

Advanced Data Analysis

DATA 71200

Class 2

Question Set 2

Géron (p. 22–30)

- ▶ **Can you name four of the main challenges in Machine Learning?**
- ▶ **If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?**
- ▶ **What is a test set and why would you want to use it?**
- ▶ **What is the purpose of a validation set?**
- ▶ **What can go wrong if you tune hyperparameters using the test set?**
- ▶ **What is cross-validation and why would you prefer it to a validation set?**

Question Set 2

- ▶ **Can you name four of the main challenges in Machine Learning?**
 - “Some of the main challenges in Machine Learning are the lack of data, poor data quality, non-representative data, uninformative features, excessively simple models that underfit the training data, and excessively complex models that overfit the data.”

Question Set 2

- ▶ **If your model performs great on the training data but generalizes poorly to new instances, what is happening? Can you name three possible solutions?**
 - “If a model performs great on the training data but generalizes poorly to new instances, the model is likely overfitting the training data (or we got extremely lucky on the training data). Possible solutions to overfitting are getting more data, simplifying the model (selecting a simpler algorithm, reducing the number of parameters or features used, or regularizing the model), or reducing the noise in the training data.”

Question Set 2

- ▶ **What is a test set and why would you want to use it?**
 - “A test set is used to estimate the generalization error that a model will make on new instances, before the model is launched in production.”
- ▶ **What is the purpose of a validation set?**
 - “A validation set is used to compare models. It makes it possible to select the best model and tune the hyperparameters.”

Question Set 2

- ▶ **What can go wrong if you tune hyperparameters using the test set?**
 - “If you tune hyperparameters using the test set, you risk overfitting the test set, and the generalization error you measure will be optimistic (you may launch a model that performs worse than you expect).”
- ▶ **What is cross-validation and why would you prefer it to a validation set?**
 - “Cross-validation is a technique that makes it possible to compare models (for model selection and hyperparameter tuning) without the need for a separate validation set. This saves precious training data.”

Typical Machine Learning Project Steps

- ▶ “You studied the data.”
- ▶ “You selected a model.”
- ▶ Feature Engineering
- ▶ “You trained it on the training data (i.e., the learning algorithm searched for the model parameter values that minimize a cost function).”
- ▶ “Finally, you applied the model to make predictions on new cases (this is called inference), hoping that this model will generalize well.”

Main Challenges

- ▶ **Insufficient training data**
 - Quantity and/or quality and/or non-representative
- ▶ **Irrelevant features**
- ▶ **Overfitting training data**
- ▶ **Under-fitting training data**

