Introduction aux mégadonnées en sciences sociales (FAS 1001)

Hiver 2025

Horaire du cours: Jeudi, 8h30 - 11h29

Local de cours: pavillon Marie-Victorin local G-415_555B

Disponibilité: Sur rendez-vous

Nom du chargé de cours: Laurence-Olivier M. Foisy

Courriel: lomf0@ulaval.ca

Prérequis

Ce cours ne demande aucun prérequis en statistiques avancées ou en programmation, mais avoir des connaissances au préalable demeure un atout. Il est fortement conseillé de suivre en parallèle le cours FAS 1003 - *Visualisation des données*, car la visualisation graphique demeurera un élément important dans la présentation de vos travaux. Ce cours comportera une section magistrale ainsi que des ateliers pratiques en classe si le temps le permet. Les élèves sont encouragés à poser des questions et à participer activement aux discussions.

Description du cours et objectifs

Jamais autant de données n'ont été disponibles pour comprendre les comportements humains. De nouveaux outils de recherche nous permettent désormais de quantifier des données difficilement analysables auparavant, tels que de larges corpus de textes, des images ou des audios en provenance de vidéos. Néanmoins, comment collecter, traiter et analyser ces nouvelles données? Comment combiner ces nouvelles données avec celles déjà largement utilisées en sciences sociales comme les sondages? Quels sont les enjeux techniques et éthiques que

soulèvent l'utilisation de ces données en recherche dans le contexte d'un développement important de l'intelligence artificielle et de l'application croissante de l'apprentissage par machines en sciences sociales? Ce cours aborde ces nombreux enjeux avec une ***approche pratique*** à l'utilisation des mégadonnées en sciences sociales.

À la fin de ce cours, les objectifs suivants seront remplis:

- Avoir une connaissance globale des différentes sources de données disponibles pour étudier les phénomènes sociaux.
- Développer l'autonomie nécessaire pour collecter, gérer et analyser quantitativement des bases de données et les présenter avec un projet de recherche les mobilisant.
- Faire preuve d'une compréhension des enjeux entourant la mobilisation et l'utilisation de volumineuses bases de données en sciences sociales.

Pédagogie

Le langage de programmation utilisé pour ce cours est `R`. Ce dernier est téléchargeable gratuitement ici. Bien que plusieurs options soient possibles, il vous est également demandé de télécharger `RStudio` ici. RStudio est l'environnement de développement intégré de prédilection pour coder en `R`. Il vous permettra d'éditer et de déboguer plus facilement votre code en plus de vous donner les outils nécessaires pour transformer, prévisualiser et analyser vos données efficacement. Les étudiant-es devront, si ce n'est pas déjà fait, créer un compte `GitHub` lors du premier cours, afin de remettre leurs codes de travaux pratiques et de travaux de session. `GitHub` est un service web d'hébergement et de gestion de développement de logiciels qui est largement utilisé en industrie pour le partage de codes. Cela permettra de socialiser les étudiant-es à son utilisation.

Le bon déroulement de ce cours nécessite que vous ayez un ordinateur, notamment pour les nombreux ateliers pratiques lors des classes. Si vous n'avez pas d'ordinateur, une solution sera trouvée, afin que vous puissiez participer aux activités en classe.

Les élèves doivent télécharger `Slack` ici pour communiquer avec le chargé de cours et les autres étudiant-es. `Slack` est une plateforme de communication collaborative qui permet de partager des fichiers, de poser des questions et de discuter en temps réel. Ce canal de communication sera privilégié aux courriels pour les questions en dehors des heures de cours.

Site du cours

Le site web du cours est disponible au https://fas1001.com. Tous les informations concernant le cours y seront téléversés. Les diapositives, les travaux pratiques, le plan de cours et les dates importantes.

Évaluations

Sur un total de 100 points:

Travaux pratiques (60/100)

Les travaux pratiques devront être mis sur votre compte `GitHub` **une journée avant le cours suivant, soit mercredi avant minuit**.

- Introduction à git, GitHub et Quarto : 22 janvier 2025
- Travail de mi-session : 12 mars 2025
- Travail de session : 30 avril 2025

Introduction à `git`, `GitHub`, et `Quarto` (10/100)

10% des points seront alloués à **l'introduction à `GitHub`**.

Travail de mi-session (20/100)

20% des points seront alloués à la remise d'un **plan de recherche de votre travail de session de 2-4 pages**. Le plan de recherche permettra d'avoir un regard du chargé de cours sur l'évolution du travail de session et de corriger la trajectoire de ce dernier si nécessaire. Le plan de recherche consistera principalement en la présentation de votre question de recherche, les raisons motivant votre recherche, les données que vous comptez utiliser et la ou les méthode-s mobilisée-s. Ce travail sera la base de votre travail de session final.

