Alina Orlioglu
CS 1030
Final Project – Data Analytics
Khan Academy Progress Report

## Lesson One: Data Tools

Data storage is vital within the digital world. In order to make any data useful, there are many ways to store and access data. Common data storage examples include text files, spreadsheets, and databases. Once data has been stored, we can compute statistics that describe the data set. There are many tools available for analyzing and organizing data, and the tools used will be dependent on our needs. In addition to being able to view how much data we have, statistical functions can provide insightful information such as calculating the average, largest and smallest values, calculating the total in a column, data filtering options, and summarizing by group. One reason to collect data is to study trends, pattern, or find correlations in data sets. When put together and analyzed in a chart or graph form, these patterns in data can help quickly visualize what is happening with the data, whether it is trending up or down, help visualize statistical fluctuations, and can be used to make predictions about where the data will go next based off of previous patterns and trends.

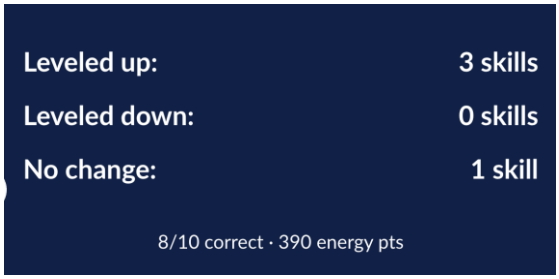
## Lesson Two: Big Data

Any time we participate in the digital world, we are contributing to data collection by any online service. With the growing number of people and cities that are connected to the internet, data collection sets are increasing in size. With data sets being so large nowadays, our traditional ways of storing and processing are not capable of handling it, which presents challenges to computer scientists and data engineers. But, because we have so much data, there are new analyzing opportunities that would not have been possible with smaller data sets. The way that we are able to collect so much is data is because it is coming from multiple sources. Data can be collected and stored from scientific research, some of which may be subsidized by government funding, which allows both scientists and hobbyists to access this data and perform analyses and provide new insight. Digital libraries are another example; digital libraries collect and store large numbers of documents, artifacts, and media. User facing applications are also large sources for data, Facebook for example can generate 4 new petabytes of data every day. Storage and processing are two main considerations when dealing with massive amounts of data. When an organization needs to manage thousands of hard drives, they can store the hardware in a data center, where the infrastructure is able to maintain proper temperatures and provide electricity, but also contains the necessary components for networking. Because almost all data is related to people in some way, companies need to be critically careful with their handling of protected information.

## Lesson Three: Bias in Machine Learning

Machine learning algorithms will automatically improve themselves based on experience, by processing more data and improving itself based on the data. The three general methods for machine learning are reinforcement learning, unsupervised machine learning, and supervised machine learning. Supervised machine learning is the most common approach and utilizes labeled data. The algorithm analyzes the data and learns how to map input data to an output label. One popular method for supervised learning is the use of neural networks. Algorithms trained by this method are fed large amounts of labeled data, so if the machine's purpose is to recognize and classify images, a training data set may contain thousands of images labeled as "car" or "dog". When it comes to accuracy in recognizing and classifying images, the accuracy of the algorithm is dependent on the training data sets it received; the size and diversity of the training sets will correlate with the accuracy of the algorithm. Diversity in training sets is vital because machine learning algorithms can become accidentally biased against a group of people based off the data that is provided in training sets. Examples of bias in machine learning include risk assessment score algorithms becoming biased against Black people, resume screening algorithms preferring men over women, certain facial recognition software having large accuracy discrepancies in gender classification, and even biases in gendered language translation.

## Unit Test




Based on my performance on the test, it seems as though I didn't fully grasp machine learning bias the first time around. My first instinct during the test was to select answers that dealt with the size and samples of the testing data – particularly on the questions that offered larger training sets as a solution to machine bias. To me it seemed obvious that more data means a better trained algorithm, but this is not always the case. This is because a larger training set does not necessarily mean a more diverse and better data set. Reading scatter plot charts is a bit hard for me to do, it does take me a bit longer to process the data and understand fully what I am seeing, while this did not affect my performance on the test, I still noticed that the scatter plot questions took me longer to answer than others. Khan Academy report states I improved on all three skills except for the Bias in Machine Learning, which consisted of no change