



**Hasso  
Plattner  
Institut**

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

19.4.2010

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](mailto: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

- <http://www.hpi.uni-potsdam.de/~naumann/>

# Information Systems Team

6

project **ViQTOR**



Paul Führung



Katrin Heinrich

**DQ Annotation & Assessment**



Prof. Felix Naumann

**Data Fusion**



Silke Trißl

**Duplicate Detection**



Uwe Draisbach



Dustin Lange

**Schufa**

**Entity Search**

**Information Integration**

project **HumMer**



Johannes Lorey

**IBM**

**Data Profiling & Cleaning**

**Information Quality**



Armin Roth

project **System P**

**Peer Data Management Systems**

**Data Profiling**

**Data Integration for Life Science Data Sources**



Christoph Böhm

**ETL Management**

**Matching**

project **PoSR**

**Service-Oriented Systems**

project **Aladin**

**Ontologies, Profiling**



Tobias Vogel

**Forschungskolleg**



Mohammed AbuJarour



Frank Kaufer



Jana Bauckmann

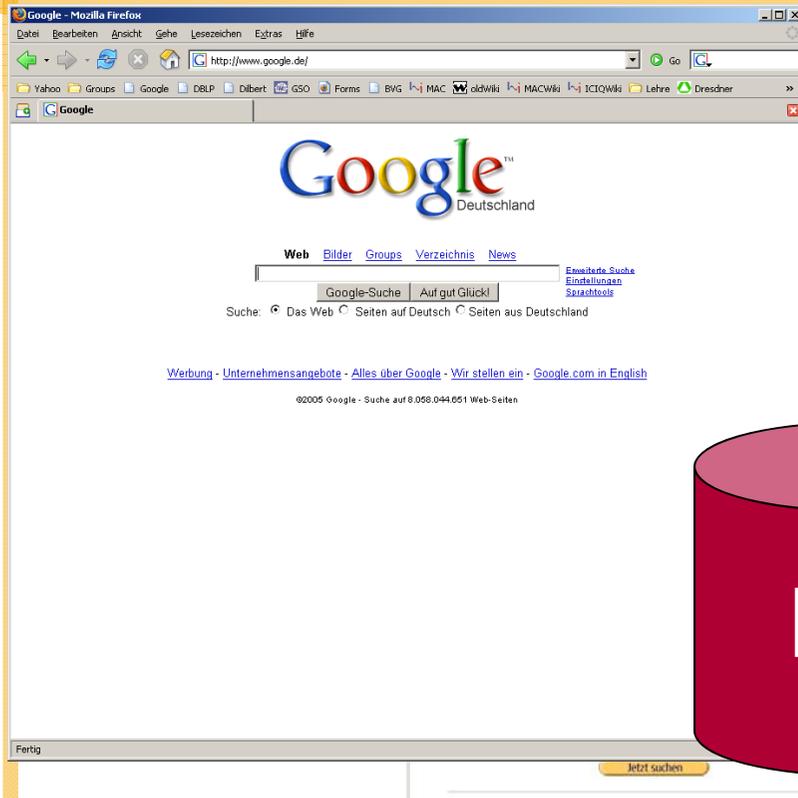


Alexander Albrecht

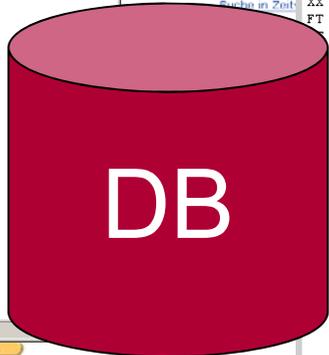
**Data Profiling for Schema Management**

