Andreas Holzinger

VO 709.049 Medical Informatics

04.11.2015 11:15-12:45

Lecture 04

Biomedical Databases: Data Acquisition, Storage, Information Retrieval and Use

a.holzinger@tugraz.at

Tutor: markus.plass@student.tugraz.at

http://hci-kdd.org/biomedical-informatics-big-data

Schedule

- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 2. Back to the future: Fundamentals of Data, Information and Knowledge
- 3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)
- **4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use**
- 5. Semi structured and weakly structured data (structural homologies)
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- 9. Intelligent Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Info Systems: System Design, Usability & Evaluation

Keywords of the 4th Lecture

- Bayes' Rule
- Biomedical data warehouse
- Business hospital information system
- Clinical workflow
- Data integration
- Enterprise data modeling
- Information retrieval (IR)
- Probabilistic Model
- Quality of information retrieval
- Set theoretic model
- Vector Space Model (VSM)

Advance Organizer (1/4)

- **Business intelligence (BI)** = a type of application software designed to report, analyze, and present information on real-time management dashboards, i.e., integrated displays of metrics that measure the performance of a system;
- **Cassandra** = an open source and free database management system designed to handle huge amounts of data on a distributed system. This system was developed at Facebook and is now managed as a project of the Apache Software foundation.
- **Cladogram** = a phylogenetic tree to show evolutionary relationships with species represented by nodes and lines of descent represented by links (unrooted or rooted);
- **Classification system** = arbitrary in nature, there is no standard measure of difference that defines a species, genus, family, or order;
- **Cloud computing** = a computing paradigm in which highly scalable computing resources, often configured as a distributed system, are provided as a service
- **CPOE (Computerized physician order entry)** = a process of electronic entry of medical practitioner instructions for the treatment of patients (particularly hospitalized patients) under his or her care;
- **Data Mart (DM)** = access layer of a data warehouse environment that is used to get data to the users. The DM is a subset of the DW, usually oriented to a specific business line or team to provide data to users usually through business intelligence tools;
- **DBGET** = a data retrieval tool (simpler than ENTREZ) from the Kyoto University, which covers more than 20 databases, related to the Kyoto Encylopedia of Genes and Genomes
- **Distance matrix method** = work by two most closely related taxa in a distance matrix and clustering them;

Advance Organizer (2/4)

- **EnsEMBL** = database format;
- **ENTREZ** = a dedicated data retrieval tool;
- **Extract, transform, and load (ETL)** = Software tools used to extract data from outside sources, transform them to fit operational needs, and load them into a database or data warehouse;
- **Federated data base system** = type of meta-database management system, which integrates multiple autonomous database systems into a single federated database;
- **Genetic algorithm** = a technique used for optimization inspired by the process of natural evolution or "survival of the fittest." Often described as a type of "evolutionary algorithm," these algorithms are well-suited for solving nonlinear problems;
- **Genomes OnLine Databases (GOLD)** = a general genomics gateway;
- **Hadoop** = An open source (free) software framework for processing huge datasets on certain kinds of problems on a distributed system. Its development was inspired by Google's MapReduce and Google File System.
- **Hbase** = An open source (free), distributed, non-relational database modeled on Google's Big Table. It was originally developed by Powerset and is now managed as a project of the Apache Software foundation as part of the Hadoop.
- **Information Extraction (IE)** = automatic assignment of meaning to elementary textual entities and complex structured information objects;
- **Information Retrieval (IR)** = indexing and retrieval of information in documents;
- **KEGG** = Kyoto Encyclopedia of Genes & Genomes, a combined database containing information on types of proteins (receptors, signal transduction components, enzymes)

Advance Organizer (3/4)

- **MapReduce** = A software framework introduced by Google for processing huge datasets on certain kinds of problems on a distributed system.32 Also implemented in Hadoop;
- **Mashup** = An application that uses and combines data presentation or functionality from two or more sources to create new services. These applications are often made available on the Web, and frequently use data accessed through open application programming interfaces or from open data sources;
- **MEDLINE** = Literature data bank;
- **Metadata** = Data that describes the content and context of data files, e.g., means of creation, purpose, time and date of creation, and author;
- **MMMDB** = Molecular Modeling Database, can be accessed at the NCBI (National Center for Biotechnology information) using ENTREZ;
- **Natural language processing (NLP)** = a set of machine learning techniques from computer science and linguistics that uses computer algorithms to analyze human (natural) language;
- **Neural networks** = computational models, inspired by the structure and workings of biological neural networks (i.e., the cells and connections within a brain), that find non linear patterns in data;
- **Non-relational database** = A database that does not store data in tables (rows and columns). (In contrast to relational database);
- **Online Mendelian Inheritance in Man (OMIM)** = a database as resource for the study of human genetics and human molecular medicine;
- **PDB** = Protein Data Bank contains data derived from X-ray crystallography and NMR (nuclear magnetic resonance) studies;

Advance Organizer (4/4)

- **Phylogenetics** = similarities and differences among species can be used to infer evolutionary relationships (=phylogenies); Examples for phylogenetic software: PAUP, PHYLIP;
- **PROSITE** = database containing sequence patterns associated with protein family membership, specific protein functions an post-translational modifications;
- **R** = An open source (free) programming language and software environment for statistical computing and graphics;
- **Relational database** = a database made up of a collection of tables (relations), i.e., data are stored in rows and columns. Relational database management systems (RDBMS) store a type of structured data. SQL is the most widely used language for managing relational databases (see there);
- **Semi-structured data** = Data that do not conform to fixed fields but contain tags and other markers to separate data elements. Examples of semi-structured data include XML or HTML-tagged text. Contrast with structured data and unstructured data.
- **Similarity table** = distance table;
- **SQL** = Originally an acronym for structured query language, SQL is a computer language designed for managing data in relational databases. This technique includes the ability to insert, query, update, and delete data, as well as manage data schema (database structures) and control access to data in the database;
- **SRS** = Sequence Retrieval System, a data retrieval tool based on open source software
- **SWISS-PROT** = is a databank containing a collection of confirmed protein sequences with annotations relating to structure, function and protein family assignment;
- **UniGene** = experimental facility for the clustering of GenBank sequences and is related to EST (expressed sequence tag) data;

Glossary

- ACeDB = A *C* elegans Data Base
- ADE = adverse drug events
- CDSS = clinical decision support system
- CPOE = computerized physician order entry
- DBMS = Data Base Management System
- EMAC = electronic medication administration chart
- EO = electronic order
- ERT = error registration table
- GFR = glomerular filtration rate
- HIS = Hospital Information System (DE: KIS = Krankenhaus Informations System)
- HWO = handwritten order
- NICU = neonatal intensive care unit
- NOE = nurse order entry (followed by physician's verification and countersignature)
- PBMAC = paper-based medication administration chart
- POE = physician order entry
- RR = rate ratio
- UniProt = Universal Protein Ressource

- … have an overview about the general **architecture of an Hospital Information System**
  - *(Details in lecture 10: Medical Information Systems and Biomedical Knowledge Management!);*
- … know some principles of **hospital databases**;
- … have an overview on some important **biomedical databases**;
- … are familiar with some basic methods of **information retrieval**;

- Increasingly large and *complex data* sets due to **data intensive biomedicine** [1]
- Increasing amounts of **non-standardized** and **un-structured information** (e.g. “free text”)
- Data **quality,** data **integration**, universal **access**
- **Privacy,** security, safety, data protection, data ownership, fair use of data (see →Lecture 11) [2]
- **Time** aspects in databases [3]

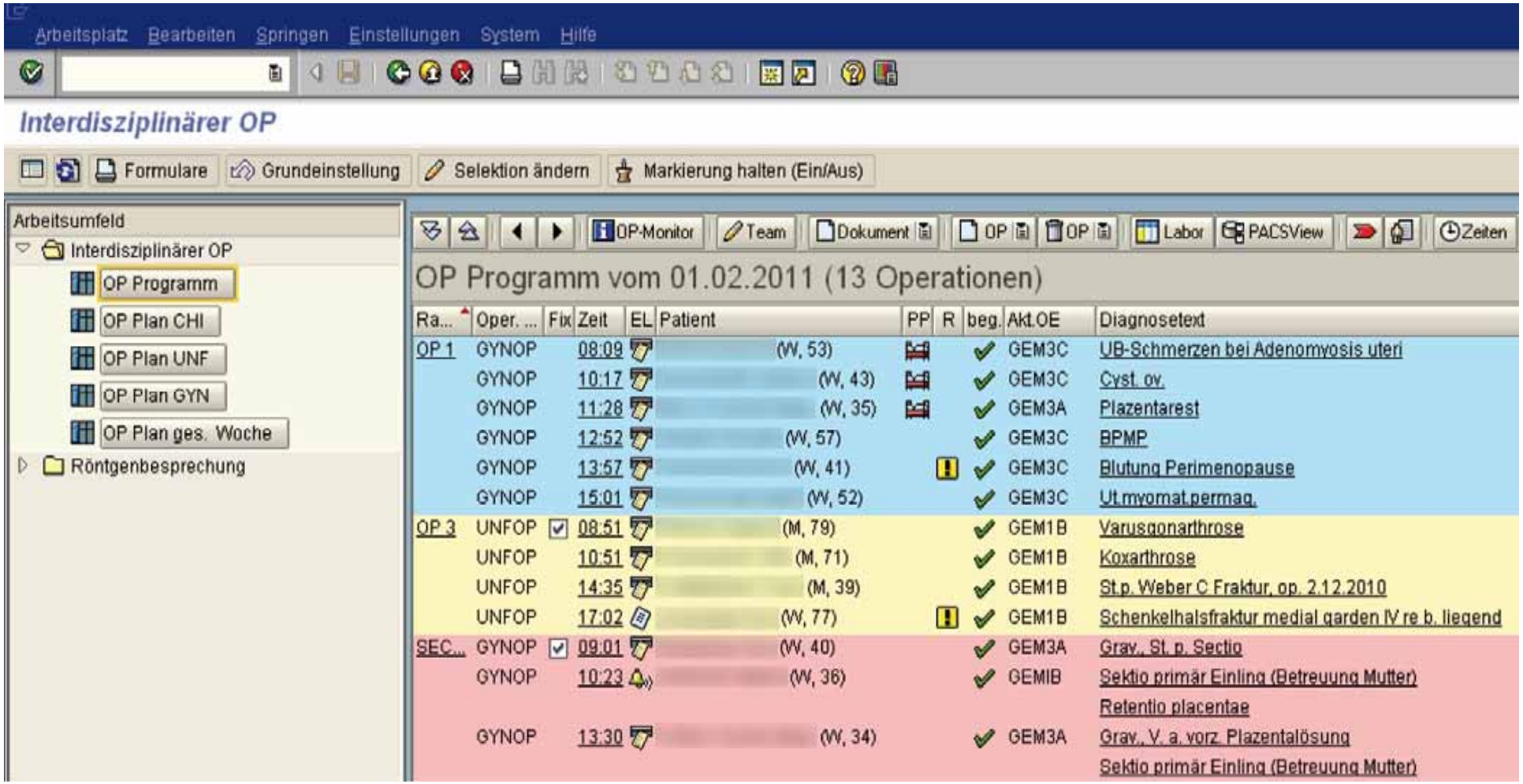
[1] Shah, N. H. & Tenenbaum, J. D. 2012. The coming age of data-driven medicine: translational bioinformatics' next frontier. Journal of the American Medical Informatics Association, 19, (E1), E2-E4.
[2] Kieseberg, P., Hobel, H., Schrittwieser, S., Weippl, E. & Holzinger, A. 2014. Protecting Anonymity in Data-Driven Biomedical Science. In: LNCS 8401. Berlin Heidelberg: Springer pp. 301-316..
[3] Gschwandtner, T., Gärtner, J., Aigner, W. & Miksch, S. 2012. A taxonomy of dirty time-oriented data. In: LNCS 7465. Heidelberg, Berlin: Springer, pp. 58-72.

Let us start with a look into the Hospital …

Slide 4-1 Hospital Information System: Typical Scenario

G'sund Net, Ausgabe 45, März 2005

Slide 4-2 HIS: Typical View on the Clinical Workplace

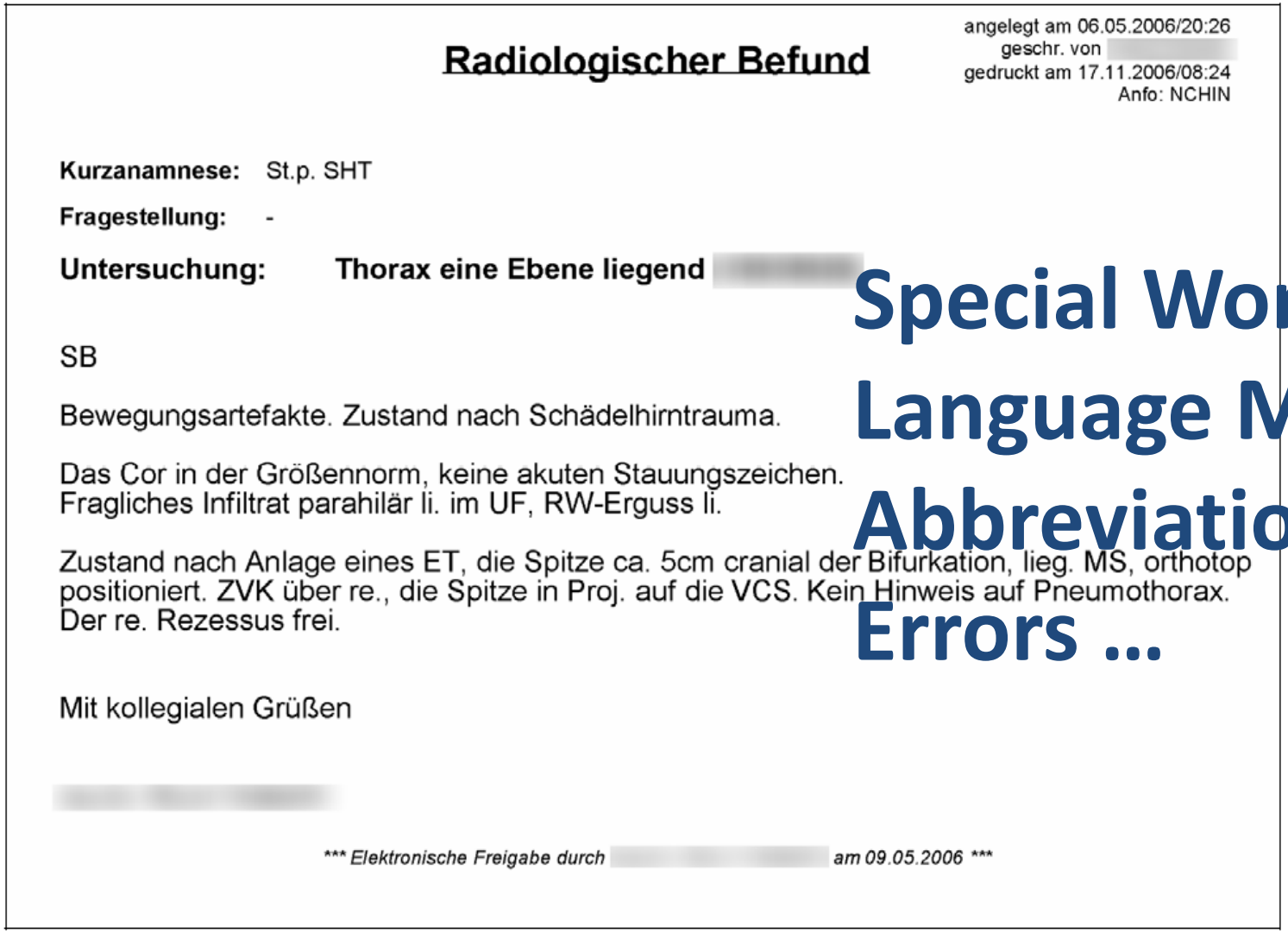


G'sund Net, Ausgabe 70, Juni 2011

- *… and requires a lot of communication and information exchange …*

Holzinger, A., Geierhofer, R., Ackerl, S. & Searle, G. (2005). *CARDIAC@VIEW: The User Centered Development of a new Medical Image Viewer. Central European Multimedia and Virtual Reality Conference, Prague, Czech Technical University (CTU), 63-68.*

Slide 4-4: The medical report is the most important output



Radiologischer Befund

angelegt am 06.05.2006/20:26
geschr. von
gedruckt am 17.11.2006/08:24
Anfo: NCHIN

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

Das Cor in der Größennorm, keine akuten Stauungszeichen.
Fragliches Infiltrat parahilär li. im UF, RW-Erguss li.

Zustand nach Anlage eines ET, die Spitze ca. 5cm cranial der Bifurkation, lieg. MS, orthotop positioniert. ZVK über re., die Spitze in Proj. auf die VCS. Kein Hinweis auf Pneumothorax.
Der re. Rezessus frei.

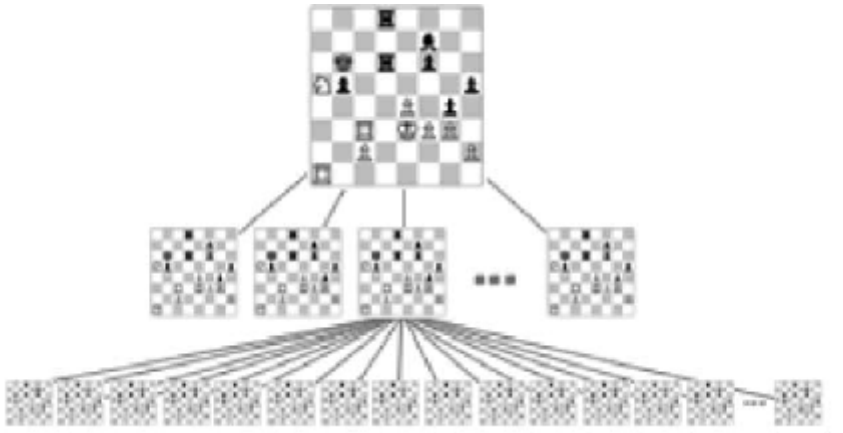
Mit kollegialen Grüßen

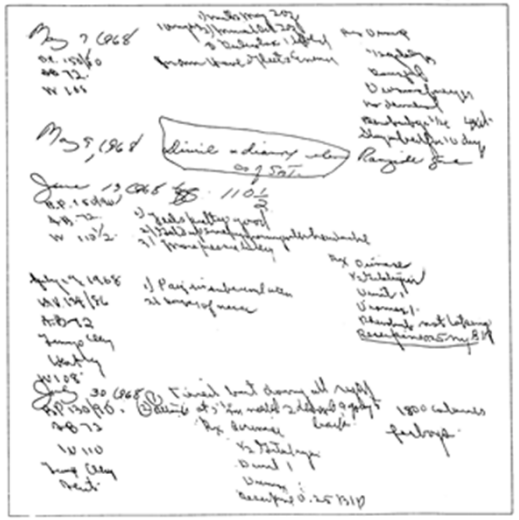
**** Elektronische Freigabe durch am 09.05.2006 ****

Special Words
Language Mix
Abbreviations
Errors …

Holzinger, A., Geierhofer, R. & Errath, M. 2007. Semantische Informationsextraktion in medizinischen Informationssystemen. *Informatik Spektrum, 30, (2), 69-78.*

Slide 4-5 Excursus: Chess Game versus Natural Language





http://stanford.edu/~cpiech/cs221/apps/deepBlue.html

German Example: Synonymity and Ambiguity

„die Antrumschleimhaut ist durch Lymphozyten infiltriert“

„lymphozytäre Infiltration der Antrummukosa“

„Lymphoyteninfiltration der Magenschleimhaut im Antrumbereich“

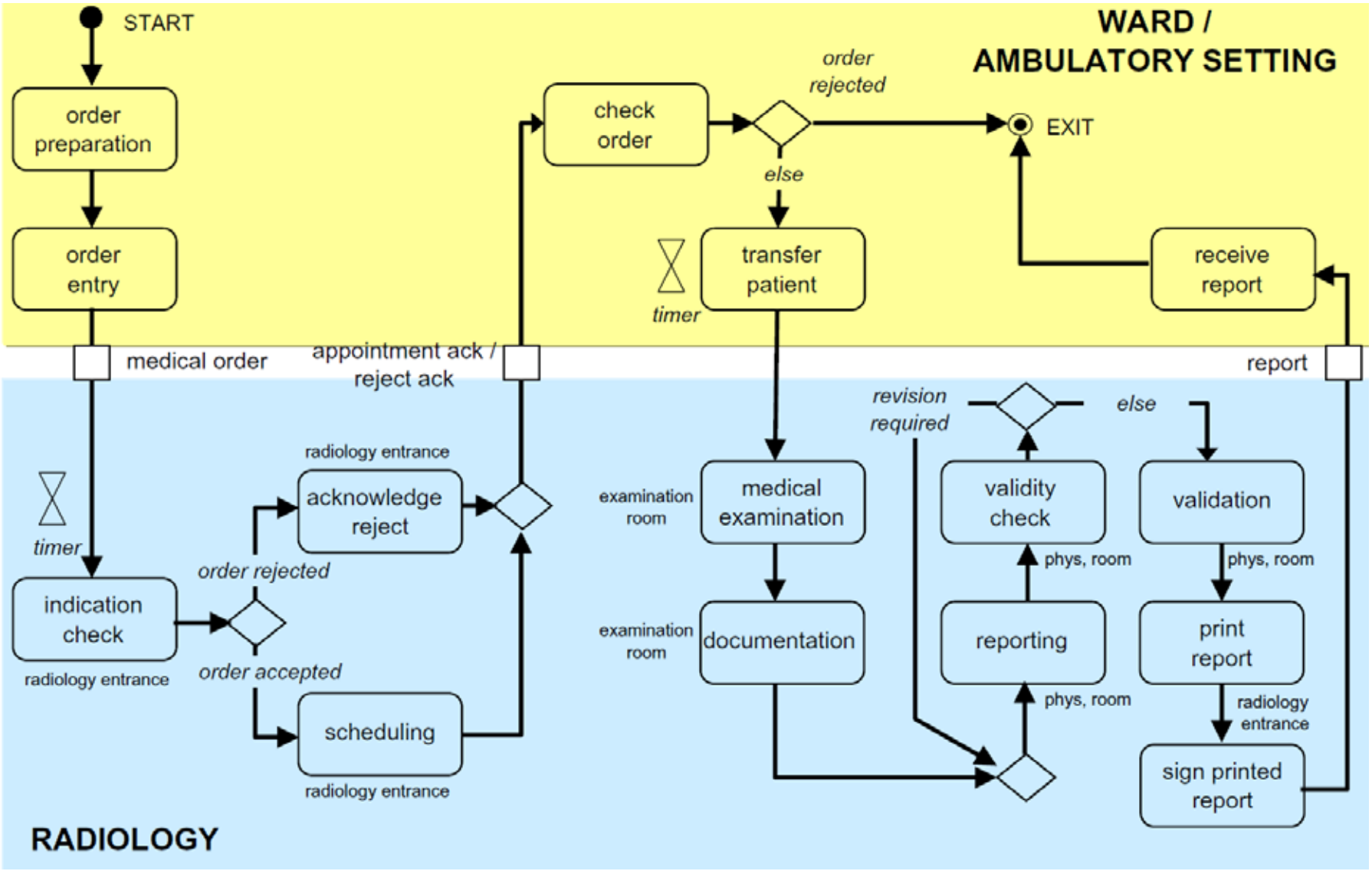
HWI = Harnwegsinfekt, Hinterwandinfarkt, Hakenwurminfektion, Halswirbelimmobilisation, Hinterwandischämie, Hip Waist Index, Height-Width Index, Häufig wechselnder Intimpartner, Hepatitic weight index …

Leitung = Nervenleitung, Abteilungsleitung, Stromleitung, Wasserleitung, Harnleitung, …

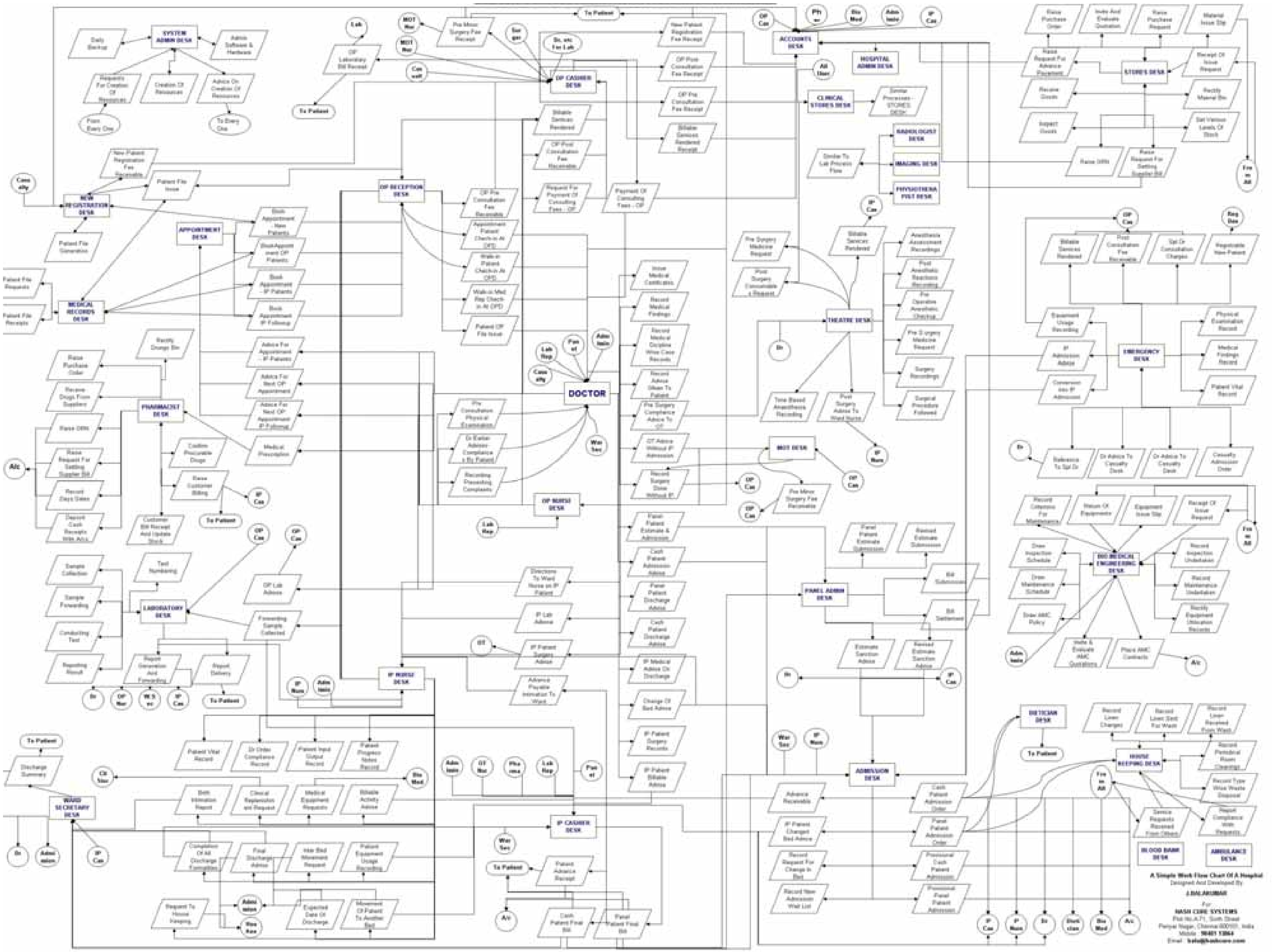
http://www.medizinische-abkuerzungen.de/suche.html

