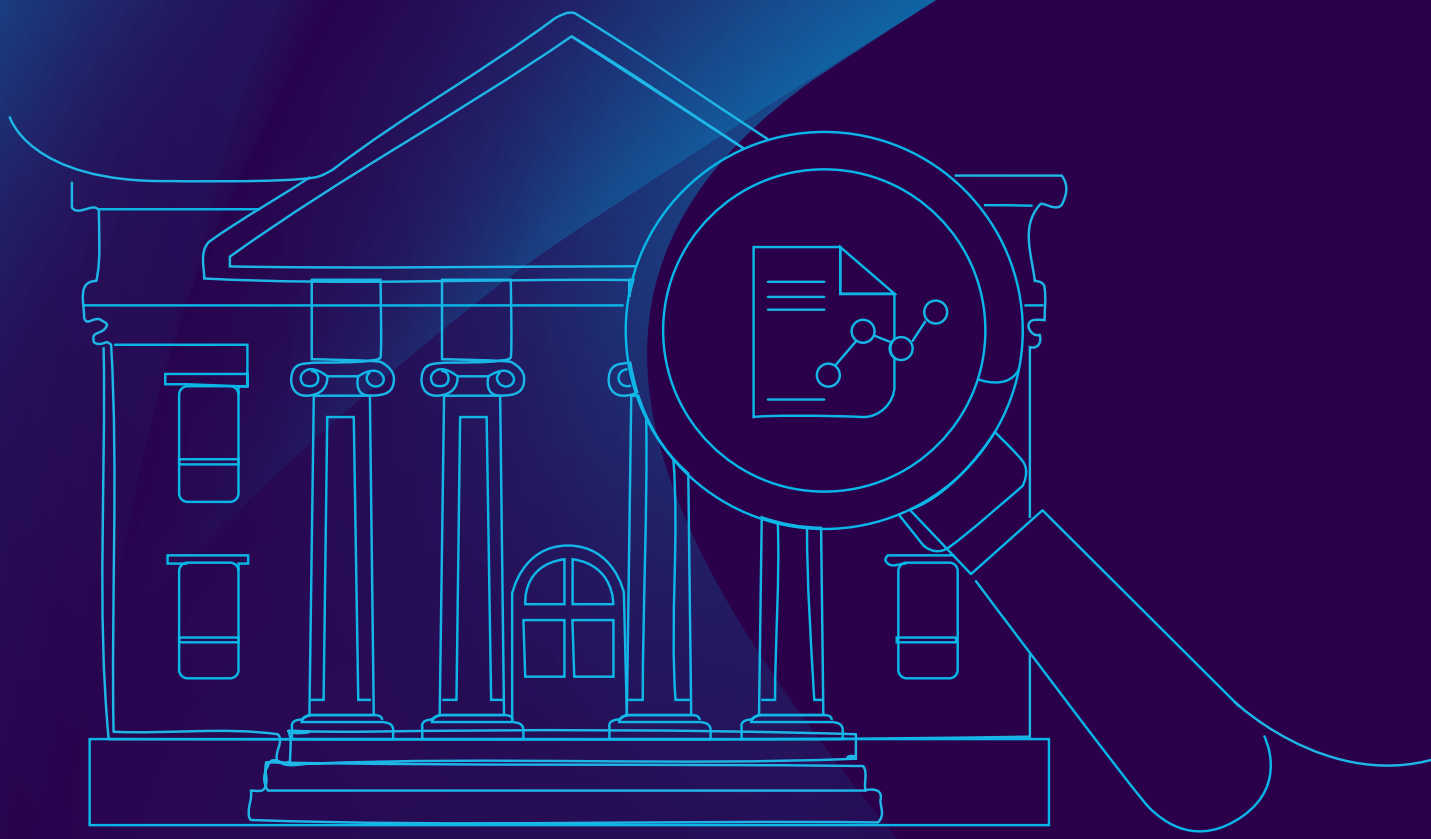
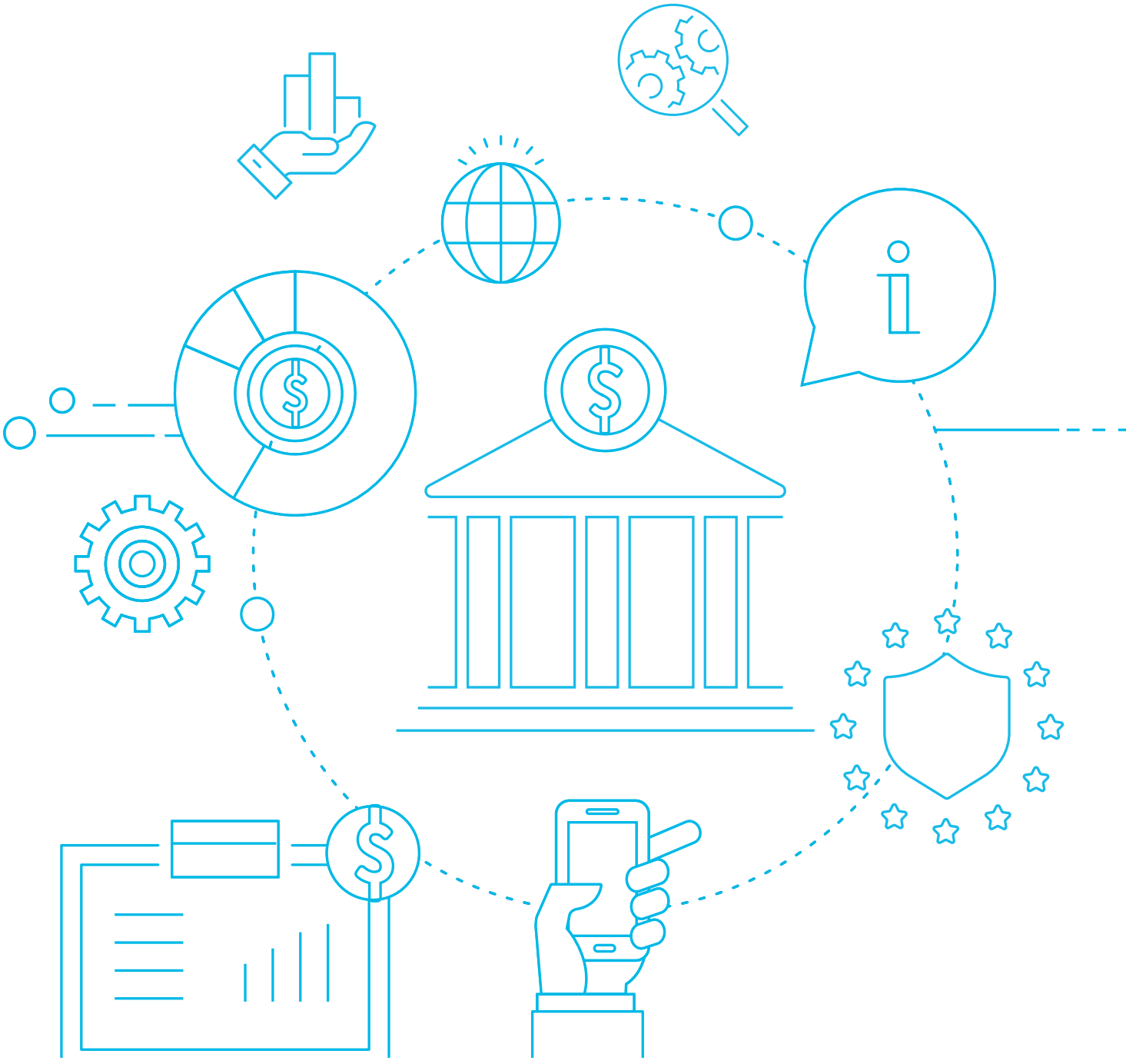
Capgemini invent

