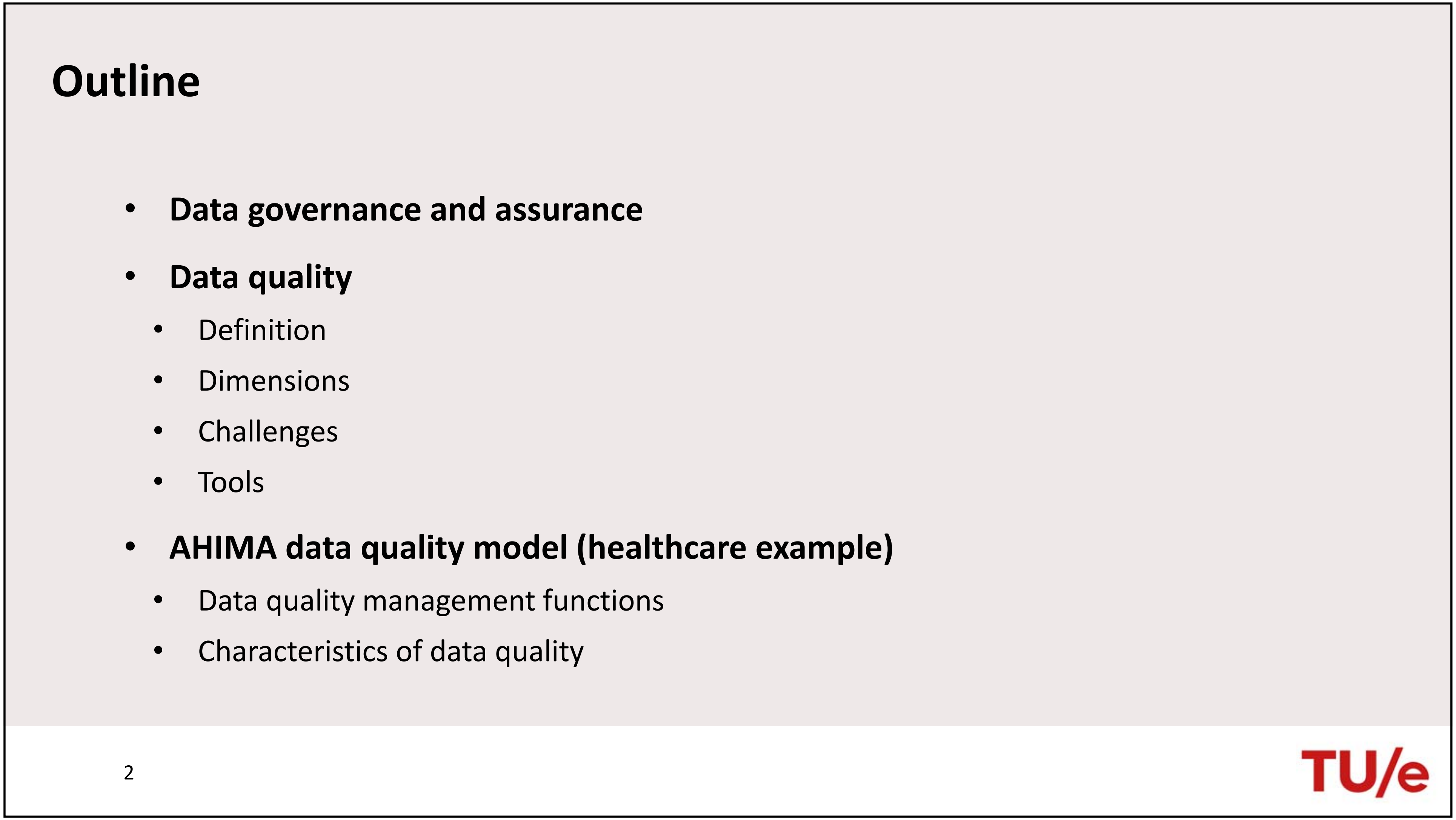
Data quality

5ARB0: DATA ACQUISITION & ANALYSIS (2022 – 2023)

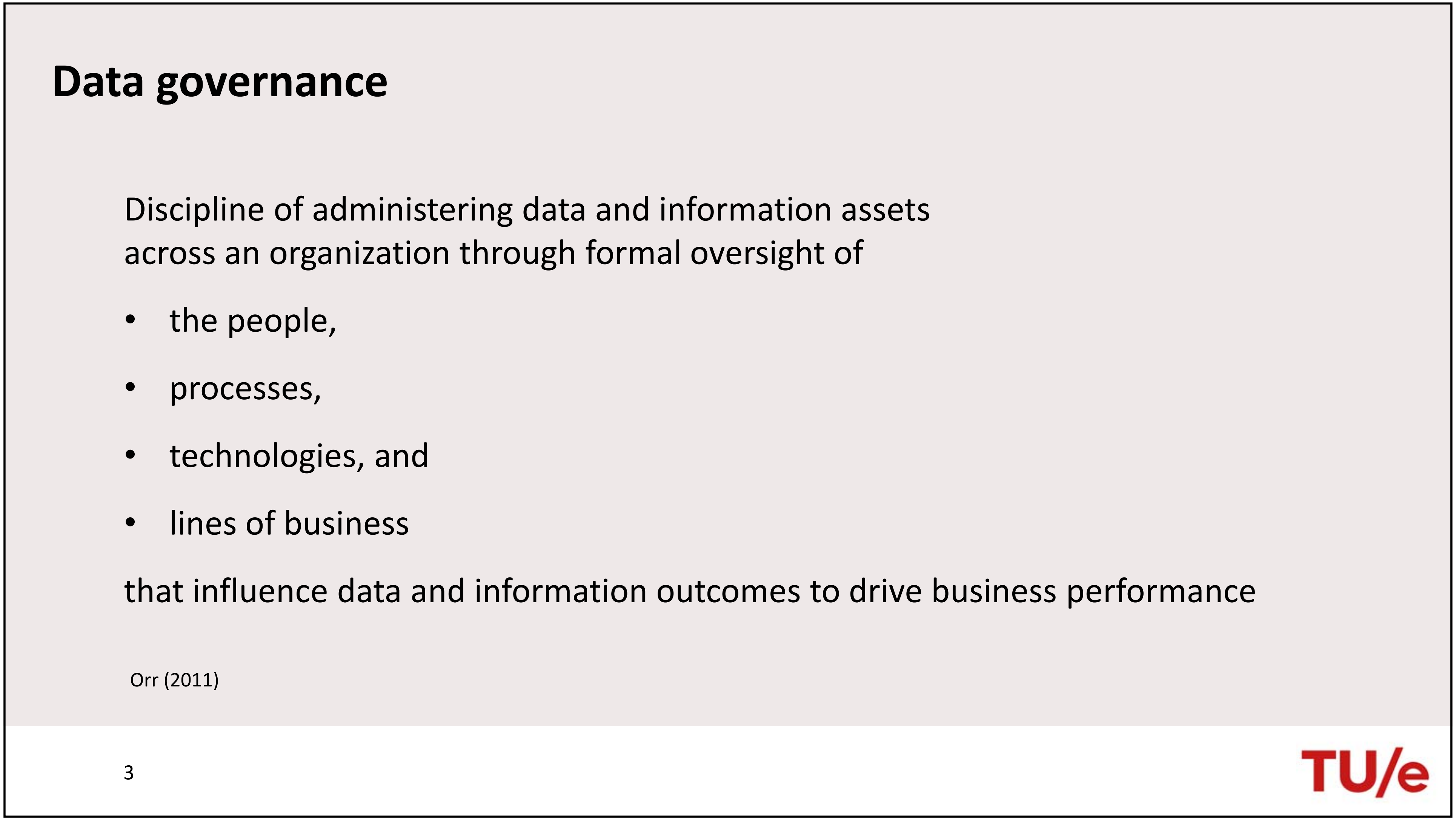
Uzay Kaymak, Jheronimus Academy of Data Science, u.kaymak@tue.nl

