GÉOSTATISTIQUE ET PROBABILITÉS APPLIQUÉES

La pratique de la Géostatistique est, d'abord, une occasion de rencontres: rencontres entre des champs d'application variés et parfois inattendus, entre des problématiques sans cesse renouvelées, et également une confrontation entre des objectifs et contraintes purement techniques d'une part et les exigences sociales, économiques, environnementales d'un monde complexe d'autre part. Autrement dit, tout en étant fiers de ce que le néologisme "géostatistique" jadis forgé à l'École des Mines ait trouvé droit de cité dans le Petit Larousse, il est satisfaisant d'observer, au fil des ans, que l'immuable définition qu'en donne le dictionnaire s'éloigne de plus en plus de la réalité, et que notre discipline trouve à s'exprimer bien au-delà de la simple estimation des gisements miniers. De fait, dans tout domaine où des jeux de données numériques présentent une organisation spatiale ou temporelle, la Géostatistique a les outils pour apporter un éclairage original, à la fois constructif et sans concession. Il semble que cet aspect transversal et non-conformiste de la Géostatistique constitue désormais son caractère dominant au regard des optionnaires, et nous ne manquerons pas dans le futur de justifier cette appréciation. Ainsi, chaque année, la diversité des vœux des étudiants constitue une chance exceptionnelle de tester des méthodes nouvelles et de parcourir des domaines nouveaux, et la garantie d'insuffler un surplus de dynamisme à l'équipe encadrante. Mais la mise en œuvre d'une Géostatistique de qualité exige en permanence d'assurer un équilibre, parfois délicat, entre des exigences souvent contradictoires: garantir une rigueur théorique indispensable à la fiabilité des résultats tout en conservant un point de vue pragmatique et réaliste afin que les conclusions abstraites trouvent à s'appliquer sur le terrain. Sans oublier une indispensable déontologie, dans des domaines où souvent les contraintes économiques ou environnementales soumettent le géostatisticien à des pressions qui ne relèvent plus de la science ou de la technique...

Ouverture et équilibre: c'est dans cet esprit que nous avons continué à proposer à la promotion actuelle un voyage de deux semaines en Guyane où, dans le contexte inhabituel et parfois tourmenté d'un DOM, les visites à des laboratoires, à des industriels et à des organismes institutionnels ont permis tout à la fois d'élargir l'horizon des optionnaires et de susciter un échange ouvert et fructueux avec nos interlocuteurs. Partie intégrante de la scolarité, la mission en Guyane constitue pour les optionnaires la phase d'initiation à la réalité du terrain.

Enfin, le souci d'ouverture s'est à nouveau exprimé au niveau des sujets de l'option. Outre la variété des champs d'application, une multiplicité de méthodes statistiques est employée. Nous avons ici l'illustration que notre démarche méthodologique peut trouver à s'appliquer dans de multiples domaines et cela souligne le caractère généraliste de l'option Géostatistique et Probabilités Appliquées, tant en ce qui concerne les champs applications abordés que les méthodes mathématiques mises en jeu.

Emilie CHAUTRU, Thomas ROMARY, Hans WACKERNAGEL

GEOSTATISTICS AND APPLIED PROBABILITY

The aim of Geostatistics is to study quantitative phenomena that are structured in space and/or time. Engineers are almost inevitably faced with problems related to geostatistical techniques such as the evaluation of natural reserves, the analysis of time series, cartography, etc., and, broadly speaking, any processing of "regionalized variables" according to the terminology of G. Matheron, the founder of geostatistics.

The Geostatistics and Applied Probability Minor gives priority to probabilistic models and statistical methods, and in particular focusses on their application to the processing of spatial and temporal data.

The training in this Minor is aimed essentially at providing students with a critical mastering of some of the tools which they might need to use during their internship. As an introduction to "geostatistics in the field", the Minor provides an opportunity of entering into contact with companies and personnel working with geostatistics in fields of application that correspond as much as possible to the themes of particular interest to the students. It is essential that the students acquire a sense of balance between an empirical approach and a mathematical approach to a problem, a sense of balance which the third-year internship will put into practice in real terms.

The presentations this year again reflect the great variety of themes and applications that are encountered in this branch of applied statistics.

Emilie CHAUTRU, Thomas ROMARY, Hans WACKERNAGEL

GÉOSTATISTIQUE ET PROBABILITÉS APPLIQUÉES
GEOSTATISTICS AND APPLIED PROBABILITY

Mardi 10 septembre *Tuesday 10th Septembre* ■ **L224**

Tom
MONNIER

10h-11h

Prédiction spatio-temporelle de la qualitée de l'air dans les villes

Spatio-temporal forecasting of air quality in cities

La pollution de l'air cause en Europe près de 500 000 morts par an et est un sujet préoccupant dans la plupart des grandes villes mondiales. Estimer et prévoir la qualité de l'air que nous respirons est déterminant pour à la fois les entreprises désirant quantifier leurs impacts, les pouvoir publics qui doivent y faire face et les particuliers souhaitant contrôler les polluants qu'ils inspirent. L'objectif du stage est de construire deux modèles. Le premier prédit à plusieurs jours les concentrations des différents polluants dans les grandes villes mondiales à partir de séries temporelles multivariées provenant de données météorologiques et de qualité de l'air. Le second prédit à l'échelle d'une rue la pollution actuelle à partir d'un modèle physique et de séries temporelles géolocalisées venant de stations de mesures fixes et des capteurs de gaz personnels connectés que vend l'entreprise. L'idée à terme est de fusionner ces deux modèles pour former un seul modèle spatiotemporel.

