

Les Modules

Descriptif et modalité de l'enseignement

Des modules TECH et THEM – Semestre 2

TECH les fondamentaux maths/info	THEM Les domaines applicatifs
	<i>Biologie : qu'analyse-t-on ?</i>
	<i>Chimie informatique sous Python</i>
<i>Data Mining : Fondements et Outils Python</i>	
<i>Data Mining avec le logiciel R</i>	<i>Analyse Spatiale Prédictive</i>
<i>Big Data avec SAS</i>	<i>Modéliser des flux avec Comsol Multiphysics</i>
<i>Big Data avec Hadoop</i>	<i>SIG Raster et 3D : initiation et modélisation environnementale</i>
<i>Méthodes et expérimentations numériques</i>	
<i>Programmation Haute Performance</i>	<i>Méthodologie de l'économétrie</i>
<i>Droit de l'informatique</i>	

Tous les modules durent 20h.

Introduction à Python et Exemples d'application



Le langage Python est utilisé et apprécié pour l'écriture de simples scripts à des logiciels complets. Nombreuses sont les entreprises ou organismes de recherche à l'utiliser 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 ensuite quelques bibliothèques spécialisées du langage, autour du calcul numérique 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 bibliothèques.

Introduction à Python et Exemples d'application

Intervenants	Mathieu Liedloff, Anthony Perez, Carine Lucas, Emmanuel Schang
Pré requis	Aucun
Période d'enseignement	mardi fin d'après midi
Forme	TP, apprentissage par l'exemple
Évaluation	Exercices de Programmation

Data Mining avec le logiciel R

Contenu

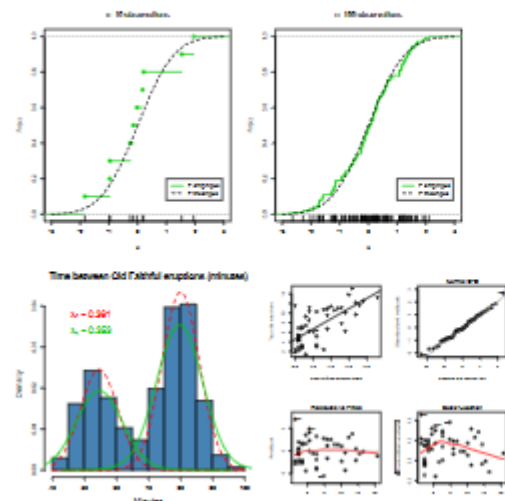
- 1 Introduction au langage  et aux outils élémentaires (*Exploratory Data Analysis*)
- 2 Méthodes de Data Mining pour données numériques ou qualitatives **volumineuses** :
 - Techniques multivariées usuelles fondées sur l'Analyse en Composantes Principales : ACP, AFC, ACM
 - Méthodes de classification (*clustering*) non supervisées et supervisées
- 3 si le temps (et le public) le permet, introduction à la manipulation des *BigData* avec  (package RHadoop)

Prérequis : Notions de statistique et d'informatique élémentaire de type Terminale scientifique.

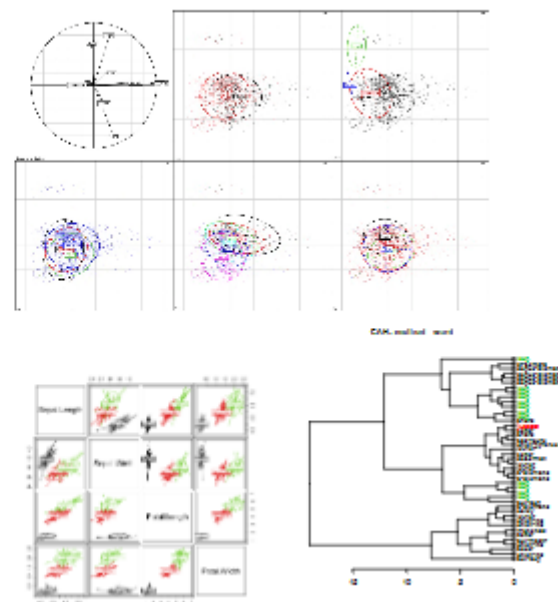
- Gratuit, Open Source, multi-plateforme
- Développé par les meilleurs experts en *Statistical Computing* : **The R Core Team**
- Langage interactif, orienté objet, extensible
- **Pensé pour l'exploration et la modélisation de données**
- L'environnement idéal pour populariser/utiliser les méthodes usuelles/nouvelles en statistique
- Plus de 4000 **packages** associés :
sites **CRAN** = **C**omprehensive **R** **A**rchive **N**etwork
- Outils de **calcul parallèle et/ou distribué** intégrés
(HPC, BigData)

<https://www.r-project.org>

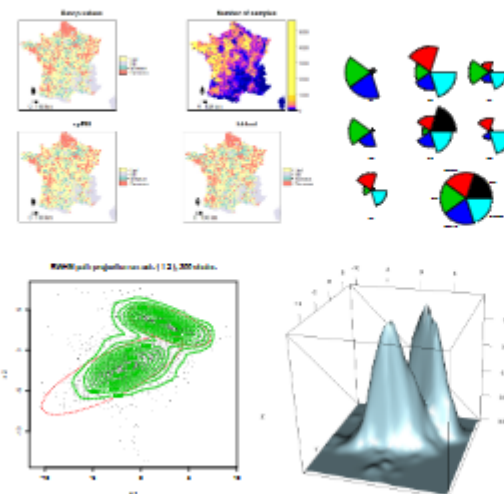
Modélisation, Estimation, Prédiction



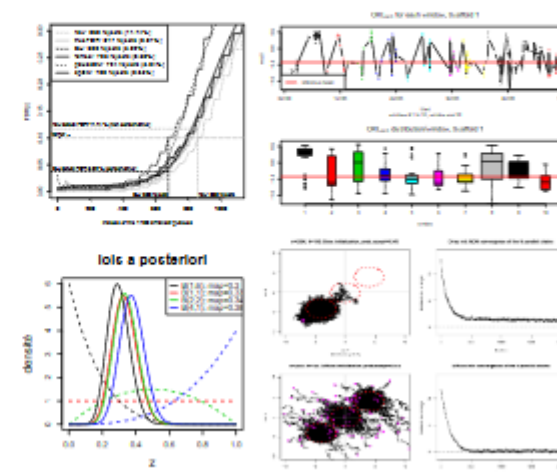
Analyses multivariées, ACP, Clustering...



