Université Saint-Joseph de Beyrouth
جامعة القدّيس يوسف في بيروت

Faculté d'ingénierie **Faculté des sciences**

MASTER en Data Sciences

ماستر في علم المعلومات

MASTER in Data Science

Janvier 2018

Adresse: Université Saint-Joseph de Beyrouth
Campus des Sciences et Technologies
Mar Roukos - Dekwaneh – Liban
B.P. 1514 - Riad El Solh
Beyrouth 1107 2050

École supérieure d'ingénieurs de Beyrouth
Courriel : fi.esib@usj.edu.lb
Téléphone : +961 (1) 421317
Télécopie : +961 (4) 532645

Faculté des sciences
Courriel : fs@usj.edu.lb
Téléphone : +961 (1) 421 368
Télécopie : +961 (4) 532 657

1. Présentation

- **INFORMATIONS GÉNÉRALES**

L'économie mondiale est bouleversée par la furie du numérique qui déferle sur notre planète. Des géants de l'économie traditionnelle disparaissent et avec eux des métiers. Simultanément, d'autres métiers éclosent et prennent leur place.

La fonction du « Data Scientist » est actuellement la star des métiers émergents. Les différents acteurs de l'économie ont enfin apprécié la valeur de l'information présente dans ce volume énorme de données sur le WEB. Le marché national, régional et international est de plus en plus à la recherche d'experts dans le domaine de data sciences pour avoir un avantage compétitif et développer des produits innovants.

- **OBJECTIFS SCIENTIFIQUE ET PEDAGOGIQUE**

Le Master « Data Sciences » s'insère dans le cadre des formations professionnelles pour préparer des spécialistes capables de développer des stratégies d'analyse et de décisions basées sur les données massives.
Le programme est à caractère professionnel et répond aux besoins du marché du travail qui est à la recherche de spécialistes dans le domaine de l'analyse et du traitement de données.
Cette formation permet également aux étudiants qui le désirent, de préparer une thèse doctorale dans ce domaine.

Ce Master vise à former :

- des spécialistes de haut niveau capables de concevoir de nouveaux outils pour collecter les données massives et les traiter à l'aide d'algorithmes adéquats,
- des chercheurs experts en informatique, mathématiques appliquées et statistique,
- des concepteurs de systèmes de gestion de bases de données pouvant garantir la qualité, la sécurité ainsi que l'accessibilité des informations,
- des consultants multidisciplinaires capables de transformer les informations en outils d'aide à la décision au sein d'une entreprise.

C'est un diplôme universitaire au Liban, auquel des établissements réputés apportent leur collaboration et leurs moyens pédagogiques et scientifiques. Le programme fait l'objet d'une collaboration entre la Faculté d'ingénierie et la Faculté des sciences de l'Université Saint-Joseph de Beyrouth. Ces deux facultés agissent en commun, sous l'égide du ministère de l'éducation nationale et de l'enseignement supérieur, pour co-habiliter la formation de haut niveau distribuée dans le cadre de ce Master. Le contenu du programme de ce master a été validé par l'École Polytechnique de Paris. Une co-diplomation avec l'X est en cours de préparation.

- **ORGANISATION GENERALE DU MASTER**

Le Master « Data Sciences » se base sur les dernières découvertes dans le domaine du traitement des données massives et les met en application.

Le master comporte 120 crédits d'une durée de deux années (M1 et M2) pour un parcours normal, répartis sur 4 semestres S1, S2, S3 et S4.

C’est une formation comprenant :

- des enseignements théoriques et pratiques,
- un stage en entreprise ou un stage de recherche donnant lieu à la rédaction d’un mémoire et à une soutenance.

Un des objectifs principaux de ce programme est de former des professionnels de « terrain » dans le domaine du traitement des données massives, opérationnels dès leur sortie du Master ; c'est pourquoi une bonne partie de la formation est consacrée à l'aspect pratique par la mise en œuvre des thèmes abordés.

- **DEBOUCHÉS ET ÉTUDES DOCTORALES**

Le Master « Data Sciences » constitue un immense réservoir d’emplois pour les années à venir. Il présente un large éventail de débouchés dans les secteurs cherchant à extraire des informations pour développer des stratégies. Il s’agit :

- des banques pour développer de nouveaux produits,
- des agences de statistiques,
- des sociétés de ventes et de distributions,
- des experts en marketing,
- des sociétés de conseils,
- business developers,
- internet des objets (IoT),
- etc.

Ce programme permet également la préparation à la recherche. En effet, les étudiants ayant validé le Master avec succès pourront effectuer une thèse de doctorat.

2. Admission et inscription

- **ADMISSION**

Les admissions se font sur dossier et éventuellement suite à un entretien.

Admission en première année du programme de Master (M1)

Sont autorisés à déposer les dossiers de candidature :
- les titulaires d'une licence en Informatique ou en mathématiques,
- les titulaires d'un diplôme reconnu équivalent par la commission des équivalences de l'USJ.

