Data Quality
Defining, Measuring and Improving

Hans - J. Lenz, Freie Universität Berlin

September 2007

All you should know about *Data Quality ...*

- ´Good enough´ is not ´good enough´
- „Even though quality cannot be defined, you know what it is“ R. Rirsig
- „It is quality rather than quantity that matters“ Seneca
- Quality Improvement is a never ending process
- Garbage in garbage out.

You should know the cost of ´poor data quality´...

- Erroneous prices in sales data led to annual disadvan-tages of clients in the USA of about 2,5 Billion $!
80 % of Barcode-Scan-errors negatively affected the consumers*
[GB 1999]

- In 2004 on the average 7% of the mail could not delivered due to wrong addresses.*
[Pierce 2004]

- In 1992 the US Treasury Dept. could not issue 100000 pay-back cheques due to erroneous addresses.
[USA 1992]

- ...

*Sourve: cited from Leser and Naumann (2007); www.information-quality.com

Data Quality – Exa

errors

No.	ISBN	Title	Name	Year	Pages
1	0-201-54329-X	An introduction to database systems	Date	**1997**	839
2	0-201-54329-X	An introduction to database systems	Date	**1995**	839
...		...			
		An introduction to spatial database systems	**Gueting**	1994	13
10		An introduction to spatial database systems	**Güting**	1994	13
...		...			
26	0-**210**-14456-5	An introduction to database systems	Date	1977	536
27	0-**201**-14456-5	An Introduction to Database Systems	Date	1977	
...		...			
30		An introduction to database systems	Date		

mistypings

variations

missing values

Poor Data Quality

- Not feasible values: Birthday = 31 Febr, 2007
- Violated constraints S = Q * P: Sales S= 100, quantity Q= 20 , price P= 4
- Missing Generalisation: Business Partners = Customers ∪ Suppliers ∪ Banks
- Wrong measurement units, i.e. *l* /100 km instead of mi/gal
- Missing Values: FamilyName = ´XXXXXXXXXX´
- Ambiguity of aggregation: German GDP with / witho
- Missing definition of an attribute: Age, i.
- Ignoring scale: avg(street_number)
- pretended precision - omitting error ra

Question:
Can we structute the
various kinds of poor
data quality using an
appropriate taxonomy?

Aug. 14, 2007, 7:03AM

German GDP Cools in 2nd Quarter

Europe's biggest economy had expanded 0.5 percent in the first three months of this year and 1 percent in last year's final quarter.

By GEIR MOULSON Associated Press Writer

Contents

1. Definition of (Data) Quality
2. History
3. Poor Data Quality
4. Indicators („Measures", „Dimensions") of Data Quality
5. Measuring, Aggregation and Using of Data Quality

1 Definition of Quality

- W. Werz (1915): Quality reflects the ability of an object to meet the purpose
- ISO Norm: Suitability for use relative to a given objective of usage
- Industry: Quality is the conformance to requirements
- Computer Science: Fitness for use - given a purpose

