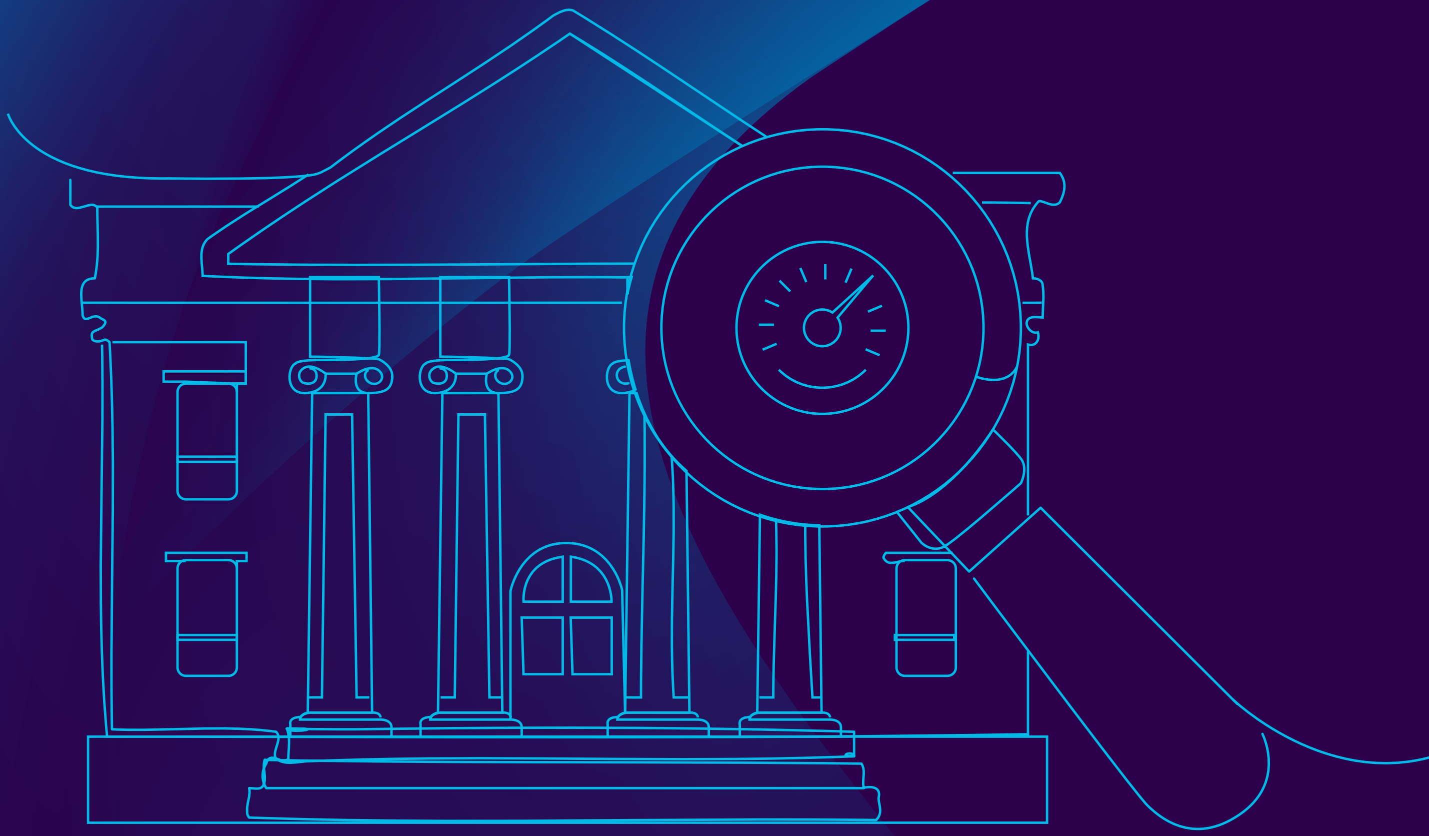
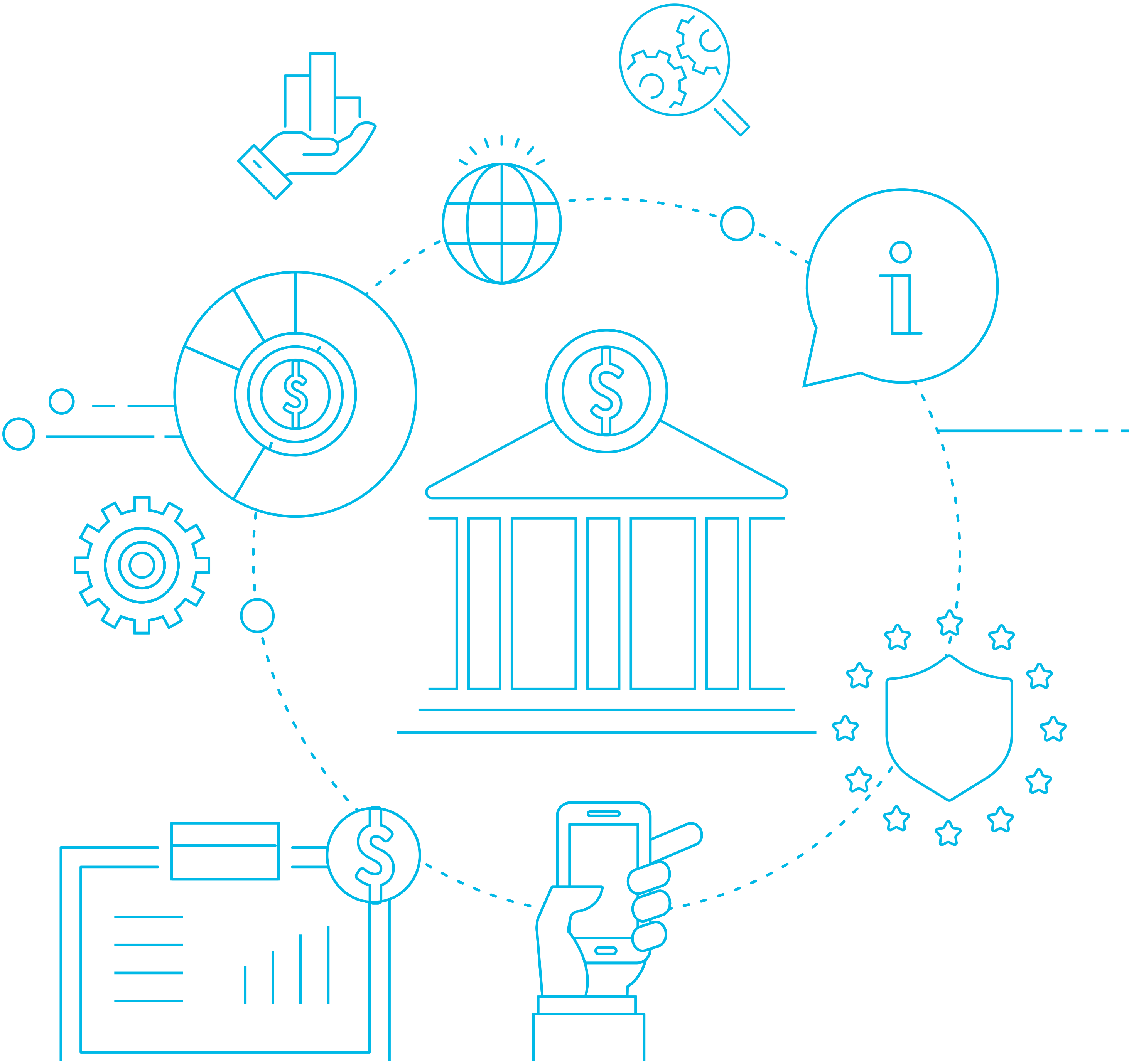
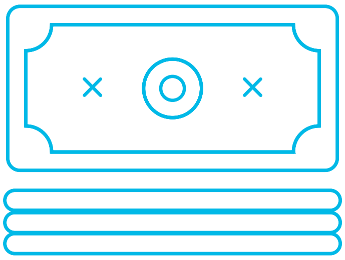
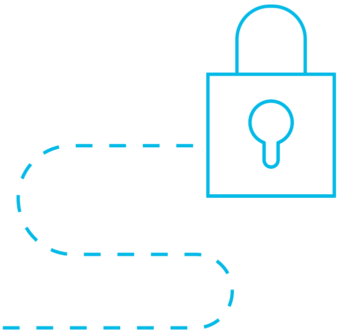
FROM TRADITIONAL TO ANALYTICAL DATA QUALITY MANAGEMENT



