


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>




A. Holzinger 709.049 1/92 Med Informatics L04


Status as of Di, 03.11.2015, 10:00

Dear Students, welcome to the 4th lecture of our course. Please remember from the last lecture: modeling of knowledge, medical Ontologies, Classification efforts and the International Classification of Diseases (ICD); Standardized Nomenclature of Medicine Clinical Terms (SNOMED CT); Medical Subject Headings (MeSH); Unified Medical Language System (UMLS);

Please always be aware of the definition of biomedical informatics (Medizinische Informatik):

Biomedical Informatics is the inter-disciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health (and well-being).

Schedule	
<ul style="list-style-type: none">▪ 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	
A. Holzinger 709.049	2/92 Med Informatics L04

Keywords of the 4 th Lecture		
<ul style="list-style-type: none">▪ 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)		
A. Holzinger 709.049	3/92	Med Informatics L04

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 Encyclopedia 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 and 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

Learning Goals ... at the end of the 4th lecture you ...



- ... 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**;


A. Holzinger 709.049

9/92

Med Informatics L04

At the end of this fourth lecture you ...

... 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 biomedical databases;
... are familiar with some basics of information retrieval.

Key Challenges	
<ul style="list-style-type: none"> ▪ Increasingly large and <u>complex</u> 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] 	
<p>[1] Shah, N. H. & Tenenbaum, J. D. 2012. The coming age of data-driven medicine: translational bioinformatics' next frontier. <i>Journal of the American Medical Informatics Association</i>, 19, (E1), E2-E4.</p> <p>[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..</p> <p>[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.</p>	
A. Holzinger 709.049	10/92
Med Informatics L04	

Amongst other problem some key challenges include:

Increasingly large and complex data sets “Big Data” due to data intensive research
 Increasing amounts of non-standardized and un-structured information (e.g. free text)

Data quality, data integration, universal access

Privacy, security, safety and data protection issues (see →Lecture 11)

Time aspects in databases (Gschwandtner, Gärtner, Aigner & Miksch, 2012),
 (Johnston & Weis, 2010).

“Big Data resources are all a waste of time and money if data analysts cannot find, or fail to comprehend, the basic information that describes the data held in the resources (Berman, 2013b)”

Data identification is certainly the most underappreciated and least understood Big Data issue. Measurements, annotations, properties, and classes of information have no informational meaning unless they are attached to an identifier that distinguishes one data object from all other data objects and that links together all of the information associated with the identified data object (Berman, 2013a).

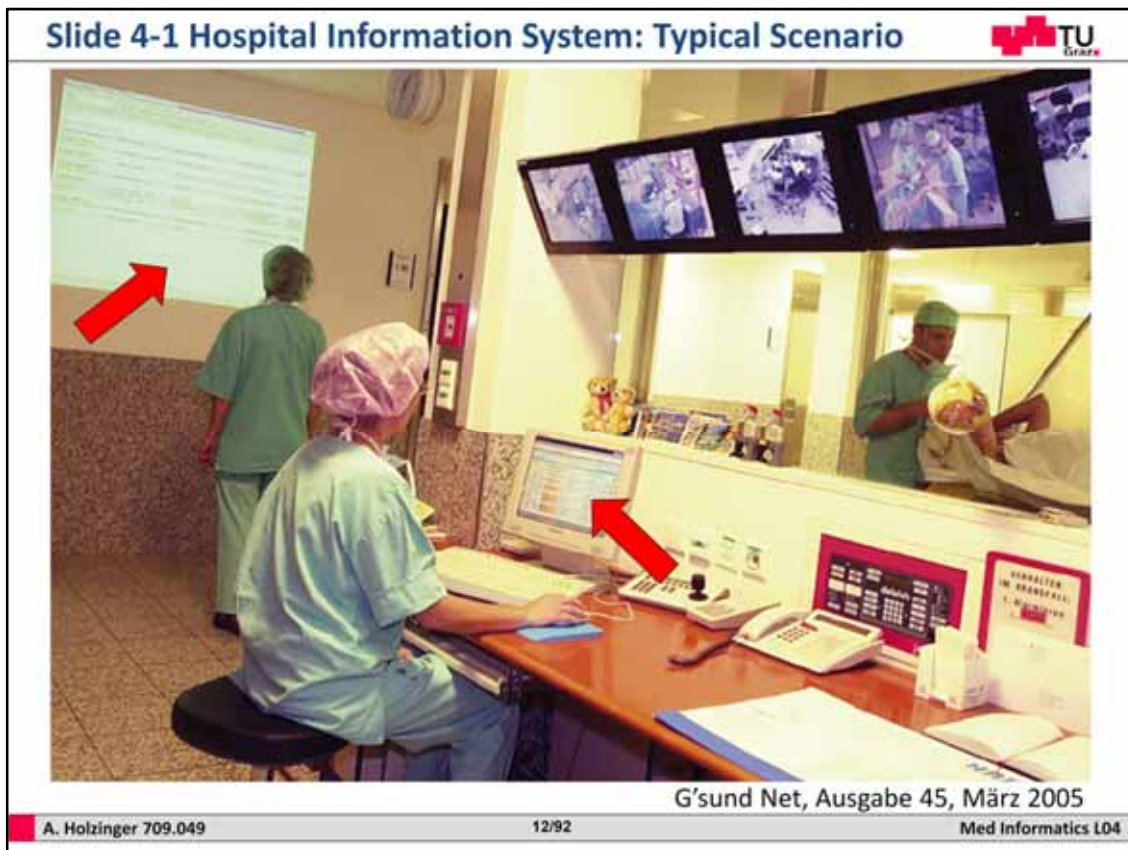
Communication of data between application systems must ensure security to avoid improper access, because trust or the lack thereof, is the most essential factor blocking the adoption of rapidly evolving Web technology paradigm such as software as service (SaaS) or data distribution services such as Cloud computing (Sreenivasaiah, 2010).



Let us start with a look into the Hospital ...


A. Holzinger 709.049 11/92 Med Informatics L04

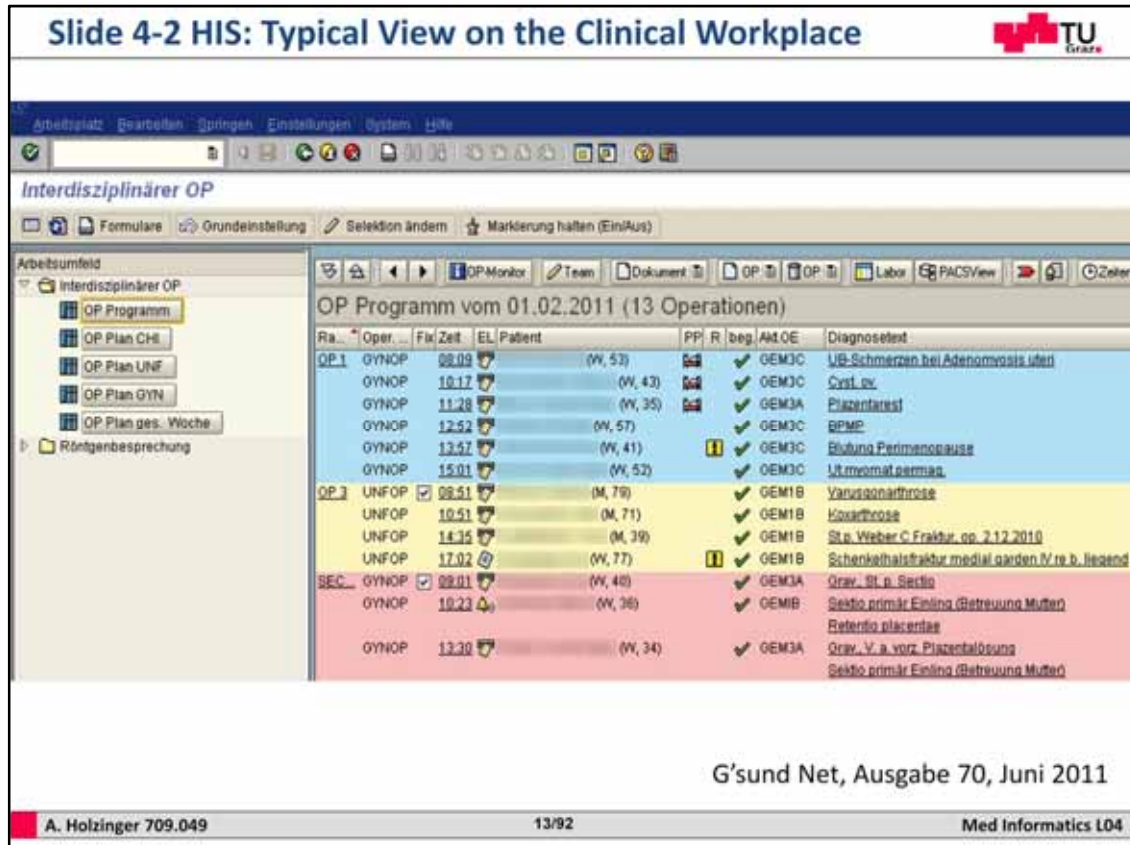
Before we discuss information systems and learn about data bases, let us start with a look into the Hospital ...



Let us start with a look into the hospital: In this slide we see a typical hospital scenario: medical professionals are surrounded by information technology. An old dream of hospital managers was always to have an “all digital hospital” to digitalize all workflows and to store all data in an electronic way – towards a paperless hospital.

Although much effort has been spent towards a paperless hospital, most hospitals worldwide are still far away from being a “all-digital hospital” (Waterson, Glenn & Eason, 2012). An interesting study: All hospitals in the province of Styria (Austria) are well equipped with sophisticated Information Technology, which provides all-encompassing on-screen patient information. Previous research made on the theoretical properties, advantages and disadvantages, of reading from paper vs. reading from a screen has resulted in the assumption that reading from a screen is slower, less accurate and more tiring. However, recent flat screen technology, especially on the basis of LCD, is of such high quality that obviously this assumption should now be challenged. As the electronic storage and presentation of information has many advantages in addition to a faster transfer and processing of the information, the usage of electronic screens in clinics should outperform the traditional hardcopy in both execution and preference ratings. In a study in the County hospital Styria, Austria, with 111 medical professionals, working in a real-life setting, they were each asked to read original and authentic diagnosis reports, a gynecological report and an internal medical document, on both screen and paper in a randomly assigned order. Reading comprehension was measured by the Chunked Reading Test, and speed and accuracy of reading performance was quantified. In order to get a full understanding of the clinicians' preferences, subjective ratings were also collected. Wilcoxon Signed Rank Tests showed no significant differences on reading performance between paper vs. screen. However, medical professionals showed a significant (90%) preference for reading from paper. Despite the high quality and the benefits of electronic media, paper still has some qualities which cannot provided electronically do date (Holzinger et al., 2011).
BTW: Graz University Hospital is the flagship hospital of the Styrian KAGES with 23 county hospitals and is amongst the largest hospitals in Europe.

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



Arbeitsplatz Bearbeiten Springen Einstellungen System Hilfe

Interdisziplinärer OP

Formulare Grundeinstellung Selektion ändern Markierung halten (Ein/Aus)

Arbeitsumfeld

- Interdisziplinärer OP
 - OP Programm
 - OP Plan CHS
 - OP Plan UNF
 - OP Plan GYN
 - OP Plan ges. Woche
- Röntgenbesprechung

OP Programm vom 01.02.2011 (13 Operationen)


Ra...	Oper...	Fix	Zeit	EL	Patient	PP	R	beg.	AM	OE	Diagnosestext
OP.1	GYNOP	08.09			(W, 53)			✓	GEM3C		UB-Schmerzen bei Adenomatosis uteri
	GYNOP	10.17			(W, 43)			✓	GEM3C		Cyst. ex.
	GYNOP	11.28			(W, 35)			✓	GEM3A		Plazentarest
	GYNOP	12.52			(W, 57)			✓	GEM3C		BPMP
	GYNOP	13.57			(W, 41)			✓	GEM3C		Blutung Perimenopause
	GYNOP	15.01			(W, 52)			✓	GEM3C		Uterinomatomat permas.
OP.2	UNFOP	09.51			(M, 79)			✓	GEM1B		Varusgonarthrose
	UNFOP	10.51			(M, 71)			✓	GEM1B		Koxarthrose
	UNFOP	14.15			(M, 39)			✓	GEM1B		Stg. Weber C Fraktur op. 2.12.2010
	UNFOP	17.02			(W, 77)			✓	GEM1B		Schenkelhalsfraktur medial garden 1/2 re b. liegend
SEC	GYNOP	09.01			(W, 40)			✓	GEM3A		Grav. III a. Sectio
	GYNOP	10.23			(W, 36)			✓	GEM1B		Sectio primär Einling (Betreuung Mutter)
											Retentio placentae
	GYNOP	13.30			(W, 34)			✓	GEM3A		Grav. V a. vorz. Plazentalösung
											Sectio primär Einling (Betreuung Mutter)

G'sund Net, Ausgabe 70, Juni 2011


A. Holzinger 709.049 13/92 Med Informatics L04

Mega issues related with hospital information systems include: data integration, data fusion, standardization issues, clinical process analysis, modeling, compliance issues, evidence based treatment and decision support, privacy, security, safety and data protection and knowledge discovery and data mining – all connected with the central topic of this lecture: databases.