Visualisation, Descriptif, Géostatistiques



Résumés graphiques d'analyses statistiques complexes





GUI → utilisée en TP du module GSON


<https://www.rstudio.com>



Interface web dynamique de visualisation

<http://shiny.rstudio.com/gallery/>



Markdown : générateur de doc. incluant $\text{L}^{\text{A}}\text{T}_{\text{E}}\text{X}$ et code 

<https://rmarkdown.rstudio.com>



: package **RHadoop** pour le *BigData*

<https://github.com/RevolutionAnalytics/RHadoop>

Nouvelles technologies sous R

Intervenants	Didier Chauveau - Laurent Delsol
Pré requis	Connaissances de base sur l'utilisation de R et la programmation dans ce langage.
Période d'enseignement	Mardi fin d'après midi
Forme	Cours et TD
Évaluation	

Data Mining : Fondements et Outils Python

Dans ce module nous introduisons les concepts de base en Machine Learning et en Data Mining et nous dressons un panorama des principales méthodes en classification supervisée et non supervisée. Nous insistons sur la nécessité de pré-traiter les données et sur la validation des modèles appris.

Nous mettons en pratique d'une part par l'utilisation d'un outil permettant de mettre en œuvre facilement une chaîne complète de Data Mining (prétraitement des données, paramétrage de l'outil, validation du modèle) et d'autre part par l'utilisation d'une librairie Python permettant d'intégrer facilement un processus de Data Mining dans un projet Python.

Il est à noter qu'il s'agit d'un module orienté Machine Learning et Data Mining et en aucun cas d'un module d'approfondissement du langage Python (uniquement utilisation de librairies Python)

Data Mining : Fondements et Outils Python

- 1- Introduction au Data Mining (types de données – tâches). Importance du pré-traitement des données et de la validation des modèles,
- 2- Classification supervisée : arbre de décision, classifieur bayésien, k-plus-proche-voisins, réseau de neurones, SVM, noyaux,
- 3- Classification non supervisée : k-moyenne, hiérarchique, clustering spectral, méthodes fondées sur la densité,
- 4- Quelques notions sur la recherche de règles d'association et de motifs fréquents,
- 5- Utilisation d'un environnement de Data Mining développé en Python (Orange: <https://orange.biolab.si/>) et de bibliothèques Python (Scikit-learn: <https://scikit-learn.org/stable/>),

Data Mining : Fondements et Outils Python

Intervenante	Christel VRAIN
Pré requis	Sans pré-requis. Une connaissance de Python peut être un plus mais nous utiliserons principalement des packages Python et des fonctions prédéfinies de ces packages.
Période d'enseignement	Semaine banalisée
Forme	Cours et TP
Évaluation	Mise on œuvre d'un processus de Data Mining sur une base de données, au choix des étudiants. Par groupes de 2 ou 3.

A regarder avant	
------------------	--

Introduction au logiciel SAS

SAS est le progiciel d'informatique décisionnelle le plus utilisé au monde. Il propose l'ensemble des outils de la chaîne analytique allant de la constitution des bases de données à la restitution en passant par la statistique, l'économétrie, la recherche opérationnelle mais aussi des approches plus récentes comme le machine learning, le deep-learning etc.

L'objet du cours sera plus particulièrement de présenter l'architecture SAS et les outils proposés par SAS afin de créer des tables SAS et les modifier. Nous présenterons aussi quelques procédures de primo-exploitation des données.

Ce cours n'est pas orienté "statistique". Il est centré sur ce qui occupe 80% du temps d'un data scientist : la donnée.

Introduction au logiciel SAS

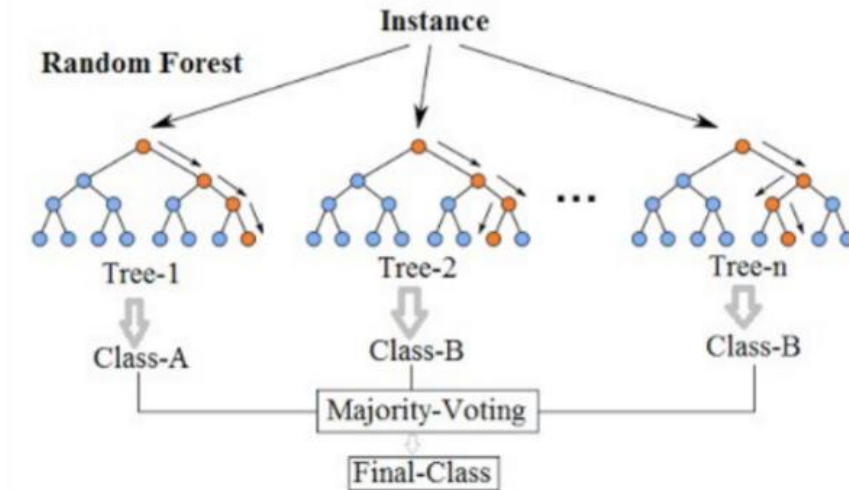
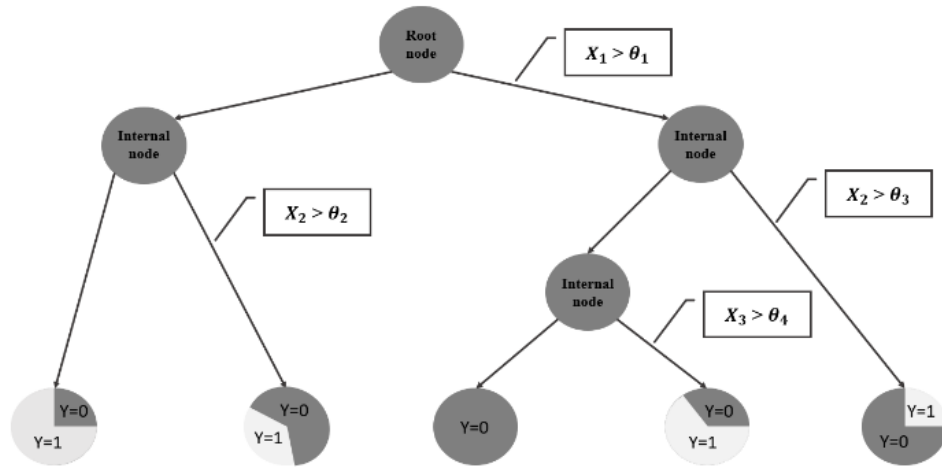
Intervenants	Sébastien Ringuedé
Pré requis	aucun
Période d'enseignement	mardi fin d'après midi – semestre 1
Forme	Salle informatique
Évaluation	Devoir de mise en application des connaissances acquises durant le cours (2 heures)

A regarder avant

Des pointeurs utiles ...

