

---

# Data/Information Quality:

## Problems, Challenges, and Techniques

(based on a tutorial given at EDBT'06 -  
joint work with Felix Naumann)

---

**Kai-Uwe Sattler**

TU Ilmenau

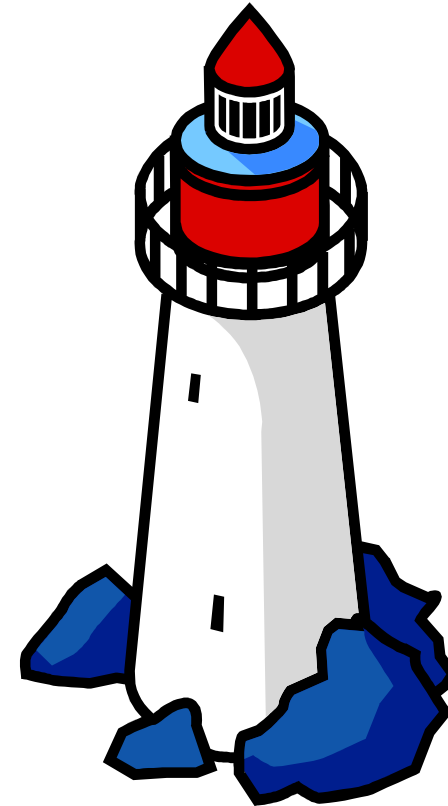
FG Datenbanken & Informationssysteme

[www.tu-ilmenau.de/dbis](http://www.tu-ilmenau.de/dbis)

---

# Overview

- ➔ ■ Motivation
- Defining IQ
  - IQ Dimensions
  - IQ Models
- IQ Assessment
  - Assessment techniques
  - IQ aggregation and ranking
- IQ Improvement
  - Profiling & Data Scrubbing
  - Outlier Detection
  - Duplicate Detection

