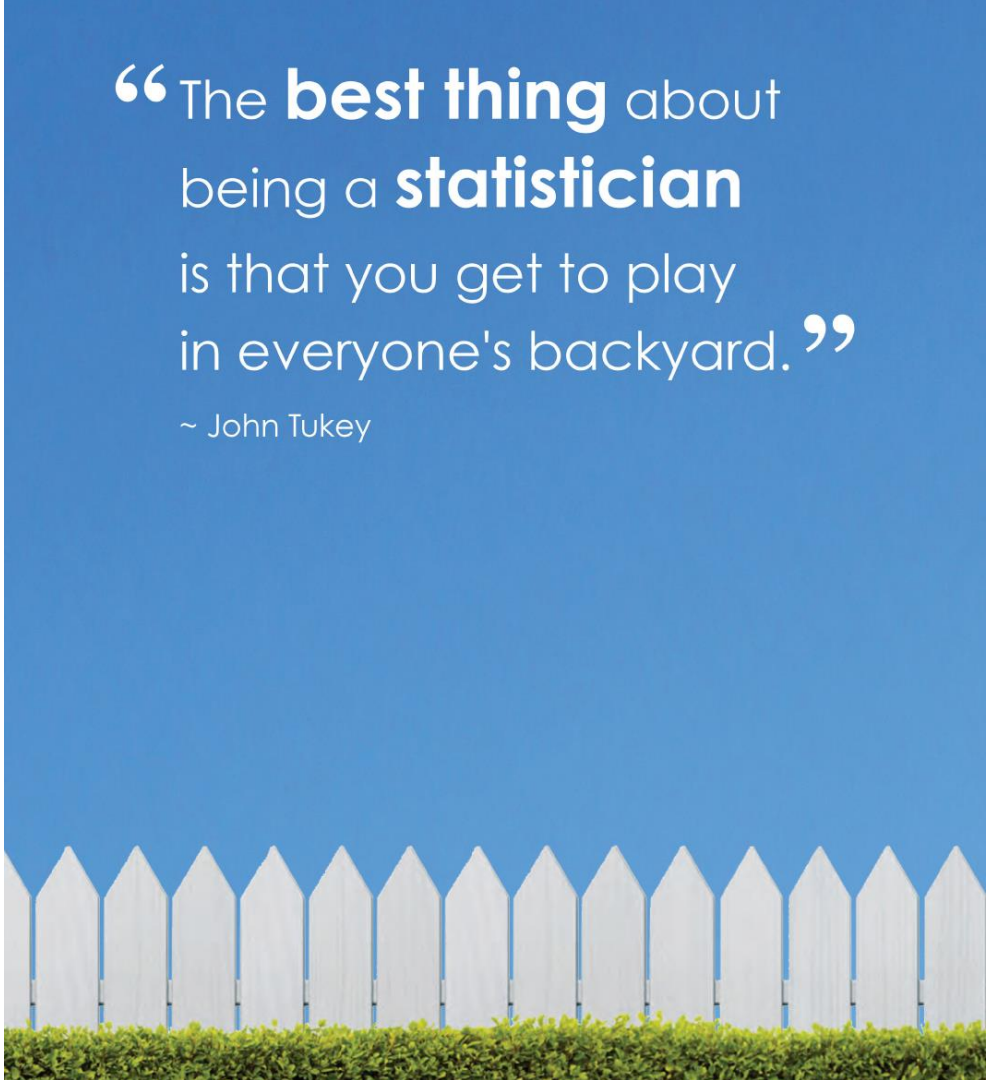
UBISOFT

JEUX VIDÉO, TECHNOLOGIE, SCIENCES DES DONNÉES... ET LA STATISTIQUE DANS TOUT ÇA?

Nicolas Grenon-Godbout


Partie I
MON PARCOURS

LA PREMIÈRE ÉTINCELLE:



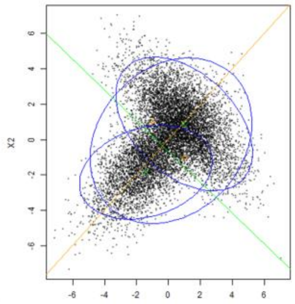
LA 2E

```
int testVic(int tableau[][SIZE])
{
    int i,j,k;
    int test=tableau[0][0];
    if (test)
    {
        for(i=1,j=1;i<SIZE;i++,j++)
            if(tableau[i][j]!=test)
            {
                test=0;
                break;
            }
        if(test) return test;
    }

    test=tableau[SIZE-1][0];
    if (test)
    {
        for(i=SIZE-2,j=1;j<SIZE;i--,j++)
            if(tableau[i][j]!=test)
            {
                test=0;
                break;
            }
        if(test) return test;
    }

    for(k=0;k<SIZE;k++)
    {
        if(test=tableau[k][k])
        {
            for(i=0;i<SIZE;i++)
                if(tableau[i][k]!=test)
                {
                    test=0;
                    break;
                }
            if(test) return test;
```

LA 3^{E}




```r
mcmc_adReg = function(rr = 1,n,t0,k,init, cible, paramCible, C,C1,C2, beta, a,b,kMoy,
{
  #calcul dimensions et tests
  if(k==1) d = length(init)
  else d = dim(init)[1]
  if(!is.matrix(C)) stop("C not a matrix")
  if(dim(C)[1] !=  dim(C)[2] ) stop("C is not square")
  if(dim(C)[1] != d) stop("dimension of C does not match the initial chain value")
  #ajustement au nb de chaines
  t0 = round(t0/k) + 1
  n = round(n/k)
  ###########
  #conteneurs
  x = array(0,c(d,k,n+1))
  cumJump = rep(0,n - t0 + 1)
  if(k==1) x[,,1] = init
  else x[,,1] = init
  #diagnostics : emplacements etc
  region=matrix(as.integer(2),k,n+1)
  regionsR = matrix(as.integer(2),k,n+1)
  regionCand = matrix(as.integer(2),k,n)
  #choix de proposition(1,2 ou 0 si proposition globale)
  prop=matrix(as.integer(3),k,n)
  #poids adaptatifs initiaux
  lambda1 = 0.5
  lambda2 = lambda1
```

Divide and Conquer: A Mixture-Based Approach to Regional Adaptation for MCMC

Yan BAI, Radu V. CRAIU, and Antonio F. DI NARZO

The efficiency of Markov chain Monte Carlo (MCMC) algorithms can vary dramatically with the choice of simulation parameters. Adaptive MCMC (AMCMC) algorithms allow the automatic tuning of the parameters while the simulation is in progress. A multimodal target distribution may call for regional adaptation of Metropolis–Hastings samplers so that the proposal distribution varies across regions in the sample space. Establishing such a partition is not straightforward and, in many instances, the learning required for its specification takes place gradually, as the simulation proceeds. In the case in which the target distribution is approximated by a mixture of Gaussians, we propose an adaptation process for the partition. It involves fitting the mixture using the available samples via an online EM algorithm and, based on the current mixture parameters, constructing the *regional adaptive algorithm with online recursion* (RAPTOR). The method is compared with other regional AMCMC samplers and is tested on simulated as well as real data examples.

$$
\begin{aligned}
&= \sup_x \sup_A \left| \int_A \left(P_{\Gamma_{t+1}}(x,y) - P_{\Gamma_t}(x,y) \right) \mathrm{d}y \right| \\
&= \sup_x \sup_A \Bigg| \int_A f_{\Gamma_{t+1}}(x,y)\alpha_{t+1}(x,y)\mathrm{d}y - \int_A f_{\Gamma_t}(x,y)\alpha_t(x,y)\mathrm{d}y \\
&\quad + \mathbb{I}(x \in A) \left(\int_S f_{\Gamma_{t+1}}(x,y)(1-\alpha_{t+1}(x,y))\mathrm{d}y - \int_S f_{\Gamma_t}(x,y)(1-\alpha_t(x,y))\mathrm{d}y \right) \Bigg| \\
&\leq \sup_x \sup_A \int_A \left| f_{\Gamma_{t+1}}(x,y)\alpha_{t+1}(x,y) - f_{\Gamma_t}(x,y)\alpha_t(x,y) \right| \mathrm{d}y \\
&\quad + \mathbb{I}(x \in A) \int_S \left| f_{\Gamma_{t+1}}(x,y)(1-\alpha_{t+1}(x,y)) - f_{\Gamma_t}(x,y)(1-\alpha_t(x,y)) \right| \mathrm{d}y \\
&\leq \sup_x \int_S \left| f_{\Gamma_{t+1}}(x,y)\alpha_{t+1}(x,y) - f_{\Gamma_t}(x,y)\alpha_t(x,y) \right| \mathrm{d}y \\
&\quad + \int_S \left| f_{\Gamma_{t+1}}(x,y)(1-\alpha_{t+1}(x,y)) - f_{\Gamma_t}(x,y)(1-\alpha_t(x,y)) \right| \mathrm{d}y
\end{aligned}
$$

UN APPEL D'UBI...


hadoop


HIVE


dataflow


cloudera
IMPALA


cassandra


APACHE
Spark
TM


Apache
Airflow
databricks
snowflake
aws

CONCLUSION

LES BASES

Les concepts mathématiques et statistiques fondamentaux sont à la base de l'édifice de la science des données. Une bonne formation dans ces domaines ne ferme aucune porte, au contraire.

