

DATA LITERACY AND CONSIDERATIONS

Learning about the essential data concepts, disciplines and activities that are crucial to DM, DG and other data-related disciplines

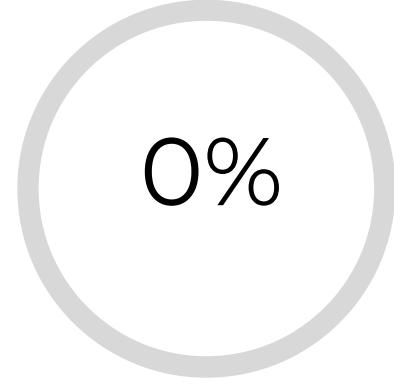


DATA LITERACY AND CONSIDERATIONS



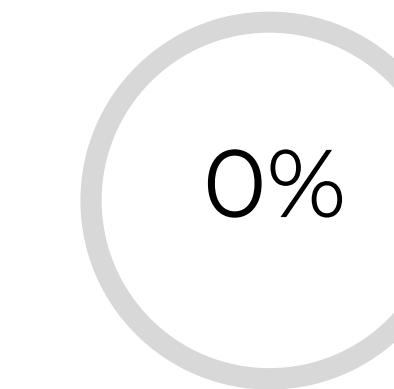
DATA LITERACY & CONSIDERATIONS

Covering the basics of data. What are the different data disciplines, what are the essential principles to handle data, what are the sophistication levels of organisations...



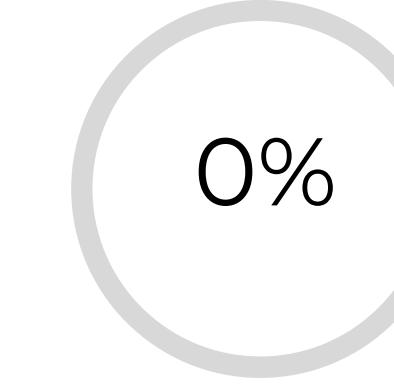
DATA GOVERNANCE

Covering how data governance works, from classifying data to setting policies and other activities, the roles and responsibilities, and the DG implementation process.



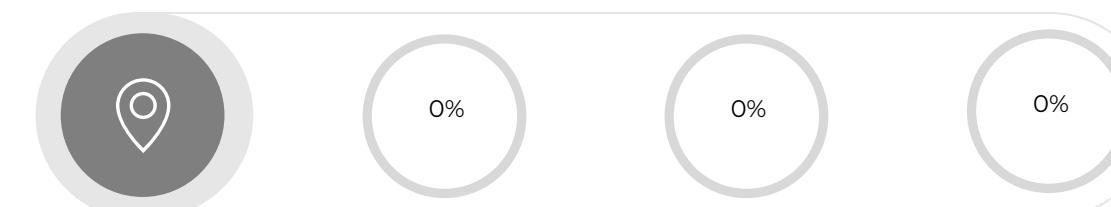
DATA AND DATA QUALITY

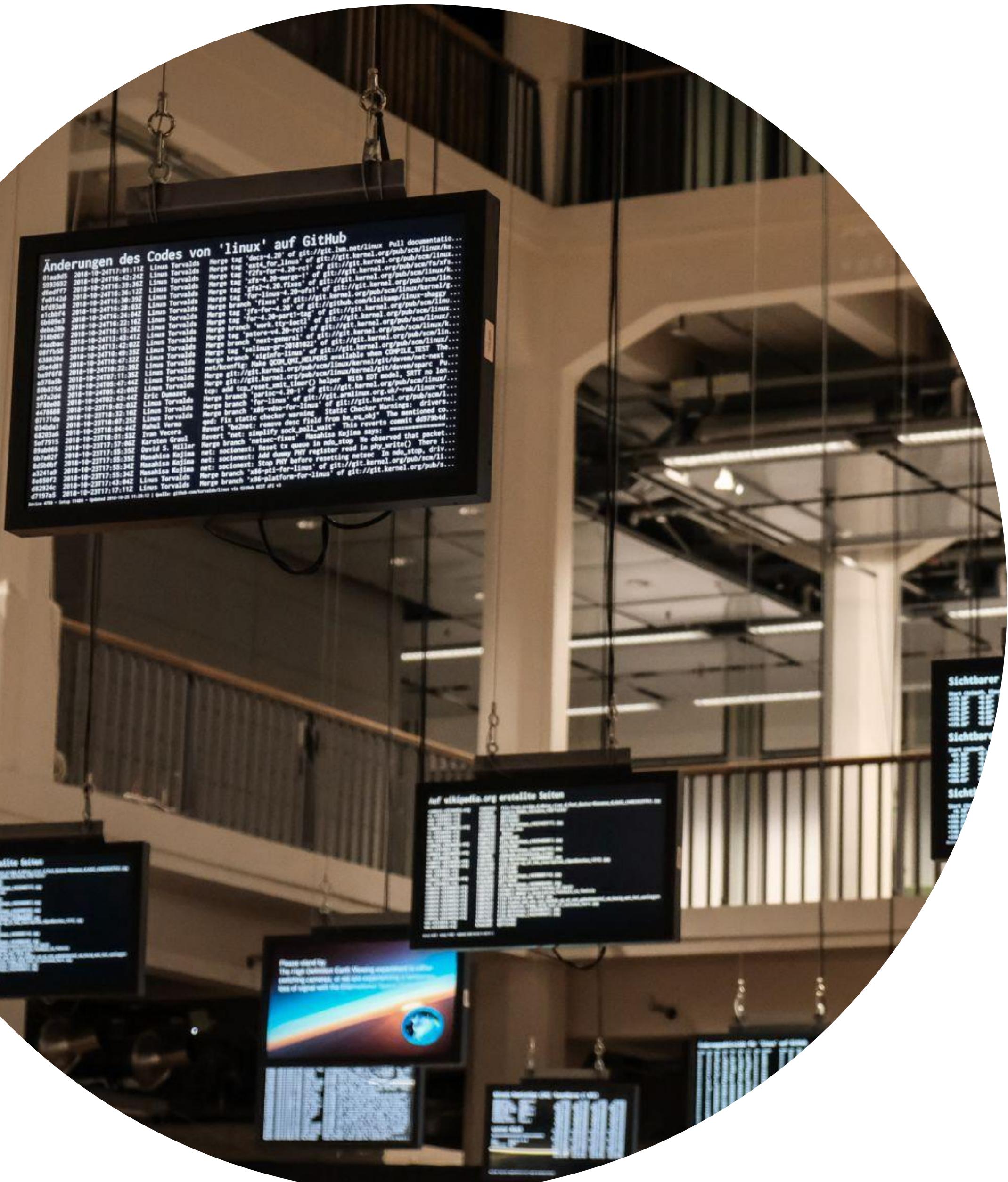
Covering specifically what data are and how to improve their quality. Data types, values, structures, and how to improve data quality through profiling and remediating.



DATA SECURITY, PRIVACY, ETHICS

Covering the different types of privacy and security controls that can be applied to data to protect them, as well as how to treat data subjects ethically.





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS GOALS

Our major goal in this module is to clarify key concepts related to data. We'll cover:

- The 4 key principles of data (data as assets, DG as a business function, and others);
- The different data disciplines and their nuances (Data Management, Data Governance, Data Stewardship, Data Science, and others);
- The usual information lifecycle, and the usual concerns and controls at each stage;
- The progression of DM/DG in an organisation, in terms of projects to programs to inherent processes;
- The different maturity and sophistication levels of an organisation in terms of data, and their manifestation;

DATA LITERACY AND CONSIDERATIONS

In this module, we will cover **six key topics** in terms of how data work and the usual practices that revolve around them:



4 KEY PRINCIPLES

Basic principles that any data activity relies on, such as data assets and monetisation



DATA DISCIPLINES

What is Data Management, Data Governance, and others, and how they are distinguished

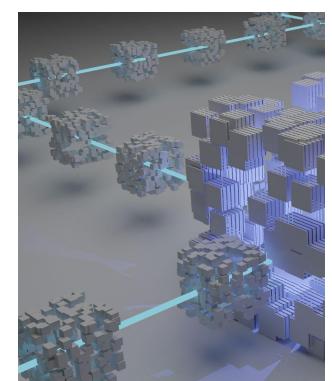


DG/DM KEY ACTIVITIES

Examples of usual activities in Data Management and Data Governance, and their differences

DATA LITERACY AND CONSIDERATIONS

In this module, we will cover **six key topics** in terms of how data work and the usual practices that revolve around them:



THE INFORMATION LIFECYCLE

What are the usual stages of data in an organisation, from creation to disposal, in detail



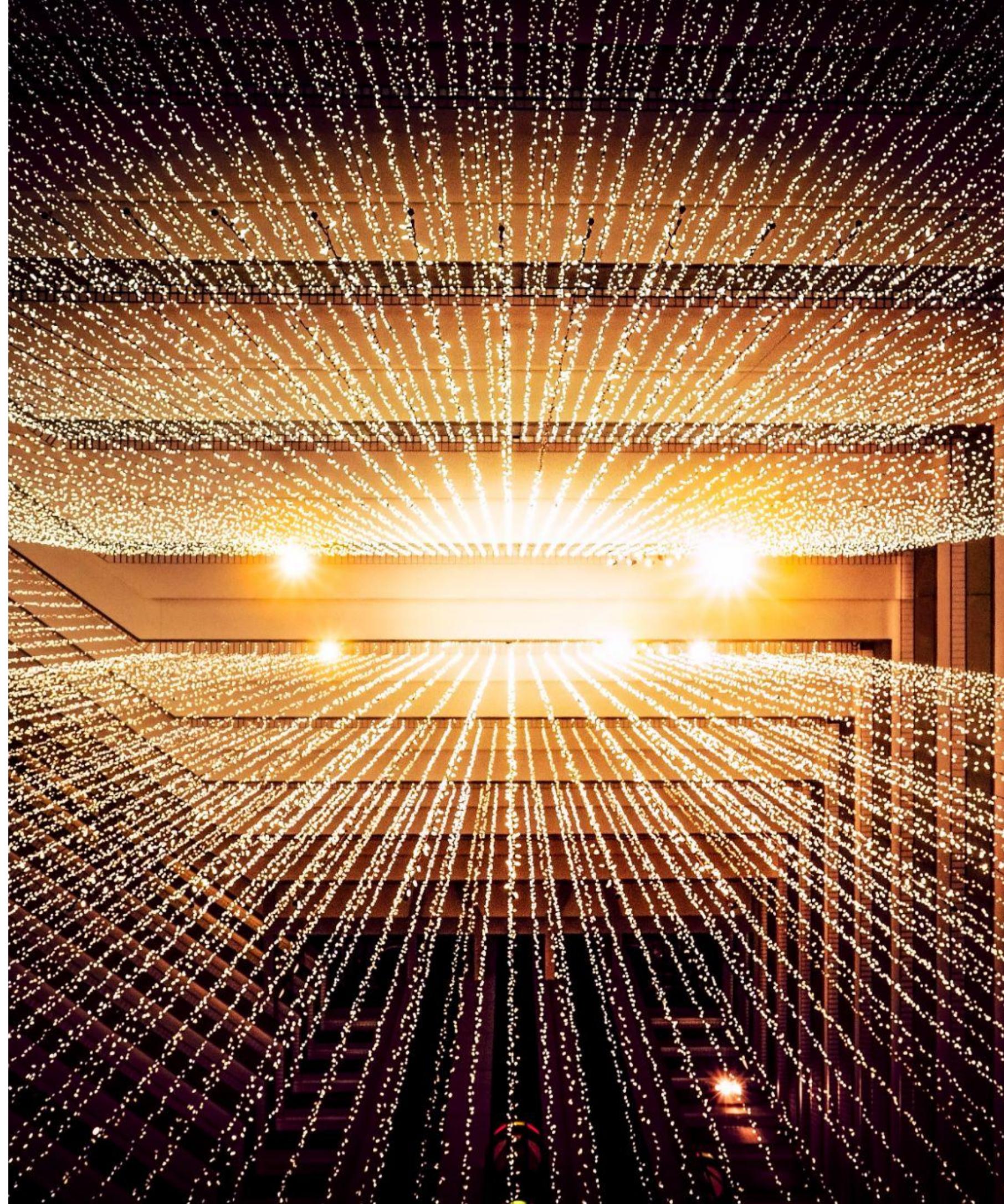
PROJECTS TO PROCESSES

How DM and DG initiatives grow within an organisation, and the usual progress



SOPHISTICATION LEVELS

The scale that organisations go through in terms of levels of sophistication, in detail



DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **4 KEY PRINCIPLES**

Throughout this course, we will cover many activities, practices and concepts, but there are certain underlying principles common to all of these, which are important to understand.

We can identify 4 main principles of relevance:

1. Data as assets
 - Measuring data in the same way as other “hard” assets;
2. Data monetisation
 - Data can have financial benefits if handled correctly;
3. Information Maturity
 - Measuring how sophisticated your organisation is;
4. Data Governance (DG) as a business function, not IT
 - Governing data is everyone’s responsibility, not an IT job;

DATA LITERACY AND CONSIDERATIONS **4 KEY PRINCIPLES**

The first concept is data as assets, and it's crucial for any DG or DM initiative to succeed. In order to benefit and profit from data, they must be seen as assets, just like inventory, materials, products, and others.

- It consists of applying the same rigor, processes and standards to measure data as you do for other assets in the company (IT systems, employees, others);
- Involves how data are created, stored, used, destroyed...
- This turns data from something intangible into something tangible, which can be measured, monitored, tracked;
- One of the key success factors when implementing a DG/DM initiative is precisely to get stakeholders thinking in this way;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS 4 KEY PRINCIPLES

The second principle stems directly from the first one. If data can be measured as hard assets, then we can extract monetary value from them.

- Usually, the benefits stemming from managing data well are the opposite of the costs of not managing them well;
- These are of three main types:
 - Increased revenue (for example, by creating more accurate products with better data);
 - Decreased operational costs (for example, saving on unnecessary tasks or activities, or reducing waste);
 - Decreased fines and regulatory risk (for example, having a lot less compliance breaches due to good data);
- Naturally, data only generate money if well-handled;

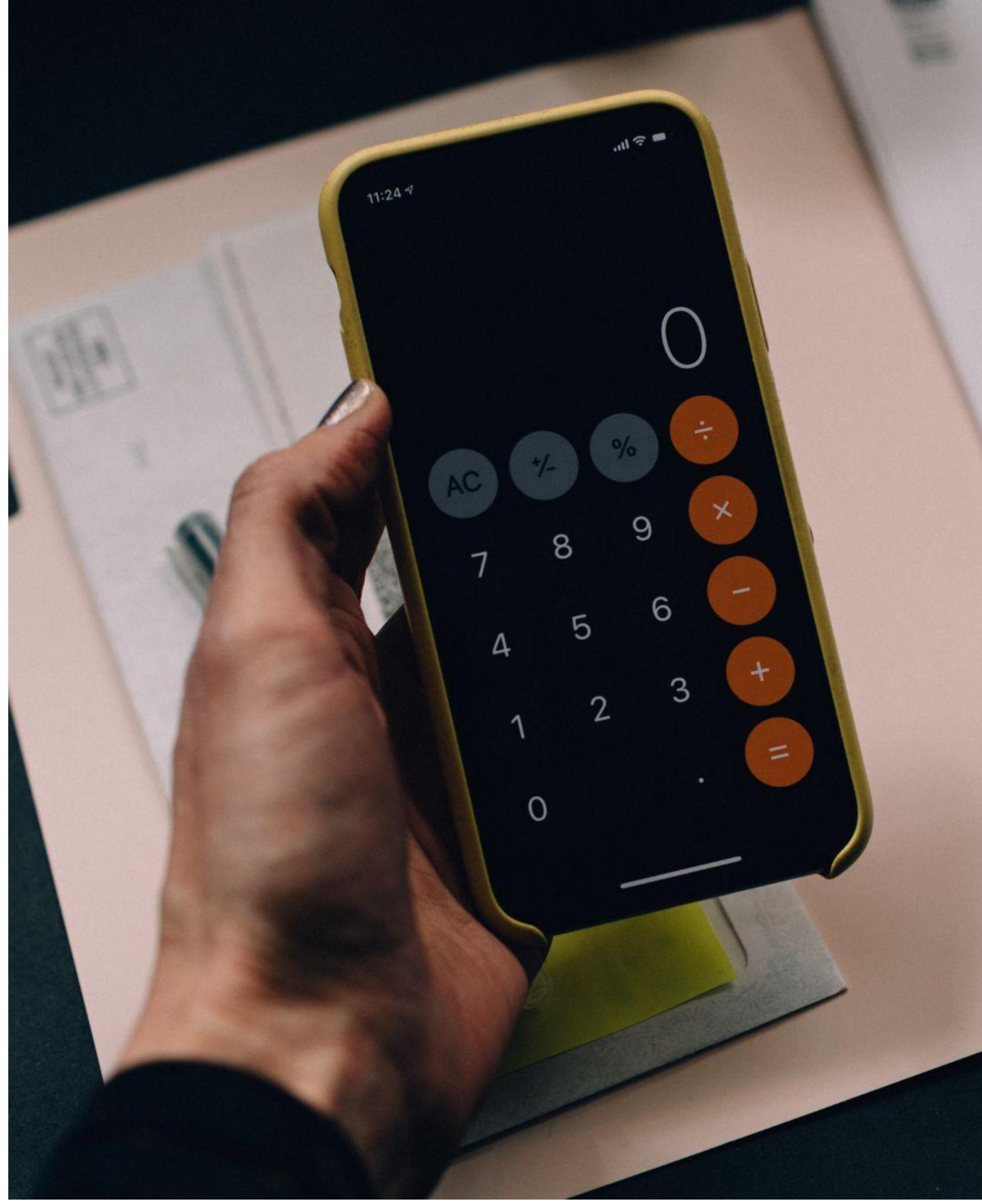
DATA LITERACY AND CONSIDERATIONS

4 KEY PRINCIPLES

The third principle is Information Maturity. That is, before starting any DG or DM activity, the organisation measuring itself in terms of where they are in terms of data.

- Information maturity is a more general term, which can encompass many different specific activities;
 - How data-literate stakeholders and executives even are;
 - How sophisticated are the policies for retaining data, for example sensitive data, including PII data (and for different departments);
 - How much information there is about data (metadata) for different apps, data stores, departments, and others;
- It's not a direct representation of how effective data are managed, but it's a powerful indicator;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS 4 KEY PRINCIPLES

Finally, regarding Data Governance (DG) in specific, it's very important to clarify that this is a business function, not IT.

- Many companies believe, erroneously, that data equal IT, and they're the sole responsibility of IT, which is an error;
- DG should be considered a business function, which must be performed by different departments in terms of their own data;
- Initially, DG may have a dedicated team, but the goal of it is to eventually "disappear" into the organisation;
- In fact, with a high level of sophistication, DG exists as just another function within each department, such as hiring, budgeting, performance management, or any other;
- Each department has a data owner, data stewards, etc;

DATA LITERACY AND CONSIDERATIONS

4 KEY PRINCIPLES

EXAMPLES

/01 INDUSTRY AS AN INDICATOR

The industry a company is in can inform about regulation, which informs about the maturity level. Companies in finance are data-aware from day one.

/02 IT LOVE-HATE

Despite the fact that data governance is not an IT function, but business, you would think IT supports it. In many cases, it's one of the toughest departments.

/03 ACTUAL MONETISATION

In specific sectors, such as marketing, besides saving in operations and improving revenue, data can actually be sold, generating revenue directly.

DATA LITERACY AND CONSIDERATIONS

4 KEY PRINCIPLES

KEY TAKEAWAYS

/01 DATA AS ASSETS

In order for data to be well-handled and managed, they must be considered assets, with the same rigor as other assets in the organisation.

/03 INFORMATION MATURITY

Information maturity is an indicator, which can be an important benchmark for an organisation to measure how it already handles data in the present.

/02 DATA MONETISATION

When data are managed as assets, they can provide financial benefit, either as increased revenue, reduced operational costs, or reduced compliance risk.

/04 DG AS BUSINESS, NOT IT

In order for Data Governance (DG) in specific to succeed, an organisation cannot consider it an IT function, but a business one for every department/LoB



DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

Before diving deep into Data Management and Data Governance, it's important to clarify the different definitions related to data, including:

- Data Quality (DQ) (the lack of which causes problems);
- Data Management (DM)/Information Management
 - Changing and manipulating data;
 - Enterprise Information Management (EIM);
 - Master Data Management (MDM);
- Data Governance (DG)
 - The policies and procedures directing DM;
 - Data Stewardship (the implementation of DG);
- Data Science;
- One application of data to obtain insights. Usually, stats.;

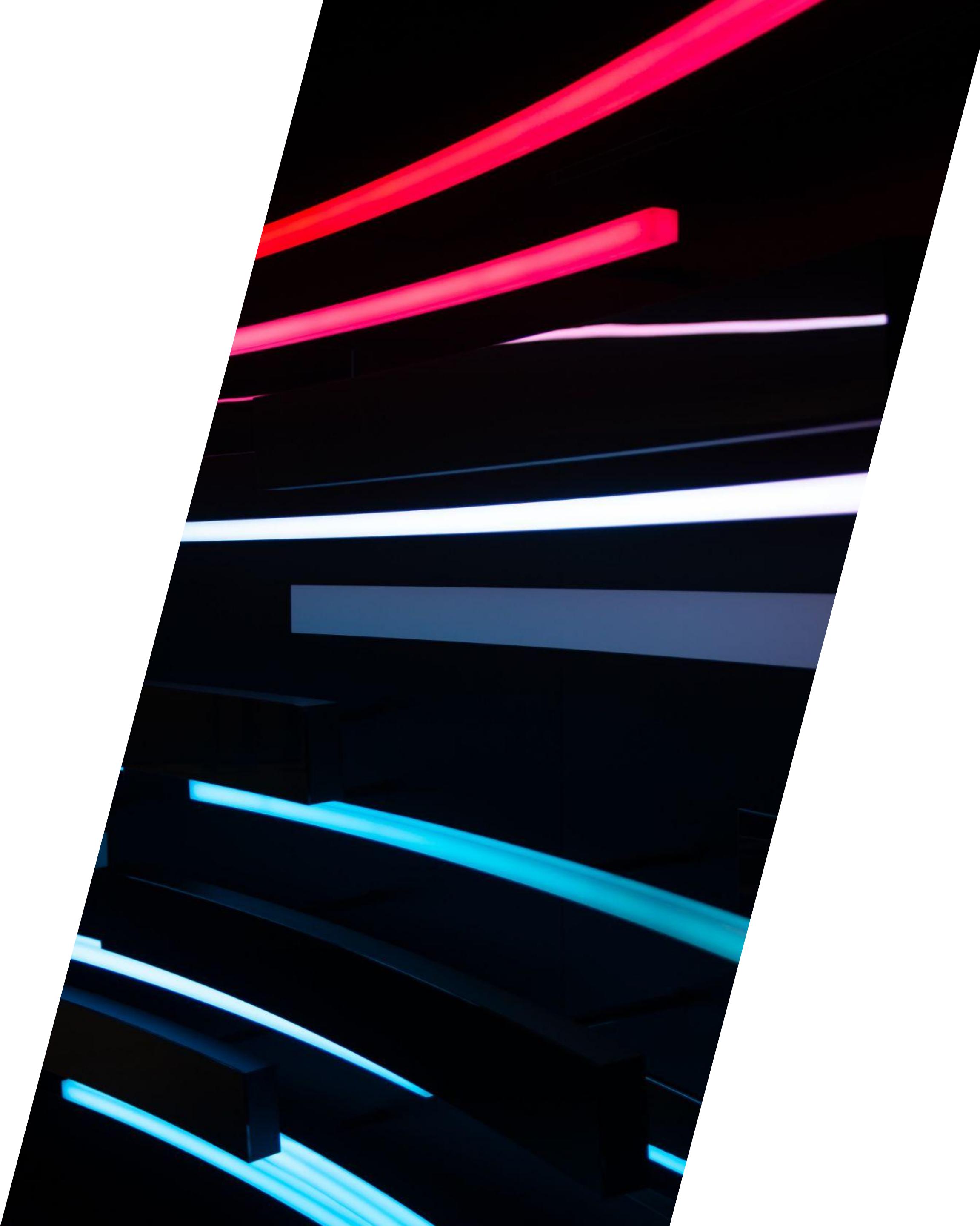
DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

Data Quality (DQ) is the name given to how appropriate data are for their use. Considering data as assets, the higher the data quality, the higher the utility or value of those data.

Lack of data quality in organisations is precisely the root cause of various problems, and it's what motivates both Data Management (DM) and Data Governance (DQ) programs.

Problems stemming from lack of DQ can include:

- An application does not reach its goal due to bad data;
- The company purchased data profiling tools... but not used;
- BI and analytics reports are not trusted by executives;
- Data scientists mostly fix data, instead of using them;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

Data Management (DM), or Information Management, is the the set of actions that directly change data in the company.

- It can impact any stage of the information life cycle. How information is acquired, processed, disposed of, etc;
- Projects to “fix” data are precisely this;
- “Data” and “information” differ, but DM/IM is done similarly;

DM can be performed “locally”, with specific projects (for example, to “fix” Data Quality problems) or impacting the whole organisation (enterprise level). The latter is known as EIM, or Enterprise Information Management.

DATA LITERACY AND CONSIDERATIONS

DATA DISCIPLINES

A specific type of DM is something called MDM, or Master Data Management. It consists of identifying the key information elements of a company and managing them as a “single source” of truth that all other data refers to.

- For most organisations, these will be customer data;

Data Governance (DG) is the set of policies and practices that direct Data Management (DM) efforts.

- Think of DM as the actions, and DG as the “guidebook” directing those actions;
- It is not done by the same people. In fact, DG should “disappear” into the organisation;
- Think of DM as the execution and DG as the auditing of it;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

Specifically related to Data Governance is the concept of Data Stewardship, which is the actual implementation of DG. In other words, guiding the policies and procedures in practice.

- If DM is the execution, then DG is the set of rules guiding that execution, and DS is the implementation of the rules;
- It usually involves analysing and validating rules, filling metadata, tracking DQ issues, and so on;

DM and DG are distinct, but related, and do interact:

- The better the policies and procedures from DG, the more effective the DM activities are (and the less DQ problems);
- Lessons learned from specific DM projects can provide insights that can be applied in “general” as DG P&Ps;

DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

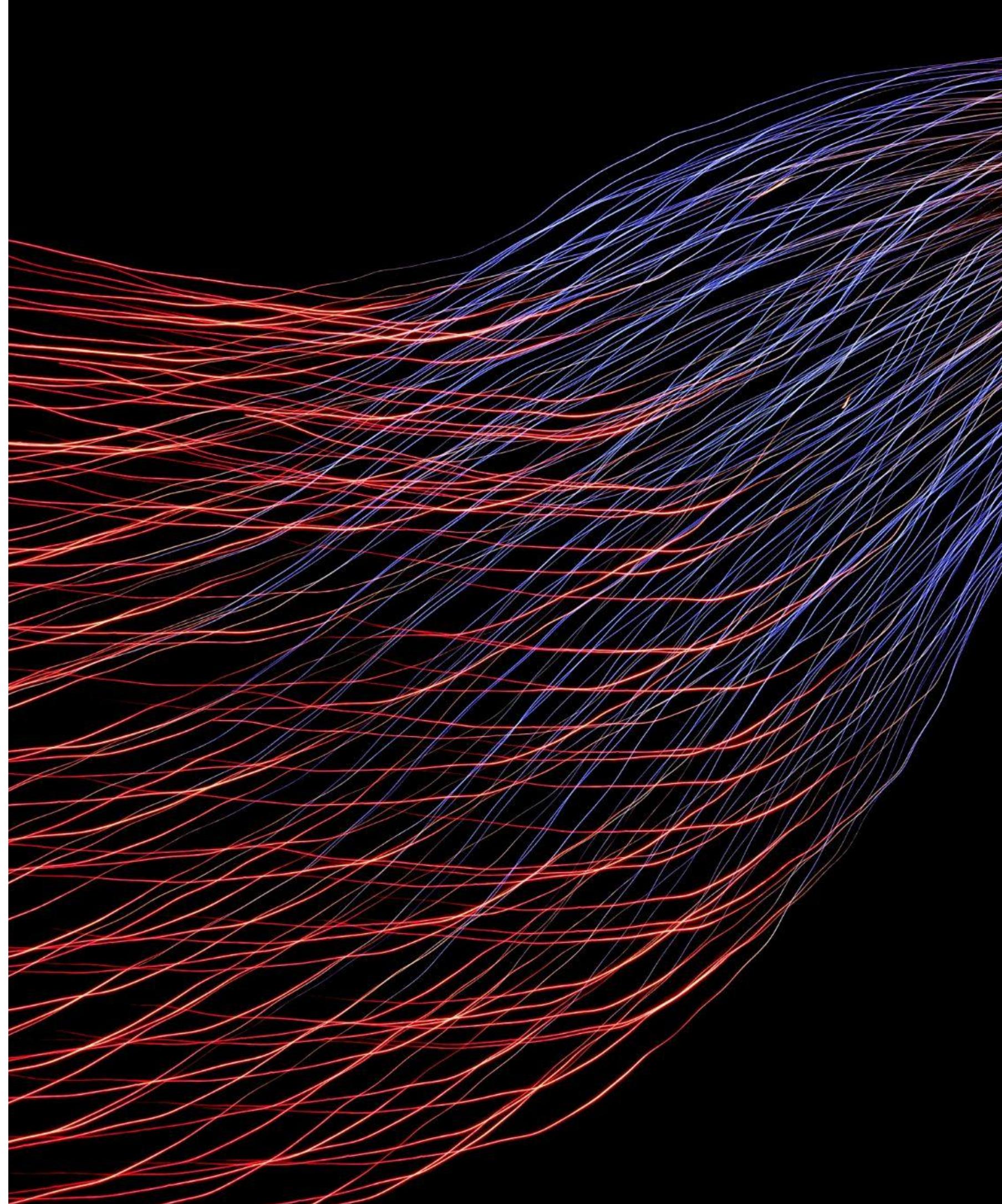
Both DG and DM usually exist as projects first, and usually expand to the enterprise level.

- “Fixing” customer contact information from an app is a specific DM project, but can branch out;
- Setting specific policies for entering data in a department database is a DG project, but can branch out;
- Both are usually part of a bigger program (DG or DM one);

Both DG and DM should not be oriented by the tools, but the business.

- Data should not be “managed for the sake of being managed” or “governed for the sake of being governed”;
- Not clear business use = no use starting a project;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DATA DISCIPLINES**

There's also common misconceptions in terms of Data Management and Project Management.

- Some people understand (wrongly) that DM is a part of PM, specifically for managing data. It is not;
- This confusion is accelerated by the fact that DM usually starts with projects;

These are two completely different disciplines:

- Project Management is the management of any type of project in a company (employee enablement, app development, marketing events, etc);
- Data Management is the set of activities for managing organisational data (mostly to improve its quality);

DATA LITERACY AND CONSIDERATIONS

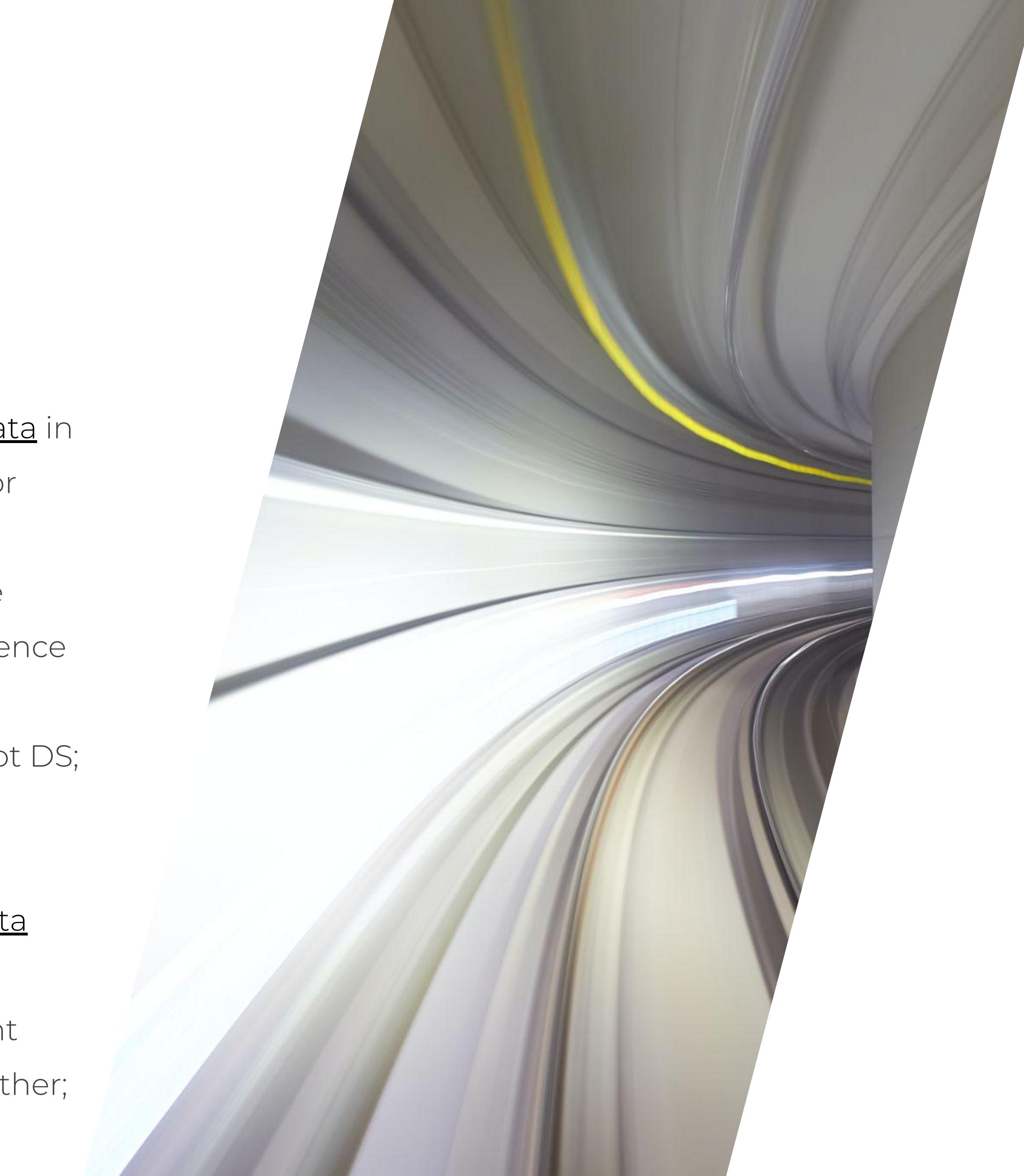
DATA DISCIPLINES

Finally, Data Science can be considered a usage of the data in the organisation, in order to obtain some sort of insight or conclusion that generates value.

- While DM and DG deal with the whole information life cycle (obtaining, storing, maintaining, using), Data Science focuses specifically on using the data;
- If due to bad organisation data scientists do DM, it's not DS;

Due to lack of DQ in many databases and systems, it's frequent for Data Scientists to actually end up "fixing" data most of their time instead of gathering insights.

- In other words, they are performing Data Management instead of Data Science. One activity "spills" onto the other;



DATA LITERACY AND CONSIDERATIONS

DATA DISCIPLINES

EXAMPLES

/01 METADATA'S IMPORTANCE

Metadata are crucial to data governance, detailing both business and technical aspects of data. Data stewards are the ones that mostly deal with metadata.

/02 DG AND COMPLIANCE

Data governance is intimately tied to compliance and many times motivated by it - in many industries, governing data properly is required to comply.

/03 MANUAL AND AUTOMATED

In many cases (but not all), Data Management is manual and Data Governance is automated. DM fixes something, DG finds an automated way to prevent it.

DATA LITERACY AND CONSIDERATIONS

DATA DISCIPLINES

KEY TAKEAWAYS

/01 DQ, DM AND DG

Data Quality is the indication of how useful data are in terms of value. Data Management is actively changing it. Data Governance is the set of P&Ps that direct it.

/02 EIM AND MDM

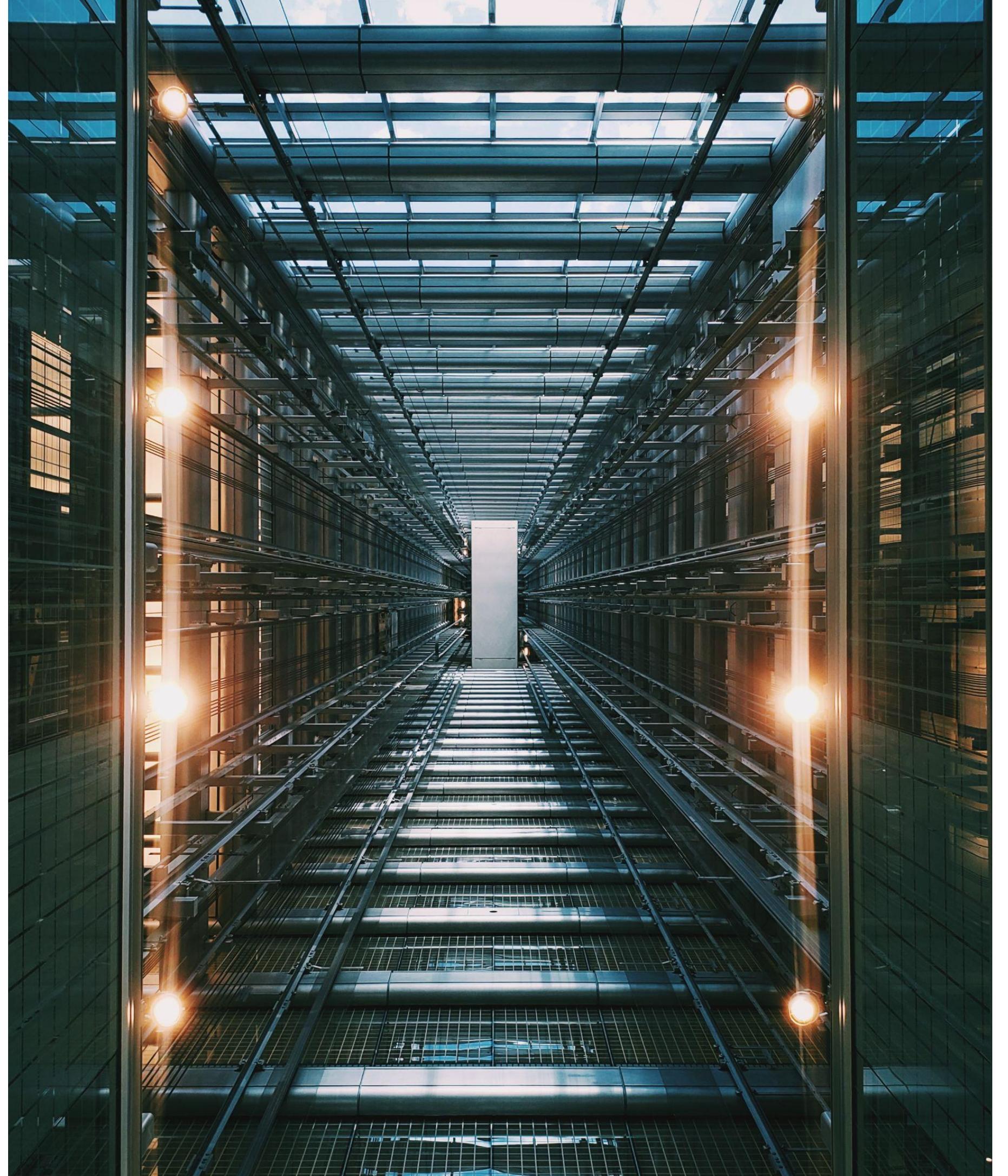
Within Data Management/Information Management, EIM is the management at the full enterprise level, and MDM is the management of Master Data.

/03 PROJECT TO PROCESS

Usually, both Data Management and Data Governance start as small projects to show value, then insights are “generalised” to the full enterprise.

/04 BUSINESS, NOT TECH

It's important to note that both DM and DG activities should be motivated by a clear business need. Not due to being in love with the technology or “just because”.



DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DG/DM KEY ACTIVITIES**

Let's clarify some of the key activities in both DM and DG:

- Data profiling (DM):
 - Examining the quality of specific data;
- Data dimension definition (DM):
 - Examining the “angle” through which DQ is analysed;
- Data remediation (DM):
 - Actually fixing low-quality data;
- Data requirement and expectation definition (DG):
 - Defining the value ranges and formats required for data;
- Data policies (incl. standards) and controls (DG):
 - Defining what's required in terms of processes;
- Data process controls and audits (DG):
 - Verifying whether processes are working;

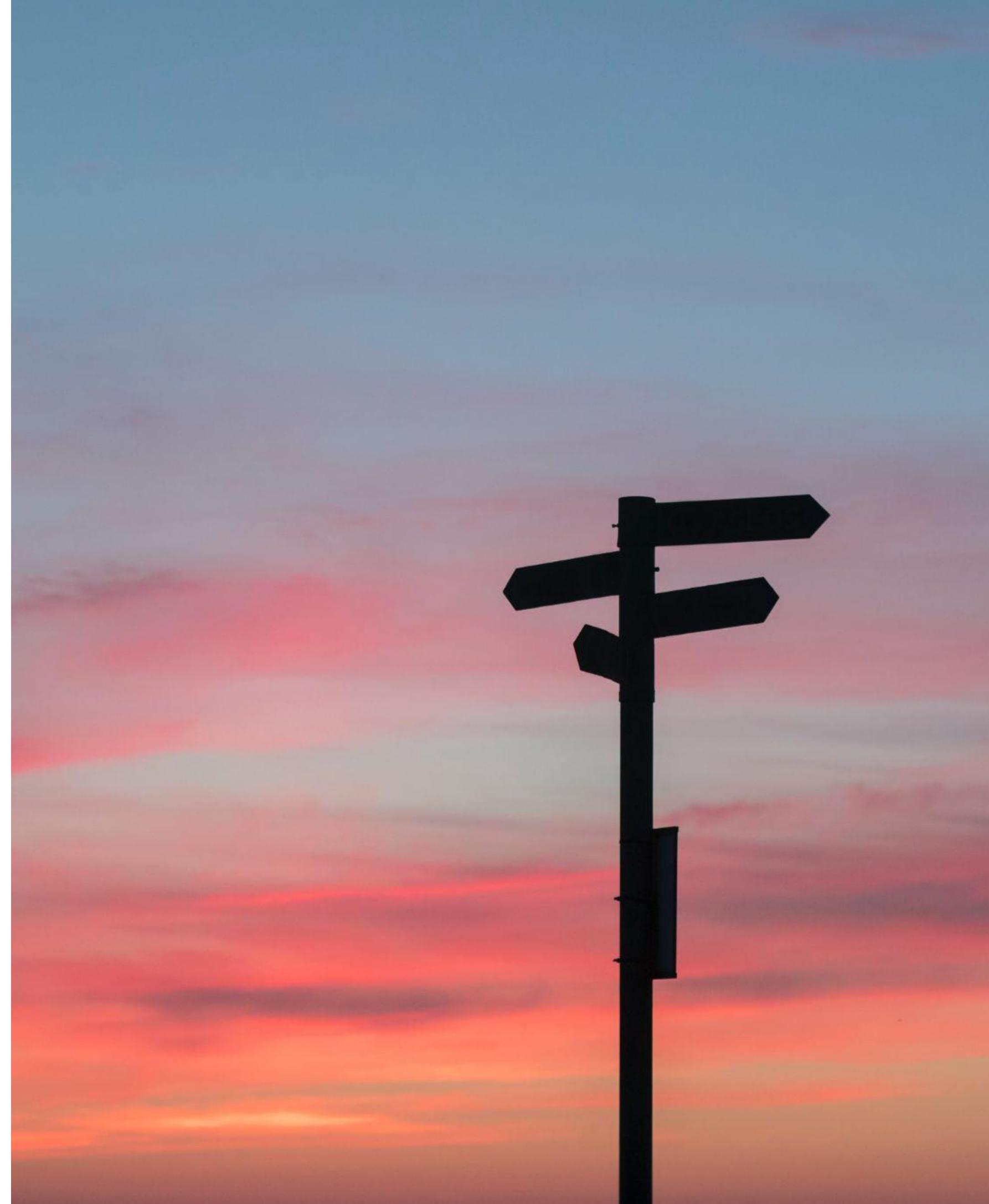
DATA LITERACY AND CONSIDERATIONS **DG/DM KEY ACTIVITIES**

Data profiling is a type of assessment activity that analyses the level of DQ in a specific DB or medium. It's usually the activity that reveals data problems, followed by remediation.

- Data profiling has to be done according to a specific dimension. For example, “Completeness” is a dimension (how much data are we missing?);
- Other important ones include “Currency” (are the data still relevant?) or “Consistency” (are these values the same in different data sources?);

Data remediation is the actual process of changing data to increase its quality (either by changing the current data, or adding new information that increases their quality).





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DG/DM KEY ACTIVITIES**

Data governance sets policies. That is, processes that have requirements embedded in terms of acquiring/using/etc data.

- The most common type of policy are standards, which are policies that define specific formats of exchange (for example, currencies, temperatures, etc are standards);
- “Customer addresses must always be inserted, containing a full postcode in a 6-digit field”;
- These are usually derived from the expectations in terms of values. DM provides feedback that states, e.g., “This temperature field must always be a number from 0 to 100”;

Then, controls and audits are simply independent activities that verify how actual operations comply with the policies.

DATA LITERACY AND CONSIDERATIONS **DG/DM KEY ACTIVITIES**

Example: Anti-Money Laundering (AML) compliance in banks:

- Data Profiling: How many transactions are missing CTRs?
- Data Dimensions: Completeness (are values missing?);
- Data Remediation: Adding CTRs to nonexistent transactions;
- Data Requirement/Expectation: “Every transaction entity instance in the DB with >\$10,000 “Value” must have an existing instance of the CTR entity associated”;
- Data Policy/Standard: “Every transaction above \$10k must have an associated CTR when processed”;
- Data Controls/Audit:
 - Automatic mechanism to alert on new >\$10k transactions without CTR in real-time;
 - Independent human audit of bank transactions;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **DG/DM KEY ACTIVITIES**

Example: Temperature readings in physics research project:

- Data Profiling: How many readings are realistic?
- Data Dimensions: Accuracy (do values make sense?);
- Data Remediation: Finding alternate sensor readings for values that don't make sense, and replace/average these;
- Data Requirements/Expectations: "Every temperature reading must be an integer between 0 and 100";
- Data Policy/Standard: "Every temperature reading inserted in the DB must have a valid value of 0 to 100 Fahrenheit";
- Data Controls/Audit:
 - Automatic alerts on temperature values that are not realistic;
 - Periodic human inspection of sensors;

DATA LITERACY AND CONSIDERATIONS

DG/DM KEY ACTIVITIES

EXAMPLES

/01 IT'S ALL ABOUT DQ

Both Data Management and Data Governance activities revolve around DQ. DM aims to directly increase it in the short term, and DG in the long term.

/02 MORE TO IT THAN “FIXING”

We'll see this later, but there are more possible actions when a DQ problem is found than just remediation. We can find the root cause, and set up monitoring.

/03 SQL DATABASE TRIGGERS

Although this is not feasible for all types of fields, DB triggers can be very useful automated tools. For number values, they can instantly trigger alerts.

DATA LITERACY AND CONSIDERATIONS

DG/DM KEY ACTIVITIES

KEY TAKEAWAYS

/01 DATA PROFILING

Data profiling is the name given to the activity of analysing specific datasets to determine the quality of the data within them.

/02 DATA DIMENSIONS

Data can be high or low quality depending on the dimension. Completeness measures how much data are missing, Currency measures how up-to-date, etc.

/03 DATA REMEDIATION

Data remediation is the name given to the process of “fixing” low-quality data through a process, which usually either replaces the data or adds more details.

/04 REQUIREMENTS + EXPECTS.

Feedback (usually, from profiling) informs of what the values ranges/formats should be for data. These are formally defined as the requirements for data.

/03 POLICIES AND STANDARDS

Policies are the actual processes of a firm that state “Data must do X” when acquired / used / other. Standards are policies made for exchange of info.

/04 AUDITS AND CONTROLS

Operations will tend to obey policies and standards, but it's crucial to verify. Controls alert of policy breaches, and audits independently verify them.



DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **PROJECTS TO PROCESSES**

Although DM and DG are two distinct disciplines (and, in a way, complementary), one common element in both practices is that both start being implemented as projects, which eventually become long-term processes.

- A business case is made that there is a DQ problem in terms of customer data;
- Two DM projects are started to remediate customer data issues in two specific departments;
- At the same time, thinking long-term, the company starts a DG project to set policies in terms of how customer data are created, stored and maintained;

The two types of projects can occur in parallel or together (for example, DM manages the data and DG “audits” it).

DATA LITERACY AND CONSIDERATIONS PROJECTS TO PROCESSES

Both DM and DG can start as small, isolated projects or as part of larger programs:

- A company that needs to radically change the level of oversight they have over data can start a DG program, with controls on how data are acquired, managed and disposed of in multiple business lines at once, with multiple projects;
- A company that detects problems with transaction data in multiple departments can start a DM program with DM projects for remediation in multiple databases;

Both of these can be for isolated departments/business lines, or they can be enterprise-wide.

- For example, within DM, MDM programs to establish a single “source of truth” for data are enterprise-wide;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **PROJECTS TO PROCESSES**

Eventually, both DM and DG projects (or programs) generate insights which are eventually converted into processes and policies for the organisation. For example:

- The DM projects executed detect problems with customer data, so the processes for customer data acquisition are changed so it never happens again;
- The DG projects executed detect that there is no oversight of how data are disposed of, so processes and policies are set in place to control data retention and disposal;

One key difference is that, as they mature, DM usually remains a discipline in the organisation, while DG “disappears” (it becomes “transversal” to all disciplines).

DATA LITERACY AND CONSIDERATIONS **PROJECTS TO PROCESSES**

Let's take an example of implementing GDPR compliance:

- The organisation wants to document all uses of Personally Identifiable Information (PII) on all applications;
- It starts with one DG project, where all uses of PII in the application are documented, with a small automated control to detect unauthorised use (e.g. an AC tool);
- The project is successful, and the project expands into 3 similar projects for apps in the same department;
 - Policies are documented, stewards are assigned;
- These are successful, and expand to a full DG program for multiple departments, affecting all applications of the enterprise with documented controls and policies;
- The program becomes enterprise-wide and is assimilated;

DATA LITERACY AND CONSIDERATIONS **PROJECTS TO PROCESSES** EXAMPLES

/01 VERTICAL OR HORIZONTAL

Projects can grow in “horizontal” terms, doing the same but for more teams/departments, or “vertically”, by doing more for the same team/department.

/02 DM REQUIRES DG

Usually, DG is “independent”. If you need DG, e.g. due to compliance, it doesn’t need DM. But in most cases, DM requires DG - to prevent the issues fixed by it.

/03 DG PIGGYBACKING

A corollary of the previous example is that DG can be easily started on top of current DM efforts. MDM, BI, analytics and others provide good DG opportunities.

DATA LITERACY AND CONSIDERATIONS

PROJECTS TO PROCESSES

KEY TAKEAWAYS

/01 PROJECTS TO PROCESSES

Although distinct disciplines (and with distinct focuses), both DG and DM usually start as projects and mature into processes (and/or policies).

/02 LOCAL OR GLOBAL

Both DG and DM can start as isolated projects, or part of bigger programs. They can focus on specific areas, or be enterprise-wide from the beginning.

/03 DG AS TRANSVERSAL

Regardless of projects or processes, DM usually remains as a “direct”, identifiable discipline, while DG “disappears”, being a requirement of all.



DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **SOPHISTICATION LEVELS**

Each organisation sits somewhere across a spectrum of sophistication in terms of DQ/DM. Grading the organisation helps it both assess where it currently is and define a vision for the future. Usually, there are 3 main stages:

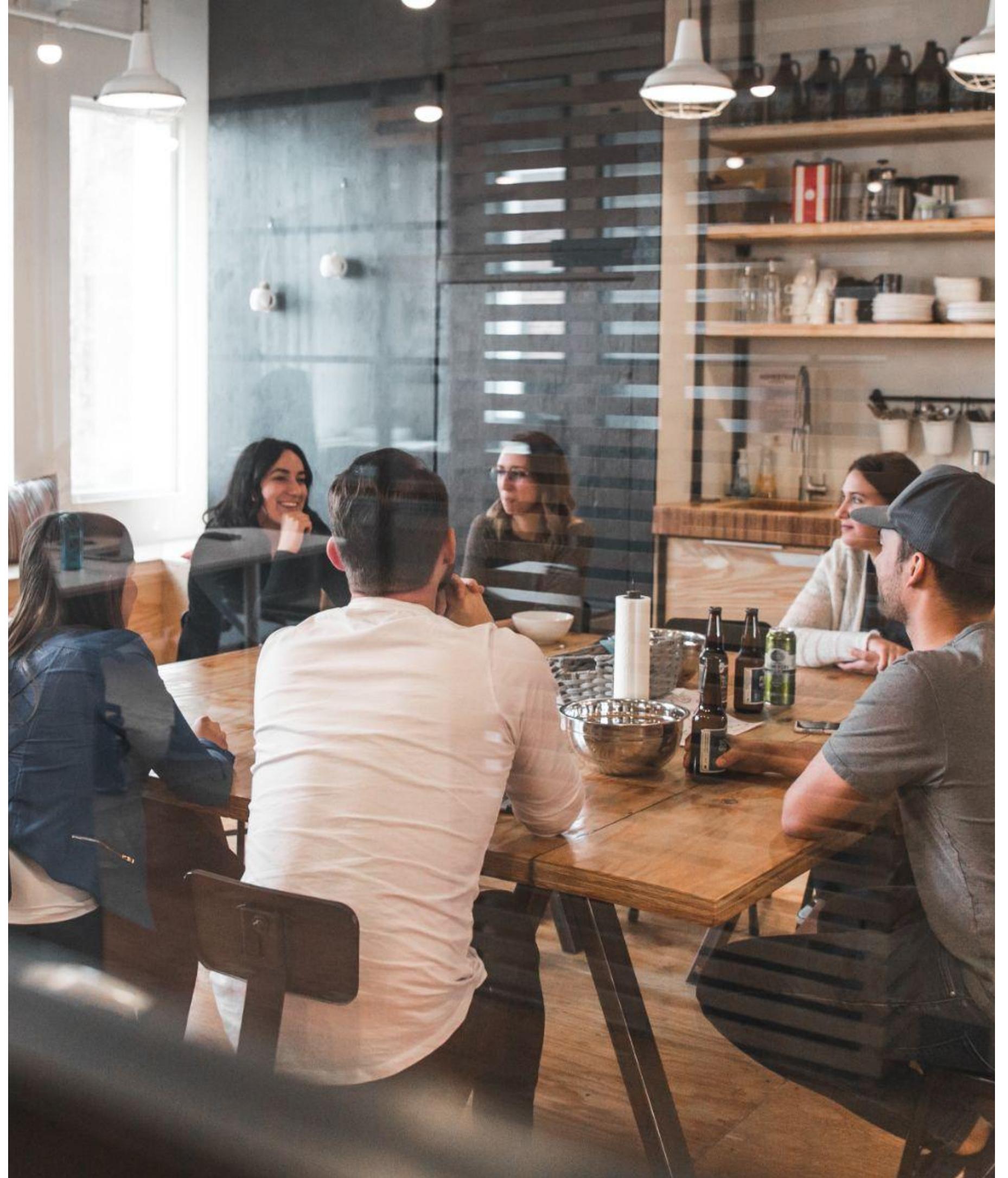
- Reactive
 - Fixes are local and immediate. There are no processes/policies, and fixes don't impact these;
- Intermediate
 - There may be local processes or policies driving DM, but fixes are mostly reactive and don't impact them;
- Proactive
 - Processes and policies are enterprise-wide, DG policies are established, tools and training are both defined;

DATA LITERACY AND CONSIDERATIONS SOPHISTICATION LEVELS

At the reactive stage, the company does not have any kind of policy or process for improving data quality.

- DM projects are usually reactionary, in real-time, with no defined tools, and vary based on the expertise of the analyst;
- There is little to no awareness of the real cost of DQ problems to the organisation (or it is ignored). There is also little to no senior management buy-in or support;
- There are usually no DG policies for any app or department;
- When DM projects remediate data problems, the fix is only immediate and temporary, and there is no root cause analysis to change the process which caused the flaw (which means, in many cases, problems are recurring);





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **SOPHISTICATION LEVELS**

At the intermediate level, there is usually a local level of standardisation. There may be small DM and DG projects causing local impact, but they're not institutionalised yet.

- DM projects don't just remediate data but their feedback is used to influence data requirements in apps/data stores;
- There may be local DG policies, with local owners, possibly stewards, and explicit DQ expectations and requirements;
- The dimensions of data and the tools used are documented, even if only at small or local levels;
- There is some level of management support, even if low. Business impact is correlated with low DQ in some ways, even if awareness/support is only local;
- DM contributes to incremental change in processes;

DATA LITERACY AND CONSIDERATIONS **SOPHISTICATION LEVELS**

At the Proactive level, DM and DG are usually fully integrated on an organisational level:

- Data flaws are not only remediated in a reactive manner, but identified beforehand and fixed at early stages. The feedback from these errors contributes to changing system design and improve DQ on an enterprise-level;
- DG policies and processes are pervasive, with expectations defined at all levels, for all stakeholders and departments, with owners and stewards, and controls/auditing;
- DM tools, training requirements, data dimensions, and data expectations/requirements are fully documented;
- There is extensive support from senior management and awareness of DQ-business impact correlation;





DATA LITERACY AND CONSIDERATIONS

DATA LITERACY AND CONSIDERATIONS **SOPHISTICATION LEVELS**

This type of sophistication level does not need to be measured at 3 levels. Organisations can (and should) create their own sophistication scale, where they can measure both their current state and their desired vision.

- Existing ones: EIM Maturity, DG Maturity, etc;

For example, a 5-step scale could be:

- Reactive, non-standardised;
- Reactive, standardised locally;
- Proactive, standardised locally;
- Proactive, standardised in large-scale;
- Proactive, standardised enterprise-wide;

Whatever the scale, grading the organisation is necessary.

DATA LITERACY AND CONSIDERATIONS **SOPHISTICATION LEVELS**

Naturally, the level of sophistication is intimately tied to the progression from projects to processes:

- A local DM project is started to identify missing firewall documentation (data profiling) and obtain it (remediation);
 - Tools are not defined, mostly ad-hoc individual effort;
- The project works and is expanded into a 4-project program for HW system documentation in 4 departments;
 - Tools are formally defined, so are data requirements
“Every system must have 1 documentation file”;
 - May feed back to DG as a policy: “Upon purchase, secure documentation for all hardware equipment”;
- The program eventually becomes enterprise-wide, with defined tools, roles, responsibilities, data requirements;



DATA LITERACY AND CONSIDERATIONS

SOPHISTICATION LEVELS

EXAMPLES

/01 CENTRAL VS. “SHADOW IT”

Shadow IT is the name given to “unofficial” databases, Excel spreadsheets, and other data stores. The more there are centralised policies, the less shadow IT.

