INTRODUCTION AU MACHINE LEARNING

Vincent Levorato

Consultant Data Scientist
Docteur en Informatique et chercheur associé au LIFO

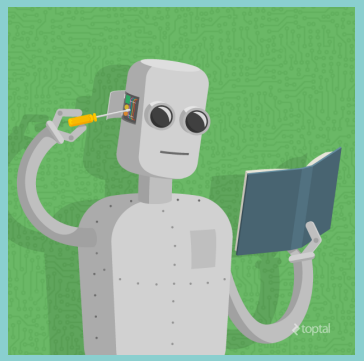
@VincentLevorato

29 MARS 2016 – MEETUP ORLEANS TECH TALKS - HTTP://ORLEANS-TECH.COM

SOMMAIRE

- Machine Learning ?
- Mythes et légendes
- Catégories
  - Apprentissage *supervisé*
  - Apprentissage *non-supervisé*
- Démo
- To continue...

01 > MACHINE LEARNING



WHAT IS MACHINE LEARNING ?

- Machine Learning ≠ Data Science

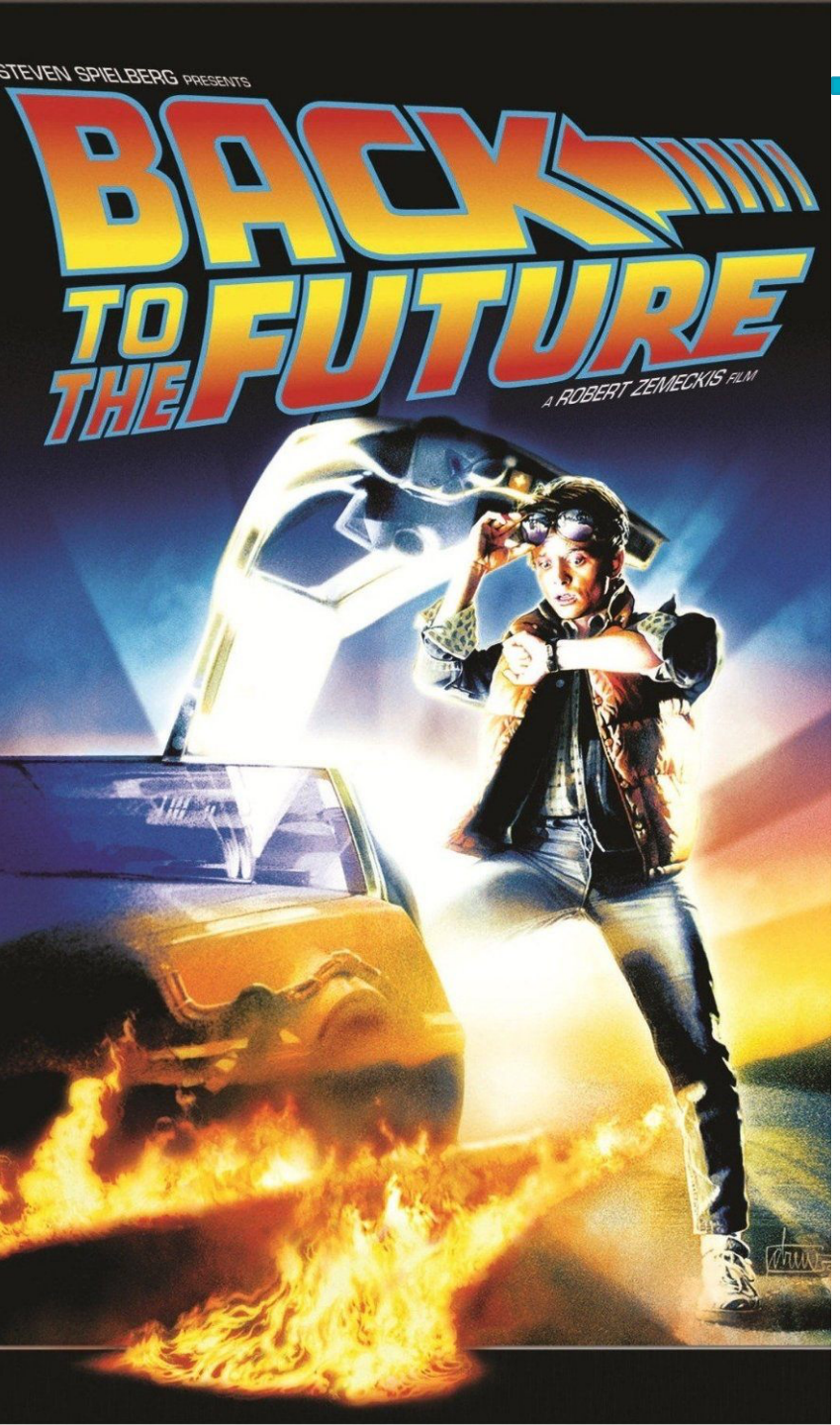
- En français: **Apprentissage Automatique**
  (~~Machine~~ Learning → Artificial Learning ≠ Natural Learning)

- Apprentissage naturel
  - Résolution de problèmes, Essai-erreurs (renforcement), par cœur, etc.

- L'apprentissage automatique (ou artificiel), qu'est-ce que c'est vraiment ?
  - Un champ d'études de la recherche académique en Informatique
  - Des algorithmes et des approches basées sur les mathématiques appliquées
  - Un sous-ensemble de l'intelligence artificielle

L'objectif du Machine Learning est de concevoir des programmes pouvant s'améliorer automatiquement avec l'expérience.

Céline Rouveirol, LIPN, Paris 13



MACHINE LEARNING IS HYPE !

- Oui, mais non (cf Mythes et légendes)
- Le Machine Learning, c'est ~~nouveau~~.
- Il n'y avait rien avant… non ?
- C'est le Big Data ! J'ai bon là ?

HISTORIQUE - STATISTIQUES

Définitions

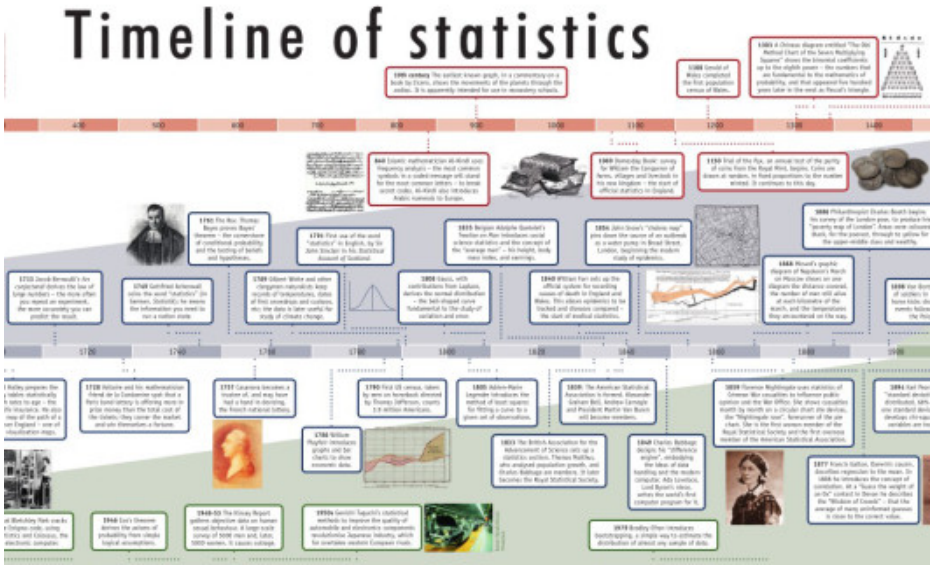
- *La Statistique* est une branche des Mathématiques qui s'est développée depuis la fin du XIXème siècle, notamment à la suite des travaux de l'école anglaise (K. Pearson, W. Gosset (Student), R. Fisher, J. Neyman...).
- Une *statistique* est une quantité définie par rapport à un modèle permettant d'inférer sur son comportement dans une situation expérimentale donnée.
- Les *statistiques* avec un « s »: apparu au XVIIIème siècle pour désigner des quantités numériques.

- Différentes familles:
  - La Statistique **descriptive**
    - représentations graphiques, résultats numériques
    - descriptif
  - La Statistique **inférentielle**
    - préciser un phénomène sur une population globale, à partir de l'observation d'un échantillon de celle-ci, modèles et hypothèses probabilistes
    - explicatif
  - L'**apprentissage** statistique
    - modèle privilégiant la prévision d'une variable ou quantitative

Philippe Besse , Exploration Statistique Multidimensionnelle, INSA

HISTORIQUE - STATISTIQUES

- Quelques faits historiques:
  - **- 450**: Hippias d'Elis a estimé la première date des jeux olympiques par la moyenne de la durée des règnes des différents rois
  - **840**: Al-Kindi, mathématicien arabe, utilise l'analyse par fréquences pour casser des codes secrets
  - **1069:** Domesday Book: rapport pour Guillaume le conquérant de l'état de son royaume
  - **1346**: Giovanni Villani propose des statistiques sur la population et les affaires de la ville de Florence
  - **1560**: Gerolamo Cardano calcule les probabilités de tirage pour les joueurs de dés
  - **1654**: Pascal et Fermat réfléchissent au partage des gains dans les jeux de hasard → théorie des probabilités
  - **1713**: Jacob Bernouilli → la loi des grands nombres
  - **1805**: Adrien-Marie Legendre → méthode des moindres carrés
  - **1808:** Gauss (et Laplace) → distribution de la loi normale
  - **1894**: Karl Pearson → écart-type
  - **XXème siècle**: Bachelier, Gossett, Zipf, Fisher, Neyman, Cox, Shannon, etc.



http://www.statsref.com

HISTORIQUE - STATISTIQUES

En bref – Probabilités et Statistiques

Théorie des probabilités

- Fondements
  - Histoire des probabilités
  - Axiomes des probabilités, Probabilité, Événement, Tribu, Espace probabilisé
  - Indépendance
- Probabilités élémentaires
  - Moment, Espérance, Variance, Écart type
  - Médiane, Quartile
- Loi de probabilité
  - Variable aléatoire, Variables iid
  - Loi de Bernoulli, Loi de Poisson, Loi uniforme, Loi normale
- Convergence de lois
  - Théorème central limite, Loi des grands nombres, Théorème de Borel-Cantelli
- Calcul stochastique
  - Marche aléatoire, Chaîne de Markov
  - Processus stochastique, Processus de Markov, Mouvement brownien
  - Martingale
  - Équation différentielle stochastique

Statistique

- Statistique descriptive
  - Échantillon
  - Moyenne, Quantile, Erreur type
  - Intervalle de confiance
  - Représentations de données, Histogramme
  - Régression linéaire, Méthode des moindres carrés
  - Analyse des données
- Statistique mathématique
  - Une statistique
  - Fonction de répartition empirique
  - Théorème de Glivenko-Cantelli
  - Inférence bayésienne
  - Maîtrise statistique des procédés
- Tests statistiques
  - Test d'hypothèse
  - Hypothèse statistique (Hypothèse nulle)
  - Estimateur
  - Test du χ^2, Test de Student, Test de Fisher
  - Signification statistique
  - Valeur p

Applications

- En théorie des probabilités
  - La mécanique quantique
  - Les probabilités et la biologie
  - Les probabilités et l'économie
  - Les probabilités et les jeux
  - La théorie de l'information
- En statistique
  - La physique statistique
  - la science actuarielle
  - Les statistiques et la biologie
  - Les statistiques et l'économie
  - La fouille de textes
  - Les statistiques et la psychologie
  - Science des données

https://fr.wikipedia.org/wiki/Portail:Probabilités_et_statistiques

WHAT IS MACHINE LEARNING ?

Finalement...

- Une définition formelle: *"A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E."* (Tom Mitchell, 1997)

- Appartient au domaine vaste de l'IA.

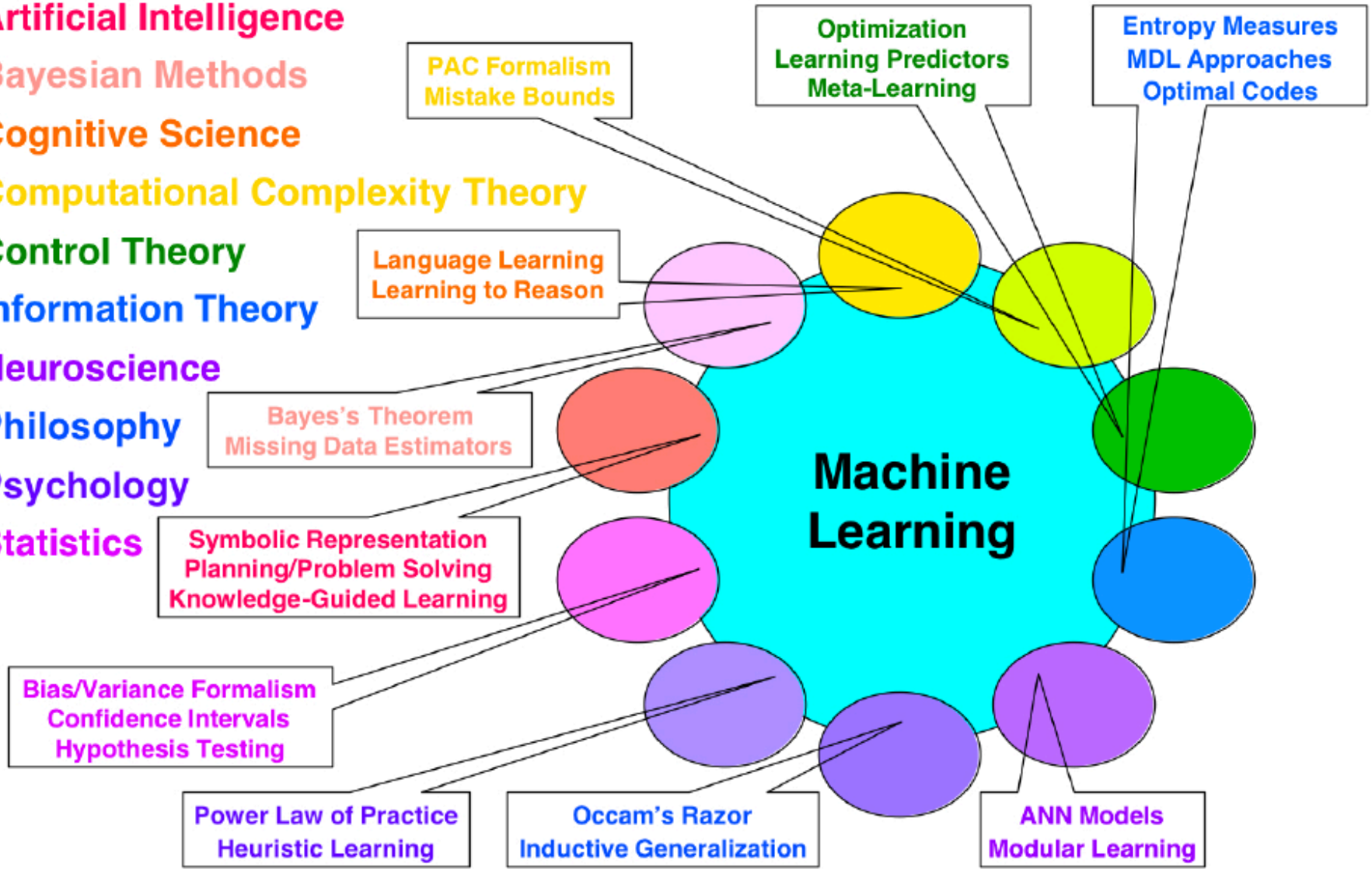
- Finalement, qu'est-ce qu'il y a dans le ML ?
  - Des algorithmes
  - Des mathématiques appliquées (et donc des statistiques)
  - Une mise en œuvre sur machine sous la forme de programmes

- Le Machine Learning fait partie de la science Informatique.

