



**DATA**

Business.fr


# DATA SCIENCE

**BIG DATA | ANALYTICS | DATAVIZ**


[www.data-business.fr](http://www.data-business.fr)

# Machine learning ?


## Machine Learning



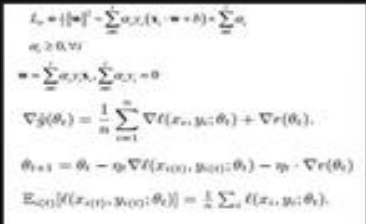
what society thinks I do



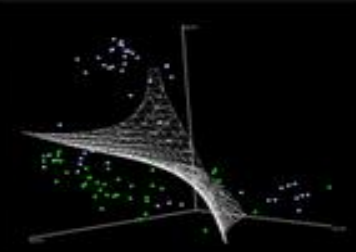
what my friends think I do




what my parents think I do



what other programmers think I do



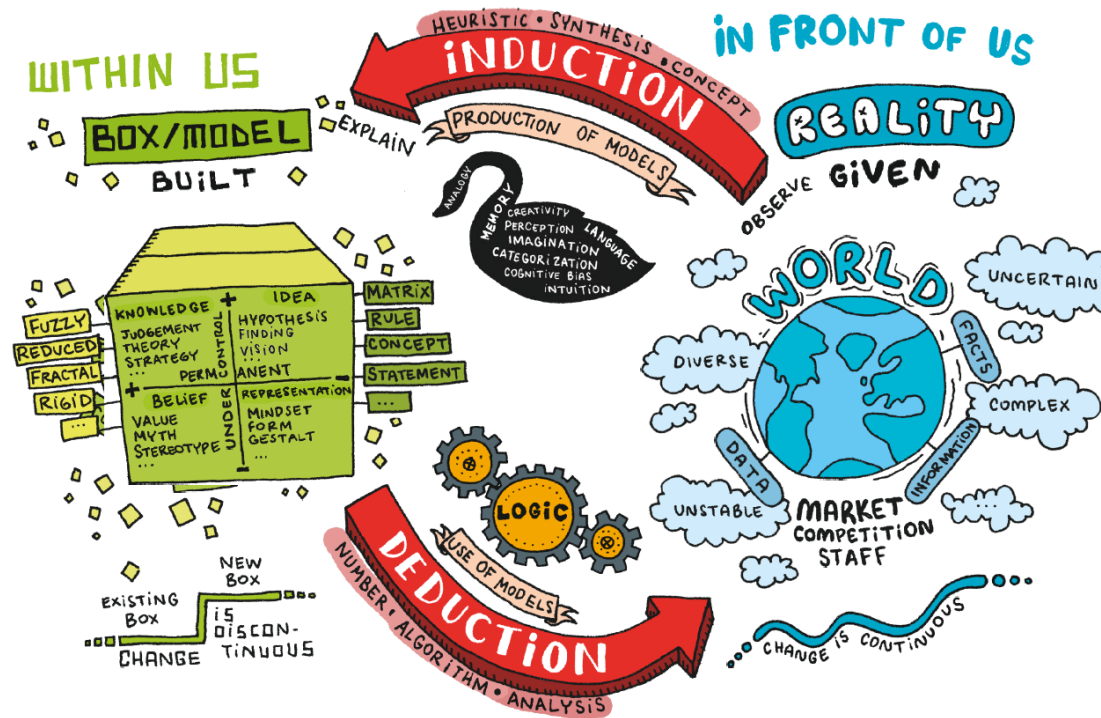
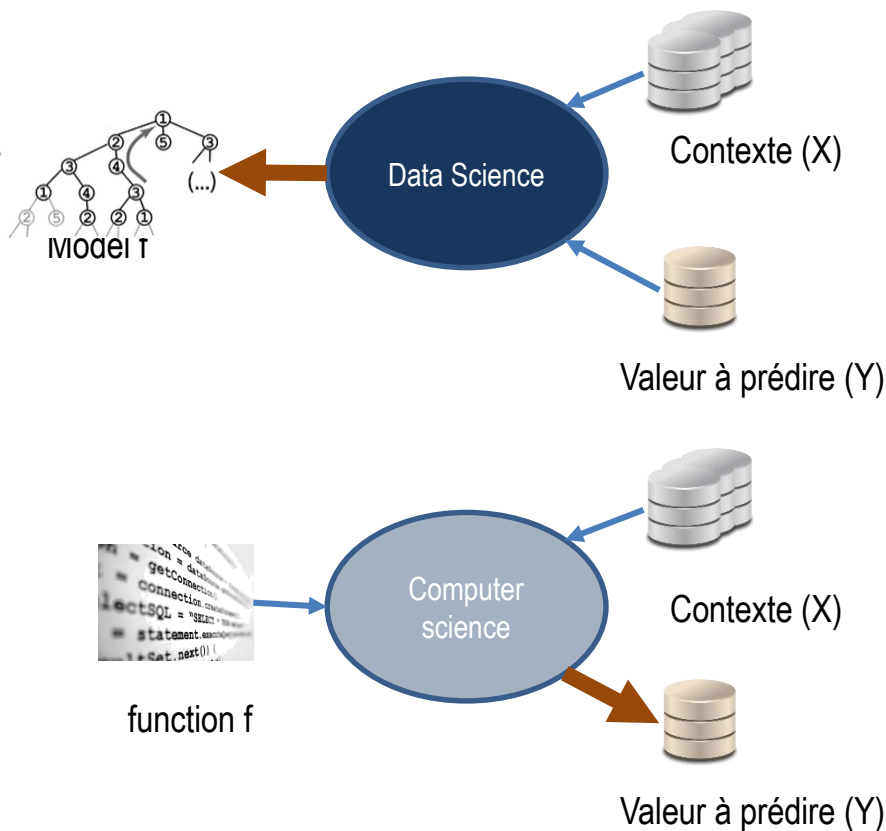
what I think I do



what I really do

# Construire des modèles un peu plus proches du monde réel

La datascience part du résultat (les valeurs réellement observées dans le monde) et cherche à en extraire une loi. Le modèle construit sera toujours une version simplifiée de la réalité



