MLlib and spark.ml

Spark's machine learning libraries are divided into two packages, MLlib and spark.ml

- MLlib is older, and is built on top of RDDs
- spark.ml is built on top of DataFrames, and can be used to construct ML pipelines
- In cases where MLlib and spark.ml offer equivalent functionality, this course will focus on spark.ml

Feature vectors for supervised learning in MLlib and spark.ml

- MLlib
  - The models in MLlib are designed to work with `RDD[LabeledPoint]` objects, which associates labels with feature vectors
- spark.ml
  - The models in spark.ml are designed to work with `DataFrames`
  - A basic spark.ml `DataFrame` will (by default) have two columns:
    - a label column (default name: "label")
    - a features column (default name: "features")