Advanced Data Analysis

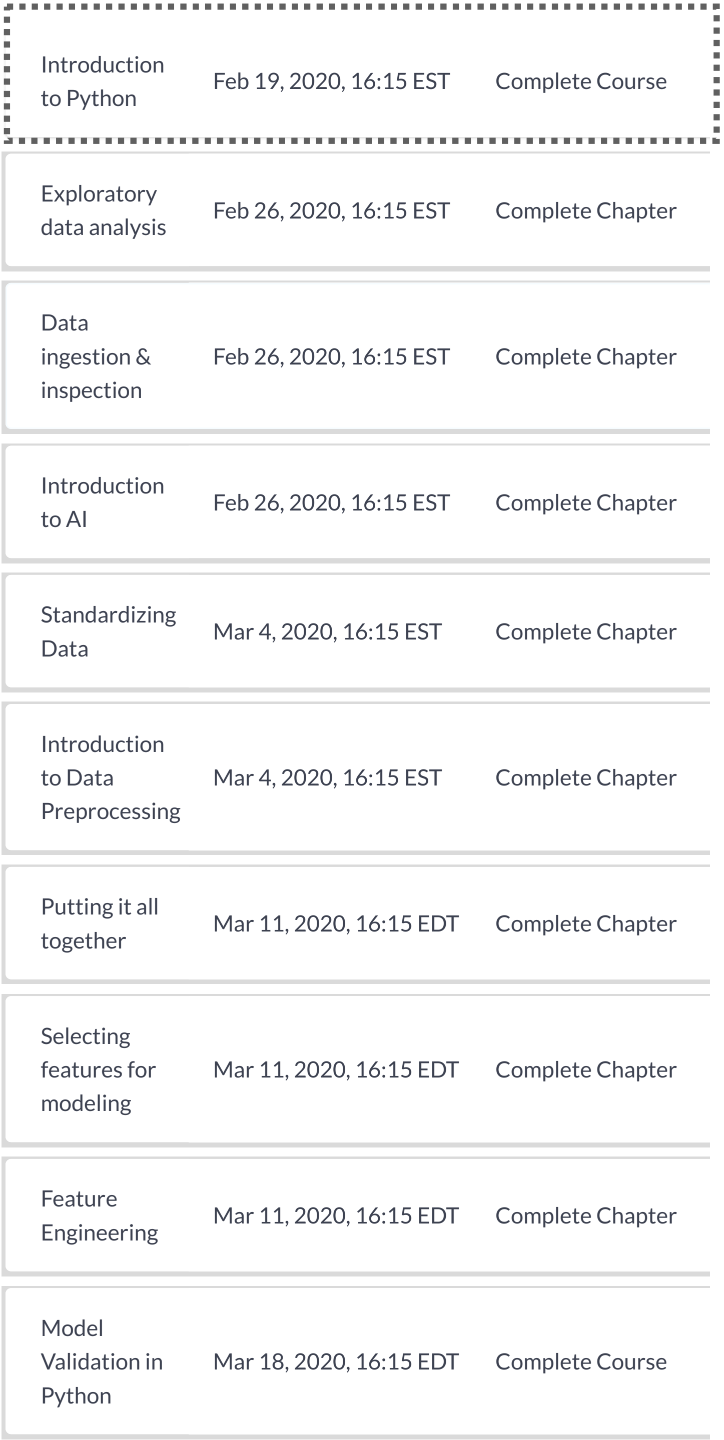
DATA 71200

Class 3

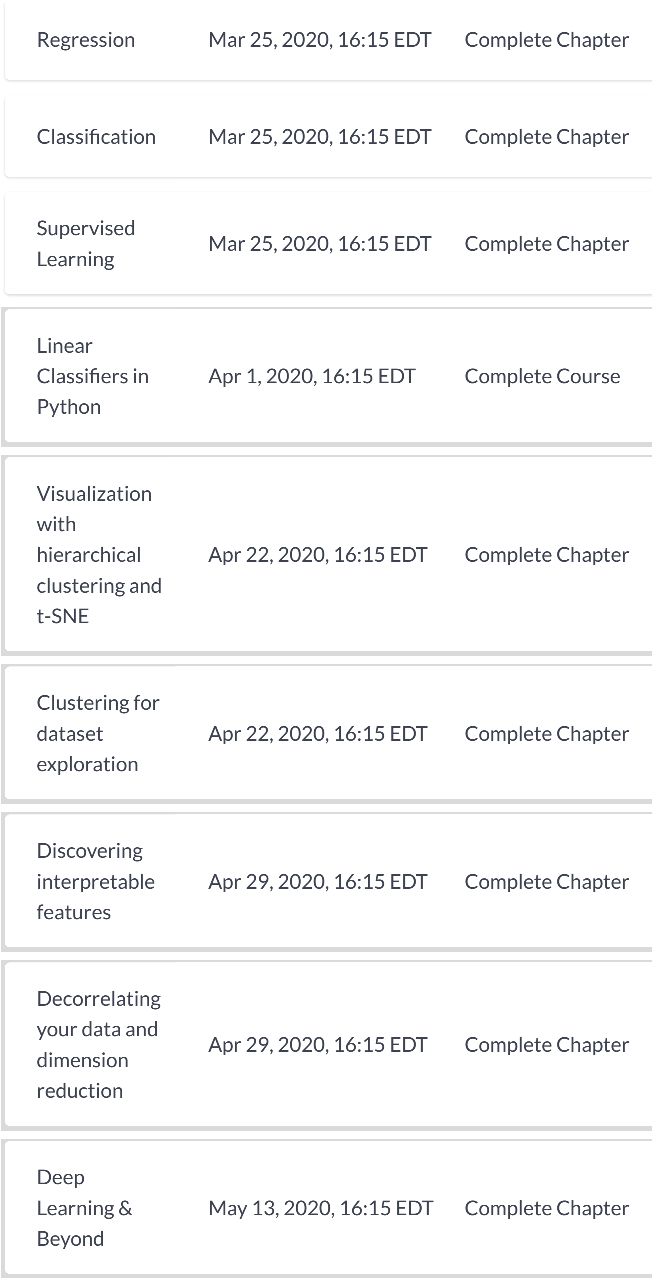
Course Schedule

19-Feb	**Getting Started with Machine Learning**
26-Feb	Inspecting Data
4-Mar	Representing Data
11-Mar	Evaluation Methods
18-Mar	Supervised Learning (k-Nearest Neighbors, Linear Models) – *Project 1 Due*
25-Mar	Supervised Learning (Naive Bayes Classifiers and Decision Trees)

