Data Mining
Definition and project assignment

1 Name of group components

- Ibáñez Pérez, Raul
- Lao Tebar, Diego
- Lázaro Costa, Carles
- 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)

4 What data is about (NEEDS REVISING)

Our data is a merge of four different datasets concerning heart disease diagnosis. 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**: