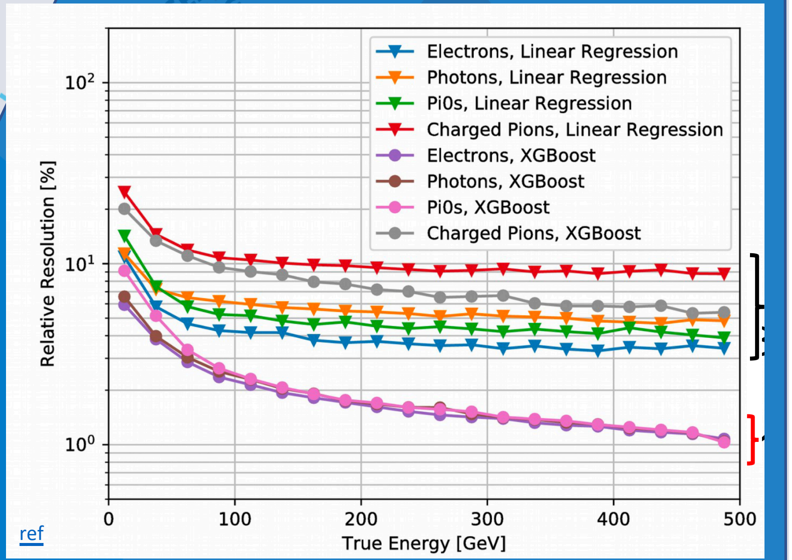
Regression
vs
Classification

<u>Regression</u>:
The target variable is continuous [ex: time, weight, mass, …]

<u>Classification</u>:
The target variable is a discrete, *ie* a category.
The simplest case correspond to a binary classification.
[ex: signal vs bkg, species, …]



<u>Example in HEP:</u>
improvement of particle energy resolution
3-5% → 1%



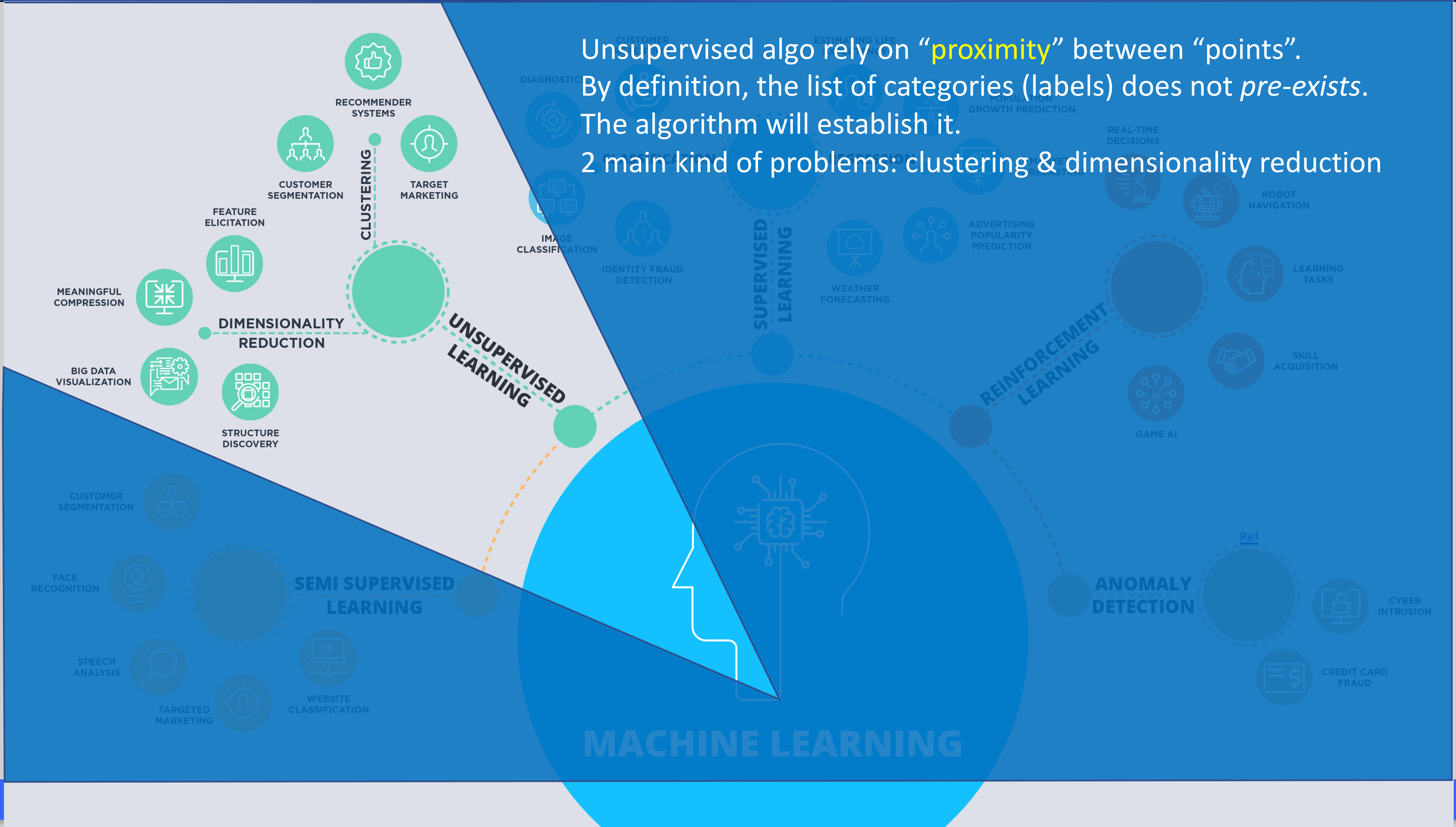
Traditional methods

ML methods
BDT

ML algo can improve "experimental performances"
Could be a cheaper/less demanding solution than building a new "apparatus" …..