© Luc de Brabandere 2012

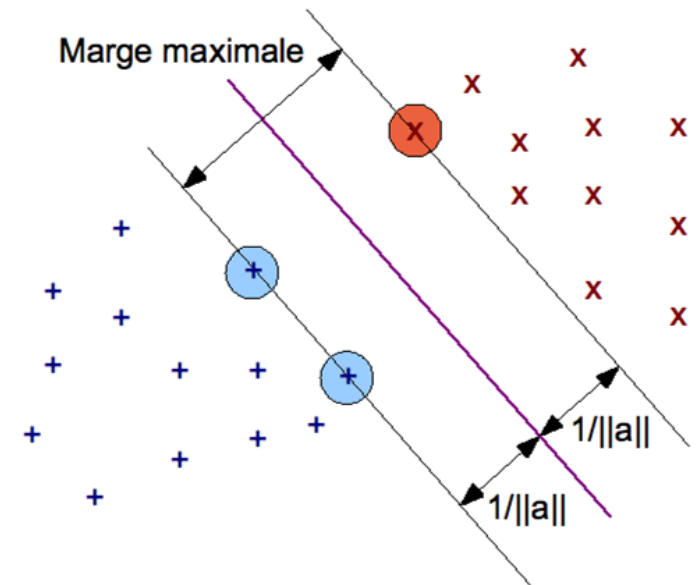
« *Le chiffre est lâche. Torturez le il vous avouera ce que vous voulez* »

*A la base des algorithmes : une alchimie ... dont voici la recette*

- prenez en compte toutes les données au niveau granulaire
- ... ne rompez pas la chaine du froid
- Utilisez plusieurs recettes bien rodées
- Mélangez vos différentes préparations entre elles en fin de cuisson
- Préférez celles qui savent s'en remettre (un peu) au hasard
- Testez tous les réglages possibles pour choisir ce qui marche le mieux pour vos données

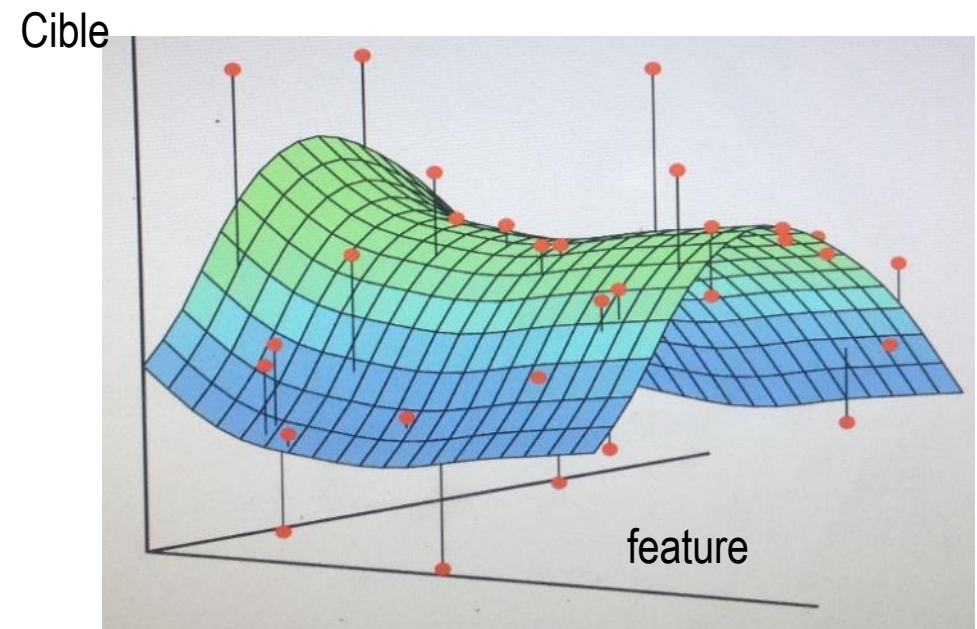
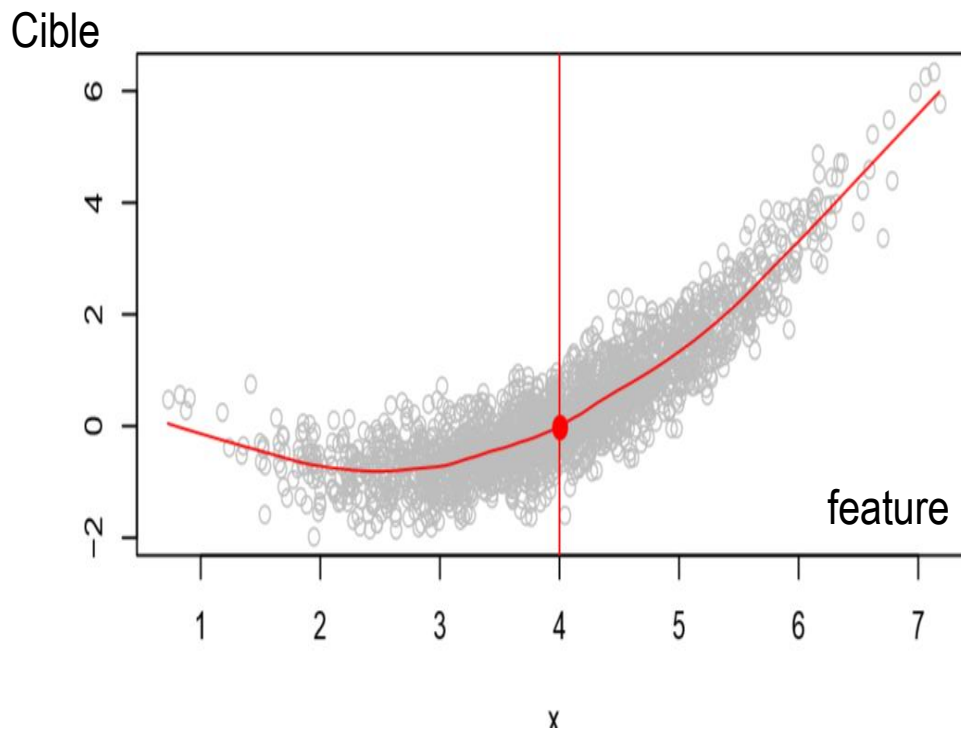
# Les algorithmes sont à votre portée

