Université de Sherbrooke

## Département d'informatique

# IFT 599 / IFT 799 – Science des données

### Plan d'activité pédagogique

Automne 2021

## Enseignant


| Shengrui Wang | |
|---|---|
| Courriel : | shengrui.wang@usherbrooke.ca |
| Local : | D4-1018-1 |
| Téléphone : | +1 819 821-8000 x62022 |
| Disponibilités : | Périodes de consultation : jeudi 14 h à 16 h |


**Responsable(s)** : Direction du département

## Site web du cours : P:\Cours\IFT599

## Horaire

| | | | |
|---|---|---|---|
| Exposé magistral : | Mardi | 10h30 à 12h20 | salle D2-1060 |
| | Jeudi | 11h30 à 12h20 | salle D7-2023 |

## Description officielle de l'activité pédagogique[1]

| | |
|---|---|
| Cibles de formation : | Comprendre et maîtriser des théories et méthodes de base pour la science des données. |
| Contenu : | Inférence statistique : procédures statistiques fondamentales, estimation des paramètres d'un modèle, tests des hypothèses liées aux caractéristiques structurelles d'un modèle, intervalle de confiance pour les paramètres de modèle. Techniques de forage de données : analyse exploratoire des données, prétraitement, visualisation, recherche et extraction des règles d'association, classification et prédiction, analyse de regroupement. Recherche d'information : principe, concepts de base, indexation, engins de recherche. Applications dans divers domaines tels que la santé, l'intelligence d'affaires, les réseaux sociaux et la finance. |
| Crédits | 3 |
| Organisation | 3 heures d'exposé magistral par semaine<br>6 heures de travail personnel par semaine |
| Préalable | IFT 436, STT 418 |
| Particularités | Aucune |

[1] https://www.usherbrooke.ca/admission/fiches-cours/ift599

# 1 Présentation

Cette section présente les cibles de formation spécifiques et le contenu détaillé de l'activité pédagogique. Cette section, non modifiable sans l'approbation du comité de programme du Département d'informatique, constitue la version officielle.

## 1.1 Mise en contexte

De nos jours, le volume de plus en plus important de données disponibles (que l'on appelle données massives, mégadonnées ou bien « big data ») et leur démocratisation combinée avec les puissances de calcul et de stockage, nous amènent vers une nouvelle ère d'analyse des données qui comporte beaucoup de défis. En effet, les mégadonnées qui ne cessent d'augmenter jour après jour sont souvent peu exploitées, alors qu'elles cachent des connaissances décisives face au marché et à la concurrence. La plupart des techniques d'analyse de données traditionnelles échouent à extraire de l'information pertinente sur ce type de données. Pour combler ce besoin, la science des données s'est rapidement développée ces dernières années. Elle s'intéresse à tous les aspects liés aux données notamment la collecte, le nettoyage, le filtrage, l'anonymisation, le cryptage, la visualisation, le stockage, le transfert, l'analyse et la modélisation (classification et prédiction), etc.. L'appellation « science des données » est issue de celle du forage de données (« Data Mining ») qui a été le porteur de la discipline avec l'apprentissage artificiel (« machine learning ») depuis une vingtaine d'années. La science des données est donc l'ensemble des algorithmes et méthodes destinés à l'exploration et l'analyse de grandes bases de données informatiques en vue de détecter dans ces données des règles, des associations, des tendances inconnues (non fixées a priori), des structures particulières restituant de façon concise l'essentiel de l'information utile, et de prédire les résultats des processus sous jacents afin de générer des valeurs pour la société.

## 1.2 Cibles de formation spécifiques

Ce cours vise à initier l'étudiante et l'étudiant aux concepts fondamentaux de sciences de données et d'appliquer ces notions à des problèmes concrets. À la fin de cette activité pédagogique, l'étudiante ou l'étudiant sera capable de :

1. comprendre les principaux concepts de la science des données et le processus d'extraction des connaissances dans les bases de données ;
2. maîtriser des concepts et techniques d'inférence statistique ;
3. maîtriser les techniques d'analyse des données transactionnelles, plus particulièrement la recherche des règles d'association ;
4. maîtriser des techniques de « clustering » et de classification les plus utilisées dans la pratique ;
5. maîtriser des techniques de prétraitement des données et de réduction de la dimension (essentiellement la gestion de la redondance et la sélection des attributs) ;
6. comprendre la problématique des données de grande dimension et maîtriser des techniques de « clustering » utilisées sur ce type de données ;
7. se familiariser avec des algorithmes de classement des pages Web tels que HITS et PageRank ;
8. se familiariser avec des algorithmes pour les systèmes de recommandation.

