## DATA SCIENCE COURSE PREWORK (~10 HOURS):

Congratulations on joining us at General Assembly for the Data Science course! We are so excited to have you and looking forward to working with you all soon!

In order to best prepare you for this course, we've gathered a list of resources that we'd like for your to complete prior to beginning the course, so when you arrive on your first day, we will all be up to speed and be familiar with the vocabulary and programming languages necessary to succeed in this course. The prework is organized into three sections: Github Account, Installation, and Familiarization. Github Account and Installation should take no more than an hour. Familiarization could take much much longer, so you must budget your time effectively and balance coverage with depth.

Also, we will be holding an "Installfest" day prior at General Assembly prior to the first day of the course in order to help with any remaining installation issues.

Throughout this course, we will be using the command line interface (CLI) for programming, as do professional data scientists. So, it is critical that you familiarize yourself with working from the command line before class starts. A great resource for learning the command line is The CLI Crash Course at learncodethehardway.org. The online HTML version is free.

Our primary programming language for this course is Python 2.7 (Note: There is a Python 3, which has significant differences from 2.7 and is not "industry standard." We will be using Python 2.7). We will use a number of Python packages such as Numpy and Scikit-learn in our work implementing machine learning algorithms. Also, we will be using SQL (Structured Query Language) to work with data in structured databases. Please work through the information below to prepare yourself to work with Python and SQL in class.

## Github Account

We will be using Git and Github heavily throughout the course. Please be sure to create your Github account before the first day of class.

### GitHub

GitHub is the most popular place for keeping most open source and many closed source projects on the web. It is a good place to keep your homework because a) like Dropbox and other cloud storage services, it will keep a backup and provides version control (so you can look at previous versions of what you did) and b) it can serve as a public portfolio of your work. Go to https://github.com/ and create an account. Then, take 15 minutes and complete the free "Try Git" tutorial on CodeSchool. Note that Git and Github are not the same thing, but are tightly integrated.

## Installation (< 1 hour)

1. Verify SQLite installation and install if necessary
2. Install the Anaconda distribution of Python and many of the libraries we will be using
3. Install Sublime Text (optional but highly recommended)

### SQLite

SQL is the basis for any inquiry into large structured data. Rumors of its demise have been greatly exaggerated and it remains the gold standard. There are many flavors of SQL out there, but for this course we will keep things simple and start with SQLite. If you own a Mac, chances are, sqlite3 is already installed. You can test this by opening a Terminal window. Terminal is under Applications → Utilities → Terminal. You'll want to put Terminal in your dock as you'll be using it a lot. In an open Terminal window, type this at the prompt:

```
sqlite3
```

You should see something like:

```
SQLite version 3.7.13 2012-07-17 17:46:21
Enter ". help" for instructions
Enter SQL statements terminated with a ";"
sqlite>
```

If so, enter:

```
.quit
```

and move on to the next section. If, on the other hand, you see something like:

```
-bash: sqlite3: command not found
```

you will need to go to http://www.sqlite.org/download.html, find the precompiled binary of the command-line shell for your computer, download and install it.

### Python & Packages

Fortunately, a company called Continuum has released Anaconda, which handles all of the installation and package dependencies for us (including installing Numpy, SciPy, Sckit-learn, and many other fun libraries). Find and read the installation instructions for your computer (e.g. “Mac Install”) on http://docs.continuum.io/anaconda/install.html. Then, download the installer appropriate for your computer from http://continuum.io/downloads.html . Run it.

From http://docs.continuum.io/anaconda/install.html :

> Due to a bug in the Mac OS X installer software, you may see a screen that says “You cannot install Anaconda in this location. The Anaconda installer does not allow its software to be installed here.” To fix this, just click the “Install for me only” button with the house icon, and the installation will work again.

Test that Anaconda & Python were installed correctly by opening a Terminal window and entering:

```
ipython notebook
```

in a few moments, your browser should open to a window titled `IPython Dashboard' with two tabs: Notebooks and Clusters. If this works, you may close your browser, return to the terminal window, and enter control-C to shutdown the notebook server (enter `y' at the prompt). If this does not work, email me or the TA for help. Remember, it is important that this be working *before* the first day of class.

### Sublime Text (Optional but highly recommended)

Sublime Text has become a very popular text editor among programmers. You can download it at http://www.sublimetext.com/. Unlike the rest of the software presented here, it is ‘nagware,’ meaning that it is free to use indefinitely, but will nag you periodically until you decide to pay for it.

## Familiarization (~ 8-9 hours)

If you are unfamiliar with any of the tools listed above, then it is highly recommended you at least go through the first few steps of the corresponding tutorial listed below. The list below is in priority order. Focus on items 1 through 4 first, then 5 and 6 if you have time.

Some of these are longer than others and it is not expected that you complete them all in their entirety. It *is* however expected that you have at least a passing familiarity with each of them before class begins. Extra time spent on these tutorials will not be wasted.

1. **Command Line Interface (Bash on OS X) —** http://cli.learncodethehardway.org
2. **Try Git --** http://trygit.codeschool.com
3. **Python —** http://www.learnpython.org/ **and** http://learnpythonthehardway.org/
4. **SQL —** http://sqlzoo.net **(Just focus on the Tutorial Section, items 1 through 8)**
5. **linear regression —** https://www.khanacademy.org/math/probability/regression
6. **Pandas —** https://bitbucket.org/hrojas/learn-pandas