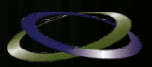

CHALLENGE
Prédire ...
Advanced Research Partners
咨询
BIG DATA - INTELLIGENCE ARTIFICIELLE – MACHINE LEARNING - INTELLIGENCE ECONOMIQUE

ALGO + LOCAUX + FORMATION + SOUTIEN BP

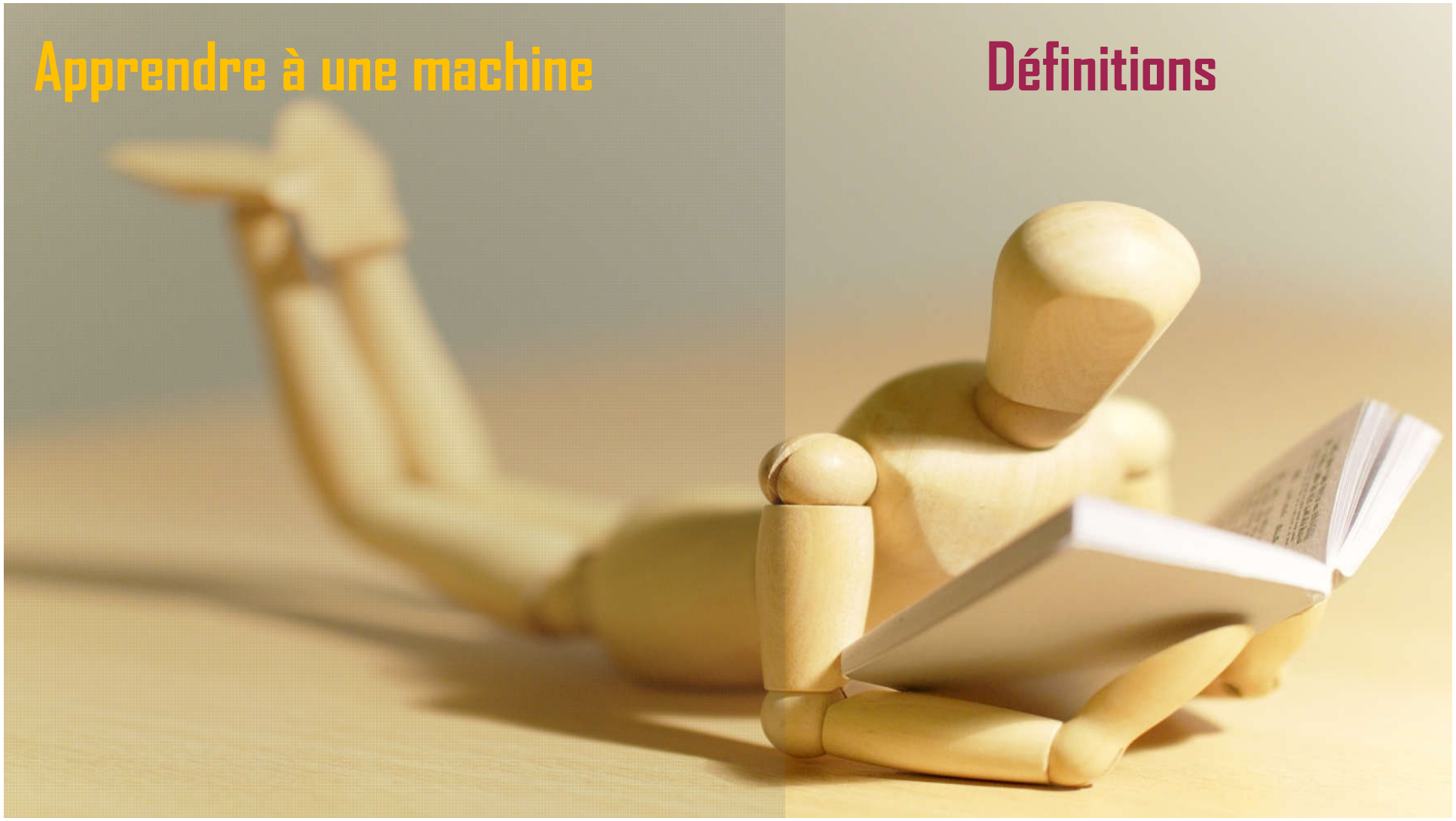
STARTUPS
LABORATOIRES
CELLULES INNOVATIONS

EQUITIES / PRESTATION DE SERVICE / RECONNAISSANCE ÉTERNELLE



Apprendre à une machine

Définitions

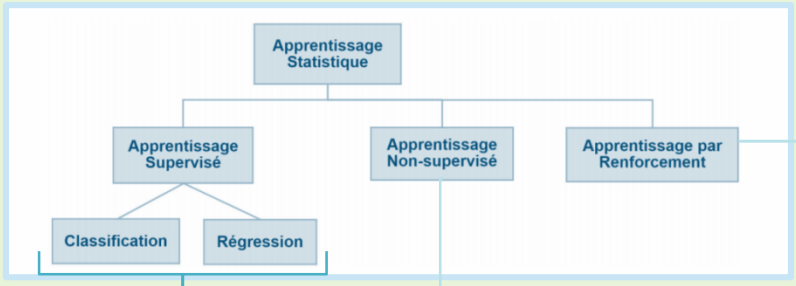


La **science des données** est l'extraction de connaissance à partir d'ensembles de données (Data Science : DS)

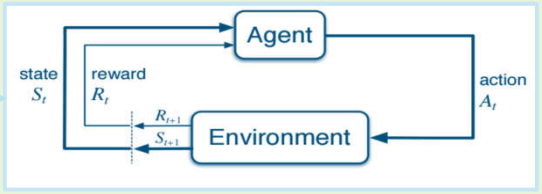
L'intelligence artificielle (IA) comporte les théories, les algorithmes et les techniques utilisables pour mettre en œuvre des machines capables de simuler l'intelligence (Artificial Intelligence : AI)

L'apprentissage artificiel par induction consiste à trouver les lois générales les plus plausibles qui expliquent des phénomènes particuliers que l'on observe (Machine Learning : ML)

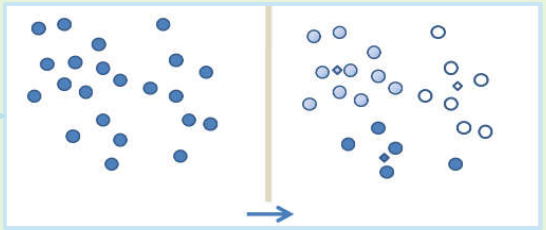
L'apprentissage artificiel est **supervisé** quand à partir d'une **liste d'observations** comportant des variables supposées explicatives et des valeurs de réponses on crée un modèle pour prédire des valeurs de réponse inconnues à partir de nouvelles observations ne comportant que les variables explicatives (Supervised Machine Learning : SML)



