SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

Introduction to Data Science

Domitilla Brandoni, CINECA

06/2023

Univerza v Ljubljani

TECHNISCHE UNIVERSITÄT WIEN

CINECA

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission.
This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

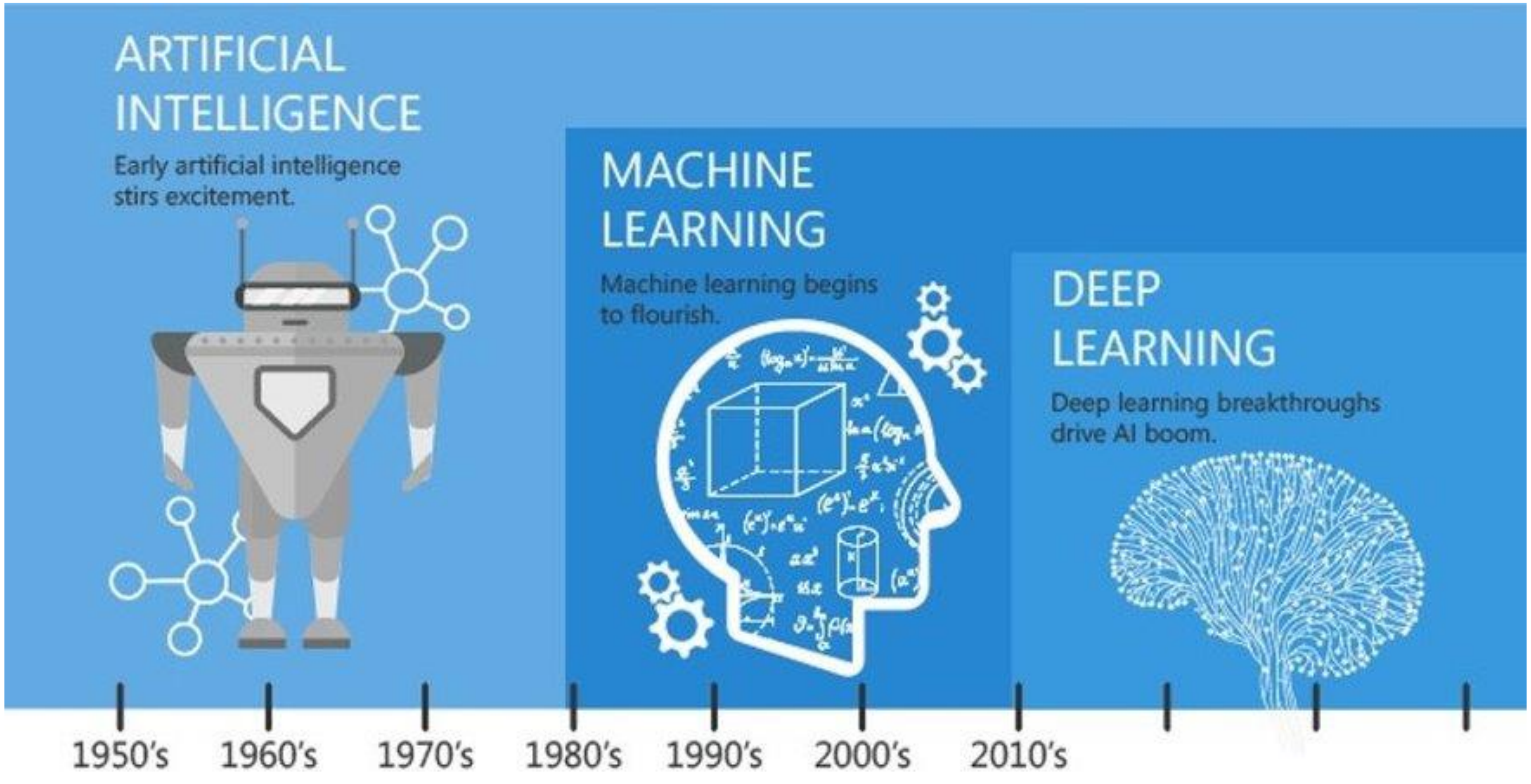
Definition of machine learning

Domitilla Brandoni, CINECA

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*.
Tom Mitchell, 1997

Example: chess playing

T = playing chess
E = playing games of checkers
P = probability to win next game
MACHINE LEARNING =LEARNING BY PLAYING


ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.
MACHINE LEARNING
Machine learning begins to flourish.
DEEP LEARNING
Deep learning breakthroughs drive AI boom.
1950's
1960's
1970's
1980's
1990's
2000's
2010's

ML in daily life

Domitilla Brandoni, CINECA

- Spam filters
- Face recognition, pattern recognition, speech recognition (Apple's FaceID, Android's Face Unlock, surveillance, medical images)
- Self-customized programs (e.g. Netflix)
- Predictive maintenance
- ChatGPT
- Language translation (deepl, quillbot)
- Agriculture
- Cybersecurity

Can machine think?

Domitilla Brandoni, CINECA

The concepts of '**learning algorithms**', '**artificial intelligence**' can be misleading but …

The aim is not creating machine that are able to think, the aim is creating machine that can **act indistinguishably from a thinker in a SPECIFIC situation,** for a SPECIFIC task
(Alan Turing, Stevan Robert Harnad)

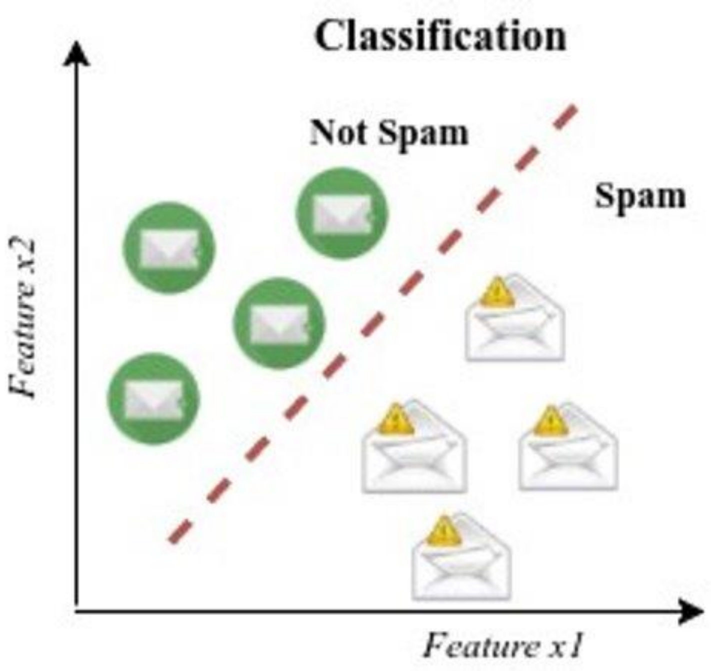
Types of machine learning

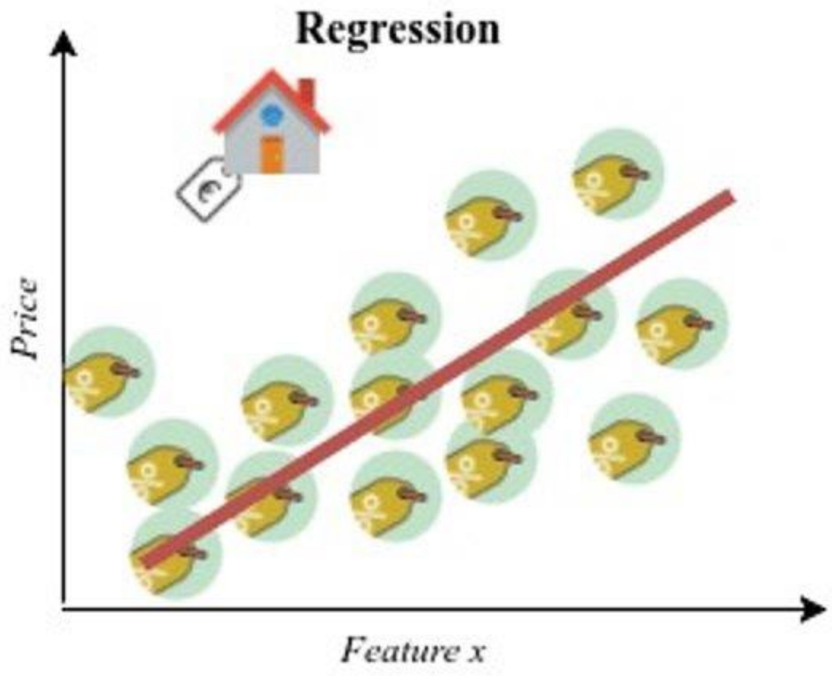
Domitilla Brandoni, CINECA

SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

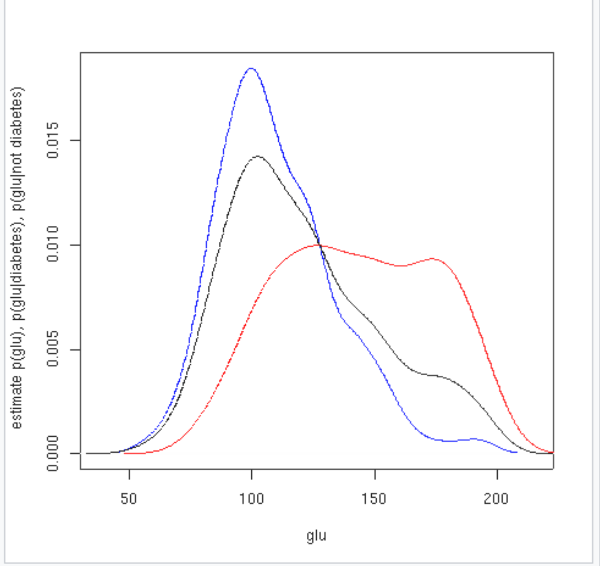
- Supervised learning
- Unsupervised learning
- Reinforcement learning
- Semi-supervised learning
- Self-supervised learning

