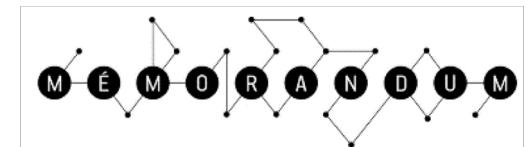


# **« Big data » et « Machine learning », en quoi le contrôle de gestion est-il concerné ?**

## Session 2

18 Avril 2016

*DataSapientia*



· 14h00 **Introduction**

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

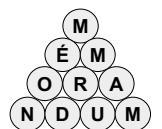
· 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

17h30 Clôture.



14h00 Introduction

- 14h15 **Cas pratique** : Comment un score peut-il permettre d'optimiser les ressources de recouvrement?

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

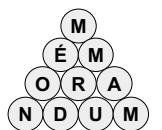
- 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

17h30 Clôture.



14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

- 14h45 **Présentation** : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

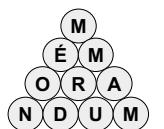
- 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

17h30 Clôture.

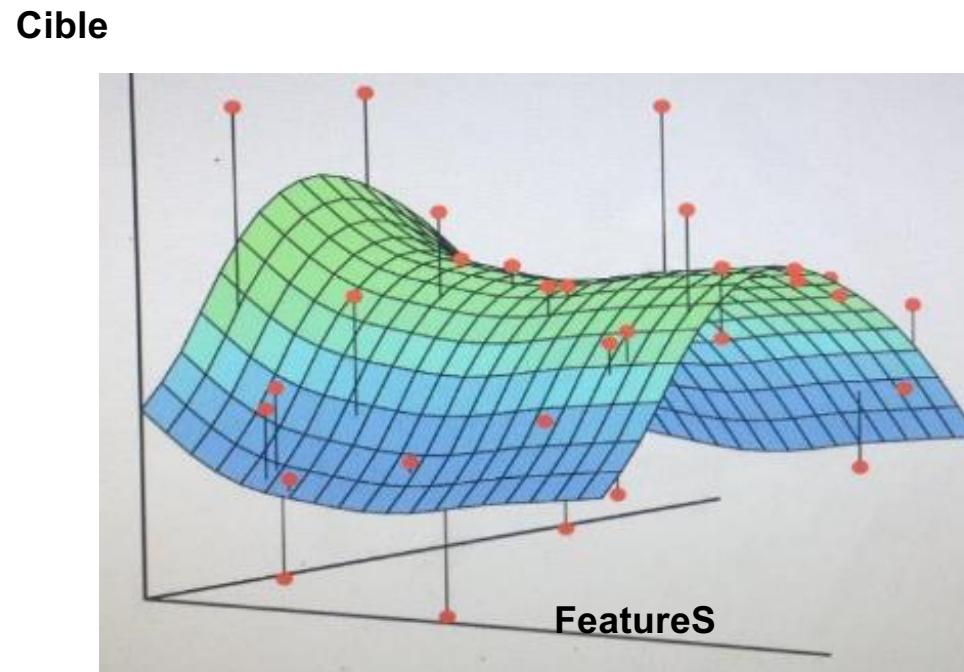
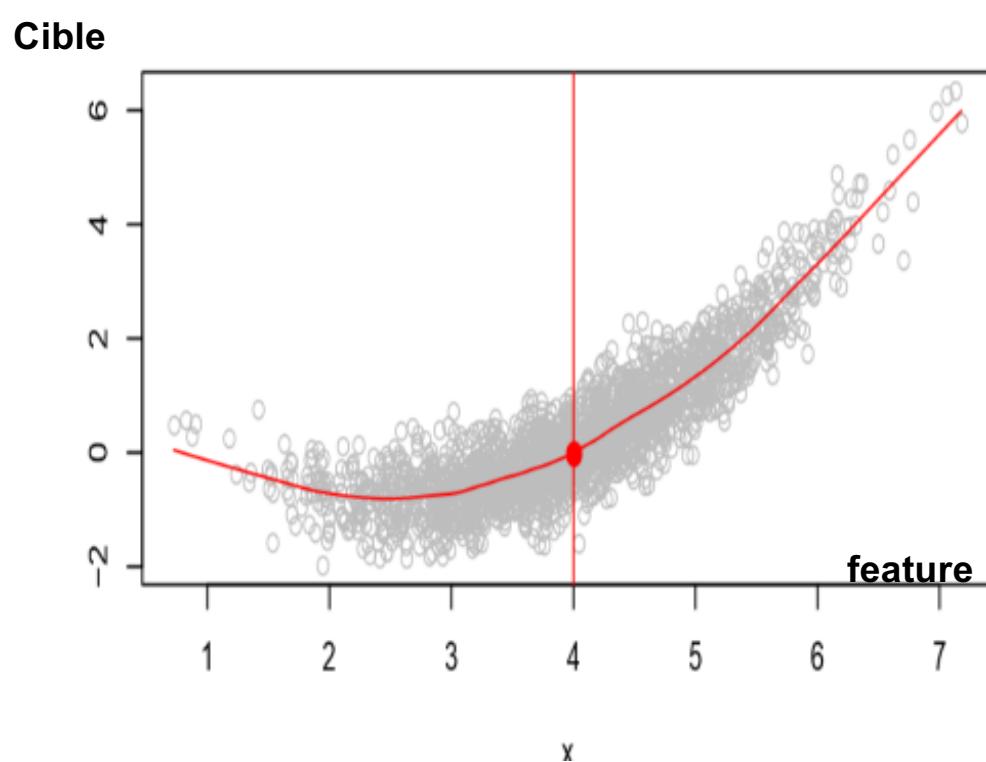


## Vue globale régression

Approcher une variable quantitative en fonction de chacun des paramètres disponibles

Fonction d'un espace de  $R^p \rightarrow R$

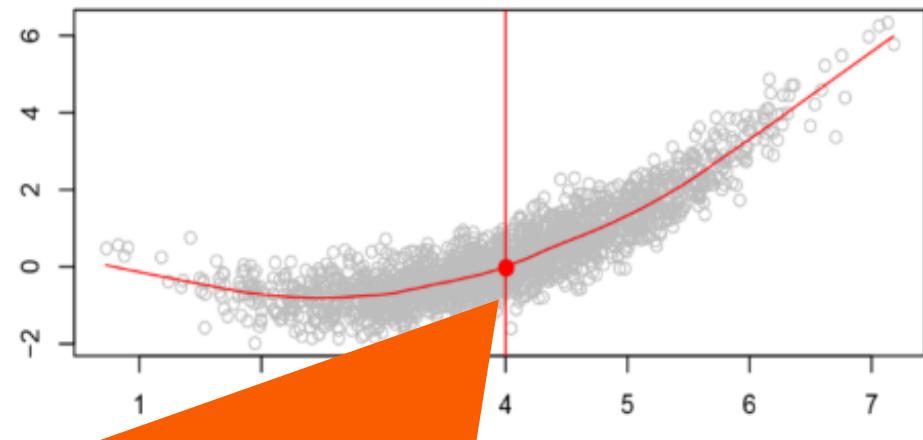
- Approcher une variable quantitative en fonction de chacun des paramètres disponibles
- Fonction d'un espace de  $R^p \rightarrow R$



## Le problème (1/2)

Idéalement

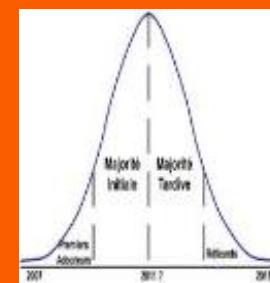
- Etre omniscient et avoir toutes les observations possibles
- Pour chaque valeur possible des features disponibles : prendre la moyenne des observations (espérance)



La dispersion autour de cette valeur moyenne peut être lié à plusieurs facteurs

Principal levier  
big data

- Il manque des facteurs explicatifs → toujours
- Il y a des erreurs de mesure → toujours
- Il y a du vrai hasard → là c'est de la philo



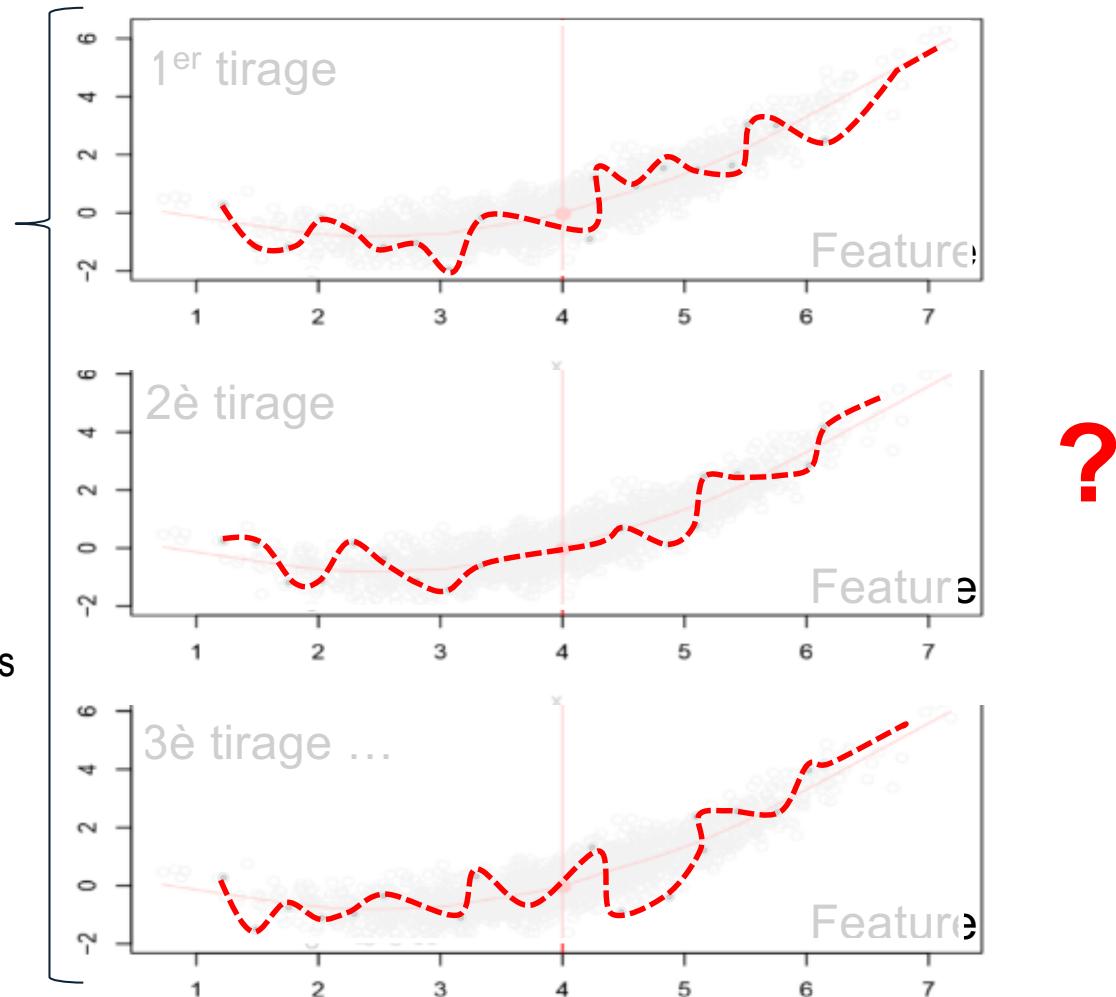
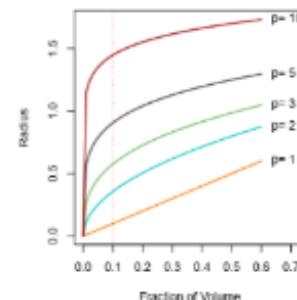
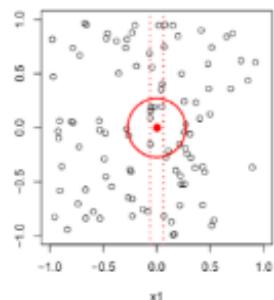
## Le problème (2/2)

**Hélas:**

- Vous ne disposez que d'un jeu de données partiel et si vous renouvez les mesures vous aurez chaque fois un autre jeu d'observation

**Hélas (bis)**

- Vous avez beaucoup d'observations.. mais encore plus de features pour chaque observation : vos êtes atteint par la malédiction de la dimension (« curse of dimensionality ») :
- dans un espace à haute dimension, vos observations sont éclatées : il n'y a plus de voisins ..



## Une démarche pleine de bon sens 1/3

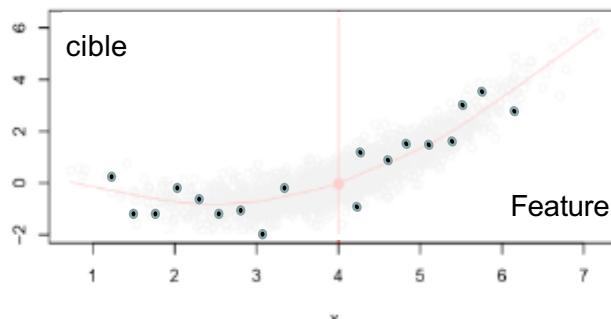
Pour bien prévoir le futur nous pouvons **simplifier le passé**

Equivalent

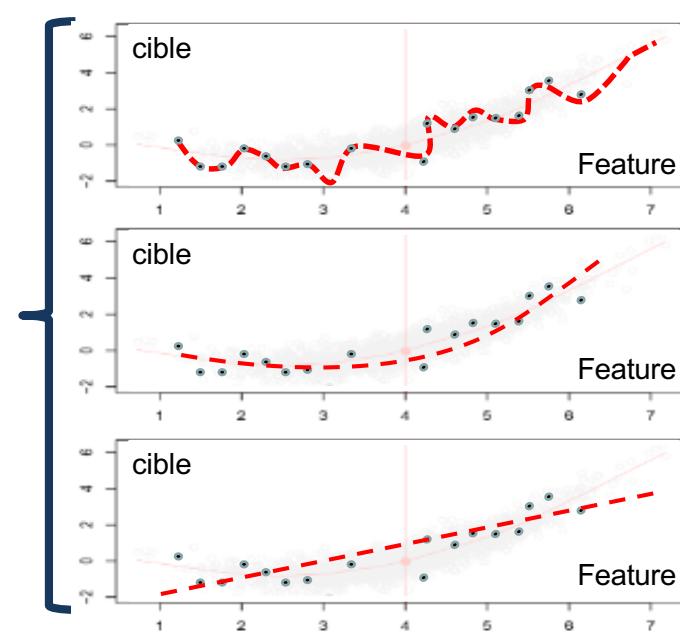
Distinguer

- le **signal** : « vraie » information apportée par les features disponibles
- du **bruit** : effet des informations (features) qui nous manquent

Données initiales



Modélisation induite



### Complex

- Parfaite description du passé
- Faible pouvoir prédictif
- « overfitting »

Un juste milieu ?

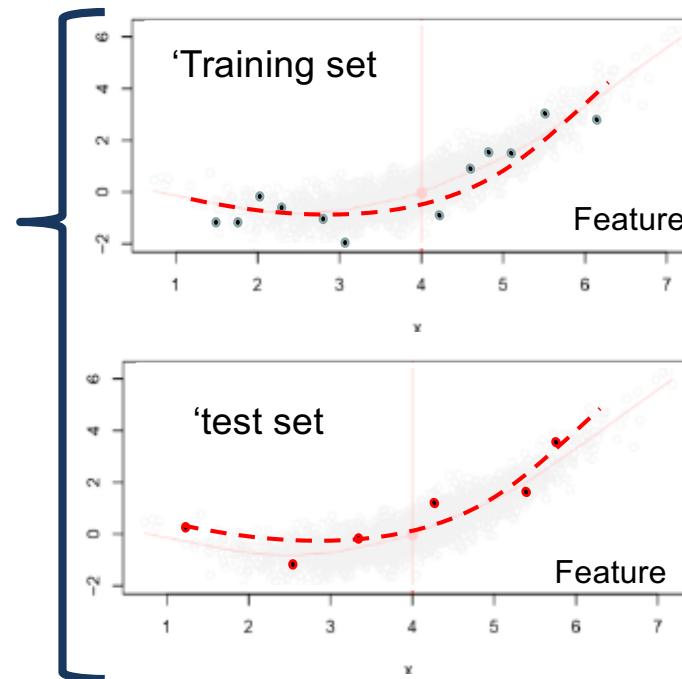
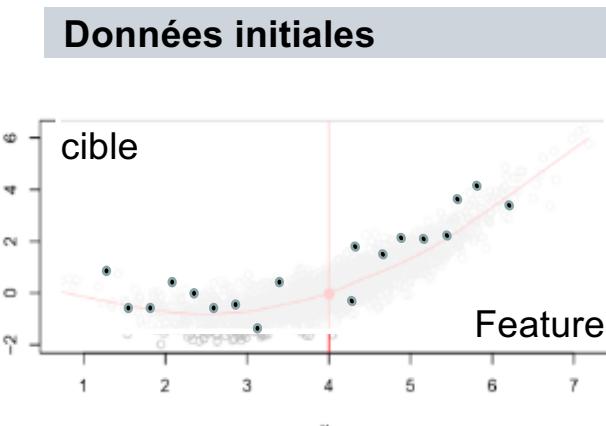
### Simple

- Grossière description du passé
- Faible pouvoir prédictif

## Une démarche pleine de bon sens 2/3

2<sup>e</sup> astuce : appliquer l'adage qui dit qu'on ne peut être juge et parti (séparation jeu d'apprentissage et jeu de test)

- L'évaluation de l'erreur d'interpolation des données connues n'est visiblement pas la métrique pertinente (sinon on va systématiquement pencher du coté « overfitting »)
- Solution « on ne peut pas être juge et partie »: les données connues sont réparties en deux lots
  - Un lot d'apprentissage
  - Un lot d'évaluation

