UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Introduction to Learning

Winter School ROBOTICA PRINCIPIA

Frederic Precioso

24/01/2019

Disclaimer

If any content in this presentation is yours but is not correctly referenced or if it should be removed, please just let me know and I will correct it.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Overview

- **Context & Vocabulary**
  - ***What represents Artificial Intelligence?***
  - ***Machine Learning vs Data Mining?***
  - ***Machine Learning vs Data Science?***
  - ***Machine Learning vs Statistics?***
- Unsupervised classification
- Explicit supervised classification
- Implicit supervised classification
- Deep Learning
- Reinforcement Learning

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

CONTEXT & VOCABULARY

UNIVERSITÉ CÔTE D'AZUR

WHAT REPRESENTS ARTIFICIAL INTELLIGENCE?

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

What is Artificial intelligence?

- The term ***Artificial Intelligence***, as a research field, was coined at the conference on the campus of Dartmouth College in the summer of **1956**, even though the idea was around since antiquity.
- For instance in the first manifesto of Artificial Intelligence, "*Intelligent Machinery*", in **1948** Alan Turing distinguished two different approaches to AI, which may be termed ***"top-down"*** or ***knowledge-driven AI*** and ***"bottom-up"*** or ***data-driven AI***

(sources: Wikipedia & http://www.alanturing.net & Stanford Encyclopedia of Philosophy)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

What is Artificial intelligence?

- The two different approaches to AI can be detailed:
  - ***"top-down"*** or ***knowledge-driven AI***
    - cognition = high-level phenomenon, independent of low-level details of implementation mechanism, first neuron (1943), first neural network machine (1950), neucognitron (1975)
    - Evolutionary Algorithms (1954,1957, 1960), Reasoning (1959,1970), Expert Systems (1970), Logic, Intelligent Agent Systems (1990)…
  - ***"bottom-up" or data-driven AI***
    - opposite approach, start from data to build incrementally and mathematically mechanisms taking decisions
    - Machine learning algorithms, Decision Trees (1983), Backpropagation (1984-1986), Random Forest (1995), Support Vector Machine (1995), Boosting (1995), Deep Learning (1998/2006)…

(sources: Wikipedia & http://www.alanturing.net & Stanford Encyclopedia of Philosophy)

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

What is Artificial intelligence?

- *AI is originally defined, by Marvin Lee Minsky, as "the construction of computer programs doing tasks, that are, for the moment, accomplished more satisfyingly by human beings because they require high level mental processes such as: learning. perceptual organization of memory and critical reasoning".*

- There are so the "artificial" side with the usage of computers or sophisticated electronic processes and the side "intelligence" associated with its goal to imitate the (human) behavior.

(sources: Wikipedia & http://www.alanturing.net & Stanford Encyclopedia of Philosophy)

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

What is Artificial intelligence?

- The concept of ***strong artificial intelligence*** makes reference to a machine capable not only of producing intelligent behavior, but also to experience a feeling of a real sense of itself, “real feelings” (whatever may be put behind these words), and "an understanding of its own arguments”.

- The notion of ***weak artificial intelligence*** is a pragmatic approach of engineers: targeting to build more autonomous systems (to reduce the cost of their supervision), algorithms capable of solving problems of a certain class, etc. But this time, the machine *simulates* the intelligence, it seems to act as if it was smart.

(sources: Wikipedia & http://www.alanturing.net & Stanford Encyclopedia of Philosophy)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Why Artificial Intelligence is so difficult to grasp?

- Frequently, when a technique reaches mainstream use, it is no longer considered as artificial intelligence; this phenomenon is described as the AI effect: "AI is whatever hasn't been done yet." (*Larry Tesler's Theorem*) -> e.g. Path Finding (GPS), Checkers game, Chess electronic game, Alpha Go…

$\Rightarrow$ "AI" is continuously evolving and so very difficult to grasp.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Machine Learning

$$\left(\mathbf{x}\right) \xrightarrow{f(\mathbf{X},\boldsymbol{\alpha})\ ?} y$$

$\left(\mathbf{x}\right)$

Face Detection

Betting on sports

Speech Recognition

y

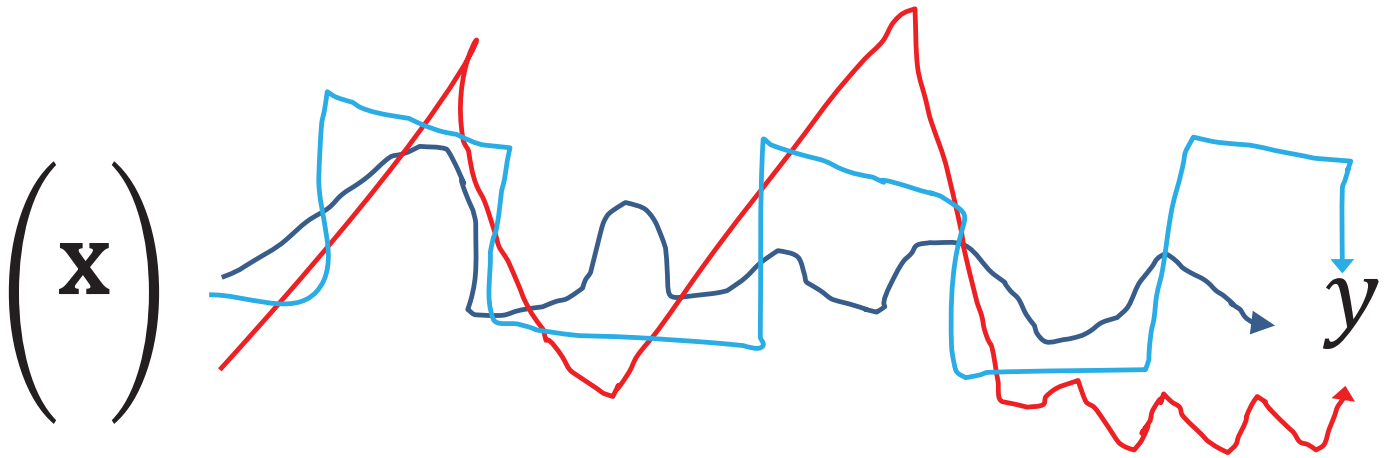
Scores, ranking…

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Machine Learning

$$\left(\mathbf{x}\right) \xrightarrow{f(\mathbf{X},\boldsymbol{\alpha})\ ?} y$$



Support Vector Machines *Random Forest* *Artificial Neural Networks*

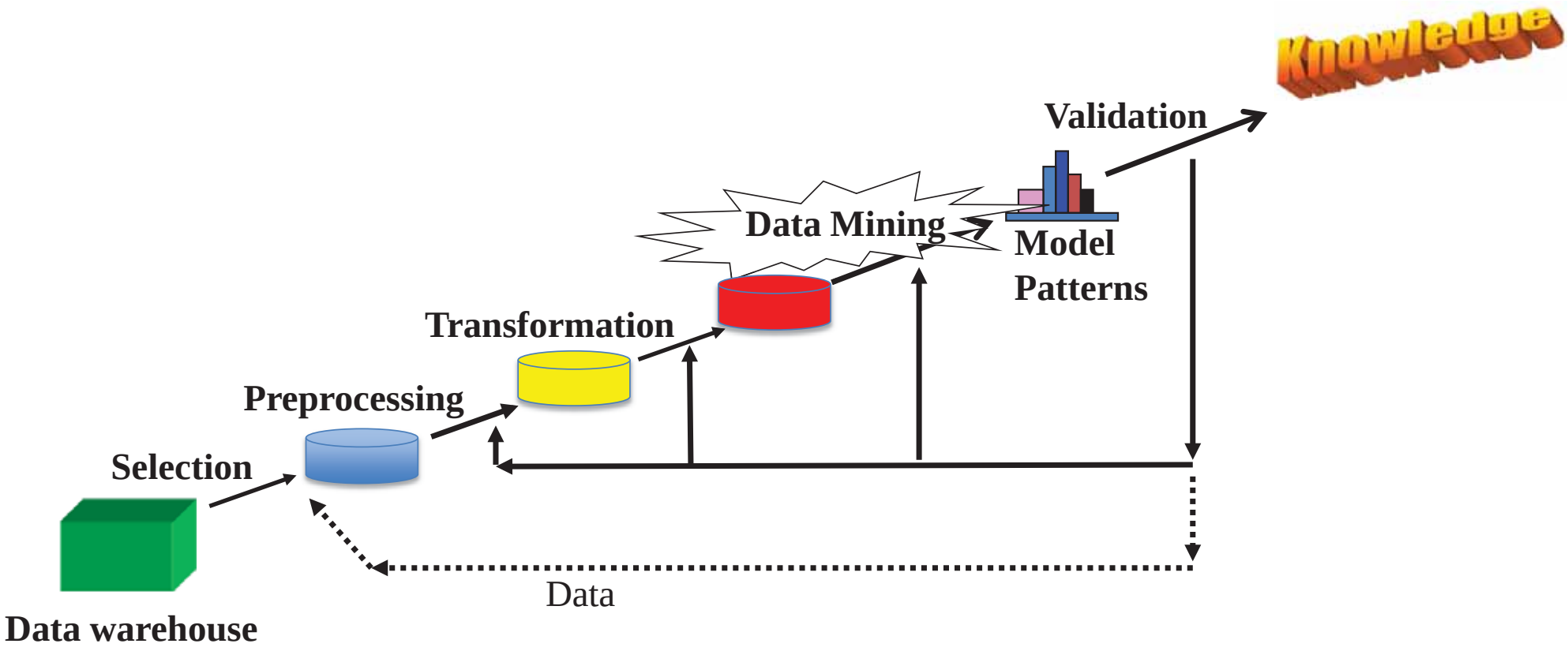
UNIVERSITÉ CÔTE D'AZUR

MACHINE LEARNING VS DATA MINING?

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

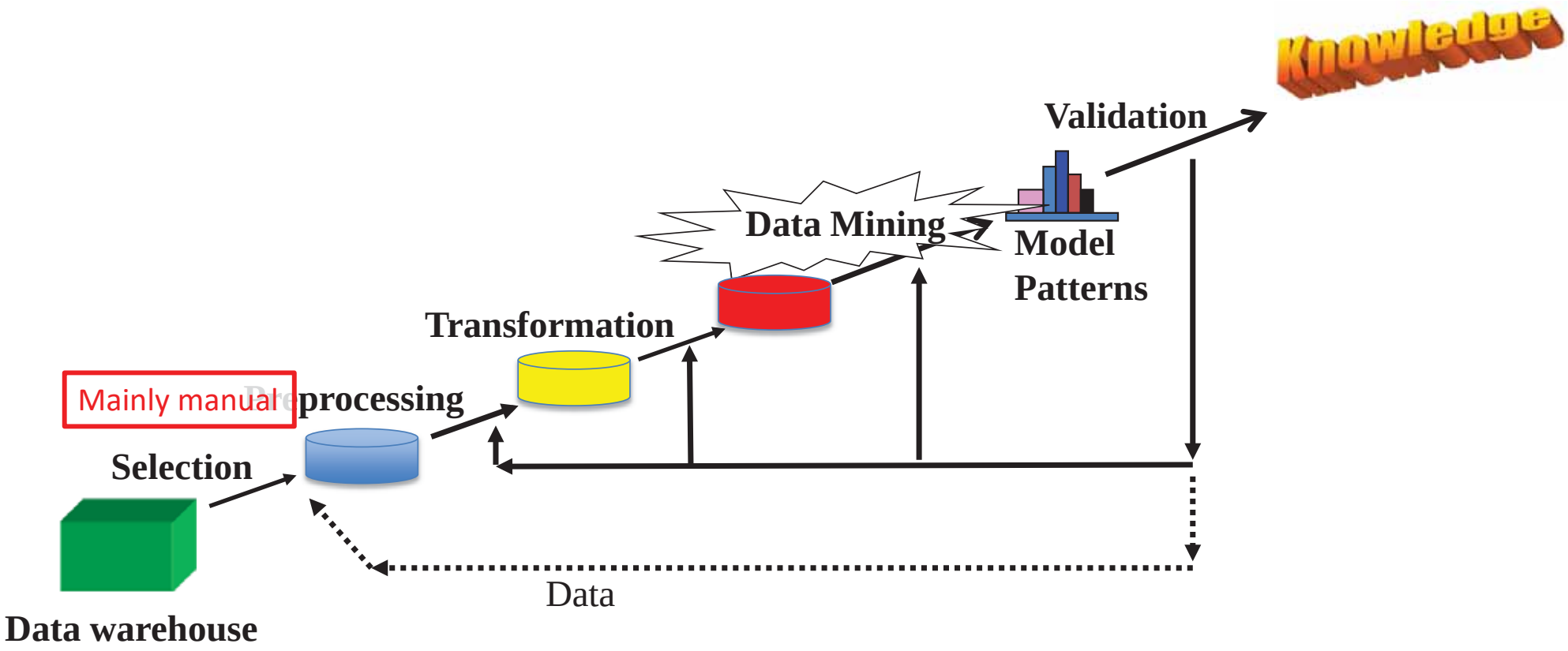
Data Mining Workflow



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

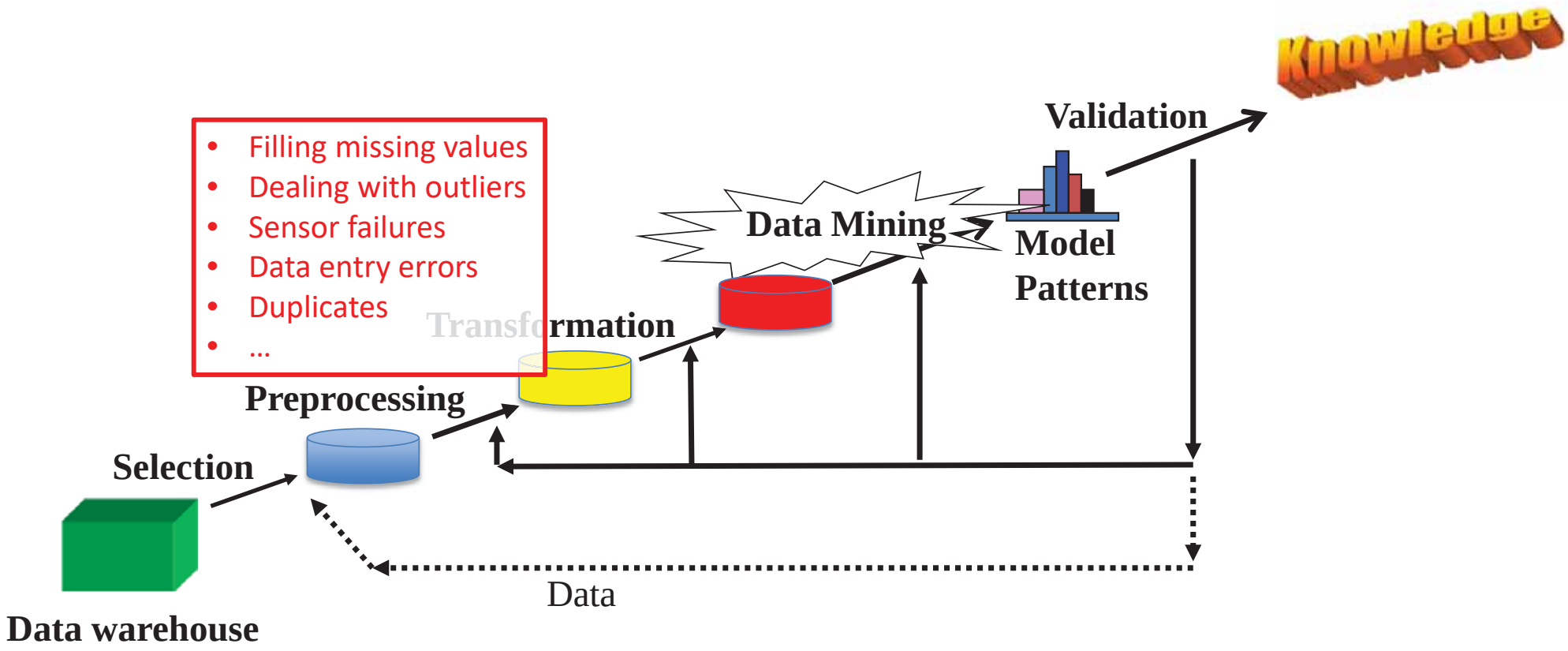
Data Mining Workflow



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

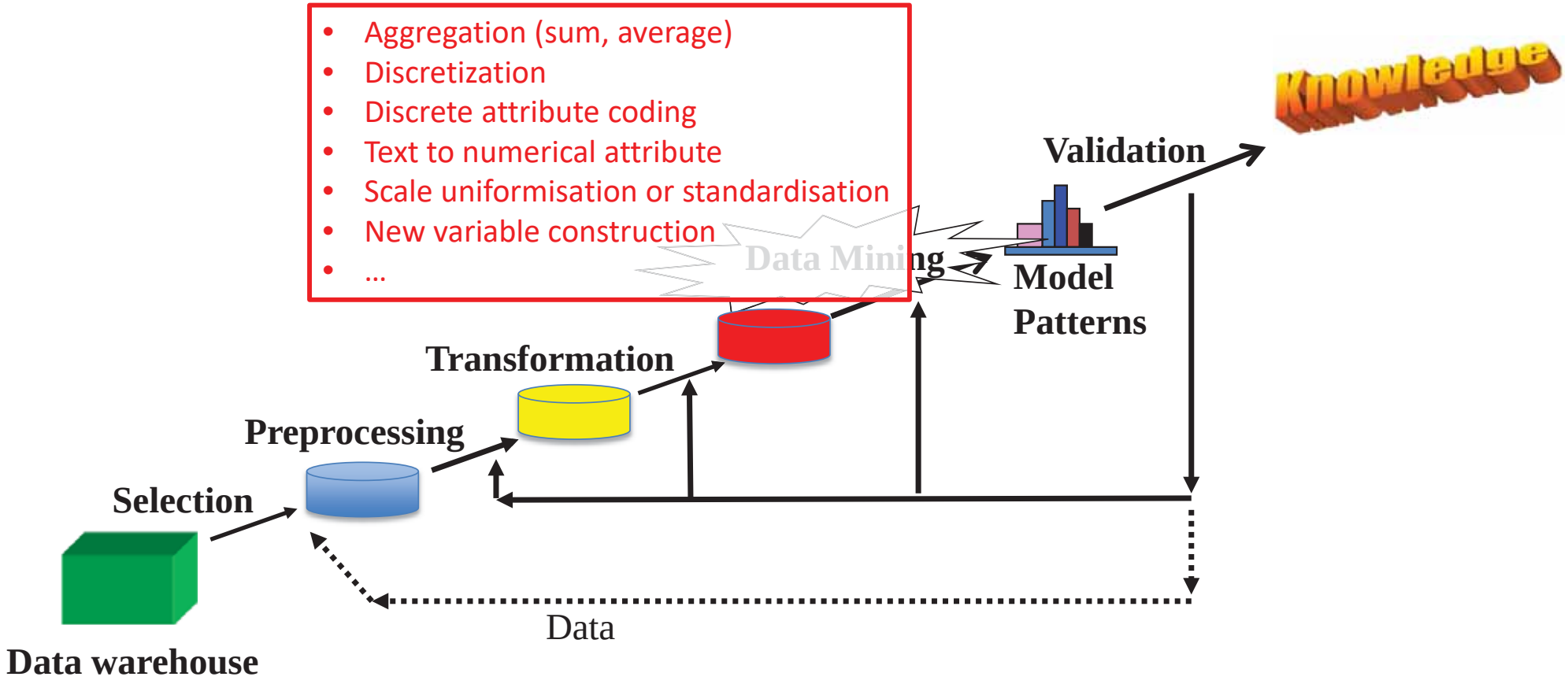
Data Mining Workflow



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Data Mining Workflow

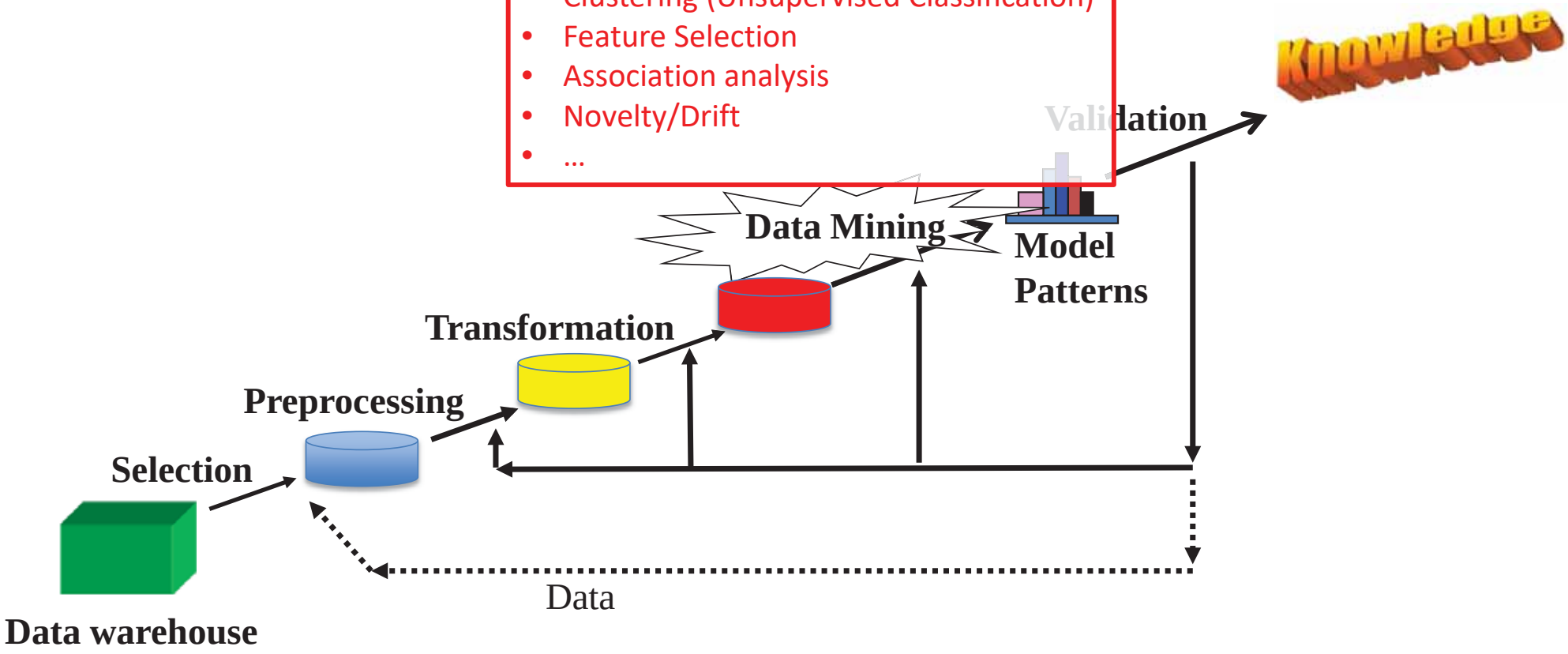


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Data Mining Workflow

- Regression
- (Supervised) Classification
- Clustering (Unsupervised Classification)
- Feature Selection
- Association analysis
- Novelty/Drift
- …

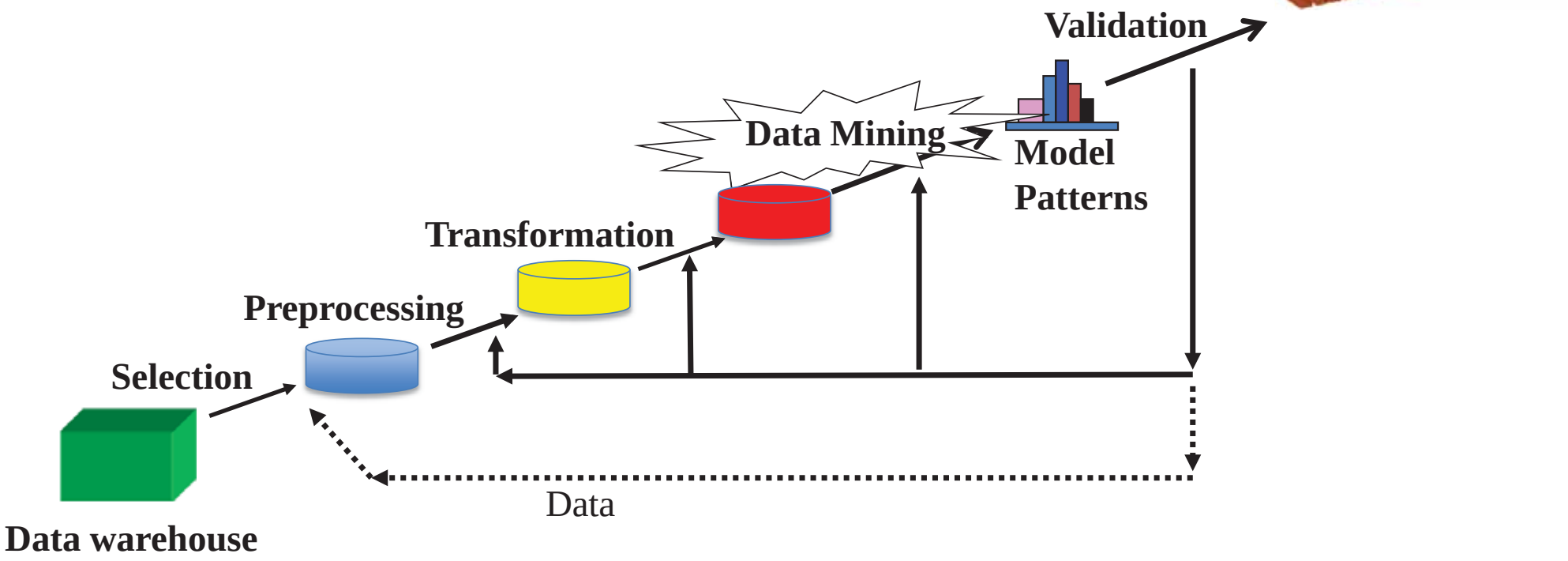


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Data Mining Workflow

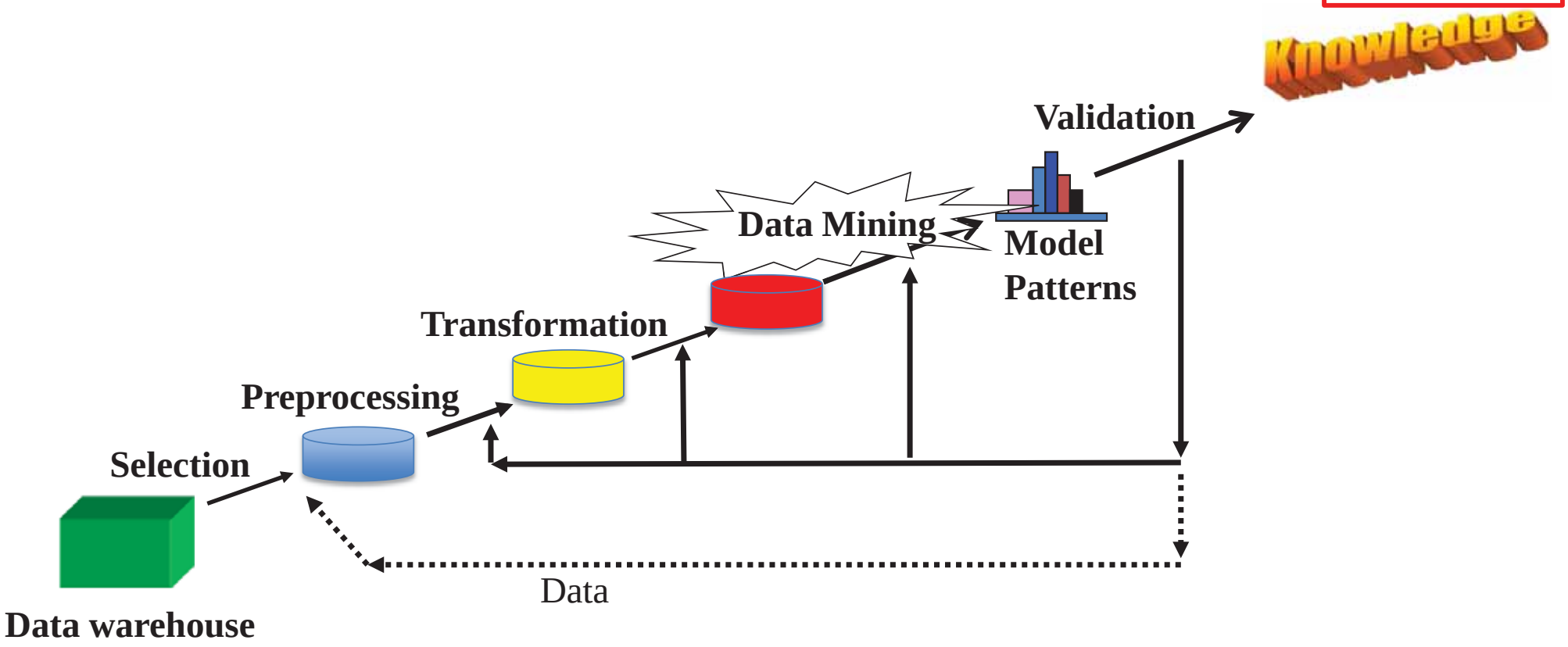
- Evaluation on Validation Set
- Evaluation Measures
- Visualization
- ...



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Data Mining Workflow

- Visualization
- Reporting
- Knowledge
- ...



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Data Mining Workflow

Problems

- Regression
- (Supervised) Classification
- Density Estimation / Clustering (Unsupervised Classification)
- Feature Selection
- Association analysis
- Anomality/Novelty/Drift
- …

Possible Solutions

- Machine Learning
  - Support Vector Machine
  - Artificial Neural Network
  - Boosting
  - Decision Tree
  - Random Forest
  - …
- Statistical Learning
  - Gaussian Models (GMM)
  - Naïve Bayes
  - Gaussian processes
  - …
- Other techniques
  - Galois Lattice
  - …

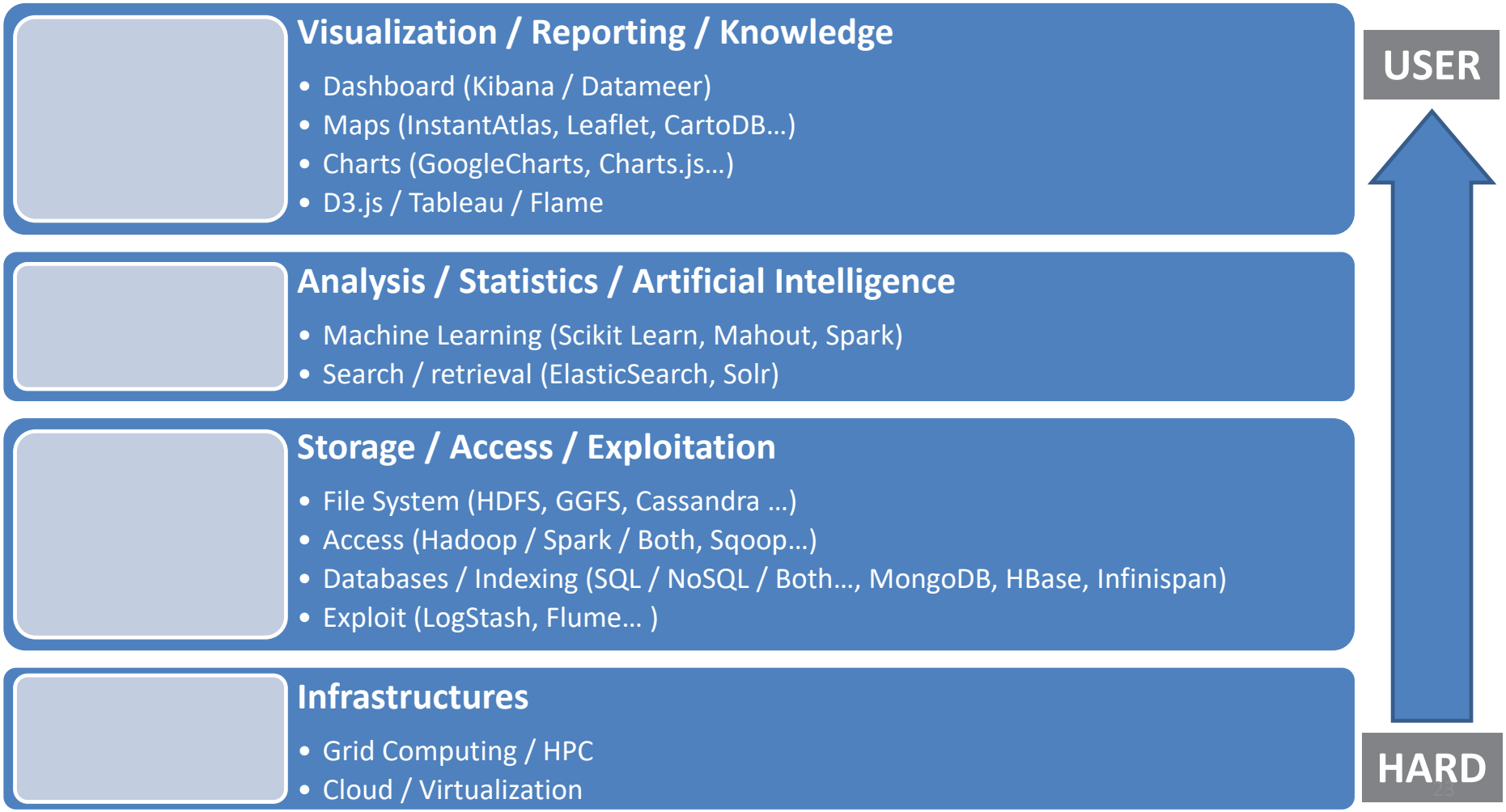
UNIVERSITÉ CÔTE D'AZUR

MACHINE LEARNING VS DATA SCIENCE?

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Data Science Stack

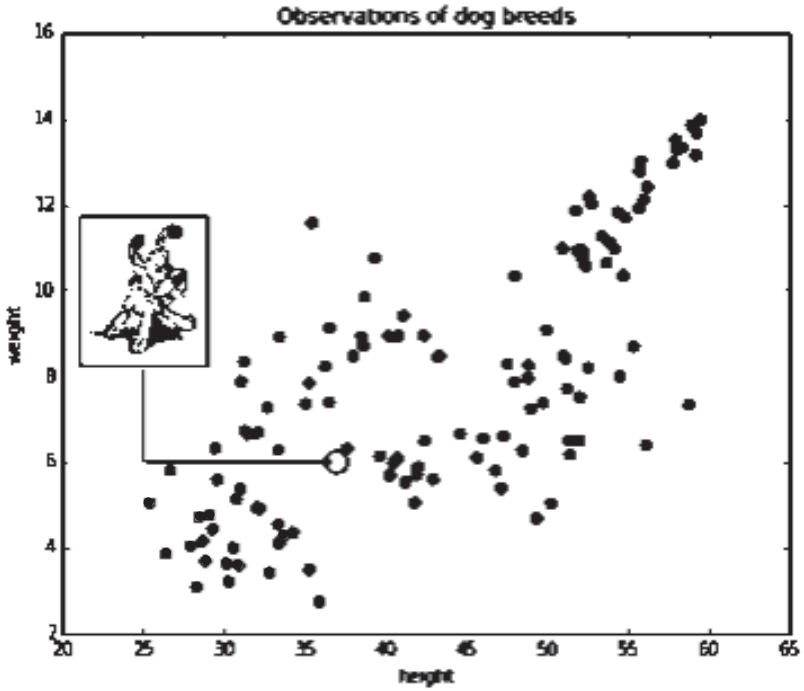


UNIVERSITÉ CÔTE D'AZUR

MACHINE LEARNING VS STATISTICS?

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

What breed is that Dogmatix (Idéfix) ?

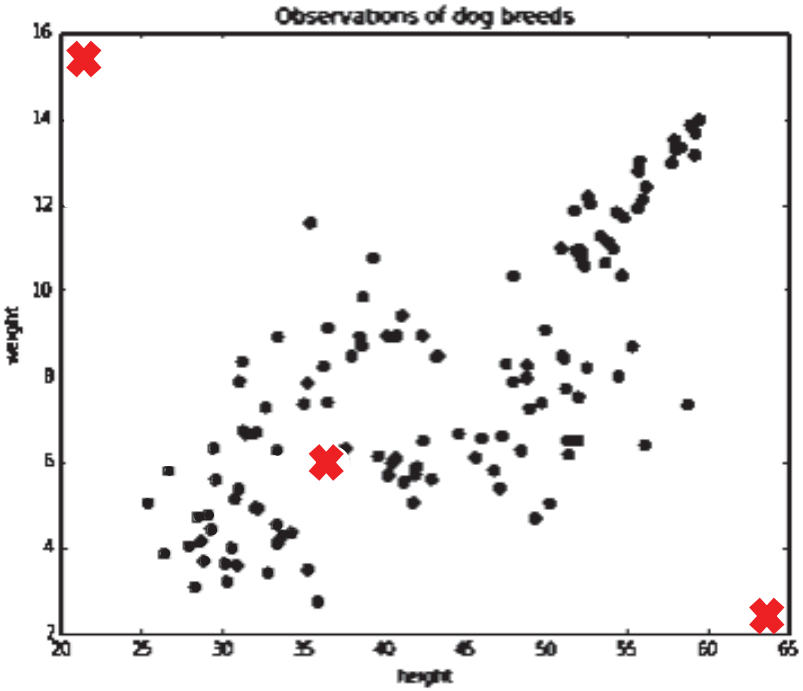


The illustrations of the slides in this section come from the blog "Bayesian Vitalstatistix: What Breed of Dog was Dogmatix?"

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

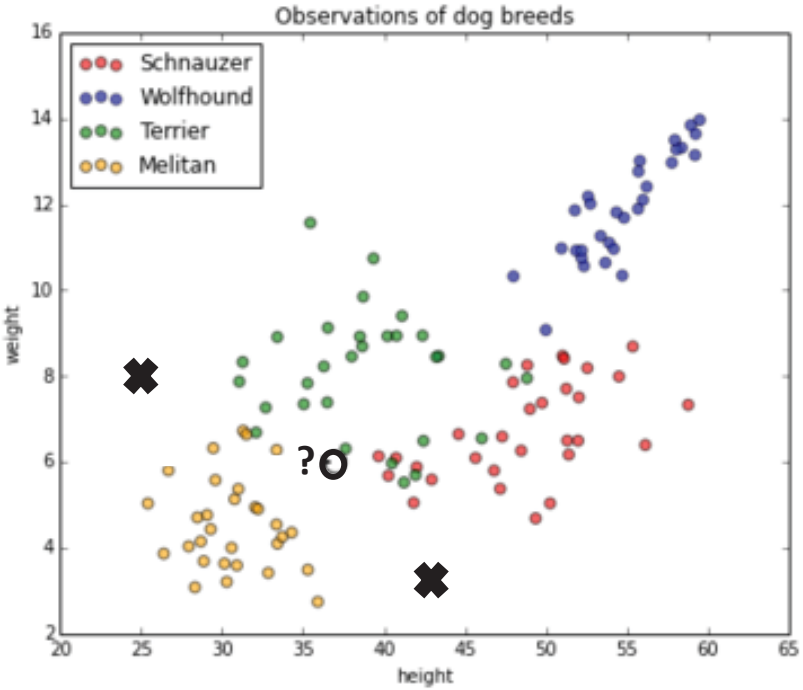
Does any real dog get this height and weight?



- Let us consider ***x***, vectors independently generated in $\boldsymbol{R}^d$ (here $\boldsymbol{R}^2$), following a probability distribution fixed but ***unknown*** $P(\boldsymbol{x})$.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

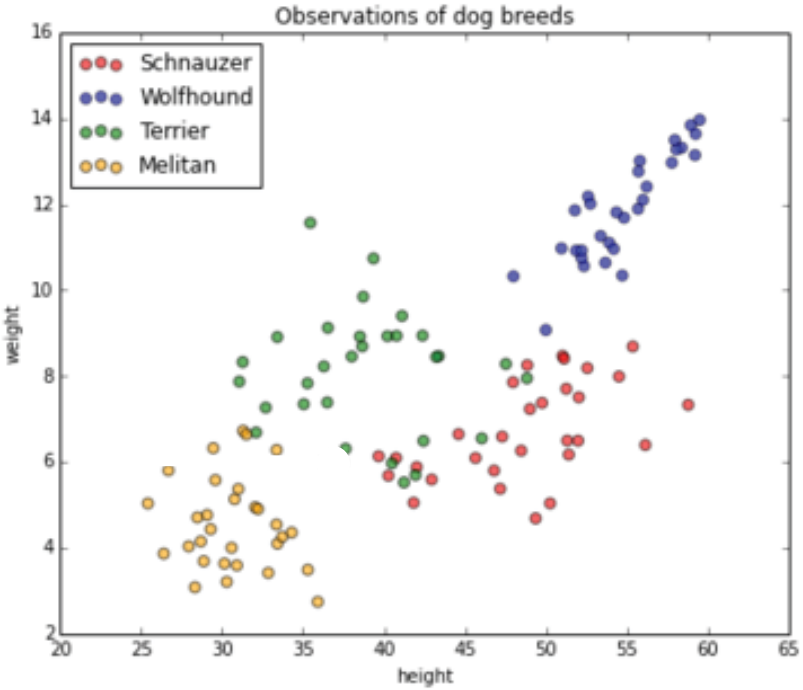
What should be the breed of these dogs?



