



# 3I026 - INTRODUCTION À L'INTELLIGENCE ARTIFICIELLE ET DATA SCIENCE

Vincent Guigue  
Christophe Marsala

Sorbonne Université

- 1 L'UE 3I026**
- 2 IA et Data Science**
- 3 L'IA au service de la data science**
- 4 Rentrions dans le sujet**

L'UE 3I026

- Début des TDs et TMEs : semaine du **24 janvier**
  - Groupe 1 : Lundi 10h45-15h45
  - Groupe 2 : Mardi 16h-19h45
  - Groupe 3 : Mercredi 14h-18h
- Tous les groupes sont (plus que) pleins... Changements impossibles

A noter :

- Ressources = moodle (URL à venir)
- lisez régulièrement vos emails **prenom.nom@etu.sorbonne-universite.fr**

## ■ Intervenants

- Cours :

- Vincent Guigue : Vincent.Guigue@lip6.fr
- Christophe Marsala : Christophe.Marsala@lip6.fr

- avec le renfort de :

- Olivier Schwander

## ■ Prérequis

- Connaissance du langage Python
- Ne pas partir en courant en face d'une équation

- La présence au cours est **obligatoire**
  - **il n'y aura aucun rappel de cours en TME**
  - le cours doit être lu et travaillé **avant** d'aller en TME
- La présence au TME est **obligatoire** (note de CC)
  - les CR de TME seront notés (3 rendus principaux)
  - ils seront à rendre à la fin du TME
- Calcul de la note d'UE
  - examen : 50% de la note
  - contrôle continu : 50% de la note = (30% Projet de fin de semestre + 20% projet intermédiaire )

## ■ Nous retenons :

- 10 % de ce que nous lisons
- 20 % de ce que nous entendons
- 30 % de ce que nous voyons
- 50 % de ce que nous entendons et voyons
- 70 % de ce que nous pratiquons

- Nous retenons :

- 10 % de ce que nous lisons
- 20 % de ce que nous entendons
- 30 % de ce que nous voyons
- 50 % de ce que nous entendons et voyons
- 70 % de ce que nous pratiquons

- En suivant ce cours : **prenez des notes !**

- noter aide à **mémoriser...**

- En rentrant chez vous ou le lendemain :

- **mettez vos notes au propre**
- refaites les exemples vus
- allez sur la page de l'UE pour trouver une copie de ce cours

- **Avant les séances** (TD/TME, cours)

- **relisez le cours et refaites les exercices**

- Liens avec la recherche :
  - Département : Données et Apprentissage
  - Equipes : Machine Learning and Information Access + Learning Fuzzy and Intelligent System
  - Autres équipes : Base de données + Agents Cognitifs et Apprentissage
  - LPSM : Probas-stats
- Liens avec les master :
  - Master DAC - Données et Apprentissage
  - Master ANDROIDE - AgeNts Distribués, Robotique, Recherche Opérationnelle, Interaction, DEcision

**Quel point commun à presque toutes ces différentes équipes/masters ?**

# IA et Data Science

# Grandes périodes – 1. Fondations

- Le Machine Learning est né en même temps que les ordinateurs
- ML = ensemble d'outils venant des statistiques, de l'optimisation et de l'algorithme



Alan Turing : Test de Turing, 1950



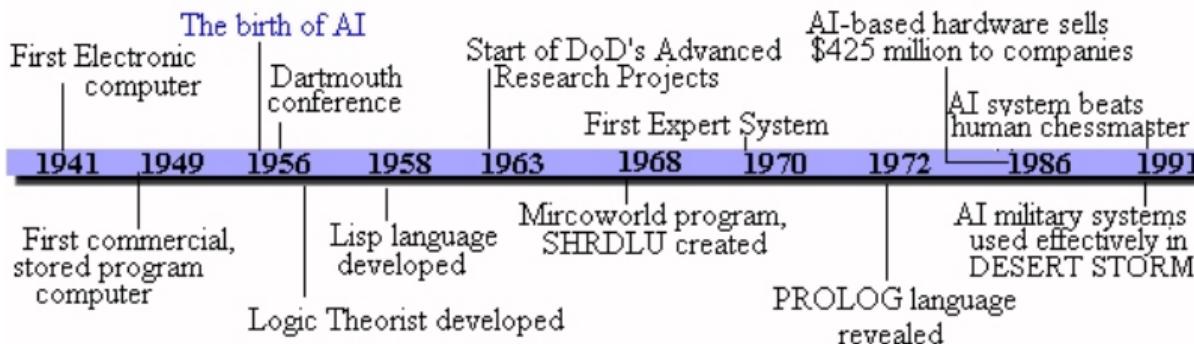
Arthur Samuel : Premier programme apprenant, 1956



Frank Rosenblatt : Perceptron, 1958

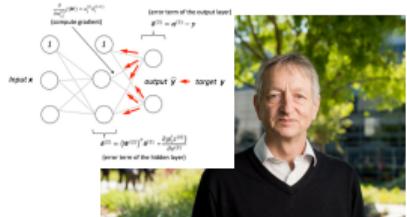


Kunihiro Fukushima : Réseaux de neurones modernes, 1979

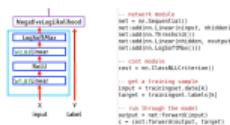


## Grandes périodes – 2. Proof of concept

- Premiers algorithmes pour les masses de données
  - Premiers succès industriels



