USJ 1875

Université Saint-Joseph de Beyrouth

جامعة القدّيس يوسف في بيروت

Faculté des sciences

Faculty of Engineering

faculty of Sciences

MASTER en Data Sciences

ماستر في علم المعلومات

MASTER in Data Science

August 2023

Address: Saint-Joseph University of Beirut
Sciences and Technologies Campus
Mar Roukos - Dekwaneh – Lebanon
PO Box 1514 - Riad El Solh
Beirut 1107 2050

École supérieure d'ingénieurs de Beyrouth (ESIB)
Email : fi.esib@usj.edu.lb
Tel : +961 (1) 421317
Fax : +961 (4) 532645

Faculty of Sciences
Email : fs@usj.edu.lb
Tel : +961 (1) 421 368
Fax : +961 (4) 532 657

1. Presentation

- **INTRODUCTION**

We live in a data-driven world that's generating huge volumes of information at ever-increasing rates, via social media, financial transactions, telecommunications, and even scientific discovery.
The emerging interdisciplinary field of data science combines areas of computer science with mathematical statistics and domain expertise to manage and analyze this data.
Incorporating statistical methods, data science puts a greater emphasis on the specialized computational skills required to manage and analyze big data from sources such as social media, sensors, mobile and transaction data.
Data scientists develop the capability to derive insight and opportunity from the vast repositories of data that many organizations collect. They help organizations in all sectors of the economy to make sense of these very large volumes of data; enabling businesses to gain a competitive edge, governments to deliver more targeted services, and research teams make new discoveries.

- **PRESENTATION OF THE PROGRAM**

The Master of Data science is a joint program between the Faculty of Sciences and the Faculty of Engineering of the Saint Joseph University. Courses are given in both Faculties taking advantage of the diversity of optional courses the students can chose from following their focus of interest. Courses are given in English.

The nature of the program prepares students for two main career strategies; either research dedicated, or market orientated. In fact, our program offers a solid theoretical background in statistics, regression and other data related mathematics topics, along with an intense practical side focusing on the mathematical and programming tools that are essential for data analysis, machine learning and big data.

In short, the program was put together by professors from science and engineering in collaboration with European colleagues from the data science field and students can find all the necessary elements they need to work in today's competitive market.

- **PREREQUISITES AND PROGRAM SUMMARY**

The program welcomes students from different scientific backgrounds, essentially; mathematical or physical sciences, engineering and computer science. During the first year, depending on the student's background, specific complementary courses are given in order to prepare the student for the second year through advanced courses in computer science, applied mathematics and statistics-core skills. The second year brings more effort on the market needed skills in data analytics, applied machine learning, big data analysis and other.

- **INTERNSHIP SCHEME**

The second and final year of the program includes a mandatory 3 to 4 months internship. We strive to offer our students the best opportunities to get in contact with the market and open the first door towards an excellent career. At this point students will have to choose between two major career strategies; an industry orientated career that can be started immediately after obtaining the Master degree or a research project with a PhD in sight. Of course, ambitious students can opt for both options in parallel by engaging in a PhD program along with a career start as a part time data scientist until the PhD is awarded.

- **Career outlook**

As data science is still a new and emerging field, the roles available for data scientists are quite varied and diverse. Along with the title of data scientist, other positions include: analytics specialist, business intelligence analyst/developer, data analyst, data architect, data engineer, data miner, research scientist and web analyst.

This program also initiates the students into AI. In fact, students who have successfully completed the Master's Degree will be eligible to pursue a PhD.

2. Admission and Registration

- **Admission**

Admission of students is based on their file and an interview might be required.

1- Admission to the first semester of the Master's program (S1)

To be authorized to submit application files, students must satisfy one of the following conditions:

- ➢ Hold a BS Degree in Computer and Communications engineering, Computer Science, or Mathematics,
- ➢ Hold an equivalent Degree recognized by USJ.

The application file is downloadable from the site of Saint Joseph University of Beirut and is to be deposited in:

Faculté des sciences de l'USJ
Mar Roukos, Mkalles
Tel : (01) 421 368

2- **Admission to the third semester of the Master's program (S3)**

To be authorized to submit application files, students must satisfy one of the following conditions:

- ➢ Hold a BE Degree in Computer and Communications Engineering or being a CCE Program Student at ESIB and earned at least 120 credits in Engineering Cycle.
- ➢ Hold a Master Degree in Computer Science, or Computer and Communication, or Mathematics.
- ➢ Hold an equivalent Degree recognized by USJ.