Ce que nous allons apprendre à faire dans ce challenge



Récompense
Conditionnement
Au fur et à mesure



Clustering
Data mining
Profils types

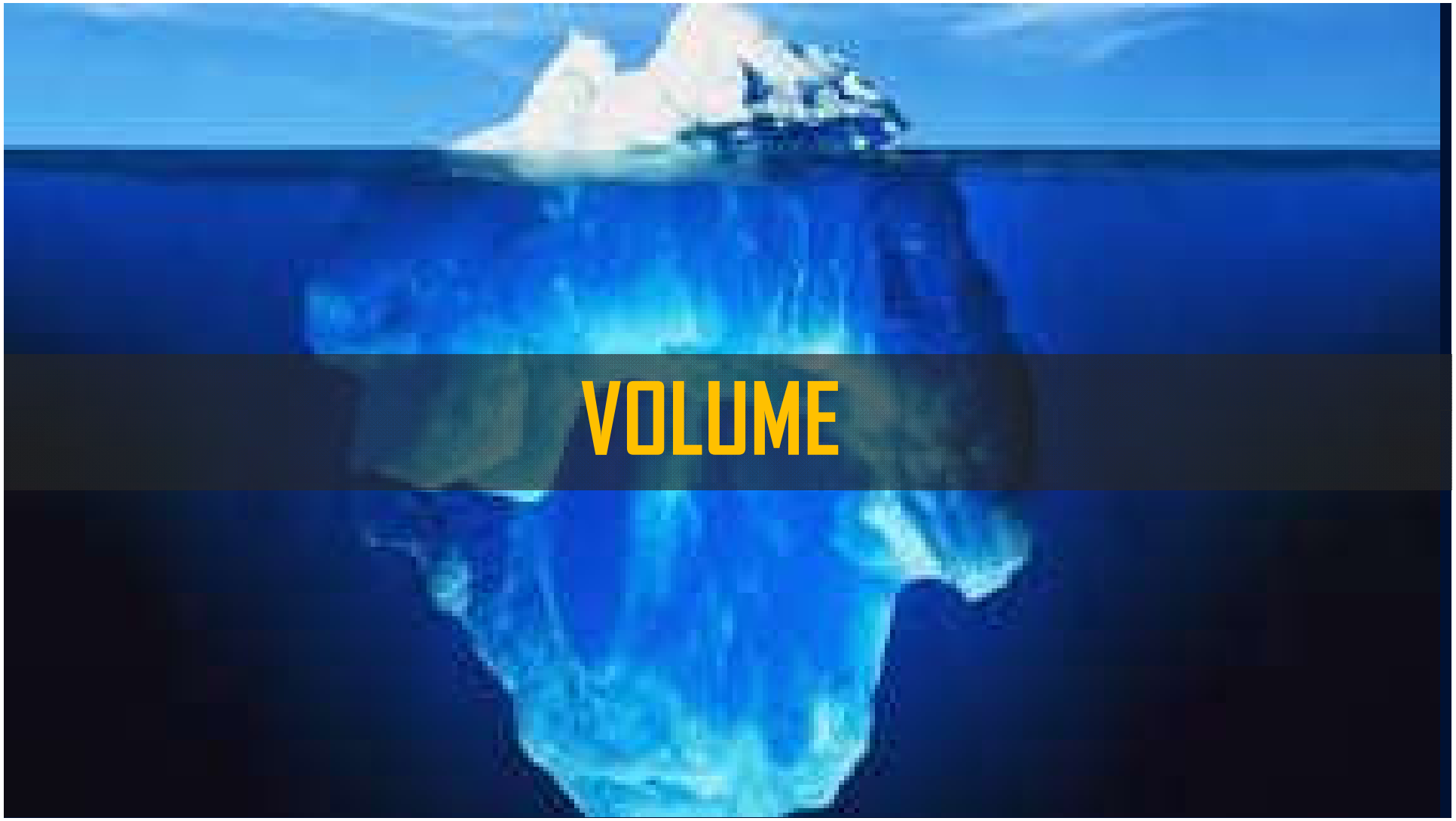
OBSERVATIONS

Variables explicatives

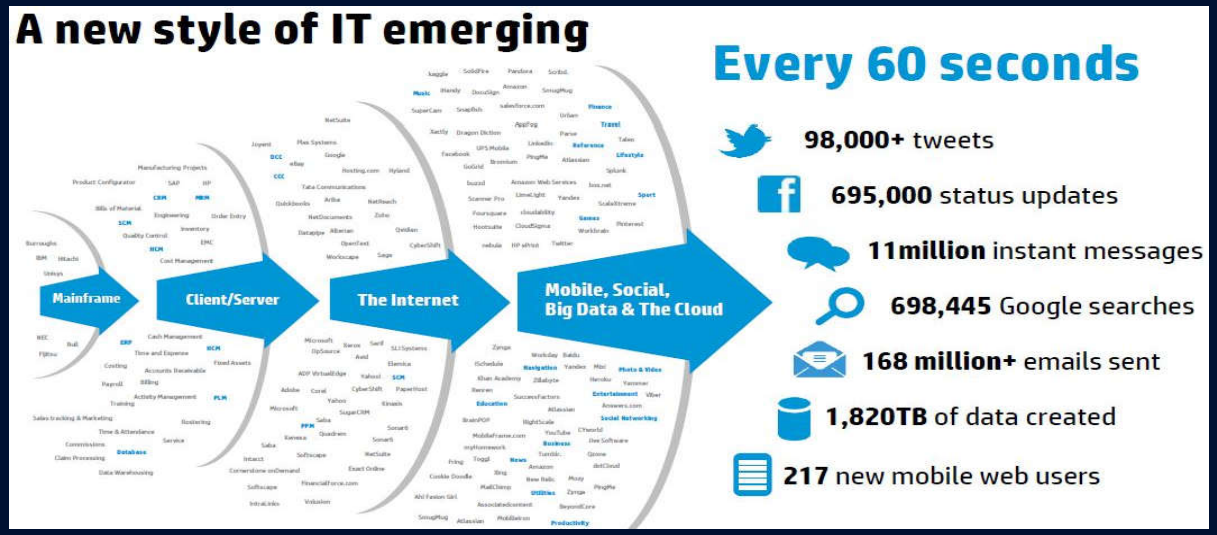
ID	LIMIT_BAL	SEX	EDUCATION	MARRIAGE	AGE	PAY_1	PAY_2	PAY_3	PAY_4	PAY_5	PAY_6	BILL_AMT1	BILL_AMT2	BILL_AMT3	BILL_AMT4	BILL_AMT5	BILL_AMT6	PAY_AMT1	PAY_AMT2	PAY_AMT3	PAY_AMT4	PAY_AMT5	PAY_AMT6	DEF
2	120000	2	2	2	26	-1	2	0	0	0	2	2682	1725	2682	3272	3455	3261	0	1000	1000	1000	0	2000	1
3	90000	2	2	2	34	0	0	0	0	0	0	29239	14027	13559	14331	14948	15549	1518	1500	1000	1000	1000	5000	0
4	50000	2	2	1	37	0	0	0	0	0	0	46990	48233	49291	28314	28959	29547	2000	2019	1200	1100	1069	1000	0
5	50000	1	2	1	57	-1	0	-1	0	0	0	8617	5670	35835	20940	19146	19131	2000	36681	10000	9000	689	679	0

