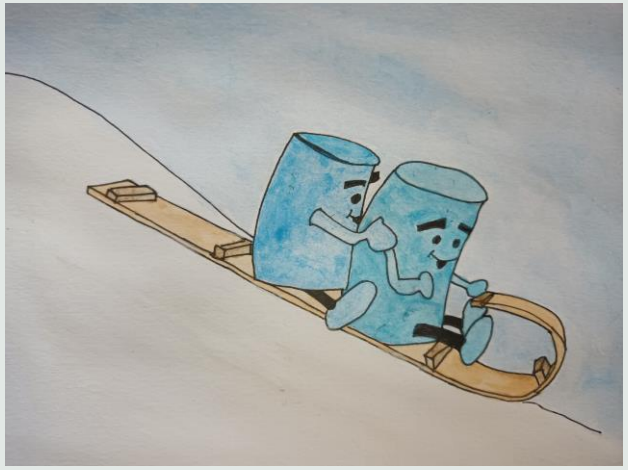
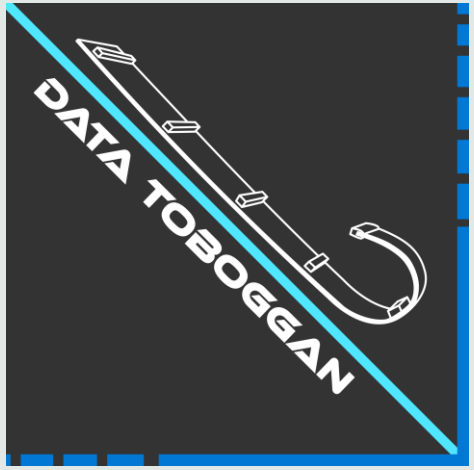
Data Quality inspiration, erudition and expedience