- Dans le monde réel:
  - Data Science = Predictive Analytics = Data Mining
  - Extraction de connaissances (KDD: Knowledge Discovery in Databases)

MACHINE LEARNING & CO.

- Artificial Intelligence
- Bayesian Methods
- Cognitive Science
- Computational Complexity Theory
- Control Theory
- Information Theory
- Neuroscience
- Philosophy
- Psychology
- Statistics



Céline Rouveirol, LIPN, Paris 13

MACHINE LEARNING & CO.

Plus concrètement, quelques applications

- Adaptive websites
- Affective computing
- Bioinformatics
- Brain-machine interfaces
- Cheminformatics
- Classifying DNA sequences
- Computational finance
- Computer vision
- Detecting frauds
- Game playing
- Information retrieval
- Internet fraud detection
- Marketing
- Machine perception
- Medical diagnosis
- Natural language processing
- Optimization and metaheuristic
- Online advertising
- Recommender systems
- Robot locomotion
- Search engines
- Sentiment analysis
- Sequence mining
- Software engineering
- Speech and handwriting recognition
- Stock market analysis
- Structural health monitoring
- Syntactic pattern recognition
- Economics and Finance

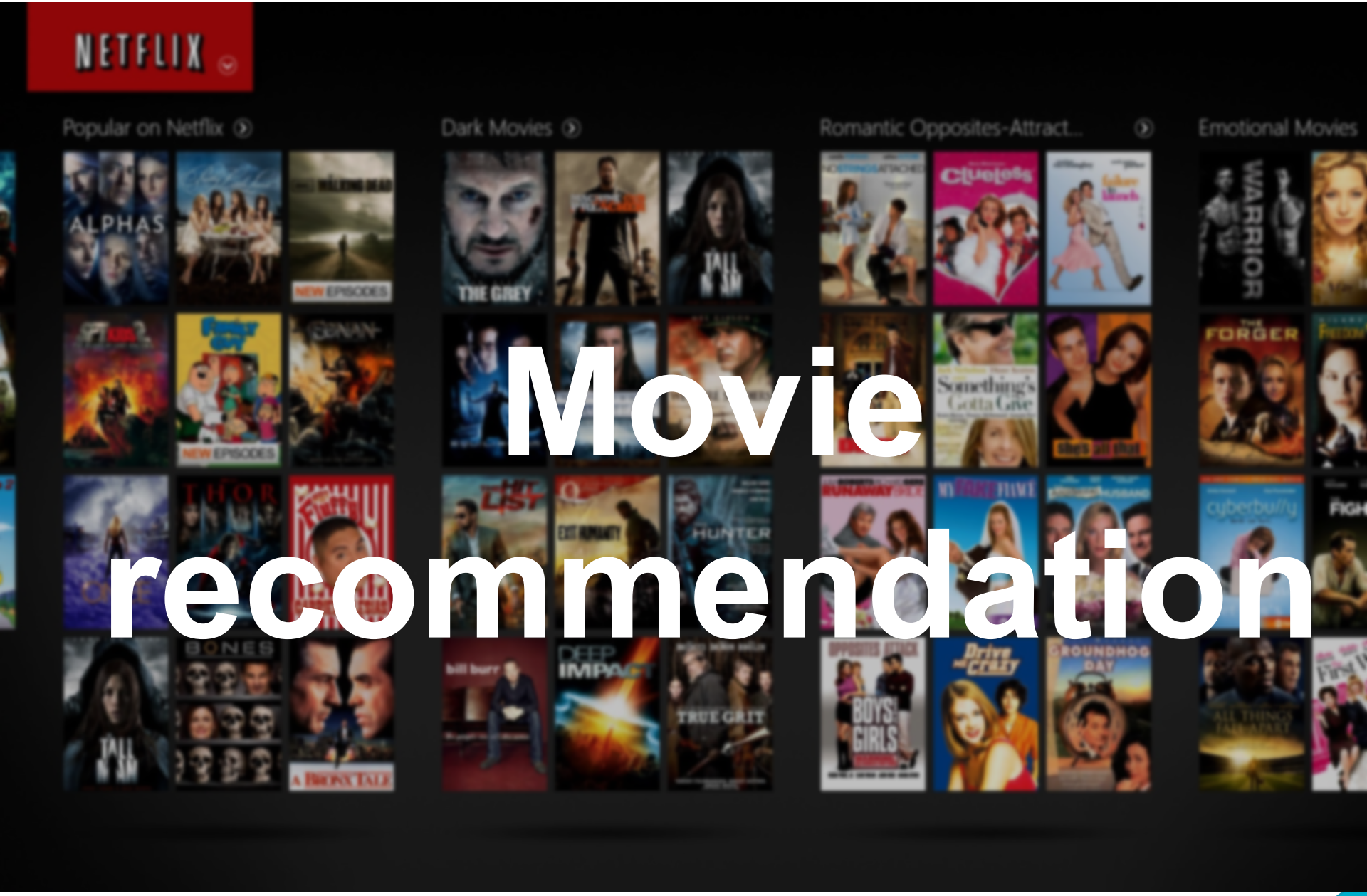
https://en.wikipedia.org/wiki/Machine_learning#Applications

Facial recognition

Spam detection

Voice recognition

ML IN YOUR EVERYDAY LIFE



Movie recommendation

LA RECHERCHE EN APPRENTISSAGE AUTOMATIQUE

En France

- Aix-Marseille, LIF - Laboratoire d'Informatique Fondamentale
- Angers, LERIA - Laboratoire d'Etude et de Recherche en Informatique d'Angers
- Annecy, LISTIC - Laboratoire d'Informatique Systèmes, Traitement de l'Information et de la Connaissance
- Avignon, LIA - Laboratoire Informatique d'Avignon
- Bordeaux, LABRI - Laboratoire Bordelais de Recherche en Informatique
- Caen, GREYC - Groupe de REcherche en Informatique, Image, Automatique et Instrumentation de Caen
- Dijon, Le2I - Laboratoire d'Electronique, Informatique et Image
- Evry, IBISC - Informatique Biologie Intégrative et Systèmes Complexes
- Grenoble, LIG - Laboratoire d'Informatique de Grenoble
- La Rochelle, L3i - Laboratoire Informatique, Image et Interaction
- Le Mans, Lium - Laboratoire d'Informatique de l'Université du Maine
- Lille, LIFL - Laboratoire d'Informatique Fondamentale de Lille
  LISIC - Laboratoire d'Informatique Signal et Image de la Côte d'Opale
- Lyon, LIRIS - Laboratoire d'InfoRmatique en Image et Systèmes d'information
  ERIC - Entrepôts, Représentation et Ingénierie des Connaissances
- Montpellier, LIRMM - Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
- Nancy, LORIA - Laboratoire lorrain de recherche en informatique et ses applications
- Nantes, LINA - Laboratoire d'Informatique de Nantes Atlantique
- Nice, i3S - Laboratoire d'Informatique, Signaux et Systèmes de Sophia-Antipolis
- Orléans, LIFO - Laboratoire d'Informatique Fondamentale d'Orléans
- Tours, LI - Laboratoire d'Informatique Université François-Rabelais, Tours
- Paris 3, LATTICE - Langues, Textes, Traitements informatiques, Cognition
- Paris 4, LaLIC - Langues, Logiques, Informatique, Cognition
- Paris 5, LIPADE - Laboratoire d'Informatique Paris Descartes
- Paris 6, LIP6 - Laboratoire d'Informatique de Paris 6
- Paris 9, LAMSADE - Laboratoire d'Analyse et Modélisation de Systèmes pour l'Aide à la Décision
- Paris 10, LRI - Laboratoire de Recherche en Informatique
- Paris 11, LIMSI - Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur
- Paris 13, LIPN - Laboratoire d'Informatique de Paris-Nord
- …
- Rennes, IRISA - Institut de recherche en informatique et systèmes aléatoires
- Strasbourg, LSiiT - Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection
- Toulouse, IRIT - Institut de Recherche en Informatique de Toulouse

Plus de 50% des laboratoires de recherche académiques ont au moins une équipe travaillant sur une thématique AA ou associée.

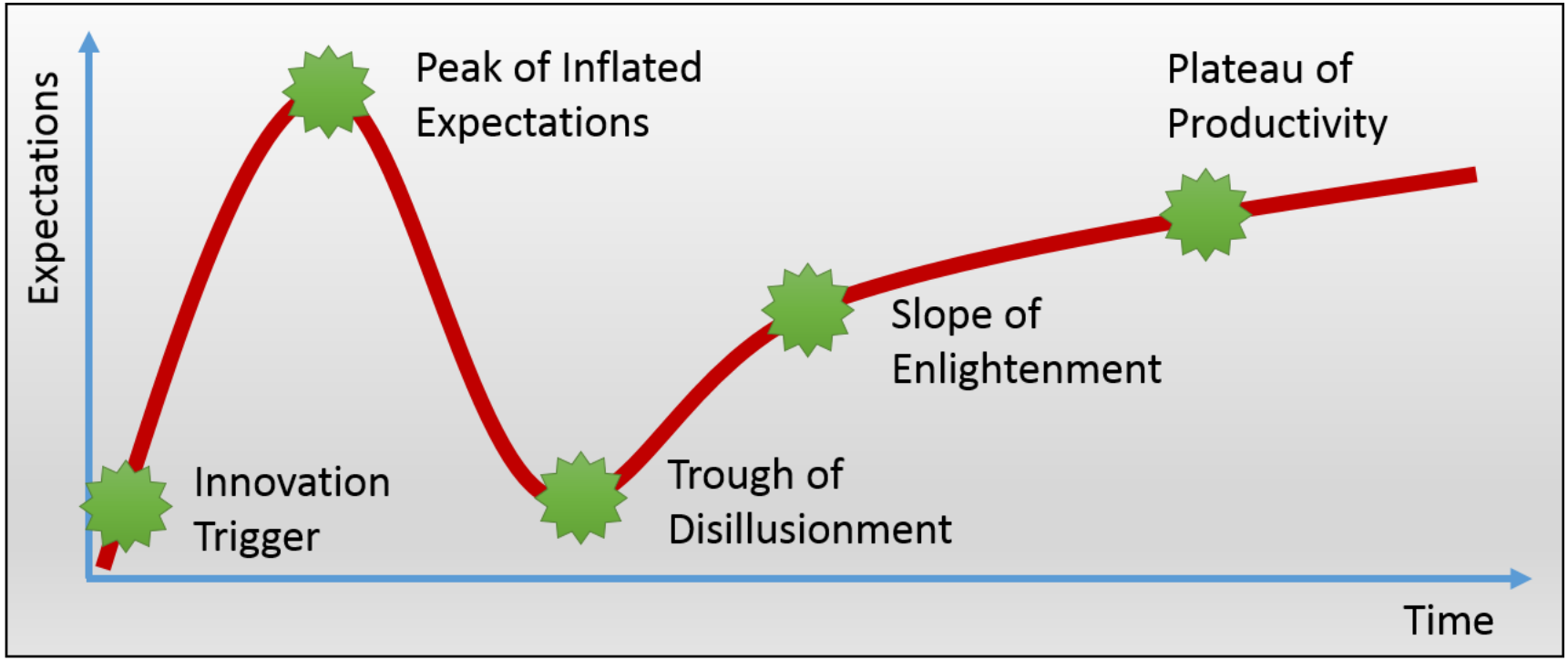
02> MYTHES ET LÉGENDES

DE NOS JOURS

Buzzwords

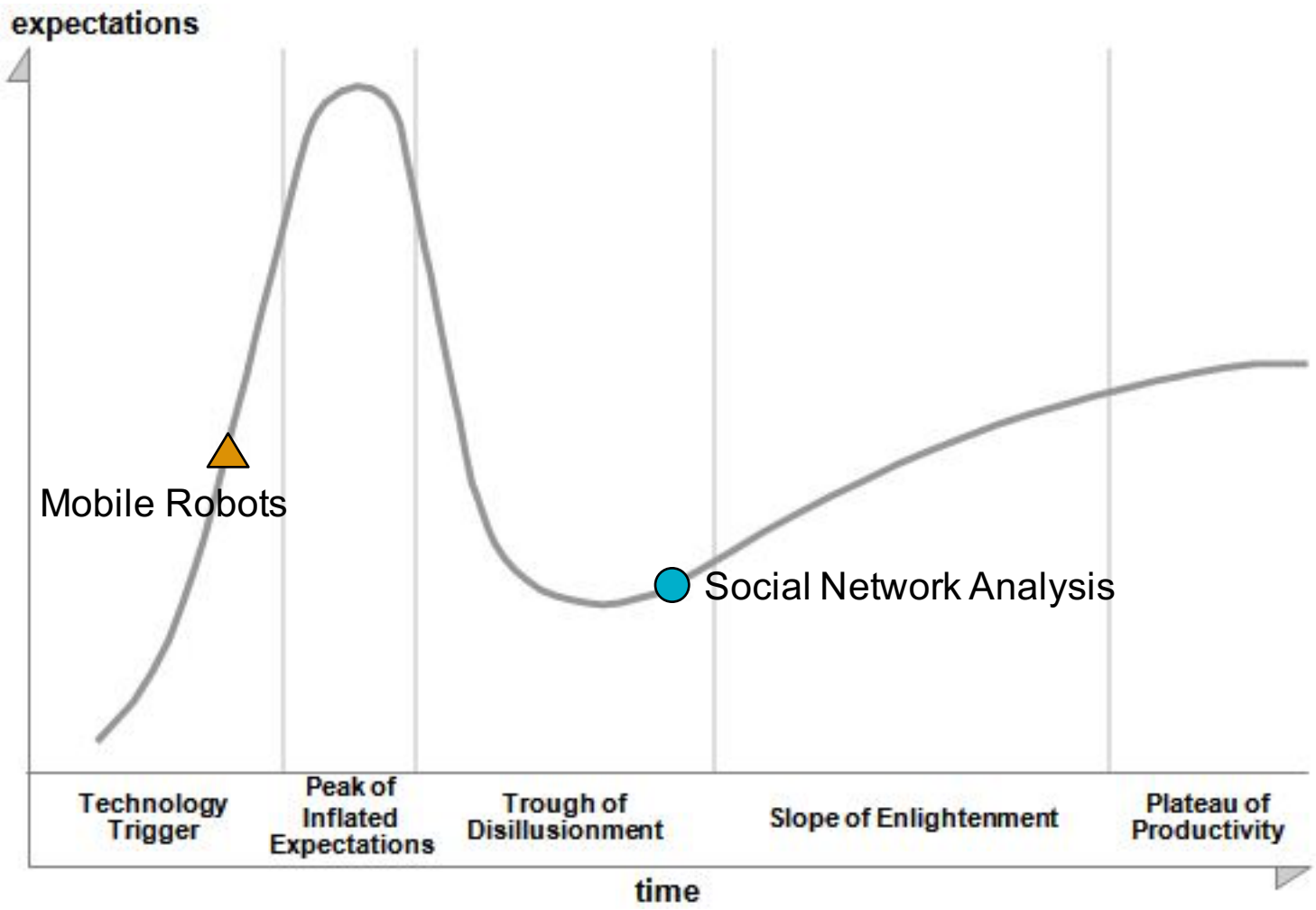
- Machine Learning, Data Science, Big Data, Deep Learning →
- Pourtant:
  - Le terme Machine Learning et les premiers travaux qu'on a classé dans ce domaine datent des années 50.
  - La statistique moderne existe depuis bien longtemps (XVIIIème siècle).
  - Quel rapport avec le Big Data ? A la base, aucun.
  - On avait DEJA de la puissance de calcul pour faire du ML il y a 15 ans.
  - On faisait déjà des statistiques sur ordinateur il y a 50 ans...
    - ...mais on manquait de données (ah ! enfin il y a un rapport avec les données massives).
  - On s'y met:
    - tout le monde en parle
    - personne n'en fait vraiment, à part les grands qui sont à des années lumière de ce qui se fait actuellement (Google, Microsoft, IBM, etc.)

