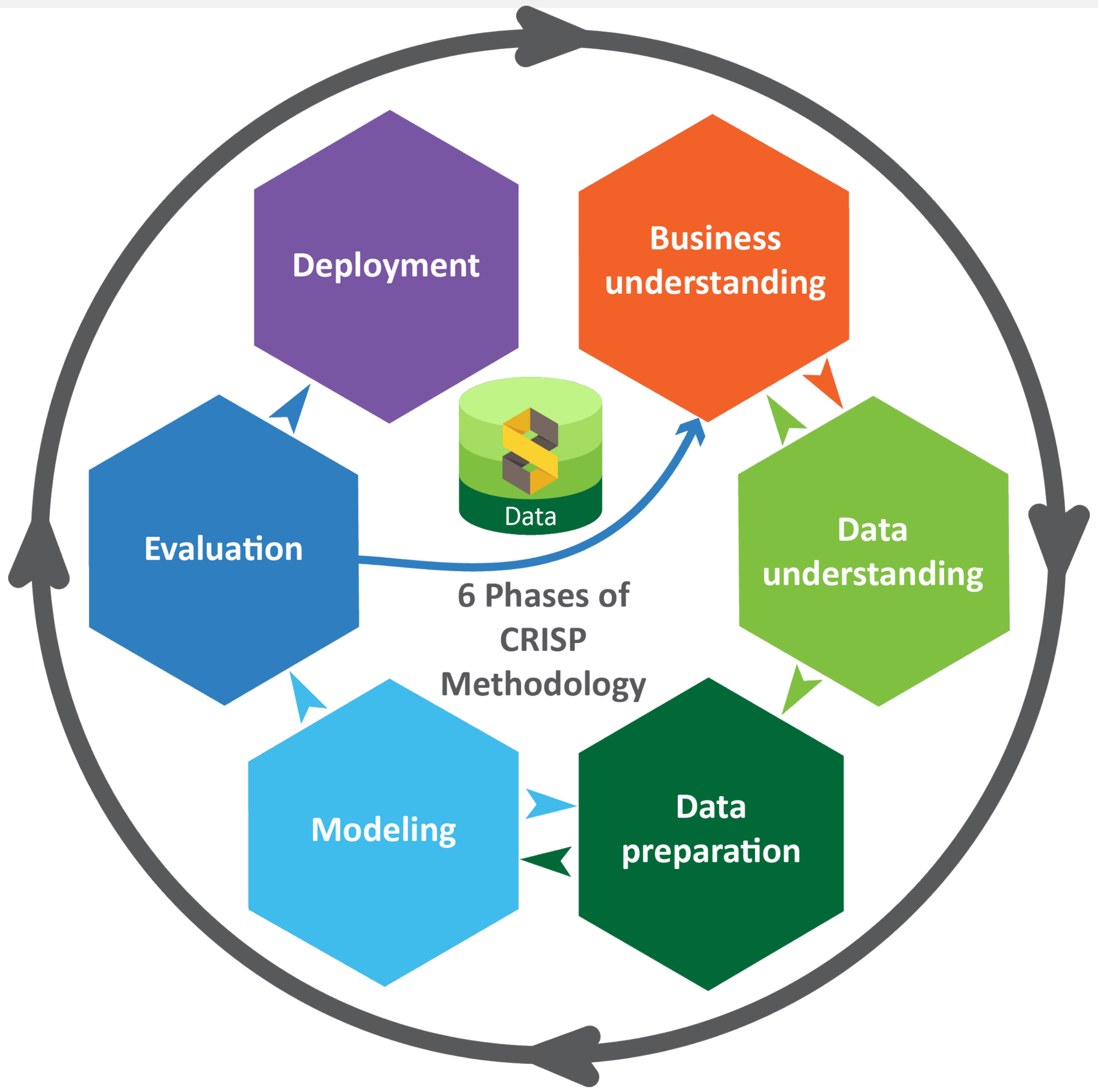
(Machine Learning)



