Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

19.4.2010

Felix Naumann

Das Motto

... 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

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

Überblick

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

Forschungsgruppe Informationsintegration

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

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

Information Systems Team

Katrin **Heinrich**

Prof. Felix **Naumann**

Silke **Trißl**

Uwe **Draisbach**

Dustin **Lange**

project **ViQTOR**

Paul **Führing**

DQ Annotation & Assessment

Data Fusion

Duplicate Detection

Schufa

Entity Search

Information Integration

project **HumMer**

project **System P**

Armin **Roth**

Peer Data Management Systems

Information Quality

Data Profiling

IBM

Data Profiling & Cleaning

Johannes **Lorey**

Matching

Data Integration for Life Science Data Sources

project **PoSR**

Christoph **Böhm**

ETL Management

Service-Oriented Systems

project **Aladin**

Tobias **Vogel**

Forschungskolleg

Mohammed **AbuJarour**

Ontologies, Profiling

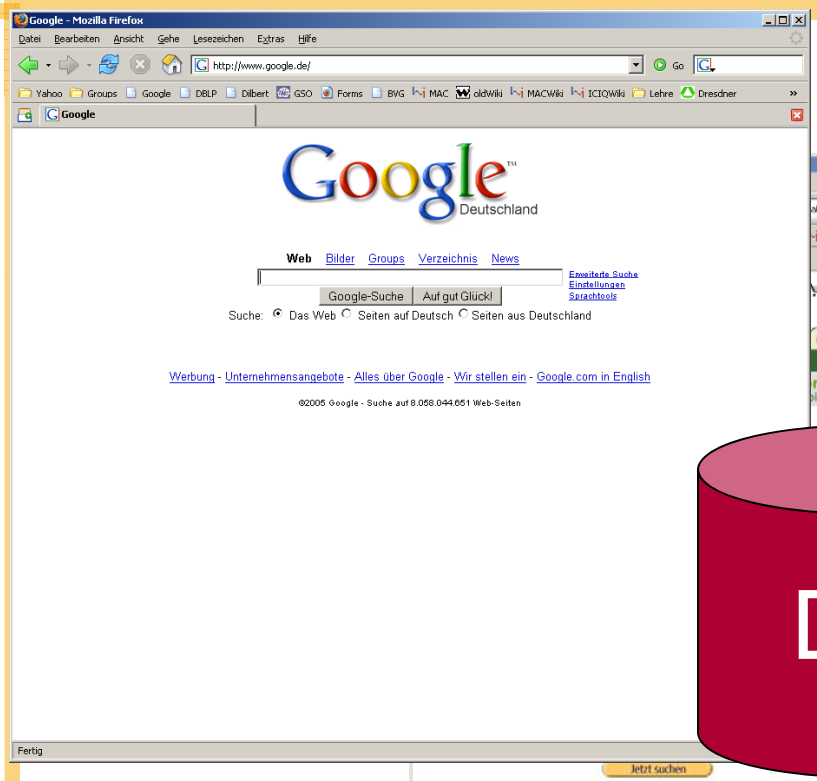
Frank **Kaufer**

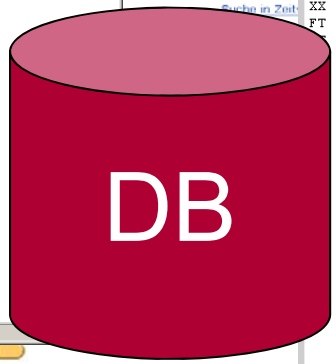
Jana **Bauckmann**

Alexander **Albrecht**

Data Profiling for Schema Management

Was sind Informationssysteme?



