Création d'une chaire *Sciences des données* au Collège de France

Stéphane Mallat, nommé titulaire, donnera sa leçon inaugurale le jeudi 11 janvier 2018, à 18h00

L'analyse automatique des données numériques est devenue un enjeu industriel, sociétal et scientifique majeur et doit faire face à la diversification et la multiplication exponentielle de ces données. Après avoir ouvert en 2009 une chaire annuelle *Informatique et sciences numériques*, puis, en 2012, une chaire *Algorithmes, machines et langages*[1], l'Assemblée du Collège de France a décidé de créer une chaire entièrement consacrée aux *sciences des données*, domaine fondamental qui non seulement bouleverse des pans entiers de nos économies et de nos sociétés, mais ouvrent aussi de vastes perspectives scientifiques et technologiques. Le Pr Stéphane Mallat a été nommé titulaire de cette nouvelle chaire.

Algorithmes d'apprentissage et réseaux de neurones artificiels

Stéphane Mallat, Professeur de mathématiques et d'informatique à l'École normale supérieure, jusqu'en 2017, a consacré sa recherche aux mathématiques appliquées au traitement du signal et plus récemment à l'étude des algorithmes d'apprentissage et des réseaux de neurones profonds.

Si la performance des algorithmes d'analyse de données et l'intelligence artificielle ont fait un bond remarquable ces dernières années, on comprend encore mal les principes mathématiques qui permettent à ces techniques de fonctionner et on ne sait donc pas toujours correctement qualifier leurs résultats. Il faut pourtant pouvoir s'assurer scientifiquement de la non-existence de comportements aberrants, quand il s'agit par exemple de continuer à développer des applications médicales ou des applications utilisées pour la conduite de voitures autonomes. Un des enjeux fondamentaux est d'être capable de généraliser, d'acquérir la certitude que l'algorithme d'apprentissage ne se trompera pas face à un cas qu'il n'a jamais analysé et donc de comprendre la nature des régularités sous-jacentes.

> « *La beauté des concepts qui se dégagent s'enracine dans la beauté des correspondances entre domaines aussi différents que la reconnaissance d'images, la neurophysiologie, la chimie quantique, la cosmologie ou l'économie. Révéler ces correspondances est une des ambitions des mathématiques appliquées »,* S. Mallat.

Le vertige de la grande dimension

Les sciences des données ont pour but de répondre à des questions à partir de données ayant un très grand nombre de variables, qu'il s'agisse d'images, de sons, de textes, de données génomiques, de liens dans des réseaux sociaux ou de mesures physiques. Dans une image les variables sont les pixels, et il y en a plusieurs millions. Cette multitude de variables ouvre un champs gigantesque des possibles, ce que l'on appel la malédiction de la dimensionnalité. Les algorithmes doivent faire face à cette malédiction, et extraire l'information pertinente en hiérarchisant les paramètres importants, grâce à des informations partielles sur la

régularité des réponses. Comprendre les principes mathématiques et la nature des régularités qui gouvernent les algorithmes d'apprentissage, c'est l'objectif des travaux de Stéphane Mallat aux frontières des mathématiques et de l'informatique, en effectuant un aller-retour constant avec les applications.

Ses cours au Collège de France permettront d'introduire les outils mathématiques et informatiques fondamentaux nécessaires pour comprendre les grandes questions et défis posés par la modélisation et l'apprentissage en sciences des données. Son cycle de cours pour l'année académique 2017/2018, *L'apprentissage face à la malédiction de la grande dimension*, aura lieu les mercredis à 9h30 à partir du 17 janvier 2018. Sa leçon inaugurale sera retransmise en direct le 11 janvier à 18h. L'ensemble de son enseignement sera disponible sur www.college-de-France.fr

Les challenges de traitement et d'analyse de données

Fidèle à une approche qui ne cesse de confronter les mathématiques à l'expérimentation, Stéphane Mallat a mis en place avec son équipe de recherche des *Challenges de données* autour de communautés en ligne. Des start-up, hôpitaux ou laboratoires scientifiques soumettent via un site internet des données et des problèmes que les participants sont appelés à résoudre. Les résultats sont évalués automatiquement par le site. Cela permet de tester la performance de différents types d'algorithmes sur des problèmes concrets. Le projet a déjà réunis plus de 2500 participants : de véritables compétitions qui associent élèves, chercheurs et ingénieurs, et auxquelles les auditeurs du cours de Stéphane Mallat seront invités à participer.

---

[1] La chaire annuelle *Informatique et sciences numériques* a été créée en 2009 dans le cadre d'un partenariat entre le Collège de France et Inria. Elle accueille chaque année un nouveau titulaire, spécialiste reconnu d'un domaine (intelligence artificielle, sécurité informatique, big data, algorithmique, …). Elle a été inaugurée par Gérard Berry, nommé depuis professeur titulaire d'une chaire pérenne d'informatique, *Algorithmes, machines et langages*.

