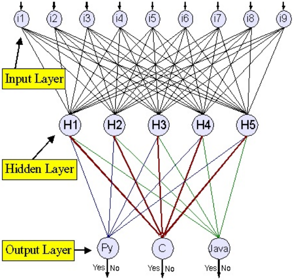
Artificial Intelligence:
A Brief Introduction



Academic team

- **Prof. Stephan Clémençon** (Telecom Paris - head of S2A)
  - Contact: `stephan.clemencon@telecom-paris.fr`
  - Profile: Research/Teaching/Consulting/Industry
  - Web: https://perso.telecom-paristech.fr/clemenco/
  - Expertises: machine learning (theory, algorithms and applications to e.g. security, finance, aeronautics, high tech, biosciences); stochastic processes
- **Hicham Janati** (Assoc. Prof at Telecom Paris - Group S2A)
  - Contact: `hicham.janati@telecom-paris.fr`
  - Expertises: machine learning, neurosciences, optimal transport
- **Matthieu Labeau** (Assoc. Prof at Telecom Paris - Group S2A)
  - Contact: `matthieu.labeau@telecom-paris.fr`
  - Expertises: machine/deep learning, natural language processing

The AI class

- Schedule
  - 14 weeks
  - Lectures on Tuesday/Thursday (04:15pm-05:30pm)
  - Labs on Thursday (02:30pm-04:00pm)
  - Zoom link for remote participation
  - Office hours: by appointment or on Thursday (05:30pm-06:30pm)
- Material
  - All the material (slides, Python notebooks) will be progressively available at https://aiteachings.github.io/NYU-AI-Fall22/ and on NYU brightspace
  - Textbooks are available on-line, reading recommendations/suggestions each week
- Evaluation: 3 labs will be graded (60%), quizz (40%) on Dec. 15

Machines with the ability to learn automatically and be smarter than us?

"AlphaGo knocks world champion Lee Se Dol out!"

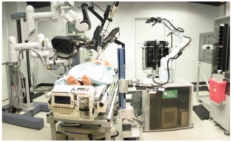
Machines with the ability to learn automatically and be smarter than us?

"Watson wins on 'Jeopardy'!"

Machines that listen or see better than us?

Intelligent systems/agents/robots that

- "read" et "categorize" digital documents
- recommend content, e.g. music, movies
- identify individuals using physical characteristics (biometrics)
- detect frauds, cyberattacks
- help pathologists diagnose a disease and find a treatment
- provide support for legal practice
- monitor the smooth operation of complex systems (*e.g.* smart grids, transportation networks)
- assist surgeons



- drive cars, explore oceans or planets...

Artificial Intelligence (AI):

the algorithms/concepts that enable machines to learn and perform tasks

During Fall, we shall try to offer you a tour in **Artificial Intelligence**

- Main **scientific principles** behind AI
- A combination of science and technology driven by the **applications**
- The nomenclature of **problems** solved by AI, popular **methods** and **use cases**
- A class **oriented towards applications**

Machine Learning (ML): now the main paradigm of AI research

The algorithms that enable machines to learn how to perform tasks

Modern technology enhances our capacity to

- collect, store and process **vast amounts of digital information**
- perform **sophisticated calculations**

Machines need **information** to learn

Information = Data

From raw digital information to machine-processable data

Digital information may take various forms

- rectangular arrays/tables of numbers (**vectors, matrices**)
- audio recordings, images, videos (**signals**)
- text (**semantics**)
- relationships between objects/individuals (**graphs**)

From raw digital information to machine-processable data

Digital information may take various forms

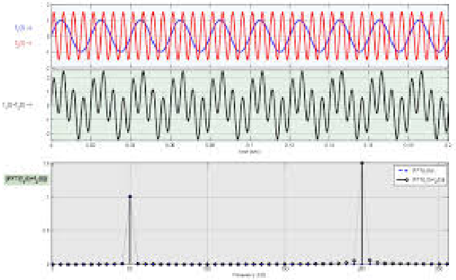
- rectangular arrays/tables of numbers (**vectors, matrices**)
- audio recordings, images, videos (**signals**)
- text (**semantics**)
- relationships between objects/individuals (**graphs**)

Mathematics provide **efficient representations** of the information

$\Rightarrow \quad X = (X^{(1)}, \ldots, X^{(D)})$ representation in a **vector space**

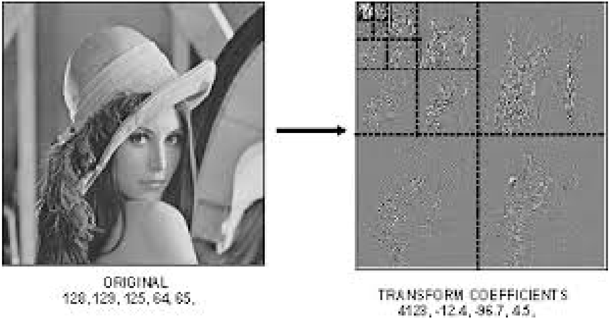
From raw digital information to machine-processable data

