

XSEDE Tutorial Proposal: Data Scientist's Python Toolbox

Abstract:

This tutorial is an intermediate level course on tackling the problems facing data scientist using Python. Python is a high-level object oriented language that has found wide acceptance in the scientific computing/ data science community. Ease of use and an abundance of software packages are some of the few reasons for this extensive adoption.

Pandas is a high-level open-source library that provides data analysis tools for Python. It provides an efficient and comprehensive platform for a large number of analytics problems. For generating sophisticated visualizations two packages: Seaborn and Plotly are introduced. While Seaborn is aimed at Statisticians, Plotly provides a rich, interactive visualization framework which is ideal for visualizing large data. Plotly also allows visualization-rich dashboards which can be shared online. To conclude, out-of-core computing with Dask/Blaze is introduced for those datasets that won't quite fit into memory. The goal of dask is to “extend the size of convenient datasets from ‘fits in memory’ to ‘fits on disk’” effectively fitting between Pandas and PySpark in the Python ecosystem for analytics.

Instructors:

Srijith Rajamohan, Ph.D, VT ARC, Computational Scientist,
Virginia Tech

Software Requirements:

Participants should bring a laptop with Anaconda Python 3.5 installed. The instructor will also bring a VM image with the necessary tools and scripts installed for users who prefer this. The data to be used for this class will be available for download before the class. Also, this will be provided on USB drives at the tutorial venue.

Tentative Agenda:

9:00 - 10.00 Introduction to Pandas for Data Science

10.00 - 10.45 Introduction to Visualization and Plotting with Seaborn and Plot.ly

11.00 - 11.45 Out-of-Core Computing with Dask/Blaze for 'Biggish' Data

Draft Slides: Draft slides are posted are available at <https://bitbucket.org/sjster/XSEDE17/downloads>. Please click 'Download repository'. The zip file should contain a pdf for the slides.

Network Outage Contingency Plans: The instructor has taught semester-long courses on Python. All data including necessary software, scripts and datasets will be available on a USB drive for users, minimal network connectivity is required.