Dr Victoria Holt

About me

Community





Data, Data, Data

DATA STRATEGY | DATA GOVERNANCE |

VIRTUAL CDO | DATA QUALITY |

DATA ARCHITECTURE | RESEARCHER

Inspiration

Data growth is relentless

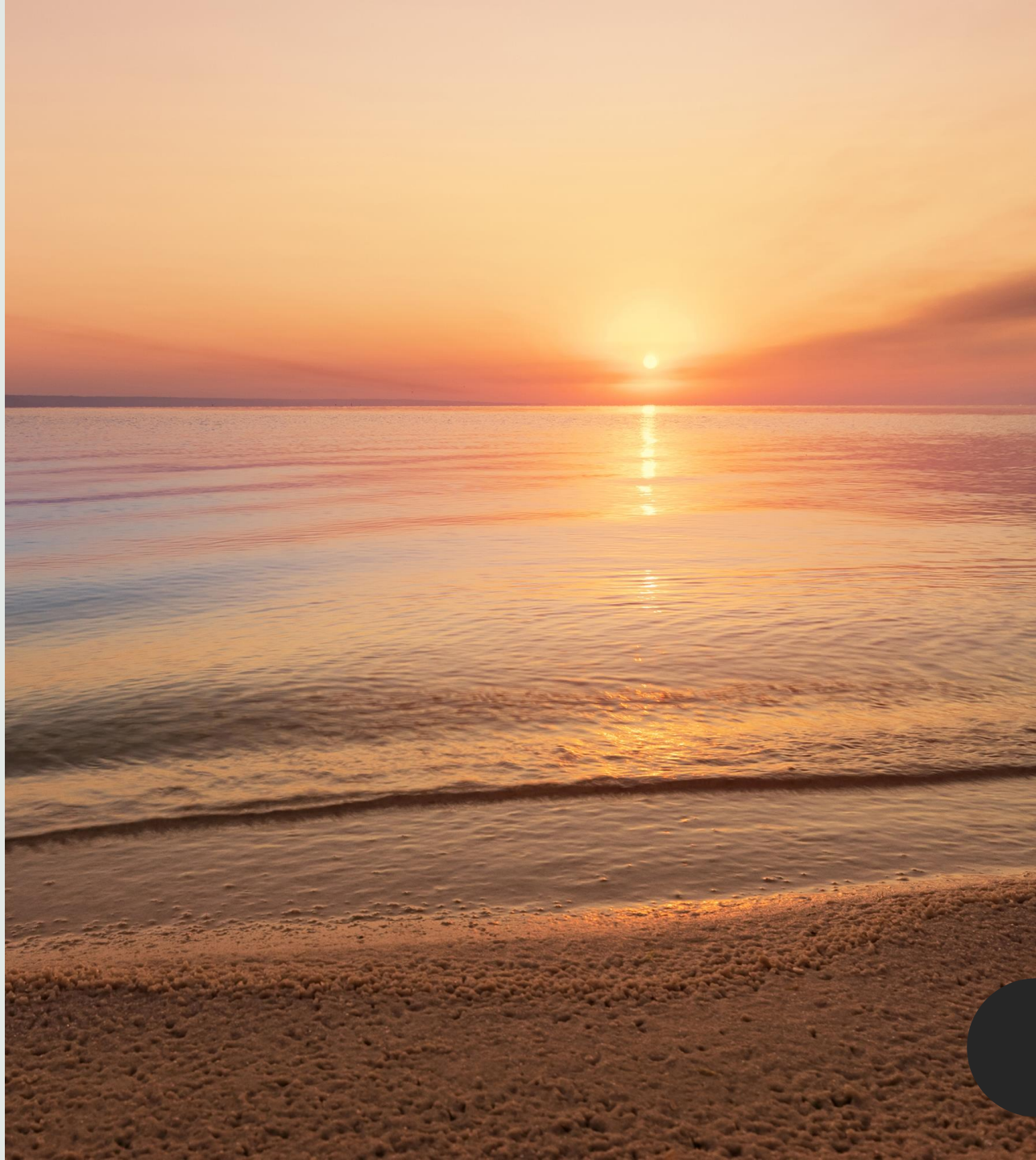
Volume – the amount of data

Velocity – the speed at which data is generated and needs to be accessible

Variety – The different types of data: structured; semi-structured; unstructured

Veracity – the uncertainty of data: data quality; cleansed data; data accuracy

Value – the business value of the data collected



Data Veracity

Uncertainty in data due to

- Inconsistency
- Incompleteness
- Ambiguities
- Latency

Data Quality Ethics

- What is the quality of the data
- From where was the data collected
- How was the data collected
- How will the data be used

Erudition

What is Data Quality?

The Data Management Body of Knowledge (DMBOK) defines Data Quality as

"the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meet the needs of data consumers."

