# Bio Data Management

Kapitel 8

## Datenintegration - Ansätze und Systeme

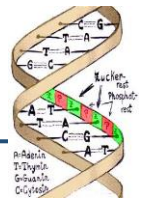
Wintersemester 2014/15

Dr. Anika Groß
Universität Leipzig, Institut für Informatik, Abteilung Datenbanken
http://dbs.uni-leipzig.de

UNIVERSITÄT LEIPZIG

# Vorläufiges Inhaltsverzeichnis

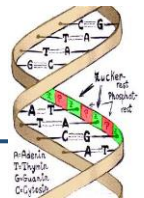
1. Motivation und Grundlagen
2. Bio-Datenbanken
3. Datenmodelle und Anfragesprachen
4. Modellierung von Bio-Datenbanken
5. Sequenzierung und Alignments
6. Genexpressionsanalyse
7. Annotationen
8. Datenintegration: Ansätze und Systeme
9. Matching
10. Versionierung von Datenbeständen



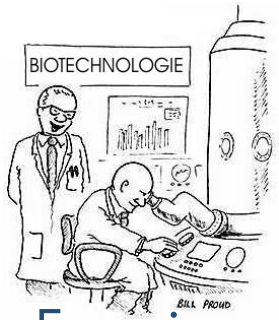
# Gliederung

- Motivation & Datencharakteristik
- Schema- und Instanzdatenintegration
- Physische vs. virtuelle Integration

- Beispiel LIFE: epidemiologische Studie



# Problembereich: Datenintegration




Differentiell exprimierte STAT3 Gene bei malignen Lymphomen von Patienten, die älter als 50 Jahre sind?

Experimentelle Daten
z.B. Microarray-Daten

Daten über biol. Objekte,
z.B. Gene, Proteine

Klinische Daten
z.B. Patientendaten

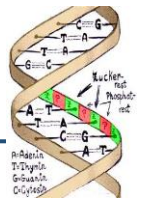
...

**Aufgabenspektrum:**

- Selektion von interessanten Daten,
- übergreifenden Datenanalyse und
- Interpretation von Analyseergebnissen

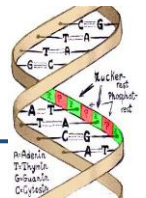
# Motivation

- Verschiedene Arten von Analysen
  - Analyse von Sequenzdaten (z.B. multiple alignments)
  - Identifikation von Transkriptionsfaktorbindungsstellen
  - Genexpressionsanalyse
  - Transkriptionsanalyse, z.B. ENCODE Projekt (http://www.genome.gov/ENCODE)
  - Functional profiling
  - Pathway Analyse und Rekonstruktion
- Viele heterogene Datenquellen
  - Experimentdaten, z.B. von Chip-basierten Techniken
  - Experimentbeschreibung (Metadaten eines Experiments)
  - Klinische Daten
  - Viele miteinander verbundene Webdatenquellen und Ontologien
  - Private vs. öffentliche Daten

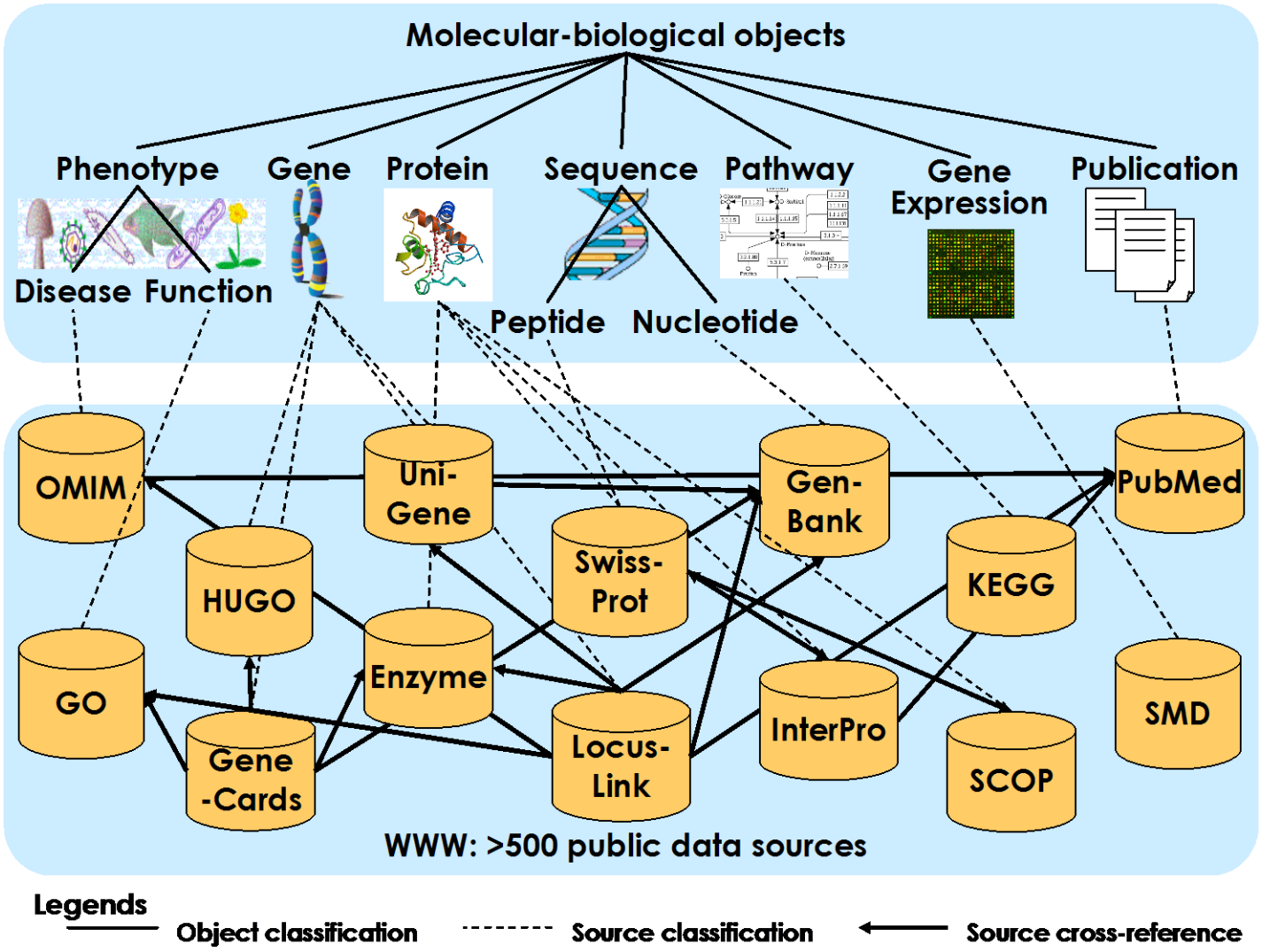


# Datencharakteristika

| Datenart | Quelle | Typ | Charakteristik | Nutzung |
|---|---|---|---|---|
| Bilddaten | Scan | binär | Sehr großes Datenvolumen, Dateien | Generierung von Expressions-, Mutations- und Sequenzdaten |
| Expressions-/ Mutationsdaten | Bildverarbeitung | numerisch | schnell wachsendes und großes Datenvolumen; aber kleiner als Bilddaten | Visualisierung, statistische und genetische Analyse, Data Mining |
| Sequenzdaten | | Text | | |
| Experiment-Annotation | Eingabe durch Benutzer | gemischt: Text, numerisch, … | oftmals Freitext, Umfang + Inhalt variiert je nach Experimenttyp | Dokumentation, Selektion, Interpretation |
| Klinische Daten | Studienmanagementsystem | gemischt: Text, numerisch, … | Umfang + Inhalt variiert je nach Studie | Selektion, Interpretation |

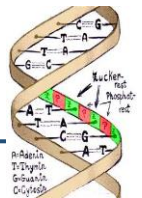


# Bio-Datenquellen



# Verschiedene Arten von Webdatenquellen

- Genomdatenquellen: Ensembl, NCBI Entrez, UCSC Genome Browser, …
  - Objekte: Gene, Transkripte, Proteine etc. verschiedener Spezies
- Objektspezifische Datenquellen:
  - Proteine: UniProt (SwissProt, Trembl), Protein Data Bank (PDB),…
  - Proteininteraktionen: BIND, MINT, DIP, …
  - Gene: HUGO (standardisierte Gensymbole für humanes Genom), MGD, …
  - Pathways: z.B. KEGG (metabolische & regulatorische Pathways)
- Publikationsquellen: Medline / Pubmed (>16 Mio Einträge)
- Ontologien: zur einheitlichen und semantischen Beschreibung (Annotation) der Eigenschaften biologscher Objekte
  - Gene Ontology: Molekulare Funktionen, Biologische Prozesse, Zelluläre Komponenten
  - Ontologie-Sammlung: Open Biomedical Ontologies (OBO)



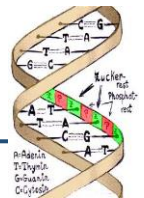
# Datenintegration

- Datenintegration = Informationsintegration
- Korrekte, vollständige und effiziente Zusammenführung von Daten und Inhalt verschiedener, heterogener Quellen zu einer einheitlichen und strukturierten Informationsmenge zur effektiven Interpretation durch Nutzer und Anwendungen*

= Zusammenfügung von Metadaten und Instanzdaten

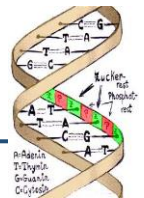
- Ziel: Mehrwert, der sich aus der Kombination von Daten ergibt
  - Bessere Ergebnisse = umfassender, qualitativ hochwertiger, abgesicherter
  - Anfragen, die nur von Daten mehrerer Datenquellen beantwortet werden können

*Leser, Naumann: Informationsintegration, dpunkt.verlag, 2007.