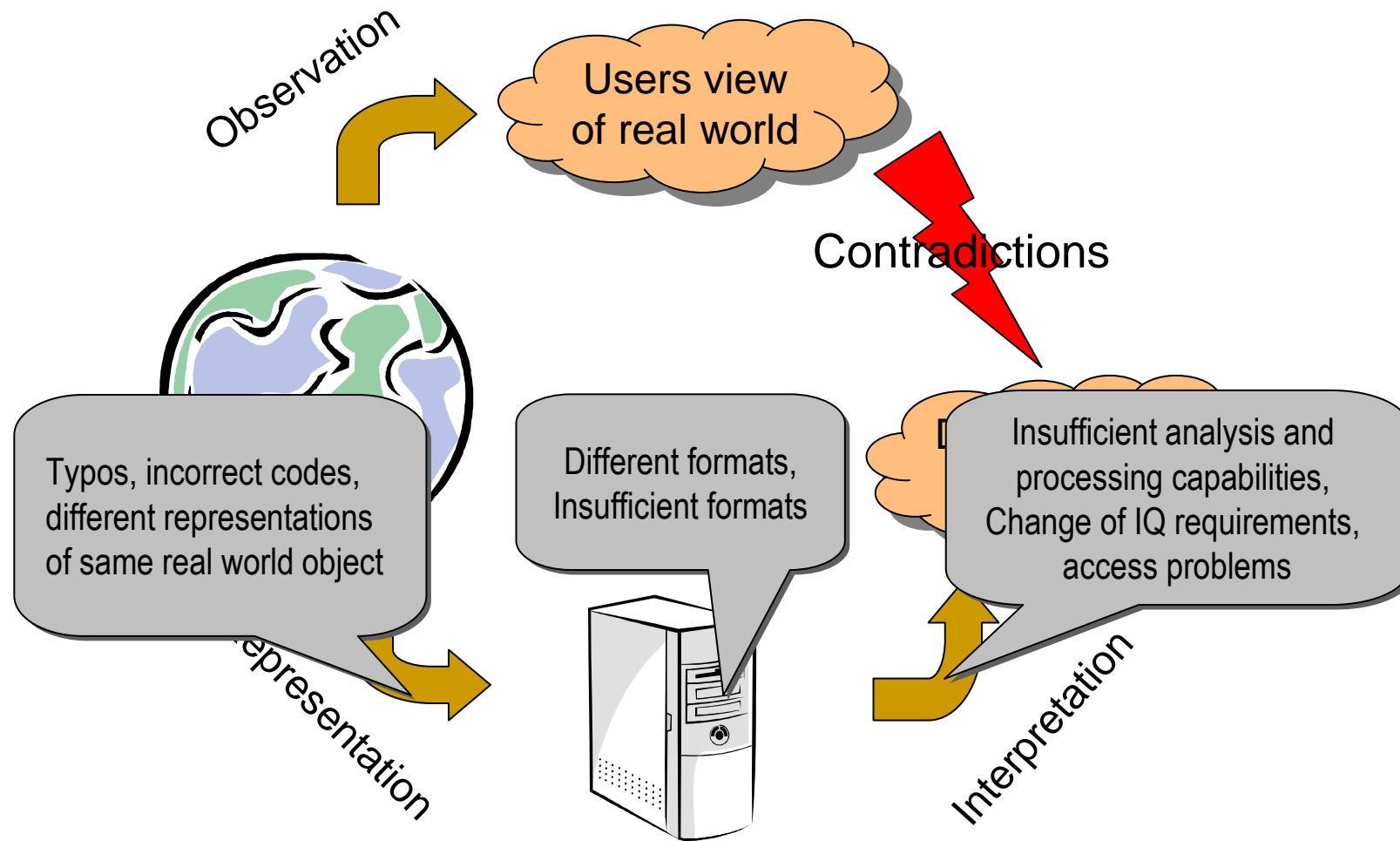


---

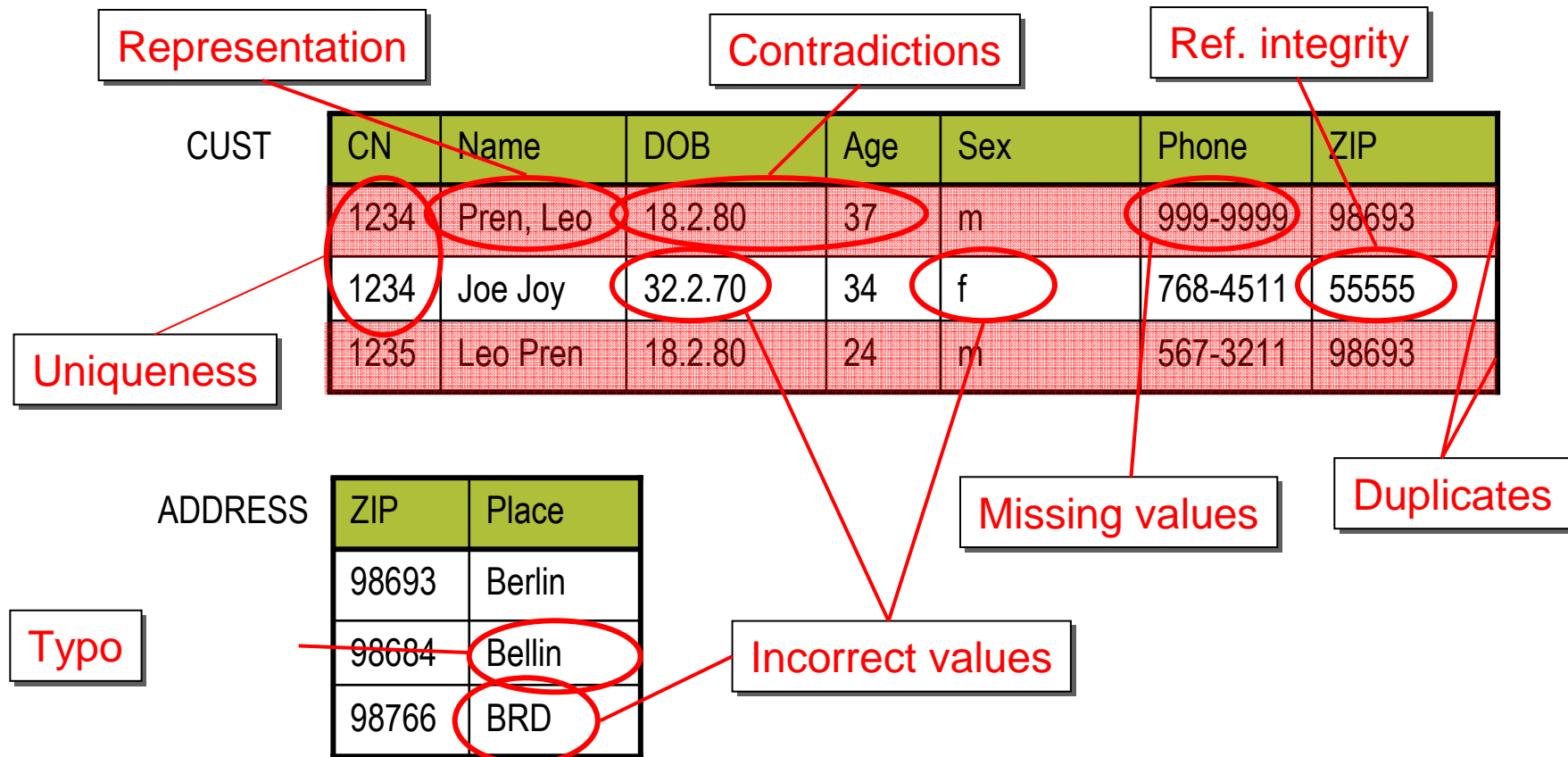
# Motivation

- Incorrect prices in inventory retail databases [English 1999]
  - Costs for consumers 2.5 billion \$
  - 80% of barcode-scan-errors to the disadvantage of consumer
- IRS 1992: almost 100,000 tax refunds not deliverable [English 1999]
- 50% to 80% of computerized criminal records in the U.S. were found to be inaccurate, incomplete, or ambiguous. [Strong et al. 1997a]
- US-Postal Service: of 100,000 mass-mailings up to 7,000 undeliverable due to incorrect addresses [Pierce 2004]
- In 2006 the Fortune 1000 companies will spend more money on IQ problems than for ERP, CRM, and BI together. [Gartner]
- More than 35% of all IT projects fail due to poor IQ. Poor IQ causes annual expenses of 2-4 billion \$ in US. [Meta Group]

# Causes of Poor Information Quality



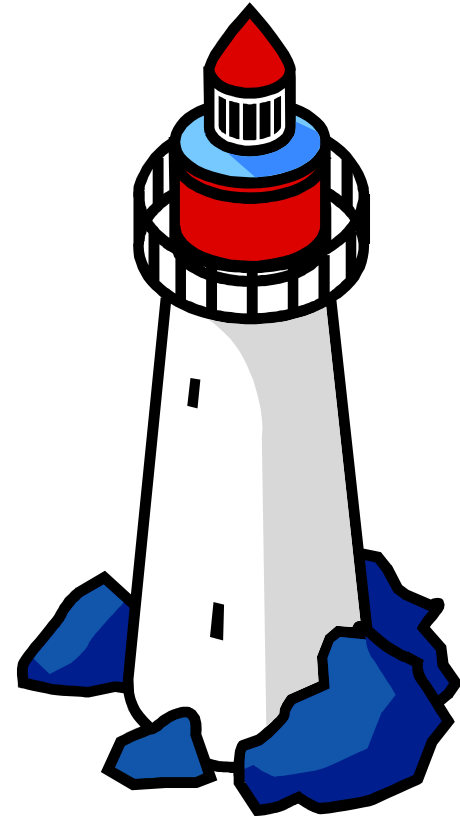
# Information Quality Problems



---

# Overview

- Motivation
- Defining IQ
  - Dimensions and Classifications
  - Models
- IQ Assessment
- IQ Improvement

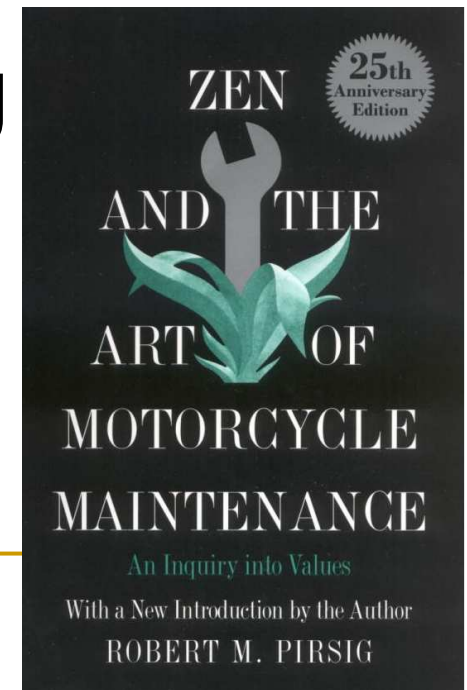


---

# Quality

***“Even though quality cannot be defined, you know what it is.”***

**Robert Pirsig**





---

# IQ from 10000 feet

- General definitions

- „excellence / value“
- „fitness for use“
- „extent to which a product successfully serves the purpose of consumers“
- „meeting / exceeding consumer expectations“
- „inexact science in terms of assessment and benchmarks“

- Observations

- Information quality is **subjective**
  - Depends on context, consumer, etc.
- Information quality is **multidimensional**
  - multiple dimensions (criteria, aspects, properties)

# IQ under the Microscope

Ability to be Joined With	Ability to Download	Ability to Identify Errors	Ability to Upload
Acceptability	Access by Competition	Accessibility	Accuracy
Adaptability	Adequate Detail	Adequate Volume	Aestheticism
Age	Aggregatability	Alterability	Amount of Data
Auditable	Authority	Availability	Believability
Breadth of Data	Brevity	Certified Data	Clarity
Clarity of Origin	Clear Data Responsibility	Compactness	Compatibility
Competitive Edge	Completeness	Comprehensiveness	Compressibility
Concise	Conciseness	Confidentiality	Conformity
Consistency	Content	Context	Continuity
Convenience	Correctness	Corruption	Cost
Cost of Accuracy	Cost of Collection	Creativity	Critical
Current	Customizability	Data Hierarchy	Data Improves Efficiency
Data Overload	Definability	Dependability	Depth of Data
Detail	Detailed Source	Dispersed	Distinguishable
Dynamic	Ease of Access	Ease of Comparison	Updated Files
Ease of Data Exchange	Ease of Maintenance	Ease of Retrieval	Ease of Correlation
Ease of Update	Ease of Use	Easy to Change	Ease of Understanding
Efficiency	Endurance	Enlightening	Easy to Question
Error-Free	Expandability	Expense	Ergonomic
			Extendibility

