Data Warehousing
Data Quality

Spring Term 2020
Dr. Andreas Geppert
geppert@acm.org

Outline of the Course

- Introduction
- DWH Architecture
- DWH-Design and multi-dimensional data models
- Extract, Transform, Load (ETL)
- Metadata
- **Data Quality**
- Analytic Applications and Business Intelligence
- Implementation and Performance
- Security and Privacy

Outline

1. **Data Quality**
2. Data Quality Management
3. Data Profiling
4. Data Cleansing
5. Data Quality in the ETL Process

Data Quality

- Data have quality, when they meet the requirements of their intended use (Olson)
- In order to meet these requirements, data need to be accurate, up-to-date, relevant, complete, understood, and trustworthy (Olson)
- The level of data quality thus depends on the intended use as well as on the data themselves
- The same data can have different levels of quality in different usage contexts

Real-world data quality problems
(not necessarily DWH-specific)

- Estimates guess that in some companies, the cost incurred by lack of data quality amounts to 15-20% of the operating result (Olson)
- According to a TDWI study in 2001, the cost of insufficient data quality in US-companies amounts to $600 bn USD (http://www.tdwi.org/display.aspx?id=6045)
- Estimates by Gartner analysts predict that up to 50% of DWH and CRM projects will not meet their requirements or fail entirely because of data quality problems
- Inmon assumes that without data quality assurance, up to 20% of the data in data warehouses are wrong, redundant or otherwise useless

Real-world data quality problems (2)

- USPS estimates that in 2001 they paid 1.8 bn USD because of undeliverable mail
- "A major ... telecommunication company ... had more than 33 different definitions for «customer churn»" (Howson)
- Eziba has sent tens of thousands of catalogues to addresses that before had been classified as "unlikely respondents". The company had to suspend operations temporarily (NYT 2005-01-24)
- It has been found out in various studies that
  - Consumer lose $1-2.5 bn because of scanner errors in supermarket cashiers
  - Scanner determine a wrong price for 6-8% of the products in certain stores (NYT 2006-01-28)

Real-world data quality problems (3)

- An employee of a Japanese brokerage unit entered a sell order over 610,000 shares at 1 Yen, instead of one share at 610,000 Yen. The resulting loss amounted to 27 bn Yen ($224 mn)
- After the presidential election in 2004, it has been found out that in New Jersey ...
  - 4,755 persons had voted that had been declared dead before;
  - 4,397 persons that had been registered in more than one county, had voted twice;
  - 6,572 persons had voted in New Jersey and one of five other states. (NYT 2005-09-16)

Real-world data quality problems (4)

- »... Alan Greenspan, former chairman of the Federal Reserve, today ... explained to a U.S. House committee what he thought went wrong. And **insufficient data** was one of the causes he pointed to.
  Greenspan has long praised computer technology as a tool that can be used to limit risks in financial markets. For instance, in 2005, he credited improved computing power and risk-scoring models with making it possible for lenders to extend credit to subprime mortgage borrowers.
  But at a hearing held today by the House Committee on Oversight and Government Reform, Greenspan acknowledged that the **data fed into financial systems was often a case of garbage in, garbage out**. Business decisions by financial services firms were based on "the best insights of mathematicians and finance experts, supported by major advances in computer and communications technology," Greenspan told the committee. "The whole intellectual edifice, however, collapsed in the summer of last year because the **data inputted into the risk management models generally covered only the past two decades** — a period of euphoria."
  He added that if the risk models had also been built to include "historic periods of stress, capital requirements would have been much higher and the financial world would be in far better shape today, in my judgment."
  And that wasn't the only bad data being crunched by IT systems. Christopher Cox, chairman of the Securities and Exchange Commission, told the committee that credit rating agencies gave AAA ratings to mortgage-backed securities that didn't deserve them. "These ratings not only gave false comfort to investors, but also **skewed the computer risk models** and regulatory capital computations," Cox said in written testimony.«
- Computerworld, October 23, 2008
  (http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9117961)

Quality Criteria

Accuracy

- DWH data correspond to source data (and real world)
- Only wanted, intended deviations
- Deviations are documented
- violations ...
  - ... because of input errors
    - typos: „Joe Cool“→ „Joe Coool“
    - misundestandings: „Joe Cool“→ „Joe Kuhl“
  - ... because of wrong input or insufficient maintenance in the source (because data are not used there)
  - ... because of deliberate wrong input
    - [„Joe Cool“, „non-smoking“]

Quality Criteria

Semantics

- The meaning of data is known and documented
- Unique and uniform data semantics exists; differing semantics are documented

Complete-ness

- DWH contains complete data set required for analysis
- Individual data elements / records are complete
- Typical reason for violations: data incomplete in source
  - Not relevant for OLTP, thus not stored in source
  - [name: „Joe Cool“, ssn: „999-99-9999“]

Quality Criteria

Consistency

- Data do not contain contradictions
- Violations through uncontrolled, redundant data storage and un-managed updates
- Example: birthday = 1992-12-19, age = 20

Uniqueness

- The same data elements exist only once in the DWH
- Violation through redundant input in one or multiple sources
- „Joe Cool“, „Joe S. Cool“, „Snoopy“

Quality Criteria

Timeliness

- Data are updated «in time»
- Depending on user requirements
- Violations because of ...
  - ... In adequate (business) processes
  - ... Missing capture
  - ... technical issues
    - Missing data delivery from sources
    - Processing issues in the ETL process

Data Quality: Exercise

- identify manifestations of real-world data quality issues
- identify root causes