- An Oracle assignes a value *y* to each vector ***x*** following a probability distribution *P(y|**x**)* also fixed but ***unknown***.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

An oracle provides me with examples?

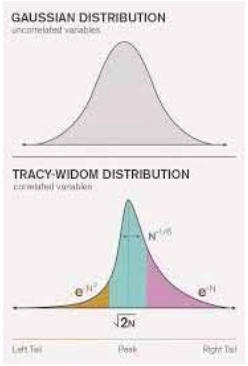


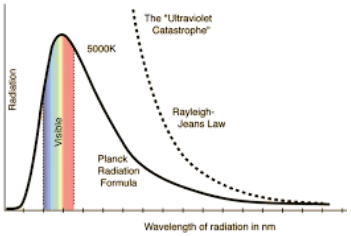
- Let S be a training set $S = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, \boldsymbol{y}_m)\}$, with m **training samples i.i.d.** which follow the **joint probability** $P(\boldsymbol{x}, y) = P(\boldsymbol{x})P(y|\boldsymbol{x})$.

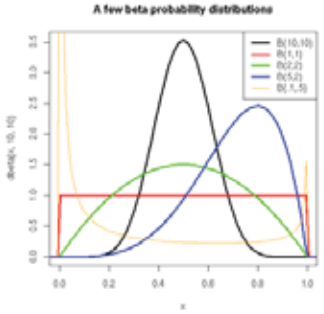
UNIVERSITÉ CÔTE D'AZUR

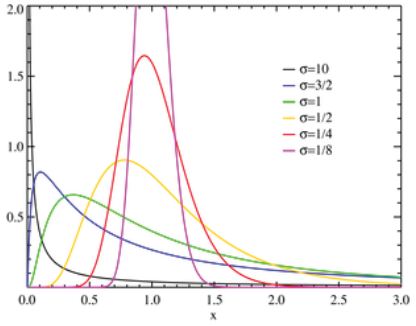
univ-cotedazur.fr

Statistical solution: Models, Hypotheses…



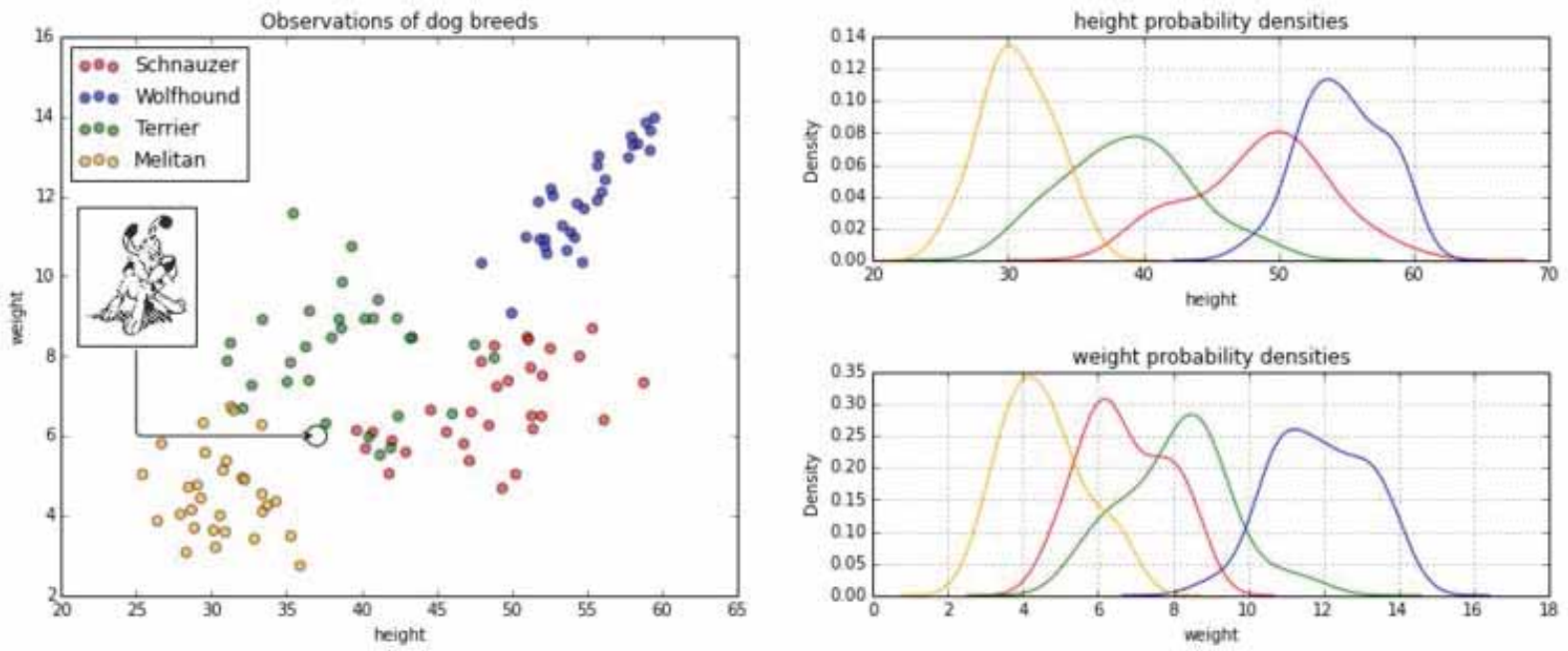






UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

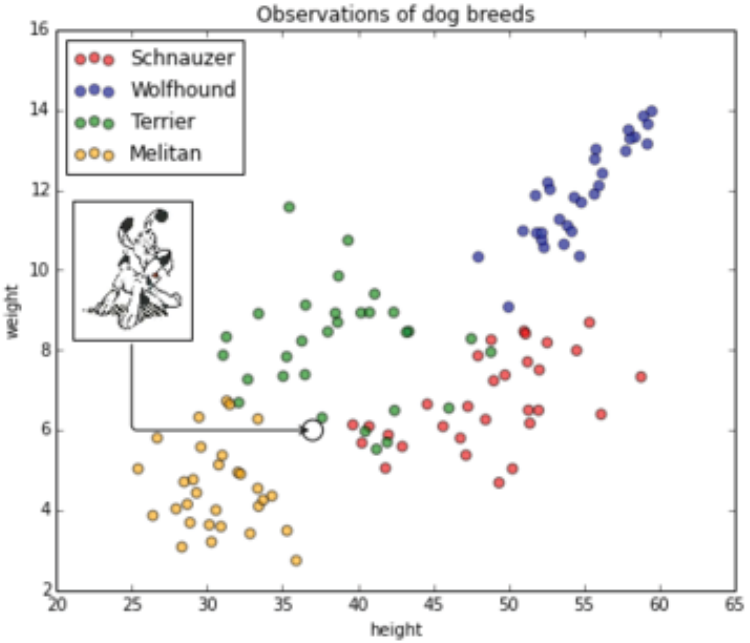
Statistical solution P(height, weight|breed)…

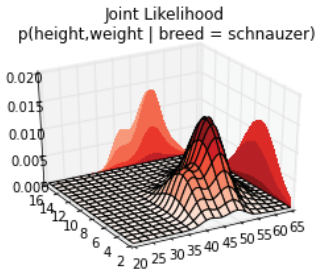


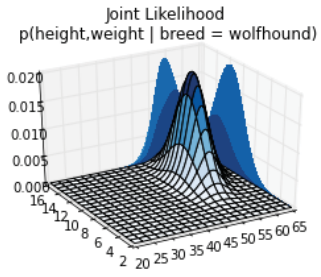
UNIVERSITÉ CÔTE D'AZUR

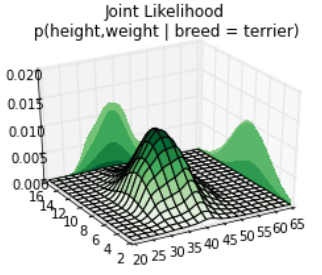
univ-cotedazur.fr

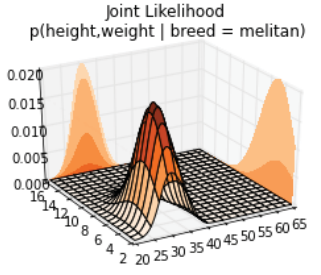
Statistical solution P(height, weight|breed)…





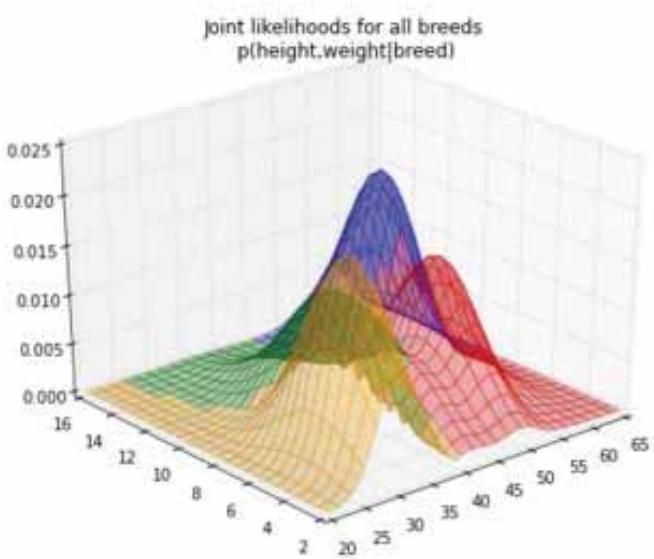


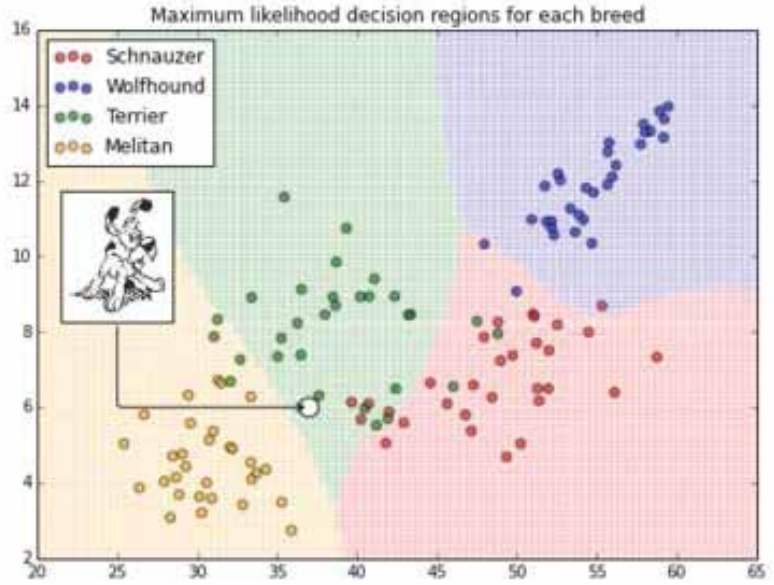




UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Statistical solution P(height, weight|breed)…

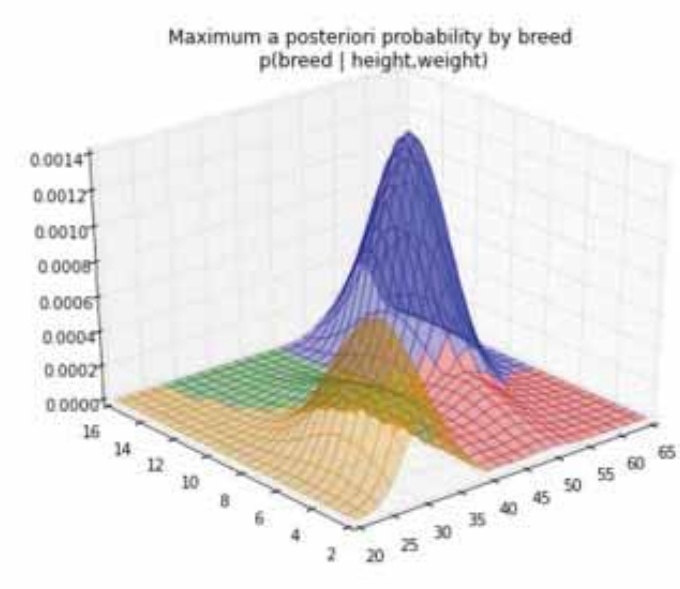


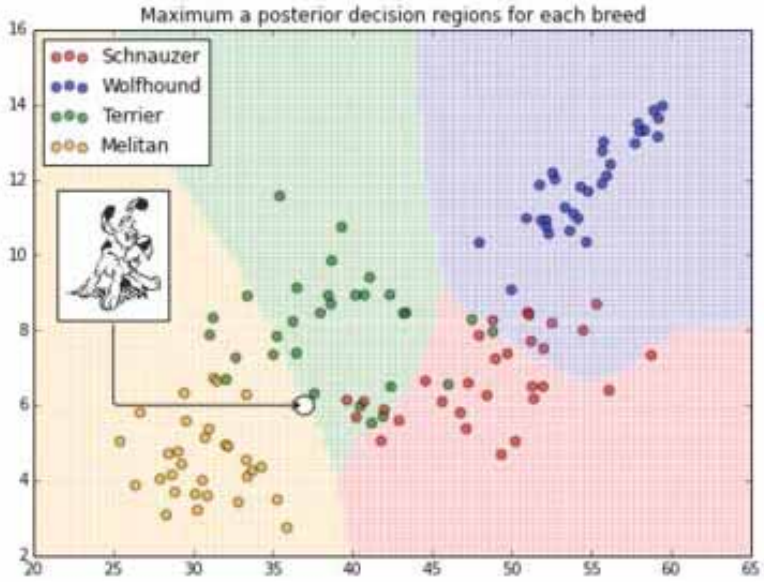


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Statistical solution: Bayes, P(breed|height, weight)…

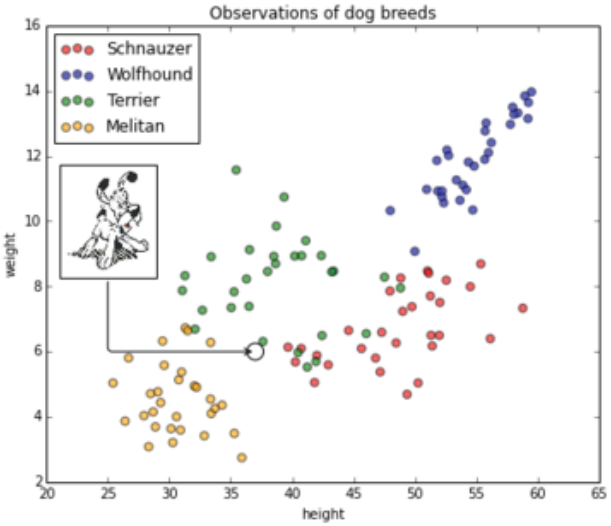




UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Machine Learning



- we have a learning machine which can provide a family of functions *{f(**x**;α)}*, where α is a set of parameters.

$$\left(\mathbf{x}\right) \xrightarrow{f(\mathbf{X},\boldsymbol{\alpha})\ ?} y$$

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

The problem in Machine Learning

$$\left(\mathbf{X}\right) \xrightarrow{f(\mathbf{X},\boldsymbol{\alpha})\ ?} y$$

- The problem of learning consists in finding ***the model*** (among the $\{f(\mathbf{x};\alpha)\}$) which provides ***the best approximation*** $\hat{y}$ of the true label y given by the Oracle.
- ***best*** is defined in terms of minimizing a specific (error) cost ***related to your problem/objectives***

  $Q((\mathbf{x}, y), \alpha) \in [a; b]$.
- Examples of cost/loss functions Q: Hinge Loss, Quadratic Loss, Cross-Entropy Loss, Logistic Loss…

Loss in Machine Learning

- ***How to define the loss L (or the cost Q)?***

You should choose the right loss function based on your problem and your data *(here y is the true/expected answer, $f(x)$ the answer predicted by the network)*.

Classification

- ***Cross-entropy loss***: $L(x) = -(y\ ln(f(x)) + (1-y)ln(1-f(x)))$
- ***Hinge Loss*** *(i.e. max-margin loss, i.e. 0-1 loss)*: $L(x) = max(0, 1-yf(x))$
- …

Regression

- ***Mean Square Error*** *(or Quadratic Loss)*: $L(x) = (f(x)-y)^2$
- ***Mean Absolute Loss***: $L(x) = |f(x)-y|$
- …

If the loss is minimized but accuracy is low, you should check the loss function. Maybe it is not the appropriate one for your task.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

The problem in Machine Learning

For Clarity sake, let us note $z = (\boldsymbol{x}, y)$.

- Thus, the objective is to minimize the ***Risk***, i.e. the expectation of the error cost:

$$R(\alpha) = \int Q(z, \alpha) dP(z)$$

where $P(z)$ is unknown.

The training set $S = \{z_i\}_{i=1,\ldots,m}$ is built through an *i.i.d.* sampling according to $P(z)$. Since we cannot compute $R(\alpha)$, we look for minimizing the ***Empirical Risk*** instead:

$$R_{emp}(\alpha) = \frac{1}{m} \sum Q(z_i, \alpha)$$

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Machine Learning fundamental Hypothesis

For Clarity sake, let us note $z = (\boldsymbol{x}, y)$.

$S = \{z_i\}_{i=1,\dots,m}$ is built through an *i.i.d.* sampling according to P(z).

Machine Learning $\Leftarrow$ ***Statistics***

Train through Cross-Validation

Machine Learning $\Leftarrow$ ***Statistics***

Training set & Test set have to be distributed according to the same law (i.e. P(z)).

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Vapnik learning theory (1995)

Vapnik had proven the following equation $\forall m$ with a probability at least equal to $1 - \eta$:

$$R(\alpha_m) \leq R_{emp}(\alpha_m) + (b - a)\sqrt{\frac{d_{VC}(ln(2m/d_{VC}) + 1) - ln(\eta/4)}{m}}$$

Training Error ***Generalization Error***

Thus minimizing the ***Risk*** depends on minimizing the ***Empirical Risk*** and the ***confidence interval*** which is linked to the term d_{VC} corresponding to the complexity of the model family chosen, i.e. the Vapnik-Chervonenkis dimension

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Vapnik learning theory (1995)

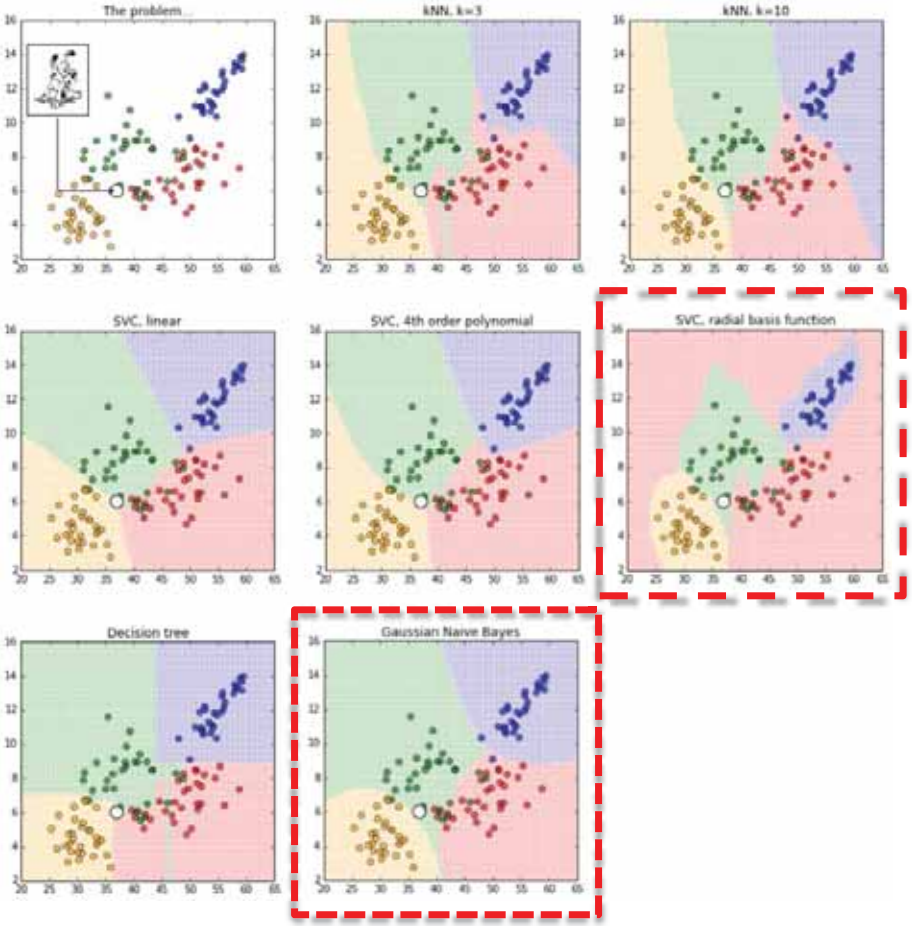
In his learning theory [Vapnik, 1995], Vapnik defines 4 fundamental steps:

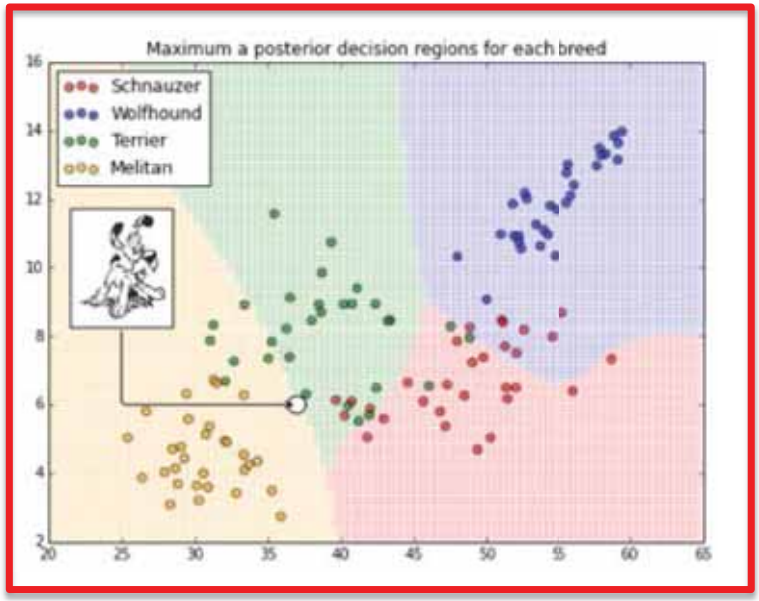
- Study the theory of consistence of learning processes
- Define bounds on convergence speed of learning processes
- Handle the generalization power of learning processes
- Design a theory to build learning algorithms in order to find a tradeoff between minimizing the ***Empirical Risk*** and the ***confidence interval*** $\Rightarrow$ minimization of the ***Structural Risk***.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Machine Learning vs Statistics





UNIVERSITÉ
CÔTE D'AZUR
univ-cotedazur.fr

UNSUPERVISED CLASSIFICATION

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Unsupervised classification

- The system or the operator has only samples, but no label
- The number of classes and their nature have not been predetermined

$\Rightarrow$ unsupervised learning or clustering.

$\Leftrightarrow$ No expert is required.

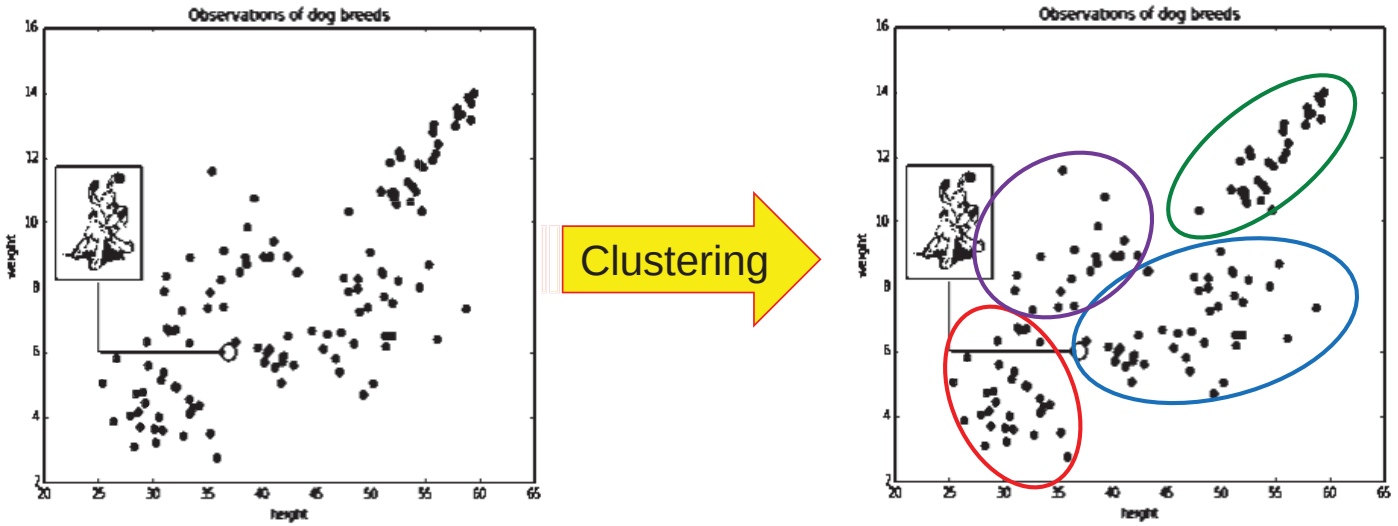
$\Leftrightarrow$ The algorithm must discover by itself more or less hidden/underlying data structure.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Clustering

Clustering: Partition a dataset into groups based on the similarity between the instances.

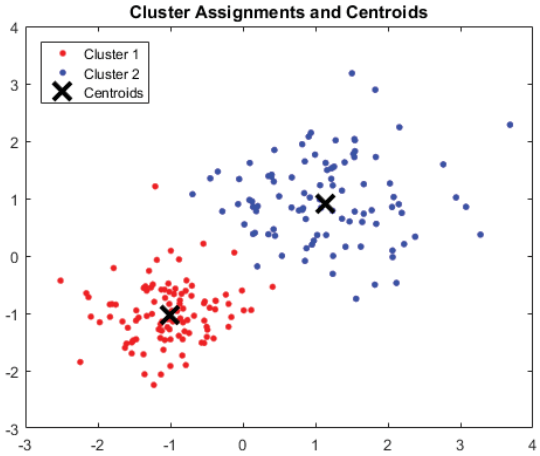


UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

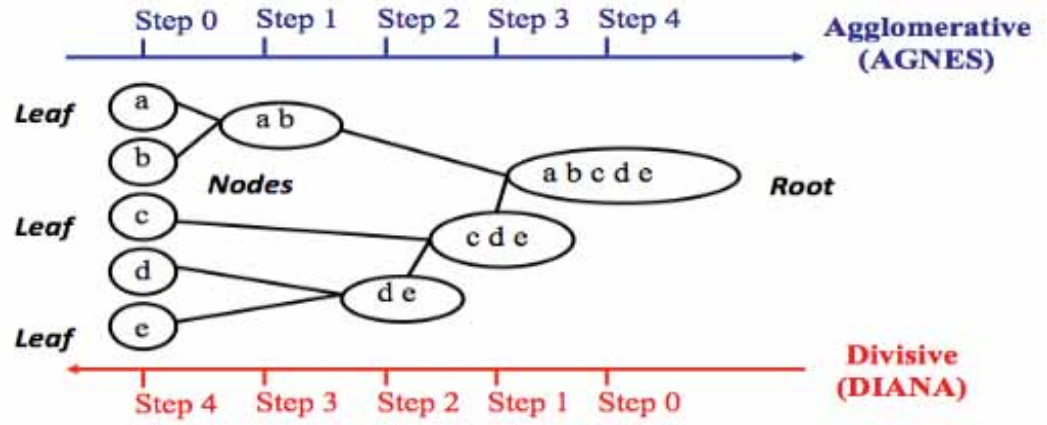
Clustering Algorithms (Partition)

- **Centroid-based (1957,1967)**

(E.g. K-means, PAM, CLARA)



- **Hierarchical:**
  - Bottom up approach.
  - Top down approach.

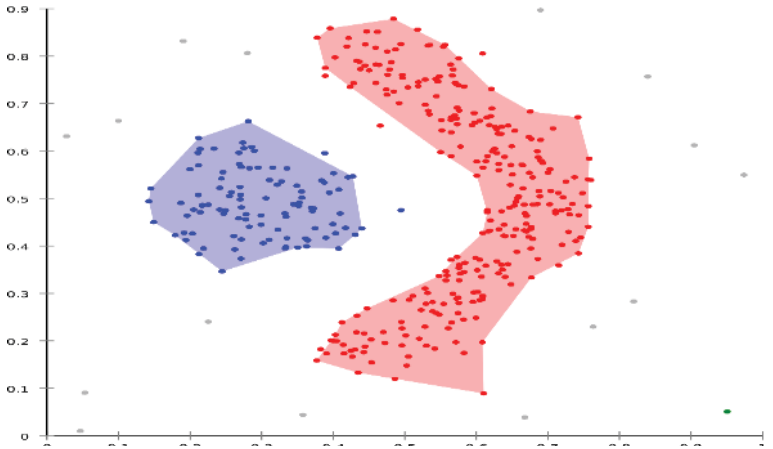


UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Clustering Algorithms (Density)

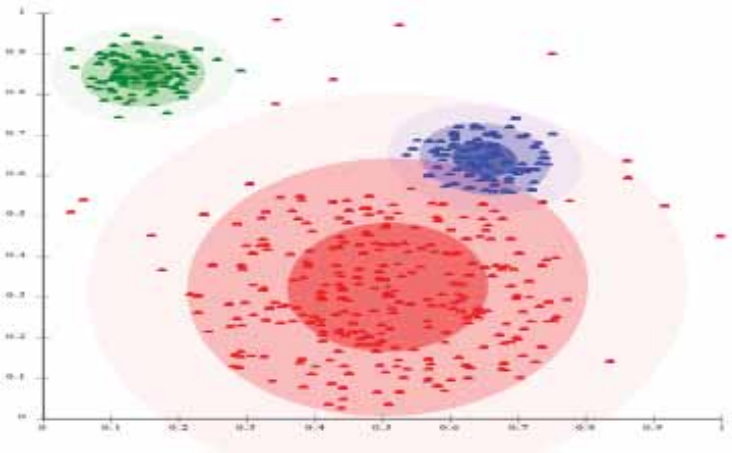
- **Density-based**

(E.g. DBSCAN, DENCLUE)



- **Distribution-based**

(E.g. EM, extension of K-means)

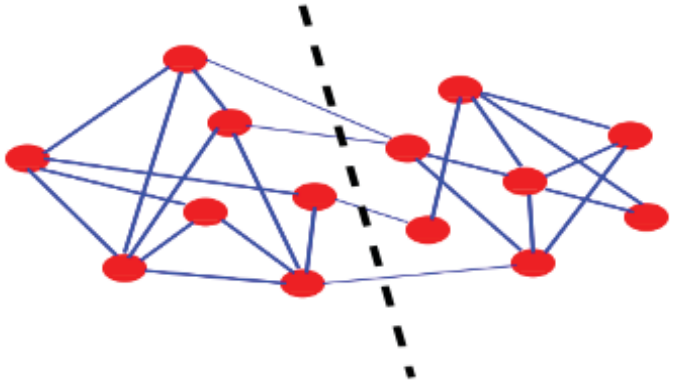


UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Clustering Algorithms (Graph)

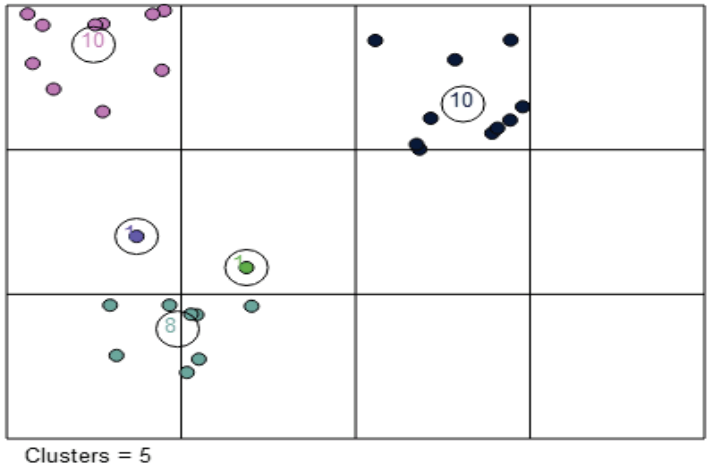
- **Graph-based**

(E.g. Chameleon)



- **Grid-based**

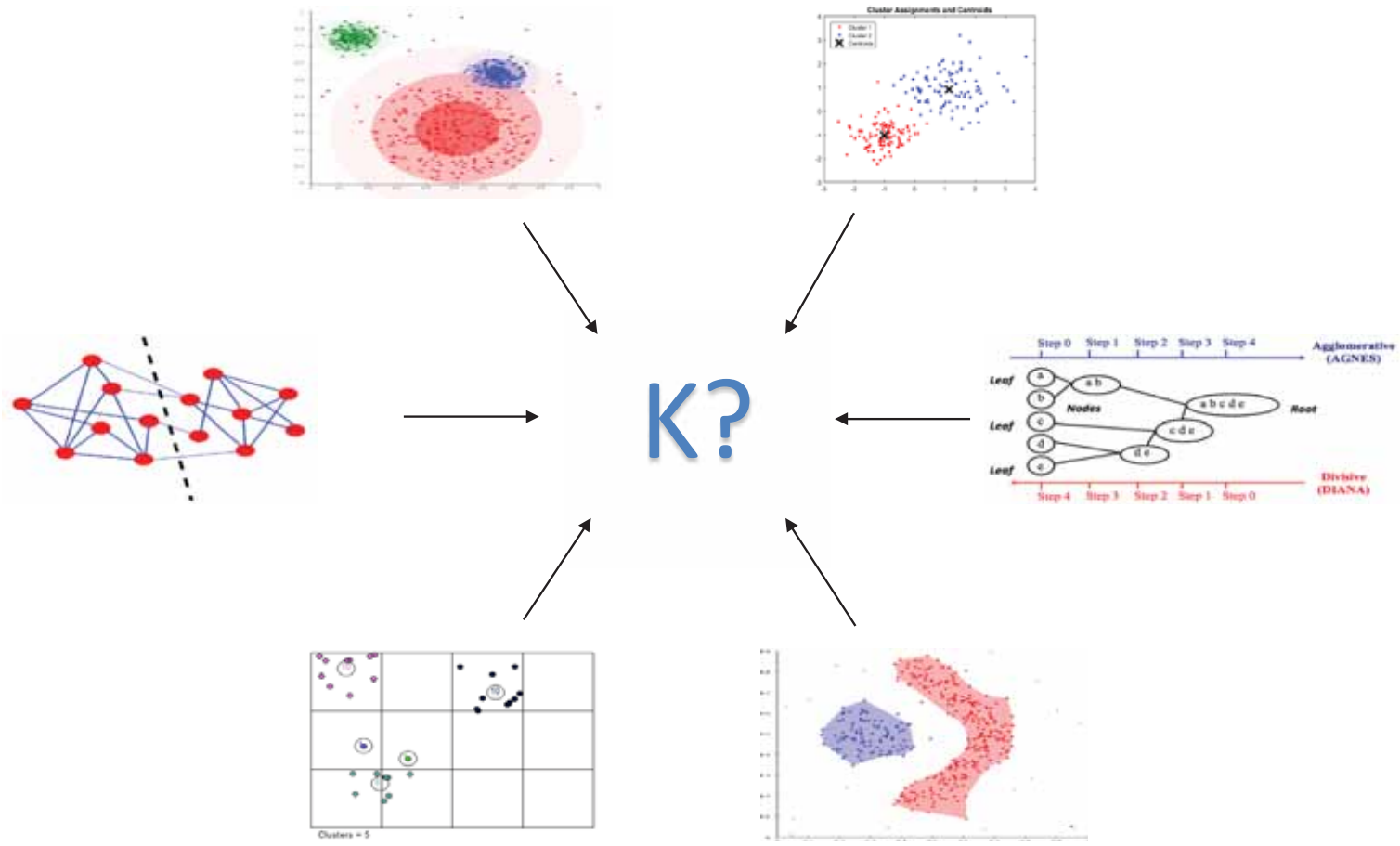
(E.g. STING, CLIQUE)



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

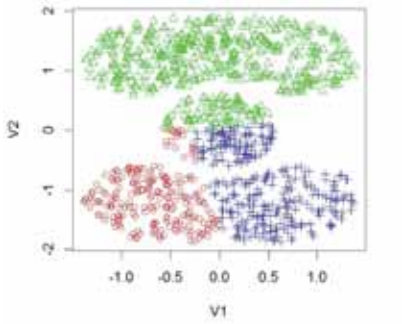
How many Clusters?



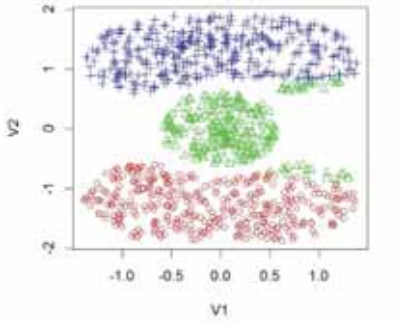
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

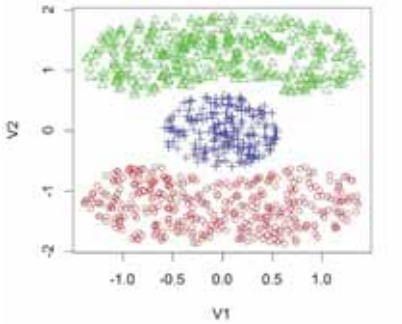
Which Algorithm?



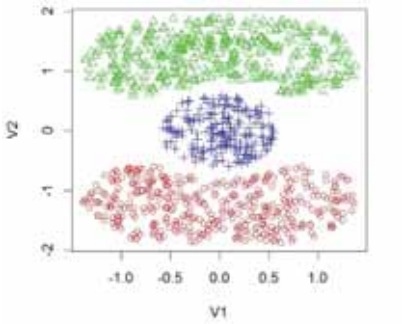
K-means



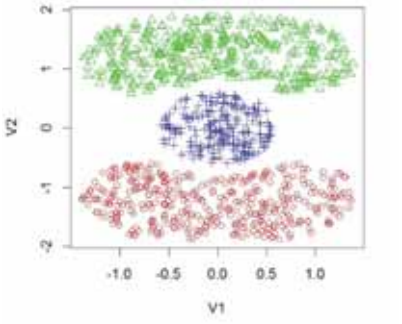
PAM



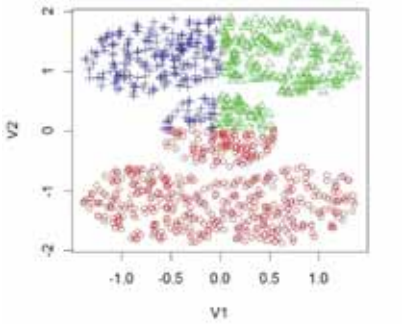
Gaussian model



DBSCAN



Average linkage

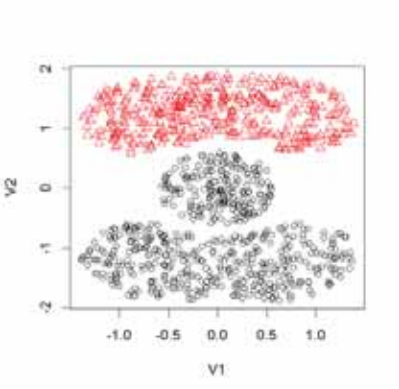


DIANA

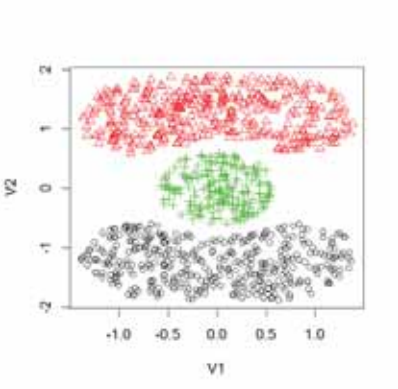
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

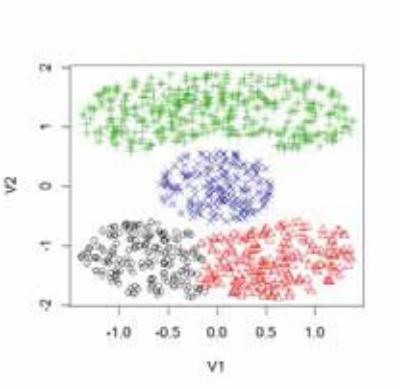
Algorithms' Parameters



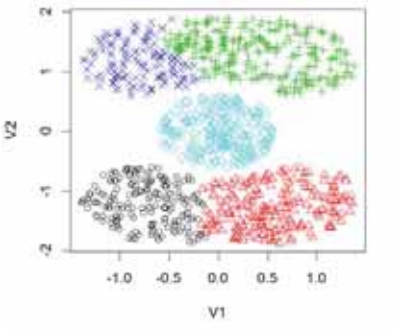
Avg. linkage: K=2



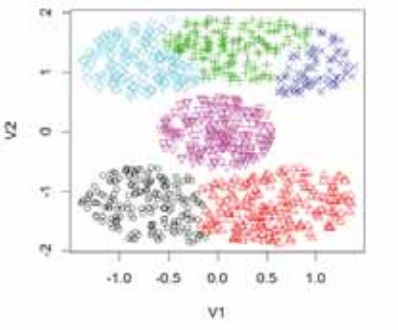
Avg. linkage: K=3



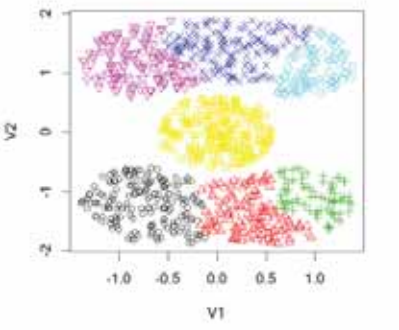
Avg. linkage: K=4



Avg. linkage: K=5



Avg. linkage: K=6



Avg. linkage: K=7

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Which metric/distance?