Machine learning problem where all the data are labelled



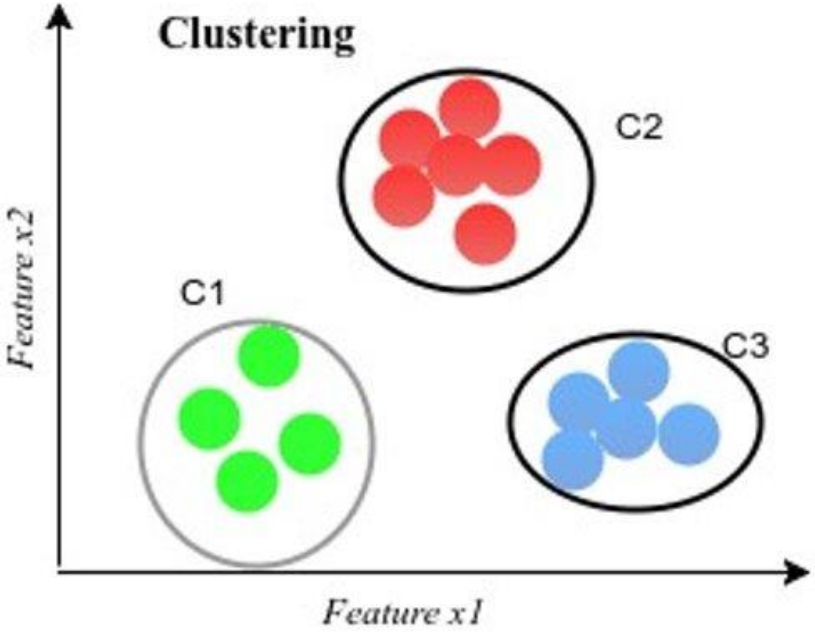


Machine learning problem where all the data are NOT labelled



Estimated density of *p* (glu | diabetes=1) (red), *p* (glu | diabetes=0) (blue), and *p* (glu) (black)

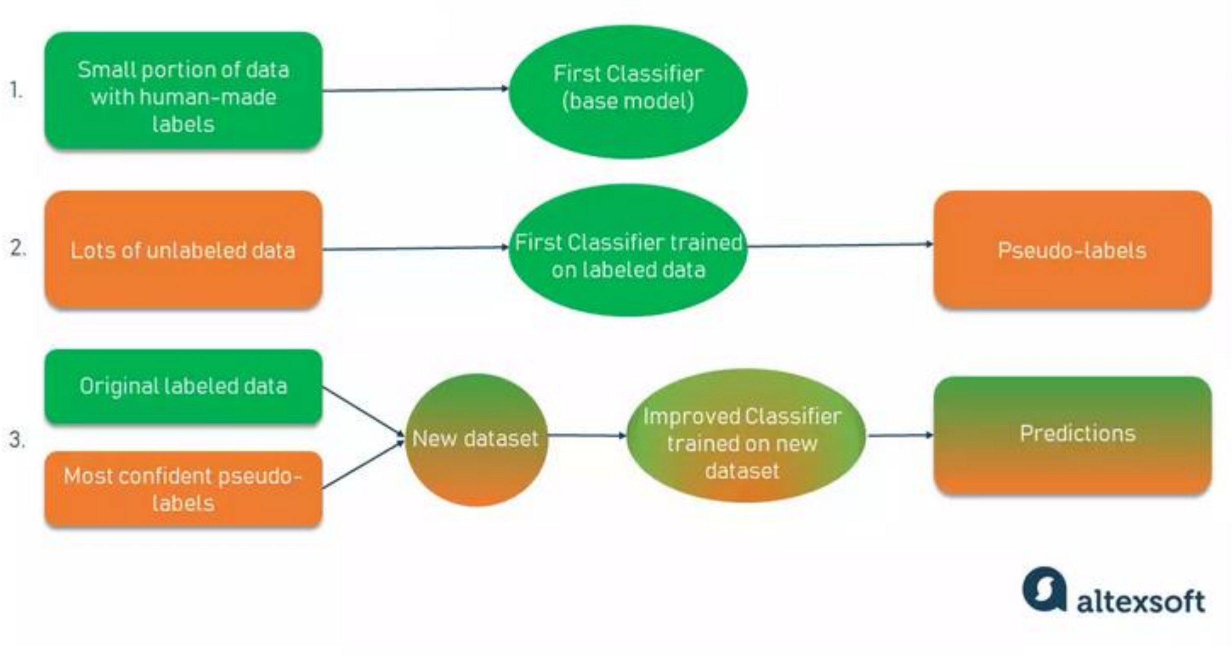
credits: Wikipedia



Semi-supervised learning

Domitilla Brandoni, CINECA

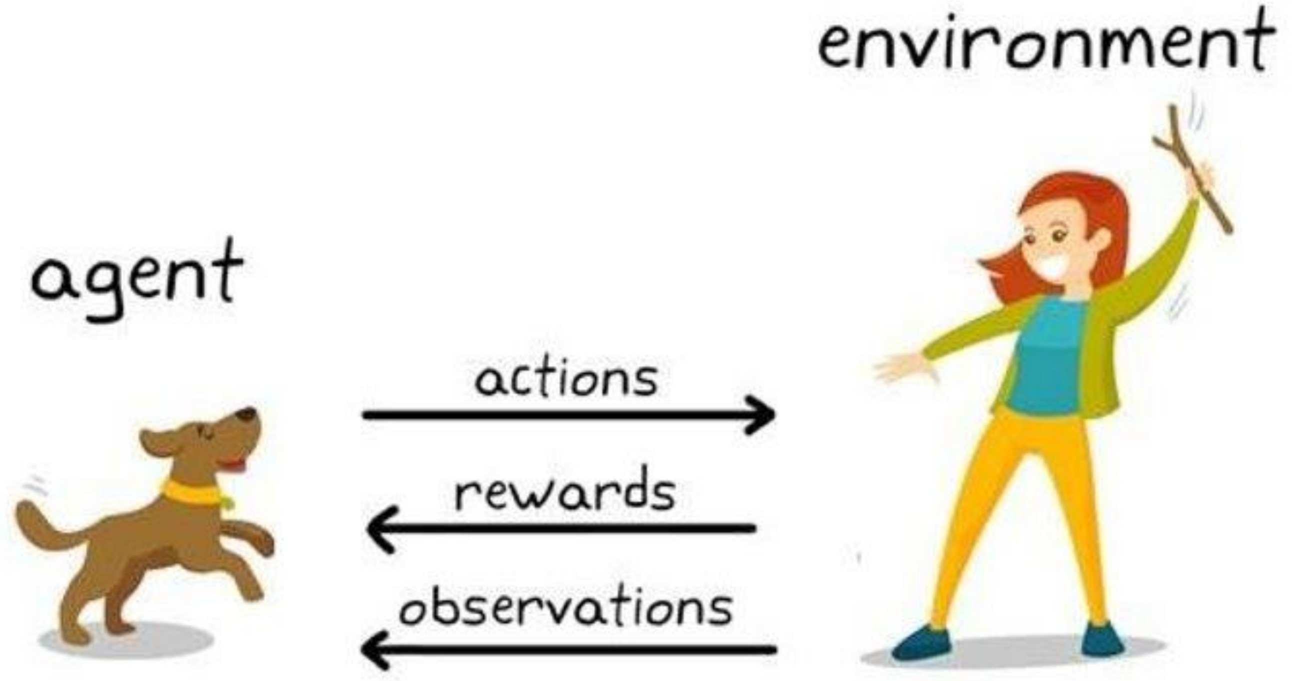
Machine learning problem where some of the data are labelled and some of the data are NOT labelled