/02 TOOLS COME LATER

There are many organisations that buy the tools first and don’t know what to do. Whether in DM or DG, it’s crucial to first define actions, and only then tools.

/03 PEOPLE TO PROCESSES

In early maturity stages, most projects may not even be official and just done by motivated personnel. Maturity comes from moving to replicable processes.

DATA LITERACY AND CONSIDERATIONS

SOPHISTICATION LEVELS

KEY TAKEAWAYS

/01 THREE MAIN LEVELS

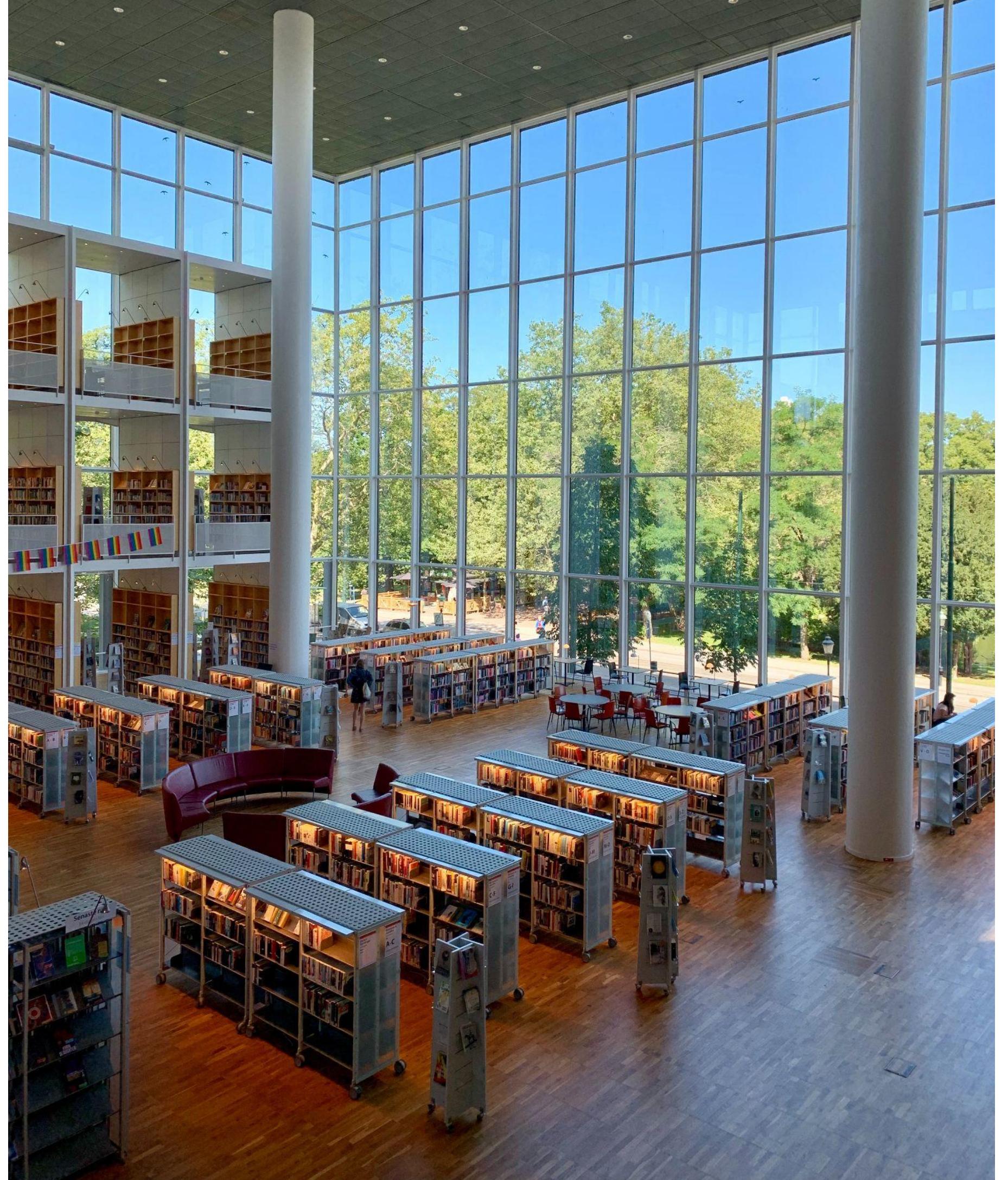
An organisation's sophistication can usually be measured in terms of 3 main levels: reactive, intermediate, and proactive.

/02 STANDARDS AND SIZE

Growing in sophistication, an organisation grows in terms of how standardised/automated DM/DG is, as well as how much of the organisation it impacts.

/03 ASSESSMENT IS CRUCIAL

Whether an organisation uses a 3-grade system, a 5-grade one, a 10-grade one, or any other, just knowing where it is (and what's the goal) is important.



LIFECYCLES

LIFECYCLES THE INFORMATION LIFECYCLE

It's important to understand how information flows through an organisation, because there may be different security and privacy policies and/or processes at different stages of the lifecycle, as well as different controls to enforce these.

A simplified version can stem from the CRUD framework:

- Create. Information is created/captured in a system.
Acquired, entered manually, or by other means;
- Read. Information is used by people. Direct use, reports, analytics, BI, AI/ML, or other data products;
- Update. Information is updated and changed. Salespeople updating client info, projects progress being updated, etc;
- Delete. Information is deleted (if unneeded/time elapses);

LIFECYCLES

THE INFORMATION LIFECYCLE

Although these are the “four basic” stages, there are more advanced frameworks, such as Danette McGilvray’s POSMAD framework:

- Plan. Architecture and design are planned for considering the information needs;
- Obtain. Information is obtained in some way;
- Store/Share. Making information available for use;
- Maintain. Information is updated, cleansed, transformed;
- Apply. Retrieving and using data. Reports, transactions, management decisions, others;
- Dispose. Archiving or deleting data due to some condition;

This is a similar model that takes the same fundamental stages but “unfolds” them into specific sub-stages.





LIFECYCLES

LIFECYCLES THE INFORMATION LIFECYCLE

If we consider the basic CRUD framework, let's examine some important factors of data at each stage:

- Create. Elements such as the lineage and consistency of information matter (e.g., Does it fit our quality needs? Formats?) as well as the trustworthiness of the source;
- Read. Elements such as the completeness and timeliness of information must be considered (e.g., Do we have the necessary data?) as well as AC and permissions for users;
- Update. Elements such as usage policies matter (e.g., Can I edit data? Will I affect DQ if I do so?);
- Delete. Elements such as privacy and regulation matter (e.g., Are we keeping data for the required time? Are we deleting them immediately once not needed anymore?);

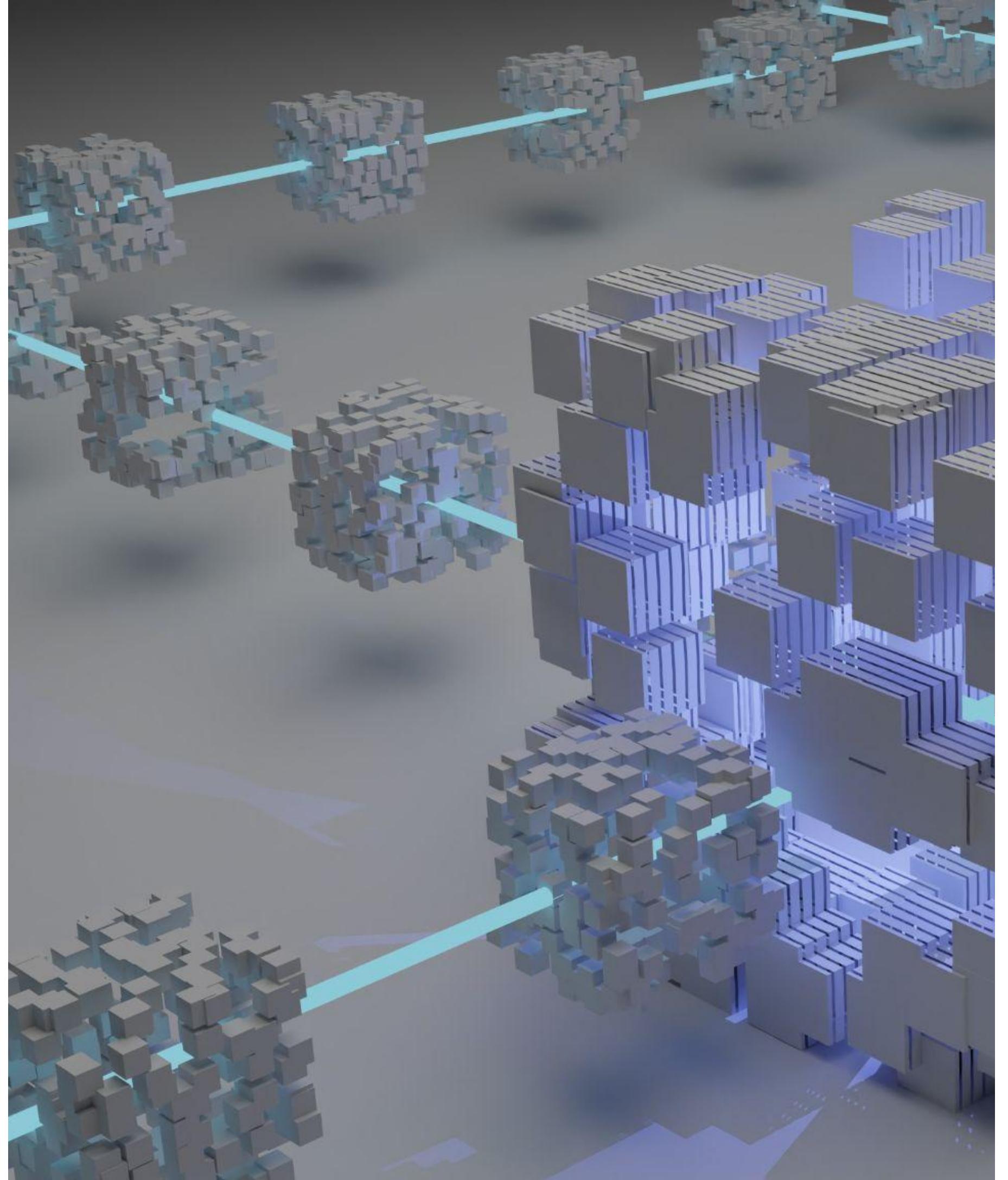
LIFECYCLES

THE INFORMATION LIFECYCLE

For example, let's say we're talking about customer data in an organisation:

- Create. Customer information may be inserted by salespeople when closing deals, either through one single solution (e.g. Salesforce), or different apps/interfaces;
- Read. Customer information may be used internally to generate invoices, analytics, direct consultation, and/or management decisions (one or multiple interfaces);
- Update. Customer information may be updated by salespeople upon account changes, by own customers if a portal is available, and/or by other users;
- Delete. Customer information may be deleted when a customer churns, and/or retention timeframes are up;





LIFECYCLES

LIFECYCLES THE INFORMATION LIFECYCLE

Another data example, for Suspicious Activity Reports (SARs) as part of AML/CFT compliance in banks:

- Create. An SAR is created by a bank employee, triggered by suspicious activity in an account. May be as data, a physical document, and through one interface or multiple;
- Read. SARs may be consulted for fraud statistics, and/or to collaborate with lawmakers during later investigations;
- Update. Since SARs are initially filed, there probably won't be a lot of changes, except maybe to add details in case of further developments by bank staff (probably using the same interface as for creation);
- Delete. SARs may be occasionally deleted if erroneous, and after the deadline for legal retention of the bank's country;

LIFECYCLES

THE INFORMATION LIFECYCLE

EXAMPLES

/01 DG POLICIES

Data Governance sets the tone for how data are obtained, used, discarded. There are usually policies at all stages, shaping how to obtain data, DQ concerns...

/02 DATA REMEDIATION

Data Management assesses the quality of the data present (data profiling). When it's not enough, a DQ issue is raised, and remediation usually follows.

/03 SYSTEMS AND SOFTWARE

When DG is embedded in an organisation, data policies shape system/SW development. For example, "how to acquire" data shapes the capturing SW.

LIFECYCLES

THE INFORMATION LIFECYCLE

KEY TAKEAWAYS

/01 THE 4 CRUD STAGES

The CRUD framework is a good place to start. Information is created, read, updated and deleted, and organisations have policies for actions at each stage.

/02 DATA QUALITY

At all stages, there are concerns regarding data quality, which may be measured in different ways during creation, editing, transformations, and so on.

/03 OTHER FRAMEWORKS

There are several different frameworks with different stages and levels of sophistication, such as the POSMAD model, but the 4 core stages remain the key.

DATA LITERACY AND CONSIDERATIONS

Learning about the essential data concepts, disciplines and activities that are crucial to DM, DG and other data-related disciplines



DATA LITERACY AND CONSIDERATIONS

In this module, we covered **six key topics** in terms of how data work and the usual practices that revolve around them:



4 KEY PRINCIPLES

Basic principles that any data activity relies on, such as data assets and monetisation



DATA DISCIPLINES

What is Data Management, Data Governance, and others, and how they are distinguished

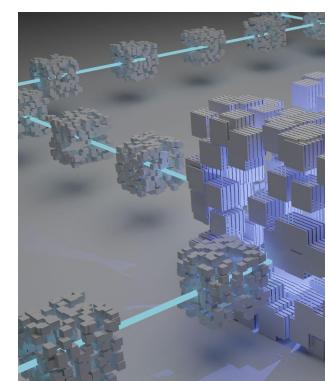


DG/DM KEY ACTIVITIES

Examples of usual activities in Data Management and Data Governance, and their differences

DATA LITERACY AND CONSIDERATIONS

In this module, we covered **six key topics** in terms of how data work and the usual practices that revolve around them



THE INFORMATION LIFECYCLE

What are the usual stages of data in an organisation, from creation to disposal, in detail



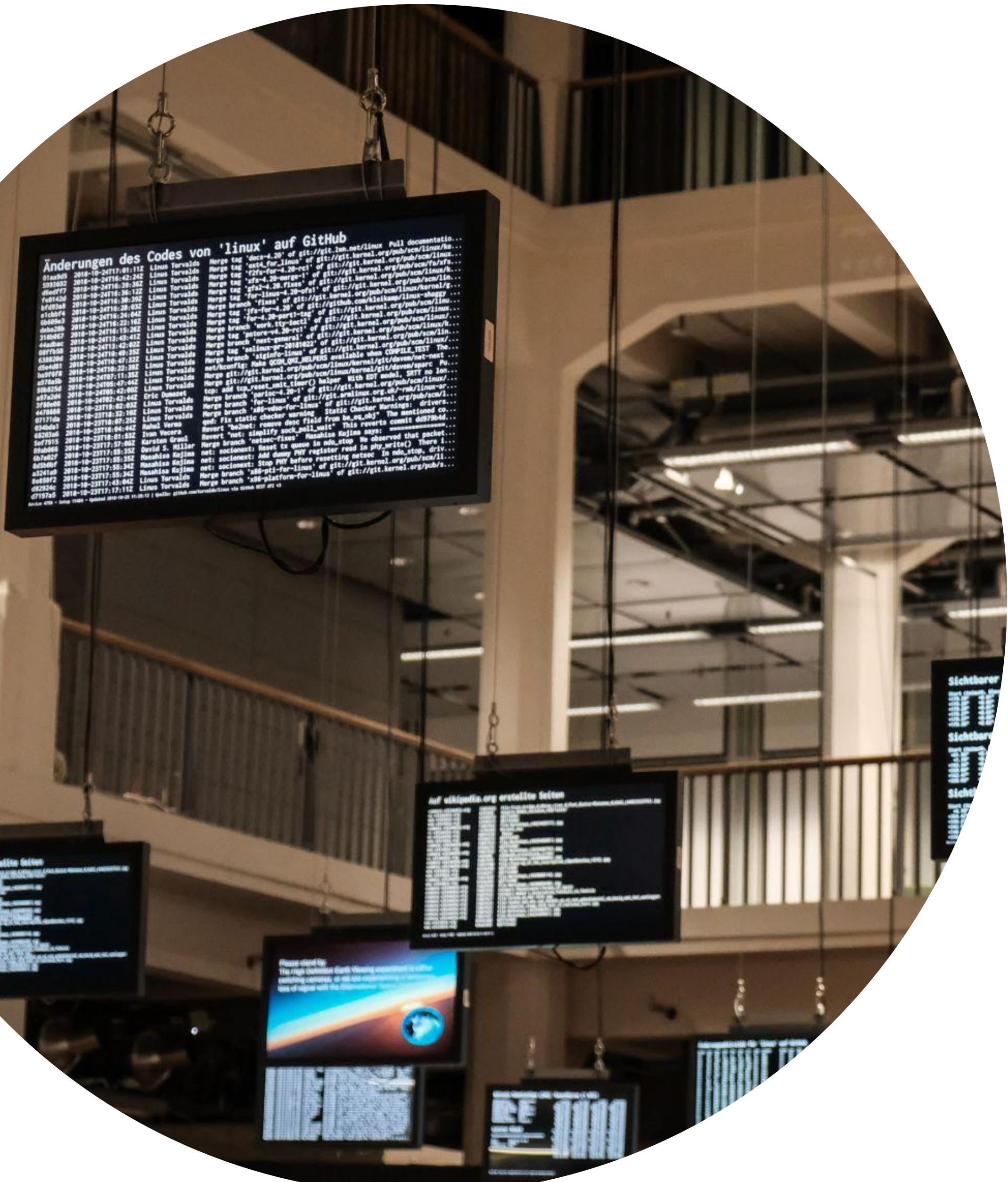
PROJECTS TO PROCESSES

How DM and DG initiatives grow within an organisation, and the usual progress



SOPHISTICATION LEVELS

The scale that organisations go through in terms of levels of sophistication, in detail



DATA LITERACY AND CONSIDERATIONS

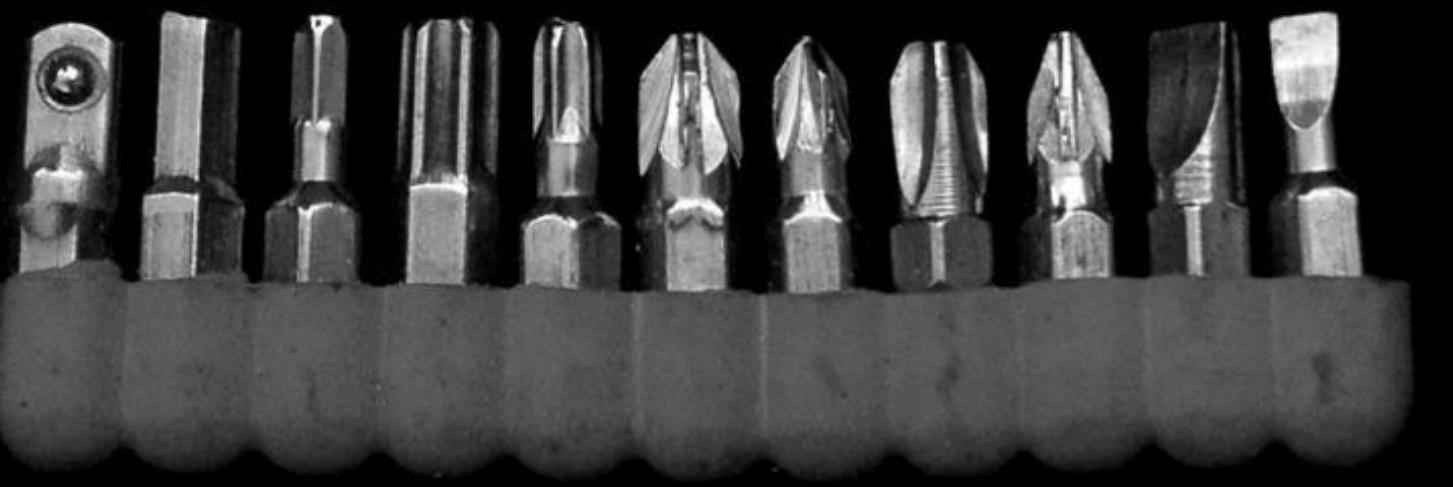
DATA LITERACY AND CONSIDERATIONS CONSOLIDATING

Some questions you can ask yourself to consolidate the knowledge in this module include:

- What specific discipline is focused on filling metadata and tracking DQ issues?
- When monitoring and prevention systems are set in place to prevent future DQ issues, is that DM or DG?
- Which specific practice, as it grows, must be absorbed by the business and become a function of each department?
- Which data discipline usually performs the direct remediation of data? And which documents the rules and validity requirements for data?
- If results from data remediation are used for automated data validity tests, which part is DM and which is DG?

DATA AND DQ

Learning about how data and Data Quality work, as well as the tools, activities, processes used to both assess and improve Data Quality

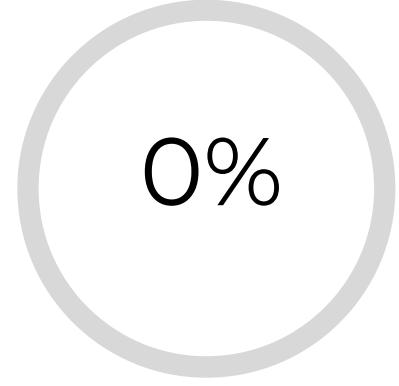


DATA AND DATA QUALITY



DATA LITERACY & CONSIDERATIONS

Covering the basics of data. What are the different data disciplines, what are the essential principles to handle data, what are the sophistication levels of organisations...



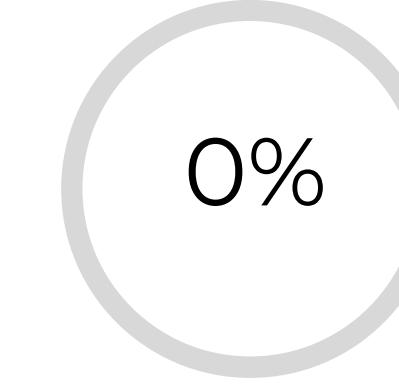
DATA GOVERNANCE

Covering how data governance works, from classifying data to setting policies and other activities, the roles and responsibilities, and the DG implementation process.



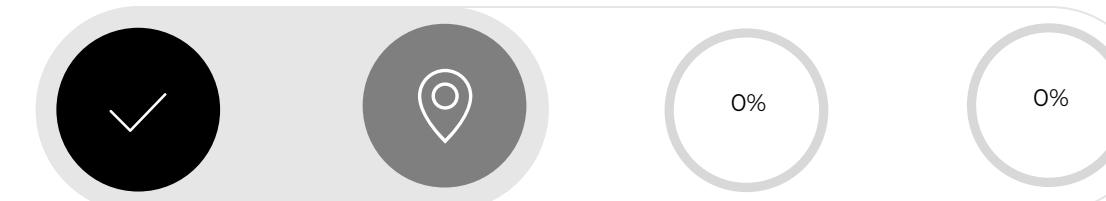
DATA AND DATA QUALITY

Covering specifically what data are and how to improve their quality. Data types, values, structures, and how to improve data quality through profiling and remediating.



DATA SECURITY, PRIVACY, ETHICS

Covering the different types of privacy and security controls that can be applied to data to protect them, as well as how to treat data subjects ethically.





DATA AND DATA QUALITY

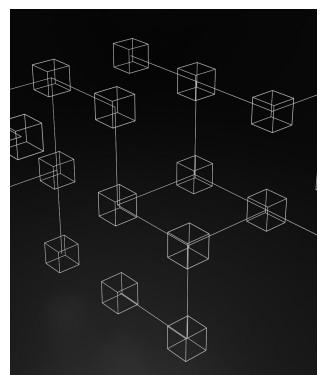
DATA AND DQ GOALS

Our major goal in this module is to clarify how data and Data Quality work. We'll cover:

- The 4 types of data and the importance of each;
- The types of DQ problems and their financial impact (lost revenue, operational costs, and/or regulatory fines);
- The DQ improvement cycle (identifying issues, tying them to costs, data profiling, issue remediation and prevention);
- The different DQ dimensions (accuracy, completeness, timeliness, lineage and other relevant ones);
- The types of DQ tools (data profiling, parsing and standardising, merging and linking, data enhancement);
- The basic elements of a DG/DM business case, and what is usually included;

DATA AND DATA QUALITY

This module is all about **how data and Data Quality work**, as well as how to assess and improve the latter:



THE 4 TYPES OF DATA

Transactional data, master data, reference data and metadata, and what they all mean

DQ PROBLEMS AND IMPACT

The common types of DQ problems, as well as the business impact they have



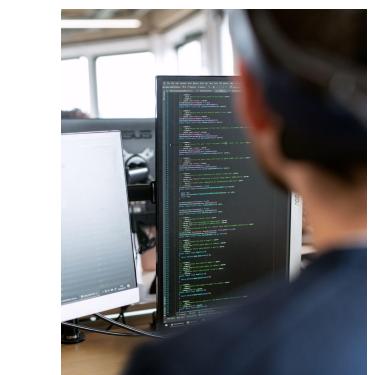
DATA QUALITY MANAGEMENT

What DQ management encompasses, including processes, data dimensions, actions



DQ TOOLS AND TECHNIQUES

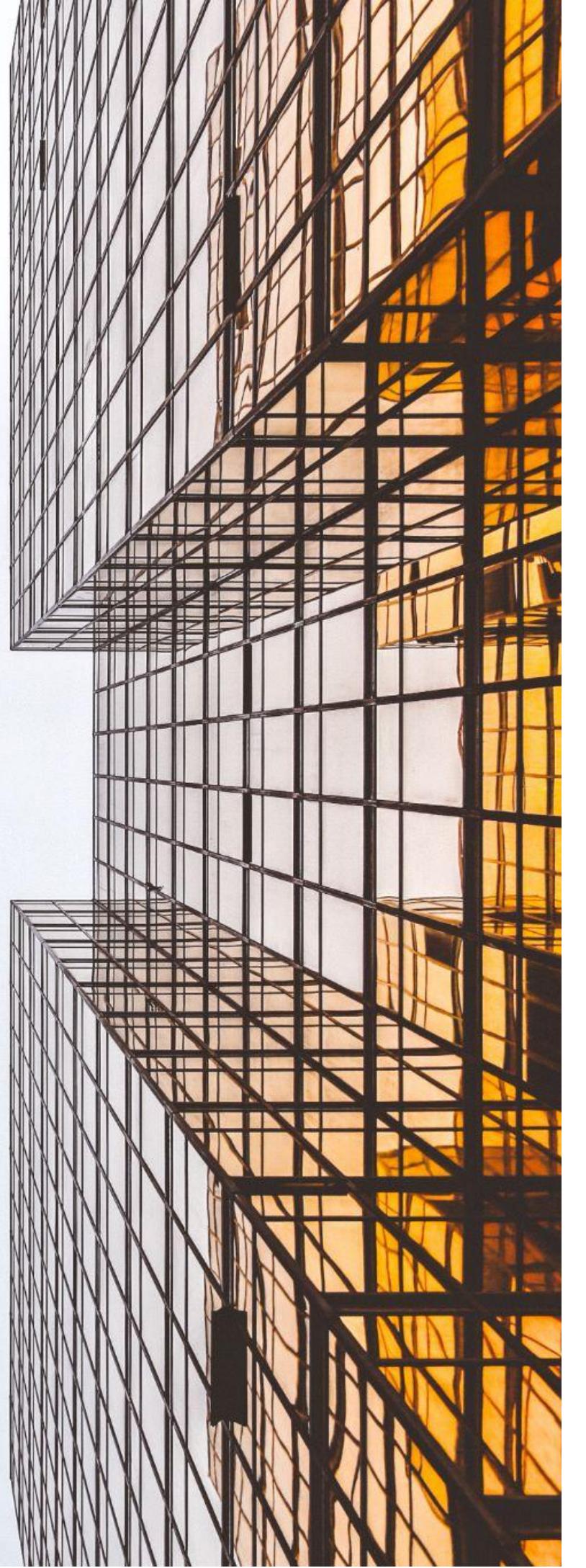
The specific tools used for profiling, parsing, standardising, merging, and more



BUSINESS CASE BUILDING

The usual elements included in a business case to show value for either DG or DM





DATA AND DATA QUALITY

DATA AND DATA QUALITY THE 4 TYPES OF DATA

Data come in all shapes and sizes. However, different types of data have different nuances, so it's important to clarify both what types of data exist, as well as the purpose of each of them. There are 4 main types of data in the “usual” classification:

- Transactional Data. Sales, orders, receipts, etc. Usually long lists, used as the basis for analytics and reports.
- Master Data. The most important concepts in the organisation. Customer, Product, Location, etc. Central;
- Reference Data. Codes or indicators used by all other data. Lists of countries, phone prefixes, hierarchies, etc;
- Metadata. “Data about data”, but much more than that. Information on data types, business uses, and a lot more;

DATA AND DATA QUALITY

THE 4 TYPES OF DATA

Let's start with Transactional Data. As the name indicates, these data are usually transactions of some type, and they are intimately tied to how that specific business is run.

- They are usually in high volume, and meaning is derived from aggregating multiple transactions;
- Can be not only sales, invoices, or receipts, but other elements that make sense to the specific business, such as consultations (healthcare), account opening and closures (banks), or others;
- They are usually useful not by themselves, but when combined to generate reports/dashboards/other data products (e.g., a dashboard of sales by region by month);





DATA AND DATA QUALITY

DATA AND DATA QUALITY THE 4 TYPES OF DATA

Master Data are the data that represent, in specific, essential entities to a business. The most frequent types are Customer and Sale, for example.

- You may think that all data are important in an organisation, so why is there a need to distinguish “Master Data” in specific? The reason is that, as a general rule, there must be only one copy of MD, usually a central one;
 - For less important data, it’s not a problem to have different data sources with different definitions or values. But for “key” information, such as Customer or Sale, this can be a big problem;
- In fact, there are entire programs dedicated to managing Master Data (MDM or Master Data Management);

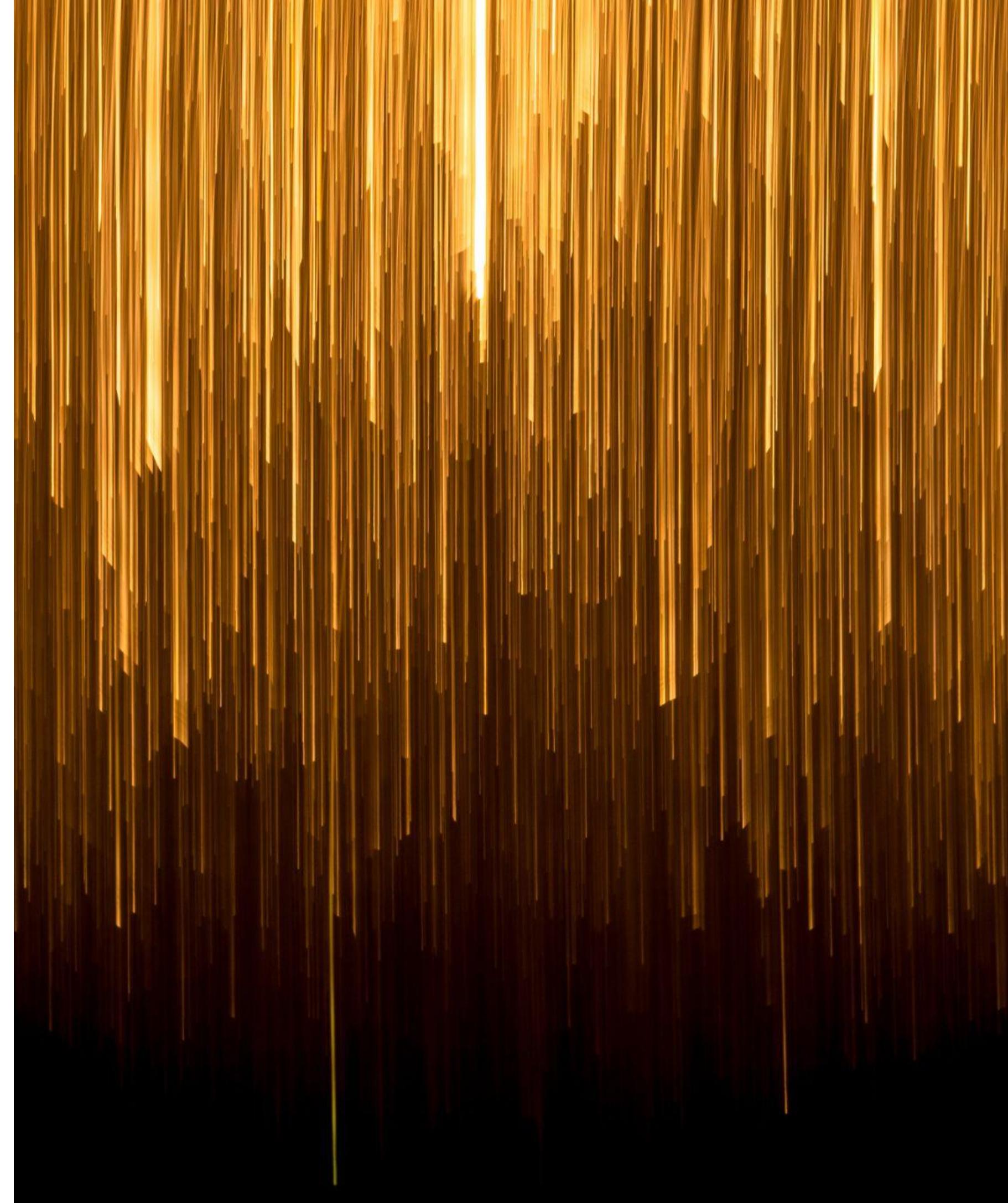
DATA AND DATA QUALITY

THE 4 TYPES OF DATA

Reference Data are data that, as the name says, are referenced by an organisation. These are very important due to the pervasiveness of their use - if one single value is wrong, and 71 applications reference it, you immediately cause 71 errors.

- They include references and codes for all areas of the organisation. Currency conversion tables, zip code tables, phone prefix lists, product hierarchies, and more;
- As a general rule, there is no “right or wrong” way to define the possible values - it’s just important these are consistent in the organisation. For example, you can list the 50 US states by abbreviation (California as “CA”) or by number (California as “5”). Just always use the same system;





DATA AND DATA QUALITY

DATA AND DATA QUALITY THE 4 TYPES OF DATA

Finally Metadata are the “data about data”. They give meaning and context to data - especially when the data themselves are missing important information.

- They're at the center of DG and DS. The “metadata hub” or “metadata repository” of an organisation contains all metadata, and data can only be properly governed if we have information about them (it may even be a prerequisite to “trust” a data source);

It's usually split between business and technical:

- Business metadata describes data uses, sensitivity classification (PII, confidential, etc), allowed users/teams...
- Technical metadata contains specifications. Is this an integer, string, etc, which length, DB layout, and so on;

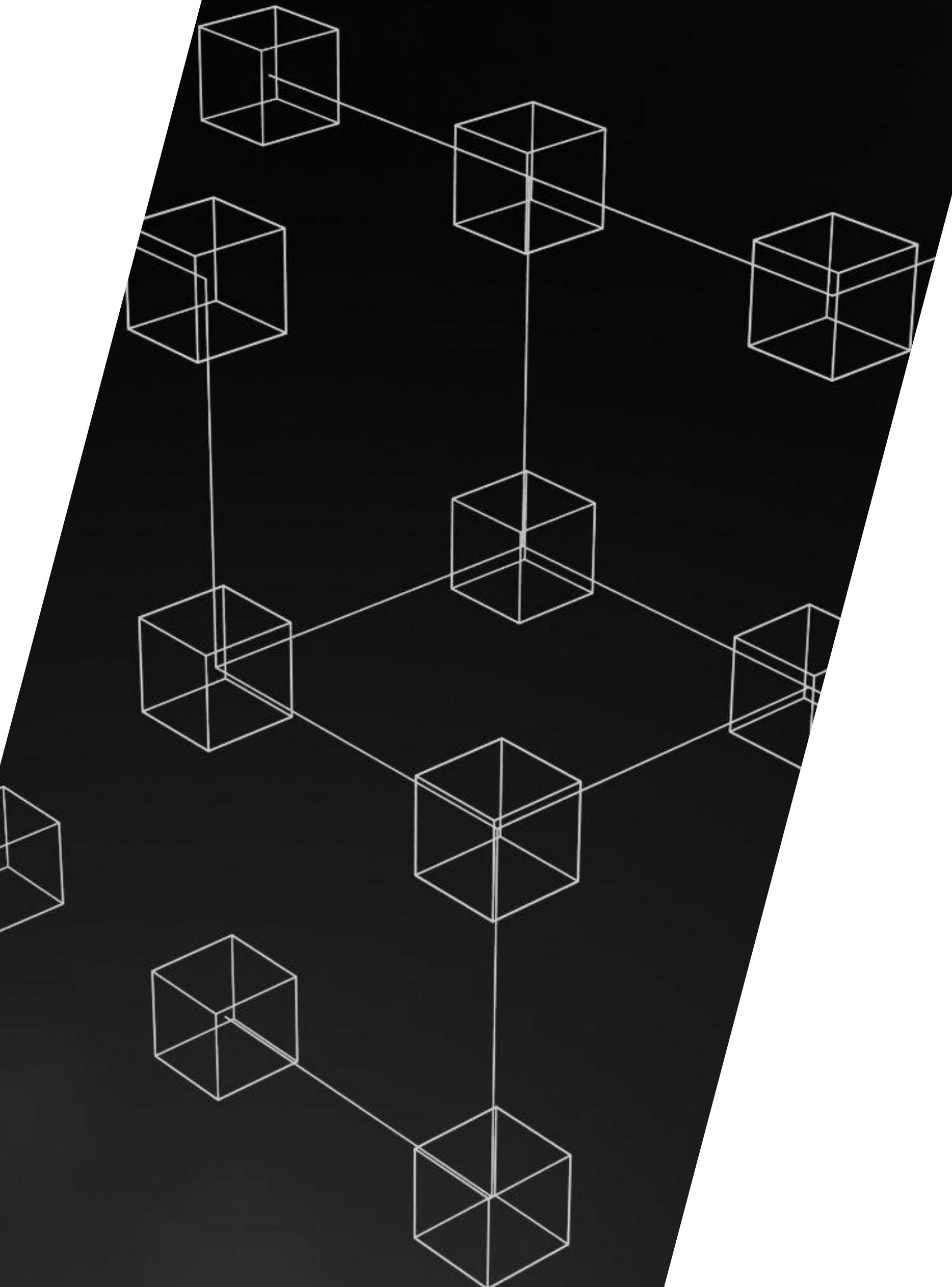
DATA AND DATA QUALITY

THE 4 TYPES OF DATA

Metadata, in specific, have a close connection to DQ. They store the information about the data formats, usual values, and expectations for any given field, DB table, full DB, etc.

They usually cover 3 elements important to DQ efforts:

- Reference data domains. Which reference data are used, from where, and what are the specific values:
 - (e.g., the “State” field must be a US state, in the State_Table table, in the format of an Integer of “1” - “50”)
- Allowed domain values. The actual values the field takes:
 - (e.g. “First Name” must be a String, max. 128 chars);
- Common value patterns. The usual patterns found:
 - (e.g. “Gender” is usually “Male”, “Female” or “Custom”);





DATA AND DATA QUALITY

DATA AND DATA QUALITY THE 4 TYPES OF DATA

It's important to note that we can have DQ problems in all types of data. Usually, Master Data problems are the most serious. Examples include:

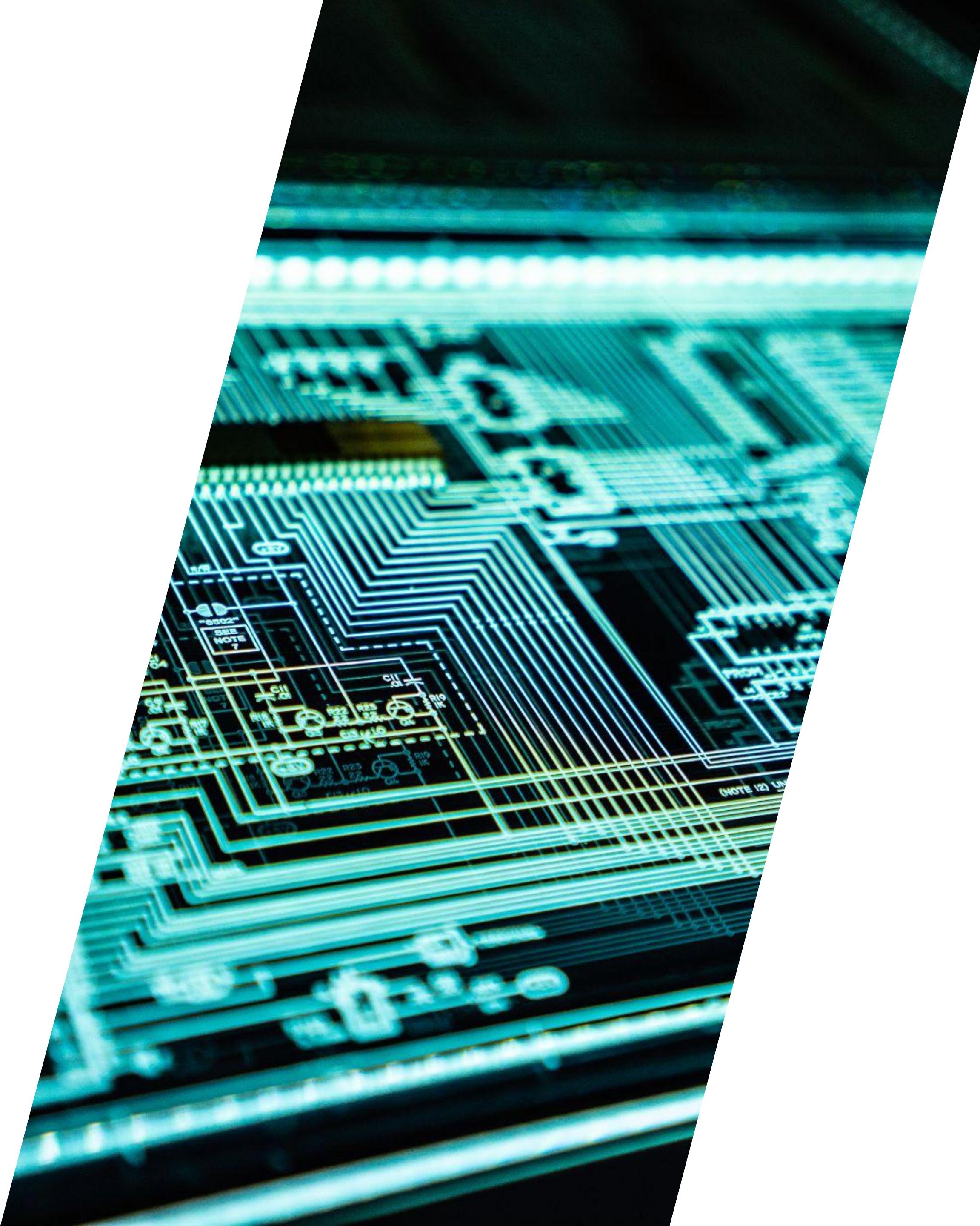
- Transactional Data. You may have a list of 70 purchases which are missing the currency, or the date of the purchase;
- Master Data. You may have a Customer with the wrong address information, affecting all communication w/ them;
- Reference Data. You may have one product code wrong in your taxonomy, affecting catalogs, transactions, website...
- Metadata. You have no information about which Customer data are PII, which must be stored with extra precaution, or which users or teams can use them;

DATA AND DATA QUALITY

THE 4 TYPES OF DATA

Despite the categorisation of data in these different types, it's important to clarify that these are interrelated and have an effect on each other.

- For example, all data depends on Reference Data in some way. If a currency is wrong, then all transactions are wrong, and key entities such as Contract may have wrong data;
- Transactional Data in specific usually depends on reference and master data. For example, to record a transaction of \$130 from Company A correctly, you need correct reference information about the "\$" currency, and correct information about the "Company A" Customer instance;
- Additionally, all data have their quality improved by Metadata, explaining them, and all contribute to them;



DATA AND DATA QUALITY

THE 4 TYPES OF DATA

EXAMPLES

/01 **MASTER DATA AS FLEXIBLE**

While master data must cover the “key” entities, there are some “borderline” ones which may be or not. For example, Contract may be transactional or master.

/02 **CLASSIFICATION IS CRUCIAL**

Classifying data (as we'll see later), especially in terms of sensitivity, is crucial to define privacy and security controls, including access control and permissions.

/03 **DQ <> DS FEEDBACK**

Data Stewards have an intimate relationship with Data Quality operations. Their metadata help profiling and remediation, which then refine the metadata.

DATA AND DATA QUALITY

THE 4 TYPES OF DATA

KEY TAKEAWAYS

/01 4 MAIN TYPES

There are usually four “formal” types of data that are involved in DM and DG: transactional data, master data, reference data and metadata.

/02 TRANSACTIONAL DATA

Transactions as part of the business. May be sales, receipts, invoices, appointments, or others. Usually in large quantity, and derive insights when aggregated.

/03 MASTER DATA

These represent the most important entities for an organisation. Customers, Sales, Locations, Employees. Usually managed to be only in one, central source.

/04 REFERENCE DATA

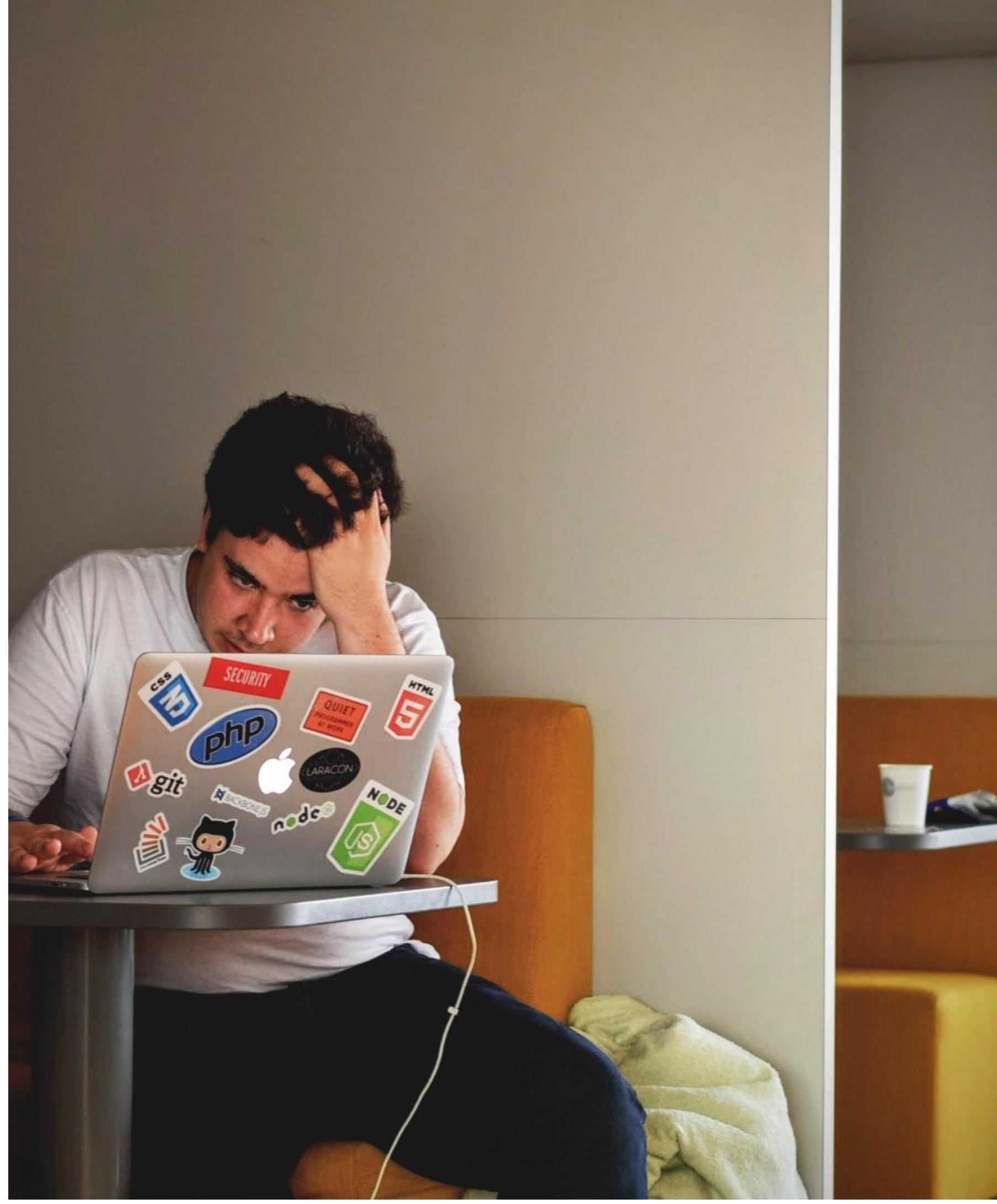
Codes, indicators and mappings that are usually referenced by all other types of data. May be lists of currencies, locations, phone numbers, or any others.

/05 METADATA

“Data about data”, but much more than that. They give context and meaning to data. Usually technical (formats and tech) and business (users and meaning).

/06 METADATA x DQ

Metadata are closely tied to DQ, and cover 3 key elements: which data are referenced, what values are actually allowed in this field, and the usual patterns.



DATA AND DATA QUALITY

DATA AND DATA QUALITY DQ PROBLEMS AND IMPACT

Lack of Data Quality (DQ) can cause multiple problems in multiple departments or business lines. Examples include:

- A new company has been bought, and there is no idea on how to integrate the new data with the current one;
- There are multiple sources of customer information (“Master Data”), which are conflicting;
- Bad data causes a bank to extend loans to the wrong people, or to severely undervalue the risks of such loans;
- The organisation obtains data from external sources, such as vendors, but they’re inconsistent, not trusted by people;
- There are BI/analytics reports which generate no insights;
- There are “big data”, but unclassified/not labeled (unused);
- Lack of data causes a breach of compliance, causing fines;

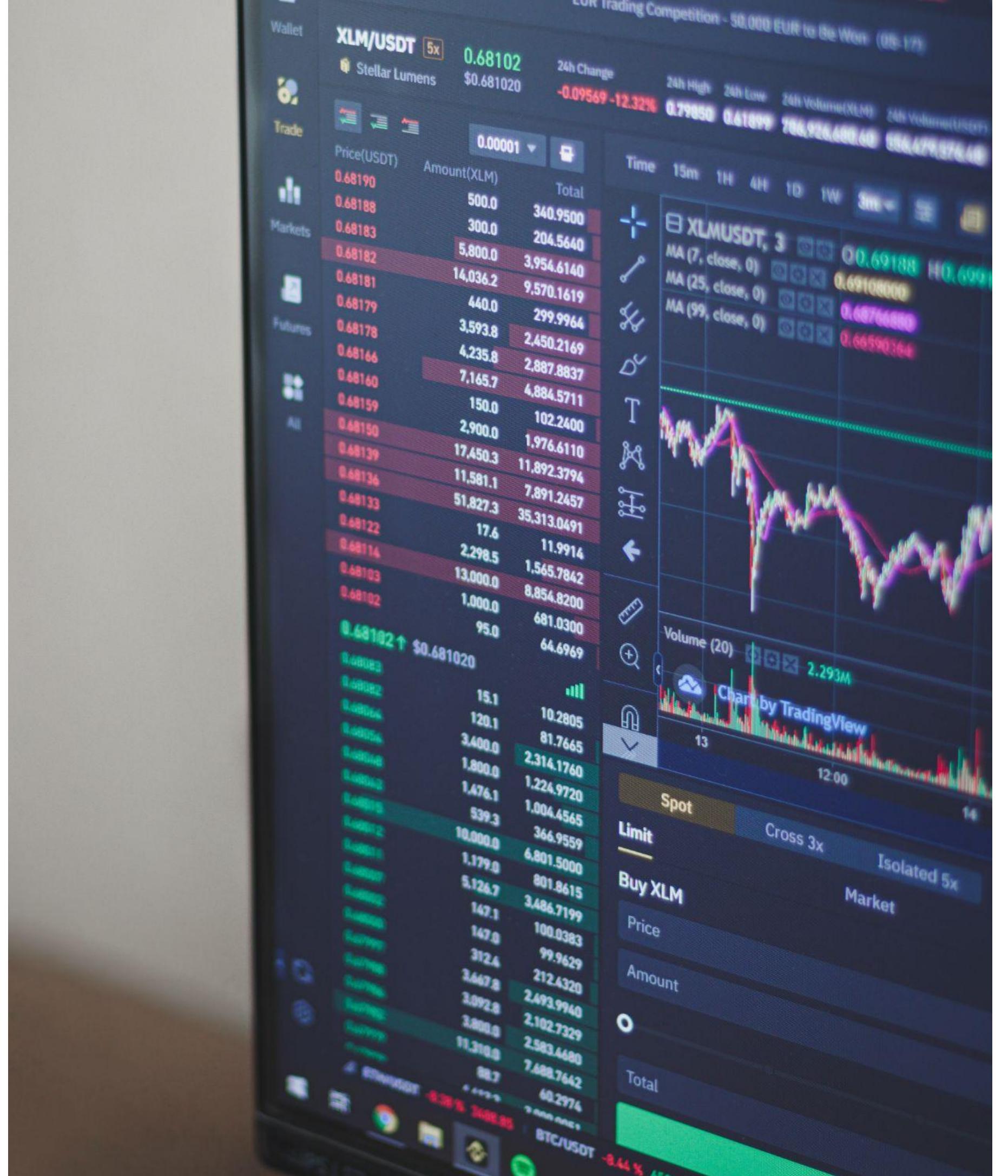
DATA AND DATA QUALITY

DQ PROBLEMS AND IMPACT

Usually, DQ problems cause three main types of impact to the organisation (when building a business case for a DM or DQ initiative, the impact must be estimated):

- Financial Impact
 - Actual financial costs from missing/incorrect data;
 - Loss of revenue, cash flow, customers, incurring fines;
- Operational Impact
 - Degradation of operations due to missing/incorrect data;
 - Project delays, app delays, reconciliations, delivery costs;
- Compliance or Risk Impact
 - Costs due to lack of compliance/risk underestimation;
 - Sanctions due to breach of AML/CFT compliance, privacy compliance, investment risk underestimation, etc;





DATA AND DATA QUALITY

DATA AND DATA QUALITY DQ PROBLEMS AND IMPACT

Usually, besides estimating the impact to the company in one of the three areas mentioned, it's crucial to link that impact to DQ issues, in order to justify taking action (usually, in the form of starting a DM or DG project). Examples:

- Linking lack of customer satisfaction data to financial impact due to losing unsatisfied customers;
- Linking lack of AML/CFT data input controls to fines and sanctions due to not having report data;
- Linking lack of project budget and execution data to financial costs from project scope creep/going over budget;

Although other elements are necessary (management buy-in, company culture, etc), linking DQ issues to costs is the start.

DATA AND DATA QUALITY

DQ PROBLEMS AND IMPACT

EXAMPLES

/01 COMPLIANCE IS FREQUENT

Especially for highly regulated industries, fines and other costs from lack of compliance can surpass all other types of costs, and become highly relevant.

/02 DG GOES WELL WITH DM

While data management projects are usually motivated by actual fixes to data, DG projects usually need a use case to “attach to”, and DM ones are good.

/03 ASSETS AND MONETISATION

The principles of data as assets and monetising data come in here. If you don't measure data with diligence and accuracy, you can't believe they can be of value.

DATA AND DATA QUALITY

DQ PROBLEMS AND IMPACT

KEY TAKEAWAYS

/01 LOW QUALITY DATA

DQ problems are problems caused by low quality data in all dimensions. It may be missing customer data, inconsistent sales data, inaccurate location data, etc.

/02 3 KEY TYPES OF IMPACT

DQ issues usually either cause actual financial impact, operational impact, or risk/compliance impact. All result in costs for the company.

/03 LINKING BOTH IS CRUCIAL

To build the business case for any DG or DM project, linking DQ problems to actual financial costs is important, as that's the first step to obtaining support.



DQ MANAGEMENT

DQ MANAGEMENT INTRODUCTION

Managing data quality is a simple (but not necessarily easy!) process, which usually starts by assessing (or “profiling”) data based on specific criteria (the “dimension”). Then, multiple actions can be taken.

It’s important to clarify the usual processes that DQ management can involve, including profiling the data, remediating them, preventing new errors, as well as what the sequence of steps is, and the dimensions involved.

DQ MANAGEMENT

We will cover **four key topics** in terms of managing Data Quality:



DQ IMPROVEMENT

The full, typical process from data profiling to remediation and prevention of further problems



DATA QUALITY ACTIONS

The three main types of actions facing DQ problems, and what each entails



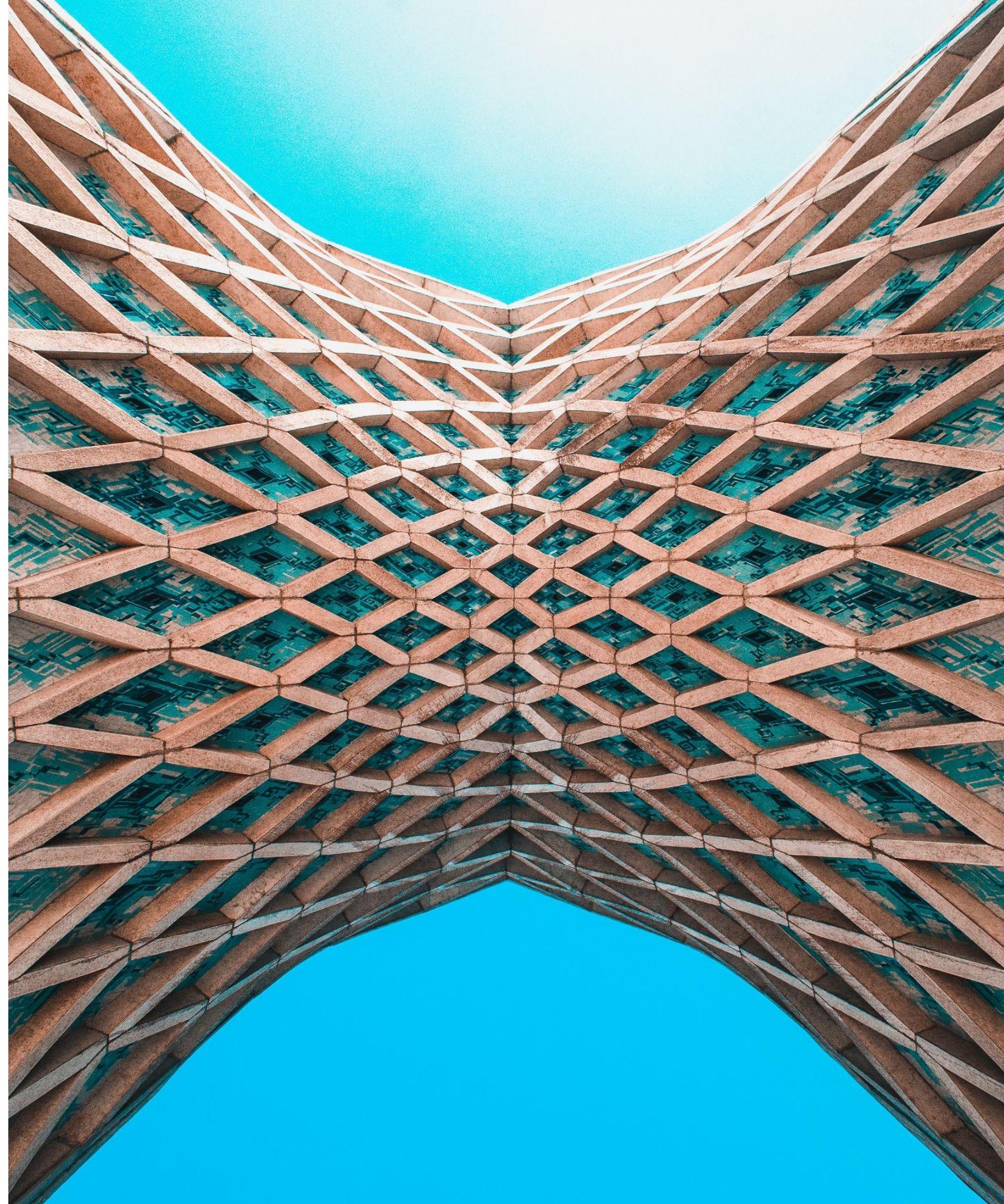
DATA DIMENSIONS

The different dimensions across which there may be data problems, and their nuances



BIG DATA AND A.I.