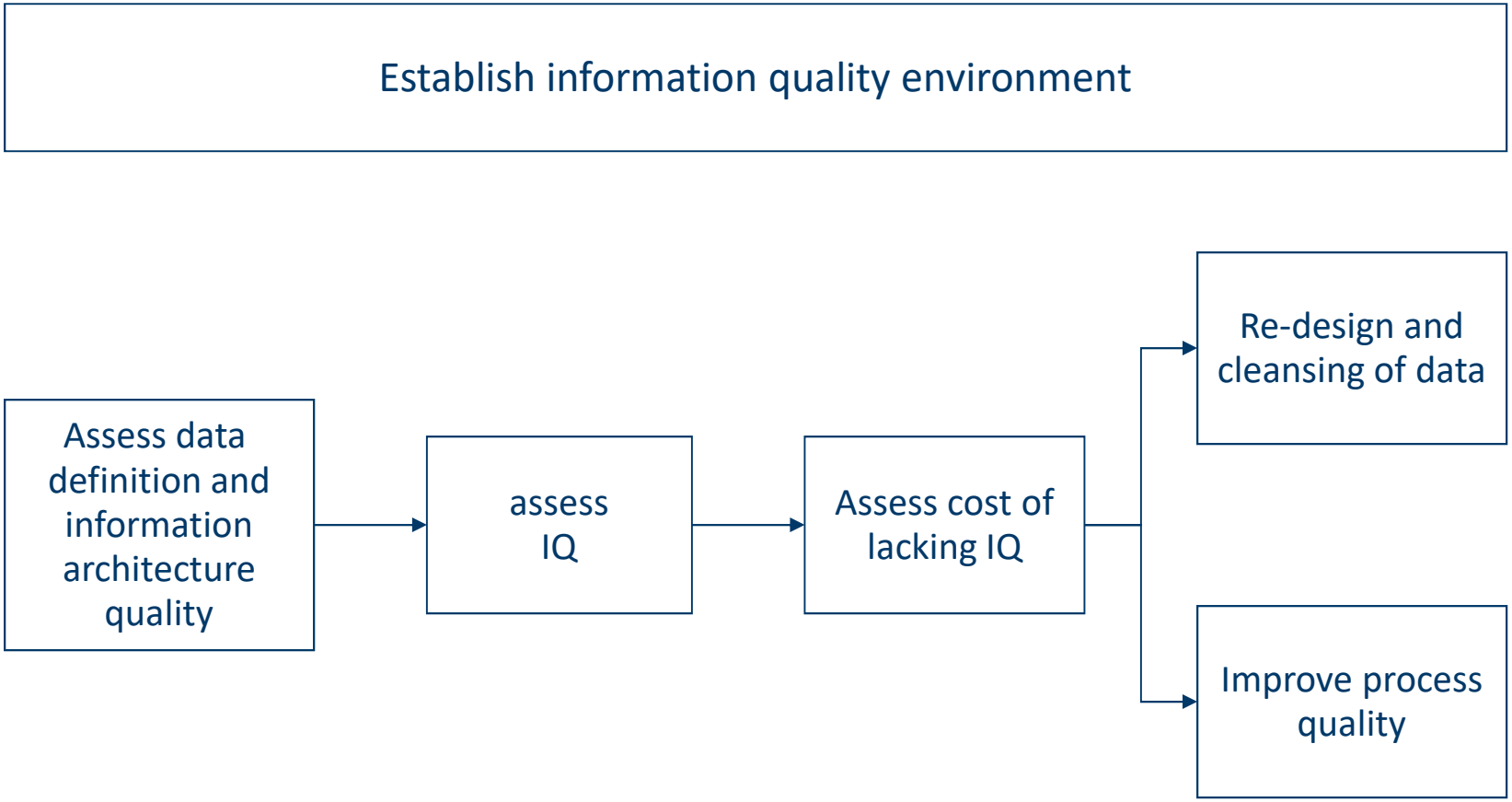
Outline

1. Data Quality
2. **Data Quality Management**
3. Data Profiling
4. Data Cleansing
5. Data Quality in the ETL Process

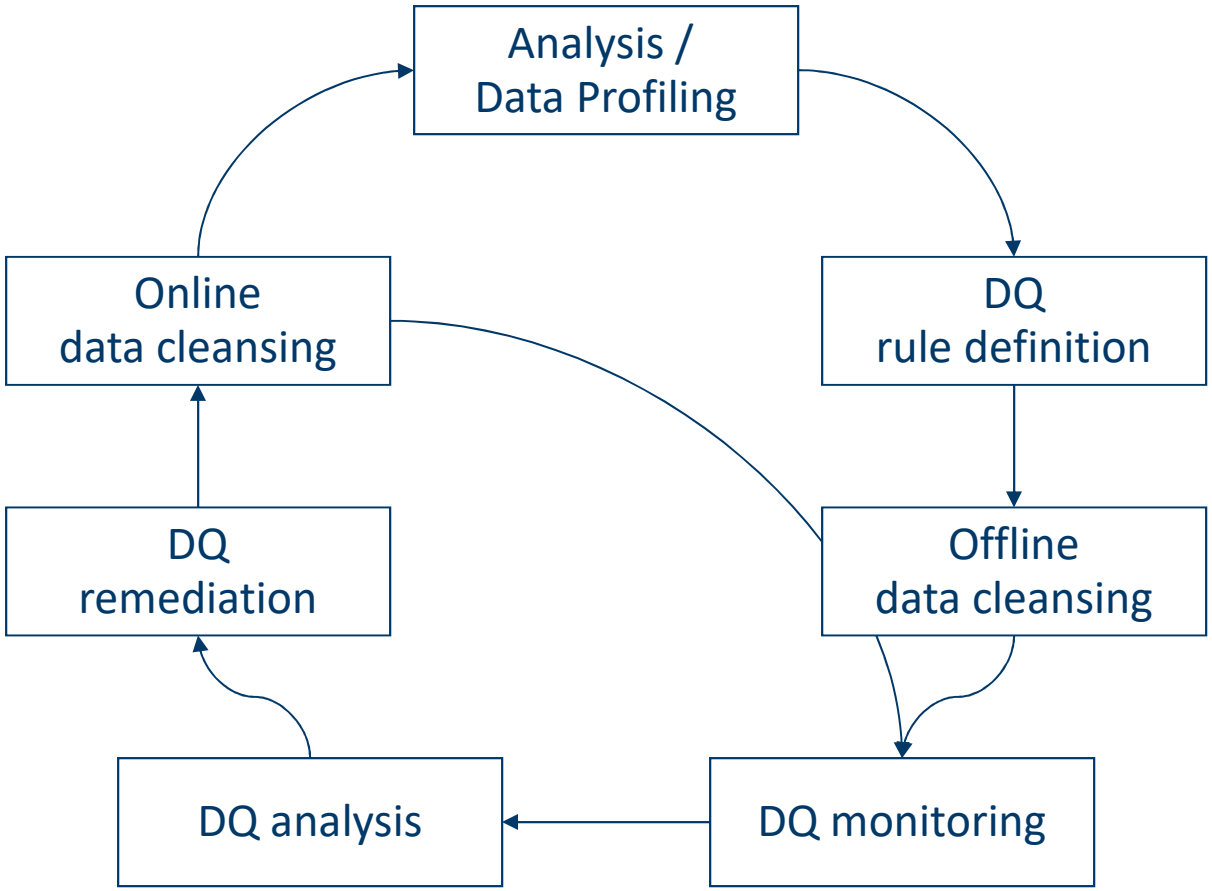
Information Quality Management

- Data quality improvement is not a one-time activity, but an ongoing process
  - Data Quality Assurance Program (Olson)
  - Total Quality Data Management
- Corresponding concepts, processes, tools, and organizational setup
- One might even have to align the entire organization, its culture, and its management, to data quality
- Buy-in of management and involvement of business (semantics !)
- Quality improvement should be integral part of all new projects, which create new data sets or migrate, replicate, or integrate data
  - improve, prevent, monitor

