Adam Ettachfini

Data Tools

Computers are used to create and store data. A computer program may store data as temporary memory but needs a permanent way to store this information. One common way is to store the data as plain text in a text file. You can organize this data in comma-separated values (CSV) to better organize it. Using the CSV format can be useful but it comes with the issues of being computationally taxing to pull specific data and the limit of how much data can be stored being the file size with it taking longer to open with the more data you have. The most common way for storing data is with a database. Databases store information in different files while keeping track of where they are. Databases use a query language to let programs or people add or read information with the most common language being Structured Query Language (SQL). The main use cases for databases are to store and analyze large datasets of information. This information can be represented in the form of a graph to clarify the overall trend of the data.

Big Data

The internet is made up of trillions of gigabytes of information which has led to computer scientists needing new ways to store this information. Much of the data comes from scientific research, digital libraries, medical records, and social media sites. Some of these sites can create multiple terabytes of data every day. Data centers are required to store the large amounts of data used. Data centers consist of multiple servers of hard drives or solid-state drives connected in parallel that are then connected to the internet.

Bias in Machine Learning

Machine learning is a type of algorithm that can improve on its training with minimal to no human involvement. The three types of machine learning are reinforcement learning, unsupervised machine learning, and supervised machine learning. Reinforcement learning is when an algorithm learns to perform an action that will get the most reward through constant feedback. Unsupervised machine learning is where an algorithm finds patterns in its training data to then analyze new data with a similar structure. Supervised machine learning is where an algorithm analyzes data that is labeled to understand the relationship between the label and the data. A supervised machine learning technique that is commonly used is a neural network. Neural networks are a model where the input is put through multiple nodes or neurons eventually going to an output. Neurons are trained on labeled datasets that change the weights of each node to fine-tune the model during training. The accuracy of a neural network is dependent on the quality of the training data. Neural networks are prone to amplifying existing biases in training data.

Unit test

Completing the unit test has helped me understand the specific use cases of data analytics and collection. I now know what data types would be more appropriate for specific cases and how to format them. It has also made me more aware of storage size and ways to optimize a database by removing unnecessary information. The unit test also goes over how to read database tables and CSV files. It specifically makes you pick out relevant information from the data to compare against other information. The different ways to analyze data are covered with it making you choose the most appropriate method for a specific example. I learned about parallel computing being faster than just using one computer because the load can be separated across multiple computers. Overall the unit test has made me think differently about the use cases of data collection and analytics as well as the different methods that can be used to efficiently complete different tasks to take advantage of the storage and computational limitations a specific system may have.