The documents required when submitting the application form are specified in the common admission file specific to Saint Joseph University of Beirut.

The submitted files will be examined by a joint Scientific Committee (from the Faculty of Engineering and the Faculty of Science), which will subsequently establish the list of admissible candidates. The Scientific Committee will decide for each application the teaching units (teaching units) validated according to the program and the results obtained previously. The selected candidates might be interviewed before their final admission.

The application file is downloadable from the site of Saint Joseph University of Beirut and is to be deposited in:

Faculté d'ingénierie de l'USJ

Mar Roukos, Mkalles

Tel : (01) 421317

3. Degree and regulation

- **Language**

All the courses are offered in English. The file of each candidate must include a written statement certifying that the student has high proficiency in English language (written by the candidate if he does not have an official certificate). If deemed necessary, the Scientific Committee might check the English level of the candidate and might require, if necessary, some remedial courses.

- **Degree requirements**

The Master's Degree in Data Sciences is awarded to candidates who have successfully passed the evaluations of the theoretical and practical Teaching Units (courses) and who show excellent level during their thesis defense. There is no provision for makeup exams in case of a missed exam or test. In the event of a serious accident, duly and seriously justified, the case will be examined by the jury to take the measures deemed appropriate.

- **Attendance**

Student attendance is compulsory for all teaching activities.

- **Conditions**

Each Teaching Unit is given a grade. Following the exam period, the jury finalizes the results. A Teaching Unit, with the exception of the Master's Thesis, is validated if its grade is greater than 10/20. Successful completion of 90 credits for semesters 1, 2 and 3 is mandatory in order to present the Master Thesis report. The priority in the choice of Master Thesis is based on the overall GPA. The Master Thesis is validated if its grade is greater than or equal to 12/20.

- **Degree**

The Master's Degree in Data Sciences is awarded to admitted students having validated all the Teaching Units of the 4 semesters M1-S1, M1-S2, M2-S3 and M2-S4. The scoring system is defined by the internal regulations of Saint Joseph University of Beirut.

4. Skills and Learning Outcomes

Upon graduation, students will be able to:

I. imagine new and valuable uses for large datasets;

II. apply creative and disciplined methods for asking questions and interpreting results;

III. retrieve, organize, combine, cleanse, and store data from multiple sources;

IV. apply machine learning and appropriate statistical techniques to identify trends and make predictions;

V. visualize and effectively communicate results;

VI. understand the ethical and legal requirements of privacy and data security.

VII. Review research articles from well-known Data Science journals and conference proceedings regarding the theories and applications of Data Science.

VIII. Perform research project and write research proposal, report and paper.

Student Outcome SO	**Key Performance Indicator KPI**
a) Ability to apply knowledge of mathematics, physics in problem solving	a1. Apply knowledge of mathematics to solve problems
	a2. Apply knowledge of statistics and data analysis to solve problems
b) Ability to design, conduct experiments, analyze and interpret data	b1. Plan experiments by applying theoretical knowledge and selecting appropriate data to measure
	b2. Perform experiments by correctly manipulating tools and measuring data
	b3. Analyze the data using the appropriate tools, and interpret the results
c) An ability to design a system, component, or process that meets the needs and realistic constraints of economic, environmental, social, political, ethical.	c1. Develop a design strategy by analyzing needs and respecting technical and non-technical constraints
	c2. Propose a solution adapted to needs, and compare it to alternative solutions
	c3. Test, improve, and implement using the appropriate tools such as modeling, prototyping, and performance testing
d) An ability to identify, formulate and solve data analysis problem	d1. Identify a data analysis problem by selecting the information used and needed
	d2. Formulate a problem by adopting an appropriate model
	d3. Solve the problem by using appropriate tools and applying technical knowledge
e) An understanding of professional and ethical responsibility	e1. Describe the professional code of conduct such as responsibility to the different actors: customers, employees, administration, society, environment

	e2. Understand the legal and security responsibilities related to the job of data analyst
f) Education needed to understand the impact of data analysis in a global, economic, environmental and societal context	f1. Describe the local and global impact of data analysis on individuals and society, identifying relevant resources and making informed judgment
g) Knowledge of contemporary subjects	g1. Quote recent developments related to the field of data analysis
h) Ability to use the modern techniques, skills, and tools needed for data analysis	h1. Use techniques necessary for professional practice such as design, prototyping and simulation
	h2. Use the skills needed to practice the profession, such as programming and tool manipulation

5. Program

The Master program is spread over 2 years. The Teaching Units (courses) are distributed over semesters S1, S2, S3 and S4.

