Outline **“Python Programming and Machine Learning for Economic Research”**, Doctoral course

The course allows doctoral students to conduct their own empirical, version-controlled projects in Python and introduces various Machine Learning models. Most units are highly practical and complemented with a training session.

Required preparation:

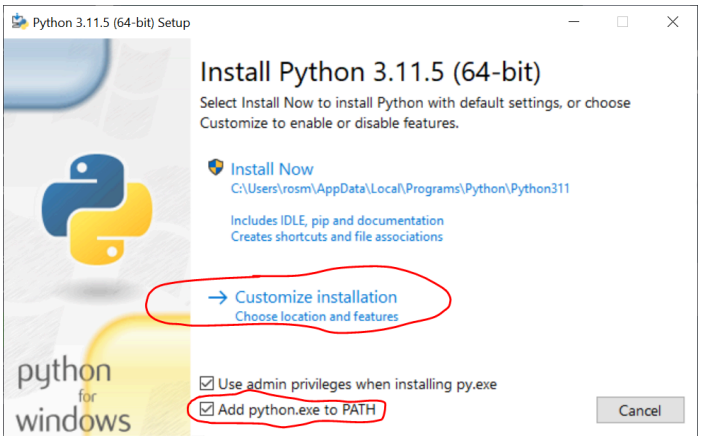
- Python beginners: Finish the self-paced, interactive online courses for Python (about 2 hours) at https://www.udemy.com/course/learn-python-3-from-scratch-python-for-absolute-beginners/ and forward a screenshot of the last page showing your name (alternatively, the certificate you receive in the end) to Michael.Rose@ip.mpg.de
- Read the required readings

Required Readings:

- Shapiro, J. and M. Gentzkow: “Code and Data for the Social Sciences: A Practitioners Guide”
- Athey, S. and G. Imbens (ARE 2019): “Machine Learning Methods That Economists Should Know About”
- Ash, E. and S. Hansen (ARE 2023): “Text Algorithms in Economics”

Required software:

- (All users:) install a modern text editor of your choice, e.g., Sublime Text 4 or Notepad++ (only Windows); Notepad will not work
- (Windows users:) Python 3.7 or higher, but NOT 3.12 - during installation, make sure both python and pip are part of the PATH (while installing Python, check “Add python.exe to PATH”, then click “Customise Installation” as in the figure on the right and on the next page check “Add pip to PATH”). Note: the use of anaconda is *discouraged* (see Notes below).
- (All users:) PyCharm Community Edition (default settings are recommended)
- (All users:) git (during installation, select the editor from step 2 as default editor) - optionally, add any of the GUI clients



Day 1: Introduction

General Introduction

- Getting to know each other
- Learning outcomes, course outline, explanation of examination
- Definitions, prerequisites and limitations of Machine Learning

Python Project Management

- The PyCharm IDE
- Introduction to pandas, matplotlib and seaborn
- Version control with git
- Collaborating with GitHub/GitLab
- Debugging

Day 2: Traditional Natural Language Processing

- Examples
- text cleaning and encoding problems, stemming
- Introduction to nltk
- Text vectors
- Plug-and-play text analysis: sentiment analysis, Readability analysis, wordclouds
- [Excourse on Latent Dirichlet Analysis]

Day 3: Unsupervised Machine Learning

- Examples
- Introduction to sklearn
- Topic modelling with text: Agglomerative clustering, Hierarchical clustering
- Other clustering methods: DBSCAN
- Cluster evaluation metrics
- [Principal Component Analysis]

Day 4: Supervised Machine Learning

- Examples
- Workflow I: Split, train, evaluate
- Distances and evaluation metrics
- Linear models and regularisation (Ridge, LASSO)
- Excourse: Machine Learning for Econometricians
- Neural networks
- Workflow II: Cross-validation, pipelines, grid search

Day 5: Advanced Natural Language Processing

- Embeddings
- BERTs and GPTs
- Introduction to transformers
- Fine-tuning BERT-based language models
- [Excourse: on ChatGPT]

Outlook

- Python Style Guide PEP8
- My own workflow
- Packages for advanced ML
- [Info re Examination]

Teacher

Michael E. Rose, PhD; Post-Doctoral Researcher at the Max Planck Institute for Innovation and Competition, Munich, Germany (Department "Innovation and Entrepreneurship", headed by Dietmar Harhoff)

- Teaching experience:
  - This course (Python and Machine Learning, for PhD/Dr students) at TUM School of Management, Kiel Institute for the World Economics, U Zurich/ETH Zurich (2x), LMU Munich, ifo institute (2x), and Georgia Tech's Scheller College of Business
  - Computational Mathematics (Matlab, SQL, VBA, Excel) for Master's students: at U Cape Town
  - Time Series Econometrics (Python, for Master's students) at U Cape Town
- Daily usage of Python, ML in own research
- Lead development of two python packages (pybliometrics and sosia)
- Co-development of Logic Mill (https://arxiv.org/abs/2301.00200) and PaECTER (pending)
- Co-PI in Academic Research Project "Tracing the flow of knowledge from science to technology using deep learning" 2021 of the European Patent Office

Related Literature

- Downey, Allen B.: "How to think like a Computer Scientist", Chapters 1-3, 5, 8, 10-12
- Shapiro, J. and M. Gentzkow: "Code and Data for the Social Sciences: A Practitioners Guide"
- Ash, E. and S. Hansen (ARE 2023): "Text Algorithms in Economics"
- Gentzkow, M., B. Kelly and M. Taddy (JEL 2019): "Text as Data"
- Mueller, A. and Sarah Guido: "Introduction to Machine Learning with Python"
- Athey, S. and G. Imbens (ARE 2019: "Machine Learning Methods That Economists Should Know About"
- Mullainathan and Spiess (JEP 2017): "Machine Learning: An Applied Econometric Approach", 31 (2).
- Athey, S. (Science 2019): "Beyond Prediction: Using Big Data for Policy Problems".

Notes:

- We are going to use **plain** Python 3, while I recommend to not use anaconda. On the first day I'll explain why. Anaconda users might not be able to replicate everything.
- The Integrated Developer Interface we use is **PyCharm Community Edition**. spyder offers less functionality, but I cannot offer support for spyder.
- Recommended text editors: **Sublime Text 2** or **Notepad++**. Do not use Windows' native editor Notepad!
- Slides, exercises and relevant data will be shared via a private shared **GitHub** repository. A (free) profile on GitHub is necessary (pick a proper username).