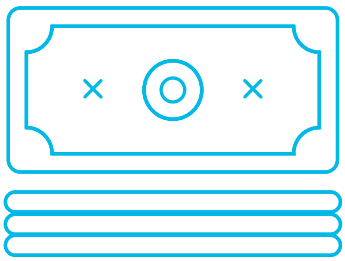
FROM TRADITIONAL TO ANALYTICAL DATA QUALITY MANAGEMENT

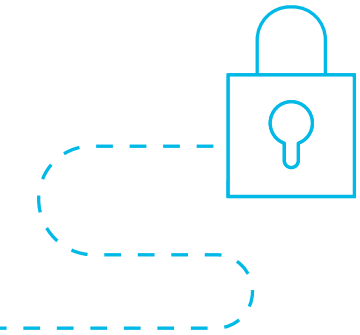
Our Point of View



THE NEW ERA OF DATA QUALITY MANAGEMENT IN BANKING IS HERE







History and context

Historically, banks were always subject to a large amount of data flowing through its systems. Since the introduction of digital devices such as smartphones, tablets and wearables, the pace of data creation has exploded. Estimates show us that more than 90% of total data has been created in only the past two years. Banks are part of this trend and acquire and process increasing amounts data. In the past, this data was gathered to a great extend in source administration systems. Over the years however, banks increasingly deal with new regulations like compliance with GDPR, TRIM, AnaCredit and BCBS239. Therefore, banks are working intensively on integral databases where uniform granular data is available. This integration poses huge challenges for the data quality management (DQM) practices of banks. In addition, banks are getting criticized and even fined for data quality issues in practices like Fraud detection and Anti money laundering.

