Big Data : volume, vélocité, variété

ECTS
6 crédits

Composante
ENSMAC

Présentation

Code interne : PB0BDATA

Description

La vitesse importante de développement de nouvelles technologies telles que les microprocesseurs, les systèmes de stockage, la 5G, les puces RFID et la blockchain a ouvert la voie à la collecte d'une multitude de données à une vitesse et un volume sans précédent. Face à cette abondance d'informations, leur analyse ou leur exploitation requiert l'utilisation de méthodes spécifiques : analyses exploratoire des données, algorithmes d'apprentissage automatique (machine learning) avec des large language model (LLM) comme ChatGPT.
C'est dans ce contexte que ce cours se propose de sensibiliser les étudiants à l'importance et à l'exploitation des jeux de données massifs. L'essentiel de ce module est tourné vers la pratique à travers des TP et projets. En acquérant ces connaissances, ils seront en mesure de jouer le rôle clé d'interface entre les chimistes/biologistes et les data scientists, favorisant une exploitation optimale des informations contenues dans ces données.

Heures d'enseignement

PRJ	Projet	20h

Pré-requis obligatoires

Cours de statistiques de première année

Syllabus

Contenu

Généralités concernant la génération, la manipulation, la représentation de jeux de données massifs (4 h)

- Acquérir et organiser : design expérimental, règles/principes, formats, type de données, de bases de données)

- Stocker, partager, protéger, archiver : enjeu du stockage, pratiques pour le partage, gestion des versions, sécurité des données contre piratage ou perte accidentelle, l'utiliser de manière éthique (RGPD, manipulation de masse)
- Manipuler : extraire, transformer, nettoyer, analyse exploratoire
- Visualiser : Rappel des règles de présentation des données et représentations graphiques de base (placer les données dans un contexte, erreur à ne pas commettre, trouver une représentation adaptée), visualisations adaptées pour jeux de données massifs (ACP, heatmaps, matrices de corrélation)

Projet analyse de donnée (20 h)

Un projet en groupe d'analyse de donnée sur un véritable jeu de données expérimentales sera à mener. L'objectif principal de ce projet sera d'explorer et d'analyser ce jeu de données afin de répondre aux questions scientifiques posées par l'expérimentateur. Cette expérience pratique permet de se confronter directement à un jeu de données massives, en mettant en œuvre de manière concrète des méthodes de manipulation, d'analyse et de visualisation. L'évaluation du projet se fera à travers la soumission d'un script ainsi qu'une présentation orale. Cette démarche permettra de développer leurs compétences en matière d'analyse de données massives tout en relevant les défis réels rencontrés dans la pratique scientifique.

Technologie Blockchain (8 h)

Bitcoin est né en 2009 et fait régulièrement le buzz dans la presse. Mais Bitcoin n'est qu'une application parmi d'autres de la technologie Blockchain. Le but de cette partie est tout d'abord de comprendre le fonctionnement de Bitcoin, de prendre conscience du changement de paradigme que la technologie blockchain peut induire (notamment en donnant la possibilité de partager/certifier des données en se passant des « tiers de confiance » habituels) et de réfléchir aux conséquences possibles et éventuellement à venir dans le monde de l'entreprise, la recherche, etc ... Les parties suivantes seront abordées :
- Principes de base de la technologie blockchain (CM, 2h)
- « TP Bitcoin » (TP, 4h)
- Ecosystème blockchain : présentation de différents projets blockchains (présentation orale, 2h)

Algorithmes d'apprentissage profond (deep learning) (18 h)

Le deep learning est une branche de l'apprentissage automatique (machine learning) qui a connu un développement spectaculaire ces dernières années. Il repose sur des architectures de réseaux de neurones artificiels profonds, capables d'apprendre des représentations de données complexes et de réaliser des tâches telles que la reconnaissance d'images, la traduction automatique, la génération de texte, etc. L'objectif de ces TP est de présenter les concepts fondamentaux et de les appliquer sur différentes thématiques.

Modalité d'évaluations

Projet et soutenance

Responsable

Emilien Peltier

Informations complémentaires

Entreprises, Métiers et Cultures

Modalités de contrôle des connaissances

Évaluation initiale / Session principale - Épreuves

Type d'évaluation	Nature de l'épreuve	Durée (en minutes)	Nombre d'épreuves	Coefficient de l'épreuve	Note éliminatoire de l'épreuve	Remarques
Contrôle Continu	Evaluation de compétences					

Seconde chance / Session de rattrapage - Épreuves

Type d'évaluation	Nature de l'épreuve	Durée (en minutes)	Nombre d'épreuves	Coefficient de l'épreuve	Note éliminatoire de l'épreuve	Remarques
Contrôle Continu	Evaluation de compétences					

Infos pratiques

Contacts

Emilien Peltier
Emilien.Peltier@bordeaux-inp.fr