

Chapter 4.

DataCamp:

Interactive Exercises: Data Types for Data Science

Interactive Exercises: Python Data Science Toolbox (Part 1)

Interactive Exercises: Introduction to Data Visualization in Python

Transforming Code into Beautiful, Idiomatic Python:

<https://www.youtube.com/watch?v=anrOzOapJ2E>

Video: PEP8 and Writing Readable Code

<https://www.youtube.com/embed/zs6BzkgHvMA?autoplay=1&controls=1&showinfo=0&rel=0>

Video: The Importance of Perseverance in Programming

<https://www.youtube.com/embed/DwQ7psiU23I?autoplay=1&controls=1&showinfo=0&rel=0>

Link: PEP8 Documentation:

<http://pep8.org/>

Video: Using Decisions in Framing Analytics Problems

https://www.youtube.com/watch?v=B_gbR6Tj5ss

<https://dssg.uchicago.edu/2013/06/26/training-data-scientists-problem-solving/>

Link: Interview Practice - The Python Data Science Stack

Practice Interview Questions

Programming Boot-Up

- What native data structures can you name in Python?
 - Of these, which are mutable, and which are immutable?
- Explain the difference between a list and a dictionary?
- In a list, what data types can be elements?
- In a dictionary, what data types can the key be? And the values? Why?
- When would you use a list vs. a tuple vs. a set in Python?
- Explain the difference between a for loop and a while loop.
- What packages in the standard library, useful for Data Science work, do you know?
- Do you know any additional data structures available in the standard library?
- Can you explain what a list or dict comprehension is?

Chapter 5 Data Wrangling

DataCamp

Interactive Exercises: Pandas Foundations

Interactive Exercises: Manipulating DataFrames with Pandas

Interactive Exercises: Merging DataFrames with Pandas

Interactive Exercises: Cleaning Data in Python

Video: Pandas from the ground up:

<https://www.youtube.com/embed/5JnMutdy6Fw?autoplay=1&controls=1&showinfo=0&rel=0>

Data Camp

Interactive Exercises: Python Data Science Toolbox (Part 2)

Interactive Exercises: Importing Data in Python (Part 1)

Interactive Exercises: Importing Data in Python (Part 2)

Project: JSON Based Data Exercise

This is the first of many mini-projects that you'll complete to consolidate your learning and apply the tools you've learned. This World Bank dataset from a school quality improvement project in Ethiopia is a good example of a real-life dataset that you'll encounter as a data scientist. Practice your data wrangling skills on this mini-project before you apply them to your capstone. Submit your code on GitHub and add a link through the submit button below.

Interactive Exercises: Learn SQL with Mode Analytics

<https://community.modeanalytics.com/sql/tutorial/introduction-to-sql/>

Interactive Exercises: Analytics Training with Mode Analytics

Now that you have some SQL under your belt, it's time to apply it to some real data! Mode Analytics has a great set of exercises in case studies using data from Yammer, a popular corporate social network tool. **Read the overview and complete at least one of the three case studies to fulfill the requirements of this course.**

<https://community.modeanalytics.com/sql/tutorial/sql-business-analytics-training/>

Project: SQL Practice

 3 - 5 Hours

Steps:

1. Download the SQL file and follow the instructions to log into the provided SQL platform.
2. Fill in your answers to the questions in the SQL file.
3. Add your SQL file to a GitHub repository and submit a link to it.

Article: Overview of NoSQL databases

<https://www.thoughtworks.com/insights/blog/nosql-databases-overview>

Collecting data from the Internet

Article: Introduction to APIs

<https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-apis-application-programming-interfaces-5-apis-a-data-scientist-must-know/>

Video: Python 'Requests' Library Tutorial: Bitcoin Price

<https://www.springboard.com/workshops/data-science-career-track/learn/#/curriculum/2794>

Project: API Mini-Project

2 - 4 Hours

This is your second mini-project and chance to practice some new data wrangling techniques.

Steps:

1. Download and unzip the API Mini-Project file.
2. Open the Jupyter Notebook and answer the questions.
3. Add your completed assignment to a GitHub repository and submit the link using the space below.

Web scraping using Python

https://www.youtube.com/embed/O_j3OTXw2_E?autoplay=1&controls=1&showinfo=0&rel=0

Project: Capstone Project 1: Data Wrangling



6 - 10 Hours

Now that you have a basic ideas of the various data wrangling steps and techniques available, let's apply it to your capstone project.

The first step in completing your capstone project is to collect data. Depending on your dataset, you may apply some of the data wrangling techniques that you learned in this unit. Some of you may be using standard datasets and sources, such as Kaggle or Yelp, where minimal or no data wrangling is required. Students often find that this part of the project takes a lot longer than they estimated, which is completely normal. The more work you put in, the more you'll learn. Data wrangling is an important tool in a data scientist's toolbox!

Steps:

1. Create a Google Doc (1-2 pages) describing the data wrangling steps you took to clean the dataset. Include answers to these questions in your submission:
 1. What kind of cleaning steps did you perform?
 2. How did you deal with missing values, if any?
 3. Were there outliers, and how did you handle them?
2. Submit a link to the document.
3. Discuss it with your mentor at the next call.
4. Revise and resubmit if needed.
5. Convert the final document to a .pdf and add it to your GitHub repository for this project.

This document will eventually become part of your milestone report.

Your project will be evaluated using this [rubric](#).

You've started building your data science portfolio! Keep it neat and well-organized to make it easier for you later on in the course and when you interview for jobs. At this point, your portfolio should include two mini-projects, a folder for a capstone project with a project proposal, and the data wrangling report. Additionally, your dataset should be ready to start exploratory data analysis.

Practice Interview Questions

Data Wrangling

Pandas

- What data structures does pandas introduce which aren't native to Python?
- **Video:** How can you deal with missing values?
- What is the difference between the .loc and the .iloc indexers?
- What file formats for storing data do you know?
- What is the standard way of marking missing values in pandas?
- **Video:** What features of pandas do you like particularly? Any that you dislike?
- What kind of indexes exist in pandas DataFrames?

SQL

- What are aggregations in SQL?
- Can you explain the different types of SQL JOINS?
- Give an example of some aggregation functions in SQL.
- Can you explain the difference between the WHERE and HAVING filters?

Chapter 6 Networking

Effective Networking : Build your network

Overview

With more technical knowledge of programming and data wrangling as part of your skill set and a data science centric LinkedIn profile, you're ready for the next phase of your job search strategy. This unit leads you through the first steps in building a data science network. We'll begin with using **meetups** as a strategy to meet new people. You'll then be introduced to the art of "**cold emailing**" as a way to grow your network. We have a networking etiquette guide that will show you important do's and don'ts of building your network as you attend events and reach out to new people. Finally, we'll address **imposter syndrome**, a common feeling that many professionals face, and ways to overcome it to grow your confidence in networking.

