USF MSDS Spring Reading List 2024

This document contains the spring reading list for the 2024-2025 cohort. We highly recommend that you read this list carefully and work on the self-identified weak areas BEFORE the beginning of the boot camp. The boot camp will have a fast pace and high intensity and doesn't leave much time or chance for you to catch up on the basic knowledge of each subject. Therefore, having good preparation beforehand is essential to success in boot camp.

This document is organized into three sections:

I. Linear Algebra
II. Probability and Statistics
III. Computer Language and Computer Tools, and Recordings of Summer 2023 MSDS 501 Computing Lectures

Linear Algebra

We strongly encourage you to finish the Linear Algebra Exam BEFORE the beginning of the boot camp so that you can focus on the three courses during boot camp.

Below is a superset of good linear algebra textbooks for review. In the linear algebra self-paced videos and exam, the instructor will draw from a combination of these books with emphasis from (1) and (5):

1. *Matrix Analysis and Applied Linear Algebra* by Carl D. Meyer
2. *Introduction to Linear Algebra* by Gilbert Strang
3. *Applied Linear Algebra and Matrix Analysis* by Thomas S. Shores
4. *Numerical Linear Algebra* by Lloyd N. Trefethen and David Bau
5. *Elementary Linear Algebra* by Anton, 11th edition, Wiley

For the time being, you can rely on your linear algebra book from college, as well as the free linear algebra book by Jim Hefferson at:

http://joshua.smcvt.edu/linearalgebra/book.pdf

In your initial review, focus on the following topics: *vectors, matrices, and associated operations, solving linear equations, determinants, vector spaces, eigenvalues and eigenvectors, and linear transformations.* Time permitting you should engage in any problems associated with computation and implementation of all these topics.

If you are looking for an online learning venue, consider taking the OCW Scholar course in linear algebra at the Massachusetts Institute of Technology at:

http://ocw.mit.edu/courses/mathematics/18-06sc-linear-algebra-fall-2011/

For further reading, Ian GoodFellow's chapter on Linear Algebra is terse, but focused on preparation for deep learning:

https://www.deeplearningbook.org/contents/linear_algebra.html

If you want a video connecting linear algebra and python data science problems,

enjoy:

https://pyvideo.org/europython-2018/from-linear-algebra-to-machine-learning.html

Probability and Statistics

In the probability and statistics boot camp (MSDS 504), the instructor will use a combination of the following books:

1. Wackerly, Dennis D., Mendenhall III, William and Scheaffer, Richard L. *Mathematical Statistics with Applications*, 7th edition.
2. Hogg, Robert V., McKean, Joseph W. and Craig, Allen T. *Introduction to Mathematical Statistics*, 8th edition.

Note that these books will not be required for the class, but free pdf versions of them can be found online.

As you review, focus on the following learning outcomes:

- Understanding the definitions of probability mass functions, probability density functions, and cumulative distribution functions;
- Knowing the properties of the most famous examples of random variables (Bernoulli, binomial, geometric, exponential, Poisson, normal, etc.);
- Mastering the underpinnings of the most common parameter estimation technique, maximum likelihood estimation;
- Understanding the difference between a sample and a population;
- Being able to state the Central Limit Theorem, understanding its importance, and applying it in a variety of basic situations;
- Understanding all elementary one- and two- sample tests of hypotheses and confidence interval constructions (e.g., means and proportions);
- Understanding the fundamental axioms, rules, and laws of probability theory;
- Defining, and working with examples related to, conditional probability;
- Understanding the importance of the concept of independence;
- Using the Law of Total Probability to prove Bayes' Theorem and deploying Bayes' Theorem in a variety of practical situations;
- Working with random vectors as well as random variables;
- Working with multivariate distributions, as well as the concepts of conditional expectation and independence in a high-dimensional setting;
- Understanding the difference between parametric and nonparametric statistics; and
- Understanding the bootstrap

If you are looking for an online and non-credit bearing opportunities to review this material, there are several you might consider. We recommend the first two courses in University of Michigan's "Statistics with Python Specialization" at Coursera. We also recommend Berkeley's three-part introduction to statistics -- addressing probability, descriptive statistics, and inferential statistics (at EdX). However, previous program participants have indicated that these learning experiences are not as rigorous as our program's boot camp review of probability and statistics.

Consequently, we are also placing onto the Canvas site for admitted students Shan Wang's class notes from a previous instance of MSDS 504. We recommend that you go through those notes quickly, marking any concepts that seem less familiar to you.

Then go back through those notes a second time, reading those areas that you marked more carefully. Note that this is just for reference to the topics and materials, and the exact lectures for the coming cohort might be different.

