Python for Big Data Machine Learning & Data Science

Thomas Schwarz, SJ
MSCS
Marquette University
Milwaukee, WI
Why Python

- Universal, accessible language
  - Clear and simple syntax
    - Python philosophy: The frequent should be easy
    - Made for reading
    - Used for fast prototyping
- Excellent support community
  - Help for beginners and experts is easily available
Why Python

- Universal Language
  - Serves many different constituencies
    - Examples:
      - Gaming: AI engine is usually in Python
      - String processing: Basis for digital humanities and data wrangling
      - Many extension modules
        - With scipy or numpy, fast programs for scientific programming
        - Use pyplot for good quality graphics
        - ...
      - Notebooks based on Python (Jupyter) integrate presentation, data, and programs
Why Python

• Python in Data Science

KDNuggets Analytics, Data Science, Machine Learning Software Poll, 2016-2018

- Python
- RapidMiner
- R
- SQL
- Excel
- Anaconda
- Tensorflow
- Tableau
- scikit-learn
- Keras
- Apache Spark

- 2018 %share
- 2017 %share
- 2016 %share
How to get Python

• If you are already working in Data Science / Big Data:
  • You probably want a package such as Anaconda

• We will use the simplest interface, because we want to learn Python first.
  • Python comes in two varieties: Python2.7 and Python3.x
    • We use the Python3.x family
How to get Python

• Go to python.org
  • Go to download
    • Python will select the right 3.x version for your OS

• Note: You will have python 2.7 already installed in MacOS, because it is used internally

• This should install Python3.6 for you
  • The interface that we are using is called IDLE
Using IDLE

• IDLE is written in Python using TkInter, one of Python’s GUI interfaces
  • Usually runs fine, be careful about non-latin characters
  • Interface depends slightly on the OS
IDLE

- IDLE is an interactive Python interpreter
- Can be used as a desk calculator
- Allows you to create new files