INTRO TO THE COURSE
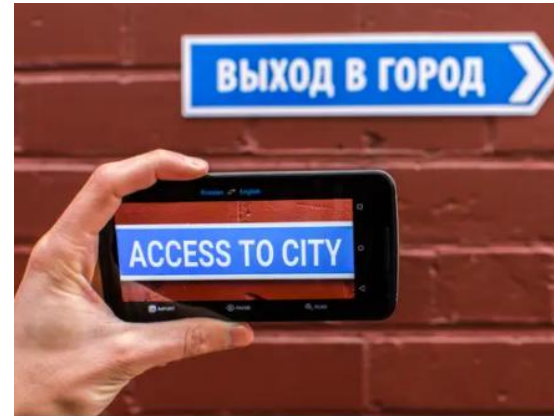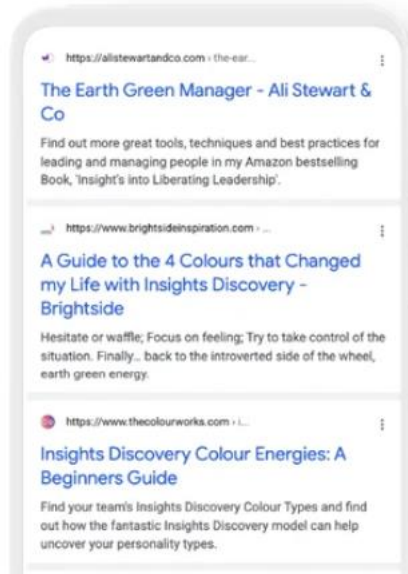
We used ML in everyday life much before ChatGPTs

How AI powers great search results

insights how to manage a green

https://allstewartandco.com › the-ear...

The Earth Green Manager - Ali Stewart & Co

Find out more great tools, techniques and best practices for leading and managing people in my Amazon bestselling Book, 'Insight's into Liberating Leadership'.

https://www.brightsideinspiration.com › ...

A Guide to the 4 Colours that Changed my Life with Insights Discovery – Brightside

Hesitate or waffle; Focus on feeling; Try to take control of the situation. Finally... back to the introverted side of the wheel, earth green energy.

https://www.thecolourworks.com › i...

Insights Discovery Colour Energies: A Beginners Guide

Find your team's Insights Discovery Colour Types and find out how the fantastic Insights Discovery model can help uncover your personality types.

ВЫХОД В ГОРОД

ACCESS TO CITY

Google

According to Netflix, the information gathered can be any of the following:

NETFLIX

How you rate titles

Your viewing history

How long you watch a video

Other users who may have similar tastes

Information about titles such as genre, actors and release year

facebook

Explainer

Why is Facebook shutting down its facial recognition system and deleting 'faceprints'?
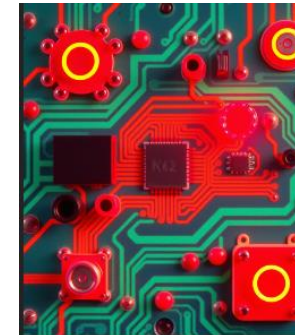
And now we use AI even more

Generate a futuristic image of ML applications

Here is a futuristic image showcasing various machine learning applications in an advanced, high-tech world, emphasizing innovation and cutting edge technology.

AI keeps advancing every day

AI keeps advancing every day

How many 'r's in strawberry?

AI keeps advancing every day

How many 'r's in strawberry?

ChatGPT 4o mini ⌄

⬆ Share

How many r in strawberry

The word "strawberry" contains **two** "r"s.

🔊 ⧉ 👍 👎 ↻⌄

DeepSeek

▷

↻⁺

New chat

8. **R** - 2nd 'r'

9. **Y**

Yes, there are two 'r's in "strawberry."

Step 8: Conclusion

After carefully reviewing and correcting my initial mistake, I conclude that the letter 'r' appears **two times** in the word "strawberry."

Final Answer: There are **2** letter 'r's in "strawberry."

⧉ 👍 👎

AI keeps advancing every day

How many 'r's in strawberry?

ChatGPT 4o mini ⌄

Share

How many r in strawberry

The word "strawberry" contains **two** "r"s.

DeepSeek

New chat

8. **R** - 2nd 'r'

9. **Y**

Yes, there are two 'r's in "strawberry."

Step 8: Conclusion

After carefully reviewing and correcting my initial mistake, I conclude that the letter 'r' appears **two times** in the word "strawberry."

Final Answer: There are **2** letter 'r's in "strawberry."

Share

How many r in strawberry

ChatGPT 4o ⌄

The word **"strawberry"** contains **three** "r"s. 🍓

Course instructor

Sabina Zejnilović

Adjunct Professor, Nova SBE

Ph.D., Carnegie Mellon University

Data Scientist, Cloudflare

sabina.zejnilovic@novasbe.pt

Renato Gonçalves
Teaching Assistant at Nova SBE

Master's in Management at Nova SBE, 2021/2022

Data Analyst at Vodafone
renato.goncalves@novasbe.pt

Class structure

VIDEO LECTURES

When: to be watched before the class

Where: moodle

What: Theoretical concepts

Assessment: Quiz on the video content (lecture 2 to lecture 10), to be done before the in-person class

IN PERSON LECTURES

When: Friday 12:30 – 14:00

Where: Room B03

What:

- Continuation and review of theoretical concepts

- Walk through coding exercises (jupyter notebooks)

- Q&A

- **No recording** (open discussion)

Course moodle

- All course materials will be available on moodle (https://moodle.novasbe.pt/)
  - Video lectures
  - Links for online office hours
  - Slides
  - Assignments
  - Forums for questions

- Please enroll in the course 2695 with the password: **ml2025.**

- Note: Teams course channel **will not be used**, only for accessing the office hours. All questions should be sent either by email or posted on the moodle course forum. Moodle forums have automatic subscription (you should receive an email if somebody posts a question).