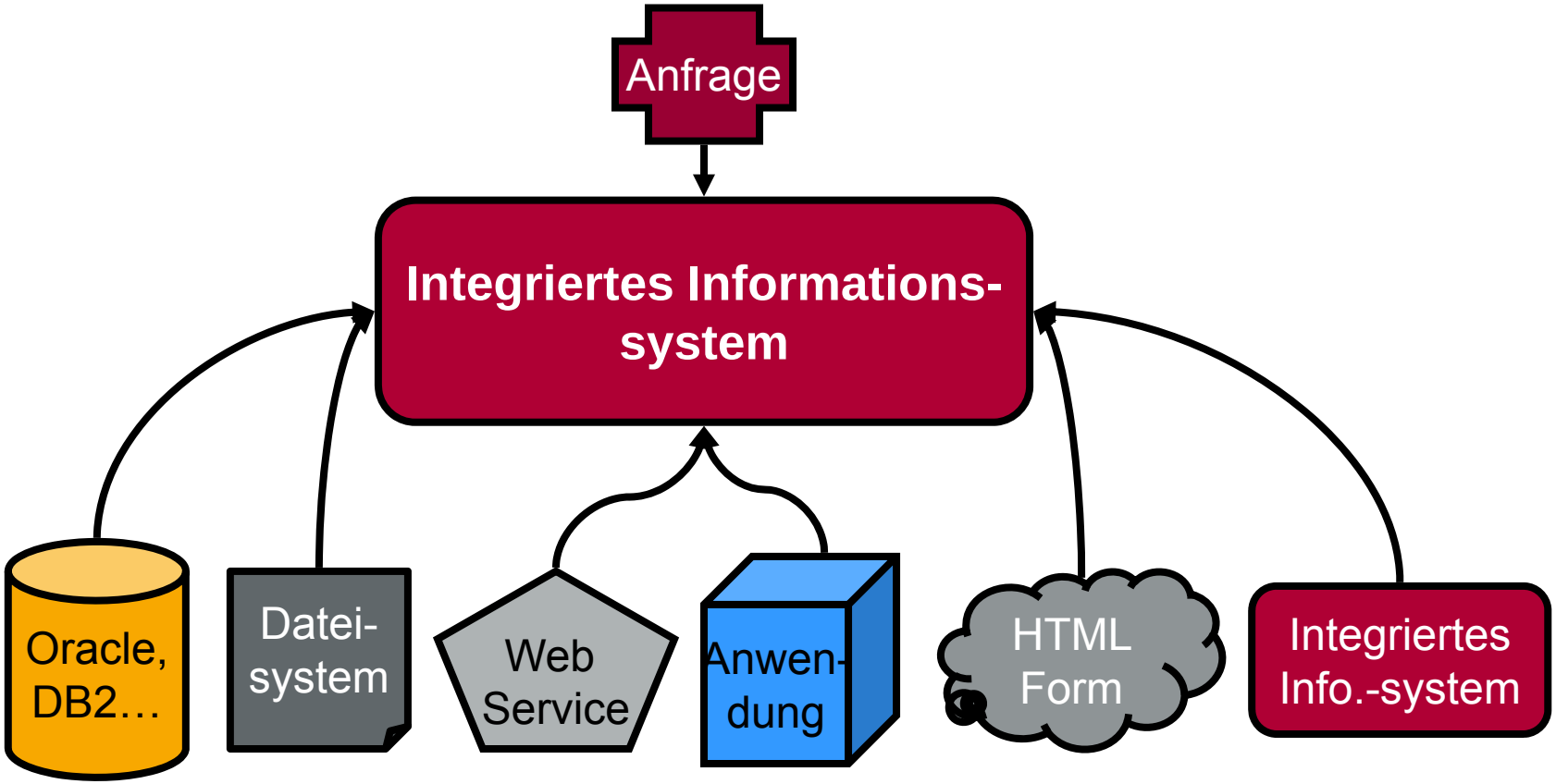



```
ID   RNGTPCHI   standard; RNA; ROD; 1016 BP.
XX
DT   01-AUG-1991 (Rel. 28, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 2)
XX
DE   Rat GTP cyclohydrolase I mRNA, complete cds.
XX
KW   GTP cyclohydrolase I.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
XX
RN   [1]
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
RT   first enzyme of the tetrahydrobiopterin biosynthetic pathway";
RL   J. Biol. Chem. 266(2):765-769(1991).
XX
FT   CDS             128..853
FT                   /codon_start=1
FT                   /db_xref="GOA:P22288"
FT                   /db_xref="SWISS-PROT:P22288"
FT                   /EC_number="3.5.4.16"
FT                   /gene="GTP cyclohydrolase I"
FT                   /product="GTP cyclohydrolase I"
FT                   /protein_id="AAA41299.1"
FT                   /translation="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK
FT                   AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI
FT                   SDVLNDAIFDEDHDEMVIVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV
FT                   EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRGVQKMNSKTVTSTML
FT                   GVFREDPKTREEFLTLIRS"
Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
gacttcgaac ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggccg        60
ccggctaagc cgagccgcag cgcttggtag caccttaggg tgtctcggga gcaatcgcgc       120
cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct       180
   ...
catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg       900
ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttatttctca       960
ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caacttt         1016
//
```

