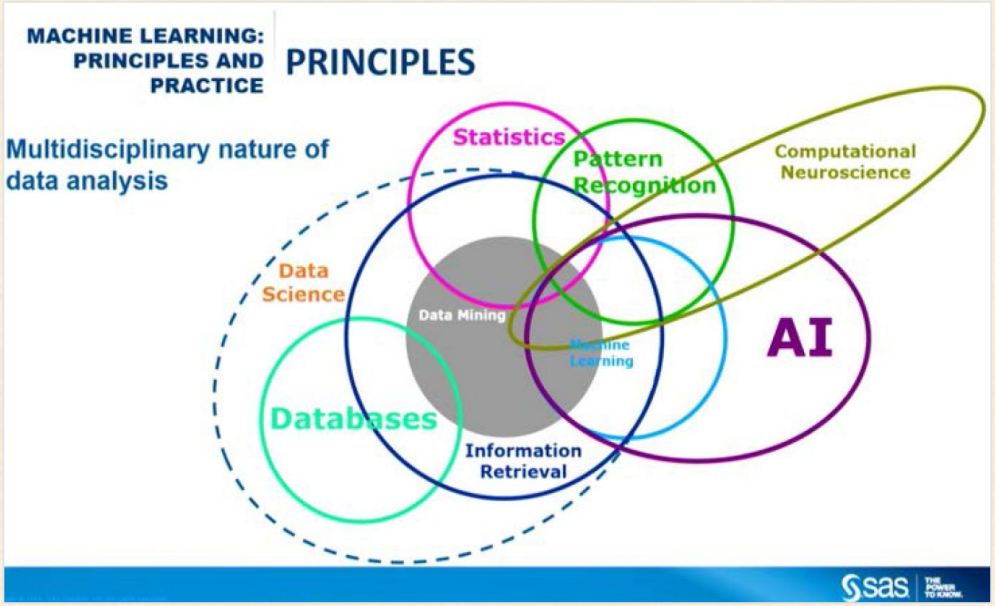
Chapitre 1 – Introduction

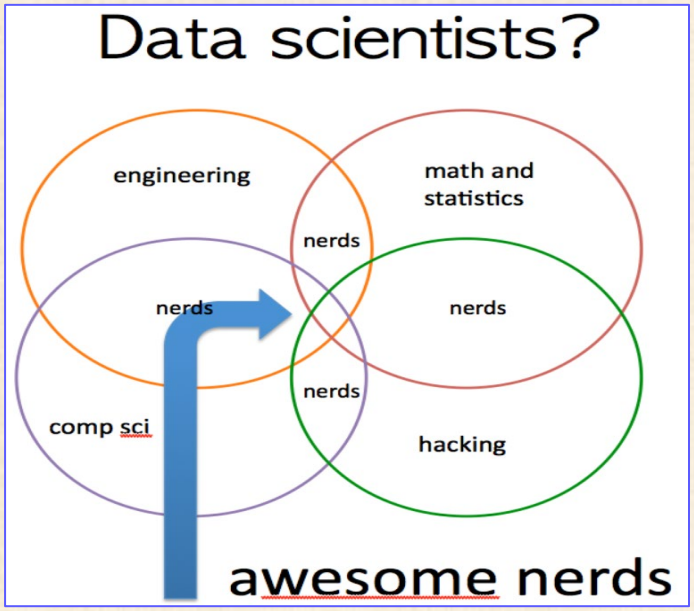
- Data Science (DS) = ?
- Machine Learning (ML) = ?
- Processus - modélisation statistique
- Étapes d'une étude statistique
- Types de variables
- Classification des modèles

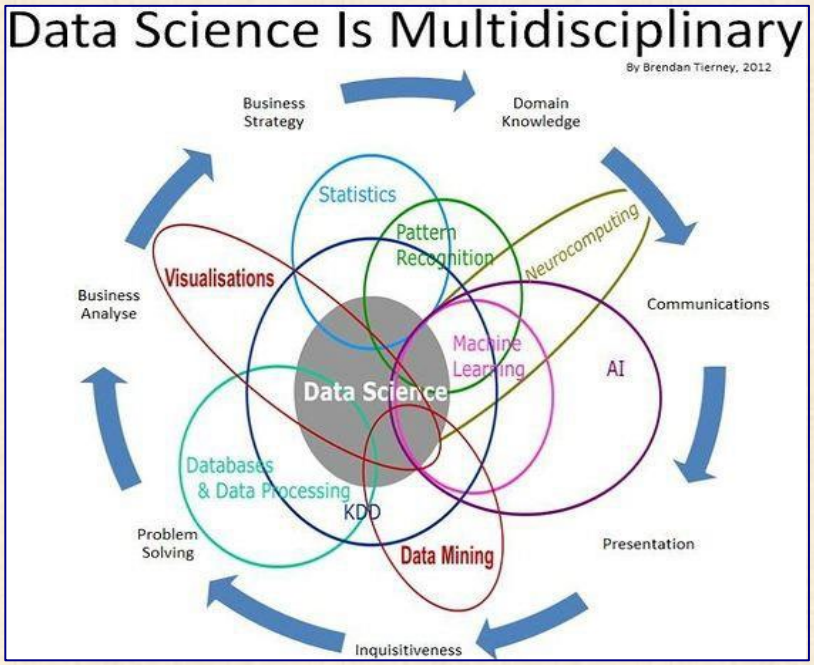
Référence

Bernard CLÉMENT, PhD
Statistique, Science des données, Intelligence artificielle
Société Statistique de Montréal : 26 avril 2018

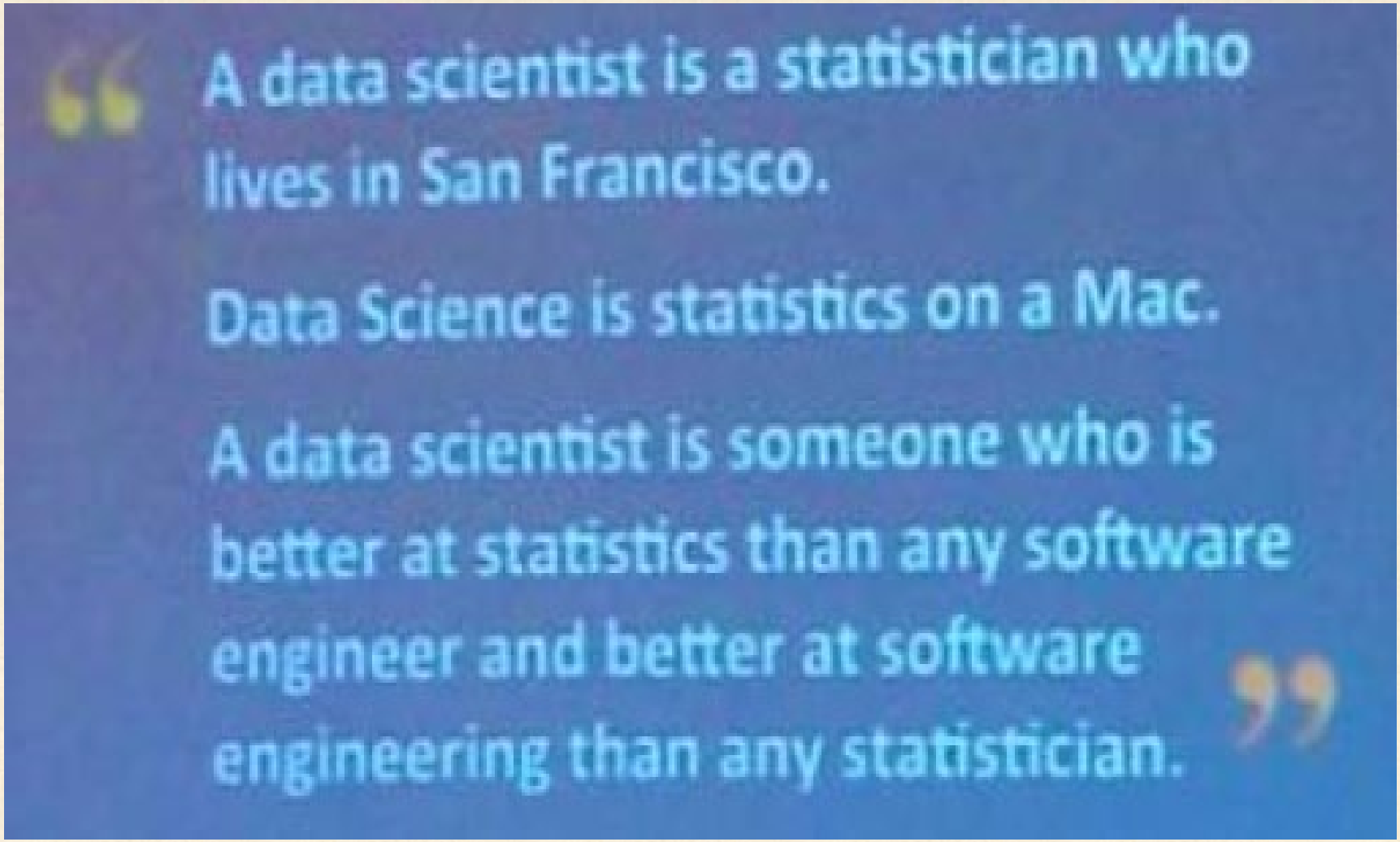
science des données = ?