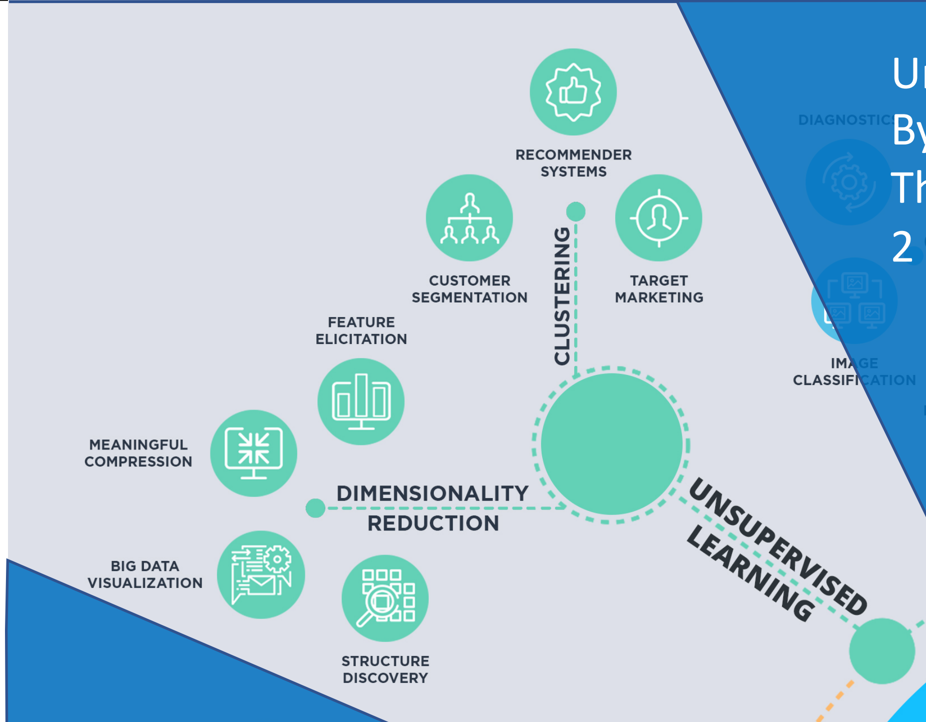
Class of problems solved by ML algorithms

Unsupervised algo rely on "proximity" between "points".
By definition, the list of categories (labels) does not *pre-exists*.
The algorithm will establish it.
2 main kind of problems: clustering & dimensionality reduction

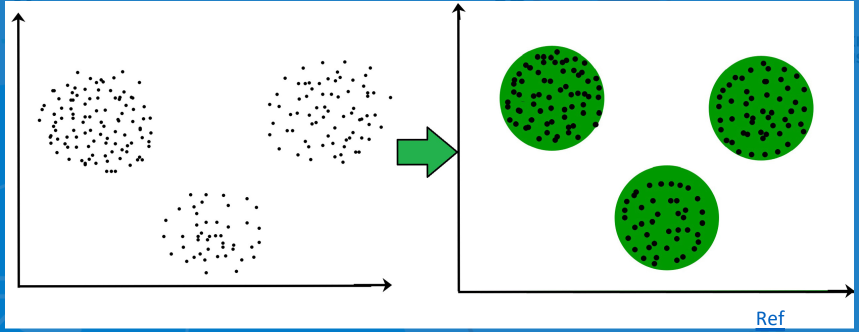


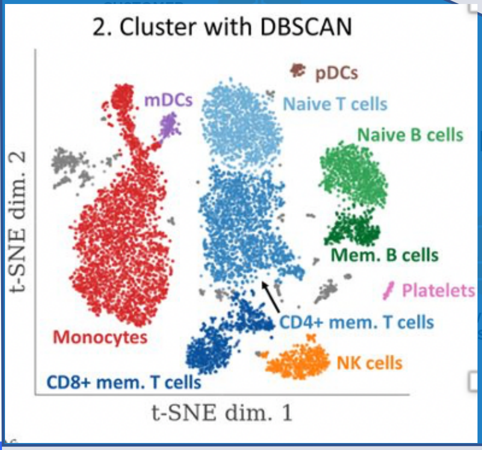
Class of problems solved by ML algorithms

Unsupervised algo rely on “proximity” between “points”.
By definition, the list of categories (labels) does not *pre-exists*.
The algorithm will establish it.
2 main kind of problems: clustering & dimensionality reduction



k-mean is an example of “geometric” algo.
Probabilistic approach also exists.





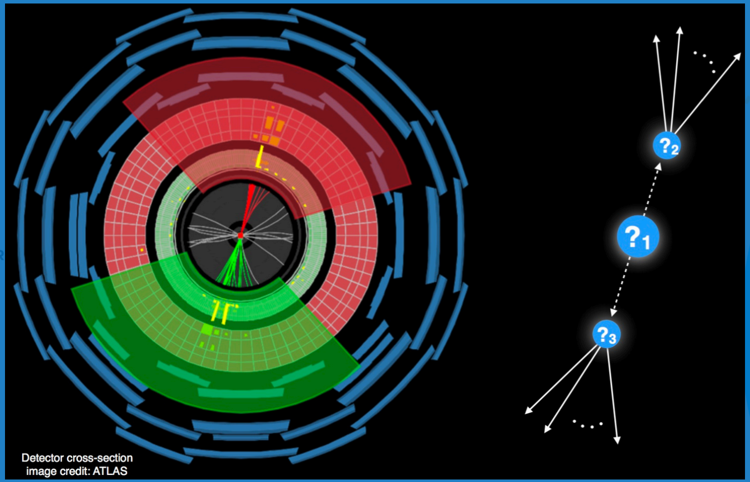
Applications by example in biology for “species” identification

Class of problems solved by ML algorithms

https://lhco2020.github.io/homepage/

Example in HEP:

- Used to detector anomalies
- Attempt to use it to search for new phenomena (*example here*)



Anomaly = unexpected experimental signature (sign of new phenomena)

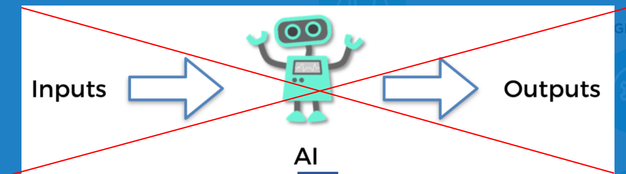
MACHINE LEARNING

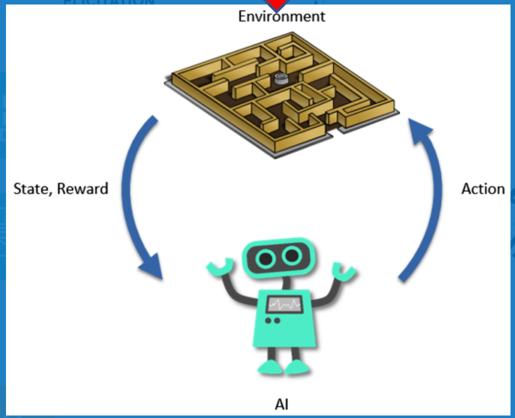
ANOMALY DETECTION

CYBER INTRUSION

CREDIT CARD FRAUD

Class of problems solved by ML algorithms





- Feedback loop
- Introduce a reward
- Continuous improvement

Example of applications in subatomic physics:

Finding optimal beam characteristics

List of questions before starting using ML

1- Should I need to use ML ?