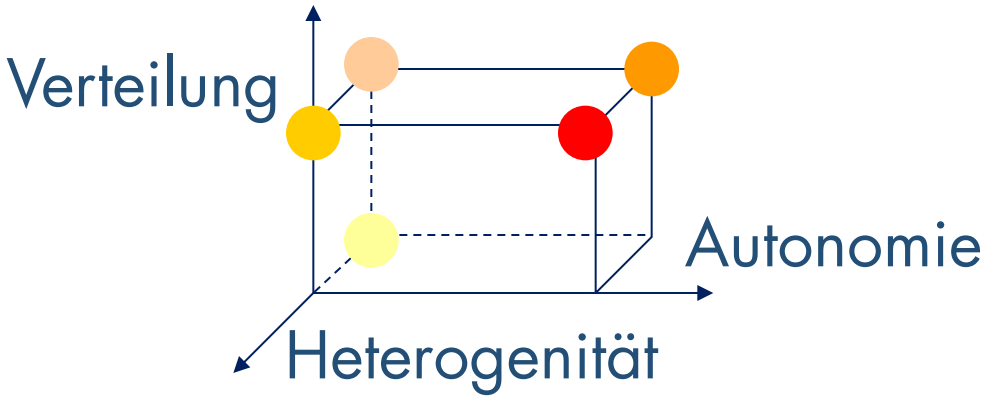
# Probleme bei einer Datenintegration

- Komplexe Fragestellungen → Nutzung vieler verschiedener Datenquellen
- Weite Verteilung der Daten
- Hohe Redundanz (überlappende Quellen)
- Heterogenität der Datenquellen bzgl.
  - Syntax
  - Schema/Struktur
  - Semantik
  - Schnittstellen
- Evolution von Daten und Schemata



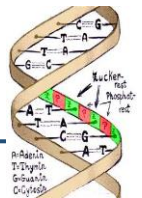
# Verteilung, Autonomie, Heterogenität*

- Architekturraum durch drei orthogonale Aspekte charakterisiert
  - **Verteilung**: physisch, logisch
  - **Autonomie**, zB bzgl. Design, Schnittstellen, Evolution, Zugriff, ...
  - **Heterogenität**




- homogene, zentrale DB
- verteilte DB
- verteilte heterogene DB
- verteilte autonome DB
- verteilt, heterogen, autonom

*Leser/Naumann: Informationsintegration, dpunkt.verlag, 2007, S. 49ff

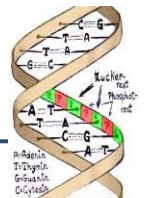


# Architekturvarianten im Überblick

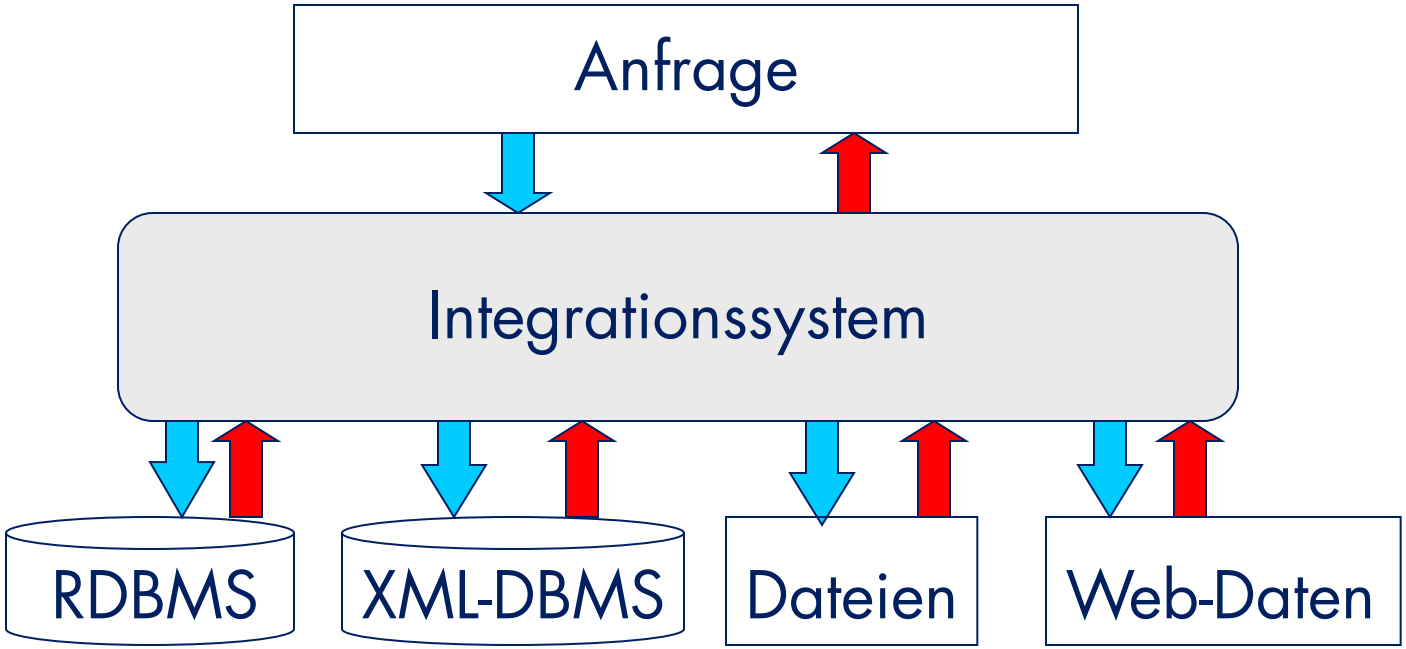
- Monolithische Datenbanken
- Förderierte Datenbanken
- Mediator-basierte Systeme
- Peer-Daten-Management-Systeme

- Suchmaschinen
- Portale

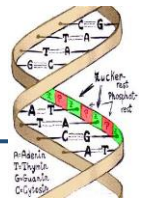
- …



# Datenintegration

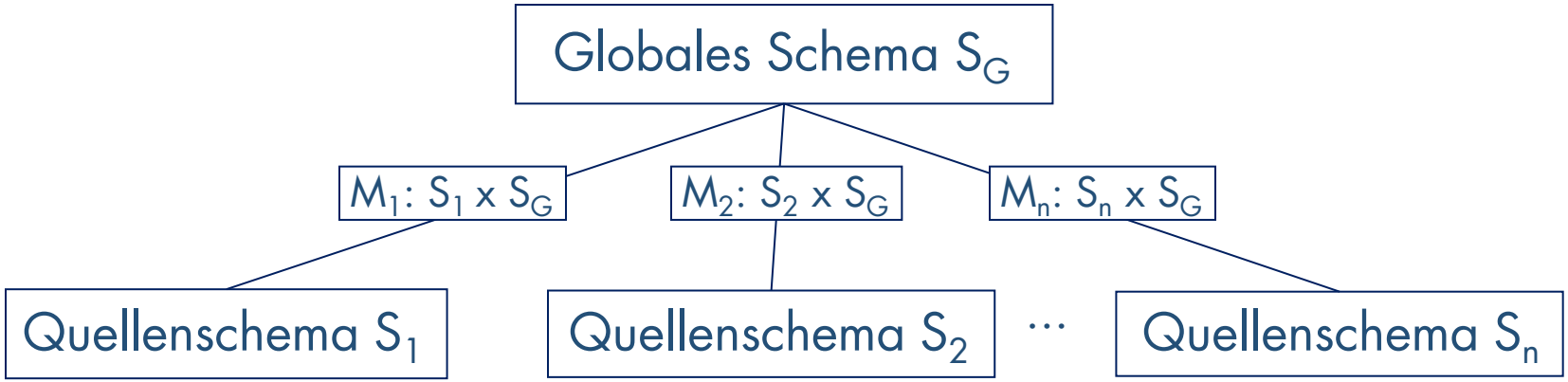



- Zwei orthogonale Aspekte:
  - Schemaintegration
  - Instanzdatenintegration



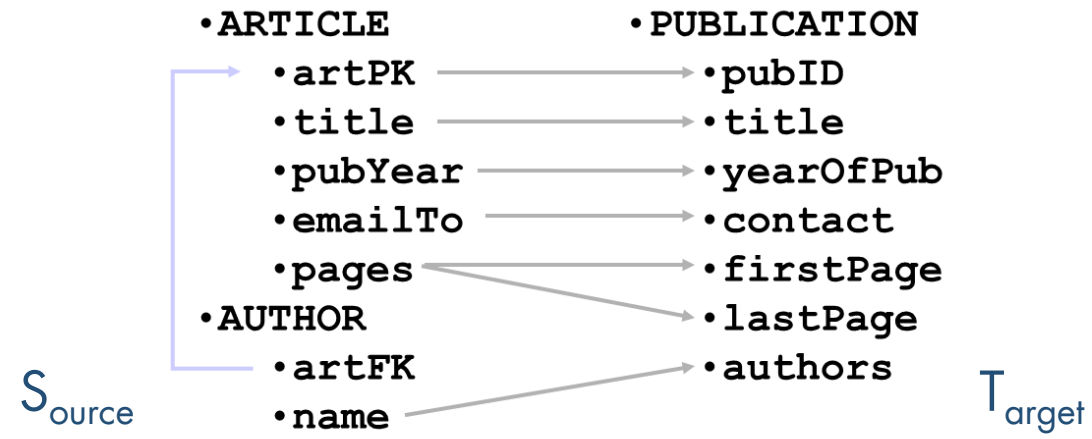
# Schemaintegration

- Schemaintegration = Metadatenintegration
- Ziel: Erstellung einer 'homogenisierten Sicht' (globales Schema) auf die zu integrierenden Datenquellen
- Globales Schema:
  - Enthält alle relevanten Schemaelemente der zu integrierenden Datenquellen
  - Schema-Mappings zwischen globalem und Quellenschemata

