

The course assumes a working knowledge of key data science topics (statistics, machine learning, and general data analytic methods). Programming experience in some language (such as R, MATLAB, SAS, Mathematica, Java, C, C++, VB, or FORTRAN) is expected. In particular, participants need to be comfortable with general programming concepts.

**Working knowledge of the Python tools ideally suited for data science tasks, including:**

- Accessing data (e.g., text files, databases)
- Cleansing and normalizing data
- Exploring data (e.g., simple statistics, correlation matrices, visualization)
- Modeling data (e.g., machine learning)

## **MODULES**

### **I. Introduction and Setting Up Your Integrated Analysis Environment**

### **II. Using Python to Control and Document Your Data Science Processes**

- Python Essentials
- Data types and objects
- Loading packages, namespaces
- Reading and writing data
- Simple plotting
- Control flow
- Debugging
- Code profiling

### **III. Accessing and Preparing Data**

- Acquiring Data with Python
- Loading from CSV files
- Accessing SQL databases

### **Cleansing Data with Python**

- Stripping out extraneous information
- Normalizing data
- Formatting data

### **IV. Numerical Analysis, Data Exploration, and Data Visualization with NumPy Arrays, Matplotlib, and Seaborn**

- NumPy Essentials
- The NumPy array
- N-dimensional array operations and manipulations
- Memory mapped files

### Data Visualization

- 2D plotting with Matplotlib
- Advanced data visualization with Seaborn

### V. Exploring Data with Pandas

- Searching for Gold in a Pile of Pyrite
- Data manipulation with Pandas
- Statistical analysis with Pandas
- Time series analysis with Pandas

### VI. Machine Learning

- Predicting the Future Can Be Good for Business
- Input: 2D, samples, and features
- Estimator, predictor, transformer interfaces
- Pre-processing data
- Regression
- Classification
- Model selection