Note: *In order to permanently ensure that the curriculum is adapted for the needs of the current job market, ENSAI reserves the right to make slight changes to the proposed curriculum and the following descriptions between the period of admission and the start of the academic year.*

SEMESTER 1

STATISTICS TRACK COURSES

Probability, Algebra, and Analysis
(Lectures: 15h)

This course introduces several essential notions in probability, algebra and analysis, required for all the following topics in Statistics. To start with, this course covers the basic concepts of probability theory such as the notions of distribution, random variable, mathematical expectation and conditional distribution. The course will continue with some basic notions of linear algebra, as well as a brief review of Hilbertian analysis.

Statistical Inference and Hypothesis Testing
(Lectures: 10h, Tutorials: 5h)

This course provides a short introduction to some basic notions of Statistics. We will first explain the notion of 'statistical model' and the purpose of statistical inference. Some classical estimation procedures, such as the method of moments or the likelihood maximization, will be presented. The notion of statistical hypothesis testing will be consequently discussed, as well as an explanation of its link with the notions of estimation and confidence interval.

Simulation and Monte Carlo Integration Methods
(Lectures: 5h, Tutorials: 5h)

This course provides a short introduction to random variable generation and to Monte Carlo integration. We will start with presenting classical methods for random numbers simulation, such as the inverse method or the accept-reject algorithm. During the second half of the course, the sampling methods from a probability distribution will be applied to numerical integration or quantile approximation.

Regression Models
(Lectures: 10h, Tutorials: 5h)

This course provides an introduction to the linear regression model and its generalizations. The course begins by introducing the basic tools for conducting a linear regression with quantitative or qualitative covariates. Students will learn how to interpret software outputs, to derive regression diagnostics, to correct any deviation to the model, and to make a selection of relevant variables. The second half of the course will be devoted to some classical methods of regression on categorical data.

Multivariate Exploratory Data Analysis
(Lectures: 10h, Tutorials: 5h)

This course provides an introduction to the main exploratory methods used to analyze multivariate data and will summarize their main characteristics. Some standard methods for dimension reduction (such as principal components analysis or multiple correspondence analysis) and for clustering (such as hierarchical clustering or the k-means clustering) will be introduced. Students will also learn how to interpret and to analyze software outputs with R, SAS or SPAD.

Basic Sampling Theory
(Lectures: 10h, Tutorials: 5h)

This course provides an introduction to basic sampling techniques used in the case of finite population sampling and to the properties of the associated estimators. Topics include: Horvitz-Thompson estimator, measures of accuracy, simple random sampling, stratified sampling and unequal probability sampling.

COMPUTER SCIENCE TRACK COURSES

Client – Server Architecture, JavaEE
(Lectures: 10h, Tutorials: 15h)

The objective of this course is to learn how to develop and to deploy a dynamic website using Java. This training allows for students to become familiar with n-tier architectures, servers and master applications, main tools, and advanced web development languages applications.

Cloud Computing
(Tutorials: 10h)

The need for software capable of storage and computing capacity has been increasing since the advent of computers. These resources are now available remotely, in the cloud. These tutorials will introduce the economic issues that gave rise to these services, as well as the technical solutions that allow for access to these web services.

JavaEE Project
(Tutorials: 10h)

This project will begin by modeling an application using methods from software engineering techniques and will continue with its implementation in Java EE.

Computer Networks
(Lectures: 20h, Tutorials: 20h)

This course aims to give students the foundations of operating systems in network architecture. Always-on connectivity, mobile devices, and connected objects are a part of our daily life. Data scientists need to take into account these new technologies. During this course, students will be given a primer on computer networks and the way they allow for new interactions.

COMMON COURSES

Aggregation Methods in Statistics and Combinatorial Complexity
(Lectures: 15h, Tutorials: 5h)

This course provides an introduction to aggregation methods used in statistical learning, such as bagging, random forest or boosting. The main goal of these methods is to combine, or to aggregate, a large number of models in order to improve the quality of the prediction in classification or regression. A short introduction to the Vapnik-Chervonenkis complexity in connection with the proposed algorithms will also be given. The methods will be illustrated by real data examples.