Travail de session (35/100)

30% des points seront alloués à **votre travail de session individuel pouvant aller de 8 à 10 pages**. Ce dernier consistera en une recherche complète avec présentation de la question de recherche, les raisons motivant cette dernière, la présentation des données utilisées, la ou les méthode-s mobilisée-s, l'analyse des données et de la présentation des conclusions de l'étude. La ou les méthode-s utilisée-s pour analyser les données n'ont pas besoin d'être poussée-s, une plus grande attention sera portée sur les données collectées, la transformation effectuée sur ces dernières et leur présentation.

Quiz (30/100)

30% des points seront alloués aux **Quiz en classe**. Les étudiant-es devront compléter deux Quiz de 1h00 sur leurs ordinateurs qui permettront d'évaluer leurs connaissances à propos de la matière apprise en classe. - Quiz 1 (Données structurés) : 13 février 2025 - Quiz 2 : 10 avril 2025

Participation en classe (10/100)

10% des points seront alloués à **la participation en classe**. L'entraide entre les étudiant-es est encouragée (évaluation et correction de code sur `GitHub` et entraide sur `Slack`.

Résumé des évaluations

Évaluations	Points	Dates
Introduction à `GitHub`	10%	22 janvier 2025 avant minuit
Quiz 1 - En classe	15%	13 février 2025
Travail de mi-session	20%	12 mars 2024 avant minuit
Quiz 2 - En classe	15%	10 avril 2025
Travail de session	30%	30 avril 2024 avant minuit
Participation en classe	10%	9 janvier au 30 avril
Total	100%	9 janvier au 30 avril

Lectures

Le livre de référence du cours sera celui du Professeur Rohan Alexander de l'Université de Toronto *Telling Stories with Data* et disponible gratuitement sur son site web.

Pour un livre ressource en méthodes quantitatives, il est conseillé de consulter le livre du Professeur Vincent Arel Bundock *Analyse causale et méthodes quantitatives* disponible gratuitement sur le site des Presses de l'Université de Montréal.

Calendrier

Cours 1 (9 janvier) : Introduction

Objectifs principaux

- Présenter la structure du cours, les ressources et le système d'évaluation
- Introduire Slack, le site web du cours, le github du cours.
- Télécharger R et RStudio, s'inscrire sur GitHub

Contenu

- Revue du syllabus : objectifs, notation, dates limites importantes, vue d'ensemble du parcours des données
- Pourquoi les big data (et les données en général) sont importantes en sciences sociales

Cours 2 (16 janvier) : Git, GitHub, Quarto, terminal et bonnes pratiques

Objectifs principaux

- S'assurer que les étudiants maîtrisent les bases du contrôle de version
- Leur montrer comment créer des documents reproductibles (Quarto)
- Établir une "hygiène numérique" pour l'organisation des fichiers, les conventions de nommage, les sauvegardes

Contenu

- Fondamentaux de Git : git add/commit/push/pull, branches, résolution de conflits mineurs
- GitHub : stockage des dépôts, fork, pull requests (collaboration)
- Bases de Quarto : création de documents, rendu du code R dans les fichiers .qmd
- Commandes Terminal : navigation (ls, cd, mkdir, rm), édition de texte
- Intégration de RStudio avec Git/GitHub

> **!** Important
>
> Remise du TP1 le 22 janvier 2025 avant minuit

Semaine 3 (23 janvier) : Révision R (Chargement de données, dplyr, Fusionner des BD, etc.) et introduction aux données structurées.

Objectifs principaux

- Donner un aperçu ou une révision sur la manipulation des données R (dplyr)
- S'assurer que tout le monde peut gérer CSV/rds/sav, faire des fusions/jointures, des résumés basiques

Contenu

- Importation de données
- Verbes dplyr : select, filter, mutate, summarize, group_by
- Pivotement/manipulation de base avec tidyr
- Les types de données (caractère vs. facteur, numérique, date)

Semaine 4 (30 janvier) : Données de sondages, nettoyage, mesure d'un concept latent et éthique des données

Objectifs principaux

- Introduire les données structurées typiques en sciences sociales
- Explorer comment mesurer les variables latentes avec l'analyse factorielle, l'alpha de Cronbach, etc.
- Introduire l'éthique des données

Contenu

- Conception de sondages, types de questions (Likert, échelle, etc.)
- Spécificités du nettoyage des données
- Analyse factorielle simple ou vérifications de fiabilité pour mesurer une variable latente
- Éthique de base des données

Semaine 5 (6 février) : Données de sondages, PCA et clustering

Objectifs principaux

- Approfondir l'apprentissage non supervisé (clustering) dans le contexte des données de sondage

- Montrer comment interpréter les clusters (segments de groupe, par ex., typologie politique)

Contenu

- Révision des statistiques descriptives pour les données multi-variables
- K-means ou clustering hiérarchique sur données d'enquête (par ex., attitudes des répondants)
- Interprétation des centres de clusters, scores silhouette, ou diagnostics de base des clusters

Semaine 6 (13 février) : Données de sondages, apprentissage machine et Quiz 1

> **!** Important
>
> Quiz 1 en classe. Vous aurez 1 heure pour le compléter.