Semi-supervised self-training method

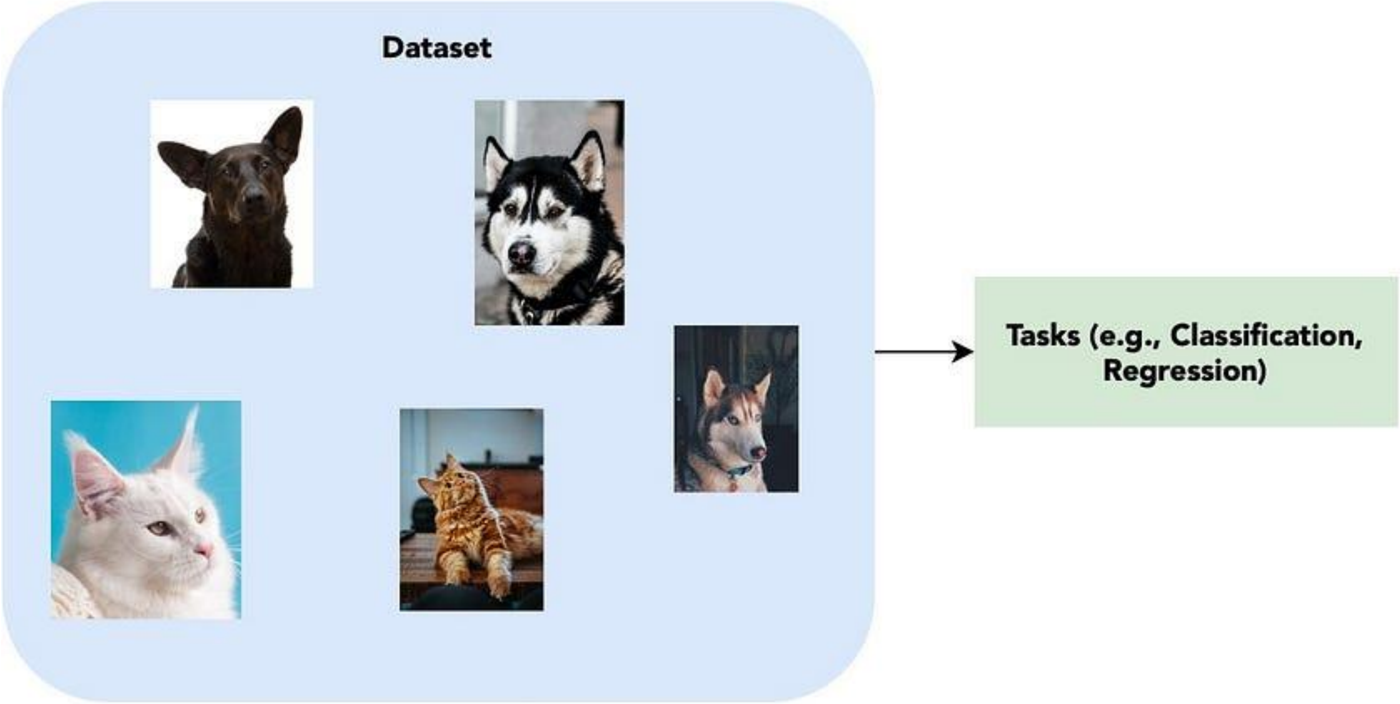
SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

Machine learning problem focused on interaction between agent and environment



Domitilla Brandoni, CINECA

No labels, supervised tasks



https://towardsdatascience.com/supervised-semi-supervised-unsupervised-and-self-supervised-learning-7fa79aa9247c

REGRESSION ERRORS

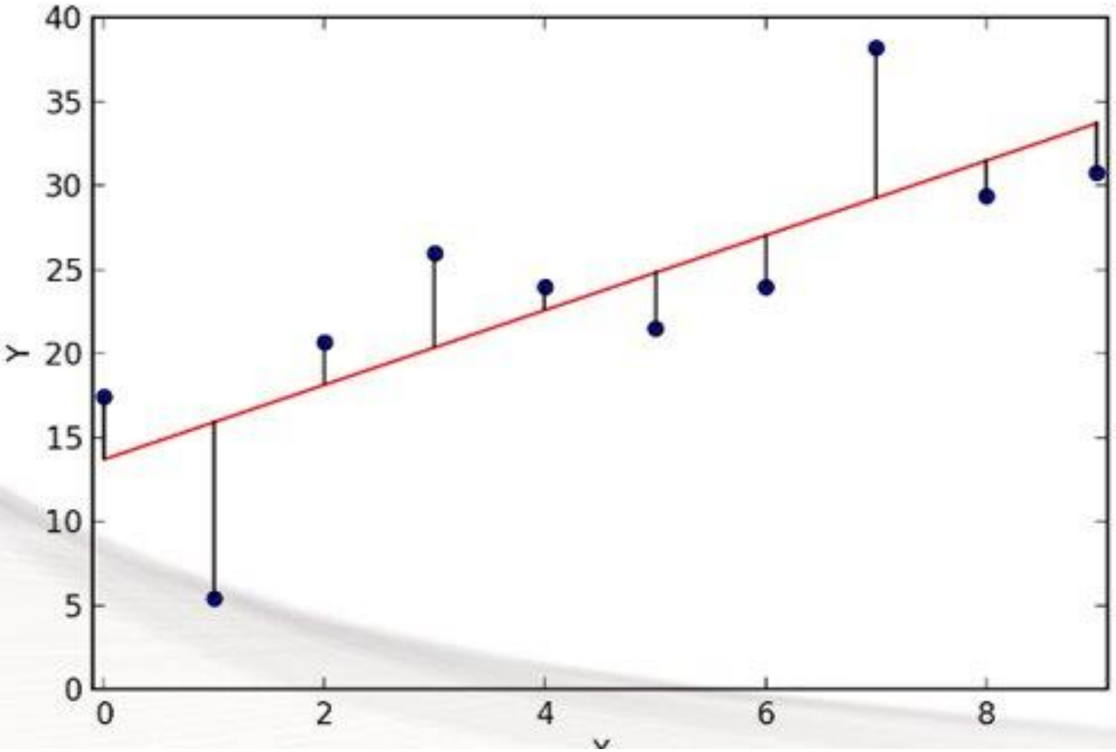
- Mean Squared Error
- Mean Average Error
- R-squared

CLASSIFICATION ERRORS

- Precision
- Accuracy
- Recall
- F1

MSE

Domitilla Brandoni, CINECA



$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

TP,TN,FP,FN

Domitilla Brandoni, CINECA

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP),** hit	**False negative (FN),** type II error, miss, underestimation
	Negative (N)	**False positive (FP),** type I error, false alarm, overestimation	**True negative (TN),** correct rejection

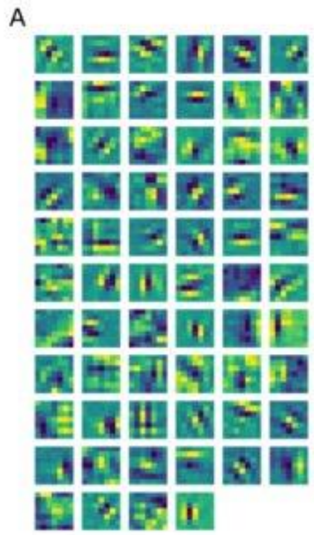
Wikipedia

Hands-on

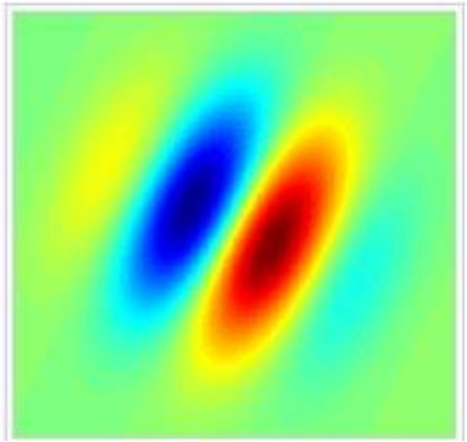
Domitilla Brandoni, CINECA

supervised_unsupervised.ipynb

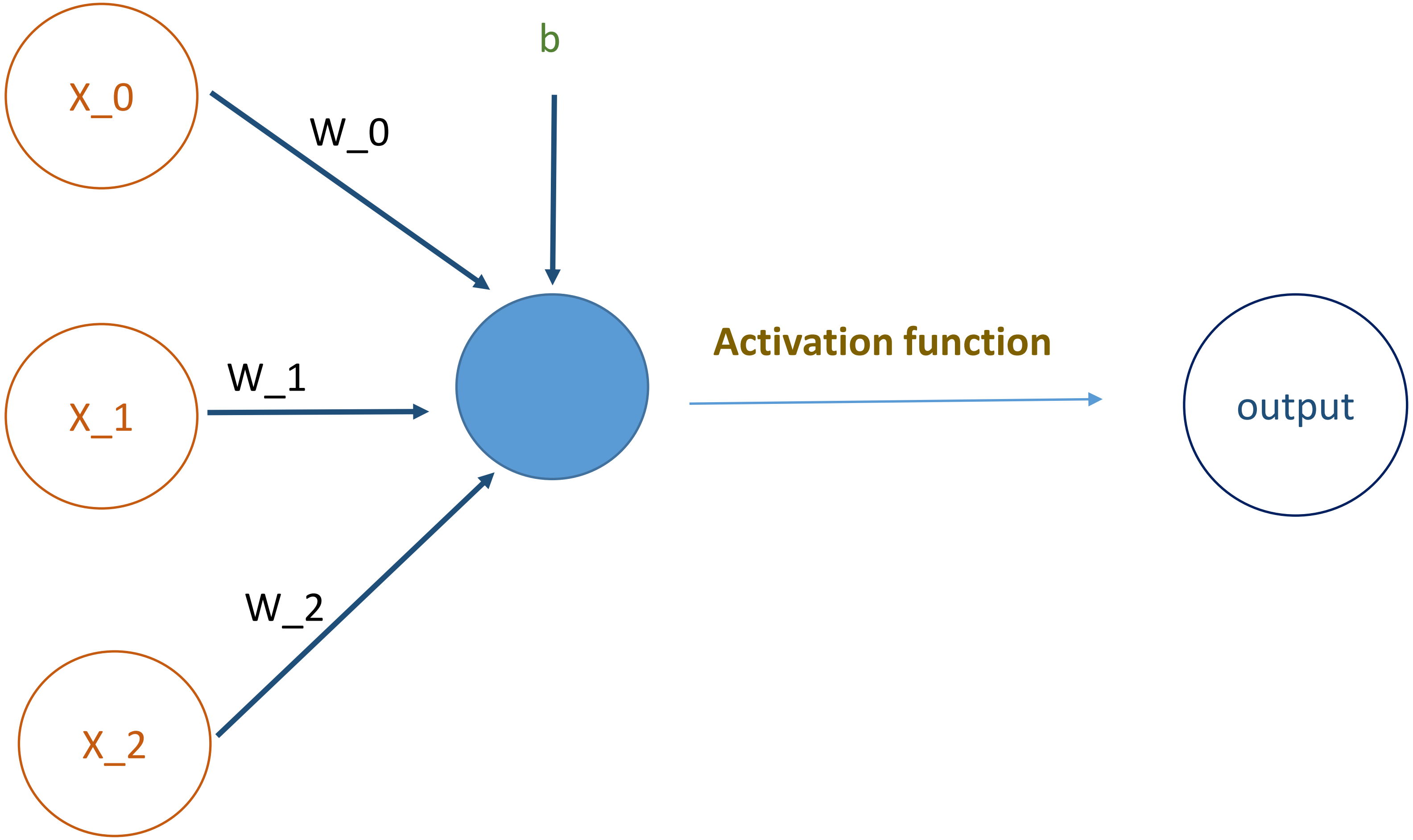
Try to mimic HNN in ANN by using simmetries specific to certain areas of the human brain (LGN, V1) within ANNs