Air pollution in Europe causes around 500,000 deaths a year and is a cause for concern in most major world cities. Estimating and predicting the quality of the air we breathe is crucial for companies wishing to quantify their impacts, public authorities that must deal with them, and individuals wishing to control the pollutants they inspire. The objective of the internship is to build two models. The first predicts multi-day concentrations of different pollutants in major global cities from multivariate time series derived from meteorological and air quality data. The second predicts on a street scale the current pollution from a physical model and geolocated time series from stationary measurement stations and connected personal gas sensors sold by the company. The ultimate idea is to merge these two models to form a single spatiotemporal model.

Plume Labs
Paris

Hugo
ROLLIN

11h-12h

Reconnaissance et mise en lien d'images et de textes dans des photographies de documents

Image-text recognition and matching in documents photographs

Parmi la richesse d'objets produits par l'activité artistique, les catalogues d'exposition sont une source principale pour retracer la création et l'histoire des oeuvres dans le domaine de l'art. L'objectif du stage est d'appliquer des méthodes d'analyse d'images afin d'extraire à grande échelle les informations structurées contenues dans ces catalogues. Dans un premier temps, il sera nécessaire de développer et d'évaluer des méthodes pour automatiquement extraire les images et reconnaitre le texte au sein des documents. On explorera en particulier la généralisation des méthodes d'apprentissage profond pour la segmentation entrainées sur des données synthétiques afin d'être appliquées sur des documents numérisées. Par la suite, des approches permettant de mettre en lien le texte et les images extraites seront étudiées et implémentées. Enfin, d'autres aspects connexes de recherche tels que l'analyse de la mise en page ou la reconnaissance de textes manuscrits seront envisagés.

Among the wealth of artifacts produced by artistic activity, exhibition catalogues are a prime source for retracing the making of facts and knowledge within the art field. The goal of the internship is to apply computer vision methods to extract structured information from these catalogues at large scale. First, it will be necessary to develop and evaluate methods to automatically extract images and recognize the text from the documents. In particular, we will study the generalization of deep learning based image segmentation approaches trained on synthetic images on real scans. Thereafter, methods to accurately match the text with corresponding images will be studied and developed. Finally, other connected paths of research such as document layout analysis or handwritten text recognition would be considered.

ENPC
Marne-la-Vallée

Jean-Rémi CONTI

13h30-14h30

Méthodes d'apprentissage statistique pour données spatiales

Statistical learning methods for spatial data

Le cadre de validité des algorithmes de machine-learning stipule en général que les données d'apprentissage sont des réalisations i.i.d. d'un vecteur aléatoire générique. Dans de nombreuses situations, auxquelles les applications en geosciences se confrontent en particulier, les données sont de nature spatiale. Si la variabilité de ces données est traditionnellement estimée au moyen de techniques de krigeage, les problématiques prédictives inhérentes à l'exploitation de ces données désormais disponibles en masse à l'heure du déploiement généralisé de capteurs suggèrent d'étendre les approches algorithmiques de minimisation de risque empirique au cadre spatial, comme cela est fait pour les techniques de moyennes locales fondées sur le partitionnement récursif de l'espace des observations. Ce stage visera à définir un cadre mathématique rigoureux permettant d'établir la capacité de généralisation des règles obtenues par minimisation du risque empirique, ainsi qu'à étudier d'autres extensions au contexte spatial d'algorithmes populaires (SVM, ensemble learning), d'un point de vue théorique et expérimental à la fois.

The validity framework of machine-learning algorithms generally states that the learning data are i.i.d. realizations of a generic random vector. In many situations, to which applications in geosciences are confronted in particular, the data are of a spatial nature. While the variability of these data is traditionally estimated by means of kriging techniques, the predictive problems inherent in the exploitation of these data now available in mass at the time of the generalized deployment of sensors suggest to extend the algorithmic approaches of minimization of empirical risk to the spatial framework, as is done for 'local averages' techniques based on recursive partitioning of the observation space. This internship will aim at defining a rigorous mathematical framework allowing to establish the ability of generalization of the rules obtained by minimization of the empirical risk as well as to study other extensions to the spatial context of popular algorithms (SVM, ensemble learning), from a theoretical and experimental point of view.

Télécom ParisTech
Paris

Ouns EL HARZLI

14h30-15h30

Apprentissage statistique pour la détection d'amas de galaxie

Machine learning for galaxy cluster detection

L'application des techniques d'apprentissage statistique nous permettent d'analyser les très grandes quantités de données obtenues par les sondages du ciel à grande échelle, et les futurs sondages, dans le but d'améliorer notre compréhension du modèle standard de la cosmologie. En effet, les amas de galaxies, qui définissent la distribution de matière dans l'Univers, sont un outil puissant pour tester le modèle standard et contraindre ses paramètres. Les futurs sondages du ciel, par des télescopes terrestres et spatiaux, comme la mission spatiale Euclid, fourniront des catalogues toujours plus complets des amas de galaxies comme produits scientifiques. Nous explorerons plus particulièrement les méthodes d'apprentissage profond pour détecter des amas dans les sondages à grande échelle.