## 1.3 Contenu détaillé

| Thème | Contenu | Nbr. d'heures | Objectifs | Travaux | Lectures |
|---|---|---|---|---|---|
| 1 | Concepts de base : Processus d'extraction des connaissances dans les bases de données ; Prétraitement des données ; Mesures de similarités ; Techniques de réduction de la dimension (ACP : Analyse de composantes principales) ; Techniques de sélection d'attributs ; Concepts de base pour la détection des anomalies ; Visualisation ; Introduction de Weka (progiciel d'analyse des données). | 8 | 1 et 5 | ✔ | [5] [6] chapitres 1 et 2 |
| 2 | Analyse d'association et analyse des séquences : Introduction aux données transactionnelles ; Principe de l'algorithme *Apriori* et, si le temps permet, l'algorithme *FP-growth* ; Recherche et extraction des règles d'associations ; Extraction et utilisation des patrons significatifs pour l'analyse des séquences. | 8 | 3 | ✔ | [5] [6] chapitres 5 et 6 |
| 3 | Inférence statistique : Procédures statistiques fondamentales ; Estimation des paramètres d'un modèle ; Tests des hypothèses liées aux caractéristiques structurelles d'un modèle ; Intervalle de confiance pour les paramètres de modèle. | 4 | 1 et 2 | | [5] |
| 4 | Clustering (méthodes de base) : Clustering par partition et hiérarchique ; Clustering basé sur la densité ; Clustering des données catégoriques et transactionnelles ; Validation des résultats de clustering. | 6 | 4 et 5 | | [5] [6] chapitre 7 |
| 5 | Clustering (méthodes avancées) : Détection des anomalies (*outlier detection*) ; Clustering des données de grandes dimensions et des données complexes. | 6 | 4, 5 et 6 | ✔ | [5] [6] chapitre 8 [1] [7] |
| 6 | Systèmes de recommandation : Principes et éléments de bases ; Principe des algorithmes basés sur le contenu ; Approches basées sur le filtrage collaboratif ; Méthodes basées sur la ressemblance directe ; Méthodes basées sur la sémantique latente. | 6 | 5 et 8 | ✔ | [5] [3] |
| 7 | Web mining : Principes et éléments de bases ; Algorithmes de classement des pages Web (HITS et PageRank) ; Extraction des communautés dans le Web. | 2 | 7 | | [5] [2] [4] |

1. Le cours doit comprendre au moins quatre travaux pratiques couvrant tous les sujets marqués «✔» dans le tableau.
2. Les lectures indiquées ne sont là qu'à titre indicatif. L'enseignant est libre de choisir un autre document de référence.

# 2 Organisation

Cette section propre à l'approche pédagogique de chaque enseignante ou enseignant présente la méthode pédagogique, le calendrier, le barème et la procédure d'évaluation ainsi que l'échéancier des travaux. Cette section doit être cohérente avec le contenu de la section précédente.

## 2.1 Méthode pédagogique

La matière sera donnée sous forme de cours magistraux. L'enseignement se fera en présentiel. Toutefois, les cours seront transmis en simultané pour les étudiantes et étudiants qui ne peuvent se présenter en classe.

Compte tenu du contexte actuel (pandémie due au COVID-19), il se peut que le cours ait lieu en totalité ou en partie à distance d'une façon différente de ce qui est énoncé ci-dessus. Notez que vous en serez informés rapidement si tel est le cas.