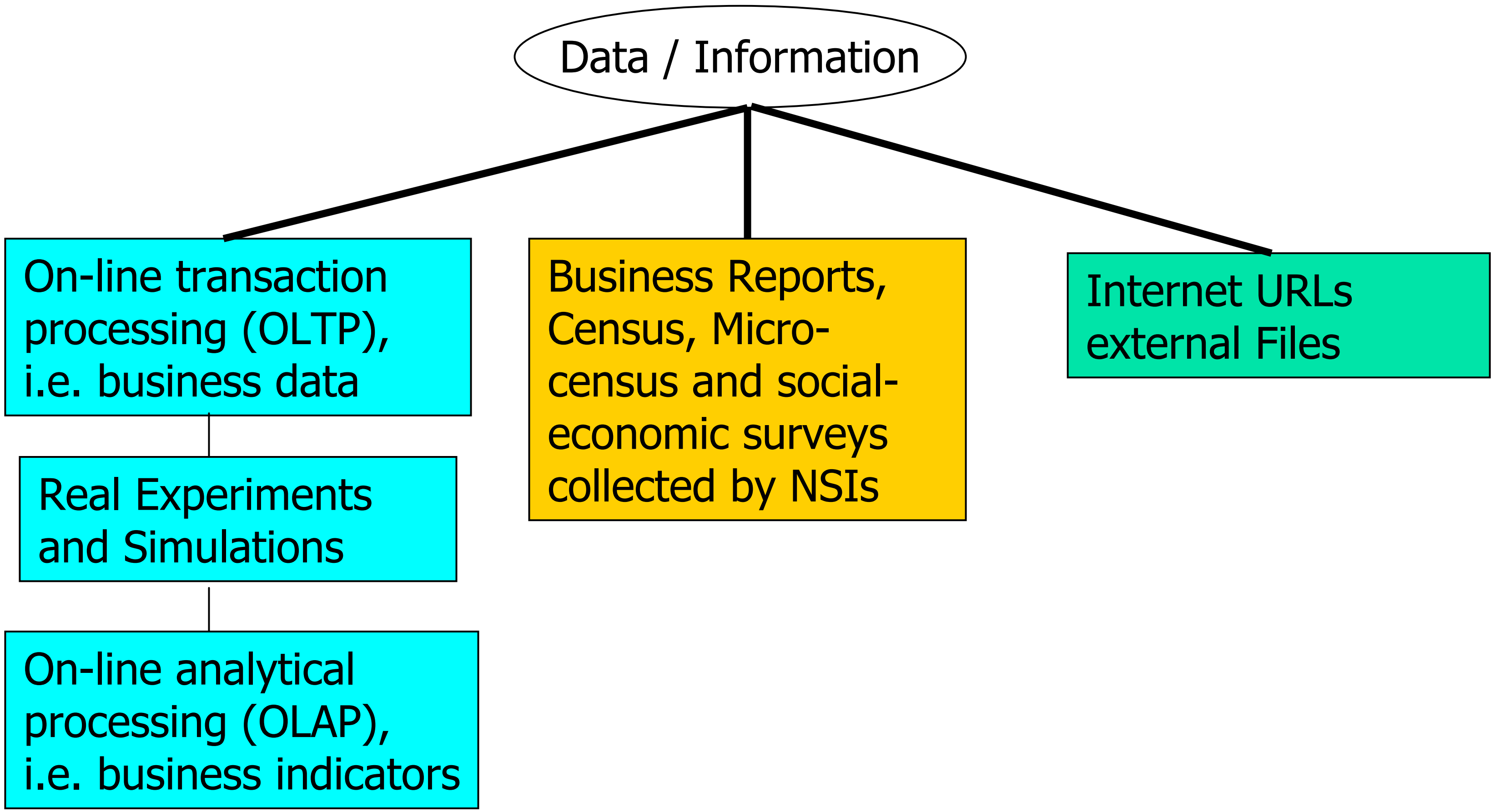
2 A short History of QC

- The quality movement can trace its roots back to guilds in the late 13th century (craftsmanship model).
- The factory system, with its emphasis on product inspection, started in Great Britain in the mid-1750s and grew into the Industrial Revolution in the early 1800s.
- In the early 20th century, manufacturers began to include quality processes in quality practices.
- Walter Shewhart's (1922) statistical process control techniques.
- World War II (1939-1945), quality became a critical component of the war effort: Bullets manufactured in one state, for example, had to work consistently in rifles made in another.
- QC - Heros: Americans Joseph M. Juran, W. Edwards Deming and H. Taguchi
- Data Quality Control: in Statistical Offices since about 1930
in Business since about 1990

Source: www.asq.org

3 Data Sources

Where do business and economic data come from?



3 Kinds of Data Quality

Quality

Quality of Design

- Database schema
- Repository (Meta DB)
- Experimental Design
- Surveys, Reports, Census
- Application Software

