The Foundation of Any Data Analysis

Big Data

Big Data. You've probably heard of it. Do you ever wonder how Facebook, Google, YouTube, and all other websites are able to really hone in on exactly the specific type of advertisement that you would be interested in? This is exactly how it's done. All through the data created by your usage.

This buzz word is thrown around so frequently in the media that it is often associated with just having huge sets of data to make decisions. Its applications can extend beyond this into the most basic uses of data of just learning about a specific topic or even solving problems. Much of the focus stems from the predictive power that many businesses and data technology companies utilize to make accurate choices and monetize off them. However, even before someone, let alone a business, tries to use the data, proper formatting of the data is necessary. Otherwise, the data is useless or may just misrepresent the area that is researched. In order for Big Data to be useful, a framework must be made in how to structure it.

Background to Data Wrangling

Often times, data that is collected has missing values or possibly laid out in a way that does not give any insights. Data wrangling describes this process of cleaning up the data acquired so that people may efficiently use the data and understand it to make useful analysis on it. For instance, for many businesses, a lot of data comes from their business intelligence department, and when it collects information, the data may not be in order, some values may be missing, maybe even have some typos. All these things are what data wrangling aims to make readable.

A simple example of this can be shown below:

```
## Loading required package: readr
```

```
# Showing how to change the class of a column to make it more efficient when analyzing.
nba_stats <- read_csv("nba2017-player-statistics.csv",
                      col_types = cols(Position = col_factor(NULL)))
class(nba_stats$Position)
```

```
## [1] "factor"
```

Data Wrangling as a Service

The task of data wrangling may seem like a job that can be done in-house, but as data accumulates and harder to manage, many companies outsource this type of activity to professional organizations that specialize in this exact software to wrangle data for them, such as Trifacta. This company not only offers firms the ability to efficiently make data management easy, but also it enables them to focus more on their own business by making decisions on their data analysis directly rather than having to wait for the data to be ready. With more data on the rise with the expansion of cloud computing services, the need for such data wrangling services will be essential. The types of businesses that will arise from the advent of this technology is endless. More metrics will be used to measure performance for a firm and as data get more complex, more intelligent software is needed to take that raw data and convert it into something readable and usable for company use, as basic software would no longer be equipped to handle such operations.

Future of Data Wrangling

This field is gaining so much traction that Microsoft is looking to get into the market by offering its Pendleton software. The purpose for this tool is exactly how data wrangling is described: data preparation and cleaning. For instance, it allows a simpler way to format columns, handling missing values, or separating values into different columns. To provide a better edge to its competitors, Microsoft has included a deep learning and Artificial Intelligence capability to allow different technologies to seamlessly work together when transferring data around amongst the various platforms.

Trends Overall in Data Science

As technology continues on its path of growth, the media almost always cover the big giants in the industry: Facebook, Google, Intel, and Apple to name a few. Who it fails to mention are the data science companies that help power these big-name brands. They also deserve this highly praised recognition as they give these giants the clean data that allows them to continue running their highly profitable business. This trend will begin to intersect companies, like Trifacta, Oracle, Salesforce, and Marketo, as Big Data may as well be the underlying value that fuels modern day firms.

References

https://www.forbes.com/sites/gilpress/2016/03/14/top-10-hot-big-data-technologies/#6788ca9265d7

https://www.forbes.com/sites/bernardmarr/2017/08/14/want-to-use-big-data-why-not-start-via-google-facebook-amazon-etc/#2f3ba04f3d5d

https://www.forbes.com/sites/tiriasresearch/2017/10/04/microsoft-machine-learning-and-data-wrangling-ml-leverages-business-intelligence-for-b2b/#5fec98b57640

https://www.trifacta.com/products/

https://www.rstudio.com/resources/webinars/data-wrangling-with-r-and-rstudio/

https://www.tableau.com/about/blog/2016/12/top-10-bi-trends-2017-63208

https://www.inc.com/bill-carmody/former-facebook-data-scientist-shares-how-to-wrangle-your-data .html

http://www.zdnet.com/article/microsoft-aims-to-take-the-work-out-of-data-wrangling-with-coming-pendleton-tool/