Information Quality Management



Data Quality Assurance Cycle



Outline

1. Data Quality
2. Data Quality Management
3. **Data Profiling**
4. Data Cleansing
5. Data Quality in the ETL Process

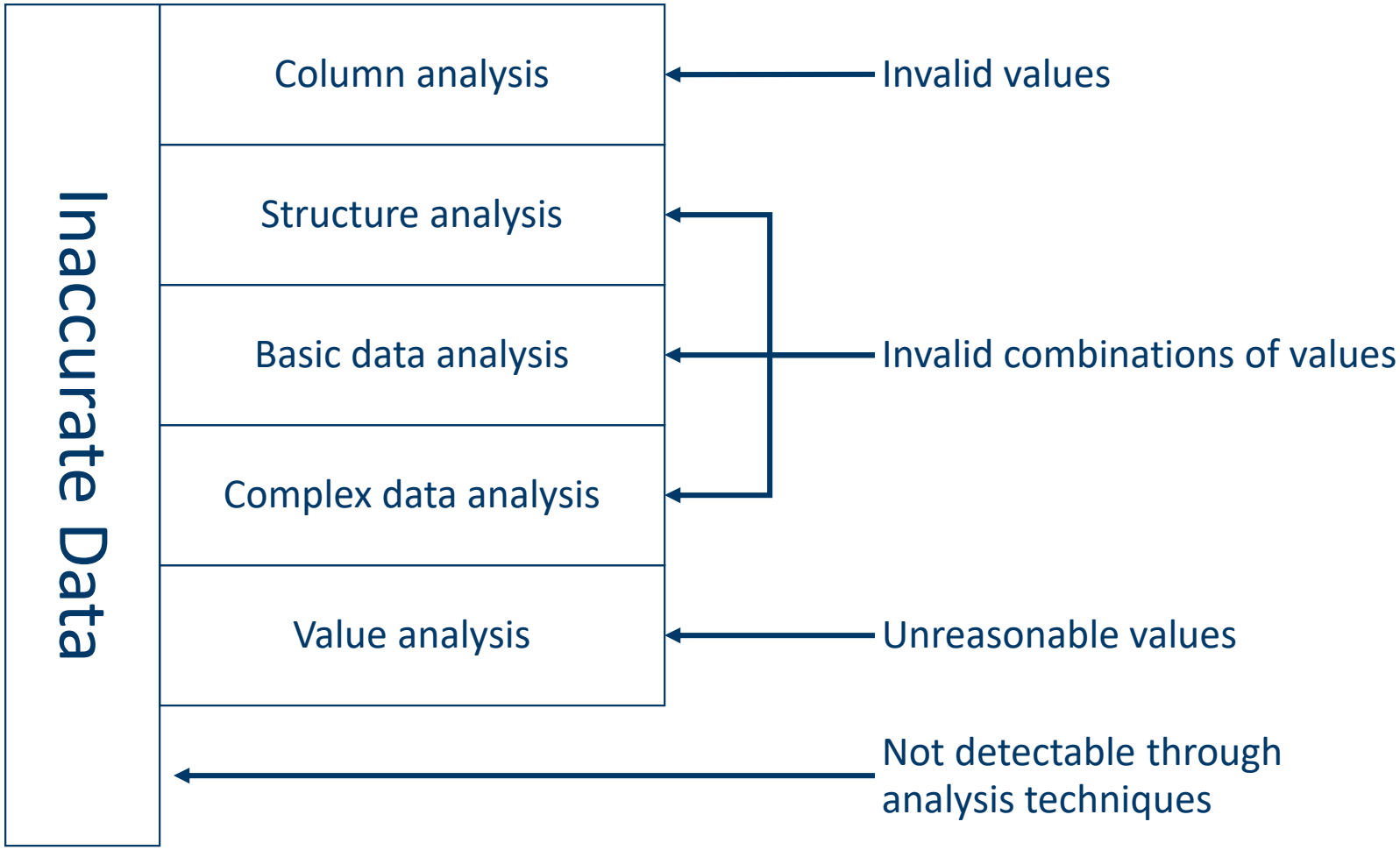
Data Profiling

- Usage of analytic techniques to discover valid structures and content and to assess quality of data sets
- Takes existing metadata and instance data as input
- Outputs accurate, more complete metadata and further information about meaning and quality of data

Data Profiling: Input

- Metadata, depending on the type of data source (database, file, …)
- Data definitions:
  - Table definitions, column definitions, incl. data types (← DB schema)
  - Type definitions (← program source code)
  - Type definitions (← Cobol copybooks, PL1-includes)