Big Data avec SAS

- 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 (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 : octroi de crédits, renouvellement de campagne publicitaire, etc ...
- Le logiciel d'application considéré dans ce cours est SAS. Un rappel d'utilisation de celui-ci sera effectué durant la semaine de cours.

Big Data avec SAS

Intervenant	Sullivan HUE
Pré requis	<ul style="list-style-type: none">- Connaissance des méthodes de régression (MCO) et de classification (logistique)- Pratique d'un langage de programmation (R, Matlab, Stata, ...)
Période d'enseignement	Semaine banalisée de Janvier
Forme	Cours
Évaluation	Projet à rendre

Big Data avec Hadoop

- **Comment manipuler et traiter de grande masses de données ?**
- Ce module présente des outils et méthodes de traitement de gros volumes de données (Big Data) en utilisant Hadoop et d'autres frameworks Big Data adaptés. Chaque framework étant bien adapté pour un usage particulier. Hadoop est un ensemble de services et d'applications permettant de stocker et d'administrer de très grandes masses de données. Hadoop est utilisé par des entreprises comme Google, Facebook, Amazon, Ebay, ...
- L'objectif 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 (HDFS) en passant par l'utilisation en pratique (pendant les séances de TPs) de différents frameworks Big Data tel que Hadoop, Hbase, Hive, PigLatin et Giraph.

Big Data avec Hadoop

Intervenant	Mostafa BAMHA
Pré requis	Connaissance de Java et Unix
Période d'enseignement	semaine banalisée de Janvier
Forme	Cours - TP (entièrement en salle machine)
Évaluation	Exercices pendant la semaine plus exercices à rendre sous Celene 15 jours plus tard.

Méthodes et expérimentations numériques

La simulation numérique s'avère être un véritable outil d'expérimentation dans bien des domaines. Ce module aborde deux types de 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, les choix sociaux (les votes), la diffusion des neutrons, ...

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 et proche de la bibliothèque Numpy de Python). La première séance permettra de prendre en main Scilab.

Méthodes et expérimentations numériques

Intervenants	C. Lucas, M. Ribot, J.-L. Rouet
Pré requis	Aucun, sinon le goût pour la programmation
Période d'enseignement	semaine banalisée
Forme	Cours/TP
Évaluation	Rapport écrit sur un cas à traiter

A regarder avant

Des pointeurs utiles ...

Programmation Haute Performance

Comment paralléliser un problème pour une mise en œuvre sur des architectures haute performance ?

- **Le principe du parallélisme est simple.**
 - Il s'agit d'exécuter en même temps le maximum d'instructions indépendantes d'un code de calculs.
- **La mise en œuvre nécessite de connaître**
 - Les architectures des machines parallèles
 - Les techniques de parallélisation
 - Les techniques de programmation
- **Ce module GSON a pour objectifs**
 - A partir de nombreux exemples d'introduire les différentes techniques de parallélisation indépendamment de tout langage de programmation.
 - De mettre en œuvre la parallélisation de codes calculs en utilisant Python et `mpi4py`. 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, ...).

Programmation Haute Performance

Intervenants	Sébastien LIMET et Sophie ROBERT
Pré requis	Python
Période d'enseignement	semaine banalisée de janvier
Forme	Cours - TP (entièrement en salle machine)
Évaluation	Exercices pendant la semaine plus exercices à rendre sous Celene 15 jours plus tard.

Biologie : qu'analyse-t-on ?

 Nouveauté GSON 2019/20

 Introduction - Fonction du génome au sein de la cellule, biologie cellulaire

 Nature du génome ADN, un alphabet à 4 lettres - Séquençage de l'ADN

 L'ARN, un intermédiaire pour ne jamais travailler avec l'original

 La protéine, un alphabet à 20 lettres, formant des cassettes modulables

 Evolution moléculaire et Phylogénie

 Autres exemples de domaines d'applications, en dehors du génome *

 Une demi-journée de TP

* Thème suggéré par les étudiants de GSON 2018/2019

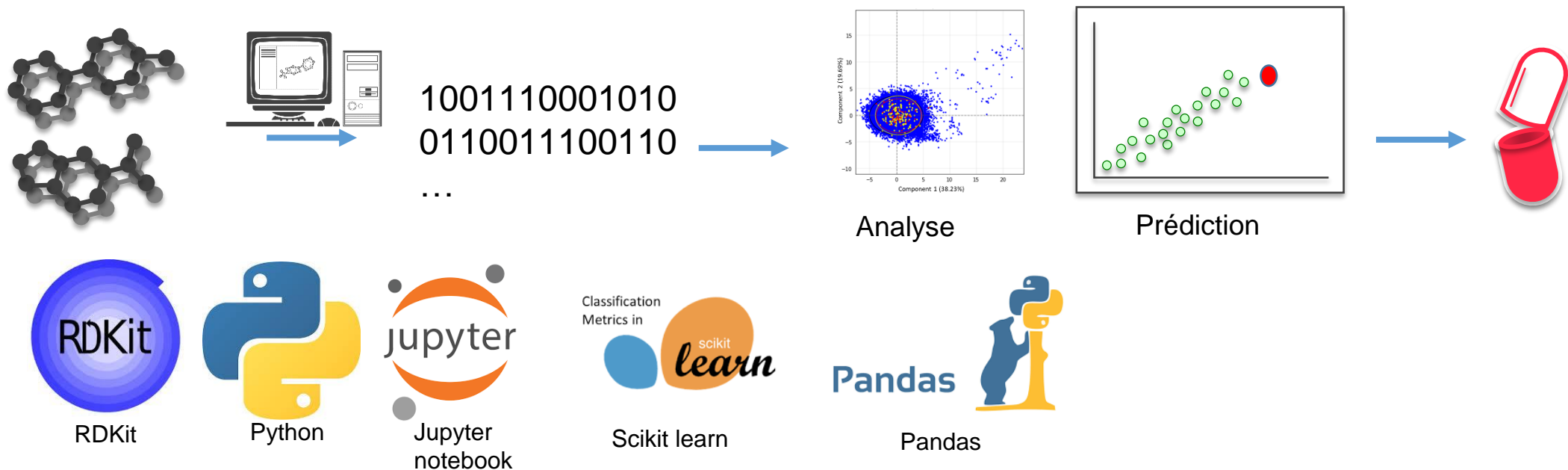
Lucile Mollet, Alain Legrand, Thierry Normand, Fabienne Brulé