[Wang Strong 1996]

---

# Finding the right Dimensions

**IQ** := {Understandability, Reputation,  
Reliability, Timeliness,  
Availability, Price,  
Consistency, Coverage,  
Response time, Density,  
Completeness, Amount,  
Accuracy, Relevancy, ... }

---

# Selected IQ Dimensions – Completeness

- The extent to which data are of sufficient **breadth**, **depth** and **scope** for the task at hand
  - [Wang Strong 1996]
- Coverage denotes the estimated **portion** of the **intended complete relation** that is actually present.
  - Trio System [Widom 2005]
- A subset of a database is complete if it includes a representation of **every occurrence** in the real world environment that it models.
  - [Motro 1986]
- Soundness measures the proportion of the stored information that is true, and completeness measures the **proportion of the true information** that is stored.
  - [Motro Rakov 1998]
- Coverage is the **probability** that a source has some answer to a given query.
  - [Florescu et al. 1997]

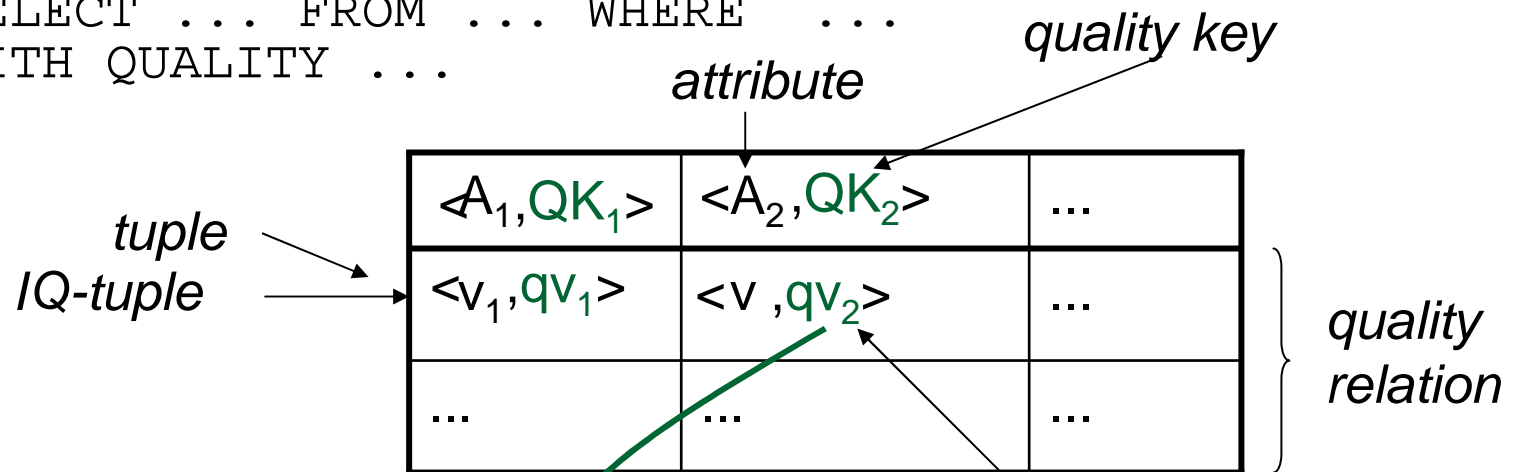
---

# IQ Models

- Data models
  - Common theme: Enrich conventional data model with elements to represent and analyze IQ.
  - Conceptual modeling
    - ER-Extension: Quality ER-Model [Storey Wang 1998]
  - Logical modeling
    - Extension of relational model
      - Polygen [Wang Madnick 1990]
      - Attribute-based model [Wang et al. 1995]
    - Trio DBMS for data, accuracy, and lineage [Widom 2005]
    - Extended XML-Model: D2Q [Scannapieco et al. 2004]
- Process model
  - Model for data production process
    - IP-MAP [Shankaranarayanan et al. 2000, Wang et al. 2003]

# Attribute-based Model

- Extension of the relational model: „cell tagging“
  - Attributes with different levels of quality indicators
- Extension of the relational algebra
- Queries against quality indicators
  - `SELECT ... FROM ... WHERE ...`  
`WITH QUALITY ...`



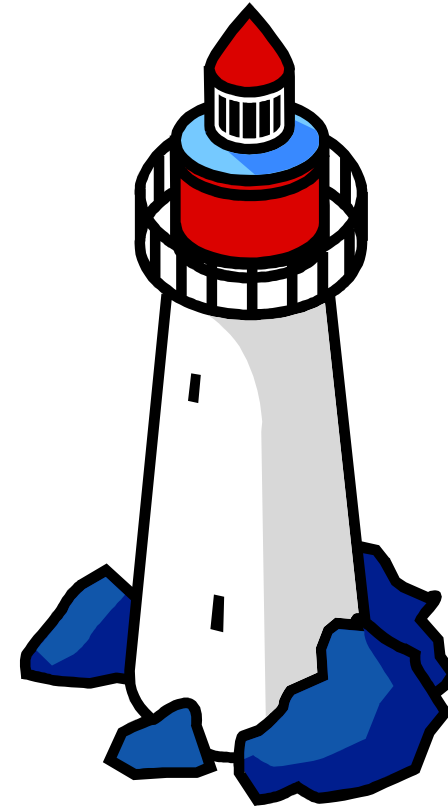
[Wang et al. 1995]



---

# Overview

- Motivation
- Defining IQ
- ➔ ■ IQ Assessment
  - Assessment techniques
  - IQ aggregation and ranking
- IQ Improvement



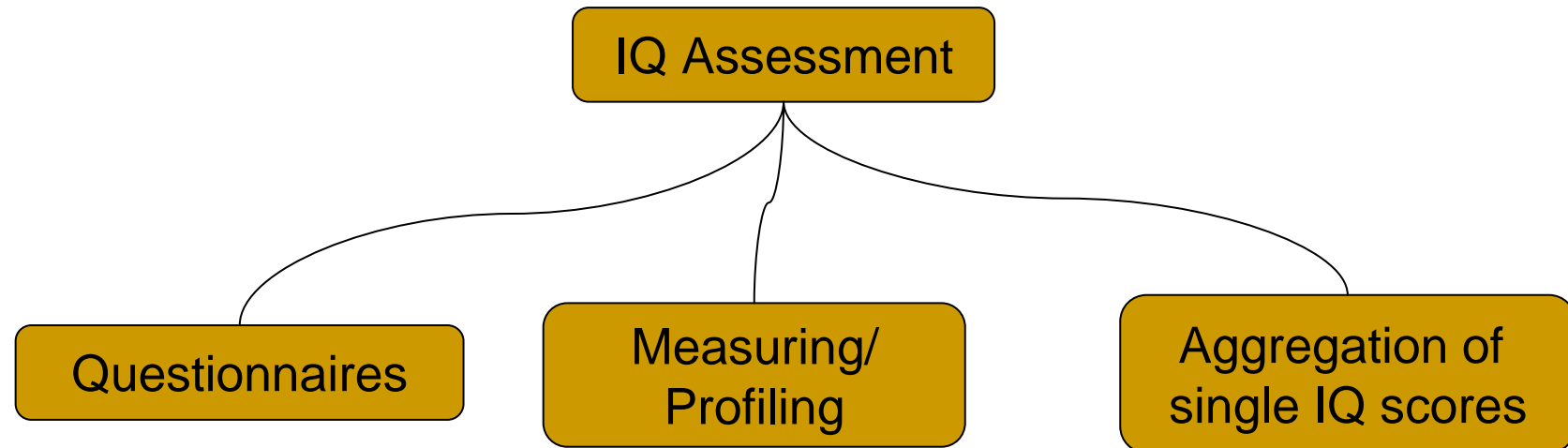
---

# IQ Assessment

- “You cannot control what you cannot measure” [DeMarco 82]
  - Why assess IQ?
    - Estimating quality, relevance, significance, ... (“GIGO”)
    - Need for improvement? Cost-benefit ratio?
  - Measurement: quantitative comparison between an observation and a reference value
  - Metrics:  $f(\text{IQ dimension}) \rightarrow \text{IQ score}$
  - Requirements: Understandable, combinable, precise, feasible, efficient
  - But:
    - Context-specific issues  $\rightarrow$  subjective measures
    - IQ values are rarely published, high data volume, frequent updates, ....
-