The consequences of both big data and AI to the data management process



DQ MANAGEMENT

DQ MANAGEMENT DQ IMPROVEMENT

Regardless of whether we are in the Data Management (DM) or Data Governance (DG) discipline, it usually comes down to improving Data Quality (DQ).

- Both DM and DG programs stem from DQ problems, and both seek to address it - hands-on/hands-off, respectively;

The process of improving DQ usually follows these steps:

- Identifying DQ Problems. What is the problem?
- Correlation with Business Impact. What is it costing us?
- Triage and Prioritisation. Which problems do we tackle?
- Profiling/Remediation. How can we fix this for now?
- Defining Validity Rules. What are our future expectations?
- Enforcement of Rules. How we can make data obey these?

DQ MANAGEMENT

DQ IMPROVEMENT

We always start by identifying DQ problems. That is, if sales reports are low-quality, there must be an issue there. If we obtain bad insights from our marketing BI tool, there may be a problem there.

- Usually, a DM team starts a project to perform data profiling, to find out exactly what is the problem, in which data. It also usually involves selecting the dimension;

After the problem is identified, we can't just create a DM project to fix it (or a DG project to create policies) without being able to justify the business value of it.

- Bad data must be tied to losses, and good data to returns. For bigger initiatives, we must create a business case;
- (E.g., “Bad marketing insights are costing us \$10M in sales”);





DQ MANAGEMENT

DQ MANAGEMENT DQ IMPROVEMENT

It's also likely that the team dealing with DQ problems has a backlog of issues (either detected by profiling, by the team, or reported by data consumers), so they must always prioritise.

- Our goal is always to solve the biggest problems with the lowest resources possible;

Triage can be performed, evaluating factors such as:

- Impact. How much is this costing the business?
- Required resources. How long will this take to fix, and by how many people?
- Probability of remediation. What are the chances we can really fix it?
- Probability of prevention. Same for preventing it again;

DQ MANAGEMENT

DQ IMPROVEMENT

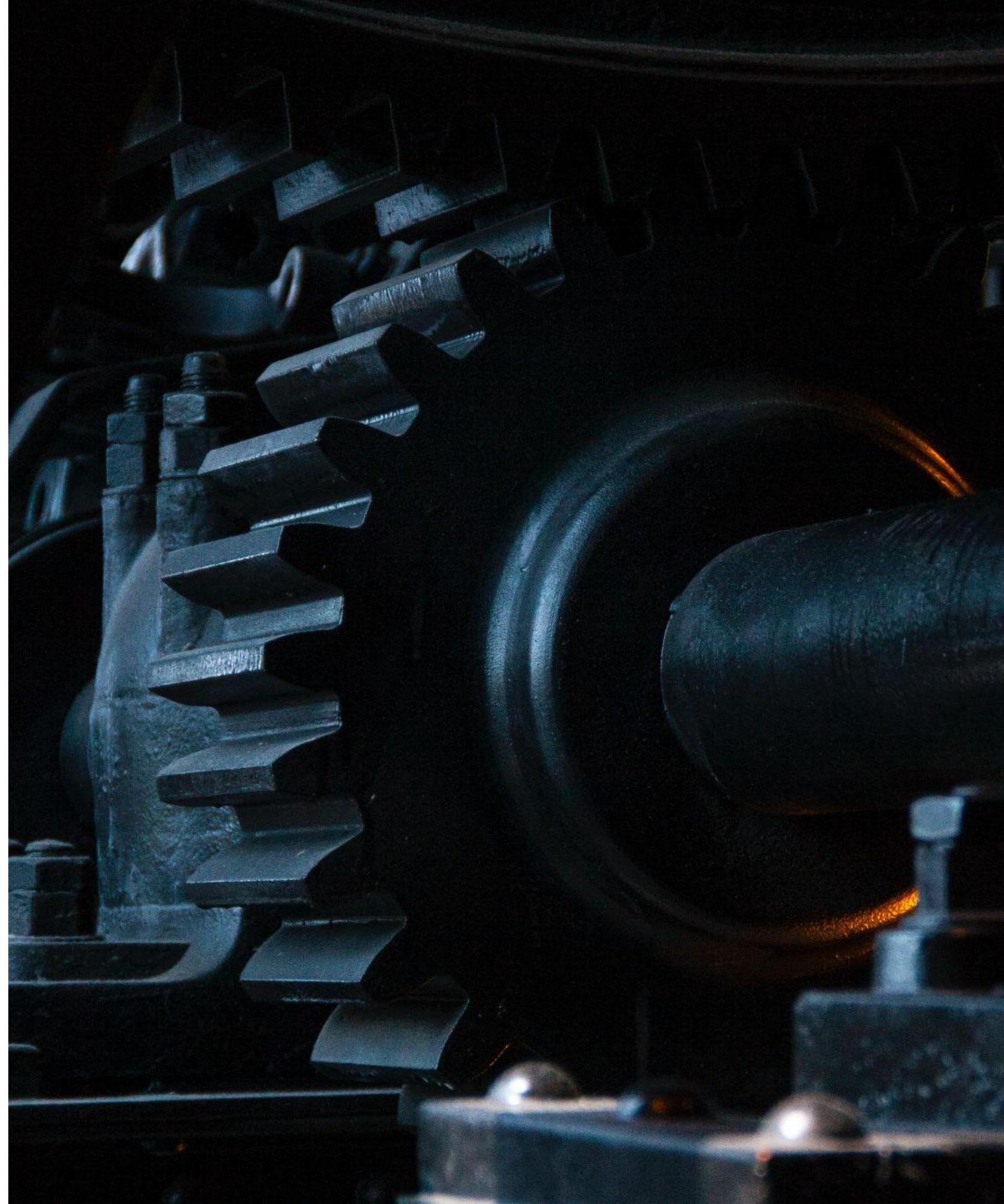
After we know what the problem is and the specific costs of it, we may take one of two paths (or both at the same time):

- DM performs remediation of the data. They go in and “fix” them in some way (“We added missing customer names to 140 transaction records”);
- DG defines processes and controls to manage the quality of the data. That is, to prevent them from becoming LQ (“Future data entered must have customer names”);

These usually converge in the creation of validity rules/expectations for the data. That is, it's not enough to correct the missing data, but we must prevent new data from having the same problem. This is usually a DG area:

- “Auto-validation will be performed on transaction creation”;





DQ MANAGEMENT

DQ MANAGEMENT DQ IMPROVEMENT

Finally, when the validity rules are set for HQ data (in other words, what formats, value ranges and patterns must the data obey), these are usually enforced either through automatic or manual mechanisms:

- Manual: “The Marketing Data Stewards will verify new transaction data every 3 days for inconsistencies and raise a Data Quality issue if one is found”;
- Automatic: “Our Sales application has been updated, requiring the mandatory filling of these fields in a specific format that is accepted”;

As a general rule, if DQ problems are “temporarily remedied”, but no future rules are enforced, they will return.

DQ MANAGEMENT DQ IMPROVEMENT

It's important to note that these steps can take different formats, and DM/DG may intersect (or not) with different intensity:

- For example, if an MDM (Master Data Management) program is already in place, and we suspect bad Customer Data, there is probably already a DM team that can perform the profiling, and can remedy the data. There may also be a DG function that can enforce new rules for data;
- Another example is that we may have a situation which was “addressed”, but coming back in new ways. Maybe we fixed the sales data by forcing salespeople to enter transactions with all mandatory fields, but the company created a new web interface that doesn't require it;



DQ MANAGEMENT DQ IMPROVEMENT EXAMPLES

/01 BEHAVIORAL CHANGE

In many DG and DM programs, enforcing controls requires behavioral change. Expect resistance when new demands are made of people entering/using data

/02 DATA STEWARDS ARE KEY

Data stewards are very important in making the bridge here. They are the ones that track the issues that DM resolves, as well as consolidating metadata.

/03 ROOT CAUSE MATTERS

There are cases where data are remediated late in the lifecycle, and all previous stages still contain the error. Performing root cause analysis, if possible, works well.

DQ MANAGEMENT DQ IMPROVEMENT KEY TAKEAWAYS

/01 IDENTIFY, FIX, MAKE A RULE

Fixing DQ problems may involve the DM and DG disciplines in different ways, but it's usually about identifying a problem, fixing it, and preventing it.

/03 DG SETS RULES/POLICIES

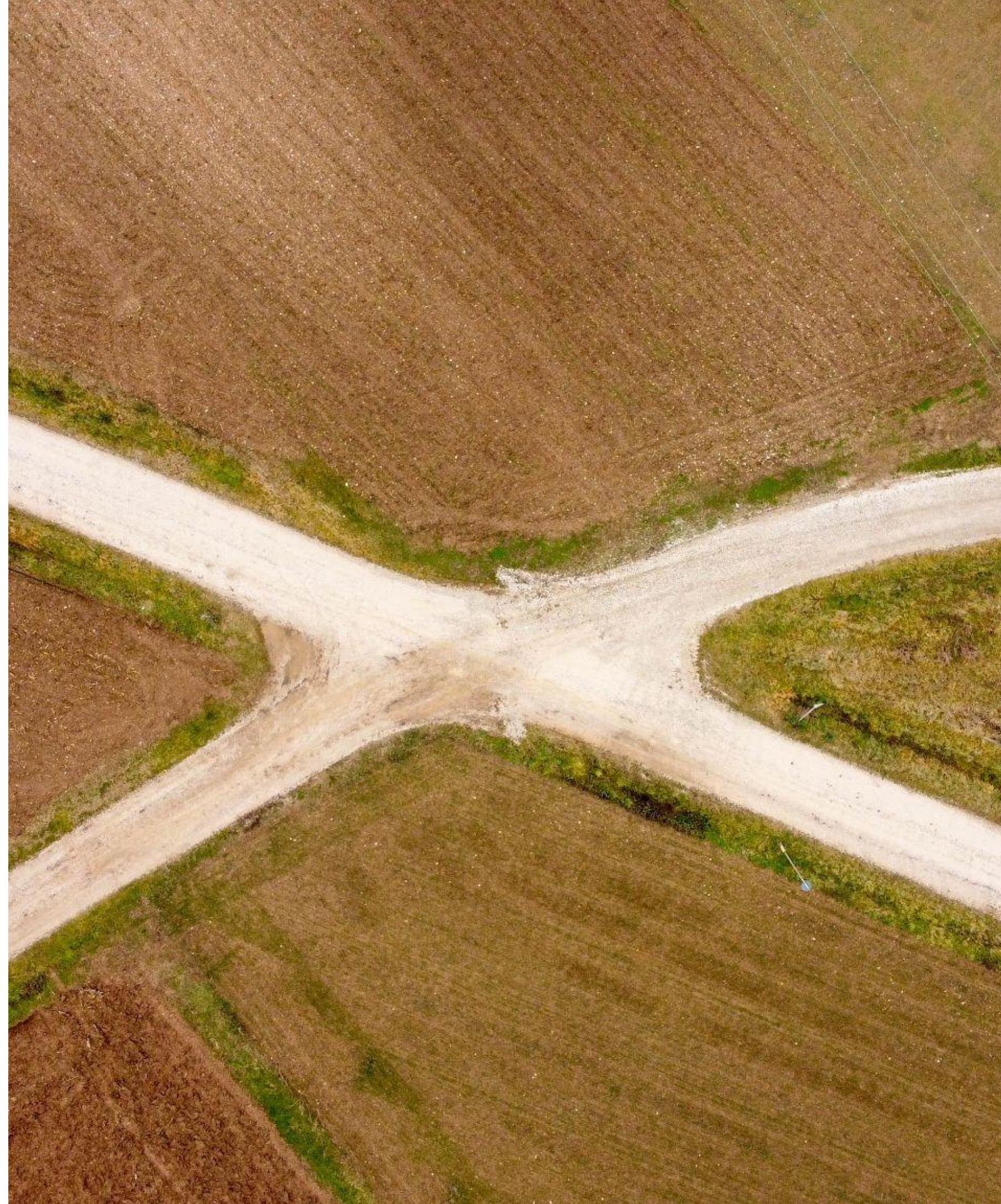
In this process, the DG function looks to create validity rules based on the insights found, as well as enforce controls/policies to ensure that validation is made.

/02 DM PROFILES/REMEDIATES

In this process, the DM function usually profiles the data (identifying specific data problems) and remediates the data (increasing their quality).

/04 THE BUSINESS CASE

Where both DM and DG intersect is that, regardless of which project/program you are trying to authorise, you must show the cost of LQ data. That's a business case.



DQ MANAGEMENT

DQ MANAGEMENT DATA QUALITY ACTIONS

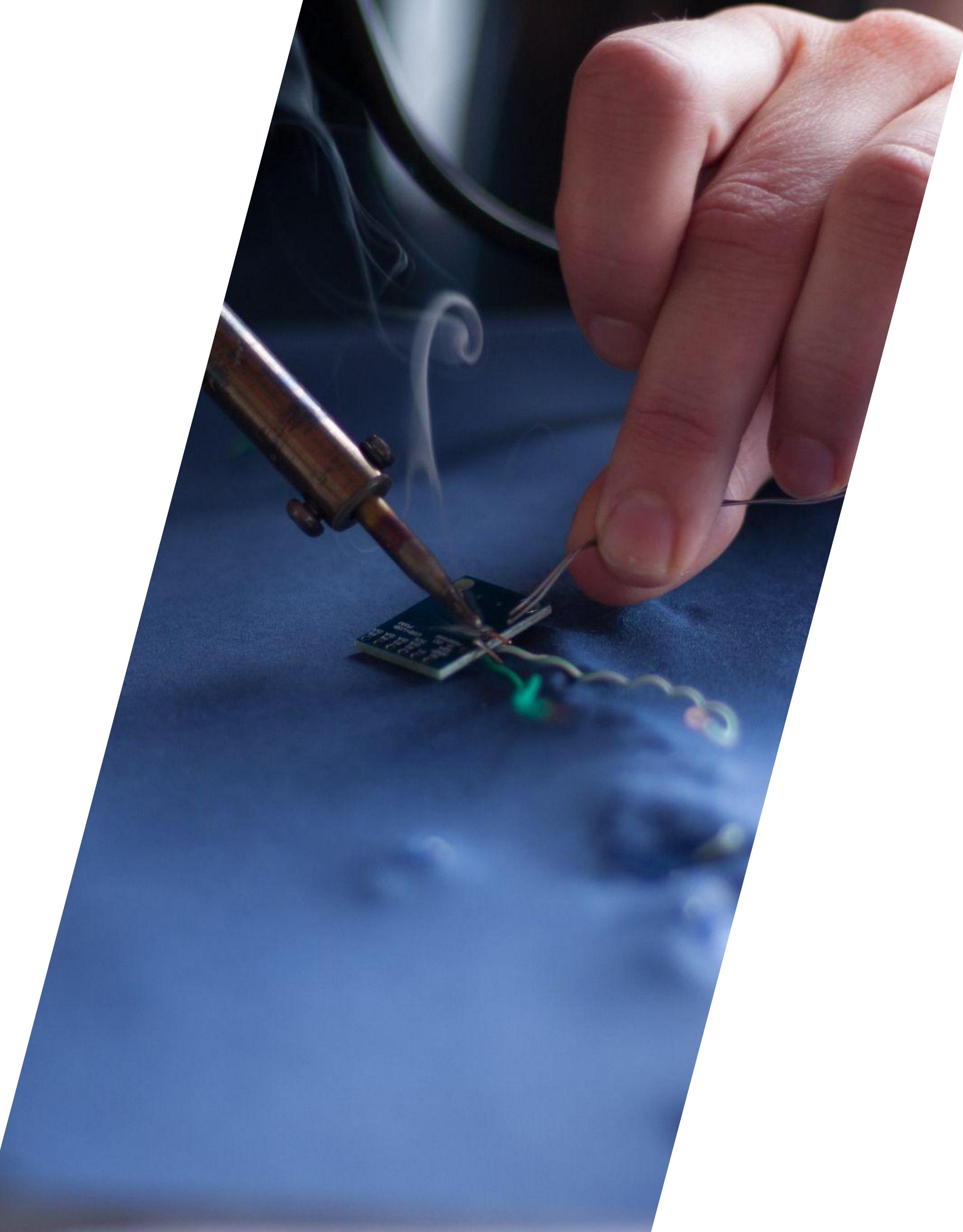
When faced with a DQ problem, it's important to notice that there are usually 3 distinct types of actions that can be taken:

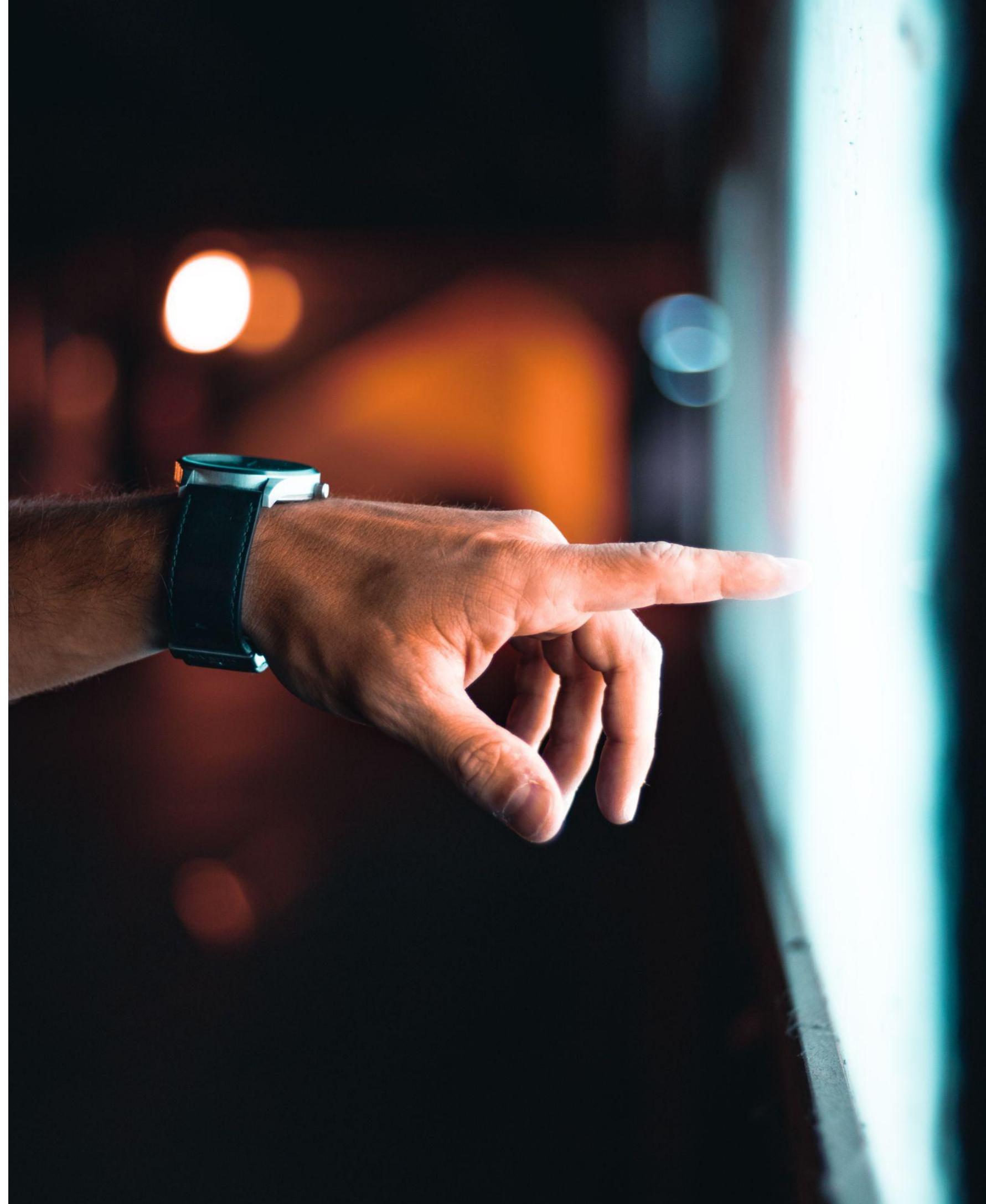
- Actual correction and remediation of the data. That is, taking “wrong” data and increasing quality in some way;
- Performing root cause analysis. That is, the data problem found may already be in a downstream stage of the data lifecycle. Fixing it here won’t fix it in the earlier stages, nor will it prevent the problem from occurring again;
- Instituting prevention and monitoring. Even if the problem is addressed, and even if it’s addressed at the root cause stage, it doesn’t mean it can’t happen again. Monitoring allows you to automatically monitor DQ, and raise alerts if a problem occurs, preventing it (or addressing it early on);

DQ MANAGEMENT DATA QUALITY ACTIONS

Data correction and remediation simply address a specific DQ problem (maybe data are missing, maybe they are not within range, etc). But when it's performed, it's important to understand the consequences, both upstream and downstream:

- Are there data at an earlier stage that contain the same error, and that need fixing just like the dataset addressed?
 - For example, “fixing” the data in a report, but leaving the original data untouched, which will create more errors;
- Are there data at later stages that are affected by this remediation?
 - Transformations may depend on data as is, and changing format “crashes” them (even if fixing data!);





DQ MANAGEMENT

DQ MANAGEMENT DATA QUALITY ACTIONS

A concern that emerges is determining the source of an error.

Remember, DQ errors can occur at any stage of the information lifecycle:

- Data are created with errors (e.g., typos);
- Data are transformed with errors (e.g., ETL processes);
- Data are edited with errors (e.g., wrong information);

Performing root cause analysis becomes an important process. This is nothing more than determining the earliest stage at which an error is present.

- It always necessitates some kind of data map or flow (where they are created, where do they move, how they are transformed, etc);

DQ MANAGEMENT DATA QUALITY ACTIONS

Armed with a quality data flow document, root cause analysis becomes simple (but not necessarily easy), and the source of data problems can be identified by a simple comparison process.

- For example, let's say Customer data has only three stages: Creation (form entry), Transformation (ETL to store in a DW) and Usage (Customer Reports). We simply analyze it for errors at all stages (such as a wrong customer address);
 - We check the Creation stage. Addresses are OK;
 - We check the Transformation stage. Addresses are not OK. An error was introduced in this stage;
 - We check the Usage stage. Addresses are still not OK in the report;





DQ MANAGEMENT

DQ MANAGEMENT DATA QUALITY ACTIONS

Naturally, it's important to consider that certain data flows have dozens of steps, and they may be used by thousands of users concurrently.

- Verifying DQ at every single step may not be viable in all situations;
- Additionally, the question of whether to check the whole dataset or just a sample becomes relevant with gigantic datasets that are computationally expensive;

There is no right answer, and each organisation can afford different intensities, but they can compromise and check a number of relevant information stages, as well as using relevant sample sizes that are doable, but still significant.

DQ MANAGEMENT DATA QUALITY ACTIONS

Finally, monitoring and prevention intersect with DG and metadata. In short, you will come to the conclusion that DQ problems occur because data don't obey the rules or requirements for them.

- For example, “Each Customer must have one Address”. All Customer records with no address are a DQ problem;
- Through many sources (data profiling tools, data steward metadata entry), you’ll determine these data validity rules;

Monitoring is nothing more than automating the testing of data against these rules. For example, testing every entered Customer record for an Address.

- Monitoring mechanisms can prevent the entering of incomplete data, or at least alert you when it occurs;



DQ MANAGEMENT DATA QUALITY ACTIONS EXAMPLES

/01 AC AND PERMISSIONS

Since data managers/analysts will be profiling different types of data for different issues, permissions and access control become very important here.

/02 BIG DATA COMPLICATE

The volume that big data bring makes early remediation a lot harder, if not impossible. In many cases, remediating after processing is reasonable.

/03 INTEGRATION W/ INFOSEC

The Data Governance prevention and monitoring tools and techniques can be easily integrated into a bigger suite of monitoring as part of Information Security.

DQ MANAGEMENT

DATA QUALITY ACTIONS

KEY TAKEAWAYS

/01 FIX, ANALYZE, PREVENT

There are usually 3 main types of actions when faced with a DQ problem. Fixing or remediating the data, analysing the root cause, and preventing future issues.

/03 CHECKING WHICH STAGES

Root cause analysis is performed by simply detecting a DQ error, and then checking for it at different stages of the data lifecycle, and finding the earliest one.

/02 MANY STAGES IMPACTED

Remember that remediating data at one stage does not mean it's fixed everywhere. There may be earlier errors, and the fixing may cause later errors.

/04 ALL ABOUT AUTOMATION

Finally, monitoring and prevention are about automated tests. Data must obey certain rules. By automating that verification, alerts can be raised.



DQ MANAGEMENT

DQ MANAGEMENT DATA DIMENSIONS

When we're talking about data quality, and about data being "high-quality" or "low-quality", we may mean different things. DQ can be measured according to different criteria, which are called the data dimensions. The most relevant ones include:

- Completeness (How many values are missing. The "usual" data dimension that comes to mind to most people);
- Accuracy (Do the values make sense? Are they within range?);
- Consistency (Throughout different sources, are data the same? Are they same in raw format and in data products?);
- Timeliness (Are these data updated? Are they frequently refreshed?);
- Lineage (Where are they from? What happened to them?);

DQ MANAGEMENT DATA DIMENSIONS

The Completeness dimension is easy to explain. In a list of values, you want to know how many are missing (or which have a specific field missing).

- For example, in a list of 100 customers, 15 of them missing the “Address” field is a completeness problem;
- Another example is, for 100 critical trials, only having documentation for 80 of them. Completeness issue as well;

The Accuracy dimension regards whether values present make sense or not, which can be subjective to measure.

- For example, for a “Zip Code” field that must be 7 digit numbers, out of 100 fields, 10 being in a different format;
- Or, for example, having a “Temperature” field in Fahrenheit where some values are below 0 or above 100;





DQ MANAGEMENT

DQ MANAGEMENT DATA DIMENSIONS

The Consistency dimension measures, for different sources that contain the same data, whether the data match or not.

- Even with MDM/other frameworks, there are always copies;
- For example, a list of employees in a central HR database, and a list of employees in a local training DB. Are they the same?
- Sales data in a report versus the original DB. Same values?

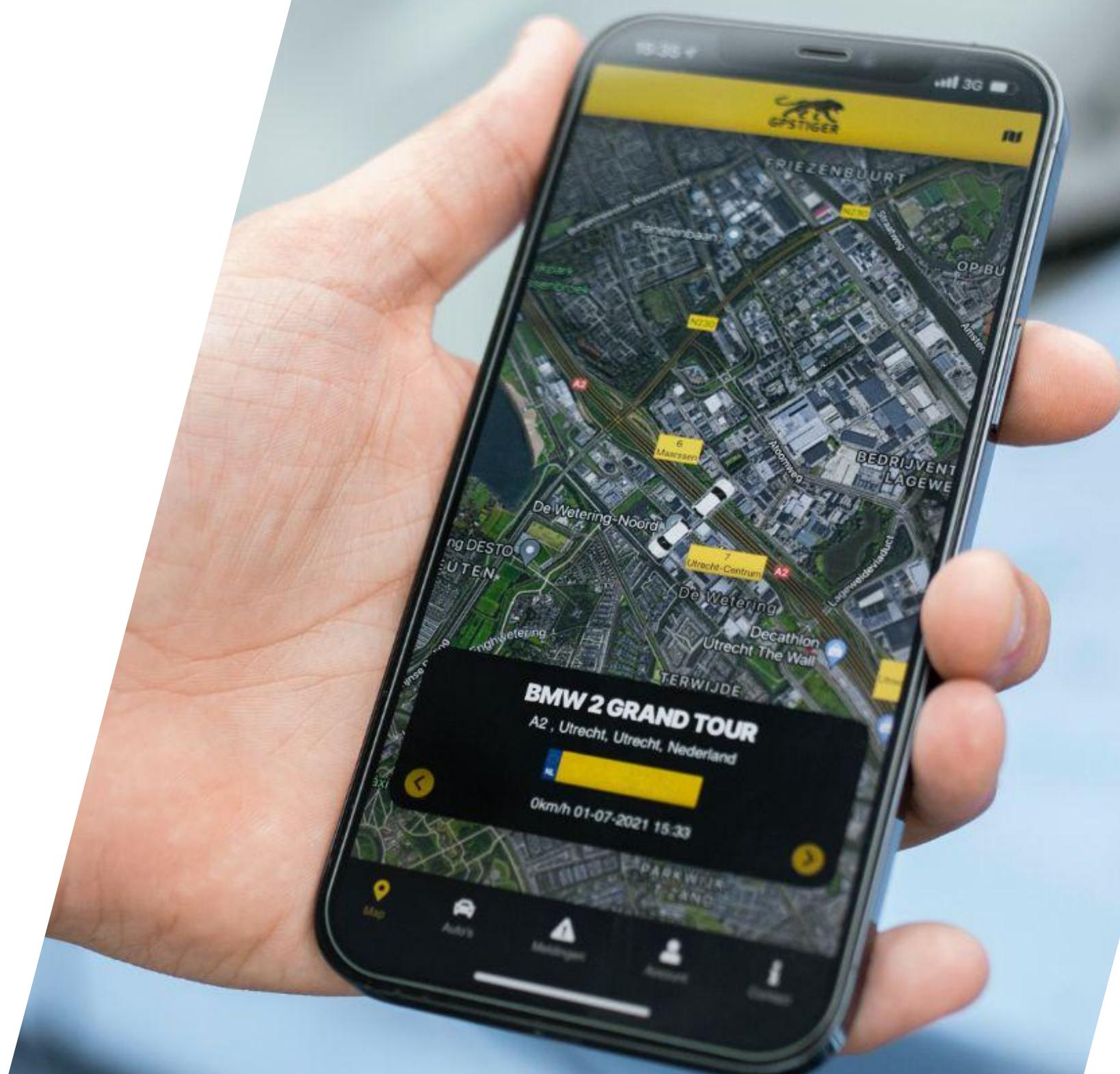
The Timeliness dimension measures how updated data are. This can be important in fast-changing environments, or when data have retention periods or are heavily regulated.

- For example, in a trade database, processing thousands of transactions per second, a transaction that was reverted 5 minutes ago not having been updated yet is a problem;

DQ MANAGEMENT DATA DIMENSIONS

The Lineage dimension cares about the full path of the data from origin to present. It includes their creation/capture, transformations applied, and people/systems involved.

- For example, if data were obtained from a third-party vendor, then cleaned up and loaded into an operational DB, then went through an ETL process and loaded in a DW, do we even have that information?
- In many cases, it's not just about knowing where data has been, but applying a "trustworthiness score" to the data sources (Is this vendor trustworthy? Is this source?);
- For regulated industries using model calculations, it may even be demanded that snapshots of calculations at the time, with all data involved, be saved and kept aside;





DQ MANAGEMENT

DQ MANAGEMENT DATA DIMENSIONS

It's important to be aware of these different data dimensions because projects that aim to improve DQ may be referring to different dimensions of data:

- Are our BI reports producing bad results because we don't have enough data (Completeness)? Is it because the values don't make sense (Accuracy)? Is it because we have old data (Timeliness)?
- Are we getting fines from regulators, as a bank, because our regulatory capital models use wrong numbers (Accuracy)? Is it because we didn't record the numbers used at the time (Lineage)?

Usually, projects to improve DQ may intervene at any of these.

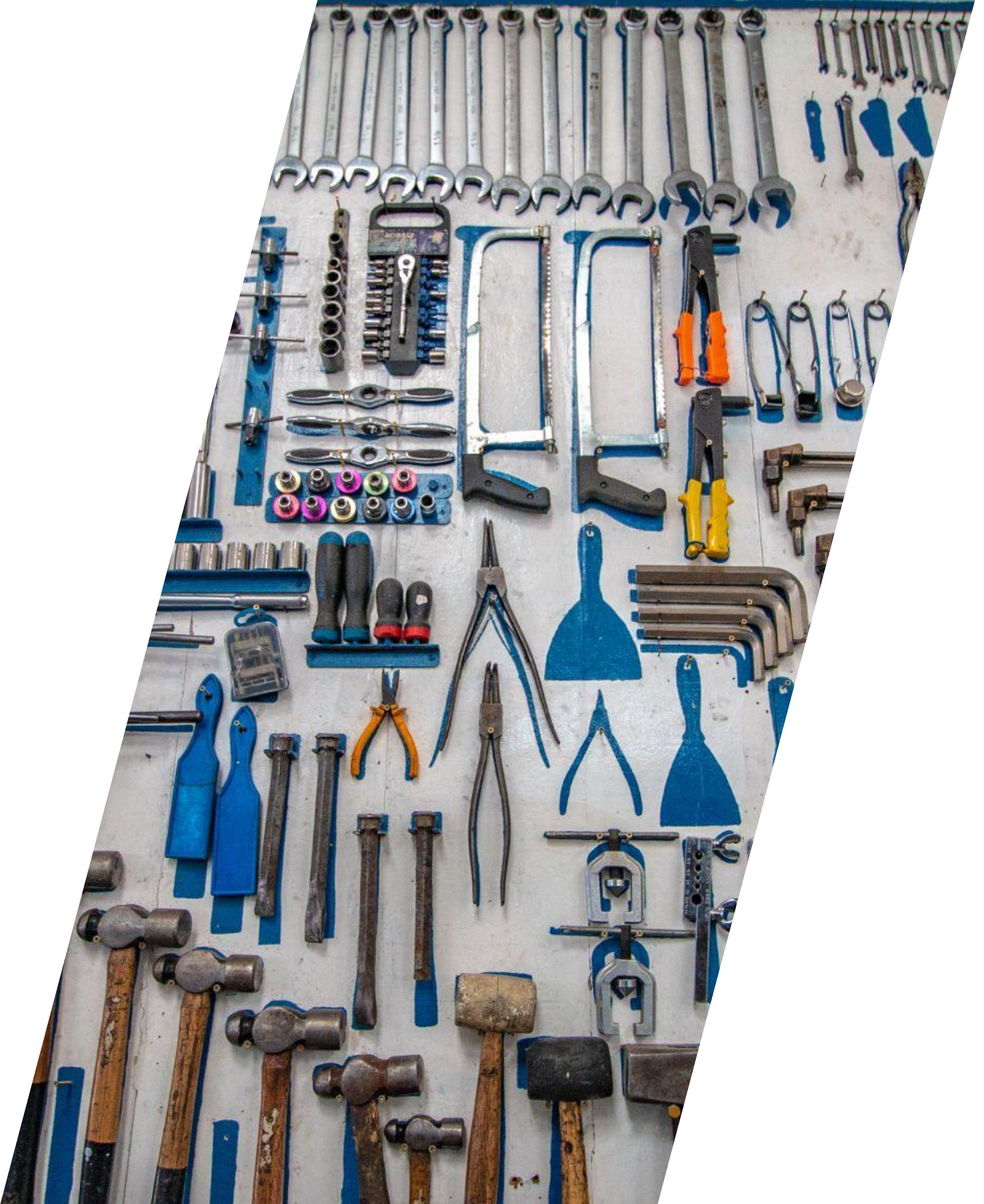
DQ MANAGEMENT DATA DIMENSIONS

It's important to note that, despite these being the most frequent dimensions, this is not a complete list by any means. Other dimensions may include:

- Identifiability and Accountability. Whether each datum has a specific user/group other associated with it or not;
- Presentability. Whether the data presented are easy to visualize or not, regardless of their values;
- Uniqueness. Whether the data are unique or duplicates exist within a specific database or storage format;
-

Also, depending on organisational needs, custom dimensions may be created, usually as combinations of existing ones.

- E.g. measuring data Completeness + Accuracy as one;





DQ MANAGEMENT

DQ MANAGEMENT DATA DIMENSIONS

What are some examples of DQ problems, in different dimensions, for a Customer entity, for example?

- Completeness. Multiple Customer rows are missing;
- Accuracy. Many rows have numbers in the “Name” field;
- Consistency. We have 3 sources of data for Customer, and all have different addresses for the same customers;
- Timeliness. All the Customer data is from years ago;
- Lineage. We have a Customer report, but we have no idea how the data were modified, or by whom. We may know it was modified 3 times, but no idea what happened in each;
- Uniqueness. We have several duplicate Customers;
- Identifiability. We do not know who in the organisation is the data owner of Customer, or who's been modifying it;

DQ MANAGEMENT DATA DIMENSIONS EXAMPLES

/01 CLINICAL TRIALS

Clinical trials are one area where healthcare companies must not only save the final results in terms of data, but in many cases intermediate steps.

/02 ORGANISATIONAL ISSUES

Consistency is an important dimension to make sure that all departments/LoBs in a company use the same data. If they get different results = consistency issue.

/03 ACCURACY

Accuracy may be the most subjective data dimension, because in many cases whether a value makes sense or not is entirely subjective. For example, outliers.

DQ MANAGEMENT

DATA DIMENSIONS

KEY TAKEAWAYS

/01 DATA QUALITY DIMENSIONS

DQ is not universal. Data may be of high quality according to one dimension, and of low quality according to another one. Issues are specific to these.

/02 COMPLETE / ACCURATE

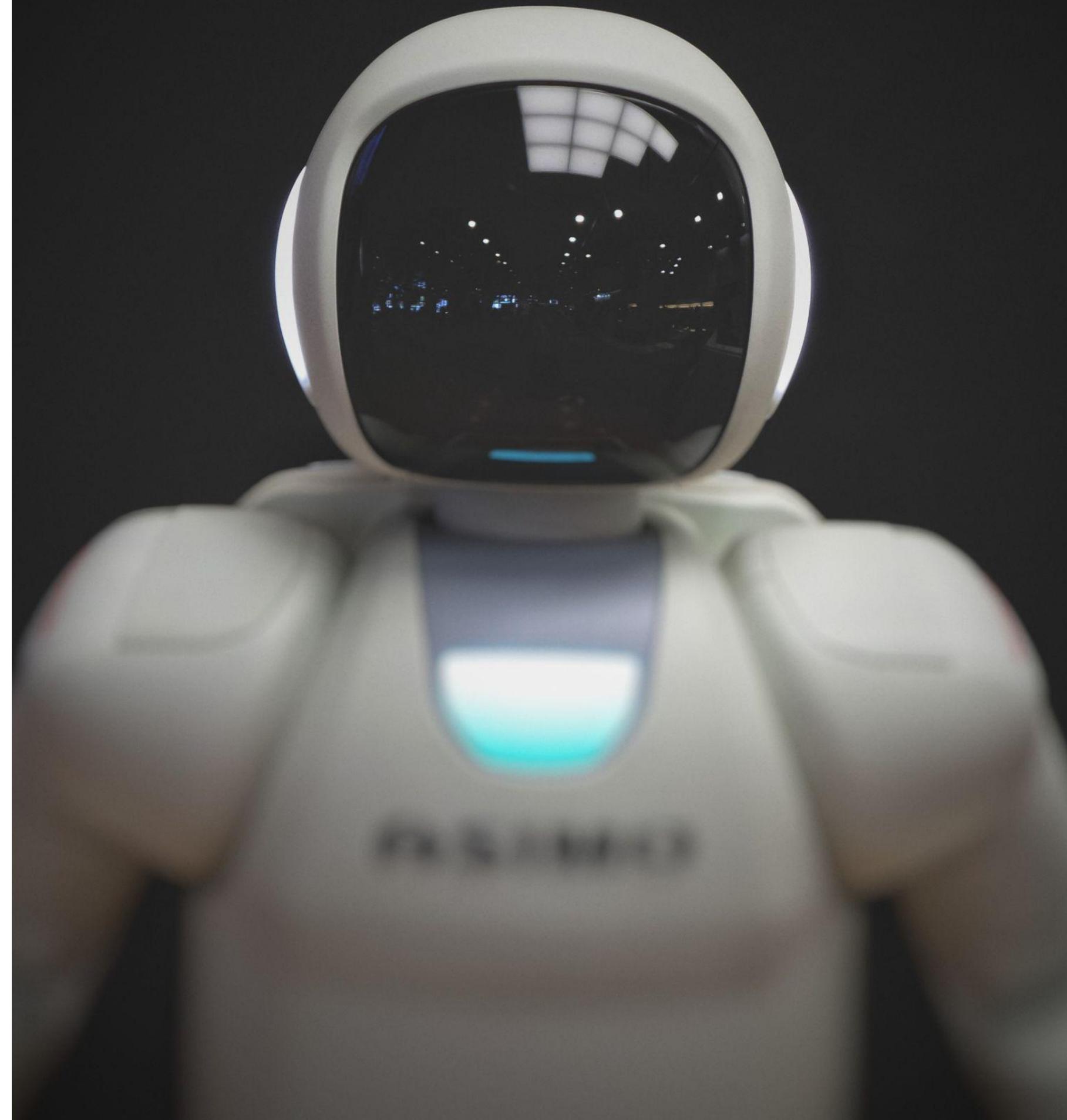
The Completeness dimension measures whether there are data missing or not at all. The Accuracy dimension measures whether the values make sense.

/03 CONSISTENT / DUPLICATE

The Consistency dimension measures whether values are the same across different data sources. The Uniqueness dimension measures duplicate presence.

/04 HAVING LINEAGE / OWNER

Lineage is all about tracking all operations on data, and by whom. Identifiability and Ownership are all about assigning an accountable owner to data.



DQ MANAGEMENT

DQ MANAGEMENT BIG DATA AND AI

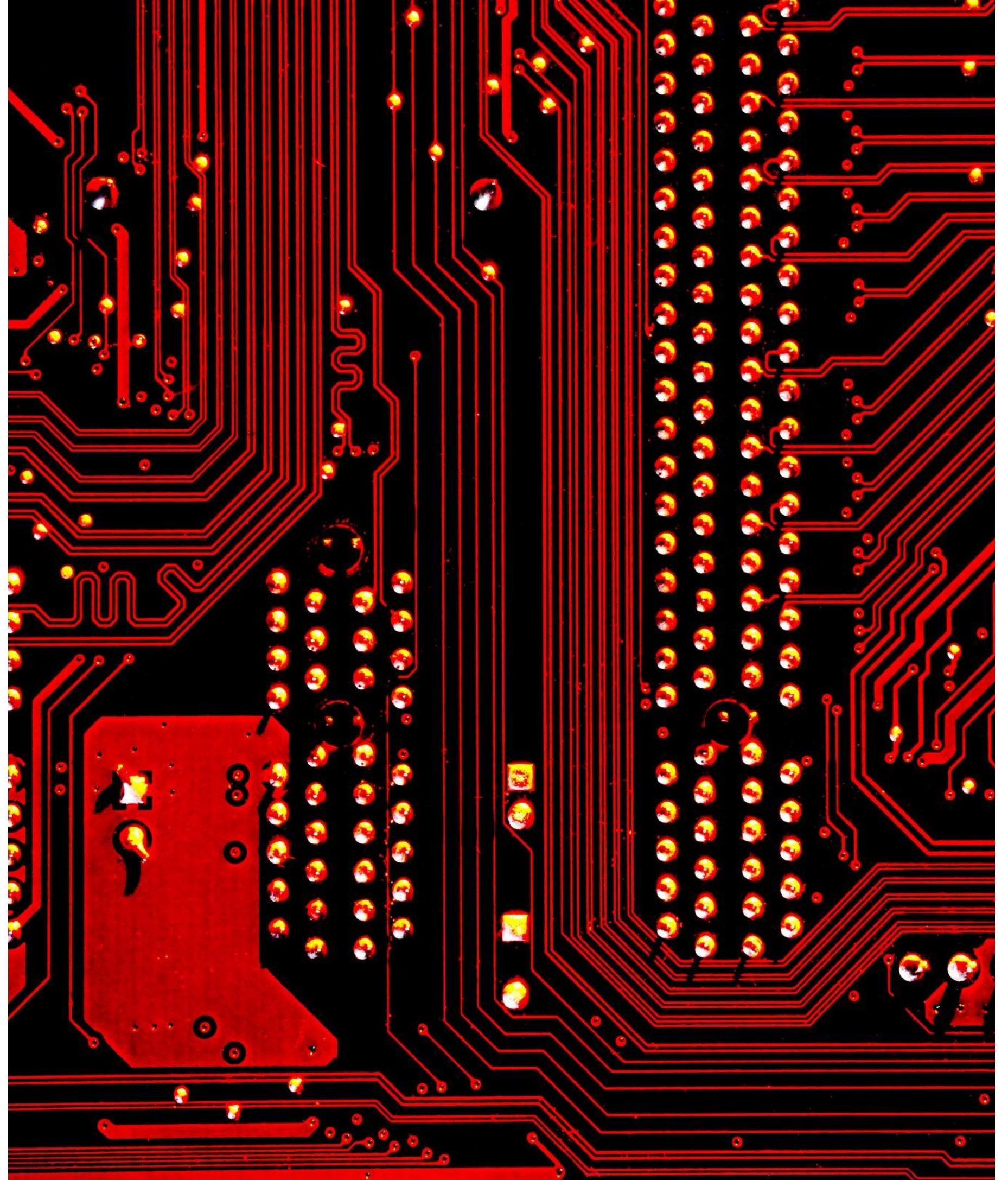
DQ issues become especially relevant when we talk about Big Data and Artificial Intelligence. The reasons for these are the following:

- In terms of Big Data, the volume of data is naturally much higher. This means that any issues or flaws in terms of DQ will be multiplied in terms of the reports and data products generated. Also, they are harder to fix (more intrusive);
- AI possibly presents an even bigger problem, which is that, if data are used to train a model (such as a Machine Learning model), then wrong data will create wrong results:
 - And if these results are fed back to the model, errors can be multiplied and amplified;

DQ MANAGEMENT BIG DATA AND AI

Additionally, two other factors which are also important in terms of Data Quality are the data volume and whether a Data Lake or Data Warehouse are being used:

- In terms of data volume, naturally, the higher the volume, the lower the chances of cleansing data upfront. A large volume of data brings with it the necessity to prioritise critical data and leave the other for later;
- In terms of a DL vs a DW, the choice inherently defines the data that go in. Data Warehouses only accept structured data by definition, so they must be structured when loaded (schema-on-write), while a Data Lake allows you to store unstructured data and structure it when you need to use it (schema-on-read). But DWs are less flexible afterwards;



DQ MANAGEMENT

DQ MANAGEMENT BIG DATA AND AI

Regardless of the platform, it's important to understand that Extract-Transform-Load (ETL) processes in specific can be a major propagator of DQ problems, for multiple reasons:

- If the data are of low-quality before going through an ETL process, we may propagate the errors within those data, or at least hide them in aggregates;
- For example, temperatures from 0 to 100. There are values of -1, representing errors. Those are not excluded, but averaged in aggregates that are loaded in the DW;
- If the ETL process (or DW) is not properly configured, it can take HQ data and kill granularity itself
 - For example, the DW has a credit card chargeback reason code field with 1 digit. But it's 2 digits. You lose 1;

DQ MANAGEMENT BIG DATA AND AI EXAMPLES

/01 BANK VaR

In banking, the Value at Risk model determines risk for a given time period. But banks have selected too positive periods to “fool” the model (not anymore).

/02 DIVIDE AND CONQUER

A common practice, due to the volume of big data, is to segment important data streams to be managed and governed, and ignore the rest until later.

/03 ETL AS AN EXAMPLE

While we are mentioning the dangers of ETL in specific, all transformations on data can be dangerous if not properly monitored and DQ tested before/after.

DQ MANAGEMENT BIG DATA AND AI KEY TAKEAWAYS

/01 VOLUME AND FEEDBACK

Big data and artificial intelligence present two distinct, but equally important challenges to DQ - a high volume of data and the propagation of data in models.

/02 VOLUME AND STRUCTURE

In terms of big data, the volume determines how much cleansing can be done beforehand. The structure (DL vs DW) determines speed/flexibility.

/03 ETL BEWARE

ETL processes, which take data from the operational DBs to a DW, can be dangerous. They can hide or propagate DQ errors, or can even create new ones.



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES

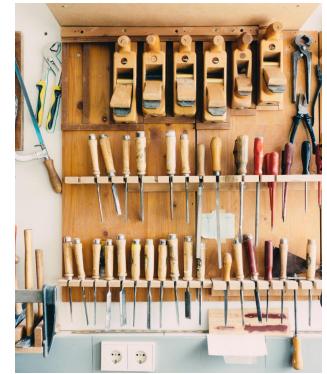
INTRODUCTION

In order to achieve higher Data Quality, a varied array of tools and techniques is used, with different purposes and implementations. These usually cover all stages of the DQ improvement lifecycle, including the detection of issues, the remediation of these, and the prevention of further problems.

These tools and operations can, in many cases, be used in combination, to achieve a higher improvement of the quality of data, and some commercial tool suites enable more than one type of functionality at once.

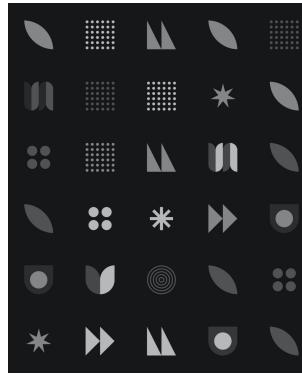
DQ TOOLS/TECHNIQUES

We are going to cover **five key topics** in terms of the tools and techniques used to improve Data Quality:



DQ TOOL OVERVIEW

An overview of the four main types of tools leveraged, as well as what each is used for



DATA PROFILING

Tools to assess and analyse data, detecting problems with values or formats



CLEANSING AND STANDARDISATION

Tools that transform values in different patterns to one common patterns and/or cleanse data



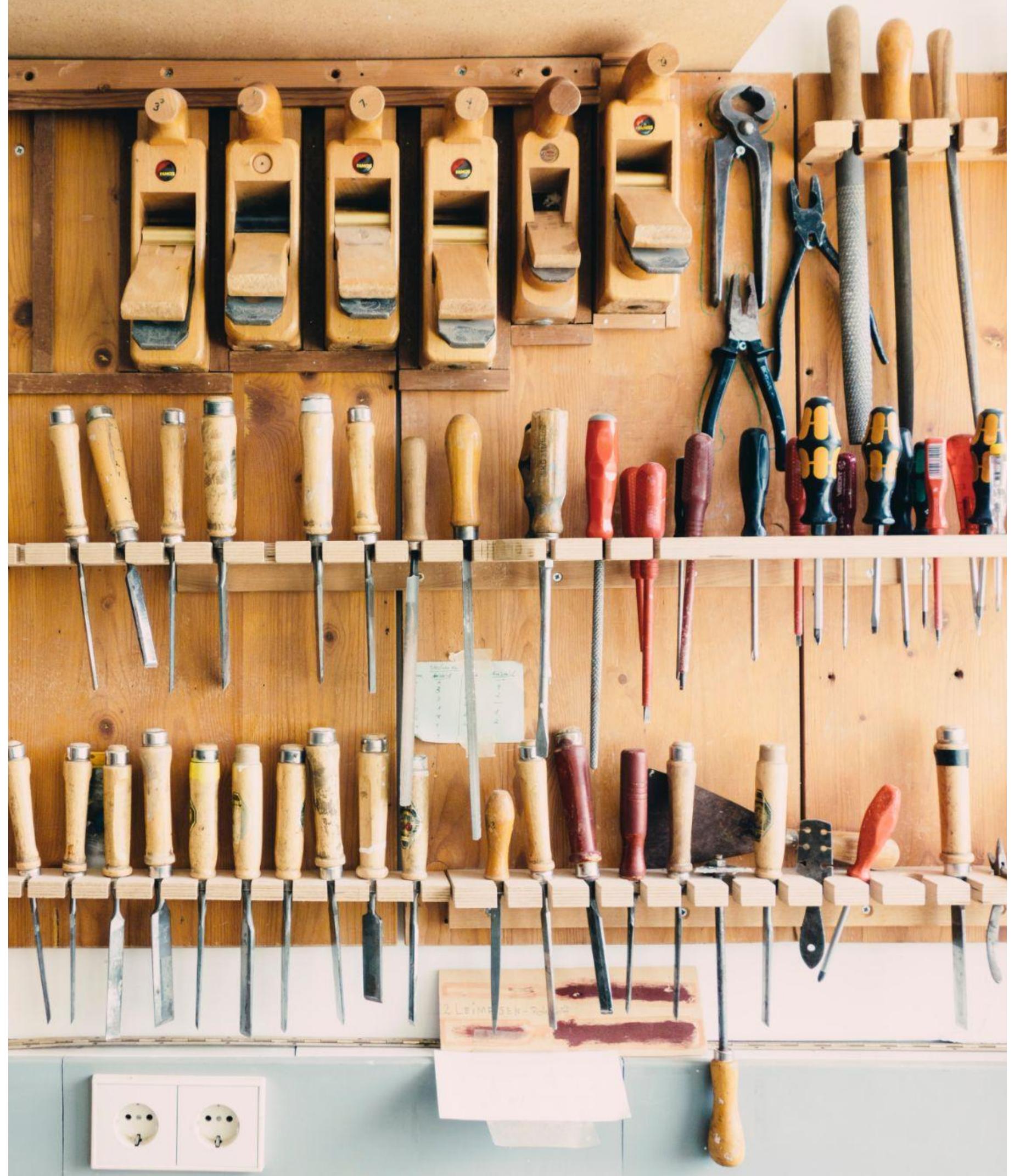
MERGING AND LINKING

Tools that match different records to link them to the same entity and/or merge them



DATA ENHANCEMENT

Tools that allow annotation and/or enrichment of current data by adding more useful data



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DQ TOOL OVERVIEW

In terms of Data Quality, there are multiple tools that can be used - not just to identify data problems, but in many cases to remediate them. These are considered Data Management tools.

Although tools vary, there are, in general, 4 key categories:

- Data profiling and auditing. These tools analyse the quality of data, in a manual or automated manner;
- Data cleansing and standardisation. These tools “clean” data, parse value formats, and standardise information;
- Data merging/linking. These tools merge datasets, merge records of an entity, or remove duplicates;
- Data enhancement/annotation. These tools allow you to add additional information, increasing quality through this;

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

The first family of tools are data profiling and auditing tools. As the name indicates, these tools take a dataset or data storage, and allow you to analyse its underlying quality.

- Most of the tools available allow different dimensions. So you can calculate the accuracy, completeness, consistency or other dimensions of a database or other storage;
- A common practice is to create a DQ scorecard before the profiling (or after its feedback), with the expected value ranges, which is used to score the data sources later;
- One of the most relevant questions in profiling is what do outliers represent (and whether to delete/keep them);
- This can be done manually, but also automatically, which is the source of DQ monitoring in DG, for example;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DQ TOOL OVERVIEW

Then we have tools for data cleansing and standardisation.

Many of these tools are used at the beginning of a data pipeline to increase DQ from the beginning. These include:

- Data parsing and standardisation. That is, taking data sources with values that represent the same thing, but may be in different formats, and converting them all to the same format. Currencies, zip codes, product categories...
- Parsing is the interpretation of values in different formats (\$7, \$7.00, 7 USD) and standardisation is the conversion of all to the same format (\$7.00);
- Cleansing removes invalid values and replaces them with more “reasonable ones”. For example, for temperatures from 0 to 100, removing values of -1 (machine errors);

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

Another category of tools are those that perform data merging and linking. These consist of merging fields that represent the same thing (or even entire datasets), including:

- Identity matching (the set of algorithms/comparison tools to determine that two records belong to the same entity);
- Duplicate record merging/deletion. Identifying records that represent the same element (usually builds on identity matching). Then removing these (or taking other actions);
- Duplicate dataset merging/deletion (taking two datasets with complementary information and merging them);

All of these techniques usually build on cleansing and standardisation.





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

Finally, another group of tools includes the ones that allow data enhancement or annotation. This is more frequent for data that are not structured and where it's not easy to change individual fields, or data that don't have a quality issue, but can be improved with more information anyway

- In these cases, it's still possible to add a field with notes explaining that the data are of high quality, and why;
- In other words, if you can't change the current data to improve their quality, add new information;
- In relational databases, think of it as SQL "JOIN"s;

This is very frequent for temporary situations (until data are structured), or when working with large, unstructured data (e.g. in Data Lakes).

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

It's not uncommon for multiple of these processes to occur:

- You perform data profiling, focused on accuracy, and find out most client zip codes are in different formats;
- You need to parse the different formats they are in and standardise them. You may also have to deal with outliers (for example, zip code “0”, where there is no address or the zip code was not provided), and figure out what to replace these values with. Fill them out by hand? Exclude them?
- You may find out that there is another table that contains customer phone numbers, but only that, so it's necessary to merge these two datasets, but first they must be standardised to make sure you can identify the same customer in both (identity matching with attributes);





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DQ TOOL OVERVIEW

There are some additional, important considerations here:

- As a general rule, the earlier you can increase data quality, the better. It's much easier to work on a dataset that was just entered, rather than a report that was generated from data having undergone 4 different transformations;
 - A great offender here are aggregations, which can remove details from data;
- This may not always be possible, though, e.g., with big data, in some cases the volume is so high the data is loaded first, and processed later (ETL comes after arriving at the DW);
- The difference between DG and DM becomes apparent here. DM tries to correct the flawed data, while DG defines rules so it must be complete to begin with;

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

It's important to note that, just like DM and DG are different disciplines, so are the tools used. The tools mentioned so far are primarily for DM. For DG, tools focus on the governance aspect, and may include, for example:

- Metadata management. Tools to create glossaries, rules, policies, metrics and others facilitating governance. Usually include functionality to categorise and classify data;
- Lineage and provenance. Considering lineage is very relevant to DG, some solutions provide this. These tools analyse data throughout creation, transformations and movements, and capture that information;
- Administrative. These tools help define policies, workflows, roles and responsibilities to ensure proper governance;



DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

EXAMPLES

/01 MATCHING ATTRIBUTES

The set of attributes used to compare identities is based on many criteria. The value range they can have, how frequently they are filled, how unique they are...

/02 PROFILERS VARY

While profiling can “technically be considered DM, depending on the organisation, different people can do it. In some cases, Data Stewards do the profiling.

/03 OVERLOADED CODES

One specific meaning of outliers may be that the field is used as a code for something else, which is “overloading” - the field has an additional meaning.