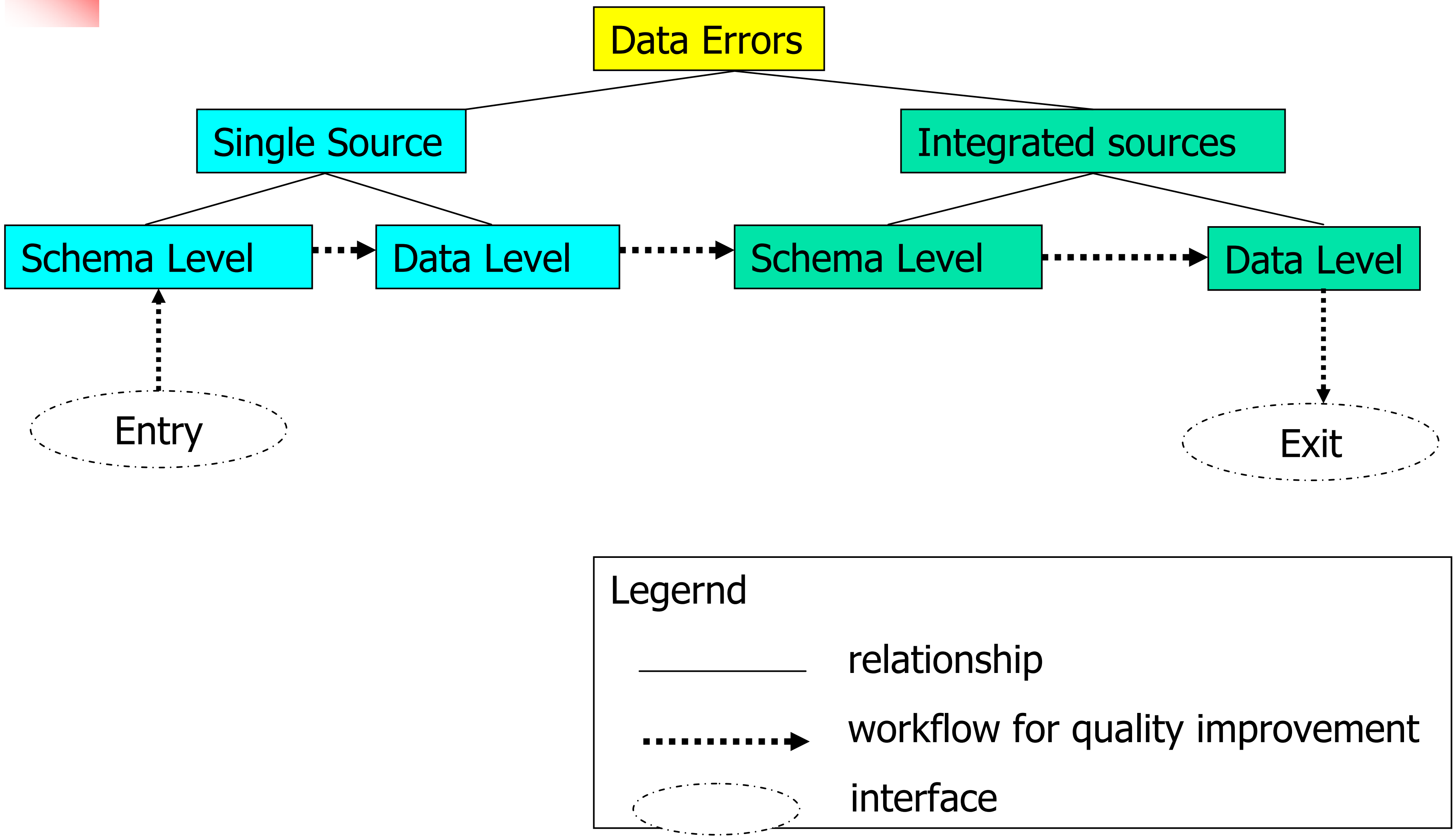
Quality of Performance

- ETL, Integration
- Data cleansing
- Insert, Delete, Update
- Querying
- Disclosure Control

3 Data Errors as trouble makers

Top Scorer of German Bundesliga in the last 5 years				
Jahr	Name	Verein	Treffer	Nationalität
2003	Elber	FC Bayern	21	Brasilien
2003	Christiansen	VfL Bochum	21	Dänemark
2004	Ailton	Werder Bremen	2004	Braslien
2005	Mintál	1.FC Nürnberg	24	Slowakai
2006	Klose	FC Bayern	25	Polen
2007	Neumeister	VfL Bochum	20	Griechenland

3.1 Data Error Classification



Source: Naumann (2007); Rahm and Do (2000)

3.1 Data Error Classification

- Data Errors
  - Single Source
    - Schema Level
      - infeasible value
      - attribute dependency ignored
      - attribute uniqueness ignored
      - referential integrity
    - Data Level
  - Integrated sources
    - Schema Level
    - Data Level

ICs of DB schema
not satisfied !!!

Source: Naumann (2007); Rahm and Do (2000)

