Doing Data Science

David Stroud

September, 2019

E-mail: david@davidstroud.me
Slack Page: TBA
Office: ...

GitHub Page: **www4.tba/ username**
Class Hours: TBA
Class Room: ...

Lab Room: ...

Lab Hours: TBA

Course Description

A Data Scientist combines statistical and machine learning techniques with both the Python and R programming languages to analyze and interpret complex data. We will cover Python and R functions and data types, then tackle how to operate on vectors and when to use advanced functions like sorting. You will learn how to apply general programming features like if-else, and for loop commands, and how to wrangle, analyze and visualize data.

We cover concepts like probability, inference, regression, and machine learning. We help you develop a skill set that includes Python and R programming languages, data wrangling, data visualization, file organization with UNIX/Linux, version control with git and GitHub, and reproducible document preparation with Jupyter notebook and RStudio.

Required Materials

- All course material will me made available on our GitHub page and Slack channel.

Prerequisites/Corequisites

Students should have a basic understanding of Statistics and some exposure to programming languages.

Course Objectives

Successful students:

1. Be able to form a testable hypothesis from an unstructured problem.
2. Apply the principles of reproducible research in testing said hypothesis.
3. Use tools such as Python, Jupyter, R, R-Studio and Github to organize and document research so that others can reproduce and/or continue your work.
4. Apply basic statistics and graphics to explore data.
5. Use the principles of data cleaning to create clean data sets from messy ones using Python and R.
6. Communicate the findings of a project in a clear, concise, and scientific manner.

Course Structure

Class Structure

We will use a combination of live lecture, GitHub, and Slack as our methods of content delivery. Heavy use of our Slack channel is encouraged. The first four weeks of the course will use the Python programming language and the last two weeks we will be using the R programming language.

Assessments

Four labs and a Capstone project in Python. Two labs and a Capstone project in R.

Labs

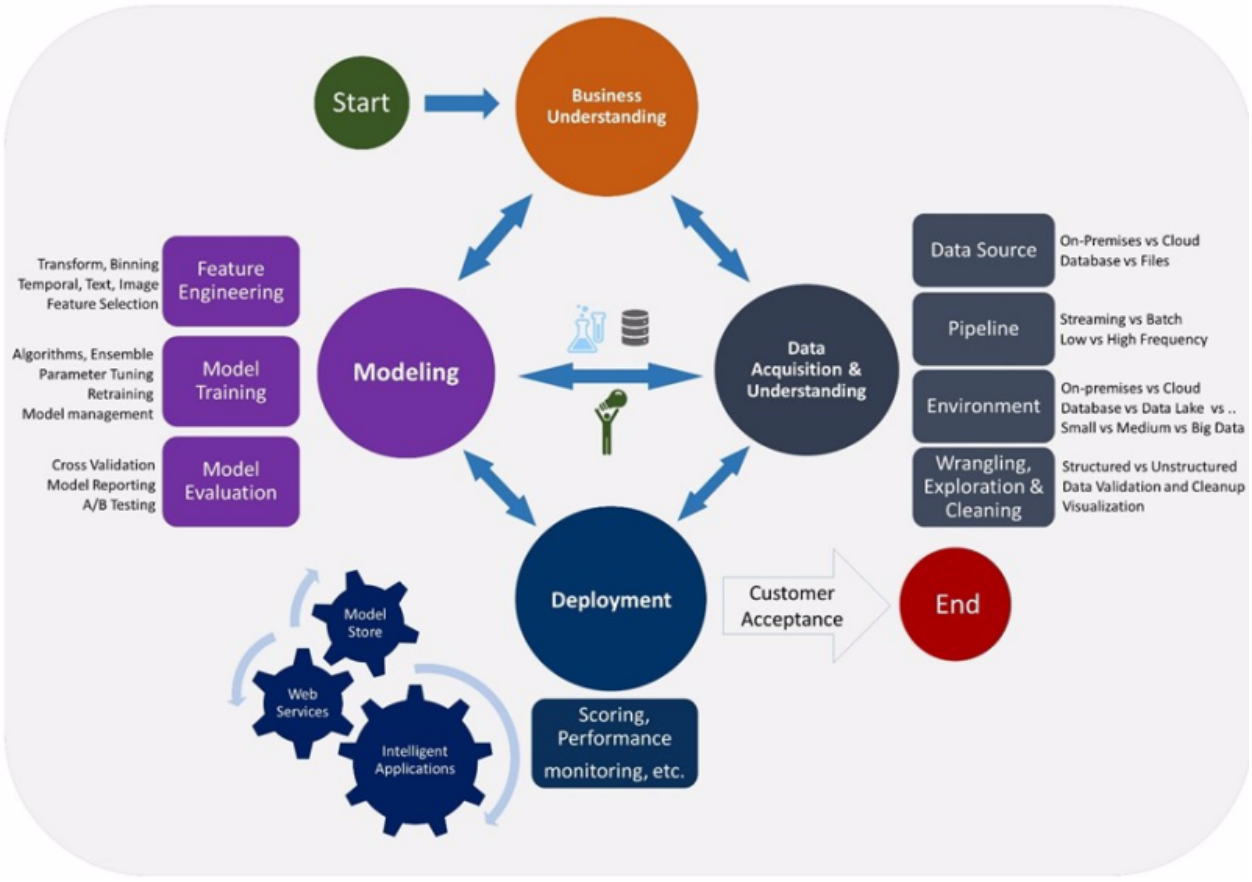
The labs will consist of mini-projects designed to demonstrate practical application of the concepts taught during lecture.

