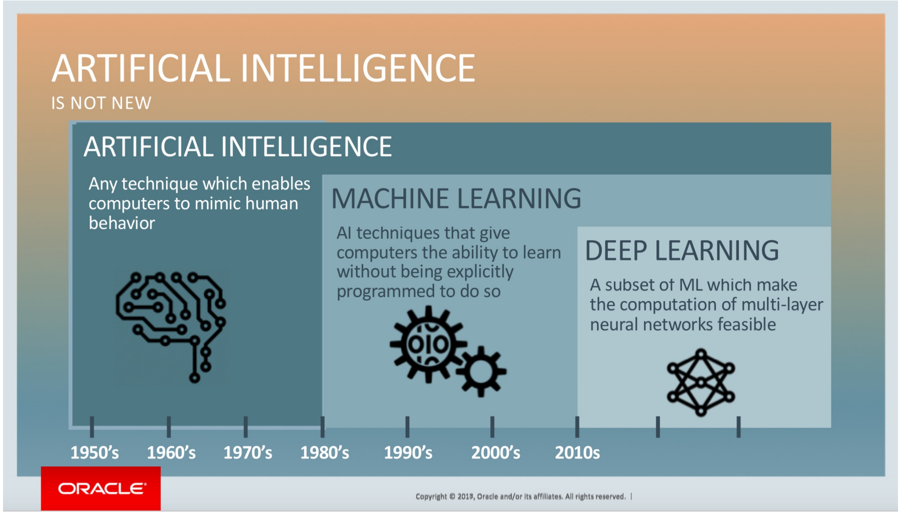
Big Data Processing

1/2023
Thanh-Chung Dao Ph.D.

Machine learning



ARTIFICIAL INTELLIGENCE
IS NOT NEW

ARTIFICIAL INTELLIGENCE
Any technique which enables computers to mimic human behavior

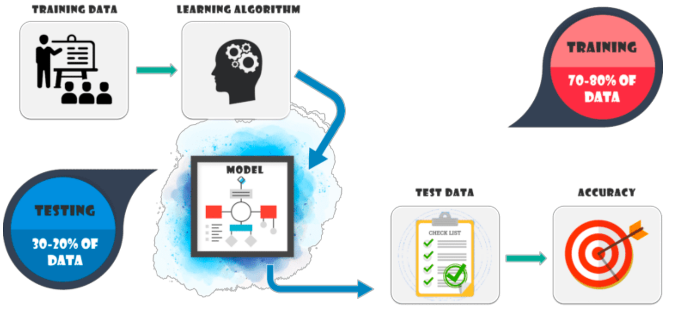
MACHINE LEARNING
AI techniques that give computers the ability to learn without being explicitly programmed to do so

DEEP LEARNING
A subset of ML which make the computation of multi-layer neural networks feasible

From [1]

Machine Learning Lifecycle

- Two major phases
  - **Training Set**
    - You have the complete training dataset
    - You can extract features and train to fit a model.
  - **Testing Set**
    - Once the model is obtained, you can predict using the model obtained on the training set



TRAINING DATA

LEARNING ALGORITHM

MODEL

TRAINING 70-80% OF DATA

TESTING 30-20% OF DATA

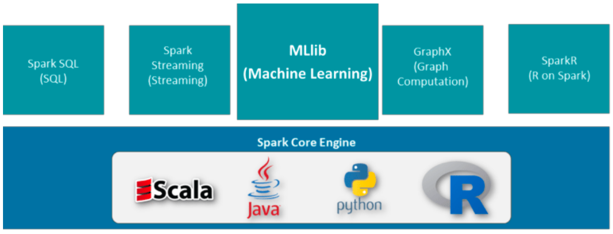
TEST DATA

ACCURACY

From [2]

Spark ML and PySpark

- Spark ML is a machine-learning library
  - Classification: logistic regression, naive Bayes
  - Regression: generalized linear regression, survival regression
  - Decision trees, random forests, and gradient-boosted trees
  - Recommendation: alternating least squares (ALS)
  - Clustering: K-means, Gaussian mixtures (GMMs)
  - Topic modeling: latent Dirichlet allocation (LDA)
  - Frequent item sets, association rules, and sequential pattern mining
- PySpark is an interface for using Python



Spark SQL (SQL)

Spark Streaming (Streaming)