Regular online office hours:

- Sabina: Mondays 12.30-13.30 (Teams meeting link available on moodle)

- Renato: Thursdays 12.30-13.30 (Teams meeting link available on moodle)

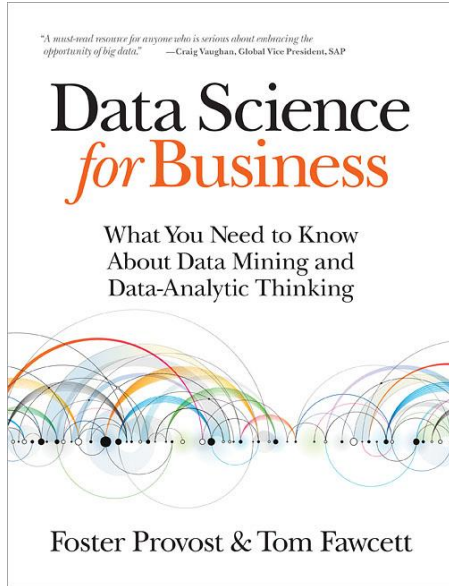- additional hours may be available on request (the teams link **will be the same** as for the regular hours)

Emails:

- Communication should be directed both to Sabina and Renato

- Email subject should include the course code **2695**

- Questions should be sent at least 24 hours before the deadline to ensure a timely response.
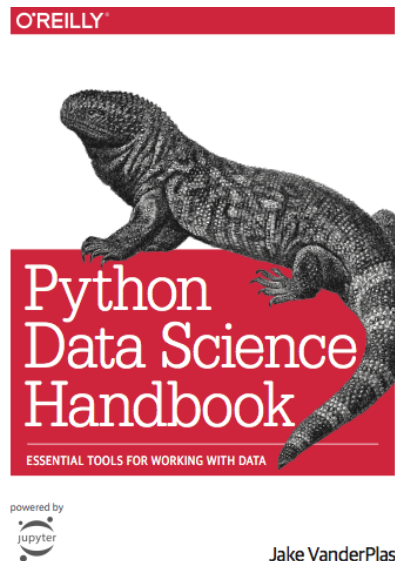
If you are struggling with some concept or code, or would like to discuss something, or you have some general question or feedback:

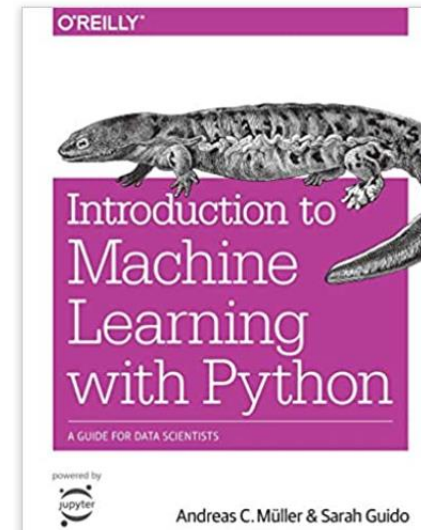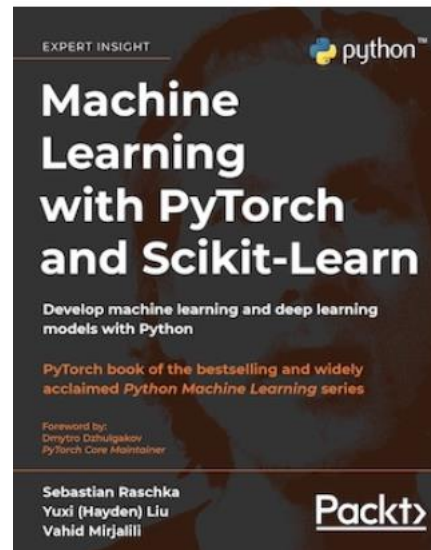<div align="center">

Let us know!

</div>

Some reference books

https://data-science-for-biz.com/

https://github.com/amueller/introduction_to_ml_with_python

https://jakevdp.github.io/PythonDataScienceHandbook/

https://github.com/rasbt/machine-learning-book

Check the links on slides and Jupyter notebooks for more references.

