Berkeley Engineering | BerkeleyHaas

PROFESSIONAL CERTIFICATE IN MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE

Module 3
Introduction to Data Analysis

Office Hours with Viviana Márquez
September 14, 2023

AGENDA

- Slack
- Expectations
- Required activities for Module 3
- Content review Module 3: Introduction to Data Analytics
- Questions

Slack

#cohort-august-2023

Slack Workspace Invitation

Expectations

- Content is released every Wednesday
  - I will provide you with a content review and share industry insights
- Please submit a ticket for questions that are unique to you
- In order to make the most of our time together, I encourage you to let me know in advance if you have questions about a specific activity

- Everyone can come to any of the Office Hours (not mandatory but highly encouraged!). I will be answering questions/grading assignments for **Section A**

Office hours: Selected Thursdays at 4 PM UTC (5 PM London / 9 AM California)
Always check Canvas for the most up-to-date information!
Tool to convert to your timezone: https://www.worldtimebuddy.com/

Required Activities for Module 3

- Discussion 3.1: THE APPLICATIONS OF THE DATA SCIENCE LIFECYCLE
- 🤖 Codio Activity 3.1: PANDAS BASICS
- 🤖 Codio Activity 3.2: THE BASICS OF DATA VISUALIZATION
- 🤖 Codoi Activity 3.3 : REPLICATING DATA VISUALIZATIONS
- Try-It Activity 3.1: CREATING DATA VISUALIZATIONS
- 🤖 Codio Activity 3.4: AGGREGATION OPERATIONS
- 🤖 Codio Activity 3.5: SORTING AND AGGREGATING
- 🤖 Codio Activity 3.6: INDEXING
- 🤖 Codio Activity 3.7: Filtering
- 🤖 Codio Activity 3.8: COMBINING DATA ANALYSIS TECHNIQUES
- 🤖 Quiz 3.1: INTRODUCTION TO DATA ANALYSIS

Content review Module 3: Introduction to Data Analysis

- The Data Science Lifecycle
- Pandas
- Visualization libraries
- Data aggregation
- Data sorting
- Data indexing
- Data filtering

Content review Module 3: Introduction to Data Analysis

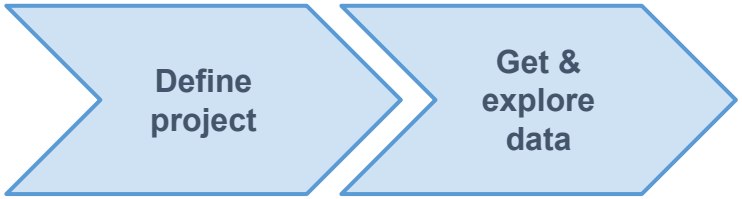
- The Data Science Lifecycle
- Pandas
- Visualization libraries
- Data aggregation
- Data sorting
- Data indexing
- Data filtering

The Data Science Lifecycle

Define project

- Specify business problem
- Acquire domain knowledge

The Data Science Lifecycle



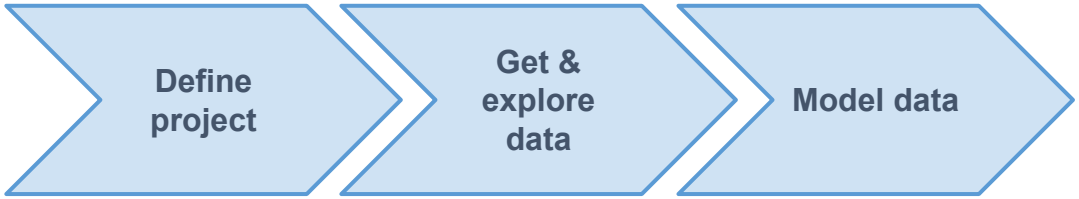
Define project

- Specify business problem
- Acquire domain knowledge

Get and explore data

- Find appropriate data
- Exploratory Data Analysis
- Clean and pre-process data
- Feature engineering

