

## MODULES GSON 2018-2019

### LES TROIS MODULES DU PREMIER SEMESTRE (Mardi 16h00-19h00 sur 8 séances)

#### Modules TECH : Apprendre des méthodes, des techniques et des logiciels (non spécifiques à un domaine d'application)

Nom	Enseignant Référent	Descriptif à destination des étudiants	Pré-requis
<b>Introduction à Python pour le calcul scientifique</b>	Mathieu Liedloff Anthoy Perez Emmanuel Schang Anaïs Lefeuvre-Halftermeyer & Konrad Hinsen	Le langage Python est utilisé et apprécié pour l'écriture de simples scripts à des logiciels plus conséquents. Nombreuses sont les entreprises ou organismes de recherche à s'en servir et à recruter des profils ayant l'expérience de ce langage. Intuitif et simple d'apprentissage, nous vous donnerons toutes les clés pour débiter avec Python. Nous découvrirons également quelques bibliothèques spécialisées du langage, autour du calcul scientifique et du traitement des langues. A la fin de ce module, vous serez autonome, vous pourrez développer vos premiers programmes et approfondir les spécificités du langage et de ses nombreuses bibliothèques.	Sans pré-requis
<b>Introduction au logiciel SAS</b>	Sébastien Ringuedé	Le cours est une initiation à la manipulation des données et leur traitement statistique et à la rédaction des rapports à l'aide du logiciel SAS®. Dans sa première partie, ce cours a pour but d'expliquer le fonctionnement de l'étape DATA, comment organiser les données en entrée et en sortie, travailler avec différents types de données, lire des données sous différents formats, manipuler et modifier des fichiers, créer des rapports. Au cours de la seconde partie, certaines procédures fondamentales pour la manipulation et l'analyse descriptive des données sont présentées : PROC FREQ, PROC TABULATE, PROC MEANS...	Aucun prérequis.

#### Module THEM : Utilisation des Data sciences dans différents domaines des Masters participant à GSON

Nom	Enseignant Référent	Descriptif à destination des étudiants	Pré-requis
<b>Analyse de données par des cas pratiques</b>	Thierry Dudok de Wit	Ce cours offre un panorama de différentes méthodes statistiques couramment utilisées dans l'analyse de données. Il est résolument axé sur la mise en pratique et non sur la théorie ou sur la programmation. En partant d'exemples concrets issus de diverses disciplines nous verrons quels sont les outils les mieux adaptés pour extraire des informations, comment les mettre en pratique et quels sont les erreurs les plus courantes. Au programme : incertitudes, leur caractérisation et leur propagation, tests d'hypothèse, ajustement de fonctions, classification. La mise en pratique se fera à l'aide du logiciel Scilab.	Pratique d'un langage de programmation (Python, Matlab, C, ...) et notions de statistiques de niveau licence.

