Quantitative Big Data Analysis Limitations: When Numbers Fail to Tell the Full Story!

Author, Michael Baron

A Data Science Foundation White Paper

March 2020

----

www.datascience.foundation

It’s no secret that Data Analytics is significantly easier to handle when it focuses on quantitative data analysis methods rather than qualitative ones. When dealing with Big Data, challenges of the qualitative data collection and mining approaches are becoming particularly transparent. The Qualitative Data Analysis methods and tools tend to be expensive and complex to implement, while accuracy of the analysis may nevertheless be compromised by a wide range of factors (data validity, data interpretation, context interpretation etc.). However, Quantitative analysis is no panacea from mistakes and discrepancies. The purpose of this White Paper is to consider contemporary challenges of the Quantitative Big Data analysis activities.

Quantitative Analysis is usually defined as analysis by the means of ‘’complex mathematical and statistical modelling’’. Therefore, this definition looks way beyond trivial number-crunching and also incorporates mining data sets for patterns and correlations. In other words, it can be used for any project big or small as long as the data can be represented via numerical values. Some of the Advantages of using Quantitative Analysis Methods in Data Analytics are obvious. They are not only cost effective and generally easier to implement as opposed to the qualitative methods and tools, but also tend to produce clear output that appears to be easy to validate (by using established analysis tools and processes where all the steps can be confirmed), automate (AI systems are particularly good at number-crunching) and classify (classifying ‘’numbers’’ is easier than classifying non-numerical values). However, the more we get ourselves engaged into the world of the Big Data, the greater shadows of concern emerge about even some of the most reliable Quantitative Methods.

The critical limitations to consider prior to employment of Quantitative Analysis Methods for Big Data analysis are:

- Complexity of the Big Data Environments
- Big Data Life Cycle Sensitivity
- Data Labelling
- Deregulation of the Big Data Standards
- Interdependency of the Data Values

Complexity of the Big Data Environments

Quantitative analysis methods work best in single-format environments. A single-format environment ensures that the data values are consistent and well-defined rather than ‘’open to interpretation’’. However, it is the very complexity that makes the data ‘’Big’’. Even if the values are compatible, analysts should also consider how these values were collected, additional factors that impact the data environment (those will vary depending on the data formats) and consistency of the data collection tools and methods used.

Big Data Life Cycle Sensitivity

Data life-cycles are getting shorter and shorter. On top of the generic principles of the diminishing value of data (e.g. 90/90), There are also significant concerns on bringing current and historic Big Data to a common denominator. Furthermore, the data environments (as evident from the discussion above) are changing very fast and so do the consequent data values. The number of data parameters also keeps increasing. Traditionally, number of the quantitative data parameters was usually limited to *8-10* even when dealing with the most complex industries cases (financial data etc.). Today, many of the Big Data Analytics projects require creation of a far greater range of the parameters.

Once the data becomes outdated, it has to be removed from the respectful data set. Sometimes, the entire data sets have to be removed from the analytics environment. However*, removing entire data sets is not as much of a logistics challenge as going through all of the data sets and religiously reviewing all of the data that may/many not fit under the newly emerging parameters/requirements*. It may turn out that it is difficult to do so consistently for all of the data sets as in cases of the Big Data, each of the data sets may be having own initial formatting (prior to being converted into a common shared format).

Data Labelling

Data Labelling makes data sorting easier and more consistent, particularly when dealing with multiple data sets. But what should those labels be? Should the matching data be grouped on the basis of matching ALL of the parameters? Or a single parameter? Or should the grouping be data set-based?

In other words, the larger and the more complex the data environment is, the harder it is to handle the labelling accurately. Studies that use generic labelling as the basis for the data processing should be taken with a grain of salt. Generic labels make data sorting easier but significantly less accurate. It can be compared to putting all of the poultry products on to a single shelf and assuming that other than all of the items being classified as ‘’poultry’’, no further sorting is required!

Deregulation of the Big Data Standards

Increasing complexity of the data sets (common trend with Big Data Analytics) also leads to deregulation of the Big Data Standards. From a technical angle, Analysts would love to have consistent standards to follow as it would make the entire data analytics process more consistent. Many efforts have been made to develop such standards to be shared across industries/projects. For example, the IEEE has come up with some Big Data Standards. At first glance, at least some of the standards appear to be quite comprehensive (some others are still ‘’under development’’). However, it turns out that following those standards and applying them consistently throughout the analytics projects is not always possible. Furthermore, for the standards to become a norm, they have to first become accepted across the globe. Given that this is private companies/industries domain rather than a governments’ one – it is not going to happen in foreseeable future.

Even within a single organisation, implementation of consistent Big Data Standards requires continuous monitoring. This can rarely be achieved through automation alone. As the author has pointed out in his recent Data Science White Paper on Data Processing Automation Challenges, automated monitoring is often unreliable and ongoing ''manual'' reviews of the Big Data Standards (along with the consequent testing/monitoring how these standards are being followed) will certainly increase complexity of the Big Data projects.

Interdependency of the Data Values

Interdependency of the Data Values refers to scenarios where changes to a single data value is going to impact other value(s) considered. The Data Value changes may even cause a ''Chain Reaction''. It is particularly important to note Implicit chain reactions where interdependency between the values is indirect rather than direct.

All it takes for a quantitative Big Data Analysis to be compromised is a single value inaccuracy/unsolicited or unaccounted change – and the Butterfly Effect will take place. The Bigger the Data is (and with some projects, data comes in 100s of formats), the higher the probability of such discrepancies is going to be. Understanding the interdependency is therefore one of the keys to being able to analyse the data successfully, but then again – errors may still happen!

If Numbers Fail to tell the Full Story, Who Can?

Based on the discussion of the Quantitative Big Data Analysis limitations above, it is easy to start wondering whether the quantitative analysis methods do work?

They certainly do! Limitations and challenges are no reason not to employ the quantitative Big Data Analysis methods but based on the authors' recent experiences with technology-driven Quantitative Data Analytics projects – it is always good to supplement testing and validation procedures with a Qualitative touch. It is the balancing of the Data Analysis methods that secures success of the DA projects.

It should also be noted that to a large extent, the problems occur NOT because of the limitations of the quantitative data analysis methods but because of our failure to employ these methods properly. Before we question the methods, we should ask ourselves: ''Do we understand the Data Analysis process that we are getting engaged into clearly or do we simply rely on the tools to do the job?'' Usually, Quantitative Data Analysis studies incorporate several stages. With each of the stages completed, we must review ''work in progress'' and validate it prior to moving on. It is the Butterfly Effect that compromises validity of the analysis most! Making errors is inevitable and having at least some discrepancies throughout the data analysis processes is unavoidable. It is our ability to fix those mistakes that makes the Data Analysis projects a success!

DataScience®
foundation
MAKE BETTER DECISIONS

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:admin@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation