MASTER IN DATA SCIENCE

Langue principale d'enseignement
Français ○ Anglais ☑ Arabe ○

Campus où le programme est proposé : CST

OBJECTIFS

Le Master in Data Science s'insère dans le cadre des formations professionnelles visant à préparer des spécialistes capables de développer des stratégies d'analyse et de prendre des 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.

COMPÉTENCES

Appliquer des connaissances en mathématiques, en statistiques et en analyse de données.
- Concevoir et exécuter des expériences, à analyser et interpréter des données.
- Concevoir un système, composant ou processus répondant aux besoins et respectant des contraintes réalistes d'ordre économique, environnemental, social, politique, éthique, sécuritaire.
- Identifier, formuler et résoudre des problèmes d'analyse de données.
- Comprendre la responsabilité professionnelle et éthique.
- Comprendre l'impact des outils d'analyse de données dans un contexte global, économique, environnemental et sociétal.
- Avoir une connaissance des sujets contemporains.
- Utiliser les techniques, compétences, et outils modernes nécessaires pour analyser des données.

CONDITIONS D'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
- Les étudiants inscrits en cursus ingénieur informatique et communications – option Génie logiciel et ayant validé au moins 120 crédits
- 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 ou Faculté des sciences de l'USJ.**

UE/CRÉDITS ATTRIBUÉS PAR ÉQUIVALENCE

Les ingénieurs diplômés en génie informatique et/ou communications, les titulaires d'une Maîtrise ou d'un Master en informatique ou en technologies de l'information, les étudiants en cinquième année GIC à l'ESIB et les titulaires d'un diplôme équivalent reconnu, peuvent valider par équivalence un maximum de 60 crédits du programme : sur proposition du directeur du Département des études doctorales, le jury d'admission fixera pour chaque étudiant admis directement en M3, les matières et modules validés en fonction de son cursus et de ses résultats préalables et définira son parcours au master dans la spécialité concernée, incluant éventuellement des matières complémentaires pré-requises. La proposition de la validation de la formation antérieure est soumise à l'approbation de la commission des équivalences de l'USJ.

EXIGENCES DU PROGRAMME

UE obligatoires (114 crédits), UE optionnelles fermées (6 crédits)

UE obligatoires (114 crédits)
Applied Regression and Time Series Analysis (4 Cr.). Big Data Frameworks (4 Cr.). Cloud and Digital Transformation (6 Cr.). Data Visualization and Communication (4 Cr.). Enterprise Data Management (6 Cr.). Foundations of Decision Modeling (6 Cr.). Graph Theory and Operational Research (6 Cr.). Inferential Statistics (6 Cr.). Legal. Policy and Ethical Considerations for Data Scientists and AI (2 Cr.). Machine Learning and Deep Learning (6 Cr.). Master Thesis (30 Cr.). Mining Massive Data Set (6 Cr.). Natural Language Processing (4 Cr.). Programming for Data Science and Artificial Intelligence (6 Cr.). R Language (2 Cr.). Regression Models (4 Cr.). Social Big Data (4 Cr.). Theoretical Guidelines for High-Dimensional Data Analysis (4 Cr.). Web Mining (4 Cr.).

UE optionnelles fermées (6 crédits)
Une UE à choisir dans la liste suivante :
Mathematics for Data Science and AI (6 Cr.), Relational Database (6 Cr.).

PLAN D'ÉTUDES PROPOSÉ

Semestre 1

Code	Intitulé de l'UE	Crédits
020CTDIM1	Cloud and Digital Transformation	6
048DSTGM1	Graph Theory and Operational Research	6
048DSSIM1	Inferential Statistics	6
048DSPMM1	Programming for Data Science and Artificial Intelligence	6
	Elective course: Mathematics for Data Science and AI (020IAMAM1) or Relational Database (020BDREM1)	6
	Total	**30**

Semestre 2

Code	Intitulé de l'UE	Crédits
020INTDM2	Enterprise Data Management	6
020IADMM2	Foundations of Decision Modeling	6
020MLDLM2	Machine Learning and Deep Learning	6
020FOBDM2	Mining Massive Data Set	6
048DSLRM2	R Language	2
048MBCMM2	Regression Models	4
	Total	**30**

Semestre 3

Code	Intitulé de l'UE	Crédits
048DSARM3	Applied Regression and Time Series Analysis	4
020BDFRM3	Big Data Frameworks	4
020DVCOM3	Data Visualization and Communication	4
020IALPM3	Legal, Policy and Ethical Considerations for Data Scientists and AI	2
020IANLM3	Natural Language Processing	4
048DSSBM1	Social Big Data	4
048DSTGM3	Theoretical Guidelines for High-Dimensional Data Analysis	4
020WEMIM3	Web Mining	4
	Total	**30**

Semestre 4

Code	Intitulé de l'UE	Crédits
020STGEM4	Master Thesis	30
	Total	**30**

DESCRIPTIFS DES UE

Semester S1 (30 credits)

020CTDIM1 Cloud and Digital Transformation 6 Cr.

