

MASTER IN DATA SCIENCES

Main Language of Instruction:

French ☐ English ☒ Arabic ☐

Campus Where the Program Is Offered: CST

OBJECTIVES

The Master in Data Science aims to train:

- Specialists who can design and use new tools for collecting massive data and process it using appropriate algorithms.
- Researchers with expertise in computer science, applied mathematics, and statistics.
- Designers of database management systems that can ensure the quality, security, and accessibility of information.
- Multidisciplinary consultants capable of transforming information into decision support tools within a company.

PROGRAM LEARNING OUTCOMES (COMPETENCIES)

1. Acquire advanced knowledge and methods in the discipline

- Apply key theoretical and practical concepts.
- Apply theoretical knowledge to practical problems.

2. Address critical issues and demonstrate expertise in key areas

- Identify and assess key challenges.
- Apply advanced mathematics and science to solve complex problems.
- Demonstrate expertise in specialized areas.

3. Apply new theoretical and experimental methods

- Learn and apply new methods and technologies.
- Use advanced analytical tools to solve complex problems.
- Integrate new technologies to improve system performance.

4. Communicate effectively at an advanced level

- Produce clear, concise, and well-structured technical reports.
- Present complex information clearly and effectively.

ADMISSION REQUIREMENTS

Admission is based on the application file and may require an interview.

- For applications to the first semester of the Master's program, students must hold a Bachelor in Computer and Communications Engineering, Computer Science, or Mathematics, or hold an equivalent degree recognized by USJ.
- For applications to the third semester of the Master's program, students must hold a Bachelor of Engineering in Computer and Communications Engineering, be a CCE student at ESIB and have earned at least 120 credits in the Engineering Cycle, or hold a Master in Computer Science, or Computer and Communications, Mathematics or an equivalent degree recognized by USJ.

COURSES/CREDITS GRANTED BY EQUIVALENCE

ESIB CCE graduates, holders of a Bachelor's or professional Master's degree in Computer Science or Mathematics, fifth-year ESIB CCE students, and holders of a recognized equivalent diploma, can validate up to 60 credits of the program by equivalence, depending on the courses they have already validated in their previous or current program, and depending on the decision of the admission jury.



PROGRAM REQUIREMENTS

120 credits: Required courses (116 credits), Institution's Elective Courses (4 credits).

Required Courses (116 Cr.)

Cloud and Digital Transformation (4 Cr.), Graph Theory and Operational Research (4 Cr.), Inferential Statistics (6 Cr.), Natural Language Processing (4 Cr.), Optimization for AI (4 Cr.), Programming for Data Science and Artificial Intelligence (4 Cr.), Data Visualization and Communication (4 Cr.), Enterprise Data Management (5 Cr.), Machine Learning (4 Cr.), Marketing Data Science (4 Cr.), Mining Massive Datasets (4 Cr.), Regression Models (4 Cr.), Social Big Data (4 Cr.), Data-Driven Marketing with AI (6 Cr.), Applied Regression and Time Series Analysis (4 Cr.), Big Data Frameworks (5 Cr.), Generative AI (4 Cr.), Legal, Political and Ethical Considerations for Data Scientists and AI (4 Cr.), Machine Learning Operations (4 Cr.), Theoretical Guidelines for High-Dimensional Data Analysis (4 Cr.), Master Thesis (30 Cr.).

Institution's Elective Courses (4 Cr.), to be chosen from the list below:

Mathematics for AI and Machine Learning (4 Cr.), Relational Database (4 Cr.).

SUGGESTED STUDY PLAN

Semester 1

Code	Course Name	Credits
020CTDIM1	Cloud and Digital Transformation	4
048DSTGM1	Graph Theory and Operational Research	4
048DSSIM1	Inferential Statistics	6
048DSPOM1	Programming for Data Science and Artificial Intelligence	4
020OPAIM1	Optimization for AI	4
020IANLM1	Natural Language Processing	4
020IAMAM1 or 020BDREM1	Institution's Elective Courses Mathematics for AI and Machine Learning or Relational Database	4
	Total	30

Semester 2

Code	Course Name	Credits
020INTDM2	Enterprise Data Management	5
020DVCOM2	Data Visualization and Communication	4
020IAMLM2	Machine Learning	4
020BDFRM2	Big Data Frameworks	5
020FOBDM2	Mining Massive Datasets	4
048DSSBM1	Social Big Data	4
048MBCMM2	Regression Models	4
	Total	30

