Data/Information Quality Requirements – What is it and Why Does it Matter

MOHAMAD GHARIB

UNIVERSITY OF TARTU

Outline

- What is Data?
- The cost/benefits of low/high quality data
- What is high quality data?
- DQ Models
- DQ dimensions
- DQ in the wild
- Dealing with DQ requirements
- Dealing with DQ requirements – an example

What is Data?

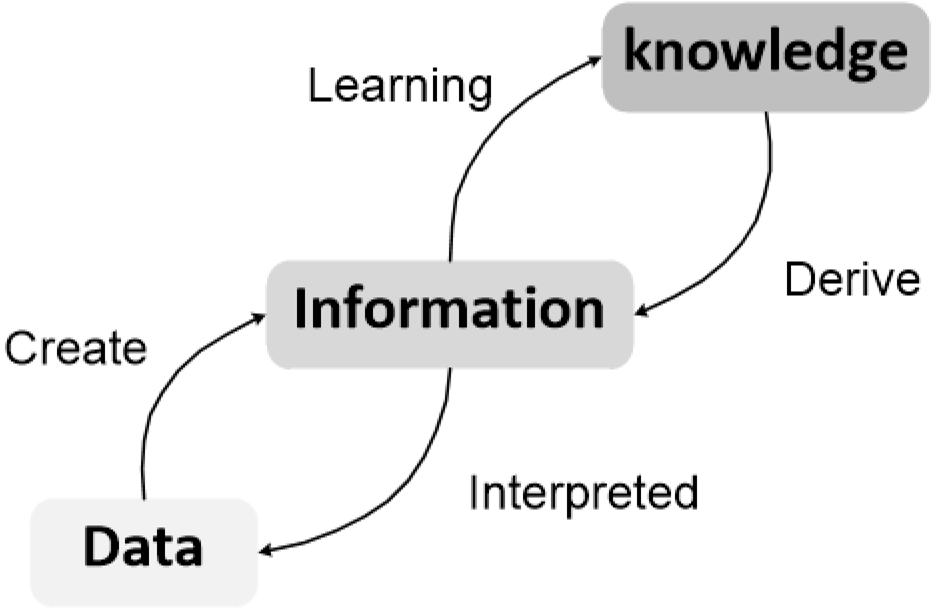
What is Data?

Data, **Information** and **Knowledge**:

Data can be defined as a **raw entity** with no meaning,

Information is **data** that has been given a **meaning**,

Knowledge is an appropriate collection of **information** along with the context on how it can be used to infer new information/knowledge.



The cost of low quality data

The **F-35 Joint Strike Fighter programs** - incorrectly detect targets in Formation (**inaccurate data due to** data fusion).

The **May 6, 2010, flash crash -** lasted for only 36 minutes and loses were around **$1 trillion** in the market value - (**inaccurate, incomplete, and inconsistent data**).

The bombing of the Chinese embassy in Belgrade – 1999 - (**out-dated data**).

Data quality cost	Costs caused by low-data quality	Direct cost	Verification cost
			Re-entry cost
			Compensation cost
		Indirect cost	Verification cost
			Re-entry cost
			Compensation cost
	Costs of improving or assuring data quality	Prevention cost	Training cost
			Monitoring cost
			Development cost
		Detection cost	Analysis cost
			Reporting cost
		Repair cost	Repair planning cost
			Reporting implimentation cost

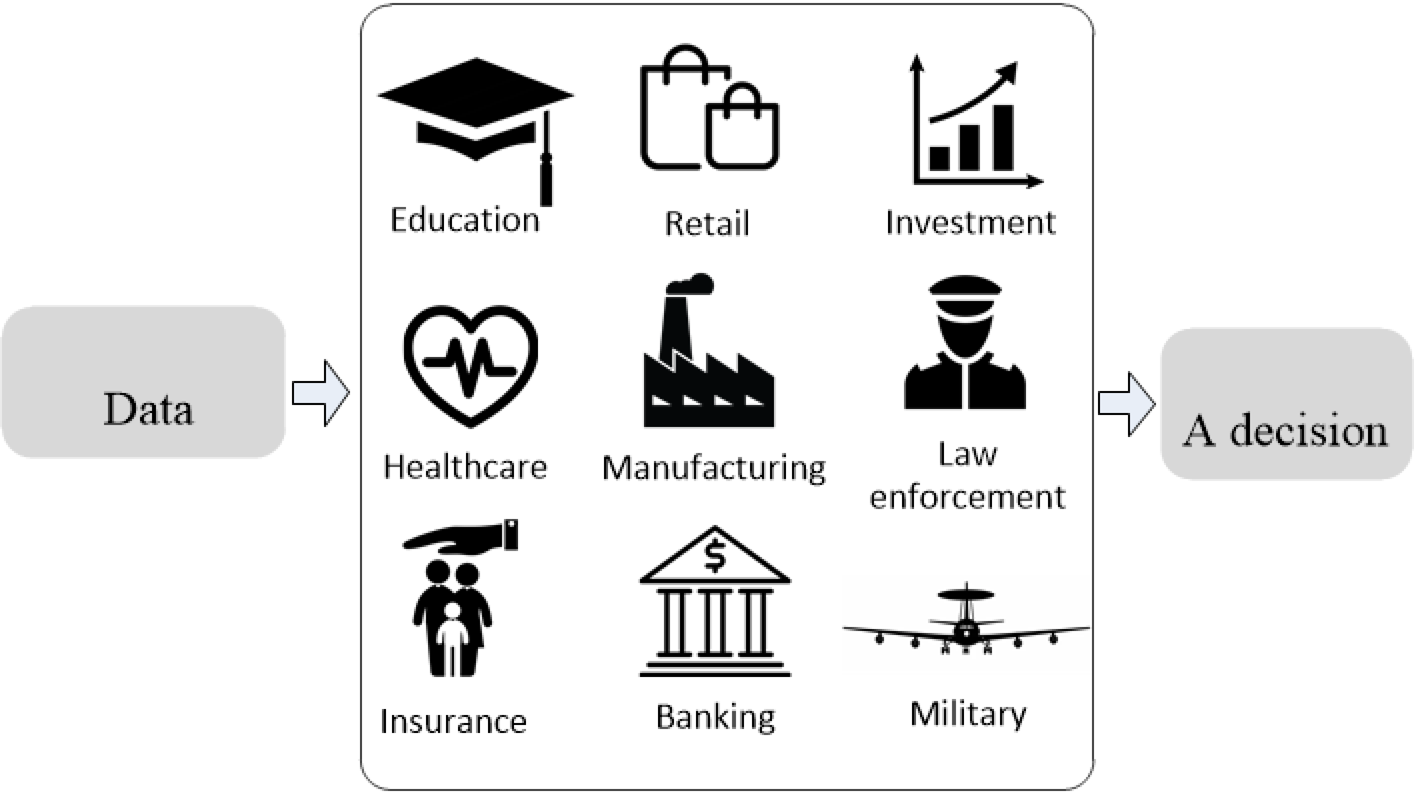
A data quality cost taxonomy[1]

[1] Eppler, M., & Helfert, M. (2004). *A classification and analysis of data quality costs*. MIT International Conference on Information Quality, November 5-6, 2004, Boston.

The need for high quality data

Almost all current organizations/complex systems require **data** to function properly.

These data is used to support **strategic**, **tactical** and **operational** decisions.



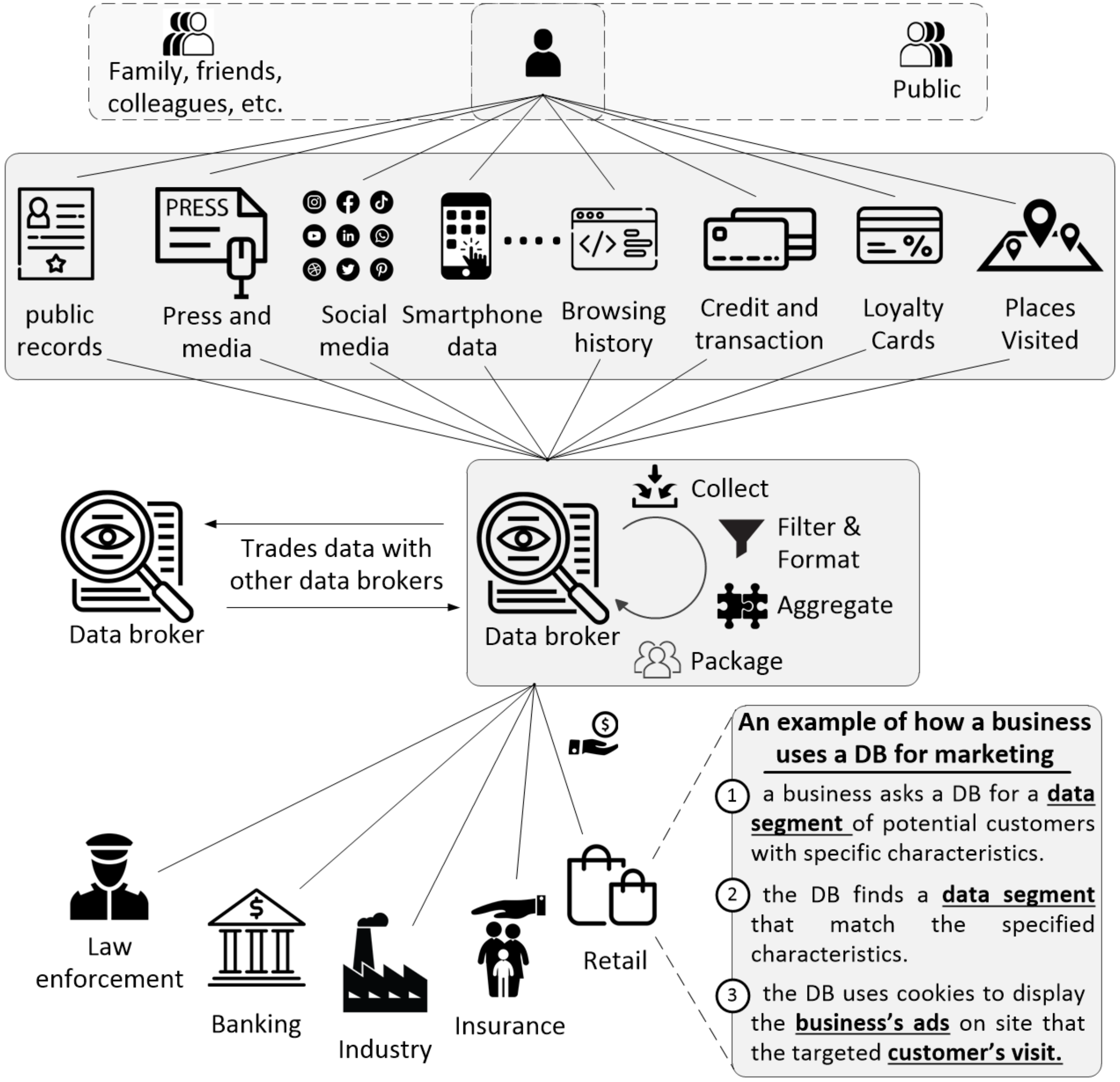
The need for high quality data - DBs

A data broker is a legal entity specializes in collecting [personal] data and selling or licensing such information to third parties for a variety of uses.

