My path to data science has been an academic and personal journey filled with many meandering paths. My dream for many years was to pursue a life dedicated to scientific discovery however two years into my undergraduate degree I decided to change majors and follow my passions. I was deeply interested in language and logic and looked to find a way to pursue this interest in a meaningful and impactful way. After taking a leave of absence from school and reflecting on what path was right for me, I realized that I craved a hands-on approach to education. More specifically, I wanted the ability to produce something meaningful from my skills. My desire to find a practical use for my education, eventually led me to taking introductory courses computer science. However, programming skills alone are not enough to make you a good Data Scientist.

The moment my obsession with learning about data began was the day I realized how important it was to have the skills necessary to take advantage of the most abundant resources of the 21st century, DATA. I can recall one day in particular during the height of the covid-19 pandemic, two months into quarantining at home. My parents, both Mexican immigrants and English language learners, were amazed with the ease of use that speak-to-chat technology afforded. Yet even though they were both saying the right activation words, neither of them was able to utter the phrase “ok google” well enough to activate my phone’s speech assistant. Although my parents brushed it off as a quirky computer bug, it left a lasting impression in me and made me question “could I make this work for them?”. This desire to fix something like voice recognition lead me down a rabbit whole of YouTube videos, vlog posts, online forums, and GitHub repos all trying to explain the basic data science and machine learning pipeline. After some time of exploring the basics of voice model training I learned the reason why the google voice assistant was not responding to my parents. First of all, voice assistants like any other machine learning model were trained on large amounts of data. This data is stored in data sets which depending on the use case are either collected from the of in the case of speech-recognition saved from voice recordings. Large companies like Google or Amazon have access to millions of voices enabled devices that store and analyze voice data from millions of users daily, yet there is still a huge issue in understanding people with accents. This is because when the speech assistant is trained the majority of users speak somewhat standard English and even though companies like Google are trying to capture English samples with accents data points from these users in particular are pretty sparse. So, a lack of diversification in the data collection process combined with models trained on this data makes understanding users with accents difficult for speech assistants.

But my understanding of the issues only went so far, and without an in depth understanding of machine learning models and a thorough understanding of the python and statistics that form these models. My only hopes to improve these voice assistants was limited to adding voice data rather than improving the model themselves. So, the last step that began my journey into Data Science was enrolling to the Flat Iron School in New York City. As I sit here writing this post, I am only two weeks into the 15-week data science program, yet in these two weeks, I have spent over 400 hours learning, practicing, and honing my skills as a Data Scientist. I will be updating this vlog bi-monthly to share my thoughts, progress, and insights into the program. But I can ensure that by the end of the program I will have the skills necessary to make a meaningful impact*.*