The Data Science Lifecycle



Define project

- Specify business problem
- Acquire domain knowledge

Get and explore data

- Find appropriate data
- Exploratory Data Analysis
- Clean and pre-process data
- Feature engineering

Model data

- Determine ML task
- Build candidate models
- Select model based on performance metrics

The Data Science Lifecycle

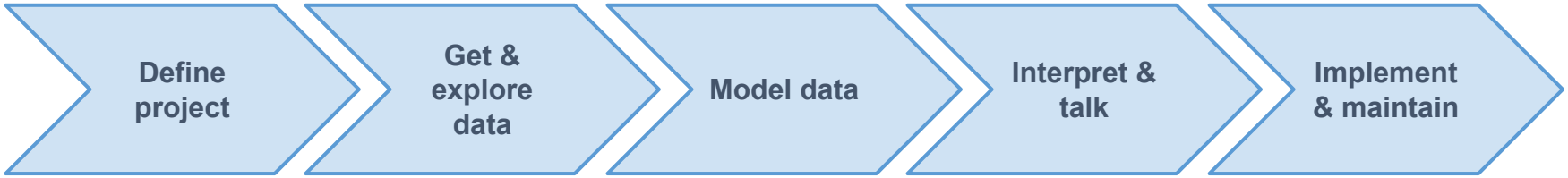
Define project
- Specify business problem
- Acquire domain knowledge

Get and explore data
- Find appropriate data
- Exploratory Data Analysis
- Clean and pre-process data
- Feature engineering

Model data
- Determine ML task
- Build candidate models
- Select model based on performance metrics

Interpret & talk
- Interpret model
- Communicate model insights

The Data Science Lifecycle



Define project
- Specify business problem
- Acquire domain knowledge

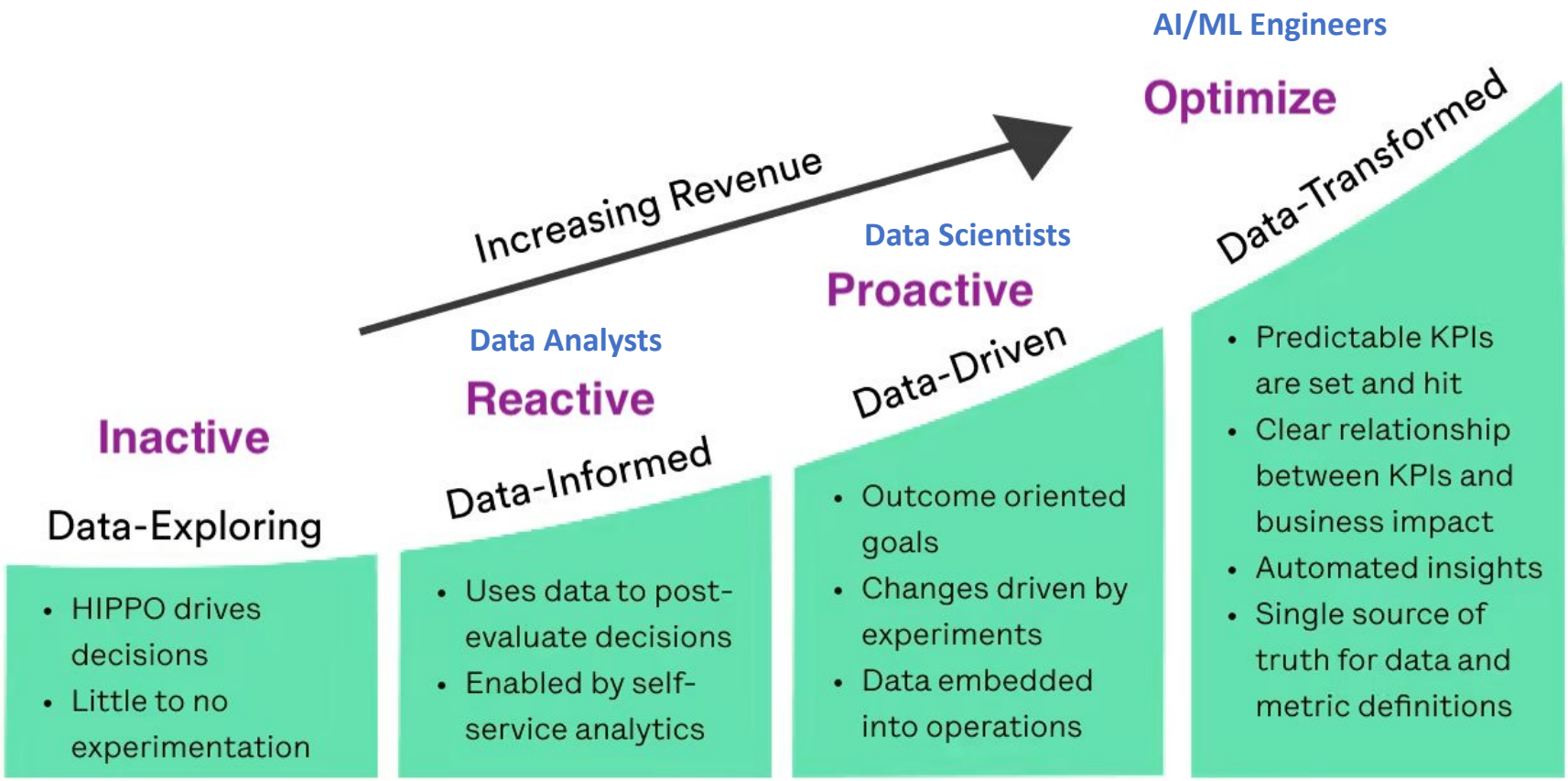
Get and explore data
- Find appropriate data
- Exploratory Data Analysis
- Clean and pre-process data
- Feature engineering