Application of machine learning techniques to cosmology permits us to analyze the large quantity of data obtained with current surveys and expected from future surveys with the aim of improving our understanding of the cosmological model.
Galaxy clusters are powerful probes of the standard cosmological model, and future ground- and space-based surveys, such as the Euclid space mission, will provide catalogs of newly discovered galaxy clusters as one of their primary science products.
We will explore deep learning techniques to detect galaxy clusters in large-scale surveys

Caltech, *Etats-Unis*
Observatoire de Paris, *Paris*

Yoann PRADAT

15h30-16h30

Étude d'un modèle de représentation de surfaces en 3D avec une topologie de sphère

Studying a model for the 3D representation of surfaces with a sphere topology

L'objectif du projet est d'étudier un modèle de géométrie différentielle pour la représentation 3D de surfaces avec une topologie de sphère. Le modèle

The aim of the project is to study a differential geometry model for the 3D representation of surfaces with a sphere topology. The model of interest will

d'intérêt sera conçu pour incorporer un contrôle explicite sur la courbure locale de la surface paramétrique. Il sera construit comme une extension d'un modèle d'ordre inférieur préexistant permettant un contrôle local sur les premières dérivées directionnelles d'une forme.
Les deux objectifs du projet sont de caractériser mathématiquement le modèle et de mettre en œuvre des exemples de validation du concept pour le paramétrage de volumes 3D stéréotypés.

be designed to incorporate explicit control over the local curvature of the parametric surface. It will be built as an extension of a preexisting lower-order model allowing local control over the first directional derivatives of a shape.
The two objectives of the project are to mathematically characterize the model and implement proof-of-concept examples of its use when parametrising stereotypical 3D volumes.

European Bioinformatics Institute
Hinxton, Royaume-Uni

Audrey KERVELLA

16h30-17h30

Science des données et interprétation médicale

Data Science and medical interpretation

L'analyse de données est aujourd'hui un indéniable levier d'amélioration de performances pour les entreprises. Il s'agit de traiter les données de manière assez pertinente afin qu'elles jouent un rôle moteur dans la prise de décision en entreprise. Artefact, un des principaux leaders en la matière, apporte son aide à de nombreux groupes français et internationaux, avec, dans le domaine de la data science, des missions variées : de la prédiction de ventes dans le domaine du luxe, au yield management dans le tourisme, ou encore à l'optimisation d'un réseau routier pour la télécommunication.
Poussé par les progrès dans le domaine de la reconnaissance d'images médicales, Artefact a également pris part à un projet dans le domaine de la science de données médicales. Dans le cadre de mon stage, j'ai intégré la cellule AI vs Lymphoma. L'objectif de ce projet est d'aider les chercheurs de Lysarc à différencier les lymphomes folliculaires (LF) des hyperplasies folliculaires (HF) qui, contrairement aux lymphomes, ne sont pas considérées comme dangereuses. Le but est de mettre en avant, à travers des techniques alliant algorithmes de deep learning et statistiques, un modèle de classification sur images ainsi qu'une méthodologie particulière. Cette méthodologie permettrait aux chercheurs d'établir des distinctions précises entre LF et HF en fonction de points définis comme pertinents sur une image donnée.

Data analysis has become an undeniable performance lever for companies. Behind the technical breakthrough of data science, lies the strategic challenge of treating the data in a fairly relevant way so that they play a leading role in business decision-making. Artefact, one of the leaders in the field, provides assistance to many French and international groups, who wish to make the best use their data. Artefact has contributed to various projects: from sale forecasting in the luxury branch, to yield management in tourism, to optimization of road networks for telecommunications.
Driven by advances in the field of medical image recognition, Artefact has also taken part in a project in the field of medical data science. The purpose of my project is to help Lysarc researchers differentiate follicular lymphoma (LF) from follicular hyperplasia (HF) which, unlike lymphomas, are not considered dangerous. The aim is to highlight, through techniques combining deep learning algorithms and statistics, a classification model on images as well as a methodology, allowing researchers to make specific distinctions, based on relevant points on given images.

Artefact
Paris

Yanis TAZI

9h-10h

Machine learning statistique pour la recherche sur le cancer

Statistical machine learning for cancer research

La leucémie aiguë myéloblastique ou « LAM » est un cancer qui prend naissance dans les cellules souches du sang. En se développant, les cellules souches du sang deviennent des cellules blastiques autrement appelées blastes et considérées comme des cellules sanguines immatures. Dans le cas de la LAM, il y a une surproduction de cellules blastiques.
L'objectif de ce projet de recherche, dans un premier temps, est d'utiliser et développer des outils d'analyses statistiques et d'apprentissage non-supervisés, afin de redéfinir les différents groupes de patients atteints de cette maladie en se basant sur leurs données génétiques, phénotypiques ainsi que cytogénétiques. Cette redéfinition des groupes permettra d'analyser l'importance des variables de chaque groupe et de mettre en place un nouveau protocole de diagnostic entre la « Myélodisplasie » et la « LAM » qui sont aujourd'hui différenciées l'une de l'autre par un simple cut-off du taux de blastes médullaires.
Dans un second temps, il faudra s'aider des groupes ainsi formés afin de mettre en place des algorithmes d'analyse de survie pour que les médecins puissent accompagner au mieux chaque patient durant la phase de traitement. L'objectif de cette deuxième partie portera sur la réflexion de la mise en place de nouveaux algorithmes de pénalisation qui prennent en compte la censure et la troncature de données.

Acute myeloid leukemia or AML is a cancer starting in blood stem cells. While growing, those blood stem cells become blast cells considered as immature blood cells. In the case of AML, there is an overproduction of blast cells.
In the first instance, the goal of this research project is to develop and use statistical tools and unsupervised learning in order to refine the patient groups developing AML based on their genomic, phenotypic and cytogenetic data. By redefining those different groups, we will be able to find important features for each group and to set up a new diagnostic protocol to distinguish AML from MDS (myelodsiplastic syndromes) who are nowadays differentiated by a simple blast rate cut-off.
Second, we will use those groups to set up survival analysis algorithms in order for the doctors to better assist each patient during treatment phase. The purpose of this second part will also be to work on new penalization algorithms taking into account censored and truncated data.