Molecule type
Name
Date of creation and last update
Free text description
Keywords describing the molecule
Organism
Article the sequence was published in
Structural annotation (coding sequence)
Link to functional annotation of resulting protein
Translated protein sequence
Sequence of bases

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

Integrierte Informationssysteme



Schematische und Daten-Heterogenität

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

Variante 1

Männer

Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen

Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen

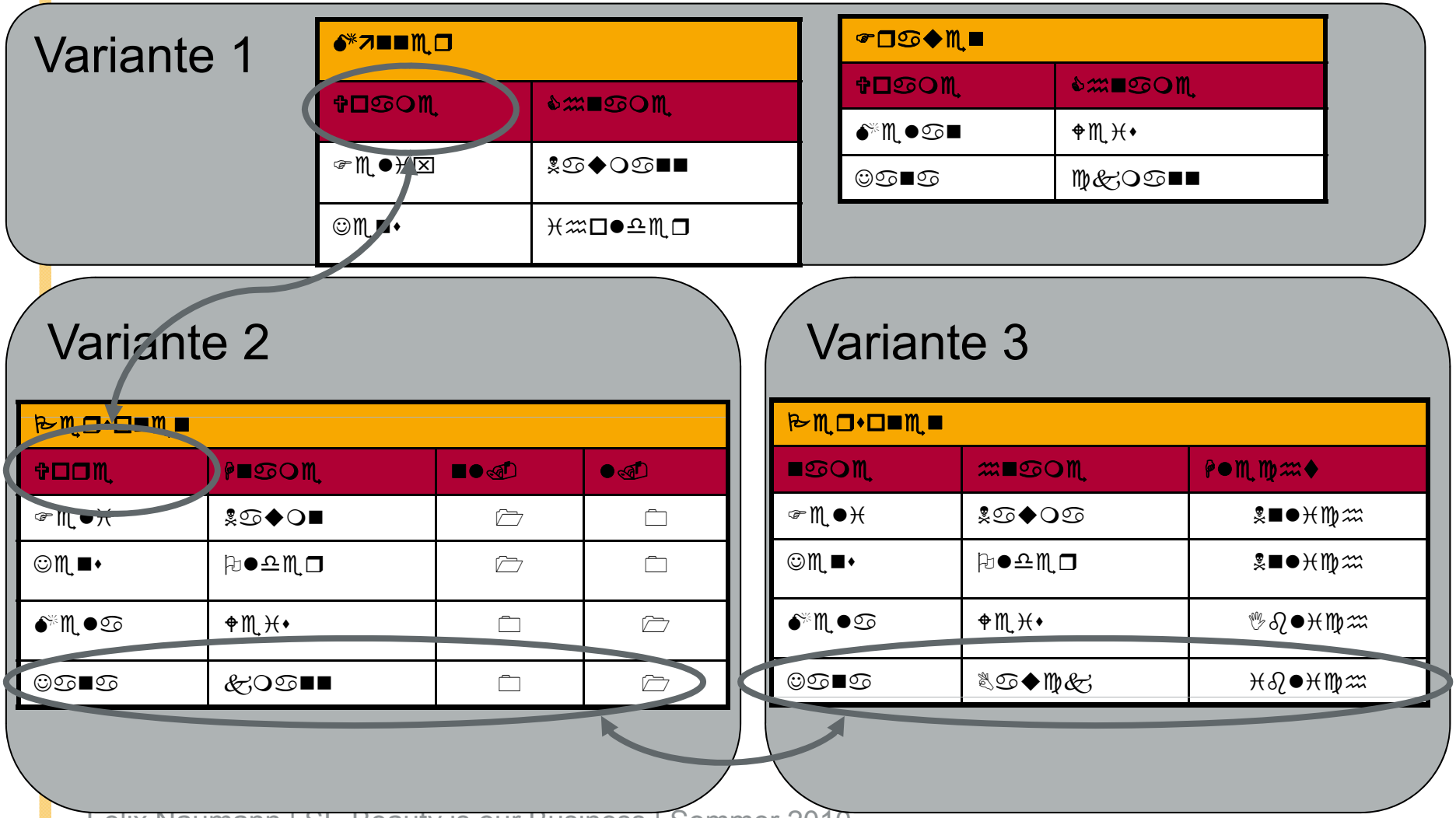
FirstNa	Name	male	femal
Felix	Naumann	Ja	Nein
Jnes	Bleiho.	Ja	Nein
Melanie	Weiß	Nein	Ja
Jana	baukman	Nein	Ja