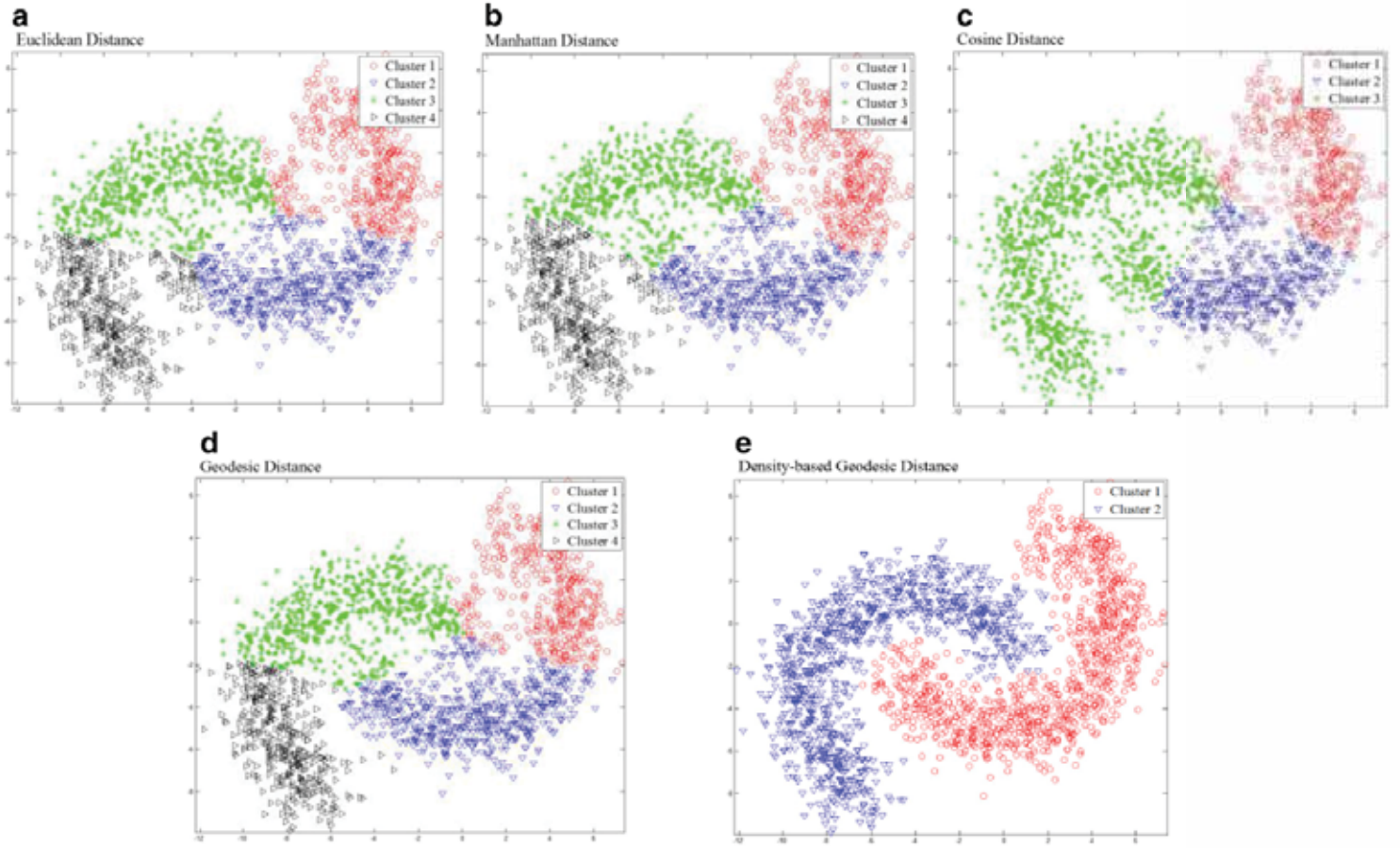
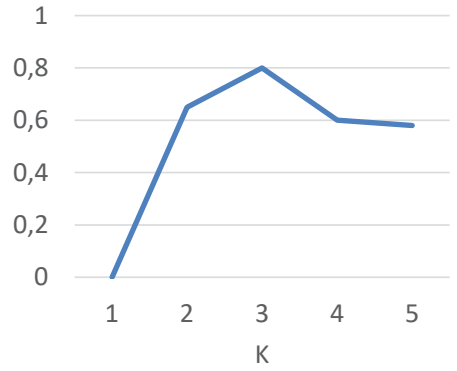


Fig. 7. PAM clustering results of Scenario 2: (a) Euclidean distance ($K = 4$), (b) Manhattan distance ($K = 4$), (c) cosine distance ($K = 3$), (d) geodesic distance ($k = 8$, $K = 4$), and (e) density-based geodesic distance ($k = 30$, $K = 2$).

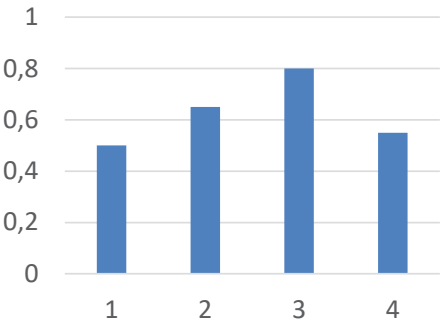
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Clustering validation measures

- ***Internal validation:*** Separation? Compactness?
  E.g. Dunn, DB, and Silhouette indexes.
- ***Problems:***
  - Different performance WRT existence of noise, variable densities, and non well-separated clusters.
  - Overrate the algorithm that uses the same clustering model.



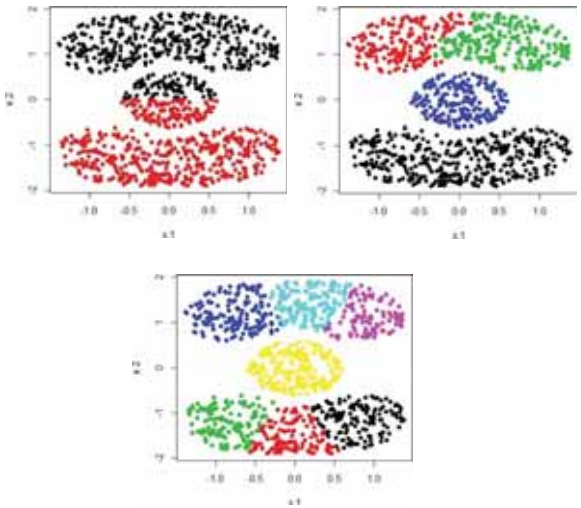
- ***External validation:*** == Class labels?
  E.g. Rand, Jaccard, Purity, MI, VI indexes.
- ***Problem:*** Some class labels, at least, have to exist.



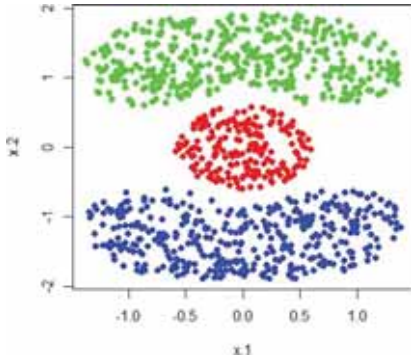
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Consensus clustering



Consensus Clustering



Ensemble of 3 base clusterings

Consensus solution

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Overview

- Unsupervised classification
- **Explicit supervised classification**
  - **Decision Trees**
  - **Random Forest**
- Implicit supervised classification
- Deep Learning
- Reinforcement Learning

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

EXPLICIT SUPERVISED CLASSIFICATION

UNIVERSITÉ CÔTE D'AZUR

DECISION TREES

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Decision tree to decide playing tennis or not

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Objective

2 classes: yes & no

Prediction if a game will be played or not

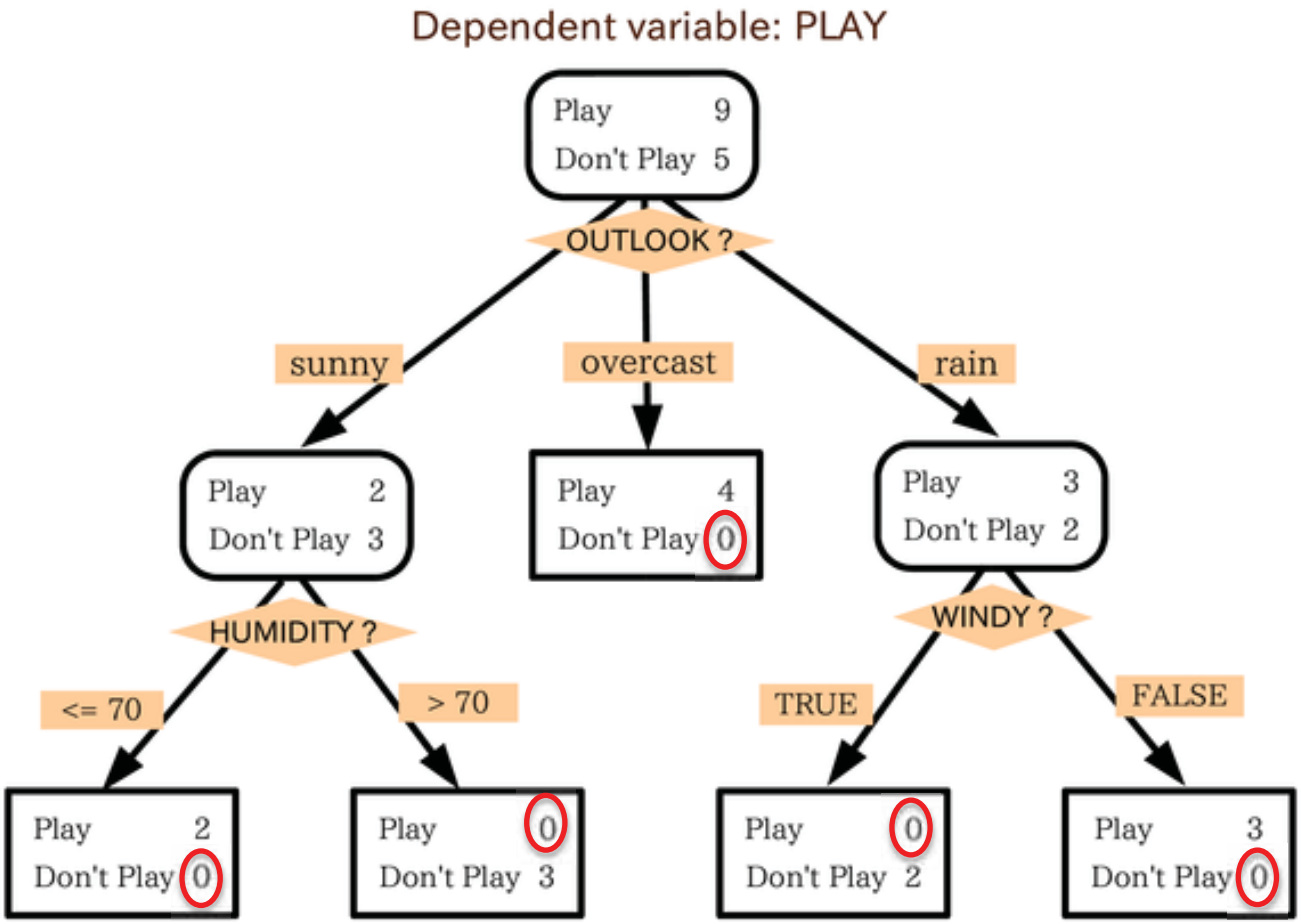
Temperature will be easily converted into numerical

I.H. Witten and E. Frank, "Data Mining", Morgan Kaufmann Pub., 2000.

UNIVERSITÉ CÔTE D'AZUR

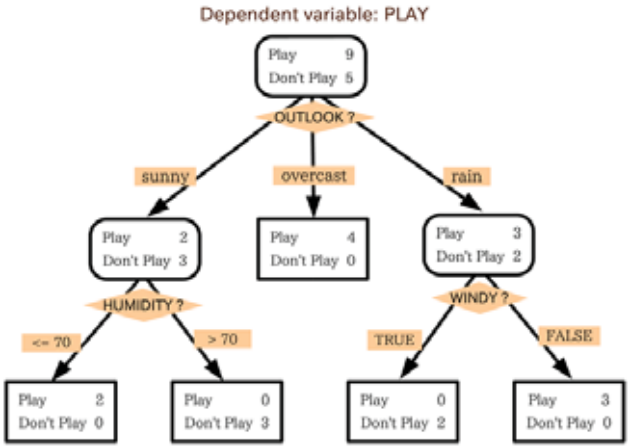
univ-cotedazur.fr

A simple example



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Example - explanations



- On the nodes
  - Distribution of the variable to predict
- The first node is segmented with the variable outlook (sunny, overcast, rainy): creation of 3 sub-groups
  - The first group contains 5 observations, 2 yes and 3 no
- The tree can be translated in a set of decision rules without loosing any information
  - Example: if outlook = sunny and humidity = high then play = yes

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Basic algorithm

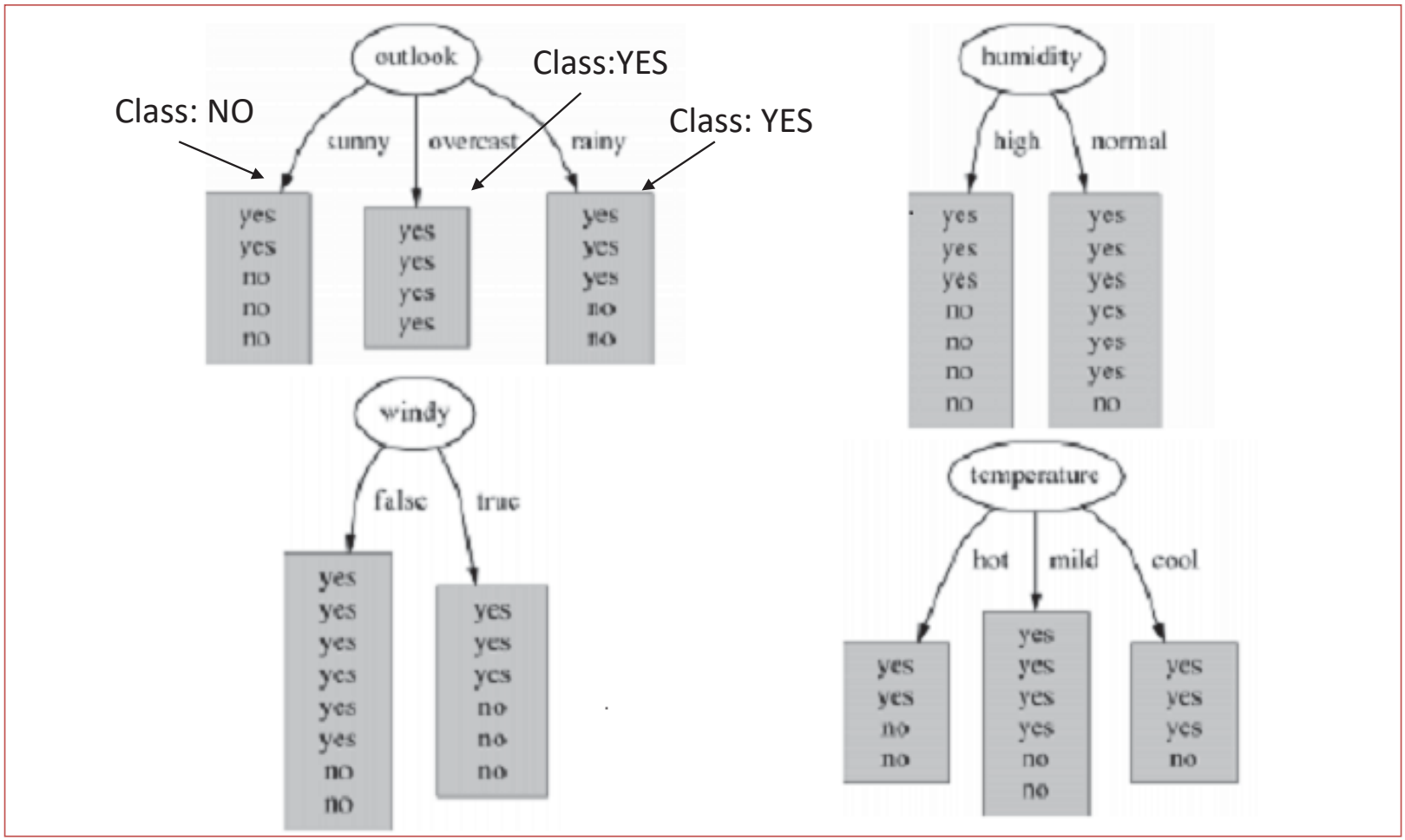
- A= BestAttribute(Examples) `// Best attribute means more //homogenous results`
- Assign A to the root
- For each value of A, create a new sub-node of the roof
- Classify all the examples in the sub-nodes
- If all examples of a sub-node are **homogeneous**, assign their class to the node, if not repeat this process from this node

- ***Question: How to measure homogeneity?***
  Entropy, Gini, Information Gain…

UNIVERSITÉ CÔTE D'AZUR

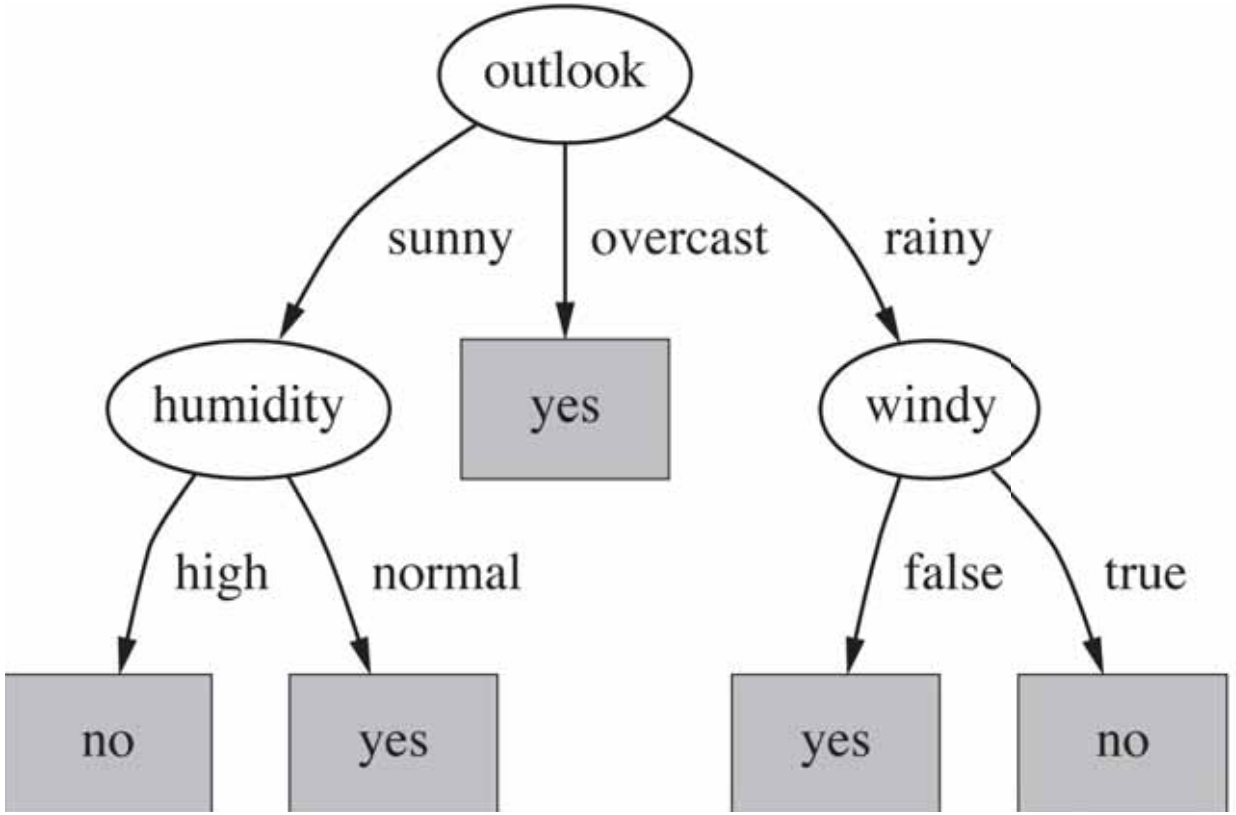
univ-cotedazur.fr

Decision tree to decide playing tennis or not



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

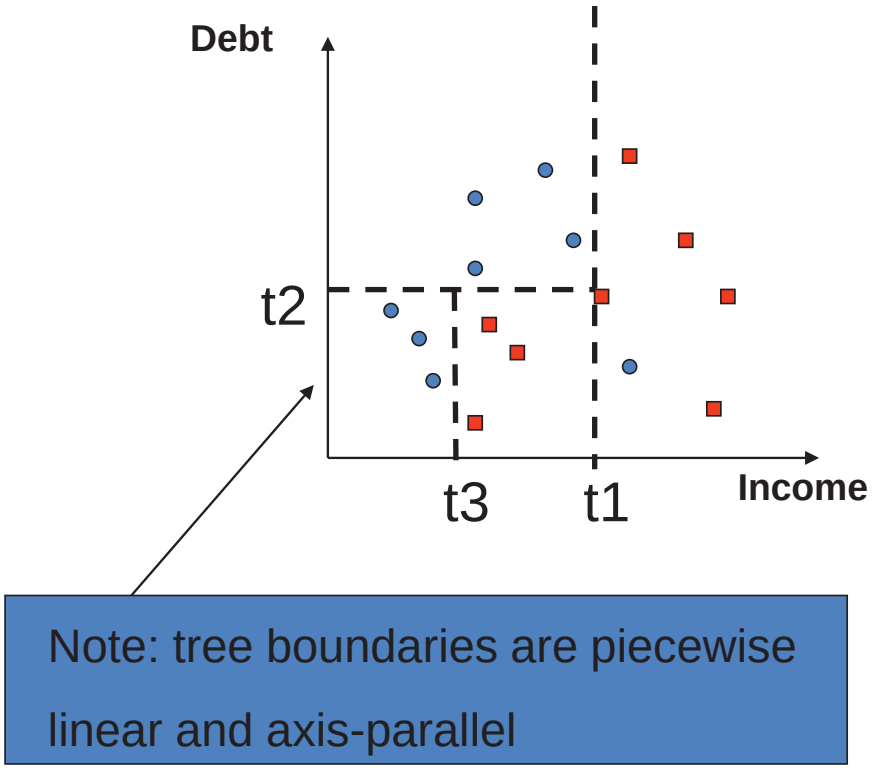
Final decision tree



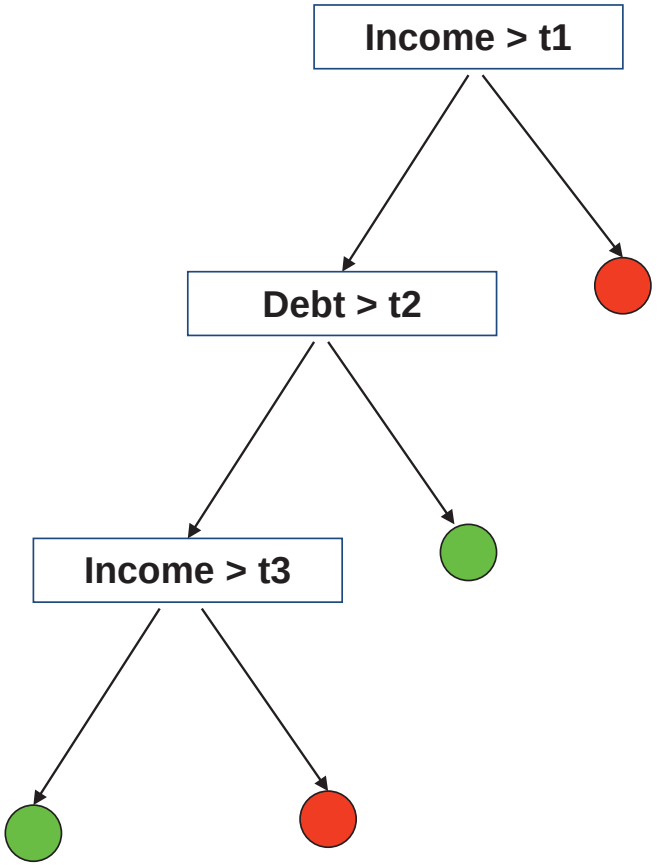
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Decision tree example



Note: tree boundaries are piecewise linear and axis-parallel



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Advantages of decision trees

- Simple and easily interpretable rules (unlike implicit decision methods)
- No need to recode heterogeneous data
- Processing with missing values
- No model and no presupposition to meet (iterative method)
- Fast processing time

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Drawbacks

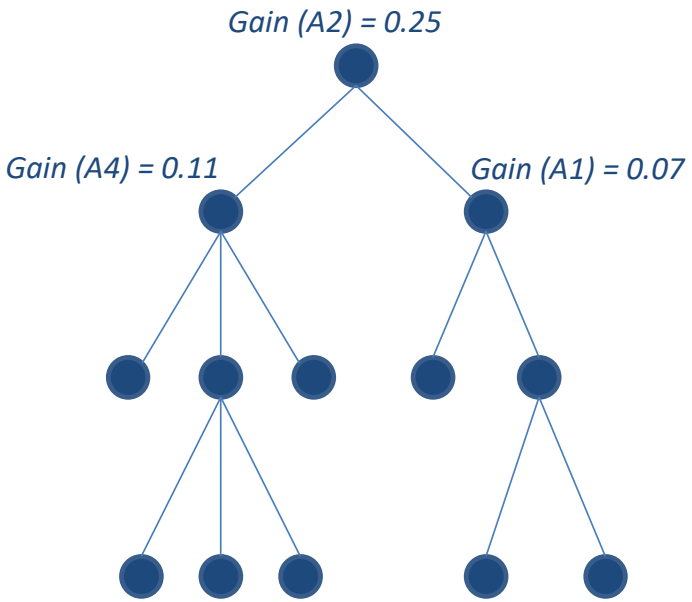
- The nodes of level n + 1 are highly dependent on those of level n (the modification of a single variable near to the top of the tree can entirely change the tree)
- We always choose the best local attributes, the best global information gain is not at all guaranteed
- Learning requires a sufficient number of individuals
- Inefficient when there are many classes

- **No convergence…**

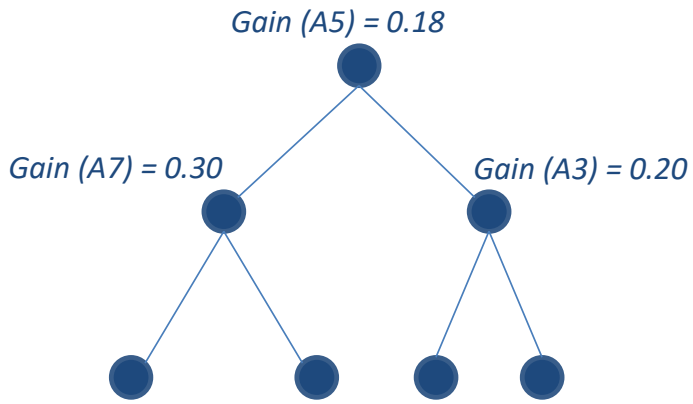
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Drawbacks

- We always chose the best local attributes, the best global information gain is not at all guaranteed



Solution chosen by the algorithm



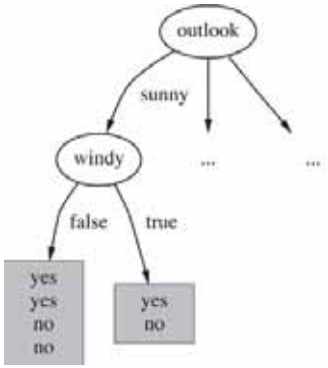
Solution that should be chosen

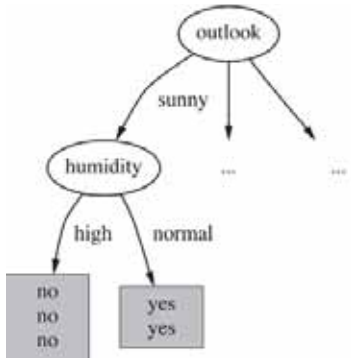
UNIVERSITÉ CÔTE D'AZUR

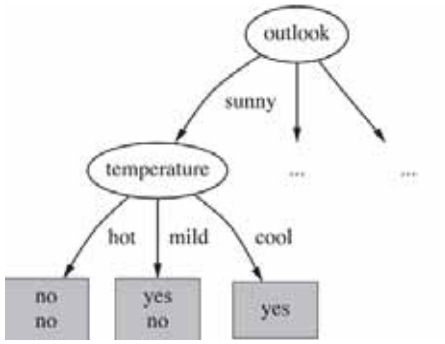
univ-cotedazur.fr

Decision trees do not converge?
"Plant" a forest

outlook	temperature	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no



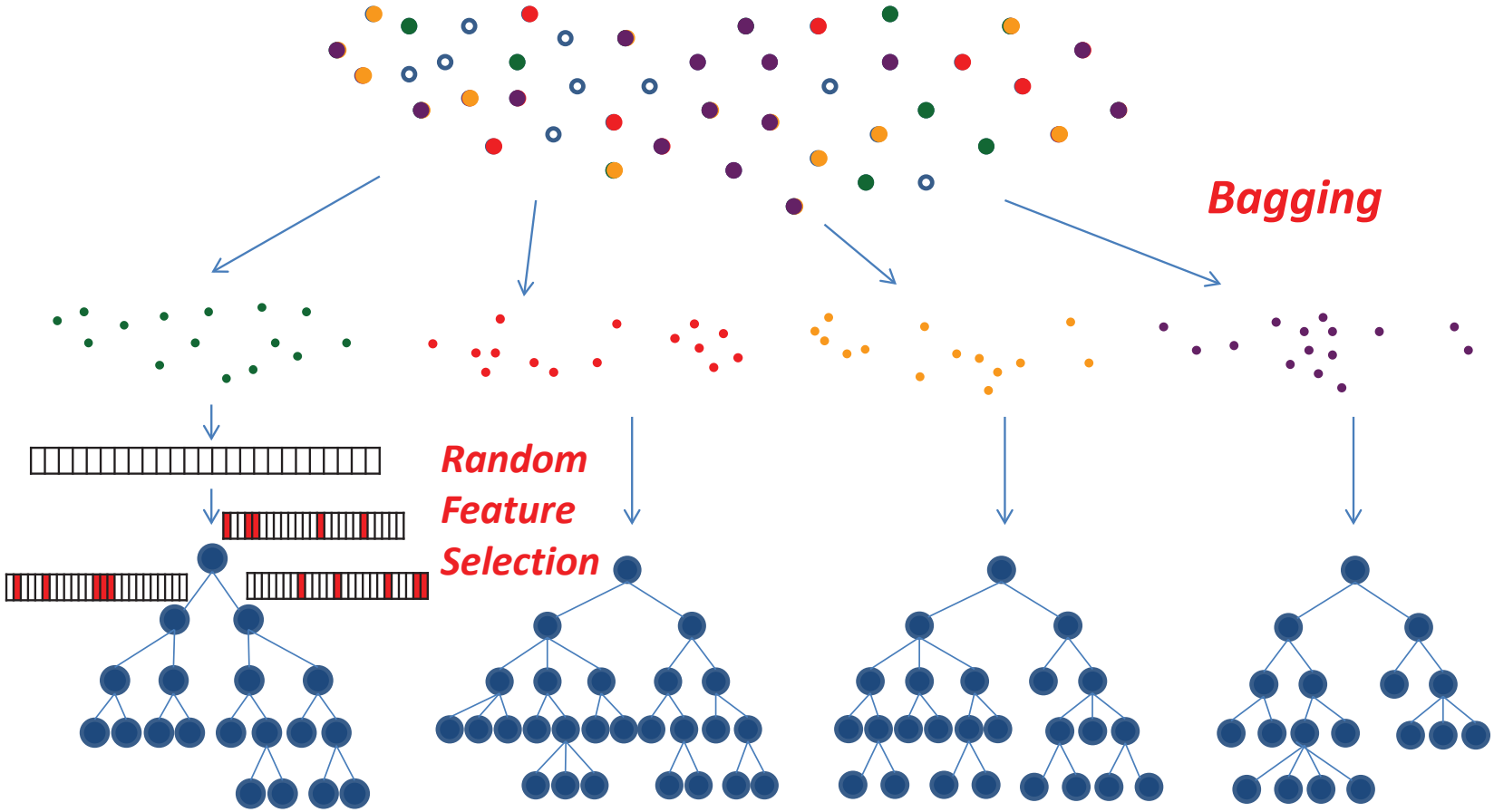




UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Standard Random Forests



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Error of generalization for Random Forest

- Error of generalization of RF can be bounded by:

$$R(RF) \leq \frac{\rho\left(1 - s^2\right)}{s^2}$$

where

- ρ is the mean correlation between two decision trees
- $\boldsymbol{s}$ is the quality of prediction of the set of decision trees

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Success story: Kinect

https://www.youtube.com/watch?v=lntbRsi8lU8

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Success story: Kinect

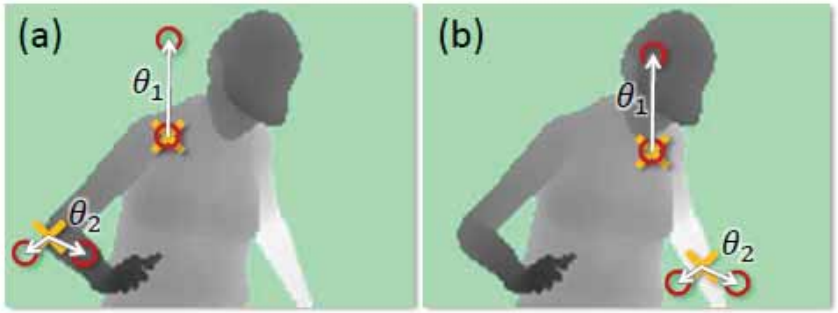


Figure 3. **Depth image features.** The yellow crosses indicates the pixel x being classified. The red circles indicate the offset pixels as defined in Eq. 1. In (a), the two example features give a large depth difference response. In (b), the same two features at new image locations give a much smaller response.

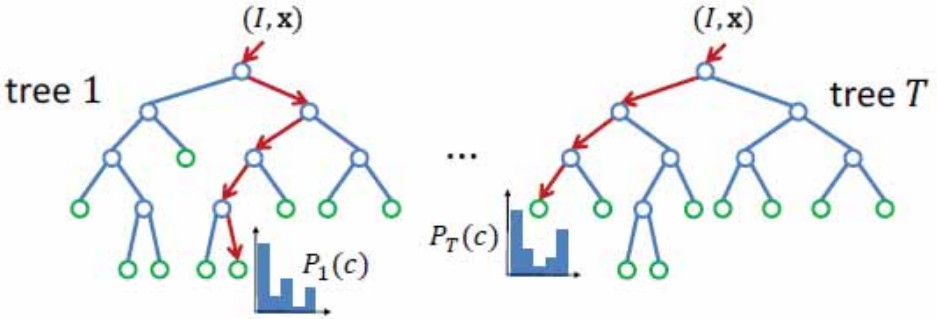
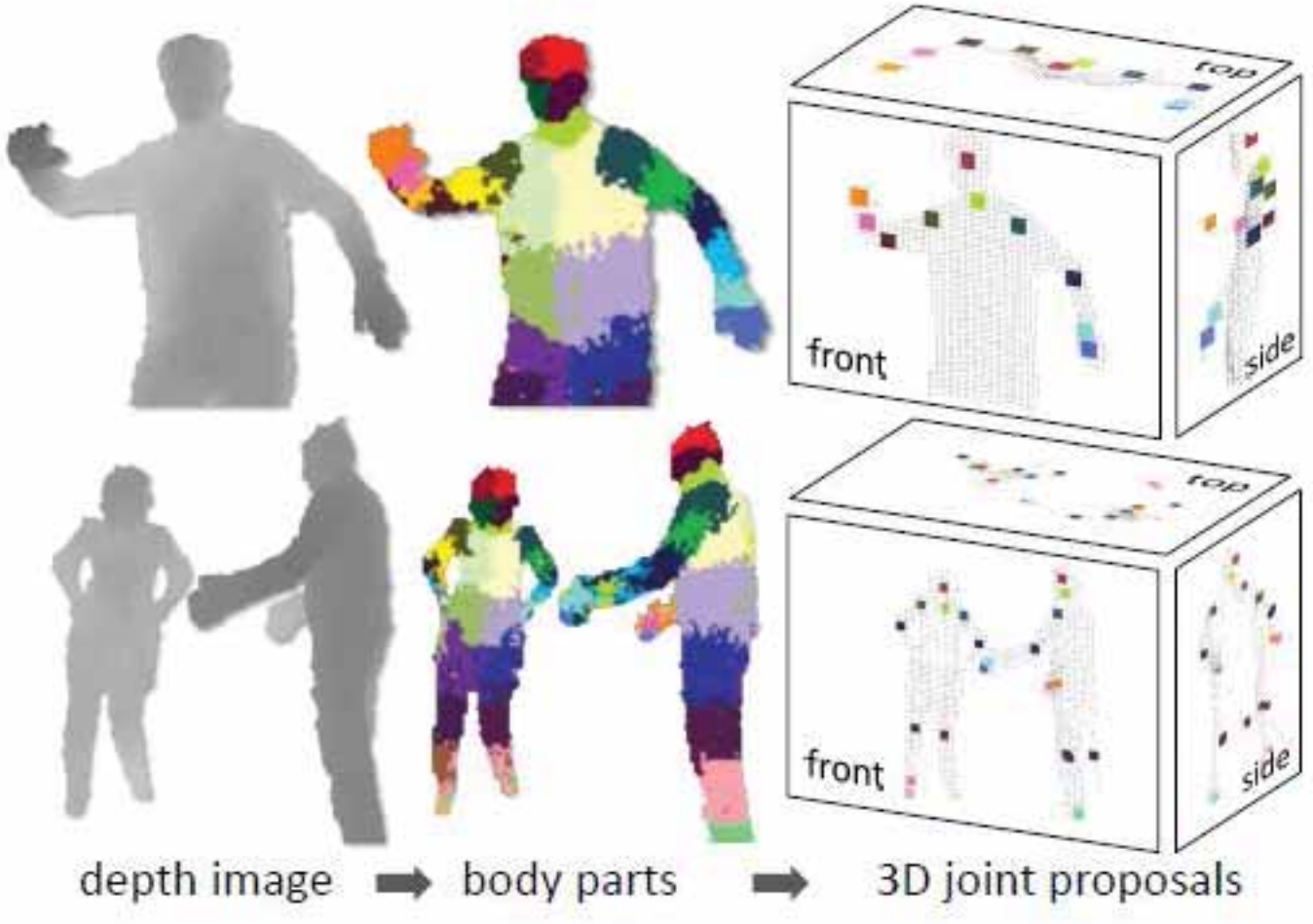


Figure 4. **Randomized Decision Forests.** A forest is an ensemble of trees. Each tree consists of split nodes (blue) and leaf nodes (green). The red arrows indicate the different paths that might be taken by different trees for a particular input.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Success story: Kinect



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Overview

- Unsupervised classification
- Explicit supervised classification
- **Implicit supervised classification**
  - **Multi-Layer Perceptron**
  - **Support Vector Machine**
- Deep Learning
- Reinforcement Learning

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

IMPLICIT SUPERVISED CLASSIFICATION

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Thomas Cover's Theorem (1965)
"The Blessing of dimensionality"

Cover's theorem states: A complex pattern-classification problem cast in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space.

