MASTER OF SCIENCE IN STATISTICS FOR SMART DATA
CURRICULUM- COURSE DESCRIPTIONS

Note: In order to ensure that the curriculum is adapted to the needs of the current job market and its students, ENSAI reserves the right to modify the proposed curriculum and the following descriptions at any time during the academic year.

BEFORE SEMESTER 1

Before the main courses start, some preliminary modules are organized. The list of these courses could change from one year to another. There is no ECTS credit associated with these preliminary modules. The preliminary courses only allow students to complete the prerequisites for the MSc. Depending on their background, students will be asked to take some or all of these courses.
Tentative list of preliminary courses:

- Statistical languages: R, Python (18h)
- Multivariate Data Exploration (12h)
- Markov Chains (12h)
- Simulation Based Inference (6h)
- Topics in Time Series (6h)

SEMESTER 1

Innhomogeneous Markov Models & Applications
(Lectures and Tutorials: 30hrs)

Homogeneous Markov chains are exploited in a broad range of applications. Nevertheless, in some situations homogeneous transition probabilities do not adequately model real processes. In these situations, Markov models with inhomogeneous rates, i.e., rates that are time-varying functions or that depend on covariates, could be much more appropriate. In the first part of this course, the theory of these processes will be described and will be illustrated with applications in several areas (financial, aeronautics, meteorological…). The second part of this course will be devoted to Hidden Markov Models (HMM). After presenting the basic HMM, the framework is extended by considering nonhomogeneous hidden Markov models and Markov-switching models. Such models are used in finance, electricity prices, genomics... Bayesian methods, such that MCMC and particle filters, and the Expectation Maximization algorithm will be introduced and applied to infer in these models. Several real data applications will be used to illustrate the methods.

Graphical Models & Dynamic Networks
(Lectures and Tutorials: 18hrs)

Due to the ability to represent complex phenomena, such as large-scale networks, the probabilistic graphical models have received a lot of attention in the last two decades. The most commonly used types of graphical models are introduced and their basic properties are presented. The common algorithms for inference and learning with such graphical models are studied. Extensions to dynamic graphical models will be considered. Next, some approaches for community detection in networks (finding dense sub-networks within a larger network) will be presented. Such methods are usually based on stochastic block models, for which there has been a surge of interest in recent years. Large-scale social or biological networks examples will illustrate the algorithms presented in this course.

Dynamic Data Visualization
(Lectures and Tutorials: 12hrs)

Data visualization is a fundamental ingredient of data science as it “forces us to notice what we never expected to see”. In this course, we show through examples and case studies that graphical methods are powerful tools for revealing the structure of the data, patterns and (ir)regularities, groups, trends, outliers… Dataviz is relevant for data analysis, when the analyst wants to study data, but also, as any statistics, to question the data. It is also a tool for communication and, as such, a visual language with a theory of the functions of signs and symbols used to encode the visual information. All along the course, we’ll focus on methods, tools and strategies to represent simple and then complex or high-dimensional datasets, highlighting the growing development of dynamic and interactive tools.

Machine Learning: Features Selection & Regularization Methods
(Lectures and Tutorials: 18hrs)

Starting from classical notions of shrinkage and sparsity, this course will cover regularization methods that are crucial to high-dimensional statistical learning. The syllabus includes feature selection and model selection, linear and nonlinear techniques for regression and for classification. The course will focus on methodological and algorithmic aspects, while trying to give an idea of the underlying theoretical foundations. Practical sessions will give the opportunity to apply the methods on real data sets using either R or Python.

Deep Learning
(Lectures and Tutorials: 30hrs)

This course will start by presenting the most common methods of classification (boosting algorithms, bagging trees, random forests,…). Next, the course provides an introduction to neural networks, and their extension now known as deep learning. We will start with deep feedforward networks, give some insight about regularization and optimization (e.g., through backpropagation) of these networks. Finally, a quick overview of convolutional networks and of recurrent networks will also be presented. All these notions will be illustrated on real data sets, using either some R packages or some Python libraries.

Parallel Computing with R and Python
(Lectures and Tutorials: 12hrs)

These tutorials will show how to implement parallel computing on a computer equipped with multiple cores or on a computer cluster. Dealing with memory issues caused by large data sets will also be examined. The Map/Reduce framework will be explored, as well as distributing computations with the aid of the Graphical

Processing Unit. All these techniques will be applied on real or simulated big data sets using either R or Python.

Foundations of Smart Sensing
(Lectures and Tutorials: 36hrs)

Although most signals (audio, images,…) of interest belong to a very large dimensional ambient space, many of them possess some structure that contains the useful information and makes them intrinsically low dimensional or compressible. Such structures can be modeled and exploited to reduce the cost of their acquisition and their processing. Sparse representations are a powerful tool to express and represent such signals, typically by a small number of nonzero coefficients in an appropriate basis or dictionary. Combined with some appropriate design of the sensing devices, they allow to acquire and to reconstruct signals with an extremely reduced number of measurements. This course will present theoretical and algorithmic frameworks and tools for low-dimensional representations of large-scale data and their compressed acquisition and recovery.

Advanced Topics in Smart Sensing
(Lectures and Tutorials: 24hrs)

This module is split into three or four independent parts where various applications of the compressive sensing will be presented.

