Data Science Bootcamp
Preliminary Syllabus (subject to change)

Prerequisites:

- At least some math background (i.e., calculus in high school, linear algebra in college)
  - Linear algebra or proof-based math would be preferable, although I can adjust the amount of time we spend on math preliminaries based on the backgrounds of the students.
- Some experience with programming would be preferable, although it's not required.

Week 1: Computer Preliminaries

- Intro to programming in Python
- Intro to dataset manipulation using pandas
- Plotting
- Basic Unix commands
- git
- Web scraping
- SQL
- API's

HW: Programming project.

Weeks 2-3: Math, Statistics, and Computer Science Preliminaries

- Mathematical notation
- Calculus review
- Linear algebra primer
- Optimization primer
- Time complexity
- Intro to statistics
  - Probability
  - Bayes' Rule
  - Maximum Likelihood Estimation
  - Normal distribution and other common distributions
  - Hypothesis Testing
  - Confidence Intervals / Confidence sets
  - Linear Regression (OLS)
  - Bayesian vs. frequentist statistics
  - Correlation vs. Causation
  - Bias / Variance tradeoff
  - cross-validation and overfitting
  - Model assumptions
  - Bootstrapping

HW: Math and Statistics exercises (both programming and on paper)

Week 4-6: Intro To Supervised Machine Learning Algorithms

- Intro to scikit-learn
- Measurement of regression and classification error
  - When to choose each measure
- Classification vs. regression
- regularized regression (lasso, ridge, elastic net)
- K-nearest neighbors

- Naive Bayes
- Decision Trees
- Random Forest
- Support Vector Machines
- logistic regression
- Poisson regression
- Gradient descent
- Stochastic gradient descent
- Neural networks

HW: Download a dataset via an API or web scraping and find the come up with the best algorithm to predict the target variable of interest. Write up a report explaining your choices, including nice plots.

Week 5: Unsupervised Learning

- k-means clustering
- Dimensionality-reduction using Principal Component Analysis
- Independent Component Analysis

HW: TBD

Week 6: Natural Language Processing

- Featurization of corpora
  - Bag-of-words / n-grams
  - Cleaning the data (changing to lowercase, removing stopwords, etc.)
  - Stemming
- Topic Modeling
- Part-of-speech tagging
- NLP for non-English languages

HW: Sentiment analysis project.

Week 7: Time Series Analysis

- How to handle time series in pandas and R
- ARIMA modeling and forecasting
- VARIMA modeling and forecasting
- Volatility modeling and forecasting (ARCH, GARCH, etc.)
- Peculiarities of analyzing financial data

HW: Financial time series forecasting project. Work on capstone project.

Week 8: Big Data Techniques

- Hadoop
- MapReduce
- Hive
- Spark

HW: Work on capstone project. Additional homework may be assigned if there is time.

Week 9: Special Topics

- Density estimation
- More efficient ways of optimizing hyperparameters

HW: Work on capstone project. Additional homework may be assigned if there is time.

Weeks 10-11: Additional Special Topics (Tentative)

- Density estimation
- Expectations-Maximization
- Advanced topic modeling
- R programming and useful R packages
- Computer vision
- Causal Analysis

HW: Work on capstone project. Additional homework may be assigned if there is time.

Week 12: Capstone Project Presentations and Critiques

Students present the results of their capstone projects, which is a prediction project of their choice. The rest of the class then critiques them, and they will have to defend their modeling decisions. At the end of the week, the class votes on who presented the the best and most interesting model. The first, second, and third place winners will get a prize. The class will also vote on who asked the best questions, and the winner will get a prize.