INTRODUCTION

A warm welcome to the Springboard Data Science Intensive Workshop.

Our team of industry experts has picked the best learning resources on Data Science and structured them into a logical curriculum below for you. Note that the individual resources need not be authored in-house, but are curated from the best in the field. We do provide the glue that stitches everything together and fills gaps, if any. We believe this is our unique strength because we stand on the shoulders of giants, instead of being restricted to in-house content.

Think of us as a college professor who creates a curriculum with the most formidable book chapters, research papers, projects instead of teaching from a single textbook.

**Working through materials**

As you work through the material, click "Mark Complete”. This allows both you and your mentor a way to keep track of your progress. We have included some Electives and Optional/Additional Resources, in case you'd like to go deeper on some topics. They are not required for completion of the course hence do not have the "Mark Complete” button.

**Turning in projects**

Once you complete your projects, you can submit projects by clicking the “Submit” button.

**Editing projects**

If you submitted a project but need to re-submit it with revisions based on feedback, please edit your links and let your mentor know so they can provide feedback.

We have a recommended schedule (~10-15 hours/week) so that the class can follow a rhythm and maximize benefit from discussions. However, this is a self-paced course, so feel free to go faster or slower as you please.

Week 1 & 2: Programming Bootup, Data Wrangling

Weeks 3 - 5: Data Story & Inferential Statistics

Weeks 6: Capstone Milestone Report

Weeks 7-9: Machine Learning

Weeks 10-12: Capstone Project

In addition to the core material, there are a bunch of additional resources and electives in case some topics catch your fancy and you would like to dive deeper there. Feel free to pace it and work with your mentor to tailor the course to your meet your goals and interests!

Without further ado, let's dive in!

Completed

 10 Minutes

[**Student Guide**](http://help.springboard.com/help/article/link/start-here)

Get ready for the workshop with this guide, which contains important information about the wor...more

Completed

 5 Minutes

[**How to Join Your Weekly Calls**](https://youtu.be/lgCPnnMOT6Y)

This video will give you a quick overview of how to join your weekly calls with your mentor.

Completed

 10 Minutes

[**Prepare for your first Mentor Call**](http://help.springboard.com/help/article/link/your-first-mentor-call)

Your first mentor call is coming up soon! This link will bring you back to the Student Guide s...more

Completed

 15 - 30 Minutes

[**Read the Capstone Project Guidelines**](https://docs.google.com/document/d/1cpA-lMcLGwCQIst76Eeo8onxbXqu_TnvoyPnWLnGfuY/edit?usp=sharing)

Submitted

 1 Hour

**Submit 3 potential Capstone project ideas and discuss with your mentor**

Project Link: [https%3A//docs.google.com/document/d/1wAyyUfELYsbynoTlRH3Z7y6OUD\_RqOJLVseHDEulGoE/edit%3Fusp%3Dsharing](https://www.springboard.com/workshops/data-science-intensive/learn/current)

Your Capstone project is one of the most important parts of this course - it will give you pra...more

 6.5+ HOURS

PROGRAMMING BOOTUP

As promised earlier, we’ll be switching between technical and career materials in this program. Now that you’ve laid the foundation for your job search, it’s time to start beefing up your technical skills.

In this module, you will start using an ecosystem of useful and powerful tools for doing data science with Python - primarily Jupyter Notebook, Pandas, and matplotlib (scikit-learn is another, but needs familiarity with machine learning, and will be introduced later in the course). The best approach is to start using them together to gain basic familiarity, and then continue learning by using them in the rest of the course. The goal is basic familiarity, hence the short time period suggestions. You will get more practice with each of them in the modules to follow. You can refer back to these resources as you need them, and spend more time as you make progress through the course.