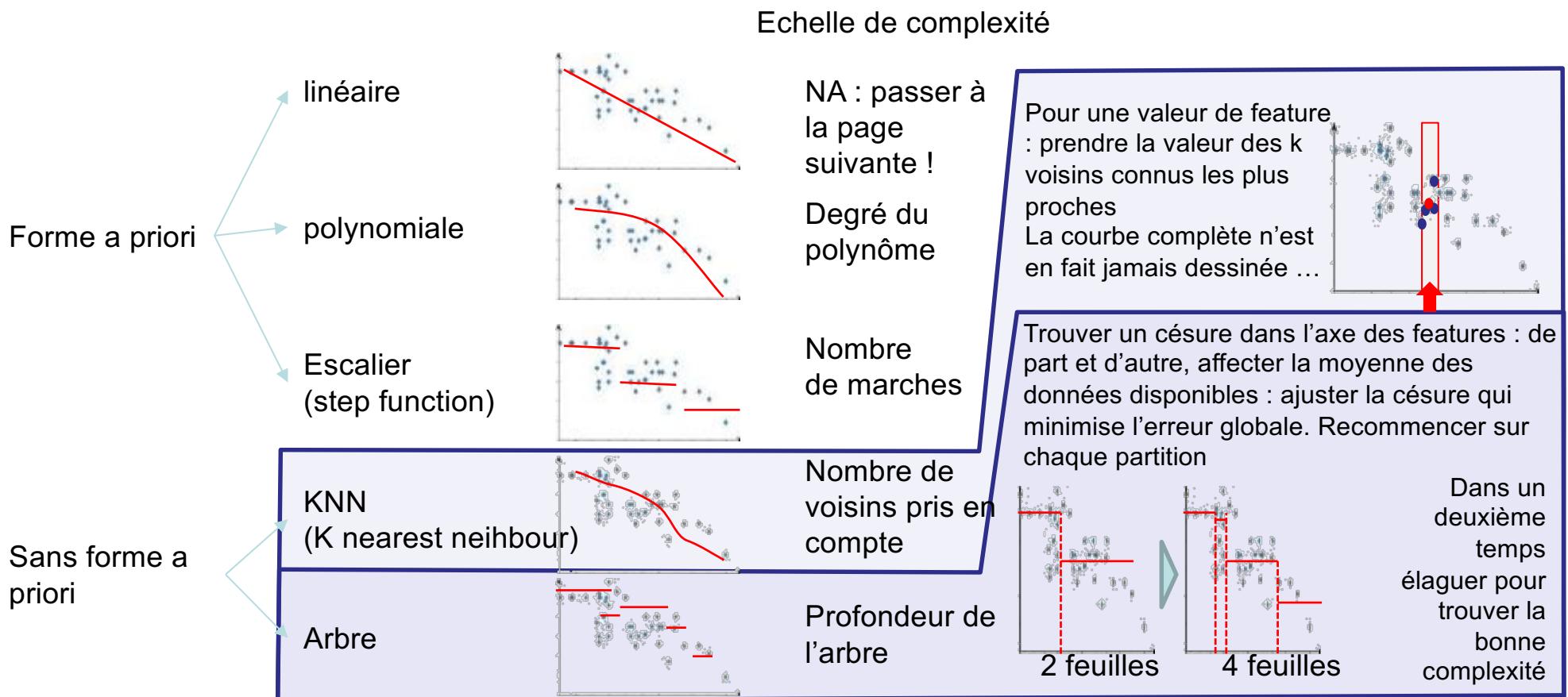


Construction d'un modèle pur un niveau de complexité donné

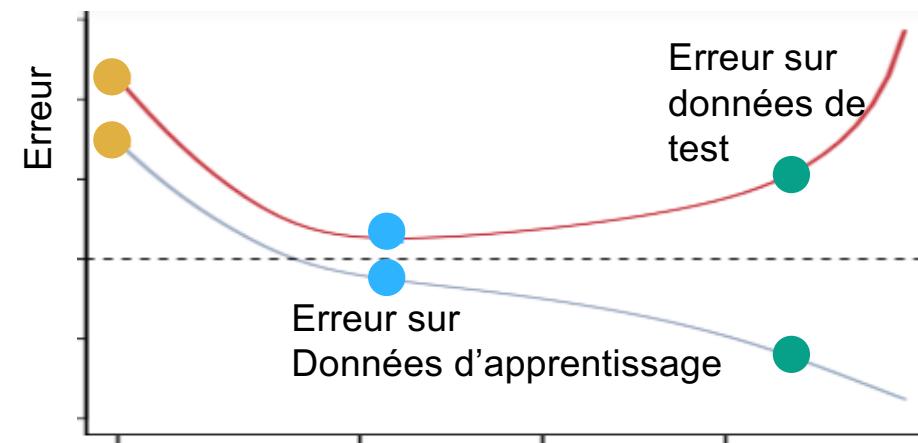
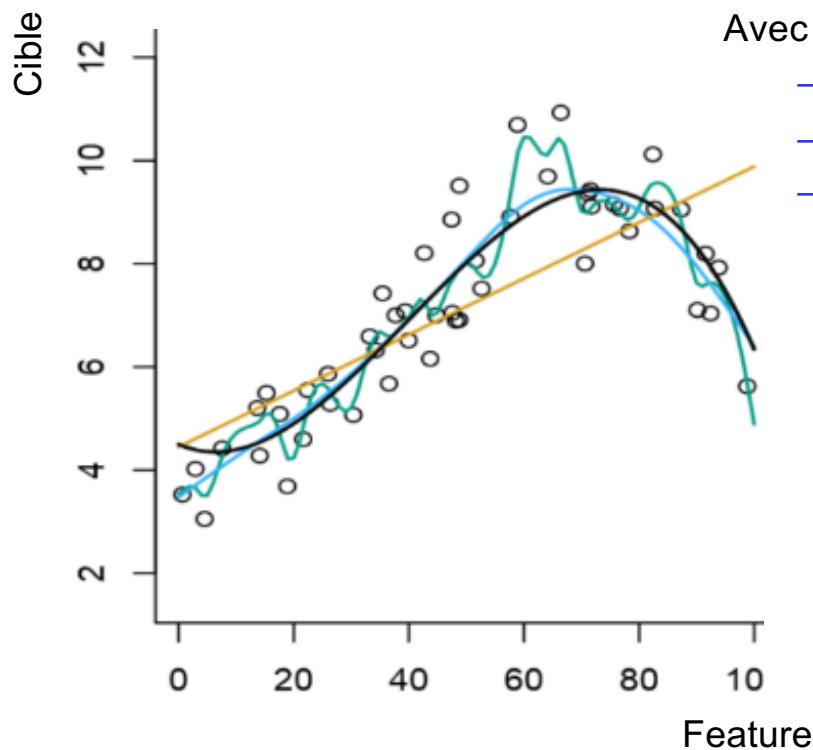
Evaluation : le modèle construit est évalué avec les points qui n'ont pas servi à la construction

## Une démarche pleine de bon sens 3/3

Ce choix de forme est un a priori, potentiellement guidé par la visualisation des données ou par l'expérience (attention : expérience est une lanterne dans le dos !)



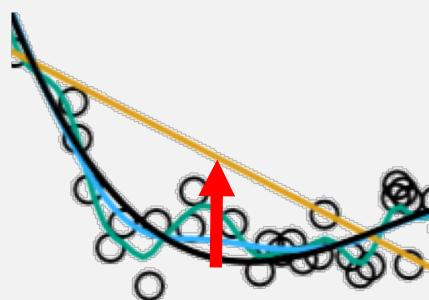
## Zoom : effet de la complexité sur l'erreur de prédiction (1/2)



## Zoom : effet de la complexité sur l'erreur de prédiction (2/2)

### Erreurs de biais

- Lié à la « raideur » du modèle
- peu sensible au jeu de données disponibles



Il existe un juste équilibre ...  
et on va le trouver !



Erreur sur  
données de test

Echelle de complexité du modèle  
(exemple degré de polynôme d'interpolation)

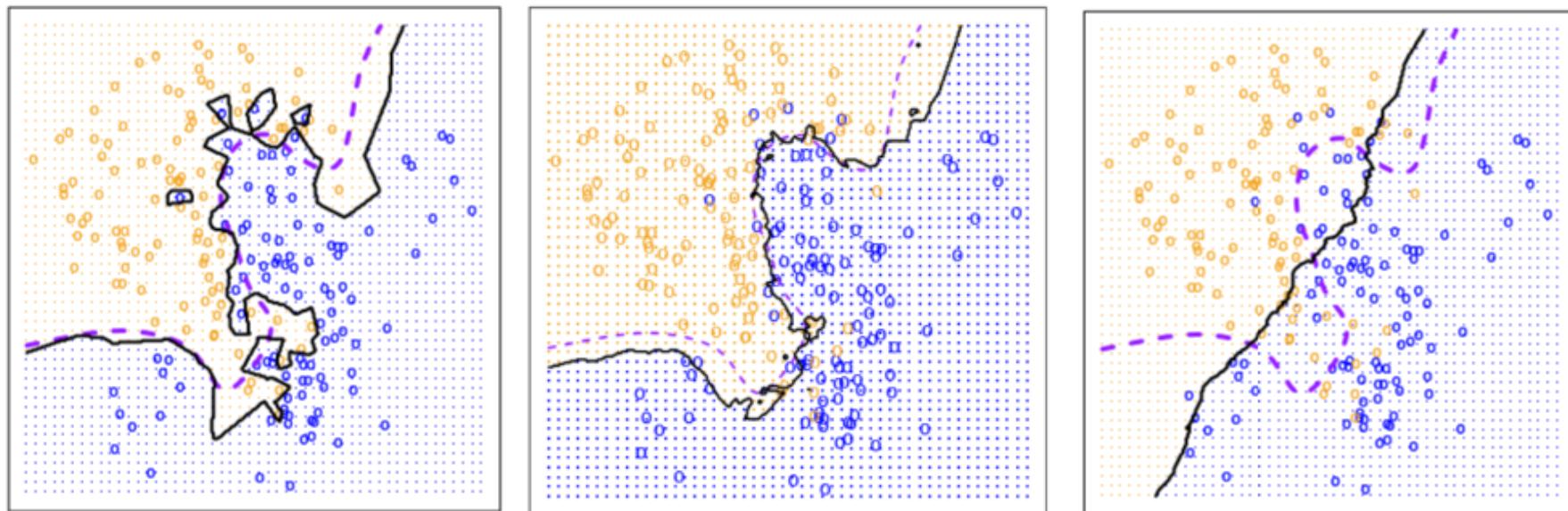
### Erreurs de variance

- Lié à la contingence des données disponibles : juste en moyenne par définition mais toujours versatiles !



Le bon réglage de la simplification du passé n'est pas fournie par les mathématiques c'est un travail d'artisan !

## K nearest neighbour



14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

• 15h15 Atelier : « Provisions, autres idées de sujets ?».

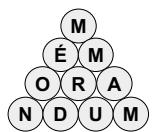
• 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

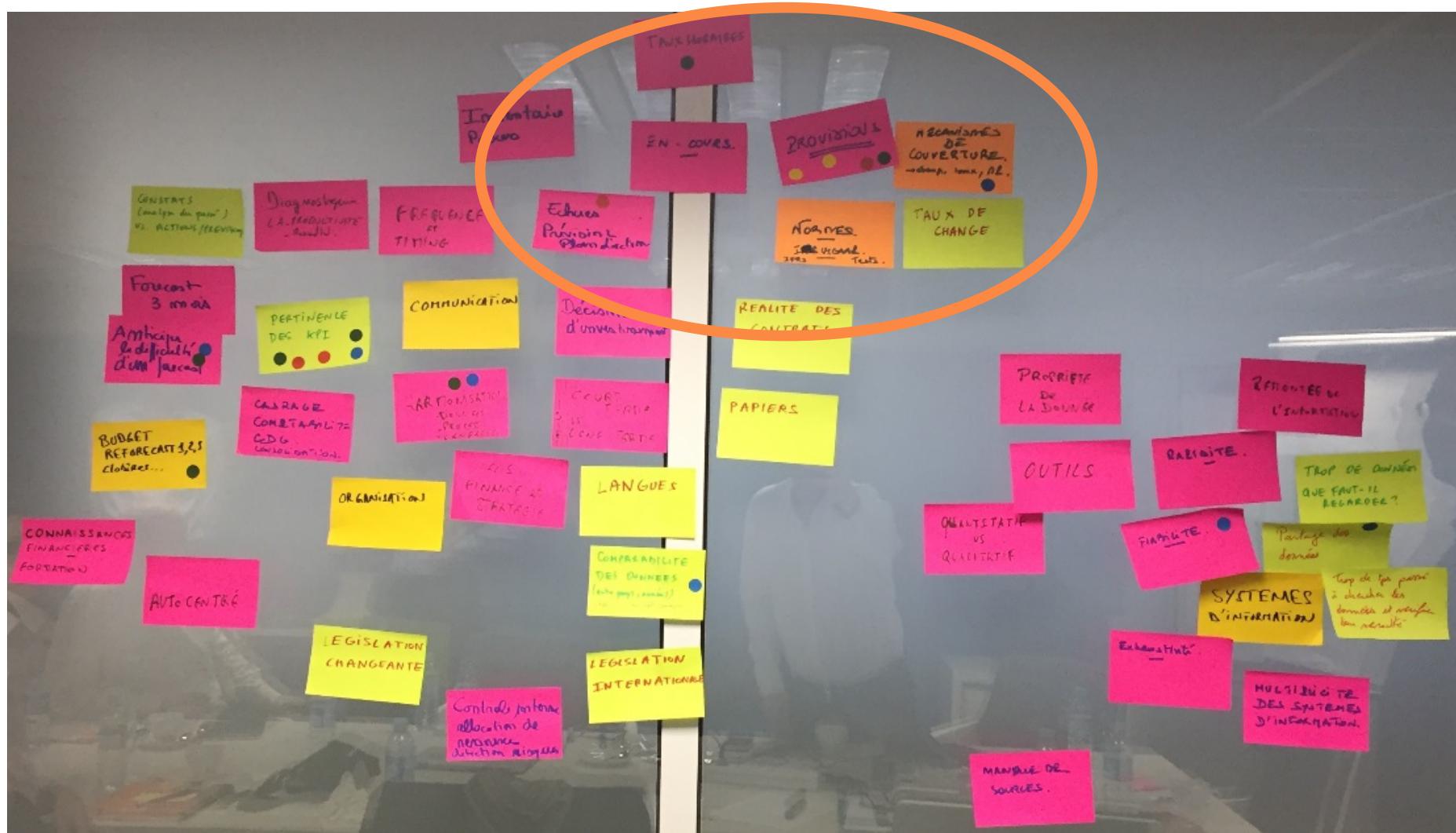
16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

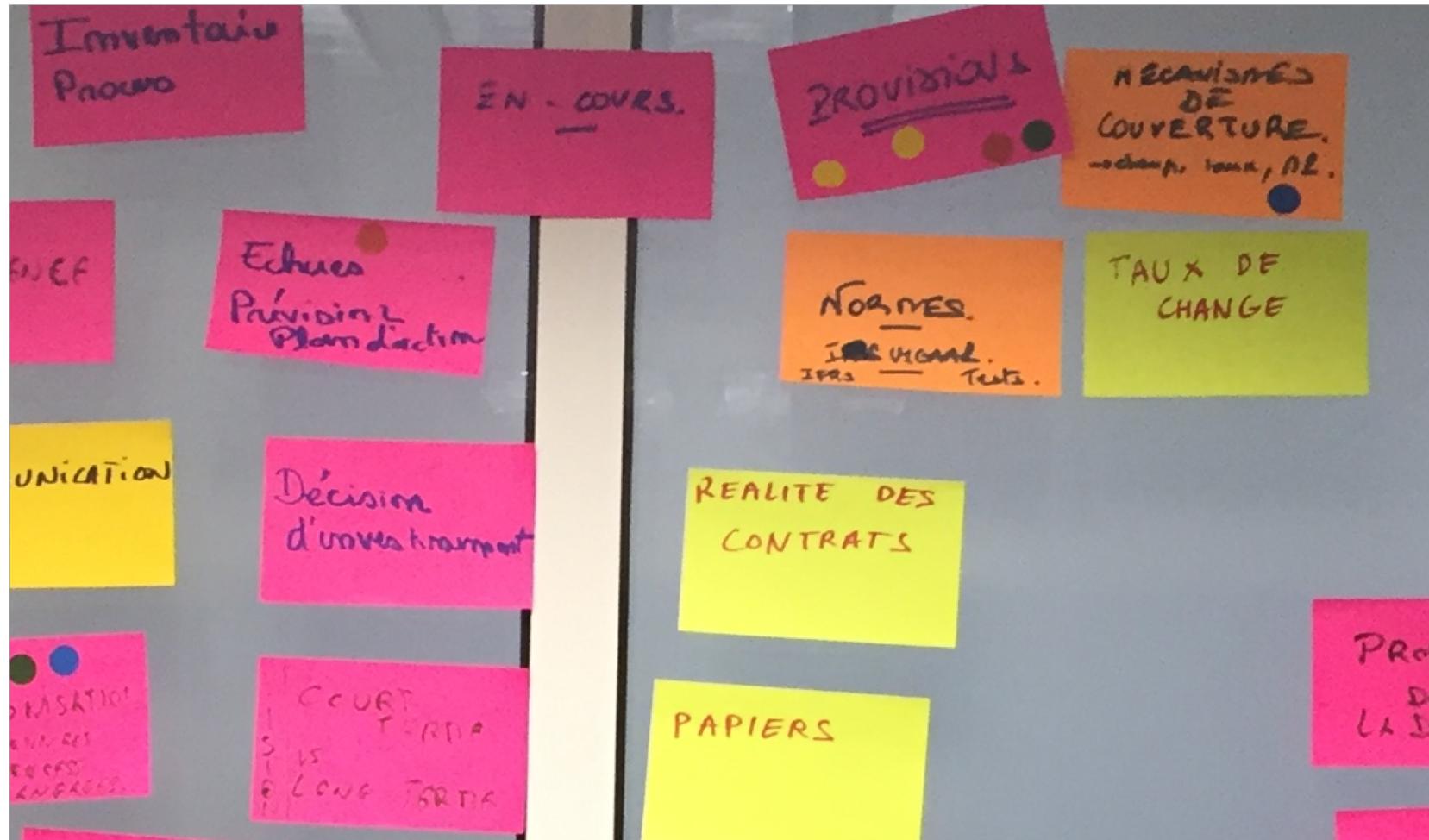
17h30 Clôture.



## Thème « Provisions » : Retour sur les brainstorming du 25 mars



## Thème « Provisions » : Retour sur les brainstorming du 25 mars



## Bien formuler la question : l'enjeu principal !

Parce que c'est la phase initiale qui conditionne la suite

Parce que peu documentée et outillée, contrairement à la partie analytique elle même

The screenshot shows the homepage of L'Usine Digitale. At the top, there's a navigation bar with links for 'TOUTE L'INFO', 'L'USINE NOUVELLE', 'INSCRIVEZ-VOUS À LA NEWSLETTER', 'DIGITAL AVENUE', and a search bar. Below the navigation, there's a main article titled 'Organiser un concours de datascientists en 5 étapes'. The article includes a sidebar with a screenshot of a software interface and a section titled '2. BIEN FORMULER LA QUESTION' containing text about defining metrics for data science challenges.

**Organiser un concours de datascientists en 5 étapes**

Par Sylvain Arnal - Publié le 02 mai 2015, à 14h00

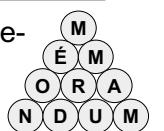
Les plateformes qui proposent des challenges aux datascientists sont de plus en plus utilisées par les entreprises. Monde d'emploi

**2. BIEN FORMULER LA QUESTION**

Comme pour tous les services d'appel à la foule (crowdsourcing), faut poser la bonne question.

Dans le cas du big data : quel est le phénomène précis que l'entreprise cherche à prévoir ? Une fois la problématique clairement déterminée, il faut identifier les données pouvant être utiles pour résoudre le problème, puis transcrire cette question métier en question mathématique. "Il faut définir une métrique (mathématique, objective, calculable) pour mesurer la qualité prédictive des modèles proposés", conseille Arnaud Laroche, le PDG de l'agence Bluestone et cofondateur de DataScience.net. Cela permettra de départager les participants. L'entreprise doit aussi décider quels jeux de données elle met à leur disposition et leur mise en forme. Cette phase très technique de collecte et de distillation des données permettra aux participants de se concentrer sur le challenge stricto sensu.

<http://www.usine-digitale.fr/editorial/organiser-un-concours-de-datascientists-en-5-etapes.N326894#xtor=EPR-4>



## Poser une question : aussi délicat que de trouver une réponse

**Une question ...**

**Exemple**

Repérer les churners ?

**.... Qui a besoin d'être précisée**

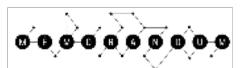
Churn ? résiliation administrative ? A  
d'usage d'un service, ... ?

Tous ? Ou un  
un certain

sûr :  
ation de la  
change la réponse

**Peut être cherché**

- Optimiser le processus
- Comprendre le comportement des churners :
- ....
- Que fera-ton avec la réponse ? (courrier, appel, offre, ...)
- Comment mesurera-ton la performance ?
- Questions de base :
  - churn = choix rationnel du client ou résultat d'un démarchage
  - Combien de temps avant le client prend-il sa décision



**Vous avez une question précise ?**



**Imaginez que vous avez la réponse : que faites vous avec ?**



**Quelles données pouvez vous utiliser pour y répondre**

## Typologie des provisions

*(construction collaborative)*



## Formulation des questions

*(Post-it, votes, reformulation)*

## Brainstorm Rules

### Defer judgment

There are no bad ideas at this point. There will be plenty of time to judge ideas later.

### Encourage wild ideas

It's the wild ideas that often create real innovation. It is always easy to bring ideas down to earth later!

### Build on the ideas of others

Think in terms of 'and' instead of 'but.' If you dislike an idea, challenge yourself to build on it and make it better.

### Stay focused on the topic

You will get better output if everyone is disciplined.

### Be visual

Try to engage the logical and the creative sides of the brain. A quick sketch can help make your idea more understandable to someone else.

### One conversation at a time

Allow ideas to be heard and built upon.

### Go for quantity

Set a big goal for number of ideas and surpass it! Remember there is no need to make a lengthy case for your idea since no one is judging. Ideas should flow quickly.

## It's time to brainstorm

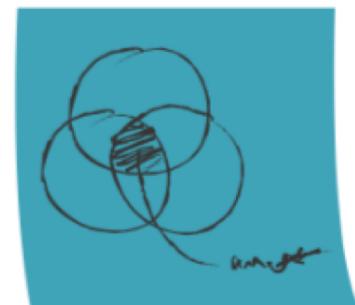
Quelles sont les problématiques que vous rencontrez régulièrement dans l'exercice des métiers de la finance d'entreprise (*contrôle de gestion, comptabilité, contrôle interne, consolidation, tax, ...*)



Pacing



Use of post-its



Sketching

14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

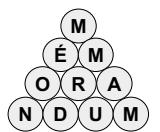
• 15h45 Pause

• 16h00 **Atelier : « Pertinence des KPI » - Exploration du thème.**

16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

17h30 Clôture.



## Thème « Provisions » : Retour sur les brainstorming du 25 mars



## Explicitation / exploration du thème *(discussion collaborative)*



## Formulation des questions *(Post-it, votes, reformulation)*

14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

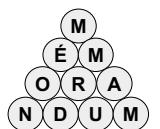
- 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

- 16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

17h30 Clôture.