Hospital workflows are also complex …

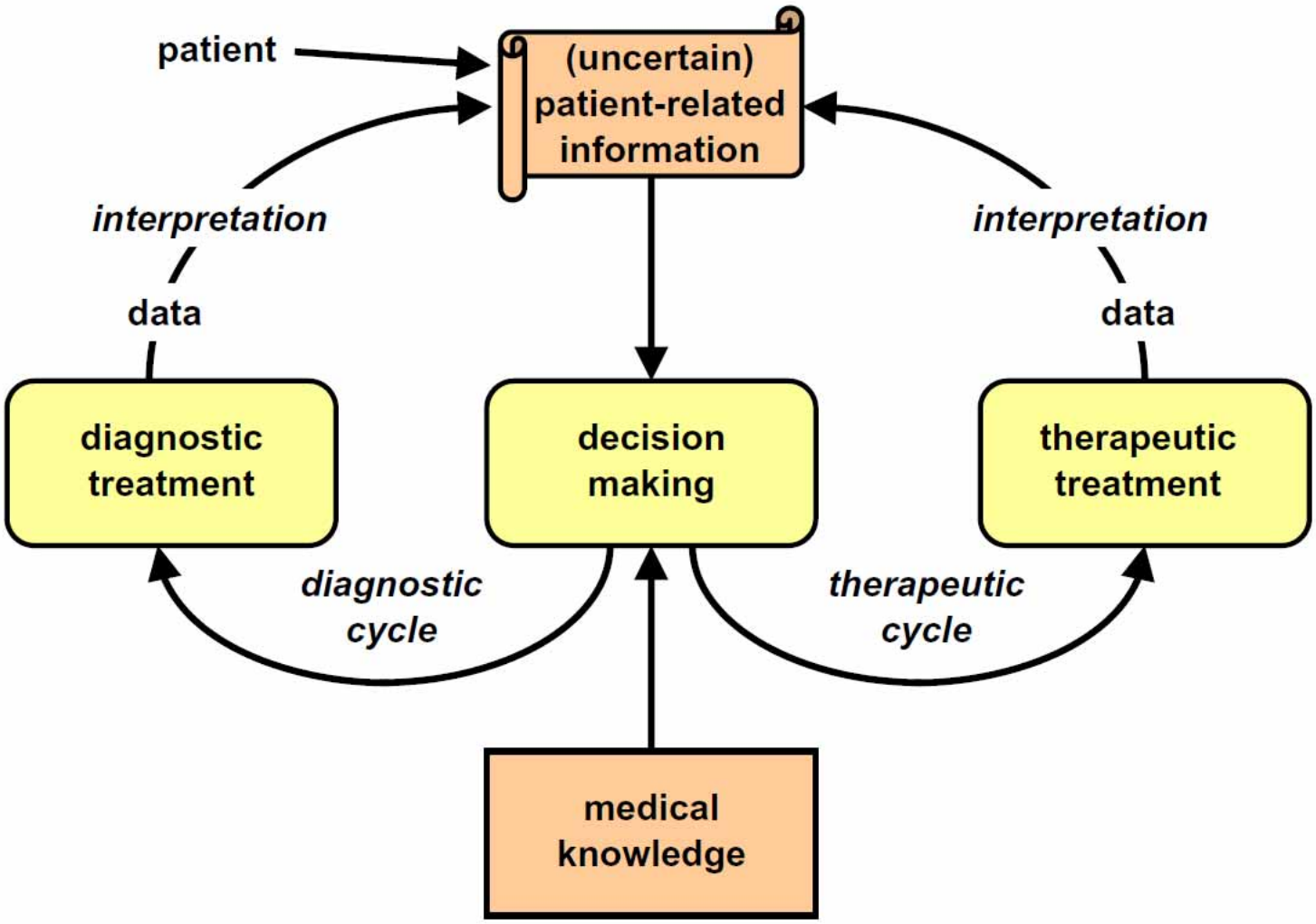
Slide 4-6 Typ. Workflow: order entry and result reporting



Lenz, R. & Reichert, M. (2007) IT support for healthcare processes-premises, challenges, perspectives. *Data & Knowledge Engineering, 61, 1, 39-58.*


SYSTEM ADMIN DESK
NEW REGISTRATION DESK
APPOINTMENT DESK
MEDICAL RECORDS DESK
OP RECEPTION DESK
OP CASHIER DESK
ACCOUNTS DESK
HOSPITAL ADMIN DESK
CLINICAL STORES DESK
STORES DESK
RADIOLOGIST DESK
IMAGING DESK
PHYSIOTHERAPIST DESK
THEATRE DESK
EMERGENCY DESK
PHARMACIST DESK
DOCTOR
OP NURSE DESK
MOT DESK
LABORATORY DESK
PANEL ADMIN DESK
BIO MEDICAL ENGINEERING DESK
IP NURSE DESK
DIETICIAN DESK
ADMISSION DESK
HOUSE KEEPING DESK
WARD SECRETARY DESK
IP CASHIER DESK
BLOOD BANK DESK
AMBULANCE DESK
To Patient
A Simple Work Flow Chart Of A Hospital
Designed And Developed By
J.BALAKUMAR
For
HASH CORE SYSTEMS

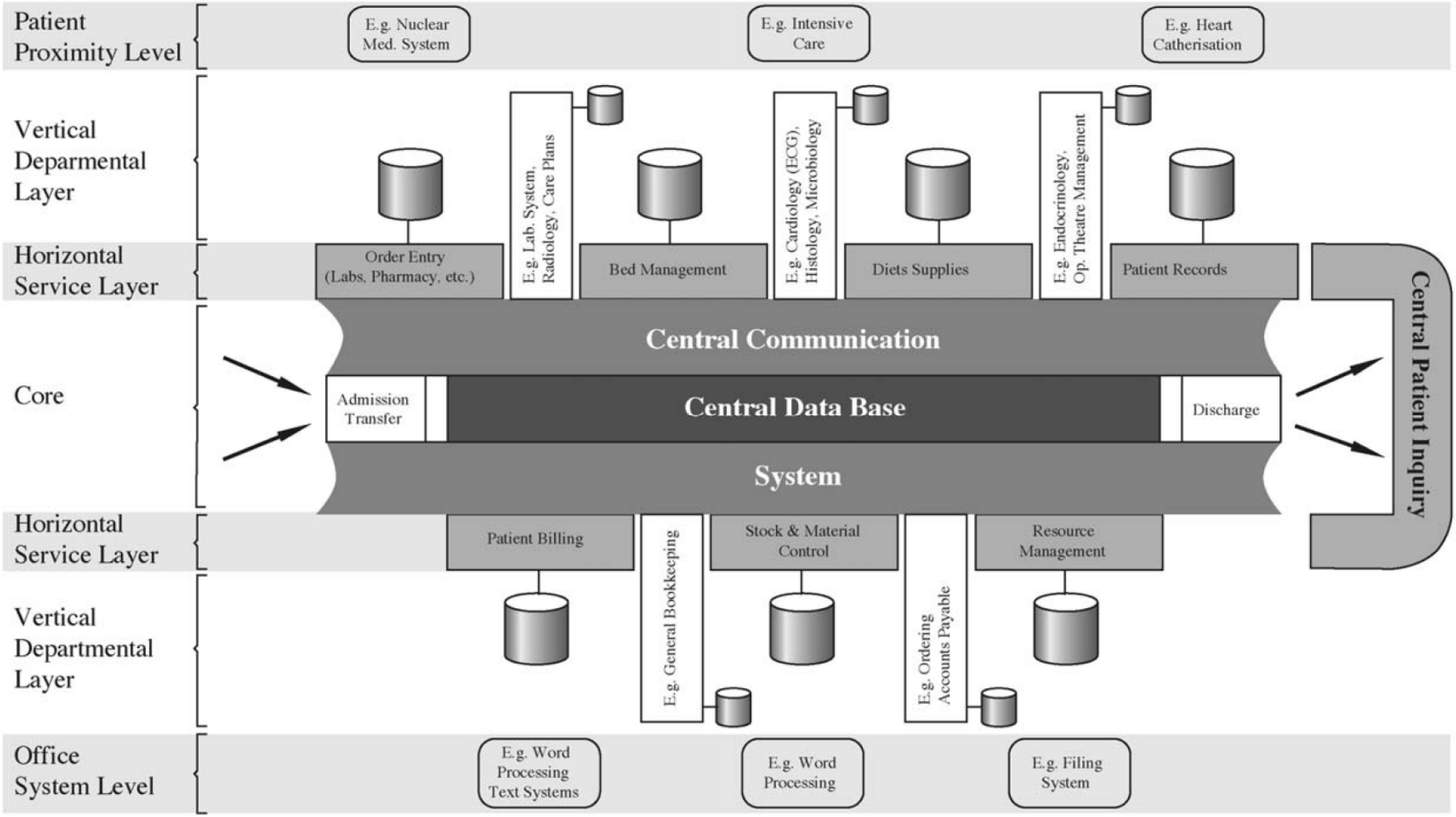
Slide 4-7: Remember the diagnostic-therapeutic cycle



Lenz, R. & Reichert, M. 2007. IT support for healthcare processes-premises, challenges, perspectives. Data & Knowledge Engineering, 61, (1), 39-58.

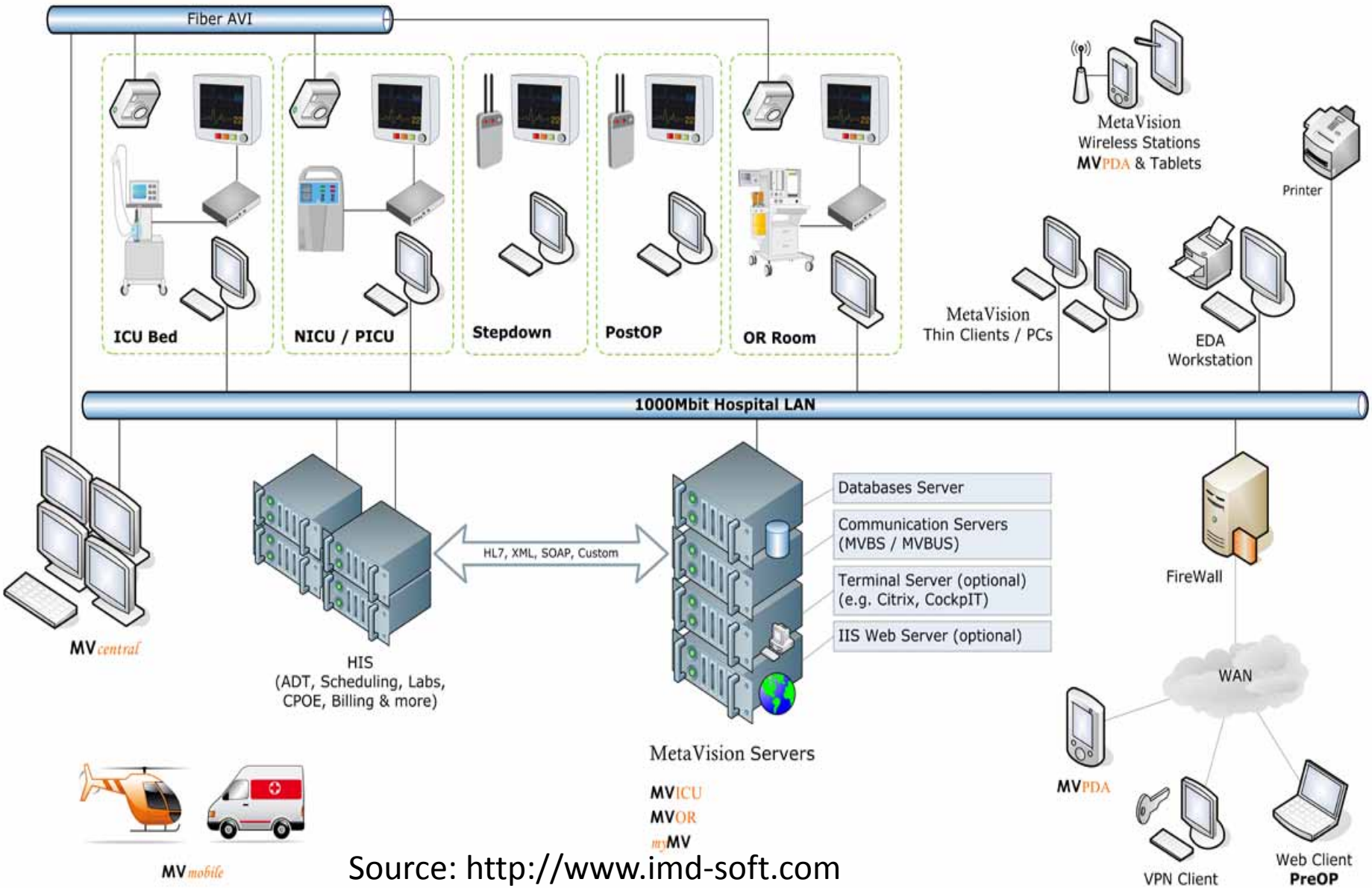
What is the architecture of an hospital information system?

Slide 4-8 HIS: Classic Conceptual Model (in use since 1984)



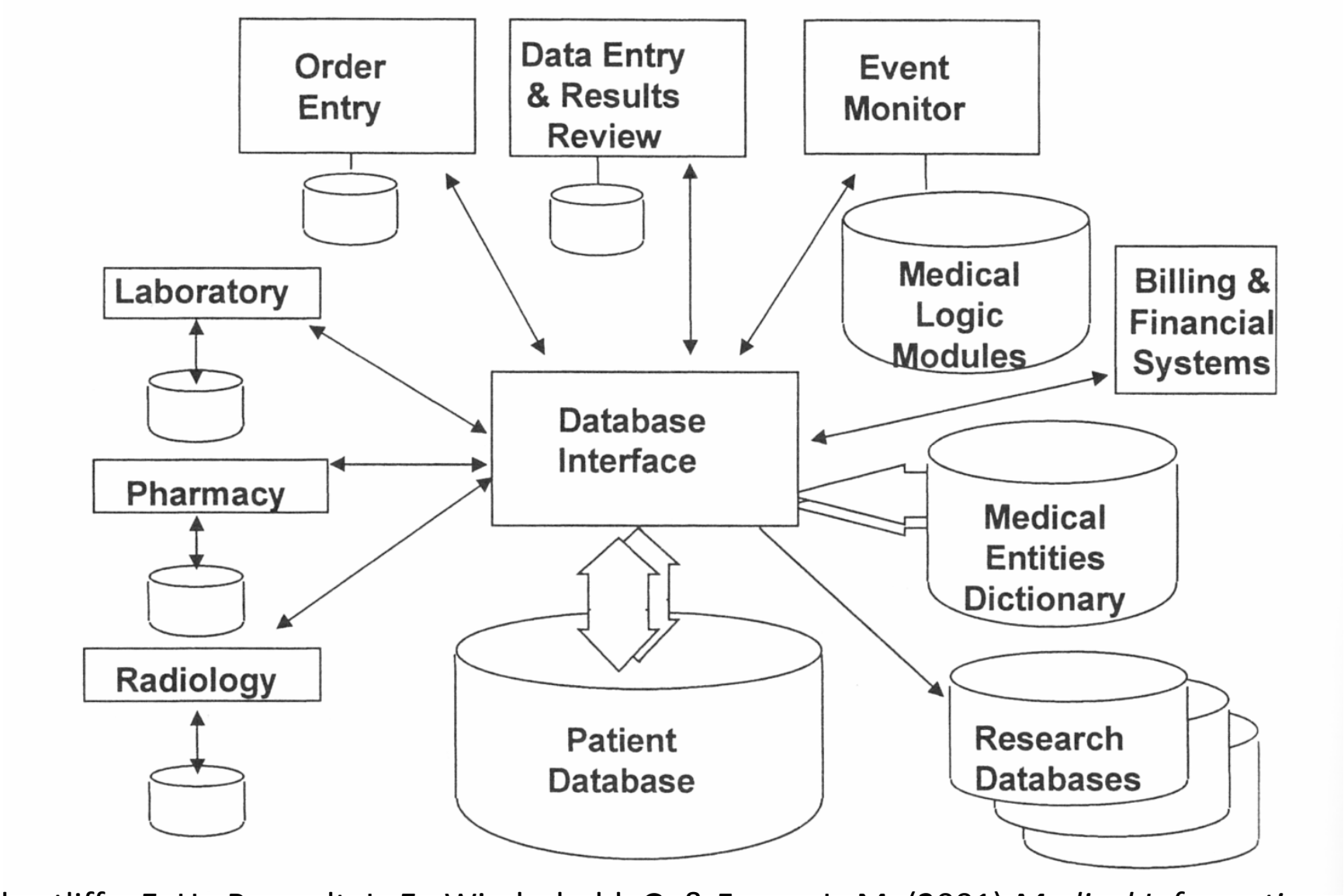
Reichertz, P. L. (2006) Hospital information systems - Past, present, future. *International Journal of Medical Informatics, 75, 3-4, 282-299.*

Slide 4-9: Modern Enterprise HIS: Sample Architecture



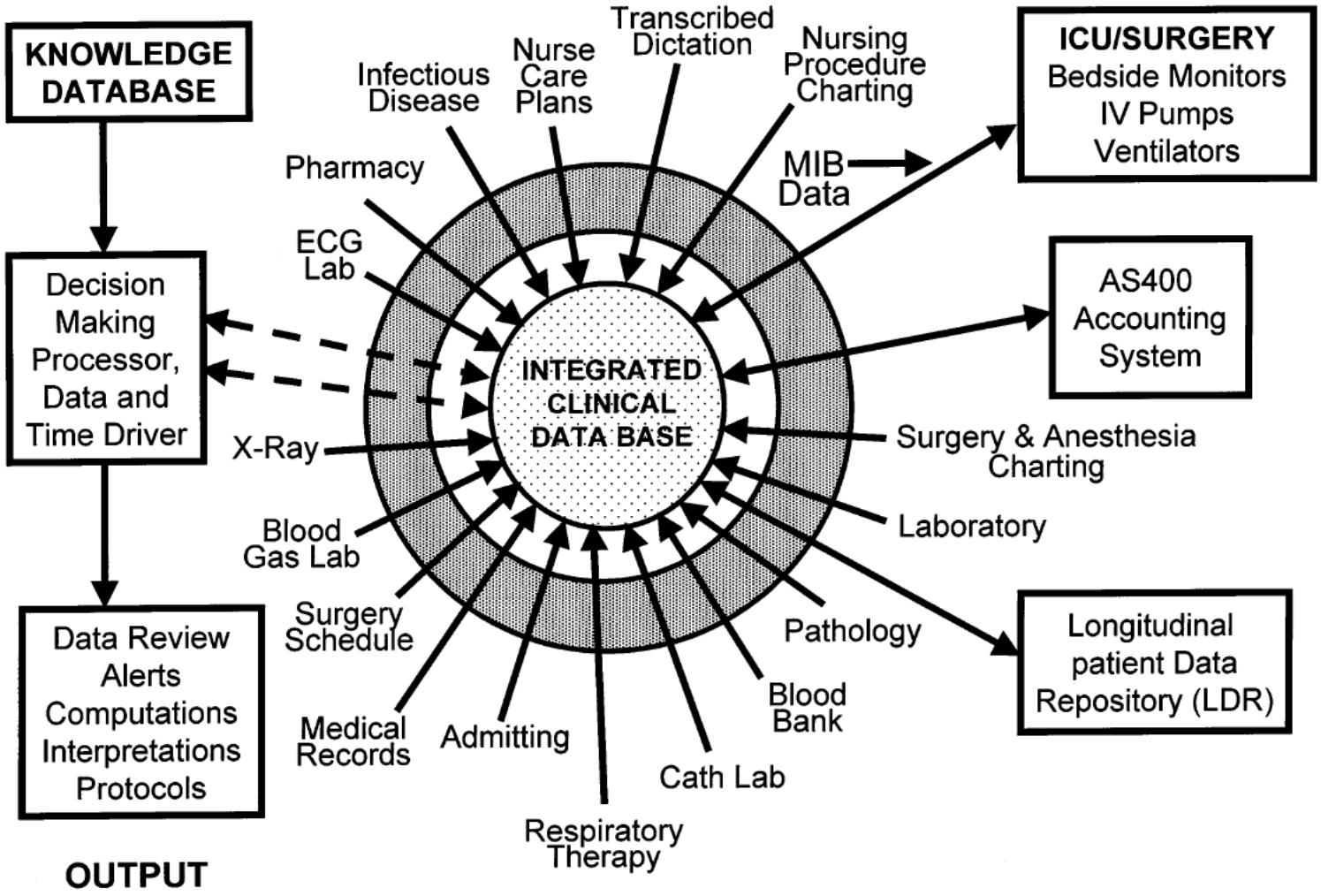
Source: http://www.imd-soft.com

Slide 4-10: HIS Central components: Databases



Shortliffe, E. H., Perrault, L. E., Wiederhold, G. & Fagan, L. M. (2001) *Medical Informatics: Computer Applications in Health Care and Biomedicine. Second Edition. New York, Springer.*

Slide 4-11 Historical Example: The HELP System (1967)



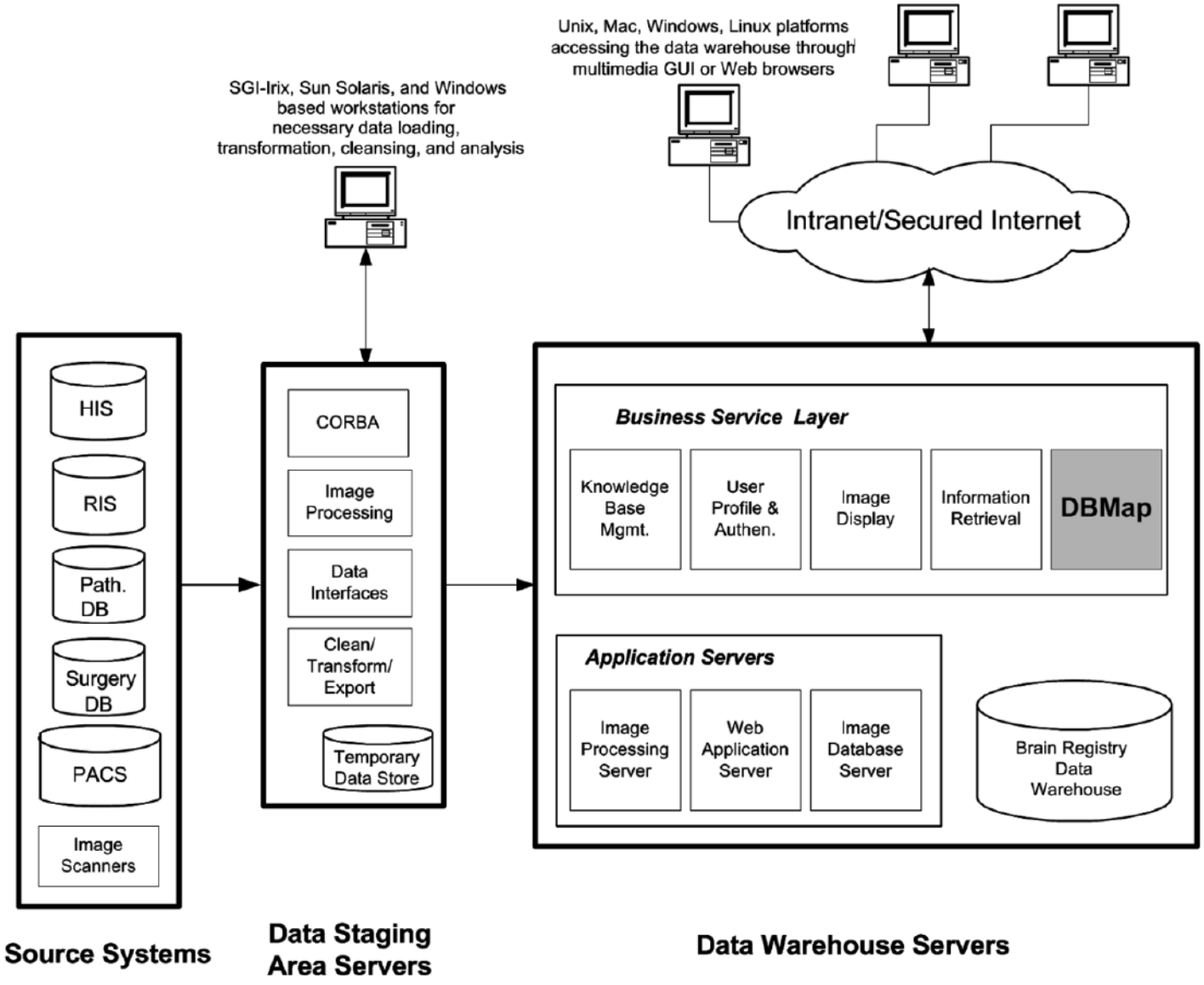
Gardner, R. M., Pryor, T. A. & Warner, H. R. (1999) The HELP hospital information system: update 1998. *International Journal of Medical Informatics, 54, 3, 169-182.*

