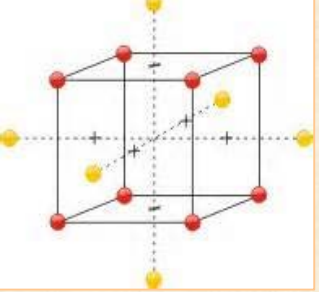


Génistat Conseils

POLYTECHNIQUE MONTRÉAL

Statistique, Science des données, Intelligence artificielle : liens, différences, convergence

Bernard CLÉMENT, PhD

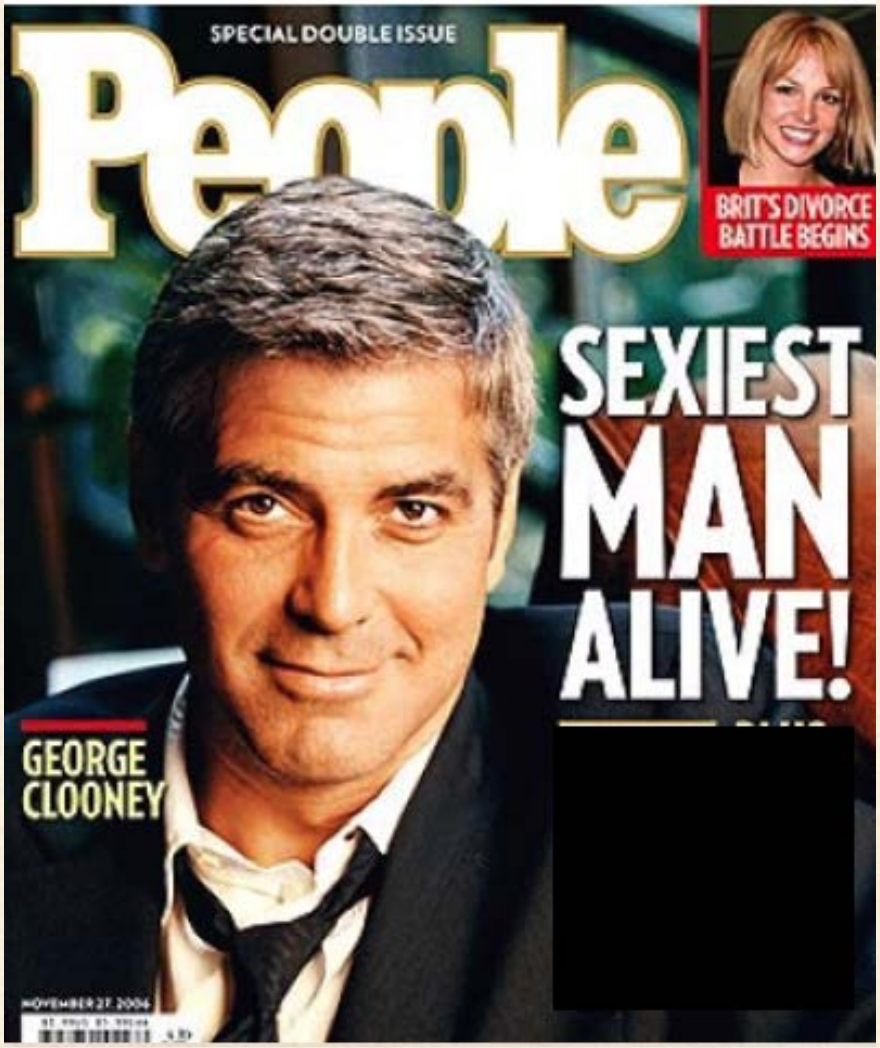
Société Statistique de Montréal : 26 avril 2018

P L A N

- Questions et Objectif
- Buzzwords et Terminologie
- Statistique Classique (SC)
- Science des Données (DS)
- Mégadonnées (Big Data) (BD)
- Apprentisage Machine (ML)
- Intelligence Artificielle (AI)
- Rôle du statisticien
- Réponses et Conclusions

" I keep saying the sexy job in the next ten years will be statisticians
People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s ? "

Hal Varian, PhD, chief economist, Google



" I keep saying the sexy job in the next ten years will be statisticians
People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s ? "

Hal Varian, PhD, chief economist, Google

Questions

Q1 : quel est l'avenir de la science statistique ?

Q2 : le métier de statisticien est-il en train de changer ?

Q3 : comment doit-on former les futurs statisticiens ?

Q4 : l'inférence statistique a-t-elle encore une place à l'ère des mégadonnées ?

Q5 : quelle contribution la statistique peut-elle apportée dans le domaine de l' intelligence artificielle ?

Q6 : comment faire pour rehausser l'appréciation et la reconnaissance du rôle du statisticien dans la société ?

Q7 : …. vos questions ….

Objectif : proposer des éléments de réponse

Buzzwords - Terminologie

DS : Data Science … ensemble des méthodes et outils orientés visant à apprendre avec les données et résoudre des problèmes … compréhension / utilisation
… science essentiellement **plusridisciplinaire**

ML : Machine Learning (apprentissage machine) … ensemble d'algorithmes, méthodes, outils, pour développer des modèles visant à améliorer le processus d'apprentissage avec des données … orientés **prédiction**

DM : Data Mining (fouille des données) … *« nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data »* G. Piatesky-Shapiro
…. **DM** maintenant remplacé par **ML**

AP : Analyse Prédictive … analyse pour prédire quelque chose
méthode **pourrait être non statistique**

BD : Big Data (mégadonnées) … données trop grosses en **taille** (à définir) et **complexité** nécessitant leur traitement informatique / statistique avec technologies sur des systèmes distribués en parallèle (Hadoop et autres)
…. **BD** terme galvaudé trop emphase **quantité**
pas assez sur la **qualité** (véracité)

Buzzwords - Terminologie

IA : Intelligence Artificielle … systèmes électronique qui ressemble à l'intelligence humaine d'une certaine manière
…. différentes interprétations de ce qu'est IA
test de A.Turin

voir *"Donner un sens à l'intelligence artificielle"* (rapport Cédric Villani)

autres

apprentissage supervisé …. apprentissage non supervisé

données structurées … données non structurées (images, textes,..)

apprentissage profond … réseaux de neurones

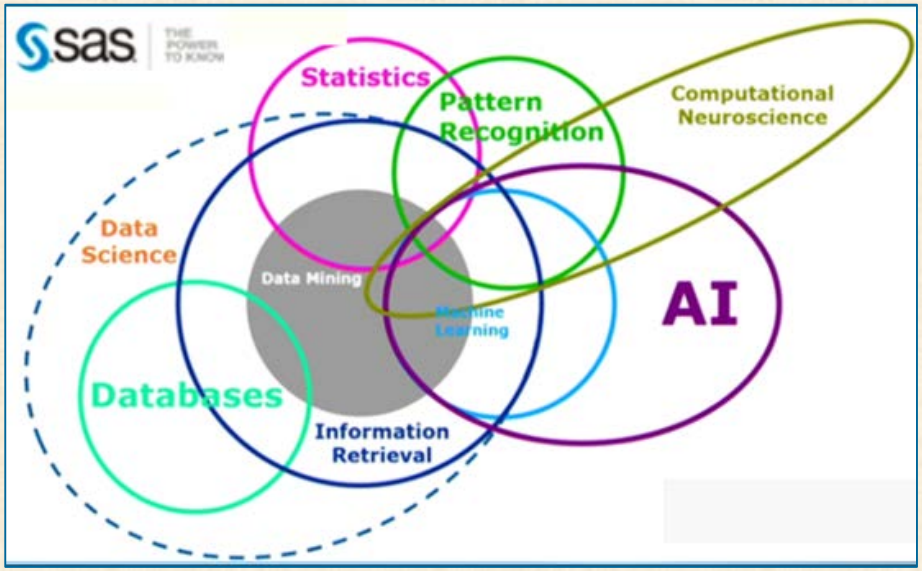
infonuagique ... systèmes distribués

IOT … internet des objets … réseaux de capteurs

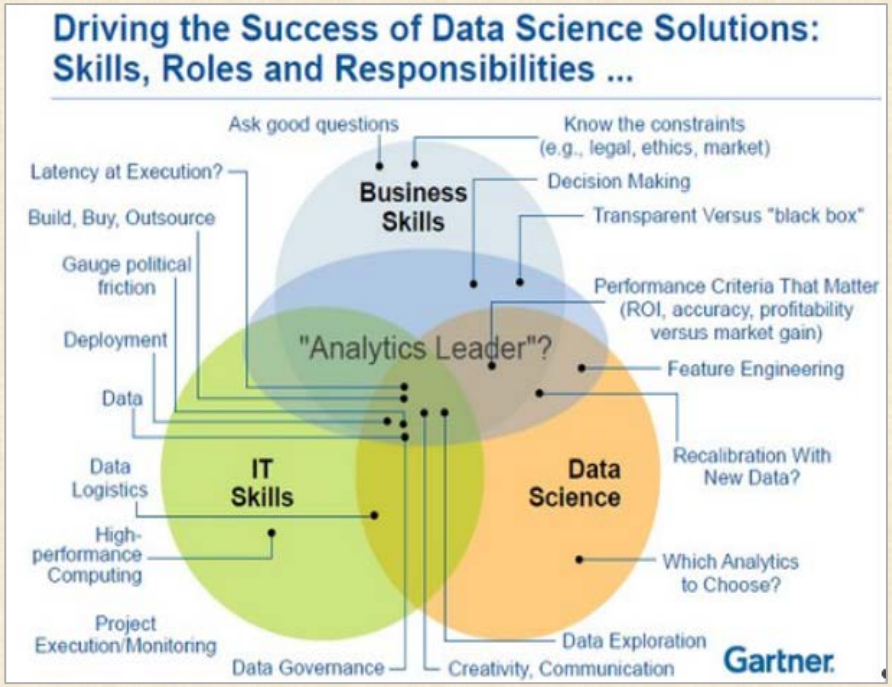
open source software ... R , Python , …

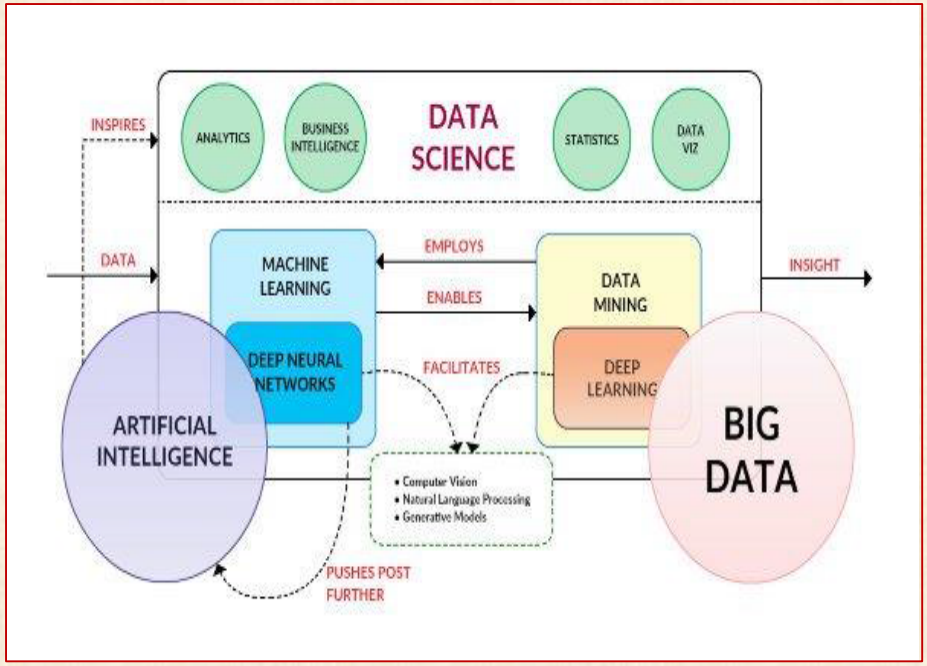
technologies … GPU (puces graphiques traitement parallèle)

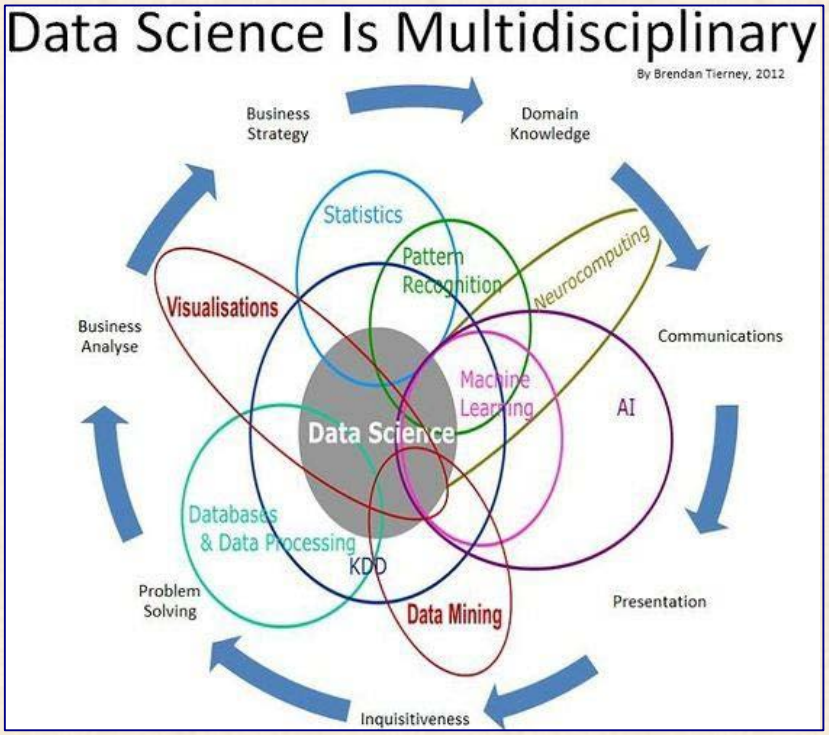
économie numérique ... GAFAM (Google Amazon Facebook Alibaba Microsoft)

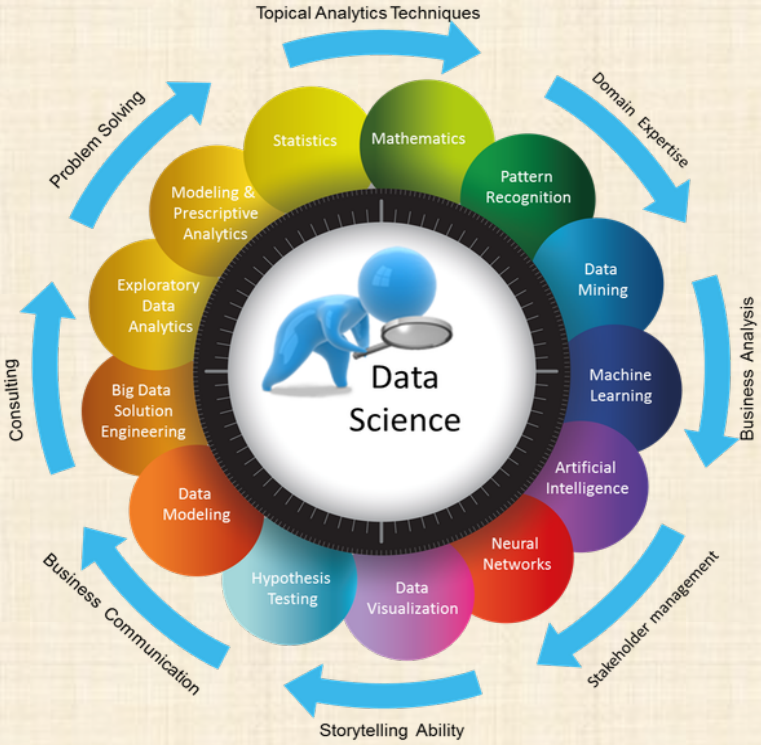

sas
THE POWER TO KNOW
Statistics
Pattern Recognition
Computational Neuroscience
Data Science
Data Mining
Machine Learning
AI
Databases
Information Retrieval


data
science
faculty
students
environment
scientists
support
research
bridge
education
activities
institute
sciences
fellows
program


Driving the Success of Data Science Solutions: Skills, Roles and Responsibilities ...
Ask good questions
Know the constraints (e.g., legal, ethics, market)
Latency at Execution?
Business Skills
Decision Making
Build, Buy, Outsource
Transparent Versus "black box"
Gauge political friction
Performance Criteria That Matter (ROI, accuracy, profitability versus market gain)
Deployment
"Analytics Leader"?
Feature Engineering
Data
Recalibration With New Data?
Data Logistics
IT Skills
Data Science
High-performance Computing
Which Analytics to Choose?
Project Execution/Monitoring
Data Exploration
Data Governance
Creativity, Communication
Gartner.