High-dimensional Time Series
(Lectures and Tutorials: 30hrs)

To model and forecast multivariate time series, practitioners often face a high dimensional problem due to the large number of parameters involved in the dynamics. The regularization techniques, originally introduced for linear regression models, could be particularly useful for fitting vector autoregressive models to the data. The first part of the course will present such approaches. Next, estimation of high dimensional correlation matrices will be considered. This problem also requires special attention due to the lack of precision of the empirical covariance matrix, especially when the inverse of this matrix is the object of interest, as is the case in some applications. An overview of some existing methods for getting more accurate estimates of these covariance matrices in a high dimensional setup will be presented. Finally, the increasing size of available databases has led to the development of factor models, especially in Econometrics. Such models are a versatile approach to summarize information contained in large vectors of data. In the last part of the course, the fundamental factor models and the common inferences approaches will be presented and illustrated with real datasets, with a focus on dynamic factor models.

Functional Data Analysis
(Lectures and Tutorials: 30hrs)

Functional data analysis (FDA) is about the analysis of information on samples of curves or functions. Such data naturally arise when recording electric power consumption, temperature or pollution levels during the day, daily brain activity, etc. Students will learn the ideas of different methods and the related theory, and also the numerical and estimation routines to perform functional data analysis. Several functional regression models, such as generalized functional linear regression and Gaussian process regression analysis, will be presented. The course will demonstrate applications where functional data analysis techniques have clear advantage over classical multivariate techniques. Recent extensions to functional time series data analysis will also be discussed.

IT Tools 1: GNU Linux & Shell Scripting, Hadoop & Cloud Computing
(Lectures and Tutorials: 30hrs)

One of the goals of this module is to present the concepts that a data scientist should understand before starting with GNU/Linux. During the first part of the module, students will install a distribution on their computer and learn how to interact with the shell, from basic tasks (navigation, file edition, network configuration) to more advanced operations with shell scripting. GNU/Linux is essential in particular when using and developing Big Data technologies.
Another goal of this module is to give a brief introduction to Cloud Computing: definitions, types of clouds (IaaS/PaaS/Saas, public/private/hybrid), challenges, applications, main cloud players (Amazon, Microsoft Azure, Google etc.), and cloud enabling technologies (virtualization). Then data processing models and tools used to handle Big Data in clouds such as MapReduce, Hadoop, and Spark will be explored. An overview on Big Data including definitions, the source of Big Data, and the main challenges introduced by Big Data, will be presented. The MapReduce programming model will be presented as an important programming model for Big Data processing in the Cloud. Hadoop ecosystem and some of major Hadoop features will then be discussed.

IT Tools 2: NoSQL, Big Data Processing with Spark
(Lectures and Tutorials: 30hrs)

One of the goals of this module is to understand the fundamentals of NoSQL databases capabilities, features and the specific challenges NoSQL databases are addressing, compared to classic SQL databases. It is as well to get some introduction to deploying and using NoSQL databases, such as MongoDB or CouchDB. Another goal of this module is to understand key concepts and get practice of distributed data processing frameworks such as Apache Spark. All steps of a typical data science project using large volumes of data will be covered: accessing data sources, preparing and processing data, storing them, but also using distributed machine learning libraries such as Apache Spark MLlib and H20 to train and fire models. Emphasis will be set on practice & hands-on sessions.

Energy Transitions: Quantitative Aspects
(Lectures and Tutorials: 12hrs)

This lecture will provide a quantitative economic perspective on energy transitions. Using micro, macroeconomics and

econometric tools, the lecture will present some proven and potential impacts on society of transforming energy in the past, present and future.

Smart Data Project
(equivalent of 24hrs of lectures)

The main part of courses focuses on studying several aspects of Statistics, Machine Learning, and Computer Science, according to the Big Data paradigm. One of the main objectives of this project is to apply all the new knowledge learned to a unique application. The project is supervised by specialists or researchers from academic, industrial, or business fields. The Smart Data project puts into practice theoretical methods studied in different courses, starting with project management. The learning objective is not limited to applying the theory learned in other courses, it also aims to raise awareness on other aspects linked to project management among students, such as communication (between students and also with the client that proposed the project). This project should provide additional support, be carried out by an expert of the field, according to the needs of students.

Topics & Case Studies in Data Science (conferences)
(24hrs)

Several conferences held by specialists or researchers from the academic, industrial or business world will be organized. “Smart Data” is becoming a major issue in modern society. The purpose of these conferences is to provide an up-to-date review of the ongoing data revolution, on the stakes for analyzing the information in a smart way, on presenting recent case studies, and on providing complementary perspectives (economic, business, management) on the Smart Data.

French Summer Program
(August - Duration: 4 weeks)

Non-French speakers arrive 1 month early to France for a mandatory, intensive French language and culture course, while being hosted with a French family. This allows students to acquire vital skills for daily life and cultural integration.

Courses for Non-French Speakers: Written and/or Oral French Language Courses
(Duration: 2 or 4 hours/week over the 1st semester)

Designed specifically for foreign students, these weekly evening courses give students practical written and/or oral French skills, necessary for everday life in France.

SEMESTER 2

End-of-Studies Internship
(Duration: 4 to 6 months from the end of February)

This final phase of the MSc in Statistics for Smart Data program involves a four to six-month paid internship, which can take place either in France or abroad, in either the professional world or academic/research laboratories. This experience should allow for the student to apply the statistical, machine learning, and computer science methods that they have learned during the first semester of coursework. The student must write an Internship Report and defend it in front of a jury in September.