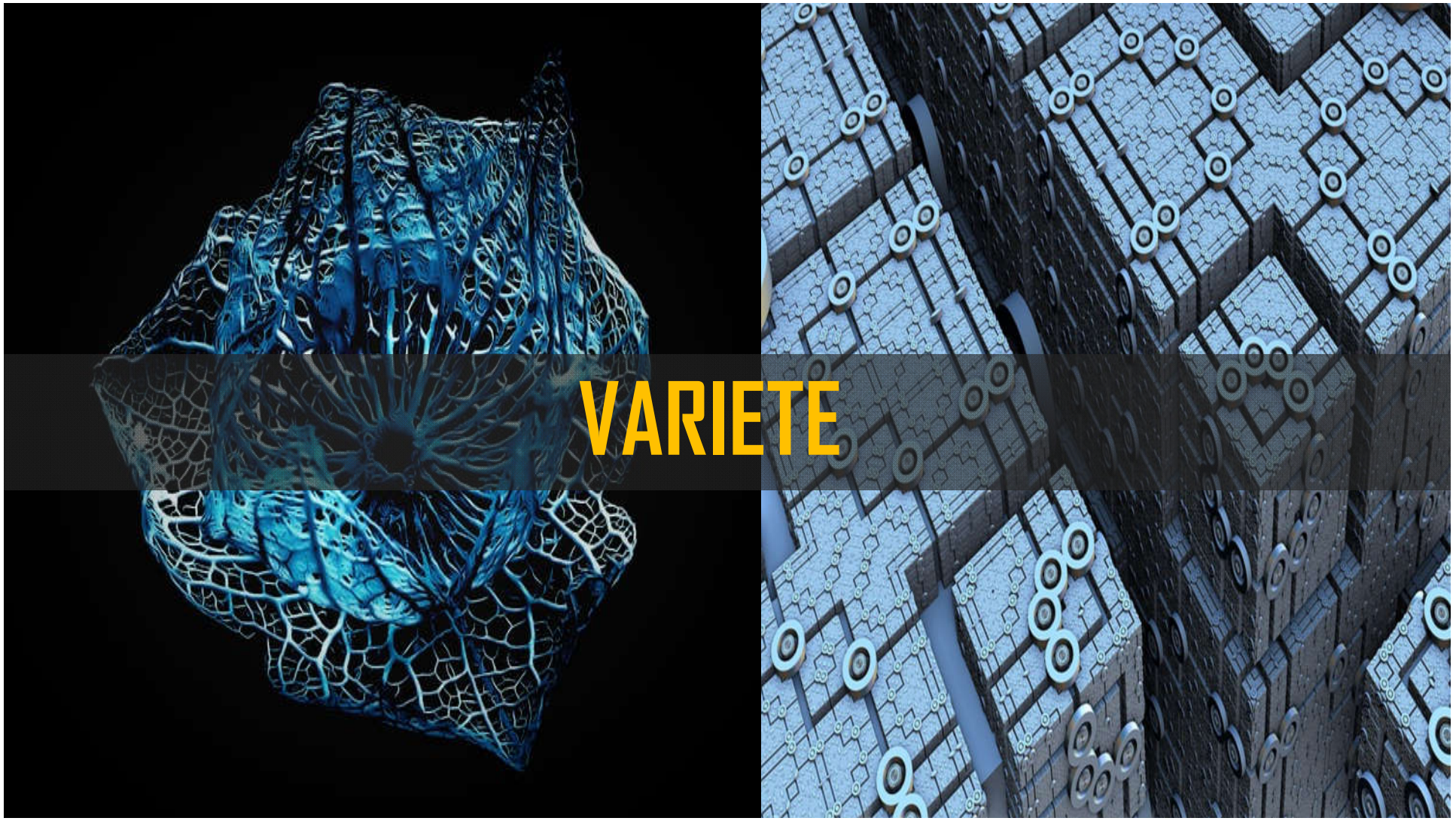
Réponse = variable expliquée


R
DATA SCIENCE vs BIG DATA


VOLUME

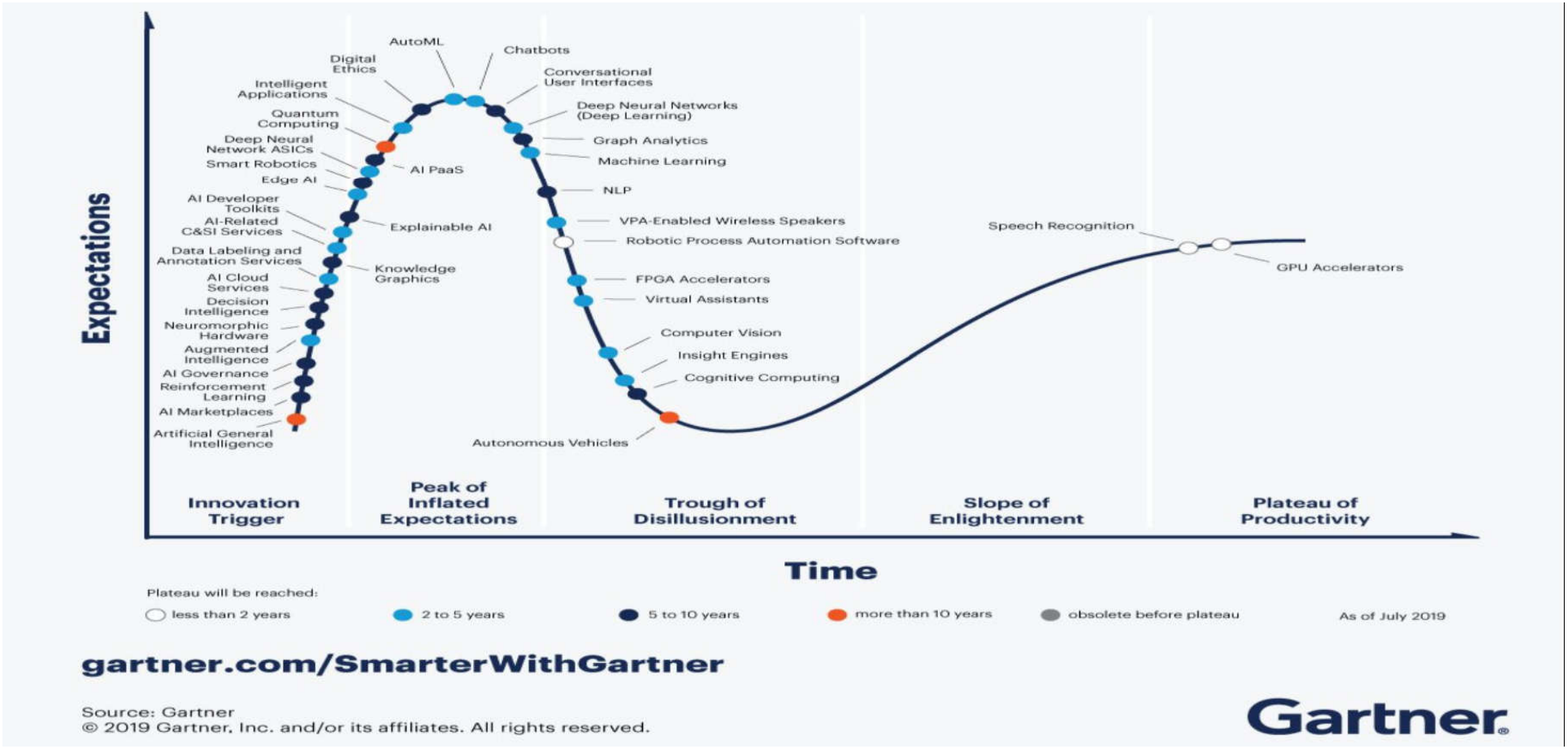
VELOCITE


A new style of IT emerging
Mainframe
Client/Server
The Internet
Mobile, Social,
Big Data & The Cloud
Every 60 seconds
98,000+ tweets
695,000 status updates
11million instant messages
698,445 Google searches
168 million+ emails sent
1,820TB of data created
217 new mobile web users