What on earth is a data scientist ?

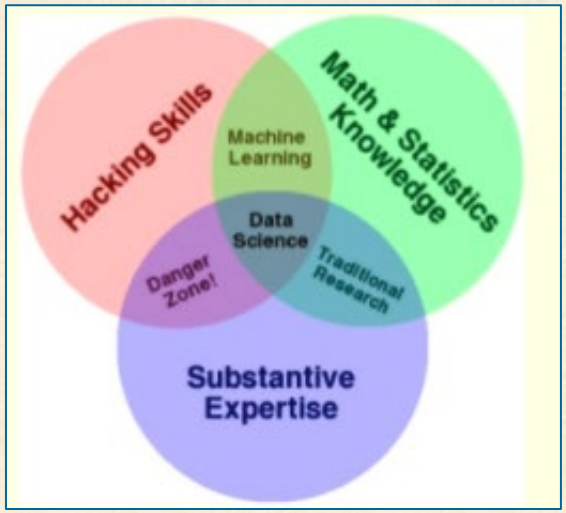


DIFFÉRENCE entre STATISTIQUE et DATA SCIENCE ?

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

idée ⬇ données

idée ⬆ données

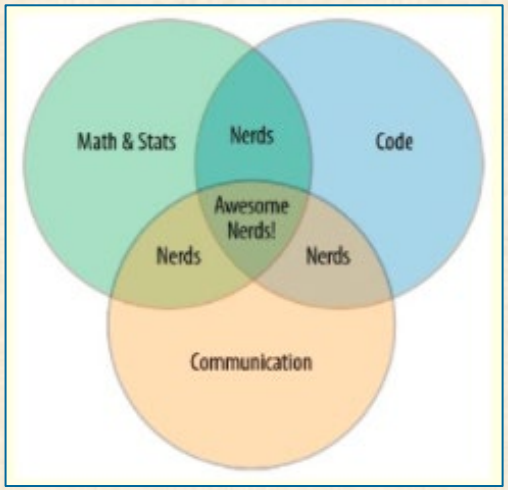


" Data Science is much older than Kepler .. It is the second oldest profession "

Gregory Piatetsky-Shapiro

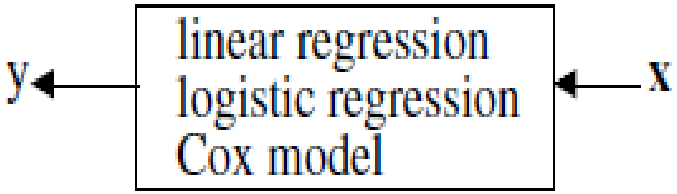
" Statistics has been the most successful information science. Those who ignore statistics are condemned to re-invent it "

Brad Efron



génie logiciel-informatique … **2 cultures** … statistiques-mathématiques

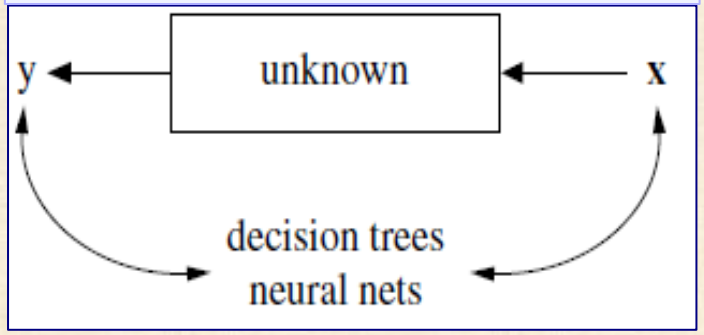
Stochastic Data Modeling Culture

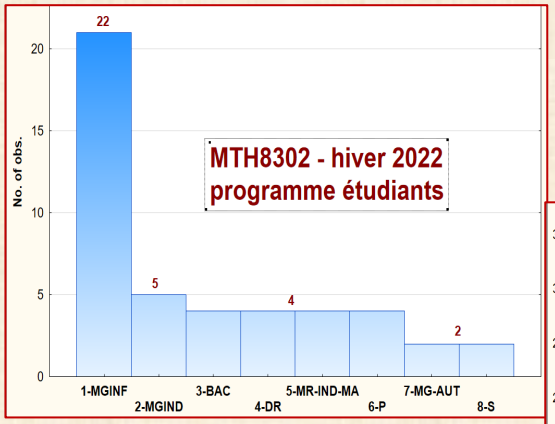


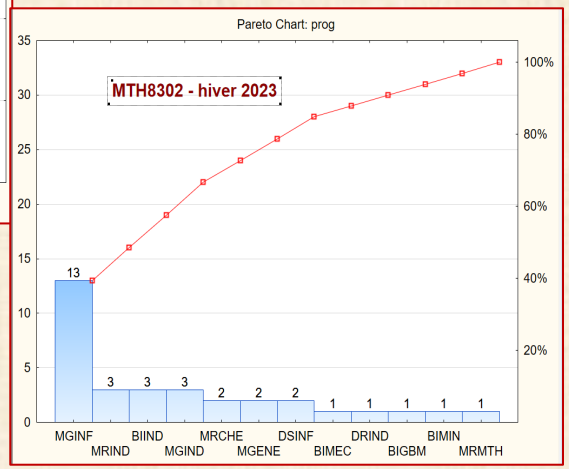
Statistician : 98%
Computer Scientists : 2%
dép. MAGI

Statistician : 2%
Computer Scientists : 98%
dép. GIGL

Algorithmic Modeling Culture







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

Breiman-two-cultures.pdf (ku.dk)

Terminologie

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

… science essentiellement **pluridisciplinaire**

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

le terme DM est maintenant remplacé par ML

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**

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

BD : Big Data (mégadonnées) (données massives) … 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)**

Terminologie

apprentissage supervisé	….	apprentissage non supervisé
données structurées	…	données non structurées (images, textes,..)
apprentissage profond	…	réseaux de neurones multicouches pour IA
infonuagique (cloud)	...	systèmes distribués
IOT	…	internet des objets … réseaux de capteurs
Open Source Software	...	R , Python , Weka, Julia …
technologies	…	GPU (puces graphiques - traitement parallèle)
économie numérique	...	GAFAM (Google Amazon Facebook Amazon Microsoft)

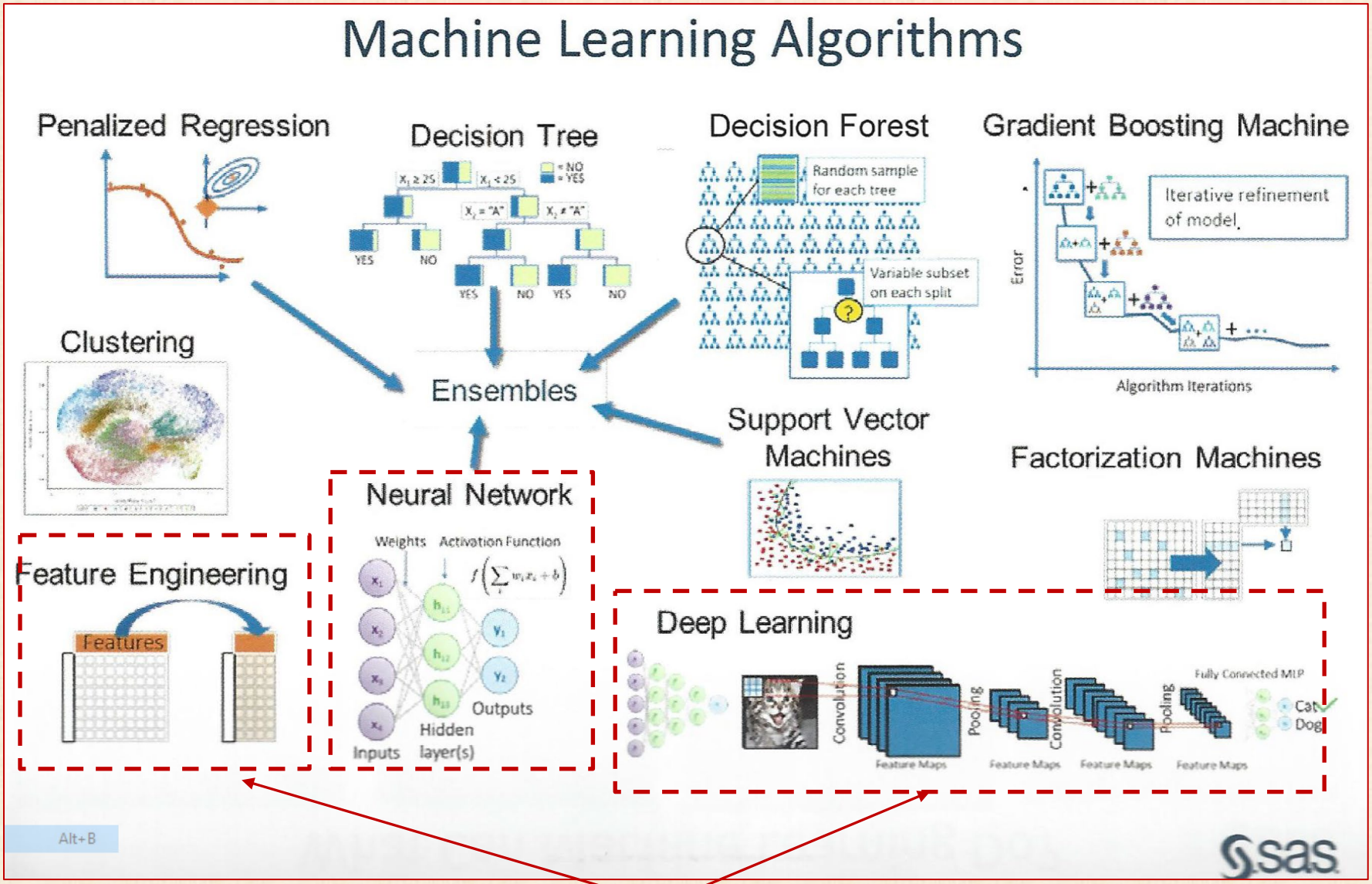
Data Mining

Machine Learning (ML)