Biologie : qu'analyse-t-on ?

Intervenants	Lucile Mollet, Alain Legrand, Thierry Normand, Fabienne Brulé
Pré requis	Aucun
Période d'enseignement	Du lundi au vendredi, lors de la semaine banalisée (janvier)
Forme	14 h de cours/TD : Lundi, mardi, mercredi : 2h matin, 2h après-midi + Jeudi 2 h matin 4 h de TP : Jeudi après-midi 2h questions ouvertes, évaluation de l'enseignement/suggestions par les étudiants, évaluation des étudiants : Vendredi après-midi
Évaluation	30 min, 8 à 10 petites questions balayant l'ensemble du cours

Chimie informatique sous Python

La Chémoinformatique est un domaine qui utilise les outils informatiques pour comprendre la chimie ou prédire des structures chimiques. La chémoinformatique nécessite un encodage des structures chimiques sous forme de données numériques. Ce cours utilisant les notebooks Jupyter en langage Python se décompose en plusieurs chapitres dont : une introduction à Python et aux notebooks Jupyter, une introduction aux outils de chémoinformatique appliqués à des bases de données publiques et à l'utilisation d'outils d'analyse et de visualisation de données chimiques (clustering, ACP). Ce cours apporte aux étudiants une ouverture d'esprit sur l'utilisation de l'informatique appliquée à des problématiques de chimie rencontrées en cosmétique, en chimie des médicaments ou des matériaux.



Chimie informatique sous Python

Intervenants	Doctorant, Chercheur post-doctorant, chercheur CNRS et/ou professeur des universités
Pré requis	Une connaissance dans un langage de programmation (Python serait un plus)
Période d'enseignement	Semaine banalisée ou le mardi fin d'après midi
Forme	Cours/TP
Évaluation	Application des outils sur un exemple similaire à celui vu en cours. Evaluation par QCM sur les notions abordées lors de la formation

A regarder avant

Data sciences et langage

Dans ce module nous étudierons les problématiques liées au traitement automatique des données langagières. Tout le long du semestre nous travaillerons sur le corpus du « Grand Débat National », corpus qui contient environ 1,9 million de contributions. Nous étudierons des phénomènes linguistiques complexes qui sont à prendre en compte lors de l'analyse automatique de textes. Nous discuterons également des méthodes utilisées en Traitement Automatique des Langues pour « essayer » de traiter cette complexité. Un projet en petit groupe (interdisciplinaire) sera réalisé tout le long du module sur les données du grand débat.

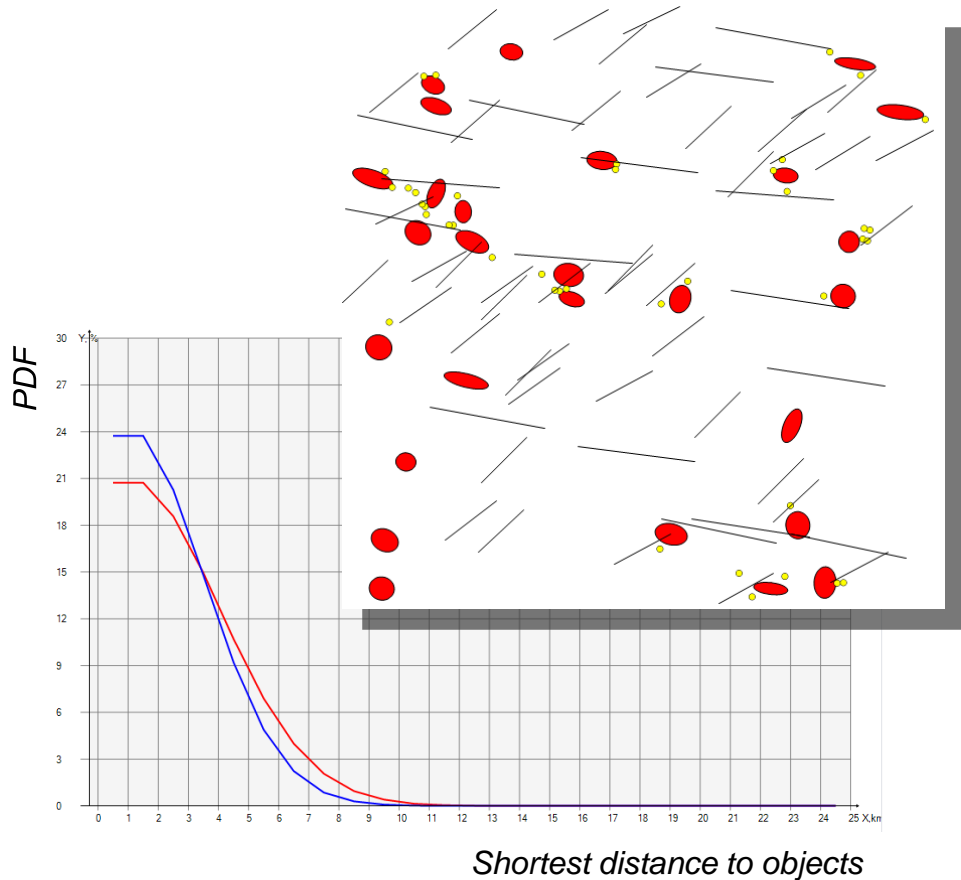
Data sciences et langage

Intervenants	Anne-Lyse Minard, Flora Badin, Caroline Cance, Katja Ploog, Emmanuel Schang
Pré requis	Aucun
Période d'enseignement	le mardi 16h-18h30
Forme	Cours - TD - TP
Évaluation	Projet à rendre en janvier