- Algorithm complexity, dataset size, computing resources used

are not guarantees that the performances will be better than what you could achieve with more "**traditional methods**" !

- **ML is not the solution to all problems**: If you have an analytical model (*scientifically motivated*) describing your data, why should you spend resources to approximate it with ML ? (*discussion about GAN*)

→ Useful to go beyond current knowledge in a data-driven way

- Gain vs effort
  - → Can I expect a large improvement compare to current method ?
  - → How much effort is needed to implement a ML solution?
  - → Do I have enough resources to perform the training ?

List of questions before starting using ML

1- Should I need to use ML ?

- Algorithm complexity, dataset size, computing resources used

are not guarantees that the performances will be better than what you could achieve with more “traditional methods” !

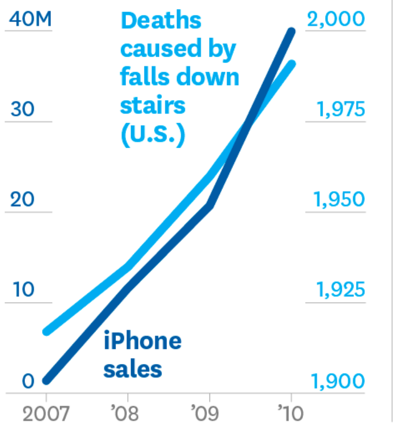
- **ML is not the solution to all problems:**
- **Gain vs effort**

ML CANNOT be “blindly used”:

ML does not replace the human knowledge / judgment

- Needs human guidance in the process of applying ML techniques
  - Ex: ”spurious correlation in large datasets” (p-score)
- For interpretation:
  - correlation does imply not causal relation
  - Remark: hard to interpret the “ML model”

MORE IPHONES MEANS MORE PEOPLE DIE FROM FALLING DOWN STAIRS

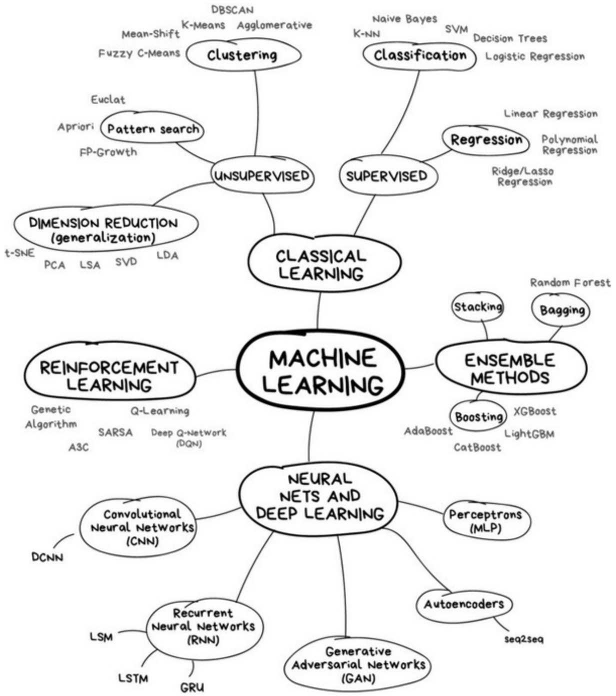


SOURCE TYLERVIGEN.COM
FROM “BEWARE SPURIOUS CORRELATIONS,” JUNE 2015

List of questions before starting using ML

2- Which category of ML algorithms should I use ?

- We have presented 5 class of algorithms
- Subcategories for some of the classes
- Many algorithms are devoted to a given class of problem



3- What is the "nature" of the data to analyze ?

4 types of data:

- Numerical data
- Categorical data
- Time series (video, sound, …)
- Text

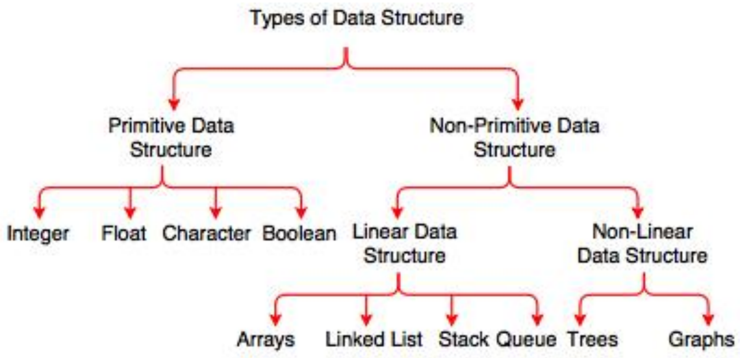


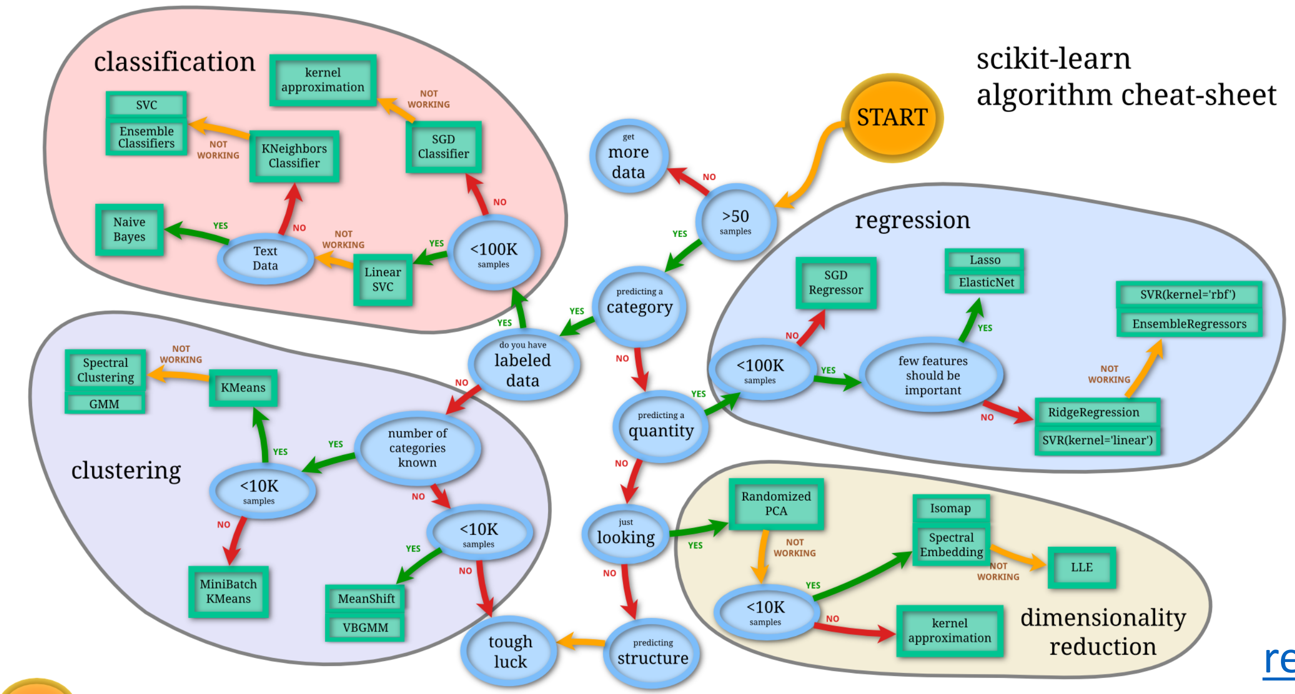
Fig. Types of Data Structure

List of questions before starting using ML

4- How large is our sample ?

- Choose an algorithm adapted to our sample
- Deep learning requires very large datasets

You should now have a list of algorithms that may suit our needs.



Back

ref

What's next ?

ML implementation

<u>Answer:</u> finding a (free / open source /maintained) implementation !

- (Google) Tensorflow (2015)/Keras
  - Fast-growing easy-to-use python lib (but also in C++, …)
  - Allows applications of deep-learning models
  - Interface to Tensorflow, Theano backends → GPU support

- Scikit-learn (2007)
  - Python lib that implements many (non-deep) techniques
  - A lot of data preprocessing & statistics tools

- TMVA (used in subatomic physics – ROOT based)

But also:
- Torch / PyTorch
- Caffee (C++/python)
- Accord.net (C++)
- R
- …

Can be discussed this afternoon

What's next ?

Hardware

<u>Answer:</u> where to run the ML training ?

- **Use our laptop:**
  several ML-analyses can be done on (from few min to few hours)
  - Multi-threading

- **Use dedicated GPU(s)**
  reasonable dataset – large computing time
  - Several algorithms profit a lot from GPU parallelization
  - From few 100's to few 1000's euros

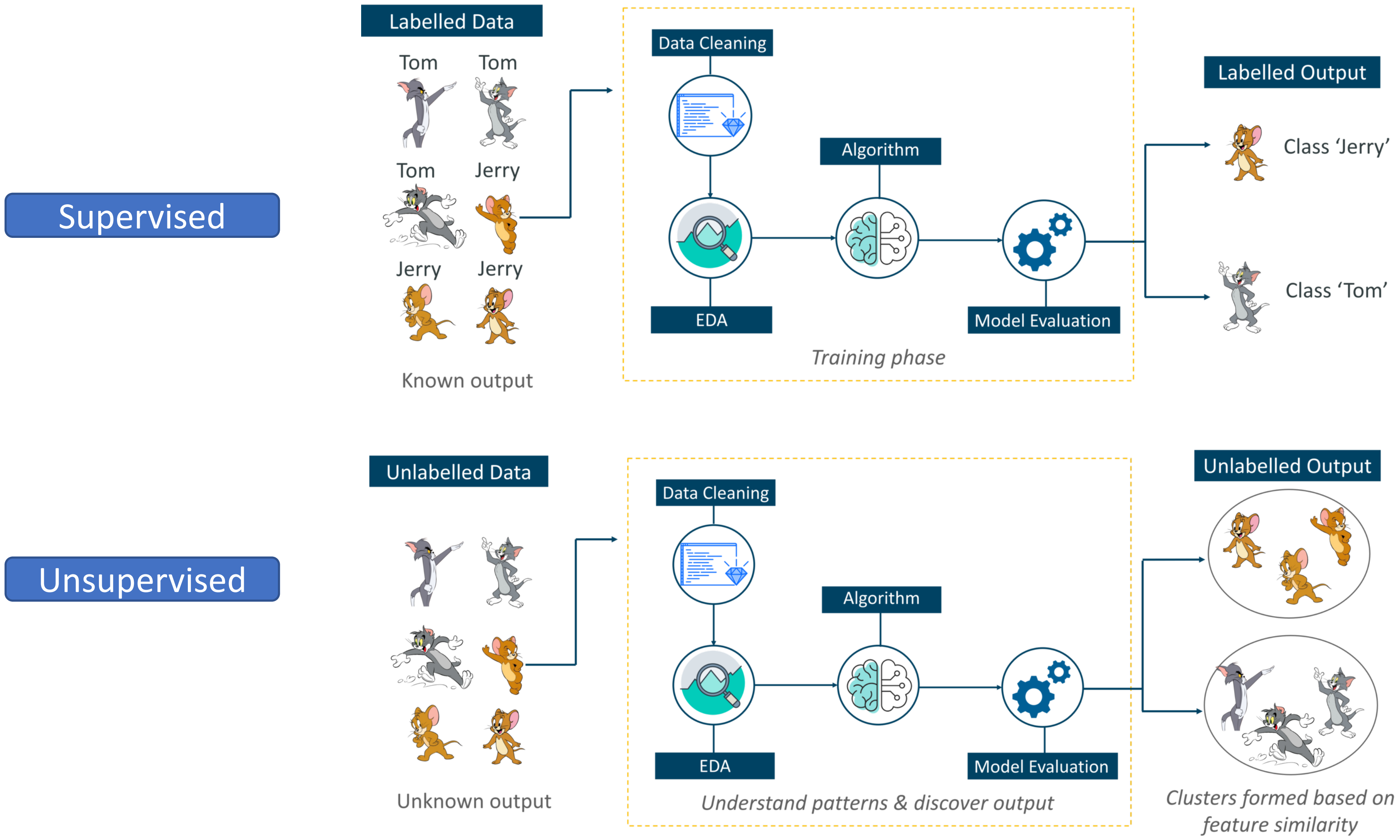
- **Use server/computing center**
  large data / complex models
  - Tier2
  - Mesocentre
  - CC-in2p3
  - CERN
  - ….