Association Rules Mining
(Lectures: 5h, Tutorials: 5h)

The detection of association rules consists in finding high probability subsets for a finite dimensional random vector. This kind of problem occurs frequently in practice, in particular in market basket analysis, and then identifying customers' most frequently purchased products as well as the association rules between some subsets of products. The corresponding random vector is typically high dimensional, due to the large number of products. The goal of this course is to introduce the appropriate methods and algorithms for dealing with such problems.

Data Visualization
(Tutorials: 10h)

The purpose of data visualization is the visual representation of data through graphical means. Graphical tools are useful to represent complex or high-dimensional data and to communicate information effectively. These tutorials provide an introduction to some of these graphical tools for various data sets using in particular some R software packages, as well as Gephi or GGobbi software.

Olap, Multidimensional Databases
(Lectures: 5h, Tutorials: 10h)

This course presents the multidimensional approach giving direct access to information according to multiple input points. It is used in Business Intelligence to facilitate decision making and to publish reports. The course involves implementing an OLAP hypercube that uses its own data extraction functions.

"Big Data" Databases
(Lectures: 5h, Tutorials: 10h)

This course compares conventional approaches for dealing with Big Data issues, such as using datawarehouses. It will offer an overview of the very large databases that are available to face the new challenges of Big Data: velocity and variety. Problems involving the building and the optimizing of a database will also be discussed.

NoSQL
(Tutorials: 10h)

These tutorials explain to students the principles and foundations of the approach Not Only SQL, as well as a technical and practical overview of NoSQL used technologies, such as BigTable, Cassandra, Redis and MongoDB.

Penalized Regression
(Lectures: 15h, Tutorials: 10h)

For regression models, high dimensional statistics refer to the situation when the number of predictors is one or several orders of magnitude larger than the sample size. For example, in genomic studies, predicting a response to chemotherapy involves several thousands of gene expression measurements across the genome. This course provides an introduction to the penalized regression methods (such as LASSO, ridge regression or Dantzig selector) that are relevant for outcome prediction in this high dimensional framework.

Variable Selection Methods
(Lectures: 10h, Tutorials: 5h)
Variable selection in a high-dimensional setting has received considerable attention in recent years, in particular for regression models involving a large number of possible predictors. Some classical methods such as penalized likelihood or supervised principal components have been adapted to this context. This course provides an introduction to some of these methods together with their practical use.

Unix (shell script)
(Lectures: 5h, Tutorials: 15h)

During this intensive workshop-style course, students will be walked through an installation of the most recent version of Linux on their machines. This will teach students how to use this system in depth, vital for their courses throughout the program. Linux being central to Big Data information systems, a sound and in-depth understanding of Linux is therefore required to succeed in this MSc program.

Parallelized Systems
(Lectures: 10h, Tutorials: 10h)

The founding principles of parallelism will be presented in order to design systems that simultaneously use different distributed resources. These systems are synchronized in their calculations and share other resources. Various other concepts including producer-consumer, reader-writer, monitor, and semaphores will also be addressed.

French Summer Program ("Université d'été")
(July-August at CIREFE, the International Center of French Studies for Foreign Students in Rennes)
(Duration: 9 weeks: Classroom: 200h, Cultural activities: 100h)

Non-French speakers arrive 2 months early to France for a mandatory, intensive French language and culture course, while being hosted with a French family. This allows for students to acquire vital skills for daily life and cultural integration.

Courses for foreigners: Written and/or Oral French Language Courses (at CIREFE)
(Duration: 2 or 4 hours/week over 11 weeks)

Designed for foreign students who are following a full-time academic program in Rennes, these weekly evening courses give students practical written and/or oral French skills, necessary for practical life in France.

SEMESTER 2

COMMON COURSES

Functional Data Analysis
(Lectures: 15h, Tutorials: 10h)

This course provides an introduction to the modeling and the statistical analysis of functional data, and it also investigates the way functional data could be recovered from discretized observations. The extension of standard multivariate exploratory methods such as principal components analysis is presented. Classical regression models and standard prediction tools will be reviewed and generalized to include functional observations, responses and predictors. The methodologies will be illustrated by real data examples treated with specific software packages.