Apprentissage statistique

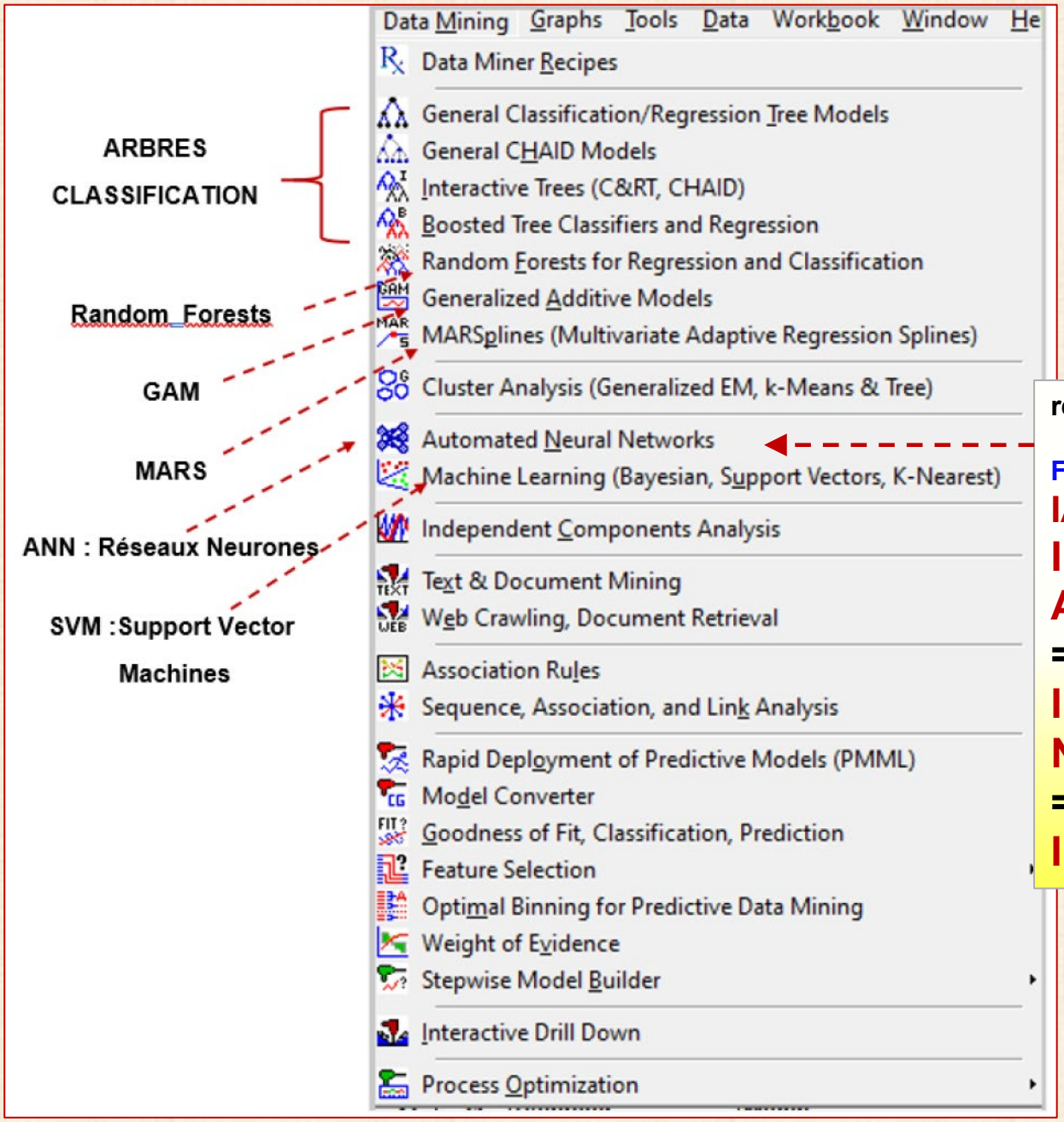
Supervisée	Non supervisée	Semi-supervisée
Régression	classification hier.	classification & prédiction
Arbres de décision (CRT)	Composantes principales (PCA)	EM
Réseau Neurones	Factorisation	Feature engineering
Random Forest	Support Vector Machine (SVM)	Auto encoder
	Bayes Network	
	Processus Gaussien	

Algorithmes du ML



IA : Intelligence Artificielle

ML Avec Statistica



réseau neurones

Fondement de

IA =

Intelligence Artificielle

=

Intelligence Numérique

=

IN

DIFFÉRENCES de 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 $

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 →		

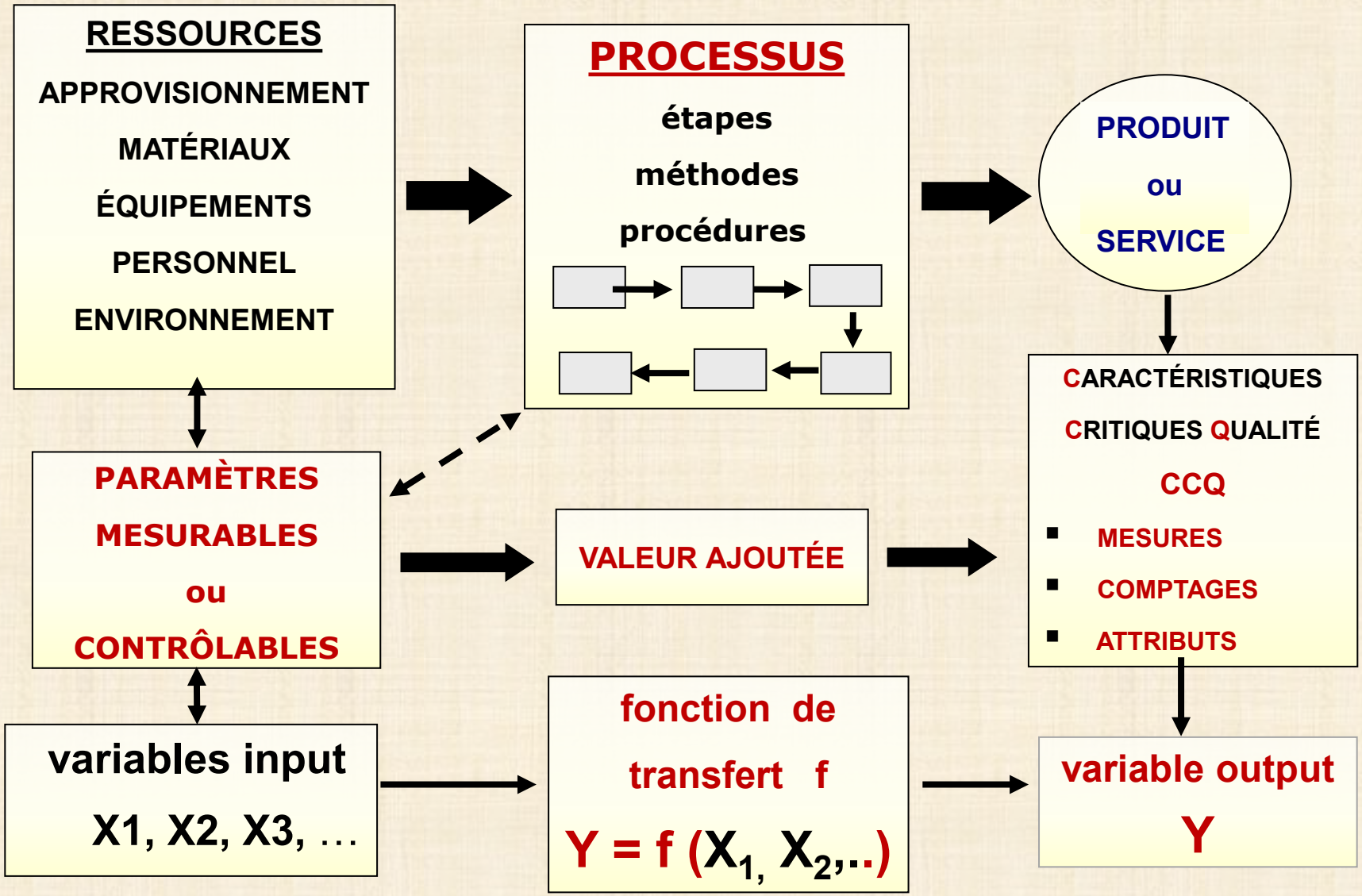
PROCESSUS / SYSTÈME

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

exemples en intelligence artificielle (IA)

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 ingénieur

passif

MTH8301

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 .....
(small data)

Sondages, enquêtes, recensements
=
études énumératives

plan d'échantillonnage des unités statistiques
pas de traitements appliqués aux unités

sciences humaines, sciences sociales, ...

