Einsatz von Datenbanken im Forschungslabor

Workflow und Data Mining

Friedrich-Alexander-Universität Erlangen-Nürnberg
Technische Fakultät, Institut für Informatik
Lehrstuhl für Informatik 6 (Datenbanksysteme)

Prof. Dr. Klaus Meyer-Wegener

Datenbanken – weshalb?

- **Speichergeräte**
  - werden immer größer und preiswerter
- **immer mehr Vorgänge**
  - werden mit Rechnern gesteuert und verwaltet
  - z.B. Paketversand, Zahlungsverkehr
- **digitale Aufzeichnung**
  - z.B. Photos, Messwerte in Physik und Astronomie, Genom
- **Datenvolumen wächst enorm**
  - nach Mega und Giga und Tera (10^{12}) und Peta (10^{15})
- **zunächst alles Dateien**
  - Suche? Auswertung? (Handel mit Daten?)

Datenbanken – wann?

bei einer großen Menge von Daten!

auch, aber vor allem:

- **vielseitig verwendbar (offen für neue Anwendungen)**
- **wohlstrukturiert**
- **redundanzfrei**
- **flexibel abfragbar (recherchierbar)**
- **von mehreren Anwendungen gleichzeitig nutzbar, bei hoher Aktualität der Daten**
- **ausfallsicher**