Cloud computing and big data are currently the two main technological developments driving companies' growth in the digital sector. Big data, achieved through the collection and analysis of large amounts of data, represents the potential for new activities in many sectors. Cloud computing allows anywhere and on-demand access to digital services, resulting in significant cost reductions. These two subjects are closely linked: cloud computing is the only technology capable of supporting the computation of problems defined by big data.
This course introduces cloud-based big data solutions, such as AWS's big data platform. Students will learn how to utilize existing cloud services to process data using the vast ecosystem of tools, how to create big data environments and apply the best practices to secure those environments in an economical approach.

048DSTGM1 Graph Theory and Operational Research 6 Cr.

This course introduces graph theory and operational research as modeling and decision-making tools for the data scientist. By the end of the course, students will be able to make mathematical and computer representations of graphs, apply graph traversal algorithms, calculate the shortest path, maximize flow problems, analyze complex networks, use the NetworkX Python library, use Markov chains to solve real-world problems, understand the Simplex algorithm and linear programming, and use numerical tools for solving optimization problems.

048DSSIM1 Inferential Statistics 6 Cr.

Statistical inference consists in predicting the unknown characteristics of a population based on a sample drawn from this population. Thus, the objective of statistics is symmetrical to that of probability. By the end of this course, students will be able to conduct a complete statistical study: spanning from selecting appropriate statistical models, to estimating unknown quantities and making informed decisions. The applications attributed during this course are led using the R language software mainly for data manipulation, implementing statistical procedures, plotting graphics and functions and presenting results in a comprehensible way.

020IAMAM1 Mathematics for AI & Machine Learning 6 Cr.

Artificial Intelligence has gained importance in the last decade with a lot depending on the development and integration of AI in our daily lives. The progress that AI has already made is astounding with the self-driving cars, medical diagnosis and even betting humans at strategy games like Go and Chess.
The future for AI holds tremendous promise, potentially leading to the creation of robotic companions. Consequently, many developers are now diving into AI and ML programing, recognizing its significance. However, mastering AI and ML algorithms demands a strong understanding of mathematics.
Mathematics plays an important role as it builds the foundation for programming for these two streams. This course will help students master the mathematical foundation required for writing programs and algorithms for AI and ML.
The course covers three main mathematical theories: Linear Algebra, Multivariate Calculus and Probability Theory.

048DSPOM1 Programming Languages for Data Science and Artificial Intelligence 6 Cr.

This course equips students with the necessary tools for developing advanced-level programs understanding the Object-Oriented Programming (OOP) approach. The first part of the course focuses on the C++ language while the second part delves into Python and its functionalities relevant to data science. In the final part, students are introduced to machine learning examples using Python, allowing exploration of the power of the libraries provided by the Python community.

020BDREM1 Relational Database 6 Cr.

This course introduces the design, creation and management of databases. It allows students to master the concept of "Database", designing a database for a given Information System (IS), understanding the Relational Model, acquiring skills in creating and managing a database using SQL language, and understanding the techniques of database management systems.

Semester S2 (30 credits)

020INTDM2 Enterprise Data Management 6 Cr.

"Enterprise Data Management (EDM) is the ability of an organization to precisely define, easily integrate and effectively retrieve data for both internal applications and external communication. EDM focuses on the creation of accurate, consistent, and transparent content."
This course addresses the challenges of enterprise data management at scale, primarily focusing on data architecture, data modeling and data integration, both on-premise and in the cloud. It covers different enterprise data architectures such as Data Warehouses, and Data Lakes. Additionally, it details various data models (including structured, semi-structured (XML), unstructured, and semantic data with RDF/OWL/SPARQL. The course also describes various NoSQL databases (key-value, column, document or graph-oriented databases), as well as various big data formats (Avro, ORC and Parquet). The course explains different data integration approaches: integration according to a materialized view (Data Warehouses/OLAP) and integration according to a virtual view (Mediators/GAV-LAV).
This course also covers Stream, and Batch processing using big data architectures such as Lambda architecture as well as integration and processing pipelines, using appropriate tools such as Talend Big Data Integration Studio, and Azure Data Factory.

