Introduction to Machine Learning

Fabien Baradel
PhD Student - INSA Lyon
fabienbaradel.github.io

What is your definition of Machine Learning ?

Definition: Machine Learning

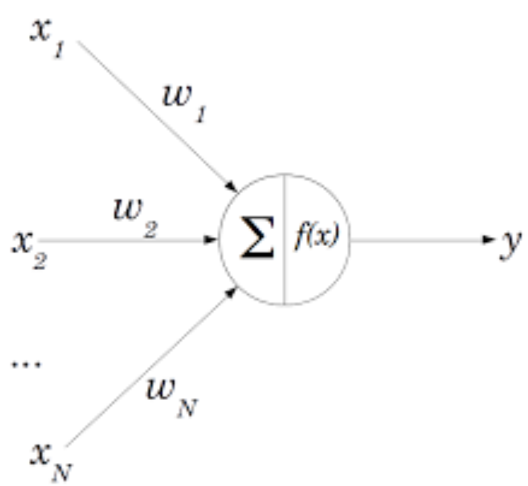
Statictics?
Maths?
Computer Science?
Big Data?
Artificial Intelligence?

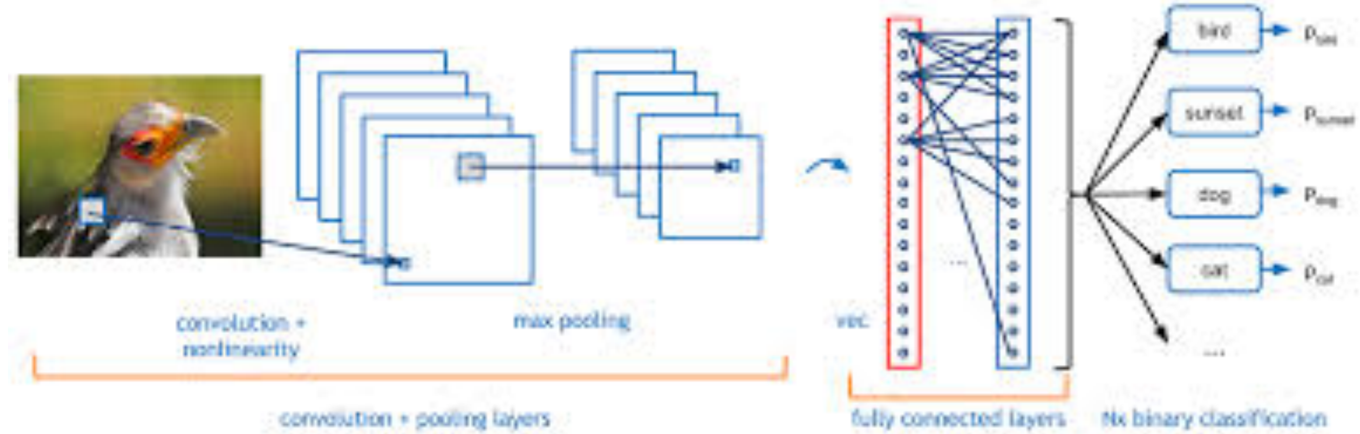
« Field of study that gives computers the ability to learn without being explicitly programmed »
Arthur Samuel, 1959