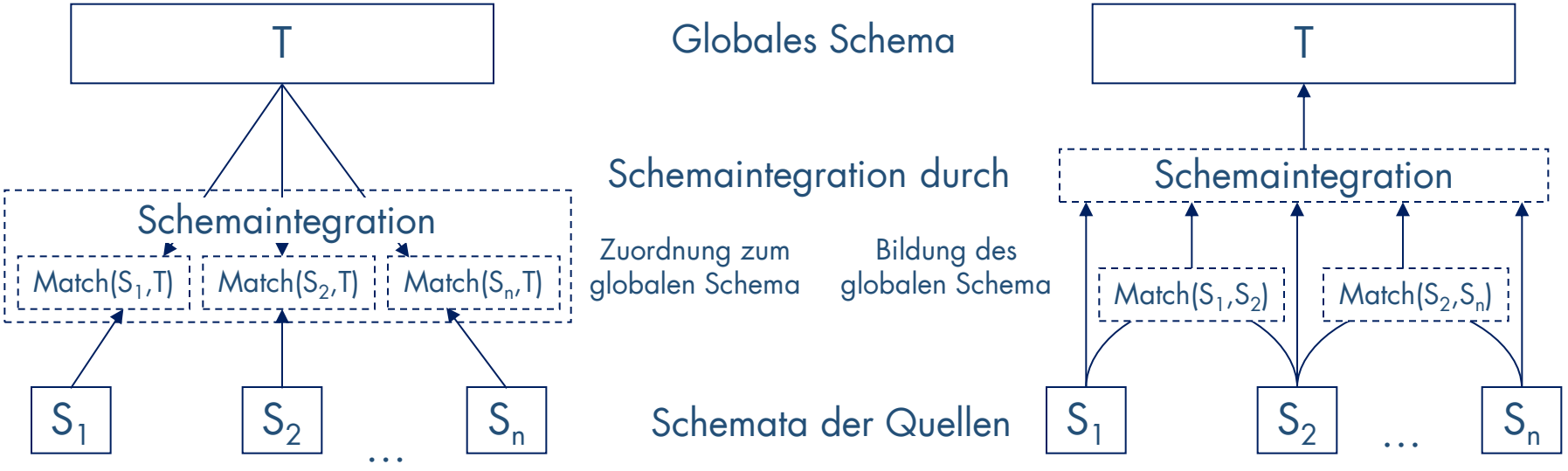


# Schema-Mapping

- Entstehung durch Schema-Matching (Prozess)
  - Eingabe: Schemata $S_{ource}$, $T_{arget} \in S_{1,...,n}$, Algorithmus
  - Ausgabe: Schema-Mapping M: S x T
- Mapping-Eigenschaften:
  - Semantik: oftmals Äquivalenzrelation
  - bidirektional
  - Suche nach 1:1 Beziehung zwischen den Elementen (aber nicht immer möglich: Name ↔ Vorname, Nachname)
  - Problem: Transformationen, z.B. Aggregation von Daten

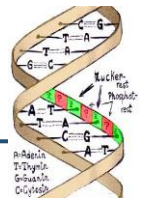


# Top-Down vs. Bottom-Up



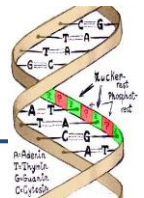

**Top-Down-Integration**

**Bottom-Up-Integration**

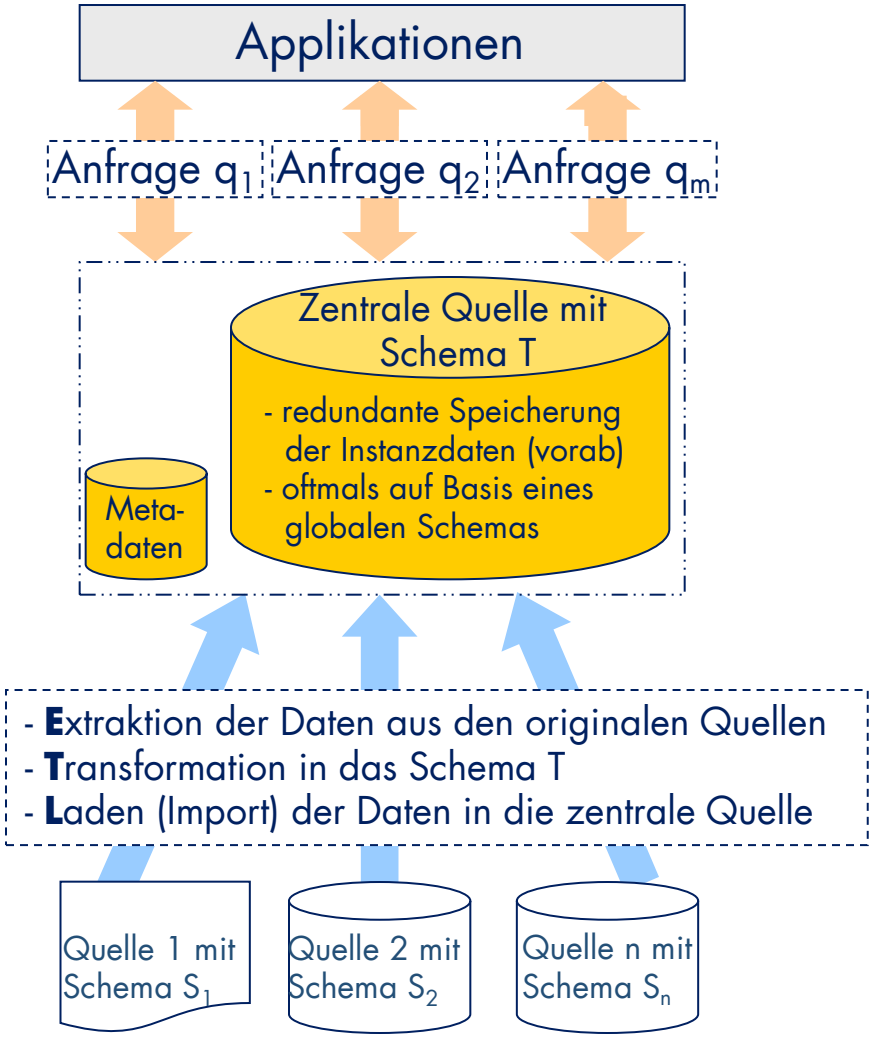


# Instanzdatenintegration

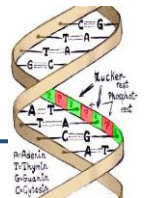
- Instanzdatenintegration: Zusammenfügen der Daten aus den verschiedenen Datenquellen
  - Materialisiert: Prozess der Vorverarbeitung (ETL)
  - Virtuell: zur Beantwortung einer Anfrage
- Basis: Schema-Mappings



# Physische/materialisierte Integration

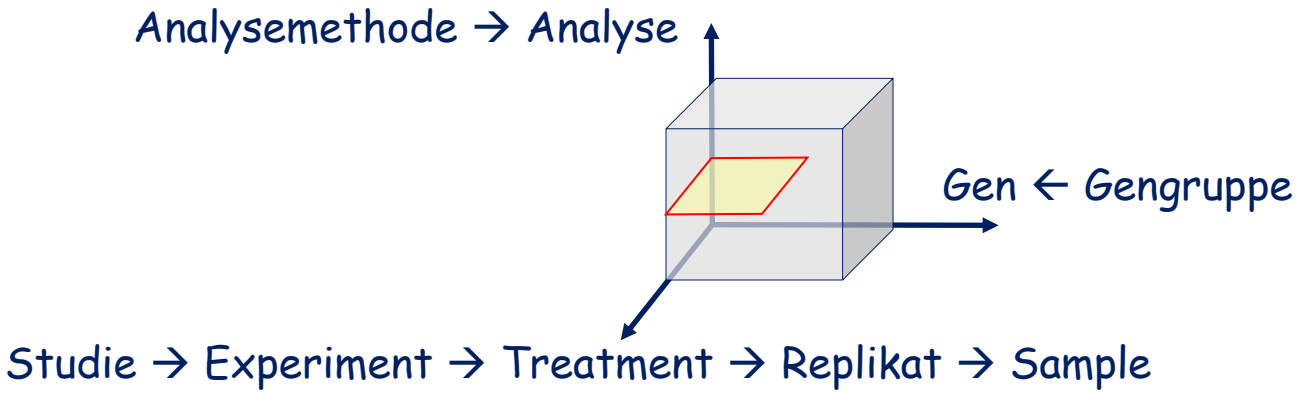



- Globales Schema mit zentraler Datenhaltung
- Separate Extraktion, Transformation und Laden der Daten
- Formen
  - Data Warehouse mit mögl. zusätzlicher Data-Mart-Schicht