INSPIRES
ANALYTICS
BUSINESS INTELLIGENCE
DATA SCIENCE
STATISTICS
DATA VIZ
DATA
MACHINE LEARNING
EMPLOYS
ENABLES
DATA MINING
INSIGHT
DEEP NEURAL NETWORKS
FACILITATES
DEEP LEARNING
ARTIFICIAL INTELLIGENCE
BIG DATA
• Computer Vision
• Natural Language Processing
• Generative Models
PUSHES POST FURTHER





?

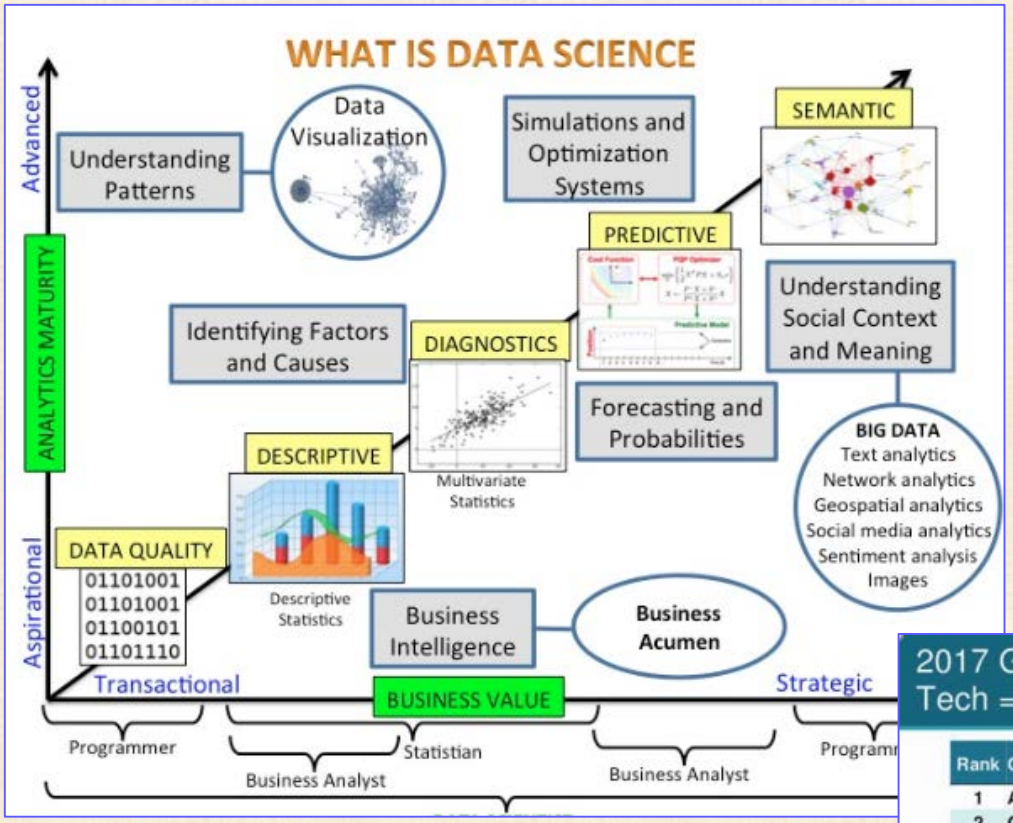
data science = statistique **réponse = O U I !!!**

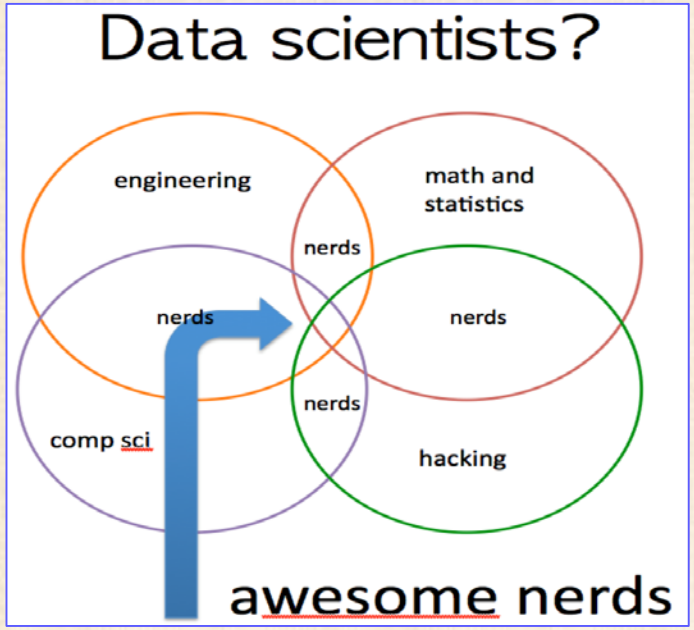
Statistique

science de d'apprentissage / compréhension d'un système avec des données en leur donnant du sens et en mesurant et contrôlant l'incertitude.

L'incertitude est mesurée en unités de probabilité (= monnaie de l'incertitude)

Data Science





STATISTICIEN

GAFAM

2017 Global Market Capitalization Leaderboard = Tech = 40% of Top 20 Companies…100% of Top 5…

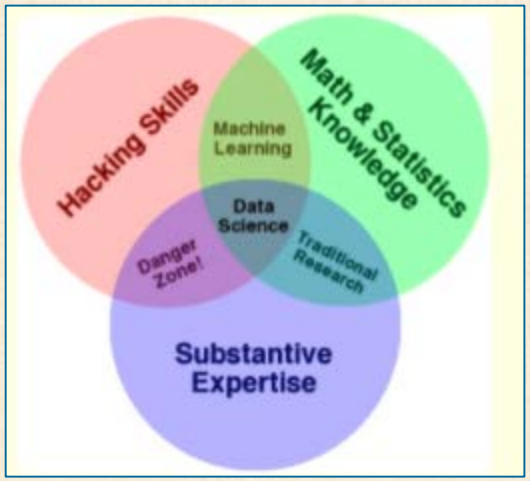
Rank	Company	Region	Industry Segment	Current Market Value ($B)	2016 Revenue ($B)
1	Apple	USA	Tech – Hardware	$801	$218
2	Google / Alphabet	USA	Tech – Internet	680	90
3	Microsoft	USA	Tech – Software	540	86
4	Amazon	USA	Tech – Internet	476	136
5	Facebook	USA	Tech – Internet	441	28
6	Berkshire Hathaway	USA	Financial Services	409	215
7	Exxon Mobil	USA	Energy	346	198
8	Johnson & Johnson	USA	Healthcare	342	72
9	Tencent	China	Tech – Internet	335	22
10	Alibaba	China	Tech – Internet	314	21
11	JP Morgan Chase	USA	Financial Services	303	90
12	ICBC	China	Financial Services	264	85
13	Nestlé	Switzerland	Food / Beverages	263	88
14	Wells Fargo	USA	Financial Services	262	85
15	Samsung Electronics	Korea	Tech – Hardware	259	168
16	General Electric	USA	Industrial	238	120
17	Wal-Mart	USA	Retail	237	486
18	AT&T	USA	Telecom	234	164
19	Roche	Switzerland	Healthcare	233	51
20	Bank of America	USA	Financial Services	231	80
Total				$7,207	$2,497

DIFFÉRENCES entre STATISTIQUE et DATA SCIENCE ?

DOMAINE	éléments
STATISTIQUE (classique)	idées, hypothèses, évaluation analyse : primaire , haut vers le bas confirmatoire données : à recueillir
DATA SCIENCE (data mining)	génération d'hypothèses, création idées analyse : secondaire, bas vers le haut exloratoire (après coup) données : historiques

idée ⬇ données

idée ⬆ données

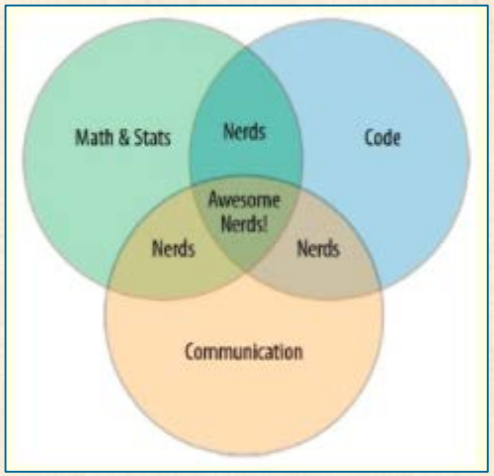


" Data Science is much older than Kepler .. It is the second oldest profession "

Gregory Piatetsky-Shapiro

" Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it "

Brad Efron



DATA SCIENCE : Tukey, Breiman, …

Un peu d'histoire …

1961 John Tukey called for a reformation of academic statistics.

`The Future of Data Analysis', he pointed to the existence of an as-yet unrecognized science, whose subject of interest was learning from data, or `data analysis'.

20 years ago, John Chambers, Bill Cleveland , Leo Breiman independently once again urged academic statistics to expand its boundaries

Chambers called for more emphasis on data preparation and presentation rather than statistical modeling;

Breiman (2001) called for emphasis on prediction rather than inference

Cleveland suggested the catchy name " Data Science "

Breiman

"... there are 2 cultures in the use of statistical modeling to reach conclusions form data and solve problems"

Culture ONE = data are generated by a given STOCHASTIC DATA MODEL. (statistique classique)

Culture TWO = uses ALGORITHMIC MODELS and treats the data mechanism as unknown.

The statistical community has been committed to culture ONE, kept statisticians from working on a large range of problems.

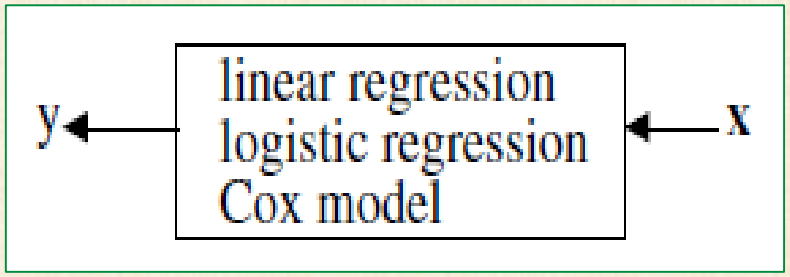
Algorithmic modeling has developed in fields outside statistics with LARGE complex data sets .... a more accurate and informative alternative to data modeling on SMALL data sets.

move away from exclusive dependence on data models and adopt a more diverse set of tools. "

Stochastic Data Modeling Culture

Statisticians : 98%

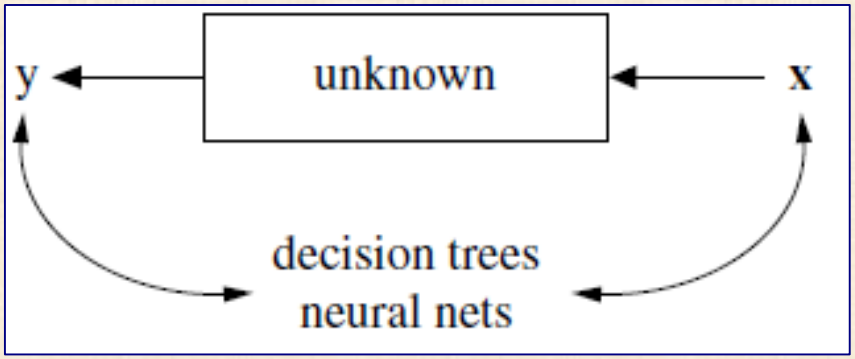
Computer Scientists : 2%



Algorithmic Modeling Culture

Statisticians : 2%

Computer Scientists : 98%



Leo Breiman (U Berkely) *Statistical Science* 2001, pp. 199–231

DIFFÉRENCES et TERMINOLOGIE

STATISTIQUE	INGÉNIERIE / INFORMATIQUE
stat / math	informatique / computer science
analyse statistique	machine learning (ML)
régression / classification	apprentissage supervisé
clustering / estimation densité	apprentissage non supervisé
modèles	réseaux, graphiques
tests / résidus	généralisation
paramètres	poids
variable input	features , classe
variable output / réponse	target, label, features
observation	instance, cas, exemple
méthodes	algorithmes
inférence = oui	inférence = non

subvention = 20 000 $ ..... subvention = 1 000 000 $

Mégadonnées (Big Data) — or noir du 21ième sciècle

- grande quantité de données qui ne peuvent pas être traitées par des outils / méthodes de gestion informatiques tradidionnels
- infrastucture / architecture matérielles et logiciels qui permet de manipuler et analyser les données massives avec des technologies informatiques appropriées

5 V : Volume - Variété – Vitesse – Véracité - Valeur

Volume : GRAND nombre d'observations (n) ou de variables (p)
combien ? pétabyte ….

Variété : données sous touts les formats, images, videos, audio …, web, banques de données gouvernementales, industrielles, disponibilité croissante …. OPEN DATA SCIENCE

Vélocité : données en mouvement, défilement continu,…

Véracité : fiabilité, qualité dans le temps, validité ..
…. très important …. problèmes de gouvernance

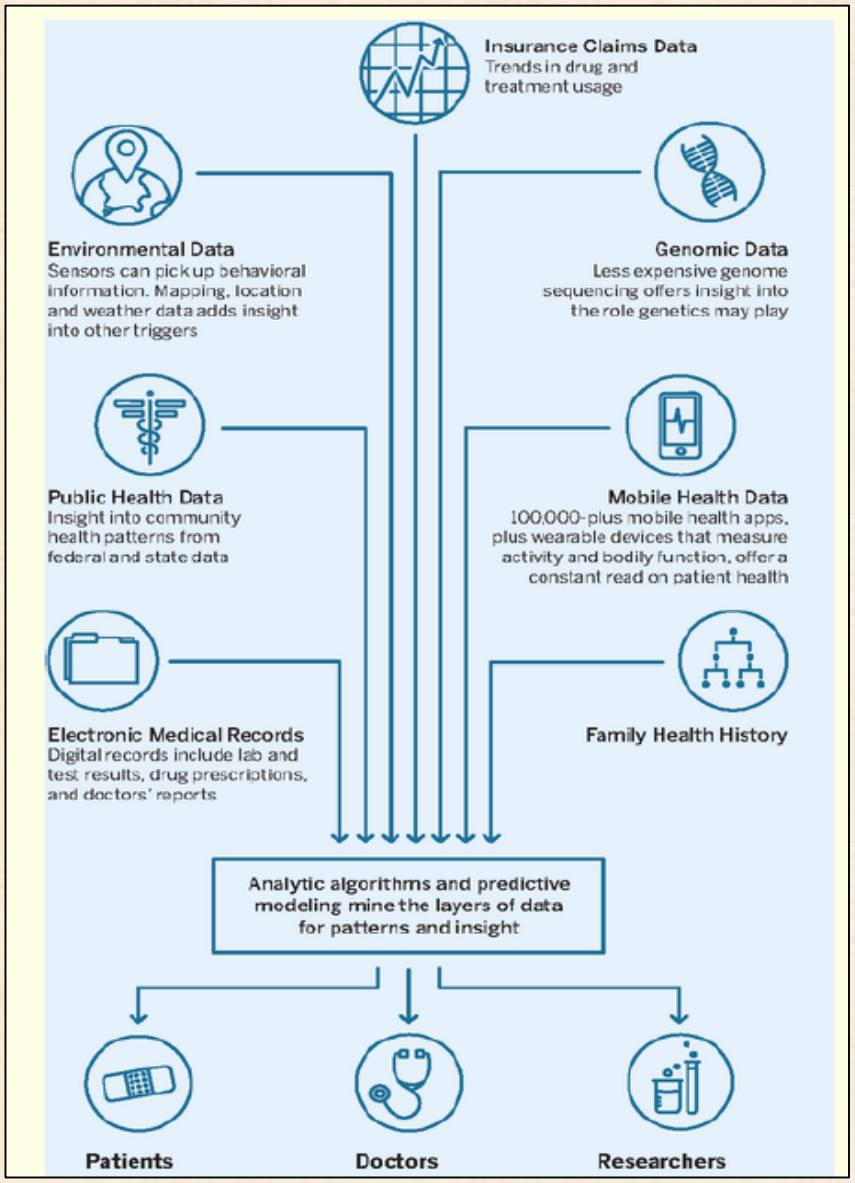
Valeur : création de valeur ajoutée …

APPLICATIONS : santé, génomes, senseurs, réseaux sociaux, gouvernements, transport, …

- très grande visibilité (exagérée ?, promesses ?) du Big Data et l'IA
- défis et opportunities statistiques : l'inférence classique est-elle inutile ?

