



**Hasso
Plattner
Institut**

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

16.4.2008

Felix Naumann

Das Motto

2

... Wenn wir uns klarmachen, dass der Kampf gegen Chaos, Durcheinander, und unbeherrschte Kompliziertheit eine der größten Herausforderungen der Informatik ist, müssen wir zugestehen:

„Beauty is our Business“.

Edsger W. Dijkstra, 1978

Auch:

“Computer Science is no more about computers than astronomy is about telescopes.”

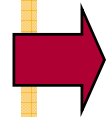
Motivation

3

- Vorbereitung auf das Master Studium
 - Wissenschaftlich Arbeiten
- Vorbereitung auf den Beruf
 - Ideen „verkaufen“
- Interesse an den Themen

Überblick

4



- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen



5

Felix Naumann

- naumann@hpi.uni-potsdam.de
- Raum A-1.13 (über Frau Pamperin)
- HU, IBM Almaden, DFG/JP/HU, HPI

Universität Potsdam, Hasso-Plattner-Institut

- Fachgebiet „Informationssysteme“

Forschungsthemen (gleich mehr):

- Datenqualität
- Informationsintegration
- Peer Data Management
- Search
- <http://www.hpi.uni-potsdam.de/~naumann/>

Arbeitsgruppe Informationssysteme

6

project **ViQTOR**



Paul Führung



Patricia Hobro

DQ Assessment



Prof. Felix Naumann

Data Fusion



Jens Bleiholder

project **fusem**



Karsten Draba

Information Integration

project **HumMer**



Data Cleaning

Information Quality

Duplicate Detection

Peer Data Management Systems

Melanie Weis & Sascha Szott



Armin Roth

Matching

Data Integration for Life Science Data Sources

Service-Oriented Systems

project **XClean**

project **System P**



Mohammed AbuJarour

Ontologies



Frank Kaufer

project **Aladin**



Jana Bauckmann



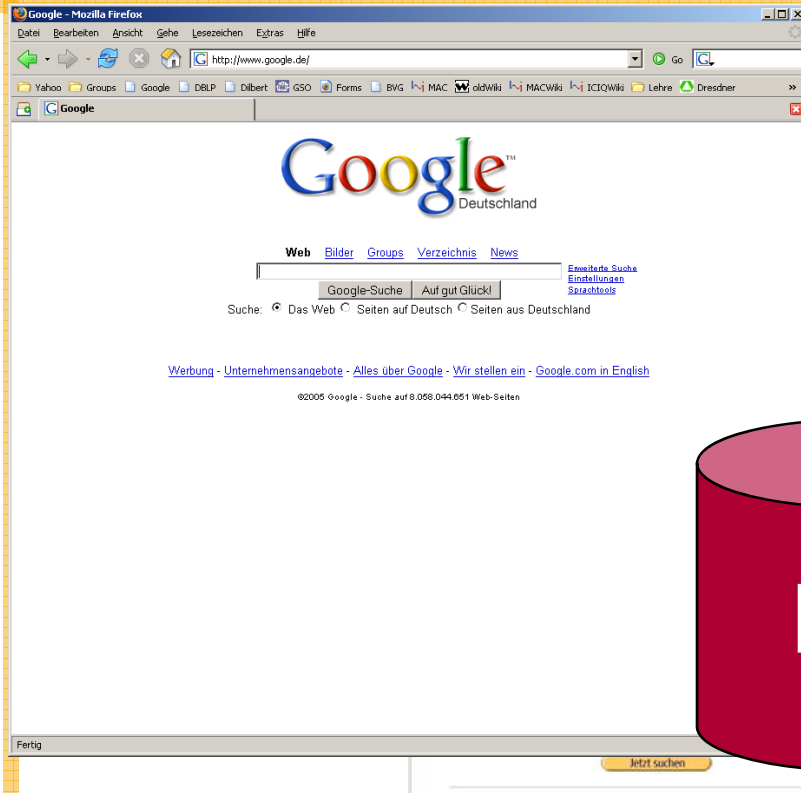
Alexander Albrecht

Personal Information Management

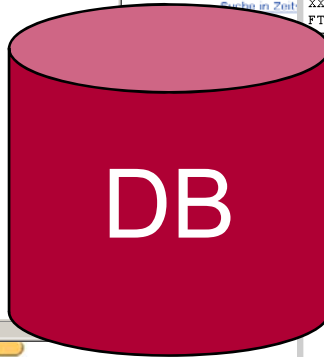
Data Profiling for Schema Management

Was sind Informationssysteme?