14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

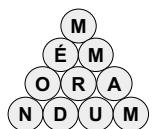
- 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

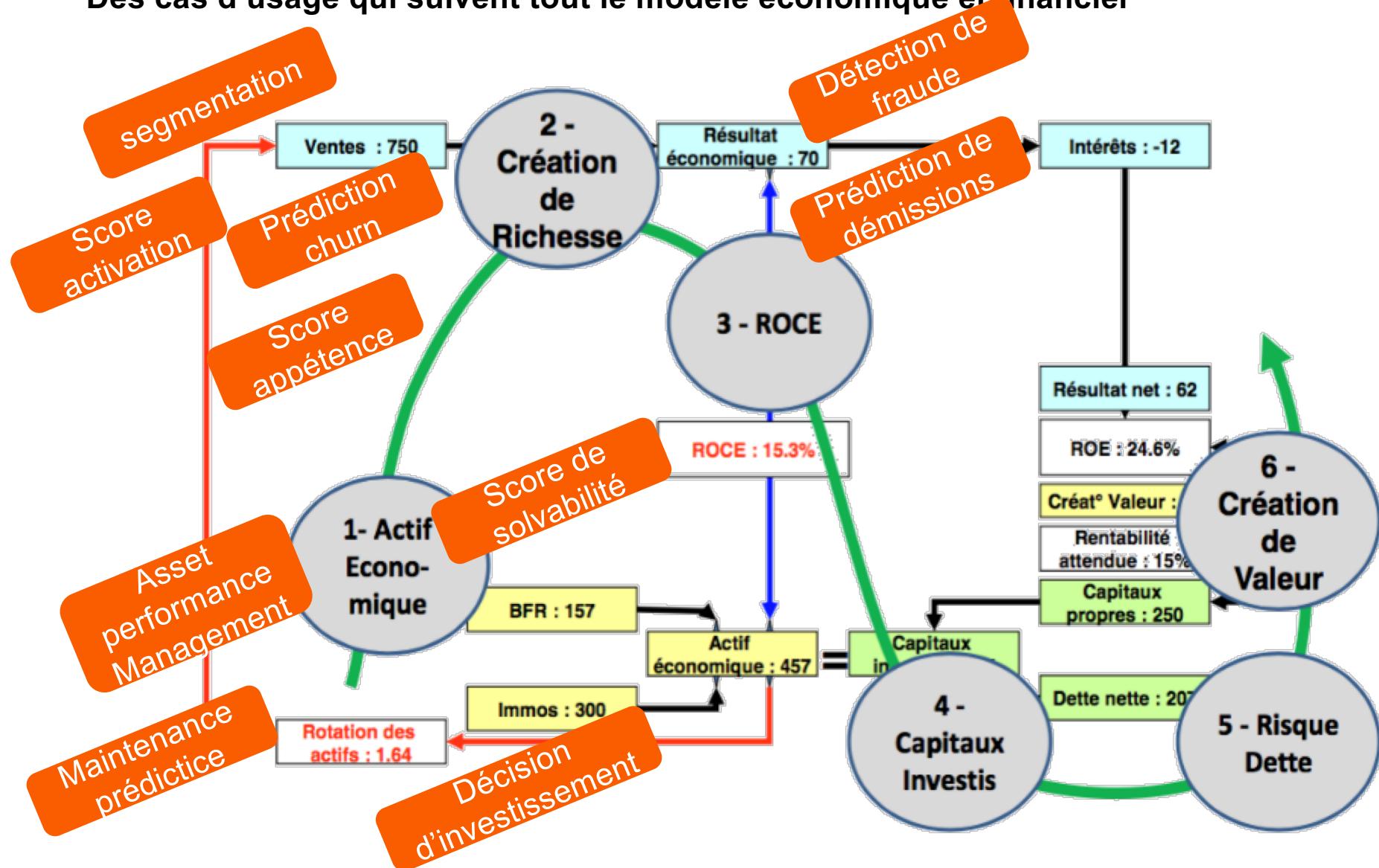
16h30 Cas pratique : « Pertinence des KPI ».

- 17h00 Bilan.

17h30 Clôture.



## Des cas d'usage qui suivent tout le modèle économique et financier



14h00 Introduction

14h15 Cas pratique : « Comment un score peut-il permettre d'optimiser les ressources de recouvrement? ».

14h45 Présentation : « Comment ça marche ? Ouvrons le moteur! ».

15h15 Atelier : « Provisions, autres idées de sujets ?».

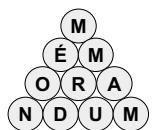
• 15h45 Pause

16h00 Atelier : « Pertinence des KPI » - Exploration du thème.

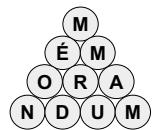
16h30 Cas pratique : « Pertinence des KPI ».

17h00 Bilan.

• 17h30 Clôture.

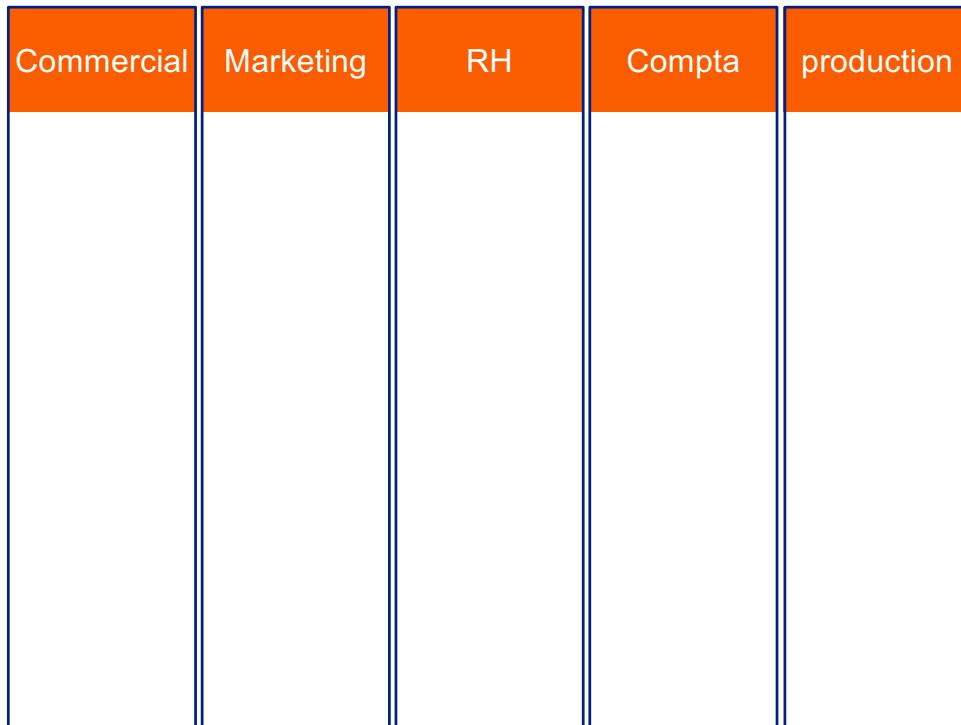


# Annexes



## Managez vos données comme vous managez vos équipes

**Vos données sont rangées comme vos organisations**



**Vos données doivent être animées comme votre management : transverse**



**Pas d'un projet informatique mais de management. Objectif managers :**  
 – Être garant de la disponibilité de ses données  
 – réutiliser les données des autres directions

**Vos données ne travailleront ensemble que sous l'impulsion de votre  
Comité de Direction**

## Plus globalement La bonne recette

1

### « La valeur n'attend pas le nombre des octets »

Une valorisation efficace des données dépend moins du nombre d'observations (quelques milliers de clients ou de dossiers représentent déjà une bonne base) que de la richesse de ses observations (pour un client avoir ses caractéristiques, ses transactions, ses courriers, ...)

2

### « Exploiter la donnée est un projet métier, pas un projet informatique »

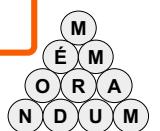
L'initialisation d'une démarche de valorisation des données ne nécessite aucune infrastructure :

- Des ordinateurs de bureau (jusque quelques giga de données)
- Des logiciels d'analyse open source (R / Python)
- Des algorithmes d'analyse libres de droit, prédéveloppés et prêt à l'emploi

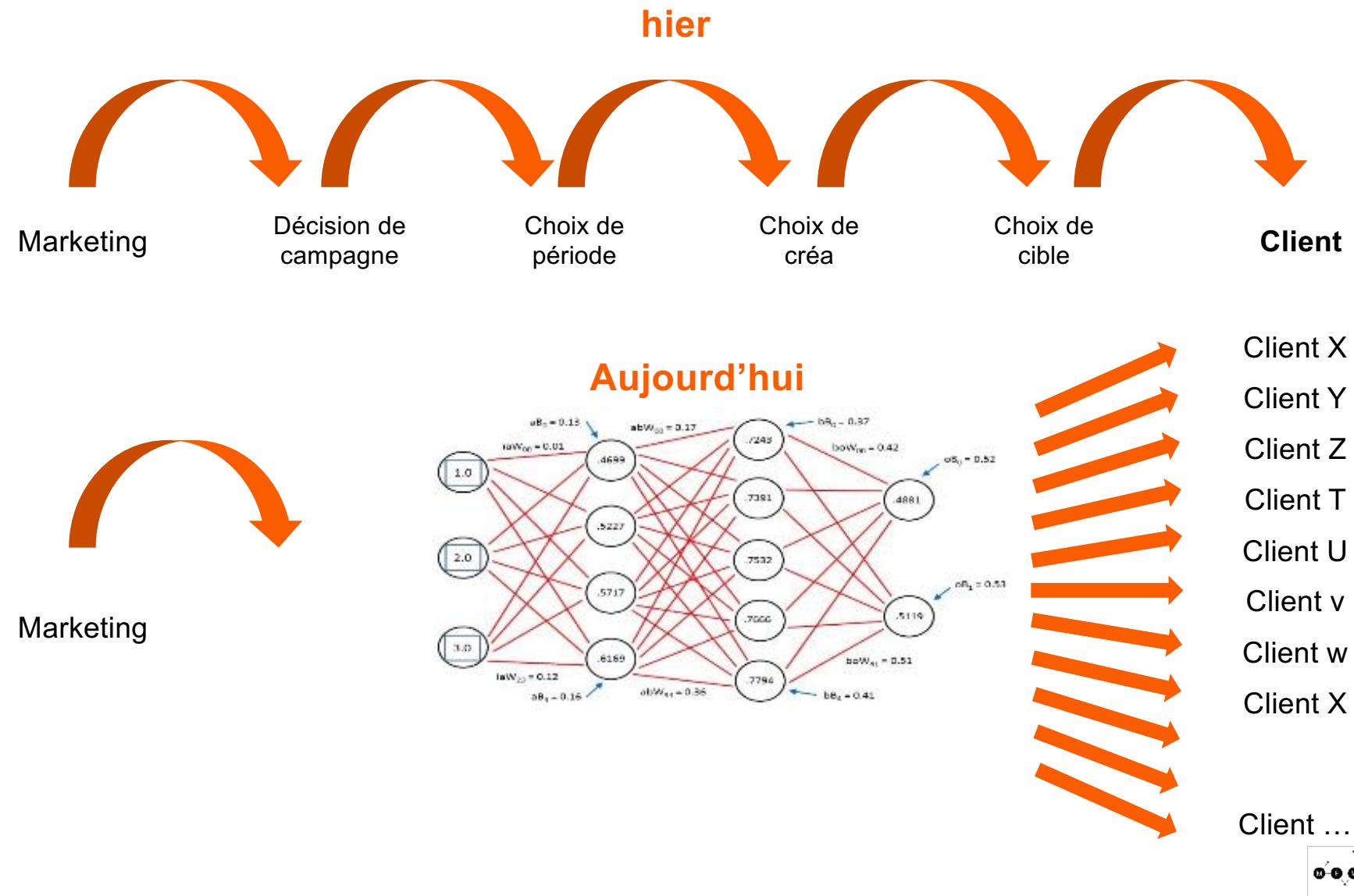
3

### « Fail fast & learn »

Les démarches de valorisation des données se font par itérations, il est donc important de savoir avant de se lancer que certaines analyses potentiellement non concluantes (mais auront fait progresser)

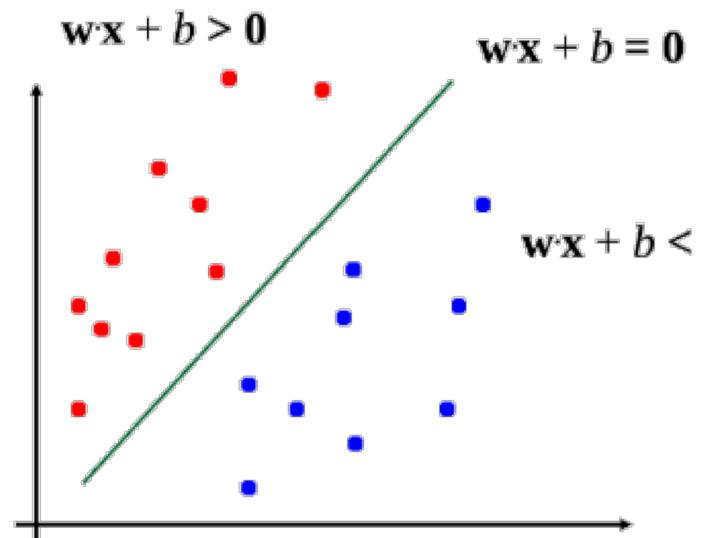


## Un circuit de décision transformé



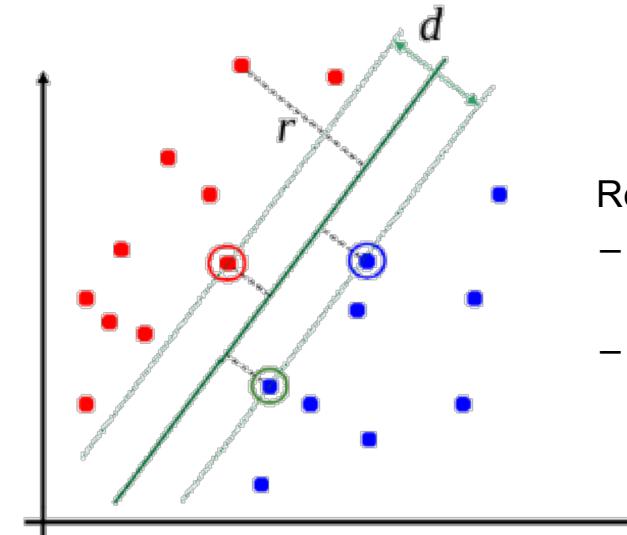
## SVM : Support vector machine / séparateur à vaste marge

Initialement le perceptron juge simplement l'erreur induite sur le jeu de test ne conduisant pas à un optimum unique



Une approche développée par Vapnik en 95

- Mathématiquement : démarche d'optimisation sous contrainte, faisant intervenir une transformation de Lagrange
- Intuitivement : quelle position d'hyperplan donne une séparation avec la meilleure marge de sécurité en fonction des points connus



Réglages

- Fonction de coût (nombre de vecteur supports)
- noyau

## Arbres de décision

Principe : séparation due l'espace de manière itérative, de plus en plus précise

Première étape :

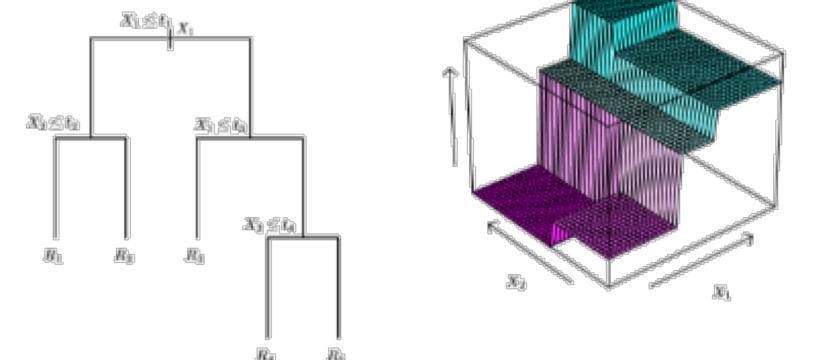
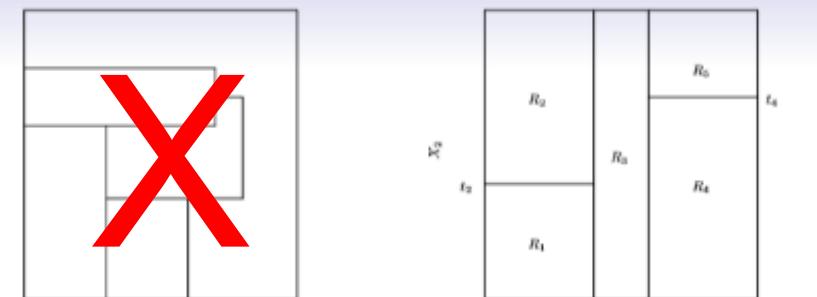
- Pour chaque feature, chercher le seuil qui sépare le mieux dataset (selon fonction de coût)
- Retenir la feature générant la meilleure séparation

Étapes suivantes

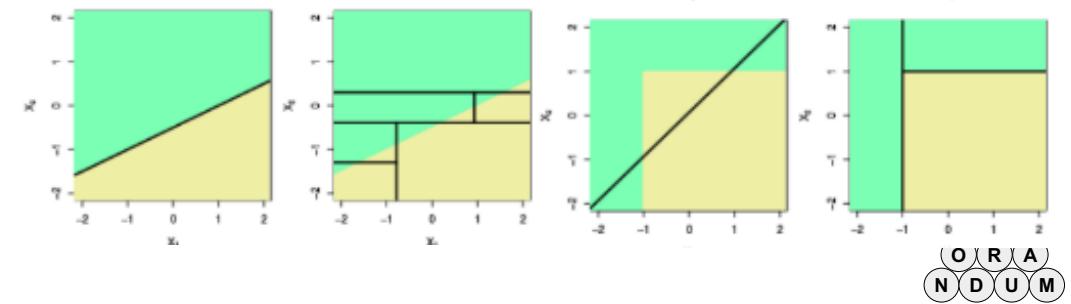
- Recommencer de la même manière sur chaque sous dataset obtenu à l'étape précédente

Cette démarche peut continuer jusque ce que les feuilles (sous dataset de la dernière itération soient des observations uniques

Puis on taille les branches (marche arrière) (« pruning ») et on s'arrête quand on atteint



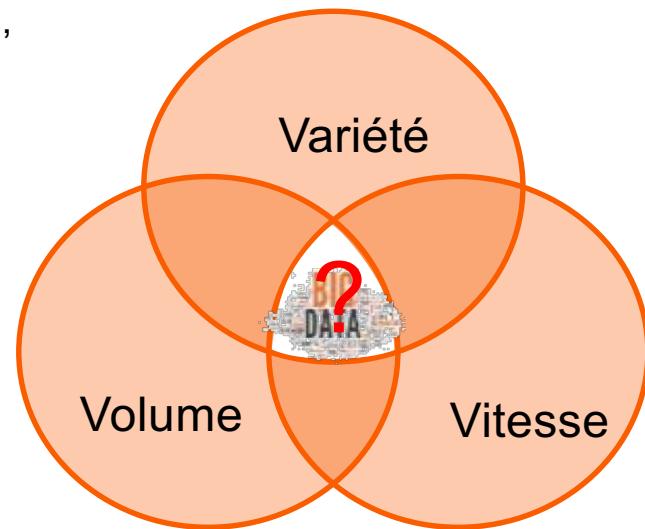
Arbre / régression linéaire



## A partir de quel volume est on éligible aux techniques du big data ?

**Une définition fréquente : Au moins 2 des 3 « V »**

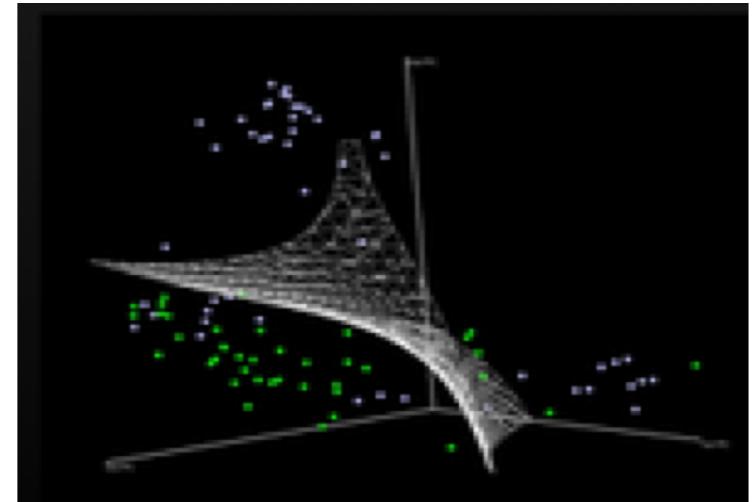
- Volume,
- Vitesse,
- Variété



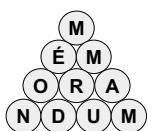
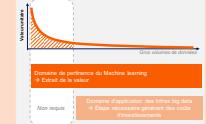
*Les 3V ne sont pas un plancher d'accès mais le symbole d'un plafond constamment repoussé par la technologie*

**Une définition plus juste de la puissance de l'approche → 0 V**

- ci-dessous : 150 observations, 3 caractéristiques → une modélisation fine !