4 M du big data : Make Me More Money !

Exemple du BIG DATA : écosystème santé



+

IOT (I nternet O T hings)

collecte des données personnelles

via des

- **appareils connectés sur la personne**
- **des senseurs stationaires**

fournit de l'information sur les activités

des consommateurs et patients

Données

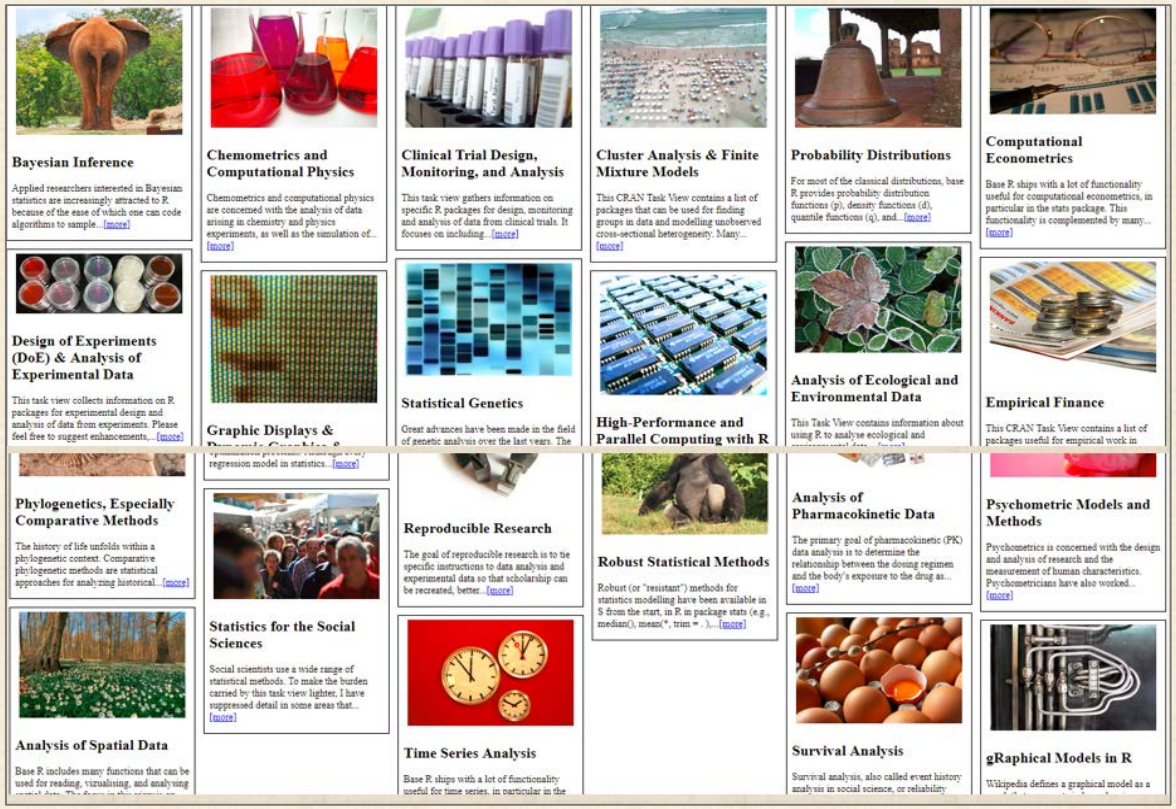
- **environnementales**
- **dossiers médicaux**
- **santé publique**
- **génomiques**
- **familiales**
- **chercheurs**
- **patients**
- **médecins**
- **hôpitaux**

Statistique : définition - applications

Statistique

collecte, analyse, interprétation, présentation, visualisation de données numériques et autres pour augmenter la connaissance en vue de …

But prendre des décisions basées sur des **données existantes** (historiques) ou **données à recueillir** dans des conditions **d'incertitude** et de **variabilité**



Applications

TOUS les domaines de l'activité humaine

science interdisciplinaire

" Statistical thinking will one day as necessary for efficient citizenship as the ability to read and write. "

H.G. Wells (1866-1946)

PROCESSUS / SYSTÈME) : S I P O C

Suppliers **Input : X** **Processus** **Output : Y** **Clients**

FACTEURS	GÉNÉRAL	RÉPONSES
Personnel →	mélange d'inputs qui génèrent un output	indicateurs en relation avec
Matériaux →		
Équipement →		→ qualité produit
Politiques →		→ qualité service
Procédures →		
Méthodes →		→ exécution tâche
Environnement →		

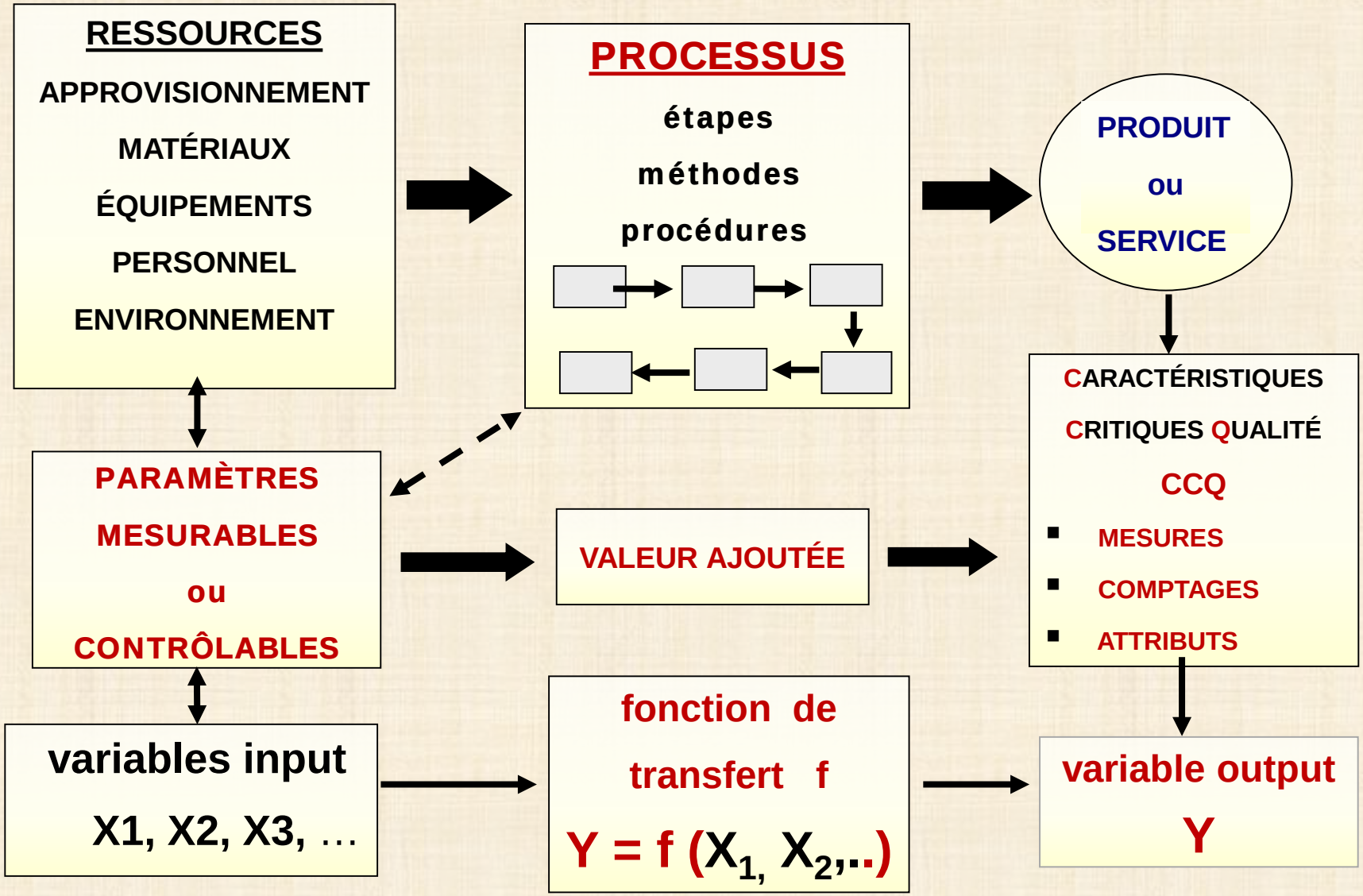
PROCESSUS / SYSTÈME

- DESIGN (CONCEPTION)
- FABRICATION
- MESURAGE
- TRANSACTIONEL
- ADMINISTRATIF

exemples en intelligence artificielle

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

PROCESSUS / SYSTÈME



Type d'études statistiques

actif — rôle statisticien — **passif**

Expériences planifiées

traitements appliqués aux unités expérimentales selon un protocole (design)

structure traitements

design expérimental : randomisation, blocage, répétitions

biostatistique, pharmaceutique, sciences physiques, sciences exactes, expériences avec sujets humains / animaux .....

Sondages, enquêtes, recensements

=

études énumératives

plan d'échantillonnage des unités statistiques

pas de traitements appliqués aux unités

sciences humaines, sciences sociales, ...

Études observationnelles

données collectées au fil du temps / temps réel

unité statistiques

=

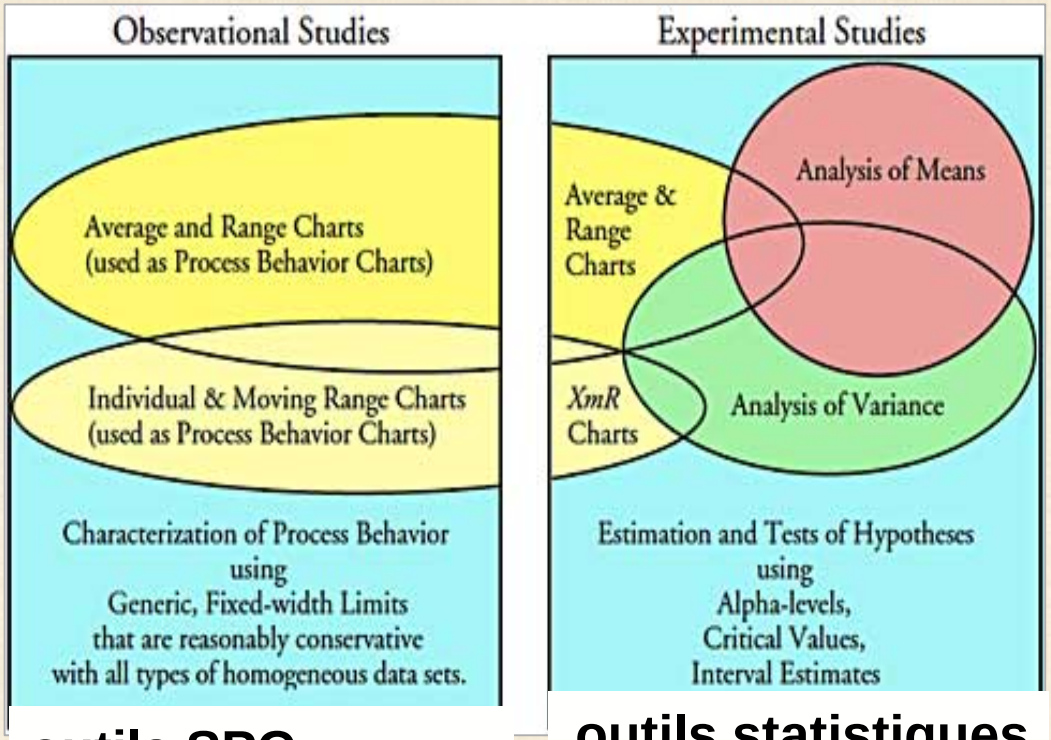
instants d'observations

peu / pas de planification statistique

banques de données, mégadonnées (big data)

Type d'études statistiques

observationnelle **expérimentale** **énumérative** **analytique**



outils SPC
cartes comportement processus

outils statistiques
tests, ANOVA, régression, etc.

Two Types of Studies

Enumerative Studies:	Analytic Studies:
Census	Selection of suppliers
Inventory	Poll to determine strategy
Exit survey at polls	Experiment to improve performance
Acceptance sampling	Drug testing

W. E. Deming's Two Types of Studies

Chapter 7 from *Some Theory of Sampling*, 1950

The aim of any experiment is to provide a rational basis for action

Enumerative study: an experiment in which action will be taken on the universe.

Analytic study: an experiment in which action will be taken on a cause system to improve performance in the future.

nouvelle distinction

CLASSIQUE (traditionnelle) : base inférentielle

MÉGADONNÉES (nouvelle) : base algorithmique

Planification étude statistique

Identification des VARIABLES

Nature: continue - catégorique

Rôle: explicatives (X = input) - à expliquer (Y = output = réponse)

Liste des X complète? p = nombre OK?

Mesure de Y - processus de mesure / erreur? justesse?

STRUCTURE et le PLAN de collecte des données

expérience planifiée - quel plan statistique ?

- combien de données ? n ?

données observées sans plan expérimental – qualité ?

Terme d'erreur expérimentale - distribution normale? Importance ?

préoccupation obsessive de la normalité !

Forme de f - connue – linéaire / non linéaire (cas plutôt rare)

- inconnue - quelle approximation ? – polynomiale ?

- techniques de sélection des variables pour modéliser

- qualité du modèle ajusté ? Critères ?

Ajustement du modèle - analyse de sensibilité des X

Évaluation de qualité du modèle - analyse des résidus

ÉTAPES ÉTUDE STATISTIQUE CLASSIQUE

1.	Identification	processus / problème / variables
2.	Observation	plan collecte des données
3.	Spécification	modèle pour analyse
4.	Estimation	paramètres du modèle
5.	Décomposition	variabilité (ANOVA), test F
6.	Validation	tests, ratio-F, analyse résidus
7.	Exploitation	optimisation / résolution problème décision / action

ÉTAPES

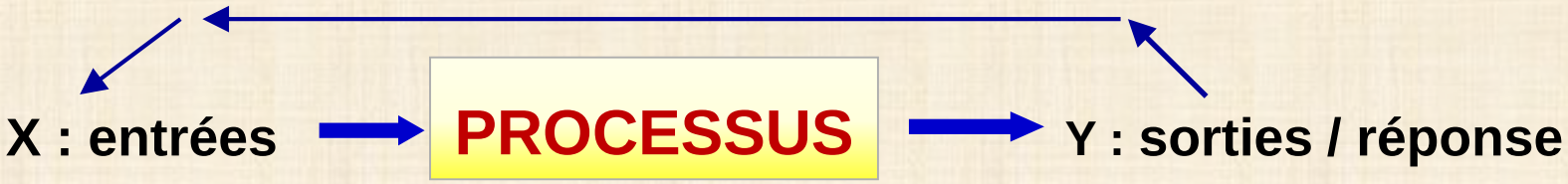
ANALYSE

STATISTIQUE

CLASSIQUE

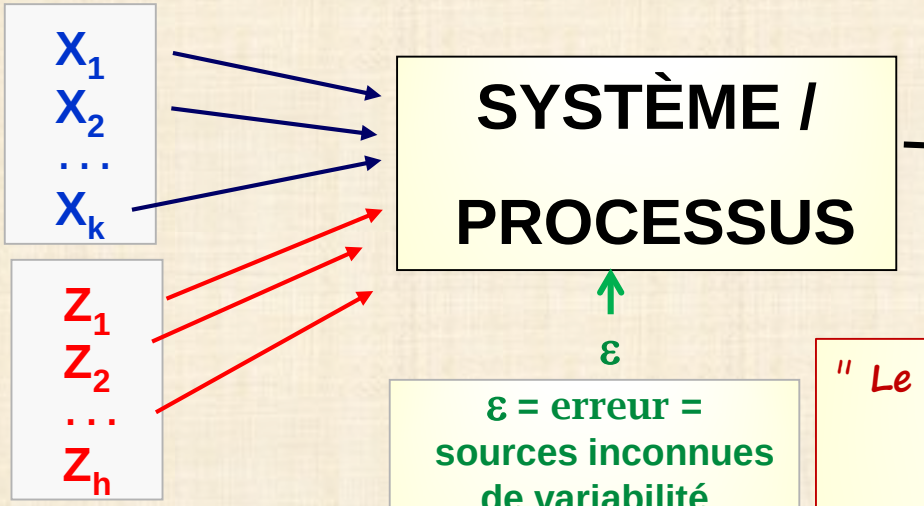
1. Spécification d'un modèle statistique
2. Estimation des paramètres du modèle
3. Décomposition de la variabilité : ANOVA
4. Tests d'hypothèses sur les paramètres
5. Analyse diagnostique des résidus
   - vérification des hypothèses de base
   - identification d'observations influentes
   - transformation Box-Cox de réponse Y
6. Si nécessaire : itération des étapes 1 à 5
7. Optimisation de la réponse (s'il y a lieu)
8. Graphiques de la réponse

Étude des relations entrées-sorties



COMPARAISON	Modèle de prédiction	Modèle d'analyse de variance
But	développement d'un modèle prédictif de la réponse	identification des effets significatifs sur la réponse
Source des données	historiques / observationnelles	résultat d'un plan d'expérimentation
Nombre d'observations	grand: centaines, milliers…	petit : dizaines
Variables d'entrée	continues / quantitatives	catégoriques / qualitatives
Nombre de valeurs distinctes des variables d'entrée	autant qu'il y a d'observations	nombre restreint généralement moins de 10
Utilisation des variables indicatrices (0-1)	occasionnelle	employées systématiquement pour représenter les modalités
Emphase et difficulté	forme et la qualité du modèle	spécification du modèle reflétant la complexité du plan expérimental
Structure des données	simple	complexe

ANALYSE STATISTIQUE CLASSIQUE : comprendre / prédire / optimiser



" Le doute n'est pas une condition agréable, mais la certitude est absurde."

