**Data Science**

Similar to the realm of Information Studies itself, the field of Data Science is heavily multidisciplinary requiring a knowledge of mathematics to construct models, programming to implement them, and subject matter expertise to contextualize any insights discovered and convey them to the business. While the necessary weighting of each of these aspects of Data Science may change with a given project, it stands that true mastery of the field as a whole requires a balance of all three. Similarly, due to the fact that virtually any field can benefit from the implementation of data science practices, the variety of implementation cases are endless. As such, it is difficult to determine a single definition with which to define the entire practice of data science and I would argue that it is better described as a process which encompasses the collection, storage, and leveraging of data in order to generate meaningful insight.

Originally predicted by John W. Tukey in 1962, the field of Data Science was initially born from his observations that data analysis should place more emphasis on scientific method than previously determined. (WISCONSSIN) However, the implementation of Data Science today would be likely to astound Tukey who’s remarks came at a time when the world had yet to see its first computer. Even with the advent of desktop computers and home computing possibilities, the modern implementation of analyzing rows of data quantified in exabytes as well as the methods and systems necessary to do so demonstrates the exponential growth thus far.

Having completed my Undergraduate degree at the School of Information Studies at Syracuse University with a concentration in Data Analytics, I was eager to expand upon my newly established skillset and better cement my programmatic, mathematic, and communication skills through a Master’s in Applied Data Science. The outcome-based approach of the program similarly appealed to me due to the hands-on nature of the field as a whole, and I can now confidently say that I have an understanding of the core tenants of the program:

* Describe a broad overview of the major practice areas of data science.
* Collect and organize data.
* Identify patterns in data via visualization, statistical analysis, and data mining.
* Develop alternative strategies based on the data.
* Develop a plan of action to implement the business decisions derived from the analyses.
* Demonstrate communication skills regarding data and its analysis for managers, IT professionals, programmers, statisticians, and other relevant professionals in their organization.
* Synthesize the ethical dimensions of data science practice (e.g., privacy).

Over the course of this paper, it is my goal to prove both by own development in these respective areas and to discuss how they fit into the broader context of Data Science.

**Data Collection & Organization**

After determining a question or problem of interest, data collection and organization is the first step in the data science process. The process of data collection can vary significantly from simply downloading an already cleaned dataset from a Source like Kaggle, to sifting through messy (and sometimes, incorrect) livestreaming data from sources such a light sensor, to collecting in-person survey responses from customers regarding their feedback on a product.

After data collection comes the determination of how best to house it for analysis. Over the course of the Advanced Database Administration Concepts & Database Management offering with Professor Fudge we addressed concepts such as how to determine the best medium and format for storing data, what implementation of a given database type is most effective, and how to optimize the data through indexing for the best possible query performance.

In practice of these concepts, the storage format of the data should be suited to both the data itself and the use case for it. When considering the medium with which to store the data, the use case is quite important. For instance, if I wanted to make a collection of data available to not only myself, but others within my organization for querying, it makes more sense to store said data in some form of database rather than having to distribute a static flat file. From a structure standpoint, it is important that the schema is suited to the data itself; if I was interested in storing high velocity live tweet data from Twitter, a relational database model is unlikely to make sense compared to a non-relational database offering.