## G. Hinton : Rétropropagation du gradient, 1986



Y. LeCun : premier succès industriel des RN, 1990

## Example Output from MUC-4 (1992)

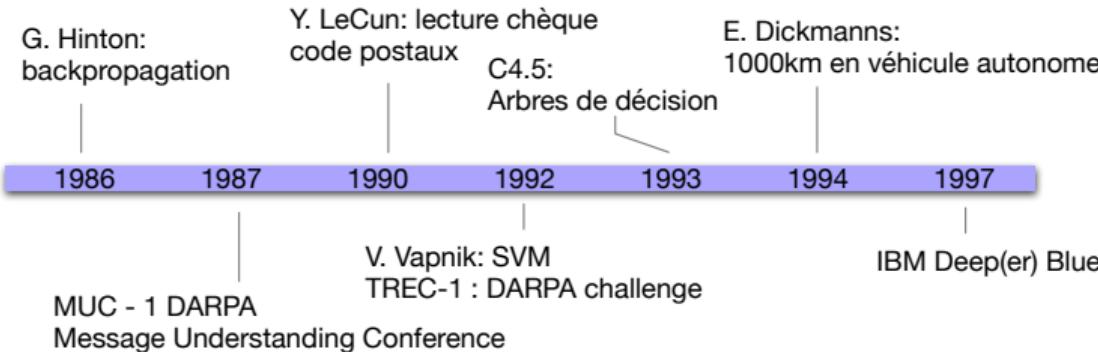
FIGURE 1. Extracted Terrorism Template

Template Slot ID	Fill Value
MESSAGE_ID	DEV-MUIC-00117 (SCC08C)
MESSAGE_TEMPLATE	3
INCIDENT_DATE	07 JAN 98
INCIDENT_LOCATION	CHILE: MOLINA (CITY)
INCIDENT_TYPE	ROBBERY
INCIDENT_STAGE_OF_EXECUTION	ACCOMPLISHED
INCIDENT_INSTRUMENT_ID	+
INCIDENT_INSTRUMENT_TYPE	SUN: ~
PERP_INVESTIGATION_CATEGORY	TERROIST ACT
PERP_INDIVIDUAL_ID	"ARMED INDIVIDUALS"/ "GROUP OF ARMED INDIVIDUALS WEARING SKI MASKS"/ "MEN"

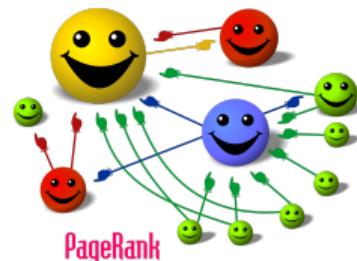
2

IBM : Deeper Blue

DARPA : TREC / MUC



- Multiplication des outils efficaces [& disgrâce des RN]
- Industrialisation des applications



Brin & Page : PageRank, 1998



Leo Breiman : Random Forest, 2001

 SYSTRAN

  
DRAGON  
NATURALLY SPEAKING



Lee & Seung : Non negative Matrix Factorization




T. Joachims:  
SVM light

1994 1997

WEKA

Lafferty:  
CRF

C.J. Lin  
libSVM

Pang & Lee  
Sentiment classification

L. Breiman  
Random Forest

Lee & Seung  
Factorisation matricielle

M. Jordan:

2001

2002

2004

2006 2007



J. Langford :  
Vowpal Wabbit

Netflix prize

DARPA:  
Grand Challenge

# Grandes périodes – 4. Succès & explosion

- Traitement des masses de données, parallélisation [Retour des RN]
- Multiplication des tâches, exploitation des réseaux sociaux



S. Thrun : 1er vainqueur DARPA  
Grand Challenge  
Pilier des futures Google car



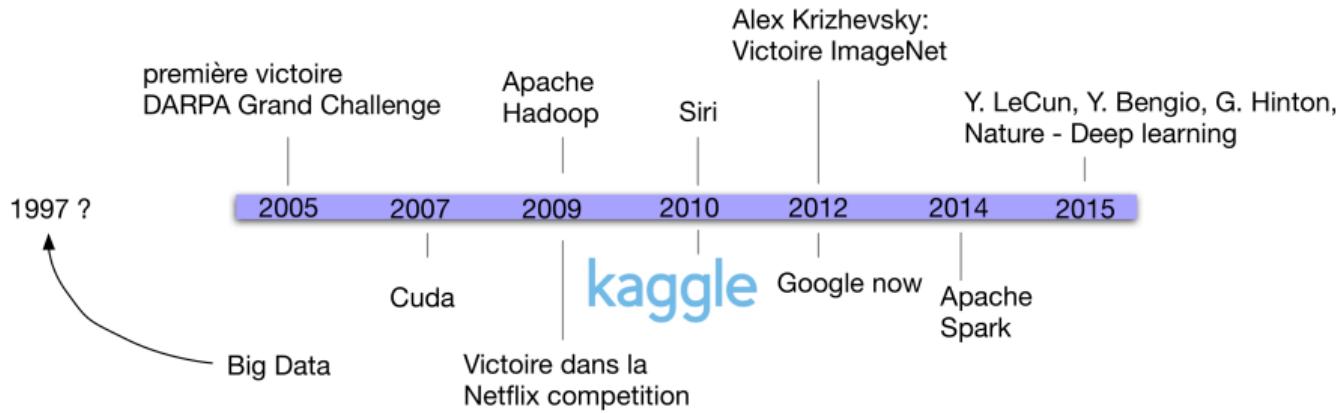
NVidia : CUDA, 2007



Algorithmes distribués, MapReduce

Rank	Name	Error rate	Description
1	U. Toronto	0.15315	Deep learning
2	U. Tokyo	0.26172	Hand-crafted
3	U. Oxford	0.26979	features and learning mod
4	Xerox/INRIA	0.27058	Bottleneck.

A. Krizhevsky, I. Sutskever, G. Hinton  
AlexNet



- Des outils matures, accessibles, partagés



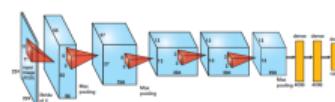
Python/C++, Scikit-learn  
(2007)