Completed

 10 Minutes

[**Get Anaconda**](http://continuum.io/downloads)

Get Anaconda, a collection of Python packages. It has a lot more than you will probably use, b...more

Completed

 30 Minutes

[**Python, of course!**](http://www.stavros.io/tutorials/python/)

We assume you are already comfortable programming in at least one language. If you are unfamil...more

Completed

 5 - 10 Minutes

[**4 Python built-in help functions you should know**](http://www.linuxnix.com/python-builtin-helpdir-help-type-and-___doc_-functions/)

When you find yourself wondering, "What does this weirdly Python object let me do?", the built...more



**Elective: Brush up your Python with DataCamp**

If you feel you need a quick Python refresher, this online course from DataCamp should help yo...more

 0.5+ HOURS

**JUPYTER NOTEBOOK**

Jupyter (formerly iPython) notebook is an interactive programming environment, that allows for coding, data exploration, and debugging in the web browser. It comes bundled with the Anaconda distribution, so you can start right away!

***Note:*** The main difference between iPython and Jupyter is that while iPython is tied to Python, Jupyter can support multiple languages. For Python programmers, it makes little difference in practice, so if you're already familiar with iPython, you can simply treat Jupyter as its latest version.

Completed

 20 Minutes

[**Get started with Jupyter notebook**](https://jupyter.readthedocs.io/en/latest/install.html)

Completed

 10 Minutes

[**How to run code in Jupyter notebook**](http://nbviewer.ipython.org/github/ipython/ipython/blob/3.x/examples/Notebook/Running%20Code.ipynb)

Completed

 20 Minutes

[**A video walkthrough on how to use the Jupyter notebook**](https://www.youtube.com/watch?v=qb7FT68tcA8)



**Additional resources**

Read some more about Jupyter, and look out for a listing of keyboard shortcuts [here](http://ipython.readthedocs.org/en/stable/).

Completed

 30 Minutes

[**Pandas - Quick Intro**](http://pandas.pydata.org/pandas-docs/stable/10min.html)

Pandas extends Python with data structures and operations for data manipulation and analysis.

 1.5+ HOURS

**MATPLOTLIB**

Matplotlib is the main plotting library for Python. The name originates from its (superficial) similarity to Matlab. It's extremely powerful, but can also require a bit of a learning curve to really use it well. Newer libraries such as Seaborn have emerged to abstract away some of its complexity, but it's still worth learning the fundamentals of Matplotlib.

Completed

 1 - 2 Hours

[**Basic plotting, default settings, and formatting**](http://nbviewer.ipython.org/github/jrjohansson/scientific-python-lectures/blob/master/Lecture-4-Matplotlib.ipynb)

Completed

 30 Minutes

[**Practice commonly used plot types**](http://jakevdp.github.io/mpl_tutorial/tutorial_pages/tut3.html)

You will have a chance to play with some more amazing visualization tools later in the course.

Completed

 20 - 30 Minutes

[**A quick tutorial on Seaborn**](https://youtu.be/E8OQAdQlljE)

As you may have noticed, matplotlib can be a bit clunky and repetitive at times, and the defau...more

 1+ HOURS

**GIT AND GITHUB**

If you don’t have a Facebook account, some reasonable people may doubt your existence in the Multiverse. Ditto for a Github account in the Tech Universe.

Completed

 1 - 2 Hours

[**Start your Github profile**](http://kbroman.org/github_tutorial/)

Stake your cyber-estate claim for free with the first four sections (upto ‘Start a new reposit...more

Submitted

 5 Minutes

**Submit a link to your Github profile**

Project Link: [https%3A//github.com/csnyd/2017](https://www.springboard.com/workshops/data-science-intensive/learn/current)

Submit a link to your Github profile. It should contain a new repository created for code from...more



**Additional Programming Resources**

If you would like to spend extra time on programming in Python, Learn Python and Code Academy...more

 17.5+ HOURS

DATA WRANGLING

Data wrangling is the process of taking data in its ‘raw’ form and manipulating it in various ways into a ‘useful’ form. The ‘raw’ data might come to you from an original source or another intermediary source - data sources can vary from unstructured/semi-structured text files (.txt) and delimited/structured/nested format files (excel, csv, json, xml) to relational databases (SQL) and non-relational databases (‘NoSQL’). The ‘useful’ form of the data will be dictated by your current and/or anticipated needs, such as statistical inference or inputs to machine learning algorithms.  
  