## 2.2 Calendrier

| Semaine | Date | Thème | Travaux pratiques |
|---|---|---|---|
| 1 | 2021-08-30 | 1 | |
| 2 | 2021-09-06 | 1 | |
| 3 | 2021-09-13 | 1 | Remise TP1 |
| 4 | 2021-09-20 | 2 | |
| 5 | 2021-09-27 | 2 | Remise TP2 |
| 6 | 2021-10-04 | 4 | |
| 7 | 2021-10-11 | Révision et 4 | |
| 8 | 2021-10-18 | Examen périodique | |
| 9 | 2021-10-25 | Relâche | |
| 10 | 2021-11-01 | 5 | Remise TP3 |
| 11 | 2021-11-08 | 5 | |
| 12 | 2021-11-15 | 6 | |
| 13 | 2021-11-22 | 6 | Remise TP3 |
| 14 | 2021-11-29 | 3 | |
| 15 | 2021-12-06 | Révision et 3 | |
| 16 | 2021-12-13 | Examen final | |

## 2.3 Évaluation

| | |
|---|---|
| Travaux pratiques (4) | 30 % |
| Examen intra | 30 % |
| Examen final | 40 % |

### 2.3.1 Qualité de la langue et de la présentation

Conformément à l'article 17 du règlement facultaire d'évaluation des apprentissages[2] l'enseignante ou l'enseignant peut retourner à l'étudiante ou à l'étudiant tout travail non conforme aux exigences quant à la qualité de la langue et aux normes de présentation.

### 2.3.2 Plagiat

Le plagiat consiste à utiliser des résultats obtenus par d'autres personnes afin de les faire passer pour sien et dans le dessein de tromper l'enseignante ou l'enseignant. Vous trouverez en annexe un document d'information relatif à l'intégrité intellectuelle qui

[2] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/Etudiants_actuels/Informations_academiques_et_reglements/2017-10-27_Reglement_facultaire_-_evaluation_des_apprentissages.pdf

fait état de l'article 9.4.1 du Règlement des études[3]. Lors de la correction de tout travail individuel ou de groupe une attention spéciale sera portée au plagiat. Si une preuve de plagiat est attestée, elle sera traitée en conformité, entre autres, avec l'article 9.4.1 du Règlement des études de l'Université de Sherbrooke. L'étudiante ou l'étudiant peut s'exposer à de graves sanctions qui peuvent être soit l'attribution de la note E ou de la note zéro (0) pour un travail, un examen ou une activité évaluée, soit de reprendre un travail, un examen ou une activité pédagogique. Tout travail suspecté de plagiat sera transmis au Secrétaire de la Faculté des sciences. Ceci n'indique pas que vous n'ayez pas le droit de coopérer entre deux équipes, tant que la rédaction finale des documents et la création du programme restent le fait de votre équipe. En cas de doute de plagiat, l'enseignante ou l'enseignant peut demander à l'équipe d'expliquer les notions ou le fonctionnement du code qu'elle ou qu'il considère comme étant plagié. En cas d'incertitude, ne pas hésiter à demander conseil et assistance à l'enseignante ou l'enseignant afin d'éviter toute situation délicate par la suite.

### 2.4 Échéancier des travaux

Les dates de remise des travaux seront indiquées sur les énoncés.

### 2.5 Utilisation d'appareils électroniques et du courriel

Selon le règlement complémentaire des études, section 4.2.3 [4], l'utilisation d'ordinateurs, de cellulaires ou de tablettes pendant une prestation est interdite à condition que leur usage soit explicitement permise dans le plan de cours.

Dans ce cours, l'usage de téléphones cellulaires, de tablettes ou d'ordinateurs est autorisées. Cette permission peut être retirée en tout temps si leur usage entraîne des abus.

Tel qu'indiqué dans le règlement universitaire des études, section 4.2.3[5], toute utilisation d'appareils de captation de la voix ou de l'image exige la permission de la personne enseignante.

**Note :** L'utilisation du courriel est recommandée pour poser vos questions.

Un délai de plus de 24 heures est possible pour que je réponde aux questions posées par courriel. Il est fortement conseillé de profiter des périodes de consultations pour poser des questions. La consultation peut se faire à distance, par exemple en utilisant Teams.

## 3 Matériel nécessaire pour l'activité pédagogique

Il est fortement recommandé que chaque étudiant ou étudiante ait son propre ordinateur ayant une bonne puissance de calcul et une bonne quantité de mémoire RAM (64 Gbytes).

## 4 Références


[1] Bouguessa, Mohamed and Wang, Shengrui : Mining Projected Clusters in High-Dimensional Spaces. *IEEE Transactions on Knowledge and Data Engineering*, 21(4):507–522, apr 2009.

[2] Brin, Sergey and Page, Lawrence : The Anatomy of a Large-scale Hypertextual Web Search Engine. *Computer networks and ISDN Systems*, 30(1-7):107–117, apr 1998.

