

# Data Science Bootcamp Curriculum

NYC Data Science Academy



NYC DATA SCIENCE  
**ACADEMY**

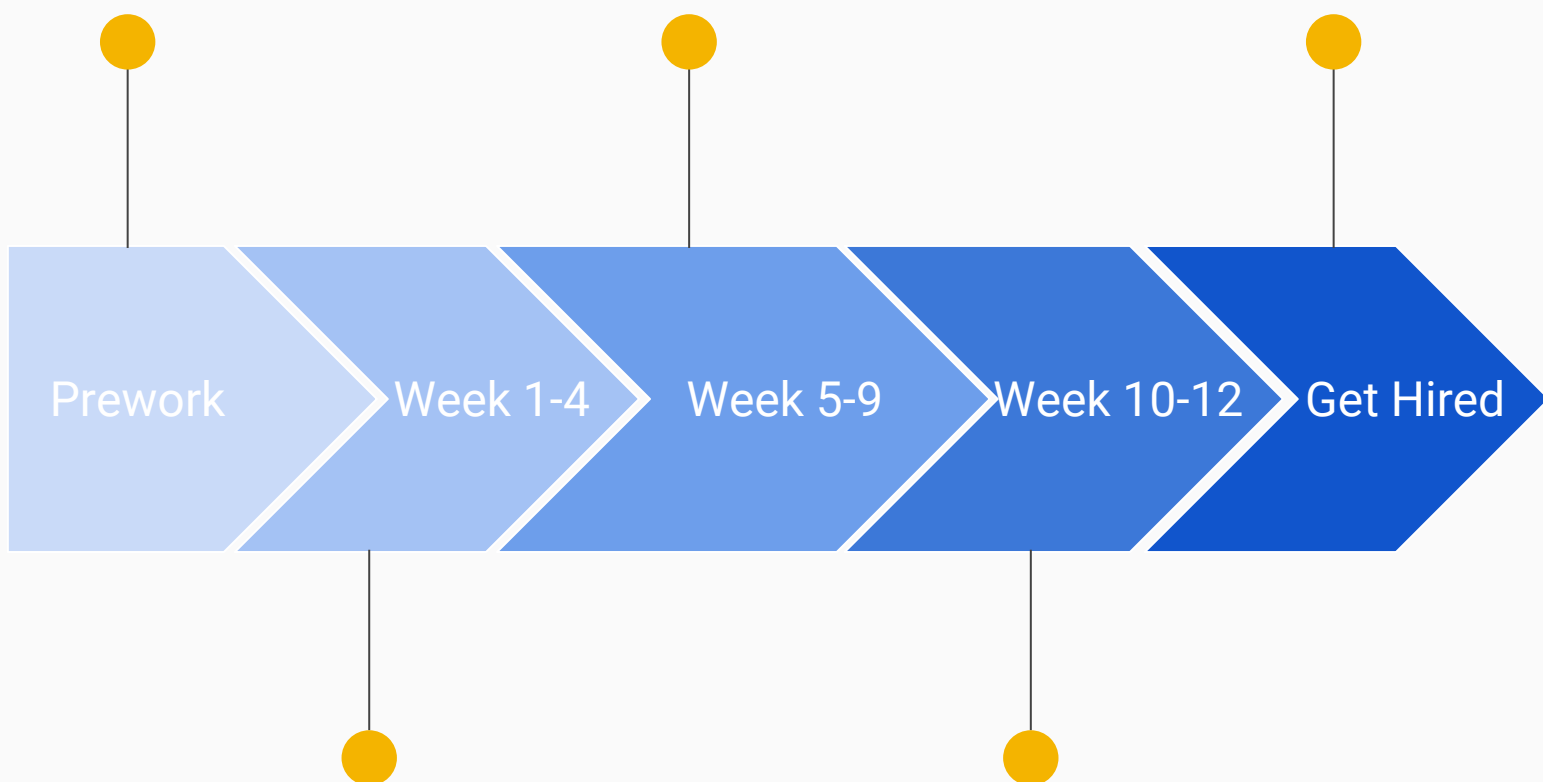
100+ hours free,  
self-paced online course.  
Access to part-time  
in-person courses hosted  
at NYC campus

## Machine Learning with R and Python

Foundations of statistics,  
regressions, classifications,  
model selections,  
unsupervised learning, etc.

