

**Data Science Fundamentals with Python for Healthcare**

**Course Number:** PYTH-218
**Duration:** 5 days

**Overview**

This Data Science Fundamentals with Python for Healthcare training course teaches quantitative professionals (engineers, statisticians, analysts, and others) how to successfully apply data science methods to analyze and visualize real-world problems in healthcare data.

**Prerequisites**

All attendees should have prior programming experience and an understanding of basic statistics.

**Materials**

All attendees receive comprehensive courseware.

**Software Needed on Each Student PC**

* Anaconda Python 3.6 or later
* Spyder IDE and Jupyter notebook (comes with Anaconda)

**Objectives**

* Understand and implement key Python concepts (data types, functions)
* Use libraries to import dynamic EHR (Electronic Health Record) data and static data
* Parse unstructured clinical text data into structured data
* Apply functions in Pandas and NumPy to quickly clean and explore data
* Understand techniques to assess missingness in patient data
* Extend cleaning techniques to reshaping data for use in advanced analytics
* Explore and clean clinical text data
* Apply regular expressions to manipulate and extract data from text
* Understand rules-based Natural Language Processing (NLP) approaches for information extraction, such as diagnoses or medications
* Identify tests for group differences using inferential statistics
* Implement linear regression to model and forecast clinically relevant data
* Using non-linear terms, as well as understanding confounding and interaction terms for more advanced system modeling
* Apply logistic regressions to model non-numeric outcomes, such as patient follow-up

**Outline**

* Overview of Data Science in Healthcare
	+ Limitations of EHR data
	+ Importance of NLP methods
	+ Overview of advanced data science work in healthcare (image recognition and temporospatial modeling)
* An Accelerated Introduction and Overview to Python for Data Science
	+ Review of course and computing environment
	+ Explanation of Integrated Development Environments (IDEs) Jupyter and Spyder
	+ Python syntax essentials
		- Primitive data types
		- Collection variable types
		- Control flow operations
		- Function syntax
		- Error handling
		- Managing libraries
* Reading and Manipulating Datasets with Libraries (NumPy and Pandas)
	+ Overview of NumPy
		- Data types in NumPy
		- Array masks
		- Manipulation and broadcasting
		- Random number generation
	+ Data processing methods with Pandas
		- Using DataFrames and Series
		- Creating calculated columns
		- Discretizing data
		- Filtering and indexing syntax
		- Merging datasets
		- Melting/pivoting DataFrames
* Exploratory Data Analysis (EDA) and Graphics Fundamentals
	+ Statistical summaries, and outlier detection for both univariate and multivariate variables using graphical and numeric methods
	+ Visualization crash course with Seaborn and Matplotlib
	+ Generating publication-quality documents with Jupyter
* Applied NLP Techniques for Clinical Text
	+ Unstructured data fundamentals
	+ Implementing regular expressions for basic information extraction
	+ Applying MedSpaCy for advanced processing of clinical text
	+ Measuring accuracy and limitations in rules-based methods
	+ Using Term Frequency Inverse Document Frequency (TF-IDF) techniques for term importance
* Applying Statistical Models for Analysis in Python
	+ Explanation of statsmodels library of functions
	+ Inferential and descriptive statistics refresher
	+ Implementing A/B tests for detecting group differences
	+ Applying linear regressions
	+ Overview of generalized linear models (GLMs) and the link function
	+ Applying logistic regression
	+ Discussion of confounding, interaction terms and model building approaches
* Conclusion