A regarder avant	
------------------	--

Analyse Spatiale Statistique / Prédictive

L'**objectif** de ce module est d'apprendre à **caractériser** et **quantifier** (i) la distribution spatiale d'objets dans un espace géométrique et (ii) les relations de recouvrement, de proximité entre des objets de nature différente.



Ce type d'approche s'applique en particulier à l'analyse des objets cartographiques mais peut être appliqué en 3D ou à d'autres espaces de projection.

La prédictivité consiste à produire une représentation graphique (carte le plus classiquement) montrant un degré de **favorabilité** (ou d'**aléa**) à ce qu'une occurrence d'un phénomène puisse être trouvée.

Au programme :

Les outils d'analyse spatiale

- Analyse de dispersion
- Analyse de dispersion multi-échelles
- Analyse de proximité ou recouvrement entre objets

Les outils de prédictivité

- Weight Of evidence
- Fuzzy logic
- *Etc.*

Analyse Spatiale Statistique / Prédictive

Intervenant	Charles GUMIAUX (ISTO / OSUC)
Pré requis	Statistiques, notions de cartographie
Période d'enseignement	semaine banalisée courant Janvier
Forme	Cours et TP sur logiciel SIG
Évaluation	Compte rendu sur les manip de TP

A regarder avant

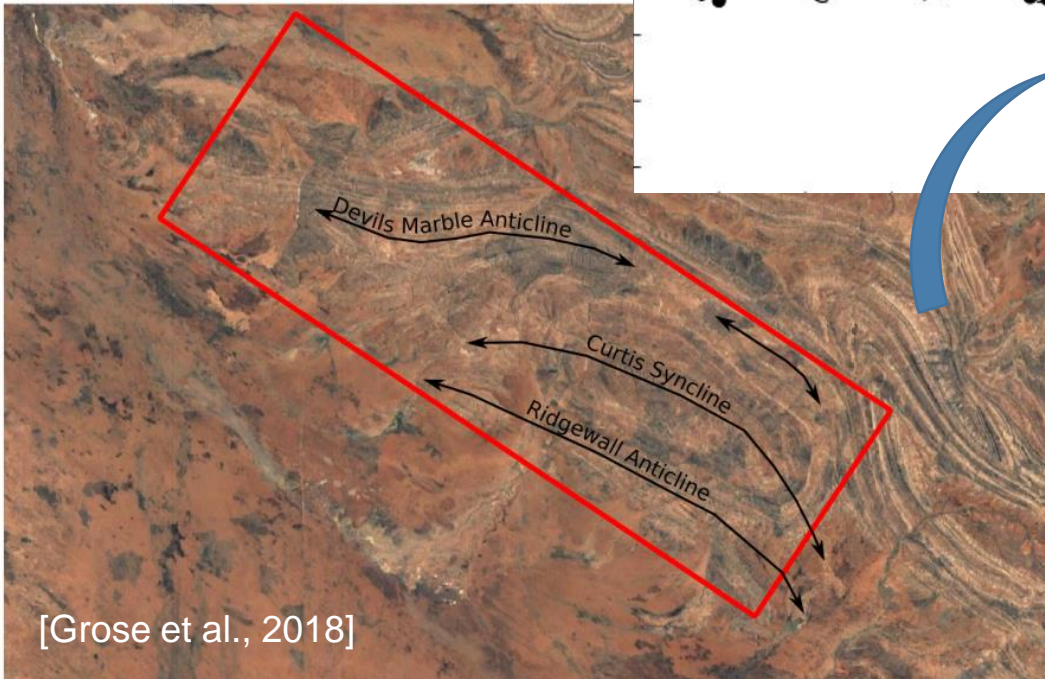
Des pointeurs utiles ...

Imagerie et interpolation des structures géométriques 3D

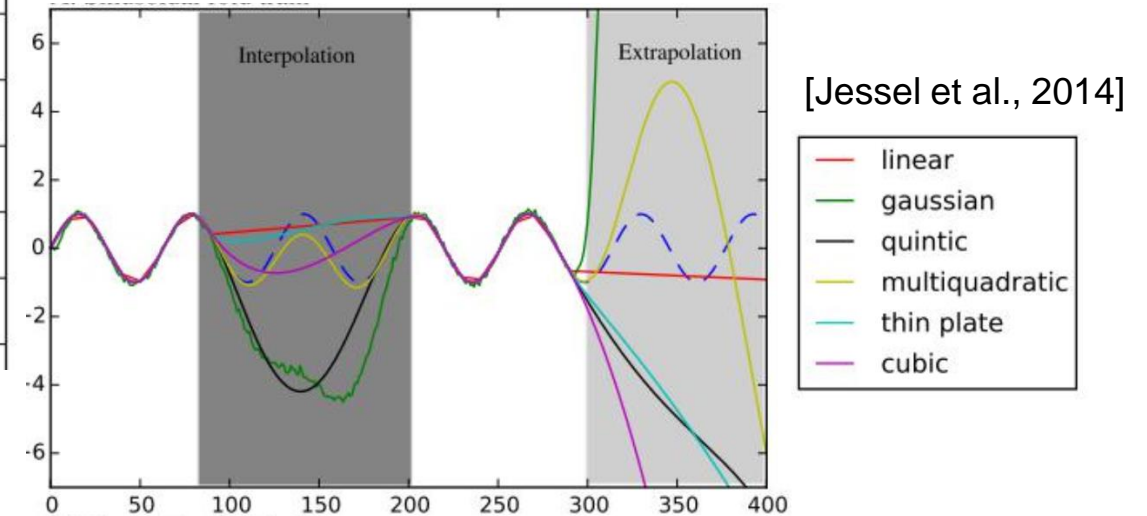
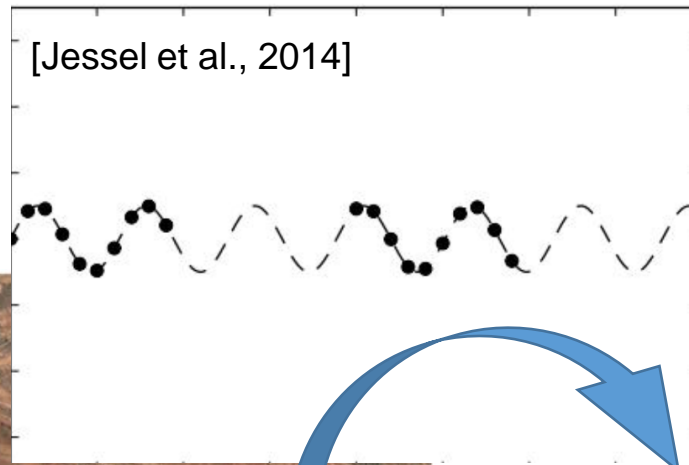
Échantillonnage