Data Integration
Data Fusion
Data Curation

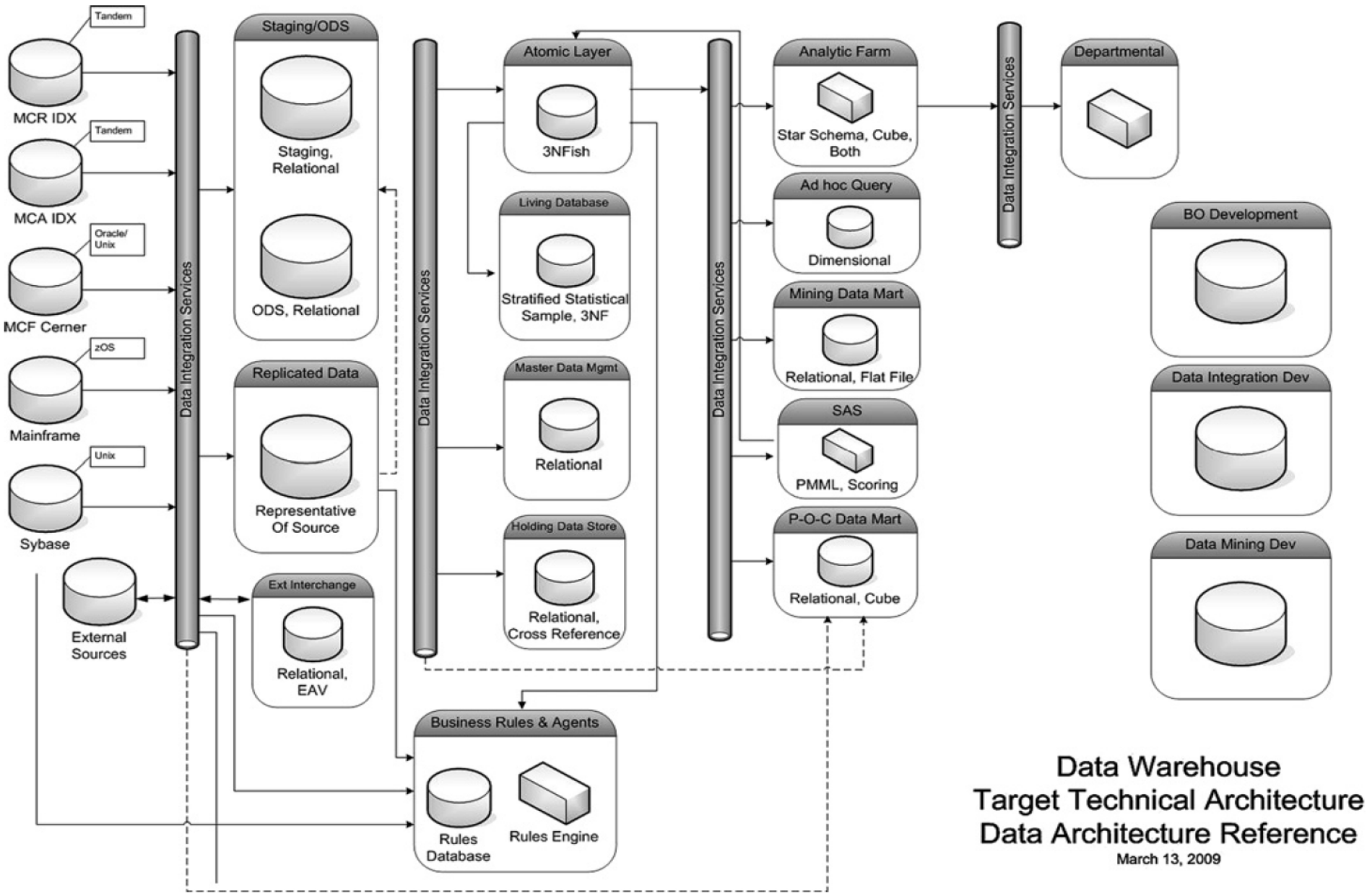
- **Database** (DB) is the organized collection of data through a certain data structure (e.g. hash-table, adjacency matrix, graph structure, etc.).
- **Database management system** (DBMS) is software which operates the DB. Well known DBMSs include: Oracle, IBM DB2, Microsoft SQL Server, Microsoft Access, MySQL, SQLite. Examples for Graph Databases include InfoGrid, Neo4j, or BrightstarDB.
- The used DB is not generally portable, but different DBMSs can inter-operate by using standards such as SQL and ODBC.
- **Database system** (DBS) = DB + DBMS. The term database system emphasizes that data is managed in terms of accuracy, availability, resilience, and usability.
- **Data warehouse** (DWH) is an integrated repository used for reporting and long term storage of analysis data.
- **Data Marts** (DM) are access layers of a DWH and are used as temporary repositories for data analysis.

Slide 4-13: Example Hospital data warehouse

Zhang, M., Zhang, H., Tjandra, D. & Wong, S. T. C. (2004) DBMap: a space-conscious data visualization and knowledge discovery framework for biomedical data warehouse. *Information Technology in Biomedicine, IEEE Transactions on, 8, 3, 343-353.*

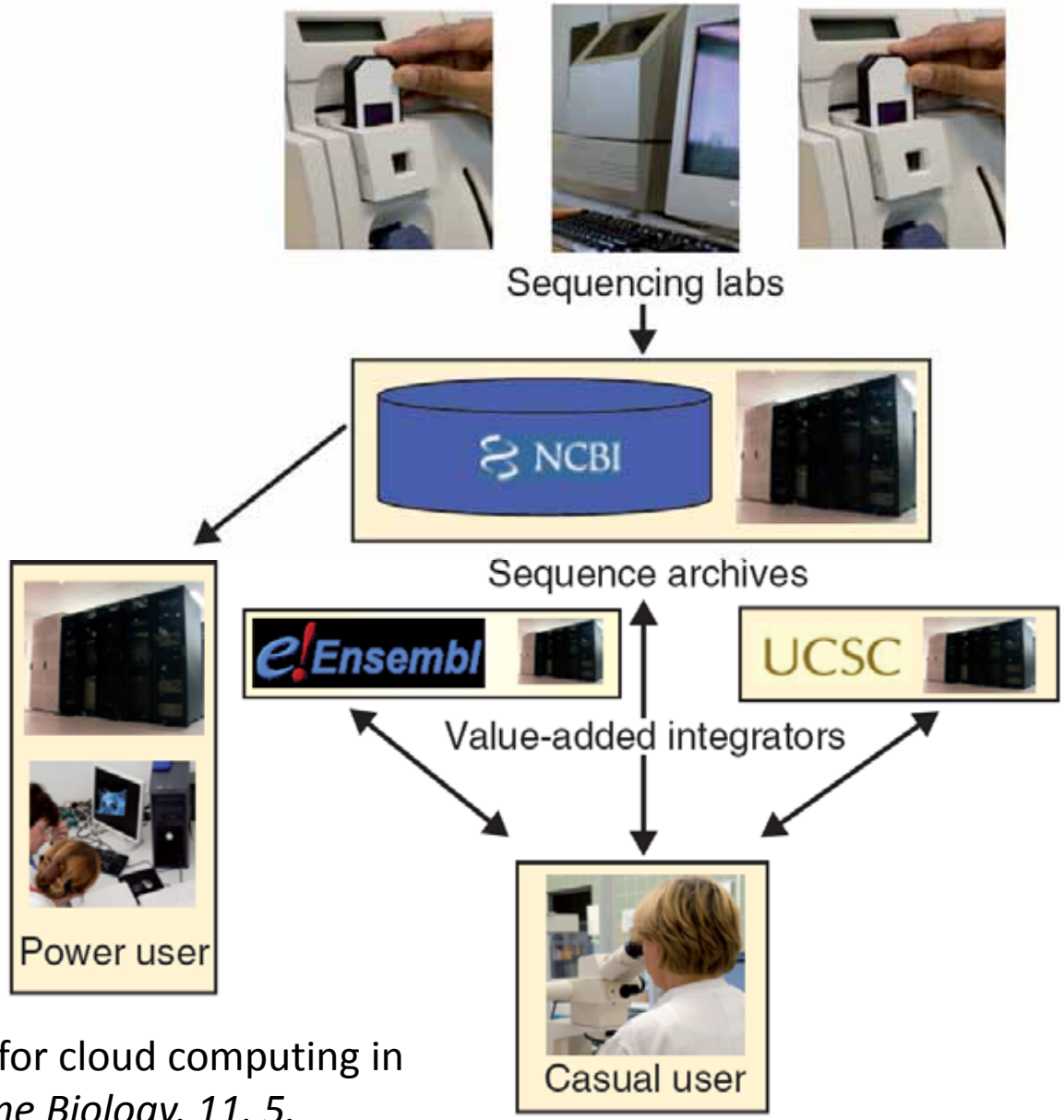


Slide 4-14 Example: Mayo Clinics Data Warehouse



What about cloud-based Information Systems?

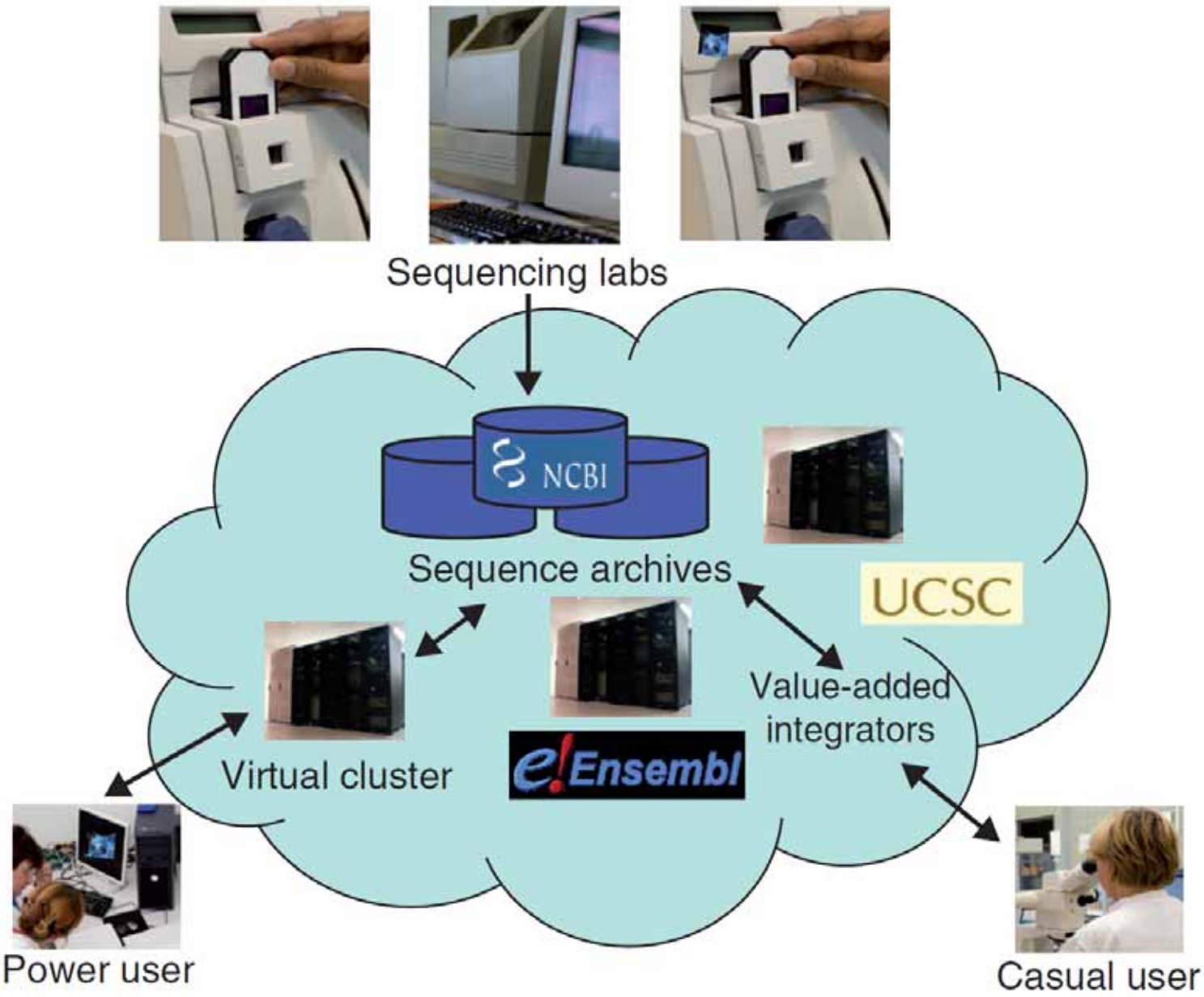
Slide 4-15: Traditional Genome Information System



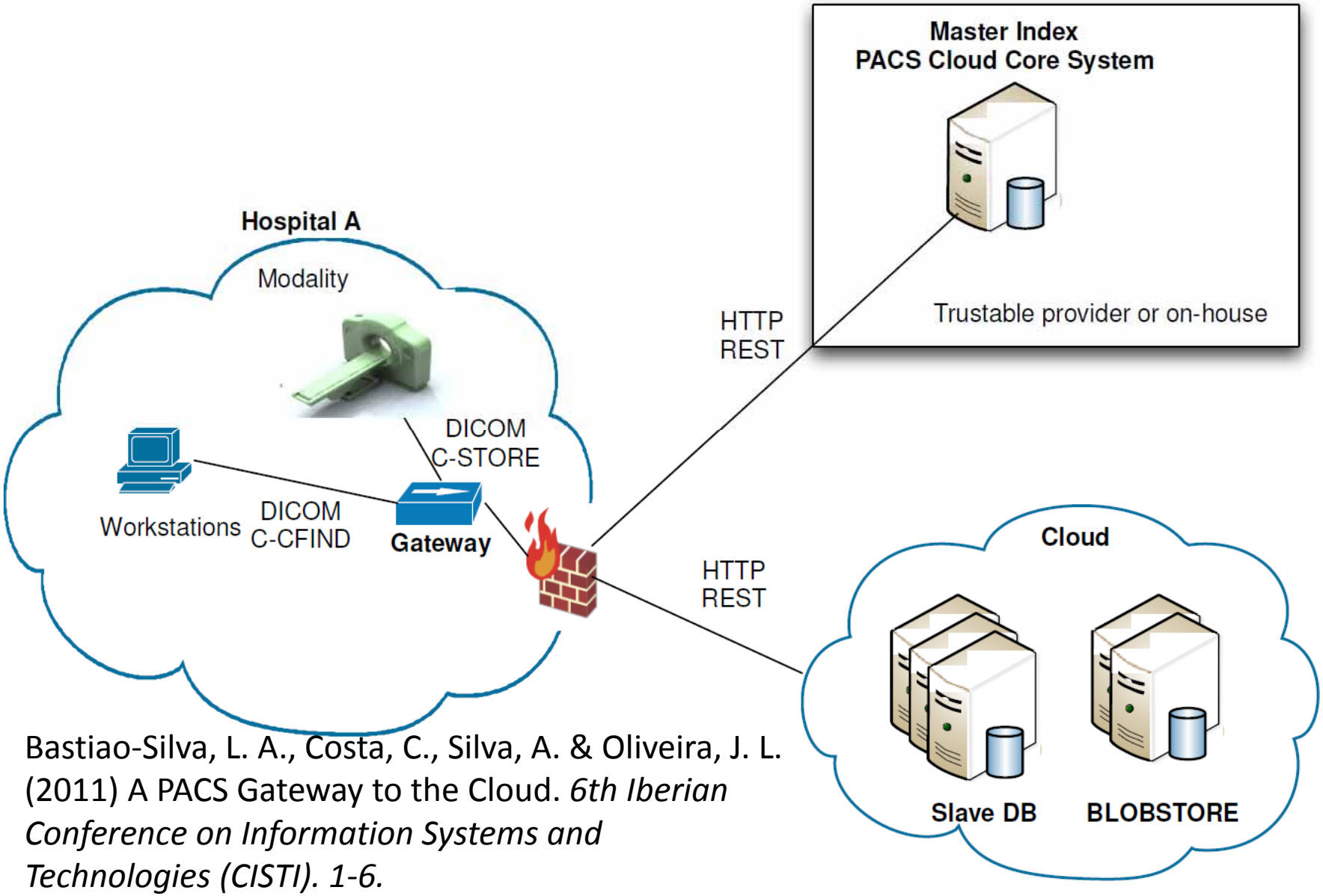
Stein, L. D. (2010) The case for cloud computing in genome informatics. *Genome Biology, 11, 5.*

Slide 4-16: Genome Info Ecosystem cloud computing

Stein, L. D. (2010) The case for cloud computing in genome informatics. *Genome Biology, 11, 5.*

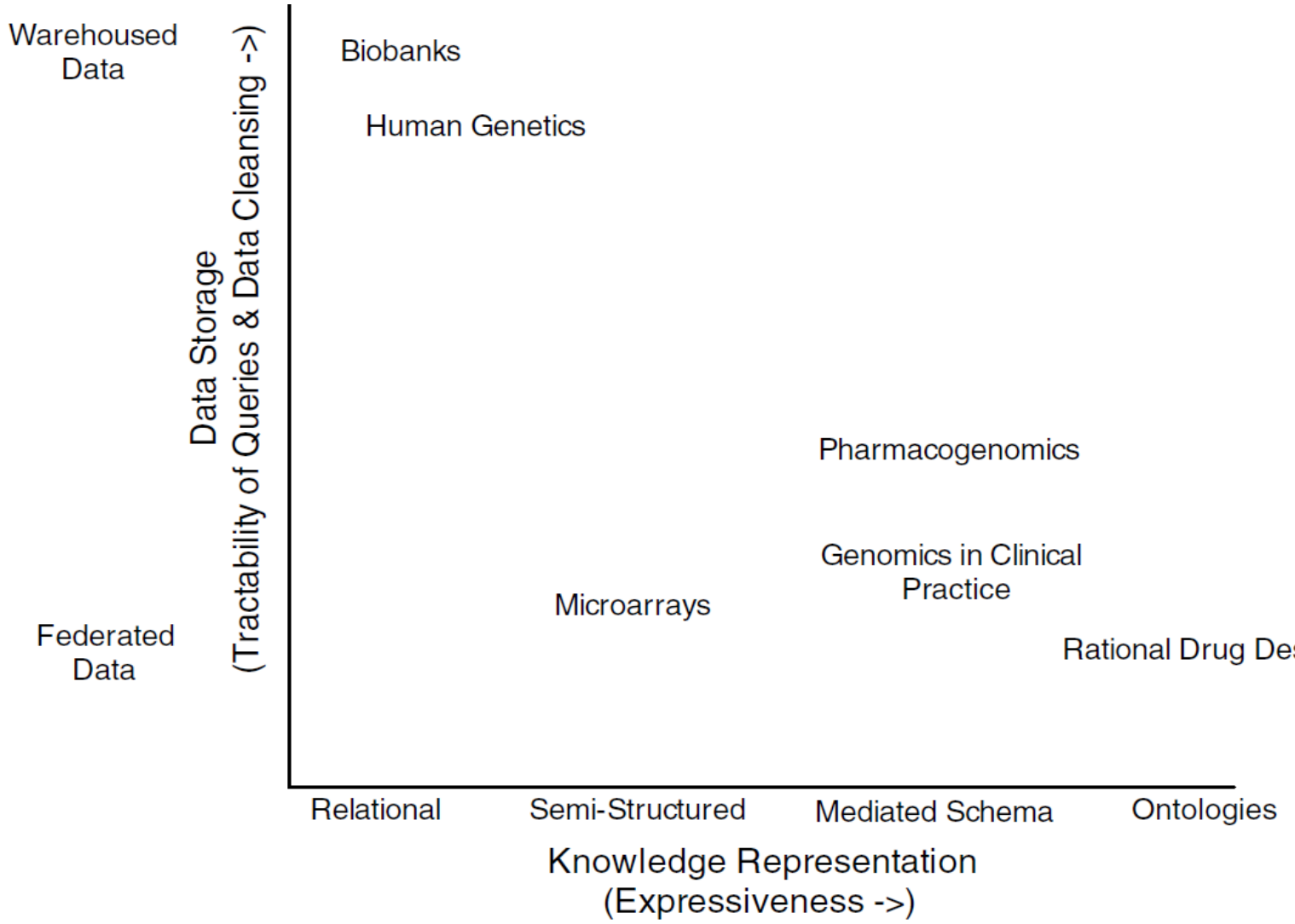


Slide 4-17: Example Clinical Cloud Computing: PACS Cloud



Bastiao-Silva, L. A., Costa, C., Silva, A. & Oliveira, J. L. (2011) A PACS Gateway to the Cloud. *6th Iberian Conference on Information Systems and Technologies (CISTI). 1-6.*

Slide 4-18: Federated Data vs. Warehoused Data

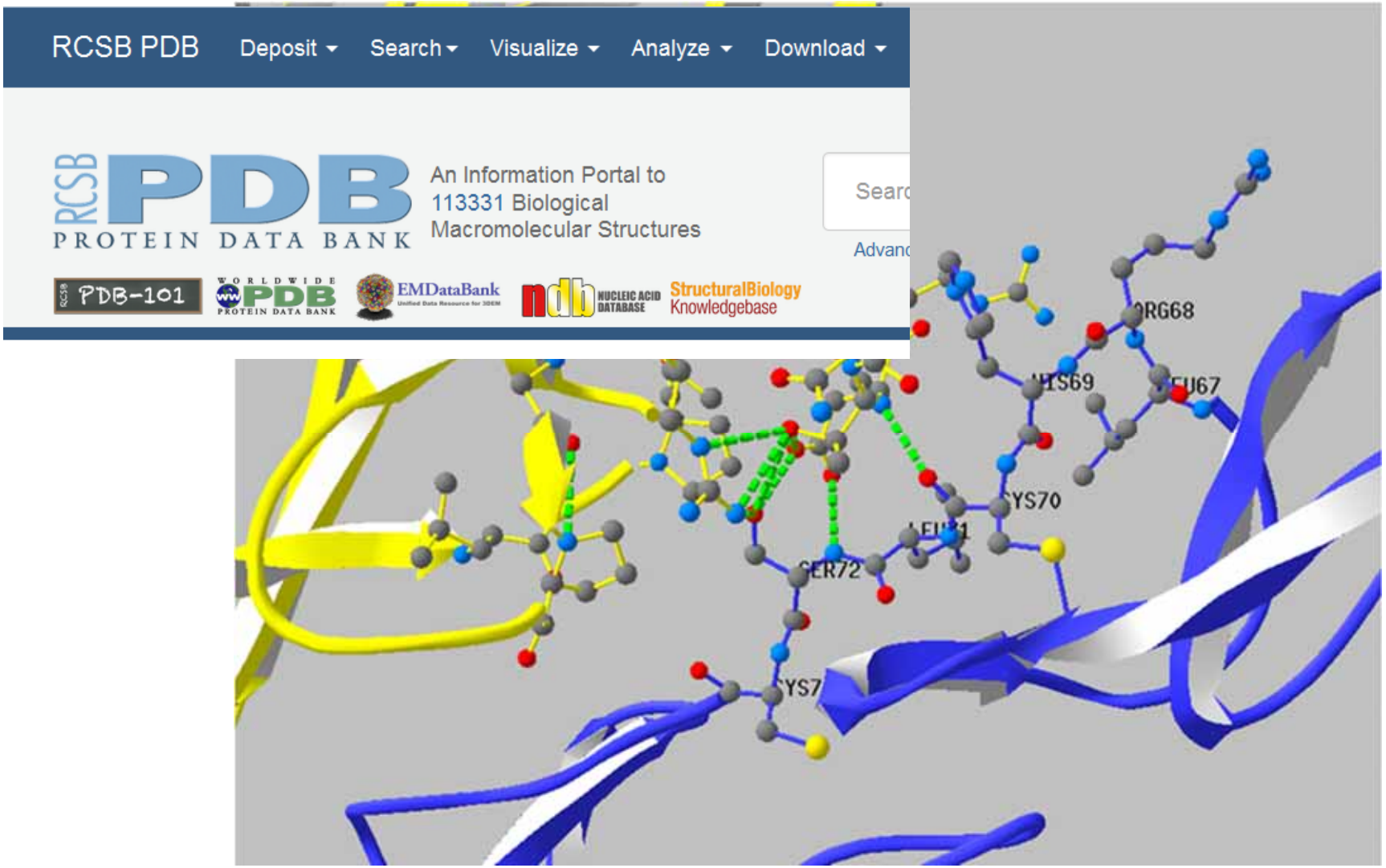


Louie, B., Mork, P., Martin-Sanchez, F., Halevy, A. & Tarczy-Hornoch, P. (2007) Data integration and genomic medicine. *Journal of Biomedical Informatics, 40, 1, 5-16.*

What is the difference between hospital databases and Biomedical databases?