Capstone Project Python

In this project, you will explore the market capitalization of Bitcoin and other cryptocurrencies. This project will test your skills in data manipulation, data visualization, data importation and data cleaning.

Capstone Project R

In this project, you get to work with the data from a large number of taxi journeys in New York from 2013. You will use regression trees and random forests to predict the value of fares and tips, based on location, date and time.


Start
Business Understanding
Modeling
Data Acquisition & Understanding
Deployment
Customer Acceptance
End
Transform, Binning
Temporal, Text, Image
Feature Selection
Feature Engineering
Algorithms, Ensemble
Parameter Tuning
Retraining
Model management
Model Training
Cross Validation
Model Reporting
A/B Testing
Model Evaluation
Data Source
On-Premises vs Cloud
Database vs Files
Pipeline
Streaming vs Batch
Low vs High Frequency
Environment
On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data
Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleanup
Visualization
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

Schedule and weekly learning goals

The schedule is tentative and subject to change. The learning goals below should be viewed as the key concepts you should grasp after each week, and also as a study guide for reinforcement.

Week 01, 09/02 - 09/06: Introduction to Data Science in Python

- What is Data Science?
- Statistical Inference and the Data Science Process.
- Introduction to Python
- Importing Data in Python
- Correlation and Regression

Week 02, 09/09 - 09/13: Data Cleaning and Visualization

- Introduction to Numpy and Pandas
- Exploratory Data Analysis in Python
- Visualization Best Practices in Python
- Data Visualization with Matplotlib

Week 03, 09/16 - 09/20: Machine Learning

- From Linear Regression to Classification
- Logistic Regression and Support Vector Machines
- Decision Trees and Random Forests
- Model Evaluation and Ensemble Classification

Week 04, 09/23 - 09/27: Case Study: Explore the market capitalization of Bitcoin and other cryptocurrencies.

- Build an end to end Machine Learning Project in Python
- Understanding the foundations of Pandas
- Manipulating Data Frames with Pandas
- Cleaning data in Python

Week 05, 09/30 - 10/04: Introduction to Data Science in R

- Introduction to the Tidyverse
- Data manipulation in R with dplyr
- Data visualization with ggplot2
- Supervised Learning in R:Regression

Week 06, 10/07 - 10/11: Case Study: Predict Taxi Fares with Random Forests

- Build an end to end Machine Learning Project in R
- Use Regression Trees and Random Forests in R
- Supervised Learning in R
- Cleaning data in R