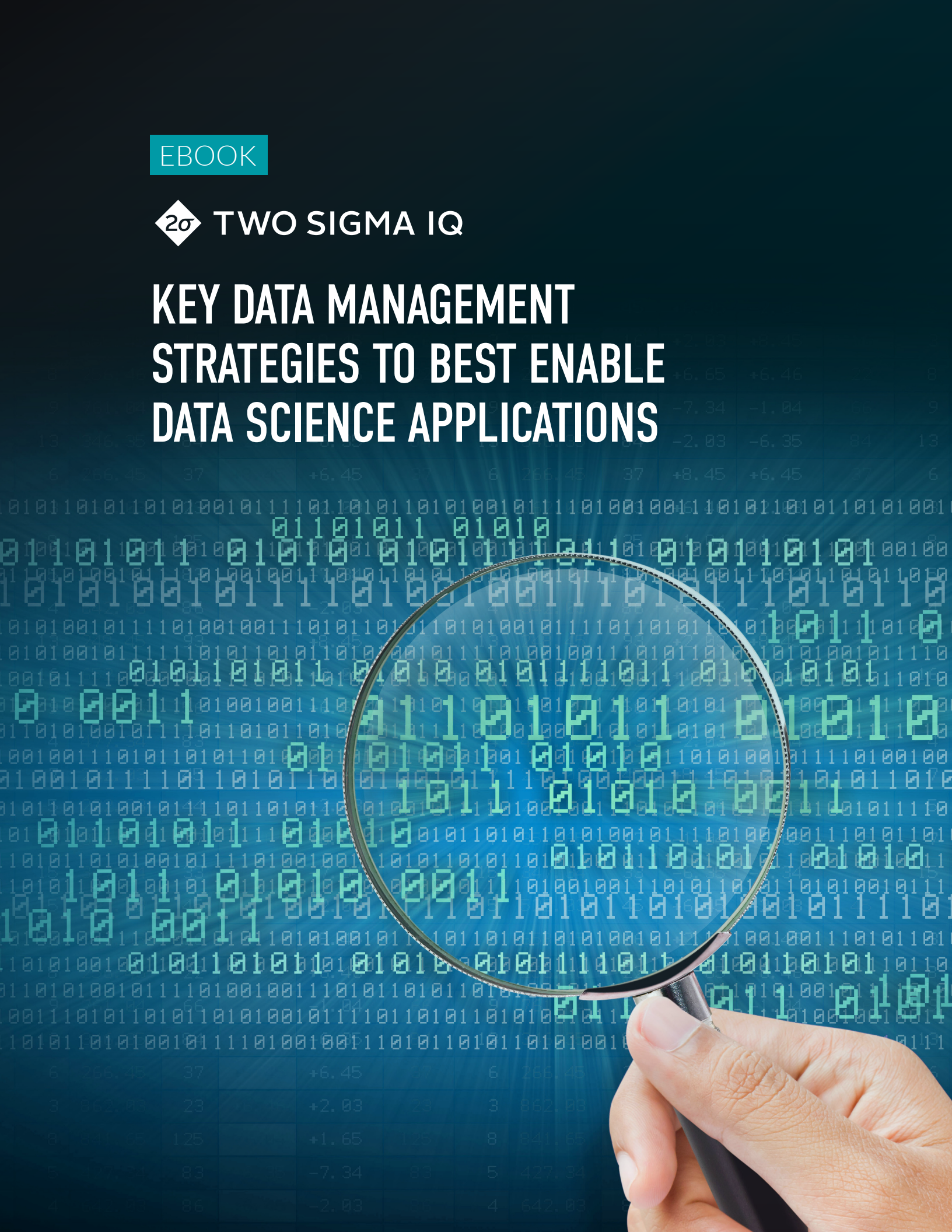
EBOOK

TWO SIGMA IQ

# KEY DATA MANAGEMENT STRATEGIES TO BEST ENABLE DATA SCIENCE APPLICATIONS



# INTRODUCTIONS

The explosion of data, recent advances in new technology and increasingly competitive market conditions are driving more and more insurance companies to investigate and invest in the application of data science to help grow their businesses. Companies are looking to better harness their own internal data and are seeking unique new sources of third-party data to help gain a competitive advantage. The application of data science to various parts of the insurance value chain has begun to prove worthwhile for some, but the full enterprise potential has yet to be realized as many projects are done in silos or as one-offs. Many data science projects are limited or fail because of issues with data quality and quantity.