- Les mécaniques à l'œuvre dans un algorithme peuvent presque tout le temps être appréhendées sans formalisme mathématique particulier
- Les 'recettes bien rodées' sont disponibles gratuitement dans des packages d'outils open sources (R, Python, ...)
- Leur manipulation est de plus en plus simple avec l'émergence d'outil graphiques (Dataiku, azure ML, ..)
- La valeur ajoutée du datascientist réside largement dans son imagination en amont (travailler les données avant) et l'interprétation en aval (une courbe ROC n'est pas directement exploitable par une direction marketing)



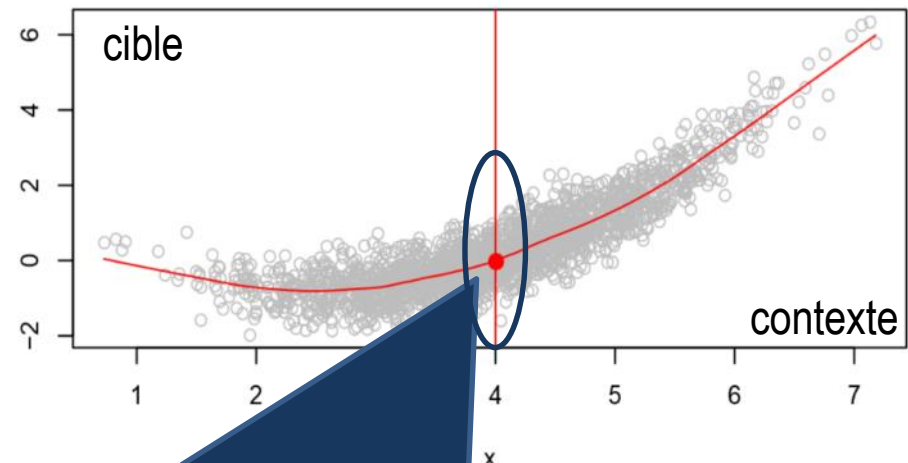
# Exemple d'approche : la régression

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



## Idéalement ...

- Etre omniscient et avoir toutes les observations possibles
- Pour chaque valeur possible des éléments de contextes connus : prendre la moyenne des observations (espérance)



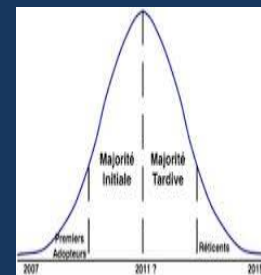
La dispersion autour de cette valeur moyenne peut être lié à plusieurs facteurs

Principal  
levier big data

**Il manque des facteurs** explicatifs → toujours

Il y a des **erreurs de mesure** → toujours

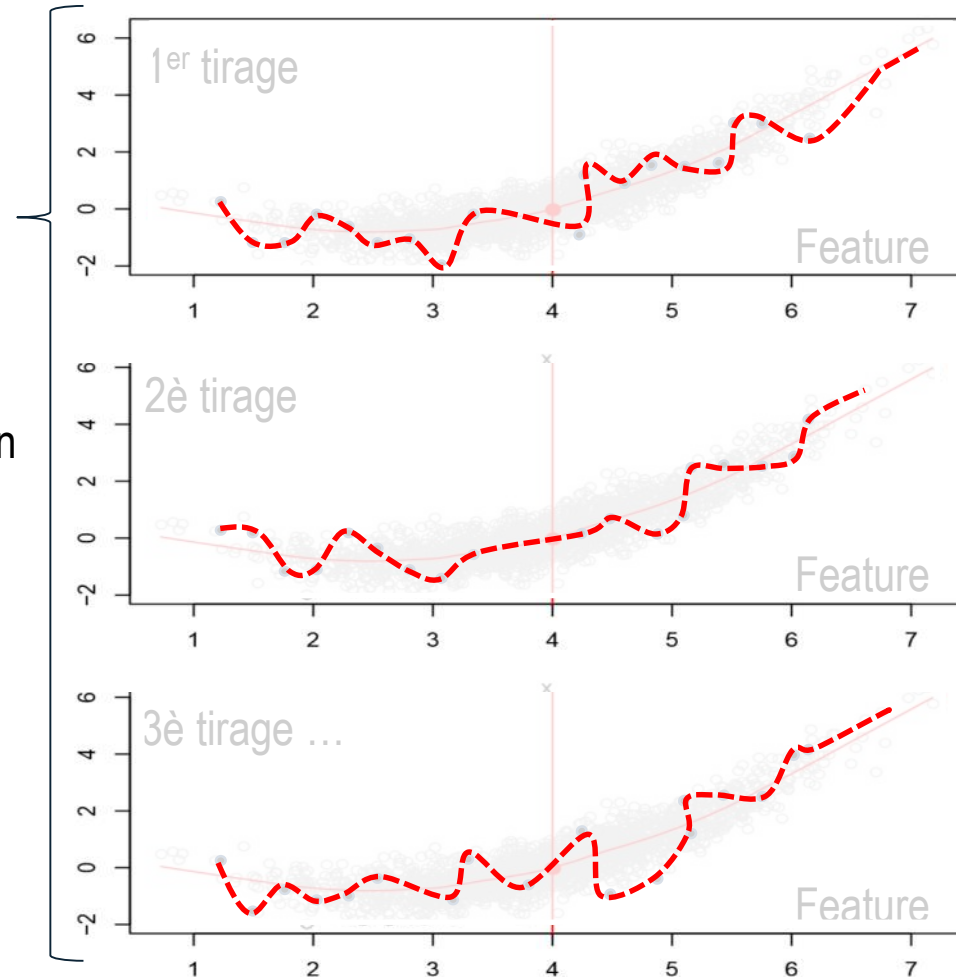
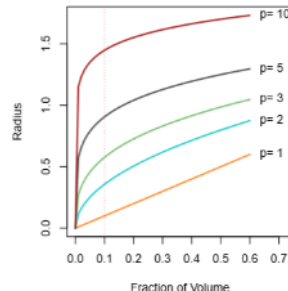
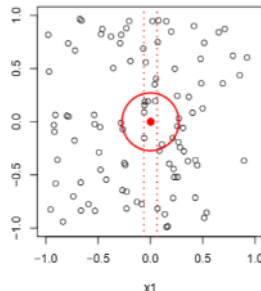
Il y a du **vrai hasard** → là c'est de la philo





## .. Hélas (et hélas bis)

- Vous ne disposez que d'un jeu de données partiel et si vous renouvelez les mesures vous aurez chaque fois un autre jeu d'observation
- Vous avez beaucoup d'observations.. mais encore plus de features pour chaque observation : vous 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 ..