Interpolation
Extrapolation

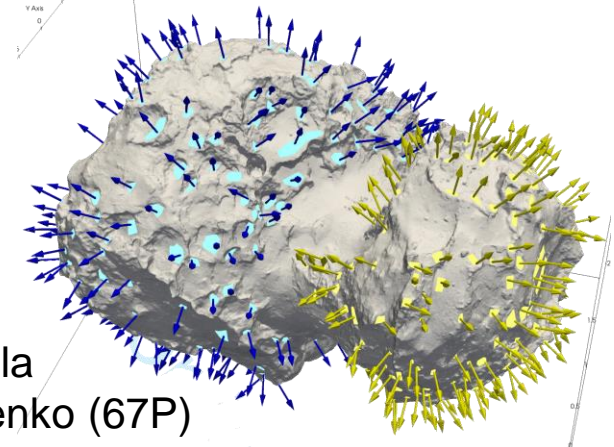
Observations
d'objets naturels



Vue satellite de structures géologiques plissées
(Davenport, Australie)



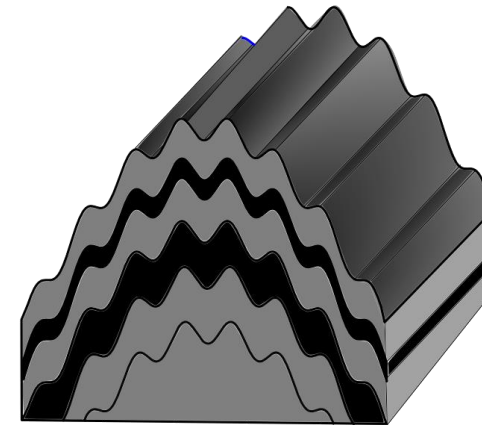
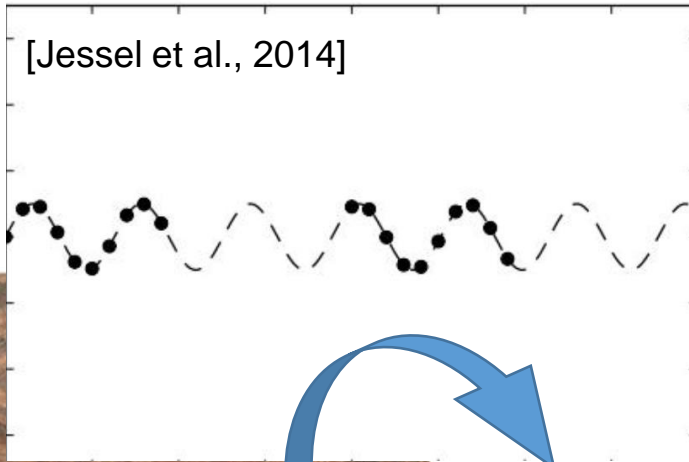
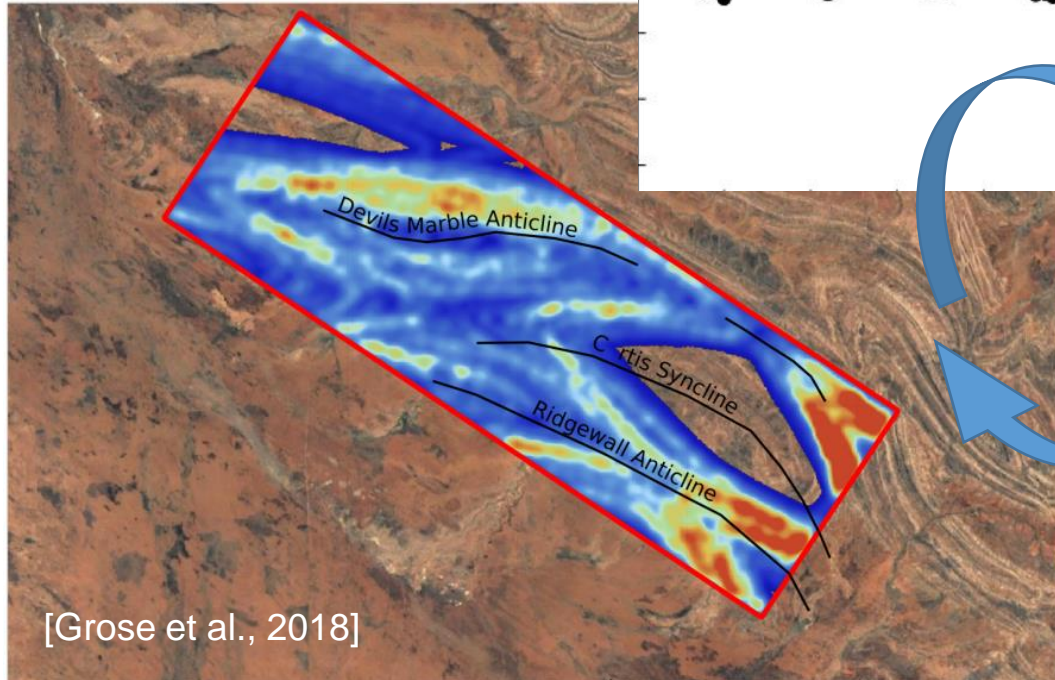
Et aussi, d'autres domaines
d'application que la géologie :
e.g., Mesures structurales de la
comète Churyumov-Gerasimenko (67P)