DAMA - DMBOK

Six core data quality dimensions

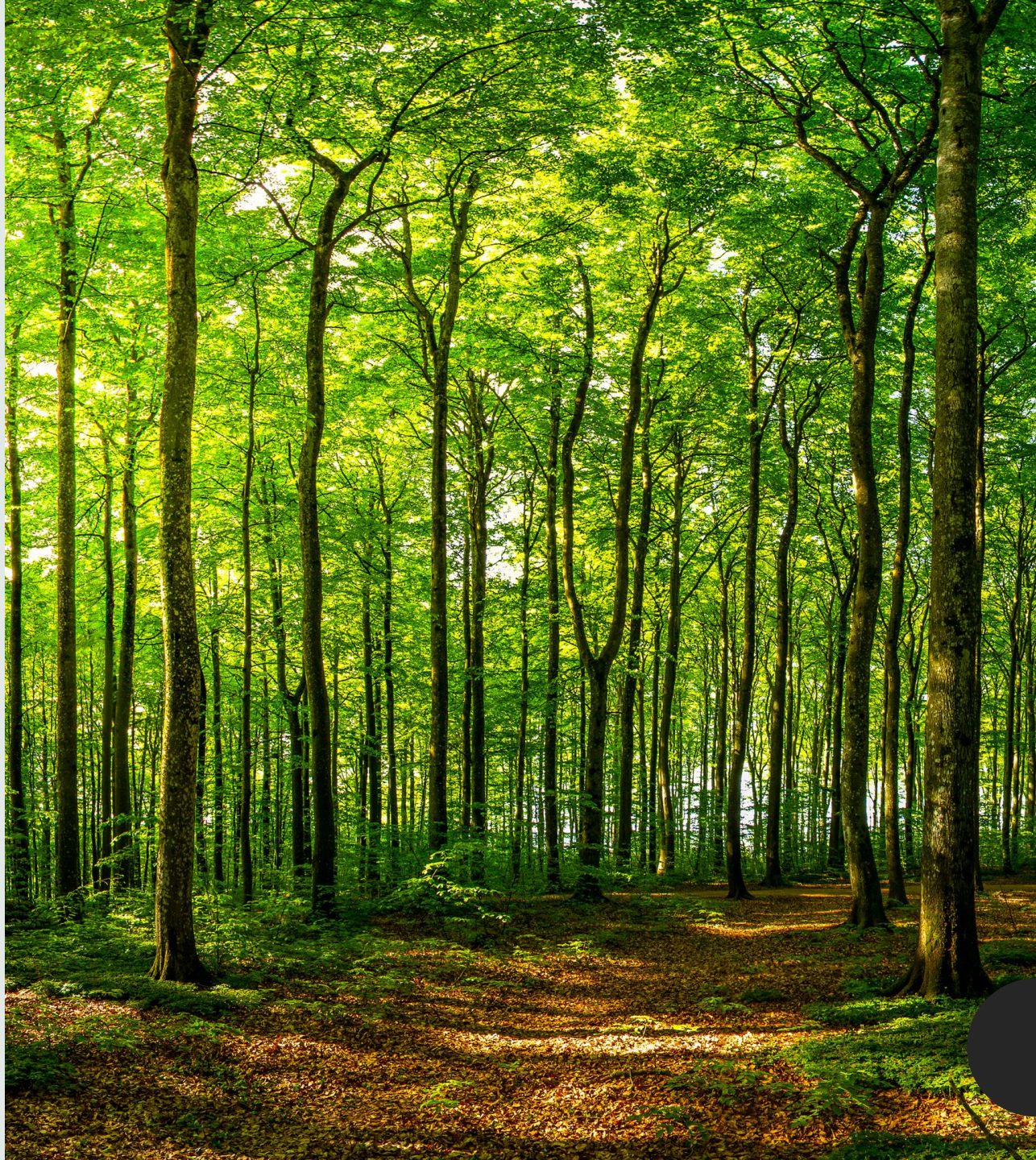
- Timeliness – an accurate reflection of the period it represents, and the values are up to date with the time lag between collection and availability meeting the intended use
- Completeness –the degree to which records are present, all records are included and not truncated
- Uniqueness – the degree to which there is no duplication in records, each value stored once
- Consistency – the degree to which values do not contradict other value representing the same entity (e.g. data of birth the same in 2 data sets)
- Validity – degree to which data is in the range and format expected (e.g. a data of birth is not in the future)
- Accuracy – degree to which matches reality (e.g. date of birth in UK (DDMM) and US (MMDD) entry. Bias in data may impact accuracy

From: DAMA UK and the Data Quality Dimensions Working Group

ISO 8000-8 Information and data quality

Concepts and measuring defines three categories for data quality measurements syntactic, semantic, and pragmatic. Provides a foundation for measuring information and data quality.

