# Spark - Exercises

# Exercise #66

- Predict the variety of iris plants in real-time
- Inputs:
  - Training data
    - Static input file: training.csv
  - New data
    - A stream of records about new iris plants
- Output
  - Predicted class label/variety for each new iris plant using only column “sepallength” and “sepalwidth”

# Exercise #66

- Training data schema
  - sepallength: double
  - sepalwidth: double
  - petallength: double – not considered
  - petalwidth: double – not considered
  - variety: integer
    - 1 -> Setosa category
    - 2 -> Versicolor category
    - 3 -> Virginica category

# Exercise #66

- New streaming data schema
  - sepallength: double
  - sepalwidth: double
  - petallength: double
  - petalwidth: double
  - variety: it is always null for the new data

# Exercise #66

- Example training data

  sepallength,sepalwidth,petallength,petalwidth,variety

  5.1,3.5,1.4,0.2,1

- Example new streaming data

  5.2,3.2,1.4,0.2,