Contacts presse : Marie Chéron / Cécile Barnier : 01 44 27 12 72 - cecile.barnier@college-de-france.fr

Sciences des données et mathématiques appliquées

Par Stéphane Mallat

Nous assistons à un déluge de données numériques, sous la forme d'images, de sons, de textes, de mesures physiques ainsi que de toutes les informations disponibles sur Internet. La performance des algorithmes d'analyse de données a fait un bond ces dernières années, grâce à l'augmentation des capacités de calcul et aux masses de données, mais aussi grâce à l'évolution rapide des algorithmes d'apprentissage. Ce bond est à l'origine de la renaissance de l'intelligence artificielle. En particulier, les réseaux de neurones ont récemment obtenu des résultats spectaculaires pour la classification d'images complexes, la reconnaissance vocale et de musique, pour la traduction automatique de textes ou la prédiction de phénomènes physiques et même pour battre le champion du monde de Go. Ils sont utilisés dans des applications industrielles et médicales.

Un algorithme d'apprentissage prend en entrée des données, par exemple une image, et estime la réponse à une question, par exemple trouver le nom de l'animal dans l'image. Ces algorithmes d'apprentissage ne sont pas entièrement déterminés à l'avance. Ils incluent de nombreux paramètres qui sont optimisés avec des exemples, lors de la phase d'apprentissage. Pour la reconnaissance d'animaux, on donne à l'algorithme des exemples d'images et le nom des animaux dans chaque image. L'apprentissage assure que l'algorithme ne fasse pas d'erreur sur les exemples d'entrainement. Cela n'a d'intérêt en soit que si l'on peut garantir que le résultat se généralise et donc que l'algorithme est capable de prédire le bon résultat sur des données qu'il n'a jamais vues. Cette généralisation est liée à l'existence de régularités, que l'algorithme utilise pour relier le résultat sur une donnée inconnue avec les exemples connus.

La complexité du problème vient du très grand nombre de variables dans chaque donnée. Ainsi une image a typiquement plus d'un million de pixels, et donc plus d'un million de variables dont il faut tenir compte pour répondre à une question. L'interaction de ces variables produit un nombre gigantesque de possibilités. C'est la malédiction de la dimensionnalité. Pour faire face à cette malédiction, l'algorithme utilise une connaissance partielle des régularités des réponses, afin de contraindre le calcul de la bonne réponse. Comprendre la nature de ces régularités en grande dimension est un enjeu fondamental qui fait appel à de nombreuses branches des mathématiques, dont les statistiques, les probabilités, l'analyse et la géométrie

Sciences des données et mathématiques appliquées

« La chaire s'intitule « *Sciences des données* » par opposition au singulier « la science des données » car ce domaine est une « auberge espagnole », où cohabitent des approches et culture scientifiques totalement différentes, qui s'enrichissent mutuellement. Cela comprend les mathématiques et notamment les statistiques, mais aussi l'informatique et l'intelligence artificielle, le traitement du signal et la théorie de l'information, et toutes les sciences comme la physique, la biologie, l'économie ou les sciences sociales, qui traitent et modélisent des données. Apporter une vision et un langage commun au-delà des spécificités de chaque domaine est la vocation des mathématiques. C'est ce point de vu qui sera développé, tout en restant enraciné dans les applications qui sont sources de problèmes nouveaux, de créativité et d'idées. Cet aller-retour entre

Contacts presse : Marie Chéron / Cécile Barnier : 01 44 27 12 72 - cecile.barnier@college-de-france.fr

mathématiques et applications, qui efface progressivement les frontières entre expérimentations et théorie, est au cœur de la démarche des mathématiques appliquées. La beauté des concepts qui se dégagent ne s'enracinent pas seulement dans leur pureté, comme celle d'un diamant qui se suffirait à lui même, mais dans la beauté des correspondances entre domaines aussi différents que la reconnaissance d'images, la neurophysiologie, la chimie quantique, la cosmologie ou l'économie. Révéler ces correspondances est aussi l'ambition des mathématiques appliquées.

En sciences des données, il s'agit de comprendre le lien entre les applications, l'algorithmique, les expérimentations numériques et la compréhension mathématique du traitement de masses de données. Les mathématiques sont importantes pour garantir la robustesse des résultats, notamment pour des usages critiques comme la médecine ou la conduite de voitures autonomes.

Cette chaire a également pour objectif de mieux faire comprendre les avancées des algorithmes et des mathématiques de l'apprentissage et de l'intelligence artificielle, à un plus large public. Diffuser la connaissance dans ce domaine est important car ces technologies auront probablement un impact croissant sur l'industrie, la médecine mais aussi sur certains aspects de notre organisation économique et sociale. Il faut y réfléchir bien au-delà des cercles scientifiques.