(repeated sequence of Bernoulli trials)

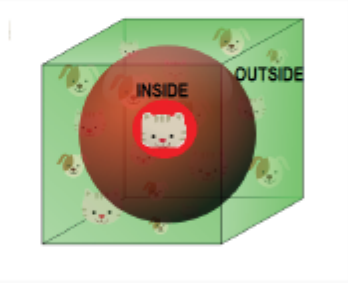
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

The curse of dimensionality [Bellman, 1956]

- Euclidian distance is not relevant in high dimension: d $\geq$10
  1. look at the examples at distance at most r
  2. the hypersphere volume is too small: practically empty of examples

$$\frac{volume\ of\ the\ sphere\ of\ radial\ r}{hypersphere\ of\ 2r\ width} \rightarrow_{d\rightarrow\infty} 0$$



  3. need a number of examples exponential in d

Remark

Specific care for data representation

UNIVERSITÉ CÔTE D'AZUR

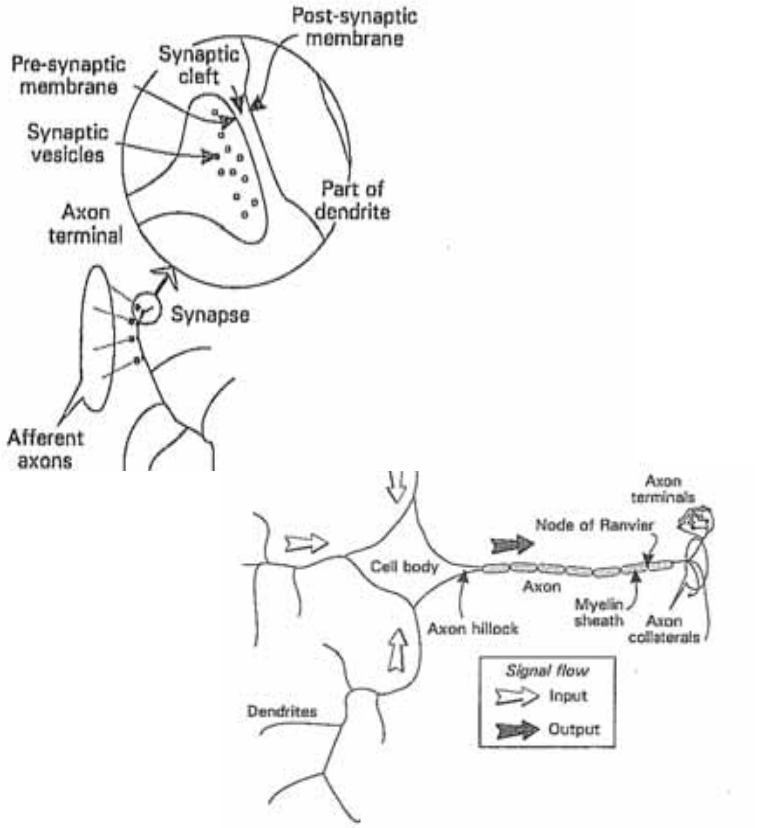
MULTI-LAYER PERCEPTRON

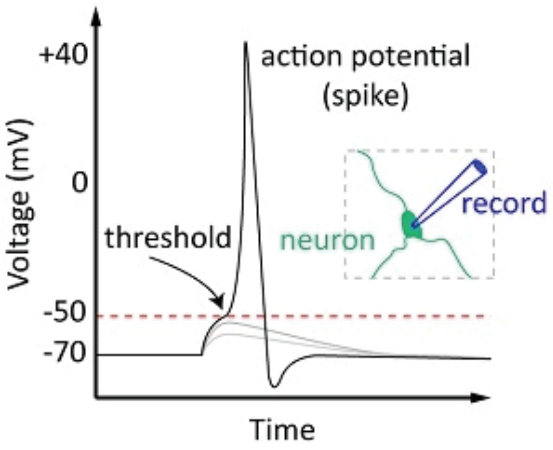
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

First, biological neurons

- Before we study artificial neurons, let's look at a biological neuron





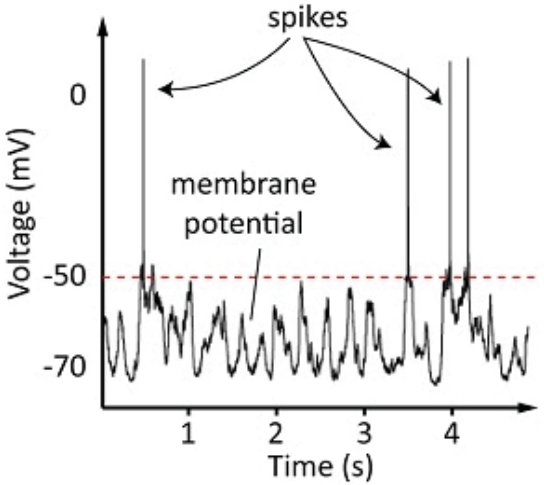
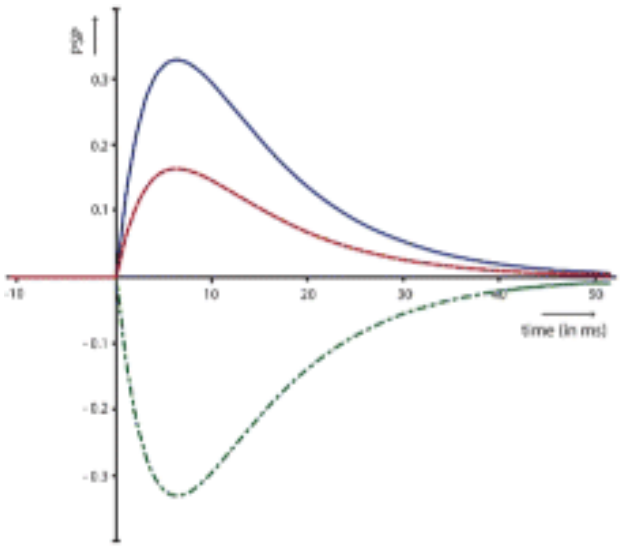


Figure from K.Gurney, An Introduction to Neural Networks

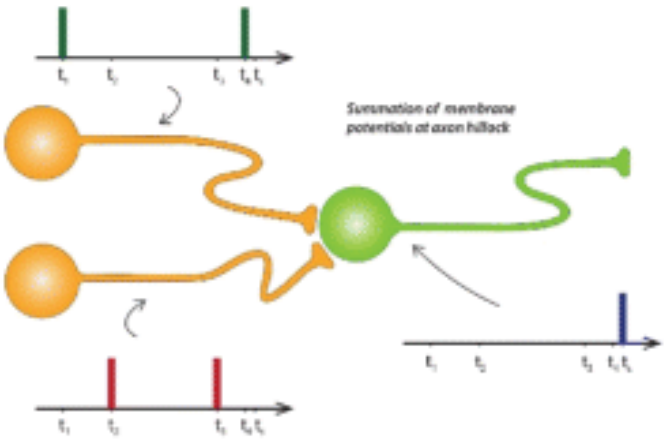
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

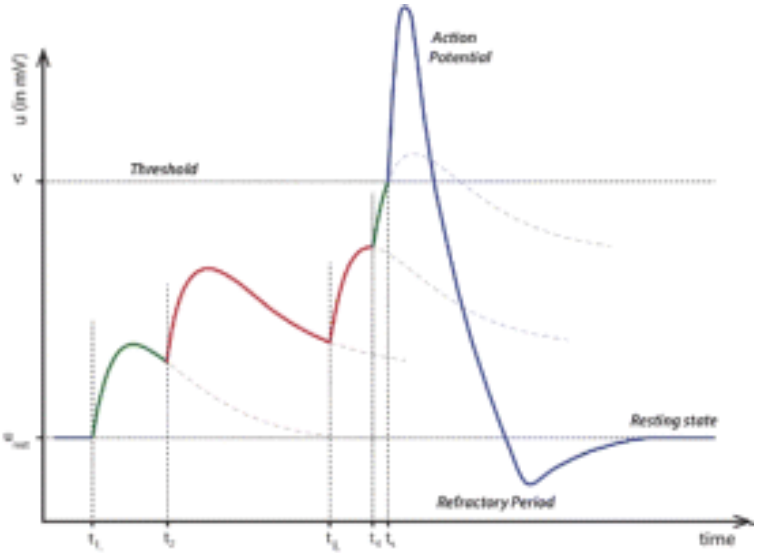
First, biological neurons



A



B

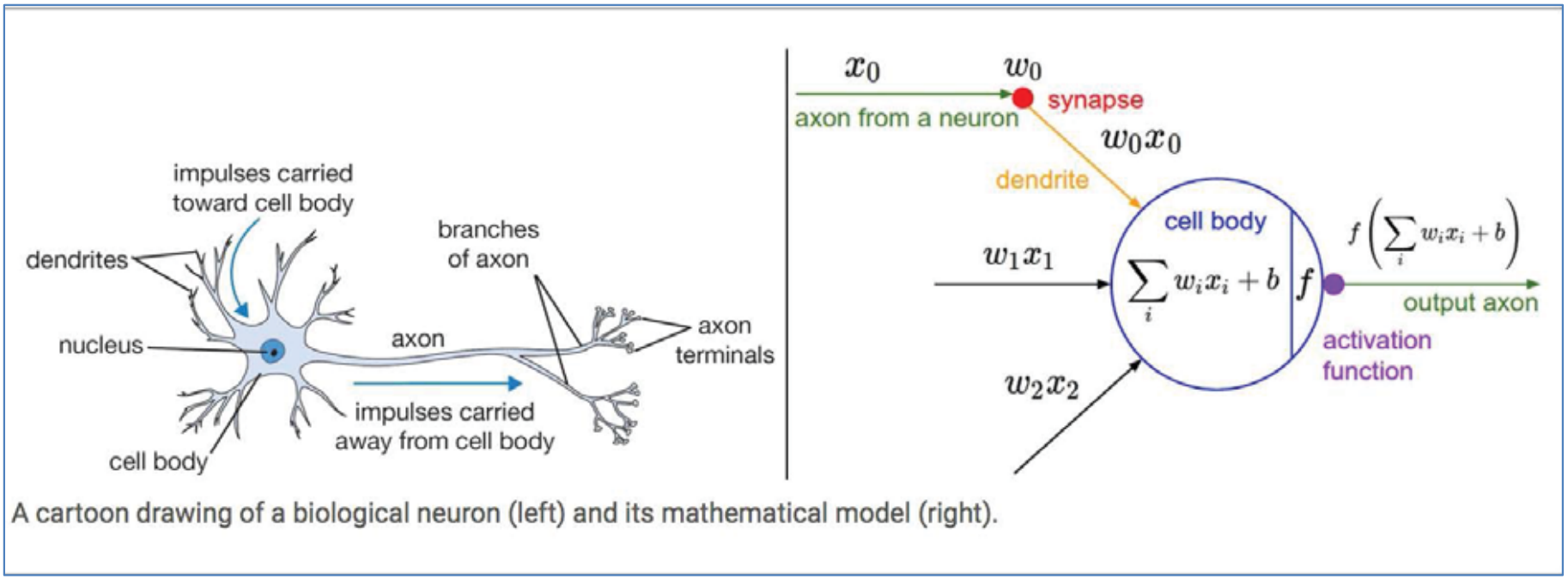


C

Postsynaptic potential function with weight dependency, as a function of time (ms) and weight value, being excitatory in case of red and blue lines, and inhibitory in case of a green line.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Then, artificial neurons



A cartoon drawing of a biological neuron (left) and its mathematical model (right).

Pitts & McCulloch (1943), binary inputs & activation function *f* is a thresholding

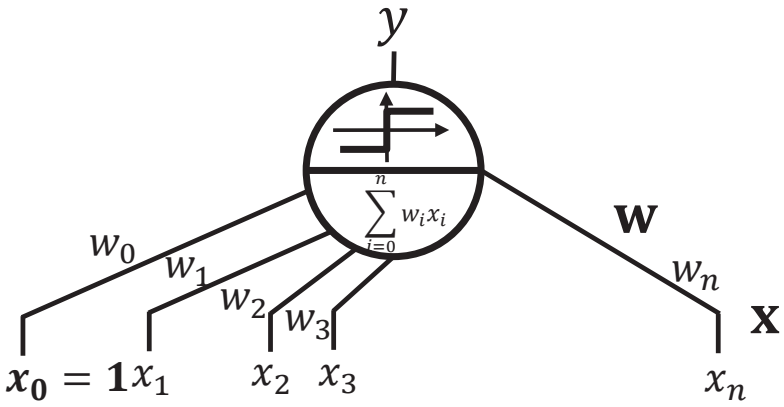
Rosenblatt (1956), real inputs & activation function *f* is a thresholding

(Schéma : Isaac Changhau)

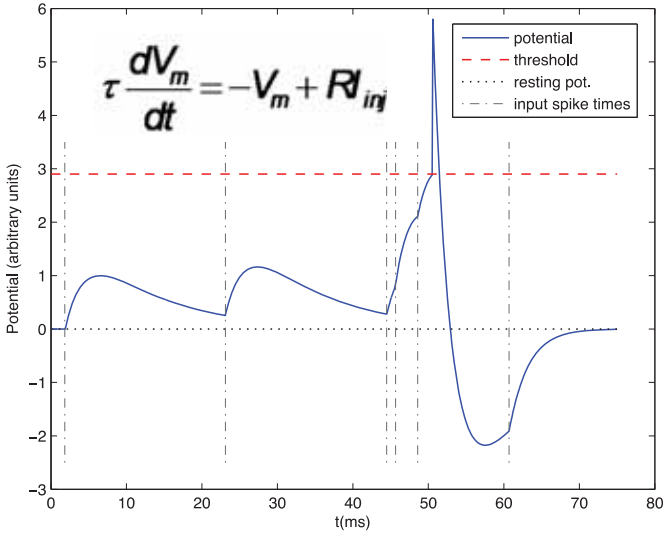
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Artificial neuron vs biology



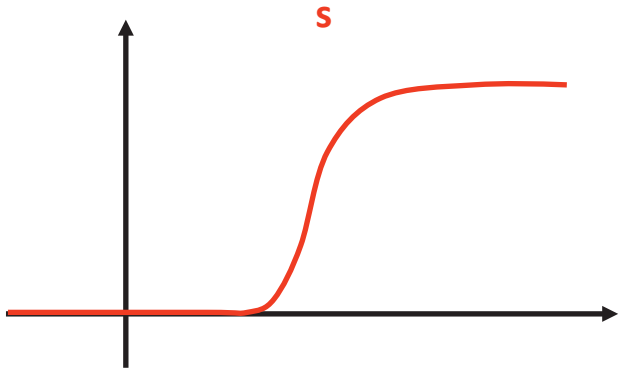
Spike-based description



Gradient descent: KO

Rate-based description

Steady regime



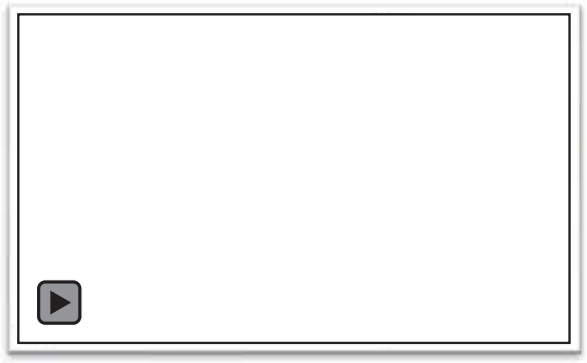
$y = s(\Sigma w_i x_i)$

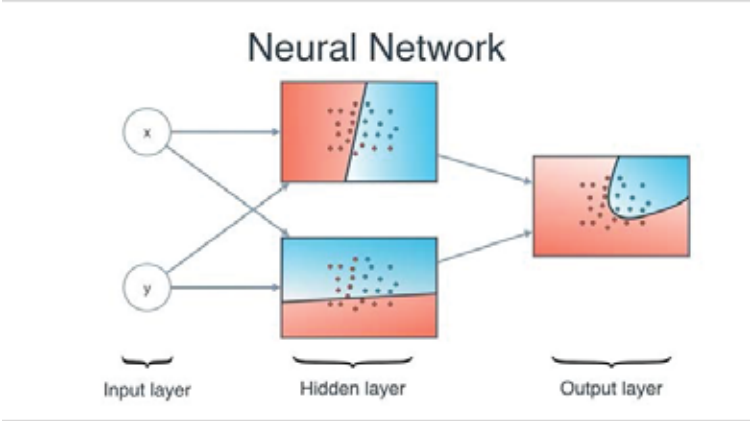
Gradient descent: OK

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

From perceptron to network



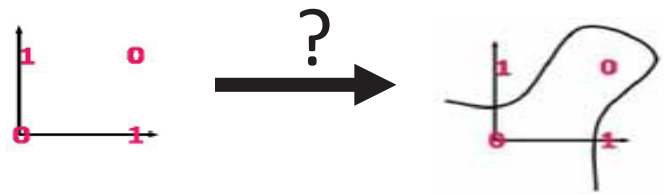


@tachyeonz: A friendly introduction to neural networks and deep learning.

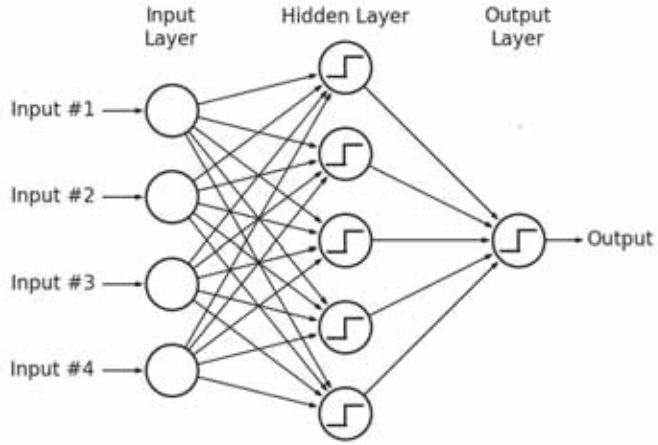
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Single Perceptron Unit

- **Perceptron** only learns linear function [Minsky and Papert, 1969]



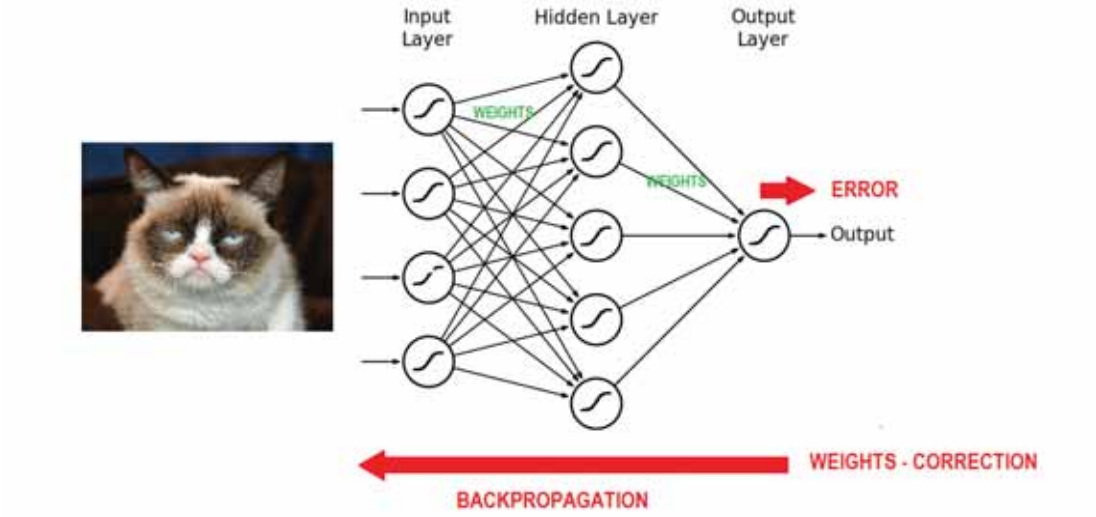
- Non-linear function needs layer(s) of neurons → Neural Network
- Neural Network = input layer + hidden layer(s) + output layer



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Multi-Layer Perceptron

- Training a neural network [Rumelhart et al. / Yann Le Cun et al. 1985]
- **Unknown parameters: weights on the synapses**
- Minimizing a cost function: some metric between the predicted output and the given output

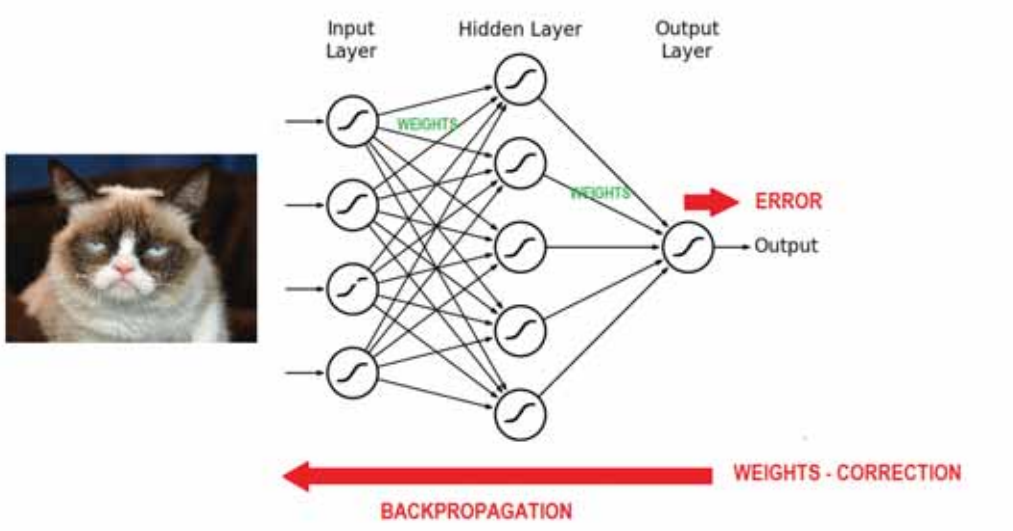


- Step function: non-continuous functions are replaced by a continuous non-linear ones

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Multi-Layer Perceptron

- Minimizing a cost function: some metric between the predicted output and the given output



- Equation for a network of 3 neurons (i.e. 3 perceptrons):

$$y = s(w_{13}s(w_{11}x_1 + w_{21}x_2 + w_{01}) + w_{23}s(w_{12}x_1 + w_{22}x_2 + w_{02}) + w_{03})$$

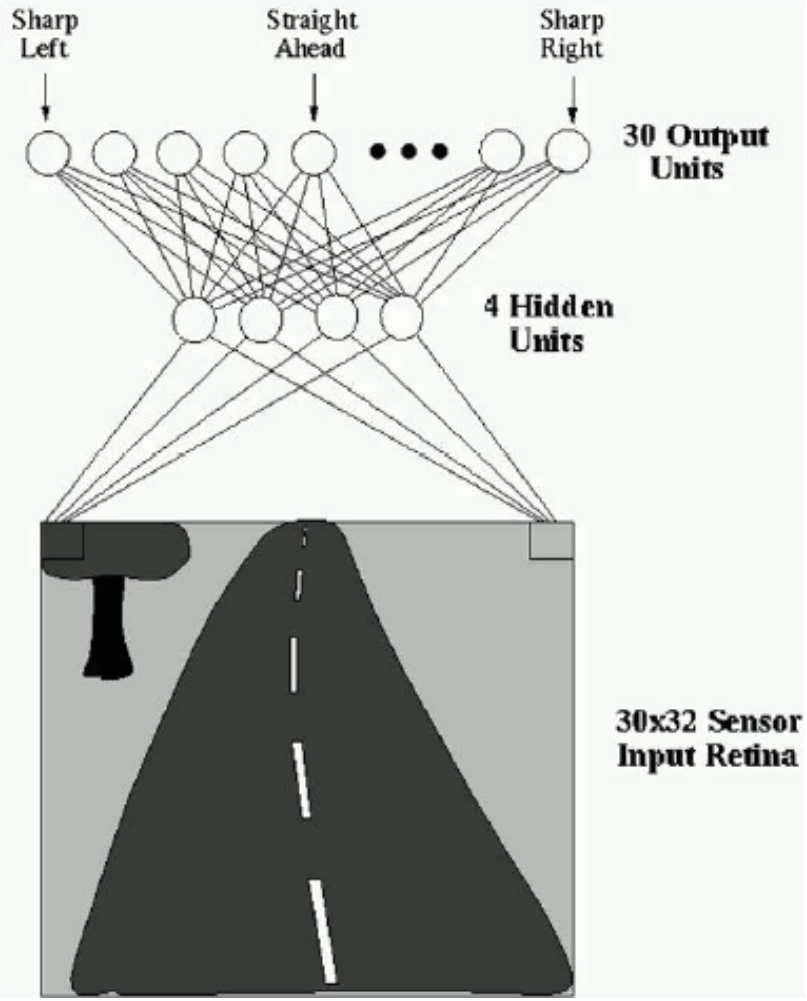
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Autonomous Land Vehicle In a Neural Network (ALVINN)

- ALVINN is an automatic steering system for a car based on input from a camera mounted on the vehicle.
  - Successfully demonstrated in a cross-country trip.

ALVINN (1989)

- The ALVINN neural network is:
  - 960 inputs (a 30x32 array derived from the pixels of an image),
  - 4 hidden units and
  - 30 output units (each representing a steering command).



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Multi-Layer Perceptron

Theorem [Cybenko, 1989]

- A neural network with one single hidden layer is **a universal approximator**: it can represent any continuous function on compact subsets of $\mathbf{R}^n$
- 2 layers is enough ... theoretically:

  "…networks with one internal layer and an arbitrary continuous sigmoidal function can approximate continuous functions with arbitrary precision providing that no constraints are placed on the number of nodes or the size of the weights"
- **But *no efficient learning rule* is known and the size of the hidden layer is *exponential* with the complexity of the problem** (which is unkown beforehand) **to get an error ε , the layer must be infinite for an error 0.**

UNIVERSITÉ CÔTE D'AZUR

SUPPORT VECTOR MACHINE

Partly based on "A Gentle Introduction to Support Vector Machines in Biomedicine", A. Statnikov, D. Hardin, I. Guyon, C. F. Aliferis, AMIA 2010.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Thomas Cover's Theorem (1965)
"The Blessing of dimensionality"

Cover's theorem states: A complex pattern-classification problem cast in a high-dimensional space nonlinearly is more likely to be linearly separable than in a low-dimensional space.

(repeated sequence of Bernoulli trials)

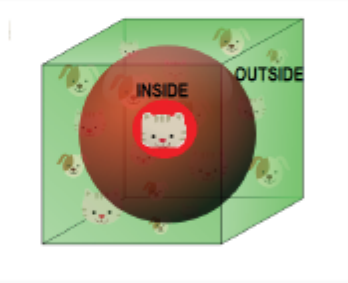
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

The curse of dimensionality [Bellman, 1956]

- Euclidian distance is not relevant in high dimension: d $\geq$10
  1. look at the examples at distance at most r
  2. the hypersphere volume is too small: practically empty of examples

$$\frac{volume\ of\ the\ sphere\ of\ radial\ r}{hypersphere\ of\ 2r\ width} \rightarrow_{d\rightarrow\infty} 0$$



  3. need a number of examples exponential in d

Remark

Specific care for data representation

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

SVM vs ANN

"*SVMs have been developed in the reverse order to the development of neural networks (NNs). SVMs evolved from the sound theory to the implementation and experiments, while the NNs followed more heuristic path, from applications and extensive experimentation to the theory.*"

"Support Vector Machines: Theory and Applications" by *Lipo Wang*, in Studies in Fuzziness and Soft Computing, Springer, 2005.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

The Support Vector Machine (SVM)

- Support vector machines (SVMs) is a binary classification algorithm.
- Extensions of the basic SVM algorithm can be applied to solve problems of regression, feature selection, novelty/outlier detection, and clustering.
- SVMs are important because of (a) theoretical reasons:
  - **Robust to very large number of variables and small samples**
  - **Can learn both simple and highly complex classification models**
  - **Employ sophisticated mathematical principles to avoid overfitting**

  and (b) superior empirical results.

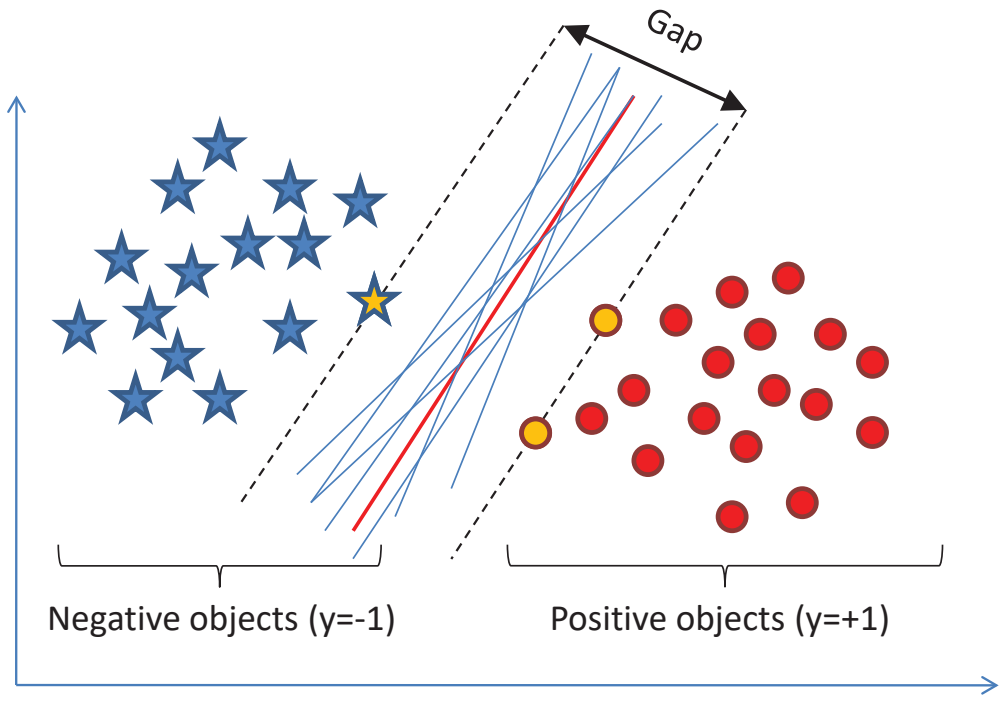
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Linearly separable data, "Hard-margin" linear SVM

Given training data: $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N \in R^n$

$y_1, y_2, \ldots, y_N \in \{-1, +1\}$



- Want to find a classifier (hyperplane) to separate negative objects from the positive ones.
- An infinite number of such hyperplanes exist.
- SVMs finds the hyperplane that maximizes the gap between data points on the boundaries (so-called "support vectors").
- If the points on the boundaries are not informative (e.g., due to noise), SVMs may not do well.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Kernel Trick

https://www.youtube.com/watch?v=-Z4aojJ-pdg

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Popular kernels

A kernel is a dot product in *some* feature space:

$$K(\vec{x}_i, \vec{x}_j) = \Phi(\vec{x}_i) \cdot \Phi(\vec{x}_j)$$

Examples:

$K(\vec{x}_i, \vec{x}_j) = \vec{x}_i \cdot \vec{x}_j$	Linear kernel
$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2)$	***Gaussian kernel***
$K(\vec{x}_i, \vec{x}_j) = \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|)$	Exponential kernel
$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q$	Polynomial kernel
$K(\vec{x}_i, \vec{x}_j) = (p + \vec{x}_i \cdot \vec{x}_j)^q \exp(-\gamma \left\| \vec{x}_i - \vec{x}_j \right\|^2)$	Hybrid kernel
$K(\vec{x}_i, \vec{x}_j) = \tanh(k\vec{x}_i \cdot \vec{x}_j - \delta)$	Sigmoidal

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

How to build a kernel function ?

$$k(\mathbf{x},\mathbf{y}) = k_1(\mathbf{x},\mathbf{y}) + k_2(\mathbf{x},\mathbf{y})$$

$$k(\mathbf{x},\mathbf{y}) = \alpha \cdot k_1(\mathbf{x},\mathbf{y})$$

$$k(\mathbf{x},\mathbf{y}) = k_1(\mathbf{x},\mathbf{y}) \cdot k_2(\mathbf{x},\mathbf{y})$$

$$k(\mathbf{x},\mathbf{y}) = f(\mathbf{x}) \cdot f(\mathbf{y})$$

with $f()$ a function from input space to $\mathbb{R}$

$$k(\mathbf{x},\mathbf{y}) = k_3(\Phi(\mathbf{x}),\Phi(\mathbf{y}))$$

$$k(\mathbf{x},\mathbf{y}) = \mathbf{x}\mathbf{B}\mathbf{y}^T$$

with $\mathbf{B}$ a matrix $N \times N$ symetric, semi-definite positive

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Complex kernels on Video Tubes

Video object extraction and description

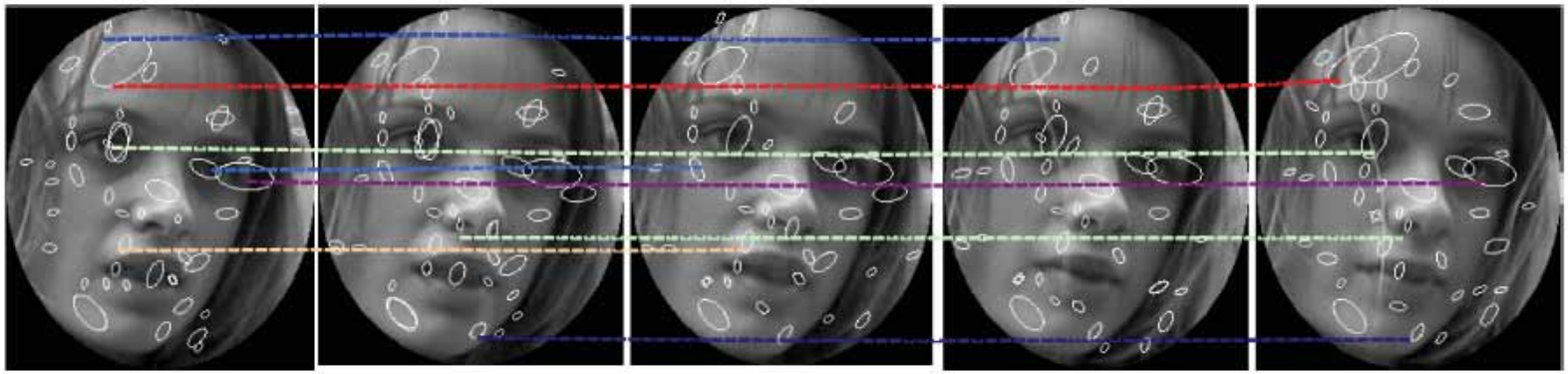
- RoI = face tubes
  - Frame face detection
  - Face region grouping in shots

Example of a tube:

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Complex kernels on Video Tubes



The major kernel on tubes is then defined as:

$$K'_{pow}(T_i, T_j) = \left(\sum_r \sum_s \frac{|C_{ri}|}{\sqrt{|T_i|}} \frac{|C_{sj}|}{\sqrt{|T_j|}} k'(C_{ri}, C_{sj})^q \right)^{\frac{1}{q}} \qquad (1)$$

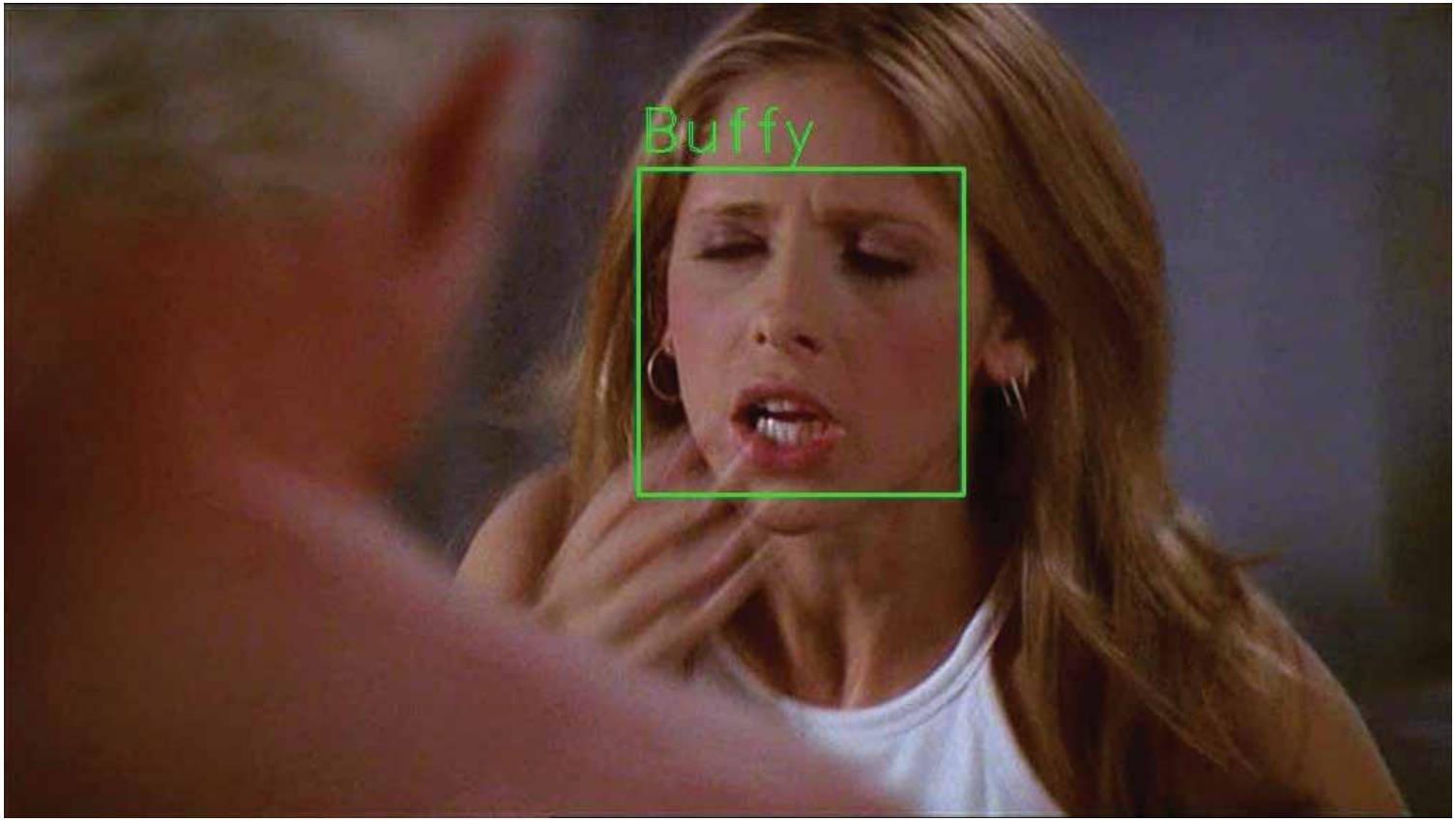
with the following minor kernel on chains:

$$k'(C_{ri}, C_{sj}) = exp\left(-\frac{1}{2\sigma^2} \chi^2 \left(\overline{C}_{ri}, \overline{C}_{sj} \right) \right) e^{-\frac{\left(\overline{x}_{ri} - \overline{x}_{sj}\right)^2 + \left(\overline{y}_{ri} - \overline{y}_{sj}\right)^2}{2\sigma_2^2}}$$

