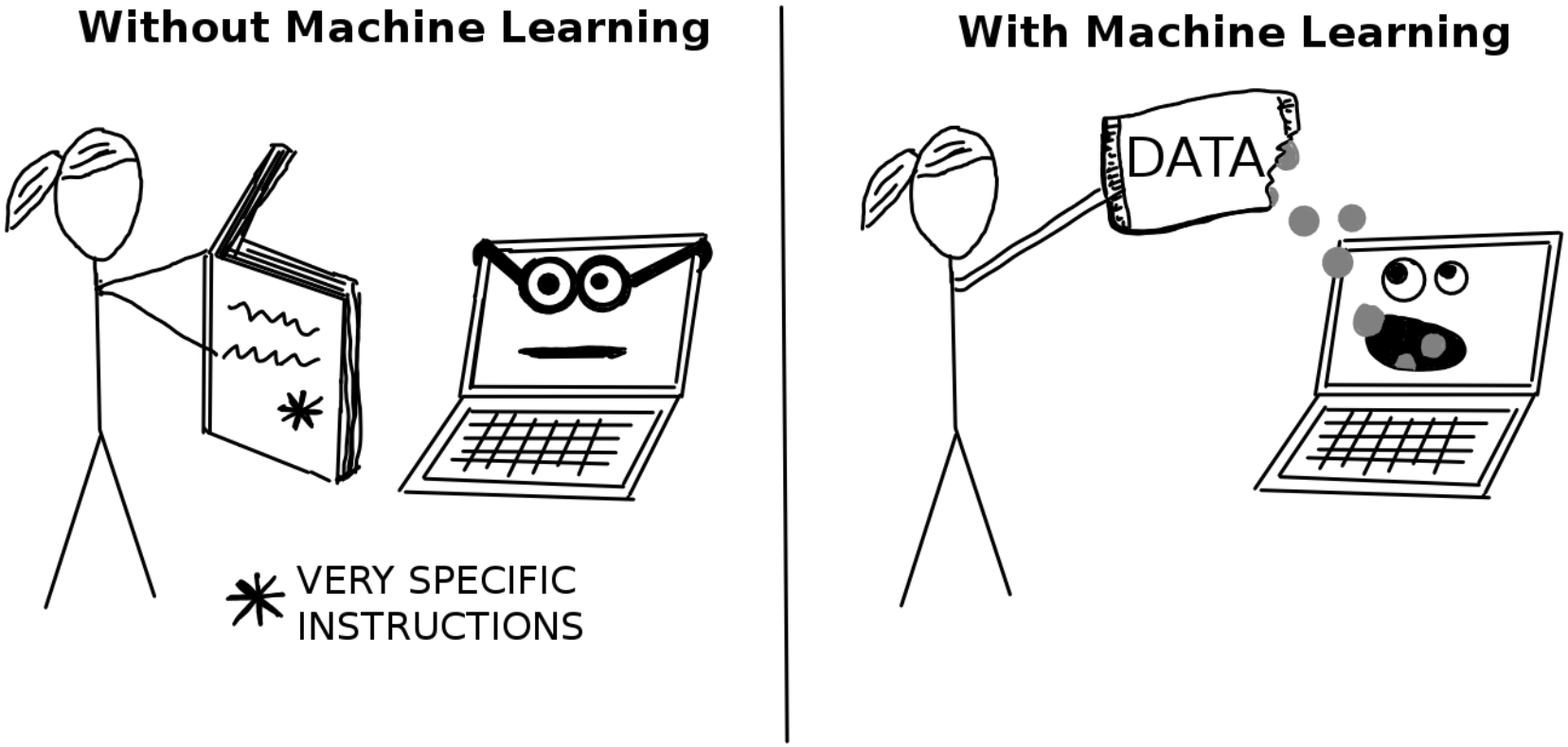
Python data science

Nicolas Rousset

Formation Data Science, day 1

Rappel sur le machine learning

Principe du machine learning



3 types d'apprentissage

- Apprentissage supervisé
- Apprentissage non supervisé
- Apprentissage par renforcement

Apprentissage supervisé

C'est le principal type de machine learning

- On donne en entrée des données et le résultat attendu, il s'agit de trouver les résultats pour des cas où on a seulement les données. Exemple :
- prévoir la production électrique éolienne
- prévoir la survie à bord du titanic
- reconnaître l'écriture manuelle
- faire des diagnostiques médicaux

Apprentissage non supervisé

Apprentissage où on ne donne pas la bonne réponse, du coup on cherche des structures. Il existe principalement deux types de problématiques :

- clustering, détecter des groupes d'élément similaires - réduction de dimensionnalité

Par exemple clustering, principal component analysys, independant component analysis, auto-encoder avec les réseaux de neurone

Apprentissage par renforcement

AlphaGoZero, utilisé dans le contrôle robotique et les jeux.
Le modèle apprend face à un environnement qui renvoie une récompense.