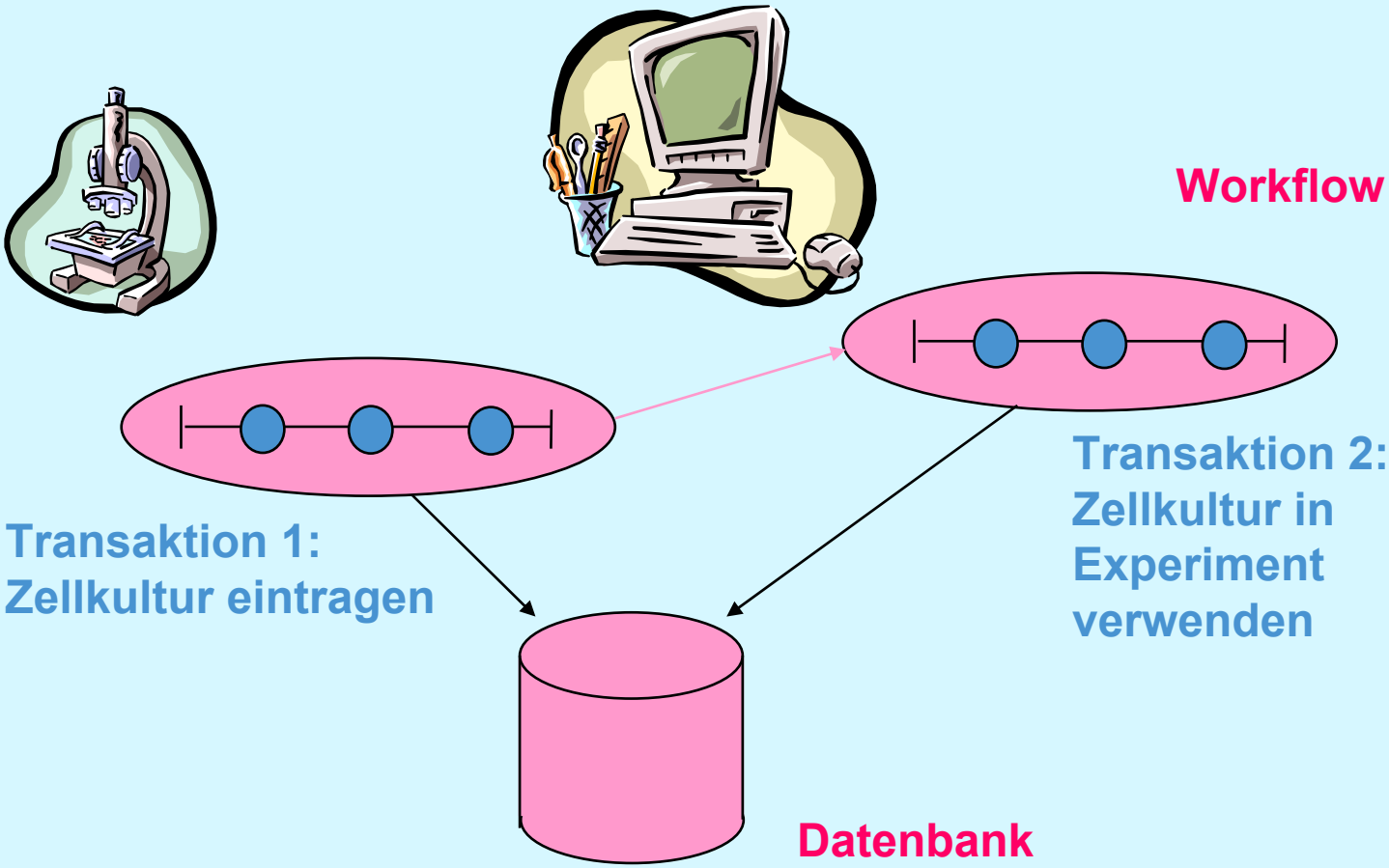
Dienste von Datenbanken

- **Datenmodell und Schema**
  - Datenmodellierung (DB-Design) schwierig und aufwändig
  - enge Kooperation zwischen Entwickler und Anwender notwendig
- **Normalisierung**
  - Richtlinien für "gute" Schemata
- **Anfragesprache (query language)**
  - Selektionen, Verknüpfungen, Aggregationen
- **Synchronisation der Zugriffe**
  - fiktiver Ein-Benutzer-Betrieb
  - tatsächlich aber gleichzeitige Zugriffe
- **Transaktionen**
  - "Alles oder nichts" für eine Sequenz von Änderungen
  - automatische Reparatur von inkonsistenten Zuständen

Transaktionen

- **unteilbare Übergänge für zusammengehörige Änderungen**
  - Banküberweisung:
    - Beginn der Transaktion
    - Abbuchen Konto 1
    - Zubuchen Konto 2
    - Ende der Transaktion
- **Konsistenzerhaltung**
- **Vollständigkeit der Eingaben / Änderungen**
- **Garantie des Datenbanksystems**
- **mehrere zusammengehörige Transaktionen verschiedener Benutzer: Workflow**

