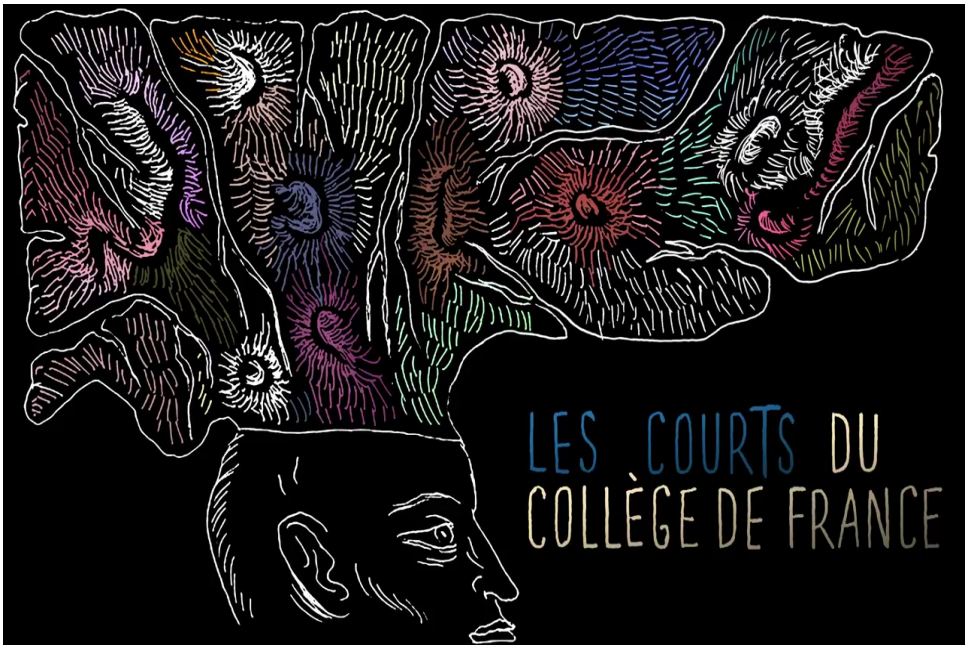
# L'apprentissage face à la malédiction de la grande dimension

**Stéphane Mallat présente son cours de l'année dans la série les courTs du Collège de France.**




**Transcription de la vidéo :**

Cette année, au Collège de France, je vais introduire cette nouvelle chaire de « Sciences des données » et ce premier cours va introduire les outils mathématiques et informatiques fondamentaux pour comprendre les grandes questions autour de la modélisation et de l'apprentissage en sciences des données.

Actuellement, le point de vue posé par les mathématiques et l'informatique, c'est la possibilité de développer des algorithmes automatiques. Les résultats de ces algorithmes, on les retrouve dans la vie de tous les jours : c'est par exemple la reconnaissance d'images avec votre téléphone portable qui est capable de reconnaître votre visage ou votre voix, c'est aussi le champion du monde de go qui a été récemment battu par une machine ou même la traduction automatique de textes. L'enjeu est à la fois de comprendre les algorithmes mais aussi de comprendre les principes mathématiques qui font que ces algorithmes, effectivement, peuvent marcher ou peuvent, de temps en temps, ne pas fonctionner et avoir des erreurs. Ces algorithmes commencent par une phase d'apprentissage où les paramètres sont ajustés en fonction des exemples qu'on donne, par exemple une image où l'on dit « c'est un chien », une autre « c'est un chat », une troisième « c'est un cerf », etc. et l'algorithme apprend à répondre de manière à ne pas faire d'erreur sur les exemples. Évidemment, l'enjeu fondamental, c'est d'être capable de généraliser. Ça veut dire que, si l'on donne une nouvelle image que l'algorithme n'a jamais vue, il faut qu'il soit capable de donner la bonne réponse.

Alors pourquoi est-ce qu'on peut généraliser ? La raison est que les phénomènes sous-jacents ont de la régularité. Par exemple, si vous faites une expérience en chimie, vous allez faire quelques expériences puis prédire les conditions que vous ne connaissez pas en traçant, par exemple, une courbe régulière entre les points de l'expérience. Eh bien c'est la même chose qu'un tel algorithme va faire : il va utiliser une forme de régularité pour prédire les valeurs inconnues à partir des valeurs connues.

Pourquoi est-ce que le problème est très difficile ? Le problème est très difficile parce que les données que l'on va considérer, que ce soit des textes, des images, des sons, ont énormément de variables Et là on fait face à ce que l'on appelle la malédiction de la dimensionnalité, c'est-à-dire l'explosion des possibles, et pour pouvoir effectivement généraliser, il faut une régularité qui est très forte et l'enjeu mathématique est de comprendre la nature de cette régularité.

Quels sont les grands principes qui permettent d'analyser cette régularité ? D'abord, une idée très fondamentale en sciences, c'est l'idée de parcimonie. On peut généraliser si le nombre d'hypothèses ou le nombre d'attributs que l'on extrait est petit. Une deuxième propriété fondamentale est le fait que le monde est structuré avec des éléments hiérarchiques à travers les échelles. Pensez à la physique : vous passez des particules aux atomes, aux molécules, aux cristaux polymères, jusqu'aux galaxies. Cette hiérarchie est fondamentale pour comprendre la complexité et il y a toute une théorie, que l'on appelle la théorie des ondelettes, qui permet de comprendre mathématiquement cette organisation multi échelles hiérarchique et qui apparaît dans cette compréhension de la généralisation et de la régularité.

Mon propre chemin de recherche est passé par la construction de ces bases orthogonales d'ondelettes, les algorithmes rapides, la compréhension de cette structure mathématique, qui a notamment donné lieu à des applications comme la compression d'image, et on retrouve ce type de structure dans les réseaux de neurones et il y a un véritable mystère mathématique pour comprendre comment cela se fait que ces réseaux de neurones avec une architecture assez bien déterminée sont capables de répondre à des questions aussi diverses que celles portant sur des problèmes de traitement d'image, de son, de physique quantique. Si vous prenez, par exemple, la voiture autonome où il y a actuellement des réseaux de neurones qui sont présents ou même des applications médicales, on voit bien que l'enjeu d'avoir des algorithmes robustes maîtrisés mathématiquement est tout à fait fondamental.