Variante 3

Personen

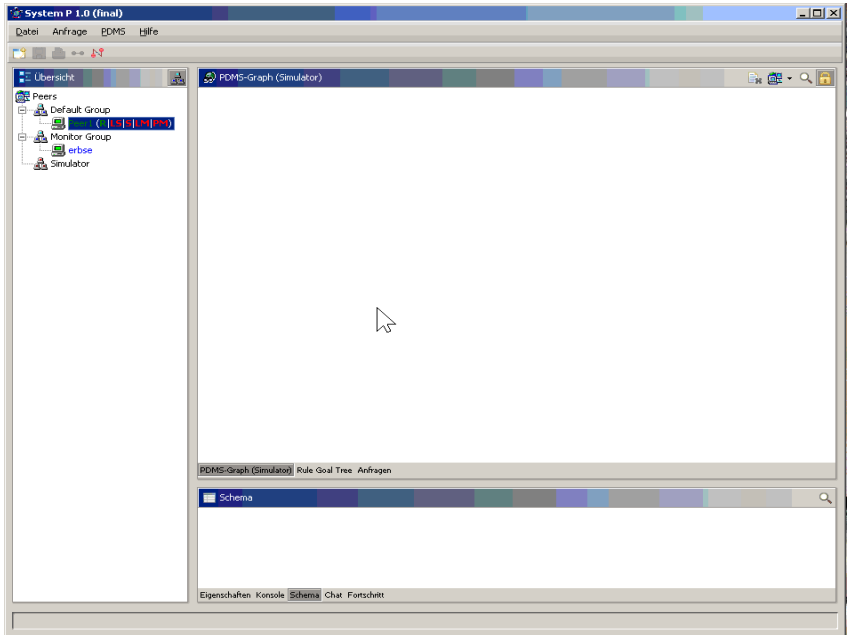
VN	NN	SEX
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Schematische und Daten-Heterogenität