Memorial Sloan Kettering Cancer Center
New York, États-Unis

Alexandre BANQUET

10h-11h

Application du traitement de langage naturel et du deep learning à la recherche d'informations biomédicales

Applying natural language processing and deep learning for open-domain biomedical question answering

Chaque année, 8 millions de requêtes sur les médicaments sont soumises à des moteurs de recherche. La formulation de ces questions varie fortement et les informations que les utilisateurs obtiennent proviennent de sources très variées, qui sont bien souvent non-structurées et peu dignes de confiance. En utilisant les méthodes de traitement de langage naturel, POSOS, une startup créée en 2017, développe une technologie capable, sur le long terme, de comprendre n'importe quelle question posée sur un médicament et de générer automatiquement une réponse en croisant différentes sources de données fiables. Cette technologie, qui pourra être utilisée par des professionnels de santé, vise à réduire les risques liés à une mauvaise posologie de médicaments, qui entraine chaque année l'hospitalisation de 144 000 personnes en France.
Les systèmes de question-réponse constituent un

Eight billion drug queries are submitted each year in search engines. The way these questions are asked varies significantly and the information the applicant hopes to obtain comes from a very large number of sources, most of which are unstructured and not trustworthy. Using new natural language processing techniques, POSOS, a start-up created in 2017, develops a technology able, in the long term, to understand any questions asked about a drug and automatically generate an answer by crossing different reliable data sources. This technology, which can be used by many healthcare professionals, aims at reducing the risk of misuse of medicines, which still causes more than 144,000 hospitalizations each year in France.
Question answering is a broad domain, including Information Retrieval (IR), Machine Reading Comprehension (MRC), answer generation tasks. The

vaste domaine, comprenant la Recherche d'Informations (IR), la Compréhension de Lecture Automatisée (MRC) et la génération automatique de réponses. L'objectif de ce stage est de se concentrer sur la partie requête de documents pertinents du système de question-réponse. La Recherche d'Informations est un domaine où des techniques robustes et efficaces ne reposant pas sur l'apprentissage artificiel ont été utilisées pendant longtemps, notamment dans de nombreux moteurs de recherche comme Google. Cependant, ces systèmes sont parfois adaptés manuellement, pas toujours très évolutifs, ni spécifiques à un domaine. Par conséquent, des réseaux de neurones commencent à être utilisés pour effectuer des requêtes de documents et pour classer leur pertinence.

goal of this internship is to focus on the document retriever part of the question answering system. Information retrieval is a domain where efficient and robust non-machine learning techniques have been used for a long time, in first-class web search services such as Google. However, these systems are sometimes heavily hand-crafted, not always very scalable, nor domain specific. As a consequence, end-to-end neural networks are starting to be used to do query-documentmatching and ranking.

Posos
Paris

Clément ACHER

11h-12h

Détection de commandes vocales sur large vocabulaire avec contraintes en temps de calcul et en mémoire

Small footprint algorithm for keyword spotting on large vocabulary

La détection de commande vocale à l'aide de modèles contraints en temps de calcul et en mémoire est un domaine de recherche actif en reconnaissance vocale. Les assistants vocaux comme Google Home ou Alexa utilisent ce type d'algorithme en continu pour détecter leurs wake words respectifs ("Hey Google", "Alexa") pour n'envoyer les paroles enregistrées au cloud qu'après détection de la commande. Snips, qui développe un assistant vocal sans utilisation de cloud dans un soucis de respect de la vie privée propose également des solutions de détections de commandes vocales plus générales pouvant être intégrées dans des micro-contrôleurs peu onéreux pour répondre à des dizaines de commandes comme "allume la lumière".
L'approche classique pour ce genre de problème consiste à constituer un jeu de données d'exemples audio annotés de la commande vocale correspondante et d'entrainer à partir de ceux-ci des réseaux de neurones récurrents. Si les modèles qui découlent d'une telle stratégie ont l'avantage d'être performants et d'avoir une mise en place relativement simple, ils ne sont utilisables que pour la tâche pour laquelle ils ont été entrainés. C'est un inconvénient majeur car la constitution d'un jeu de données est à la fois onéreux et chronophage. On peut donc envisager une approche différente, plus proche des techniques de reconnaissance automatique de la parole : non plus tenter de reconnaître directement les commandes mais plutôt des séquences de phonèmes. Si une séquence de phonèmes correspondant à une commande donnée est détectée, on peut alors inférer que la commande a été prononcée. L'ajout d'une nouvelle commande au modèle, en général un réseau de neurones récurrents, ne demande dans ce cas que l'addition de la liste de phonèmes correspondante. Le contexte impose cependant des contraintes matérielles très fortes, l'algorithme étant destiné à des micro contrôleurs dont les caractéristiques sont plus que modestes. Comment donc implémenter et optimiser les performances d'un tel modèle, plus général et qui nécessite des étapes supplémentaires de décodages, dans ces contraintes de calculs et de mémoire?

Small-footprint keyword spotting systems is an active field of research in speech recognition. Voice assistants like Google Home or Alexa use such algorithms to spot their respective wake words ("Hey Google", "Alexa") in order to stream to the cloud only after detecting a command. Snips also develops a voice assistant that runs only on device and offers solutions to spot voice commands on cheap micro-controller units that can spot a dozen of them such as "turn on the light".
A common approach to tackle this kind of keyword spotting problem is to gather annotated datasets at the command level and use them to train recurrent neural network models. The classifiers obtained with this strategy are performing well and are rather easy to implement, but they can only be used for the task they have been trained for. It is an important downside as crafting datasets is both expensive and time consuming. We can imagine another approach, more similar to what is done in automatic speech recognition : instead of trying to directly classify the keywords (either the stream contains one or not), the model can have the goal to transcript the sequence of phonemes from the audio stream. If the sequence corresponds to a voice command to be spotted, we can infer that the command has been said. The addition of another keyword or command to the model, usually a recurrent neural network, doesn't require more than adding the corresponding sequences of phonemes. The context requires the algorithm to be as light as possible as it is expected to run on micro controller units that have very small computational power. How can this more complex model be implemented and optimised to make them work on device with low computational cost and memory footprint?