Course schedule

Lecture	Date	Content	Quiz done before the class	HW	Project
1	Friday Feb 7	ML concepts	-		
2	Friday Feb 14	Regression	1	HW1 out	Project discussion
3	Friday Feb 21	Classification	2		Project preferences
4	Friday Feb 28	Data preprocessing ML pipeline	3		Projects assigned
5	Friday Mar 7	Decision Trees Ensemble learning	4	HW1 due HW2 out	
6	Friday Mar 14	Unsupervised Learning	5		

2-week break from the lectures

Course schedule

Lecture	Date	Content	Quiz	HW	Projects
7	Friday Apr 4	Ethics & model interpretability	6	HW2 due HW3 out	
8	Friday Apr 11	Neural networks	7		
		2-week break from the lectures			
9	Friday May 2	Text mining	8	HW3 due	
10	Friday May 9	Intro to Large Language Models	9		
11	Monday May 12 (makeup from 25.4.)	Applications of LLM	-		Final projects video due
12	Friday May 16	Final project discussion MANDATORY ATTENDANCE	-		Final project code due

Assessment

Type	Quantity	Description	Weight
Quizzes	9	Weekly on video content, due before in person class (best 8/9)	20%
Homework	3	Individual assignment	25%
Project	Code submission + video presentation, Q&A	Group up to 4 students	25%
Exam	1	Theory and coding Tuesday 27.5.2025. 11:30 (2h open book)	30%

- Example of solving a business problem with ML.

- A list of possible project ideas will be available on Moodle.

- You can also choose your own topic and dataset, BUT the project proposal should be approved before proposal submission. This is done by sending an email to the TA & instructor with a brief project description.

- You should submit your top 3 project preferences on Moodle by Friday, February 21st .

- An assignment of students to projects will be available in week 4.

- All students without a group will be assigned to groups based on their student number and project preferences. We will do our best to accommodate everyone's interest, and we will solve ties by giving preference to earlier submissions.

- **Myth 1**: This is a beginner's course for Machine Learning, this is a business school, and with so AI assistants available, there is no need for theory and math.

- **Reality**:   If we just simply copy the steps to apply the algorithm in practice without the basic understanding of how the algorithm works, we may end up:
    - doing things that don't make sense, due to that lack of understanding.
    - only being able to copy code from others, not being able even to adapt the code to our needs.
    - not knowing how to apply algorithms to new kinds of data, without someone showing us how first.

    - Quizzes will require knowledge of theory.

    - Homework will require knowledge of theory.

    - Project will require knowledge of theory.

    - Exam will require knowledge of theory.

https://lazyprogrammer.me/why-do-you-need-math-for-machine-learning-and-deep-learning/

- **Myth 2**: This is a business school, so this class will be all theoretical, with no need for programming.

- **Reality:** This is a hands-on course where students will be expected to use Python and Jupyterlab to apply Machine Learning concepts to implement solutions to various business problems.

  - Homework will require coding.

  - Project will require coding.

  - Half of the exam will require coding.

- **Myth 3**: This is an elective course, it's an introduction course, so it should be less work than a mandatory or non introduction course.

- **Reality**:
  - We will study Machine Learning! This is not an easy topic!
  - We will learn the basic theoretical concepts and apply them.
  - There will be video lectures, in person classes, quizzes, homework, project, exam.
  - **This course is a lot of work!**
  - But it will be worth it, just wait to see your final projects ☺

- **Myth 4**: This course will give teach me all that is needed to become a Deep Learning expert with complete knowledge of the state of the art in Generative AI.

- **Reality**:
  - Machine learning is a vast field.
  - Deep learning is also a vast field.
  - We will have in total 12 lectures to cover the most important concepts for Machine Learning.
  - We can't jump to neural networks without understanding the basics of how to prepare the data, what is a machine learning model, how to evaluate it, and, as we will see, neural networks are not always the best solution for business problems.

- **Myth 5**: This course will teach me how to use the latest AI tools.

- **Reality**:
  - Knowing the latest AI tools can speed up many process, but this is not the goal of this course.
  - We want you to learn the basic concepts of Machine Learning, and these will help you later to understand the capabilities and limitations of AI tools.
  - With a solid grasp of ML, you can critically analyze the results produced by AI tools.
  - The field of AI tools is so rapidly growing, it's impossible to know all the latest tools. Feel free to share with the class your favorites.

Can you use ChatGPT and similar tools for assignments

YES!

BUT You should not simply copy paste the code from the code assistants!

DO:
- Use code assistant to help you understand the code
- Use code assistants to help you debug the code
- Use code assistants to help you get unstuck

DON'T
- Don't just copy paste the code without understanding it, running it and adjusting if necessary

NO POINTS AT ALL will be given for an assignment if we see that:
- the code was just pasted, and not even executed
- the code does not answer the exact question asked, but something similar that the code assistant assumed
- the code is overly complicated, even though a much simpler approach was covered in the course material

- Do I have to attend all classes?
  - We will note take attendance, but we strongly suggest attending. Attending the last lecture is mandatory.

- What is the late assignment policy?
  - Assignments need to be submitted by their due date and will lose 20% of points for each late day.
  - Example: if the assignment is late for 3 days, student may get 40% of points at maximum. If the assignment is late for 6 days, it will be returned without evaluation.

ML CONCEPTS

What is Machine Learning?

Machine Learning (ML) is the science and art of programming computers so they can learn from data.