**LES MODULES DE LA SEMAINE BLOQUÉE DU 14 JANVIER**

**Modules TECH : Apprendre des méthodes, des techniques et des logiciels (non spécifiques à un domaine d'application)**

Nom	Enseignant Référent	Descriptif à destination des étudiants	Pré-requis
<b>Algorithmes pour la résolution de problèmes</b>	Mathieu Liedloff & Anthoy Perez	<p>Les algorithmes sont au coeur de la résolution de problèmes complexes.  Même si les ordinateurs sont de plus en plus rapides, certains problèmes ont besoin d'idées algorithmiques astucieuses pour être résolus rapidement.  Dans ce module, nous commencerons par étudier quelques approches algorithmiques élémentaires mais efficaces, avant d'étudier des algorithmes bien plus ingénieux et peut-être même surprenants !  Nous aborderons également les possibilités de modélisation d'un problème par des outils tels que les graphes et discuterons d'approches algorithmiques exactes, approchées ou probabilistes.  Le contenu de ce cours est susceptible de s'adapter au public.</p>	Connaître un langage impératif (C, C++, Java, Python ou autre) est nécessaire
<b>Programmation haute performance</b>	Sophie Robert & Sébastien Limet	<p>Comment paralléliser un problème pour une mise en œuvre sur des architectures à mémoire distribuée (grappe de PCs, machines massivement parallèles) ou des machines à mémoire partagée (machine multicœurs) ?  Introduction à la programmation avec la librairie MPI et l'API OpenMP.  Des TP seront proposés pour mettre en pratique quelques exemples mais un accent fort sera mis sur les techniques de parallélisation qui peuvent être appliquées dans différents types de calculs ou de traitement de données (parallélisme de données, calculs stencils, ...) quelque soit le domaine dont est issu le problème (physique, économie, biologie, traitement du langage, ...).  En fonction des étudiants inscrits à ce module, une introduction à la programmation en C pourra être proposée.</p>	Savoir programmer dans un langage impératif (C, Fortran, ...)
<b>Data Mining avec le logiciel R</b>	Didier Chauveau & Laurent Delsol	<p>Introduction au langage R et aux outils de statistiques descriptives ("data exploratory analysis") avec R.  1. Méthodes de data mining multivariées pour données numériques ou qualitatives: Analyse en composantes principales (ACP), Analyse factorielle des correspondances (AFC), Analyse des Correspondances multiples (ACM) .  2. Méthodes de classification (clustering) non supervisées (kmeans, classification hiérarchique, à base de modèles probabilistes de mélange)et supervisées (régression logistique,...).  NB: Ce module s'intitule "Data Mining" (fouille de données) et pas "Big Data" car ces méthodes sont adaptées à des données éventuellement très volumineuses, mais qui "tiennent" dans un ordinateur courant, sans nécessiter de système ni de calcul distribué type Hadoop/MapReduce.  Cependant, si le temps le permet, une introduction au manipulation des BigData avec R (package RHadoop) sera proposée.</p>	<p>Sans prérequis.   (mais une connaissance de base du logiciel R et des statistiques élémentaires est préférable).</p>

<p><b>Data Mining : Fondements et Outils Python</b></p>	<p>Christel Vrain</p>	<p>* Introduction au Data Mining (types de données - tâches)  * Classification supervisée : arbre de décision, classifieur bayésien, k-plus-proche-voisin, réseau de neurones, SVM, noyaux  * Classification non supervisée : k-moyenne, hiérarchique, clustering spectral, méthodes fondées sur la densité  * Quelques notions sur la recherche de règles d'association et de motifs fréquents.  Utilisation d'un environnement de Data Mining développé en Python (Orange) et de bibliothèques Python comme Scikit-learn.  Importance du pré-traitement des données et de la validation des modèles.  <b>Important</b> : il s'agit d'un module, principalement Machine Learning et Data Mining et non d'un module de développement Python.</p>	<p>Sans prérequis.</p> <p>Une connaissance de Python peut être un plus mais nous utiliserons principalement des packages Python et des fonctions prédéfinies de ces packages.</p>
<p><b>Expérimentations numériques</b></p>	<p>Jean-Louis Rouet  Carine Lucas,  &amp; Magali Ribot-Barre</p>	<p>La simulation numérique s'avère être un véritable outil d'expérimentation dans bien des domaines.  Ce module aborde deux méthodes, les méthodes déterministes et celles de Monte Carlo.  Dans le premier cas, on programmera des méthodes efficaces qui trouvent des applications, par exemple, en dynamique des populations et en physique pour  - interpoler des mesures, calculer une intégrale,  - visualiser la solution d'une équation différentielle,...  Les méthodes stochastiques seront illustrées et mises en œuvre pour le dépôt de couches minces, la théorie des votes, la biologie ou encore l'écoulement des fluides, ...  Le module laisse une large place à la programmation à l'aide du logiciel libre Scilab, référence pour le calcul scientifique (syntaxe similaire à celle de Matlab).  La première séance permettra de prendre en main Scilab.</p>	<p>Goût pour la programmation, connaissances en algorithmique et programmation.</p>
<p><b>Big Data avec SAS</b></p>	<p>Sullivan Hué</p>	<p>Ce cours a pour objet l'étude d'un modèle d'apprentissage supervisé connu sous le nom d'arbres de décision.  Les méthodes d'agrégation des arbres de décisions, telles que les forêts aléatoires ou Random Forest (Breiman, 2001) et les méthodes de Boosting (Freund et Schapire, 1996) seront également abordées.  Ces méthodes sont très utiles pour la prédiction dans le contexte des données massives.  Le logiciel d'application est SAS.</p>	<p>Connaissance des méthodes de régression (MCO) et de classification (logistique).  Pratique d'un langage de programmation (R, Matlab, Stata, ...)</p>
<p><b>Big Data avec Hadoop</b></p>	<p>Mostafa Bamha</p>	<p>Le but de ce module, est de permettre aux étudiants d'acquérir des connaissances à la fois théoriques et pratiques dans la gestion, le stockage et la manipulation de grandes masses de données en utilisant le modèle MapReduce, le système de fichiers distribués en passant par l'utilisation en pratique de différents frameworks Big Data tel que Hadoop, Hbase, Hive, PigLatin et Giraph permettant de manipuler de très grandes masses de données sur des systèmes à grande échelle.</p>	<p>Connaissance de Java et Linux.</p>