Stéphane Mallat

Contacts presse : Marie Chéron / Cécile Barnier : 01 44 27 12 72 - cecile.barnier@college-de-france.fr

Biographie

Stéphane Mallat, né en 1962 à Suresnes, est élève à l'École polytechnique de 1981 à 1984, puis à l'École nationale supérieure des Télécommunications en 1985. Il fait un Ph.D. en traitement du signal à l'Université de Pennsylvanie de 1986 à 1988 et soutient sa thèse d'habilitation en mathématiques à l'université de Paris Dauphine en 1992.

De 1988 à 1996, il est professeur d'Informatique et de Mathématiques à l'Institut Courant de l'Université de New York puis revient en France comme professeur en Mathématiques Appliquées à l'École polytechnique (1995/2012), où il préside le département de 1998 à 2001. De 2001 à 2007 il cofonde et dirige une start-up de traitements d'images. Il est de 2012 à 2017, professeur d'Informatique et de Mathématiques à l'École normale supérieure de la rue d'Ulm.

En 2012, Stéphane Mallat reçoit un *Advanced Grant* du Conseil européen de la recherche. Il reçoit, en 2013, la médaille de l'innovation du CNRS. Il est membre de l'Académie des sciences depuis 2014 et membre de la *National Academy of Engineering* Américaine depuis 2017.

Biographie complète, prix et distinctions : http://www.college-de-france.fr/site/stephane-mallat/index.htm

Recherche et Activités Scientifiques :

Les travaux de Stéphane Mallat portent sur les mathématiques et le développement d'algorithmes en traitement du signal et en apprentissage statistique.

Dans les années 90, il introduit les principes et les premiers algorithmes d'analyse de signaux par ondelettes (images, sons, enregistrements médicaux, vibrations, etc,), par sa *théorie des multi-résolutions* et un algorithme de calcul rapide des coefficients d'ondelettes. Ces coefficients représentent des signaux et des images avec un nombre réduit de variables. Ces travaux débouchent sur le standard de compression d'image JPEG-2000 ainsi que sur de nombreuses applications en traitement du signal.

À partir de 1993, il développe des modèles de calcul de représentations parcimonieuses dans des dictionnaires redondants. La parcimonie de ces représentations est utilisée en apprentissage et pour restituer des signaux à partir d'un nombre restreint de mesures. Avec ses collaborateurs et étudiants, il introduit des dictionnaires de bandelettes qui épousent la géométrie des images. Il fonde en 2001 une start-up (qu'il dirige jusqu'en 2007), « *Let It Wave* », dédiée au design de circuits novateurs de traitement du signal pour la vidéo haute définition. L'objet est d'implanter ces dictionnaires de bandelettes dans des puces électroniques, pour améliorer la résolution des images de télévision haute-définition.

Depuis 2008, les recherches de Stéphane Mallat portent sur les propriétés mathématiques des algorithmes d'apprentissage et des réseaux de neurones profonds, pour des données incluant un grand nombre de variables.

Enseignement du Pr Stéphane Mallat au Collège de France

Face à la malédiction de la dimensionnalité

Cours les mercredis à 9h30 (à partir du 17 janvier 2018) **:**

Pour sa première année d'enseignement au Collège de France, Stéphane Mallat introduira les outils algorithmiques et mathématiques liés à la généralisation pour des données incluant un grand nombre de variables. Il approfondira la notion de régularité qui est centrale dans ce domaine, et son utilisation par des différents types d'algorithmes, y compris des réseaux de neurones. Le cours commencera par la notion de régularité pour des données en basse dimension, à travers la transformée de Fourier et la transformée en ondelettes, avec des applications pour le débruitage et la compression de signaux. Il considérera ensuite l'apprentissage supervisé, les algorithmes à noyaux, et la performance des réseaux de neurones à une couche cachée.

Chaque séance de cours sera suivie d'un séminaire présentant l'état de l'art dans différents domaines d'applications.

Des challenges de données seront proposés aux participants, et présentés lors des premières séances. Au menu de cette année, plus de 10 challenges, pour l'économie d'énergie, le diagnostic de cancer à partir de données génomiques, la prédiction en finance, l'analyse de questionnaires, la reconnaissance d'images de célébrités ou la prédiction de scores de football. Ces challenges seront disponibles sur la page web : *challengedata.ens.fr*

L'ensemble de l'enseignement de Stéphane Mallat, ouvert à tous, sera également disponible sur notre site www.college-de-france.fr

Contacts presse : Marie Chéron / Cécile Barnier : 01 44 27 12 72 - cecile.barnier@college-de-france.fr