Data point/ Instance / example

- Described by a set of **features (and possibly a label)**

Features/attributes/predictors/explanatory/independent variables/input

- Variables describing our data

Label/ target/outcome/response/dependent variables

- Unknown value of interest, variable we are predicting

Basic Feature types

- Numeric: anything that has some order
- Categorical (symbolic): features that do not have an order

Model/ Estimator/ Algorithm

- A *formula/function* created for a specific purpose: (for example: estimating the label)

    Machine automatically "learns" the formula on its own, instead of knowing it a priori.

    - **Training**

        Creating or learning the model from data

    - **Inference**

        Applying the trained model to make predictions **on *new (unseen)* data**

Prediction

- Estimate of an unknown value (target)

- In this context it does not imply forecasting/prediction of the future, it estimates the value of interest

How much and what type of supervision is needed for model training?

- **Supervised learning**

    Given: training data + desired outputs (labels)

- **Unsupervised learning**

    Given: training data (without desired outputs)

- **Semi-supervised learning**

    Given: training data + a few desired outputs

- **Reinforcement learning**

    Rewards from sequence of actions

Supervised learning

Model is trained with examples that include labels.

- **Task:**

Given a set of features and labels learn a model that will predict a label to a new feature set

- **Example:**

Given a dataset with customers' attributes and information whether they churned or not, learn to predict whether the new clients will churn or not

Given a dataset with house characteristics and respective market prices, learn to predict the market price of houses not present in the dataset
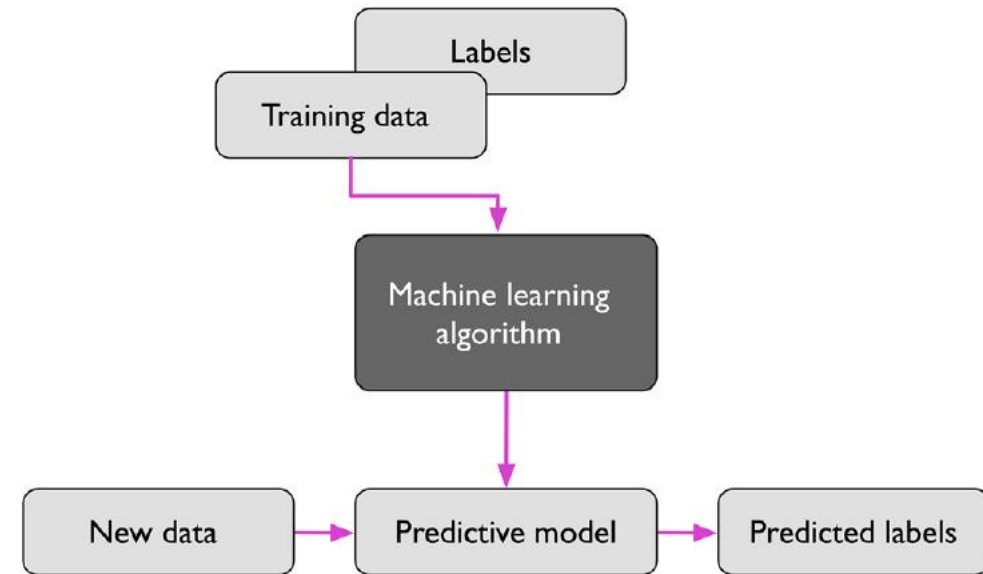
Types of supervised methods: Classification vs. Regression

Classification model: predict a **category/class** of instances among two or more discrete classes:

Will a client churn YES or NO

Regression model: a **continuous** value

How much will this house cost?

Input Data
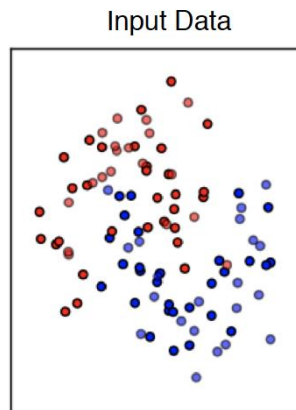
How to classify this dataset into 2 categories ?

red and blue
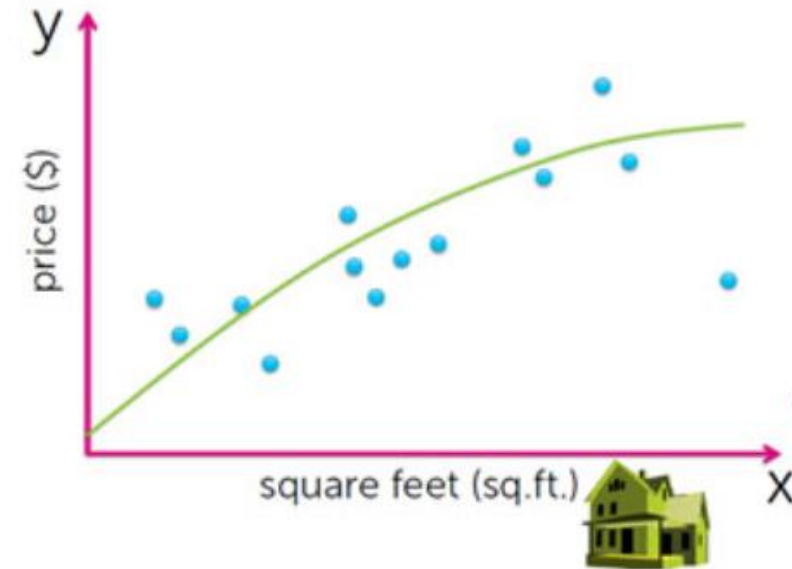
Unsupervised learning

Model is trained without labeled data

- **Task:**

Discover patterns in data

- **Example:**

Given a dataset with customers' attributes, group customers into segments

Given a list of transactions, find anomalous (fraudulent) ones

Market segmentation

Semisupervised learning

Model is trained with a lot of unlabeled data and a little bit of labeled data

- **Task:**

Given a set of features and a small set labels learn a model that will predict a label to a new feature set, with a model better than if developed using only a small set of labeled data

