Data Science Resources

This repo is intended to provide open source resources to facilitate learning or to point practicing/aspiring data scientists in the right direction. It also exists so that I can keep track of resources that are/were helpful to me and hopefully for you.

I aim to cover the full spectrum of data science and to hopefully include topics of data science that aren't either actively covered or easy to find in the open-source world. For instance, I haven't focused on in-depth machine learning theory since that is well covered. If you are looking for ML theory I would look to some of the online courses, books or bootcamps. There is a lot of theory information available online, some is linked lower on this page here, here and other info is available with many purchasable books.

Keep in mind that this is a constant work in progress. If you have anything to add, any feedback, or would like to be a contributor - please reach out. If there are any mistakes or typos, be patient with me, but please let me know.

Lastly, I would add that a large portion of data science is exploratory data analysis and properly cleaning your data to implement the tools and theory necessary to solve the problem at hand. For each problem there are many different ways and tools to execute a successful solution - if one method isn't working re-evaluate, re-work the problem, try another approach and/or reach out to the community for support. Good luck and I hope this repo helpful!

Table Of Contents

1. Data Science Getting Started
    - Data Science Courses
2. Data Pipeline & Tools
    - Python
    - Stats/Engineering Libraries
    - Databases/Frameworks
    - Data Acquisition
    - Processing & EDA
    - Machine Learning
    - Model Selection
    - Model Evaluation
    - Feature Engineering
    - Additional Tools or Processes
    - Data Visualization
    - ipython Notebook Tutorials
    - Data Sources
    - New Data Tools
3. Product
    - Product Metrics

    - Team Communication & Business Tools
    - Best Practices
4. Career Resources
    - Data Science Career Path
    - Types of Data Scientists
    - Data Science Applications/Use Cases
    - Data Science Websites/Books
    - Data Science Meetups in the Bay Area
    - Data Science Blogs
    - Design Blogs
    - Data Science Conferences
    - Data Science Presentations
    - Relevant Business Processes
5. Open Source Data Science Resources
    - Additional Open Source Content
    - Auxiliary Content & Apps
6. About Me

Section of the data pipeline & resources:

Data Science Getting Started

Data Science is a multidisciplinary field covering at the very minimum - statistics, programming, machine learning Drew Conway's venn diagram. These topics are covered throughout this repo. I personally find the best way to learn a topic is to get my hands dirty quickly - with that in mind I would probably get to work in python and then implement different tools or theory into my toolkit as I understand each element. If you haven't used python before I would strongly urge you to use the codecademy course to familiarize yourself with the content and how to program. Good luck and have fun.

Starting

- Data Science Pipeline - Detailed overview of data pipeline from MachineLearningMastery.com
- Intro to ipython - A curation of Ipython Notebooks great for introductory level to python, programming, comp sci, data science and other topics.
- How do I Become a Data Scientist? - Some more great starting points from William Chen.

Data Science Courses:

- Coursera - Data Science Specialization at Coursera - many other courses available as well.
- Udacity - Online MOOCs that are the Data Science related courses. by I
- Data Science Bootcamps - A collection of all bootcamps currently on the market as of April 5, 2014 by Ikechukwu Okonkwo.
- Coursera Machine Learning Course - Andrew Ng's pinnacle Machine Learning course.
- Edx - EDX courses related to data science.

Data Pipeline & Tools

Python

Python is my workhorse language specifically as it has many data science and statistic library, the ability to work in production environments, and work on other problems outside of data science. There are many other languages that could be useful but are not covered here: Julia, R, Cython, Pig, Scala, Java, etc.

- Python @ Codecademy - If you have never used Python, right this way..
- The Python Wiki - Good resource with lots of info about Python.
- Python for Data Science Tutorial - Kaggle - Stepping into Data Science with Kaggle and installing some libraries.
- Introduction to Data Processing with Python - Just as the name says - some introductory level information and exercises.
- Git tutorial - Git for Version Control. Simple tutorial for Git from Github.