Audio signals et **Fourier transform**



From raw digital information to machine-processable data

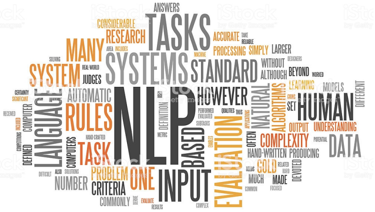
Images and **wavelet analysis** (jpeg 2000)

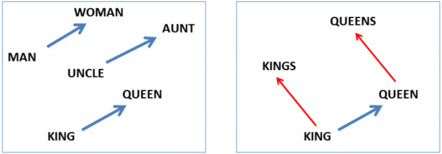


Medical imaging

Medical imaging

Natural Language Processing (NLP)





(Mikolov et al., NAACL HLT, 2013)

Representations through deep learning based '**word embedding**'

From data to 'intelligence': machine learning

- 'Intelligence'= ability to perfom tasks **autonomously** in a **predefined** framework with a **quantifiable** performance
- **General idea:** use machines to analyze massive datasets and elaborate decision/predictive rules, that can be implemented by machines as well
- **Examples** of tasks:
  - play a video game
  - turn speech into text
  - recognize the presence of a certain object in an image or a video
  - ...

Various disciplines involved in machine learning, mainly:

- Computer science
  - databases
  - algorithmics
  - knowledge representation
  - infrastructures, frameworks
  - ...
- Mathematics
  - **probability / statistics**
  - **optimisation**
  - linear algebra
  - numerical analysis
  - signal processing
  - ...
- Cognitive sciences, ...

The prototypical (supervised) task: pattern recognition

The machine tries to assign/predict a label Y ('the output') related to an object/individual described by a vector X ('the input')

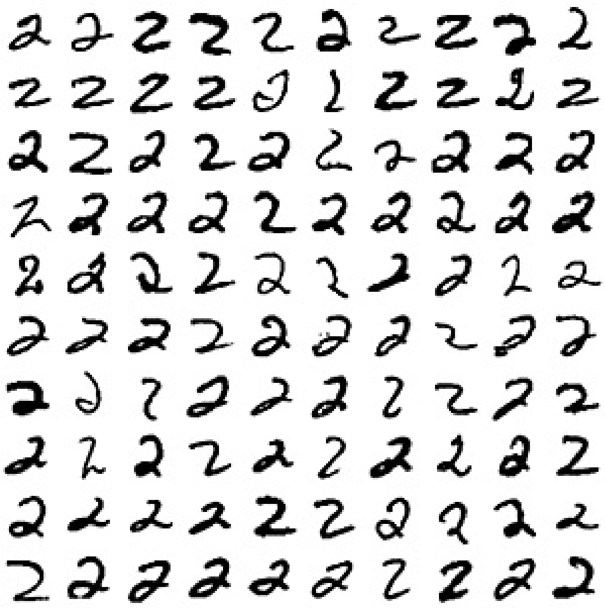
- **Handwritten digit recognition**

  - The input information $X \in \mathbb{R}^D$ is an array describing a pixelated image (*e.g.* $D = 640 \times 480$)
  - The label Y is a digit: $0, 1, \ldots, 9$
  - Databases $\mathcal{B}$ with many *labeled examples* are available

$$(X_1, Y_1), \ldots, (X_n, Y_n), \quad n = 60\,000$$

e.g. www.nist.gov

Pattern recognition - classification



Handwritten digit recognition

The machine uses the information gathered in $\mathcal{B}$ to learn how to **predict** the digit Y represented by any new handwritten character X

- The **predictive rule** takes the form of a **function**

$$g : x \in \mathbb{R}^D \mapsto g(x) \in \{0, \ldots, 9\}$$

- The computation of $g(x)$ must be **programmable** (*i.e.* machine feasible)

- Risk minimization: ideally, the rule g should minimize the probability of error

$$L(g) = \mathbb{P}\{Y \neq g(X)\}$$

over all possible pairs (X, Y) to which it could be applied in the future

Learning by observing: the machine ‘learns‘ from examples

- **The main paradigm of machine learning:** the frequentist principle of **empirical risk minimization**
- The empirical risk of rule g is the **frequency of errors** over the training examples