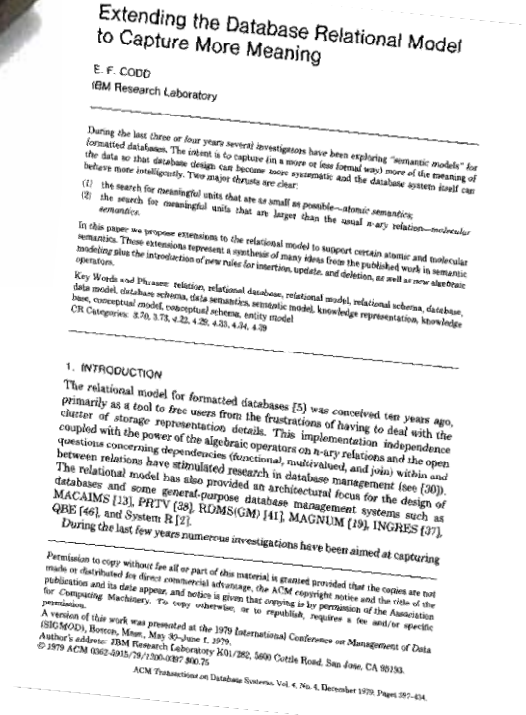


Weitere Forschungsthemen

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



Other courses in this semester



Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

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

Projektseminar
Mobile Application Development (MAD)

- (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

Überblick

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

Vorstellung Teilnehmer

Seminarleistungen

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

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

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

Wissenschaftliche Texte lesen

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

Wissenschaftlichen Vortrag halten

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

Ausarbeitung schreiben

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

Zeitlicher Ablauf – Termine

- 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

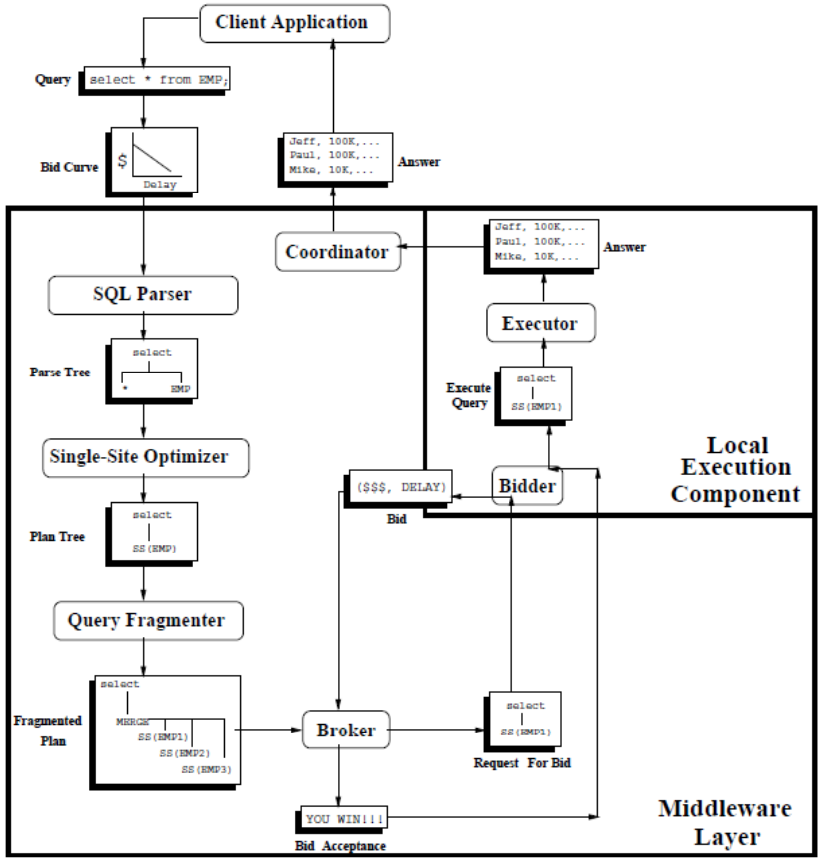
- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen

Themenüberblick

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

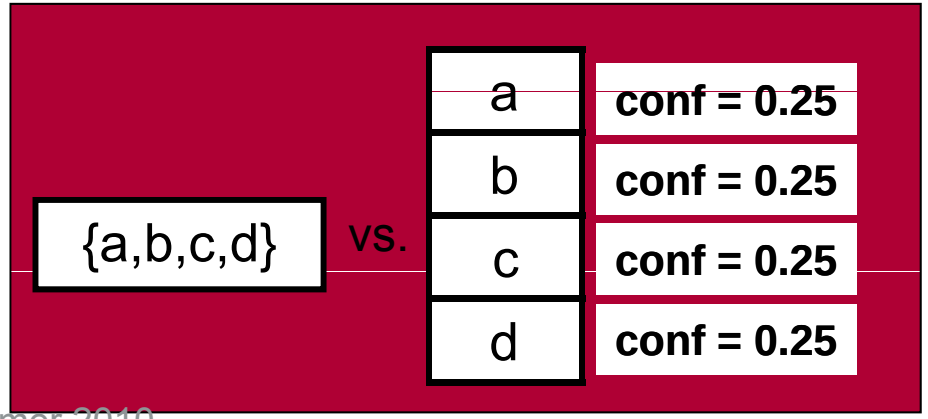
Mariposa

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

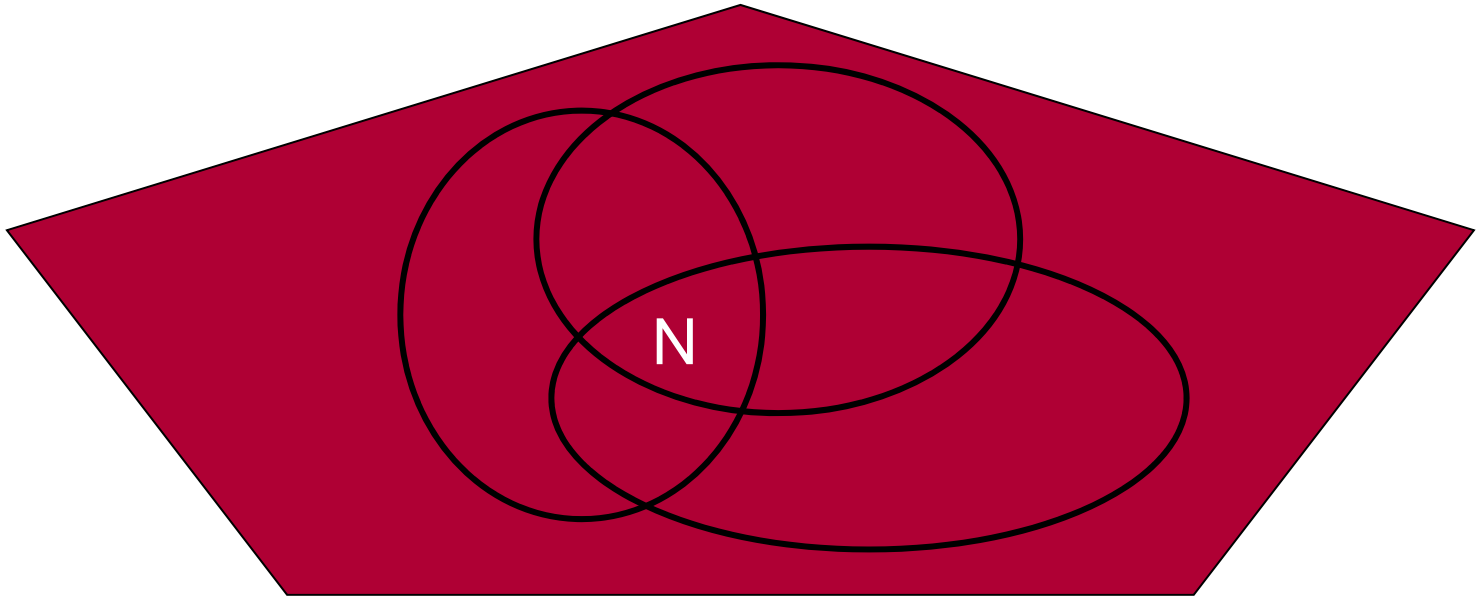


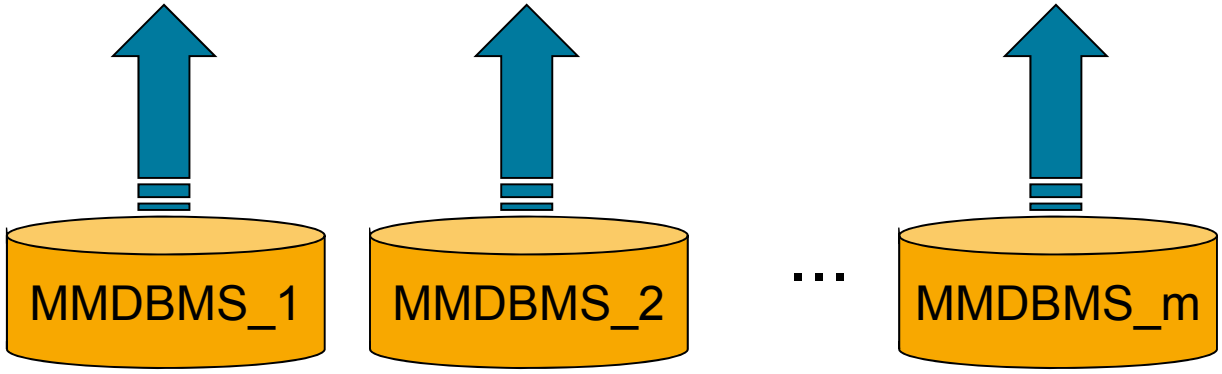
Trio

