ECG Heartbeat Categorization Dataset
(From Kaggle)

The Data

- 2 distinct datasets:
  - Arrythmia: 5 categories, 109446 observations
    - 1 observation = (1x187) vector + class
    - Train: (72471, 2223, 5788, 641, 6431) => 87554
    - Test: (18118, 556, 1448, 162, 1608) => 21892
  - PTB: 2 categories, 14552 observations
    - 1 observation = (1x187) vector + class
    - 10506 normal, 4046 abnormal
- Unbalanced data

OBJECTIVES

- Build a classification model for these datasets

- *Bonus 1:* See if the model performs well when trained on one dataset and tested on the other
- *Bonus 2:* Try to build an unsupervised clustering to see if we can spot the 4 pathologies in the first dataset

STRATEGY

- Start with a baseline classifier, using KNN
- Neural network
- SVM + kernel trick

Compare those 2

- Metric for KNN: L2 norm
- Tuning the parameters:
  - Train/Dev/Test sets for the big dataset
  - Cross validation for the other one