Example: GDP and Happiness

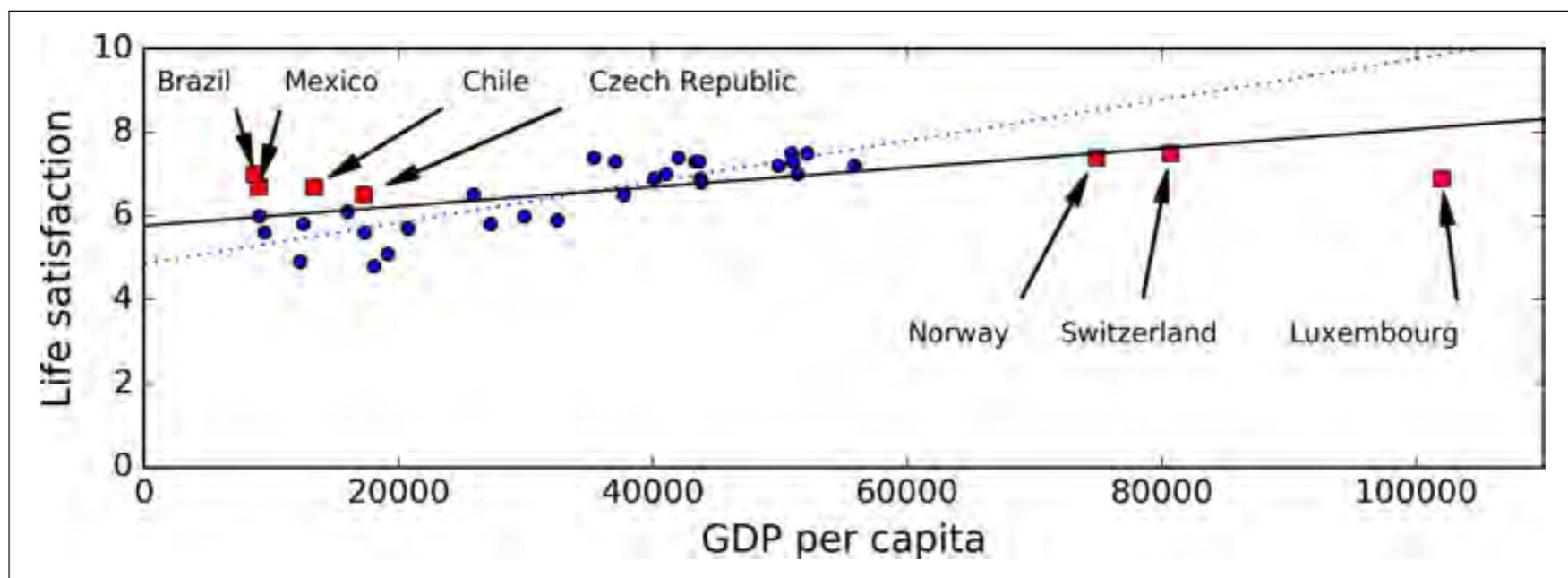


Figure 1-21. A more representative training sample

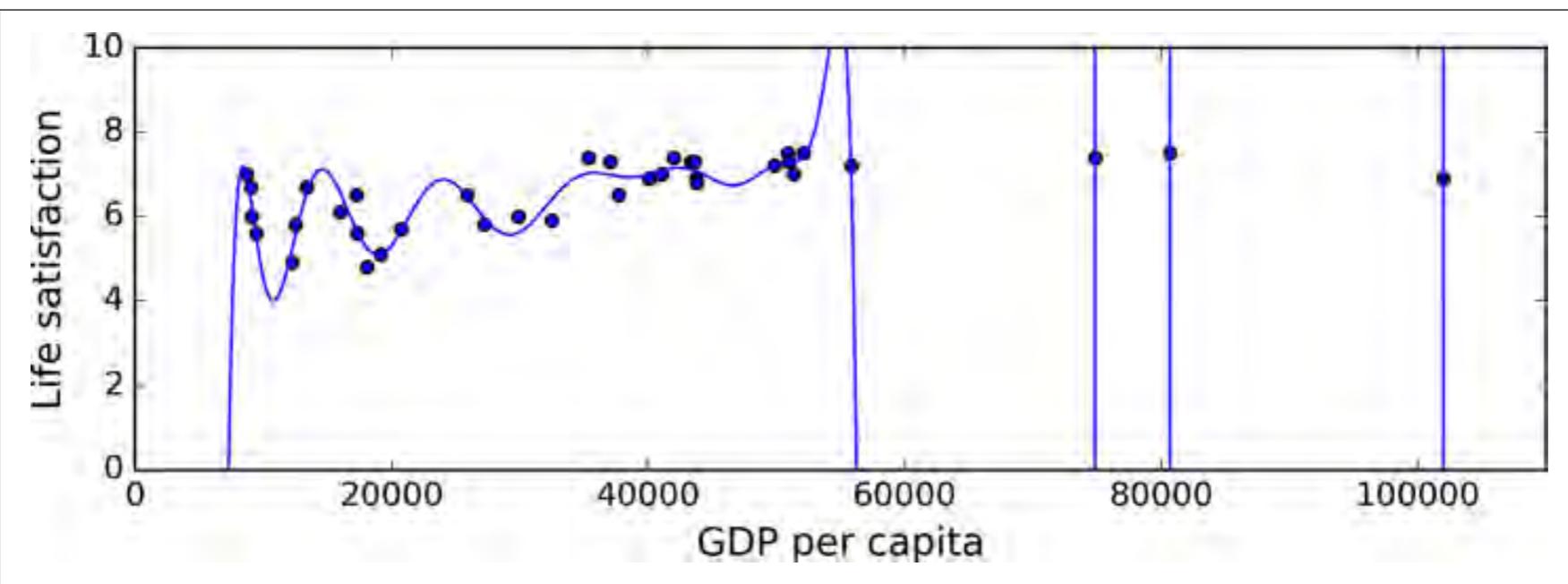


Figure 1-22. Overfitting the training data

Example: GDP and Happiness

► regularization

- “constraining a model to make it simpler and reduce the risk of overfitting”

► hyperparameter

- “amount of regularization to apply during learning”
- “needs to be set before training”