In order to make the most of data science techniques, large amounts of data are needed, as larger volumes of data and more varieties of types of data sets help to produce more interesting insights and results. Insurance carriers are interested in using satellite imagery data, data from drones, telematics data and other data sets that can help them better understand their customer base and the risks they are asked to underwrite. With a sharp increase in the volume of data comes challenges in managing it all. Data is often trapped, not centralized or not easily accessible. Old or historical data can lack consistency or not be well-defined, and data quality often varies. Without an effective strategy to manage data, the success of machine learning and other techniques will be limited. Poor data management can limit and even hurt the success of analytics and data science.

To make the most out of data science applications, insurers need to be thoughtful and strategic in managing their data. Key elements of an effective strategy should include early alignment of data scientist and business user requirements; saving all data; tracking and recording the provenance of data; ease of accessibility; and ensuring data quality. An effective data management strategy will lead to deeper and more accurate insights from machine learning and other applications and, in turn, better risk and overall business decision-making.

# 5 KEY ELEMENTS OF AN EFFECTIVE DATA MANAGEMENT STRATEGY

## 1 ENGAGE DATA SCIENTISTS IN THE PLANNING STAGE

Insurers need to take a strategic approach to managing their data and applying data science. Bringing the right expertise early into the planning stages will ensure that the right types of data, analysis and insights are utilized at the right points within a workflow. When considering the use of data science applications in the insurance underwriting process, for example, a data scientist with industry experience should be equally part of the planning stage as an underwriter. First having a general understanding of the purpose of a particular project or application, such as auto-ingesting third-party data sources, will allow the data scientist to determine what sources of data are appropriate and how best the data can be used. The data scientist can help guide data selection, how it can be best used and design potential data science solutions to solve the particular business problem.

Insurance carriers are becoming increasingly interested in utilizing third-party data sources to help improve the underwriting process, especially around submission ingestion and risk assessment. However, bringing new data sources into a workflow is more complicated than it seems and needs to be strategically planned. Most new external and even some existing internal data sets are rarely in useful formats. A data scientist can analyze and help plan out how to ensure that the right data is available in the appropriate format. They can also advise as to the best points within the workflow where the data sets can be applied.

When using data science techniques to help solve workflow issues, such as those in the underwriting process, bringing a data scientist in at the planning stages will help to ensure the right data is available in the right format at the right point in the process. Applying their expertise later in the design/redesign process hinders the ability of the data scientist to make the most effective improvements to the underwriting workflow.


# KEEP ALL DATA, NEVER DELETE DATA

Data science applications yield the best results from large amounts of data - the more data the better. If an insurer wants to take advantage of machine learning applications, it is wise to save all of its data and never delete any of it. A data set that seems to have no use today may become useful later, especially since risk profiles and lines of business change over time. Having the ability to utilize not only current internal system data, but also historical decision-making and transactional data can help insurers analyze why and how a risk was underwritten. Saving all of the information involved in an underwriting decision can allow a carrier to understand what may have led to a claim. Having the ability to analyze the steps that were taken in underwriting specific risks can provide transparency as to any errors that may have been made in the process or help to better understand if any additional data and analysis could have been used in the underwriting process or risk analysis.

Unfortunately, the way many core systems are designed today, data is not captured in the most usable format. Insurance companies have typically been focused on only capturing data for reporting purposes and not for modeling purposes. Typical policy administration systems don't capture records of change in enough granularity to support data science techniques mainly due to the perceived lack of need since qualified data scientists have been scarce and data storage and converting information into usable form has historically been expensive. However, advanced technology and growing interest in more data has changed the demands and cost structure.