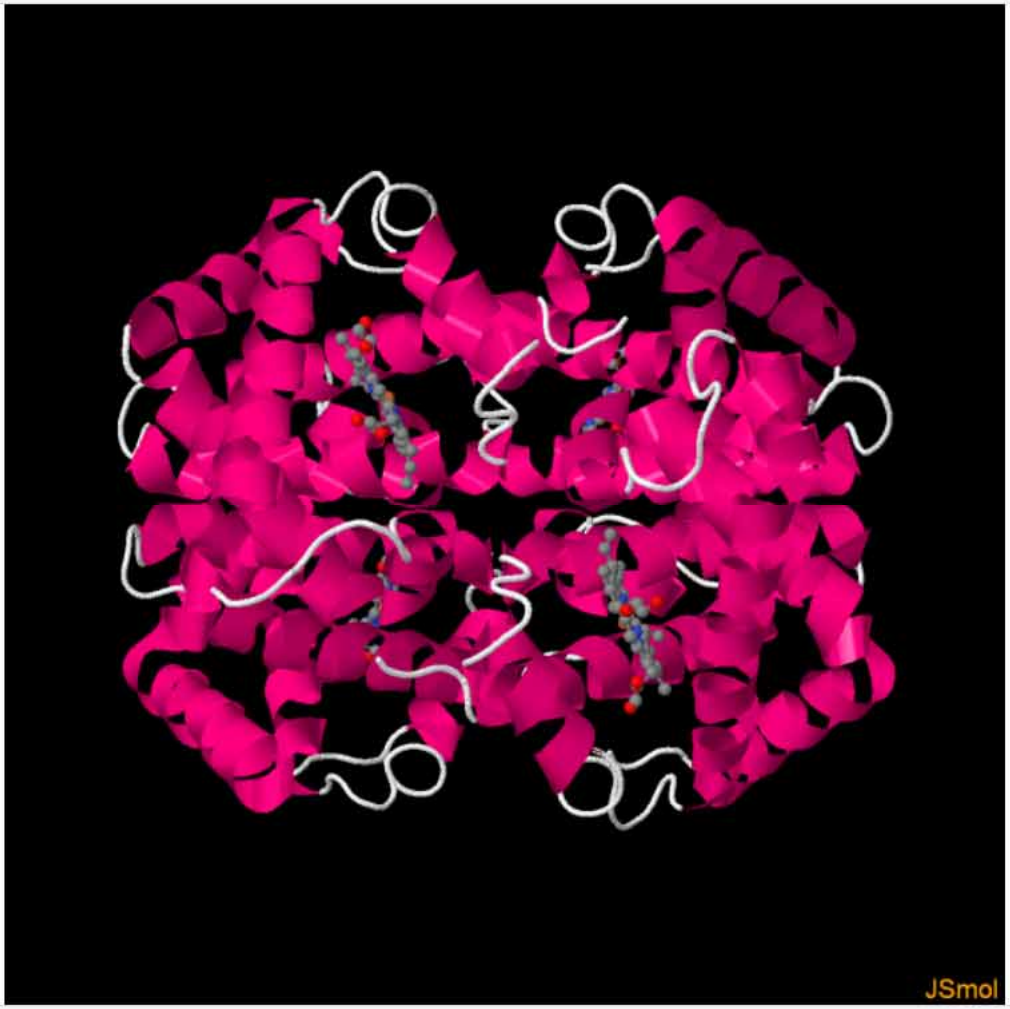
- … are libraries of life science data, collected from scientific experiments and computational analyses.
- … contain (clinical, biological, …) data from clinical work, genomics, proteomics, metabolomics, microarray gene expression, phylogenetics, etc.
- Examples:
  - Text: e.g. PubMed, OMIM (Online Mendelian Inheritance in Man);
  - Sequence data: e.g. Entrez, GenBank (DNA), UniProt (protein).
  - Protein structures: e.g. PDB, Structural Classification of Proteins (SCOP), CATH (Protein Structure Classification);

Slide 4-20 Example Database: PDB



Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

NOTE: Use your mouse to drag, rotate, and zoom in and out of the structure. Help



Biological assembly 1 assigned by authors and generated by PISA

Select a Viewer

JSmol (JavaScript)

Structure Details

Structure	Biological Assembly 1
Symmetry Type	Global Symmetry
Symmetry	C2
Stoichiometry	A2B2

Select Orientation

Front C2 axis

Select Display Mode

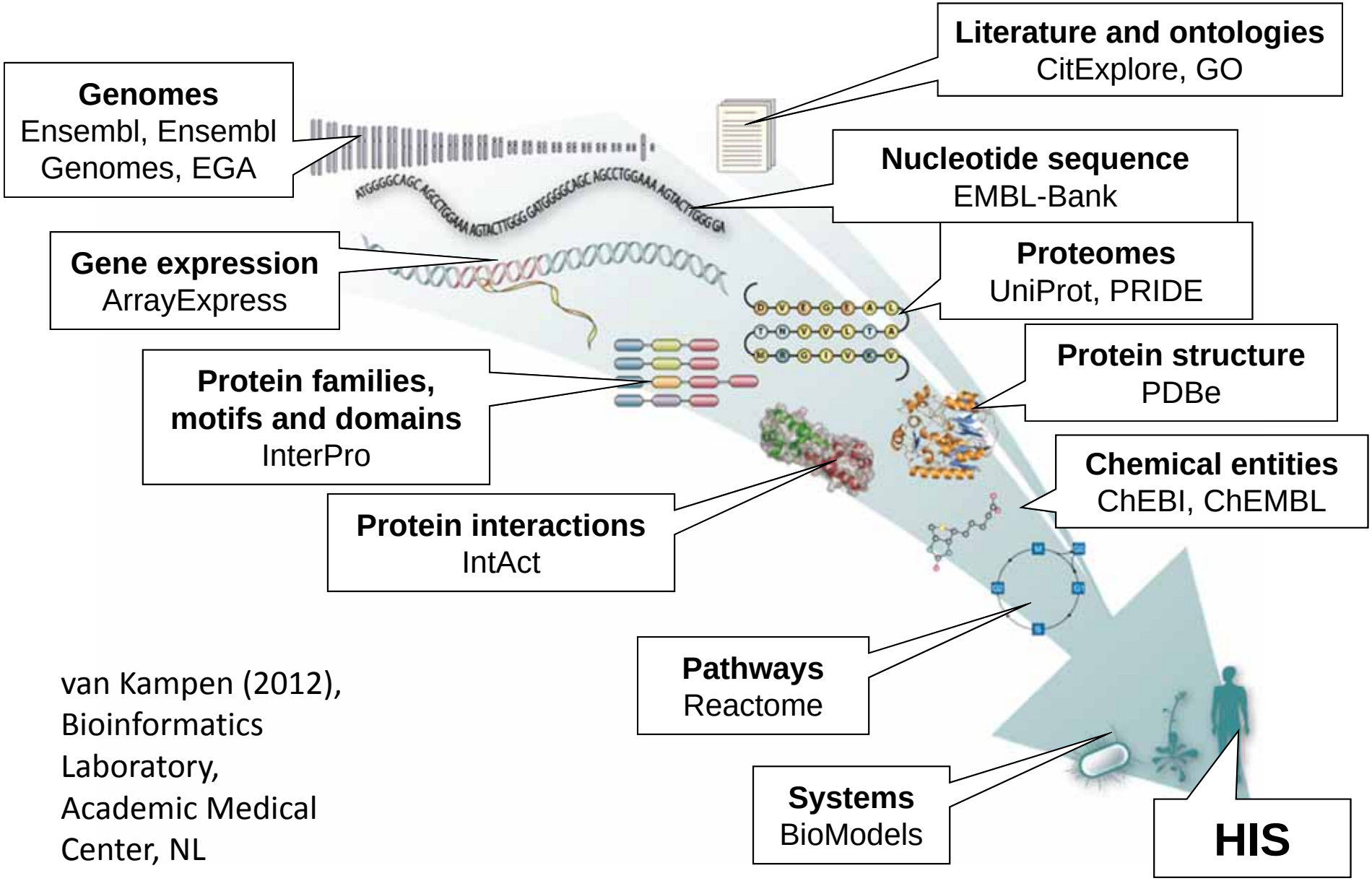
Secondary Structure | Subunit | Symmetry

Display Options

Style	Cartoon
Color	Secondary Structure
Surface	None

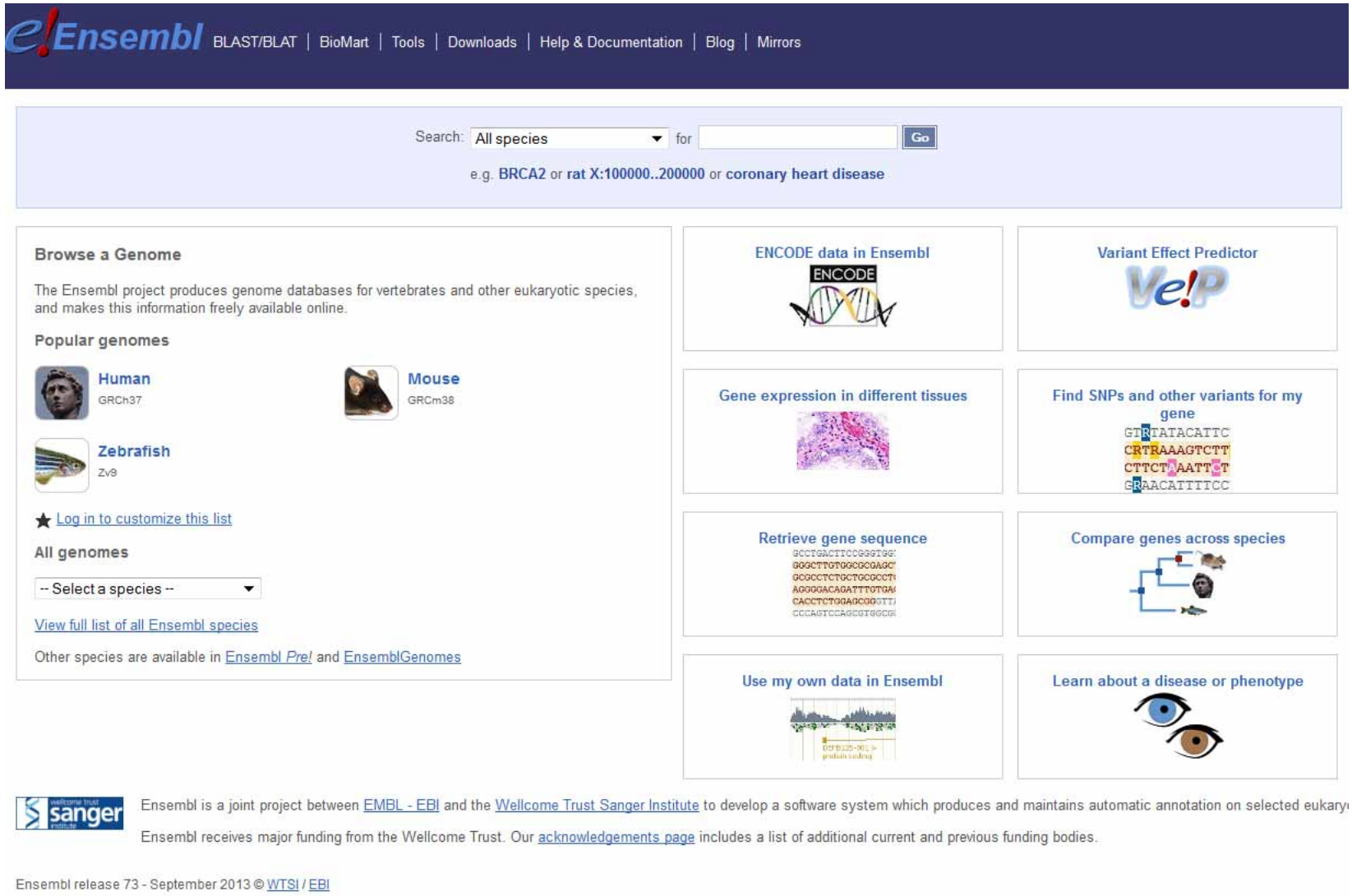
- [ ] H-Bonds
- [ ] Rotation
- [ ] Polyhedron
- [ ] SS Bonds
- [x] Black Background
- [ ] Axes

Slide 4-21 Databases: From Molecules to Systems



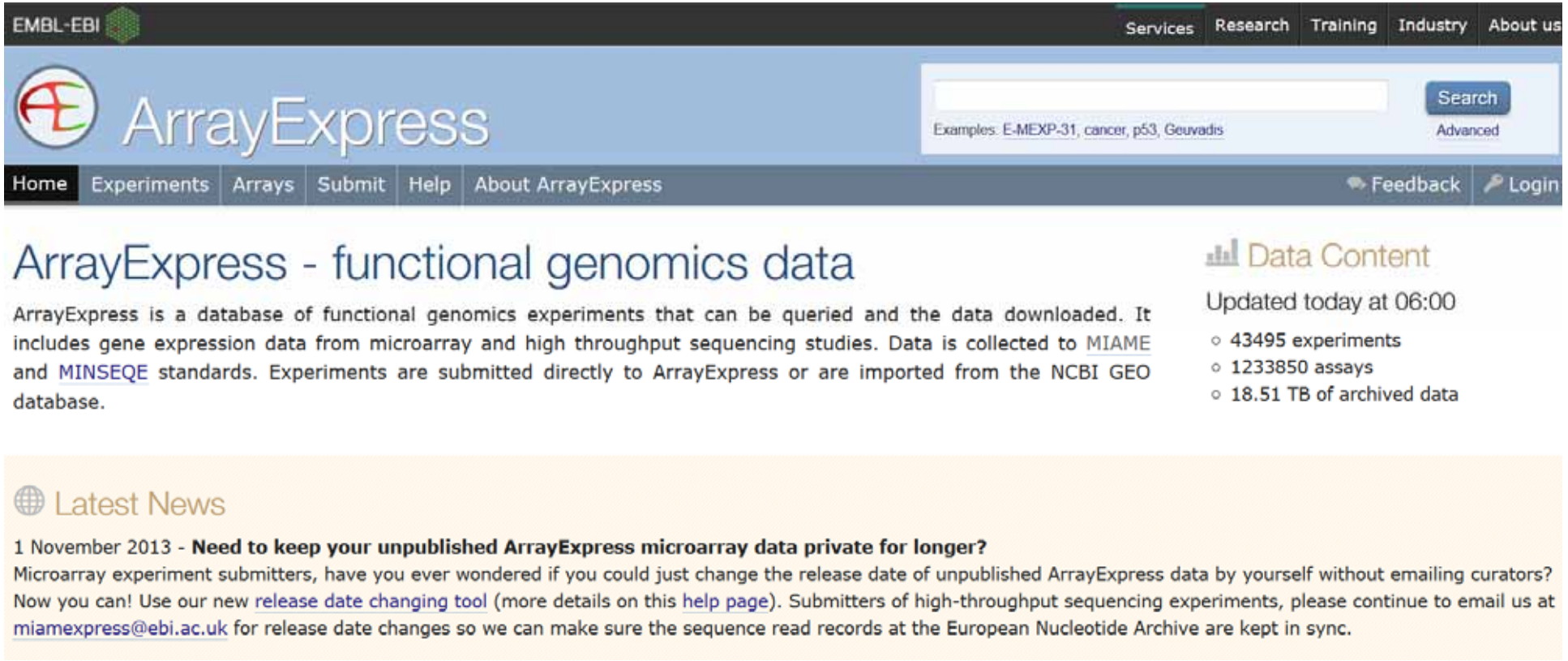
van Kampen (2012), Bioinformatics Laboratory, Academic Medical Center, NL

Slide 4-22: Example Genome Database: Ensembl



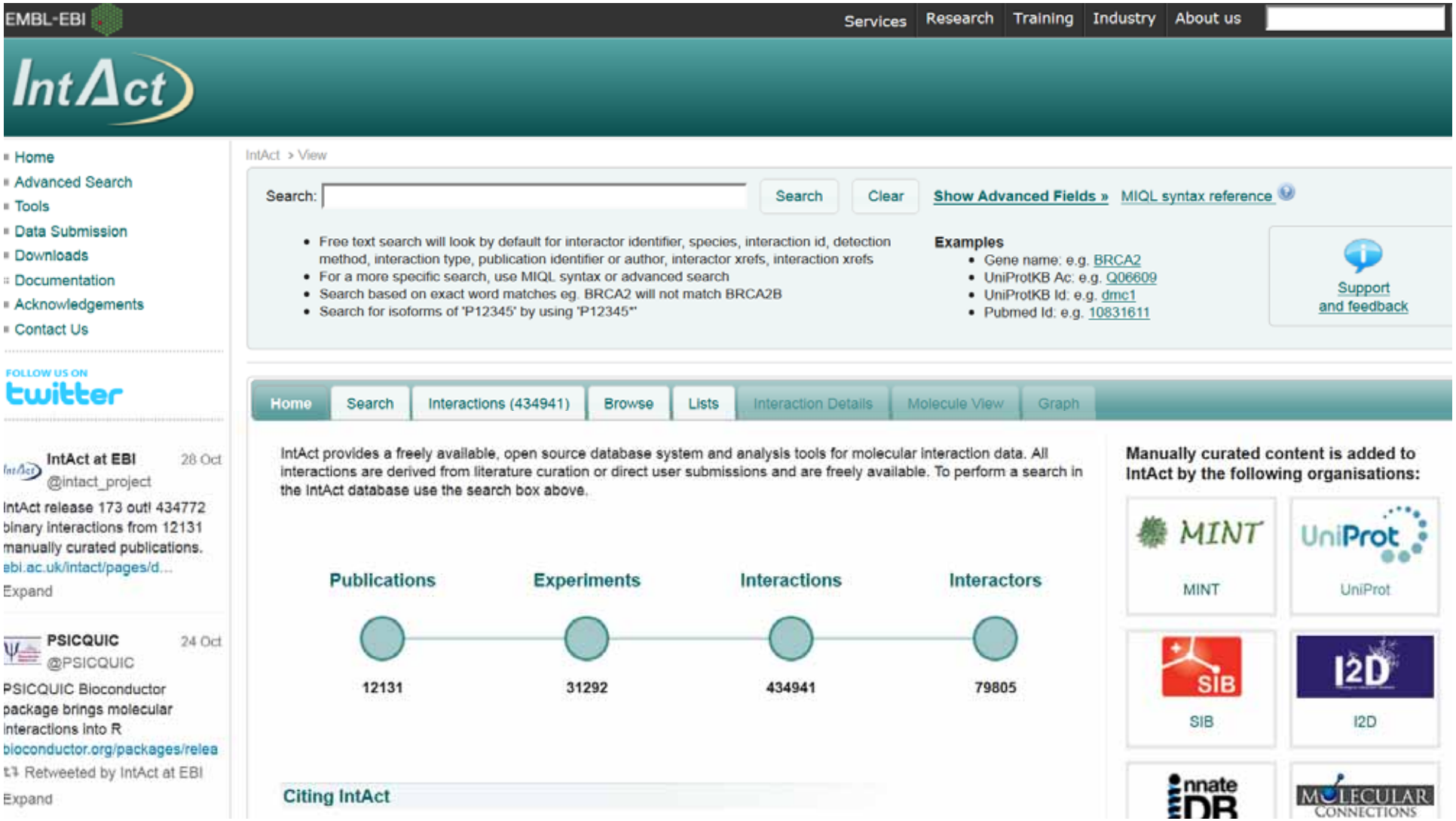
http://www.ensembl.org/index.html

Slide 4-23 Ex. Gene Expression Database: ArrayExpress



http://www.ebi.ac.uk/arrayexpress/

Slide 4-24: Example Protein Interaction Database: IntAct



http://www.ebi.ac.uk/intact/

Slide 4-25: Example for Systems Database: BioModels

EMBL-EBI

Services | Research | Training | Industry | About us

BioModels Database

Search

Advanced

BioModels Home | Models | Submit | Support | About BioModels | Contact us

McAuley et al., (2012). A whole-body mathematical model of cholesterol metabolism and its age-associated dysregulation.

October 2013, model of the month by *Nick Juty*
Original model: BIOMD0000000434

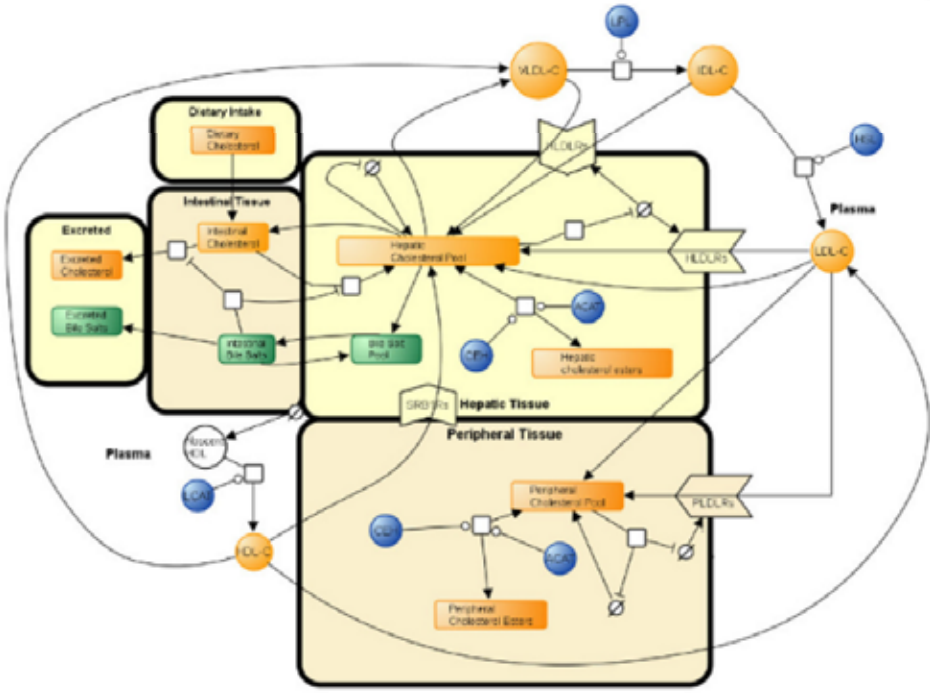
Cardiovascular disease is by far the most prevalent disease in ageing populations. Correlated with alterations in lipid metabolism profiles, it has estimated incidence rate of 30-40% in the UK population, over the age 85. Low Density Lipoprotein Cholesterol (LDL-C), a prominent component in lipid metabolism, stands out as a major contributory factor. Furthermore, it is apparent that neither nutritional status nor physical activity have any effect on the rising levels of LDL-C with age.

Besides its well publicized detrimental effects, cholesterol is also an important component of all cell membranes, being a hormone precursor and playing a crucial role in absorption of lipid soluble vitamins. It's absorption from the gut is documented as being inefficient, and also displays high variability between individuals (30-80%). The precise transport and enzymatic mechanisms involved, particularly pertaining to how cholesterol traverses enterocyte membranes, is not well established.

The hepatic system is central in cholesterol metabolism, with the liver able to synthesize VLDs (very low density lipoproteins), which are converted into IDLs (intermediate density lipoproteins (IDLs) through the action of lipoprotein lipase (LPL). LPLs can be taken up by the liver directly, or further hydrolysed into LDLs, the main cholesterol carrier in the blood. LDLs may also be taken up through the LDL-receptor (LDLR), which is highly expressed in the liver, and expressed in peripheral tissues. The hepatic receptor is transcriptionally regulated by intracellular cholesterol levels.

It has been demonstrated that: a) There is age-associated decline in the clearance rate of LDL-C from the blood, as well as a decrease in the number of hepatic LDLRs. b) Intestinal cholesterol absorption increases with age in some species.

In this paper, the authors take a mechanistic approach to construct a model, with these observations in mind, making extensive use of published experimental measurements over the last seventy years. The model incorporates dietary cholesterol absorption in the intestine, and hepatic LDL-C clearance from the plasma [1, BIOMD0000000434]. It consists of 6 compartments (Figure 1), and is composed of a series of coupled ODEs.