DataCamp

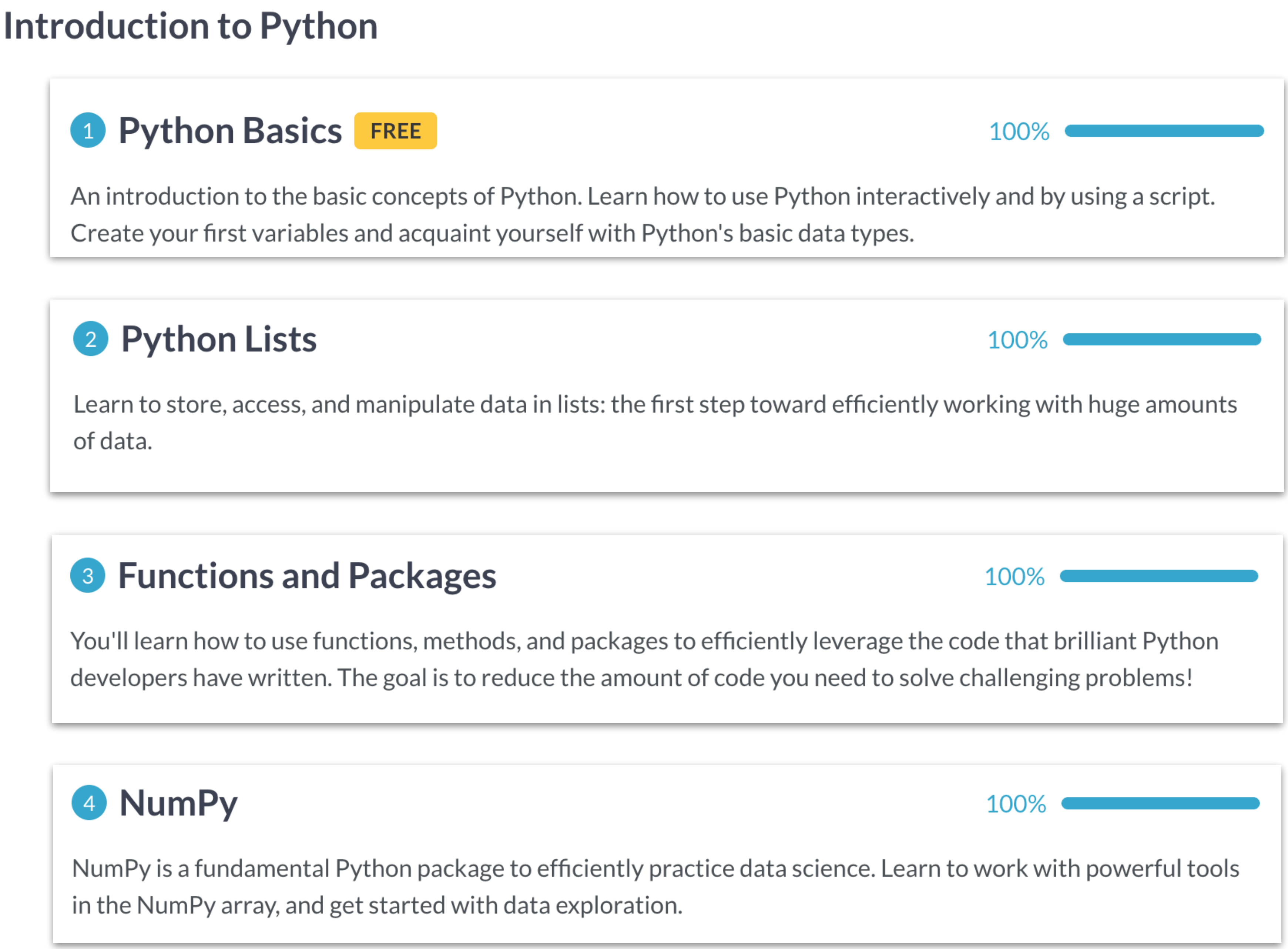


Introduction to Python	Feb 19, 2020, 16:15 EST	Complete Course
Exploratory data analysis	Feb 26, 2020, 16:15 EST	Complete Chapter
Data ingestion & inspection	Feb 26, 2020, 16:15 EST	Complete Chapter
Introduction to AI	Feb 26, 2020, 16:15 EST	Complete Chapter
Standardizing Data	Mar 4, 2020, 16:15 EST	Complete Chapter
Introduction to Data Preprocessing	Mar 4, 2020, 16:15 EST	Complete Chapter
Putting it all together	Mar 11, 2020, 16:15 EDT	Complete Chapter
Selecting features for modeling	Mar 11, 2020, 16:15 EDT	Complete Chapter
Feature Engineering	Mar 11, 2020, 16:15 EDT	Complete Chapter
Model Validation in Python	Mar 18, 2020, 16:15 EDT	Complete Course



Regression	Mar 25, 2020, 16:15 EDT	Complete Chapter
Classification	Mar 25, 2020, 16:15 EDT	Complete Chapter
Supervised Learning	Mar 25, 2020, 16:15 EDT	Complete Chapter
Linear Classifiers in Python	Apr 1, 2020, 16:15 EDT	Complete Course
Visualization with hierarchical clustering and t-SNE	Apr 22, 2020, 16:15 EDT	Complete Chapter
Clustering for dataset exploration	Apr 22, 2020, 16:15 EDT	Complete Chapter
Discovering interpretable features	Apr 29, 2020, 16:15 EDT	Complete Chapter
Decorrelating your data and dimension reduction	Apr 29, 2020, 16:15 EDT	Complete Chapter
Deep Learning & Beyond	May 13, 2020, 16:15 EDT	Complete Chapter

DataCamp



Introduction to Python

1 Python Basics FREE

100%

An introduction to the basic concepts of Python. Learn how to use Python interactively and by using a script. Create your first variables and acquaint yourself with Python's basic data types.