- Des architectures accessibles, des modèles pré-entraînés



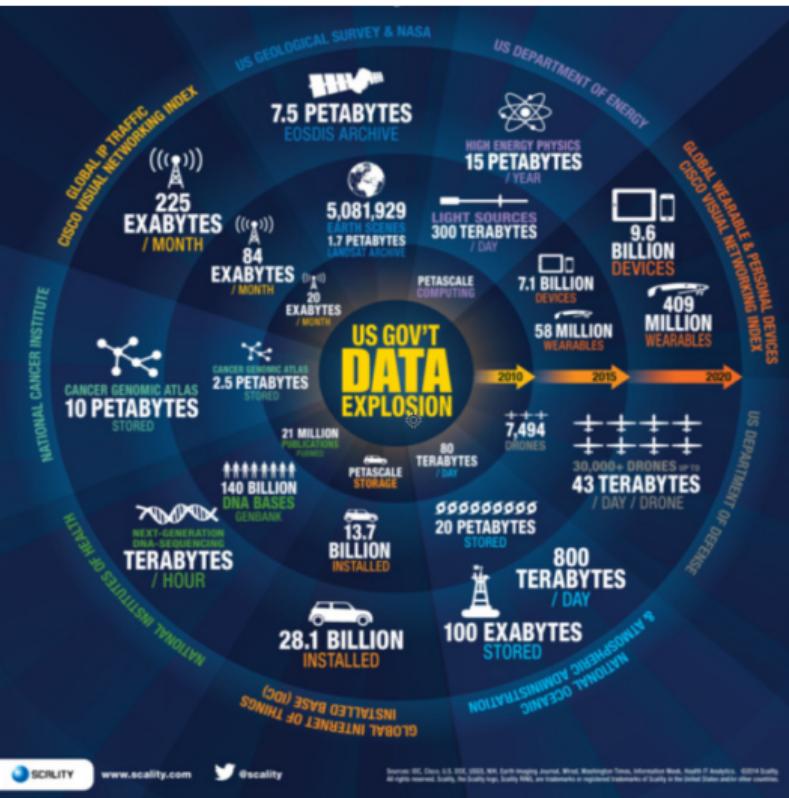
PC + carte graphique



Modèles pré-entraînés / expériences reproductibles

- Stockage bon marché
- ... Un alignement astral favorable !

- ↘ coûts des **capteurs**
  - Multiplication des sources
  - Politique **open-data**
- ↘ coûts de **stockage de données**
- ↘ coûts de **calcul**
  - CPU... Mais surtout GPU
- ↗ maturité (+support) des outils
  - Accès aux **algorithmes**



- IA (39-45) : Cryptographie, machine de Turing
- IA 56 : Tout algorithme malin
- IA 58 (a posteriori) : le perceptron de Rosenblatt
- IA 65 : Système expert
- IA 80 : invention des réseaux de neurones
- IA 80 bis : optimisation & logistique
- IA 2000 : Big data
- IA 2010 : Data-sciences
- IA 2020 : xAI

⇒ l'IA reflète des concepts radicalement différents, associées à des applications industrielles variables

## Une prise de conscience

- Un investissement public grandissant
- Un investissement massif des géants du Web



The screenshot shows a news article from Le Monde. At the top left, there's a navigation bar with links like "Actualités", "Dossiers", "Analyses & Réactions", "Habits", and "Aidez-nous". Below that, a sub-navigation bar includes "Blog", "Industrie", "Geopolitique", "HC Performance à l'assurance", and "Vos blog". The date "APRIL 10, 2016" is shown. The main headline reads "The Race For AI: Google, Facebook, Amazon, Apple In A Rush To Grab Artificial Intelligence Startups" with a sub-headline "Race To AI: Major Acquisitions In Artificial Intelligence". A "TECHNO" tag is visible. The main text discusses the race for AI startups between major tech companies. On the right side of the article, there's a sidebar titled "DOSSIER L'INTELLIGENCE ARTIFICIELLE C'EST DÉJÀ UN BUSINESS... ET CERTAINS L'ONT BIEN COMPRIS" which includes a sub-headline "Ruée sur l'intelligence artificielle. un business de 11 milliards de dollars en 2024". At the bottom left, there's a footer with "TOUTE L'ACTUALITÉ", "LOGICIEL", "MACHINE LEARNING", "SELON LE CEO DE GOOGLE, LE MACHINE ...", and the date "Le 22 Avril 2016". The bottom right of the article features another sidebar titled "DOSSIER INTELLIGENCE ARTIFICIELLE : RÉELLE MENACE OU TECHNOPHOBIE ?" with a sub-headline "Toyota investit 50 millions de dollars dans l'Intelligence artificielle".

- National Big Data R&D Initiative de la **maison blanche** en 2012 > 200 M \$
- NIST crée en 2013 le "Big Data Working Group"
- US National Research Council's Committee – National Academy of Sciences – Document de référence **Frontiers in Massive Data Analysis**
- Rapport Lauvergeon sur l'innovation : ambition 7 "La valorisation des données massives (Big Data)"
- CNRS : Les 10 instituts du CNRS mettent le Big Data comme une priorité – 2015 année des data sciences
- Appel(s) à projet Big Data
- ...

## Une prise de conscience

- Un investissement public grandissant
- Un investissement massif des géants du Web

### L'IA, priorité d'investissement des entreprises selon Gartner



**Google se préparerait à lancer un fonds d'investissement pour l'intelligence artificielle**

**INTELLIGENCE ARTIFICIELLE : LE SECTEUR DÉCOLLE, VOICI COMMENT INVESTIR**

PUBLIÉ LE 31/08/2017 À 14H12 | MIS À JOUR LE 05/09/2017 À 12H00

DOSSIER L'INTELLIGENCE ARTIFICIELLE C'EST DÉJÀ UN BUSINESS... ET CERTAINS L'ONT BIEN COMPRIS

Ruée sur l'intelligence artificielle... un business de 11 milliards de dollars en 2024

*AI accelerates in Canada: Record high investments in AI in the past five years, according to the MoneyTree Canada Report*

- Le Canada investit plusieurs centaines de millions de \$ dans le hub Montréal/Toronto, Google & Facebook suivent.
- Google a investi 400 M£ dans Deepmind – 2014
- Les géants du web investissent entre 20 et 30 Milliards de \$ dans l'IA en 2016
- D'après McKinsey, l'high-tech, les télécommunications et les services financiers seront les domaines en tête dans l'utilisation de l'IA au cours des trois prochaines années.