HYPENESS ILLUSION

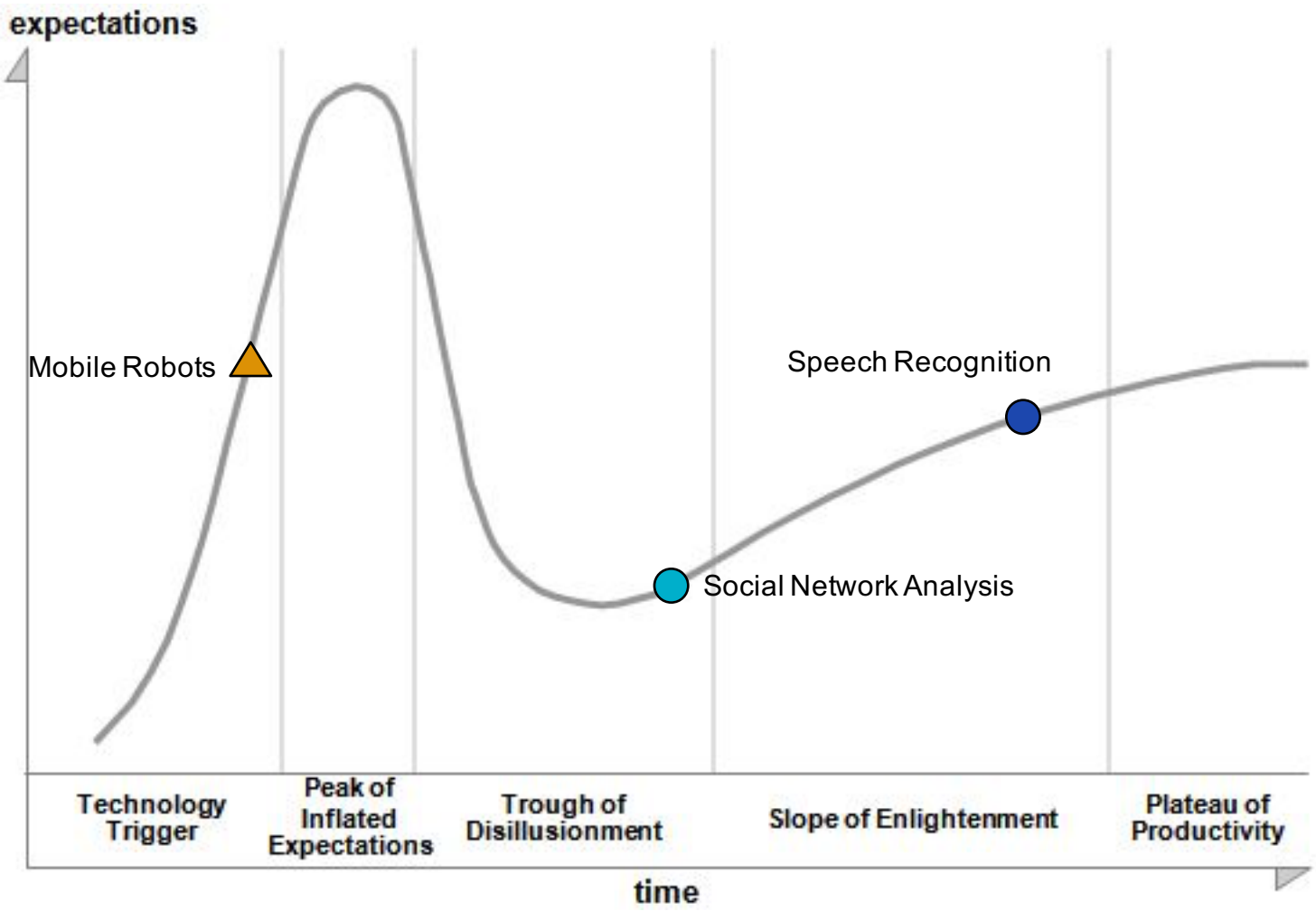


Gartner Hype Cycle for Emerging Technologies

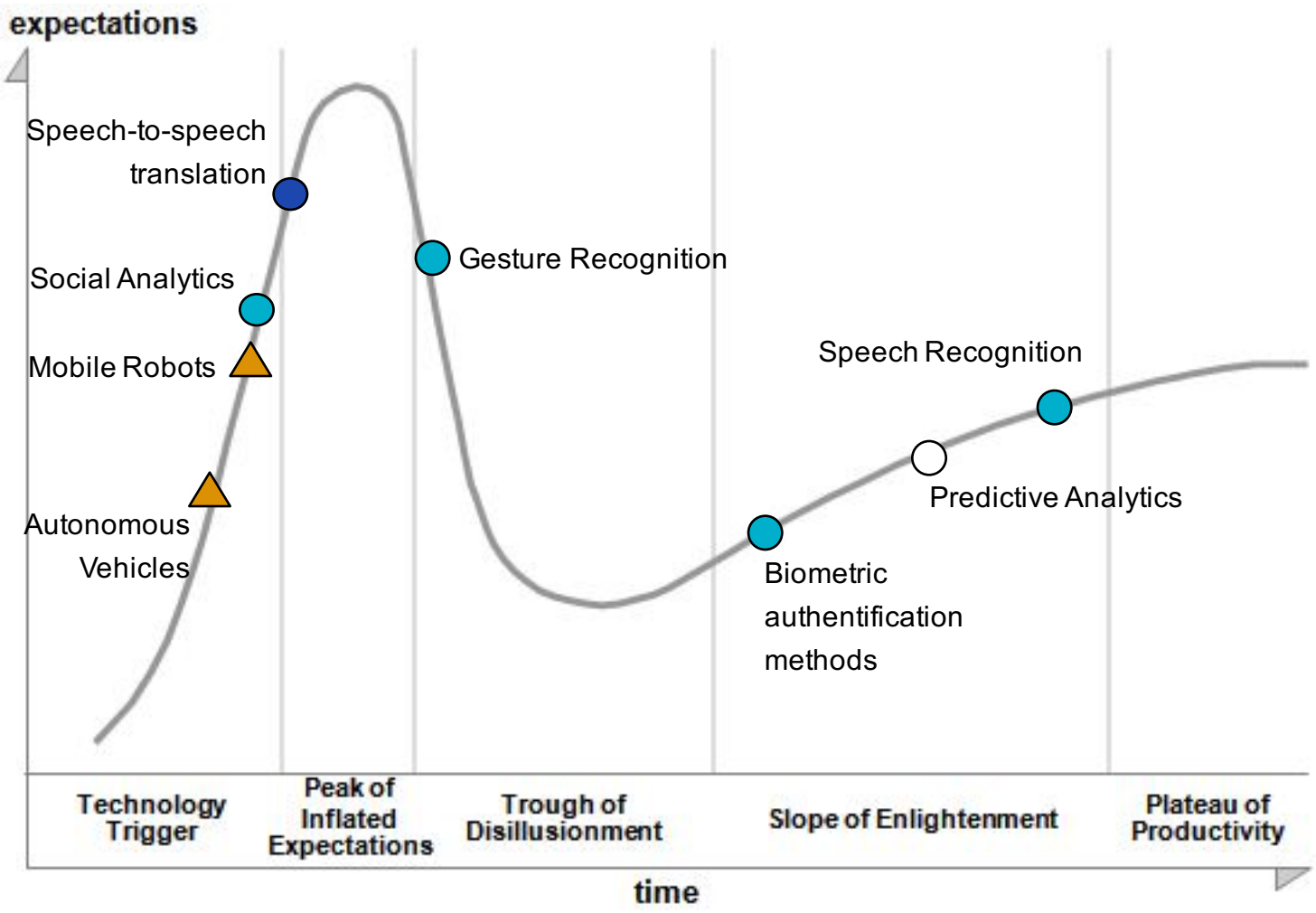
2008


expectations
Mobile Robots
Social Network Analysis
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

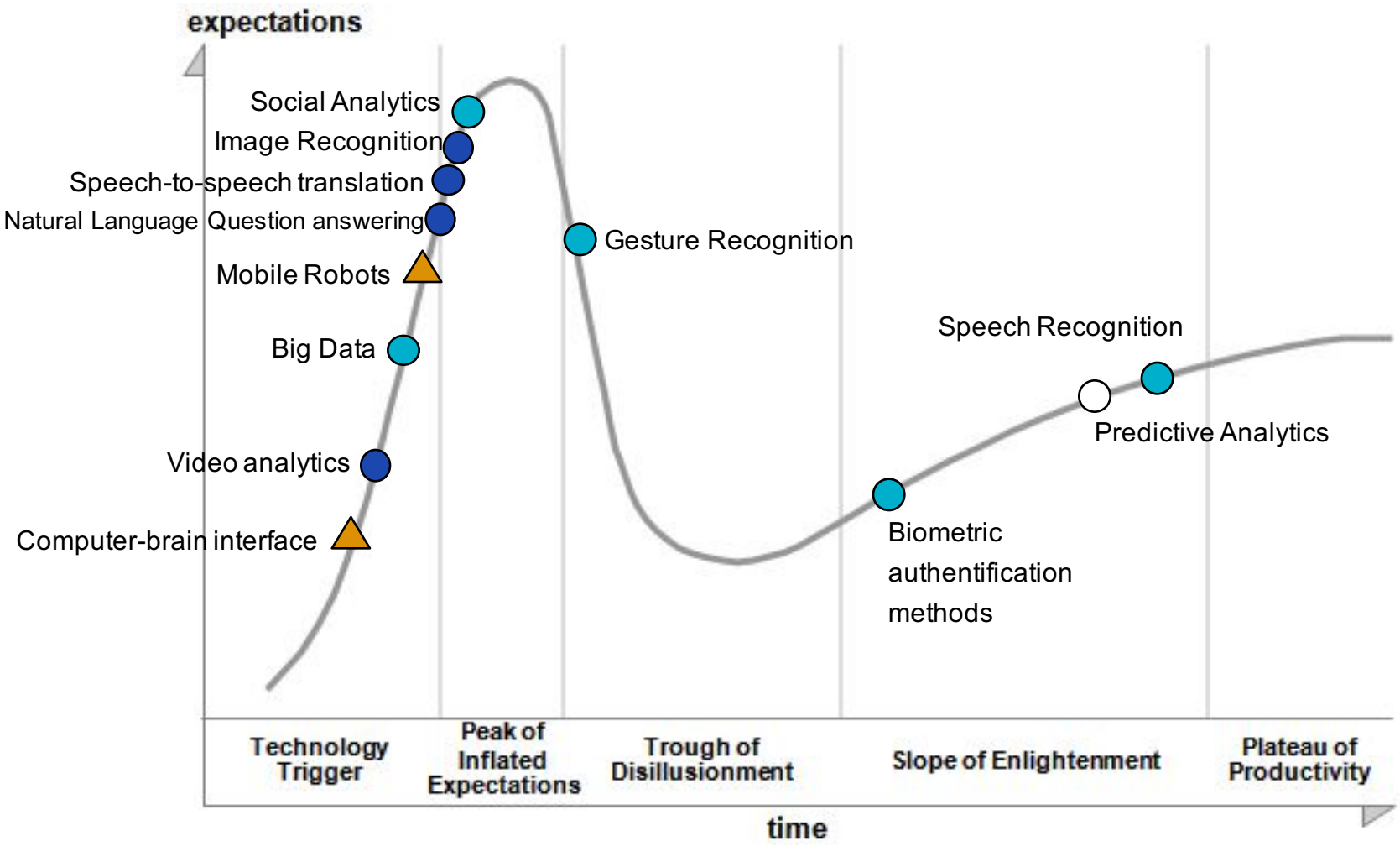
2009


expectations
Mobile Robots
Speech Recognition
Social Network Analysis
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

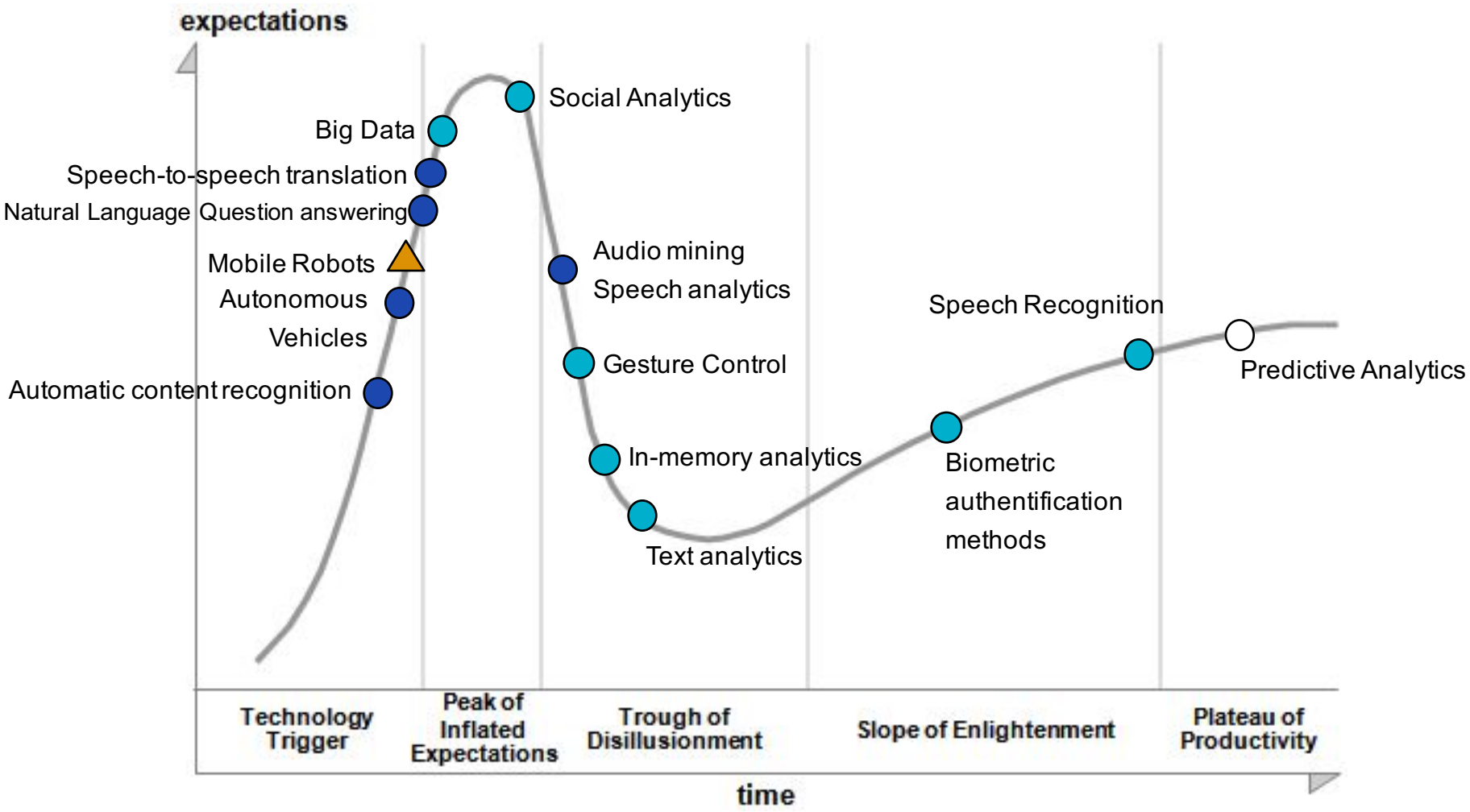
2010


expectations
Speech-to-speech translation
Social Analytics
Mobile Robots
Autonomous Vehicles
Gesture Recognition
Speech Recognition
Predictive Analytics
Biometric authentification methods
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

