Data, from Small Solutions to Big Problems

In the beginning, man created bits. This in itself was amazing, however before long the bit was insufficient and the first digital evolution took place. The bits were fused together in groups of eight, they were now a collective entity that behaved in all the same manner as the original plus more. Bytes where then group in the thousands. Bytes to kilobytes, then megabytes, gigabytes, terabytes, and so on. These events occurred in a relatively small amount of time and have resulted in the monsters that exists today. Commonly referred to as Big Data, the exponential growth of computing power and communications have given rise to a need for large amounts of storage. At first there were few masters of these machine and they help facilitate a healthy ecosystem for these delicate creatures to flourish. After time there came Individuals who did not fully comprehend the interworking’s of these creations. They poked and prodded until these codes bent to their will. Over time these electric slaves became very misunderstood and their masters became less understanding. Eventually leading to separation between men into two groups the “pokers” and the “caretakers”. The pokers contribute in a negative sense and the caretakers help protect machine and men from the ones who would tinker without knowing. We must be careful to not upset these delicate creatures as we become more dependent on them, for in the future, we may be slaves to them.

In an attempt to understand this better, let us start back at the beginning and move forward. The first known bit was not electronic, but rather a hole punched through a stiff piece of paper that was fed into a mechanical loom to produce specific patters by Basile Bouchon in 1732[1]. This technique was modified a couple times and later applied to informatics and then numeric counting by Charles Babbage. Babbage’s machine is considered to be the first computer because of its ability to process and store data for computing purposes[2]. The encoding of text occurred for the first time with the invention of Morse Code in 1844. This is the first time bits were paired. Let us jump ahead to 1928 with the introduction of logarithmic bits. This allowed complex data to be abstracted into almost any number of bit. In the next several decades, events started to snowball and things started to get out of hand. Problems started to arise as the digital evolution took off. The people in charge of these creations implemented standards and control formats. As man is lazy and imperfect, some of these standards were not adhered to. Small anomalies started to arise in the behavior as the size and complexities of the machines grew. Fast forward to the 1990’s and here comes the internet. Now more than ever the evolution is in full force and more humans have the ability to modify digital processes and also get more reliant on them. This gives birth to the monster in question, Big Data. Large pieces of software are now keeping accounts of large fractions of the world population. This poses more problems as the data monster grows there is a lower ratio of man to machine.

More specifically speaking Big Data can be classified into different areas. Black Box Data consists of flight records such as aircraft components, voice recordings, GPS signals, trajectory, and more. Social Media Data, such as Facebook or Twitter, keep track of information on connections, views, and large amount of text. The rest include Stock Exchange Data, Power Grid Data, Transportation Data, and Search Engine Data.[3] These information banks are so vastly large that traditional data processing techniques have once more become obsolete. The data contained in these records can be one of three different types Structured, Semi Structured, and Unstructured. Structural data, such as SQL, has relationships that tie them together making for better access to records that may be disconnected. Semi structured data has some relational data only for basic organization. Finally, unstructured data has no relationships at all and is used in raw data processing. Big data is important in providing detailed analysis on large scale data sets in real time. This gives companies better decision making facts to optimize performance of the company.

With big data comes big problems. The biggest being access, with such large data sets it is very difficult to find one single item in the million, billions or more records contained in the set. This is not very easily solved but more easily managed with the differing big data types mentioned earlier. The issue of concurrent access is another major problem. That is to say when two or more entities modify the same records simultaneously, which one should be kept. Again no real solution exists but at the moment records are checked for differences and merged together, another technique is to notify and reload the data to the user that attempted to save last[4].

This leads us to our next topic. What happens when these monsters get out of hand?

According to DigitalGuardian.com, the largest data breech in history occurred around March of 2012.[5] A company called Experian had acquired a smaller company called Court Ventures, this company processed public records for analytics. In the process acquiring this company, there were certain detail that were not taken into consideration. This is a perfect example of careless pokers without the foresight to predict what would happen. Had the proper precaution been taken by fully analyzing the company’s new assets, this entire event would never have taken place. That being said, the smaller company “Ventures” held a contract with another company called “U.S. Info Search” which allowed Ventures to search Info Search’s records of sensitive data for court proceedings. After Experian acquired Ventures there was information sold to a number of third parties. One of which included selling access to a company called “Vietnamese fraudster service”. In this deal Experian had inadvertently given a foreign company rights to look up American names, address, social security numbers, and more. The Secret Service notified Experian of what had transpired, but at that point it was too late. There is no way to know for a fact, but its estimated that more than 200 million records were “stolen” over the course of ten months.

Education and careful analytics are the key to preventing future disasters. When technology is handed from on person to the next, even without intention, a caretaker can become a poker. Not fully understanding what that person has been given, it is easy to achieve an undesired effect. If this continues who knows what the future holds for the symbiotic relationship between man and machine.