Introduction to Big Data

Course title - Intitulé du cours	Introduction to Big Data
Level / Semester - Niveau /semestre	M1 / S2
School - Composante	Ecole d'Economie de Toulouse
Teacher - Enseignant responsable	RUIZ-GAZEN Anne
Other teacher(s) - Autre(s) enseignant(s)	PAINDAVEINE Davy
Other teacher(s) - Autre(s) enseignant(s)	LAURENT Thibault
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Other teacher(s) - Autre(s) enseignant(s)	
Lecture Hours - Volume Horaire CM	12
TA Hours - Volume horaire TD	0
TP Hours - Volume horaire TP	18
Course Language - Langue du cours	English - Anglais
TA and/or TP Language - Langue des TD et/ou TP	English - Anglais

Teaching staff contacts - Coordonnées de l'équipe pédagogique :

The course is made of three parts. The first part (12 hours) is taught by Anne Ruiz-Gazen (professor), office MF206, contact by email (anne.ruiz-gazen@tse-fr.eu) for appointment. The second part (12 hours) is taught by Davy Paindaveine (professor), contact by email (davy.paindaveine@tse-fr.eu). The third part (6 hours) is taught by Thibault Laurent (research engineer), contact by email (thibault.laurent@tse-fr.eu).

Preferred means of interaction : email, prior appointment.

Course Objectives - Objectifs du cours :

This course is particularly relevant for students who are interested in pursuing their studies and career as data scientists. It does not contain advanced theory but all methods and algorithms are described and implemented using *tableau* or *R* and results are analyzed in detail. The course is not difficult but requires much work throughout the semester.

The course is divided into three parts.

The first part presents some descriptive tools using the tableau software and some clustering methods using R. It also presents Discriminant Analysis using R and introduces the Classification and Regression Tree procedure for supervised problems (Principal Components Analysis is assumed to be already known by the students).

The second part of the class describes methods to deal with big data for supervised classification problems. The course mainly covers the bootstrap and bagging approaches (including random forests). The students are expected to develop skills in computational statistics and to be able to combine efficient programming with relevant statistical methods. The class will therefore include a large amount of practical applications.

The third part focuses on an introduction to parallel computing to deal with big data.

Prerequisites - Pré requis :

Proficient R programming, knowledge of descriptive statistics and principal components analysis.

Practical information about the sessions - Modalités pratiques de gestion du cours :

For each of the first two parts, there are 4 weekly sessions of 3 hours. The slides are made available to students but it is highly recommended not to miss any session in order to be able to implement the statistical methods and interpret the results. The third part consists in 2 sessions of 3 hours.

Personal laptops are accepted at the student's own risk (some sessions take place in a computer room). Students are expected to actively participate to the class (that is mostly based on what students are able to program). Late arrivals or missing students will be reported and can result in a grade penalty.

Grading system - Modalités d'évaluation :

Each of the first two parts corresponds to 50% of the final grade.

The first part is evaluated through some Multiple Choice Questionnaires (30%; during the lectures of the first part) and a project (70%; after these four lectures).

The second part is evaluated through some Multiple Choice Questionnaires (30%; during the lectures of the second part) and an exam (70%; after these four lectures).

Bibliography/references - Bibliographie/références :

For the first part:

- An introduction to Applied Multivariate Analysis with R by B. Everitt and T.Hothorn, UseR!, Springer.
- Data Mining with Rattle and R by G. Williams, Use R!, Springer.
- Exploratory Multivariate Analysis by Example using R, by F. Husson, S. Lê, and J. Pagès, Chapman & Hall/CRC Computer Science & Data Analysis.

For the second part:

- An Introduction to Statistical Learning, with Applications in R, by G. James, D. Witten, and T. Hastie and R. Tibshirani, Springer.

Session planning - Planification des séances :

To be determined.

Distance learning – Enseignement à distance :

Distance learning can be provided when necessary by implementing interactive virtual classrooms, MCQ tests and other online exercises / assignments, chatrooms and forums.