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)	VIALANEIX Nathalie
Other teacher(s) - Autre(s) enseignant(s)	
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	18
TP Hours - Volume horaire TP	0
Course Language - Langue du cours	Anglais
TA and/or TP Language - Langue des TD et/ou TP	Anglais

Teaching staff contacts - Coordonnées de l'équipe pédagogique :

The course is made of two parts. The first part is taught by Anne Ruiz-Gazen (professor), office MF206, contact by email (anne.ruiz-gazen@tse-fr.eu) for appointment. The second part is taught by Nathalie Vialaneix (researcher at INRA), contact by email (nathalie.vialaneix@inra.fr).

Modes d'interactions privilégiés : email et / ou demande de rendez-vous.

Course's Objectives - Objectifs du cours :

This course is particularly relevant for students who are interested in pursuing his/her studies and career as a data scientist. 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 a lot of work all along the semester.

The course is divided in two parts.

The first part presents some descriptive tools using the tableau software and some classical unsupervised multivariate statistical techniques such as Principal Components Analysis, Correspondence Analysis and Clustering using R. It also introduces the Classification and Regression Tree procedure for supervised problems.

The second part of the class describes methods to deal with big data for supervised classification problems. The course includes description of the bootstrap and bagging approaches, introduction to parallel computing and an overall description of various strategies (subsampling, divide and conquer, online) to deal with big data. 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

Prerequisites - Pré requis :

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

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

For each of the two parts, there are 5 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.

Personal laptops are accepted at your own risk. 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 part corresponds to 50% of the final grade.

The first part is evaluated through some Multiple Choice Questionnaires (30%) and a project (70%) which will take place on **Saturday February 16 (during the whole day for all students)**.

For the second part, a main exam and a group homework. The final grade will mainly consist into the grade of the main exam which is planned on March 27 in a computer room, plus a bonus or malus depending on the homework quality. Late homeworks are not accepted.

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, J.Pagès, Chapman & Hall/CRC Computer Science & Data Analysis.

For the second part, references are listed on the: http://www.nathalievialaneix.eu/teaching/m1se/

Session planning - Planification des séances :

The first part of the course is organized into 5 sessions of 3 hours from January 9 to February 6. The last 3 sessions take place in a computer room.

The second part of the course is organized into 5 main sessions of 3 hours in a computer room from February 13 to March 20, each one covering a specific topic and including a practical application and a group homework.