Stats/Engineering Libraries

A collection of workhorse libraries that are elemental for any python data scientist. * Pandas Wes McKinney's pandas library for EDA on small to medium sized data sets when you don't want to put the infrastructure for SQL or when it isn't necessary. It has many other great applications other than just better than SQL on small to medium data sets. * Numpy/Pandas/Scipy Cheatsheet - self explanatory * SciPy - Open-source software for mathematics, science and engineering. * NumPy - Fundamental package for scientific computing with Python. * StatsModels - Module that allows users to explore data, estimate statistical models and perform statistical tests. * PyMC - Bayesian estimation useful for Markov chain Monte Carlo analysis (among other things). * Probalistic Programming and Bayesian Methods for Hackers - Github Repo all about the namesake. * The only probability Cheatsheet you'll ever need - Self explanatory - (thanks William Chen @ http://datastories.quora.com/) for pointing me this great cheat sheet out - wish I had that back at college.

Data Acquisition

Libraries that are very helpful for abstracting away some of the complications of scraping or working with HTTP. * BeautifulSoup - A python library to make web-scraping HTML easier. * Beautiful Soup Cheat Sheet * Requests - HTTP for Humans - python library that makes working with http and api's more effortless

Processing & Exploratory Data Analysis

A collection of documents explaining some of the ways to do processing & EDA. * Unix for Processing - sed & awk for data processing. * Pandas - Already mentioned is great for data processing - cleaning, filtering and getting rid of nan's, normalizing, scaling, replacing values, etc. * SciKit Learn for Preprocessing - Doc on sklearn's preprocessing methods. * Regular Expressions - Regex explained.

Databases/Frameworks

A collection of databases & frameworks that are helpful for data management and are the industry standard. * SQL - SQL Database - I linked to Postgres since that is the version I use. * Psycopg - Python <> Postgres. Able to adapt PostgreSQL for the python environment. * SQL Cheet Sheet * SQLZoo - Develop your skills * SQLSchool - Develop your skills MongoDB - NoSQL database * PyMongo - Python Mongo Driver. * MongoDB - cheatsheet - Cheat sheet for MongoDB * Apache Hive - Uses Hive Query Language (HQL) - similar to SQL for data at scale. * Hive Cheatsheet - Self Explanatory. * ElasticSearch - For scalable, fast text search/analysis. * Neo4j - Leading graph database. * Redis - Key-value open source data structure server. * Redshift - AWS petabyte-scale data warehouse solution. * Hadoop - the definitive guide - Hadoop ecosystem. * Spark - Lightening fast cluster computing.

Machine Learning

There is a lot of information available online about the theory, mathematical intuition, tuning for this discipline. I am not trying to cover it in that depth, at least not at this current time. These are some high level knowledge posts and toolkits. * SciKit-Learn - Simple and efficient machine learning tools for data mining and data analysis * NLTK - Natural Language Toolkit to work with human languages data. * Tour of Machine Learning Algorithms - Blog post about some of the high level ML methods * VIDEO - How to get started w/mL - Melanie Warrick @ PyCon 2014. * Some ML methods classified - Classification for some sample ML algorithms by Melanie Warrick. * SciKit-image - Algorithms for image processing. * Deeplearning4j - Deep Learning in Java. * Machine Learning CheatSheet - I would actually say this is more than just a cheat sheet given that there are > 100 pages of notes.

Model Selection

Resources about how to decide on your model. * SciKit Learn Flow Chart for Model Selection - A helpful for a starting point selecting SKlearn algorithms.

Model Evaluation

Resources to help with understanding model evaluation. * Evaluating ML Algorithms - Blog Post from MachineLearningMastery about how to evaluate your performance. * Cross-Validation - Critical concept to evaluate the performance of your models. * K-fold & Grid Search in Scikitlearn - Demo on how to implement kfold cross validation and grid-search using scikit-learn. * Scikit-learn Cross Validation doc - Self explanatory title.

