INSTITUT DE FRANCE
Académie des sciences

Inria

VILLE DE NICE

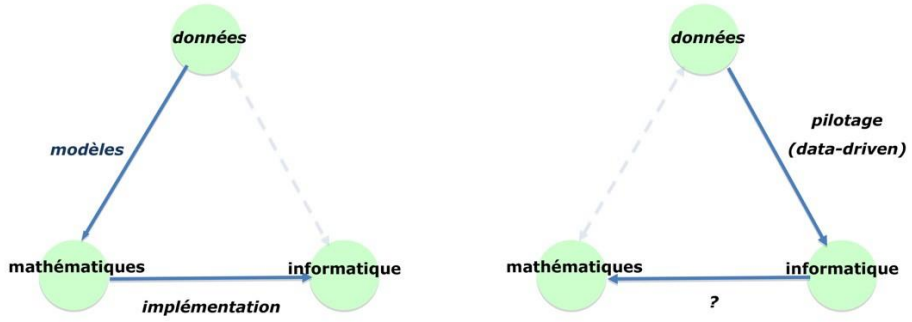
UNIVERSITÉ CÔTE D'AZUR

Académie en région
à Nice et Sophia Antipolis

Quelques réflexions sur la science des données

Patrick Flandrin
CNRS, École normale supérieure de Lyon et Académie des sciences

Ce que l'on appelle aujourd'hui « science des données » se situe à la confluence de trois grands champs thématiques : la « physique » au sens large (pouvant inclure la biologie comme les artefacts technologiques, en se référant aux mondes d'où viennent et où se déploient les données), les mathématiques et l'informatique. De l'équilibre de ces trois piliers naît l'efficacité de l'analyse et de l'action, tant pour la représentation et la modélisation des données elles-mêmes que pour la mise en œuvre et le contrôle d'algorithmes permettant d'en extraire de l'information et de les manipuler. Ce schéma qui a longtemps prévalu — avec l'analyse de Fourier comme une de ses plus éclatantes illustrations, et celle en ondelettes comme un de ses avatars les plus récents — est aujourd'hui à reconsidérer à l'aune des bouleversements observés depuis quelques années. Du côté de la science, les données de plus en plus massives et leurs trois V de Volume, Vitesse et Variabilité, ont permis l'explosion de méthodes d'apprentissage (en particulier, profond, ou *deep learning*) pouvant substituer l'usage d'exemples à la modélisation, au prix parfois de l'interprétation et de l'explicabilité. Eu égard aux résultats spectaculaires qui ont été obtenus récemment, la tentation est grande de se conformer à ce glissement de paradigme par pragmatisme, voire à se ranger sous cette bannière nouvelle faute d'être déconsidéré (ce qu'on pourrait appeler ironiquement « le complexe du *deep* ».) Il n'en reste pas moins que le défi est bien réel de mieux comprendre (en particulier, d'un point de vue mathématique) le pourquoi de cette efficacité (cf. Figure 1).



Schématisation du glissement de paradigme entre l'approche « classique », ancrée sur la modélisation, et celle plus « moderne » s'appuyant sur l'apprentissage (voir texte).

Du côté des données, c'est une nouvelle richesse qui est apparue, avec à la clé les deux V nouveaux de Valeur et de Véracité, liés à des questions d'acquisitions, de biais, de propriété, de confidentialité, de partage, de sécurité… Les données sont en fait bien souvent des « obtenues » (pour citer Bruno Latour) et leur usage n'est bien souvent pas neutre. À côté des questions éthiques liées à leur confidentialité ou aux biais possibles qui sont inhérents à leur sélection lorsqu'il s'agit d'apprentissage, leur usage immodéré est aussi source de coûts énergétiques et environnementaux pouvant être considérables. Ces questions transportent la science des données vers des enjeux sociétaux qui ne doivent pas être disjoints des questions proprement scientifiques et techniques mais envisagés et anticipés de manière globale et partagée.