LES TECHNOLOGIES

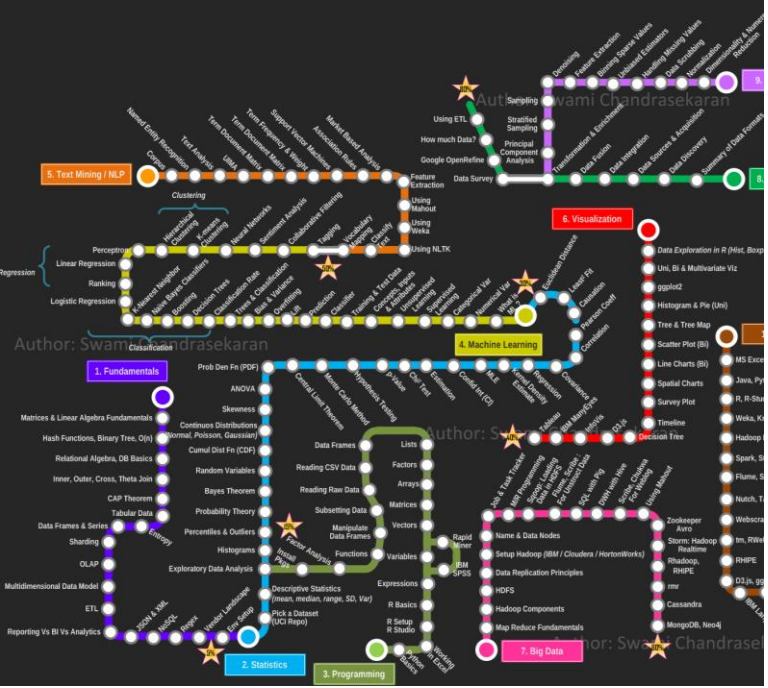
Il ne faut pas tout connaître d'entrée de jeu, seulement avoir le potentiel de pouvoir y arriver. Reste que des acquis conceptuels et pratiques en informatique sont un gros avantage pour savoir s'adapter.

LE PLUS IMPORTANT

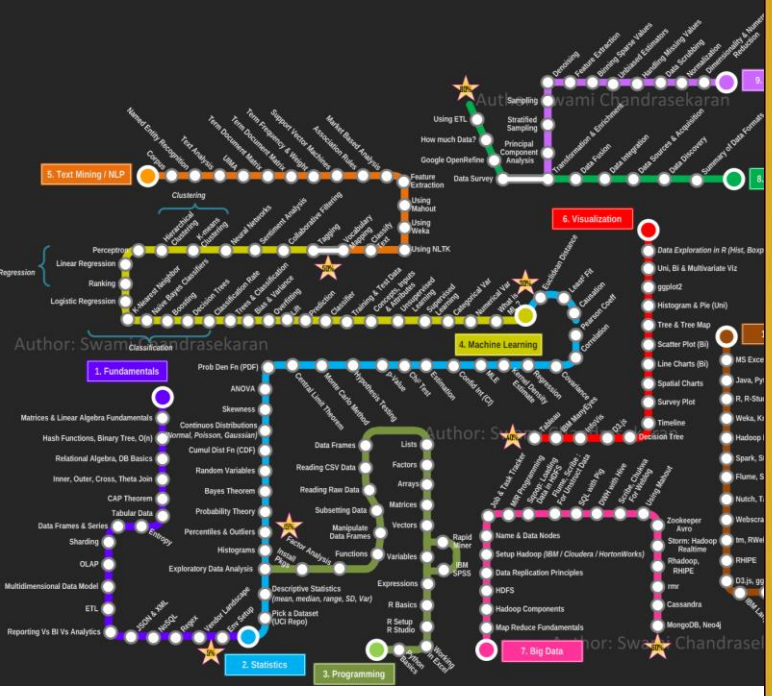
Développer une capacité d'apprendre, de autonomie et une curiosité sans borne, car l'université n'est que le début de notre formation.

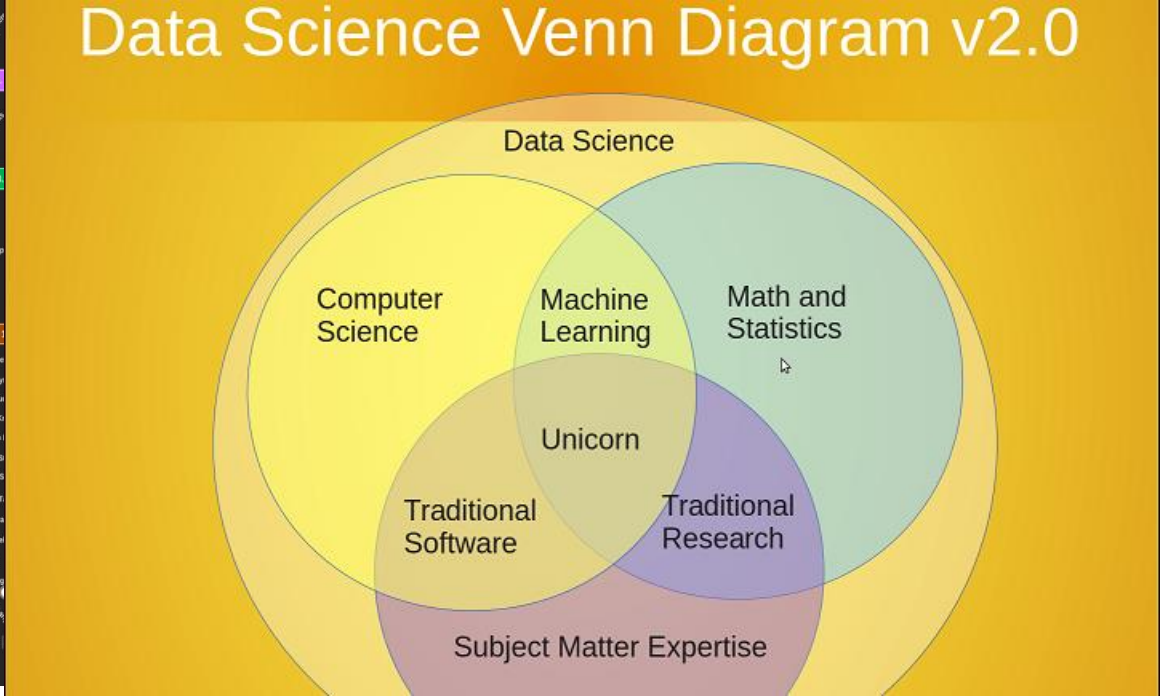

Partie II
DS @ UBISOFT

Science des données ?