THE NEW ERA OF DATA QUALITY MANAGEMENT IS HERE







History and context

Historically banks have always had a large amount of data flowing through their systems. In recent times, however, the pace of data creation has exploded due to the introduction of digital devices, such as smart phones, tablets, and wearables. The pace of data creation has exploded. In fact, recent estimates find that more than 90% of total data has been created in the last two years alone. This trend has clearly involved the banking industry, with banks acquiring and processing increasing amounts of data. In the past, bank data has been largely gathered in source administration systems. This is changing, as banks are now increasingly challenged to comply with data compliance regulations like GDPR, TRIM, AnaCredit and BCBS239. Because of this banks are intensively working to create integral databases in which granular data is uniformly available. This integration poses huge challenges for banks, particularly with regards to data quality (DQ) and data quality management (DQM). Failure to meet these requirements has caused banks to be criticized in the press and even fined for having substandard or faulty data quality issues, such as in practices like fraud detection and anti-money laundering (AML).

Traditional data quality management

Traditional semi-automated DQM often consists of 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'. These new methods allow us to identify anomalies in the data and determine the root cause, thus allowing us to enable the implementation of more refined DQ rules. ADQM brings critical benefits for firms in form of data efficiency and accuracy, thereby supporting their overall success by minimizing financial costs, increasing productivity, ensuring regulatory compliance, accelerating organizational innovation, improving decision-making processes, and preventing damage to a firm's reputation.

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.

FINANCIAL COSTS	PRODUCTIVITY	REGULATORY COMPLIANCE	ORGANIZATIONAL INNOVATION	DECISION MAKING	REPUTATION
• Average financial impact of poor data quality for banks is significant • Poor-quality data costs approx. 30% of revenues	• Tasks required to correct weak data are prone to error, 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 to integrate new data standards required by regulations • Unable to meet continuously evolving regulatory reporting requirements • Regulatory fines due to non-compliance	• 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	• Misleading assumptions create customer satisfaction issues • Media leakages of bad encounters harm the firm's image • Employees question the validity of data

1

TRADITIONAL DATA QUALITY MANAGEMENT

Data Quality Management starts with defining business rules. It builds on meta data and the knowledge of subject matter experts. In its traditional form, DQM is about creating data quality rules, the right controls and custom dashboards to improve the level of data quality. This also includes the implementation of frameworks like Six Sigma method and the creation of DQ KPIs.

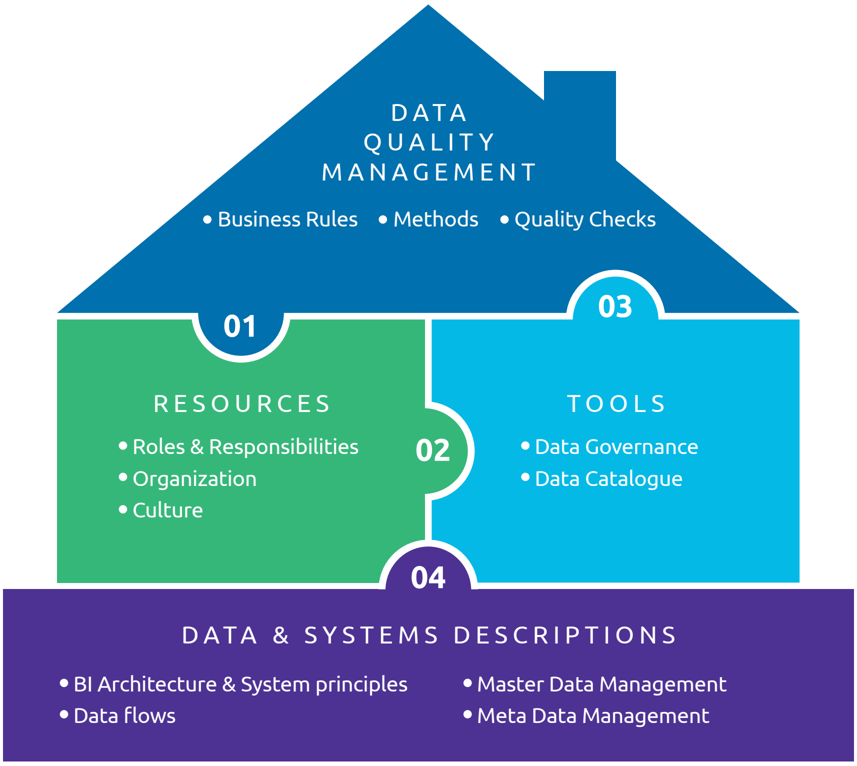


Figure 2: Core elements of traditional Data Quality Management

Core elements

Traditional Data Quality Management (DQM) requires data quality policies defined by business that detail why data quality matters and which data issues should be avoided. Data rules enable business to design and establish controls on the data. The adoption of Sigma Six methods and their integration into the DQM process enable a continuous improvement process.