Mastertrack: Artificial Intelligence & Engineering Systems



Outline

- **Data governance and assurance**
- **Data quality**
  - Definition
  - Dimensions
  - Challenges
  - Tools
- **AHIMA data quality model (healthcare example)**
  - Data quality management functions
  - Characteristics of data quality



Data governance

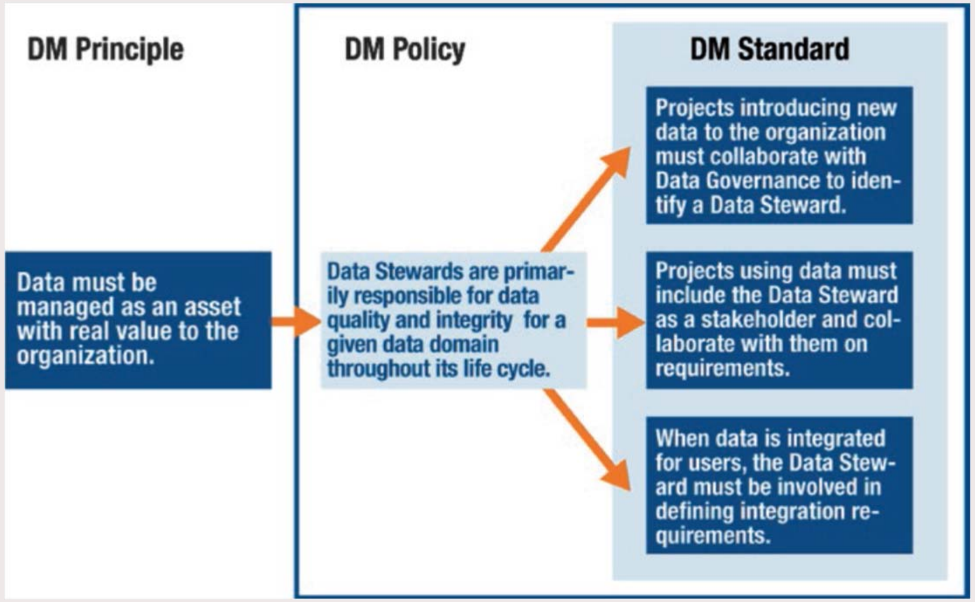
Discipline of administering data and information assets
across an organization through formal oversight of

- the people,
- processes,
- technologies, and
- lines of business

that influence data and information outcomes to drive business performance

Orr (2011)

Organizational principles, policies and standards



DM Principle

Data must be managed as an asset with real value to the organization.

DM Policy

Data Stewards are primarily responsible for data quality and integrity for a given data domain throughout its life cycle.

DM Standard

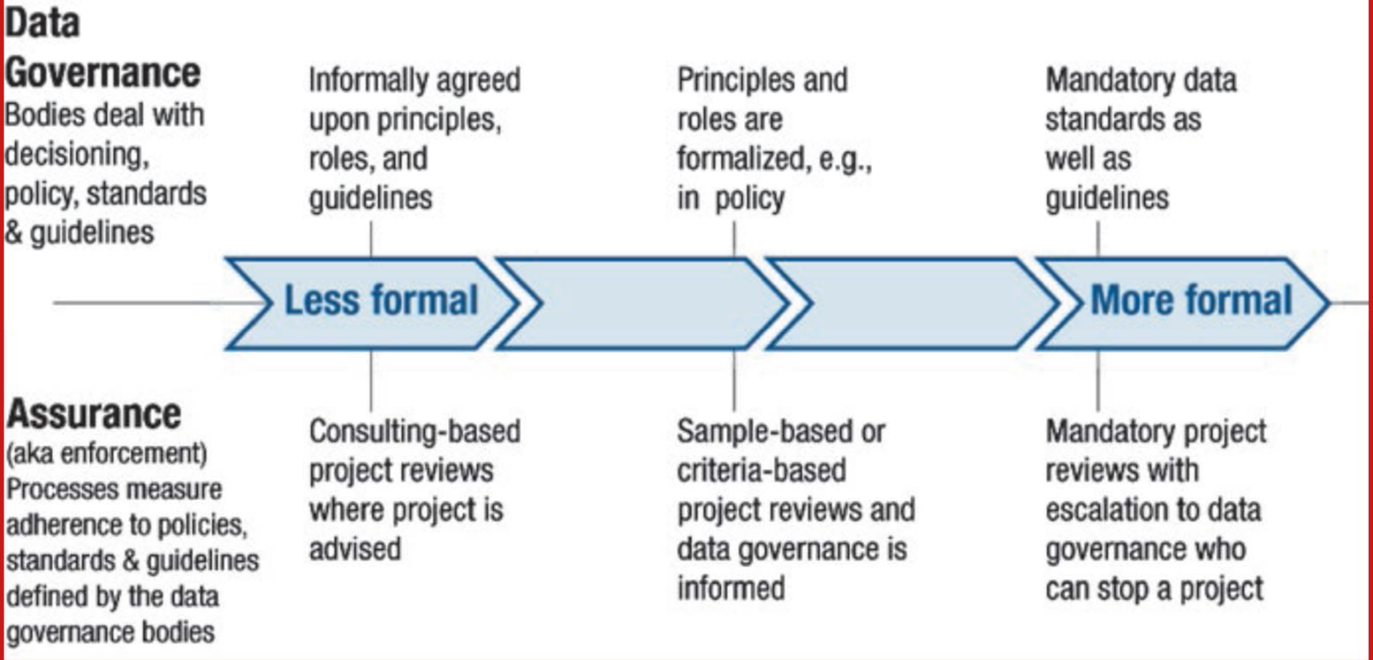
Projects introducing new data to the organization must collaborate with Data Governance to identify a Data Steward.