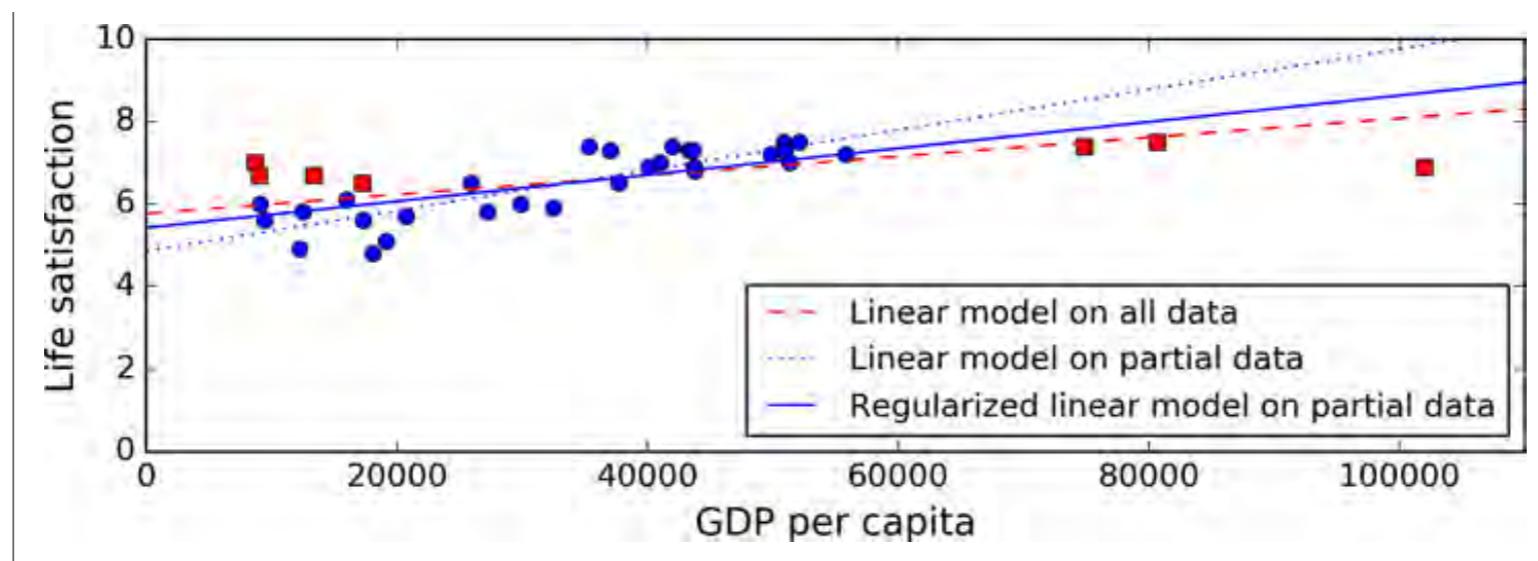


Figure 1-23. Regularization reduces the risk of overfitting

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 training data and testing on the rest (typically 80/20: 5-fold cross-validation)

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.”

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.”

GitHub - Cloning a Repository

 jcdevaney / onssen
forked from speechLabBcCuny/onssen

 Watch ▾ 0  Star 0  Fork 23

 Code  Pull requests 0  Actions  Projects 0  Wiki  Security  Insights  Settings

An open-source speech separation and enhancement library 

[Manage topics](#)

 28 commits  2 branches  0 packages  0 releases  1 contributor  GPL-3.0

Branch: master ▾ [New pull request](#) [Create new file](#) [Upload files](#) [Find file](#) [Clone or download ▾](#)

This branch is even with speechLabBcCuny:master.

 nateanl Create LICENSE Latest commit 0479d78 on Nov 29, 2019

 configs Add batch_norm after rnn, refactorize training, add readme 3 months ago

 data Add batch_norm after rnn, refactorize training, add readme 3 months ago

Clone with HTTPS  [Use SSH](#)

Use Git or checkout with SVN using the web URL.

`https://github.com/jcdevaney/onssen.c` 

[Open in Desktop](#) [Download ZIP](#)

GitHub Desktop

Current Repository **onssen** Current Branch **master** Fetch origin Last fetched just now

An updated version of GitHub Desktop is available and will be installed at the next launch. See [what's new](#) or [restart GitHub Desktop](#).