Chapter 7 Data Storytelling

Video: Exploratory Data Analysis

Lecture 3 from here:

<http://cs109.github.io/2015/pages/videos.html>

Video: Storytelling and Effective Communication

Lecture 6 from here:

<http://cs109.github.io/2015/pages/videos.html>

Project: Capstone Project 1: Data Story

Apply Data Storytelling

Now that you've seen some examples of great data storytelling, let's apply these techniques to your capstone project!

1

Project: Capstone Project 1: Data Story



10 - 20 Hours

How do you create a data story? You've learned the basics, but the information is probably a bit abstract at this point. Keep in mind that storytelling is an art, so you have to get your imagination bubbling. In this project, you'll learn some pointers to get those creative juices flowing. In the following sections, we'll work step-by-step to create your first data story.

Steps:

1. Ask the following questions and look for the answers using code and plots:
 1. Can you count something interesting?

2. Can you find trends (e.g. high, low, increasing, decreasing, anomalies)?
 3. Can you make a bar plot or a histogram?
 4. Can you compare two related quantities?
 5. Can you make a scatterplot?
 6. Can you make a time-series plot?
2. Looking at the plots, what are some insights you can make? Do you see any correlations? Is there a hypothesis you'd like to investigate further? What other questions do the insights lead you to ask?
 3. Now that you've asked questions, hopefully you've found some interesting insights. Is there a narrative or a way of presenting the insights using text and plots that tells a compelling story? What are some other trends/relationships you think will make the story more complete?

Submission: Submit links to a GitHub repository containing a Jupyter Notebook. The Notebook should contain:

- The questions you asked
- The trends you investigated
- The resulting visualizations and conclusions

You will be evaluated using this [rubric](#).

These results will go into your final portfolio and presentation. Organize your work as you go along to make it easier to compile later. Create slides and/or a presentation (.ppt) about your emerging data story.

In case the dataset is too large to commit to GitHub, please include a link to the dataset inside the Jupyter Notebook.

Discuss these results with your mentor at the next call. If you're having trouble with your code for this unit, you can reach out to your course TA for help by emailing projects@springboard.com, or post questions in the community forum.

Wrap-Up: Data Storytelling

1

Recap: Data Storytelling

In this unit, you **learned** the basic principles of effective storytelling and communicating through data. You **incorporated** your learning by creating a presentation, visualizing ideas and concepts, and adding a data story to your capstone project.

Data storytelling is a standard part of communicating insights, and it's a powerful skill to have as you begin making proposals and decisions based on analysis.

You've completed another key unit of the course! Keep up the momentum and reach out to your Springboard support system for any feedback and advice.

By now, you should've attended a data science meetup and spoken to your career coach about the event. If you haven't had a chance to attend one yet, take some time to attend an upcoming meeting and submit a short write-up on your experience. In case you haven't scheduled your second career call, do that now. Checking in with your career coach regularly will help you stay on track and keep you motivated in your job search.

Next, we'll focus on **inferential statistics**, a fundamental part of the Exploratory Data Analysis Theme.

Inferential Statistics

Overview

This is the second unit in the **Exploratory Data Analysis** theme, and it focuses on the basics of statistical inference, hypothesis testing, regression, correlation, and their applications, such as A/B testing. Descriptive statistics is useful for discovering and communicating insights from data, while inferential statistics is useful for drawing conclusions and predicting outcomes. Both are a part of exploratory data analysis and used to understand data stories.

This is a large unit which is broken into many parts. First, we'll brush up on the basics of probability and descriptive statistics. Then, we'll learn to apply them using Python. You'll practice these skills through a series of mini-projects, before applying them to your capstone project. Review the "What will help" section of the Unit Plan for more details, and reach out to your mentor and course TA if you struggle with the material.

[Unit Plan](#) (What you'll learn, Words to know, What will help)

Work to Submit:

- Mini-project on human body temperature dataset
- Mini-project on racial discrimination dataset
- Mini-project on hospital readmissions dataset

- Report on the inferential statistics methods used on the capstone project and its results

Foundations of Statistical Inference

The following resources from Khan Academy give you a solid foundation in statistical inference, which can be a bit dry, so please learn the material at your own pace and reach out to your TA with any questions. You may be tested on some of these concepts in your technical job interviews, so take the time to understand them thoroughly.

Random Variables

<https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library>

Sampling Distributions

<https://www.khanacademy.org/math/statistics-probability/sampling-distributions-library>

Course: One Sample Confidence Intervals

<https://www.khanacademy.org/math/statistics-probability/confidence-intervals-one-sample>

Course: One Sample Significance Tests

<https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample>

Course: Two sample inference for the difference between groups

<https://www.khanacademy.org/math/statistics-probability/significance-tests-confidence-intervals-two-samples>

Course: Inference for categorical data

<https://www.khanacademy.org/math/statistics-probability/inference-categorical-data-chi-square-tests>

DataCamp:

Interactive Exercises: Statistical Thinking in Python (Part 1)

Interactive Exercises: Statistical Thinking in Python (Part 2)

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 walk you through how hypothesis testing can be used to elicit insights from data and create good data stories. You'll use concepts that you learned from both the Exploratory Data Analysis units (Units 7 and 8).

The three mini-projects in this unit have similar structure, but require different techniques. You may need to perform some cleaning and wrangling of the data and then perform some visual analysis, along with some statistical tests to answer the posed problem. Finally, you'll put everything together into a coherent story

summarizing your approach and conclusion. At the end of these mini-projects, you should have a clearer idea of how various visualization and statistical techniques can work together to create a data story.

Submit your results using the links below and discuss them with your mentor on the next call. Remember, if you're feeling stuck, you can always reach out to your course TA for feedback on technical questions and code reviews.

1

Project: Analyze Human Body Temperature Using EDA

 2 - 4 Hours

In this exercise, you'll analyze a dataset of human body temperatures and employ the concepts of hypothesis testing, confidence intervals, and statistical significance. View the data and instructions in the linked Jupyter notebook.

Project: Examine Racial Discrimination Using EDA

2 - 4 Hours

In this exercise, you'll perform a statistical analysis to establish whether race has a significant impact on the rate of callbacks for resumes. View the data and instructions in the linked Jupyter Notebook.

Project: Reduce Hospital Readmissions Using EDA

 2 - 4 Hours

In this exercise, you'll critique a preliminary analysis of data and recommendations for reducing hospital readmissions rates and construct a statistically sound analysis to make recommendations. View the data

and instructions in the linked Jupyter Notebook.

[Download the Mini-Project Evaluation Rubric here.](#)

[Download project file\(s\) here.](#)

A/B Testing

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

<https://www.shopify.com/blog/12385217-the-beginners-guide-to-simple-a-b-testing>

Article: **3 Real-Life Examples of Incredibly Successful A/B Tests**
Written by [Robin Johnson](#)

Apply Inferential Statistics

Now that we've learned the basics of inferential statistics and hypothesis testing, let's apply your knowledge to the capstone project.

1

Project: Capstone Project 1: Exploratory Data Analysis



4 - 12 Hours

At this point, you've obtained the dataset for your capstone project, cleaned, and wrangled it into a form that's ready for analysis. It's now time to apply the inferential statistics techniques you've learned to explore the data.

