Data Quality

Introduction

Real World

In the *real world,* activities are implemented in the field. These activities are designed to produce results that are quantifiable.

Data Management System

***An information system* represents these activities by collecting the results that were produced and mapping them to a recording system.**

- Data Quality (DQ) is often defined as "fitness for use," which implies the relative nature of the concept
- Why Data Quality?
  - High quality data coupled with advanced technology will yield added values

Introduction

- It is possible that data can be of high quality in one decision context, but perceived to be poor in another decision context.
- DQ is a multi-dimensional concept and every dimension represents a single aspect and also comprises both objective and subjective aspects.
- Hence, it’s useful to define DQ in terms of its dimensions.

Data Quality Dimensions

Category	**Dimension**	**Definition**
Intrinsic	Accuracy	Data are regarded as correct
	Believability	Data are accepted as true, real and credible
	Objectivity	Data are unbiased and impartial
	Reputation	Data are trusted or highly regarded in terms of their source and content
Contextual	Value-added	Data are beneficial and provide advantages for their use
	Completeness	Data values are present
	Relevancy	Data are applicable and useful for the task at hand
	Appropriate amount of data	The quantity or volume of available data is appropriate
Representational	Interpretability	Data are in appropriate language and unit and the data definitions are clear
	Ease of understanding	Data are clear without ambiguity and easily comprehended
Accessibility	Accessibility	Data are available or easily and quickly retrieved
	Security	Access to data can be restricted and hence kept secure

Data Quality - Accuracy

- Accuracy indicates whether the data stored are the correct values.
- For example if my birthdate is February 27, 1975, for a database that expects dates in USA format, 02/27/1975 is the correct value.
- However, for a database that expects a European representation, the date 02/27/1975 is incorrect; Instead 27/02/1975 is the correct value.

Data Quality - Completeness

- Schema completeness refers to the extent to which entities and attributes are not lacking from the schema
- Column completeness verifies whether a column of a table has missing values or not
- Population completeness refers to the degree to which members of the population are not present (See the following example)

Example: Population Completeness

ID	Name	Surname	Birth Date	Email
1	Monica	Smith	04/10/1988	smith@gmail.com
2	Yuki	Pitt	04/03/1968	Null [a]
3	Rose	David	02/01/1975	Null [b]
4	John	Edward	05/09/1990	Null [c]

a – Not existing
b – Existing but unknown
c – Not known if existing

- Tuple 2: Since the person represented by tuple 2 has no email address, we can say that the tuple is complete.
- Tuple 3: Since the person represented by tuple 3 has an email, but its value is not known, we can say that the tuple is incomplete.
- Tuple 4: If we do not know the person represented by tuple 4 has an email or not, incompleteness may not be the case.

Data Quality - Believability

- The extent to which data is regarded as true and credible.

Data Quality - Accessibility

- Accessibility refers to how easy the data can be located and retrieved. (It is important that the data can be accessed and delivered on time, so as to not needlessly delay important decisions.)

Data Quality - Consistency

- Can be addressed from various perspectives:

i. Presence of redundant data (e.g. name, address etc.) in multiple data sources.

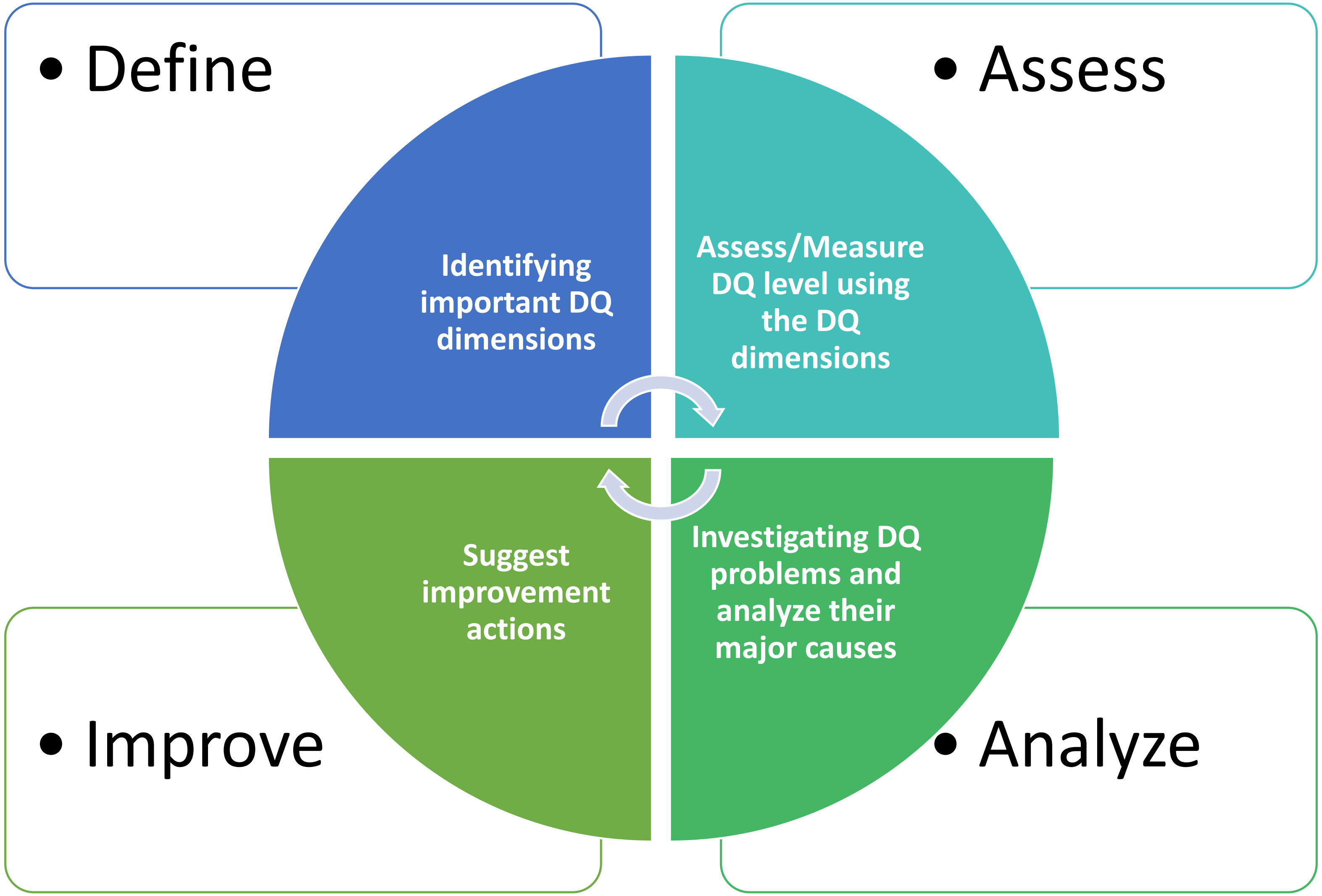
ii. Consistency between related data attributes. For example, city name and zip code should be corresponding.

