# Advanced Data Analysis with Python

M. Fuat Kına
mkina@ku.edu.tr

# Course objectives

- Learn Python 3 syntax
- Understand basic programming concepts
- Understand advanced data analysis problems and the needed tools to solve them
- Establish a basic understanding of machine learning concepts and algorithms

# Why Python?

- Great for beginners and advanced use
  - Easily readable code
  - Online resources
- Widely used, especially in scientific computing
- Powerful
  - Advanced data analysis techniques
  - Machine learning modules
- Open-source
- Alternatives for data analysis: R, STATA, SPSS, GIS programs, (ArcGIS, QGIS, Geoda) etc.

# Course content

- Python basics (data types, lists, sets, dictionaries, basic operations, if statements, functions, and loops)
- Data collection (web scrapping, APIs)
- Working with data, creation and manipulation (numpy, matplotlib, pandas)
- Advanced data analysis techniques
  - Conventional models
  - Causal Inference
  - Spatial Analysis
  - Bayesian statistics
  - Machine Learning

# Course materials

- VanderPlas, Jake. 2016. Python Data Science Handbook: Essential Tools for Working with Data. O'Reilly Media. Available at: https://jakevdp.github.io/PythonDataScienceHandbook/
- Shaw, Zed A. 2017. Learn Python 3 the Hard Way: A Very Simple Introduction to the Terrifyingly Beautiful World of Computers and Code (Zed Shaw's Hard Way Series). 1st Edition. Addison-Wesley. Available at: https://learnpythonthehardway.org/python3/
- The Official Documentation for Python. Available at: https://docs.python.org/3/
- Weekly readings and links for some online sources and Python scripts
- Two more books:
  - Angrist, J.D., & Pischke, J.S. (2009). Mostly Harmless Econometrics: An Empiricist's Companion, Princeton: Princeton University Press.
  - LeSage, J., & Pace, R. K. (2009). Introduction to Spatial Econometrics. Chapman and Hall/CRC.

# Datasets

When we start data analysis, we are going to use two datasets:

- Brazil MST data: Bolsa Familia ~ Land occupations, poverty, infant mortality, etc.
  - Geographical data
  - Instrumental variable analysis
- Twitter data: University ranking ~ Twitter habits, network, followers, unique words, etc.
  - Good for Machine Learning problems

# Grading

- Three assignments: %15 x 3
  - Solve the assigned problem
  - Scrape data of your final project or build a regressor function for OLS
  - Shared task: Create different ML models
- Midterm report: %5
- Final project: %40
  - Presentation
  - Github page
  - Written report
  - Python code
- Presentation of a weekly reading: %10
- Prepare yourself before class (must)

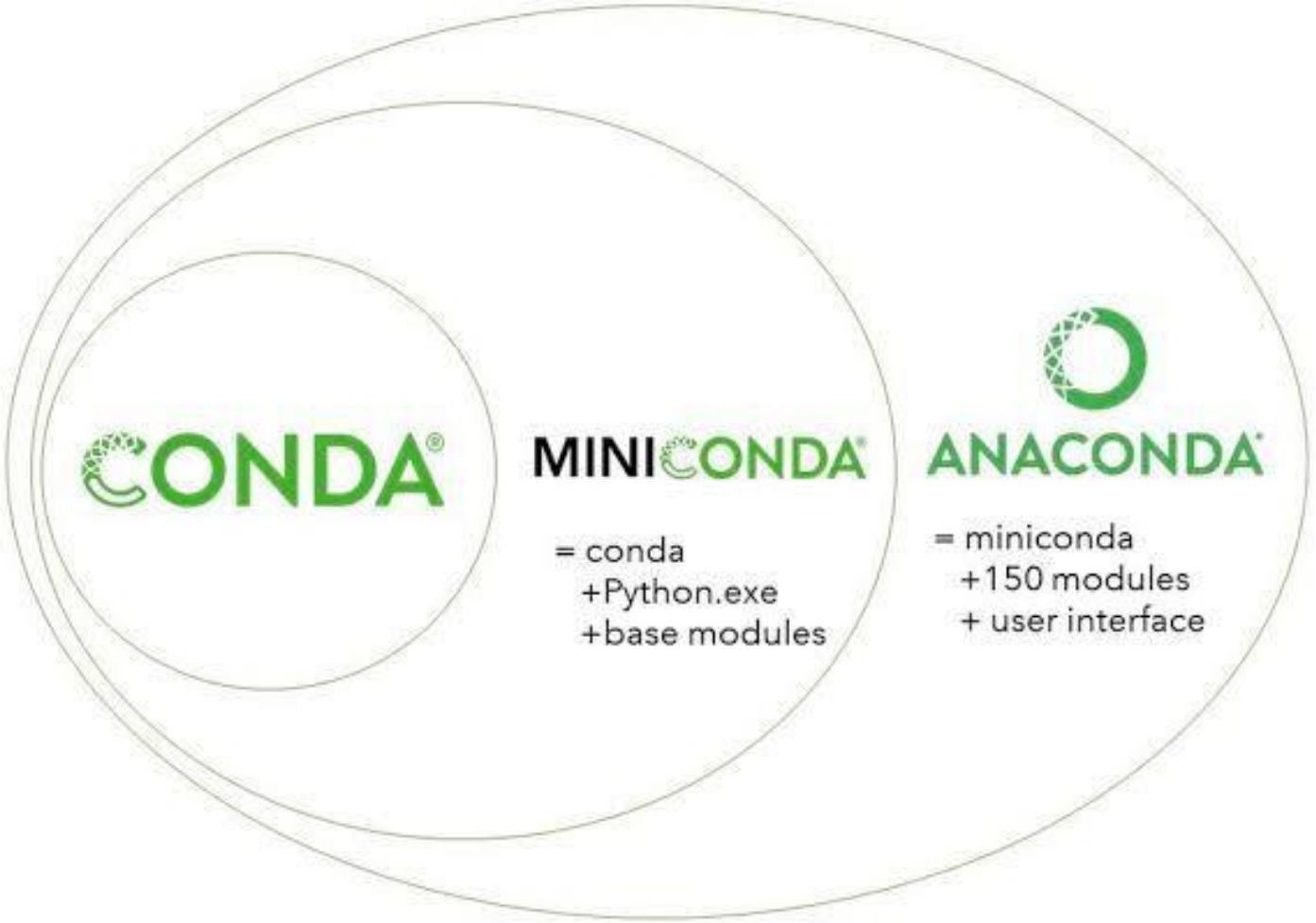
# Let’s look at the syllabus

Let’s look at installation guideline!

You can run python without downloading anything

https://colab.research.google.com/

# Conda, miniconda, anaconda



# Conda environments

- What is an environment?
  - A conda environment is a directory that contains a specific collection of conda packages that you have installed. For example, you may have one environment with NumPy 1.7 and its dependencies, and another environment with NumPy 1.6 for legacy testing.
  - https://docs.conda.io/projects/conda/en/latest/user-guide/concepts/environments.html#:~:text=A%20conda%20environment%20is%20a,NumPy%201.6%20for%20legacy%20testing.
- How to manage an environment?
  - https://docs.conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html#id2

# How to use Python?

- Go for the installation guidelines
- Download the Python 3
- Install Anaconda
- Inspect Google Colab
- Play with conda
- Work on Shaw, Exercises 0-40