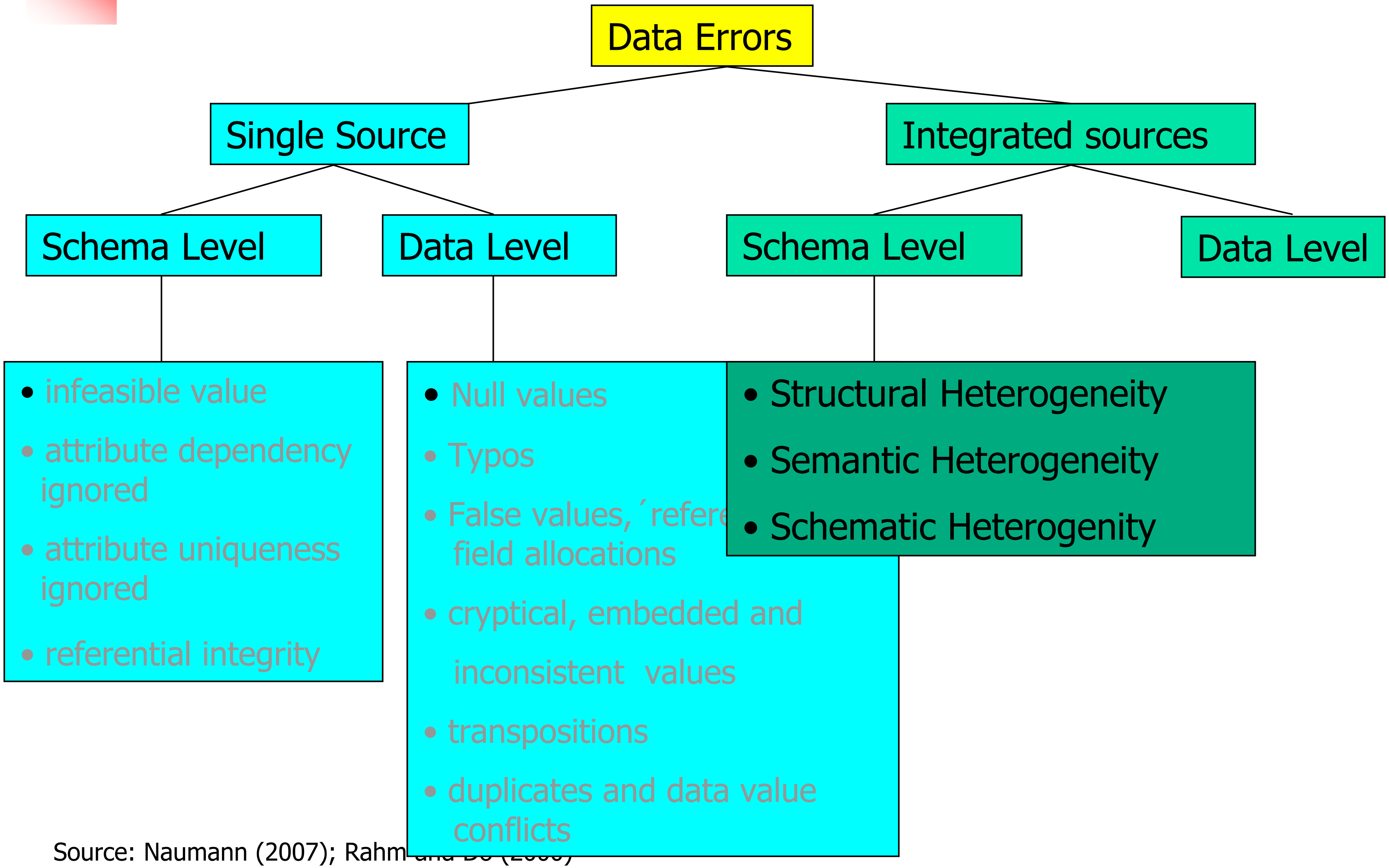
3.1 Data Error Classification

- Data Errors
  - Single Source
    - Schema Level
      - infeasible value
      - attribute dependency ignored
      - attribute uniqueness ignored
      - referential integrity
    - Data Level
      - Null values
      - Typos
      - False values,´references, and field allocations
      - cryptical, embedded and inconsistent values
      - transpositions
      - duplicates and data value conflicts
  - Integrated sources
    - Schema Level
    - Data Level

Errors not resolvable by schema!

