At the same time as regulators are focusing attention on data quality and data risk management processes, the complexity involved in managing data risks has become greater than ever.

The growth in data volumes needed for regulations and capital rules such as MiFID and FRTB, as well as the additional complexity in processing that data, such as mapping transaction data to risk factors and proxying non-modelable risk factors, has resulted in increased organizational exposure to incomplete, inconsistent or inaccurate data. However, these data risks are small when compared to the data risks coming from the adoption of AI. New risks such as the use of biased data, poisoned data (malicious data injected in the model), and dirty data can have severe legal and reputational ramifications.

A very illustrative example of the use of dirty data, although from outside Financial Services, is the use of Predictive Policing Systems that forecast criminal activity and allocate resources accordingly. A report from the AI Now Institute shows that in numerous cases these models are built using data from periods with flawed, racially biased, and sometimes unlawful practices. If AI learns from dirty data, predictions are going to display the same behavior.

A recent example from our industry is gender bias in the Apple Card algorithm. A male user was given a credit limit twenty times higher than his wife despite her higher credit score. The husband’s tweet went viral and resulted in a pending investigation over claims of discrimination.

One can assume that removal of gender in training of the algorithm will resolve gender bias. However, it is rarely that simple. Feature engineering is often performed to improve the predictive power of models. Even though gender is removed, other features and/or combinations thereof often can predict gender and therefore result in models still containing gender biases. As these types of data risks are relatively new, many data scientists are not yet aware. A complicating factor is that data science is a vastly growing field and most practitioners do not yet have sufficient experience to deal with these ‘nuances’.

If social media, and unstructured data in general, is used to train ML algorithms, data risks multiply again. From social media insights relating to gender, religion, political inclination, sexual identity can be easily derived and used as features in models without the data scientist’s knowledge, especially if deep neural networks are used. For example, from an image it is easy to derive the ethnicity of an individual, and if these images are positive for one group, and negative for another, issues may arise. But other risks such as cyber attacks, data poisoning and fake news become relevant as well.
How should organizations manage these data risks?

Although data volumes are larger, with more relationships, and less structure, the same principles apply when managing these new data risks. Organizations need to put in place data quality frameworks (See Figure 1. Data Quality Framework).

Organizations must outline a data quality policy that establishes clear data objectives, for example, data needs to be ethical. The policy needs to define data quality and target levels, and put in place data governance, including processes and procedures, responsibilities and data ownership. The next step is identifying the critical data elements, the risks (errors, gaps, bias) in that data, the flows needed to process and distribute that data, the risks in the data flows, and the controls that are needed to mitigate risks identified.

The important thing is that organizations should be aware of the different data risks that can impact model outcomes. Once risks are identified, the organization can put in place controls to mitigate. For these new types of data risks, the process is not different, only the controls may be a bit more advanced, e.g. the controls may require the use of AI and even Machine Learning (ML). Sometimes the models that the data feeds into are used to test for fairness and/or attacks, thereby blurring the line between data risks and model risks.

To illustrate how ML can be used for controls, there are many standard open-source libraries such as AI Fairness 360 and FairML that can be used to test for fairness in datasets and models, and to mitigate bias in those datasets and models.

For cyber attacks, such as data poisoning and evasion attacks, anomaly detection, the use of multiple data sets based on time windows, and the impact of newly added data may be used to identify issues, although these approaches are not 100% accurate as it may be very difficult to distinguish malicious from normal data.

As mitigating data risks and producing high quality data becomes more difficult, this is the time for organizations to put a data quality framework in place. To gain competitive advantage, improve efficiency and reduce overall cost, organizations can adopt a Data Quality Intelligence approach that provides further insights in data quality, processes, and operations, allowing the organization to move to a continuous improvement model (See Figure 2. Data Quality Discovery Process).
As easily 80% of effort for model development goes into data preparation and management, the impact of process improvements is significant, not only on cost, but also on turnaround times. Without a data quality framework and Data Quality Intelligence, organizations will have difficulty keeping up with the speed of innovation and will lose competitive advantages along the way.

**Figure 2: Data Quality Discovery Process**

---

**sources**


---

**author**

**Boyke Baboelal**

Boyke Baboelal is Strategic Solutions Director Americas at Asset Control, where he ensures the company maintains its position as the industry’s best-informed and most flexible partner for financial data management. Boyke brings 15 years of risk and data management knowledge to Asset Control. With his financial engineering background and hands-on experience, Boyke understands the importance and value of good data. Boyke holds a master’s degree in Econometrics, a postgraduate degree in IT and an MBA from Rotterdam School of Management.