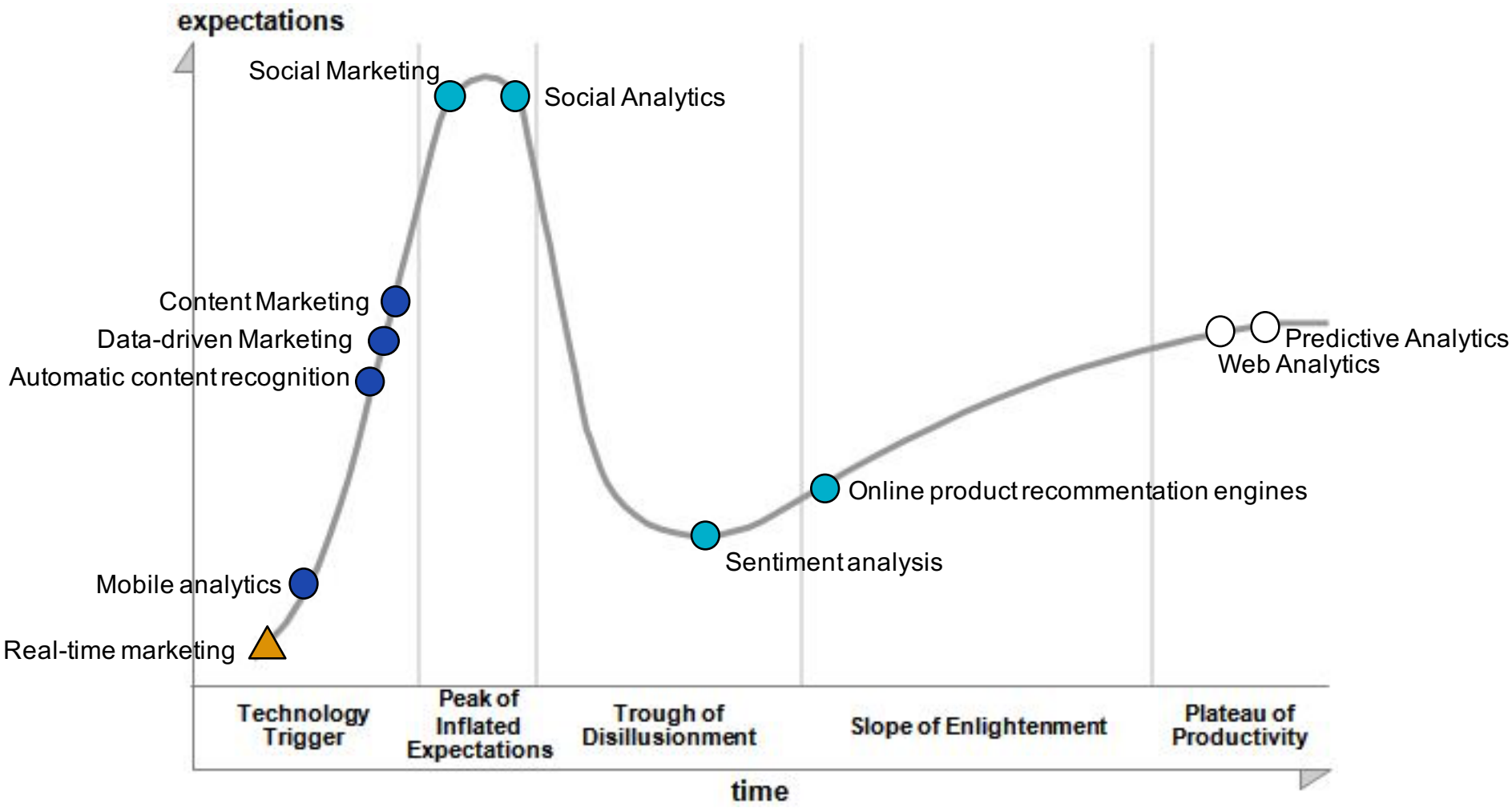
2011


expectations
Social Analytics
Image Recognition
Speech-to-speech translation
Natural Language Question answering
Mobile Robots
Big Data
Video analytics
Computer-brain interface
Gesture Recognition
Speech Recognition
Predictive Analytics
Biometric authentification methods
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

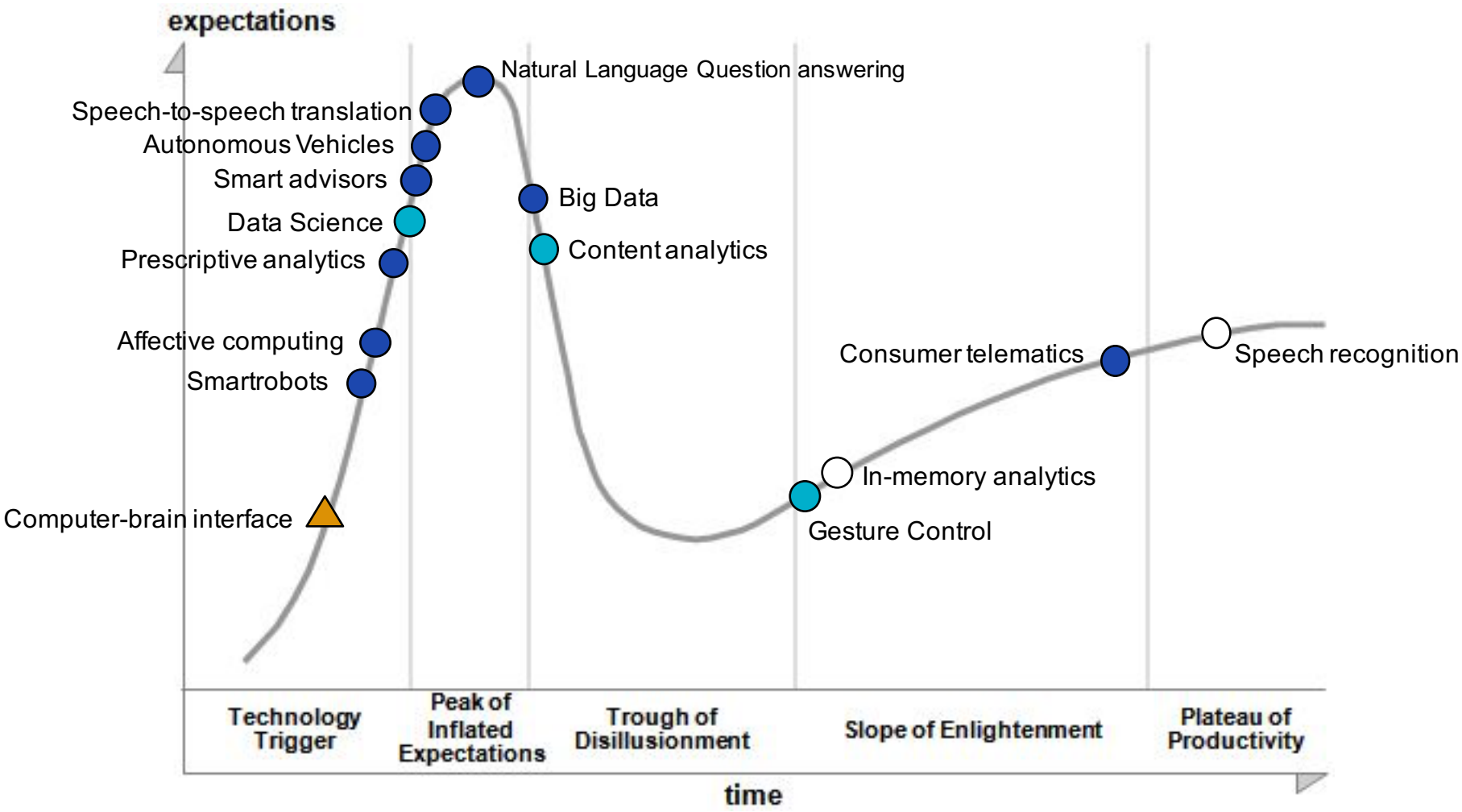
2012


expectations
Social Analytics
Big Data
Speech-to-speech translation
Natural Language Question answering
Mobile Robots
Autonomous Vehicles
Automatic content recognition
Audio mining
Speech analytics
Gesture Control
In-memory analytics
Text analytics
Speech Recognition
Biometric authentification methods
Predictive Analytics
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

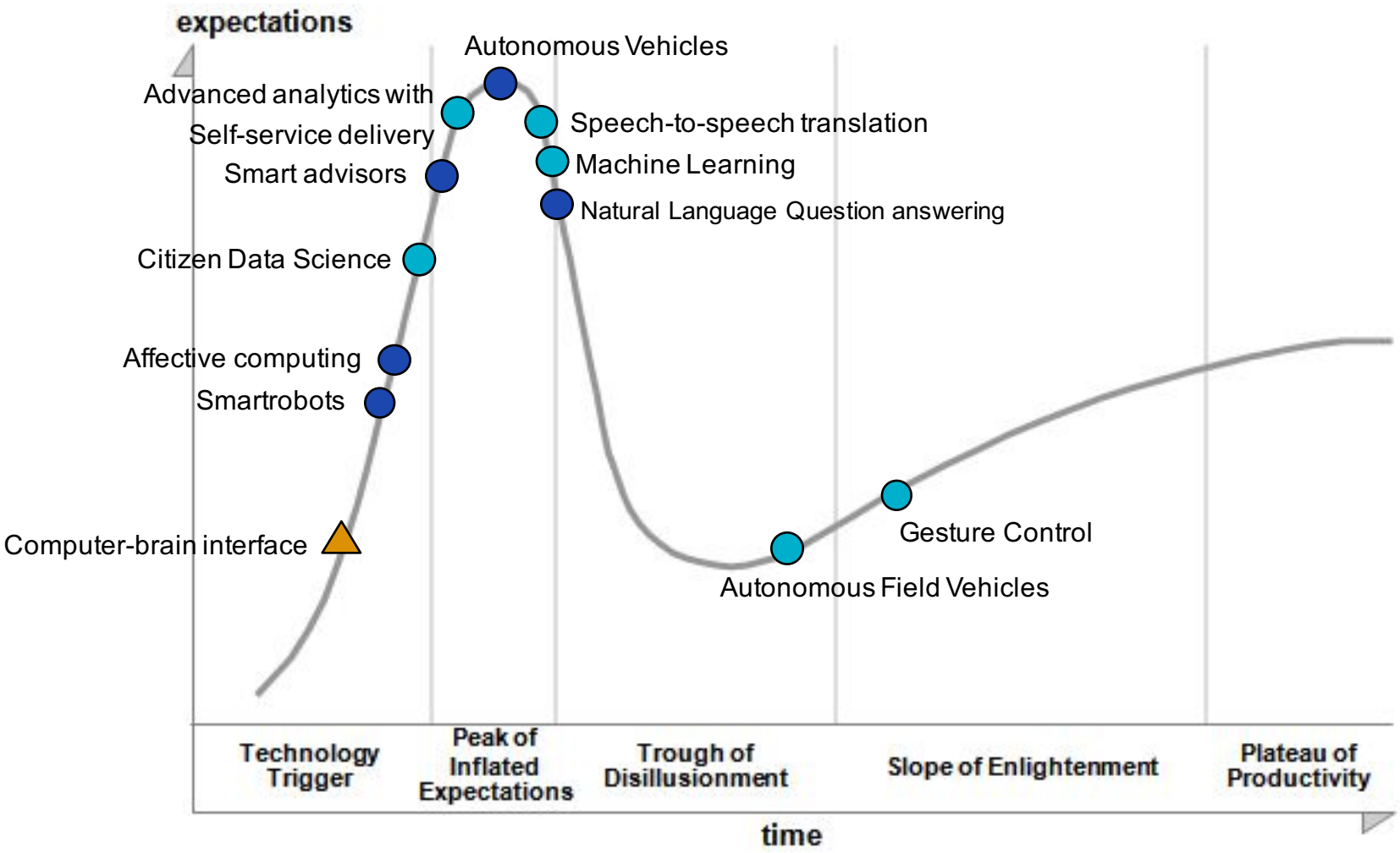
2013


expectations
Social Marketing
Social Analytics
Content Marketing
Data-driven Marketing
Automatic content recognition
Mobile analytics
Real-time marketing
Predictive Analytics
Web Analytics
Online product recommentation engines
Sentiment analysis
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

2014


expectations
Natural Language Question answering
Speech-to-speech translation
Autonomous Vehicles
Smart advisors
Big Data
Data Science
Content analytics
Prescriptive analytics
Affective computing
Smartrobots
Consumer telematics
Speech recognition
In-memory analytics
Computer-brain interface
Gesture Control
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

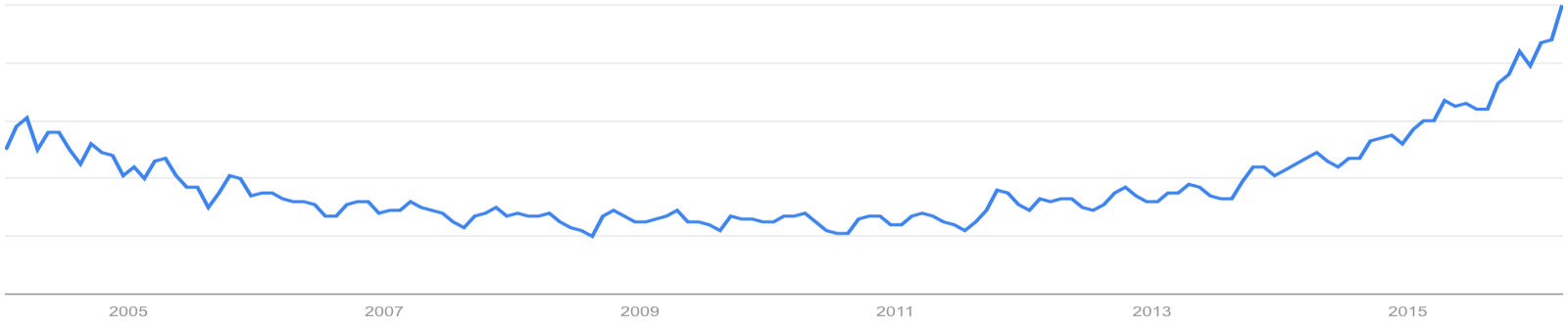
2015


expectations
Autonomous Vehicles
Advanced analytics with
Self-service delivery
Smart advisors
Speech-to-speech translation
Machine Learning
Natural Language Question answering
Citizen Data Science
Affective computing
Smartrobots
Computer-brain interface
Gesture Control
Autonomous Field Vehicles
Technology Trigger
Peak of Inflated Expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of Productivity
time