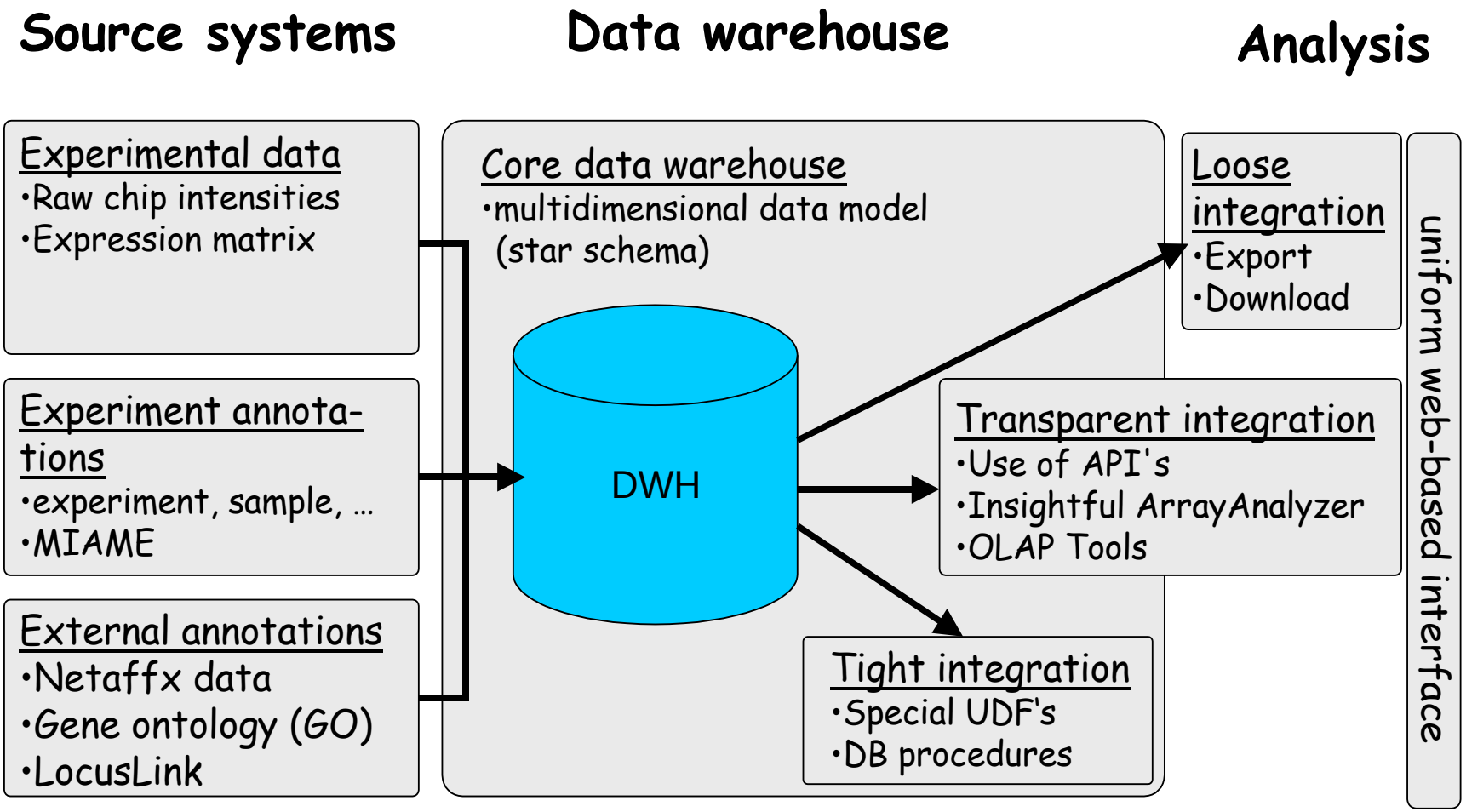


# Multidimensionales Datenmanagement

- Dimensionen vs. Fakten
- Skalierbarkeit und Erweiterbarkeit
  - Hinzufügen neuer Chip-Typen, Gene, Analysemethoden und assoziierter Fakten ohne Schemaänderungen
  - Hinzufügen neuer Dimensionen und Fakten
- Multidimensionale Analyse
  - Einfache Selektion, Aggregation und Vergleich von Werten
- Basis für anspruchsvolle Analysemethoden
  - Fokussierte Selektion und Erstellung von Matrizen



# Beispiel: GeWare

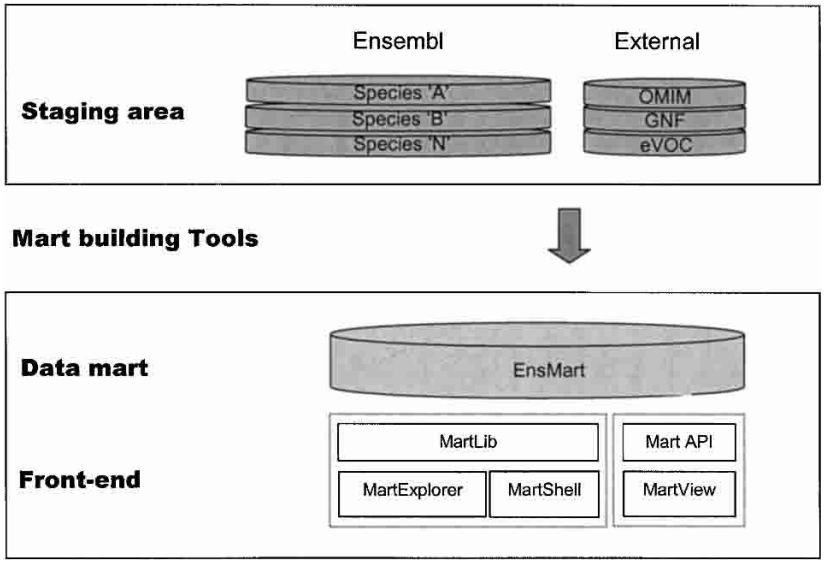



Quelle: Do, H.H., Kirsten, T., Rahm, E.: Comparative Evaluation of Microarray-based Gene Expression Databases. Proc. 10. Fachtagung Datenbanksysteme für Business, Technologie und Web (BTW 2003), Leipzig, Feb. 2003

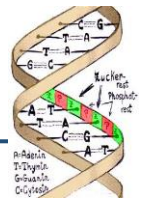
# Beispiel: Ensembl BioMarts

- Data-Warehouse-basierte Datenintegration
- Data Marts auf Basis von Ensembl
- Gezielte Suche und Wiedergabe von
  - Ensembl-eigenen biol. Objekten: Gene, Transkripte und Polypetide
  - Assoziierten Annotationen
  - Referenzierte biol. Objekte
- Multidimensionales Schema
  - Ensembl-eigene biol. Objekte als "Fakten"
  - Instanz-Mappings zu referenzierten Objekten als beschreibende Dimensionen
- Spezies-spezifische Data Marts (verknüpft über Homologie-Mappings)

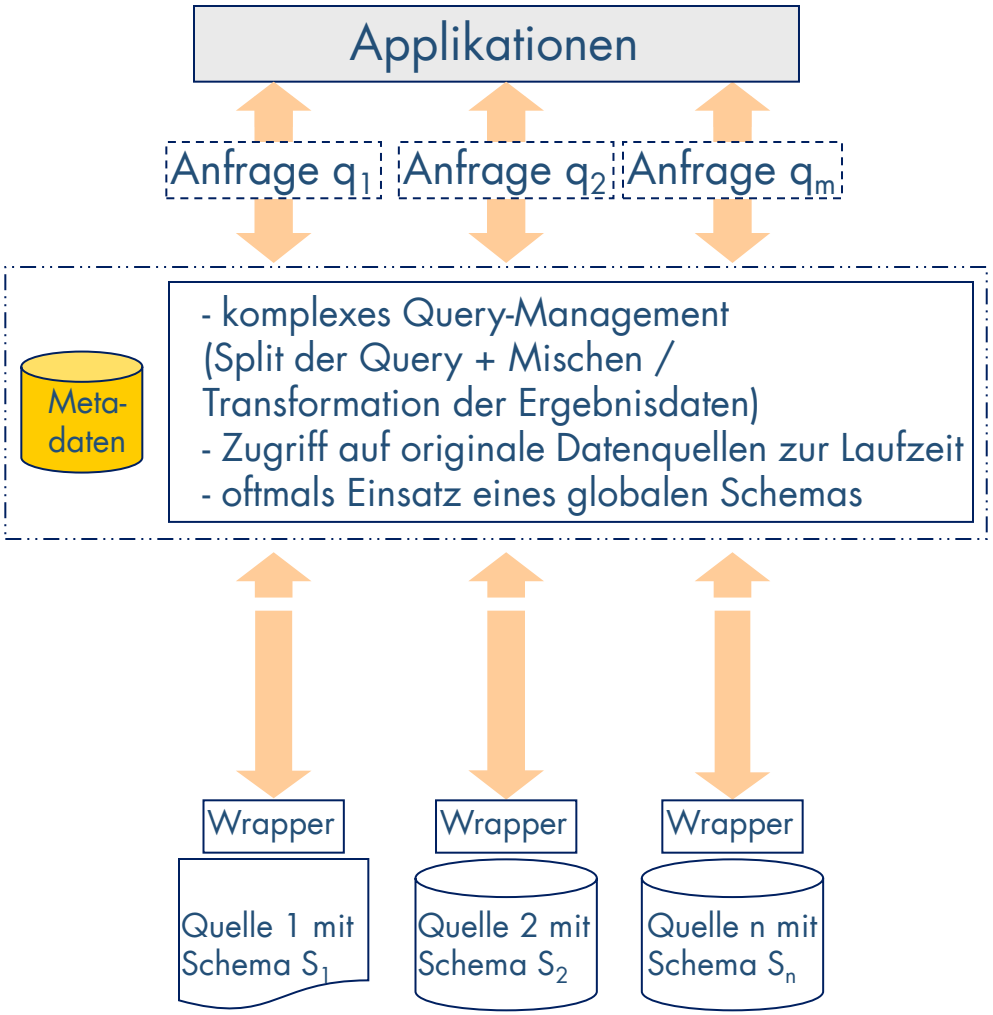
**EnsMarts Architektur**



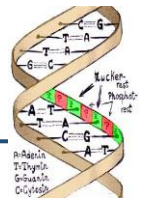

- http://www.ensembl.org/biomart



# Virtuelle Integration

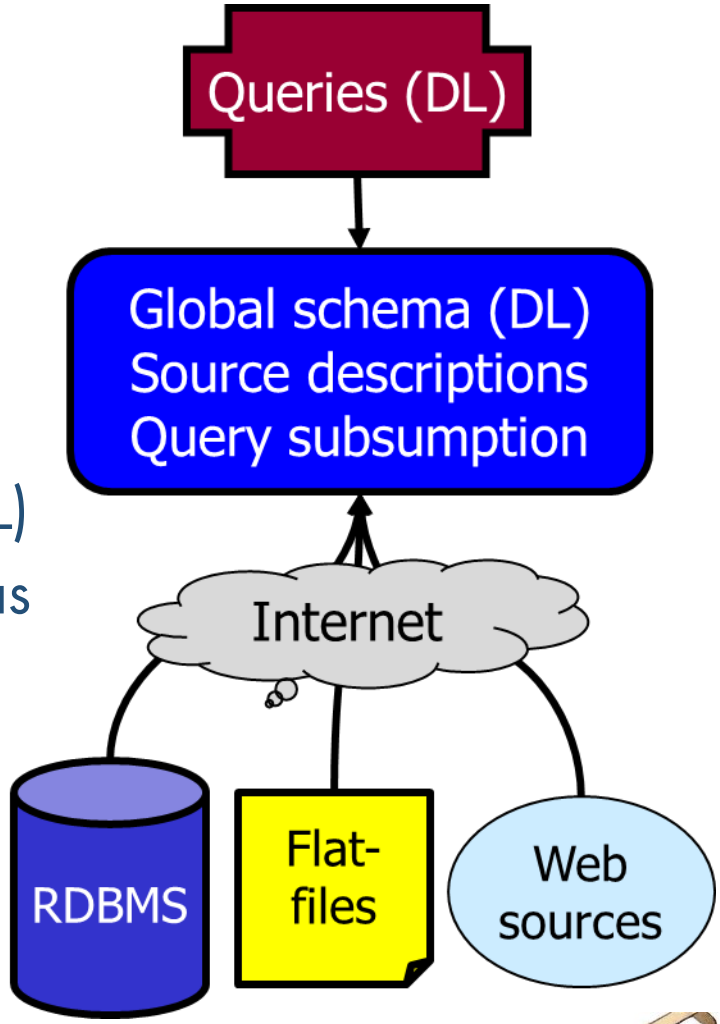



- Globales Schema mit Mappings zu den Quellschemata (Metadaten)
- Transparenter Zugriff auf originale Datenquellen
- Wrapper
  = quellenspezifische API
- Formen
  - förderierte DBMS
  - Mediatoren