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

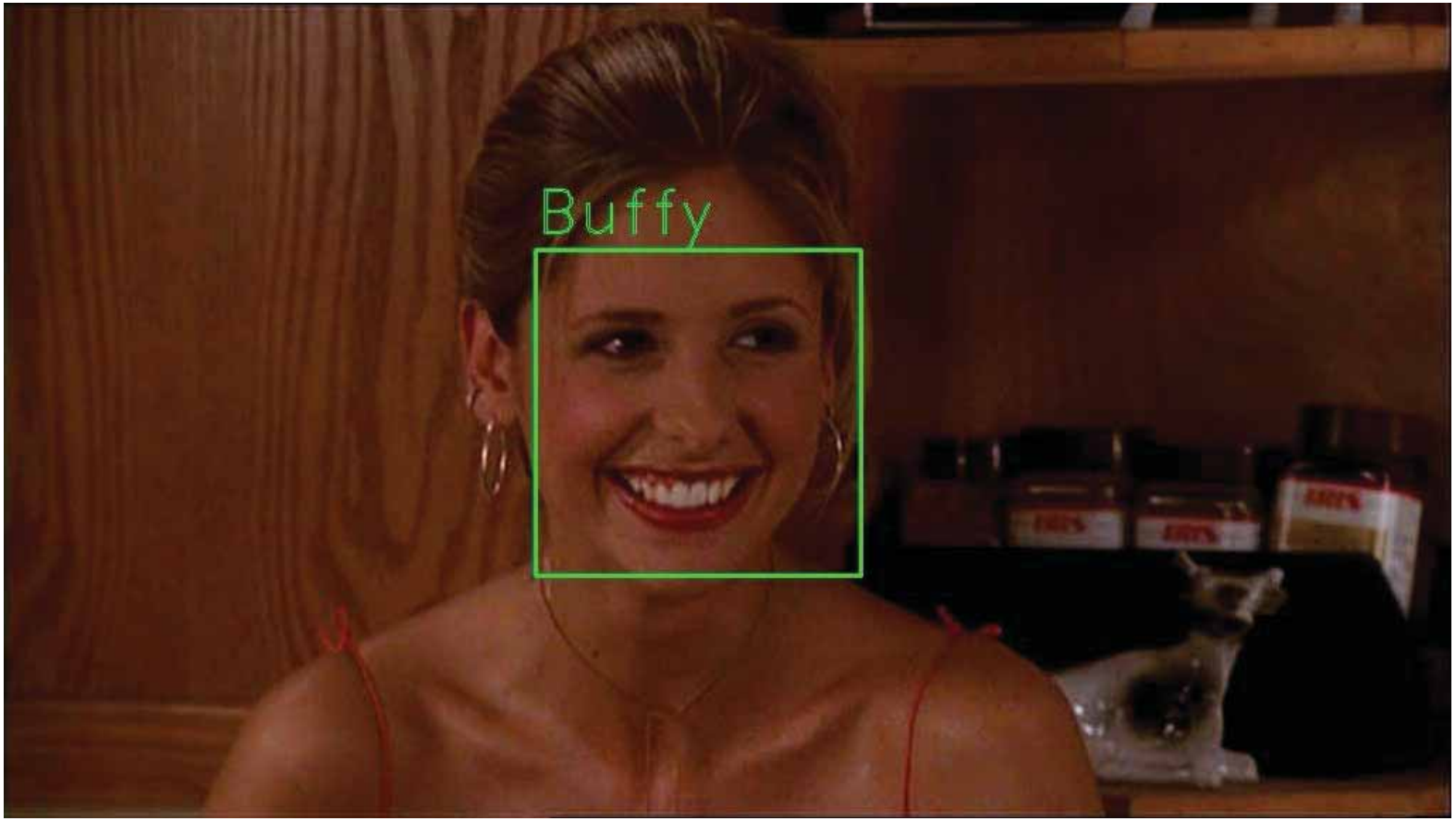
Classification



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

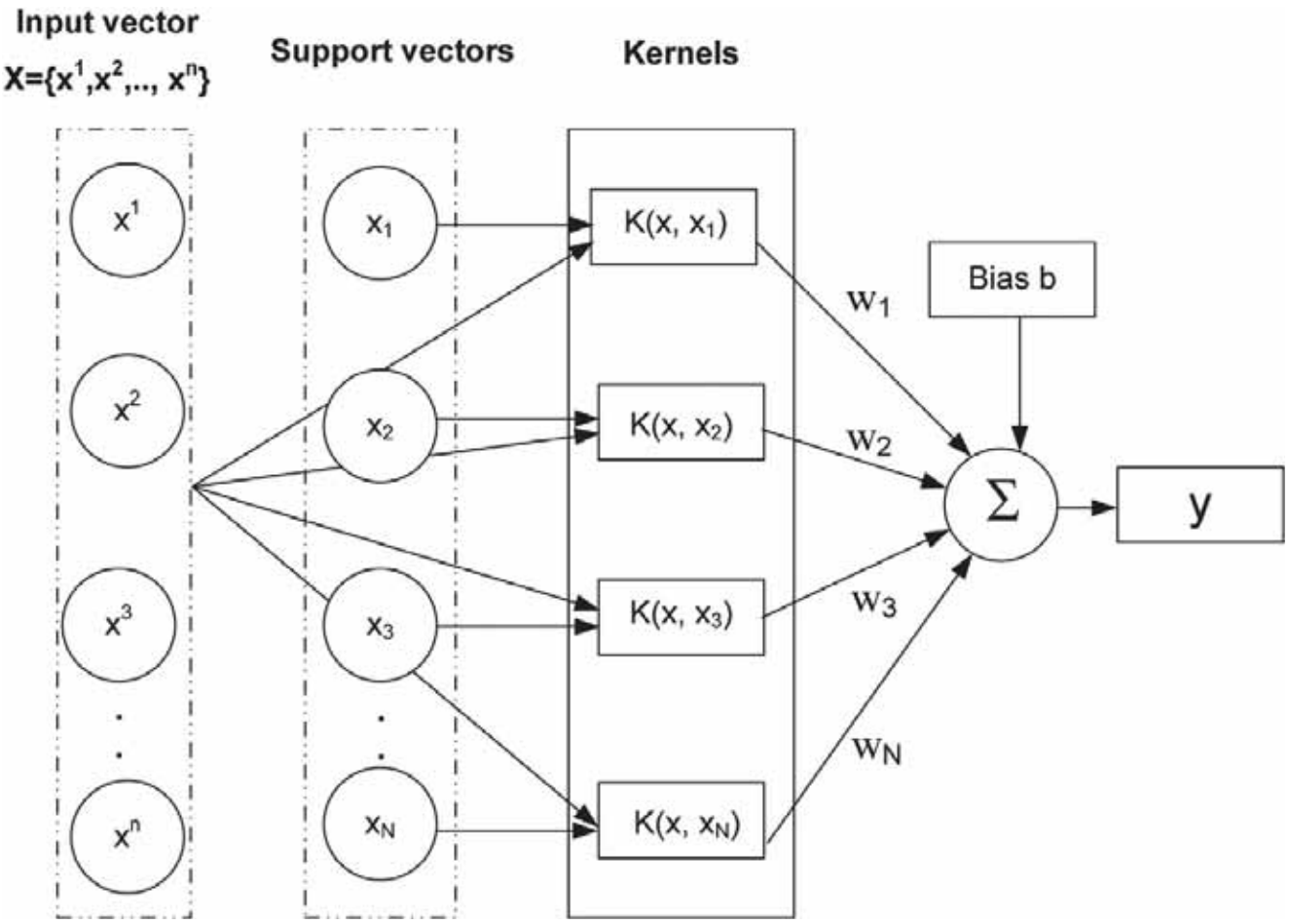
Classification



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

SVM are ANN



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Overview

- Unsupervised classification
- Explicit supervised classification
- Implicit supervised classification
- **Deep Learning**
  - **Convolutional Neural Networks (CNN)**
  - **Generative Adversarial Networks (GAN)**
  - **Stacked Denoising AutoEncoder (SDAE)**
  - **Reccurent Neural Networks (RNN)**
- Reinforcement Learning

UNIVERSITÉ CÔTE D'AZUR

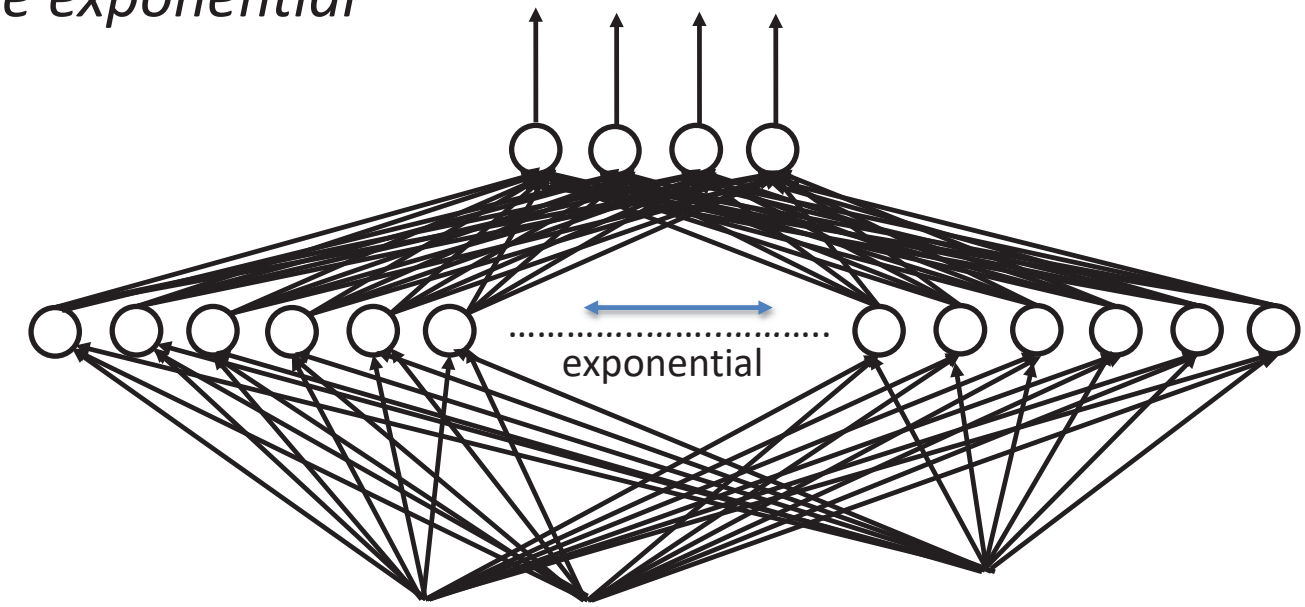
univ-cotedazur.fr

DEEP LEARNING

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deep representation origins

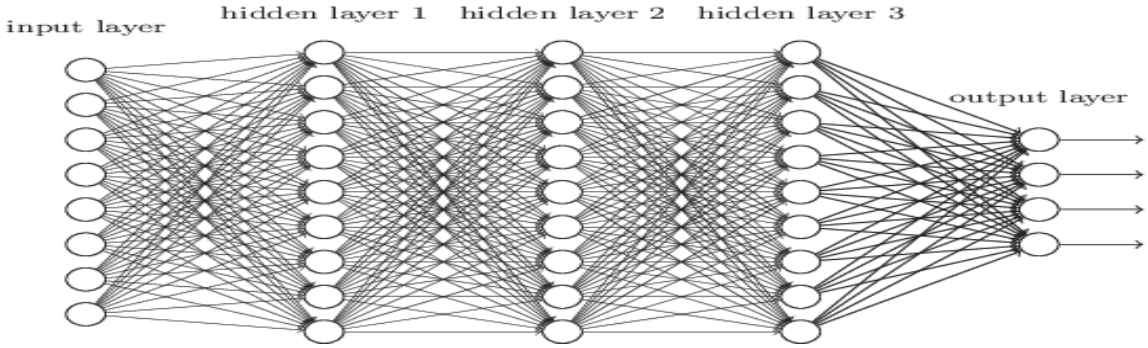
- **Theorem Cybenko** (1989) *A neural network with one single hidden layer is a universal "approximator", it can represent any continuous function on compact subsets of* $\boldsymbol{R}^n \Rightarrow$ *2 layers are enough…but hidden layer size may be exponential*



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deep representation origins

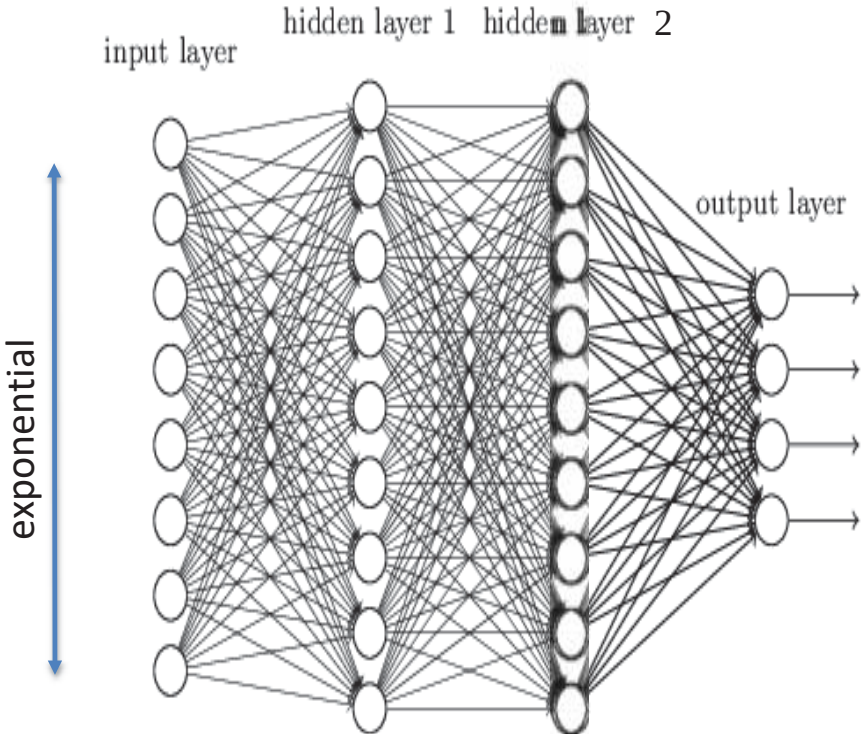
- **Theorem Hastad** (1986), **Bengio** et al. (2007) Functions representable compactly with *k* layers may require exponentially size with *k-1* layers



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deep representation origins

- **Theorem Hastad** (1986), **Bengio** et al. (2007) Functions representable compactly with *k* layers may require exponentially size with *k-1* layers



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Enabling factors

- Why do it now ? Before 2006, training deep networks was unsuccessful because of practical aspects
  - faster CPU's
  - parallel CPU architectures
  - advent of GPU computing

- Hinton, Osindero & Teh « A Fast Learning Algorithm for Deep Belief Nets », *Neural Computation*, 2006
- Bengio, Lamblin, Popovici, Larochelle « Greedy Layer-Wise Training of Deep Networks », *NIPS'2006*
- Ranzato, Poultney, Chopra, LeCun « Efficient Learning of Sparse Representations with an Energy-Based Model », *NIPS'2006*

- Results…
  - 2009, sound, interspeech + ~24%
  - 2011, text, + ~15% without linguistic at all
  - 2012, images, ImageNet + ~20%

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Structure the network?

- Can we put any structure reducing the space of exploration and providing useful properties (invariance, robustness…)?

$$y = s(w_{13}s(w_{11}x_1 + w_{21}x_2 + w_{01}) + w_{23}s(w_{12}x_1 + w_{22}x_2 + w_{02}) + w_{03})$$

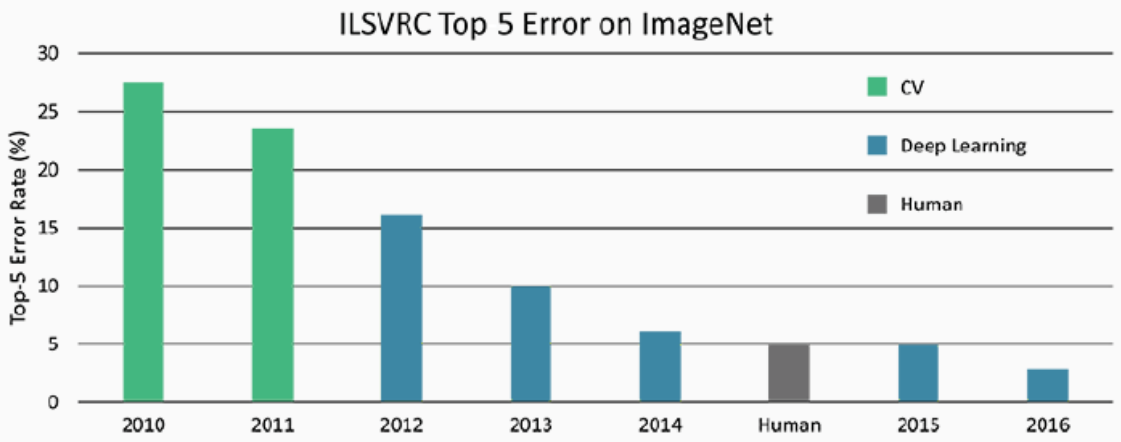
UNIVERSITÉ CÔTE D'AZUR

CONVOLUTIONAL NEURAL NETWORKS (AKA CNN, CONVNET)

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Convolutional neural network

- Deep Networks are as good as humans at recognition, identification...

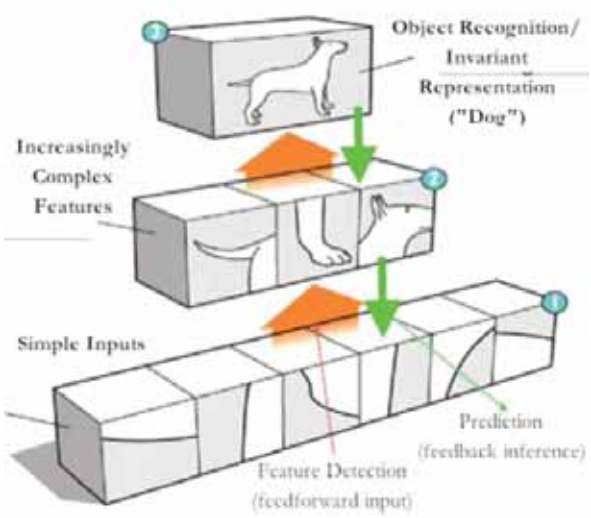


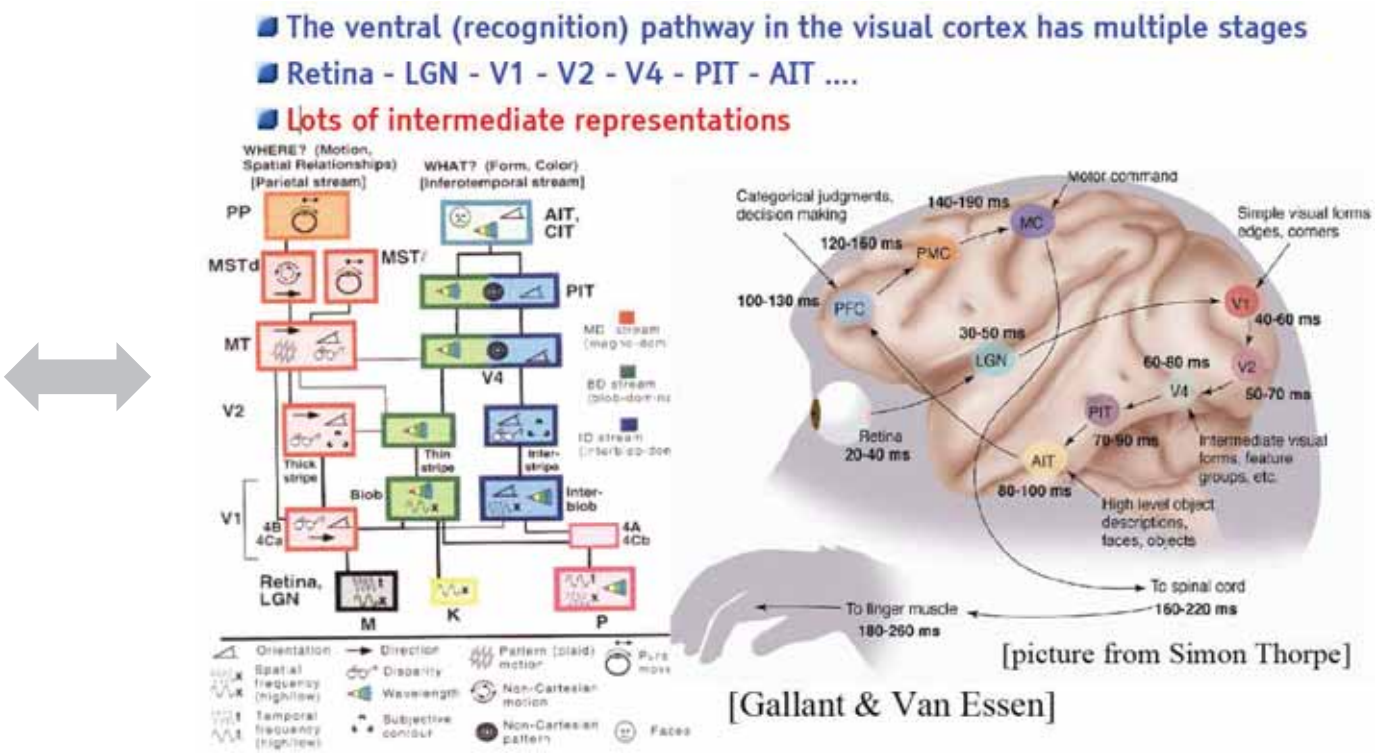
How much does a deep network understands those tasks?

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Deep representation by CNN

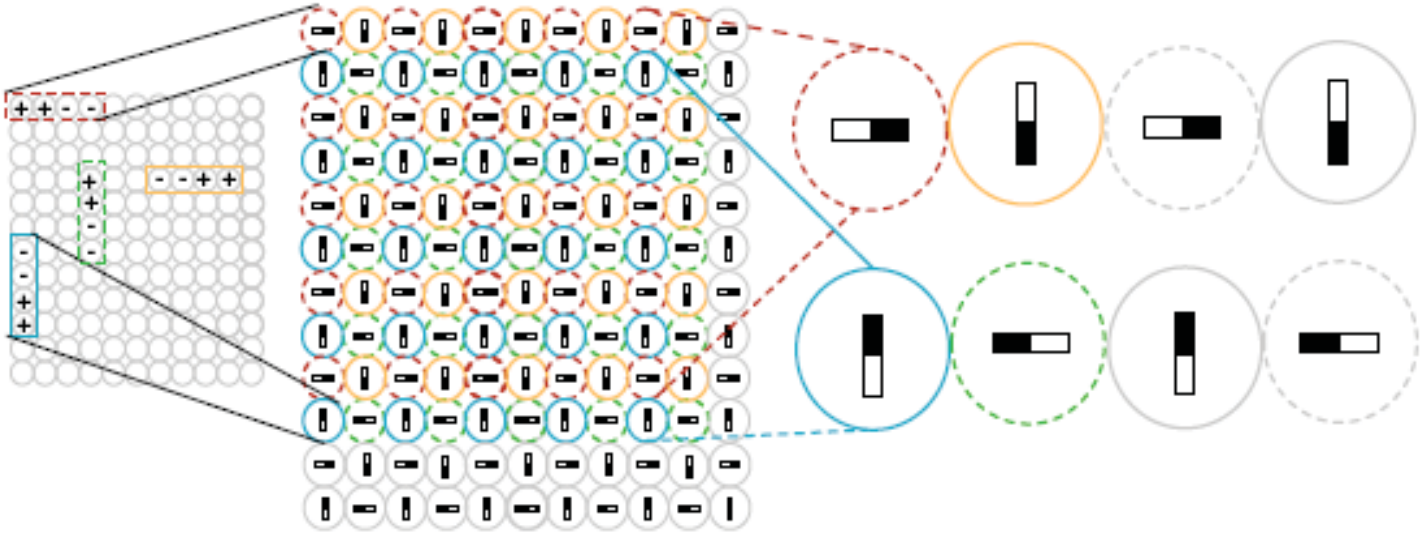




UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Convolution in nature



Octopus Retina Neurons

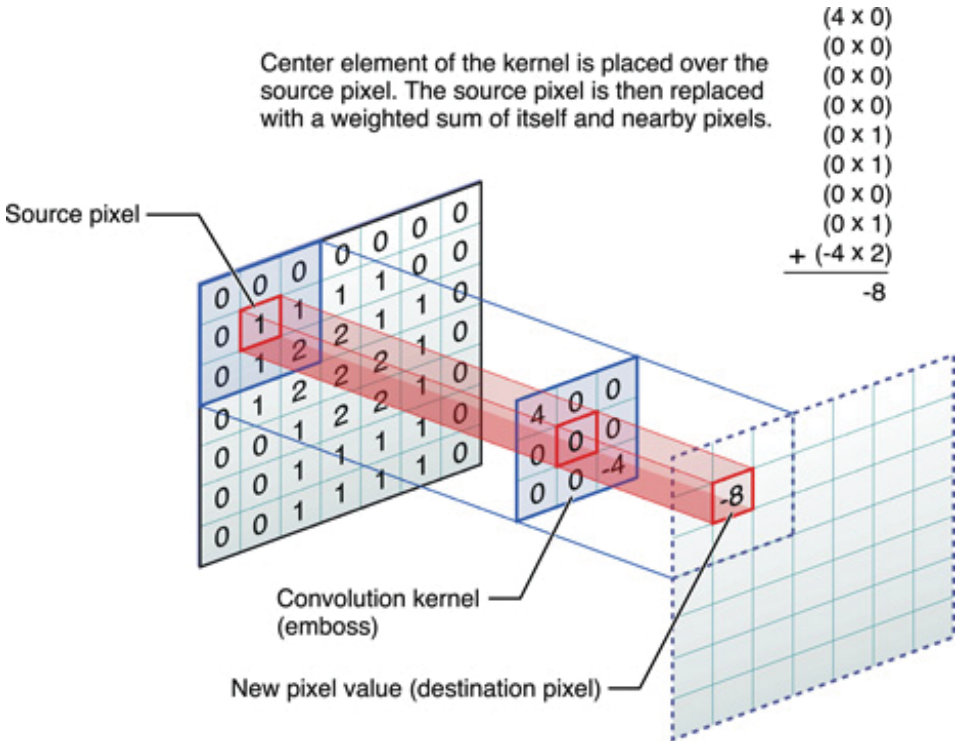
Simple Cells (IF Chip 0)

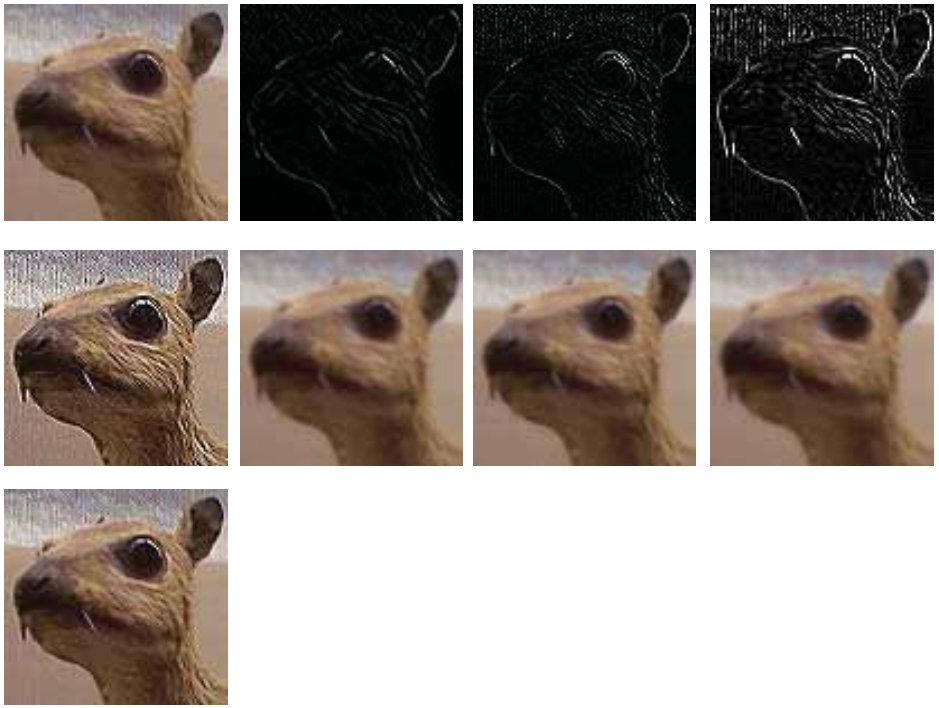
Complex Cells (IF Chip 1)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Convolution

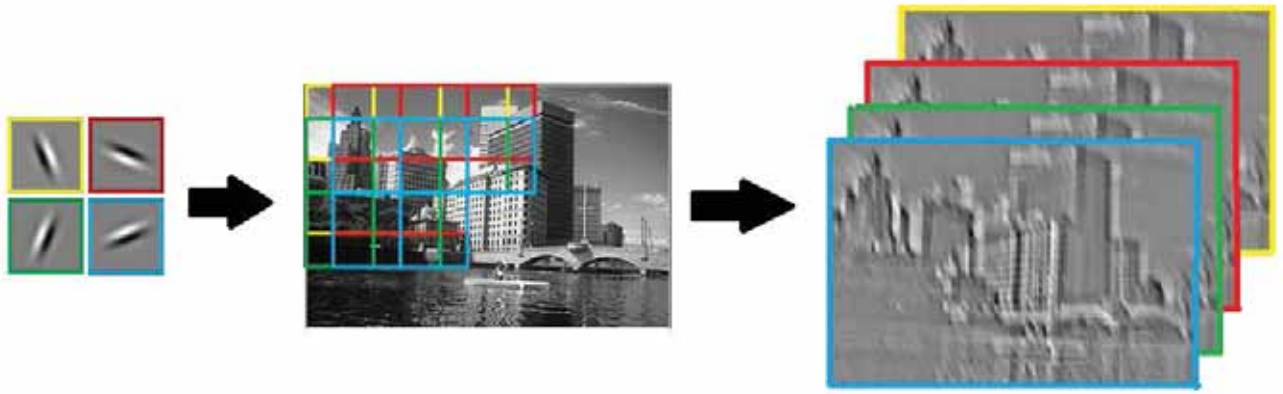




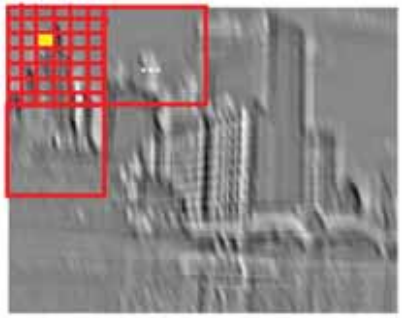
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Convolution in nature

1. Hubel and Wiesel have worked on visual cortex of cats (1962)
2. Convolution



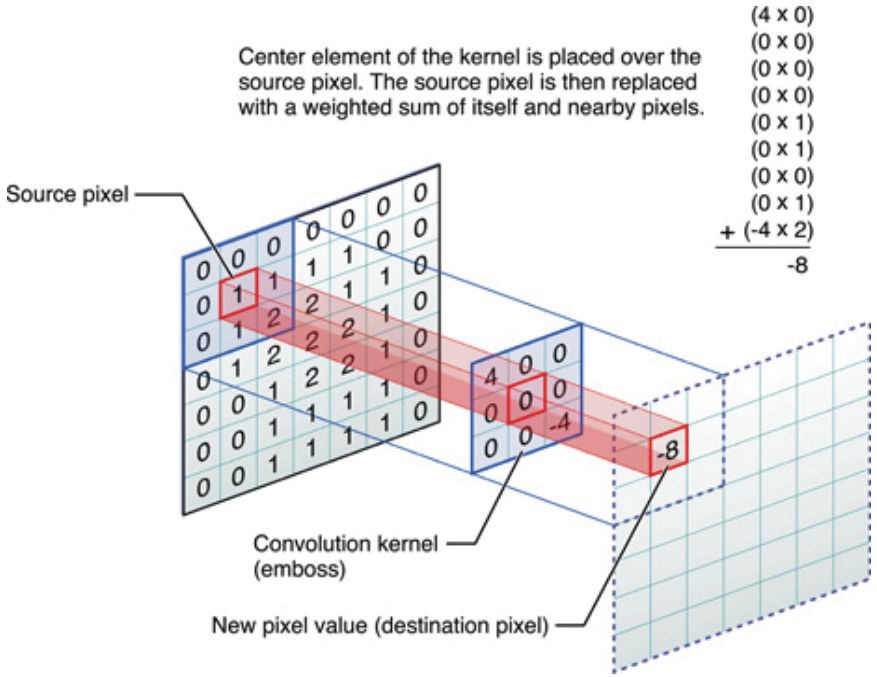
3. Pooling

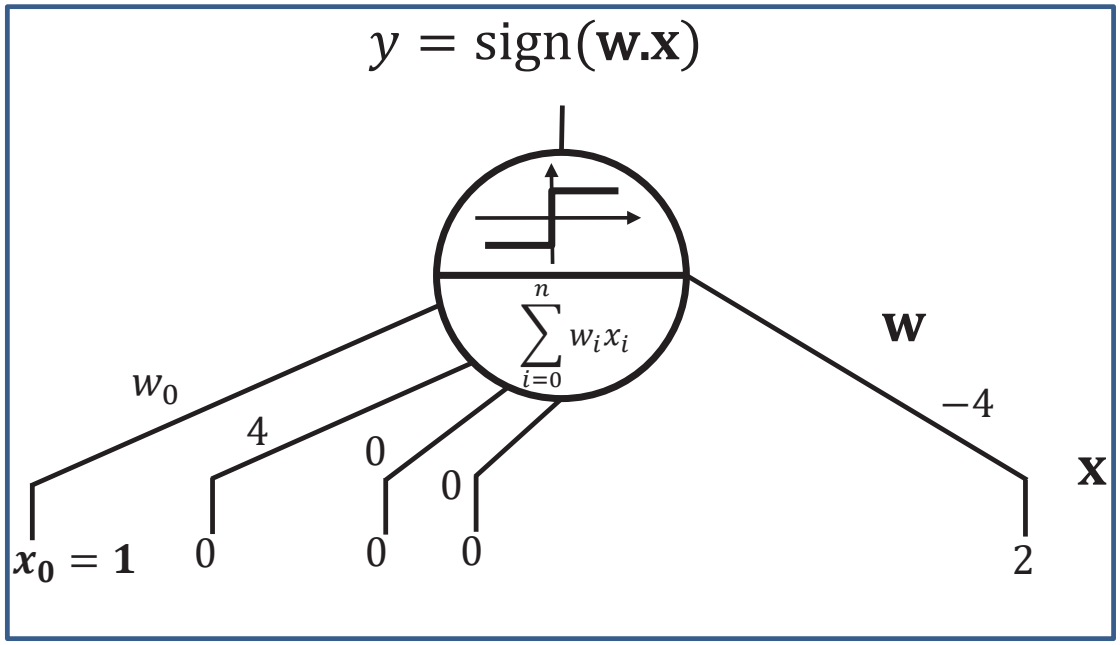


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Convolution = Perceptron

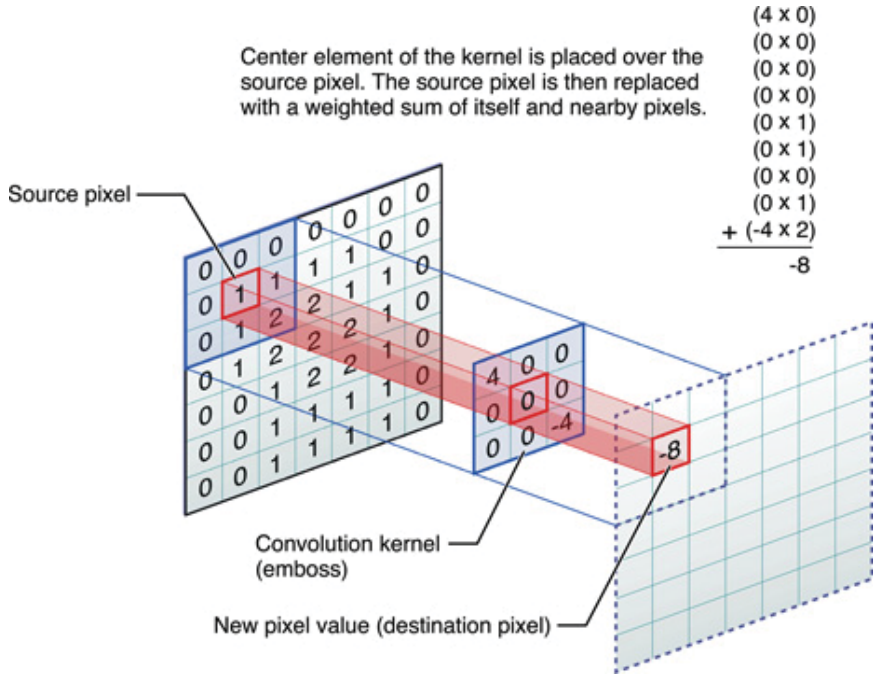


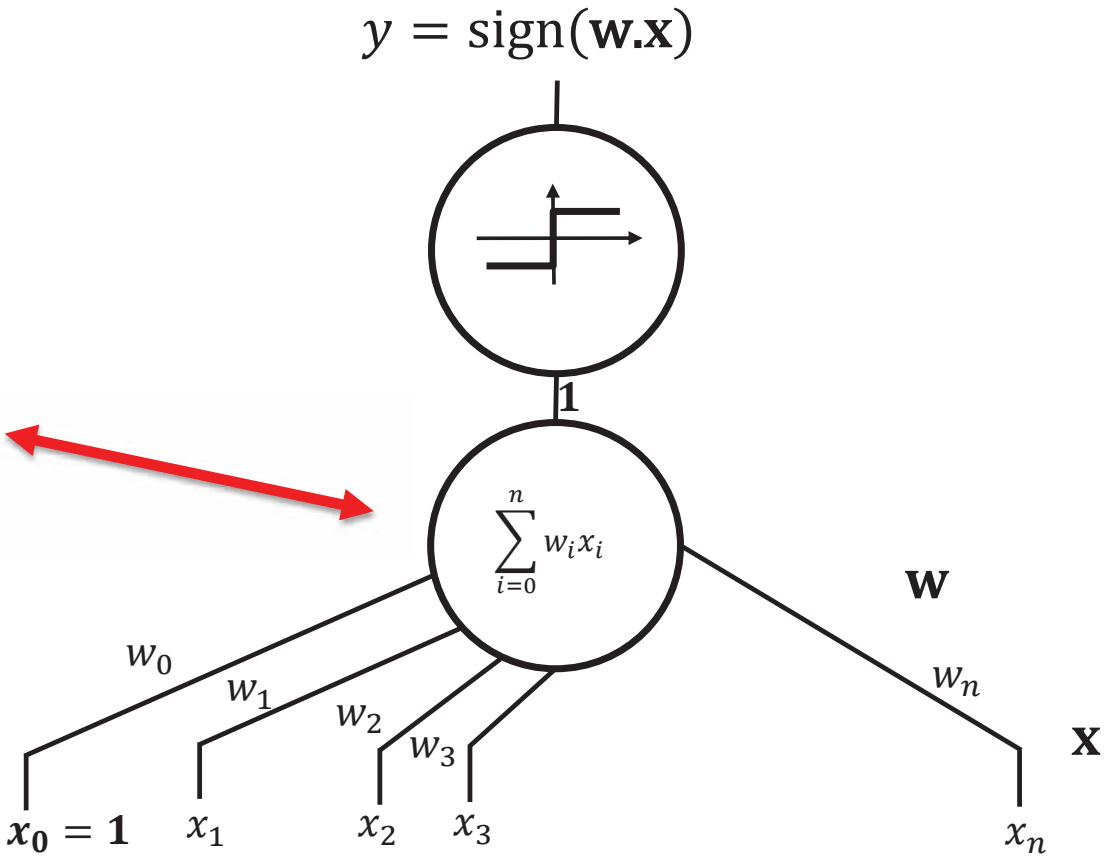


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Convolution = Perceptron





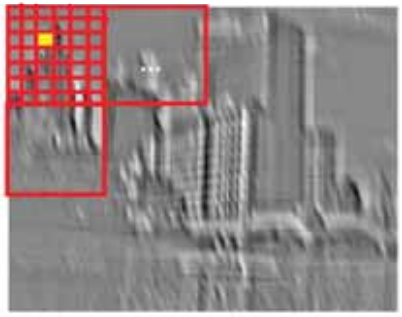
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

If convolution = perceptron

1. Convolution

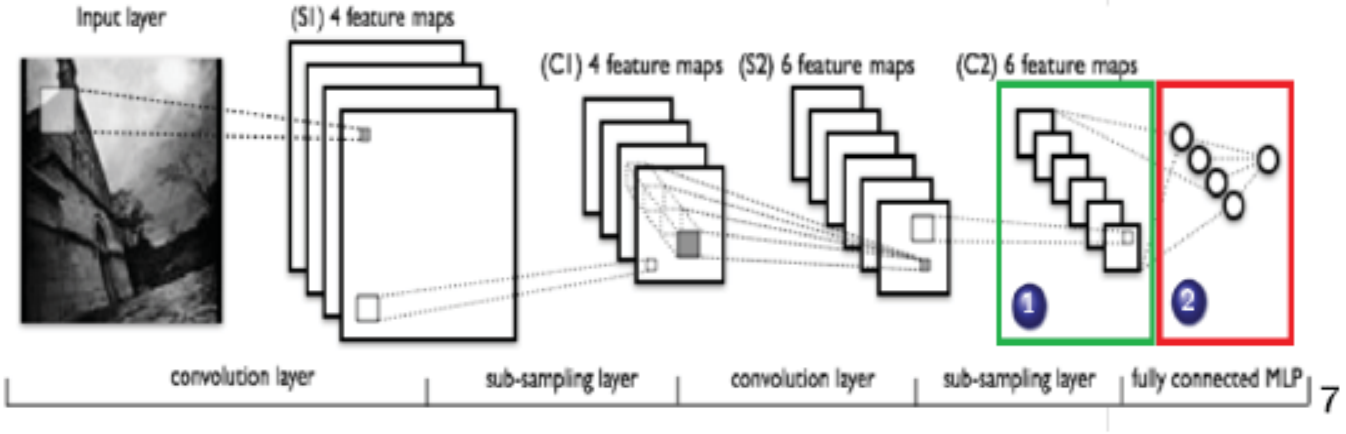
2. Pooling



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deep representation by CNN

- feature map = result of the convolution
- convolution with a filter extract characteristics (*edge detectors*)
- extract parallelised characteristics at each layer

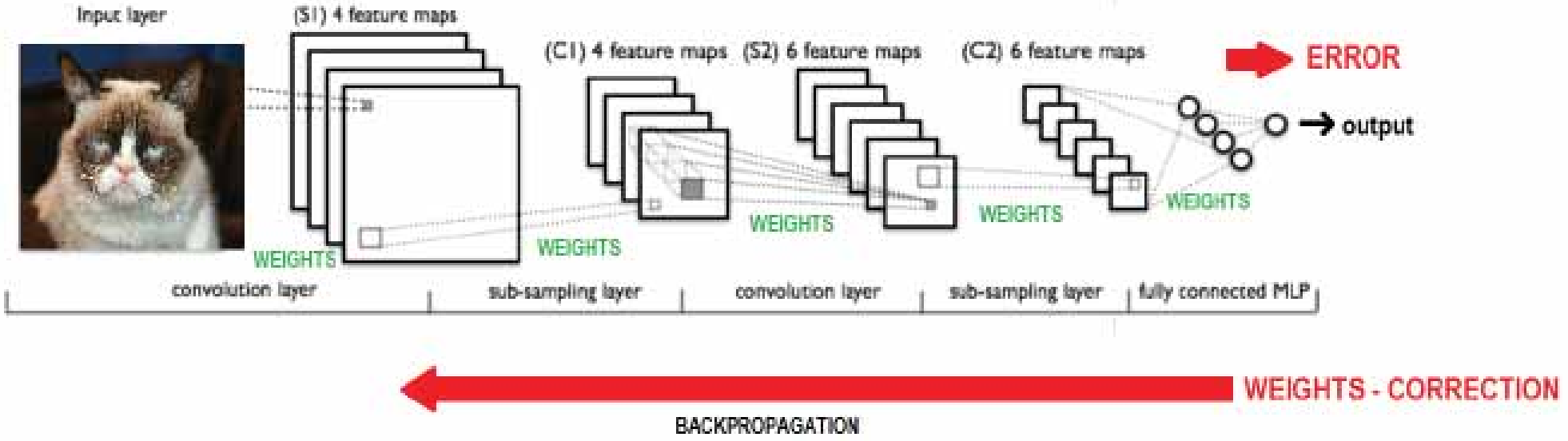


1. final representation of our data
2. classifier (MLP)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