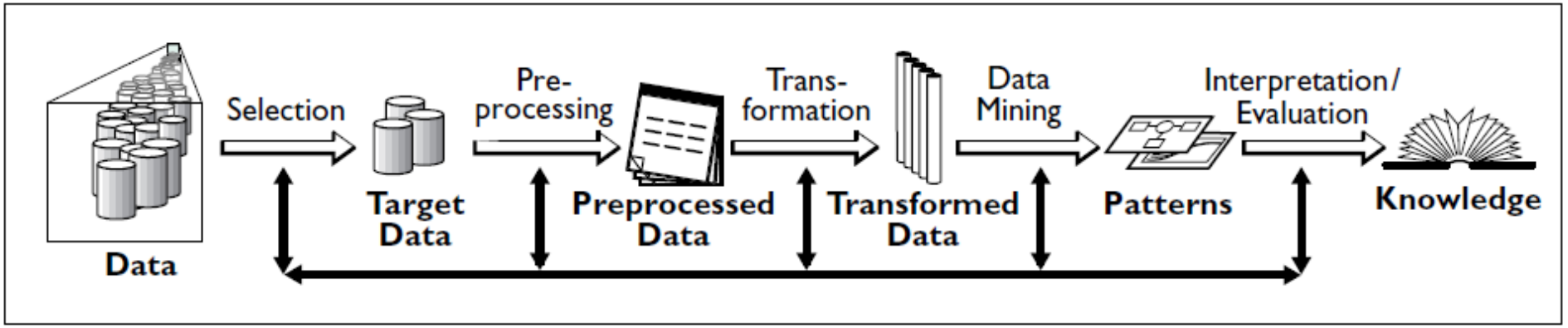


http://www.ebi.ac.uk/biomodels-main/

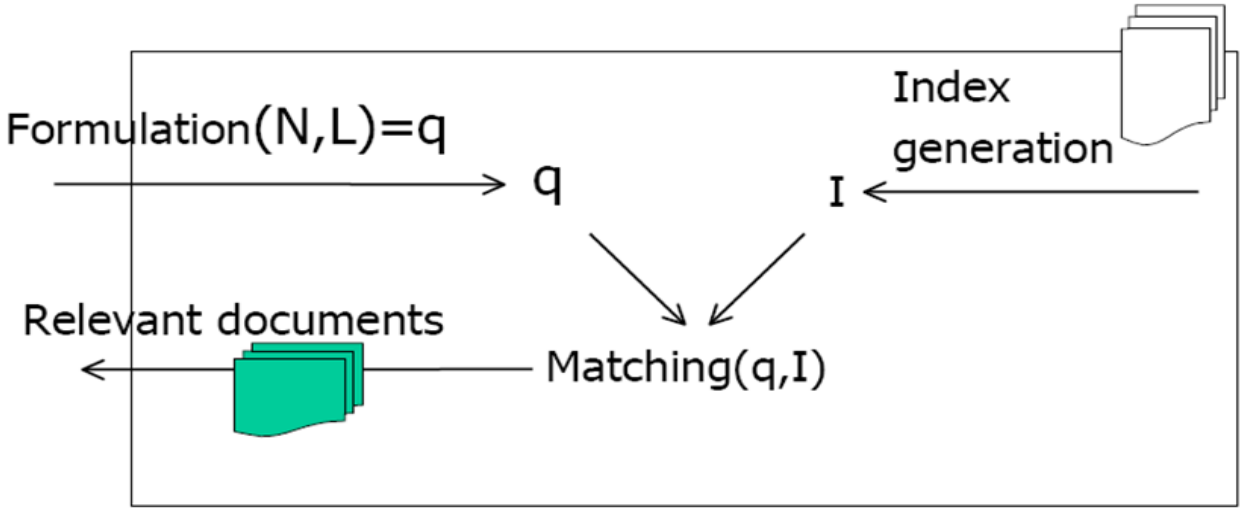
What is this animal doing?

What is the difference between retrieval and discovery?

Slide 4-26: Data Mining/KDD versus Information Retrieval



Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. *Communications of the ACM, 39, (11), 27-34.*



Baeza-Yates, R. & Ribeiro-Neto, B. 2011. *Modern Information Retrieval: The Concepts and Technology behind Search, Harlow et al., Pearson.*

What is the difference between data retrieval and information retrieval?

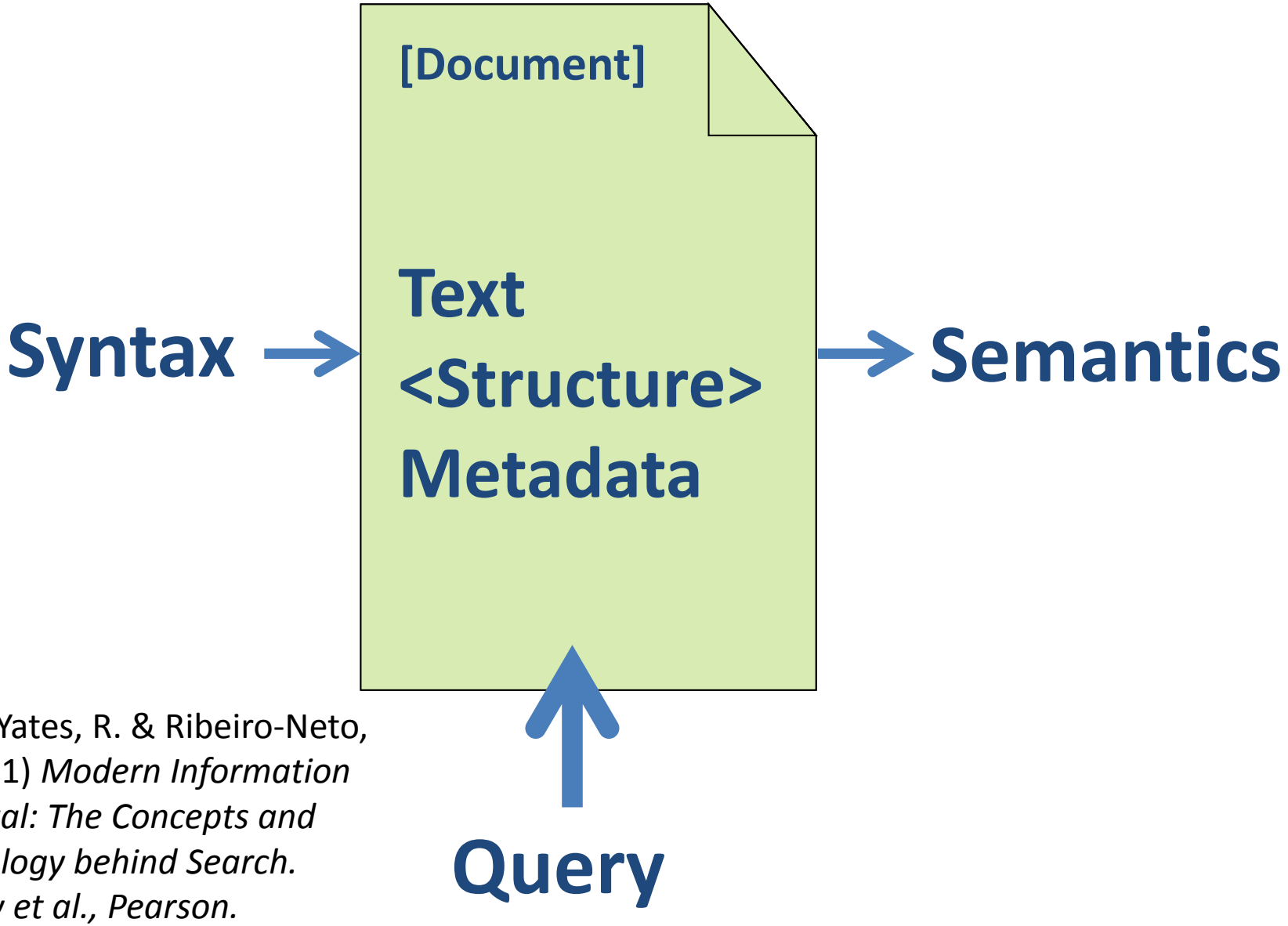
Slide 4-27: Data retrieval (DR) vs. Information retrieval (IR)

- IR is used to satisfy the end-users' information needs.
- Def.: IR deals with the representation, storage, organization of and access to information objects.

Factor	Data Retrieval (DR)	Information Retrieval (IR)
Model	Deterministic	Probabilistic
Matching	Exact match	Partial (best match)
Inference	Deduction	Induction
Classification	Monothetic*	Polythetic**
Query language	Artificial (abstract)	Natural
Query specification	Must be complete	Can be incomplete
Items wanted	matching	relevant
Error response	sensitive	insensitive

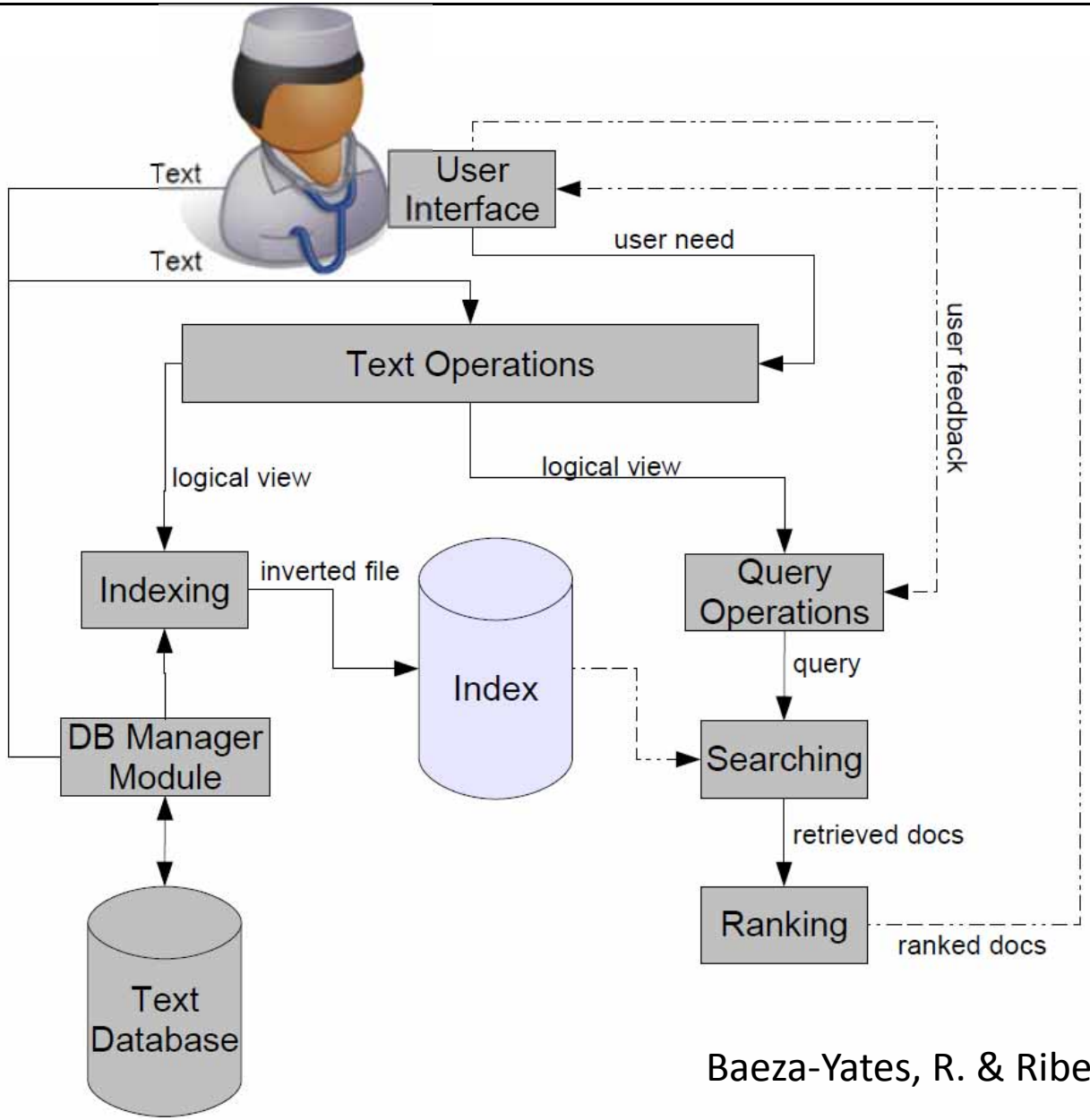
*Monothetic = type in which all members are identical on all characteristics;
**Polythetic = type in which all members are similar, but not identical;

Van Rijsbergen, C. J. (1979) *Information Retrieval (Second Edition). London, Butterworths.*



Baeza-Yates, R. & Ribeiro-Neto, B. (2011) *Modern Information Retrieval: The Concepts and Technology behind Search. Harlow et al., Pearson.*

Slide 4-29: IR Process principle



Baeza-Yates, R. & Ribeiro-Neto, B. (2011)

Slide 4:30: Formal Description of IR Models

Definition: Let the **IR Model** be a quadruple {**D**, **Q**, **F**, $R(q_i, d_j)$ }

- **D** is a set composed of logical views (representation component) of the **documents** within a collection;
- **Q** is a set of logical views (representation component) of the user information needs (these are called "**queries**");
- **F** is a framework for modeling document representations, queries and their relationships (reasoning component);

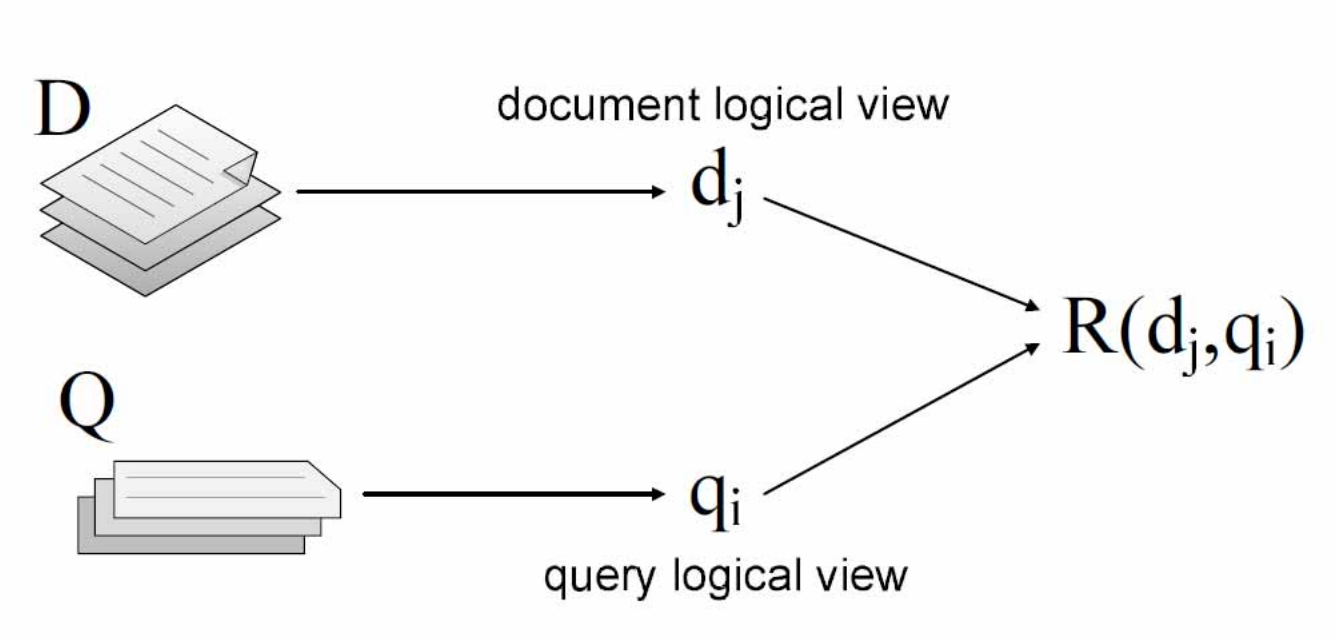
This includes sets and Boolean relations, vectors and linear algebra operations, sample spaces and probability distributions;

- $R(q_i, d_j)$ is a ranking function that associates a real number with a query representation $q_i \in$ **Q** and a document representation $d_j \in$ **D**.

Such ranking defines an ordering among the docs with regard to the query q_i

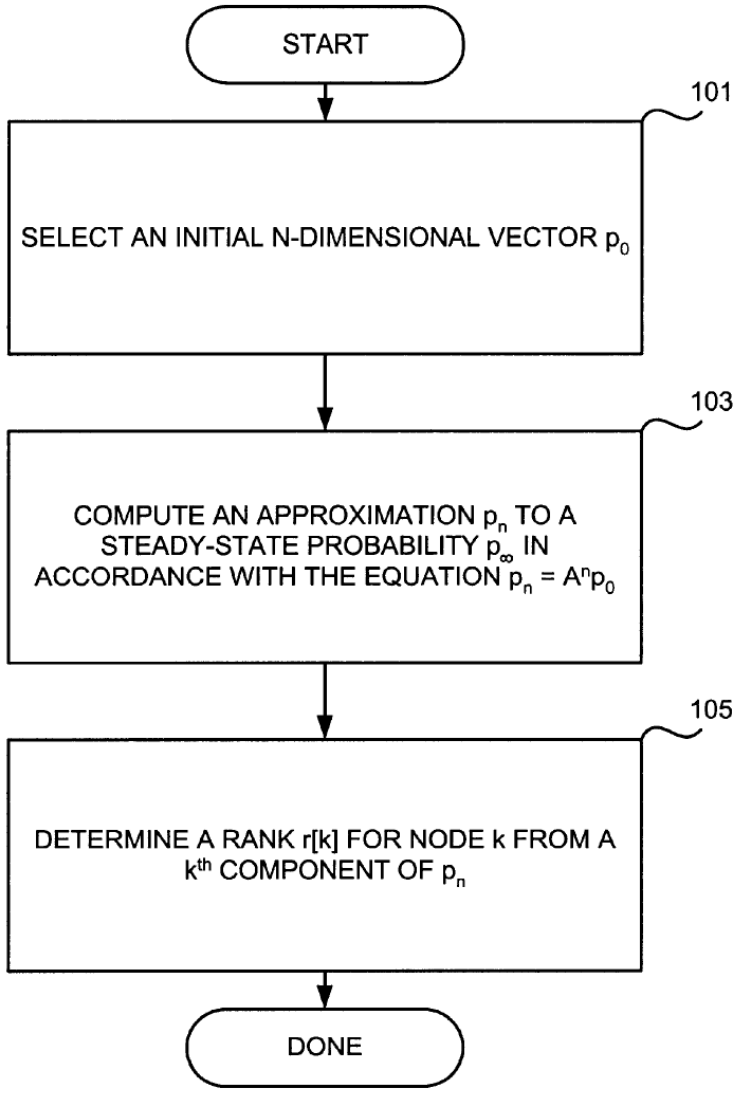
Baeza-Yates, R. & Ribeiro-Neto, B. (2011) *Modern Information Retrieval: The Concepts and Technology behind Search. Harlow et al., Pearson.*

Slide 4-31 Logical view (query Q and document D)



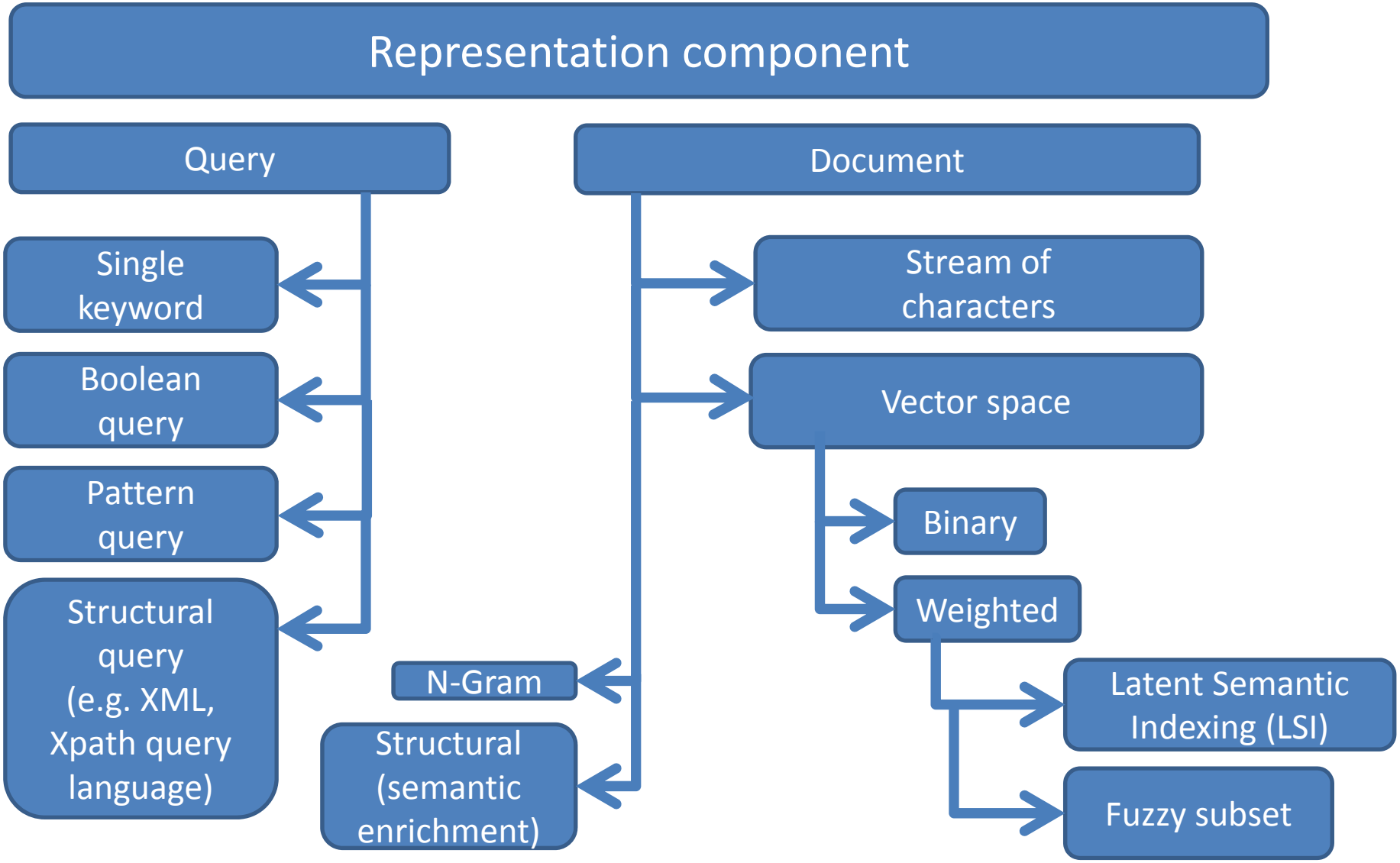
Baeza-Yates, R. & Ribeiro-Neto, B. (2011) *Modern Information Retrieval: The Concepts and Technology behind Search. Harlow et al., Pearson.*

Slide 4-32 A famous Example



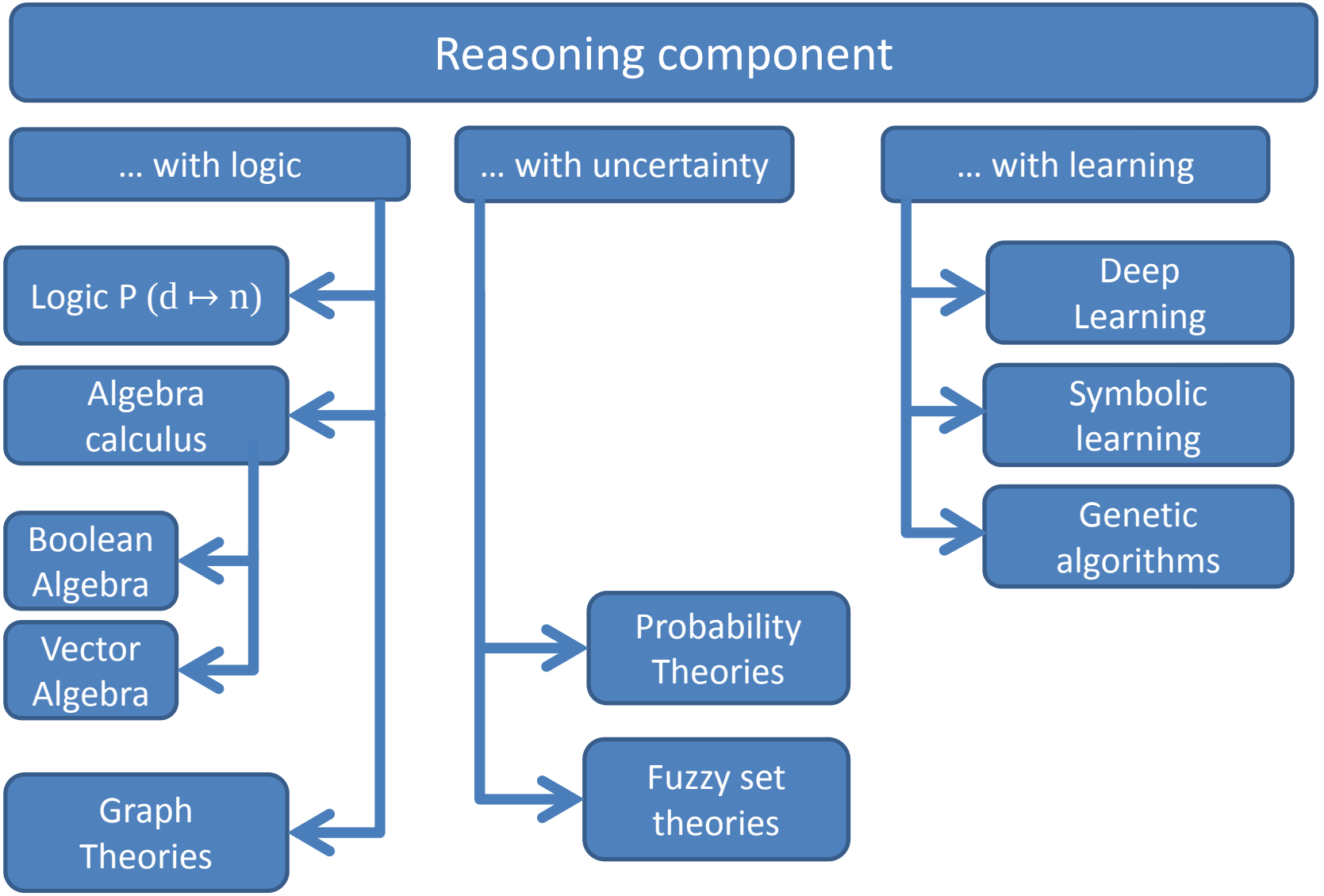
Remember: We have two components: Representation and Reasoning component