Machine learning theory  
defense, Capstone  
project presentations.

Code reviews, resume  
workshop, mock  
interviews, career day



## Data Analysis and Visualization

Linux system, Git, SQL  
Data analysis and visualization with R  
and Python  
R Shiny  
Web scraping with Python

## Big Data with Hadoop & Spark

Spark, Spark SQL, Spark MLlib, AWS,  
Hadoop and MapReduce, Hive

## Deep Learning

Neural Network, TensorFlow, Machine  
Vision, NLP, Time Series Analysis,  
Reinforcement Learning

# Pre-work

Once students are enrolled in the bootcamp, they are granted access to our online, self-paced pre-work materials:

- 20-30 hours: Introductory Python (Optional)
- 35-45 hours: Data Analysis and Visualization with R
- 20-30 hours: Data Analysis and Visualization with Python

Students are also invited to join their cohort's Slack channel, where they meet their future classmates, instructors, and get support on pre-work assignments.

Enrolled bootcamp students can also choose to take part-time, beginner-level courses hosted at our NYC campus. 100% tuition credited to bootcamp.

**Week 1****Data Science Toolkit – Linux, Git, Bash, and SQL****Data Science with R – Data Analytics – Part I**

- Linux system
  - Operating Systems and Linux
  - File System and File Operations
  - Text-processing commands
  - Other useful commands
- Git
  - What is Version Control and Git?
  - Installing Git
  - Getting Started with Git
  - Git Tips
  - Undoing Changes
  - What is Github?
  - Working With Remotes
- SQL
  - Intro to SQL
  - Tables and schemas
  - SQL queries – SELECT
  - MySQL database management
  - Joins
- Programming foundation in R I
  - Introduction to R
  - Introduction to RStudio
  - R objects
  - Functional programming: apply
- Programming foundation in R II
  - More data types
  - Control statements
  - Functions
  - Data Transformations

**Week 2****Data Science with R – Data Analytics – Part II**

- Data manipulation with “dplyr”
  - Introduction to dplyr
  - Built-in functions

- Join data sets
- Groupwise operations
- Data Visualization with "ggplot2"
  - Why ggplot2?
  - The "Grammar of Graphics"
  - Constructing a ggplot2 plot
  - Scatterplots
  - Bar charts
  - Histograms
  - Visualizing big data
  - Saving Graphs
  - Customizing Graphics
- Lab: Data Visualization from Scratch
- Introduction to Shiny
  - Shiny introduction
  - Design the User-interface
  - Control widgets
  - Build reactive output
  - Use data table in Shiny Apps
  - Use R scripts, data and packages
  - UI and server for the App
  - Make Shiny perform quickly
  - Matrix-based visualizations
  - Use reactive expressions
  - Share and deploy Shiny apps
- Lab: Build a Shiny app from Scratch

### **Week 3**

#### **Data Science with R – Machine Learning – Part I**

#### **Data Science with Python - Data Analytics – Part I**

- Foundations of Statistics
  - All About Your Data
  - Statistical Inference
  - Introduction to Machine Learning
  - Review
- Get Started with Python
  - Installing and using iPython
  - Simple values and expressions

- Lambda functions and named functions
- Lists
- Functional operators: map and filter
- Strings and Data Structures
  - String operations
  - File Input and Output
  - Searching in files
  - Data Structures
- Conditionals and Control Flows
  - Conditionals
  - For loops
  - List Comprehensions
  - While loops
  - Errors and Exceptions
- Project Day: Exploratory Visualization & Shiny