MTH8302

Études observationnelles
données collectées au fil du temps / temps réel
unités statistiques
=
instants d'observations
peu / pas de planification statistique

banques de données, mégadonnées
(big data)

Type d'étude Statistique

CARACTÉRISTIQUE	OBSERVATIONNELLE (mode passif)	EXPÉRIMENTALE (mode actif)
provenance des données	- historiques - suivi de processus dans des conditions normales d'opération; -le processus n'est pas manipulé (volontairement perturbé);	on fait varier le processus sous différentes conditions (variables) dans le but d'obtenir des données
quantité de données	-généralement abondantes; -on peut obtenir des observations additionnelles	fixe et limitée
qualité de données	peut présenter des difficultés : changements non documentés, données manquantes etc.	excellente
coût	généralement faible	généralement élevé
but	modélisation et exploration	détecter des changements
hypothèse sous-jacente	homogénéité des données (*)	hétérogénéité causée par les perturbations induites
méthodes d'analyse	- carte données individuelles et étendues mobiles XmR; - carte Xbar&R ou Xbar&S afin de vérifier l'homogénéité des données - méthodes de Data Mining	- analyse de la variance - analyse de régression - autres méthodes - cartes XmR, Xbar&R,…

Concepts

Facteurs
- primaires
- secondaires
- blocs

Design plan

Protocole expérimental

Unités statistiques

Randomisation

Répétition

……..

* L'homogénéité des données est fondamentale lors de leur l'analyse.
Cette question est clarifiée dans l'article suivant :
Wheeler, Donald J. (2009) *The four Questions of Data Analysis*

http://www.qualitydigest.com/inside/quality-insider-column/four-questions-data-analysis.html

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 !

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

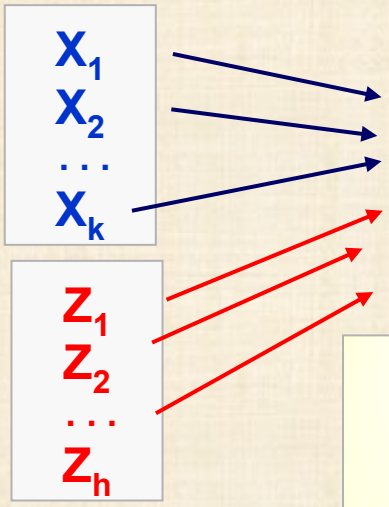
ÉTAPES ÉTUDE STATISTIQUE CLASSIQUE

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

ANALYSE STATISTIQUE CLASSIQUE : comprendre / prédire / optimiser



réponse - sortie mesurée

$Y = f(X) + g(Z) + \varepsilon$

peut avoir plusieurs Y

ε = erreur
=
sources inconnues
de variabilité

$\varepsilon \sim (0, \sigma^2)$

" 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 (Voltaire)

Aucune restriction concernant la nature des X et Y

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

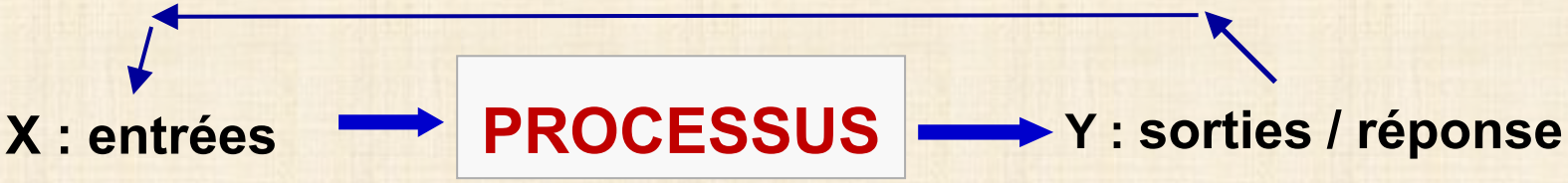
Algorithmes du Machine Learning

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 !

Étude des relations entrées-sorties : Analyse supervisée



COMPARAISON	Modèle de régression	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

approche processus

X : entrées → PROCESSUS → Y : sorties / réponse

Quelles sont les variables CRITIQUES X affectant les variables de réponse Y ?	IDENTIFICATION
Quelle est la FONCTION de TRANSFERT f entre les variables critiques X et la variable de réponse variable Y ?	MODÉLISATION f X → Y = f (X)
Comment CONTRÔLER la réponse Y à un niveau désiré nominal - maximum - minimum en fixant les variables X à des niveaux spécifiques ?	CONTRÔLE et OPTIMISATION

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 | mixte = effets fixés + effets aléatoires

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

modèle mixte $Y = X\beta + Zu + \varepsilon$ $u \sim N(0,G)$ $\varepsilon \sim N(0,R)$ $Cov[u, \varepsilon] = 0$

- Y vector of responses
- X known design matrix of the fixed effects
- β unknown vector of fixed effects parameters to be estimated
- Z known design matrix of the random effects
  - if Z = 0 modèle effets fixés
- u unknown vector of random effects
- ε unobserved vector of random errors

valeurs individuelles $Yi = Xi\beta + Zi\ ui + \varepsilon i$ $i = 1, \ldots, N$

- Yi ni×1 vector of responses subject i
- Xi ni×p design matrix of fixed effects subject i (p is the number of columns in X)
- β p×1 vector of regression parameters
- Zi ni×q design matrix of the random effects subject i
- ui q ×1 vector of random effects for subject i which has means of zero and covariance matrix Gsub
- εi ni×1 vector of errors subject i with zero mean and covariance Ri
- ni number of repeated measurements subject I
- N number subjects
- ei vector of residuals for subject *i* (ei = yi - Xiβ)

VARIABLES et MODÈLE

VARIABLES

Nature: continue - catégorique

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

Liste des X complète? k = nombr e 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?

importance obsessive sur 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

- validation croisée

Classification des modèles statistiques

Modèle général $Y = \varphi(X_1, X_2,\ldots, X_k; \beta_0, \beta_1, \beta_2,\ldots, \beta_p) + \varepsilon \quad \varepsilon \sim N(0,\sigma^2)$ (1)

Modèle LINÉAIRE dans les β si

$\varphi(X_1, X_2,\ldots, X_k; \beta_0, \beta_1, \beta_2,\ldots, \beta_p) = \Sigma\, \beta_j\ f_j(X_1, X_2,\ldots,X_k)$ (2)

$f_j(X_1, X_2,\ldots, X_k) = U_j$ ne dépend pas de paramètre inconnu (3)

alors $Y = \Sigma\, \beta_j\, U_j + \varepsilon$ (4)

Modèle sans variable explicative: $Y = \beta_0 + \varepsilon$

Modèle de régression par l'origine: $Y = \beta_1 X + \varepsilon$

Modèle de régression linéaire simple: $Y = \beta_0 + \beta_1 X + \varepsilon$

Modèle de régression linéaire multiple: $k \geq 2$ ou plus variables explicatives

Modèles intrinsèquement linéaires: linéaires après transformations sur X et ou Y

exemple: $Y = \beta_0 \exp(\beta_1 X + \varepsilon) \longrightarrow Y^* = \ln(Y) = \beta_0' + \beta_1 X + \varepsilon$

Modèles intrinsèquement non linéaires: équations (2) et (3) non satisfaites et aucune transformation sur X ou Y ne permet de se ramener à ce cas

exemple: $Y = \beta_0 + \beta_1 \exp(\beta_2 X) + \varepsilon$

Modèles linéaires généralisés (GLZ) $g(Y) = \varphi(X_1, X_2,\ldots, X_k; \beta_0, \beta_1, \beta_2,\ldots, \beta_p) + \varepsilon$

Modèles d'analyse de la variance présence de variables catégoriques

ALGORITHMES (méthodes) (Machine Learning)

SUPERVISÉES : X et Y

- Régression multiple ordinaire
- Régression non linéaire
- Régression linéaire généralisée
- Régression avec contraintes: Ridge, Lasso
- Régression splines (MARS)
- Régression généralisée additive
- Régression réseaux neuronaux
- Flux Tenseur
- Arbres de classification (CRT)
- Forêts Aléatoires
- Méthodes gradient non-convexe
- Algorithmes génétiques
- Méthodes ensemblistes
- Régression boosted
- XGBoost
- …

NON SUPERVISÉES : X

- Réduction dimension (PCA)
- Clustering
- K-Means
- K-Neighbour
- Classification hiérarchique
- Réseaux Baysiens
- Modèle de Markov
- ….

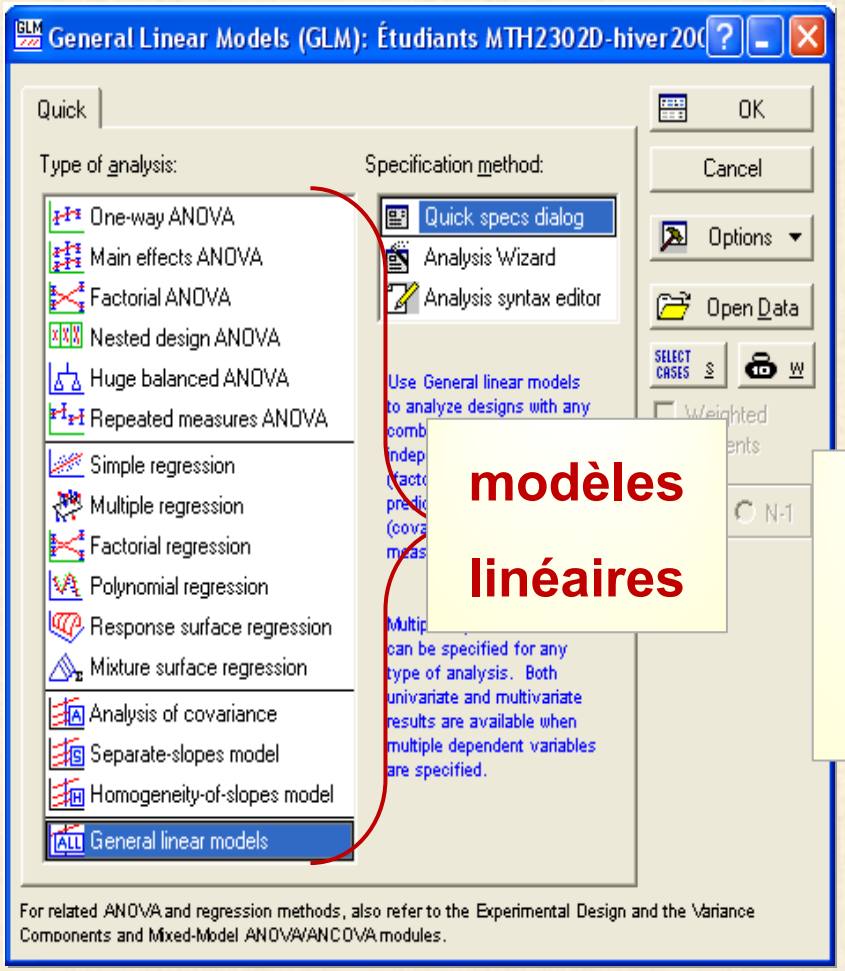
SÉRIES CHRONOLOGIQUES

Deep Learning
= Apprentissage profond
= réseaux neurones multicouches
= intelligence artificielle (AI)

modèles statistiques classiques

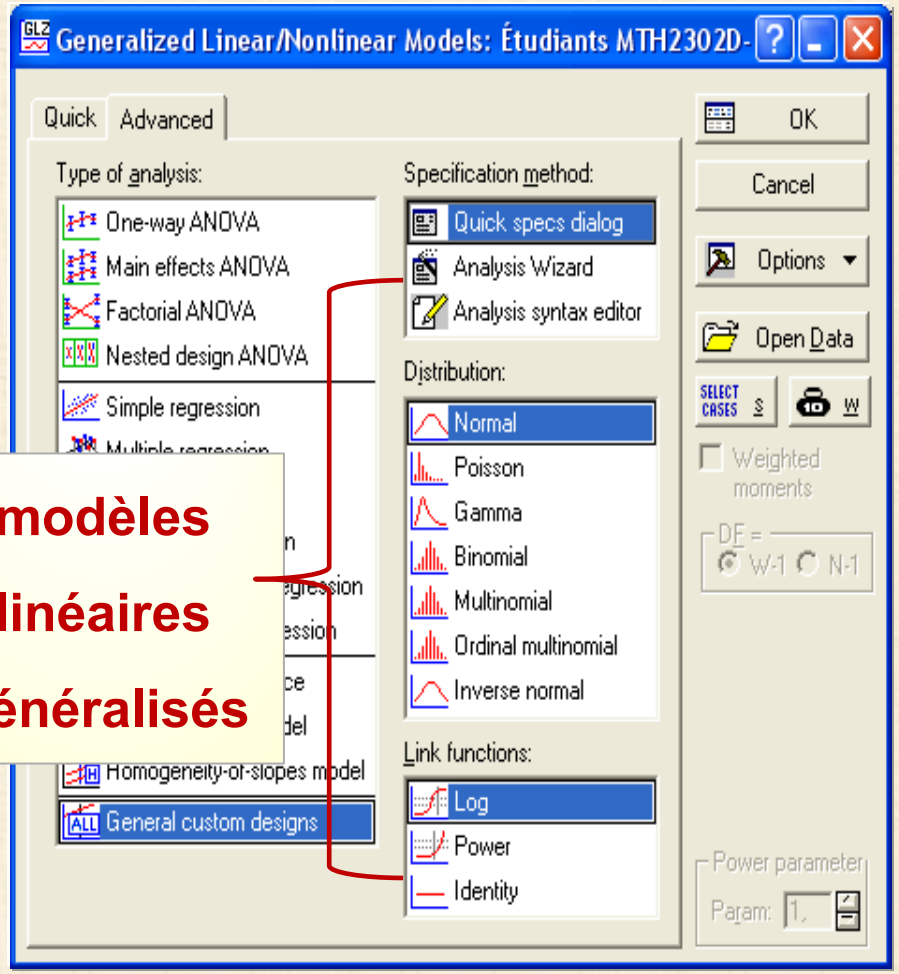
logiciel Statistica

GLM : General Linear Model



modèles linéaires

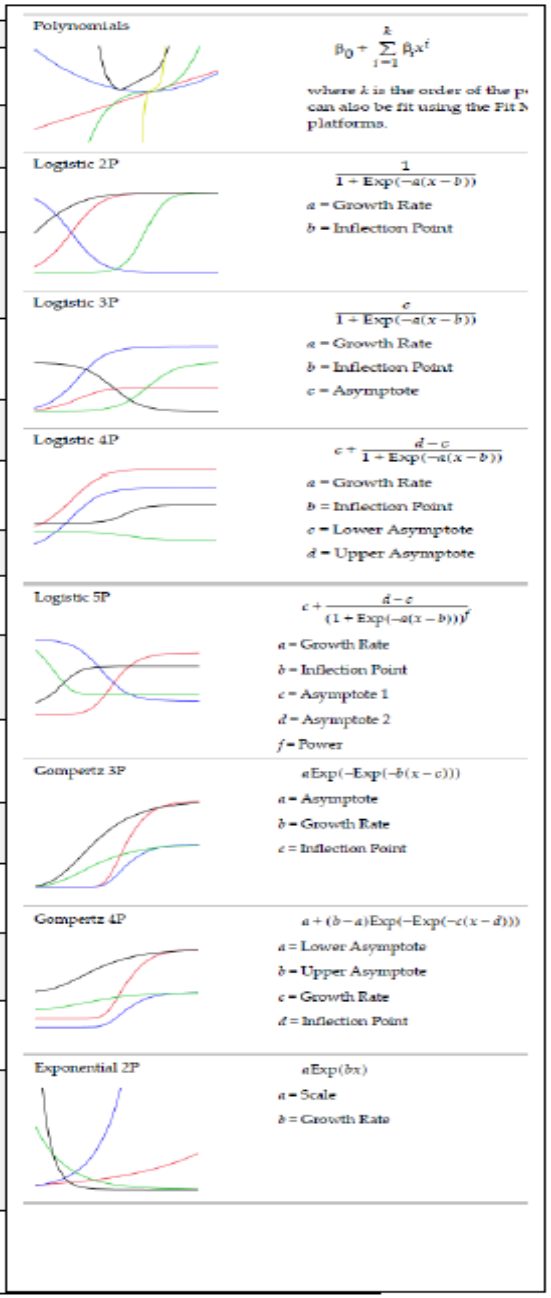
GLZ : Generalized Linear/Nonlinear Model