Changes History [New](#)

0 changed files

No local changes

There are no uncommitted changes in this repository. Here are some friendly suggestions for what to do next.

Open the repository in your external editor
Select your editor in [Preferences](#)
Repository menu or ⌘ ⌘ A

View the files of your repository in Finder
Repository menu or ⌘ ⌘ F

Open the repository page on GitHub in your browser
Repository menu or ⌘ ⌘ G

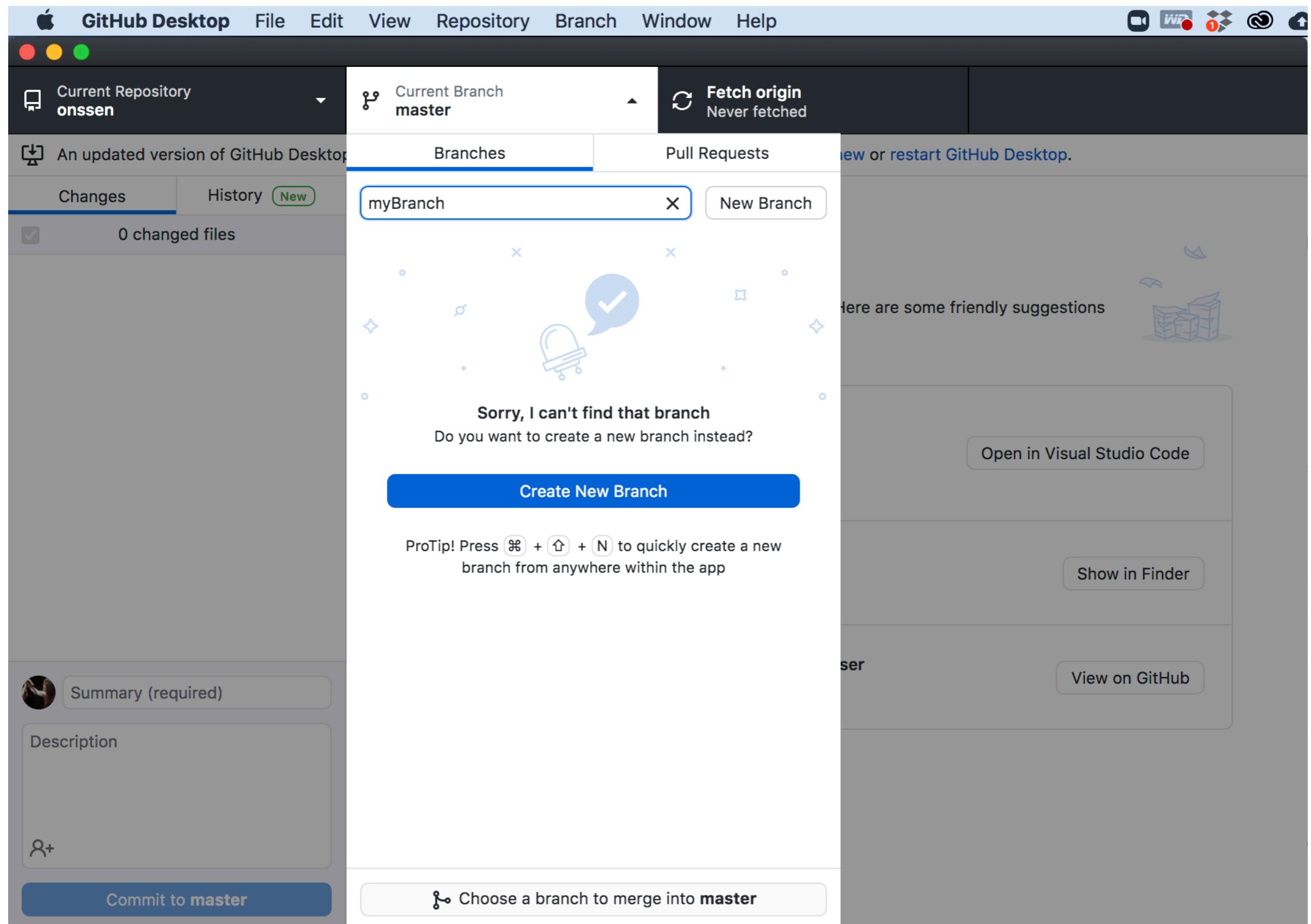
Summary (required)