Projects using data must include the Data Steward as a stakeholder and collaborate with them on requirements.

When data is integrated for users, the Data Steward must be involved in defining integration requirements.

Data governance is
a management function

Data governance and assurance



Data Governance
Bodies deal with decisioning, policy, standards & guidelines

Informally agreed upon principles, roles, and guidelines

Principles and roles are formalized, e.g., in policy

Mandatory data standards as well as guidelines

Less formal → More formal

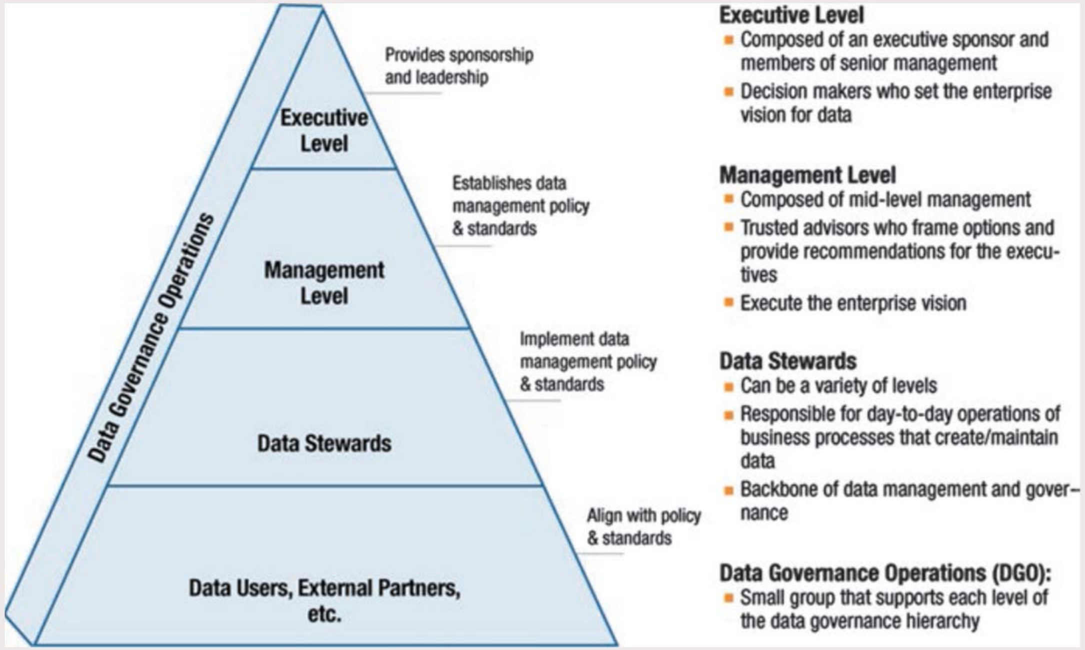
Assurance
(aka enforcement) Processes measure adherence to policies, standards & guidelines defined by the data governance bodies

Consulting-based project reviews where project is advised

Sample-based or criteria-based project reviews and data governance is informed

Mandatory project reviews with escalation to data governance who can stop a project

Data governance hierarchy


Data Governance Operations
Executive Level
Management Level
Data Stewards
Data Users, External Partners, etc.
Provides sponsorship and leadership
Establishes data management policy & standards
Implement data management policy & standards
Align with policy & standards

Executive Level

- Composed of an executive sponsor and members of senior management
- Decision makers who set the enterprise vision for data

Management Level

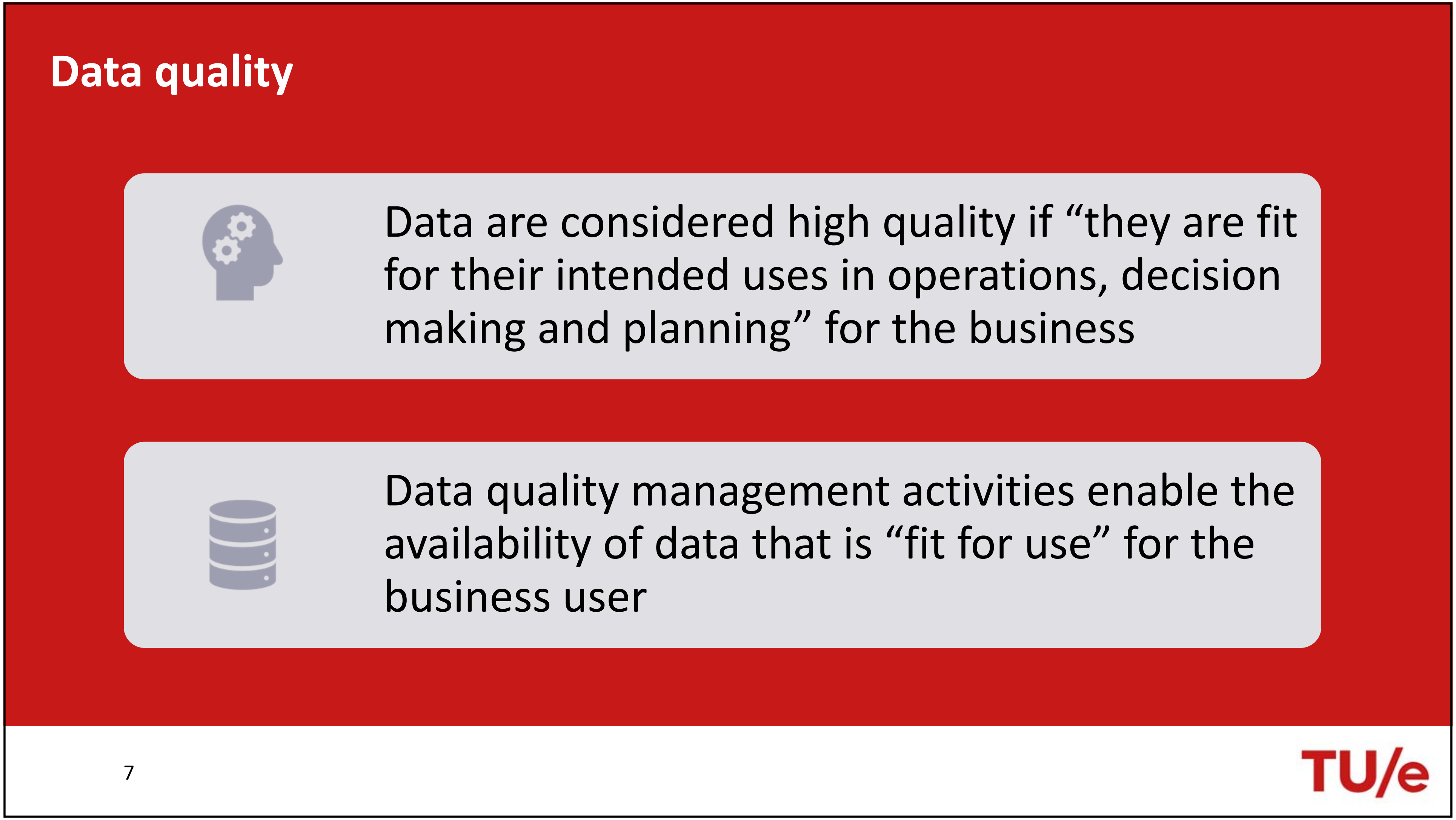
- Composed of mid-level management
- Trusted advisors who frame options and provide recommendations for the executives
- Execute the enterprise vision