Admission en deuxième année du programme de Master (M2)

Sont autorisés à déposer les dossiers de candidature :
- les titulaires d'une maîtrise ou d'un Master M1 en Informatique ou en mathématiques,
- les titulaires d'un diplôme d'ingénieur informatique et communications de l'ESIB,
- les titulaires d'un diplôme reconnu équivalent par la commission des équivalences de l'USJ.

La sélection des candidats est faite par un Comité Scientifique commun (de la Faculté d'ingénierie et de la Faculté des sciences) dans la limite des places disponibles. Le Comité Scientifique décidera pour chaque candidature les UE (Unités d'enseignement) validées en fonction du programme et des résultats préalablement obtenus.

Les documents requis lors du dépôt du dossier de candidature sont précisés dans le dossier d'admission commune propre à l'Université Saint-Joseph de Beyrouth.

Les dossiers seront examinés par le Comité Scientifique qui établira la liste des candidats admis à suivre cette formation. Les candidats retenus pourraient être soumis à un entretien avant leur admission finale. Le dossier de candidature est téléchargeable à partir du site de l'Université Saint-Joseph de Beyrouth et est à déposer dans l'une des deux institutions :

Faculté d'ingénierie de l'USJ

Mar Roukos, Mkalles

Tel : (01) 421317

ou

Faculté des sciences de l'USJ

Mar Roukos, Mkalles

Tel : (01) 421 368

3. Conditions et Diplôme

- **LANGUE D'ENSEIGNEMENT**

L'ensemble de cette formation sera dispensé en français et en anglais.

- **CONTROLE DES CONNAISSANCES**

Le diplôme de Master "Data Sciences" est délivré aux candidats qui ont passé avec succès les contrôles portant sur les enseignements théoriques et pratiques et qui justifient d'un niveau suffisant lors de la préparation et de la soutenance du mémoire. En principe, en cas d'absence, il n'est pas prévu de rattrapage des examens. En cas d'accident grave, dûment et sérieusement justifié, le cas sera examiné par le jury de fin d'année en vue de prendre les mesures jugées convenables.

- **PRÉSENCE**

Toutes les activités d'enseignement sont obligatoires. Des contrôles sont périodiquement effectués.

- **ÉVALUATION ET REUSSITE**

A chaque UE est affectée une note. À la suite des examens de chaque semestre, un jury est réuni et arrête les résultats.

A la fin du semestre S3, une moyenne générale est calculée à partir des notes des UE des trois premiers semestres (S1, S2 et S3), pondérées par le nombre de crédits. Les UE sont validées si les notes de toutes ces UE sont supérieures à dix.

Sont autorisés à présenter le rapport de stage ou le mémoire de recherche les étudiants qui ont validé les UE des trois premiers semestres. La priorité dans le choix des stages est fonction de la moyenne générale.

Le mémoire de recherche est validé si sa note est supérieure ou égale à 12/20.

- **DIPLÔME**

Le diplôme Master en "Data Sciences" est délivré aux étudiants admis ayant validé toutes les UE des 4 semestres S1, S2, S3 et S4 portant sur les enseignements et la soutenance de leur mémoire. Le système de notation est défini dans le règlement intérieur des études de l'Université Saint-Joseph de Beyrouth.

Compétences et résultats d'apprentissage

A l'issue de cette formation, les étudiants seront capable :

I. d'imaginer des utilisations nouvelles et précieuses pour les grands ensembles de données ;
II. d'appliquer des méthodes créatives et disciplinées pour poser des questions et interpréter les résultats ;
III. de récupérer, organiser, combiner, nettoyer, et stocker des données provenant de sources multiples ;
IV. d'appliquer l'apprentissage automatique ainsi que des techniques statistiques appropriées pour identifier les tendances et faire des prédictions ;
V. de visualiser et communiquer efficacement les résultats ;
VI. de comprendre les exigences éthiques et juridiques de la vie privée et la sécurité des données.

Compétence	**Résultats d'Apprentissage niveau Programme (RAP)**
a) Une capacité à appliquer des connaissances en mathématiques, en statistiques et en analyse de données	a1. Appliquer des connaissances en mathématiques pour résoudre des problèmes a2. Appliquer des connaissances en statistiques et en analyse de données pour résoudre des problèmes
b) Une capacité à concevoir et exécuter des expériences, à analyser et interpréter des données	b1. Planifier les expériences en appliquant les connaissances théoriques et en sélectionnant les données appropriées à mesurer b2. Exécuter des expériences en manipulant correctement les outils et en mesurant les données b3. Analyser les données en utilisant les outils appropriés, et interpréter les résultats
c) Une capacité à concevoir un système, composant, ou processus répondant aux besoins et respectant des contraintes réalistes d'ordre économique, environnemental, social, politique, éthique, sécuritaire	c1. Elaborer une stratégie de conception en analysant les besoins et en respectant les contraintes techniques et non techniques c2. Proposer une solution adaptée aux besoins, et la comparer aux solutions alternatives c3. Tester, améliorer et implémenter en utilisant les outils appropriés tels que la modélisation, le prototypage, et les tests de performance
d) Une capacité à identifier, formuler et résoudre des problèmes d'analyse de données	d1. Identifier un problème d'analyse de données en sélectionnant les informations utilisées et nécessaires d2. Formuler un problème en adoptant un modèle approprié d3. Résoudre un problème d'analyse de données en utilisant des outils appropriés et en appliquant les connaissances techniques
e) Une compréhension de la responsabilité professionnelle et éthique	e1. Décrire le code de conduite professionnel tel que la responsabilité envers les différents