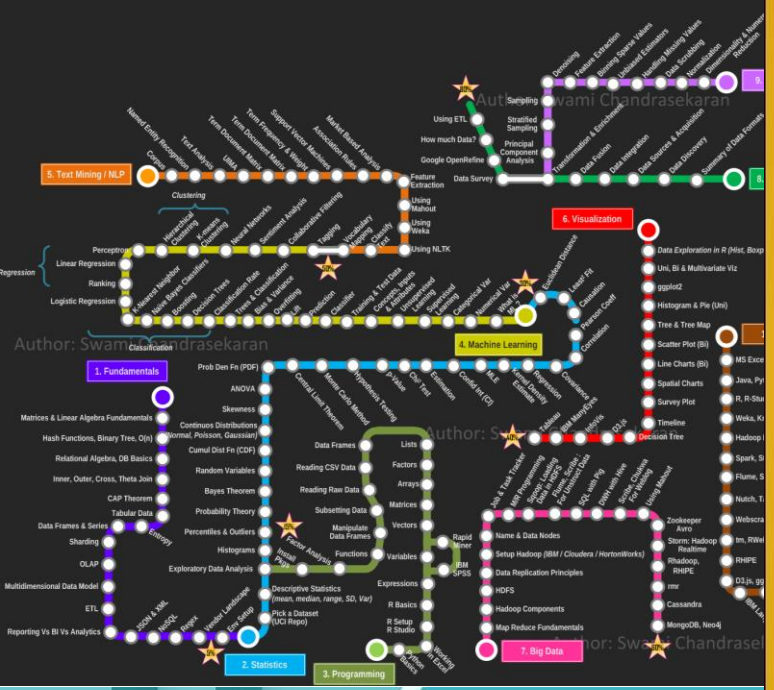


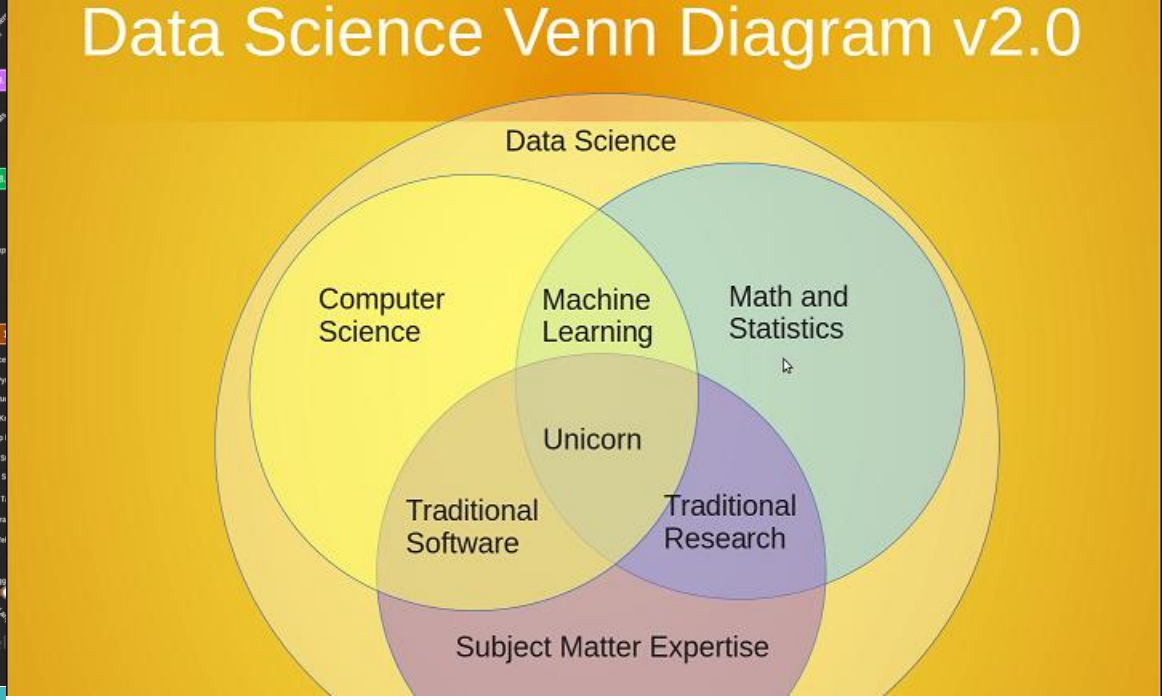
Science des données ?

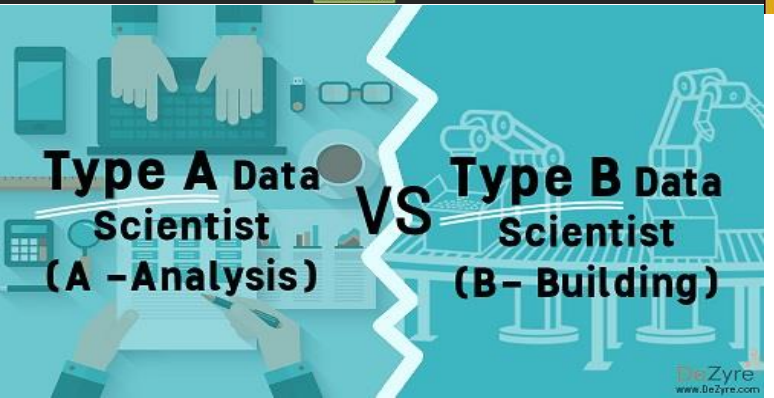




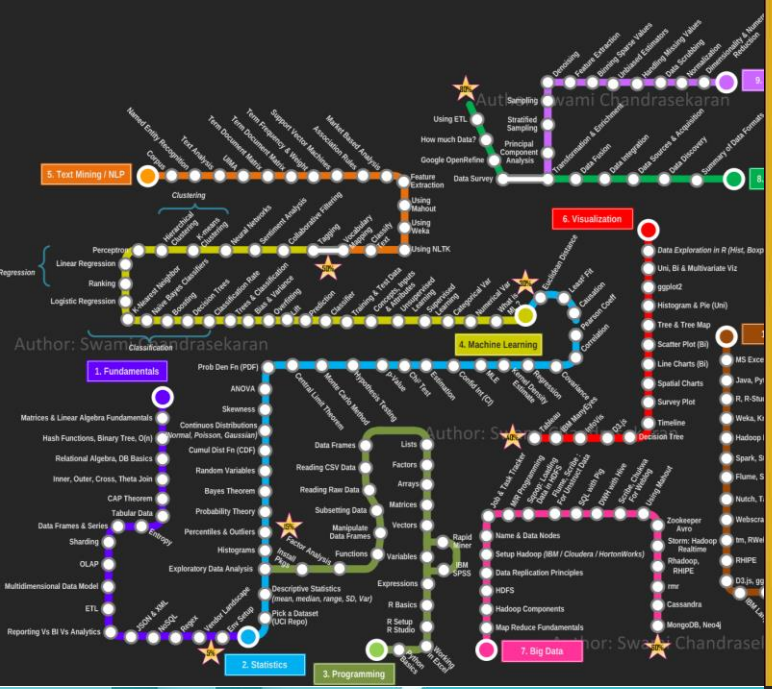
Science des données ?

