

**SIO-6003**

Techniques de forage de données

**SECTION Z1****Hiver 2011****Enseignant: Alexandru Toma, PhD, MBA****Horaire du cours:**• **Eliminate (plate-forme d'enseignement à distance) :**

Tous les 12:30 à

Mardis 14:00

Du 11 Janvier 2011, au

17 Avril, 2011

**Site web du cours:** <http://www.webct.ulaval.ca>


---

**Information générale**


---

Enseignant: [Alexandru.Toma@sio.ulaval.ca](mailto:Alexandru.Toma@sio.ulaval.ca)

Alexandru Toma Après avoir obtenu l'accès à WebCT, toute communication avec l'enseignant doit se faire par WebCT (email and forums)

Support technique: Comptoir d'aide APTI (CAA)  
Pavillon Palasis-Prince, Room 2215[caa@fsa.ulaval.ca](mailto:caa@fsa.ulaval.ca)

☎ 418-656-2131 ext. 6258

<http://www.fsa.ulaval.ca/azimut/>**Horaire : automne et hiver**

|          |                |
|----------|----------------|
| Lundi    | 08:15 to 21:45 |
| Mardi    | 08:15 to 21:45 |
| Mercredi | 08:15 to 21:45 |
| Jeudi    | 08:15 to 21:45 |
| Vendredi | 08:15 to 21:45 |
| Samedi   | 09:00 to 16:00 |
| Dimanche | 09:00 to 16:00 |

**Horaire: été**

|          |                |
|----------|----------------|
| Lundi    | 08:30 to 21:00 |
| Mardi    | 08:30 to 21:00 |
| Mercredi | 08:30 to 21:00 |
| Jeudi    | 08:30 to 21:00 |
| Vendredi | 08:30 to 21:00 |
| Samedi   | Fermé          |
| Dimanche | Fermé          |

## DESCRIPTION DU COURS

---

### Introduction

---

D'une manière générale, le forage de données (« Data Mining » dans la littérature anglophone) est défini comme un processus d'exploration de vastes ensembles de données dans le but de faire émerger les connaissances et les informations significatives qu'ils peuvent contenir.

Dans le monde des affaires électroniques, l'objet du forage de données est d'orienter l'action et la prise de décision, notamment à partir de l'étude systématique des bases de données opérationnelles de l'entreprise et du comportement des utilisateurs de ses sites web.

Le forage de données se situe ainsi au cœur de la problématique centrale d'acquisition et de fidélisation des clients car il permet de répondre à des questions telles que :

- Quels sont les comportements d'achat des clients (dont la connaissance est essentielle pour orienter les décisions commerciales et envisager une individualisation de la relation-client) ?
- Quel est l'impact des actions promotionnelles en fonction des profils des clients ?
- Peut-on identifier à l'avance les clients qui risquent de se révéler défaillants (par exemple, ceux qui ne renouvelleront pas leur abonnement à un service en ligne) ?
- Quels sont les comportements des visiteurs d'un site (par exemple, pour adapter l'interface et l'organisation du site aux besoins de navigation de l'utilisateur connus à partir de l'analyse de ses précédents comportements de visite) ?

Ce cours s'adresse principalement aux étudiants en systèmes d'information, et il sera plus utile à ceux qui ont déjà une expérience professionnelle. Il n'y a pas de pré-requis pour ce cours, cependant des connaissances de statistiques sont nécessaires et des connaissances en modélisation de données seront bien utiles. Le cours fait un grand usage de logiciels (Open Modelsphere, R) donc il est important d'être à l'aise avec l'utilisation des outils informatiques, y compris en mode ligne de commande (le cas de R). Ces logiciels seront introduits dans le cours, leur connaissance préalable n'est pas indispensable. Bien que la réalisation des travaux ne nécessite pas de programmer, une expérience de programmation en langages procédurales sera aussi utile.

**Remarque concernant la charge de travail :** ce cours universitaire de deuxième cycle exige en moyenne 12 heures de travail par semaine. Soyez donc bien conscients qu'il est essentiel pour votre apprentissage et pour la réussite du cours d'avoir du temps à y consacrer.

