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 (9h)
- Topics in Bayesian Inference (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 as 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)

Among the various statistical methods involving graphs, this course aims to give an introduction to two vast fields of Statistics extremely popular at the era of big data. First, random graphs, which are used when the data itself is available as a graph whose nodes are fixed and edges random. In this case, the objective is to adjust a model to unravel some particular organization of the data. Second, probabilistic graphical models, which give a compact and analytically useful representations of joint distributions over a large number of variables, using graphs. Each graph represents a family of distributions – the nodes of the graph represent random variables, the edges encode independence assumptions. Large-scale of 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. Upon completing this course, students should be able to: select the appropriate methods; implement these statistical methods; compare leading procedures based on statistical arguments; assess the prediction performance of a learning algorithm; apply these key insights into class activities using statistical software.

Deep Learning
(Lectures and Tutorials: 30hrs)

The course starts with a general introduction to machine learning and deep learning giving a brief overview of the problems which may be addressed using different kind of approaches. For the machine learning part we begin with a review of classification and regression trees and then focus on aggregation methods like bagging, random forests (RF), and boosting (AdaBoost algorithm and the gradient boosting optimisation). Support vector machine (SVM) are also discussed.
The second part of the course is devoted to neural network (NN) architectures and their extension known as deep learning. Beforehand, the stochastic gradient descent algorithm and the back-propagation - its application to feedforward neural networks - are introduced to be further used as the learning basis. This is followed by the study of most spread NN architectures for regression and classification. Among those, convolutional neural networks (CNN) are investigated in detail and other structures like Recurrent Neural Networks, Restricted Boltzmann machines (RBM) and the contrastive divergence

algorithm (CD-k) are examined. Further practical aspects will be addressed about the usage of Deep Learning to resolve typical problems like object recognition or tracking, and Image segmentation.
Presented material shall be motivated by the theoretical background together with real data illustrations. There will be specific labs for each topic majorly held in R, along with sparse Tensorflow and Python illustrations.

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

First, an overview of architectures for distributed computing is proposed: those based on central processing units (CPU) including multicore processor, multiprocessor node (single-node cluster), computing cluster; and the graphics processing unit (GPU). The content of the course is devoted to implementation of calculation and memory distribution for the mentioned architectures.
For the CPU parallel computation, the focus is put on two prevailing frameworks: Message Passing Interface (MPI) and the Map/Reduce model. First, parallelising using MPI shall be demonstrated in R with the Rmpi and extending it packages and ScientificPython library in Python. Second, realisation of the Map/Reduce model in Python and using the RHadoop packages are discussed. For the GPU-parallelisation, gpuR package shall be explored, a handy R-wrap to the OpenCL - an open computing language supporting GPU next to NVIDIA's CUDA.
(As the tools that are in the scope of the course evolve rapidly, the professor preserves the right to adapt the content to recent developments.)

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 course collects some advanced topics in compressive sensing and related techniques. Although self-contained, a prerequisite to this course is the course of foundations of smart sensing, that focuses on the notion of vector sparsity.
In many applications, signals are structured in many ways that go beyond the basic notion of sparsity for vectors. We will thus introduce notions of sparsity for matrices, focusing on low rank matrices, and for groups. We will also describe some emerging techniques of compressive learning, where data compression is oriented toward a subsequent learning task. In this context, data compression is also called sketching. Examples of sketched PCA and sketch k-means will be presented.
This course will present theoretical and algorithmic frameworks and discuss applications that are in the scope of the methods.

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)

This course introduces ideas and methodology in functional data analysis (FDA) as well as the use of software. Students will learn the idea of different methods and the related theory, and also the numerical and estimation routines to perform functional data analysis. Students will also have an opportunity to learn how to apply FDA to a wide array of application areas. The course will demonstrate applications where FDA techniques have clear advantage over classical multivariate techniques. Some recent development in FDA 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)

Today's energy transition raises multiple issues, questioning political choices but also industrial firms' strategy and consumers' behavior. This course aims at illustrating the key role of energy in our society.
Firstly, we will present the complex history of energy transitions, closely tied to the economic growth history and feeding the successive industrial revolutions - starting by the steam power, followed by the use of oil and finally the rise of electricity. Secondly, this course will shed light on the microeconomic organization of energy markets and the macroeconomic threats that energy resources and energy use imply for countries – especially, the Dutch Disease and the Environmental Kuznets curve will be evaluated using panel data. Thirdly, industrial organization concepts and micro-econometric analysis will be applied to recent structural changes of the energy markets: oil price fall, renewable energies entry and the uncertain future of nuclear power. Finally, the course will present the upcoming energy transition multiple challenges, focusing on the environmental taxation impact on energy prices (using the example of the EU-ETS), the investment issue in networks and the energy-efficiency paradox. All covered topics will be illustrated using empirical data.

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.