

Project 2

Conrad Polkosnik

January 30, 2018

1 Tools of a Data Scientist

1.1 Data Collection (Measurement)

In data collection, as economists and statisticians, we want to make sure our data is of complete and important value to what we are interested in studying. We want to understand the distribution of the data set, the sample size, the variance, as well as the kurtosis of the data if possible.

1.2 Data Storage/Manipulation

In regard to storage, we want to make sure we store the data properly following the syntax of our programming language. We want to make sure the computer can understand the means by which we input our data. In regard to data manipulation, we do not want to 'over-present' what the data may tell us. It is improper as economists to 'over-exaggerate' a causality by which the data could be limited to.

1.3 Visualization

In my opinion, I believe it is important to 'dumb-down' a topic of interest as much as possible. Doing so, for the average person who has some sort of interest, will be able to easily understand the data being presented. The way in which we present the data (i.e. charts, graphs, tables) is significantly important as well. Such charts, graphs, and tables provide one with interest to easily visualize the outcomes of many of the complex statistical models by which the programmer has completed. It can be very hard for someone with no coding experience to understand fully, how the model has come up with such numbers, especially without comments in the code for guidance.

1.4 Statistical Programming/Modeling

Statistical programming and modeling can explain all sorts of deep learning, machine learning, and causality in the spectrum of big data and data science.

There are many developers out there today who run all sorts of complex models using programming languages that, personally, I have no experience with. But, what I do understand from thorough research, specifically in the realm of blockchain technology, is that big data is today, a hassle in every industry. Companies have hundreds of thousands, if not millions of lines of data stored away which they have a hard time organizing and finding causal inference from. A specific blockchain company that I am currently invested in comes to mind, named Snovio. Snovio is a decentralized lead generation platform which functions on crowdsourced data collection and the blockchain. Basically, Snovio solves the issues of low-quality and outdated leads in the marketplace. It also provides data suppliers a guarantee of the revenue from data sales to companies who want to use their service. A service like this can be used in literally almost any sector of business (i.e. marketing, advertising, oil and gas land acquisition) and so forth. In summary, as a computer programmer or data scientist, it is important to see real-world application of the statistical models or programming that continue to be built. Such models may one day change the world.