?



# Des astuces : Simplifier le passé

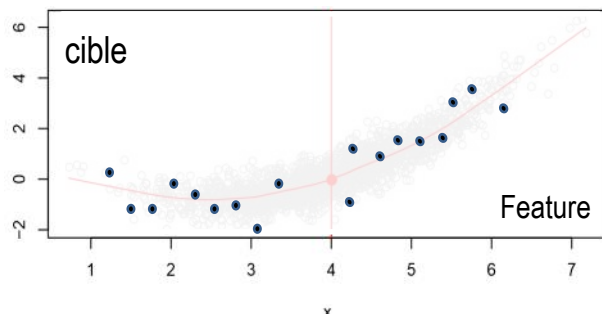
Pour bien prévoir le futur nous pouvons **simplifier** le passé



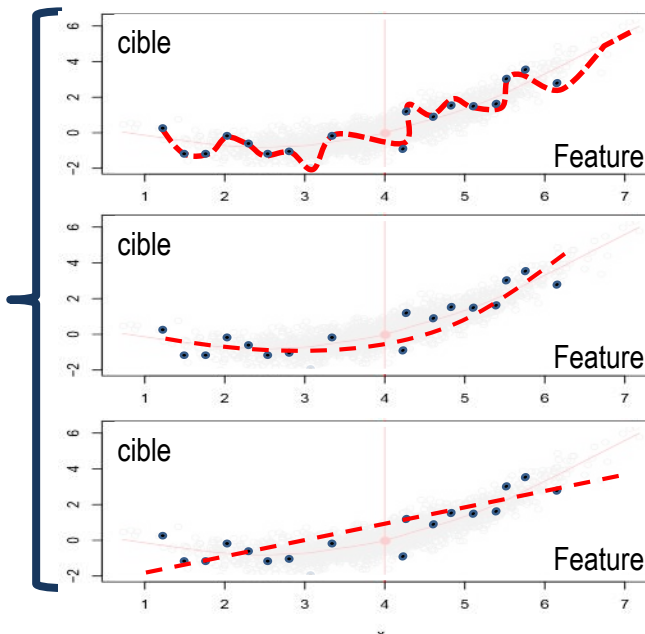
Distinguer

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

## Données initiales



## Modélisation induite



### Complexe

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

Un juste milieu ?

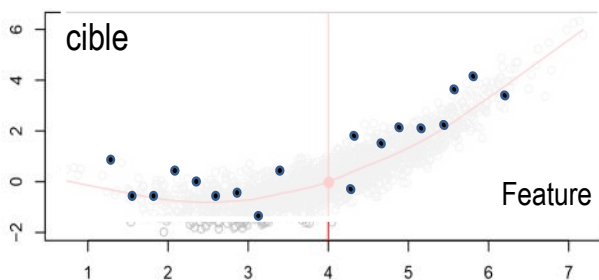
### Simple

- Grossière description du passé
- Faible pouvoir prédictif

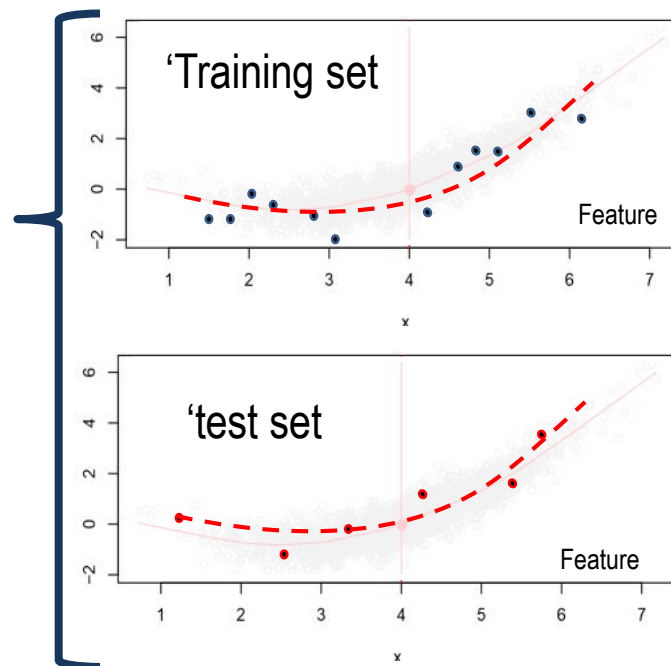
# Des astuces : ne pas être juge et partie

