Paxton Luther
Professor Wirfs-Brock
CS 215
Dec 13, 2023

Personal Data Manifesto

Data has become increasingly prevalent in both the world and in my personal life. I often see the usage of data in the world and am sometimes exposed to my own personal data or the effects of someone using my personal data. With this understanding it is becoming increasingly important to have knowledge of data science. I personally view data almost entirely through the lens of computer science and generally understand data to be differentiated from simple statistics and information by its scale, means of collection and how it is interpreted and used. Data is often in very large datasets and so traditional methods of analysis without using a computer are ineffective, this combined with the fact that many datasets are created by computers necessitates the usage of computer science in the understanding of data science. Data science however is not focused on all types of data equally some are more easily processed then others, so the data we manipulate is generally qualitative in nature. I believe that to be successful in data science one must consider both the question that they are asking of the data as well as how they will utilize whatever tools they have at their disposal to answer the question.

Part of using a dataset involves understanding which questions are feasible or meaningful to try to find answers to and which are either not feasible or generally not particularly meaningful. This is one of the most important parts of data science as the kinds of questions you ask will determine the effectiveness of your approach and your understanding of the dataset. For instance, in my project to analyze the Ao3 dataset there were several questions for which the answer was either prohibitively challenging to find or for which the answer was not particularly

insightful. An example of the latter was finding the largest single work of fanfiction in the dataset as this did not reveal much about the rest of the dataset as the largest fanfiction doesn't have much relation to the sizes of the other works. This of course is not always the case as there are many datasets in which simply knowing the maximum values can be illuminating as in our exploration of wildfires. Generally finding the right questions can begin almost as soon as or even before you have found your dataset as various dataset allow for different questions to be asked. It can be useful to try to connect more than one type of data, i.e. date and the presence of an entry, to find questions that have answers that are particularly meaningful or insightful.

To aid in answering the questions and understanding the data in greater depth a visualization is often a very effective tool. Generally with large datasets it is almost impossible to get a full picture of the shape of the data without the usage of a visualization method of some sort. For example trying to track the growth of the dataset overtime is fairly difficult to do without a visualization and the result would be hard to interpret, but with a visualization even as simple as a line graph the data is not only more accessible and readily worked with but the interpretation of the data is generally much clearer. Generally seeing the relationship between two different variables or columns is fairly easy with a visualization such as a scatter plot.

Of course to best understand the data one must first understand the context surrounding the data. Without context much of the data that we look at is largely just meaningless numbers. For instance in my Ao3 project I calculated the percentage of complete works in that dataset. Without understanding what Ao3 is or what defines a complete work that percentage is meaningless. Understanding that Ao3 is a fanfiction website and that a complete work is simply a work that has been marked complete by the author and generally, but not always, means that the work has ended and will no longer be updating allows one to understand why I did not

simply accept 84 percent at face value. To be able to do analysis of any data one must first know what that data means. It is not simply enough to find various connections between the data if you don't understand the implications of those connections or what you are connecting.

Perhaps the most important thing to do before any data analysis is to make sure that the data is clean and accurate. Performing what should be fairly simple operations can become much more complicated if the data in a column is not all of the same type. For example in my Ao3 project when I attempted to find the growth of certain fandoms over time some of the works had zero tags and thus trying to separate out the fandom tags failed as there was nothing to separate. One must also be careful that your dataset is itself accurate though this is harder to gauge from a simple coding perspective and requires best judgment.

In short I believe that data science can be defined as the study and application of large amounts of somewhat varied information and data for the purpose of gaining a better understanding of that data and the world as a whole. To do this as a data scientist it is important to ask the kinds of questions to your data that will allow you to glean the most insight from the data. Of course the types of questions you ask must be dependent and fully aware of the relevant context surrounding the data and its creation or else you run the risk of either asking poor questions or in misinterpretation of your results. Alongside context, understanding the actual data itself and determining whether or not it is useful in its current state and if the data is clean and collected properly. Often in data science the management of the data is more important and more difficult than the actual analysis. To perform that analysis I believe that a helpful tool is to create a data visualization so as to provide easy access to any trends or patterns in the data and to easily understand the whole scope of the dataset.