*CNN filters, Federico Bertoni,Noemi Montobbio, Alessandro Sarti e Giovanna Citti**



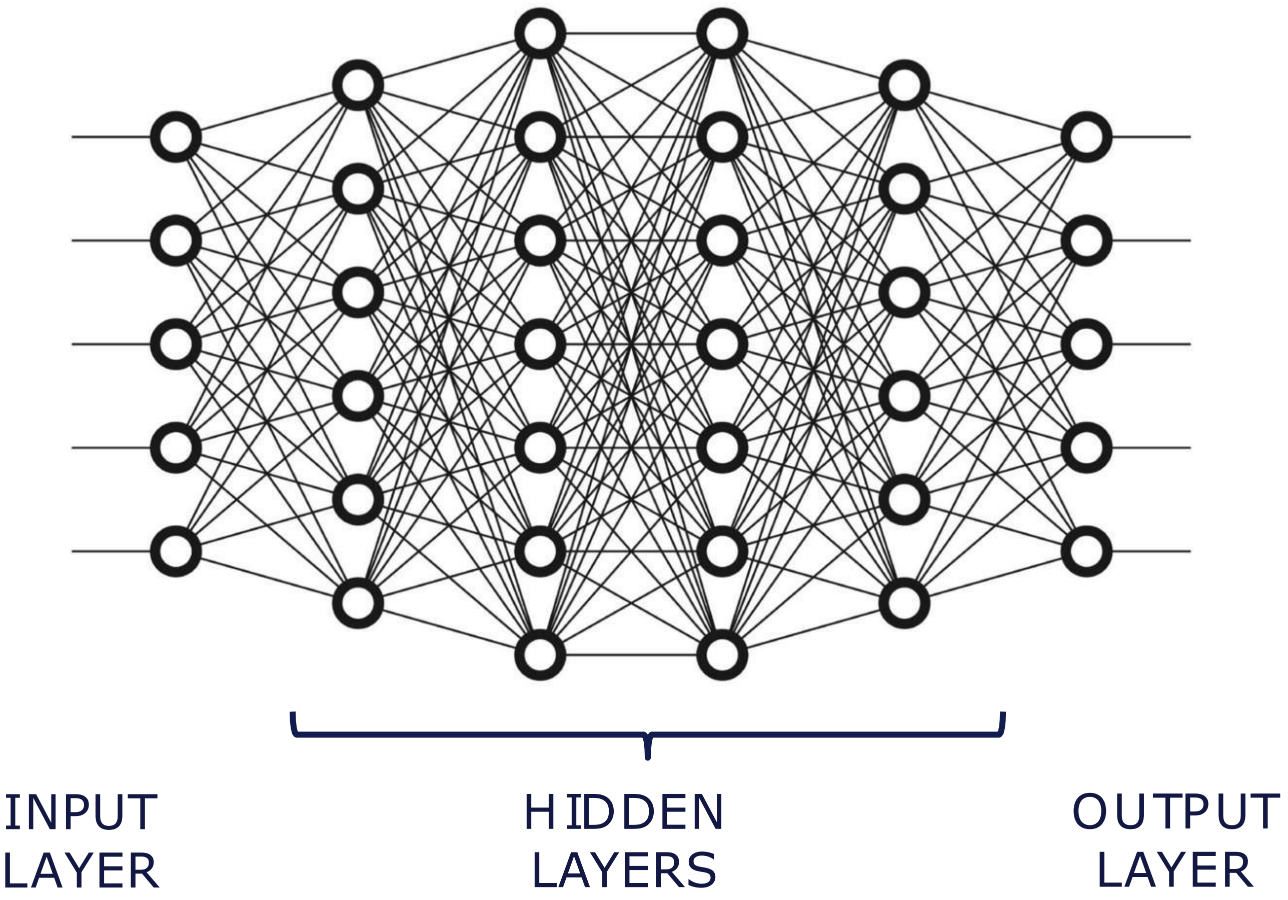
*Emergence of Lie symmetries in functional architectures learned by CNNs *Federico Bertoni,Noemi Montobbio, Alessandro Sarti e Giovanna Citti*



Inputs (x_i)
Weights (w_i)
Bias (b)

Architecture of a Neural Network

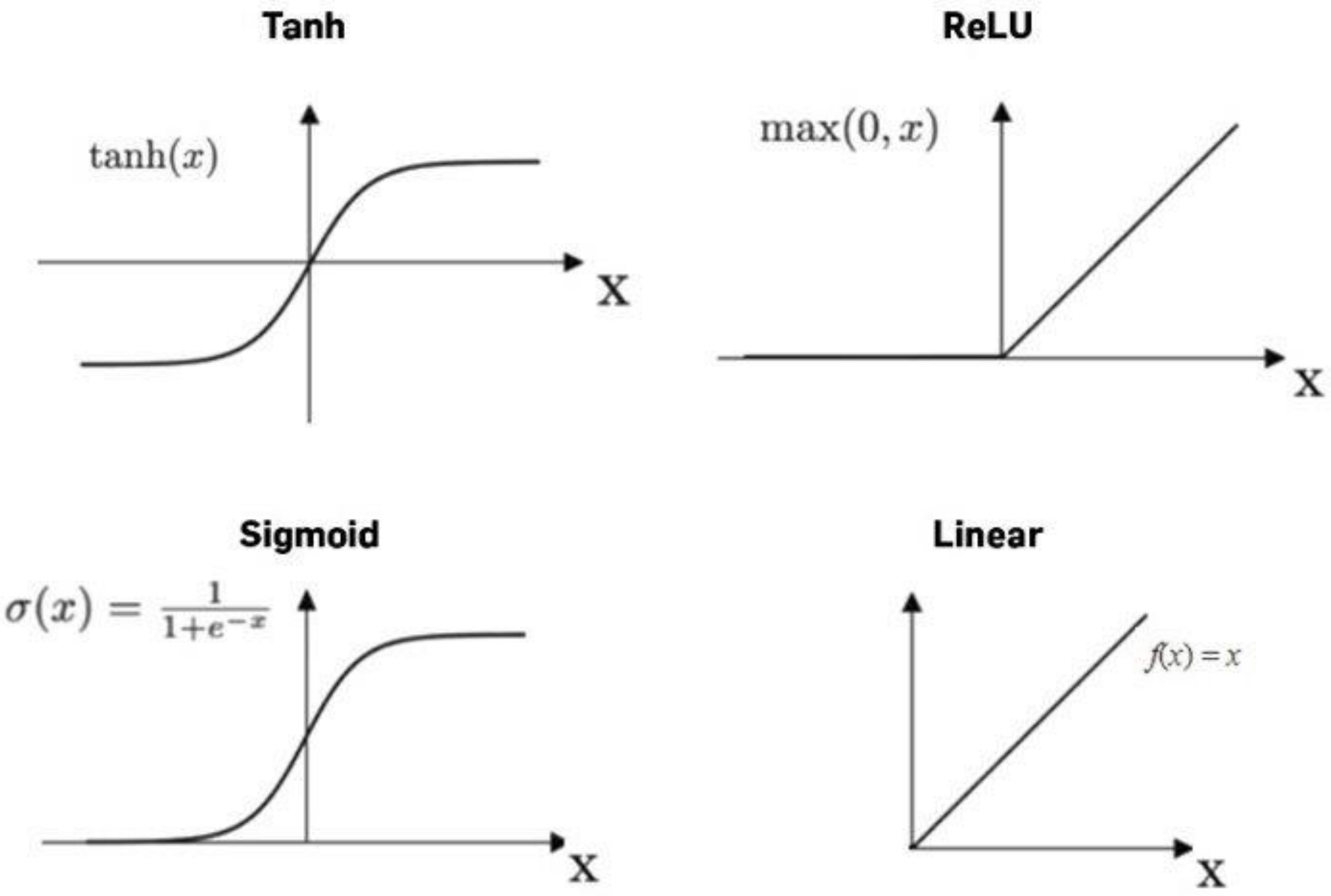
Domitilla Brandoni, CINECA



- Many layers (deep)
- Many nodes
- Activation functions
- input/output

Activation functions

Domitilla Brandoni, CINECA



No activation, no DL

Domitilla Brandoni, CINECA

Many layers **without** activation function
=
one layer with many weights

Many layers **with** activation function
=
Deep Learning

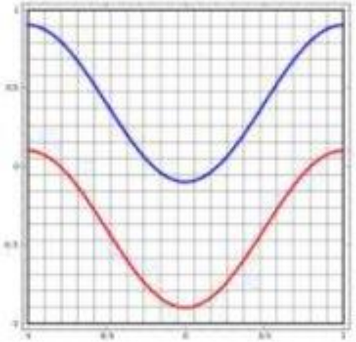
Why NN?