Imagerie et interpolation des structures géométriques 3D

Échantillonnage

Observations
d'objets naturels

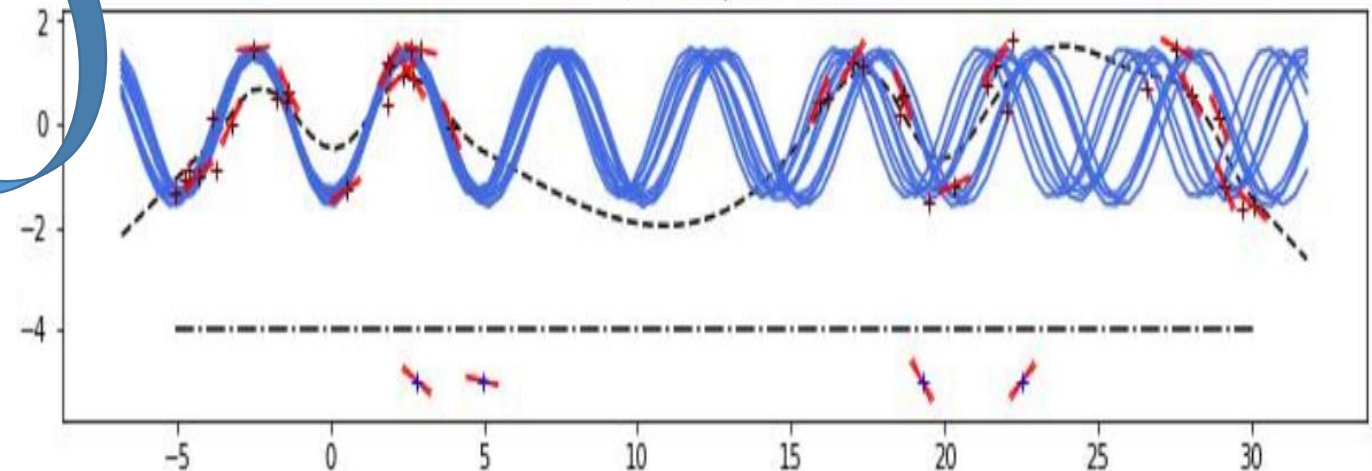


Modèle
conceptuel

[Grose et al., 2018]

Incertitudes

Simulations



Imagerie et interpolation des structures géométriques 3D

Intervenant	Gautier Laurent (ISTO/OSUC)
Pré requis	Notions de géométrie et de programmation python recommandées (mais non obligatoires)
Période d'enseignement	semaine banalisée
Forme	Introduction en CM + TP classique puis classe inversée
Évaluation	Mini projet et présentation avec co-évaluation par les pairs

A regarder avant

Des bases en python (eg., sur OpenClassroom)

SIG Raster et 3D : initiation et modélisation environnementale

Le module s'adresse particulièrement aux étudiants ayant une solide base en environnement qui souhaitent s'initier à la représentation de la donnée en mode image. La formation revient dans sa partie théorique sur les fondamentaux de la géomatique avec un glissement progressif vers le SIG image et ses principales différences avec les SIG vecteur. Au total 4 parties seront abordées durant la formation

- le formatage et l'intégration de la donnée dans un SIG Raster
- retour sur la notion de données en mode Raster et interrogation des images
- passage en revue des principales fonctionnalités de manipulation des données
- application de quelques traitements simples
- introduction à l'analyse spatiale et à la modélisation (interpolation et prédiction « markov »)
- représentation 3 D et valorisation

SIG Raster et 3D : initiation et modélisation environnementale

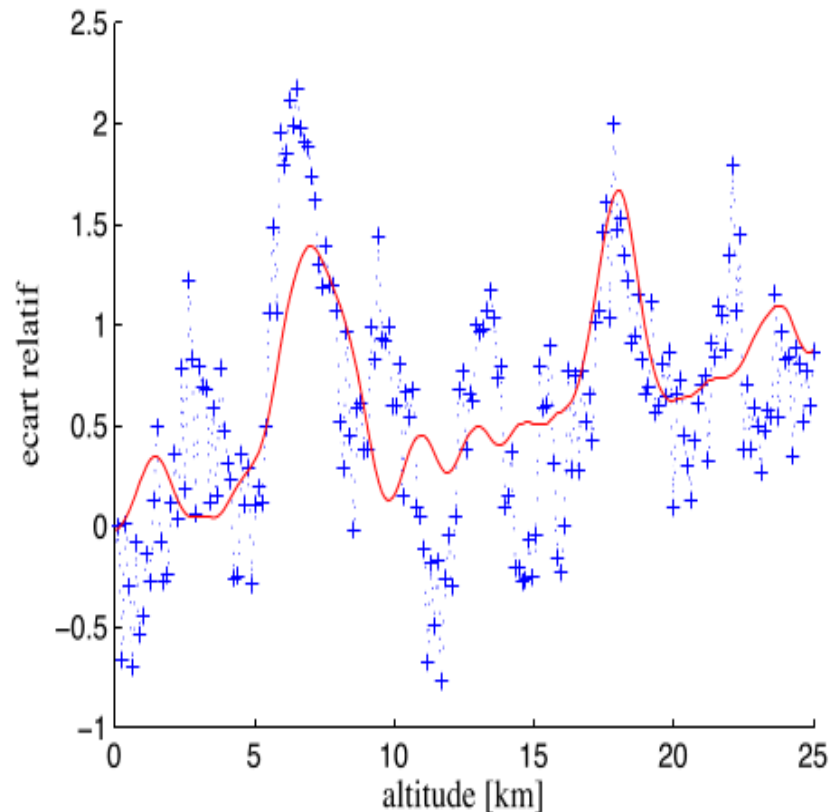
Intervenants	
Pré requis	
Période d'enseignement	semaine banalisée ou le mardi fin d'après midi
Forme	Cours ? TD ? TP ?
Évaluation	

A regarder avant

Des pointeurs utiles ...