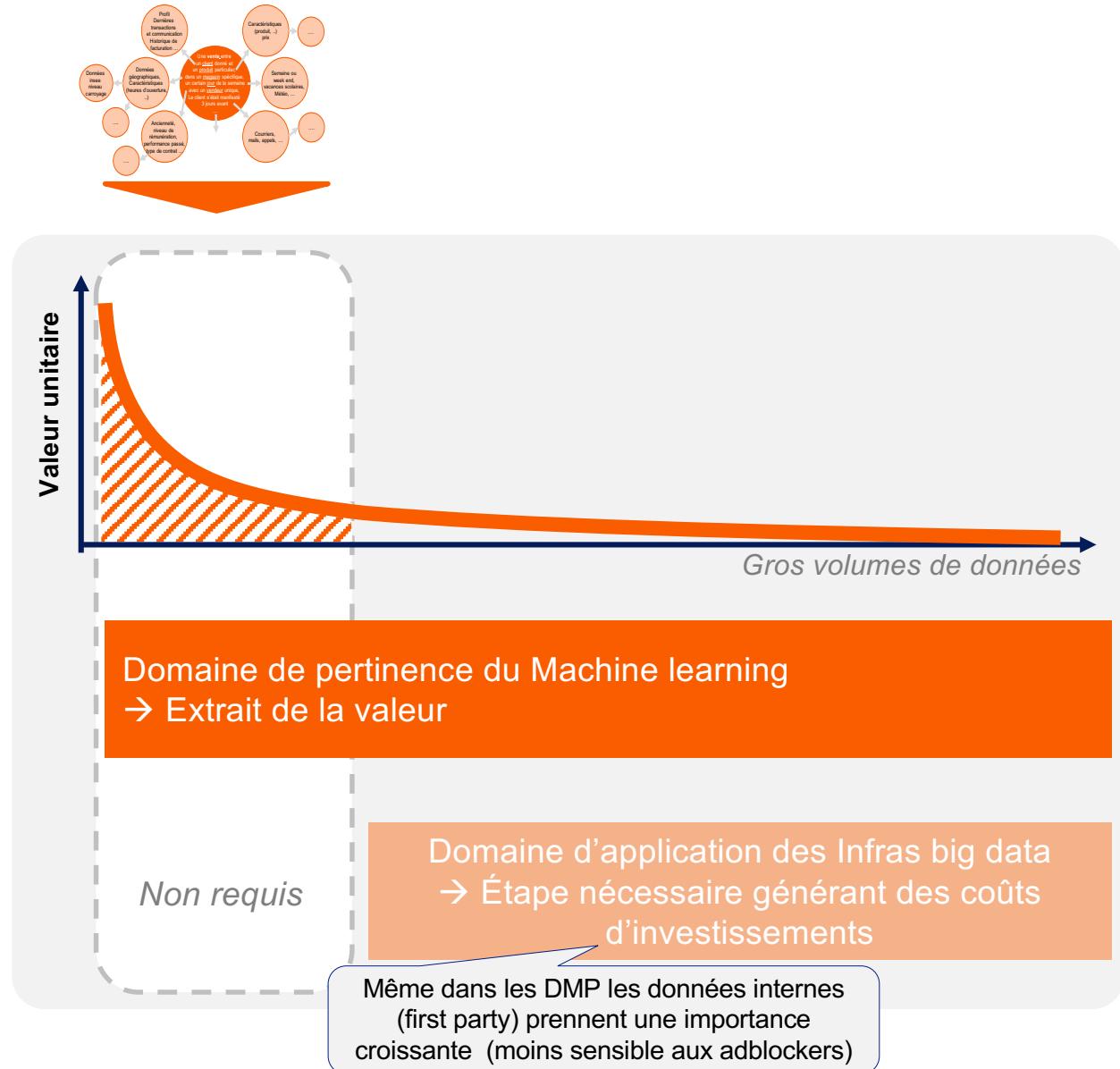


**Votre entreprise a assez de données pour tester l'application des algorithmes prédictifs open source (ceux qui marchent !)**



## L'exploitation des données internes de l'entreprise cumule 3 avantages

- 1) Travail sur des données naturellement riches et accessibles
- 2) Pertinentes pour les algorithmes d'analyse
- 3) Sans besoin d'infrastructure



## **Vous êtes plus riche (en données) que vous ne le croyez**

Toutes nos transactions laissent des traces

Objets et humains en produisent ... même sans transactions

Etat et entreprises ouvrent leurs données

non structurées

internes

CRM et ensemble des bases de données de 'entreprises'

Carte au trésor

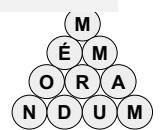
externes

Open data

Croisement de données interentreprises

Third parties

Réseaux sociaux

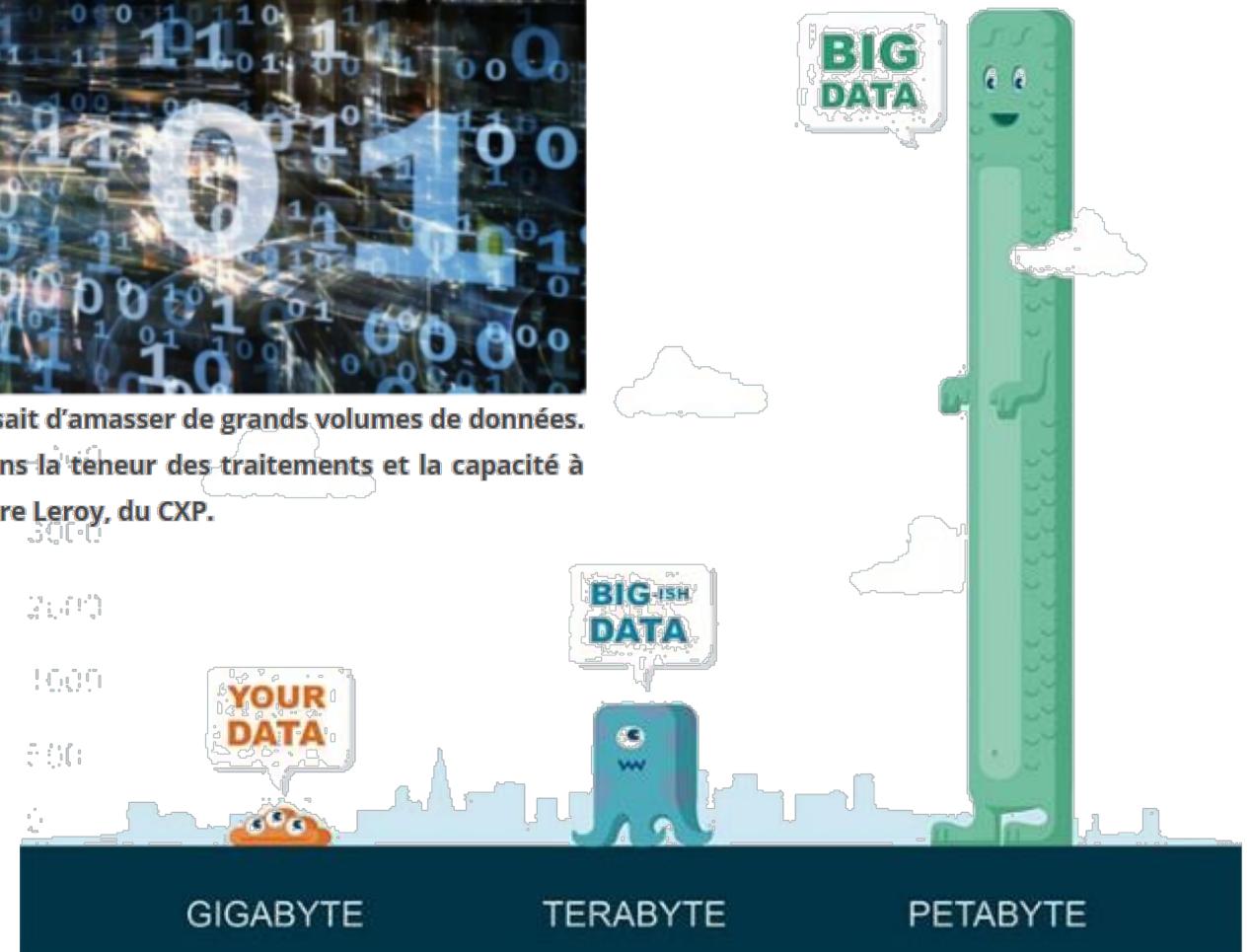


## Big Data : une affaire de traitement et non de volume (tribune)

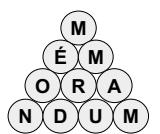
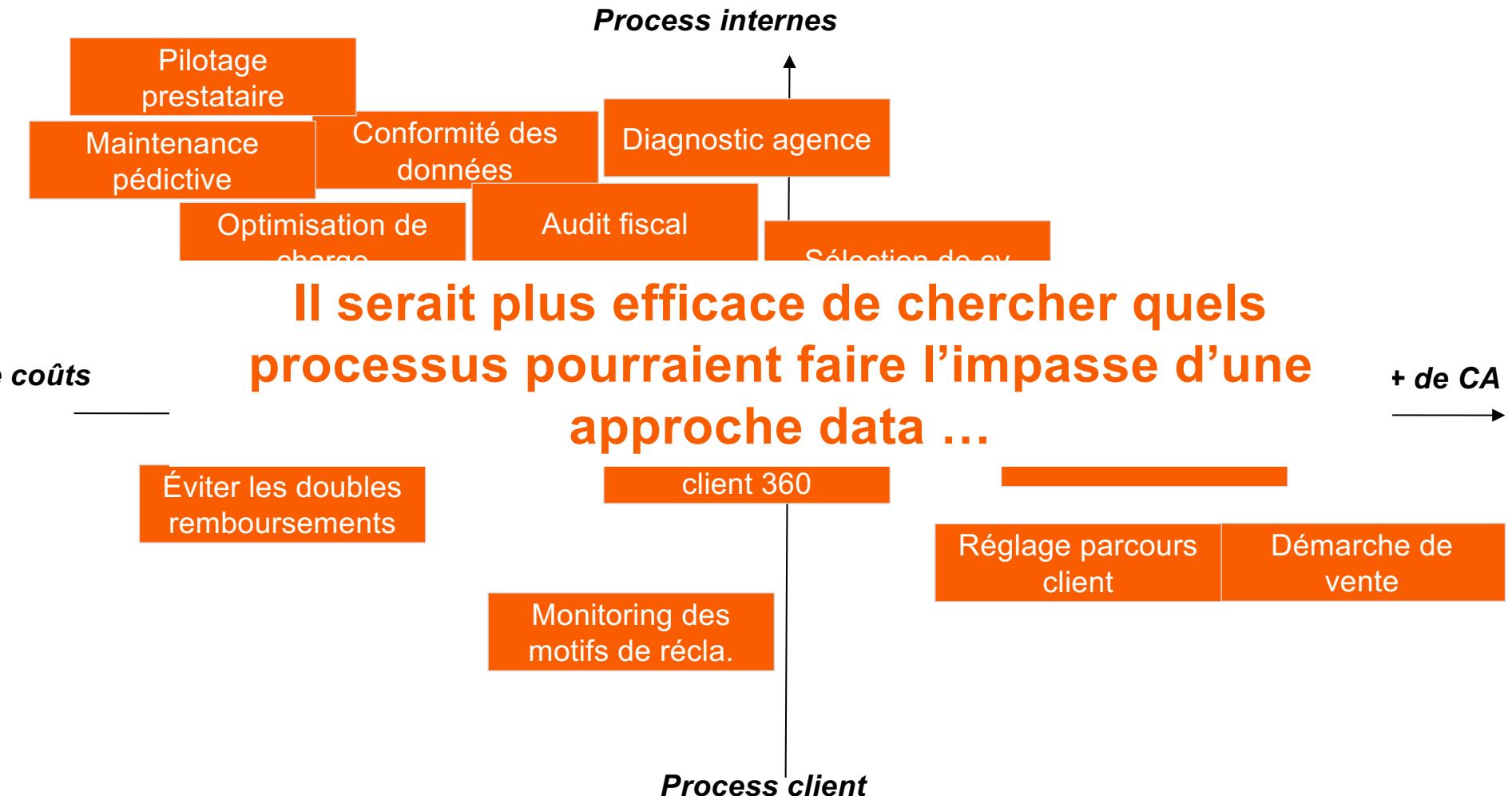
La Rédaction, 26 novembre 2015, 9:32



Le terme de Big Data a pu laisser penser qu'il s'agissait d'amasser de grands volumes de données. Mais la valeur de ces applications réside plutôt dans la teneur des traitements et la capacité à évaluer la pertinence des informations, analyse Claire Leroy, du CXP.



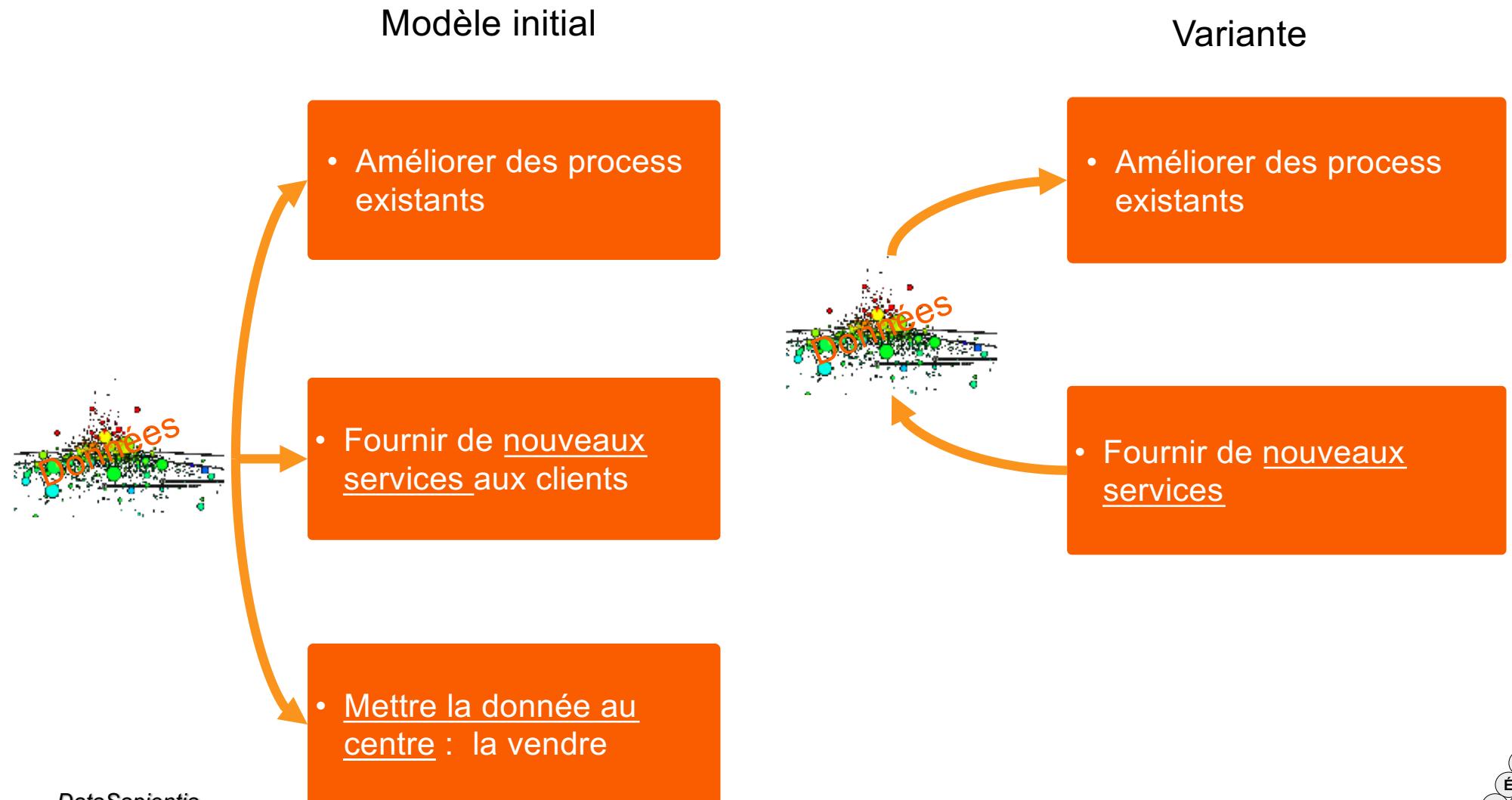
## Applications business : Toutes les fonctions/process de l'entreprise sont touchés



# Exemple thématique → les données clients (1/2)

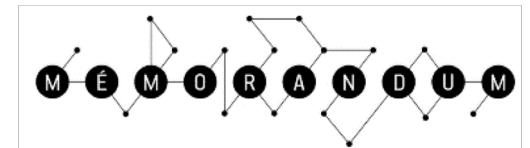


# D'une stratégie à l'autre : inverser la logique proposer des services qui produisent des données



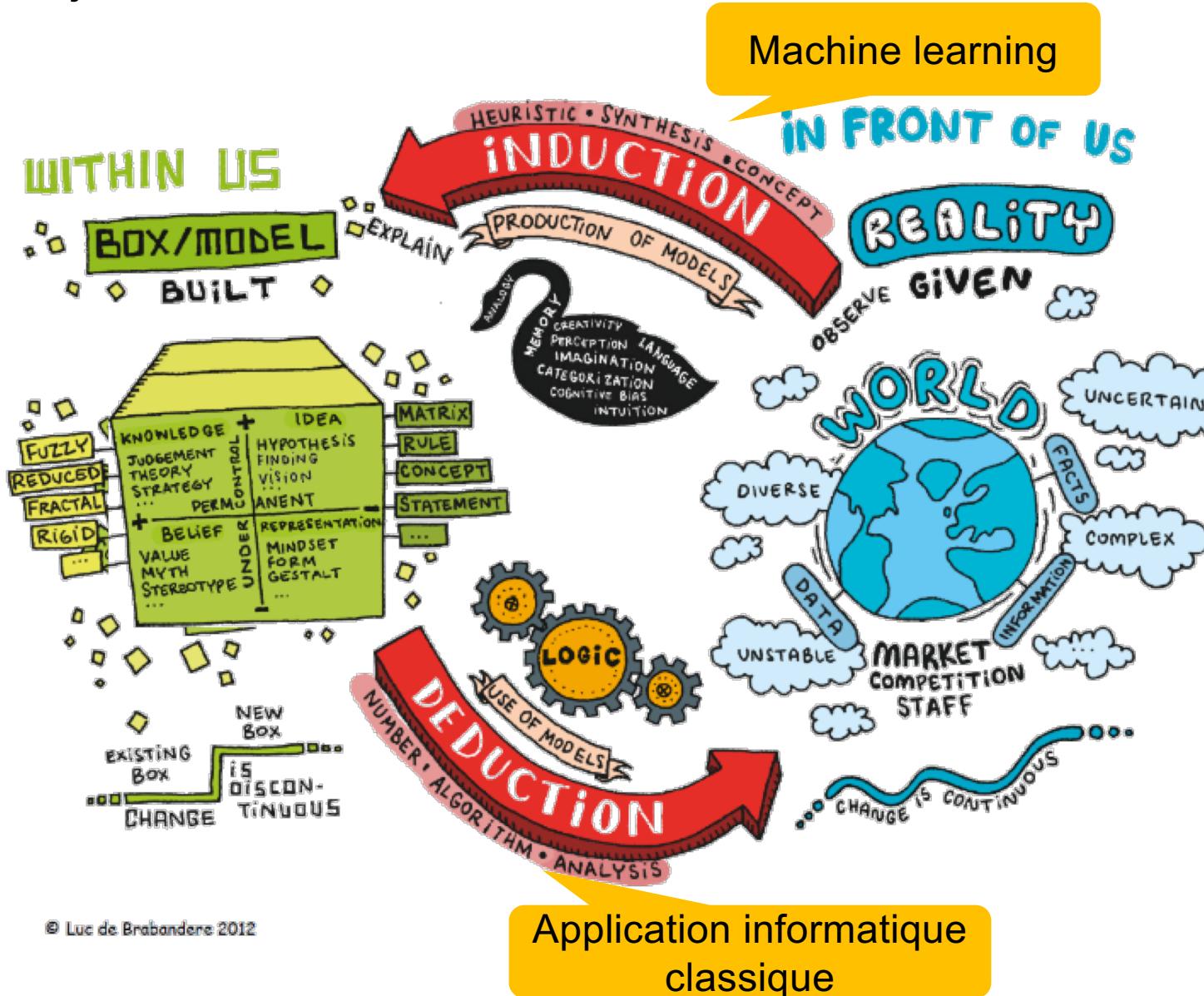
**Et vous qu'avez-vous fait avec vos  
données aujourd'hui ?**

11 avril 2016



Comment ca marche ?

Datascience : ni plus ni moins ce que fait votre cerveau tous les jours

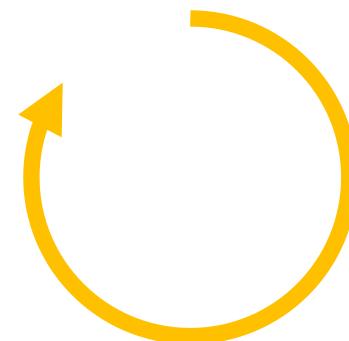


## Comment ca marche ? Datasience : un art culinaire

... ne rompez pas la  
chaine du froid

Mélangez vos différentes  
préparations entre elles en  
fin de cuisson

Testez tous les réglages  
possibles pour choisir ce qui  
marche le mieux pour vos  
données



prenez vos données au  
niveau granulaire

Utilisez plusieurs recettes bien  
rodées (toutes disponibles  
gratuitement !)

Préférez celles qui savent s'en  
remettre (un peu) au hasard

Maîtriser l'erreur d'apprentissage  
 → Etape 4 : philosopher

**No silver bullet** : Impossible de viser juste du premier coup

**No free lunch theorem** : Il n'y a pas de meilleur classifieur dans l'absolu : tout changement de contexte impose un nouveau réglage

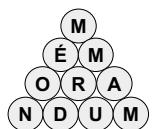
*Un algorithme qui améliore un score dans une zone .. Le dégrade dans une autre (exemple de l'overfitting)*

**Le machine learning est un travail d'artisan**

...

