Projet de chaire CFM-ENS sur *Modèles et Sciences des Données*

Chaire portée par Florent Krzakala et Stéphane Mallat

1. DES MODELES ET DES DONNEES

Un des enjeux de la science moderne est de modéliser des systèmes complexes incluant un très grand nombre de variables non homogènes (des milliers ou des millions). La turbulence des écoulements, la maintenance prédictive de grands systèmes, la classification d'images, la génétique, ainsi que l'analyse de réseaux sociaux, sont quelques exemples typiques. Les enjeux du traitement des données sont fondamentaux pour de nombreux champs scientifiques comme la médecine, la physique, la géophysique, la biologie, l'économie, la sociologie... et bien sûr les mathématiques et l'informatique.

Même lorsque les équations d'évolution d'un système sont connues, la complexité due à l'interaction d'un grand nombre de variables rend très difficile l'analyse de son comportement macroscopique. Trouver des macro-variables stables capables de spécifier les propriétés du système est un enjeu majeur, qui est au cœur de nombreux domaines scientifiques, en particulier de la physique statistique dont c'est le but historique, et de l'analyse de données en grande dimension.

Des résultats numériques remarquables ont été obtenus ces dernières années, grâce au développement de la puissance des ordinateurs, à l'utilisation de masses de données ainsi qu'à de nouveaux algorithmes hautement non-linéaires comme des réseaux de neurones. Par exemple, chose inimaginable il y a encore 5 ans, des algorithmes de classification d'images sont maintenant plus efficaces que le cerveau humain pour reconnaitre des visages dans des scènes complexes, sur des images de plusieurs millions de pixels.

Le cœur des problèmes d'apprentissage en grande dimension est de construire des approximations de basse dimension afin de pouvoir calculer les paramètres du système à partir des données existantes. Ces approximations peuvent être considérées comme des modèles. Il ne s'agit donc pas seulement de « valider » des modèles préalablement établi dans une discipline, comme cela a surtout été fait jusqu'à maintenant, par exemple pour la recherche du Boson de Higgs au CERN ou pour la recherche des ondes gravitationnelles. Il s'agit véritablement de construire des modèles à partir de données et d'informations a priori sur le système, tels que l'existence d'invariants et de contraintes.

Un premier axe de recherche original de cette chaire est de créer un lien étroit entre les modèles de systèmes complexes développés dans chaque discipline, comme en physique, biologie ou géosciences, et les modèles mathématiques issus de l'analyse de données. La chaire *Modèles et Sciences des Données* de l'ENS est donc une chaire pluridisciplinaire qui fonctionnera à travers une collaboration des départements de Biologie, Économie, Informatique, Géophysique, Mathématiques, Physique, Mathématiques, Sciences Cognitives ainsi que plusieurs départements de Sciences Humaines et Sociales.

Un second axe de recherche sera basé sur l'objectif d'une compréhension détaillée des algorithmes utilisés dans la science des données. Ce domaine étant en rapide expansion, on comprend encore parfois très mal pourquoi certains algorithmes fonctionnent sur des problèmes complexes. On contrôle encore plus mal leur précision, et il peut leur arriver de produire parfois des erreurs flagrantes et dommageables. Etant donné l'impact important de ces algorithmes sur l'ensemble des sciences, sur le monde économique et plus largement la société dans son ensemble, cette compréhension et ce contrôle sont pourtant fondamentaux. La chaire des Modèles et Sciences de Données sera par conséquent l'occasion d'une approche interdisciplinaire de ces questions fondamentales, avec des méthodes issues des mathématiques, de l'informatique, et de la physique statistique, sciences qui partagent toutes trois cette ambition.

Nous présentons ci-dessous quelques directions de recherche importantes dans ces directions à l'interface entre analyse de données et étude de systèmes complexes.

2. ORIENTATIONS DE RECHERCHE

A. RESEAUX DE NEURONES PROFONDS

Les réseaux de neurones profonds (« deep networks ») ont récemment obtenu des résultats remarquables sur un large éventail de problèmes d'apprentissages en grande dimension. Ils obtiennent l'état de l'art en classification d'images avec des milliers de classes, en reconnaissance de la parole, sur des applications bio-médicales mais aussi pour la reconnaissance du langage naturel. Ils sont aussi étudiés comme modèles neurophysiologiques pour la vision et l'audition. Les capacités d'approximation en grande dimension de ces réseaux ouvrent des problèmes fondamentaux en informatique, en statistique et dans d'autres domaines des mathématiques. Au-delà des aspects numériques, établir un lien entre ces techniques d'apprentissage et les modèles scientifiques en physique, biologie, géosciences ou en économie ouvre de nombreuses perspectives. Le but de la chaire est aussi de fédérer la recherche dans ce domaine, à travers une collaboration des chercheurs concernés dans les différents départements de l'ENS, ainsi que des chercheurs invités, experts du sujet.