VARIETE


VERACITE

HYPE CYCLE



gartner.com/SmarterWithGartner

Source: Gartner

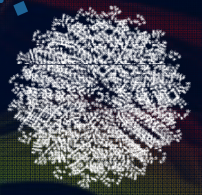
Gartner

UNE METHODE

LE PROBLEME A RESOUDRE

DATA

INFORMATION



CONNAISSANCE

ACQUERIR
NETTOYER

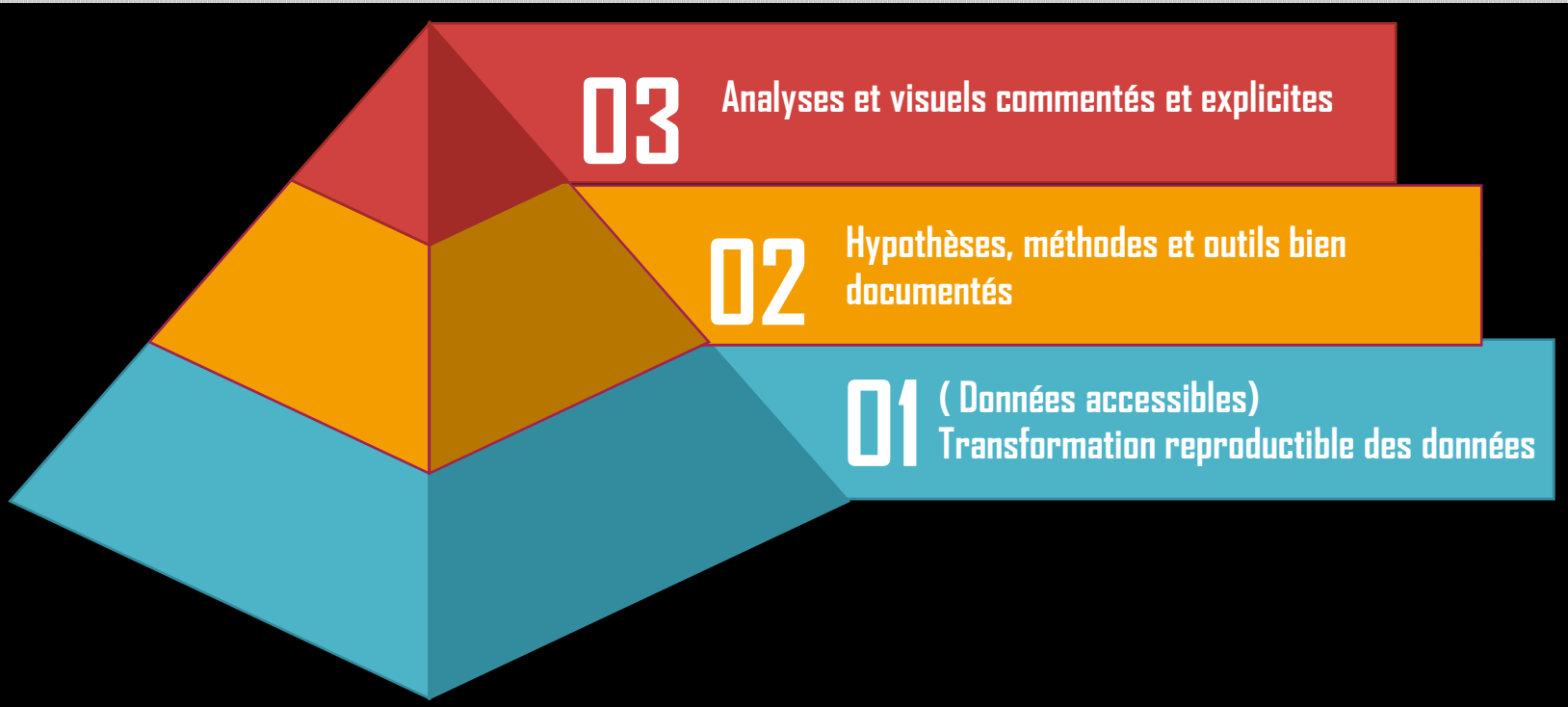
DONNER
DU
SENS

CREER
UN
MODELE


/!\ REPRODUCTIBLE

DS REPRODUCTIBLE :

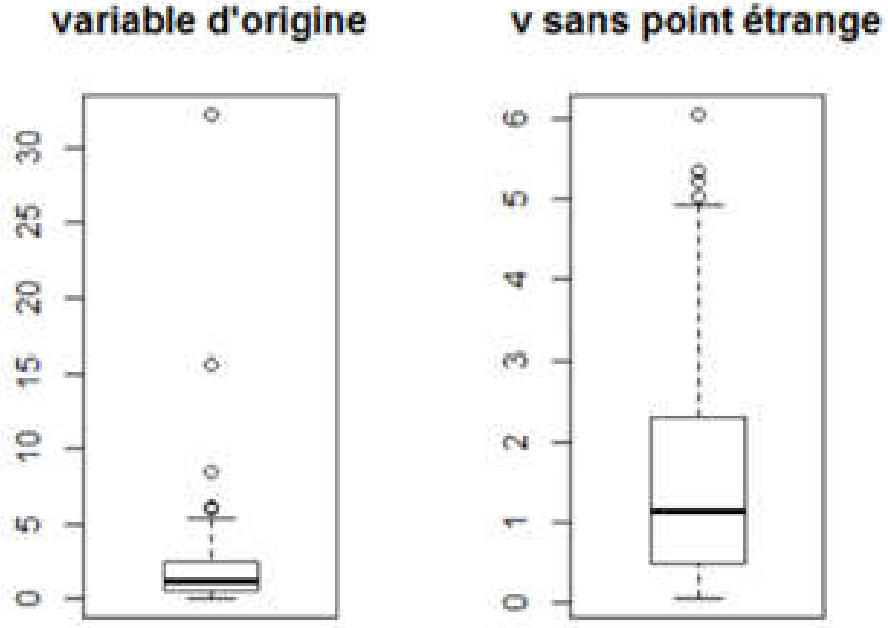
Évaluation du document d'accompagnement



TYPES DE DONNEES

- Variables catégorielles nominales, par exemple :
  - *femme, homme*
  - *cyan, magenta, jaune*
  - *oui, non*
- Variables catégorielles ordinales, par exemple :
  - *pas d'accord, indifférent, d'accord*
  - *sur le podium, dans le peloton de tête, dans le peloton de queue*
- Variables quantitatives d'intervalles, par exemple :
  - *années (de -2000 à + 3000)*
  - *niveau par rapport à la mer*
- Variables quantitatives de ratios, par exemple :
  - *comptage de la population des départements français*
  - *consommation d'électricité en kW*

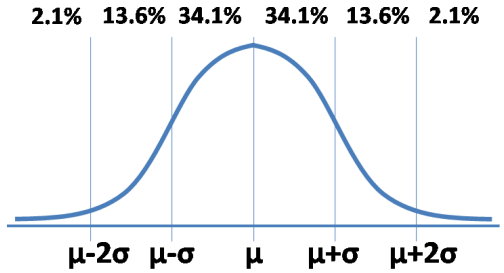
NETTOYAGE DES DONNEES



POINTS ABERRANTS (OUTLIERS)

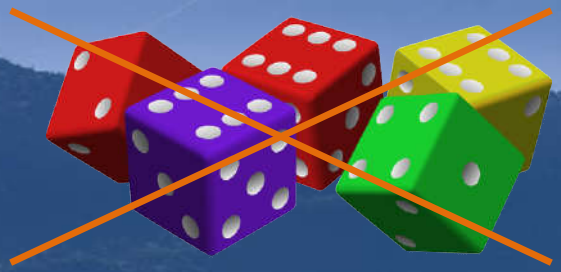
HASARD OU PAS ?