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. Creating these roles has the added benefit of fostering a data-driven culture. Additionally, the right tools should be implemented to assist employees in their work. There are numerous tools that 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 to the data sourcing layer. Another example is the use of data hubs as fixtures for storing and integrating data that passes through the quality gate. The implementation of master data management systems can provide common services to all applications and ensure access to master data. Consequently, a meta data repository serves as a single point of truth for the definition of data elements and reports, which can be managed from all enterprise systems.

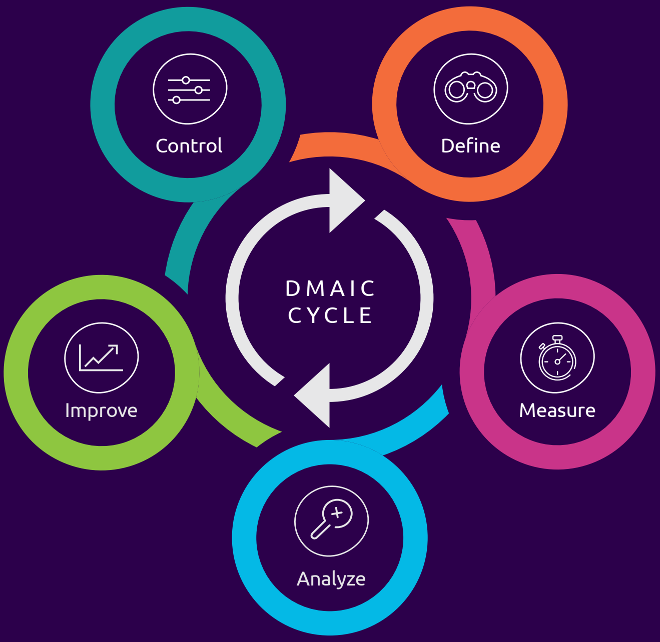


Figure 3: DMAIC cycle of Six Sigma

DMAIC cycle

The Six Sigma DMAIC Method describes the traditional DQM process to reach continuous data quality improvements. The first step is to define data quality rules based on meta data and come to an agreement on the business concept. Secondly, the implementation and execution of measurements must aggregate results to KPIs. A detailed analysis of measurement results aids to identify root causes and hence prepares data quality improvements by deriving appropriate measures. Subsequently, the execution of planned measures improves data quality. Data quality can be controlled and verified based on KPI dashboards if improvement activities show anticipated results.

Challenges with traditional DQM

Even with correct data governance, dedicated resources and solid processes, there remain huge challenges to control current and future data flows within firms. In fact, the effectiveness of traditional DQM nears its limits given today's unprecedented technological changes and the rise of new digital solutions. These challenges are wide-ranging.

Firstly, the requirement to have access to granular data poses new challenges for data quality management. In the past, firms could report data from their administration on aggregated level. Nowadays, however, a higher granularity of data is requested by the regulator. This comes in combination with an enormous growth in the volume of data. New digital devices and user behavior have accelerated the pace of data creation and decoupled the data volume from 2013 to 2020. Integration of this data throughout the IT landscape requires many checks and data quality rules. Hence, traditional DQM, which consists of only 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 by how they can deal with increasing complexity and interdependency of the data. Static rules are the most common rules, which work fine for well-defined data sources. However, in modern data environments, these simple rules are challenged to keep up with the complexity of data.

Of course, it is possible for data stewards to come up with complex rules. These rules can be manually hard coded, but such efforts can require extensive resources. This can be time consuming for the subject matter experts who need to provide data definitions, and also for the developers who must code the rules. This makes achieving scalability a costly proposition. This means in practice that many firms opt to keep simple rules and just engage in the cost-inducing activity of data cleansing. This short-term problem solving delivers a quick fix for local data quality issues, but skips the core of the problem. It may appear to be an easy and cheap solution, but in the long term the total cost of keeping control of data increases immensely. Firms can be excused for this, however, as remedying the underlying data problems is very difficult. Even with solid DQM in place, the traditional approach to DQM is limited because users must identify the root causes of the problems.