DQ TOOLS/TECHNIQUES

DQ TOOL OVERVIEW

KEY TAKEAWAYS

/01 IMPROVING DQ

Tools for DQ focus on, naturally, improving DQ, regardless of dimension. They both assess the current problems and allow practitioners to fix them.

/03 MORE THAN ONE

It's very possible that, during DQ work, more than one tool has to be used. For example, parsing, cleansing and standardising, de-duplicating, then merging.

/05 THERE IS ANOTHER

Although profiling and remediation techniques are in scope of DM activities, DG tools also exist. These, instead, focus on policies, classification, and admin.

/02 4 MAIN FAMILIES

DG tools are usually for profiling/auditing, detecting problems, for cleansing/standardising, improving DQ, for merging and linking, or for annotating information.

/04 THE EARLIER THE BETTER

Considering the number of transformations that data may go through, the earlier in the lifecycle that these tools can work, the better the situation downstream.



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA PROFILING

Data profiling is the analysis and assessment of values and content in datasets. It's, in many cases, the first step in a DQ improvement process, as it brings problems to light.

- It's also, usually the most frequent DQ type of technique (and what most people associate with DQ improvement);

There are usually 4 types of profiling operations:

- Data value validation (whether values are reasonable);
- Anomaly discovery (identifying values that stand out);
- Data format and model validation (validating the format of the data field, format, or other specifications);
- Business rule discovery and/or validation (deriving business rules from existing values, or validating these);

DQ TOOLS/TECHNIQUES

DATA PROFILING

Validation of the data values is a frequent type of profiling. In other words, you gauge whether values in a given DB column, table (or other format) are as expected, or not at all.

Data values can be in compliance (or not) with a couple of key areas:

- Type compliance (That is, do values match the type? For example, having characters in a numeric field does not);
- Domain compliance (That is, are values within the expected range? Do we have -1 values in a 0 to 10 field?);
- Business rule compliance (Are values obeying specific, custom rules? For example, if a Customer requires at least one valid Address, do we have Customers without one?);





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA PROFILING

Anomaly analysis is the detection of values that are significantly different from the majority.

- It's important to note that this can, in fact, indicate errors, but these can also be legitimate outlier values.

Anomaly analysis usually first determines a “baseline” of values (or the domain of acceptable values), and detects outliers through one of various ways:

- Performing a frequency distribution and detecting the values that are in very low frequency;
- Examining the variance of values and detecting the values that are in the extremes;

Anomaly analysis can usually be aided by data from other columns (or tables).

DQ TOOLS/TECHNIQUES

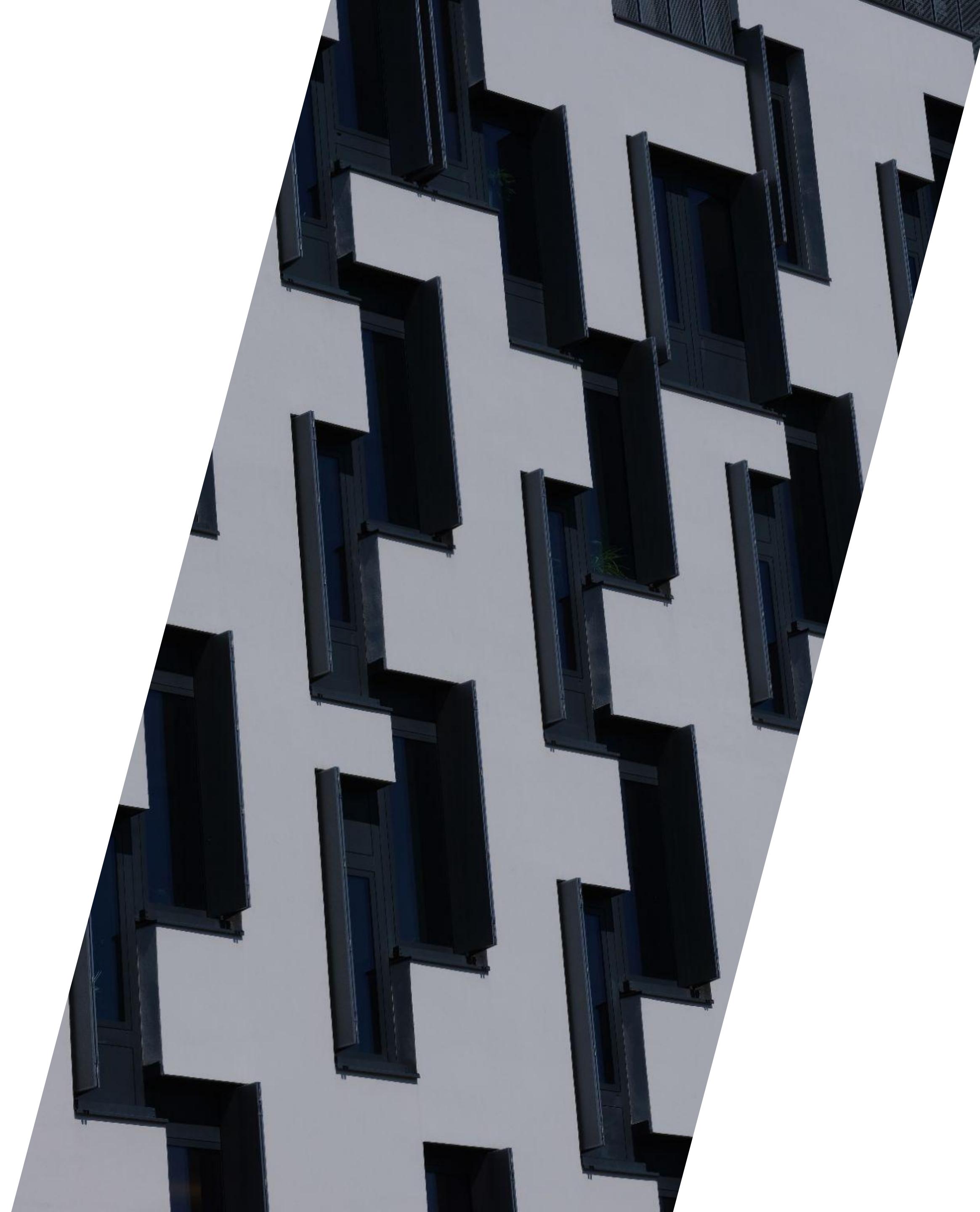
DATA PROFILING

Data format and model validation is usually the reverse of data value validation, but uses a similar process.

- Data value validation tells you that the rule is for values to be from 0 to 100, and this one is outside, so it's "wrong";
- Data format and model validation tells you that most values appear to be outside 100, so the rule of 0 to 100 is "wrong" itself;

Usually, the same three types apply:

- Data types: The types used may not be the specified ones;
- Data domains. The value ranges may not be the specified ones, but may have different ranges not documented;
- Data business rules. The data present may not obey the specified rules, but maybe other ones not mentioned;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES **DATA PROFILING**

Data profiling can also be used to discover business rules, or even to validate these in datasets (there is some degree of overlap with data format/model validation).

- For example, if we have no metadata for the Customer class, but profiling shows that all Customers have the Phone Number field filled, and with a specific format (e.g. "111 1111-1111", the rule that "Every Customer must have a valid Phone Number in the format 111 1111-1111" can be suggested;

Usually, if there is no metadata regarding business rules, profiling can help discover new rules, that are present in the data but are not formal.

- If there already are rules, these can be validated;

DQ TOOLS/TECHNIQUES

DATA PROFILING

It's important to note that data profiling usually has two major uses, one related to data themselves, and the other to metadata:

- First, profiling can be used to raise DQ issues:
 - What you think of when you think of profiling. Looking at the data and saying “These are probably wrong”;
 - Either because values are outliers, don't follow the rules established, many of them are missing, or others;
- Profiling can also be used to enrich metadata:
 - Metadata should (ideally) contain information on all business rules, all data formats, fields, value domains, and more. But it may not be updated (or even missing);
 - Data profiling can help discover rules and formats;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA PROFILING

In terms of techniques, data profiling usually leverages statistical and mathematical techniques in order to derive insights. These include, for example:

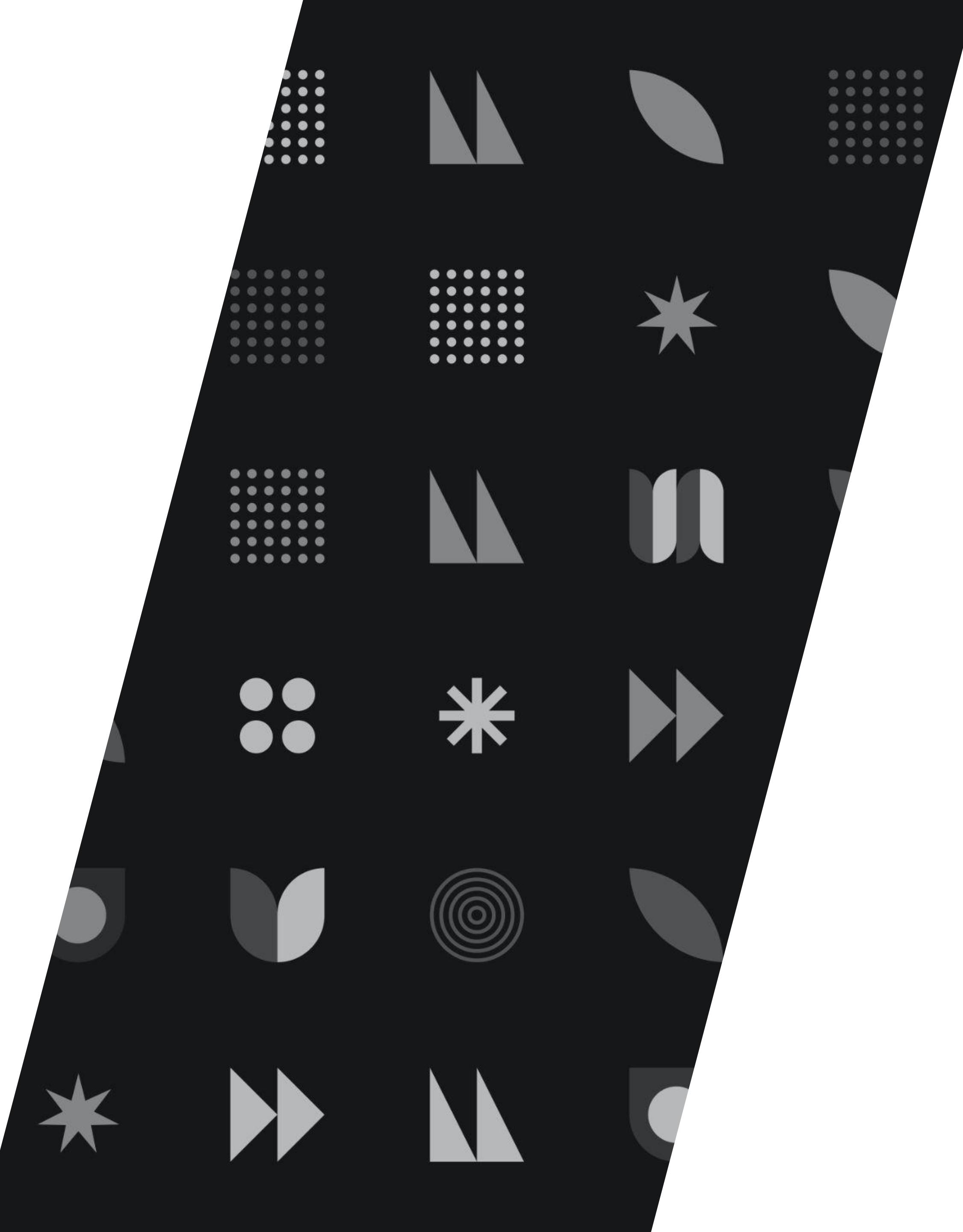
- Determining ranges. Are these values constrained within a range, and if so, what is the range?
- Determining absent and missing values (how many values are empty, and what shape do they take when empty?);
- Determining cardinality (how many distinct values are there?) and frequency distribution (for each distinct value, how many times does it occur?);
- Additionally, for numerical values, other calculations such as the minimum and maximum, the mean and median, and the standard deviation may also be helpful;

DQ TOOLS/TECHNIQUES

DATA PROFILING

Some of these analytical techniques, in specific, provide interesting insights regarding rules and formats:

- If most (or all) values in a column are unused, that means this column is not used (or used for a different purpose), or could be represented in a different way;
- If a field is used for multiple purposes, it's probably an overloaded element or composed element, which contains more than one type of information, (e.g. "Price with Tax"), which could be split into two fields (or more);
- If a field's values all fit a certain pattern, that should be a business rule or constraint (e.g., phone numbers are always in the (111) 1111-1111 format);
- If a field is derived from others, such as a "full name";





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA PROFILING

When dealing with data present in multiple tables, profiling can also be useful to detect dependencies and correlations. This can be useful for multiple purposes:

- Detecting the presence or absence of either primary or foreign keys. For example, detecting if all records in a table have a unique identifier (PK) and if records referencing other table records are referencing a valid object (FK);
 - Can detect both non-existing “parents” and “children”;
- Detecting dependencies and correlations among data, which can indicate a dependency relationship.
 - For example, a Customer may only have a Delivery Address if they have at least one Purchase, indicating that one probably determines the existence of the other;

DQ TOOLS/TECHNIQUES

DATA PROFILING

EXAMPLES

/01 AUTOMATED TESTING

Although data profiling is used in a manual manner to detect DQ issues, once the rules are set, automated tests can be run to monitor current data quality.

/02 DIFFERENT MEANINGS

Anomalies are very important, because they can occur for a myriad reasons. They may be wrong values, they may be exceptional situations, or may be other fields!

/03 DIFFERENT FUNCTIONALITY

Although some tools only perform data profiling, others include other DQ operations such as remediation, so that both can be integrated.

DQ TOOLS/TECHNIQUES

DATA PROFILING

KEY TAKEAWAYS

/01 FOUR MAJOR PURPOSES

Data profiling analyses data, with 4 major purposes. Validating the data, finding anomalies, validating the formats, and/or discovering business rules.

/02 VALUE ANALYSIS

Data profiling can reveal whether data values are valid or not, usually in three major areas. Whether they fit the expected format, the range, or specific rules.

/03 ANOMALY ANALYSIS

Data profiling can reveal anomalies by detecting which values do not fit the ranges (whether formal ones, such as rules, or informal ones - infrequent ones)

/04 MODEL/FORMAT ANALYSIS

Data profiling can validate the data models or formats used, that is, if the existing data fit them or not. Usually also in terms of format, range, or specific rules.

/05 BUSINESS RULES

Finally, data profiling can be used to derive business rules from the values present. Or, if there are already formal business rules, it can help validate these.

/06 NUMBERS AND STATISTICS

Most profiling methods are numerical or statistical. Frequency distributions, counts, value ranges, missing numbers, standard deviation, and others.



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES **CLEANSING AND STANDARDISATION**

Data cleansing, parsing and standardisation are all operations that may be necessary to make data more valuable and allow them to be used more easily.

- It becomes especially important when the data are of low-quality or come from a third party, which may not have treated the information beforehand;

There are several types of activities in this family, but they focus on solving issues such as:

- Typos, wrong abbreviations (or abbreviation expansions), which create variation in otherwise similar text values;
- Wrong formats, such as having numbers in a text field, not allowing the field to be used in its true, “native” format;
- Overloaded fields, which contain more than 1 piece of data;

DQ TOOLS/TECHNIQUES

CLEANSING AND STANDARDISATION

Usually, when we talk about cleansing, parsing and standardisation, it's important to define what each term, individually, means:

- Parsing is the division of existing values into smaller values (known as “tokens”), which are later reorganised;
- Standardisation is the conversion of values in different formats to values in the same format;
 - It usually comes after (and requires) parsing beforehand;
- Cleansing is the removal of invalid or empty values in information, which makes the information homogeneous;

Note that one or more of these processes may be necessary to treat information. For example, you may just need to cleanse “wrong” values, or cleanse + standardise them.





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES CLEANSING AND STANDARDISATION

Parsing is usually the first step in any cleansing process, and it consists of taking existing values and breaking them into basic elements, known as “tokens”:

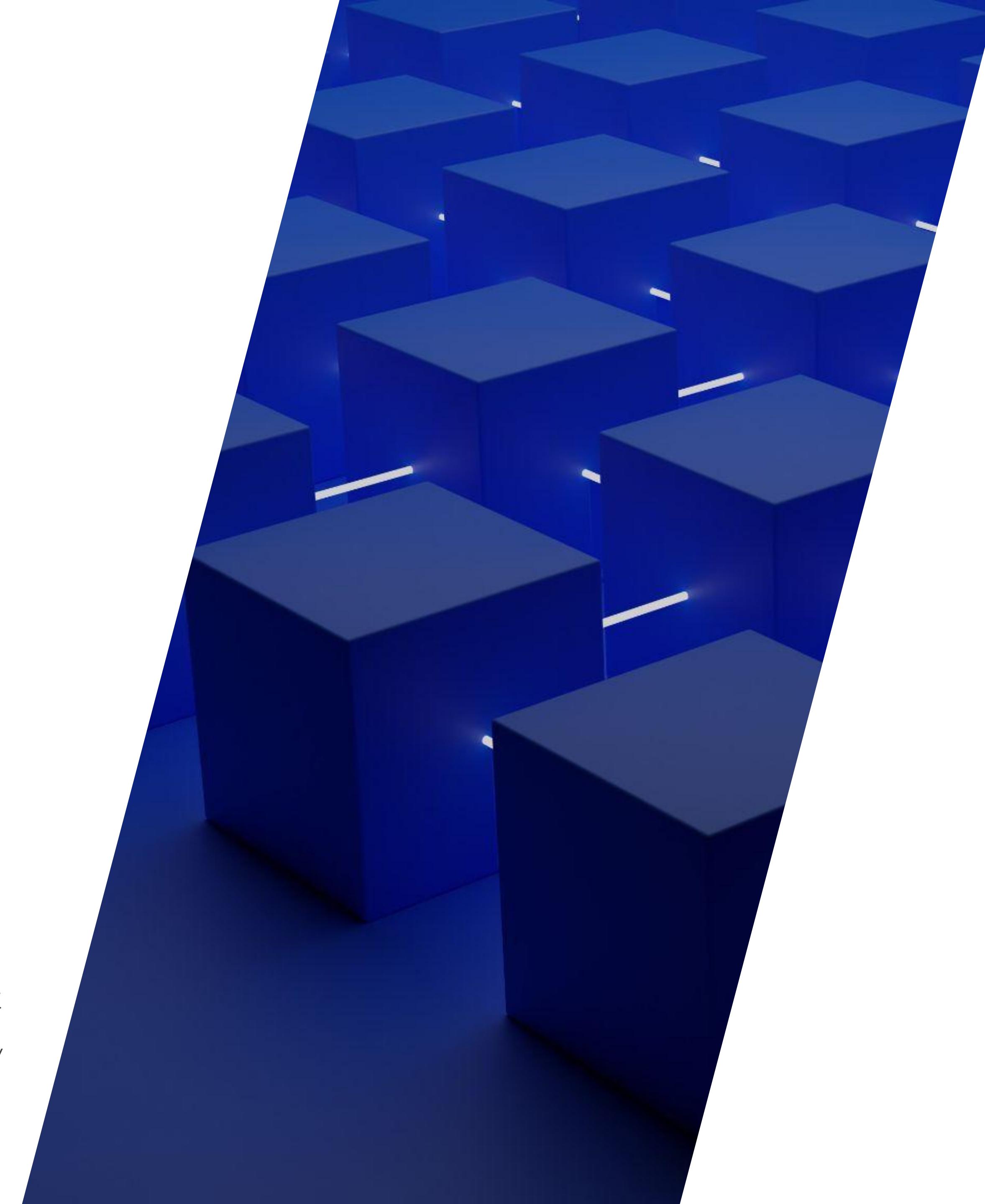
- For example, a street address can be broken down into Street Name + Number + Floor + Zip Code + Country;
 - Possibly, additional fields;
- The parsing tool or technology used can usually take data in multiple different formats:
 - “21 Oak Street, 12345 New York, USA”
 - “Oak Street, 21, 1245 NY”
 - “21, Oak Street, 12345, New York, United States”;
- The output of the parser is usually a list of ordered tokens:
 - [“Oak Street”, “21”, “12345”, “New York”, “USA”];

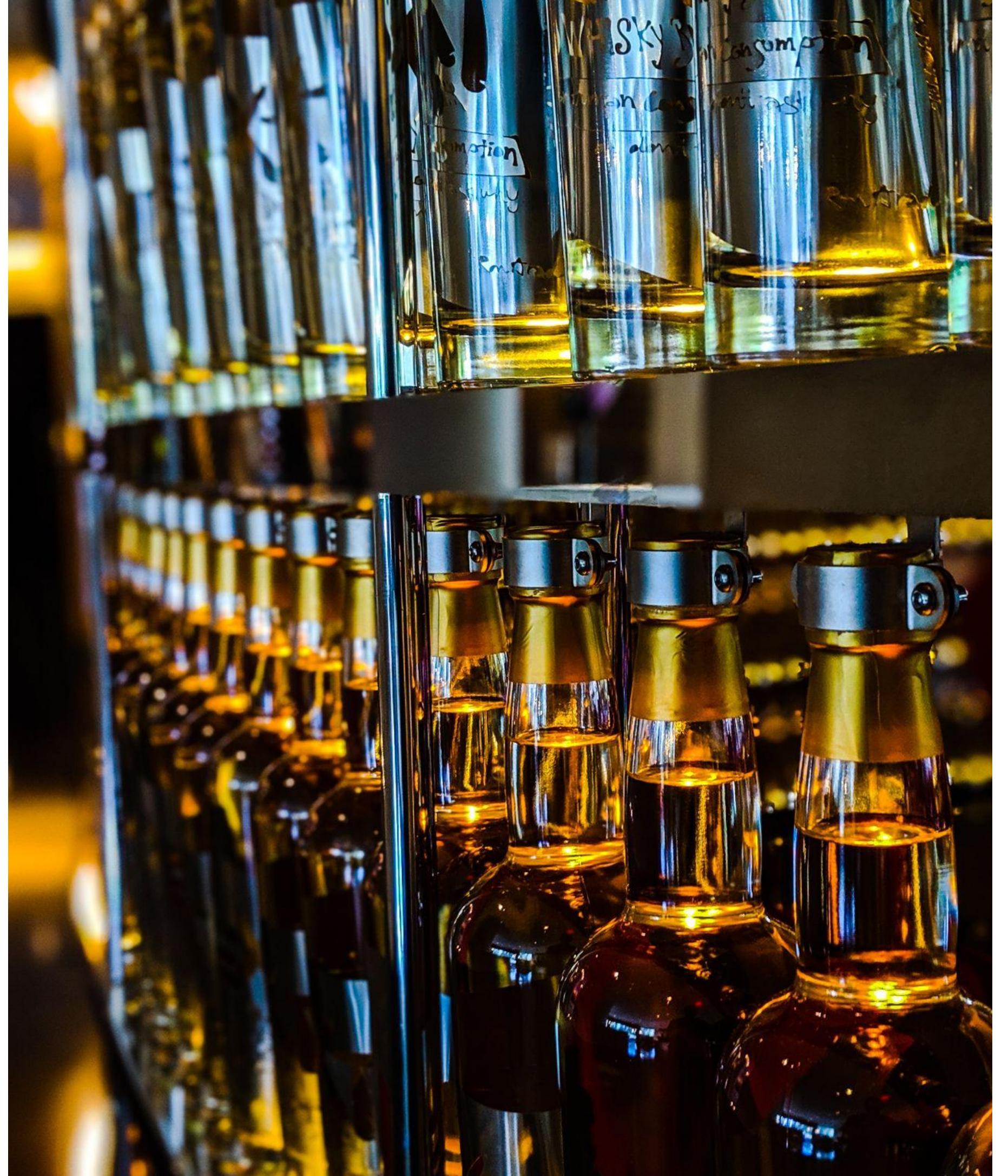
DQ TOOLS/TECHNIQUES

CLEANSING AND STANDARDISATION

Both cleansing and standardisation usually come after parsing, and they deal with the existing tokens:

- Cleansing usually focuses on eliminating or changing invalid tokens;
 - For example, in a list of 20 addresses, one may have an invalid zip code but valid city. The cleansing tool may search the zip code by street name and state, and replace the zip code token with the correct one;
- Standardisation focuses on rearranging tokens to place them in a determined format, which all records will obey;
 - For example, placing all addresses in the format “(Street Name) (Number), (Zip Code), (State), (Country)”, possibly each one in their own field;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES CLEANSING AND STANDARDISATION

Both parsing and standardisation are usually two very technical operations, which rely on specialised tools that can perform both the parsing and standardisation of different data types, including frequent examples such as:

- Standardising customer names into first + middle + last name (or other variations including title and so on);
- Standardising product codes and materials into codes used by a specific taxonomy;
- Standardising addresses into address + zip code + country (and possibly others, such as a flag for apartment/home);

In either case, the process remains the same, with two key stages: breaking data into tokens and reassembling them.

DQ TOOLS/TECHNIQUES

CLEANSING AND STANDARDISATION

The process of cleansing data may differ in terms of:

- Earlier vs. later: Just like other data remediation techniques, the earlier standardisation can be performed in an information flow, the better;
- Standardising data in a report whose data have gone through 8 transformations means doing it at all levels (and possibly with different results). Doing it at the source is potentially a lot better;
- Naturally, this depends on where the data come from (e.g., 3rd parties) and the volume of data;
- Synergies with profiling: Data profiling efforts can indicate value domains and formats for data, which can, for example, be used to standardise the data that don't fit;



DQ TOOLS/TECHNIQUES

CLEANSING AND STANDARDISATION

EXAMPLES

/01 REGULAR EXPRESSIONS

Regular expressions are one of the most powerful ways to describe different patterns. They are not easy to master, but are very flexible in what they can do.

/02 STRING - LIST - STRING

Usually, parsing a string of characters splits a character string into tokens, so we end up with a list of smaller strings. Then, these are collated into the target string.

/03 REFERENCE DATA COUNTS

In many cases, standardisation is paired with reference data. For example, replacing zip codes as strings by references in an existing zip code table.

DQ TOOLS/TECHNIQUES

CLEANSING AND STANDARDISATION

KEY TAKEAWAYS

/01 SIMPLER AND EASIER

Cleansing, parsing and standardisation have the goal of reducing irregularities in data and making them easier to understand, and overall, easier to use.

/02 3 DIFFERENT OPERATIONS

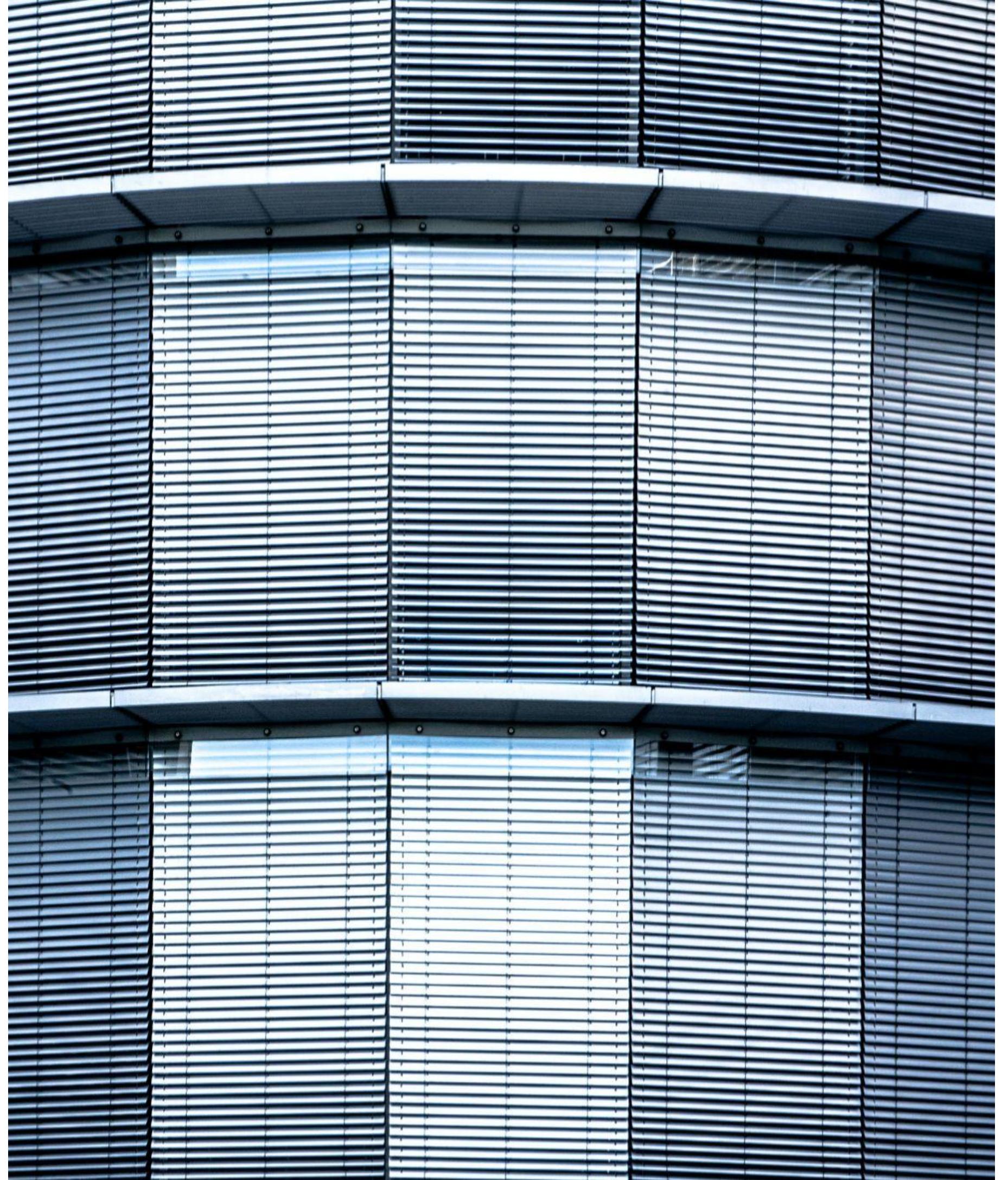
Technically, the 3 terms are different. Parsing is breaking data into tokens. Cleansing is changing the invalid ones. Standardising is reassembling them.

/03 TECHNICAL AND VARIED

Parsing and standardising are usually technical operations performed by specialised applications, and can apply to customers, products, and other data.

/04 GOES WITH PROFILING

Parsing, cleansing and standardisation are operations that go well with profiling, as the latter can reveal rules and formats that can then be used.



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES **MERGING AND LINKING**

Merging and linking are two specific types of operations that deal with one specific problem: record duplication. They are important to centralise data - especially, critical data.

- They are a staple of activities such as Master Data Management (MDM), which aims to centralise the core entities in terms of data;
- It's important to note that duplication doesn't always come in the form of records that are exactly the same - in some cases, we may have records with "similar" fields (equal but in different formats), or that reference the same entity in different ways (name versus ID, for example);

In both cases, we will need to link different records which represent the same entity, and later merge them.

DQ TOOLS/TECHNIQUES MERGING AND LINKING

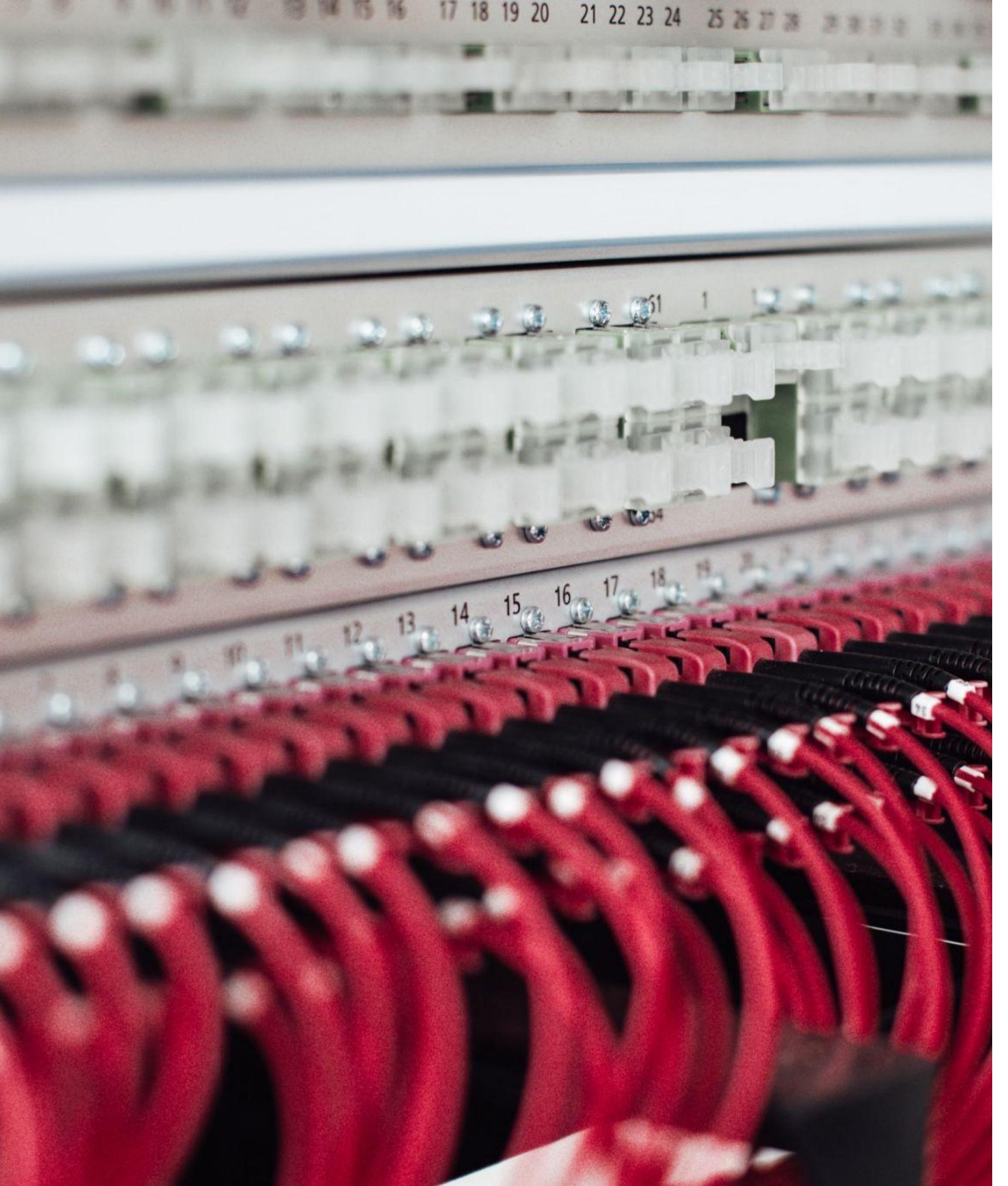
Duplicates and redundant information can come from countless sources (especially in big companies):

- Acquisitions and purchased data;
- Having multiple, non-integrated applications where users enter data in parallel (especially if with weak controls);
- Lack of standards in terms of data entry;

Both linking and merging records usually rely on one activity, which is identity matching or identity resolution. That is, the capacity to compare two records and conclude they represent the same entity (or not)

- The sophistication of the matching algorithm (together with the data quality) defines the quality of the matching;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES MERGING AND LINKING

In order to perform identity matching, we need to define which attributes are unique to an entity (as is necessary for any comparison algorithm).

- For a person, it may be a date of birth combined with a full name. Or date of birth + place of birth. Or many others;
- In DB terms, these generate the Primary Key of a table;

Therefore, the process of identifying the attributes compared is very important to identity matching.

- It's why, in many cases, identity matching can benefit from parsing and standardising the data first, so you are sure that these attributes are all in the same format;
- Also, remember not all data are contained in DBs with PKs;

DQ TOOLS/TECHNIQUES MERGING AND LINKING

There are several elements that can make certain attributes more or less unique, and therefore more or less useful as identifiers for an entity (whether alone or in combination).

Three key ones are:

- The structure of the field (That is, a numeric field will probably obey stricter rules than a short text one, which will be better than a long text one, and so on);
- The completeness and value distribution of a field (That is, a field that is blank in most rows, or that is the same in most rows will not be very unique);
- The presence of the field in other entities (E.g, Employee may have a “Birth date” field which also exists in “Customer, Partner, etc, and an “Employee ID” one, which doesn’t);





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES MERGING AND LINKING

Usually, specialised matching algorithms are used to classify records as belonging to the same entity or not. These are usually either binary (Producing a “Match”/“No match” result), or probabilistic (Outputting a number between 0 and 100%).

There are two important things to consider:

- What do about values in the middle. Some records are clearly the same entity. And others are clearly a different one. But what about the others that are in the middle?
 - May require iterating and refining the algorithm;
- Whether false positives or false negatives are preferred. We must err on one side, so would we prefer to consider entities the same if they’re not, or the opposite?

DQ TOOLS/TECHNIQUES MERGING AND LINKING

Finally, an important topic is the topic of which information is prioritised. That is, in case we have two distinct records that clearly refer to the same entity, but with different values for some fields, which end up in the final, merged entity.

- This process is also called survivorship;
- For example, two customer records. “John Smith” of 17 Oak Street, New York, and “John Smith” of 27 Oak Street, NY are the same. Which address remains in the final version?

In this case, it's important to establish a ranking of the data sources by quality. This allows a simple heuristic of, in case of conflict, prioritising data from the highest-ranking source.



DQ TOOLS/TECHNIQUES MERGING AND LINKING EXAMPLES

/01 INTERNAL COMPARISONS

The actual algorithms used for comparison can use several techniques, including string comparison, edit distance, and others, depending on the data types.

/02 DIFFERENT OR SAME

Record merging can be in terms of actual records (merging a Customer with a Customer) or different parts of one (associating a Customer and their Sales).

/03 COMMERCIAL OR CUSTOM

As with parsing and standardisation, commercial solutions can be used, but for specific needs, custom comparison algorithms may also be leveraged.

DQ TOOLS/TECHNIQUES

MERGING AND LINKING

KEY TAKEAWAYS

/01 LINK AND MERGE

To deal with duplicates and redundancy, linking and merging are frequent operations. Linking links two records to the same entity, merging merges them.

/02 IDENTITY MATCHING

Linking two records to the same entity relies on identity matching, which is done through a comparison algorithm, depending on the attributes.

/03 FORMAT, VALUE, PRESENCE

The selection of the uniqueness attributes is crucial for matching. Values are good if they have a strict format, unique values, and/or are present in only that entity.

/04 MATCH / NO MATCH

Matching algorithms usually detect either a match or a non-match. But we must deal with inconclusive results, and prioritise false positives or negatives.

/05 ALL THAT REMAINS

It's very possible for different data sources to contain conflicting information on the same entity, and selecting which information remains is important.



DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA ENHANCEMENT

Data enhancement is the process of appending additional information to a dataset (or data product) to increase its quality. It's usually common in two specific scenarios:

- When the data cannot be changed, but additional information about the data can (and should) be noted
 - For example, a dataset that can't be changed due to heavy use by critical processes. But a column can be standardised and added as a new one;
 - In particular, data about data is known as annotation;
- When the data are already of high quality, but additional data can be added to improve their usefulness;
 - For example, we have marketing information with name/age. But we can add an “interest” estimate to it;

DQ TOOLS/TECHNIQUES

DATA ENHANCEMENT

Data enhancement can either be done by modifying the actual dataset or generating a new data product. For example:

- If we have data that cannot be changed right now due to critical processes, but we want to standardise the values in a column, we create a new one with the standardised values. The original data are used, unchanged, but the dataset now contains new information;
- If we have an algorithm that generates a probability of purchase, we can take a list of customers, and append, for each, the probability of purchase;
 - Since this probability may change frequently, a new report or view should be generated (not saved in data);





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA ENHANCEMENT

You may notice that, unlike other Data Management (DM) operations that directly modify data, data enhancement is partly an effort of metadata and documentation, and not just “pure data editing”.

- In some cases, the additional annotations and information obtained are actually stored in the metadata repository and not the data structure itself;
- E.g., User logs. Do we store info in the files? DB? Others?

Additionally, this is a type of data quality operation that, similar to other ones, relies on standardisation.

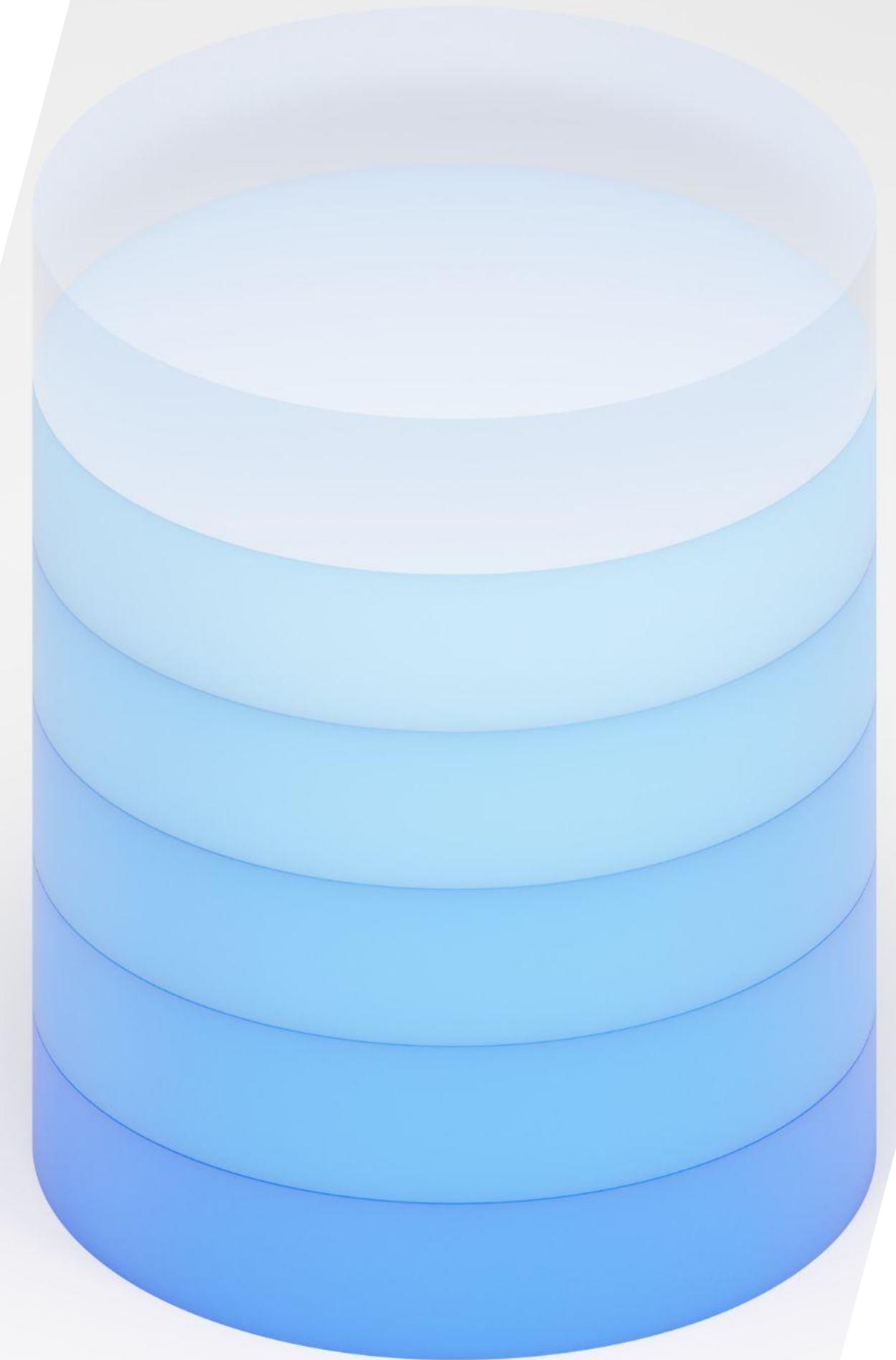
- If you’re appending a “Purchase %” that is based on age and address, the more standardised these are, the better;

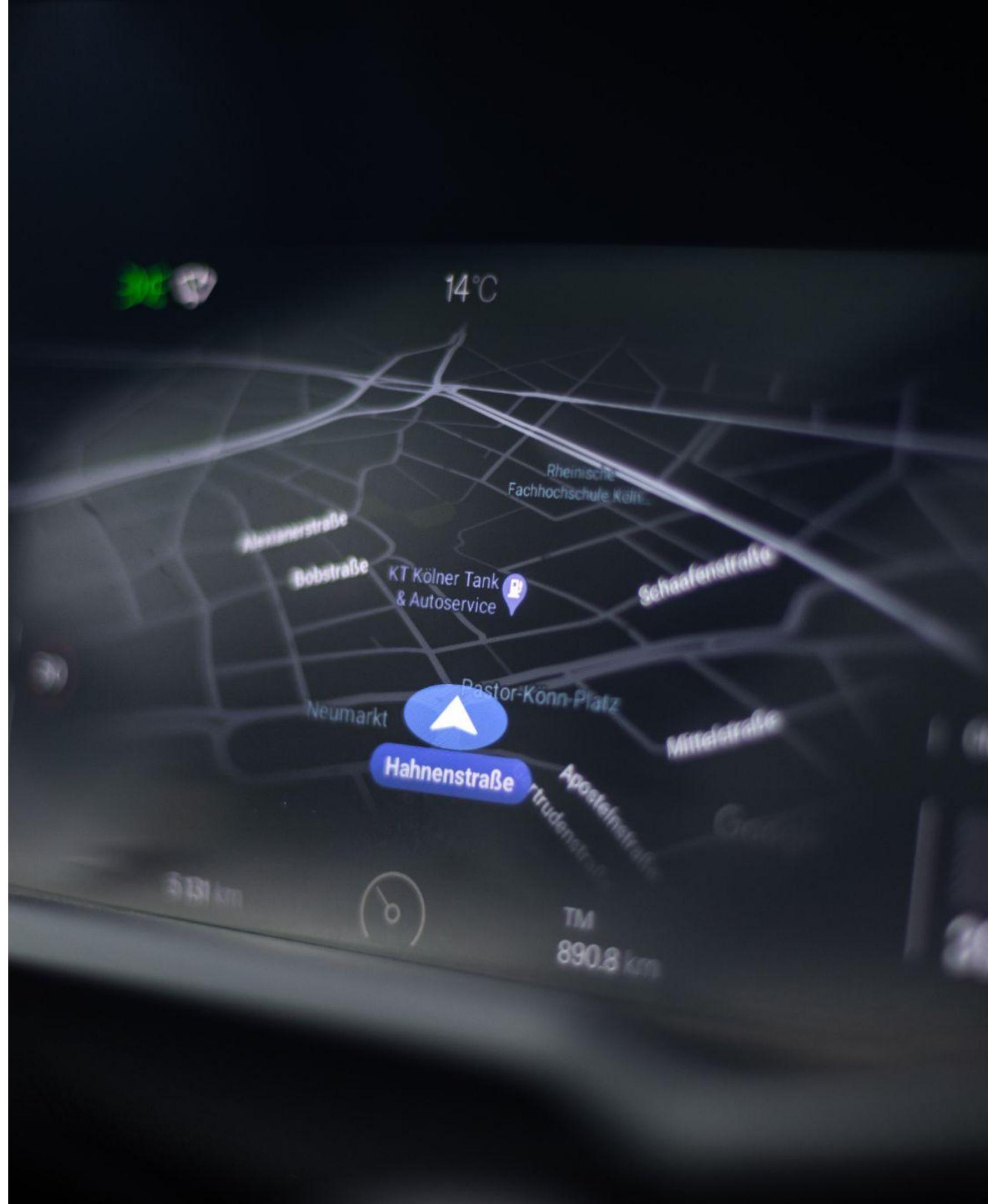
DQ TOOLS/TECHNIQUES

DATA ENHANCEMENT

In relational SQL databases, data enhancement is frequently thought of in the analogy of a “JOIN” statement.

- In other words, you “pull” data from another table, which is added to the current results;
- For example, take a user activity event (for Information Security logging purposes). You have the base table, with “User ID”, “Event Type”, “Timestamp”;
 - For example, “User 32”, “Login”, “2022/01/07 14:10”;
- If you have another table, with the last IP address for every, with “User ID” and “IP Address”, you can pull from it;
- You end up having “User ID”, “Event Type”, “Timestamp”, “IP”;
 - For example, “User 32”, “Login”, “2022/01/07 14:10”, “192.168.1.1”;





DQ TOOLS/TECHNIQUES

DQ TOOLS/TECHNIQUES DATA ENHANCEMENT

There are several types of additional information sources that can usually be used to add meaning and value to a given data source:

- Associated location data. IP addresses, geographical coordinates, proxy locations, others;
- Associated device information. Device type, cookie data, device ID, MAC address, others;
- Associated activity. Products this person has bought, companies they follow, posts they have liked, stores they have physically visited, others;
- Associated users and entities. Who registered a purchased, who created a dataset, who is the owner, others;

Naturally, depending on the purpose, others may be relevant.

DQ TOOLS/TECHNIQUES

DATA ENHANCEMENT

When performing data enhancement, it's very important to understand what the attached data mean. In our example of attaching a User IP to event logs:

- Table with "Most Recent IP of User":
 - Is it even useful? Will have changed in the meantime;
- Let's say you have a table of "User Login IPs":
 - Are you obtaining the IP address with the same timestamp of the event? Clarify. "User IP at Event Time";
 - But we only similar timestamps for logins (All Login events will have an IP, all others empty);
 - You can obtain the first IP address after the event. But name it "First User IP After Event Time". Also, include the timestamp difference. May be 60 min, may be 900;



DQ TOOLS/TECHNIQUES

DATA ENHANCEMENT

EXAMPLES

/01 LOGGING AND METADATA

Data enhancement is very useful for creating audit trails and establishing metadata. It allows you to associate locations, users, previous activity to events.

/02 FRAUD PREVENTION

Similar to the previous example. By appending information from other transactions, other merchants to one specific activity, it's easier to gauge authenticity.

/03 SALES AND MARKETING

One of the most frequently used applications of data enhancement is to add additional information to leads and customers, including locations, demographics...

DQ TOOLS/TECHNIQUES

DATA ENHANCEMENT

KEY TAKEAWAYS

/01 CAN'T OR WON'T CHANGE?

Data enhancement works by appending data to existing data. It's used when we can't change the current data, or when we don't need to, as it's of HQ.

/02 DATA OR DATA PRODUCTS

After the information is appended, the enhanced data can actually be saved in the dataset, or a data product generated - what makes most sense.

/03 DATA OR METADATA?

Since in a lot of cases, the "supporting" data is data about data, it may be considered data or metadata, and it can therefore be stored in different formats.

/04 SQL "JOIN"s ANALOGY

In relational databases, the SQL JOIN analogy is a great example. For each record in a table, you pull more data from another table to add information.

/05 DIFFERENT CATEGORIES

There are several types of data which can enhance other data, such as locations, user activity, associated devices, and other contextual information.

/06 CLARIFY NEW DATA

When you add new information, it usually has very specific meaning, and the name and documentation should reflect it. "Last location" vs. "Current location"...



DATA AND DATA QUALITY

DATA AND DATA QUALITY BUSINESS CASE BUILDING

A business case is a document that is necessary for most (if not all) data-related initiatives (whether DM or DG). Naturally, all initiatives must prove value and be profitable, so a business case is involved in this process.

- But more than that, it's important to notice that, for many executives, data seem intangible. It's hard to make them think of them as assets, with possible financial gain (or they may just resist) A business case helps crystallise these gains;

Regardless of whether DG or DM, there are common parts:

- Identifying the current costs of low-quality data;
- Detailing the benefits of DG/DM;
- Proving the financial viability of DG/DM;

DATA AND DATA QUALITY BUSINESS CASE BUILDING

The first component of the business case is simply putting a number to the common DQ problems.

- As we've mentioned, impact usually belongs to 1 of 3 areas. Either DQ issues cause financial costs, prevent further revenue, or present regulatory/compliance costs;
- Naturally, the bigger the costs, and the more relatable to the people involved, the easier the business case is to "sell";

Examples include:

- For a company with inconsistent sales reports, selling DG/DM as a way to increase profit with good insights;
- For a company with recent regulatory fines, selling DG/DM as a way to prevent further regulatory sanctions;





DATA AND DATA QUALITY

DATA AND DATA QUALITY BUSINESS CASE BUILDING

When describing the costs, it may be useful to present the cost of “data debt”. That is, the cost of applying “quick fixes” to data now, which will have consequences later.

- It's similar to technical debt, but for data;
- For example, instead of reformatting a DB structure to allow for a new table (new data in different format), instead spreadsheets are created and the data are saved there;
 - It's an easy fix in the short term... but you will pay later;
- Simulating the costs of fixing DQ issues for these types of problems (which inevitably exist in most organisations) can also be added to the costs of not having DG/DM;
 - You can consider both the “one-off” cost + maintenance;
- The costs of all “shadow IT” should be included;

DATA AND DATA QUALITY BUSINESS CASE BUILDING

Then, just like any business case, it's all about making the numbers work. No DG/DM program is without its costs, but at the end of the day, having a select set of people and processes to govern/manage data will cost a lot less than the spend of fixing the problems afterwards instead.

- Costs with people and tools should be included, and the gains from proper DG/DM should more than make up for these;
- It may be the case - especially for smaller initial projects - that the initial financial benefit from DG/DM is not that big (existing, but not that big), but it will increase as the program expands to more projects;





DATA AND DATA QUALITY

DATA AND DATA QUALITY BUSINESS CASE BUILDING

Other important aspects to include in the case can be:

- The risks of a DG/DM program. Although the goal is to succeed, these can fail. It's important to include risks;
- The vision of a successful DG/DM implementation. This helps make it real for executives. "A day in the life";
- The specific projects interacting with DG/DM;

When presenting the business case, it's also important to keep in mind:

- DM can be an IT capability under the CIO, but DG is a business function - specifically, a control/audit function;
- While both DM and DG can prove value as isolated projects, they have to go enterprise-wide soon. Otherwise, they are just isolated success cases with no real impact;

DATA AND DATA QUALITY BUSINESS CASE BUILDING

Besides these, success with a business case presentation comes down to good communication and presenting:

- Never falling into the trap of too much jargon or complex language. Illustrating, in simple terms, what DG/DM is about. Who will do what, and how do you make money;
- Showing clear numbers is important. For example, current costs as \$40k/month, costs of DG/DM as \$10k/month. Anyone can understand that;
- Some (or many) executives may need to be more data-literate before fully understanding the business case, so reinforcing the fundamental principles in private first can help (data as assets, data monetisation, DG as a business function, etc);



DATA AND DATA QUALITY BUSINESS CASE BUILDING EXAMPLES

/01 NOT NEEDED AT FIRST

For small projects, you may not need a business case from the beginning, but you will need at least calculations - with no clear gain, no one will support it.

/02 DG FOR REGULATION

While DM is more focused on fixing data of low quality, DG is many times motivated by regulation, which focuses on privacy, security, data lineage, etc.

/03 CLEAR SEPARATION

If DG and DM are being performed in parallel, it's crucial to separate both responsibilities. The people doing DM should NOT be doing DG as well.

DATA AND DATA QUALITY BUSINESS CASE BUILDING KEY TAKEAWAYS

/01 COSTS AND BENEFITS

Good reactions to a DG/DM business case are very similar to those to any other business case. Show the costs, show the benefits, and suggest the solution.

/02 THREE MAIN COSTS

Remember DQ problems cause problems in 1 of 3 major categories. They either cause actual financial cost, they prevent revenue, or cause compliance fines.

/03 COSTS OF CURRENT ISSUES

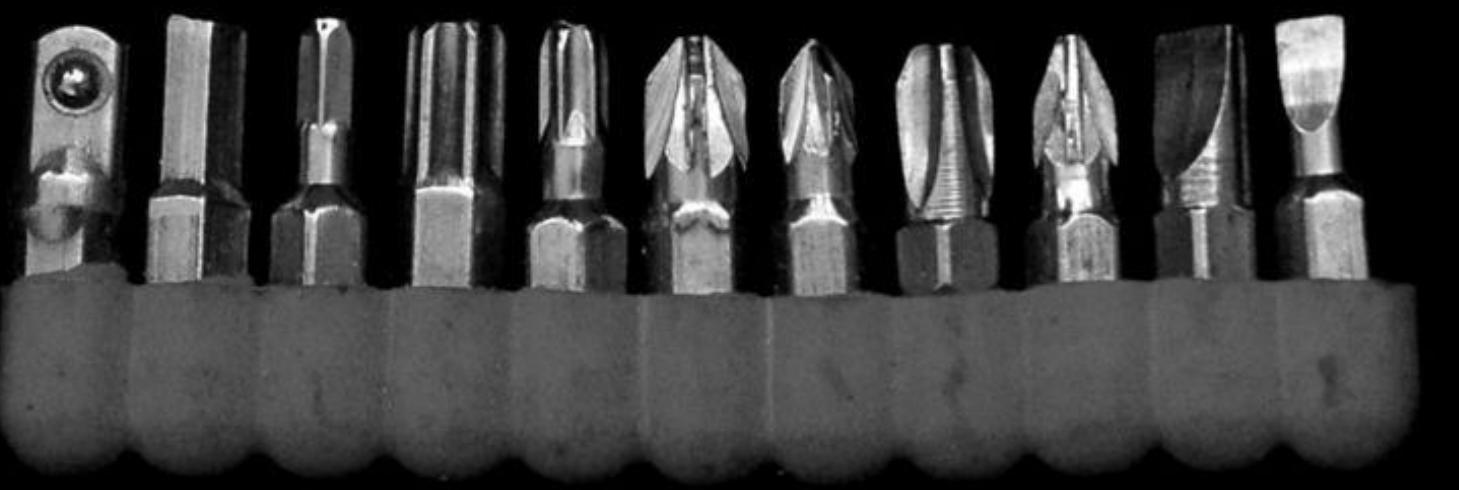
The costs of the current issues should represent high numbers, including the costs of data debt, shadow IT, and the maintenance of all of these data sources/tools.

/04 NUMBERS AND WORDS

A DG/DM business case has a higher chance of succeeding if you make the numbers work first - low costs, high gains - and present it effectively.

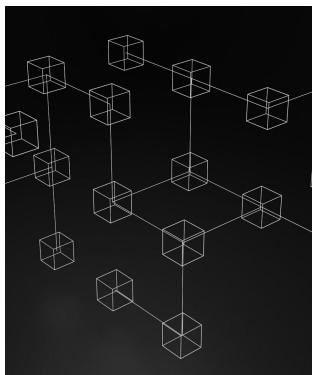
DATA AND DQ

Learning about how data and Data Quality work, as well as the tools, activities, processes used to both assess and improve Data Quality



DATA AND DATA QUALITY

This module was all about **how data and Data Quality work**, as well as how to assess and improve the latter:



THE 4 TYPES OF DATA

Transactional data, master data, reference data and metadata, and what they all mean

DQ PROBLEMS AND IMPACT

The common types of DQ problems, as well as the business impact they have



DATA QUALITY MANAGEMENT

What DQ management encompasses, including processes, data dimensions, actions



DQ TOOLS AND TECHNIQUES

The specific tools used for profiling, parsing, standardising, merging, and more



BUSINESS CASE BUILDING

The usual elements included in a business case to show value for either DG or DM





DATA AND DATA QUALITY

DATA AND DQ CONSOLIDATING

Some questions you can ask yourself to consolidate the knowledge in this module include:

- What are some criteria to define good attributes for identity matching? Does standardising beforehand help?
- Between remediation and monitoring, which is usually a DM function and which is usually a DG one?
- Data profiling can only be used to assess if values fit a format, but not if a format fits the values. True or false?
- Do big data usually mean that data can be remedied earlier or later in the pipeline?
- In which two situations do you usually perform data enhancement?
- Which specific dimension tracks how “fresh” data are?