2 Python Lists

100%

Learn to store, access, and manipulate data in lists: the first step toward efficiently working with huge amounts of data.

3 Functions and Packages

100%

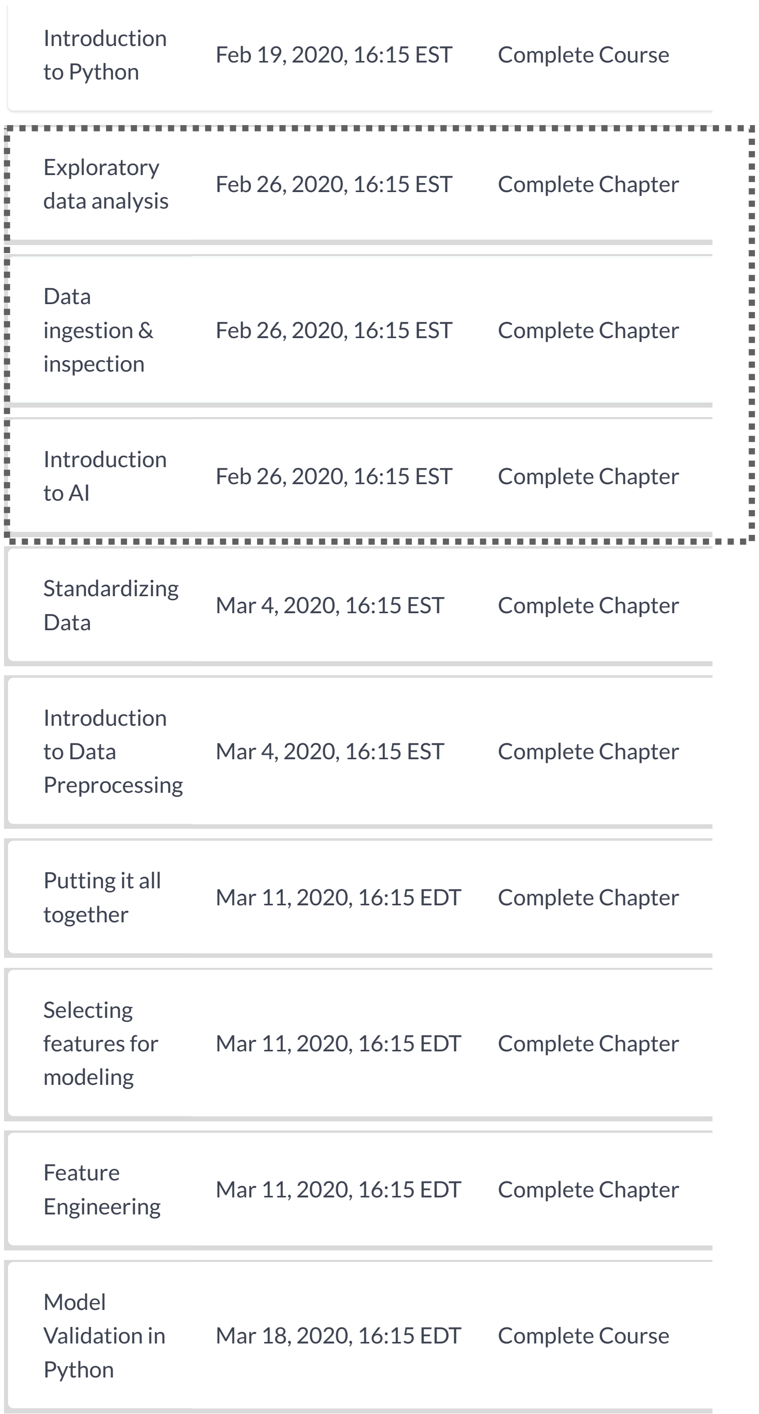
You'll learn how to use functions, methods, and packages to efficiently leverage the code that brilliant Python developers have written. The goal is to reduce the amount of code you need to solve challenging problems!