Data Stewards

- Can be a variety of levels
- Responsible for day-to-day operations of business processes that create/maintain data
- Backbone of data management and governance

Data Governance Operations (DGO):

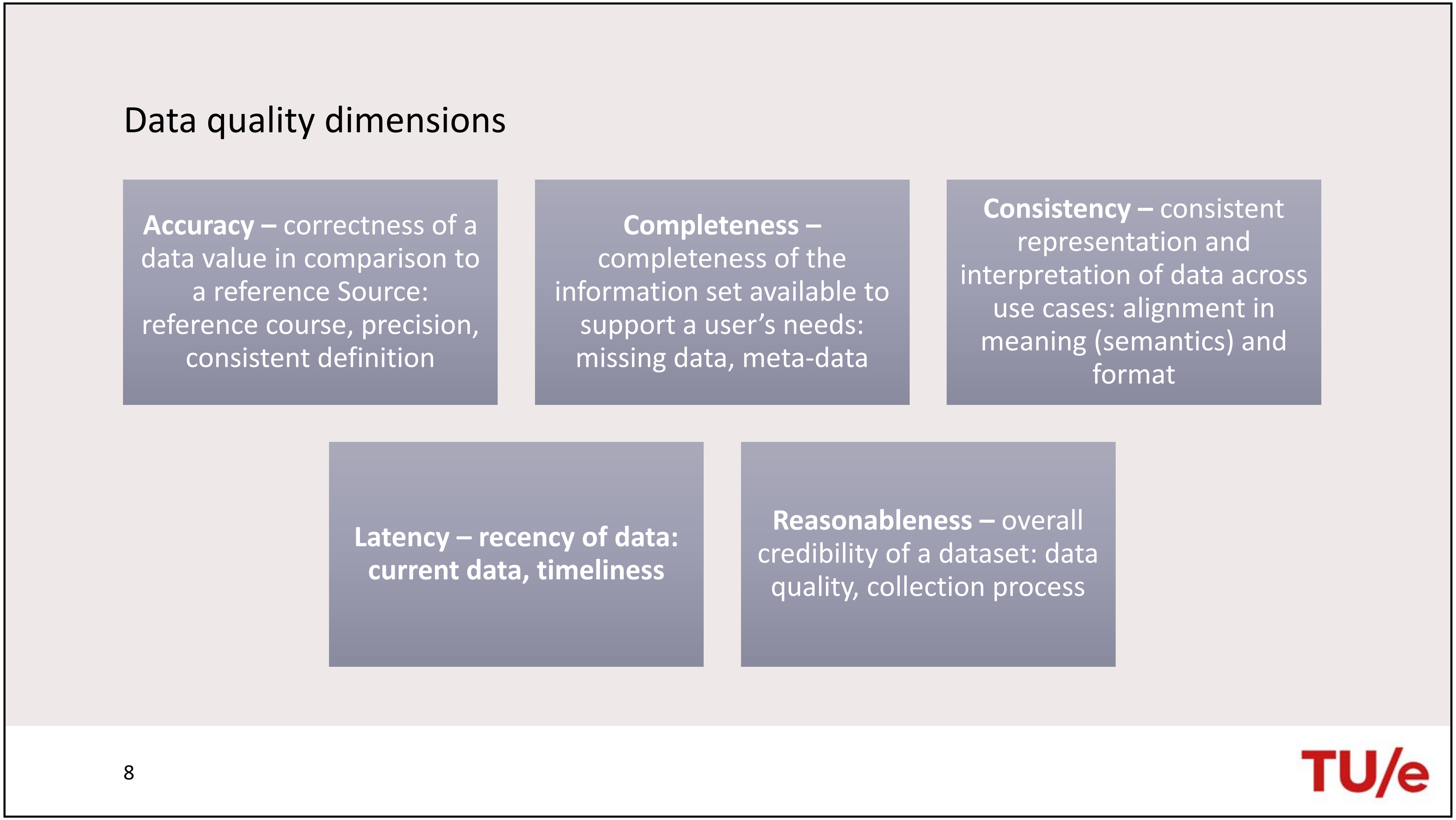
- Small group that supports each level of the data governance hierarchy



Data quality

Data are considered high quality if “they are fit for their intended uses in operations, decision making and planning” for the business

Data quality management activities enable the availability of data that is “fit for use” for the business user