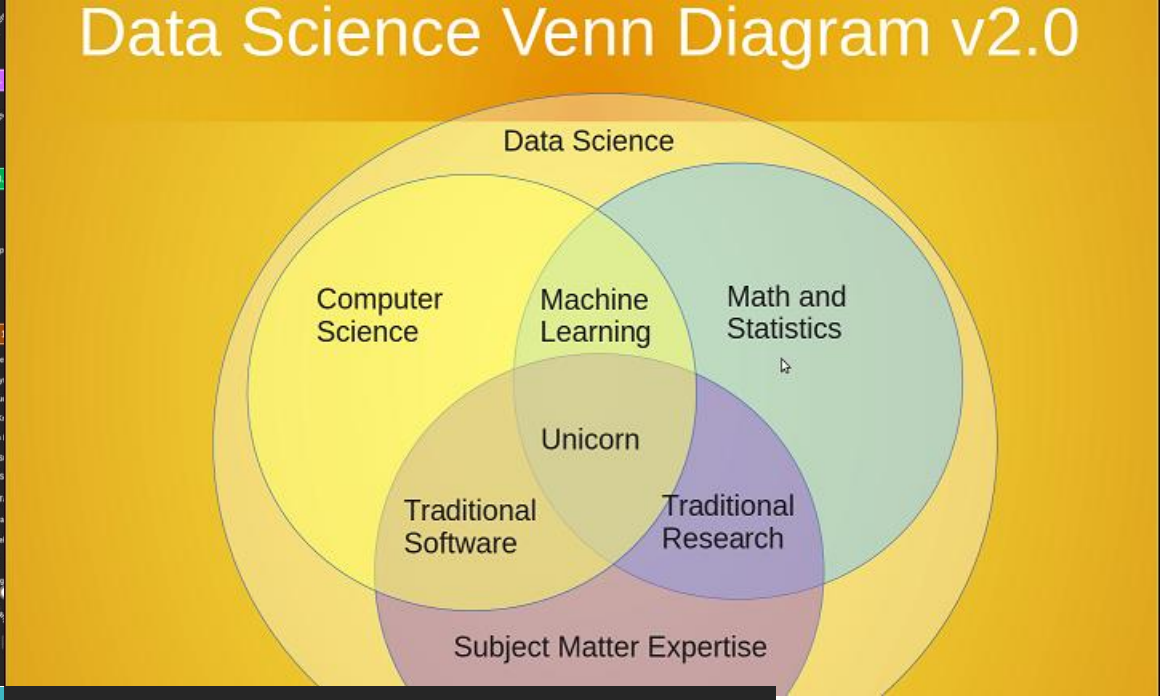


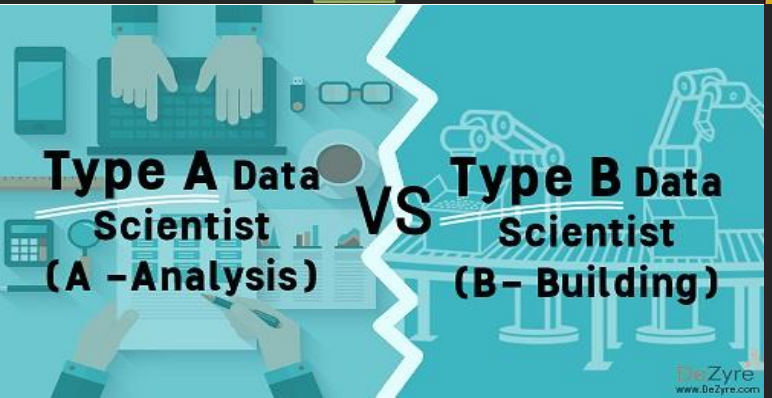




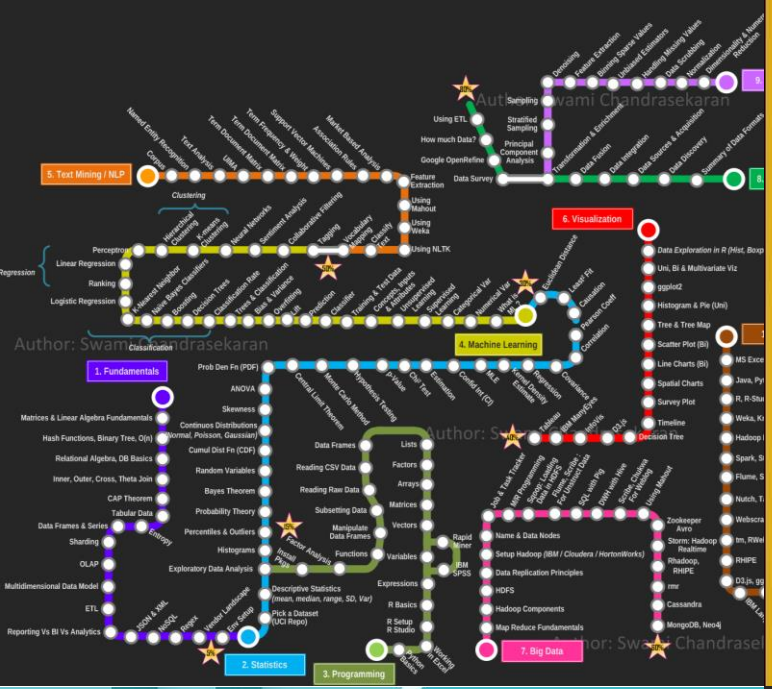
Science des données ?

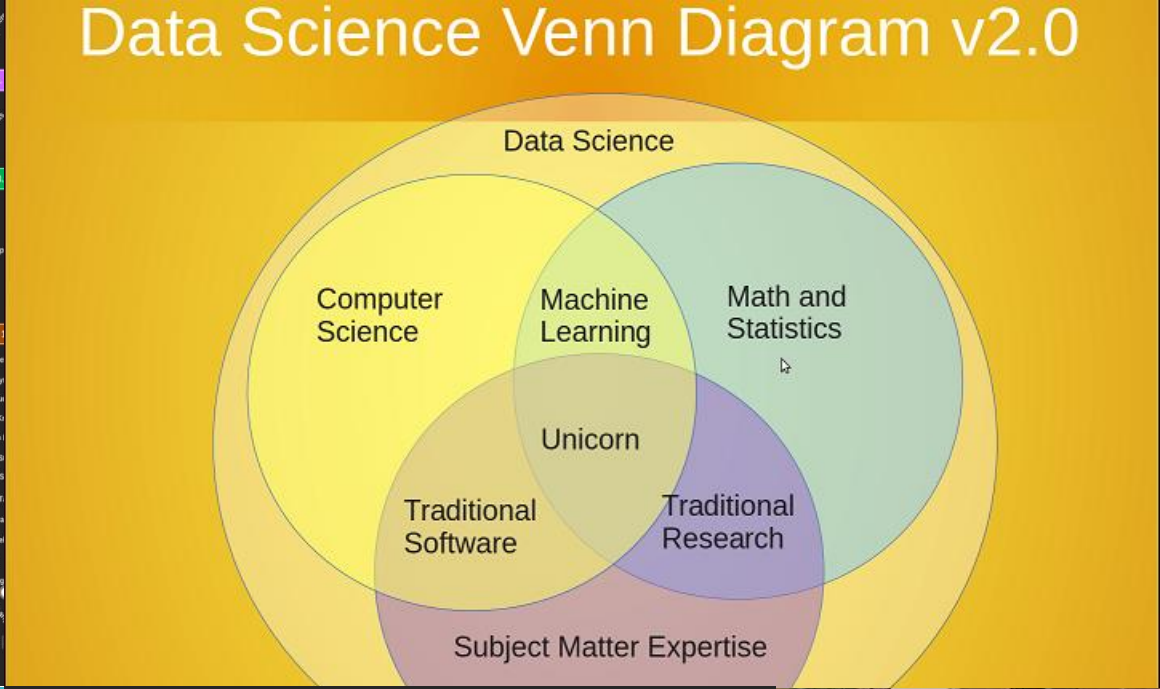

5. Text Mining / NLP
6. Visualization
4. Machine Learning
1. Fundamentals
2. Statistics
3. Programming
7. Big Data
Author: Swami Chandrasekaran

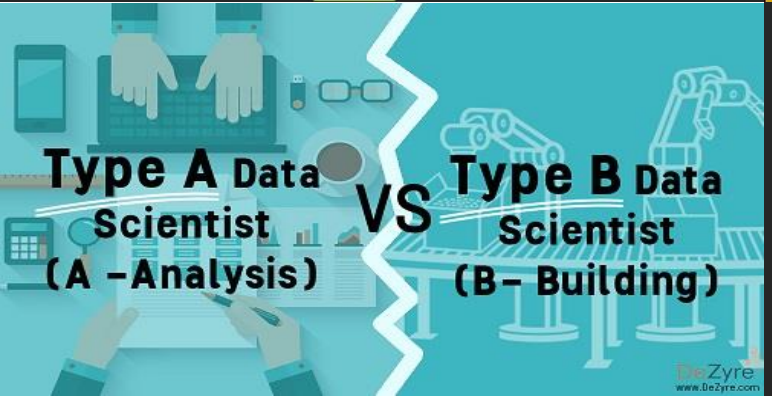

Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise


Type A Data Scientist (A -Analysis)
VS
Type B Data Scientist (B- Building)
DeZyre
www.DeZyre.com

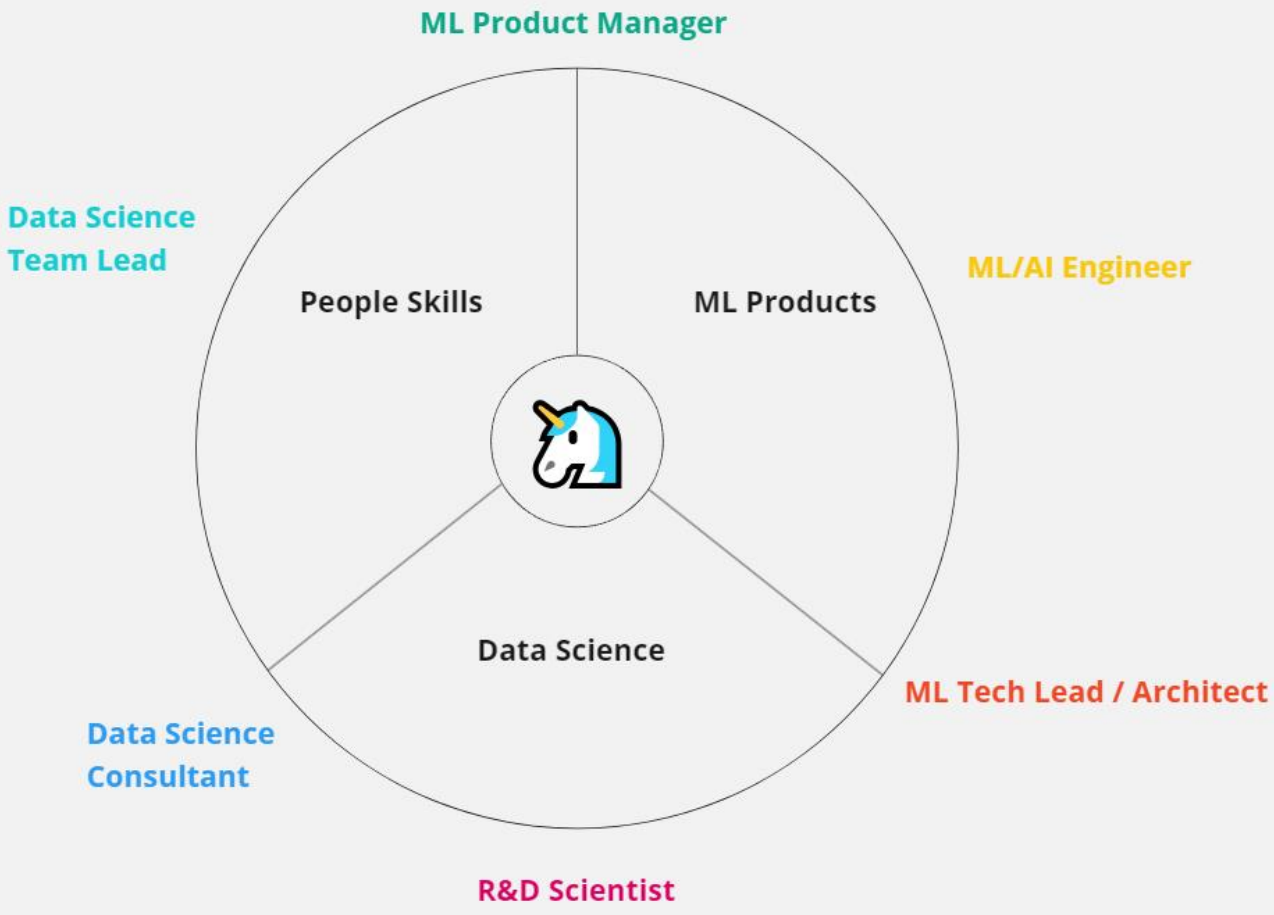

Data Science Unicorn


Author: Swami Chandrasekaran
5. Text Mining / NLP
6. Visualization
4. Machine Learning
1. Fundamentals
2. Statistics
3. Programming
7. Big Data