BTW: The KAGES uses openMEDOCS based on ish.med which is based on SAP R3, an overview about different business hospital information systems vendors can be found here:

Slide 4-3: Much of hospital work is teamwork ... 

- ... 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.

A. Holzinger 709.049 14/92 Med Informatics L04

The teamwork in the hospital requires a lot of communication and information exchange. The vision of a business enterprise hospital information system is to cover all workflows, organizational processes and information flows electronically.

Note: The quality of the work of physicians is heavily influenced by the usability of their available equipment. In the slide you see a typical work meeting of medical professionals, where they discuss the patient cases jointly. It is important to study and understand the workflows of the end users and to involve them into the development of information systems as early as possible by a user-centered design process (Holzinger, 2003). Experiments showed that by studying the workflows the engineers get deep insights into how to develop an appropriate application for a specified target end user group (Holzinger, Geierhofer, Ackerl & Searle, 2005).

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

Radiologischer Befund

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

Kurzanamnese: St.p. SHT
Fragestellung: -
Untersuchung: Thorax eine Ebene liegend [redacted]

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

Das Cor in der Größennorm, keine akuten Stauungszeichen.
Fragliches Infiltrat parahilar 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 [redacted] am 09.05.2006 ***

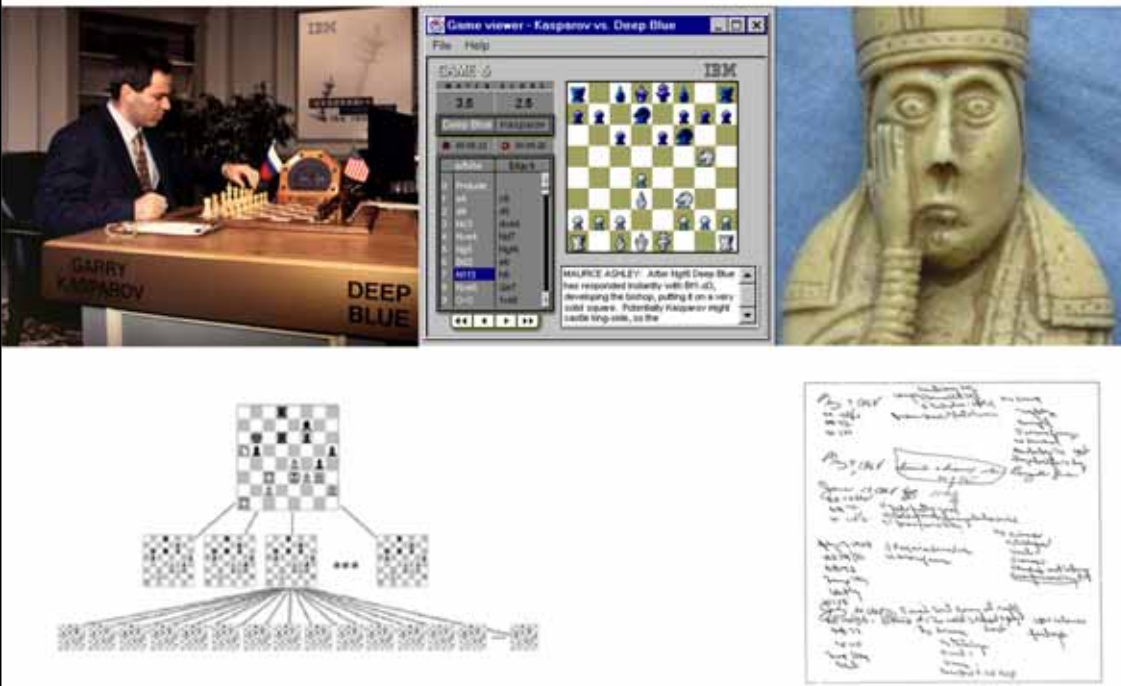
Special Words
Language Mix
Abbreviations
Errors ...

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

A. Holzinger 709.049
15/92
Med Informatics L04

The aforementioned goal of an “all-digital hospital” produces “big data” and remarkably much of the data is unstructured text. Interestingly, the main and most important output is the medical report (Arztbrief): In the example it is the report of a medical image – not the image itself is the relevant issue – it is the report (Holzinger, Geierhofer & Errath, 2007b). The handling with unstructured data is a mega challenge and brings along a lot of challenges for computers.

Slide 4-5 Excursus: Chess Game versus Natural Language 



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

A. Holzinger 709.049 16/92 Med Informatics L04

Let us briefly compare human intelligence with machine intelligence.

A good example on the complexity which we are facing in hospital information processing are the differences between chess and human natural language processing:

Whereas chess is a finite, mathematically well-defined search space, hence we have a well defined computational space, with limited numbers of moves and states and grounded in explicit, unambiguous mathematical rules, human language is exactly the opposite: Ambiguous, contextual and implicit; grounded in the human cognitive space, with a seemingly infinite number of ways to express one and the same meaning.

Note: IBM Deep Blue defeated the World Chess Champion Garry Kasparov in a six-game match in 1997. There were a number of factors that contributed to this success, including: a single-chip chess search engine, a massively parallel system with multiple levels of parallelism, a strong emphasis on search extensions, a complex evaluation function, and effective use of a Grandmaster game database. Technically, Deep Blue was a massively parallel system designed for carrying out chess game tree searches. The system was composed of a 30-node IBM RS/6000 SP computer and 480 single-chip chess search engines, with 16 chess chips per SP processor. The SP system consists of 28 nodes with 120 MHz P2SC processors, and 2 nodes with 135 MHz P2SC processors. The nodes communicated with each other via a high speed switch and all nodes had 1 GB of RAM, and 4 GB of disk. During the 1997 match with Kasparov, the system ran the AIX 4.2 operating system. The chess chips in Deep Blue were each capable of searching up to 2.5 million chess positions per second, and communicate with their host node via a fast micro channel bus (Campbell, Hoane & Hsu, 2002).

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,
Hepatitis weight index ...

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

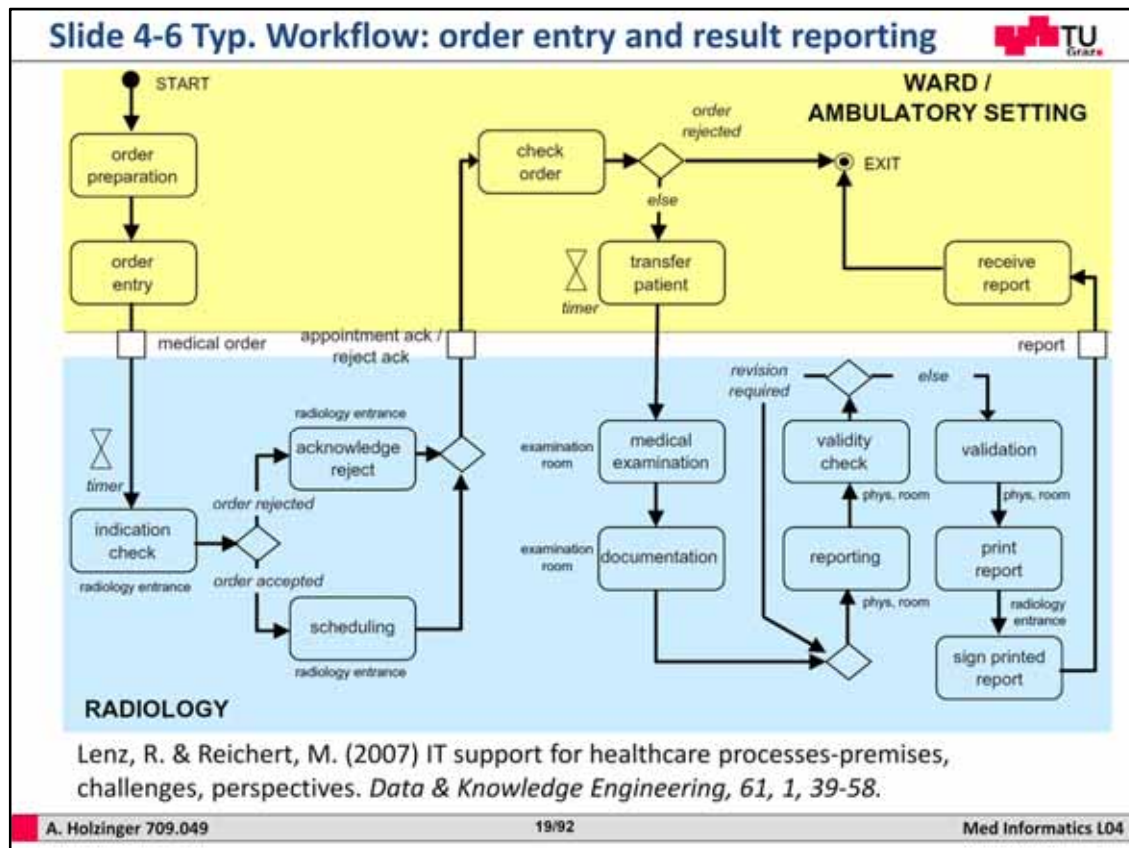
<http://www.medizinische-abkuerzungen.de/suche.html>



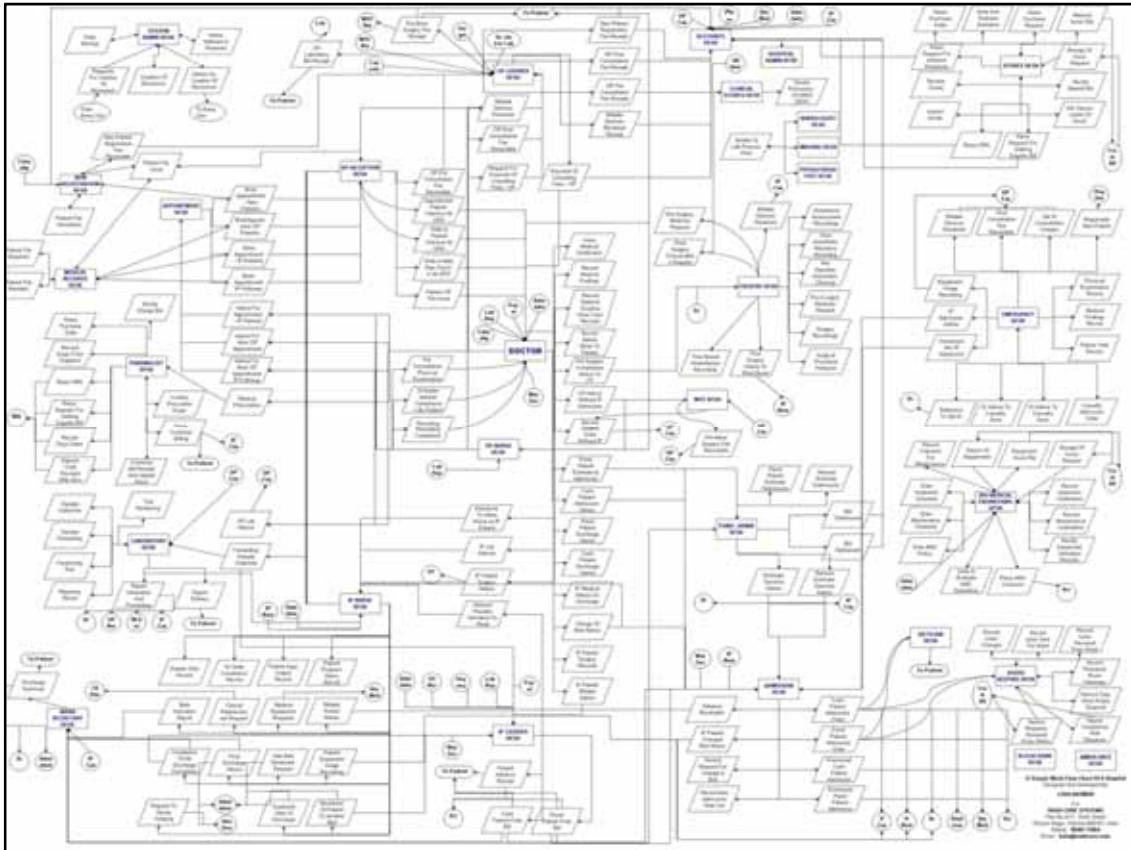
Hospital workflows are also complex ...