Traditional Data Quality Management

Traditional human based Data Quality Management (DQM) often consists of only semi-automatic data monitoring and highly manual data cleaning processes. Processing the growing bulk of complex interdependent data is becoming increasingly costly and time-consuming as more and more DQ rules are required. As a result, crude rules are often applied which only touch the surface of data quality issues.

New approach

Our approach, Analytical Data Quality Management (ADQM), deals with these challenges by incorporating advanced analytical methods such as 'Anomaly detection' and 'Root cause analysis'. Methods like these allow us to identify anomalies in the data quickly and allow for the implementation of more refined DQ rules because the root cause is known. The efficiency and accuracy benefits of ADQM can be crucial for the success of banks, thereby supporting to minimize financial costs, increase productivity, safeguard regulatory compliance, increase organizational innovation, improve the decision-making process and prevent loss of reputation.

FINANCIAL COSTS	PRODUCTIVITY	REGULATORY COMPLIANCE	ORGANIZATIONAL INNOVATION	DECISION MAKING	REPUTATION
• Average financial impact of poor data quality on banks is significant • Poor-quality data costs approx. 30% of revenues	• Tasks due to weak data are prone to errors, require manual corrections and more cleaning time • One third of analysts spend 40% on validating and vetting data before decision making and Data scientists spend 50-80% of their time collecting and preparing data	• Failure of integration new data standards required by regulations • Unable to meet continuously evolving regulatory reporting require-ments • Fines from the regulator for being unable to be compliant	• Unable to rely on data for insight driven business innovation and AI- Predictive banking • Data quality issues hindering expansion of open banking	• Weak business decisions • Missed opportunities • Poorly phrased business strategies in banking • Inaccurate customer information	• Mis-leading assumptions create customer satisfaction issues • Media leakages of bad encounters harm the banks image • Employees question the validity of data customers for validating services or goods which deteriorate trust to products

Figure 1: The importance of data quality: data quality issues increase cost, decrease productivity, worsen reputation, negatively impact organizational innovation, weaken decision making and avoid regulatory compliance of organizations.

1

TRADITIONAL DATA QUALITY MANAGEMENT

Traditionally, Data Quality Management starts with defining business rules. It builds on metadata and the knowledge of subject matter experts. Traditional DQM is about creating data quality rules, the right controls and custom dashboards to improve the level of data quality. This also includes implementation of things like Six Sigma Method and DQ KPIs.