modèles linéaires généralisés

Exemples de modèles de régression non-linéaires

MODÈLE	FORMULE Y =	
Polynômes	$\beta_0 + \sum_{i=1}^{k} \beta_i x^i$	
Logistique 2P	$\frac{1}{1+\mathrm{Exp}(-a(x-b))}$	a taux croissance b point d'inflexion
Logistique 3P	$\frac{c}{1+\mathrm{Exp}(-a(x-b))}$	a taux croissance b point inflexion c asymptote
Logistique 4P	$c+\frac{d-c}{1+\mathrm{Exp}(-a(x-b))}$	a taux croissance b point inflexion c asymptote inférieure d asymptote supérieure
Logistique 5P	$c+\frac{d-c}{(1+\mathrm{Exp}(-a(x-b)))^f}$	a taux croissance b point inflexion c asymptote inférieure d asymptote supérieure f puissance
Gompertz 3P	$a\mathrm{Exp}(-\mathrm{Exp}(-b(x-c)))$	a asymptote b taux croissance c point inflection
Gompertz 4P	$a+(b-a)\mathrm{Exp}(-\mathrm{Exp}(-c(x-d)))$	a asymptote inférieure b asymptote supérieure c taux croissance d pont inflection
Exponentielle 2P	aExp(bx)	a échelle b taux croissance
Exponentielle 3P	a + bExp(cx)	a asymptote b échelle c taux croissance
Bi-Exponentielle 4P	aExp(– bx) + cExp(– dx)	a échelle 1 b taux décroissance 1 c échelle 2 d taux décroissance 2
Bi-Exponentielle 5P	a + bExp(– cx) + dExp(– fx)	a asymptote b échelle 1 c taux décroissance 1 d échelle 2 f taux décroissance 2
Croissance mécanique	a(1 – bExp(– cx))	a asymptote b échelle c taux croissance
Sommet Gauss	$a\mathrm{Exp}\left(-\frac{1}{2}\left(\frac{x-b}{c}\right)^2\right)$	a sommet b point critique c taux croissance
Sommet Lorentz (Cauchy)	$\frac{ab^2}{(x-c)^2+b^2}$	a sommet b taux croissance c point critique
dose orale 1 compartiment	$\frac{abc}{c-b}(\mathrm{Exp}(-bx)-\mathrm{Exp}(-cx))$	a aire courbe b taux élimination c taux absorption
dose Bolus 2 compartiments	$\frac{e}{\alpha-\beta}((\alpha-b)\mathrm{Exp}(-\alpha x)-(\beta-b)\mathrm{Exp}(-\beta x))$ $\alpha=\frac{1}{2}(b+c+d+\sqrt{(b+c+d)^2-4bd})$ $\beta=\frac{1}{2}(b+c+d-\sqrt{(b+c+d)^2-4bd})$	a concentration initiale b taux transfert in c taux transfert out d taux élimination
Michaelis-Menten	$\frac{ax}{b+x}$	a taux max réaction b infinité adverse



mesures répétées

7 CHOLESTEROL	8 patient	9 treatment	10 Y_AprilAM	11 Y_AprilPM	12 Y_MayAM	13 Y_MayPM	14 Y_JuneAM	15 Y_JunePM
Cholesterol	p1	A	278,0	280,0	204,0	208,0	171,3	175,0
Cholesterol	p2	A	278,0	281,0	195,2	199,0	185,0	189,0
Cholesterol	p3	A	276,0	280,0	213,4	219,0	179,0	181,0
Cholesterol	p4	A	276,0	281,0	201,3	211,0	183,0	187,9
Cholesterol	p5	A	279,0	285,0	188,0	192,0	170,3	174,0
Cholesterol	p6	B	266,0	270,0	220,0	224,0	180,0	184,0
Cholesterol	p7	B	280,0	284,0	228,0	232,0	200,0	204,0
Cholesterol	p8	B	284,0	288,0	233,0	237,		
Cholesterol	p9	B	273,0	277,0	215,0	219,		
Cholesterol	p10	B	281,0	285,0	237,0	241,		
Cholesterol	p11	Control	278,0	282,0	273,0	277,		
Cholesterol	p12	Control	273,0	277,0	274,0	278,		
Cholesterol	p13	Control	282,0	285,0	276,0	281,		
Cholesterol	p14	Control	274,0	277,3	284,5	289,		
Cholesterol	p15	Control	277,0	280,6	279,1	283,		
Cholesterol	p16	Placebo	279,0	283,0	278,0	284,		
Cholesterol	p17	Placebo	277,0	279,0	291,0	291,		

plan central-composite
3 facteurs X - 4 réponses Y

43 TIRETREAD	44 tire	45 X_Silica	46 X_Silane	47 X_Sulfur	48 Y_abrasion	49 Y_modulus	50 Y_elong	51 Y_hardness
tiretread	t1	0,70	40,00	2,80	102	900	470	67,5
tiretread	t2	1,70	40,00	1,80	120	860	410	65,0
tiretread	t3	0,70	60,00	1,80	117	800	570	77,5
tiretread	t4	1,70	60,00	2,80	198	2294	240	74,5
tiretread	t5	0,70	40,00	1,80	103	490	640	62,5
tiretread	t6	1,70	40,00	2,80	132	1289	270	67,0
tiretread	t7	0,70	60,00	2,80	132	1270	410	78,0
tiretread	t8	1,70	60,00	1,80	139	1090	380	70,0
	t9	0,38	50,00	2,30	102	770	590	76,0
	t10	2,02	50,00	2,30	154	1690	260	70,0
	t11	1,20	33,67	2,30	96	700	520	63,0
	t12	1,20	66,33	2,30	163	1540	380	75,0
	t13	1,20	50,00	1,48	116	2184	520	65,0
	t14	1,20	50,00	3,12	153	1784	290	71,0
	t15	1,20	50,00	2,30	133	1300	380	70,0
	t16	1,20	50,00	2,30	133	1300	380	68,5
	t17	1,20	50,00	2,30	133	1300	380	68,5
	t18	1,20	50,00	2,30	142	1090	430	68,0
	t19	1,20	50,00	2,30	145	1260	390	69,0
	t20	1,20	50,00	2,30	142	1344	390	70,0

S. Bisgaard - J. Q. Tech. vol 32 (2000) no 1 p. 39-56
Box, Bisgaard, Quality Engineering, vol 8 (1996) no 4, p. 705-708
Les facteurs A B C D définissent les WholePlot - E est le facteur SplitPlot

	1 StdOrder	2 RunOrder	3 bloc	4 WholePlot	5 A_Pressure	6 B_Power	7 C_GasFlow	8 D_GasType	9 E_PaperType	10 Y_WetAngle
1	5	1	1	1	-1	-1	1	Oxygen	E1	37,6
2	21	2	1	1	-1	-1	1	Oxygen	E2	43,5
3	2	3	1	2	1	-1	-1	Oxygen	E1	41,2
4	18	4	1	2	1	-1	-1	Oxygen	E2	38,2
5	10	5	1	3	1	-1	-1	SiCl4	E1	56,8
6	26	6	1	3	1	-1	-1	SiCl4	E2	56,2
7	14	7	1	4	1	-1	1	SiCl4	E1	47,5
8	30	8	1	4	1	-1	1	SiCl4	E2	43,2
9	11	9	1	5	-1	1	-1	SiCl4	E1	25,6
10	27	10	1	5	-1	1	-1	SiCl4	E2	33,0
11	3	11	1	6	-1	1	-1	Oxygen	E1	55,8
12	19	12	1	6	-1	1	-1	Oxygen	E2	62,9
13	13	13	1	7	-1	-1	1	SiCl4	E1	13,3
14	29	14	1	7	-1	-1	1	SiCl4	E2	23,7
15	6	15	1	8	1	-1	1	Oxygen	E1	47,2
16	22	16	1	8	1	-1	1	Oxygen	E2	44,8
17	16	17	1	9	1	1	1			
18	32	18	1	9	1	1	1			
19	9	19	1	10	-1	-1	-1			
20	25	20	1	10	-1	-1	-1			
21	15	21	1	11	-1	1	1			
22	31	22	1	11	-1	1	1			
23	1	23	1	12	-1	-1	-1	Oxygen	E1	48,6
24	17	24	1	12	-1	-1	-1	Oxygen	E2	57,0
25	8	25	1	13	1	1	1	Oxygen	E1	48,7
26	24	26	1	13	1	1	1	Oxygen	E2	44,4
27	7	27	1	14	-1	1	1	Oxygen	E1	47,2
28	23	28	1	14	-1	1	1	Oxygen	E2	54,6
29	4	29	1	15	1	1	-1	Oxygen	E1	53,5
30	20	30	1	15	1	1	-1	Oxygen	E2	51,3
31	12	31	1	16	1	1	-1	SiCl4	E1	41,8
32	28	32	1	16	1	1	-1	SiCl4	E2	37,8

Expérience en parcelles divisées (SplitPlot) 32 essais

The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine.
1599 red wine 4898 white wine
http://www.vinhoverde.pt/en/
Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available.
(e.g. there is no data about grape types, wine brand, wine selling price, etc.).
The inputs include objective tests (e.g. pH values) and the output is based on sensory data (median of at le
Each expert graded the wine quality between 0 (very bad) and 10 (excellent).
Data mining methods were applied to model the dataset under a regression approach.
Variables
input variables X (based on physicochemical tests)
1 - fixed acidity 2 - volatile acidity 3 - citric acid 4 residual sugar
5 - chlorides 6 - free sulfur dioxide 7 - total sulfur dioxide
8 - density 9 - pH 10 - sulphates 11 - alcohol
Output variable Y (based on sensory data) Y = QUALITY : score between 0 (=very bad) and 10 (=excellent)