Automatic discover of patterns in data by a computer

Different from rule-based methods

Brief history of Machine Learning and AI





1950: Turing test

1952: 1st computer learning program by Arthur Samuel

1957: Perceptron by Frank Rosenblatt

1957: Nearest neighbor algorithm

1981: Gerald Dejong introduces concept of Explanation Based Learning

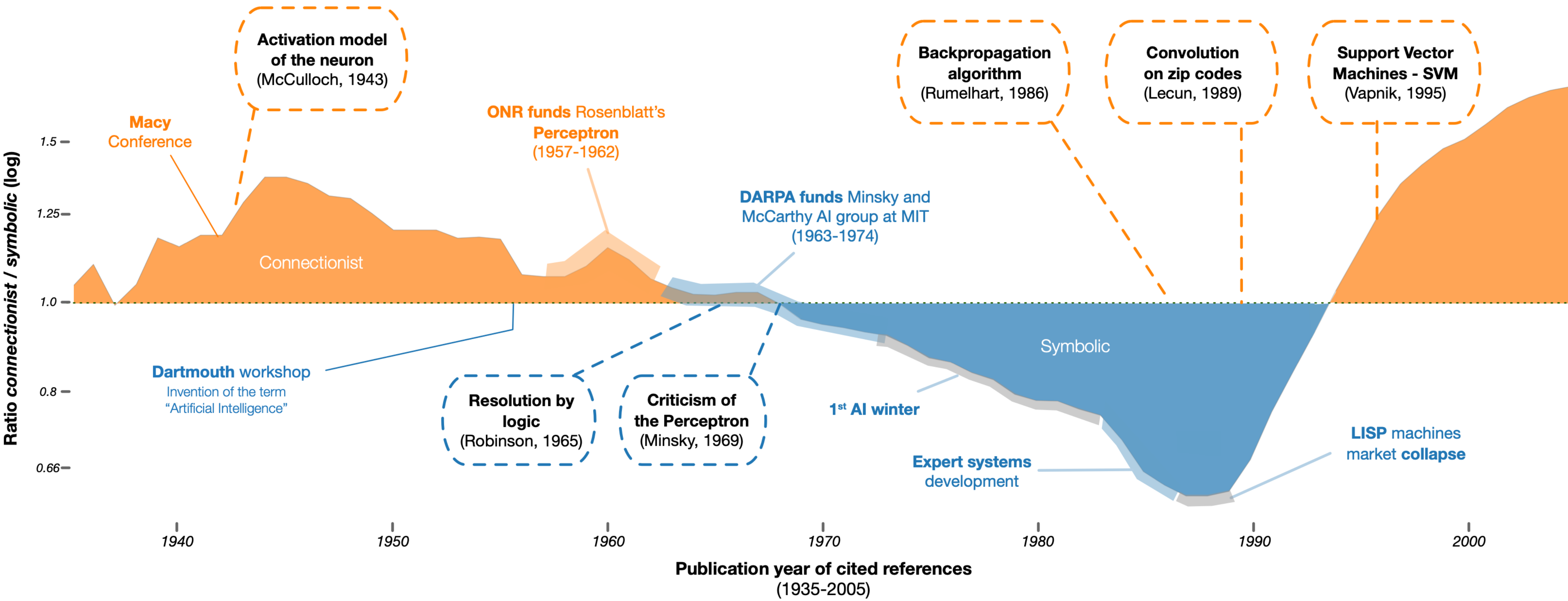
1990: ML shifts from knowledge-driven to data driven

1997: IBM Deep Blue

2006: Geoffrey Hinton coins the term « deep learning »

2012: AlexNet won Imagenet with CNN

Timeline AI



https://neurovenge.antonomase.fr

What do we want to learn?