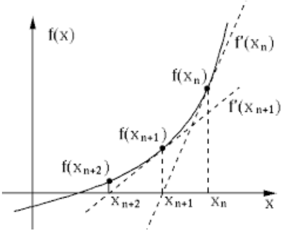
$$\widehat{L}_n(g) = \frac{\mathbb{I}\{Y_1 \neq g(X_1)\} + \ldots + \mathbb{I}\{Y_n \neq g(X_n)\}}{n}$$

- The machine searches for a predictive rule in a ‘catalogue’ $\mathcal{G}$ by solving the **minimization problem**

$$\min_{g \in \mathcal{G}} \widehat{L}_n(g)$$

Learning by optimizing

- **Continuous optimisation:** minimize the loss function incrementally by **gradient descent**



- **Discrete optimisation:** enumerate 'cleverly' using structural properties of the class $\mathcal{G}$

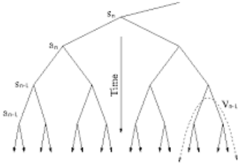
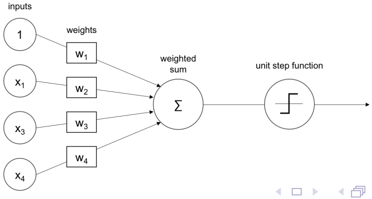


Figure 8.3: Hierarchical optimisation tree schematic. In each of the N stages, state s is realised before taking action a. Each path from top to bottom corresponds to a unique realisation of outcomes and decisions.

1957: the perceptron of F. Rosenblatt

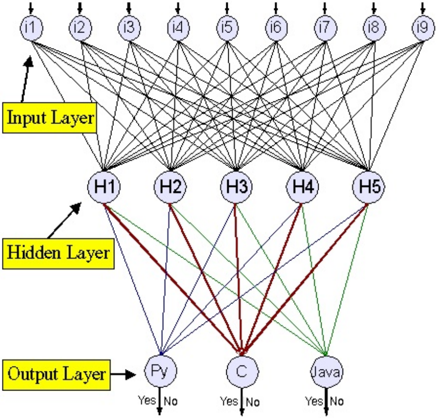
- The first machine learning algorithm. The machin: an IBM 704

- $Y \in \{-1, +1\}$ binary label - class $\mathcal{G}$ of linear rules $g(x) = sgn(a + \langle b, x \rangle)$ - incremental and on-line minimization by means of stochastic gradient descent



Modern machine learning algorithms

- (deep) neural networks



- support vector machines

Does the statistical learning principle work?

- The **theoretical** optimisation problem

$$\min_{g} L(g)$$

has been replaced by a **statistical** counterpart computed from the training data

$$\min_{g \in \mathcal{G}} \widehat{L}_n(g)$$

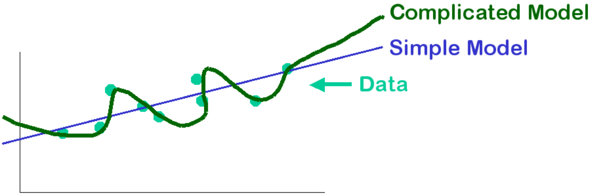
- The class $\mathcal{G}$ must be '**rich/complex**' enough to fit the data
- Probabilistic theory of statistical learning: the deviation between the risk $L(g)$ and the empirical version $\widehat{L}_n(g)$ must be 'small' uniformly over the class $\mathcal{G}$

Does the statistical learning principle work?

Predict labels of **past data**

vs

Predict labels of **future data**



Does the statistical learning principle work?

- Yes, whatever the distribution of the data, provided that the class $\mathcal{G}$ is **not too rich/complex**

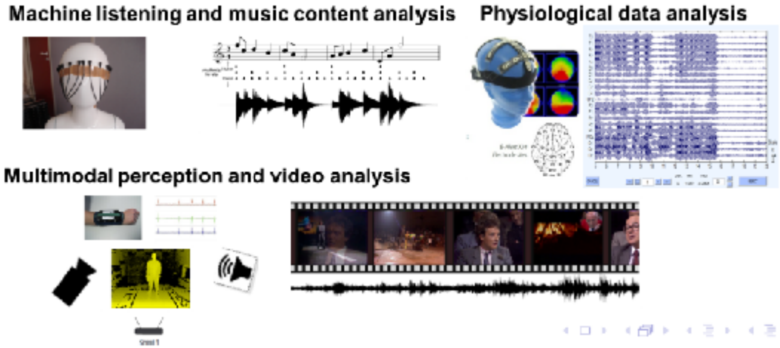
A. Chervonenkis & V. Vapnik

The statistical learning principle works!

With enough training data and computing power!