# Was sind Informationssysteme?

7

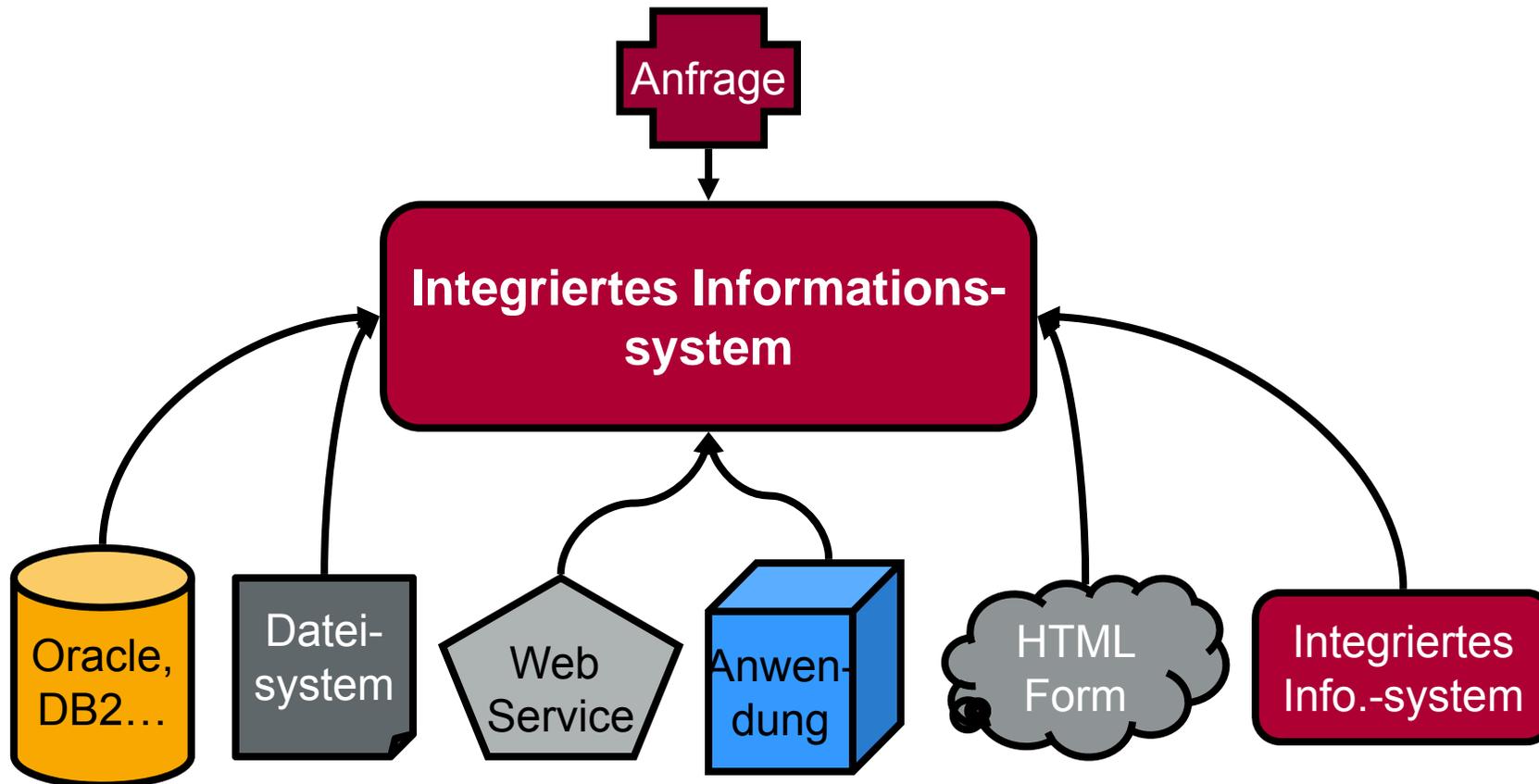


ID	RNGTPCHI	standard; RNA; ROD; 1016 BP.	Molecule type
XX			Name
DT	01-AUG-1991	(Rel. 28, Created)	Date of creation and last update
DT	04-MAR-2000	(Rel. 63, Last updated, Version 2)	Date of creation and last update
XX			
DE	Rat	GTP cyclohydrolase I mRNA, complete cds.	Free text description
XX			
KW	GTP cyclohydrolase I.		Keywords describing the molecule
XX			
OS	Rattus norvegicus	(Norway rat)	Organism
OC	Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.		
XX			
RN	[1]		Article the sequence was published in
RP	1-1016		
RX	MEDLINE; 91093270.		
RX	PUBMED; 1985963.		
RA	Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;		
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).		
XX			
FT	CDS	128..853	Structural annotation (coding sequence)
		/codon_start=1	
		/db_xref="GOA:P22288"	Link to functional annotation of resulting protein
		/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 AGRPRSEEDNELNLPNLAAAYSSILRSLGDPQRQGLLKPURAAATMQFFTKGYQETI SDVLND&I&F&D&H&D&M&V&I&K&D&I&M&F&M&C&E&H&L&V&P&F&V&G&R&V&H&I&G&Y&L&P&N&K&Q&V&L&G&L&S&K&L&A&R&I&V E&I&Y&S&R&L&Q&V&E&R&L&T&K&I&A&V&A&I&E&A&L&Q&A&G&V&G&V&I&E&A&T&H&M&C&M&V&M&R&G&V&Q&M&N&S&K&T&V&T&S&T&M L&G&V&F&R&E&D&P&K&T&R&E&E&F&L&I&R&S"	Translated protein sequence
	Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;		
	gacttgcgac ctcattcggt gcagaactcc tgtcccgggt acagccacag gtcgaagccg	60	
	ccggctaagc cgagcccgag cgcttgtag caacctaggg tgtctcggga gcaatcgcc	120	
	cggtccatg gagaagccgc ggggtgtaag gtcaccaat ggggtccccg agcgggagct	180	
	...		
	catcaggagc tgaactccg tgtgcgagcc ccggtttgca gacccccgt gaggccagc	900	
	ttatctgtct cgattgtaca ttccagttcc agttggata ctgtccaact ttatttcca	960	