DISTRIBUTION

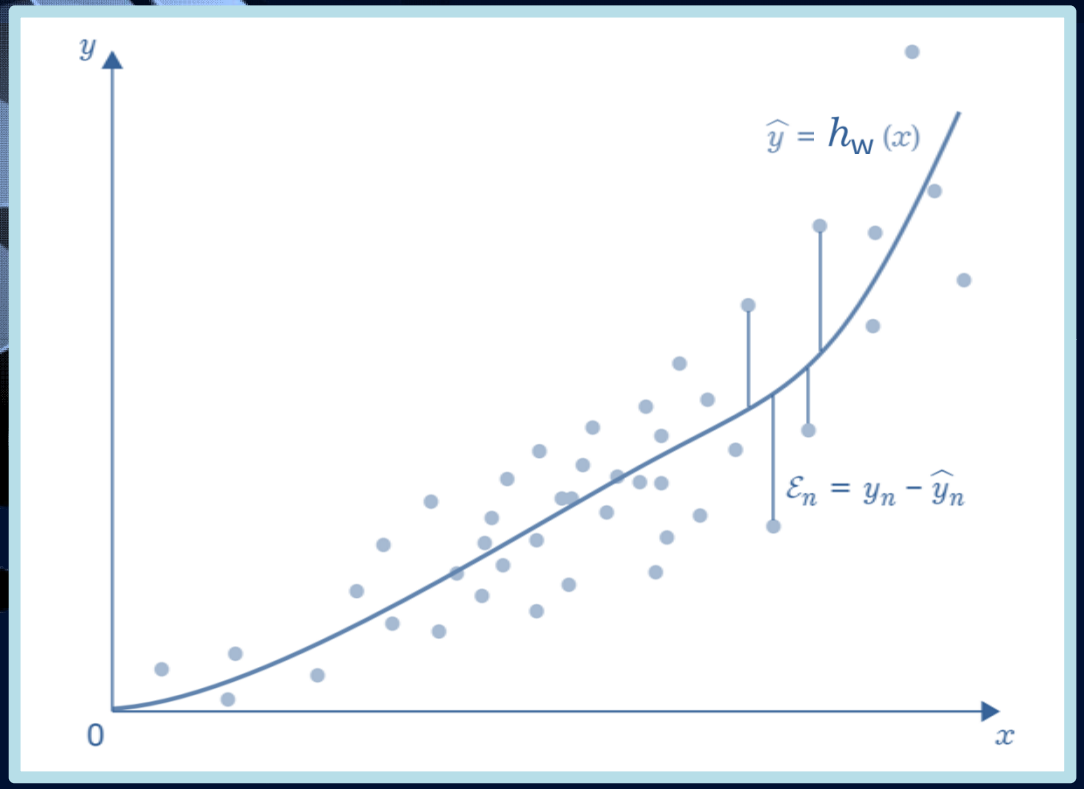


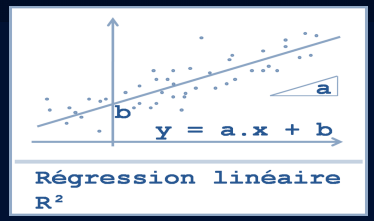
Loi Normale
Moyenne,écart type

Dissymétrie
Skewness



MODELISATION = trouver un modèle hypothétique *h*

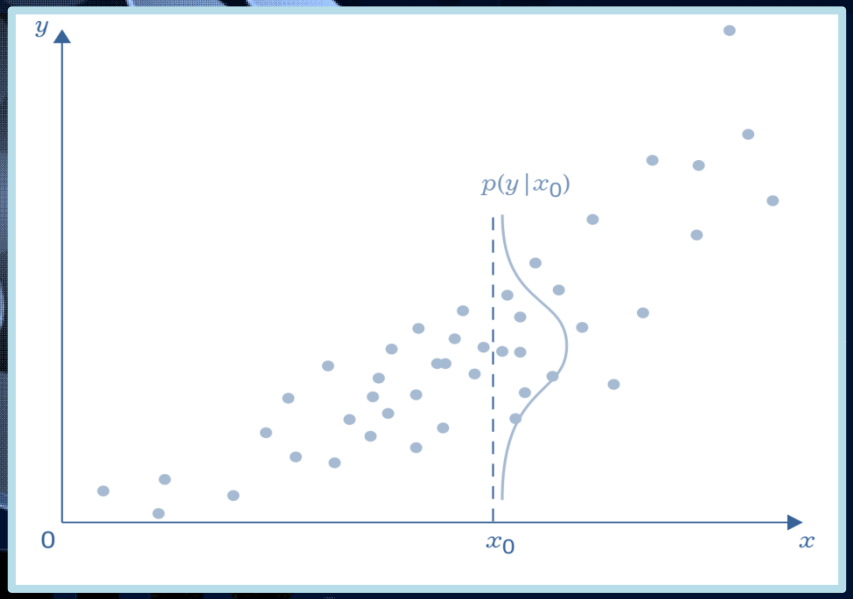




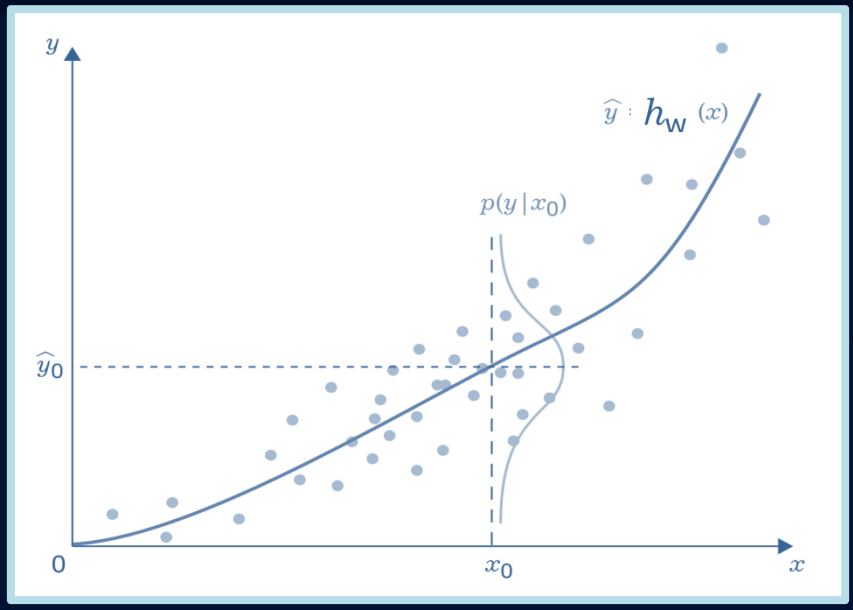
Minimiser
la somme des erreurs

$$E(w) = \sum_{n=1}^{N} \varepsilon_n^2 = \sum_{n=1}^{N} \left(y_n - h_w(x_n)\right)^2$$

MODELISATION



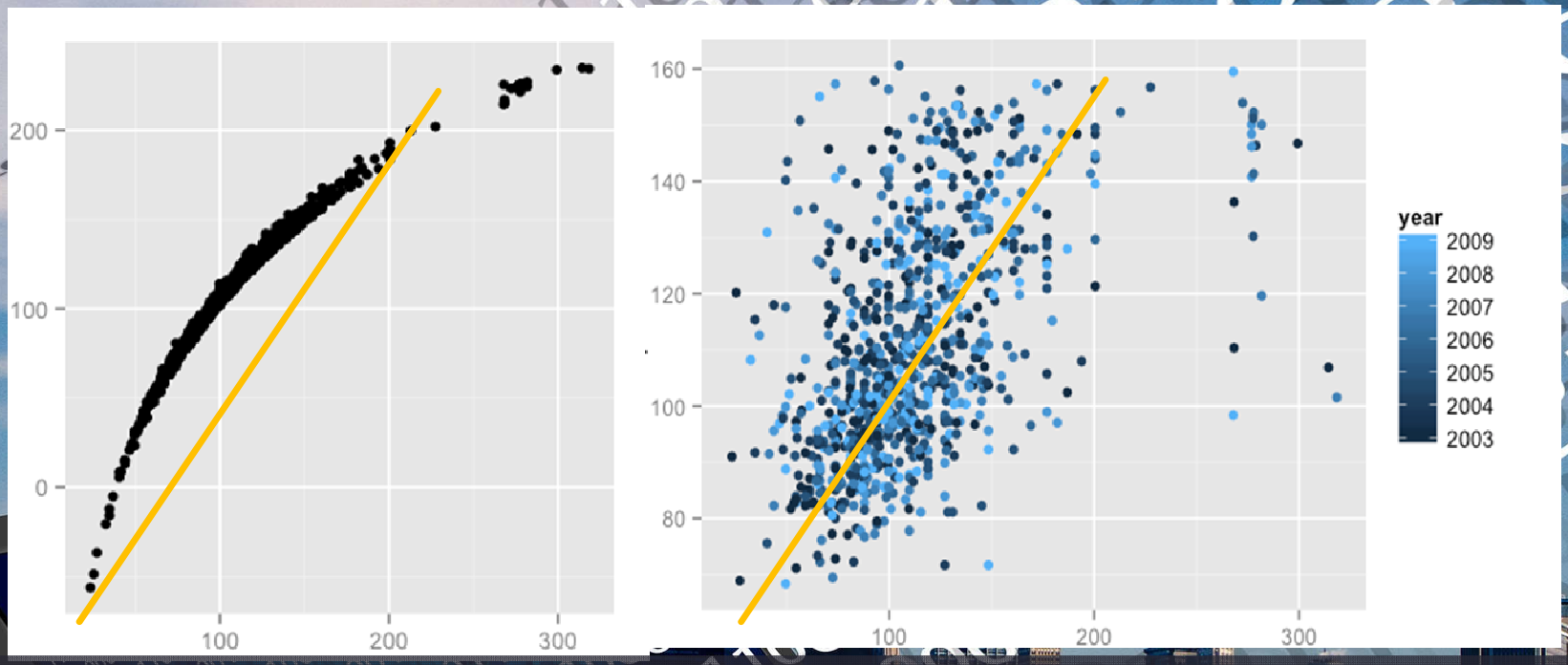
Probabilité d'une valeur *y* pour un *x* connu =
Valeur maximale de la densité conditionnelle



On cherche les paramètres *w* (weights) d'une fonction *h*
d'un type de modèle que l'on pense intuitivement plausible