A. Holzinger 709.049 18/92 Med Informatics L04

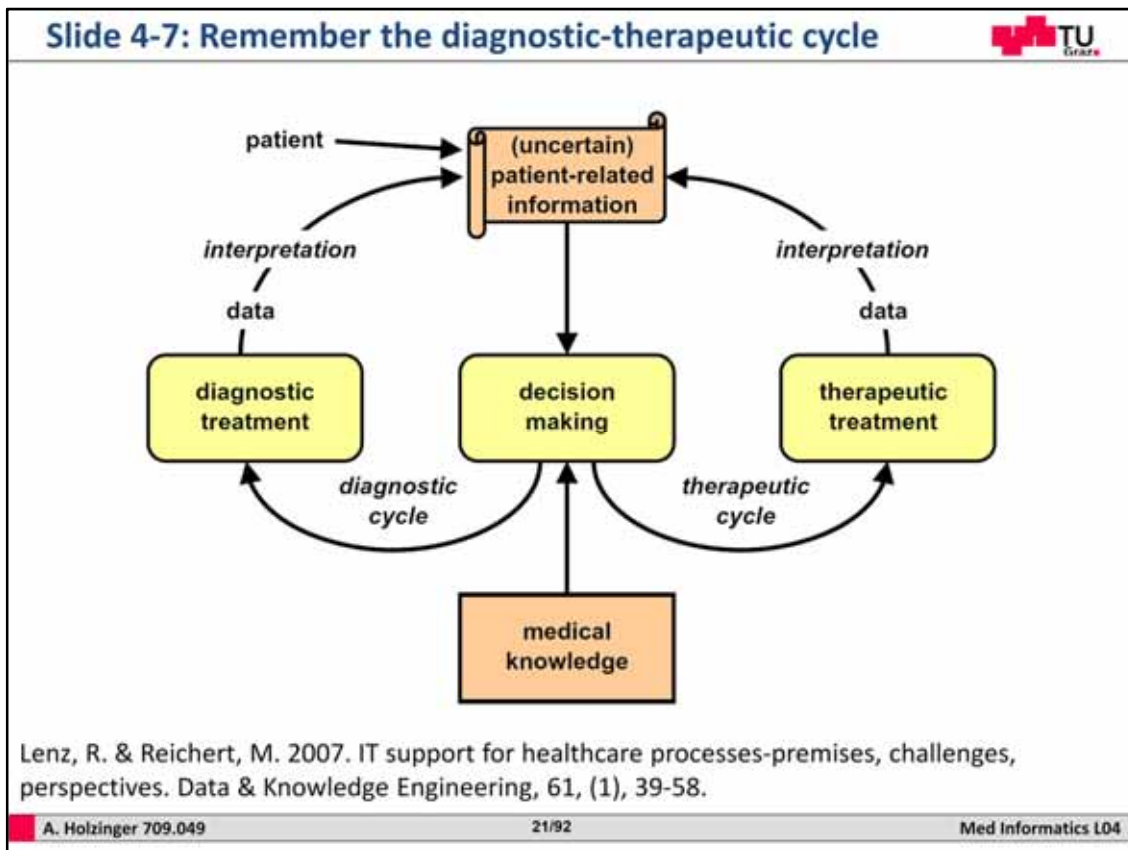
You can remember what we learned last lecture about workflows and workflow modelling ..



Healthcare processes require the cooperation of different organizational units and various medical disciplines. In such an environment optimal process support becomes crucial. In this slide we see a typical organizational process for medical order entry and result reporting, which is used to coordinate the inter-departmental communication between a ward (ambulatory setting) and the radiology unit. The depicted process is not tailored to a specific clinical pathway, but shows an example for a characteristic organizational procedure of the hospital: An order (in German: Anweisung, Verschreibung) is placed by a physician at the ward or at an ambulatory setting. The indication is checked in the radiology department and depending on the result the order placer is informed whether the request has been rejected or scheduled. The actual radiological examination and corresponding documentation is done in the examination room. The radiology report is generated afterwards, which has to be validated by the physician with his signature. The report is sent back to the order placer. This is an example for a fundamental process of clinical practice and captures the organizational knowledge necessary to coordinate the healthcare process among different people and organizational units; i.e., focus is on the support of core organizational processes (Lenz & Reichert, 2007).



This is just that you have an idea, how complicated such processes can be and you can imagine how difficult it is to digitalize all involved data

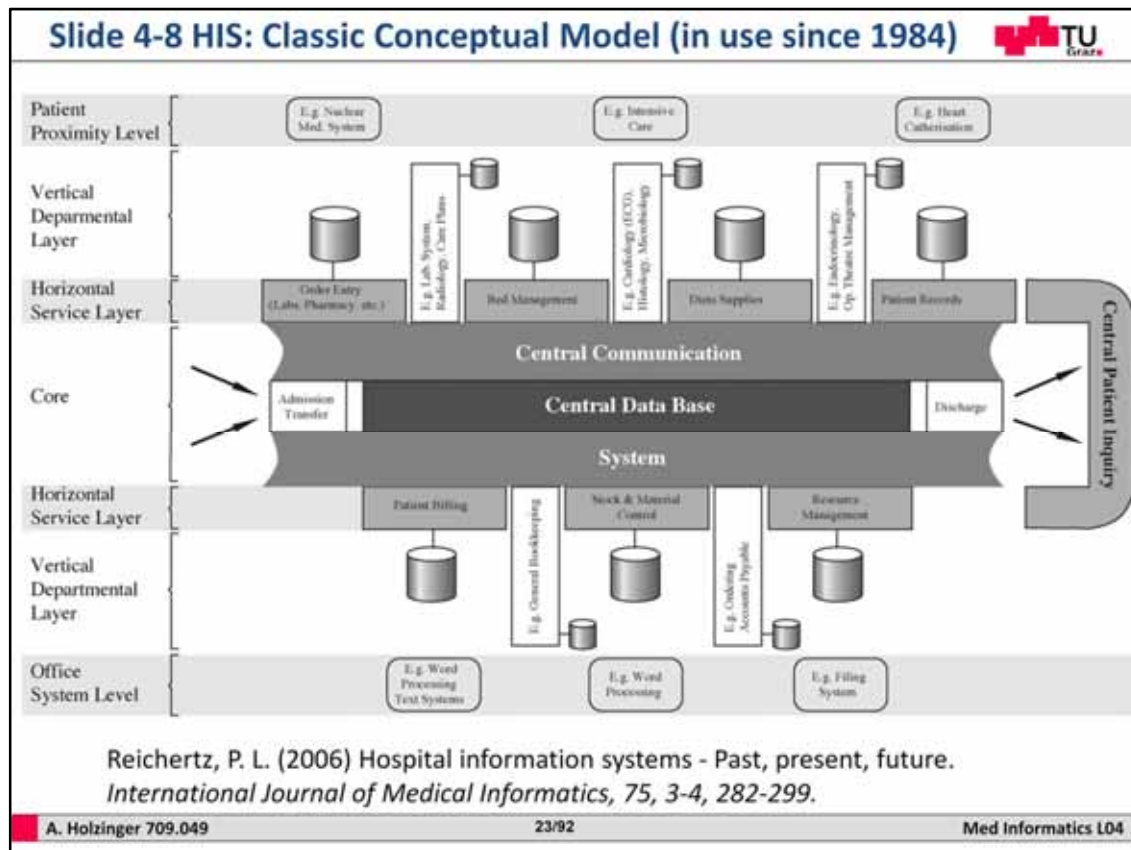


The medical treatment process is often described as diagnostic–therapeutic cycle (Bemmel & Musen, 1997) including: observation, reasoning, and action. Please remember that in medicine we deal with uncertain information (Holzinger & Simonic, 2011) and each pass of the diagnostic-therapeutic cycle can be seen as a step in decreasing the uncertainty about the patient’s disease. Consequently, the observation process always starts with the patient history (“looking into the past”) and proceeds with diagnostic procedures which are selected based on available information. The aim of the HIS is to assist healthcare personnel in making informed decisions. Maybe the most important question to be answered is how to determine what is relevant. Availability of relevant information is a precondition for (good) medical decisions – and the medical knowledge guides these decisions (Lenz & Reichert, 2007).

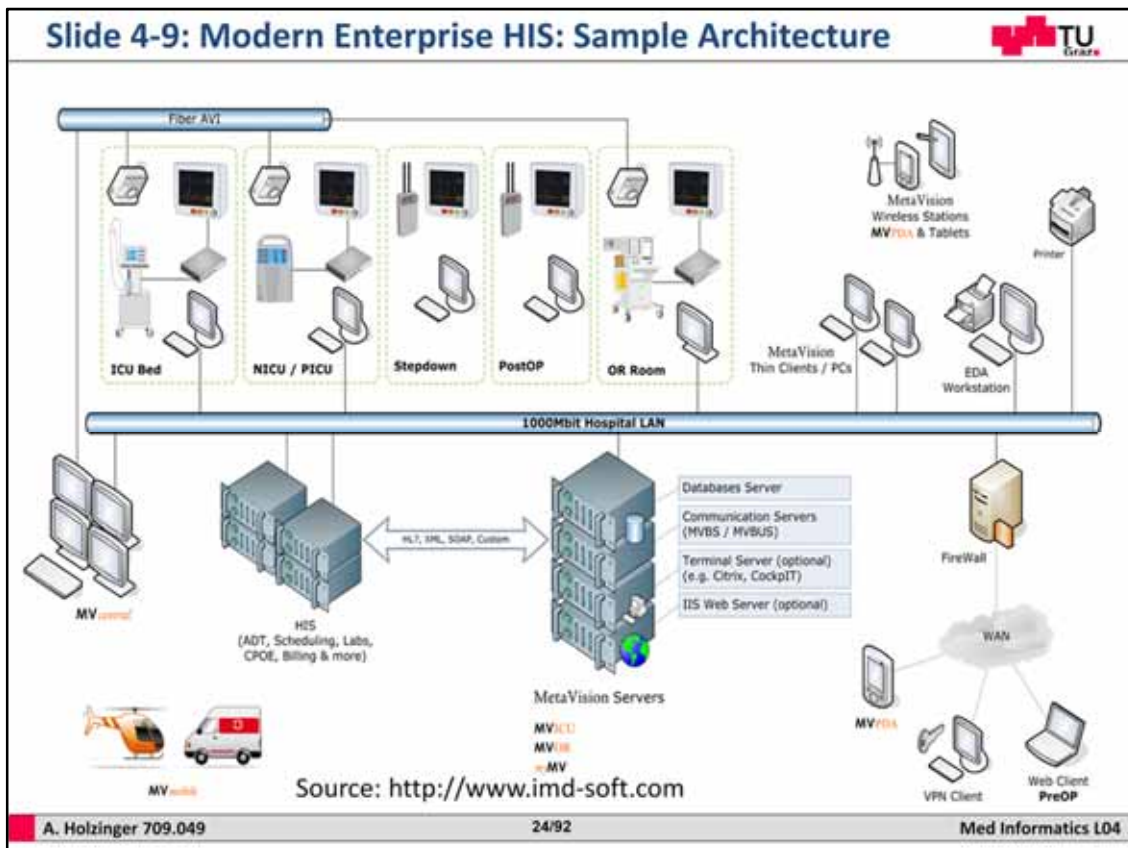
Following the principles of Evidence based medicine (EBM) physicians are required to formulate questions based on patients’ problems, search the literature for answers, evaluate the evidence for its validity and usefulness, and finally apply the (new) information to patients treatment (Hawkins, 2005). The limiting factor is the short time a clinician has to make a decision (Gigerenzer & Gaissmaier, 2011).



What is the architecture of an hospital information system?



This slide shows a classical conceptual model: The heart is a central data and communication structure. The patients “enter” (logically) the system through the admission on the left side, transfer and discharge functions of the core and leaves the system, at least partially, through the right side. In the main focus is a central data base, although alternative solutions have opted for a more distributed construction of data bases; nonetheless central ordering principles have to be kept to achieve the necessary integration of information and the distribution to the various points where it is needed, be it in the area of hospital management or in the field of care provision. This central data base is serving the central operational purposes of the hospital in the context of its dual goals (Haux et al., 1998), (Reichertz, 2006), (Haux, 2006).