Semester 1 (S1)	**Type**	**Credits**
Cloud and digital transformation (020CTDIM1)	Mandatory	**6**
Graph theory and operational research (048DSTGM1)	Mandatory	**6**
Inferential statistics (048DSSIM1)	Mandatory	**6**
Programming for Data Science and Artificial Intelligence (048DSPMM1)	Mandatory	***6***
Elective Courses (One of the following two courses)		
Mathematics for Data Science and AI (020IAMAM1)	Elective	**6**
Relational Database (020BDREM1)	Elective	**6**
		30

Semester 2 (S2)	**Type**	**Credits**
Enterprise Data Management (020INTDM2)	Mandatory	**6**
Foundations of Decision modeling (020IADMM2)	Mandatory	**6**
Machine learning and Deep learning (020MLDLM2)	Mandatory	**6**
Mining Massive Data Set (020FOBDM2)	Mandatory	**6**
R langage (048DSLRM2)	Mandatory	**2**
Regression models (048MBCMM2)	Mandatory	**4**
		30

Semester 3 (S3)	**Type**	**Credits**
Applied regression and time series analysis (048DSARM3)	Mandatory	**4**
Big Data Frameworks (020BDFRM3)	Mandatory	**4**
Data visualization and communication (020DVCOM3)	Mandatory	**4**
Legal, Policy and Ethical considerations for Data scientists and AI (020IALPM3)	Mandatory	**2**
Natural Language Processing (020IANLM3)	Mandatory	**4**
Social Big Data (048DSSBM1)	Mandatory	**4**
Theoretical guidelines for high-dimensional data analysis (048DSTGM3)	Mandatory	**4**
Web mining (020WEMIM3)	Mandatory	**4**
		30

Semester 4 (S4)	**Type**	**Credits**
Master Thesis (020STGEM4)	Mandatory	**30**

6. Courses description

The Master in Data Science is designed to prepare data science leaders. This program introduces best practices for collecting, storing, and retrieving data - and how these factors influence the speed, accuracy, and reliability of large-scale storage models. The program prepares students to apply the latest statistical and computational methods to identify patterns and extract knowledge from complex data and eventually predict certain actions. They especially learn to communicate the results of data analysis by effectively using visualization tools and are exposed to ethical dilemmas and legal requirements related to the use of real-world data.

Semester S1 (30 credits)

Cloud and digital transformation (020CTDIM1 – 6 Credits)

Cloud computing and big data are currently the two main technological developments and the main growth drivers for companies in the digital sector. Big data, 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, thereby resulting in a significant reduction in expenses. These two subjects are closely linked: cloud computing is the only technology capable of supporting the computing 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 use existing cloud services to process data using the vast ecosystem of tools. Students will also learn to create Big Data environments and apply the best practices in order to those environments in a secure and economical approach.

Graph theory and operational research (048DSTGM1– 6 Credits)

This teaching unit introduces students to the graph theory and operational research as modeling and decision-making tools for the data scientist. Therefore students will learn to make a mathematical and computer representation of graphs, apply the algorithms for traversing the graphs, calculate the shortest path, maximize a flow problem, analyse complex networks, use the Networkx Python library, use Markov chains to slove real-world problems, understand the Simplex algorithm and linear programming, use numerical tools for solving optimization problems.

Inferential statistics (048DSSIM1– 6 Credits)

Statistical inference consists in predicting the unknown characteristics of a population from a sample from this population. Thus, the objective of the statistics is symmetrical to that of the probability. At the end of this course, the student is able to conduct a complete statistical study: from the choice of the statistical model, to the estimation of unknown quantities and concluding with a decision. The applications attributed during this course are led using the R language software mainly for data manipulation, use of statistical procedures, plotting graphics and functions and presenting results in a comprehensible way.