Semester 3

Code	Course Name	Credits
048DSARM3	Applied Regression and Time Series Analysis	4
020DDMAM3	Data-Driven Marketing with AI	6
020MADSM2	Marketing Data Science	4
020IAGIM3	Generative AI	4
020IALPM3	Legal, Political and Ethical Considerations for Data Scientists and AI	4
020SEAIM3	Machine Learning Operations	4
048DSTGM3	Theoretical Guidelines for High-Dimensional Data Analysis	4
	Total	30

Semester 4

Code	Course Name	Credits
020STGEM4	Master Thesis	30
	Total	30

COURSE DESCRIPTION

020CTDIM1	Cloud and Digital Transformation	4 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 cost-effectively.

048DSTGM1	Graph Theory and Operational Research	4 Cr.
------------------	--	--------------

This course introduces students to graph theory and operations 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 in 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.
------------------	-------------------------------	--------------

This course covers statistical inference, focusing on predicting unknown population characteristics from sample data. It enables students to conduct complete statistical studies, including selecting appropriate models, estimating unknown quantities, and making informed decisions. Applications use R software for data manipulation, implementing statistical procedures, plotting graphics, and presenting results clearly. Students will apply these methods to analyze data and interpret statistical findings effectively.

048DSPOM1	Programming for Data Science and Artificial Intelligence	4 Cr.
------------------	---	--------------

This course equips students with the necessary tools for developing advanced-level programs using 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.

020IAMAM1	Mathematics for AI and Machine Learning	4 Cr.
------------------	--	--------------

This course covers the mathematical foundations essential for programming in Artificial Intelligence (AI) and Machine Learning (ML). It focuses on three main areas: linear algebra, multivariate calculus, and probability theory. Students will master these mathematical concepts to develop and implement AI and ML algorithms effectively, providing a strong basis for advanced programming and algorithm design in these fields.

020IANLM1	Natural Language Processing	4 Cr.
------------------	------------------------------------	--------------

This course goes beyond the phase of gathering large amounts of data by focusing on how machine learning algorithms can be rewritten and scaled to work on petabytes of both structured and unstructured data simultaneously, to generate sophisticated models used for making predictions. This course is divided into two parts. The first part covers deep learning and key network architectures, including: convolutional neural networks, autoencoders, recurrent neural networks, and long short-term memory (LSTM) networks. This part also addresses stochastic networks, conditional random fields, Boltzmann machines, stochastic and mixed deterministic models, as well as deep reinforcement learning. The second part focuses on natural language processing (NLP): Research in automatic natural language processing is a subfield of artificial intelligence aimed at developing automated techniques for manipulating linguistic data. Immediate applications of these techniques include the development of more natural textual interfaces, automatic document translation, spam detection, information retrieval from document collections based on queries, question/answer systems, and more. This part introduces students to the following topics: Introduction to the problem of automatic natural language processing and its applications. The relationship between natural language and formal languages: the problem of ambiguity. An overview of current linguistic theories. Speech analysis and synthesis. Morphological analysis: structure of the dictionary and suffix analysis. Syntactic analysis: ATN parsers, unification grammars, and the representation of the semantics of natural languages: formal logic and frameworks. Semantic interpretation. World knowledge and speech context. Applications.

020OPAIM1	Optimization for AI	4 Cr.
------------------	----------------------------	--------------

This course delves into the mathematical optimization techniques essential for developing and refining machine learning algorithms and AI applications. Focusing on theoretical foundations, this course explores deep neural network initialization, gradient descent techniques, automatic differentiation and backpropagation, and adaptive learning rate algorithms such as Adam and RMSProp. Additionally, it covers principal component analysis (PCA), density estimation algorithms, and support vector machines (SVM). Students will learn to solve unconstrained and constrained optimization problems, apply these methods to neural networks, and enhance model performance. The course provides a comprehensive understanding of optimization's role in AI, equipping students with the theoretical knowledge to tackle complex challenges in various AI domains.

020BDREM1	Relational Database	4 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.

020INTDM2	Enterprise Data Management	5 Cr.
------------------	-----------------------------------	--------------

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.

020DVCOM2	Data Visualization and Communication	4 Cr.
------------------	---	--------------

This course introduces concepts of data visualization, analysis, and visual representation, essential for managing and interpreting large data flows. It covers designing and implementing complementary visual and verbal representations to convey results, answer questions, support decision-making, and provide evidence-based insights. Students will apply these techniques to create effective tools and applications for analyzing and communicating big data.