Supervised Learning

Object classification

$\xrightarrow{f(\,.\,)}$ cat

Human Pose estimation

Unsupervised Learning

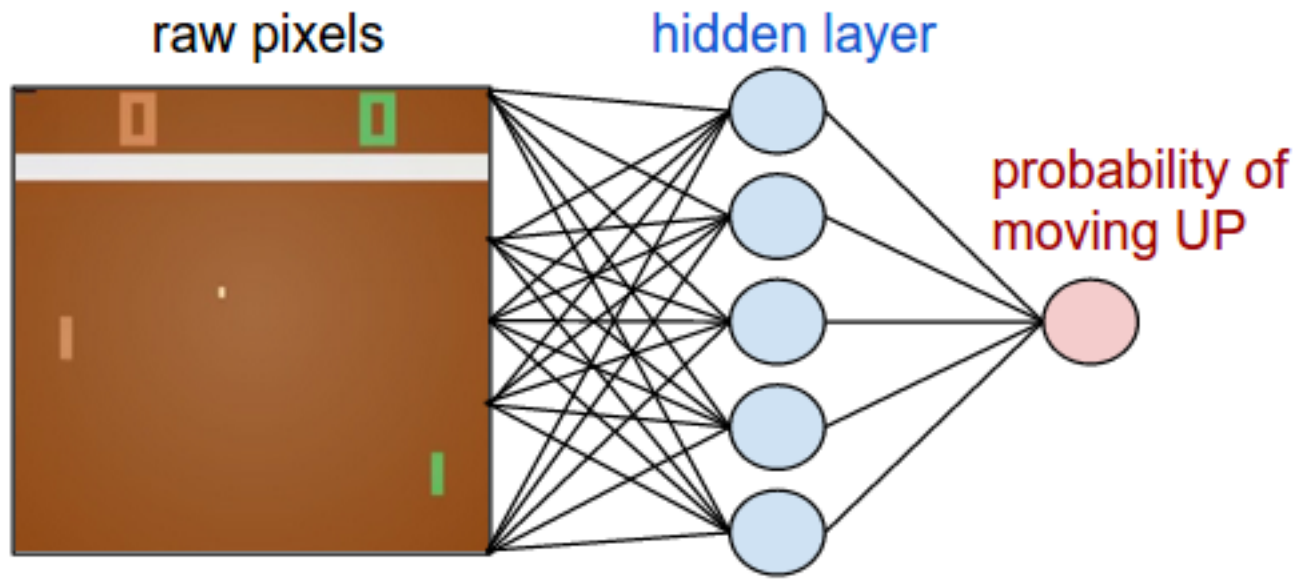
Image Generation

Future forecasting

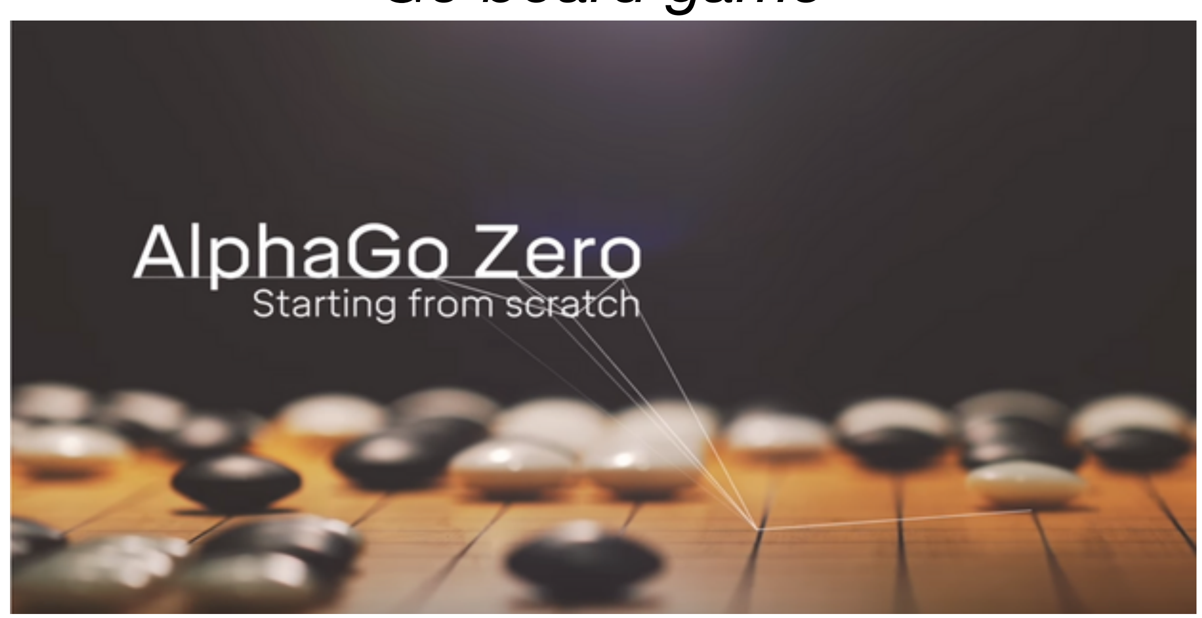
What do we want to learn?