- **Example:**

Given many text documents, with only of few having their topic identified (label), find a topic of a new document

Any task where labels are expensive to obtain, but we have a lot of unlabeled data

Reinforcement learning

Model is learnt in a special context:

- agent observes the *environment*,

- selects and performs *actions*

- receives *rewards* based on performed actions
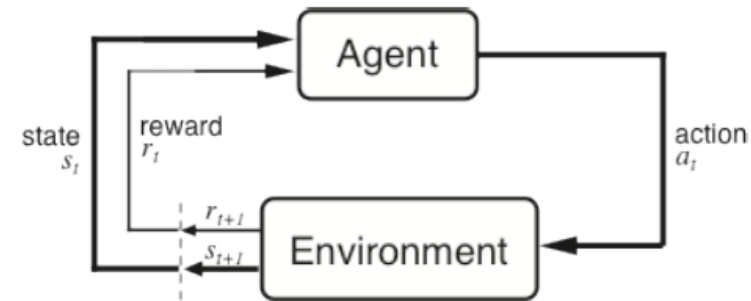
Task:

Given *{environment, actions, rewards}* learn a policy (strategy) and utility (rewards) functions

- **Example:**

Given stock price and relevant attributes decide whether to hold, buy or sell stock to maximize profit
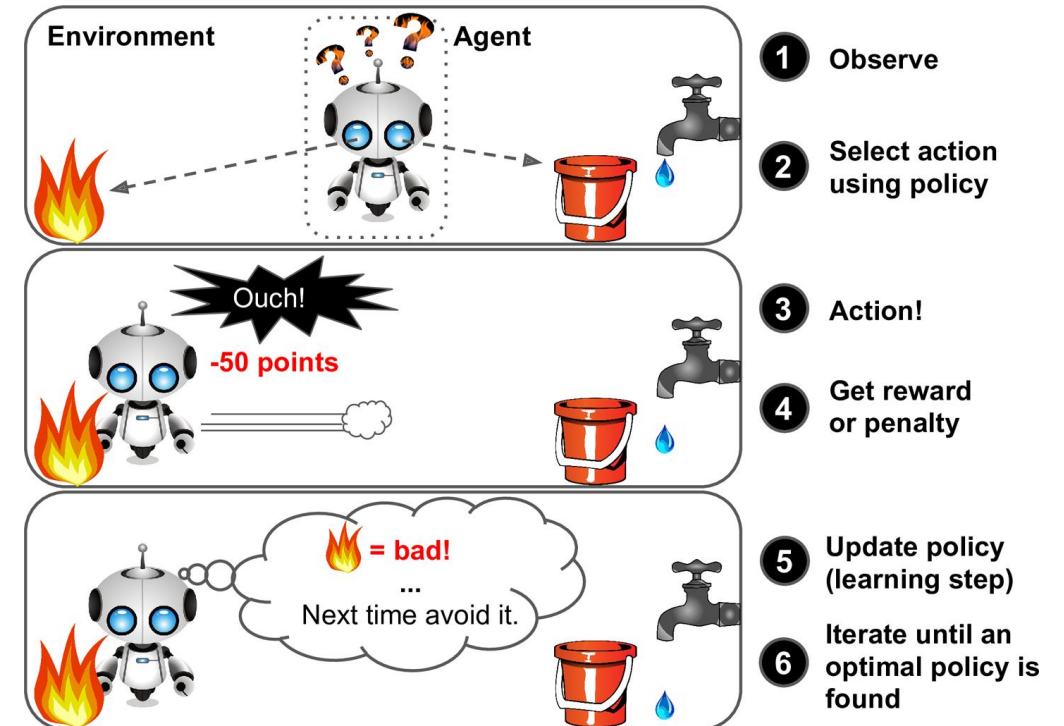
PYTHON SETUP

Python is becoming the most popular programming language for ML

Source: TechVidvan

Python libraries

StatsModels
Statistics in Python

NetworkX

xarray

TensorFlow

scikits-image
image processing in python

django

PyTorch

scikits
learn
machine learning in Python

DASK

And many,
many more...

Keras

pandas
$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$

matplotlib

plotnine

SciPy

bokeh

NumPy

jupyter

plotly

IP[y]:
IPython

python

spyder

Seaborn

Write your first python program interactively

- IPython is a powerful interactive shell for Python programming.

- In 2014, Jupyter project (http://jupyter.org/) was created as a spin-off project from IPython. It is language-agnostic.
  - JUlia + PYthon + R

- Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations.

- Jupyter notebook system allows you to author content in Markdown to create a rich documentation with code and text.

https://github.com/adam-p/markdown-here/wiki/Markdown-Cheatsheet

- We will be using **JupyterLab,** the next generation of the Jupyter Notebook.