- Syntactic – the degree to which data conforms to its specified syntax (Format)
- Semantic – the degree to which data corresponds to what it represents (Meaning)
- Pragmatic - the degree to which data is found suitable and worthwhile for a particular purpose (Usefulness)

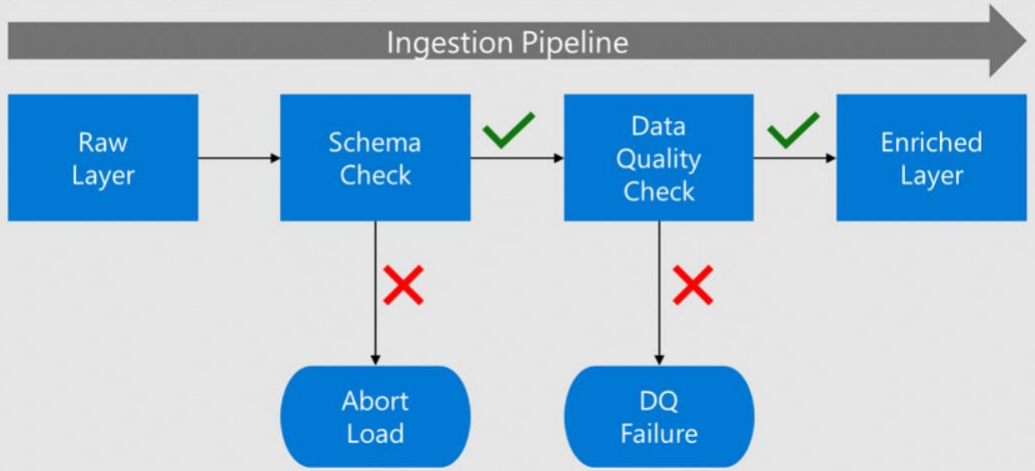


Data Profiling

Minimum metrics Microsoft recommend to increase data quality

- Completeness
- Uniqueness
- Consistency
- Validity
- Accuracy
- Linkage

Validate early



https://learn.microsoft.com/en-us/azure/cloud-adoption-framework/scenarios/cloud-scale-analytics/govern-data-quality

Expedience

DAMA Measure and Monitor Data Quality

Measurement can be described at 2 levels

- The details related to the execution of individual rules
- Overall results aggregated from the rules

Key
DQI – Data Quality Index
R - represents rule being tested

Rules: KPI indexes to be set
-Standard
-Target
-Threshold

Percentage of correct data

$$\text{ValidDQI}(r) = \frac{\text{TestExecutions}(r) - \text{ExecptionsFound}(r))}{\text{ExecptionsFound}(r))}$$

Percentage of exceptions

$$\text{InvalidDQI}(r) = \frac{\text{ExecptionsFound}(r)}{\text{TestExecutions}(r)}$$

Example 10,000 business rules tested with 560 exceptions

ValidDQI(r) 9440/10000 = 94.4% InvalidDQI(r) 560/10000 = 5.6%

DAMA Data Quality Metric Examples

Dimension and Business Rule	Measure	Metrics	Status Indicator
Completeness Business Rules 1: Population of field is mandatory	Count the number of populated records; compare to total number of records	Divide obtained number of records where data is populated by the total number of records in the table or database and multiply by 100 to get percent complete	Unacceptable: <80% populate >20% not populated
Example: Post code must be populated	Count populated: 700,000 Count not populated: 300,000 Total count:1,000,000	Positive measure: 700,000/1,000,000*100 = 70% populated Negative measure: 300,000/1,000,000 *100 = 30% not populated	Example result Unacceptable
Uniqueness Business Rule 2: There should be only one record per entity instance in a table	Count the number of duplicate records identified; report on the percent of records that represent duplicates	Divide the number of duplicate records by the total number in the table or database and multiple it by 100	Unacceptable: Above 0%
Example: There should be one and only one current row per postal code on the postal codes master list	Count of duplicates: 1,000 Total Count: 1,000,000	10,000/1,000,000*100 = 1.0% of postal codes are present on more than one current row	Example result Unacceptable