The global data brokers market was valued at **240.3** $ Bn in **2021**, and it is expected to reach **462.4** $ Bn by the end of **2031**.

Transparency Market Research Inc.

https://www.transparencymarketresearch.com/data-brokers-market.html



Family, friends, colleagues, etc.

Public

public records

Press and media

Social media

Smartphone data

Browsing history

Credit and transaction

Loyalty Cards

Places Visited

Data broker

Trades data with other data brokers

Data broker

Collect

Filter & Format

Aggregate

Package

Law enforcement

Banking

Industry

Insurance

Retail

An example of how a business uses a DB for marketing

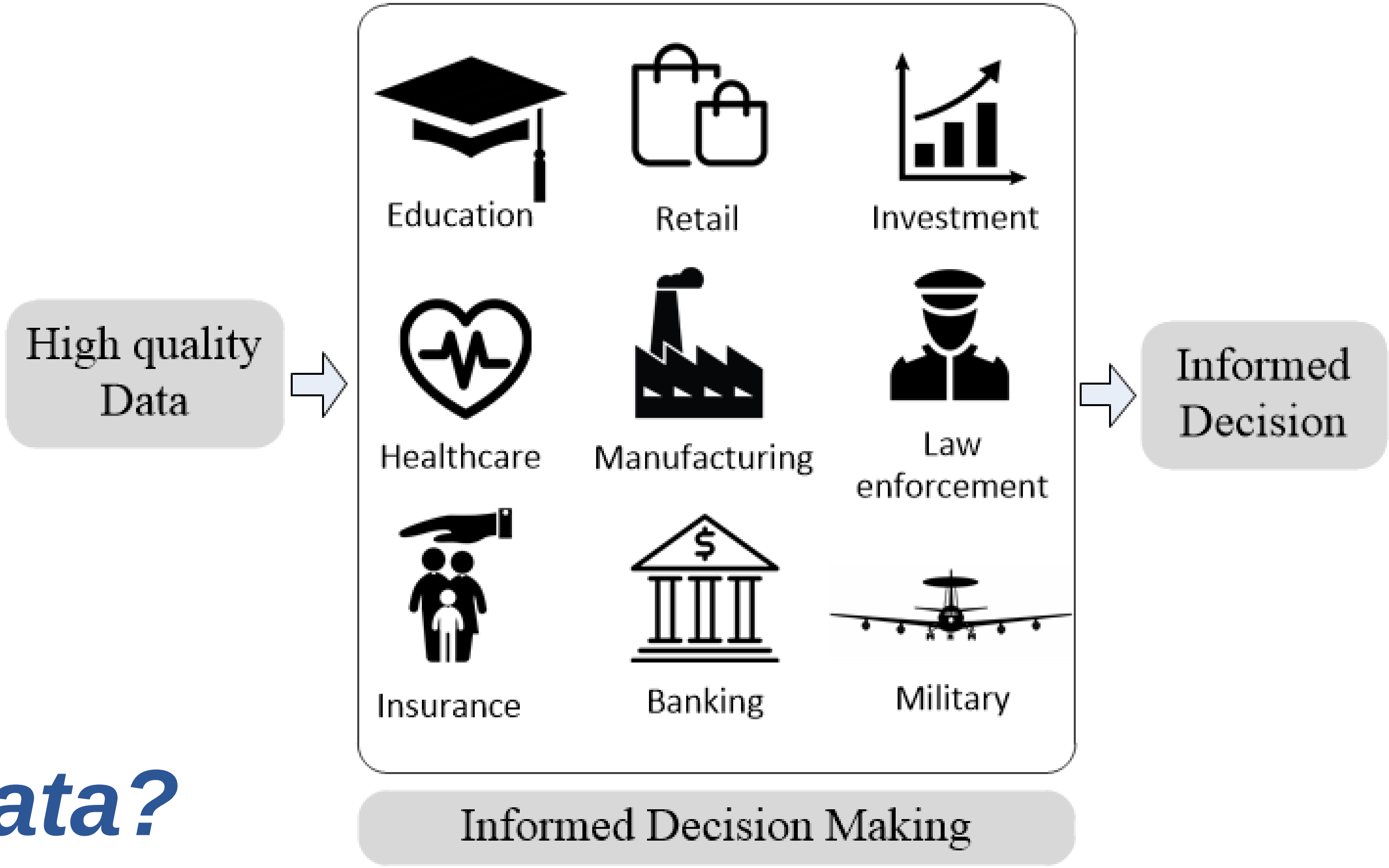
1. a business asks a DB for a **data segment** of potential customers with specific characteristics.
2. the DB finds a **data segment** that match the specified characteristics.
3. the DB uses cookies to display the **business's ads** on site that the targeted **customer's visit.**

What is high quality data?

Quality can be defined as "***fitness for use***", or the conformance to specifications.

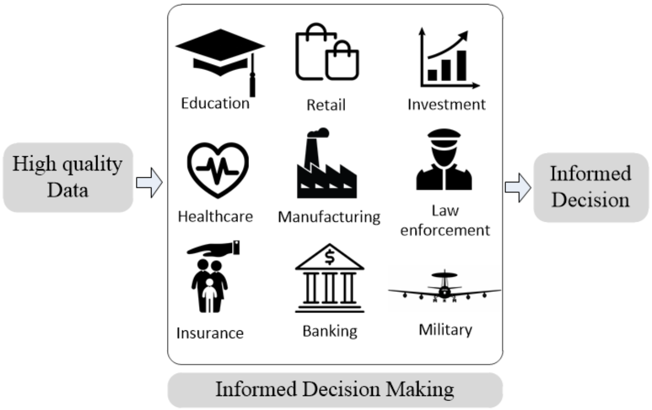
Determining whether data is of high or low quality depends on its "***fitness for use***".

What is high quality data?



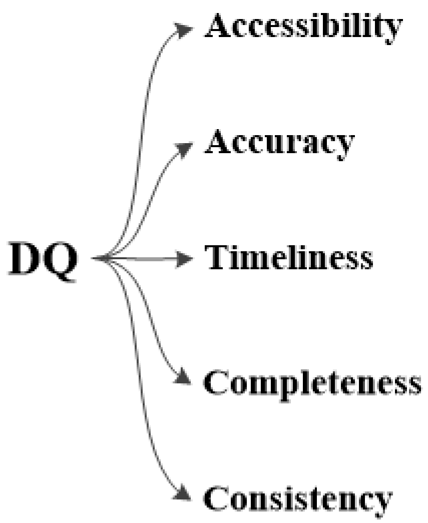
DQ Models

DQ Models

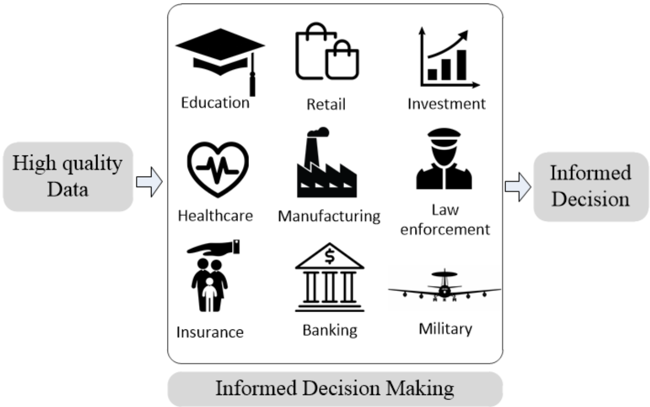

High quality Data
Education
Retail
Investment
Healthcare
Manufacturing
Law enforcement
Insurance
Banking
Military
Informed Decision
Informed Decision Making

DQ is a **hierarchical multi-dimensional concept that** can be analyzed depending on different dimensions and sub-dimensions.

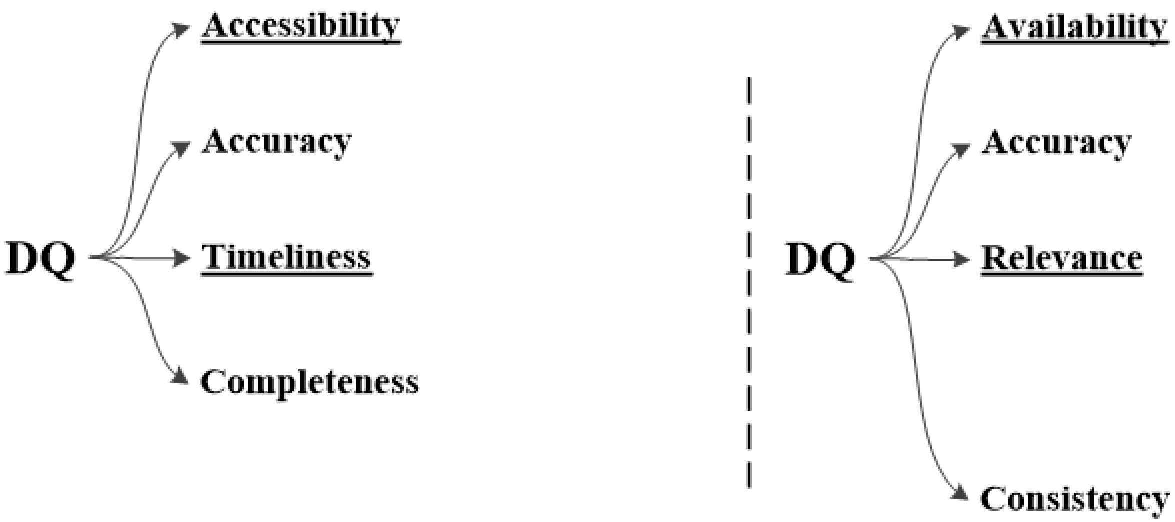
Many DQ **models** have been proposed.