ÉTUDE OBSERVATIONNELLE
Vins du Portugal
n=6496 p=13

	1 ID6497	2 QUALITY	3 color	4 fixed acidity	5 volatile acidity	6 citric acid	7 residual sugar	8 chlorides	9 free sulfur dioxide	10 total sulfur dioxide	11 density	12 pH	13 sulphates	14 alcohol
1	1	5	red	7,4	0,70	0,00	1,90	0,076	11	34	0,9978	3,51	0,56	9,4
2	2	5	red	7,8	0,88	0,00	2,60	0,098	25	67	0,9968	3,20	0,68	9,8
3	3	5	red	7,8	0,76	0,04	2,30	0,092	15	54	0,9970	3,26	0,65	9,8
4	4	6	red	11,2	0,28	0,56	1,90	0,075	17	60	0,9980	3,16	0,58	9,8
5	5	5	red	7,4	0,70	0,00	1,90	0,076	11	34	0,9978	3,51	0,56	9,4
6	6	5	red	7,4	0,66	0,00	1,80	0,075	13	40	0,9978	3,51	0,56	9,4
7	7	5	red	7,9	0,60	0,06	1,60	0,069	15	59	0,9964	3,30	0,46	9,4
8	8	7	red	7,3	0,65	0,00	1,20	0,065	15	21	0,9946	3,39	0,47	10,0
9	9	7	red	7,8	0,58	0,02	2,00	0,073	9	18	0,9968	3,36	0,57	9,5

	1 ID2	2 ID	3 couleur	4 fixed_acidity	5 volatile_acidity	6 citric_acid	7 residual_sugar	8 chlorides	9 free_sulfur_dioxide	10 total_sulfur_dioxide	11 density	12 pH	13 sulphates	14 alcohol
6475	6475	5684	blanc	6,5	0,33	0,24	14,50	0,048	20,0	96	0,99456	3,06	0,30	11,50
6476	6476	5718	blanc	6,5	0,43	0,31	3,60	0,046	19,0	143	0,99022	3,15	0,34	12,00
6477	6477	5767	blanc	6,3	0,17	0,32	1,00	0,040	39,0	118	0,98886	3,31	0,40	13,10
6478	6478	5795	blanc	7,1	0,45	0,24	2,70	0,040	24,0	87	0,98862	2,94	0,38	13,40
6479	6479	5932	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6480	6480	5933	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6481	6481	5934	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6482	6482	5935	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6483	6483	5936	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6484	6484	5937	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6485	6485	5938	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6486	6486	5939	blanc	6,8	0,24	0,29	2,00	0,044	15,0	96	0,99232	3,23	0,64	10,40
6487	6487	5940	blanc	7,3	0,19	0,27	13,90	0,057	45,0	155	0,99807	2,94	0,41	8,80
6488	6488	6365	blanc	5,2	0,30	0,34	1,50	0,038	18,0	96	0,98942	3,56	0,48	13,00
6489	6489	6366	blanc	6,4	0,32	0,25	5,00	0,055	28,0	138	0,99171	3,27	0,50	12,40
6490	6490	6386	blanc	4,4	0,32	0,39	4,30	0,030	31,0	127	0,98904	3,46	0,36	12,80
6491	6491	6387	blanc	3,9	0,23	0,40	4,20	0,030	29,0	118	0,98900	3,57	0,36	12,80
6492	6492	6402	blanc	5,8	0,28	0,34	2,20	0,037	24,0	125	0,98986	3,36	0,33	12,80
6493	6493	2374	blanc	9,1	0,27	0,45	10,60	0,035	28,0	124	0,99700	3,20	0,46	10,40
6494	6494	2420	blanc	6,6	0,36	0,29	1,60	0,021	24,0	85	0,98965	3,41	0,61	12,40
6495	6495	2427	blanc	7,4	0,24	0,36	2,00	0,031	27,0	139	0,99055	3,28	0,48	12,50
6496	6496	2476	blanc	6,9	0,36	0,34	4,20	0,018	57,0	119	0,98980	3,28	0,36	12,70

Financial data of 40 UK companies 1983

J. Jobson *Applied Multivariate Data Analysis*, vol 1, Regression and Experimental Design 1991, Springer-Verlag

Y_RETCAP Return on capital employed
X1_WCFTCL Ratio of working capital flow to total current liabilities
X2_WCFTDT Ratio of working capital flow to total debt
X3_GEARRAT Gearing ratio (debt-equity ratio)
X4_LOGSALE Log to base 10 of total sales
X5_LOGASST Log to base 10 of total assets
X6_NFATAST Ratio of net fixed assets to total assets
X7_CAPINT Capital intensity (ratio of total sales to total assets)
X8_FATTOT Gross sixed assets to total assets
X9_INVTAST Ratio of total inventories to total assets
X10_PAYOUT Payout ratio
X11_QUIKRAT Quick ratio
X12_CURRAT Current ratio

ÉTUDE OBSERVATIONNELLE:
données financières
40 entreprises bourse de Londres

	1 ordin	2 Y_RETCAP	3 WCFTCL	4 WCFTDT	5 GEARRAT	6 LOGSALE	7 LOGASST	8 NFATAST	9 CAPINT	10 FATTOT	11 INVTAST	12 PAYOUT	13 QUIKRAT	14 CURRAT
1	1	0,26	0,25	0,25	0,46	4,11	4,30	0,10	0,64	0,12	0,74	0,07	0,18	1,53
2	2	0,57	0,33	0,33	0,00	4,25	4,00	0,12	1,79	0,15	0,27	0,30	1,26	1,73
3	3	0,09	0,50	0,20	0,24	4,44	4,88	0,94	0,36	0,97	0,01	0,57	0,39	0,44
4	4	0,32	0,23	0,21	0,45	4,71	4,44	0,29	1,86	0,52	0,29	0,00	0,69	1,23
5	5	0,17	0,21	0,12	0,91	4,85	4,75	0,26	1,26	0,54	0,33	0,31	0,90	1,76
6	6	0,24	0,37	0,25	0,26	5,61	5,42	0,42	1,54	0,57	6,00	0,15	1,23	1,44
7	7	0,53	0,59	0,40	0,52	4,83	4,30	0,14	3,34	0,21	0,00	0,21	0,83	0,83
8	8	0,26	0,44	0,37	0,24	4,49	4,35	0,40	1,38	1,04	0,36	0,16	0,58	1,45
9	9	0,13	0,21	0,21	0,19	4,13	4,17	0,06	0,91	0,11	0,29	0,39	1,95	2,89
10	10	0,16	0,21	0,18	0,29	4,40	4,17	0,21	1,70	0,40	0,58	0,46	0,56	2,13
11	11	0,06	0,01	0,01	0,85	4,30	4,09	0,23	1,60	0,38	0,34	0,00	0,73	1,31
12	12	0,07	0,70	0,70	0,02	3,62	4,45	0,54	0,15	0,63	0,00	0,00	4,51	4,57
13	13	-0,18	-0,58	-0,32	0,76	4,13	4,35	0,54	0,60	0,84	0,00	0,00	0,47	0,47
14	14	0,12	0,11	0,11	0,39	4,11	3,74	0,41	2,34	0,97	0,49	0,00	0,14	0,85
39	39	0,14	0,39	0,39	0,02	3,99	3,85	0,52	1,38	0,82	0,24	0,78	0,66	1,37
40	40	0,13	0,30	0,20	0,23	4,54	4,37	0,32	1,49	0,46	0,28	0,58	1,47	2,49

exemple : données observationnelles
prix résidences vs caractéristiques

Harrison, D, Rubenfeld, D. (1978) Hedonic Prices and the Demand For Clean Air
J. of Environmental Economic and Management, v.5, 81-102
Combined information from 10 separate governemental and education sources
506 census tracts (CT) in city of Boston of the year 1970
But: modéliser la relation entre 12 indicateurs de la qualité de vie (X) et la valeur d'une propriété Y = MV
Le fichier de 506 observations est divisé en 2 groupes
GROUP = M pour le développement des Modèles (405 observations: 80% des observations)
les données M sont surlignées (vert) et constituent un filtre;
la modélisation analyse statistique est basée sur ces données (405 obs.)
voir *Tools...selections conditions...edit*
GROUP = T pour Tester le modèle (101 observations: 20% des observations)
pour inclure ce groupe dans une prédiction, éditer le filtre

INDICATEURS
X1 CRIM = CRIMe Rate Per Capita by town
X2 NOX = Nitric OXide concentration (parts per 10 million)
X3 AGE = Proportion of owner occupied units built prior to 1940
X4 DIS = Weighted DIStances to five Boston employment centers
X5 RM = Average number of RooMs per dwelling
X6 LSTAT = % of the Lower STATus of the population
X7 RAD = Index of accessibility to RADical highways
X8 CHAS = CHASrles river dummy variable (1 if census tract bounds the river; 0 otherwise)
X9 INDUS = Proportion of non-retail INDUStrial business acres per town
X10 TAX = Full value property TAX rate per $10,000
X11 PT = Pupil-Teacher ratio by town
RLZ = Proportion of Residential Land Zoned for lots over 25,000 sq.ft.