DAMA-DMBOK example

DAMA Data Quality Metric Examples (2)

Dimension and Business Rule	Measure	Metrics	Status Indicator
Timeliness Business Rules 3: Records must arrive within a scheduled timeframe	Count the number of records failing to arrive on time from a data service for business transactions to be complete	Divide the number of incomplete transactions by the total number of attempted transactions in a time period and multiple by 100	Unacceptable: <99% completed on time > 1% not completed on time
Example: Equity market record should arrive within 5 minuets of being transacted	Count of incomplete transactions: 2000 Count of attempted transactions: 1,000,000	Positive measure: (1,000,000 – 2000) / 1,000,000*100 = 99.8% of transactions records arrived within defined timeframe Negative measure: 2000/1,000,000*100=0.20% of transactions did not arrive within defined timeframe	Example result : Acceptable
Validity Business Rule 4: If field X = value 1, then field Y must = value 1-prime	Count the number of records where the rule is met	Divide the number of records that meet the condition by the total number of records	Unacceptable: Below 100% adherence to the rule
Example: Only shipped orders should be billed	Count of records where status for shipping = Shipped and status for billing = Billed: 999,000 Count of total records: 1,000,000	Positive: 999,000/1,000,000*100 = 99.9% of records conforms to the rule Negative: (1,000,000-999,000)/1,000,000*100= 0.10% do not conform to the rule	Example result Unacceptable DAMA-DMBOK example

