# BECOMING A DATA SCIENTIST - (WORK IN PROGRESS)

March 30, 2022

## Contents

# 1 Introduction

## 1.1 Primary Objective

- Prepare learners for internship role in Data Science & Machine Learning within 6-8 months.
- Prepare learners for Full-Time role as professional Data Scientist within 1.5 years.

## 1.2 Structure

Critical courses and practical projects are emphasized. We will emphasize projects and team work once basic course work is completed with the goal of preparing learners for real world problems in industry. Topics such as cloud computing, programming fundamentals, mathematical foundations are treated with an equal level of importance.
**NOTE**: This document is meant to be used as a guide. Not all courses and linked videos are needed nor meant to be covered. For beginners it is highly advised to seek the guidance of an expert or working data science professional prior to beginning the courses. This is mean to be a fully self contained guide, however it is not perfect, and sometimes assumed prior knowledge might have been overlooked (although I try hard to avoid this).

## 1.3 Computer Setup

Please follow instructions in appendix: Computer Environment Preparation and Software Installation for Scientific Computing

# 2 First Block

## 2.1 Computer Science

### 2.1.1 Introduction to Computer Science

Video lectures: Introduction to Computer Science and Programming (6.00)

### 2.1.2 Structure and Interpretation of Computer Programs

Video lectures:Structure and Interpretation of Computer Programs (6.001)

### 2.1.3 Introduction To Algorithms

Video lectures: Introduction to Algorithms (6.006)

Alternative Introduction to Computation Theory

## 2.2 Math and Stats

### 2.2.1 Linear Algebra

Video lectures: Linear Algebra (18.06)

### 2.2.2 Statistics and Probability

Video lectures: Introduction to Probability and Statistics - A (UCI)
Video lectures: Introduction to Probability and Statistics - B (UCI)
Video lectures: Probability (Harvard - Stats110)

Alternative (lecture notes and materials) - Introduction to Probability and Statistics (18.05)

## 2.3 Misc + Additional Resources

### 2.3.1 Git Fundamentals

Git Tutorial for Beginners
– Book: Pro Git 2nd Edition

### 2.3.2 Regular Expressions

Onur's Regex Intro Videos and Thorough Jupyter Notebook Tutorials

### 2.3.3 Unix/Linux Navigation and Shell Scripting

Onur's Linux/Unix Navigation and Shell Scripting Intro Videos and Thorough Jupyter Notebook Tutorials

# 3 Second Block

## 3.1 Computer Science

### 3.1.1 Introduction to Computational Thinking and Data Science

Video lectures: Introduction to Computational Thinking and Data Science (6.0002)
Additional or Alternative Videos: Introduction to Data Science in Python (UMich Coursera)

### 3.1.2 Introduction to Computational Thinking (Julia Based Computational/Applied Math Class)

Video lectures: Introduction to Computational Thinking (18-s191)

## 3.2 Math and Stats

### 3.2.1 Math For CS

Video lectures: Mathematics for Computer Science (6.042)
Video Lectures: Discrete Mathematical Structures (Clemson MATH-4190)

### 3.2.2 Probability Theory

Lecture Notes: Probability and Random Variables (18.440)

## 3.3 Misc + Additional Resources

### 3.3.1 Introduction to Databases

Video Lectures: Introduction to Databases – SQL, Database Design, Database Connectivity
Single Video Lecture: RDF, SPARQL and Entities

### 3.3.2 Data Visualization in Python

Applied Plotting, Charting  Data Representation in Python (UMich Coursera)

### 3.3.3 Reading Material

3 Big Problems with Datasets in AI and Machine Learning

# 4 Project 2

Resume building project

# 5 Third Block

## 5.1 Computer Science

### 5.1.1 Introduction to Statistical Learning and ML (Less Technical)

Video Lectured: Introduction to Machine Learning (MIT 6.036)
Video Lectures: Learning From Data (CalTech Short Course)
Alternative or Aditional lectures: Introduction to Machine Learning and AI (DeepMind)