Slide 4-33: Taxonomy of Information Retrieval Models 1/3



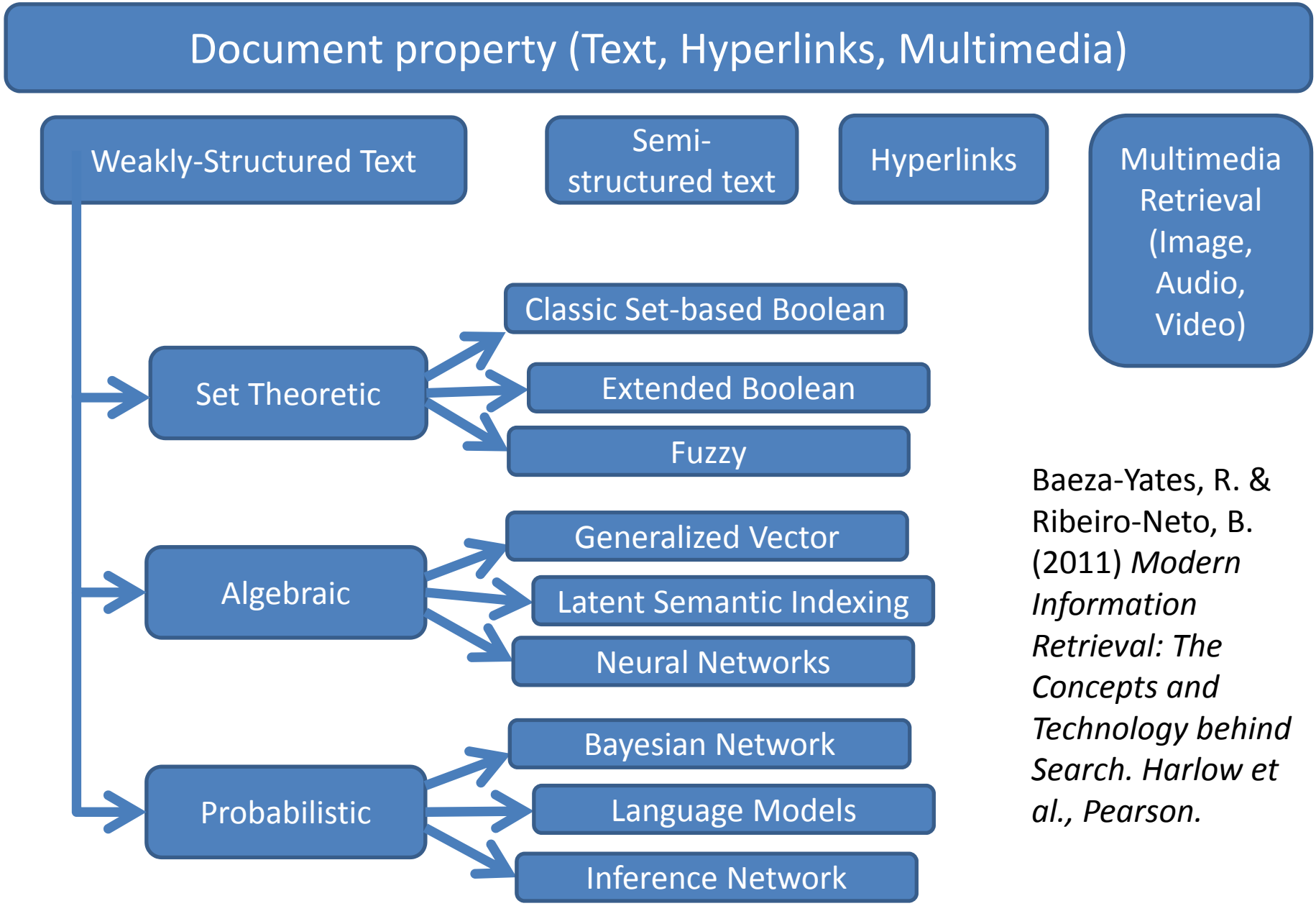
Canfora, G. & Cerulo, L. (2004) A Taxonomy of Information Retrieval Models and Tools. *Journal of Computing and Information Technology (CIT), 12, 3, 175-194.*

Slide 4-34: Taxonomy of Information Retrieval Models 2/3



Canfora, G. & Cerulo, L. (2004) A Taxonomy of Information Retrieval Models and Tools. *Journal of Computing and Information Technology (CIT), 12, 3, 175-194.*

Slide 4-35: Taxonomy of Information Retrieval Models 3/3



Baeza-Yates, R. & Ribeiro-Neto, B. (2011) *Modern Information Retrieval: The Concepts and Technology behind Search. Harlow et al., Pearson.*

Slide 4-36: Set Theoretic Example: Boolean Model

- Documents and queries are represented as a set of index terms; the queries are Boolean expressions (AND, OR, NOT);

"For the Boolean model, the index term weight variables are all binary i.e., $\omega_{i,j} \in \{0,1\}$. *A query* q *is a conventional Boolean expression. Let* $\vec{q}_{dnf}$ *be the disjunctive normal form for the query* q. *Further, let* $\vec{q}_{cc}$ *be any of the conjunctive components of* $\vec{q}_{dnf}$. *The similarity of a document* d_j *to the query* q *is defined as*

$$sim(d_j, q) = \begin{cases} 1 & if\ \exists\ \vec{q}_{cc} | (\vec{q}_{cc} \in \vec{q}_{dnf}) \wedge (\forall_{k_i}, g_i(\vec{d}_j) = g_i(\vec{q}_{cc})) \\ 0 & otherwise. \end{cases}$$

If $sim(d_j, q) = 1$ *then the Boolean model predicts that the document* d_j *is relevant to the query* q *(it might not be). Otherwise, the prediction is that the document is not relevant."*

Baeza-Yates, R. & Ribeiro-Neto, B. (2011)

Slide 4-37: Set Theor. Model - Boolean Model Pros & Cons

Advantages	Disadvantages
Easy to understand	No partial matches
Exact formalism	The “bag-of-words” representation does not accurately consider the semantics of documents *
Query language is expressive	Query language is complicated
	Retrieved documents cannot be ranked

*) refer to: Vallet, D., Fernández, M. & Castells, P. (2005) An Ontology-Based Information Retrieval Model. In: Gómez-Pérez, A. & Euzenat, J. (Eds.) *The Semantic Web: Research and Applications. Berlin, Heidelberg, Springer, 103-110.*

D = $\langle d_1, d_2, \ldots\ d_n \rangle$ (collection of medical docs)

d_i =$t_1, t_2, \ldots t_k$ (every document consists of terms)

Now we carry out a document transformation and get vectors:

$$w_{i,j} = \begin{cases} 1, & t_i \in d_j \\ 0, & t_i \notin d_j \end{cases} \rightarrow d_j = (0, 1, 1, 0, 1, \ldots, 1)^T$$

Now we count the frequency of the terms and get:

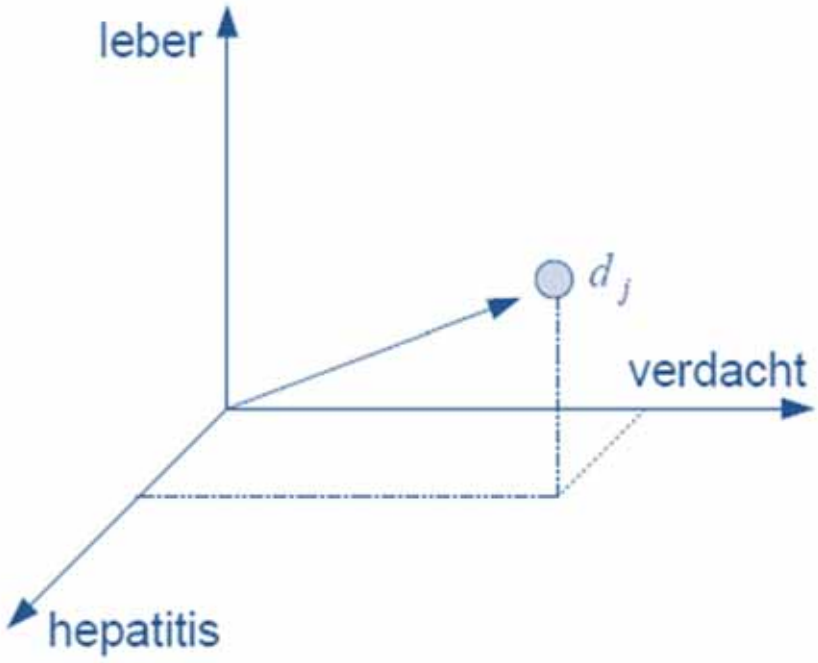
$$w_{i,j} = \begin{cases} (1 + \log f_{i,j}) * \log \frac{N}{n_i}, & \text{if } f_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases}$$

Slide 4-39: As a result we get a matrix …

$$D_{m\times n} = \begin{Bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n-1} & w_{1,n} \\ w_{2,1} & w_{2,2} & & w_{2,n-1} & w_{2,n} \\ \vdots & & \ddots & & \vdots \\ w_{m-1,1} & w_{m-1,2} & & w_{m-1,n} & w_{m-1,n} \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n-1} & w_{m,n} \end{Bmatrix}$$

Salton, G., Wong, A. & Yang, C. S. 1975. Vector-Space Model for automatic indexing. *Communications of the ACM,* 18, (11), 613-620.

Slide 4-40: dj can thus be seen as a point in n-dim space



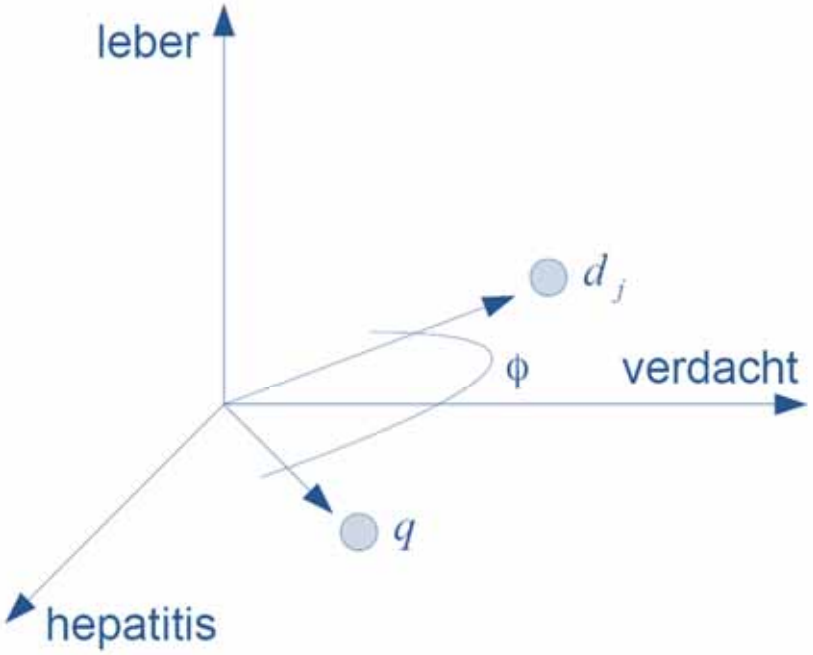
One of the biggest obstacles to making full use of the power of computers is that they currently understand very little of the meaning of human language.

Recent progress in search engine technology is only scratching the surface of human language, and yet the impact on society and the economy is already immense.

Vector space models (VSMs) are likely to be a part of these new semantic technologies.

Turney, P. D. & Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. Journal of artificial intelligence research, 37, (1), 141-188.
Survey article 922 citations yet …

Slide 4-41: Use the cos-similarity for ranking similar docs



$$\cos(\phi) = \frac{q.d_j}{\| q \| \| d_j \|}$$

$$sim\left(\vec{d}_j, \vec{q}\right) = cos\left(\Phi\right) = \frac{\vec{d}_j \bullet \vec{q}}{\left|\vec{d}_j\right| \times |\vec{q}|}$$

$$= \frac{\sum_{i=1}^{t} \omega_{i,j} \times \omega_{i,q}}{\sqrt{\sum_{i=1}^{t} \omega_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} \omega_{i,q}^2}}$$

Slide 4-42: Algebraic Vector Space Model: Pros & Cons

Advantages	Disadvantages
Easy to understand	Higher effort to calculate similarity
Partial matches possible	The “bag-of-words” representation does not accurately consider the semantics of documents *
Sorting of documents by rank	
Using term weighting schemes	

*) refer to: Vallet, D., Fernández, M. & Castells, P. (2005) An Ontology-Based Information Retrieval Model. In: Gómez-Pérez, A. & Euzenat, J. (Eds.) *The Semantic Web: Research and Applications. Berlin, Heidelberg, Springer, 103-110.*

Slide 4-43: Example: Probabilistic Model (Bayes' rule)

"For the probabilistic model, the index weight variables are all binary i.e., $\omega_{i,j} \in [0,1]$, $\omega_{i,q} \in [0,1]$. *A query* q *is a subset of index terms. Let* R *be the set of documents known (or initially guessed) to be relevant. Let* $\overline{R}$ *be the complement of* R *(i.e., the set of non-relevant documents). Let* $P(R|\vec{d}_j)$ *be the probability that the document* d_j *is relevant to the query* q *and* $P(\overline{R}|\vec{d}_j)$ *be the probability that* d_j *is non-relevant to* q. *The similarity* $sim(d_j, q)$ *of the document* d_j *to the query* q *is defined as the ratio*

$$sim(d_j, q) = \frac{P(R|\vec{d}_j)}{P(\overline{R}|\vec{d}_j)}$$

$$sim(d_j, q) = \frac{P(\vec{d}_j|R) \times P(R)}{P(\vec{d}_j|\overline{R}) \times P(\overline{R})}$$

Rev. Thomas Bayes (1702-1761)

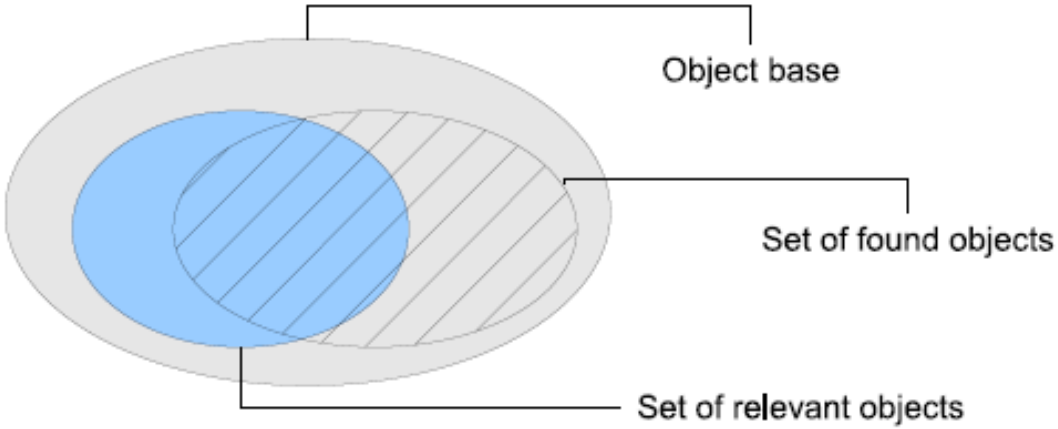
$$sim(d_j, q) \sim \frac{(\prod_{g_i(\vec{d}_j)=1} P(k_i|R)) \times (\prod_{g_i(\vec{d}_j)=0} P(\overline{k_i}|R))}{(\prod_{g_i(\vec{d}_j)=1} P(k_i|\overline{R})) \times (\prod_{g_i(\vec{d}_j)=0} P(\overline{k_i}|\overline{R}))}.$$

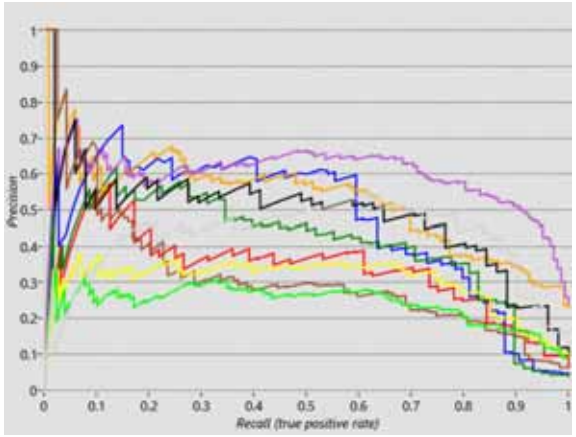
Slide 4-44: Probabilistic Model: Pros & Cons

Advantages	Disadvantages
Documents can be ranked by relevance	It is a binary model (→ binary weights)
	The index terms are assumed to be independent and a lack of document normalization
	There is a need to guess the initial separation of documents into relevant and non-relevant sets

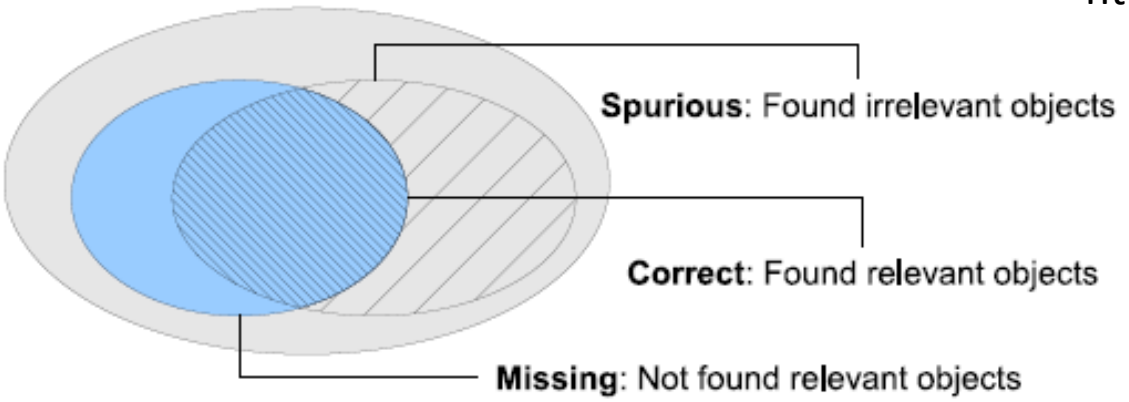
How can we measure the quality of the IR?

Slide 4-45: Measuring the Quality of Information Retrieval





http://www.mbmlbook.com/

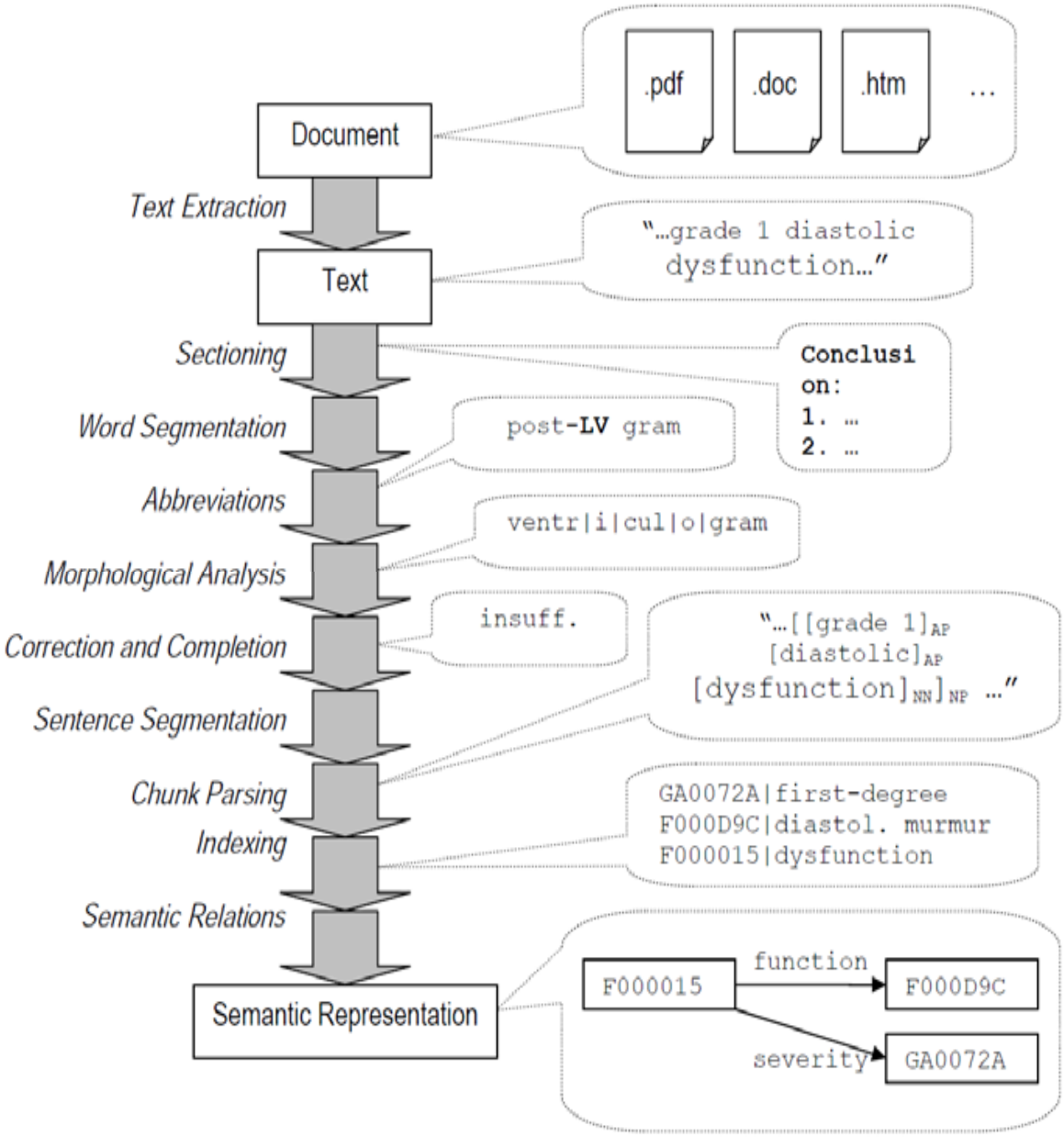


$$Recall = \frac{Correct}{Correct + Missing}$$

$$Precision = \frac{Correct}{Correct + Spurious}$$

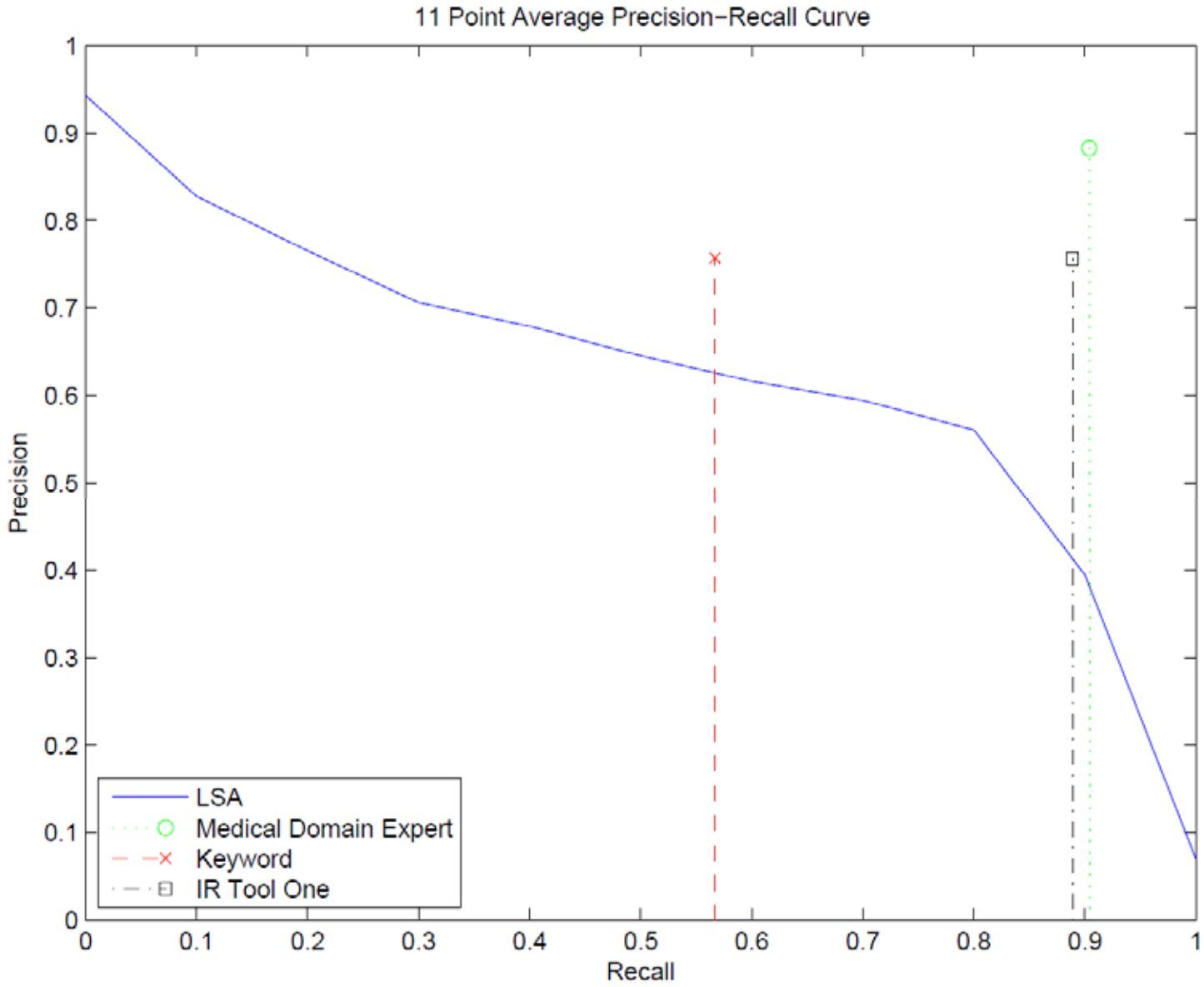
Slide 4-46: Example: Linguistic Processing Pipeline

Kreuzthaler, M., Bloice, M. D., Faulstich, L., Simonic, K. M. & Holzinger, A. (2011) A Comparison of Different Retrieval Strategies Working on Medical Free Texts. *Journal of Universal Computer Science, 17, 7, 1109-1133.*



Slide 4-47: Example: Point Average Precision-Recall Graph

Kreuzthaler, M., Bloice, M. D., Faulstich, L., Simonic, K. M. & Holzinger, A. (2011) A Comparison of Different Retrieval Strategies Working on Medical Free Texts. *Journal of Universal Computer Science, 17, 7, 1109-1133.*

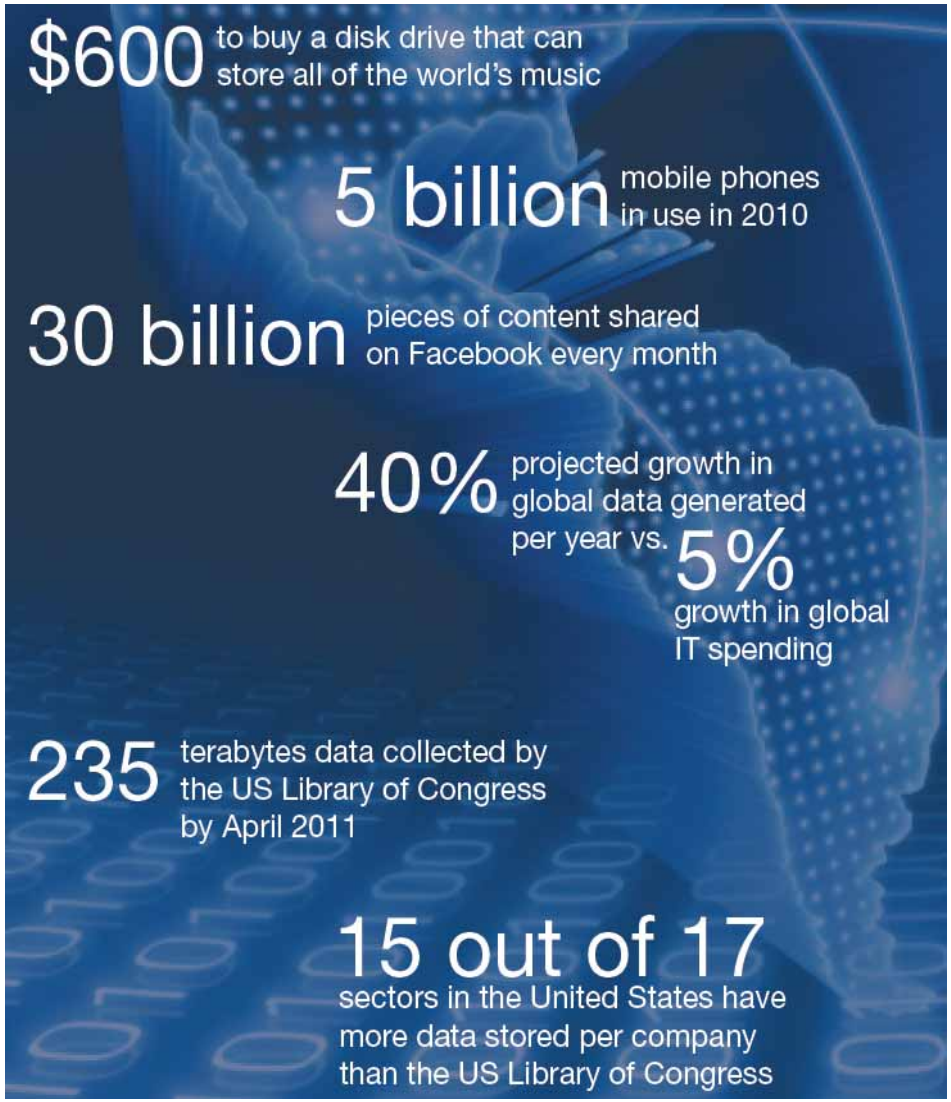


Slide 4-48: Big data – a growing torrent in the future

McKinsey Global Institute

June 2011

Big data: The next frontier for innovation, competition, and productivity



Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

Complex Data!