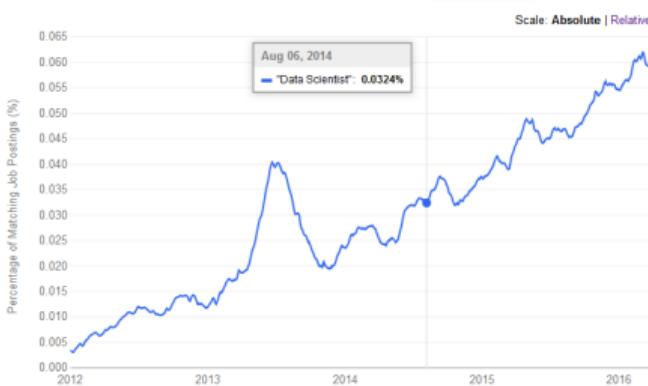
## Une formation académique en pleine explosion

- Un nombre croissant de formations
- Un nombre croissant d'étudiants A+ (= la nouvelle finance)

### Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

FROM THE OCTOBER 2012 ISSUE



- New York University : Center for Data Science.
- University of Washington : eScience Institute.
- Berkeley : Institute for Data Science.
- The Moore and Sloan foundations announce a five-year 37.8M\$ cross-institutional initiative to support the three previous institutes.
- Columbia : Institute for Data Sciences and Engineering.
- University of Rochester : 100M\$ commitment to create and house its Institute for Data Science.
- Center for Data Science Paris-Saclay
- SCAI - Sorbonne Université + DAC<sup>a</sup>

a. <http://dac.lip6.fr/master/>

## Data driven science : le 4e paradigme (Jim Gray - Prix Turing)

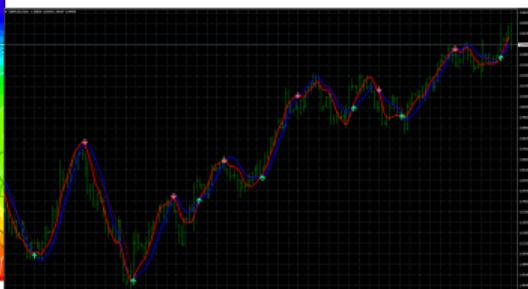
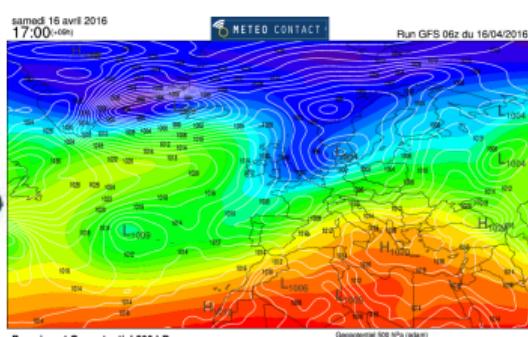
Extrait : "A l'heure actuelle, la science vit une révolution qui conduit à nouveau paradigme selon lequel 'la science est dans les données', autrement dit la connaissance émerge du traitement des données [...] **Le traitement de données et la gestion de connaissances représentent ainsi le quatrième pilier de la science après la théorie, l'expérimentation et la simulation.** L'extraction de connaissances à partir de grands volumes de données (en particulier quand le nombre de données est bien plus grand que la taille de l'échantillon) , l'apprentissage statistique, l'agrégation de données hétérogènes, la visualisation et la navigation dans de grands espaces de données et de connaissances sont autant d'instruments qui permettent d'observer des phénomènes, de valider des hypothèses, d'élaborer de nouveaux modèles ou de prendre des décisions en situation critique"

## Modélisation

- On va décrire (implémenter) dans le système les connaissances humaines du phénomène.
- Vision Algorithmique : Réaliser une tâche = décrire les différentes étapes nécessaires.
- Simulation = modéliser un phénomène physique complexe...  
et utiliser –quelques– données pour régler les degrés de libertés

## Data Science

- On va fournir au système un ensemble de données qui décrivent la tâche à faire
- +**Mécanisme général** pour apprendre à réaliser la tâche



- Learning is making useful changes in mind - [Marvin Minsky, 1985]
- Learning is any change in a system that allows it to perform better the second time on repetition of the same task or another task drawn from the same population - [Herbert Simon, 1983]
- Learning is the organization of experience - [Scott, 1983]
- Learning is constructing or modifying representations of what is being experienced - [Riszard Michalski, 1986]

## ■ Plusieurs définition (écoles) de l'apprentissage :

### ■ Cognitivistes :

- Postulat : Assimilation de l'organisme à un ordinateur, à une machine à traiter des informations
- Apprentissage : L'apprentissage vise à créer un nouveau programme

### ■ Théories écologiques :

- Postulat : Le comportement dynamique traduit l'adaptation du système aux contraintes qui pèsent sur lui
- Apprentissage : Intégrer de nouveaux états stables

### ■ Mathématicien :

- Postulat : tout peut se réduire à une formulation de problème d'optimisation
- Apprentissage : ... La résolution du problème d'optimisation

## ■ Informaticien / algorithmique

### ■ Algorithmes accumulatifs

Allez voir les spécialistes....

C'est quoi la science de l'apprentissage ?

- On étudie les apprentissages :
  - naturels
  - artificiels
- Des théories
- Des modèles
- Des Algorithmes

Différentes écoles :

- Les Symbolistes
- Les Connexionnistes (rejoint par les statisticiens)
  - Statistical Machine Learning