- 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

## Données initiales



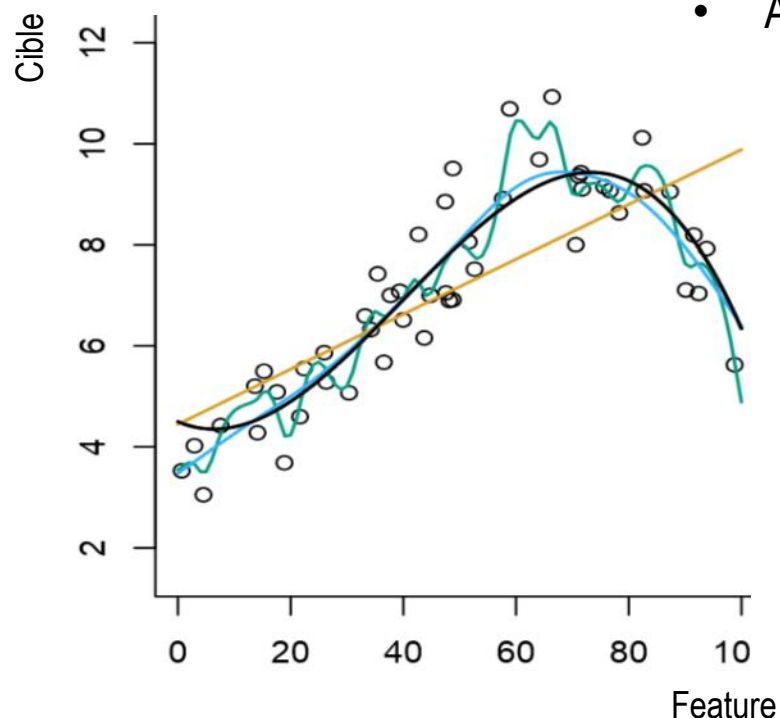
**Exemple de répartition entre apprentissage et test 2/3 apprentissage, 1/3 test**  
**Plus sophistiqué : cross validation**



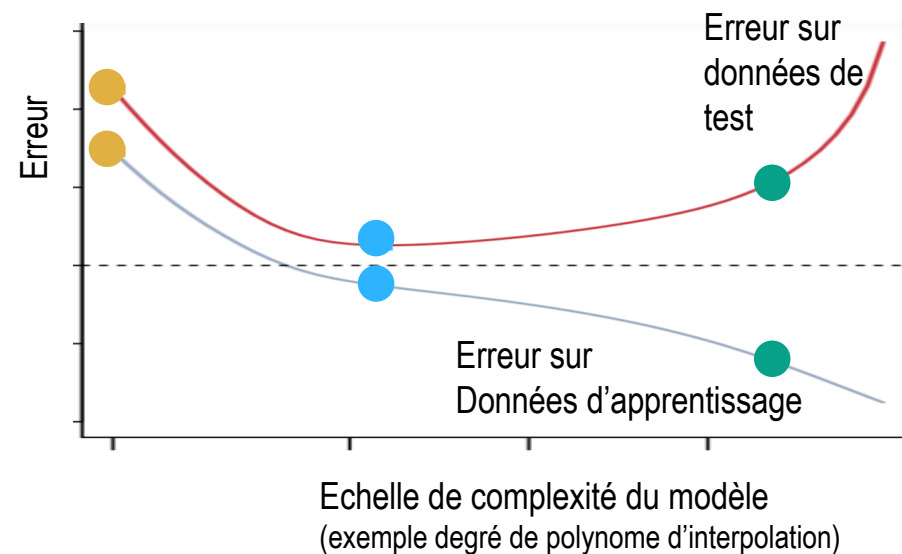
Construction d'un modèle pur un niveau de complexité donné

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

## 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)

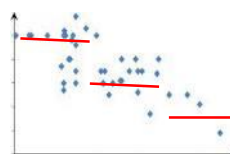
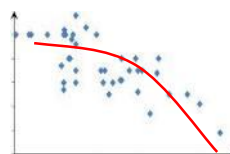
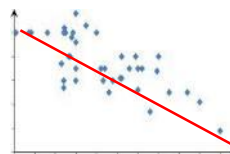
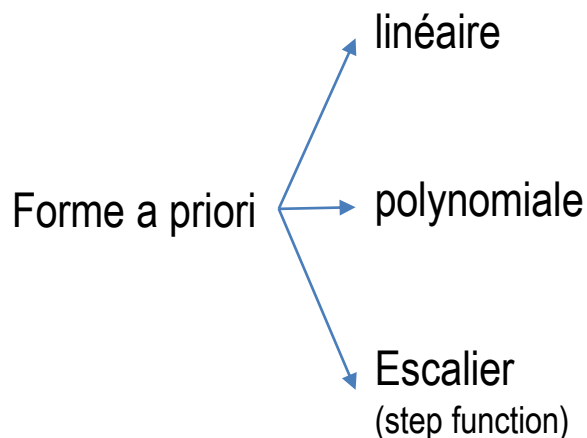


# Des astuces : choisir une forme générale de la solution puis l'ajuster

## Choisir une gamme de formes, à priori, et s'y tenir

- 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 !)

### Echelle de complexité

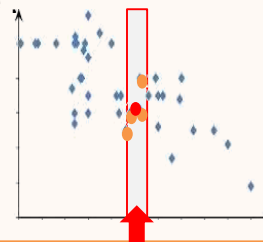


NA : passer à la page suivante !

Degré du polynôme

Nombre de marches

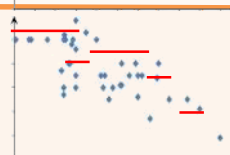
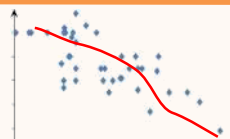
Pour une valeur de feature : prendre la valeur des k voisins connus les plus proches  
La courbe complète n'est en fait jamais dessinée ...



Sans forme a priori

KNN (K nearest neighbour)

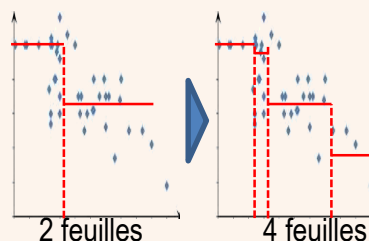
Arbre



Nombre de voisins pris en compte

Profondeur de l'arbre

Trouver un césure dans l'axe des features : de part et d'autre, affecter la moyenne des données disponible : ajuster la césure qui minimise l'erreur globale. Recommencer sur chaque partition