What is interesting?
What is relevant?

Thank you!

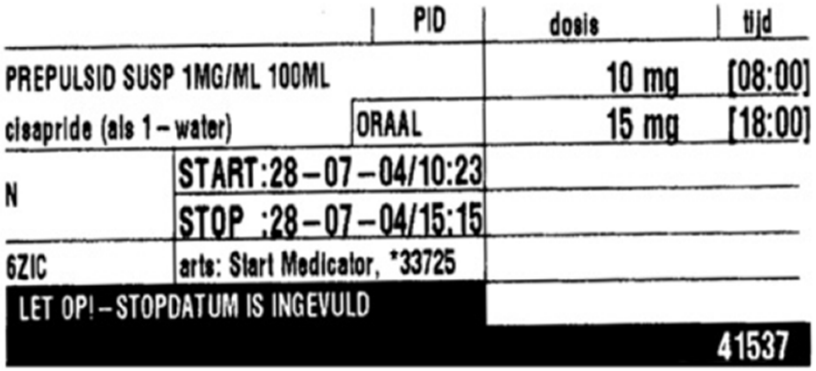
Sample Questions (1)

- What is typical for medical workflows?
- How is the workflow in the clinical control loop?
- What does each shell in the Hospital Activity Shell model express?
- Of which main parts does the classic conceptual model of a Hospital Information System consist?
- What is a data mart?
- Why is the physician order entry a critical process?
- What is business intelligence in the context of a HIS?
- What is the difference between Information Extraction and Information Retrieval?
- Which differences exist between Data Retrieval and Information Retrieval?
- What advantages/disadvantages does cloud computing in health care have?
- What is a PACS cloud?

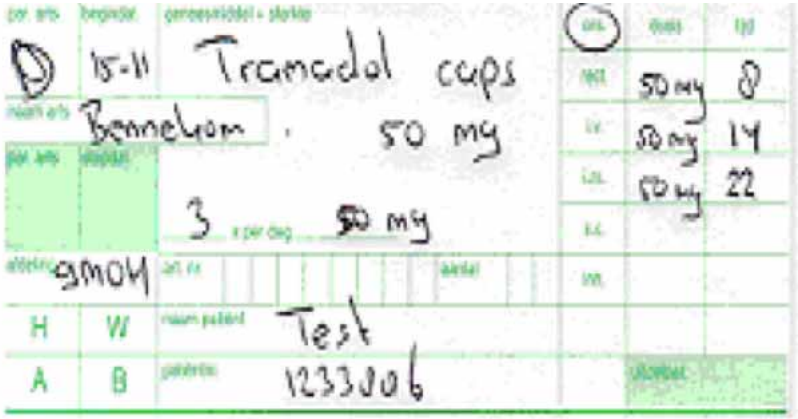
Sample Questions (2)

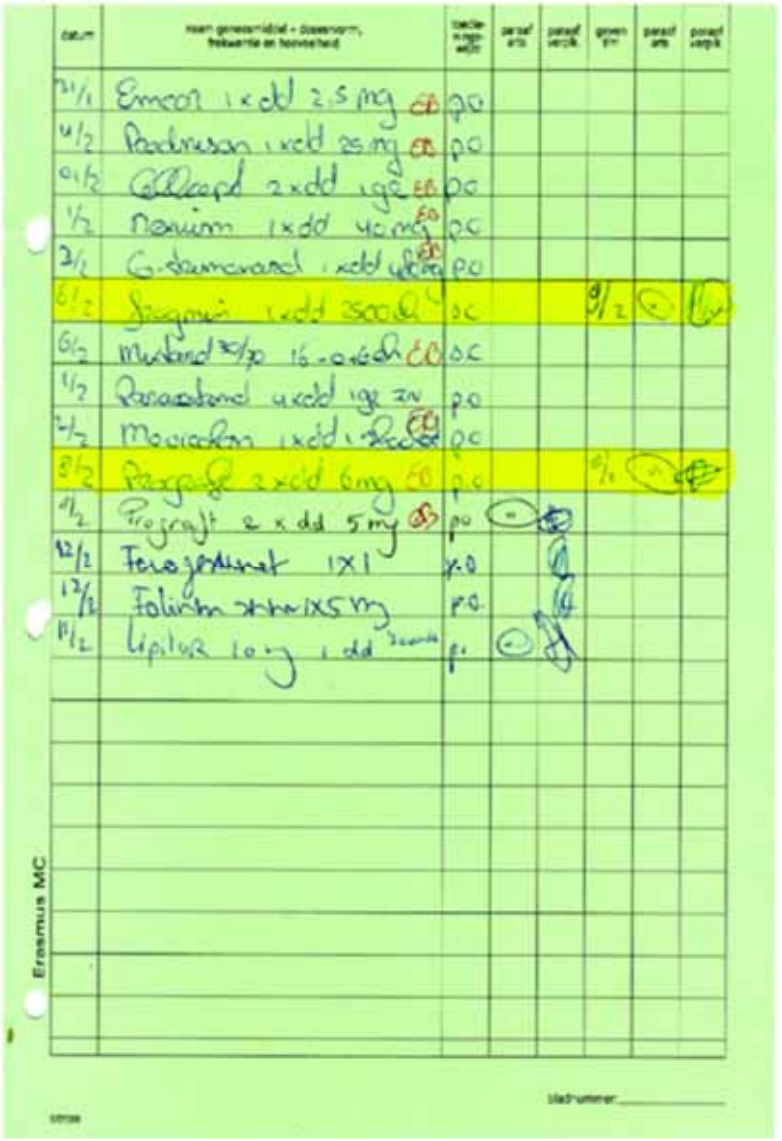
- What is the purpose of the Protein Structure Database (PDB)?
- What advantages does a integrated HIS offer?
- What is the difference between monothetic data types and polythetic data types?
- What is the purpose of medical documentation?
- How does a typical medical document look?
- What are the big difficulties in medical documents?
- How can an Information Retrieval Model be formally described?
- What is the difference between a representation component and a reasoning component?
- What advantages/disadvantages does the Boolean model have?
- Describe the principles of the Vector space model!
- Which advantage does the Probabilistic model offer?
- What is the big disadvantage of an Ontology-Based Model?
- How can you determine the quality of information retrieval?

Backup Slide: Example: Physician Order Entry (Paper)



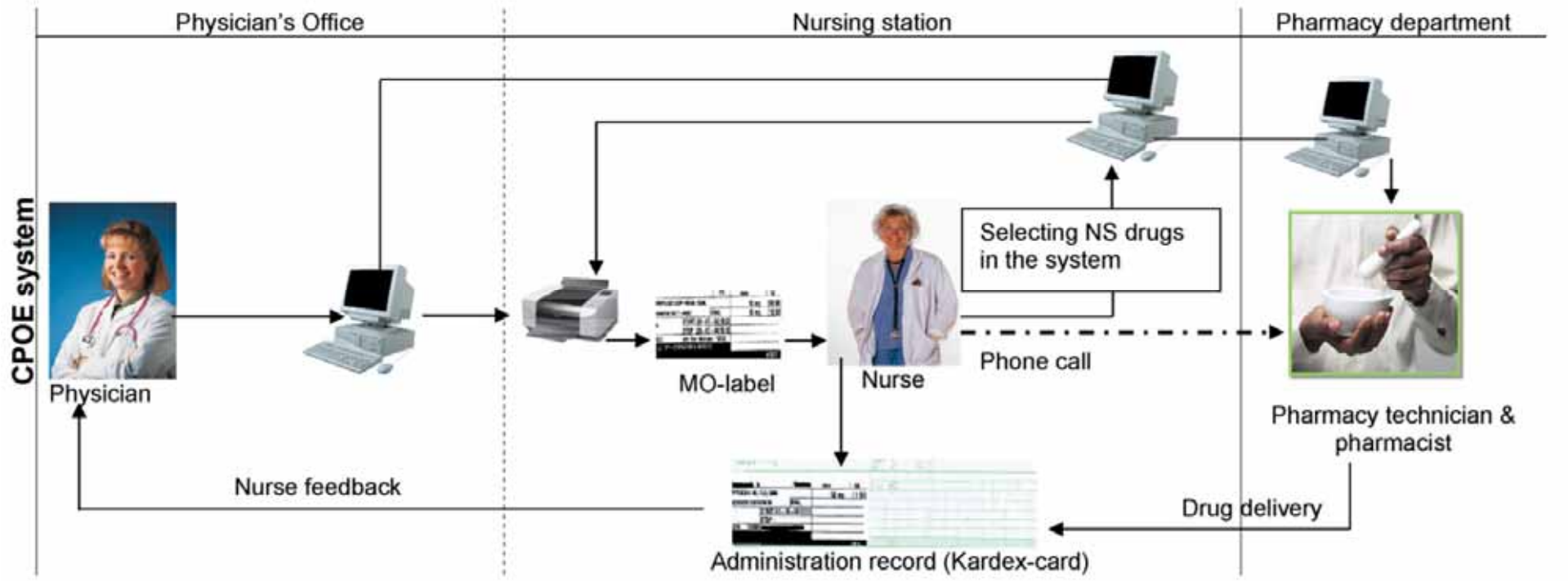
	PID	dosis	tijd
PREPULSID SUSP 1MG/ML 100ML		10 mg	[08:00]
cisapride (als 1 – water)	ORAAL	15 mg	[18:00]
N	START:28 – 07 – 04/10:23		
	STOP :28 – 07 – 04/15:15		
6ZIC	arts: Start Medicator, *33725		
LET OP! – STOPDATUM IS INGEVULD			
			41537





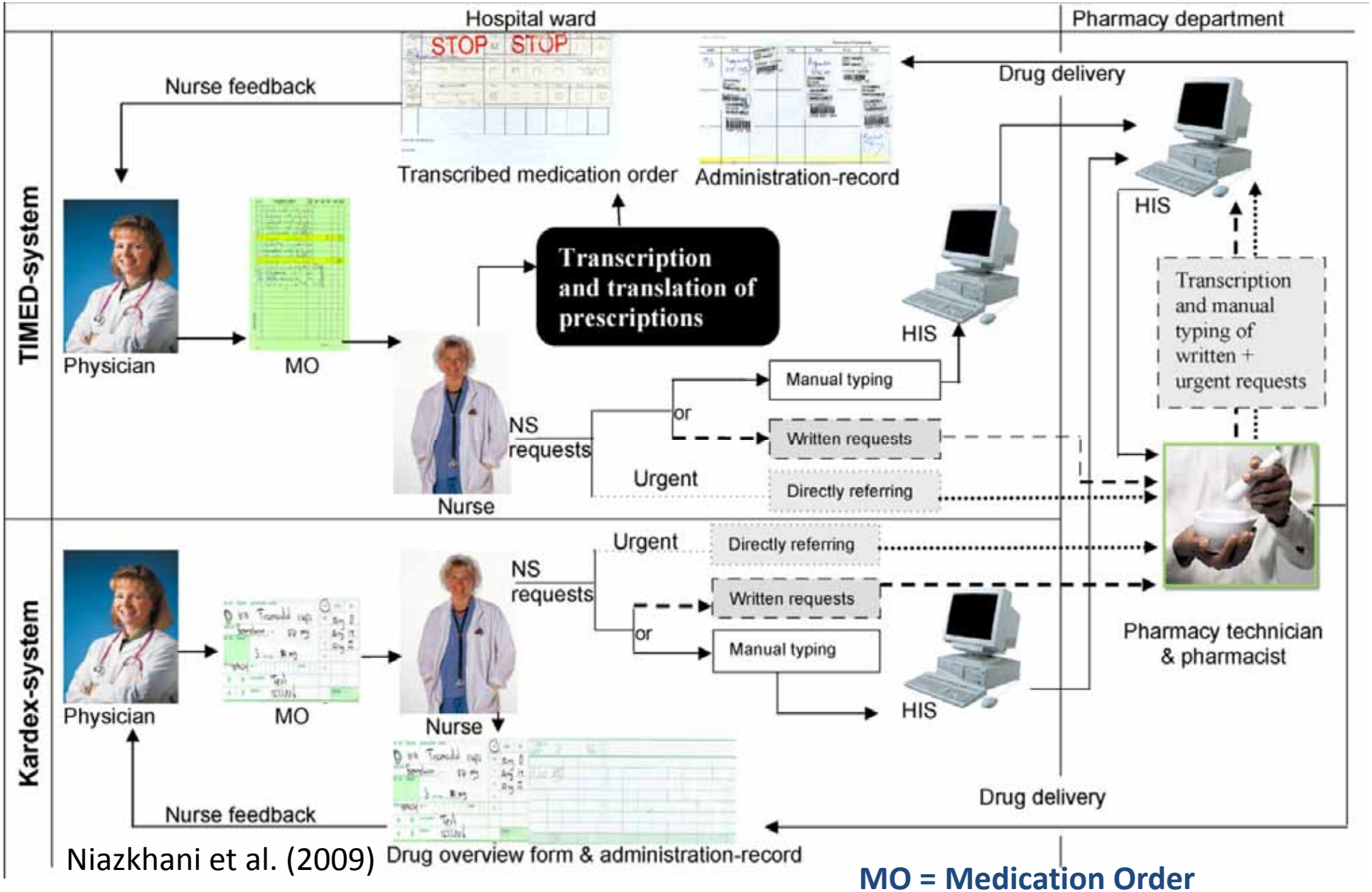
Niazkhani, Z., van der Sijs, H., Pirnejad, H., Redekop, W. K. & Aarts, J. (2009) Same system, different outcomes: Comparing the transitions from two paper-based systems to the same computerized physician order entry system. *Int. Journal of Medical Informatics, 78, 3, 170-181.*

Computerized Physician Order Entry (e-Medication)



Niazkhani et al. (2009)

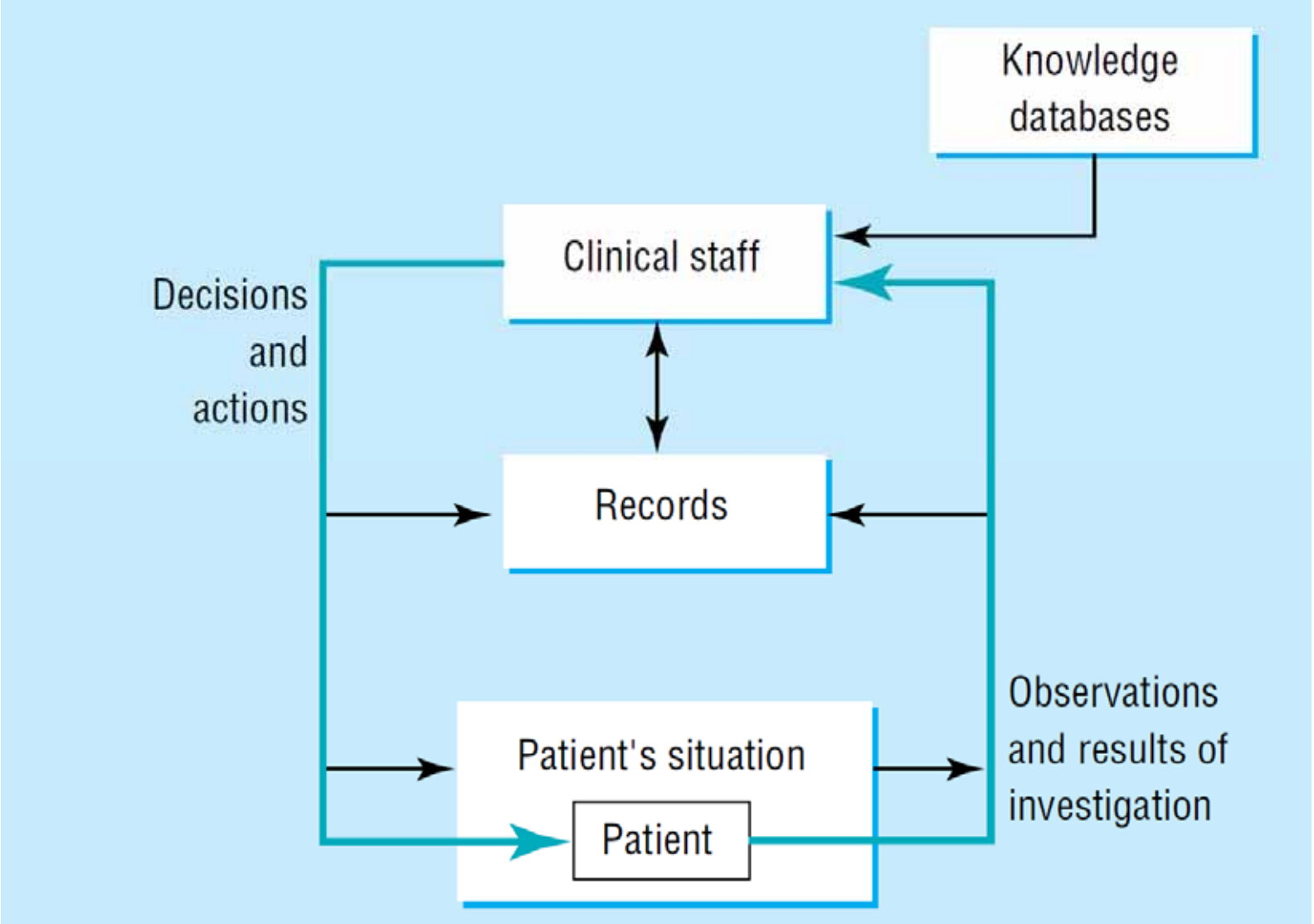
Example: Physician Order Entry (2)



Niazkhani et al. (2009)

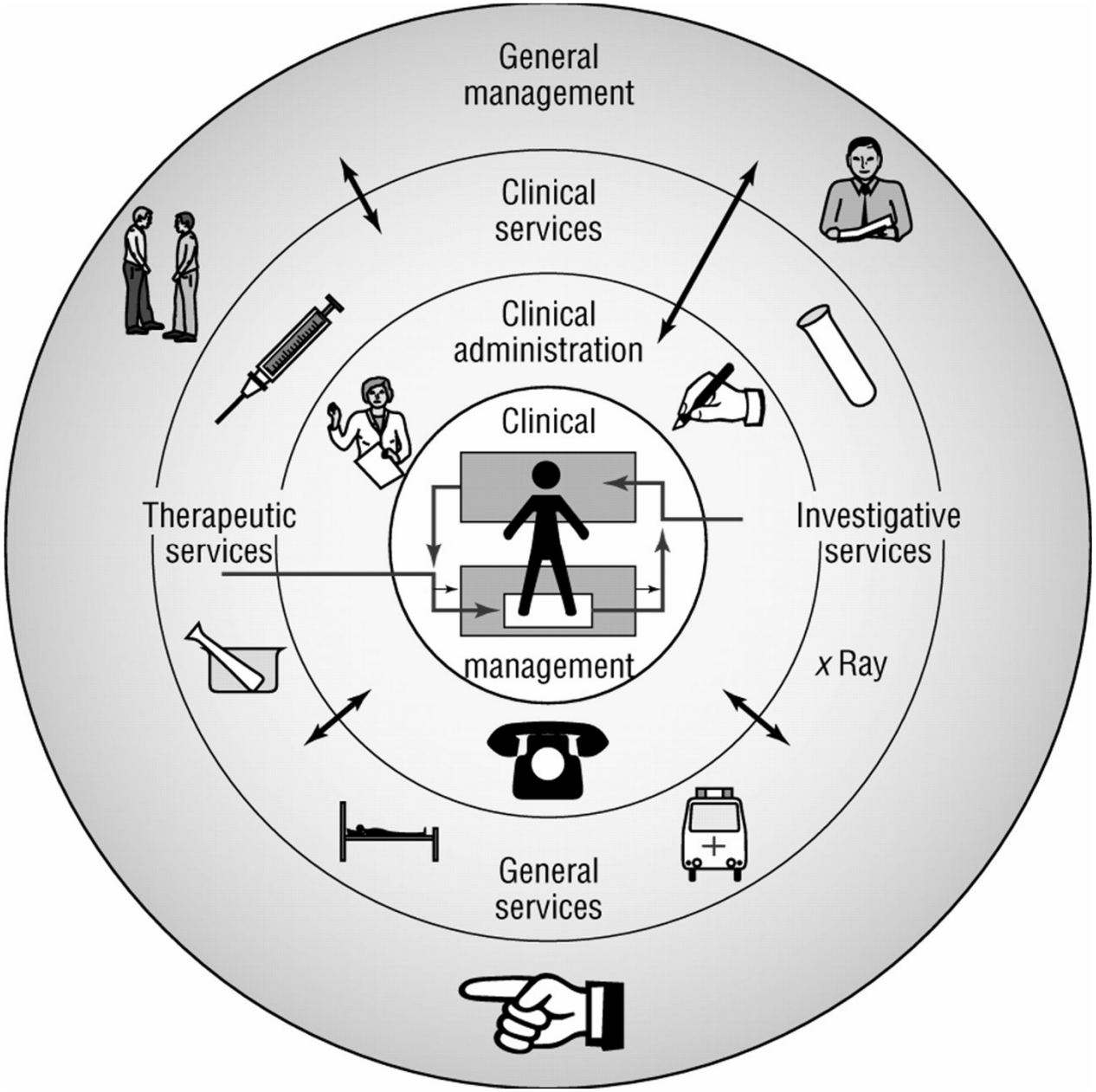
MO = Medication Order

HIS - Clinical control loop



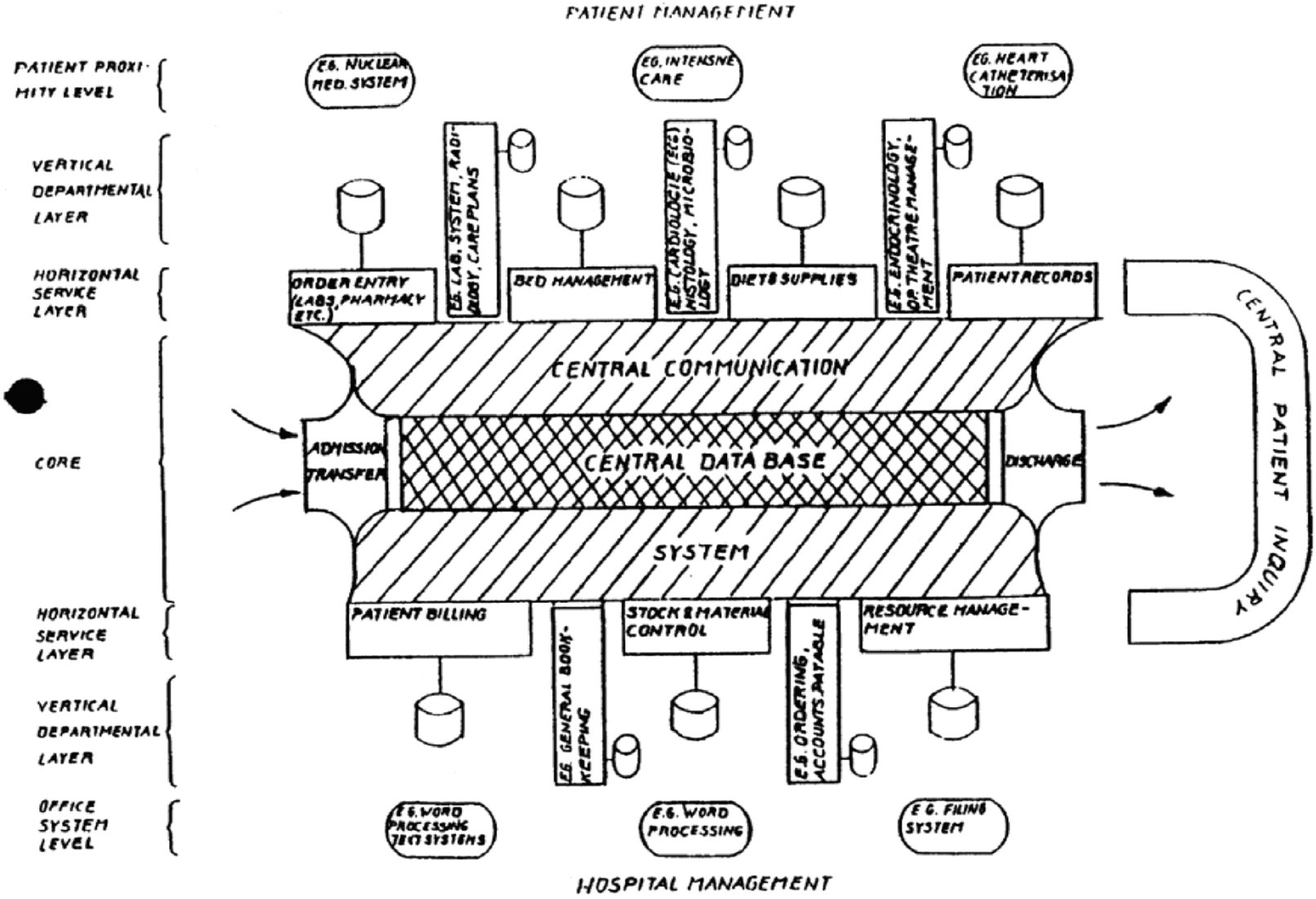
Simpson, K. & Gordon, M. (1998) The anatomy of a clinical information system. *British Medical Journal, 316, 7145, 1655-1658.*