### 5.1.2 Machine Learning and Applied ML (More Technical)

Video Lectures: Applied Machine Learning in Python (UMich Coursera)
Video Lectures: Applied Machine Learning (in Python) (Columbia - W4995)
– Course Materials: Course Materials
Supplemental Lectures: Mathematics of Big Data and Machine Learning

## 5.2 Math and Stats

### 5.2.1 Applied Stats

Video lectures: Statistics for Applications(18.650)

### 5.2.2 Applied Probability

Video lectures: Probabilistic Systems Analysis and Applied Probability (6.041sc)

### 5.2.3 Information Theory

Video lectures: Introduction to Information Theory
Video Lectures: Information Theory, Pattern Recognition, and Neural Networks (Cambridge)

## 5.3 Misc + Additional Resources

### 5.3.1 Cloud Computing and Data/Web Mining

- Onur's online video lectures – AWS Cloud computing
- Onur's online video lectures – Data Retrieval from the Web

# 6 Project 3

Resume building project.

# 7 Forth Block

## 7.1 Computer Science

### 7.1.1 Deep Learning

Video Lectures: Introduction to Deep Learning (MIT 6.S191)
Video lectures: Deep Learning (CMU: 11-785)

## 7.2 Math and Stats

### 7.2.1 Advanced Linear Algebra

Video lectures: Advanced Linear Algebra (Graduate level - Clemson 8530)
Course Materials: Advanced Linear Algebra Course Page

### 7.2.2 Applied Linear Algebra

Video lectures: Matrix Methods in Data Analysis, Signal Processing, and Machine Learning (18.065)

### 7.2.3 Computational Science and Engineering I (aka Mathematical Methods for Engineers I)

Video lectures: Computational Science and Engineering I (18.085)

### 7.2.4 Computational Science and Engineering II (aka Mathematical Methods for Engineers II)

Video lectures: Computational Science and Engineering II (18.086)

# 8 Project 4

Resume building project. (Propose Final All Encompassing Project based on all Skills Learned Over the Past 1-1.5 Years)

# 9 Fifth Block - 'Electives'

## 9.1 Deep Learning Block

- Advanced Deep Learning (Yann LeCun)
- NLP with Deep Learning (Stanford)

## 9.2 ML and Graphs

- Machine Learning on Graphs (All Video Lectures - UPENN)
- Machine Learning with Graphs (CS:224W CMU)

## 9.3 NLP Block

- Intro to NLP (Stanford)
- NLP with Deep Learning (Stanford)

## 9.4 Time Series and Finance Block

- Signal and Systems (6.003)
- Topics in Mathematics with Applications in Finance (18.s096)

## 9.5 Business Analytics and Data Science Block

- Introduction to Analytics (15.071)
- Data Science for Business
- Social Media Data Analytics
- Designing, Running, and Analyzing Experiments
  – Experiment Design Concepts in a Simple A/B Test

## 9.6 ML For Health Sciences Block

- Machine Learning For Healthcare (6.s897)
- Computational Systems Biology: Deep Learning in the Life Sciences
- Biomedical Image Analysis in Python

## 9.7 Network Science Block

- Networks: Friends, Money, and Bytes (Princeton - Coursera)
- Network Science
- Network Science Book and Course Notes
- Machine Learning with Graphs (CS:224W CMU)

## 9.8 Image and Video Processing & Computer Vision Block

- Many resources and lecture course links
- Biomedical Image Analysis in Python
- Accelerated Computer Vision: A Free Course From Amazon

### 9.9 Machine Learning For Physics and Physical Sciences Block

- Machine Learning For Physicists
- Advanced Machine Learning for Physics, Science, and Artificial Scientific Discovery

### 9.10 Reinforcement Learning Block

- Deep Reinforcement Learnin (CS:285 CMU)
- Reinforcement Learning (Deepmind)
- Reinforcement Learning - ICTP (QLS-RL)

### 9.11 Recommendation Engine Block

- Introduction to Recommender Systems: Non-Personalized and Content-Based
- Nearest Neighbor Collaborative Filtering
- Recommender Systems: Evaluation and Metrics
- Matrix Factorization and Advacned Techniques

# 10 Project 5 - Special Project/Capstone

(Project to be based on the Elective Specialization Track chosen by Student) Resume building project.