Reinforcement Learning

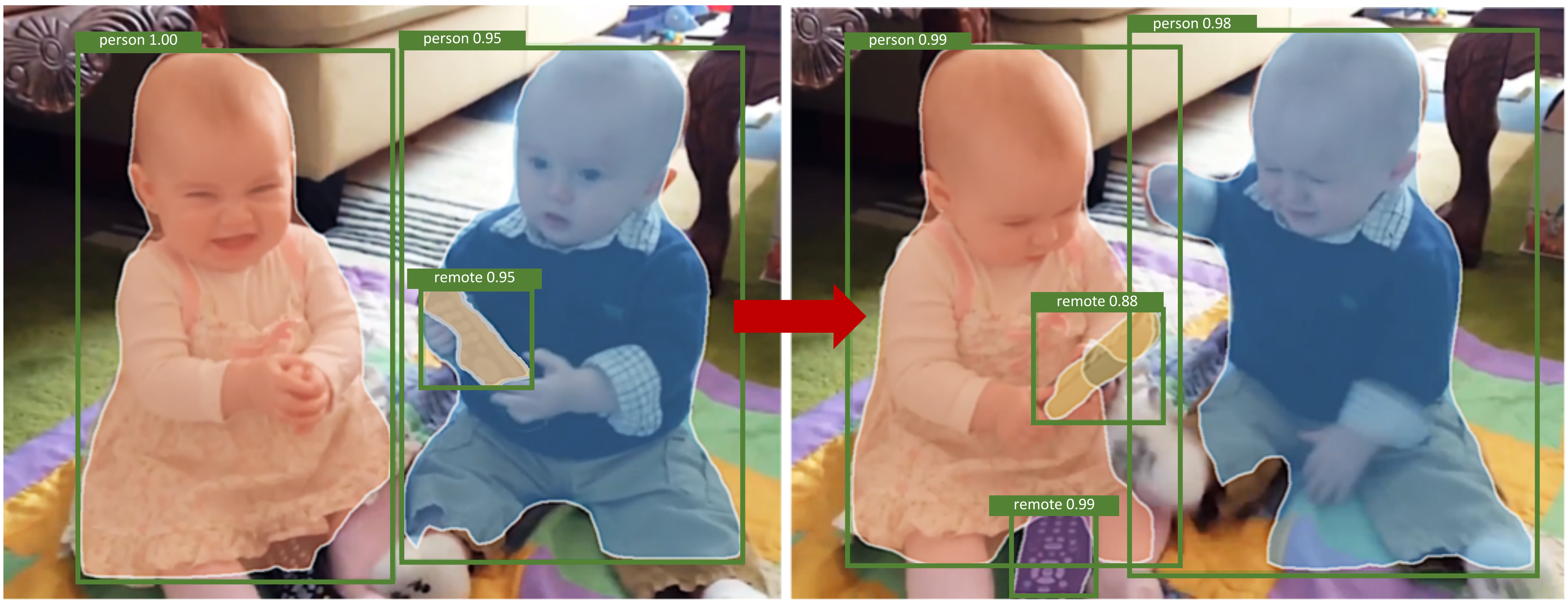
Superhuman performance in video game



Go board game



Video Understanding



Causal Reasoning
Spatio-Temporal Interactions

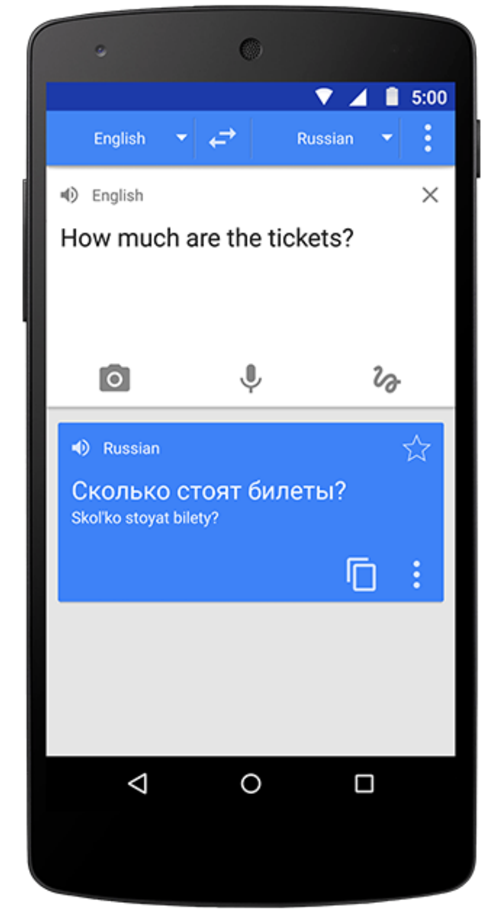
Strong AI

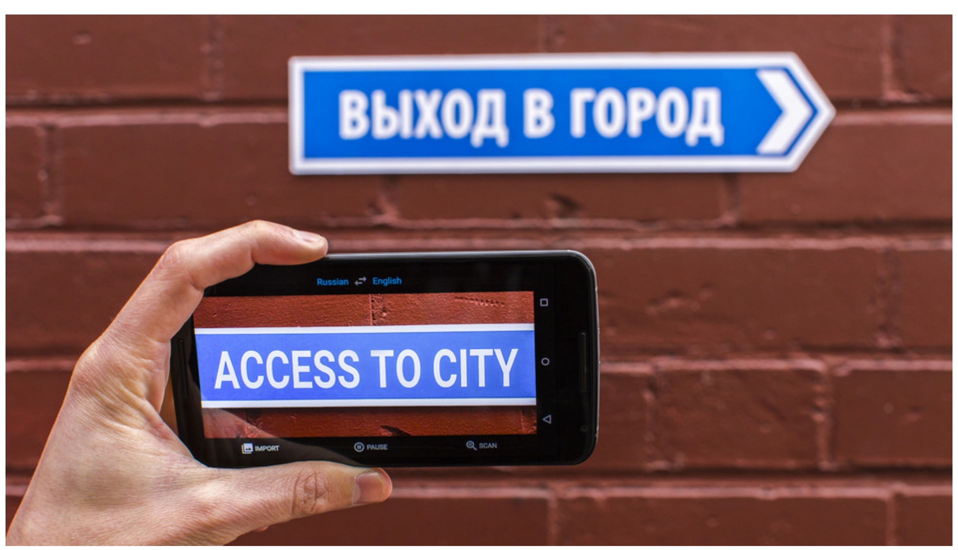
Specific task
Memorization
Shift between train and test
Adaptation to new task

Real World Applications

Tesla

Google





NETFLIX

Criteo

Industry

Huge investment
R&D Centers
USA - Canada - China - Europe (France!)