Snips
Paris

Romain
ROLLIN

13h30-14h30

Optimisation multi-objectif dans un espace moléculaire continu construit à partir de données non labellisées

Multi-objective optimization in a data-driven continuous representation of molecules

L'élaboration de nouveaux médicaments requiert à l'heure actuelle énormément de temps, de connaissances spécialisées et d'expériences biologiques excessivement coûteuses. Au cours de la dernière décennie, de nouvelles approches utilisant l'IA ont suscité un intérêt croissant dans le milieu du « drug-design » parce que ces méthodes pourraient réduire la dimensionnalité de l'espace de recherche et le temps nécessaire pour trouver une nouvelle molécule candidate prometteuse. L'application de l'IA au « drug-design » est un problème d'optimisation dans lequel l'algorithme recherche les molécules qui maximisent une fonction objective. Cependant, l'optimisation dans l'espace moléculaire est extrêmement difficile car l'espace de recherche est très grand, discret et non structuré. De plus, en pratique, les entreprises pharmacologiques recherchent des molécules qui répondent à de nombreux critères différents. Ces molécules doivent être efficaces mais aussi non dangereuses pour l'être humain. Pourtant, la communauté scientifique s'est principalement concentrées sur l'optimisation d'une seule fonction objective, en combinant de manière assez arbitraire les différents critères. Au court du stage, nous combinerons deux nouvelles approches récemment développées. Tout d'abord, nous construirons une représentation moléculaire multidimensionnelle continue qui sera élaborée à partir de données, en utilisant un auto-encodeur. Ce nouvel espace, couramment appelé espace latent moléculaire, présente plusieurs avantages : d'abord, il est vraiment plus facile et pratique de travailler dans un espace continu, puis de retourner dans l'espace moléculaire en utilisant le décodeur. Deuxièmement, l'espace latent est construit en utilisant de grands ensembles de composés chimiques non étiquetés, ce qui est un atout important puisque l'étiquetage des bases de données chimiques est en pratique très coûteux. Enfin, la construction de cet espace latent ne nécessite a priori pas d'avoir une expertise en chimie ou biologie. Cet espace latent nous permettra ensuite de tester différentes techniques d'optimisation multi-objectifs pour ensuite les comparer à l'état de l'art.

Many of today's drug discoveries require a huge amount of time, expertise knowledge and excessively expensive biological experiments for identifying the chemical molecular properties. In the past decade, new approaches using AI caught a growing interest in the drug design community because they could reduce the dimensionality of search space and the time to find a new promising candidate molecule. Combining AI and « drug design » is an optimization problem in which the algorithm search for the molecules that maximize an objective function. However, optimization in molecular space is extremely challenging because the search space is very large, discrete, and unstructured. Moreover, in practice the pharmacological companies are looking for molecules that complete many different criteria - they must be efficient, but also non dangerous for a human being. Yet, people have mainly focus on optimizing one single objective function, combining quite arbitrarily the different criteria. In this internship, we combine two new approaches that were recently developed. First, we build a molecular multidimensional continuous representation that is data-driven, using an auto-encoder. This new space called latent space has several advantages; first, it is really easier and convenient to work in a continuous space and then go back to the molecular space using the decoder. Second, the latent space is built using large sets of unlabeled chemical compounds which is an important asset since labeling chemical databases is in practice really expensive. Third, the construction of this space does not require any expert knowledge in chemical biology. This latent space will then enable us to test different multi-objective optimization techniques in order to compare them to state-of-the-art existing techniques.

Iktos
Paris

Tristan
DOT

14h30-15h30

Analyse multimodale de la locomotion: détection et reconnaissance de motifs

Multimodal analysis of locomotion: pattern detection and recognition

Ce stage s'inscrit dans le contexte de la collaboration entre l'UMR CMLA (ENS Paris-Saclay) et l'UMR COGNAC-G (Université Paris Descartes), constituée de médecins, d'ingénieurs et de mathématiciens et visant l'étude du comportement humain et animal grâce à la constitution et à l'analyse de bases de données.
Dans ce cadre, la marche de plusieurs sujets a été enregistrée de façon synchronisée grâce à un tapis de mesure et des accéléromètres triaxiaux, permettant une analyse multimodale des signaux. Le but du stage consiste à utiliser et développer des outils de reconnaissance de formes et d'apprentissage pour le dépouillement de cette cohorte. Il s'agira tout à la fois de valider et développer des algorithmes de détection de pas (reconnaissance de formes, analyse de séries temporelles: par déformation temporelle dynamique, décomposition en ondelettes, etc), d'étudier les liens existants entre les motifs d'accélération et les motifs de pression, et d'analyser automatiquement et classifier les motifs observés (par apprentissage de dictionnaire, notamment).

