DATA SCIENCE IMMERSIVE SYLLABUS

COURSE OVERVIEW

By the end of this course, students will be able to:

- Collect, extract, query, clean, and aggregate data for analysis
- Perform visual and statistical analysis on data using Python and its associated libraries and tools
- Build, implement, and evaluate data science problems using appropriate machine learning models and algorithms
- Use appropriate data visualization tools to communicate findings
- Create clear and reproducible reports to stakeholders
- Identify big data problems and understand how distributed systems and parallel computing technologies are solving these challenges
- Apply question, modeling, and validation problem-solving processes to datasets from various industries in order to provide insight into real-world problems and solutions

WEEK 1: MATH, & PROGRAMMING FUNDAMENTALS

Project 1	*Provided with a dataset of SAT scores from across the United States, students will perform exploratory analysis using NumPy, Matplotlib, and Tableau to apply basic summary statistics.*
Python & NumPy	*Demonstrate introductory programming concepts using Python and NumPy as a tool to navigate data sources and collections*
UNIX	*Utilize UNIX commands to navigate file systems and modify files*
git	*Maintain a git repository in order to keep track of changes and iterations as your project evolves.*
Descriptive Statistics	*Define and apply descriptive statistic fundamentals*
Intro to Plotting and Visualization	*Practice using plot.ly, iPython notebook and Tableau to plot and visualize data*

WEEK 2: EDA, PANDAS & SCIPY

Project 2	*Using a dirty dataset of Billboard hits, students will use Pandas to munge data, create a problem statement, and perform exploratory analysis for a local music publisher.*
Experiment Design	*Plan experimental study design with a well thought out problem statement and data framework*
Pandas & Pivot Tables	*Use Pandas to read, clean, parse, and plot data using functions such as boolean, indexing, math series, joins, and others*
SciPy	*Review statistical testing concepts (p-values, confidence intervals, lambda functions, correlation/causation) with SciPy*

Topics and projects subject to change based on the instructor and local job market requirements.

DATA SCIENCE IMMERSIVE SYLLABUS

WEEK 3: LINEAR REGRESSIONS, SCIKIT-LEARN, GRADIENT DESCENT, & MODEL FIT

Project 3	*Given access to state liquor sales data, students will choose between performing market research or conducting tax audits, using Pandas, stats models, and sklearn to transform data, perform linear regression, and plot results.*
Linear regression, stats models, and scikit learn	*Use scikit learn and statsmodels to run linear regression models and evaluate model fit*
Bias-Variance Tradeoff	*Understand bias-variance trade-off to evaluate machine learning models*
Gradient Descent	*Look behind the hood at the math and theory of how gradient descent helps to optimize the loss function for models*
Regularization & Optimization	*Learn to apply regularization and optimization when evaluating model fit*

WEEK 4: LOGISTIC REGRESSION, NLP, AND WEB SCRAPING

Project 4	*Posing as a federal contractor, students will scrape website data and use Pandas, Statsmodels, and NLTK to clean and analyze data, perform logistic regression, and evaluate correlation coefficients.*
Web Scraping	*Learn to scrape website data using popular scraping tools*
Logistic regression	*Build, evaluate, and refine a logistic regression model for a given business case study*
NLP	*Get introduced to natural language processing through sentiment analysis of scraped website data.*
Intro to Capstone Project	*Get introduced to the final project. Begin to plan your experiment design and look for data sources*

WEEK 5: SQL, DATABASES, & CLASSIFICATION

Project 5	*As a researcher for a disaster response agency, students will pull remote data on Titanic survivors in order to build a local database, run a logistic regression classification model, and validate results from test subsets.*
SQL & Remote Databases	*Get introduced to different types of databases, review SQL commands, and practice connecting to and pulling data from a remote AWS database*
Feature Selection	*Use feature selection to deepen knowledge of model evaluation*
kNN & SVMs	*Begin to look at classification models through an application of the kNN algorithm and learn how SVMs can simplify the process of analyzing data for supervised learning algorithms*

DATA SCIENCE IMMERSIVE SYLLABUS

WEEK 6: APIS, TREES & ENSEMBLE METHODS

Project 6	*Acting as a member of the Netflix data science team, students will collect data from IMDB's API and use SQL in order to construct a random forest model that identifies ratings indicators and correlates these findings with viewer sentiment analysis.*
JSON & APIs	*Learn to pull JSON data from APIs as another potential data source*
Ensemble Models	*Build and evaluate ensemble models, using decision trees, random forests, bagging, and boosting*

WEEK 7: PCA, CLUSTERING, K-MEANS & AWS

Project 7	*Working as an airport operations consultant, students will analyze plane delay data in US airports, performing a principle component analysis.*
Clustering	*Define clustering and it's advantages and disadvantages from classification models*
K-Means	*Practice building and evaluating a k-means algorithm*
PCA	*Convert a set of observations or variables into principal components in order to improve predictive analysis.*
PostgreSQL	*Learn to build and maintain your own postgreSQL database*

WEEK 8: BAYESIAN INFERENCE & LDA

Capstone Project, Part 1	*During the first part of the capstone project, students will select and verify topics and datasets, in order to develop problem statements, goals, and success criteria.*
Bayesian Methods	*Build a linear regression model with bayesian methods*
LDA	*Refine data using latent dirichlet allocation (LDA)*
Naive Bayes	*Learn how Naive Bayes can simplify the process of analyzing data for supervised learning algorithms*

Topics and projects subject to change based on the instructor and local job market requirements.

DATA SCIENCE IMMERSIVE SYLLABUS

WEEK 9: WORKING WITH TIME SERIES

Capstone Project, Part 2	*During the second part of the capstone project, students will create their own database, acquire and clean their data, and perform exploratory data analysis using real world tools and processes.*
Time Series & Autocorrelation	*Analyze and visualize time series data using Pandas, plot.ly and Tableau*
ARIMA Model	*Use the ARIMA model to make predictions with time series data*

WEEK 10: INTRO TO BIG DATA AND SPARK

Capstone Project, Part 3	*During the third part of the capstone project, students will build and tune a predictive model from their data using machine learning algorithms to analyze, classify, and plot their data.*
Hadoop & MapReduce	*Get introduced to the history and use of Hadoop as well as the advantages and disadvantages of using parallel or distributed systems to store, access, and analyze big data*
Hive & Spark	*Gain an introductory understanding of how Hive interacts with Hadoop and learn about Spark's advantages through big data case studies*

WEEK 11 – 12: JOB PREP & CAPSTONE

Capstone Project, Part 4 & 5	*Students will create a detailed technical model notebook to share with peer audiences, while adapting their findings into a visual narrative for non-technical stakeholders.*
Advanced Topics	*Specific advanced topics will vary by market and instructor*
Portfolio Development	*Work with career coaches to create and polish your portfolio for employers*
Interview Prep	*Practice data science case studies to prep for job interviews*