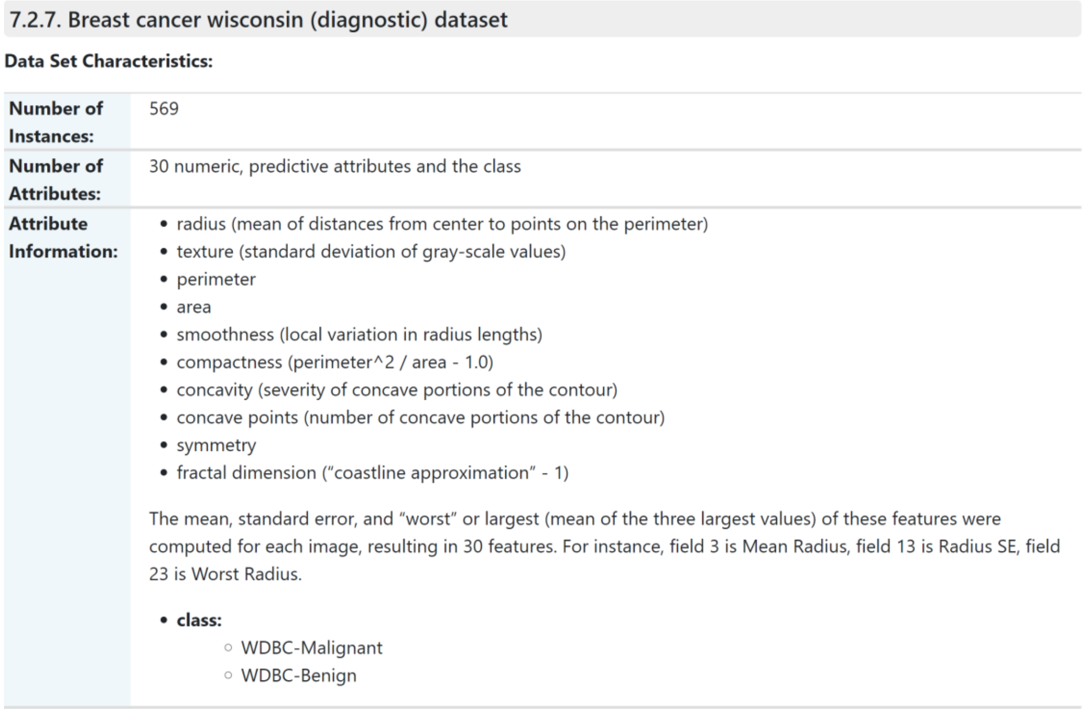
Challenge 1: Goal

- A Machine Learning system is based on these steps:
  1. Understand the problem and get the data
  2. Understand the data and analyze the features
  3. Train the model
  4. Evaluate the performance
- In this challenge we will focus on phase 2
- The goal is to learn the basics of Scikit-learn library focusing on feature engineering, which we analyzed in the theoretical lectures

Challenge 1: Material

- For this challenge, we use a very famous toy dataset, the "Breast Cancer Wisconsin (Diagnostic)" dataset available at:
  - UCI Machine Learning Repository
    - https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
  - Scikit-learn
    - https://scikit-learn.org/stable/datasets/index.html



7.2.7. Breast cancer wisconsin (diagnostic) dataset

Data Set Characteristics:

Number of Instances:	569
Number of Attributes:	30 numeric, predictive attributes and the class
Attribute Information:	• radius (mean of distances from center to points on the perimeter) • texture (standard deviation of gray-scale values) • perimeter • area • smoothness (local variation in radius lengths) • compactness (perimeter^2 / area - 1.0) • concavity (severity of concave portions of the contour) • concave points (number of concave portions of the contour) • symmetry • fractal dimension ("coastline approximation" - 1) The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. • **class:** ○ WDBC-Malignant ○ WDBC-Benign

Challenge 1: Method and evaluation

- Study the features and apply, if necessary:
  - transformations
  - dimensionality reduction
  - selection
- Train and test the given classifier model
- Which accuracy could you reach just acting on features without modifying the model?