Analyse de données scientifiques

Les nouveaux enjeux planétaires relatifs à l'écologie, au climatique, aux impacts de la pollution, à la prévention des risques (sismique, énergétique...) mettent en œuvre des systèmes fortement complexes. Les disciplines associées sont les sciences de la terre, la météorologie, la tectonique des plaques et les sciences des équilibres océaniques… Ces sciences ne peuvent s'appuyer que sur de nombreuses observations. Ainsi, les données brassées par ces sciences sont colossales et se comptent en Térabits. De même l'astronomie (astrophysique, cosmologie) développe des moyens d'accès aux limites de l'Univers de plus en plus pointues, ce qui implique des flots de données gigantesques. L'enjeu du XXI ème siècle réside dans l'exploitation intelligente de toutes ces données, peut-êtte une question de survie pour notre espèce. Le Big Data et la data science sont-ils des éléments de réponse ?
Aujourd'hui, la plupart des applications se situent au niveau du business (Entreprise, finance, marketing, grande consommation) du monde de la santé et du développement d'objets technologiques (Vision, reconnaissance de la parole, jeux vidéo, transports intelligents …).
L'apprentissage automatique a récemment contribué à de nombreuses découvertes majeures en physique, en particulier en physique des particules et en astronomie. En physique des particules, les données du collisionneur sont massives et contiennent beaucoup de bruit. L'application de l'apprentissage automatique pour passer au crible les données afin de déceler d'éventuels cas d'une particule ou d'un processus (comme le boson de Higgs) permet aux chercheurs de se concentrer sur des résultats intéressants dans une pile de trillions de données, ce qui permet de faire des découvertes plus rapides et de perdre moins de temps. Il permet également aux chercheurs de passer au crible des expériences non directement liées au processus qu'ils étudient, mais dont la présence est possible dans cette expérience.

En astronomie, la science des données et l'apprentissage automatique sont souvent utilisés pour identifier les processus ou les caractéristiques intéressantes des données astronomiques (nouvelles galaxies, supernovas, trous noirs...). Encore une fois, les données sont habituellement massives et les

signatures sont faibles.

Au cours des dernières années, le domaine de la physique s'est beaucoup développé. Surtout la physique expérimentale. Les données recueillies manuellement il y a quelques décennies le sont maintenant avec un ordinateur ou un groupe d'ordinateurs et ont permis de capturer plus de points de données et de paramètres. Une seule seconde de données peut contenir entre quelques centaines et des millions de points de données. À mesure que le nombre de ces données augmente, la data science est le moteur qui extrait l'information ténue, la cohérence rare, les liens précieux au sein de cet océan de données.

Hélas ces applications ne sont que trop rares et localisées à quelques secteurs (défense, astrophysique …). Il serait judicieux d'étendre ce type d'approche aux enjeux actuels :

- Encourager la trans disciplinarité et le transfert de techniques d'une discipline à l'autre
- Exploiter la littérature scientifique du XX ème siècle dans l'ensemble des disciplines : beaucoup d'idées ont déjà été pensées.
- Elargir les techniques de data science qui sont pour beaucoup linéaires à une approche de la complexité (non linéaire)

Dans la plupart des applications en data science, le chercheur se préoccupe de la répartition des cas les plus probables et reproductibles. Cette zone d'intérêt correspnd à une courbe en cloche : la gaussienne dans la zone es 3 écarts types. Ainsi les méthodes linéaires (filtrage, estimation, prédiction) s'avèrent suffisantes. Cependant, dans certaines applications de la physique (systèmes complexes), les grands écarts à la moyenne ne sont pas rares et ce sont souvent eux qui sont à l'origine des risques. Ainsi, dans ces applications, le chercheur doit caractériser le comportement asymptotique des distributions, le monde des grandes valeurs loin de l'équilibre, donc soumises à la non linéarité.

Dans ce cas, le modèle linéaire gaussien ne suffit plus. On rentre dans la science non linéaire, seule capable de calculer avec suffisamment de précision les grandes valeurs des états des systèmes complexes loin de l'équilibre.

L’exemple en astrophysique est la répartition de masse dans l’univers aux grandes distances, pour laquelle on n’a que peu de données et qui nécessiterait des développements de modèle d’univers non linéaires, permetant d’extrapoler avec le peu de données que nous pouvons capter.