Saving large quantities of data today is relatively inexpensive and simple compared to the past due to the low cost and abundant and flexible capacity cloud storage provides. The tools now available in the cloud and the ability to scale computational and graphics processing unit resources up or down enable data scientists to now analyze data at a scale that was prohibitively expensive in the past. Cloud vendors can tier storage costs by level of availability. Less accessible or less needed data is often less expensive to store. With the growing interest in machine learning and artificial intelligence applications, the need for large volumes of data and creating a true system of record is more important than in the past. Having the ability to save historical internal and external third-party data will help enable carriers to better apply data science applications to improve their underwriting process by allowing them to troubleshoot potential problems and errors and add additional insights and tools. Data is becoming an increasingly valuable asset and should be preserved.


# ENSURE AND TRACK PROVENANCE OF DATA

In addition to having access to large volumes of data, in order to create a reliable system of record, insurers should track the provenance of their data to ensure its veracity. In other words understanding and having a record of the where, what and when of each data set/point is important when applying data science. Information around internal and external data transformations is extremely helpful for data science research purposes. Knowing where data came from, its source and format, and the transitions it went through helps to answer how a data set fits into a data ecosystem and what features and fields it contains.

Knowing the source of data helps with determining its validity and reliability - data becomes more trustworthy and reliable. The more reliable a data set, the better the insights that can be drawn. In the case of an insurer interested in applying data science to the underwriting process, more informed underwriting decisions can be made. If data sources are known, a full audit trail can be created and errors can be more easily traced back to causes. Data bias issues can also be found and internal data can be checked and validated against external data which will ultimately give an underwriter more confidence in data science application results.

Creating a data catalog to store all of the provenance of data is beneficial. Challenges do exist when recording data provenance and creating a data catalog. Data sets are unique and some data sets are harder to track than others. In addition legacy core systems don't typically have data catalogs as high volumes of data and tracking data can slow a system down. In order to best capture data provenance, a modern cloud computing architecture with a data catalog is a must.

> Investments in data collection and curation capabilities will be a key differentiating factor.
>
> ***Swiss Re Institute, No. 5/2020,*** Machine-intelligence in insurance: insights for end-to-end enterprise transformation

# 4 MAKE DATA EASILY ACCESSIBLE, ESPECIALLY FOR RESEARCH

Having large amounts of reliable and cataloged data is of greatest value when it is easily accessible and centralized. Data scientists, underwriters and other business users will best benefit from data and the insights it produces if it is accessible at the point of decision-making. The more easily a data scientist can access data, especially for research purposes, the better. Having the ability to research and test available data is an important part of creating meaningful data science applications.

Easily available data in various formats, even overlapping data sets, are incredibly valuable for data science research. Raw data, processed versions of data and data in Apache Parquet files are useful formats (Parquet is an open source file format that is most preferred by data scientists. Apache Parquet is ™ the Apache Software Foundation, https://parquet.apache.org/). Data scientists find it useful to research and experiment with a variety of types of data sets. Having the ability to easily access a variety of data helps data scientists to identify patterns, perform and identify clustering, run models or algorithms and in general get a better understanding of the data and what solutions and insights might be created and drawn.

Making data easily accessible to an experimental environment is also compelling. Having an underwriting system with a flexible testing environment that doesn't impact real data, but has easy access to historical data is incredibly valuable for testing and experimental purposes. For example, testing models, researching new signals in loss data, experimenting with changes in rating and underwriting criteria provides for valuable experimental learning. Being able to access data to test proposed rating changes and the impact on a product and portfolio level allows an insurer to ultimately save time in bringing new products to market.