Software

ANACONDA

Python Data Science Platform
Conda - environnement

NVIDIA

SciPy

NumPy

matplotlib

IP[y]: IPython
Interactive Computing

pandas

scikit learn

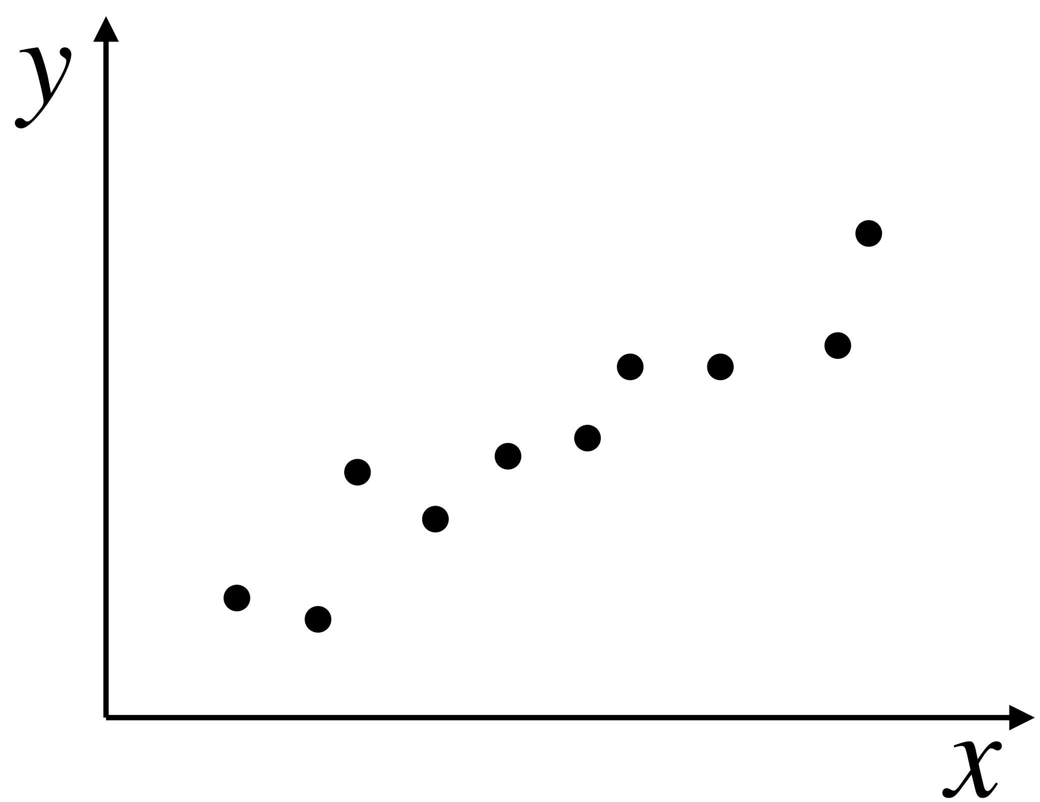
Supervised Learning

Regression

Problem Statement

$$D = \{(x_i, y_i)\}_{i=1}^{N}$$

$$y = f(x)$$



Solution

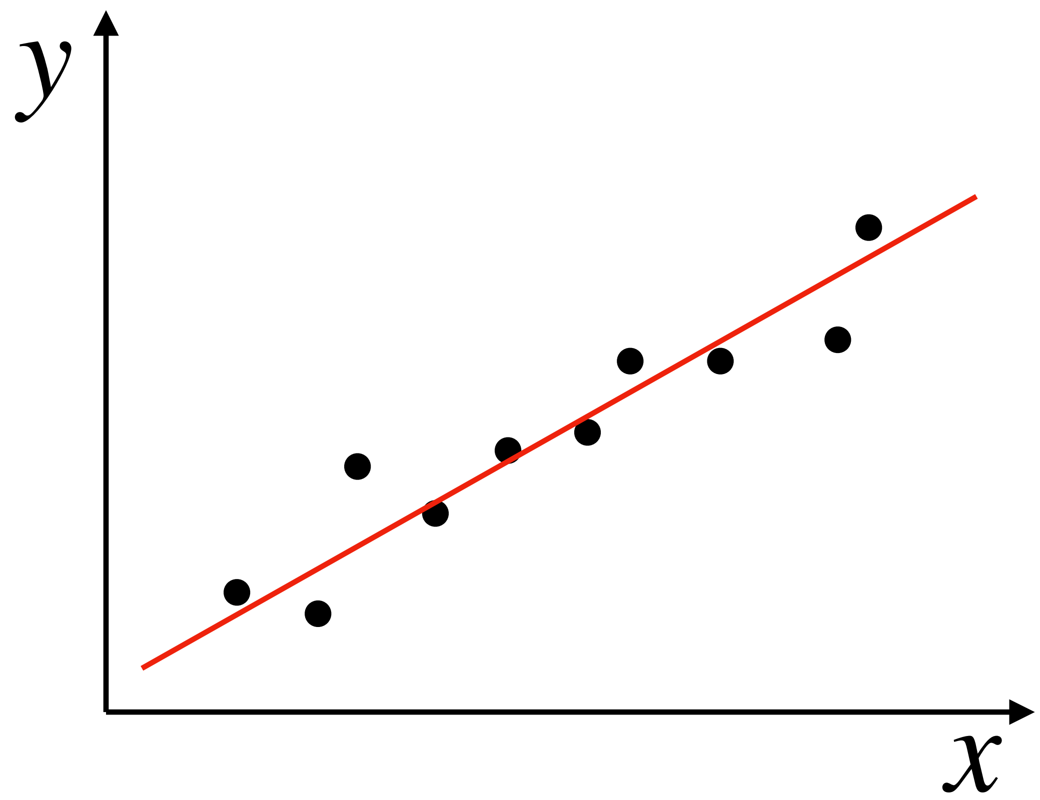
Model

$$f_{w,b}(x) = wx + b$$

Prediction

$$y = f_{w,b}(x)$$

Loss function $(f_{w,b}(x_i) - y_i)^2$



Solution

Model

$$f_{w,b}(x) = wx + b$$

Prediction

$$y = f_{w,b}(x)$$

Loss function $(f_{w,b}(x_i) - y_i)^2$

Objective $$\min_{w,b} \frac{1}{N} \sum_{i=1}^{N} (f_{w,b}(x_i) - y_i)^2$$

Closed-form solution

We set $\beta = [b \quad w]$ $X = \begin{bmatrix} 1 & x_1 \\ \cdots & \cdots \\ 1 & x_N \end{bmatrix}$ $Y = \begin{bmatrix} y_1 \\ \cdots \\ y_N \end{bmatrix}$