B. INFORMATIQUE

Le traitement de données massives a explosé grâce au développement considérable de l'informatique. Il a bouleversé de nombreuses disciplines comme la vision par ordinateur ou le traitement de la parole. Le département d'informatique de l'ENS est fort de 3 équipes de chercheurs concentrées sur ces questions. L'équipe *Willow* qui travaille en vision par ordinateur. L'équipe *Sierra* qui se concentre sur les problèmes d'apprentissage et d'optimisation en grande dimension. L'équipe *Data* qui travaille sur les problèmes de représentations de données et leurs aspects mathématiques. Ces trois équipes comptent environ 30 étudiants de doctorats, post-doc et chercheurs qui se consacrent à ces questions. L'étude des réseaux de neurones est un problème central qui est approché du point de vu de l'optimisation numérique, des applications en vision, et pour étudier leurs propriétés mathématiques d'approximation en grande dimension.

C. MATHEMATIQUES

Le traitement de données en grande dimension est devenu un domaine majeur des statistiques et des probabilités, mais ouvre aussi de nouveaux problèmes en analyse fonctionnelle, en analyse harmonique, en géométrie et plus particulièrement en théorie des groupes. Les problèmes d'apprentissage supervisés ont pour but d'approximer une fonction f(x) où x est un vecteur dans un espace de très grande dimension d, à partir d'un nombre limité p d'exemples f(xi) . Il s'agit donc de construire des approximations de fonctionnelles f(x) dont le nombre de degrés de libertés est défini par le nombre d'exemples. En général, la « malédiction de la dimensionnalité » prédit que le nombre d'exemples doit augmenter exponentiellement en fonction de la dimension d, ce qui n'est pas possible dés que d est typiquement plus grand que 20. En pratique d est souvent plus grand qu'un million.
Construire des approximations de basse dimension sur des fonctions de très grande dimension nécessite de mettre en évidence des régularités très fortes. Ces régularités sont notamment définies par des invariances sur des groupes particuliers, ce qui fait le lien avec la théorie des groupes.
La théorie de l'approximation est un sous-domaine de l'analyse fonctionnelle, où l'analyse harmonique joue aussi un rôle important pour caractériser la régularité, et l'utiliser pour réduire la dimensionnalité des approximations. La transformée de Fourier et la transformée en ondelettes jouent ici un rôle important. Les représentations multi-échelles sont des outils de modélisation qui sont aussi utilisés en physique, biologie, économie et géosciences. La construction de modèles capables de capturer les interactions à travers les échelles est un problème ouvert qui est au centre de beaucoup de questions de la physique, et notamment la turbulence pleinement développée, mais aussi en biologie ou dans les réseaux sociaux.
Les problèmes de grande dimension donnent lieux à des phénomènes de concentration. De nombreuses questions ouvertes sur la construction de modèles de processus stochastiques, non-Gaussiens, non-Markovien, ayant des dépendances de longue portée, apparaissent au travers des applications.
Tous ces aspects font partie des outils qui doivent être intégrés par les statistiques de grande dimension, puisqu'il s'agit de construire des estimateurs capables d'intégrer les données d'apprentissage, pour élaborer des modèles que l'on espère être consistants.

D. PHYSIQUE STATISTIQUE ET PHYSIQUE

À l'origine, l'objectif principal de la physique statistique a été l'étude du comportement collectif émergent d'un ensemble de composants élémentaires simples. Il se trouve que cette problématique se pose naturellement dans d'autres domaines que la physique, et en particulier dans les problèmes d'informatique aussi divers que l'optimisation combinatoire, l'inférence statistique, l'apprentissage automatique et le traitement du signal. Cette approche est devenu au cours des dernières années un domaine en évolution rapide à l'interface de plusieurs disciplines, et notamment de la science des données.