Text Mining, Image Analysis
(Lectures: 10h, Tutorials: 5h)

This course provides an introduction to the analysis of some specific data such as textual or image data. Text mining is the set of methods used for the automatic processing of natural language text data available in computer files. Customer relationship management, document organization or classification, and email spam filters are examples of text mining applications. The course begins by explaining how to use certain classification and clustering methods to analyze such large data sets. Image analysis is another field where the configurations space is generally very large. Pattern recognition and edge detection are classical problems occurring in this field. The course continues by providing an introduction to statistical learning methods used in this context.

Compressive sensing
(Lectures: 15h, Tutorials: 5h)

Compressive sensing exploits the sparsity of a signal and proposes mathematical models that allow for acquiring and reconstructing signals using a few non-zero coefficients in a suitable basis or dictionary. This course provides an overview of some recent advances in compressive sensing. The theory is illustrated by some emerging applications.

Parsimonious Representations
(Lectures: 10h, Tutorials: 10h)

This course presents additional mathematical models and algorithms for low-dimension representations of large scale data. Several applications will be considered, for instance the pattern recognition and the analysis of large sets of images or videos.

Foundations of Big Data using MapReduce
(Lectures: 10h, Tutorials: 10h)

This course presents IT issues arising in the real-time processing of massive and heterogeneous data. It teaches students the foundations of parallel processing technologies (especially statistical) on massive and changing data.

Hadoop Technologies (batch/real time processing), Storm, HD File System
(Lectures: 5h, Tutorials: 15h)

This course explains to students the principles and fundamentals of Hadoop followed by a technical and practical overview of used technologies directly related to Hadoop, such as Pig, Hive, Hbase, ZooKeeper, Mahoot, Spark, etc.

Programming with Big Data in R using Distributed Memory
(Tutorials: 20h)

These tutorials show how to use the main R packages used in Big Data: some are for parallel computing, some are for working with data sets that are too large to be loaded into memory, some are for Map/Reduce programming, and some are for adding code in C, C++ or Fortran to R. The use of these packages will be illustrated with examples of data sets.

Statistical Libraries for Big Data (Mahout, SAS, HPA)
(Tutorials: 20h)

These tutorials present alternatives to R for Big Data, with commercial solutions (SAS High-Performance Analytics) and Apache Mahout, which is a popular library of machine learning algorithms scalable to large data sets and mainly implemented on Hadoop.

Secure Pairing, Security Services against Piracy, Cryptography
(Lectures: 10h, Tutorials: 20h)

Computer security is currently a particularly hot topic in the news, with the subject of heavy attacks (eg. viruses, intrusion) and e-commerce being most often discussed. The purpose of this course is to introduce the main principles of information security, two of these principles in particular: cryptography, one of the protection tools preventing disclosure, modification or illegitimate data access, and secure pairing, which preserves the anonymity of aggregated data.

Privacy
(Lectures: 5h, Tutorials: 5h)

Privacy is a cornerstone in our digital economy. This course presents the regulations and laws in several countries relative to the privacy protection and personal data access protection. The course will present the technical solutions that make such protections possible.

Big Data Project
(Lectures: 5h, Tutorials: 35h)

This project aims to deepen students' knowledge of topics both previously learned in their courses and new. In small groups, students will work on projects centered on a current issue. They will apply their theoretical knowledge and practical skills to concrete, high-volume data provided by a professional organization. The project culminates in a final report.

Courses for foreigners: Written and/or Oral French Language Courses (at CIREFE)
(Duration: 2 or 4 hours/week over 11 weeks)

Designed for foreign students who are following a full-time academic program in Rennes, these weekly evening courses give students practical written and/or oral French skills, necessary for practical life in France.

End-of-Studies Internship

(Duration: 5 months from May to September)

This final phase of the MSc in Big Data program involves a five-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 and computer science theory and methods that they have learned during the two semesters of coursework.
The student must write an Internship Report and defend it in front of a jury the following Autumn.