Datenerfassung



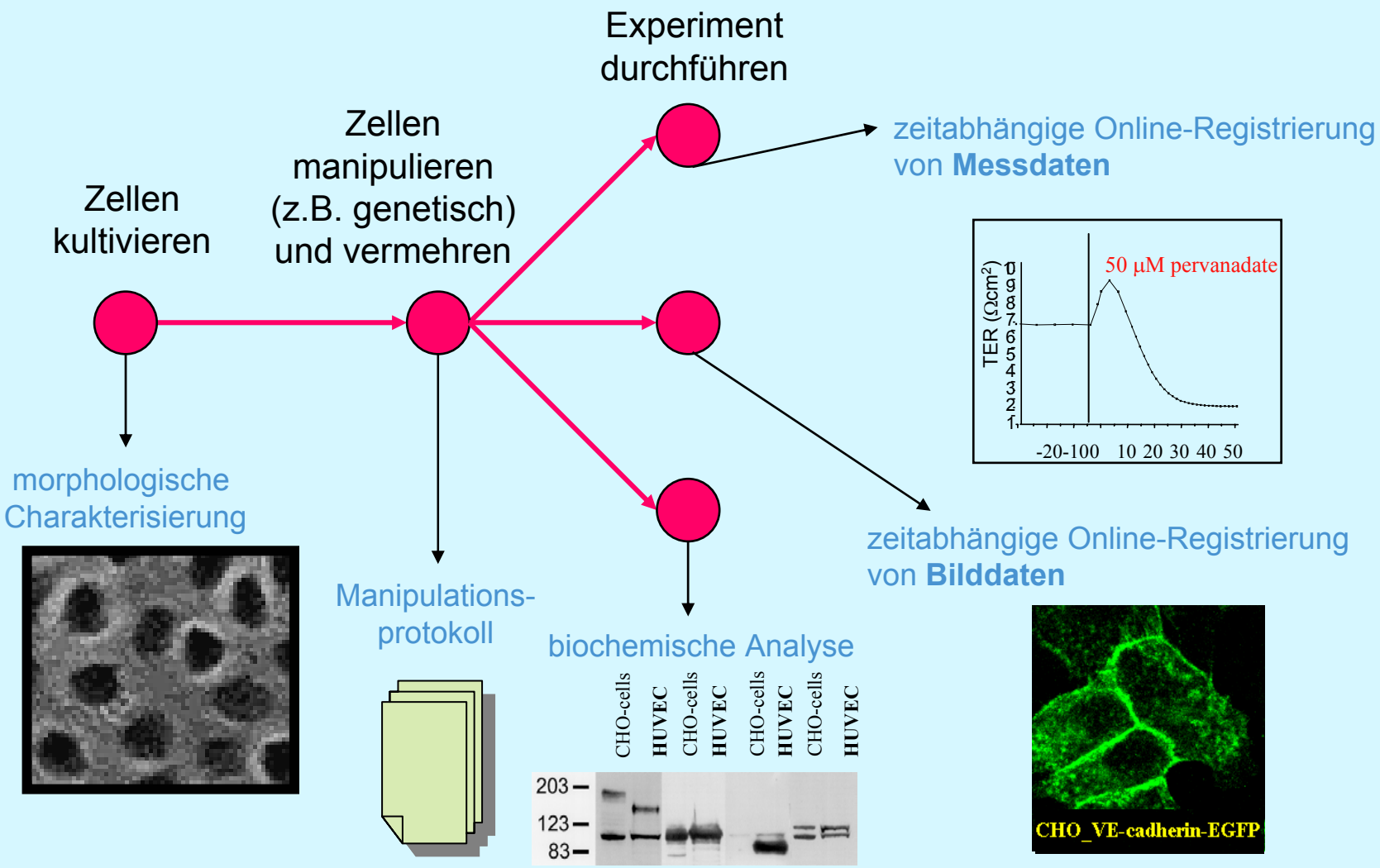
Workflows

- **Information**
  - Wie macht man das (in dieser Firma)?
    - Beschaffung, Dienstreise, Laborversuch, ...
  - Wer muss gefragt werden?
  - Was muss man alles angeben?
- **Vollständigkeit**
  - an alles denken
- **Reihenfolge**
  - technische Präzedenzen u.a.
- **Kontrolle**
  - Zwang, Überwachung ("Assistent" oder "Polizist"?)
  - interne Überprüfbarkeit, Nachvollziehbarkeit:
    Good Laboratory Practice

Workflow-Systeme

- **Abläufe (Prozesse)**
- **Definition**
  - Schritte, Bedingungen, Akteure, Daten, ...
- **Durchführung**
  - Initiierung
  - automatische Weiterleitung
  - Eingangskorb bzw. Aufgabenliste bei den Klienten
  - Bereitstellung von Anwendungen (**Transaktionen!**) und Daten
  - Zustandsinformation im Management
- **Einsatz für Erfassung, Datenpflege, operatives Geschäft**