- Structure information
  - Primary keys, foreign keys, constraints (← DB-Schema)
  - Semantic/implicit relationships (← trigger and stored procedures source code)
  - Application constraints when inserting, modifying, deleting data (← application source code)

Data Profiling: Outline



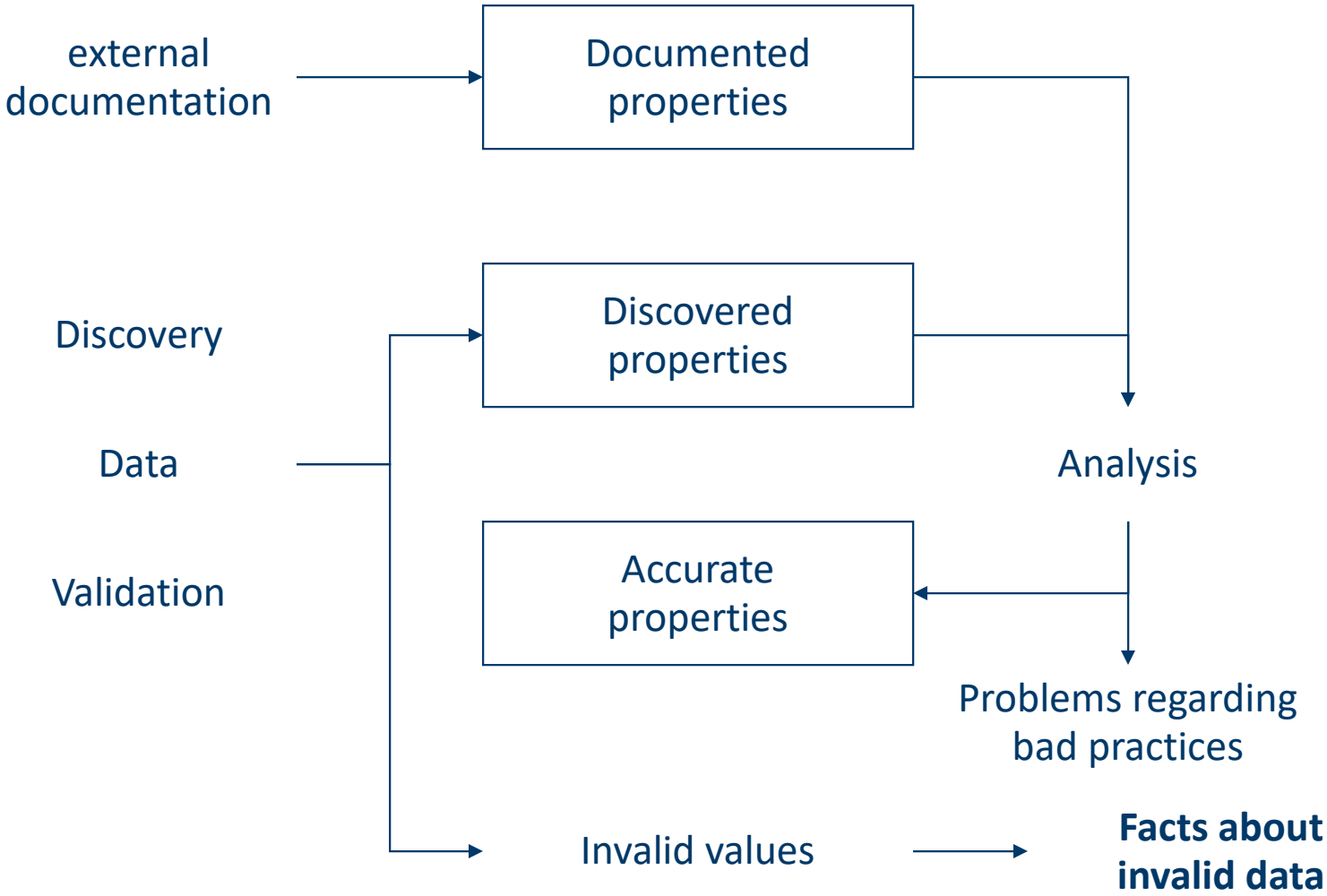
Analysis Techniques

- Discovery
  - Discovers previously unknown facts
  - see Data Mining
- Rule and Assertion Checking
  - Assertions are input for analysis
  - Analysis checks validity and identifies data elements that violate assertions
- Visual inspection
  - Usually against pre-processed data sets
  - Frequency distributions histograms), groupings, compare values/ aggregated across data sources, threshold values
  - Example: "Mickey Mouse" is the most frequent user name