---

# Assessment Techniques



control matrices,  
AIMQ, ...

Objective measures, MADM techniques  
Data profiling, sampling,  
...

# Techniques: Questionnaire

- For subjective, non-functional criteria
- Comparison to real-world state
- Exploiting human expertise
- Example: control matrices [Pierce 04]
  - Matrix for steps in data production process affecting IQ

## IQ problem

	duplicates	typos	missing values
check	yes		
Check #1			
Check #2		8%	
Check #3		12%	45

rating (yes/no, category, score)

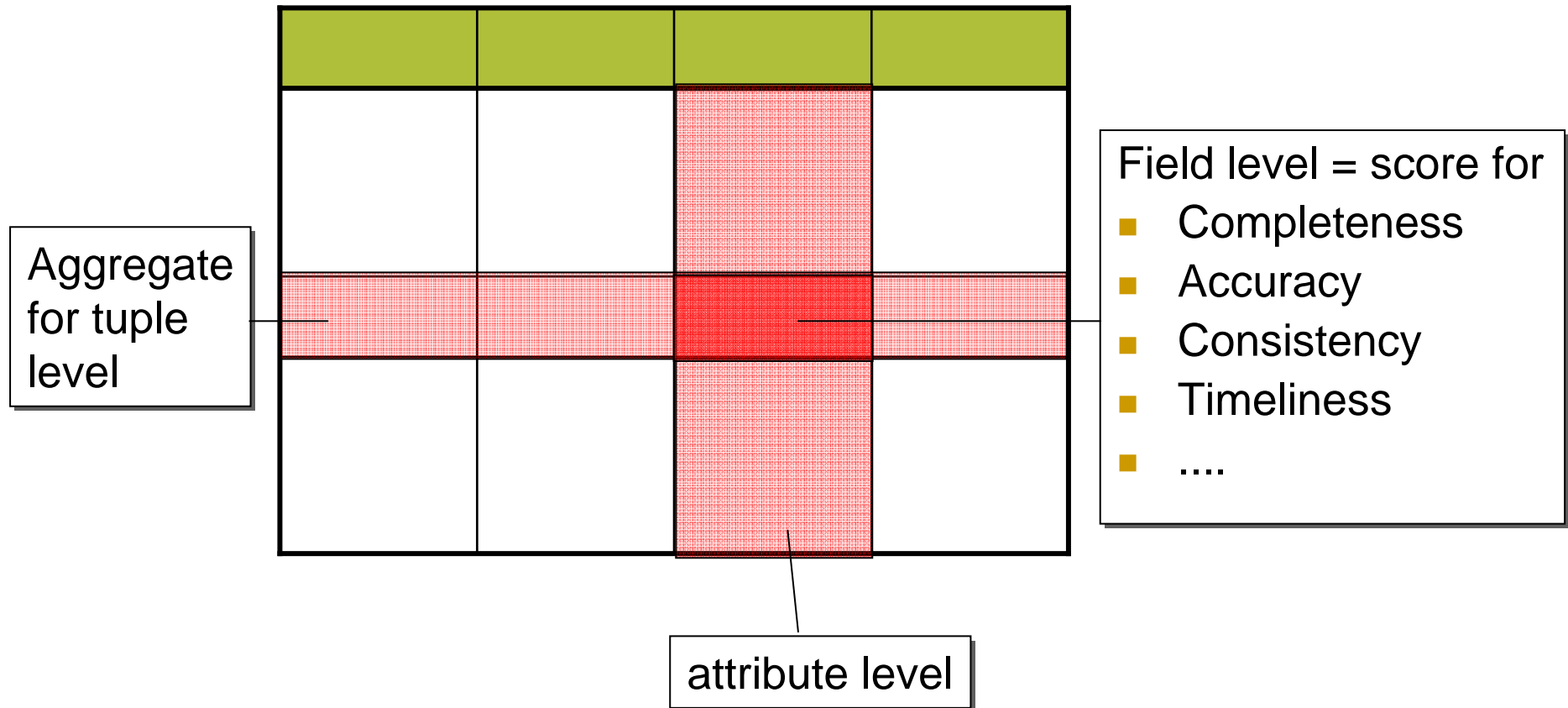
Estimating overall IQ scores by combining ratings

---

# Techniques: Measuring

- Completeness: absence of null values, but beware of semantics of null
- Accuracy: distance between current value  $w$  and correct value  $w'$ 
  - Syntactic distance: numeric values  $|w-w'|$ , string values `edit_distance( $w, w'$ )`
  - Semantic distance: Munich=München, BMW=Bayerische Motorenwerke
- Consistency: ratio of correct values wrt.
  - Integrity rules, business rules, ...
- Timeliness:  $1/(\text{update frequency} \cdot \text{age})$

# Measuring



# Measuring /2

- Example: completeness of relation  $r$  with  $R(A_1, A_2, \dots, A_n)$

- Non-null values of  $A_i$ :

$$N_A = \{ t \in r \mid \text{NotNull}(t.A) \}$$

- Completeness for  $A_i$ :  $\frac{|N_A|}{|r|}$

- Completeness for  $A_1, \dots, A_k$ :  $\frac{|N_{A_1, \dots, A_m}|}{|r|}$

- With attribute weighting:  $\frac{\sum_{t \in r} (\sum_{i=1}^n w_i \cdot \text{NotNull}(t(A_i)))}{|r|}$

---

# Aggregation of Measurements

- Ratio: non-null values vs. total cardinality
  - For completeness, accuracy, ...
- Minimum/maximum
  - For timeliness, response time, ...
- Sum
  - For access costs, ...
- Product
  - For availability, ...