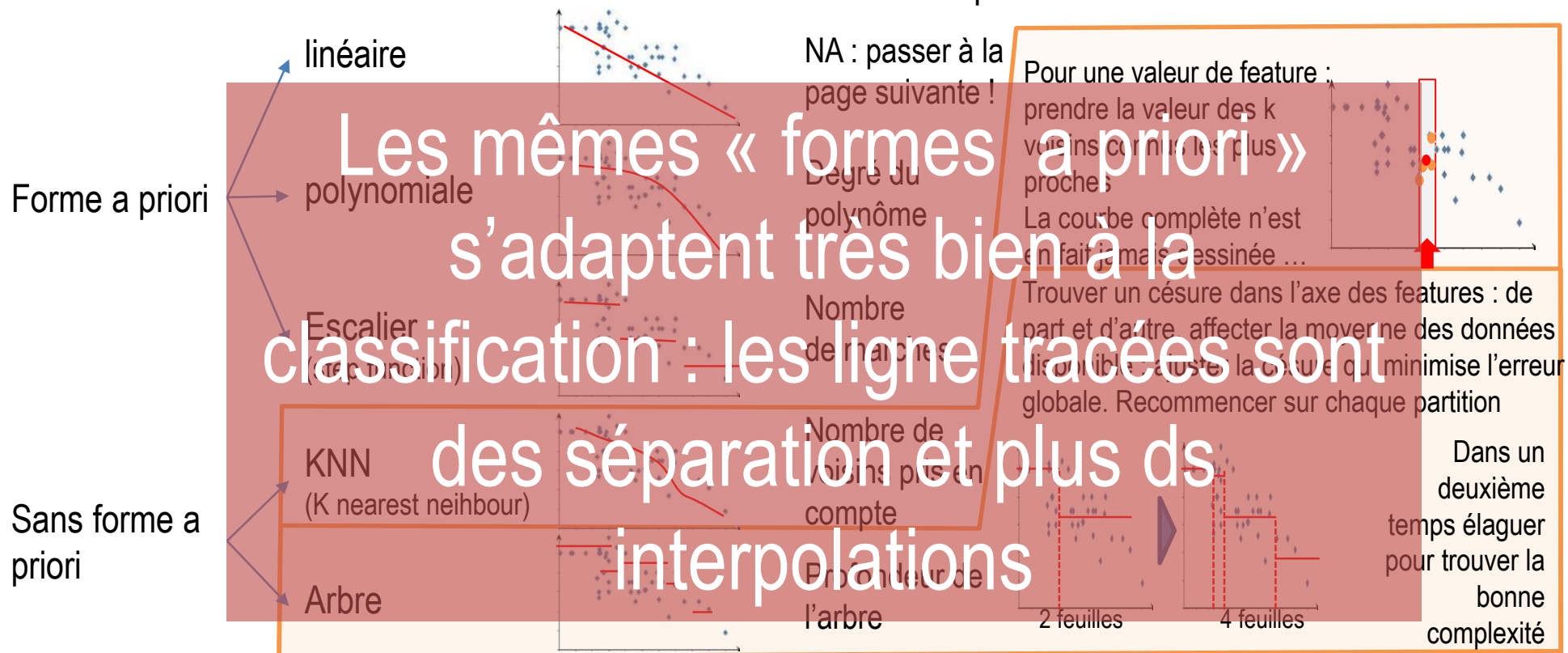
Dans un deuxième temps élaguer pour trouver la bonne complexité

# Et pour la classification ?

## Choisir une gamme de formes, à priori, et s'y tenir

- 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 !)

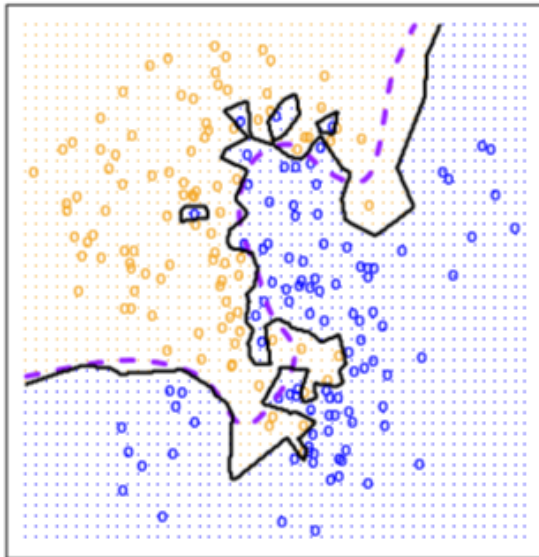
Echelle de complexité



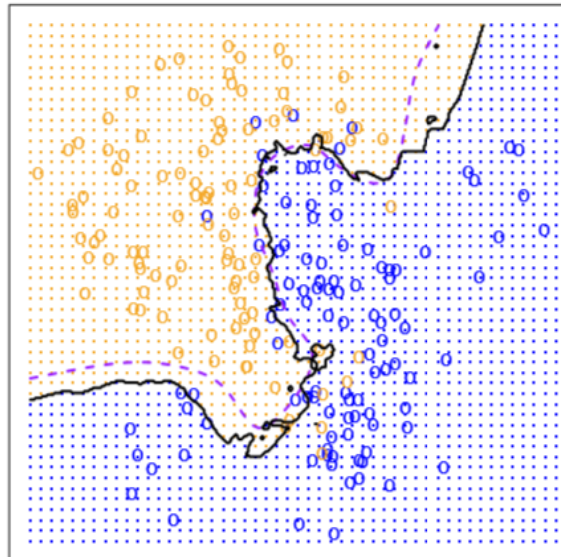
# K nearest neighbour

■ Chaque nouveau pint est évalué en fonction des k plu proches valeurs connues

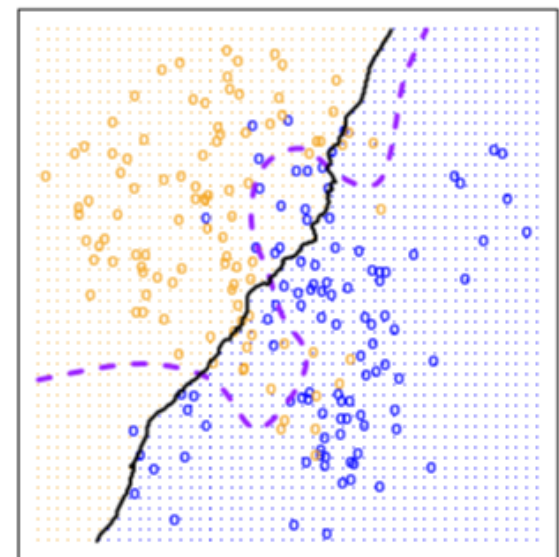
K=1



K=10



K=100



Source : Stanford -

# Un vrai travail de créativité

Le machine learning trouve des réponses à des questions précises. Aucun algorithme n'aide à trouver cette question

Comprendre le problème

Comprendre n'est pas une nécessité algorithmique. Cela donne des idées d'enrichissement, facilite le dialogue avec le client final et accensement donne de l'intérêt

Comprendre les données : contenu et signification

Un data set n'est jamais complet, et nombre de modèles ne supportent pas les « trous » ... à vous de les combler !