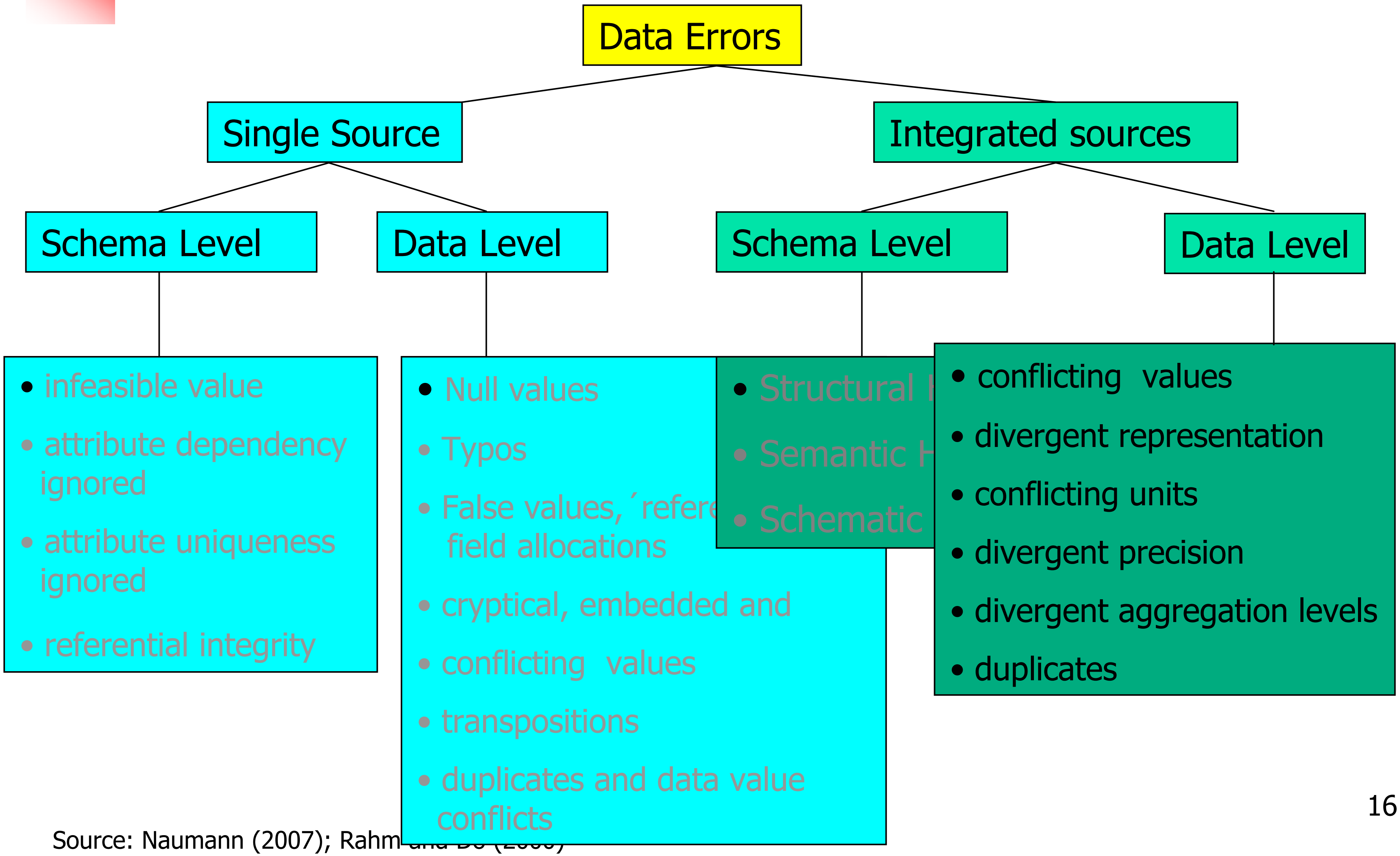
Source: Naumann (2007); Rahm and Do (2000)

3.1 Data Error Classification



Source: Naumann (2007); Rahm and Do (2000)

3.1 Data Error Classification



Source: Naumann (2007); Rahm and Do (2000)

3.2 Data Quality:
Focus and Dimensions

- Data Quality is not a single (scalar) quantity, but
  - **Data Quality is a multivariate indicator**
  - The components are called „dimensions“
- Data Quality is not only focussing on
  - entities (characteristics carriers)
  - single attributes
  - records
  - tables(relations)
  - databases, but even on
  - **various data sources** (internal OLTP-DB, external files or Web) and **data consumers.**

Taxonomy of Data Quality by classes and dimensions

Class	**Dimension**
Intrinsic Data Quality	believabilitiy accuracy objectivity reputation
Contextual Data Quality	value-added relevancy timeliness completeness amount of data
Representational Data Quality	interpretability understandability representational consistency representational conciseness
Accessibility Source: Wand and Strong(1996), Naumann: (2007)	accessibility access security

Taxonomy of Data Quality by classes and dimensions

Class	**Dimension**	**Short Explanation**
Intrinsic Data Quality	Believabilitiy	accepted, true, , real credible
	Accuracy	correct, error-free, reliable
	Objectivity	not manipulated, impartial
	Reputation	trusted w.r.t. content, source
Contextual Data Quality	value-added	beneficial, advantageous
	Relevancy	useful, applicable for task
	Timeliness	data age is sufficient for task
	Completeness	data depth, breadth, scope okay
	Appropriate amount of data	right sized volume
Representational Data Quality	Interpretability	language, units, DEF okay
	Understandability	no ambiguity, comprehendable
	representational consistency	unique fixed data format
	representational conciseness	compactness of representation
Accessibility	Accessibility	available, easily& quickly retrieved
	access security	disclosure control effective

Source: Wand and Strong(1996), Naumann: (2007)

3.3 Single Dimension

Hard Quality Dimensions:

- Accuracy
- Completeness
- Time-related Dimensions
  - Currency
  - Volatility
  - Timeliness
- Consistency ...

Soft Quality Dimensions: (not covered here)

- Believability
- Objectivity
- Reputation ...

3.3.1 Accuracy

Objective: Indicator of overall correctness of objects

- DEF.: Let $D \subseteq D`$ be multi-variate domains and $\mathbf{x} \in D`$ an observation or measurement and $\mu \in D$ a corresponding value.