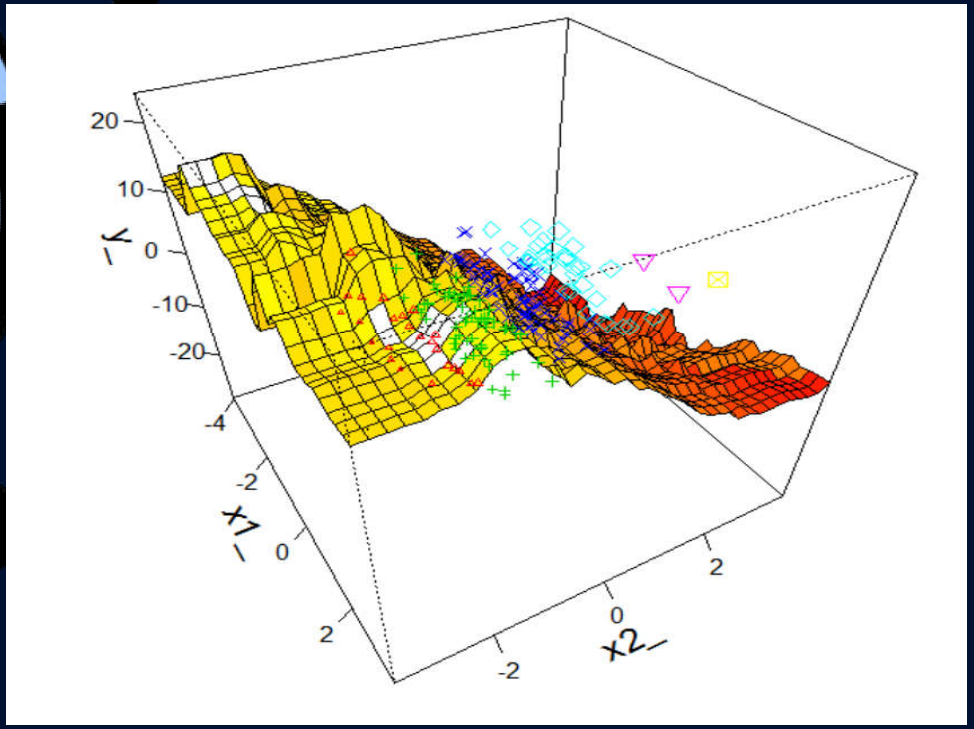
PREDICTION

Apprentissage supervisé



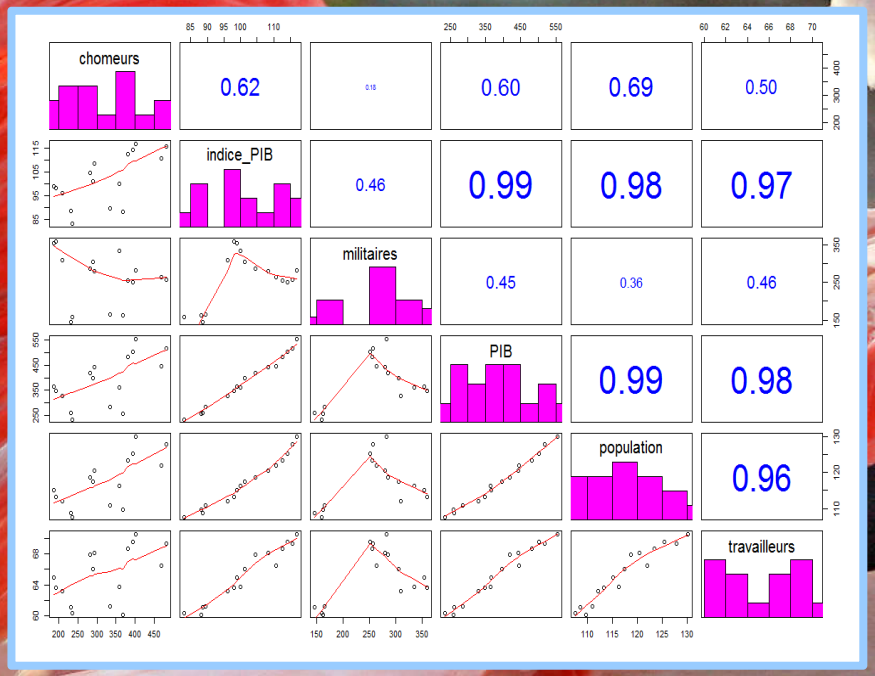
UNE MACHINE APPREND QUAND ELLE FAIT EVOLUER SON PARAMETRAGE AUTOMATIQUEMENT EN FONCTION DES CIRCONSTANCES (DATA)

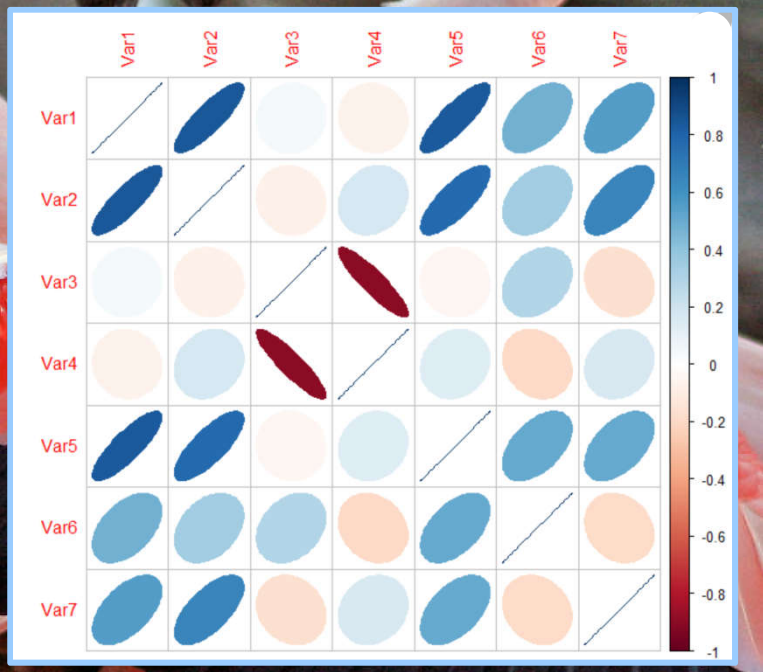
DIMENSIONS



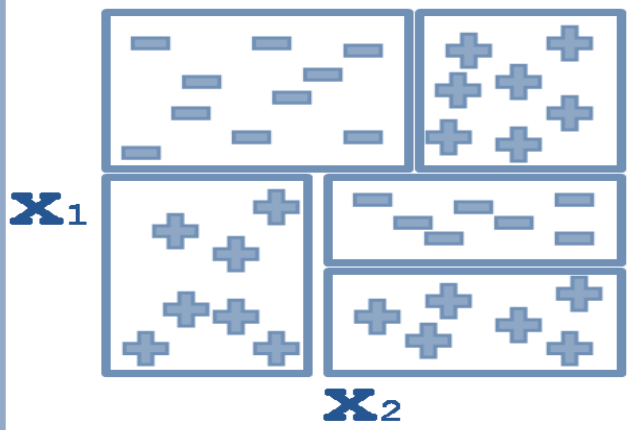
RELATION ENTRE FEATURES - 2 A 2

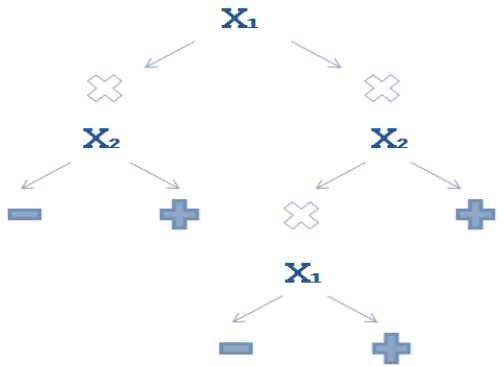
Corrélation – dépendance – redondance ?