Lettre à Frédéric II, roi de Prusse 6 avril 1767
de François Marie Arouet dit Voltaire

Aucune restriction concernant la nature des X et Y

X: catégorique, entière, continue, contrôlées, aléatoires
Y: binaire(0, 1), multinomiale, entière, continue

Algorithmes du Machine Learning

linéaire, linéaire généralisé, arbres, réseaux neurones, PLS, etc. ..

p = nombre de variables n = nombre d'observations

on peut avoir plus de variables que d'observations !

VARIABLES et MODÈLE

VARIABLES

RÔLE

Y : réponse , output, à expliquer
peut être: binaire (0, 1), multinomiale, continue, multidimensionnelle

X, Z : explicatives, régresseurs, input
inter / intra relativement aux unités expérimentales

NATURE

X (fixées) : continues, catégoriques (facteurs)
Z (aléatoires) : continues, catégoriques

INFLUENCE

X : affecte la centralité (moyenne) de Y : effets fixes
Z : affecte la dispersion (variance) de Y : effets aléatoires

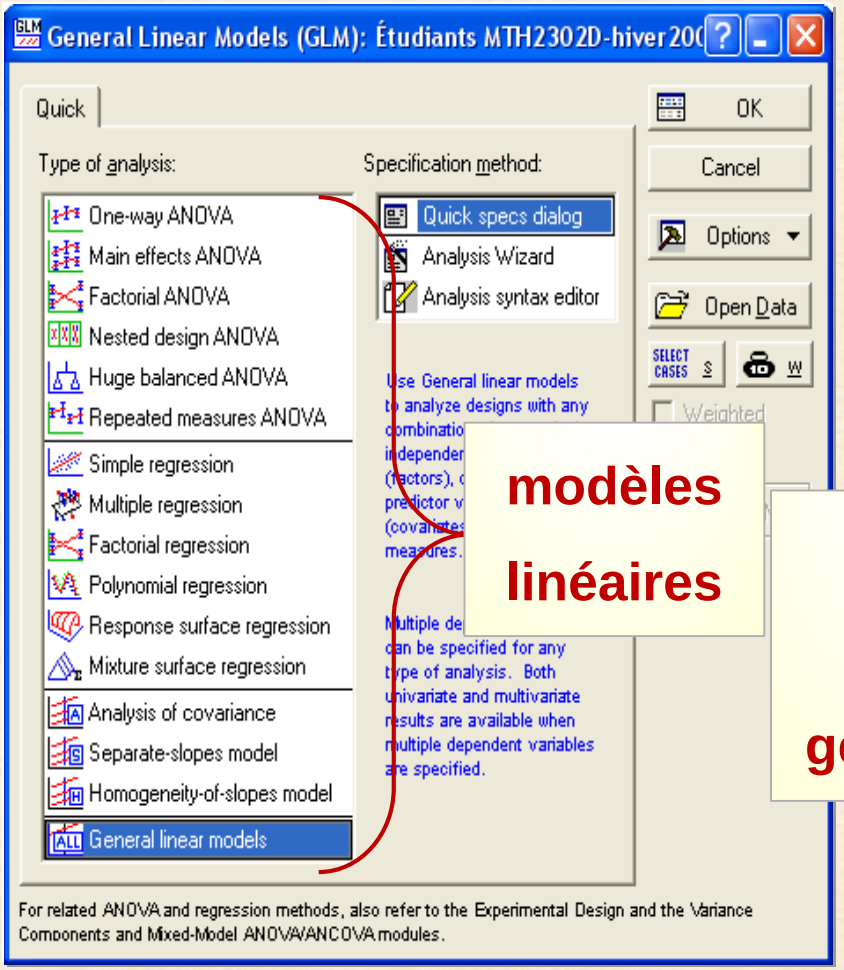
MODÈLES effets fixés | effets aléatoires | mixtes (fixés , aléatoires)

$$Y = f(X_1, X_2, \ldots, X_k ; \beta_0, \beta_1, \beta_2, \ldots) + g(Z_1, Z_2, .., Z_h ; \sigma_1^2, \sigma_2^2, \ldots) + \varepsilon(0, \sigma^2)$$

Statistica

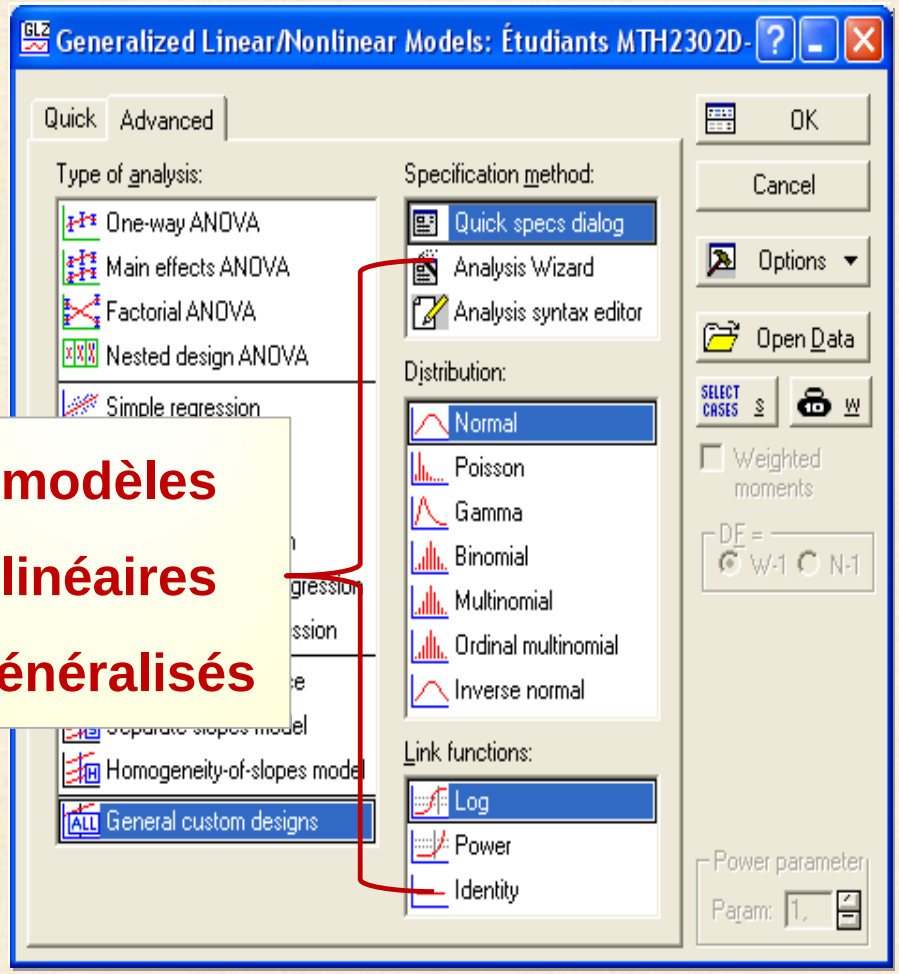
modèles statistiques classiques

GLM : General Linear Model



modèles linéaires

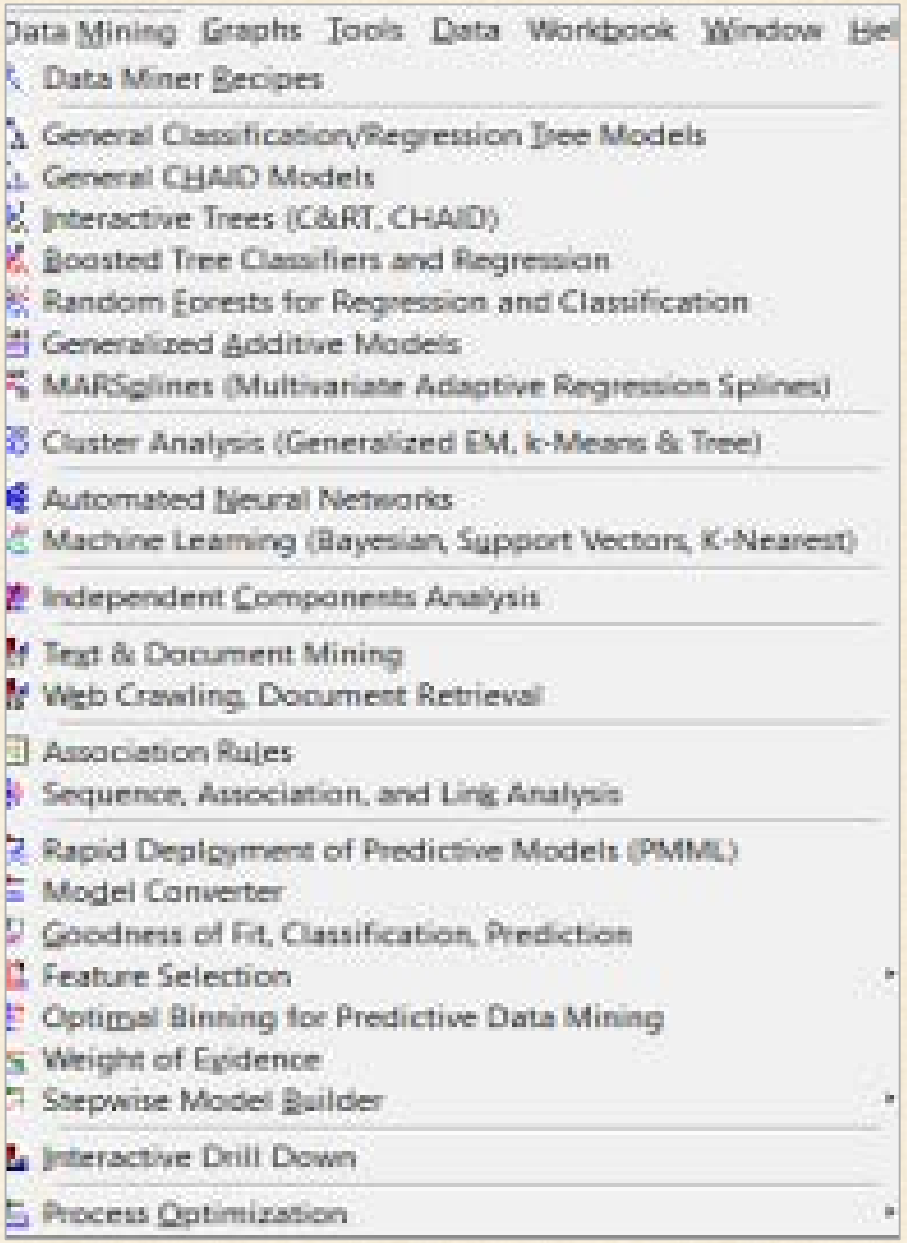
GLZ : Generalized Linear/Nonlinear Model



modèles linéaires généralisés

Statistica

Algorithmes du ML



Class Regression Trees (CRT)

Bosting (boostrap)

Ensembles

Random Forests

GAM

MARSplines

Clustering

Bayesian Networks

Automated Neural Networks

Support Vector Machine (SVM)

Text Mining

Web Crawling

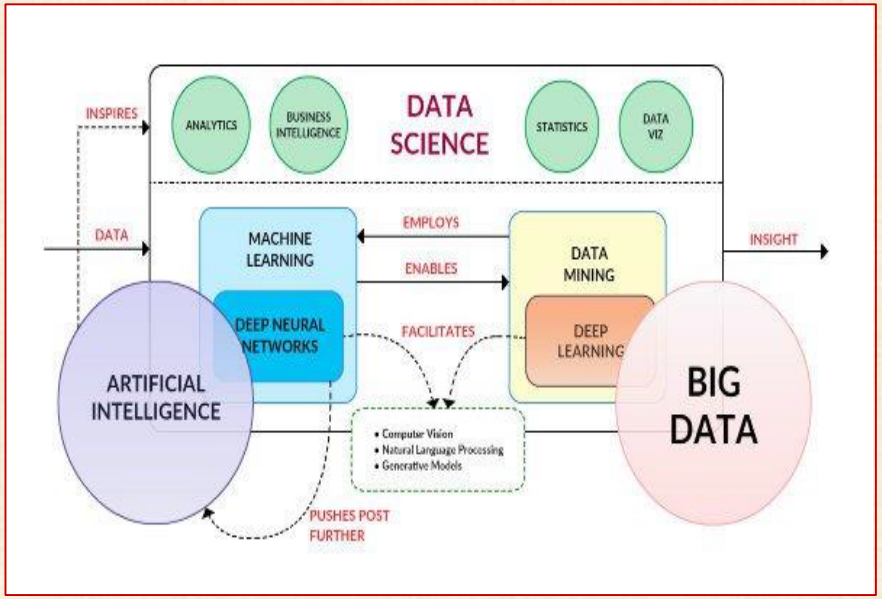
Rapid Deployment (PMML)

Optimal Binning

Stepwise Model Builder

Process Optimization

BIG DATA



Unités d'octets v · d · m

Ordre de grandeur	Système international (SI)			Préfixes binaires		
	Unité	Notation	Valeur	Unité	Notation	Valeur
1	**octet**	o	1 octet	**octet**	o	1 octet
10^3	kilooctet	ko	10^3 octets	kibioctet	Kio	2^{10} octets
10^6	mégaoctet	Mo	10^6 octets	mébioctet	Mio	2^{20} octets
10^9	gigaoctet	Go	10^9 octets	gibioctet	Gio	2^{30} octets
10^{12}	téraoctet	To	10^{12} octets			2^{40} octets
10^{15}	pétaoctet	Po				2^{50} octets
10^{18}	exaoctet	Eo	10^{18} octets	exbioctet	Eio	2^{60} octets
10^{21}	zettaoctet	Zo	10^{21} octets	zébioctet	Zio	2^{70} octets
10^{24}	yottaoctet	Yo	10^{24} octets	yobioctet	Yio	2^{80} octets

big data

DÉMARCHE Machine Learning / Data Mining

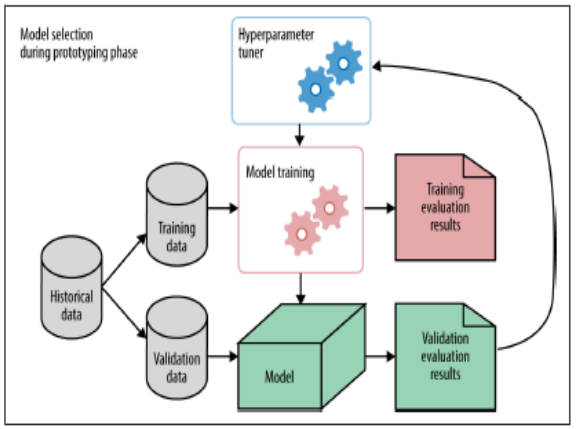


Figure 3-1. The prototyping phase of building a machine learning model

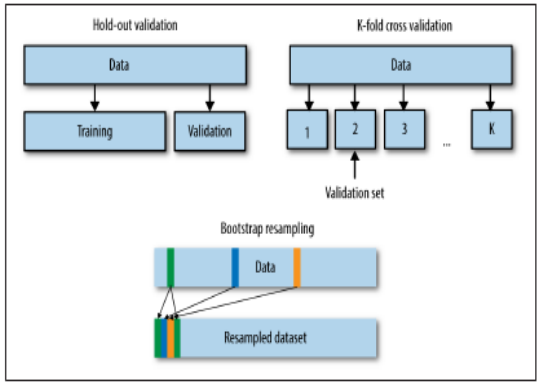


Figure 3-2. Hold-out validation, k-fold cross-validation, and bootstrap resampling

