Data Analytics Integrity: Challenges to Implementation of the Automated Data Collection Processes

Author, Michael Baron

A Data Science Foundation White Paper

March 2020

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: admin@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

In recent months, my company (Baron Consulting) has been proactively involved in setting up Data Collection Systems for a range of Private and Public Organisations we are servicing. Some of the data collection challenges have already been discussed in our recent *Raw Data Collection 2020: Principles and Challenges* White Paper. While the RDC (Raw Data Collection) paper analysed the current state of the everchanging Data Collection requirements, it did not have the scope to address technicalities of the RDC processes along with the specific Data Collection tools and methods. The purpose of this paper is to fill the void by looking into implementation of the automated data collection processes.

Automated Data Collection Processes: Cutting Down Cost of the Data Collection

The cost of the data collection tools appears to be stable and within the affordability limits. Furthermore, it is not hard to establish the costs required to acquire these tools and come up with accurate costing projections prior to commencement of the data collection processes. Furthermore, even the ''freebies'' and low-cost apps can usually acquire the ''live''/stored data to be managed fairly efficiently. Therefore, it is the ''cost of labour'' that increases overall cost of the data acquisition dramatically. In ideal scenarios, the data collection processes should be automated to the point of being tool-driven rather than person-driven. Once set-up, they would be able not only to collect the data required, but even to handle initial (basic) sorting of the Data.

On top of collecting ''live'' data, the automated processes can easily extract data from pre-existing documents. Needless to say, the document data extraction is cheaper, faster and more accurate when automated as opposed to human-driven alternatives! Automated processes do not make ''typos'' and do not deviate from pre-set parameters. Therefore, given a choice to automate or not to automate, there appear to be very few if any doubts for the data analysts to pounder over.

Automation Challenge of the Decade: Eliminating ''Human Touch'' From the Data Collection Processes

Despite obvious benefits of the automation outlined above, it is also linked to a number of challenges. With rapid development of the AI, it may appear that the automated processes have capacity to fully replace ''human involvement'' with the Data Collection. The AI proponents claim that contemporary machines act and react like humans. Unfortunately, with the Data Collection processes and tools, there are still some ''grey areas'' where the AI systems need to be administered with great care.

More specifically, the are 5 processes that appear to be the greatest challenges of all when implementing Automated Data Collection processes:

- *Pattern Recognition*
- *Data Verification*
- *AI Self-Awareness Programming Discrepancies*
- *Multi-Level Automation*
- *Skill Shortages*

Pattern Recognition

Automated data collection is based on a range of pre-set parameters. Optimal accuracy and relevance of the data can only be achieved via simultaneous application of multiple parameters, including pattern-based parameters. Patterns are far more difficult to establish and manage than standard quantitative data as they often incorporate qualitative factors. Traditionally, the patterns were to be identified by ''humans''. An experienced data analyst can continuously monitor and adjust the parameters as required. However, automated processes are harder to adjust. Furthermore, customizing data collection applications for the purpose of Pattern Recognition is often difficult.

One example of Pattern Recognition challenges is setting up automated data collection systems to monitor customer satisfaction levels. Many of the patterns and trends are difficult to define. Customer ''happiness'' can seldom be expressed through numerical values. Quantitative methods such as Likert Scale may be effective to measure distance travelled, or money spent but identifying patterns for measuring emotions-driven factors and finding ways to ''program'' these patterns for the applications to use could be problematic.

Data Verification

Automated accuracy and relevance verifications of data are usually multi-level processes. Data verification problems may start occurring at the initial stage of identifying validity of the data sources. For instance, our (Baron Consulting) projects focus predominantly on User-Centred Design (UCD). In order to carry out the data collection, parameters for the valid users need to be not only identified but also enforced throughout the collection process. When dealing with perse user population, automated systems sometimes struggle to do so. If the data is collected without users' input, it is not clear how the automated systems can establish their age, background, level of technical expertise, task range required etc., since even the most intelligent systems can not read human minds or to provide accurate observational assessments.

With the user-driven inputs – validation may still be a problem as automated systems are not able to verify data that the users provide. If the users provide (knowingly or unknowingly) false information, automated data collection systems often fail to note the lack of authenticity and disqualify corrupt data from the data banks.

Likewise, it is often difficult to establish and implement consistent data verification processes across a range of data sources – particularly if these sources belong to different data channels. Random checks

and tests may reduce the risks of collecting large amounts of invalid data but even if a small proportion of the data sets is based on false/invalid data, the consequent stages of the data mining (turning raw data into information & information into knowledge) are likely to be corrupted.

AI Self-Awareness Programming Discrepancies

Contemporary AI tools can be developed with in-built self-programming capacities. Unfortunately, as far as data collection tools are concerned, such automated adjustments to the tools and processes are difficult to implement. Majority of the AI systems can be classified into 4 types*: reactive machines, limited memory, theory of mind and awareness*. In order to maintain integrity of the data collections, the data collection tools need to be capable of the self-awareness-driven adjustments.

The Self-awareness capacity requires (ideally) data collection tools to form representations about themselves. Unfortunately, as far as the data collections are concerned, the mainstream data collection tools appear to be lacking the self-awareness capacity for at least some of the tasks they are required to undertake. While the AI systems are never easy to utilise, self-awareness is particularly challenging to implement with the data collection tasks involving user/object validation, seasonal or conditional adjustments and parameter deviations.

Skill Shortages

No matter how automated the data collection processes are, there are always people (data analysts) behind the automation. AI data collection systems are often tricky to manage and ‘’the human factor’’ is particularly critical during initial stages of the systems’ implementation as setting up correct parameters, testing and validation all require proactive involvement by the subject matter experts. Also, given complexity of the systems, generic knowledge of automated data collection processes may not be sufficient for completing the tasks. AI-driven platforms and applications are perse from one another, so mastering new ones can not be achieved overnight. Furthermore, many data collection processes are industry-sensitive so require involvement of data professionals who have already had sufficient exposure to projects within the very same industry. Baron Consulting for instance has been handling Data Collection and Analysis for several Universities and Colleges but would we have to implement and manage a Data Collection process for a financial institution, we would be faced with significant challenges and have to ‘’upskill’’ ourselves first!

As discussed above (see the *AI Self-Awareness Programming Discrepancies* section of the paper), Data Collection systems require high level of expertise. Given the ever-growing demand for the experts, significant shortage of skilled and experienced specialists is transparent. Head-hunting Data Analytics gurus requires not only ‘’expanded’’ budget but even more importantly – very good understand of the exact project requirements.

Another dimension to address is timing of the jobs’ appointments. Setting up automated data collection

processes should be planned with understanding that they are often multi-stage tasks and with large projects in particular – different stages require involvement of different specialists (or may even involve engagement of a number of different consultancies/organisations). Finding the right people is only half-of-the-job done. It is also essential to ensure that these specialists are available ‘’ad hock’’ if required.

To sum up, in a ‘’perfect world’’, automation of the data collection processes is an optimal solution to all the data collection needs. In reality, the automation should definitely be considered for implementation and over the time, the AI systems are obviously very likely to minimize the human touch required. However, the 5 risks/challenges outlined above should be acknowledged and addressed. With some of the data collection projects, automation may appear to be lucrative...yet not feasible to implement!

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