Data Science Venn Diagram v2.0
Data Science
Computer Science
Machine Learning
Math and Statistics
Unicorn
Traditional Software
Traditional Research
Subject Matter Expertise


Type A Data Scientist (A -Analysis)
VS
Type B Data Scientist (B- Building)
DeZyre
www.DeZyre.com


Data Science Unicorn


ML Product Manager
Data Science Team Lead
ML/AI Engineer
People Skills
ML Products
Data Science
ML Tech Lead / Architect
Data Science Consultant
R&D Scientist

ROLLER CHAMPIONS

ASSASSIN'S CREED

FAR CRY

FOR HONOR

RAINBOW SIX

WATCH DOGS

CHILD OF LIGHT

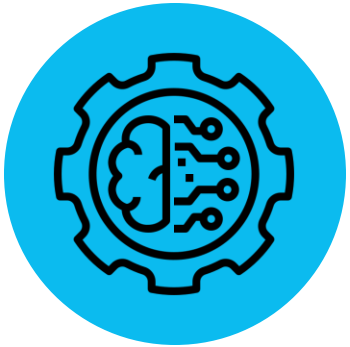
SPLINTER CELL

PRINCE OF PERSIA

WHAT WE DO

GAME INSIGHTS

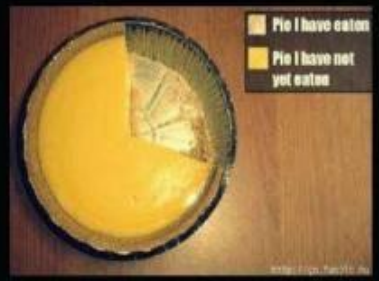
PRODUCTION ASSISTANCE



LIVE ML APPLICATIONS

STATISTICIAN

What my friends think I do



What my mom thinks I do

What society thinks I do

What my boss thinks I do

What I think I do

What I actually do

STAT=?DS

One doesn't need to look far to see click-bait capitalizing on the befuddlement about this new state of affairs:

- *Why Do We Need Data Science When We've Had Statistics for Centuries?*
  Irving Wladawsky-Berger
  Wall Street Journal, CIO report, May 2, 2014

- *Data Science* **is** *statistics.*
  `When physicists do mathematics, they don't say they're doing number science. They're doing math. If you're analyzing data, you're doing statistics. You can call it data science or informatics or analytics or whatever, but it's still statistics. ... You may not like what some statisticians do. You may feel they don't share your values. They may embarrass you. But that shouldn't lead us to abandon the term ''statistics''.`
  Karl Broman, Univ. Wisconsin[6]

On the other hand, we can find pointed comments about the (near-) irrelevance of statistics:

- *Data Science without statistics is possible, even desirable.*
  Vincent Granville, at the Data Science Central Blog[7]

- *Statistics is the least important part of data science.*
  Andrew Gelman, Columbia University [8]

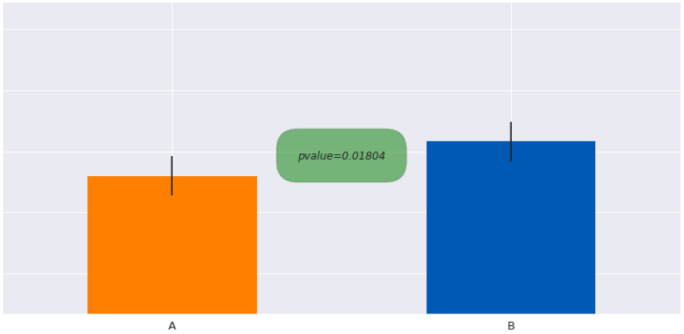
LA STAT DANS LA DS

CONSULTATION

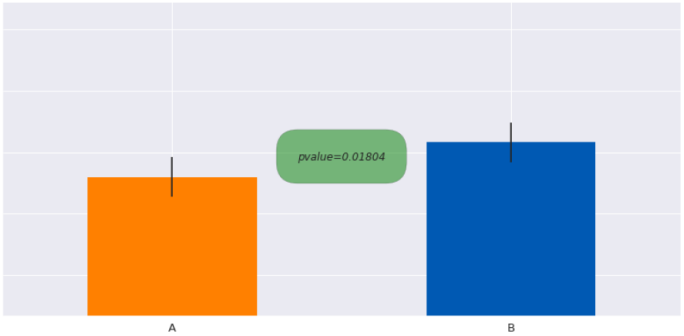
LES PETITS +

LES APPLICATIONS FONDAMENTALES

ÊTRE STATEU.SE DE SERVICE

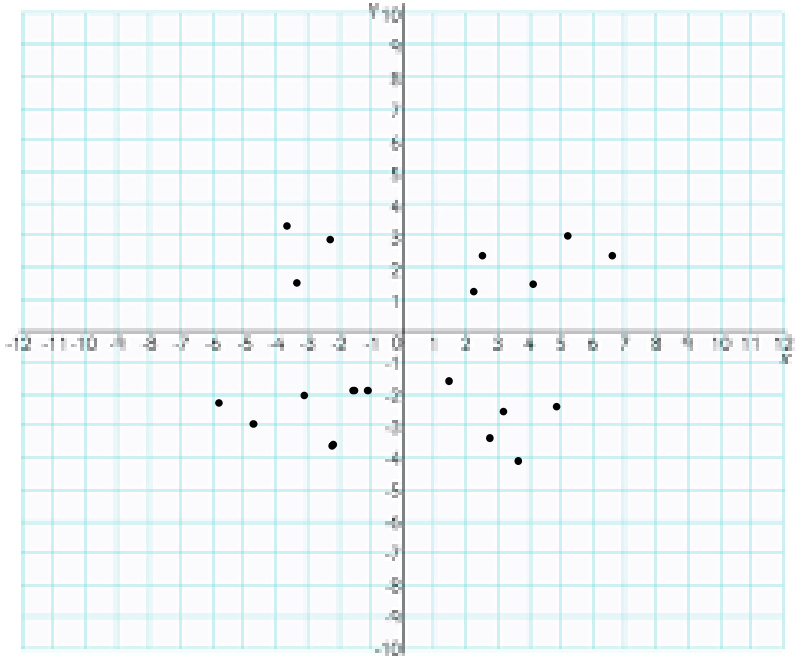


ÊTRE STATEU.SE DE SERVICE

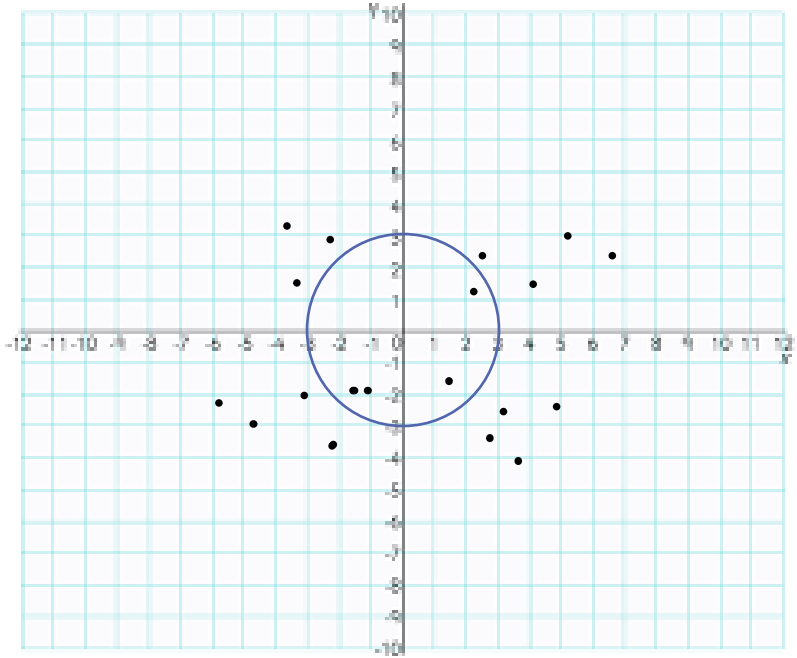


0.37701159	0.070547117	0.326791863	0.771936443	1.11E-62	0.018040747
0.033413119	0.130700809	0.19883805	0.993519492	8.87E-39	0.002261022
0.491265754	0.23050403	0.11990524	0.071253792	1.80E-21	0.000250067
0.461195004	0.609647209	0.480817436	0.046986079	0.000139121	0.013004543

