UNIVERSITÄT LEIPZIG

Data Cleaning

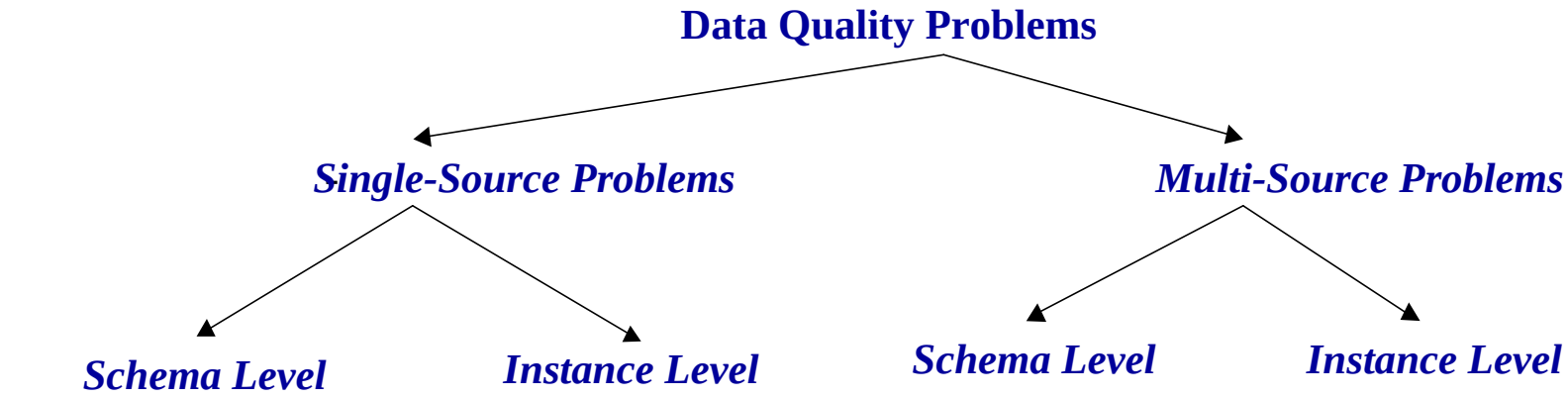
Problemseminar WS 2006/07

http://dbs.uni-leipzig.de

Data Cleaning („Datenreinigung“)

- Erkennung und Behebung von Inkonsistenzen, Fehlern, Informationsdefiziten in Daten
- Verbesserung der Datenqualität
- Sicherstellung von aussagekräftigen Datenanalysen („garbage in, garbage out“)
- Probleme innerhalb einzelner Datenquellen
- Probleme zwischen mehreren Datenquellen

Klassifikation Datenqualitätsprobleme*



(Lack of integrity constraints, poor schema design)

- Uniqueness
- Referential integrity
…

(Data entry errors)

- Misspellings
- *Redundancy/duplicates*
- Contradictory values
…

(Heterogeneous data models and schema designs)

- Naming conflicts
-Structural conflicts
…

(Overlapping, contradicting and inconsistent data)