	acteurs : clients, employés, administration, société, environnement e2. Comprendre les responsabilités juridiques et sécuritaires relatives au métier d'analyste de données
f) Une éducation nécessaire pour comprendre l'impact des outils d'analyse de données dans un contexte global, économique, environnemental et sociétal	f1. Décrire l'impact local et global de l'analyse de données sur les individus et la société, en identifiant les ressources pertinentes et en faisant un jugement informé
g) Une connaissance des sujets contemporains	g1. Citer des développements récents liés au domaine de l'analyse de données
h) Une capacité à utiliser les techniques, compétences, et outils modernes nécessaires pour analyser des données	h1. Utiliser les techniques nécessaires à la pratique professionnelle comme la conception, le prototypage et la simulation h2. Utiliser les compétences nécessaires à la pratique de la profession, telles que la programmation, la manipulation des outils

4. Programme prévisionnel

Le programme de ce Master est réparti sur 2 années d'études (M1 et M2).
Les UE sont réparties sur les semestres S1, S2, S3 et S4 suivant les tableaux ci-dessous.

Semestre 1 (S1)	**Type**	**Crédits**
Programmation orientée objet et Python pour Data Scientist	Obligatoire	**6**
Social Big Data & Langage R	Obligatoire	**6**
Cloud & Transformation digitale	Obligatoire	**4**
Programmation parallèle	Obligatoire	**4**
Deux cours optionnels parmi les cours qui suivent		**12**
Probabilité & Statistique[1]	Optionnel	6
Mathématiques pour Data Scientist[1]	Optionnel	*6*
Structures de données et algorithmes[2]	Optionnel	*6*
Base de Données relationnelles[2]	Optionnel	*6*
		32

Semestre 2 (S2)	**Type**	**Crédits**
Théorie des graphes et recherche opérationnelle	Obligatoire	**4**
Intégration des données	Obligatoire	**4**
Environnements distribués	Obligatoire	**4**
Les fondamentaux des données massives (Big Data)	Obligatoire	**4**
Modèles de régression	Obligatoire	**4**
Intelligence artificielle	Obligatoire	**4**
Exploitation de données (Data mining)	Obligatoire	**4**
		28

[1] Obligatoire pour les candidats qui n'ont pas fait une spécialisation en Mathématiques
[2] Obligatoire pour les candidats qui n'ont pas fait une spécialisation en Informatique

Semestre 3 (S3)	**Type**	**Crédits**
Theoretical guidelines for high-dimensional data analysis	Obligatoire	**4**
Big Data Frameworks	Obligatoire	**4**
Legal, Policy and Ethical considerations for data scientists	Obligatoire	**2**
Statistical learning theory	Obligatoire	**4**
		14

Semestre 4 (S4)	**Type**	**Crédits**
Web Mining	Obligatoire	**4**
Applied regression and time series analysis	Obligatoire	**4**
Machine learning, Deep learning and application to Natural language processing and Graphs	Obligatoire	**4**
Data visualization and communication	Obligatoire	**4**
		16

Stage d'été	**Type**	**Crédits**
Stage d'entreprise ou mémoire de recherche	Obligatoire	**30**

5. Contenu du programme

Le Master en data sciences est conçu pour préparer des data sciences leaders. Ce programme présente les meilleures pratiques pour la collecte, le stockage et la récupération des données - et l'influence de ces facteurs sur la rapidité, la précision et la fiabilité des modèles de stockage à grande échelle. Le programme prépare les étudiants à appliquer les dernières méthodes statistiques et informatiques pour identifier les modèles et l'extraction de connaissances à partir de données complexes et éventuellement prévoir certaines actions. Ils apprennent surtout à communiquer les résultats de l'analyse de données en utilisant efficacement les outils de visualisation et sont exposés à des dilemmes éthiques et les exigences juridiques liées à l'utilisation des données du monde réel.