Based on your dataset, the questions that interest you, and the results of the visualization techniques that you used previously, you might end up using only a few of the inferential techniques that you've learned. Your specific situation determines how much time it'll take you to complete this project. Talk to your mentor to determine the most appropriate approach to take for your project. You may find yourself revisiting the analytical framework that you first used to develop your proposal questions. It's fine to refine your questions more as you get deeper into your data and find interesting patterns and answers. Remember to stay in touch with your mentor to remain focused on the scope of your project

Think of the following questions and apply them to your dataset:

- Are there variables that are particularly significant in terms of explaining the answer to your project question?
- Are there strong correlations between pairs of independent variables or between an independent and a dependent variable?
- What are the most appropriate tests to use to analyse these relationships?

Submission: Write a 1-2 page report on the steps and findings of your inferential statistical analysis. Upload this report to your GitHub and submit a link. Eventually, this report can be incorporated into your milestone report.

Project: Capstone Project 1: Milestone Report



10 - 20 Hours

Think of a milestone report as an interim report that you may be asked to share with your client to keep them updated on your findings. It's also an opportunity for you to take stock of how far you've come, what you've found, and practice your data storytelling skills. This is similar to an early draft of the final capstone project 1 report.

The milestone report compiles all the reports that you've been writing throughout the course. Hopefully, you've been keeping your findings organized and documenting in a systematic manner. You should not need to do any new data analysis for this report.

Steps:

1. Write a your capstone project 1 milestone report (Google Doc, 5-6 pages) and include the following:
 1. Problem statement: Why it's a useful question to answer and for whom (get this from your proposal)
 2. Description of the dataset, how you obtained, cleaned, and wrangled it (get this from your data wrangling report)
 3. Initial findings from exploratory analysis (get this from your data story and inferential statistics reports)
 1. Summary of findings
 2. Visuals and statistics to support findings
2. Update your presentation slides.
3. Update your GitHub repository with the capstone project 1 code, milestone report, document, and slides .
4. Use the link below to share your report with your mentor for feedback, and update as needed.
5. Convert to .pdf and add to your repository. Share with your peer community

Wrap-Up: Inferential Statistics

Practice Interview Questions

Inferential Statistics

- What descriptive measure in statistics do you know?
- What's the difference between the mean, the median and the mode?
- What is the difference between Type I and Type II error?
- What is a p-value, and what is it used for?
- What are the null and alternative hypothesis in a statistical test?
- What is a t-test? Do you know what is its relationship to the z-test?
- What is the power of a statistical test?
- What is the standard deviation?
- What is a confidence interval, and what is it used for?
- What is bootstrapping/resampling used for?
- Do you know the difference between frequentist and Bayesian statistics?
- What are outliers? How can you check for them?
- What is a correlation coefficient? What range of values can it take?
- What is the Central Limit Theorem?
- What is the normal distribution? How can you test if data is normally distributed?
- Can you explain the major types of plots - histogram, bar chart, box plot, and scatter plot?
- What is a correlation matrix?

Recap: Inferential Statistics

In this unit, you **learned** the foundations of statistical Inference and hypothesis testing, basics of A/B Testing, and how Python can be utilized in Inferential statistics. You **explored** this information through three data analysis projects and **incorporated** your skills with inferential statistics into your capstone project.

You've addressed core elements of one of the four major themes: Exploratory Data Analysis. Congratulations on making significant progress on your journey to becoming a data scientist!

If you have any open questions on inferential statistics, especially on how you apply this unit's content to your capstone project, speak to your mentor during your next discussion.

We'll continue to refer back to this topic in future units, so it's a great time to get feedback.

Also, take a moment to "mark as complete" all the resources that you've worked through to help us track your progress and support you. Your capstone project is a central asset in your

professional portfolio and a core part of demonstrating your skill set as a data scientist during your career search.

In the next unit, we'll continue to help you develop your job search strategy by tackling **informational interviews**.

Informational Interviews

Overview

Informational interviews are one of the best ways to quickly gather information (beyond what's available through your internet searches) about a company. This unit delves into improving social skills, participating in an informational interview, and developing long-lasting professional relationships

Unit Plan (What you'll learn, Words to know, What will help)

Work to Submit:

- Short summary of your Capstone project
- Summary write up about your practice informational interview session

Let's start gathering information about informational interviews and helpful social skills to jumpstart your career search!

Machine Learning

Overview

This unit covers core machine learning (ML) techniques. It explores **supervised and unsupervised learning algorithms** and best practices for applying machine learning. The curriculum in this unit represents a balance between technical rigor and practical applications with ML techniques that are most widely used today.

You'll learn a number of machine learning techniques in this unit and apply them to mini-projects for practice. You'll also apply some of them to your capstone project. Keep your data stories in mind while you build your predictive models. Working through this unit will also help you develop a better sense of the learning track you want to choose.

Work to Submit:

- Linear Regression using Boston housing data set
- Heights and weights using Logistic Regression
- Predictive movie ratings from reviews using naive bayes
- Customer segmentation using clustering

Linear and Logistic Regression

Imagine you have some initial data that's labeled "True/False" or "Spam/Not Spam" and you want to extract "features" from the data that, when passed through a function, generate the labels as accurately as possible. To find this function, you'd use a

classification algorithm to automatically generate labels for those that don't have one.

In this unit, you'll learn classification algorithms.

To get started with machine learning, you'll learn about regression, a technique to predict unknown values when the values are real numbers. For example, you'd use regression to predict the amount of time a customer spends on a website, given data about characteristics and behavior of past customers. In this module, you'll study the simplest regression approach, linear regression, through a Harvard University course. View the presentation slides [here](#).

Please pay close attention both to the different types of bias that can arise, which are discussed during the first 15 minutes of the talk and to the derivation of the linear model, which starts at 50:00.

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=afe70053-b8b7-43d3-9c2f-f482f479baf7>

Video: Regression (continued)

<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, one of the simplest approaches to classification. For example, given data about characteristics and behavior of past customers, you might use classification to predict whether a customer will actually make a purchase, a binary outcome instead of a real number. This Harvard University lecture covers concepts that are critical both to your understanding of machine learning as well as relevant to job interviews. Please pay close attention to concepts such as collinearity (15:00), odds ratios (25:00), Curse of Dimensionality (40:00), and Lasso vs. Ridge regularization (1:00:00). View the presentation slides [here](#).

Questions related to these topics come up frequently in interviews, so make sure to understand them well and discuss with your mentor if you have further questions.

Video: Classification, kNN, Cross-validation, Dimensionality Reduction

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=c322c0d5-9cf9-4deb-b59f-d6741064ba8a>

Datacamp:

Interactive Exercises: Supervised Learning with Scikit-Learn

Project: Linear Regression Using Boston Housing Data Set

Instructions:

Please download and open the zipped file and work in the Jupyter Notebook in the unzipped directory. In the notebook, the phrase “Your turn” indicates sections where you need to fill in the code. After you’re done, please add the entire directory on your GitHub and submit a link to the completed Jupyter Notebook.