Objectifs principaux

- Montrer comment l'apprentissage automatique supervisé peut être utilisé pour prédire un résultat à partir de variables d'enquête
- Garder un niveau élevé : comment faire les divisions train/test, mesurer la précision, interpréter les résultats

Contenu

- Révision rapide de la régression linéaire vs. logistique en R
- Introduction aux arbres de décision ou random forest pour la classification
- Surapprentissage, validation croisée, métriques de performance de base

Semaine 7 (20 février) : Rencontres individuelles

Objectifs principaux

- Fournir un retour personnalisé sur le TP 1 et le Quiz 1
- Discuter des travaux de mi-session

Semaine 8 (27 février) : Introduction aux Données Textuelles + Infrastructures de Données

Objectifs principaux

- Transition vers des données moins structurées (texte)
- Donner un aperçu du stockage/infrastructure des données au niveau conceptuel : du CSV local aux bases de données relationnelles.

Contenu

- Pourquoi les données textuelles sont importantes en sciences sociales (tweets, actualités, transcriptions)
- Idées de base des bases de données relationnelles vs. non-relationnelles
- Considérer la taille des données : stockage local vs. cloud

Semaine 9 (6 mars) : Semaine de Lecture

Objectifs principaux

- Compléter le travail de mi-session

> **!** Important
>
> Remise du TP2 le 12 mars 2025 avant minuit

Semaine 10 (13 mars) : Données textuelles et méthodes d'analyse

Objectifs principaux

- Fournir plus de détails sur la manipulation de texte (tokenisation, mots vides)
- Analyse de sentiment et modélisation de sujets

Contenu

- Approche tidytext en R (tokenisation, suppression des mots vides, etc.)
- Analyse de sentiment de base
- (Optionnel) Introduction à la modélisation de sujets (LDA) si le temps le permet

Semaine 11 (20 mars) : Web Scraping, Introduction à HTML, API

Objectifs principaux

- Enseigner le scraping de base (structure HTML, utilisation de rvest ou similaire en R)
- Manipulation des données JSON d'une API
- Limites de taux, directives éthiques de scraping, et CGU

Contenu

- Utilisation du package rvest
- Accéder aux API REST avec httr ou curl, analyse JSON avec jsonlite
- Éthique & légalité du scraping (robots.txt, confidentialité, etc.)

Semaine 12 (27 mars) : LLM via API

Objectifs principaux

- Montrer comment utiliser les grands modèles linguistiques avec R
- Discuter de l'ingénierie de prompt de base, interpréter les résultats
- Aborder les biais potentiels

Contenu

- Création de prompts structurés pour la génération ou classification de texte
- Prix et utilisation des tokens—contraintes du monde réel

Semaine 13 (3 avril) : Rencontres individuelles

Objectifs principaux

- Répondre aux questions ou blocages

Contenu

- Pas de nouveau cours. Consultations à propos du travail final
- Préparation pour le Quiz 2

Semaine 14 (10 avril) : Données non structurées (Images et audio) et Quiz 2

> **!** Important
>
> Quiz 2 en classe. Vous aurez 1 heure pour le compléter.

Objectifs principaux

- Introduction à l'analyse d'images (détection d'objets, classification) et audio (parole-en-texte, traitement du signal de base)

Contenu

- Démonstration rapide : lecture d'images en R (package magick ou imager)
- Pipeline conceptuel de base pour la classification d'images

Semaine 15 (17 avril) : Introduction à R Shiny

Objectifs principaux

- Enseigner comment faire une application web interactive ou tableau de bord en R pour présenter des analyses
- Résumer l'ensemble du cours en se concentrant sur la communication des résultats

Contenu

- Bases de Shiny : UI + serveur, expressions réactives et graphiques
- Déploiement sur shinyapps.io ou autres solutions d'hébergement

> **!** Important
>
> Remise du TP Final le 30 avril 2025 avant minuit

Rappel de règlements pédagogiques

Veuillez prendre note que le trimestre commence le 8 janvier et se termine le 30 avril 2024 (incluant la période des examens) et que la présence physique est attendue à tous les cours. Aucune demande d'examen différé ne sera acceptée sans motif valable. Nous entendons par motif valable, un motif indépendant de votre volonté, tel que la force majeure, le cas fortuit ou une maladie attestée par un certificat de médecin.