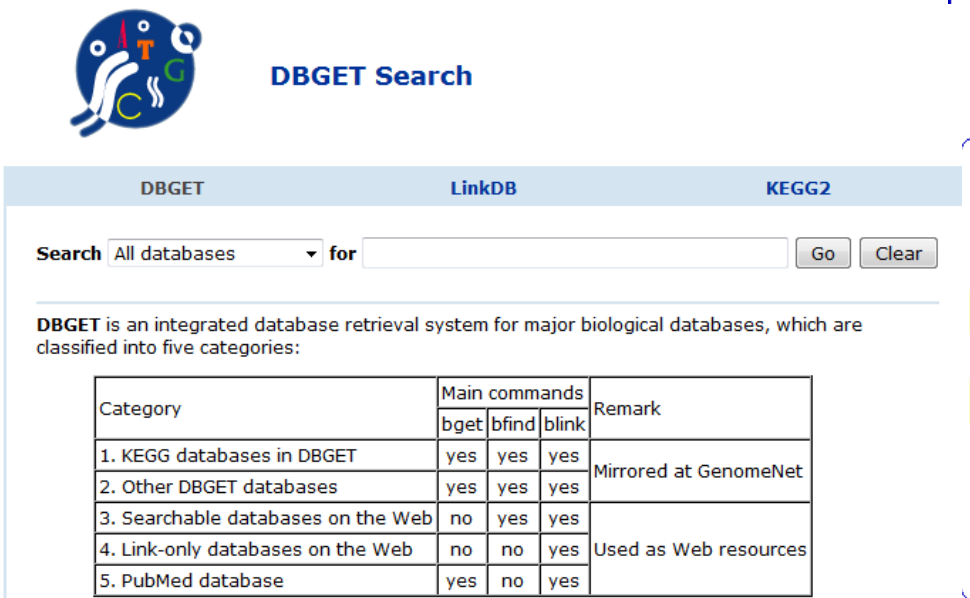
# TAMBIS

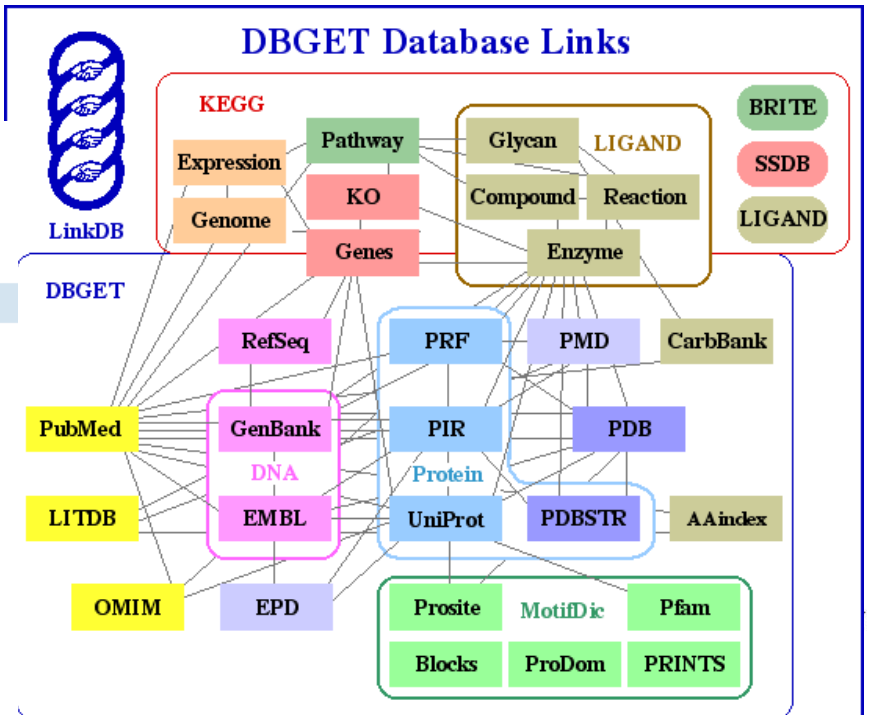
- Architektur
  - Mediator-basiert
  - Fokus: Integration durch Ontologie-basierte Anfrageverarbeitung und -umformulierung
- Features
  - Nutzung von Beschreibungslogik (DL)
  - Semantische Integration der Schemas
  - Keine Dublettenbehandlung, keine Datenfusion



# Beispiel: DBGET / LinkDB

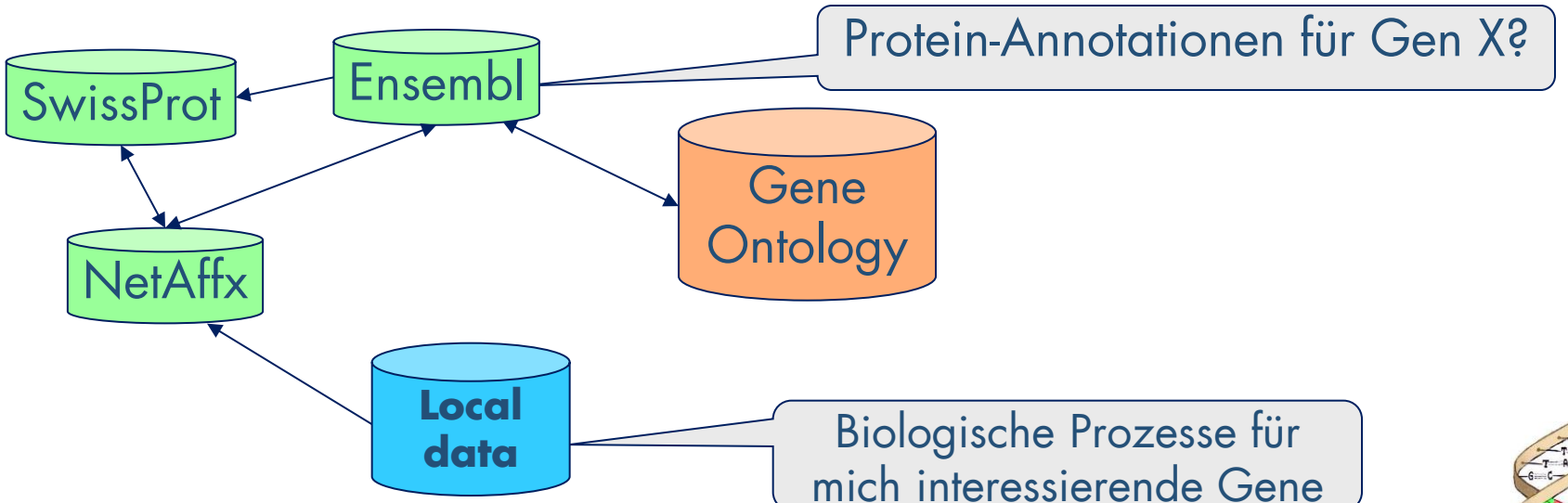
- Integration mit verfügbaren Web-Links
- Web-Link = URL einer Datenquelle + ID (accession number) des Objekts
- Einfacher Integrationsansatz
  - Wenig Integrationsaufwand
  - Aber: Analyse eines Objekts zu einer Zeit
- DBGET / LinkDB: Sammlung von Links zwischen verschiedenen Quellen
  - Verwaltung von quellenspezifischen Objektreferenzen (ID) u. Instanz-Mappings
  - Keine expliziten Mapping-Typen



# P2P-like Integration

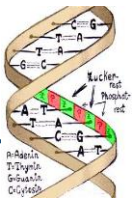
- Virtuell, kein globales Schema
- Bidirektionale Instanz-Mappings zwischen autonomen Datenquellen
  - Wiederverwendung dieser Mappings (Menge von Korrespondenzen zwischen Objekten)
- Anfragen an eine Datenquellen u. deren Propagierung an relevante Peers
- Einfachere Hinzufügung neuer Datenquellen
- Unterstützung lokaler Datenquellen



# Beispiel: BioFuice

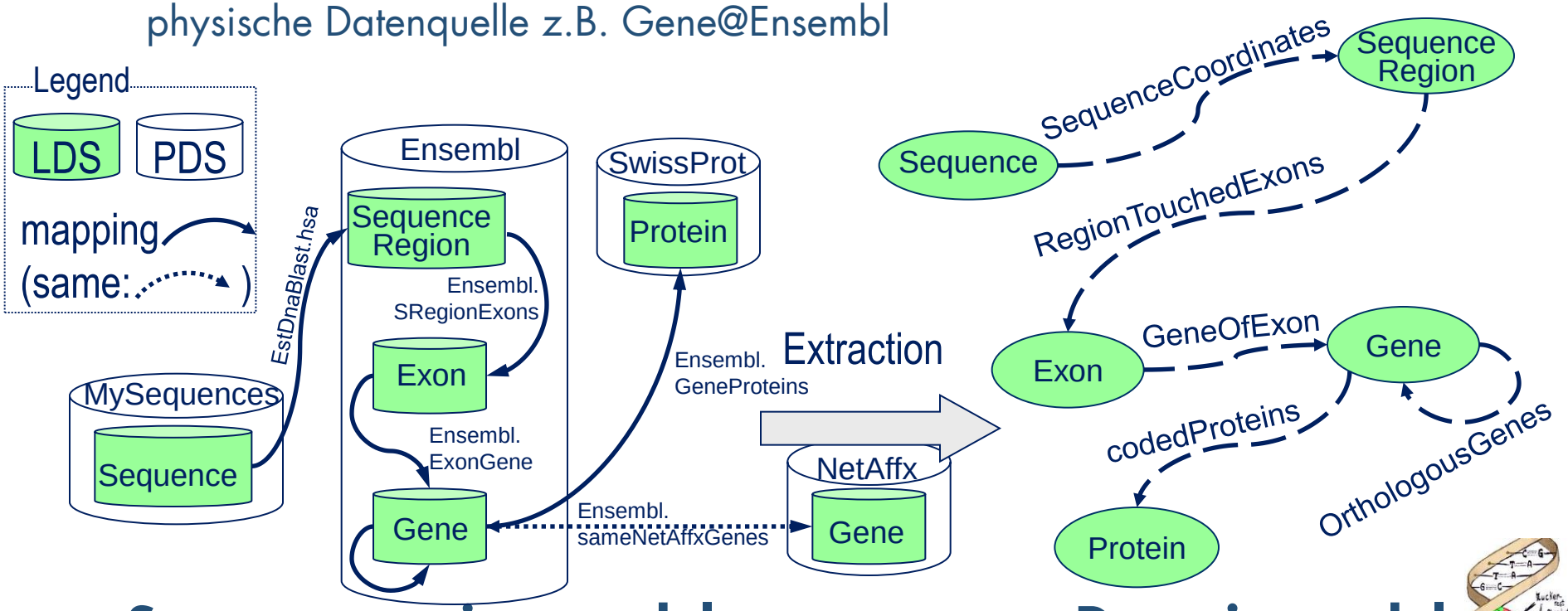
- ***Bio****informatics information* ***f****usion* ***u****tilizing* ***i****nstance* ***c****orrespondences and p****ee****r mappings*
- Mediator
  - Zur Steuerung der Mapping- und Operatorausführung
  - Nutzung eines anwendungsspezifischen **semantischen** Domänenmodells
- **High-level Operatoren** (mengenorientiert), z.B.
  - Single source: queryInstances, searchInstances, …
  - Navigation: traverse, map, compose, …
  - Navigation + Aggregation: aggregate, aggregateTraverse, …
  - Universell: diff, union, intersect, …