Can be discussed this afternoon

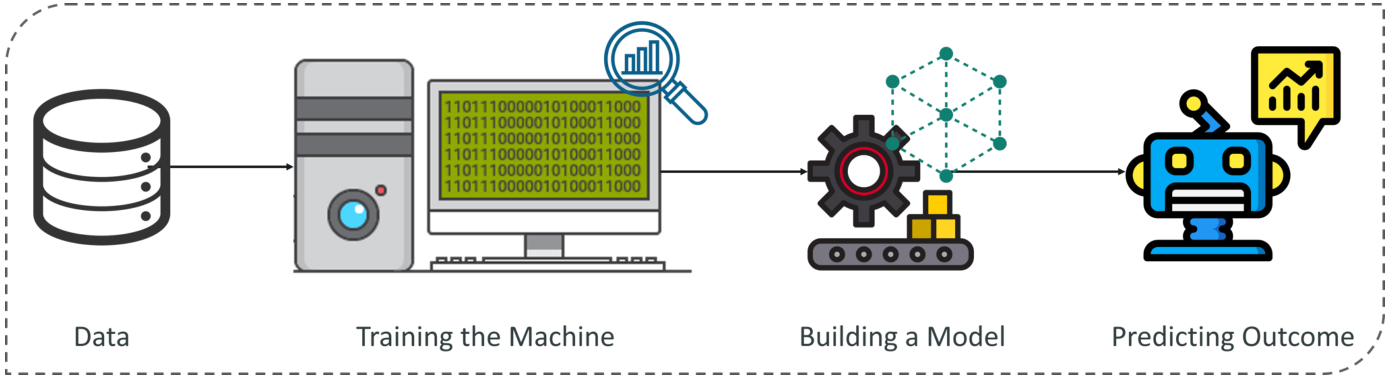
What's next ?

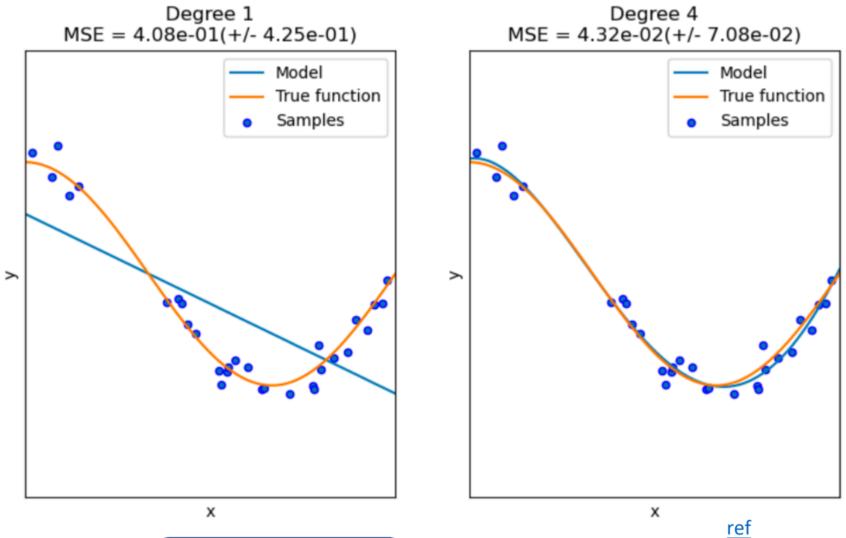
ML workflow

Ref



Supervised ML workflow

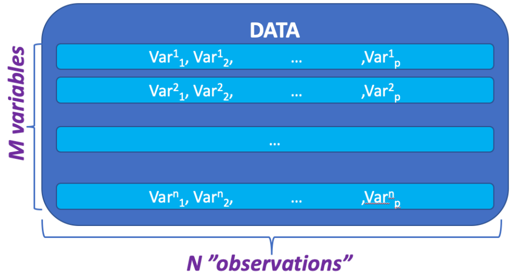




ref

Inputs (X) → Model (f) → Output (Y)

- N “samples”.
- M “features” per sample [dimension]
- (*predicator variables*)



- Predictive function: y = f(x)
- Depends on the choice of the algo
- Configuration (ex: NN: #layers, #nodes, …)
- Have **many** parameters to be determined

Find an function f(x)
that approximates the function
F such that
Y = F(x)

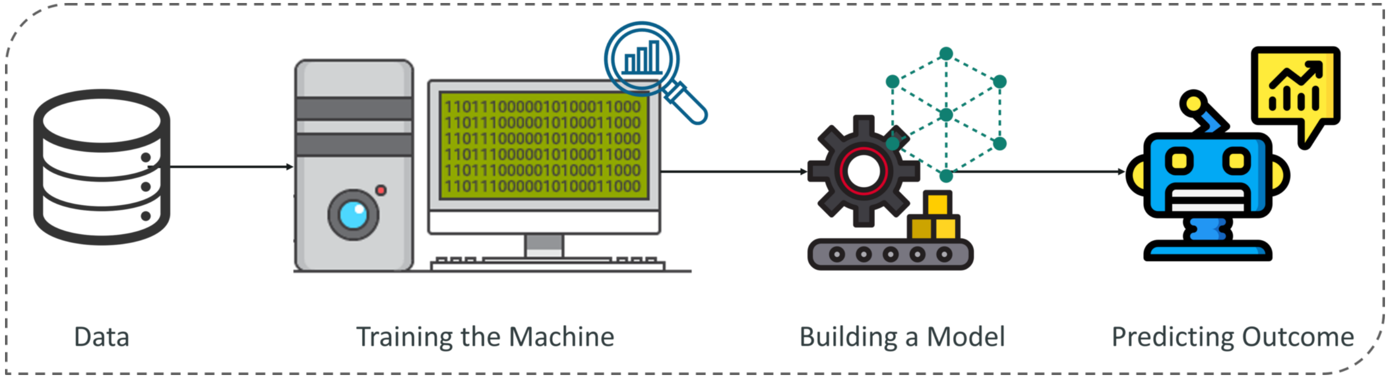
- One (or few) response (*target*) variable
- Can be discrete or continuous