Data from original sources is often

* ‘messy’ or ‘dirty’ in the sense that it might contain values that are invalid, missing, corrupted, inconsistent or non-uniform
* ‘coarse’, in the sense that it needs to be refined, transformed or combined with other data.

For the above reasons, in order to derive meaningful conclusions from the data, it needs be ‘cleaned’, ‘scrubbed’, ‘munged’, ‘wrangled’.

 4+ HOURS

**PANDAS DEEP DIVE**

Pandas is often a tool of choice for data wrangling operations. You’ve already had a taster, and now it is time to dive deeper and wider with this hands-on tutorial.

Completed

 15 Minutes

[**Download source files**](https://github.com/brandon-rhodes/pycon-pandas-tutorial)

Detailed instructions for the tutorial can be found [here](https://github.com/brandon-rhodes/pycon-pandas-tutorial#detailed-instructions).

Mark complete

 4 - 5 Hours

[**Watch the video and code along**](https://www.youtube.com/watch?v=5JnMutdy6Fw)

 1 HOUR

**DATA CLEANING WITH PANDAS**

Mark complete

 30 Minutes

[**Other people's messy data**](https://www.youtube.com/watch?v=_eQ_8U5kruQ)

Watch this talk by Mali Akmankalp (PyCon 2015) on the different ways in which data can be mess...more

Mark complete

 30 Minutes

[**Working with missing data**](http://pandas.pydata.org/pandas-docs/stable/missing_data.html)

Reinforce your bag of tricks for missing data by going through official Pandas documentation.

 4 HOURS

**WORKING WITH DATA IN FILES**

Data sources can vary from unstructured or semi-structured text files (.txt) and delimited, structured or nested format files (excel, csv, json, xml). While working with data stored in files, the basic operation is to read files into a Pandas data frame. Most of the formats have standard row and column tables, and are relatively easy to work with. JSON and XML are nested formats and need some more work.

Mark complete

 30 Minutes

[**Practice working with different file formats**](http://pandas.pydata.org/pandas-docs/stable/io.html)

For JSON, pay particular attention to normalization.

Submit project

 1.5 Hours

[**Work on JSON based data exercises and submit on your Github**](https://www.springboard.com/static/project_files/data_wrangling_json.zip)

Instructions for the exercise are in the Jupyter notebook.

Mark complete

 30 Minutes

[**Working with XML files in Python**](http://luisartola.com/software/2010/easy-xml-in-python/)

XML does not have a direct reader in Pandas as yet, here is how to deal with it.

Submit project

 1.5 Hours

[**Work on XML based data exercises and submit on your Github**](https://www.springboard.com/static/project_files/data_wrangling_xml.zip)

 6.5+ HOURS

**WORKING WITH DATA IN DATABASES**

There are two kinds of databases - relational (most commonly based on SQL), and more general (‘NoSQL’ or ‘Not Only SQL’).

Mark complete

 4 - 6 Hours

[**Mode Analytics: Learn SQL**](https://sqlschool.modeanalytics.com/toc/)

The wonderful folks at Mode Analytics have created a great SQL learning tool (you'll need a fr...more

Mark complete

 2 - 4 Hours

[**Mode Analytics: Analytics Training**](https://sqlschool.modeanalytics.com/analytics-training/yammer/yammer/)

Now that you have some SQL under your belt, it's time to apply it to some real data! Mode Anal...more

Mark complete

 30 Minutes

[**Overview of NoSQL databases**](http://www.thoughtworks.com/insights/blog/nosql-databases-overview)

The last few years have seen a steep growth in applications for which relational databases are...more



**Additional resources**

* [3 hour Pandas tutorial](https://www.youtube.com/watch?v=w26x-z-BdWQ).
* Udacity "[Data Wrangling with MongoDB](https://www.udacity.com/course/data-wrangling-with-mongodb--ud032)" course.