*Kirsten, T; Rahm, E: *BioFuice: Mapping-based data integration in bioinformatics.* Proc. 3rd Intl. Workshop DILS, July 2006



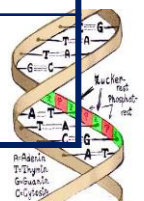
# Metadatenmodelle

- Verwendung durch Mediator zur Mapping- und Operatorausführung
- **Domänenmodell** stellt relevante Objekttypen und Beziehungen (=Mappingtypen) zwischen ihnen bereit
- Physische Datenquelle (PDS): Öffentliche, private und lokale Daten (Genliste, …), Ontologien …
- Logische Datenquelle (LDS): Bezieht sich auf einen Objekttyp und eine physische Datenquelle z.B. Gene@Ensembl



# Vergleich physische vs. virtuelle Integration

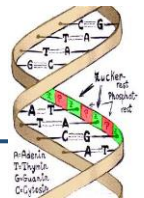
| | Physische I. (Warehouse) | Virtuelle Integration | |
|---|---|---|---|
| | | Mediatoren | Peer Data Mgmt |
| Schemaintegration | A priori | A priori | Nicht zwingend |
| Instanzdatenintegration | A priori | Zur Laufzeit der Anfrage | Zur Laufzeit der Anfrage |
| Unterstützung der Datenqualität | + | o | o/- |
| Analyse großer Datenmangen | + | - | - |
| Resourcen-anforderungen (HW) | - | o | o |
| Datenaktualität | o | + | + |
| Autonomie der Datenquellen | o | + | + |
| Skalierbarkeit (#Datenquellen) | - | - | o |



# Anwendung: Integration klinischer Daten

- Studienmanagementsystem zur einheitlichen Erfassung der klinischen Daten (Metadaten + Instanzdaten)

- Anforderungen
  - **Datenintegration:** Patient-bezogene Daten + Chipbasierte genetische Daten
  - **Privacy** Aspekte: Keine „eine Person identifizierende Daten“
  - Nutzung **existierender** Software für Datenmanagement und Analyse



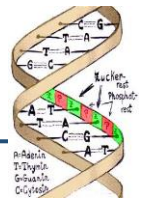
# Beispiel: LIFE-Studie*

Leipziger Forschungszentrum für Zivilisationserkrankungen

- Leipziger Forschungszentrum für Zivilisationserkrankungen
- Sächsische Exzellenzinitiative
- Ziel: Evaluierung klinischer und genetischer Faktoren für ausgewählte Zivilisationserkrankungen
  - Einbeziehung von Lebensstil, soziales & Wohnumfeld
  - Rekrutierung aus der Leipziger Bevölkerung

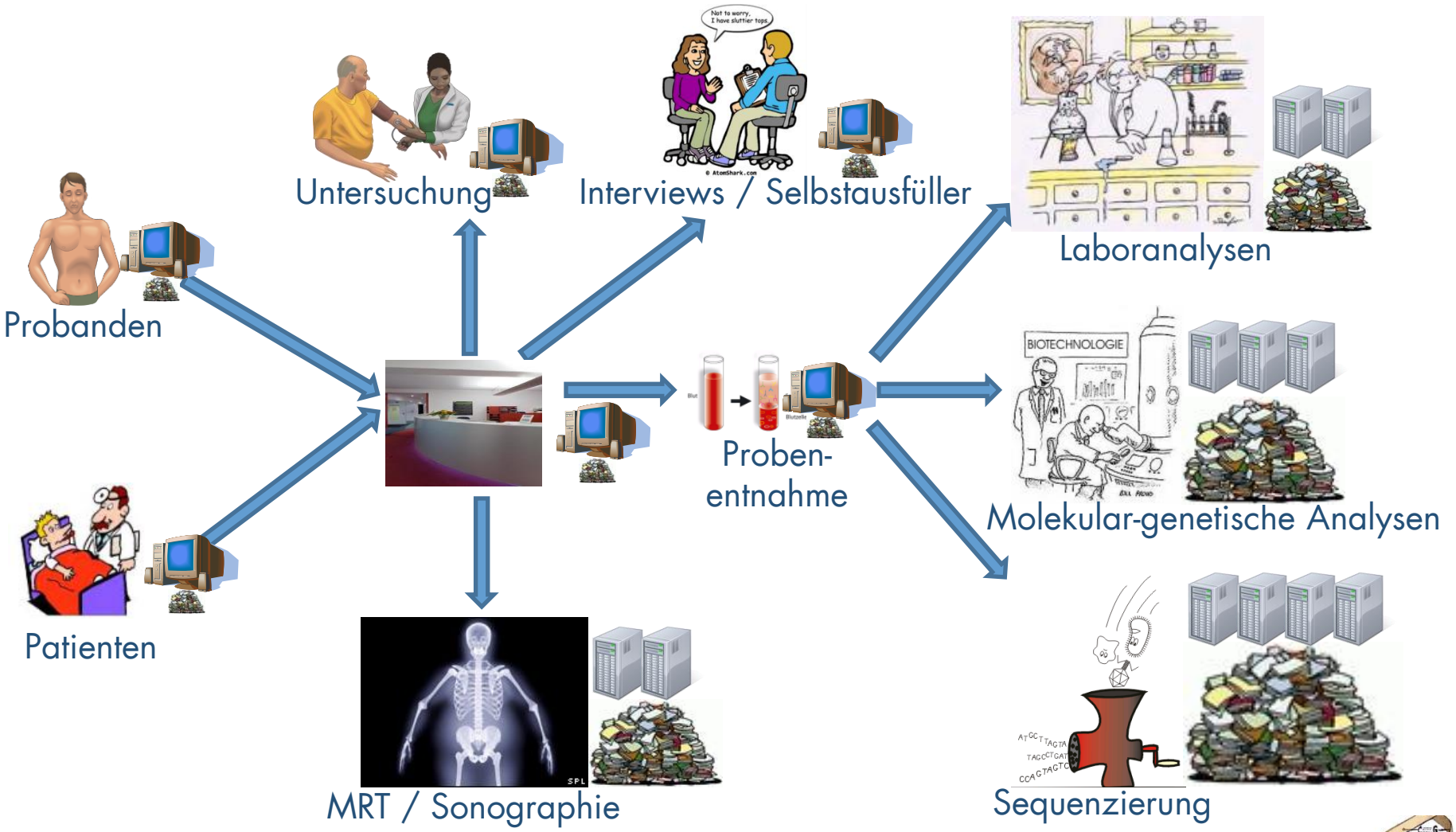
**> 22.000 Teilnehmer (Stand 11/2014)**

- Erwachsene
  - Bevölkerungsquerschnitt: **10.000**
  - Herzbezogene Erkrankungen: **8.000**
- Kinder
  - Bevölkerungsquerschnitt: **3.000**
    inkl. Adipositas, Neugeborene
  - Depressive Erkrankungen: **1.000**



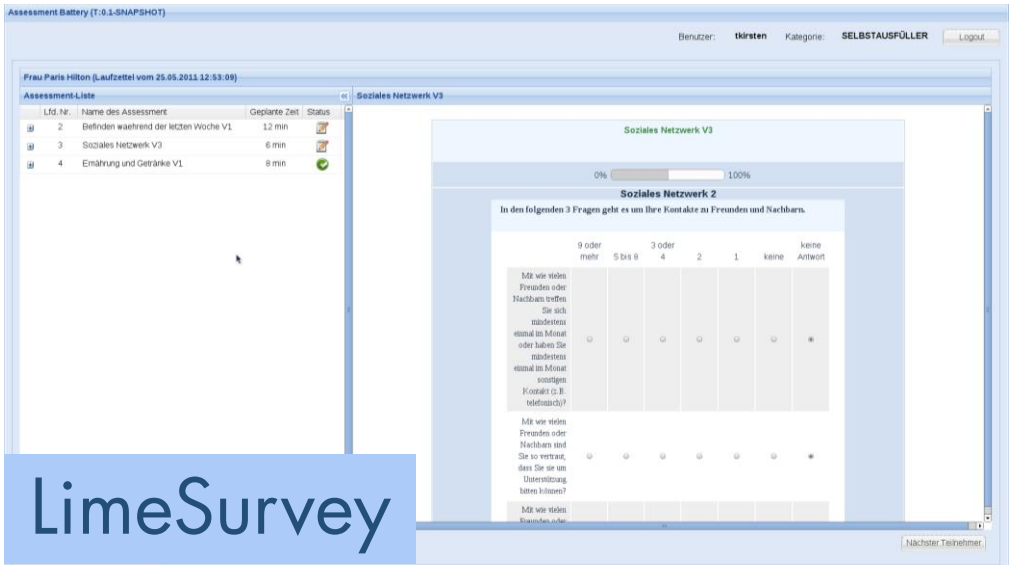
** Folien zu LIFE: von Dr. Toralf Kirsten, Leipzig Research Centre for Civilization Diseases (LIFE)*