7



ID	RNGTPCHI	standard; RNA; ROD; 1016 BP.	Molecule type
DT	01-AUG-1991	(Rel. 28, Created)	Name
DT	04-MAR-2000	(Rel. 63, Last updated, Version 2)	Date of creation and last update
DE	Rat GTP cyclohydrolase I mRNA, complete cds.		Free text description
KW	GTP cyclohydrolase I.		Keywords describing the molecule
OS	Rattus norvegicus (Norway rat)		Organism
OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.		
RA	Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;		Article the sequence was published in
RT	"Cloning and sequencing of cDNA encoding rat GTP cyclohydrolase I: The first enzyme of the tetrahydrobiopterin biosynthetic pathway";		
RL	J. Biol. Chem. 266(2):765-769(1991).		
CDS	128..853		Structural annotation (coding sequence)
	/codon_start=1		Link to functional annotation of resulting protein
	/db_xref="GO:P22288"		
	/db_xref="SWISS-PROT:P22288"		
	/EC_number="3.5.4.16"		
	/gene="GTP cyclohydrolase I"		
	/product="GTP cyclohydrolase I"		
	/protein_id="AAA41299.1"		
	/translation="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGQPADAWK AGRPSEEDNENLNLPLNLAAYSSILRSLGDPQRQGLLKPURAAATMQFFTKGYQETI SDVLND&IFDEHDENVIVKIDIMFSMCEHHLVFPVGRVHIGYLPNKQVLGSLK&ARIV EIYSRRLQVQERLTKQIAVAITEALQPA&GVGVVIEATHMCMVMR&GVQKMN&SKVT&STML GVFREDPKTREF<LIRS"		Translated protein sequence
	Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other:		
	gacttcgaac ctcattcggc gcagaactcc tgcctccggtg acagccacag gtcacggccg	60	Sequence of bases
	cgggctaagc cgagccgcag cgcttggttag caacttaggg tgcctcggga gcaatcgcc	120	
	cgggtccatg gagaagccgc ggggtgtaag gtcaccaat ggggtccccc agcgggagct	180	
	...		
	catcaggagc tgaactccg tgtgcgagcc cgggtttgca gacccccgt gaggccagcg	900	
	ttatctgtct cgattgtaca ttccagttcc agttggata ctgtccaact ttattctca	960	
	ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caactt	1016	

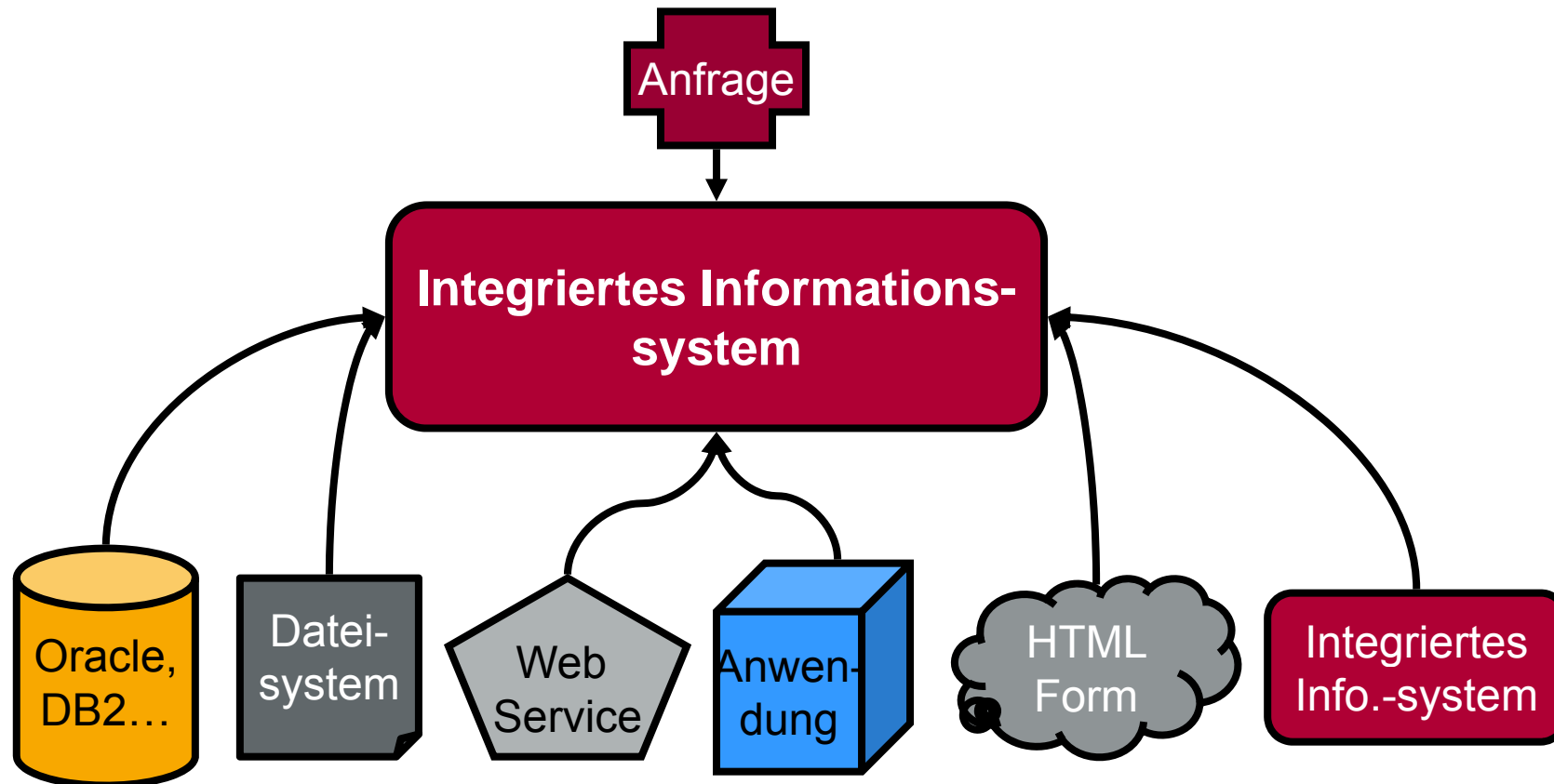


```
<buch>
  <isbn>0-201-318051</isbn>
  <titel>XML und Datenbanken</titel>
  <autor>Klettke/Meyer</autor>
</buch>
```



