**Big data : Comprendre big data et ces enjeux**

**Introduction**

**Le plan du cours**

**Installation de R:** <http://www.ics.uci.edu/~jutts/110/InstallingRandRStudio.pdf>

**Exemple de Wal-Mart (**million de transactions client par heure**) :**

1. **Les enjeux du bigdata**

* Augmentation de la capacité de stockage
* Méga données, données massives difficile à gérer avec les outils classiques de base de données
* Analyse prédictive (Données climatiques, la bourse, l’économie, la science, l’assurance des biens et des personnes, la sécurité…)
* Dimension (les 3V)

**Volume :**les données numériques créées dans le monde seraient passées de 1,2 [zettaoctets](https://fr.wikipedia.org/wiki/Zettaoctet) (1021 [octets](https://fr.wiktionary.org/wiki/octets)) par an en 2010 à 1,8 [zettaoctets](https://fr.wikipedia.org/wiki/Zettaoctet) en 2011[37](https://fr.wikipedia.org/wiki/Big_data#cite_note-37), puis 2,8 [zettaoctets](https://fr.wikipedia.org/wiki/Zettaoctet) en 2012 et s'élèveront à 40 [zettaoctets](https://fr.wikipedia.org/wiki/Zettaoctet) en 2020

**Variété :**les données non-structurées rendent utilisation des outils traditionnels très difficile

**Vélocité :**fréquence à laquelle les données sont générées, capturées, partagées et mises à jour

* Trouver un modèle mathématique dans les données

1. **Notions de mathématiques**
2. Equations de droites

• Elles admettent une équation de la forme y = mx+ p.

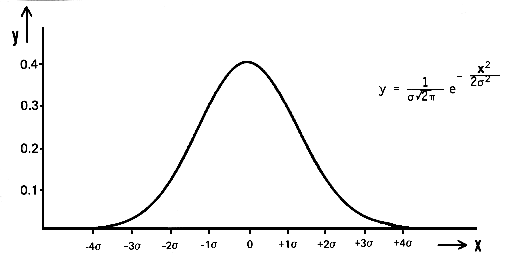
m: est le coefficient directeur et

p: est l’ordonnée à l’origine.

Dire qu’un point A appartient à la droite d’équation y = mx+ p signifie que ses coordonnées vérifient l’équation, c’est à dire que yA = mxA + p

2) Les matrices (chap 6 page149)

3) Les distributions (Normale)

Exemple : Tir à l’arc, heure d’arrivée

* Les calculs mathématiques sont plus simples
* Les évènements de la vie courante, en science ont une distribution normale

4) Les régressions

* Régression linéaire simple

Le résidu suit une loi normale.

Y : variable dépendante

X : variable indépendante

On veut trouver la droite d’équation

qui représente le mieux les nuages. Nous mesurons l’écart entre le nuage et la droite par la somme Q des carrés des distances verticales entre les points du nuage et la droite.

La droite des moindres carrés est celle qui minimise Q.

* Régression multiple

+…++

En écriture matricielle; y=X

* GLM (Generalized linear model)
* Les modèles de série temporelle

1. **Big data : approches**

* Collecte des données (continues, catégoriales,)
* Nettoyage et préparation des données (division des données pour l’estimation et validation des données, normalisation …)
* Analyse des données
* Estimation des paramètres du modèle
* Evaluation du modèle (Underfit and overfit, RootMeanSquaredError RMSE, R^2, analyse du résidu)
* Utilisation du modèle

1. **Comprendre et analyser les données : Boston housing data**

* <https://archive.ics.uci.edu/ml/machine-learning-databases/housing/>
* Descriptions des variables

1. CRIM: per capita crime rate by town   
2. ZN: proportion of residential land zoned for lots over 25,000 sq.ft.   
3. INDUS: proportion of non-retail business acres per town   
4. CHAS: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)   
5. NOX: nitric oxides concentration (parts per 10 million)   
6. RM: average number of rooms per dwelling   
7. AGE: proportion of owner-occupied units built prior to 1940   
8. DIS: weighted distances to five Boston employment centres   
9. RAD: index of accessibility to radial highways   
10. TAX: full-value property-tax rate per $10,000   
11. PTRATIO: pupil-teacher ratio by town   
12. B: 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town   
13. LSTAT: % lower status of the population   
14. MEDV: Median value of owner-occupied homes in $1000's

**Exemples:**

housing<-read.csv2(choose.files(), header = TRUE)

fix (housing)

str(housing)

summary(housing)

boxplot(housing$AGE)

with(housing, boxplot(AGE))

with(housing, boxplot(CRIM,INDUS))

with(housing, plot(AGE, TAX, xlab="AGE",ylab="TAX"))

with (housing, hist(RM)) loi normale

with (housing, hist(AGE)) plus de maisons agées

with(housing, hist(DIS)) loi lognormale

with(housing, hist(RAD)) TAX et RAD se ressemblent beaucoup

with(housing, hist(LSTAT))la loi lognormale

pairs(housing[,c("RM","LSTAT","MEDV")])

Test de normalité

result <- with(housing, shapiro.test(RM))

print (result)

Régression linéaire

mod1<- lm(MEDV ~RM, data = housing)

summary(mod1)

newdata1<-data.frame(RM=7)

predicted.values<-predict(mod1,newdata1)

print(predicted.values)

mod2<- lm(MEDV ~RM+AGE, data = housing)

summary(mod2)

newdata2<-data.frame(RM=7, AGE=60)

predicted.values<-predict(mod2,newdata2)

print(predicted.values)

1. **Comment choisir un bon modèle (*A compléter*)**
2. **Conclusion et pour aller plus loin (*A compléter*)**

* Besoin de bâtir de nouvelles infrastructures
* Trouver un modèle mathématique dans les données
* Droit à la vie privée ([Orwell](https://fr.wikipedia.org/wiki/George_Orwell), [Edward Snowden](https://fr.wikipedia.org/wiki/Edward_Snowden) )
* Mise en place d’une régulation forte pour nous protéger des dérives
* Opportunité d’emploi considérable dans les prochaines années