020IADMM1 Foundation of Decision Modeling 6 Cr.

Preferences are present and pervasive in many situations involving human interaction and decisions. Preferences are explicitly or implicitly expressed in numerous applications and relevant decisions should be made based on these preferences. This course aims at introducing preference models for multicriteria decisions. It covers concepts and methods for preference modeling and multicriteria decision making, as well as the presentation of stochastic processes and estimators.

020MLDLM3 Machine Learning and Deep learning 6 Cr.

This course goes beyond the phase of collecting large volumes of data by focusing on how machine learning algorithms can be rewritten and extended to scale for petabytes of structured and unstructured data. Also sophisticated models for predictions are included. The course is divided into three main parts.
The first part deals with the design and development of algorithms allowing the behavior of computers to evolve based on empirical data, such as databases or sensory data. We also define supervised, unsupervised and reinforcement learning.
The second part introduces deep learning as well as key network architectures including: convolutional neural networks, autoencoders, recurrent neural networks, long-term short-term networks "LSTM". This part also covers deep reinforcement learning.
The third part deals with the processing of natural languages: Indeed, research in the automatic processing of natural languages is a field of artificial intelligence aiming at the development of automated techniques for the manipulation of language data, in textual or sound forms. Immediate applications include developing more natural textual interfaces, automatic document translation, spam detection, information retrieval, question-answering systems, among others. This part introduces the student to the following subjects: Introduction to the problem of automatic processing of natural language and its applications.

020FOBDM2 Mining Massive Data Set 6 Cr.

This course covers the fundamentals of designing dedicated software systems for big data analytics.
The course begins with principles of designing relational database systems for analyzing business data, including declarative queries, query optimization and transaction management. It also covers the evolution of basic data systems to support complex analytical problems and scientific data management.
The course then explores fundamental architectural changes necessary for processing data beyond the limits of a single computer. This includes parallel databases, "MapReduce", column storage, distributed key value, and enabling low-latency analytical results from real-time data streams. Finally, this course examines advanced data management systems supporting various data models including tree structure (XML and JSON), structured data graphics (RDF), new workloads such as machine learning tasks (Spark), and mixed workloads (such as Google Cloud Dataflow).

048DSLRM1 R Language 2 Cr.

This course introduces R language programming, its basic concepts and essential functionalities for data treatment.

048MBCMM2 Regression Models 4 Cr.

This course covers the fundamentals of regression, including linear regression, its approach, and its applications in practical studies. It also includes ANOVA techniques and logistic regression. The course alternates between theoretical presentations and computer exercises, utilizing the R language.

Semester S3 (30 credits)

048DSARM3 Applied Regression and Time Series Analysis 4 Cr.

This course introduces the following subjects: Visualization techniques for time series data, key concepts in probability and mathematical statistics, classical linear regression models, variable transformation, model specification, causal inference, variable estimation, autoregressive (AR) instrumental models, moving average, autoregressive moving average (ARMA), integrated average autoregressive (ARIMA), (GARCH) models, vector autoregression (VAR), statistical forecast, and regression with time series data.

020BDFRM3 Big Data Frameworks 4 Cr.

This course is conceptually divided into two parts.
The first part covers the fundamental concepts of MapReduce parallel computing, focusing on Hadoop, MrJob and Spark. It delves deeply into Spark, data frames, Spark Shell, Spark Streaming, Spark SQL, MLlib. Students use MapReduce for industrial applications and deployments across various fields, including advertising, finance, health, and search engines.
The second part focuses on algorithmic design and development in parallel computing environments (Spark). It covers algorithmic development (learning decision tree), graphics processing algorithms (such as PageRank and short path), Newton algorithms, and support vector machines.

020DVCOM3 Data Visualization and Communication 4 Cr.

Access to data is exponentially increasing while human capacity to manage and understand it remains constant. Communicating clearly and effectively about the models found in the data is a key skill for a successful data scientist. This course introduces basic concepts of visualization, analysis, and visual representation of data, necessary for the creation of suitable applications and tools that allow students to manage and analyze big data flows. It involves designing and implementing complementary visual and verbal representations of patterns and analyses to convey results, answer questions, drive decisions, and provide convincing evidence supported by data.