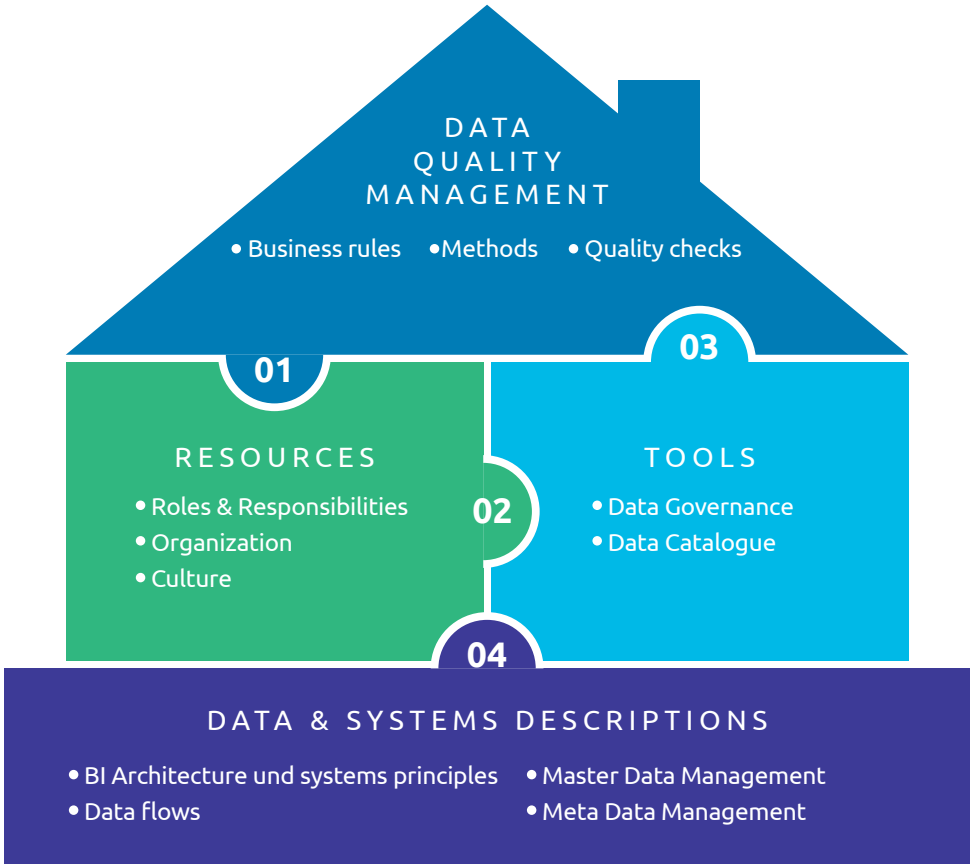


Figure 2: Core elements of traditional data quality management

Core elements

Traditional data quality management requires data quality policies where the business defines why data quality matters and what issues in the data should be avoided. Data rules are designed for the business to establish controls on the data. Data quality management is a continues process, therefore adaptation of the six sigma methods in data quality management can help in the cycle of improvements.

Resources and tools should be in place as well. In order to steer and control this process, the right data governance should be set up. Defining resources and data-related roles such as the Chief Data Officer (CDO), Data Steward, Business Owner, Data Owner and Data Custodian will lead to clear responsibilities and creates ownership of the data. This will aid to establish a data-driven culture. To assist the people in their work, the right tools should be implemented. There are numerous tools can facilitate easy custom dashboards and quality controls.

In addition, it is important to structure and design the data-architecture and system principles. An IT-architecture cartography of the key systems helps organizations to establish quality checks at the data entry interfaces and define principles which promote automation, system harmonization and simplification of the business processes. One example of a popular change in the data architecture is the creation of a data quality gate that works as a gateway in the data sourcing layer. Another example is the usage of data hubs. Data that pass the quality gate can be integrated in the data hubs. The implementation of master data management systems can provide common services to all applications and ensures access to master data. Consequently, a metadata repository serves as a single point of truth for the definition of data elements and reports which can be managed from all enterprise systems.

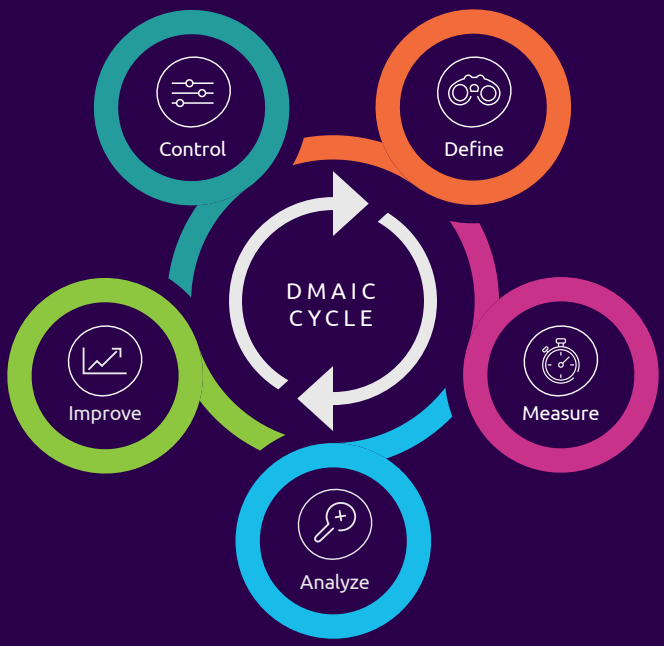


Figure 3: DMAIC cycle of Six Sigma

DMAIC cycle

The Six Sigma DMAIC Method reflects the traditional DQM process to reach continuous data quality improvements. In the beginning a definition of data quality rules based on meta data and an agreement of a business concept needs to be done. Secondly, the implementation and execution of measurements aggregate results to KPIs. A detailed analysis of measurement results aids to identify root causes and hence prepare data quality improvements by deriving appropriate measures. Subsequently, the execution of planned measures improves data quality. Based on KPI dashboards data quality can be controlled and verified if improvement activities showed expected results or repeat analysis and initiate corrective actions is required.

Challenges with traditional DQM

Even with correct data governance, dedicated resources and solid processes, there are huge challenges with controlling current and future data flows within banks. Due to unprecedented changes in the technological environment and the continuous rise of new digital solutions, the traditional DQM reaches its limits.