4 NumPy

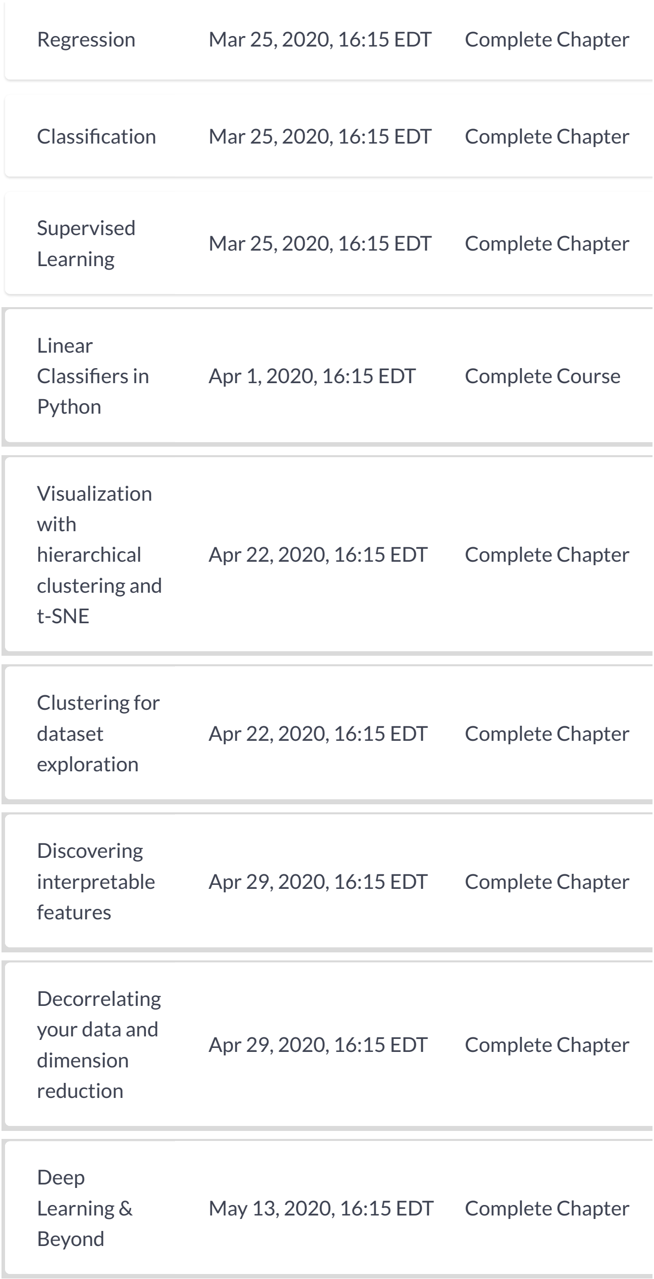
100%

NumPy is a fundamental Python package to efficiently practice data science. Learn to work with powerful tools in the NumPy array, and get started with data exploration.

DataCamp



Introduction to Python	Feb 19, 2020, 16:15 EST	Complete Course
Exploratory data analysis	Feb 26, 2020, 16:15 EST	Complete Chapter
Data ingestion & inspection	Feb 26, 2020, 16:15 EST	Complete Chapter
Introduction to AI	Feb 26, 2020, 16:15 EST	Complete Chapter
Standardizing Data	Mar 4, 2020, 16:15 EST	Complete Chapter
Introduction to Data Preprocessing	Mar 4, 2020, 16:15 EST	Complete Chapter
Putting it all together	Mar 11, 2020, 16:15 EDT	Complete Chapter
Selecting features for modeling	Mar 11, 2020, 16:15 EDT	Complete Chapter
Feature Engineering	Mar 11, 2020, 16:15 EDT	Complete Chapter
Model Validation in Python	Mar 18, 2020, 16:15 EDT	Complete Course