# 11 Appendix

## A Jupyter Tutorials

- My Instructional Jupyter Notebooks

## B Onur's Video Lecture's and Workshops

- My Video Lecture's and Workshops

## C Onur's Topic Specific Mentee Crash/Mini Courses and Tutorial Videos

- Intro to Spark and Distributed Big Data Processing – Mini Course
- Intro Linux/Unix OS and Shell Scripting Deep Dive – Mini Course
- Crawlers and Scraping Deep Dive – Tutorial Series
- Basics of Data Acquisition and Sanitization – Tutorial Series
- Introduction to Outliers and Anomalous Data –Mini Course
- Monte Carlo Method and Simulation: A Deep Dive – Mini Couse
- AWS: Basics of EC2, S3, RDS, Neptune, EMR, Lambda, Dynamo DB, EMR, and Sage Maker – Mini Course
- Regular Expression: From 0 to Hero – Tutorial Series
- Version Controls – What Are they and why are they important? – Tutorial Series
- Time Series Analysis – Mini Course
- Data Visualization and Story Telling with Data for Python Deep Dive – Mini Course
- Introduction TO NLP in Python (mostly spacy and hugging face)– Mini Course
- Creating Interactive Data Science Dashboard Applications with Streamlit – Tutorial Series
- Introduction to the GDELT Project – Big Data Anyone? – Tutorial Series

## D Online Coding Practice

- HackerRank
- LeetCode
- My Job/Interview Prep Jupyter Notebooks

## E Addtional Learning Materials

### E.1 Core Data Science and ML Subjects

- Colab Tutorial
- Python Numpy Tutorial (with Jupyter and Colab)
- Introduction to Pandas for Deep Learning | Part of Larger Lecture Series
- Python for Data Analysis: Pandas  NumPy
- Sk-Learn Short Course
- Learning Pytorch with Examples
- Deep Learning from Scratch with PyTorch Tutorial | 3hr Video
- Machine Learning with TensorFlow  Scikit-learn | Video Lectures

### E.2 NLP

- Advanced Spacy Course
- Hugging Face Course

### E.3 CV

- Open CV Tutorials
- Vision Transformer ViT Tutorial

### E.4 Misc

- BigQuery in Colab
- GDELT 2.0: Our Global World in Realtime (I love this)
- GDELT Visual
- AlgoExpert
- Mathematical Methods of Physics (GaTech)
- Chaos and Nonlinear Dynamics (GaTech)
- Group Theory (GaTech)
- Python + AWS Lambda (Video Series and Tutorial Deployments)

# F Computer Environment Preparation and Software Installation for Scientific Computing

## F.1 Introduction

First and foremost make sure you computer is fully updated in terms of software and you operating system version. Check for updates, and make sure you installed and suggested updated.

## F.2 3rd Party Software

### F.2.1 Text Editors

Get simple **text editor** – Light weight quick and simple code modification/editing, inspection, and creation. I suggest downloading/installing BBEdit, or SublimeText. These are much better than than Macos build in text editor because of formatting issues.
NOTE: Due to serious formatting issues, you never want to use MS Word or things of that nature when writing or analyzing code.

### F.2.2 Integrated Development Environments

Download an **IDE** (Integrate Development Environment). There ew of the popular ones, and PyCharm is the one I use. Other popular IDEs include (there are many more, but no need to explore them):

1. Spyder
2. VSCode
3. Eclipse with PyDev
4. LeoEditor

### F.2.3 Terminal Alternative

For Mac users, I highly recommend downloading ITerm2. iTerm2 is a replacement for Terminal and the successor to iTerm. It works on Macs with macOS 10.14 or newer. iTerm2 brings the terminal into the modern age with features you never knew you always wanted. Why Do I Want It? Check out the impressive features and screenshots. If you spend a lot of time in a terminal, then you'll appreciate all the little things that add up to a lot. It is free software and you can find the source code on Github.