# Combining Multiple IQ Dimensions

- IQ score = vector of (completeness, exactness, ...)
- How to compare IQ scores?

$$\boxed{0.1, 0.3, 0.9, 0.4, \dots} < \boxed{0.2, 0.2, 0.8, 0.45, \dots} \quad ?$$

IQ dimensions with different

- scales
- ranges
- importance

Therefore

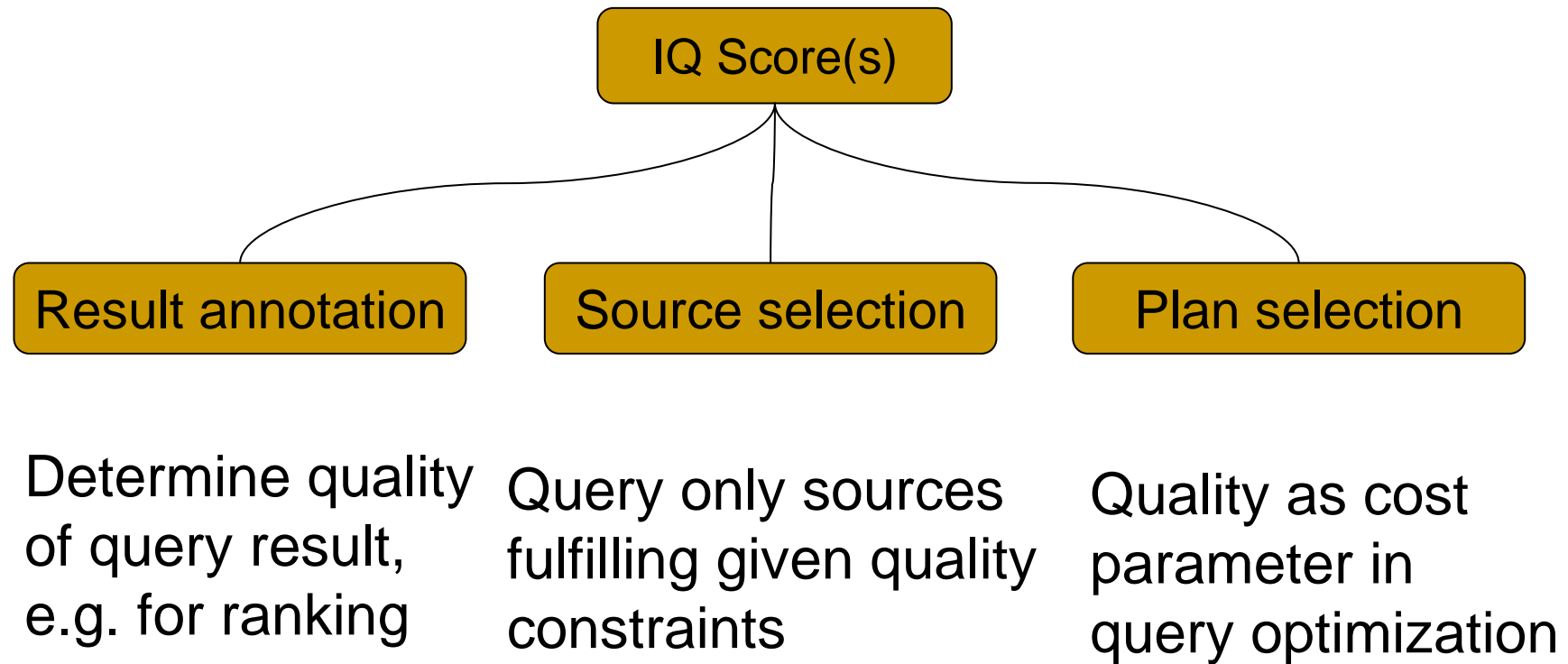
- convert
- scale/normalize
- weight

Multi attribute decision making (statistical techniques)

E.g. simple additive weighting, data envelopment analysis

---

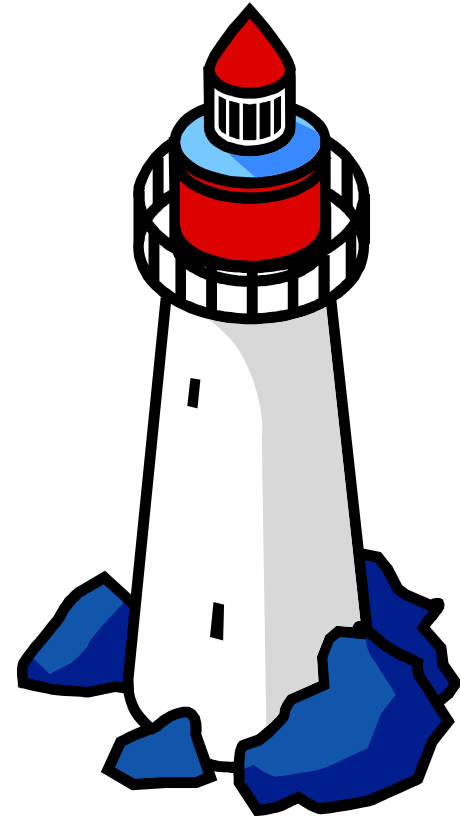
# IQ Interpretation



---

# Overview

- Motivation
- Defining IQ
- IQ Assessment
- ➔ ■ **IQ Improvement**
  - Cleaning Steps
  - Profiling and Data Scrubbing
  - Outlier Detection
  - Duplicate Detection



---

# Data Cleaning

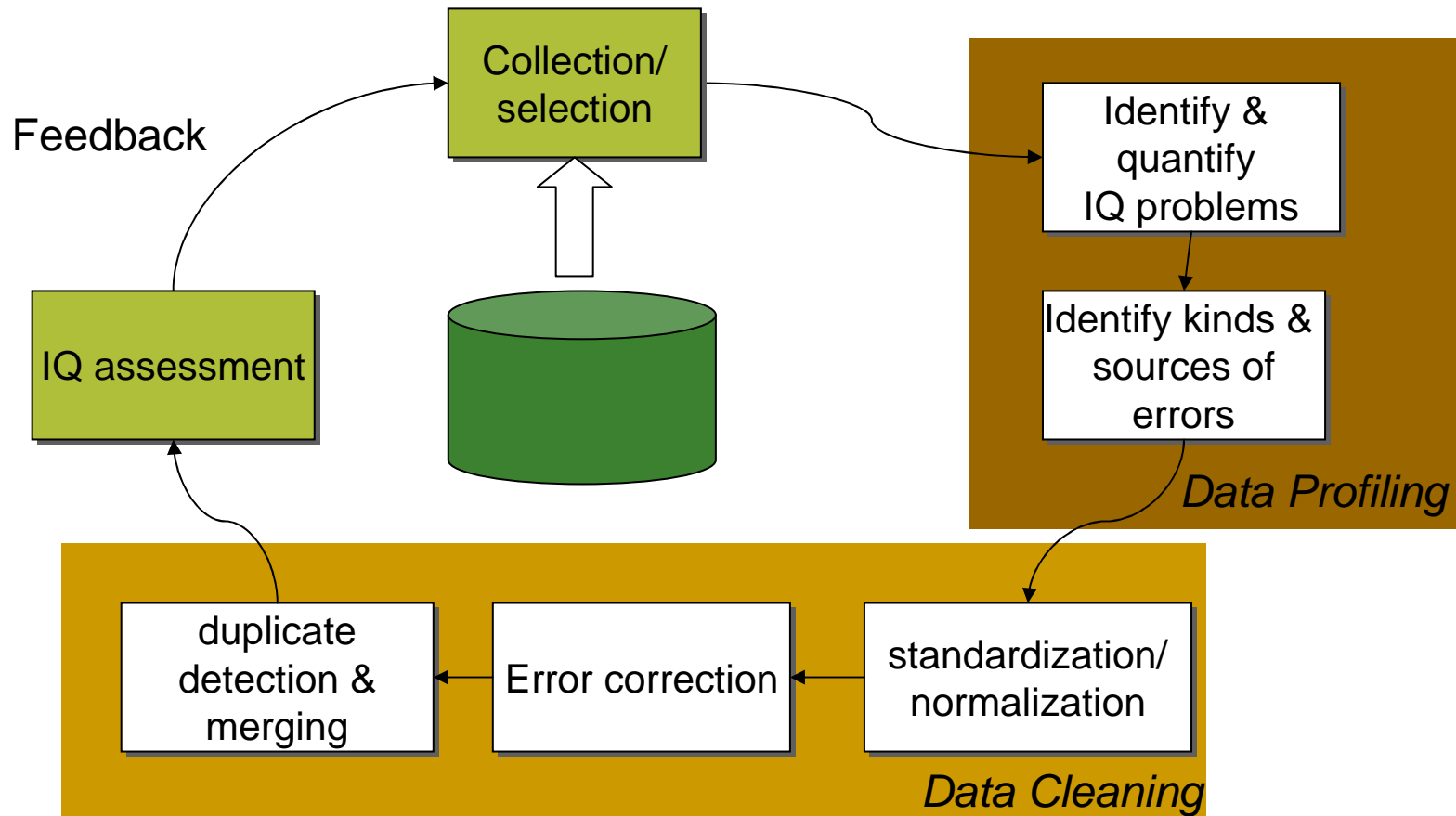
- Identifying & eliminating inconsistencies, discrepancies and errors in data in order to improve quality
  - aka „data cleansing“ or „data scrubbing“
  - Up to 80% of costs in DW projects
- Possible effects of dirty data (e.g. duplicates)
  - Example: Portfolio Management Offers
  - Credit maximum not detected
  - No quantity discount for multiple orders
  - Multiple mailings of same catalog to same household
- General problems
  - Additional, unnecessary IT expenses
  - Low customer satisfaction