Data quality dimensions

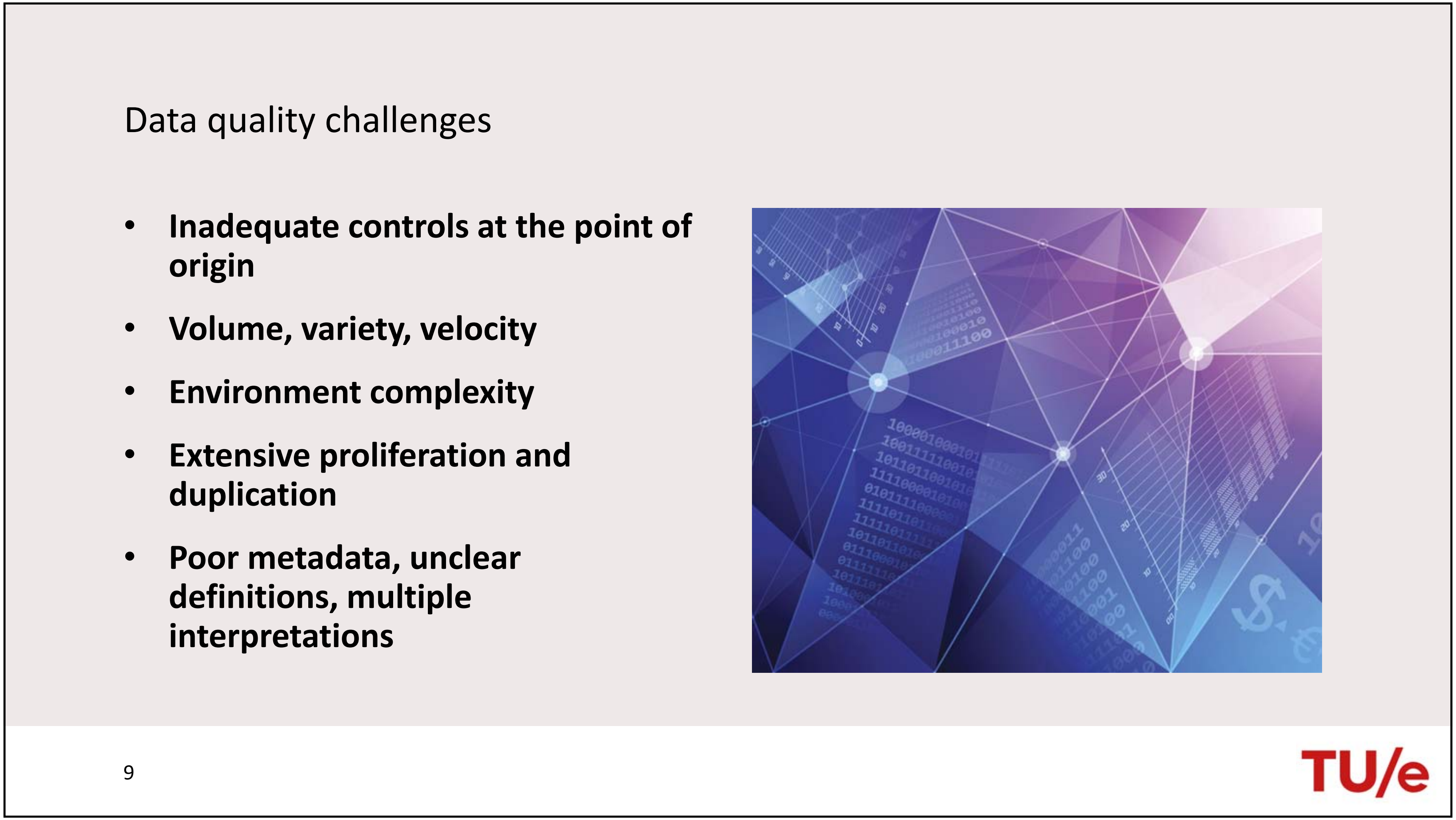
Accuracy – correctness of a data value in comparison to a reference Source: reference course, precision, consistent definition

Completeness – completeness of the information set available to support a user's needs: missing data, meta-data

Consistency – consistent representation and interpretation of data across use cases: alignment in meaning (semantics) and format

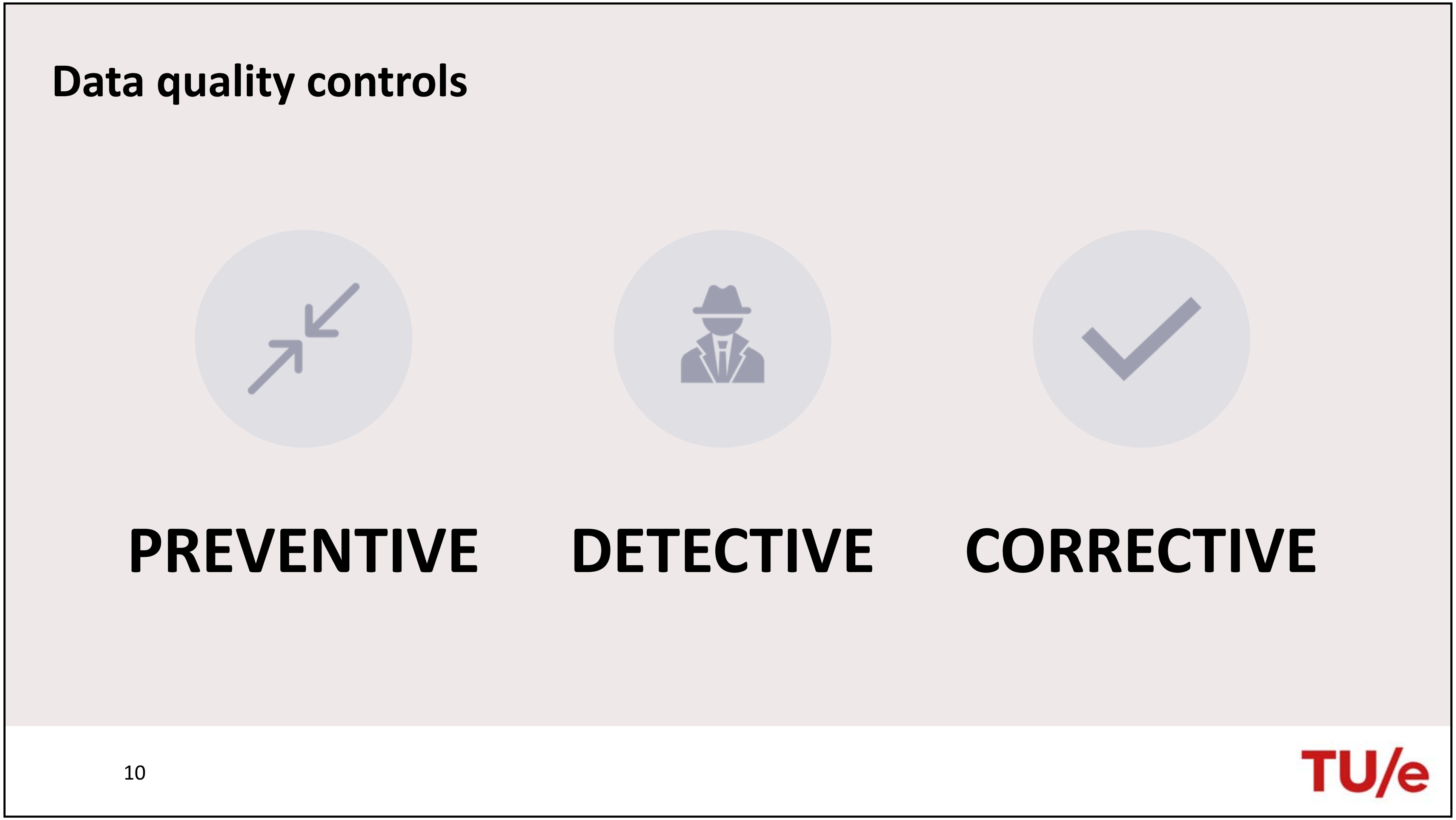
Latency – recency of data: current data, timeliness

Reasonableness – overall credibility of a dataset: data quality, collection process



Data quality challenges

- **Inadequate controls at the point of origin**
- **Volume, variety, velocity**
- **Environment complexity**
- **Extensive proliferation and duplication**
- **Poor metadata, unclear definitions, multiple interpretations**