Mathematics for AI & Machine Learning (020IAMAM1 – 6 Credits)

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 is extremely promising and it isn't far from when we have our own robotic companions. This has pushed a lot of developers to start writing codes and start developing for AI and ML programs. However, learning to write algorithms for AI and ML isn't easy and requires extensive programming and mathematical knowledge.

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.

Programming languages for Data Science and Artificial Intelligence (048DSPOM1 - 6 Credits)

This course gives the student the necessary tools to develop advanced level programs understanding the Object Oriented Programming (OOP) approach. The first part of the course focuses on the C++ language and the second part on Python and its functionalities that are related to Data Science. The final part of the course shows an introduction to machine learning examples using Python allowing the student to explore the power of the libraries provided by the Python community.

Relational database (020BDREM1 – 6 Credits)

This teaching unit aims to introduce students to the design, creation and management of databases. It allows students to master the "Database" concept, design a Database for a given Information System (IS), understand the Relational Model, know how to create and manage a Database using SQL language and understand the techniques of database management systems.

Semester S2 (30 credits)

Enterprise Data Management (020INTDM2 – 6 Credits)

"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." (Wikipedia).

This course addresses the challenges of enterprise data management at scale, mainly at the level of the data architecture, data modeling and data integration, on-premise as well as on the cloud. It covers different enterprise data architectures i.e DataWarehouses, and DataLakes. It details various data models (structured, semi-structured (XML), unstructured and semantic data with RDF/OWL/SPARQL, and describes various NoSQL databases (key-value, Column, Document or Graph Oriented Databases), as well as various Big Data Formats (Avro, ORC and Parquet). It describes 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. "

Foundation of Decision Modeling (020IADMM1 – 6 Credits)

Preferences are present and pervasive in many situations involving human interaction and decisions. Preferences are expressed explicitly or implicitly in numerous applications and relevant decision should be made based on these preferences. This course aims at introducing preference models for multicriteria decisions. We will present concepts and methods for preference modelling and multicriteria decision making. The course also presents stochastic processes and estimators.

Machine learning and Deep learning (020MLDLM3 – 6 Credits)

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 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. The immediate applications of these techniques include the development of more natural textual interfaces, the automatic translation of documents, the detection of spam, the search for information in a collection of documents, the systems of questions / answers, and several others.

This part introduces the student to the following subjects: Introduction to the problem of automatic processing of natural language and its applications.

Mining Massive Data Set (020FOBDM2 – 6 Credits)

This course covers the fundamentals of designing dedicated software systems for big data analytics.

The course begins with the principles of design of relational database systems for the analysis of business data, including declarative queries, query optimization and transaction management, as well as the evolution of basic systems of data to support complex analytical problems and scientific data management.

The course then looks at fundamental architectural changes to the scale of data processing beyond the limit of a single computer, including parallel databases, "MapReduce", column storage and distributed key value, and allows the calculation of low latency analytical results from real-time data streams. Finally, this course examines advanced data management systems to support models of various data including tree structure (XML and JSON) and structured data graphics (RDF) and new workloads such as learning tasks. Automatic (Spark) and mixed workloads (Google Cloud data flow).

R language (048DSLRM1 – 2 Credits)

In this course students are introduced to the R language programming, basic concepts and essential functionalities for data treatment.

Regression models (048MBCMM2 – 4 Credits)

This course addresses the fundamentals of regression, i.e. linear regression, its approach, and its applications to practical studies. ANOVA techniques and logistic regression are also included. The course alternates theoretical presentations and computer exercises. The exercises are carried out using the R language.

Semester S3 (30 credits)

Applied regression and time series analysis (048DSARM3 – 4 Credits)

This course introduces the student to 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, regression with time series data.

Big Data Frameworks (020BDFRM3 – 4 Credits)

Conceptually, the course is divided into two parts.
The first covers the fundamental concepts of MapReduce parallel computing, through the eyes of Hadoop, MrJob and Spark, while delving deep into Spark, data frames, Spark Shell, Spark Streaming, Spark SQL, MLlib. Students will use MapReduce for industrial applications and deployments for various fields, including advertising, finance, health, and search engines.
The second part focuses on algorithmic design and development in parallel computing environments (Spark), development of algorithms (learning decision tree), graphics processing algorithms (pagerank / short path), Newton algorithms, and support vector machines.

