Data Quality

What is data quality?

Data quality is a perception or an assessment of data's fitness to serve its purpose in a given context.
It is described by several dimensions like

•Correctness / Accuracy : Accuracy of data is the degree to which the captured data correctly describes the real world entity.

•Consistency: This is about the single version of truth. Consistency means data throughout the enterprise should be sync with each other.

Contd…

•Completeness: It is the extent to which the expected attributes of data are provided.

•Timeliness: Right data to the right person at the right time is important for business.

•Metadata(元数据): Data about data.

Meaning of Data Quality

- Generally, you have a problem if the data doesn't mean what you think it does, or should
  - Data not up to spec : garbage in, glitches, etc.
  - You don't understand the spec : complexity, lack of metadata.
- Many sources and manifestations
- Data quality problems are expensive and pervasive
  - Resolving data quality problems is often the biggest effort in a data mining study.

Example

T.Das|97336o8327|24.95|Y|-|0.0|1000
Ted J.|973-360-8779|2000|N|M|NY|1000

- Can we interpret the data?
  - What do the fields mean?
  - What is the key? The measures?
- Data glitches (数据故障)
  - Typos, multiple formats, missing / default values
- Metadata and domain expertise
  - Field three is Revenue. In dollars or cents?
  - Field seven is Usage. Is it *censored*?
    - Field 4 is a censored flag. How to handle censored data?

What are:

- Valid data?
- Reliable data?
- Complete data?
- Precise data?
- Timely data?
- Data with integrity?

? = Dimensions of Data Quality

Validity	Valid data are considered *accurate*: They measure what they are intended to measure.
Reliability	The data are measured and collected consistently; definitions and methodologies are the same over time.
Completeness	Completely inclusive: the DMS represents the *complete* list of eligible names and not a fraction of the list.
Precision	The data have sufficient detail; in this case the "accuracy" of the data refers to the *fineness* of measurement units.
Timeliness	Data are up-to-date (current), and information is available on time; the DMS produces reports under deadline.
Integrity	The data are protected from deliberate bias or manipulation for political or personal reasons.

Conventional Definition of Data Quality

- Accuracy
  - The data was recorded correctly.
- Completeness
  - All relevant data was recorded.
- Uniqueness
  - Entities are recorded once.
- Timeliness
  - The data is kept up to date.
- Consistency
  - The data agrees with itself.

Problems …

- Unmeasurable
  - Accuracy and completeness are extremely difficult, perhaps impossible to measure.
- Context independent
  - No accounting for what is important. E.g., if you are computing aggregates, you can tolerate a lot of inaccuracy.
- Incomplete
  - What about interpretability, accessibility, metadata, analysis, etc.
- Vague
  - The conventional definitions provide no guidance towards practical improvements of the data.

Data Quality Constraints

- Many data quality problems can be captured by *static* constraints based on the schema.
  - Nulls not allowed, field domains, foreign key constraints, etc.
- Many others are due to problems in workflow, and can be captured by *dynamic* constraints
  - E.g., orders above $200 are processed by Biller 2
- The constraints follow an 80-20 rule
  - A few constraints capture most cases, thousands of constraints to capture the last few cases.
- Constraints are measurable. Data Quality Metrics?