Semestre S1 (32 crédits)

Programmation orientée objet et python (6 crédits)

Le but principal de cette unité d'enseignement étant de donner aux étudiants les outils nécessaires pour l'élaboration de programmes de niveau avancé en utilisant le concept d'objets dans leurs programmes. En effet cette approche de programmation offre une flexibilité et une portabilité exceptionnelles, ce qui rend cette UE essentielle pour des étudiants en formation mathématique. Cette UE vient en complément à une formation préalable en informatique spécialisée utilisant le langage C++.
D'autre part, Python est un langage de programmation orienté objet interprété. Outre les librairies standards, un grand nombre de paquetages (packages) développés par des contributeurs indépendants donne accès à des fonctionnalités spécialisées performantes. Ils nous donnent la possibilité de programmer des applications dans quasiment tous les secteurs de l'informatique. Nous nous intéresserons en particulier à la programmation statistique, au machine learning, au big data et au data science.

A l'issue de cette unité d'enseignement les étudiants seront capables de:

- Identifier et définir les différents éléments de bases pour établir un algorithme suivant le concept de la programmation utilisant des objets
- Décrire un problème pratique de programmation à travers des étapes logiques définissants les classes à utiliser
- Ecrire et Interpréter un algorithme relatif à une modélisation d'un phénomène donné
- Concevoir et Ecrire un programme en langage C++ utilisant des classes des objets et des membres publics et privés.
- Mise en œuvre des techniques statistiques inférentielles (tests d'adéquation, tests de normalité, tests de conformité à un standard, tests de comparaisons de populations, tests pour échantillons appariés, mesures d'association...) et exploratoires (essentiellement la classification automatique, k-means, classification ascendante hiérarchique) en Python.
- Econométrie, régression linéaire multiple, estimation des paramètres, moindres carrés ordinaires, tests de significativité, diagnostic, analyse des résidus, détection des points atypiques et influents, prédiction ponctuelle et par intervalle en Python.

Social Big Data & Langage R (6 crédits)

Les objectives de cette unité d'enseignement sont les suivants :

- Utiliser les différents services d'interface de programmation applicative (API) pour collecter des données provenant de différentes sources de médias sociaux.
- Utiliser différents outils pour collecter, analyser et explorer les données de médias sociaux à des fins de recherche et de développement.
- Exploiter les données tirés et analysé pour améliorer la présence et la stratégie sur les réseaux sociaux.

Cloud & Transformation digitale (4 crédits)

Le cloud computing et le big data sont à l'heure actuelle les deux principales évolutions technologiques et les principaux leviers de croissance pour les entreprises du secteur numérique.
Le big data, par le recueil et l'analyse de grandes masses de données, représente un potentiel d'activités nouvelles dans de nombreux secteurs d'activités. Le cloud computing permet un accès en tout lieu et « à la demande » aux services numériques, entrainant ainsi une réduction significative des dépenses. Ces deux sujets sont intimement liés : le cloud computing est la seule technologie à même de supporter l'informatique sous-jacente au bouleversement engendré par le big data.
Cette unité d'enseignement présente les solutions de Big Data basées sur le cloud, comme la plate-forme de Big Data d'AWS. Dans le cadre de ce cours, les étudiants vont découvrir comment utiliser les services clouds existants afin de traiter des données grâce au vaste écosystème d'outils. Les étudiants apprendront également à créer des environnements de Big Data et à appliquer au mieux les bonnes pratiques afin de concevoir des environnements de Big Data sécurisés et économiques.

Programmation parallèle (4 crédits)

Cette UE permet de compléter des connaissances en langage C par une formation approfondie sur les mécanismes d'accès au système d'exploitation. L'accent sera particulièrement sur les fichiers, les processus, l'allocation de mémoire, la synchronisation, la communication, le parallélisme et les librairies systèmes.
Cette UE permet à l'étudiant de maitriser tous les aspects de la programmation applicative et d'apprendre les outils de base, la gestion du système de fichiers, de la mémoire, des processus, des threads, des sockets ainsi que la programmation parallèle en utilisant la librairie MPI.
Elle couvre les thèmes suivants :
Gestion des processus: Création des processus, Héritage du processus fils, Terminaison d'un processus, Synchronisation père/fils, Recouvrement
Les signaux: Envoi d'un signal, Signaux délivrés, Traitement des signaux, Masquage des signaux
Gestion des fichiers: Opérations, redirections
Tubes de communication: Tube anonyme, Tube nommé, Le balayage
IPC du system V : Files de messages, Segments de Mémoire partagée, Les sémaphores
Les Sockets: Création d'un socket, Attachement d'un socket, Communication en mode non connecté, Communication en mode connecté.
Introduction au parallélisme : Classification des architectures parallèles, Architecture des machines parallèles à mémoire partagée, Architecture des machines parallèles à mémoire distribuée, Réseaux d'interconnexion, Les communications dans les réseaux statiques.

La librairie MPI : Initialisation de l'environnement d'un programme MPI, Communicateurs, Domaines de communication, Communication Point à Point, Communications collectives.