### **Project 1 Due: Exploratory Visualization & Shiny**

#### **Week 4**

#### **Data Science with Python – Data Analytics – Part II**

- Advanced Topics
  - Multiple-list operations: map and zip
  - Functional operators: reduce
  - Object Oriented Programming
- Introduction to Web Scraping
  - Regular Expressions
  - Introduction to HTML
  - Basics of BeautifulSoup
  - Examples
- Introduction to Scrapy
  - An example
  - Getting Started
  - Items/spider/pipelines/settings.py
  - In Class Lab
- Introduction to Numpy and Scipy
  - Ndarray
  - Subscripting and slicing
  - Operations
  - Matrix and linear algebra

- Random Sampling

### **Shiny Project Presentations**

#### **Week 5**

##### **Data Science with Python - Data Analytics – Part III**

##### **Data Science with R - Machine Learning – Part I**

- Introduction to Pandas
  - Data Structure
  - Data Manipulation
  - Handling missing data
  - Grouping and aggregation
- Matplotlib & Seaborn
  - In-class Lab
- Missingness & Imputation
  - Missing Data
  - Basic Methods of Imputation
  - K-Nearest Neighbors
  - Review
- Linear Regression I
  - Simple Linear Regression
  - Assumptions & Diagnostics
  - Transformations
  - The Coefficient of Determination  $R^2$
- Project Day: Web Scraping

### **Project 2 Due: Web Scraping**

#### **Week 6**

##### **Data Science with R - Machine Learning – Part II**

- Linear Regression II
  - Multiple Linear Regression
  - Assumptions & Diagnostics
  - Research Questions of Interest
  - Extending Model Flexibility
  - Review
- Generalized Linear Models
  - Logistic Regression
  - Maximum Likelihood Estimation



- Model Interpretation
- Assessing Model Fit
- Review
- The Curse of Dimensionality
  - Ridge Regression
  - Lasso Regression
  - Cross-Validation
  - Bias/Variance Tradeoff
- Tree Methods I
  - Decision Trees
  - Bagging
  - Random Forest
  - Boosting
  - Variable Importance

### **Web Scraping Project Presentations**

#### ***Week 7***

#### **Data Science with R - Machine Learning – Part III**

#### **Data Science with Python - Machine Learning – Part I**

- Tree Methods II
  - Decision Trees
  - Bagging
  - Random Forest
  - Boosting
  - Variable Importance
- Support Vector Machines
  - Maximal Margin Classifier
  - Support Vector Classifier
  - Support Vector Machines
  - Multi-Class SVMs
  - Review
- Association Rules & Naïve Bayes
  - Association Rule Mining
  - Naïve Bayes
  - Review
- Python - Linear Regression
  - What is Machine Learning



- Introduction to Scikit-Learn
- Simple Linear Regression
- Multiple Linear Regression
- Statsmodels

### **Week 8**

#### **Data Science with Python - Machine Learning – Part II**

#### **Data Science with R - Machine Learning – Part IV**

- Python - Classification Part I
  - Limitation of Linear Regression
  - Logistic Regression
  - Discriminant Analysis: Motivation
  - Discriminant Analysis: Models
  - Naïve Bayes
- Python - Model Selection
  - Cross-Validation
  - Bootstrap
  - Feature Selection
  - Regularization
  - Grid Search
- Python - Classification Part II
  - Support Vector Machines
  - Tree-Based Methods
- Principal Component Analysis
  - Taking a New Perspective
  - Dimension Reduction
  - Vectors of Highest Variance
  - The PCA Procedure
- Project Day: Machine Learning

#### **Project 3 Due: Machine Learning**

### **Week 9**

#### **Data Science with R and Python - Machine Learning (Continued)**

- Cluster Analysis
  - Intro to Cluster Analysis
  - K-Means Clustering



- Hierarchical Clustering
- Clustering Takeaways
- Review
- Python - Unsupervised Learning
  - Intro to Unsupervised Learning
  - Principal Component Analysis
  - Clustering
- Python – Advanced Regression
- Python – Advanced Classification