ARBRES (TREES) TROUVER DES REGLES

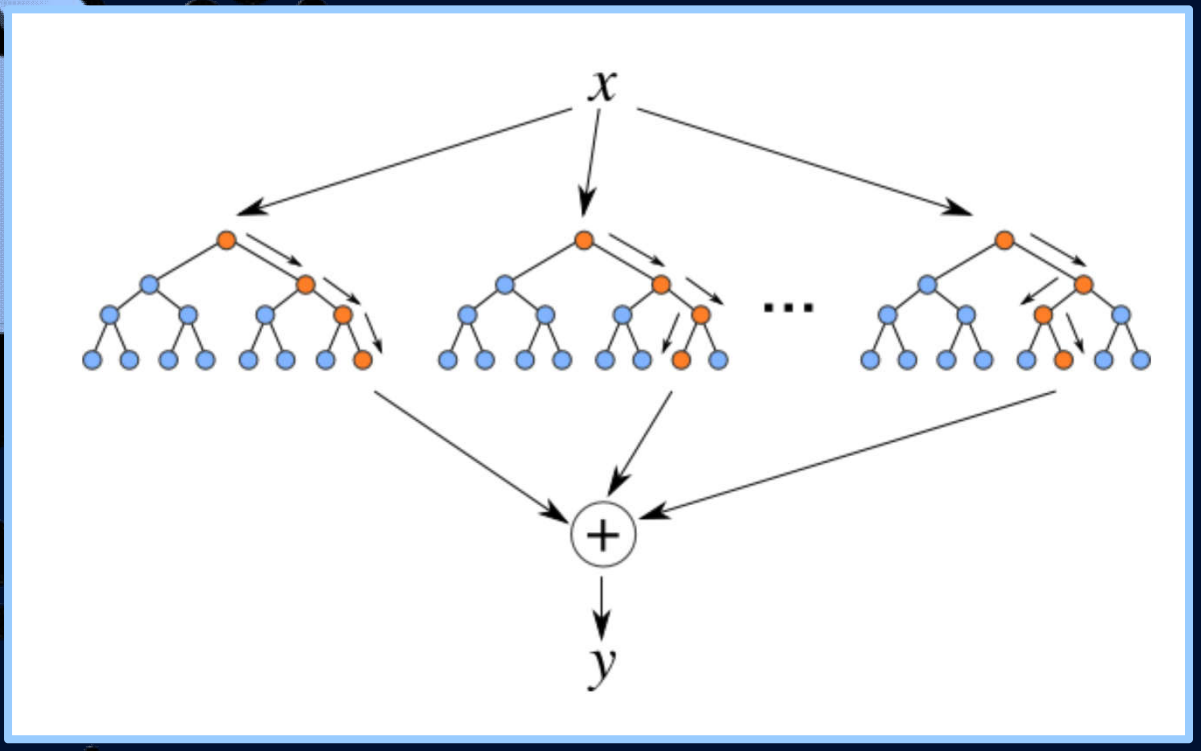




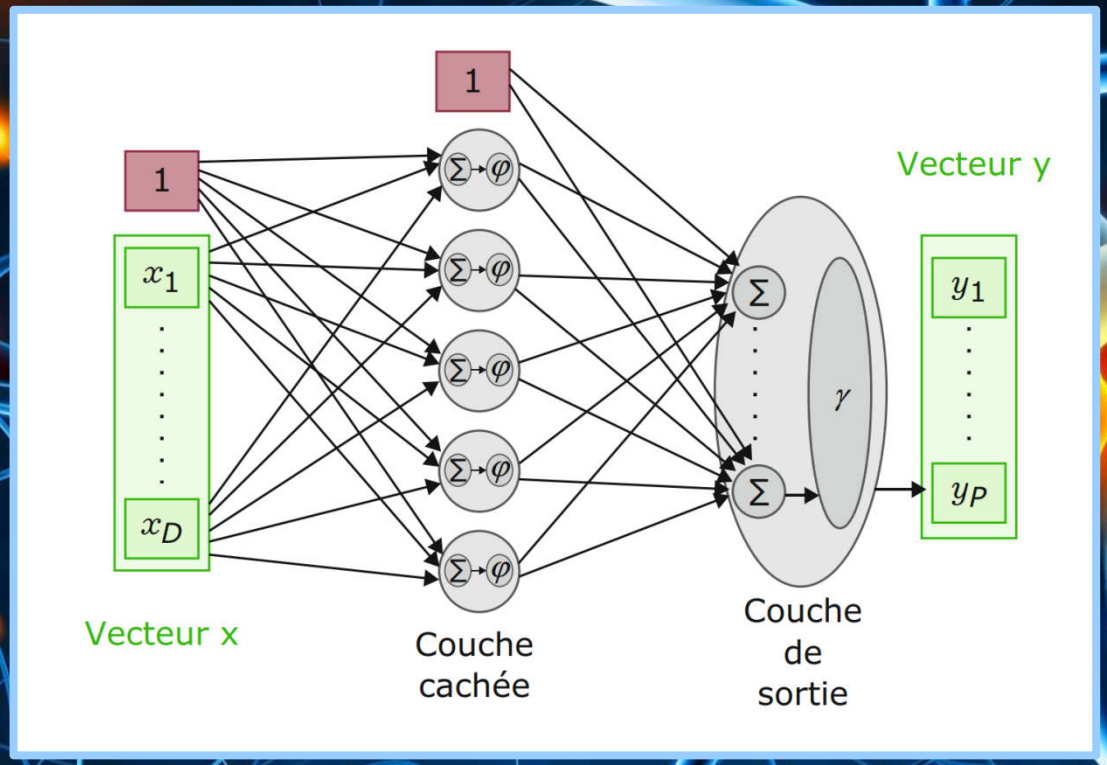
Arbres de décision

Données -> arbre

FORETS : DEMOCRATIE ENTRE ARBRES

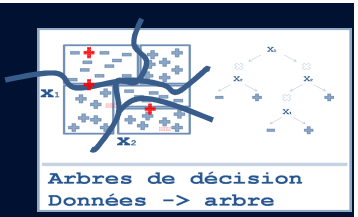


RESEAUX NEURONAUX & DEEPLEARNING




QUESTIONNER SES CERTITUDES

MESURER LES ECARTS



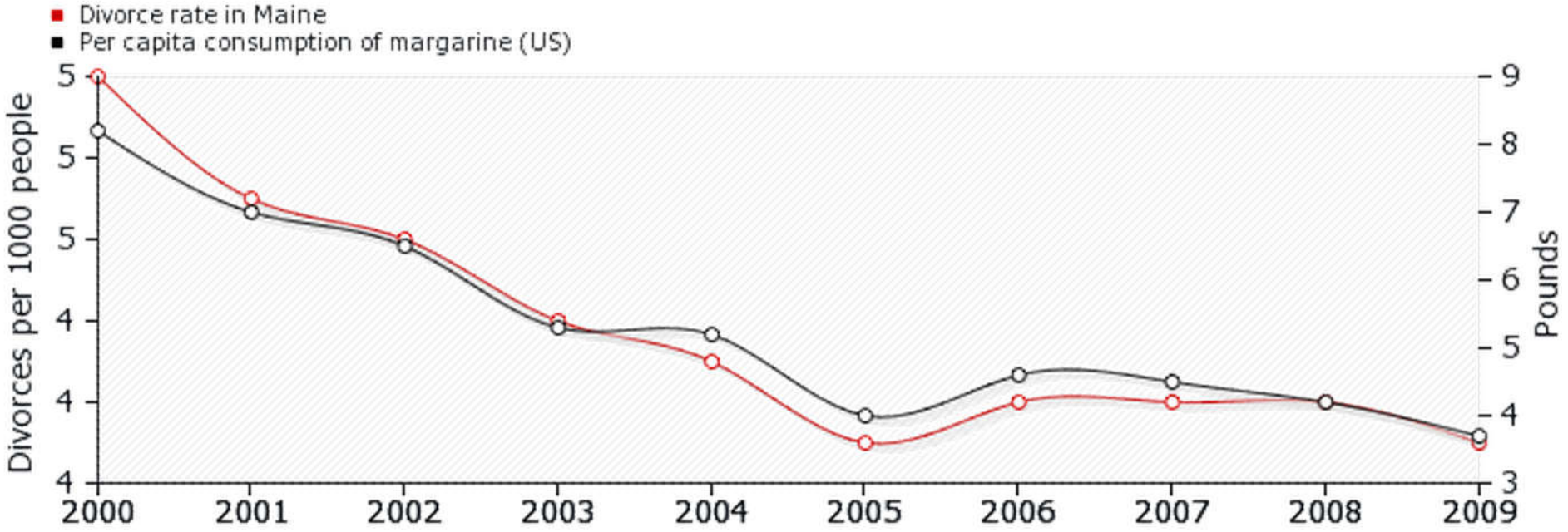
Prédiction \ Réalité	−	+
−	Vrais −	Faux −
+	Faux +	Vrais +