- Beaucoup de preprocessing (préparer l'information pour la rendre digeste)
- Des algorithmes paramétrables : à ajuster à la main
- Un étalonnage empirique

**... qui s'automatise de plus en plus et devient de plus en plus accessible à tout un chacun**



Un peu de vocabulaire: approches fréquentistes ou séquentielles  
(pour les modèles paramétriques)

Input

- Un modèle paramétrique de distribution statistique (gaussienne par exemple) D
- Une distribution de valeurs observées : X

**Approche fréquentiste** : maximiser la vraisemblance du modèle sachant D

- trouver les paramètres de D qui maximisent  $p(X|D)$
- risque inhérent d'overfitting

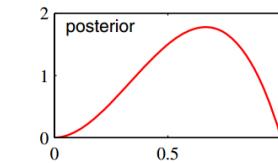
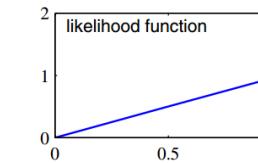
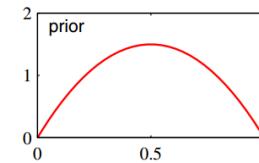
**Approche séquentielle (bayésienne)** : partir d'une distribution a priori des paramètres et corriger avec les valeurs observées

posterior

Vraisemblance  
Likelihood

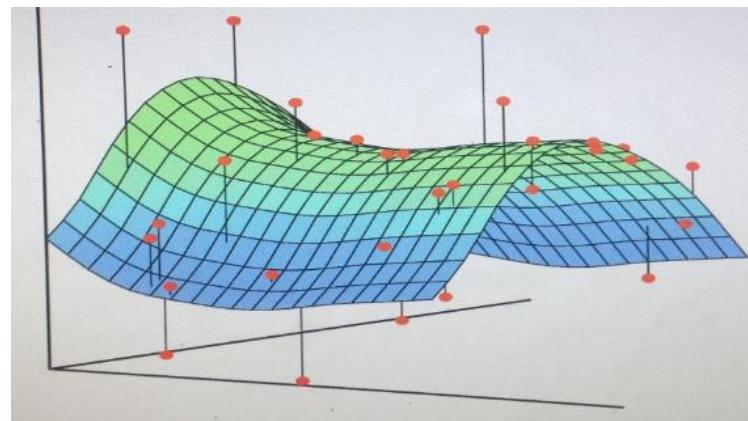
prior

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})}$$

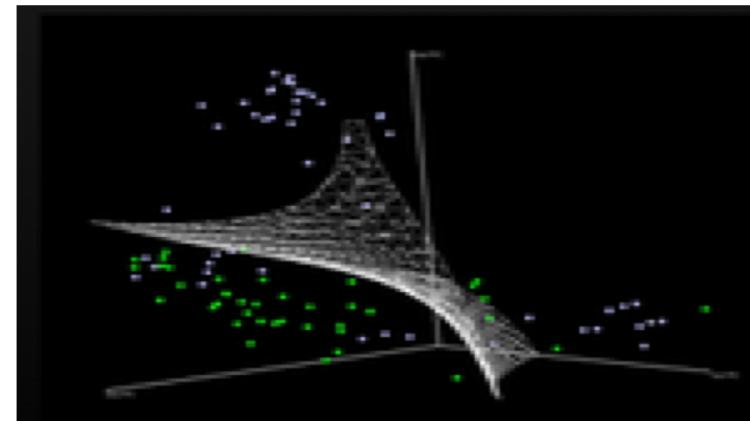


## Apprentissage supervisé : deux grandes familles

**Quantitatif : régression,**



**Qualitatif : classification**



Objectif

Réunir les points dn un même hyperplan    Séparer les observations le plus proprement  
(droite si une feature, plan si 2 features, ... possible)

résultat

Une valeur numérique

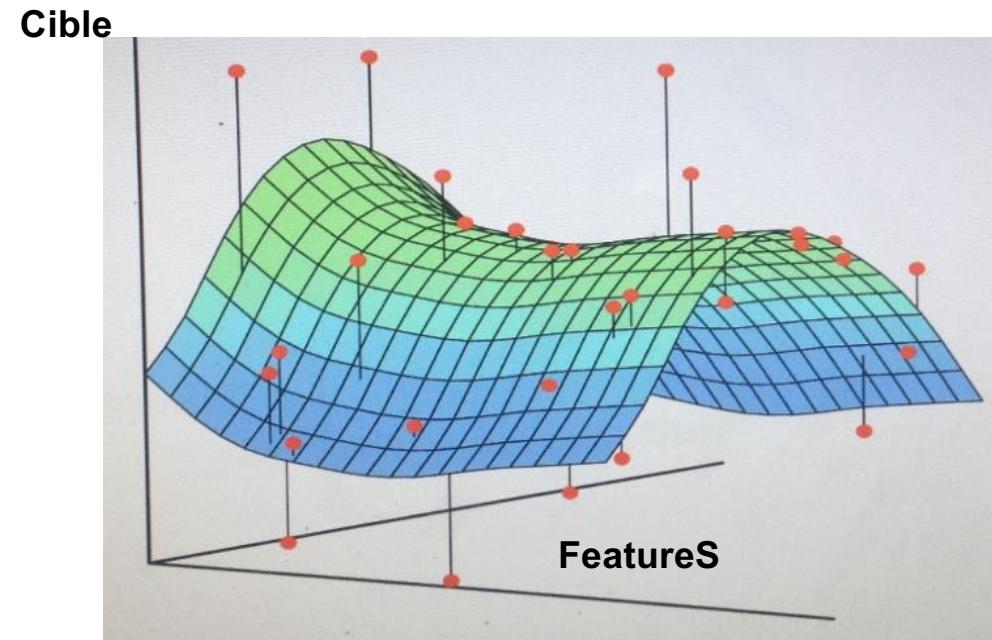
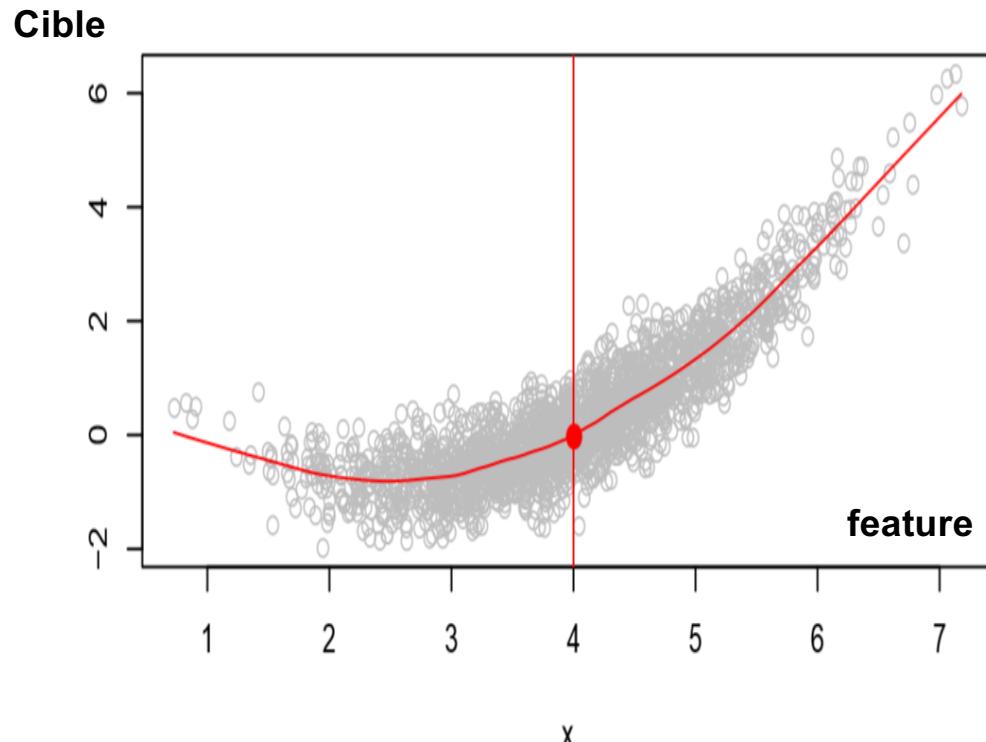
Une classe

## Vue globale régression

Approcher une variable quantitative en fonction de chacun des paramètres disponibles

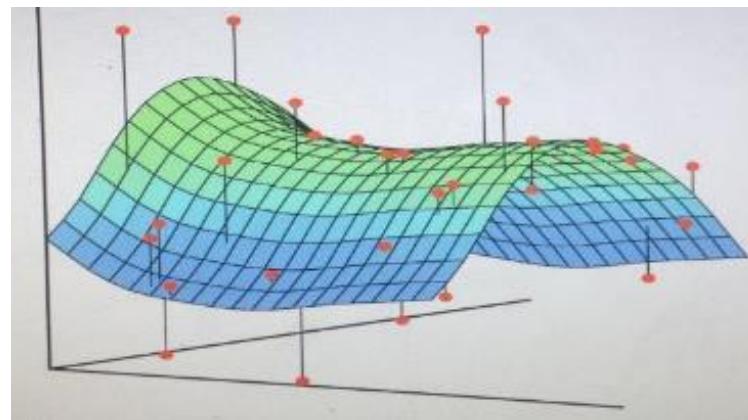
Fonction d'un espace de  $R^p \rightarrow R$

- Approcher une variable quantitative en fonction de chacun des paramètres disponibles
- Fonction d'un espace de  $R^p \rightarrow R$

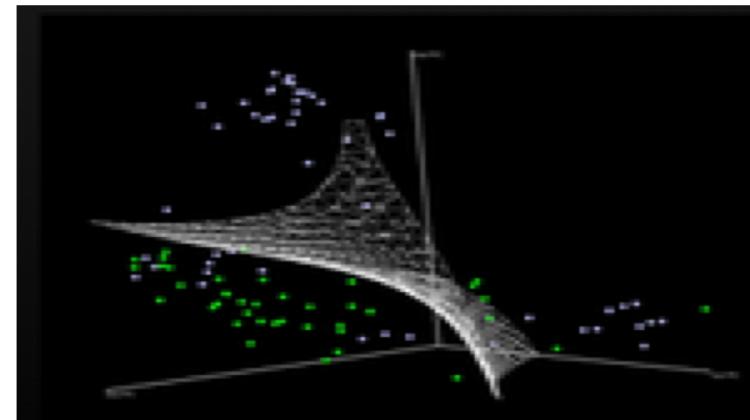


# Apprentissage supervisé : deux grandes familles

**Quantitatif : régression,**



**Qualitatif : classification**



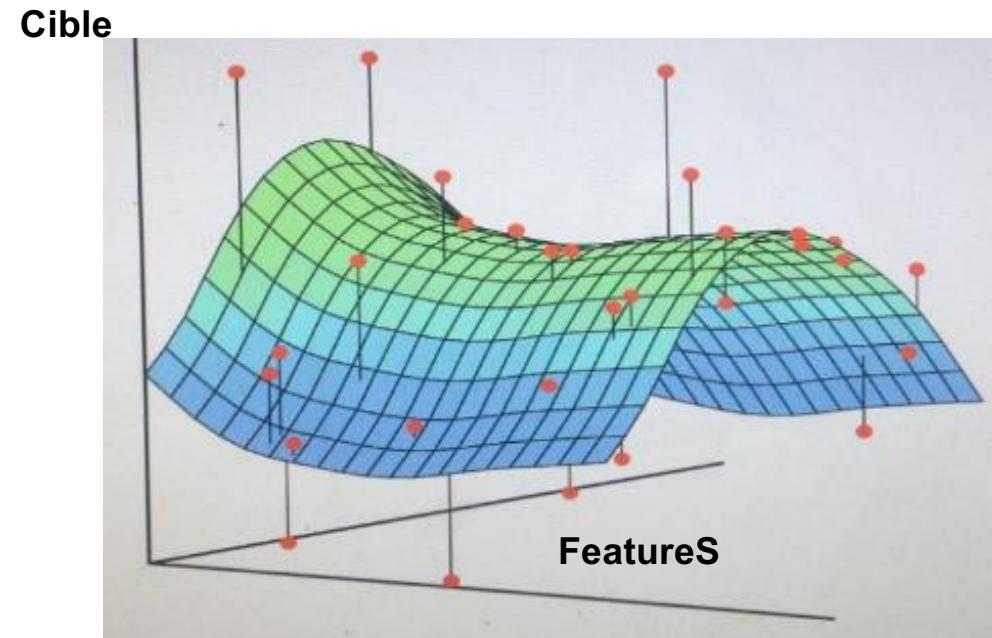
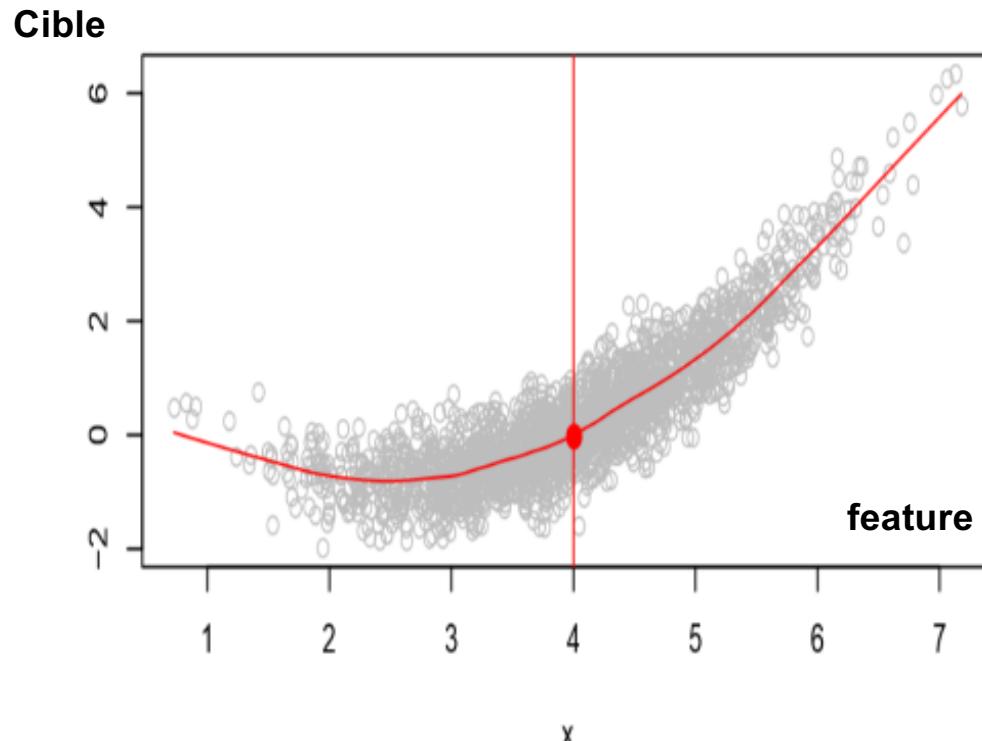
Objectif	<u>Réunir</u> les points dn un même hyperplan (droite si une feature, plan si 2 features, ... possible)	<u>Séparer</u> les observations le plus proprement possible
résultat	Une valeur numérique	Une classe

## Vue globale régression

Approcher une variable quantitative en fonction de chacun des paramètres disponibles

Fonction d'un espace de  $R^p \rightarrow R$

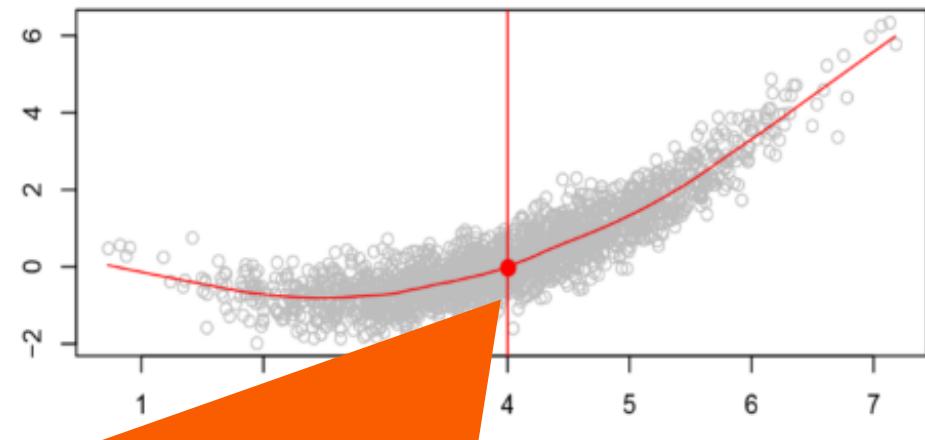
- Approcher une variable quantitative en fonction de chacun des paramètres disponibles
- Fonction d'un espace de  $R^p \rightarrow R$



## Le problème (1/2)

Idéalement

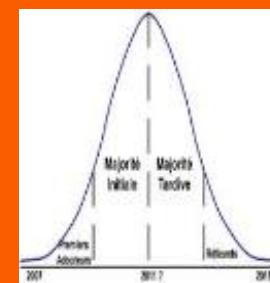
- Etre omniscient et avoir toutes les observations possibles
- Pour chaque valeur possible des features disponibles : prendre la moyenne des observations (espérance)



La dispersion autour de cette valeur moyenne peut être lié à plusieurs facteurs

Principal levier  
big data

- Il manque des facteurs explicatifs → toujours
- Il y a des erreurs de mesure → toujours
- Il y a du vrai hasard → là c'est de la philo



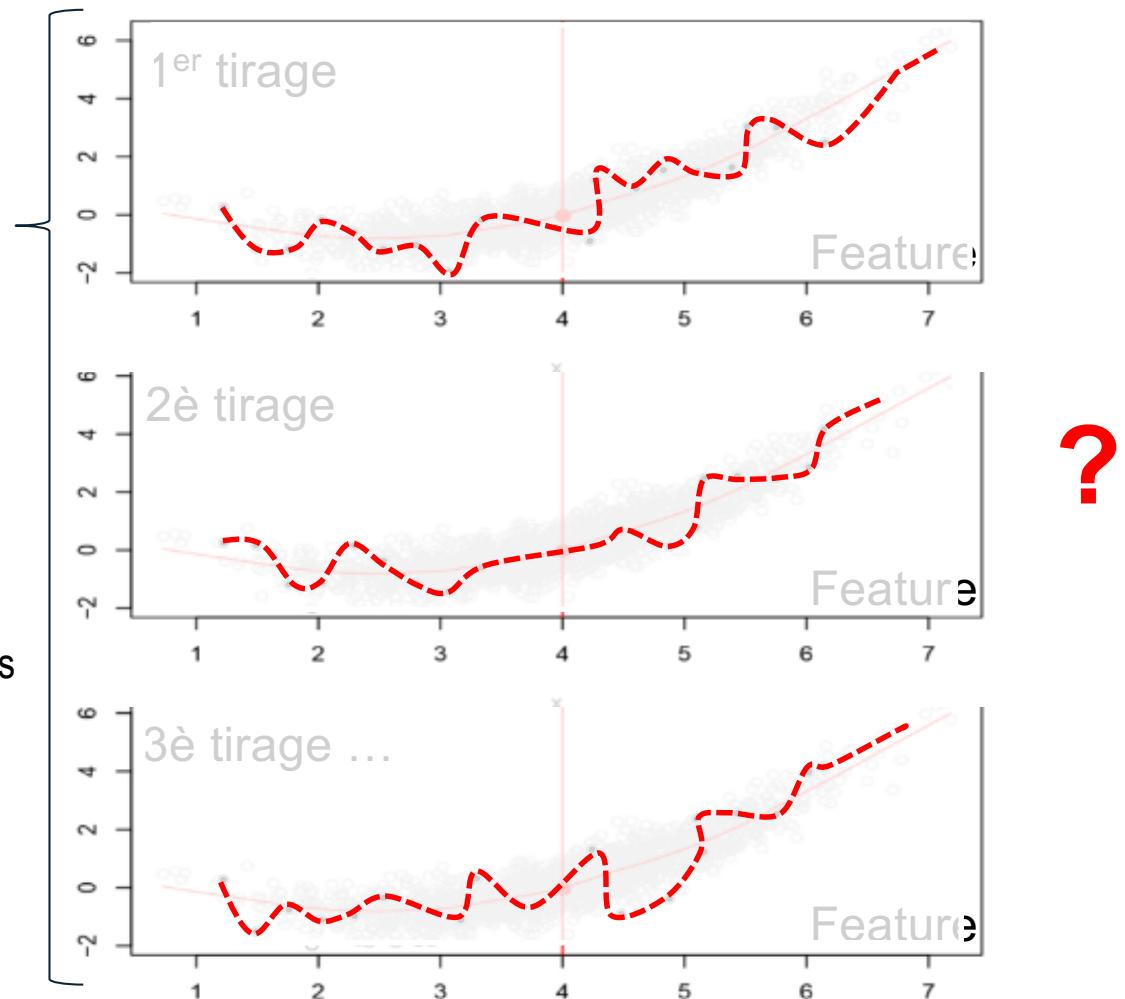
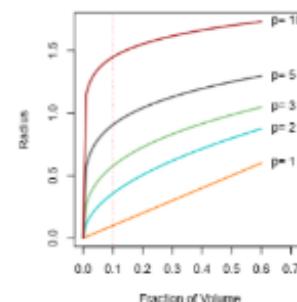
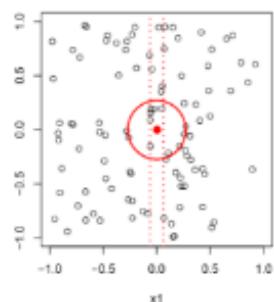
## Le problème (2/2)

**Hélas:**

- Vous ne disposez que d'un jeu de données partiel et si vous renouvez les mesures vous aurez chaque fois un autre jeu d'observation

**Hélas (bis)**

- Vous avez beaucoup d'observations.. mais encore plus de features pour chaque observation : vos êtes atteint par la malédiction de la dimension (« curse of dimensionality ») :
- dans un espace à haute dimension, vos observations sont éclatées : il n'y a plus de voisins ..