Submit project

 1 Hour

**Submit your Capstone project proposal**

Finalize one Capstone idea based on the feedback you got from mentor(s) and peers on your Sect...more

 9.5+ HOURS

DATA STORY

DJ Patil, the first Chief Data Scientist of the US, repeatedly says “Data science is about telling stories and data scientists are the storytellers.” A data story is a powerful way to present insights, by combining visualizations and text into a narrative. In this module, you will use the awesome programming and data wrangling skills you’ve gained to analyze an interesting dataset and come up with a data story.

 4+ HOURS

**CORE RESOURCES FOR DATA STORY**

Mark complete

 1.5 Hours - 2 Hours

[**Exploratory Data Analysis**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=a4e81697-fd86-415c-9b29-c14ea7ec15f2)

Exploratory Data Analysis (EDA) is an approach for summarizing and visualizing the important c...more

Mark complete

 1.5 Hours - 2 Hours

[**Storytelling and Effective Communication**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=7f968df9-404a-46a2-ae5f-e35479875f95)

Data science is not only about the data and the algorithms, it’s also about the story you tell...more

Mark complete

 1.5 Hours - 2 Hours

[**Effective Presentations**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=d213b305-65a8-4ee8-95d3-4dfc4a437205)

You know how to tell a good story, awesome! In most workplaces as a data scientist, you'll be ...more

Submit project

 5 - 10 Hours

**Work on your Data Story**

How does one go about creating a data story? You have some pointers from the material you've j...more



**Additional Reading**

Lead with story Making beautiful visualizations with Matplotlib A brief intro to visualizati...more

 16.5+ HOURS

INFERENTIAL STATISTICS

Descriptive statistics are useful for discovering and communicating insights from data. Inferential statistics are useful for drawing conclusions and predicting outcomes. In this course, the focus is on inferential statistics. For students unfamiliar or rusty with probability and descriptive statistics, [Khan Academy’s probability track](https://www.khanacademy.org/math/probability) is a good refresher.

We begin with a wonderful workshop by Allen Downey from Pycon 2015, followed by some material from Khan Academy's Inferential Statistics course. Some reading material on A/B testing follows, which is a form of inference, and its use in the context of website features.

 4+ HOURS

**BASIC STATISTICAL INFERENCE**

You have a data sample, and would like to be able to say something about the population or process that generated this sample. You have a hunch, or a hypothesis, say something like, doing X results in greater benefit or harm compared to doing Y. How can you establish this with confidence, and convince all comers? How much confidence do you have?

Mark complete

 4 - 6 Hours

[**Statistical Inference with Computational Methods - Allen Downey**](https://www.youtube.com/watch?v=5Vjrqnk7Igs)

Allen Downey, an amazing statistics educator, Professor at Olin College and author (Think Stat...more

 6+ HOURS

**A DEEPER DIVE INTO HYPOTHESIS TESTING - KHAN ACADEMY**

At this point, you have a really solid grounding in inferential statistics. While we agree with Prof. Downey that a large part of the data science world is slowly moving away from p-values as a method of evaluating significance, it's still widely used in both academia and industry. As a result, we recommend that you to acquire deeper knowledge of this topic, using a few specific sections of Khan Academy's Inferential Statistics course.

Please complete all the videos in the following sections. We've linked below to the first video of each section.