Project: Heights and Weights Using Logistic Regression

Instructions:

Please download and open the zipped file and work in the Jupyter Notebook in the unzipped directory. In the notebook, the phrase “Your turn” indicates sections where you need to fill in the code. After you're done, please add the entire directory on your GitHub and submit a link to the completed Jupyter Notebook.

SVM and Trees

The algorithms that we’ve studied so far are only the simplest ones in machine learning. The algorithms assume that the dataset they work on has a relatively straightforward structure. For example, both linear and logistic regression assume that data is mostly described by drawing a straight line. But, what if that’s not true?

Video: SVM and Evaluation

SVM and Trees

The algorithms that we've studied so far are only the simplest ones in machine learning. The algorithms assume that the dataset they work on has a relatively straightforward structure. For example, both linear and logistic regression assume that data is mostly described by drawing a straight line. But, what if that's not true?

Video: SVM and Evaluation

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=f21fcc8f-93a8-49f6-9ff8-0f339b0728bd>

Video: Decision Trees

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=8892a8b7-25eb-4bc5-80b6-47b9cf681a05>

Video: Using Random Forests in Python

Video: Ensemble Methods

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=4831ebf0-7832-42c5-9339-5b5e08dd3e92>

Article: Gradient Boosting from Scratch

<https://medium.com/mlreview/gradient-boosting-from-scratch-1e317ae4587d>

Bayesian Methods and Text Data

Bayesian methods are a powerful suite of techniques that are gaining traction in the world of data science. Unlike most other classification methods, which are *discriminative* (i.e. they give you a classification boundary), Bayesian methods are *generative* (i.e. they give you a model to generate the data, allowing you to infer

statistical properties of the data). In practice, Bayesian methods are often used in text analysis and spam/fraud detection.

Video: Bayes Theorem and Bayesian Methods

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=233f6c34-306f-481b-8ea5-be33076eb6a8>

Video: Sentiment Classification Using Scikit-Learn

Ryan Rosario, a Springboard mentor and Facebook Data Scientist, demonstrates practical text analysis and machine learning.

Project: Predicting Movie Ratings from Reviews Using Naive Bayes

Instructions:

Please download and open the zipped file and work in the Jupyter Notebook in the unzipped directory. In the notebook, the phrase "Your Turn" indicates sections where you need to fill in the code. After you're done, please add the entire directory on your github and submit a link to the completed Jupyter Notebook.

Best Practices

Bayesian methods are a powerful suite of techniques that are gaining traction in the world of data science. Unlike most other classification methods, which are discriminative (i.e. they give you a classification boundary), Bayesian methods are generative (i.e. they give you a model to generate the data, allowing you to infer statistical properties of the data). In practice, Bayesian methods are often used in text analysis and spam/fraud detection.

At this point, we've learned different techniques and methods, both for performing supervised learning as well as for evaluating machine learning models. How do we put them all together? What are some common tips and tricks for well-designed and effective models?

Some of the concepts in the following sections require a basic understanding of linear algebra. If you'd like a refresher on linear algebra, here's a [quick summary](#).

Video: Best Practices in Supervised Learning

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=b33eec92-d049-4353-a904-5054eb718aff>

Video: Best Practices (Continued)

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=afee45b9-dcf5-4f29-bc60-871aa78f1cf8>

Introduction to Unsupervised Learning

What do you do if you have data but no labels, and you want to find some structure in the data by defining your own classes? It's time for unsupervised learning!

Video: Clustering

AWS Clusters

- New and updated instructions for Spark 1.5 are on Piazza:

<https://piazza.com/class/icf0cypdc3243c?cid=1369>

DataCamp:

Interactive Exercises: Unsupervised Learning in Python

Scikit

https://www.youtube.com/watch?v=HjAB45qsx_c

Capstone Project Report

You've learned several techniques to use supervised and unsupervised learning to help build your predictive models. As you work through the interactive learning resources and mini-projects, apply these techniques to your capstone project.

1

Project: Capstone Project 1: In-Depth Analysis



20 - 30 Hours

The techniques you'll use in this project depend on your dataset so if you are unsure, ask your mentor to guide you on selecting the most appropriate methods.

Because each student's individual dataset and project needs determine which techniques are used, we've found that the amount of time a student spends on this section varies quite a bit. Don't be surprised if you find yourself spending more or less than the estimated time. Stay in touch with your mentor to ensure that you're on track and defining a reasonable scope of work.

Write up your findings and use the link below to share your progress with your mentor.

Your project will be evaluated using this [rubric](#).

Video: Effective Presentations

CS109 Video

Project: Capstone Project 1: Final Project

You're now ready to finalize and complete your first capstone project! By this point, you should have a pretty clear idea of the main points that you'd like to make and the predictive models appropriate to your story.

There's often no clear point at which you can say, "I'm done." It's likely that there are still questions and refinements to be made. This is a good time to discuss a stopping point for your first capstone project with your mentor. It's a good idea to list the open questions and refinements as future projects or further work to be done. The goal is to get a good capstone project with sufficient breadth and depth to demonstrate your newly acquired skill set.

Steps:

1. Put together all of your project components and review them with your mentor. Your report should include:
 - a. Code for the project
 - b. Presentation slides deck (and an optional blog post if you want to really stand out)
 - c. Consolidated report (Google Doc, 10-12 pages) based on summary reports from all your previous project submissions, including:
 - i. Proposal with problem statement
 - ii. Data collection and wrangling summary
 - iii. Exploratory data analysis summary (visualization and inferential statistics)
 - iv. Results and In-depth analysis using machine learning
2. Incorporate any revisions and feedback from your mentor.

3. If you'd like, you may also write a blog post to submit with your capstone project report and presentation.
4. Submit your final capstone project 1 for evaluation to your mentor, and share with your peers.

If you would like to get your code reviewed before you submit your final project, you can email your TA at projects@springboard.com

You've reached an important milestone in your learning, and your capstone project will be the centerpiece of your portfolio when you interview for jobs. Please review the submission expectations for the report, slides, and code in the capstone project guidelines and the evaluation rubric that your mentor will use.

Link: Interview Practice - Machine Learning

Practice Interview Questions

Machine Learning

Fundamentals

- Can you make the distinction between an algorithm and a model? *Model is a function representing a data set, algorithms are a way to obtain that function*
- What's the difference between supervised and unsupervised learning? *Labeled vs. unlabeled data*
- What's the difference between a regression and a classification problem? How about clustering?
- Why do we use a train/test split? *Avoid overfitting and better generalization*
- What is cross-validation used for? What types of cross-validation do you know?
- What is generalization error?
- What is the bias-variance trade off?
- What is the difference between overfit and underfit?
- **Video:** What accuracy metrics do you know, both for classification and for regression? When would you use one metric vs the other?
- What is the curse of dimensionality?
- Why do you need to set the random seed prior to running certain ML algorithms?