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.

Digital Distribution & Sales
Analysis
▲+89%
AR Consumption growth
Average user base
Social Networks growth in marketing
Product relative effectiveness
Social networks influence
Community involvement
Content Consumption

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

Detecting anomalies and finding root causes complement traditional Data Quality Management to improve its robustness and efficiency.

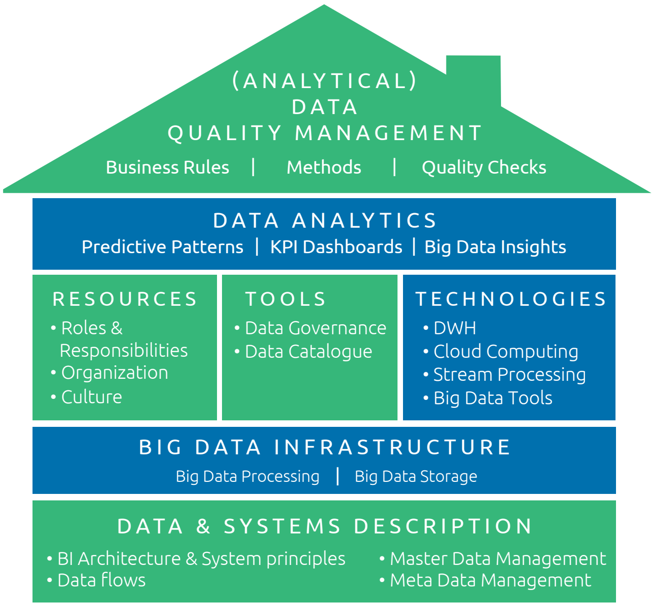

(ANALYTICAL) DATA QUALITY MANAGEMENT
Business Rules | Methods | Quality Checks
DATA ANALYTICS
Predictive Patterns | KPI Dashboards | Big Data Insights
RESOURCES
• Roles & Responsibilities
• Organization
• Culture
TOOLS
• Data Governance
• Data Catalogue
TECHNOLOGIES
• DWH
• Cloud Computing
• Stream Processing
• Big Data Tools
BIG DATA INFRASTRUCTURE
Big Data Processing | Big Data Storage
DATA & SYSTEMS DESCRIPTION
• BI Architecture & System principles
• Data flows
• Master Data Management
• Meta Data Management

Figure 4:

🟦 The green elements are those of traditional Data Quality Management.

🟩 The blue elements are the additional elements of Analytical Data Quality Management.

Methods of ADQM

'Anomaly Detection' and 'Root Cause Analysis' are the two main methods which represent the core of ADQM and shall be explained in more detail. In both cases, either supervised or unsupervised machine learning methods are used to enable algorithmic pattern detection.

'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 such as hidden issues. Once the anomalies are identified, this insight can then be used to alter the original data set.