## Problématique :

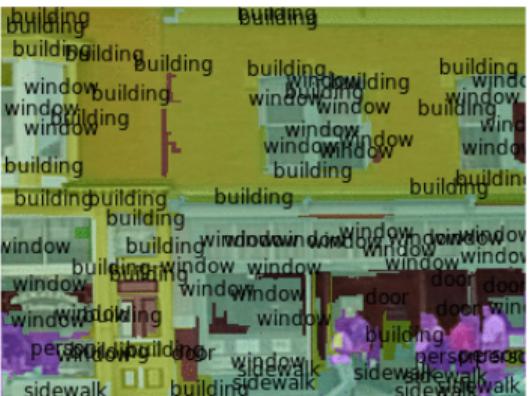
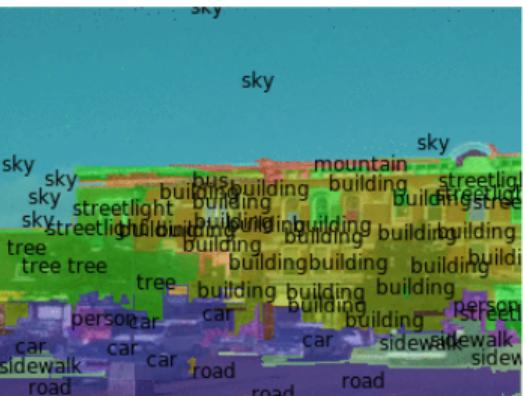
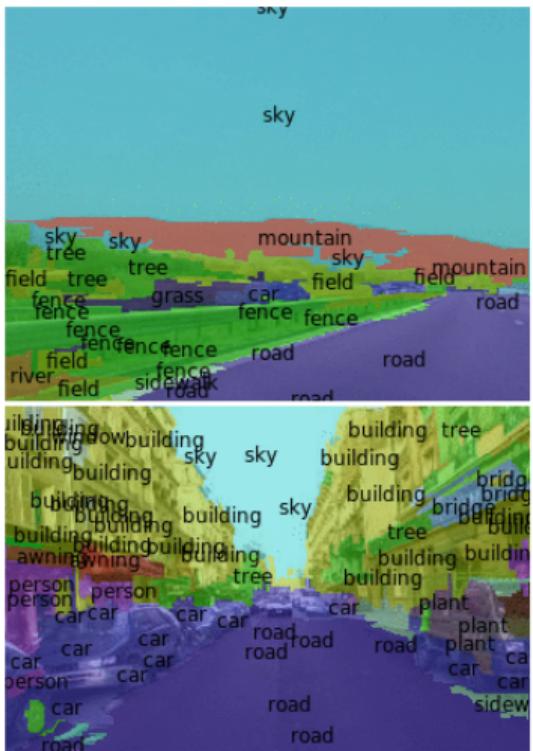
- Nous souhaitons avoir des ordinateurs
  - intelligents
  - adaptatifs
  - avec un comportement robuste
- Programmer de tels comportement est souvent impossible
  - Par exemple : Intelligence artificielle dans les jeux (scripts)

## Solution :

- Faire un ordinateur capable de se programmer lui-même
- à partir d'exemples (apprentissage classique / par imitation)
- à partir de son "expérience" (apprentissage par renforcement)

## Exercice

Ecrivez un programme JAVA permettant de ....



## Exercice

Ecrivez un programme JAVA permettant de ....



# Exercice

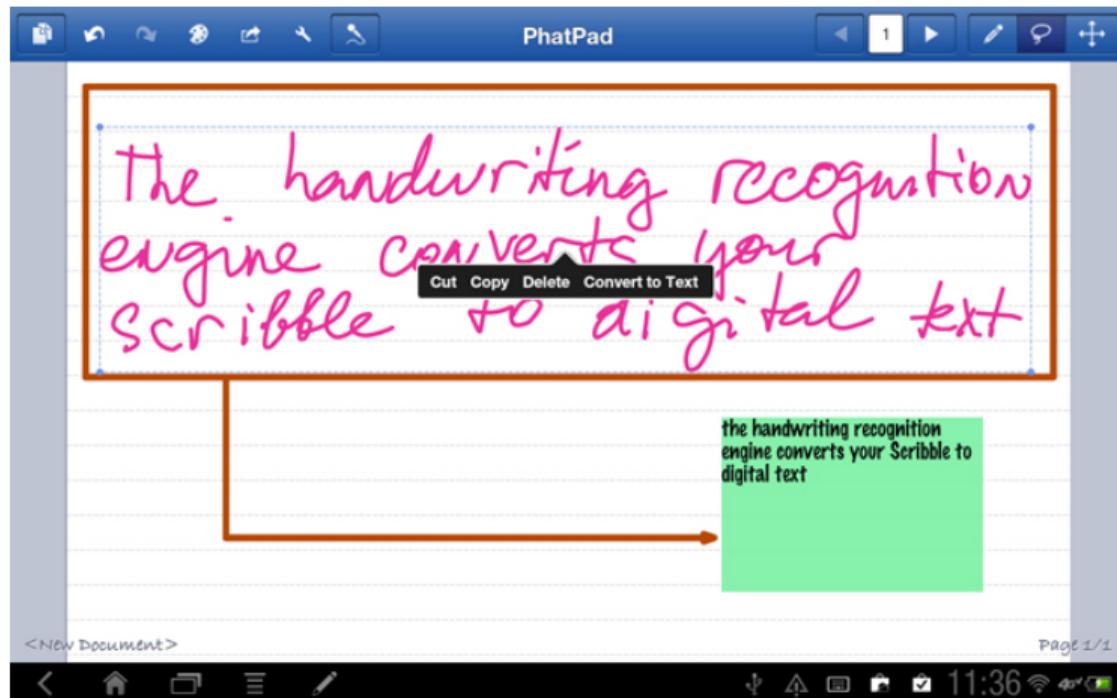
Ecrivez un programme JAVA permettant de ....