DQ
Accessibility
Accuracy
Timeliness
Completeness
Consistency

DQ Models

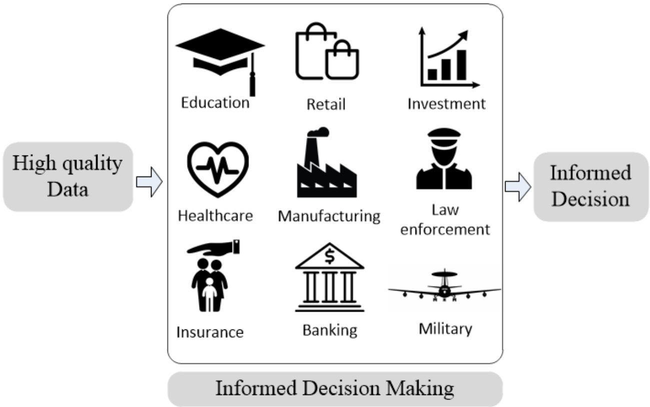


- **DQ** is a **hierarchical multi-dimensional concept that** can be analyzed depending on different dimensions and sub-dimensions.
- Many DQ **models** have been proposed.



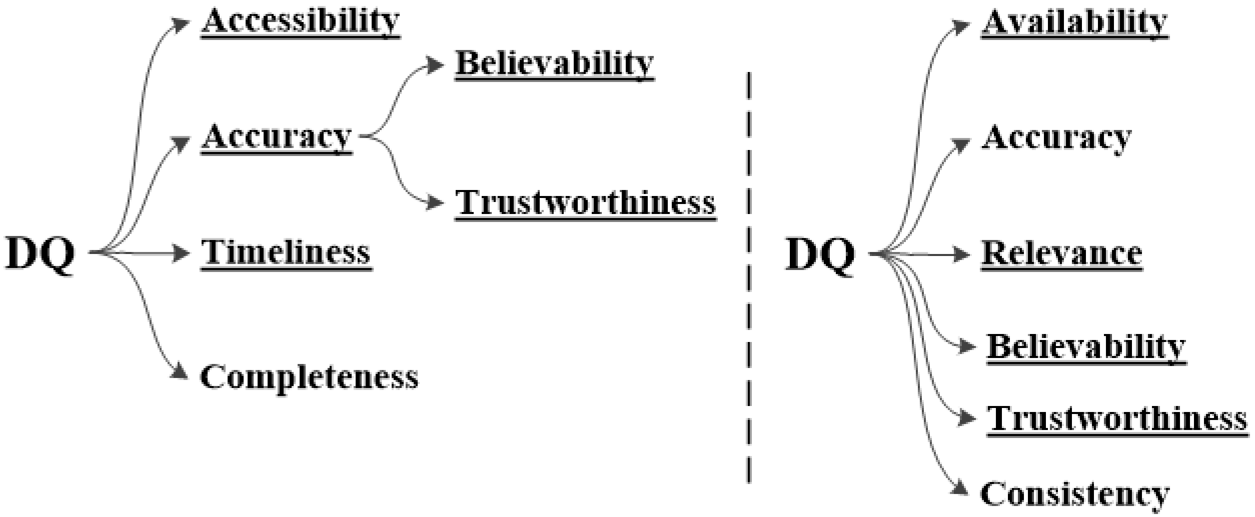
- Most **models** are not **consist** among the **dimensions they consider**, the **inter-relationships** among these **dimensions**, and even the **definitions** of these **dimensions** and how they can be **analyzed**.

DQ Models



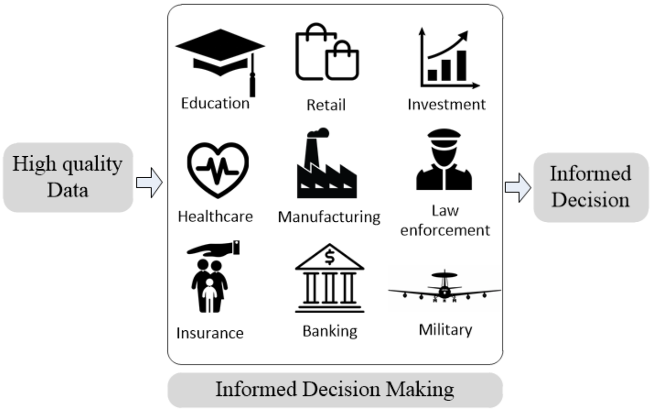
DQ is a **hierarchical multi-dimensional concept that** can be analyzed depending on different dimensions and sub-dimensions.

Many DQ **models** have been proposed.

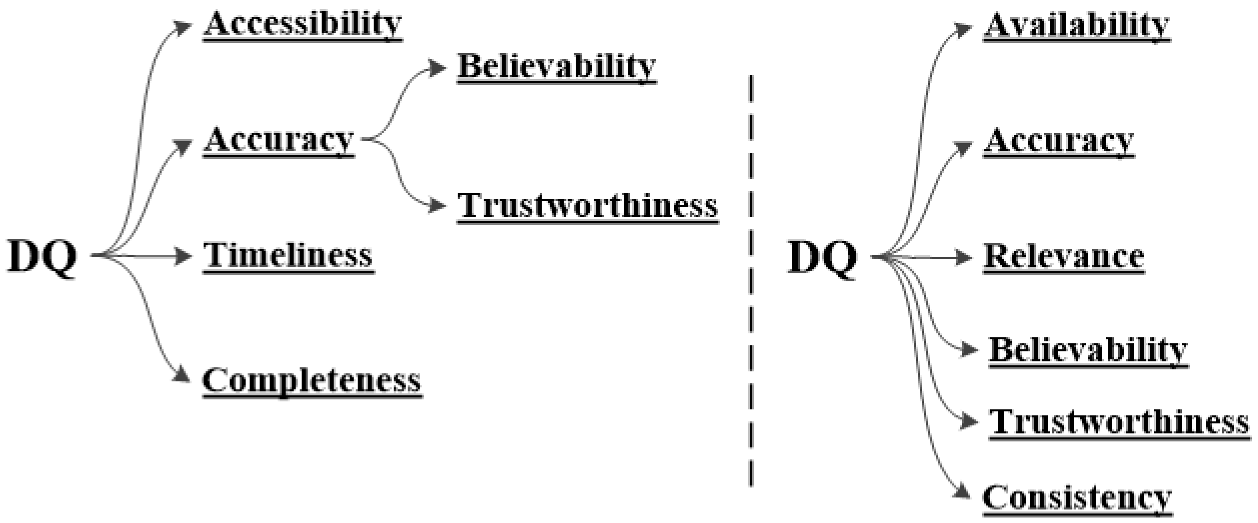


Most **models** are not **consist** among the **dimensions they consider**, the **inter-relationships** among these **dimensions**, and even the **definitions** of these **dimensions** and how they can be **analyzed**.

DQ Models



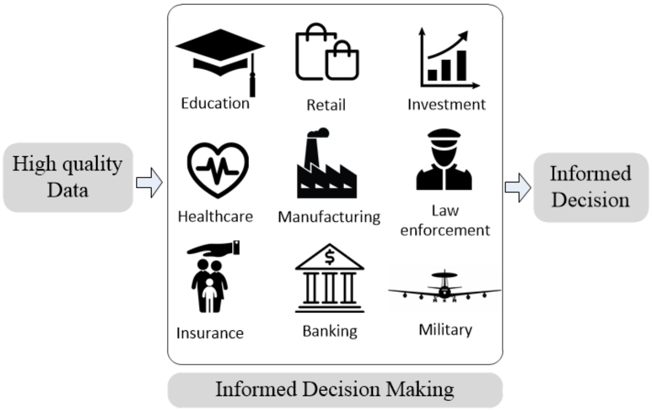
DQ is a **hierarchical multi-dimensional concept that** can be analyzed depending on different dimensions and sub-dimensions.



Many DQ **models** have been proposed.

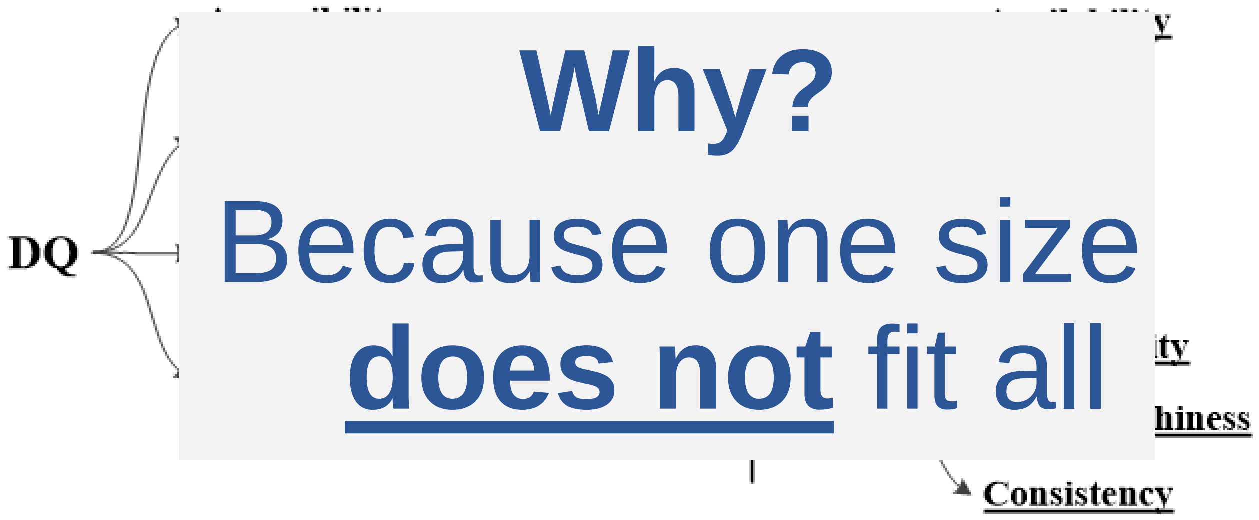
Most **models** are not **consist** among the **dimensions they consider**, the **inter-relationships** among these **dimensions**, and even the **definitions** of these **dimensions** and how they can be **analyzed**.

DQ Models



DQ is a **hierarchical multi-dimensional concept that** can be analyzed depending on different dimensions and sub-dimensions.

Many DQ **models** have been proposed.



Why?