Description

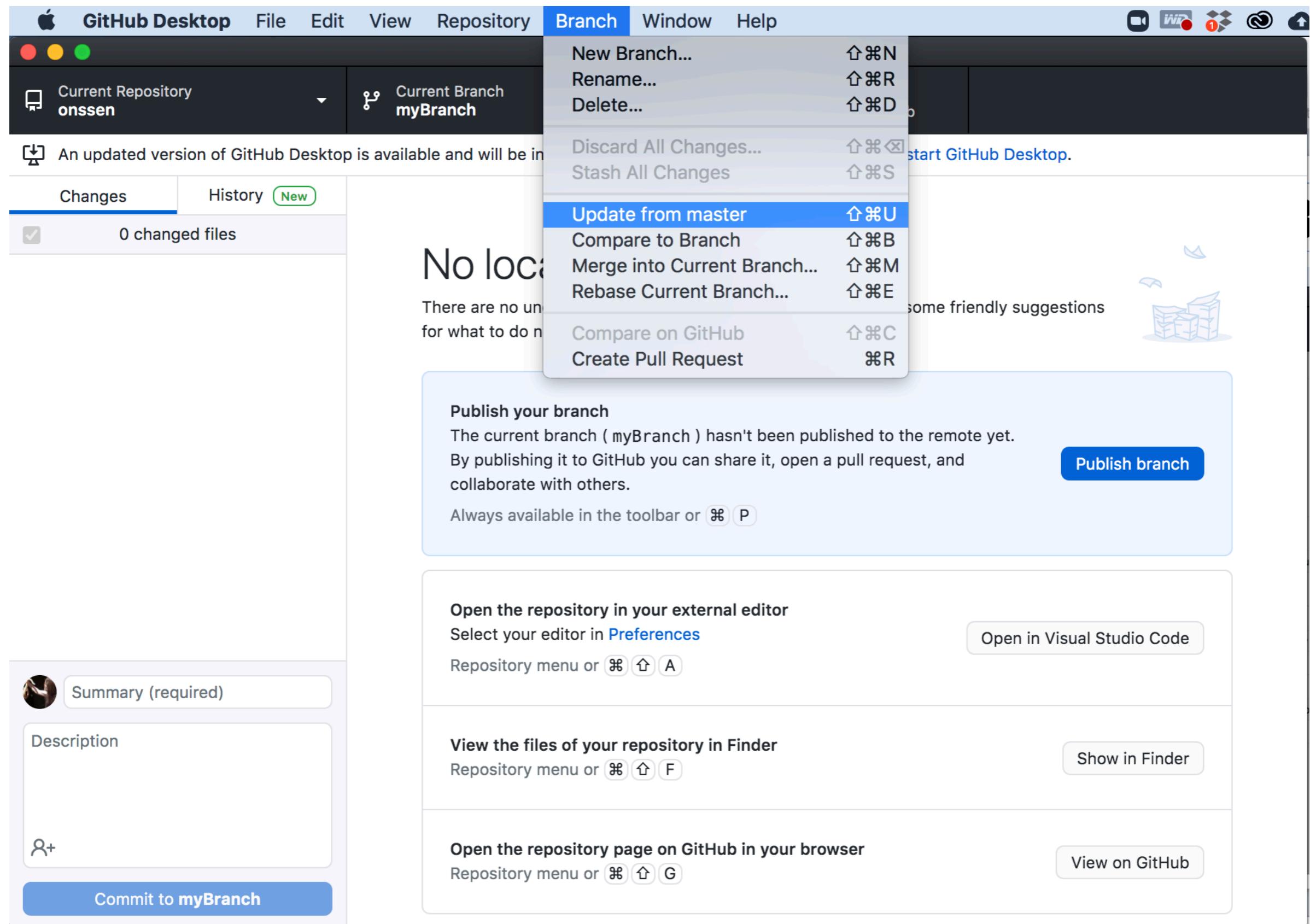
Commit to master



GitHub Desktop - Create a Branch



GitHub Desktop - Update a Branch



Coding

- ▶ **Jupyter notebooks**

- Clone the following repositories
 - <https://github.com/jcdevaney/data71200su21>
 - 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 plt
- import pandas as pd

**Jupyter Notebook
01-introduction.ipynb
[2-8]**

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”

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”