	ccatgaattg tattaataa ttatttatag agatgtcaaa taaaggtgat caactt	1016	Sequence of bases



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





# 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

**Variante 1**

☛☛☛☛☛	
☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛☛

☛☛☛☛☛	
☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛
☛☛☛☛	☛☛☛☛☛

**Variante 2**

☛☛☛☛☛			
☛☛☛☛	☛☛☛☛	☛☛☛☛	☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛	☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛	☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛	☛☛☛☛

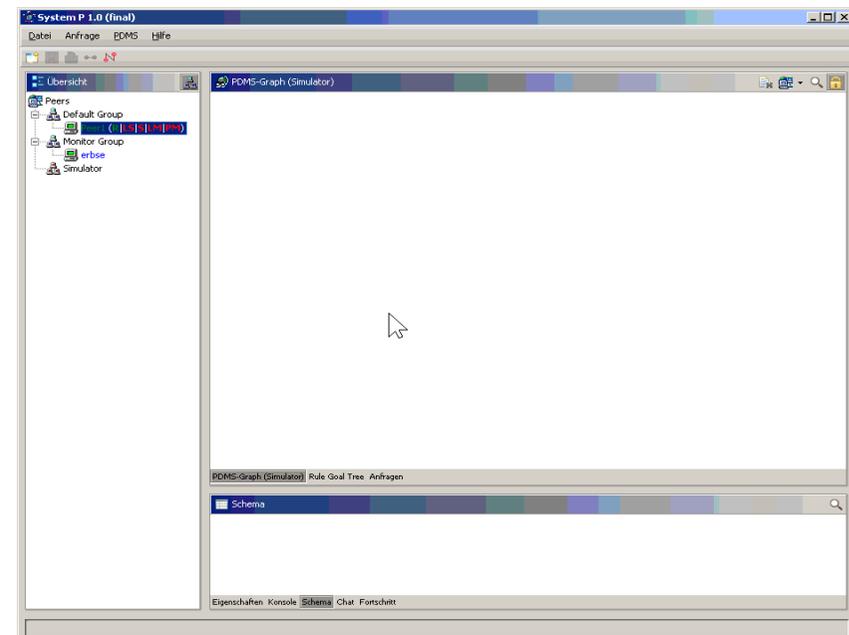
**Variante 3**

☛☛☛☛☛		
☛☛☛☛	☛☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛☛
☛☛☛☛	☛☛☛☛☛	☛☛☛☛☛

# Weitere Forschungsthemen

12

- Informationsintegration
  - Schema Matching
  - Duplikaterkennung
  - Datenfusion
- Datenqualität
- Peer Data Management
- Life Sciences: Aladin
- Search



# Other courses in this semester

13

## Lectures

- DBS I
- Information Integration

## Seminars

- Bachelor: Beauty is our Business
- Bachelor: Mobile Application Development
- Master: Large-scale Data Analysis in the Cloud
- Master: Entity-centric Information Retrieval
- Master: Similarity Search Algorithms
- Forschungsseminar

## Bachelorprojects

- ETL-Prozess-Management für BMW Financial Services
- IBM Midas: Extreme Web Data Integration for Government Data



### 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 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 operations.

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

CR Categories: 3.70, 3.73, 4.22, 4.26, 4.33, 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 clutter 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], INGRESS [37], QBE [46], and System R [2].