Data Profiling: Column Analysis

- Investigates values in individual columns, independent from values in other columns
- Metadata specify properties of admissible values
  - Name
  - (Business) Meaning
  - Domain name
  - Data type
  - Character Set
  - Length restrictions (min, max)
  - Acceptable values (discrete list; value set; character/digit patterns)
  - Null constraints
  - Uniqueness constraints
  - Consecutivity constraints

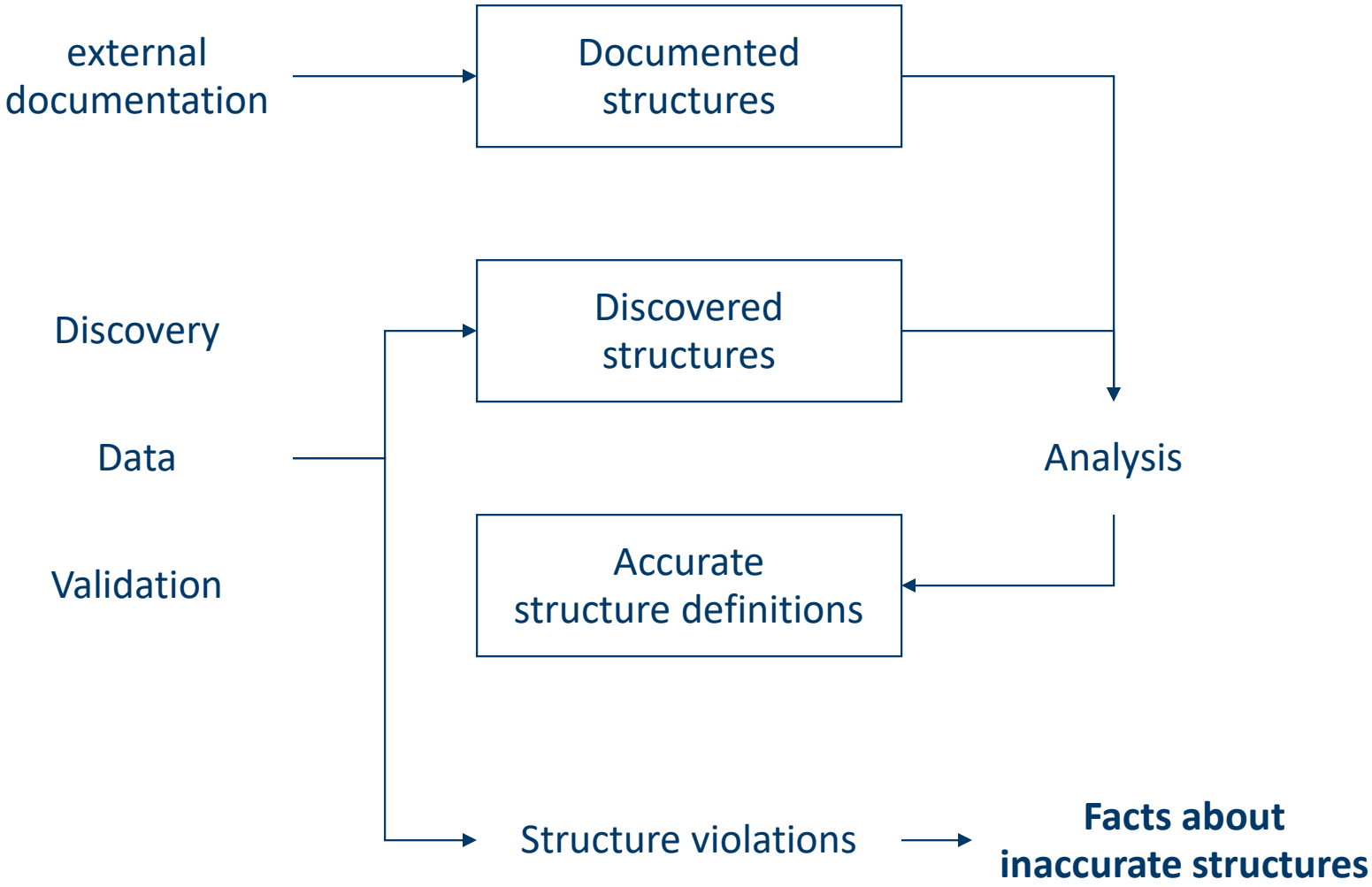
Data Profiling: Column Analysis



Data Profiling: Structure Analysis

- Looks for table structures (between columns) and object structures (between tables)
- Uses these relationships to assess data quality
- Determines functional dependencies as basis for subsequent steps
- Identifies primary and foreign key constraints, redundant and derived columns, synonyms
- Relies on understanding of semantics to decide whether data are inaccurate or rules invalid in case of rule violations
- Problem when using samples: false positives