### F.2.4 XCode (Macusers Only)

For Max users we will need XCode. Alternatively you can sign in to your Apple account and download directly from apple.com.

### F.2.5 Command Line Tools

Once you've install XCode you need to install Command Line Tools. You have two options to do so:
**Option 1:** - Install through XCode Application

1. Start Xcode on the Mac.
2. Choose Preferences from the Xcode menu.
3. In the General panel, click Downloads.
4. On the Downloads window, choose the Components tab.
5. Click the Install button next to Command Line Tools.

6. You are asked for your Apple Developer login during the install process.(if you don't already have account create one, it's free)

**Option 2:** - Install through the Web (You can download the Xcode command line tools directly from the developer portal as a .dmg file)

1. On the Mac, go to Apple Developers Page
2. You are asked for your Apple Developer login during the install process. (if you don't already have account create one, it's free)
3. On the "Downloads for Apple Developers" list, select the Command Line Tools entry that you want.
4. Once downloaded, find the .dmg file, double click it, and follow the installation instructions.

#### F.2.6 Homebrew (Mac Only)

Homebrew is a powerful packmanager for Unix (i.e. MacOSC) and Linux operating systems. Homebrew installs the stuff you need that Apple (or your Linux system) didn't. To install, open a Terminal (iTerm2) window and copy and paste the following line and press 'return':
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/install.sh)"

### F.3 Installing Python

#### F.3.1 For Macs

Follow the directions Here to install Python 3 on you Mac or Linux machine

#### F.3.2 Windows

The easiest way to get the most up to date version on python on your Windows machine (as far as I remember) is installing and downloading the software found Here.

### F.4 Jupyter Notebook and Jupyterlab

#### F.4.1 Installing Jupyter

1. step 1 blah...

#### F.4.2 Launching Jupyter Locally

### F.5 Virtual Environments

Using pyenv will be very beneficial.

## G Version Control

### G.1 Free Online Version Control Options

Visit HERE for a list of common/popular free version control tools (we'll use git)

## G.2 Git Setup

### G.2.1 Github Registration

– Visit Github.com and register for a free account.

– Alternative: Bitbucket

## G.3 Setting Up Your Secure Connection

Visit: Generating a new SSH key and adding it to the ssh-agent

# H AWS Setup

## H.1 Setting up AWS account

Register for a free AWS. account Here. It requires a credit card number, but don't worry, unless you decide to use services which I will explicity tell you not use, you will be eligible for the 'Free Tier' account.

## H.2 AWS CLI

Adding Private Credentials to simply access to S3 and other services.

## H.3 Creating and EC2 Instance and Accessing Remote Machine

Visit: Introduction to AWS EC2 tutorial.

## H.4 PuTTy (Windows Users Only)

You will need a Linux machine to access you cloud environment. So for Windows users you will need the PuTTy emulator to do so.

## H.5 Accessing a Remote Jupyter Instance on Local Browser

**Creating Secure Socket - Tunnel to AWS Instance**

- In the remote machine terminal:
  1. Activate torch env on aws DEEP LEARNING AMI (if desired):
     source activate pytorch_latest_p37
  2. Only need to do this step once per instance:
     - cd
     - mkdir ssl
     - cd ssl
     - openssl req -x509 -nodes -days 365 -newkey rsa:2048 -keyout mykey.key -out mycert.pem (might need chmod )
  3. jupyter notebook −−certfile=∼/ssl/mycert.pem −−keyfile ∼/ssl/mykey.key

- On your local machine terminal:
  ssh -i ∼/.ssh/some.pem -N -f -L 8888:localhost:8888 ec2-user@ec2-3-91-63-179.compute-1.amazonaws.com

- Access remote jupyter notebook:
  go to your browser, and type the following address https://localhost:8888
- When finished:
  - shut down remote server: open the remote terminal, and press:
    CONTROL C
  - open the local terminal, enter the following then hit RETURN:
    lsof -ti:8888 | xargs kill -9