Data visualization and communication (020DVCOM3 – 4 Credits)

Access to data is exponentially increasing while the human capacity to manage and understand it remains constant. Communicating clearly and effectively on the models we find in the data is a key skill for a successful data scientist. This course introduces the basic concepts of visualization, analysis and visual representation of data. These concepts are necessary to create suitable applications and tools that allow the student to manage and analyze big data flows. It involves the design and implementation of complementary visual and verbal representations of patterns and analysis in order to convey results, answer questions, drive decisions, and provide convincing evidence supported by data.

Legal, Policy and Ethical considerations for Data scientists and AI (020IALPM3 – 2 Credits)

This course provides an introduction to the ethics, politics, and ethical implications of data in general, including personal data. The course will examine the legal, political, and ethical issues that arise throughout the entire life cycle of data science from collection, storage, processing, analysis and use, including, privacy, surveillance, security, classification and discrimination. Moreover, a brief introduction will be given about law and Labor law in general. Case studies will be used to explore these issues in various fields such as criminal justice, national security, health, marketing, politics, education, automotive, employment, athletics, and the development. Particular attention will be paid to legal and political constraints and considerations which are set in specific areas.

Natural language processing (020IANLM3 - 4 credits)

This course goes beyond the phase of gathering large amounts of data by focusing on how machine learning algorithms can be rewritten and scaled up to work on petabytes of data, at the same time. both structured and unstructured, to generate sophisticated models used to make predictions. Conceptually, the course is divided into two parts.

The first part deals with deep learning and key network architectures including: convolutional neural networks, autoencoders, recurrent neural networks, short-term long-term memory networks LSTM. This part also covers stochastic networks, conditional random fields, Boltzmann machines, stochastic and mixed deterministic models as well as deep reinforcement learning.

The second part deals with the processing of natural languages: Indeed, research in automatic natural language processing (NLP) is a field of artificial intelligence aiming at the development of automated techniques for the manipulation of linguistic data. Immediate applications of these techniques include the development of more natural textual interfaces, automatic document translation, spam detection, search for information in a collection of documents from queries, question / answer systems, and several others. This part introduces the student to the following topics: Introduction to the problem of automatic processing of the natural language and its applications. The natural language in relation to formal languages: the problem of ambiguity. Overview of current linguistic theories. Analysis and synthesis of speech. Morphological analysis: structure of the dictionary and suffix analysis. Syntax analysis: ATN parser, unification grammars and representation of the semantics of natural languages: formal logic and frameworks. Semantic interpretation. Knowledge of the world and speech context. Applications.

Social Big Data (048DSSBM1 – 4 Credits)

The main objective of this course is to introduce students to the structures and types of data present on social networks (Facebook, Twitter, Instagram…) as well as the forms of data collection and analysis based on application areas under the R language. Students will learn to use different application programming interface services (API) to collect data, to analyze and explore social media data for research and development purposes and thus be able to use the data drawn and analyzed to improve the presence and strategy on social networks.

Theoretical guidelines for high-dimensional data analysis (048DSTGM3 – 4 Credits)

The purpose of this course is to provide students with an introduction to the different types of quantitative research methods and statistical techniques for analyzing data. We start with an emphasis on measurement, statistical inference and causal inference. Next, we explore 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.

Web mining (020WEMIM3 – 4 Credits)

This course is divided into 3 parts:

- The first part essentially aims to introduce the 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)

Master Thesis (020STGEM4 – 30 Credits)

During the last semester, students must complete a professional project in a company or research work in a laboratory for a period of 4 months.

The project can take place in Lebanon or abroad. Scientific responsibility for the project is shared between the company and a teacher from the USJ or a partner university. This project, of a minimum duration of one semester, aims to develop the skills of the student preparing him for the Data Science work field.

The student can also choose to contribute in an academic research project. The research work can take place in a laboratory either in Lebanon or in an external establishment.

A report detailing, the project or the research of the student should be presented in the form of a dissertation and defended publicly with the presence of professors from USJ.
Only students who have validated all the courses of the first year and the first semester of the second year of the Master are authorized to present the project report or the research dissertation.
The project or the report should include a bibliographic part and a technical part.
The evaluation of the project takes into account three elements:
- evaluation of the trainee's scientific initiative,
- evaluation of the report,
- evaluation of the oral defense.