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 25-June 1, 1979.

Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95133.

© 1979 ACM 0362-5915/79/1200-0387 \$06.75

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 387-434.

# Projektseminar

## Mobile Application Development (MAD)



14

- (Datenbank-)Themen
  - **Data & information retrieval** (camera, audio, GPS, compass, speed, personal data, ...)
  - **Peer-to-peer communication** (Google Talk)
  - **Data management** (local vs. remote, caching, power efficiency)
  - **Stream processing** (web data, location data, ...)
  - **Mobile databases** (SQLite)
  - **Visualisierung** von Daten (Grafiken, Animationen), tbc.
- Ablauf
  - Jedes Team entwickelt eine Android App zu einem Thema
  - Präsentation zum Semesterende
  - 8 Teams à 2 Studenten à 1 G1 Dev Phone
  - Wöchentliche Veranstaltungen (Techniken, Programmierung, Beispiele, Übungen)
  - [http://code.google.com/android/adc/gallery\\_winners.html](http://code.google.com/android/adc/gallery_winners.html)



# Überblick

15

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



# Vorstellung Teilnehmer

# Seminarleistungen

17

## Lesen

- Paper lesen und verstehen
- Verwandte Literatur lesen und verstehen
- Mindestens eine individuelle Besprechung mit mir
- Aktive Teilnahme an anderen Vorträgen

1/5 Note

## Vortragen

- Kurzvorstellung der Literatur (je 2 min)
- Mindestens eine Folien-Besprechung mit mir
  - Spätestens 1 Woche vor Vortrag
- 30 min. Vortrag am jeweiligen Termin
  - + 15 min Diskussion

2/5 Note

## Ausarbeitung

- Gemeinsame Gliederungs-Besprechung
- 8-10-seitige Ausarbeitung bis zum 17.9.2010
- Unter Verwendung der LaTeX-Vorlage im WWW

2/5 Note

**Aktive Teilnahme an sämtlichen gemeinsamen Terminen**

# Feedback

18

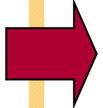
Fragen bitte jederzeit!

- Während des Seminars
- Während individueller Besprechungen
  - Termin bei Katrin Heinrich ausmachen
    - ◇ office-naumann@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

19

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



# Wissenschaftliche Texte lesen

20

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

# Wissenschaftlichen Vortrag halten

21

- Gliederung
  - Die Kunst des Weglassens
- Foliengestaltung
  - Powerpoint, Keynote, usw.
- Zeit einhalten
- Techniken zur Vorbereitung
- Techniken während des Vortrags



# Ausarbeitung schreiben

22

- Gliederung
- Schreibstil
  - Objektives
  - Subjektives
- Plagiate
- LaTeX
  - denn „Beauty is our Business“