Data Quality Metrics

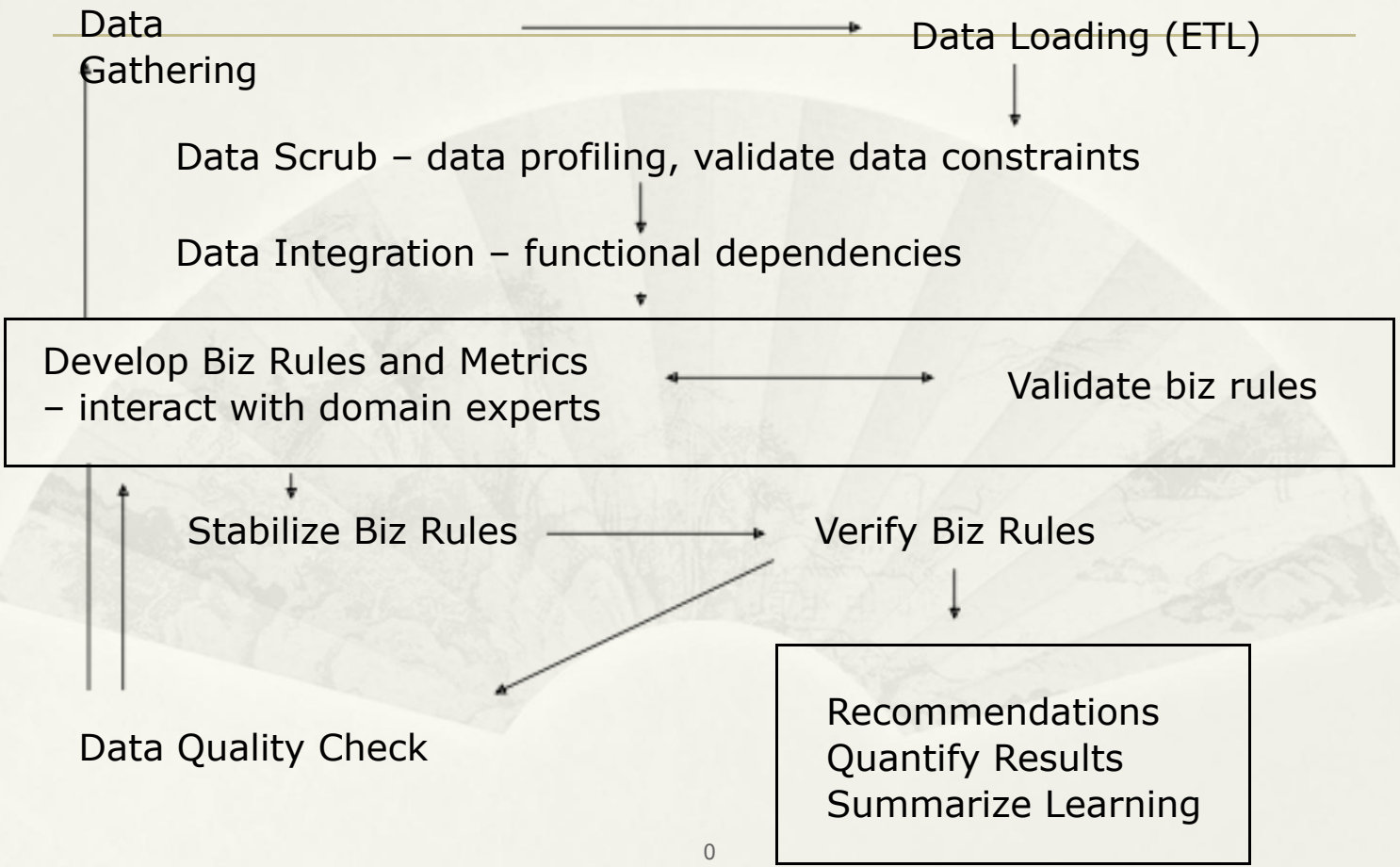
Data Quality Metrics

- We want a measurable quantity
  - Indicates what is wrong and how to improve
  - Realize that DQ is a messy problem, no set of numbers will be perfect
- Types of metrics
  - Static vs. dynamic constraints
  - Operational vs. diagnostic
- Metrics should be *directionally correct* with an improvement in use of the data.
- A very large number metrics are possible
  - Choose the most important ones.

Examples of Data Quality Metrics

- Conformance to schema
  - Evaluate constraints on a snapshot.
- Conformance to business rules
  - Evaluate constraints on changes in the database.
- Accuracy
  - Perform inventory (expensive), or use proxy (track complaints). Audit samples?
- Interpretability
- Glitches in analysis

Data Quality Process



Missing Data

- Missing data - values, attributes, entire records, entire sections
- Missing values and defaults are indistinguishable
- Truncation/censoring - not aware, mechanisms not known
- Problem: Misleading results, bias.