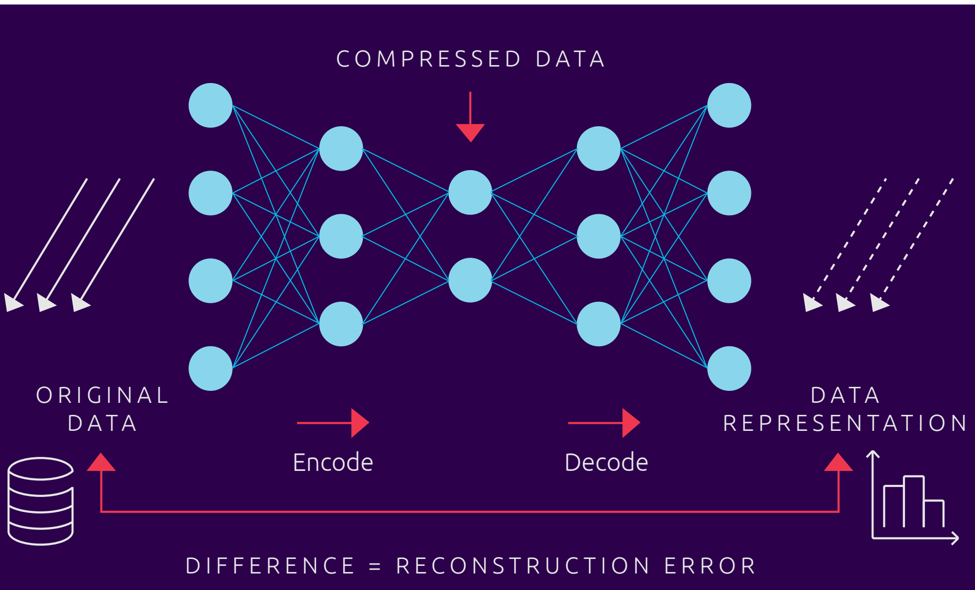


Figure 5: Visualization of autoencoder technology, a neural network that can find patterns in the data and creates a representation. The differences between the original and the representation are the anomalies.

‘Root Cause Analysis’ is a statistical model that enables 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 enables investigations 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.

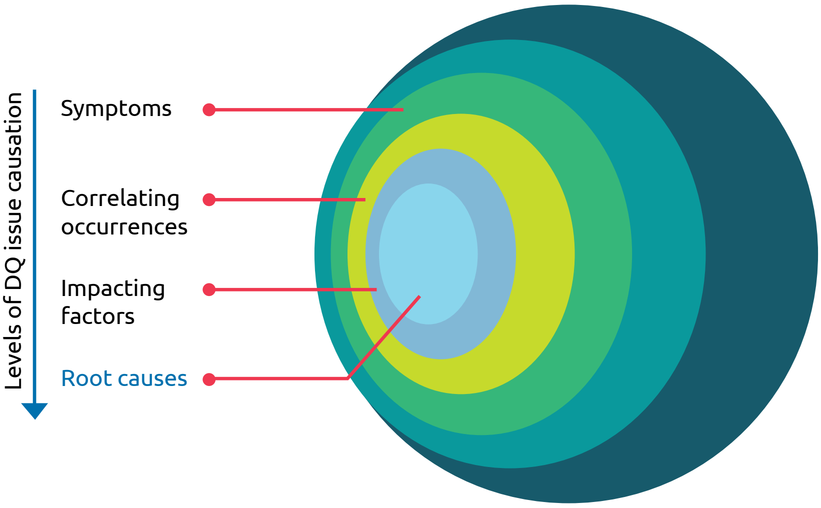
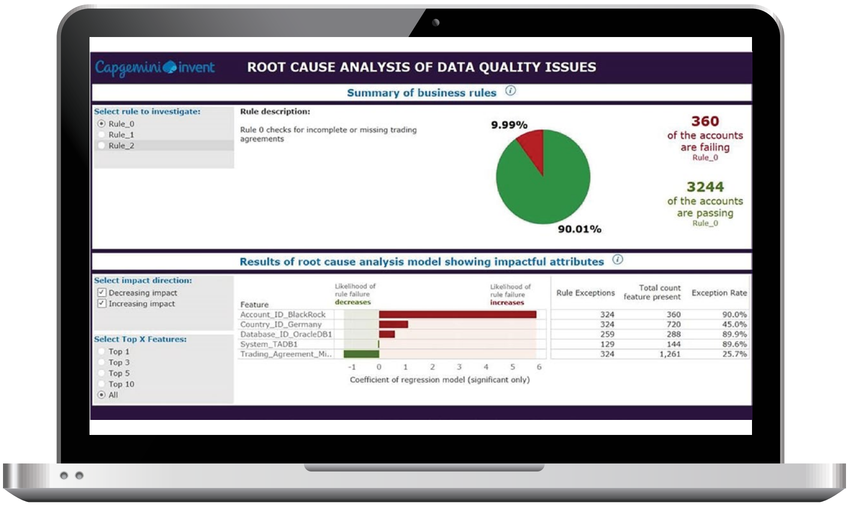


Figure 6: ‘Root Cause Analysis’ investigates the deepest level of a data quality issues

Before one or both methods can be applied there are some prerequisites. The data needs to be prepared. Firstly, the firm should have implemented an integrated data model. Structured and labeled data helps in the preperation 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 meta data of location, user, system and timing of entry or changes. Lastly, an analytical base table needs to be created to enable the successful implementation of the algorithm.