Le département de physique de l'école normale a joué un rôle de précurseur dans ce domaine, étant à l'origine, pour une large part, de l'approche de physique statistique des

codes correcteurs d'erreur ou encore des problèmes d'optimisation combinatoire. Enfin, historiquement, les réseaux de neurones multi-couches tels que ceux étudiés dans les réseaux profonds ont longtemps été un sujet fondamental dans le département, certains travaux ayant eu un impact importants (par exemple les calculs de capacité du perceptron, ou encore les travaux sur les algorithmes d'apprentissage qui ont influencé la création des Machine à Support de Vecteur).

Toutes ces thématiques font partie des modèles qui doivent être re-imaginés dans une problématique moderne, dans le cadre des réseaux profonds ou des statistiques de grande dimension. Le département de physique possède plusieurs équipes de chercheurs concentrées sur ces questions, de part sa renommée internationale dans le physique des systèmes désordonnés : en particulier l'équipe SPHINX et le groupe « réseaux complexes et systèmes cognitifs » dans le laboratoire de physique statistique (LPS), mais aussi plusieurs chercheurs dans le laboratoire de physique théorique (LPT). Ces équipes représentent un large nombre d'étudiants de doctorats, de post-docs et de chercheurs qui se consacrent entièrement à ces questions. L'un des objectifs de cette chaire sera aussi de favoriser le développement de ces approches en physique statistique, ainsi que la discussion entre ces équipes autour d'une thématique commune « données et modèles ».

Au-delà de la physique statistique, il existe de plus en plus d'expériences de physique pour lesquelles les problèmes d'analyse sont très proches de ceux qui existent dans les problématiques d'analyse de données a grande dimension. Mais les connections peuvent prendre des visages inattendus ! Un exemple, parmi tant d'autres, est donné par les expériences d'optique a haut débit au laboratoire Kastler-Brossel permettant de réaliser, avec un système physique, les mêmes projections aléatoires que celles utilisées en informatique théorique dans la réduction dimensionnelle des de systèmes de grande dimension. Il y a là encore une interface très riche avec des travaux effectués, par exemple, dans le département d'informatique.

E. CHIMIE QUANTIQUE

Les simulations numériques se sont imposées en chimie quantique au cours des dernières années, mais elles sont toujours fortement limitées par les capacités de calculs et de mémoire des ordinateurs. Il s'agit en effet de simuler des systèmes incluant un très grand nombre de particule. Cela se fait en réduisant la dimensionnalité des modèles comme par exemple en « Density Functional Theory ».
L'apprentissage à partir de bases de données chimique est une nouvelle orientation de ce domaine, dont les succès récents sont prometteurs puisqu'ils nécessitent beaucoup moins de calculs que les méthodes traditionnelles. Par exemple, l'énergie de molécules dans leur état stable peut être calculée par régression sur des bases de données d'exemples de molécules, avec des techniques utilisant certains invariants fondamentaux de la physique, mais sans modèle quantique. Certaines approches caractérisent les interactions multi-échelles d'un système à partir de dictionnaires d'invariants liés à des réseaux de neurones.
Un enjeu important est de pouvoir lier les modèles physico-chimiques avec les modèles de traitement de données. La collaboration de chimistes, de physiciens et de mathématiciens sur ce type de questions peut non-seulement faire avancer ces techniques d'apprentissages

mais aussi potentiellement aider à mieux comprendre la nature des macro-variables de ces systèmes, et par la même mieux comprendre leurs propriétés physico-chimique.

F. BIO-PHYSIQUE

Au cours des dernières années, le développement de thématiques proches de la biologie chez les physiciens a été à l'origine d'un regain d'intérêt majeur des méthodes d'analyse de données par les physiciens, souvent proches de la physique statistique. En particulier, plusieurs groupes au LPS et LPT sont parmi les plus actifs dans cette direction, sur des thématiques qui exploite largement la science des données pour la biophysique. L'un des exemples les plus étudiés est l'étude des protéines.

Les récents développements dans les méthodes de séquençage permettent de déterminer toujours plus rapidement le profil génétique de nombreuses espèces. Ces progrès ont notamment permis d'identifier de nombreuses protéines partageant une grande similarité de séquence, ce qui montre qu'elles partagent une origine évolutive commune. Les protéines de ces familles réalisent toutes des fonctions proches et permettent donc de comparer différentes solutions trouvées par l'évolution à un ensemble de problèmes fortement similaires. En effet, au cours de l'évolution de l'ARN, les séquences de protéines vont subir des mutations : Les statistiques de ces mutations et de leurs corrélations reflètent les contraintes structurelles et fonctionnelles exercées sur les biomolécules. Un défi fondamental est d'extraire et d'exploiter ces informations à partir des bases de données de séquences en croissance rapide. L 'étude de la coévolution de biomolécules, du point de vue de l'analyse de séquence, de la prédiction de structure et de la dynamique évolutive est une thématique extrêmement prometteuse, et bien représentée au département de physique. Ces champs sont à la croisée de la biologie moléculaire, structurelle, et évolutive, de la biochimie, de l'inférence statistique, de la physique statistique et de la bioinformatique.