Supervised ML workflow

Ref

Requirements:

Step 1: Define the **objective**
→choice of algo/implementation
→choice of hardware chosen
Step 2: Data gathering

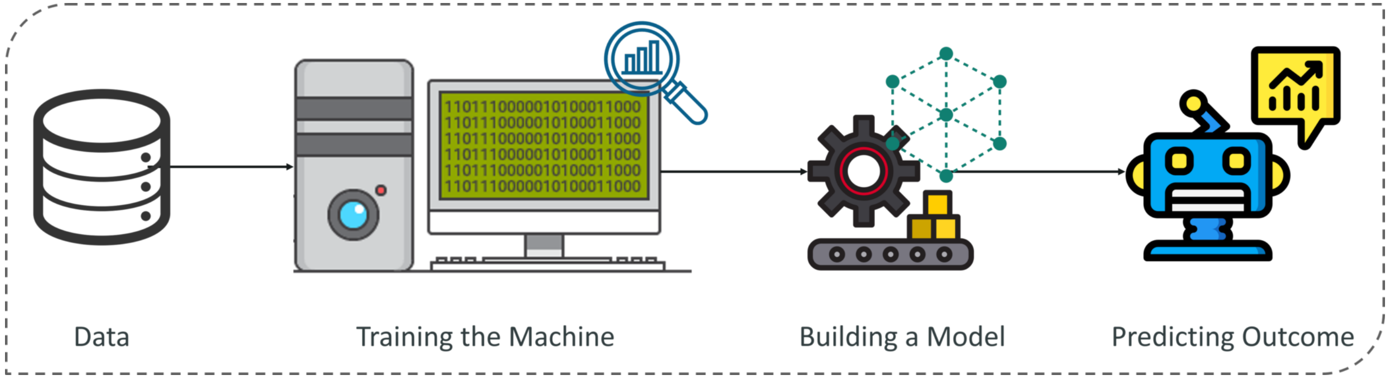


Step 3: **Data preparation**
data cleaning, filtering, binning, transformation …

Step 4: Exploratory Data Analysis (EDA)

Understand patterns, trends, correlations, ..
Goal: Selection input variables (features) & "samples"
Advices:
→ Avoid variables too sensitive to noise, uncertainties, …
→ Avoid variable badly described by our simulation if the dataset is simulation based
→ Use a representative "samples"

Supervised ML workflow



CPU
- Small models
- Small datasets
- Useful for design space exploration

GPU
- Medium-to-large models, datasets
- Image, video processing
- Application on CUDA or OpenCL

TPU
- Matrix computations
- Dense vector processing
- No custom TensorFlow operations

Tensor PU
Dev. By google
2016

FPGA
- Large datasets, models
- Compute intensive applications
- High performance, high perf./cost ratio

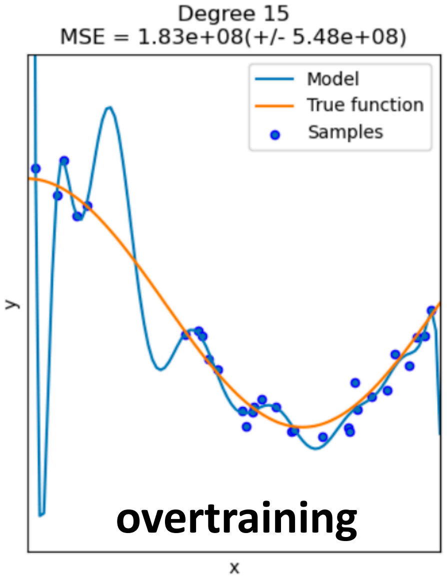
Step 5: Building a Machine Learning Model == Model training
- Define a (large) subset of data == **training sample**
- Most CPU/GPU demanding task
- **Goal**: determining the parameters of the model
- **Advices**:

→ sample should be large enough (depends on model complexity)

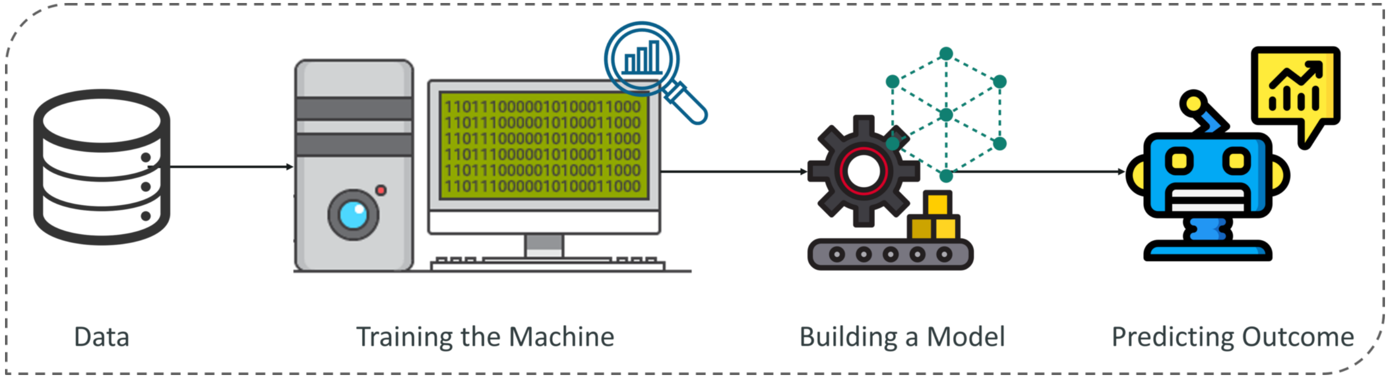
→ Computation time can be speed-up depending on both the hardware architecture chosen & the model

→ Remark: processing does not necessarily linearly scale with #threads !