Mark complete

 1 - 2 Hours

[**Khan Academy: Confidence Intervals (One Sample)**](https://www.khanacademy.org/math/statistics-probability/confidence-intervals-one-sample)

How do you know how good your descriptive statistics are in actually describing the population...more

Mark complete

 2 - 3 Hours

[**Khan Academy: Significance Tests (One Sample)**](https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample)

How do we know whether some claim made about a sample apply to the population as a whole, or i...more

Mark complete

 2 - 3 Hours

[**Khan Academy: Significance Tests and Confidence Intervals**](https://www.khanacademy.org/math/statistics-probability/significance-tests-confidence-intervals-two-samples)

Let's apply what you know about confidence intervals and significance tests to situations that...more

Mark complete

 1 - 2 Hours

[**Khan Academy: Inference for categorical data**](https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests)

This section covers Chi-square tests, which are a family of significance tests that give us wa...more

 4.5+ HOURS

**EXPLORATORY DATA ANALYSIS PROJECTS**

Now that you have a foundation in Inferential Statistics and Hypothesis Testing, it's time to see those ideas in action! The following mini-projects will walk you through how hypothesis testing can be used to elicit insights from data and create good data stories.

Submit project

 1.5 Hours

[**EDA: Human Body Temperature**](https://www.springboard.com/archeio/download/b7bfb45ac9894b57b999c263c87395bc/)

In this exercise, you will analyze a dataset of human body temperatures and employ the concept...more

Submit project

 1 Hour

[**EDA: Examine Racial Discrimination**](https://www.springboard.com/archeio/download/154e33ef54ae4197abacdea3cfa7c8b0/)

In this exercise, you will perform a statistical analysis to establish whether race has a sign...more

Submit project

 2 Hours

[**EDA: Reduce Hospital Readmissions**](https://www.springboard.com/archeio/download/5ba938b305ee4adab59b66791f8ca268/)

In this exercise, you will: critique a preliminary analysis of data and recommendations for r...more

 2 HOURS

**REGRESSION AND CORRELATION**

You see that the amount of A depends on the quantity of B. How can you statistically quantify this dependence? How can you model a relationship between a dependent variable and independent variables?

Mark complete

 2 Hours

[**Khan Academy: Describing relationships in quantitative data**](https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data)

This section covers various techniques to find relationships between multiple variables in a d...more

 15+ MINUTES

**A/B TESTING**

A/B Testing is a form of hypothesis testing, a randomized experiment with two variants, that has gained prominence for web and mobile design in recent years.

Mark complete

 5 Minutes

[**Introduction to A/B testing**](http://20bits.com/article/an-introduction-to-ab-testing)

Mark complete

 10 Minutes

[**A/B Testing with websites**](http://blog.hubspot.com/marketing/a-b-testing-experiments-examples)



**Additional Resources**

Broader applications: Randomized Control Trials for measuring social impact Udacity A/B Testi...more

 5+ HOURS

CAPSTONE MILESTONE REPORT

You have proposed a project, collected a data set, cleaned up the data and explored it with descriptive and inferential statistics techniques. Now’s the time to take stock of what you’ve learned. The project milestone is an opportunity for you to practice your data story skills. Your milestone will be reached when you produce an early draft of your final Capstone report. This is a slightly longer (3-5 page) draft that should have the following:

* An introduction to the problem: What is the problem? Who is the Client? (Feel free to reuse points 1-2 from your proposal document)
* A deeper dive into the data set:
  + What important fields and information does the data set have?
  + What are its limitations i.e. what are some questions that you cannot answer with this data set?
  + What kind of cleaning and wrangling did you need to do?
  + Are there other datasets you can find, use and combine with, to answer the questions that matter?
* Any preliminary exploration you’ve performed and your initial findings. Test the hypotheses one at a time. Often, the data story emerges as a result of a sequence of testing hypothesis e.g. You first tested if X was true, and because it wasn't, you tried Y, which turned out to be true.
* Based on these findings, what approach are you going to take? How has your approach changed from what you initially proposed, if applicable?

Add your code and milestone report to the github repository. As before, once your mentor has approved your milestone document, please share the github repository URL on the community and ask the community for feedback.

While we require only one milestone report, we encourage you and your mentor to plan multiple milestones, especially for more complex projects.

