CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2020
Doc 4 SciPy
Jan 28, 2020

Copyright ©, All rights reserved. 2020 SDSU & Roger Whitney, 5500 Campanile Drive, San Diego, CA 92182-7700 USA. OpenContent (http://www.opencontent.org/openpub/) license defines the copyright on this document.

SciPy

Part of Anaconda installation

https://scipy.org

NumPy
- N-dimensional homogeneous array
- Array manipulation, indexing, shape, slicing
- Linear algebra, Fourier transform, random number

Pandas
- Data structures & data analysis

Matplotlib
- 2D plotting

Sympy
- Symbolic math

SciPy
- Scientific computing
- Integration, optimization, signal processing, Sparse graphs
- Linear algebra, Statistics, multidimensional image processing

Other Libraries of Interest

Statsmodels

- http://www.statsmodels.org/

scikit-learn

- Machine Learning
- https://scikit-learn.org/

sklearn-pandas

TenserFlow

- https://www.tensorflow.org
- Numerical computation using data flow graphs
- Targets CPU, GPU, server, mobile, etc

Visualization

Altair
    Declarative statistical visualization

Bokeh
    Interactive, web

Seaborn
    High level, based on matplotlib

yhat/ggpy
    Grammar of Graphics,
    R's ggplot2

Ploty
    Interactive, web shareable
    Convert matplotlib, ggplot, Seaborn to interactive web-based plots

Pandas

DataFrame

- Think SQL table without SQL

Data IO

- CSV, text files, MS Excel files
- SQL databases
- HDF5

Missing data

Reshaping, slicing, subsetting, column inserting-deleting

Group by, merging data sets

Time series

Panda Data Structures

Series

- 1D labeled array with index

DataFrame

- 2D labeled data structure with columns
- Columns can have different data types
- Think spreadsheet or SQL table
- Most commonly used