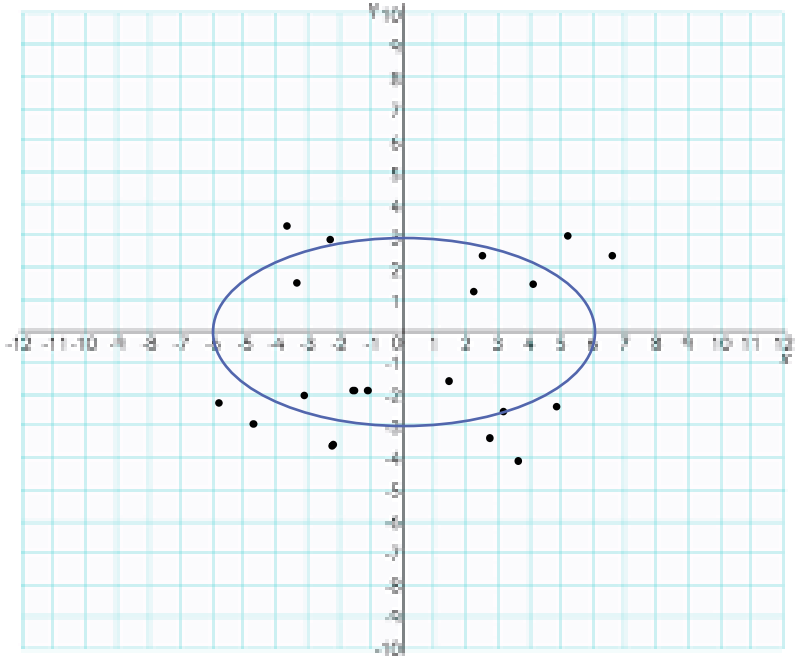
ÊTRE MATHEU.SE DE SERVICE (X2)



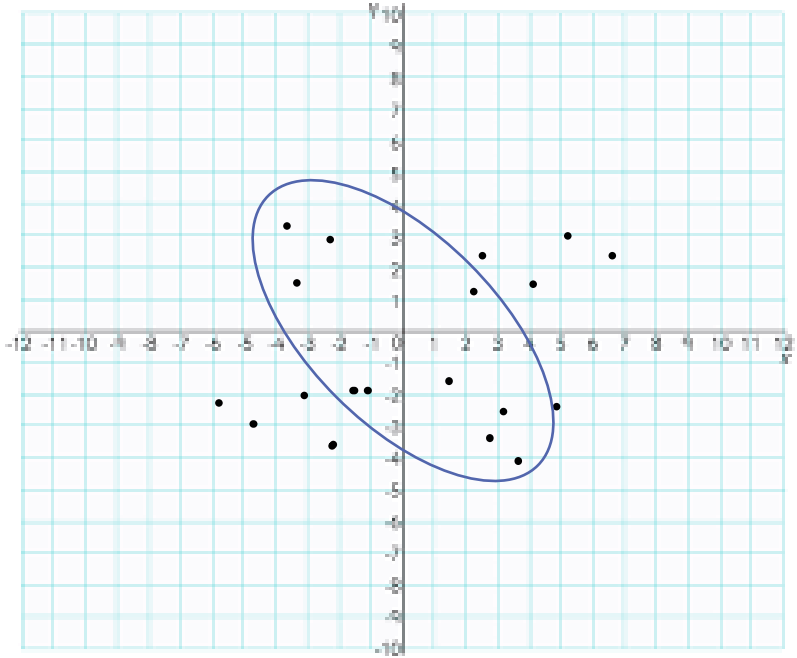
ÊTRE MATHEU.SE DE SERVICE (X2)



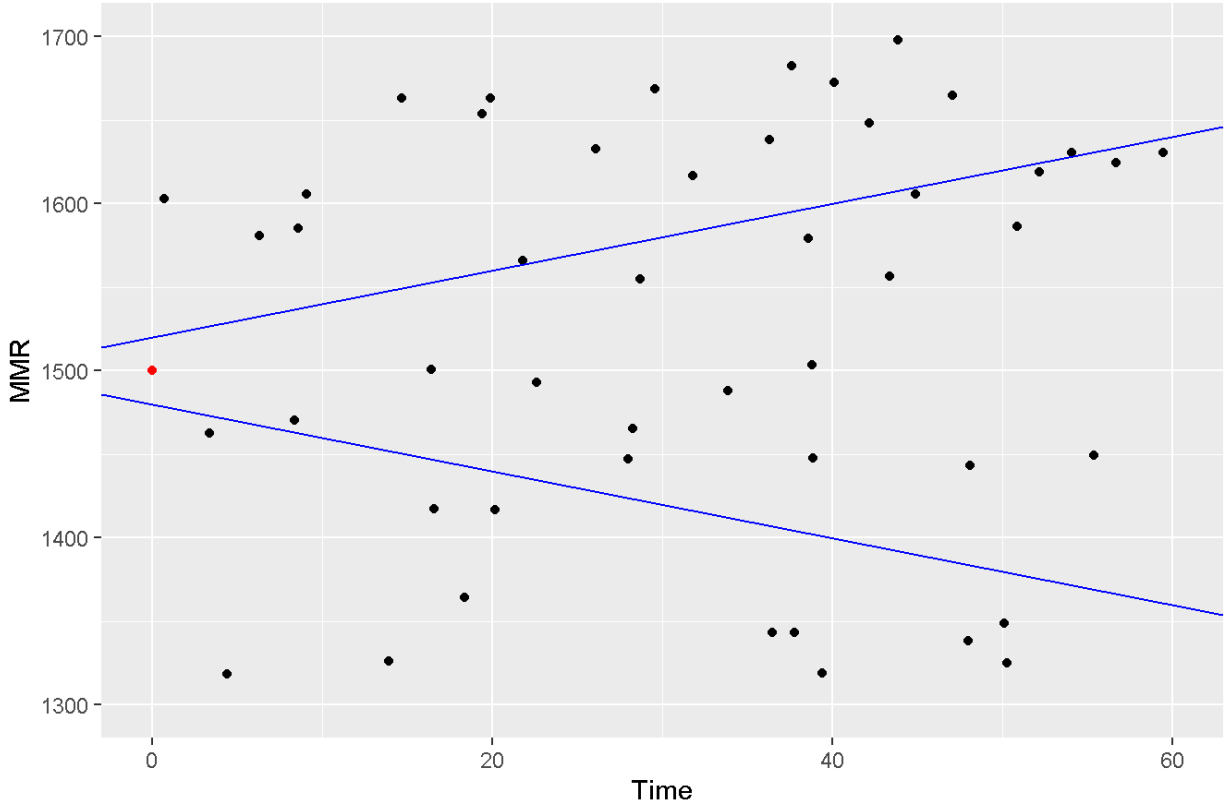
ÊTRE MATHEU.SE DE SERVICE (X2)



ÊTRE MATHEU.SE DE SERVICE (X2)