Submit project

 5 - 10 Hours

**Submit your Capstone Milestone Report**

 21+ HOURS

MACHINE LEARNING

By now, you’ve gained impressive skills in many of the core steps of the data science process: data wrangling, storytelling, and inferential statistics. What you'll build next is some lean and mean machine learning muscle. Combining aspects of computer science and statistics to extract useful insights and predictions from data,  machine learning is a key skill for data scientists. You should get a good grip on what the strengths and limitations of each machine learning algorithm are, assumptions underlying them, how to evaluate their performance, and how to arrive at the right algorithm to use in different problem scenarios.

We use material from Harvard's amazing [CS109 course](http://cs109.github.io/2015/). We find this course to be a great balance between technical rigor and practical applications.

**Note to students:** Our curriculum aims to provide learners with a good coverage of ML techniques that are most widely used today, while making sure that it is practical to learn in a relatively short timeframe. Don't worry if some of the material, especially the math, feels a bit difficult. As long as you have a good intuitive understanding of the algorithms, you'll be OK!

 18+ HOURS

**SUPERVISED LEARNING**

You have some initial data and it is ‘labeled’, say, as ‘True/False’ or ‘Normal/Abnormal’, ‘Class 1, Class 2, Class 3’ etc. You want to extract some ‘features’ from the data that, when passed through a function, will generate the labels as accurately as possible. How do you find this function? By using a classification algorithm! And why do you want to find this function? So that when you get data for which the labels are not available, you can automatically generate the labels. Let's start with learning the classic classification algorithms.



 9+ Hours

**Linear and Logistic Regression**

Mark complete

 2 - 3 Hours

[**Bias and Regression**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=afe70053-b8b7-43d3-9c2f-f482f479baf7)

We start with Regression, a technique to predict unknown values, when the values are real numb...more

Mark complete

 2 - 3 Hours

[**Regression (contd)**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=664f668e-e008-4f44-8600-e09ee6d629b0)

We finish up with Linear Regression, and start  exploring linear Logistic Regression. Even tho...more

Mark complete

 2 - 3 Hours

[**Classification, kNN, Cross-validation, Dimensionality Reduction**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=c322c0d5-9cf9-4deb-b59f-d6741064ba8a)

Now you have a supervised regression or classification model trained with your data. How do yo...more

Submit project

 2 - 3 Hours

[**Linear Regression Mini-Project: Boston Housing**](https://www.springboard.com/archeio/download/18350417e3834bbd8ed17f66fec496c3/)

Instructions: Please download and open the zipped file and work in the Jupyter notebook in the...more

Submit project

 1 - 2 Hours

[**Logistic Regression: Heights and Weights**](https://www.springboard.com/archeio/download/e15045f38bb54344a952bef3ed291540/)

Instructions: Please download and open the zipped file and work in the Jupyter notebook in the...more



 4+ Hours

**SVM and Trees**

The algorithms that we have studied so far are only the simplest ones in machine learning. The...more

Mark complete

 2 - 3 Hours

[**SVM and Evaluation**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=f21fcc8f-93a8-49f6-9ff8-0f339b0728bd)

There are many advanced algorithms to handle more complex data sets, and Support Vector Machin...more

Mark complete

 2 - 3 Hours

[**Decision Trees and Random Forests**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=8892a8b7-25eb-4bc5-80b6-47b9cf681a05)

Tree-based algorithms (decision trees, random forests) are some of the most popular and effect...more



 2+ Hours

[**Elective: Ensemble Methods**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=4831ebf0-7832-42c5-9339-5b5e08dd3e92)

You've now learned about many learning algorithms, each with their own pros and cons. So, whic...more



 3+ Hours

**Bayesian Methods and Text Data**

Bayesian methods are a powerful suite of techniques that are gaining more and more traction in...more

Mark complete

 2 - 3 Hours

[**Bayes Theorem and Bayesian Methods**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=233f6c34-306f-481b-8ea5-be33076eb6a8)