So, how do we get started with python?

- A **software distribution** is a pre-built and pre-configured collection of packages that can be installed and used on a system.

- A **package manager** is a tool that automates the process of installing, updating, and removing packages.

- **Anaconda** is a Python and R distribution platform, contains a package manager **Conda**.

- Many data science packages come preinstalled with Anaconda.

- You can download a free Anaconda Distribution at:
  https://www.anaconda.com/products/distribution

What does ANACONDA interface look like?

Anaconda comes with a suite of graphical tools called Anaconda Navigator

- **JupyterLab** is an interactive programming environment (execute-explore vs. edit-compile-run): *experiment and evaluation*

- **Spyder** is an integrated development environment (IDE): *module development*
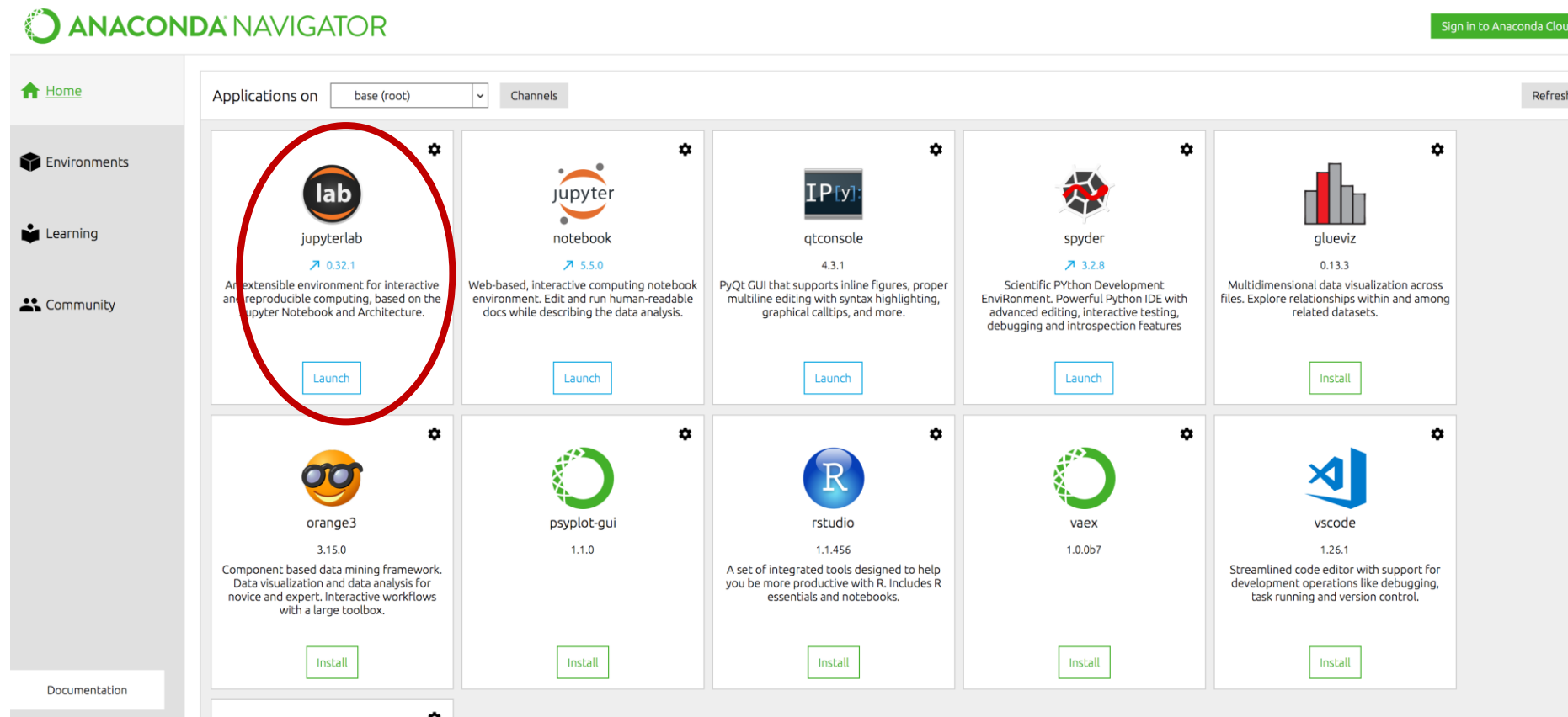
- **Pip** (Pip Installs Packages) is Python's officially-sanctioned package manager.
- Pip vs Conda:
  - Pip is a general-purpose manager for Python packages; conda is a language-agnostic cross-platform environment manager
  - Pip installs python packages within any environment; conda installs any package within conda environments.
  - For *our* use, pip and conda are mostly interchangeable.

Nice explanations of anaconda details:
https://jakevdp.github.io/blog/2016/08/25/conda-myths-and-misconceptions/

Virtual environments

- Python has different versions, packages have different versions. What if for two projects you need different versions of the same package?

- **Virtual environment** is an isolated environment that allows us to keep these dependencies in separate "sandboxes".

- We can have many different environments, as they take up little space, each with separate package versions.