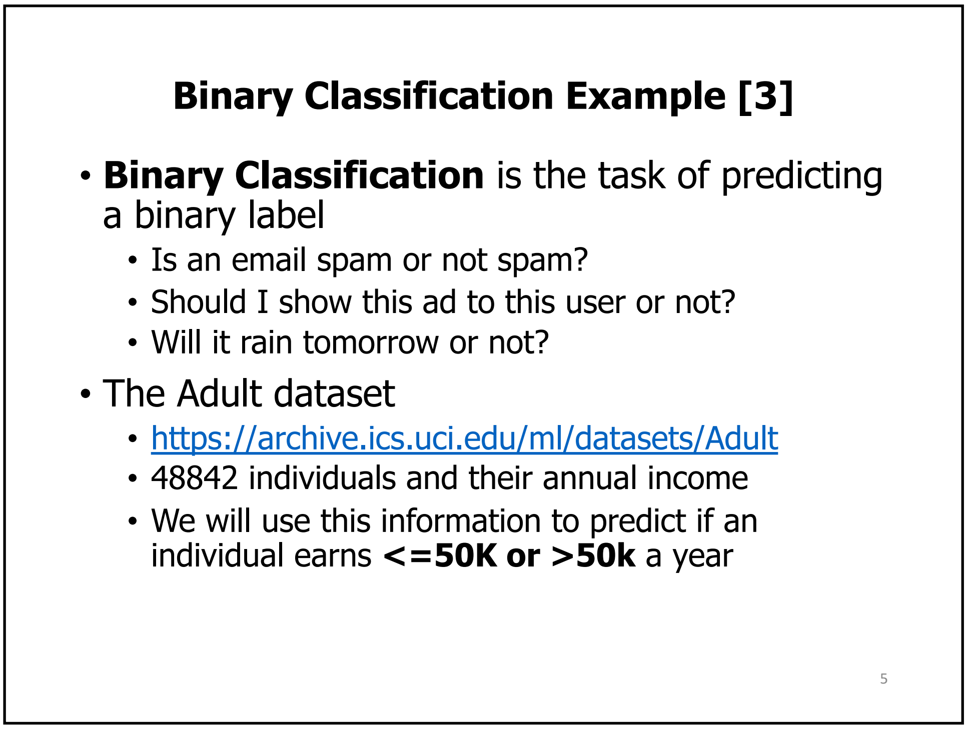
MLlib (Machine Learning)

GraphX (Graph Computation)

SparkR (R on Spark)

Spark Core Engine

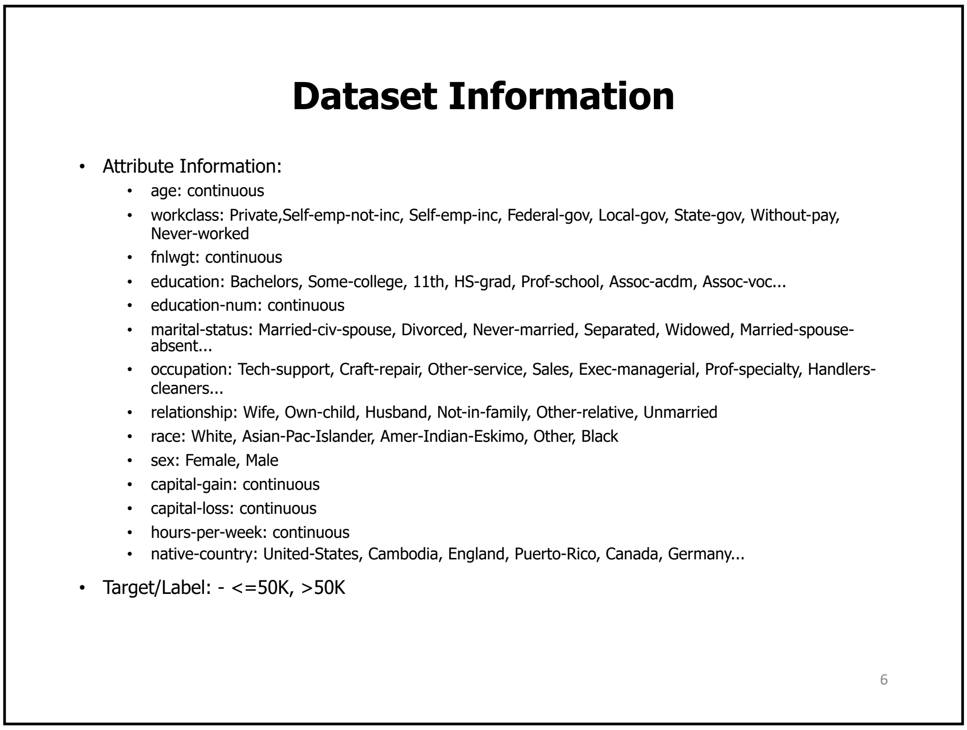
From [2]



