Fondements de la sciences des données 2

ECTS : 6

Description du contenu de l'enseignement :

Lors des séances prévues, il s'agira de présenter les outils disponibles en R pour la prise en main et l'analyse des données complexes. En particulier, seront abordées les techniques exploratoires pour l'analyse des données comportant un caractère temporel, une structure en réseau, ou des informations textuelles.

Dans un premier temps, il sera question de données tabulaires et de leur analyse exploratoire : méthodes factorielles et réduction de dimension, traitement des données manquantes, recherche de motifs, clustering.

Dans un deuxième temps, seront traitées les données ayant un caractère temporel, et notamment les données longitudinales.

Dans un troisième temps, il sera question d'analyse de graphes (caractérisation d'un réseau via différents indices, comparaison à des graphes aléatoires, recherche de communautés, exploration de réseaux temporels).

Enfin, dans un quatrième temps, sera abordée la question de la récupération des données du web (web scrapping, utilisation des API), et des données textuelles (analyse basique, sacs de mots, topic models).

Selon le temps disponible, d'autres points pourront être abordés : science ouverte et reproductibilité, visualisation interactive, création d'interfaces, traitement des données massives.

Les différents outils et méthodes seront illustrés sur des jeux de données réelles, et en utilisant des packages R existants.

Compétence à acquérir :

Gagner en autonomie sur le traitement des données avec R.

Connaitre et savoir appliquer des méthodes récentes pour l'analyse des données complexes.

Comprendre les enjeux de la reproductibilité des résultats, et avoir une démarche scientifique en ce sens.

Mode de contrôle des connaissances :

Projet à réaliser individuellement ou en groupe.