**Data Science: Transforming**

**Information into Understanding**

**Ryan French**

Applied Data Science

Syracuse University

School of Information Studies

Data Science is a 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 business 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 discussed 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 (UW). 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 collections of data quantified in exabytes as well as the methods and systems necessary to do so demonstrates the exponential growth of the field 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 the Master’s of Science in Applied Data Science program. 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**

Unless data is readily available, data collection and organization are often the first steps in the data science pipeline. 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 as IoT sensors, to collecting in-person survey responses from customers regarding their feedback on a product. With an endless number of data sources in a world where data is all around us, the avenues for data collection are essentially limitless.

After data collection comes the determination of how best to house said data for analysis. Over the course of the Advanced Database Administration Concepts & Database Management course offering 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.

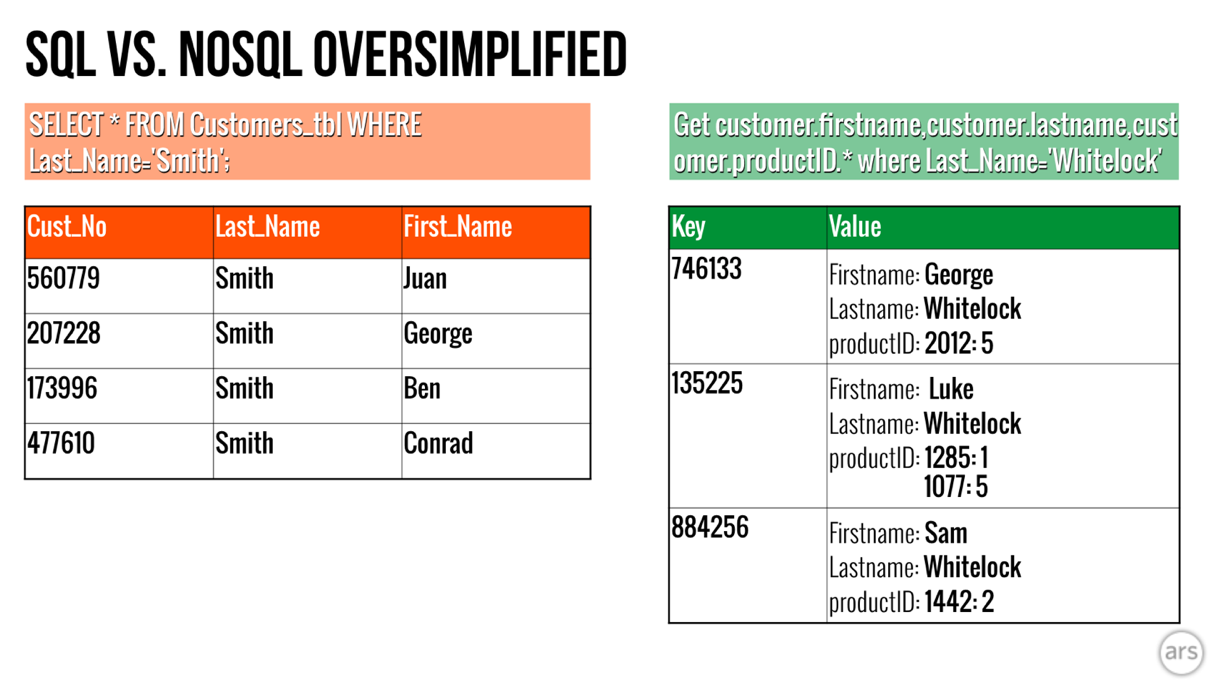


Figure 1 - A simplified example of SQL vs Key Value Pair (NoSQL) databases.

**Exploratory Data Analysis & Pattern Identification**

After collecting the necessary data, it is good practice examine the gathered data in a process commonly known as Exploratory Data Analysis. This can help us to better understand the context of our problem, what information we have, and what cleaning and/or feature engineering might be necessary in order to prepare our data for further analysis.

Typically, I begin this process by viewing some basic descriptors of the data in question such as its shape, data types, and whether there are null values. This information is helpful to frame the size and shape of the data in question, what it looks like, and how much cleaning and engineering might be required.

**A close up of text on a black background

Description automatically generated**

Figure 2 - Example Output of Data Exploration

From there, I find plotting a simple and highly effective way to generate insight about the distribution of the data and what variables might be most useful for further analysis. These plots often have the additional benefit of being useful for answering questions about the data and delivering high level analytical summaries.