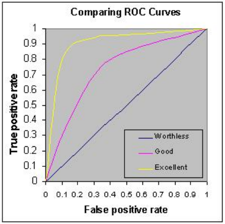
Numerous applications of this principle

Numerous applications of this principle



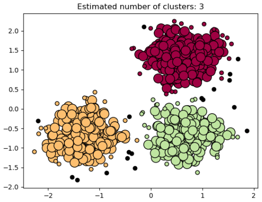
The ERM principle can be applied to many other problems

- **Regression:** continuous label $Y \in \mathbb{R}$ - quadratic risk $\mathbb{E}[(Y - g(X))^2]$ - ex: sales forecasting
- **Ordinal regression:** ordinal label $Y \in \{0, 5\}$ - quadratic risk $\mathbb{E}[(Y - g(X))^2]$ - ex: prediction in a five-star rating system
- **Scoring:** binary label Y - compute a score $s(X)$ that is an increasing transform of the posterior probability $\mathbb{P}\{Y = +1 \mid X\}$ - performance is measure by the ROC curve - ex: credit risk scoring

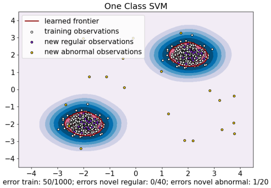


A wide variety of machine learning problems

- **Unsupervised problems:** 'learn without a teacher' - no label Y
  - **clustering**



  - **anomaly/novelty detection**



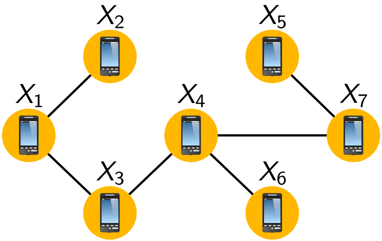
- Semi-supervised, on-line, multitask, one-pass, single/few-shot learning and many other problems . . .

Artificial Intelligence - A few milestones

- 1943: Artificial neuron model - McCulloch & Pitts
- 1957: Single layer perceptron - F. Rosenblatt
- 60's: Data-mining - John Tukey
- 1971: Probabilistic theory of statistical learning, V. Vapnik A. Chervonenkis
- 1974, 1986: Backpropagation and neural networks
- 80's: CART (Breiman, Friedman, Stone, Olshen), SVM
- 90's: A revival in ML algorithms: kernels, Boosting, random forests, *etc.* Before that the main paradigm in AI was symbolic AI
- 2000's: Web data, recommending systems, search engines, site retargeting, *etc.*
- 10's: The re-birth of neural networks (GPU, Deep Learning): computer vision, machine listening, ...

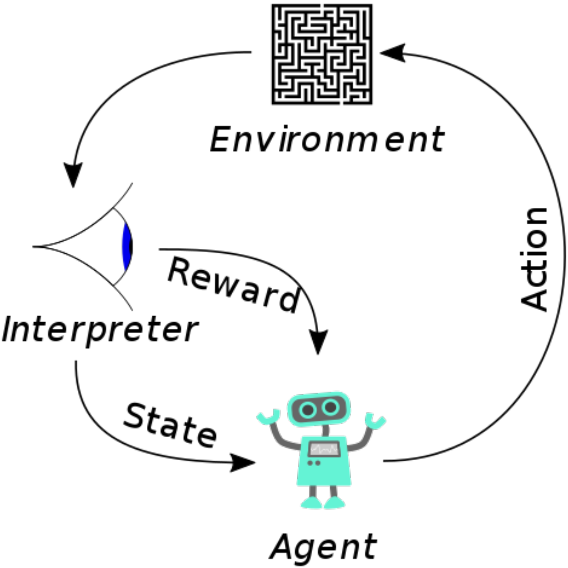
The end? No, an ongoing story!

- In the era of IoT... Can ML work in a distributed way efficiently?



- To deploy ML on a smartphone, compress deep neural networks?
- Learn how to detect weak signals (health monitoring)
- Many ongoing issues: interpretability, explainability, reliability/robustness, bias and fairness, privacy preserving ML, *etc.*

Reinforcement learning: when the machine must explore to learn



Our agenda

- **Pattern Recognition**: a 'basic' problem at the heart of many successful applications of AI
  - Algorithms: perceptron, linear discriminant analysis, naive Bayes, decision trees, neural nets, SVM, ensemble learning
  - Applications: computer vision, speech recognition, medical diagnosis, etc.
- **More advanced supervised problems**: similarity learning, collaborative filtering
  - Biometrics
  - Recommending systems
- **Learning without a teacher** (unsupervised problems)
  - Clustering, vector quantization
  - Anomaly/novelty detection
- **Learning on a trajectory**, reinforcement learning
  - Sequential learning, stochastic bandits
  - Q-learning
- **Natural Language Processing**
  - Symbolic AI
  - Deep Learning