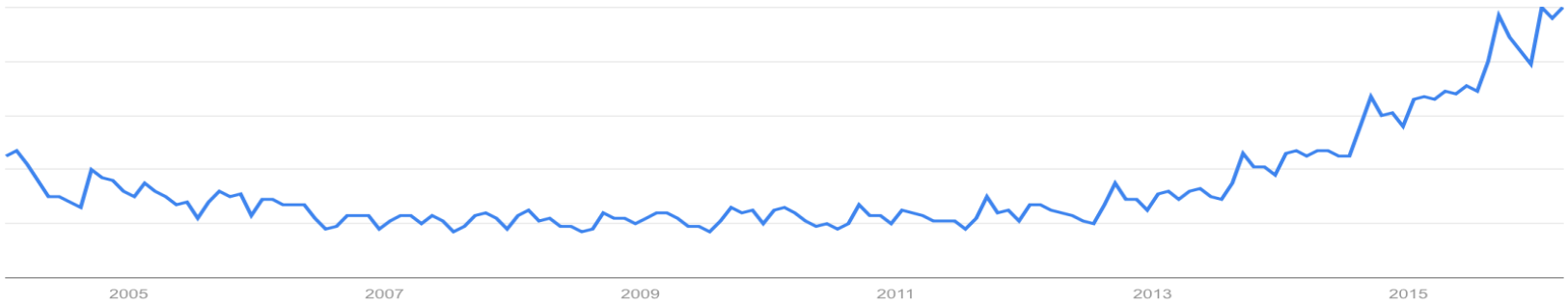
SERIOUSLY ?

Let me ask Google trends

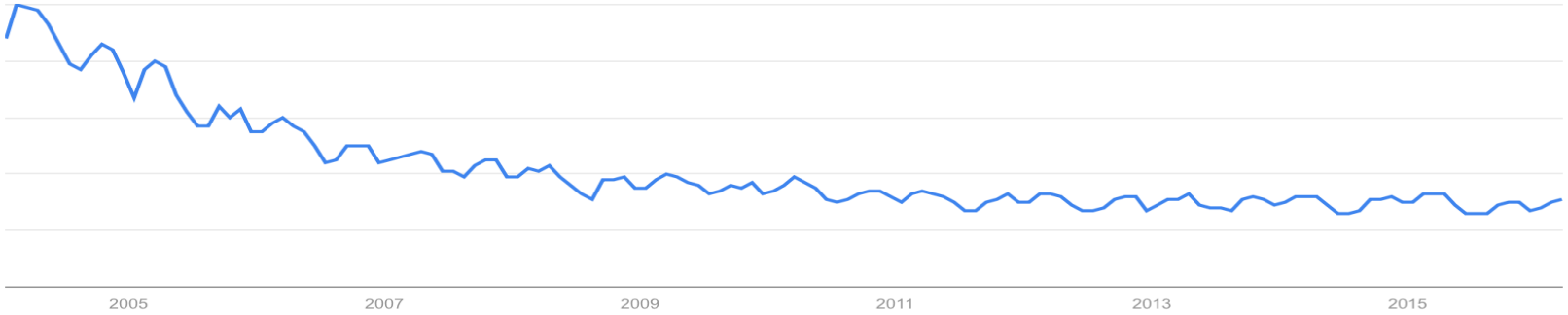
- *Machine Learning* – autant de recherches en mars 2015 qu'en mars 2004.



- *Data Science* – autant de recherches en mai 2014 qu'en février 2004.



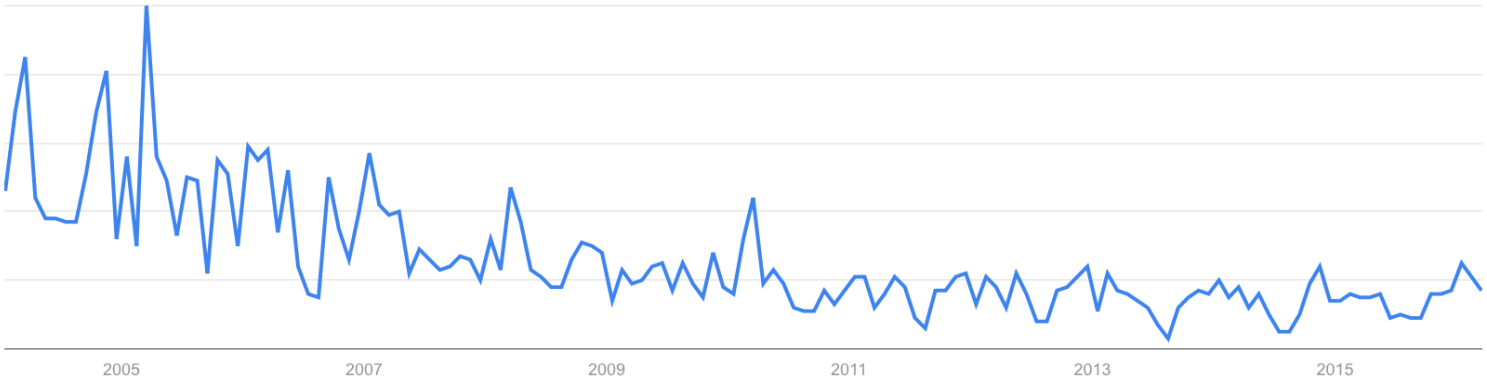
- *Data Mining* – c'est la chute !



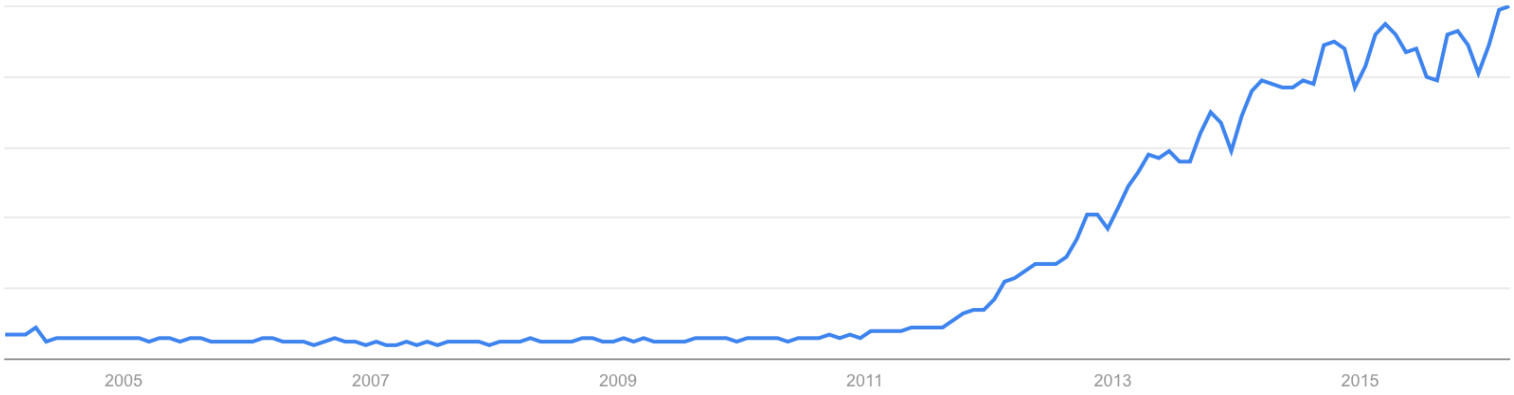
SÉRIOUZELI ?

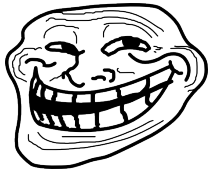
Demandons à Gougueule trènedes

- *Apprentissage automatique, science de la donnée, fouille de données....*



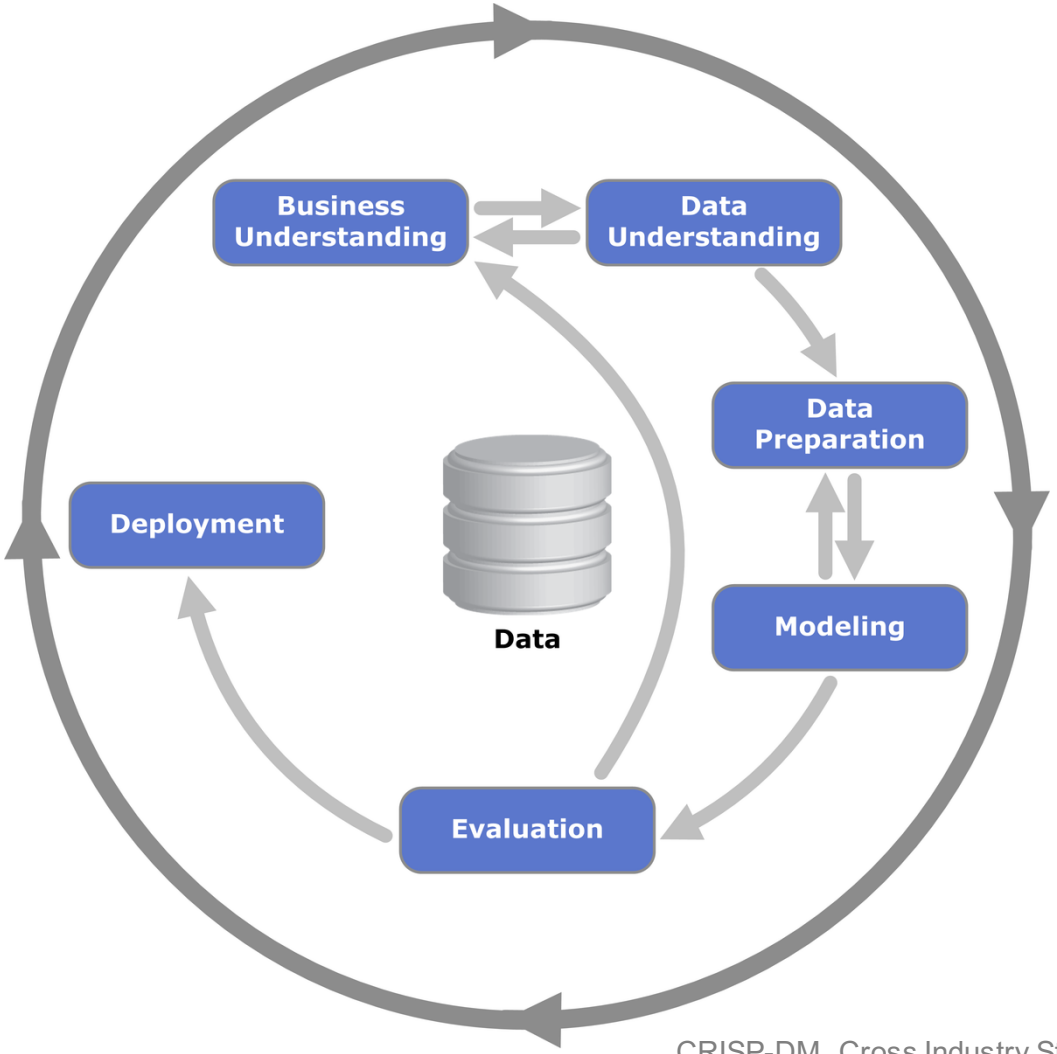
- Au fait, et *Big Data* ?