Matrice de confusion

Erreurs de classification

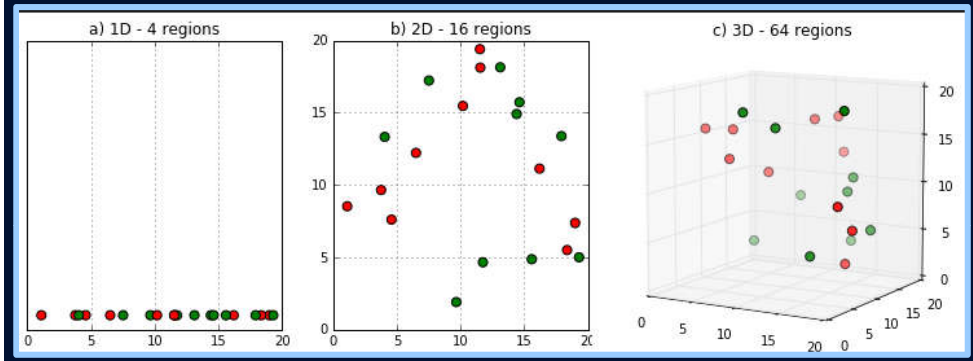
(60 % Entrainement + 20 Test) + 20 Validation

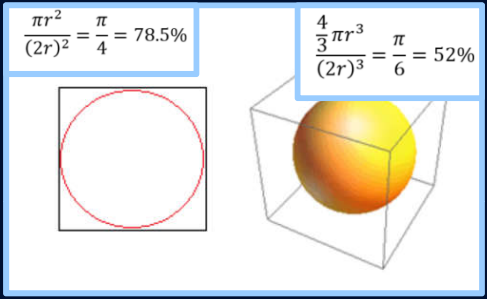
SERIES TEMPORELLES

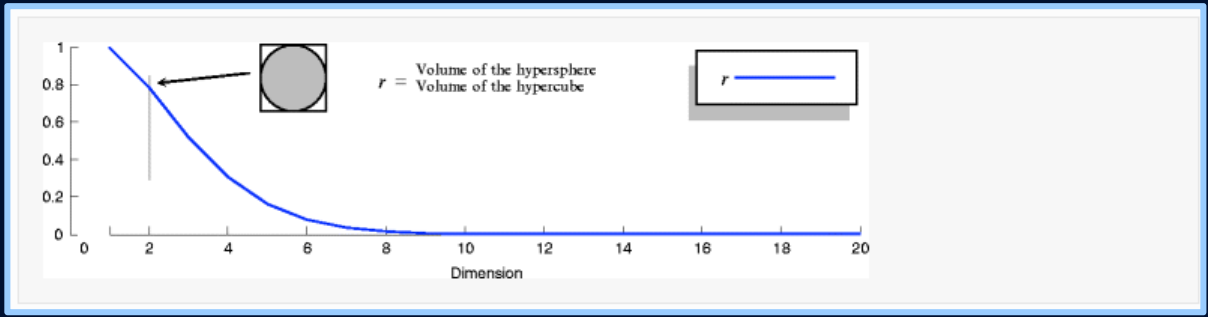


GARDER UN PEU DE RECUL /!\

CURSE OF DIMENTIONALITY

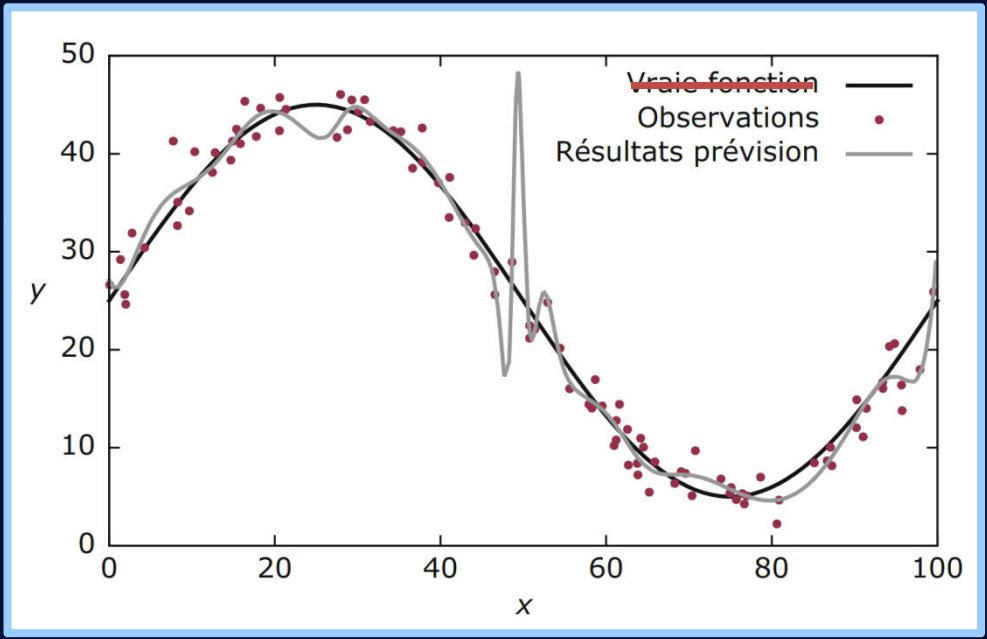


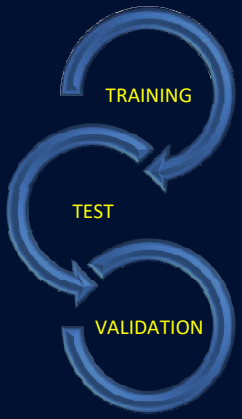




OVERFITTING

SURAPPRENTISSAGE





TECHNIQUES DE REDUCTION DE DIMENSION

•ENLEVER DES VARIABLES A IMPACT FAIBLE SUR LE MODELE

•ENLEVER LES VARIABLES REDONDANTES

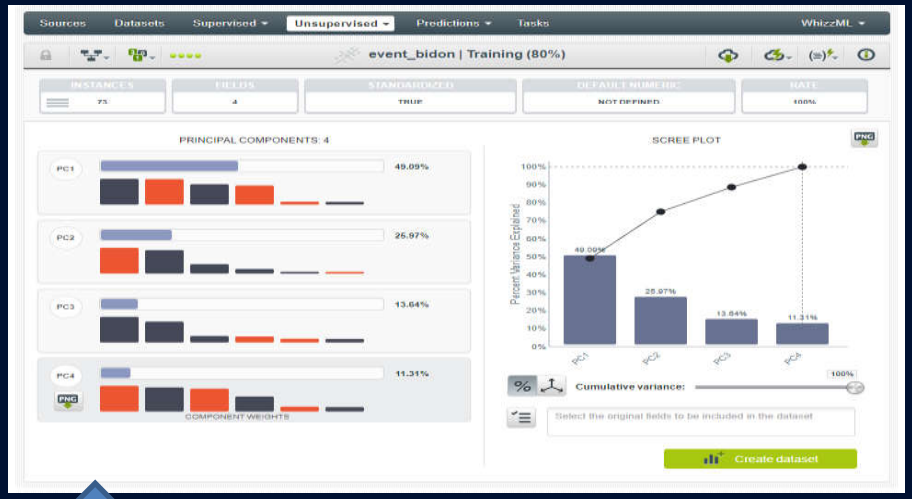
- •DEPENDANCES
- •SEMANTIQUE SIMILAIRE

•SYNTHETISER DES COLONNES

- MOYENNE
- ECART TYPE
- VARIANCE
- MAX
- MIN
- SEUILS → CLASSES
- FORME D'UNE SERIE TEMPORELLE

ALGO PROFESSIONNELS

- ENTROPIE
- ANALYSE EN COMPOSANTES PRINCIPALES (PCA)




Version numérique
OFFERTE !
www.editions-eni.fr
Data Scientist
et langage R
Guide d'autoformation
à l'exploitation intelligente
des Big Data
2e édition
Henri LAUDE
Eva LAUDE
Fichiers complémentaires
à télécharger
eni
epsilon
MERCI