In this lecture, we get introduced to Bayes Theorem, learn about algorithms like Naive Bayes i...more



 1+ Hours

[**Tutorial: Sentiment Classification using scikit-learn**](https://youtu.be/y3ZTKFZ-1QQ)

An amazing demonstration of practical text analysis and machine learning by one of our mentors...more

Submit project

 1 - 2 Hours

[**Naive Bayes: Predicting movie ratings from reviews**](https://www.springboard.com/archeio/download/02471e657fa246fa8a2b28ad45fdd21d/)

Instructions: Please download and open the zipped file and work in the Jupyter notebook in the...more



 2+ Hours

**Best Practices**

At this point, we have learned a lot of different techniques and methods, both for performing ...more

Mark complete

 2 - 3 Hours

[**Best Practices in Supervised Learning**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=b33eec92-d049-4353-a904-5054eb718aff)

NOTE: Slide deck for this lecture. This particular video does not have the slides, but you can...more



**Elective: Recommendation Systems**

Recommendation systems are everywhere; from Amazon recommending books and products, Netflix re...more



 2+ Hours

[**Recommendation Systems**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=afee45b9-dcf5-4f29-bc60-871aa78f1cf8)

We start with a lecture on recommendation systems from the Harvard CS 109 class.



 1+ Hours

[**Tutorial: Building a recommendation system in Python**](https://www.youtube.com/watch?v=F6gWjOc1FUs)

Now that you have the fundamentals of how recommendation systems work, how do you actually bui...more

 3+ HOURS

**UNSUPERVISED LEARNING**

You just have data and no labels, but you’re not the kind who gives up, and want to find some structure in the data, and define your own classes, perhaps? Unsupervised learning time!

PS: Some of the concepts in the following sections require a basic understanding of Linear Algebra. In case you have difficulty with the math in the material that follows, here's a [quick summary](https://www.coursera.org/learn/machine-learning/supplement/NMXXL/linear-algebra-review).



 3+ Hours

**Introduction to Clustering**

Mark complete

 2 - 3 Hours

[**Clustering**](https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=27aebc37-3c53-4ce8-a7cb-b2be397aa2f0)

Linear and Logistic Regression are both examples of learning algorithms that are “supervised” ...more



 1+ Hours

[**Tutorial: Clustering with scikit-learn**](https://youtu.be/-J9ZICyev5E)

A deeper dive into the various clustering algorithms available in scikit-learn.

Submit project

 1 - 2 Hours

[**Clustering: Customer Segmentation**](https://www.springboard.com/archeio/download/645b500146aa41f0bb810175de8e20a0/)

Instructions: Please download and open the zipped file and work in the Jupyter notebook in the...more



 2+ Hours

[**Elective: Anomaly Detection**](https://www.youtube.com/playlist?list=PLnnr1O8OWc6b_KudrHI0Vcf6sEu6xs9Ky)

When you have a large and complex system, one way to find out if everything is working well is...more



**Additional Scikit-learn resources**

* [The basics](http://scikit-learn.org/stable/tutorial/basic/tutorial.html)
* [1 hr tutorial](https://github.com/ogrisel/sklearn_pycon2014)
* [3 hr tutorial](https://www.youtube.com/watch?v=HjAB45qsx_c)

 20+ HOURS

CAPSTONE PROJECT

Time to bring everything together with the Capstone! Discuss a stopping point for the project with your mentor, and submit the following:

1. Code for your project, well-documented on github.
2. A final paper explaining the problem, your approach and your findings in complete technical detail. Include ideas for further research, as well as up to 3 concrete recommendations for your client on how to use your findings.
3. A slide deck or a blog post which presents your analysis to your clients (e.g. non-technical and business teams) in an easy to understand, but compelling way. As a data scientist in a company, you’ll be frequently called upon to produce these kinds of materials.
4. Extra credit: Actually present or send your report/slide deck to your designated client and let us know what kind of response you received from them.