Feature Engineering

A critical element of Data Science to improve your performance but minimally talked about. * Ipython Notebook for Feature engineering - Some discussion about Feature Engineering. * CS Princeton Course - Course content on Feature Engineering. * Blog Post about Feature Engineering / Data Exploration - Blog post about topic.

Additional Tools or Processes

Resources on other topics that are very helpful for data scientists and product. * A/B Testing - Blog about A/B testing. * A/B Testing - And how you are screwing it up. * Bloom Filters - Python notebook about bloom filters. * Bloom filters - Bloom Filters. * Reservoir Sampling - A primer on Reservoir Sampling. * Reservoir Sampling Again * Monte Carlo for the Monty Hall Problem - Hyon Chu puts on a good explanation to MC for the Monty Hall Problem. * Markov Chain Monte Carlo - Opening the black box of MCMC. * Multithreading and Queues - How to build multithreading and queues. * Basics of Multithreading and queses - More about multithreading. * Building a Recommender System - Quora answer to this question. Helpful starting point.

Data Visualization

Collection of the best libraries that I know for easy and powerful data visualizations. * ggplot - ggplot for python ported by the team at yhat. * matplotlib - Awesome plotting library for python. * d3 - Mike Bostock's viz library - the de facto gold standard for polished visualization - in js, steep learning curve but beautiful outcomes. * bokeh - Interactive visualization library. * d3py - Another library for data viz. * vincent - Help with python for d3. * seaborn - Clean statistical data visualization library.

Other available Visualization Resources. * Scott Murray's D3 Tutorials Tutorials from *Interactive Data Visualization for the Web* * tributary.io - live code visualization platform designed specifically for D3.js * plot.ly - A web visualization and data processing platform * blockspring - Share code and visualizations through a single platform * dot.append - Ian Johnson (enjalot) goes through several live-coding examples using D3

Design Theory

The importance of design theory in data visualization and presentations could not be understated. Through better understanding of design theory and principles, a data scientist can convey more information and meaning in their presentations. * Accelerating Understanding Through Data Visualization - Accenture White paper on Data Visualization

Ipython Notebook Tutorials

Collection of ipython notebooks that are helpful as examples to either using tools or to explain certain topics. * Pandas Tutorial - Basic intro to Pandas in notebook form. * Scipy Tutorial - Basic Scipy Tutorial. * Numpy Tutorial - Basic Numpy Tutorial. * Multiple Regressions using Statsmodels - Using statsmodels for regression. * Intro to PyMC - Intro to PyMC. * More on PyMC - More PyMC. * Kaggle Titanic Comp Tutorial - Kaggle Titanic Tutorial using RandomForests. * Psycopg2 tutorial in Python - How to use Psycopg2. * SQL in iPython - SQL in Python. * Mongo in Python - Mongo in Python. * Beautiful Soup Tutorial - Beautiful Soup! * Sci-Kit Learn Basics - Machine Learning Basics with scikit-learn. * MatPlotLib - Some of the possibilities of data-viz with MatPlotLib. * Choosing the right priors - Bayesian - Bayesian statistics and prior selection. * Some Basic Data Analysis in Python - Basic data analysis with python. * Crash Course in Python for Scientists - Ipython Notebook for

Scientists! * Regular Expressions - Regex to match patterns in strings - very powerful. * MapReduce - Classes, inheritance and map-reduce exercises. * Recursion Notebook visualization recursion "The single most powerful idea in algorithms". * Recursion More about Recursion and Functional Programming

Data Sources

Collection of sites to access data if you want to build out a project or just use some of the tools for EDA. * Data.Gov - The US government portal to open data. * California Water Resources - California's water resource data. * Data for Cool DS projects * Academic Torrents - Sharing Data is hard, torrents make it easier for academics. * Data Basin - Science based mapping and analytics platform. * Open Energy Data Initiative - Over 800 data sets covering energy issues. * UCI Machine Learning Datasets - Data for machine learning - lots of labeled data and description of the problem types.