Model data
- Determine ML task
- Build candidate models
- Select model based on performance metrics

Interpret & talk
- Interpret model
- Communicate model insights

Implement & maintain
- Set up function to predict on new data
- Document process
- Monitor and maintain model

Data maturity at an organization

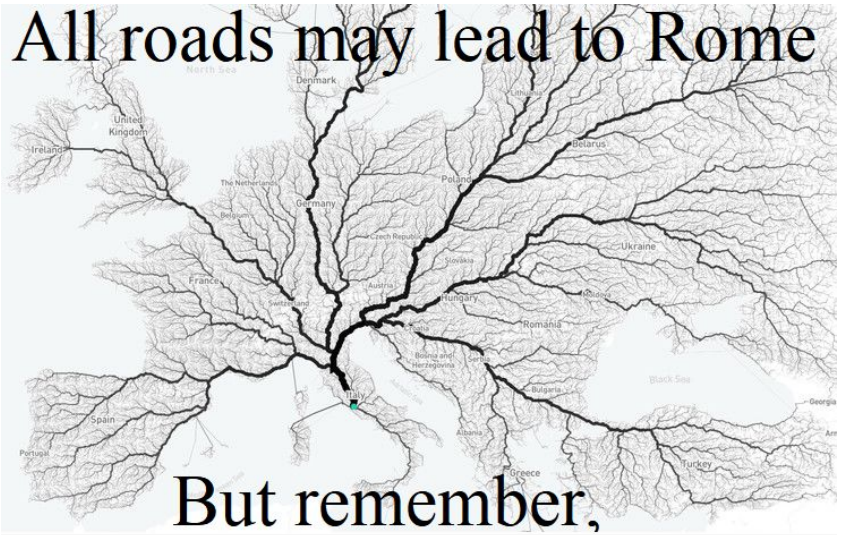


Content review Module 3: Introduction to Data Analysis

- ✅ The Data Science Lifecycle
- Pandas
- Visualization libraries
- Data aggregation
- Data sorting
- Data indexing
- Data filtering

Content review Module 3: Introduction to Data Analysis

- ✅ The Data Science Lifecycle
- Pandas
- Visualization libraries
- Data aggregation
- Data sorting
- Data indexing
- Data filtering



