7.2 Apply Data Storytelling - In progress : /data\_story\_telling

8. Inferental Statistics- In progress : /Inferential Statistics

8.1 Basic Statistical Inference – In progress

<https://www.youtube.com/watch?time_continue=1822&v=5Vjrqnk7Igs>

[Allen Downey](http://www.allendowney.com/wp/), an amazing statistics educator, Professor at Olin College and author (Think Stats, Think Bayes, Think Python) delivered this workshop on inferential statistics for Python developers at PyCon 2015. As per the description: "Statistical inference is a fundamental tool in science and engineering, but it is often poorly understood.  This tutorial uses computational methods, including Monte Carlo simulation and resampling, to explore estimation, hypothesis testing and statistical modeling.  Attendees will develop understanding of statistical concepts and learn to use real data to answer relevant questions."

The workshop is highly interactive, using Jupyter notebooks, and meant to be hands-on. All of the code for the workshop is available in this [Github repository](https://github.com/allendowney/compstats).

8.2 A Deeper Diver into Hypothesis Testing with Khan Academy

8.2.1 One Sample Confidence Intervals: <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 as a whole? Confidence Intervals are one metric that statisticians often use in practice.

8.2.2 One Sample Significance Tests: <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 is just a fluke? Significance tests give us a systematic way to make that assessment.

8.2.3 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 involve comparing two samples to see if there is a significant difference between the two populations.

8.2.4 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 ways to test hypotheses about distributions of categorical data.

8.3 Exploratory Data Analysis Project: /Inferential Statistics

8.4 Regression and Correlation: <https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data>

This section covers various techniques to find relationships between multiple variables in a data set, including scatter plots, correlation coefficients, and simple linear regression.

8.5 A/B Testing

8.5.1 Beginers Guide to A/B Testing: <https://www.shopify.com/blog/12385217-the-beginners-guide-to-simple-a-b-testing>

A quick introduction to the basics of A/B testing.

8.5.2 A/B Testing with Websites: <https://blog.hubspot.com/marketing/a-b-testing-experiments-examples>

How do you actually apply the principles of A/B testing to websites?

8.5.3 Additional resources: Broader Applications: <https://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty/up-next>

<https://www.udacity.com/course/ab-testing--ud257>

8.6 Apply Inferential Statistics:

Project: Capstone Project 1 - Inferential Statistics

https://www.springboard.com/assets/assets/icon-clock.13397d8ee8aa861b48ee7037f14eebaa.png2 - 6 Hours

At this point, you have obtained the data set 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 have learned to explore the data. For example, 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?

**Submission:**Write a short report (1-2 pages) on the inferential statistics steps you performed and your findings. Check this report into your github and submit a link to it. Eventually, this report can be incorporated into your Milestone report.

9. Capstone Project 1- Milestone Report: \Capstone1

10. Effective Networking \EffectiveNetworking2

10.4 Project: Conduct Informational Interviews – In Progress

11. Machine Learning \MachineLearning

11.1 Linear and Logistic Regression – In Progress

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.