Optimization $\min_{\beta} ||\beta X - y||^2$

Optimal solution $\beta* = \hat{\beta} = (X^T X)^{-1} X^T y$

Pros and Cons

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

We can add statistical hypothesis
Robust modelling

Model checking
Invertibility
Difficult to compute in some case

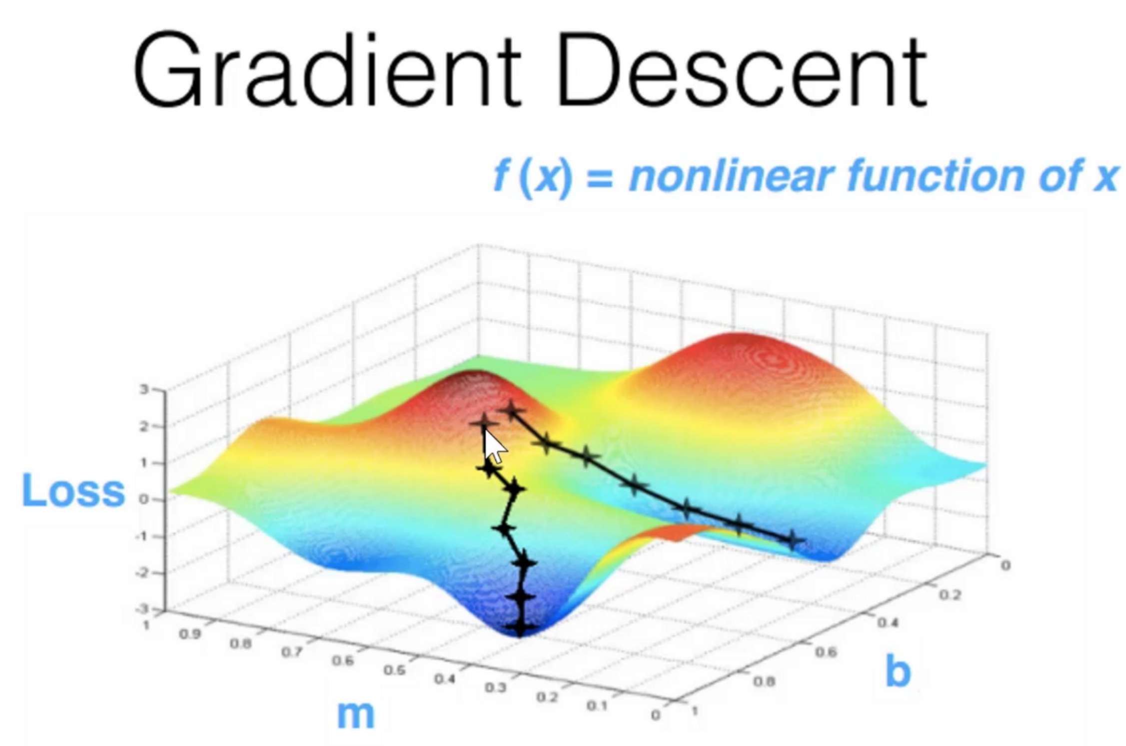
Linear regression with Gradient Descent

$$J(w,b) = \frac{1}{N}\sum_{i=1}^{N}(f_{w,b}(x_i) - y_i)^2$$

Optimization problem