Nettoyer / compléter

La fonction de coût peut être étalonnée sur une prédiction naïve : permettra d'évaluer le gain réel de vos efforts. Vous pouvez aussi en choisir deux

Définir la fonction de coût

Complexifiez progressivement vos approches :  
-> chaque modélisation reprend toutes les étapes d'optimisation détaillées dans ce support

Tester des modélisations

Une courbe de ROC n'est pas forcément un livrable explicite pour une direction métier ...

Présenter des résultats actionables

Enrichissement :

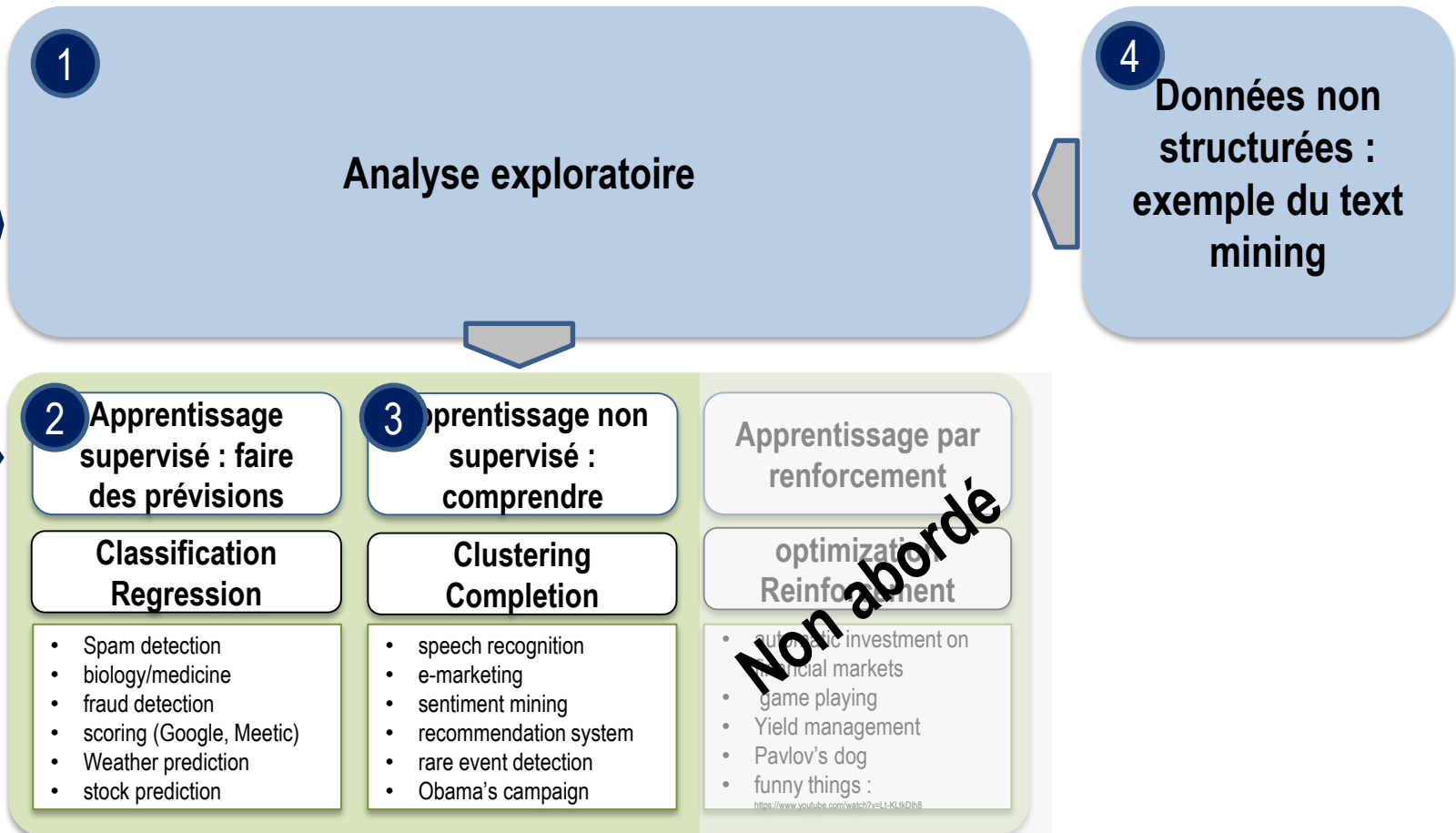
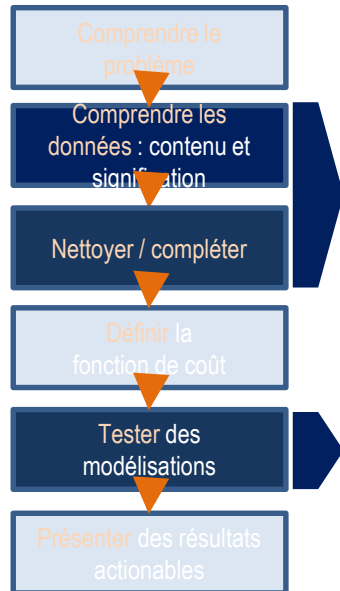
- Depuis le dataset lui-même :
  - calcul de durée, flux, autorégression
  - Composition de variables
- Avec des données externes