## Une démarche pleine de bon sens 1/3

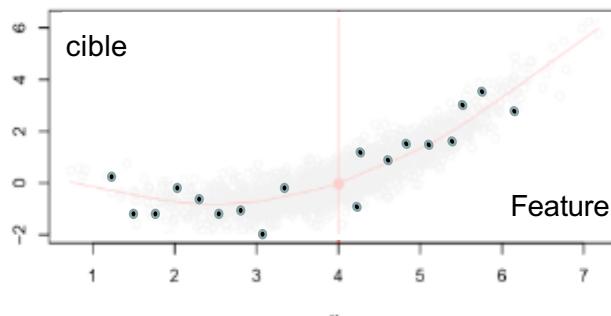
Pour bien prévoir le futur nous pouvons **simplifier le passé**

Equivalent

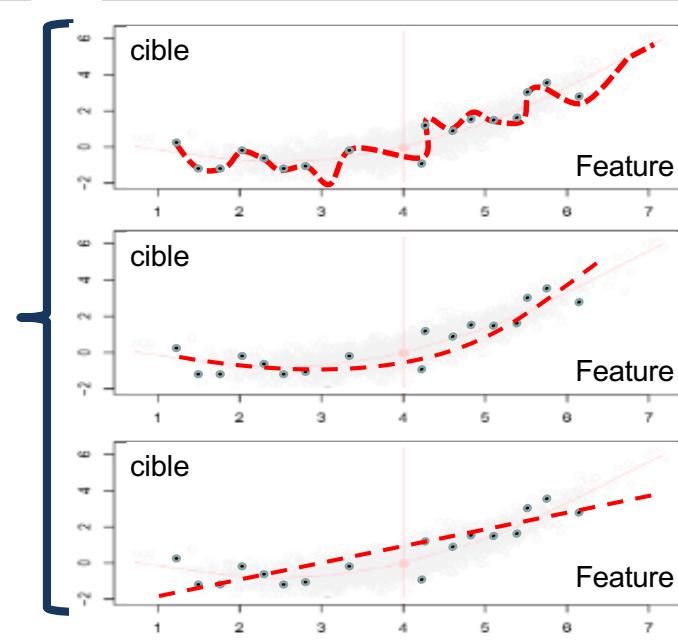
Distinguer

- le **signal** : « vraie » information apportée par les features disponibles
- du **bruit** : effet des informations (features) qui nous manquent

Données initiales



Modélisation induite



### Complex

- Parfaite description du passé
- Faible pouvoir prédictif
- « overfitting »

Un juste milieu ?

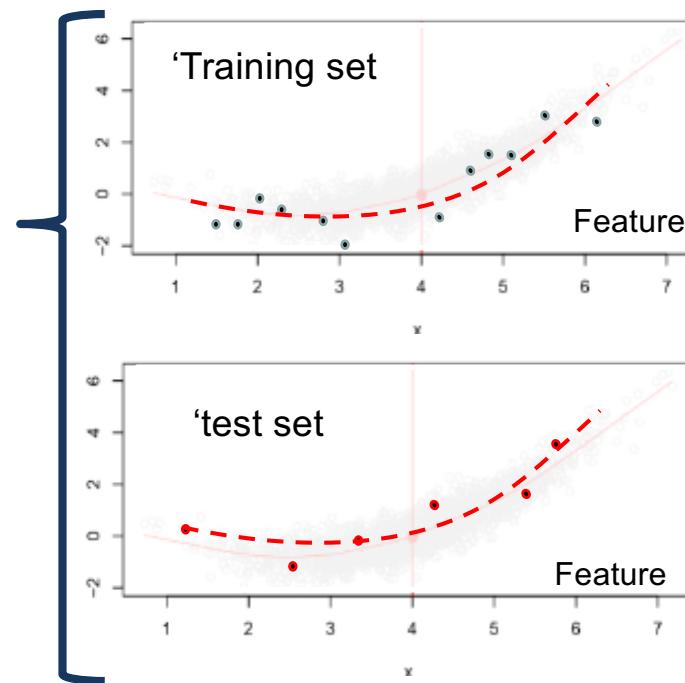
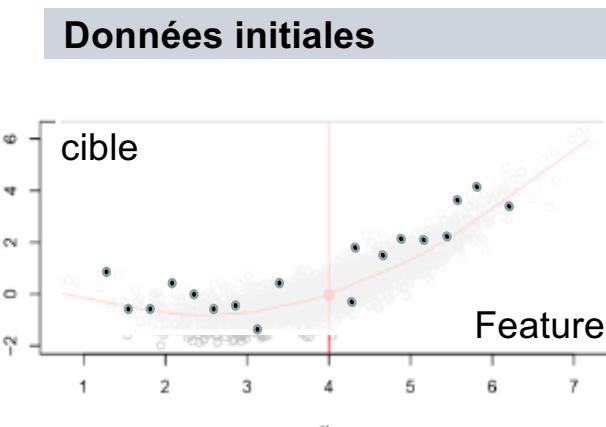
### Simple

- Grossière description du passé
- Faible pouvoir prédictif

## Une démarche pleine de bon sens 2/3

2<sup>e</sup> astuce : appliquer l'adage qui dit qu'on ne peut être juge et parti (séparation jeu d'apprentissage et jeu de test)

- L'évaluation de l'erreur d'interpolation des données connues n'est visiblement pas la métrique pertinente (sinon on va systématiquement pencher du côté « overfitting »)
- Solution « on ne peut pas être juge et partie »: les données connues sont réparties en deux lots
  - Un lot d'apprentissage
  - Un lot d'évaluation

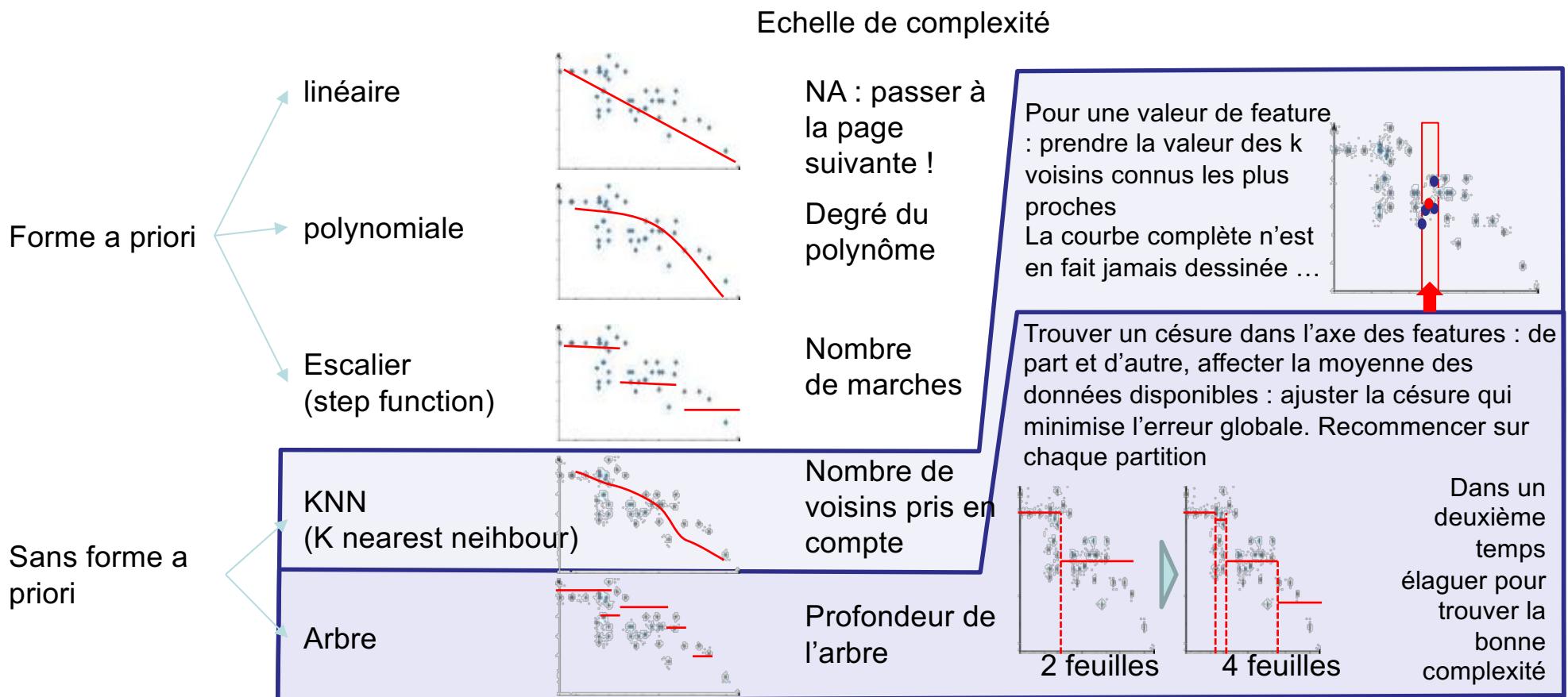


Construction d'un modèle pur un niveau de complexité donné

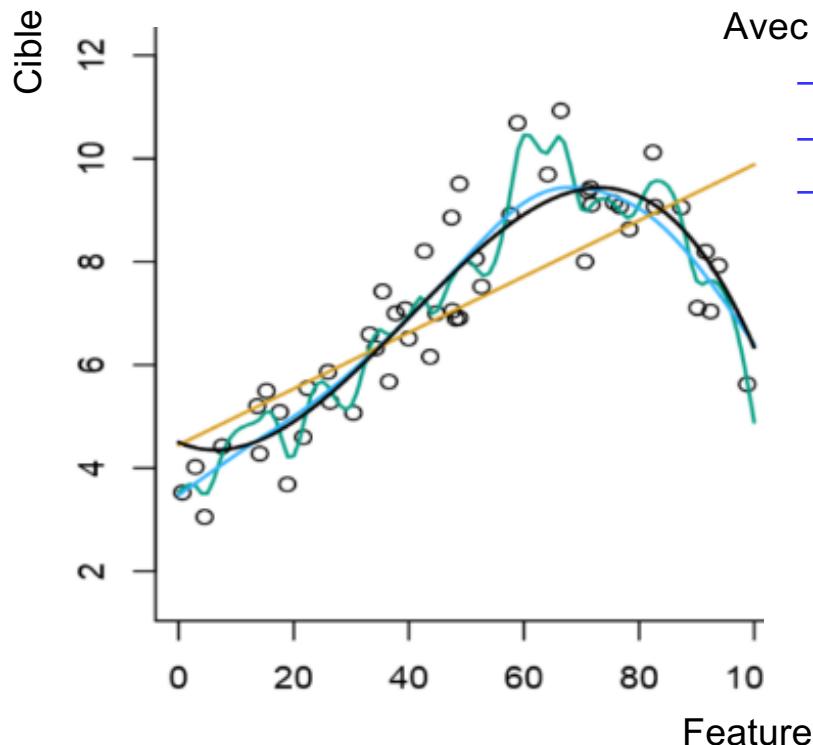
Evaluation : le modèle construit est évalué avec les points qui n'ont pas servi à la construction

## Une démarche pleine de bon sens 3/3

Ce choix de forme est un a priori, potentiellement guidé par la visualisation des données ou par l'expérience (attention : expérience est une lanterne dans le dos !)

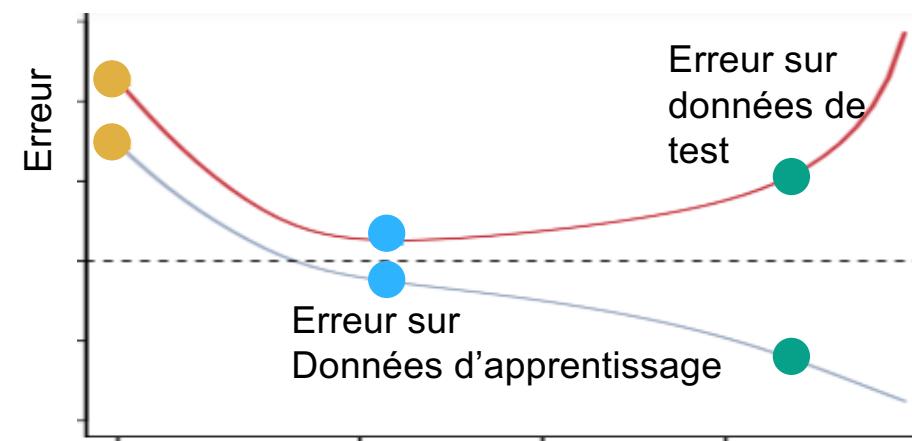


## Zoom : effet de la complexité sur l'erreur de prédiction (1/2)



Avec un jeu de données, 3 modèles sont représentés sur ce graphe

- Un modèle complexe : s'approche de près des données observées
- Un modèle plus simple
- Un modèle grossier (ligne droite)

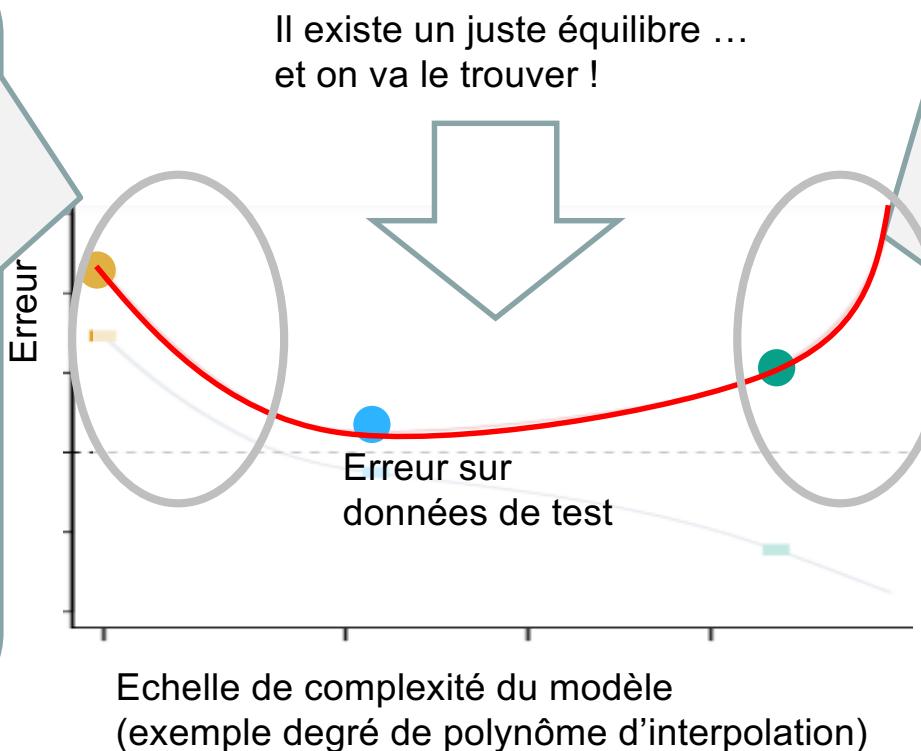
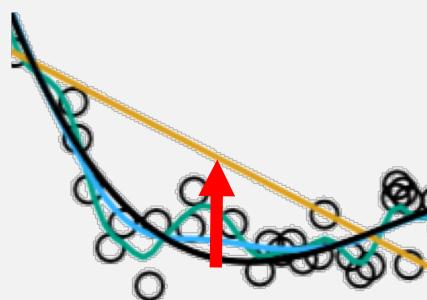


Echelle de complexité du modèle  
(exemple degré de polynôme d'interpolation)

## Zoom : effet de la complexité sur l'erreur de prédiction (2/2)

### Erreurs de biais

- Lié à la « raideur » du modèle
- peu sensible au jeu de données disponibles



- ### Erreurs de variance
- Lié à la contingence des données disponibles : juste en moyenne par définition mais toujours versatiles !

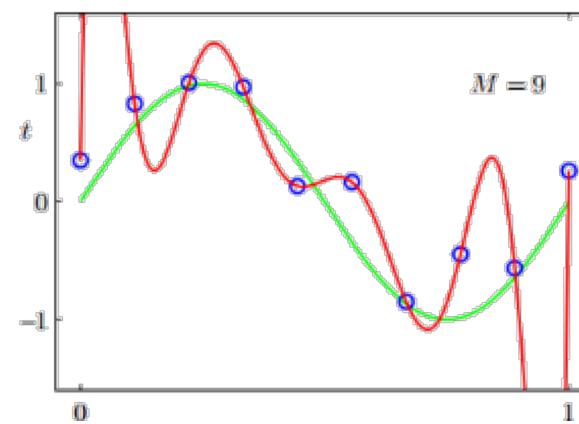
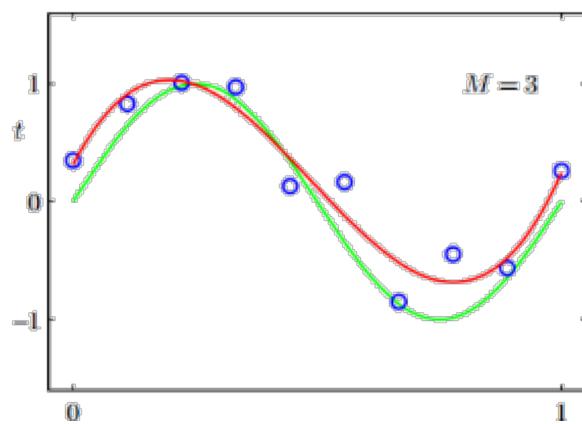
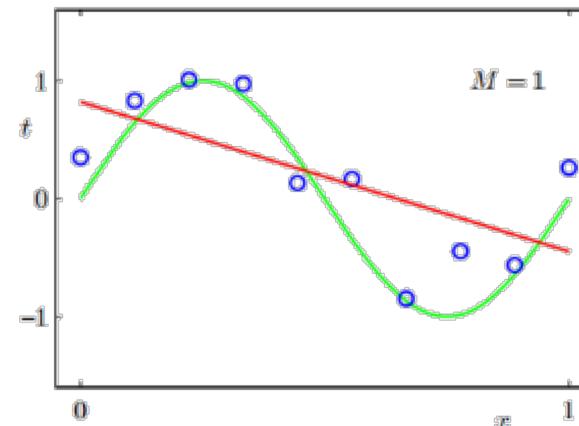
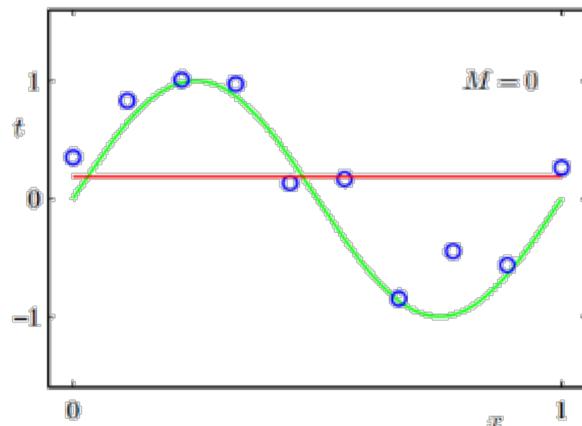


Le bon réglage de la simplification du passé n'est pas fournie par les mathématiques c'est un travail d'artisan !