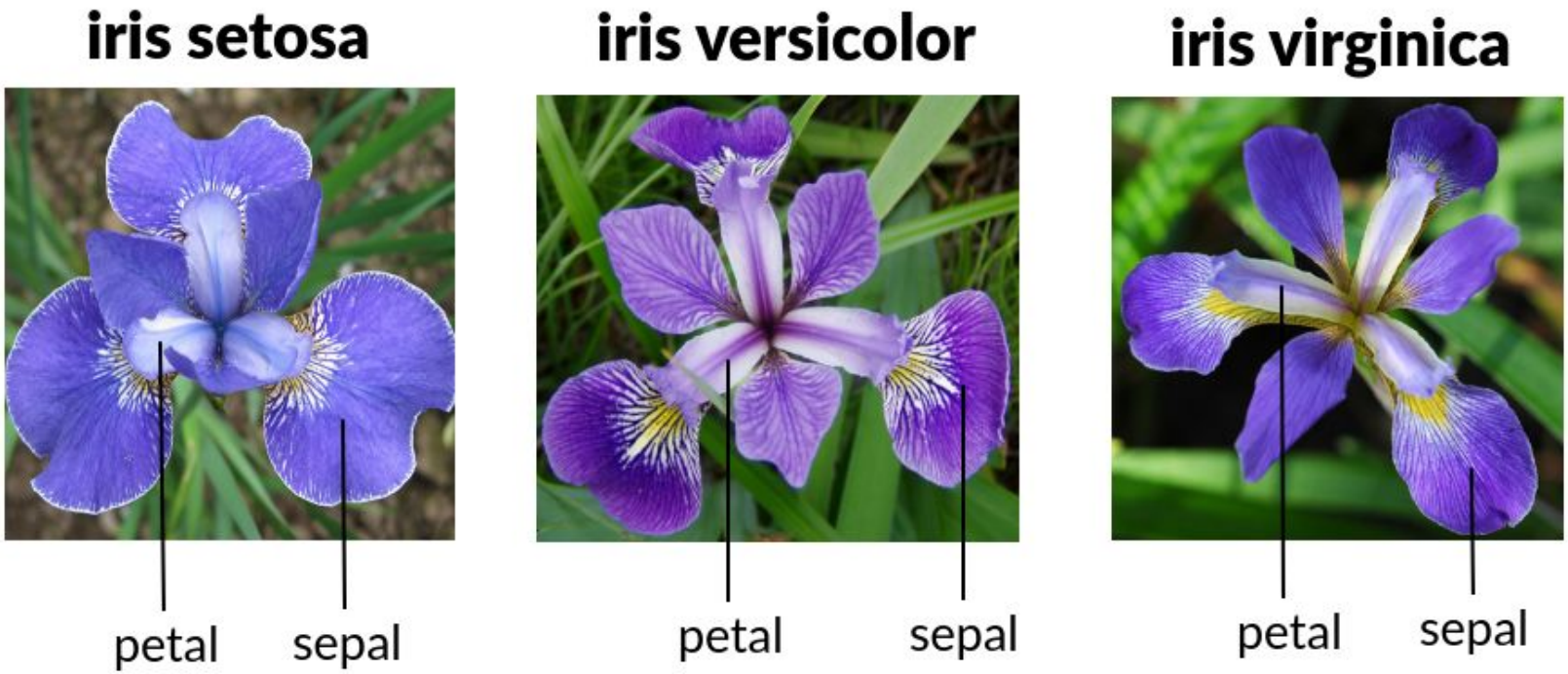
Notebook time!

[12]

```python
# Method 1: Direct Pandas method
mean_1 = iris['sepal length (cm)'].mean()

# Method 2: Using numpy on the Pandas Series
import numpy as np
mean_2 = np.mean(iris['sepal length (cm)'])

# Method 3: Manual calculation using Pandas
mean_3 = iris['sepal length (cm)'].sum() / len(iris

print(mean_1, mean_2, mean_3)
```

```
5.843333333333334 5.843333333333334 5.843333333333334
```

Content review Module 3: Introduction to Data Analysis



QUESTIONS?