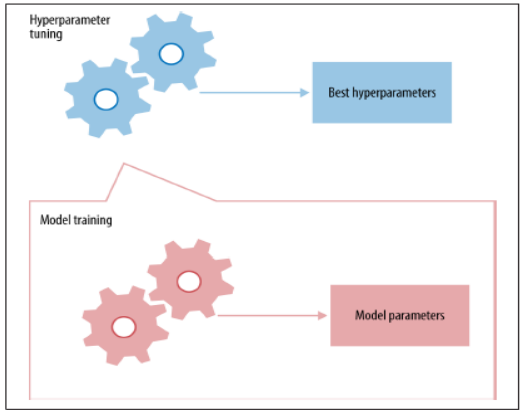


Figure 4-1. The relationship between hyperparameter tuning and model training

Statistique traditionnelle (classique)

expert du domaine apprend les bonnes variables ("features“) et apporte au statisticien un petit tableau de données nettoyées

étapes : specification modèle + analyse des résidus + critères d’évaluation

Approche Data Mining / Machine Learning

départ avec un gros ensemble de données
(beaucoup de variables / d’observations) (pourrait être un big data)
lance plusieurs algorithmes pour identifier plusieurs « bons » modèles
data séparé en 2 / 3 parties : entrainement + test + prédiction

étapes : plusieurs algorithmes
+
critères d’évaluation : courbe ROC, AUC, Lift chart, bootstrap …

EXAMPLES

Click-through rate

Based on the search term, knowledge of this user (IPAddress), and the Webpage about to be served, what is the probability that each of the 30 candidate ads in an ad campaign would be clicked if placed in the right-hand panel.

Logistic regression with billions of training observations.
Each ad exchange does this, then bids on their top candidates, and if they win, serve the ad within 10ms!

Adverse drug interactions

US FDA (Food and Drug Administration) requires physicians to send in adverse drug reports, along with other patient information, including disease status and outcomes. Massive and messy data.

found drug interactions associated with good and bad outcomes.

Social networks

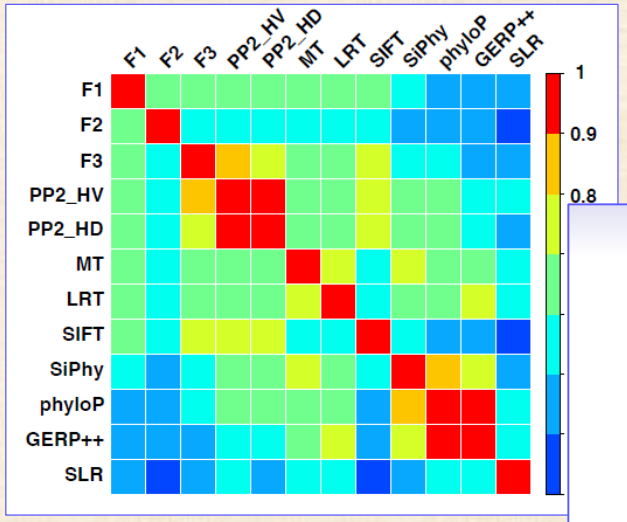
Based on who my friends are on Facebook or LinkedIn, make recommendations for who else I should invite. Predict which ads to show me.

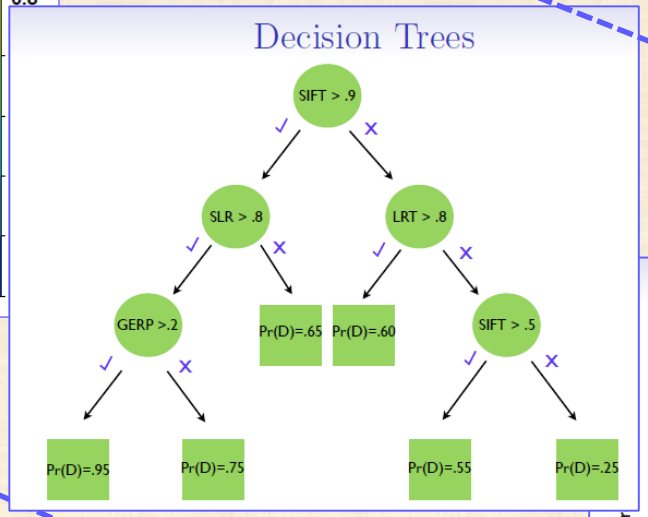
more than a billion Facebook members, and two orders of magnitude more connections.
Knowledge about friends informs our knowledge about you.

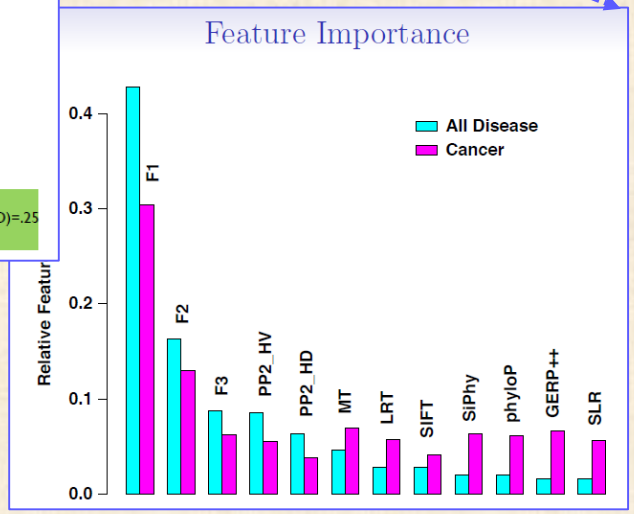
CANCER PROSTATE

- 12 scores existants (variables de réponse) pour la maladie ne sont pas toujours en accord
- Données : 52 000 observations : 31 00 sujets sains + 21 000 sujets malades
- Algorithme Forets Aléatoires intégrer les scores dans un critère de decision de maladie

matrice de corrélations



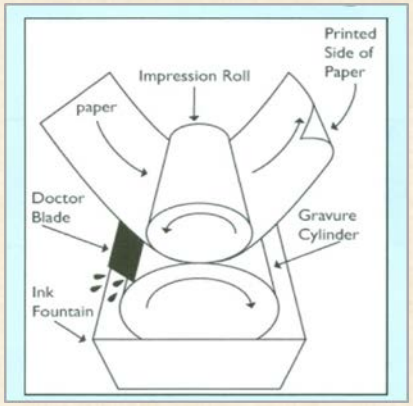




IMPRIMERIE

combinaison des 2 approaches : data mining + planification d'expérience (DOE)

Example printing industry - process description
Problem defect on cylinder = « banding » image of poor quality occurring 40% of time
Action stop production – repair cylinder
many hours - cost 10 000$ per incident
Project what conditions lead to banding?
Data 540 observations x 39 variables Y = band occured (red)
X : 11 categorical var (blue) + 17 continuous var (green)

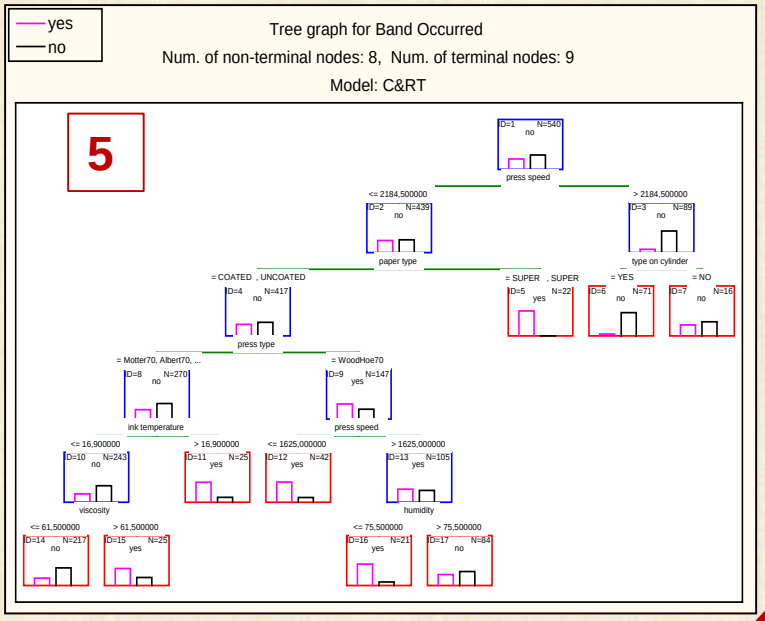


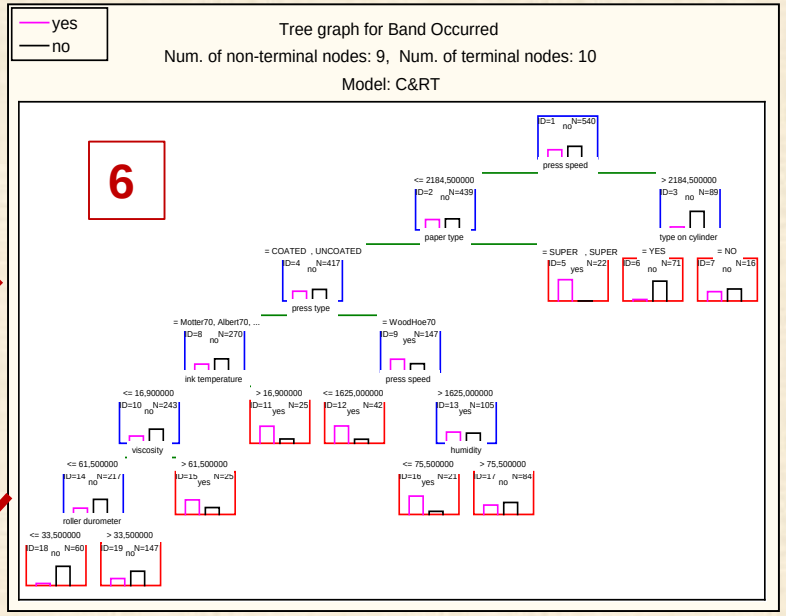
data

	Year	Month	Day	Job Number	Cylinder No	Customer	Y band occurred	grain screened	proof on ctd ink	paper type	ink type	solvent type	type on cylinder	press type	press	cylinder size	location	plating tank
1	1990	03	30	23040	X750	GUIDE POSTS	yes	YES	YES	UNCOATED	UNCOATED	LINE	NO	Motter94	821	TABLOID		
2	1990	04	09	34683	G467	ECKERD	no	NO	YES	COATED	COATED	LINE	YES	Wood Hoe70	815	TABLOID	North US	1910

	proof cut	viscosity	caliper	ink temperature	humidity	roughness	blade pressure	varnish pct	press speed	ink pct	solvent pct	wax	hardener	roller durometer	current density	anode space ratio	chrome content
1	50,0	59	0,33	14,5	71	0,63	20	11,1	1650	55,5	33,3	2,5		40			100
2	40,0	38	0,30	16,0	92	0,63	25	1,1	1900	53,8	45,2	2,5		30	33	100,0	100

Phase 1 : CRT utilisé pour identifier 8 variables input critiques
Phase 2 : Planification une expérience pour modéliser et optimiser





8 critical controlable input variables (X) identified

1. press speed
2. paper type
3. type on cylinder
4. press type
5. ink temperature
6. grain screened
7. viscosity
8. humidity

Follow up study DOE

Design Of Experiment
8 input X variables
24 runs - new data on Y
Model Y with X then
optimal values of X found
such that Y = yes (banding)
occurs 4,5% now
vs. 40% before

MÉGADONNÉES : défis associés à des problèmes statistiques

- **Ethique :** utilisation et liaison de données personnelles, assurer la confidentialité
  voir General Data Protection Regulation (GDPR) de UE (mai 2018)
- **Mégadonnées :** seulement un échantillon à un temps particulier d'intérêt dans un processus – pas un échantillon aléatoire – pas toutes les données !
- **Qualité des données :** erreurs, omission, biais, censures, données manquantes, duplication, erreus de mesure. observations atypiques, **VARIABLES MANQUANTES**, **hétérogénéité des données !** **pourquoi pas le SPC ?**
- **Visualisation des données :** développement de nouvelles procedures graphiques
- **Fausses associations :** accroissement associations fausses avec l'accroisement de la taille de l'échantillon - habileté à trouver des varies relations causales
  correlation n'est pas cause !
- **Development de méthodes statistiques valides** pour résoudre le problème de tester un grand nombre d'hypothèses simultanées
- **Development de stratégies statististiques valides** pour la réduction du fléau de la dimensionalité / selection de variables
  curse of dimensionality ! data sparse out as the dimensionality increases
- **Réplication des résultats :** est-ce que des expériences indépendantes visant les mêmes questions produiront des résultats cohérents ?
- **Reproducibileté du résultat :** habileté à recalculer les résultats avec les mêmes données
- **Généralisation :** interprétation d'une étude exploratoire comme une étude prédictive

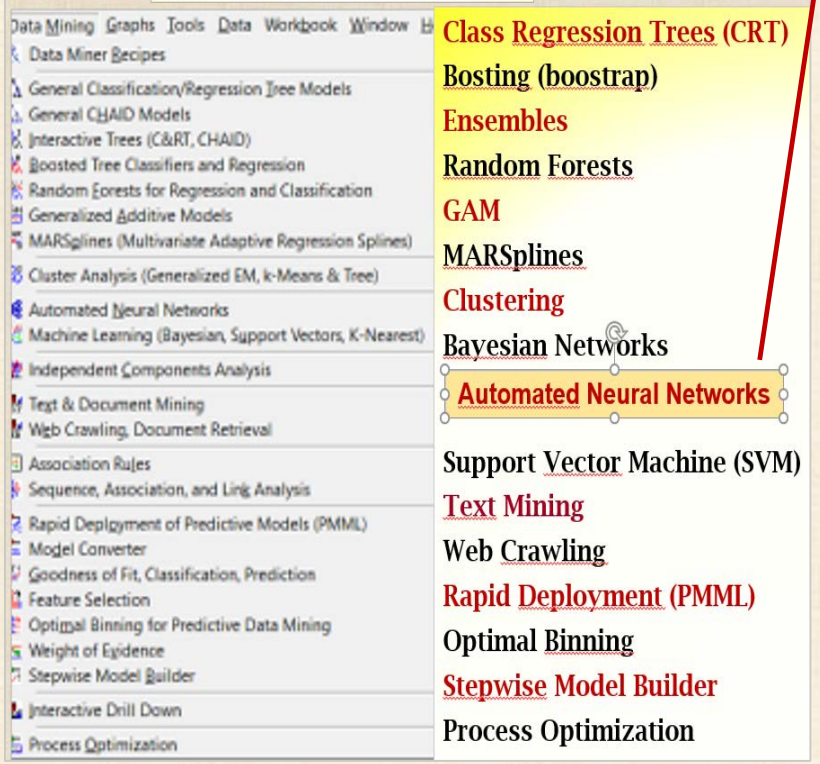
Conclusions et défis

- Les décisions basées sur l'intuition sont chose du passé; elles doivent reposées sur les données.
- Les mégadonnées sont là pour rester; elles impacteront tous les domaines de l'activité humaine ; l'âge d'or de la statistique est maintenant.
- Les principes statistiques et la rigueur sont nécessaires pour justifier le saut entre les données et la connaissance scientifique.
- Le manque d'expertise en statistique peut conduire à des erreurs fondamentales.
- Une grande quantité de données combinées à des analyses sophistiquées ne garantit pas le succès.
- Les données historiques n'assurent pas la performance future.
- Les éléments clés pour l'exploitation réussie des mégadonnées et le développement de la science sont la rigueur et les principes statistiques.
- La statistique, les mégadonnées, les algorithmes de la science données sont des aides pour penser et non pas des remplacements.
- L'utilisation des connaissances du domaine d'application est absolument nécessaire pour aider à définir le problème, appuyer le plan de collecte des données et guider l'analyse des résultats et leur interprétation.