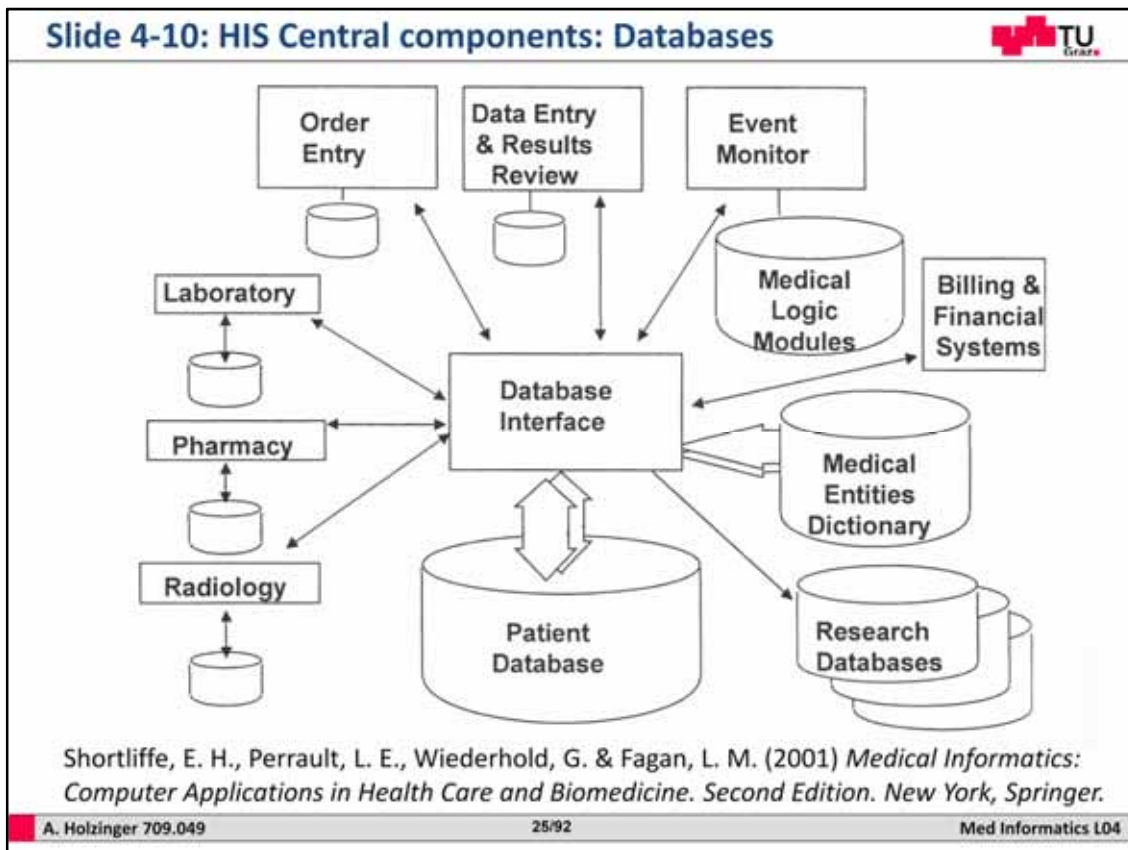
Here you see the typical architecture of such a system

ICU = Intensive Care Unit

NICU = Neonatal Intensive Care Unit

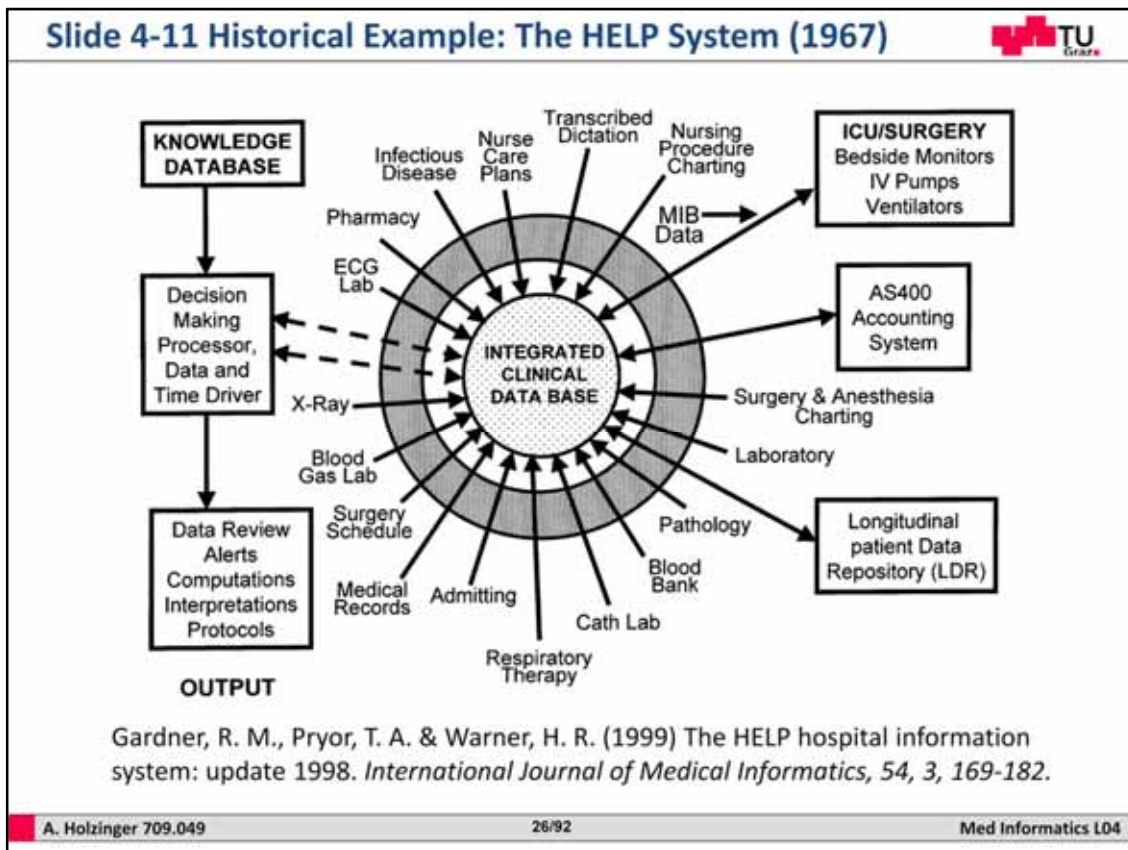
PICU = Pediatric Intensive Care Unit

There are many different application architectures in use, and we will come to it back later, in →Lecture 10, so here just ONE example for a enterprise business hospital information system as it is called professionally. However, we want now to concentrate on some technical issues of databases.



In a hospital there are data, data, data, ...

In this classical image by (Shortliffe, Perrault, Wiederhold & Fagan, 2001) it becomes very obvious that data bases are central components for an hospital information system. A very interesting slide is the next, where we see an historical example from the “stone-age” of computer science.




This picture by (Gardner, Pryor & Warner, 1999) is insofar interesting as it shows us clearly a mega issue up to the present: to integrate and fusion different data and to make it accessible to the clinician. While there is much research on the integration of heterogeneous information systems, a shortcoming is in the integration of available data. Just to clarify the differences between data integration and data fusion:

Data integration involves combining data residing in different distributed sources and providing users with a unified view of and access to these data. It has become the focus of extensive theoretical and practical work, and numerous open problems remain unsolved (Lenzerini, 2002).

Data fusion is the process of merging multiple records representing the same real-world object into a single, consistent, accurate, and useful representation (Bleiholder & Naumann, 2008).

The trend towards P4 medicine (Predictive, Preventive, Participatory, Personalized) has resulted in a sheer mass of the generated (-omics) data, hence a main challenge is in the integration and fusion of heterogeneous data sources, especially in the integration of data from the clinical domain with sources from the biological domain.

Key issues are ...



Data Integration

Data Fusion

Data Curation


A. Holzinger 709.049

27/92

Med Informatics L04

Integration – data fusion – for data analysis – the central goal to support decision making processes – data virtualization – abstract layer – business intelligence – service oriented architecture

Slide 4-12 Database – fundamental terms and definitions



- **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.

A. Holzinger 709.049

28/92

Med Informatics L04

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.

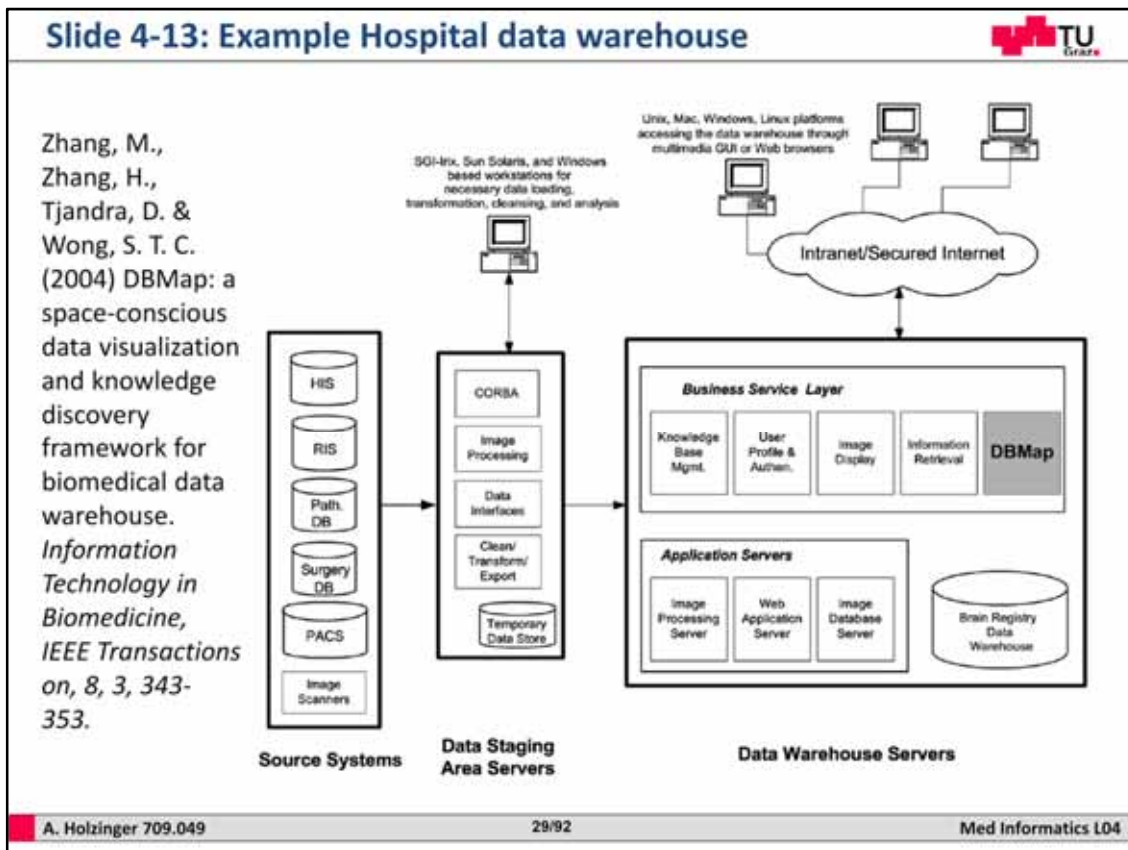
Data Marts (DM) are access layers of a DWH and are used as temporary repositories for data analysis.

Recommendable Reading include: (Plattner, 2013), (Robinson, Webber & Eifrem, 2013):

Robinson, I., Webber, J. & Eifrem, E. 2013. Graph Databases, O'Reilly Media.

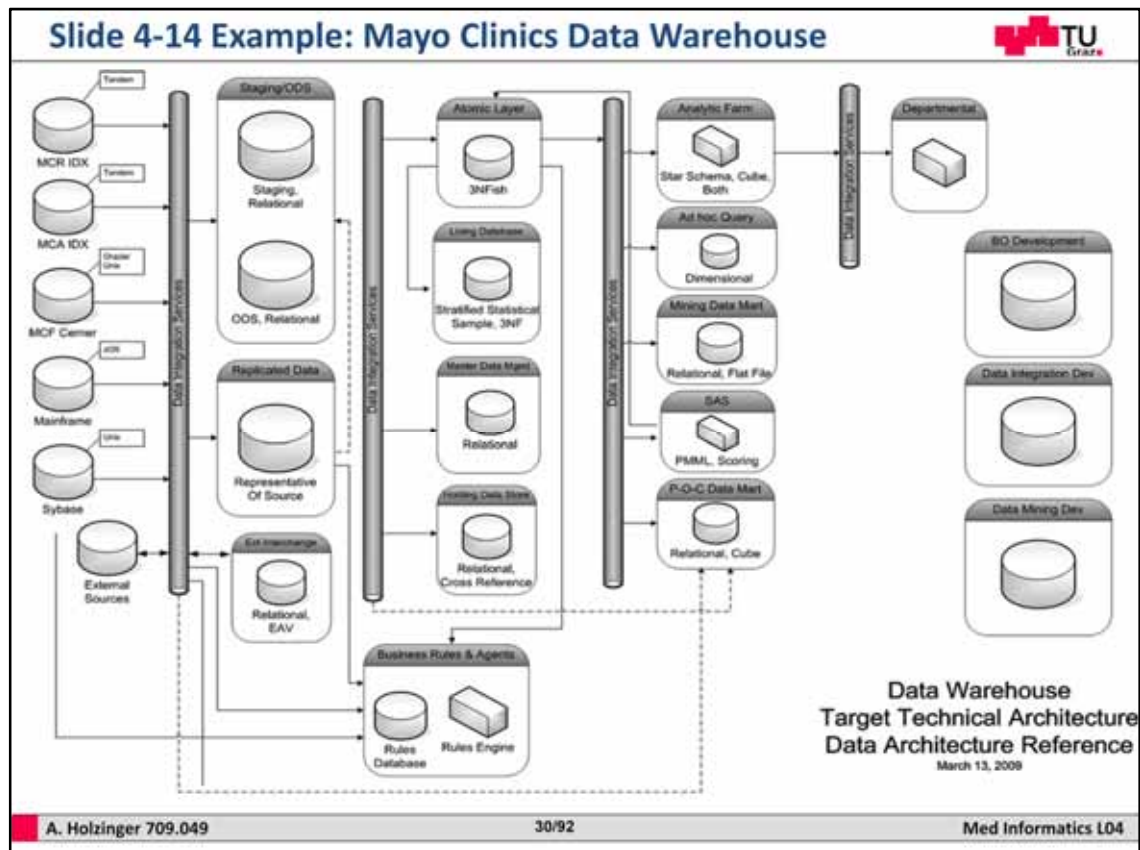
Plattner, H. 2013. A Course in In-Memory Data Management: The Inner Mechanics of In-Memory Databases, Heidelberg New York Dordrecht London, Springer.