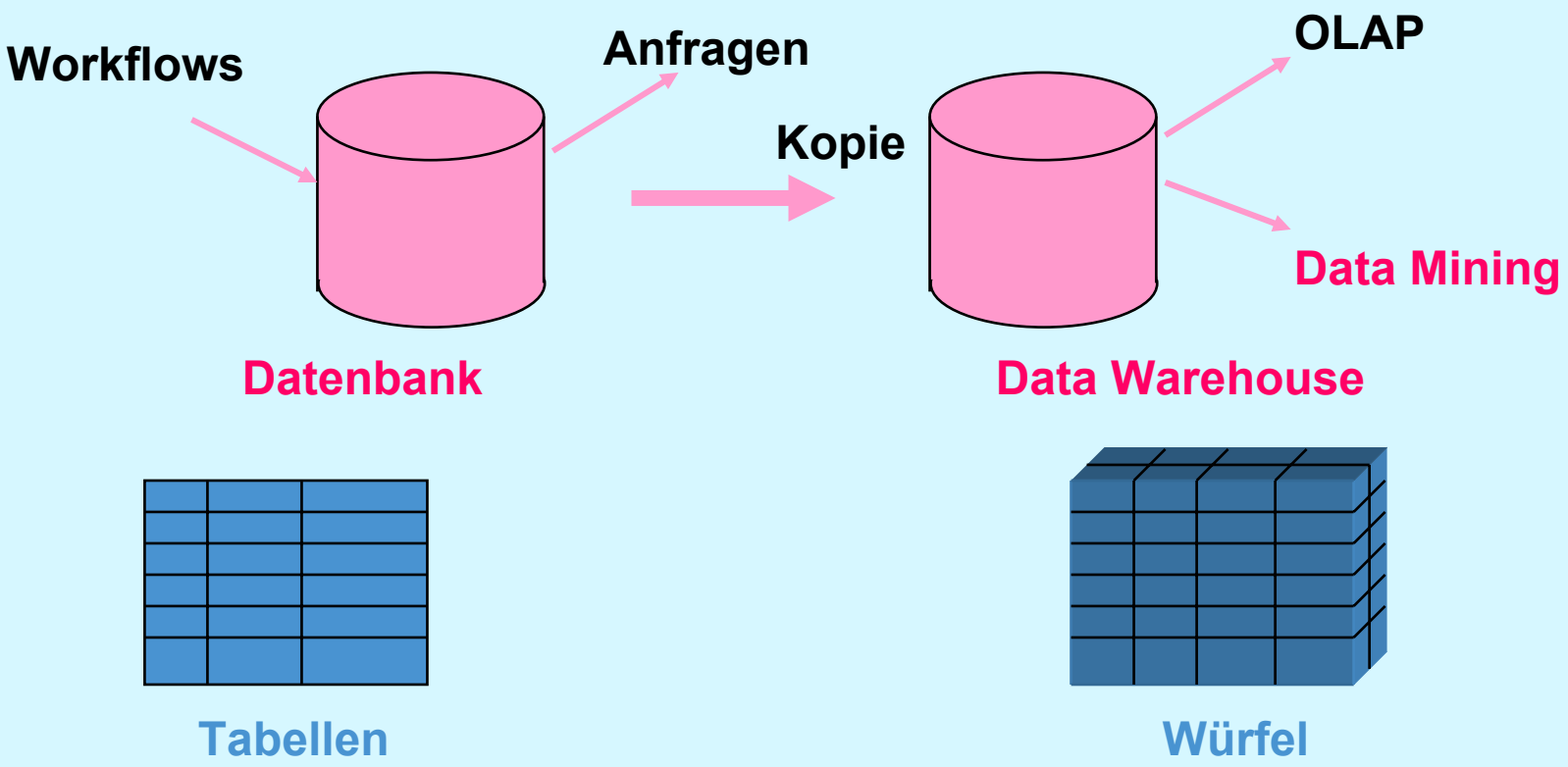
Workflow-Beispiel



Auswertung: Data Warehouse, OLAP

- **große Datenmengen**
- **historische Daten (z.B. Messreihen)**
- **gesucht: Entwicklungen, Trends, Zusammenhänge**
- **nicht auf demselben Datenbestand wie Erfassung**
  - Leistungsfähigkeit: gegenseitige Behinderung vermeiden
  - Umstrukturierung für Optimierung
  - Verteilung der Originaldaten
  - Bereinigung
- **zusammenfassende Kopie: Data Warehouse**

Data Warehouse



Data Warehouse: Modelle

- **Stern-Schema:**
  - Faktentabelle und Dimensionstabellen
- **Fakten:**
  - Verkäufe, Messwerte, ...
- **Dimensionen:**
  - Zeit, Ort, Beteiligte, Einstellungen, ...
- **auch hierarchisch: Uhrzeit – Tag – Woche – Monat**
- **typische Anfragen: Gruppierung und Aggregation**
  - Summe, Mittelwert, Maximum, Minimum usw.
- **Drill-Down: mehr Detail**
  - von den Monatssummen zu den einzelnen Tagen
- **Roll-up: mehr Verdichtung**
- **wenn man so will: Statistik**