Binary Classification Example [3]

- **Binary Classification** is the task of predicting a binary label
  - Is an email spam or not spam?
  - Should I show this ad to this user or not?
  - Will it rain tomorrow or not?
- The Adult dataset
  - https://archive.ics.uci.edu/ml/datasets/Adult
  - 48842 individuals and their annual income
  - We will use this information to predict if an individual earns **<=50K or >50k** a year

5



Dataset Information

- Attribute Information:
  - age: continuous
  - workclass: Private,Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked
  - fnlwgt: continuous
  - education: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc...
  - education-num: continuous
  - marital-status: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent...
  - occupation: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners...
  - relationship: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
  - race: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
  - sex: Female, Male
  - capital-gain: continuous
  - capital-loss: continuous
  - hours-per-week: continuous
  - native-country: United-States, Cambodia, England, Puerto-Rico, Canada, Germany...
- Target/Label: - <=50K, >50K

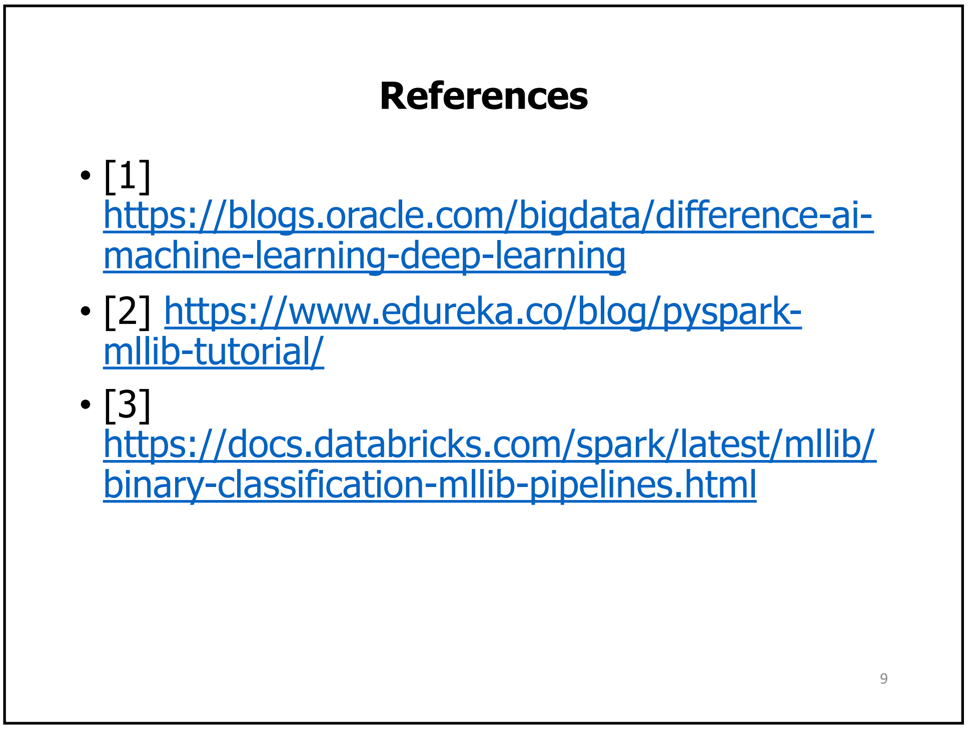
6

Analyzing Flow

- Load data
- Preprocess Data
- Fit and Evaluate Models
  - Logistic Regression
  - Decision Trees
  - Random Forest
- Make Classification

Lab: Running Binary Classification on Zeppelin

- Get the prepared notebook
- Run and try to understand algorithms



References

- [1] https://blogs.oracle.com/bigdata/difference-ai-machine-learning-deep-learning
- [2] https://www.edureka.co/blog/pyspark-mllib-tutorial/
- [3] https://docs.databricks.com/spark/latest/mllib/binary-classification-mllib-pipelines.html