Minimize a loss function using a certain model

Find the parameters which are minimizing the loss function



Linear regression with GD

Cost function

$$J(w, b) = \frac{1}{N} \sum_{i=1}^{N} (f_{w,b}(x_i) - y_i)^2$$

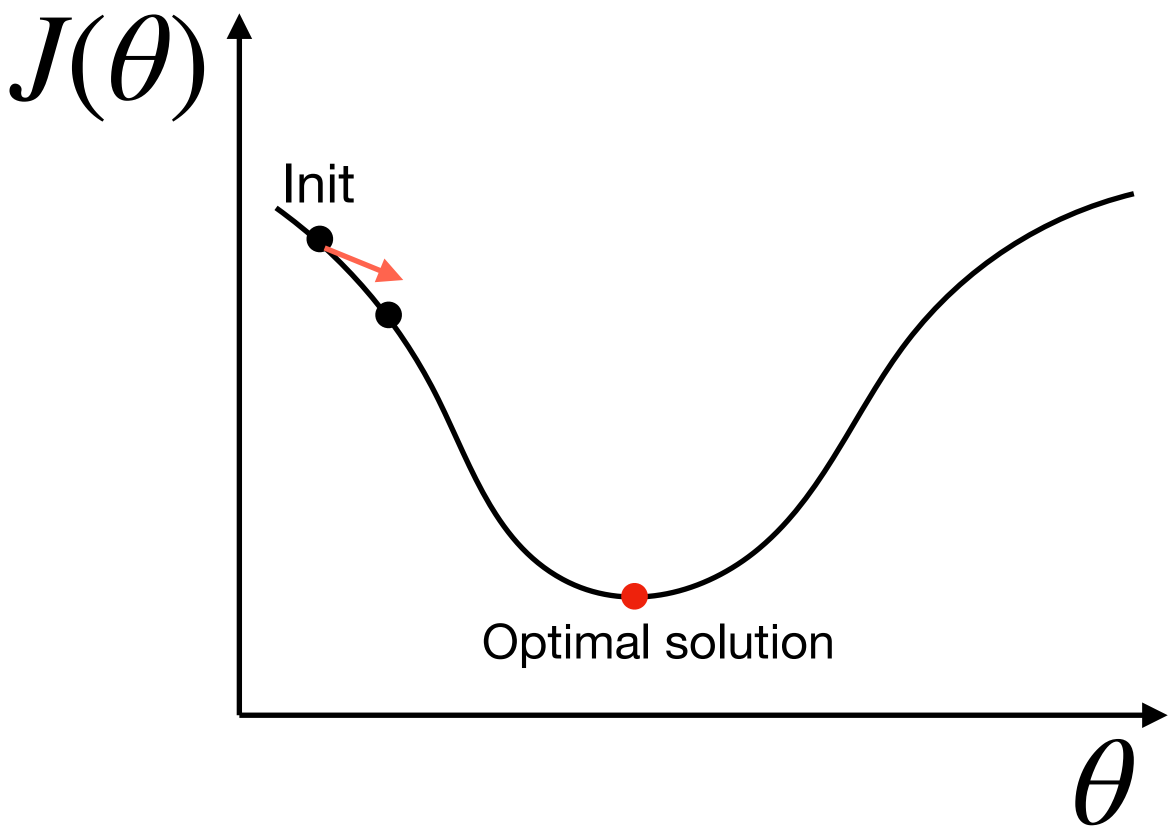
Compute derivatives

$$\frac{\partial J}{\partial w}(w, b) = \frac{1}{N} \sum_{i=1}^{N} - 2x_i(y_i - (wx_i + b))$$

$$\frac{\partial J}{\partial b}(w, b) = \frac{1}{N} \sum_{i=1}^{N} - 2(y_i - (wx_i + b))$$

