Introduction

— Slides : link

1 Qu'est-ce que l'analyse de données ?

L'analyse de données est un ensemble de méthodes permettant de retirer de l'information d'un jeu de données. On parle aussi d'apprentissage statistique (*statistical learning*). L'idée est d'utiliser des modèles statistiques pour comprendre comment les données sont structurées et comment elles intéragissent l'une avec l'autre.

> Exemple
>
> Imaginons que vous êtes employé par l'Organisation des Nations Unies (ONU). Votre mission est d'analyser l'espérance de vie à travers le monde. Pour cela, vous disposez d'une mesure de l'espérance de vie dans chaque pays membre de l'ONU, bien sûr, mais aussi le PIB par habitant, les montants des dépenses liés à la santé, le taux de fertilité, le taux d'urbanisation, le niveau d'éducation du pays, etc. Le but de l'analyse de données est de trouver des liens entre ses différentes variables et la variable d'intérêt, l'espérance de vie, de visualiser ces données, et éventuellement de prédire l'espérance de vie à partir des autres variables.

2 Objectifs du cours

Dans ce cours, on cherche à introduire des méthodes qui permettent une étude d'un jeu de données de "haute dimension" (dans le sens où l'on ne peut pas faire un simple graphique de l'ensemble des variables pour chaque observation) sans avoir recours à un modèle probabiliste. Les différentes techniques que l'on va voir peuvent servir à :

— visualiser les données ;
— réduire la dimension des données ;
— identifier certains liens entre les variables ;

— diviser le jeu de données en groupes/classes.

Ce cours n'a pas vocation à être exhaustif, dans le sens de présenter toutes les méthodes possibles. Ce cours n'a pas non plus vocation à être à l'état de l'art, dans le sens où on ne s'intéressera pas aux derniers développements en apprentissage machine. Ce cours n'est pas non plus un cours de programmation.

Pour finir cette indroduction, voici un passage de *Statistical Rethinking* de Richard McElreath (McElreath 2020) trouvant résonnance dans ce cours.

> Statistics courses […] tend to resemble horosscopes. There are two senses to this resemblance. First, in order to remain plausibly correct, they must remain tremendously vague. This is because the targets of the advice, for both horoscopes and statistical advice, are diverse. But only the most general advice applies to all cases. A horoscope uses only the basic facts of birth to forecast life events, and a […] statistical guide uses only the basic facts of measurement and design to dictate a model. It is easy to do better, once more detail is available. In the case of statistical analysis, it is tipically only the scientist whho can provide that detail, not the statistician. Second, there are strong incentives for both astrologers and statisticians to exaggerate the power and importance of their advice. No one likes an astrologer who forecasts doom, and few want a statistician who admits the answers as desired are not in the data as collected. Scientists desire results, and they will buy and attend to statisticians and statistical procedures that promise them. What we end up with is too often *horoscopic* : vague and optimistic, but still claiming critical importance.

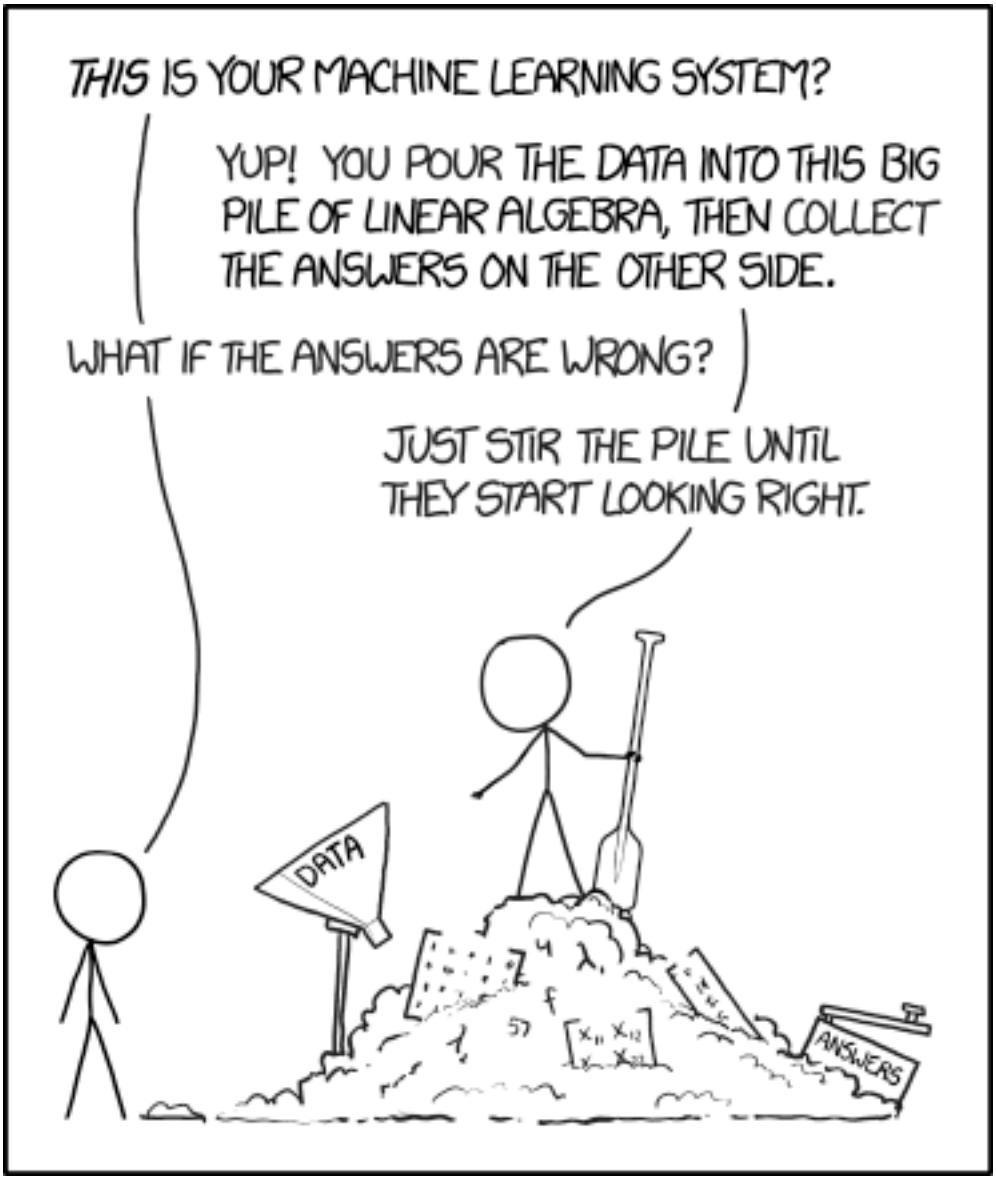

FIGURE 1 – Machine learning (xkcd :1838).

Références

McElreath, Richard. 2020. *Statistical Rethinking : A Bayesian Course with Examples in R and STAN*. 2 éd. New York : Chapman and Hall/CRC. https://doi.org/10.1201/9780429029608.