Probabilité & Statistique (6 crédits)

La théorie des probabilités est conçue comme un modèle mathématique pour le «hasard». Elle définit un cadre d'études adéquat pour les phénomènes aléatoires. Dans cette unité d'enseignement, il est question de mener une étude assez complète sur les variables aléatoires, laquelle débouche sur les grands théorèmes de convergence qui sont abordés en détails.
L'étudiant ayant suivi cette unité d'enseignement est en mesure de manipuler les variables aléatoires en déterminant leur loi, en calculant leurs moments, en analysant leur indépendance... Il pourra aussi étudier la convergence d'une suite de variables aléatoires et appliquer la loi forte des grands nombres et le théorème de limite central.
D'autre part, l'inférence statistique consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Ainsi, l'objectif du statisticien est symétrique de celui du probabiliste.
A la sortie de cette unité d'enseignement, l'étudiant est capable de mener une étude statistique complète : du choix du modèle statistique, à l'estimation de quantités inconnues, à la prise de décision.

Mathématiques pour Data Scientist (6 crédits)

Cette unité d'enseignement est un rappel des notions mathématiques utiles pour un Data Scientist. Elle couvre les thèmes suivants :

- Calcul différentiel pour une ou plusieurs variables
  - Rappels d'algèbre linéaire
  - La dérivée d'une fonction
  - Dérivée partielle d'une fonction multi-dimensionnelle
  - Lien dérivées partielles et gradient
  - Dérivées partielles d'ordre deux
  - Approximation quadratique
  - Les fonctions quadratiques en dimension deux
- Convexité et fonction convexe
  - Ensembles convexes
  - Fonction convexe
  - Convexité de l'épigraphe
  - Inégalité de convexité
  - Fonctions convexes réelles
  - Fonctions convexes multi-dimensionnelles
- Optimisation sans contrainte
  - Exemples de problèmes d'optimisation en apprentissage / statisiques
  - L'optimisation : minimisation
  - Optimisation, apprentissage et statistique
  - Résolution de systèmes matricielle
  - Condition d'existence d'un minimum
  - Minimum local, minimum global
  - Convexité et minimum
  - Condition du premier ordre pour un minimum local
  - La moyenne comme problème d'optimisation

- Optimisation avec contrainte
  - Exemples de problèmes avec contraintes
  - Condition d'existence d'un minimum
  - Condition du premier ordre
  - Projection sur les convexes fermés
  - Optimisation avec contraintes et Lagrangien
  - Conditions de Karush-Khunn-Tucker (KKT)
  - Conditions de Slater
- Algorithme pour l'optimisation sans contrainte
  - La descente de gradient
  - Recherche linéaire
  - Détour par la méthode de Newton
  - Méthode de Newton pour la minimisation

Structures de données et algorithmes (4 crédits)

Le cours de structures de données et algorithmiques a pour objectif de développer la capacité de résolution de problèmes chez l'étudiant en lui présentant l'ensemble des structures de données usuelles et les algorithmes relatifs. Ce cours traite les structures de données élémentaires (Listes chaînées, Tableaux, Files et Piles), les problèmes de recherche (séquentielle, dichotomie), les problèmes de tris (tris élémentaires, tri rapide, tri par fusion), les arbres (caractéristiques, structure, parcours), les algorithmes de recherche sur les chaines de caractères, les files de priorité, le tri maximier, la récursivité et la programmation dynamique.

Bases de données relationnelles (4 crédits)

Cette unité d'enseignement vise à initier les étudiants à la conception, à la création et à la gestion de Base de Données. Il permet aux étudiants de :
- maîtriser le concept « Base de Données »,
- concevoir une Base de Données à partir d'un Système d'Information (SI) donné,
- comprendre le Modèle Relationnel,
- savoir créer et gérer une Base de Données en utilisant le langage SQL,
- comprendre les techniques des systèmes de gestion de bases de données.

Semestre S2 (28 crédits)

Théorie des graphes et recherche opérationnelle (4 crédits)

Cette unité d'enseignement introduit la théorie des graphes et la recherche opérationnelle comme des outils de modélisation et de prise de décision pour l'ingénieur.
A l'issue de cette unité d'enseignement les étudiants seront capables :
- de faire une représentation mathématique et informatique des graphes,
- d'appliquer les algorithmes de parcours des graphes,
- de savoir calculer le plus court chemin,
- de savoir maximiser un problème de flot,
- d'appliquer les graphes à la gestion des projets,
- de comprendre l'algorithme du Simplexe et la programmation linéaire.

Intégration des données (4 crédits)

Cette unité d'enseignement détaille les spécifications XML, tel que les Namespaces, la validation avec une DTD, la validation avec un XMLSchema, XPATH, XSLT, XQUery, les parsers XML JAXP (SAX, DOM), Les pipelines d'intégration, les spécifications relatives au web sémantique RDF, OWL, SPARQL.
Cette unité d'enseignement constitue la brique de base qui traite la variété des données dans un contexte Data warehouse ou Data Lake.