- 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



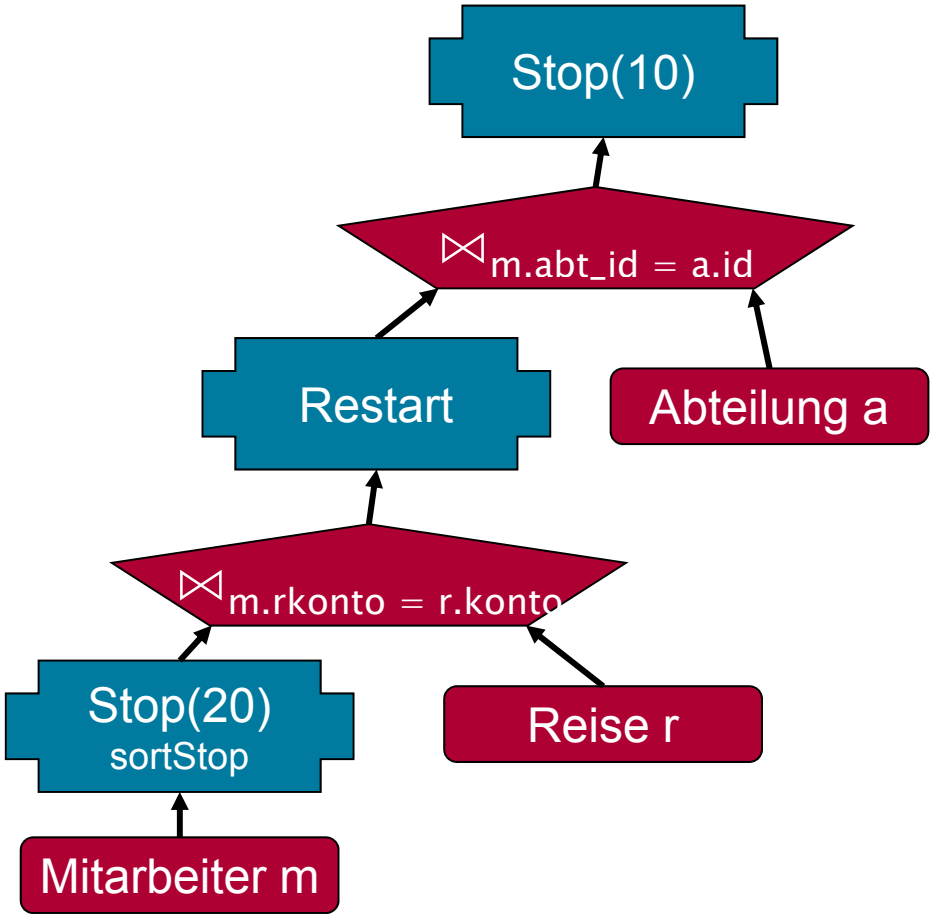
Fagins Algorithmus





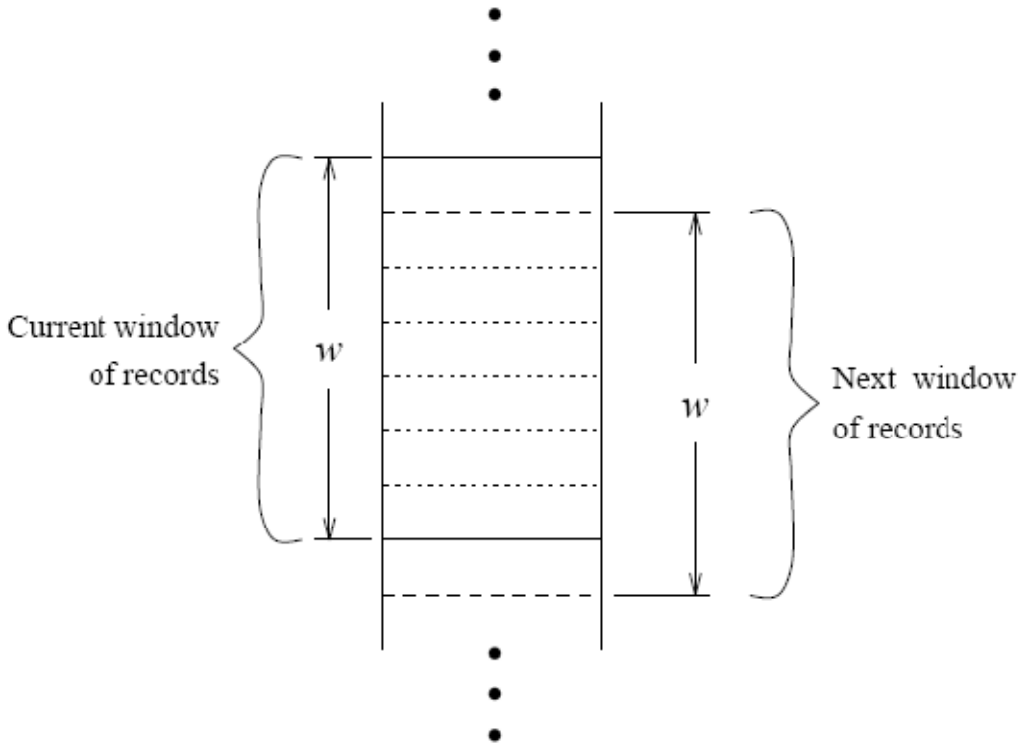
„Enough Already“ in SQL

- First-N und Top-N Techniken in SQL
  - Syntax & Semantik
  - Neue Operatoren
  - Optimierung
  - Evaluation



Sorted Neighborhood

- Datenreinigung und Duplikaterkennung
  - Effizienz
  - Effektivität



Apriori: Data Mining

Fast Algorithms for Mining Association Rules

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

Nächste Schritte

- 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“