Absence à un examen

Il est de votre responsabilité de motiver, en remplissant le formulaire disponible dans le Centre étudiant, toute absence à une évaluation ou à un cours faisant l'objet d'une évaluation continue dès que vous serez en mesure de constater que vous ne pourrez pas vous présenter à une évaluation. Vous devez obligatoirement fournir les pièces justificatives dans les sept jours suivant l'absence.

Délais pour la remise d'un travail

Vous devez motiver, en remplissant le formulaire disponible dans le Centre étudiant, toute demande de délai pour la remise d'un travail et fournir les pièces justificatives dès que vous êtes en mesure de constater que vous ne pourrez pas remettre à temps le travail.

La pénalité imposée pour les retards dans la remise des travaux est de 10 points de pourcentage par jour. Cette pénalité est calculée en déduisant 10 points de pourcentage à la note obtenue pour le travail en question. Il s'agit de la politique « par défaut » du Département; le corps enseignant est libre d'imposer une pénalité plus élevée s'il le désire. La personne étudiante qui remet son travail après 23h59 sur Studium le jour de la remise est réputé les avoir remis le matin du jour ouvrable qui suit et les jours non ouvrables sont comptés comme des jours de retard.

Prévention du plagiat

Le Département porte une attention toute particulière à la lutte contre le plagiat, le copiage ou la fraude lors des examens. Le plagiat consiste à utiliser de façon totale ou partielle, littérale ou déguisée le texte d'autrui en le faisant passer pour sien ou sans indication de référence à l'occasion d'un travail, d'un examen ou d'une activité faisant l'objet d'une évaluation. Cette fraude est lourdement sanctionnée.

Toutes les personnes étudiantes sont invitées à consulter le site web http://www.integrite.umontreal.ca/ et à prendre connaissance du Règlement disciplinaire sur le plagiat ou la fraude concernant les étudiants. Plagier peut entrainer un échec, la suspension ou le renvoi de l'Université.

Bibliothécaire et règles bibliographiques

Il est obligatoire de respecter les règles de présentation et de citations/références (modèle de Chicago pour les travaux et examens-maison du Département de science politique. Deux guides à cet effet sont disponibles sur le site du département aux adresses suivantes: Pour la présentation des travaux:

https://bib.umontreal.ca/economie-politique-relations-industrielles/science-politique

Pour les citations et références:

https://bib.umontreal.ca/citer/styles-bibliographiques/chicago

N'hésitez pas à profiter des services de la bibliothécaire spécialisée en science politique Julia Généreux Randall. Vous pouvez la rejoindre à son bureau (local 3017 de la Bibliothèque des lettres et sciences humaines, Pavillon Samuel-Bronfman) ou lui envoyer un courriel. La BLSH met aussi à disposition un Guide internet, point de départ idéal pour toute recherche documentaire en science politique.

Le harcèlement, y compris à caractère sexuel

Il incombe à chaque membre de la communauté universitaire de se conduire avec respect en tout temps envers tout le monde. En particulier, le Département de science politique s'engage à créer un milieu accueillant et sécuritaire pour toutes et tous, quelle que soit leur identité.

Les documents suivants ont des démarches pratiques à suivre : Si vous pensez que vous vivez du harcèlement : https://respect.umontreal.ca/obtenir-de-laide/vous-vivez-une-situation-difficile/. Si on s'est confié à vous ou si vous êtes témoin de harcèlement : https://respect.umontreal.ca/obtenir-de-laide/vous-avez-ete-temoin-dune-situation/. Pour toute autre question : https://respect.umontreal.ca/accueil/

Besoin d'écoute? Situation de détresse?

Vous pouvez faire appel à plusieurs **lignes d'écoute** ou d'urgence. Vous avez accès à un service 24 heures/7 jours offert par l'Alliance pour la santé étudiante au Québec. Le numéro

est le suivant : 1-833-851-1363. Vous retrouverez les services d'aide disponibles du le site du Service à la vie étudiante : https://toutlemondeadesbas.ca/

Vous pouvez aussi faire appel à une **sentinelle**. La sentinelle est employée par l'UdeM, formée et disponible pour vous accueillir, vous écouter et vous orienter vers les bonnes ressources. Son accueil est spontané, respectueux et strictement confidentiel. Le service est offert en plusieurs langues. Bottin des sentinelles : http://cscp.umontreal.ca/activiteprevention/sentinelle.htm

Si vous souhaitez discuter avec des pairs du stress que peut occasionner la vie étudiante, le local du **PASPOUM** au C-3144 est ouvert (3e étage, Pavillon Lionel-Groulx). Une personne étudiante formée à l'écoute active pourra vous orienter vers des ressources appropriées. Le local du PASPOUM est aussi un espace où vous pouvez déconnecter pendant quelques instants. Consultez les heures d'ouverture et les activités du PASPOUM sur la page Facebook. Vous pouvez vous abonner au compte Instagram du même nom pour suivre les actualités.