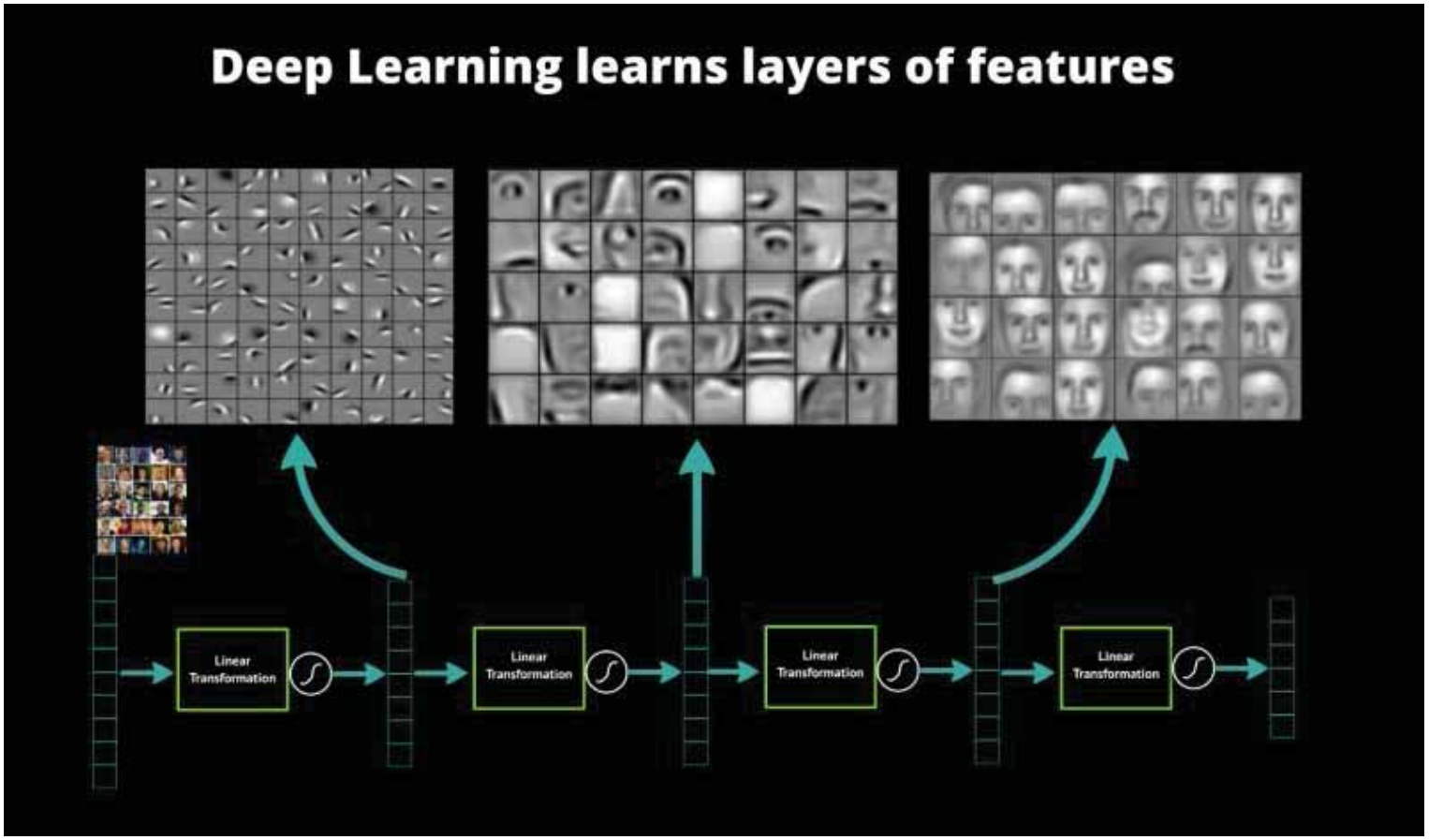
Deep representation by CNN



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

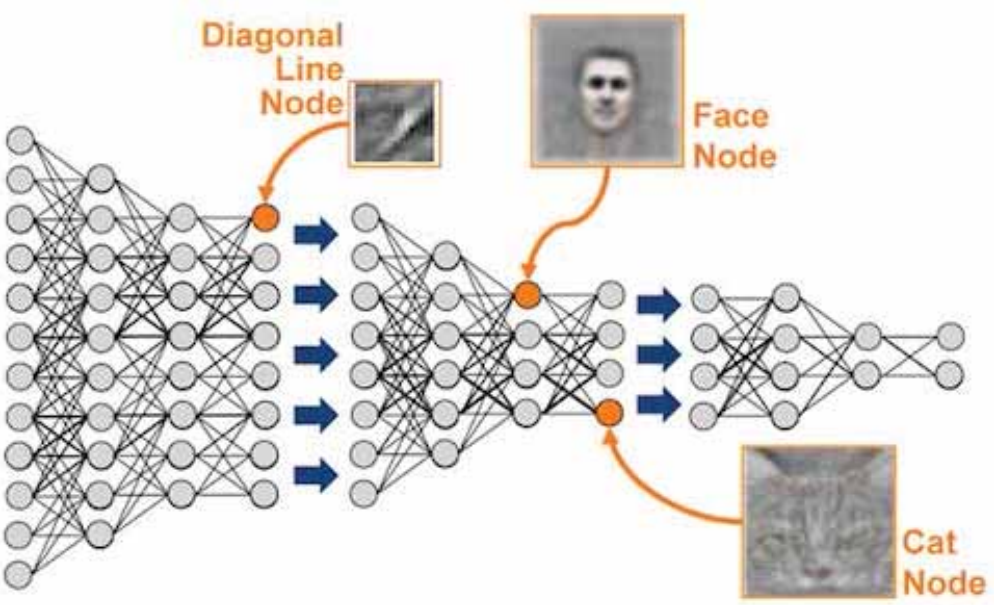
Deep representation by CNN

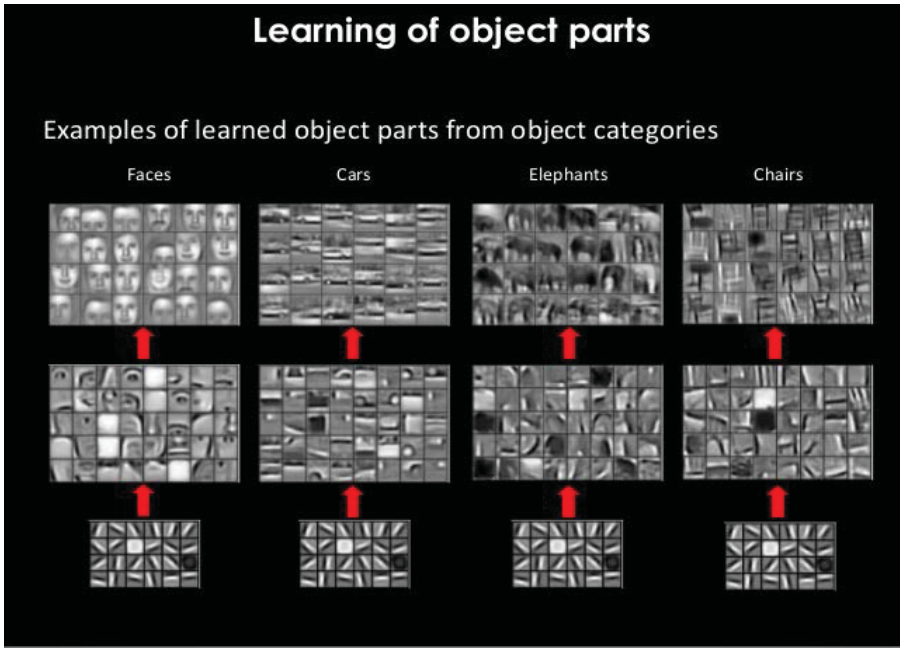


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Deep representation by CNN





UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Transfer Learning!!

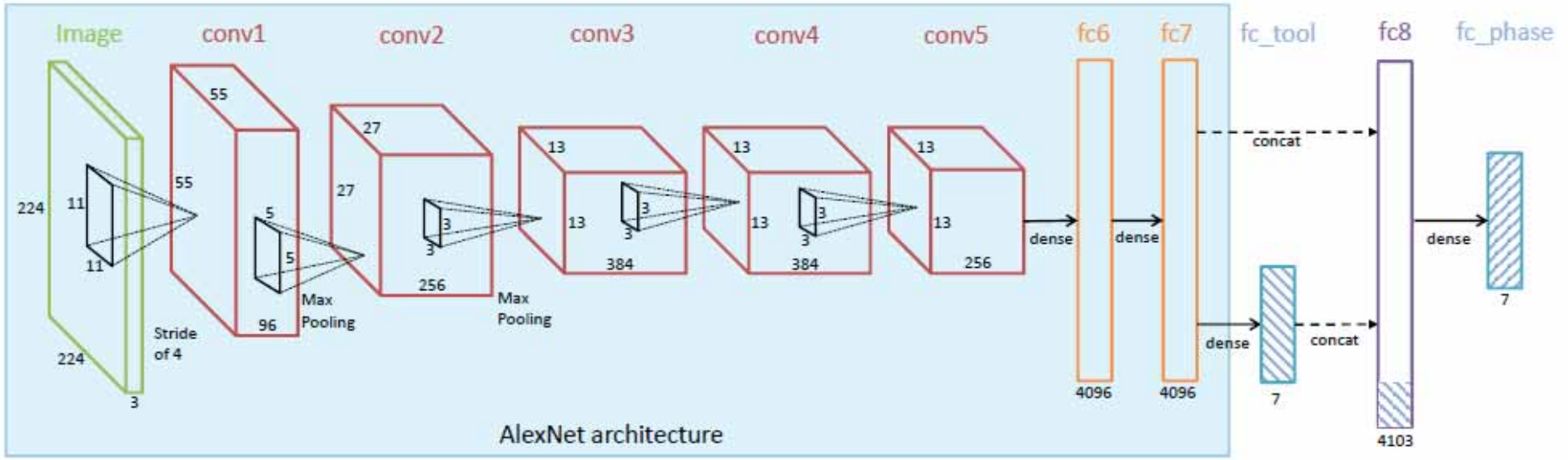
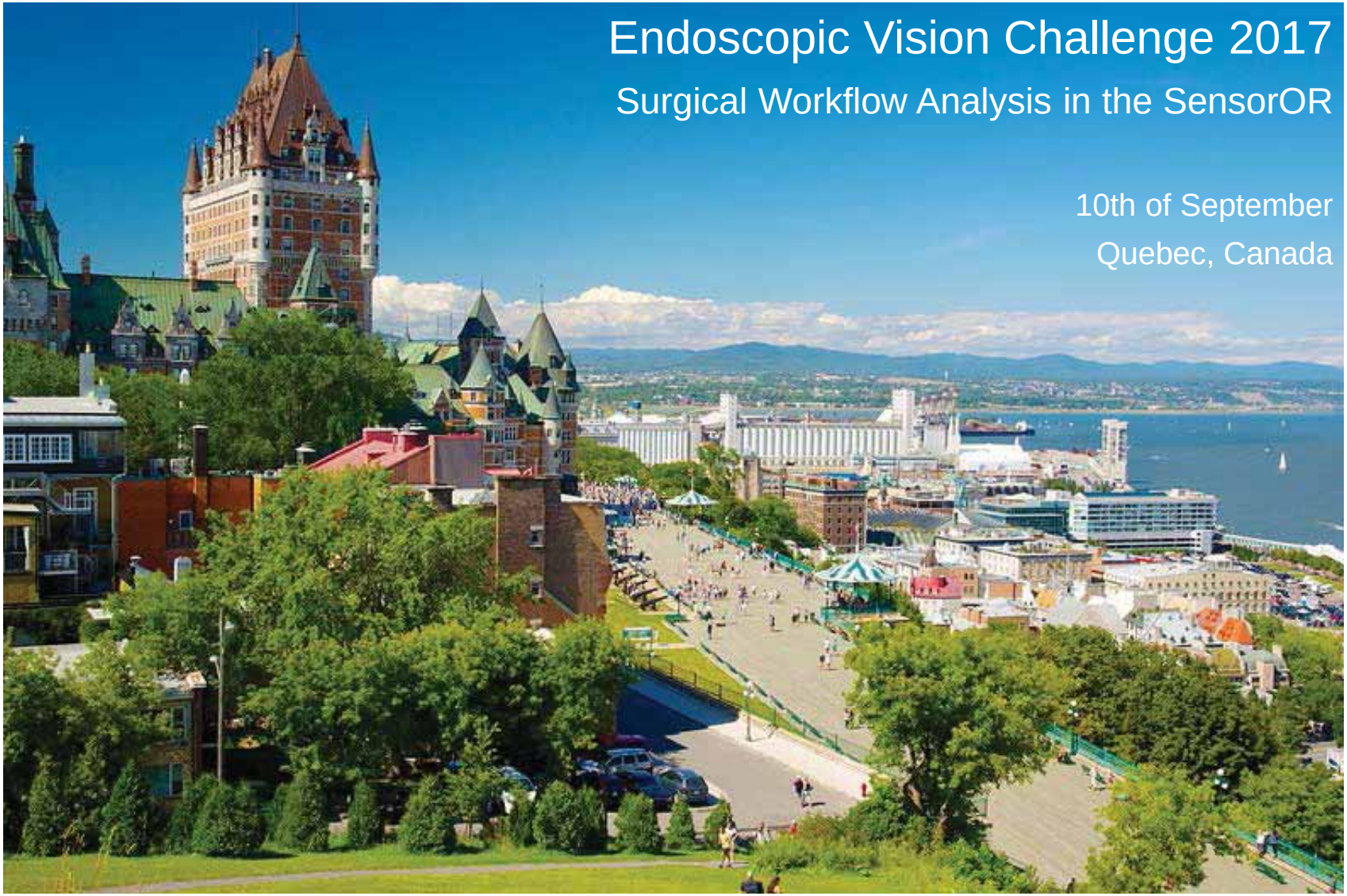


Fig. 2: EndoNet architecture (best seen in color). The layers shown in the turquoise rectangle are the same as in the AlexNet architecture.

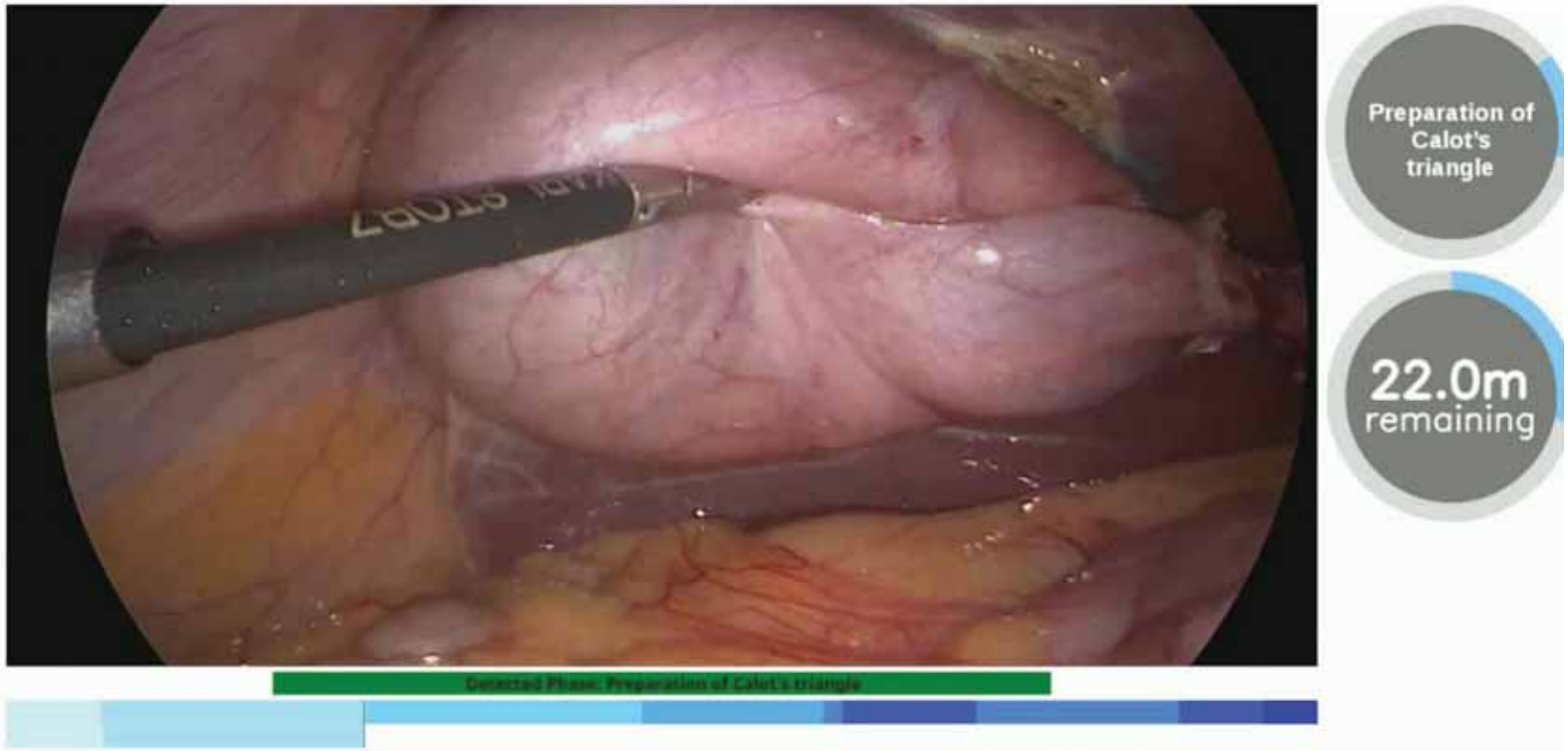

Endoscopic Vision Challenge 2017
Surgical Workflow Analysis in the SensorOR
10th of September
Quebec, Canada

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Clinical context: Laparoscopic Surgery

Surgical Workflow Analysis



UNIVERSITÉ CÔTE D'AZUR

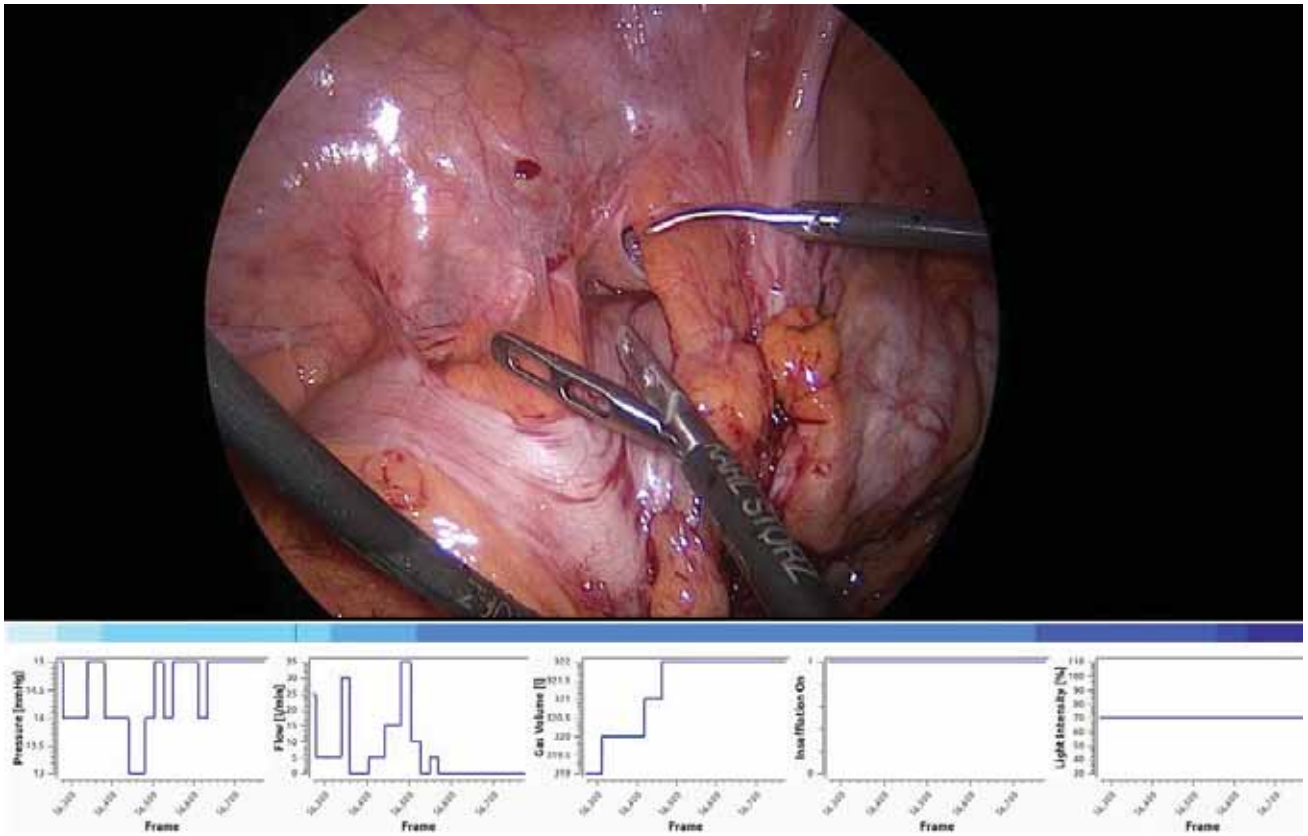
univ-cotedazur.fr

Task

Phase segmentation of laparoscopic surgeries

Video

Surgical Devices



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Dataset

30 colorectal laparoscopies

- Complex type of operation
- Duration: 1.6h – 4.9h (avg 3.2h)
- 3 different sub-types
  - 10x Proctocolectomy
  - 10x Rectal resection
  - 10x Sigmoid resection

Sensor data recorded in integrated OR (Karl Storz OR1)

- Laparoscopic image stream
- Surgical devices

Recorded at

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Annotation

Annotated by surgical experts, 13 different phases

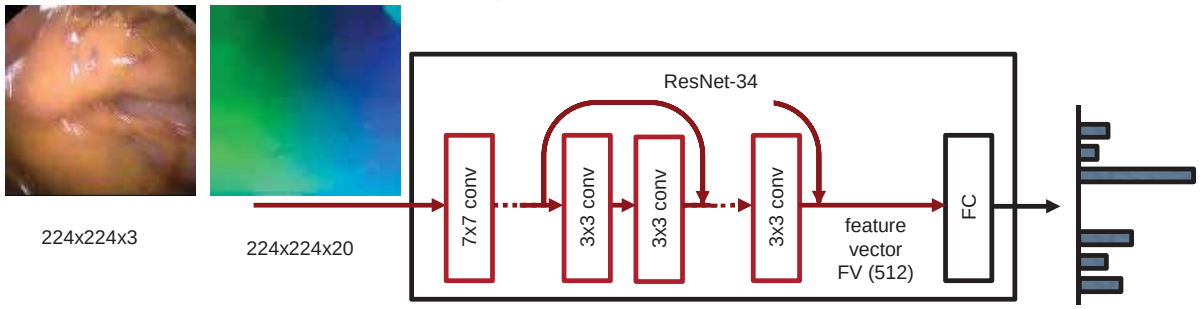
Phase ID	Phase
0	Preparation and orientation at abdomen
1	Dissection of lymphnodes and blood vessels
2	Retroperitoneal preparation to lower pancreatic border
3	Retroperitoneal preparation of duodenum and pancreatic head
4	Mobilizing the sigmoid and the descending colon
5	Mobilizing the spenic flexure
6	Mobilizing the tranverse colon
7	Mobilizing the ascending colon
8	Dissection and resection of rectum
9	Preparing the anastomosis extraabdominally
10	Preparing the anastomosis intraabdominally
11	Placing stoma
12	Finishing the operation
13	Exception (will be ignored during evaluation)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Method

Temporal Network



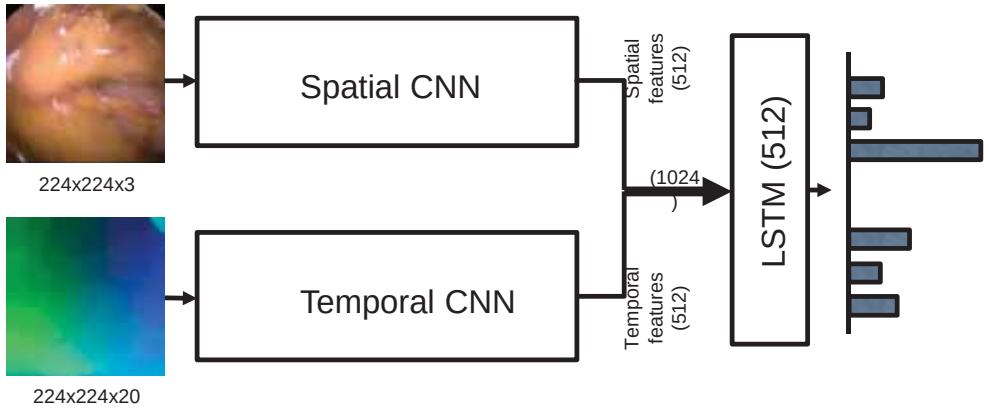
Number of target classes:
Rectal resection: 11
Sigmoid resection: 10
Proctolectomy: 12

Spatial network accuracy:
Rectal resection: 62.91%
Sigmoid resection: 63.01%
Proctolectomy: 63.26%

Temporal network accuracy:
Rectal resection: 49.88%
Sigmoid resection: 48.56%
Proctolectomy: 46.96%

Spatial Network

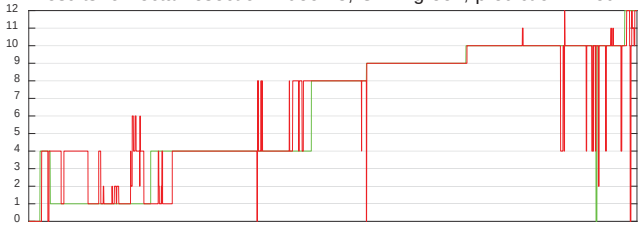
Final Network



Final Network Accuracy:

Rectal resection (8): **80.7%**	sigmoid resection (7): **73.5%**	Proctocolectomy (1): **71.3%**
Rectal resection (6): **79.9%**	sigmoid resection (1): **54.7%**	Proctocolectomy (4): **73.9%**

Results for rectal resection video #8; GT in green, prediction in red



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

And the winner is...

	Data used	Average Jaccard	Median Jaccard	Accuracy	
1	Video	40%	38%	61%	Team UCA
2	Video + Device	38%	38%	60%	Team NCT
3	Video	25%	25%	57%	Team TUM
4	Device	16%	16%	36%	Team TUM
5	Video	8%	7%	21%	Team FFL

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Why Deep Learning?

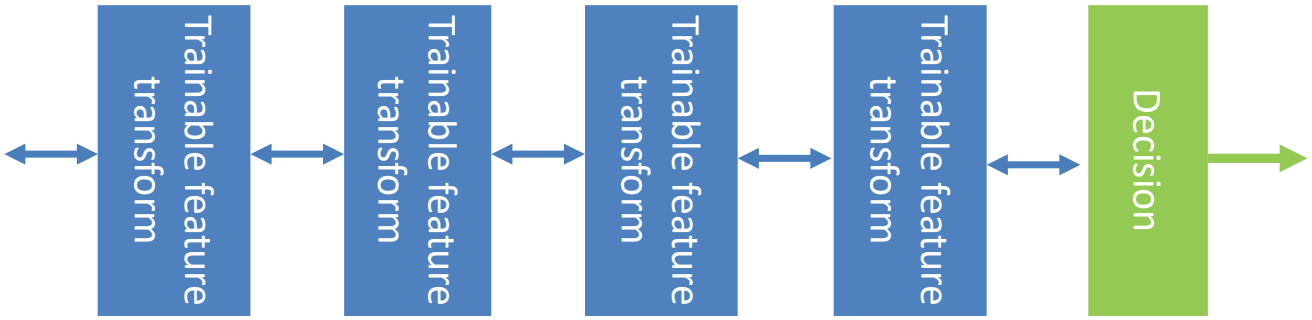
Before Deep Learning

Hand-crafted/Engineered features

- Image recognition
  - Pixel → edge → texton → patterns → part → object
- Text
  - Character → word → word group → clause → sentence → story
- Speech
  - Sample → spectral band → sound → … → phone → phoneme → word

Since Deep Learning

→ Hierarchy of representations with increasing level of abstraction
→ Each stage is a kind of trainable feature transform…
as long as you have enough data to train the hierarchy



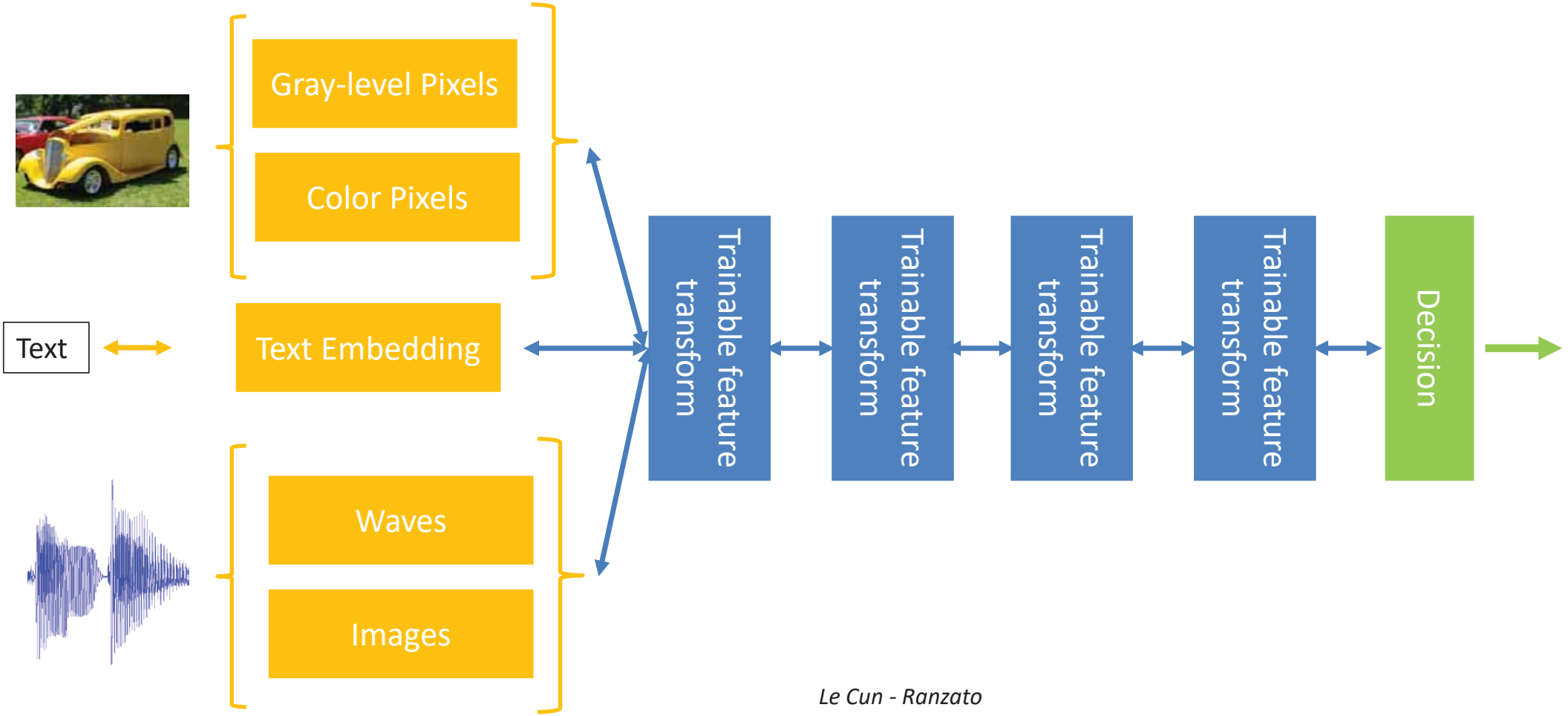
Le Cun - Ranzato

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

How Deep Learning?

Start from raw data OR from a first level representation?



Le Cun - Ranzato

AMAZING BUT…

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Amazing but…be carreful of the bias in the initial data

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings

Tolga Bolukbasi[1], Kai-Wei Chang[2], James Zou[2], Venkatesh Saligrama[1,2], Adam Kalai[2]

[1]Boston University, 8 Saint Mary's Street, Boston, MA

[2]Microsoft Research New England, 1 Memorial Drive, Cambridge, MA

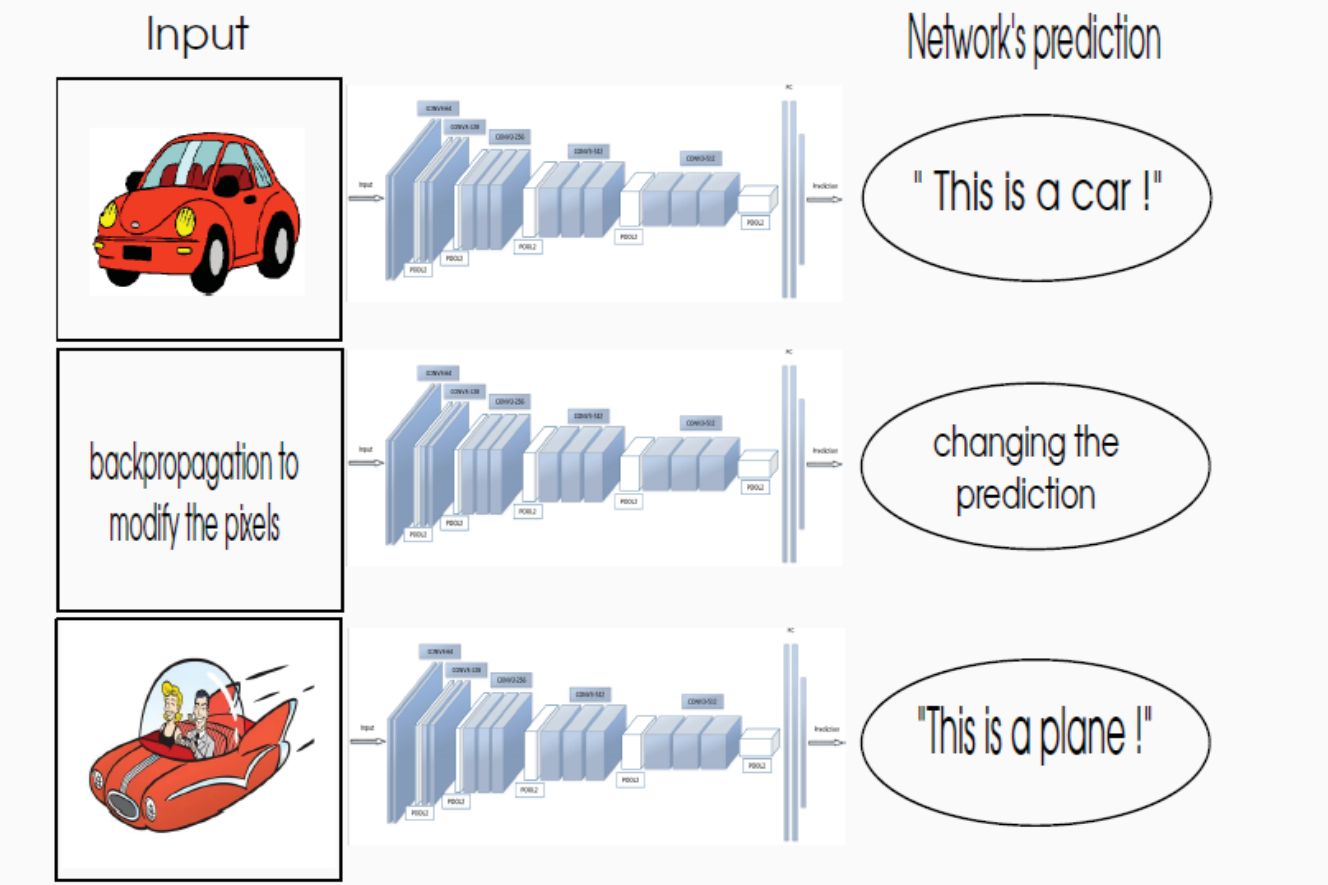
tolgab@bu.edu, kw@kwchang.net, jamesyzou@gmail.com, srv@bu.edu, adam.kalai@microsoft.com

Abstract

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

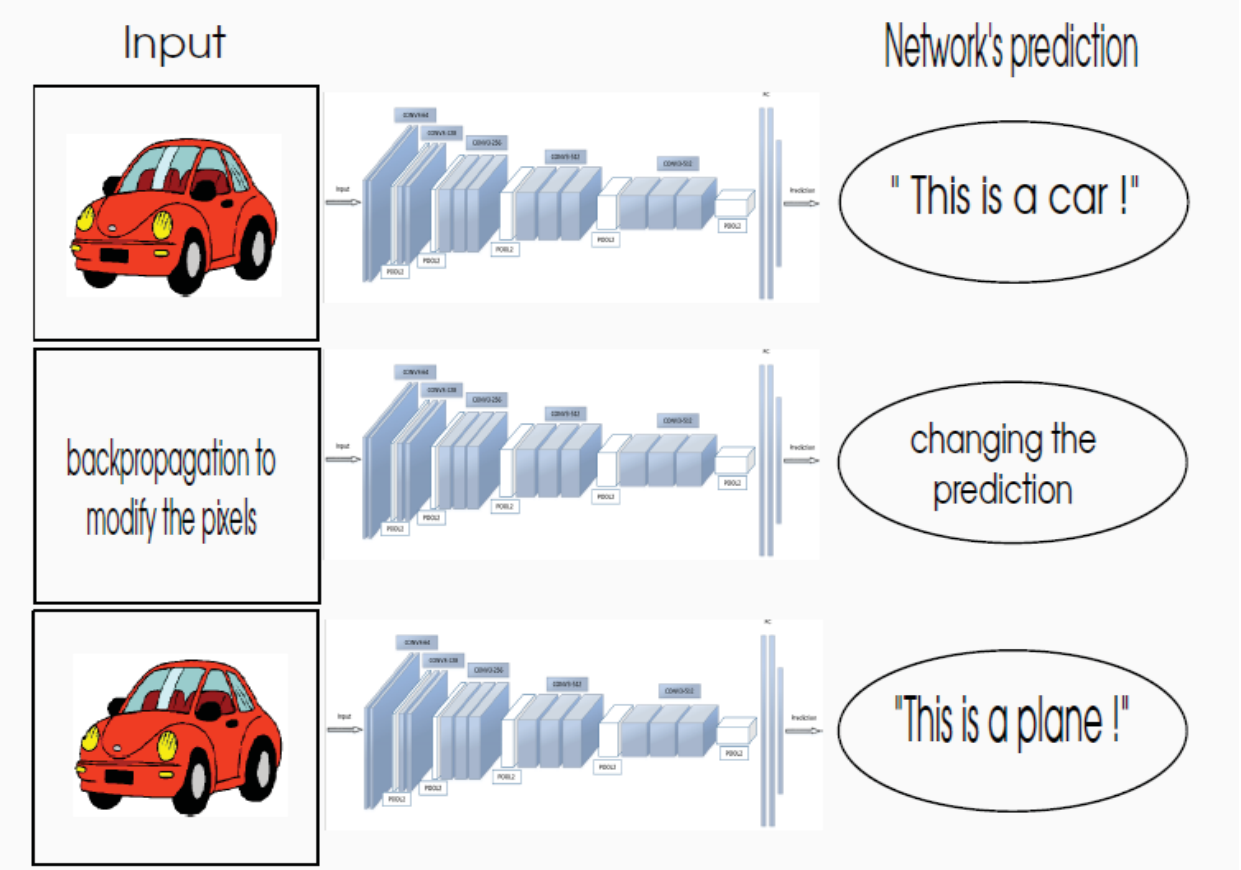
Amazing but…be careful of the adversaries (as any other ML algorithms)