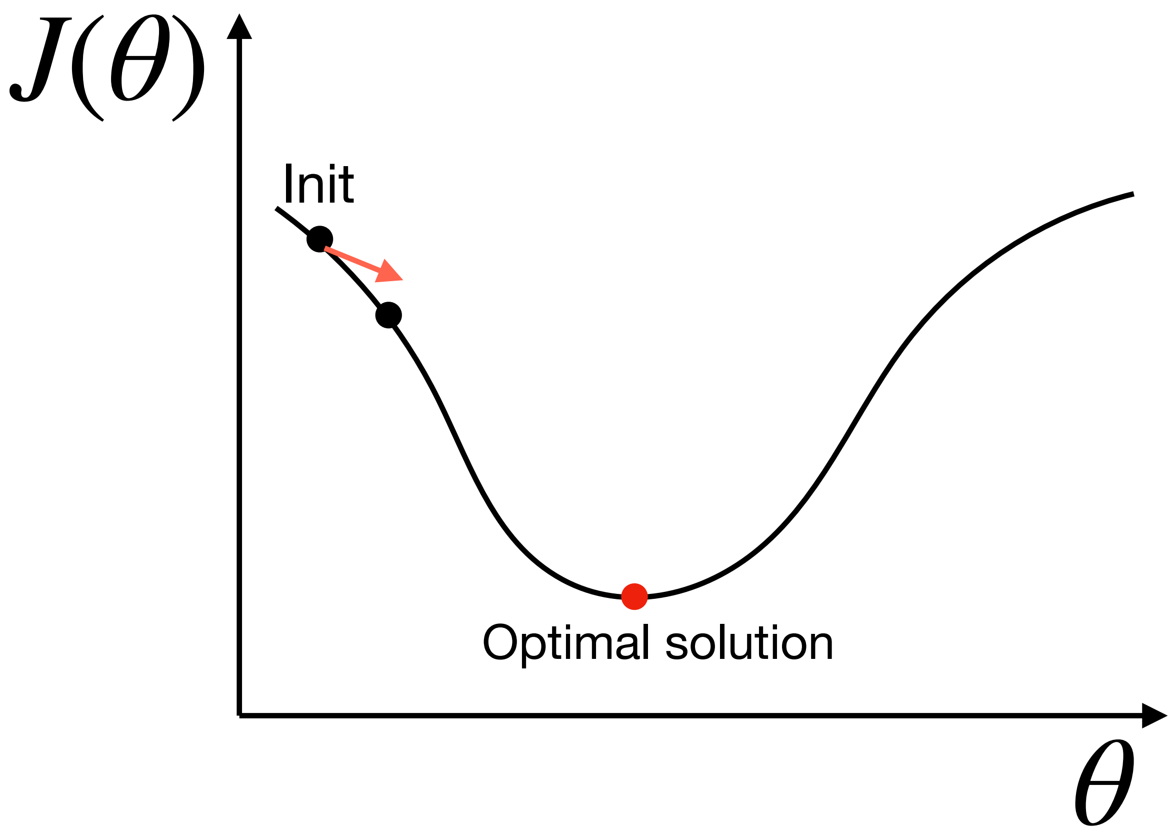
And update parameters iteratively

Gradient Descent



Goal: minimization of a function
Random initialization of parameters
At time t, gradient = slope fo the function
Iterative process
Updating parameters in the positive direction with a learning rate
Repeat until convergence

Gradient Descent



- Init $\hat{\theta}$ randomly
- Choose a learning rate η
- for t in range(nb_iter):
  - Update parameter

$$\hat{\theta} \leftarrow \hat{\theta} - \eta \frac{\partial J}{\partial \theta}(\hat{\theta})$$

Linear Regression with GD

$$J(w, b) = \frac{1}{N}\sum_{i=1}^{N}(f_{w,b}(x_i) - y_i)^2$$

- Init $\hat{w}, \hat{b}$ randomly
- Choose a learning rate η
- for t in $1 \ldots T$:
  - Update parameters

$$\hat{w} \leftarrow \hat{w} - \eta \frac{\partial J}{\partial w}(\hat{w}, \hat{b})$$

$$\hat{b} \leftarrow \hat{b} - \eta \frac{\partial J}{\partial b}(\hat{w}, \hat{b})$$

Linear Regression with Stochastic GD

$$J(w, b) = \frac{1}{N} \sum_{i=1}^{N} (f_{w,b}(x_i) - y_i)^2$$

- Init $\hat{w}, \hat{b}$ randomly
- Choose a learning rate η
- for t in $1 \ldots T$:
  - Sample some points from the data
  - Update parameters

$$\hat{w} \leftarrow \hat{w} - \eta \frac{\partial J}{\partial w}(\hat{w}, \hat{b})_D$$

Only on a subset D of the dataset

$$\hat{b} \leftarrow \hat{b} - \eta \frac{\partial J}{\partial b}(\hat{w}, \hat{b})_D$$

Exercise

Implementing Gradient Descent
LinearRegression: Closed-form vs GD vs SGD