Challenges in Data Quality

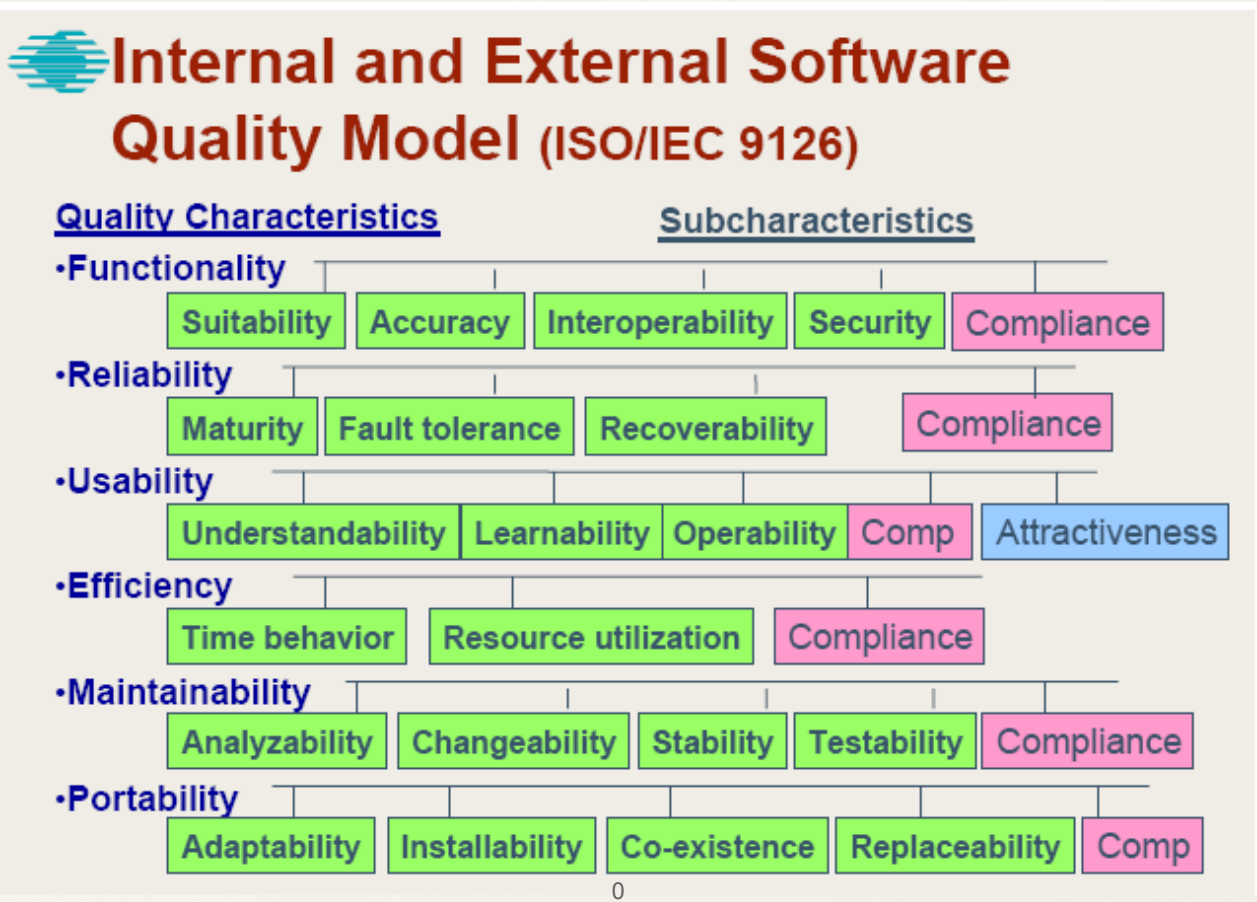
- Multifaceted nature
  - Problems are introduced at all stages of the process.
    - but especially at organization boundaries.
  - Many types of data and applications.
- Highly complex and context-dependent
  - The processes and entities are complex.
  - Many problems in many forms.
- No silver bullet
  - Need an array of tools.
  - And the discipline to use them.