This internship is part of the collaboration between the CMLA laboratory (ENS Paris-Saclay) and the COGNAC-G laboratory (University Paris Descartes), made up of doctors, engineers and mathematicians, and aiming at the study of human and animal behavior through the creation and analysis of databases. In this context, the walking of several subjects was recorded synchronously thanks to a pressure-sensitive walkway and triaxial accelerometers, allowing a multimodal analysis of the signals. The purpose of the internship is to use and develop form recognition and learning tools in order to robustly analyze this cohort. The goal will be, at the same time, to validate and develop algorithms of step detection (by pattern recognition and time series analysis: using dynamic time warping (DTW), wavelet decomposition, etc.), to study the existing links between the acceleration and pressure patterns, and to automatically analyze and classify the patterns observed (thanks to dictionary learning, in particular).

ENS *Paris Saclay*
et Université Paris Descartes

Matthieu
WITTIG

15h30-16h30

Analyse de la gestion de l'équilibre offre/demande sur les marchés de l'électricité

Analysis of the demand/supply equilibrium management on power markets

La DOAAT (Direction de l'Optimisation Amont-Aval et Trading) est chargée de gérer l'équilibre offre-demande sur le périmètre EDF : elle ajuste à tous les horizons de temps production et consommation, et définit les programmes de production des actifs d'EDF tout en respectant leurs contraintes techniques de son parc et en optimisant économiquement leur usage.
Le stage est intégré à l'équipe journalière, en charge des activités opérationnelles à l'horizon J-1. Pour ces activités à fort enjeu, l'équipe génère et s'appuie sur une grande quantité de données.
L'objet du stage et de développer des outils d'analyse destinés à appuyer les équipes opérationnelles dans leur activité quotidienne. Les analyses tournent essentiellement autour des signaux de prix, qui sont une donnée clef pour l'optimisation. Parmi ces outils : un module de reprise de facturation dans des cas spécifiques ; une méthodologie pour quantifier les erreurs de prévision des prix spot de l'électricité ; une analyse de sensibilité d'un modèle de construction d'offres de flexibilité nucléaire sur le marché spot ; et une analyse du marché infra-journalier de l'électricité avec des données disponibles à la sortie de l'enchère

Within power producer EDF, the Division for Upstream-Downstream Optimisation and Trading (DOAAT) oversees the management of the demand-supply equilibrium: at all time horizons, it adjusts production and consumption and defines the production planning for EDF's assets considering technical constraints as well as economic optimality. The internship is more specifically linked to the short-term operational team in charge of the day-ahead optimisation. For their high-stakes activities, large amounts of data are generated and handled.
The aim of the internship is to develop analytical tools to support on a daily basis the operational teams, with a focus on price signals, which are a key input to power optimisation. The projects undertaken include: a tool to recalculate bills in specific cases; a methodology to quantify uncertainties linked to spot price predictions; a sensitivity analysis for a model offering power flexibility to the spot market; and an analysis of the intraday markets with the data available just after the day-ahead fixing. Those modules will be implemented according to the ongoing harmonisation of IT tools and data sources used by team.

spot. Les implémentations se feront dans le cadre de l'harmonisation des outils informatiques et des sources de données utilisées par l'équipe.

EDF
Saint-Denis

Jeudi 12 septembre *Thursday 12nd Septembre* L312

Reda MANSOURI

9h-10h

Structuration et pricing de produits dérivés et structurés actions et hybrides

Structuring and pricing of equity and hybrid derivative as well as structured products

Un produit dérivé est un instrument financier qui dépend d'un autre produit, appelé le sous-jacent. Un produit dérivé exotique, par opposition à un produit dérivé vanille, est un produit financier dont la fonction de flux terminal (payoff function) g(ST) à la date d'échéance T n'est pas une fonction simple de la valeur du sous-jacent ST. En général, les produits structurés et dérivés font intervenir d'une part différentes classes d'actifs tels que les actions, les devises, les commodités ou les taux d'intérêt et d'autre part un composant optionnel sophistiqué destiné à améliorer la performance du produit. Les produits dérivés dit hybrides mélangent plusieurs classes d'actifs, comme les actions et les devises par exemple.

Les produits structurés peuvent être des outils de diversification ou d'optimisation du taux de retour sur investissement. Mais ils peuvent également servir à couvrir des risques ou à spéculer. Les banques d'investissement vendent en général des produits structurés à deux types de clients: les institutions (compagnies d'assurance, fonds de pension) ou les banques de détail. Dans une banque proposant des produits structurés et des produits dérivés exotiques, on retrouve plusieurs métiers distincts qui interviennent au cours des différentes étapes de la vie du produit en question. Au niveau du front office, les vendeurs, les ingénieurs financiers (structureurs) et les traders ont une importance capitale dans le développement de ce business.

Ce stage se concentre sur le rôle de l'ingénieur financier, qui comprend d'une part la création de nouveaux produits structurés et d'autre part le pricing (évaluation de la valeur) de ces produits. Le rôle de l'ingénieur financier dans le pricing de produits structurés nécessite d'analyser leurs risques avant que le trade ne soit exécuté. L'ingénieur financier travaille en étroite collaboration avec les traders afin de se mettre d'accord sur la marge à inclure dans le prix du produit et qui est destinée à couvrir la prise de certain risques. Le processus de pricing comprend des simulations informatiques, des analyses de risque ainsi que le développement et la mise en place d'outils de pricing.

A derivative is a financial instrument derived from another asset, called the underlying asset. An exotic derivative, as opposed to a vanilla derivative, is a financial product whose payoff function g(ST) at maturity date T is not a simple function of the underlying value ST. Structured products and derivatives usually combine equities, currencies, commodities or interest rates with a more sophisticated optional component to boost performance. Hybrid derivatives mix several asset classes, such as equities and currencies for instance.

Structured products can serve as diversification or yield enhancement vehicles, and also as specifically tailored hedging or speculative tools. Investment banks typically sell structured products to retail clients and institutionals. In a platform of structured products and exotic derivatives, we can find distinct roles involved in the different stages of the life of a product. On the front office side, the sales people, the structurers and traders are all of central importance in the development of this business.