**Modules THEM : Utilisation des Data sciences dans différents domaines des Masters participant à GSON**

Nom	Enseignant Référent	Descriptif à destination des étudiants	Pré-requis
<b>Biologie : qu'analyse-t-on?</b>	Lucile Mollet Intervenants : Lucile Mollet (1 cours) Fabienne Brulé (2 cours) Thierry Normand (2 cours) Alain Legrand (2 cours) Kévin Moreau (1 cours)	Descriptif : Nature, fonction et expression du génome. L'ADN, un code à 4 lettres formant des signatures. Séquençage de l'ADN, assemblage des morceaux et bar-coding. L'ARN, ne jamais travailler avec l'original. Traduire un ARN en protéine, règle et prédictions. La protéine, un code à 20 lettres formant des cassettes modulables. Séquence protéique : prédictions de structures et fonctions. Analyses de n séquences : alignements et motifs consensus. Phylogénie : évolution moléculaire, un arbre de distance.	Aucun prérequis.
<b>Data sciences et langage</b>	Anne-Lyse Minard Anaïs Lefeuvre-Halftermeyer & Emmanuel Schang	Dans ce module nous traiterons des problématiques associées au traitement automatique des langues. Nous aborderons entre autres les questions liées au traitement des données issues du domaine médical et issues des réseaux sociaux. Une séance sera également consacrée à l'étude de phénomènes linguistiques complexes à analyser comme les coréférences.	Aucun prérequis.
<b>Du CRM au BigData</b>	Xavier Driane & Pascal Maloti	Si vous considérez le client comme un véritable « actif » de l'entreprise, vous êtes prêt à élaborer une stratégie orientée client. Cette session vous apporte tous les éléments pour comprendre l'intérêt du CRM et mener à bien votre relation client.	PC
<b>Chimie informatique sous python</b>	Colin Bournez & Gautier Peyrat	La Chemoinformatique est un nouveau domaine qui utilise les outils informatiques pour comprendre la chimie ou prédire des structures chimiques. La chemoinformatique nécessite un encodage des structures chimiques sous formes de données numériques. Ce cours utilisant les notebook Python se décompose en plusieurs chapitres dont : une introduction à Python notebook, une introduction aux outils de chemoinformatique appliqués à des bases de données publiques et à l'utilisation d'outils d'analyse et de visualisation de données chimiques (clustering, PCA). Ce cours apporte aux étudiants une ouverture d'esprit sur l'utilisation de l'informatique appliqué à des problématiques de chimie rencontrées en cosmétique, en chimie des médicaments ou des matériaux.	Sans prérequis.  (mais une connaissance de base du langage de programmation Python est préférable).
<b>Modéliser des flux avec Comsol Multiphysics</b>	Yannick Branquet	Le module vise à initier l'étudiant à la physique et la modélisation des flux de chaleur (par exemple: géothermie) et/ou de matière (par exemple: polluants dans une nappe d'eau) dans un milieu continu ou poreux. Après une présentation en cours des flux et de leur mise en équation, les étudiants choisiront un problème qui les intéressent et devront apporter des solutions ou faire des tests via l'outil de résolution numérique Comsol Multiphysics. La méthode de résolution numérique des Équations aux Dérivées Partielles par Éléments Finis sera présentée brièvement. La seconde partie du module constitue un mini-projet encadré en salle. Ce mini-projet fait donc initialement l'objet d'une discussion et d'un dimensionnement avec l'enseignant afin de bien cerner les objectifs de l'expérience numérique qui sera tentée. Des allers/retours systématiques seront réalisés entre résultats des simulations numériques, physique des flux et équations les décrivant, afin d'éviter le piège de "boite noire" que constituent ces logiciels de modélisation.	* Le module est organisé de sorte <b>qu'aucun pré-requis fondamental ne soit nécessaire.</b> * Des bases de physique des transports (thermiques et/ou moléculaires, diffusion vs. advection, conservation de l'énergie et de la masse) ainsi que des phénomènes géologiques simples (e.g. magmas s'intrudant dans une roche encaissante, pollution dans une nappe phréatique) constituent un plus pour la réalisation du mini-projet.