LES ENTREPRISES NE FONT PAS DU MACHINE LEARNING

Elles font du Data Mining, du KDD ; à la limite, on dit Data Science

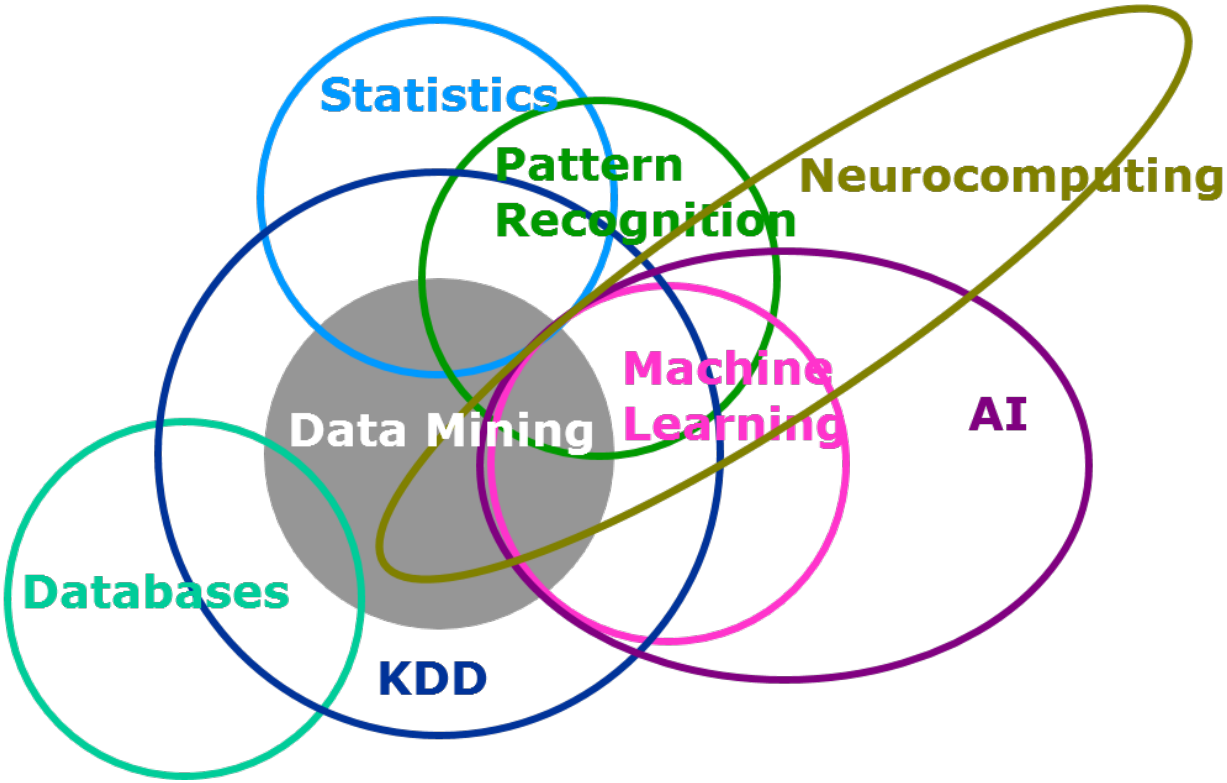


CRISP-DM, Cross Industry Standard Process for Data Mining

CC Kenneth Jensen, 2012

DATA MINING

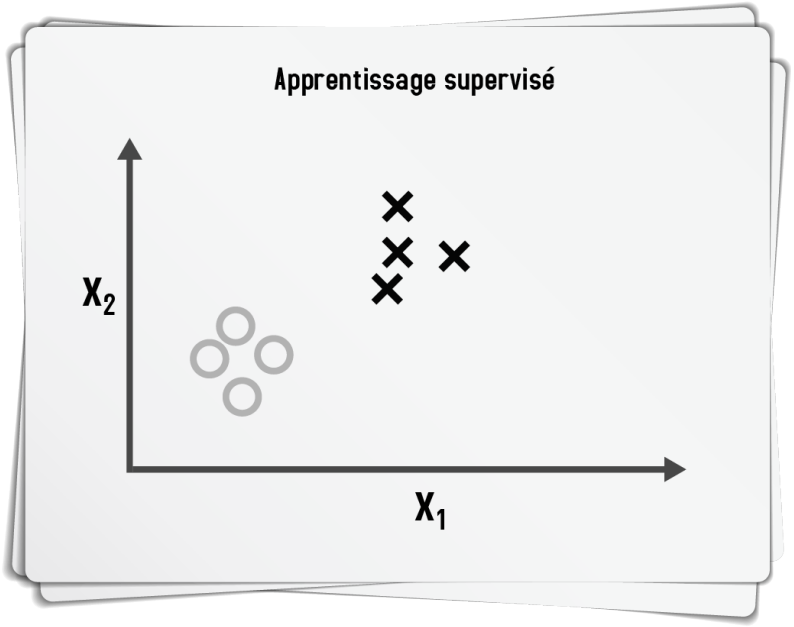
Tiré d'un document SAS de 1998

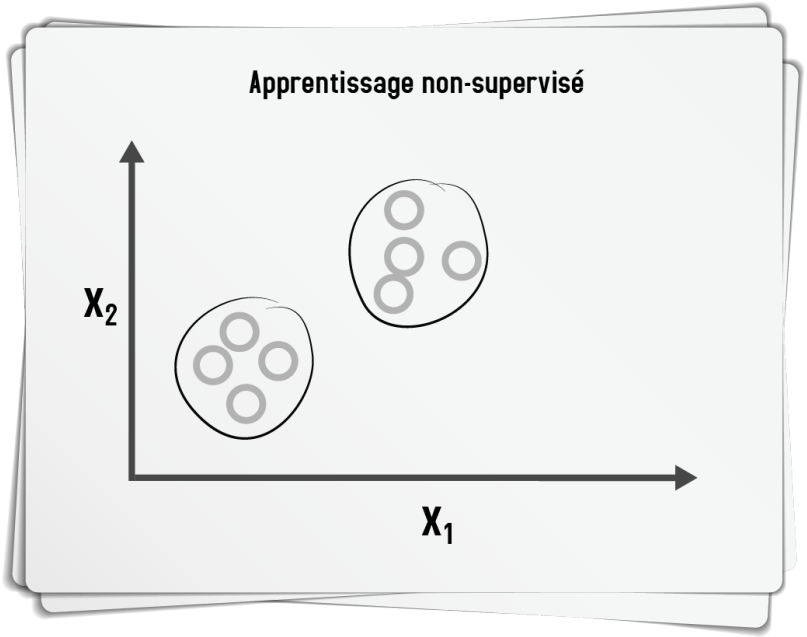


03> CATÉGORIES

CATÉGORIES D'ALGORITHMES DANS LE MACHINE LEARNING

supervisé, non-supervisé, (semi-supervisé)





APPRENTISSAGE SUPERVISÉ

- On dispose d'exemples permettant l'apprentissage.

- Habituellement, on sera confronté à un problème de:
  - Régression
  - Classification

- En vrac:
  - Réseaux de neurones
  - Modèles bayésiens
  - Arbres de décision
  - SVM
  - K plus proches voisins
  - Modèles de Markov
  - Descente de gradient
  - ...
  - Et bien sûr toutes les techniques classiques de la statistique (régressions, ANOVA, ARIMA, etc.).

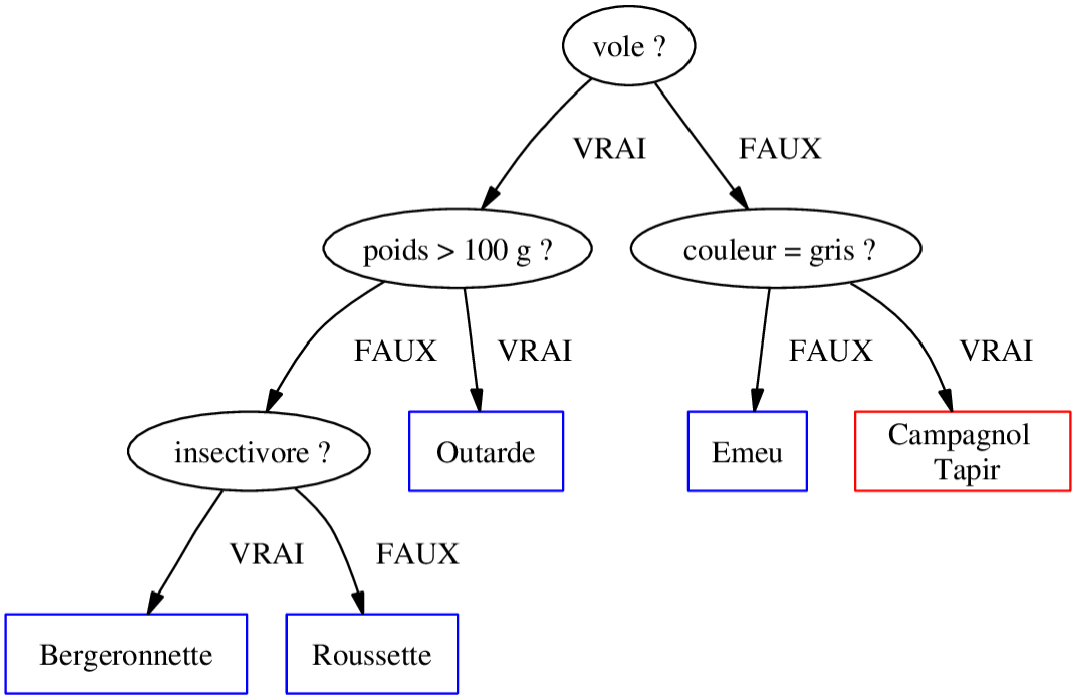
APPRENTISSAGE SUPERVISÉ

Arbre de décision

- Exemple: on veut développer un système capable d'apprendre à reconnaître des animaux et on dispose de descriptions.
- Algorithmes: ID3, C4.5, ...

VOLE	POIDS	COULEUR	NUTRITION	PEAU	ANIMAL
oui	1kg	roux	granivore	plumes	outarde
oui	20g	gris et jaune	insectivore	plumes	bergeronnette
non	100kg	noir et blanc	omnivore	plumes	émeu
non	5g	gris	granivore	poils	campagnol
non	40kg	gris	herbivore	poils	tapir
oui	60g	noir	frugivore	poils	roussette

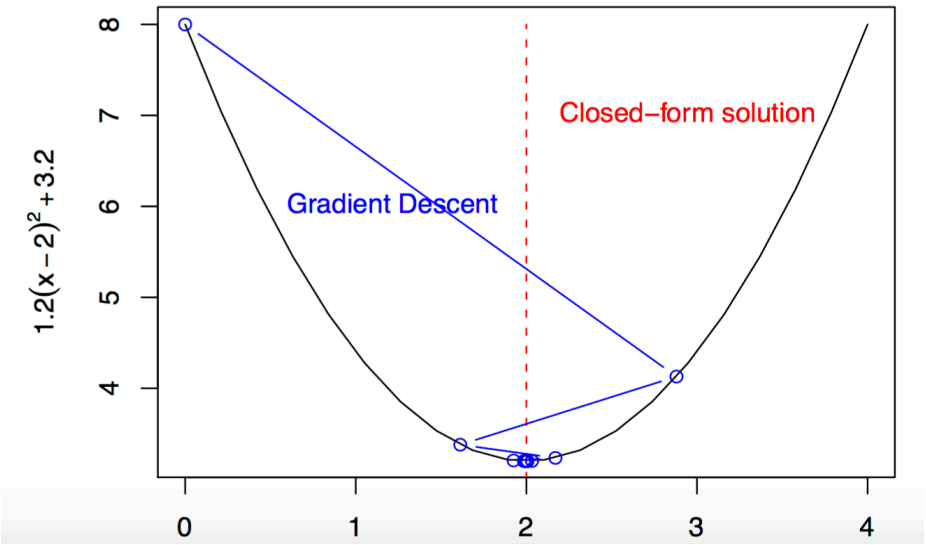
Frédéric Fürst, Univ. de Picardie

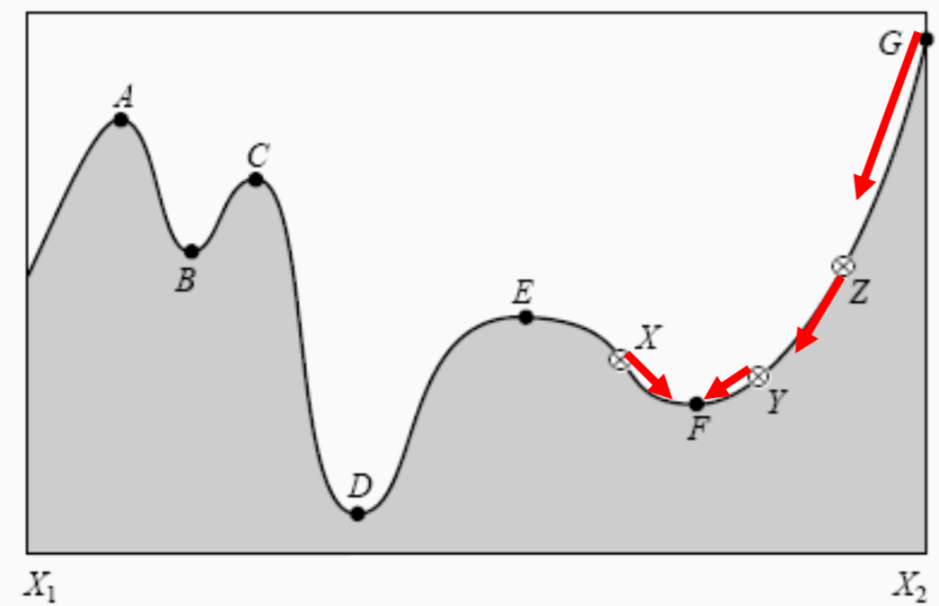


APPRENTISSAGE SUPERVISÉ

Descente du gradient

- Problème d'optimisation.
- Objectif: trouver la solution « optimale » par rapport à une certaine fonction.
- Etapes:
  - On part d'un point quelconque
  - Tant que le critère n'est pas satisfait
    - On détermine une direction
    - On détermine un pas
    - On met à jour
- On tombe souvent sur un minimum local
- L'utilisation de méthodes stochastiques permet d'améliorer les résultats (recuit simulé, gradient stochastique)



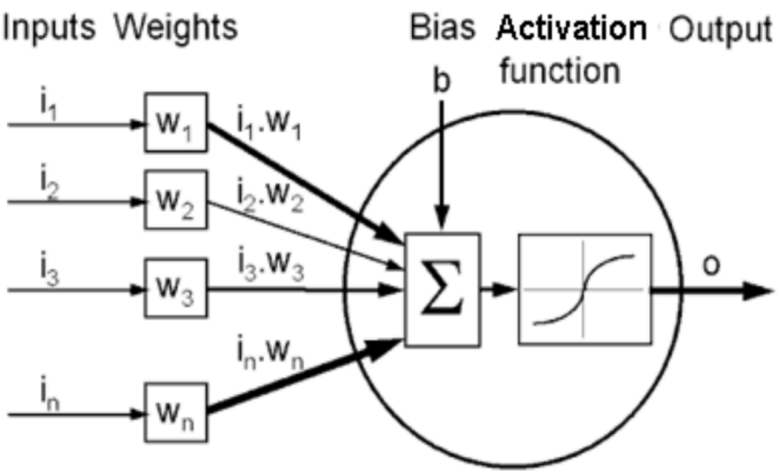


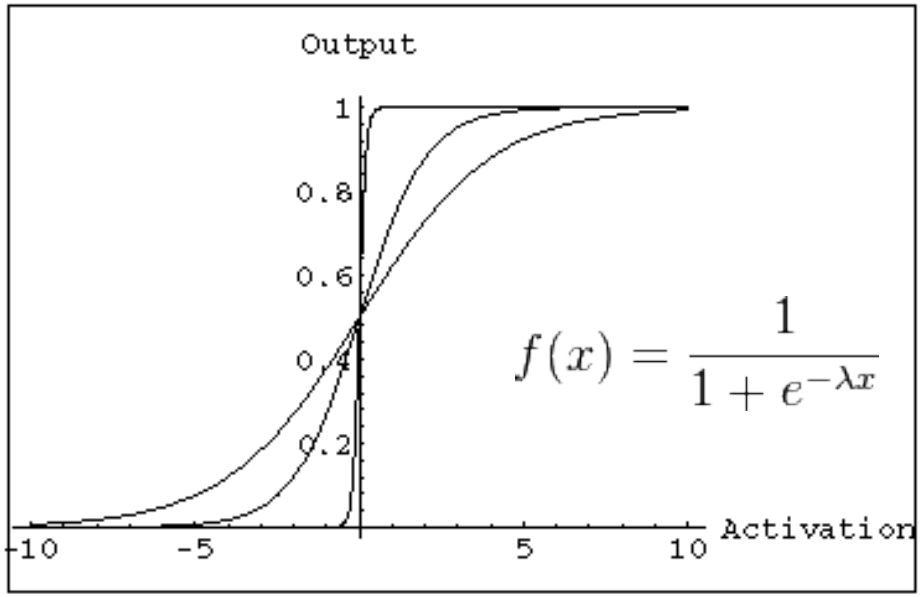
APPRENTISSAGE SUPERVISÉ

Réseaux de neurones

- Modèles inspirés du vivant.
- Le modèle le plus simple est le *perceptron* (1960's).
- En général, une fonction dite « sigmoïde » (ou logistique) est utilisée comme fonction d'activation.
- Deux phases:
  - Apprentissage
  - Evaluation
- Un perceptron est un *classifieur*.

$$o = f\left(\sum_{j=1}^{n} w_j x_j\right)$$





APPRENTISSAGE SUPERVISÉ

Réseaux de neurones

- Exemple: soit un perceptron:
  - avec 2 entrées booléennes: X1, X2 ∈ {0, 1}
  - une sortie booléenne o ∈ {0, 1}