This internship is about the role of the structurer, which involves creating new structures as well as pricing these structures. The role of the structurer in pricing structured products involves analysing their risks before the trade can be done. The structurer will work closely with traders to agree on the levels they charge for taking on certain risks, and reflect these when making prices and considering new payoff structures. The pricing process involves simulations, risk analysis as well as the development and implementation of pricing tools.

Société Générale
Tokyo, Japon

Louis
LAPASSAT

10h-11h

Utilisation de méthodes statistiques pour développer des signaux de trading sur le marché d'actions Européen

Use of statistical methods to develop trading signals on the European equities market

Un signal de trading est le déclencheur d'une action généré par analyse, que ce soit pour l'achat ou la vente d'un actif ou d'un autre produit financier. Cette analyse peut être produite par un trader notamment en se basant sur des indicateurs techniques, ou bien en utilisant des algorithmes mathématiques fondés sur le dynamisme du marché ou d'autres facteurs alternatifs tels que les annonces, la comptabilité, etc. Le but du stage est de se concentrer sur cette dernière approche en utilisant des méthodes statistiques, extrait d'articles de recherche, pour construire des signaux. Dans un premier temps il s'agira de construire un environnement sain pour tester et évaluer le pouvoir des signaux trouvés. Ensuite on pourra progressivement, en partant d'une stratégie de base, améliorer et complexifier les modèles. Pour cela on pourra utiliser diverses méthodes de régression (lasso, ridge, regression linéaire multiple) ou de machine learning (Xgboost, ACP, ...).

A trading signal is a trigger for action generated by analysis, either to buy or sell an asset or other. That analysis can be human generated using technical indicators, or it can be generated using mathematical algorithms based on market action, possibly in combination with other market factors such as economic indicators. The aim of this internship is to focus on this last approach by using statistical methods, extracted from research papers, in order to build trading signals. In first place, it is a matter of establishing an environment to backtest
trading signals to judge performance. Then gradually starting with a basic strategy we will improve and complexify the models. For that purpose, various regression methods (lasso, ridge, multiple linear regression) or methods from machine learning (Xgboost, PCA, ...) could be used.

BNP
Paris

Yosri
SAKLY

11h-12h

Analyse de séries financières en utilisant des techniques d'apprentissage profond

Financial times series analysis using deep learning techniques

La prévision du prix des actions est la clé du succès des investissements. Ils sont très proches du hasard et cette tâche est donc très difficile. Cependant, les rendements boursiers présentent une certaine prévisibilité. Des techniques modernes d'apprentissage en profondeur et des machines, associées à des données complétées par quelques fonctionnalités intelligentes, pourraient permettre de repousser les limites d'approches plus classiques pour la prévision des rendements des actions. Durant ce stage à la Société Générale, j'ai développé des fonctionnalités avancées liées aux stratégies de trading, susceptibles d'ajouter des informations au modèle. La première partie de mon stage consiste à trouver des patterns et à développer un modèle d'apprentissage profond basé sur les CNN / LSTM afin de reconnaître les patterns importants qui nous aident à prévoir les tendances futures du cours des actions.
La deuxième partie consiste en un traitement du langage naturel permettant de tirer parti des nouvelles et de quantifier les données qualitatives à partir de nouvelles. En effet, la grande quantité d'informations textuelles et quantitatives impliquées dans les nouvelles peut aider notre modèle à anticiper les tendances haussières ou baissières. Une fois les fonctionnalités développées, je tenterai ensuite de les agréger pour prévoir les prix des actions en tenant compte des relations entre les actions. Cela se fera par le biais de techniques d'apprentissage approfondi telles que LSTMS / CNN pour apprendre les caractéristiques et essayer de prédire le cours des actions suivant dans un délai raisonnable.

Predicting stock prices is the key to doing successful investments. They are very close to random, and this task is thus very hard. However, stock returns do exhibit some predictability. Modern machine and deep learning techniques, coupled with data augmented with some clever features based on it, might be able to push the limits of more classic approaches for predicting stock returns. During this internship at Société Générale, I will develop some advanced features linked to trading strategies which may add information to the model. The first part of my internship consists in the feature engineering and developing a deep learning model based on CNNs/LSTMs to recognize important trading patterns that help us predict the future trend of the stock price.
The second part consists in natural language processing to take advantage of the news and to quantify qualitative data from news. Indeed, the vast amount of textual and quantitative information involved in news can help our model to anticipate bullish or bearish tendencies. Once the features developed, I will then aggregate them to predict the stocks prices considering the relationships between stocks. This will be done through deep learning techniques to predict the next stock prices in a convenient time scale.

Société Générale
Paris

Raphaël TEBOUL

13h30-14h30

Élaboration d'un système de recommendation

Setting up a recommendation system

Pour toutes les entreprises dont l'activité est exclusivement la publication de contenu en ligne, les campagnes de mailings constituent une chaîne d'acquisition de nouveaux clients capitale, et se révèle être des arguments de poids à présenter à de nouveaux investisseurs car cela témoigne de la santé de l'entreprise. Mais pour ce faire il est nécessaire d'avoir de bonnes statistiques d'ouverture et de clique des sus-dits mails. Il va donc s'agir à partir des données brutes de reception et de cliques d'élaborer un modèle prédictif de l'intérêt porté à la campagne par les utilisateurs, afin de pouvoir optimiser le contenu en améliorant la répartition des sujets abordés, mais également de mener une étude prospective des utilisateurs les plus fidèles, afin de fournir des recommendations sur la ligne éditoriale de l'entreprise. L'entreprise étant implantée à travers l'intégralité des Etats-unis, il va être intéressant d'employer des méthodes de géostatistique pour créer des cartes de chaleur explicitant où résident les utilisateurs les plus fidèles.
Dans un second temps, il va s'agir de créer un système de recommendation d'articles pour le site internet et pour cette campagne de mailing afin de personnaliser l'expérience utilisateur. Pour ce faire il va d'abord falloir mener une analyse ontologique de chaque article - différentes techniques de NLP seront donc au programme. Enfin, il s'agira de tester différents algorithmes de recommendation.