Integrierte Informationssysteme

8



Schematische und Daten-Heterogenität

9

Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen	
Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen			
Vorname	Nachname	Männl.	Weibl.
Felix	Naumann	Ja	Nein
Jens	Bleiholder	Ja	Nein
Melanie	Weis	Nein	Ja
Jana	Bauckmann	Nein	Ja

Variante 3

Personen		
Vorname	Nachname	Geschlecht
Felix	Naumann	Männlich
Jens	Bleiholder	Männlich
Melanie	Weis	Weiblich
Jana	Bauckmann	Weiblich

Schematische und Daten-Heterogenität

10

Variante 1

Männer	
Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen

Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

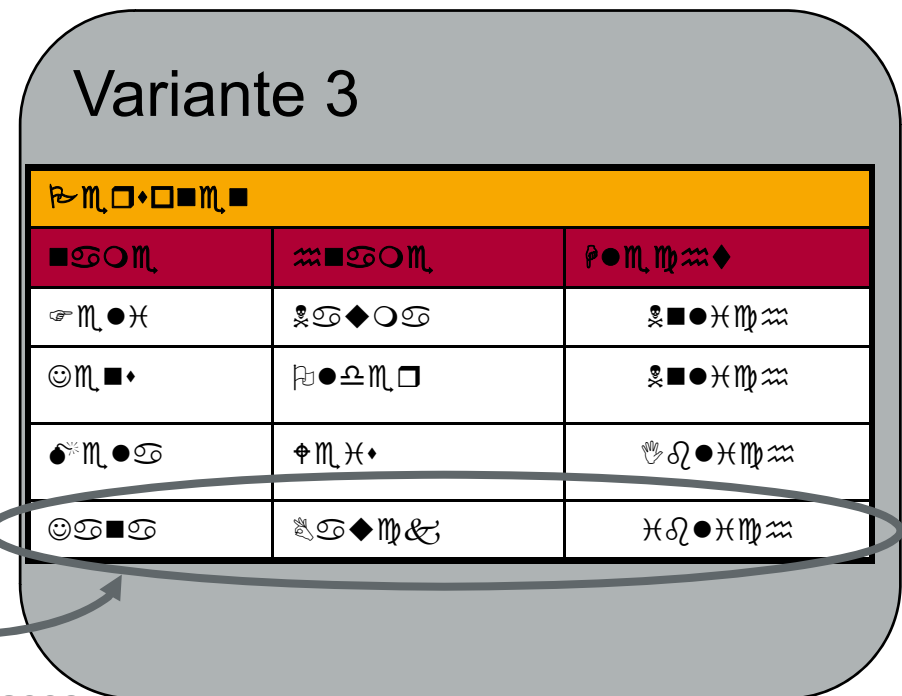
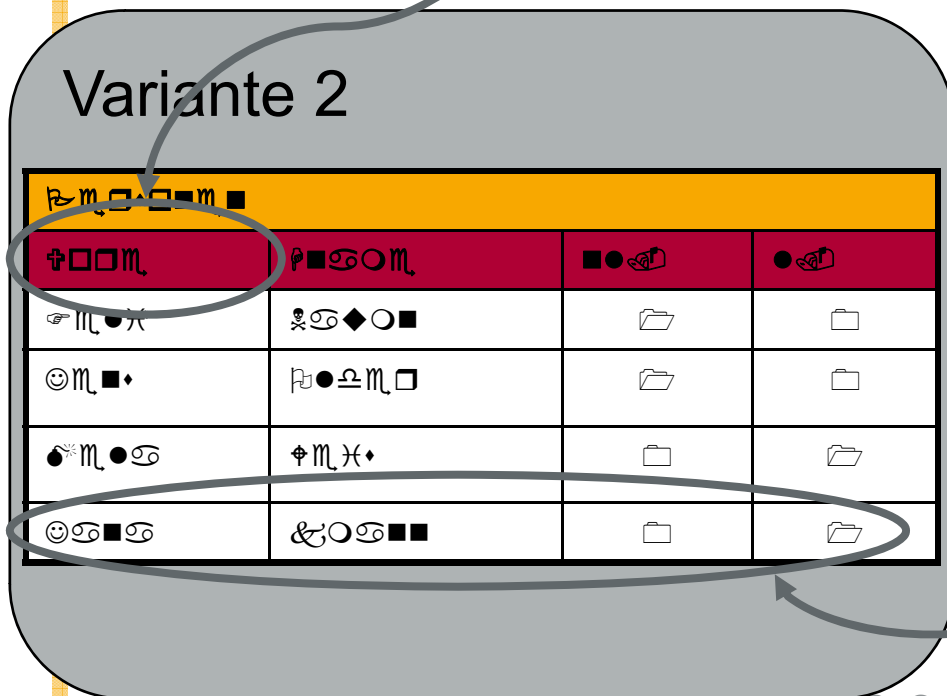
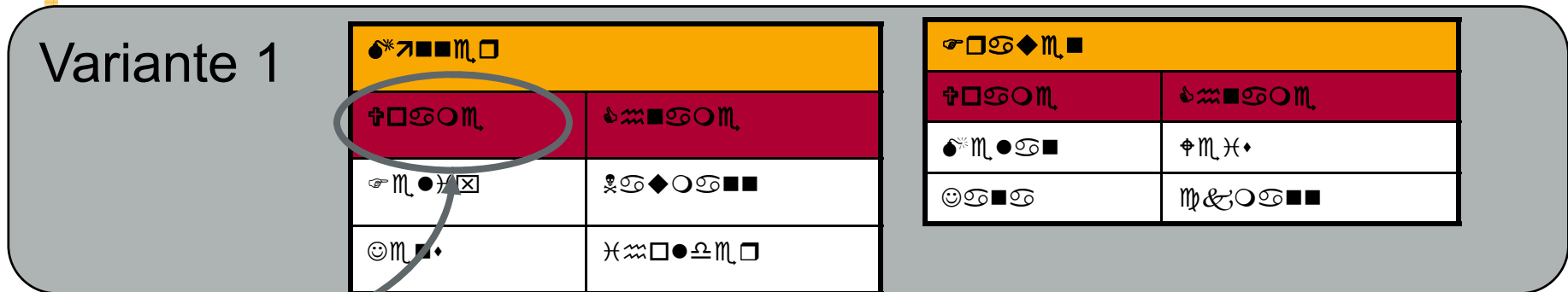
Personen			
FirstNa	Name	male	femal
Felix	Naumann	Ja	Nein
Jnes	Bleiho.	Ja	Nein
Melanie	Weiß	Nein	Ja
Jana	baukman	Nein	Ja

Variante 3

Personen		
VN	NN	SEX
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Schematische und Daten-Heterogenität