DATA GOVERNANCE

Learning about how Data Governance works, in terms of both its processes and policies, but also its specific implementation and scaling in an organisation.



DATA GOVERNANCE



DATA LITERACY & CONSIDERATIONS

Covering the basics of data. What are the different data disciplines, what are the essential principles to handle data, what are the sophistication levels of organisations...



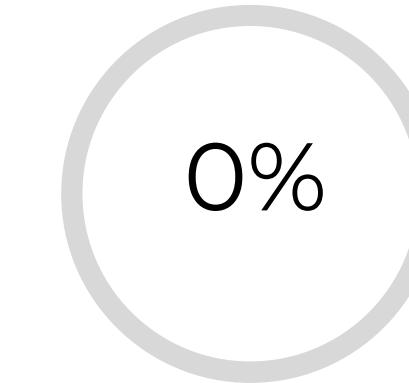
DATA GOVERNANCE

Covering how data governance works, from classifying data to setting policies and other activities, the roles and responsibilities, and the DG implementation process.



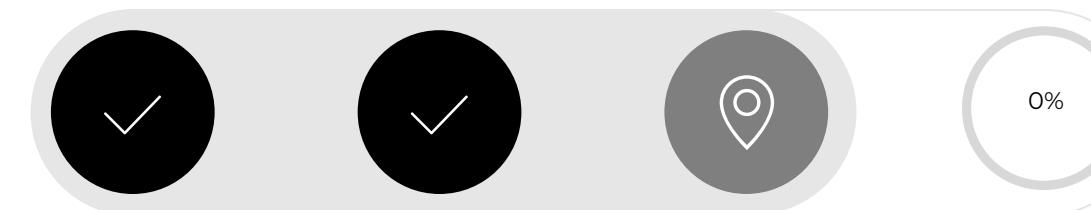
DATA AND DATA QUALITY

Covering specifically what data are and how to improve their quality. Data types, values, structures, and how to improve data quality through profiling and remediating.



DATA SECURITY, PRIVACY, ETHICS

Covering the different types of privacy and security controls that can be applied to data to protect them, as well as how to treat data subjects ethically.





DATA AND DATA QUALITY

DATA GOVERNANCE GOALS

Our major goal in this module is to clarify how Data Governance works. We'll cover:

- The major capabilities enabled by DG (data lineage tracking, data classification, data privacy controls, and many others);
- How to assess an organisation and define the scope of the initial DG program/project;
- How to prioritise data-centric projects that go well with DG;
- How DG affects (and is affected by) company culture;
- How to plan for, roll out, sustain and scale a DG program;
- What is Data Stewardship in specific, and its relation to DG;
- The usual types of data classes and categories in DG;

DATA GOVERNANCE

We will cover **two main topics** in terms of Data Governance:



THE DG FUNCTION

The essentials of what Data Governance is, who does it, and for what goals in specific.



DG IMPLEMENTATION

How to plan, prepare for, roll out and scale DG across an organisation.



THE DG FUNCTION

THE DG FUNCTION INTRODUCTION

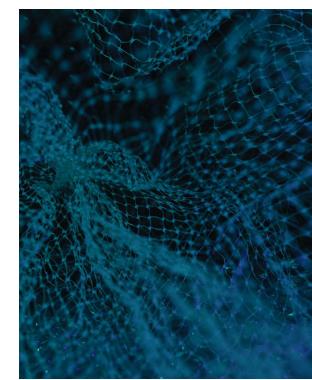
In order to cover how Data Governance (DG) works, it's important to cover the essentials of the function before anything else.

There are a lot of different possible capabilities for DG, including tracking data lineage, setting privacy controls, classifying data, monitoring user access, and many other possible ones, and these are achieved by a variety of people and using a variety of tools.

- It's important to clarify the activities, the tools and the people involved;

THE DG FUNCTION

In order to clarify what Data Governance is in practice, we will cover
five distinct topics:



COMMON FUNCTIONS / CAPABILITIES

The common functions of DG. Data classes, validity rules, InfoSec, lineage, others.



ROLES AND RESPONSIBILITIES

Who does DG. Executives, the Data Council, data owners, data stewards, others.



DATA CLASSIFICATION

The different classes of data (according to 3 different systems), and their consequences.



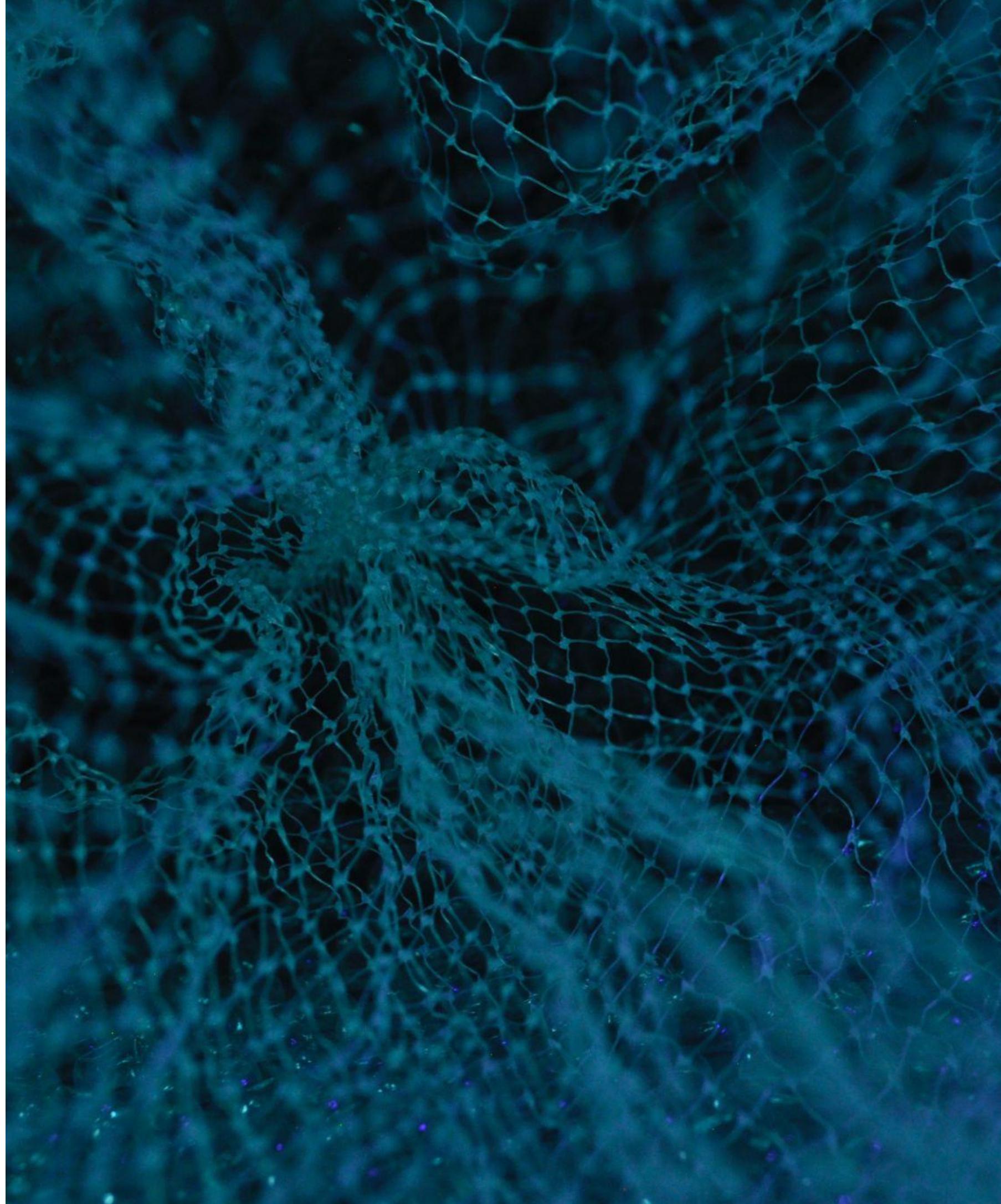
DATA STEWARDSHIP

What is Data Stewardship in specific, and how it bridges DG and reality.



A DAY IN THE LIFE

A day in the life of DG for different roles in an organisation, and with different activities.



THE DG FUNCTION

THE DG FUNCTION COMMON FUNCTIONS/CAPABILITIES

At its core, Data Governance (DG) consists of defining policies and processes for governing data. This can include several possible capabilities, depending on program and priorities:

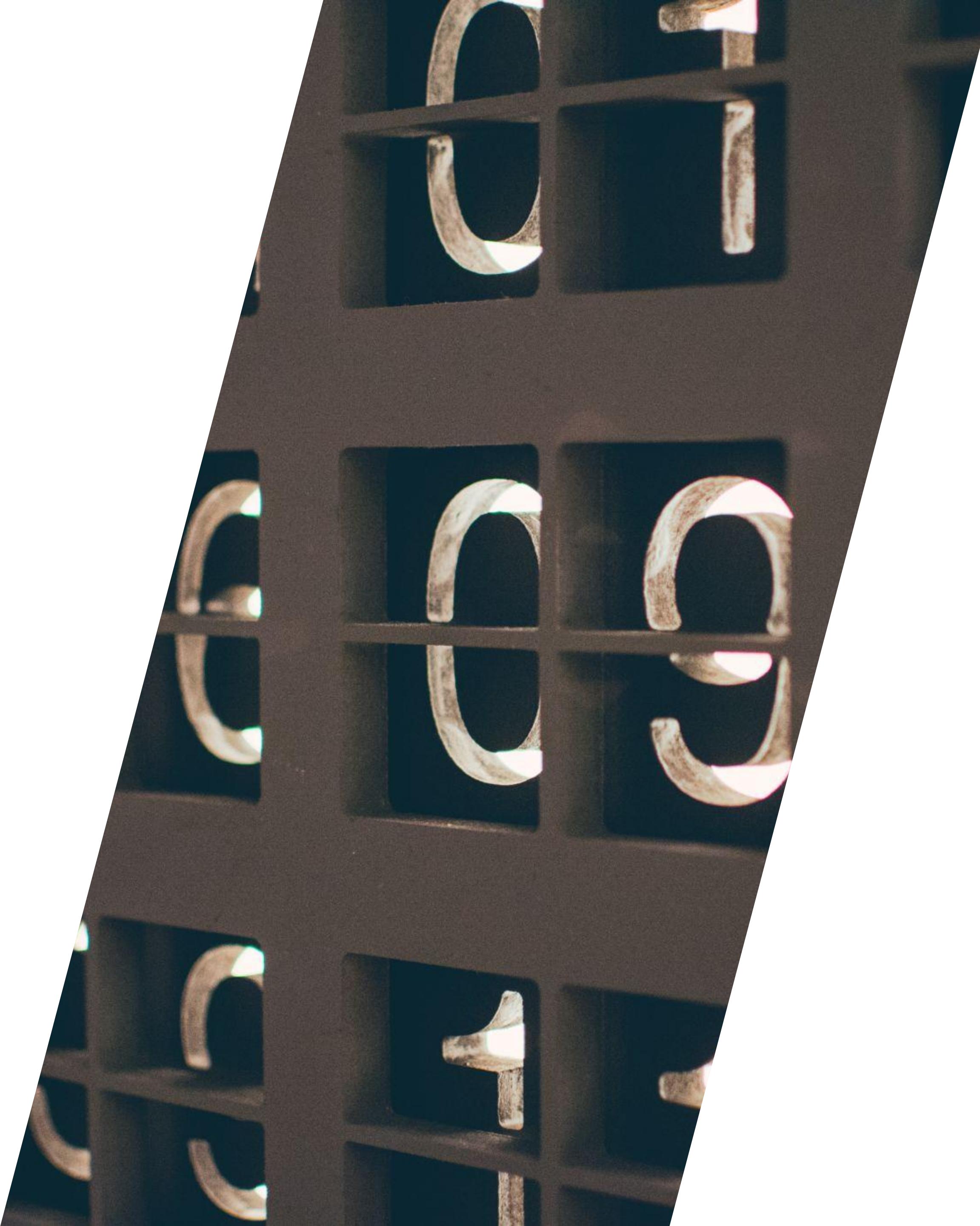
- Defining the different data classes that exist;
- Defining expectations and validity for data;
- Defining expectations and validity for metadata;
- Defining policies for data use and access;
- Defining privacy and security measures;
- Defining technology and tools used;
- Defining required lineage information for data;
- Defining monitoring and reporting for DQ;

These are just some of the most common capabilities. In essence, DG can govern data at all stages of the lifecycle.

THE DG FUNCTION COMMON FUNCTIONS/CAPABILITIES

One of the most important capabilities of DG is to define and catalog the different data classes in the organisation.

- For example, defining the priority of data based on their importance. Low-, medium- and high-priority + critical;
 - Usually, critical data are data to be protected at all costs. Personally Identifiable Information (PII), financial transactions, healthcare information... Usually regulated;
 - Based on the class, data are subject to different controls;
 - All data in the organisation must be catalogued and classified;
 - Additionally, data sources are usually considered “Validated” or “Pending” based on their data quality;
 - Also possible by stage (Raw, Profiled, Cleansed, Vetted);





THE DG FUNCTION

THE DG FUNCTION **COMMON FUNCTIONS/CAPABILITIES**

Another function is to define expectations and validity rules for data. This is intimately related to DQ. These are the rules for data to have high quality and be labeled “Trustworthy”;

- For example, “All rows in the Customer table must have the Address field populated”. This defines the quality that table must comply with;
- These usually stem from the problems that led to DG. For example, if the DG program was created because BI doesn’t have timely sales information, a validity rule for the sales database may be “Information must be updated within 24 hours of a sale being closed”;
- Rules may cover all data dimensions (accuracy, completeness, etc), and usually define sources as “Trusted”;

THE DG FUNCTION

COMMON FUNCTIONS/CAPABILITIES

Just like there must be expectations and requirements for data, you should have them for metadata as well. Metadata are more than just “data about data”. They include both technical and business metadata. Usual components include:

- Data asset name, DB/file type, column/field names, data types, valid value range, meaning of codes, flags, indicators;
- Assigned department, data owner, data steward;
- Data class, processing stage, sensitivity;
- Retention period, applicable regulations and privacy;
- Description of data usage and users (e.g. “Reports for BAs”);
- Associated models or reports + primary consumers;

This information is crucial in order to educate new users who may not be familiar with the data (and formalize it).





THE DG FUNCTION

THE DG FUNCTION COMMON FUNCTIONS/CAPABILITIES

Another common function of DG is to determine who can use the data and what for. This is not just in terms of user Access Control and identification, but also in terms of purpose.

- Usually, there is a component of AC here, and usually based on a combination of role and group.
 - E.g. 2 BAs. Marketing + R&D. Same role. Both would have access to an R&D DB. There's also an "R&D Group".
Access is a combination of "BA" Role + "R&D" Group;
 - This because many regulations (e.g. GDPR) demand that data can only be accessed for certain use cases, and the same role may have multiple use cases;
- Also includes data Usage Agreements. Employees must comply with policies (training, advance notice, NDAs...);

THE DG FUNCTION

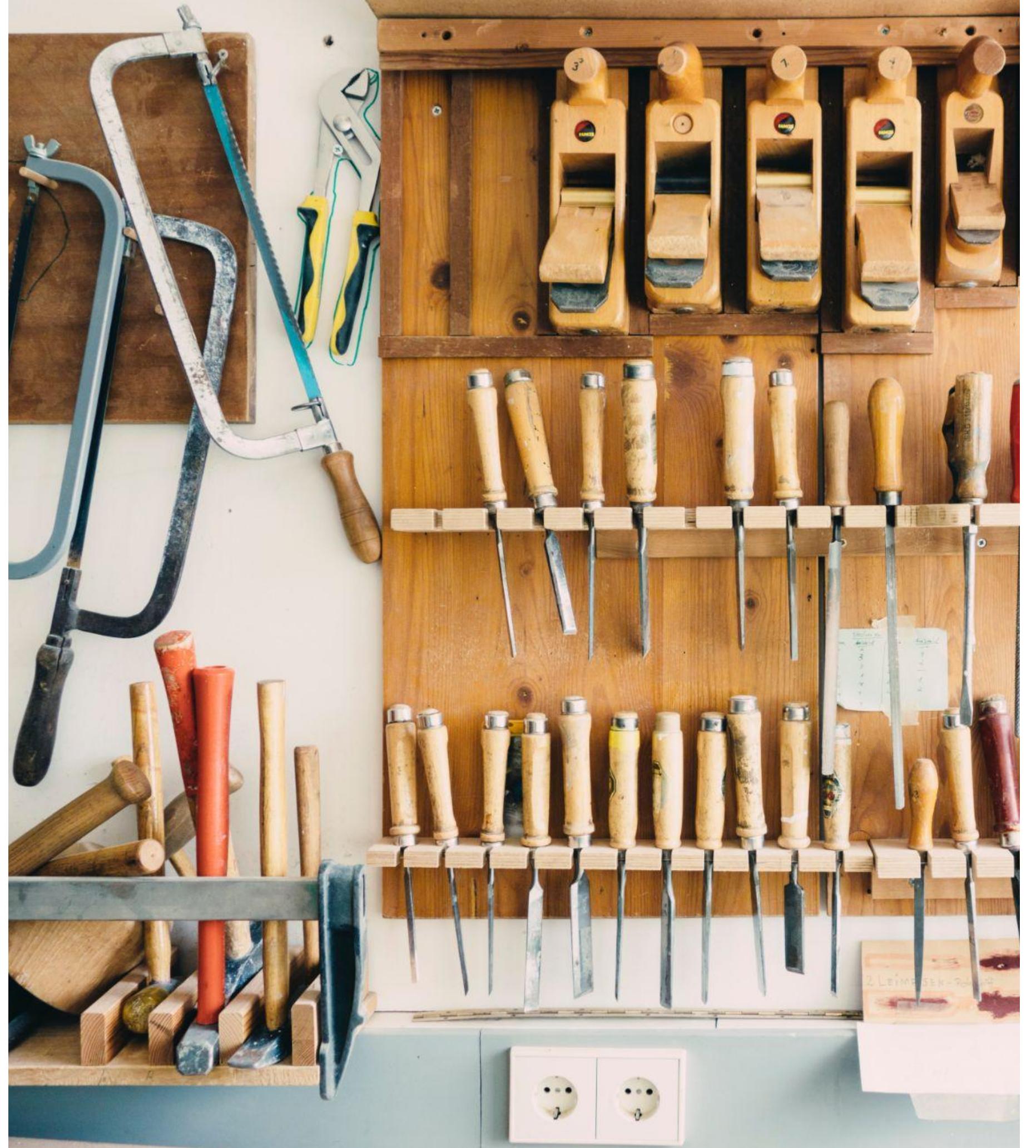
COMMON FUNCTIONS/CAPABILITIES

Another function are privacy and security measures, which have a high degree of overlap with the Information Security area, and consist of keeping information secure and private. it can include:

- AC from the previous point, possibly with MFA, specific times of day, request forms for data access, and more;
- Firewalls, IDS/IPS systems, FIM, audit trail logging, and other systems to alert in case of data leaks/breaches;
- Perimeter security (protecting offices, premises, rooms with guards, CCTV, credentials, others);
- Network security (network segregation, blocking traffic, preventing data exfiltration, IP address auditing, etc);

Can be considered InfoSec requirements for data in specific;





THE DG FUNCTION

THE DG FUNCTION COMMON FUNCTIONS/CAPABILITIES

The technology and tools used are also relevant in a DG program. There is no one solution that is superior to others, it's just important to define which solutions and tools are used.

- Which tools are used for DQ efforts (profiling, annotation, merging, remediation, etc);
- Which tools are used for cataloging/classifying data (such as an Enterprise Dictionary or Metadata Hub);
- Which tools are used for BI/analytics, as well as big data structures (Data Lakes or Data Warehouses, and ETL processes used);
- Which InfoSec tools are used for the different requirements (AC, network security, data encryption, media destruction);

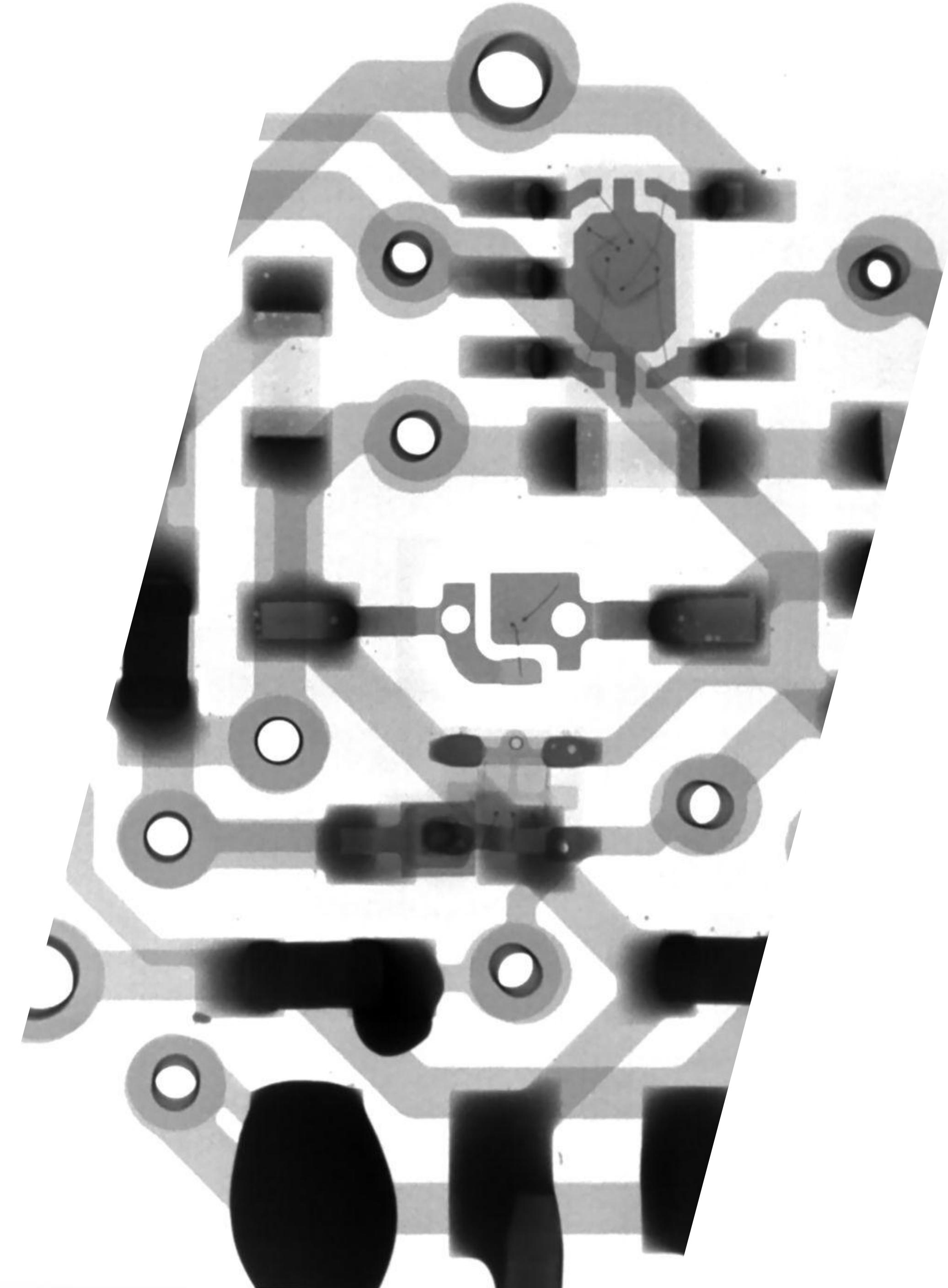
The DG portfolio can be one integrated tool, or “smaller” ones.

THE DG FUNCTION

COMMON FUNCTIONS/CAPABILITIES

Capturing lineage information is another important component of many DG programs. Although this is only a part of the required metadata, it's a crucial part of it.

- Lineage should encompass all stages of the information lifecycle. How was the data obtained? Is the source trusted? What transformations did it undergo? Did we lose information in the process? Did we merge datasets with different qualities?
- Lineage is crucial to be able to trust the data we have. Both to use trusted data, but to label our data as trusted as well;
- This is especially important for regulated information (e.g. transaction information used to decide on a loan), and it may be required to “save snapshots” of data for decisions;





THE DG FUNCTION

THE DG FUNCTION **COMMON FUNCTIONS/CAPABILITIES**

Finally, the monitoring and reporting structure is another function of the DG capability - and it's usually a key one, as showing actual performance relies on it. The DG program must report on:

- What are the current DQ issues, remediation plans, and deadlines?
- What data sources are trusted, and which are pending validation? Which are clearly untrusted?
- Are we in breach of compliance? How many data sources?
- What lineage information do we have for different data sources? Which is required?

This allows both early warning in case of problems, and also demonstration of value of the DG program itself.

THE DG FUNCTION

COMMON FUNCTIONS/CAPABILITIES

EXAMPLES

/01 PHYSICAL/LOGICAL

In terms of privacy and security, it's important to notice that data must be protected not just in logical form, but physical form as well. Paper, HDDs, USBs...

/02 CLASS COMBINATIONS

The classifications mentioned can (and frequently do) combine. So you may have "Trusted" Low/Med/High-Priority and "Pending" Low/Med/High-Priority data.

/03 OWNER DISSEMINATION

While the "initial" DG team is dedicated, as DG expands, each department or team becomes accountable for their own data, with owners and DSs.

THE DG FUNCTION

COMMON FUNCTIONS/CAPABILITIES

KEY TAKEAWAYS

/01 MULTIPLE AREAS

A DG program can contain multiple functions or capabilities depending on scope, and each one is important in its own way.

/03 CLASSIFICATION & PRIORITY

As part of the metadata, usually a classification of the different data types and sources is included, with priority and sensitivity classification, among others.

/03 PRIVACY AND SECURITY

Usually, DG also encompasses privacy and security measures to ensure there are no leaks of data, especially sensitive and critical data.

/02 METADATA + LINEAGE

Usually, having a specific set of metadata information is also required, as is, in specific, tracking the lineage of data to “trust” them, especially for compliance.

/04 ACCESS AND USE

There are usually policies regarding the use of data, by whom, and under what circumstances, with specific roles and use cases for different types of data.

/04 DQ EXPECTATIONS + ISSUES

Usually, setting expectations and rules for the required data quality is included, as is tracking DQ issues and having remediation of these.



THE DG FUNCTION

THE DG FUNCTION ROLES AND RESPONSIBILITIES

In terms of DG, there are several roles involved, and it's important to clarify them. Let's start with the executives:

- The CDO (Chief Data Officer). If it exists in a company, usually the person driving a DG program;
- The CIO (Chief Information Officer). Not to be confused with the CDO. Deals with IT infrastructure and handles DM tasks. Sometimes wrongly assume data is under them;
- Other executives. These may need to “buy in” or provide resources for a successful DG program. They may include the CEO, CIO, CISO, and specific department executives;
- The Data Council. The group of executives that steers DG, and approves/gives feedback on policies and projects;
- Accountable Executives. Responsible for DG in their LoB;

THE DG FUNCTION ROLES AND RESPONSIBILITIES

It's specifically important to clarify the difference between CDO and CIO. Although the names may seem similar, they are very different in nature.

- The short answer is that the CIO is usually associated with DM, and with the IT side, while the CDO is associated with DG, and with the business side;
- If you have specific DM programs, such as MDM or EIM, at the enterprise level, the CIO is usually driving these;
- However, many organisations assume that data governance is the responsibility of the CIO and DM. It isn't;
- The CDO drives DG, across all different departments and teams, and the CIO manages the DM projects and the IT infrastructure that supports and powers these projects;





THE DG FUNCTION

THE DG FUNCTION ROLES AND RESPONSIBILITIES

Once DG spreads to the organisation (and even when it's small), it's frequent to have data owners and data stewards at each department or Line of Business:

- The data owner is the person that is responsible for data in that department/LoB. For example, there will be a data owner for Marketing data, another one for Sales data, etc;
- If these are the executives, then the data owner is the accountable executive (e.g. Sales data may be owned by the CRO);
- The data stewards are the people guiding the implementation of DG in that specific department or team (e.g., Sales DSs help implement policies for Sales data);

THE DG FUNCTION ROLES AND RESPONSIBILITIES

It's important to understand that these roles are not "fixed", and that some people may end up having more than one role. In general, DG is a practice that is usually understaffed and sees people wear more than one "hat":

- For example, the same person may be doing Data Management (DM) and Data Governance (DG), both setting policies for DQ and actually remedying data. This is not recommended, as there must be a clear distinction of DG/DM;
- The data owners of different departments may have different backgrounds. They may have a legal background, an InfoSec background, or others. The only element in common is the responsibility of the role;





THE DG FUNCTION

THE DG FUNCTION ROLES AND RESPONSIBILITIES

Depending on the size and topology of the company, it may also be possible to have data owners and stewards organised by different dimensions:

- By department or team. This is the most frequent model. Sales, Marketing, Finance, HR, etc;
- By geography. For a big, multinational company, especially with different regulations for different countries, there may be a US team, a UK team, a Germany team, and so on;
- By app or database. Especially for companies with monolithic apps or DBs, these may merit their own teams. You may have a gigantic ERP with a data owner for the Customer/Sales area, another for the Products/R&D area, another for the Vendor/Service area, and so on;

THE DG FUNCTION ROLES AND RESPONSIBILITIES EXAMPLES

/01 MULTIPLE COUNCILS

In many organisations, there is only one Data Council, which steers DG from the top. But for especially complex ones, there may be “sub-councils” under it.

/02 HISTORICAL CDO/CIO ROLES

The reason why the CIO is associated with DM and the CDO with DG is because historically there was only the CIO, and only DM. DG came later, and so did the CDO.

/03 OWNERSHIP IS CRUCIAL

We can see by the pattern of “accountable executive”, “data owner” and similar roles that data will only be governed well if there are people accountable.

THE DG FUNCTION ROLES AND RESPONSIBILITIES KEY TAKEAWAYS

/01 MULTIPLE EXECUTIVES

There are many important stakeholders in DG that provide resources and approval, including CEO, CIO, CISO, and other relevant ones.

/02 CDO VS. CIO

Although they seem similar, the CDO and CIO are different. The CIO is related to tech and DM, while the CDO is related to the business and DG.

/03 OWNERS AND STEWARDS

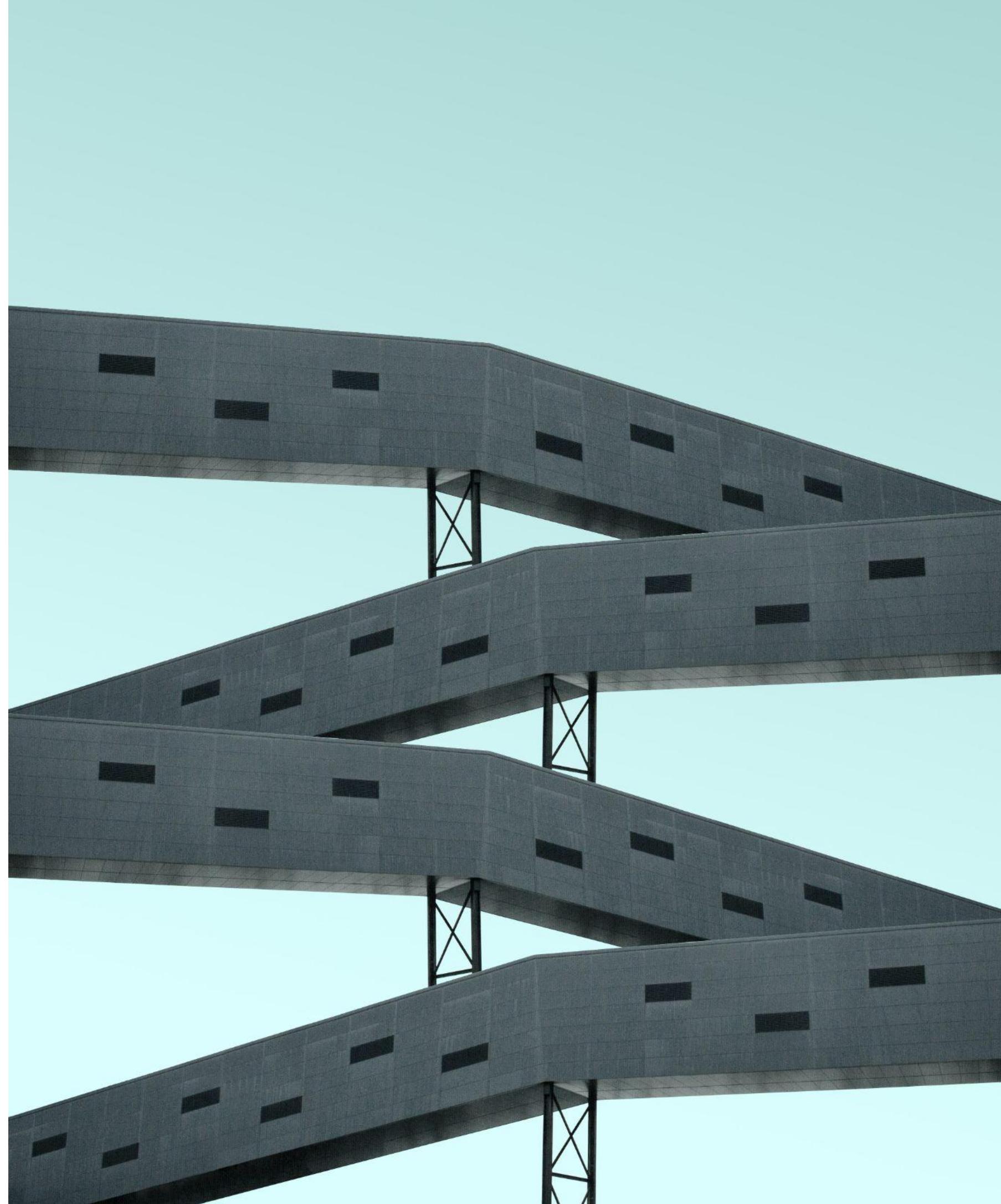
Usually, for a given department or group, there is a data owner, accountable for the quality of the data, and data stewards, guiding implementation of DG.

/02 MULTIPLE HATS

It's very frequent, in a DG effort, to see people wear multiple hats. The data owner may also be a steward, they may be an executive or not, and DG/DM may mix.

/05 DIFFERENT TAXONOMIES

DG teams don't need to be organised by team or department. Depending on company characteristics, it may be by geography, data area, or other criteria.



THE DG FUNCTION

THE DG FUNCTION DATA CLASSIFICATION

One of the core activities of Data Governance is cataloguing and classifying data in the organisation (in the scope of the DG program, but ideally, to cover the whole enterprise sooner or later), which is usually done by Data Stewards.

There are many ways to classify data, but there are usually 3 systems of classification:

- By priority/business impact (prioritising the data that are the most important, from very high to very low priority);
- By data sensitivity (usually related to privacy requirements - PII, healthcare information, financial information, etc);
- By processing stage (raw data, cleaned data, analysed data, and possibly other stages);

THE DG FUNCTION DATA CLASSIFICATION

In terms of priority, data are usually classified as being high-priority, medium-priority and low-priority, usually with the addition of a “critical” category. In detail:

- Critical data are data that are crucial to decision making and have devastating consequences if leaked. Think of individual healthcare information, financial information, etc. These are usually regulated;
- High priority data are important to decision making. Severe consequences if leaked, but not extreme. R&D plans, employee information, etc;
- Medium priority data are somewhat important to decision making. Sales aggregates, strategy documents, etc;
- Low priority data are usually day-to-day operational data;





THE DG FUNCTION

THE DG FUNCTION DATA CLASSIFICATION

The sensitivity classification classifies data based on how sensitive they are - how protected their privacy should be.

There are no “standard” classifications, but usual ones include:

- Classified. Three layers of sensitivity, frequently used in military/intelligence/governmental organisations;
 - Least to most sensitive: Confidential, secret, top secret;
- Privileged. Data that are not public and must be protected, with legal consequences if not (e.g. attorneys, psychiatrists);
- Open or public. Data that can be shared freely. Website, social media, etc;
- Specific sensitivity classifications may exist for regulations:
 - CHD (Cardholder Data). Credit card data for PCI-DSS;
 - HIPAA. Health information (e.g. patient data);

THE DG FUNCTION DATA CLASSIFICATION

And finally, another classification system that may be used regards the processing stage of the data. This is particularly useful in terms of defining permissions and uses for the data:

- Raw data. These are usually data that have been received from a data source, but not analysed/validated yet. Limited access, and no data products can be generated by these;
- Cleaned or parsed data. These are usually data that have been processed (usually de-identifying sensitive data), as well as having some metadata. More users can access these, and data products may or not be generated;
- Trusted/completed data. Completely documented in terms of metadata (including other classes). Can be used by many users, and data products can be generated;





THE DG FUNCTION

THE DG FUNCTION DATA CLASSIFICATION

In order to maximise documentation and governance of data, it's recommended the 3 classifications are used, although some organisations may not use all three. Also, naturally, the classifications are related in some ways:

- For example, in a heavily regulated company dealing with credit cards, credit card information can be considered critical in terms of priority, have a specific high-sensitivity classification (e.g. PCI-DSS CHD) and need to be trusted/completed before being used for any report;
- On the other hand, low-priority data, for example daily work hour totals, may also not be sensitive information, and may have lower priority in terms of cleaning and trusting those data for use in reports;

THE DG FUNCTION DATA CLASSIFICATION

In practice, these data classes and categories will be defined by the Data Stewards of the corresponding department or line of business in the organisation's metadata hub, and ideally classes will exist for all three systems. In practice:

- The priority categories allow you to prioritise processing of the data. For example, when using big data for analytics, it's more likely you won't have the availability to process all, and this provides a heuristic;
- The sensitivity categories allow you to define security and privacy controls. The more sensitive, the stricter the measures;
- The processing stage classes allow you to monitor how much data are governed, ready for reports, and monetised;



THE DG FUNCTION DATA CLASSIFICATION EXAMPLES

/01 CONTROLS BY CATEGORY

Data classification is very useful for defining security and privacy controls, as they may vary. Some data may require locked rooms, servers, strict auth and more.

/02 COMPLETENESS REQUIRED

In terms of processing stage, in many cases the processing must be “complete” for the data to be used. This prevents using noisy data in reports/others.

/03 NOT THE ONLY FACTOR

While classification affects the prioritisation of data processing, other factors may come in, such as the volume of the data, and whether they can be modified

THE DG FUNCTION DATA CLASSIFICATION KEY TAKEAWAYS

/01 3 CLASSIFICATION SYSTEMS

Data are usually classified by Data Stewards, in the metadata hub, as belonging to specific classes or categories. There are usually 3 distinct systems.

/03 BY SENSITIVITY

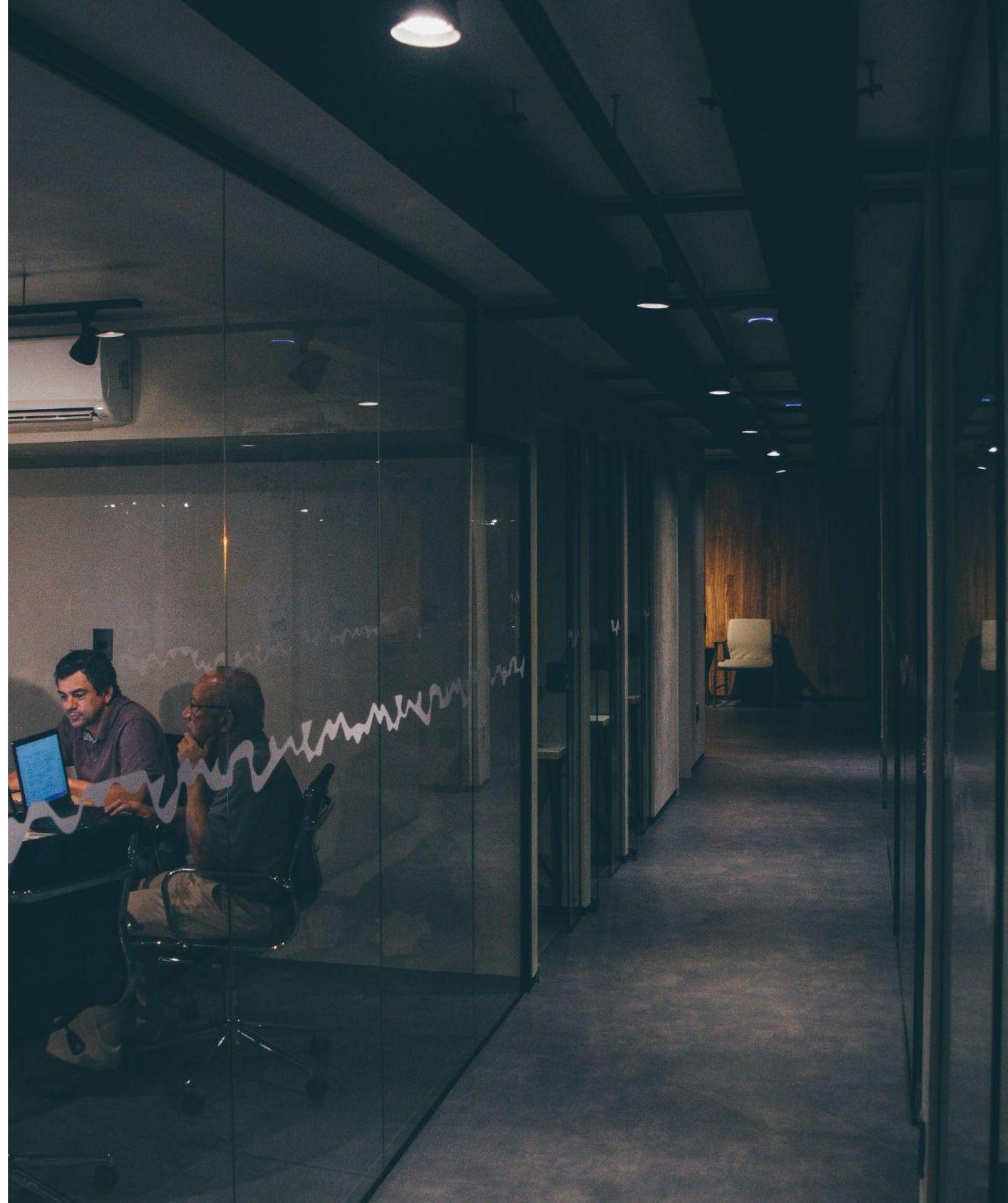
Classifying data by sensitivity is analogous, but it's about the privacy requirements of the data in specific, usually with regulation and compliance attached.

/02 BY PRIORITY

Classifying data by priority is all about the impact that they have on decision making to the business, and how serious a leak of these data would be.

/04 BY PROCESSING STAGE

Classifying data by the processing stage is all about how treated and cleaned they are, and how "safe" they are to consume (by both users and data products).



THE DG FUNCTION

THE DG FUNCTION DATA STEWARDSHIP

Although Data Governance (DG) and Data Stewardship (DS) are many times confused, they are actually two halves of the same whole. DG sets the policies, and DS guides their implementation.

Among other tasks, Data Stewards take care of:

- Classifying and categorising data, including critical data;
- Entering and tracking metadata in the metadata hub (reference data codes, business terms, data users, data purposes, business rules for DQ expectations, etc);
- Creating business rules for the profiling that DM will do, defining acceptable values and validation criteria;
- Defining acceptable data uses and by which users;

THE DG FUNCTION DATA STEWARDSHIP

We can split common DS responsibilities and activities into 5 major areas:

- Data Quality (data validity rules, data remediation activities and timelines);
- Master Data (defining the rules for matching data, creating rules, ensuring data uniqueness);
- Reference Data (defining codes and enumerations, managing the mappings with data products);
- Metadata (both business and technical - defining business terms, associating master data with them, others);
- Security and Privacy (defining sensitive data classes, defining acceptable uses and users);





THE DG FUNCTION

THE DG FUNCTION DATA STEWARDSHIP

In specific, Data Quality activities have to do with ensuring that data are of high quality. It's about defining, for the different data sources (regardless of format), what is expected in terms of completeness and the values contained, and remediating data if those criteria are not met.

- Also involves tracking and reporting on DQ issues;

In terms of Master Data, activities revolve around consolidating a unique source of data for key concepts (Customer, Product, etc). Activities revolve around defining the matching activities for the different data sources, which allow consolidation of these data, ensuring there are no duplicates, and ensuring Master Data quality.

THE DG FUNCTION DATA STEWARDSHIP

Reference data activities have to do with ensuring the quality of data which are referenced by the organisation. Reference data are especially important since one wrong definition propagates to all data sources using it. Activities revolve around validating the codes used, and the mappings to different applications and data products.

Metadata activities have to do with entering information in the metadata hub, especially in terms of business definitions and meanings (e.g., what is a “Sale”?), because these derive the requirements for all data related to them. It may also involve the actual workflows related to approvals and contribution of stakeholders to those definitions.





THE DG FUNCTION

THE DG FUNCTION DATA STEWARDSHIP

In terms of Security and Privacy, activities usually have to do with identifying which data are the most sensitive (such as PII, healthcare information, etc), and defining acceptable uses for them - and the acceptable users. It may also have an overlap with InfoSec policies for user AC and controls themselves.

Naturally, different organisations have different definitions of Data Stewardship, and a professional may end up not performing activities in all of these areas.

- Furthermore, ideally the DG function is divided by LoB / department, so a Data Steward may work only within that department (e.g. only for Marketing data in specific);

THE DG FUNCTION DATA STEWARDSHIP

Particularly, in some organisations the distinction is made between business data stewards and technical data stewards.

There may also be operational data stewards. If this occurs:

- Business data stewards are the ones that guide the “business” side of DG. Business definitions, workflows, communication with involved stakeholders, etc;
- Technical data stewards are very similar to data engineers or IT experts, but in a DG framework. They create/update DBs and schemas, assign/remove users and permissions, and perform other technical tasks;
- Operational data stewards deal with the day-to-day DQ issues and operations. They may sit down with data users to collect feedback and/or with DM to monitor tasks done;



THE DG FUNCTION DATA STEWARDSHIP EXAMPLES

/01 MAKING THE BRIDGE

Data Stewards usually connect DM and DG practices. They can cascade DG policies into specific rules, and they can take specific DM feedback and generalize it.

/02 PARALLEL TO INFOSEC

In Information Security, there are “acceptable uses” for devices (no piracy, no social media, etc). For data, the equivalent exists (no marketing use, restricted AC, etc).

/03 DQ ISSUE TRACKING

While DM is who actually “fixes” DQ issues (profiling, standardising, merging, etc), Data Stewards are usually the ones who track issues and their resolution.

THE DG FUNCTION DATA STEWARDSHIP KEY TAKEAWAYS

/01 IMPLEMENTING DG

Data Stewardship can be considered implementing Data Governance. While DG sets the policies and principles, DS guides their implementation.

/02 DQ / MASTER DATA

Data Quality deals with expectations and fixes for the quality of data in different sources. Master Data activities help consolidate a single source of data.

/03 REFERENCE / METADATA

Reference data activities safeguard the data referenced by the whole organisation. Metadata manages business definitions that inform all data.

/04 SECURITY AND PRIVACY

Security and privacy activities deal with the protection of data, especially sensitive and/or highly-regulated data.

/05 DIFFERENT MODELS EXIST

Data Stewards do not necessarily take care of activities in all 5 areas, depending on the organisation. They may also be assigned to specific teams/departments.



THE DG FUNCTION

THE DG FUNCTION A DAY IN THE LIFE

In order to better understand how DG works - and to be able to paint that picture to an executive you're trying to pitch - it's important to be able to explain what happens, in terms of DG, in an "usual" day in the life.

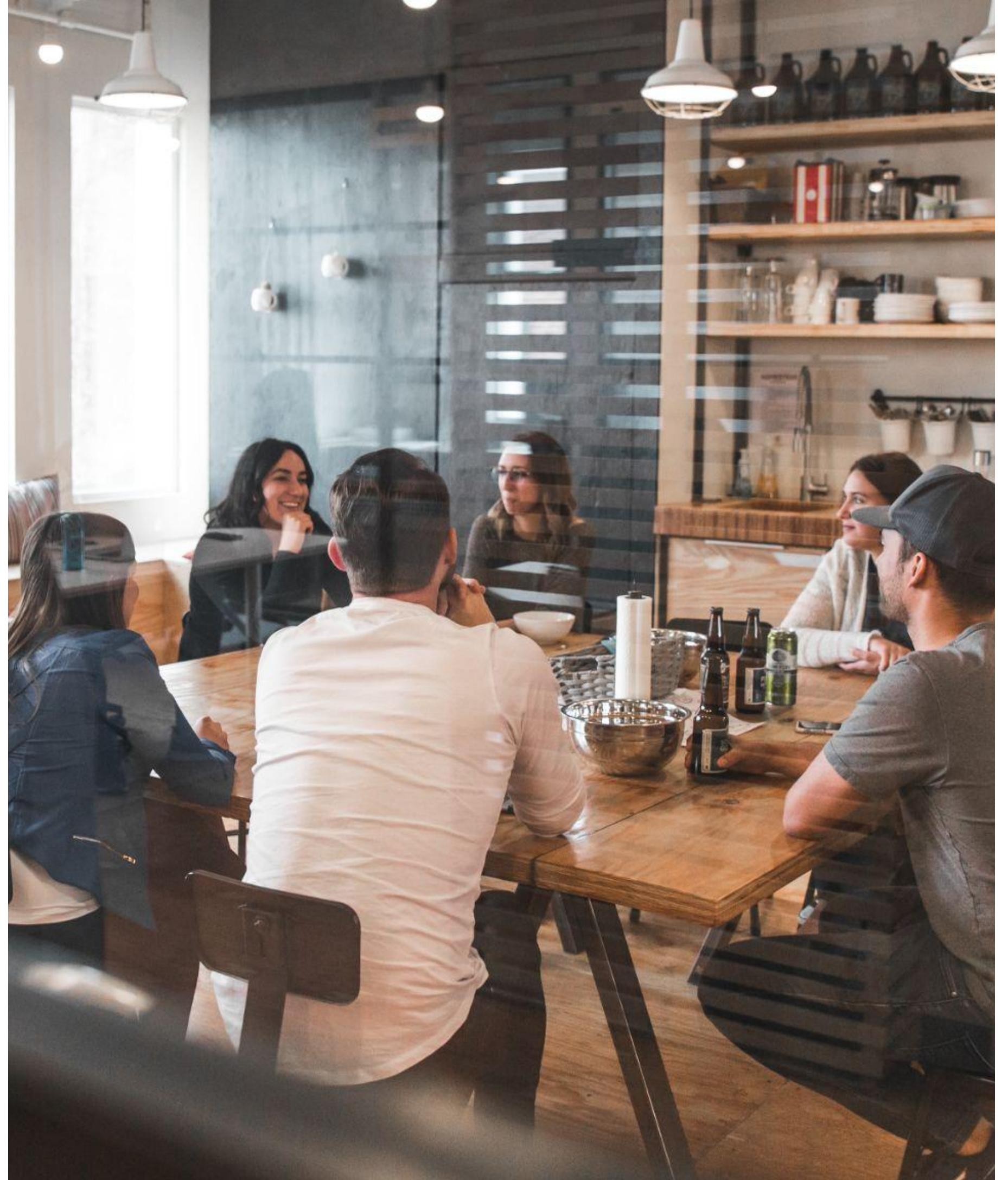
Let's start with the data owner. Let's say they are the CMO, responsible for data in the Marketing department.

- They receive their reports from Data Stewards, who indicate there may be occasional Data Quality (DQ) issues;
 - For example, issues with data regarding a prospect list from a specific event, which has the prospect address in the wrong format;
 - They provide a remediation suggestion + deadline;

THE DG FUNCTION A DAY IN THE LIFE

The Marketing executive receives this report and may review the expected Data Quality rules for this prospect list, since an important report must be generated from it.

- The rules are: “Every Prospect must have a full address across 3 DB fields. An address text field, a house number field, and a zip code number sequence field with a specific length”;
- The data steward report suggests that the data from this prospect list are 80% complete, but the other 20% are lacking a valid address, or it’s in the wrong format;
- The Marketing executive needs this DQ issue fixed to consider this a “Trusted” data source and generate reports;



THE DG FUNCTION

THE DG FUNCTION A DAY IN THE LIFE

In order to remedy these data, the Data Steward (DS) coordinates with the Data Management (DM) team, who creates a project to parse and standardise addresses, and for prospects with missing addresses, they request business analysts present in the event to get in touch with the person and confirm the address.

After this project is complete, the Data Stewards mark the source as being “Trusted”, with 95% completeness, and a report on prospects by job title and company location is generated by the Marketing Department.

- The Data Stewards oversee the report generation process, as well as tracking data lineage information along the way;

THE DG FUNCTION A DAY IN THE LIFE

The same Data Stewards (DS) may do a random quality check on the report data, specifically a Consistency one (matching if the same data are present in the raw DB versus the report), and they find out that some information has been transformed - that is, the total prospects by country are not the same in the DB and the report.

- Maybe in the DB, we have 660 US prospects, and in the report we have 620;

The DSs coordinate with DM, who clarify that a truncation was made of some duplicate prospects, which were not detected - therefore, the report is the correct source.

- DS informs the data owner (the CMO) of this, who suggests a new policy to filter duplicates at the data capture stage;





THE DG FUNCTION

THE DG FUNCTION A DAY IN THE LIFE

This new policy allows the Marketing department to eliminate duplicate prospects in future situations.

- In order to get this approved, however, the CMO needs the approval of the Data Council, which in this case may include the CEO, CIO, CISO and Legal Counsel;

In their monthly meeting, where each data owner presents a list of DQ issues prioritised by severity, and with remediation roadmaps and deadlines, the CMO brings up the new policy to prevent duplicate data at the creation stage.

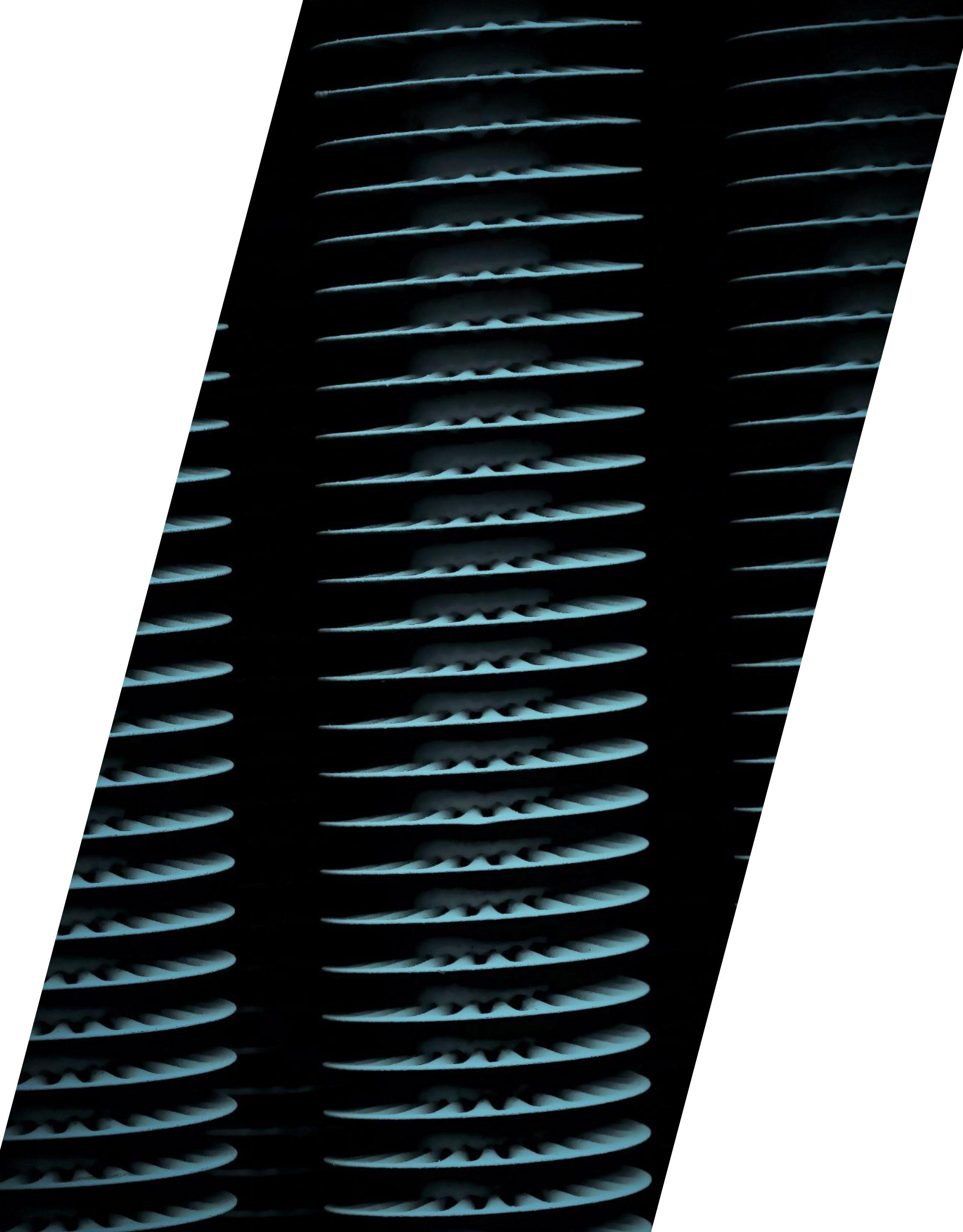
- This is validated by the Council, and the CMO coordinates with the CIO, who tells the DM team to reconfigure the automated DQ monitoring tool to validate the data;

THE DG FUNCTION A DAY IN THE LIFE

This change is made, and the Marketing Data Stewards oversee the rules in action for future data creation, when new events occur, concluding the process works well and DQ has been increased.

- The CMO can, in fact, take the performance metrics, which indicate that prospect data from events have increased in DQ, on average, from 80% to 87% at the creation stage, through this new policy, and uses it show DG value, tying it to sales obtained from prospects with accurate data;

Later that month, Legal Counsel informs, at the monthly Data Council meeting, there is a new regulation affecting retention times for prospect and customer data.





THE DG FUNCTION

THE DG FUNCTION A DAY IN THE LIFE

Each data owner, including the CMO, informs their Data Stewards of the new policy, requesting them to document it in the enterprise metadata hub, and ask them to oversee data disposal operations to reflect this.