# Avoiding dirty data in DBMS

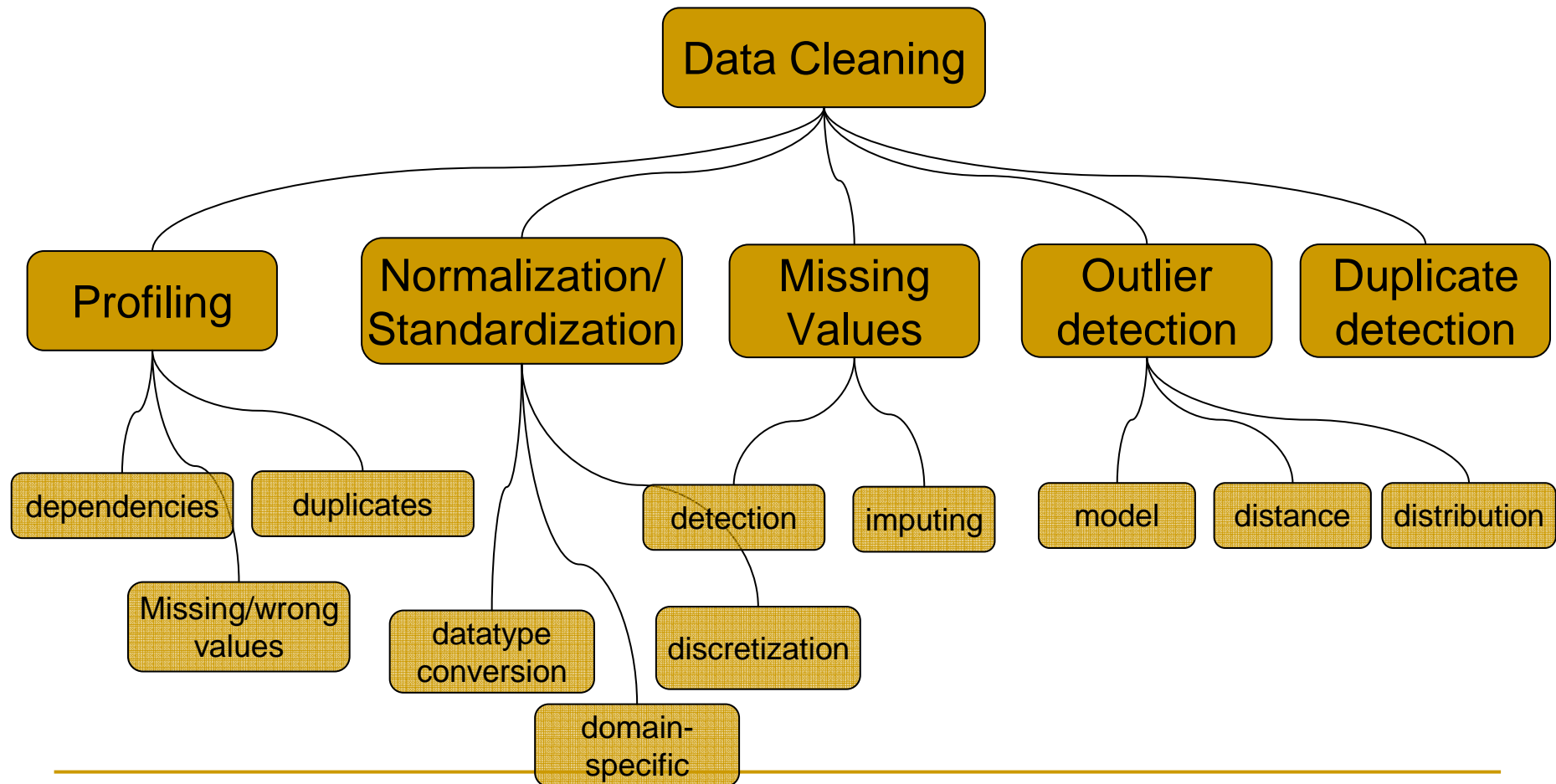
avoiding	by
wrong data types	type definition, DOMAIN constraints
wrong values	CHECK
missing values	NOT NULL
invalid references	FOREIGN KEY
duplicates	UNIQUE, PRIMARY KEY
inconsistencies	ACID transactions
outdated data	replication, materialized views

- So, why is data still dirty?
  - ❑ Missing metadata, integrity constraints, ...
  - ❑ Data from „foreign“ sources
  - ❑ Non-DBS sources
  - ❑ typos, lack of knowledge, ...
  - ❑ Multi source problems, heterogeneities

# Steps of Data Cleaning



# Cleaning Tasks



---

# Profiling

- Analysis of content and structure of attributes
  - Data type, domain, data distribution and variance, occurrence of null values, uniqueness, pattern (e.g. mm/dd/yyyy)
- Analysis of dependencies between attributes of a single relation
  - Functional dependencies, primary key candidates, „fuzzy“ dependencies
- Analysis of overlapping attributes from different relations
  - Redundancies, foreign keys

---

# Profiling /2

- Missing or wrong values
  - current vs. expected cardinality (e.g. number of shops, gender of customers)
  - frequency of null values, minimum / maximum, variance
- Data and input errors
  - Sorting and manual inspection
  - Similarity checks
- Duplicates
  - Number of tuples vs. Cardinality of attribute domain
- „Fuzzy“ keys, functional dependencies and joins
  - no explicitly defined integrity constraints
  - But satisfied in most cases

---

# Profiling with SQL

- SQL queries for basic profiling tasks

- schema, data types: querying data dictionary

- Domain of data

- ```
SELECT MIN(A), MAX(A), COUNT(DISTINCT A)
FROM DataTable
```