**Project Evaluation**

For Springboard to consider your workshop complete, and issue a certificate of completion, your mentor needs to approve your final project submission per the rubric described below. In case the project is not approved, please discuss the feedback from your mentor and resubmit in case improvements are necessary for the approval. Your Student Advisor will not be able to process your workshop completion until your project is approved by your mentor!

**Capstone Project Rubric for Data Science**

We use the following rubric for evaluating final Capstone Projects. Please take a good look at it and make sure you discuss with your mentor to agree on success criteria.

[View the Capstone Project Rubric here](https://docs.google.com/spreadsheets/d/1P4n5G-_uw_AiAgkWL6cuEfAuHU2PmKc1J1iCqB01nXo/edit?usp=sharing)

The first tab is the rubric itself, and the next two tabs are sample projects that have been graded using the rubric. The rubric consists of several evaluation criteria, each graded on a 3-point scale: Below Expectations, Meets Expectations, or Exceeds Expectations. Your project is considered Complete if you get a Meets Expectations or Exceeds Expectations on ALL of the criteria.

Because your mentor decides your grade, it’s vital that you work with your mentor during the entire project to understand and agree on what the bar is for each criterion, and incorporate their feedback during the intermediate stages.

For now,  your mentor will be using a copy of this sheet to grade your project and returning the graded sheet by email along with your assessment.

Share these with and optionally present to your friends and mentors in this class to get a feel for the life of a Data Scientist. With this, you are all set to step out in the wild and vast world of Data Science. Happy exploring, and wish you the best!

Submit project

 20 - 30 Hours

**Complete & Submit Capstone Project**

ELECTIVE 1: ADVANCED DATA VISUALIZATION

We recommend the following three resources for an advanced exposure to data visualization.



[**Communicating with Data**](https://www.udacity.com/course/data-visualization-and-d3js--ud507)

Udacity’s course on “Communicating with Data” which is based on d3.js. It includes visualizati...more



[**Interactive Data Visualisation for the Web**](http://chimera.labs.oreilly.com/books/1230000000345/index.html)

Scott Murray’s book titled “Interactive Data Visualisation for the Web” is an excellent place ...more



[**Stay up-to-date with this weekly newsletter**](https://www.dashingd3js.com/data-visualization-and-d3-newsletter)

Sebastian Gutierrez’ weekly newsletter titled “DashingD3.js” - is a great curated resource to ...more

ELECTIVE 2: DATA SCIENCE AT SCALE

When your data gets real big, you will need to use cluster computing. Apache Spark is a popular and powerful cluster computing framework for working with data at scale.



[**Get started with this Tutorial**](https://districtdatalabs.silvrback.com/getting-started-with-spark-in-python)



[**Introductory course on Spark**](https://www.edx.org/course/introduction-big-data-apache-spark-uc-berkeleyx-cs100-1x)



[**Machine learning with Spark**](https://www.edx.org/course/scalable-machine-learning-uc-berkeleyx-cs190-1x)

CAREER RESOURCES

1. Preparing for interviews: Find some great questions for practice [here](http://blog.udacity.com/2015/04/data-science-interview-questions.html) and [here](http://www.datasciencequestions.com/)
2. [Interviews with data scientists](http://www.thedatasciencehandbook.com/)
3. [Building data products](http://firstround.com/review/everything-we-wish-wed-known-about-building-data-products/)
4. [Design Thinking for Data Scientists](https://www.linkedin.com/pulse/design-thinking-data-science-george-roumeliotis)
5. Advice for strengthening portfolio:
   * Participate in competitions: [Kaggle](http://www.kaggle.com/), [CrowdAnalytyx](https://www.crowdanalytix.com/)
   * Take on freelance projects on [Odesk](https://www.odesk.com/), [Elance](https://www.elance.com/), [Experfy](https://www.experfy.com/)
   * Participate in Data hackathons.