Dans son ouvrage sur la science du XXI ème siècle Roger Penrose montre qu’aujourd’hui beaucoup de théories de physique fondamentale , comme la théorie des cordes, sont de nature théorique et validées plus par un sentiment d’élégance et d’esthétique mathématique que par l’expérience et le réel. Je pense que là aussi, le data scientist se doit de changer son fusil d’épaule en faisant de sa discipline une science expérimentale. En effet, la nature et les systèmes complexes sont régis par des lois d’efficacité et de simplicité qui, à priori, ne sont pas corrélés à un esthétisme mathématique. L’approche de cette simplicité, de cette efficacité passe par l’observation et la démarche expérimentale.

Nous avons donc fait le pont entre la méthode expérimentale et la data science, comme vecteur de cette démarche expérimentale. Puis nous illustrons cette correspondance par une applications dans le monde de la physique (radar) qui illustrent l’apport des sciences non linéaires et de la complexité. Le tableau suivant décrit chacune des étapes de notre démarche.

Démarche empirique A. D. de Groot	Data Science	Analyse de la rétrodiffusion de la mer
Historique : recherche de l’ensemble des travaux sur le sujet	Econométrie	Travaux de calcul de prévision de houle, de calcul du spectre spatio temporel de la mer. Travaux de modélisation de rétrodiffusion de mer dans plusieurs bandes de fréquence dont radars et sous différentes conditions météo (ENSIETA/ Brest, Alyotech) Travaux d’analyse non linéaires (chaos dynamique) de signaux de

		fouillis de mer (Université de Deft) Campagnes d ‘enregistrements de données de radar naval (radar de surveillance côtière) à faible incidence
Observation : il s’agit de mettre en évidence un phénomène. Ce peut être avec un instrument de mesure ou par recherche bibliographique ou historique.	Compréhension du problème. Définition des objectifs Comment le problème peut-il être résolu avec les donnés disponibles? Compréhension et analyse des données (Visualisation, statistique descriptive, …) Préparation des données (manquantes, extrêmes, aberrantes …): nettoyage, mise en forme (préfiltrages)	Analyse statistique des données radar (Energie rétrodiffusée par la mer) : Nous obtenons des lois de répartition à queue Lourde (Les grandes valeurs ne sont pas rares) Comparaison des histogrammes à diverses lois : lognormale, K, Rayleigh. Aucun de ces modèles ne semble concluant.
Induction : il s’agit d’établir une hypothèse la plus probable sur le pourquoi de ce que l’on a observé. On procède par induction. Plusieurs hypothèses peuvent surgir. La plus simple (rasoir d’Occam) sera privilégiée.	Poser les hypohèses. Recherche bibliographique. Choix d’outils et de modèles	Recherche bibliographiqueDans certaines configurations, les vagues de gravité se comportent comme des miroirs réfléchissant une grande part de l’énergie émise : facettes quasi verticales. Comme les cas de fortes rétrodiffusion ne sont pas rares, nous considérons que la répartition des « pentes » de vagues obéit à une loi Lévy stable de paramètre de forme DLe paramètre de forme D ne dépend que de la vitesse du vent.
Déduction : Si on pousse le raisonnemet, quels sont les phénomènes qui découlent naturellement de l’hypothèse. Comment les mettre en évidence et les observer ?	Conséquences logiques du modèle et des hypothèses. Définition d’expérimentations supplémentaires pour étayer les hypothèses et le modèle.	Nous justifions ces hypothèses par 1 Des données expérimentales de la littérature (expressions empiriques des coefficients de rétrodiffusion) 2 Des calculs : répartition des pentes et périodes d’un signal quasi monochromatique
Tests : Validation des phénomènes déduits	Validation du modèle et des outils sur les	Validation à partir de données et modèles empiriques de la littérature

par l'expérience.	données.	
Evaluation : C'est la construction du modèle et des éléments théoriques.	Formulation d'une théorie: on peut construire un simulateur. Un simulateur est un outil de calcul complexe qui permet de prédire des configurations du système en fonction des paramètres d'entrée (petit nombre de paramètres avec un sens physique et facilement interprétables.	Construction 'un simulateur avec les paramètres d'entrée suivants : 1 La surface de la mer interceptée par le radar se compose de points brillants (vagues de gravité), associées à des facettes de pente aléatoire 2 Loi de répartition des facettes de mer (Lévy) : paramètre de forme D et paramètre d'échelle 3 Configuration géométrique du radar (incidence) et forme de l'onde incidente 4 D fonction de la vitesse du vent
Application : ce sont les applications dans le domaine de la science ou de la technologie.	Prédiction	Tracking de cibles de faible pouvoir réflecteur dans un environnement naval (mise au point d'algorithmes, analyse et prédiction de performances)