# Exemple

## Exercice

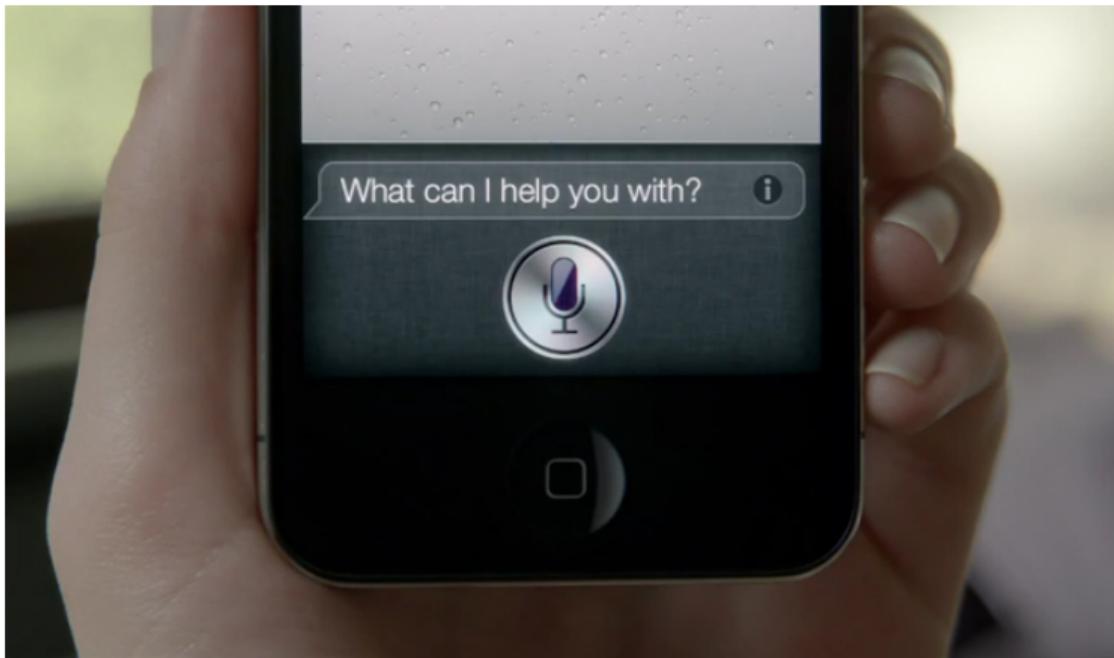
Ecrivez un programme JAVA permettant de ....



# Exemple

## Exercice

Ecrivez un programme JAVA permettant de ....



# Exemple

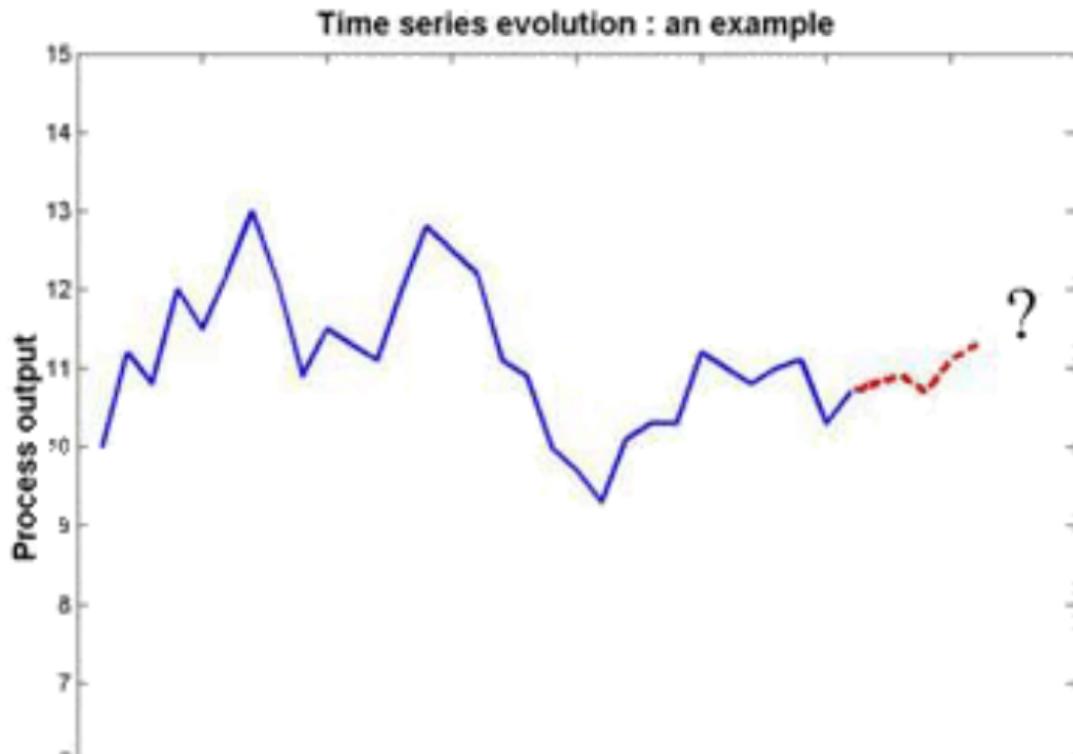
## Exercice

Ecrivez un programme JAVA permettant de ....



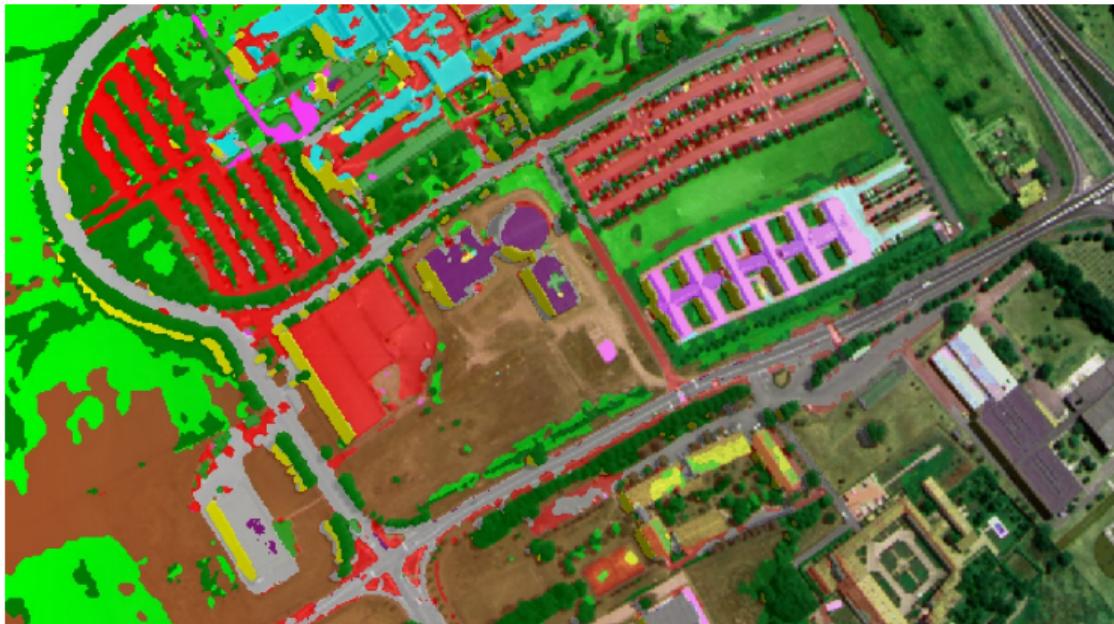
## Exercice

Ecrivez un programme JAVA permettant de ....