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

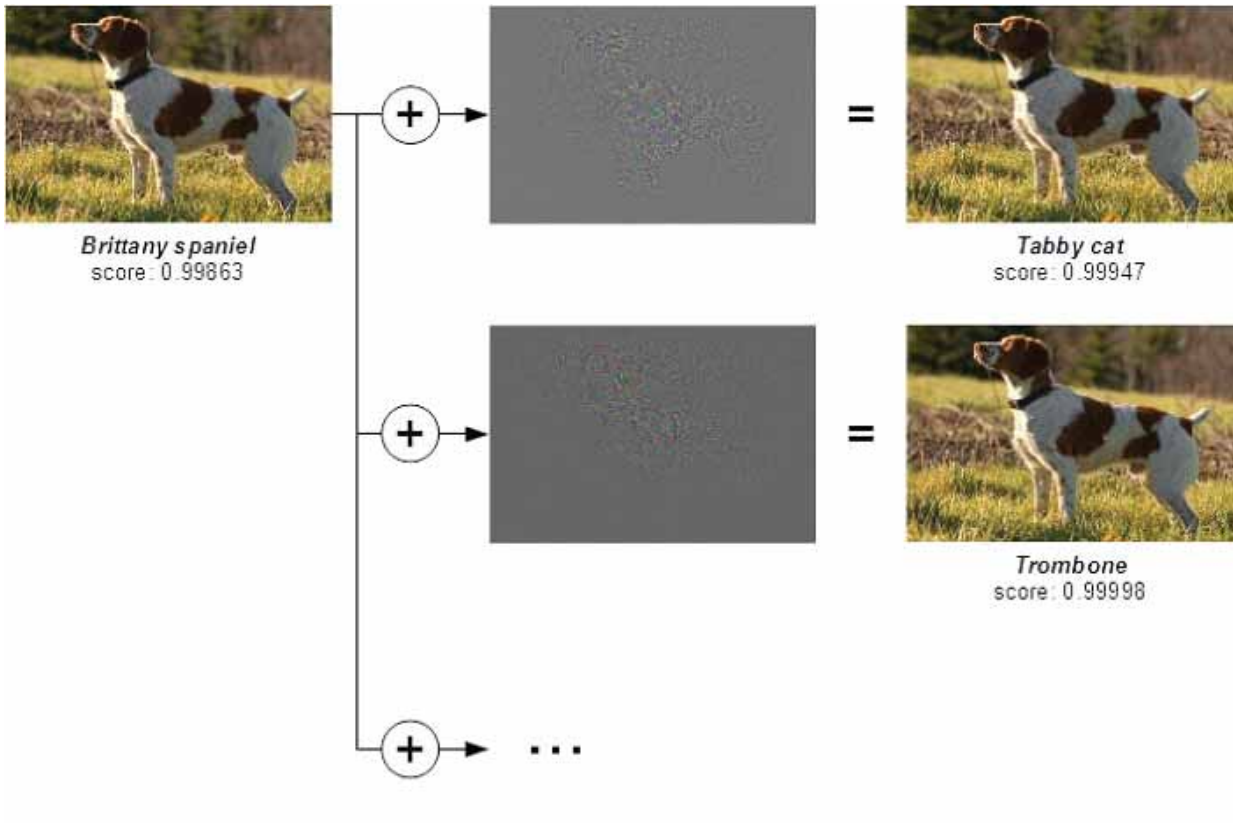
Amazing but…be careful of the adversaries (as any other ML algorithms)



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Amazing but…be careful of the adversaries (as any other ML algorithms)



From Thomas Tanay

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Amazing but…be careful of the adversaries (as any other ML algorithms)

Red Light Modified to Green after 18 white pixels. Probability: 59%

Red Light Modified to Green after 9 green pixels. Probability: 50.9%

Red Light Modified to Green after 9 green pixels. Probability: 53%

No Light Modified to Green after 4 green pixels. Probability: 51.9%

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Amazing but…be careful of the adversaries (as any other ML algorithms)

stop	30m speed limit	80m speed limit	30m speed limit	go right	go straight
	Confidence	0.999964	0.99		

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Amazing but…be careful of the adversaries

https://nicholas.carlini.com/code/audio_adversarial_examples/

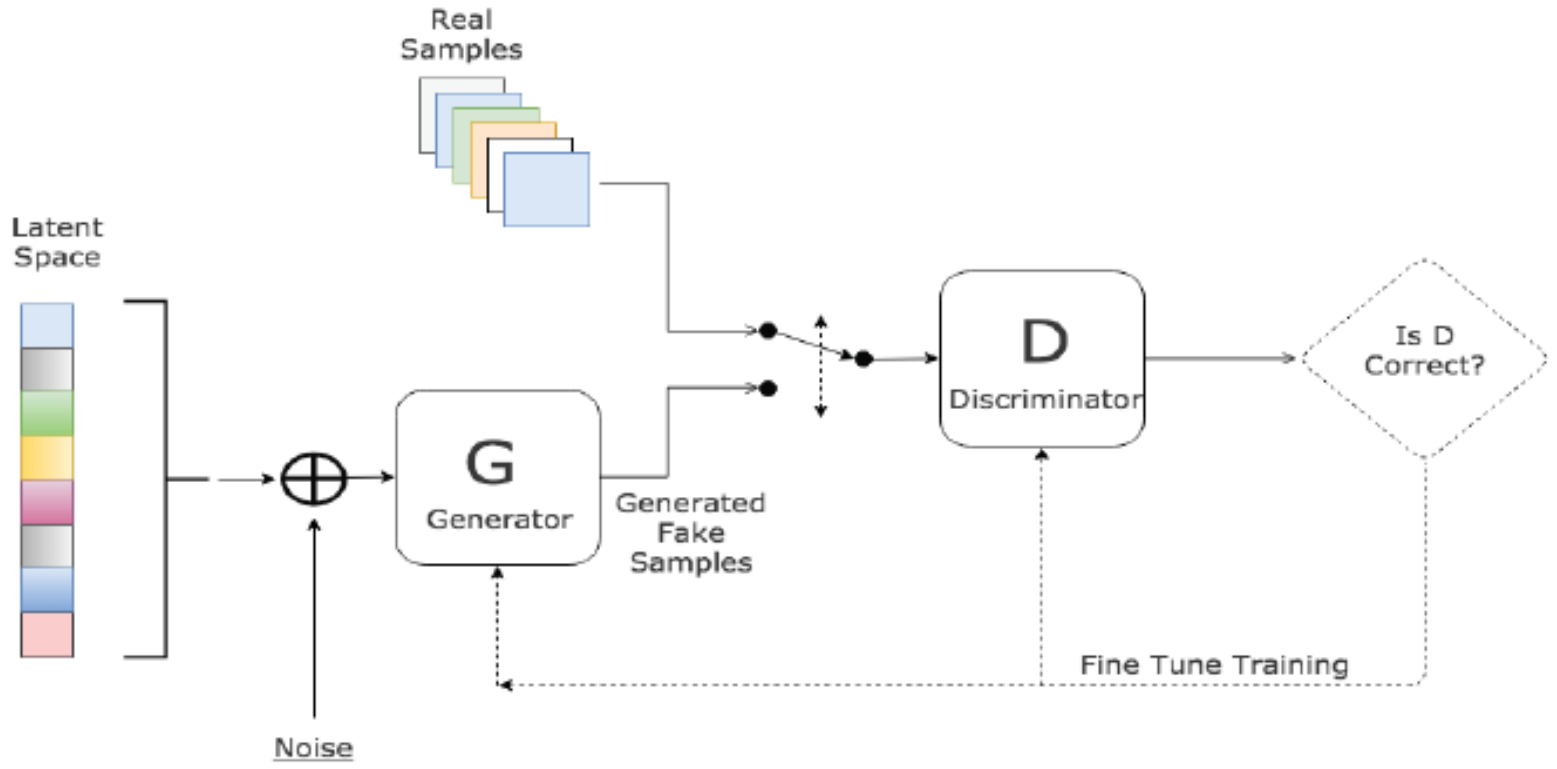
UNIVERSITÉ CÔTE D'AZUR

GENERATIVE ADVERSARIAL NETWORKS

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

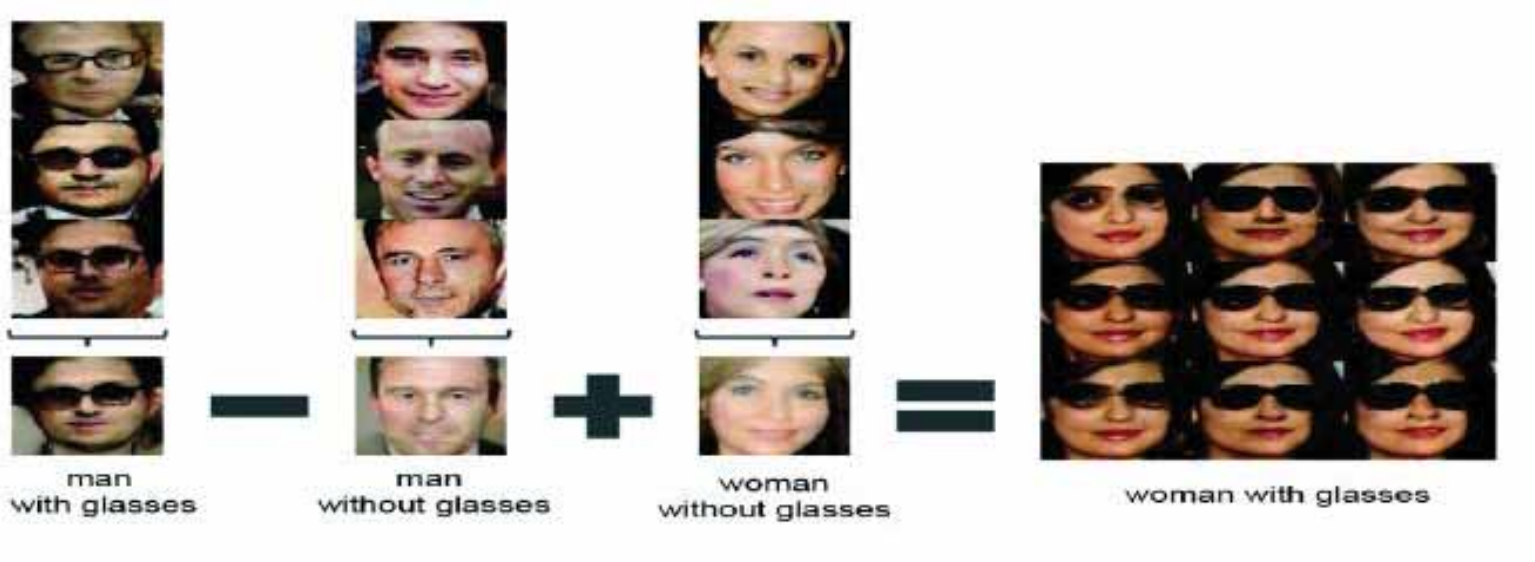
How to solve it?
Generative Adversarial Networks



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

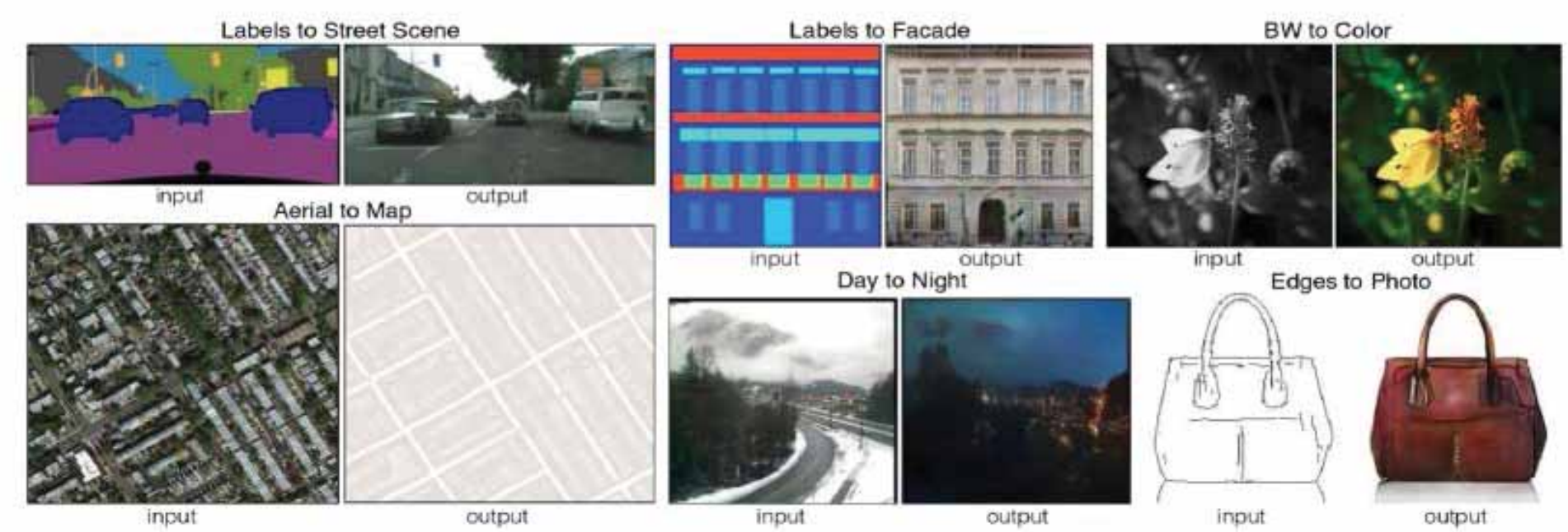
It finally did not solve adversarial, but…

Operations between latent representations (manifold)



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

It finally did not solve adversarial, but…



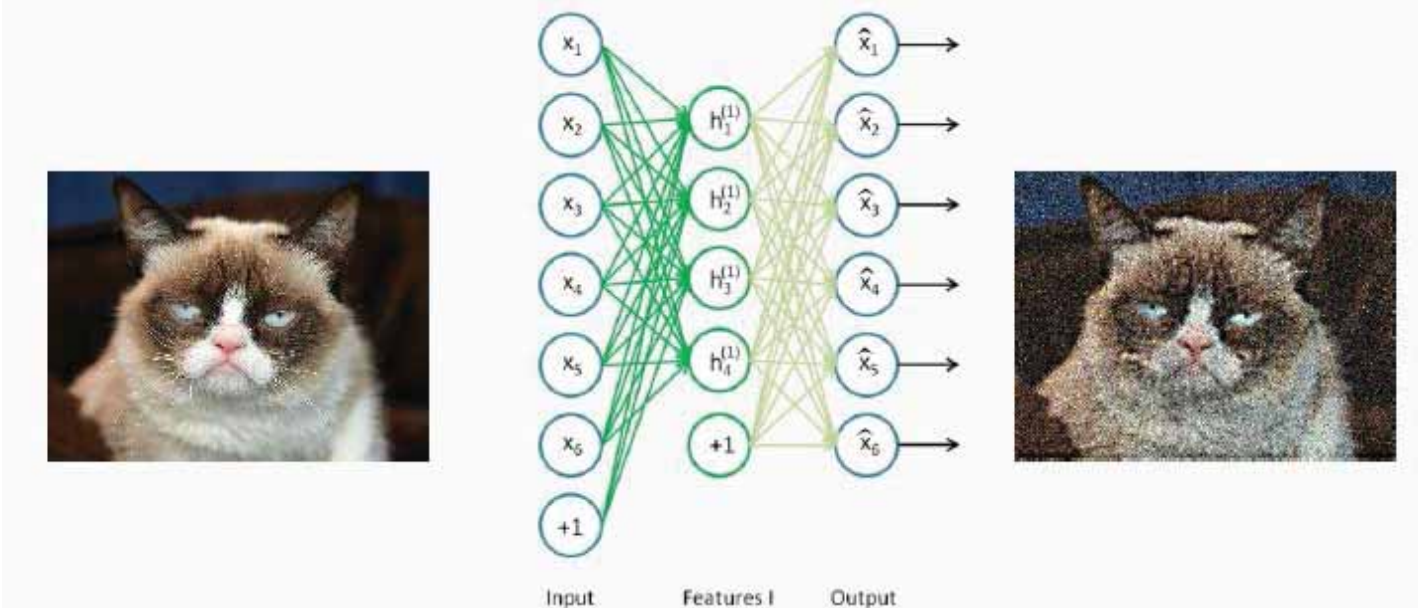
UNIVERSITÉ CÔTE D'AZUR

(DENOISING) STACKED AUTOENCODER

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Autoencoder: unsupervised!

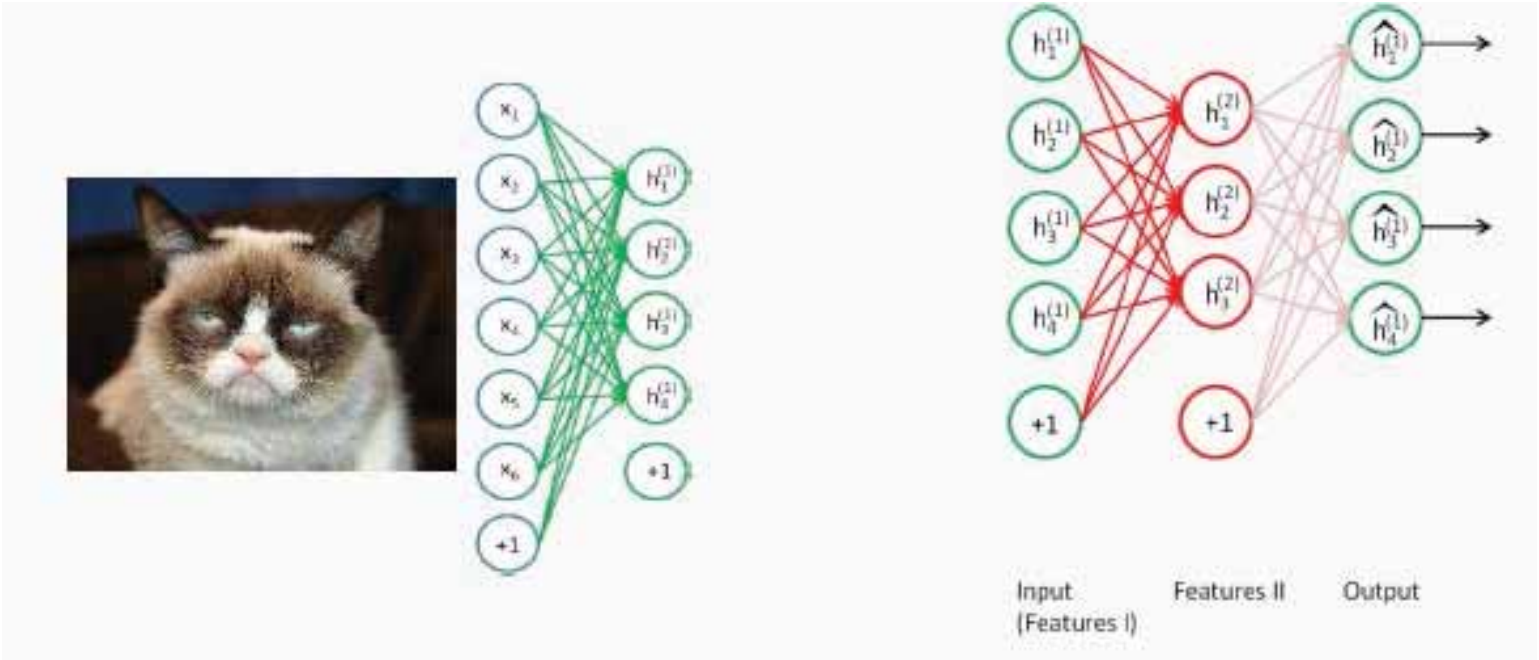
Learning a compact representation of the data (no classification)
First we train an AutoEncoder layer 1.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Autoencoder: unsupervised!

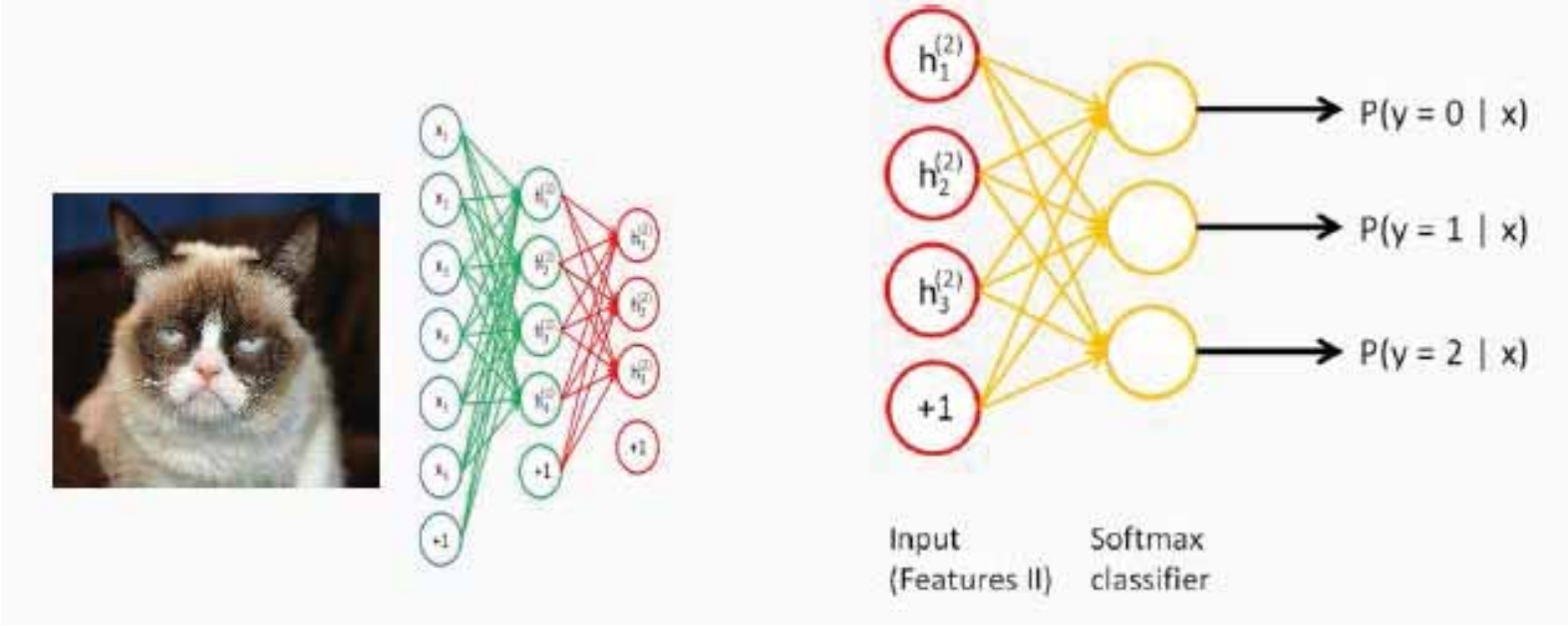
Second we train an AutoEncoder layer 2.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Autoencoder -> Supervised

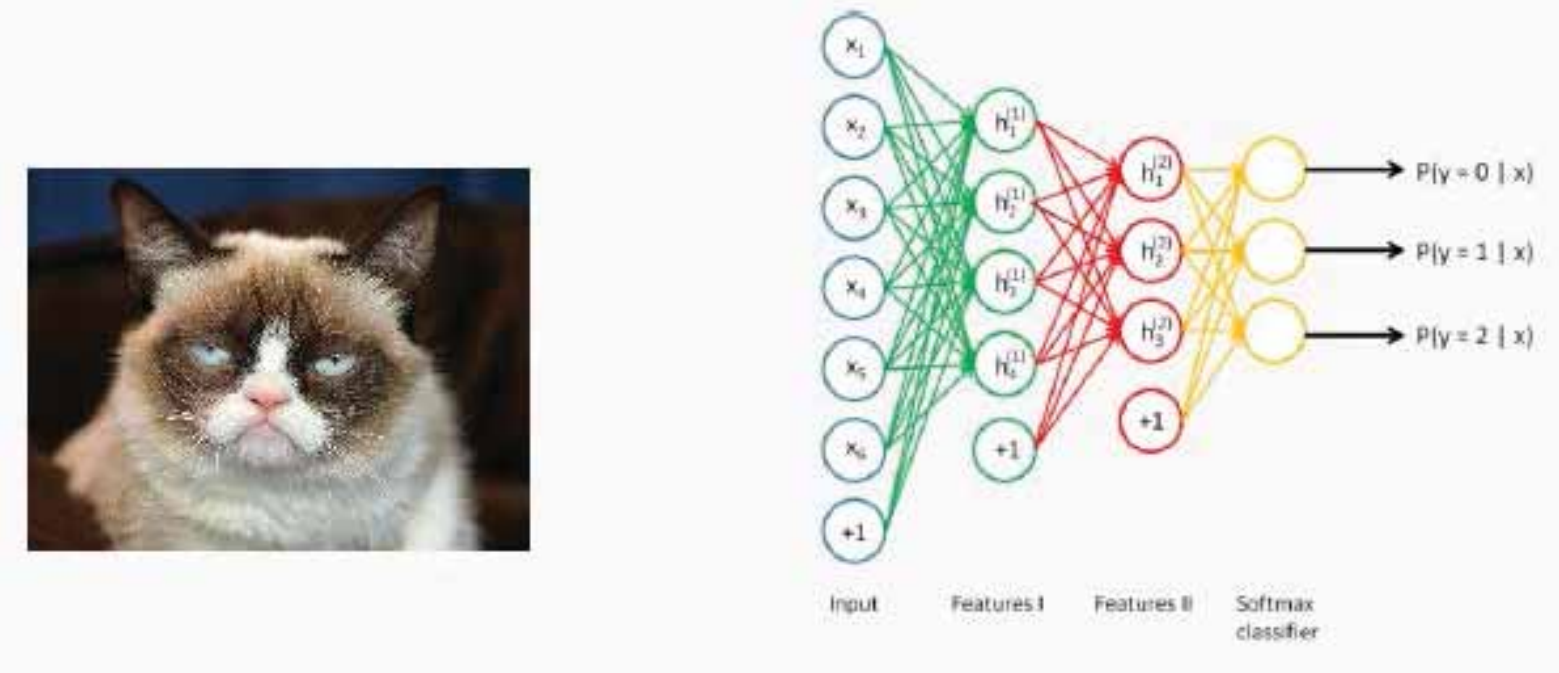
Then we train an output layer of non-linearities based on softmax.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Autoencoder -> Supervised

Finally, we fine-tune the whole network in a supervised way.

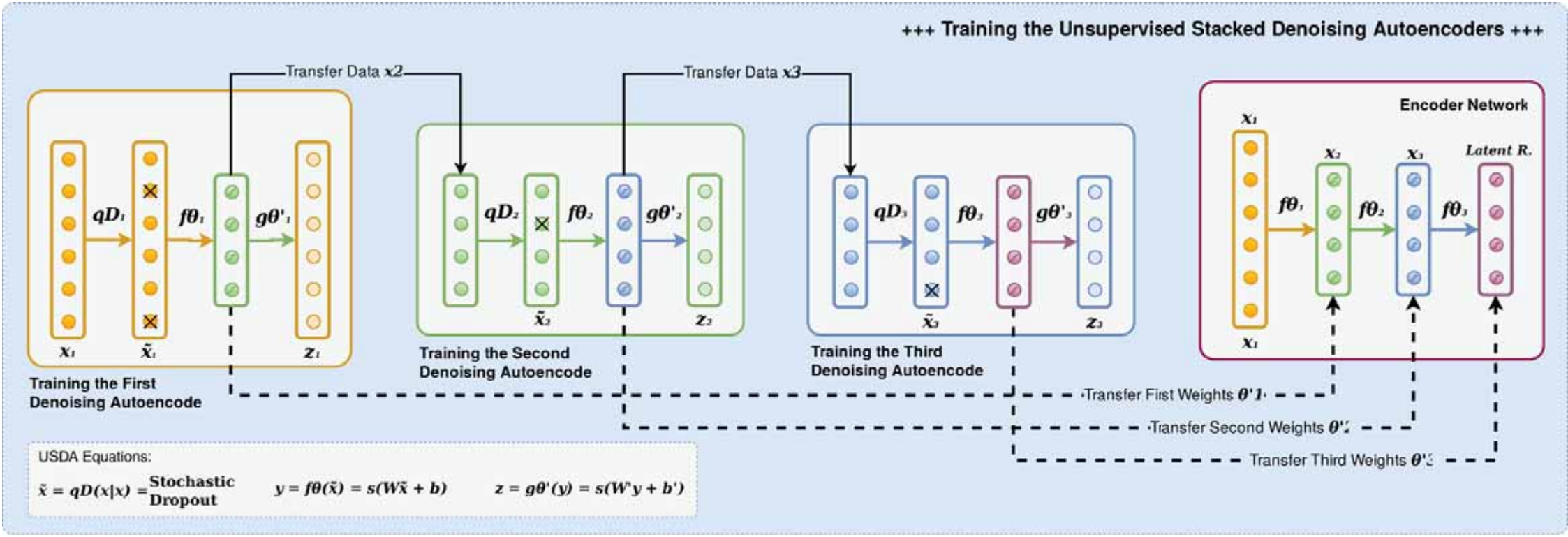


UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Denoising stacked Autoencoder: unsupervised

Result = a new latent representation



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Denoising stacked Autoencoder: example

Model *Deep Patient*, Published in Nature, 2016

Stage 1.

Data-mining stage & Feature extraction:

Driving Electronic Health Records to build a binary phenotype representation.

Stage 2.

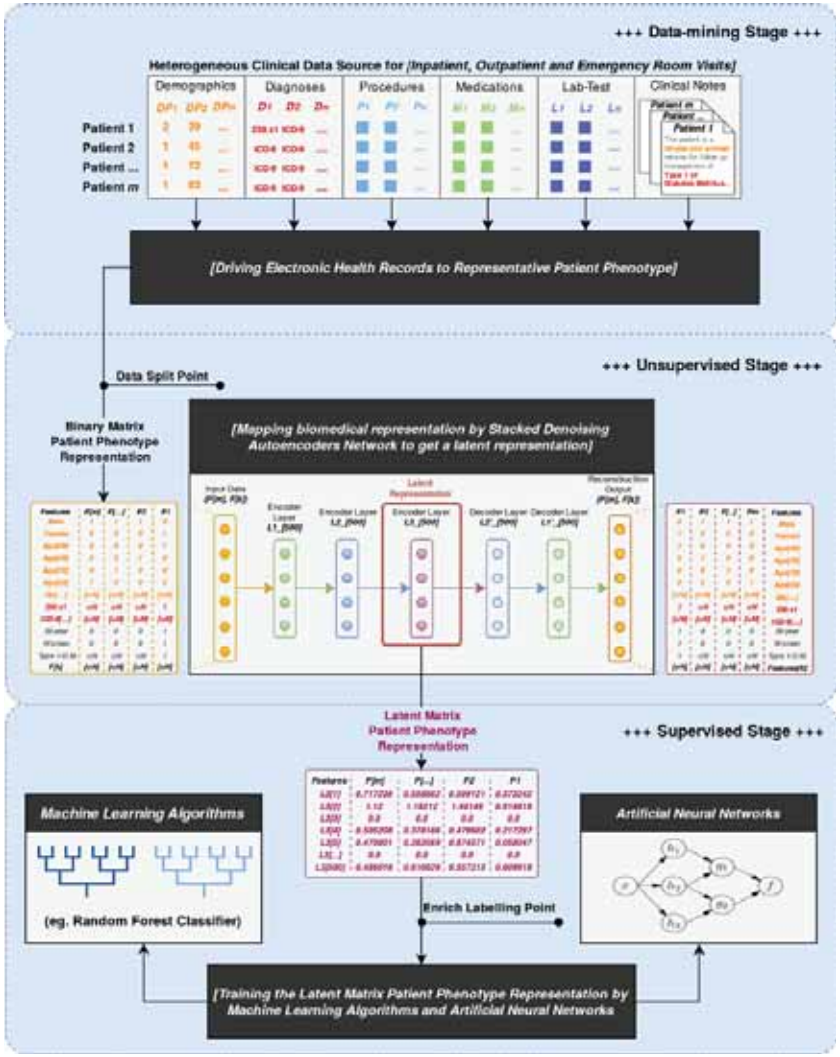
Unsupervised stage:

Mapping the Binary Patient Representation to get a new space call Deep Patient (or Latent Representation) Using Stacked Denoising Autoencoders.

Stage 3.

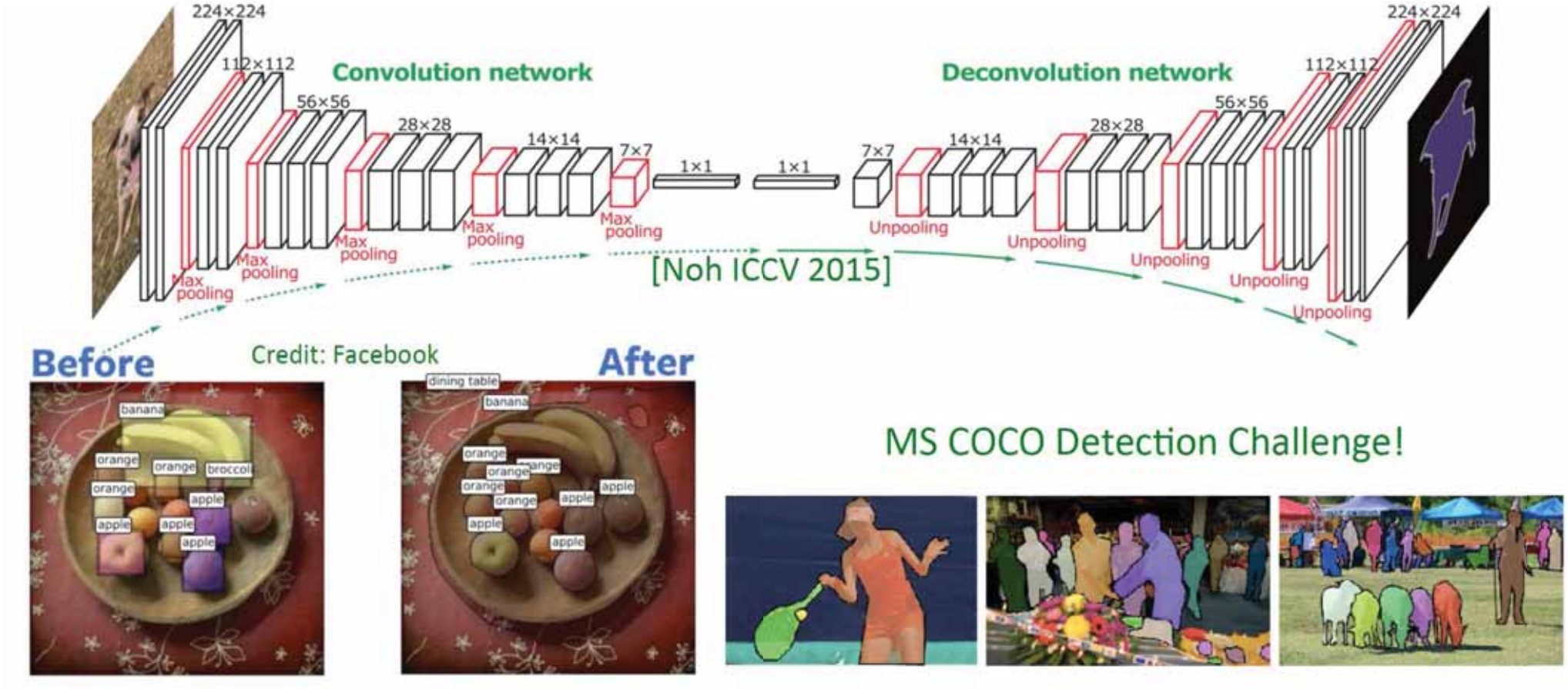
Supervised stage:

Labeling Medical Target and training the *Latent Representation* by Machine Learning algorithms for classification and prediction of patient's disease.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Supervised Image Segmentation Task



Credits Matthieu Cord

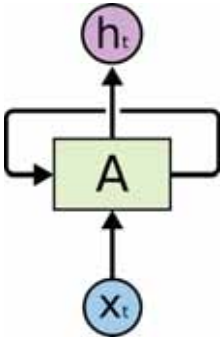
UNIVERSITÉ CÔTE D'AZUR

Partly from COLAH's Blog http://colah.github.io/posts/2015-08-Understanding-LSTMs/

RECURRENT NEURAL NETWORK

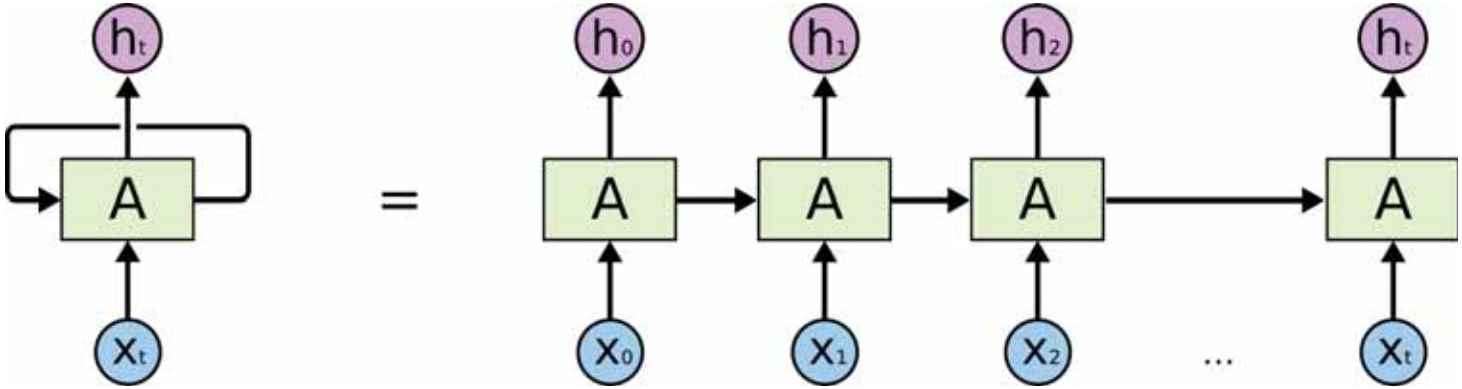
UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Recurrent Neural Networks have loops.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

An unrolled recurrent neural network.

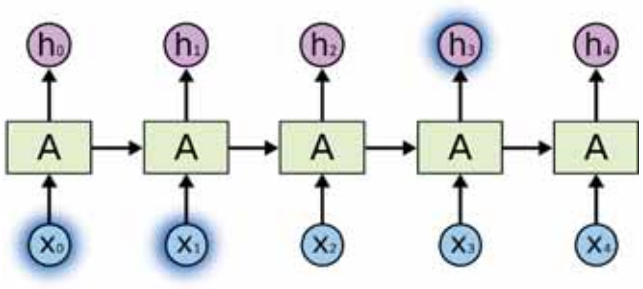


In the last few years, there have been incredible success applying RNNs to a variety of problems: speech recognition, language modeling, translation, image captioning…

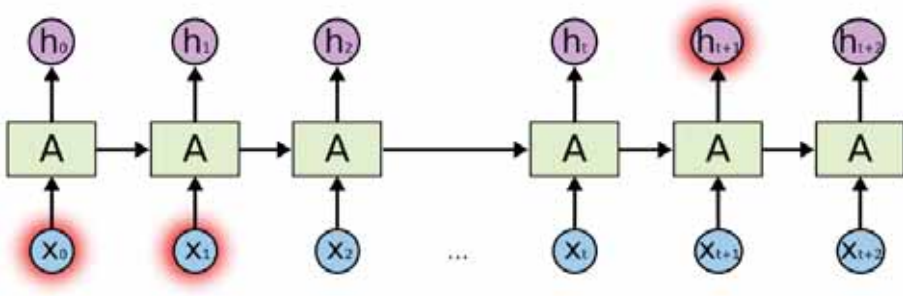
UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

The Problem of Long-Term Dependencies



*If we are trying to predict the last word in "the clouds are in the **(sky)**," we don't need any further context – it's pretty obvious the next word is going to be sky.*

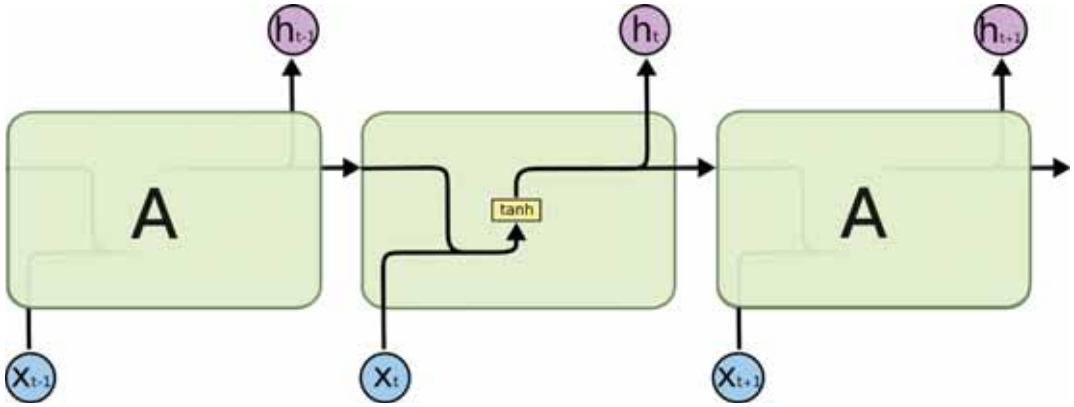


*Consider trying to predict the last word in the text "I grew up in France… I speak fluent **(French)**." Recent information suggests that the next word is probably the name of a language, but if we want to narrow down which language, we need the context of France, from further back.*

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

The repeating module in a standard RNN contains a single layer.