- L'ensemble d'apprentissage est donné par:

x_1	x_2	$o(x_1, x_2) = x_1 \wedge x_2$
1	*1*	*1*
1	*0*	*0*
0	*1*	*0*
0	*0*	*0*

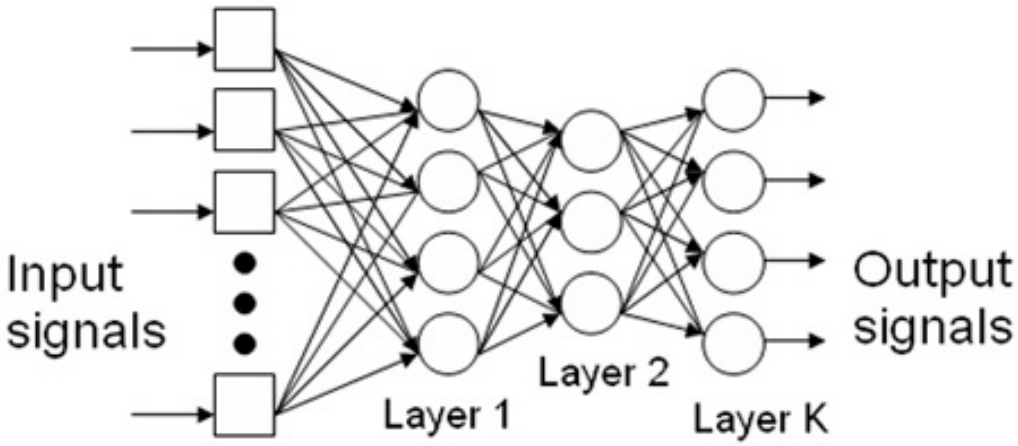
Fonction ET logique

- Objectif: trouver les poids w1 et w2, ainsi que la valeur seuil b du biais pour obtenir la sortie désirée
- Une solution: w1 = w2 = ½ et b = 0,6

$$o(x_1, x_2) = \begin{cases} 1 & \textit{if } x_1/2 + x_2/2 \geq 0.6 \\ 0 & \textit{otherwise} \end{cases}$$

APPRENTISSAGE SUPERVISÉ

Réseaux de neurones



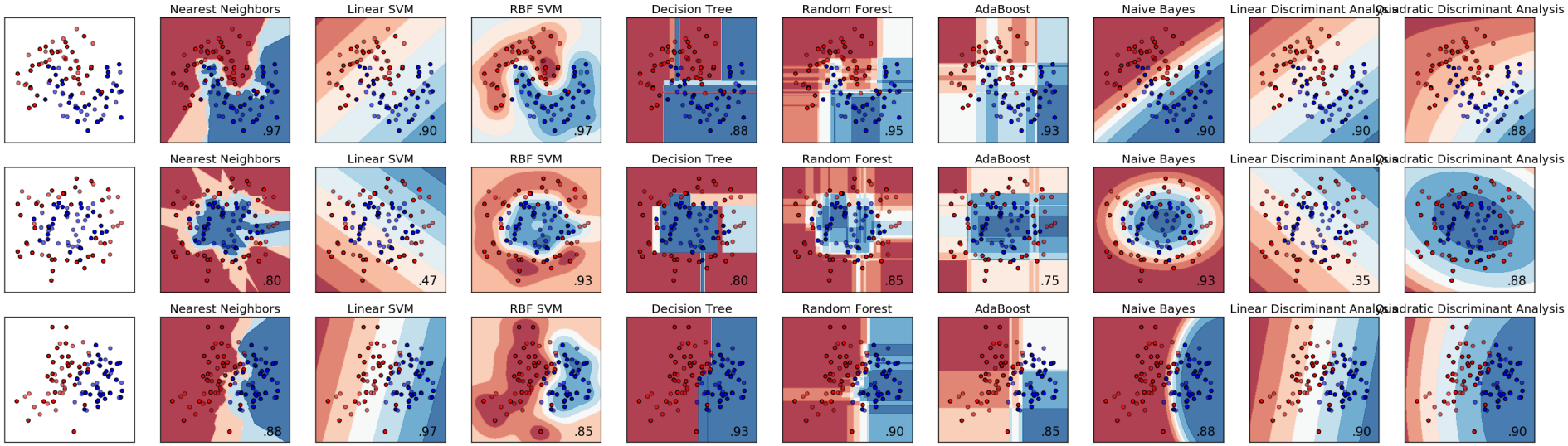
- Un réseau de neurones se comporte comme un ensemble de perceptrons.

- En général, on utilise l'algorithme de « Backpropagation »
  - Foward propagation
  - Backward propagation
  - Mise à jour des poids

- Référence avec un exemple concret:
  http://home.agh.edu.pl/~vlsi/AI/backp_t_en/backprop.html

APPRENTISSAGE SUPERVISÉ

En vrac – Algorithmes de classification



http://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

APPRENTISSAGE NON SUPERVISÉ

- On ne dispose pas d'exemples permettant l'apprentissage.

- Habituellement, on sera confronté à un problème de:
  - Clustering
  - Apprentissage de modèles à variables latentes

- En vrac:
  - Réseaux de neurones
  - Clustering hiérarchique
  - K-means
  - Gaussian mixture models
  - Hidden Markov Models
  - ...

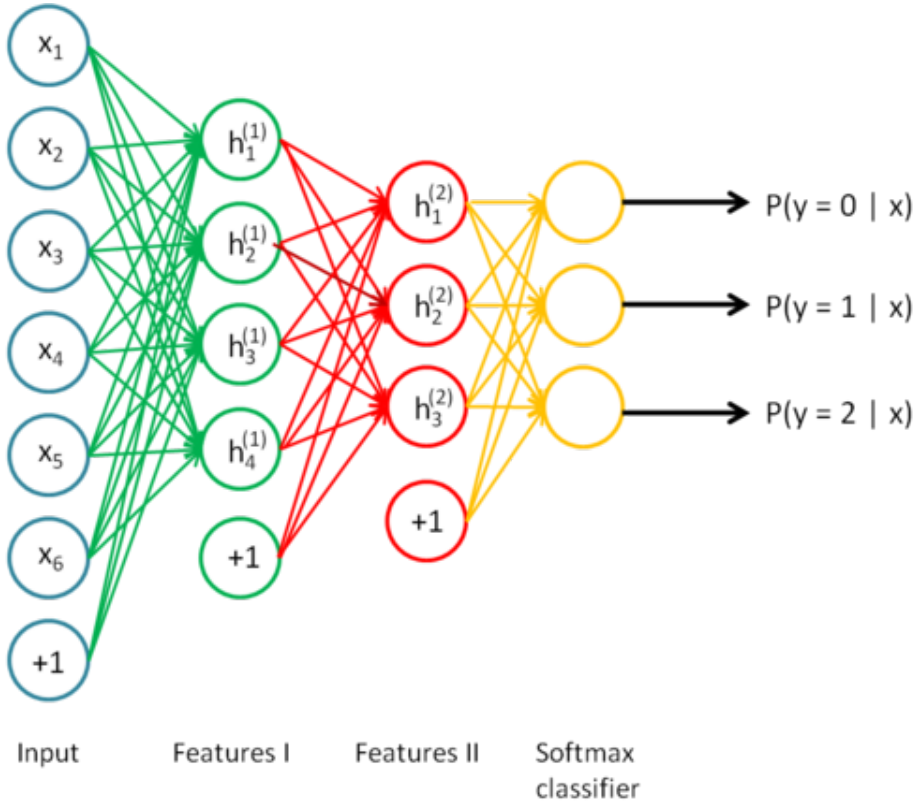
  - Et bien sûr toutes les techniques classiques de la statistique (ACP, AFC, analyse multi-variée, EM, etc.).

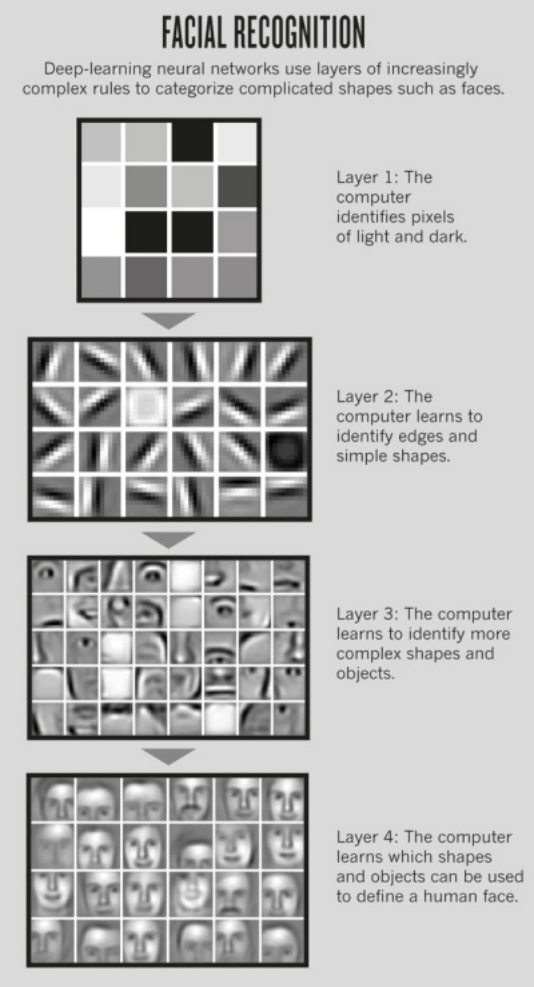
  - Cas particulier: apprentissage par renforcement

APPRENTISSAGE NON SUPERVISÉ

ANN Auto-Encodeurs – Exemple de Deep Learning

- Un réseau de neurones avec un grand nombre de couches cachées.
- Complétement non-supervisé, mais nécessite un apprentissage.
- Un certain nombre de « features » doit être extrait.
- Une grande quantité de données est nécessaire pour de bonnes performances.

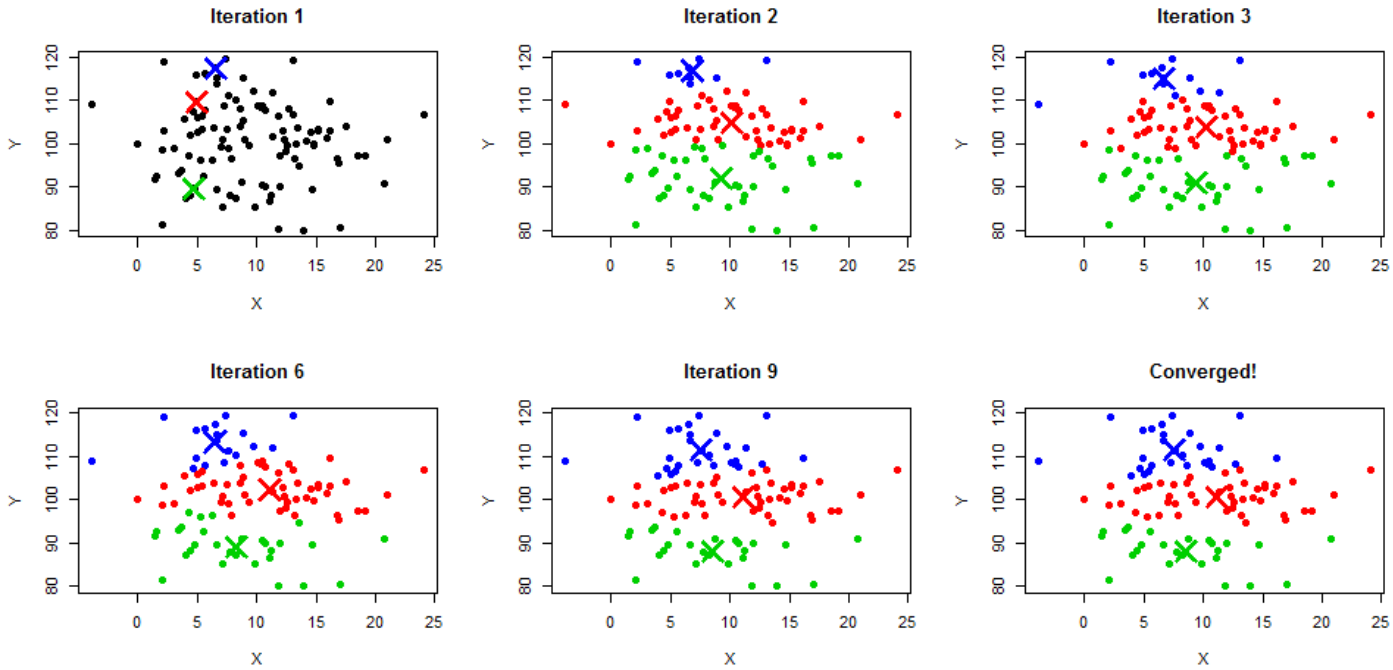




APPRENTISSAGE NON SUPERVISÉ

K-Moyennes

- Partitionner les données en K « clusters » (K est fixé).
- Sensible au bruit, aux intrus (outliers).
- Minimum local et clusters convexes.
- L'algorithme minimise la somme des distances entre les données et les centres des clusters.

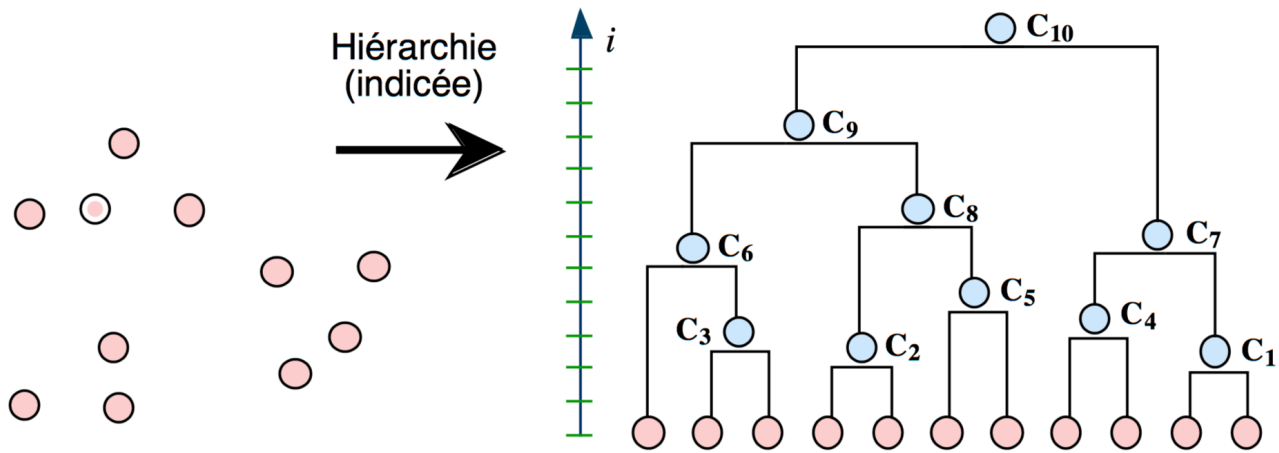


APPRENTISSAGE NON SUPERVISÉ

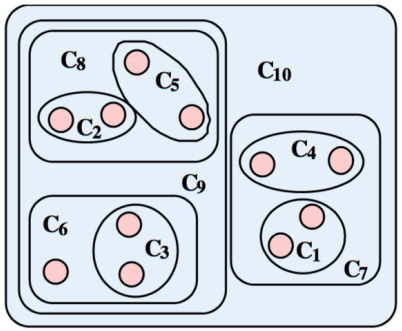
Clustering hiérarchique

- **Ascendant**
  - Chaque point ou cluster est progressivement "absorbé" par le cluster le plus proche.
  - Trouver une coupe « idéale »

