

# 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

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

Each data record is a copy of the patient medical results. All the centers used the same unit representation for the medical analysis to keep the data.

### 4 What data is about

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 data was collected from:

- Hungarian Institute of Cardiology, Budapest by Andras Janosi, M.D.
- University Hospital, Zurich, Switzerland by William Steinbrunn, M.D.
- University Hospital, Basel, Switzerland by Matthias Pfisterer, M.D.
- V.A. Medical Center, Long Beach and Cleveland Clinic Foundation (The data is corrupted - Discarded) by Robert Detrano, M.D., Ph.D.

The Cleveland data is corrupted, so we discarded it.

## 5 Structure of data matrix

- **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:** 55
- **Number of binary variables:** 15
- **Number of qualitative variables:** 6
- **Number and % of missing data per each variable:**
  - *ITEM #Entities MissPercentage*
  - ID 0 0%
  - CCF 0 0%
  - AGE 0 0%
  - SEX 0 0%
  - PAINLOC 0 0%
  - PAINEXER 0 0%
  - RESTREL 4 0,64829822%
  - PNCADEN 617 100%
  - CP 0 0%
  - TRESTBPS 59 9,5623987%
  - HTN 34 5,51053485%
  - CHOL 30 4,86223663%
  - SMOKE 387 62,72285251%
  - CIGS 415 67,26094003%
  - YEARS 427 69,20583468%
  - FBS 90 14,58670989%
  - DM 545 88,33063209%
  - FAMHIST 422 68,39546191%
  - RESTECG 2 0,32414911%
  - EKGMO 53 8,58995138%

- EKGDAY 54 8,75202593%
- DIG 53 8,58995138%
- PROP 66 10,69692058%
- NITR 64 10,37277147%
- PRO 63 10,21069692%
- DIURETIC 61 9,88654781%
- PROTO 80 12,96596434%
- PROTO 112 18,15235008%
- THALDUR 56 9,07617504%
- THALTIME 384 62,23662885%
- MET 105 17,0178282%
- THALACH 55 8,91410049%
- THALREST 56 9,07617504%
- TPEAKBPS 63 10,21069692%
- TPEAKBPD 63 10,21069692%
- DUMMY 59 9,5623987%
- TRESTBPD 59 9,5623987%
- EXANG 55 8,91410049%
- XHYPO 58 9,40032415%
- OLDPEAK 62 10,04862237%
- SLOPE 308 49,91896272%
- RLDV5 143 23,17666126%
- RLDV5E 142 23,01458671%
- CA 606 98,2171799%
- RESTCKM 617 100%
- EXERCKM 616 99,83792545%
- RESTEF 589 95,46191248%
- RESTWM 587 95,13776337%
- EXEREF 615 99,67585089%
- EXERWM 612 99,18962723%
- THAL 475 76,98541329%
- THALSEV 487 78,93030794%
- THALPUL 573 92,86871961%
- EARLOBE 616 99,83792545%
- CMO 11 1,7828201%
- CDAY 9 1,45867099%

- CYR 9 1,45867099%
- NUM 0 0%
- LMT 275 44,57050243%
- LADPROX 236 38,24959481%
- LADDIST 246 39,87034036%
- DIAG 276 44,73257699%
- CXMAIN 235 38,08752026%
- RAMUS 285 46,19124797%
- OM1 271 43,92220421%
- OM2 290 47,00162075%
- RCAPROX 245 39,7082658%
- RCADIST 270 43,76012966%
- LVX1 19 3,07941653%
- LVX2 19 3,07941653%
- LVX3 19 3,07941653%
- LVX4 19 3,07941653%
- LVF 16 2,59319287%
- CATHEF 306 49,59481361%
- JUNK 498 80,71312804%
- NAME 617 100%
- **% of missing data in the whole data matrix: 33,84%**