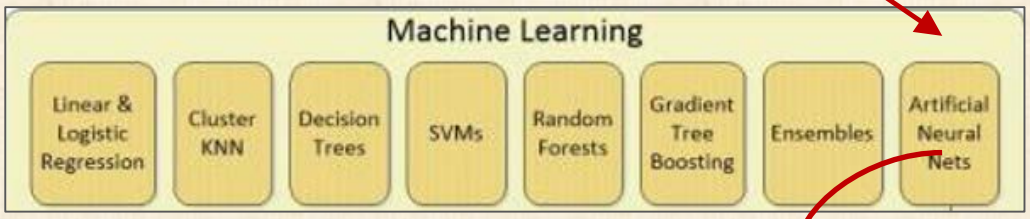
Rôle pour la statistique en IA ?

Quelques éléments sur les réseaux de neurones ...

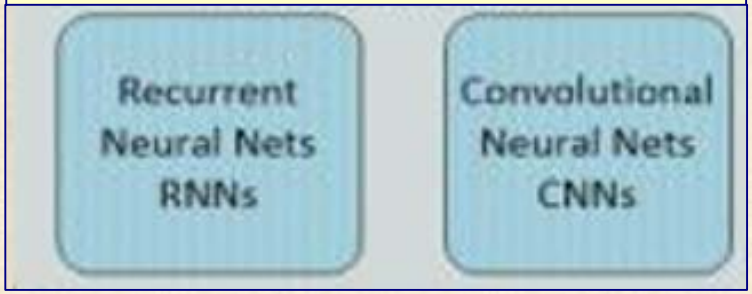
Algorithmes du ML



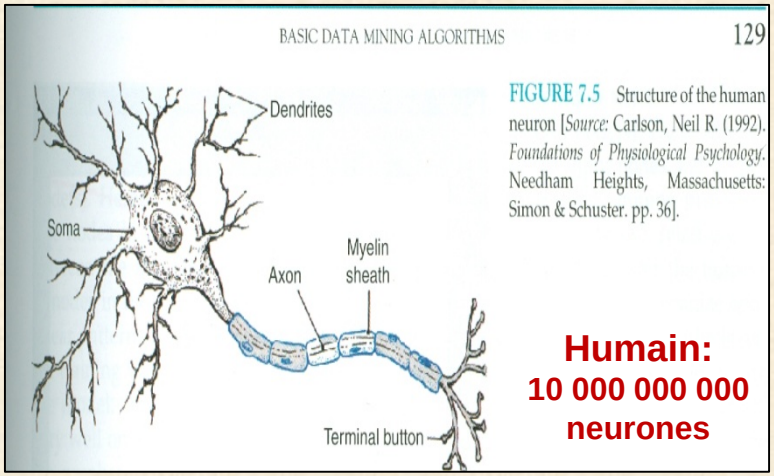
Automated Neural Networks



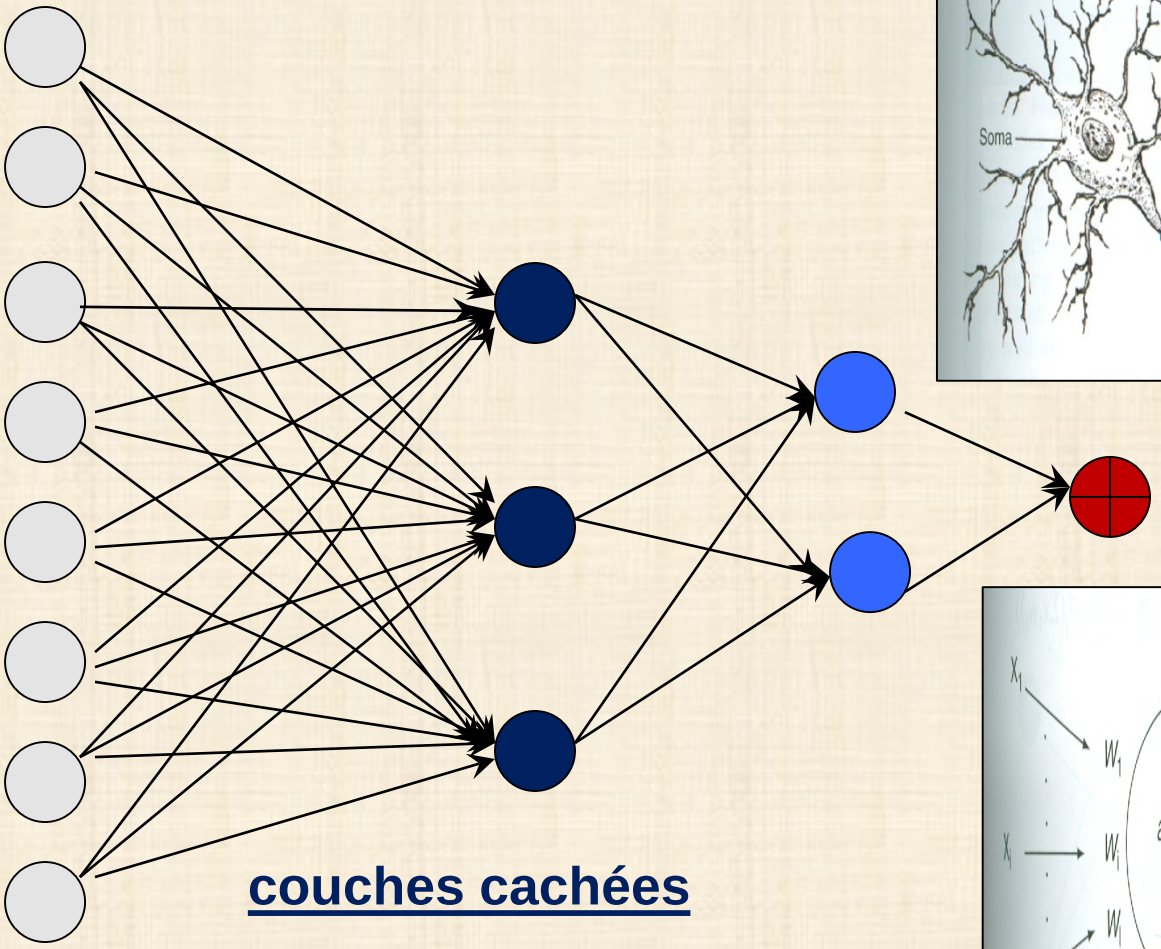
apprentisssage profond : deep learning



Base des Réseaux Neurones



Humain:
10 000 000 000 neurones



Y variable cible

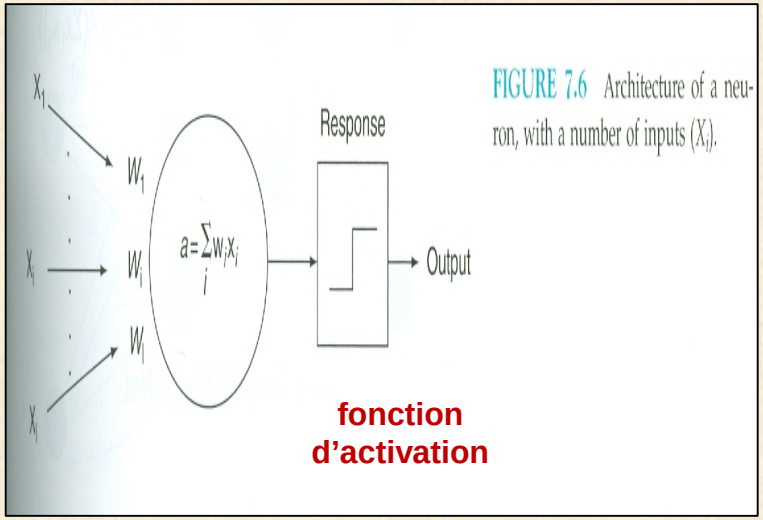
couches cachées

nombre de neurones = ?

nombre de couches cachées ?

n = 1 cas usuel

variables X prédicteurs



fonction d'activation

source: Nisbet et all p.129

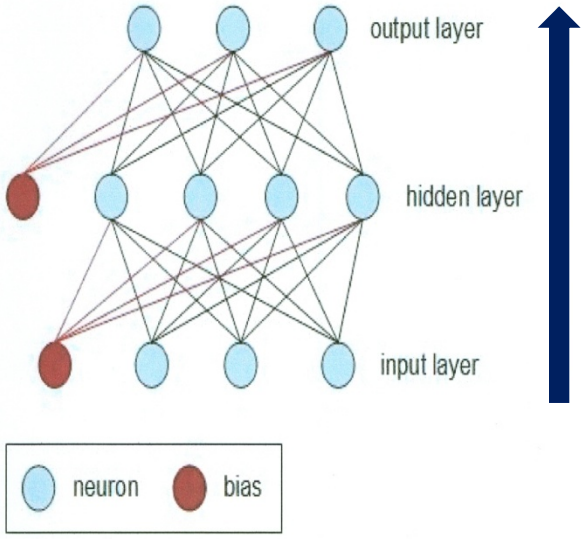
Types de réseaux de neurones

MLP Multi Layer Perceptron

SANN Overviews - Network Types

SANN Overviews - Network Types

The Multilayer Perceptron Neural Networks



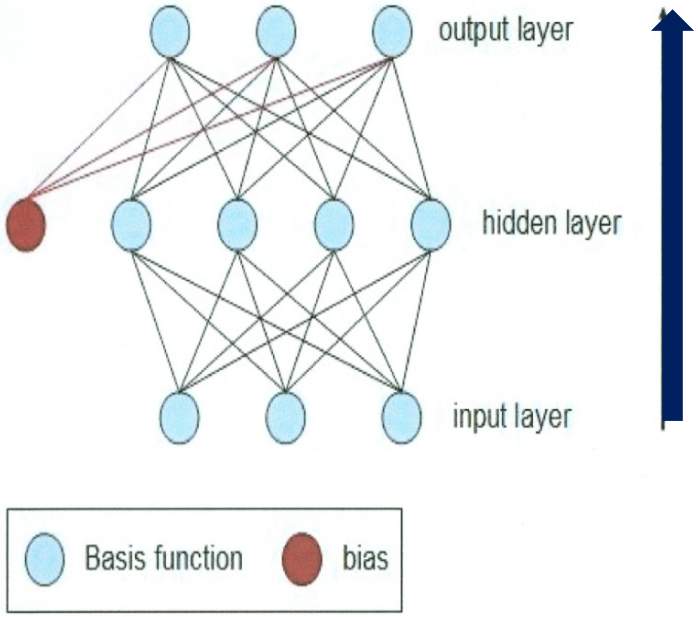
A schematic diagram of a fully connected MLP2 neural network with three inputs, four hidden units (neurons), and three outputs. Note that the hidden and output layers have a bias term. Bias is a neuron that emits a signal with strength 1.

Source : Statistica

type le plus employé

RBF Radial Basis Function

The Radial Basis Function Neural Networks

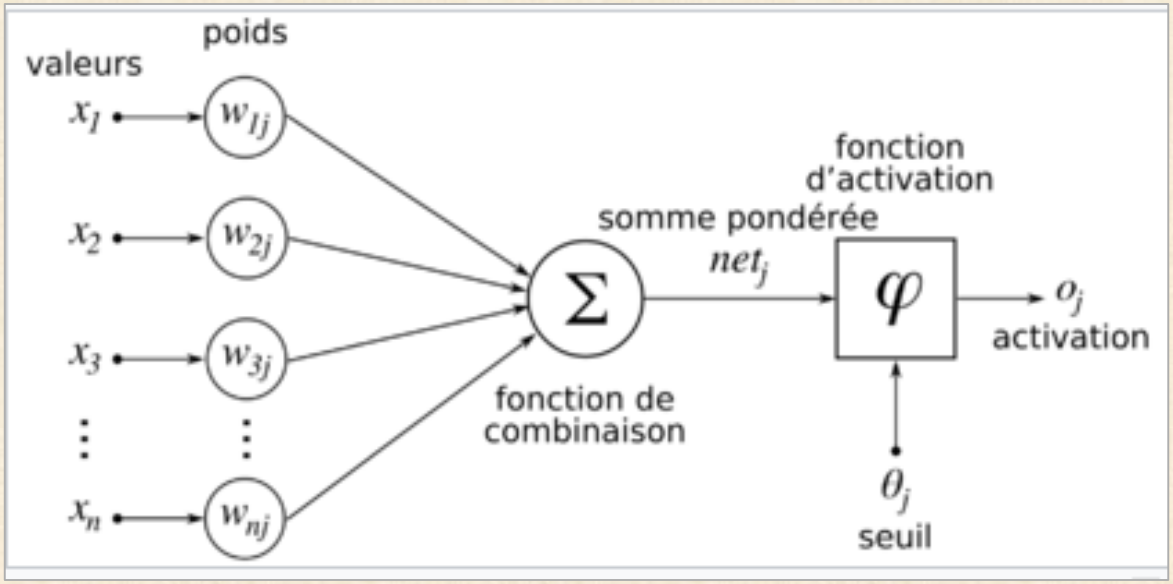


A schematic diagram of an RBF neural network with three inputs, four radial basis functions and 3 outputs. Note that, in contrast to MLP networks, it is only the output units that have a bias term.

type pas très recommandé pour des inputs catégoriques

Élements des réseaux de neurones

Perceptron Multi Couches



- Les solutions ne sont pas calculées directement : un algorithme d'optimisation calcule itérativement les solutions
- Le modèle est construit sur une base d'apprentissage et testé à chaque étape sur une base de validation
- L'apprentissage du réseau s 'effectue en mode supervisé : les poids W_i et les valeurs V_i synaptiques des neurones de chaque couche (cachée) sont évalués pour minimiser une fonction d'erreurs
- Les entrées (variables explicatives) et sorties (variables à expliquer) du réseau sont connues

Éléments des réseaux de neurones

Interprétation modèles et traitements

- valeur Y_j d'un individu j est une function
- composée des valeurs prises par X_i

- fonctions d'activation ……

fonctions d'activation

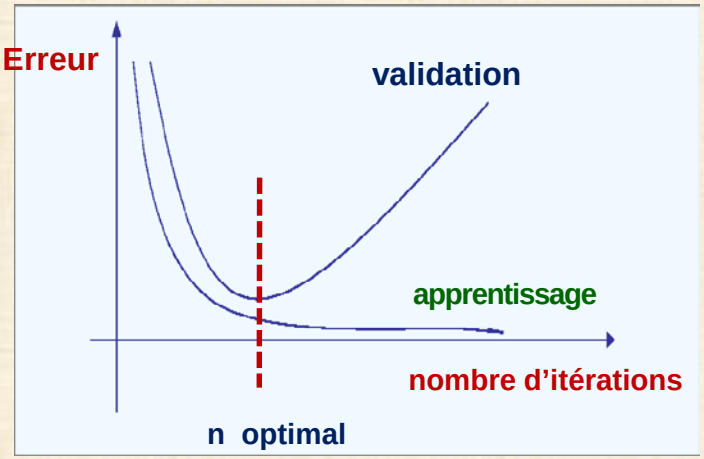
- poids W associés aux neurones

- Évaluation complexe + présence de colinéarités

variables, poids, nombre de couches, nombre de neurons

éviter le « sur apprentissage »

Function	Definition
Identity	a
Logistic sigmoid	$\frac{1}{1+e^{-a}}$
Hyperbolic tangent	$\frac{e^{a}-e^{-a}}{e^{a}+e^{-a}}$
Exponential	e^{-a}
Sine	$\sin(a)$
Softmax	$\frac{\exp(a_i)}{\sum \exp(a_i)}$
Gaussian	$\frac{1}{\sqrt{2\pi\sigma}}\exp\left[\frac{(x-\mu)^2}{2\sigma^2}\right]$



y_i : prédictions t_i : target

Erreur : Sum Of Square
$= E_{SOS} = \sum (y_i - t_i)^2$

t_i continue

Erreur : Cross Entropy
$= E_{CE} = -\sum t_i \ln(y_i / t_i)$

t_i catégorique

Algorithmes

- BFGS Broyden-Fletcher-Goldfarb-Shanno
- Scaled Conjugate Gradient

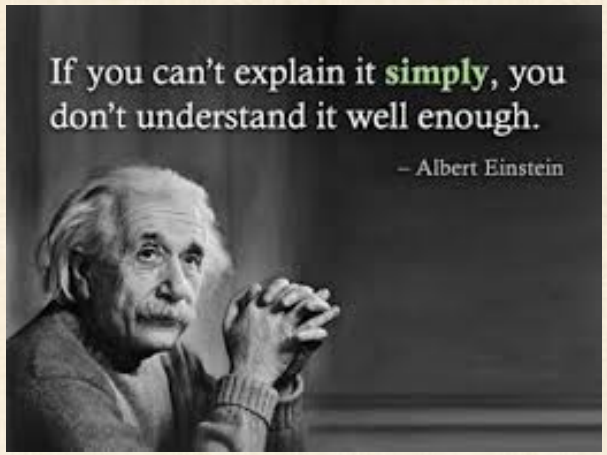
Commentaires sur les réseaux de neurones

- performance bonne à très bonne en général
- boite noire : difficile à interpréter
- analyse de sensibilité possible mais limitée
- utiles domaine des données non structurées

INPUT A	RESPONSE B	APPLICATION
Picture	Are there human faces? (0 or 1)	Photo tagging
Loan application	Will they repay the loan? (0 or 1)	Loan approvals
Ad plus user information	Will user click on ad? (0 or 1)	Targeted online ads
Audio clip	Transcript of audio clip	Speech recognition
English sentence	French sentence	Language translation
Sensors from hard disk, plane engine, etc.	Is it about to fail?	Preventive maintenance
Car camera and other sensors	Position of other cars	Self-driving cars

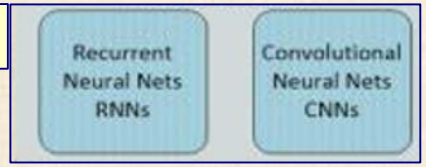
sortie en C / C++ interprétation ?

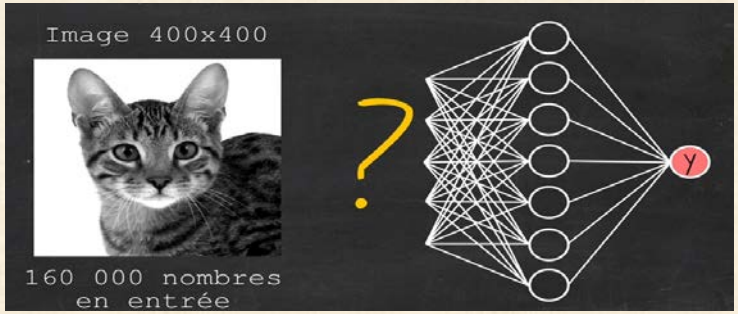
```
double 4bar_linkage2_sta_in_RéseauxNeurones_1_MLP_5_10_4_input_hidden_weights[10][5]=
{
 {2.02837435462445e-003, -9.19300736782718e-002, -8.33394542388945e-003, 5.96815120368463e-003, -9.72173081422972e-001 },
 {-4.10541254965628e+001, -1.57048702434030e+001, 7.24298697308776e+000, 7.54136136116508e-001, 7.23575930276387e+000 },
 {-4.27953674542435e-003, 5.22558713365225e-002, -2.02630063174860e-002, 3.55526806539767e-001, 4.38086865729618e-001 },
 {8.36988375566669e-003, 6.51471505482633e-002, -1.87374240864233e-005, -2.93668874571122e-002, 4.28066974047903e+000 },
 {-4.51758787016232e-003, -1.65295098364779e-002, -9.55504020044409e-003, 3.37743795686801e-002, -3.92503675526061e+000 },
 {8.09715832028303e-003, 1.40380167160264e-001, -5.11574576386343e-002, 4.63819910138810e-002, -4.14377599772541e+000 },
 {2.64573937244475e-002, 1.38171451159571e-001, -5.05816891278269e-002, 3.03640534484363e-002, -4.99478506036909e+000 },
 {3.56819229312110e+001, -1.44427879618125e+001, 5.05330645369423e+000, 6.79476064151221e+000, -1.20589227870316e+001 },
 {1.63237325776004e-002, 9.93417867044676e-002, -4.79318114375762e-003, -2.87432514207203e-002, 3.81634117448330e+000 },
 {5.06016857284542e-003, -6.65190232893867e-003, 1.98066912357954e-002, -3.29635020222122e-002, 3.65513674103006e+000 }
};
```

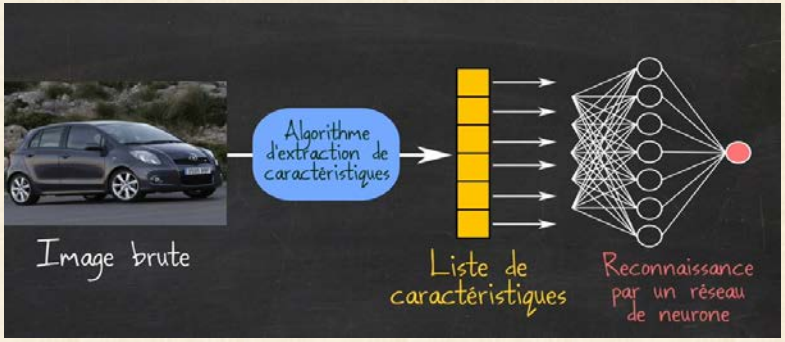


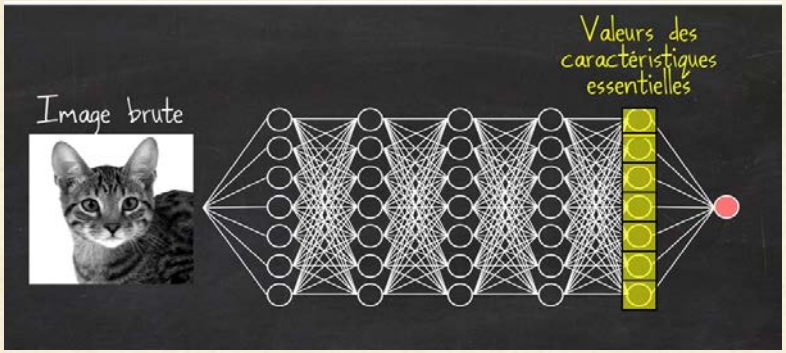
IA = réseaux neurones multi couches (réseau profond)

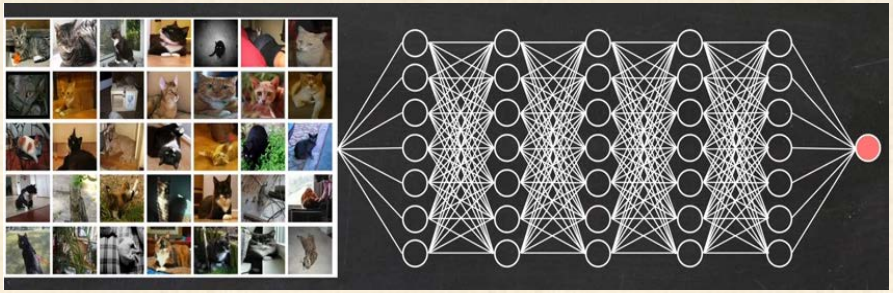
apprentisssage profond : deep learning = IA

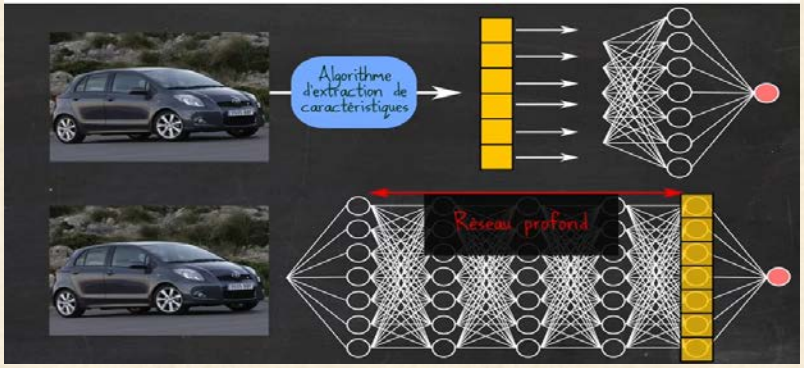








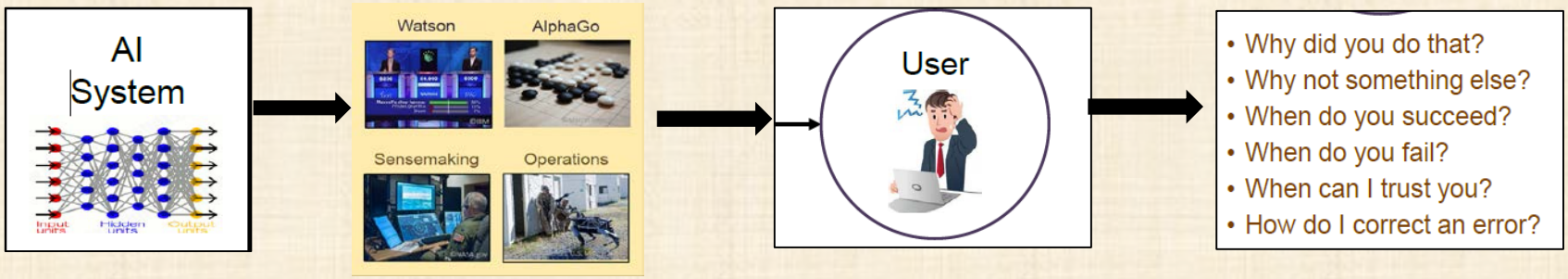




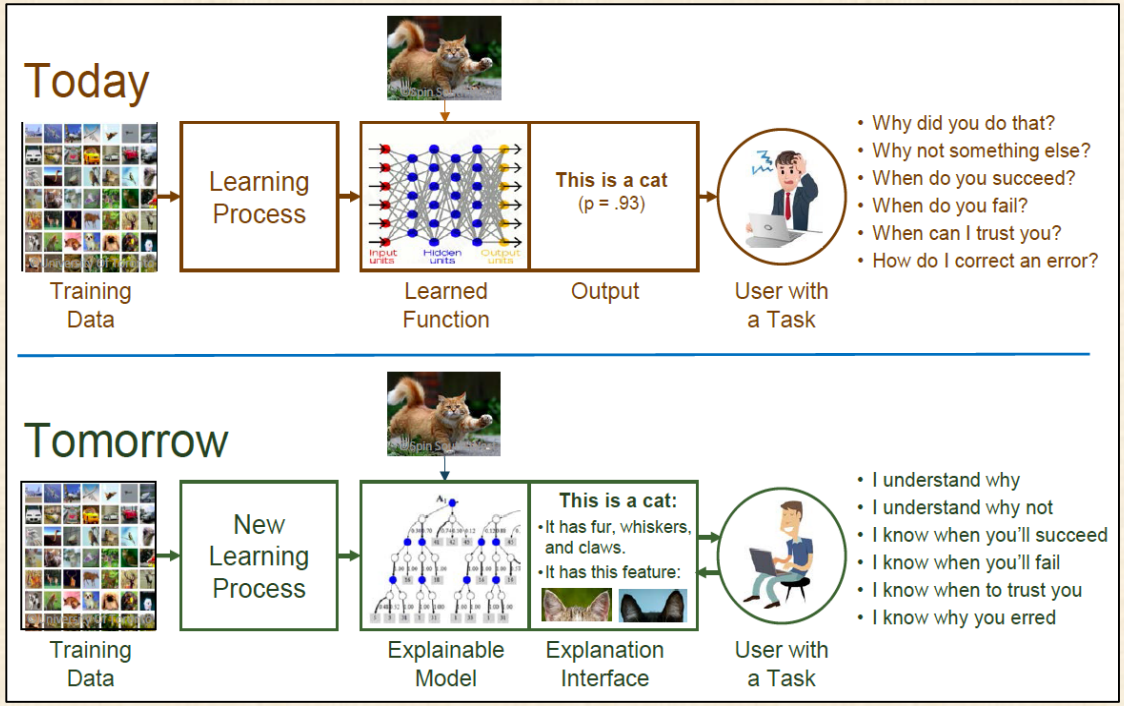
découverte des caractéristiques communes : hauteur, largeur, nombre de roues, …

algorithmes : RNN CNN ..

mais la boite noire est encore plus fortement incompréhensible …



EXplainable AI (XAI) - What Are We Trying To Do?



REMARQUE

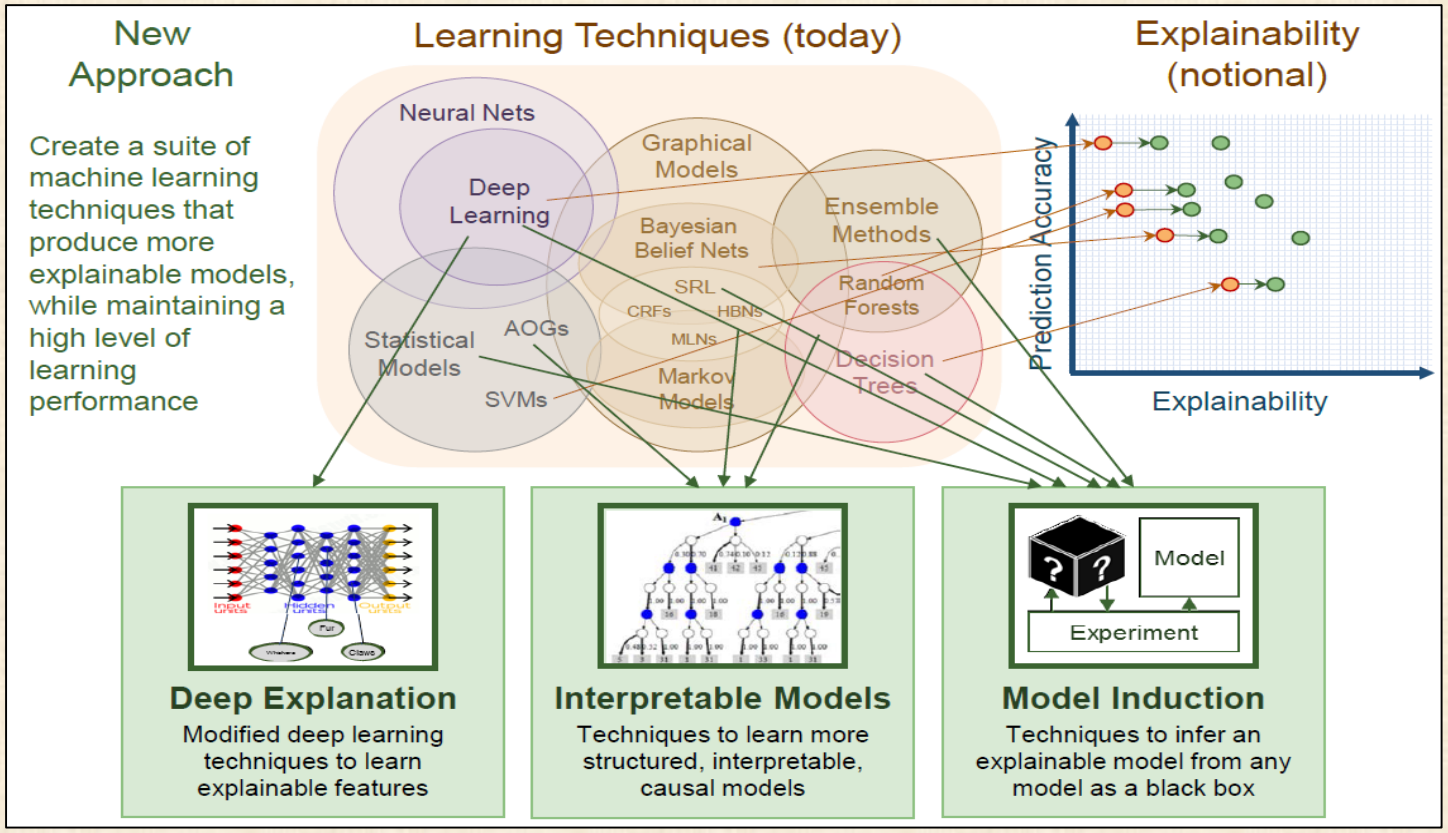
l'explication de l'action / décision va au-delà de la seule prédiction (pensons traitement médical)