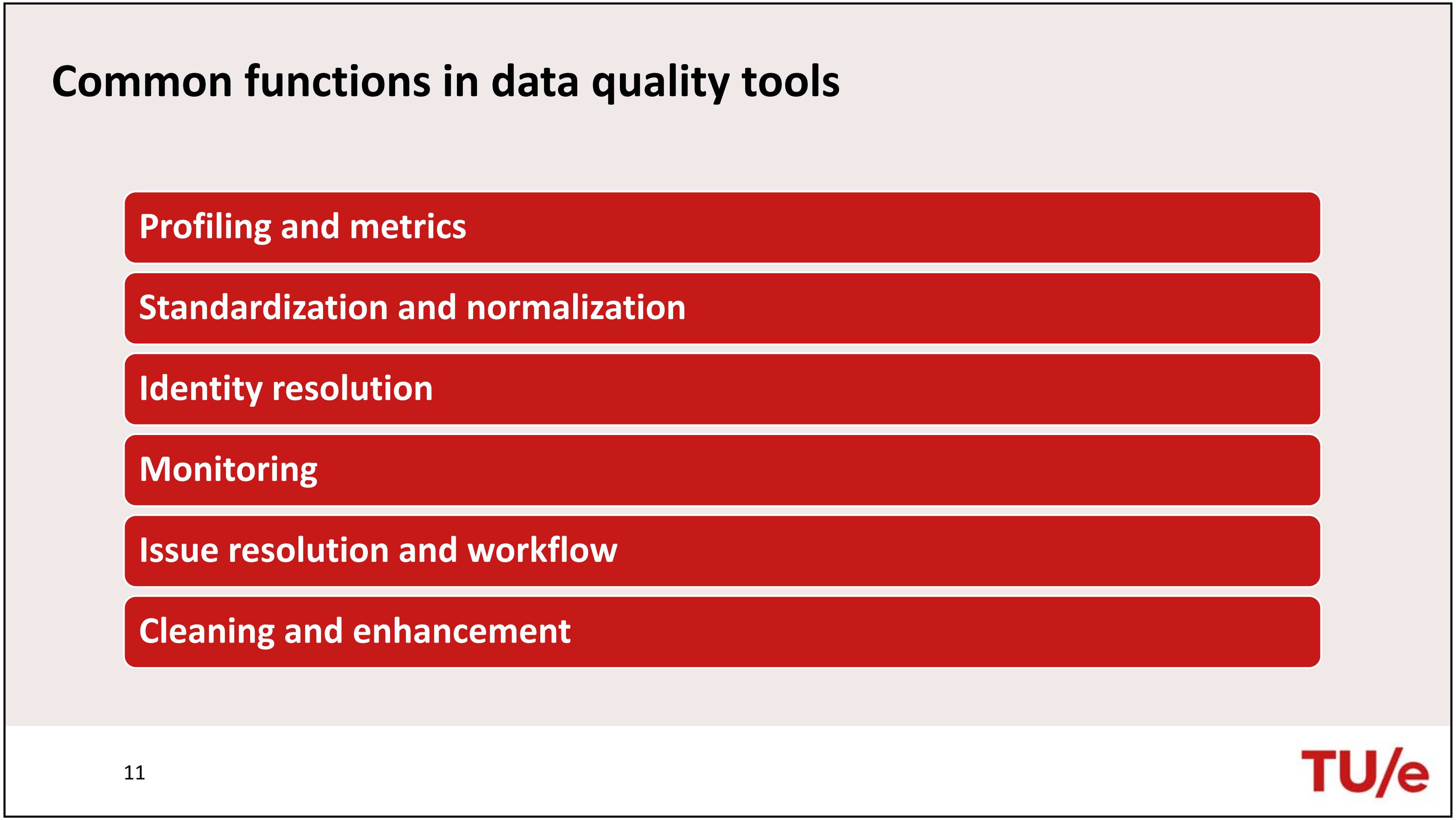


Data quality controls

PREVENTIVE

DETECTIVE

CORRECTIVE



Common functions in data quality tools

- Profiling and metrics
- Standardization and normalization
- Identity resolution
- Monitoring
- Issue resolution and workflow
- Cleaning and enhancement

AHIMA Data Quality Model

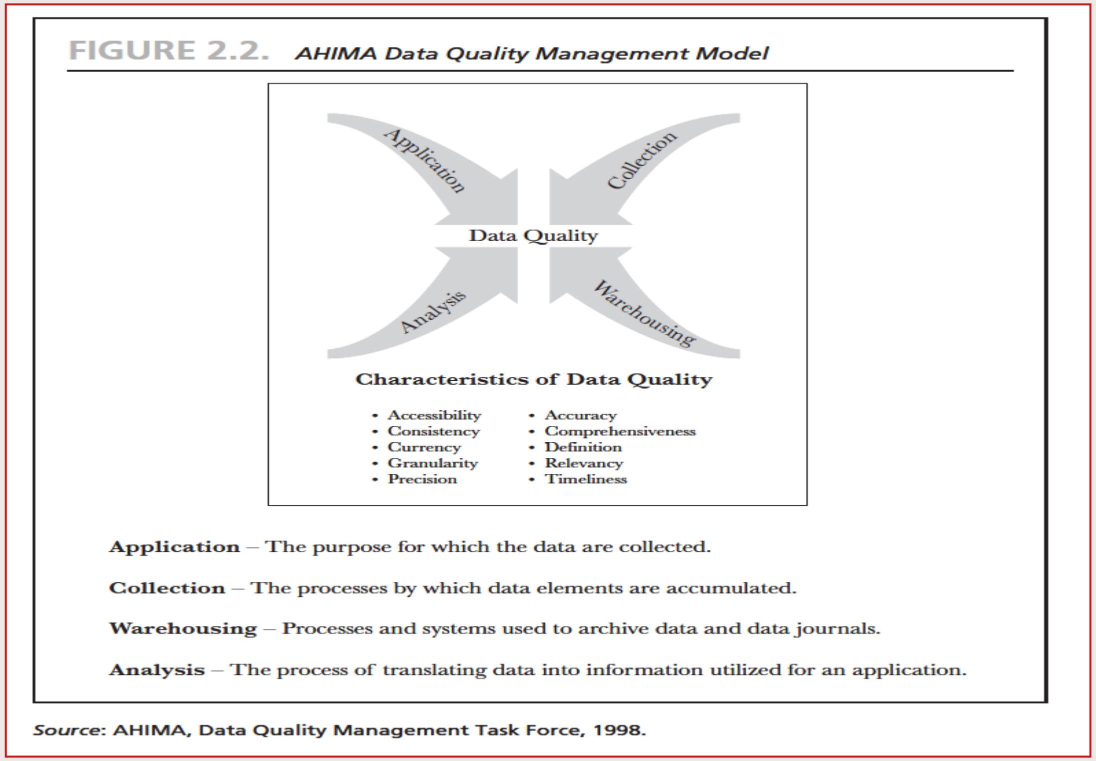


FIGURE 2.2. *AHIMA Data Quality Management Model*

Characteristics of Data Quality

- Accessibility
- Consistency
- Currency
- Granularity
- Precision
- Accuracy
- Comprehensiveness
- Definition
- Relevancy
- Timeliness