11.1.2 Regression

11.1.3 Classification, kNN, Cross-validation, Dimensionality Reduction

11.1.4 Project: Linear Regression Using Boston Housing Data Set: /MachineLearning/LinearRegressionProject

11.1.5 Project:Heights and Weights Using Logisting Regression: \MachineLearning\LogisticRegression

-Reading In Progress

11.2 SVM and Trees

11.2.1 SVM and Eval

11.2.2 Decision Trees and Random Forests

11.2.3 Elective: Ensemble Methods

11.3 Bayesian Methods and Text Data

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

11.3.1 Bayes Theorem and Bayesian Methods

11.3.2 Optional: Sentiment Classification using Scikit-Learn: <https://www.youtube.com/watch?v=y3ZTKFZ-1QQ>

### 11.3.3 ProjectProject: Predicting Movie Ratings from Reviews Using Naive Bayes

\MachineLearning\BayesianMethodsTextData

11.4 Best Practices

11.4.1 Best Practices in Supervised Learning: \MachineLearning\Best Practices

11.5 Introduction to 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).

11.5.1 Clustering

11.5.2 Tutorial: Clustering with scikit-learn : Optional

<https://www.youtube.com/watch?v=-J9ZICyev5E>

11.5.3 Project: Customer Segmentation Using Clustering: \MachineLearning\Clustering

>>>Check on Machine Learning Harvard Online Course

12. Capstone Project 1 – \Capstone1\Final Submission

14 Advanced Topics in Machine Learning : \AdvancedMachine

### Project: Submit 3 Ideas for Your Second Capstone Project

https://www.springboard.com/assets/assets/icon-clock.13397d8ee8aa861b48ee7037f14eebaa.png1 - 4 Hours

Congratulations, you've learned the fundamentals of Machine Learning and completed a Capstone project! In this program, we encourage you to continue building your technical skills and add to your portfolio. The more full-length projects you have in your portfolio, the more impressive it looks to potential employers. So it's time to work with your mentor and pick a topic of choice for your second Capstone Project.

**Submission:**Similar to your first Capstone Project, submit a few (up to 3) ideas for your second project and get your mentor to review them.

14.1 Recommendation Systems

Recommendation systems are everywhere; from Amazon recommending books and products, Netflix recommending movies, to Facebook and Twitter recommending posts that you might like. But how do they work? Let's find out!

14.1.1 Recommendation Systems

14.1.2 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 build one in Python? This Pydata hands-on tutorial walks through step by step to creating a basic recommendation system in Python. Please download the code and data and follow along. You'll be asked to submit your code at the end of this tutorial.

**Note:**The [code and exercises](https://www.springboard.com/archeio/download/fe04bb75b2ff411485f22c186bdfe607/) file: 1489104021\_pycon2015\_tutorial322

### 14.1.3 Project: Submit your solutions to the exercises from the PyData recommendation systems tutorial

https://www.springboard.com/assets/assets/icon-clock.13397d8ee8aa861b48ee7037f14eebaa.png5 Minutes

As part of the recommendation system tutorial that you just finished, you worked through several exercises on building a Python recommendation system. Please complete those exercises and submit a link to your completed notebook(s) on github.

14.2 Time Series Analysis

Time series analysis problems occur commonly in data science. For example, sales for a retail organization based on past sales, network traffic for a big IT company based on past traffic -- are all important problems to apply predictive analysis to.

14.2.1 Time Series Analysis with Python: <https://www.youtube.com/watch?v=JNfxr4BQrLk>

In this SciPy 2016 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.

Materials: <https://github.com/AileenNielsen/TimeSeriesAnalysisWithPython>

### 14.2.2 Project: Submit your completed notebooks for the Time Series Analysis tutorial

https://www.springboard.com/assets/assets/icon-clock.13397d8ee8aa861b48ee7037f14eebaa.png5 Minutes

Submit your completed notebooks from the tutorial by Aileen Nielsen that you just finished. You may submit it as a link to a github repository. You need to complete only the notebooks covered in the tutorial, the remaining are recommended but optional.

14.3 Miscellaneous topics

Besides recommendation systems and time series analysis, there are some other advanced machine learning topics that we'd like you to have a good grasp of. These include anomaly detection, neural networks and density-based clustering.

14.3.1 Anomaly Detection : <https://www.youtube.com/watch?list=PLnnr1O8OWc6b_KudrHI0Vcf6sEu6xs9Ky&v=DEdrSPCdnVE>

When you have a large and complex system, one way to find out if everything is working well is to look for anomalies. For example, credit card companies often find fraud by detecting unusual transactions. How does that actually work?

14.3.2 Modern NLP in Python: <https://www.youtube.com/watch?v=6zm9NC9uRkk>

Beyond the basic text analysis that we worked in while studying Bayesian Inference, Python has a vast toolset for natural language processing. While you don’t need to be an expert in computational linguistics, a basic awareness of NLP techniques is quite critical in this age of unstructured data. This PyData talk by Patrick Harrison covers the commonly used NLP techniques and tools in Python.