G. NEURO-PHYSIOLOGIE ET SCIENCES COGNITIVES

Le cerveau reste bien sûr un système de traitement de masses de données étonnamment performant. On assiste actuellement à des convergences entre des modèles neurophysiologiques de la vision, de l'audition et des architectures de traitements de données avec des réseaux de neurones profonds. Les réseaux de neurones profonds sont en particulier utilisés pour interpréter la chaine de traitement de données visuelles dans les aires corticales V1, V2, V4 et pour l'audition dans la cochlée et l'aire auditive A1.
Le département de sciences cognitives à l'ENS développe de nombreux programmes de recherche, qui touchent la vision, l'audition, et le langage naturel en essayant de faire la jonction entre des observations neuro-physiologiques et des modèles algorithmiques. Il y a donc une interface très riche avec les travaux faits dans les autres départements sur l'étude de systèmes de grande dimension.

H. FINANCE

Il ne fait aucun doute que les marchés financiers sont générateurs de données assez massives, et que leur compréhension est un enjeu scientifique majeur. Un exemple : De façon intéressante, il existe de nombreuses analogies entre l'étude des corrélations dans l'évolution des protéines du paragraphe précèdent et le comportement des actions en finance. En effet, dans les deux cas, l'analyse montre l'existence de « secteurs ». Dans les protéines, ce sont des ensembles d'aminoacide fortement couplés dans l'évolution, alors que dans le cas des marchés financiers, les secteurs reflètent les corrélations entre la baisse et la hausse de différentes actions provenant des secteurs d'activités économiques. Ces analyses sont fondamentales dans l'analyse du risque en finance. La encore, ces considérations montrent l'intérêt d'une approche interdisciplinaire.

3. ORGANISATION ET FINANCEMENT DE LA CHAIRE

Le but de la chaire est de promouvoir des travaux de recherches pluridisciplinaires à l'interface entre le traitement de données et l'étude de systèmes de grande dimension. Afin de créer une communauté scientifique dans ce domaine à l'ENS, la chaire propose de développer plusieurs initiatives :

- Organisation d'un séminaire commun aux différents départements scientifiques, sur les thématiques de la chaire.
- Recrutement de post-doctorants travaillant sur cette interface.
- Aide au financement de missions sur ces thématiques
- Aide au financement pour l'organisation d'école ou de conférences sur les thèmes de la chaire.
- Invitation de chercheurs et professeurs juniors ou seniors, pour des périodes allant de 3 mois à 1 ans, pour participer à des programmes de recherche et faire des enseignements dans cette spécialité
- Aide au financement de moyens de calcul. Financement de projets d'ingénierie de traitement de données, afin d'acquérir des données, ou de les organiser pour des programmes de recherche ou d'enseignement.
- Développement d'interfaces entre des problématiques de recherche industrielles de traitement de données, et des groupes de recherche au sein de l'ENS

La chaire sera organisée par un comité de pilotage, composé d'un membre par département scientifique de l'ENS. Ce comité de pilotage sera responsable de l'allocation du financement sur les diverses activités de la chaire. La chaire sera co-portée par Florent Krzakala (directeur du programme de chaire) et Stéphane Mallat (directeur adjoint du programme de chaire)

L'ENS s'engage à soutenir les activités de recherche liées aux thématiques de cette chaire.

La Fondation CFM pour la recherche (http://www.fondation-cfm.fr) financera le budget annuel de la chaire, d'un montant de 200 000€ TTC. Ce financement pourra être interrompu

avec un préavis de deux ans. Les activités de la chaire seront présentées tous les ans au conseil d'administration de la fondation CFM. Elles seront examinées tous les deux ans par un comité scientifique composé de trois membres extérieurs, nommés d'un commun accord par la direction de l'ENS et la fondation CFM. Ce comité extérieur donnera un avis sur les orientations de la chaire, avis qui sera examiné par la fondation CFM ainsi que par la direction et le Conseil scientifique de l'ENS.