# Zeitlicher Ablauf – Termine

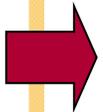
23

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

# Überblick

24

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



# Themenüberblick

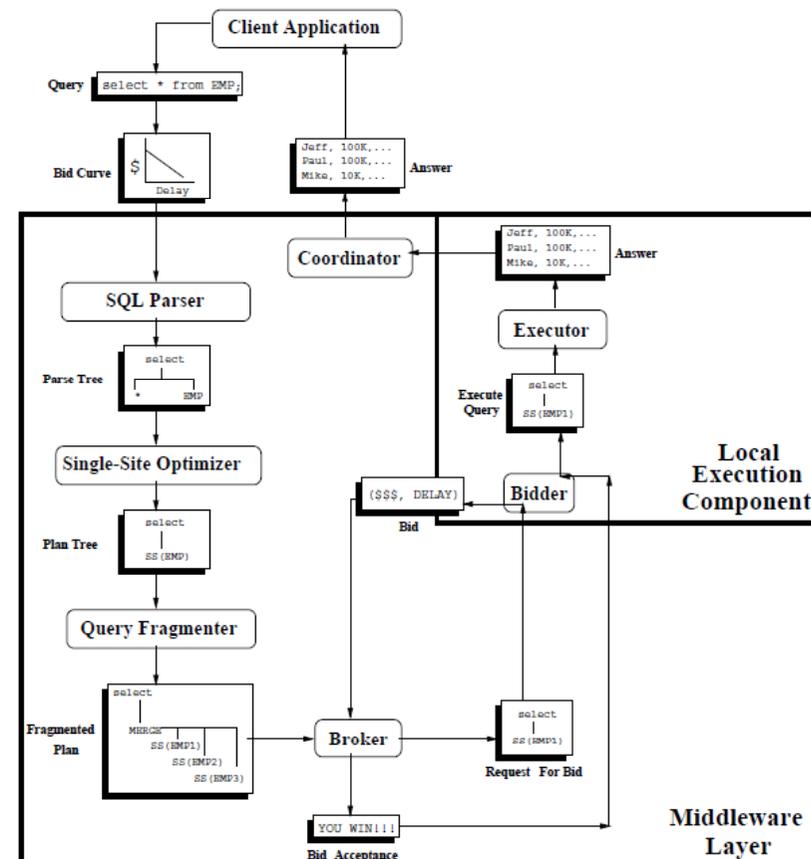
25

- Systems
  - Mariposa ([pdf](#))
  - Trio ([pdf](#))
  
- Top-K
  - Fagins Algorithmus ([pdf](#))
  - Enough Already in SQL ([pdf](#))
  
- Algorithms
  - Sorted Neighborhood ([pdf](#))
  - Data Mining ([pdf](#))

# Mariposa

26

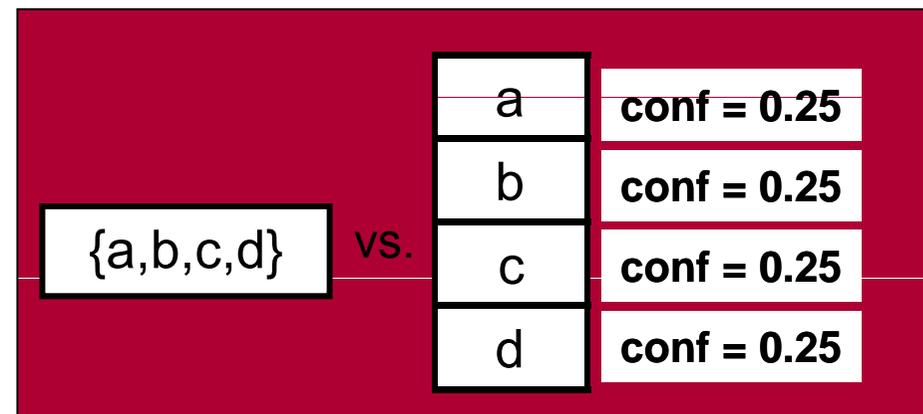
- Neuartige Architektur für ein weit verteiltes Informationssystem
- Mikroökonomisches Prinzip zur Anfrageoptimierung
  - Datenhaltung
  - Anfragebearbeitung