- In the Marketing department, the Data Stewards may even find out that the required retention dates are not included in the datasets, so they suggest a policy to include that information in these, which is validated by the CMO and later approved by the Data Council;

This new information makes the job easier for the DM team, who can, when purging a data set, simply look at the metadata, showing the retention deadline, performing the deletion with more accuracy and trust.

THE DG FUNCTION A DAY IN THE LIFE EXAMPLES

/01 DATA STEWARDS BRIDGING

Once again, we see Data Stewardship as the “bridge” between DM and DG, here in practice. They help implement DG, and generalize feedback from DM.

/02 DATA COUNCIL DECISIONS

While many aspects, such as regulation and technology, can affect data policies, the Data Council is who makes decisions that cascade down from there.

/03 ATTENTIVE DATA OWNERS

In this example, we also noticed how the data owner (the CMO in this case) must pay attention to DQ issue resolution, to DG implementation, and a lot more.

THE DG FUNCTION

A DAY IN THE LIFE

KEY TAKEAWAYS

/01 DQ ISSUE MONITORING

In everyday operations, a significant part of the time for a data owner will be monitoring DQ problems. What are they, how are they fixed, and by when?

/02 DATA STEWARD INSIGHTS

DSs play an important role here, by mediating between the data owner, which directs policies and processes, and providing feedback from issues found.

/03 UP AND DOWN THE SCALE

Feedback collected bottom-up by DSs goes up the data owner to the other executives and Data Council, and the decision is reflected top-down again.

/04 ANY STAGE OF THE LC

DG issues (and changes) may occur at any stage of the data life cycle. Creation, storage, usage, disposal, or any combination of these.



DG IMPLEMENTATION

DG IMPLEMENTATION INTRODUCTION

Implementing a Data Governance project (or program) has several stages. It's important to be aware of how to select the people, tools, and projects to support, as well as how to define what will happen in practice.

Along the way, several other factors become important, such as defining metrics to track performance, obtaining executive support at various levels, or defining the federation level of DG across the organisation.

DG IMPLEMENTATION

We'll clarify the distinct phases of DG implementation
with **five main topics**:



PLANNING

Defining the scope of DG,
architecting operations and
defining processes.



ROLL-OUT

Implementing DG, selecting
projects to support and laying out
operations.



ITERATION

Maintaining and supporting DG,
ensuring behavioral change
management and support.



MONITORING

Defining metrics, at various levels,
to track the performance and
efficiency of DG.



CULTURAL ASPECTS

Which elements of company
culture affect (or are affected by)
DG, and how.



PLANNING

PLANNING INTRODUCTION

The Planning stage is very important in DG, because it's when we select the scope of our project or program, as well as who is involved and how we will tackle the actual governance of data.

At this stage, there are several things that need to be done, from assessing the state of the organisation, to securing executive support, defining the scope of the program, and several other factors.

PLANNING

To clarify the Planning stage of DG, we will cover
three main topics:



ASSESSMENT AND SCOPE

Assessing the organisation and defining the scope of the DG program.



ENGAGEMENT AND BUY-IN

Securing engagement from the different executives involved in the DG program.



ARCHITECTURE AND FRAMEWORK

Defining the processes and operations for the DG program in practice.



PLANNING

PLANNING ASSESSMENT AND SCOPE

Two important considerations when planning a DG program are performing an assessment of the current state of the organisation, as well as defining the scope of the DG program.

- Assessing the organisation is important to know where it currently is. Frequently, but not necessarily, some kind of data maturity assessment is performed;
- The goal is to know where the organisation is, in terms of managing data, which informs on where you can take it;
- Change management assessments can also be valuable;
- After assessing the organisation, defining the scope of the DG program is also valuable. Which people, processes, policies, departments, projects will be involved;
- DG must show value, but not take on too much at once;

PLANNING ASSESSMENT AND SCOPE

In terms of assessments, the goal is to understand where the organisation is, as this informs what you can actually change:

- An organisation that already manages data well, but doesn't have automated processes will demand a very different DG program from one that doesn't even know which data it has in the first place:
 - Frameworks such as John Ladley's IMM (Information Management Maturity) consolidate this in quick surveys;
- Other assessments, such as change capacity assessments, can be useful for a different reason. They don't necessarily have to do with data themselves, but they measure how nimble the organisation can be in terms of change;
 - Remember, DG at its core requires behavior change;





PLANNING

PLANNING ASSESSMENT AND SCOPE

It's important to note that, while different assessments serve as indicators of how sophisticated the organisation is in terms of managing data, they are not deterministic.

- For example, an organisation that lacks sophistication on all fronts, with low-quality data (that may even not be inventoried) is not necessarily one who can't change. It may embrace DG;
 - But naturally, an indicator is still an indicator;
 - This is also why other factors, such as the company's openness to change, and the culture count as well;
- For smaller DG programs, assessments may not even be necessary at the beginning;
- But they will be at some point, when expanding;

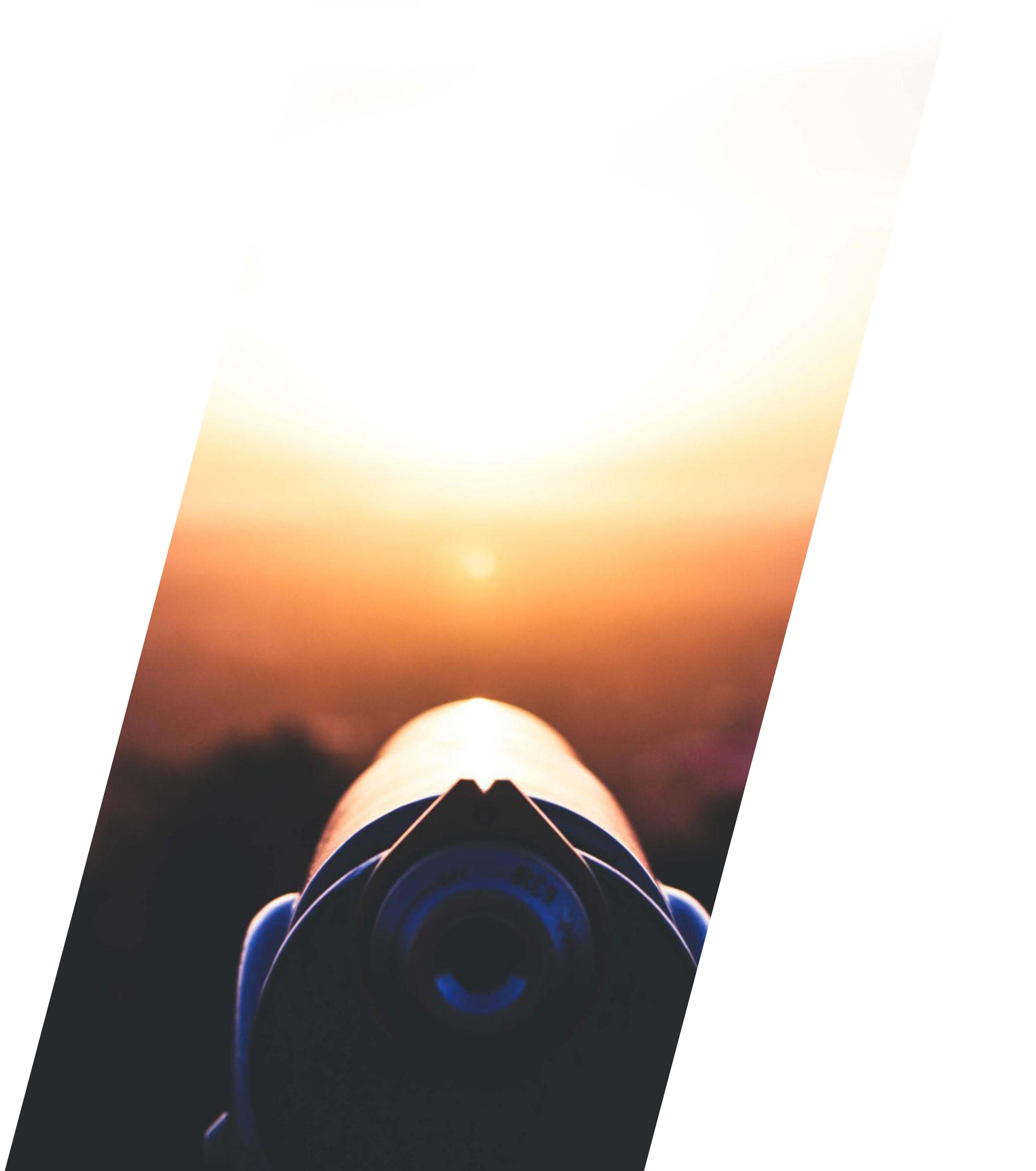
PLANNING ASSESSMENT AND SCOPE

Then, defining the scope of the DG program is also important.

There will likely be a team of personnel from various teams in various roles, and it's important to define structure and goals.

Elements include:

- Which projects will DG support. DG usually “piggybacks” on existing projects and provides controls/policies on data;
 - It's important to select the right projects to show value;
- Who are the people involved. Both in operational terms (the data owners, data stewards, etc) but also the sponsor and other executives involved (including a possible Data Governance Council to steer the program);
- The technology and tools used. DQ issue tracking, metadata management, administrative, and more;



PLANNING ASSESSMENT AND SCOPE EXAMPLES

/01 PIGGYBACKING IS EASY

Since the goal of DG is to show value fast and expand from there, taking existing DM projects and “coupling” DG is a great way to some quick results, if possible.

/02 MANY POSSIBLE OBSTACLES

An organization may have low information management maturity or low change capacity due to multiple factors. Culture, incentive structures, tech...

/03 TOOLS COME LAST

First, the processes and actions that DG will accomplish must be defined. Tools come last. Being in love with the tools wastes resources with no results.

PLANNING ASSESSMENT AND SCOPE KEY TAKEAWAYS

/01 ASSESSING AND SCOPING

These two activities are important when planning a DG program. One defines where the organisation is, the other defines what it needs to do.

/03 JUST AN INDICATOR

Assessments are just indicators of the organisational state. They don't define the possible outcomes. They may also not be necessary at the beginning stages.

/02 ASSESS DATA + CHANGE

Relevant assessments include two important categories: assessing the organisation's sophistication in managing data, and how easily it adopts change.

/04 PEOPLE, PROJECTS, TOOLS

In terms of the scope of DG, it's important to define who will be involved, what projects will be supported, and which tools will be used, to know what to do.



PLANNING

PLANNING ENGAGEMENT AND BUY-IN

Engaging with executives in the organisation is very important to get a DG program off the ground (and to later sustain and expand it). This is due to many reasons, including but not limited to:

- Making them data-literate. A precursor to the others. The executive must see data as assets, possibly monetised;
- Securing commitment/resources. For executives owning the projects that DG will support, it's critical that these give DG authority, so the team will support and accept changes;
- Making benefits understood. They must understand the real bottom line of good DG, and what they benefit from it;
- Fighting resistance. Engaging and communicating with executives allows you to counter their doubts/questions;

PLANNING ENGAGEMENT AND BUY-IN

Specifically, one important hurdle to DG acceptance by other executives will be effort. Specifically, in terms of changing behavior. You may frequently hear “this is too much effort”, or even “we don’t have the time for this”. Several tactics can help:

- Having a smaller scope. The less DG tries to do in one go, the more reasonable the effort will be, and the less effort it appears to be;
- State it’s “the usual”, in a different way. Many DG operations are probably already done, with the same effort - just not well done;
- Leveraging your DG champion. The executive driving DG must not be afraid to request favors or spend political capital to overcome resistance from these executives;





PLANNING

PLANNING ENGAGEMENT AND BUY-IN

Additionally, it's important to clear the misconception of data as IT in specific. Many executives see data as a part of IT, and the Chief Data Officer as a role of the Chief Information Officer, which are very different at the end of the day. It's important to clarify that:

- Data is a business function, and the goal of DG is to “disappear” into the organisation, with each department owning their data and being accountable for them;
 - DG is just like hiring, budgeting, training, or others;
- Data Management (DM), usually done by IT, is not the same as Data Governance (DG), which is a business capability;
 - So IT can work together with DG, but it does not own it;

PLANNING ENGAGEMENT AND BUY-IN

It's also important to gauge the reactions of different executives, as many of them may not welcome the changes brought by DG (especially when we're talking about bringing accountability to the executives who don't manage data well).

- The only thing to do here is to persuade them to accept the changes just like you would persuade an executive of any other thing you need from them. The sponsor may, again, ask for a favor or spend political capital;

Ironically, while everyone thinks that IT owns the data, IT executives are many times the ones most resisting DG, and especially because it changes how they do business:

- Many times, DG controls involve changing apps, DB schemas, imposing additional controls, and so on;



PLANNING ENGAGEMENT AND BUY-IN EXAMPLES

/01 A LOVE-HATE RELATIONSHIP

Since IT is the department most affected by DG, despite being the one who most understands the value of it, it's many times who raises most resistance.

/02 CHAMPIONS MATTER

Having a DG champion is important to break through resistance. Presenting well and having results does help, but sometimes political capital is needed.

/03 DATA AS BUSINESS

A very important, but very elusive distinction to impress upon executives is data as business, not IT. Every team must do it - like hiring or budgeting.

PLANNING ENGAGEMENT AND BUY-IN KEY TAKEAWAYS

/01 UNDERSTAND + SUPPORT

Engaging with different executives is crucial so that they both understand the advantages of DG, and so they support the program in its execution.

/02 DEALING WITH EFFORT

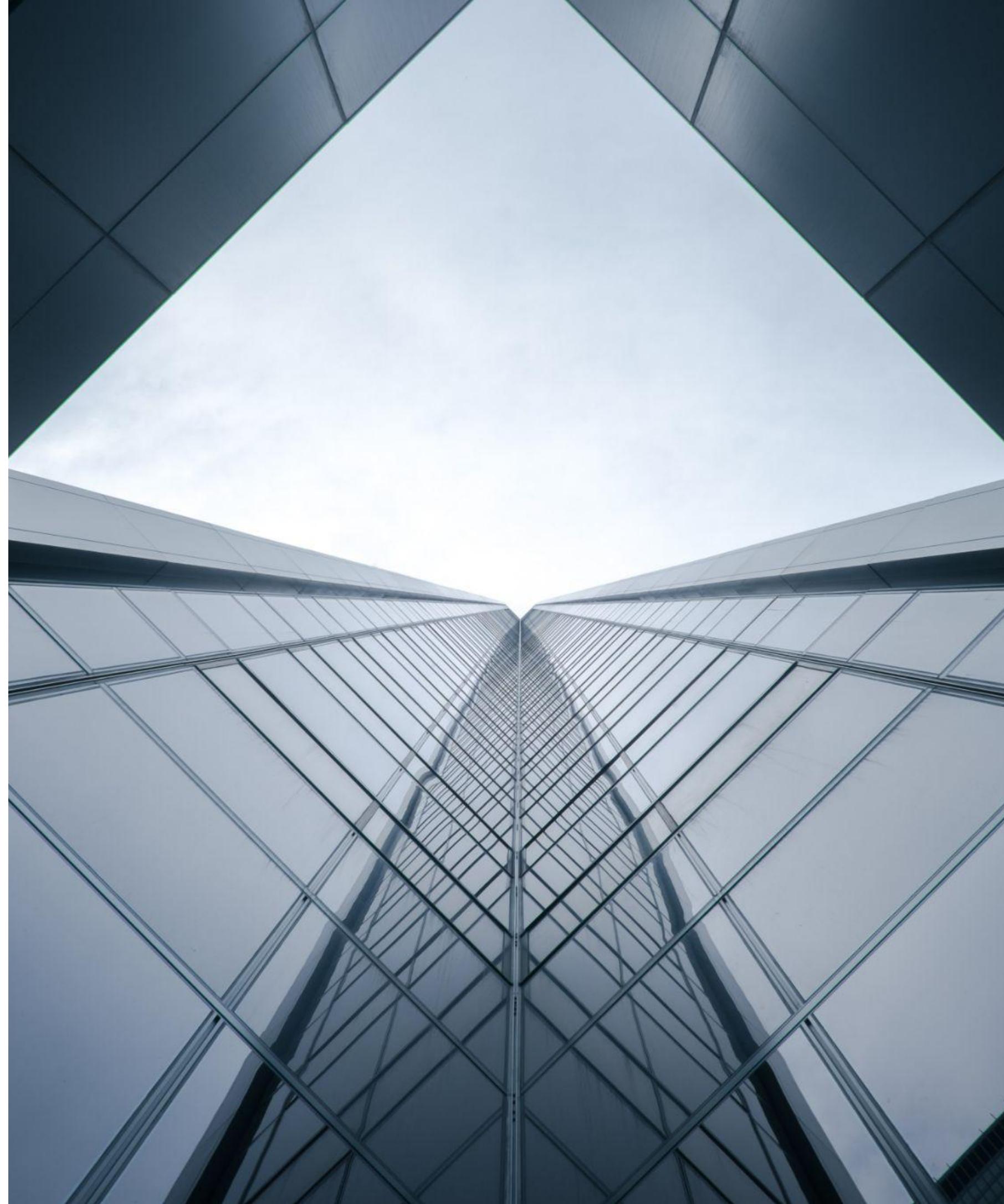
The thought of changing what they do, and especially with more effort, can be an obstacle with some executives, and it's important to prove why it's not so.

/03 DATA AS BUSINESS (NOT IT)

One of the most frequent misconceptions is that data are the responsibility of IT, and it's important to show they are a business function, with accountable owners.

/04 DG NOT WELCOME HERE

There may be executives that fear the changes brought by DG - especially if they are some of the major data offenders. Ironically, IT is a frequent one.



PLANNING

PLANNING ARCHITECTURE AND FRAMEWORK

In order to allow it to run, the architecture and the framework of DG must be put into place. Key areas include:

- Tech/tooling. Which technology will support the DG program. In many cases, it may overlap with DM tech;
- Processes. Which processes will be covered by DG. Is it about monitoring information entry? Enforcing data privacy and security? Monitoring data retention periods and disposal? All of them?
- Roles and responsibilities. Who will be the data owner(s) and data steward(s)? What specific function will every single one of them do?
- Operating model. Bringing it together. Who does what, how do they do it, and using which tools? A summary;

PLANNING ARCHITECTURE AND FRAMEWORK

There are many different tools that can be used. DG tools, however, have a different focus from DM tools (such as for profiling, parsing, de-duplicating, annotating, etc).

DG tool functionality usually includes:

- Metadata management. Tools that enable a metadata repository, and functionality such as business rules, a data glossary, documentation of DG metrics, and more;
- Provenance and lineage. Tools that help track where data come from, how trusted they are, and every transformation they go through - as well as creating audit trails;
- Administrative. These are tools to actually define DG processes, roles, workflows, policies, and others;





PLANNING

PLANNING ARCHITECTURE AND FRAMEWORK

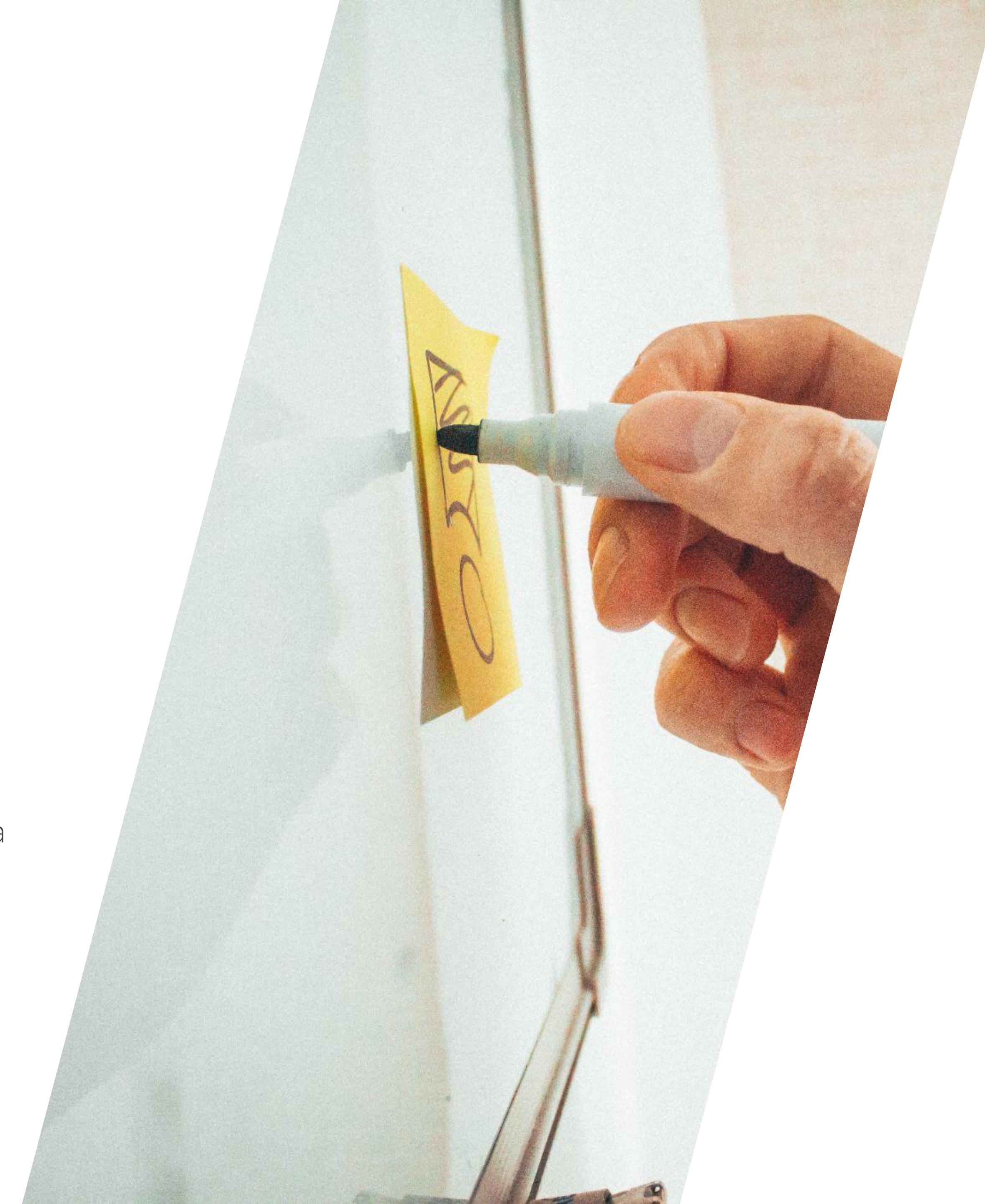
One big trap that many organisations fall into is to buy tools before you have a use for them. Just like DQ issues must be inevitably tied to business impact, tools must be inevitably tied to a use case. For example:

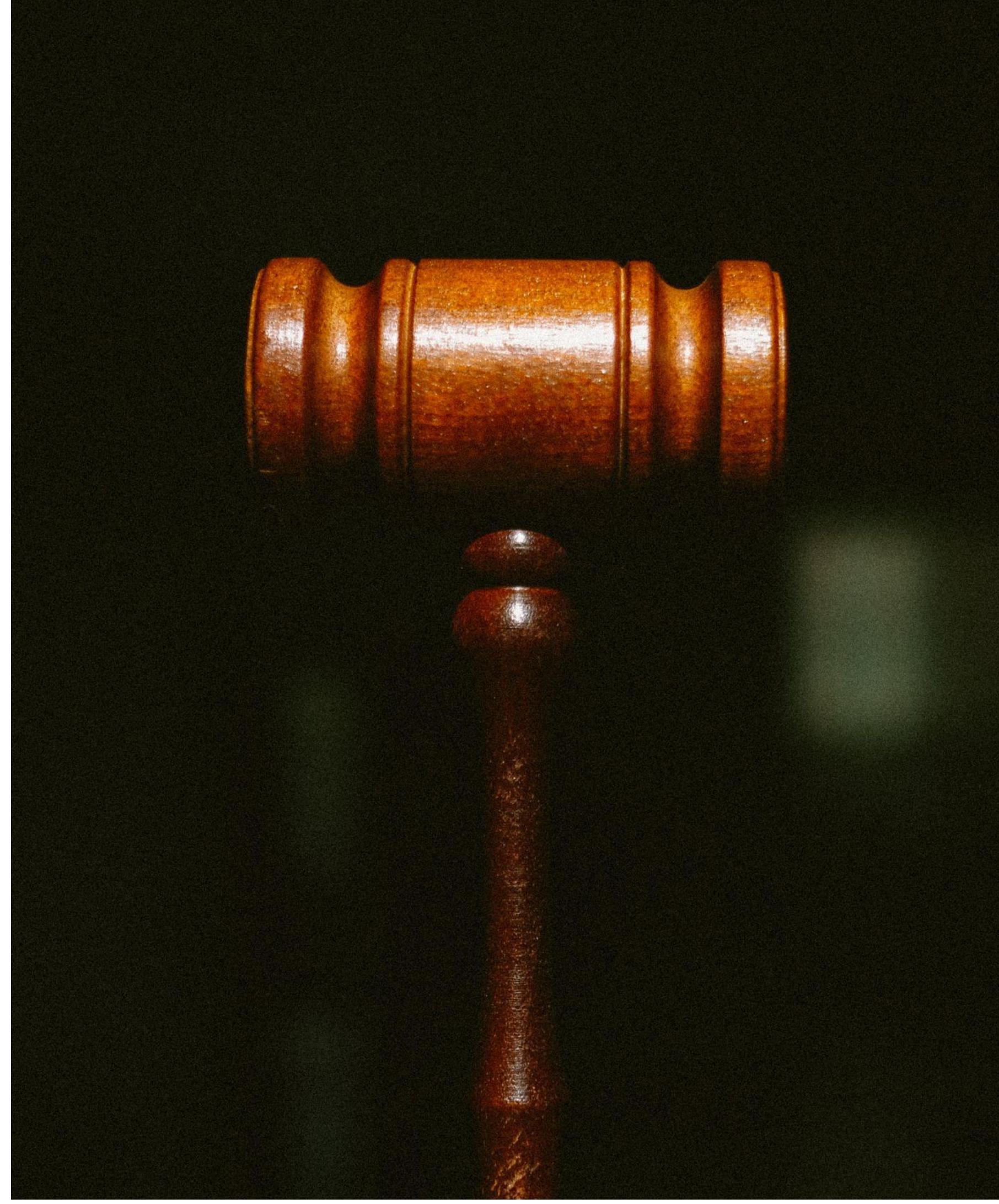
- Bad: “We will use Tool ABC to manage data lineage”;
 - Good: “The DG project will oversee PII retention + disposal. We will use Tool ABC to manage lineage of PII”;
 - Bad: “We will use Tool ABC as a metadata repository”;
 - Good: “The DG project will piggyback on MDM to define Customer and Supplier data validity rules. We will use Tool ABC as a metadata repository to store these business rules and the data expectations for all DB columns and tables”;

PLANNING ARCHITECTURE AND FRAMEWORK

Then come processes. Just like tools, they can't be "generic", but instead must be tied to a specific use case. Examples include:

- Bad: "We will track disposal of cardholder data (CHD);
 - Good: "We will track disposal of cardholder data (CHD). We do this to comply with the PCI-DSS deadlines and disposal requirements, and achieve compliance";
- Bad: "We will enforce data requirements for Customer data entry in the Salesforce connecting application";
 - Good: "We will enforce data requirements for Customer data entry in the Sales connecting application. We will do this to ensure high-quality Customer data, reflected in higher sales and marketing campaign revenue";





PLANNING

PLANNING ARCHITECTURE AND FRAMEWORK

In general, you can think of processes as enforcing rules at some stage of the information lifecycle. For example:

- Creation:
 - “We will ensure DQ requirements when information is entered/integrated from a vendor”;
- Updating:
 - “We will enforce Access Control on data editing”;
 - “We will test data for compliance with DQ requirements after editing”;
- Usage:
 - “We will enforce Access Control on users with access to data (raw or in data products);
 - “We'll monitor network traffic for possible leaks / exfil”;

PLANNING ARCHITECTURE AND FRAMEWORK

Naturally, the scope of the DG program (or specific projects) varies based on use case but there will usually be a group of related processes. For example, for a DG project to protect PII:

- Creation: “We will capture PII data lineage information at the creation stage, as well as data owner information”;
- Usage: “We will enforce Access Control on all PII data, requiring credentials + department + purpose”;
- Security and Privacy: “All PII data will be encrypted / masked / tokenised at rest and in transit”;
- PII Data Subject: “There will be a public contact form so a data subject can request their data or have them deleted”;
- Deletion: “Data will be deleted after the retention period expires, and using secure, approved methods”;





PLANNING

PLANNING ARCHITECTURE AND FRAMEWORK

The assignment of roles and responsibilities is important. Although we've covered the "general" roles and functions, it's important to consolidate them for the current program/project.

- For example, you have a DG project to ensure DQ for Customer data, to support an MDM initiative:
 - Who will discover the business rules for Customer data?
 - Who will monitor the controls that enforce these?
 - Who will record information about both in the metadata repository?
 - Who will periodically track DG metrics?
 - Who will be informed of outstanding issues and expected resolution times?

PLANNING ARCHITECTURE AND FRAMEWORK

Then, finally, all of this comes together in the operating model for the DG program. For example, for an MDM Customer data initiative:

- Tech:
 - Metadata repository tool for storing rule information;
 - Data profiling tool for discovering Customer DQ issues;
- Processes:
 - Monitoring Customer DQ at creation;
 - Using AC for Customer data editing + DQ testing after;
- Roles and responsibilities:
 - Person ABC as Data Steward, to discover rules, fill them in the metadata hub, as well as the controls used;
 - Person DEF as Data Owner, to receive DQ issue reports;



PLANNING ARCHITECTURE AND FRAMEWORK EXAMPLES

/01 WHERE TO GROW?

Once you've run a successful DG project, you may ask, "Where to grow?". You can tackle more stages of data, or more projects with the same stage. Both can work.

/02 MORE OR LESS FORMAL

While it's important to define specific processes, tools, stages, roles, this can be less formal in smaller projects. But as DG expands, it must become formal.

/03 EVERYBODY ON BOARD

It's crucial to make everyone aware of their responsibilities in the operating model. Someone not aware of what they're agreeing to just causes issues.

PLANNING ARCHITECTURE AND FRAMEWORK KEY TAKEAWAYS

/01 TOOLS, PROCESSES, ROLES

The architecture of a DG project - summarised as the operating model - usually defines what tools to use, for which processes, and involving whom.

/02 SPECIFIC USE CASES

Whether tools or processes, it's crucial these serve specific use cases and not generic ones. Generic means they will not be adopted or used (a waste).

/03 DATA LIFECYCLE STAGES

Depending on the DG project, different data stages may be affected. Some DG projects may just affect disposal, some may just affect creation, and/or others.

/04 THE OPERATING MODEL

The operating model is the summarisation of everything. Who will do what, at which stages of data, and using which tools? It brings it all together.



ROLL-OUT

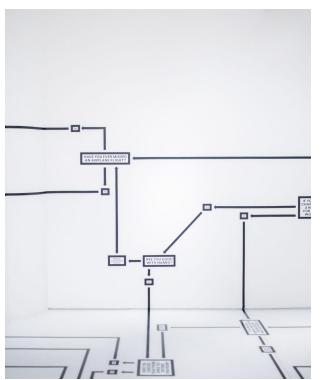
ROLL-OUT INTRODUCTION

The roll-out stage of DG is when things start happening. We put the plan into practice, start putting our processes in place, and effectively executing DG.

At this stage, we will make DG real, including defining a roadmap and milestones, dealing with resistance by stakeholders in practice, and prioritising the projects that go well with DG and allow us to prove value quickly.

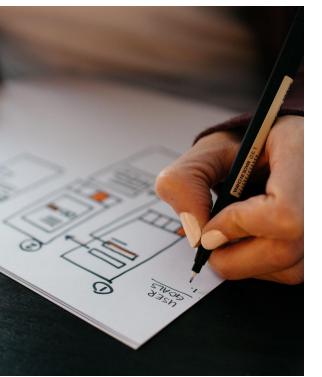
ROLL-OUT

We will cover the roll-out of DG with **two topics**:



INITIAL DEPLOYMENT

What to keep in mind when initially deploying DG to ensure a successful implementation.



COMPLEMENTARY INITIATIVES

Which initiatives DG goes well with, as well as how to prioritise them.



ROLL-OUT

ROLL-OUT INITIAL DEPLOYMENT

The roll-out stage of DG is when planning turns into action, and it's important that the necessary elements are in place. Regardless of the size of the program, and the people involved, it's important to have a minimum set of processes to put into place to assure DG.

Besides this, some key elements at the deployment stage include:

- The basic, minimum roles and activities included;
- A roadmap with calendarised activities;
- A set of metrics to measure program performance, especially tied to data monetisation;
- A change management plan to prevent reverting;

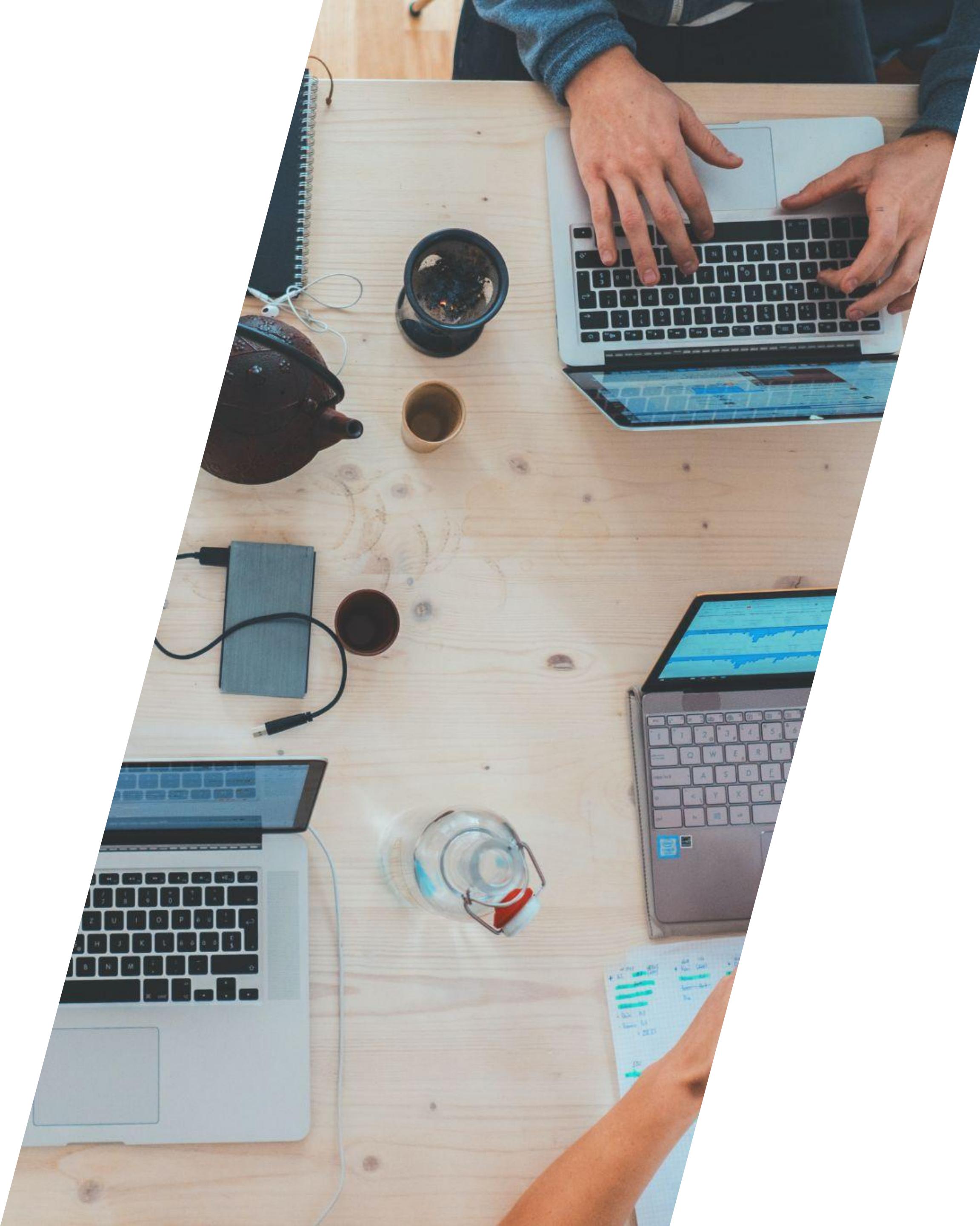
ROLL-OUT

ROLL-OUT INITIAL DEPLOYMENT

We start with the roles and activities of the DG program. We want to define a “minimum” framework because there will be resistance (DG demands behavioral change), and we don’t want to “ruffle feathers” at an initial stage until value is proven.

Let’s say the DG program starts with a scope of PII data usage and disposal within the Marketing department:

- Data owner: Possibly the CMO or senior marketing manager;
- Data steward tasks: Classify PII as sensitive data in metadata hub, govern all PII lifecycle operations from creation to disposal, embed lineage information;
- Data owner tasks: Collect progress metrics, report to DC;





ROLL-OUT

ROLL-OUT INITIAL DEPLOYMENT

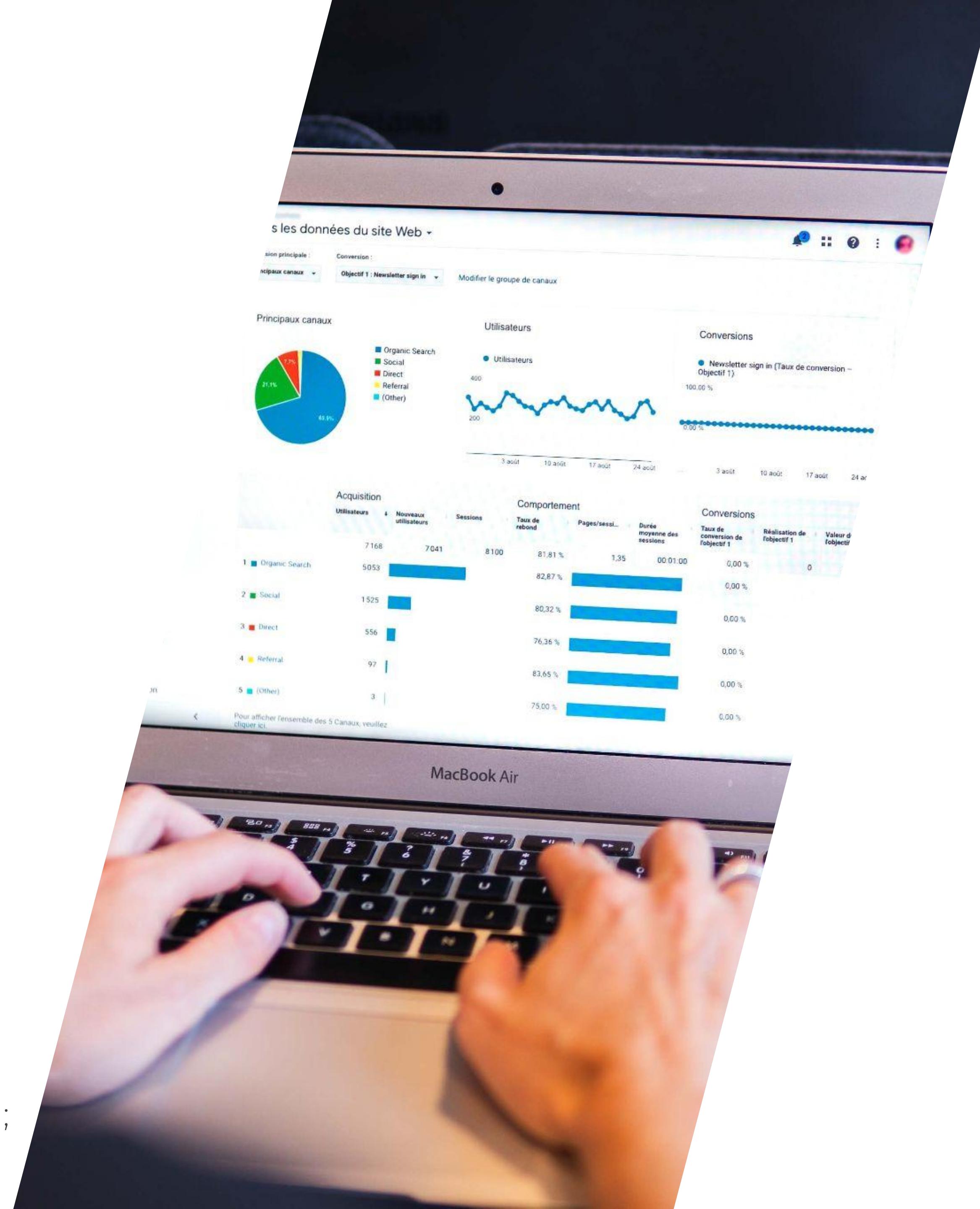
The DG roadmap is crucial to illustrate what projects or programs DG will oversee. This is crucial to know where we will track metrics and create impact.

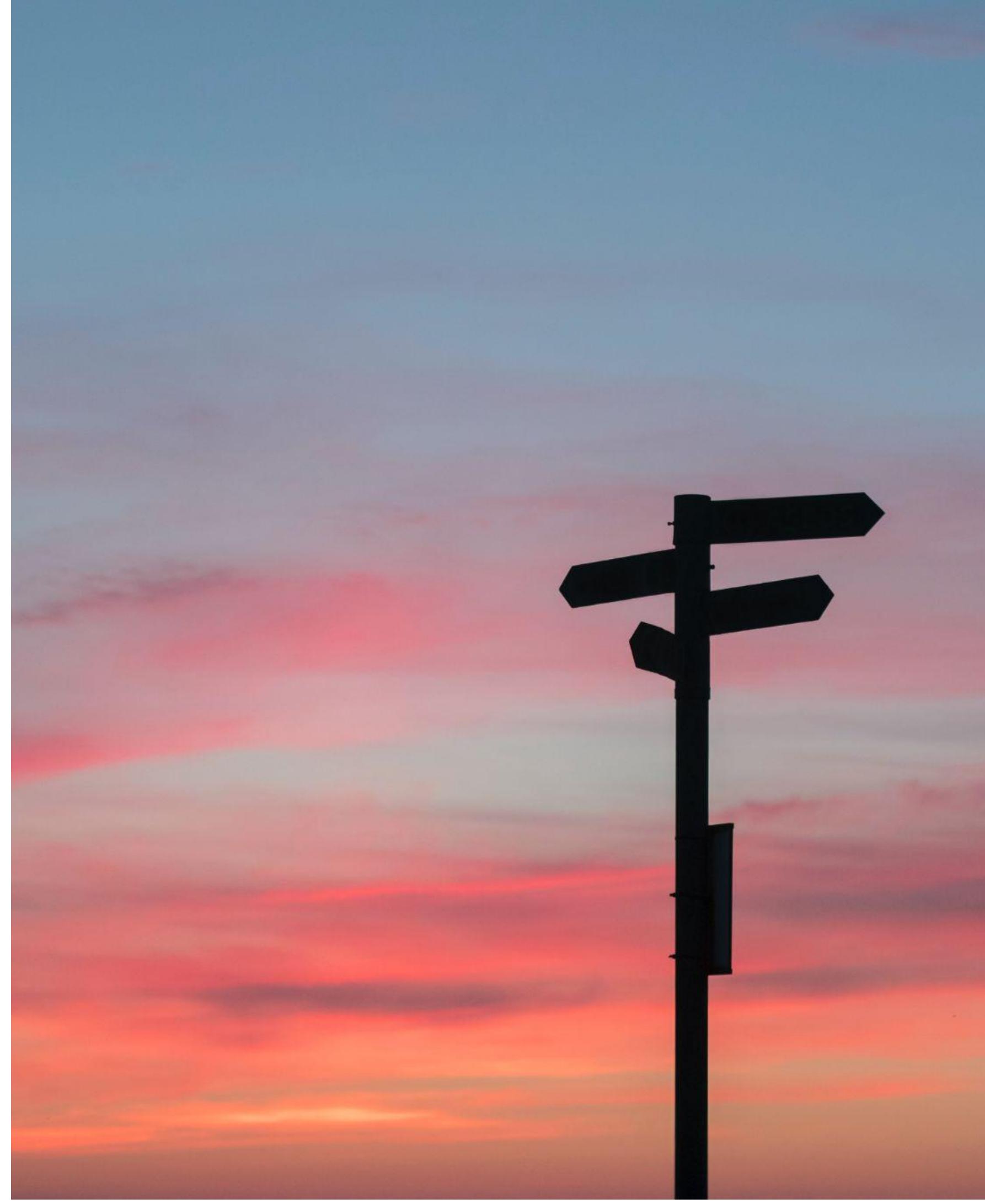
- If “data-centric” projects or programs such as MDM, EIM, big data, analytics or BI exist, these are great candidates;
- Ironically, the areas that most resist DG are IT, including app delivery/project delivery, considering it a “hindrance”;
- Example: Govern PII in Marketing, then Sales, then R&D... Implementing the roadmap will make resistance very clear (whose project/program owners are on board, and which are not really committing), and they provide feedback to it.
- In early stages, it may be interesting to temporarily “give up” on resisted projects and show value with the others;

ROLL-OUT INITIAL DEPLOYMENT

Although this is an early stage, collecting DG success metrics is imperative in order to show value from the beginning and allow the program to be sustained and expanded.

- Metrics vary, but remember they always tie back to data monetisation, which is of 1 of 3 main types: financial gain, reducing operational costs, or reduced compliance costs;
- In our example of governing PII in Marketing, we can include metrics such as:
 - % of data sources in compliance with PII regulation;
 - Expected savings from regulatory fines/data breaches;
 - Expected customer trust (and possible gains from it);
- Metrics may be not just about the financial gains with DG, but also operational ones (training progress, project costs...);





ROLL-OUT

ROLL-OUT INITIAL DEPLOYMENT

Finally, some sort of change management plan will be necessary to sustain DG. Do not underestimate this component.

- The rationale behind it is that DG is being established for a reason - which is that the current managing of data was not good. So behavior must be changed;
 - The problem is... no one wants to change behavior;
 - A formal change management plan is necessary, even if a simple one, to formalise the necessary changes in behavior;
 - Examples include:
 - PII collection process: “PII must be classified in terms of sensitivity, and masked, at creation time;”
 - PII disposal: “PII retention rate must be in metadata”;

ROLL-OUT INITIAL DEPLOYMENT EXAMPLES

/01 GAINS MAY COME LATER

DG may not produce a lot of gains in the initial project, but much larger ones later. This is because tools/people are the same, but impact grows very fast.

/02 APPARENT CHANGE ISSUES

The roadmap and behavioral changes required make resistance clear. Some people are willing to adopt it, while others may openly say, "I refuse to do this".

/03 BACK TO SQUARE ONE

In cases where a DG program does not obtain support, it's very frequent for the successes and changes to easily revert, as nothing is enforcing them anymore.

ROLL-OUT INITIAL DEPLOYMENT KEY TAKEAWAYS

/01 4 KEY ELEMENTS

To properly deploy a DG program, there are usually 4 key elements: roles and responsibilities, a roadmap, metrics for success, and a change management plan.

/02 CHANGE + VALUE

At their core, the goals of a DG program are to change organisational behavior, to better govern data, and to show value by doing so.

/03 PRIORITISATION IS KEY

As with anything in life, no roadmap survives reality. Some managers will resist, some behaviors will not stick. A nascent DG program must stick to what works.



ROLL-OUT

ROLL-OUT COMPLEMENTARY INITIATIVES

Despite any type of process benefitting from DG, there are specific initiatives which present a high level of synergy - in some cases, these won't even succeed without DG. Examples include:

- Data-centric programs. Master Data Management (MDM), Enterprise Information Management (EIM);
- Analytics, big data and/or AI projects which rely on data to generate insights;
- ERP implementation. Covers a lot (if not all) functions within a company. DG may be included in the package;
- In most cases, the owners of these projects will realise that lack of high-quality data is a barrier, and that DG is crucial to their proper functioning;

ROLL-OUT

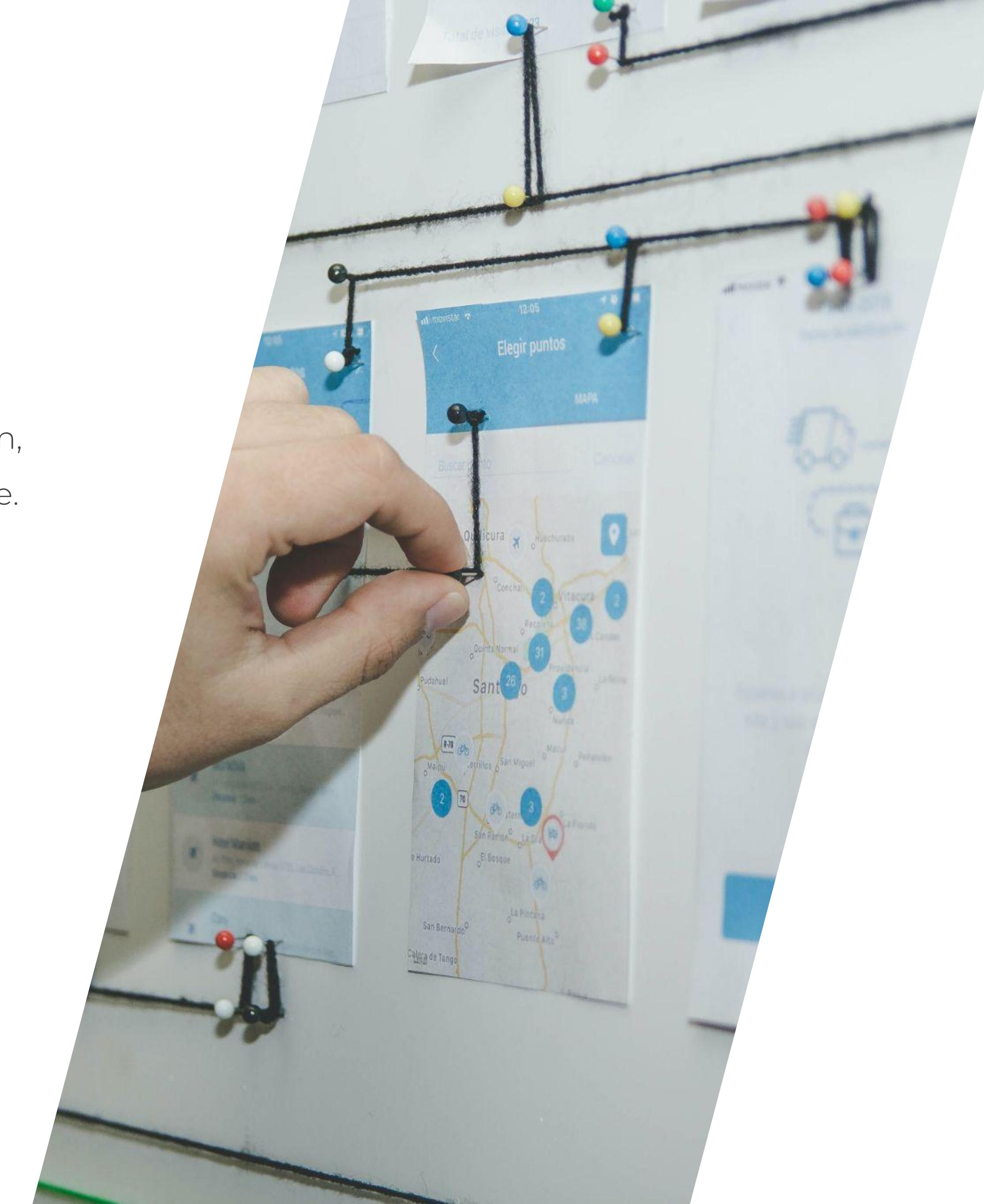
ROLL-OUT COMPLEMENTARY INITIATIVES

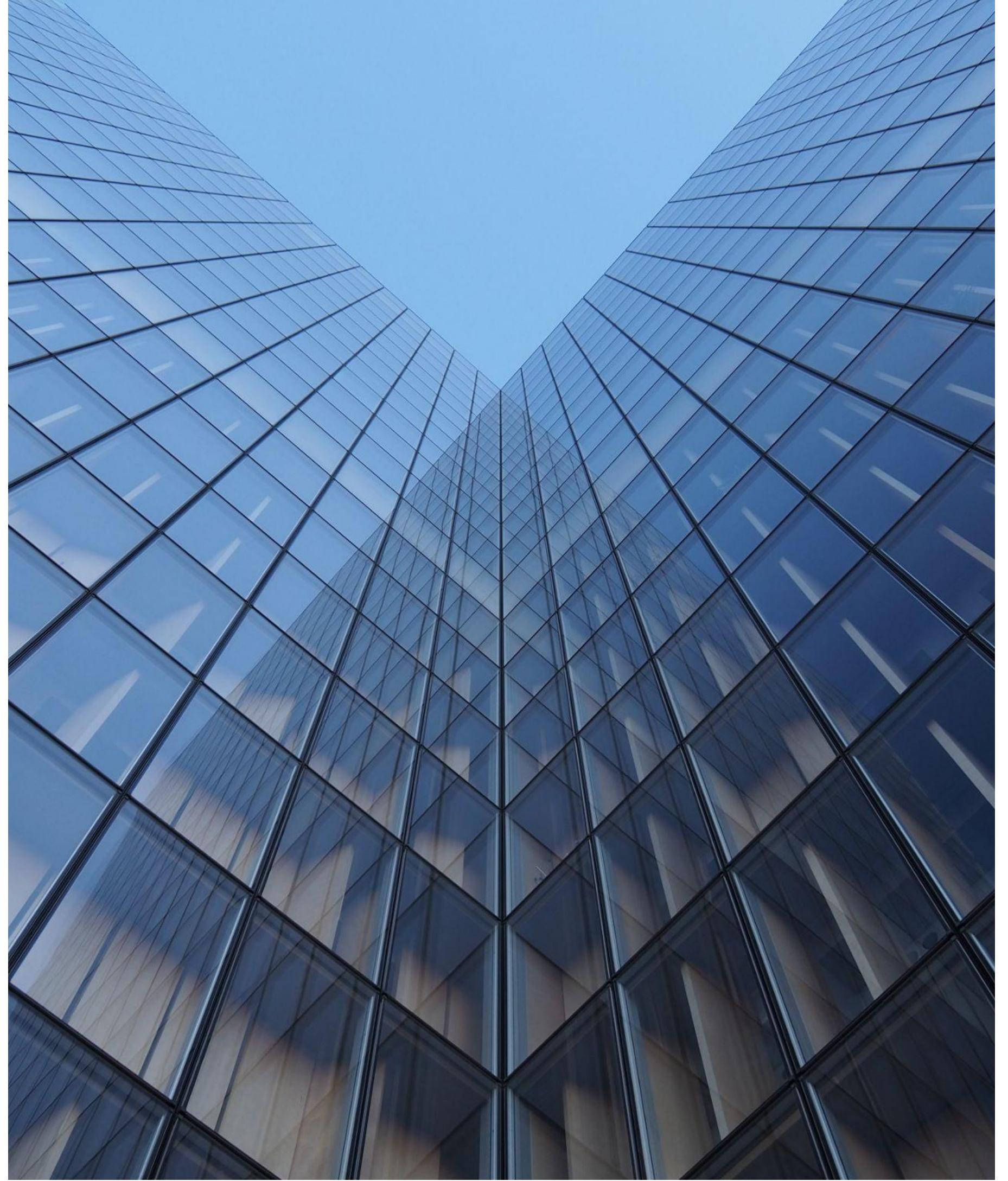
These cases present visible opportunities for implementation, because DG can “piggyback” on these projects to show value.

- For example, in an MDM project, metrics can include the uniqueness and completeness of customer data, which can be easily monitored and optimised with DG;

It's important, however, to align with the project owners to find the right place for DG:

- In many ERP implementations, DG is “sold” as a service attached, provided by an external third party, and can be instead brought inside;
- In an MDM/EIM project, IT may be owning the DG portion of it, considering it DM, and it must be brought outside;





ROLL-OUT

ROLL-OUT COMPLEMENTARY INITIATIVES

Regardless of the project that DG is supporting, it's crucial to maintain the enterprise-wide mindset.

- That is, DG cannot be considered a "local" function, but be something the whole organisation performs;
 - Naturally, it can start locally, but it must branch out sooner or later, and spread to other departments;
- The key is to prioritise the projects where DG shows the most value, and then slowly expand to the "resistance";

In many cases, this also facilitates implementing DG, as a lot of assessments (including business impact analyses, or even whole business cases) already exist for these initiatives, and DG can leverage these vs. building from scratch

ROLL-OUT COMPLEMENTARY INITIATIVES EXAMPLES

/01 **SOME PROJECTS NEED DG**

DG has an easy time providing value to data-centric projects, as these require high-quality data - which DG ensures. This provides great opportunities.

/02 **WHAT EXISTS (AND WHERE)**

The first step to proposing DG projects to go with these initiatives is to know if anything is already being done in terms of DG, and if so, in what format.

/03 **ENTERPRISE-WIDE**

These projects also remind us, again, that DG is a function that must be enterprise-wide and of each department. It must grow, sooner or later, to this scale.

ROLL-OUT COMPLEMENTARY INITIATIVES KEY TAKEAWAYS

/01 3 TYPES THAT GO WELL

There are 3 types of projects that DG has great synergies with: data-centric (MDM, EIM...), analytics/AI, and ERP implementation. They all depend on HQ data

/02 PIGGYBACKING WORKS

In these cases, instead of starting from scratch, DG can both attach itself to the performance metrics of these projects, and use analyses from them to show value.

/03 ALWAYS ENTERPRISE

It's important, however, to never fall into the trap of just doing DG as an isolated project supporting these, but spread to the organisation eventually.



ITERATION

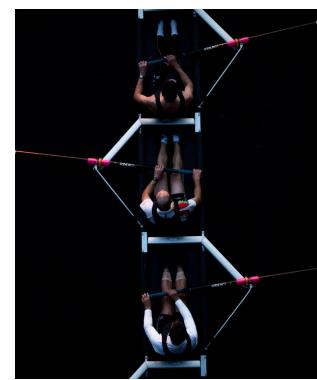
ITERATION INTRODUCTION

As a DG program dives into a new iteration - whether to optimise the current program, or to expand to support more projects - there are things that can be improved and fine-tuned.

When iterating, it's important to maintain current processes, such as training or communication, while at the same time ensuring support from new stakeholders if new projects are involved, among other elements.