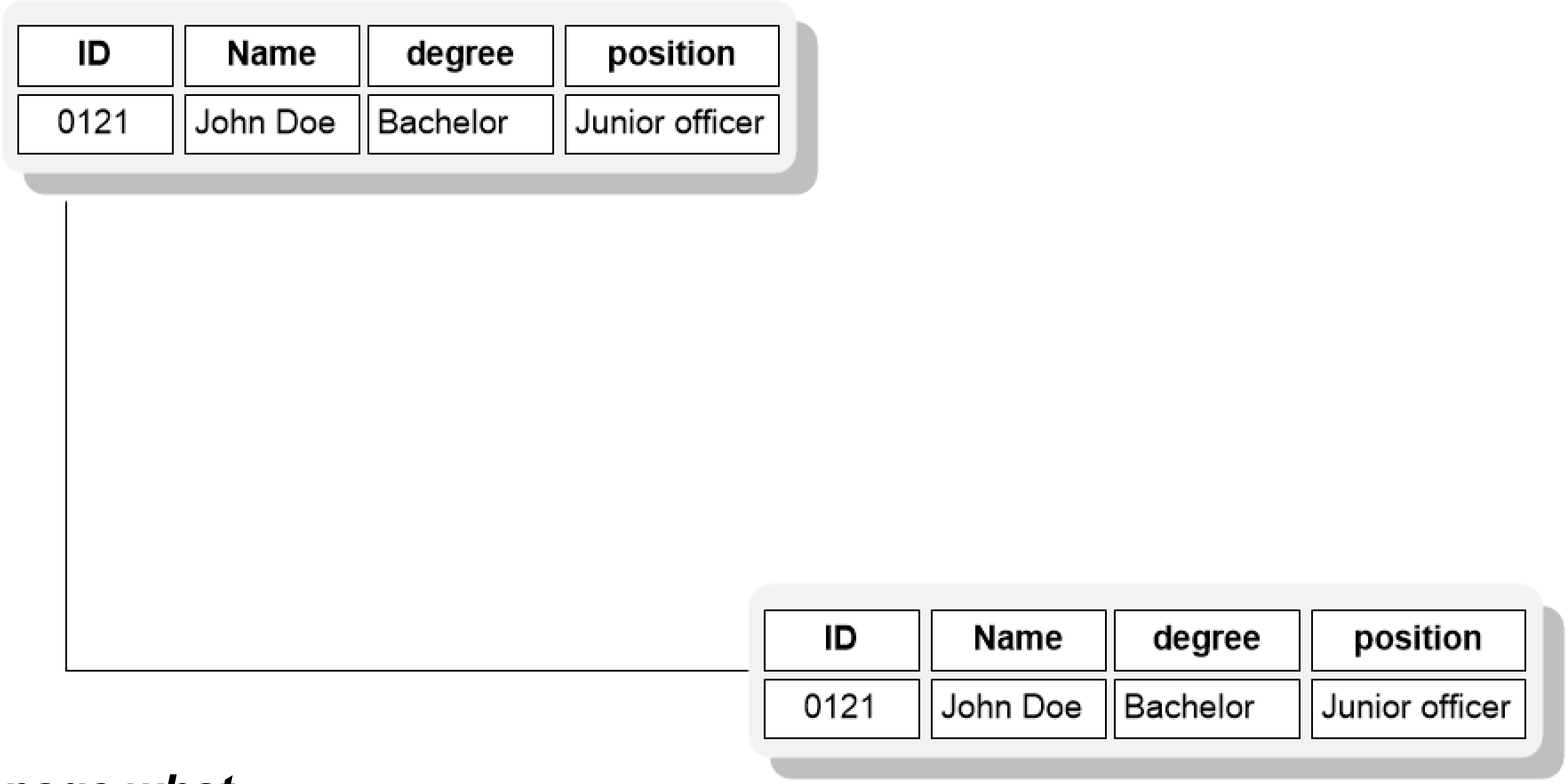
Because one size **does not** fit all

Most **models** are not **consist** among the **dimensions they consider**, the **inter-relationships** among these **dimensions**, and even the **definitions** of these **dimensions** and how they can be **analyzed**.

DQ dimensions

DQ dimensions – definitions & problems

- Reference data -

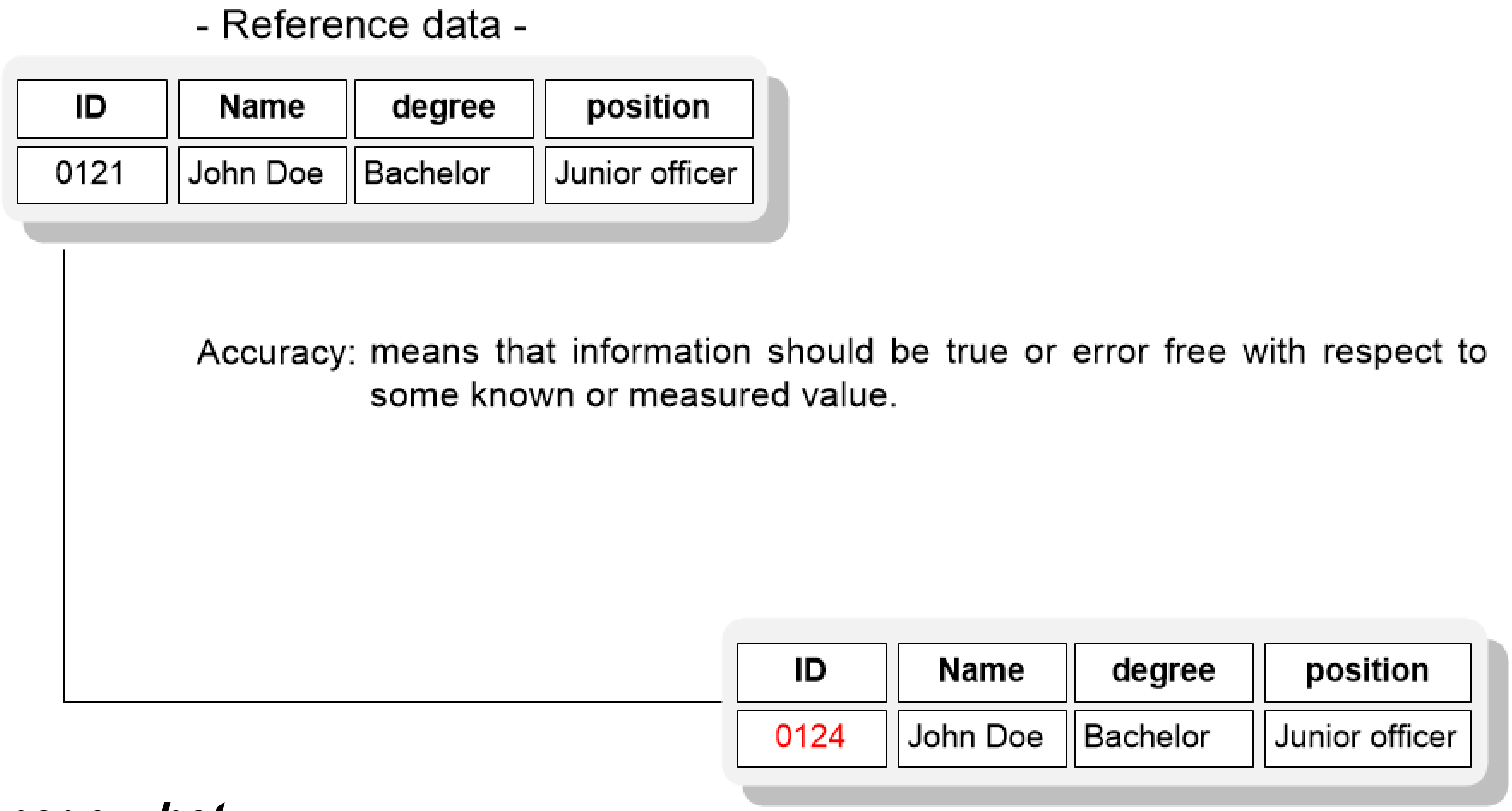


ID	Name	degree	position
0121	John Doe	Bachelor	Junior officer

ID	Name	degree	position
0121	John Doe	Bachelor	Junior officer

“You can't manage what you can't **measure*”***
- Edwards Deming

DQ dimensions – definitions & problems



- Reference data -

ID	Name	degree	position
0121	John Doe	Bachelor	Junior officer

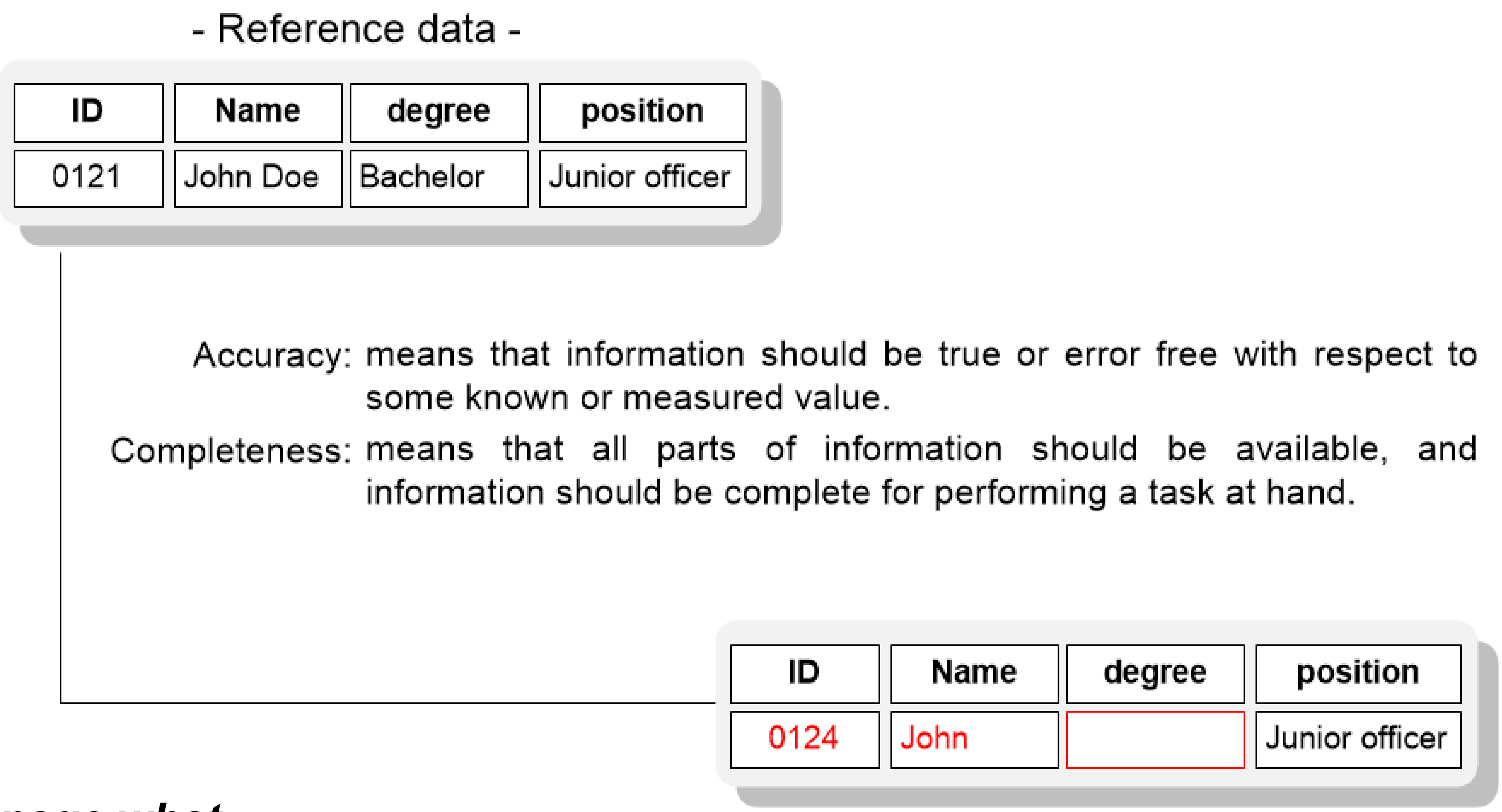
Accuracy: means that information should be true or error free with respect to some known or measured value.