[3] Desrosiers, Christianand Karypis, George : *A Comprehensive Survey of Neighborhood-based Recommendation Methods*, pages 107–144. Springer US, Boston, MA, 2011.

[4] Kleinberg, Jon M. : Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, sep 1999.

[5] Shengrui Wang : Acétates du cours et des documents supplémentaires . Répertoire public du DI, 2019.

[6] Tan, Pang-Ning and Steinbach, Michael and Kumar, Vipin : *Introduction to Data Mining, (Second Edition)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2018. Ce livre contient plus que 65% de la matière du cours (thèmes : 1, 2, 3 et 4). Voir le site web du livre`http://www-users.cs.umn.edu/~kumar/dmbook/index.php`.

[7] Wu, Shu and Wang, Shengrui : Information-theoretic outlier detection for large-scale categorical data. *IEEE transactions on knowledge and data engineering*, 25(3):589–602, 2013.


[3] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

[4] https://www.usherbrooke.ca/sciences/fileadmin/sites/sciences/documents/Intranet/Informations_academiques/Sciences_Reglement_complementaire_2017-05-09.pdf

[5] https://www.usherbrooke.ca/registraire/droits-et-responsabilites/reglement-des-etudes/

UNIVERSITÉ DE SHERBROOKE

## L'intégrité intellectuelle passe, notamment, par la reconnaissance des sources utilisées.

## À l'Université de Sherbrooke, on y veille!

**Extrait du *Règlement des études* (Règlement 2575-009)**

**9.4.1 DÉLITS RELATIFS AUX ÉTUDES**

Un délit relatif aux études désigne tout acte trompeur ou toute tentative de commettre un tel acte, quant au rendement scolaire ou une exigence relative à une activité pédagogique, à un programme ou à un parcours libre.

Sont notamment considérés comme un délit relatif aux études les faits suivants :

a) commettre un plagiat, soit faire passer ou tenter de faire passer pour sien, dans une production évaluée, le travail d'une autre personne ou des passages ou des idées tirés de l'œuvre d'autrui (ce qui inclut notamment le fait de ne pas indiquer la source d'une production, d'un passage ou d'une idée tirée de l'œuvre d'autrui);

b) commettre un autoplagiat, soit soumettre, sans autorisation préalable, une même production, en tout ou en partie, à plus d'une activité pédagogique ou dans une même activité pédagogique (notamment en cas de reprise);

c) usurper l'identité d'une autre personne ou procéder à une substitution de personne lors d'une production évaluée ou de toute autre prestation obligatoire;

d) fournir ou obtenir toute aide non autorisée, qu'elle soit collective ou individuelle, pour une production faisant l'objet d'une évaluation;

e) obtenir par vol ou toute autre manœuvre frauduleuse, posséder ou utiliser du matériel de toute forme (incluant le numérique) non autorisé avant ou pendant une production faisant l'objet d'une évaluation;

f) copier, contrefaire ou falsifier un document pour l'évaluation d'une activité pédagogique;

[…]

**Par plagiat, on entend notamment :**

- Copier intégralement une phrase ou un passage d'un livre, d'un article de journal ou de revue, d'une page Web ou de tout autre document en omettant d'en mentionner la source ou de le mettre entre guillemets;
- reproduire des présentations, des dessins, des photographies, des graphiques, des données… sans en préciser la provenance et, dans certains cas, sans en avoir obtenu la permission de reproduire;
- utiliser, en tout ou en partie, du matériel sonore, graphique ou visuel, des pages Internet, du code de programme informatique ou des éléments de logiciel, des données ou résultats d'expérimentation ou toute autre information en provenance d'autrui en le faisant passer pour sien ou sans en citer les sources;
- résumer ou paraphraser l'idée d'un auteur sans en indiquer la source;
- traduire en partie ou en totalité un texte en omettant d'en mentionner la source ou de le mettre entre guillemets ;
- utiliser le travail d'un autre et le présenter comme sien (et ce, même si cette personne a donné son accord);
- acheter un travail sur le Web ou ailleurs et le faire passer pour sien;
- utiliser sans autorisation le même travail pour deux activités différentes (autoplagiat).

## Autrement dit : mentionnez vos sources

Document informatif V.3 (août 2017)