Capgemini invent
9.99%
90.01%
Likelihood of rule failure decreases
Likelihood of rule failure increases
-1
0
1
2
3
4
5
6
Coefficient of regression model (significant only)

ROOT CAUSE ANALYSIS OF DATA QUALITY ISSUES

Summary of business rules

Select rule to investigate:
● Rule_0
○ Rule_1
○ Rule_2

Rule description:
Rule 0 checks for incomplete or missing trading agreements

360 of the accounts are failing Rule_0

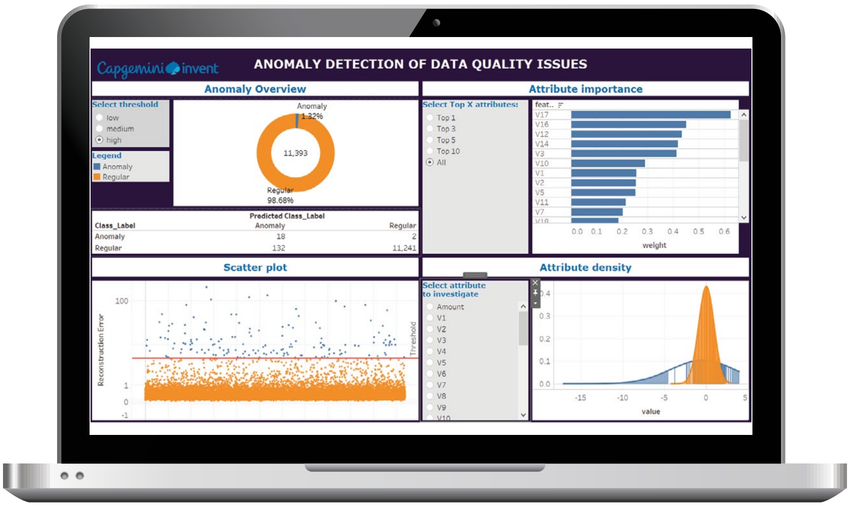
3244 of the accounts are passing Rule_0

Results of root cause analysis model showing impactful attributes

Select impact direction:
☑ Decreasing impact
☑ Increasing impact

Select Top X Features:
○ Top 1
○ Top 3
○ Top 5
○ Top 10
● All

Feature	Rule Exceptions	Total count feature present	Exception Rate
Account_ID_BlackRock	324	360	90.0%
Country_ID_Germany	324	720	45.0%
Database_ID_OracleDB1	259	288	89.9%
System_TADB1	129	144	89.6%
Trading_Agreement_Mi..	324	1,261	25.7%



ANOMALY DETECTION OF DATA QUALITY ISSUES

Anomaly Overview

Select threshold
○ low
○ medium
● high

Legend
🟦 Anomaly
🟧 Regular

	Predicted Class_Label	
Class_Label	Anomaly	Regular
Anomaly	18	2
Regular	132	11,241

Attribute importance

Select Top X attributes:
○ Top 1
○ Top 3
○ Top 5
○ Top 10
● All

Scatter plot

Attribute density

Select attribute to investigate
○ Amount
○ V1
○ V2
○ V3
○ V4
○ V5
○ V6
○ V7
○ V8
○ V9
○ V10

Figure 7: Screenshots of the prototypes for 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