A screenshot of a cell phone

Description automatically generatedA screenshot of a cell phone

Description automatically generated

Figure 3 - Example Digestible vs Uninterpretable Variables in Exploratory Plotting

For instance, consider a machine learning project that I completed for the Scripting for Data Analysis course which involved predicted highly scoring apps on an app store. In Figure 3, we can see the ‘Content Ratings’ variable which is readily interpretable and might be used for further analysis compared to the ‘Distribution of App Current Versions’ variable which is quite messy and contains many categories. If I was interested in using this variable for further analysis, I would have to heavily clean it first. On the other hand, it is readily apparent that there is a pattern of applications being rated predominantly ‘4+’.

Once satisfied with Exploratory Data Analysis, one can then proceed to data cleaning and feature engineering. This process involves steps such as:

* Handling missing data
* Removing bad data or outliers
* Discretizing or bucketing
* Creating dummy variables
* Removing unhelpful/unintelligible variables

To continue with the earlier example of the app store score prediction problem, some of the steps that I took included removing text-based values from the variable related to the app version and truncating each version to the first digit in its release cycle (version 3.1.2 becoming 3). Additionally, I created a dummy variable which served as an indicator of whether an app was highly rated (greater than or equal to 4) or not.

Following the completion of feature engineering and data cleaning, one is then ready to create statistical models or generate insight through other means such as plotting.

**Generating & Communicating Actionable Insight**

Although the process of conducting data science is an enjoyable and interesting exercise, it is often of little practical use to a business without the ability to synthesize meaningful insights from what is discovered. It is for this reason that the ability to both generate insights and communicate them in a way that is understandable and useful to the stakeholders in question is so important.

When selecting what insights to share I have noticed three primary factors to consider:

* The goal of the project (what the stakeholder is interested in) and how it can be accomplished (classification vs inference)
* The familiarity of the stakeholder with the subject matter
* The familiarity of the stakeholder with the data science techniques utilized

By considering these three factors, one can be certain that they are presenting the relevant information in the right context with the right degree of detail. For instance, recently for work I conducted some forecasting analysis for web traffic in the coming quarters utilizing a Linear Regression model for our VP of Sales. Because he knew specifically what he wanted (the predicted numbers), was quite familiar with the subject matter, but did not have much time to spend on the discussion, I focused on explaining what was predicted and to what degree of certainty rather than the minutiae of how I arrived at the optimum model. While I included this information in the report which I sent to him later, this was the most efficient way to utilize his time; providing him only necessary information while still providing him the option to look into the methodology later if he was so inclined.

Although I selected a Linear Regression for this use case in particular, there are a variety of models that might have worked in this case. When it comes to selecting the best model for a given problem, the tradeoff of interpretability vs accuracy is important to keep in mind as some models are better certain to one outcome than the other.

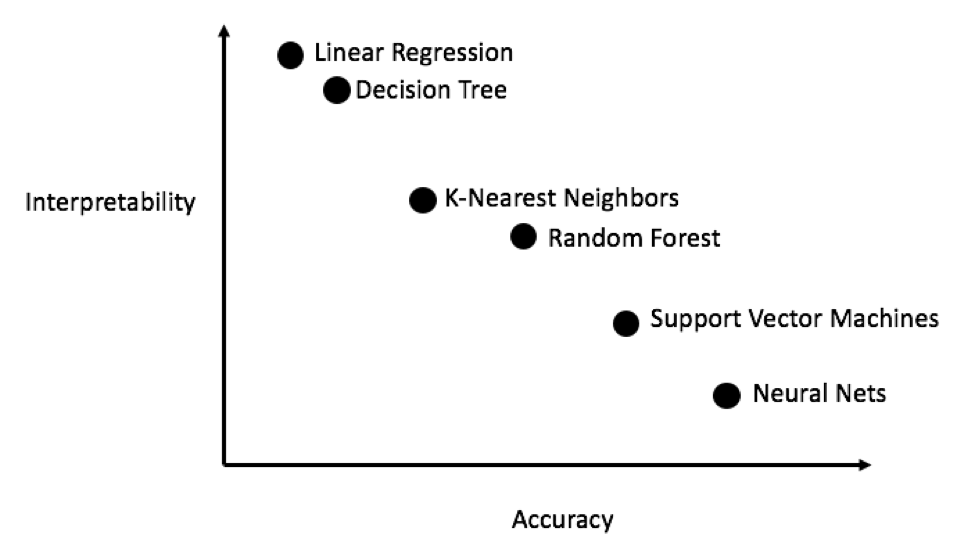


Figure 4 - The Interpretability vs Accuracy Model Tradeoff

Due to this trade off, one model that I frequently utilize is the Random Forest due to its ability to both provide robust accuracy and offer the ability to generate insight as to how decisions are being made with relative ease. One good example where these two traits are of great importance can be seen in my undergraduate honors capstone project where I conducted an analysis of crimes which occurred in the city of Syracuse and whether or not they would result in an arrest.