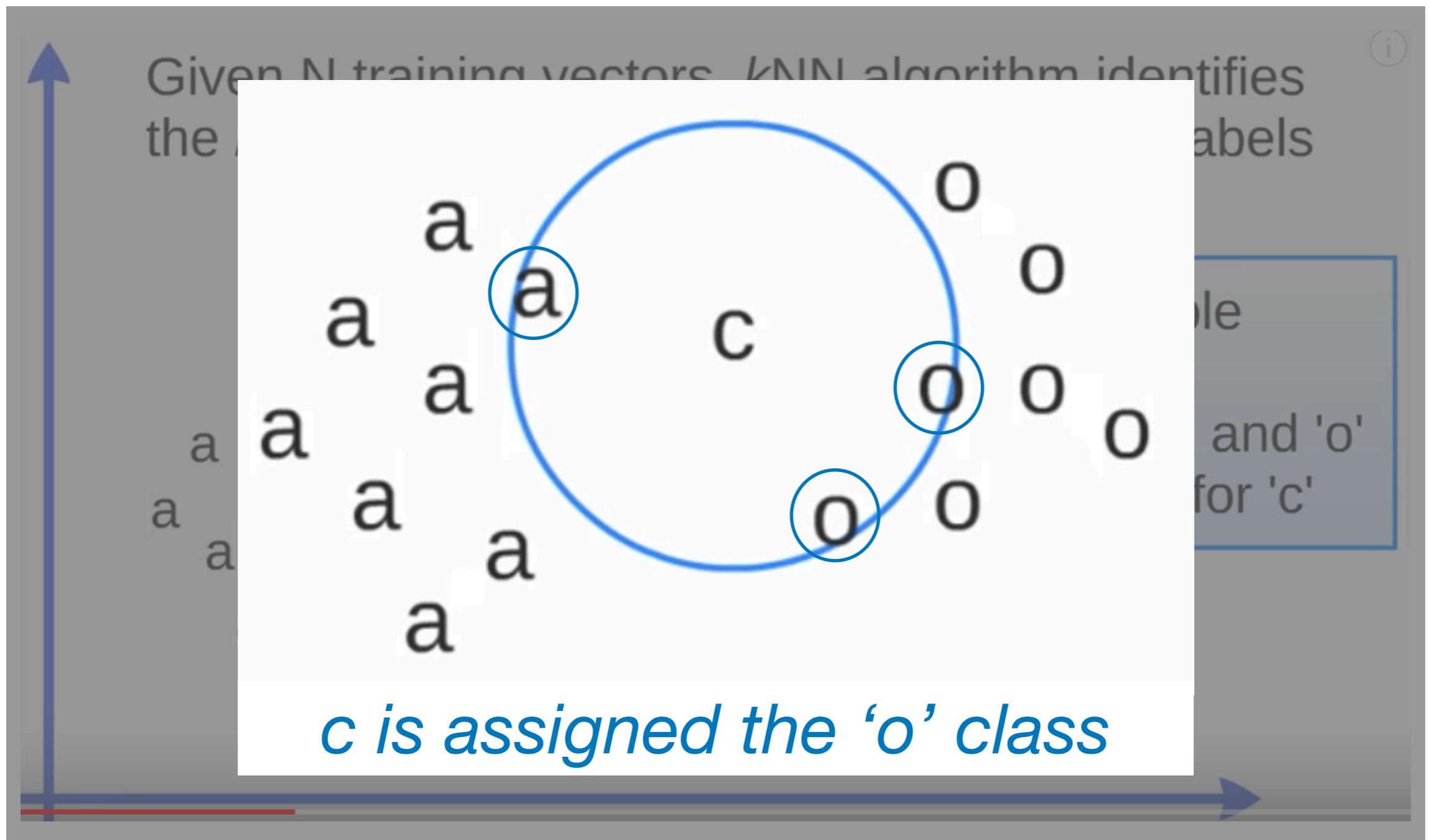
**Jupyter Notebook
01-introduction.ipynb
[10-24]**

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”

**Jupyter Notebook
01-introduction.ipynb
[25-28]**

k Nearest Neighbor (kNN)



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**

Jupyter Notebook
01-introduction.ipynb
[29-32]

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 June 14**
- ▶ **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**

- *Introduction to Python (If Needed)*
- AI Fundamentals
 - Introduction to AI
- Data Manipulation with pandas
 - Transforming Data
 - Aggregating Data
 - Slicing and Indexing
 - *Creating and Visualizing DataFrames (Optional)*
- *Writing Efficient Code with pandas (Optional)*

► **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. 33–66