### **Machine Learning Kaggle Project Presentations**

#### ***Week 10***

#### **Advanced Topics: Parallel Computing, Hadoop, and Spark**

#### **Advanced Topics: Deep Learning**

- Hadoop and MapReduce:
  - What is Hadoop
  - HDFS
  - MapReduce
  - Combiner
  - Hadoop Monitoring Ports
- Apache Hive:
  - Databases for Hadoop
  - Hive
  - Compiling HiveQL to MapReduce
  - Technical aspects of Hive
  - Extending Hive with TRANSFORM
- Introduction to Spark
  - What is Apache Spark
  - Initializing Spark
  - RDDs, Transformations and Actions
  - Working with Key-Value Pairs
  - Performance & Optimization
- Introduction to Spark SQL
  - Overview
  - Spark Session
  - Working with DataFrames
  - Using HiveQL in Spark SQL



- Spark Mllib
  - Spark Machine Learning Workflow
  - How ML Pipeline Works
  - ML Pipeline Example: Predicting Diamonds Price
  - Extracting, transforming and select features
  - Train Validation Splitting
  - Building the ML Pipeline with DecisionTreeRegressor
  - Model Evaluation
  - Model Tuning
- How Deep Learning Works
  - Neural Units
  - Neurons in TensorFlow
  - Cost Functions, Gradient Descent, and Backpropagation
  - Fitting Models in TensorFlow
  - Interactive Visualization of a Deep Neural Network
  - TensorBoard and Interpretation
- TensorFlow Lab
  - Random Initialization and Stochastic Gradient Descent
  - Introduction to Convolutional Neural Networks for Visual Recognition
  - Dropout and Regularization
  - Tuning Hyperparameters
- Machine Vision
  - Classic ConvNet Architecture I: LeNet-5
  - Classic ConvNet Architecture II: AlexNet
  - Classic ConvNet Architecture II: VGGNet
  - Transfer Learning
  - Dogs vs Cats Kaggle Competition
- Natural Language Processing
  - Word Vectors: word2vec and Vector-Space Embedding
  - Build a recommendation system with doc2vec
  - Sentiment Analysis using Convolutional Neural Network
- Time Series Analysis
  - The Nature of Time Series Analysis
  - Learn from the Examples
  - Decomposition of Time Series Data
  - Examples of Stationary Non-White-Noise Time Series
  - ARMA and ARIMA Models
  - Assessing Model Fit

**Week 11****Advanced Topics: Parallel Computing, Hadoop, and Spark****Advanced Topics: Deep Learning****SQL, R, & Python Code Review****Machine Learning Theory Defense**

- Big Data on AWS
  - Creating a Hadoop Cluster using EMR
  - Submitting MapReduce / Hive Jobs via Web Console
  - Working with AWS CLI
  - Accessing to EMR Master Node using SSH
  - Running Self-Contained Spark Applications
- Database Management Tools
  - AWS cloud services (IAM, S3, EC2, RDS.)
  - MySQL / AWS RDS
  - GUI Tool: MySQLWorkBench
  - MySQL Python Connector
- NoSQL Databases and MongoDB
  - Intro to NoSQL
  - Installing MongoDB on AWS EC2
  - Common database commands
  - GUI tool: MongoDB Compass
  - pyMongo
- Time Series Analysis with Deep Learning
  - Recurrent Neural Networks
  - Long Short-Term Memory Units
  - Forecasting with Financial Time Series Data
  - Web Traffic Time Series Forecasting Kaggle 1st Place Solution
- Reinforcement Learning
  - Applications of Reinforcement Learning
  - Essential Theory of Reinforcement Learning
  - OpenAI Gym
  - Two Sigma Halite Competition

**Week 12****SQL, R, & Python Code Review****Machine Learning Theory Defense**

**Capstone Project Presentations**

- A/B Testing
- Capstone Project Presentations
- Machine Learning Theory Defense
- SQL Code Challenge

From the beginning of Bootcamp, you will work on hands-on projects. Now your Capstone Project lets you create your own data product that showcases your interests and talents. Students are free to use anything covered in class on this project.