

FEDS - Modelling and Analysing of Forestal and Environmental Data

An introduction to data science

How to learn?

Do not expect the teachers to teach you.

They will present some information to you, but it is entirely 100% up to you to either make the most of it, or waste your time here, and go home and get a normal dumb job.

How I will present you the informations!

- I will jump between websites, slides, coding ... will be intense
- I offer you to use some of your time and do challenges
- I get you to know stuff you have maybe never seen before
- I ask you to read and try some things just on your own
- I am here to answer questions popping up
- but you are also here to answer questions :)

Topics we will touch in this course

Setting up a data science environment

Learn to manage data, analyse data, visualise data, and model data

Learn the Python language and data science libraries - today's most used data science tools

Learn to use Jupyter notebooks and Jupyter lab as interactive work tool

Learn to use machine learning methods

Learn to create computational essays


Why Python?



for data processing

- We can use all of them.
- Some are better fitting.
- We need sometimes to combine them.
- There are more.

<https://www.tiobe.com/tiobe-index/>



TIOBE

the software quality company

[About us](#)

[Join TIOBE](#)

[News](#)

[Coding Standards](#)

[TIOBE Index](#)

















[Contact](#)

Products

Quality Models

Markets

Schedule a demo

Jan 2023	Jan 2022	Change	Programming Language		Ratings	Change
1	1			Python	16.36%	+2.78%
2	2			C	16.26%	+3.82%
3	4	▲		C++	12.91%	+4.62%
4	3	▼		Java	12.21%	+1.55%
5	5			C#	5.73%	+0.05%
6	6			Visual Basic	4.64%	-0.10%
7	7			JavaScript	2.87%	+0.78%
8	9	▲		SQL	2.50%	+0.70%
9	8	▼		Assembly language	1.60%	-0.25%
10	11	▲		PHP	1.39%	-0.00%
11	10	▼		Swift	1.20%	-0.21%
12	13	▲		Go	1.14%	+0.10%
13	12	▼		R	1.04%	-0.21%
14	15	▲		Classic Visual Basic	0.98%	+0.01%
15	16	▲		MATLAB	0.91%	-0.05%
16	18	▲		Ruby	0.80%	-0.08%

Is it just a hype?

No!

Long-term history of programming language popularity

Programming Language	2023	2018	2013	2008	2003	1998	1993	1988
Python	1	5	8	7	13	28	17	-
C	2	2	1	2	2	1	1	1
Java	3	1	2	1	1	17	-	-
C++	4	3	4	3	3	2	2	6
C#	5	4	5	8	12	-	-	-
Visual Basic	6	15	-	-	-	-	-	-
JavaScript	7	7	10	9	8	21	-	-
Assembly language	8	12	-	-	-	-	-	-
SQL	9	-	-	-	7	-	-	-
PHP	10	8	6	5	6	-	-	-
Objective-C	16	18	3	45	47	-	-	-
Ada	29	27	17	18	15	7	8	2
Lisp	31	31	13	15	14	9	5	3
Pascal	242	128	15	20	99	11	3	7
(Visual) Basic	-	-	7	4	4	3	6	5

<https://www.tiobe.com/tiobe-index/>

How to set up a data science environment for use with Python?

Local computer

Desktop/Laptop/Server



Install Docker or Docker Desktop



Choose Jupyter Docker image

<https://jupyter-docker-stacks.readthedocs.io/en/latest/>



Run Jupyter notebook or Jupyter lab in your browser

Pro's:

- Image/Container with up-to-date libraries
- Multiple instances
- Fast testing possible

Con's:

- Need to learn Docker basics



Install Miniconda or Anaconda



1. Create virtual environment
2. install needed 3rd-party libraries
3. install Jupyter



Run Jupyter notebook or Jupyter lab in your browser

Pro's:

- Full control on the installed libraries
- Multiple venv's possible

Con's:

- Need to learn venv basics
- Need to keep libraries manually up-to-date

Cloud service

Google Colab



Access Google Colab via your browser
<https://colab.research.google.com/>



1. Create a GitHub repository to exchange data files
2. Google drive etc is also possible

Pro's:

- No installation of Python
- Instantly available
- Fast testing possible

Con's:

- Need some repository for data in/out
- Need paid version if you want to use it persistently (better choice for resources like CPU/GPU)

What we will use?

- We will use the Google Colab (cloud service)



You need to have a Google account!

- We also use a Github repository to organise the course



You need to have a Github account!

Jupyter notebooks



The concept of computational documents

- Combines text and code
- Let you organise your data processing
- Let you visualise your results (fast)
- You have a good documentation of your work
- The paper is almost written on the fly 😊

Let's see an example

SimpleExample.ipynb

File Edit View Insert Runtime Tools Help Last saved at 20:52

Table of contents

- A simple example how to use Colab with Python
- Getting the data and make an overview what we got
 - 2.1 Loading and visualising the data
 - 2.2 Load the data
 - 2.3 Plot a scatter plot of the data
- Section

[9] 1 import pandas as pd
2 import matplotlib.pyplot as plt

2.2 Load the data

[8] 1 df = pd.read_csv("./elusad_puud.csv")
2 df.head()

	H	HV	HKO	PL	D	ER	PIND	MOODETUD_P	KKT	DBH	H_VALEM	V	S
0	8.2	2.6	0.0	KS	7.85	10	0.64010	0.64010	JO	7.85	10.694	0.02270	1.2016
1	18.9	5.9	3.3	KU	20.20	7	1.36565	0.09597	MO	20.20	19.916	0.31732	8.0531
2	26.1	14.7	9.2	KU	30.25	10	0.64010	0.64010	JO	30.25	23.294	0.90391	15.8536
3	10.9	4.9	0.5	KU	11.40	10	0.64010	0.64010	JO	11.40	14.352	0.06580	2.7994
4	16.8	4.6	1.0	KU	23.25	10	0.64010	0.64010	JO	23.25	21.161	0.37236	8.1901

2.3 Plot a scatter plot of the data

[13] 1 plt.scatter(df.DBH, df.H)
2 plt.xlabel("diameter at breast height (cm)")
3 plt.ylabel("tree height (m)");

[] 1

Make it happen...

Instead of too much theory



we will cook



and experiment