Chris Levy Remesha Ngabo

Prof.Jordan Wirfs-Brock

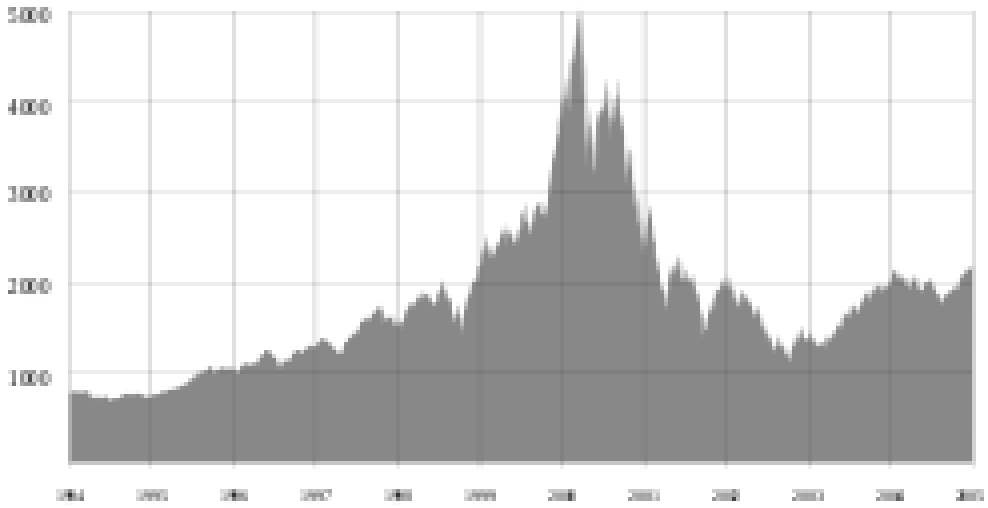
CS-215/MATH-215

15 May 2023

Personal Data Manifesto

In my mother-tongue, Kirundi, Data simply means father. I grew up using this formal word to both call my father and occasionally the heavenly father. Taking from this meaning, data is a being that has answers to my questions, possibly. My dad surely had some answers to some questions of mine and the heavenly father knows everything. I grew up idolizing my father and I still do. A father, as well as a mother, plays a huge role in the development of their children and provides insights, teachings, values and norms. However, as the african proverb says, it takes more than parents to raise a child, it takes a whole village to raise a child. What my parents and the surrounding community shared with me are the norms and knowledge, they sold me a way of living.

The second meaning of data, I heard, is more related to the dot-com boom of the 1990's. The boom came up with a way of making the globe a more connected entity, more like a village some people say. I think the doc-com shaded light up on a new world, a virtual one but yet so real. Everything that was done on the internet was kept, stored in extractable forms. This “Everything” is the data, unimaginably big. The creation of more sophisticated devices such as mobile devices provided an increase in data collection which propelled data-related fields.



The NASDAQ Composite index spiked in the late 1990s and then fell sharply as a result of the dot-com bubble

For me, this everything, data, is an accumulation of less meaningful raw figures (words, numbers, etc) which can be organized, analyzed, or processed in a meaningful way.

Therefore to be a data scientist means having the ability to ask questions that make sense of that data. Data Science is the study of data from its collection to the possible insights that can be generated from. Data is too raw and once processed, one can do wonders.

Since my manifesto comes with a cookbook, I am going to make a joke about cooking. When I am asked if I know how to cook food (I don't really know how to cook in reality, I can only cook what I can eat when I am hungry and presented with no other options) I reply, " The fire does most of the cooking".

Yes, the libraries and the packages do most of the cooking or the processing of data. I think that it is such a privilege that we have data scientists who built so many libraries that simplify the work. This part is really underrated in my opinion, I wonder how people used to process the data without most of those libraries. I am thankful for those who contributed and laid foundations for us while simplifying the hard part.

I think that libraries and packages can't do everything on their own and there has to be a cook with some skills, there has to be a data scientist capable of asking questions and getting desired answers from the data.

Based on my novice experience from this class, I can provide an answer to what process I applied while approaching data. Here's an outline of my process:

- The data has to be in convenient form.
- Opening the dataframe and exploring it.
- Framing the Questions.
- Data Wrangling and Data Visualization.
- How can this be improved and What else can be done.

In the class, we worked with Data that was in .csv and .excel formats which is already in a format so the point one can be ignored. However, when it came to web scraping, the first point really needed a certain degree of attention.

Opening the data frame includes reading it and using the .min(), .max(), .describe(), and .info() to list a few. These will allow us to know what kind of data we have and help us into framing questions. The frame of Questions is a crucial part in the process, knowing what kind of questions you can get answers and those out of context defines the boundaries of our dataframes. For instance, trying to answer questions out of the context of the dataframe you have is wrong. After knowing the questions one wants to answer, one can start doing the data wrangling which involves using the columns and rows of your choice. This wrangling ought to be followed with a production of visualizations. Visualizations reveal patterns and trends. The last point is to know that however the work is accomplished, trying to think of ways to improve it is also very important.

That being said, there are many points that I feel like I should raise. The data in general is a source of information, there's a possibility that data can raise imperfections through its collection or processing. Also, the interpretation of data all alone might not be all that is to know. Data generates understanding of the world around us but that understanding should be also questioned.

In conclusion, Data should be seen as an ally as Chetima (a classmate) said. It is a tool that can help us by providing insights to help us make better decisions. In general, I hope to get to work with bigger data sets and try to apply this knowledge into what's out there. I hope to be responsible while working with data and frame clearly my questions so as to reduce the generation of biases.