Environnements distribués (4 crédits)

Initier les étudiants à la notion de middleware et leur apprendre à utiliser les architectures tels que CORBA de l'OMG, Java RMI, les composantes distribuées (Enterprise Java Beans) et les services web en vue d'implémenter une solution distribuée.

Les fondamentaux des données massives (Big Data) (4 crédits)

Cette unité d'enseignement couvre les principes fondamentaux de conception de systèmes de logiciels dédiés pour les traitements analytiques des grosses données.
Le cours débute par les principes de conception des systèmes de base de données relationnelle pour l'analyse des données des entreprises, y compris les requêtes déclaratives, l'optimisation des requêtes et gestion des transactions, ainsi que l'évolution des systèmes de base de données à l'appui des problèmes analytiques complexes et de la gestion des données scientifiques.
Le cours se penche ensuite sur les modifications architecturales fondamentales à l'échelle du traitement des données au-delà de la limite d'un seul ordinateur, y compris des bases de données parallèles, « MapReduce », stockage colonne et valeur clé distribuée, et de permettre aussi le calcul des résultats analytiques de faible latence à partir des flux de données en temps réel. Enfin, ce cours examine des systèmes de gestion avancée des données pour soutenir des modèles de diverses données y compris la structure arborescente (XML et JSON) et graphique structuré de données (RDF) et de nouvelles charges de travail tels que les tâches d'apprentissage automatique (Spark) et les charges de travail mixtes (flux de données Google Cloud).

Modèles de régression (4 crédits)

L'objectif de ce cours est d'introduire la régression linéaire et non linéaire (régression logistique et modèles linéaires généralisés). Les méthodes de régression jouent un rôle clé dans de nombreux problèmes et il est absolument essentiel pour un analyste de données de comprendre la théorie et la pratique de l'analyse par régression. C'est également une approche importante face aux défis statistiques : sélection de modèle, pénalisation, robustesse rééchantillonnage (bootstrap, validation croisée), détection des valeurs aberrantes et évaluation des écarts par rapport à un modèle hypothétique. Il s'agira aussi d'affiner la compréhension des techniques statistiques, notamment les tests et les estimations.

Intelligence artificielle (4 crédits)

Étude des agents intelligents : résolution de problèmes, algorithmes de recherches en longueur et en largeur, programmation des jeux : minimax, exptimax, savoir et raisonnement, planification, apprentissage, traitement du langage naturel, vision, robotique, les mécanismes d'inférence, les réseaux de Bayes, les processus de markov, le « Reinforcement learning » et leurs algorithmes : TD et Q.

Contenu :

- Communication: Traitement du langage naturel, vision
- Apprentissage renforcé
- Agent intelligents
- Incertitude, savoir et raisonnement
- Apprentissage: Bases de connaissance
- Apprentissage par observation
- Planification, recherche et programmation des jeux
- Résolution de problèmes
- Raisonnement : Logique de premier ordre
- Prise de décision

Exploitation des données (Data mining) (4 crédits)

Étude des algorithmes de Data Mining et paradigmes informatiques qui permettent aux ordinateurs de trouver des modèles et des régularités dans les bases de données, effectuer la prédiction et la prévision et généralement améliorer leur performance grâce à l'interaction avec les données.
Le cours portera sur le processus d'exploration de données en se basant sur des exemples de données volumineuses. Les technologies importantes d'usage dans ce cas seront également abordées ; data warehousing et traitement analytique en ligne (OLAP).
Les objectifs principaux du cours sont :

- Présenter aux étudiants les concepts de base et les techniques d'exploration de données.
- Développer leurs compétences en utilisant des logiciels récents de data mining pour résoudre des problèmes pratiques.
- Développer des algorithmes d'exploration de données pratique pour les données volumineuses
- Acquérir une expérience de recherche et d'étude indépendante.

Semestre S3 (30 crédits)

Theoretical guidelines for high-dimensional data analysis (4 credits)

Le but de ce cours est de fournir aux étudiants une introduction aux différents types de méthodes de recherche quantitative et des techniques statistiques pour analyser les données. Nous commençons avec un accent sur la mesure, inférence statistique et l'inférence causale. Ensuite, nous allons explorer une gamme de techniques et de méthodes statistiques en utilisant le langage des statistiques open-source, (R ou Python). Nous allons utiliser des techniques différentes pour l'analyse et la visualisation de données, avec un accent sur l'application de ces connaissances à des problèmes de données du monde réel. Les techniques incluses sont: statistiques descriptives et déductives, échantillonnage, la conception expérimentale, tests paramétriques et non paramétriques de la différence, régression des moindres carrés, et de régression logistique.