ID	Name	degree	position
0124	John Doe	Bachelor	Junior officer

“You can't manage what you can't measure”

- Edwards Deming

DQ dimensions – definitions & problems



- Reference data -

ID	Name	degree	position
0121	John Doe	Bachelor	Junior officer

Accuracy: means that information should be true or error free with respect to some known or measured value.

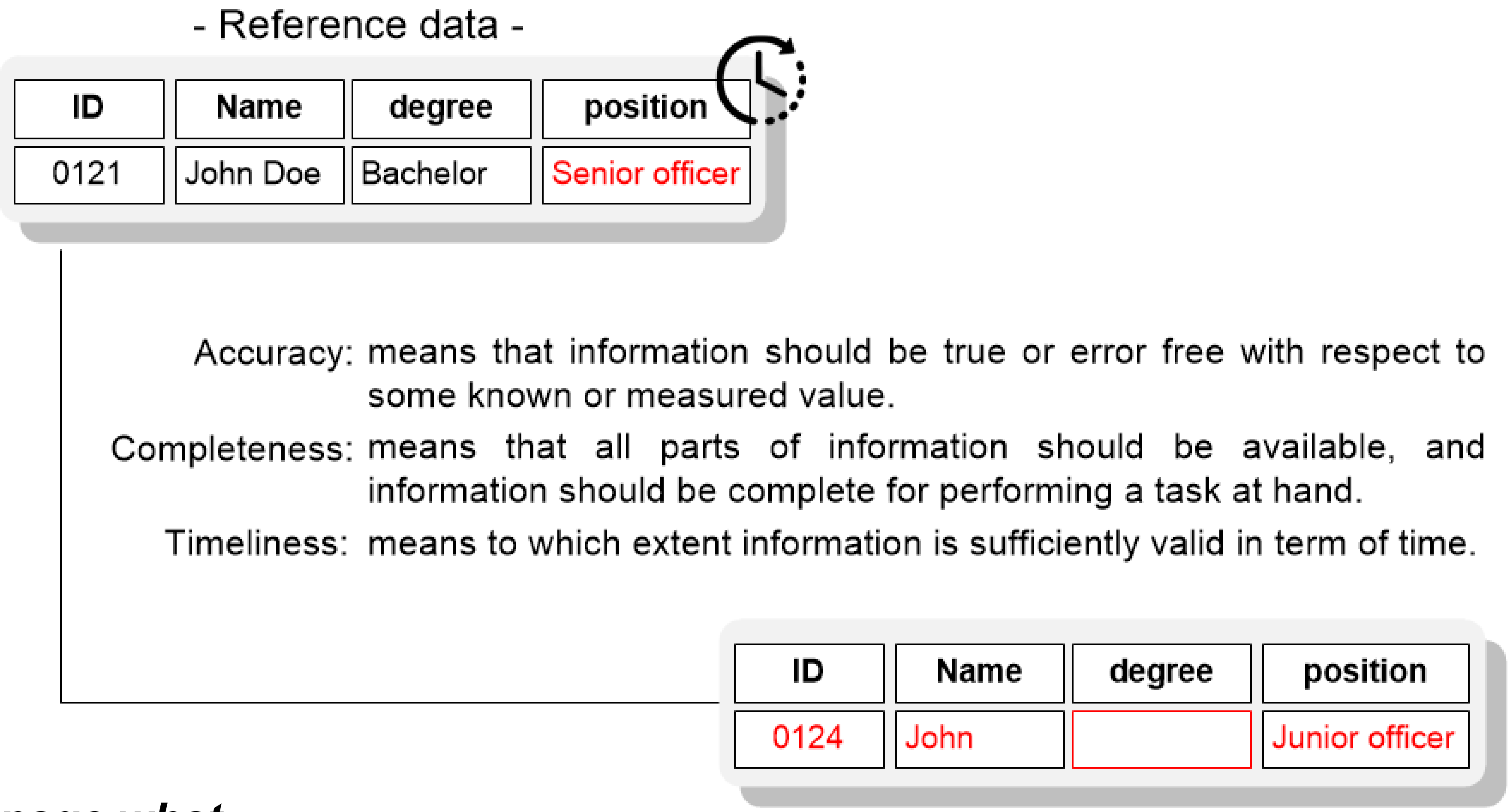
Completeness: means that all parts of information should be available, and information should be complete for performing a task at hand.

ID	Name	degree	position
0124	John		Junior officer

“You can't manage what you can't measure”

- Edwards Deming

DQ dimensions – definitions & problems



- Reference data -

ID	Name	degree	position
0121	John Doe	Bachelor	Senior officer

Accuracy: means that information should be true or error free with respect to some known or measured value.

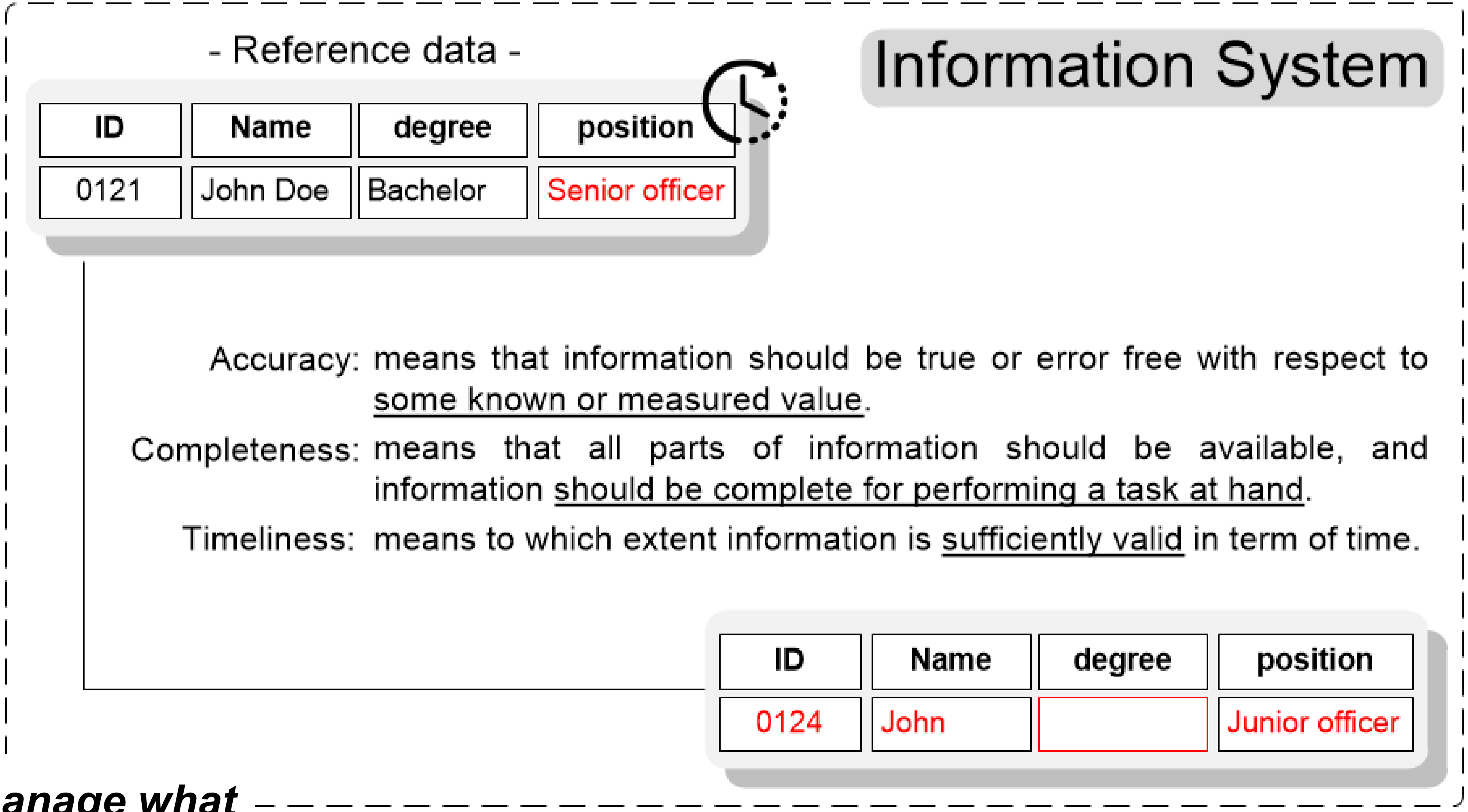
Completeness: means that all parts of information should be available, and information should be complete for performing a task at hand.

Timeliness: means to which extent information is sufficiently valid in term of time.

ID	Name	degree	position
0124	John		Junior officer

*“You can't manage what you can't **measure**”*
- Edwards Deming

DQ dimensions – definitions & problems



- Reference data -

ID	Name	degree	position
0121	John Doe	Bachelor	Senior officer

Information System

Accuracy: means that information should be true or error free with respect to some known or measured value.

Completeness: means that all parts of information should be available, and information should be complete for performing a task at hand.

Timeliness: means to which extent information is sufficiently valid in term of time.