Unfortunately, as that gap grows, RNNs become unable to learn to connect the information (cf. vanishing gradients)

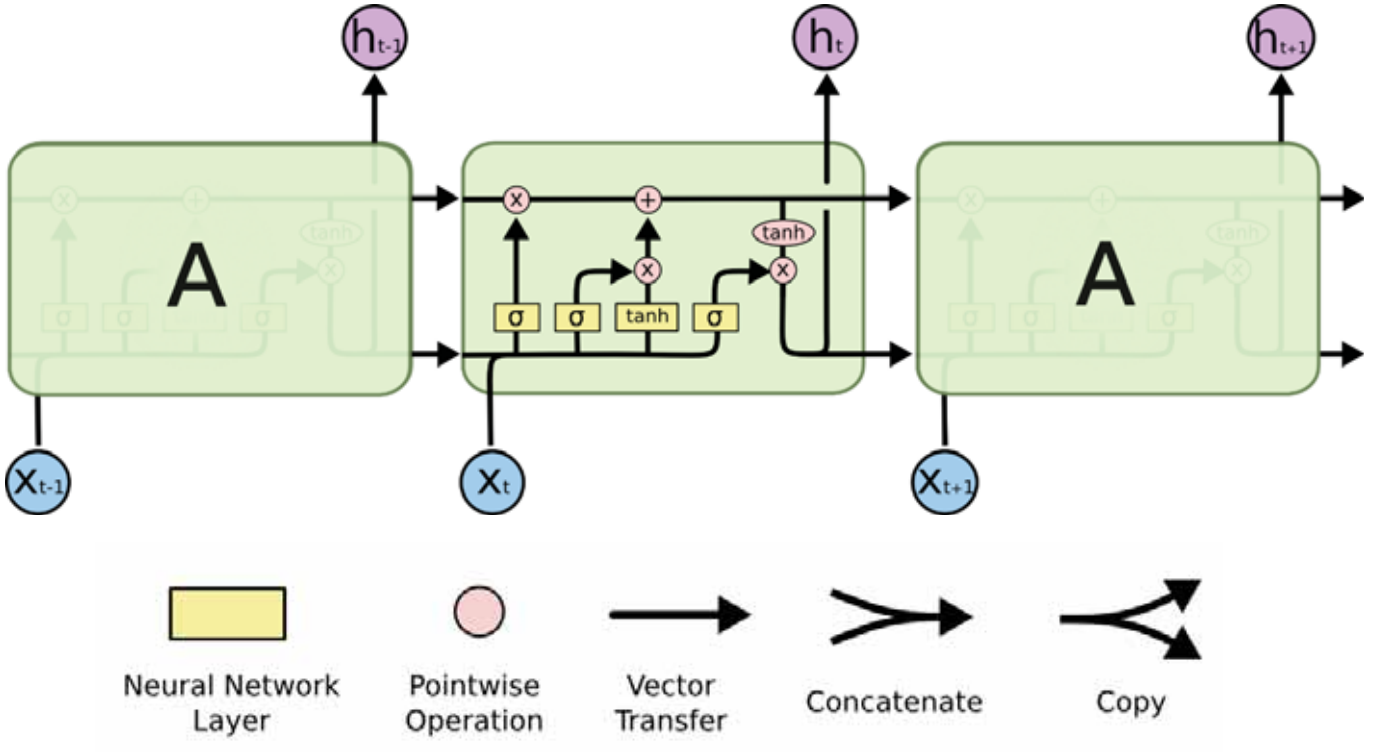
The problem was explored in depth by Hochreiter (1991) and Bengio, et al. (1994), who found some pretty fundamental reasons why it might be difficult.

Thankfully, LSTMs/GRUs do not have this problem!

UNIVERSITÉ CÔTE D'AZUR

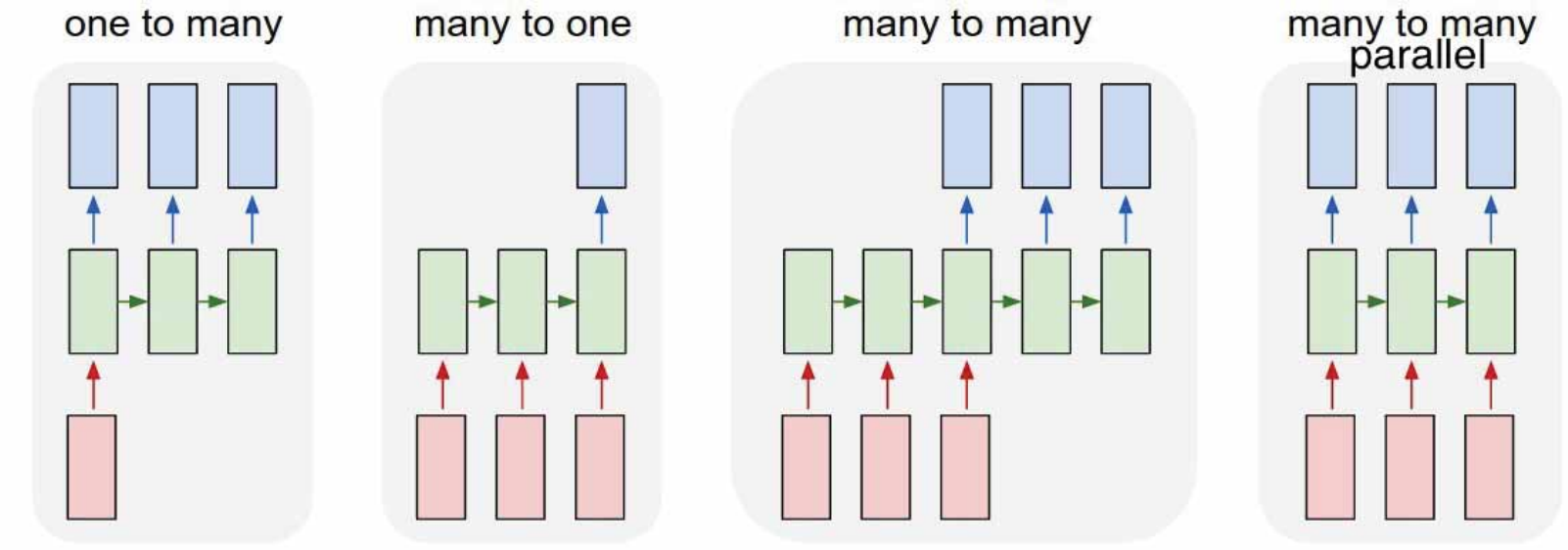
univ-cotedazur.fr

The repeating module in an LSTM/GRU contains four interacting layers.



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

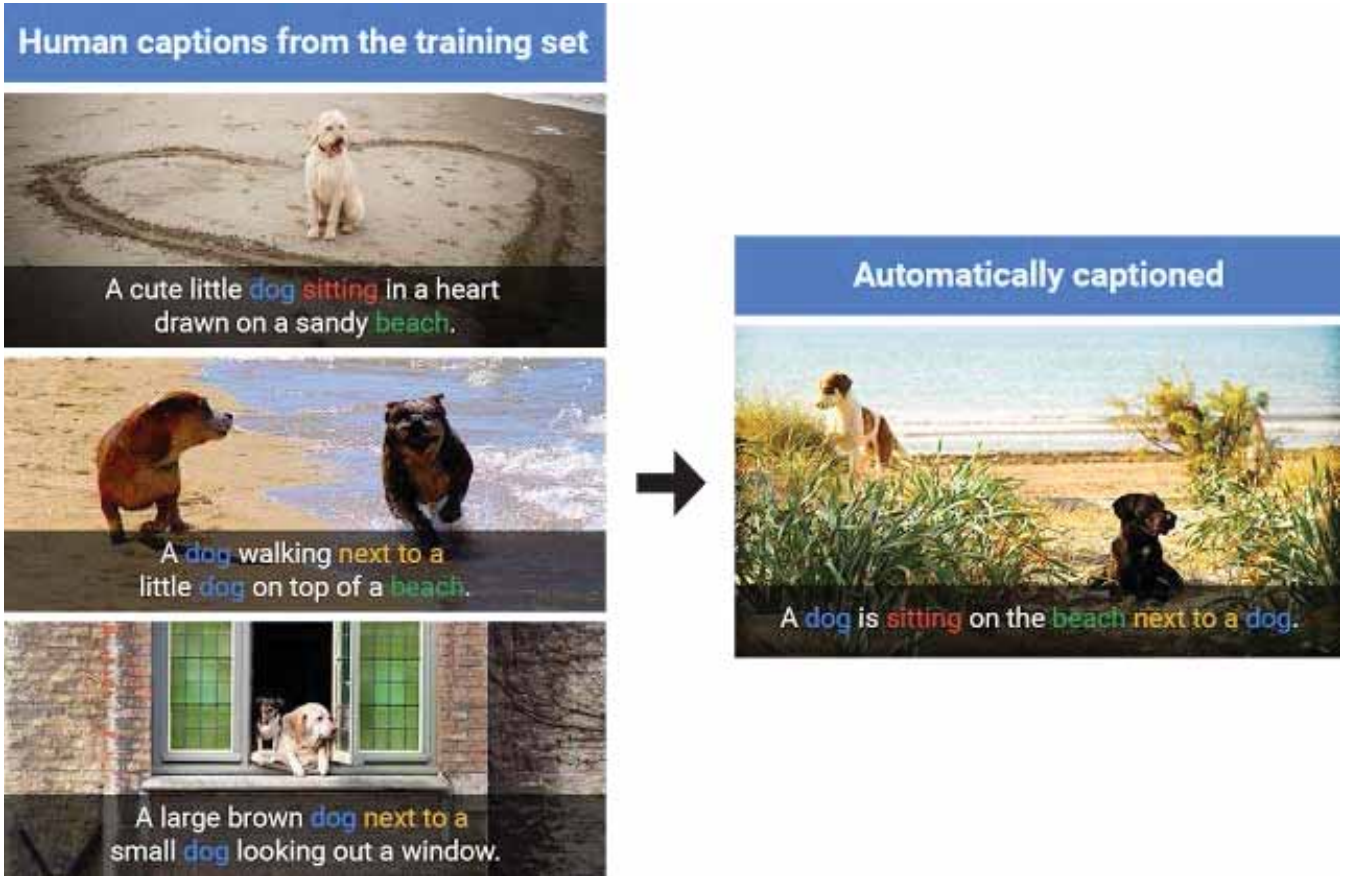
Sequence modeling with RNNs



UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

One to Many - Image captioning



[Xu et al. 2015]

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Many to Many Parallel - Char-nn

Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma ??. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

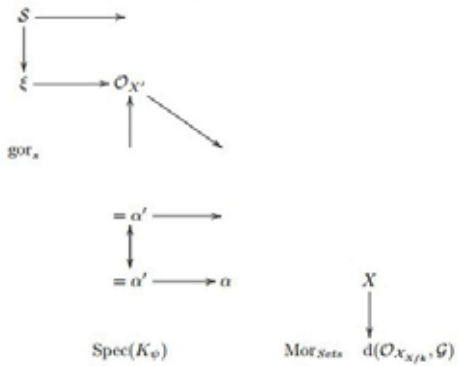
be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □

This since $\mathcal{F} \in \mathcal{F}$ and $x \in \mathcal{G}$ the diagram



is a limit. Then $\mathcal{G}$ is a finite type and assume S is a flat and $\mathcal{F}$ and $\mathcal{G}$ is a finite type f_*. This is of finite type diagrams, and

- the composition of $\mathcal{G}$ is a regular sequence,
- $\mathcal{O}_{X'}$ is a sheaf of rings.

□

Proof. We have see that $X = \mathrm{Spec}(R)$ and $\mathcal{F}$ is a finite type representable by algebraic space. The property $\mathcal{F}$ is a finite morphism of algebraic stacks. Then the cohomology of X is an open neighbourhood of U. □

Proof. This is clear that $\mathcal{G}$ is a finite presentation, see Lemmas ??.
A *reduced above* we conclude that U is an open covering of $\mathcal{C}$. The functor $\mathcal{F}$ is a "field

$$\mathcal{O}_{X,x} \longrightarrow \mathcal{F}_{\overline{x}} \quad \text{-1}(\mathcal{O}_{X_{\acute{e}tale}}) \longrightarrow \mathcal{O}_{X_t}^{-1}\mathcal{O}_{X_\lambda}(\mathcal{O}_{X_\eta}^{\overline{s}})$$

is an isomorphism of covering of $\mathcal{O}_{X_\iota}$. If $\mathcal{F}$ is the unique element of $\mathcal{F}$ such that X is an isomorphism.
The property $\mathcal{F}$ is a disjoint union of Proposition ?? and we can filtered set of presentations of a scheme $\mathcal{O}_X$-algebra with $\mathcal{F}$ are opens of finite type over S.
If $\mathcal{F}$ is a scheme theoretic image points. □

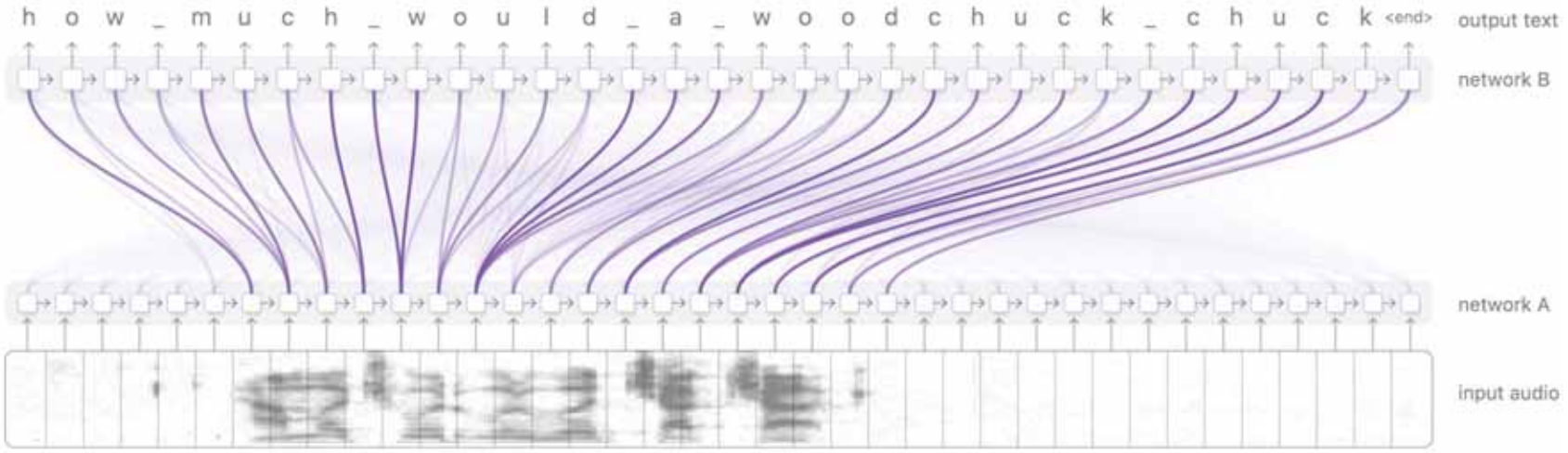
If $\mathcal{F}$ is a finite direct sum $\mathcal{O}_{X_\lambda}$ is a closed immersion, see Lemma ??. This is a sequence of $\mathcal{F}$ is a similar morphism.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

Many to Many - Machine translation speech2text

- Input : Audio mp3 (natural language)
- Output : "How much would a woodchuck chuck"



[Chan et al. 2015] [Olah et Carter 2016]

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Overview

- Context & Vocabulary
- Unsupervised classification
- Explicit supervised classification
- Implicit supervised classification
- Deep Learning
- **Reinforcement Learning**

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

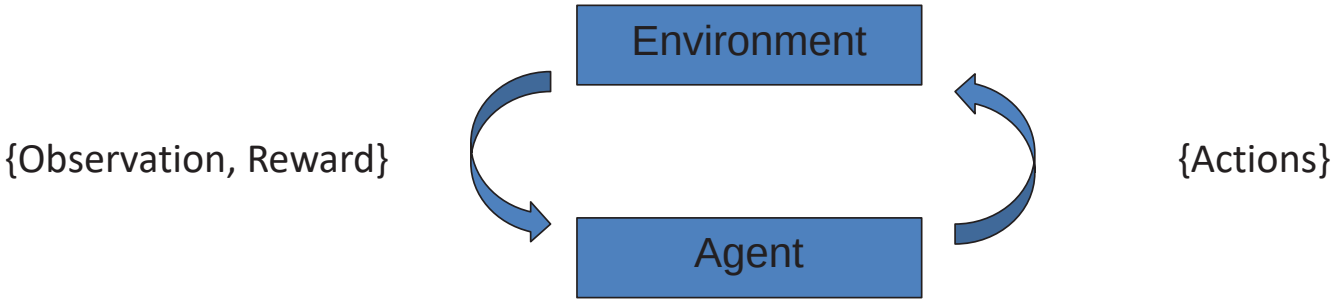
Based on "Introduction to Deep Learning", by Professor Qiang Yang, from The Hong Kong University of Science and Technology

REINFORCEMENT LEARNING

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Reinforcement Learning

- What's Reinforcement Learning?



- Agent interacts with an environment and learns by maximizing a scalar reward
- No labels or any other supervision
- Previously suffering from hand-craft states or representation

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Policies and Value Functions

- Policy π is a behavior function selecting actions given states
(it defines the probability of each possible action regarding the state *s*)

$$a = \underset{all\ possible\ actions}{\text{argmax}}\ \boldsymbol{\pi}(\text{s})$$

- Value function Q^{π}(s,a) is expected total reward r from state s and action a under policy π

$$Q^{\pi}(s,a) = \mathbb{E}\left[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots \mid s, a\right]$$

"How good is action a in state s?"

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Approaches To Reinforcement Learning

- **Policy-based RL**
  - Search directly for the optimal policy π^*
  - Policy achieving maximum future reward
- **Value-based RL**
  - Estimate the optimal value function Q^*(s,a)
  - Maximum value achievable for the best policy π^*
- **Model-based RL**
  - Build a transition model of the environment
  - Plan (e.g. by look-ahead) using model

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Bellman Equation

- Value function can be unrolled recursively

$$Q^{\pi}(s,a) = \mathbb{E}[r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots | s,a]$$
$$= \mathbb{E}_{s'}\left[r_t + \gamma \max_{a'} Q^{\pi}(s',a') | s,a\right]$$

- Optimal value function Q∗ (s, a) can be unrolled recursively

$$Q^{*}(s,a) = \mathbb{E}_{s'}\left[r_t + \gamma \max_{a'} Q^{*}(s',a') | s,a\right]$$
$$V_{\pi}(s) = \max_{a'} Q^{\pi}(s',a')$$

- Value iteration algorithms solve the Bellman equation

$$Q_{i+1}(s,a) = \mathbb{E}_{s'}\left[r_t + \gamma \max_{a'} Q_i(s',a') | s,a\right]$$

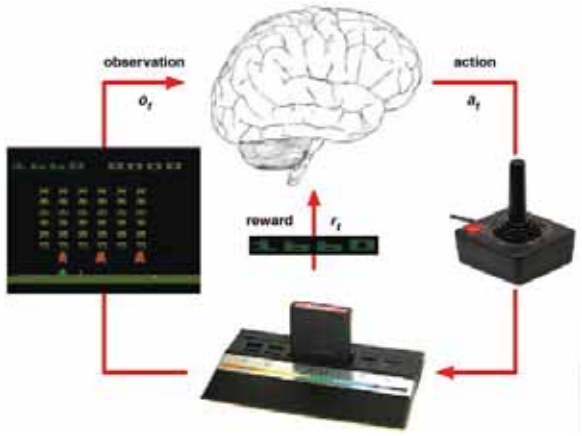
This last equation corresponds to how we should update Q ideally, i.e. knowing the state distribution BUT we do not know the state distribution Prob(s' | s,a) and the complexity is not even polynomial in the number of states

UNIVERSITÉ CÔTE D'AZUR

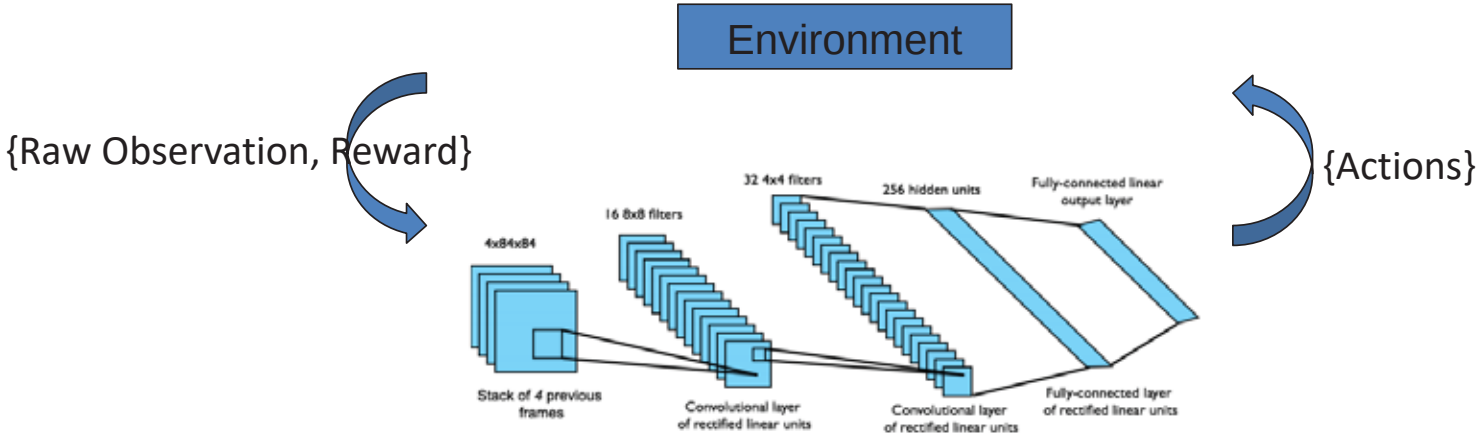
univ-cotedazur.fr

Deep Reinforcement Learning

- Human



- So what's **DEEP** RL?



UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Deep Reinforcement Learning

- Represent value function by deep Q-network with weights w

$$Q(s, a, \boldsymbol{w}) = Q^{\pi}(s, a)$$

- Define objective function by mean-squared error in Q-values

$$\mathcal{L}(\boldsymbol{w_i}) = \mathbb{E}\left[\underbrace{r_t + \gamma \max_{a'} Q(s', a', \boldsymbol{w_{i-1}})}_{\text{target}} - Q(s, a, \boldsymbol{w_i})\right]$$

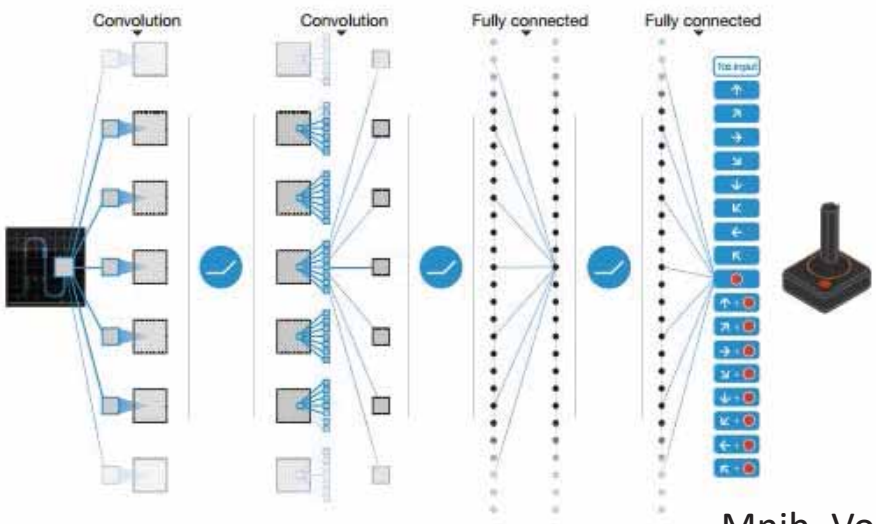
- Leading to the following Q-learning gradient

$$\frac{\partial \mathcal{L}(\boldsymbol{w_i})}{\partial \boldsymbol{w}} = \mathbb{E}\left[\left(\underbrace{r_t + \gamma \max_{a'} Q(s', a', \boldsymbol{w_{i-1}})}_{\text{target}} - Q(s, a, \boldsymbol{w_i})\right) \frac{\partial Q(s, a, \boldsymbol{w_i})}{\partial \boldsymbol{w}}\right]$$

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

DQN in Atari

- End-to-end learning of values $Q(s, a)$ from pixels
- Input state s is stack of raw pixels from last 4 frames
- Output is $Q(s, a)$ for 18 joystick/button positions
- Reward is the change in the score for that step

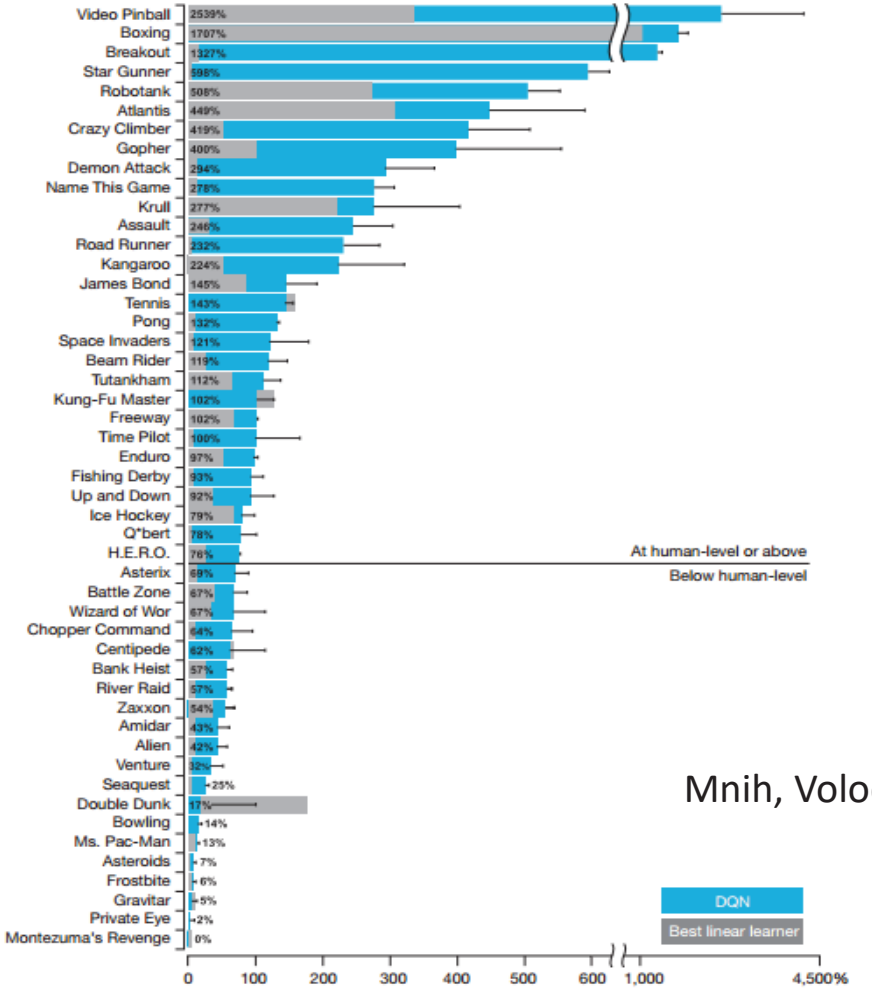


Mnih, Volodymyr, et al. 2015.

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

DQN in Atari : Human Level Control

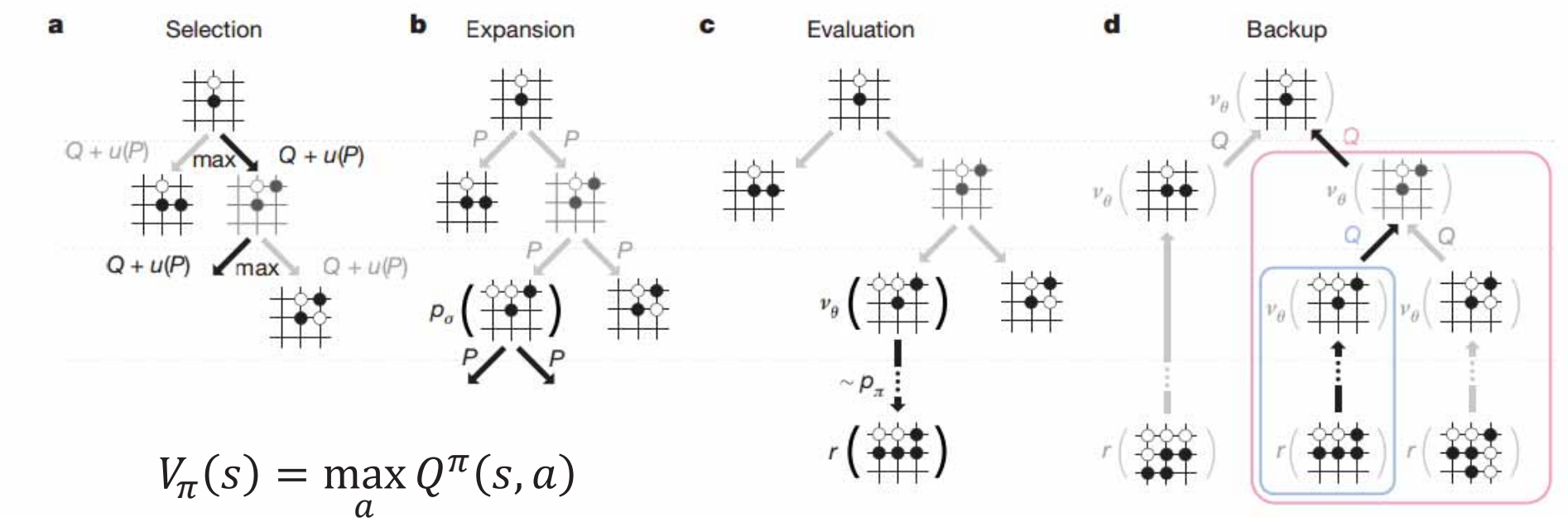


Mnih, Volodymyr, et al. 2015.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

AlphaGO: Monte Carlo Tree Search

- MCTS: Model look ahead to reduce searching space by predicting opponent's moves



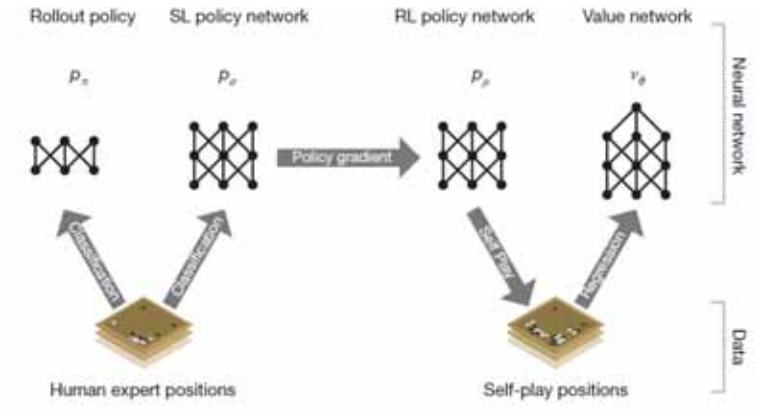
$$V_\pi(s) = \max_a Q^\pi(s,a)$$

Silver, David, et al. 2016.

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

AlphaGO: Learning Pipeline

- Combine SL and RL to learn the search direction in MCTS

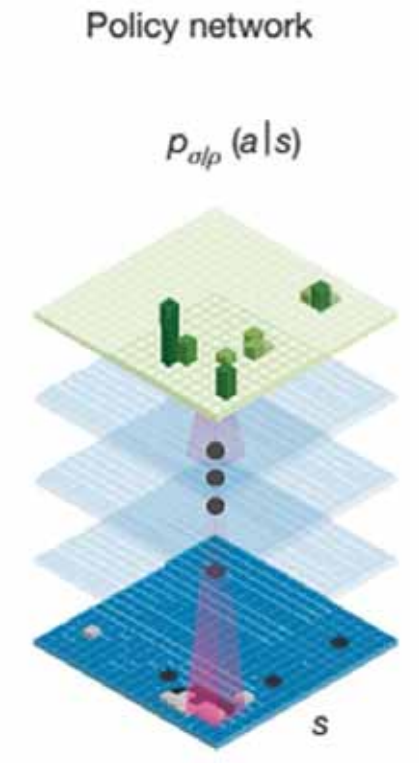


Silver, David, et al. 2016.

- SL policy Network
  - Prior search probability or potential
- Rollout:
  - combine with MCTS for quick simulation on leaf node
- Value Network:
  - Build the Global feeling on the leaf node situation

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Learning to Prune: SL Policy Network



- 13-layer CNN
- Input board position s
- Output: $\mathrm{p}_\sigma\ (a|s)$, where a is the next move

Extended Data Table 2 | Input features for neural networks

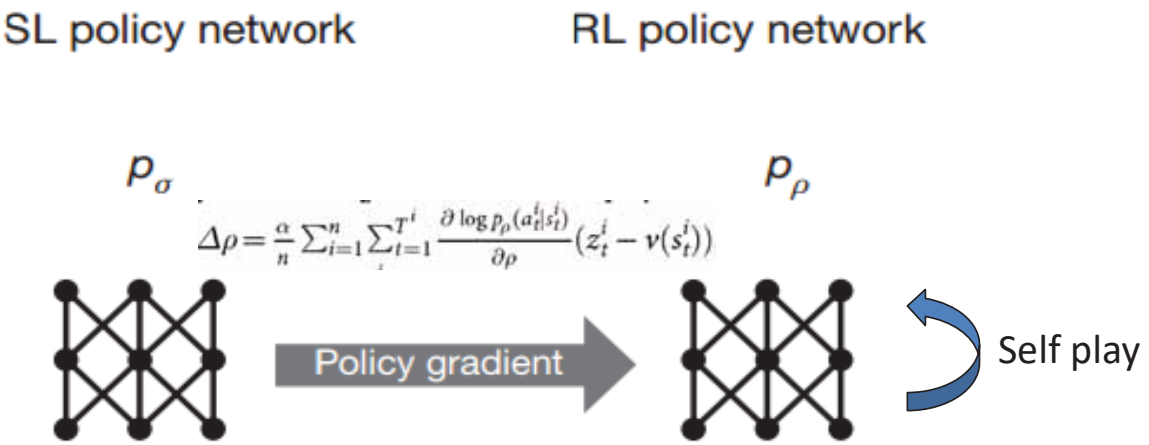
Feature	# of planes	Description
Stone colour	3	Player stone / opponent stone / empty
Ones	1	A constant plane filled with 1
Turns since	8	How many turns since a move was played
Liberties	8	Number of liberties (empty adjacent points)
Capture size	8	How many opponent stones would be captured
Self-atari size	8	How many of own stones would be captured
Liberties after move	8	Number of liberties after this move is played
Ladder capture	1	Whether a move at this point is a successful ladder capture
Ladder escape	1	Whether a move at this point is a successful ladder escape
Sensibleness	1	Whether a move is legal and does not fill its own eyes
Zeros	1	A constant plane filled with 0
Player color	1	Whether current player is black

Feature planes used by the policy network (all but last feature) and value network (all features).

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

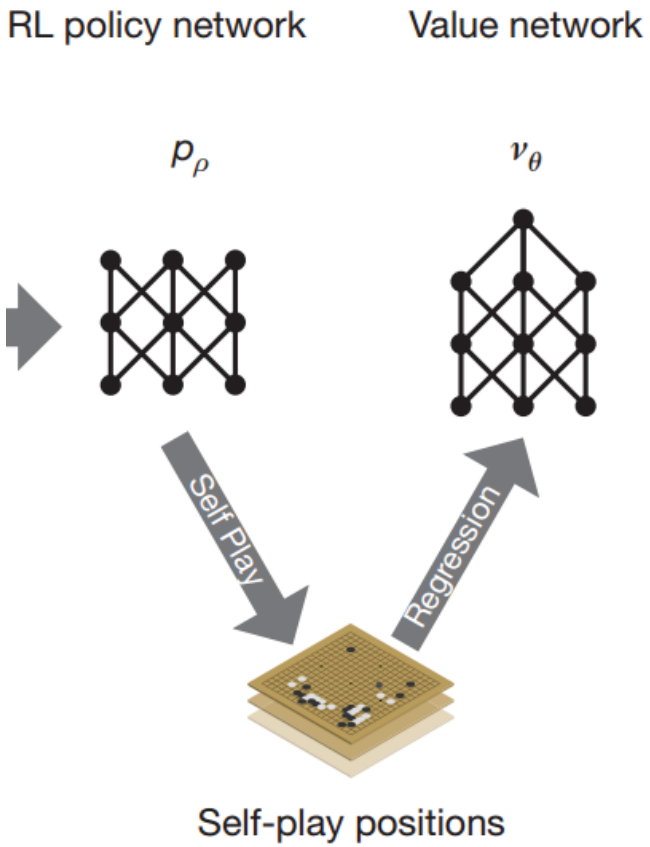
Learning to Prune: RL Policy Network



- 1 Million samples are used to train.
- RL-Policy network *VS* SL-Policy network.
  - RL-Policy alone wins 80% games against SL-Policy.
  - Combined with MCTS, SL-Policy network is better
- Used to derive the Value Network as the ground truth
  - Making enough data for training

UNIVERSITÉ CÔTE D'AZUR
univ-cotedazur.fr

Learning to Prune: Value Network



- Regression: Similar architecture

$$v^p(s) = \mathbb{E}[z_t | s_t = s,\ a_{t...T} \sim p]$$

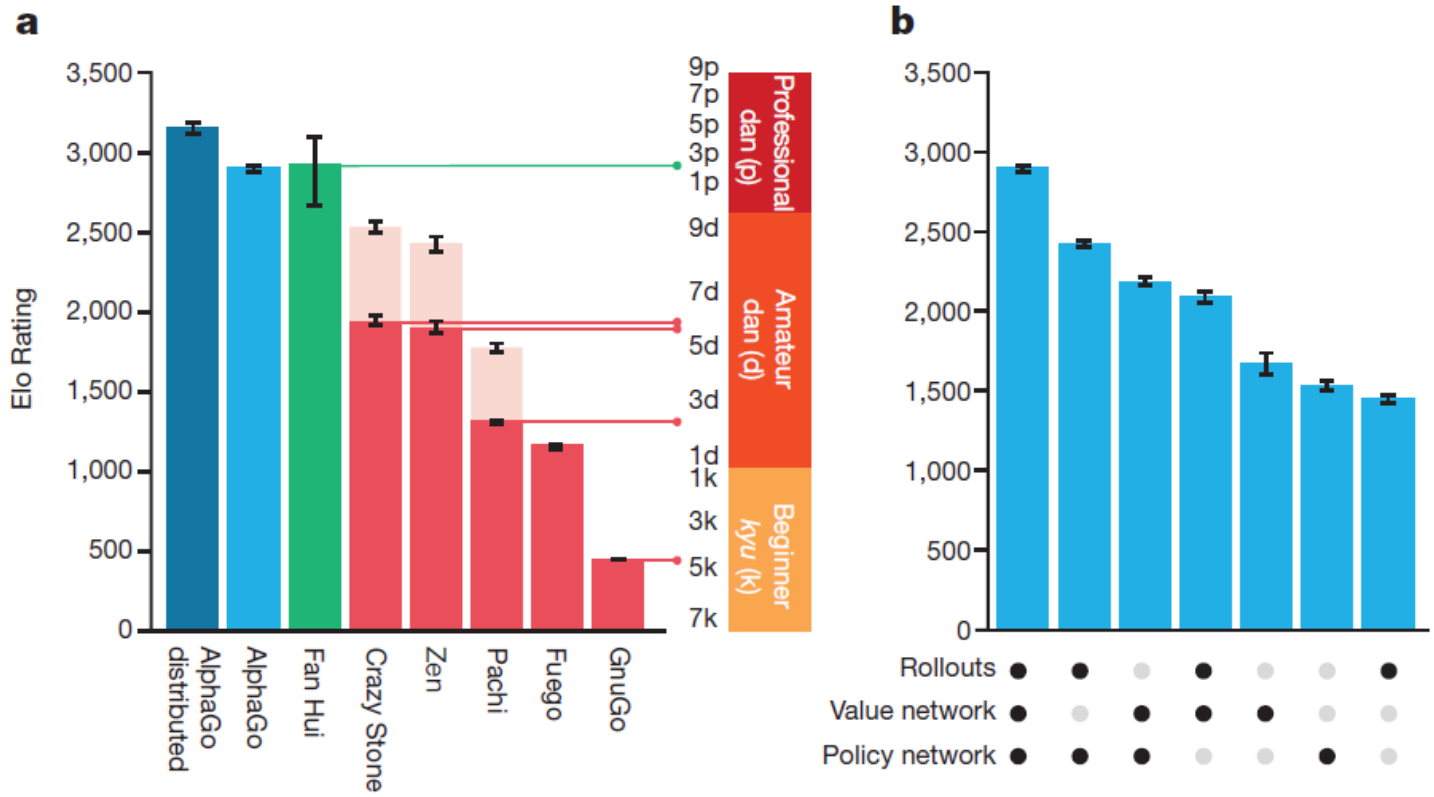
- SL Network: Sampling to generate a unique game.
- RL Network: Simulate to get the game's final result.

- Train: 50 million mini-batches of 32 positions (30 million unique games)

UNIVERSITÉ CÔTE D'AZUR

univ-cotedazur.fr

AlphaGO: Evaluation



The version solely using the policy network does not perform any search

Silver, David, et al. 2016.

UNIVERSITÉ
CÔTE D'AZUR

QUESTIONS?