Of the models I built to predict whether or not an arrest would occur based on a given situation, I found that the Support Vector model that I utilized was performing marginally better than the Random Forest model which I was comparing it to. However, due to the sensitive nature of the problem at hand and the importance of understanding how the prediction of an arrest occurring might take place, I would have recommended the use of the Random Forest model due to its accuracy being relatively close and the ability to use it for more readily interpretable insight.

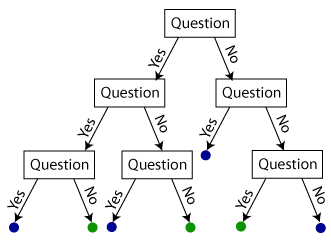
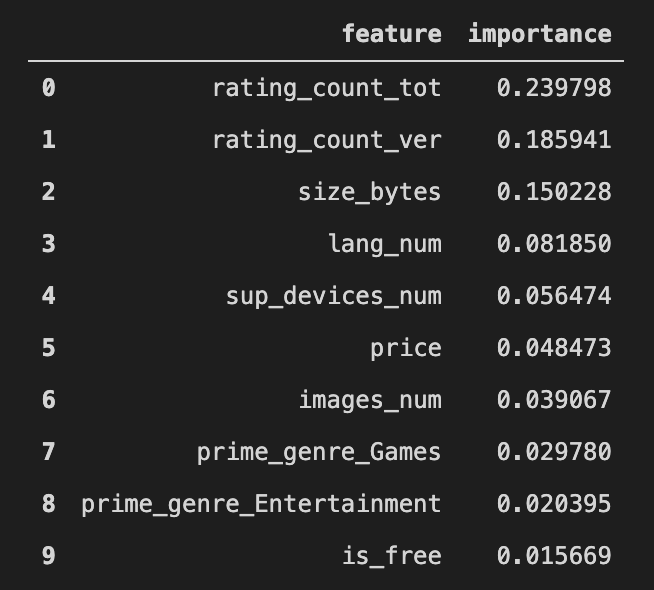
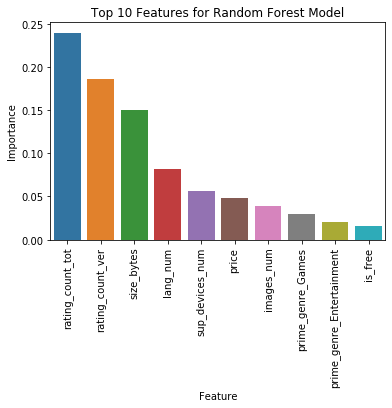


Figure - Example Output of a Tree Within a Random Forest

In regard to presenting the information itself, over the course of my classes and internships I have found that a visual approach is often the most effective way to relay information and insights. By presenting said results visually, it makes them more attractive, easier to digest, and it is often possible to pack more information into a smaller form factor. However, that is not to say that text-based information does not have its place; including text within visualizations can help to highlight or better explain key points and including more in-depth printouts of the data and insights in an appendix or footer allows interested parties to examine them later.

Figure 6 - A Textual vs Visual Approach to Information Presentation



Over the course of the Information Visualization course I was able to both experiment with a wide variety of unique and interesting visualization methodologies that I had not seen previously while simultaneously honing the principles of good visual design and how to apply them to practices with which I was already familiar. The culmination of our semester of work took the form of a poster displaying a variety of visualizations on our subject of choice.

As a video game enthusiast, I took this opportunity to tell the story of the video game industry over the last few decades in a format that was both sound as far as visual principles and also suited the subject matter in its retro-digital art and color scheme.

**A picture containing newspaper, meter, parking, machine

Description automatically generated**

Figure 7 - 'Digitizing the Gaming Industry', Information Visualization Final Project

**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; it is also possible for data science practices and models to perpetuate biases, impact finances, and have a dramatic and concrete impact on the lives of others. 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.

The problem of ethics in data science is one that I feel is well covered in Cathy O’Neil’s *Weapons of Math Destruction: How Big Data Increases Inequality and*

*Threatens Democracy* 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 my analysis could have if this was the case.

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.

**Works Cited**

O'Neil, Cathy. *Weapons of Math Destruction: How Big Data Increases Inequality and*

*Threatens Democracy.* Penguin Books, 2018.

UW Data Science Team. “A Modern History of Data Science.” *University of Wisconsin Data Science Degree*, 10 July 2017, datasciencedegree.wisconsin.edu/blog/history-of-data-science/.

Vaas, Lisa. “To SQL or NoSQL? That's the Database Question.” *Ars Technica*, 30 Mar. 2016, arstechnica.com/information-technology/2016/03/to-sql-or-nosql-thats-the-database-question/.