Regression

- Can you explain the difference between Linear and Logistic Regression?
- How are the coefficients in a Linear Regression interpreted?
- How is the intercept in a Linear Regression interpreted?
- Can the coefficients in a Logistic Regression be directly interpreted?
- What is the Adjusted R-Squared? What range of values can it take?
- Why is the Adjusted R-Squared a better measure than the regular R-Squared?
- How does Logistic Regression work "under the hood"? Can you explain Gradient Descent?

SVM

- Can you explain how a Support Vector Machine works?
- What is the kernel trick?
- Where does the “support vector” term come in the SVM name?
- What kind of kernels exist for SVMs?

Trees

- How do Decision Trees work?
- What criteria does a tree-based algorithm use to decide on a split?
- How does the Random Forest algorithm work? What are the sources of randomness?
- How is feature importance calculated by the Random Forest?

Other Supervised Learning

- What is the difference between Bagging and Boosting?
- How do Gradient Boosted Machines work?
- What is Regularization, and what types do you know? *Avoid overfitting, L1 and L2 regularization. L1 used as dim reduction, L2 better for overall generalization, bonus points for ElasticNet*
- Can a Random Forest and a GBM be easily parallelized? Why/why not?

Unsupervised Learning

- How does PCA work? What are the uses cases for it?
- How can you determine the optimal number of principal components?
- How does the K-Means algorithm work? What are its limitations?
- What other clustering algorithms do you know?
- How can you assess the quality of clustering?
- Can you explain in detail any other clustering algorithms besides K-Means?

Recap: Machine Learning

In this unit, you **learned** about the core techniques of machine learning and **explored** various supervised and unsupervised learning algorithms. You also discovered best practices for applying machine learning, the strengths and limitations of each, and how to evaluate their performance. You **incorporated** your knowledge through two interactive exercises and three projects, as well as by finalizing your capstone project.

The fundamentals of machine learning are a major milestone in the course, giving you a good balance of theory and application. **Congratulations** getting through this critically important theme of the data science course!

Note: This is the last unit focusing on the technical aspects of data science before you select a specialization. Your capstone project relates your specialization. Speak to your mentor about your experience with the first part of the course, your capstone project, and your interests to select the best specialization for you. You'll need to **choose a specialization** soon.

Next, you'll continue to work on your job search strategy by **identifying dream job titles and companies**.

Chapter 11

Find the Right Job Title and Companies

Overview

A dream job is the intersection of two phenomena: the right role at the right company. There are several job titles and a diverse array of companies in the data science domain. In this unit, you'll research roles within the data science domain that might be a good fit and identify industries and companies that offer those roles.

Unit Plan (What you'll learn, Words to know, What will help)

Work to Submit:

- A short paper describing at least three job titles that are right for you
- A list of 40-50 companies you'd like to join

The Right Job Title

This unit provides a step-by-step process to identify a few job titles in the data science field that you'd like to hold upon graduation.

Project: Find 2-3 Job Titles



2 - 6 Hours

Use the step-by-step process in the [presentation](#) to identify 2-3 job titles that you'd like to hold after graduation.

Steps:

1. Identify 5-10 job postings, focusing on just the role and job descriptions, not on the company. Be ambitious!
2. Put the text of the job postings through a word cloud generator. What comes up as common phrases?
3. Use these common phrases to search for more positions


Submit: A link to a 1 page (max) Google Doc, containing a short paragraph describing each of your 2 - 3 job titles and why you think these are the right fit for you.

The Right Companies

Now that you've squared away your job title, it's time to find your dream company. Many students that we've talked to feel that good data science jobs are only available in a few well-known companies. The fact is that data science is now crucial for companies of all sizes and industries. A thorough search, like the one we ask you to do in this section, will uncover many hidden gems.

Video: How to Find the Right Companies

Project: Identify 40-50 Dream Companies

 2 - 8 Hours

Using the process in the presentation, identify 40-50 companies you'd like to join. For every Tier 1 company, include 4-5 Tier 2 companies. Make a copy of this [Google Spreadsheet](#) track companies that you're researching. Feel free to modify it to your needs and track companies that you're researching on Tab 1.

Submit: A link to a Google Spreadsheet tracking the companies, with three columns: Tier 1 or 2, names of contacts, and research notes.

Link: Schedule a call with your Career Coach to review job titles and companies

Mark as complete



Schedule Call For Review

Wrap-Up: Find the Right Title and Companies

1

Recap: Find the Right Job Title and Companies

In this unit, you **learned** about the importance of focusing on specific job roles and companies as part of honing your job search strategy. You **explored** different roles and **incorporated** your work into a list of three positions and several target companies.

You'll choose your **Learning Tracks** in the next unit. Reach out to your mentor to talk through

your interests and determine which is the best path for you. You can also reach out to your student advisor for additional input.

The Generalist Track: Advanced topics in Data Science

Overview


This unit covers several advanced machine learning techniques and applications, including recommendation systems, anomaly detection, time series analysis and basic neural networks. The goal of this unit is to acquaint you with the basics of these applications, not necessarily to master them. As you work through the Generalist Track, you'll develop ideas for your second capstone project, helping you develop a project proposal by the end. Keep in mind how the new techniques introduced in this unit may apply to your second capstone project.

Second Capstone Project

You're now ready to develop your second capstone project. It's time to take the experience you've learned so far and exercise more independence in this project. The steps you need to follow for this project are the same ones you took to complete the previous capstone. You're responsible for progressing at a steady pace and getting the help you need. Speak to your mentor and agree on the scope, timeline, and deliverables. There are two mandatory check-ins or milestone reports before your final submission.

Start your second capstone project by selecting three ideas that allow you to showcase the advanced skills that you'll learn in this section. By the end of this unit, you'll narrow down your ideas and write a proposal.

Project: Capstone Project 2: Initial Project Ideas

 2 - 3 Hours

In this program, you'll continue building your technical skills and adding to your portfolio. The more full-length projects you have in your portfolio, the more impressive it looks to potential employers. Work with your mentor to pick a topic of choice for your second capstone project.

Note: To make sure you stay on track, review the [capstone project guidelines](#) and the [rubrics](#) which will be used to evaluate your project. which will be used to evaluate your project. Lean on your mentor during the entire project to understand and agree on what's expected and to incorporate feedback. You can also reach out to your course TA for intermediate code reviews or ask for help if you feel stuck.

Submission: Submit up to three ideas for your second project and review them with your mentor.

Steps:

1. Write a description of three capstone project ideas. Your ideas can be broad and high-level. The descriptions should address the problem and identify the data you'll use. You don't need to talk about specific methods and techniques.
2. Submit a Google Doc link through the space below and remember to enable sharing permissions to "comment." Please do not submit .pdfs, .ppts, or markdowns.
3. Review your ideas with your mentor during the next call.
4. Post your idea (title and description) to the student community for feedback.

Social Network Analysis

How do you know who the most influential people are in a group? What does information about your friends say about you? Which people in a group are most

likely to become friends (or date each other)? Social network analysis is a fascinating area of data science that has become critical in the world of apps and social media.