020MADSM2	Marketing Data Science	4 Cr.
------------------	-------------------------------	--------------

This course provides students with foundations in various modeling methodologies commonly used in marketing, including Attribution Modeling, Marketing Mix Modeling (MMM), and other advanced techniques such as Bayesian methods. Through a combination of theoretical lectures, hands-on practical exercises, and real-world case studies, students will develop the skills necessary to analyze marketing data, derive actionable insights, and make data-driven decisions to optimize marketing strategies.

020IAML2	Machine Learning	4 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. Sophisticated models for predictions are also 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: research in the automatic processing of natural languages is a field of artificial intelligence aimed at developing automated techniques for the manipulation of language data, in textual or audio form. 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 Datasets	4 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 structures (XML and JSON), graph-structured data (RDF), new workloads such as machine learning tasks (Spark), and mixed workloads (such as Google Cloud Dataflow).

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.

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 using the R language. Students will gain proficiency in utilizing different application programming interface services (APIs) 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.

020DDMAM3	Data-Driven Marketing with AI	6 Cr.
------------------	--------------------------------------	--------------

This course applies AI and machine learning to marketing decision-making across the customer journey. Students will use predictive and generative AI for segmentation, personalization, recommendation systems, forecasting, pricing, and campaign optimization. Through case studies and a hands-on project, students will design, evaluate, and optimize data-driven marketing strategies while addressing fairness, transparency, and privacy in the use of AI and customer data.

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), autoregressive integrated moving average (ARIMA), GARCH models, vector autoregression (VAR), statistical forecasting, and regression with time series data.

020BDFRM2	Big Data Frameworks	5 Cr.
------------------	----------------------------	--------------

This course is 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 will 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 trees), graphics processing algorithms (such as PageRank and shortest path), Newton algorithms, and support vector machines.

020IAGIM3	Generative AI	4 Cr.
------------------	----------------------	--------------

This course offers a comprehensive understanding of generative models in AI, like ChatGPT and diffusion models, and the practical application of these technologies. The course emphasizes open-source models, training techniques, and fine-tuning practices.

Objectives:

- Understand the architecture and principles of generative models.
- Learn the process of training and fine-tuning AI models.
- Explore open-source AI models and their applications.
- Develop hands-on experience in building AI-driven solutions.

020IALPM3	Legal, Political and Ethical Considerations for Data Scientists and AI	4 Cr.
------------------	---	--------------

This course examines the ethical, legal, and political dimensions of data, including personal data, throughout the data lifecycle—from collection and storage to processing, analysis, and use. It addresses key issues such as privacy, surveillance, security, classification, and discrimination, and introduces fundamental concepts of law and labor law. Through case studies in criminal justice, national security, healthcare, marketing, politics, education, the automotive sector, employment, sports, and development, students will analyze the ethical implications and the legal and policy considerations specific to each field.

020SEAIM3	Machine Learning Operations	4 Cr.
------------------	------------------------------------	--------------

This course provides a comprehensive exploration of software engineering principles tailored for AI applications. It covers the entire software development lifecycle (SDLC) for AI projects, including requirements engineering, design patterns for machine learning applications, and software design for AI systems. The course delves into development tools and techniques essential for AI software development, and emphasizes machine learning operations (MLOps) such as model training and deployment pipelines, model monitoring and performance evaluation, version control and management of machine learning models, and responsible AI practices focusing on bias, fairness, and explainability.

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 (such as 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 inferential statistics, sampling, experimental design, parametric and non-parametric difference tests, least squares regression, and logistic regression.

020STGEM4	Master Thesis	30 Cr.
------------------	----------------------	---------------

During the final semester, students must complete a four-month professional internship in a company or undertake a research project in a laboratory, either in Lebanon or abroad.

The professional internship is jointly supervised by a company representative and a faculty member from USJ or a partner university and is designed to develop the professional and technical skills required in data science. Alternatively, students may contribute to an academic research project conducted in a laboratory.

Both options culminate in a written thesis or internship report and a public oral defense before a panel of USJ faculty members. Students may submit their internship report or defend their research thesis only after successfully completing all courses in the first year and the first semester of the second year of the Master's program.

The thesis or internship report must include a literature review and a technical component.

The project is evaluated based on:

- The student's scientific initiative.
- The written report.
- The oral defense.