- Erroneous data, default values

- ```
SELECT City, COUNT(*) AS Cnt
FROM Customer
GROUP BY City ORDER BY Cnt
```

- ascending: typos, e.g. Illmenau: 1, Ilmenau: 50

- descending: undocumented default values, e.g. AAA: 80

---

# Data transformation and normalization

- Data type conversion: varchar → int
- Normalization: mapping into a common format
  - date: 03/01/05 → 01-MAR-2005
  - currency: \$ → €
  - Uppercase strings
  - tokenizing: „Date, Chris“ → „Date“, „Chris“
- Discretization of numerical values
- Domain-specific transformations
  - Codd, Edgar Frank → Edgar Frank Codd
  - St. → Street
  - Address transformation using address databases
  - Domain-specific product names/codes (e.g. in pharmacy)

---

# Detecting missing values

- Missing data:
    - Treating null values: missing value or default value
    - Biased data, e.g. caused by null values
  - Basic analysis:
    - Number of null values, duplicates, mean, frequency, ...
    - Comparing with expected values
    - Analyzing order of tuples
      - No sales information during 03/01...03/04?
  - Incomplete data, e.g. truncated and censored data
    - Sales with  $< 1$  € are not collected in the dataset
    - Sales with  $> 100$  € stored as 100 €
  - Detection: by analyzing data distribution but often domain knowledge required
-

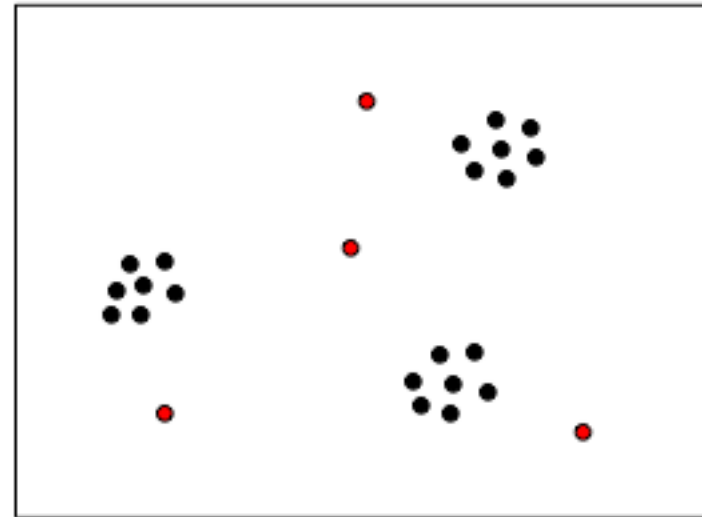
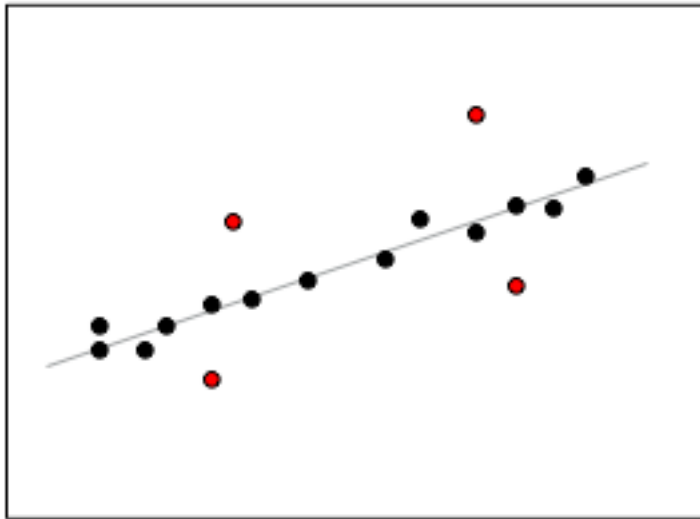
---

# Imputing missing values

- „Unbiased estimators“
  - Estimating missing values without changing characteristics of existing dataset (mean, variance, ...)
  - E.g.: 1, 2, 3, \_\_, 5 → (mean: 2.75; variance: 4.659)
- Exploiting functional dependencies
  - E.g.: #Bedrooms → Income
- Techniques from statistics
  - Linear regression:  
income =  $c \cdot \text{\#Bedrooms}$
  - techniques for non-linear dependencies:
    - Neural networks, ...

# Outlier detection

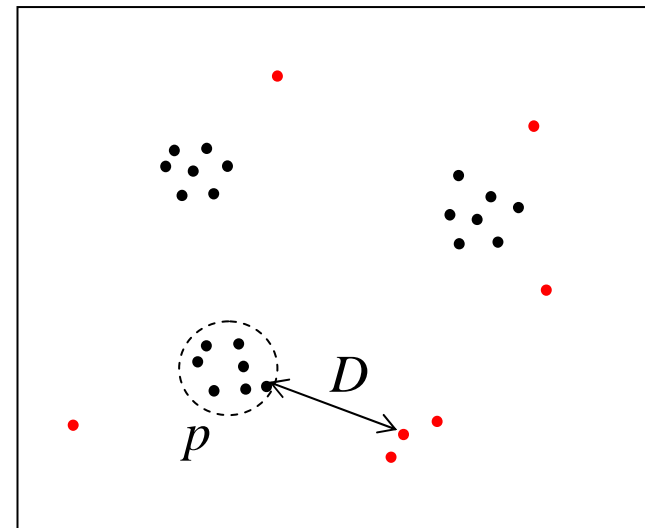
- Outlier: „suspicious“ observation that deviates too much from other observations
- Issues:
  - detection: distribution, „geometry“, distance, time series
  - interpretation: data or observation error vs. real event



# Distance-based outliers

- Object  $o$  in dataset  $T$  is a  $DB(p,D)$ -outlier, if at least a fraction  $p$  of  $T$  lies greater than distance  $D$  from  $o$   
[Knorr Ng 1998]

Outlier = object with not enough neighbors  
Parameter  $p$  for determining „cluster of outliers“