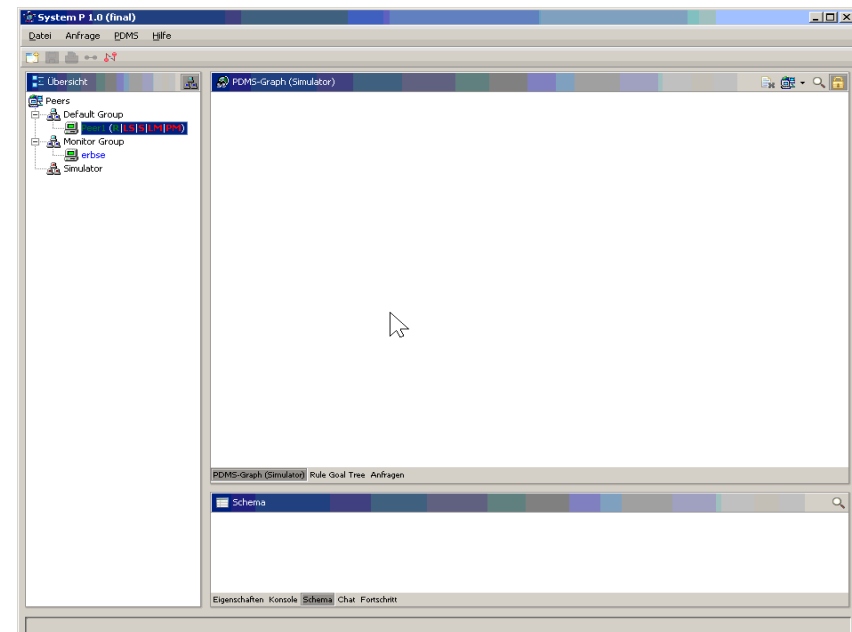
11



Weitere Forschungsthemen

12

- Informationsintegra
 - ◇ Schema Matchii
 - ◇ Duplikaterkennung
 - ◇ Datenfusion
- Datenqualität
- Peer Data Management
- Life Sciences: Aladin
- Search



Lehrveranstaltungen in diesem Semester

13

Vorlesungen

- DBS II
- Informationsintegration

Seminare

- Bachelor: Beauty is our Business
- Bachelor: www.ligageschichte.de
- Master: Duplikaterkennung
- Forschungsseminar



Extending the Database Relational Model to Capture More Meaning

E. F. CODD
IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

- (1) the search for meaningful units that are as small as possible—atomic semantics;
- (2) the search for meaningful units that are larger than the usual n -ary relation—molecular semantics.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, concept model, conceptual schema, entity model

CR Categories: 3.70, 3.73, 4.22, 4.28, 4.35, 4.34, 4.39

1. INTRODUCTION

The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the cluster of storage representation details. This implementation independence coupled with the power of the algebraic operators on n -ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37].

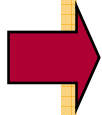
During the last few years numerous investigations have been aimed at capturing permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.
Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.
© 1979 ACM 0362-5915/79/1390-0387 \$00.75.

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 397–424.

Überblick

14



- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen



Vorstellung Teilnehmer

15

- Felix Elliger
- Alexander Kuscher
- Fabian Lindenberg
- Jan Oberst
- Stefanie Reinicke
- Stefan Richter

Seminarleistungen

16

Lesen

- Paper lesen und verstehen
- Verwandte Literatur lesen und verstehen
- Mindestens eine individuelle Besprechung mit mir
- Kurzvorstellung der Literatur (je 2 min)

1/5 Note

Vortragen

- Mindestens eine Folien-Besprechung mit mir
 - Spätestens 1 Woche vor Vortrag
- 30 min. Vortrag am jeweiligen Termin
 - + 15 min Diskussion
- Aktive Teilnahme an anderen Vorträgen
 - Jeweils ein *designated diabolus*

2/5 Note

Ausarbeitung

- Mindestens eine Gliederungs-Besprechung mit mir
 - Spätestens 2 Wochen vor Abgabetermin
- 8-10-seitige Ausarbeitung bis zum 15.8.
- Unter Verwendung der LaTeX-Vorlage im WWW

2/5 Note

Aktive Teilnahme an sämtlichen gemeinsamen Terminen

Feedback

17

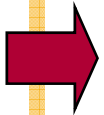
Fragen bitte jederzeit!

- Während des Seminars
- Während individueller Besprechungen
 - Termin bei Pat Hobro ausmachen
 - ◇ brigitte.hobro@hpi.uni-potsdam.de
 - ◇ 0331 / 5509 280
 - Muss nicht zur Sprechstunde sein!
- Sprechstunde
 - Dienstags 15:00 – 16:00
 - Raum A-1.13
 - Am liebsten mit Anmeldung
- Email: naumann@hpi.uni-potsdam.de

Überblick

18

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen



Wissenschaftliche Texte lesen

19

- Fachartikel
 - Journale, Konferenzen und Workshops
 - Entstehung:
 - ◇ Forschungsvorhaben
 - ◇ Begutachtungsprozess
- Struktur eines Artikels
 - Kritisches Lesen
 - Experimente
- Literatur
 - (Online-) Recherche
- Englisch

Wissenschaftlichen Vortrag halten

20

- Gliederung
 - Die Kunst des Weglassens
- Foliengestaltung
 - Powerpoint
 - Overhead
- Zeit einhalten
- Techniken zur Vorbereitung
- Techniken während des Vortrags

Ausarbeitung schreiben

21

- Gliederung
- Schreibstil
 - Objektives
 - Subjektives
- Plagiate
- LaTeX

Zeitlicher Ablauf – Termine

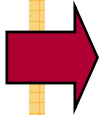
22

- Part 1: Lesen
 - Termin zur Literaturkritik
 - *Elevatorpitch*
- Part 2: Vortragen (Zwei pro Termin)
 - Jeweils 30 Minuten Vortrag + Diskussion
- Part 3: Schreiben
 - Ausarbeitung / Seminararbeit
- Vortragsreihenfolge
 - Mariposa, Trio, Fagin, Stop, SNM, Apriori
- Siehe auch Web-Page
- Dazwischen Vorträge von mir
 - Lesen wissenschaftlicher Texte
 - Vortragstechniken
 - LaTeX (bei Bedarf)
 - Tipps zur Gliederung und Ausarbeitung