Bisson, 2001



Dendogramme

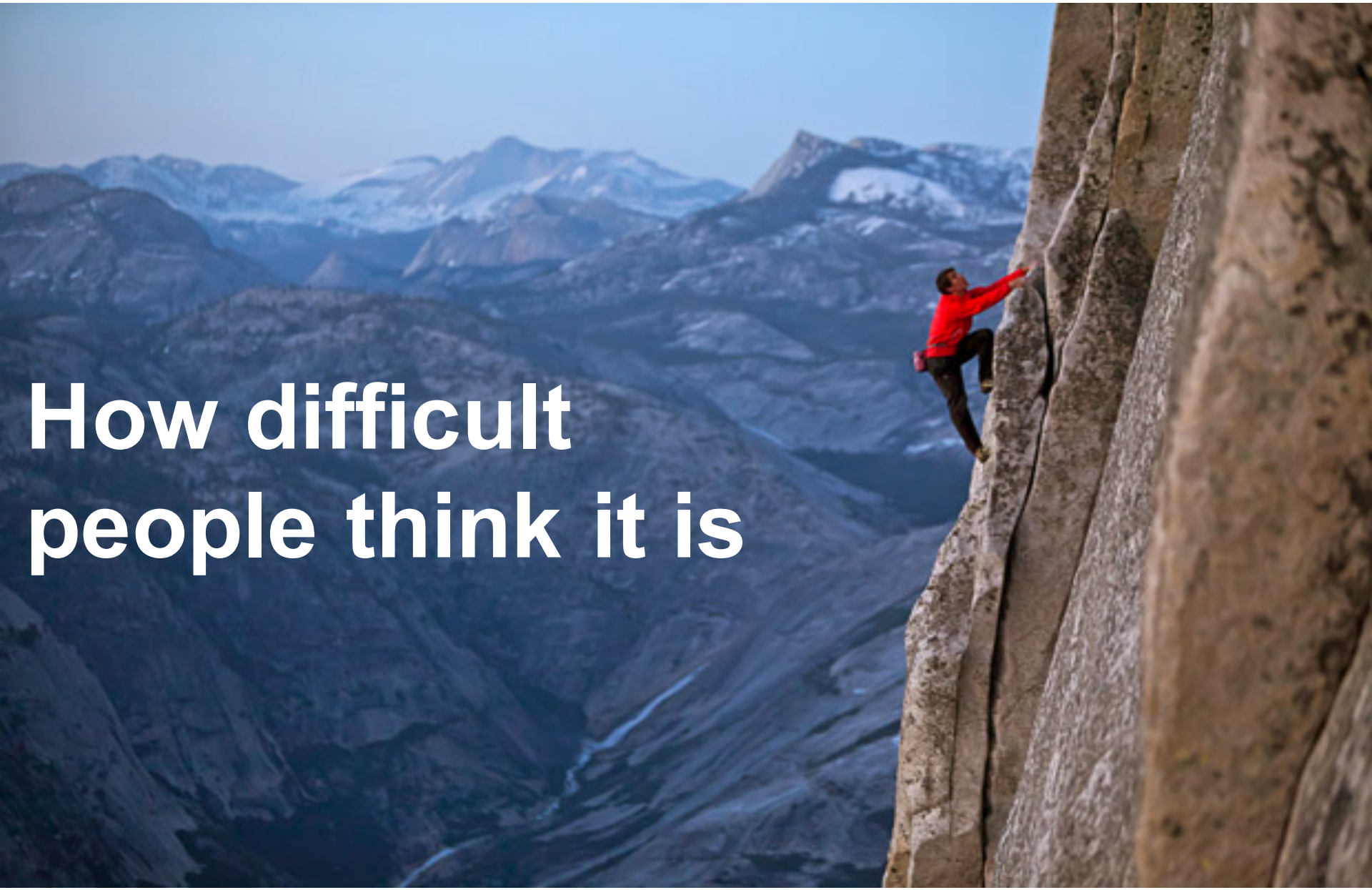


- **Descendant**
  - Sélection du cluster le moins cohérent
  - Subdivision du cluster le moins cohérent

04> DÉMO

"""icon.py - draw file ic
import Image, ImageDraw,
SHADOW = True
doc = Image.open('code.png')
draw = ImageDraw.Draw(doc)
draw.text((32, 8), py)
if SHADOW:
s = ImageOps.gra
e = ImageEnhance
s = e.enhance(0.
s.paste(doc, (0,
doc= s
logo = Image.open('brand.
doc.paste(logo, (120,96),
doc.save('py.ico')

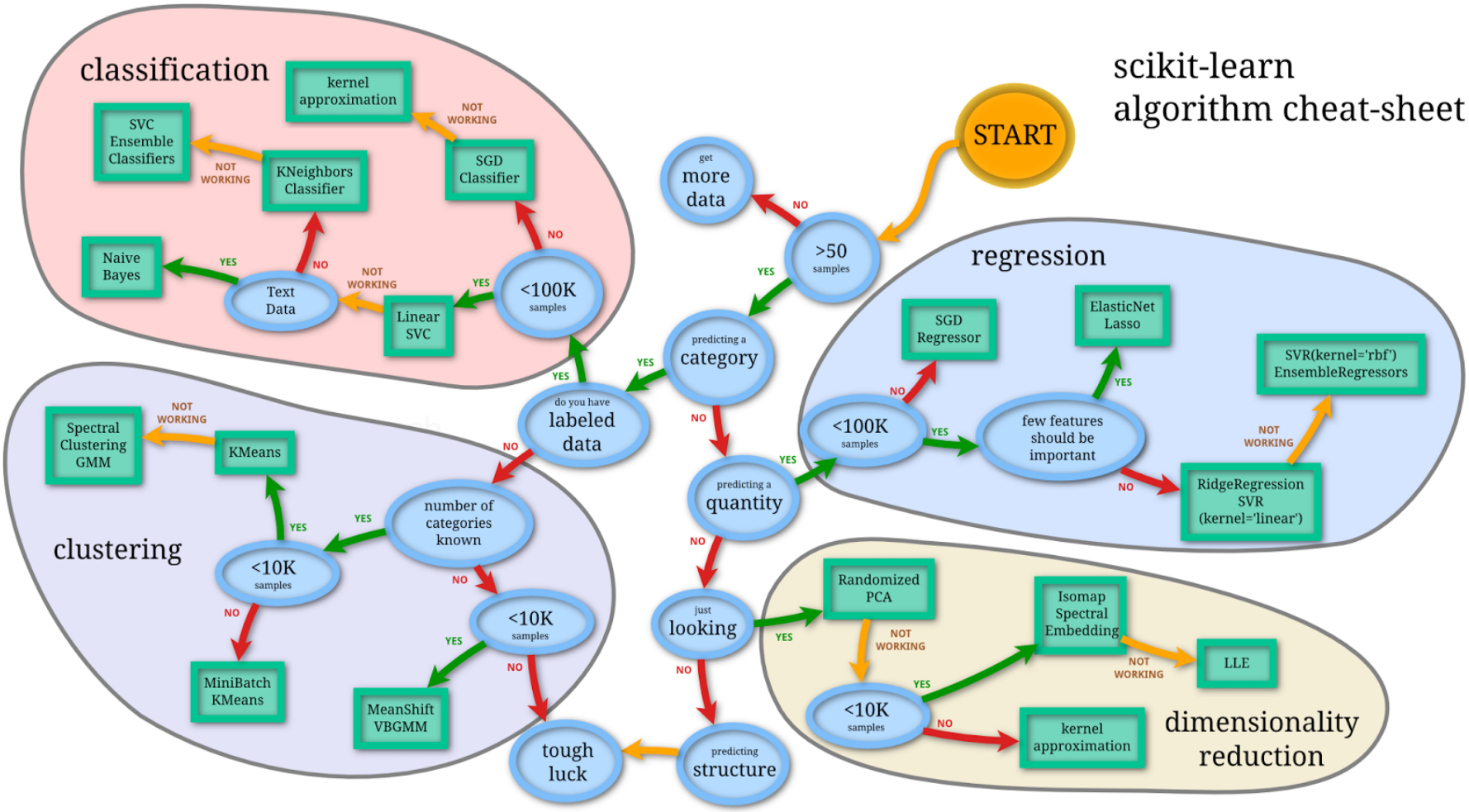
How difficult people think it is



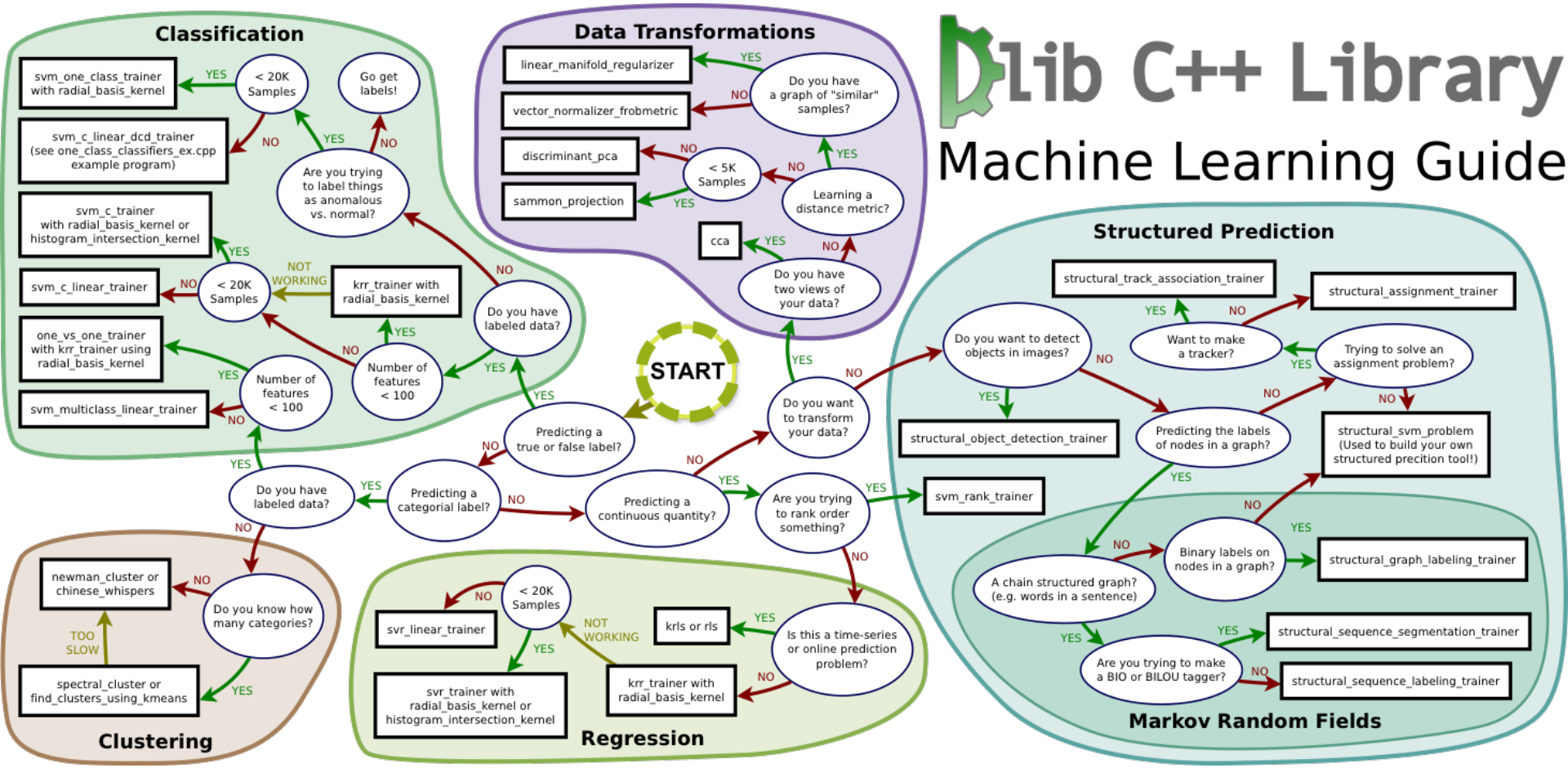
How it really is :

```
from sklearn.ensemble import RandomForestClassifier
clf = RandomForestClassifier(n_estimators=10)
clf = clf.fit(X, Y)
clf.predict(data)
```

QUELLE MÉTHODE CHOISIR ?



QUELLE MÉTHODE CHOISIR ?



QUELLE MÉTHODE CHOISIR ?

Microsoft Azure Machine Learning: Algorithm Cheat Sheet

This cheat sheet helps you choose the best Azure Machine Learning Studio algorithm for your predictive analytics solution. Your decision is driven by both the nature of your data and the question you're trying to answer.

START

Finding unusual data points

Discovering structure

Predicting categories

Predicting values

Three or more

Two

ANOMALY DETECTION

- **One-class SVM** — >100 features, aggressive boundary
- **PCA-based anomaly detection** — Fast training

CLUSTERING

- **K-means**

MULTI-CLASS CLASSIFICATION

- Fast training, linear model — **Multiclass logistic regression**
- Accuracy, long training times — **Multiclass neural network**
- Accuracy, fast training — **Multiclass decision forest**
- Accuracy, small memory footprint — **Multiclass decision jungle**
- Depends on the two-class classifier, see notes below — **One-v-all multiclass**

REGRESSION

- **Ordinal regression** — Data in rank ordered categories
- **Poisson regression** — Predicting event counts
- **Fast forest quantile regression** — Predicting a distribution
- **Linear regression** — Fast training, linear model
- **Bayesian linear regression** — Linear model, small data sets
- **Neural network regression** — Accuracy, long training time
- **Decision forest regression** — Accuracy, fast training
- **Boosted decision tree regression** — Accuracy, fast training, large memory footprint

TWO-CLASS CLASSIFICATION

- **Two-class SVM** — >100 features, linear model
- **Two-class averaged perceptron** — Fast training, linear model
- **Two-class logistic regression** — Fast training, linear model
- **Two-class Bayes point machine** — Fast training, linear model

- Accuracy, fast training — **Two-class decision forest**
- Accuracy, fast training, large memory footprint — **Two-class boosted decision tree**
- Accuracy, small memory footprint — **Two-class decision jungle**
- >100 features — **Two-class locally deep SVM**
- Accuracy, long training times — **Two-class neural network**

© 2015 Microsoft Corporation. All rights reserved.

Created by the Azure Machine Learning Team

Email: AzurePoster@microsoft.com

Download this poster: http://aka.ms/MLCheatSheet

Microsoft

TITANIC

Prédire les survivants

```
#titanic {
  float: none;
}
```

0.2%
98.4%

05> TO CONTINUE...

TO CONTINUE...

Quelques références

- Pour apprendre
  - Livres
    - **Data Science: fondamentaux et études de cas**
      Eric Biernat, Michel Lutz. Eyrolles, 2015
    - **Machine Learning**
      Tom Mitchell. McGraw Hill, 1997.
  - MOOC
    - https://www.coursera.org/learn/machine-learning
  - Kaggle
    - https://www.kaggle.com

- Techs
  - SQL
  - Python → Anaconda (ScikitLearn, NumPy, Pandas, SciPy, Matplotlib ...)
  - R
  - Matlab, Scilab
  - ...

- Tools
  - PyCharm IDE
  - IPython/Jupyter