14.3.3 A Gentle Introduction to Neural Networks and making your own with Python: <https://www.youtube.com/watch?v=2sevic5Vy4E>

After decades of being in machine learning wilderness, Neural Networks have made a stunning comeback over the last few years in the form of Deep Learning and other technologies. While Deep Learning is a bit too advanced for this workshop, this PyData talk by Tariq Rashid covers the fundamentals of Neural Networks, and how to build one for yourself in Python.

Files: <https://github.com/makeyourownneuralnetwork/makeyourownneuralnetwork>

14.3.4 Density-based clustering methods in Python: <https://www.youtube.com/watch?v=5cOhL4B5waU>

Density-based clustering methods are based on the intuition that clusters are regions where many data points lie near each other, surrounded by regions without much data. This is a bit different from K-Means, which assumes that points close to each other are likely to be in the same cluster, but don’t take into account any information about the empty space between clusters. This PyData talk by Brian Kent covers common density-based methods such as DBSCAN.

Files here: <http://nbviewer.jupyter.org/github/papayawarrior/public_talks/blob/master/pydata_nyc_dbscan.ipynb>

<http://nbviewer.jupyter.org/github/papayawarrior/public_talks/blob/master/pydata_nyc_DeBaCl.ipynb>

### 14.3.5 Project: Submit Your Second Capstone Project Proposal

https://www.springboard.com/assets/assets/icon-clock.13397d8ee8aa861b48ee7037f14eebaa.png1 Hour

Finalize one Capstone idea for your second capstone project based on the feedback you got from your mentor and peers in the community. Submit a project proposal - a short (1-2 page) document that answers the following questions:

1. What is the problem you want to solve?
2. 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 otherwise?
3. What data are you going to use for this? How will you acquire this data?
4. In brief, outline your approach to solving this problem (knowing that this might change later).
5. What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

The proposal will be part of a github repository for your project. All code and further documentation you write will be added to this repository.

Once your mentor has approved your proposal, please share the github repository URL on the community and ask for feedback.

14.3.6 Machine Learning Cheat Sheets: <https://startupsventurecapital.com/essential-cheat-sheets-for-machine-learning-and-deep-learning-researchers-efb6a8ebd2e5>

16 Advanced Data Visualization

16.1 Getting started with bokeh: <https://www.youtube.com/watch?time_continue=32&v=XBiS0oBzX3o>

Bokeh is another up-and-coming Python library that allows you to create interactive web-friendly visualizations directly in Python (without relying on Javascript). This introduction gets you started with this powerful library.

Files: <https://github.com/bokeh/bokeh-notebooks/tree/master/tutorial>

16.2 Introduction to D3.js: <https://www.youtube.com/watch?v=8jvoTV54nXw>

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.

16.3 Optional: 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 keep up to date on the use cases, codes and exemplars on using d3.js delivered as a weekly email newsletter

17 Capstone2 Milestone

19 Data Science at Scale

19.1 Introduction to Spark with Python – Orlando Karam

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

In this PyCon 2015 tutorial, Orlando Karam covers the basics of writing spark programs in python (initially from the pyspark shell, later with independent applications). He also discusses some of the theory behind spark, and some performance considerations when using spark in a cluster.

files: <https://github.com/okaram/spark-pycon15>

19.2 Intro to Machine learning on Apache Spark MLlib (Cloudera) : <https://www.youtube.com/watch?v=qKYpMPPL-fo>

Juliet Hougland, a Senior Data Scientist at Cloudera, presents this tutorial on Spark MLlib, a library for performing machine learning and associated tasks on massive datasets. With MLlib, fitting a machine-learning model to a billion observations can take only a few lines of code, and leverage hundreds of machines. This talk will demonstrate how to use Spark MLlib to fit an ML model that can predict which customers of a telecommunications company are likely to stop using their service.

Files: <https://github.com/jhlch/ds-for-telco>

19.3 Project: MapReduce with Spark: files: 1504727178\_spark\_mini\_project\_v1

20 Capstone 2 Final

9 out of 38 Projects submitted

Data Camp

NYU

Tableau- web doc – up to date with the latest version:

The big book of dashboards

Tableau annual meetings in NY

Tutorials and network

Documentation of interface of Python and Tableau