# Trio

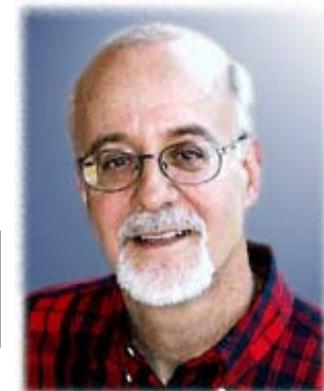
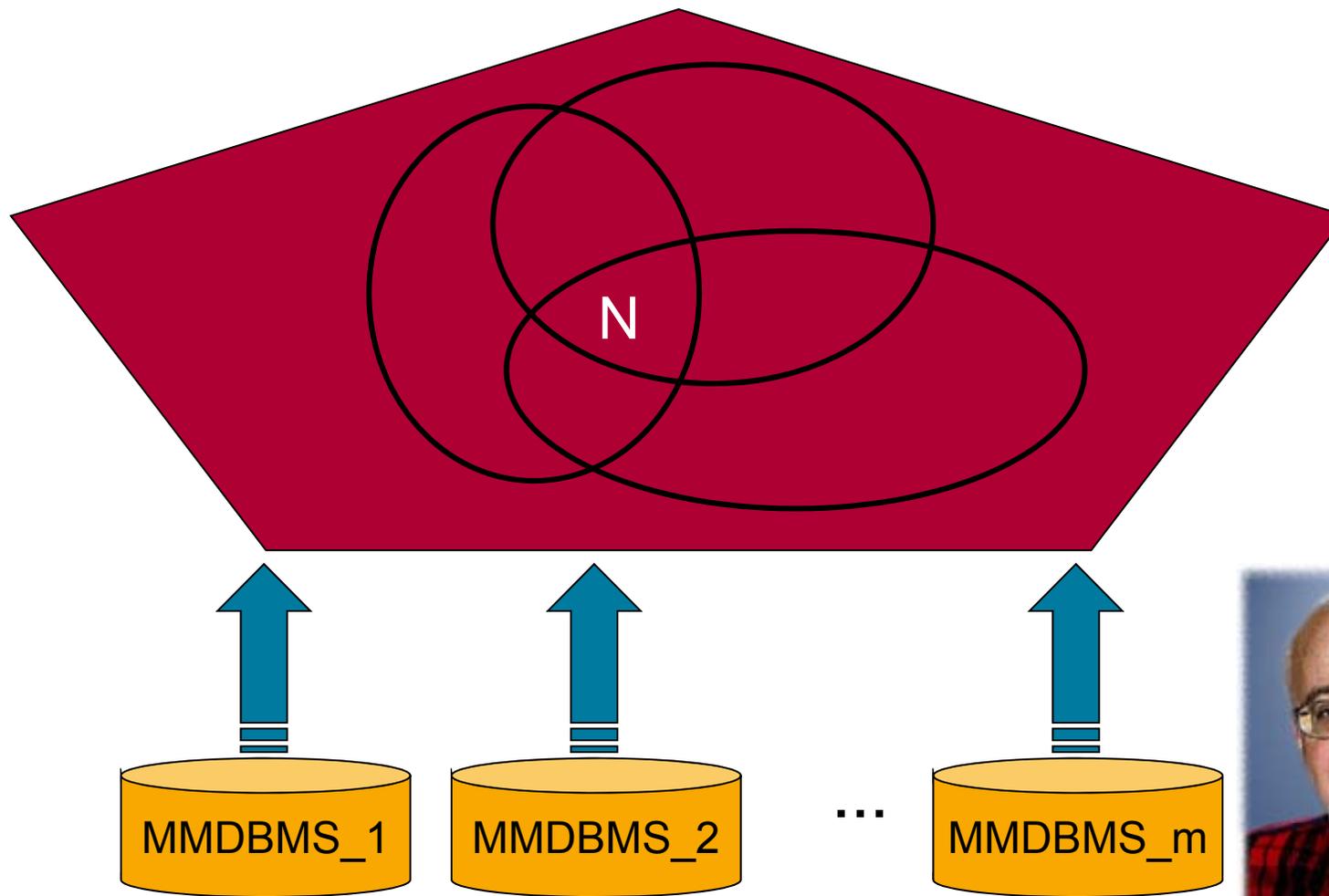
27