Let's get started

STEP 1: **Install Anaconda**

- Install Anaconda environment on your laptop
    - Download open-source Individual Edition Anaconda distribution for Python according to your OS (https://www.anaconda.com/download/)
    - It is free for solo practitioners, students, and researchers
    - Follow the installation instruction (https://docs.anaconda.com/anaconda/install/)
    - If your computer is short in storage, you may also consider Miniconda (no packages pre-installed)
    - To verify everything is working, open Anaconda prompt and write:
        python --version
        conda -- version
    - Confirm that you have the latest version of conda:
        conda update conda
        (Update any package, if necessary, by typing y to proceed)

Step 2: Create conda virtual environment for this class

- To **create a virtual environment** named *ml2025,* type from Anaconda Prompt

  conda create -n ml2025 python=3.12

  when conda asks you to proceed, type y

- Activate your environment:

  conda activate ml2025

  the active environment---the one you are currently using---is shown in parentheses () or brackets [] at the beginning of your command prompt:

- List all packages in environment

  Environment active: *conda list*

  Environment not active   *conda list -n ml2025*

- Check if a specific package is installed:

  Environment active: *conda list <package name>*

  Environment not active   *conda list -n ml2025 <package name>*

```
-n, --name
```

Name of environment.

Step 2: Create conda virtual environment for this class

- Deactivate your environment

  conda deactivate

- Delete an environment (no need to do now)

  conda env remove -n ml2025

- List all environments

  conda env list

Step 3: Install the packages we will use during this course

- Install a package

  Environment active: *conda install <package_name>*

  Environment not active   *conda install –n ml2025 <package name>*

- Install mutiple packages

  conda install pandas numpy

- Upgrade a package

  conda update <package name>

- Install a package with a specific version

  conda install <package name> = <version number>

- Remove a package

  conda remove <package name>

Some of the packages we will use during this course

- **pandas**   data manipulation and analysis

- **numpy** mathematical functions

- **scikit-learn** machine learning (`conda install conda-forge::scikit-learn`)

  Referred to as *sklearn* when importing, example: from **sklearn**.metrics import mean_squared_error

- **xgboost, lightgbm, catboost** gradient boosting

    Xgboost Installed as: conda install -c conda-forge py-xgboost

    lightgbm installed as: conda install -c conda-forge lightgbm

    catboost installed as: as conda install -c conda-forge catboost

- **pytorch** and **torchvision**   neural networks   (conda install pytorch torchvision cpuonly -c pytorch)

- **shap** interpreting ML models

    Installed as: conda install -c conda-forge shap

- **matplotlib, seaborn** data visualization

- **spacy, nltk** text analysis

    conda install -c conda-forge spacy

    conda install nltk

- **imbalanced learn** classification with imbalanced classes

    conda install -c conda-forge imbalanced-learn

- **aequitas** Bias and Fairness Audit Toolkit

    Installed as: pip install aequitas

Important note

- As each student may have their own hardware and software configuration, we cannot guarantee that it will be possible to install all the packages on all the configurations.

- If a problem is encountered:
  - Follow the recommended steps in the notebook or lecture notes
  - Google for similar errors (sometimes a package should be upgraded or downgraded to work within a specific setting)
  - Consider making a new environment to test installation of a new package before you start downgrading or upgrading already installed packages in a working environment
  - Ask TA or instructor for help
  - Use Google colab notebook, it will be accepted for homework and project submissions.

Step 4: Use the conda environment in your jupyter notebook

- Install ipykernel in *ml2025* environment (allows Jupyter to recognize the environment as a kernel)

  conda install ipykernel

- Add the *ml2025* environment as a kernel for Jupyter

  python -m ipykernel install --user --name=ml2025

- Deactivate the environment

  conda deactivate

To list existing kernels

  jupyter kernelspec list

Note: A notebook **kernel** is a "computational engine" that executes the code contained in a Notebook document

Step 4: Use the conda environment in your jupyter notebook

- From Anaconda Navigator Launch **JupyterLab**
  (not Jupyter Notebook!)
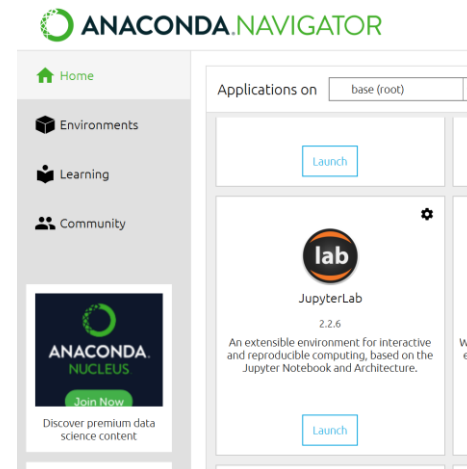
- Select from list of kernels the kernel *ml2025*

Scikit-learn: Machine Learning in python

Good reference on coding how-to's

https://jakevdp.github.io/PythonDataScienceHandbook/

Python Data Science Handbook

Jake VanderPlas

O'REILLY®

Python
Data Science
Handbook

ESSENTIAL TOOLS FOR WORKING WITH DATA

powered by
jupyter

Jake VanderPlas

3. Data Manipulation with Pandas

- Introducing Pandas Objects
- Data Indexing and Selection
- Operating on Data in Pandas
- Handling Missing Data
- Hierarchical Indexing
- Combining Datasets: Concat and Append
- Combining Datasets: Merge and Join
- Aggregation and Grouping
- Pivot Tables
- Vectorized String Operations
- Working with Time Series
- High-Performance Pandas: eval() and query()
- Further Resources