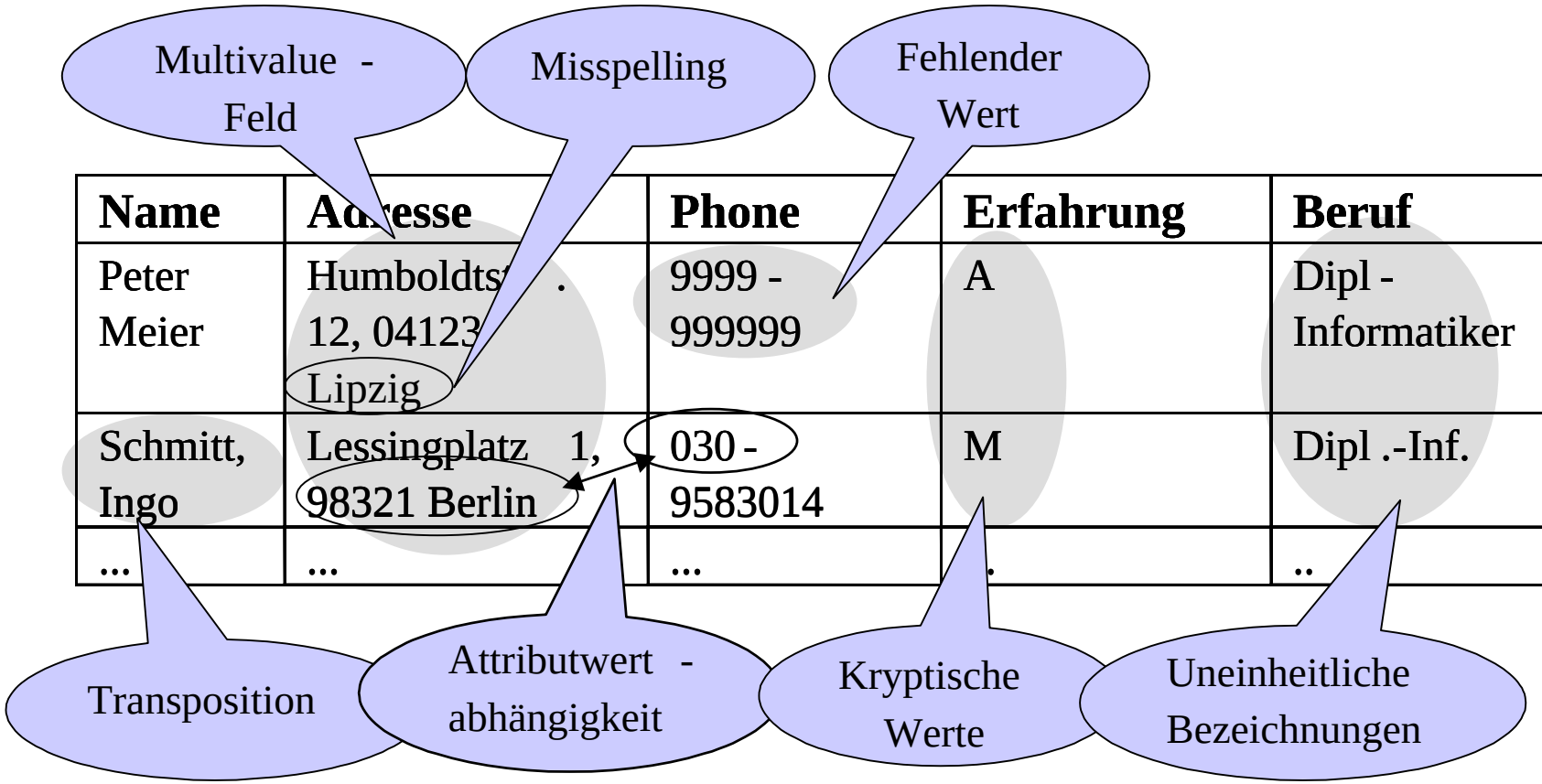
- Inconsistent aggregating
- Inconsistent timing
…

Object matching
(entity resolution, duplicate detection, reference reconcilation …)

* E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bull. Data Eng., Dec. 2000

Single-Source Probleme

- Ursachen:
  - Eingabefehler
  - Fehlen von Schemata (z.B. bei dateibasierter Datenverwaltung)
  - Unzureichende Einhaltung von Integritäts-Constraints
  - Unterschiedlichen Änderungsstände



Multi-Source-Dateninkonsistenzen

Customer (source 1)

CID	*Name*	*Street*	*City*	*Sex*
11	Kristen Smith	2 Hurley Pl	South Fork, MN 48503	0
24	Christian Smith	Hurley St 2	S Fork MN	1

Client (source 2)

