Scientific Computing using Orange - Ankit Mahato

About the speaker:

Ankit is a Product Manager with 3+ years of industrial experience in machine learning, quantitative modelling, data analytics and visualization. Over the years, he has developed an expertise in handling the entire data analytics pipeline comprising – ingestion, exploration, transformation, modeling and deployment. He is a polyglot programmer with an extensive knowledge of algorithms, statistics and parallel programming. He has shipped multiple releases of DB Lytix, a comprehensive library of over 800 mathematical and statistical functions used widely in data mining, machine learning and analytics applications, including "big data analytics".

A die hard Pythonista, Ankit is an open source contributor and a former Google Summer of Code 2013 scholar (under Python Software Foundation). Currently, he is contributing to the following open source projects:
1. opendatagroup/hadrian - Implementations of the Portable Format for Analytics (PFA)
2. Fuzzy-Logix/AdapteR - Advanced analytics package that enables R users to perform in-database analytics

An IIT Kanpur alumnus, Ankit is also an active researcher with publications in international journal and conferences. He is actively working in the domain of IoT Analytics and recently presented his work - "In-database Analytics in the Age of Smart Meters" in the 5th IIMA International Conference on Advanced Data Analysis, Business Analytics and Intelligence, 2017.

Previous Workshop Experience:
- Making Machine Learning Fruitful and Fun using Orange in PyCon India 2017, New Delhi.
- High Performance Computing, IIT Kanpur, 2013.

LinkedIn - https://www.linkedin.com/in/ankitmahato

Abstract

IPython/Jupyter notebook is widely used for scientific computing among the research community. This notebook style programming belongs to an imperative paradigm which is linear in nature. In the past decade, Visual programming paradigm has gained a lot of popularity as it is user-centric in nature and driven by data streams.

In this workshop, we will visually uncover the various aspects of a Scientific Computing Pipeline using Orange 3, a Python based open source interactive data analysis, machine learning and data visualization workbench. Its simple "drag-and-drop" based workflow design interface makes it ideal for novices, and its modular design, extensibility and python integration makes it powerful for advanced users.

The workshop will begin with the building of basic scientific computing pipeline using built-in Orange widgets, which will further evolve into complex pipeline covering advanced topics like - GPU computing (pyCUDA), In-database analytics, Using external ML toolkit, Custom widget development, etc. For these advanced topics the audience will be made familiar with the GUI and computational concepts involved in the development of add-on (custom-built) widgets for

Orange.

Hands-on experience of the various aspects will be provided in this workshop:

Data Access (files & external data sources)
Data Exploration
Data Transformation/Filtering
Data Analysis
Basic and advanced Visualization (in-built, matplotlib)
Report Generation

Real life use cases (Internet of Things, Monte Carlo Simulation (Finance), Sentiment Analysis) will be selected for the workshop.

Workshop Breakup -
Introduction Dataflow Programming & Orange - 45 mins
Exercise 1 (Data Access)- 15 mins
Exercise 2 (IoT)- 30 mins
Exercise 3 (Monte Carlo/PyCUDA)- 30 mins
Exercise 4 (Supervised Learning) - 30 mins
Exercise 5 (Text Mining) - 30 mins

These timings are tentative and based on the interest of audience, the workshop can go deeper into the some topics like advanced visualization, GPU computing, etc.

Knowledge prerequisites:

This workshop has no prerequisites, but it would be great to know Basic Python Programming (development of simple functions and classes) for widget development section.

Software prerequisites:

Install Python 3.5 or 3.6 (Python 2.7 is not recommended as the latest development and release of Orange is in Python 3)
Install the following packages:

```
pip install Orange3 matplotlib Orange3-Text twython PyQt5
```

Make sure Orange Canvas is up and running:

```
python -m Orange.canvas
```

Optional Setup:
pycuda - python library for gpu computing.
This will require installation of CUDA toolkit and Microsoft Visual C++ 2015 Build Tools (for Windows)