Firstly, the requirement to have access to granular data poses new challenges for data quality management. In the past, banks could report data from their administration on aggregated level. Nowadays a higher granularity of data is requested by the regulator. In combination with the of the enormous growth of data volume, this poses new challenges in data quality management. New digital devices and user behavior has accelerated the pace of data creation and decuples the data volume from 2013 to 2020. Integration of this data throughout the IT landscape requires many checks and data quality rules. Hence, the traditional human based DQM, which consists of only a semi-automatic data monitoring and highly manual data cleaning processes becomes more time-consuming and less accurate.

Secondly, traditional data governance and data quality management platforms have proven to be constrained to deal with increased complexity and interdependency of the data. Static rules are the most common rules which work fine in well-defined data sources. However, in modern data environment, the simple rules are challenged to keep up with the complexity of data.

Of course, it is possible to come up with complex rules by data stewards. These rules can be hard coded manually, but that takes time. For subject matter experts to define the rules and then for developers to code them can be time consuming. This makes the scalability a costly challenge. What this means in practice is that many banks keep simple rules and just engage in the cost-inducing activity of data cleansing once every while. This short-term problem solver delivers a quick fix on local data quality issues but skips the core of the problem. It may appear an easy and cheap solution locally, but on the long term, the total cost of keeping in control of their data becomes bigger and bigger. Banks can be excused, as understanding the underlying data problems is very difficult. Even with a solid data quality management in place, the traditional way of DQM gives a human limitation of finding the root causes.

2

OUR APPROACH: ANALYTICAL DATA QUALITY MANAGEMENT

Machine learning methods offer a new realm of possibilities when it comes to identifying, analyzing and solving data quality issues.

Analytical Data Quality Management (ADQM) has three additional core elements: Data Analytics, Technology and Big Data Infrastructure. With these elements build of the existing core elements, ADQM is addressing the shortcomings of traditional data quality management. Analytical methods of ADQM can identify anomalies in data without a hard coded DQ rule. Additionally, a root cause for known DQ issue can be localized without a complete knowledge of the data lineage and dependencies.

Detecting anomalies and finding root causes complements traditional Data Quality Managements to improve robustness and efficiency.

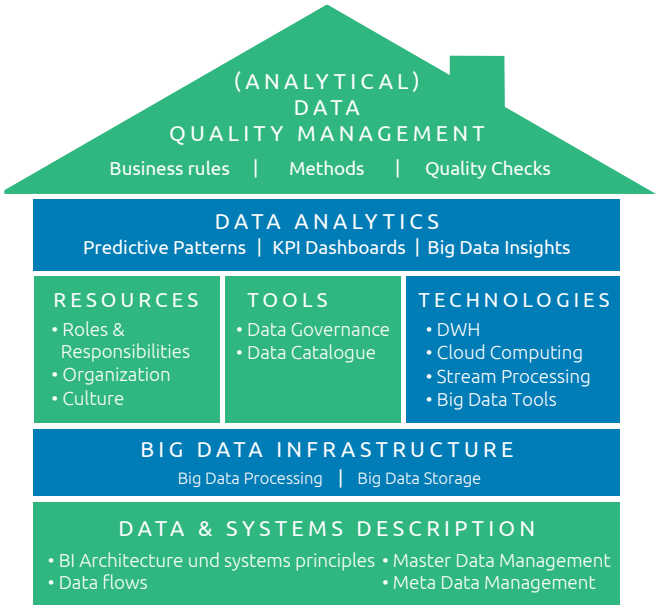


Figure 4: The green elements are those of the Traditional Data Quality Management. The blue elements are the additional elements of Analytical Data Quality Management.

Methods of ADQM

The two main methods 'Anomaly Detection' and 'Root Cause Analysis' pose the core of ADQM and shall be explained in more detail. For both cases, adapted supervised and unsupervised data analytics methods are used that are based on nothing more than the data itself.

The first method 'Anomaly Detection' finds unknown data quality problems through historical data patterns, which is possible without the knowledge of semantics. First, the dataset is encoded and compressed, thereby forcing the algorithm to find patterns. Secondly, the data is decoded again thereby creating a representation of the original data. This method is also called "autoencoder". ADQM uses the Autoencoder technology to investigate the reconstruction error between original data and data representation. Autoencoders learn standard patterns of historical data to subsequently distinguish between regular and irregular records i.e. hidden issues. This means the difference between the original data and the data representation are the anomalies. Once the anomalies are known, this insight can then be used to alter the original data set.