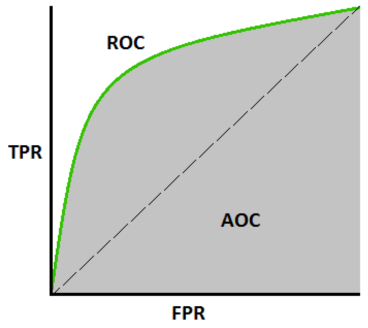
→ Check that there is no **overtraining**



Supervised ML workflow



[Ref](#)



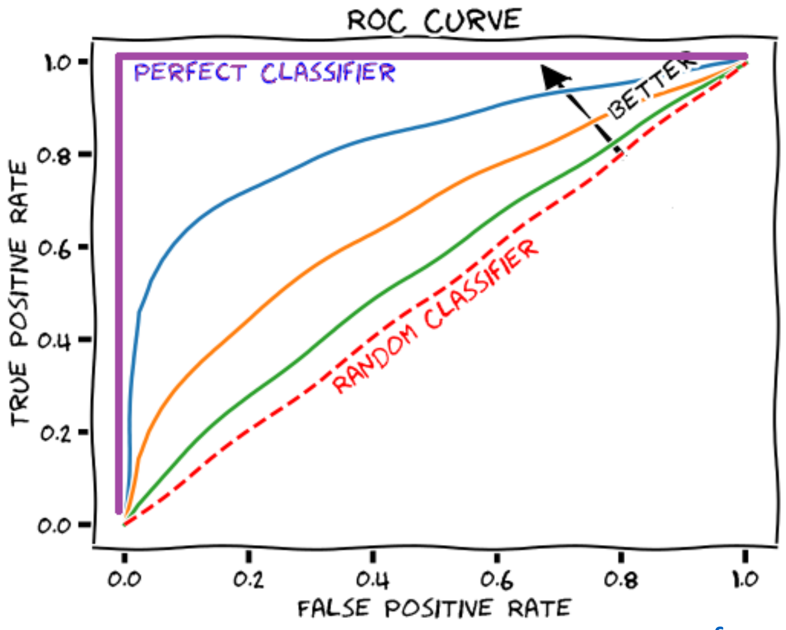
Step 6: Model Evaluation

- Define a (small / independent) subset of data == **test sample**
- Evaluate the performances
  - True vs False Positive Rage
  - ROC: Receiver Operating Characteristic
  - AUC: Area under the curve
- Cross-validation

  → Avoid overtraining

Cross-validation

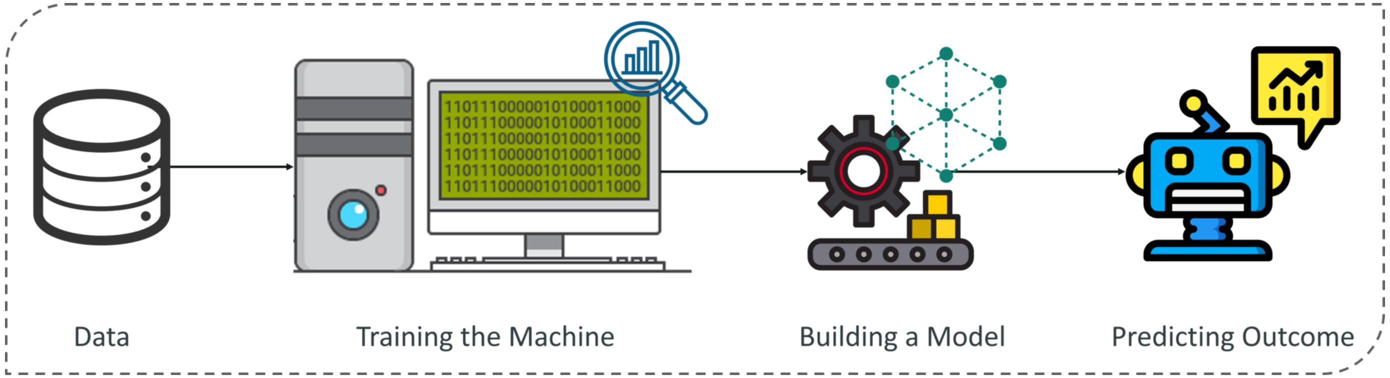
Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test



[ref](#)

Supervised ML workflow

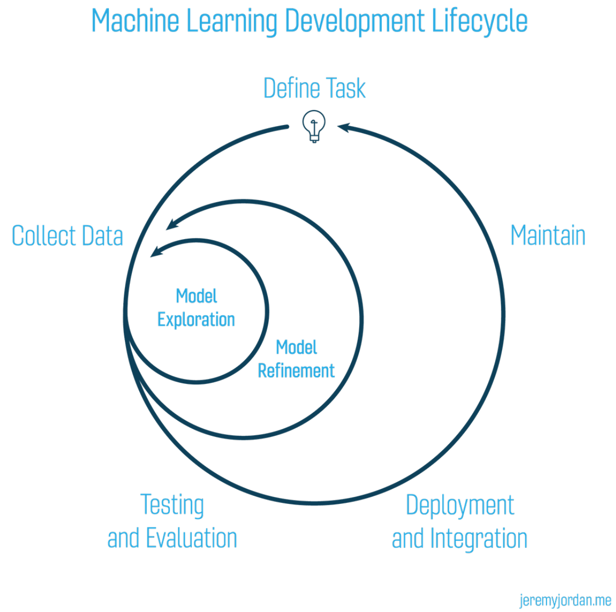
Ref



Step 7: **Model optimization (refinement)**

- (Hyper-)parameter tuning (ex with NN: #layers, #nodes, ..)
- Test/Optimize feature selection (ex: change/add input var)
- **Imply retraining** (CPU/GPU demanding)

→ This **loop** can be done several times …



Step 8: Predictions … !
→ algo deployment/integration/maintenance

Remark: Retraining can be needed (even if inputs/parameters are fixed) if the model is applied on data collected in new conditions (new sensors/calibration/alignment etc …)