Cno	*LastName*	*FirstName*	*Gender*	*Address*	*Phone/Fax*
24	Smith	Christoph	M	23 Harley St, Chicago IL, 60633-2394	333-222-6542 / 333-222-6599
493	Smith	Kris L.	F	2 Hurley Place, South Fork MN, 48503-5998	444-555-6666

Duplikate in (integrierten) Web-Datenquellen

A **survey** of approaches to automatic schema matching - group of 27 »
E **Rahm**, PA **Bernstein** - The VLDB Journal The International Journal on Very Large ..., 2001 - Springer
... E. **Rahm**, PA **Bernstein**: A **survey** of approaches to automatic schema matching 335 ... 336
E. **Rahm**, PA **Bernstein**: A **survey** of approaches to automatic schema matching ...
Cited by 616 - Related Articles - Web Search - Import into BibTeX

[CITATION] A **survey** of approaches to automatic schema matching
PA **Bernstein**, E **Rahm** - VLDB Journal, 2001
Cited by 14 - Related Articles - Web Search - Import into BibTeX

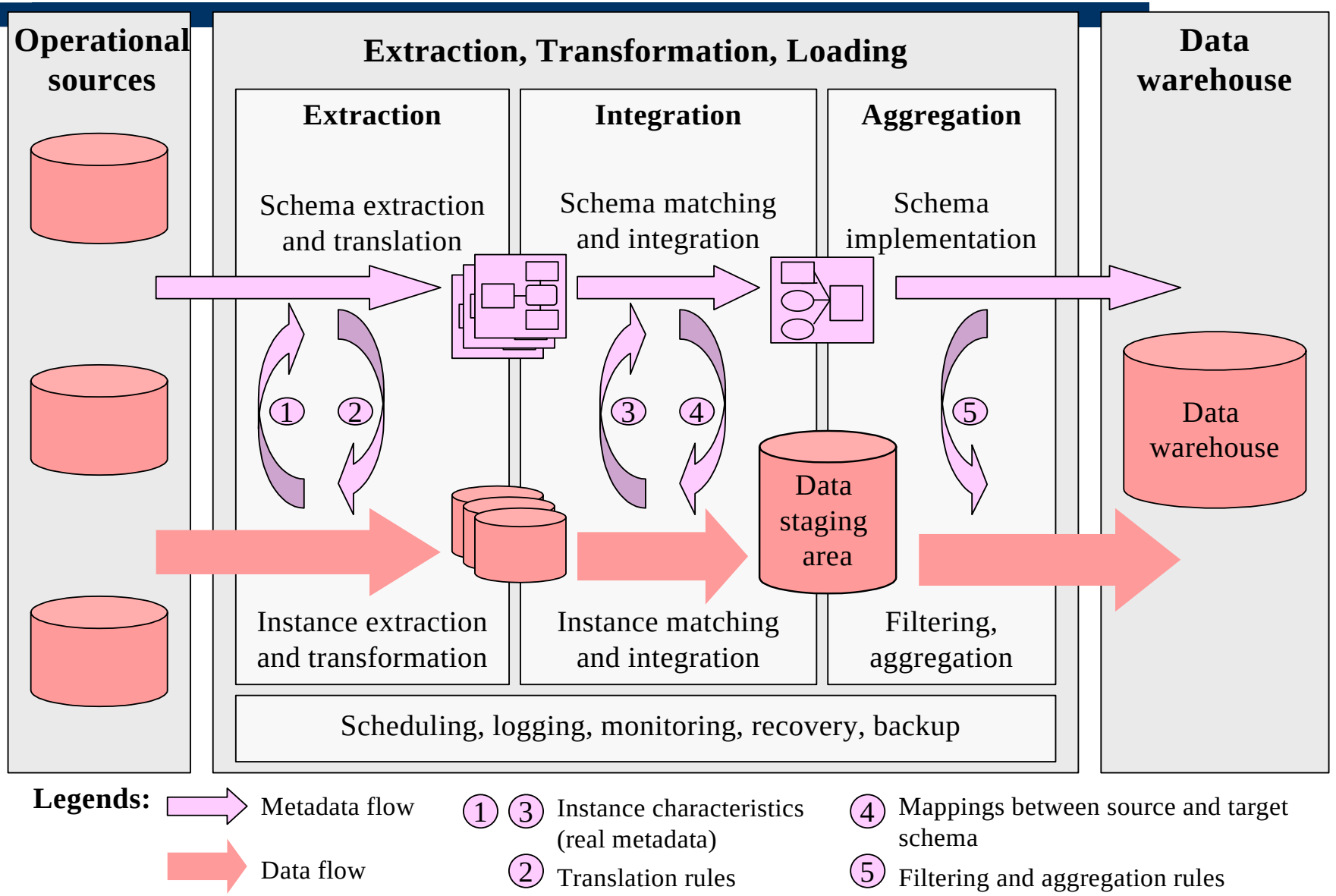
[CITATION] A **survey** of approaches to automatic schema mapping
E **Rahm**, PA **Bernstein** - The VLDB Journal, 2001
Cited by 5 - Related Articles - Web Search - Import into BibTeX