020IALPM3 Legal, Policy and Ethical Considerations for Data Scientists and AI 2 Cr.

This course introduces ethics, politics, and ethical implications of data, including personal data. It examines the legal, political, and ethical issues that arise throughout the entire lifecycle of the science of data collection, storage, processing, analysis and use, including, privacy, surveillance, security, classification and discrimination. Additionally, a brief introduction will be provided about law and Labor law in general. Case studies will be used to explore these issues in various areas such as criminal justice, national security, health, marketing, politics, education, automotive, employment, athletics, and development. Particular attention will be paid to legal and political constraints and considerations specific to each area.

020IANLM3 Natural Language Processing 4 Cr.

This course delves into advanced machine learning techniques, shifting focus from data collection to scaling machine learning algorithms for processing petabytes of both structured and unstructured data, enabling the creation of sophisticated predictive models. Conceptually, it is divided into two parts.
The first part deals with deep learning and key network architectures, such as convolutional neural networks, autoencoders, recurrent neural networks, and short-term long-term memory networks (LSTM). Additionally, it covers stochastic networks, conditional random fields, Boltzmann machines, stochastic and mixed deterministic models, and deep reinforcement learning.

The second part focuses on natural language processing (NLP), a field of artificial intelligence dedicated to automating linguistic data manipulation. Immediate applications include developing more natural textual interfaces, automatic document translation, spam detection, information retrieval, question-answering systems, among others. This part introduces students to various topics, including the problem of NLP and its applications, natural language ambiguity, linguistic theories, speech analysis and synthesis, morphological analysis (dictionary structure and suffix analysis), syntax analysis (ATN parser, unification grammars and representation of the semantics of natural language: formal logic and frameworks), semantic interpretation, knowledge of the world and speech context, and applications.

048DSSBM1 Social Big Data 4 Cr.

This course introduces the structures and data types found on social networks (such as Facebook, Twitter, Instagram, etc.). It covers various methods of data collection and analysis based on application areas under the R language. Students gain proficiency in utilizing different application programming interface services (API) to collect data, analyze and explore social media data for research and development purposes. Ultimately, students will use the data drawn and analyzed to improve their presence and strategy on social networks.

048DSTGM3 Theoretical Guidelines for High-Dimensional Data Analysis 4 Cr.

This course introduces the different types of quantitative research methods and statistical techniques for analyzing data. It starts with an emphasis on measurement, statistical inference and causal inference. Next, it explores a range of statistical techniques and methods using the language of open-source statistics (using R or Python). Different techniques for data analysis and visualization are introduced, with a focus on applying this knowledge to real-world data problems. The techniques included are descriptive and deductive statistics, sampling, experimental design, parametric and non-parametric difference tests, least squares regression, and logistic regression.

020WEMIM3 Web Mining 4 Cr.

This course is divided into 3 parts:

- The first part essentially aims to introduce students to the opportunities offered by the adequate extraction of information from large-scale textual data. It then delves into the Vector Space Model (VSM) and applies to it algorithms such as Term Frequency (TF) – Inverse Document Frequency (IDF) in the context of word association mining. It goes on to explore probabilistic topic models and delves into the implementation of algorithms such as Expectation-Maximization (EM) and Probabilistic Latent Semantic Analysis (PLSA). It also covers text clustering and categorization algorithms as well as opinion mining and sentiment analysis.
- The second part discusses the analysis of large graphs. It views the web as a directed graph and essentially explores link analysis and PageRank.
- The third and final part delves into recommender systems with a focus on content-based and collaborative filtering algorithms.

Semester S4 (30 credits)

020STGEM4 Master Thesis 30 Cr.

During the 4th semester, students must complete a professional project in a company or research work in a laboratory for 4 months.
The projects will take place in companies in Lebanon or abroad. The scientific responsibility for the project is provided jointly by the company and an instructor from USJ or a partner university. This project, of a minimum of one semester, aims to develop all the skills necessary in the data science field.
Student can also choose to contribute in an academic research project. It takes place in a laboratory either in Lebanon or in an external institution.
The project or research work is the subject of a dissertation and a public defense in the presence of USJ professors.
Students who have validated all the courses of the first year and the first semester of the second year of the Master's program are authorized to submit their project report or present their research dissertation.
The project or the report includes a bibliographic part and a technical part.
The evaluation of the project considers three elements:

- Evaluation of the trainee's scientific initiative.
- Evaluation of the report.
- Evaluation of the oral defense.