Überblick

23

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen



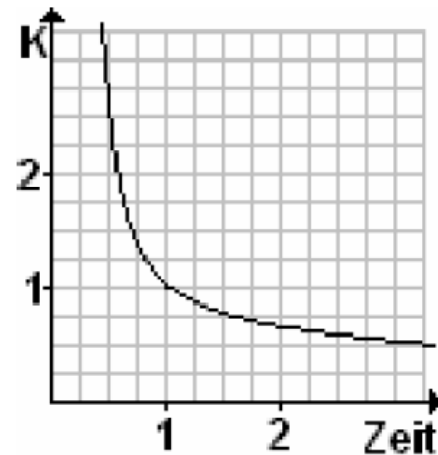
Allgemeines

24

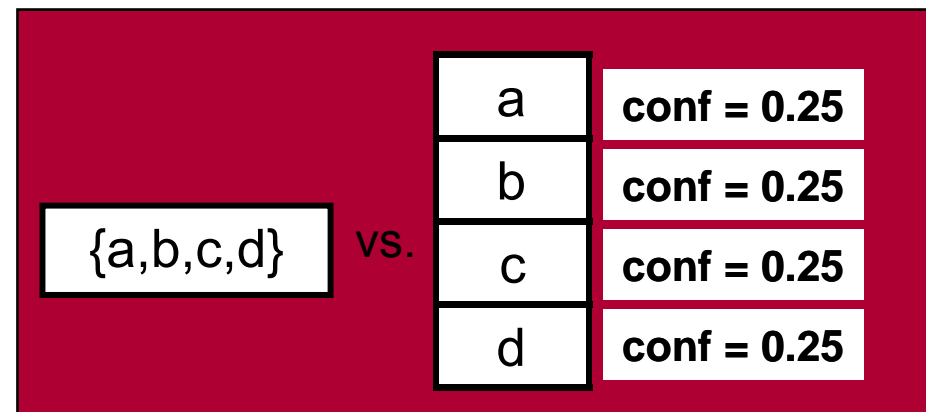
- Themen aus dem Umfeld
 - Datenbanken (DB)
 - Informationsintegration (II)
 - Business Intelligence (BI)
- Sehr gute, oft wegweisende paper
 - Jeweils mindestens eine wirklich gute Idee
 - Gut geschrieben

 - Dies sind nicht die „besten Paper der Informatik“.

- Neuartige Architektur für ein weit verteiltes Informationssystem
- Mikroökonomisches Prinzip zur Anfrageoptimierung
 - Datenhaltung
 - Anfragebearbeitung