iii. Data format used. For example, gender can be encoded as male/female, M/F, or 0/1.

Various DQ Problem Causes

- Multiple data sources: Multiple sources of the same data may produce duplicates; a consistency problem.
- Subjective judgment: Subjective judgment can create data bias; objectivity problem.
- Limited computing facilities: Lack of sufficient computing facilities limits data access; accessibility problem.
- Size of data: Big data can give high response times; accessibility problem.

How to Improve Data Quality?



- Define

Identifying important DQ dimensions

- Assess

Assess/Measure DQ level using the DQ dimensions

Investigating DQ problems and analyze their major causes

- Analyze

Suggest improvement actions

- Improve

Data Quality Management Program

Privacy

- 2 major problems
  1. First, data about individuals can be collected without these individuals being aware of it
  2. Second, people may be aware that data is collected about them, but have no say in how the data is being used

Additional Concerns

- Conventional: Simple data collection and data retrieval from databases
- Data Analytics: Uses massive amounts of data—possibly combined from several sources, including the Internet to mine for hidden patterns.

Digital Footprint
Items purchased
Transaction amount

Status: May qualify for a loan.

More Digital Footprint
Items purchased
Transaction amount
Spending Pattern
Sentiment Analytics

Status: May be blacklisted for a loan.

Note: 3 independent pieces of information about a certain customer lead to the customer being classified as a long-term credit risk, whereas the individual pieces of information would never have led to this conclusion. It is exactly this kind of discovery of hidden patterns that forms an additional threat to citizens’ privacy.

How to Guard Privacy?

- The privacy of an individual is breached when an attacker can learn anything extra about a record owner, possibly with the presence of any background knowledge from other sources.
- Quasi-identifier: Pieces of information that are not of themselves unique identifiers but can be combined with other quasi-identifiers to create a unique identifier. (Capable of identifying a personal information)

Quasi-Identifier

Table A

Name	Age	Sex	Zipcode	Disease
Bob	23	M	11000	Pneumonia
Ken	27	M	13000	Flu
Jane	24	F	15000	Gastritis
Linda	25	F	14500	Bronchitis
Sam	41	M	13100	Flu

Table B

Name	Age	Sex	Zipcode	Occupation
Bob	23	M	11000	Engineer
Ken	27	M	13000	Accountant
Jane	24	F	15000	Student
Linda	25	F	14500	Consultant
Sam	41	M	13100	Programmer

