**What programming, scripting, databases, or statistical languages have you worked with, and in what capacity?**

Please make a detailed list. For each language, database, or software, please thoroughly describe how much time you spent working with it, as well as the related work and results.

Python: This summer, I spent approximately 300 hours (~30 hours a week for 10 weeks) on my side project scraping, visualizing, and modeling Bay Area home prices in preparation for this program. Earlier this year, I also spent something like 50 hours (~12 hours a week for a month) developing image analysis scripts to automate the measurement of features of interest in microscopy images of flash memory devices. This work significantly reduced engineer time spent manually measuring microscopy images.

MATLAB: During my PhD, I spent approximately 250 hours (~20 hours a week for 12 weeks) writing code to identify the centers of nanocrystals, measure the separation between them, and count the number of nearest neighbors in electron microscopy images of nanocrystal arrays. This code was the basis for my J.  am. Chem. Soc. publications in 2014 and 2015.

R: During my postdoc, I spent approximately 30 hours (two weekends) visualizing income, expenses, and cash flows pulled from my credit card transaction history.

**Please describe any relevant side projects you have completed that highlight your transferable skills (unrelated to coursework).**

Descriptions and public URLs of side projects you have worked on outside of course work. Write 'N/A' if you have not completed a relevant side project.

To hone my data science skill set and satiate my curiosity to understand the dynamics of the housing market in the San Francisco Bay Area, I scraped several thousand single family home listings from a popular real estate website, visualized them geographically, brought in additional features of commute time and school quality, and fit the data with a multiple linear regression model to uncover underpriced listings/neighborhoods as potential investment opportunities.

Please see the report “Bay Area Housing - Michael Boles - August 2019.pdf” on my GitHub (www....).

**Please describe any relevant coursework you have completed that highlight your transferable skills.**

Please include a short description of any relevant courses and related projects. Write 'N/A' if you have not taken any relevant courses.

Most recently I’ve viewed and worked through some of the lectures and exercises in the Machine Learning and Statistics for Data Science courses on Udemy.  
  
During my PhD and postdoc I attended several software carpentry workshops. These daylong events on coding in R and Python, and navigating the Unix shell, provided me with the foundation for writing scripts for work and side projects.

**What statistical methods or machine learning models have you worked with and in what capacity?**

Please include a short description of any relevant courses and related projects. Write 'N/A' if you have not taken any relevant courses.

Making use of some concepts I was exposed to in the Udemy Machine Learning and Statistics for Data Science courses, I made use of Scikit-learn and Statsmodels packages in my side project analyzing Bay Area real estate. With these tools I fit my data set to a multiple linear regression model, evaluated statistical significance of features within the model, and eliminated some of them from the model based on tests of multicollinearity.

**Insight essay questions (1000 characters)**

***General questions***

*Describe your professional experience and/or research to a non-expert.*

I’m currently an engineer at Lam Research, a Silicon Valley semiconductor capital equipment manufacturer. I work with leading chipmakers (e.g., Micron, Intel, Samsung) to design the chemical/physical processes that shape the nanoscale features which ultimately become the bits used in future generations of NAND (nonvolatile, e.g., flash drives) and DRAM (volatile, e.g., main memory) devices.

As PhD student at UChicago, and later a postdoc at Stanford, I developed solution-processible inorganic materials for transistors, LEDs, and solar cells. Complementing my experimental work was a heavy emphasis on developing MATLAB code to extract various geometric properties of nanoscale objects contained within electron microscopy images.

These experiences have given me a somewhat unusual perspective at the interface of hardware and software, from bottom-up assembly of individual atoms, to monitoring data accumulated by thousands of sensors, and ultimately extracting insights from image data.

*What programming, scripting, databases, or statistical languages have you worked with, and in what capacity?*

Python. My side project analyzing single-family home prices in the Bay Area

Python, R, and MATLAB – describe projects for each.

*Please describe any relevant side projects you have completed that highlight your transferable skills (unrelated to coursework).*

*Please describe any relevant coursework you have completed that highlight your transferable skills.*

Most recently - Udemy: machine learning, statistics courses

PhD research – MATLAB for analysis of microscopy images – published JACS 2014

During PhD/Postdoc – Python and MATLAB courses given by UChicago Library, software carpentry workshops offered at Stanford

**Data science**

*What statistical methods or machine learning models have you worked with and in what capacity?*

*What excites you about becoming a data scientist and working in the tech industry?*

1. Thrill of writing code, and 2. Epiphany during data analysis.

I find at least two aspects of a data science career to be particularly exciting: (1) the thrill of writing code, and (2) the epiphanies that pop up during the data analysis.

In the first case, I can say without hesitation that seeing a snippet of code work as I’d envisioned, particularly after a long struggle, is about the most exciting feeling I’ve experienced. For instance, learning to scrape the web for my real estate side project was a big victory for me. After scouring StackOverflow and posting several questions of my own, watching the dataframe of home listings grow to several thousand entries was enough to send me dancing around the house.

I also love to see how insights begin to emerge during the analysis of a large dataset. In that same project, by plotting color-coded scatter points on top of a map of the Bay Area, I saw in one image what I’d been trying to understand over years of browsing Zillow and driving around. My love of maps, California, and numbers converged into one graphic that I stared at for longer than I care to admit.

[old answer – make more personal]

As an engineer at a Silicon Valley semiconductor company, I’ve seen first-hand how data collection, storage, and analysis is reshaping the world. Because global data storage demand is growing at > 50% annually, the data economy is the primary driver of revenue for chipmakers. It is also the largest opportunity to improve the efficiency of the semiconductor manufacturing process, since speed and uniformity gains attained by machine learning can increase output volume at minimal added cost.

Motivated by the desire to acquire the relevant skillset, I’ve used my PhD coding experience as a springboard to build a side project focused on a personal passion of mine: figuring out how I might become a homeowner in this part of the country. Frankly, I found tremendous enjoyment in learning to scrape data from the web, visualize it with maps and scatter plots, and build a model that explains how factors such as commute time, school quality, and crime rate influences the market value of a house.

*Any comments or additional information you wish to share with us?*