ID	Name	degree	position
0124	John		Junior officer

“You can't manage what you can't ***measure****”*
- Edwards Deming

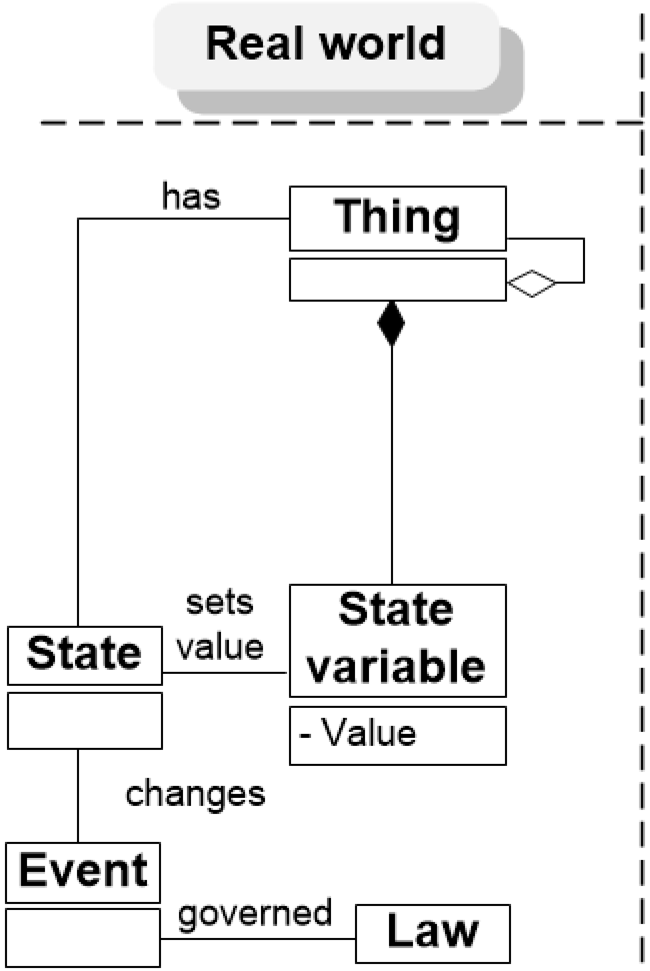
DQ in the wild

DQ in the wild

A ``**real world**'' is made up of **things** that can be **composite**.

A **thing** has **state(s)** that are represented by **state variable(s)**.

A **state** can be **changed** by an event, which is governed by a **law**.



[2]**A simplified and partial mapping between the real world and an IS**

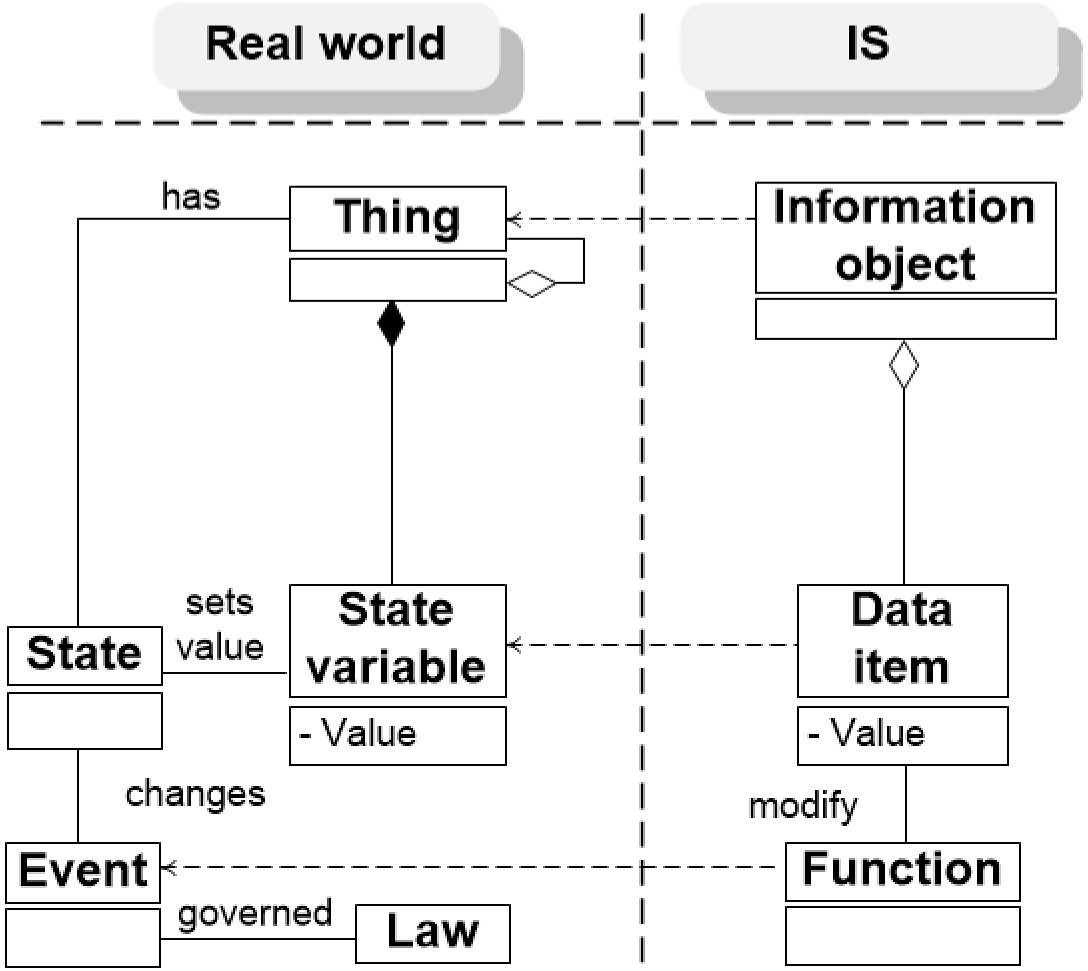
[2] Bunge, M. Treatise on Basic Philosophy: Ontology I/II, Reidel, 1977/1979.

DQ in the wild

A **thing** can be **represented** in **IS** by **information objects**.

An **information objects** has a defined set of **data items**, whose **value** should reflects the **value** of a corresponding **state variable**.

An **event** is reflected by a **functions** of the **IS**.



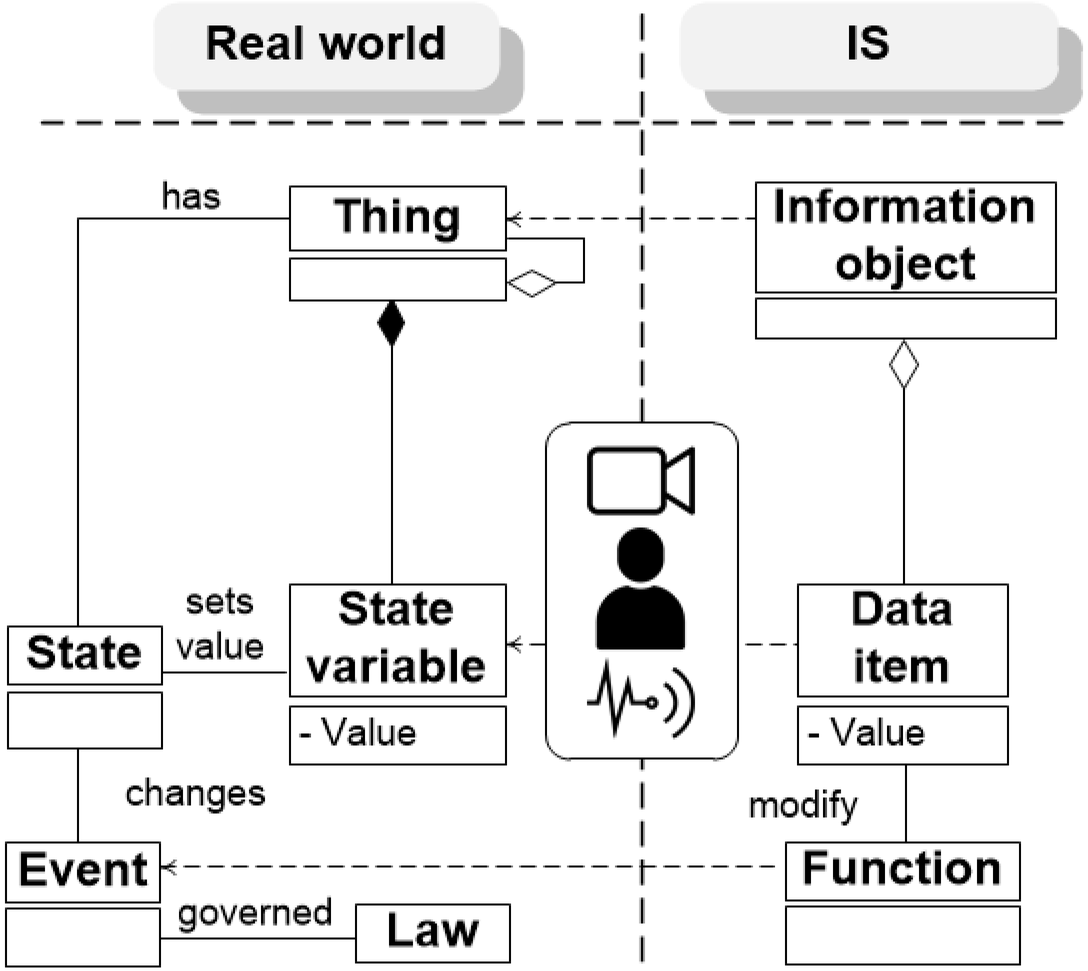
[2]A simplified and partial mapping between the real world and an IS

[2]Bunge, M. Treatise on Basic Philosophy: Ontology I/II, Reidel, 1977/1979.

DQ in the wild

Various means can be used for acquiring the **value** of a State variable and assign it to the **value** of a Data item (e.g., sensors, cameras, and even humans).

Each of these means might has its **advantages** and **disadvantages**.