Networks & training

Networks

- ML collaborations @ IN2P3
- IML (LHC ML working group)
- …

Training

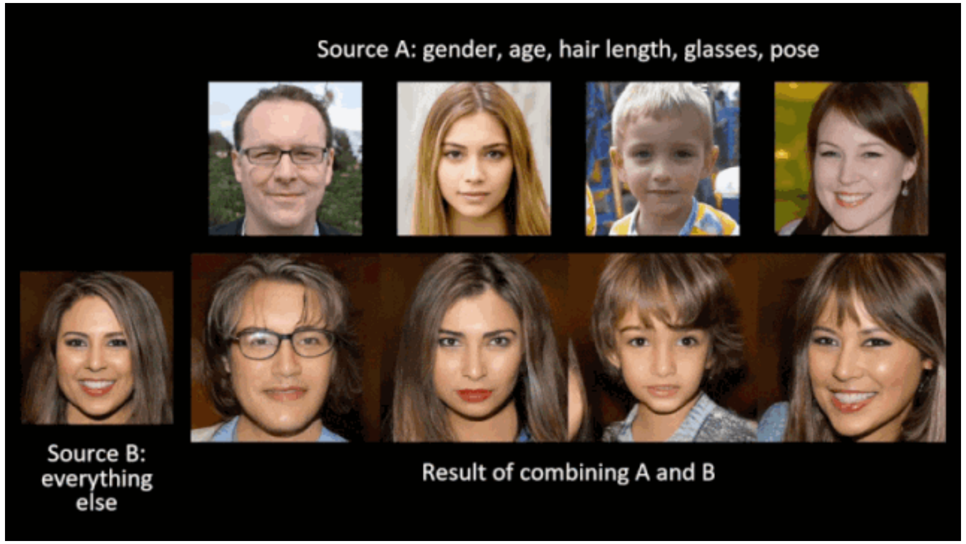
- Master "Big data and ML" @ unistra
- IN2P3 school of statistics
- Ecole doctorale (ED 182)
- MOOC
- …

Can be discussed this afternoon

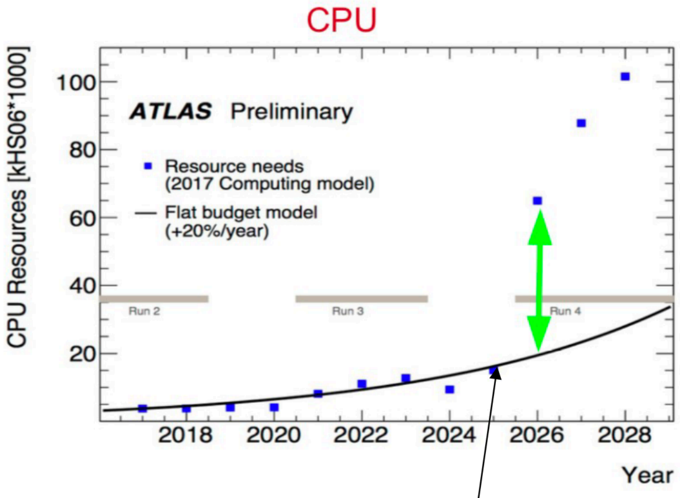
GAN: Generative Adversarial Network

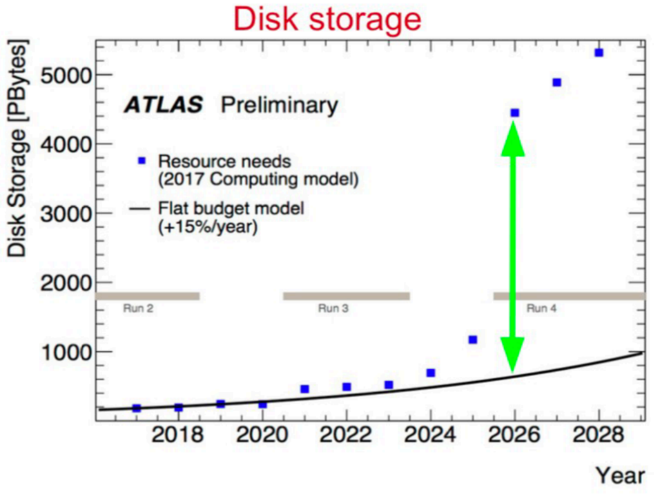
Technique used to generate “realistic human face”

Can have application in HEP to **speed-up simulation**
- Simulation of proton-proton collision @ LHC
- Simulation of detector response



Ref





Dominated by : calorimeter simulation and tracking

09:00 → 09:20 **Accueil et Introduction** 20m
Orateurs: Eric CHABERT (IPHC/UDS), Pierre Van Hove (CNRS)

09:20 → 10:20 **Exploitation des données de détecteurs**

09:20 **Bio-loggers, utilisation de "Convolutional Neural Network"** 15m
Orateur: Lorène Jeantet

09:35 **Exploitation des données des bio-loggers** 25m
Orateur: Marie-Amélie Forin-Wiart

10:00 **Capteurs CMOS et réseau de neurones** 20m
Orateur: auguste besson (Institut Pluridisciplinaire Hubert Curien)

10:20 → 11:00 **Particules et collisions**

10:20 **Reconstruction des événements à l'aide de ML en physique des particules** 25m
Orateur: Emery Nibigira (LPC Clermont)

10:45 **Traitement global d'événements sur collisionneur: utilisation de DNN** 15m
Orateur: Giulio Dujany (CNRS - IPHC)

11:00 → 11:20 **Pause pommes** 20m

11:20 → 11:55 **Traitement d'image**

11:20 **Reconnaissance d'espèces: image et vidéo** 20m
Orateur: Claire Saraux

11:40 **Projet dessin** 15m
Orateurs: Cédric Sueur (Institut Pluridisciplinaire Hubert Curien), Marie Pelé

11:55 → 12:15 **Spectrométrie de masse**

11:55 **Utilisation de ML et études mathématiques pour le traitement des données de spectrométrie de masse** 20m
Orateurs: Alexandre Burel (CNRS), Marie Chion

12:15 → 12:35 **Support**

12:15 **Informations, conseils et développements du service informatique** 15m
Orateurs: Jerome Pansanel (IPHC - CNRS), sebastien geiger (IPHC IN2P3)

12:35 → 14:00 **Déjeuner** 1h 25m

14:00 → 16:00 **Table ronde**

Fils de discussions envisagés:
-Prochaines étapes
-Retour d'expériences
-Conseils/suggestions du service info

Présidents de session: Eric CHABERT (IPHC/UDS), Pierre Van Hove (CNRS)

Beyond activities presented this morning, additional are done

Non exhaustive list:

- Clustering in CMOS pixels with NN (*DRHIM, Finck & al*)
- Use of AI to improve detector performances in the context of PET (tomography) (*DRHIM, Brasse & al*)
- Automatic ML-based localisation of radioactive contamination zone with an autonomous drone (learning based on MC) (*DRS, Arbor & al*)