New Data Tools

Aim to keep track of developing trends and new tech that is helpful for the practicing Data Scientist. New might be a misnomer. * BigML - machine learning for the everyday user, also useful for EDA. * GraphLab - graph-based, high performance, distributed computation framework. They just implemented deep learning onto their platform. * ModeAnalytics - platform to share analysis/data science. * Apache Mahout - Scalable machine learning library. Not in python. * Apache Hadoop - Open-source software for reliable, scalable, distributed computing.

Product

Product Metrics

Understanding product, user behavior, and product metrics is helpful for data scientists in industry. Being able to help your product manager and team execute on strategies by understanding the problem, metrics and what they understand facilitates a more fruitful relationship. * Actionable Metrics - Funnel reports, cohort analysis, actionable metrics. * Analytics for Product Managers - Everything a PM needs to know about analytics - or the minimum amount your PM should know about analytics as a Data Scientist. * Startups, you are doing data science wrong! - High level explanation about how to use data science in a start-up company. * Product Psychology - Understanding user behavior.

Team Communication & Business Tools

There are some very innovative new companies that are producing very effective tools to minimize and abstract away inefficient processes at companies. While it isn't strictly data science related, these products could be very help to integrate with your teams to improve overall productivity. * Aha! - Clean product roadmapping software for PMs. * Slack - Amazing team communication tool - abstracting away unnecessary e-mails. * Harvest - Effortless time tracking for business. * Trello - Helping organize everything - great for project management. * Zapier - Bringing together Harvest +

Slack + Trello and a lot more... * Thoughtbot Playbook - A detailed account of how thought book runs is software consulting company talking about guiding principles, design sprints, code reviews to sales and operations. A content packed post. * IFTTT - 'Putting the internet to work for you'. Great for small companies to automate social media, marketing or to have your own personal recipes set up. * Github - Clearly a great product - 'Build software better, together'. * Web Analytics & Reporting Software: * Google Analytics - In depth real-time analytics. * Mixpanel - provides real-time analytics and solid cohort analysis. * Clicky - Pride themselves on ease of use.

Best Practices

Source control and keeping accurate documentation so that you and your colleagues can follow and reproduce your work is very important. I will add some best coding practices & data science practices.
* Python Code Style - Allows for better understanding for everyone involved on the project.
* Slide Deck for BMPs - Slide deck about best practices for coding or the repo.
* Engineering Practices in Data Science A blog post about the lack of source control in Data Science. It's a challenging topic - I believe mode analytics is trying to solve it.

Career Resources

Data Science Career Path

- Data Science @ Google - Quora answer about Data Science career trajectory @ google.

Types of Data Scientists

Not all Data Scientists are the same and it's critical for organizations to understand what it is they need, and how best to fill those roles and/or complement the skills of their team. Finding the organizational structure that enables the data scientists/data engineers within the organization and generates better results is also crucial. It should be given thorough consideration. * Kind's of Data Scientist - O'Reilly's classification of 4 different data scientists. * Data Science For Startups - Which of the Five Types of DS does your startup need? Different classification from O'Reilly. * Building Data Science Teams - posted from 2011 about how to build data science teams. * Data Science Team Building - The Power of Collaborative Analytics - Post post about different team org structures, difference between DS & BI.

Data Science Applications/Use Cases

Data Science has so many different applications and use cases within industry - many are continuously discovered. These resources provide some potential ideas. * Kaggle Data Science Use Cases - Helpful to generate ideas for new uses in different industries * Data Science for each Industry - Description of uses for different industries. * Big Data Analytics News - use Cases - For Big Data but that's almost synonymous with Data Science.