---

### Objectifs généraux

---

- Comprendre les relations entre modélisation dimensionnelle, l'entreposage de données (*data warehousing*) et le forage de données (*data mining*).
- Comprendre le processus de forage de données et être capable d'interagir avec des spécialistes techniques en forage de données.
- Acquérir de l'expérience dans l'utilisation d'un logiciel de forage de données.
- Comprendre le processus de développement et implantation d'un projet de forage de données.
- Comprendre le rôle du forage de données dans les organisations

---

## Objectifs spécifiques

---

- Comprendre des principes de base de modélisation dimensionnelle de données
- Comprendre le processus itératif de forage de données (CRISP-DM) dans toutes ses phases
- Identifier les principales tâches de forages de données (description, estimation, prévision, classification, *clustering* et association) et les principaux algorithmes (statistiques, k-plus proches voisins, arbres de décision, réseaux de neurones)
- S'initier à la pratique du forage de données par un projet de session en groupe de travail
- Se familiariser avec le logiciel R, outil « open source » d'analyse statistique, exploration et forage de données
- Acquérir des compétences dans le développement l'implantation de projets de forage de données

---

## Liens avec les buts et objectifs du programme

---

|   | Degré d'atteinte dans le cours | Méthode d'évaluation utilisée              |
|---|--------------------------------|--|
| 1. Résoudre des problèmes complexes en contexte d'incertitude.  | Amorcé                         | Projet en groupe, travaux individuels      |
| 2. Communiquer efficacement.  | En développement               | Forums, bureaux virtuels, projet en groupe |
| 3. Gérer des équipes de travail.  | Amorcé                         | Projet en groupe                           |
| 4. Reconnaître les principaux enjeux sur les scènes locales et internationales.   | Amorcé                         | Projet en groupe, forums                   |
| 5. Démontrer des aptitudes de leadership.   | En développement               | Projet en groupe, Forum                    |
| 6. Utiliser les technologies de l'information et de la communication dans la conception, le design, le développement et la gestion des organisations. | Intégré                        | Projet en groupe, exercices                |
| 7. Favoriser l'adoption d'un comportement socialement responsable.  | Amorcé                         | Projet en groupe                           |

---

## Approche pédagogique

---

Toutes les séances seront diffusées en simultané et enregistrées sur la plate-forme Elluminate (voir [www.bsp.ulaval.ca/classevirtuelle/aide/](http://www.bsp.ulaval.ca/classevirtuelle/aide/) pour plus d'informations sur Elluminate). Le contenu des présentations sera aussi disponible sur le site WebCT du cours, qui contiendra aussi l'énoncé des évaluations hebdomadaires et le projet en groupe de travail.

Après la séance introductive, la semaine type se déroulera de la façon suivante : l'étudiant lit le texte obligatoire (défini sur WebCT) et livre un travail hebdomadaire. La livraison du travail

hebdomadaire se fait le Dimanche soir, par l'intermédiaire des boîtes de dépôt (liens définis sur WebCT). La séance sur Elluminate commente ensuite les travaux, en insistant sur les points moins compris, fournit plus d'informations sur le contenu, répond aux questions et prépare la séance suivante. **Cette approche vous demandera un effort continu au long du trimestre.** Les étudiants peuvent en tout moment communiquer avec l'enseignant et entre eux par les outils de WebCT (forums et courriel, voir aussi le point Encadrement plus bas). Vers la fin du trimestre quelques semaines seront réservées pour le projet en équipe, qui est une synthèse des acquis au long du trimestre.

## Livre obligatoire et logiciels

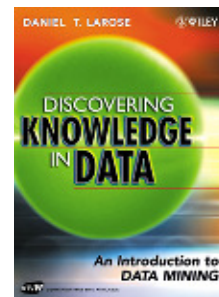
- **Livre obligatoire (Data mining):**

**Titre:** *Discovering Knowledge in Data-An Introduction to Data Mining*

**Auteur:** Daniel T. Larose

**Publisher:** Wiley Inter-Science, **Year:** 2005, **ISBN-**978-0-471-66657-8.

Le livre est aussi disponible en format électronique à l'adresse suivante: <http://arianeweb.ulaval.ca> (la bibliothèque de l'Université Laval)



- **Logiciel de modélisation en SIO (nous utiliserons la partie de modélisation de données) :** OpenModelSphere ( <http://www.modelsphere.org/> )
- **Logiciel d'analyse statistique et forage de données:**
  - R (<http://www.r-project.org/> )

## ENCADREMENT

Une rétroaction sur les travaux hebdomadaires sera présentée dans chaque séance. Pour le projet final, une rétroaction sera fournie suite au travail 10, qui est en fait un plan de ce que vous comptez faire pour ce projet. L'enseignant lira régulièrement les forums et les courriels, répondant sous 24 heures à vos messages (possiblement plus cependant en fin de semaine).