One of the standard textbooks is the 6th edition of "Database System Concepts" by (Silberschatz, Korth & Sudarshan, 2010).




A DWH is an integrated system, specifically designed for enterprise business decision support and can be used in hospitals and in biomedical applications. In Slide 4-13 we see an example of a hospital data warehouse: On the left there are the (heterogeneous) data sources, such as PACS (Picture Archiving & Communication System) and RIS (Radiological Information System), and apart from the core HIS, some special data bases which can also include proprietary and legacy systems. For the data staging and area servers the Common Object Request Broker Architecture (CORBA) is used, a standard defined by the Object Management Group (OMG) that supports multiple platform interoperability (Zhang, Zhang, Tjandra & Wong, 2004).

This is a standard hospital information architecture and – typically - with no integration of laboratory data sources and most of all no Omics-data integration, as for example from the pathology or a bio-bank.



A DWH can be subdivided into so-called data marts (DM), which can be seen as specific access layer of a DWH, oriented to a specific team. Slide 4-14 shows the architecture of the Mayo clinic DWH, which is incrementally instantiating each component of the architecture on demand. Data integration proceeds from left to right (leftmost you see the primary data sources; moving right, the data are integrated into staging and replication services, with further refinement). The layers are:

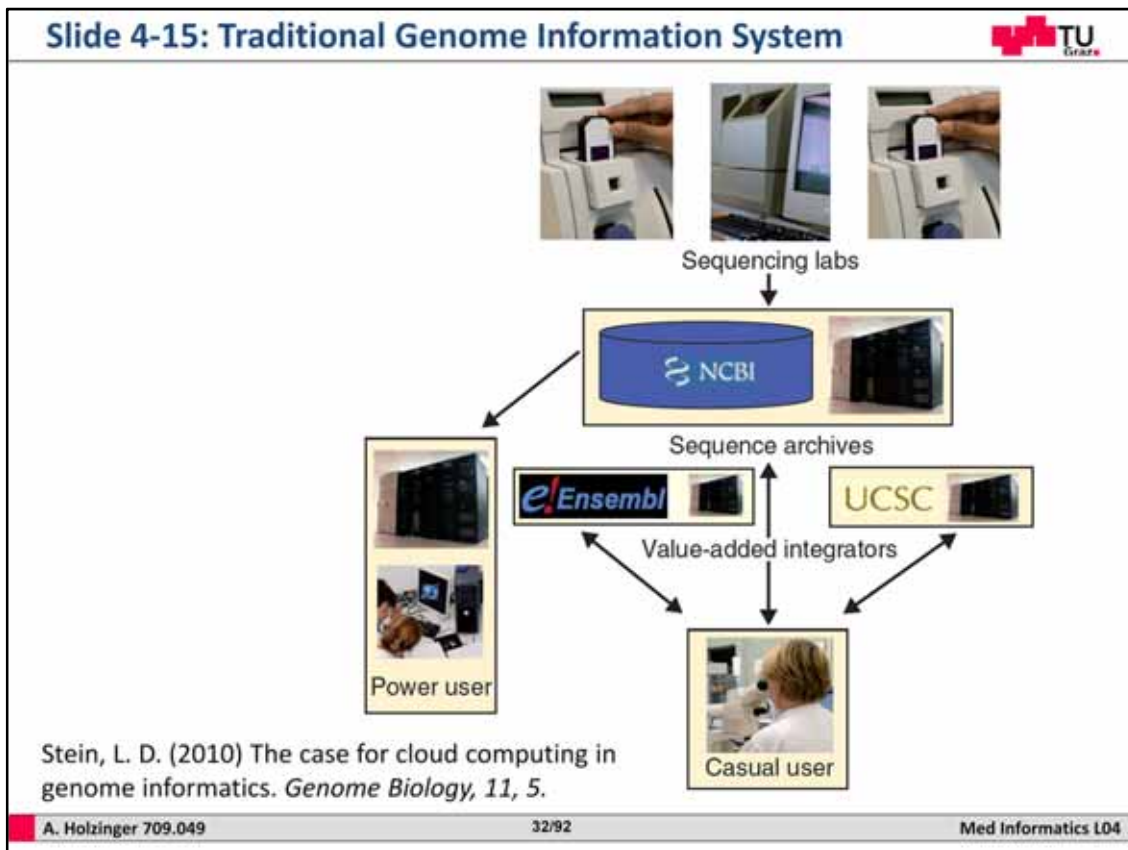
- 1) Subjects = the highest level areas that define the activities of the enterprise (e.g. Individual);
 - 2) Concepts = the collections of data that are contained in one or more subject areas (e.g., Patient, Provider, Referrer, etc.);
 - 3) Business Information Models = the organization of the data that support the processes and workflows of the enterprise's defined Concepts.
- (Chute, Beck, Fisk & Mohr, 2010)



What about cloud-based Information Systems?

A. Holzinger 709.049 31/92 Med Informatics L04

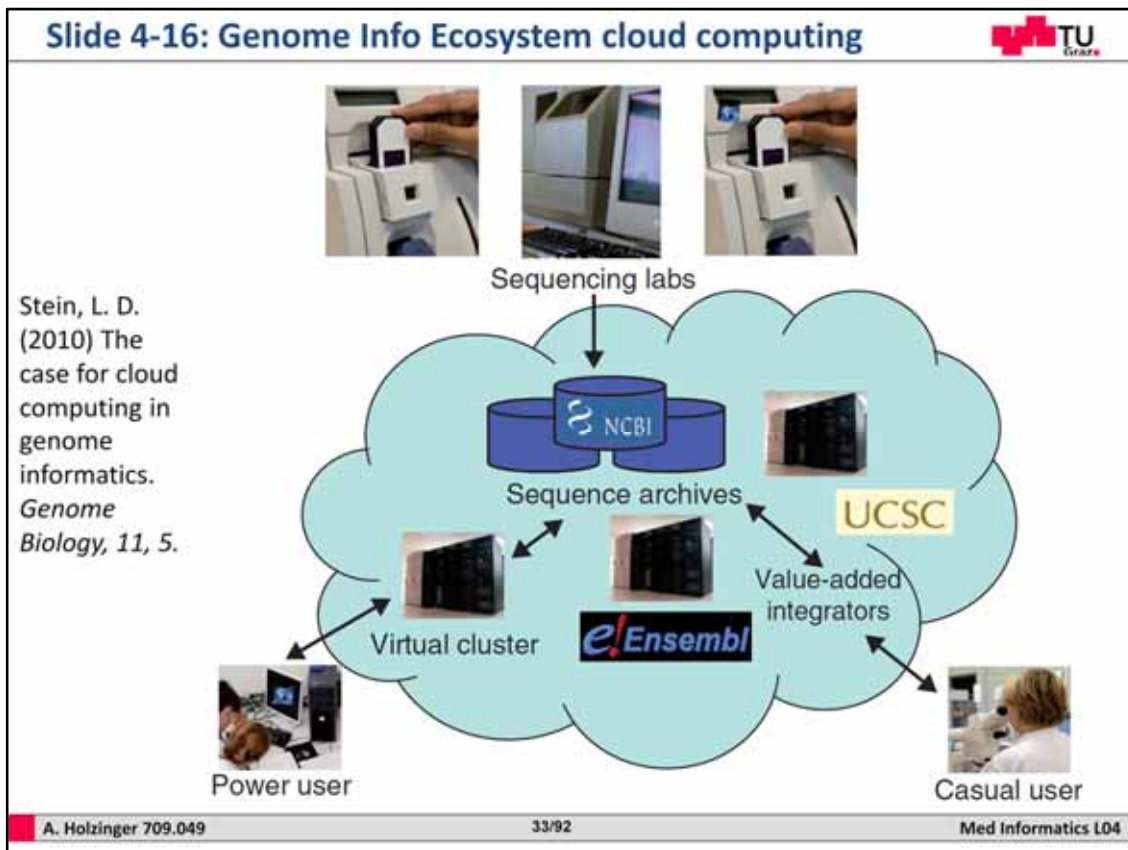
Cloud computing is a good example for Software as a service – flexible space via network – this reminds us to the early days of computing with mainframe computing and thin-client terminals.



A standard environment for production and processing of genomic data can be seen in Slide 4-15: Sequencing labs submit their data to large databases, e.g. GenBank, National Center of Biotechnology Information (NCBI); European Bioinformatics Institute (EMBL) database; DNA Data Bank of Japan (DDBJ); Short Read Archive (SRA); Gene Expression Omnibus (GEO) or Microarray database Array Express. These maintain, organize and distribute the sequencing data. Most users access the information either through web-based applications or through integrators, such as Ensembl, the University of California at Santa Cruz (UCSC) Genome Browser or Galaxy. The end users have to download genomic data from these primary and secondary sources (Stein, 2010).

Remember: Sequencing is the process of determining the precise order of nucleotides within a DNA molecule to determine the order of the four bases—adenine, guanine, cytosine, and thymine—in a strand of DNA. The advent of rapid DNA sequencing methods has greatly accelerated biological and medical research and discovery and produces large data sets. Sequencing has become indispensable for basic biological research, and in numerous applied fields such as diagnostics and biotechnology.

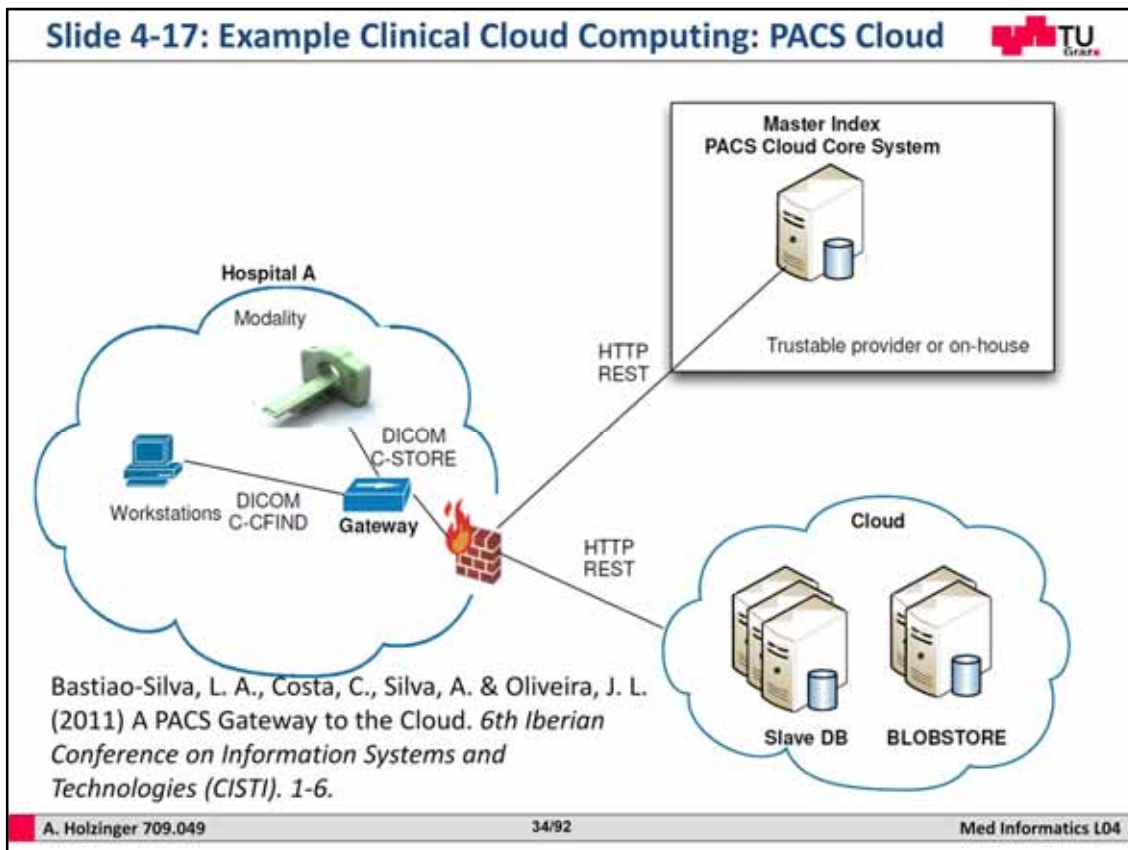
Note: A biobank is a physical place which stores biological specimens – and in some cases also data (Roden et al., 2008).