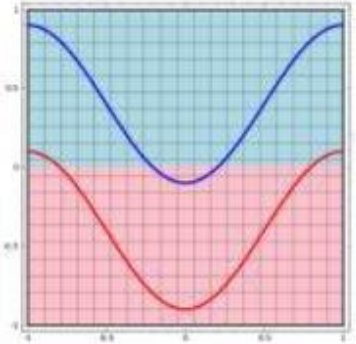
Domitilla Brandoni, CINECA

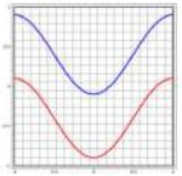
A neural network with a single hidden layer and a non linear activation function is a **“Universal Function Approximator”**

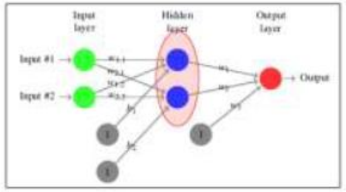
The universal approximation theorem states that a feed-forward network with a single hidden layer containing a finite number of neurons can approximate any continuous functions on a compact subsets of R (under mild assumptions on the activation function)

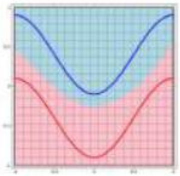
A feed-forward network is a powerful deep learning tool as a universal function approximator able to model any complex function

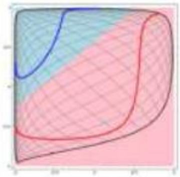


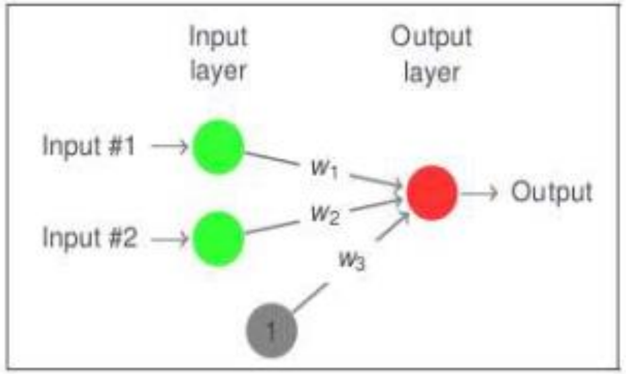












DL workflow

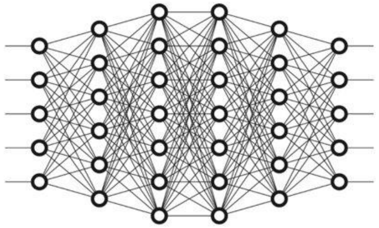
Domitilla Brandoni, CINECA

Model selection

Error function

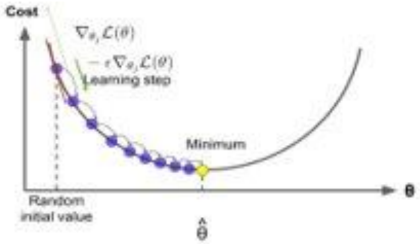
Optimization algorithm

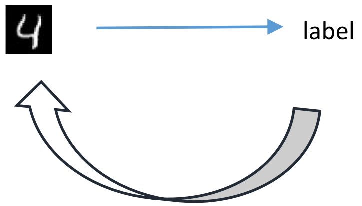
Iteration on data



$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

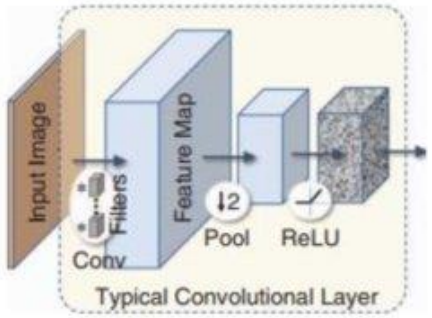
$$MAE = \frac{\sum_{i=1}^{n} | y_i - \hat{y}_i |}{n}$$

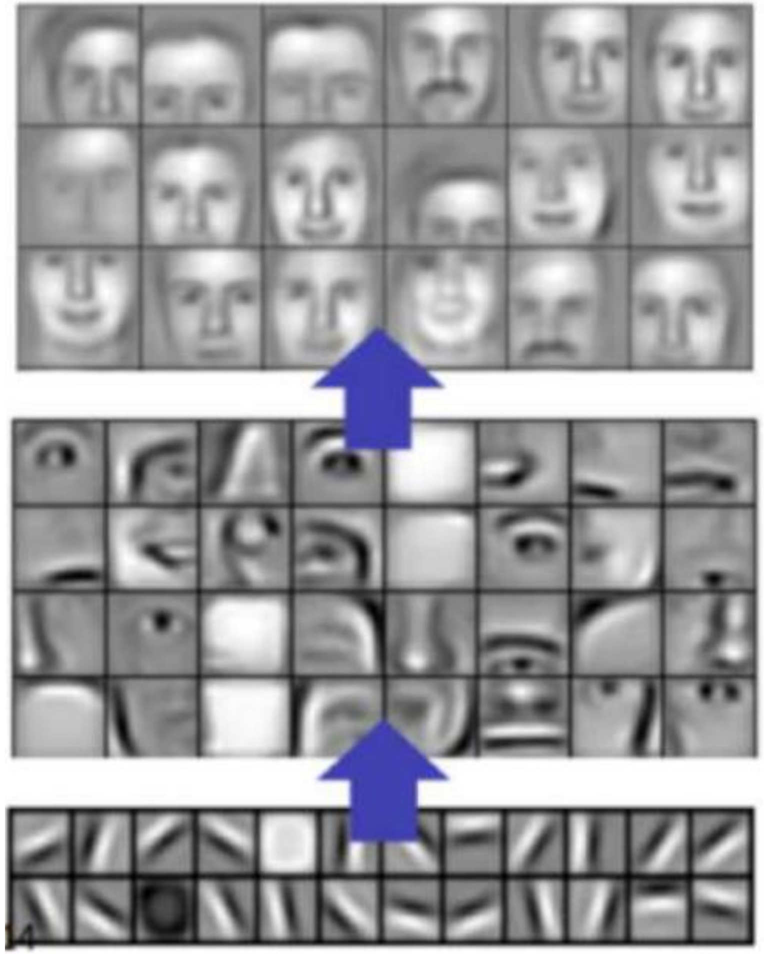




Domitilla Brandoni, CINECA

- **Convolution layer** : **convolutions of data** (an application of a spatial filters) which **extract specific features** (one per filter)
- **Pooling layer**: **downsamples the feature map** to introduce Translation invariance and reduce parameters (i.e. overfitting)
- **Last layer**: feature identifier/**classifier**





Layer 3

Layer 2

Layer 1

Convolutional Neural Networks

Domitilla Brandoni, CINECA

$$\begin{bmatrix} 60 & 7 & 98 & 14 & 19 \\ 165 & 159 & 147 & 196 & 169 \\ 187 & 204 & 165 & 41 & 111 \\ 209 & 30 & 201 & 23 & 203 \\ 58 & 79 & 218 & 59 & 118 \end{bmatrix} * \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 0 & -1 \end{bmatrix} =$$

Credits: towardsdatascience

GAN

Domitilla Brandoni, CINECA

Le Generative Adversarial Network (GAN) sono costituite da due reti neurali: generatore e discriminatore