Current status

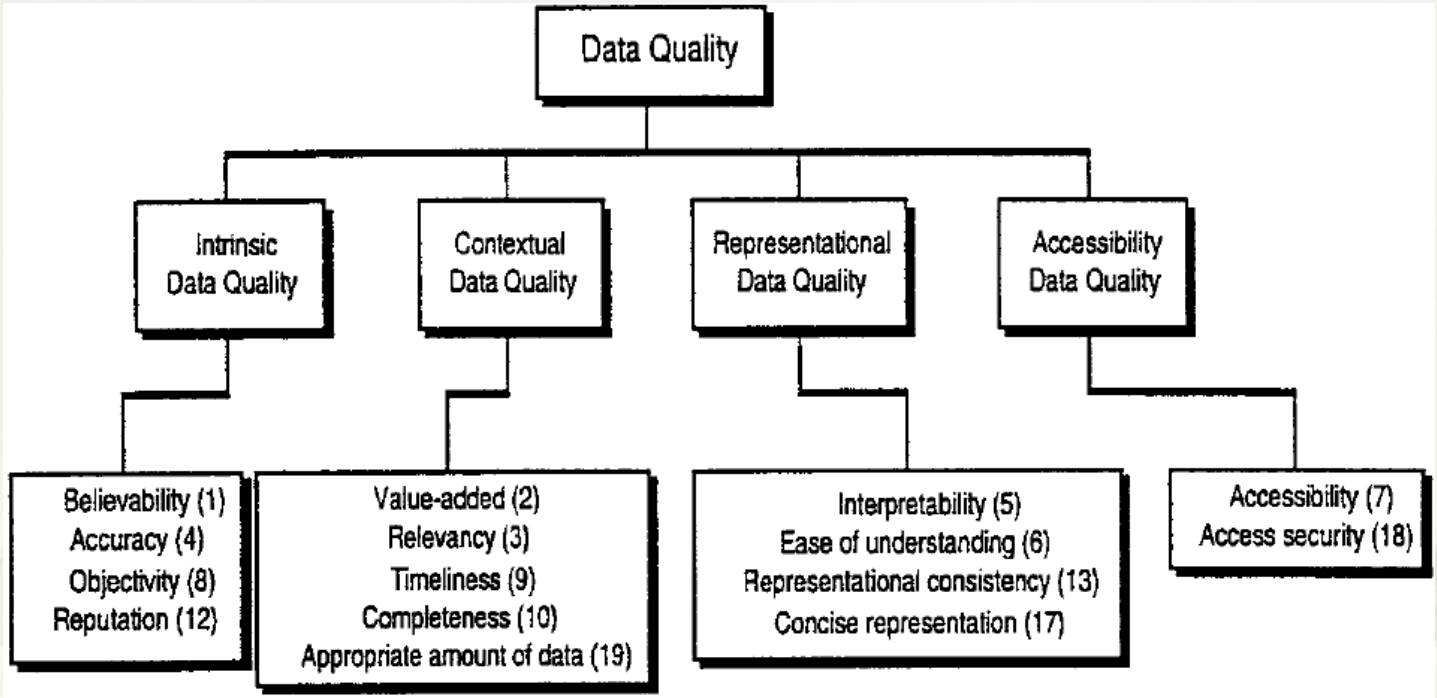
Lack of a unified definition of data quality

- 1. Data Quality refers to the degree of excellence exhibited by the data in relation to the portrayal of the actual phenomena.
- 2. The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.
- 3. The totality of features and characteristics of data that bears on their ability to satisfy a given purpose; the sum of the degrees of excellence for factors related to data.
- 4. Information Quality : the fitness for use of information; information that meets the requirements of its authors, users, and administrators.
- 5. Data quality: The processes and technologies involved in ensuring the conformance of data values to business requirements and acceptance criteria

Current status



Current status



A conceptual framework of data quality, (RICHARD Y. WANG AND DIANE M. STRONG .Beyond Accuracy: What Data Quality Means to Data Consumers)

Problems

Duplicated , inconsistent , ambiguous, incomplete.

So there is a need to collect data in one place and clean up the data.

Maintenance of data quality

Data quality results from the process of going through the data and scrubbing it, standardizing it, and de duplicating records, as well as doing some of the data enrichment.

1. Maintain complete data.
2. Clean up your data by standardizing it using rules.
3. Use fancy algorithms to detect duplicates. (Eg: ICS and Informatics Computer System, are the same)
4. Merge existing duplicate records.

…….

Context

In process of data warehouse design, many database professionals face situations like:

1.Several data inconsistencies in source, like missing records or NULL values.

2.Or, column they chose to be the primary key column is not unique throughout the table.

3.Or, schema design is not coherent to the end user requirement.

4.Or, any other concern with the data, that must have been fixed right at the beginning.