<b>Imagerie et interpolation des structures géométriques 3D</b>	Gautier Laurent	<p>Ce module permettra aux étudiants d'explorer la problématique de la reconstruction d'objets tridimensionnels (3D) à partir de données géométriques partielles.</p> <p>Ce processus vise à produire une image de la géométrie possible d'un objet en complétant les observations disponibles, notamment par des approches d'interpolation et d'extrapolation.</p> <p>Très présente en infographie, cette problématique prend une tournure particulière en géoscience et dans l'exploration spatiale car la géométrie complexe de certains objets peut s'avérer un véritable défi méthodologique lorsque l'information disponible est incomplète ou mal répartie spatialement. Il est parfois difficile voire impossible de contrôler l'échantillonnage et ce sont alors les outils d'interpolations qui doivent s'adapter aux données.</p> <p>Ce cours donnera un aperçu des méthodes numériques permettant d'aborder ces problèmes d'imagerie et d'interpolation.</p>	Sans prérequis.
<b>Analyse Spatiale Prédictive</b>	Charles Gumiaux	<p>Ce cours vise à comprendre et savoir utiliser les méthodes et outils d'analyse spatiale.</p> <p>Ces approches ont été développées en particulier pour l'analyse des objets cartographiques mais elles peuvent être appliquées à l'étude de la distribution de toute entité projetée dans un espace donné.</p> <p>Une analyse visuelle peut permettre sous certaines conditions de qualifier une organisation particulière d'objets en carte (dispersion, regroupement, alignement,...) ou de faire ressortir des liens de proximité entre des objets de différentes natures.</p> <p>Au delà de l'analyse purement visuelle, des traitements statistiques/numériques particuliers permettent de quantifier ces schémas d'organisation et de relation entre objets.</p> <p>Contenu :</p> <ul style="list-style-type: none"> <li>- projection des données en carte et gestion des systèmes de projection sous SIG ;</li> <li>- caractérisation statistique de la répartition d'objets ponctuels ;</li> <li>- quantification des liens spatiaux pouvant exister entre les entités cartographiques (points,lignes,polygones) ;</li> <li>- traitement et interpolation des variables régionalisées (introduction aux géostatistiques).</li> </ul> <p>Ces différents outils peuvent être appliqués notamment au calcul de cartes prédictives (risques ou favorabilité) basées sur des modèles statistiques.</p>	Trigonométrie; statistiques; notions de cartographie.
<b>SIG Raster et 3D : initiation et modélisation environnementale</b>	Rachid Nedjai	<p>Il s'agit d'introduire auprès des étudiants la dimension Raster et son utilité en environnement.</p> <p>Cette approche sera poursuivie par une initiation à la 3ème dimension avec des exemples de modélisation des cours d'eau, leur classification et calcul des principaux paramètres hydrologiques associés.</p>	Avoir des connaissances en SIG
<b>Droit de l'informatique</b>	Marie Luce Mariani	<p>Le droit de l'informatique s'étend désormais au droit de la donnée c'est-à-dire à l'ensemble des informations qui circulent sur le réseau. Il s'agit, notamment des données personnelles qui font l'objet d'une réglementation européenne spécifique.</p> <p>Le cours, à partir des concepts juridiques de base comme ceux qui régissent le contrat, présente le cadre juridiques des nouvelles technologies et de leur usage.</p> <p>Un accent particulier est mis sur le droit d'auteur, le droit de l'internet et fait une brève présentation des problématiques liées à l'intelligence artificielle.</p>	<p>* Intérêt pour l'actualité</p> <p>* Aisance de compréhension de textes écrits</p>