[CITATION] A (2001) A **survey** of approaches to automatic schema matching
E **Rahm**, PA **Bernstein** - The International Journal on Very Large Data Bases (VLDB)
Cited by 1 - Web Search - Import into BibTeX

Duplicates due to
- Order of authors
- Different titles
- Extraction error (title)
- Typos ...

http://scholar.google.com

Data Warehouses: ETL-Prozess*



* E. Rahm, H. H. Do: *Data Cleaning: Problems and Current Approaches.* IEEE Techn. Bull. Data Eng., Dec. 2000

Related Work

- Surveys
  - Rahm, Do: *Data Cleaning: Problems and Current Approaches*. IEEE Techn. Bull. Data Eng., 2000
  - Gu, Baxter, Vickers, Rainsford. *Record Linkage: Current Practice and Future Directions*. Technical Report, 2003
- Frameworks (new operators for data cleaning, user-controlled workflows)
  - AJAX (Galhardas et al., VLDB 2001)
  - Potter‘s Wheel (Raman et al., VLDB 2001)
  - TAILOR (Elfeky et al., Data Eng. 2002)
- Tools
  - DataCleanser (EDD), Merge/Purge Library (Sagent/QM Software), MasterMerge (Pitnew Bowes) ...
  - MS SQL Server 2005: Data Cleaning Operators (Fuzzy Join / Lookup)

Related Work (2)

- Attribute Similarity
  - String distance metrics; Edit Distance, Jaro-Winkler, TFIDF, SoftTFIDF, ...
  - Comparison → Cohen (KDD03-Workshop on Data Cleaning ...)
  - Learnable string distance metrics: Bilenko et al. (KDD, 2003)
- Manually specified combination of matchers
  - Rules : Hernandez et al. (SIGMOD 1995)
  - Constraints: Shen et al. (AAAI 2005) („age 2 cannot match with salary 200K“)
- Adaptive combination of matchers
  - Active Atlas (Tejada et al., Information Systems 2001)
    - Combination of multiple similarity scores for object pairs
    - Interactive decision tree learning to identify most informative example for the user to classify next
  - Multiple profilers: Doan et.al. (IIWeb, 2003)
- Context-based object matchers
  - Co-authors: Bhattacharya, Getoor (DMKD 2004)
  - Warehouse hierarchies: Ananthakrishna et.al. (VDLB 2002)
  - XML hierarchies: Weis et al. (SIGMOD 2005)
  - XML graphs: Dong et al. (SIGMOD 2005)

Publication Categorizer on Data Cleaning

Research Area

- Data cleaning (94)
  - Duplicate/matching (59)
    - Similarity functions (13)
  - Evaluation/benchmark (6)
    - Synthetic datasets (1)
  - Data analysis/outliers (6)
  - Self-Tuning (7)
  - Applications (2)