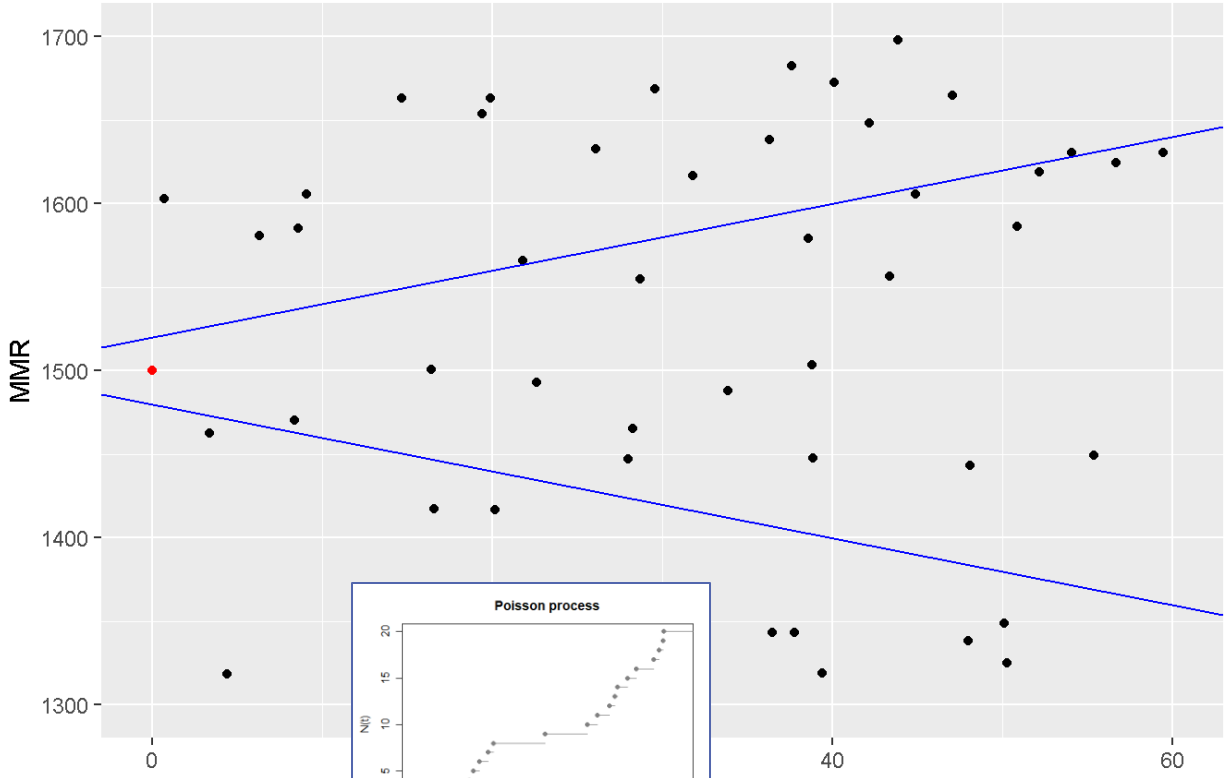


LES PETITS +: MATCHMAKING



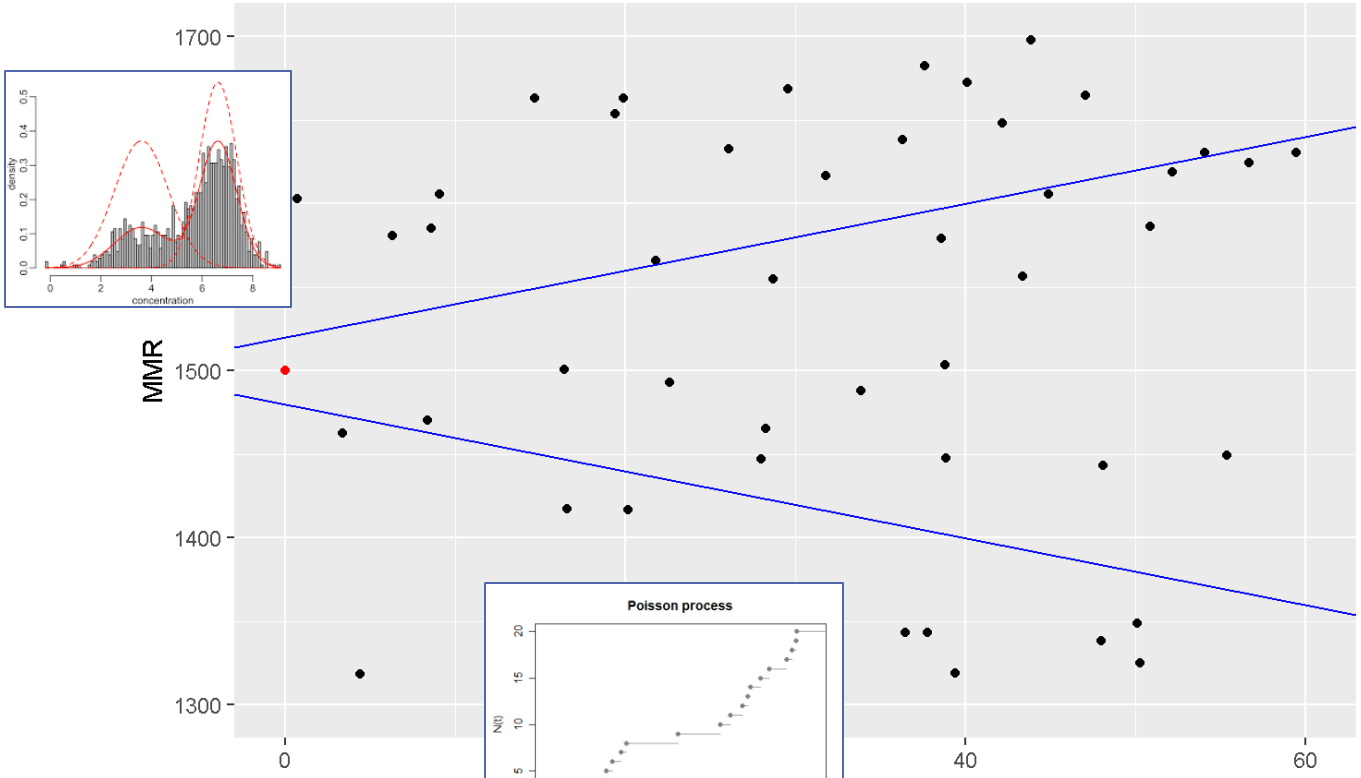
High activity, MMR limit in blue: because of a relatively rapid MMR limit expansion, the red player quickly found an acceptable match, but with a stricter limit, he would have found a closer match some seconds later.

LES PETITS +: MATCHMAKING



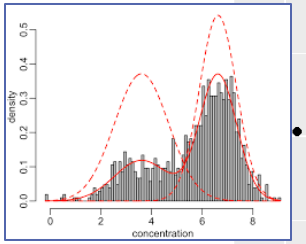
High activity, MMR limit in b limit expansion, the red player quickly found an acceptable match, but with oser match some seconds later.

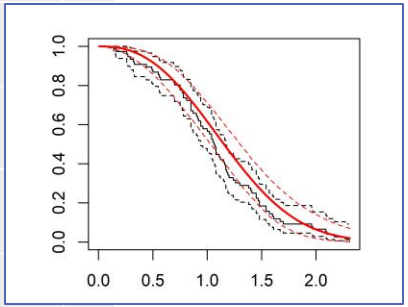
LES PETITS +: MATCHMAKING



High activity, MMR limit in b limit expansion, the red player quickly found an acceptable match, but with oser match some seconds later.

LES PETITS +: MATCHMAKING





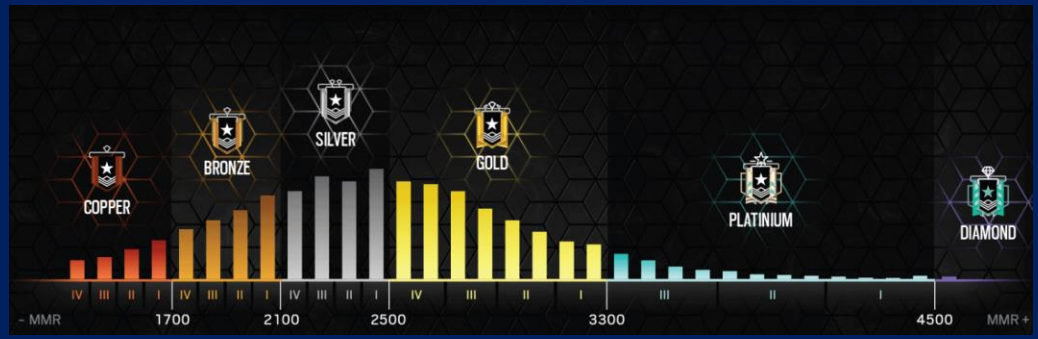
High activity, MMR limit in b limit expansion, the red player quickly found an acceptable match, but with oser match some seconds later.

LES APPLICATIONS FONDAMENTALES: SKILL RATING

Goal : Assign a **rating** to each player in a population, based on **matches' outcome**.

Objective : The rating should represent their relative **proficiency at winning games.**

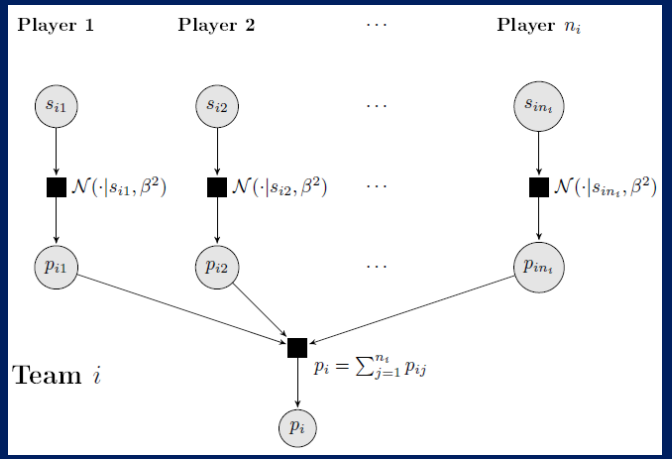
Usage :
- Provide an **overall ranking** of the players.
- Use the rating to **match players** of similar skill together.