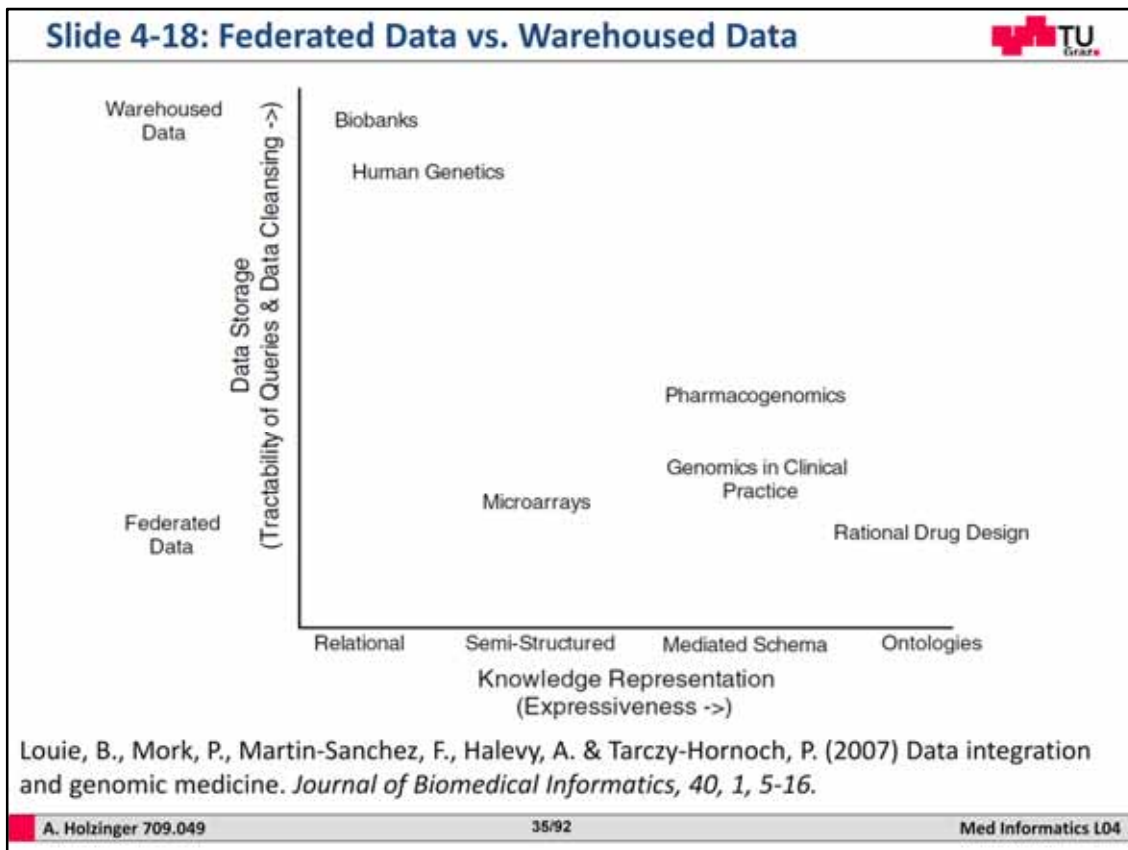
Here we see a cloud-based genome informatics system. Instead of separate genome datasets stored at various locations, the data sets are stored in the cloud as virtual databases. Web services run on top of these data sets, including the primary archives and the integrators, running as virtual machines within the cloud. Casual users, who are accustomed to accessing the data via the NCBI, DDBJ, Ensembl or UCSC, work as usual; the fact that these servers are located inside the cloud is invisible to them. Power users can continue to download the data, but have an attractive alternative. Instead of moving the data to the computational cluster, they move the computational cluster to the data (Stein, 2010).

Note: Cloud computing is based on sharing of resources to achieve coherence and economies of scale over a network (similar to the electricity grid). At the foundation of cloud computing is the broader concept of converged infrastructure and shared services. Cloud providers offer their services according to several fundamental models


- 1) Infrastructure as a service (IaaS),
- 2) Platform as a service (PaaS),
- 3) Software as a service (SaaS)



Just an example for a cloud based service: The Master Index is the PACS Cloud core entity and contains information about other modules, including Gateways and Cloud Slaves (repository and database). It also provides authentication services to institutional gateways and all identifiable information related with patients are stored in a master index database, fundamental to ensure solutions for confidentiality and privacy. The Cloud Slaves provide, on one hand, storage of sightless data (objects repositories) and, on other hand, a database containing all no identifiable metadata extracted from DICOM studies, i.e. the most demanding task concerning computational power (Bastiao-Silva, Costa, Silva & Oliveira, 2011).



We have to determine between federated data and warehoused data. A federated database system is a meta-database management system, which transparently maps multiple heterogeneous and autonomous database systems into a single federated database and this can be a “virtual database” – without data integration as it is in data warehouses. In the slide we can see on the y-axis the data integration architecture and on the x-axis the knowledge representation methodologies and where current data integration systems lie along this continuum. The essence of this image is that there is no “best-solution”: A system designed to have full control of data and fast queries can have difficulty expressing complex biological concepts and integrating them. Systems that employ highly expressive knowledge representation methodologies such as Ontologies are more able to represent and integrate complex biological concepts but have much less tractable queries (Louie et al., 2007).



What is the difference between hospital databases and Biomedical databases?

A. Holzinger 709.049 36/92 Med Informatics L04

Obviously there is a difference between the databases for the Hospital Information System and the databases which are used for scientific work.

Slide 4-19: 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);

Whereas databases for the use in HIS are process centered and central for the electronic patient record, biomedical databases are libraries of all sorts of life science data, collected from scientific experiments and computational analyses. Such databases contain experimental biological data from clinical work, genomics, proteomics, metabolomics, microarray gene expression, phylogenetics, pharmacogenomics, 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);

An overview can be found here: (Masic & Milinovic, 2012),

Online open access via: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3544328>

Note: Pharmacogenomics is the technology for the analytics of how genetic makeup affects an individual's response to drugs – so it deals with the influence of genetic variation on drug response in patients by correlating gene expression or single-nucleotide polymorphisms with efficacy and toxicity. The central aim is to optimize drug therapy to ensure maximum effectiveness with minimal adverse effects and is a core towards personalized medicine.

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

A. Holzinger 709.049 38/92 Med Informatics L04

A good video can be seen here: https://www.youtube.com/watch?v=DSHhep_w6pk

The Protein Data Bank archive-information about the 3D shapes of proteins, nucleic acids, and complex assemblies helps students and researchers understand all aspects of biomedicine, from protein synthesis to health and disease.

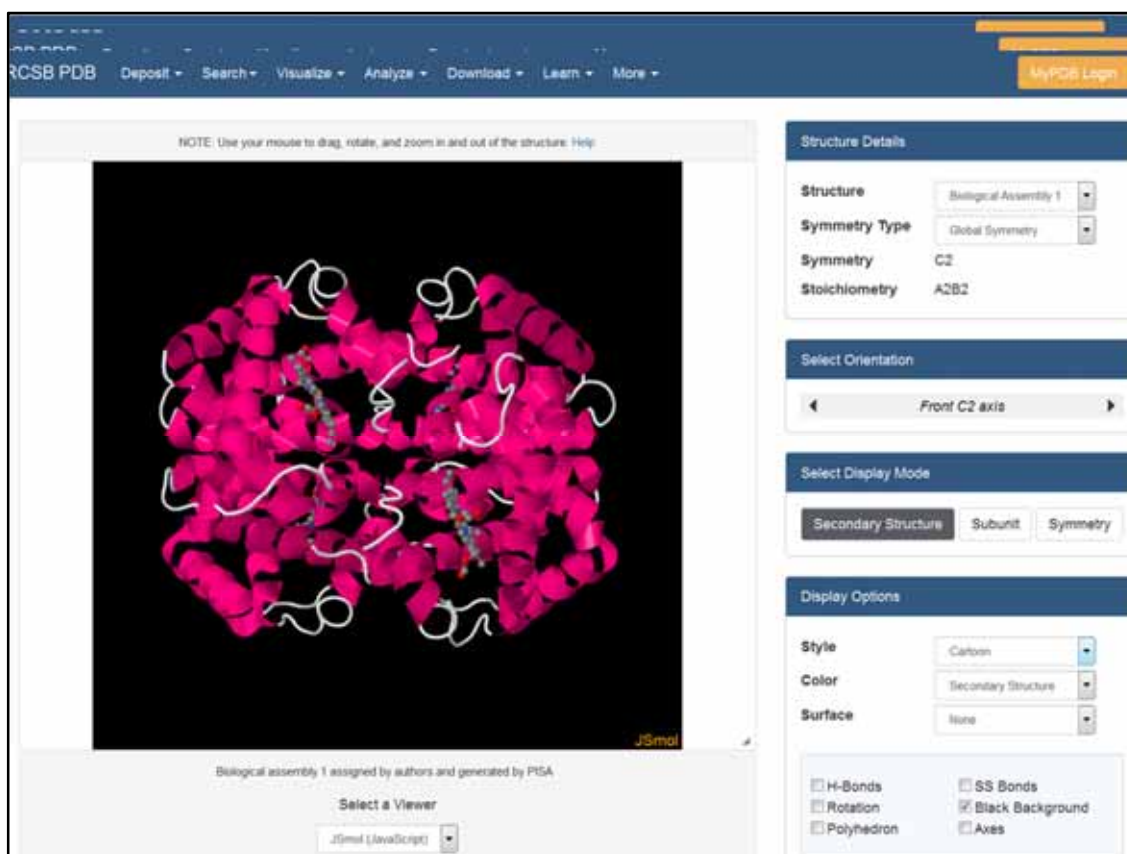
As a member of the wwPDB, the RCSB PDB curates and annotates PDB data.

The RCSB PDB builds upon the data by creating tools and resources for research and education in molecular biology, structural biology, computational biology, and beyond.

Remember: Proteins are the molecules used by the cell for performing and controlling cellular processes, including: degradation and biosynthesis of molecules, physiological signaling, energy storage and conversion, formation of cellular structures etc.

Protein structures are determined with crystallographic x-ray methods or by nuclear magnetic resonance spectroscopy. Once the atomic coordinates of the protein structure have been determined, a table of these coordinates is deposited into the protein database (PDB), an international repository for 3D structure files: <http://www.rcsb.org/pdb/>

This database is handled by the RCSB (Research Collaboratory for Structural Biology) at the Rutgers University and UC San Diego. PDB is the most important source for protein structures. Before a new structure of a protein is added, a careful examination of the data must be carried out to guarantee the quality of the structure. The PDB data file contains, among others, the coordinates of all the atoms of the protein (Wiltgen & Holzinger, 2005), (Wiltgen, Holzinger & Tilz, 2007).



A PDB structure entry should be cited with its PDB ID and primary reference. For example:

PDB ID: 1O2L

D.W. Heinz, W.A. Baase, F.W. Dahlquist, B.W. Matthews (1993) How Amino-Acid Insertions are Allowed in an Alpha-Helix of T4 Lysozyme *Nature* 361:561.

An entry without a published reference can be cited with the PDB ID, author names, and title:

PDB ID: 1C10

W. Shi, D.A. Ostrov, S.E. Gerchman, V. Graziano, H. Kycia, B. Studier, S.C. Almo, S.K. Burley, New York Structural GenomiX Research Consortium (NYSGXRC). The Structure of PNP Oxidase from *S. cerevisiae*

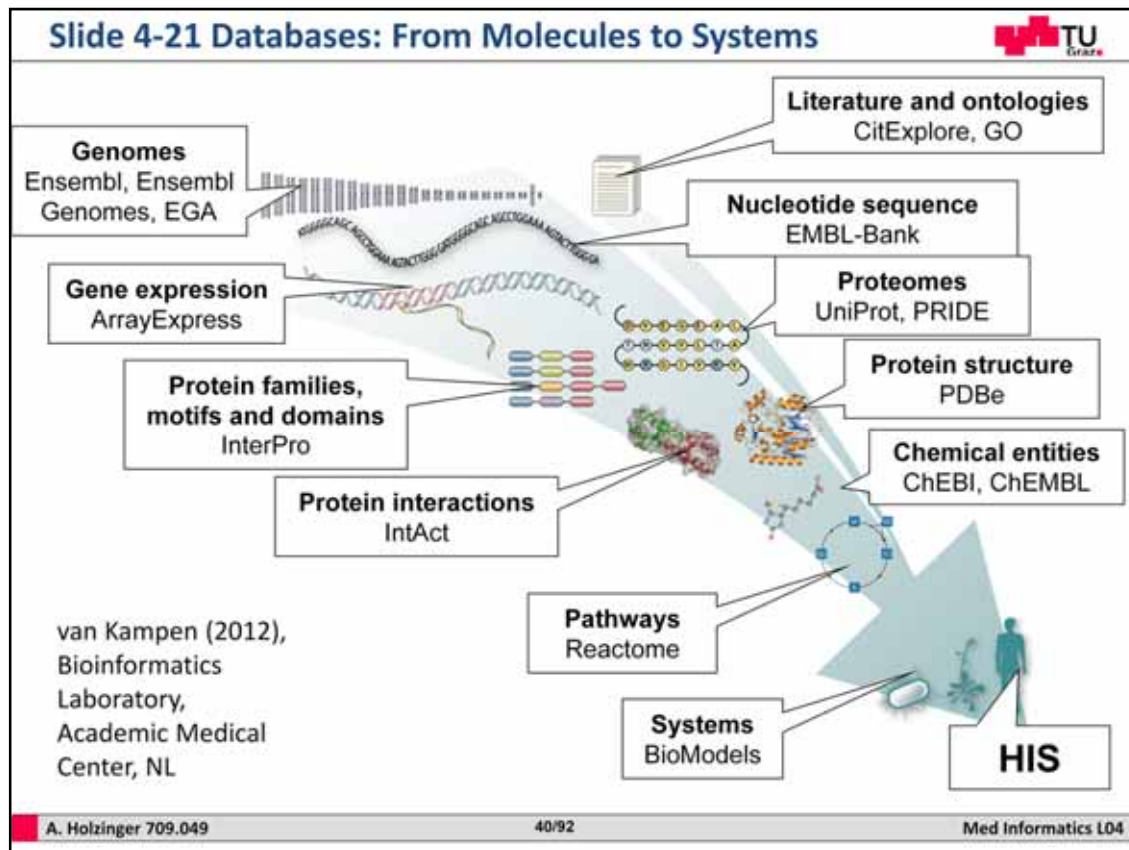
An entry may also be referenced using its Digital Object Identifier (DOI). The DOIs for PDB entries all have the same format: 10.2210/pdbXXXX/pdb, where XXXX should be replaced with the desired PDB ID. The DOI can be used as part of a URL to obtain this data file (<http://dx.doi.org/10.2210/pdb4hhb/pdb>), or can be entered in a DOI resolver (such as <http://www.crossref.org/>) to automatically link to [pdb4hhb.ent.gz](http://ftp.wwpdb.org/pdb4hhb.ent.gz) on the main PDB ftp archive (<ftp://ftp.wwpdb.org>). For example, the DOI for PDB entry 4HHB is "10.2210/pdb4hhb/pdb". This links directly to the entry in the PDB file format on the FTP server.