[2]A simplified and partial mapping between the real world and an IS

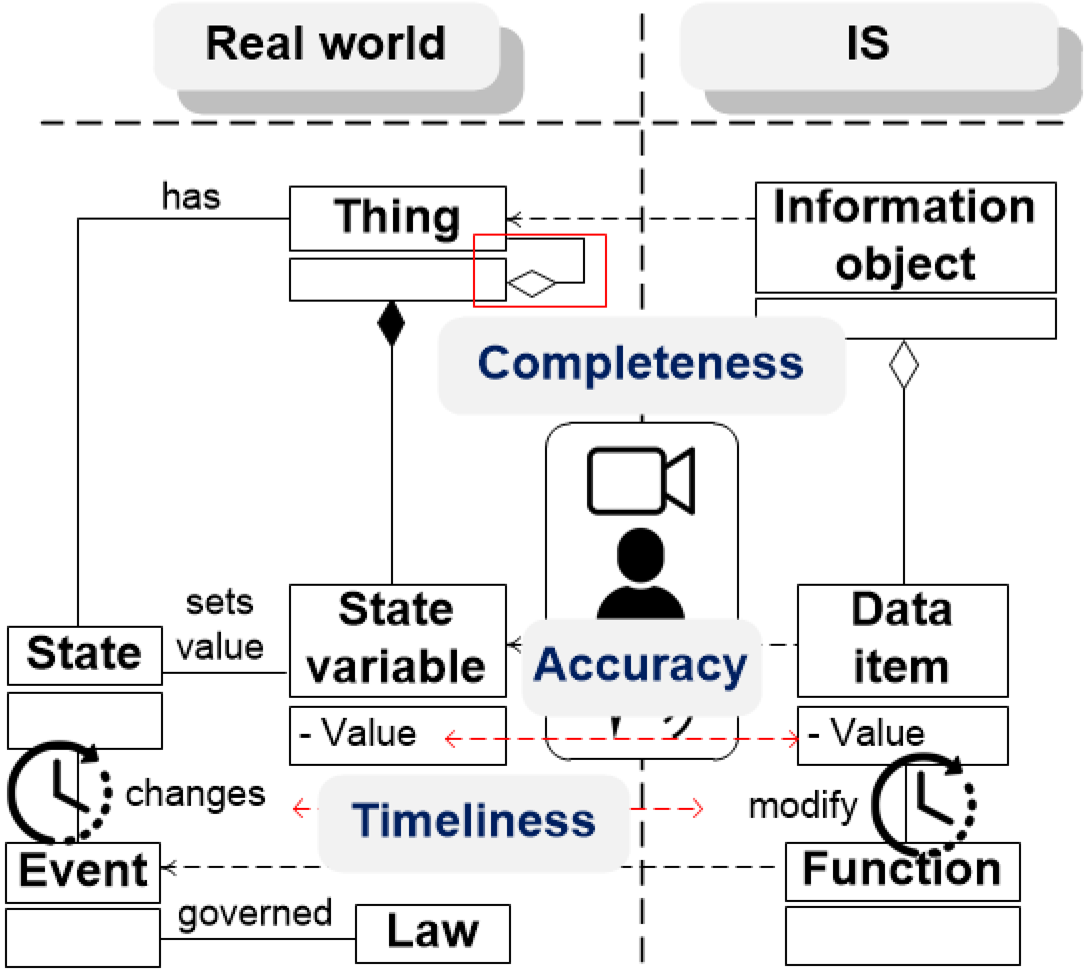
[2]Bunge, M. Treatise on Basic Philosophy: Ontology I/II, Reidel, 1977/1979.

DQ in the wild

Accuracy can be determined by analysing whether the value of a data item correctly representing the value of its corresponding State variable.

Completeness can be determined by analysing whether data is complete for the purpose of use.

Timeliness can be determined by analysing whether the currency (age) of the value of a data item is smaller than the volatility interval of the value of its corresponding State variable.



[2]A simplified and partial mapping between the real world and an IS

[2]Bunge, M. Treatise on Basic Philosophy: Ontology I/II, Reidel, 1977/1979.

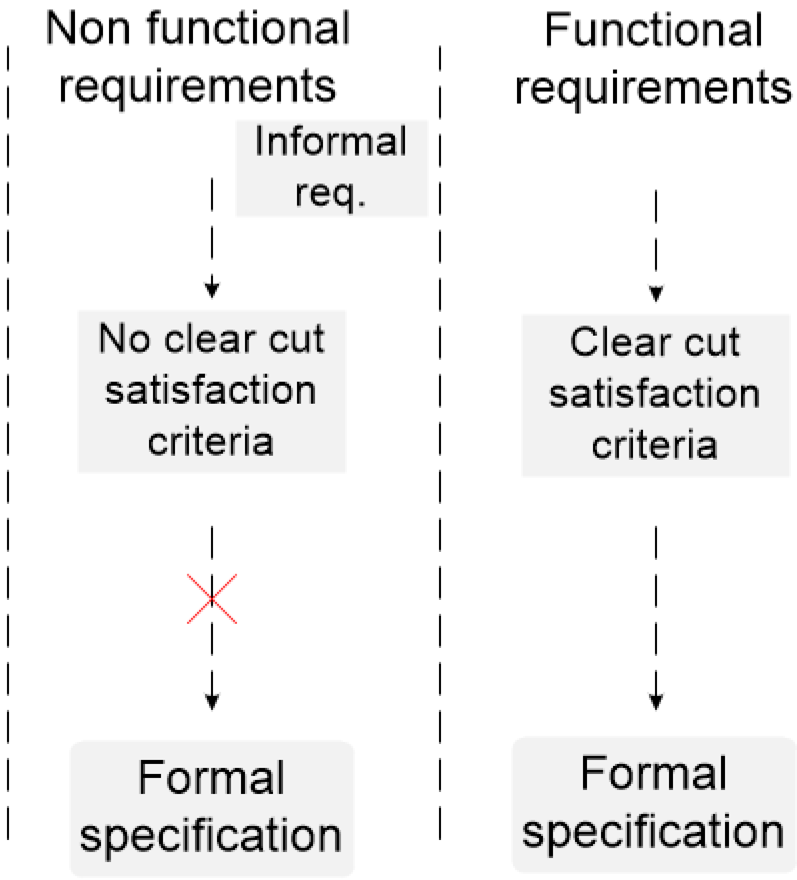
Dealing with DQ requirements

Dealing with DQ requirements

Requirements can be classified under functional and non-functional (quality) requirements

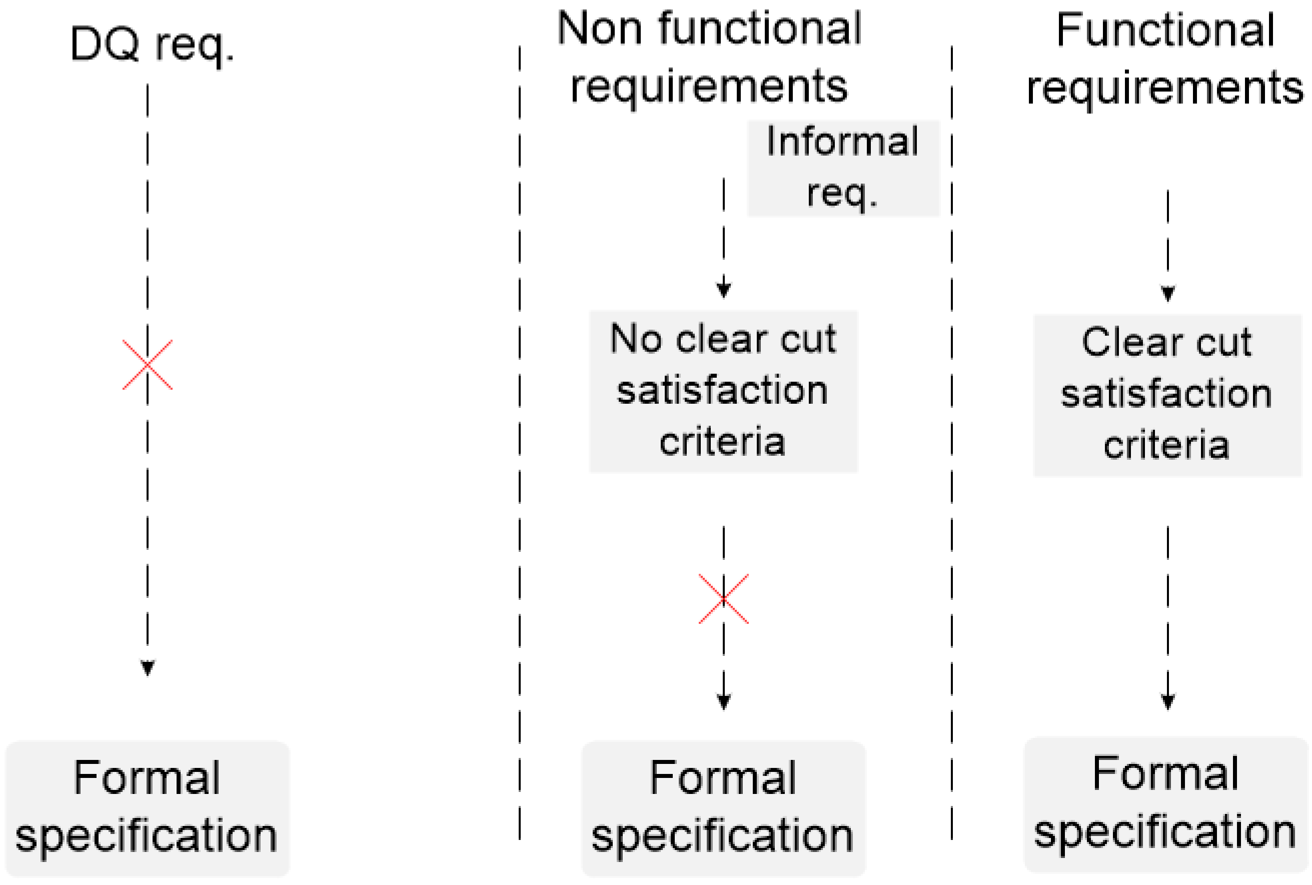
FR refers to the functionalities/services that the system should deliver

NFR refer to qualities that the system needs to satisfy while delivering the aforementioned services.