Web mining (4 crédits)

Ce cours est divisé en trois parties :

- L’objectif principal de la première partie est de sensibiliser les étudiants à la puissance des données textuelles de grande quantité et aux méthodes de calcul pour trouver des modèles au sein de textes volumineux. Cette partie présentera les applications des technologies de regroupement de texte dans :
  - L’organisation de l’information et l’accès
  - L’intelligence d’affaires (« Business intelligence »)
  - L’analyse du comportement social
  - « Digital humanities »
- L'objectif de la deuxième partie de ce cours ce cours est de présenter une information aux modèles graphiques probabilistes pour le traitement d'information non-structurées et d’offrir une introduction pratique aux activités d’exploration graphique évolutive. Il présentera les principales classes de modèles (dirigés, non dirigés), ainsi que les principaux algorithmes d'inférence exacte et approchée, en s'appuyant sur des applications concrètes : analyse linguistique, traduction automatique, catégorisation et clustering de documents, analyse d'opinions, modélisation graphique, analyse de centralité, détection de communauté, partitionnement, visualisation, compression, exploration de sous-graphe et plusieurs autres problèmes d’optimisation.
- 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

Big Data Frameworks (4 crédits)

Conceptuellement, le cours est divisé en deux parties.

- La première couvre les concepts fondamentaux de l'informatique parallèle MapReduce, à travers les yeux de Hadoop, MrJob et Spark, tout en plongeant profondément dans Spark, des trames de données, le Spark Shell, Spark Streaming, Spark SQL, MLlib. Les élèves utiliseront MapReduce pour les applications et les déploiements industriels pour différents domaines, y compris la publicité, la finance, la santé et les moteurs de recherche.
- La deuxième partie se concentre sur la conception algorithmique et le développement dans les environnements informatiques parallèles (Spark), le développement d'algorithmes (arbre de décision apprentissage), des algorithmes de traitement graphique (pagerank / court chemin), les algorithmes de newton, les support vector machines.

Legal, Policy, and Ethical Considerations for Data Scientists (2 credits)

Ce cours offre une introduction à la morale, la politique, et les implications éthiques de données. Le cours examinera l’aspect juridique, politique, et les questions éthiques qui se posent tout au long du cycle de vie complet de la science des données de la collecte, au stockage, le traitement, l'analyse et l'utilisation, y compris, la vie privée, la surveillance, la sécurité, la classification et la discrimination. Des études de cas seront utilisées pour explorer ces questions dans divers domaines tels que la justice pénale, la sécurité nationale, la santé, le marketing, la politique, l'éducation, l'automobile, l'emploi, l'athlétisme, et le développement. Une attention particulière sera accordée à des

contraintes et des considérations juridiques et politiques qui se fixent à des domaines spécifiques.

Applied Regression and Time Series Analysis (4 crédits)

Ce cours introduit l'étudiant aux sujets suivants : Les techniques de visualisation pour les données de séries chronologiques / concepts clés en probabilité et statistique mathématique / modèles de régression linéaire classique / transformation de variable / Spécification du modèle / inférence causale / estimation des variables / autorégressifs (AR) modèles Instrumental / moyenne de moyenne mobile / Moyenne mobile autorégressive (ARMA) / Autorégressive moyenne intégrée (ARIMA) / (GARCH) modèles / autorégression vectorielle (VAR) / prévision statistique / régression avec des données de séries chronologiques

Machine Learning, Deep learning and application to Natural language processing and Graphs (4 crédits)

Ce cours va au-delà de la phase de collecte de gros volumes de données en mettant l'accent sur la façon dont les algorithmes d'apprentissage machine learning peuvent être réécrites et étendu à l'échelle de travailler sur pétaoctets de données, à la fois structurées et non structurées, pour générer des modèles sophistiqués utilisés pour faire des prédictions. Conceptuellement, le cours est divisé en deux parties.

- La première partie porte sur l'apprentissage automatique qui est une discipline scientifique qui traite la conception et le développement d'algorithmes permettant aux comportements des ordinateurs d'évoluer en se basant sur des données empiriques, tels que des bases de données ou des données d'un capteur. Un axe majeur de recherche en apprentissage machine est de rendre la machine capable de reconnaître et apprendre des motifs complexes et de prendre des décisions intelligentes basées sur les données captées ; la difficulté réside dans le fait que l'ensemble de tous les comportements possibles compte tenu de toutes les entrées possibles est trop complexe pour le décrire en utilisant des langages de programmation. Cette partie mettra l'accent sur la compréhension des concepts importants en apprentissage machine et présentera les principaux paradigmes et les méthodes qui forment la base de l'apprentissage de la machine moderne. Cela implique l'étude spécifique des algorithmes d'apprentissage ainsi que l'expérimentation empirique des algorithmes.
- La deuxième partie porte sur l'apprentissage profond (« deep learning ») ainsi que sur les architectures clés de réseau incluant : les réseaux de neurones convolutifs, les autoencodeurs, les réseaux neuronaux récurrents, les réseaux de longue mémoire à court terme « LSTM ». Cette partie couvre également les réseaux stochastiques, les champs aléatoires conditionnels, les machines de Boltzmann, les modèles stochastiques et déterministes mixtes ainsi que l'apprentissage par renforcement profond.
- La troisième partie porte sur le traitement des langues naturelles : En effet, la recherche en traitement automatique des langues naturelles (TALN) est un domaine de l'intelligence artificielle visant le développement de techniques automatisées pour la manipulation de données langagière, sous une forme textuelle ou sonore. Les applications immédiates de ces techniques incluent le développement d'interfaces textuelles plus naturelles, la traduction automatique de documents, la détection de