## Exemple : approximation par une fonction polynomiale

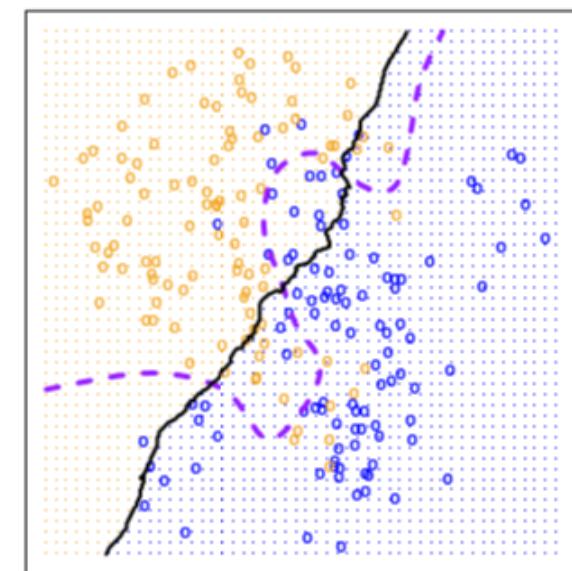
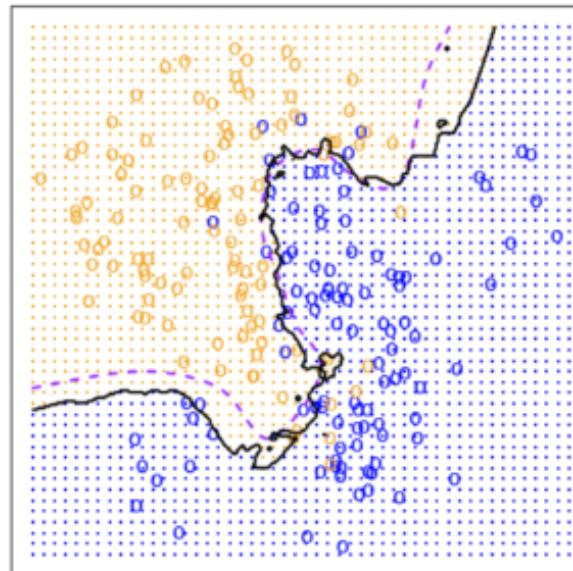
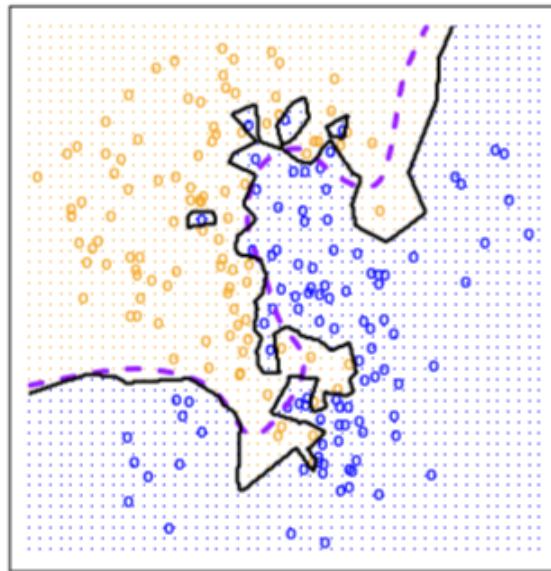
$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M = \sum_{j=0}^M w_j x^j$$

$$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^N \{y(x_n, \mathbf{w}) - t_n\}^2$$



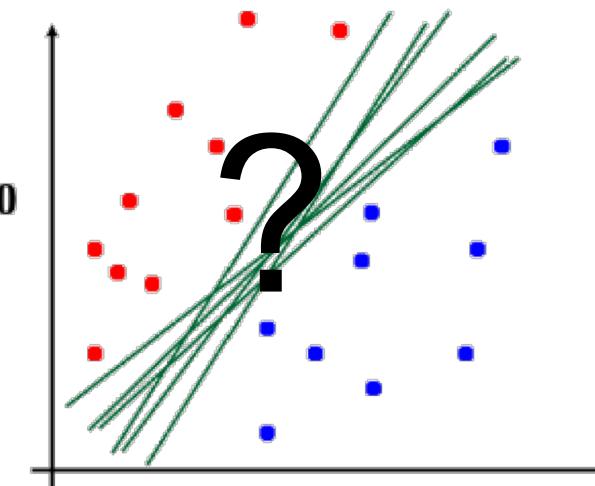
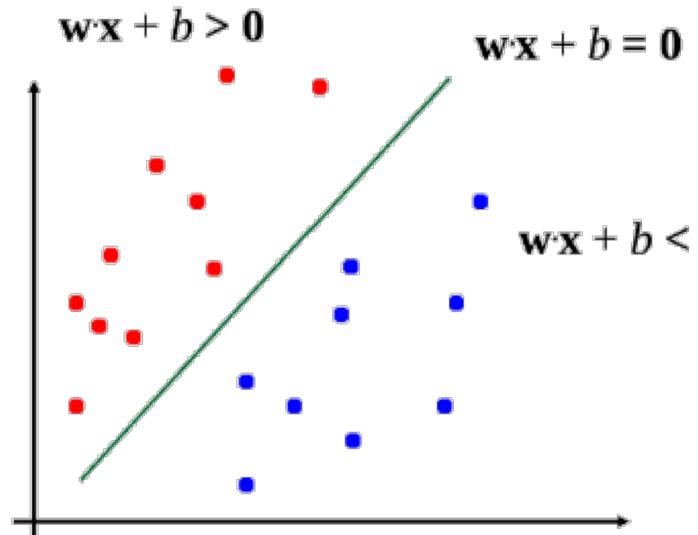
	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^*$	0.19	0.82	0.31	0.35
$w_1^*$		-1.27	7.99	232.37
$w_2^*$			-25.43	-5321.83
$w_3^*$			17.37	48568.31
$w_4^*$			-231639.30	
$w_5^*$			640042.26	
$w_6^*$			-1061800.52	
$w_7^*$			1042400.18	
$w_8^*$			-557682.99	
$w_9^*$			125201.43	

## K nearest neighbour



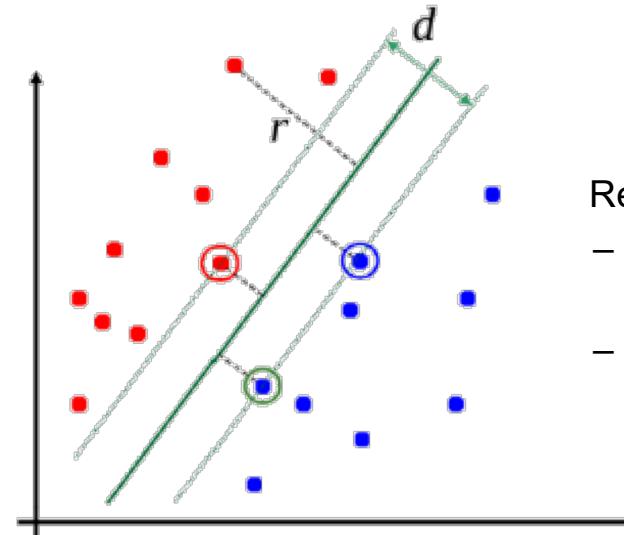
## SVM : Support vector machine / séparateur à vaste marge

Initialement le perceptron juge simplement l'erreur induite sur le jeu de test ne conduisant pas à un optimum unique



Une approche développée par Vapnik en 95

- Mathématiquement : démarche d'optimisation sous contrainte, faisant intervenir une transformation de Lagrange
- Intuitivement : quelle position d'hyperplan donne une séparation avec la meilleure marge de sécurité en fonction des points connus



Réglages

- Fonction de coût (nombre de vecteur supports)
- noyau

# arbres

Principe : : séparation due l'espace de manière itérative, de plus en plus précise

Première étape :

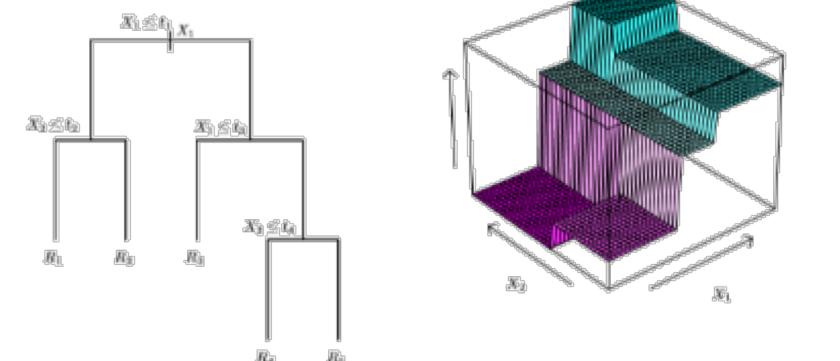
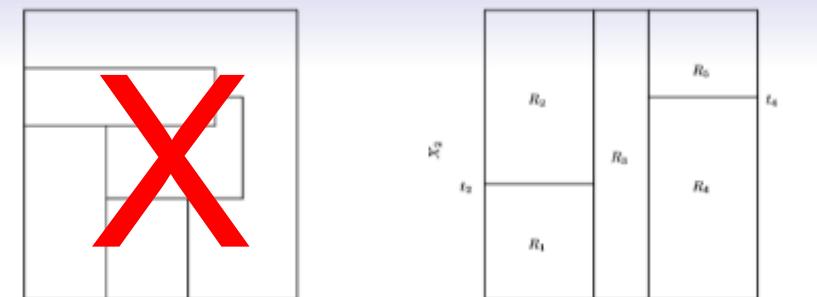
- Pour chaque feature, chercher le seuil qui sépare le mieux dataset (selon fonction de coût)
- Retenir la feature générant la meilleure séparation

Étapes suivantes

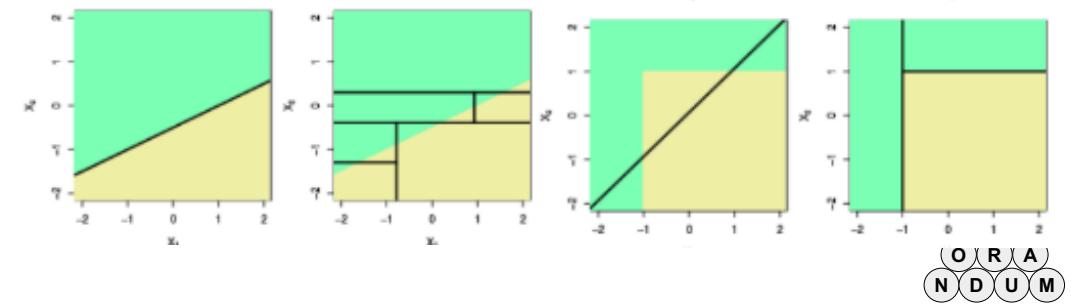
- Recommencer de la même manière sur chaque sous dataset obtenu à l'étape précédente

Cette démarche peut continuer jusque ce que les feuilles (sous dataset de la dernière itération soient des observations uniques

Puis on taille les branches (marche arrière) (« pruning ») et on s'arrête quand on atteint

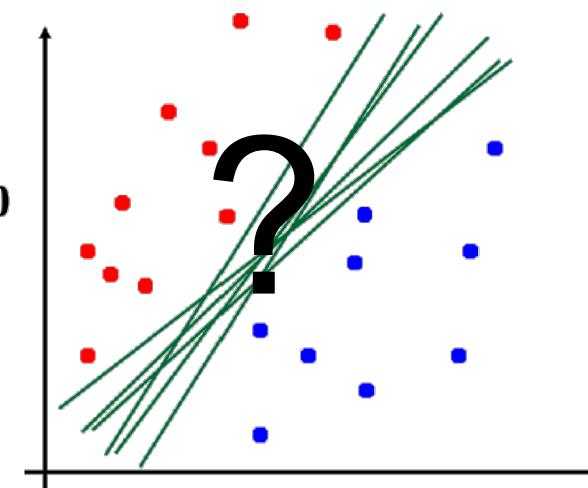
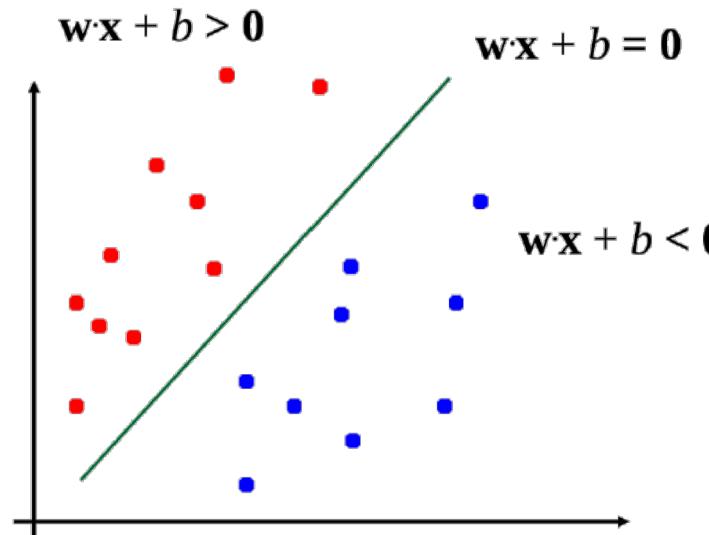


Arbre / régression linéaire



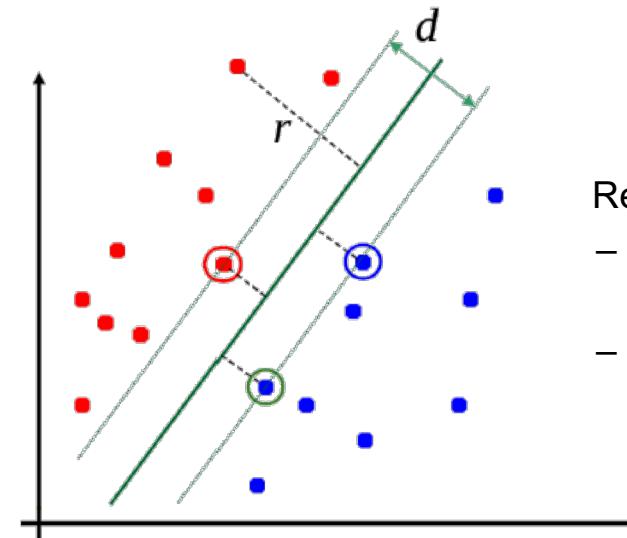
## SVM : Support vector machine / séparateur à vaste marge

Initialement le perceptron juge simplement l'erreur induite sur le jeu de test ne conduisant pas à un optimum unique



Une approche développée par Vapnik en 95

- Mathématiquement : démarche d'optimisation sous contrainte, faisant intervenir une transformation de Lagrange
- Intuitivement : quelle position d'hyperplan donne une séparation avec la meilleure marge de sécurité en fonction des points connus



Réglages

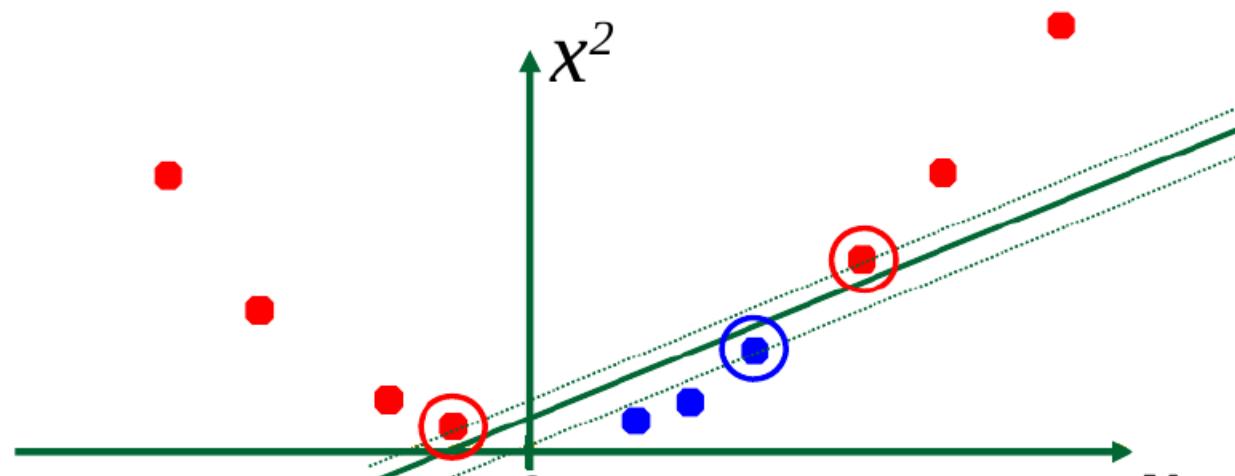
- Fonction de coût (nombre de vecteur supports)
- noyau

## SVM : rôle des noyaux

Que se passe-t-il si l'ensemble d'apprentissage est intrinsèquement non séparable ?



Pourquoi ne pas plonger le problème dans un espace de plus grande dimensionnalité ?



## arbres

Principe : : séparation due l'espace de manière itérative, de plus en plus précise

Première étape :

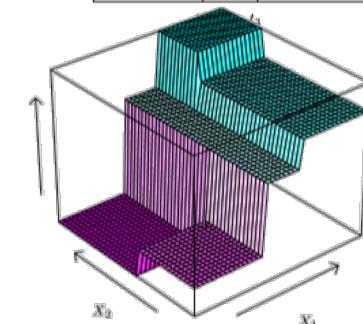
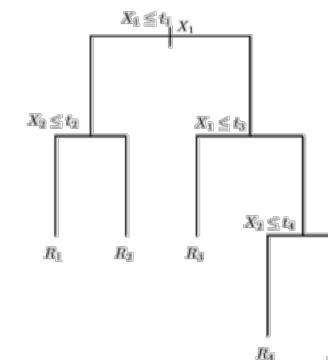
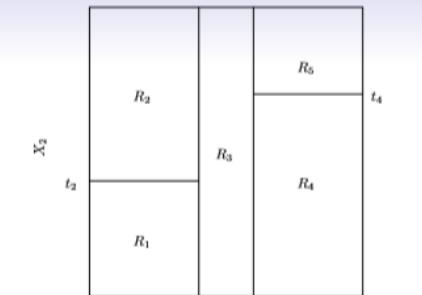
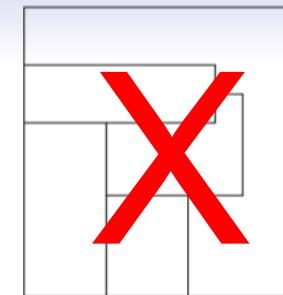
- Pour chaque feature, chercher le seuil qui sépare le mieux dataset (selon fonction de coût)
- Retenir la feature générant la meilleure séparation

Étapes suivantes

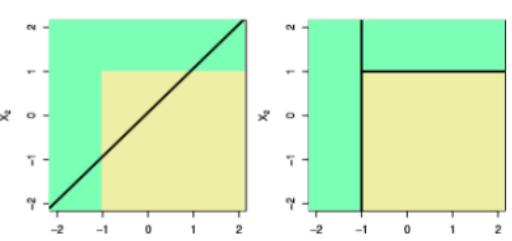
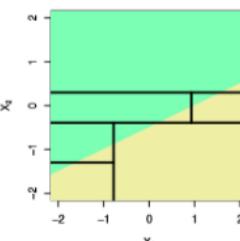
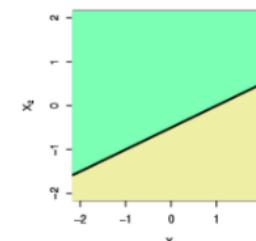
- Recommencer de la même manière sur chaque sous dataset obtenu à l'étape précédente

Cette démarche peut continuer jusque ce que les feuilles (sous dataset de la dernière itération soient des observations uniques

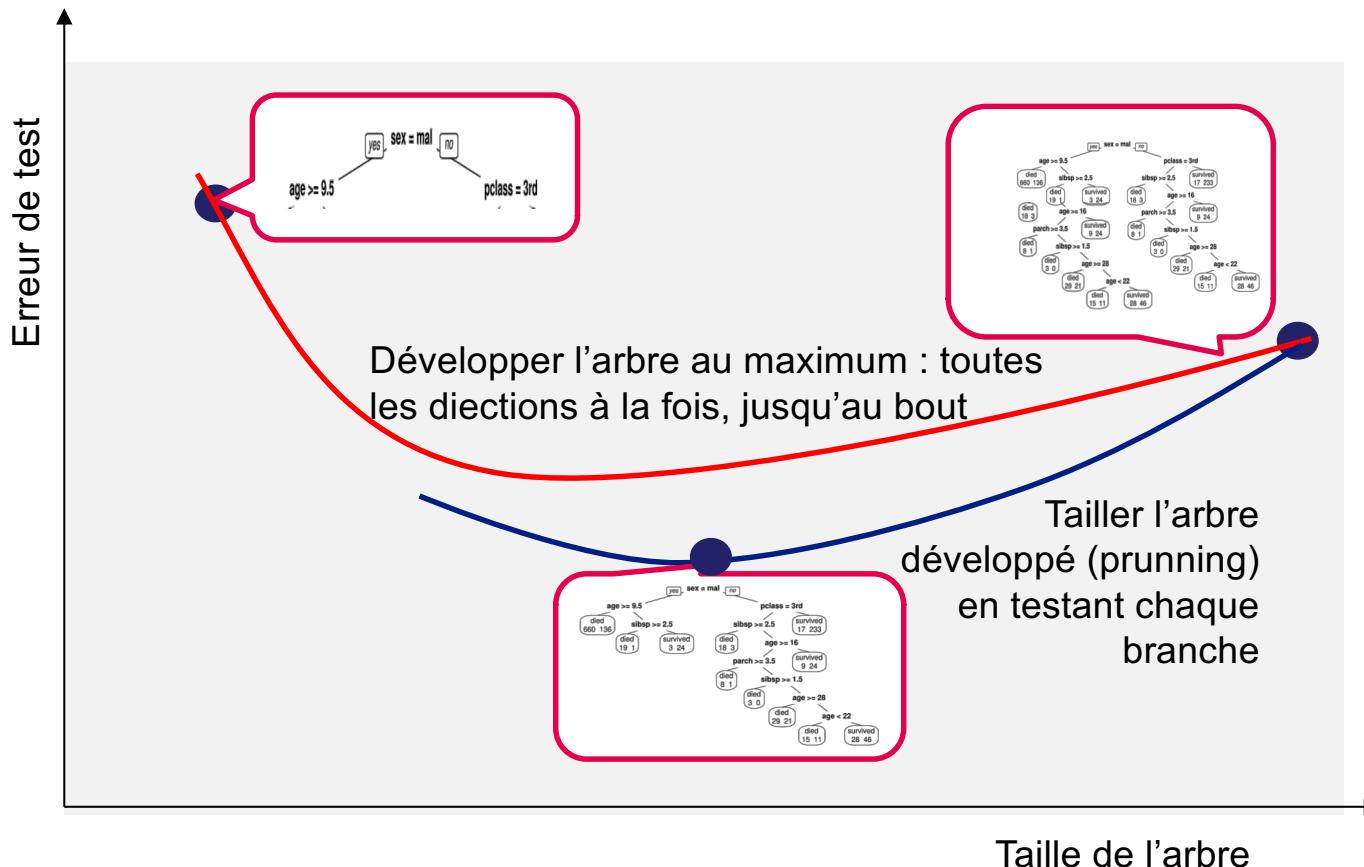
Puis on taille les branches (marche arrière) (« pruning ») et on s'arrête quand on atteint



