

# Coding for Data Science and Data Management

Data Science and Economics (DSE), 2019/2020, The University of Milan

*N. Cesa-Bianchi*<sup>1</sup>, *E. Guidotti*<sup>2</sup>, *S. Montanelli*<sup>3</sup>

## Course Overview

The course aims at providing technical skills about coding/scripting aspects for data analysis and to manage persistent data storage of sources and results involved in analysis. On the one side, the Python programming language and the R framework are illustrated. The goal is to deal with essential notions about data structures and control structures of both Python and R. On the other side, the goal is to present the core notions of relational databases, such as keys, integrity, and primary/foreign key constraints, as well as the SQL language for data definition, manipulation, and query. Recent and innovative NoSQL solutions are also discussed, with special focus on a document-oriented system called MongoDB.

## Course Content (R Module)

### 1. Introduction to the R framework and R Studio

- Installation
- Comments
- Variable assignment
- Functions
- Extension packages
- R help, StackOverflow, Google

### 2. Basic Data Types

- `numeric`
- `integer`
- `complex`
- `logical`
- `character`

---

<sup>1</sup>[nicolo.cesa-bianchi@unimi.it](mailto:nicolo.cesa-bianchi@unimi.it) (Python Module)

<sup>2</sup>[emanuele.guidotti@unine.ch](mailto:emanuele.guidotti@unine.ch) (R Module)

<sup>3</sup>[stefano.montanelli@unimi.it](mailto:stefano.montanelli@unimi.it) (DB Module)

3. **Basic Data Structures**
  - vector
  - matrix
  - data.frame
  - list
  - environment
4. **Basic operations**
  - Subsetting
  - Arithmetics
5. **Time Series**
  - The zoo package
  - The xts package
6. **Control Structures**
  - if-else
  - for loops, while loops, repeat, break
  - apply, lapply, sapply
7. **Custom Functions**
  - Code quality and modularity
  - How to implement R functions
8. **Speeding up the code**
  - The Rcpp package
  - References to parallel computing
9. **Data Acquisition**
  - Local & Remote files
  - R packages
  - RESTful APIs
  - Web scraping
10. **Data visualization**
  - Base plotting system
  - The ggplot2 package
  - The plotly package

### 11. Building interactive interfaces, documents and websites

- R Shiny
- Rmarkdown - Rbookdown - Rblogdown

### 12. Building R packages

- Roxygen
- Base packages
- Rcpp packages
- R Shiny packages

### 13. If time allows...

- Debugging in Rstudio
- Connect R to Python: the `reticulate` package
- Connect R to databases: the `RMySQL`, `RSQLite`, `RMongo` packages
- Q&A

## References (R Module)

1. Introduction to R  
<https://storage.guidotti.dev/tutorial/introduction-to-r.html>
2. Data Acquisition in R  
<https://storage.guidotti.dev/tutorial/data-acquisition-in-r.html>
3. R Plotly  
<https://plot.ly/r/>
4. R Shiny  
<https://shiny.rstudio.com/tutorial/>
5. R Markdown  
<https://bookdown.org/yihui/rmarkdown/>