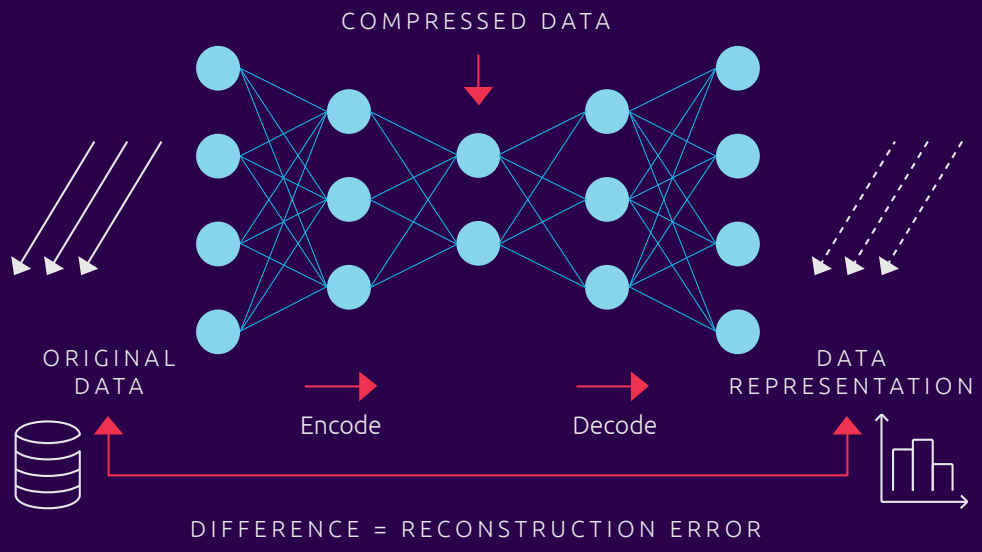


Figure 5: Visualisation of autoencoder technology, a neural network that can finds patterns in the data and creates a representation. The differences between the original and the representation are the anomalies.

The second method 'Root cause analysis' is a statistical model that enable a broader perspective on data quality issues by filtering for significant impact factors. The model finds data quality problems on multidimensional layers. Using supervised learning, Root-cause-analysis enable investigations of several layers below the symptoms of data quality issues. First, patterns are identified in the symptoms by using the best fitting separators in the data. Next, a partial dependency plot and model coefficients are created to approximate causation. ADQM does not merely look at correlating occurrences but analyses the impacting factors. In this sense, the deeper underlying problem is investigated. Lastly, the critical impact factors are specified.

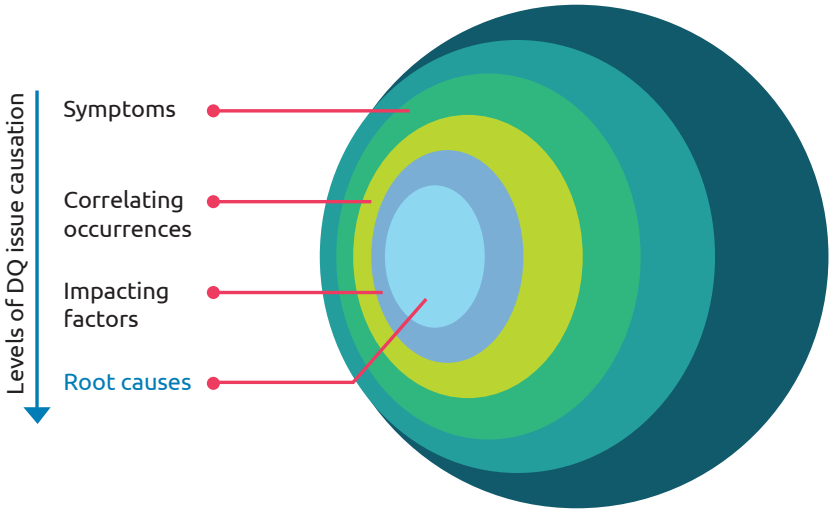


Figure 6: 'Root Cause Analysis' investigates the deepest level of a data quality issues

Before one or both methods can be applied there are some prerequisites. The data needs to be prepared. Firstly, the bank should have implemented an integrated data model. Structured and labeled data helps in preparation of the use of algorithms. Next to that, a data hub needs to be in place where uniform data is stored. Ideally, process data should be captured next to the content data. This includes data on location, user, system and timing of entry or changes. Lastly, an analytical base table needs to be created to enable the algorithm to work its magic.

Figure 7: Screenshots of the prototypes of the "Root Cause Analysis" and "Anomaly Detection" methods of Capgemini Invent

Analytical Data Quality Management significantly increases the robustness, efficiency and proactiveness of traditional Data quality management

Benefits of AI and Analytics for DQM

Banks can increase their robustness by using ADQM. Banks that adopt these statistical methods will create a resistance to outliers. 'Anomaly detection' facilitates continuous data quality improvements and proactive remediation by finding data quality issues before causing process failures and errors occur. Root-cause-analysis enables organization understanding of interdependencies between different data sources and avoid gaps during the data creation or integration process. In practice, this means that organizations can be in true control of their data, thereby enables regulatory compliance, avoiding data and process errors and costly regulatory fines.