ITERATION

We will touch on **two key topics** related to iterating a DG program:



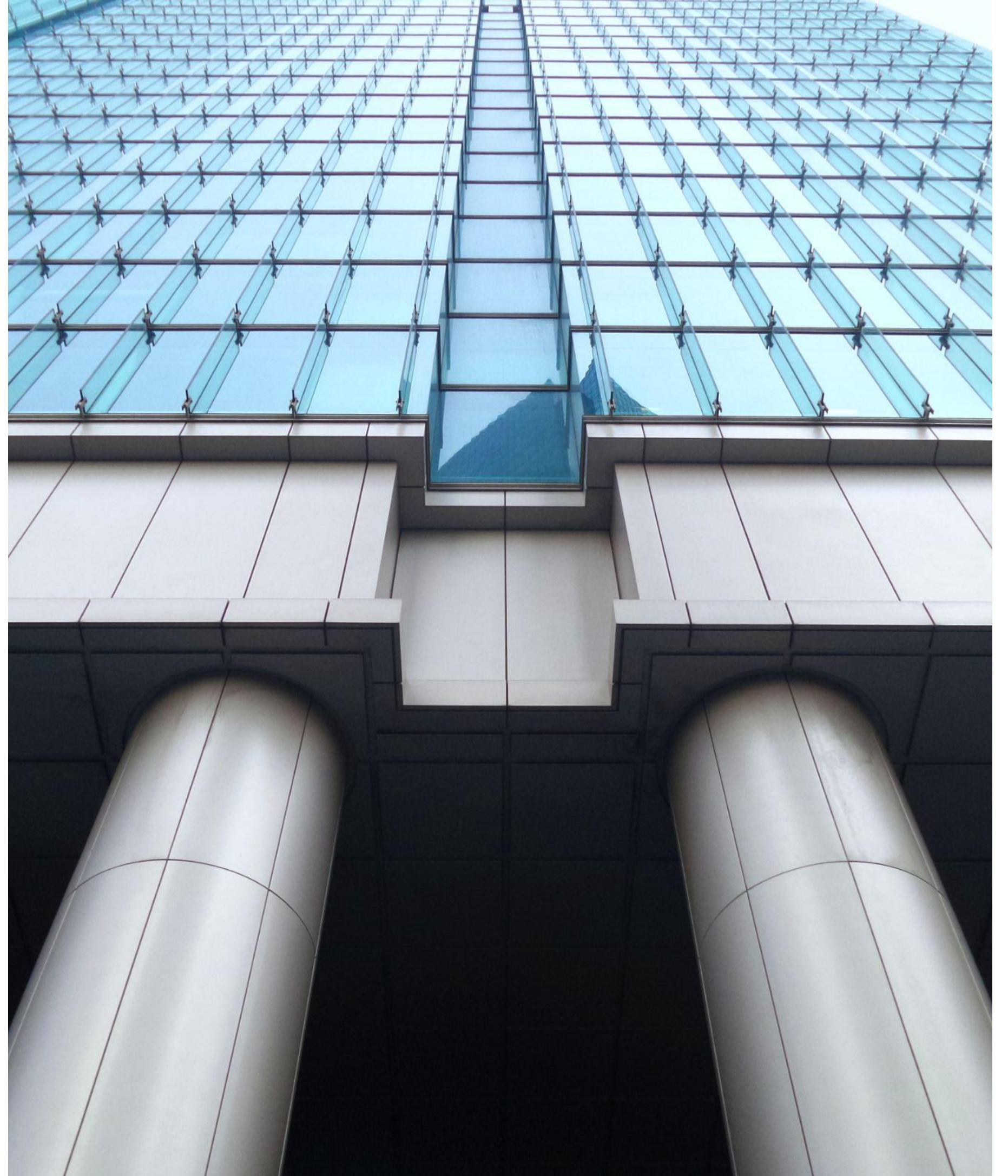
SUSTAINING

What is needed to sustain a DG program. Ensuring processes are continued + support is given.



SCALING

Properly scaling a DG program, selecting the correct projects and dealing with resistance.



ITERATION

ITERATION SUSTAINING

Whether a DG program has already reached maximum scope or still has room to grow, something that is essential to realise is that some effort is needed to sustain it. Even if not “actual work”, but at least in terms of expectations and alignment.

There are usually two levels at which effort is needed:

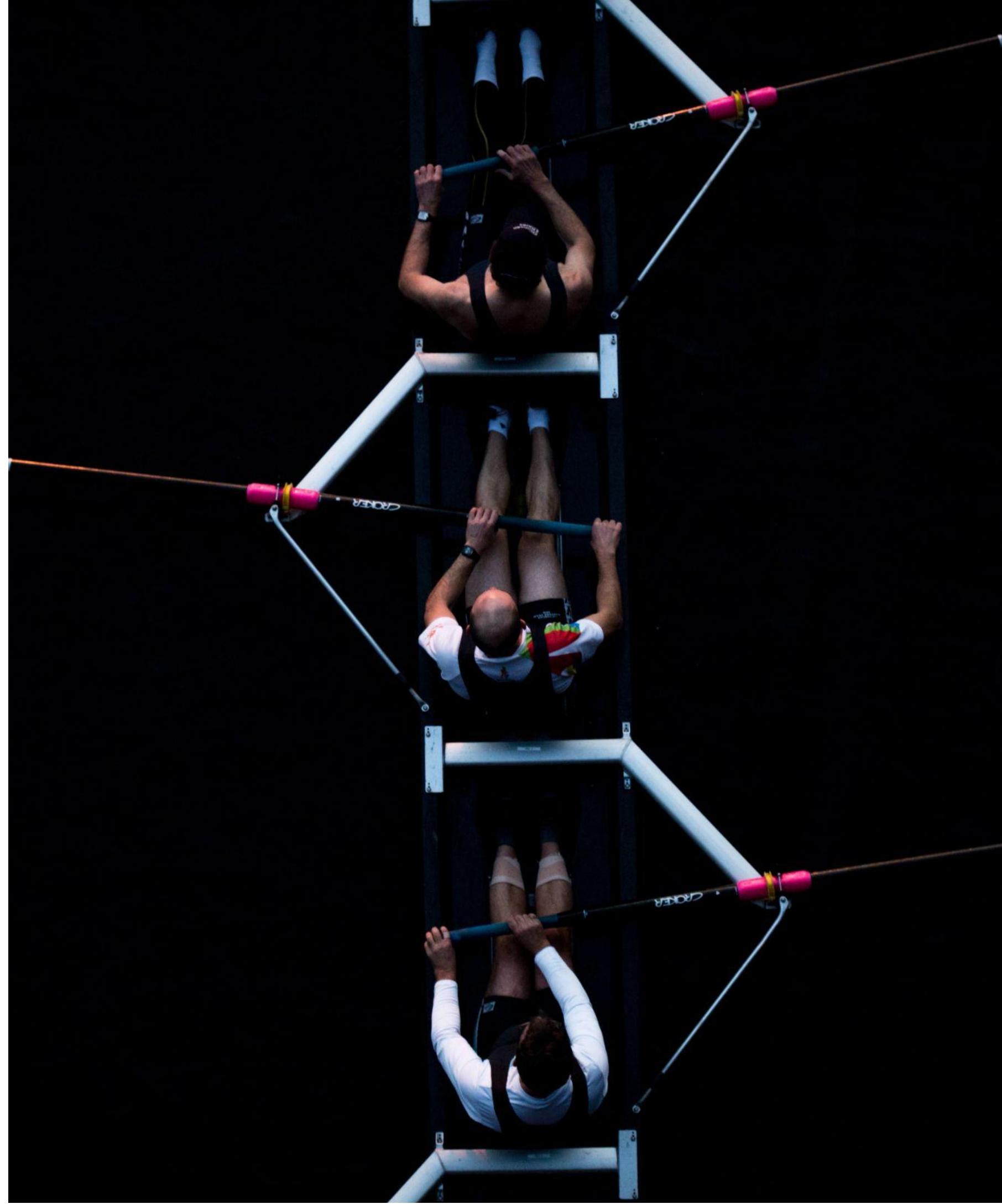
- Strategic: Maintaining alignment with protractors and dealing with resistance from detractors;
 - Essential to overcome roadblocks to implementation;
- Operational: Ensuring changes are implemented, as well as ensuring that training, communication, literacy continue;
 - Letting these be forgotten is a surefire way to make people care less and less about DG;

ITERATION SUSTAINING

On a strategic level, it's all about executive communication and commitment. The DG sponsors and supporting executives need to commit time and resources for the program to work - and convince other executives to do it. This includes:

- Having a champion that promptly deals with resistance (if someone questions the value or metrics of DG, who will address it - and immediately?);
- Making sure that your DG champion keeps performing (are they still evangelizing? Are they translating DG performance into business success metrics?);
- Ensuring that the value of DG is articulated (many executives consider data “esoteric” - change their mind);





ITERATION

ITERATION SUSTAINING

On the operational level, it's important to understand that DG brings a change (or several!), and it's important to enforce these. As with any other type of project, there is a danger of forgetting the processes and documents when the project begins.

- It's important, therefore, to maintaining the training and communication processes defined, and to improve them based on feedback. They're not a "one-off" thing;
- Also, the changes in behavior (AC to data, data encryption, acceptable uses, and many others) are consolidated in change management processes and data-related controls. It's important to ensure these are enforced, and not just "forgotten" once people start doing things right;

ITERATION SUSTAINING EXAMPLES

/01 START AND CONTINUE

Not unlike other types of changes to a company, DG processes must not only be started, but continued as time goes by, to prevent regressing into old behaviors.

/02 EXECUTIVE SUPPORT

The DG champion is crucial to keep breaking resistance and allowing DG to expand. Someone that can leverage political capital is needed at the C-Suite.

/03 BEFORE AND AFTER

As with other types of change management, we define the before and the after, the changes to make the transition, and enforce the changed behaviors.

ITERATION SUSTAINING KEY TAKEAWAYS

/01 IT TAKES EFFORT

A DG program is not self-sustaining by nature, regardless of scope. Some effort is needed to keep obtaining support and keep people behaving well.

/02 STRATEGY + OPERATIONS

Sustaining a DG program involves commitment at two different levels - strategy and operations. Ensuring support at the executive level, and in operations, too.

/03 CHANGE MGMT + CONTROLS

In specific, the day-to-day changes brought by DG are usually enforced by controls on data, and by the change mgmt. process leading to the new behaviors.



ITERATION

ITERATION SCALING

The process of scaling a DG program is related to sustaining one, but it may bring some additional challenges, especially related to resistance.

- Besides the effort that should be put into maintaining the behavioral changes and controls effected, the inclusion of new projects, new departments and new teams may bring new types of metrics to track, and new stakeholders to convince.

There are two important aspects in terms of expansion:

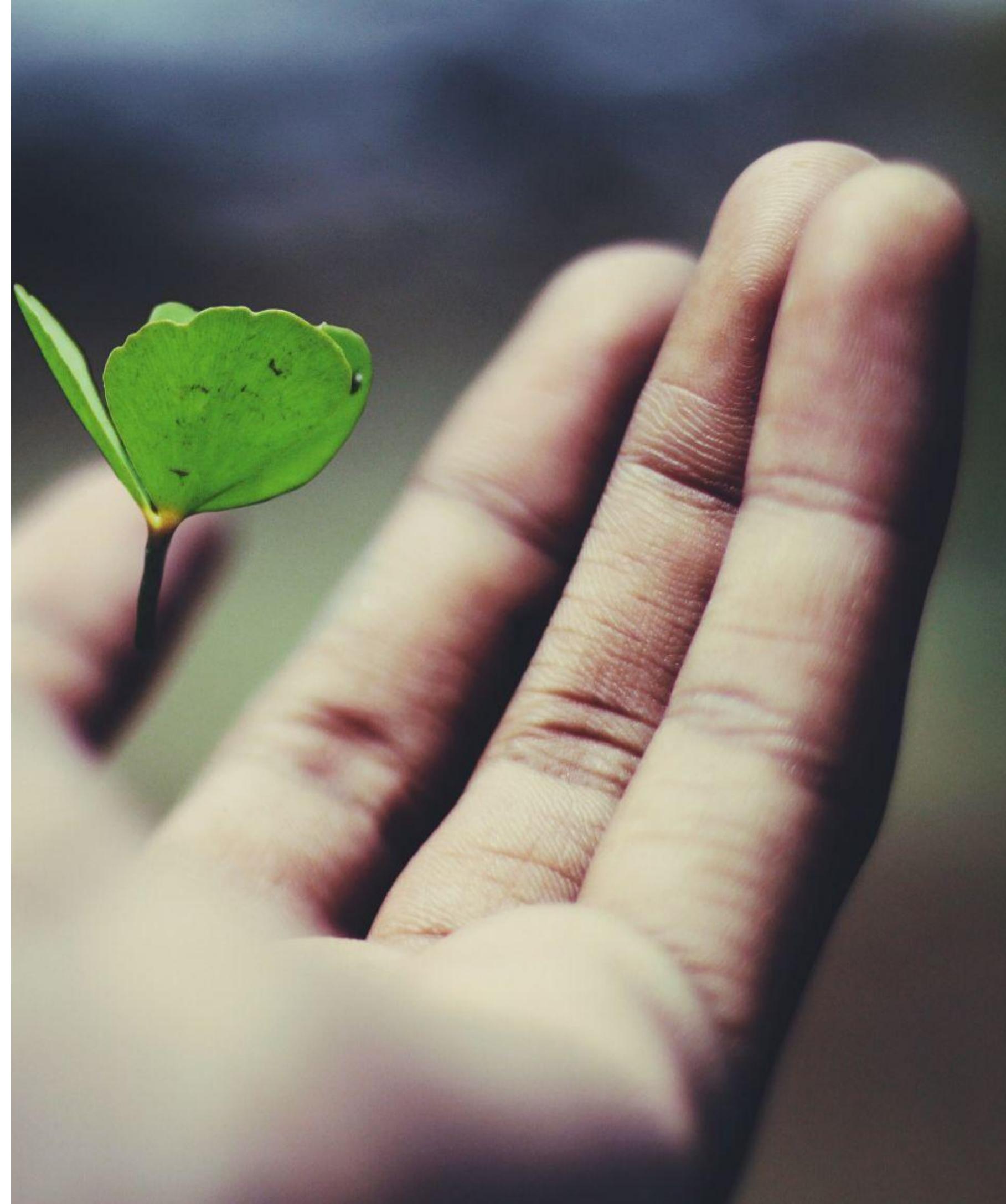
- Properly dealing with resistance from new stakeholders;
- Defining the level of federation and standardisation across the organisation;

ITERATION SCALING

Dealing with resistance from new departments and teams is similar to dealing with it in sustaining the current program. Some stakeholders, at any level, will resist (openly or in a passive-aggressive manner). Important factors include:

- The champion's commitment and effort. The DG champion must be willing to spend some political capital and persuade different executives until they're all on board;
- The company culture. Whether there are consequences for people not aligning with required behaviors (and which);
- The prioritisation of expansion (DG will have multiple project candidates that can be supported - picking the "easiest" ones decreases the probability of resistance);
- The value articulation for different departments/teams;





ITERATION

ITERATION SCALING

As mentioned, as DG becomes more sophisticated, its goal is to “disappear” into the organisation. There shouldn’t be a “DG team”, but instead DG becoming a function of every team.

One element that will become very important as DG expands is the federation. That is, how harmonised and standardised DG is across the organisation.

- Since many companies are present in multiple countries, obeying different regulations, and with radically different datasets, we can't just set the same processes and policies for the entire organisation;
- But we can set them for certain teams, countries, geographies or groups - that is federation;

ITERATION SCALING

Federation, in specific, becomes important when you have big, complex organisations with heterogeneous landscapes.

For example:

- Companies dealing with different geographies, where each one has different regulation, data retention deadlines, etc;
- Variety and resistance of different teams and business lines. If all activities are similar, DG can be similar for them. If some are very different - or leaders resist - not so much;
- The federation of the actual organisation. That is, if it tends to have centralised decisions and controls for other areas, chances are that DG will also be centralised by nature;
- The variety of architecture and apps also counts. Small, dispersed app repositories may result in dispersed DG;



ITERATION SCALING EXAMPLES

/01 FROM DG TO DEPARTMENTS

Once DG makes the “final step” to becoming a transversal company function, a lot of “DG employees” become “department employees” doing DG.

/02 A FEDERATED CHALLENGE

DG is most efficient when the policies/processes are global, but local regulation and legislation force fragmentation. Companies must find a balance.

/03 CULTURE AFFECTS IT ALL

The company culture impacts whether DG can expand. But to a deeper level, it affects change capacity, data literacy and other crucial mindsets.

ITERATION SCALING KEY TAKEAWAYS

/01 RESISTANCE + FEDERATION

Two important topics to take into account when scaling are resistance by different teams and departments, and the federation of DG across the org.

/02 RESISTANCE IN PRACTICE

Elements such as the culture of the company, the commitment of the DG champion, and the specific departments involved define the resistance faced.

/03 ABSORPTION/FEDERATION

As it grows, DG should be absorbed by the different business functions. There should also be a level of federation, that is, how standardised DG controls are.

/04 FEDERATION ELEMENTS

There are many different elements affecting the level of federation of DG, usually in terms of varied the organisation is. Geography, regulation, apps, culture...



DG IMPLEMENTATION

DG IMPLEMENTATION MONITORING

Monitoring is a process that is crucial since the initial deployment of DG, regardless of scope. And this is because, in order to show financial benefit from the DG program, performance must be tracked.

Regardless of the activities (and the current stage) of the DG program, monitoring is important for several reasons:

- Compliance and regulation. Monitoring whether data are in compliance allows remediation, preventing sanctions;
- Alerts and responses. If something goes wrong with data (for example, LQ critical data), you can take early action;
- Transparency and pervasiveness. Monitoring allows you to know what data exist in all apps/stores/departments/LoBs;

DG IMPLEMENTATION MONITORING

Naturally, the areas to monitor depend on the capabilities of DG that have been implemented. However, some key ones are:

- Metadata quality monitoring. Monitoring the quality of metadata information in the Metadata Hub, and, for example, number of “Complete” data sources vs. “Pending”;
- Lineage monitoring. Monitoring information on origin, transformations, trustworthiness, others about data;
- Data Quality (DQ). Actually monitoring DQ. Are data complete? Are they accurate? Etc;
- Compliance and privacy monitoring. Are we obeying data retention regulation? Are data used only for authorised purposes?





DG IMPLEMENTATION

DG IMPLEMENTATION MONITORING

Specific metrics to track may include:

- Data Quality
 - Number of open DQ issues/average remediation time;
 - Number of “Trusted” datasets (full metadata);
 - Data volume by data owner/department/LoB;
- Metadata Monitoring (incl. Lineage)
 - Percentage of Datasets with Full Metadata;
 - Average Undocumented Lineage Steps;
 - Average Metadata Completeness Percentage;
- Compliance and Regulation
 - Number of Compliance Breaches/Violations;
 - Amount in Fines/Sanctions Due to Data Breaches;
 - Percentage of Compliant Data;

DG IMPLEMENTATION MONITORING

Usually, the monitoring of specific metrics is intimately tied to the tools that are involved in them.

- For example, privacy and compliance metrics may be collected by the security systems that you have in place to protect the data themselves;
- Or, for example, DQ monitoring may already be performed by the tools that are actually used to profile and/or remediate the data of multiple sources;

It is possible that, in some cases, monitoring systems have to be developed from scratch, but in most cases, monitoring is usually an integration of existing metrics collected by different systems.





DG IMPLEMENTATION

DG IMPLEMENTATION MONITORING

Naturally, monitoring is only done for the DG capabilities implemented, and it grows with them.

- Project 1: Improving completeness of metadata related to uses of PII;
 - Metrics: Metadata completeness as %, # of accesses;
- Project 2: Project 1 + complying with PII retention;
 - Metrics: Above + Compliant deletion timestamps %;

As the DG program grows, new metrics are adopted to prove success for each iteration of the program. It's crucial, however, to start monitoring from the beginning.

- In many cases, if DG doesn't prove traction from the first iteration... it won't get a second one;

DG IMPLEMENTATION MONITORING

It's important to clarify that not all monitoring is equal.

Depending on the classification of data, higher priority data must be monitored more closely, with stricter rules and more severe alerts;

- For example, HIPAA-compliant data vs. public website data will have completely different priorities and controls;

Additionally, it's possible that monitoring may be only done
on a portion of the data, especially if big data are involved.

- DG and DM are already understaffed areas in most cases, and it may not be viable to monitor/remediate all data;
 - In many cases, random quality checks are a cost-effective manner to monitor multiple sources at different stages;



DG IMPLEMENTATION MONITORING EXAMPLES

/01 CONTROLS BY DATA CLASS

Monitoring is one of the reasons why security controls must be based on data classes. For example, sensitive data breach alerts take priority over public data ones.

/02 INFOSEC INTEGRATION

For companies with a sophisticated Information Security architecture, many control metrics can be “siphoned” here. AC logs, secure data disposal, etc.

/03 THE “GOVERNANCE” IN DG

Monitoring emphasizes the “governance” component of “data governance”. It’s where we track DQ issues, security issues, and other key metrics related to data.

DG IMPLEMENTATION MONITORING KEY TAKEAWAYS

/01 CRUCIAL TO PROVE VALUE

Monitoring of a DG program is crucial in order to show that the program is succeeding. Metrics should be compared against desired benchmarks.

/02 ALERT, COMPLY, KNOW

Monitoring is useful for multiple purposes, including alerting on problems, providing information on compliance (or not), and knowing what you have.

/03 MULTIPLE AREAS

Monitoring can be performed for multiple capabilities and functionality, but frequent dimensions are DQ, lineage information, metadata quality, and privacy.

/04 FROM THE BEGINNING

Monitoring is crucial to show effectiveness of the DG program and monetisation of the data, and it must be in place from the beginning to be taken seriously.

/05 DIFFERENT PRIORITIES

Monitoring is done for different classes of data, and naturally, these should not be treated the same. Critical or sensitive data must be under close watch.



DG IMPLEMENTATION

DG IMPLEMENTATION CULTURAL ASPECTS

The culture of a company is something that both affects DG success, and can also be affected by DG itself. Specifically:

- Culture affects:
 - Who will govern the data, and in which roles;
 - Which projects and departments are prioritised;
 - The quantity (and type) of resistance you will face;
- Culture is affected by:
 - How literate and committed executives and data owners are, at the end of the day;
 - How trained people are in terms of DG, as well as how frequent communication is (especially about success);
 - How behavior is changed (especially, with change management processes);

DG IMPLEMENTATION CULTURAL ASPECTS

When we're talking about consolidating a corporate culture that cares about data and properly governs data, we need to ensure a couple of main elements:

- That we have DG priorities:
 - It may be PII retention periods, high customer DQ, etc;
- That there are people accountable:
 - There must be owners for every data source/domain;
- That we have enough people and resources:
 - Not overloading people, or mixing DG/DS/DM (or more!);
- That people have the proper skills and knowledge:
 - Proper training, continued, is key, as is data literacy;
- That behavioral change is formally managed:
 - Being able to track adoption of new behaviors;





DG IMPLEMENTATION

DG IMPLEMENTATION CULTURAL ASPECTS

The role of the DG champion (and “bought in” executives) is essential here, because in most cases, culture flows top-down:

- If an executive is fully committed to DG, chances are they will be accountable, allocate enough people, hold those people accountable, set formal processes, track success metrics, force behavioral change, and ensure adoption;
- If an executive is so-so, chances are they will allocate some people, maybe adopt some processes and track success to a point, but they won’t be very accountable;
- If an executive is not committed (or even resisting), chances are they will not allocate enough people, not track behavioral change seriously, and may not even measure success metrics at the end of the day;

DG IMPLEMENTATION CULTURAL ASPECTS

A final, important aspect to mention is that culture becomes especially important outside the defined processes and structure. That is, even if a DG program works perfectly with the current structure, if there are demands beyond that structure, that is when culture shines (or not).

- An example is a regulatory change. Let's say there's a new period for retaining a specific type of data:
 - How soon will this be implemented? Will the data owners come forward to get ahead of the change? Will they be organised? Will executives support them?
- Or, for example, key data people leave the company:
 - Is their knowledge reflected in metadata? Can the team effectively reorganize itself? Can DG be sustained as-is?



DG IMPLEMENTATION CULTURAL ASPECTS EXAMPLES

/01 DG IS JUST A SYMPTOM

In many cases, the way an organisation reacts to DG is just a symptom of something bigger. How does it deal with changes? Do executives resist new initiatives?

/02 TONE AT THE TOP COUNTS

As with many other elements, the way executives deal with something is the way their teams deal with them. So it's crucial to obtain support and buy-in for DG.

/03 CULTURE CAN CHANGE

Regardless of the starting point, with proper data literacy, training and communication, the culture can improve towards adopting and promoting good DG.

DG IMPLEMENTATION CULTURAL ASPECTS KEY TAKEAWAYS

/01 CULTURE AND DG

The company culture has a twin relationship of affecting and being affected by DG. It both defines how people deal with DG, but DG also affects it.

/03 AN EXECUTIVE ORDER

The role of executives is key in culture, because everything flows from them. Whether DG is taken seriously, is staffed, is tracked, and other aspects.

/02 RESOURCES + RESPONSIBLE

Although culture affects DG implementation in many ways, the two key ones are in terms of DG having enough resources + people being responsible (or not).

/04 OUTSIDE THE NUMBERS

Finally, culture is important to codify the behaviors that are not demanded by processes and numbers. Who steps forward? Who reacts? How do they do it?

DATA GOVERNANCE

Learning about how Data Governance works, in terms of both its processes and policies, but also its specific implementation and scaling in an organisation.



DATA GOVERNANCE

We covered **two main topics** in terms of Data Governance:



THE DG FUNCTION

The essentials of what Data Governance is, by whom, and for what purpose.



DG IMPLEMENTATION

How to plan, prepare for, roll out and scale DG across an organisation.



DATA AND DATA QUALITY

DATA AND DQ CONSOLIDATING

Some questions you can ask yourself to consolidate the knowledge in this module include:

- What are the 3 main systems of classification for data?
- What are examples of some functionality present in DG tools?
- What are some activities within metadata management?
- What does a “Trusted” data source mean, compared to a “Pending” one? In terms of both metadata + processing?
- What are some of the three usual types of data stewards, and what does each one do?
- In what ways does culture affect how DG is done within a company?
- What is the difference between a role and a purpose in AC?

DATA SECURITY, PRIVACY, ETHICS

Covering how both data and data subjects are protected in various data disciplines, both in terms of actual InfoSec controls, but also in terms of being treated fairly.



DATA SECURITY, PRIVACY, ETHICS



DATA LITERACY & CONSIDERATIONS

Covering the basics of data. What are the different data disciplines, what are the essential principles to handle data, what are the sophistication levels of organisations...



DATA GOVERNANCE

Covering how data governance works, from classifying data to setting policies and other activities, the roles and responsibilities, and the DG implementation process.



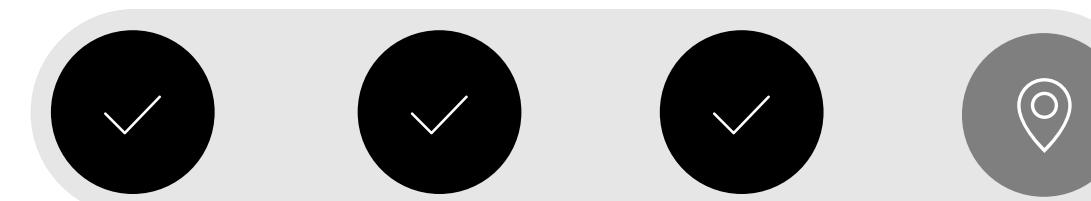
DATA AND DATA QUALITY

Covering specifically what data are and how to improve their quality. Data types, values, structures, and how to improve data quality through profiling and remediating.



DATA SECURITY, PRIVACY, ETHICS

Covering the different types of privacy and security controls that can be applied to data to protect them, as well as how to treat data subjects ethically.





DATA SECURITY, PRIVACY, ETHICS

DATA SECURITY, PRIVACY, ETHICS **GOALS**

Our major goal in this module is to clarify how to protect and treat data ethically. In further detail, we will cover, for example:

- How different governance structures and regional regulation change how protected data are;
- How algorithms can discriminate against - and hurt - data subjects;
- How data can be protected in both logical and physical mediums (encryption, masking, media destruction, etc);
- How security controls are set based on different data sensitivity classes;
- How ethics can be implemented in an organisation, with specific data dimensions, compliance structures and more;

DATA SECURITY, PRIVACY, ETHICS

We will cover **three groups of topics** related to protecting data:



DATA SECURITY

How to protect data with controls such as locking rooms, encrypting data, and others.



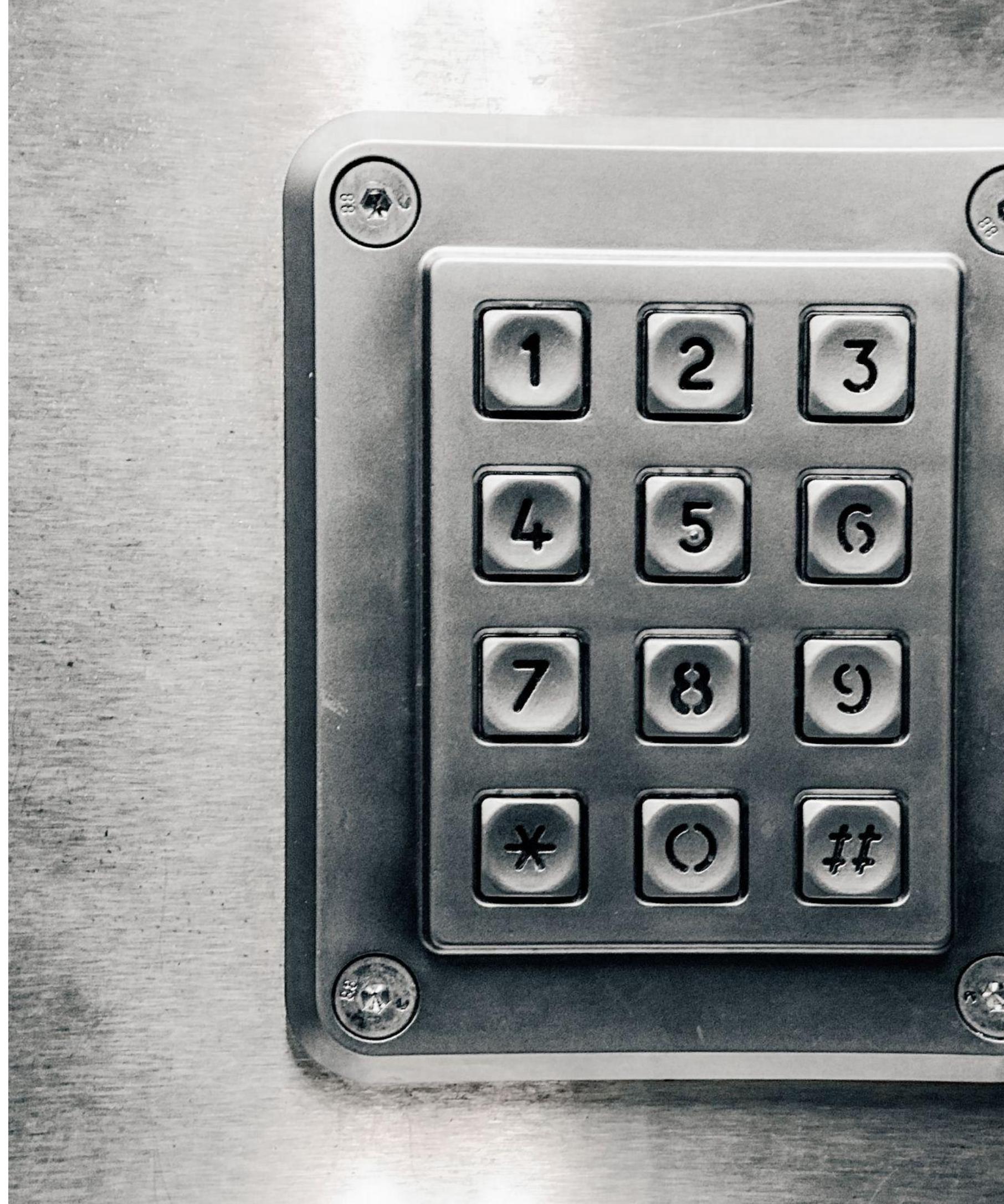
DATA PRIVACY

How data can be protected with data governance structures, controls by data classes, and more.



DATA ETHICS

How data can be treated ethically with transparent algorithms, data purposes, and others.



DATA SECURITY

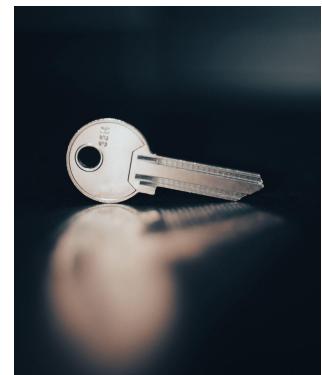
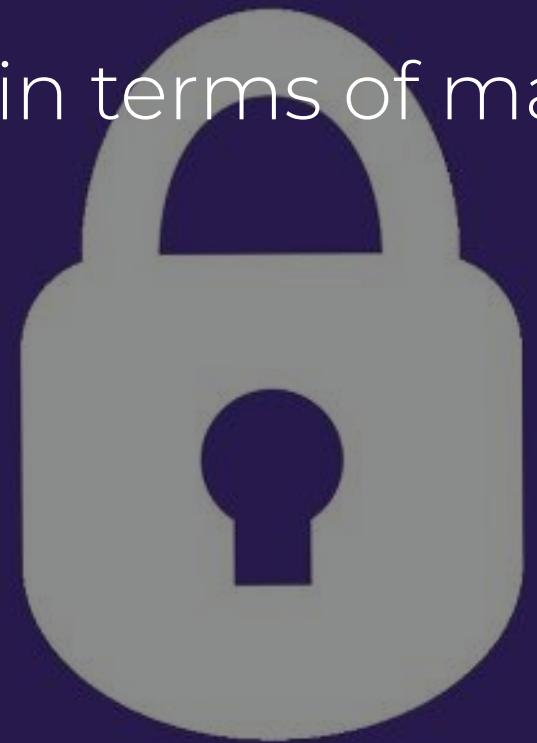
DATA SECURITY INTRODUCTION

Measures to keep data protected are important in the course of any data discipline (DM, DG, DS, or others). These should be applied to both data in physical and logical formats, using a variety of controls.

Despite measures to actively protect data, from locked rooms to encryption and others, it's important to take into account other elements such as the retention period of the data and how they are disposed of, as is being aware of protection at different stages, including protecting data while at rest, while in transit, and while being destroyed.

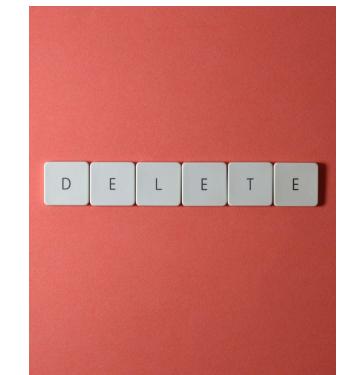
DATA SECURITY

We will cover **five key topics** in terms of maintaining data security:



CRYPTOGRAPHIC PROTECTION

Protecting data with encryption, as well as things to keep in mind.



DATA RETENTION AND DISPOSAL

Having policies for retaining data and when to dispose of them (as well as how).



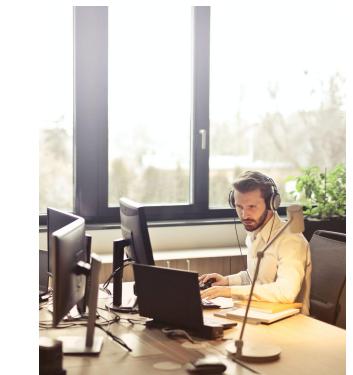
LOCKED ROOMS / DEVICES / PORTS

Locking actual rooms, or devices and ports to create physical obstacles to attackers.



PHYSICAL MEDIA PROTECTION

Protecting physical media at rest, in transit and during destruction, regardless of format.



PROVIDER ASSESSMENT

Initially assessing - and continuously monitoring - service providers.



SECURITY CONTROLS

SECURITY CONTROLS CRYPTOGRAPHIC PROTECTION

Encryption (formally termed cryptographic protection) is used to protect data from attacks - even if an attacker obtains the information, they cannot decrypt it without the proper keys.

- Usually, encryption is both performed for data in transit and for data at rest;
 - The former is done by using strong encryption protocols (SSL, TLS) in transit, whether on WiFi or other mediums, while the latter is done by encrypting information on databases or files;
- Usually, the more sensitive the data, the stronger the encryption used;
 - Confidential or personal data will be protected with much higher strength than public/open data;

SECURITY CONTROLS CRYPTOGRAPHIC PROTECTION

In terms of encryption, two important aspects to take into account are the version of the protocol (in transit) and the management of encryption keys:

- When transmitting information, even secure protocols such as SSL or TLS can be vulnerable if an older version is used (since vulnerabilities are made public). This makes not only the protocol, but the version used the true judge of whether data are protected;
- In terms of encrypting data at rest, the encryption keys must be properly managed. This may mean storing them in protected places (physically or digitally), having a specific cryptoperiod (time after which they are renewed), and possibly split knowledge (2 users, each knows half the key);





SECURITY CONTROLS

SECURITY CONTROLS CRYPTOGRAPHIC PROTECTION

While the encryption of data touches on almost all areas of an organisation, there are two specific controls which are important in relation to it: contingency planning and access control:

- Contingency Planning. In case the encryption keys are stolen or compromised, there must be a process to renew them, isolate the encrypted data, or otherwise protect the stored data;
- Access Control. Users with access to the encryption keys must have much more sophisticated methods of access control, since vulnerabilities in terms of their accounts can be critical;

SECURITY CONTROLS CRYPTOGRAPHIC PROTECTION EXAMPLES

/01 DATA OPERATIONS

It's frequent to encrypt data for the purposes of data science or data management. Someone drawing conclusions from 500 users will see names encrypted.

/02 WEAK POINT: COMMS

In many cases, the weak point are communications. For example, a user accesses highly encrypted data in a protected laptop... and then sends it by email.

/03 THE CRYPTOPERIOD

Just like renewing a password, encryption keys must be changed with some frequency. Given enough time, any encryption can be "cracked".

SECURITY CONTROLS

CRYPTOGRAPHIC PROTECTION

KEY TAKEAWAYS

/01 ENCRYPTING DATA

Encrypting data makes it impossible to decipher by attackers. Data are usually encrypted in transit and at rest, and proportionally to their level of sensitivity.

/02 VERSIONS AND KEYS

Not only is the encryption process important, but the version of the protocol also, due to vulnerabilities, and the management of the keys used in the encryption.

/03 CP/AC ALSO IMPORTANT

Contingency planning is also important, in terms of having an action plan in case of attack, as well as access control for people who deal with the keys.



SECURITY CONTROLS

SECURITY CONTROLS DATA RETENTION/DISPOSAL

One of the most important controls in terms of data privacy and security are the controls for data retention and disposal.

There are two essential distinctions:

- The minimum data possible should be retained. Any data can be considered a vulnerability (they can be stolen or leveraged);
 - The best way to protect data is to not store them in the first place, if possible;
- Data should be disposed of as soon as possible, and through secure means:
 - If we only need to hold Personally Identifiable Information for 1 year, ideally, they should be deleted on day 365 with no delay;

SECURITY CONTROLS

DATA RETENTION/DISPOSAL

Usually, the retention period for data has two opposite forces shaping it. One is the philosophy of deleting data as soon as they are not needed anymore. But on the other hand, there are mandatory retention periods as part of certain regulations.

- For example, in the PCI-DSS certification, credit card data logs (a type of Personally Identifiable Information) must be stored for a minimum of 1 year;
 - This means that these data cannot be deleted before a year passes, but after that, must be deleted ASAP;
- Usually, this can only be optimised by having deep knowledge of data flows and an elaborate data diagram - to know where data are in each step of processing;





SECURITY CONTROLS

SECURITY CONTROLS DATA RETENTION/DISPOSAL

It's important to notice that, as time passes and organisations change, the type of data necessary also changes.

- Certain personal information needed one day may not be needed the next day;
 - Ideally, data diagrams, retention periods and scheduled disposals should reflect that;
 - (E.g., a USA bank extends loans in Europe and must save PII as proof of decisions made. One day leaves the business in Europe. Now what?);

In terms of data disposal, it must be done in a secure way, regardless of whether in logical or physical mediums (e.g. zeroing out logically, de-magnetizing HDDs physically, etc);

SECURITY CONTROLS DATA RETENTION/DISPOSAL EXAMPLES

/01 DECISION DATA COUNT

One of the main types of PII stored is for making decisions. Extending loans, deciding on criminal sentences, etc. Data are kept for some time as proof.

/02 MINIMISING DATA/FIELDS

Minimising data retained is a good practice. But in some cases, even within the same data, we may retain just some fields (E.g., person's name but not birthdate)

/03 REMEMBER DERIVATIVES

Personal data are not just the “raw” data, but any data product made from them. That means reports, logs, dashboards and others containing PII are also PII.

SECURITY CONTROLS

DATA RETENTION/DISPOSAL

KEY TAKEAWAYS

/01 RETENTION AND DISPOSAL

There are two very important distinctions in terms of data: their retention must be minimised, and their disposal must be as soon as possible, and secure.

/03 CHANGING WITH TIMES

As organisations change, the data needed also change, and data retention and disposal should reflect this - the more promptly, the better.

/02 TWO OPPOSITE FORCES

Data retention is shaped by two opposite forces. Certain regulations demand a minimum retention period, and good sense dictates deletion ASAP.

/04 PHYSICAL OR LOGICAL

Whether the data are stored in a physical or a logical format, they must be securely erased in some form, as any other data would be.



SECURITY CONTROLS

SECURITY CONTROLS LOCKED ROOMS/DEVICES/PORTS

In terms of physical protection, a very simple but very effective type of control is to simply prevent access to certain locations, certain servers, certain ports. This prevents intruders from tampering with devices, as they simply cannot reach them.

In specific, we can usually find three types of “locked” elements:

- Locked rooms (preventing unauthorised users from entering them);
- Locked devices (for example, wiring closets, preventing access to servers or routers);
- Locked ports (for example, locked USB/Ethernet ports);

SECURITY CONTROLS LOCKED ROOMS/DEVICES/PORTS

While in the case of locations, the only way to protect them is to prevent access (locked doors, external guards, ID badges, etc), for devices, ports can also be disabled instead of locked.

- In fact, in order to minimise vulnerabilities, a security best practice is precisely to disable all ports which are not used, as they only present vulnerabilities;

Additionally, an important distinction is that inventorying and tracking is important to ensure protection:

- Keeping visitor logs, not just in terms of raw data, but also unusual patterns in actions and people is important to ensure locations are protected;
- Keeping IT inventories is crucial to know what ports exist;





SECURITY CONTROLS

SECURITY CONTROLS LOCKED ROOMS/DEVICES/PORTS

As with other types of security controls, “locking” locations and devices is not just for the purpose of preventing data theft, but also to prevent actual attacks to the integrity of locations and devices. This includes:

- Having measures against floods, power surges, and other environmental hazards;
- Having protection on power cables and sources to prevent power cuts;
- Having emergency lighting in case power to lighting is cut;
- Having emergency exits in case primary exits are compromised;

In this case, these factors are already accounted for when picking the location, and not afterwards.

SECURITY CONTROLS LOCKED ROOMS/DEVICES/PORTS EXAMPLES

/01 WIRING CLOSETS

Wiring closets are usually locked, containing servers or network devices. The idea is that only authorised administrators can unlock them and make changes.

/02 VARIOUS AUTH CONTROLS

In the case of locked rooms, multiple controls can be used, as long as they provide access control. Guards, smart cards, ID badges, SMS codes, and/or others.

/03 GATEWAYS TO MUCH MORE

Physical locations must be protected not only due to the initial attacks that may be caused, but physical access can be a precursor to much bigger attacks.

SECURITY CONTROLS

LOCKED ROOMS/DEVICES/PORTS

KEY TAKEAWAYS

/01 ROOMS, DEVICES, PORTS

Both rooms, devices and specific ports of devices can be locked in order to block access by possible attackers. This prevents tampering with the devices.

/02 LOCK OR DISABLE

In the specific case of device ports, locking them is not the only alternative available, as they can be disabled as well - and should be, if not used.

/03 TRACKING & INVENTORYING

To protect both locations and devices, tracking is crucial. Tracking visitors to find out unusual patterns, and inventorying devices to know what ports exist.

/04 PHYSICAL PROTECTION

Locking rooms and devices is not only a measure against device tampering and data theft, but also integrity attacks, such as to power or lighting.



SECURITY CONTROLS

SECURITY CONTROLS PHYSICAL MEDIA PROTECTION

Physical media can be considered any form of physical storage of data. They can be as sophisticated as servers, HDDs or flash drives, but also just plain paper records.

- Regardless of the format, these must be protected just like logical data;

There are usually three main types of controls regarding physical media protection:

- Protecting media at rest (locked rooms, doors, others);
- Protecting media in transit (tracking transportation, having records, others);
- Protecting media when destroyed (demagnetizing data on flash drives, exploding or burning hard drives, others);

SECURITY CONTROLS

PHYSICAL MEDIA PROTECTION

Controls to properly protect physical media have a lot in common with other types of physical protection:

- Measures such as security guards, identity badges for room access, and others are relevant, in this case to protect data in specific:

Usually, just like logical data are categorised in terms of sensitivity, so are physical media. And marked accordingly.

- Hard drives containing public website files will not be protected by the same controls as hard drives containing credit card data, for example;
- Usually, both the security controls and the access control requirements are proportional to the marked sensitivity;





SECURITY CONTROLS

SECURITY CONTROLS PHYSICAL MEDIA PROTECTION

In specific, physical media protection is usually correlated with safe locations. Some operations may only be performed in certain locations and not others:

- For example, allowing maintenance operations on a hard drive with sensitive data only in a certain room;
- Or allowing USB drives that contain personal information to only be used on-premises, not taken home by employees;
- Or, for example, only allowing the destruction of servers at specific secure locations of a given provider;

In most cases, these secure locations have strict entry and access controls applied, and only allowing specific roles.

SECURITY CONTROLS

PHYSICAL MEDIA PROTECTION

EXAMPLES

/01 DESTRUCT. CERTIFICATES

When destroying media through external providers, one thing that is almost always required is a certificate of destruction, ensuring the medium is not “out there”.

/02 OFFLINE ROOMS

Even with state-of-the-art AV, network connections pose vulnerabilities, so some sensitive data are stored in offline rooms with servers that are “air-gapped”.

/03 EXFILTRATION TOO

While a lot of the security controls focus on preventing intruders from entering locations, it’s also important to focus on data exfiltration - how do they get data out?

SECURITY CONTROLS

PHYSICAL MEDIA PROTECTION

KEY TAKEAWAYS

/01 REST, TRANSIT, DELETION

Physical media must be protected just like logical media. Three main stages are of importance: when at rest, when in transit, and when being destroyed.

/03 MEDIA MARKING

Just like logical data are categorised in terms of how sensitive are the data within, physical media are too, in the form of markings (annotations, labels, or others).

/02 SAME SECURITY MEASURES

A lot of the same security measures as for premises protection are applied - guards, ID badges, visitor logs, etc - but in this case, to prevent access to media.

/04 SAFE LOCATIONS

In specific, in many cases physical media are considered safe in certain locations with heavy controls, and all operations must be performed there.



SECURITY CONTROLS

SECURITY CONTROLS PROVIDER ASSESSMENT/MONITORING

External providers must be initially assessed, when making purchase decisions, but also continually monitored, to prevent both supply chain ruptures but attack vectors.

External providers can be used for many products and services, but important ones include:

- Providers of data storage/processing locations (who you hire a data center from, or locations with other functionality);
 - Providers of cloud storage (when using cloud solutions);
- Providers of systems or parts (who you buy a router or firewall from, or replacement parts);
- Providers of maintenance services (who fixes systems);

SECURITY CONTROLS

PROVIDER ASSESSMENT/MONITORING

Usually, assessing a service provider has to do with assessing their ability to comply with your security/privacy demands.

- For example, if you use MFA to authenticate access to all personal data, and you're considering an external provider to process your personal data, do they use MFA as well, and with the same strength?

Usually, the service provider will have their own SLAs in terms of what they can guarantee;

- Managing their vulnerabilities, internally, is usually the service provider's responsibility, but managing vulnerabilities in the interface between your organisation and them is your responsibility;





SECURITY CONTROLS

SECURITY CONTROLS PROVIDER ASSESSMENT/MONITORING

Passing the initial assessment and having a contractual relationship, as a service provider, does not exclude them from continuous evaluation. In fact, many factors can change:

- The provider may identify vulnerabilities in their systems/software/platforms, and it's important to understand how they deal with them;
 - (E.g., your cloud provider detects a vulnerability that may have affected your account. Now what?);
- The provider may change their supply chain, which may create supply chain risks for your organisation itself
 - (E.g., they change manufacturers, and now router replacement parts do not arrive on time anymore. Now what?);

SECURITY CONTROLS PROVIDER ASSESSMENT/MONITORING EXAMPLES

/01 RUPTURE OR ATTACK

Provider problems usually come in two types. Rupture or attack. That is, either they can't provide products or services, or these are attacked by malicious actors.

/02 EMBEDDED IN SCRM

The discipline of Supply Chain Risk Management is relatively new, but supplier assessments are crucial to it, in order to identify your own SC vulnerabilities.

/03 ETHICAL CONSIDERATIONS

Although supplier assessment is mostly focused on technical qualities and flaws, ethical problems may cause firing a provider (e.g., using modern slavery).

SECURITY CONTROLS

PROVIDER ASSESSMENT/MONITORING

KEY TAKEAWAYS

/01 DATA, SYSTEMS, SERVICES

External providers usually are used for data storage or processing, for system or system part purchases, or for services such as maintenance.

/02 A COMPLIANCE MATTER

Usually, selecting a service provider boils down to whether they can comply with the same security controls as your own organisation or not.

/03 ATTACKS OR RUPTURES

Usually, external provider failures come down to two situations - attacks, through exploited vulnerabilities, and supply chain ruptures, affecting deliveries.



DATA PRIVACY

DATA PRIVACY INTRODUCTION

Protecting the privacy of data is essential regardless of data discipline. And, in most cases, the classes of data requiring the highest level of privacy will be sensitive data classes, namely personal data.

There are many elements that define the privacy of data and how well they are protected, from the overall Data Governance structures and/or the regions the organisation is present in, to more operational actions and controls such as downgrading/redacting of physical media with sensitive data, or the de-identification of datasets themselves.

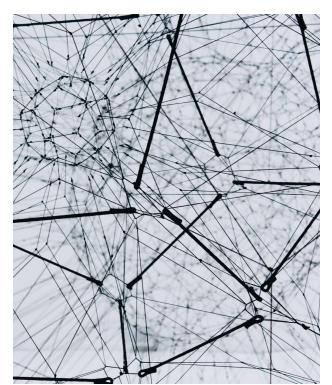
DATA PRIVACY

We will cover **five key topics** related to data privacy:



GEOGRAPHICAL REGULATION

How regulation in different countries affects data privacy for multinational organisations.



DATA GOVERNANCE STRUCTURES

The specific structures that can be set in place to govern data, and how they work.



CONTROLS BY DATA CLASSIFICATION

Defining InfoSec controls by data classification, usually related to sensitivity levels.



MEDIA DOWNGRADING

How media downgrading works - removing sensitive data to reduce the media's sensitivity class.



DATA DE-IDENTIFYING / ANONYMISATION

Removing specific personal data from datasets in order to allow them to be processed with no risk.



DATA PRIVACY

DATA PRIVACY GEOGRAPHICAL REGULATION

Organisations which are present in multiple countries or continents are bound to face different types of privacy regulation. This determines several factors:

- Whether the company is present or not in a specific geography;
 - E.g. US companies leaving Europe due to GDPR;
- Whether the data subjects are protected or not in that given geography;
 - E.g., Apple turning off VPN / Google censoring in China;
- Whether the data themselves are secure in that given geography;
 - E.g., having data centers in countries that turn over data to their government;

DATA PRIVACY GEOGRAPHICAL REGULATION

Regulation in various countries or continents can affect the company's reputation, mostly due to how data subjects are treated, but can be promising revenue opportunities:

- A company is considering entering a country with censorship. Should we comply, making money? Or risk reputational harm due to complying with censorship?
- A company uses data centers in countries which may be hostile to their “mother” country. Should we risk it for the cost savings? Do we risk being sanctioned by our country?

It's important to note that regulation does not just apply to the organisation's local data but also external data centers (e.g., complying with GDPR means your external data center vendor also complies with GDPR - otherwise you fail);





DATA PRIVACY

DATA PRIVACY GEOGRAPHICAL REGULATION

In many cases, data subjects can be harmed without even being aware of this:

- For example, certain countries demand companies to provide end user data on demand, for example as part of criminal investigations;
 - The company may do it without the user ever knowing;
- Certain countries may actually have direct access to data centers and providers;
 - Organisations dealing with these providers may not even know all their data can be accessed by that government on demand;
- And of course, in some cases organisations themselves sell data to external third parties without users being aware;

DATA PRIVACY GEOGRAPHICAL REGULATION EXAMPLES

/01 CENSORING COUNTRIES

One of the biggest ethical dilemmas for many Western companies has to do with entering countries with censorship. What would we need to sacrifice?

/02 CAMBRIDGE ANALYTICA

The Facebook and Cambridge Analytica case is a great example of how companies, without regulation against it, can sell personal data with no one knowing.

/03 HQs VS. ENTITIES

Whether a company (e.g., service provider) is registered in a country or has physical offices there is an important factor. Can we accept this country?

DATA PRIVACY GEOGRAPHICAL REGULATION KEY TAKEAWAYS

/01 GEO AFFECTS REGULATION

Companies present in multiple geographies must make concessions. Different countries may have demands, not protect data, or actively harm subjects.

/02 PRESENCE & PROTECTION

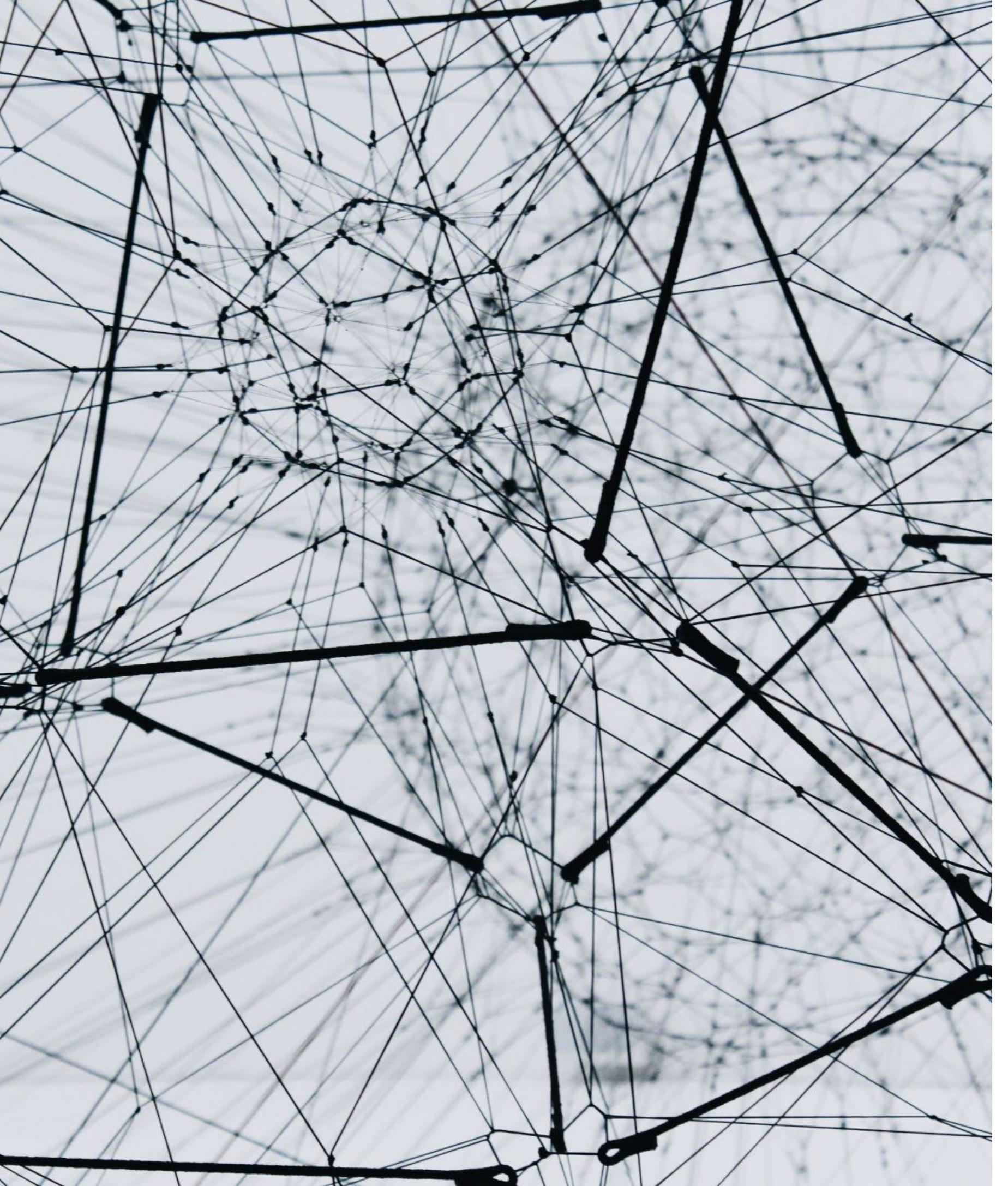
Usually, countries with hostile alignments or censorship define two things. Whether a company can be present there, and whether data can be safe.

/03 REPUTATIONAL TRADE-OFFS

Being present in questionable countries with questionable regulation is usually a trade-off. Do we choose to make money, or protect data subjects?

/04 NONE THE WISER

In many cases, unfortunately, data subjects are not even aware that their data may not be protected, or may be used against them behind their backs.



SECURITY CONTROLS

SECURITY CONTROLS DATA GOVERNANCE STRUCTURES

There may be one or more governance structures for the handling of data in an organisation. This accomplishes several purposes:

- It allows decisions to flow from the executive level;
- It establishes accountable executives for every team, department or data source;
- In the case of multiple boards, it allows each to focus on specific problems (ethics, data integrity, others);

Usually, there are three main types of data-related boards:

- Data Governance Board (directs all organisational DG);
- Data Ethics Board (focuses on ethical data processing);
- Data Integrity Board (focuses on data matching/sharing);

SECURITY CONTROLS DATA GOVERNANCE STRUCTURES

Usually, all of these structures relate to Data Governance (DG).

This is a data discipline that focuses on policies and processes for data, for several possible purposes:

- To track the lineage of data from creation to disposal;
- To make sure that retained data are minimised and that they are promptly disposed of;
- To make sure that all data are classified, in order to prioritise security controls and Access Control to them;
- To bridge technology and business, by deriving technical data validity rules from business requirements;
- Possibly, many others;

Data Governance (DG) should not be mistaken with DM (Data Management), which is about the actual data management.





SECURITY CONTROLS

SECURITY CONTROLS DATA GOVERNANCE STRUCTURES

In a bigger organisation, it's common to have a hierarchical structure to boards, with a top-level steering committee and one board per department/region/team.

- For example, a big, multinational company may have a top-level Data Governance Steering Committee, where all “global” decisions about policies and processes are made;
- Then, there may be three Data Governance Councils per continent (for each the company is present in). One for EMEA, one for APAC and one for AMER;
 - Within each of these, each continental branch office may have individual data councils per department (APAC sales, APAC marketing, etc);
- Decisions cascade from steering committee to councils;

SECURITY CONTROLS DATA GOVERNANCE STRUCTURES EXAMPLES

/01 ONE OR MORE

In smaller companies, a single Data Governance Council or Board is more than enough. But as companies grow in size, more become necessary.