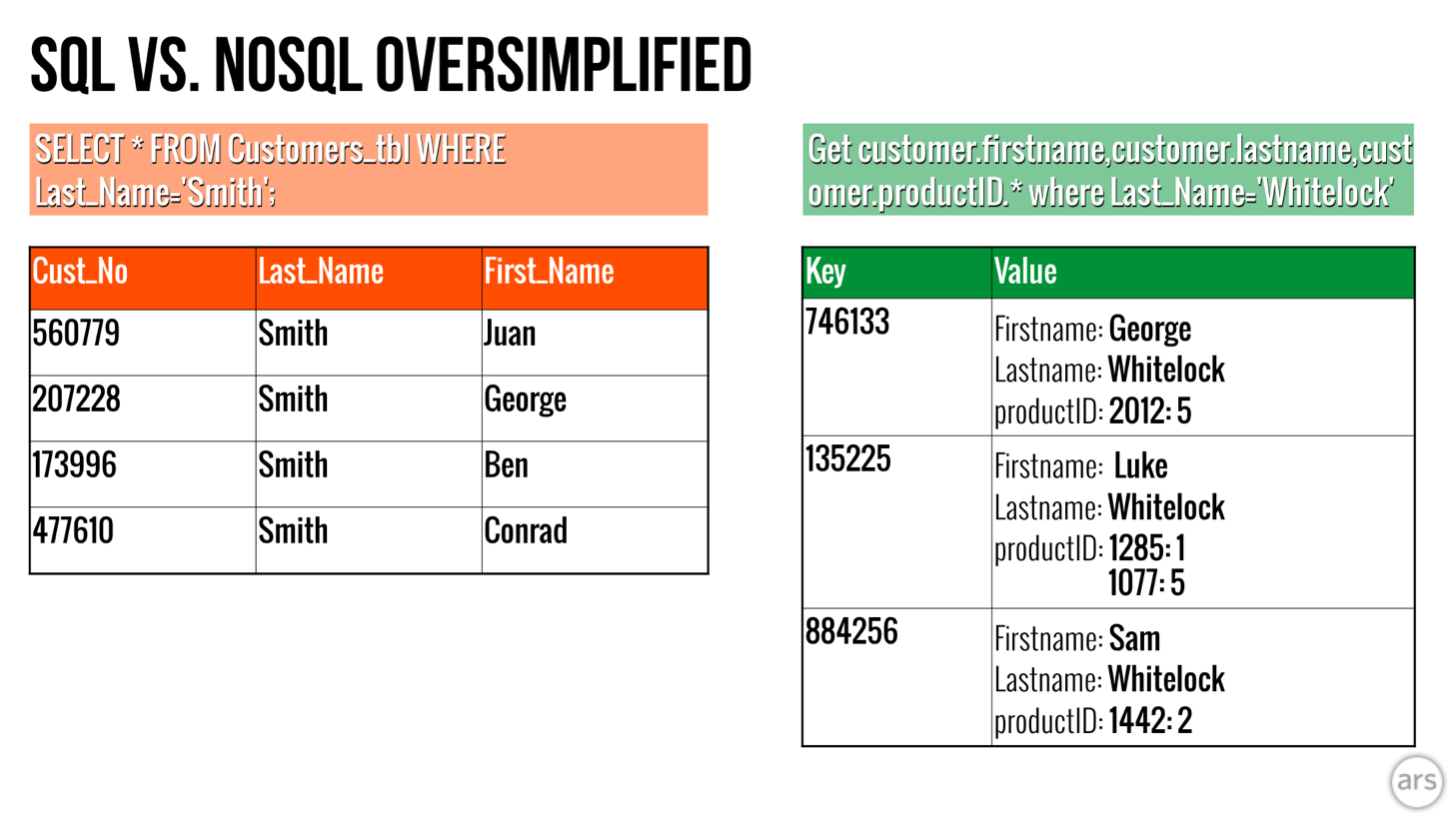


Figure - A simplified example of SQL vs Key Value Pair (NoSQL) databases.

**Pattern Identification**

**Generating & Communicating Actionable Insight**

**Ethical Ramifications**

There are many examples of Data Science techniques and methodologies being put to use for the good of the general public.

However, as with any set of tools, it is the wielder who determines how they are utilized and so the opposite is also true. Whether intentional or not as practicing Data Scientists we must be cognizant of the potential negative impact that our work can have both in our organization and in the wider community as well.

This is a problem that I feel is well covered in Cathy O’Neil’s ‘Weapons of Math Destruction’ in which she discusses a number of examples in which the ethical repercussions of Data Science are readily apparent. These models, which she refers to as “WMD’s” or “Weapons of Math Destruction” are those which seek to solve a valid problem such as predicting recidivism (the likelihood to reoffend) for prisoners, but which ultimately end up having unintended consequences. To continue with the recidivism example, O’Neil sites the use of factors such as race and income or proxies such as residential area and how they contribute to perpetuate a system of inequality, in this case that of racial divide. Often, as is the case here, these models also have the problem of being recursive; their own output consequently affects their input, leading to a progressively more polarizing output.

Over the course of conducting my analysis of crime in the Syracuse area this and other such examples were at the forefront of my mind. While the data that I was working with did not contain factors such as race or income level, I was conscious of the potential unintended consequences that basing too much of my analysis on geography may have.

Although it is inevitable that at some point, we are likely to make a mistake, simply being aware of the potential pitfalls of ethics in the data science world as well as the circumstances when such situations have previously arisen can help to reduce the likelihood of history repeating itself. On the other hand, when such problems do present themselves, taking responsibility for the problem at hand, moving swiftly to mitigate the damage, and sharing the outcome with others so that they may learn as well would be the responsible course of action. Expediency over ethics has no place within the realm of responsible Data Science.

<https://arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/>

<https://datasciencedegree.wisconsin.edu/blog/history-of-data-science/>

Weapons of Math Destruction