MS-CAPP First Year Curricular Overview

This document is intended to provide an overview of the technical skills developed in the first year of the MS-CAPP degree. Carefully reading this document should help inform you as to what internships you are well-suited for, as well as how to market yourself in cover letters and interviews for those internships.

1. Data Manipulation and Exploratory Data Analysis

In several different programming languages, you will learn to manipulate and explore data sets. This includes working with tabular datasets, often called dataframes, in: (a) R's tidyverse (that is the set of packages used in R for Data Science), (b) Python's pandas library, and (c) STATA. Further, you will employ a wide range of descriptive statistics and basic exploratory data visualization.

2. Databases and SQL

Databases are how organizations store large amounts of data in a structured way. You will learn to design, query, and build databases, including cloud-based databases in Amazon Web Services. In Aaron Elmore's course, you will first learn to use SQL (pronounced *sequel* or S-Q-L) to create and interact with *relational* databases, which contain many tables. A table is similar to a tabular dataset that you might normally store in a .csv or open in Excel. Being able to write queries for these databases with SQL enables you to make effective use of this data.

You will also learn to use NoSQL databases, and query these with NoSQL (Not-Only SQL). NoSQL databases store large amounts of less structured data, each unit of which is called a *document*. A common example of a document you might store in NoSQL is a tweet. You can read more about SQL and NoSQL databases here.

3. Causal Inference Statistics

Your statistics courses, especially Statistics for Data Analysis II / Mathematical Statistics II and Program Evaluation are oriented towards teaching causal inference. In causal inference statistics, we use a large group of methods to explore whether we can determine if one variable had a *causal* impact on an outcome variable of interest. A huge portion of policy and social science research is based in causal inference, so do not underestimate its tremendous importance in this field.

Causal inference statistics are inherently retroactive – we are looking backwards in time to determine what has happened in the past. Common examples are linear regression (for a continuous outcome) and logistic regression (for a binary outcome). For those of you who have the Economics waiver and can take Program Evaluation before summer, you will also be able to run important models like difference-in-difference, fixed and random effects, and instrumental variables. Program evaluation is one of the common tasks for which we use causal inference in public policy research – we ask did this policy change have an effect (e.g. did expanding Medicaid improve health outcomes?).

4. **Machine Learning and the Underlying Math**

Machine learning focuses on using algorithms to predict outcomes based on data. Conversely to causal inference statistics, it almost always cannot and rarely even attempts to provide causal information. Instead, machine learning aims to accurately predict future outcomes based on prior data (called a training dataset). You will learn to use Python's SciKit-Learn package to use algorithms like decisions trees, random forests, support vector machines, and k-nearest neighbor. Further, you will learn how to validate models (ensure they are right) using cross validation and holdout datasets.

You will also learn to ethically apply machine learning to real world problems in Rayid Ghani's course. This is very important because there are many real risks of bias and harm that can come from a naïve approach to machine learning. It is a meaningfully distinguishing aspect of this degree that you are prepared for using machine learning in the context of governance and policy, where we are obligated to hold higher ethical standards than the private sector. Jens Ludwig's article on the subject is a good place to learn more.

Further, you will learn the underlying mathematics of machine learning in Amitabh Chaudhary's The Math for Computer Science and Data Analysis course. This includes a focus on linear algebra/matrix algebra.

5. **Programming Fundamentals and Computational Thinking**

As you have surely noticed, the MS-CAPP degree takes very seriously the teaching of fundamental skills in programming and computational thinking. Over time, you will learn to appreciate how this allows you to learn much faster than people who have not learned these fundamentals.

When you discover different Python or R packages, encounter foreign algorithms or models, or even come across entirely new programming languages, you will be dramatically better prepared to learn how they work armed with a comprehensive understanding of the underlying pieces at work. To an employer, this means you will be very prepared to tackle new technical challenges that you have not seen before.

It may also be beneficial to note (do this briefly or in passing) that you have learned important tools, like Git for version control and collaborative development, as well as and Linux command line, which constantly come up when working in applied settings.

6. **Economic Analysis**

Your economic coursework, in Microeconomics I and II, prepares you to model individual and organizational incentives mathematically. These courses teach a rigorous approach to exploring behavior and the effects of public policy on people and markets. Further, these methods inform data-driven models like those based on microsimulations (e.g. the Urban Institute's Tax Policy Center model) and agent-based modeling.