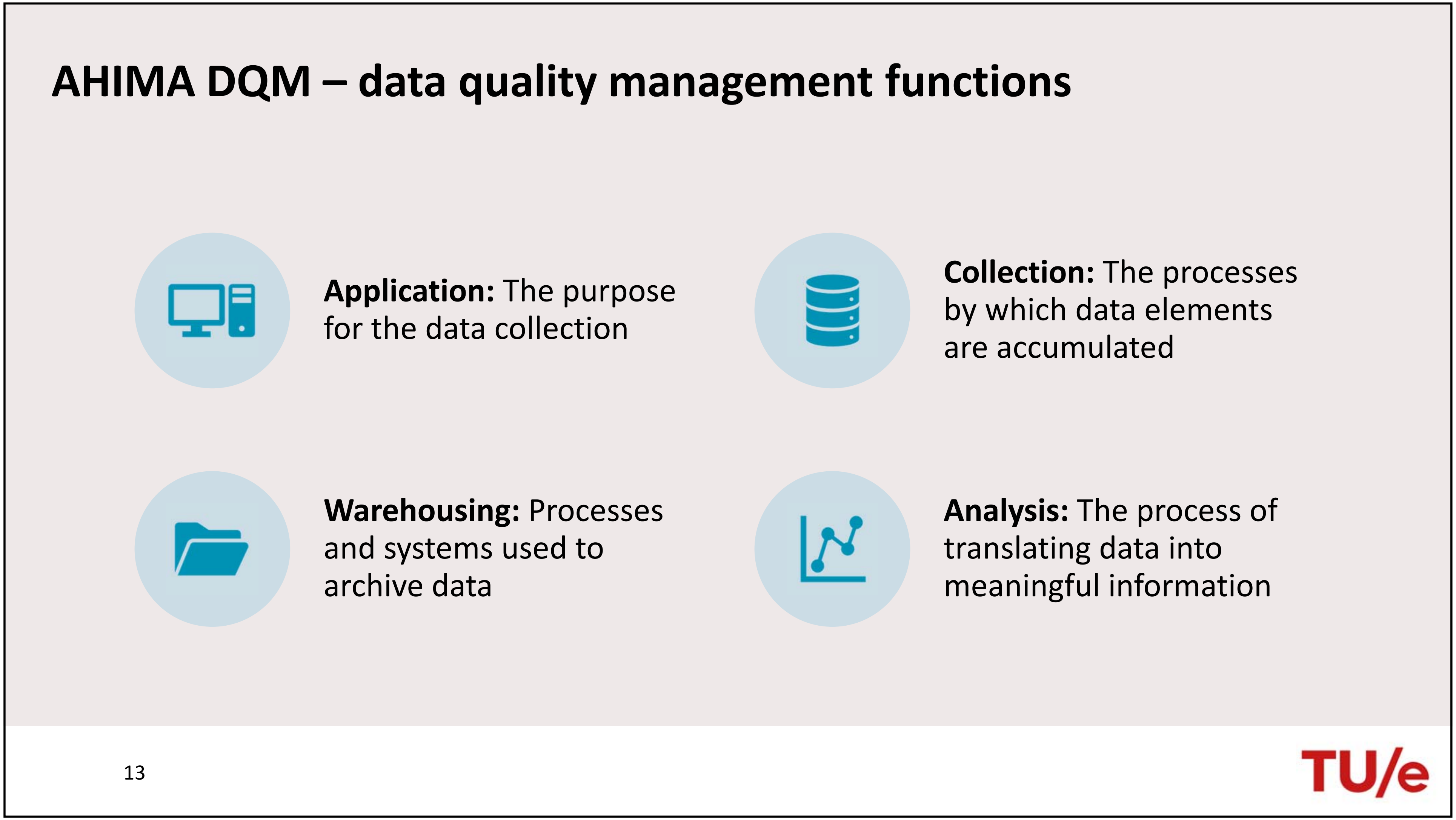
Application – The purpose for which the data are collected.

Collection – The processes by which data elements are accumulated.

Warehousing – Processes and systems used to archive data and data journals.

Analysis – The process of translating data into information utilized for an application.

Source: AHIMA, Data Quality Management Task Force, 1998.



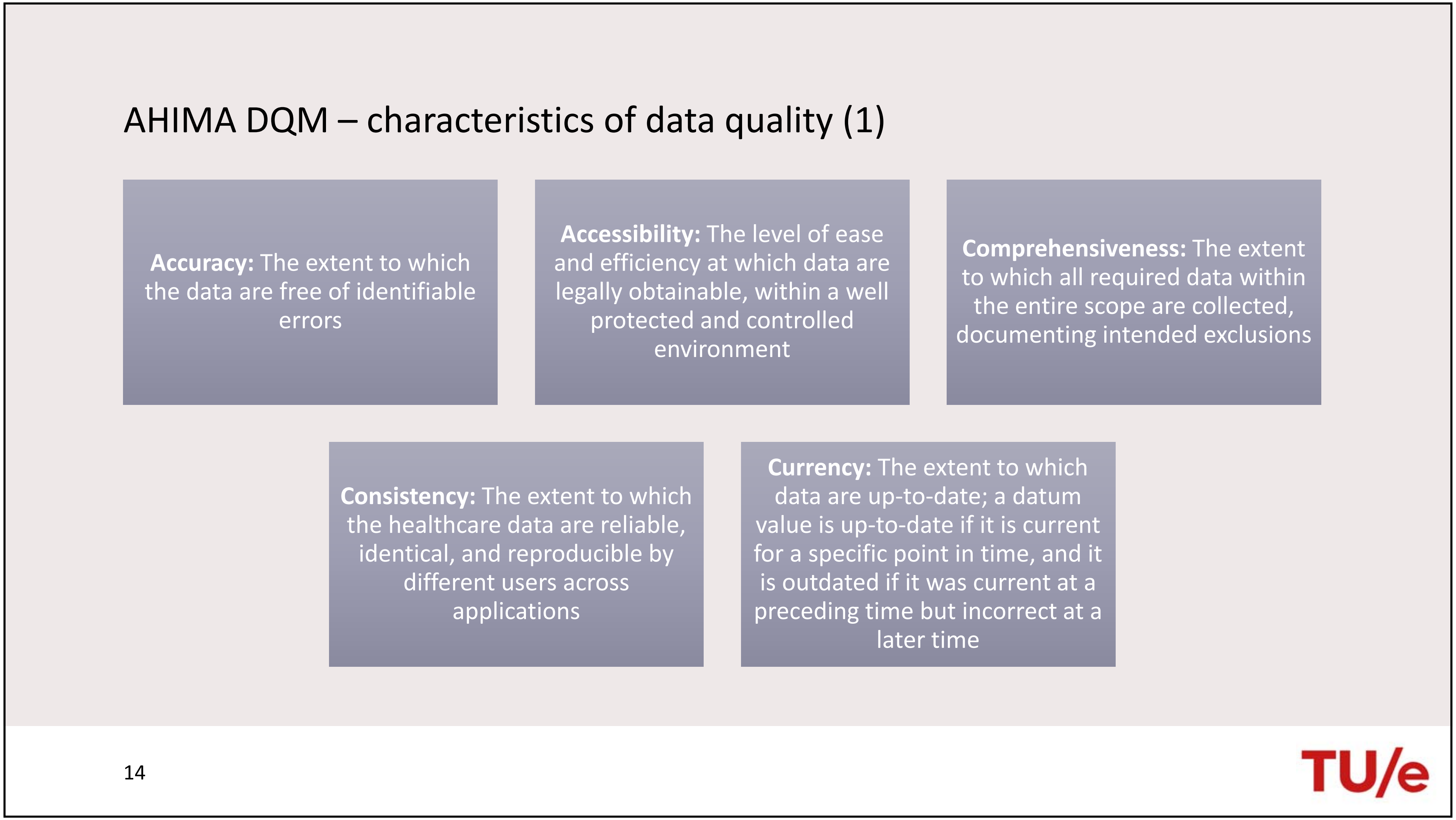
AHIMA DQM – data quality management functions

