

# Machine Learning

Rodrigo Fernandes de Mello

Invited Professor at Télécom ParisTech

Associate Professor at Universidade de São Paulo, ICMC, Brazil

<http://www.icmc.usp.br/~mello>

[mello@icmc.usp.br](mailto:mello@icmc.usp.br)



# TP using the R Statistical Software

- You are suggested to start studying R:
  - In French:  
[https://cran.r-project.org/doc/contrib/Paradis-rdebuts\\_fr.pdf](https://cran.r-project.org/doc/contrib/Paradis-rdebuts_fr.pdf)
  - In English:  
[https://cran.r-project.org/doc/contrib/Paradis-rdebuts\\_en.pdf](https://cran.r-project.org/doc/contrib/Paradis-rdebuts_en.pdf)

- Employ K-Means to codify audio signals
  - 1) You start implementing the K-Means algorithm in R
    - **Use a different implementation than the one provided during the class. That will be disconsidered**
  - 2) Then you record some audio signal using any tool
    - We suggest Audacity (<https://www.audacityteam.org/>) given it is an open source platform
    - You must record using the following option:
      - Audacity option “Export Audio”
        - RAW / Unsigned 8-bit PCM
        - At least 8000 observations per second
          - This is the same as 8 kHz of sampling rate
  - 3) Use our C code to translate the binary audio signal to a file with textual integers

- Employ K-Means to codify audio signals
  - 4) Read the files with textual integers in R and apply K-Means
    - using  $k=25, 15$ , and  $7$
  - 5) For every resultant data clustering you should produce a series of values substituting the original observation by its prototype value
    - This makes values assume the central value for every prototype (a.k.a. cluster or group)
  - 6) Now you should round the series and save them into some textual integers file
  - 7) Use our C code to translate those 3 executions of K-Means to a binary file
  - 8) Play the files and listen to them
  - 9) Test the compression:
    - Compress the original and each file with the algorithm you prefer (Zip, GZip, BZip2, etc.) and see the resultant sizes

- Deliverables:
  - A link to a **folder** including:
    - Your R code
    - At least **two** audio files you recorded
    - The output files after processing each of your audio files for:
      - 5 clusters
      - 7 clusters
      - 15 clusters
      - 25 clusters
    - A separated text document in the same **folder** mentioning the compression tool you used (e.g. Bzip2, Gzip2, WinZip, etc) and the resulting compression rate of your audio files
      - For example, let's suppose the original audio file had 1 Mb and that after compression it has 512 Kb
      - Consider your audio file with 5, 7, 15, and 25 clusters have the following sizes after compression: 20 Kb, 50 Kb, 250 Kb and 350 Kb, so that the rates will be:  
 $20\text{Kb}/512\text{Kb}$ ,  $50\text{Kb}/512\text{Kb}$ ,  $250\text{Kb}/512\text{Kb}$ ,  $350\text{Kb}/512\text{Kb}$