  Accuracy is defined as the closeness or similarity sim: $D` \times D \rightarrow R_{\geq 0}$
- Ex.1: x=´Hanz´; μ=´Hans´
- Ex.2: x = 100; μ=95; measurement error e_{μ} =10%.
- Synonyms: inverse error rate, integrity, correctness.

Batini and Scannapieco(2007), Mohan and Willshire(1999)

3.3.1 Syntactic Accuracy

- DEF.: Syntactic Accuracy
  Let $D \subseteq D`$ be multi-variate domains and $\mathbf{x} \in D`$ an observation or measurement and D the corresponding target domain.

  Syntactic Accuracy is the closeness of **x** to any $\mathbf{y} \in D$.

- Ex.: x=´Hans´; y=´Haus´;$d_{Edit}(x,y)=1$
- *Syntactic Accuracy* is a necessary condition for (overall) correctness.

3.3.1 Semantic Accuracy

- DEF.: Semantic Accuracy
  Let D be a multivariate domain and
  $\mathbf{x} \in D$ an observation or measurement and $\mu \in D$ the corresponding correct value.
  Semantic Accuracy („correctness") is defined as the closeness or similarity sim: $D \times D \rightarrow R_{\geq 0}$ with $s = \text{sim}(\mathbf{x}, \mu)$.
- Ex.: Nationality(Pélé)=Germany is semantically wrong, but syntactically okay
- Note: Object identification (= approximate joins) makes use of semantic accuracy

3.3.1 Accuracy Measures

Relation Accuracy

- Percentage of tupels (records) without data errors in table T, i.e.

$$\mathrm{acc}(T) = \sum_{i=1}^{|T|} \frac{\varphi(t_i)}{|T|} 100$$

where φ is an indicator function flagging errors in tupel $t \in T$

Attribute Accuracy

- covariance matrix Σ_{xx} or spur(Σ_{xx}) if T has metric data space dom_T

3.3.1 Accuracy Measure(s)

Example:

Top Scorer of German Bundesliga in the last 5 years				
Jahr	Name	Verein	Treffer	Nationalität
2003	Elber	FC Bayern	21	Brasilien
2003	Christiansen	VfL Bochum	21	Dänemark
2004	Ailton	Werder Bremen	2004	Braslien
2005	Mintál	1.FC Nürnberg	24	Slowakai
2006	Klose	FC Bayern	25	Polen
2007	Neumeister	VfL Bochum	20	Griechenland

Tupel Level: acc (T) = 2/6 * 100 ≈ 33 %

Value Level: $acc_v(T)$ =24/30 * 100 = 80%

3.3.2 Completeness

Objective: Indicator of coverage of set of objects

Two very different kinds of completeness:

- null values (missing values)
- missing entities or tupels („units" in Statistics)

DEF.: **Completeness**

Let O be an universe („population" or reference relation) and $O' \subseteq O$ a (not necessarily proper) subset of objects.
The coverage $com(O, O') = |O'|/|O| * 100$ is called completeness.

Synonyms: scope, extent

cf. Naumann (2002), Batini and Scannapieco (2006)

3.3.2 Completeness
of tuples, attributes, relations

- **Tuple completeness com_t**: Percentage of number of non-null values O´and all attribute values O of a tupel $t \in T$.

- **Attribute competeness com_A**: Percentage of number of non-null values O´and all tupels O given an attribute $a \in \boldsymbol{A}$ from schema of T.

- **Relation completeness com_T**: Percentage of number of non-null values O´in T and size of T (=#rows$_*$colums).

Note:

- We ignore „value completeness" as an indicator function whether or not a value is missing (null value).

- com_{Schema} can be defined in an analogue way

3.3.2 Completeness

Student

StudentID	Name	Surname	Vote	ExaminationDate
6754	Mike	Collins	29	07/17/2004
8907	Anne	Herbert	18	07/17/2004
6578	Julianne	Merrals	NULL	07/17/2004
0987	Robert	Archer	NULL	NULL
1243	Mark	Taylor	26	09/30/2004
2134	Bridget	Abbott	30	09/30/2004
6784	John	Miller	30	NULL
0098	Carl	Adams	25	09/30/2004
1111	John	Smith	28	09/30/2004
2564	Edward	Monroe	NULL	NULL
8976	Anthony	White	21	NULL
8973	Marianne	Collins	30	10/15/2004

$|T|=12$

$|\boldsymbol{A}|=5$

$size_T=60$

- Tupels: $com_{t1} = com_{t2} = 100\%$; $com_{t3} = 80\%$
- Attributes: $com_{Name} = 100\%$; $com_{ExamDate} \approx 33\%$
- Table: $com_{Student} = 100 * 53/60 \approx 88\%$

Source: Batini and Scannapieco (2006)

3.3.3 Time-Related Dimensions

Objective: Defining and measuring how up-to-date, stable, slowly or frequently changing data are.