UK Government Data Quality Framework

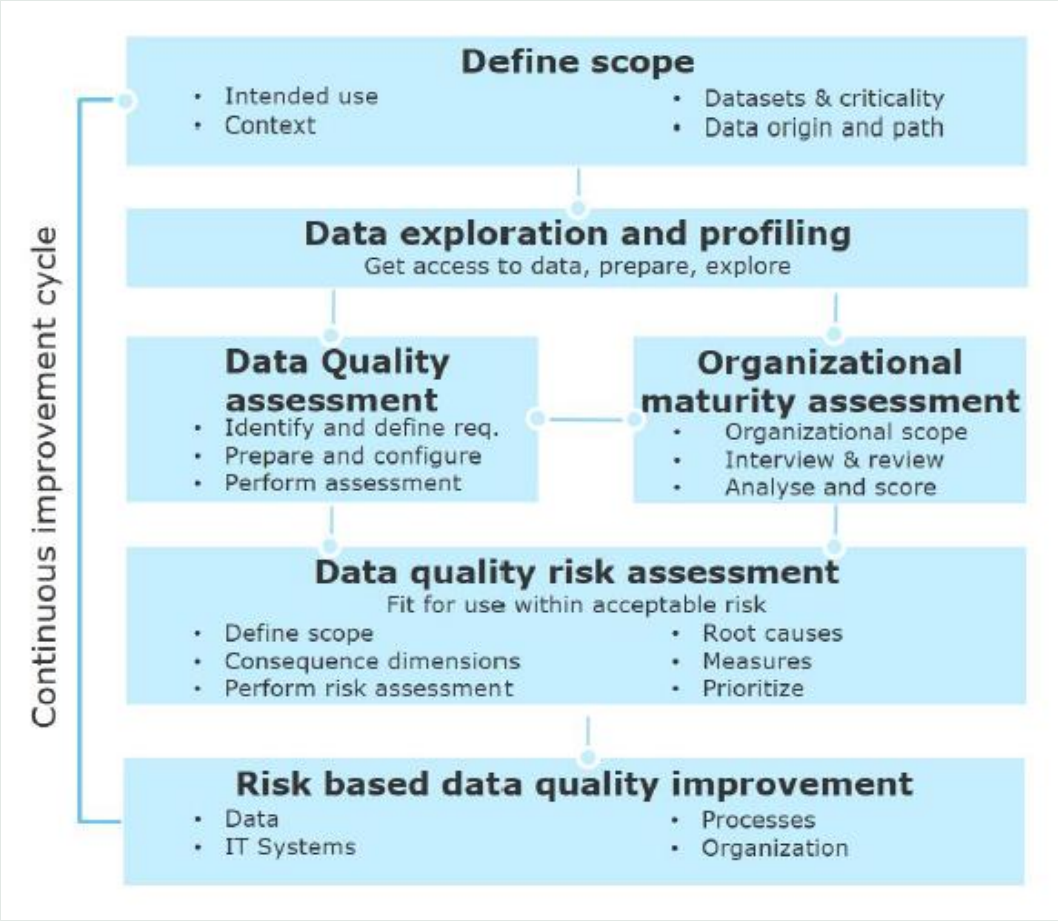
Principles

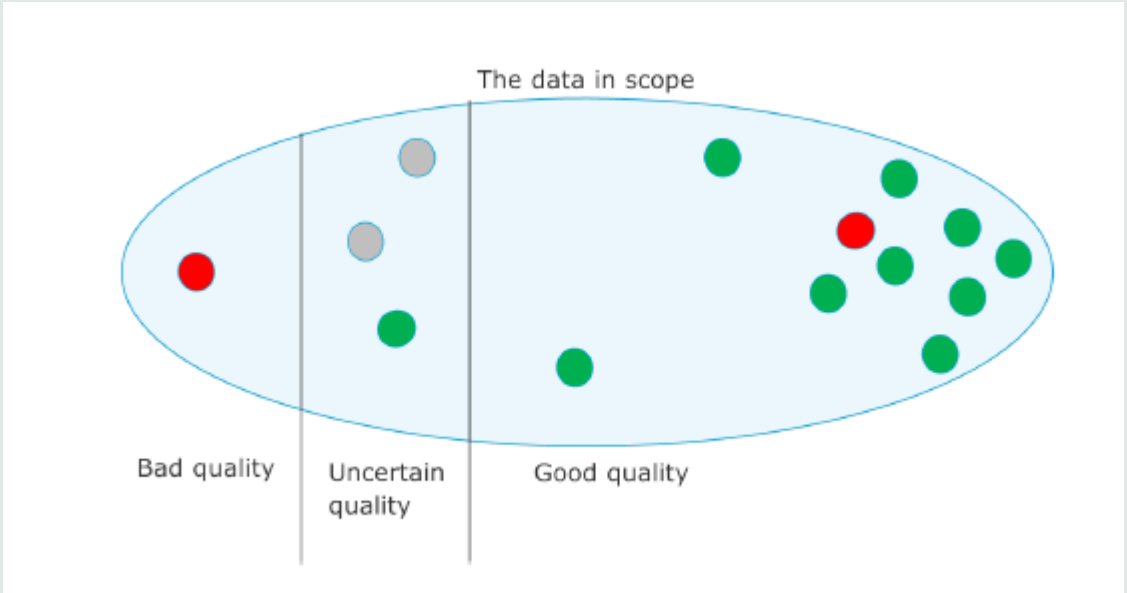
- Commit to data quality through embedding effective data management and governance, building a data quality capability with continuous improvement
- Know your users and understand their quality needs
- Assess quality throughout the data lifecycle and communicate with users and stakeholders across the lifecycle
- Communicate data quality clearly and effectively with documentation and metadata
- Anticipate changes affecting data quality and plan for the future

https://www.gov.uk/government/publications/the-government-data-quality-framework/the-government-data-quality-framework

https://www.gov.uk/government/organisations/government-data-quality-hub

Data quality process





https://rules.dnv.com/docs/pdf/DNV/rp/2017-01/dnvgl-rp-0497.pdf

The future is where?

Create a roadmap of tasks for each business use case

Understand what data exists

Assign data owners to the data sources

Understand the lineage of the data used for reports and dashboards.

Create a data quality metrics dashboard or use a tool