Hospital Activity Shell – Clinical Control Loop in the center



Simpson, K. & Gordon, M. (1998) The anatomy of a clinical information system. *British Medical Journal, 316, 7145, 1655-1658.*

Slide 4-8 HIS: Historical Architecture (Original from 1984)



Haux, R. (2006) Health information systems-past, present, future. *International Journal of Medical Informatics, 75, 3-4, 268-281.*

Example: Enterprise Data Modeling (EDM) at Mayo Clinic

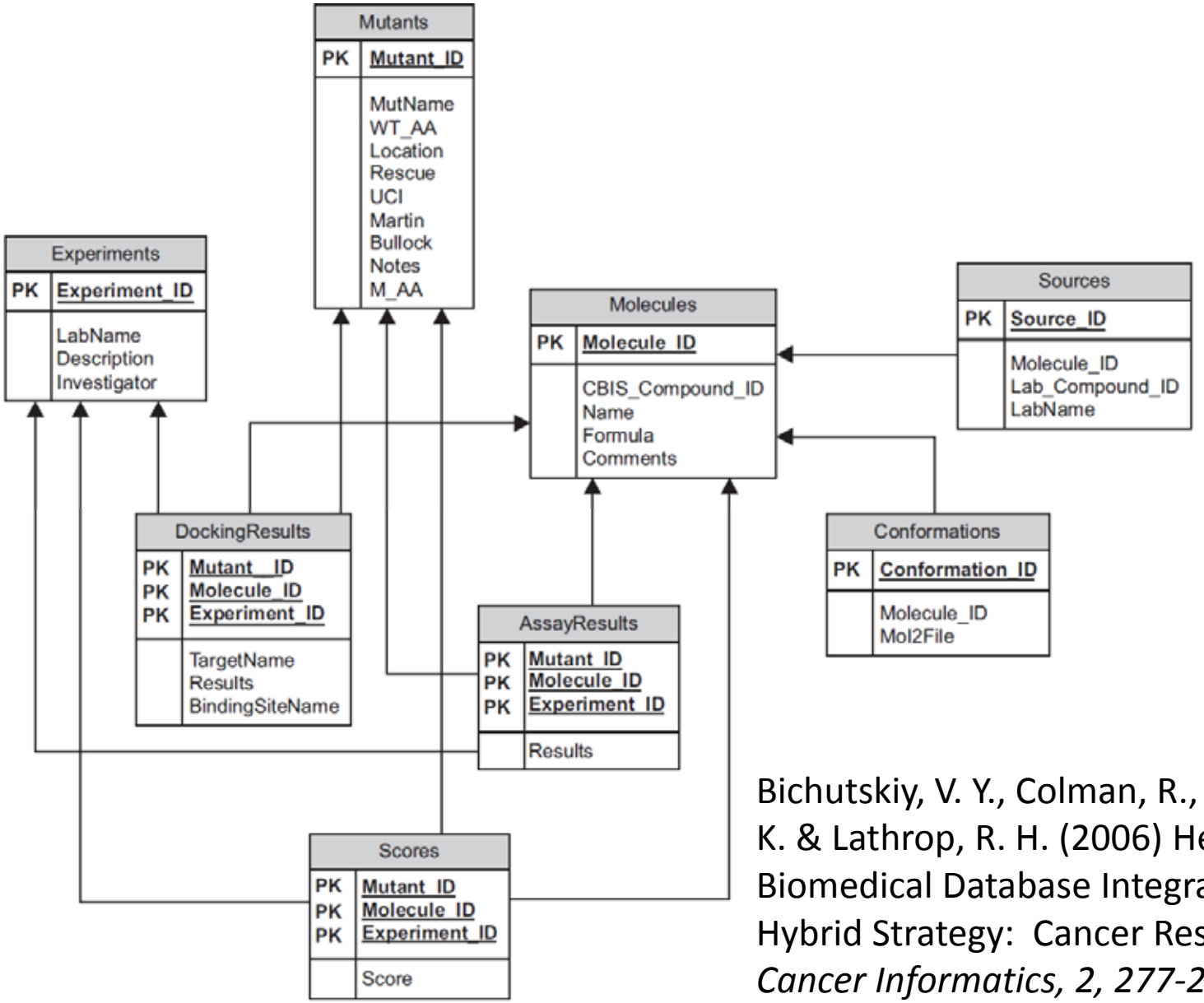
- **Subjects** = the highest level areas that define the activities of the enterprise (e.g. Individual)
- **Concepts** = the collections of data that are contained in one or more subject areas (e.g., Patient, Provider, Employee, Referrer, Volunteer, etc.)
- **Business Information Models** = the organization of the data that support the processes and workflows of the enterprise's defined Concepts.

Chute, C. G., Beck, S. A., Fisk, T. B. & Mohr, D. N. (2010) The Enterprise Data Trust at Mayo Clinic: a semantically integrated warehouse of biomedical data. *Journal of the American Medical Informatics Association, 17, 2, 131-135.*

Backup: For your own experiments: www.care2x.org

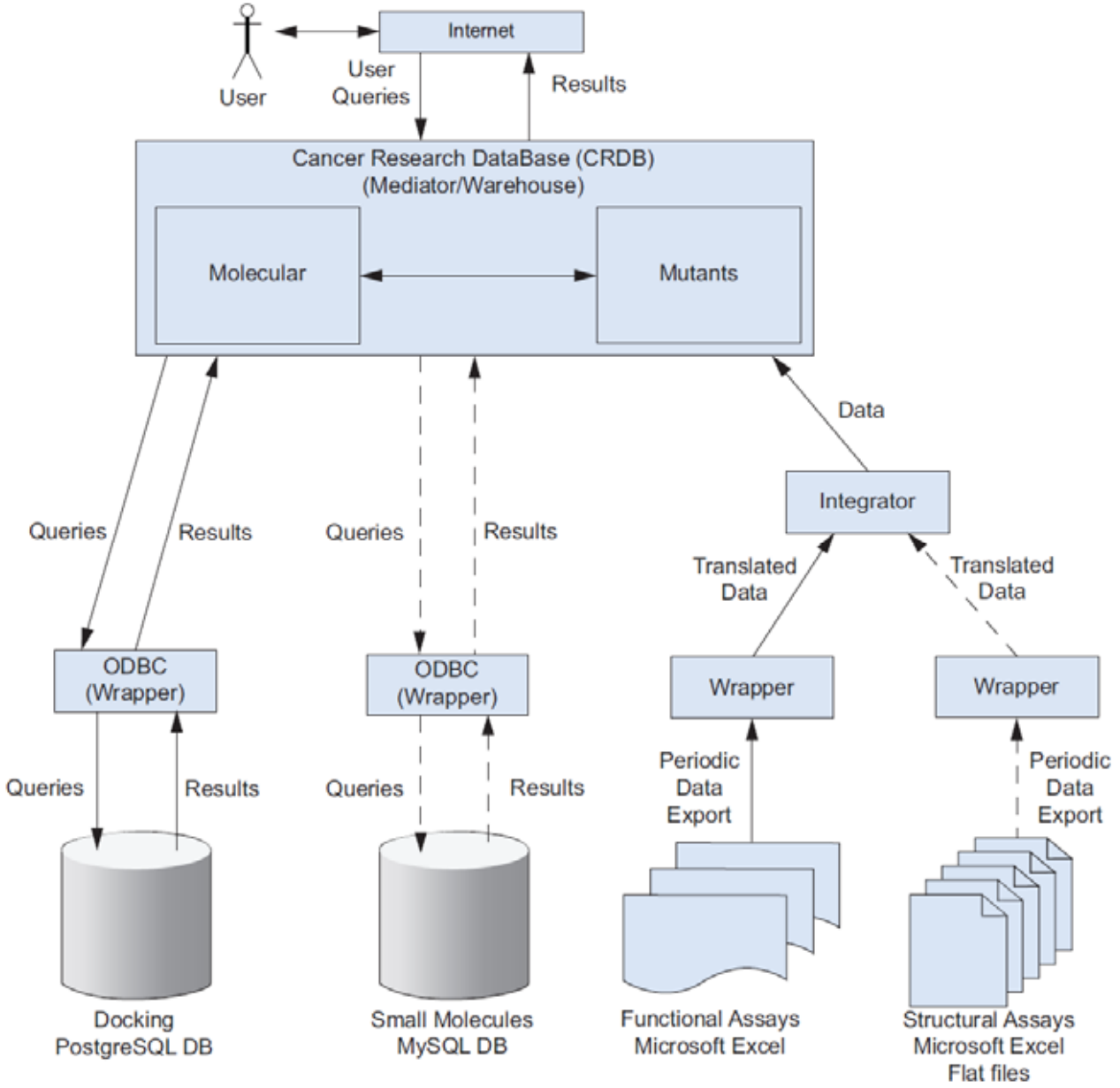
Holzinger, A., Burgsteiner, H. & Maresch, H. 2005. Experiences with the Practical Use of Care2x in Medical Informatics Education (Reverse Engineering). *In: Lazakidou, A. (ed.) Encyclopaedia of Informatics in Healthcare & Biomedicine. Hershey (PA): Idea Group Reference, pp. 81-88.*

Example: CRDB Global Data base schema



Bichutskiy, V. Y., Colman, R., Brachmann, R. K. & Lathrop, R. H. (2006) Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: Cancer Research Database. *Cancer Informatics, 2, 277-287.*

Example for a hybrid strategy for data integration

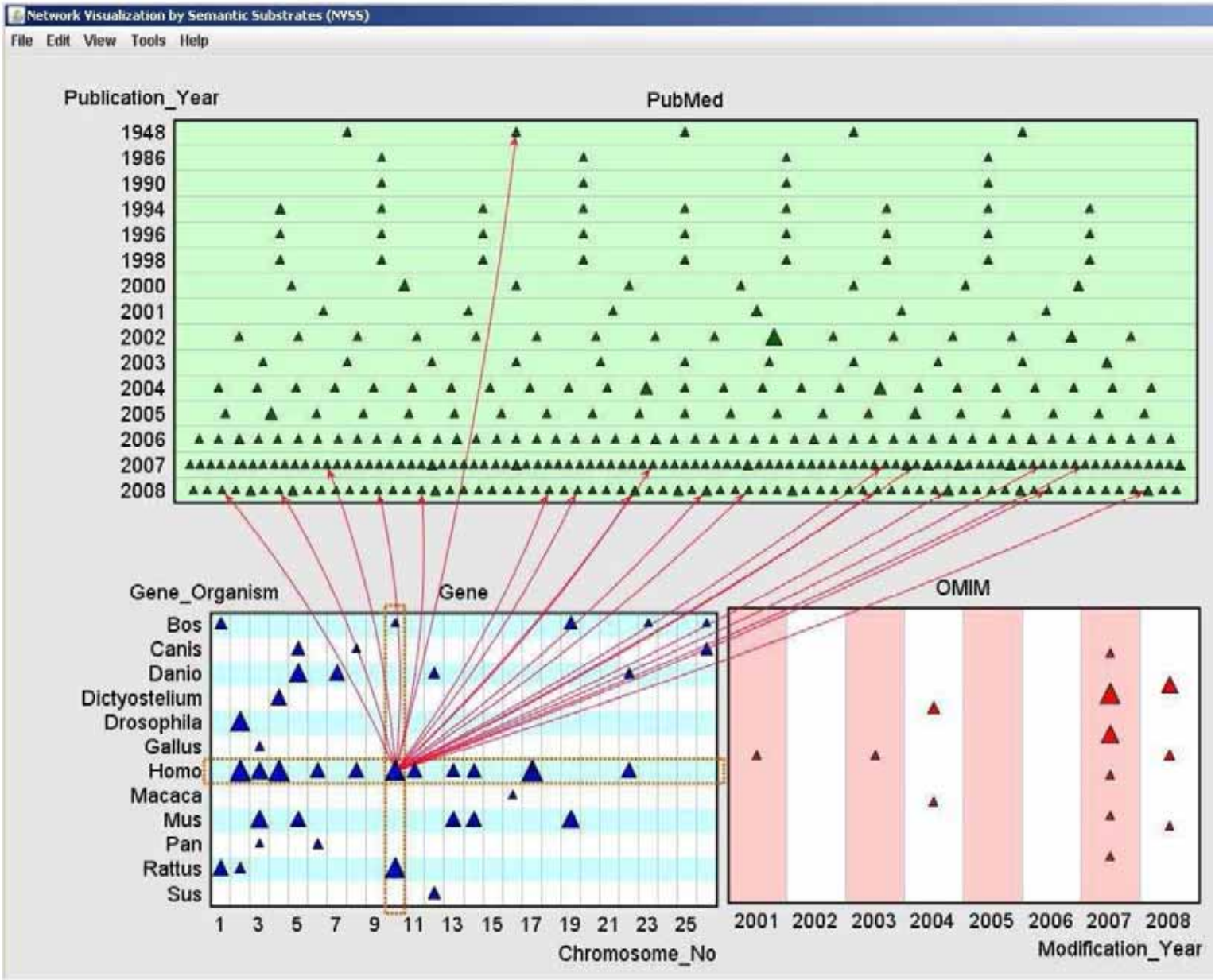


Bichutskiy, V. Y., Colman, R., Brachmann, R. K. & Lathrop, R. H. (2006) Heterogeneous Biomedical Database Integration Using a Hybrid Strategy: Cancer Research Database. *Cancer Informatics, 2, 277-287.*

Slide 4.22 Example Database: Protein Structure Data Bank

Example: Cervical Cancer query visualization

Lieberman, M. D., Taheri, S., Guo, H. M., Mirrashed, F., Yahav, I., Aris, A. & Shneiderman, B. (2011) Visual Exploration across Biomedical Databases. *IEEE-ACM Transactions on Computational Biology and Bioinformatics, 8, 2, 536-550.*



Example Patent of Scoring Documents in a linked database

US006799176B1

(12) **United States Patent**
Page

(10) **Patent No.: US 6,799,176 B1**
(45) **Date of Patent: *Sep. 28, 2004**

(54) **METHOD FOR SCORING DOCUMENTS IN A LINKED DATABASE**

(75) Inventor: **Lawrence Page**, Stanford, CA (US)

(73) Assignee: **The Board of Trustees of the Leland Stanford Junior University**, Palo Alto, CA (US)

(*) Notice: Subject to any disclaimer, the term of this patent is extended or adjusted under 35 U.S.C. 154(b) by 171 days.

This patent is subject to a terminal disclaimer.

(21) Appl. No.: **09/899,068**

(22) Filed: **Jul. 6, 2001**

Related U.S. Application Data

(63) Continuation of application No. 09/004,827, filed on Jan. 9, 1998, now Pat. No. 6,285,999.
(60) Provisional application No. 60/035,205, filed on Jan. 10, 1997.

(51) **Int. Cl.**[7] **G06F 17/30**
(52) **U.S. Cl.** **707/5**; 715/501.1
(58) **Field of Search** 707/5, 7, 10, 1–3; 715/501.1; 702/179, 181

6,014,678	A	1/2000	Inoue et al.	
6,112,202	A *	8/2000	Kleinberg	707/5
6,163,778	A *	12/2000	Fogg et al.	707/10
6,269,368	B1 *	7/2001	Diamond	707/6
6,285,999	B1	9/2001	Page	707/5
6,389,436	B1 *	5/2002	Chakrabarti et al.	707/513
2001/0002466	A1 *	5/2001	Krasle	704/270.1

OTHER PUBLICATIONS

Recker et al "Predicting document access in large multimedia repositories", ACM Transactions on Computer–Human Interaction, vol. 3, No. 4, Dec. 1996, pp. 352–375.*

Copy of claims of U.S. Serial No. 09/895,174, filed on July 2, 2001; Lawrence Page; Method for Node Ranking in a Linked Database; 8 pages.

Yuwono et al., "Search and Ranking Algorithms for Locating Resources on the World Wide Web", IEEE 1996, pp. 164–171.

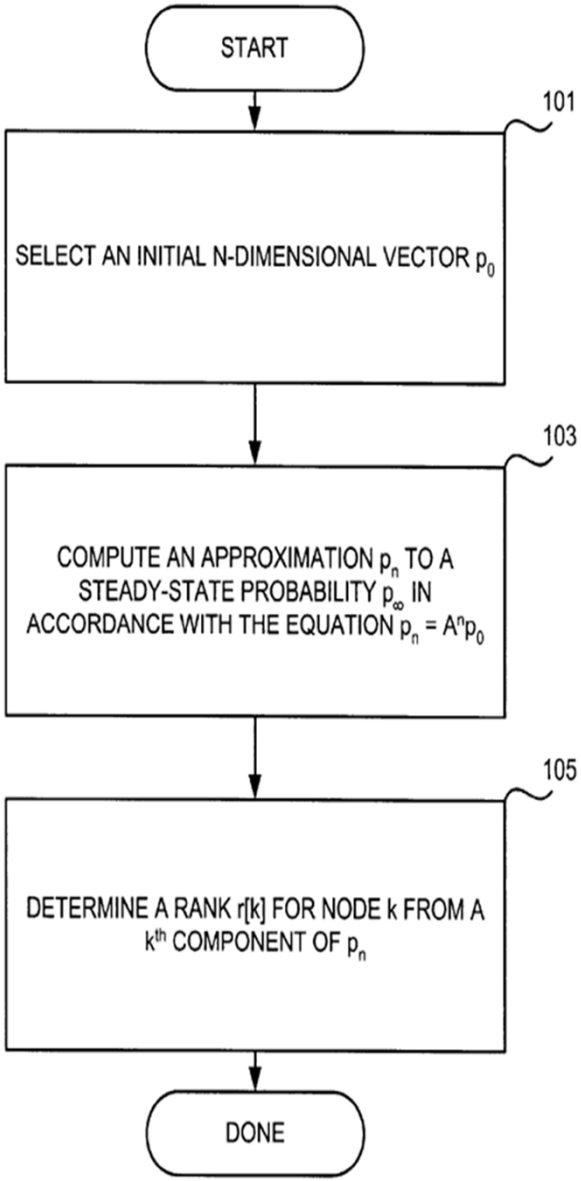
L. Katz, "A new status index derived from sociometric analysis", 1953, Psychometricka, vol. 18, pp. 39–43.

C.H. Hubbell, "An input–output approach to clique identification sociometry", 1965, pp. 377–399.

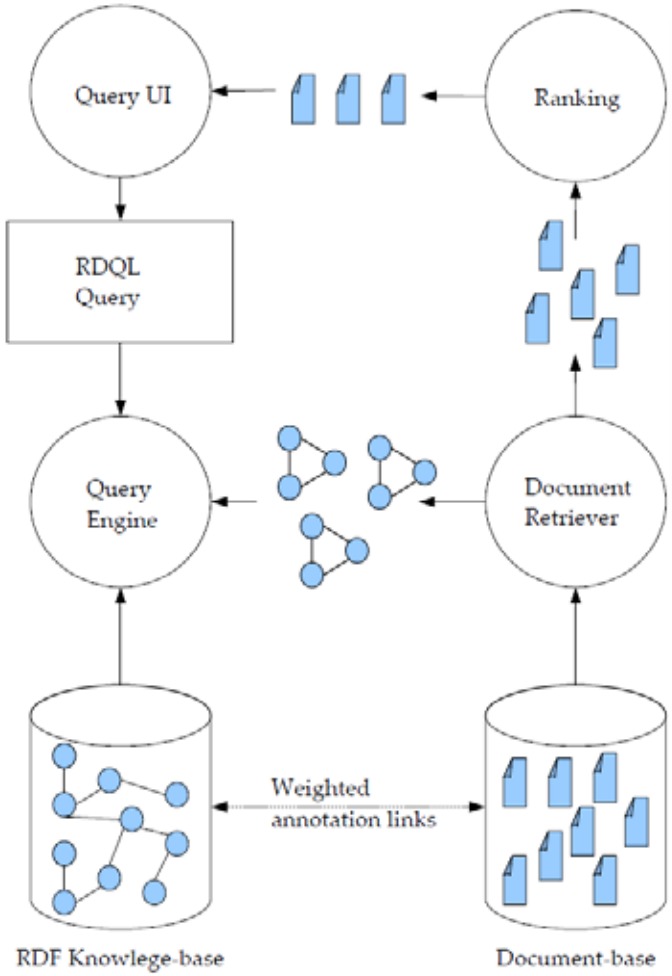
Mizruchi et al., "Techniques for disaggregation centrality scores in social networks", 1996, Sociological Methodology, pp. 26–48.

E. Garfield, "Citation analysis as a tool in journal evaluation", 1972, Science, vol. 178, pp. 471–479.

(List continued on next page.)



Backup: Compare with an Ontology-Based Model



Q — *user query*

$O = \{I_i\}_{i=1}^{M}$ — *set of all instances in the ontology*

$\{D_i\}_{i=1}^{N}$ — *set of all documents in the search space*

$(v_1, \ldots, v_k)$ — *the weights of the variables in the SELECT clause in the RDQL query Q*

$T = \{T_i\}_{i=1}^{n}$ — *list of tuples in the query result set, where* $T_i = \{T_{i,j}\}_{i=1}^{k}$ *with* $T_{i,j} \in O$

$\vec{d}_i = (d_{i,1}, \ldots, d_{i,M})$ — *document vector of* D_i*, where* $d_{i,j}$ *is the weight of the annotation of document* d_i *with* I_j *or zero otherwise*

$\vec{q} = (q_1, \ldots, q_M)$ — *extended query vector, where* $q_l = \sum_{\exists i | I_i = T_{i,j}} v_j$

$$sim(D_i, Q) = \frac{\vec{d}_i \bullet \vec{q}}{|\vec{d}_i| \times |\vec{q}|}$$

Cf with Vallet et al. (2005) and see also the work by

Spat, S. (2007) *Prototype of a Medical Information Retrieval System for Electronic Patient Records: Finding relevant information in clinical text documents (Diploma Thesis). TU Graz*

Ontology Based Model: Pros & Cons

Advantages	Disadvantages
Documents can be ranked by relevance	Works only if adequate knowledge base is available
Semantics of the documents can be considered	Only usable for already known facts – completely useless to discover new items
Model outperforms classic IR models	Big effort to build and maintain a adequate knowledge base

Some Useful Links

- http://www.library.tufts.edu/hsl/resources/dbases.html
- http://www.ncbi.nlm.nih.gov/omim
- http://lucene.apache.org/java/docs/
- http://www.dcs.gla.ac.uk/Keith/Preface.html
- http://hive.apache.org/
- http://www.cs.waikato.ac.nz/ml/weka/
- http://scikit-learn.sourceforge.net/stable/
- http://www.eecs.wsu.edu/mgd/gdb.html