- Currency (Promptness): Indicator of how promptly data are updated.
- Volatility (Valid Period, Change Frequency): Indicator for the length of time data remain valid, or of the frequency with which data vary over time.
- Timeliness (Freshness, Age): Indicator of how delayed, old or current data are at a user's disposal

Cf. Ballou et al (1998), Batini and Scannapieco (2006)

3.3.3 Time-Related Dimensions

Objective: Defining and measuring how up-to-date, stable, slowly or frequently changing data are.

- Currency (Promptness): Indicator of how promptly data are updated.
- DEF.: Currency =
  Age + Dwell_for_useTime = Age + ($Time_{Use}$ - $Time_{Disseminate}$)
- Timeliness (Freshness, Age): Indicator of how delayed, old or current data are at a user‵s disposal

Cf. Ballou et al (1998), Batini and Scannapieco (2006)

3.3.3 Time-Related Dimensions

Objective: Defining and measuring how up-to-date, stable, slowly or frequently changing data are.

- Currency (Promptness): Indicator of how promptly data are updated.
- Volatility (Valid Period, Change Frequency): Indicator for the length of time data remain valid, or of the frequency with which data vary over time.
- DEF.: Volatility = $time_{next\text{-}change} - time_{last\text{-}update}$

  Change Frequency = 1 / Volatility

Cf. Ballou et al (1998), Batini and Scannapieco (2006)

3.3.3 Time-Related Dimensions

Objective: Defining and measuring how up-to-date, stable, slowly or frequently changing data are.

- Currency (Promptness): Indicator of how promptly data are updated.
- Volatility (Valid Period, Change Frequency): Indicator for the length of time data remain valid, or of the frequency with which data vary over time.
- Timeliness (Freshness, Age): Indicator of how delayed, old or current data are at a user's disposal

DEF.: Timeliness = max{0, 1- currency / volatility}

bad Timeliness = 0; good Timeliness = 1

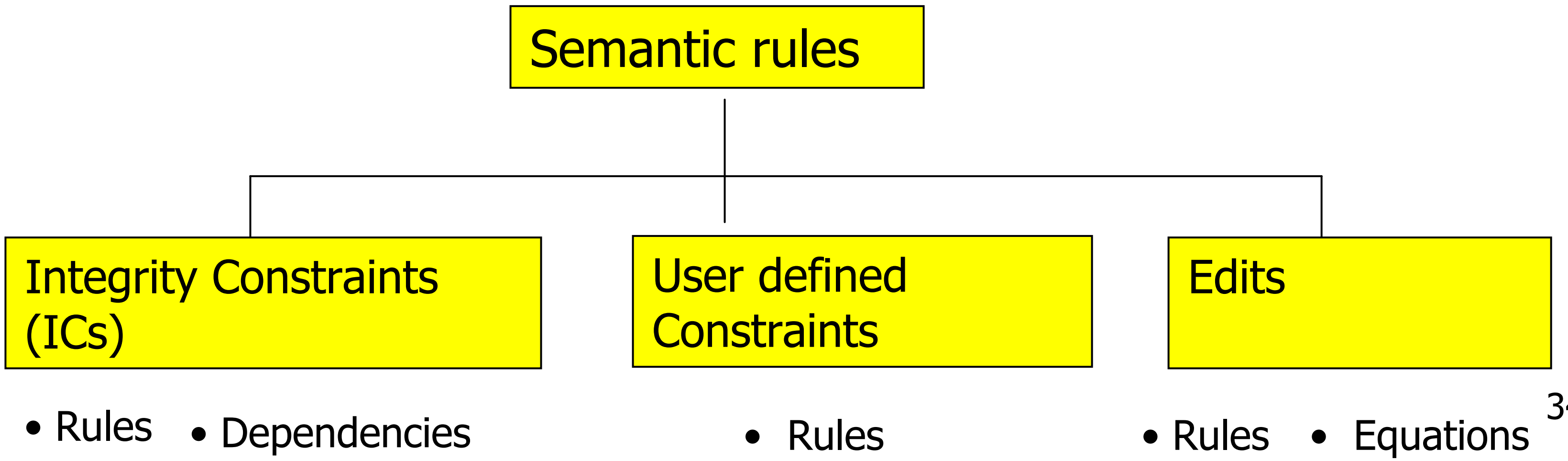
Cf. Ballou et al (1998), Batini and Scannapieco (2006)

3.3.3 Time-Related Dimensions (Example)

- Annual GDP computation; Time measured in months:
- Case A:
  - Currency = Age + ($Time_{Use}$ - $Time_{Disseminate}$) = 3 + (15th April - 1 April) = 3.5 month
  - Volatility = $time_{next\text{-}change}$ - $time_{last\text{-}update}$ = 6 months
  - Timeliness = max{0, 1- currency / volatility} = max{0, 1- 3.5/6} = 0.42 - quite bad !
- Case B:
  - Currency = 3,5
  - Volatility = 12
  - Timeliness = 0.7 - not bad !