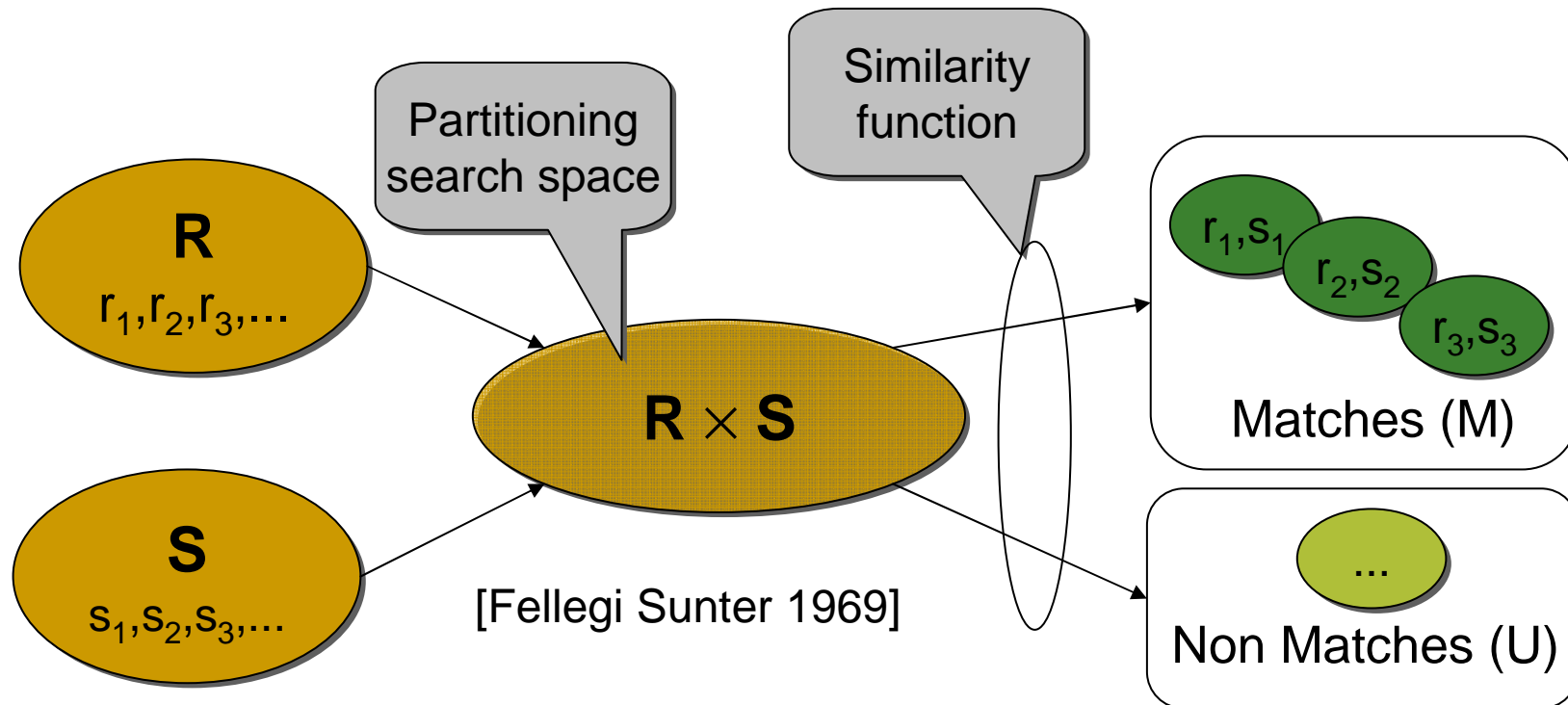
---

# Duplicate Detection

First name	Last name	Address	ID
Sal	Stolpho	123 First St.	456780
Mauricio	Hernandez	321 Second Ave	123456
Klemens	Böhm	Hauptstr. 11	987654
Sal	Stolfo	123 First Street	456789

- Many duplicate terms:
  - ❑ Duplicate detection / de-duplication
  - ❑ Record linkage
  - ❑ Object identification / object consolidation
  - ❑ Entity resolution / entity clustering
  - ❑ Householding / household matching
  - ❑ Merge/purge

# The Problem



- **Duplicate Detection (Record Linkage)**
  - Identification of semantically equivalent representations, i.e., representations of the same real-world object
- **Duplicate Elimination (Reconciliation / Fusion)**
  - Create a complete, concise, and consistent data set.

# Similarity functions

## ■ Edit-based

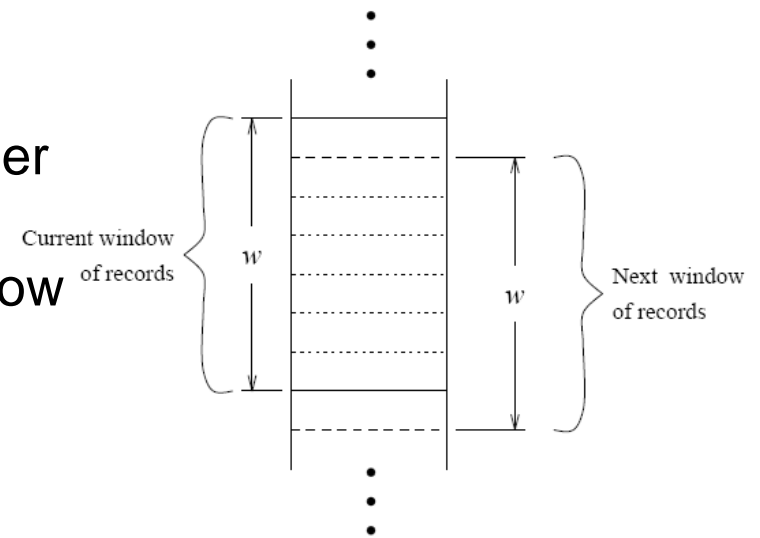
- Edit-distance / Levenshtein-distance [Levenshtein 1965]
  - Minimum number of edits from one word to the other
  - Domain-specific costing
  - Also: Smith-Waterman [Smith Waterman 1981]
    - Compensates abbreviations
- Soundex
  - 4-letter code for each word
  - SOUNDEX('Farwick ') = F620
    - Fähruschi, Feuerhake, Frass, Fricke
- Jaro [Jaro 1989] / Jaro-Winkler [Winkler 1999]
  - Common letters within  $\frac{1}{2}$  string length
  - Transposed letters
- SQL LIKE
  - Precision / Recall tradeoff
    - Fr% vs. Frick%
  - Expensive, no similarity scoring

## ■ Token-based

- Tokens
  - Words / Terms
  - n-grams
- Jaccard
  - $\frac{|\{\text{common tokens}\}|}{|\{\text{all tokens}\}|}$
- TFIDF [Cohen et al. 2003]
  - Term frequency (tf)
  - Inverse document frequency (idf)
  - Tfidf:  $\log(\text{tf}+1) \times \log \text{idf}$
  - Common words have low weight
  - Cosine similarity of term vectors weighted by tfidf
- And many more [Koudas Srivastavasa 2005]
- Domain-dependent
  - Special similarity for dates, numerical attributes, names, addresses
  - e.g. rules

# Algorithms: Sorted Neighborhood

- Idea
  - Sort tuples so that similar tuples are close to each other.
  - Only compare tuples within a small neighborhood (window)
- Generate key
  - E.g.: SSN+“first 3 letters of name“ + ...
- Sort by key
  - Similar tuples end up close to each other
- Slide window over sorted tuples
  - Compare all pairs of tuples within window
- Problems
  - Choice of Key
  - Choice of window size
- Complexity: at least 3 passes over data
  - Sorting!



[Hernandez Stolfo 1998]

---

# Data Fusion / Reconciliation

- Duplicate elimination
  - Keep any tuple
  - Keep best tuple
    - Subsumption
    - Highest quality tuple
- Duplicate fusion
  - Conflicts among duplicates
  - Conflict resolution functions

---

# Tools for data cleaning

[Barateiro Galhardas 05]

- Auditing & Profiling
  - Axio (EvokeSoft), WizWhy (WizSoft), DB-Examiner (DBE Software), ...
- Transformation
  - SQL Server 2005, Oracle Warehouse Builder, Hummingbird ETL, ...
- Cleaning & Duplicate elimination
  - Trillium, dfPower (DataFlux), WizRule & WizSame, FirstLogic, Sagent, ...

---

# Conclusions

- Modern DBMS provide many features for ensuring high quality data
  - Integrity constraints, ACID transactions, ...
  - But nevertheless: data quality problems

**Real world data is dirty!**

- Information quality has many facets
  - Notion of quality
  - Assessment & interpretation
  - Data cleaning
- But often context-dependent, requires domain knowledge