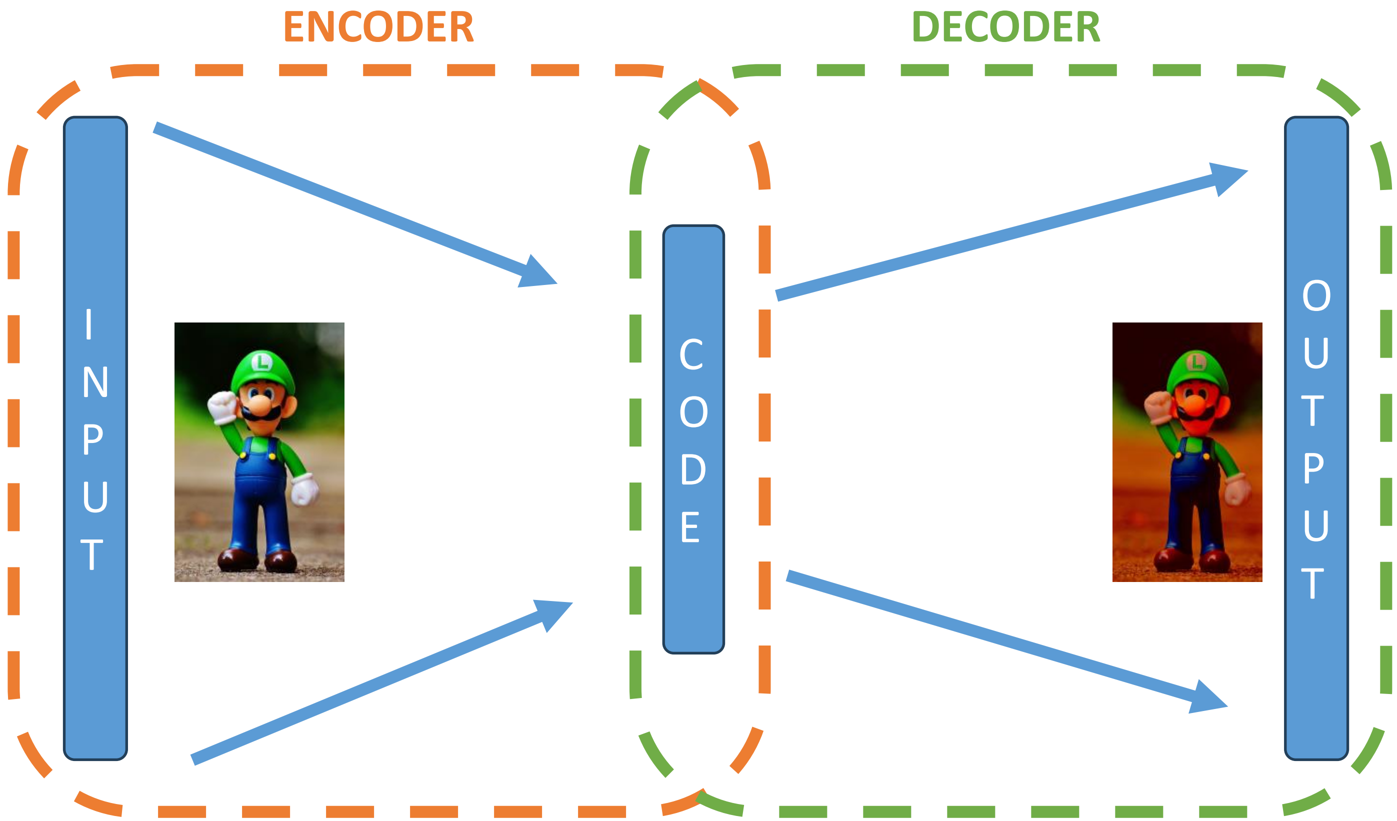
GENERATOR

Learns to generate data similar to training data. The generated instances become negative training examples for the discriminator

DISCRIMINATOR

Learns to distinguish the real data (training data) from the fake data (generator's data)

GAN learns by minimizing/maximizing a specific objective function
Optima: **generator mimics well the input and the discriminator outputs 0.5 deterministically on all inputs**



Minimize the difference between the input and the reconstruction

$$\phi : \mathcal{X} \rightarrow \mathcal{F}$$
$$\psi : \mathcal{F} \rightarrow \mathcal{X}$$
$$\phi, \psi = \underset{\phi,\psi}{\arg\min} \|\mathcal{X} - (\psi \circ \phi)\mathcal{X}\|^2$$

DL workflow

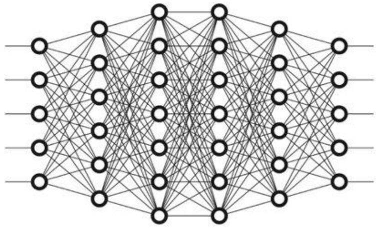
Domitilla Brandoni, CINECA

Model selection

Error function

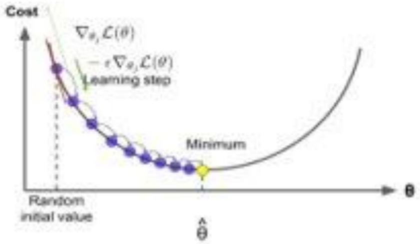
Optimization algorithm

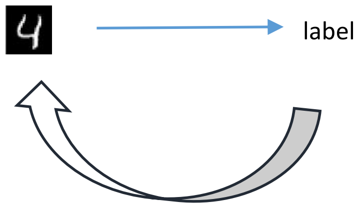
Iteration on data



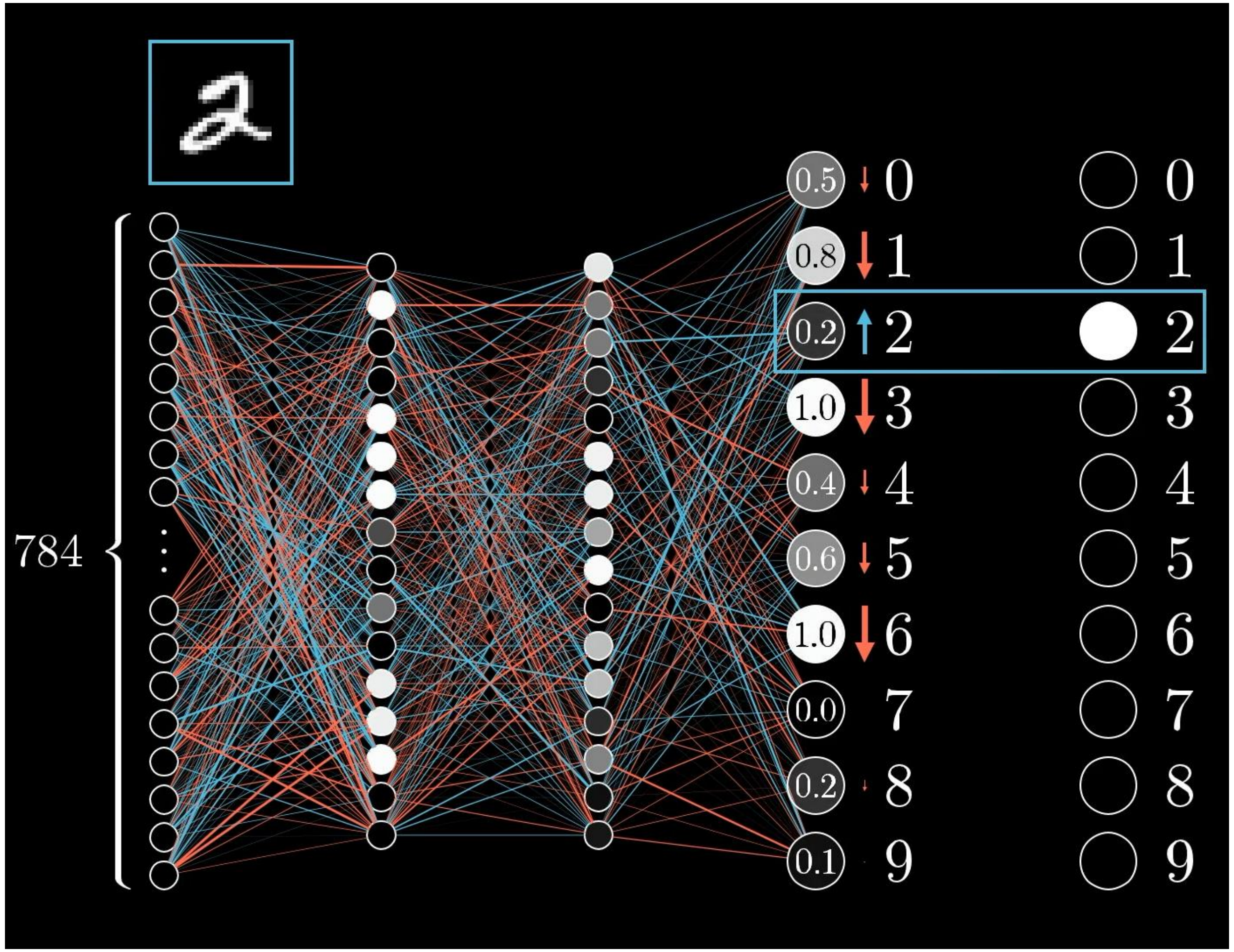
$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$$

$$MAE = \frac{\sum_{i=1}^{n} \mid y_i - \hat{y}_i \mid}{n}$$





Domitilla Brandoni, CINECA



Cost function

How much does a nudge to $w^{(L)}$ change $z^{(L)}$?

How much does that nudge to $z^{(L)}$ change $a^{(L)}$?

$$\frac{\partial C_0}{\partial w^{(L)}} = \boxed{\frac{\partial z^{(L)}}{\partial w^{(L)}}} \boxed{\frac{\partial a^{(L)}}{\partial z^{(L)}}} \boxed{\frac{\partial C_0}{\partial a^{(L)}}}$$

How much does *that* nudge to $a^{(L)}$ change C_0?