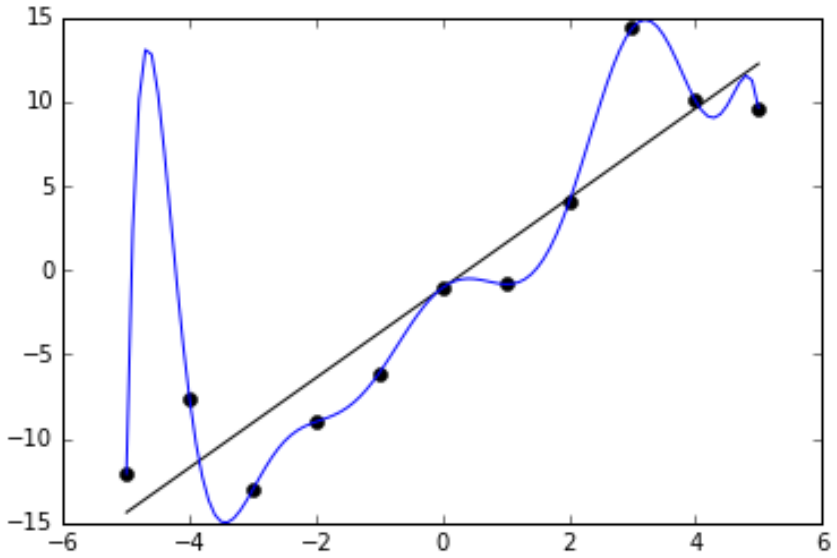
Assez marginal, car il faut un environnement interactif

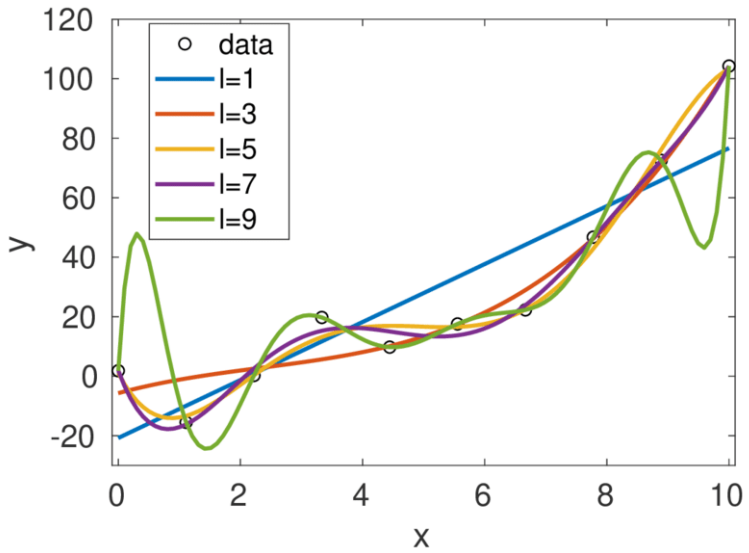
Principe du machine learning (2)

L'idée est de ne pas programmer explicitement les règles, mais de les apprendre des données

!! Le machine learning et encore plus le deep learning nécessite beaucoup de données
Et il faut qu'elles soient annotés pour l'apprentissage supervisé

Problème fondamental du sur-apprentissage





Source : wikipedia / Research gate

Importance de la formulation du modèle et de la mesure

Exemple simple : prédiction du cours de la bourse

- on peut chercher à prédire le cours de la bourse $(C_{t+1} = f(C_t))$
- on peut chercher à prédire la variation cours de la bourse $(C_{t+1} - C_t = f(C_t))$

Importance de la formulation du modèle et de la mesure (2)

Exemple simple : prédiction du cours de la bourse

- on peut chercher à prédire le cours de la bourse ($C_{t+1} = f(C_t)$)
- on peut chercher à prédire la variation cours de la bourse ($C_{t+1} - C_t = f(C_t)$)

Premier cas => Il va être très facile d'obtenir un très bon score avec $C_{t+1} = C_t$
Deuxième cas => La plus petite prévision réelle est déjà énorme