pourriels, la recherche d'information dans une collection de documents à partir de requêtes, les systèmes de questions/réponses, et plusieurs autres. Cette partie introduit l'étudiant aux sujets suivants : Initiation à la problématique du traitement automatique de la langue naturelle et de ses applications. La langue naturelle par rapport aux langages formels : le problème de l'ambiguïté. Survol des théories linguistiques actuelles. Analyse et synthèse de la parole. Analyse morphologique : structure du dictionnaire et analyse suffixale. Analyse syntaxique : analyseur ATN, grammaires d'unification et représentation de la sémantique des langues naturelles : logique formelle et cadres. Interprétation sémantique. Connaissance du monde et contexte d'élocution. Applications.

Data Visualization and Communication (4 crédits)

L'accès aux données est exponentiellement croissant alors que la capacité humaine à les gérer et à les comprendre reste constante. Communiquer clairement et efficacement sur les modèles que nous trouvons dans les données est une compétence clé pour un scientifique de données réussie. Ce cours introduit les concepts de base de la visualisation, l'analyse et la représentation visuelle des données. Ces concepts sont nécessaires pour créer des applications et des outils adaptés qui permettent de gérer et d'analyser les flux de données massives. Il porte sur la conception et la mise en œuvre de représentations visuelles et verbales complémentaires de motifs et de l'analyse afin de transmettre les résultats, répondre aux questions, conduire des décisions, et de fournir des preuves convaincantes étayées par des données.

Statistical learning theory (4 crédits)

Le cours mettra l'accent sur la compréhension des concepts importants en apprentissage machine et présentera les principaux paradigmes et les méthodes qui forment la base de l'apprentissage de la machine moderne. Cela implique la compréhension de la formulation mathématique de l'apprentissage statistique et des bases de la théorie de l'apprentissage statistique et informatique.

Semestre S4 (30 crédits)

Stage professionnel ou mémoire de recherche (30 crédits)

Durant le semestre 4, les étudiants doivent effectuer un stage professionnel dans une entreprise ou un travail de recherche dans un laboratoire pour une durée de 4 mois sur un thème lié à la Data Sciences.

- Un étudiant aura le choix entre :
  - Une mission en entreprise d'une durée de 3 à 4 mois, dans une entreprise sur un thème lié à la Data Science, conclue par la rédaction et la soutenance d'un rapport professionnel.
  - Un sujet de recherche d'une durée de 3 à 4 mois dans un laboratoire reconnu par le comité scientifique, conclue par la rédaction et la soutenance d'un mémoire de recherche.

- Les stages se dérouleront dans les entreprises au Liban ou à l'étranger. La responsabilité scientifique du stage est assurée conjointement par l'entreprise et un enseignant de l'USJ ou d'une université partenaire. Ce stage, d'une durée minimale d'un semestre, a pour objectif de développer chez l'étudiant l'ensemble des compétences nécessaires à un spécialiste :
  - recherche bibliographique,
  - étude de l'état de l'art,
  - proposition et implémentation des solutions.

- Les travaux de recherche se dérouleront dans un laboratoire soit au Liban soit dans un établissement extérieur. La responsabilité scientifique de ces travaux de recherche est assurée par le ou les enseignants - chercheurs qui les dirigent.
  Ce travail, d'une durée minimale d'un semestre, a pour objectif de développer chez l'étudiant l'ensemble des compétences nécessaires pour effectuer un travail de recherche :
  - recherche bibliographique,
  - analyse critique de l'état de l'art,
  - propositions et implémentations des solutions,
  - propositions et débouchés sur des travaux de thèse.

- Le stage ou le travail de recherche fait l'objet d'un rapport ou d'un mémoire écrit et d'une soutenance publique.
  Les étudiants qui ont validé les UE des semestres S1, S2 et S3 sont autorisés à présenter le rapport de stage ou le mémoire de recherche.
  Le mémoire ou le rapport comporte une partie bibliographique et une partie technique.
  L'évaluation du stage ou du travail de recherche tient compte de trois éléments :
  - évaluation de l'initiative scientifique du stagiaire,
  - évaluation du mémoire ou du rapport écrit,
  - évaluation de la soutenance orale.