Analyse de données par des cas pratiques

Une introduction au **Data Science** au travers d'exemples, en mettant résolument l'accent sur la mise en pratique.... et sur les erreurs à ne pas commettre !



Exemple : comparaison entre des observations (+) et un modèle (—).

- Peut-on dire qu'il y a accord ?
- Comment quantifier cela ?
- Comment exploiter au mieux les informations contextuelles ?

Au programme :

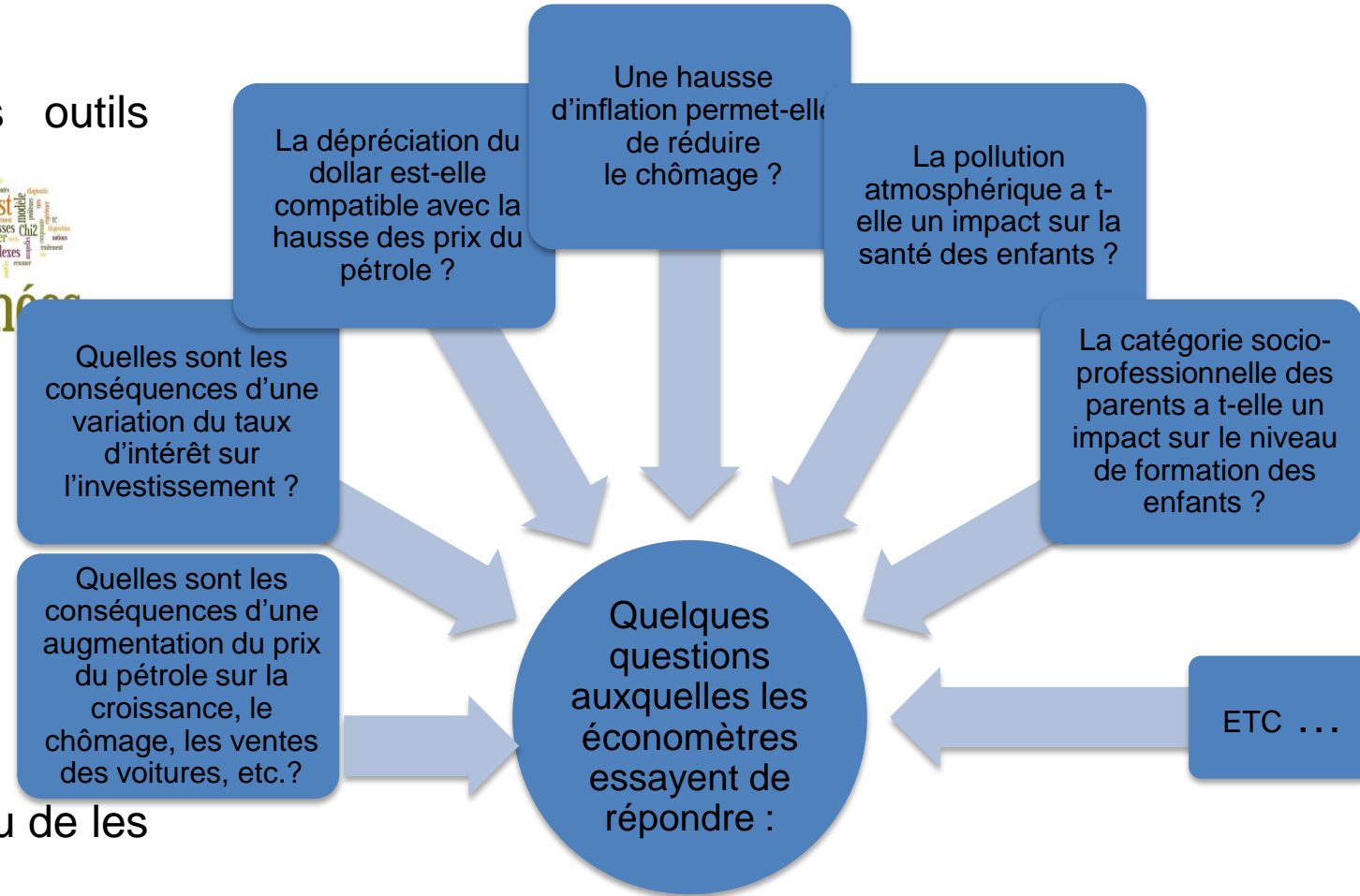
- Les incertitudes : faites-en votre meilleur allié (et comment les estimer)
- Corrélations : quand deux résultats sont-ils similaires ?
- Ajustement de courbes : du moindre carré aux modèles de Bayes
- L'analyse spectrale : un puissant diagnostic
- Exemples tirés de physique, chimie, sciences de la Terre, biologie...

Analyse de données par des cas pratiques

Intervenant	Thierry DUDOK DE WIT (OSUC)
Pré requis	Pratique courante d'un langage de programmation (Python, Matlab, C, ...) et notions de statistiques de niveau licence
Période d'enseignement	Mardi fin d'après midi
Forme	Cours-TD interactif
Évaluation	Mini-projet à rendre en Janvier
A regarder avant	Cours, exercices et divers liens utiles sont tous sur CELENE

- **L'économétrie** représente l'application des outils mathématiques et statistiques

aux théories économiques et de les corroborer ou de les réfuter. (Maddala, 1992)



Mots clés

- régression, corrélation, causalité, estimateur, paramètre, test statistique, intervalle de confiance...

Méthodologie de l'économétrie

Intervenants	Denisa BANULESCU-RADU (LEO)
Pré requis	Algèbre linéaire, notions de base de probabilités, enthousiasme & curiosité
Période d'enseignement	Semaine banalisée (Janvier 2020)
Forme	Cours – TD interactif
Évaluation	Mini-projet

A regarder avant

Le matériel disponible sur CELENE