Non functional requirements
Informal req.
No clear cut satisfaction criteria
Formal specification
Functional requirements
Clear cut satisfaction criteria
Formal specification

Dealing with DQ requirements

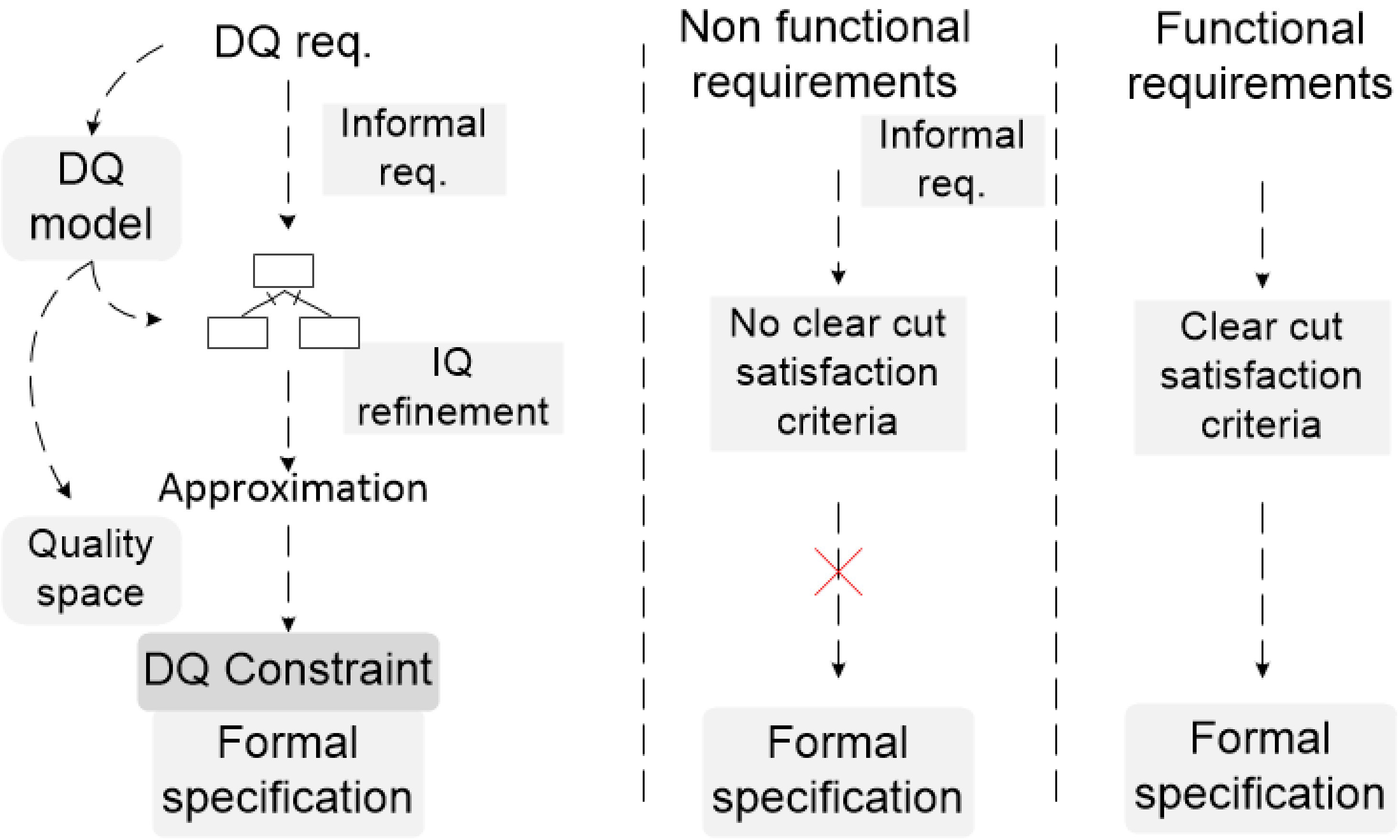
DQ requirements use to be represented as generic qualitative properties without specific methods for their analysis.



Dealing with DQ requirements

DQ requirements use to be represented as generic qualitative properties without specific methods for their analysis.

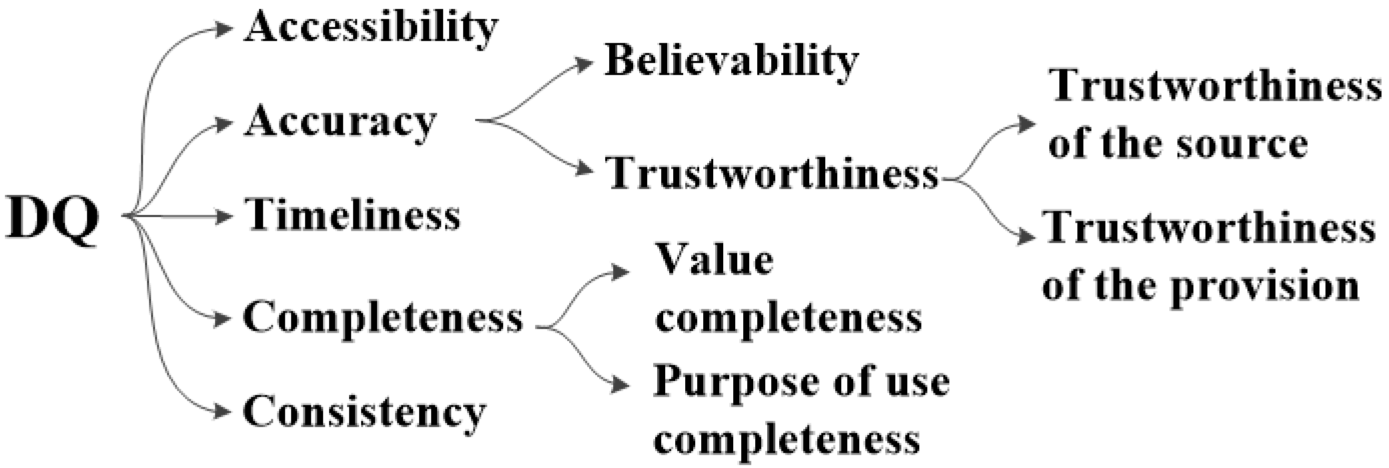
We proposed an approach for capturing DQ requirements at a high-level of abstraction and gradually refining them until they can be approximated into DQ constraints.



Dealing with DQ requirements – an example

The Flash crash

The **May 6, 2010, flash crash -** lasted for only 36 minutes and loses were around **$1 trillion** in the market value - (**inaccurate, incomplete, and inconsistent data**).

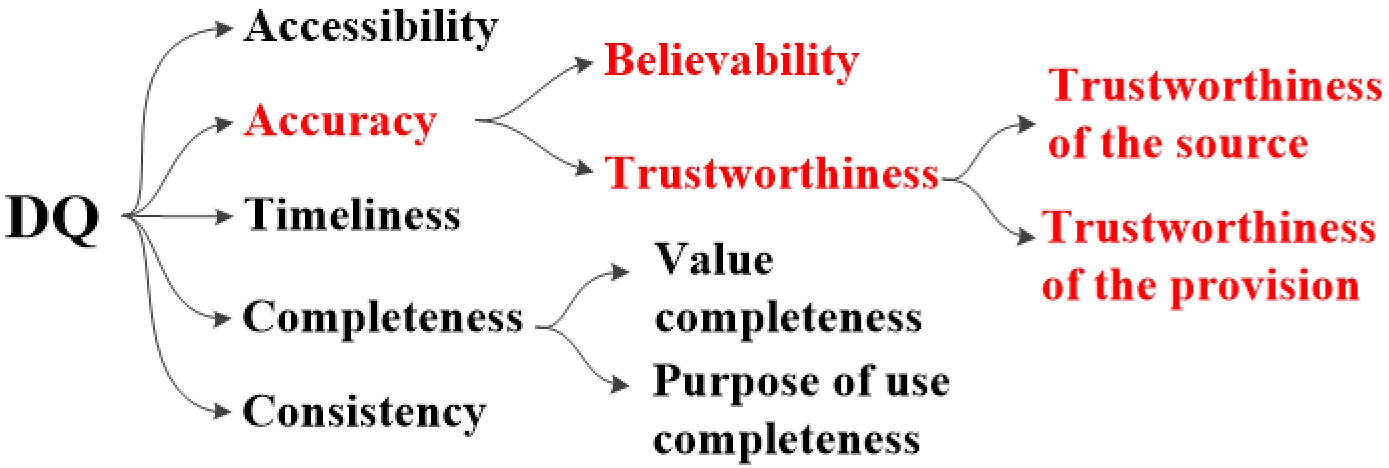


"All ***models*** *are wrong, but some are useful" - George Box*

The Flash crash

Two reasons contributed to the crash:

1. The behaviour of some High-frequency traders (**HFTs**) that used **flickering quotes** (e.g., falsified, inaccurate) to compromise the overall system performance.



DQ

Accessibility

Accuracy

Believability

Trustworthiness

Trustworthiness of the source

Trustworthiness of the provision

Timeliness

Completeness

Value completeness

Purpose of use completeness

Consistency

“If you torture the data long enough, it will confess”

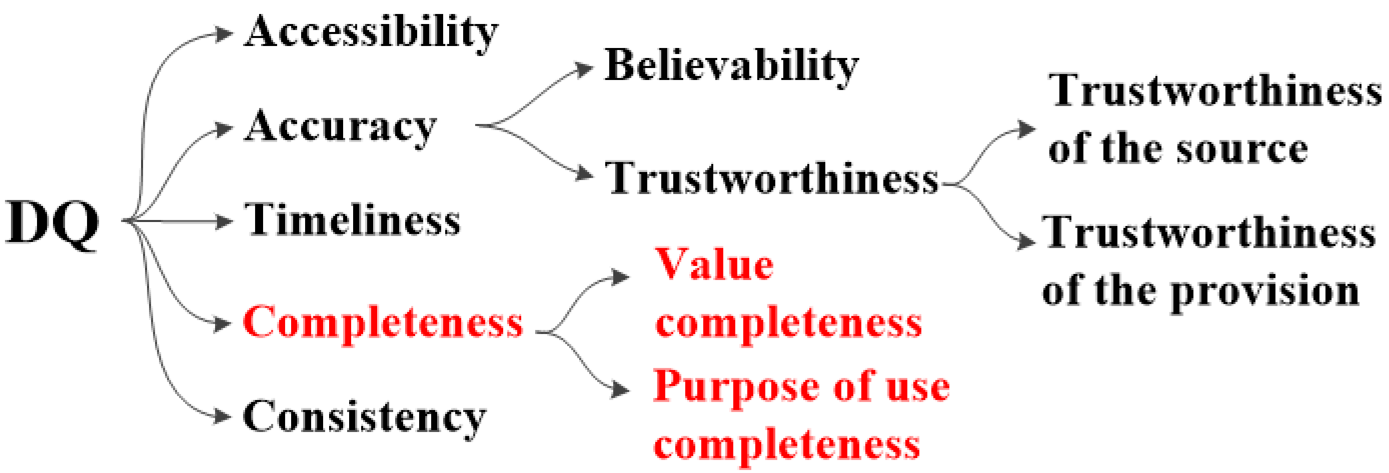
- Ronald Coase

The Flash crash

Two reasons contributed to the crash:

1. The behaviour of some High-frequency traders (**HFTs**) that used **flickering quotes** (e.g., falsified, inaccurate) to compromise the overall system performance.

2. The highly fragmented nature of the market along with the **inefficient coordination mechanisms** among the Circuit Breakers (**CBs**) of the trading markets.



DQ

Accessibility

Accuracy

Timeliness

Completeness

Consistency

Believability

Trustworthiness

Value completeness

Purpose of use completeness

Trustworthiness of the source

Trustworthiness of the provision

"If you torture the data long enough, it will confess"
- Ronald Coase

UNIVERSITY OF TARTU 1632

Thank You
for your attention

Mohamad Gharib
mohamad.gharib@ut.ee

unitartu

tartuuniversity