- 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

28

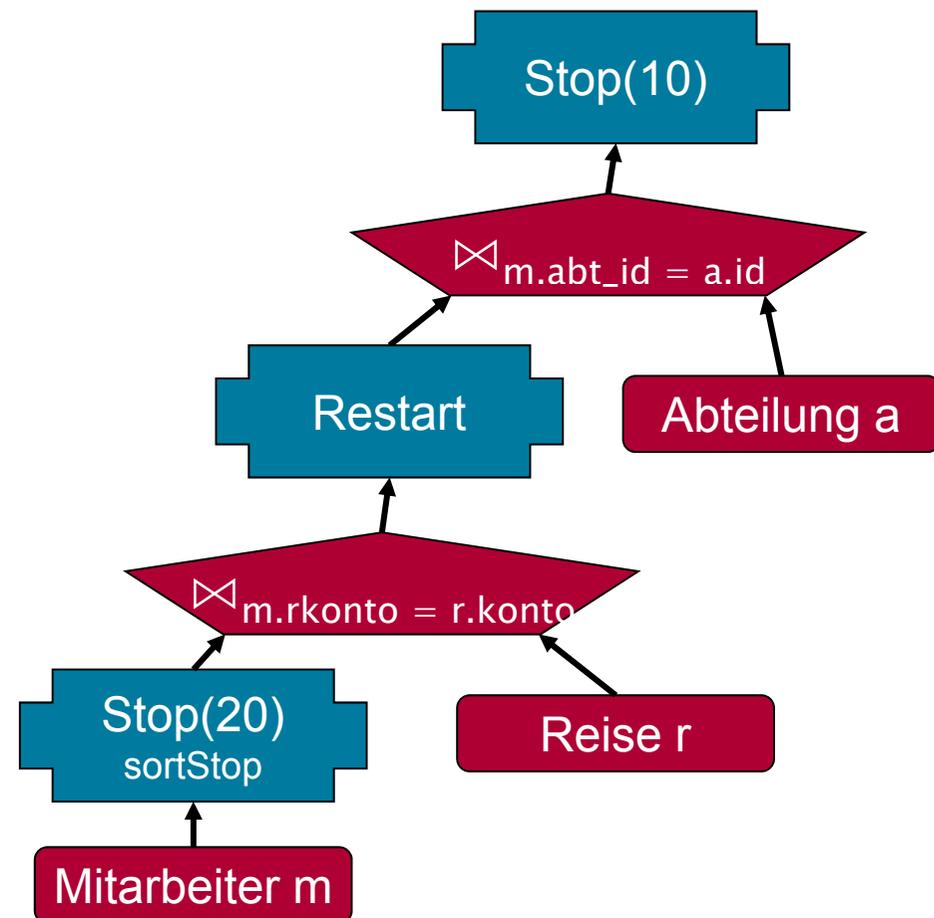


# „Enough Already“ in SQL

29

- First-N und Top-N Techniken in SQL

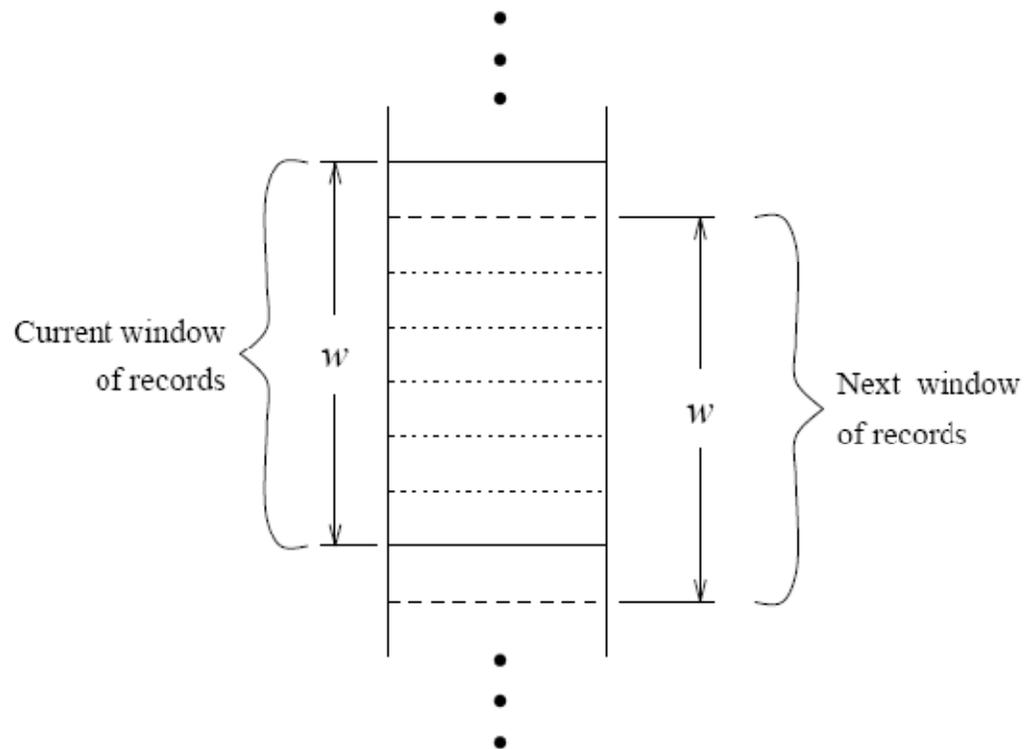
- Syntax & Semantik
- Neue Operatoren
- Optimierung
- Evaluation



# Sorted Neighborhood

30

- Datenreinigung und Duplikaterkennung
  - Effizienz
  - Effektivität



# Apriori: Data Mining

Fast Algorithms for Mining Association Rules

31

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



## Nächste Schritte

32

- Sie: E-Mail an mich
  - Mit Wunschliste (Platz 1 – 3)
    - ◇ Daraus: Themenvergabe
  - <http://www.informatik.uni-trier.de/~ley/db/index.html>
  - Bis Mitternacht
- Ich: Themenvergabe bald
  - E-Mail Benachrichtigung
- Ich: Vortrag zum „Lesen wissenschaftlicher Texte“