#datasets n

- centralized (n=1) (3)
- distributed (n>1) (10)
  - Warehousing /ETL (5)
  - virtual / mediators (1)
  - peer-to-peer (3)

Data type

- relational (9)
- XML (4)

Paper type

- Survey (5)
- Framework (2)
- Tool / product (19)

Home » FlexiLists

Authors, weighted by occurrence count

Allester Ananthakrishna Arenas Aumueller Bai Baxter Benedikt Benjelloun Bhamidipaty Bhattacharya Bilenko Bilke Bitton Bohannon Bohlen Brosy Bruns Candan Catarci **Chaudhuri** Chen Chen Chiang Chua **Cohen** Dai Dayal **Do** Doan Domingos Dong Elfeky Elkan Elmagarmid Fan Fienberg Florescu Fox Freytag Frigui Galhardas Ganesh Ganjam **Ganti** Getoor Golovin Gu Guha Halevy Han Han Hellerstein Hernandez Hirsh Hjaltason Hsu Inc Jeh John Jokinen Jurk Kailing Kalashnikov Kang Kapoor Kaushik Kautz Kementsietsidis Kintigh Kirsten Knoblock Kohavi Kothari Kotidis **Koudas** Kriegel Lee Lee Lee Lenz Leser Li Licamele Lim Liu Lu Madhavan Maletic Marathe Marcus Marian Massmann Mazeika Mead Mehrotra Menestrina Milano Miller Minton Mitra Molina Monge Mooney Motwani Moustakides Müller Narasayya **Naumann** Neiling Nelder Ooi Pinheiro Princeton Qi Quass **Rahm** Rainsford Ram Raman Ravikumar Richardson Richman Ristad Rosenthal Saita Samet Sapino Sarawagi Sayyadian Scannapieco Schonauer Seidl Shasha Shen Simon Singla Sirvastava **Srivastava** Starkey Stolfo Stöhr Sun Tejada **Thor** Ukkonen Vassilakis Venkatasubramanian Verykios Vickers Wang Weis Widom Winkler Witt Xi Yan Yan Yianilos Zhang Zhao Zhuang

http://dc-pubs.dbs.uni-leipzig.de/

Seminarbedingungen

Seminarziele

- Beschäftigung mit einem praxis- und wissenschaftlich relevanten Thema
- Erarbeitung und Durchführung eines Vortrags zu einem Thema unter Verwendung wissenschaftlicher (englischer) Literatur
- Diskussion
- Schriftliche Ausarbeitung zu dem Thema
- Hilfe und Feedback durch Betreuer / Seminarteilnehmer

- Bedingungen für Scheinvergabe / Prüfungsleistungsnachweis
  - Selbständiger Vortrag mit Diskussion
  - Schriftliche Ausarbeitung (ca. 15-20 Seiten)
    - Ausarbeitung vom Betreuer abzunehmen
    - Ausarbeitung soll zum Vortragstermin vorliegen
  - (möglichst aktive) Teilnahme an allen Vortragsterminen

Themen

Thema	Bearbeiter	Betreuer	Termin
1. Data Cleaning – ein Überblick		Sosna	9.1.07
2. Normalisierung von Daten		Aumüller	9.1.07
3. Outlier-Analyse		Hartung	16.1.
4. Ähnlichkeitsanalyse zur Duplikatsanalyse		Kirsten	16.1
5. Objekt-Matching in relationalen Daten		Köpcke	23.1.
6. Objekt-Matching in hierarchischen Daten		Thor	23.1
7. Data-Cleaning-Frameworks		Massmann	30.1.
8. Data-Cleaning-Unterstützung in kommerziellen Produkten		Weikum	30.1.
9. Bereinigung v. Web-Daten		Aumüller	23.1.
10. Evaluierung / Benchmark		Köpcke	30.1.