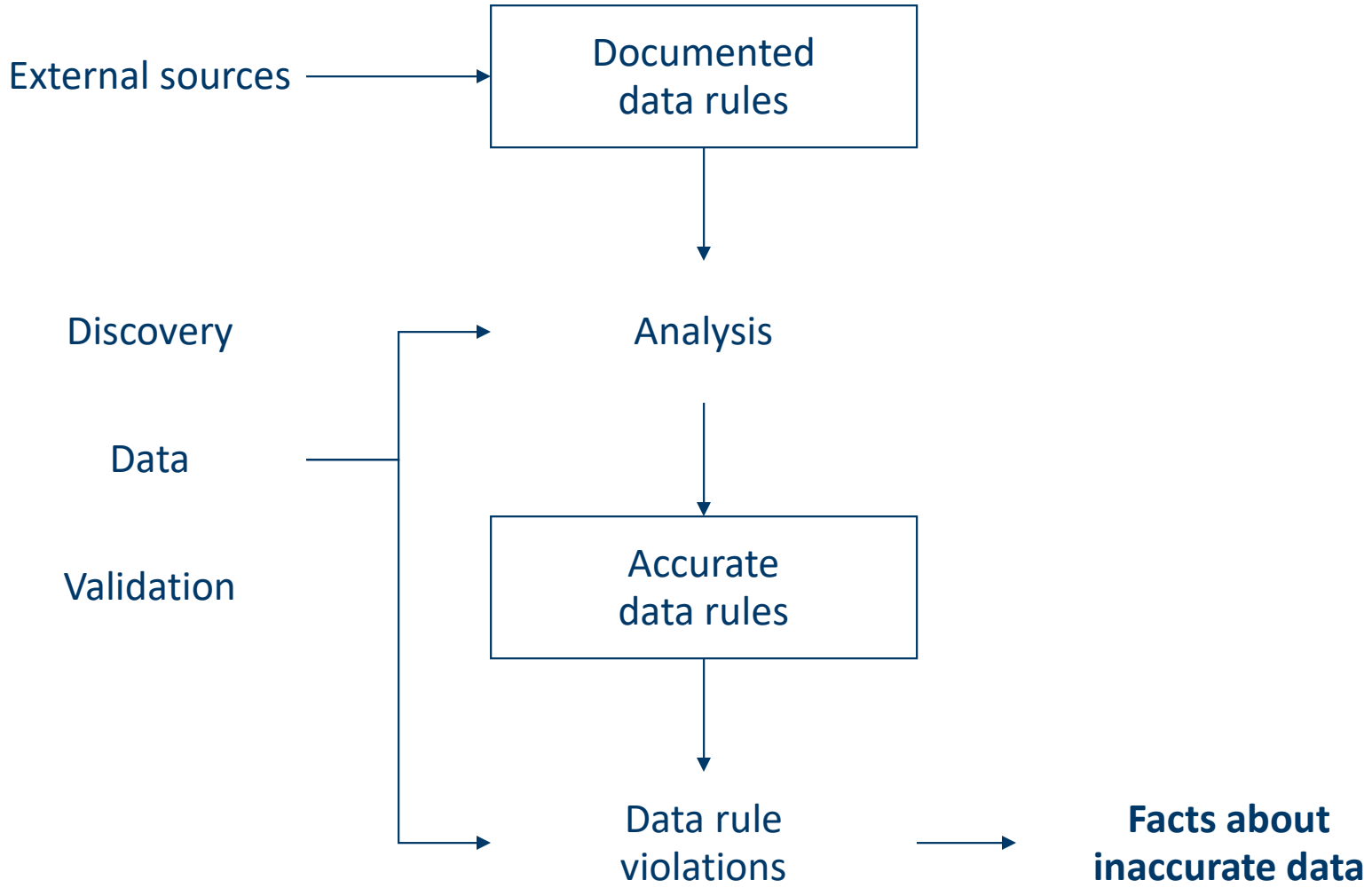
Data Profiling: Structure Analysis



Data Profiling: Basic Data Analysis

- Basic rule: restriction of the values of multiple attributes of a business object
- Soft vs. hard rules
  - Example: hire_date > birth_date vs. hire_date > birth_date + 18 years
- Formulation of rules and validation of data against rules
  - Often, an individual data value cannot be identified, only in combination with other values
  - Example: martial status = “single", maiden name = "Meier"
- Rules typically for:
  - Date and time attributes, time durations
  - Classifying attributes
  - Process status attributes ("workflow")
  - Derived attributes