Data Science Websites/Books

More resources for community based information or hard copy books. * Data Science Handbook - Not yet released but should be interesting providing stories from academia and industry about data science - go read the post for a better description! * CrossValidated - A question and answer site for people interested in statistics, machine learning, data analysis, data mining, and data visualization. * StackOverflow - Language-independent collaboratively edited question and answer site for programmers. * Kaggle - Model building competition and great resources for training and data. * O'Reilly Media - A lot of content rich books available and tutorials on using the tools. * Quora - Question and answer site - lots of data science content and career content.

Data Science Meetups in the Bay Area

A great way to meet other Data Scientists and keep up to date with best practices. * SF Data Science * Data Science for Sustainability * Python Meetup Group * USF Seminar Series in Analytics * DataKindSF * SF Bayarea Machine Learning * AirBnB Tech Talks

Data Science Blogs

The name say's it all. * Data Stories @ Quroa - William Chen's (DS@Quora) blog about data science. * FastML * FiveThirtyEight Blog - Nate Silver's blog. * Data Science Hanbook - Data Science Handbook Project (not quite a blog but it fits here). * Simply Statistics Blog * All The Things Tech * Musings in Data Science * Zipfian Data Science Blog - Zipfian Academy DS Blog. * Machine Learning Mastery * DataTau - Hackernews for Data Science. * HackerNews * Quora - Q&A site with lots of information about Data Science.

Design Blogs

- ThreeStoryBlog

Data Science Conferences

- Strata - Conference and a lot of videos from previous conferences - great resource.
- GraphLab - Another great conference.

Data Science Presentations

- Strata Collection of Presentations - Most of their conference presentations available online.

Relevant Business Processes

- Lean Startup - A method to develop product and businesses.
- Agile Development - group of software development methods to optimize for self-organizational and cross-functional teams.
- Scrum - an iterative and incremental agile software development framework for managing product development.

Open Source Data Science Resources

While the name might sound redundant this section represents other sites or repos that have aggregated information covering similar topics. Tons of great content on these sites - definitely go check them out.

Other Open Source Data Science Content

There are some really great resources linked within this section covering all of Data Science, the entire data pipeline, machine-learning, statistics, python, etc. Go check them out. * Open Data Science Masters - Clare Corthell's Open Source online blog/github with lots of resources available for data science. * A Practical Intro to Data Science - Zipfian Academy's collection of excellent resources available. * LearnDataScience - Nitin Borwankar's collection of IpythonNotebooks for Linear Regression, Logistic Regression, Random Forests, K-Means Clustering * FreeDataScienceBooks - Yu Wu's free open sourced online data science books. * Gallery of Ipython Notebooks - iPython's introduction to Python, Data Science, Economics, Comp Sci, Linguistics, and much more. * Data Science 45 Min Intros - The team @ Gnip have a collection of repos to introduce data science topics in roughly 45 minutes per topic. * Awesome Data Science - Collection of bloggers, twitter accounts, facebook accounts, MOOC's, datasets, tools. * Awesome Big Data - Onur Akpolat's curated list of awesome big data frameworks, resources and papers.

Auxiliary Content & Apps

- Markable - Let's me visualize my Markdown
  - Markdown Cheatsheet - Self explanatory.
- LightPaper - Markdown editor that I use.
- iterm2 - Terminal application for Mac.
- Oh My Zsh - Framework for managing your ZSH config. Awesome.
- Sublime Text Editor - For all your scripting needs.

ABOUT ME

I acquired my skills through programming in an on-the-job environment and then taking three months off to learn and put into practice my data science skills @ Zipfian Academy. For me taking that time off to learn, run the daily/weekly sprints, and be in a collective learning environment at Zipfian was irreplaceable. Even if Zipfian resources were open source, without taking the time off work and having the drive to learn all the necessary material would be next to impossible. I am always interested to hear what other data scientists are doing and using for tools. I am interested in a wide range of different open source &/or private projects - feel free to reach out on Twitter @sf_oak or LinkedIn. Or go check out my start-up venture capitalist recommender ~ finding the long-tail of the VC community at findyourvc.co.