DataCamp

Interactive Exercises: Network Analysis in Python (Part 1)

Interactive Exercises: Network Analysis in Python (Part 2)

Recommendation Systems

Recommendation systems are everywhere, from Amazon's recommended books and products to Twitter recommending posts. But how do these commendations work? Let's find out!

<https://matterhorn.dce.harvard.edu/engage/player/watch.html?id=afee45b9-dcf5-4f29-bc60-871aa78f1cf8>

Video: Tutorial: Building a Recommendation System in Python

Mark as complete

Capstone Project 2: Project Proposal

1

Project: Capstone Project 2: Project Proposal



2 - 4 Hours

Finalize one capstone idea based on the feedback that you received on your initial ideas and your discussions with your mentor.

The proposal should address the following questions:

- What is the problem you want to solve?
- Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?
- What data are you using? How will you acquire the data?
- Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.
- What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Submission instructions:

1. Write your proposal in a Google Doc (1-2 pages) and submit the link via the "Submit" button. Make sure your mentor has permissions to comment on the document.
2. Work with your mentor to incorporate any feedback into later drafts and submit as many times as needed.
3. Once your mentor has approved your proposal, convert the doc to a PDF file.
4. Create a GitHub repository for this project (if you haven't done so already).
5. Add the PDF to your GitHub repository for this project.
6. Share the proposal with your peer community for feedback.

Note: All code and further documentation you write will be added to this repository.

Your project will be evaluated using this [rubric](#).

Problem statement

The recommendation problem in its most basic form is quite simple to define:

user_id, movie_id	m_1	m_2	m_3	m_4	m_5
u_1	?	?	4	?	1
u_2	3	?	?	2	2
u_3	3	?	?	?	?
u_4	?	1	2	1	1
u_5	?	?	?	?	?
u_6	2	?	2	?	?
u_7	?	?	?	?	?
u_8	3	1	5	?	?



Hot Topics in Machine Learning (Electives)

This section covers some of the hot topics in data science, including deep learning and natural language processing (NLP). Additionally, you'll learn about time series analysis and anomaly detection, both of which have become increasingly important in the world of constant, streaming data. All the learning resources in this unit are optional. They are selected to give you a sense of the approaches and commonly used techniques available in these specialized fields.

Datacamp”

Interactive Exercises: Deep Learning in Python OPTIONAL

Interactive Exercises: Natural Language Processing in Python

Article: Machine Learning Cheat Sheets

<https://startupsventurecapital.com/essential-cheat-sheets-for-machine-learning-and-deep-learning-researchers-efb6a8ebd2e5>

Video: Time Series Analysis with Python - Aileen Nielsen

n this PyCon 2017 tutorial, Aileen Nielsen walks through the fundamentals of time series analysis in Python. Starting with an introduction to the time series functions in Pandas, the tutorial covers the basic principles of this area and ends with forecasting using Autoregressive algorithms.

Advanced Data Visualization

You've learned about data storytelling and understand the importance of stories in conveying your insights to stakeholders, but how do you make stories compelling? How do you ensure that you really drive home your message? As in many other

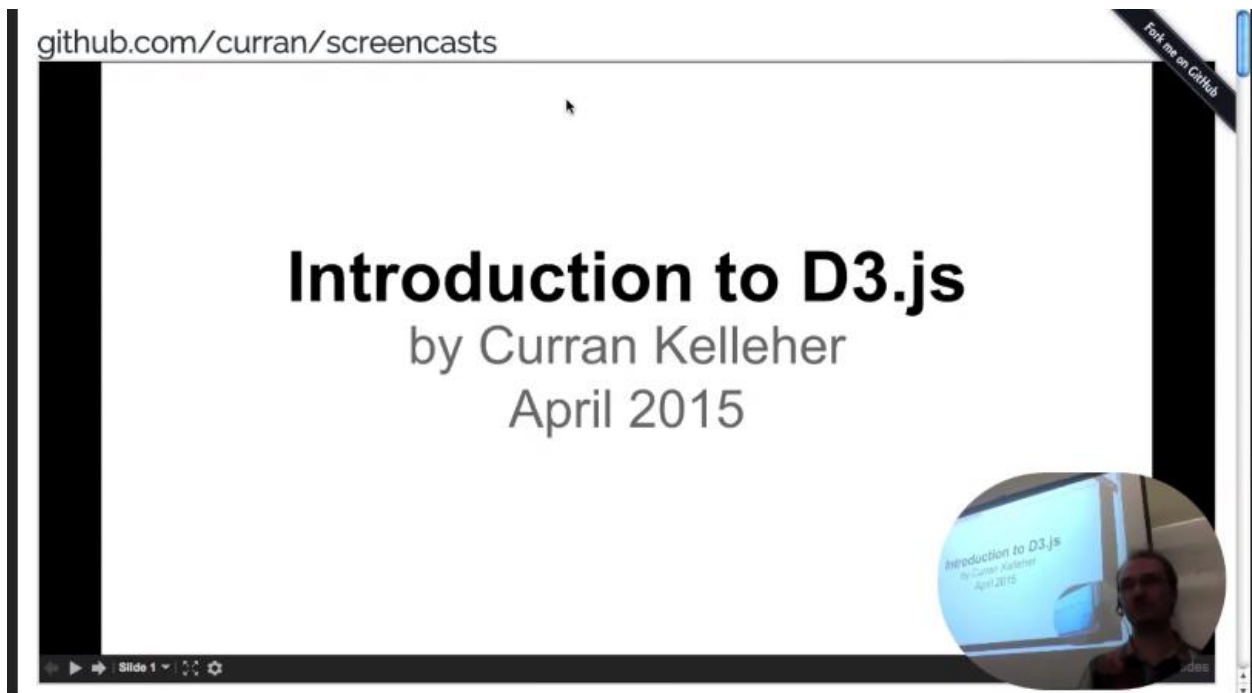
aspects of life, “a picture is worth a thousand words” in data science. Visual representations play a big role in how effectively information can be assimilated, so learning how to create engaging and meaningful visualizations is essential for a data scientist. Today, data visualization is almost a field on its own, with aspects of design, graphics, and aesthetics all merging together. For now, though, you’ll learn some important fundamentals. This unit introduces some advanced tools for creating interactive, web-friendly visualizations.

Datacamp:

Interactive Exercises: Interactive Data Visualizations with Bokeh

Video: Introduction to D3.js

This tutorial gives you a quick but thorough introduction to the foundations of D3.js, the most popular tool around for interactive data visualizations on the web.