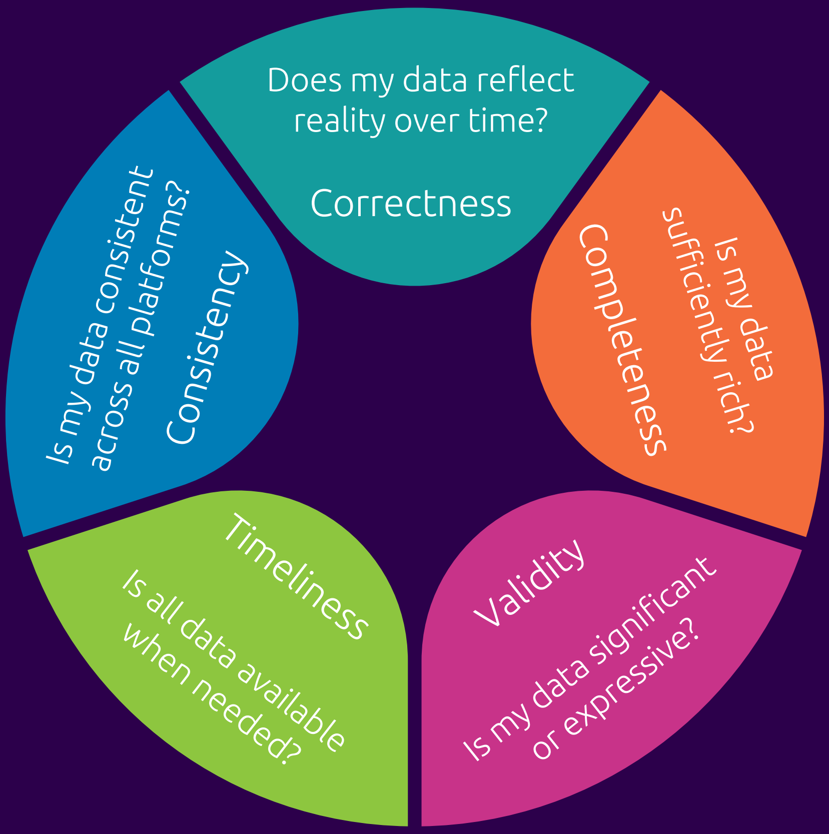
Benefits of AI and Analytics for DQM

Firms can increase their robustness by using ADQM. Those that adopt these statistical methods will create 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 organizational understanding of interdependencies between different data sources and avoids gaps during the data creation or integration process. In practice, this means that organizations can be in true control of their data, thereby enabling 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 Cause 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 speed up the process. Even though ADQM cannot automatically identify the reason for a certain group of issues, 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 resolve 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 do not yet exist. Identifying the pattern that can lead to DQ issues can allow companies to take preventive actions. This opens the door for new opportunities and enables organizational innovation. With 'Anomaly Detection' in conjunction with a big data environment, firms could react to DQ issues in real time. In some cases, it is even possible to correct these issues and to use the already cleansed data in 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 firms.


Does my data reflect reality over time?
Correctness
Is my data sufficiently rich?
Completeness
Validity
Is my data significant or expressive?
Timeliness
Is all data available when needed?
Is my data consistent across all platforms?
Consistency

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 for 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

DASHBOARD
My profile
Users 242 -4.5% last week
New users 55 +1.8% last week
Page views 908 -2.9% last week
Avg. time 5:36 +9.5% last week
Bounce rate 34% -8.2% last week
Conversion 12% +3.7% last week
Visitors countries
Traffic origin
Target completion
72%
45%
Sales
$23K
Visitors devices
Downtime
+4.1%
Last updated: 3 minutes ago

Anomaly Detection Improves Risk Domain Data

Capgemini Invent helped a bank to upgrade a dataset in the risk domain by improving data management practice with the Capgemini Data Management Flywheel framework. It built a data quality monitoring dashboard and applied 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 meaningfully analyze them.

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 exist 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 into 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 into the data quality of a specific variable. Workshops with experts were organized to gather more advanced, cross-relational data quality business rules, which were then 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 identify anomalies which help to improve the data quality. 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 the data lifecycle, especially when it comes to sensitive and retiring data
- Improvement of data usage and discovery by users
- Reduced risk costs and fines

MACHINE LEARNING TECHNIQUES HELPED TO 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 handled in these processes. 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 it might be that the John Smith that was named 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 transactional 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.

The advantage of using algorithms versus traditional methods is that manually determined rules are static and are not all-covering while algorithms detect unknown 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

FOR MORE INFORMATION, PLEASE CONTACT

Global

Aurelien Grand
aurelien.grand@capgemini.com

APAC

Samuel Levy-Basse
samuel.levy-basse@capgemini.com

Belgium

Koenraad D'Hondt
koenraad.d-hondt@capgemini.com

DACH

Ulrich Windheuser
ulrich.windheuser@capgemini.com

France

Matthieu Cirelli
matthieu.cirelli@capgemini.com

India

Gaurav Bedekar
gaurav.bedekar@capgemini.com

Netherlands

Casper Stam
casper.stam@capgemini.com

North America

Allan Frank
allan.frank@capgemini.com

Norway

Bjorn Tore Sand
bjorn-tore.sand@capgemini.com

Spain

David Julian Brogeras
david.brogeras@capgemini.com

Sweden & Finland

Johan Bergström
johan.bergstrom@capgemini.com

UK

Patrick Vance
patrick.vance@capgemini.com

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 25 creative studios around the world, its 7,000+ 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, digital transformation, technology and engineering services. 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. Today, it is a multicultural company of 270,000 team members in almost 50 countries. With Altran, the Group reported 2019 combined revenues of €17billion. People matter, results count.

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

People matter, results count.

The information contained in this document is proprietary.