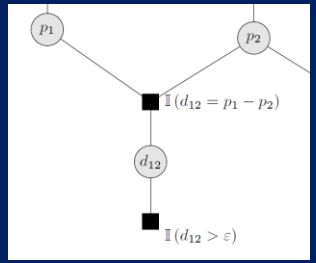


Ranking Model Assumptions

In a population of n players

1) Each player has a skill value $s \in \boldsymbol{R}$.
2) For a given match, player j in team i has a random performance value p_{ij} based on s_{ij}. Usually : $p_{ij} \sim N(s_{ij}, \beta^2)$.
3) A player's skill value contributes additively to its team's performance: $p_j = \sum_{j=1}^{n_i} p_{ij}$.
4) The team with the highest performance is ranked highest.
   So that in a Two-teams match, team 1 wins with probability
   $$\mathbb{P}(p_1 > p_2).$$





Skill Rating Model

Rating Model = Ranking Model + Inference mechanism

- Most models suppose random normal skill values : $S_{ij} \sim N(\mu_{ij}, \sigma_{ij}^2)$
- Inference is made on $(\mu_{ij}, \sigma_{ij}^2)$
- Bayesian inference :
  - θ : Collection of all skills
  - x : Match(es) outcome(s)
  - Compute $\pi(\theta|x) \propto p(x|\theta)\pi_0(\theta)$
  - $\pi_0(\theta)$: Prior skill distribution
  - $p(x|\theta)$: Ranking model

3.1.1 For $q = 1, \ldots, k; q \neq i$,

$$\delta_q = \frac{\sigma_i^2}{c_{iq}} \times \begin{cases} V\left(\frac{\mu_i - \mu_q}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) > r(i), \\ \tilde{V}\left(\frac{\mu_i - \mu_q}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) = r(i), \\ -V\left(\frac{\mu_q - \mu_i}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) < r(i), \end{cases}$$

$$\eta_q = \left(\frac{\sigma_i}{c_{iq}}\right)^2 \times \begin{cases} W\left(\frac{\mu_i - \mu_q}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) > r(i), \\ \tilde{W}\left(\frac{\mu_i - \mu_q}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) = r(i), \\ W\left(\frac{\mu_q - \mu_i}{c_{iq}}, \frac{\varepsilon}{c_{iq}}\right) & \text{if } r(q) < r(i), \end{cases}$$

where

$$c_{iq} = (\sigma_i^2 + \sigma_q^2 + 2\beta^2)^{1/2},$$

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad \Phi(x) = \int_{-\infty}^{x} \phi(u)du,$$

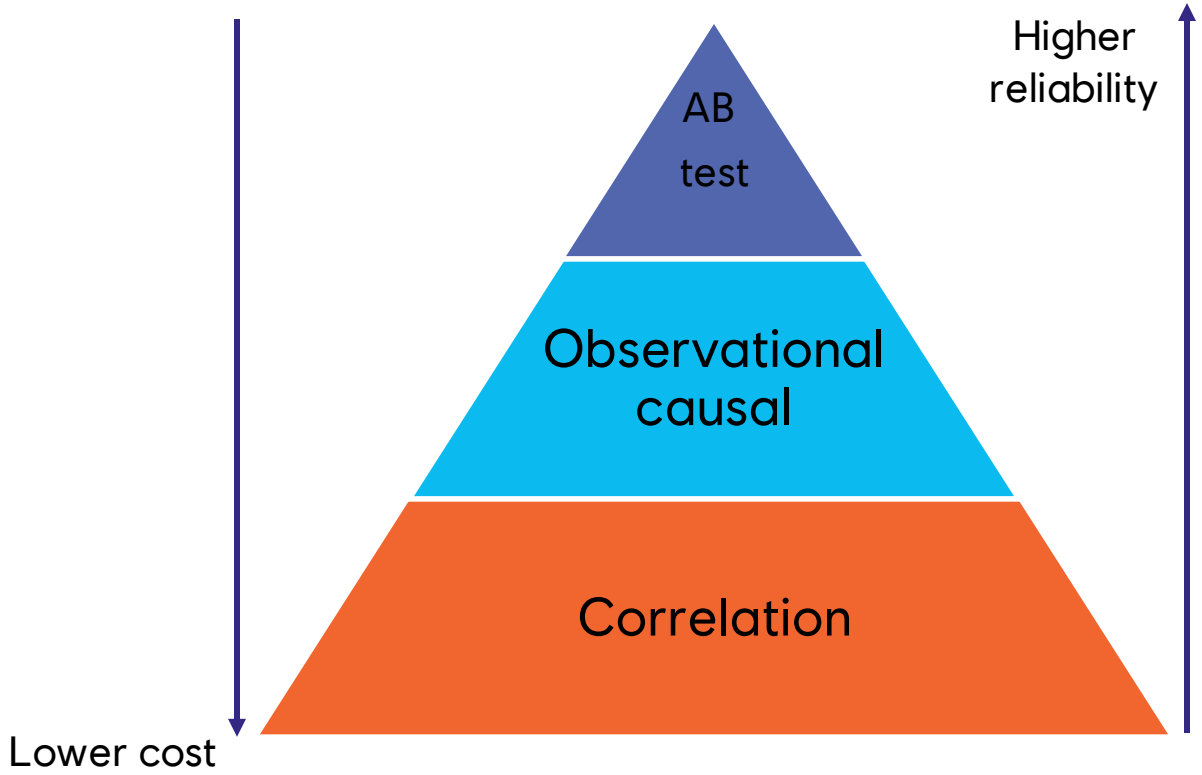
$$V(x,t) = \phi(x-t)/\Phi(x-t), \quad W(x,t) = V(x,t)(V(x,t) + (x-t)),$$

$$\tilde{V}(x,t) = -\frac{\phi(t-x) - \phi(-t-x)}{\Phi(t-x) - \Phi(-t-x)},$$

$$\tilde{W}(x,t) = \frac{(t-x)\phi(t-x) - (-(t+x))\phi(-(t+x))}{\Phi(t-x) - \Phi(-t-x)} + \tilde{V}(x,t)^2.$$

- To preserve distribution family, $\pi(\theta|x)$ is approximated by a normal.
- So that only μ_{ij}, σ_{ij}^2 are updated as the posterior mean and variance.

INFÉRENCE CAUSALE



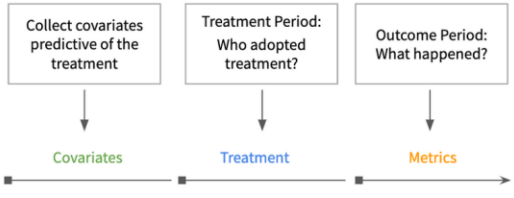
INFÉRENCE CAUSALE

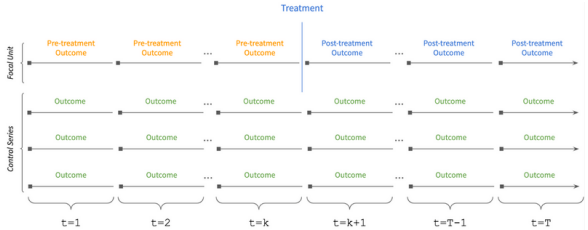
Participants (Treatment)

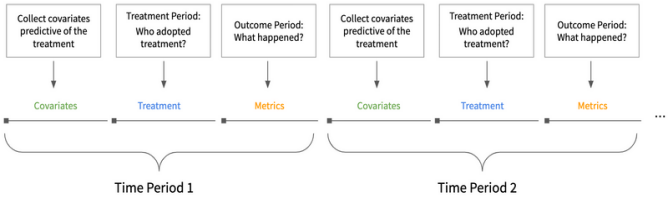
VS

Non-Participants (Control)

Inférence Causale







Cross-sectional Study

Interrupted Time Series

Panel Data Study


Partie III
TEAM LEAD DS @ UBISOFT

TEAM LEADING

COLLABORER

DISSÉMINER DE L'INFORMATION

CONTINUER DE SE FORMER

EXPÉRIMENTER ET ITÉRER

JONGLER, S'ORGANISER

TEAM LEADING

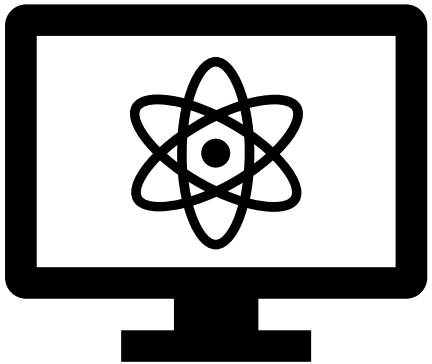
COLLABORER

DISSÉMINER DE L'INFORMATION

CONTINUER DE SE FORMER

EXPÉRIMENTER ET ITÉRER

JONGLER, S'ORGANISER




MERCI
QUESTIONS?