Data Profiling: Basic Data Analysis



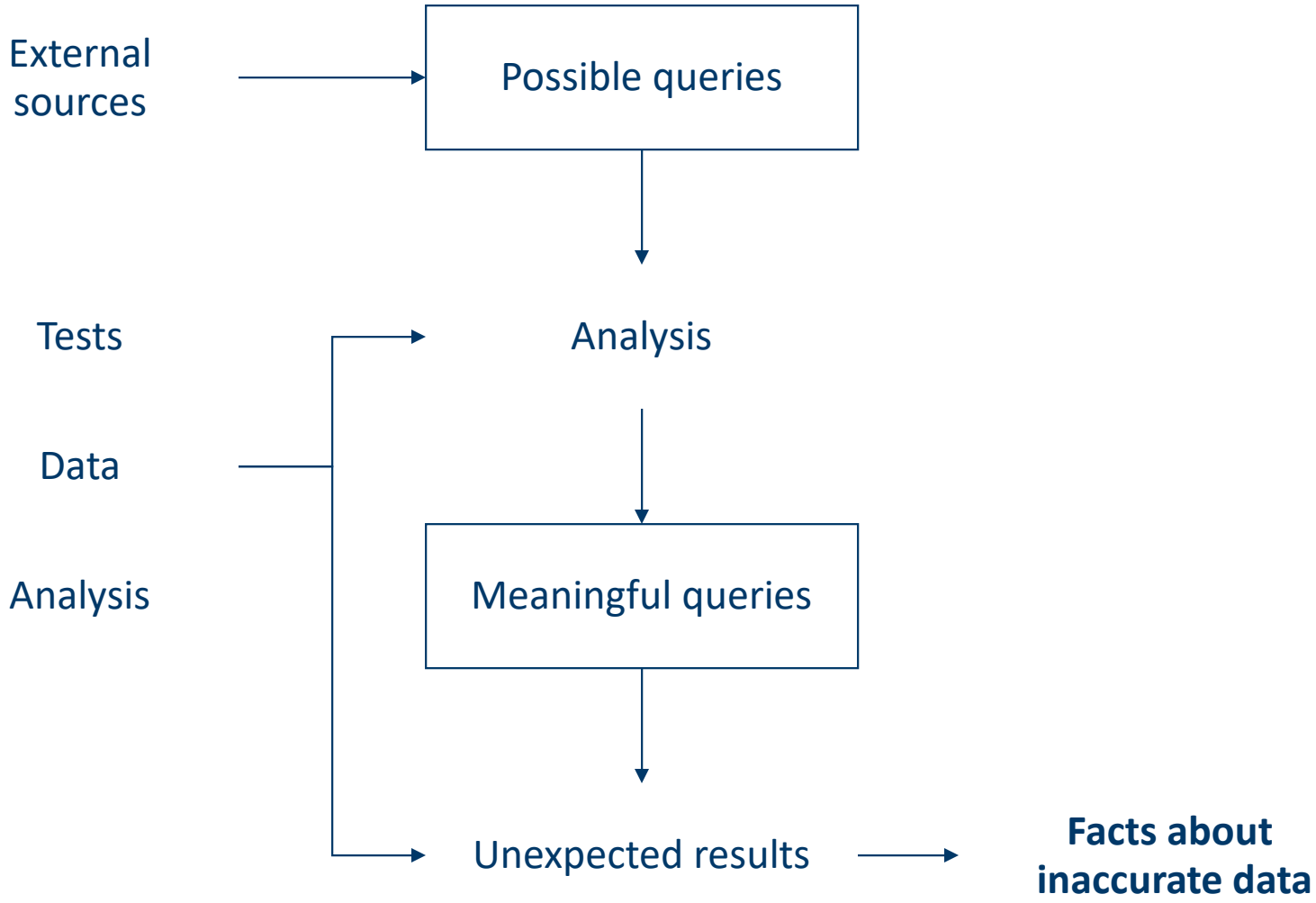
Data Profiling: Complex Data Analysis

- Complex rule: restriction of values of multiple business objects or the relationships between multiple business objects
- Otherwise very similar to basic data analysis
- Examples:
  - A book can be lent to at most one user per point in time (time and space properties)
  - Aggregations (completeness)

Data Profiling: Value Analysis

- Inaccurate data that are not detectable through clear-cut ("black/white") rules
- All individual attribute values can be valid or possible
- Value rule: calculation of one or more values over a data set
  - cardinalities
  - Frequency distribution
  - Number of distinct values
  - Extreme values
  - Percentage of null values
- Visual inspection to identify obvious outliers

Data Profiling: Value Analysis



Outline

1. Data Quality
2. Data Quality Management
3. Data Profiling
4. **Data Cleansing**
5. Data Quality in the ETL Process

Data Cleansing

- Problem detection
  - Identification of bad data (dirty data)
- Resolution of problems, data correction
- Location of data cleansing
  - Dirty data should be cleansed in sources (why ?)
  - Data modification in the DWH only in exceptional cases (which ?)
- Automated vs. manual data cleansing
  - Automation desirable especially for large data sets
  - In many cases, manual cleansing / human supervision will be necessary
- offline vs. online

Data Cleansing: Duplicate Detection / Matching

- duplicates: violate uniqueness criteria
  - e.g., multiple representation of the same person (customers, clients)
  - find duplicates and resolve (e.g., merge into a single representation)
  - but only trivial cases can be caught with database uniqueness constraints
- related: matching
  - determine whether an object (e.g. a person) is contained in a reference set
  - e.g., politically exposed persons (PEPs)
- approach:
  - determine candidates for closer inspection
    - comparing everything with everything else typically not feasible
  - calculate similarity / distance
    - treat very similar objects as duplicates, or inspect manually

Duplicate Detection: Window Construction

- Move «window» over set of objects
  - sort objects according to some key
  - key construction reflects possible data quality problems (e.g., typical mistakes)
- Compare objects «visible» through the window
  - Calculate similarity / distance
- Decide based on similarity
  - manually/visually
  - automatically
- Run multiple iterations with different keys to improve reliability

Duplicate Detection: Key Construction Example

- Key:
  - the first three consonants of the last name
  - first three letters of the first name
  - house number
  - all consonants of the street address
  - first three digits of the SSN

First	Last	Address	SSN	Key
Bill	Jansen	123 First Street	45678987	JNSBIL123FRST456
Bill	Jansen	123 First Street	45678987	JNSBIL123FRST456
Bill	Janssen	123 First Street	45678987	JNSBIL123FRST456
Bill	Jensen	123 Forest Street	45654321	JNSBIL123FRST456

