DATA SCIENCE & MACHINE LEARNING
#3 Data Manipulation and Visualization
(1 Feb 2019)

Lecturers: António Raimundo & Daria Baikova

#3 Data Manipulation

Contents (25 JAN 2019)

- Introducing Pandas Objects
- Data Indexing and Selection
- Operating on Data in Pandas
- Handling Missing Data
- Hierarchical Indexing
- Combining Datasets: Concat and Append
- Combining Datasets: Merge and Join
- Aggregation and Grouping
- Pivot Tables
- Vectorized String Operations

Python Pandas

- Open source python module that provides high-performance data structures and data analysis tools
- Runs on top of NumPy
- Very popular for Data Science

NumPy

- Low-level data structure (np.array)
- Large multidimensional arrays and matrices
- Mathematical operations

Pandas

- High level data structures (DataFrame)
- Better for handling tabular data and time series
- Data alignment, missing data, statistics, merge, join..
- Can use NumPy functions for manipulating data