Computer Programming Languages and Tools

a. Programming

The computation boot camp is a “review” course, and instructors assume that you have already learned the concepts while you were taking pre-requisite courses. Those topics include the following subjects.

- Basic Computer Architecture
- Terminal and Shell Commands
- Version Control & Git
- Unit Test (pytest)
- Code Standards
- Python Programming
    - Debugging and Error Handling
    - Conditional Statements
    - Loop
    - File I/O
    - Data Aliasing
    - Function
    - Packages, Libraries and Modules
    - Object Oriented Programming

When students encounter challenges with the programming assignments, we often inquire about their preparation leading up to the boot camp. It is commonly observed that insufficient focus on programming practice is a recurring issue. Each year, a few students face difficulties passing the computational boot camp and may need to withdraw from the program. To assist you in better preparing for success, we have developed the following guide.

The mastery of concepts in this program primarily occurs through hands-on coding. Proficiency in coding and utilizing programming tools directly impacts your ability to navigate the curriculum effectively.

How to learn

Many of you may have completed programming courses, whether in-person or online, but may not have fully grasped the concepts. To truly master coding, there is no substitute for actively engaging in writing code to solve problems. Merely observing the instructor write programs is not sufficient; you must actively write the code yourself. Just as you wouldn't learn to play a musical instrument by solely listening to music, the same applies to coding.

Do not copy and paste code while learning. Manually typing code is part of the learning process. You might make small typos which will help hone your debugging skills and overall coding writing abilities.

Also, write code in an interactive environment, see Jupyter Notebook section below, so you get immediate feedback about what works and what does not work. Later we'll write scripts in .py files and run the scripts at the command line. For now, writing scripts will slow down your learning curve.

One of the best ways of studying is to sit in front of a blank screen with a coding prompt for a problem you have already solved. Solve the problem from memory without looking at your previous solution or the internet. Only use those resources when you are completely stuck. Drilling skills from memory will reinforce what you have already learned.

One of our favorite tools is **Python Tutor,** http://pythontutor.com/. It visualizes what happens when you run Python code. It is useful to understand existing code or debug broken code. Being able to visualize code execution is a critical skill for all programmers.

We suggest everyone complete Python for Everybody (PY4E) https://www.py4e.com/lessons. As stated above, you have to learn through doing so create a login and complete the exercises which are auto-graded.

Here are additional online courses:

- https://www.coursera.org/learn/python
- https://www.coursera.org/learn/interactive-python-1

- https://www.edx.org/course/cs-all-introduction-computer-science-harveymuddx- cs005x-0

b. Tools

Throughout the MSDS program, you will use the same tools as professional Data Scientists. That means by the time you start Practicum or a job, you'll be ready to contribute to the team right away.

Please check out:

1. A quick introduction to the hardware and software elements of your machine
2. A broad overview of python and tools used in our MSDS program

For the installation of Anaconda and JupyterLab, please follow the video below. **You are required to install everything mentioned in the video by the beginning of the boot camp:**

Link: https://studystudio.ai/subscribe/jCGdHXkX/MSDS501%20-%20Computation%20for%20Data%20Science

Before arriving at orientation, you should have Anaconda installed on your laptop and have some familiarity with the command line (Terminal.app or iTerm2 etc…). Using the command line is a critical skill in this program (all the rest of the examples assume use of the command line). The command line is also called "the shell." Data Scientists use the command line every day to run scripts, manage files, or use computers in the cloud. Go through a course, such as https://guide.bash.academy/, to make sure you are familiar with the command line.

We mostly use Jupyter Notebooks and JupyterLab. For more information, see http://jupyter.org/. Notebooks combine programs and text which is nice to interweave your code with your thoughts. Data Science programming is particularly challenging because of the variety of tools we use. Notebooks are extremely helpful because data, data frames, code, graphs, and output are all in the same document. Here is a video tutorial to check out: https://www.youtube.com/watch?v=HW29067qVWk.

c. MSDS 501 Computation for Analytics class recordings from Summer 2023

Professor Diane Woodbridge has made the recordings of her MSDS 501 Computation for Analytics classes from last summer available to you. Please note that the course will be taught by a different professor this summer, so these recordings should be used as additional study material and preparation *in general*, not for specific assignments.

You may view the recordings of the Programming course in Summer 2023 in studystudio:

`Link:`
https://studystudio.ai/subscribe/jCGdHXkX/MSDS501%20-%20Computation%20for%20Data%20Science

First, you need to create an account, and then copy/paste the URL. You can see the recordings in the “Shared” tab.

Additionally, you may access the course slides and codes in github:
https://github.com/dianewoodbridge/msds501_computation_2023

Note that we have a separate document for the **Laptop Minimum Requirements**. Please make sure to read it and prepare accordingly. We look forward to meeting you soon!