Arbre / régression linéaire



## Arbre : algorithme cart



### Avantage :

- Peuvent attaquer des données massives
- Gère naturellement un mélange de données qualitatives et quantitatives
  - Qualitatif : comment faire ?
- Petits arbres faciles à interpréter

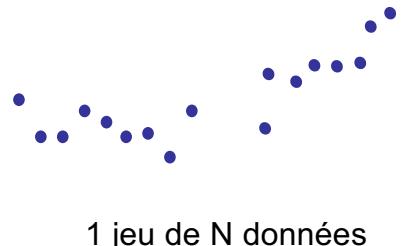
### Problèmes

- Arbres profonds difficiles à interpréter
- Performance limitée

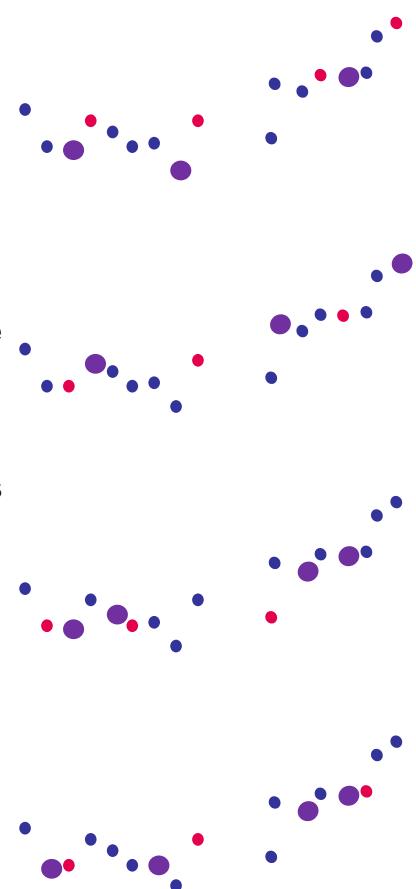
3 méthodes pour doper les arbres  
→ Bagging

Avec N données dans le jeu de train générer des centaines (ou milliers) de jeux différents

Comment faire ?



Tirage aléatoire de N données avec remise  
Certaines données ne seront pas tirées  
D'autres seront tirées 2 fois ou plus



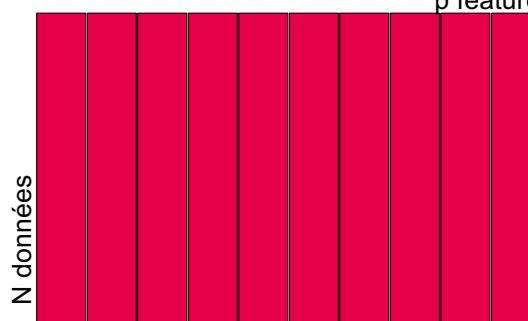
P jeux différents de N données  
Chacun donne naissance à un arbre spécifique  
Les résultats sont ensuite moyennés

## 3 méthodes pour doper les arbres → Random Forest

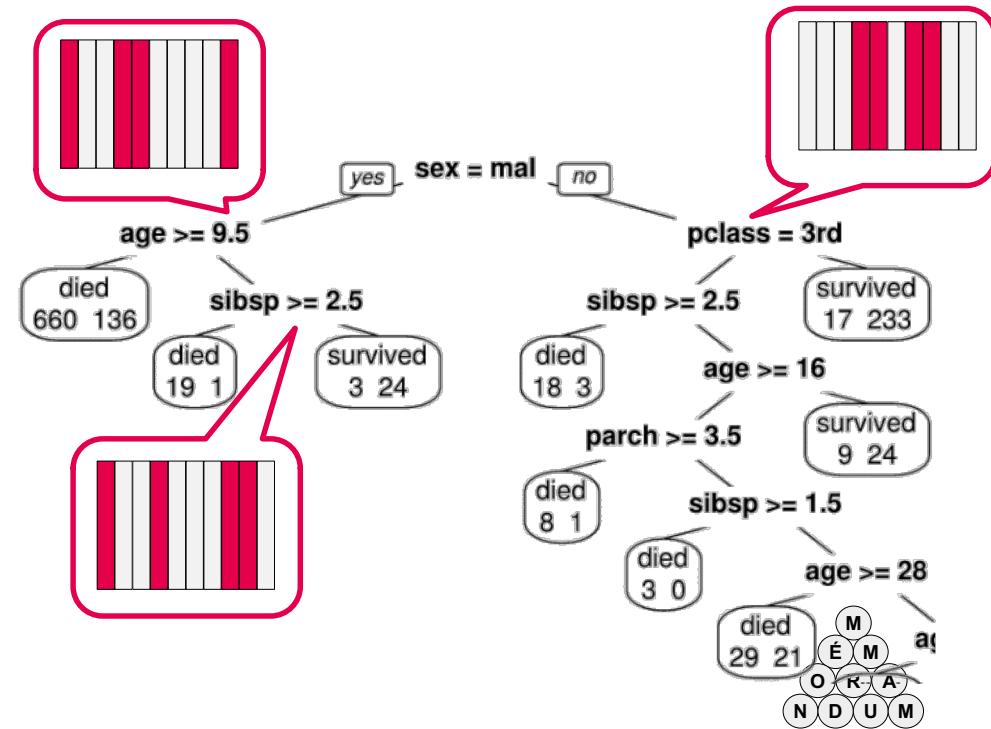
Objectif : renforcer l'indépendance de chaque arbre.

- Le levier bagging est maintenu en tirant au sort les points
- L'indépendance de chaque arbre est accentué en bridant à chaque nœud le périmètre des features pris en compte
  - De l'ordre de racine (nb de features) mais ce nombre fait l'objet d'un réglage

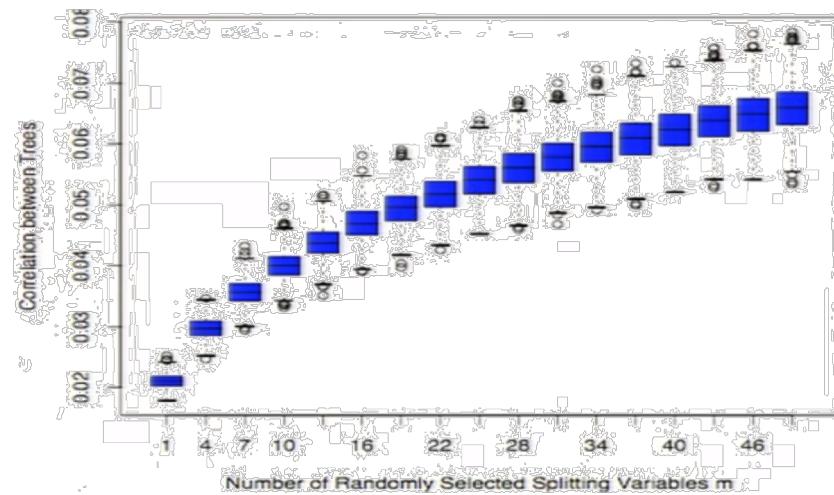
Données initiales  $p$  features



Exemple de tirage aux sorts de features

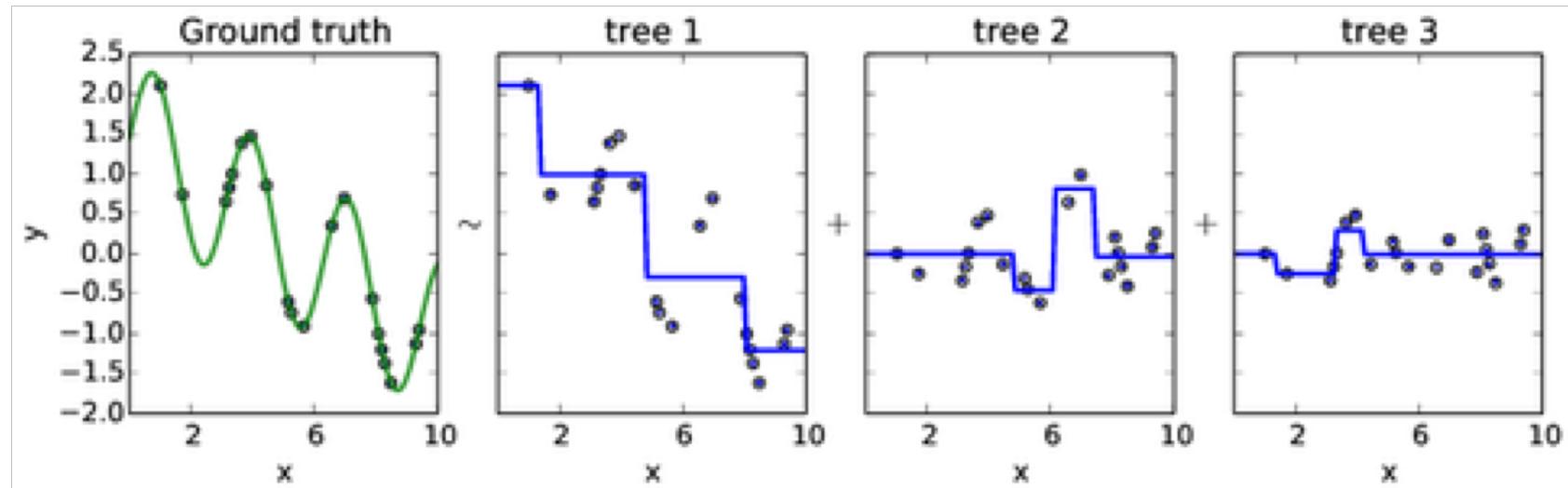


Impact du bridage des features sur la corrélation entre les arbres



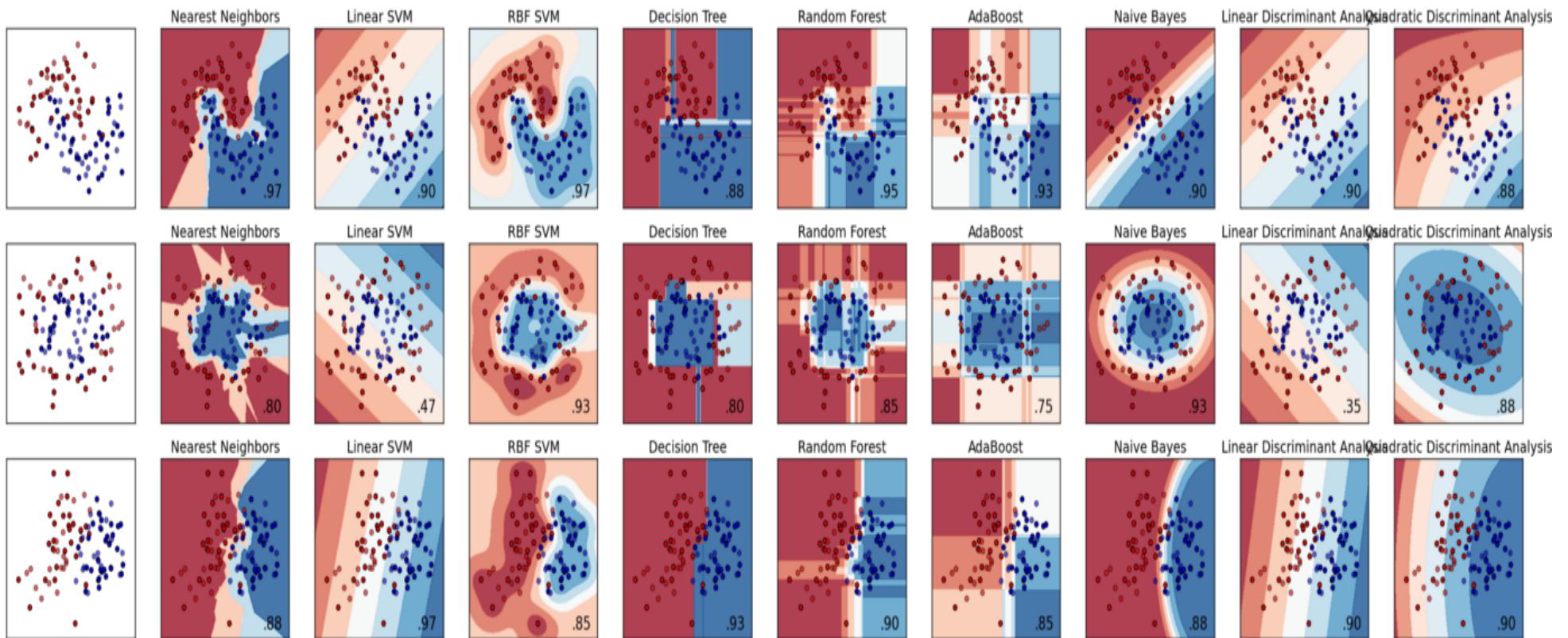
### 3 méthodes pour doper les arbres → Boosting

Cette fois les arbres ne sont plus indépendants, ils sont construits de manière séquentielle pour corriger les lacunes des précédents



1. Start with function  $F_0(x) = 0$  and residual  $r = y$ ,  $m = 0$
2.  $m \leftarrow m + 1$
3. Fit a CART regression tree to  $r$  giving  $g(x)$
4. Set  $f_m(x) \leftarrow \epsilon \cdot g(x)$  (**shrink it down!**)
5. Set  $F_m(x) \leftarrow F_{m-1}(x) + f_m(x)$ ,  $r \leftarrow r - f_m(x)$  and repeat step 2–5 many times

<http://fr.slideshare.net/DeepakGeorge5/decision-tree-ensembles-bagging-random-forest-gradient-boosting-machines>



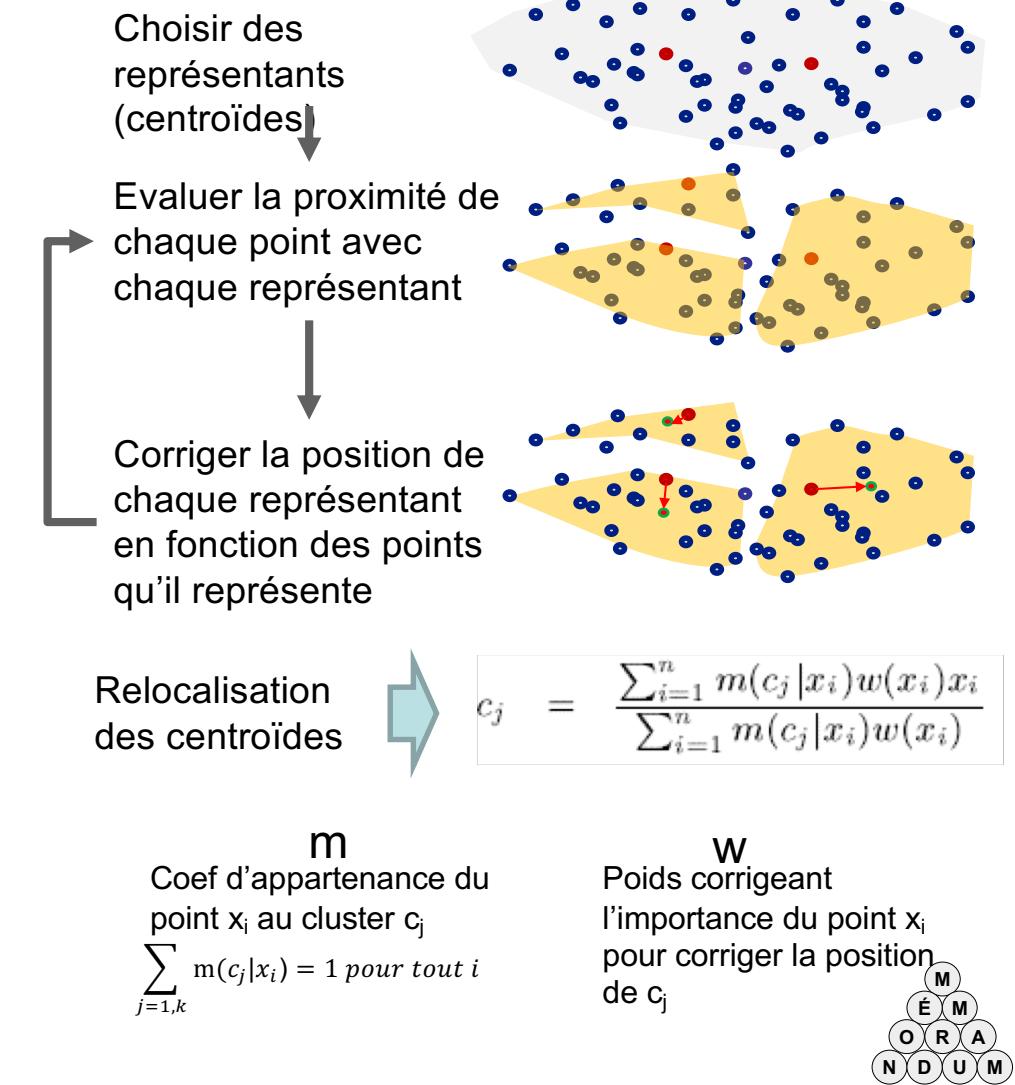
## Partitionnement à plat : k mean → principe

Une méthode itérative supposant de faire des choix aléatoires ...

- Nombre cible de segments(cluster)
- Représentants initiaux de chaque cluster (aléatoire le plus souvent)
- ... et des choix tactiques
- Fonction distance (pas forcément euclidienne)
- Affectation hard ou soft des points aux clusters
- Fonction de coût

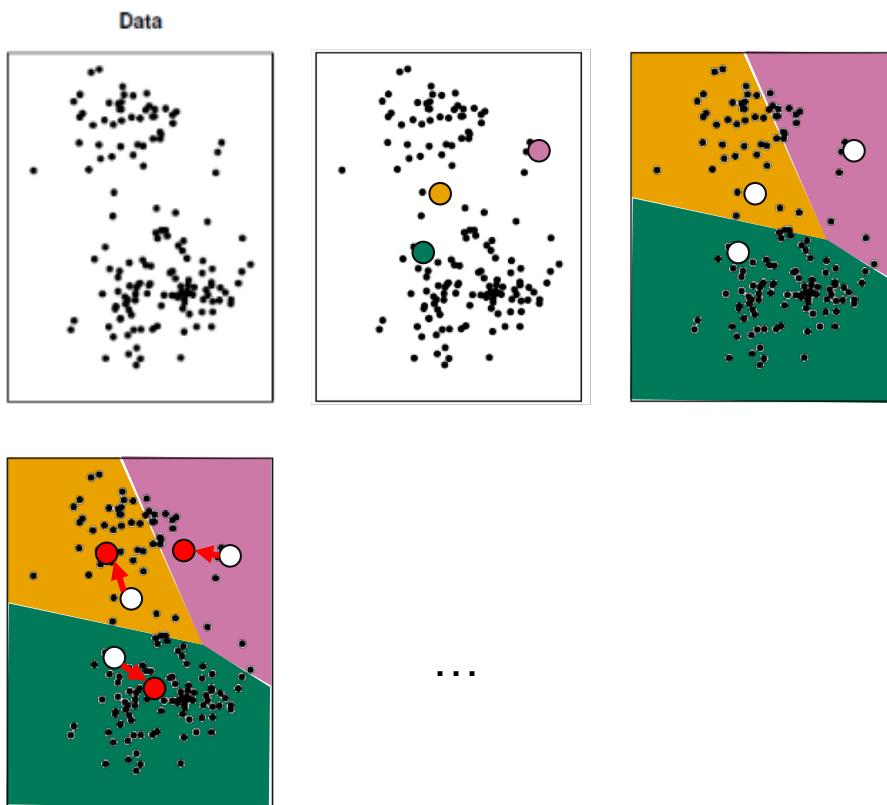
Plusieurs leviers pour optimiser l'algorithme :

- L'inspiration : Faire des choix judicieux d'initialisation (comment ?)
  - La force : faire beaucoup de tests et choisir le meilleur
  - La délicatesse : régler le moteur et le volant
- Si possible : les 3



Partitionnement à plat : k mean  
→ Comment initialiser ?

### Choix aléatoire des centroïdes initiaux



### Affectation aléatoire des clusters pour chaque point