Importance du features engineering

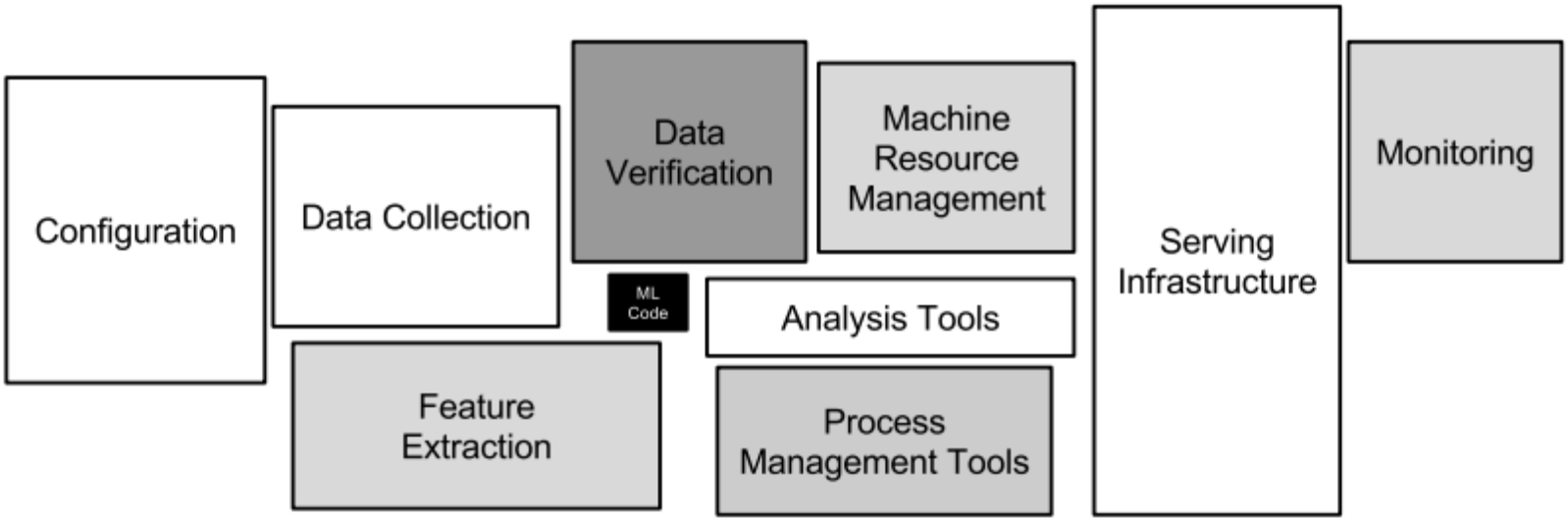
Le ŕeseaux de neurones peut apprendre toutes les fonctions(comme $f(x) = x^2$) mais cela fonctionnera beaucoup mieux si vous lui donnez directement la fonction en entrée.

Donner les bonnes entrées reste le travail du datascientist en général (discutable pour le traitement d'image)

Limites du machine learning

- Processus global
- Les données annotées
- Stationnarité
- Stabilité, fiabilité, explicabilité

Processus global



Extrait de “Hidden Technical Debt in Machine Learning Systems. Sculley et al”

Décalage entre la formation et la réalité

- On va travailler sur des jeux de données disponibles et annotés
- Les jeux de données sont assez bien conçus
- On a un feedback direct

Premier problème de la réalité => les données

- un modèle ne vaut rien à priori
- les sociétés ne donnent pas facilement accès aux données (y compris en interne)
- RGPD

Deuxième problème de la réalité => les données

- il faut qu'elle soient correctement annotés (notamment dans la reconnaissance d'image)

Troisième problème de la réalité => les données

- il faut qu'elles soient non biaisées
- beaucoup de "bad buzz", par exemple le système de lecture de CV d'amazon

Quatrième problème de la réalité => les données

- Certaines données sont plus chères que d'autres
- J'ai personnellement des doutes sur le développement du machine learning dans les domaines où les données ne sont pas générés numériquement (éducation, algriculture) à part si les gains économiques sont très importants (maintenance prédictives)

En résumé

- On ne le verra pas dans la formation, mais le machine learning est d'abord un problème de données avant d'être un problème d'algorithme

La question du feedback

- Si on travaille en local ou sur kaggle/ENS, on ne se teste que sur des données dont on connait le résultat
- En production, comment savoir si vos résultats sont bons ?
- Problème de monitoring
- Attention à la stationnarité

La stationnarité (question du biais)

- On apprend toujours sur le passé, est-il fiable pour prédir l'avenir ?
- Certains domaines nécessitent des mises à jour des modèles très régulières (cybersécurité)