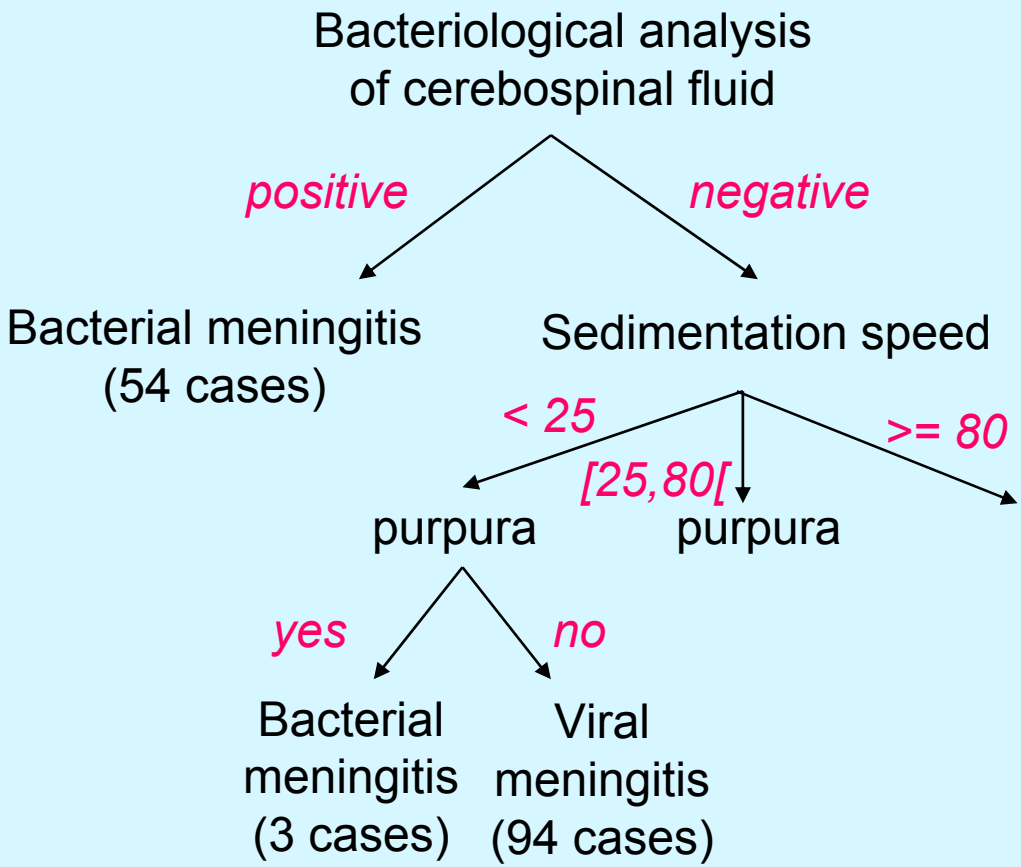
Data Mining

- **große Datenmengen (im Data Warehouse) nach bisher unbekannten Zusammenhängen durchsuchen**
  - basiert auf Knowledge Discovery (daher KDD)
- **Auswertungen über die Anfragesprache hinaus**
- **Suche nach Mustern, Regelmäßigkeiten, Ausnahmen ("das Signal im Rauschen")**
- **wichtige Methoden:**
  - Klassifikation (Entscheidungsbäume)
  - Finden von Assoziationsregeln

Klassifikationen

- **eine Form der multivariaten Analyse**
- **Herleitung aus einer repräsentativen Datenmenge**
- **Unterscheidung von Merkmalen und Klassen**
  - z.B. Symptome und Diagnose
- **Entscheidungsbäume**
  - genutzt für Vorhersagen
    bei neuen Merkmalskombinationen
  - liefert Hinweise auf Ursachen,
    aber nicht die Ursachen selbst

Klassifikationen: Beispiel



Assoziationsregeln

- **Implikationsregeln für Zusammenhänge zwischen verschiedenen Fakten:**
  - "Wenn jemand Windeln kauft, kauft er auch Bier"
- **Vertrauen (confidence):**
  - Anteil der Datensätze, bei denen Folgerung erfüllt ist, an denen, die Voraussetzung erfüllen
- **Unterstützung (support):**
  - Anteil der Datensätze, bei denen die Regel stimmt

Zusammenfassung und Ausblick

- **Datenbanken:**
  - Grundlage
    für die systematische Erfassung und Auswertung
    von Datenmengen
- **Workflows:**
  - Systemunterstützung für operative Abläufe
- **Data Warehouses:**
  - Kopien von Datenbeständen für umfangreiche Auswertungen
- **Data Mining:**
  - Finden von Zusammenhängen
- **bisher oft nur für Geschäftsdaten, zunehmend auch für Experimentaldaten ("scientific databases"):**
  - Arbeitserleichterung
  - neue Auswertungen

Literatur

- Stefan Jablonski, Markus Böhm und Wolfgang Schulze: *Workflow Management - Entwicklung von Anwendungen und Systemen*. Heidelberg : dpunkt, 1997. - ISBN 3-920993-73-X
- Usama Fayyad, David Haussler and Paul Stolorz: Mining Scientific Data. *Communications of the ACM* 39 (1996), Nr. 11, S. 51-57