Visualization and Analysis with Python

Jonathan Woodring* Los Alamos National Laboratory
Joseph A. Cottam† CREST, Indiana University
Peter Wang‡ Continuum Analytics, Inc.
Julien Jomier§ Kitware Inc.

ABSTRACT

Python is a powerful development, computational, programming environment due to the wide variety of libraries developed for it, and importantly, the enthusiastic, active development and user community. One of the areas where Python excels at is visualization and analysis of data, because of several high-quality modules for both simple and advanced visualization. This tutorial will cover the following visualization capabilities in Python: interactive plotting with *IPython* and *matplotlib*, data analysis with *numpy* and *pandas*, building web visualizations with *Bokeh*, and Python integration with 3D visualization in *VTK* and *ParaView*. This tutorial is intended for intermediate level participants that have a basic understanding of the Python language and development environment (i.e., the student ought to have an understanding of native data structures, file I/O, and is able to develop and run simple programs). Beginner participants are welcome, but Python fundamentals, such as language constructs, "hello world," and program execution will not be covered in this tutorial.

Index Terms: D.2.11 [Software]: Software Engineering—Software Architectures; E.m [Data]: Miscellaneous; I.3.4 [Computing Methodologies]: Computer Graphics—Graphics Utilities

1 INTRODUCTION

Python is a popular interpreted [1] language, currently found in the top 10 of programming language popularity lists. It has a young, active user and development community, expanding the library base to many different computational and application fields.

In particular, it has found strong popularity in the scientific and analytics community, through powerful numeric and computing libraries, like *numpy*, *scipy*, and *pandas*, visualization libraries like *matplotlib*, and interactive environments, like *IPython*. This allows Python to act as a free [2] replacement for Matlab, SAS, and other proprietary analysis tools, living along side of *GNU R*, as another popular analysis environment. In addition to those reasons, Python is a superb binding and embedded language for direct analytics of *in situ* data, as it is easy to extend existing code and libraries through C bindings and *Cython*, such as *VTK* and *ParaView*.

Thus, Python is an ideal environment for development and interactive visualization and analysis of data. The multitude of graphing and plotting tools along with many libraries for computation and data analysis, enable a huge variety of analysis in application domains. We will teach an intermediate level tutorial for Python, which will provide an introduction to several modern visualization and analysis modules for Python. Beginners to Python are welcome to attend this course, but language, data structures, and execution fundamentals will not be taught in this tutorial.

This tutorial is an expansion from the one conducted at IEEE VisWeek 2013. The expansion reflects feedback from last year's participants, and a new set of instructors not involved in last year's effort.

*e-mail:woodring@lanl.gov
†email:jcottam@indiana.edu
‡e-mail:pwang@continuum.io
§e-mail:julien.jomier@kitware.com

[1] IronPython, Jython, PyPy, Cython, C-based libraries, etc. speed up Python through native machine code, virtual machines, source-to-source translation, and just-in-time compilation.

[2] Free as in both beer and speech.

2 EDUCATIONAL GOALS

1. Generate excitement and interest for Python
2. Provide a reference for contacting the user community
3. Introduction to Interactive Plotting with Python
4. Introduction to Data Analysis and Manipulation with Python
5. Introduction to Web Visualizations with Python
6. Introduction to VTK/ParaView with Python

3 TUTORIAL OUTLINE

The tutorial will begin with an introduction of interactive plotting and visualization of data in *IPython Notebooks* using *matplotlib*. This will be followed up with *numpy* and *pandas*, for exploring data analysis, manipulation, and slicing in Python. *Bokeh* will be introduced, which can extend IPython with web analytics and generate stand-alone web visualizations. Finally, the Python bindings of *VTK* and *ParaView* will be taught, showing examples of 3D and large-scale data visualization pipelines.

Introduction: (10 minutes)

IPython and matplotlib: (50 minutes), instructor Joseph Cottam

Interactive plotting via IPython and matplotlib is one of the most widely used cases of Python visualization and analysis.

1. Introduction to the IPython Notebook environment: how to start the notebook server and load and save a notebook state
2. Loading and caching a data set: how to bring data into the IPython environment and store intermediate results
3. Plotting with matplotlib: examples of different data plots of a data set
4. Creating sharable documents with IPython: how to save the Notebook into pdf and html formats to share with others

Break (10 minutes)

numpy and pandas: (50 minutes), instructor Jonathan Woodring

numpy and pandas provide fast array and columnar data representations for C-speed manipulation and analysis of data sets.

