Data Mining
Definition and project assignment

1 Name of group components

- Ibáñez Pérez, Raúl
- Lao Tebar, Diego
- Lázaro Costa, Carlos
- López Alcácer, Albert
- Ribes Marzá, Albert
- Roldán Montaner, Carlos

2 Data Source

We got our dataset from the Machine Learning Repository of the Center for Machine Learning And Intelligent Systems (Bren School of Information and Computer Science)
Repository url: `http://archive.ics.uci.edu/ml/datasets/Heart+Disease`
Dataset url: `http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/`

We are going to use the Hungarian, Long Beach and Switzerland non processed .data sources.

3 Process to get data (TO BE DONE)

The process to get data was done by the doctors and analysts from Hungary, Long Beach and Switzerland hospitals.

The data is a copy of the patient medical results donated by different medical centers. All the centers used the same method and representation to keep the data in order to maintain the data consistency.

4 What data is about (NEEDS REVISING)

Our data is a merge of four different datasets concerning heart disease diagnosis. Each instance collects different conditions or variables (numerical, binary and qualitative) of a patient which can be used to predict the presence of a heart disease in that patient.

The locations the data was collected from are:

- Cleveland Clinic Foundation (The data is corrupted - Discarded)
- Hungarian Institute of Cardiology, Budapest
- V.A. Medical Center, Long Beach, CA
- University Hospital, Zurich, Switzerland

The principal investigator responsible for the data collection are:

- Andras Janosi, M.D.
- William Steinbrunn, M.D.
- Matthias Pfisterer, M.D.
- Robert Detrano, M.D., Ph.D.

The Cleveland data is corrupted, so we discarded it.

5 Structure of data matrix (TO BE COMPLETED)

- **Number of records**:
  - ~~Cleveland: 303~~ (Discarded)
  - Hungarian: 294
  - Switzerland: 123
  - Long Beach VA: 200
  - **Total: 617**
- **Number of variables**: 76 (including the predicted one)
- **Number of numerical variables**:
- **Number of binary variables**:
- **Number of qualitative variables**:
- **Number and % of missing data per each variable**:
- **% of missing data in the whole data matrix**: 33,84%