Week 1

Data Science Toolkit – Linux, Git, Bash, and SQL

Data Science with R – Data Analytics – Part I

- Linux system
  - Introduce Linux environment
  - Learn Linux commands
  - IO redirection and Pipe
  - Introduce server-side Linux usage
- Git
  - Introduce modern source code management
  - Learn common git operations
  - Setup github and personal portfolio page
- Other server related topics
  - Text editors and IDEs
  - ssh: how to communicate with a remote server
  - Linux environment variables
- SQL
  - Introduction to relational database
  - Introduction to structured query language
  - SQL major commands and examples
- Programming foundation in R I
  - Syntax
  - Data object: Vectors, Matrices, Data Frames, and Lists
  - Common functions
  - Rstudio environment and package management
  - Local data input/output
  - Introduction to R data visualization
- Programming foundation in R II
  - Data sorting and merging
  - String manipulation
  - Dates and times
  - Connecting to an external database

Week 2

Data Science with R – Data Analytics – Part II

- Data manipulation with “dplyr”
  - Tables in R
  - Join
  - Subset
  - Advanced manipulations with dplyr
- Data Visualization with "ggplot2"

- Histogram
- Point graphics
- Columnar graphics
- Line charts
- Pie charts
- Box plots
- Scatter plots
- Visualizing multivariate data
- Matrix-based visualizations
- Maps

- 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: Moneyball

Project 1 Due: Exploratory Data Visualization

Week 3

Data Science with Python - Data Analytics – Part I

- Python Programming Language I
  - Simple Values and Expressions
  - Functions
  - Lists
  - Conditionals
  - Functional programming: map, filter and reduce
- Python Programming Language II
  - String operations
  - File input/output and searching
  - Data Structures:
    - Mutating operations on Lists
    - Tuples, sets and dictionaries
- Python Programming Language III
  - Control flows

  - Errors and exceptions
  - Object-oriented programming
- Web scraping
  - Regular expression
  - HTML, beautiful soup and scrapy
  - NoSQL and MongoDB

Week 4

Data Science with Python – Data Analytics – Part II

- Numpy and Scipy
  - Basic data structure and operations
  - Matrices and linear algebra
  - Stats module
  - Random Sampling
- Pandas
  - Series and data frame
  - I/O of pandas data frame
  - Concatenation and merge
  - Arithmetic, drop, apply and describe
  - Selection and filter
  - Missing values
  - Grouping and aggregation
  - Time series
  - Interacting with data base
- Matplotlib and Seaborn
  - Basic plots
  - Statistical plots:
    - Scatter plots
    - Histogram
    - Boxplot
    - Barchart
  - Multiple figures
  - Advanced plots with seaborn
- Python lab: linear regression from scratch

Project 2 Due: R Shiny Interactive Applications

Week 5

Data Science with R - Machine Learning – Part I

- Foundations of Statistics
  - Descriptive Statistics

    - Measures of Centrality
    - Measures of Variability
    - Frequency, Proportion & Contingency Tables
    - Correlation
  - Hypothesis Testing
    - One Sample t-test
    - Two Sample t-test
    - F-test
    - One-way ANOVA
    - X2 Test of Independence
  - Introduction to Machine Learning
    - Supervised Learning
      - Regression
      - Classification
    - Unsupervised Learning
      - Clustering
      - Dimension Reduction
- Missingness & Imputation
  - Types of Missingness
    - MCAR
    - MAR
    - MNAR
  - Basic Methods of Imputation
    - Mean Value Imputation
    - Simple Random Imputation
    - Regression Prediction
  - K-Nearest Neighbors
    - Voronoi Tessellations
    - KNN for Classification
    - KNN for Regression
    - Distance Measures
- Linear Regression I
  - Simple Linear Regression
    - From a Mathematical Standpoint
    - Accuracy of the Coefficient Estimates
    - Performing Hypothesis Tests
    - Constructing Confidence Intervals
  - Assumptions & Diagnostics
  - Transformations
    - Power Transformation
    - Box-Cox Transformation

  - The Coefficient of Determination R^2
- Linear Regression II
  - Multiple Linear Regression
    - From a Mathematical Standpoint
  - Assumptions & Diagnostics
  - Potential Problems
  - Research Questions
  - Variable Selection
  - Factors
  - Interactions
  - Higher-Order Terms

Week 6

Data Science with R - Machine Learning – Part II

- Lab: Building Bridges
- Generalized Linear Models
  - Logistic Regression
- The Curse of Dimensionality
  - Ridge Regression
  - Lasso Regression
  - Cross-Validation
  - Bias/Variance Tradeoff
  - Density
  - Principal Component Analysis
- The Curse of Dimensionality
  - Density
  - Principal Components Analysis
- Guest Lecture: Dataiku Part I

Project 3 Due: Python Web Scraping

Week 7

Data Science with R - Machine Learning – Part III

- Classification
  - Feature Selection
  - Support Vector Machines
  - Decision Trees
  - Pruning/Purity/Entropy/GINI
  - Random Forests
  - Bagging
  - Boosting
- Cluster Analysis

- K-Means Clustering
- Agglomerative Clustering
- Hierarchical Clustering

- Neural Networks

Week 8

Data Science with R - Machine Learning – Part IV

Introduction to Natural Language Processing

- Case Study: Spam Detection
- Association Rules
  - Market Basket Analysis
- Naïve Bayes Analysis
- Introduction to Natural Language Processing
  - Creating corpus: stemming and lemmatization
  - POS tag and chunking
  - Text classification
- Time Series Analysis
  - Smoothing
  - Seasonal Decomposition
  - ARIMA
- Guest Lecture: Dataiku Part II

Week 9

Data Science with Python - Machine Learning

- Machine Learning Recap / Linear Regression
  - Introduction to scikit learn
  - Simple linear regression
  - Multiple linear regression
  - Stats module
- Classification part I
  - Logistic regression
  - Discriminant analysis
  - Naïve Bayes
- Model Selection
  - Cross-validation
  - Bootstrap
  - Feature selection
  - Regularization
  - Grid search
- Classification part II

    - Support vector machine
    - Decision tree
    - Random forest
- Unsupervised learning
    - Principal Components Analysis
    - Kmeans and Hierarchical Clustering

Project 4 Due: Machine Learning Project (It can be a Kaggle competition, a hiring partner project or a non-profit project from our partners)

Week 10

Big Data

- Parallel processing: Introduction to Hadoop and MapReduce
    - HDFS
    - MapReduce
        - Conceptual framework
        - Streaming and Python
    - Examples and lab work
- MapReduce design pattern
    - Filtering patterns
        - Simple filtering
        - Top N
    - Summarization patterns
        - Numerical summarizations
        - Inverted Index summarizations
- Apache Hive:
    - Databases for Hadoop
    - Hive
        - Select
        - Joins
    - Compiling HiveQL to MapReduce
    - Technical aspects of Hive
    - Extending Hive with TRANSFORM
- Spark
    - Basics concepts
        - RDDs, transformations and actions
        - PairRDDs
    - Examples
        - Wordcount
        - Mean and variance

Week 11

Big Data and Algorithms

- Spark MLlib
- Amazon Web Service
- Introduction to Algorithms
  - Analysis of algorithms: big-O notation
- Sorting
  - Elementary sorts
  - Merge sorts
  - Quick sorts
- Searching
  - Linear search
  - Binary search
  - Hash tables
- Machine Learning Theory Defense Practice

Week 12

Capstone Project Presentations and Review

- Machine learning theory defense practice
- SQL code review
- R code review
- Python code review
- 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.

Project 5 Due: Capstone Project