Images from Structure Summary pages should cite the RCSB PDB and the PDB entry:

Image from the RCSB PDB (www.rcsb.org) of PDB ID 1BNA (H.R. Drew, R.M. Wing, T. Takano, C. Broka, S. Tanaka, K. Itakura, R.E. Dickerson (1981) Structure of a B-DNA dodecamer: conformation and dynamics *Proc.Natl.Acad.Sci.USA* 78: 2179-2183).

Images created using PDB data and other software should cite the PDB ID and the molecular graphics program used.

Image of 1AOI (K. Luger, A.W. Mader, R.K. Richmond, D.F. Sargent, T.J. Richmond (1997) structure of the core particle at 2.8 Å resolution *Nature* 389: 251-260) created with Protein Workshop (J.L. Moreland, A. Gramada, O.V. Buzko, Q. Zhang, P.E. Bourne (2005) The Molecular Biology Toolkit (MBT): a modular platform for developing molecular visualization applications. *BMC Bioinformatics* 6:21).



Remember the structural dimensions which we discussed in Lecture 1 and Lecture 2. This Slide by (Kampen, 2013) is a very nice overview of various databases addressing the different microscopic dimensions. Additionally, the data on the level of the hospital information systems are added – so that you have a good summary of the aforementioned. If we take aside Literature databases and ontologies (in the upper right corner of this Slide) we start with:

Genome databases: Ensembl <http://www.ensembl.org/index.html>

Nucleotide sequence EMBL-Bank <http://www.ebi.ac.uk/ena/>

Gene expression: ArrayExpress <http://www.ebi.ac.uk/arrayexpress>

Proteomes: UniProt <http://www.uniprot.org/>

Proteins: InterPro <http://www.ebi.ac.uk/interpro/>

Protein structure: PDB <http://www.rcsb.org/pdb/home/home.do>

Protein Interactions: IntAct <http://www.ebi.ac.uk/intact/>

Chemical entities: ChEMBL <https://www.ebi.ac.uk/chembl/>

Pathways: Reactome <http://www.reactome.org/>

Systems: BioModels <http://www.ebi.ac.uk/biomodels-main/>

Slide 4-22: Example Genome Database: Ensembl

Slide 4-22: Example Genome Database: Ensembl

Search: All species for []

e.g. BRCA2 is at X:100000..200000 in coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Popular genomes

Human (Homo sapiens)
Mouse (Mus musculus)
Zebrafish (Danio rerio)

★ [Log in to customize this list](#)

All genomes

— Select a species —

[View full list of all Ensembl species](#)

Other species are available in [Ensembl FISH](#) and [Ensembl Genomes](#)

ENCODE data in Ensembl

Variant Effect Predictor

Gene expression in different tissues

Find SNPs and other variants for my gene

Retrieve gene sequence

Compare genes across species

Use my own data in Ensembl

Learn about a disease or phenotype

Sanger Ensembl is a joint project between [EMBL](#), [ECB](#) and the [Wellcome Trust Sanger Institute](#) to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl receives major funding from the Wellcome Trust. Our [acknowledgements page](#) includes a list of additional current and previous funding bodies.

Ensembl release 73 - September 2015 @ [GTDB](#) [VISTA](#)

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

A. Holzinger 709.049 41/92 Med Informatics L04

Ensembl (not to mix up with Ensemble ;-)) is a good example for a Genome database and is a joint project between the European Bioinformatics Institute and the Wellcome Trust Sanger Institute, which was launched in 1999 in response to the imminent completion of the Human Genome Project (Flicek et al., 2011). Its aim remains to provide a centralized resource for geneticists, molecular biologists studying the genomes of our own species and other vertebrates and model organisms. Ensembl provides one of several well-known genome browsers for the retrieval of genomic information.

Slide 4-23 Ex. Gene Expression Database: ArrayExpress

Slide 4-23 Ex. Gene Expression Database: ArrayExpress

EMBL-EBI ArrayExpress

Services Research Training Industry About us

Home Experiments Arrays Submit Help About ArrayExpress

Feedback Login

ArrayExpress - functional genomics data

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

Data Content
Updated today at 06:00

- 43495 experiments
- 1233850 assays
- 18.51 TB of archived data

Latest News

1 November 2013 - Need to keep your unpublished ArrayExpress microarray data private for longer?
Microarray experiment submitters, have you ever wondered if you could just change the release date of unpublished ArrayExpress data by yourself without emailing curators? Now you can! Use our new release date changing tool (more details on this help page). Submitters of high-throughput sequencing experiments, please continue to email us at miamexpress@ebi.ac.uk for release date changes so we can make sure the sequence read records at the European Nucleotide Archive are kept in sync.

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

A. Holzinger 709.049 42/92 Med Informatics L04

ArrayExpress is a database of functional genomics experiments that can be queried and the data downloaded. It includes gene expression data from microarray and high throughput sequencing studies. Data is collected to MIAME and MINSEQE standards. Experiments are submitted directly to ArrayExpress or are imported from the NCBI GEO database.

MIAME = Minimum Information About a Microarray Experiment. This is the data that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment (Brazma et al., 2001). The six most critical elements contributing towards MIAME are:

- 1) The raw data for each hybridisation (e.g., CEL or GPR files),
- 2) The final processed (normalised) data for the set of hybridisations in the experiment;
- 3) The essential sample annotation including experimental factors and their values,
- 4) the experimental design including sample data relationships;
- 5) Annotation of the array (e.g., gene identifiers, genomic coordinates, probe oligonucleotide sequences or reference commercial array catalog number), and
- 6) Laboratory and data processing protocols (e.g., what normalisation method has been used to obtain the final processed data); see:

<http://www.mged.org/Workgroups/MIAME/miame.html>

Slide 4-24: Example Protein Interaction Database: IntAct

The screenshot shows the IntAct website interface. At the top, there's a header with the IntAct logo and navigation links: Home, Advanced Search, Tools, Data Submission, Downloads, Documentation, Acknowledgements, and Contact Us. Below the header, there's a search bar with a 'Search' button and a 'Clear' button. To the right of the search bar, there's a link to 'Show Advanced Fields' and a note about MQL syntax reference. Below the search bar, there's a list of search tips and examples. The main content area shows a summary of interaction statistics: Publications (12131), Experiments (31292), Interactions (434941), and Interactors (79805). To the right of the statistics, there's a section titled 'Manually curated content is added to IntAct by the following organisations:' with logos for MINT, UniProt, SIB, I2D, Innoter, and Molecular Connections. At the bottom, there's a URL: <http://www.ebi.ac.uk/intact/>.

EMBL-EBI

Services Research Training Industry About us

IntAct

Home
Advanced Search
Tools
Data Submission
Downloads
Documentation
Acknowledgements
Contact Us

IntAct v1.10.0

Search: Search Clear [Show Advanced Fields](#) [MQL syntax reference](#)

- Free text search will look by default for interactor identifier, species, interaction id, detection method, interaction type, publication identifier or author, interactor xrefs, interaction xrefs
- For a more specific search, use MQL syntax or advanced search
- Search based on exact word matches, eg BRCA2 will not match BRCA2B
- Search for isoforms of 'P12345' by using 'P12345'

Examples

- Gene name: e.g. BRCA2
- UniProtKB Ac: e.g. Q00552
- UniProtKB ID: e.g. Q00552
- Pubmed ID: e.g. 10831811

[Support and feedback](#)

Home Search Interactions (434941) Browse Lists Interaction Details Molecular View Graph

IntAct provides a freely available, open source database system and analysis tools for molecular interaction data. All interactions are derived from literature curation or direct user submissions and are freely available. To perform a search in the IntAct database use the search box above.

Publications Experiments Interactions Interactors

12131 31292 434941 79805

[Citing IntAct](#)

Manually curated content is added to IntAct by the following organisations:

MINT UniProt SIB I2D Innoter Molecular Connections

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

A. Holzinger 709.049 43/92 Med Informatics L04

IntAct is an open source database for protein-protein interactions. The web interface provides both textual and graphical representations of such protein interactions, and allows exploring interaction networks in the context of the GO annotations of the interacting proteins. Moreover, a web service allows direct computational access to retrieve interaction networks in XML format. IntAct contains binary and complex interactions imported from the literature and curated in collaboration with the Swiss-Prot team, making intensive use of controlled vocabularies to ensure data consistency (Hermjakob et al., 2004).

<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 July
Original model: [BioModels00000434](#)

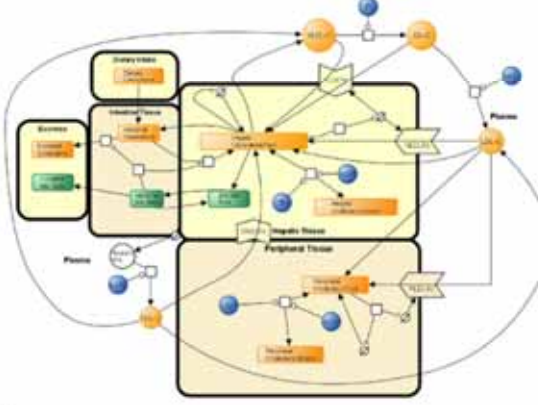
Cardiovascular disease is by far the most prevalent disease in ageing populations. Correlated with alterations in lipid metabolism profiles, it has estimated incidence rate of 30-40% in the UK population, over the age 85. Low-Density Lipoprotein Cholesterol (LDL-C), a prominent component in lipid metabolism, 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. Its 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) 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, BioModels00000434]. It consists of 6 compartments (Figure 1) and is composed of a series of coupled ODEs.



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

A. Holzinger 709.049 44/92 Med Informatics L04

The BioModels Database is a freely-accessible online resource for storing, viewing, retrieving, and analyzing published, peer-reviewed quantitative models of biochemical and cellular systems. The structure and behavior of each simulation model are thoroughly checked; in addition, model elements are annotated with terms from controlled vocabularies as well as linked to relevant data resources. Models can be examined online or downloaded in various formats and reaction network diagrams can be generated from the models in several formats. BioModels Database also provides features such as online simulation and the extraction of components from large scale models into smaller sub-models. The system provides a range of web services that external software systems can use to access up-to-date data from the database (Li et al., 2010). <http://www.ebi.ac.uk/biomodels/>


Note: Quantitative models of biochemical and cellular systems are used to answer research questions in the biological sciences and digital modeling is of growing interest in molecular and systems biology. A well-known example is the Virtual Human (Kell, 2007).



The largest monastery library of the world – a good example for a well-defined knowledge space.

Yes, perfectly correct – this Golden Retriever is bringing back the wooden stick – he is retrieving it.

This is exactly what the word to retrieve means: bringing something back.