Link: Plotly for Python

Wrap-Up: Advanced Topics in Data Science

Practice Interview Questions

Advanced Topics in Data Science

- What are stop words?
- Can you explain tokenization and lemmatization?
- Can you explain what is a Bag-of-Words?
- What is the advantage of using TF-IDF instead of simple word counts?
- Can you explain how word2vec works?
- What is topic modeling?
- Can you explain how Latent Dirichlet Allocation works?
- What ML algorithms are most commonly used for text classification tasks?
- How would you explain what a Recommender System is to a non-technical person?
- What types of Recommender Systems do you know? Can you explain the differences between them?
- What is the cold start problem?
- Where does the term “deep” come in the definition of Deep Learning?
- What is a perceptron?
- How does backpropagation work?
- How does Stochastic Gradient Descent differ from regular Gradient Descent?
- Do you know any other optimization algorithms?
- What types of layer do you know in Deep Neural Networks?
- What types of activation functions do you know?
- Why do Deep Neural Networks train faster on a GPU?
- How does a convolution work?
- Is time series modeling a regression or a classification problem?
- What is the difference between an AR and an MA component in a time series?
- How does ARIMA forecasting work? Why is the I term important?
- How can you ensure that you don’t overfit when building a time series model? Can you use regular Cross-Validation methods?
- What methods do you know for working with time series data in pandas? *Open-ended*
- How can you cluster time series data?

Recap: Advanced Topics in Data Science

In this unit, you **learned** about advanced topics in machine learning, including techniques and applications, and **explored** these areas in more detail, focusing on recommendation systems, anomaly detection, time series analysis, and basic neural networks. Through these topics, you should have an idea of how machine learning techniques can be adapted and used in a wide variety of domains. You **incorporated** your work through interactive exercises and began your 2nd Capstone Project.

You've now completed your specialization track and are ready to continue applying your new skills to your Capstone Project 2, while you continue to explore the four major themes of this course: Programming, Exploratory Data Analysis, Machine Learning, and Career Development.

Next you will focus on some **best practices** for working in your role as a data scientist with **software engineering** teams.

Software Engineering for Data Scientists

Overview

In this unit, you'll learn best practices to help you be a better engineer and to work more effectively with an engineering team. As a data scientist, no matter how many algorithms you design, how much data you crunch or charts you create, ultimately, you'll be writing software. Some companies expect their data scientists to contribute directly to the code base, others have engineers who are around to help translate prototype code to production. No matter which kind of team you're working with, it's critical to learn how to be a good citizen of the code base, so that you can make life easier for yourself and the rest of your team. The better your code is, the easier it is to deploy, and the greater the likelihood that you'll see your projects having an impact on the company!

Unit Plan (What you'll learn, Words to know, What will help)

Work to Submit:

- Short summary of your Capstone project
- Summary write up about your practice informational interview session

Let's begin learning how to write code that will be effective cross-functionally!

Write Better Code

In this section, you'll learn the best practices of software engineering that apply to data scientists, including how to write better tests, use libraries and APIs effectively to reduce duplication of work, and how to give better feedback to your colleagues about their code.

Video: So you want to be a Python expert?

<https://www.youtube.com/watch?v=nFVjLSvK5po>

<https://www.youtube.com/watch?v=cKPIPjyQrt4>

<https://www.youtube.com/watch?v=EKUy0TSLg04>

<https://www.youtube.com/watch?v=iNG1a--SIk>

<https://www.podcastinit.com/episode-111-cauldron-notebook-with-scott-ernst/>

<https://www.youtube.com/watch?v=FxSsnHeWQBY>

<https://www.youtube.com/watch?v=04paHt9xG9U>

https://www.youtube.com/watch?v=yACtdj1_IxE

Working with Production Systems

Building machine learning models in a Jupyter Notebook is one thing, but actually deploying those models in the real world requires some more work and skill. Knowing how to do that well will make you stand out to your team members and massively successful at your data science job.

<https://www.youtube.com/watch?v=f3I0izerPvc>

Interactive Exercises: Production Data Science using Git

This guide merges the gap that data scientists may have in software development practices. You'll look at the data science workflow in Python that adapts ideas from software development to ease collaborations and keep the project in a state that is easy to productionize.

<https://github.com/Satalia/production-data-science>

Video: From Model to Production Like a Pro

<https://www.youtube.com/watch?v=MKrPXfIWoc>

Wrap-Up: Software Engineering for Data Scientists

1

Recap: Software Engineering for Data Scientists

In this unit, you **learned** about the best practices in coding and working with a team in which you may hand off prototype code for production. You **explored** coding techniques to write better code, how to discuss code with your colleagues, the importance of testing and debugging and how to do each effectively, and tips and techniques for deploying machine learning models to production.

At this point, you have learned about best practices for data scientists to work cross functionally with software engineering teams. Start **incorporating** some of these practices in your second capstone project in order to see them in action. It will significantly improve the quality of your project as well as give you practice to hit the ground running in your job!

In the next unit, you will focus on **techniques for successful interviews** for your data science career.

Remember to work on your project in parallel with all the units from this point on.

Next

Preparing For and Getting Interviews

Overview

The Data Science Career Track Course has emphasized the importance of building a

professional network. This unit reviews the strategies of getting job interviews using these networks and tailoring your resume to fit a data science.

Unit Plan (What you'll learn, Words to know, What will help)

Work to Submit:

- Create a data science resume
- Get referrals to your target companies
- Write a cover letter

Create (or Update) Your Data Science Resume

Creddle.io

<https://will-stanton.com/2015/07/15/creating-a-great-data-science-resume/>

Project: Create a Data Science Resume Using Creddle.io



2 - 4 Hours

Creddle.io is a popular resume builder website that lets you create a resume that's both creative and professional.

Project: Update Your LinkedIn Profile



30 Minutes - 1 Hour

Take some time to update your LinkedIn profile to reflect accomplishments you've achieved and the skills you've learned in this course.

Submit: After you update your profile with accomplishments and skills, submit a link to your updated LinkedIn profile to your career coach.

It will be evaluated using [this rubric](#).

During the next call, you and your career coach will review the changes.

Submit: Use the website to create a resume and submit the link for feedback.

It will be evaluated by your career coach using [this rubric](#)

Video: Ramit Sethi - The Briefcase Technique

<https://www.youtube.com/watch?v=3p28MFt8RBA>

Project: Get Referrals Into Your Target Companies



2 - 4 Hours

You have built a network with people who are excited to support and advocate for you in finding a job. Now it's time to get to work.

1. Identify 5-10 companies from your list that are in your network i.e. you either have a connection there, or someone who's connected to someone in your network. Focus on Tier 2 companies first, and apply to Tier 1 companies only when you've gotten a little bit of practice.
2. If you have a connection there, reach out to them and ask them for a referral

3. If you have a 2nd connection there, ask a mutual connection to recommend you to them.

Repeat steps 1-3 above till you have received 5 referrals.

Update your tracking spreadsheet to reflect whether which companies you've reached out to, and whether you got a response.

Create Your Cover Letter

While cover letters (or emails) are not always required, they're a good thing to have on hand. A good cover letter tells the story behind your resume and often makes your application stand out in a sea of bullet points. When in doubt, send a cover letter or email with your resume/application.

Note: Sometimes an application system may not allow a cover letter to be uploaded. In that case, simply send one via your connection in that company.

<https://www.themuse.com/advice/the-8-cover-letters-you-need-to-read-now>

Project: Create a Cover Letter for One Position



2 - 4 Hours

Using the articles in this section, pick one position that you're targeting and write a cover letter for it.