RÉPONSE
Y MV = Median Value of owner occupied-homes in $1000's

15 var X 506 obs

	1 CT	2 CRIM	3 NOX	4 AGE	5 DIS	6 RM	7 LSTAT	8 RAD	9 CHAS	10 INDUS	**11 TAX**	12 PT	13 MV	14 GROUP	15 RLZ
1	1	0,00632	0,538	65,2	4,0900	6,575	4,98	1	0	2,31	296	15,3	24,0	M	18,0
2	2	0,02731	0,469	78,9	4,9671	6,421	9,14	2	0	7,07	242	17,8	21,6	T	0,0
3	3	0,02729	0,469	61,1	4,9671	7,185	4,03	2	0	7,07	242	17,8	34,7	M	0,0
4	4	0,03237	0,458	45,8	6,0622	6,998	2,94	3	0	2,18	222	18,7	33,4	M	0,0
5	5	0,06905	0,458	54,2	6,0622	7,147	5,33	3	0	2,18	222	18,7	36,2	M	0,0
6	6	0,02985	0,458	58,7	6,0622	6,430	5,21	3	0	2,18	222	18,7	28,7	M	0,0
7	7	0,08829	0,524	66,6	5,5605	6,012	12,43	5	0	7,87	311	15,2	22,9	T	12,5
8	8	0,14455	0,524	96,1	5,9505	6,172	19,15	5	0	7,87	311	15,2	27,1	M	12,5

	1 CT	2 CRIM	3 NOX	4 AGE	5 DIS	6 RM	7 LSTAT	8 RAD	9 CHAS	10 INDUS	**11 TAX**	12 PT	13 MV	14 GROUP	15 RLZ
499	499	0,23912	0,585	65,3	2,4091	6,019	12,92	6	0	9,69	391	19,2	21,2	T	0,0
500	500	0,17783	0,585	73,5	2,3999	5,569	15,10	6	0	9,69	391	19,2	17,5	M	0,0
501	501	0,22438	0,585	79,7	2,4982	6,027	14,33	6	0	9,69	391	19,2	16,8	T	0,0
502	502	0,06263	0,573	69,1	2,4786	6,593	9,67	1	0	11,93	273	21,0	22,4	T	0,0
503	503	0,04527	0,573	76,7	2,2875	6,120	9,08	1	0	11,93	273	21,0	20,6	T	0,0
504	504	0,06076	0,573	91,0	2,1675	6,976	5,64	1	0	11,93	273	21,0	23,9	T	0,0
505	505	0,10959	0,573	89,3	2,3889	6,794	6,48	1	0	11,93	273	21,0	22,0	T	0,0
506	506	0,04741	0,573	80,8	2,5050	6,030	7,88	1	0	11,93	273	21,0	11,9	T	0,0

exemple : données expérimentales
suivi de personnes
diètes

5 var X 108 obs

sujet	Xgroupe	Xsemaine	YpertePoids	YestimeSoi
s1	control	1	4	14
s2	control	1	4	13
s3	control	1	4	17
s4	control	1	3	11
s5	control	1	5	16
s6	control	1	6	17
s7	control	1	6	17
s8	control	1	5	13
s9	control	1	4	15
s10	control	1	3	14
s11	control	1	4	16
s12	control	1	5	15
s13	diet	1	6	12
s14	diet	1	5	13
s15	diet	1	7	17
s16	diet	1	6	16
s17	diet	1	3	16
s18	diet	1	5	13
s19	diet	1	4	12
s20	diet	1	4	12
s21	diet	1	6	17
s22	diet	1	7	19
s23	diet	1	4	15
s24	diet	1	7	16
s25	diet+exer	1	8	16
s26	diet+exer	1	3	19
s27	diet+exer	1	7	15
s28	diet+exer	1	4	16
s29	diet+exer	1	9	13
s30	diet+exer	1	2	16

s30	diet+exer	3	1	17
s31	diet+exer	3	1	16
s32	diet+exer	3	2	18
s33	diet+exer	3	3	18
s34	diet+exer	3	2	17
s35	diet+exer	3	4	19
s36	diet+exer	3	1	17

MODÈLES d'analyse de variance – A, B facteurs catégoriques

Design classification simple : 1 facteur catégorique A

(1) $Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}$ $\quad i = 1, 2, \ldots, I$ (= nombre groupes) $\quad k = 1, 2, \ldots, n_i$

Y_{ik} : valeur de la variable de réponse Y j-ème essai modalité i du facteur A

μ : effet général - comme β_0 dans les modèles de régression

α_i : effet différentiel de la modalité i du facteur $\quad \sum_i \alpha_i = 0$

ε_{ik} : erreur aléatoire distribuée $\quad N(0, \sigma^2)$

Design classification double avec 2 facteurs A et B croisés

(2) $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}$ $\quad i = 1, 2, \ldots, I$ / $j = 1, 2, \ldots, J$ / $k = 1, 2, \ldots, n_{ij}$

α_i : effet principal de A $\quad \sum \alpha_i = 0$

β_j : effet principal de B $\quad \sum \beta_i = 0$

$(\alpha\beta)_{ij}$: effet d'interaction entre A et B $\quad \sum_i (\alpha\beta)_{ij} = 0 \quad \sum_j (\alpha\beta)_{ij} = 0$

Design classification double avec facteur B emboité dans facteur A

(3) $Y_{ijk} = \mu + \alpha_i + \beta_{j(i)} + \varepsilon_{ijk}$ $\quad i = 1, 2, .., I$ / $j = 1, 2, \ldots, J$ / $k = 1, 2, \ldots, n_{ij}$

α_i : effet facteur A et $\sum \alpha_i = 0$

$\beta_{j(i)}$: effet facteur B et $\sum \beta_{j(i)} = 0$ pour tout i

effet d'interaction A entre A et B n'existe pas

remarque: termes d'erreur ε_{ik} ε_{ijk} .... sont emboités dans la structure la plus fine (cellules) des données

on devrait écrire $\varepsilon_{k(i)}$ eq. (1) $\varepsilon_{k(ij)}$ eq, (3)

MODÈLES d'analyse de variance – facteurs catégoriques A, B

Design d'analyse de covariance : facteur catégorique A + facteur continu X
pente β égale chaque sous-groupes de A = pas d'interaction entre A et X

(4) $Y_{ik} = \mu + \alpha_i + \beta (X_{ik} - \overline{X}_{..}) + \varepsilon_{ik}$ $\quad X_{..} = \sum\sum X_{ik}$

$i = 1, 2, \ldots, I \quad k = 1, 2, \ldots, n_i$

Design d'analyse de covariance : pentes distinctes

pentes β_i distinctes chaque sous-groupes de A = interaction entre A et X

(5) $Y_{ik} = \mu + \alpha_i + \beta_i (X_{ik} - \overline{X}_{..}) + \varepsilon_{ik}$

si $\beta_i = \beta$ modèle (5) devient modèle (4)

Design mesures répétées par tous les sujets sur toutes modalités facteur A

(6) $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$ $\quad i = 1, 2, \ldots, s$ (nb sujets) / $j = 1, 2, \ldots, n_i$

μ : effet général

α_i : effet aléatoire du sujet i - indépendantes N $(0. \sigma_\rho^2)$

β_j : effet de la modalité j du facteur fixe A et $\sum \tau_j = 0$

ε_{ij} : erreur aléatoire N $(0, \sigma^2)$

$\alpha_i, \varepsilon_{ij}$ indépendantes

remarque : exemple d'un modèle mixte - nature différentes des facteurs
facteur 1 = sujet = nature aléatoire
facteur 2 = A = nature fixe (modalités contrôlées)

MODÈLES d'analyse de variance – facteurs catégoriques A, B

Design Split-Plot (parcelles divisées) : 2 tailles d'unités exp. - 2 termes d'erreur

(7) $Y_{ijk} = \mu + \underbrace{\alpha_i + \beta_j + (\alpha\beta)_{ij}}_{\text{plot (parcelle)}} + \underbrace{\gamma_k + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\gamma\beta)_{ijk}}_{\text{subplot (sous parcelle)}} + \underbrace{\varepsilon_{ijk}}_{\text{erreur exp.}}$

$i = 1, 2, ..., r$ (facteur bloc - répétition) $j = 1, 2, ..., I$ (facteur A) $k = 1, 2, ..., J$ (facteur B)

α_i : effet facteur bloc - facteur aléatoire distribué $N(0, \sigma_\alpha^2)$

β_j : effet principal facteur A = facteur plot

$(\alpha\beta)_{ij}$: erreur plot = interaction = bloc x A

γ_k : effet principal facteur B = facteur subplot

$(\alpha\gamma)_{ik}$: interaction = bloc x B

$(\beta\gamma)_{jk}$: interaction A x B

$(\alpha\gamma\beta)_{ijk}$: erreur subplot = interaction = bloc x AB

ε_{ijk} : erreur expérimentale distribution $N(0, \sigma^2)$

version simplifiée de (7) $(\alpha\gamma)_{ik}$ et $(\alpha\gamma\beta)_{ijk}$ négligeables et incoporées avec ε_{ijk}

(8) $Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_k + (\beta\gamma)_{jk} + \varepsilon_{ijk}$