## CALENDRIER DU COURS

| Séance   | Contenu  | Date                                |
|--|--|-------------------------------------|
| S1: Introduction à l'intelligence d'affaires et au forage de données | Données, Information, Connaissances, bases de données opérationnelles et décisionnelles, intelligence d'affaires, présentation du plan de cours et de l'enseignant | Semaine 1 : pas de travail à rendre |
| S2: Généralités sur les  | Data Warehousing, Data marts, ETL, OLTP,   | Semaine 2 : travail 1 à             |

|   |  |   |
|---|--|---|
| entrepôts de données  | OLAP   | rendre le 16 Janvier  |
| S3: Modélisation dimensionnelle de données                            | Modèle en étoile, faits, dimensions  | Semaine 3 : travail 2 à rendre le 23 Janvier                      |
| S4: Introduction au forage des données et prétraitement des données   | Le cycle de vie du forage de données, tâches accomplies par le forage de données, exemples, nettoyage de données, traitement des données manquantes, identification des données aberrantes, transformation des données | Semaine 4 : travail 3 à rendre le 30 Janvier                      |
| S5: Introduction au logiciel R  | Introduire des tableaux de données dans R, aide en ligne, transformer des données, quelques types de données et syntaxe  | Semaine 5 : pas de travail à rendre                               |
| S6 : Exploration des données  | Exploration des variables qualitatives et quantitatives, traitement des variables corrélées  | Semaine 6 : travail 4 à rendre le 13 Février                      |
| S7: Exploration des données avec R                                    | Exploration des variables qualitatives et quantitatives avec R, analyse des corrélations avec R  | Semaine 7 : travail 5 à rendre le 20 Février                      |
| S8: Approches statistiques pour estimation et prévision               | Inférence statistique, intervalles de confiance, régression simple et multiple   | Semaine 8 : travail 6 à rendre le 27 Février                      |
| <b>Semaine de lecture (7-12 Mars)</b>                                 |  |   |
| S9: Algorithme des k plus proches voisins (KNN) et arbres de décision | Méthodes supervisées et non supervisées, fonction distance, choix du paramètre k, algorithmes CART et C4.5, règles de décision   | Semaine 9 : travail 7 à rendre le 13 Mars                         |
| S10: Réseaux de neurones  | Réseaux de neurones pour estimation et prévision, rétropropagation, descente de gradient   | Semaine 10 : travail 8 à rendre le 20 Mars                        |
| S11: Techniques d'évaluation des modèles                              | Évaluation des modèles pour description, estimation, prévision et classification   | Semaine 11 : travail 9 à rendre le 27 Mars                        |
| S12: Travail sur le projet final                                      | Feedback et commentaires sur la première version / plan  | Semaine 12 : travail 9 (plan du projet final) à rendre le 3 Avril |
| S13: Travail sur le projet final                                      | Travail sur le projet final  | Semaine 13 : pas de travail à rendre, ni séance                   |
| S14: Travail sur le projet final                                      | Version finale à livrer  | Semaine 12 : travail d'équipe à rendre le 17 Avril                |

## Évaluation et résultats

---

### Travaux

---

| Type d'évaluation                          | Poids       |
|--|-------------|
| 10 travaux hebdomadaires (7 points chacun) | 70%         |
| 1 travail d'équipe (final)                 | 30%         |
| <b>Total</b>                               | <b>100%</b> |

### Projet d'équipe

---

Le but de ce projet est d'appliquer les connaissances théoriques et les techniques acquises en classe sur un jeu de données réel. Chaque équipe doit travailler sur le jeu de données fourni ou choisir son propre jeu de données, sous réserve d'approbation par l'enseignant.

Les détails concernant le projet, incluant la formation des équipes, seront accessibles sur le site WebCT.

### Information détaillée sur les évaluations

---

#### Travaux (100%)

##### Description et instructions

Ces travaux visent à vérifier l'acquisition de vos connaissances ainsi que votre compétence à appliquer et à transférer les notions étudiées à des situations concrètes. Le français utilisé dans vos travaux d'évaluation doit être correct. Vous devez obligatoirement réaliser et retourner aux dates prévues (voir la section Évaluations du site du cours) les travaux notés. L'énoncé de chaque travail décrira le format de chaque travail, incluant le nombre de pages ou mots conseillés, la mise en page etc.

Les étudiants sont priés de porter une attention particulière à l'orthographe et à la clarté d'expression dans leurs travaux, cas et projet de session. Ces éléments feront partie de l'évaluation.

##### Critères d'évaluation

Un barème d'évaluation sera fourni pour chaque travail après la date de livraison, mais avant le cours suivant. Ce barème sera toutefois flexible dans le sens où vous obtiendrez souvent des points pour l'expression d'une idée générale même si elle est différente du barème. Pour chaque travail, jusqu'à 10% des points peuvent être enlevés en raison d'erreurs de grammaire ou manque de clarté dans l'expression des idées.

---

## Barème de conversion

---

| Intervalle   | Cote | Intervalle  | Cote      |
|--------------|------|-------------|-----------|
| [ 100 - 95 ] | A +  | ] 75 – 70 ] | B -       |
| ] 95 - 90 ]  | A    | ] 70 – 65 ] | C +       |
| ] 90 – 85 ]  | A -  | ] 65 – 60 ] | C         |
| ] 85 – 80 ]  | B +  | ] 60 – 0 ]  | E (Échec) |
| ] 80 – 75 ]  | B    |             |           |

---



---

## Résultats

---

- Vos résultats seront accessibles sur le site du cours.