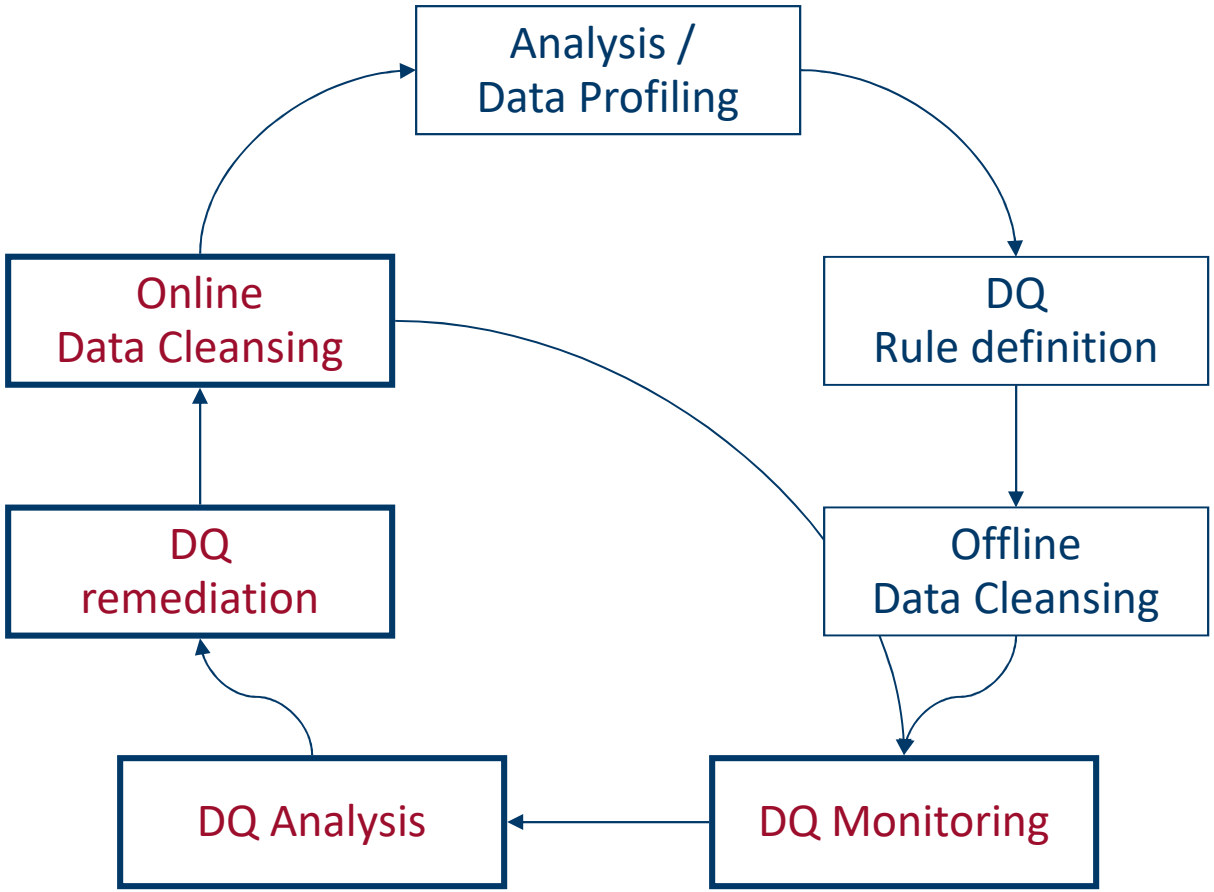
Duplicate Detection: Similarity / Distance

- Different distance functions possible
- Editing distance
  - How many changes do we need to change t1 into t2?
    - d(„Joe Cool“, „Joe Coool“) < d(„Joe Cool“, „Joe Kuhl“)
- Keyboard distance
  - Characterwise comparison of words
  - Neighboring characters on the keyboard have a smaller distance
    - d(„Joe Cool“, „Joe Fool“) < d(„Joe Cool“, „Joe Pool“)
- Phonetic distance
  - based on sond, pronounciation
  - see soundex function
    - d(„Joe Cool“, „Joe Kuhl“) < d(„Joe Cool“, „Joe Pool“)

Outline

1. Data Quality
2. Data Quality Management
3. Data Profiling
4. Data Cleansing
5. **Data Quality in the ETL Process**

Data Quality in the ETL Process



DQ Monitoring

- Data quality must be determined for newly loaded data
- (selected) rules are checked at runtime as part of the ETL process
  - Checking all rules for the entire data set will not be feasible in many cases
- Possible reactions in case of DQ rule violation:
  - Log DQ issue (Continue with ETL processing)
  - Notify data owner
  - Drop dirty data
  - Cleanse dirty data
  - Stop ETL process

DQ Analysis

- Data quality issues detected as part of monitoring must be analyzed further
- "DQ Dashboard" provides overview about data quality and data quality issues
  - Which rules have been checked for which data sets?
  - User (group) defined quality thresholds
  - Traffic light system provides information about usability of data
  - Drill-down support allows user to analyze data quality issues in more detail

DQ Remediation

- Information about DQ issues are sent to the organization responsible for rectifying quality issues (DQ Issue Management)
  - Part of Data Governance
- Detailed analysis of DQ issues
- Correction of dirty data
- Process improvements, application changes in order to improve quality in a sustainable way

Summary

- DQ issues in sources are a frequent problem in many data warehouses
- DQ issues are aggravated by integration
- DQ issues can compromise the usefullness of analysis results
- Data quality must be approached as a comprehensive and continuous process
- Quality rules are a form of metadata and must often be determined and validated
- Data profiling helps to complete and improve metadata and assess initial status of data quality
- DQ monitoring checks data quality at runtime and provides a range of possible reactions to DQ issues