5. Machine Learning

- What Is Machine Learning?
- Introducing Scikit-Learn
- Hyperparameters and Model Validation
- Feature Engineering
- In Depth: Naive Bayes Classification
- In Depth: Linear Regression
- In-Depth: Support Vector Machines
- In-Depth: Decision Trees and Random Forests
- In Depth: Principal Component Analysis
- In-Depth: Manifold Learning
- In Depth: k-Means Clustering
- In Depth: Gaussian Mixture Models
- In-Depth: Kernel Density Estimation
- Application: A Face Detection Pipeline
- Further Machine Learning Resources

Good reference on coding how-to's

https://vedraiyani.github.io/notes-1/ipynb/index.html

Preprocessing Structured Data

- Convert Pandas Categorical Data For Scikit-Learn
- Delete Observations With Missing Values
- Deleting Missing Values
- Detecting Outliers
- Discretize Features
- Encoding Ordinal Categorical Features
- Handling Imbalanced Classes With Downsampling
- Handling Imbalanced Classes With Upsampling
- Handling Outliers
- Impute Missing Values With Means

- Imputing Missing Class Labels
- Imputing Missing Class Labels Using k-Nearest Neighbors
- Normalizing Observations
- One-Hot Encode Features With Multiple Labels
- One-Hot Encode Nominal Categorical Features
- Preprocessing Categorical Features
- Preprocessing Iris Data
- Rescale A Feature
- Standardize A Feature

Trees And Forests

- Outlier Detection With Isolation Forests
- Adaboost Classifier
- Decision Tree Classifier
- Decision Tree Regression
- Feature Importance
- Feature Selection Using Random Forest
- Handle Imbalanced Classes In Random Forest

- Random Forest Classifier
- Random Forest Classifier Example
- Random Forest Regression
- Select Important Features In Random Forest
- Titanic Competition With Random Forest
- Visualize A Decision Tree

Nearest Neighbors

- Identifying Best Value Of k
- K-Nearest Neighbors Classification

- Radius-Based Nearest Neighbor Classifier

Feature Engineering

- Dimensionality Reduction On Sparse Feature Matrix
- Dimensionality Reduction With Kernel PCA
- Dimensionality Reduction With PCA
- Feature Extraction With PCA
- Group Observations Using K-Means Clustering

- Se
- Se
  TS
- Us
  Di

Feature Selection

- ANOVA F-value For Feature Selection
- Chi-Squared For Feature Selection
- Drop Highly Correlated Features
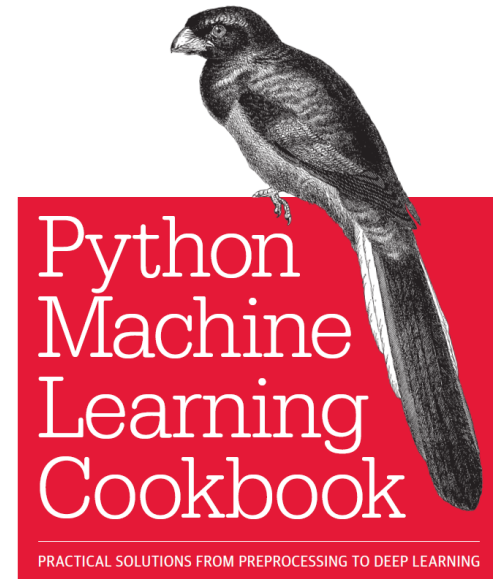
- Re
- Va
- Va

Model Evaluation

- Accuracy
- Create Baseline Classification Model
- Create Baseline Regression Model
- Cross Validation Pipeline
- Cross Validation With Parameter Tuning Using Grid Search
- Cross-Validation
- Custom Performance Metric
- F1 Score

- Ge
- Ne
- Pl
- Pl
- Plot The Validation Curve
- Precision
- Recall
- Split Data Into Training And Test Sets

Model Selection

- Find Best Preprocessing Steps During Model Selection
- Hyperparameter Tuning Using Grid Search

- Hyperparameter Tuning Using Random Search
- Model Selection Using Grid Search
- Pipelines With Parameter Optimization

Linear Regression

O'REILLY®

Python Machine Learning Cookbook

PRACTICAL SOLUTIONS FROM PREPROCESSING TO DEEP LEARNING

Chris Albon

Using colab

- Google Colab (https://colab.research.google.com/) is an extension of Jupyter notebook that runs on the Google Cloud. This platform provides various different computing resources, such as CPUs, **GPUs** free of charge.

- Colab allows you to use and share Jupyter notebooks with others without having to download, install, or run anything.

- Google Colab has a 'maximum lifetime' limit of running notebooks that is 12 hours with the browser open, and the 'Idle' notebook instance is interrupted after 90 minutes.

- Colab notebooks can be shared with other users and opened by multiple users at a time. If one person makes a change, the others will be able to see the change after a short delay. However, if two people edit the document at the same time, one person's changes must be discarded upon refreshing.

- https://research.google.com/colaboratory/faq.html