/02 GOVERNANCE x PRIVACY

Data Governance (DG) and data privacy go hand in hand. The better data are classified, and the more their lineage is tracked, the easier controls are to apply.

/03 DG INTEGRATION

One of the major reasons for a formal DG body is to integrate DG into the organisation. In many cases, DG metrics are obtained from privacy controls themselves

SECURITY CONTROLS DATA GOVERNANCE STRUCTURES KEY TAKEAWAYS

/01 GOVERNANCE BOARD(S)

Usually, DG in a company comes from one or more Data Governance Boards. In many cases, there are additional boards for ethics and for data integrity.

/02 ALL DG PURPOSES

The DG board establishes organisation-wide directives for all facets of DG, including data classification, lineage tracking, access controls, data disposal and more.

/03 THE COUNCIL OF COUNCILS

If an organisation is big enough, instead of one DG council, there may be a top-level one (usually a steering committee), and many smaller, local ones.



SECURITY CONTROLS

SECURITY CONTROLS CONTROLS BY DATA CLASSIFICATION

Not all data are created the same, and they have different levels of sensitivity. The more valuable the data, the more they must be protected.

- Usually, data are categorised, and security controls are applied with intensity proportional to their value;

Data classification affects controls in different categories:

- The level of physical protection (locked rooms, offline servers, guards and visitor logs, others);
- The level of logical protection (encryption, masking, hashing, tokenising, anonymisation, others);
- Security in both storage and in transit (encrypted databases, encrypted communications, device CM, etc);

SECURITY CONTROLS

CONTROLS BY DATA CLASSIFICATION

Usually, there are specific sensitive categories of data that required added controls:

- PII (Personally Identifiable Information). Deeply personal information, which may include health information or credit card information;
- Privileged information. Sensitive business information, such as lawyer or psychologist client data;
- Classified information. Military, intelligence agency, governmental or other organisations' information;

Usually, due to its importance, sensitive information is also under the effect of regulation (e.g. patient data under HIPAA, credit card data under the PCI-DSS, among others).





SECURITY CONTROLS

SECURITY CONTROLS CONTROLS BY DATA CLASSIFICATION

For the most sensitive types of information, we usually see the application of two Access Control techniques: multifactor authentication and the Principle of Least Privilege:

- Multifactor authentication. To access credit card data, for example, the user may need to physical connect to a network within a locked room, using strong WiFi authentication, and to access the room, they may need to scan a badge, receive an SMS token and/or pass a guard;
- Principle of Least Privilege. Each user has access to the minimum possible data, and complex operations involve individual “blind” users. For example, to modify patient data and delete wrong records, one user role can only modify them, and the other role can only delete them;

SECURITY CONTROLS CONTROLS BY DATA CLASSIFICATION EXAMPLES

/01 PCI-DSS

The PCI-DSS are a great example. CHD (cardholder data) is on the front of the card, and secure. SAD (sensitive authentication data) cannot even be stored.

/02 BACKUPS COUNT, TOO

Not only are the “base” data protected by security controls, but also backups. Certain regulations demand storing data up to 1 year - must be protected!

/03 MEDIA CONTROLS

It's very frequent to have security controls on physical media with sensitive data. USB flash drives, hard drives, servers and others. In storage and in transit.

SECURITY CONTROLS

CONTROLS BY DATA CLASSIFICATION

KEY TAKEAWAYS

/01 DIFFERENT SENSITIVITIES

Data are usually classified based on sensitivity level. The intensity of the security controls is proportional to this - the most sensitive data are the most protected.

/03 HEALTH, FINANCE, MILITARY

Usually, the most sensitive types of data are personal data, including health and financial information, and classified data, including governmental and military.

/02 ALL DIMENSIONS APPLY

Sensitive data must be protected in all ways possible. Both in the logical and the physical worlds, and both at rest and in transit - to prevent all vulnerabilities.

/04 MFA AND PoLP

The two most common types of controls for sensitive data are multifactor authentication (logical or physical) and the principle of least privilege.



SECURITY CONTROLS

SECURITY CONTROLS MEDIA DOWNGRADING/REDACTING

Since access control to media with data is usually based on the sensitivity level of the data contained, this results in certain confidential or sensitive information not being viewable by almost everyone.

- In such cases, it's possible to remove some of the sensitive aspects, to "downgrade" the sensitivity classification of information, and allow it to reach a wider audience;
- On paper, downgrading usually occurs as redactions; This control has some elements in common with sanitisation. In sanitisation, you remove all information on a medium in order to void the information (destroying paper, securely erasing digital data, etc):
 - In downgrading, we only sanitising the confidential part;

SECURITY CONTROLS

MEDIA DOWNGRADING/REDACTING

In order for downgrading to be effectively used, it must be a structured process that fulfils a set of criteria:

- The sensitivity level of the information contained in the medium must be verified before and after the downgrading;
- The specific medium must have its marking updated to reflect the new sensitivity classification;
- The systems or media that have their sensitivity changed must be updated in inventories;

Rigor in this control is important, but becomes even more important when the downgraded medium is to be shared outside the organisation - being much more vulnerable.





SECURITY CONTROLS

SECURITY CONTROLS MEDIA DOWNGRADING/REDACTING

The downgrading process goes well with - and should be reflected in - the Access Control controls in existence, regardless of what specific controls they are:

- Re-marking a USB drive with a different label/marking to reflect the downgrading;
- Re-labeling folders in a local network, or emails in an account to reflect the downgrading;
- Changing physical room access controls to the specific server/computer to reflect the downgrading;

Media downgrading can be considered a specific instance of a more general situation - when data sensitivity changes, access to them should be immediately changed accordingly.

SECURITY CONTROLS

MEDIA DOWNGRADING/REDACTING

EXAMPLES

/01 DON'T LET THEM DEDUCE

When downgrading media, it's important to make sure that the information that is left is not enough to extrapolate more data. It may be between the lines.

/02 BLACKED OUT REPORTS

It's common to see in movies or TV Shows FBI or CIA reports that have "blacked out" sections. This is an example of redacting, to share them with a public.

/03 DE-IDENTIFICATION

In specific, when we want to process personal information without identifying the person, we remove personal elements (de-identify) from them.

SECURITY CONTROLS

MEDIA DOWNGRADING/REDACTING

KEY TAKEAWAYS

/01 BEING LESS SENSITIVE

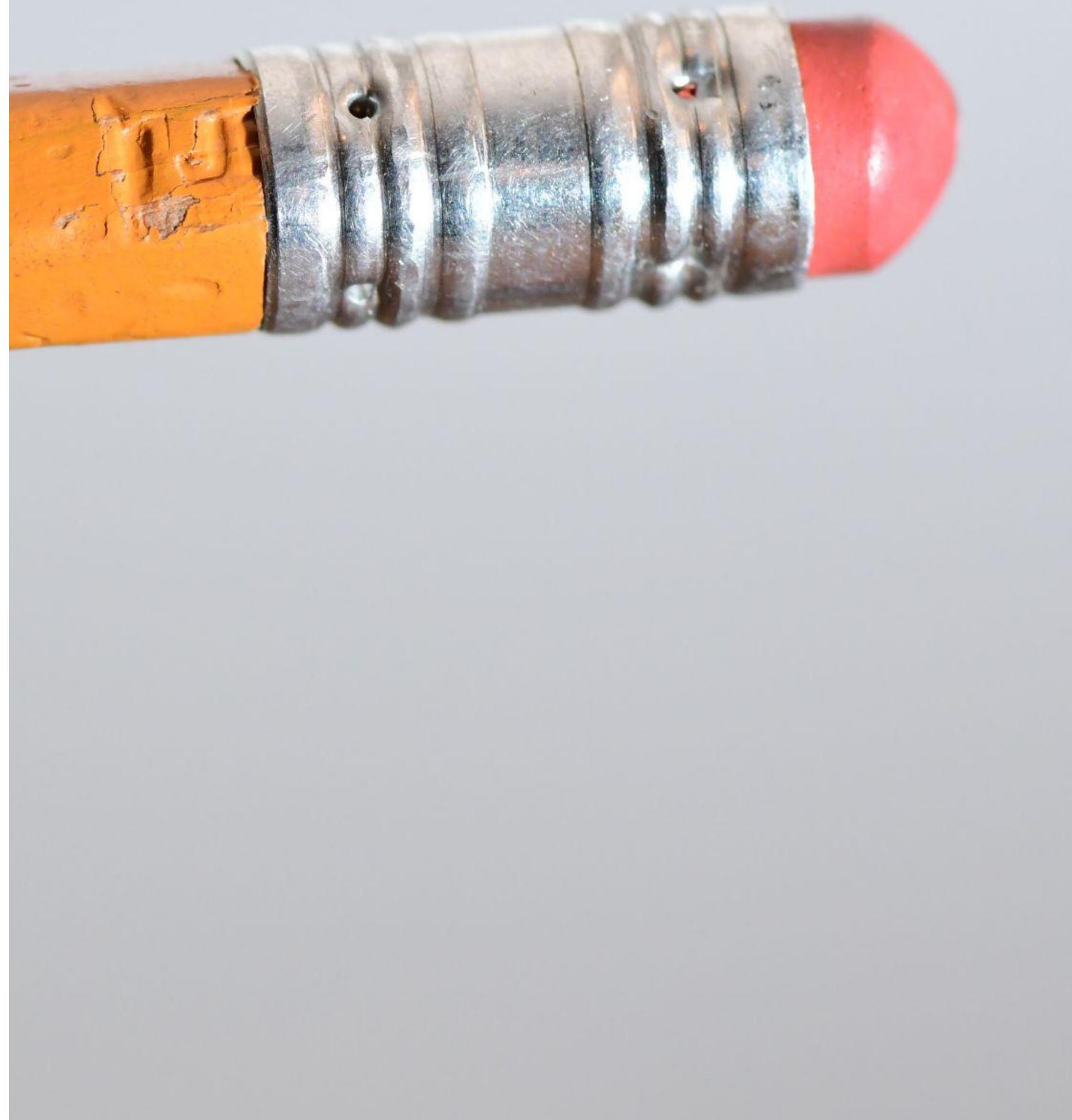
Downgrading (or redacting, on paper), consists of removing information from a medium in order to decrease its sensitivity classification, allowing sharing.

/02 VERIFY AND UPDATE

The downgrading process must be a formal process, with verification of the new sensitivity levels and updates on inventories and markings.

/03 REFLECTED ON AC

Since access to sensitive media is already done through AC controls, these must be updated to reflect the downgrading, be it in whatever form they exist.



SECURITY CONTROLS

SECURITY CONTROLS

DATA DE-IDENTIFICATION/ANONYMISATION

Data anonymisation, or de-identification is the process of changing personal information in datasets.

- For example, if we have a dataset of 5000 people with personal addresses and credit card numbers, we may remove the addresses and credit card numbers - and possibly, also abbreviate the names or convert them to initials;

Anonymisation is useful when we need to work on datasets with personal information - either to profile them for data quality operations, or to draw conclusions as part of data science projects.

SECURITY CONTROLS

DATA DE-IDENTIFICATION/ANONYMISATION

There is a variety of techniques that can be used to anonymise data, but these usually include:

- Masking. This is the replacement of a value (or a part of it) with useless data to mask it. For example, masking all digits of a credit card but the last 4 (XXXX-XXXX-XXXX-4313);
- Tokenisation. Replacing data with a “token” version that doesn’t make sense externally. For example, replacing “John Smith” with “Person431”. People working on the data only see “Person431”, and do not know how that is;
- Noise insertion. Inserting random numbers or data to “disturb” the original data. For example, replacing house numbers or street names in a list of addresses to make addresses useless;





SECURITY CONTROL

SECURITY CONTROLS

DATA DE-IDENTIFICATION/ANONYMISATION

Besides changing the original data, which is very useful for privacy but can be costly, we can also create what are called synthetic data. These are data that are not real, but created by an algorithm, and are similar to real data;

- For example, taking a list of 50 customers from a certain area with certain consumer patterns, and generating 50 “fake” customers which are highly realistic, from a very similar area and very similar consumer patterns;

Usually, statistical methods are used to ensure that the synthetic data are similar to the original data. That is, that means are the same, the maximum and minimum values are the same, and so on.

SECURITY CONTROLS

DATA DE-IDENTIFICATION/ANONYMISATION EXAMPLES

/01 ONLY ONE TECHNIQUE

Data anonymisation is only one possible privacy control. There are many more, such as encrypting data, enforcing AC, disposing of unneeded data, etc.

/02 SEPARATING ALSO WORKS

If you can't mask all data, separating it works. For example, customers with purchase history + CC data. Some users only see the CCs. Some only purchases.

/03 DATA ROLES MATTER

Data anonymisation is very important for data operations (data management, data science, etc). Data anonymisation also depends on the role's AC.

SECURITY CONTROLS

DATA DE-IDENTIFICATION/ANONYMISATION

KEY TAKEAWAYS

/01 REMOVING PERSONAL DATA

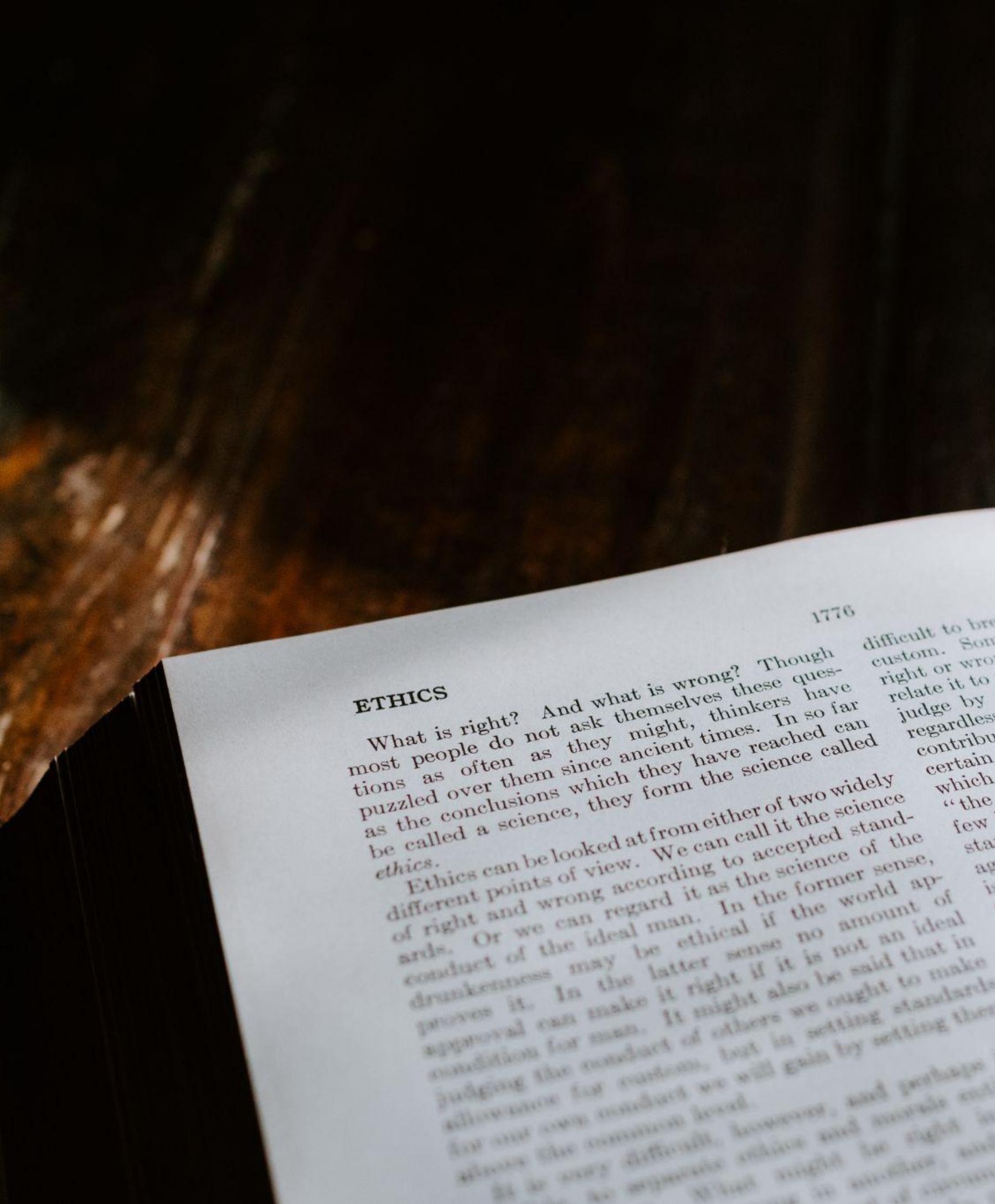
Data de-identification or anonymisation is, unsurprisingly, about removing personal information from data, allowing them to be processed safely.

/02 MASKS, TOKENS, NOISE

There are many ways to anonymise data, but the three major ones are masking (hiding data), tokenising (replacing w/ different data) or inserting noise in them.

/03 SYNTHETIC DATA

An alternative to altering the current data is to actively create new data, which have similar values and patterns, but does not represent any real person.



DATA ETHICS

DATA ETHICS INTRODUCTION

Privacy and security should not be the only concerns when it comes to data. It is completely possible to keep data secure and private, but still harm data subjects. Companies make choices on whether to treat data subjects ethically or not.

We will cover both the general framework of how to implement ethical data handling in a given organisation, but will also go deep on details, such as what criteria make algorithms/processes unfair, or what data dimensions can be used to specifically measure ethics, among other topics.

DATA ETHICS

In terms of ensuring the ethical processing of data,
we will cover **five specific topics**:



THE “COMPLIANCE” APPROACH

Considering ethics a parallel to compliance - how to comply, and what the consequences are.

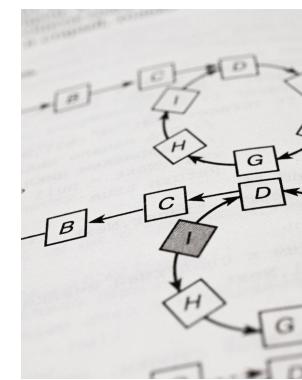
IMPLEMENTING ETHICS

How to actually implement an ethical framework to govern data in an organisation.



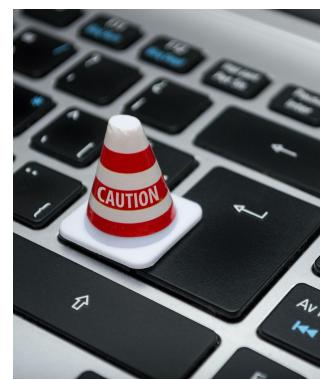
ALGORITHMS AND PROCESSES

How algorithms and processes can be unfair or discriminatory and harm data subjects.



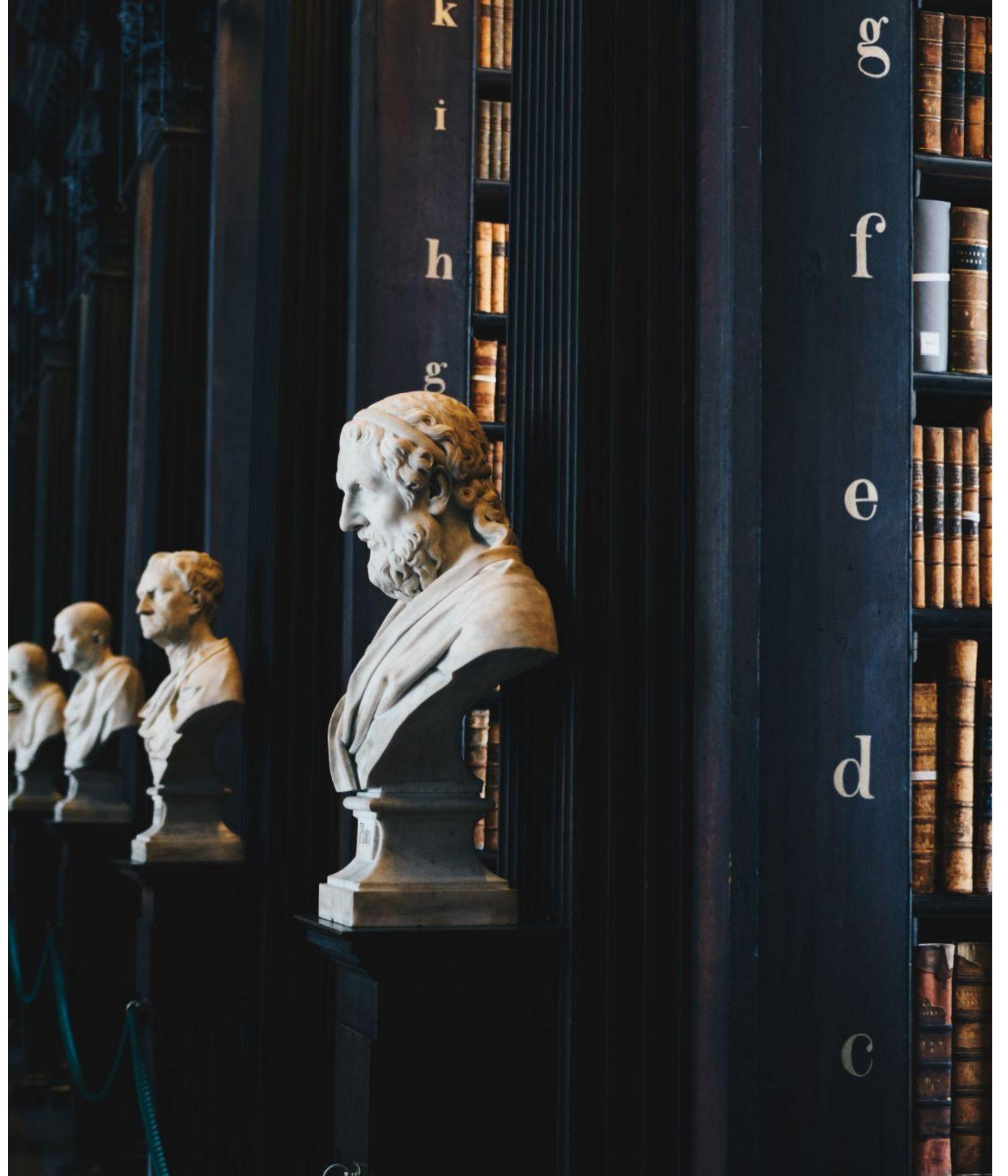
ETHICAL DATA DIMENSIONS

Defining specific data profiling dimensions for ethics, and how to measure them.



DATA USAGE PURPOSE/AUTHORITY

Defining specific purposes and authority for processing data for given roles.



DATA ETHICS

DATA ETHICS

THE “COMPLIANCE” APPROACH

An ethical approach to data (whether it's DM, DG or other disciplines) should consist of safeguarding data subjects and using data for the good of them (and society in general).

- There are usually both user expectations in terms of ethics (“I expect my data to not be used for marketing purposes without authorisation”) and society expectations (“No social media platform should actively promote hate speech”);

One of the best analogies to ethical data processing is considering it a type of compliance (such as with regulation):

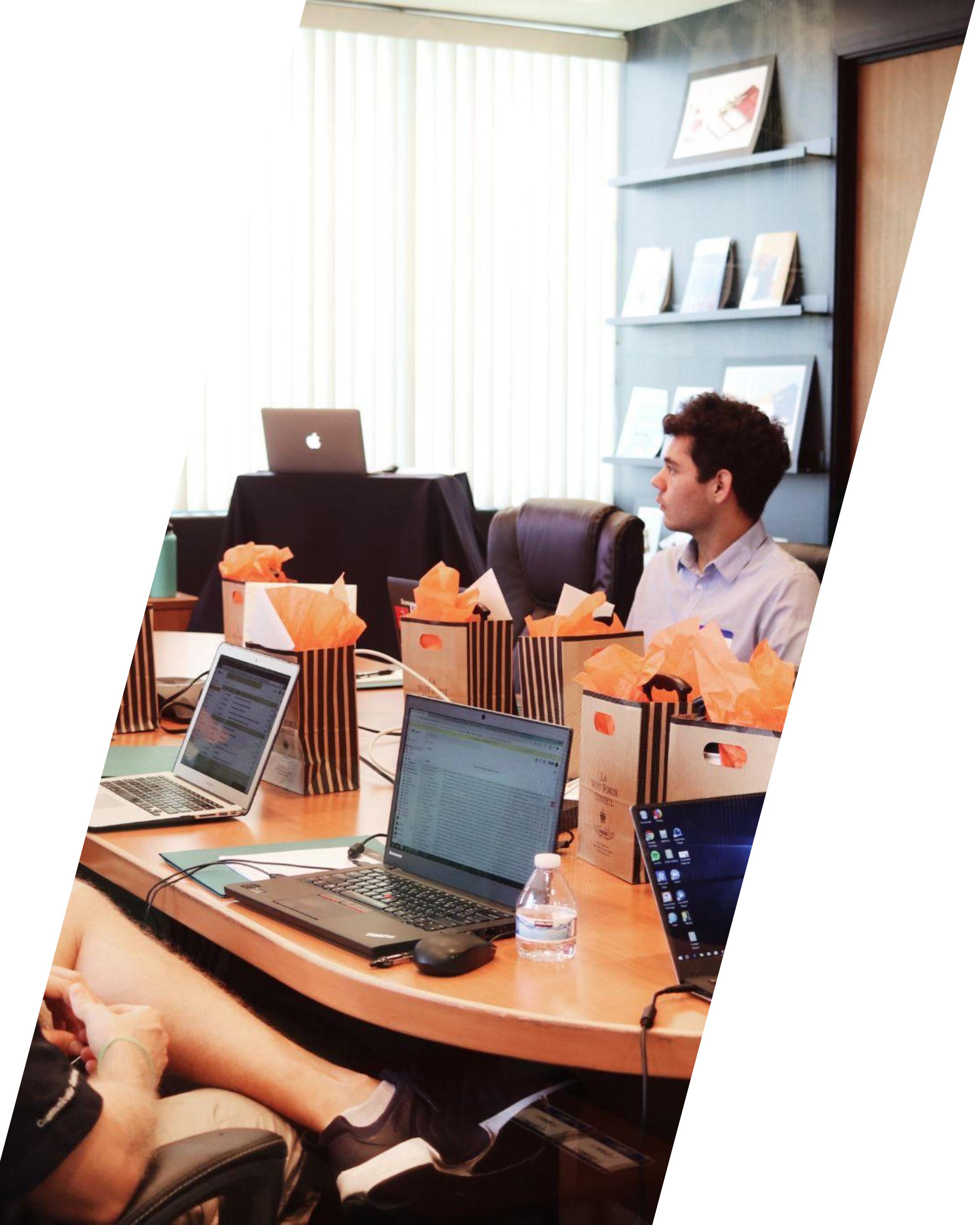
- In many cases, compliance consists of obeying certain rules and quality standards. Otherwise, you are fined/sanctioned;
- Ethical treatment of data can be considered the very same;

DATA ETHICS

THE “COMPLIANCE” APPROACH

Likewise, considering ethical data processing as a type of compliance, we realise there are several keys to a successful implementation:

- Staff must be trained and willing to comply with ethical standards (if they don't know how to do it, or the company culture doesn't care, they won't act in an ethical manner);
- Tone at the top matters (just like if top executives do not care about breaching compliance and being fined, their teams won't care either, the same happens with ethics);
- The company's “compliance appetite” must be defined (some companies do not even go near dangerous territory, others venture into “gray areas”, and others actively breach compliance and deal with fines - the same with ethics);





DATA ETHICS

DATA ETHICS

THE “COMPLIANCE” APPROACH

Perhaps more importantly, considering data ethics a type of “compliance” makes it a quantifiable, measurable metric:

- Just like treating “data” as assets, we also treat “ethical data treatment” as assets;
- We can measure metrics such as the quantity of ethical incidents (reported by employees or users), the quantity of problems raised (and what was done about it), the quantity of revenue or clients lost due to unethical behavior, and others;
- We can also quantify how trained employees are to treat data ethically, as well as how many incidents occur due to both internal and external causes;
- All of this allows us to quantify data ethics;

DATA ETHICS

THE “COMPLIANCE” APPROACH

We can take this correlation further, as in some cases, lack of data ethics actually causes breaches of compliance:

- Improperly retaining or using private customer data can result in breaching data privacy compliance, (e.g. GDPR);
- In the banking industry, discriminating individuals by using skewed models when providing loans (and with different interest rates) can result in fair lending compliance breaches;
- Not disposing of personally identifiable information (PII) after the organisation should, and suffering a data breach which reveals it can result in breaching data privacy compliance;



DATA ETHICS

THE “COMPLIANCE” APPROACH

EXAMPLES

/01 BREACH AT YOUR OWN RISK

Just like with the different types of compliance, we can consider that data ethics consists of either obeying certain standards, or risking consequences.

/02 QUANTIFYING ETHICS

As seen, the major advantage of this comparison is making data ethics quantifiable. How many incidents? How many untrained people? How many lost clients?

/03 ALGORITHMS AT FAULT

As we'll see, although it's possible to not be ethical in terms of the stored data, in most cases algorithms and processes are the ones that cause damage to people.

DATA ETHICS

THE “COMPLIANCE” APPROACH

KEY TAKEAWAYS

/01 JUST LIKE COMPLIANCE

Treating data ethics like legal compliance is a good analogy. You have a set of standards to comply with, and not complying may mean severe consequences.

/02 CULTURE + SKILLS COUNT

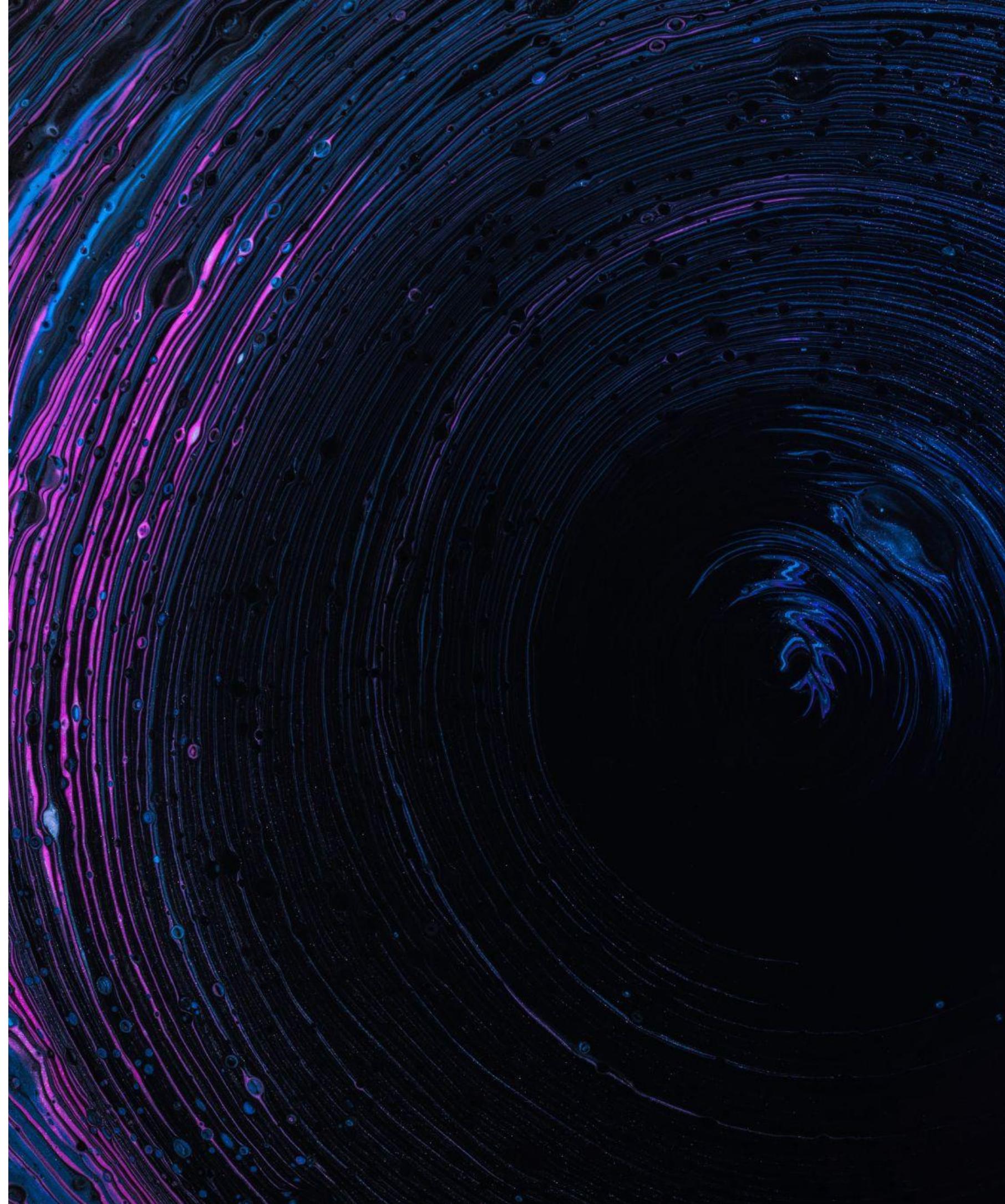
As with compliance, the company's culture and approach towards compliance, as well as the skills of employees both define how compliant the company is

/03 COMPLIANCE QUANTIFIES

A great advantage of treating data ethics like a form of compliance is that it becomes quantifiable. How many incidents? What consequences? Who is untrained?

/04 ACTUAL COMPLIANCE, TOO

In many cases, lack of data ethics actually leads to a lack of compliance as well, as a lot of regulation is made precisely to prevent unethical treatment of data.



DATA ETHICS

DATA ETHICS IMPLEMENTING ETHICS

A proper implementation of ethical data treatment in a company actually has a lot of components in common with DG. Namely, there are three important ones:

- The existence of an ethical governance body:
 - A set of executive-level individuals who make decisions regarding how data are ethically treated;
- The existence of ethical policies and processes:
 - The only way to assure ethical treatment of data at all levels is by making it mandatory, repeatable processes;
- The existence of a feedback loop:
 - If individuals detect ethical problems within a company, they must have a way to escalate them or feed them back to upper levels;

DATA ETHICS IMPLEMENTING ETHICS

Having an ethical governance body is crucial in order to assure implementation of ethical data practices across the organisation.

- Just like a council or steering committee for DG, for example, it also decides on the ethical principles/policies that are then implemented at all levels of the organisation;

Usually, ethics will be embedded into the current principles:

- Accountability and ownership (by having data owners that are held accountable, they are more likely to be ethical);
- Data as assets (only by managing, securing and measuring data as assets can they be treated ethically);
- Compliance (ethics as type of compliance to be enforced);





DATA ETHICS

DATA ETHICS IMPLEMENTING ETHICS

Specifically, in the case of ethics, the existing principles and policies may be tailored to have the data subject in mind. For example:

- Considering that data are assets not of the company, but of the subject, which the company is only temporarily using - and that must be safeguarded;
- Considering data must serve the person. That is, they should only be used for the good of the subject, and never for purposes that may harm them;
- Only considering ethics implemented if consolidated into processes. Just like DG, talking about it means nothing. Ethics are consolidated into processes, systems, behaviors respectful of data subjects;

DATA ETHICS IMPLEMENTING ETHICS

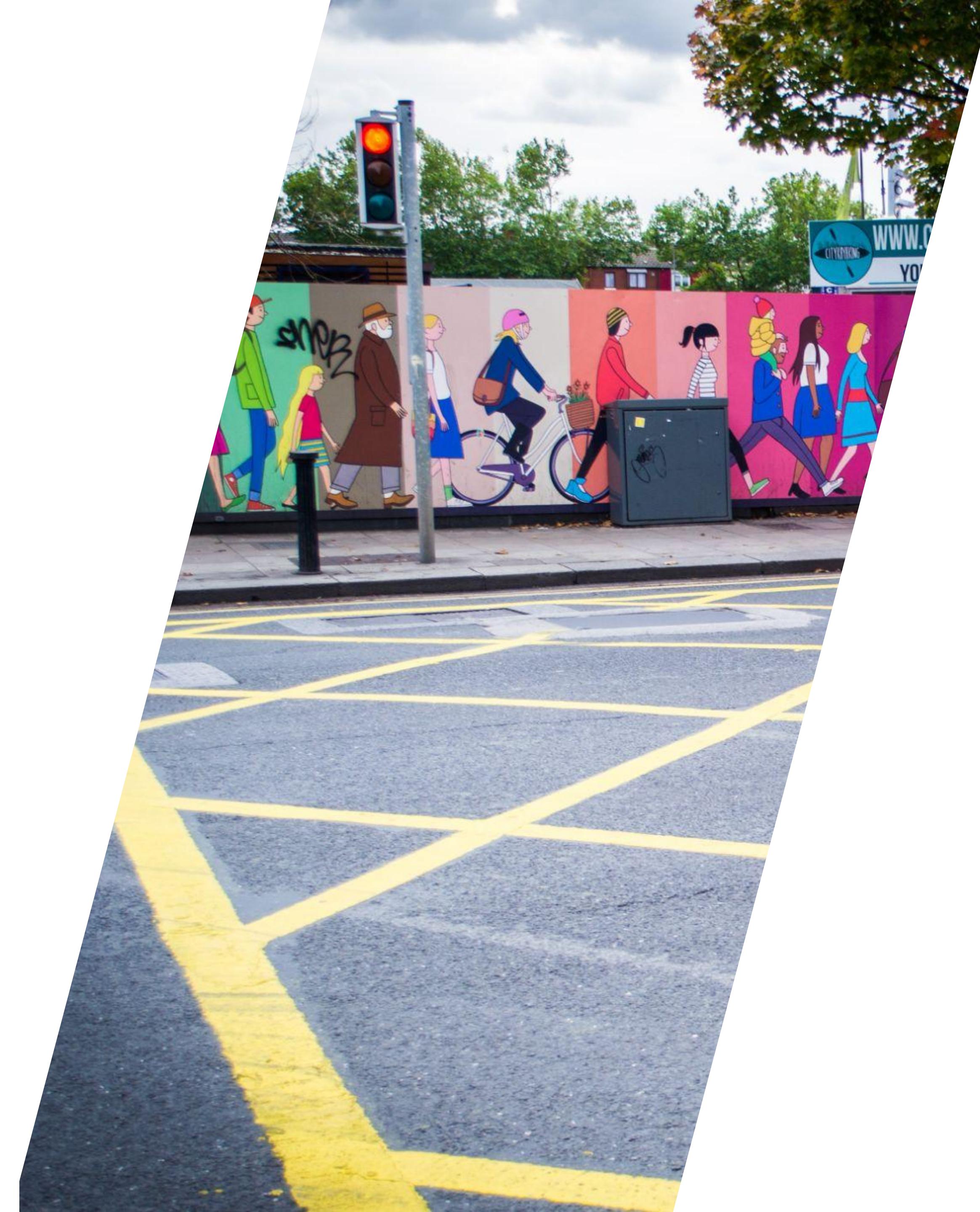
Data ethics expert Katherine O'Keefe considers there are three layers of ethical expectations that must be aligned.

Societal, organisational and individual expectations:

- Societal: what society expects the company to do;
- Organisational: what the organisation expects itself to do;
- Individual: what the employee expects the organisation to do;

Problems occur when these three are not in alignment:

- Organisational \neq societal expectations (the company doesn't behave how society expects, ethically);
- Individual \neq organisational expectations (the company doesn't behave as individual employees expect it to);



DATA ETHICS IMPLEMENTING ETHICS EXAMPLES

/01 DOWNLOAD AND REMOVE

The ability of a data subject to download their data - or request removal of it - from any organisation holding them is an example of considering them their owner.

/02 “ETHICAL APPETITE”

Different company cultures have a different ethical tolerance. Shady marketing, discrimination and other behaviors can be blocked (or not) by ethical policies.

/03 SOCIAL MEDIA PLATFORMS

Perhaps the most notorious example of unethical data treatment. Not transparent to end users, not helping their well-being, taking advantage of them...

DATA ETHICS IMPLEMENTING ETHICS KEY TAKEAWAYS

/01 3 MAIN ELEMENTS

We can consider 3 elements as crucial to implementing ethical data treatment: a governance body, processes and policies, and a feedback loop.

/03 PRIORITISING THE SUBJECT

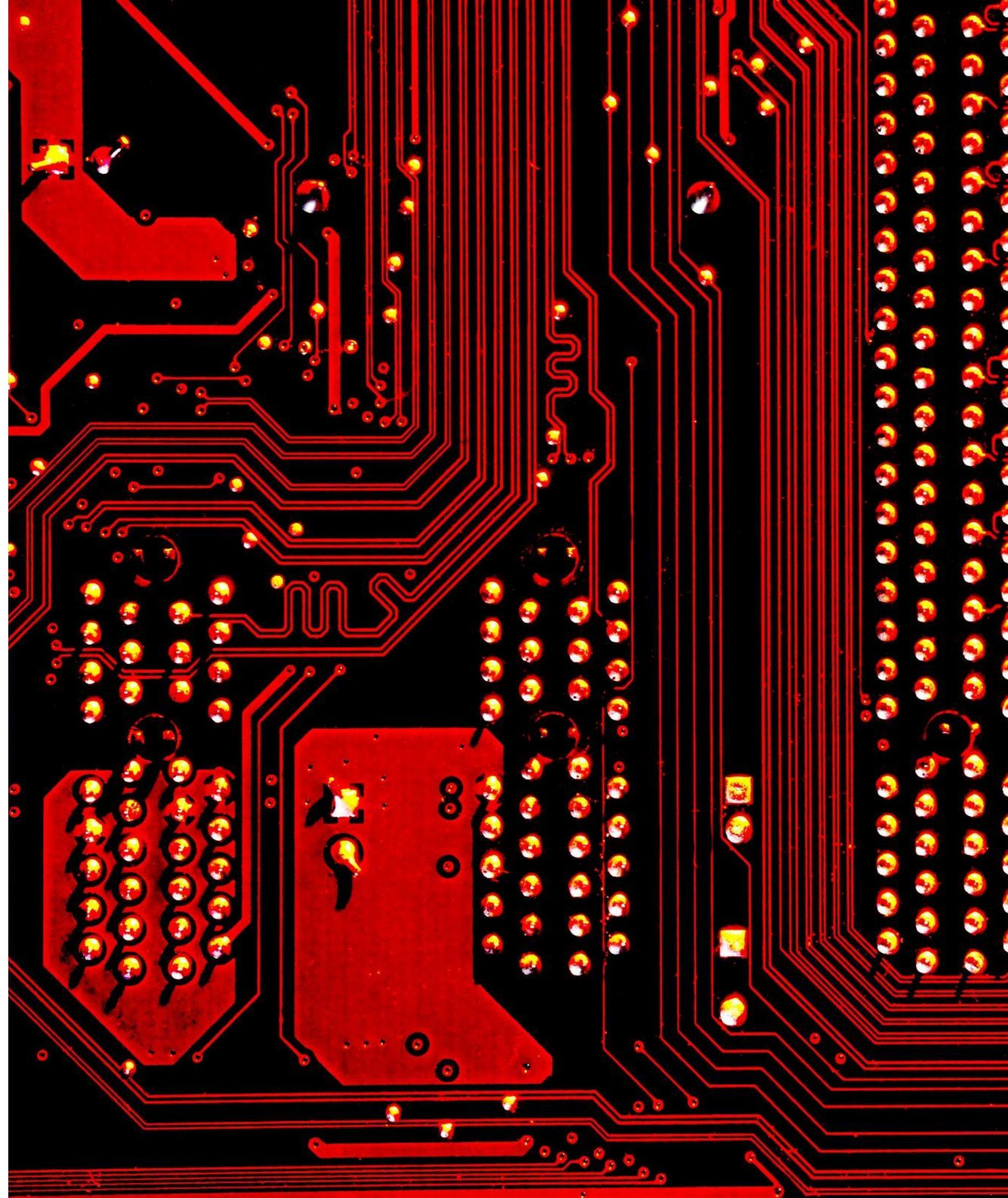
For ethical data treatment, in many cases principles are modified to have the data subject in mind. Safeguarding them, considering them the owners, etc.

/02 CURRENT PRINCIPLES

In many cases, ethical principles are not created from scratch - they modify current ones. Accountability, prioritising the user, complying with regulation, etc.

/04 3 LAYERS OF ETHICS

There are 3 layers of ethical expectations: Societal, organisational and individuals. These must be in alignment, or problems will occur.



DATA ETHICS

DATA ETHICS ALGORITHMS AND PROCESSES

In many cases, it's not stored data who hurt data subjects, but algorithms and processes. Several processes can use incorrect data (or use data the incorrect way) to (un)willingly hurt the data subject.

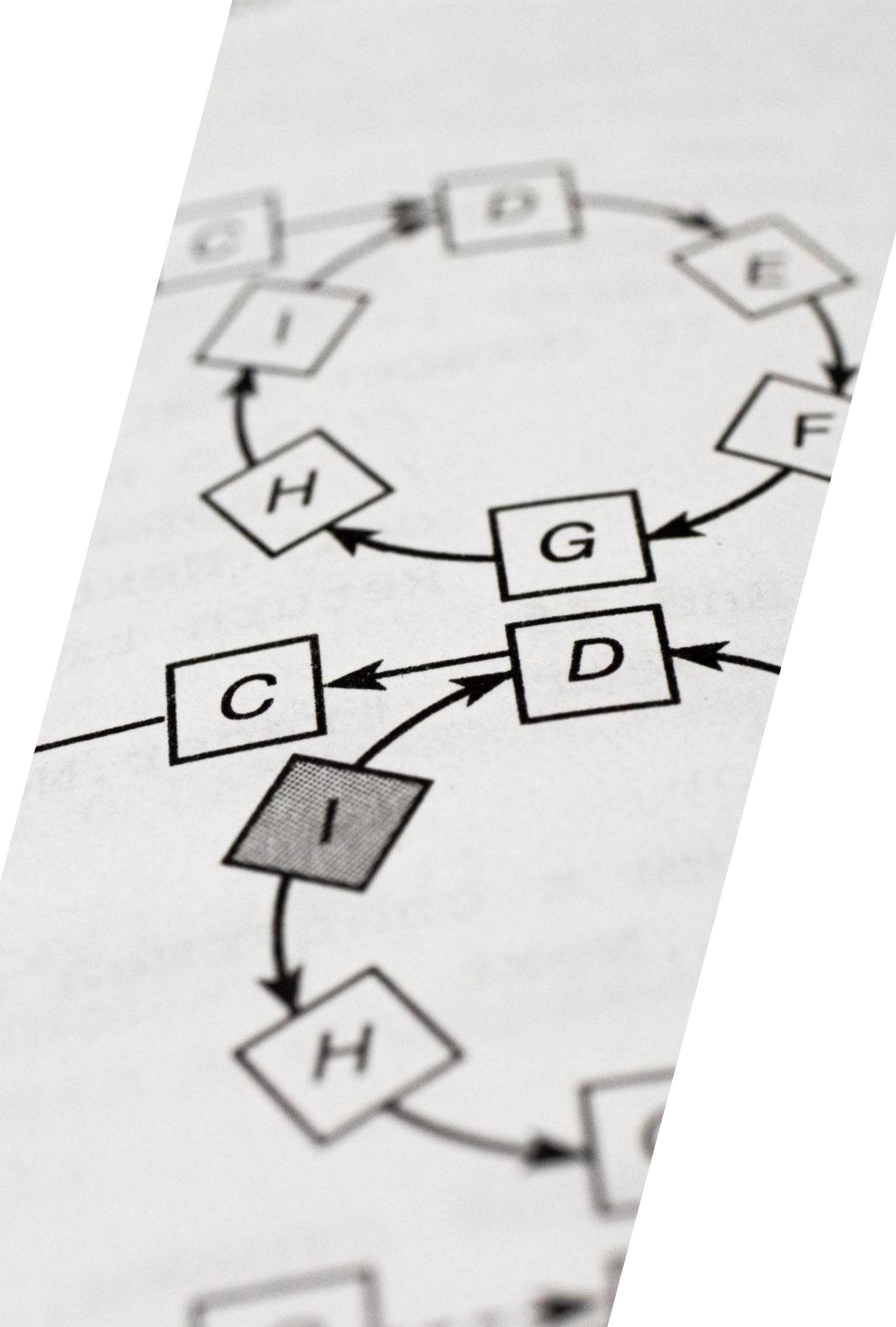
Common use cases where unfair or destructive algorithms are present include:

- Loan extension (discrimination/unfair pricing);
- Insurance pricing (unfair/random pricing);
- Performance reviews (discrimination/hidden criteria);
- Candidate evaluation (discrimination/hidden criteria);
- Defendant evaluation (discrimination/random criteria);

DATA ETHICS ALGORITHMS AND PROCESSES

Algorithm expert Cathy O'Neill calls destructive algorithms WMDs (Weapons of Math Destruction), and states their danger can be measured through three key factors:

- Opacity. Not giving transparency to a data subject (a rejected candidate never knowing why, a worker with a bad performance review never knowing why, and so on);
- Scale. The quantity of people affected (one racist criterion in a defendant evaluation algorithm that marks all US African-American as having a higher tendency for crime, versus in just one city or one state);
- Damage. The level of damage done by the algorithm (sending someone to prison, preventing them from being hired, etc);





DATA ETHICS

DATA ETHICS ALGORITHMS AND PROCESSES

In more general terms, algorithms and processes can hurt data subjects due to many different problems:

- Overefficiency. An extreme focus on numbers. Scheduling workers for peak work times with no breaks, measuring CS resolution times only, or just “reducing all to numbers”;
- Overloaded or correlated criteria. Using criteria that seem “innocent” but are complex - or unneeded (e.g., using location as a criterion for a loan may seem innocent, but may discriminate underprivileged people living there);
- Proxies. Using indicators of reality instead of reality. Using “Supervisor opinion of performance from 1 to 10” instead of an objective measure of performance from 1 to 10;
- Feedback loops. Unfair results used iteratively. Crime rates;

DATA ETHICS ALGORITHMS AND PROCESSES

In order for algorithms and processes to be fair and not hurt data subjects, perhaps the most important step is to make them transparent (both internally and externally):

- In many cases, complex or overloaded criteria are not readily apparent (e.g., location as a proxy for discriminating poor people), and can only be known if revealed;
- In other cases, the criteria may be fine, but there may be custom weights which may be “overadjusted” for certain purposes (e.g. a cancer detection image recognition program - “overadjusted” for one type... but used for many!);
- The algorithm or process may “work perfectly”, but with skewed data (e.g. facial recognition only for Caucasian faces that misses Black faces - works well... for this dataset)



DATA ETHICS ALGORITHMS AND PROCESSES EXAMPLES

/01 COMPLIANCE IN FINANCE

Finance (banking in specific) are highly regulated in terms of algorithms. Banks must keep data on lending decisions, and disclose terms of most products sold.

/02 THE FURTHER, THE WORSE

Criteria which are proxies don't represent reality - and the more removed from it, the worse. Performance becomes "1-10 rating" becomes "Subjective 1-10 rating"

/03 FEEDBACK COMPOUNDS

If errors are used as inputs, errors compound. Detecting more crimes in an area leads to more arrests and more crimes in that area. Financial models.

DATA ETHICS ALGORITHMS AND PROCESSES KEY TAKEAWAYS

/01 HURTING DATA SUBJECTS

Unfair or hidden algorithms can hurt data subjects in many areas. Extending loans, rejecting applicants, predicting someone's probability of crime, and others.

/03 NUMBERS, PROXIES, MORE

There are several other reasons why algorithms or processes may be unfair. Reducing everything to numbers, using indicators, using complex criteria, etc...

/02 3 WMD CRITERIA

Cathy O'Neill defines 3 major criteria for WMDs: Opacity, Scale and Damage. Having hidden criteria, hurting a lot of people, and hurting someone deeply.

/04 TRANSPARENCY IS CRUCIAL

Most problems with unfair algorithms and processes can be resolved with transparency. It allows detection of wrong criteria, biased training sets, and other issues.



DATA ETHICS

DATA ETHICS

ETHICAL DATA DIMENSIONS

In the Data Quality module, we covered Data Dimensions. These are the “lenses” through which you profile data, detecting if they’re high-quality or low-quality. Dimensions include:

- Completeness (how many values are missing?);
- Accuracy (do the values make sense?);
- Consistency (are they the same across mediums?);
- Timeliness (are the values recent?);

By considering ethical data treatment something quantifiable and measurable, we can also define data dimensions which measure how ethical data are.

- These focus on fairness, transparency, utility, and more;

DATA ETHICS

ETHICAL DATA DIMENSIONS

Some key “ethical data dimensions” include:

- Privacy. How much do these data invade on the privacy of the person? While in some cases it may be necessary, in others data may be infringing on privacy for no reason;
- Fairness. How much are these data, as results of an algorithm, similar to other outputs of the same algorithm? Similar results for all people means fair algorithms;
- Data subject agency. What degree of agency did the data subject have in this result? In other words, were they the ones to choose this result, or did they have no say in it?
- Impact. What impact (good or bad) do these data have on the person? Are they being used for the person or against them?





DATA ETHICS

DATA ETHICS ETHICAL DATA DIMENSIONS

Let's crystallise some of these dimensions with examples:

- Privacy
 - Data such as birth dates, home addresses, relationship statuses are highly sensitive - and possibly not needed;
- Fairness
 - Results of a lending algorithm that discriminates against poor borrowers will be different from other results;
- Data Subject Agency
 - For an email communication received by a person, with opt-out by default, agency is low. They had no say;
- Impact
 - Negative news posts shown to a person for shock may have a negative impact on their well-being;

DATA ETHICS

ETHICAL DATA DIMENSIONS

Just like with the “basic” data dimensions, we can also define other “ethical data dimensions” which may be of value:

- Transparency. How transparent are the criteria of an algorithm to the data subject? If I’m rejected as a candidate, do I have access to the specific criteria?
- Necessity. The degree to which these data are necessary to keep. Personal health information, for a hospital, may be very necessary. Political activity history, for a social media network, may not be necessary at all;
- Actionability. The degree to which these data are actionable and can be used for something (data which are not used should not be kept in the first place);

These characteristics evaluate data on how ethical they are.



DATA ETHICS

ETHICAL DATA DIMENSIONS

EXAMPLES

/01 NECESSITY/PRIVACY RATIO

In some cases, the ratio of necessity to privacy is important. Medical information is private but needed. Political affiliation data may be private but unneeded.

/02 WHOLE DATASET

Some dimensions require evaluating whole datasets. For example, for 50k transactions, you can only evaluate fairness by looking at all 50k results in total.

/03 HIGH SOPHISTICATION

Leveraging ethical data dimensions is one of the most sophisticated practices, and only companies with a very high level of data maturity usually do it.

DATA ETHICS

ETHICAL DATA DIMENSIONS

KEY TAKEAWAYS

/01 ETHICAL DIMENSIONS

Just like “normal” data dimensions classify data as high- or low-quality based on certain perspectives, these classify data based on ethical perspectives.

/02 FAIR, USEFUL, PRIVATE

Several ethical dimensions can be used, but they revolve around knowing if the data are fair, if they’re useful to the person, and if they’re private, for example.

/03 COMPOUNDING WORKS

Just like with other dimensions, ethical dimensions can be compounded. For example, privacy divided by utility, determining whether private data is needed.

/04 PRIORITISING THE SUBJECT

At the end of the day, ethical data dimensions detect whether an organisation is prioritising the data subject, using data for their good, or not at all.



SECURITY CONTROLS

SECURITY CONTROLS DATA USAGE PURPOSE/AUTHORITY

In terms of processing data, defining the purpose of the processing allows the restriction of user access to those data, as well as complying with privacy regulation.

- Usually, purposes are defined only for processing personally identifiable information (PII), but this control can be very effective for other types of data as well;

An organisation - or an individual user - has authority to process or use certain data based on their purpose:

- The same user may have authority to process certain data for one purpose, but not another one;
- For example, an after-sales professional using a customer address to ship a product (authorised), but not being able to use that address for marketing (not authorised);

SECURITY CONTROLS

DATA USAGE PURPOSE/AUTHORITY

A user or role's authority to process data may occur at any stage of the data lifecycle:

- They may be authorised to create data, to update them, to access them, and/or to dispose of them, depending on the purpose and role permissions;

Having authority to process data for a given purpose allows to more easily achieve compliance, as this information can be shared with regulators:

- “CS specialist can access customer data only for ABC purpose, and they can dispose of it only for DEF purpose”;
- “Systems administrators can access user logs to view activity, and cannot dispose of them for any purpose”;





SECURITY CONTROLS

SECURITY CONTROLS DATA USAGE PURPOSE/AUTHORITY

Defining the purpose of data processing works in synergy with two other important elements of data security and privacy: access control and data classification.

- Access Control (AC). Authorisation for a purpose is usually achieved through a combination of roles and groups;
 - After-sales expert shipping product: “After-Sales Expert” Role and “Shipping Operations” group;
 - After-sales expert doing email marketing: “After-Sales Expert” and “Email Marketing” group;
- Data Classification. Purpose definition is mostly applied to PII, so these data must be clearly identified;
 - Data stewards must classify all PII data in their metadata hub, and document which internal roles have access;

SECURITY CONTROLS

DATA USAGE PURPOSE/AUTHORITY EXAMPLES

/01 CONSENTED PURPOSES

Usually, data processing can be performed for purposes which the user has consented to, but not for others. For example, unsolicited communications.

/02 USER LOGS ARE IMPORTANT

Since the same user may have access or not to the same data based on different purposes, it's crucial to log the purpose to define whether a user has access.

/03 AUDITING IS PERFORMED

Usually, there are internal audits of the processing of PII, as any other control is audited. In this case, the specific purposes, roles and usage are evaluated.

SECURITY CONTROLS

DATA USAGE PURPOSE/AUTHORITY

KEY TAKEAWAYS

/01 AUTHORITY AND PURPOSE

An organisation - and specific users within it - may have authority to process data (usually, PII), or not, based on the specific purpose of the processing.

/02 DEFINED STAGES/PURPOSES

The only way to allow data processing by purpose is to clarify exactly what processing can be done, at what stage, by whom. Organisations must document this.

/03 AC AND CLASSIFICATION

This control usually works together with access control - which defines who can access what data - and data classification - defining which data can be accessed.

DATA SECURITY, PRIVACY, ETHICS

Covering how both data and data subjects are protected
in various data disciplines, both in terms of actual InfoSec
controls, but also in terms of being treated fairly.



DATA SECURITY, PRIVACY, ETHICS

We covered **three groups of topics** related to protecting data:



DATA SECURITY

How to protect data with controls such as locking rooms, encrypting data, and others.



DATA PRIVACY

How data can be protected with data governance structures, controls by data classes, and more.



DATA ETHICS

How data can be treated ethically with transparent algorithms, data purposes, and others.



DATA SECURITY, PRIVACY, ETHICS

DATA SECURITY, PRIVACY, ETHICS **CONSOLIDATING**

Some questions to consolidate your knowledge can include:

- What element prevents data disposal, forcing a retention period? Besides this, how soon should data be destroyed?
- Does any version of an encryption protocol keep data secure?
- What are some ethical data dimensions?
- When we define a purpose for data processing, does a specific role have authority for any purpose? Or not?
- What is the name given to media downgrading when talking about paper records in specific?
- What are some elements that can make an algorithm discriminatory or unfair?