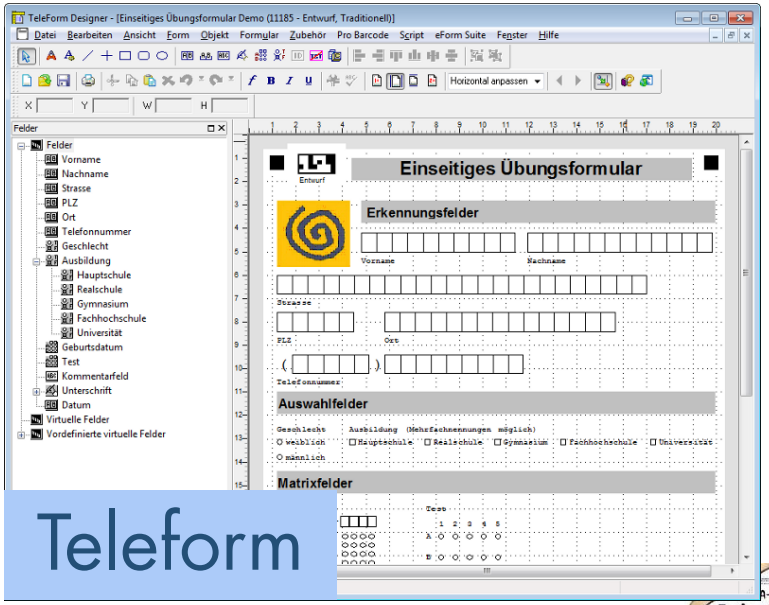
# Komplexe Workflows in LIFE



# Heterogenität: Strukturierte Daten

- Interviews, Fragebögen, Untersuchungen → RDBMS
- Ausgewählte Untersuchungen → Desktop DBS, Excel
- Instrumente in unterschiedlichen Versionen und Varianten
- Datenerfassung: LimeSurvey (875 Versionen) & Teleform (176 Versionen) mit durchschnittlich ca. 60 Items



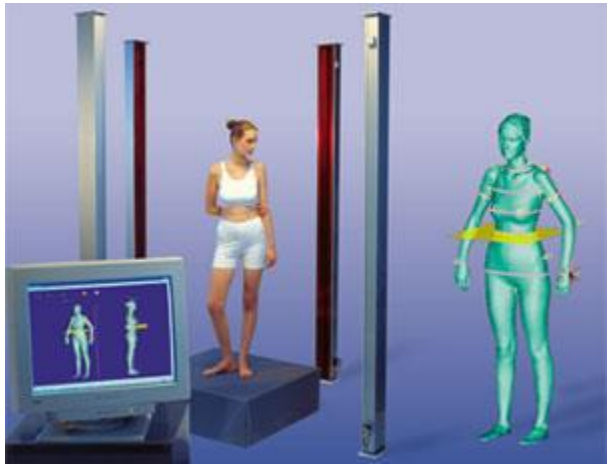



# Heterogenität: Strukturierte Daten II

- „ Ausnahme“ Labordaten
  - Herkunft:  Analyse von Proben (Blut, Speichel, Haar, Urin, …)
  - Nutzung des HL7-Formats (Health Level 7 Standard) zur Datenübertragung
  - Keine explizite Typisierung: Datentyp String
  - Wechselnde Parameter und -profile (derzeit 88 Parameterprofile)

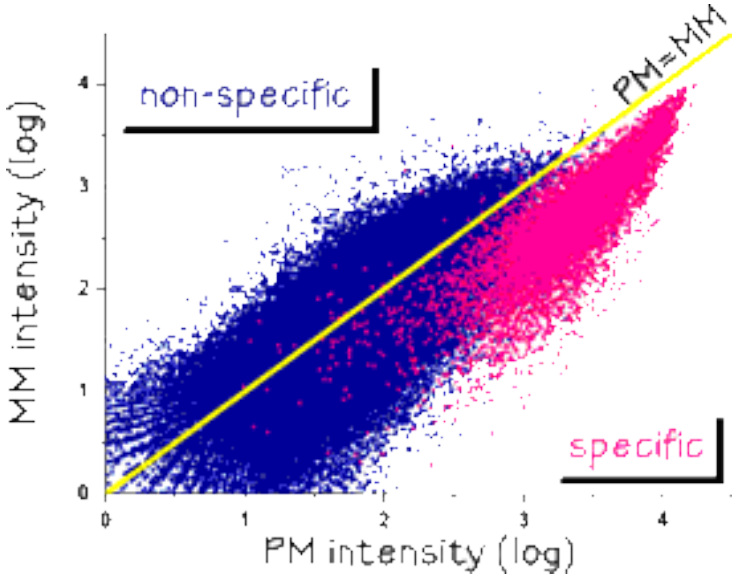
# Heterogenität: Unstrukturierte Daten

- Vornehmlich Gerätedaten von apparativen Untersuchungen
- Kaum Direktzugriff auf Daten in med. Geräten möglich
  - Manueller Export pro Untersuchung (teilweise als batch)
  - Evtl. Datentransport per USB-Stick
  - Notwendiges Preprocessing (Originaldaten → Rohdaten)
  - Formate: proprietär, Excel

# Heterogenität: Unstrukturierte Daten II

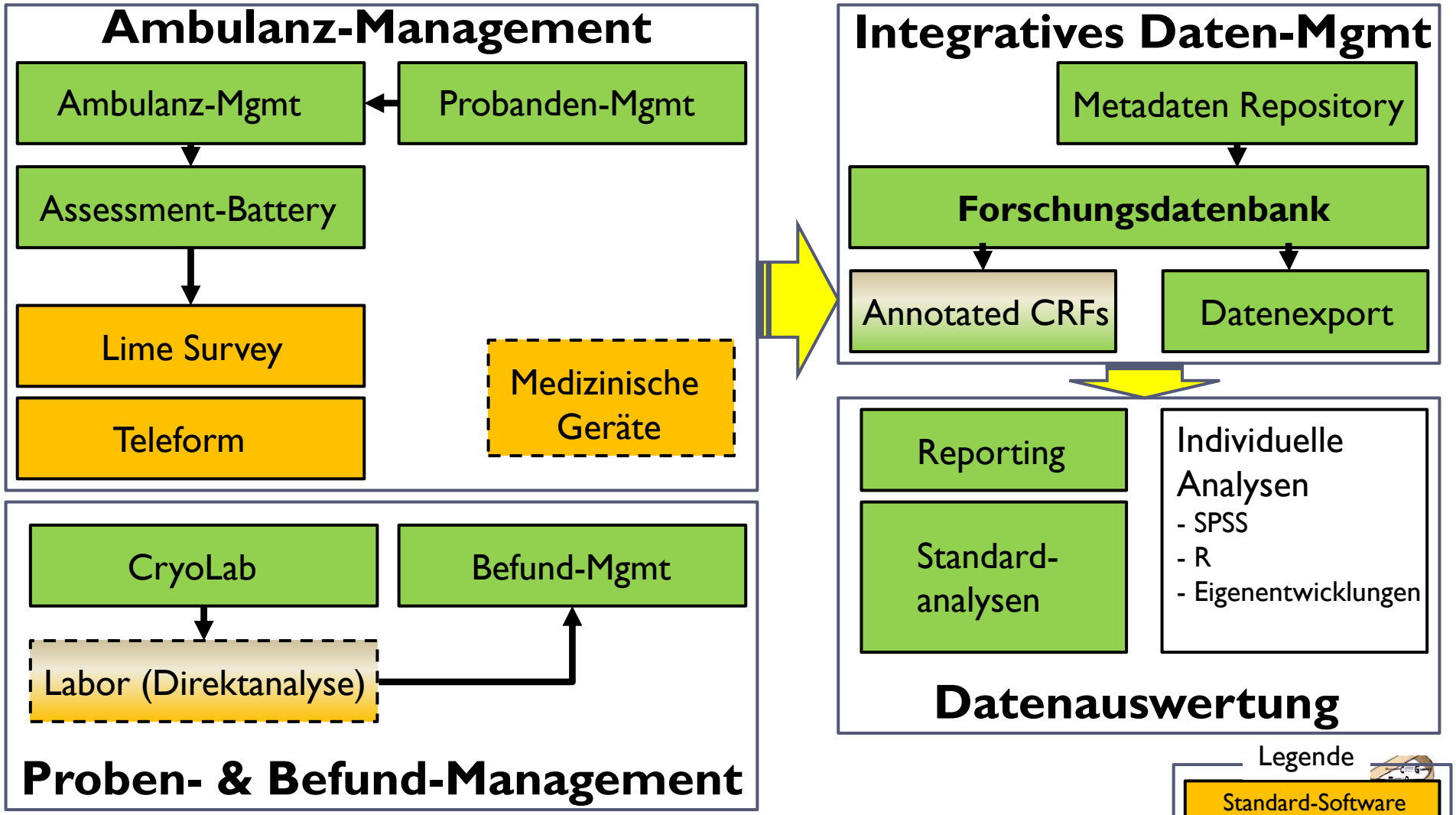
- Genetische Daten
  - Genetische Daten: Messergebnisse aus wet-lab Experimenten zur Bestimmung der Genexpression, Mutation (SNP) und Sequenz (Next Generation Sequencing)
  - Hohes Datenvolumen an Rohdaten
  - Hoher und schnell wachsender Umfang an Analysedaten



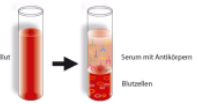
# Heterogenität: Unstrukturierte Daten III

- Multimediale Daten
  - Formen: Bilddaten, Filme, Akkustikaufzeichnungen
  - Verwaltung an verschiedenen Standorten mit unterschiedlichen Systemen: PACS, Dateisystem, …
  - Preprocessing
    - Manuelle Befundung, z.B. von MRT-Bildern
    - Abhören von Interviewmitschnitten
    - …