Application: The purpose for the data collection

Collection: The processes by which data elements are accumulated

Warehousing: Processes and systems used to archive data

Analysis: The process of translating data into meaningful information



AHIMA DQM – characteristics of data quality (1)

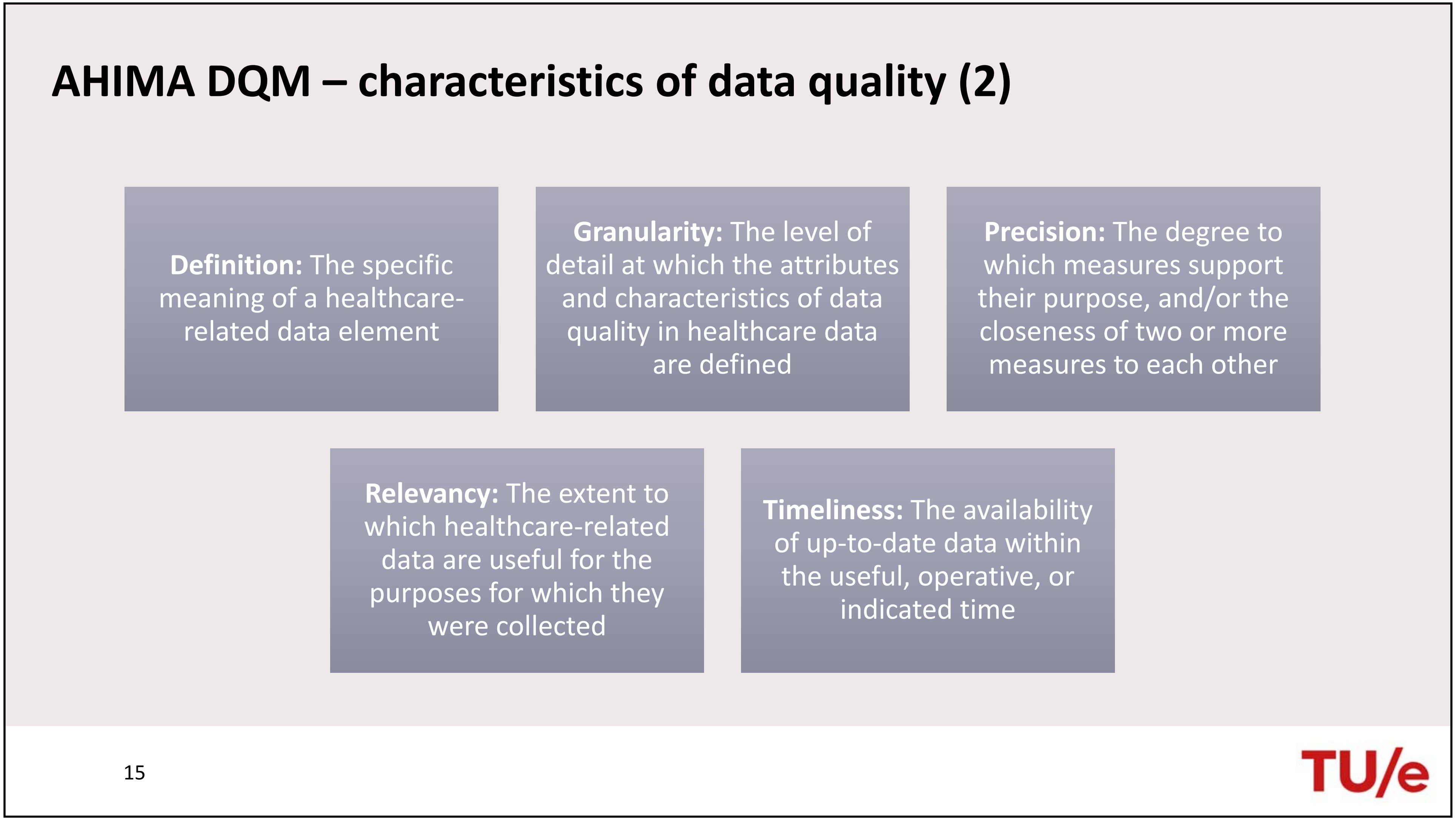
Accuracy: The extent to which the data are free of identifiable errors

Accessibility: The level of ease and efficiency at which data are legally obtainable, within a well protected and controlled environment

Comprehensiveness: The extent to which all required data within the entire scope are collected, documenting intended exclusions

Consistency: The extent to which the healthcare data are reliable, identical, and reproducible by different users across applications

Currency: The extent to which data are up-to-date; a datum value is up-to-date if it is current for a specific point in time, and it is outdated if it was current at a preceding time but incorrect at a later time



AHIMA DQM – characteristics of data quality (2)

Definition: The specific meaning of a healthcare-related data element

Granularity: The level of detail at which the attributes and characteristics of data quality in healthcare data are defined

Precision: The degree to which measures support their purpose, and/or the closeness of two or more measures to each other

Relevancy: The extent to which healthcare-related data are useful for the purposes for which they were collected

Timeliness: The availability of up-to-date data within the useful, operative, or indicated time

Exercise