Software Installation

In order to get the most out of the short session, you will need to follow along in the coding. Please install the following software packages.

1. **Anaconda for your OS**

Download Anaconda from [here](https://www.anaconda.com/distribution/). Please ensure you install the Python 3.7 version.

Anaconda will enable us to use Python and essential Python packages for data science

1. **Atom Text Editor**

Download and install Atom from[**here**](https://atom.io/)

1. **Apache Spark: Pyspark**

**Pyspark installation for Windows OS**

* 1. Ensure Anaconda is installed and that Python is running okay
  2. Download and install Java JDK from [here](https://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html)
* First, you can check if you have java 8 jdk installed by checking java version in CMD. Type "java --version" to check.
* If java is not installed, install java 8 sdk by downloading from Oracle website indicated above
* In some cases, java needs to be installed in a folder with a folder name without spaces
* Please take note of the folder where java 8 sdk is installed (e.g., C:\java\)
  1. Download apache spark from [here](https://www.apache.org/dyn/closer.lua/spark/spark-2.4.3/spark-2.4.3-bin-hadoop2.7.tgz)
* Extract into local drive and rename it to “spark”
* Download [winutils](https://github.com/steveloughran/winutils/blob/master/hadoop-2.7.1/bin/winutils.exe) and move it to “spark/bin/”
* Also take note of “spark” path
  1. Create windows environment variables as follows
* HADOOP\_HOME: set this variable to location of spark (e.g., C:\spark)
* SPARK\_HOME: set this variable to location of spark
* JAVA\_HOME: set this to location java (e.g., C:\java\)
* Add “spark\bin” to PATH variable
  1. Test Pyspark
* Go to Anaconda terminal and type: pyspark
* Run this code
  + nums = sc.parallelize([1,2,3,4])
  + nums.map(lambda x: x\*x).collect()
* If everything installed correctly, the code above should run without errors
  1. Install “findspark”. In order to use Pyspark with Jupyter notebook, install findspark as follows: in anaconda terminal, type “pip install findspark”

**Apache Spark on Mac OS and Linux based OS**

* For Mac OS, please follow instructions on this [blog](https://medium.com/@yajieli/installing-spark-pyspark-on-mac-and-fix-of-some-common-errors-355a9050f735)
* For Linux OS, follow this [tutorial](https://www.tutorialkart.com/apache-spark/install-latest-apache-spark-on-ubuntu-16/)