# IT-Systemlandschaft im Überblick



# Integratives Forschungsdaten-Management


**Probanden Mgmt**

**Lime Survey**

**Teleform**

**Proben-Mgmt. CryoLab**

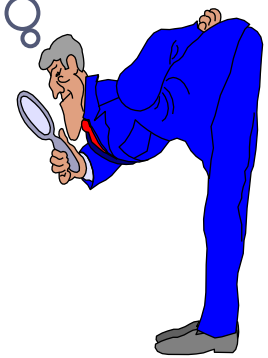
**Befunddaten ILM**

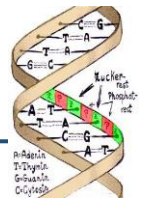
**Geräte**

**Diverse …**

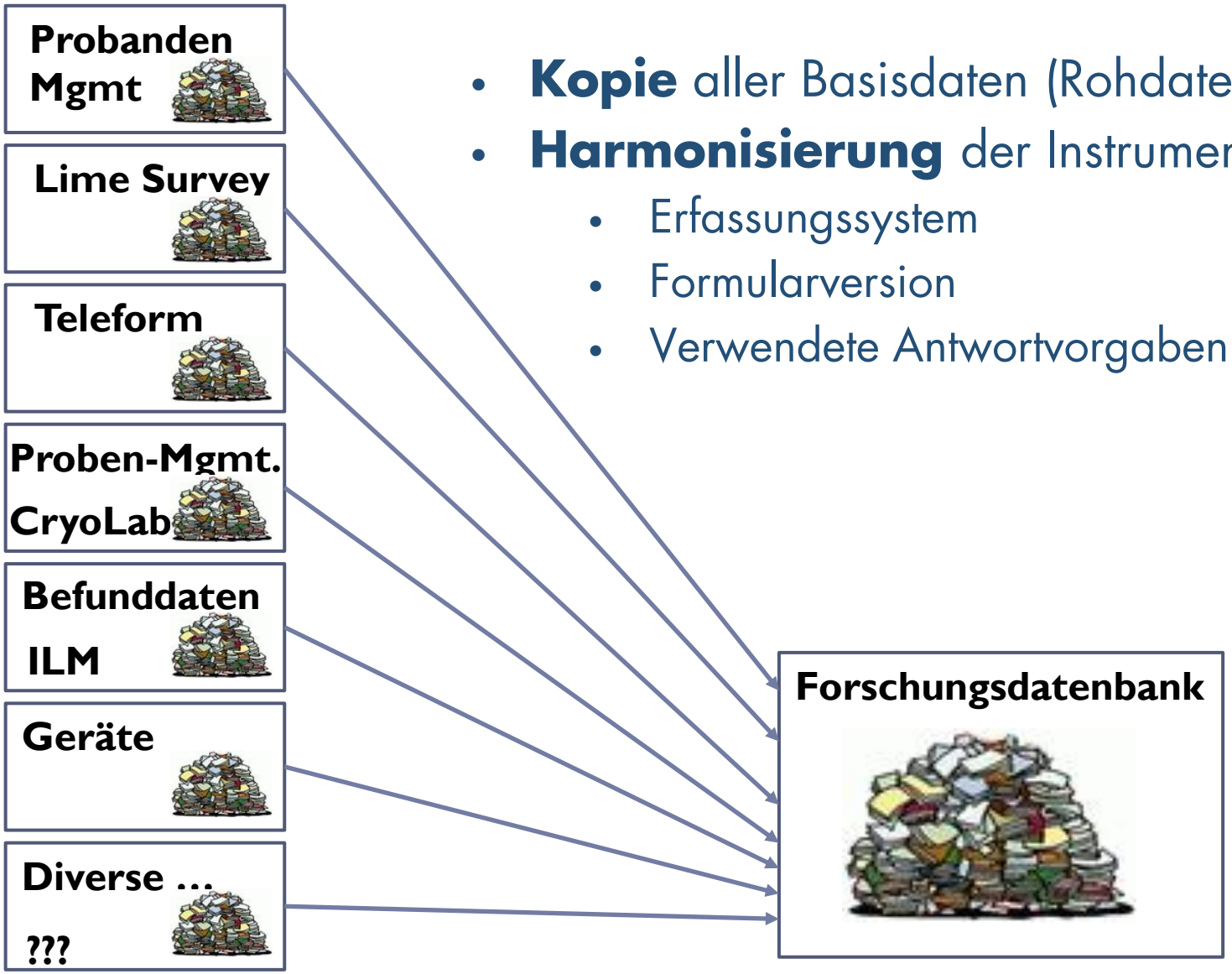
**???**

**Korreliert Stadtteil mit Bildungsniveau und Alter?**



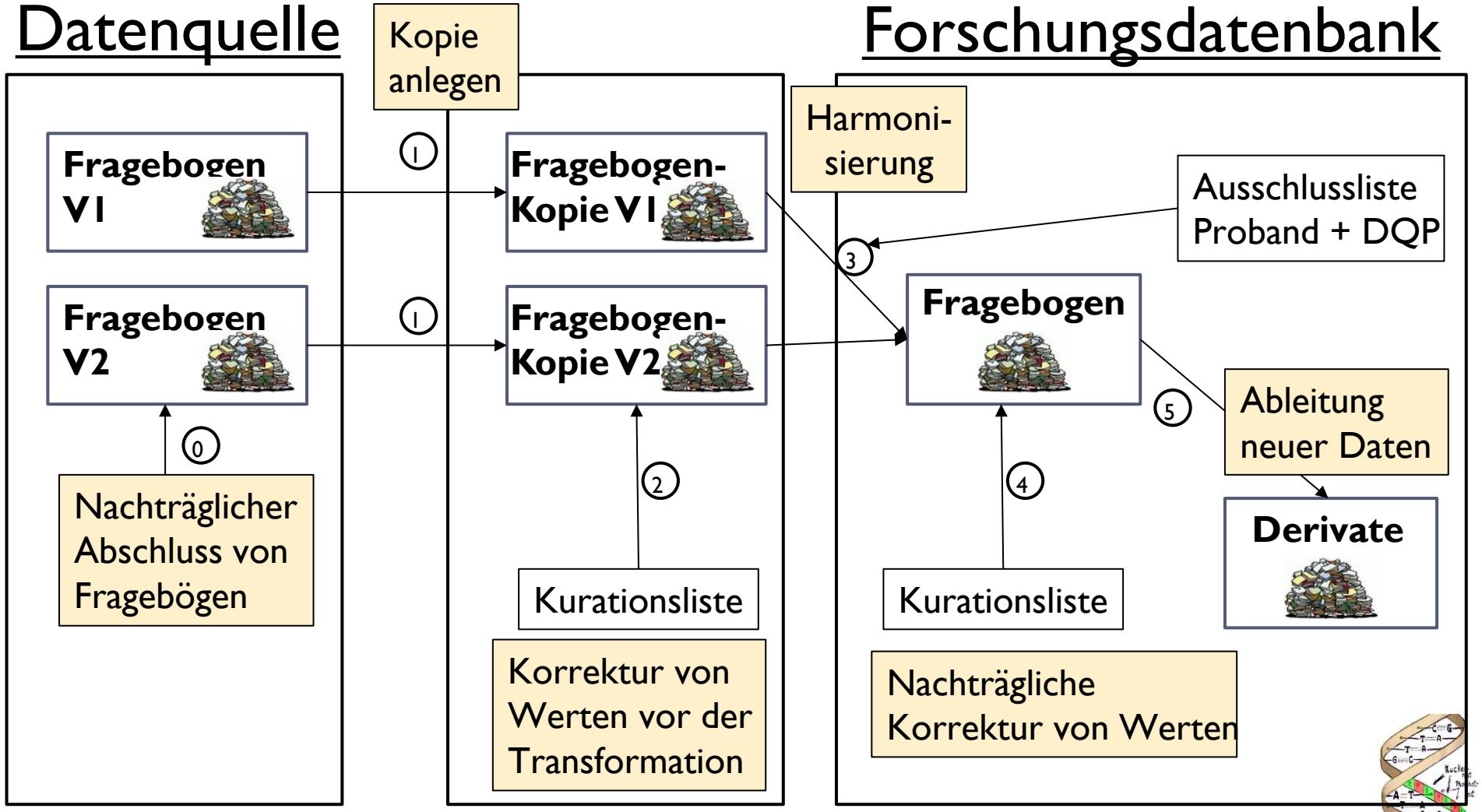


# Integratives Forschungsdaten-Management



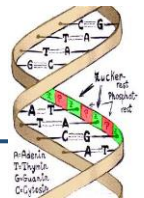

- **Kopie** aller Basisdaten (Rohdaten, …)
- **Harmonisierung** der Instrumente bzgl.
  - Erfassungssystem
  - Formularversion
  - Verwendete Antwortvorgaben

# Prozess der Datenüberführung, -harmonisierung und -aufbereitung



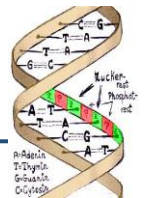
# Datenanalyse

- Anwendung unterschiedlicher Algorithmen, projektspezifisch je nach Fragestellung
- Voraussetzung: Datenbereitstellung
  - DBS-Direktzugriff
  - Reporter-System (vordefinierte Reports)
  - Datenexporte
- Unterschiedliche Analysesoftware: R, SPSS, SAS, Excel, …
- Geplant: Rückführung der Ergebnisse in die Forschungsdatenbank



# Zusammenfassung Daten-Management LIFE

- Komplexe Prozesse in Ambulanzen an Kliniken
- Hohe Heterogenität der Daten
- Sehr hoher Grad an Eigenentwicklungen
- Integratives Forschungsdaten-Management
  - Ziel: Integration relevanter Daten aus Erfassungssystemen
  - Basis: Metadaten und Beziehungen zwischen ihnen
- Relationale Datenbanktechnologie für Management der Fragebogen- und Befunddaten ausreichend



# Zusammenfassung

- Viele verschiedene Datenquellen mit unterschiedlichem Fokus
- Hauptsächlich Heterogenität als Integrationsbarriere

- Virtuelle vs. physische Integration
  - Virtuell: Mediator-Systeme, P2P-like Integration
  - Physisch: Data Warehouses
- Schemaintegration: Top-down vs. Bottom-Up
- Anwendungen
  - Integration im Kontext molekular-genetischer Experimente
  - Integration von klinischen / epidemiologischen Studiendaten

- Mehr Details: VL Datenintegration