---

## Plagiat

---

La FSA ne tolère pas les comportements non conformes à l'éthique. Le Règlement disciplinaire à l'intention des étudiants de l'Université Laval fait état de près d'une vingtaine d'infractions relatives aux études passibles de sanctions. Vous connaissez sûrement les fautes les plus courantes, mais saviez-vous que copier des phrases d'un ouvrage papier ou d'un site web sans mettre les guillemets ou sans mentionner la source constituent deux de ces infractions passibles de sanctions? Ou encore qu'il est interdit de résumer l'idée originale d'un auteur en l'exprimant dans ses propres mots sans en mentionner la source ou traduire partiellement ou totalement un texte sans en mentionner la provenance. Afin d'éviter de vous exposer à des conséquences allant de l'attribution d'un échec dans un cours au congédiement de l'Université, consultez le site Web suivant : [www.fsa.ulaval.ca/plagiat](http://www.fsa.ulaval.ca/plagiat). Vous y trouverez toute l'information utile pour prévenir le plagiat.

---

## Règles disciplinaires

---

Tout étudiant qui commet une infraction au Règlement disciplinaire à l'intention des étudiants de l'Université Laval dans le cadre du présent cours, notamment en matière de plagiat, est passible des sanctions qui sont prévues dans ce règlement. Il est très important pour tout étudiant de prendre connaissance des articles 28 à 32 du Règlement disciplinaire. Celui-ci peut être consulté à l'adresse suivante :

[http://www.ulaval.ca/sg/reg/Reglements/Reglement\\_disciplinaire.pdf](http://www.ulaval.ca/sg/reg/Reglements/Reglement_disciplinaire.pdf)

---

## Gestion des échéances et des retards

---

Le cheminement d'apprentissage proposé au calendrier doit être respecté dans la mesure du possible. **Pour un travail hebdomadaire, tout retard sera pénalisé de 5 % par heure de retard jusqu'à un maximum de 10 heures.** Après ce délai, le travail sera refusé et la note 0 sera

accordée à cette évaluation. Pour le travail d'équipe, tout retard sera pénalisé de 10% par jour jusqu'à un maximum de 5 jours.

Cependant, il est entendu que certaines circonstances exceptionnelles peuvent empêcher l'étudiant de remettre une évaluation dans les délais prescrits. Dans ce cas, il est de la responsabilité de l'étudiant d'en avertir l'enseignant le plus tôt possible afin de négocier une extension ou d'envisager des alternatives. Il est cependant entendu que ce privilège ne pourra être accordé plus d'une fois dans le cours.

---

## Évaluation du cours

---

À la 4e semaine, une évaluation formative du cours sera effectuée. Cette évaluation confidentielle sera uniquement consultée par votre enseignant afin de valider si la formule pédagogique est correcte et si certains ajustements s'imposent avant la fin de la session.

À la fin de ce cours, la Faculté procédera à l'évaluation sommative du cours afin de vérifier si la formule pédagogique a atteint ses buts et si vous êtes satisfait en recueillant vos commentaires et vos suggestions. Durant la session, un lien hypertexte sera ajouté sur la page d'accueil du site Web de cours. Ce lien vous mènera vers un questionnaire d'évaluation qui permettra d'améliorer ce cours. Cette dernière étape est très importante et les responsables du cours vous remercient à l'avance pour votre collaboration. Veuillez noter que cette évaluation est confidentielle.

---

## Quelques conseils pour réussir ce cours

---

- Familiarisez-vous avec le plan de cours, objectifs et approche pédagogique.
- Ce cours est un cours à distance. Vous devez faire vos travaux et planifier votre temps chaque semaine. Il est recommandé de planifier neuf à douze heures chaque semaine pour ce cours.
- L'approche pédagogique de ce cours est basée sur un effort constant de votre part au cours du trimestre. Vous allez devoir faire une lecture chaque semaine et préparer un travail montrant votre compréhension de la lecture. **L'évaluation est faite principalement (70%) selon ces travaux hebdomadaires, qui commencent déjà à la fin de la première semaine (le premier travail est à rendre le 16 Janvier).** Le but de cet effort continu est de vous garder en contact avec le cours, même s'il n'y a pas de rencontre en salle. Il est essentiel de comprendre cela quand vous décidez de suivre ce cours; il ne sera pas possible de retarder le travail vers la fin de la session.
- Identifiez les outils nécessaires pour les travaux et assurez-vous d'avoir accès à ces outils.
- Assurez-vous que votre ordinateur possède les capacités pour l'apprentissage en ligne.
- Une autre grande partie de l'évaluation (30%) est constituée par un travail d'équipe. Organisez-vous et assurez-vous de développer les compétences nécessaires au travail en équipe.

**Bon cours!**