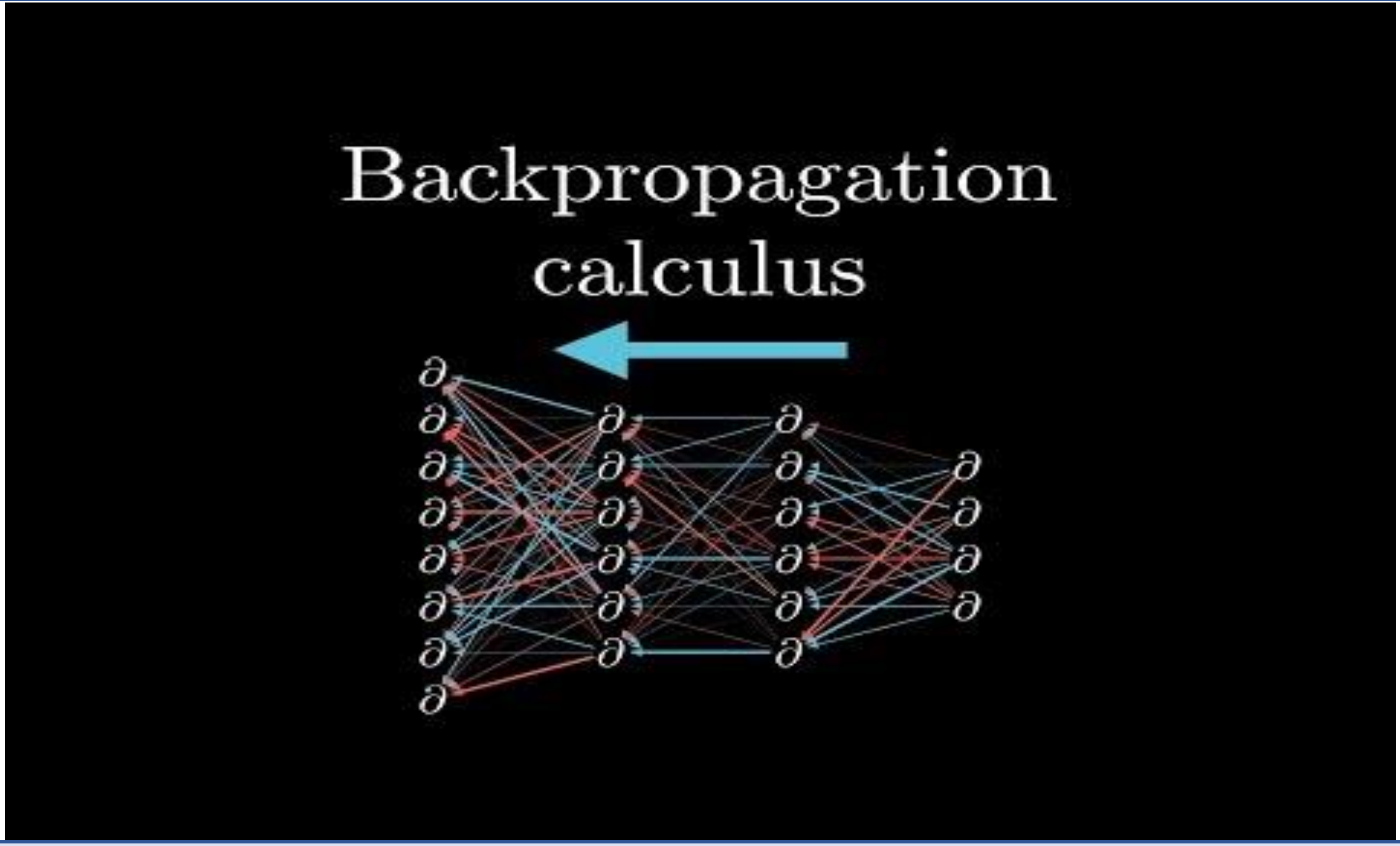
https://www.3blue1brown.com/lessons/backpropagation-calculus

Credits: 3blue1brown

Backpropagation

Domitilla Brandoni, CINECA

SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP



Credits: 3blue1brown

GD vs SGD? Batch GD

Domitilla Brandoni, CINECA

GD

Gradient descent on all data

Slow
Minimizing on all data
Higher probability to approximate better the local minima

SGD

Gradient descent on randomly selected data

Faster
Minimizing on some data
Lower probability to approximate the local minima

BATCH GD

Gradient on a batch of data

batch size: hyperparameter of the model
(High bs ~ GD)

Epoch : when the ENTIRE dataset is evaluated once from the network (backward and forward)

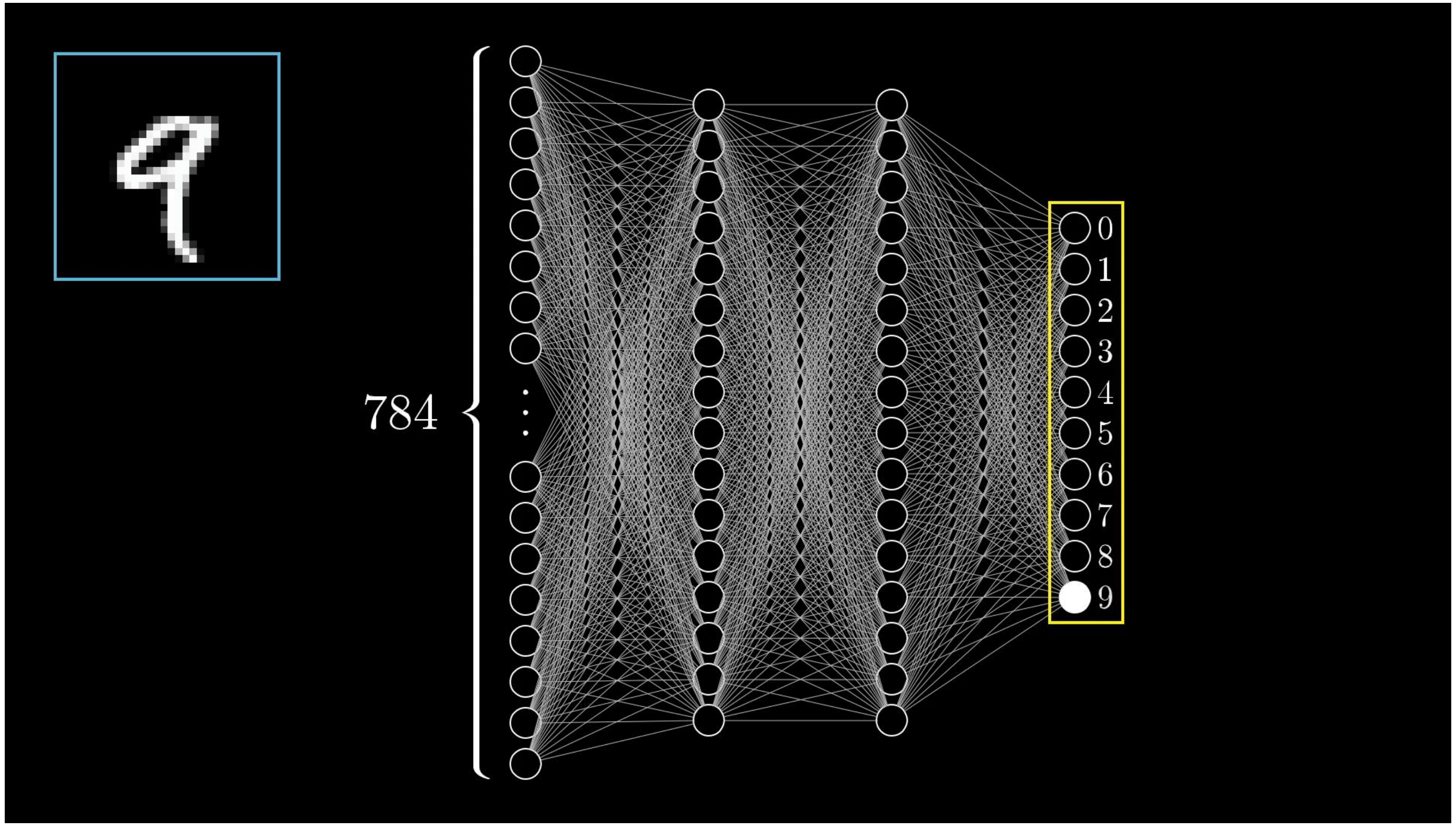
Steps per epochs: gradient descent steps per epoch (depends from the bs)

```
for i=0,….,epochs
    for batch in batches
        w = batch_GD_algorithm(w)
```

- Choose an initial vector of parameters w and learning rate η .
- Repeat until an approximate minimum is obtained:
  - For i = 1 , 2 , . . . , n , do:
    - **w := w − η ∇ C (w)** , where the gradient of the cost function is computed on the batch samples. The way in which the update is done is defined by the optimization algorithm