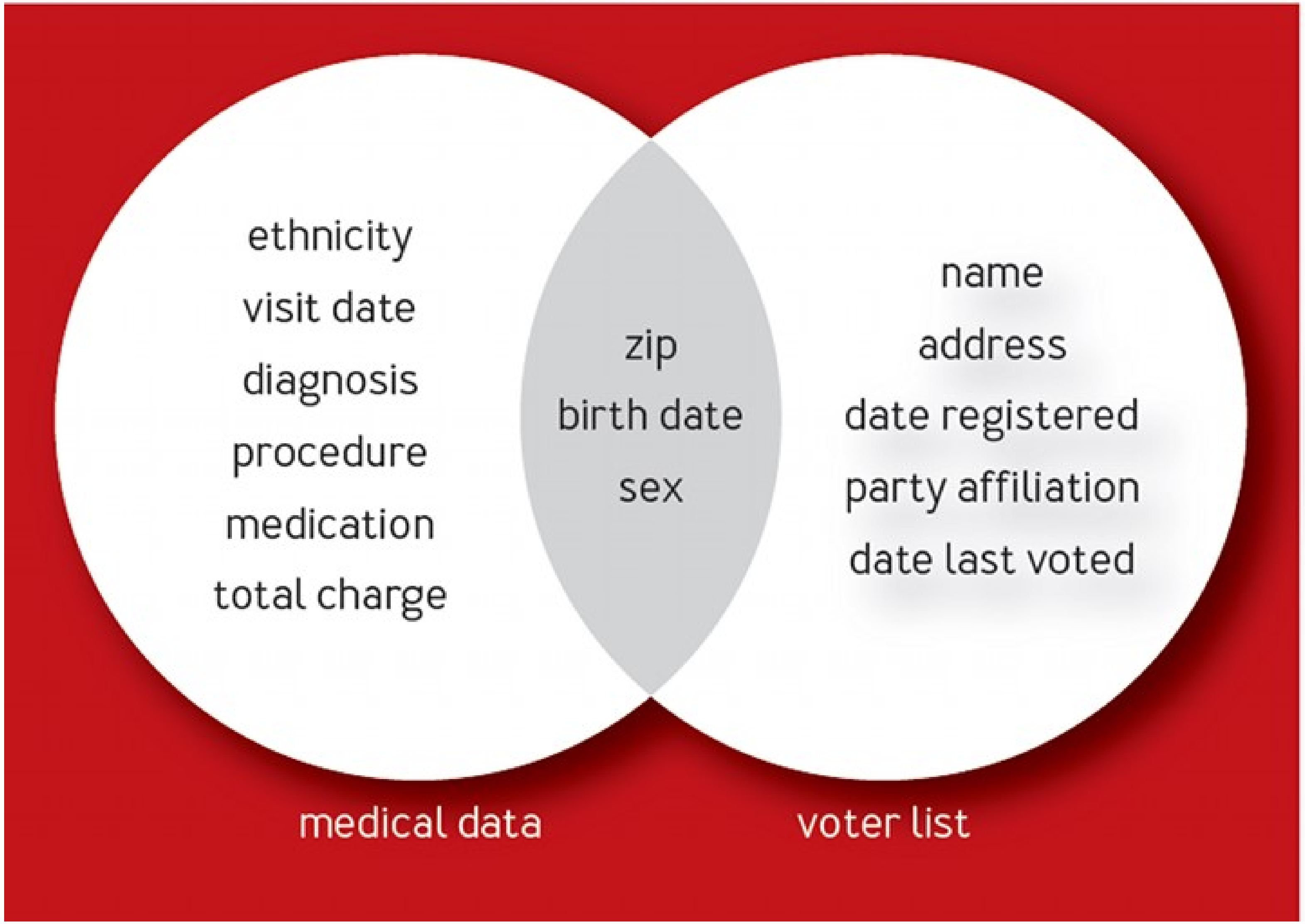
Quasi-Identifier

Age	Sex	Zipcode	Disease
23	M	11000	Pneumonia
27	M	13000	Flu
24	F	15000	Gastritis
25	F	14500	Bronchitis
41	M	13100	Flu

Published Data

Another Example

FIGURE 1: COMBINATION OF TWO DATA SETS THAT ALLOW RE-IDENTIFICATION



How to Guard Privacy?

- Several methods to anonymize data:
  1. Generalization and Suppression (*k*-anonymization) - Remove information from the quasi identifiers, until the records are not individually identifiable
  2. Anatomization and permutation - Groups and shuffles sensitive values within a QID group, in order to remove the relationship between the QID and sensitive attributes (Change the data by adding noise, swapping values, creating synthetic data)

k-anonymization

Name	Age	Sex	Zipcode	Disease
Bob	23	M	11000	Pneumonia
Ken	27	M	13000	Flu
Jane	24	F	15000	Gastritis
Linda	25	F	14500	Bronchitis
Sam	41	M	13100	Flu

where *k* = degree of anonymity

- **Suppression**: In this method, certain values of the attributes are replaced by an asterisk '*'.
- **Generalization**: In this method, values of attributes are replaced with a broader category. For example, the value '23' may be replaced by '20 < Age ≤ 30‘
- Suppose, *k*=2 and we apply suppression onto **name** and generalization onto **age**.

k-anonymization

Name	Age	Sex	Zipcode	Disease
*	$20 < \text{Age} \leq 30$	M	11000	Pneumonia
*	$20 < \text{Age} \leq 30$	M	13000	Flu
*	$20 < \text{Age} \leq 30$	F	15000	Gastritis
*	$20 < \text{Age} \leq 30$	F	14500	Bronchitis
*	$40 < \text{Age} \leq 50$	M	13100	Flu

where *k* = degree of anonymity

- **Suppression**: In this method, certain values of the attributes are replaced by an asterisk '*'.

- **Generalization**: In this method, values of attributes are replaced with a broader category. For example, the value '23' may be replaced by '$20 < \text{Age} \leq 30$'