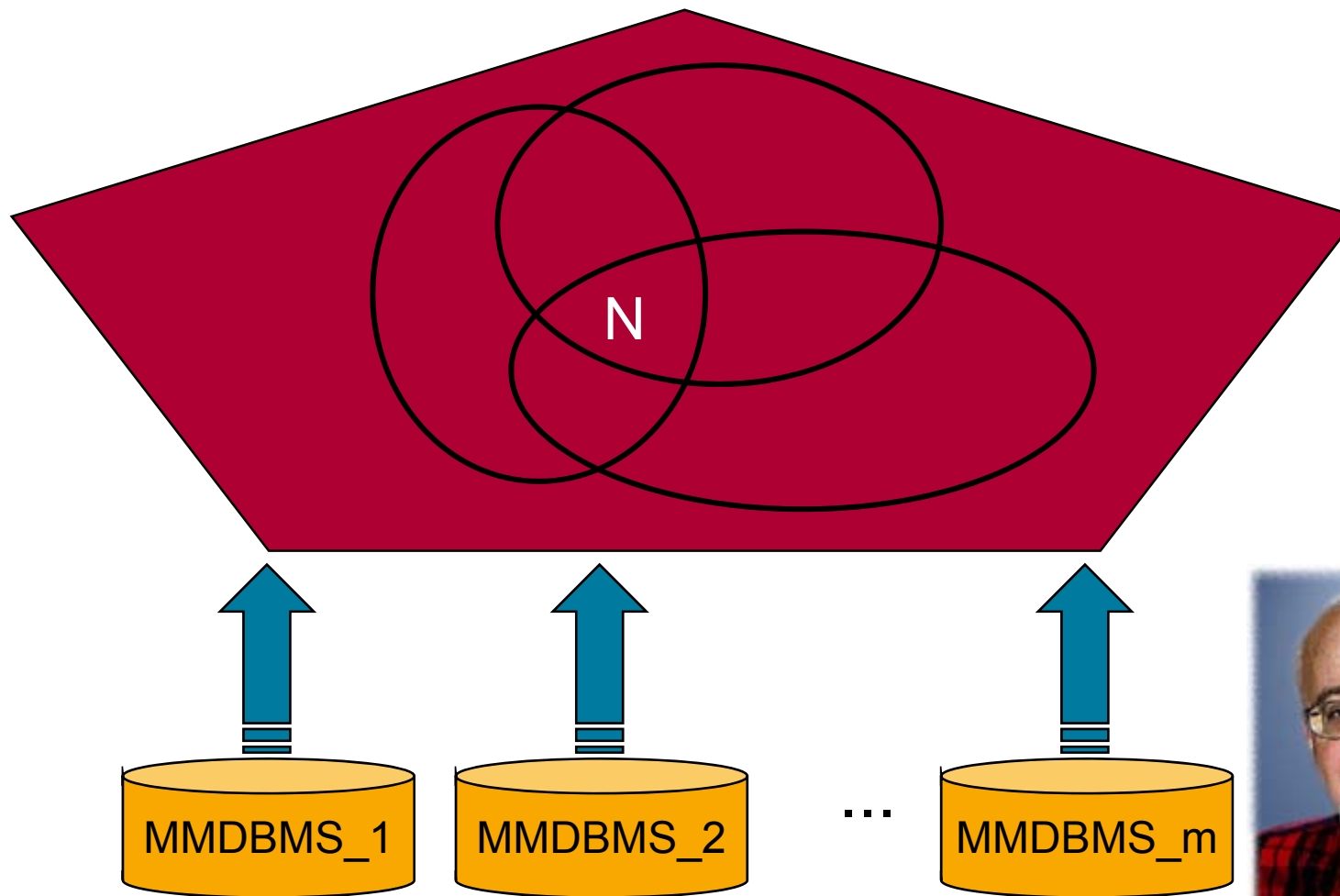


- DBMS Grundannahmen werden aufgehoben:
 - Jeder Datenwert ist entweder in der DB gespeichert oder nicht.
 - Der Datenwert ist absolut.
 - Seine Herkunft ist irrelevant
- Beispiel: Christmas Bird Count
- Trio relaxiert diese Annahmen:
 - Daten
 - Genauigkeit auf Attributebene, Tupelebene und Relationenebene
 - Herkunft
- Anfragen alle drei Aspekte
 - Schwierige Semantik