Inference

Domitilla Brandoni, CINECA



Credits: 3blue1brown

Hands-on

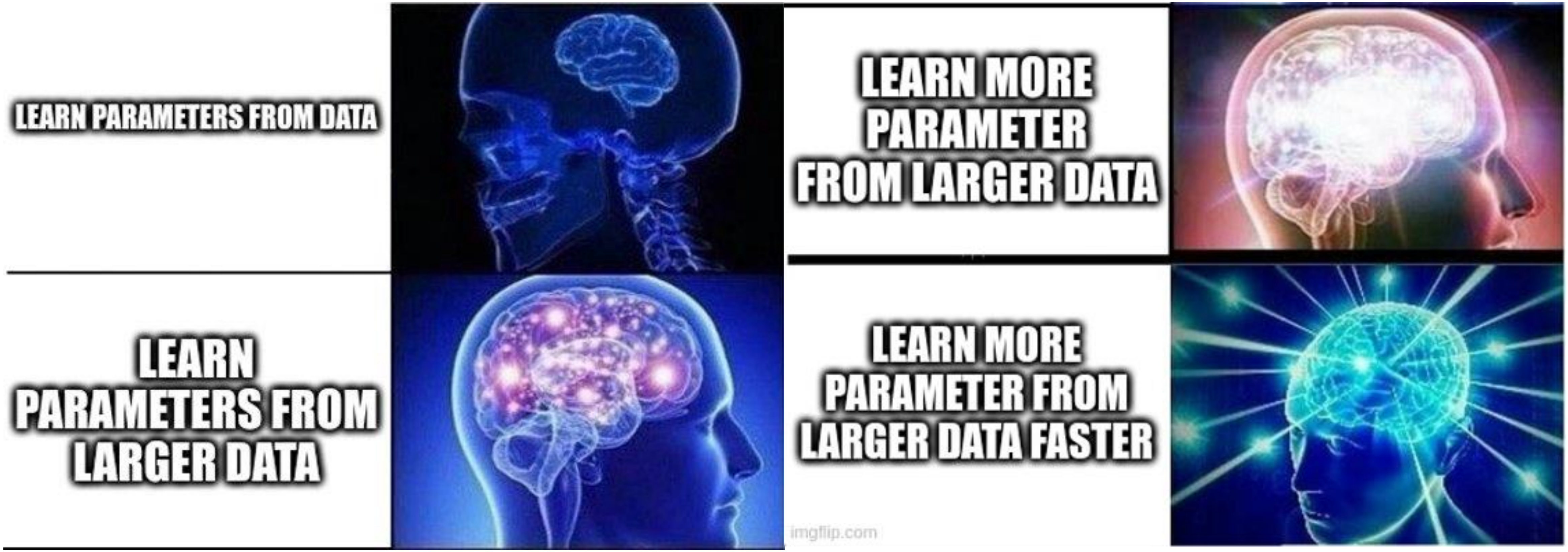
Domitilla Brandoni, CINECA

SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

NeuralNetworks.ipynb

Why should I care for HPC?

Domitilla Brandoni, CINECA



Why should I care for HPC?

Domitilla Brandoni, CINECA

SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

Training performance: NVIDIA DGX-1 32GB (8x V100 32GB)

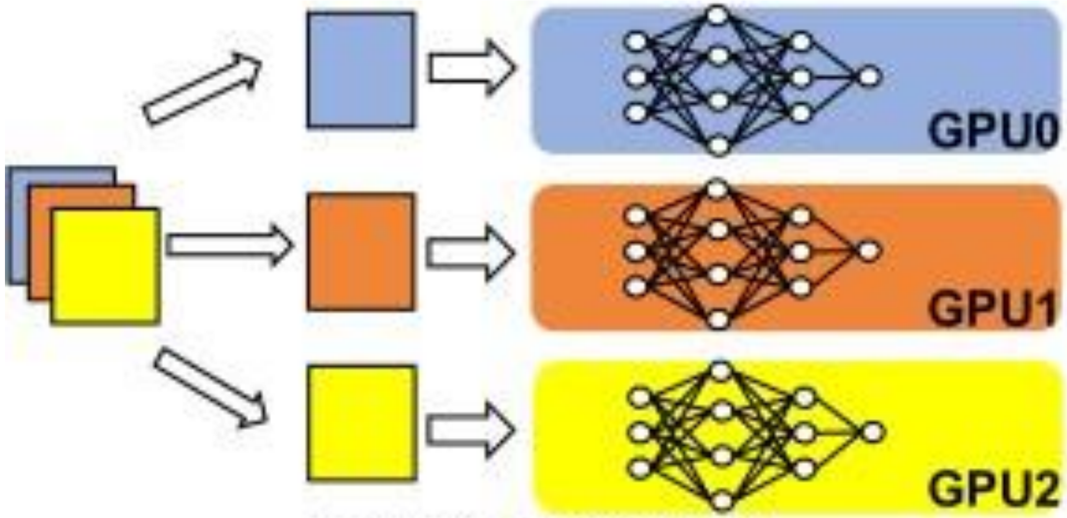
GPUs	Throughput - FP32	Throughput - mixed precision	Throughput speedup (FP32 to mixed precision)	FP32 Strong Scaling	Mixed Precision Strong Scaling	Mixed Precision Training Time (90E)	FP32 Training Time (90E)
1	356 img/s	1156 img/s	3.24 x	1.0 x	1.0 x	~30 hours	~95 hours
8	2766 img/s	8056 img/s	2.91 x	7.75 x	6.96 x	~5 hours	~13 hours

Training performance: NVIDIA DGX A100 (8x A100 80GB)

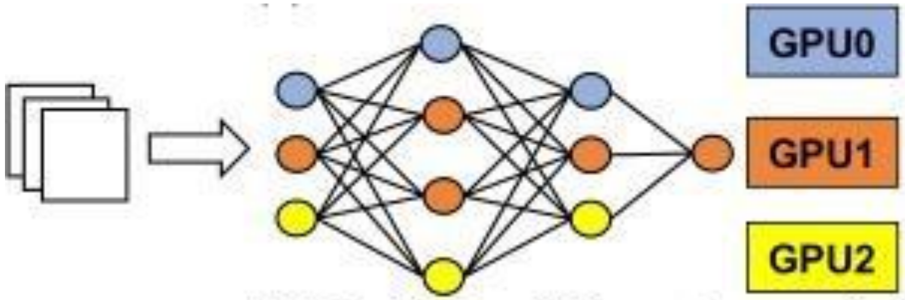
GPUs	Throughput - TF32	Throughput - mixed precision	Throughput speedup (TF32 to mixed precision)	TF32 Strong Scaling	Mixed Precision Strong Scaling	Mixed Precision Training Time (90E)	TF32 Training Time (90E)
1	938 img/s	2470 img/s	2.63 x	1.0 x	1.0 x	~14 hours	~36 hours
8	7248 img/s	16621 img/s	2.29 x	7.72 x	6.72 x	~3 hours	~5 hours

Parallel DL

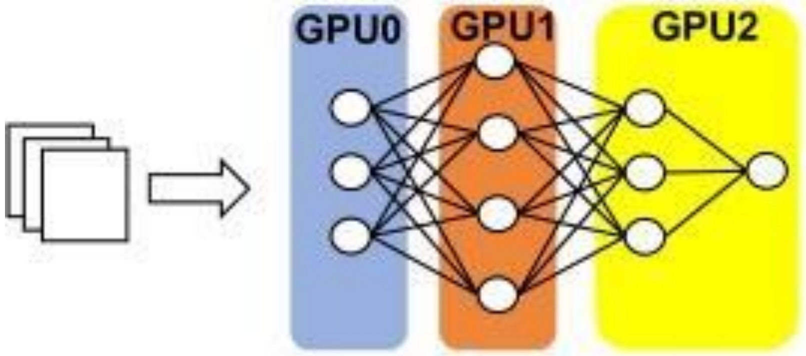
Domitilla Brandoni, CINECA



(a) Data Parallelism



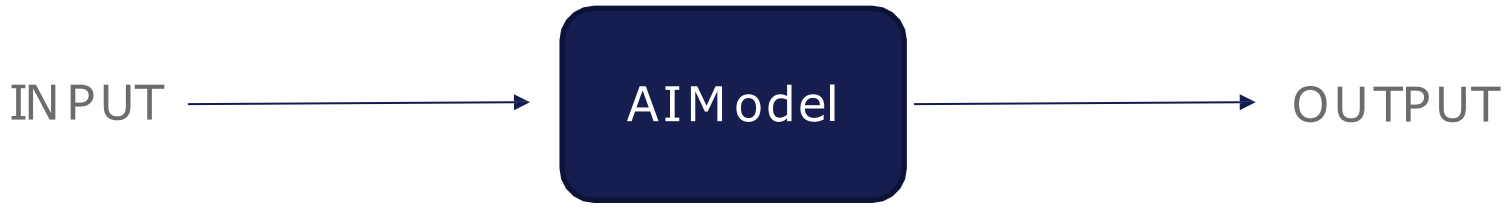
(b) Model Parallelism—tensor slicing



(b) Model Parallelism—layer-wise

https://doi.org/10.1145/3442442.3452055

Domitilla Brandoni, CINECA



AI models are usually considered as a sort of black box. Thus, it is not easy to understand how they decide -> lack of trust

Exercise

Domitilla Brandoni, CINECA

SCtrain | SUPERCOMPUTING KNOWLEDGE PARTNERSHIP

Same procedure of the hands-on session but with another dataset

- Download the dataset
- Look at the variables
- Create a neural network
- Train the model
- Test the model

Thank you for your attention!

http://sctrain.eu/

Univerza v Ljubljani

TECHNISCHE UNIVERSITÄT WIEN

CINECA

VSB TECHNICAL UNIVERSITY OF OSTRAVA | IT4INNOVATIONS NATIONAL SUPERCOMPUTING CENTER

Co-funded by the Erasmus+ Programme of the European Union

This project has been funded with support from the European Commission.
This publication [communication] reflects the views only of the author, and the Commission cannot be held responsible for any use which may be made of the information contained therein.