Enrichir le data set

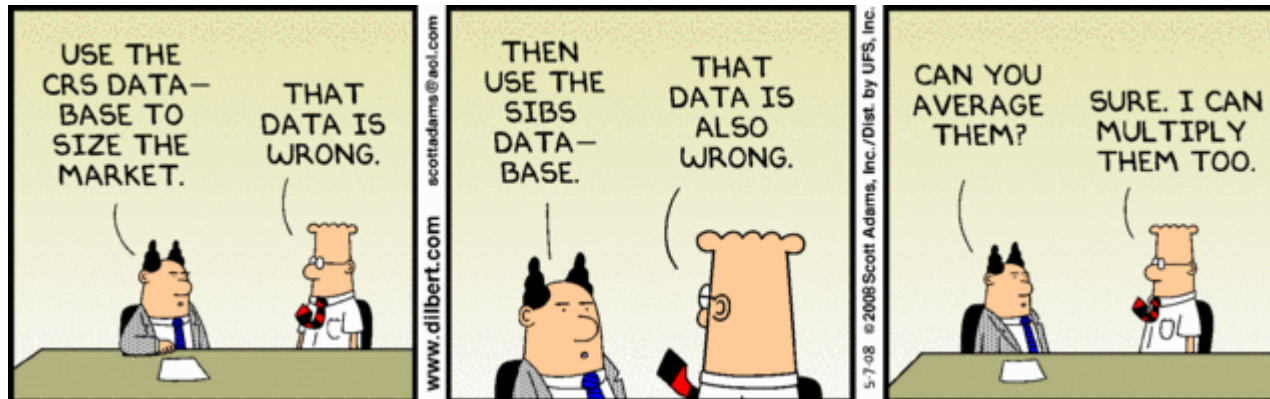


# Le cours focalise sur les phases les plus techniques de la démarche projet

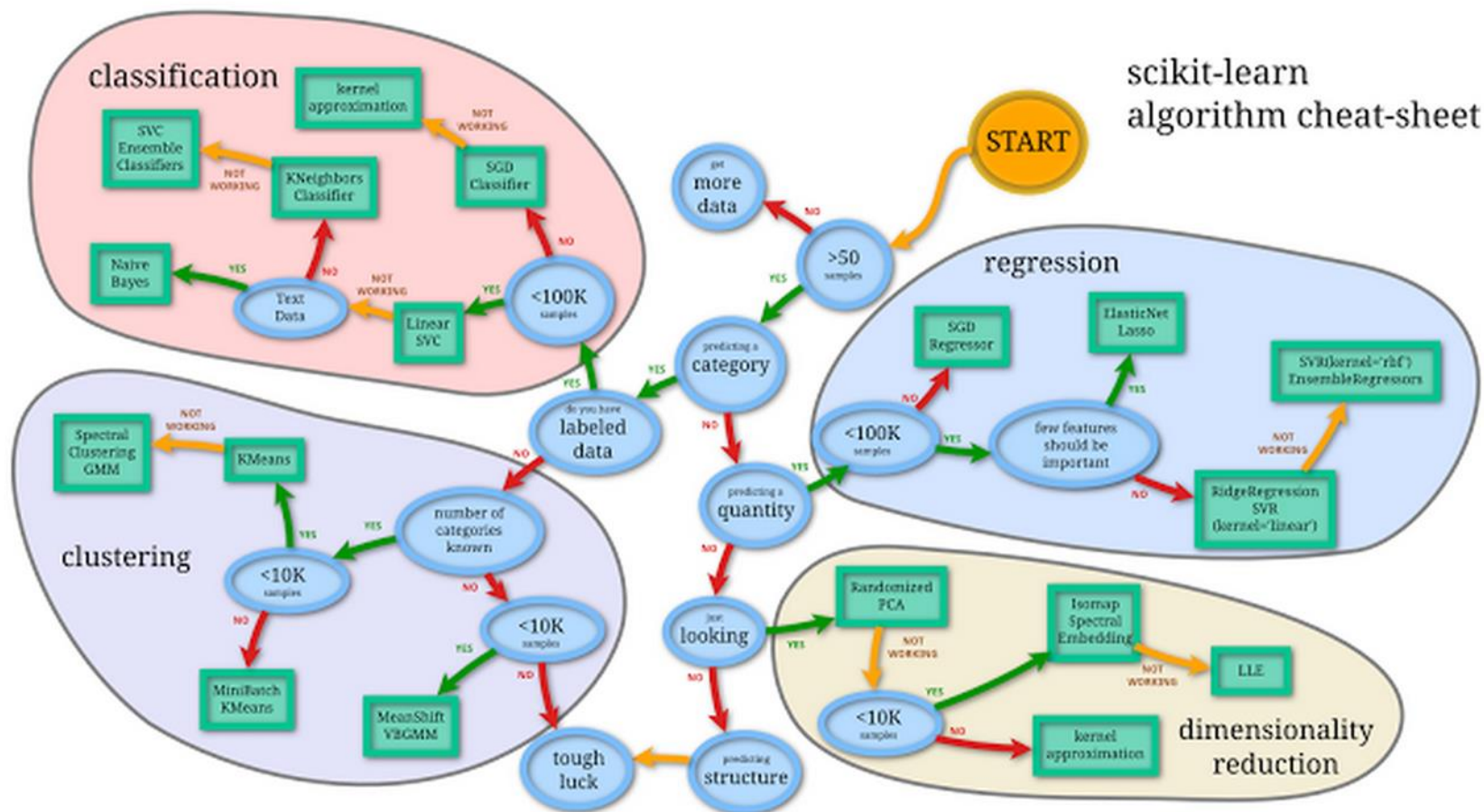
## Rappel de la démarche



# Des questions ?



- General : <http://www.datasciencecentral.com/>
- Courses : <https://www.coursera.org/course/artificialvision/>
- Material : <http://www.di.ens.fr/~fbach/>
- Machine Learning competitions : <http://www.kaggle.com/>



<http://peekaboo-vision.blogspot.fr/2013/01/machine-learning-cheat-sheet-for-scikit.html>