Laboratory 10 Project Specification

Angelica Liguori

Introduction

- Predictive models can have life-changing effects on individuals in certain situations
  - E.g., in the United States, recidivism prediction models, such as the COMPAS score, are used to guide sentencing for crimes in several states and major cities

- Project is about **COMPAS RECIDIVISM DATASET:** https://github.com/propublica/compas-analysis

- The **goal** is to understand whether a defendant had reoffended after the arrest or not

Instructions

- Analyze the dataset applying what you learnt during the course, trying to achieve your goals
    - You can use any kind of analysis tool the best fits your needs

Instructions

- Analyze the dataset applying what you learnt during the course, trying to achieve your goals
  - You can use any kind of analysis tool the best fits your needs

- Produce 2 documents:

Instructions

- Analyze the dataset applying what you learnt during the course, trying to achieve your goals
  - You can use any kind of analysis tool the best fits your needs

- Produce 2 documents:
  - A **CRISP-DM documentation** (.doc, .docx or .pdf), max 25 pages

Instructions

- Analyze the dataset applying what you learnt during the course, trying to achieve your goals
  - You can use any kind of analysis tool the best fits your needs

- Produce 2 documents:
  - A **CRISP-DM documentation** (.doc, .docx or .pdf), max 25 pages
  - A **Presentation** (.ppt, .pptx or .pdf)

Instructions

- Analyze the dataset applying what you learnt during the course, trying to achieve your goals
  - You can use any kind of analysis tool the best fits your needs

- Produce 2 documents:
  - A **CRISP-DM documentation** (.doc, .docx or .pdf), max 25 pages
  - A **Presentation** (.ppt, .pptx or .pdf)

- The maximum score you can achieve in this phase is 25/30
- If the project will be approved, you will have the oral proof

Documentation

Example

Business Understanding

Determine Business Objectives
Background
Business Objectives
Business Success Criteria

Assess Situation
Inventory of resources
Requirements, Assumptions and Constraints
Risks and Contingencies
Terminology
Costs and Benefits

Determine Data Mining Goals
Data Mining Goals
Data Mining Success Criteria

Produce Project Plan
Project Plan
Initial Assessment of Tools and Techniques

Data Understanding

Collect Initial Data
Initial Data Collection Report

Describe Data
Data Description Report

Explore Data
Data Exploration Report

Verify Data Quality
Data Quality Report

Data Preparation

Data Set
Data Set Description

Select Data
Rationale for Inclusion/Exclusion

Clean Data
Data Cleaning Report

Construct Data
Derived Attributes
Generated Records

Integrate Data
Merged Data

Format Data
Reformatted Data

Modeling

Select Modeling Technique
Modeling Technique
Modeling Assumptions

Generate Test Design
Test Design

Build Model
Parameter Settings
Models
Model Description

Assess Model
Model Assessment
Revised Parameter Settings

Evaluation

Evaluate Results
Assessment of Data Mining Results w.r.t. Business Success Criteria
Approved Models