Secondly, ADQM increases the efficiency of DQM processes as a result of 'Anomaly detection' i.e. automated data quality issue identification. When defining data quality rules based on meta data and learnings of previous DMAIC cycles, the AI supports the analysis of measurement results in order to identify root causes and prepare data quality improvements. Another significant efficiency benefit is gained in the root analysis process. Traditional DQM processes are very manual, particularly when data lineage and dependencies are not transparent and clear. With the ADQM approach, AI methods are used which speeds up the process. Even though ADQM can't identify the reason for a certain group of issue completely automatically, it can significantly reduce the number of objects and systems to check. The result is that issues can be identified much faster, which enables data stewards to solve them much faster. Furthermore, it avoids new DQ issues in the future. In this sense, ADQM enables increased productivity and decreases future cost.

ADQM also creates a better understanding of the data and its dependencies as a result of the AI driven root cause analysis. This can now even be used to predict DQ issues, that are not yet existing. Identifying the pattern can lead to DQ issue can be used to take preventive actions. This opens the door for new opportunities and enables organization innovation. With 'Anomaly detection' in conjunction with a big data environment, banks could react to DQ issues in real time. In some cases, it's even possible to correct these issues and to use the already cleansed data in the further processes.

Lastly, finding the root causes of data quality issues enables stakeholders to take sustainable actions and improve decision-making based on data. In this sense, ADQM is the next step for data-driven banks.

Consistency	The degree to which a unique piece of data holds the same value across multiple data sets • Calculate data set closeness using Probabilistic Record Linkage
Correctness	The degree of conformity of a data element or a data set to an authoritative source • Anomaly detection (unsupervised, e.g. auto encoder)
Completeness	The degree to which all required occurrences of data are populated • Missing Data imputation using causality and correlation between attributes
Validity	The measure of how a data value conforms to its domain value set (i.e. value, range of values) • Validate existing DQ rules • Identify root causes of data quality issues
Timeliness	The degree to which data is available when it is required • Identify process and system bottlenecks using process mining of log files (NLP)

Figure 8: ADQM assists to solve the five main quality issues of correctness, completeness, validity, timeliness and consistency:

3

EXAMPLES OF USE CASES IN THE BANKING INDUSTRY

Anomaly Detection Improves Risk Domain Data

Capgemini Invent helped a bank upgrade a dataset in the risk domain by improving data management practice with the Capgemini Data Management Flywheel as a framework, building a data quality monitoring dashboard and applying machine learning techniques to improve data quality.

The dataset in the risk domain (Risk management banking software includes systems such as Matlab, SAS, SCART, RiskPro, Murex) was facing challenges such as inadequate data management and poor data quality. Data objects included customers, accounts, products and services as well as transactional types and a balance history, in which each data object had attributes (type, length, scaling, integer, count, number, defined and operation code). Also, there was lack of evidence for data quality issues to communicate with data sources. The bank wanted to gain a better understanding of the relationships between its disparate pieces of data. Traditional databases found it difficult to analyze them meaningfully.

Our ADQM approach enabled risk functions to make use of structured and unstructured customer information. This increased the predictability power of the models and led to better credit risk decisions by monitoring portfolios for early evidence of existing or potential problems and detecting financial crime. For example, a fraudulent transaction may be for a product the account owner has never bought or would likely never buy or the geographical location of the person who made the purchase may not coincide with where the account owner was at the time of purchase. The machine learning algorithm can detect these inconsistencies after being trained, and so it will be more sensitive to those data points within transactions and flag them if the location data and the purchased product is suspicious.

The dashboard is developed to give frequent, periodic insights in the quality of the data, and it contains three sections: a summary that shows an overall overview of the data quality of the variables and dashboards for individual data variables that give more insight in the data quality of a specific variable. Workshops with experts were organized to gather more advanced, cross-relational data quality business rules, which were incorporated in the data quality monitoring dashboard. The dashboard brings transparency to data quality issues and can be used as evidence to communicate with stakeholders and trigger the data quality remediation process.

Machine learning algorithms were used to help identifying anomalies which helps to increase the quality of data. Anomaly detection algorithms can find strange data entries, that the usual, rule-based data quality rules would not find. Application of the algorithms helps to narrow down the potential pool of anomalies, which should lead to further investigation. Technology brought false positives down in the anti-money laundering function. This allowed for focused approaches to risk detection and avoidance.