Fagins Algorithmus

27

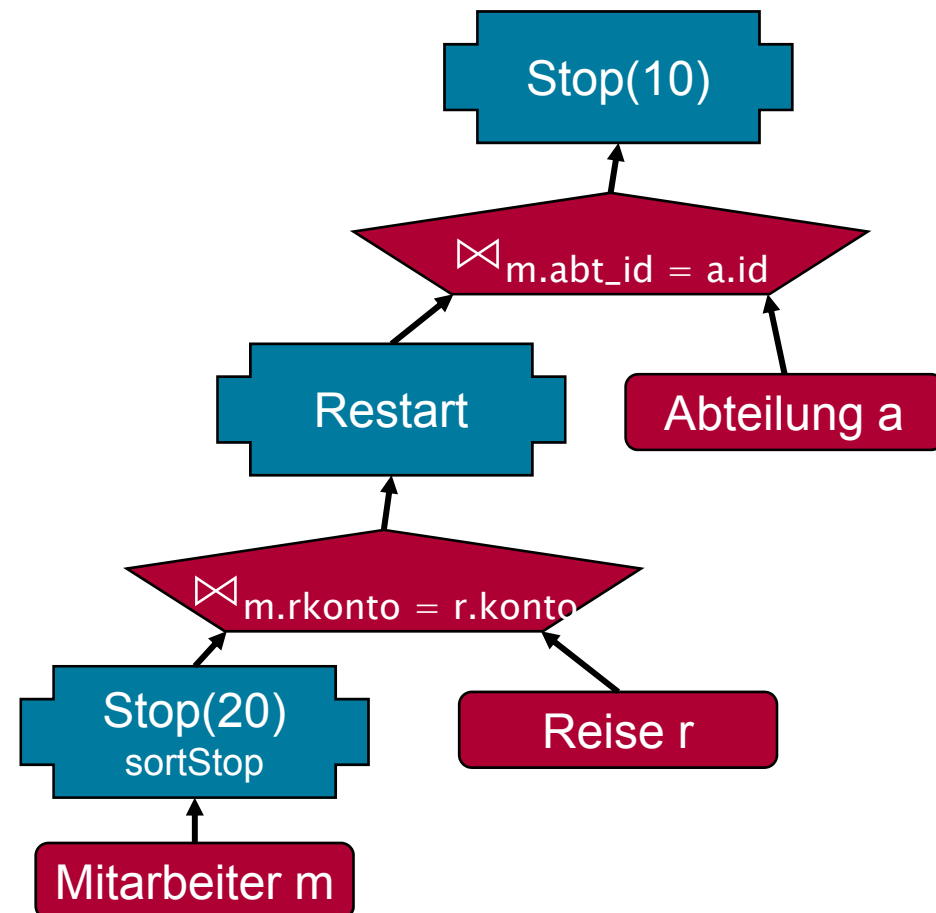


„Enough Already“ in SQL

28

- First-N und Top-N Techniken in SQL

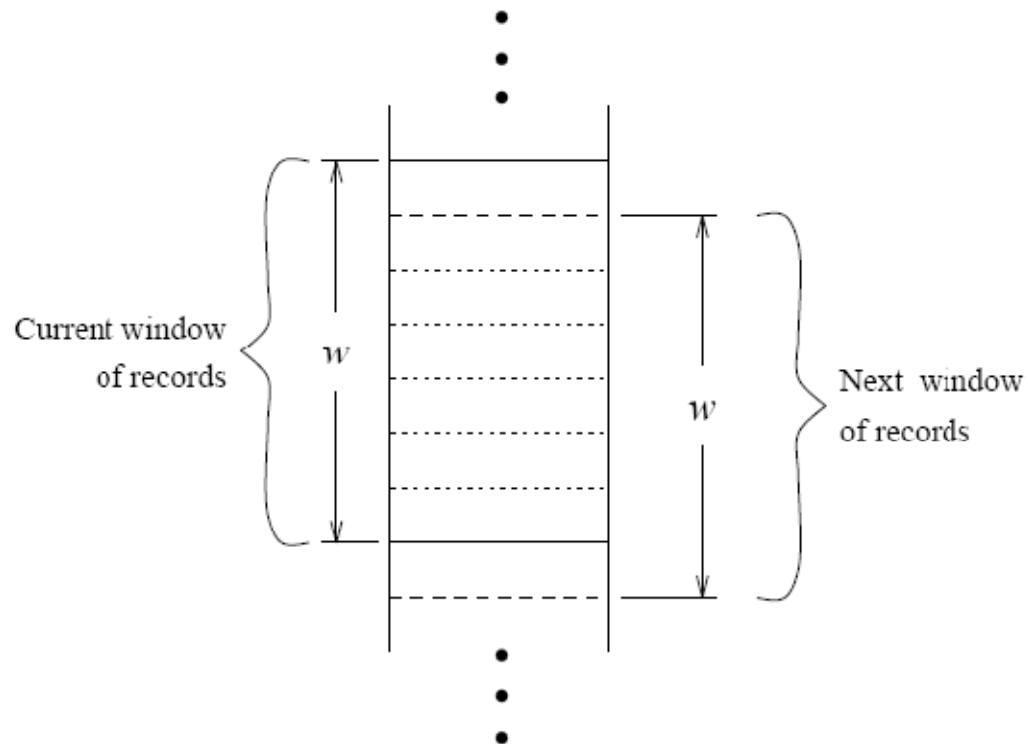
- Syntax & Semantik
- Neue Operatoren
- Optimierung
- Evaluation



Sorted Neighborhood

29

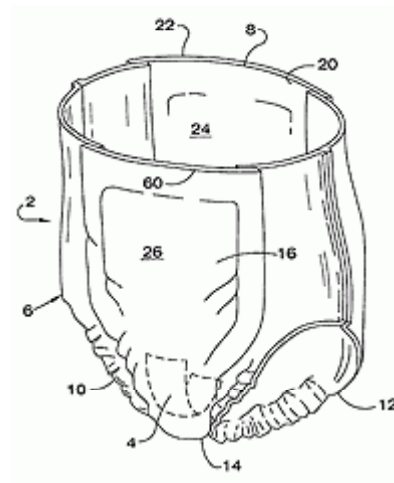
- Datenreinigung und Duplikaterkennung
 - Effizienz
 - Effektivität



Data Mining

30

- Bahnbrechendes Papier
- Setzte intensive Forschung zu Data Mining in Gang
- Zwei Algorithmen zur schnellen Entdeckung von Assoziationsregeln
 - Apriori
 - AprioriTid



Nächste Schritte

31

- Sie: E-Mail an mich
 - Mit Wunschliste (Platz 1 – 3)
 - ◇ Daraus: Themenvergabe
 - ◇ Daraus: Vergabe des Diskutantenplatzes
 - Bis Mitternacht
- Ich: Themenvergabe bald
 - E-Mail Benachrichtigung
- Ich: Vortrag zum „Lesen wissenschaftlicher Texte“