Review Process
Review of Process

Determine Next Steps
List of Possible Actions
Decision

Deployment

Plan Deployment
Deployment Plan

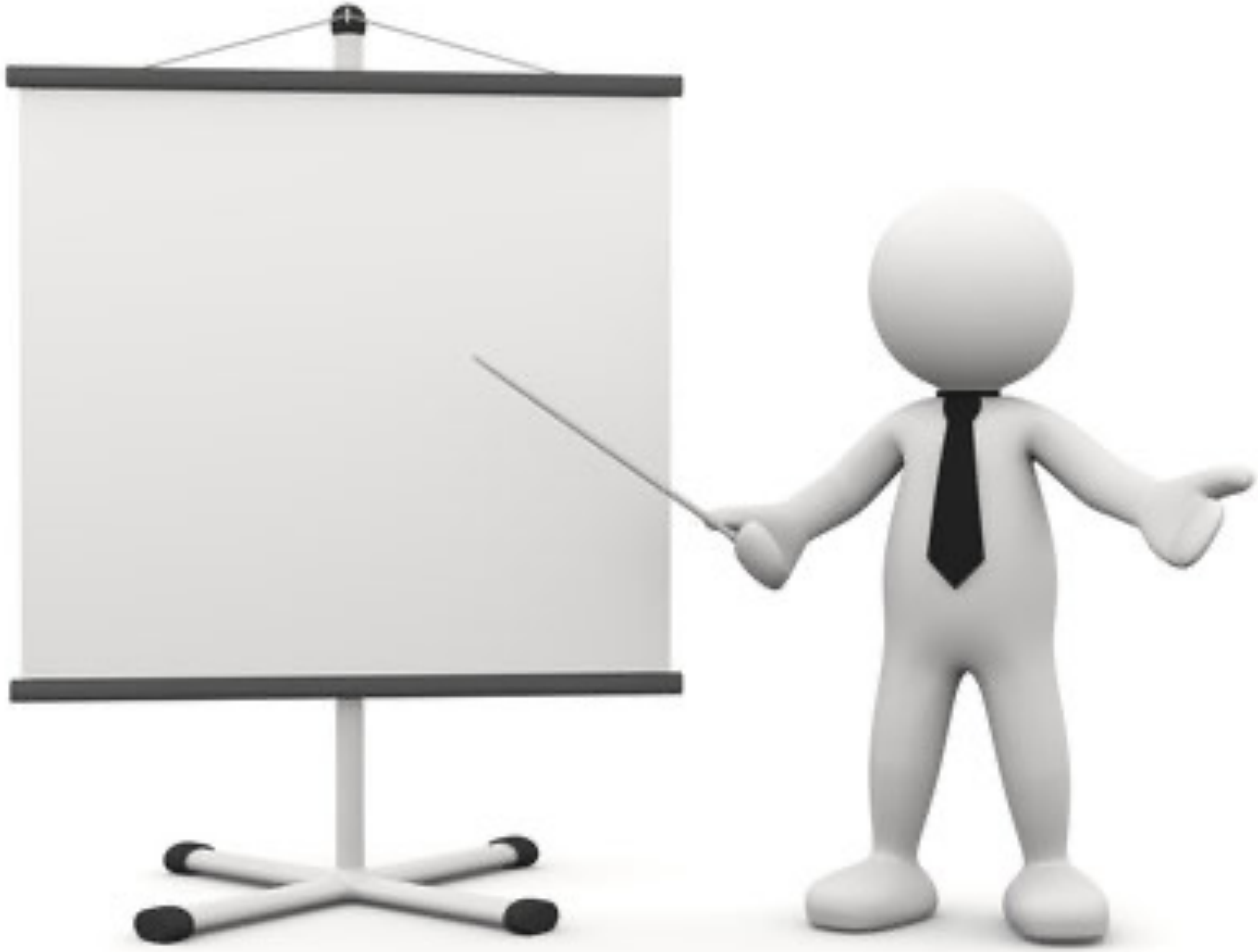
Plan Monitoring & Maintenance
Monitoring and Maintenance Plan

Produce Final Report
Final Report
Final Presentation

Review Project
Experience documentation

Presentation

- The presentation is a summary of what you did during the dataset analysis (**once again you can take inspiration from the CRISP-DM methodology**)
    - Don't forget any step
- It must last **15** minutes at most
- Focus on your analysis



General Information

- The project will last 1 year
  - You can deliver it whenever you are ready
  - Anyway, 2 week before the exam

- Official notifications will be provided in official exam periods

Information about the dataset

- The COMPAS dataset consists of several arrests logged in Broward County, Florida and contains different features describing the demographics and criminal history of the defendants

- Dataset Characteristic: Multivariate
- Number of instances: 11757
- Number of attributes: 47
- Missing value? Yes