For compagnies whose activities are merely publishing online content, mailing campaigns are a key part of the acquisition funnel, and are a key argument to present to any would-be investor as it takes the temperature of how the audience reacts to the company. But for that to be a boon and not a bane, a company has to get good open and clicked-through rates on these emails. This is why first and foremost, the key task will be to create a predictive model of the interest of the customers for the specific emails. This can be used to optimize the distribution of the subjects we deal with, and then to build a cohort analysis of our best customers to provide recommendations on the editorial line. As the company is present across the USA, it will be interesting to use geostatistical methods to create heat maps of what works where.
Finally, the aim is to create a recommendation system for the website and the mailing campaign to personalize the user experience. To do so, it has to start with an ontological analysis of the articles - several NLP methods are going to be tested. In the end several recommendation algorithms will be tested.

The Plunge
New-York, États-Unis

Abderaheman YEWGAT

14h30-15h30

Apprentissage statistique, analyse fonctionnelle, géostatistique et analyse de sensibilité dans le contexte de la production pétrolière

Statistical learning, functional data analysis, Geostatistics and Sensitivity analysis for oil and gas forecasting

Le but de mon travail est de développer des modèles d'apprentissage statistique pour prédire la production d'huile et de gaz. L'idée de mon travail est de partir des modèles géostatistiques classiques comme le krigeage universel, le co-krigeage universel ou la simulation gaussienne et aussi des modèles de machine learning comme les forêts aléatoires, et essayer de construire des modèles de prédiction pour l'huile et le gaz. Ces modèles sont généralement appliqués dans le cas où l'on a des sorties scalaires, donc pour adapter ces derniers au cas des sorties fonctionnelles on a utilisé la théorie des données fonctionnelles qui nous donne un cadre mathématique dans lequel on peut développer nos modèles d'apprentissage. La deuxième partie de mon stage est plus sur la partie analyse de sensibilité vu que dans le cas de données d'huile et de gaz. On a souvent un ensemble de paramètres qui doivent être analysés avant d'entrainer les précédents modèles.

The aim of my work is to develop statistical learning models to forecast the production of oil and gas. The idea is to start from a classical geostatistical framework like Universal Kriging, Universal Co-Kriging or Sequential Gaussian Simulation and from machine learning models like Random Forests and to try to construct learning models capable of forecasting in the case of oil and gas production. These models are generally applied for scalar outputs ; in the case of oil and gas production we are working with curves (production over years for example) rather than scalar outputs, so to adapt these models to our context we have used the functional data analysis approach which gives us a mathematical framework in which we can apply the previous approach. The second part of my work is about sensitivity analysis, because in the oil and gas context we have several parameters and we need to perform a deep analysis of these parameters before building any learning models.

Total Exploration and Production *Pau*

Juliette
ORTHOLAND

Etudes des séries temporelles et de l'anonymisation des données pour l'industrie

Study of time series and data de-identifying for industry

De nos jours l'industrie possède un stock important de données qu'elle cherche à valoriser. Un axe de valorisation est l'étude de problématiques reliées à l'analyse des séries temporelles. En effet, l'étude des anomalies et des saisonnalités des séries temporelles peut avoir des applications en maintenance ou en efficacité énergétique. D'autre part, suite à l'entrée en vigueur du RGPD, des problématiques entourant les données personnelles et leur anonymisation doivent être prises en compte. L'objectif de mon stage et d'accompagner les acteurs industriels dans la valorisation de leurs données à travers des démonstrations de faisabilité reliées à ces axes d'études. Mes travaux porteront sur l'aspect algorithmique des traitements ainsi que leurs implémentations dans les différents outils de l'écosystème technologique.

Nowadays the industry seeks to value the important amount of data it owns. Studying time series is a way to value industrial data. Indeed, the analysis of time series anomalies and seasonality can have application on maintenance and energy efficiency. On the other hand, with the entry into force of GDPR, questions around personal data and their de-identifying should be taken into account. The aim of my internship is to help industrial stakeholders to value their data through proof of concept that are linked to this axis of study. My work will be related to the data treatment and the implementation of tools in the technology environment.

EDF
Nanterre

Victor
BIAGGI

Optimisation d'algorithmes de trading

Optimization of trading algorithms

La stage consiste en l'optimisation d'algorithmes de type « machine learning » de trading. Le but du stage est de comprendre comment se fait le trading de marché automatique et de mettre en pratique le savoir acqui en statistiques et machine learning. Ce stage permet aussi de comprendre le fonctionnement d'un « order book » de trading et de savoir comment sont traités les différents produits sur indice equity du desk « Automated Market Making » de la banque BNP.

This internship at the bank BNP in Hong Kong consists in optimizing trading algorithms involving machine learning techniques. The aim of this intership is to understand how automated trading is performed. It allows me to put in practice my knowledge in statistics and machine learning. This intership is also a good opportunity to understand how an « order book » works and learn about market microstructure. It helps understanding how products on Equity Index are traded on the trading desk « Automated Market Making ».

BNP
Hong Kong, Chine