A picture containing logo

Description automatically generated

**Data Science Crash Course**

**Gabe Ketron**

**[Ask questions or connect with**

**me on** [**LinkedIn**](https://www.linkedin.com/in/gabriel-ketron/) **or** [**Twitter**](https://twitter.com/GabeKetron)**]**

**Chapter 1 – Overview & Getting Started**

Hey all!

By popular demand y’all asked me to put together a quick explainer of the different Data Science tools we’ll be using on the project and put out on the internet to introduce folks who are new to get started quickly.

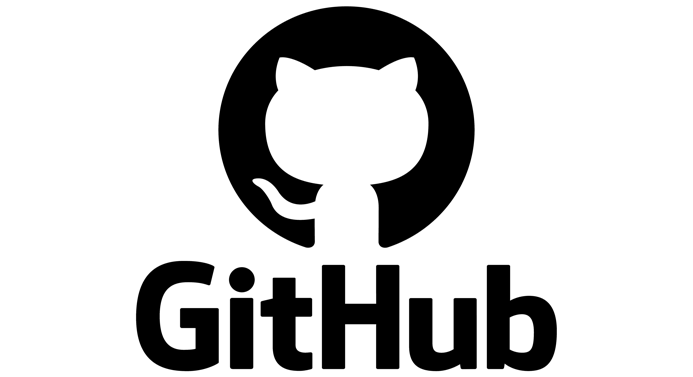
Graphical user interface

Description automatically generated

Courtesy of u/instituteprograms on r/datascience

This flow chart gives an over-arching idea of what goes into data science, and I’ll quickly highlight the minimum to get started and what we still need to dive deeper into. To get started with the bare minimum we are going to want to cover all the items in the “Python” subheading in the “Programming” category; Basics, NumPy, Pandas, Matplotlib, Seaborne, and Scikit-Learn.

Basics and NumPy cover our basic operations and equations. Pandas allows us to set up matrix math and eventually input into a database. Matplotlib and Seaborne allow us to visualize our data and results in Pandas dataframes. Lastly, Scikit-Learn will give us the tools to do machine learning with Pandas dataframes.



To get started we are going to also cover Git and Jupyter Notebooks. Git is critical to keeping record of your work, and eventually making it accessible to the public by showing your code, data, and results on Github. Jupyter Notebooks are a popular tool for data science project, since it functions as an Integrated Development Environment for a variety of coding languages and has markdown functions in the same workspace. This lets you write code, show result plots and write explanations of your results all in the same file. Jupyter notebooks can also integrate with your Github account to automatically save your progress while working on your projects.

For the next meeting, set up an account on github.com, and also install Python with Anaconda.

Anaconda is a distribution system that allows you to run python and jupyter notebooks from your computer’s terminal, and download a wide array of libraries that we will use during the project.

The process of setting up anaconda is a little more intensive than an account on github, so I will lay out the steps to get our development environment up and running.

1. Go to [www.anaconda.com](http://www.anaconda.com/) .

Graphical user interface, text, application

Description automatically generated

1. Download the 64-bit Python 3 version of Anaconda for your operating system.
2. Follow the default instruction options except if you are using Windows, you need to select the advanced option to add anaconda to your path variable so you can create virtual environments later if needed, and the default option.
3. Once the setup is complete, you can open up Jupyter Notebooks from the list of available IDEs in Anaconda

Graphical user interface, application

Description automatically generated

I hope this helps as a quick overview of what we need to cover. Over the course of the project we will also need to familiarize ourselves with machine learning techniques, and some of the ways to make our findings available to the public, whether through GitHub, a cloud service, a website or a formal publication.

Next week I’ll put up an explainer on how to use Git, how to set it to work with Jupyter notebooks, and some basics to refresh your knowledge of Python. See you then!

Gabe Ketron