The following was achieved:

- Creation and enrichment of data assets in an efficient way
- Management of a data lifecycle, especially when it comes to sensitive and retiring data
- Improvement of use of data and discovery by users

- Reduced risk costs and fines

MACHINE LEARNING TECHNIQUES HELPED IMPROVE THE ACCURACY OF RISK MODELS BY IDENTIFYING NON-LINEAR AND COMPLEX PATTERNS IN LARGE DATA SETS

THROUGH THIS, 'ANOMALY DETECTION' FINDS LESS FALSE POSITIVES, LEADING TO A REDUCED NEED TO MANUALLY VERIFY FLAGGED TRANSACTIONS

ADQM improves 'Know Your Customer' and 'Anti Money Laundering'

Know Your Customer (KYC) and Anti Money Laundering (AML) have been in the news regularly. Data is the foundation of both the KYC and AML process. Manually screening all clients and verifying all transactions is a huge challenge, due to the large amounts of data that come with this. Capgemini Invent proposes ADQM as a solution.

Financial institutions perform background checks before accepting customers, leading to a large pool of customer data. Names are checked against blacklists and sanction lists. Using Natural Language Processing (NPL) in combination with several other Artificial Intelligence (AI) techniques, the quality of the data that the bank collects can increase. For example, AI and NLP might be able to link
a news article about 'John Smith' with negative implicating information to the actual 'John Smith' that is trying to open a bank account. In this case that might be that the John Smith that was name in the news article is not allowed to become a customer. However, a second, different John Smith does not get unnecessarily flagged. Being flagged here would imply that the bank needs to further investigate a potential customer. In this way, financial institutions are able to increase the quality of their data and can make more informed decisions and take them quicker.

'Anomaly detection' can be leveraged within transaction monitoring with the goal to prevent money laundering. Banks contain a big amount of transaction data. With this data, the goal is to discover criminal activities, preferably with the smallest number of false positives as possible. A false positive here can be for example a transaction of a large sum of money to a certain country that is flagged as suspicious, however after further manual investigation it turns out that there is no proof for criminal activity.

The autoencoder technology investigates the data and learns from patterns in historical data to flag outliers in the transactions.

Anomaly detection algorithms will not only reduce false positives but can be used to detect suspicious transactions as well. The advantage of using algorithms versus traditional methods is that manually determined rules are static and are not all-covering while algorithms detect unknown real patterns. Finding suspicious transactions can be done using two types of algorithms. The first type is a supervised algorithm which needs a training data set in which the true anomalies are known. This algorithm also requires user input to learn. The second type is an unsupervised algorithm. This algorithm has a self-learning ability and can update itself based on new data.

Capgemini has implemented both Root Cause analysis and Anomaly detection on client site. There is a showcase available to demonstrate its principles and look & feel. Please reach out to us if you are interested in a demo.

4

ABOUT THE AUTHORS

Casper Stam
Head of Data, Finance, Risk and Regulatory

casper.stam@capgemini.com

Netherlands

Frank Clement
Lead Innovative Data Management

Frank.clement@capgemini.com

Netherlands

Thomas Bakker
Senior Consultant Data, Finance, Risk and Regulatory

Thomas.bakker@capgemini.com

Netherlands

Elena Paniagua Avila
Senior Consultant Data, Finance, Risk and Regulatory

elena-paniagua.avila@capgemini.com

Netherlands

Koen Wolvers
Consultant Data, Finance, Risk and Regulatory

Koen.wolvers@capgemini.com

Netherlands

Chao Bao
Senior Consultant Data, Finance, Risk and Regulatory

Chao.bao@capgemini.com

Netherlands

About
Capgemini Invent

ABOUT CAPGEMINI INVENT

As the digital innovation, consulting and transformation brand of the Capgemini Group, Capgemini Invent helps CxOs envision and build what's next for their organizations. Located in more than 30 offices and 22 creative studios around the world, its 6,000+ strong team combines strategy, technology, data science and creative design with deep industry expertise and insights, to develop new digital solutions and business models of the future.

Capgemini Invent is an integral part of Capgemini, a global leader in consulting, technology services and digital transformation. The Group is at the forefront of innovation to address the entire breadth of clients' opportunities in the evolving world of cloud, digital and platforms. Building on its strong 50-year heritage and deep industry-specific expertise, Capgemini enables organizations to realize their business ambitions through an array of services from strategy to operations. Capgemini is driven by the conviction that the business value of technology comes from and through people. It is a multicultural company of over 200,000 team members in more than 40 countries. The Group reported 2018 global revenues of EUR 13.2 billion.

Visit us at **www.capgemini.com/invent**

People matter, results count.

The information contained in this document is proprietary.
©2019 Capgemini. All rights reserved.