Submit a link to a Google Doc containing the cover letter.

Wrap-Up: Preparing For and Getting Interviews

Recap: Preparing for and Getting Interviews

In this unit, you learned about data science-focused aspects of key job search staples: resume and LinkedIn profile. You've explored what makes a resume and a professional profile stand out.

You are well on your way to completing the elements of the Career Development theme. The next focus in this area will be on interviewing.

At this stage in the course, your portfolio is filling up with new skills and experiences:

1. Key components ready to start searching for job interviews
2. A list of companies you would like to work for
3. Types of data science role you'd like to have
4. A strong data science resume
5. A list of new contacts from your network building activities (Meetups, informational interviews, and your LinkedIn professional network)
6. Technical skills, including improved skills in Python, machine learning, data visualization, and inferential statistics, which you're applying in your mini projects and capstone projects

You now have a resume that'll impress potential employers! At this point, please ensure that you have completed the following steps in the program:

1. Completed the first capstone project
2. Completed a proposal for the second project
3. Selected a specialization
4. Created a list of potential companies and had the list reviewed by your career coach
5. Scheduled a call with your career coach to review your resume

Please ensure that you've also checked off the 'Mark as Complete' box for every resource you've finished so far. This will let us know exactly how much progress you've made and ensure that you get the right support you need.

You should continue to develop your **second capstone project** in parallel. Spend some time focusing on it now, if you aren't already, and turn in your milestone report in the next unit.

Then, you'll return to enhancing your technical skills, including completing a course from Lynda.com. Please email support@springboard.com to request your Lynda.com login information so that you have access to the course at the start of the unit.

Chapter 16

Data Science at Scale

Overview

Increasingly, data scientists are expected to know the fundamentals of building web-scale, cloud-based applications. This unit teaches the fundamentals of Spark, the most popular tool used today to build data science distributed applications at scale. You'll also learn advanced topics in SQL for data scientists. Please review the Unit Plan's What Will Help section to ensure you're set up for success in this unit.

Advanced Data Wrangling for Large Datasets

When you start working with large datasets, the techniques and tools you've used so far for data wrangling may not handle that scale. For example, it might be impossible to load up your entire dataset in Pandas to look for missing values, because your computer may run out of memory. In this section, you'll learn some advanced tools and techniques to help you wrangle big data. You may not need to use it in your

second capstone project (depending on the size of your data set), but it'll certainly give you a leg up in your interviews and in the workplace.

1

Review SQL OPTIONAL

It's time to review your SQL basics. Please feel free to go back to the Data Wrangling Unit to review SQL using the Mode Analytics tutorials. In case you'd like something different, you can also do the following DataCamp resources:

Course: Advanced SQL for Data Scientists

<https://www.lynda.com/SQL-tutorials/Advanced-SQL-Data-Scientists/559183-2.html>

Video: Big Pandas

https://www.youtube.com/watch?v=YGk09nK_xnM

Datacamp:

Interactive Exercises: Introduction to Pyspark

Video: Introduction to Spark with Python - Orlando Karam

<https://www.youtube.com/watch?v=9xYfNznjCIE>

Video: Introduction to Machine Learning on Apache Spark MLlib (Cloudera)

<https://www.youtube.com/watch?v=qKYpMPPL-fo>

Project: MapReduce with Spark



3 - 6 Hours

This mini-project covers basic exercises in PySpark (Python's interface to Spark) and MapReduce. Your project will be evaluated using [this rubric](#).

Submit: Please submit a link to a GitHub repository containing your completed notebook. You'll have to install a JDK, Spark and PySpark for this assignment (instructions are in the notebook).

Capstone Project 2: Milestone Report 2

In the following sections, you'll continue to work on your capstone project 2 in parallel with the unit curriculum. As you work on your project, please stay in touch with your mentor and ensure your deliverables are submitted for regular feedback. By the time you submit a final report, everything in the project should've already been discussed and reviewed for feedback.

1

Project: Capstone Project 2: Milestone Report 2



4 - 8 Hours

For this submission, you'll utilize the work you've completed using machine learning and the more advanced machine learning or specialization specific techniques to complete your report. Once completed, you'll incorporate this report into your overall portfolio and ensure that it's organized to help you excel in interviews.

Your portfolio consists of all of your data science projects, including the code and documentation, usually in your GitHub account. Typically, a data scientist who looks at your portfolio wants to see evidence of both your technical and your communication skills. As part of your milestone report submission, ensure that your portfolio is clear and easy to navigate.

Here are a few tips that'll help your portfolio stand out:

1. Every project should be in a separate repository that is clearly named.
2. For each project:
 - Make sure you have a README page that provides an executive summary of the project (i.e. summarizes the problem, approach, and final results).
 - The README should include a list of the important files that the reader should view. The files should be clearly named and organized.
 - Clean and document your code so that your approach and methodology are clear to any technical reader. You don't need to document every line of your code but have comments or text explaining important decision points and why you chose them.
 - Include any other documents that you've created (e.g. reports or slides) in the same repository as the code. Make sure the README identifies them to the reader.
3. Ensure your portfolio is not cluttered with "junk" (i.e. repositories that are incomplete, irrelevant, or undocumented).

Overall, try to put yourself in the shoes of an experienced data scientist who has a limited amount of time to look at your portfolio. How can you ensure that you make it easy for them to get a good idea of your skills and abilities? Your mentor and the peer community should also be able to provide good feedback on your portfolio, so please use them as resources.

Submit: A link to your GitHub portfolio, which includes your capstone project milestone report.

You can reach out to your course TA for a code review at any time by emailing projects@springboard.com

Practice Interview Questions

Data Science at Scale

- What is MapReduce? What kind of problems is it used for?
- Do you know what the advantages are of Spark over MapReduce?
- Can you explain the difference between batch processing and stream processing?
- Do you know what is a Spark Driver?
- What is a Spark Executor?
- Can you explain the difference between a Spark RDD and a Spark DataFrame?
- What programming languages does Spark support with APIs?
- What Machine Learning libraries do you know for Spark?
- What is sparkContext?
- What is the difference between a Transformer and an Estimator?
- What is the Pipeline API used for?
- What is a Fully Connected Layer?
- What is a pooling layer?
- What are Recurrent Neural Networks?
- What sort of improvement does an LSTM bring over RNNs?
- Can you explain the TensorFlow programming model?
- What types of activation functions do you know?
- Why do Deep Neural Networks train faster on GPUs?

Recap: Data Science at Scale

In this unit, you **learned** the fundamentals of building web-scale, cloud-based applications.

You **explored** how Spark is utilized widely to build data science distributed applications at scale and **explored** advanced topics in SQL for data scientists. You incorporated your learning into an exercise: MapReduce with Spark.

At this point, you've completed almost all of the technical learning themes. Congratulations on your tremendous progress to date!

Next, you'll learn about **interviewing for a data science job** in preparation for putting all of your work and new skills into action.