Regression	Mar 25, 2020, 16:15 EDT	Complete Chapter
Classification	Mar 25, 2020, 16:15 EDT	Complete Chapter
Supervised Learning	Mar 25, 2020, 16:15 EDT	Complete Chapter
Linear Classifiers in Python	Apr 1, 2020, 16:15 EDT	Complete Course
Visualization with hierarchical clustering and t-SNE	Apr 22, 2020, 16:15 EDT	Complete Chapter
Clustering for dataset exploration	Apr 22, 2020, 16:15 EDT	Complete Chapter
Discovering interpretable features	Apr 29, 2020, 16:15 EDT	Complete Chapter
Decorrelating your data and dimension reduction	Apr 29, 2020, 16:15 EDT	Complete Chapter
Deep Learning & Beyond	May 13, 2020, 16:15 EDT	Complete Chapter

Coding

- **Jupytr notebooks**
  - Fork the following repositories
    - https://github.com/jcdevaney/data71200sp20
    - https://github.com/amueller/introduction_to_ml_with_python
    - https://github.com/ageron/handson-ml
- **Python 3 tools**
  - import numpy as np
  - import scipy as sp
  - import matplotlib.pyplot as pat
  - import pandas as pd

Jupytr Notebook
01-introduction.ipynb
[2-9]

Terminology Review

- **labeled training set**
  - “a training set that contains the desired solution (a.k.a. a label) for each instance.”
- **regularization**
  - “constraining a model to make it simpler and reduce the risk of overfitting”
- **hyperparameter**
  - “amount of regularization to apply during learning”

Terminology Review

- **Training set**
  - data used to train the model
- **Testing set**
  - hold out data used to estimate the generalization error on new data
- **Validation set**
  - used to compare models
- **Cross-validation**
  - iteratively holding out a subset of the data and testing on the rest (typically 80/20)

More Terminology

- **Class**
  - “One of a set of enumerated target values for a label.”
- **Classification**
  - “A type of machine learning model for distinguishing among two or more discrete classes.”

https://developers.google.com/machine-learning/glossary

More Terminology

- **Samples**
  - Individual items
  - **Label**
    - “In supervised learning, the "answer" or "result" portion of an example”
  - **Feature**
    - “An input variable used in making predictions.”

https://developers.google.com/machine-learning/glossary

Machine Learning Pipeline

- **“However simple or complex the Machine Learning problem at hand may be, it will always contain the following steps:**
  - Data loading, preparation and splitting into the train and test partitions
  - Model selection and training ("fitting")
  - Model performance assessment”

“AI Fundamentals” course, DataCamp

Machine Learning Pipeline

- **“However simple or complex the Machine Learning problem at hand may be, it will always contain the following steps:**
  - **Data loading, preparation and splitting into the train and test partitions**
  - Model selection and training ("fitting")
  - Model performance assessment”

Jupytr Notebook
01-introduction.ipynb
[10-23]