---

la connaissance scientifique va encore bien plus loin …

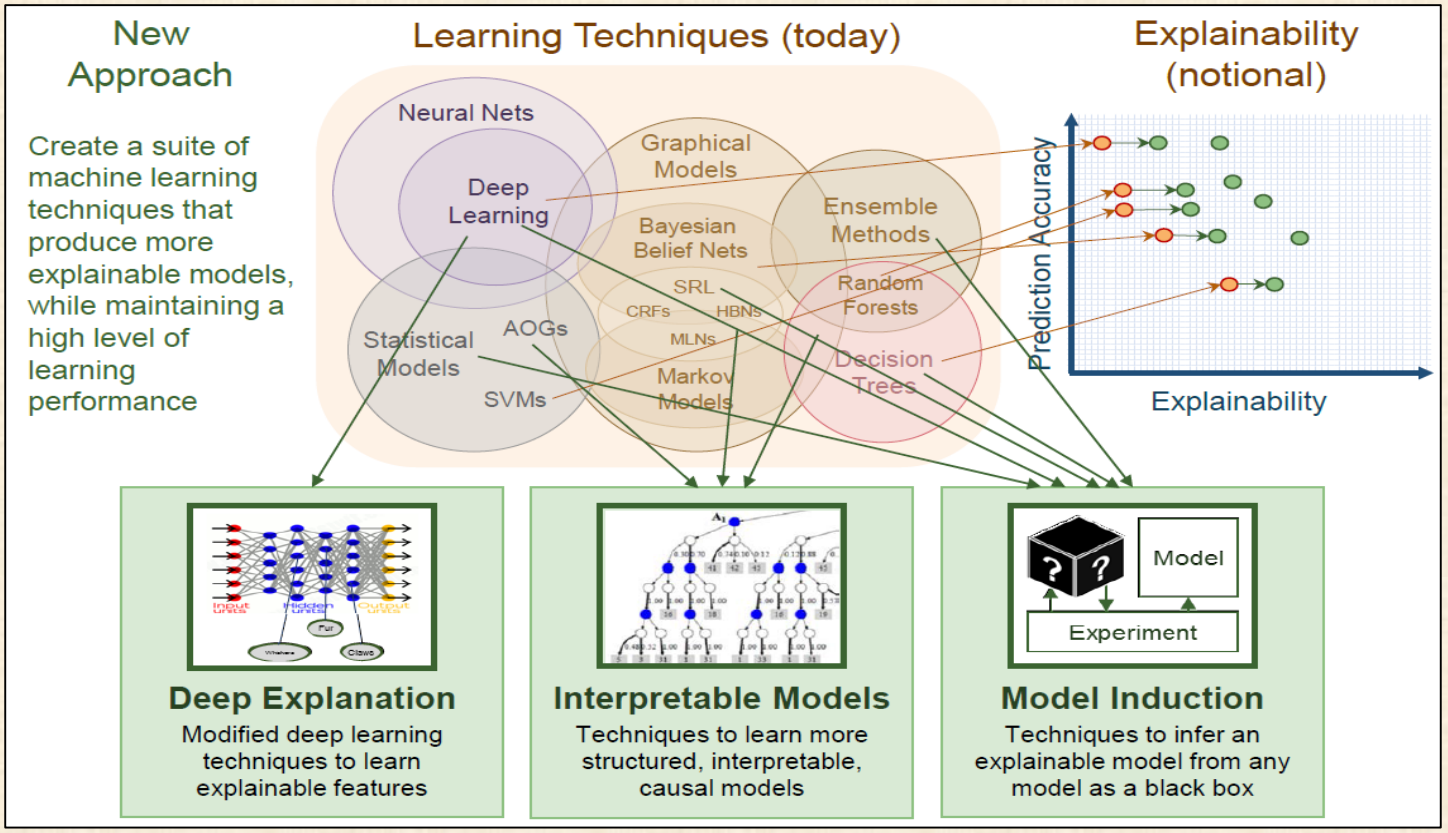
---

la vraie intelligence artificielle ce n'est pas celle qui est abondamment galvaudée aujourd'hui …



Rôle de la statistique dans le AI d'aujourd'hui de demain ?

OUI c'est ……



Rôle de la statistique dans le AI d'aujourd'hui de demain ?

OUI c'est …… l'analyse de sensibilité et de robustesse des systèmes AI particulièrement : médical … sécurité …

QUESTIONS

Q1 : quel est l'avenir de la science statistique ?

Q2 : le métier de statisticien est-il en train de changer ?

Q3 : comment doit-on former les futurs statisticiens ?

Q4 : l'inférence statistique a-t-elle encore une place dans l'ère des mégadonnées ?

Q5 : quelle contribution la statistique peut-elle apportée dans le domaine de l' intelligence artificielle ?

Q6 : comment faire pour rehausser l'appréciation et la reconnaissance du rôle du statisticien dans la société ?

Q7 : …. vos questions ….

RÉPONSES

R1 :

R2 :

R3 :

R4

R5 :

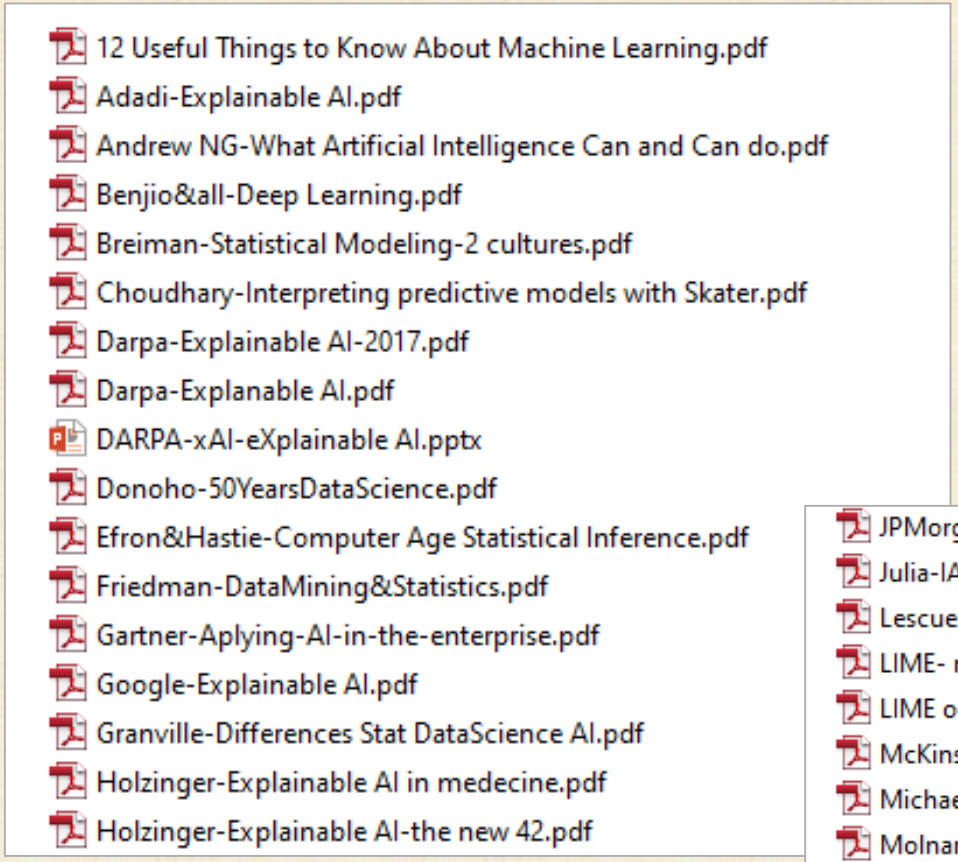
R6 :

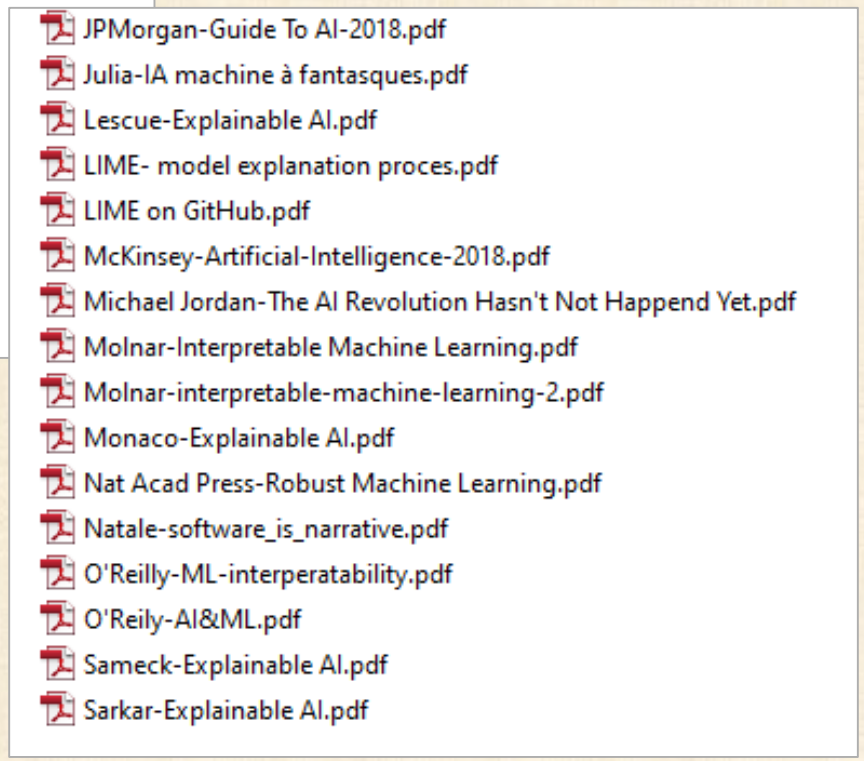
R7 : …. vos questions ….

QUELQUES CONCLUSIONS

- Tout modèle statistique fournit une représentation simplifiée du système dont on a obtenu des données.
- La statistique est avant tout un état d'esprit avec un grand coffre d'outils.
- La modélisation automatisée est risquée au-delà d'un certain niveau.
- La connaissance du domaine d'application est essentielle.
- Un statisticien professionnel habileté à concevoir une étude statistique: c'est la caractérisque principale d'un statisticien.
- Toute personne qui entreprend une étude statistique devrait absolument chercher à collaborer avec un statisticien professionnel pour assurer la plus grande qualité de l'étude. L'expert d'un domaine d'application ne doit pas être le seul responsable de la planification d'une étude statistique. Le statisticien doit être partie prenante dès le début.
- Un statisticien professionnel n'est pas un simple technicien qui analyse des données qu'on lui fournit.
- Dans une étude primaire, les données sont collectées avec un objectif (hypothèse) spécifique de départ.
- Dans une étude secondaire, on analyse des données qui furent collectées avec d'autres objectifs initialement.
- Il ne faut pas croire que l'analyse de mégadonnées rend inutile la démarche inférentielle pour l'évaluation des risques.
- L'analyse statistique inférentielle apporte une contribution essentielle en intelligence artificielle pour étudier la robustesse des systèmes développés en AI.

QUELQUES RÉFÉRENCES





POLYTECHNIQUE MONTRÉAL

LE GÉNIE EN PREMIÈRE CLASSE

2900, boul. Édouard-Montpetit
2500, chemin de Polytechnique
Montréal (Québec) Canada H3T 1J4
Adresse postale
C.P. 6079, succ. Centre-ville
Montréal (Québec) Canada H3C 3A7

BERNARD CLÉMENT, PhD
Professeur titulaire
Département de mathématiques et de génie industriel
Pavillon principal, bureau A-520.30
bernard.clement@polymtl.ca ■ www.polymtl.ca
Tél. : 514 340-4711 poste 4944 ■ Cell.: 514-677-7896

POLYTECHNIQUE MONTRÉAL

WORLD-CLASS ENGINEERING

2900, boul. Édouard-Montpetit
2500, chemin de Polytechnique
Montréal (Québec) Canada H3T 1J4
Mailing Address
P.O. Box 6079, Station Centre-ville
Montréal (Québec) Canada H3C 3A7

BERNARD CLÉMENT, PhD
Full Professor
Department of Mathematics and Industrial Engineering
Main Building, office A-520.30
bernard.clement@polymtl.ca ■ www.polymtl.ca
Tel. : 514 340-4711 ext. 4944 ■ Cell: 514 677-7896

Bernard Clément PhD

Génistat Conseils
Courriel : genistat@sympatico.ca
Tel : (514) 677-7896

Département de mathématiques
et de génie industriel
École Polytechnique de Montréal
Tél. : (514) 340-4711 poste 4944
Courriel : bernard.clement@polymtl.ca

Bernard Clément, PhD
Statisticien

Génistat Conseils

1205-80 Berlioz
Verdun, QC
Canada H3E 1N9

Cell : 514-677-7896
genistat@sympatico.ca

Bernard Clément, PhD est professeur titulaire au département de mathématiques et de génie industriel de l'École Polytechnique de Montréal affiliée à l'Université de Montréal. Il possède plus de 30 années d'expérience en enseignement des méthodes de statistiques appliquées et en management de la qualité aux ingénieurs et scientifiques.

Il a fondé Génistat Conseils Inc., une firme de consultation spécialisée en design et analyse d'études statistiques. Son produit principal est le transfert d'expertise, de connaissance et de management pour l'amélioration de la qualité des produits et des procédés.

Sa liste de clients comprend IBM, Sidbec-Dosco, Noranda Research Center, Bolting Technology Council, Nortel, Institut de Recherche en Biotechnologie, Compagnie Générale des Eaux (Vivendi), Bell, Postes Canada, DALSA semiconducteurs, Cardianove, Warnex, Camoplast, CIRANO et plusieurs établissements de recherche.

Il est membre élu de l'International Statistical Institute et membre de American Society of Quality. Il a été vice-président du Canada Quality Council, administrateur de l'Association québécoise de la Qualité (Mouvement Québécois Qualité), et il fut président de la Société Statistique de Montréal. Il a été membre du comité ISO du Standard Council du Canada.

Consultation - recherche - formation

- Planification et analyse d'expériences industrielles (DOE)
- Maîtrise statistique des processus (SPC)
- Études statistiques appliquées: design, analyse, data mining
- Management de la qualité et Six Sigma
- Ingénierie robuste de Taguchi : design de produit et de procédé
- Logiciels statistiques : STATISTICA, MINITAB, JMP, SAS, DESIGN-EXPERT

Bernard Clément, PhD

Professeur titulaire | Mathématiques et génie industriel | Polytechnique Montréal

E bernard.clement@polymtl.ca | O +1 514-340-4711 x 4944 | M +1 514-677-7896 | bureau A-520.30

2900, boul. Édouard-Montpetit, Campus de l'Université de Montréal

2500, chemin de Polytechnique, Montréal, QC, Canada H3T 1J4

Adresse postale : C.P. 6079, succ. Centre-ville, Montréal, QC, H3C 3A7

Enseignement : http://www.groupes.polymtl.ca/mth6301

Expertise : http://www.polymtl.ca/expertises/clement-bernard

CV : http://www.groupes.polymtl.ca/mth6301/mth8301/PDF/Clement-CV.pdf

Sites Internet

http://www.groupes.polymtl.ca/mth6301	**Planification et analyse d'expériences**
http://www.groupes.polymtl.ca/mth6301/MTH8302.htm	**Modèles de régression et d'analyse de variance**
http://www.groupes.polymtl.ca/ind2501	**Ingénierie de la qualité**
http://www.groupes.polymtl.ca/mth6301/STATISTICA.htm	**STATISTICA**
http://www.groupes.polymtl.ca/mth6301/JMP.htm	**JMP**

Consultation - recherche - formation

- Planification et analyse d'expériences industrielles (DOE)
- Maîtrise statistique des processus (SPC)
- Études statistiques appliquées: design, analyse, data mining
- Management de la qualité et Six Sigma
- Ingénierie robuste de Taguchi : design de produit et de procédé
- Logiciels statistiques : STATISTICA, MINITAB, JMP, SAS, DESIGN-EXPERT