## Exercice

Ecrivez un programme JAVA permettant de ....



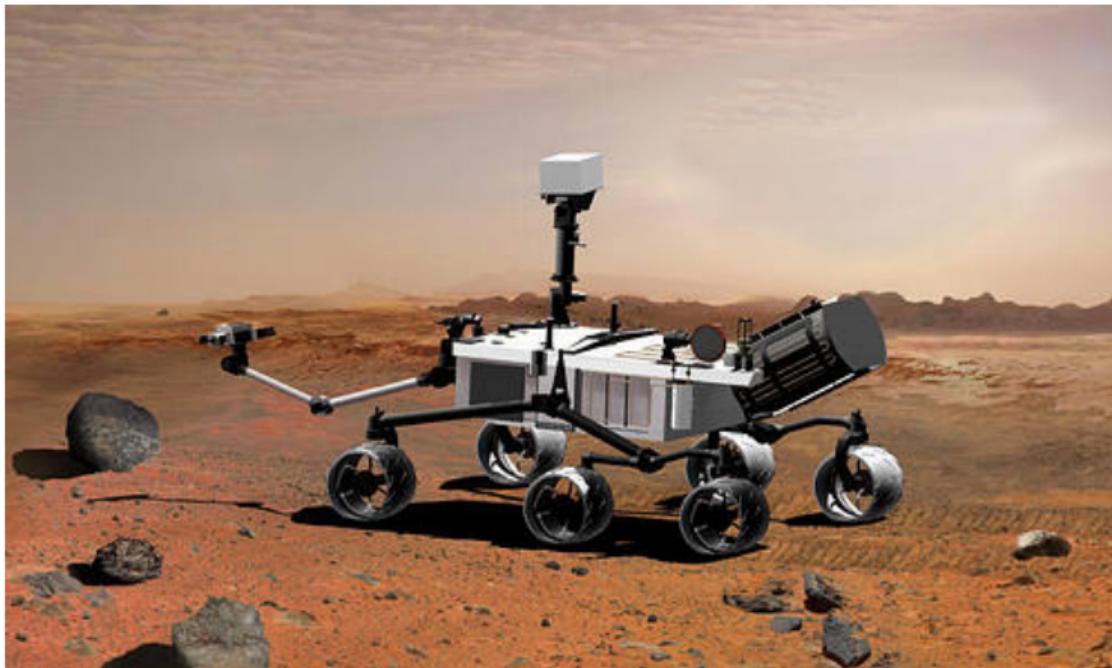
## Exercice

Ecrivez un programme JAVA permettant de ....



## Exercice

Ecrivez un programme JAVA permettant de ....



# Exemple

## Exercice

Ecrivez un programme JAVA permettant de ....

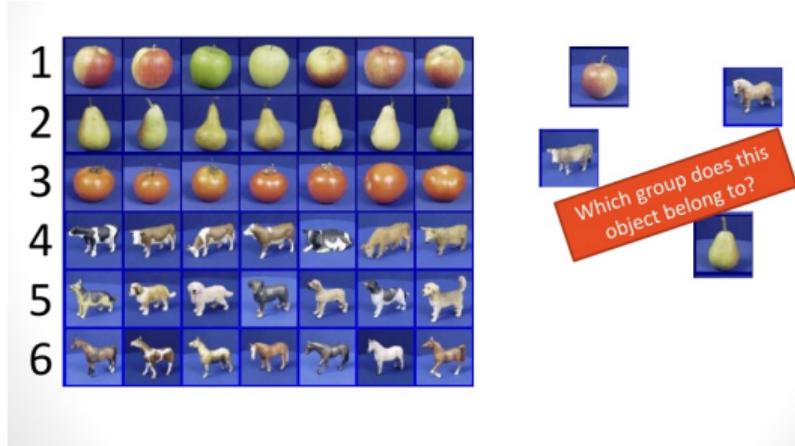


## Exercice

Ecrivez un programme JAVA permettant de ....