“AI Fundamentals” course, DataCamp

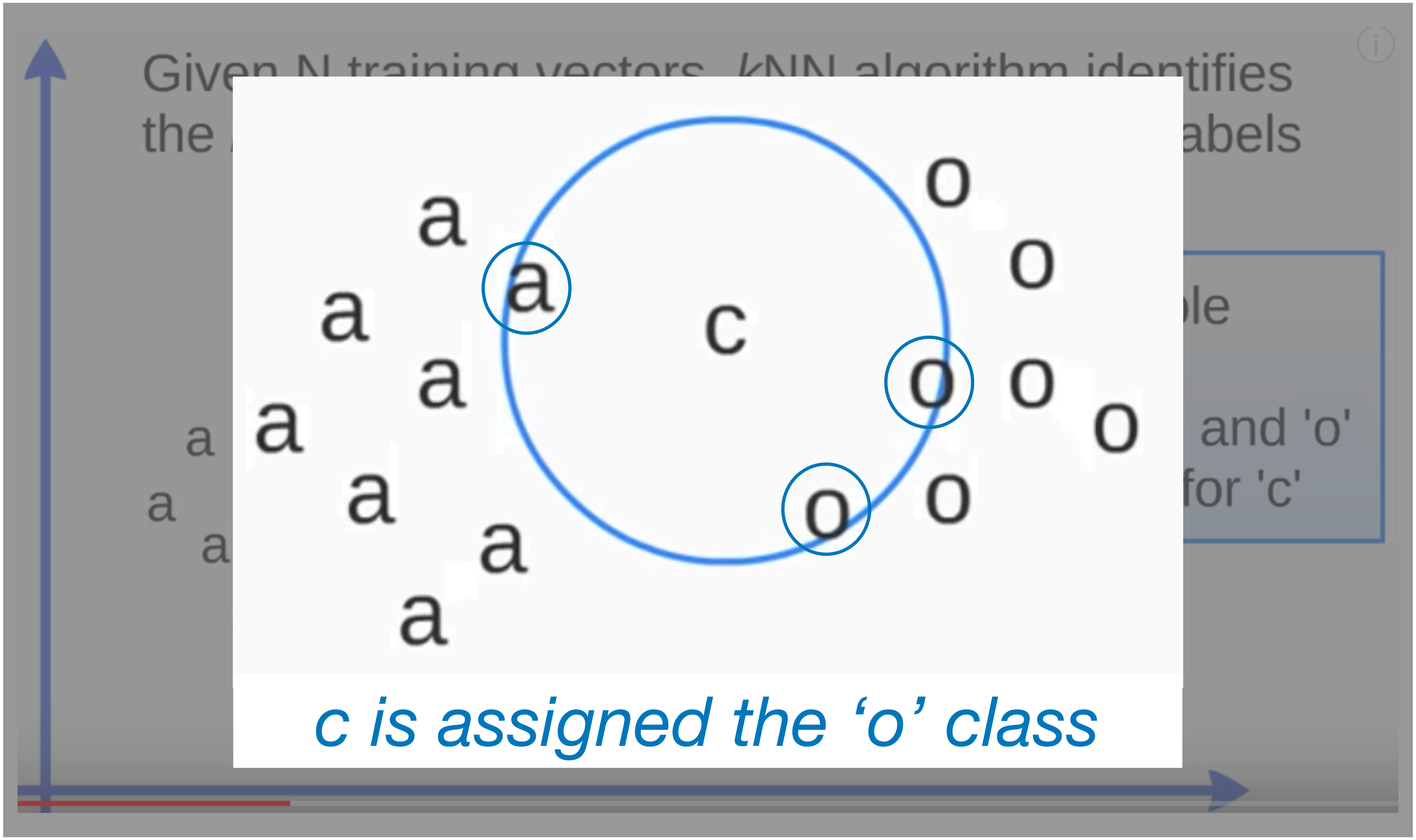
Machine Learning Pipeline

- **“However simple or complex the Machine Learning problem at hand may be, it will always contain the following steps:**
  - Data loading, preparation and splitting into the train and test partitions
  - **Model selection and training ("fitting")**
  - Model performance assessment”

Jupytr Notebook
01-introduction.ipynb
[24-27]

“AI Fundamentals” course, DataCamp

k Nearest Neighbor (kNN)



c is assigned the ‘o’ class

https://www.youtube.com/watch?v=UqYde-LULfs

Machine Learning Pipeline

- **“However simple or complex the Machine Learning problem at hand may be, it will always contain the following steps:**
  - Data loading, preparation and splitting into the train and test partitions
  - Model selection and training ("fitting")
  - **Model performance assessment”**

Jupytr Notebook
01-introduction.ipynb
[28-30]

“AI Fundamentals” course, DataCamp

Paired Question

- **What do you most want to learn to do with machine learning?**
    - What kind of data are you interested in working with?
    - What kind of questions do you want to be able to ask of your data?

Project 1

- **Due March 18**
- **Start exploring potential datasets**
  - kaggle.com
  - archive.ics.uci.edu/ml/datasets.php
  - libguides.nypl.org/eresources
  - opendata.cityofnewyork.us/data/
- **The data set will need to be labeled as you are going to use it for both supervised and unsupervised learning tasks**

Assignments for next week

- **DataCamp**
    - AI Fundamentals
        - Introduction to AI
    - pandas Foundations
        - Data ingestion & inspection
        - Exploratory data analysis
- **Reading**
    - Ch 2: End-to-End Machine Learning Project. in Géron, Aurélien. (2019). Hands-On Machine Learning with Scikit-Learn, Keras and TensorFlow' O'Reilly Media, Inc.