3.3.4 Consistency

- **Objectivity**: Asserting semantic rules to be true with respect to a database (or file).
- DEF.: Consistency
  Semantic constraints are used to ensure coherency of data and application domain knowledge (metadata).
- Using a DBMS metadata can be used as part of a repository or triggers to prohibit a violation of the semantic rules.



3.3.4 Consistency
ICs

- IC activation events:
  - on insert
  - on delete
  - on update
- IC Types:
  - single attribute ICs (Domain constraints)
  - multi attribute Ics
  - cross-table constraints

3.3.4 Consistency

IC Types - Examples

single IC: $0 \leq Years \leq 60$

multi IC: if $Years \leq 3$ then $Salary \leq 35000$

Cross IC: $D.Budget \leq E.sum(Salary)$ where $E.Dep\# = D.Dep\#$

Employee (E)

Emp#	Dep#	Name	Years	Salary
001	A	Adam	2	3500[illegible]
005	C	Parker	17	4600[illegible]
007	A	Maier	31	52000

- IC Types:
  - single attribute ICs (Domain constraints)
  - Multi attribute ICs
  - Cross-table constraints

Department (D)

Dep#	Budget
A	470000
B	125000
C	360000

3.3.4 Consistency

Dependencies in DBs

- **Key Dependency**: A key dependency holds in relation (table) T if no two tupels t_1, $t_2 \in$ T have the same value of the primary key or a candiate key.

$$t_1.\text{key} = t_2.\text{key} \Rightarrow t_1 \equiv t_2 \text{ (duplicates not allowed)}$$

- **Inclusion Dependency**: Let T_1, T_2 be tables and **A, B** nonempty subsets of the corresponding attributes. Inclusion dependency holds if T_1.**A** is contained in T_1.**B** or, alternatively, T_2.**B.**

Ex.: Referential Integrity like $\exists$ Employee.Dep# $\Rightarrow$ Department.Dep#

- **Functional Dependency** (FD): Let **A, B** be nonempty subsets of the attributes in table T. FD is satisfied in T (**A -> B**) if for all tupels t_1, $t_2 \in$ T

$$t_1.\mathbf{A} = t_2.\mathbf{A} \Rightarrow t_1.\mathbf{B} = t_2.\mathbf{B}$$

- **Multivalued Dependency** (MVD): Let **A, B, C** be nonempty subsets of the attributes in table T. MVD is satisfied in T if **A ->> B** / **C** or, equivalently, conditional independence exists: **B** $\perp$ **C | A.**

Note: Simpson Paradox may happen if MVD is ignored!

Incorrect Dicing (Marginalisation)

if MVD C->> M|Y ignored

Proposition : Slice, dice and roll-up are incorrect if MVD constraints Z->> X|Y are not preserved

color	blue		white		blue		white		ALL
year	90	91	90	91	90	91	90	91	ALL
count	255	156	88	82	174	102	222	175	1254

Source: Gray et a. (1997)

$$p(\text{chevy}|\text{blue}, 90) \approx 59\%, \quad p(\text{chevy}|\text{blue}, 91) \approx 60\%$$

$$p(\text{chevy}|\text{white}, 90) \approx 28\%, \quad p(\text{chevy}|\text{white}, 91) \approx 32\%$$

$$p(\text{chevy}|90) \approx 46\%, \quad p(\text{chevy}|91) \approx 46\%$$

Those who ignore Statistics are condemned to reinvent it! (Source: Efron, Stanford Univ.)

3.3.4 Consistency
Edits

DEF.: Added-on Rules to ensure the semantic integrity (mostly) of files or part of a data entry system.

Synonym: Sematic integrity constraints

Main Types:

- simple edits
- logical edits
- numerical edits
- probabilistic edits
- statistical edits
- fuzzy edits

Source: Lenz, Köppen, Müller (2006)

Quality remains long after the price is forgotten.

H.G. Selfridge

4. References

1. Batini, C., Scannapieco, M.: Data Quality: Concepts, Methods and Techniques. Heidelberg: Springer Verlag (2006)

2. Dombrowski, Erik und Lechtenböger, Jens: Evaluation objektorientierter Ansätze zur Data-Warehouse-Modellierung, Datenbank-Spektrum 15/2005

3. Naumann, Felix: Datenqualität, Informatik-Spektrum_30_1_2007

4. Naumann, Felix: Quality-driven Query Answering for Integrated Information systems, LNCS 2261, Springer, Heidelberg et.c, 2002

5. www. Information-quality.com