L'IA au service de la data science

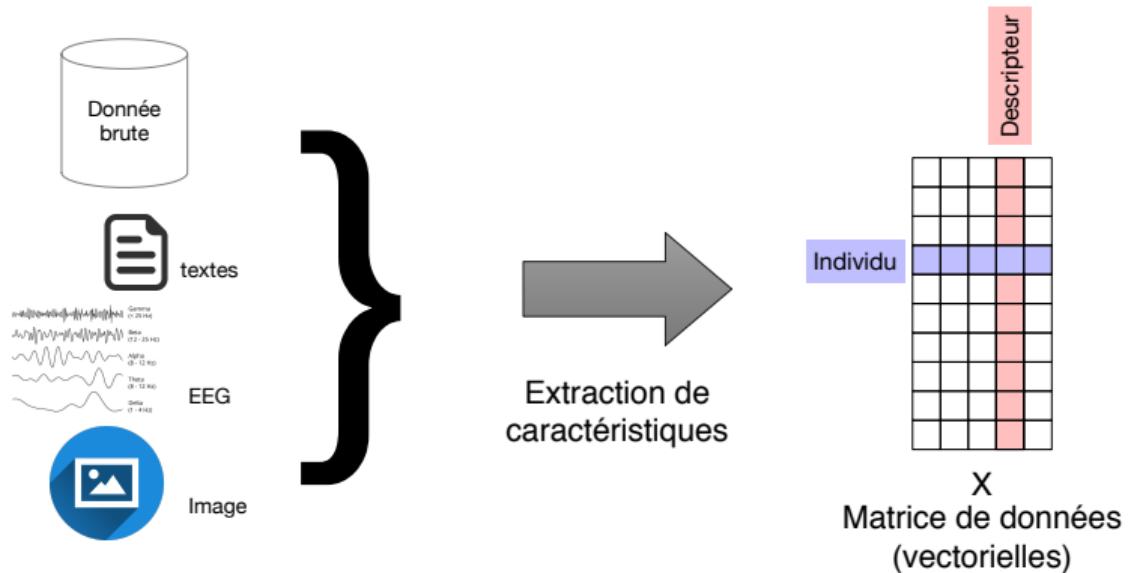


## Etapes

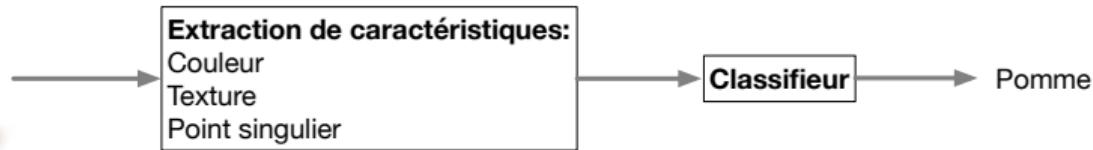
- Collecte des données (et éventuellement de la vérité terrain)
- Prétraitement des données = Extraction de caractéristiques
- Apprentissage (et détection) de modèles
- Evaluation
- Mise en production

# A la base du process : la donnée

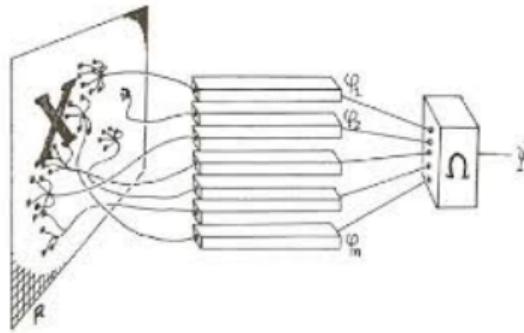
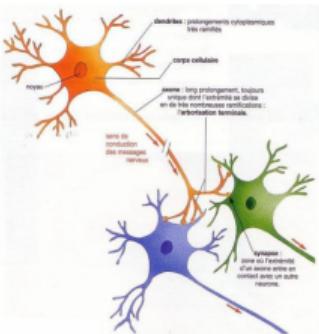
Cas plus standard : vectorisation



Exemple :



C'est le premier modèle formel de neurone (un seul neurone)



$$\text{Modèle} = y = f\left(\sum_{i=1}^n \pi_i x_i\right)$$

H.264 avi

H.264 avi

## ■ Eléments de base

- "Intelligence Artificielle : résolution de problèmes par l'homme et la machine", J.L. Laurière, 1987
- "Intelligence Artificielle", P. Winston, 1988
- "Principes d'intelligence artificielle", N. Nilsson, 1988
- "Artificial intelligence, a modern approach", S. Russel & P. Norvig, 1995

## ■ Livres plus généraux (format poche)

- "L'intelligence artificielle", J.-P. Haton & M.-C. Haton, 1993 (Que-sais-je ?)
- "Les sciences de l'artificiel", H. Simon, 1996 (Folio essais)
- "A la recherche de l'intelligence artificielle", D. Crevier, 1999 (Champs Flamm.)
- "La machine de Turing", A. Turing (& J.-P. Girard), ed. 1999 (Points Sciences)
- "L'intelligence artificielle", J.-G. Ganascia, 2007 (Le cavalier bleu)

## ■ Pour aller plus loin...

- "Gödel, Escher, Bach : Les Brins d'une Guirlande Éternelle", D. Hofstadter, 1979
- "Métaconnaissance : futur de l'intelligence artificielle", J. Pitrat, 1990
- "De la machine à l'intelligence artificielle", J. Pitrat, 1995

Rentrons dans le sujet

Soit des données  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ,  $\mathbf{x}_i \in \mathbb{R}^2$

Soit des étiquettes

$Y = \{y_1, y_2, \dots, y_N\}$ ,  $y \in \{1, -1\}$

Soit un classifieur linéaire

$$f(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i = \sum_j w_j x_{ij}$$

- 1 Dessiner le problème
- 2 Identifier les entrées et sorties du problème
- 3 Dessiner une frontière de décision
- 4 Trouver l'équation de cette frontière
- 5 Trouver les paramètres qui correspondent à cette frontière

Représenter les données  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ,  
 $\mathbf{x}_i \in \mathbb{R}^d$  dans une matrice

Soit un classifieur linéaire

$f(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i = \sum_j w_j x_{ij}$  Calculer  $f(\mathbf{x})$  pour tous les échantillons dans le cas où  $d = 2$  puis dans le cas général.

Représenter les données  $X = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ ,  
 $\mathbf{x}_i \in \mathbb{R}^d$  dans une matrice

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ \vdots & \ddots & & \\ x_{i1} & x_{i2} & \cdots & x_{id} \\ \vdots & \ddots & & \\ x_{N1} & x_{N2} & \cdots & x_{Nd} \end{bmatrix} \in \mathbb{R}^{N \times d}$$

Soit un classifieur linéaire

$f(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i = \sum_j w_j x_{ij}$  Calculer  $f(\mathbf{x})$  pour tous les échantillons dans le cas où  $d = 2$  puis dans le cas général.

Soit un classifieur linéaire

$f(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i = \sum_j w_j x_{ij}$  et des données  $X, Y$ ,  
comment optimiser les paramètres  
automatiquement ?

- 1** Qu'est ce que je cherche à faire ?
- 2** Comment je le formule ?
- 3** Comment je l'optimise ?

- Créer une liste +
- Créer une matrice +
- Additionner deux listes ≈
- Faire un produit matriciel –
- Calculer des moyennes, des écarts-types,  
enchaîner les manipulations sur des  
matrices... — — —

- Créer une liste +
- Créer une matrice +
- Additionner deux listes ≈
- Faire un produit matriciel -
- Calculer des moyennes, des écarts-types,  
enchaîner les manipulations sur des  
matrices... - - -

⇒ En route vers numpy