University of Essex

Python
Data Science Setup

Akitaka Matsuo

Overview of Week 03-04

University of Essex

- We will see the key packages
  - Week 03: NumPy + Pandas
  - Week 04: Pandas + Matplotlib (and related packages)

NumPy

University of Essex

- NumPy
  - Numerical Python
- Python is a general purpose language
  - Normal computation on Python is not that fast
- Need a package for efficient numerical computation
- Basic for all python computation oriented packages

Pandas

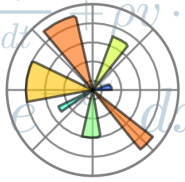
University of Essex

- Pandas: data manipulation package
- What it does?
  - Reading/writing data from various file formats
  - Create **DataFrame** object
  - Sophisticated indexing
    - Reshaping (e.g. pivoting) the data
    - Filtering (=subsetting) the data
    - Merging the data
    - Generating summaries by group

Matplotlib

University of Essex

- Graphical extension of NumPy
  - Generate various plots from NumPy objects
- Default plot options looks outdated (especially until recently)…, but
- But matplotlib is easy to extend
- Extension (or wrapper packages)
  - seaborn
  - ggplot



matplotlib

Machine learning and deep learning

University of Essex

- *scikit-learn*
  - Comprehensive machine learning package
  - Provides wide variety of methods
    - Classification with SVN, kNN, random forest
    - Regression with penalized regression
    - Clustering
  - Data-preprocessing
  - API for model selection/tuning
  - We will see a bit in the last week
- *TensorFlow*
  - A package for deep-learning with neural networks