1. Introduction to numpy data structures: examples of array and matrix data in numpy
2. Data analysis and computation: how to do computation with numpy arrays and analyses scipy functions
3. Creating data with pandas: examples of how data are represented with pandas and sqlite databases

4. Data manipulation with pandas: how to slice, join, and subset data within pandas

Bokeh: (50 minutes), instructor Peter Wang
Generation of web graphics with Bokeh and Python for interactive and collaborative sharing of data

1. Overview of web visualization options for Python: give the context of different web tools, in addition to Bokeh

2. How to build basic interactive web graphics using Bokeh: examples of generating web graphics with Python data

3. Integrating Bokeh plots with IPython notebook: examples of integrating Bokeh into IPython for interactive visualization

4. Hosting stand-alone server-based plots: how to share the web graphics with others over the internet

Break (10 minutes)
VTK and ParaView: (50 minutes), instructor Julien Jomier
VTK and ParaView are scriptable and extensible through Python, allowing for easy programmability of visual analytics

1. Load data into VTK: how to bring external data into VTK Python

2. Basic data visualization with VTK's Python bindings: examples of generating VTK pipelines with Python

3. Overview of Python scripting capabilities in ParaView: show how ParaView is extensible through Python

4. Create simple Python plugins in ParaView: examples of programming ParaView with Python filters

Conclusion/Wrapup (10 minutes)

4 Instructor Biographies

Joseph Cottam is a Post-Doctoral researcher at Indiana University in the Center for Research in Extreme Scale Technologies (CREST). He received his Ph.D. in computer science from Indiana University in 2011. His research falls in the intersection between visualization, large-scale systems, and programming languages. In particular, Joseph has built systems for large and streaming data information visualization. He is a contributor to the Bokeh visualization toolkit.

Jonathan Woodring is a research scientist at the Los Alamos National Laboratory in the CCS-7 Data Science at Scale team. He received his Ph.D. in computer science from the Ohio State University in 2009, specializing in computer graphics and scientific visualization. His primary research areas include visualization and analysis, data intensive supercomputing, and high-performance computing for large-scale scientific simulations. Jon is a Python enthusiast and advocate, and has been a user since 2000, developing the majority of his visualization and analysis research in Python over the past 14 years.

Peter Wang is the co-founder and president of Continuum Analytics, a young startup focused on taking Python analytics, scientific computing, and data visualization to the next level. Continuum Analytics develops many of the Python tools in use that enable analytics, visualization, and large-scale data processing, and teaches courses in Python for different application domains. Peter has been professionally developing with Python for almost 15 years, and is an active participant in the scientific Python and PyData communities. Peter holds a BA in Physics from Cornell University.

Julien Jomier is the director of Kitware's European subsidiary in Lyon and is one of the developers of VTK and ParaView. He is also the main architect of CDash, an open-source, distributed, software quality system, a companion to CMake and CTest. Julien received both his B.S. and M.S in Electrical Engineering and Information Processing in 2002 from ESCPE-Lyon and an M.S. in Computer Science from The University of North Carolina at Chapel Hill in 2003. Julien has taught many courses and tutorials on visualization and image processing around the world, and is leading the development of the MIDAS Journal and Insight Journal, an electronic journal promoting open-science.

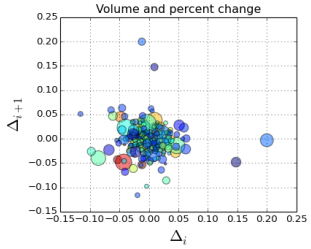

Figure 1: matplotlib

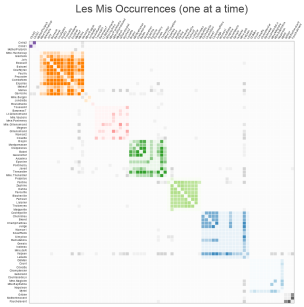

Figure 2: Bokeh

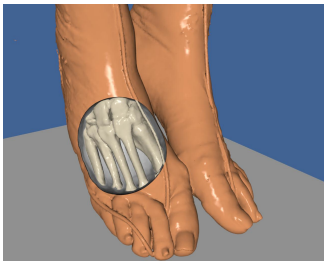

Figure 3: VTK