> Data powers all AI; the more data that inference engines can ingest, the faster and better they can learn, and the better their answers.
>
> ***Celent: Machine Learning in Insurance Fact from Fiction.***
> Donald Light


## DETERMINE AND ENSURE DATA QUALITY

It goes without saying, any data management strategy needs to include steps to ensure data quality. The quality of all data sets, both internal data and external data, is of utmost importance as it provides the foundation for advanced analytics. Insurance companies have traditionally stored data in a variety of disparate systems, many of which are legacy systems. In addition much data is also stored in spreadsheets. Having large amounts of information stuck in silos increases quality issues. Some data can be incomplete and inconsistent, some fields may not be as populated or as complete as others. Insurers often have a lot of unstructured data (such as pdf files, images, website, email, IoT data etc.) that needs to be converted. In addition curated data from third-parties can inadvertently contain discrepancies between data set updates, and vendors may switch formats and not inform the user. Some vendors use sub-vendors leaving room for more errors and/or gaps. Columns or rows could be lost, or data can even be incorrect.

A majority of data modeling and data science is data cleansing. Quality data science results depend on quality data, therefore all data sets need to be checked for accuracy, consistency and gaps. The sooner the data is prepped and cleansed the sooner it can be validated and used - speed matters. Simple algorithms can be used to efficiently and effectively check and cleanse data sets. Machine learning algorithms can be used to help to look for quality issues and discrepancies between data set updates and third-party data sets can be used to validate each other. Quality information leads to more accurate data science results and in turn more confident decision-making and improved productivity. A reliable system of record requires a robust ETL (Extract, Transform, Load) and data engineering process which ensures the highest quality data possible.

## CONCLUSION

With the proliferation of data and the increased adoption of data science, managing data effectively is becoming more and more important. Having a successful data management strategy is critical for the successful application of advanced analytics such as machine learning. Aligning data scientists and business users in the planning stage, having large volumes of easily accessible, quality data that can be sourced and tracked are all key elements of a data management strategy that best enables data science applications. In addition, an advanced integrated underwriting system built on a modern architectural platform can help both successfully manage data and apply data science at the point of decision-making. Having the ability to capture an abundance of data and make it available through a modern data warehouse will provide insurers the ability to scenario-test, create new/better insurance products, and price risk more effectively.

## HOW TWO SIGMA IQ CAN HELP

Having the right system/platform, one that can effectively access, utilize and apply a carrier's data is important in helping to implement a strong data management strategy. Many of today's core systems, however, are unable to effectively and efficiently access and utilize a carrier's data. According to Accenture up to 75% of an insurer's data may be inaccessible to automated systems, a roadblock often referred to as "Dark Data" (Shining a Light on Dark Data, 2020).

Two Sigma IQ can help. In our early experience working with customers, we were met with the challenges of extracting data from existing legacy systems in order to be able to derive risk insights using data science techniques. This led us to develop our IQ Platform, an automated underwriting platform for the future. The IQ Platform integrates data science directly into the workflow in order to provide risk insights to the underwriter at the point of decision-making. The results are improved accuracy, more consistency and greater efficiency within the entire underwriting process. The IQ Platform enhances organizational agility, accelerates time to market for new products, enables data-driven decision-making for risk analysis, improves operational visibility and streamlines processes.

# ABOUT THE AUTHOR


## GRAEME DIXON

**CHIEF TECHNOLOGY STRATEGIST**


Prior to TSIQ, Graeme worked as the Chief Architect at Dell Software and as a Distinguished Engineer in various product and research divisions at IBM. His areas of focus have been on building scalable and robust solutions and developing middleware and cloud technology.

At TSIQ, a Two Sigma initiative, his focus is now on building cloud-native solutions focused for the insurance industry.

When he is not at TSIQ he spends his free time on his bike(s) training to keep up with the local racers and taking photographs to justify his GAS for new camera equipment.

If you would like to learn more about our IQ Platform please feel free to contact one of our client representatives:


**John Paladino**
john.paladino@twosigmaiq.com

**Suzanne Daly**
suzanne.daly@twosigmaiq.com


TWO SIGMA IQ

www.twosigmaiq.com