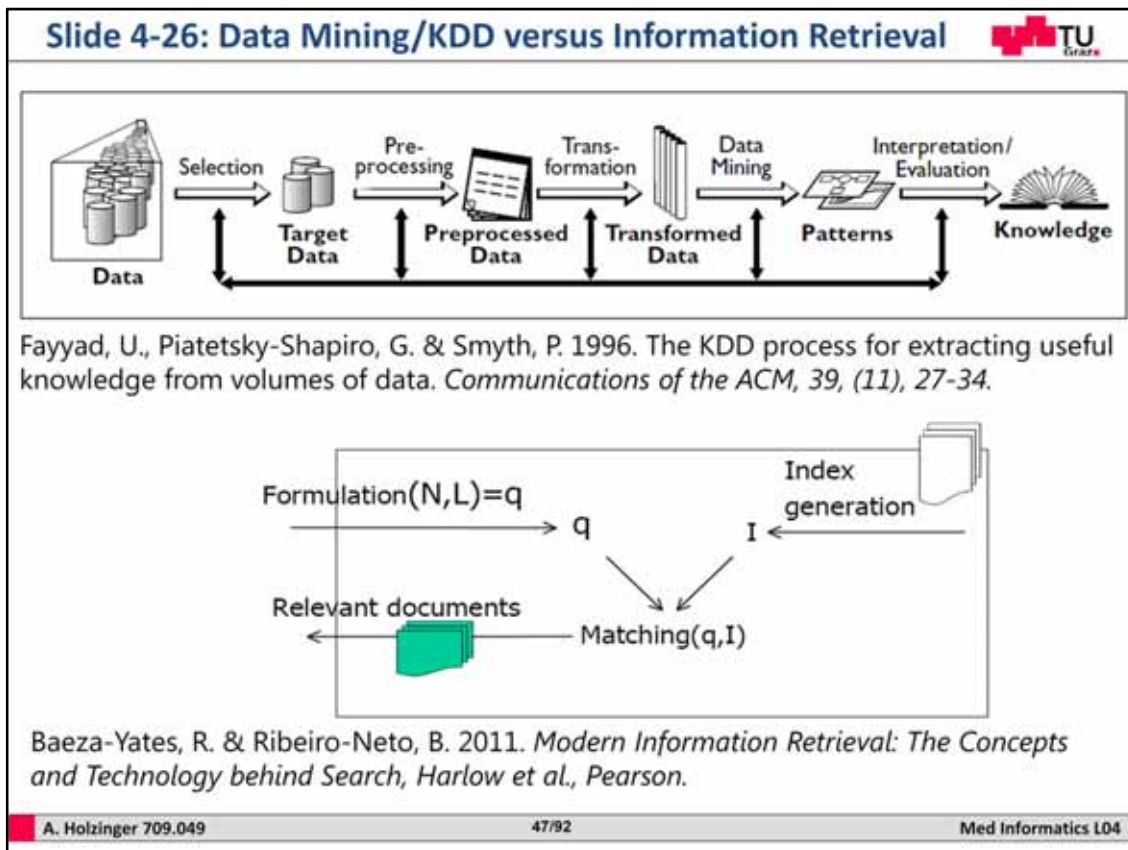
What is the difference between retrieval and discovery?

A. Holzinger 709.049 46/92 Med Informatics L04

Please remember the basic differences between retrieval and discovery: Retrieval is bringing back an already known object, whereas discovery is finding something which was previously unknown. In other words: Retrieval is dealing with known objects and Discovery/Mining is finding new things – in our case new insight (sensemaking) into data. Slide 4-26 makes it clear:


Maimon & Rokach (2010) (Maimon & Rokach, 2010) define Knowledge Discovery in Databases (KDD) as an automatic, exploratory analysis and modeling of large data repositories and the organized process of identifying valid, novel, useful and understandable patterns from large and complex data sets. Data Mining (DM) is the core of the KDD process (Witten, Frank & Hall, 2011).

The term KDD actually goes back to the machine learning and Artificial Intelligence (AI) community (Piatetsky-Shapiro, 2000). Interestingly, the first application in this area was again in medical informatics: The program Rx was the first that analyzed data from about 50,000 Stanford patients and looked for unexpected side-effects of drugs (Blum & Wiederhold, 1985). The term really became popular with the paper by Fayyad et al. (1996) (Fayyad, Piatetsky-Shapiro & Smyth, 1996), who described the KDD process consisting of 9 subsequent steps:



1. Learning from the application domain: includes understanding relevant previous knowledge, the goals of the application and a certain amount of domain expertise;
2. Creating a target dataset: includes selecting a dataset or focusing on a subset of variables or data samples on which discovery shall be performed;
3. Data cleansing (and preprocessing): includes removing noise or outliers, strategies for handling missing data, etc.);
4. Data reduction and projection: includes finding useful features to represent the data, dimensionality reduction, etc.;
5. Choosing the function of data mining: includes deciding the purpose and principle of the model for mining algorithms (e.g., summarization, classification, regression and clustering);
6. Choosing the data mining algorithm: includes selecting method(s) to be used for searching for patterns in the data, such as deciding which models and parameters may be appropriate (e.g., models for categorical data are different from models on vectors over reals) and matching a particular data mining method with the criteria of the KDD process;
7. Data mining: searching for patterns of interest in a representational form or a set of such representations, including classification rules or trees, regression, clustering, sequence modeling, dependency and line analysis;
8. Interpretation: includes interpreting the discovered patterns and possibly returning to any of the previous steps, as well as possible visualization of the extracted patterns, removing redundant or irrelevant patterns and translating the useful ones into terms understandable by users;
9. Using discovered knowledge: includes incorporating this knowledge into the performance of the system, taking actions based on the knowledge or documenting it and reporting it to interested parties, as well as checking for, and resolving, potential conflicts with previously believed knowledge (Holzinger, 2013).


In Information retrieval a query q is defined as a formulation $(N,L)=q$ and the matches with an index I $\text{Matching}(q,I)$ retrieves relevant data to satisfy the search query (Baeza-Yates & Ribeiro-Neto, 2011).



What is the difference between data retrieval and information retrieval?

A. Holzinger 709.049 48/92 Med Informatics L04

Please remember the differences between data objects and information objects – data is an abstract representation in the computational space – information is perceivable for the cognitive space (Note that it does not mean that information is automatically knowledge –for getting knowledge we must use both our perception and cognition, i.e. human intelligence)

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;

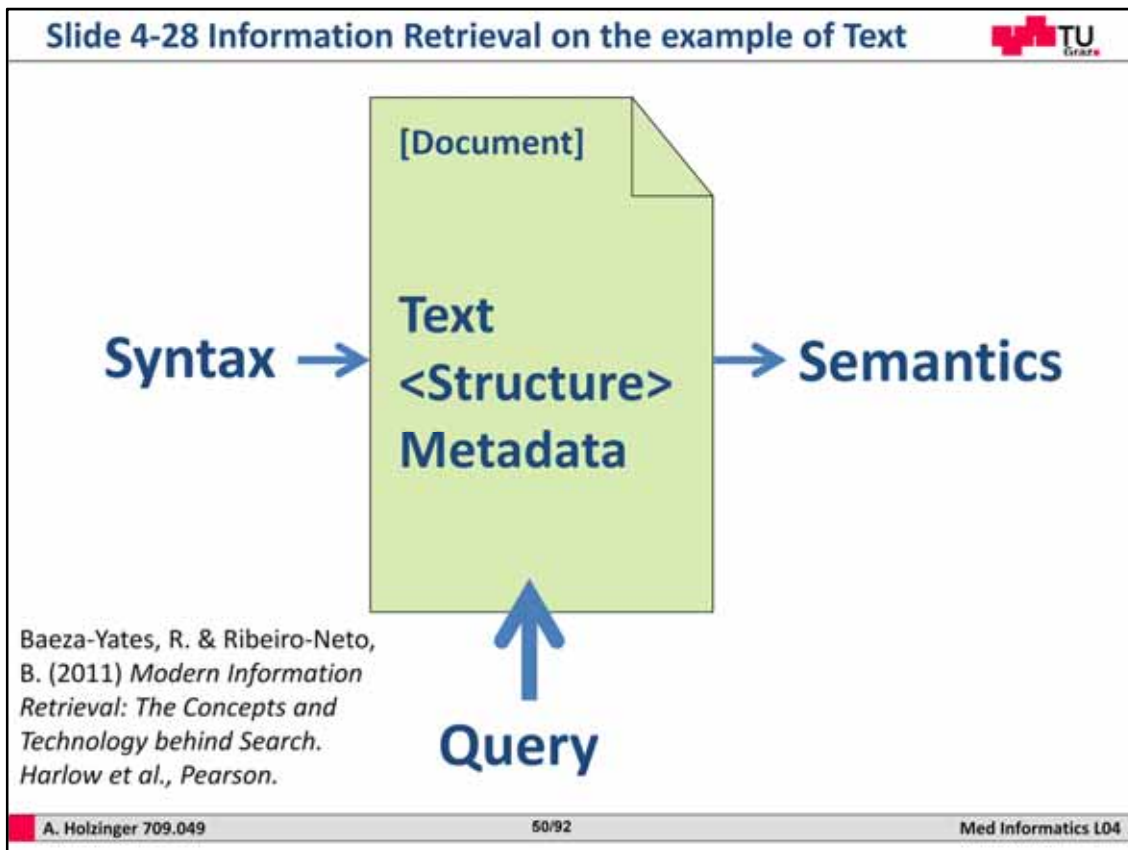
Van Rijsbergen, C. J. (1979) *Information Retrieval (Second Edition)*. London, Butterworths.

A. Holzinger 709.049 49/92 Med Informatics L04

An excellent start in the determination between DR and IR is the work of (Van Rijsbergen, 1979): The most important difference is that the data model in DR is deterministic, whereas we speak about probable information in the IR Model, hence information retrieval is probabilistic (Simonic & Holzinger, 2010).

*Monothetic = type in which all members are identical on all characteristics;

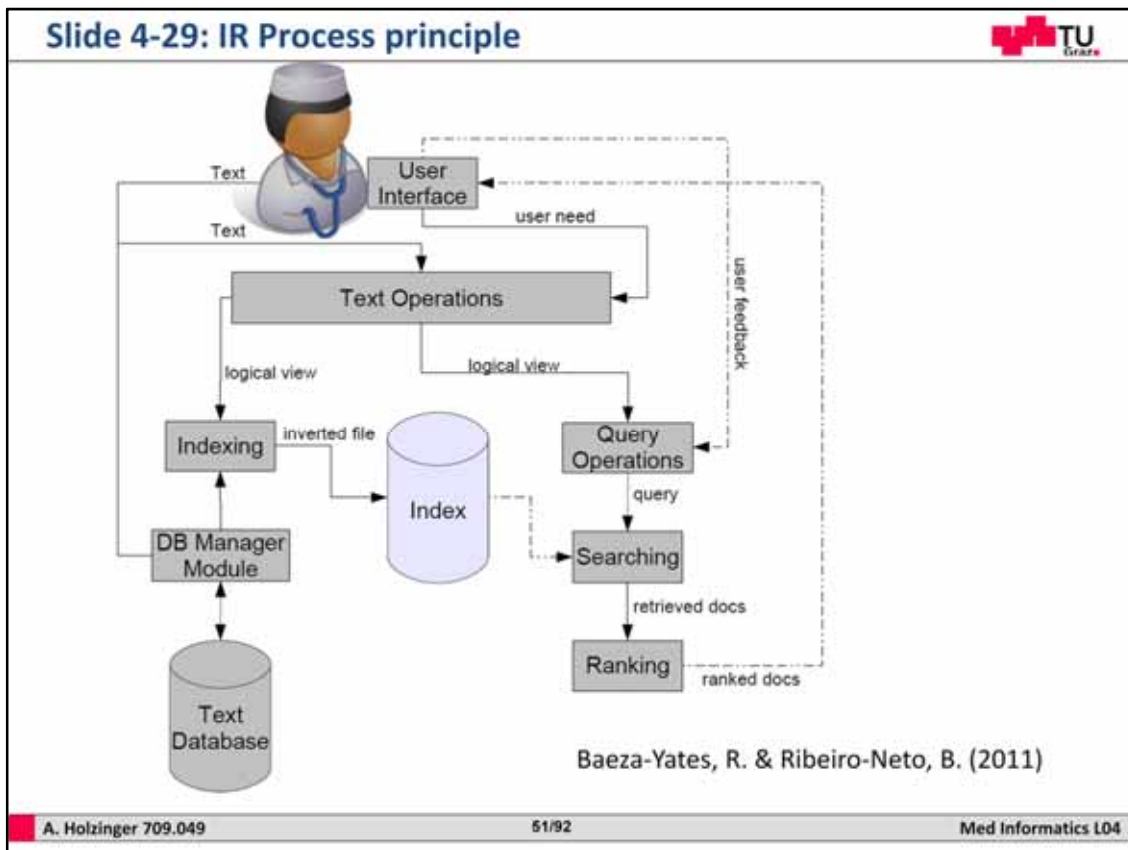
**Polythetic = type in which all members are similar, but not identical;



IR can be defined as a re-call of already existing information, not aiming at the discovery of new structures as it is the goal in Knowledge Discovery and Data Mining (see →Lecture 6). As we have already heard several times, in hospital information systems most of the data consists of medical documents, which consist mostly of unstructured information: text. But: What is text?

From a computational perspective, text consists of sequences of character strings, the syntax (Hotho, Nürnberger & Paaß, 2005), hence it is an abstract representation of natural language and the challenges are in semantics (meaning). Text processing belongs to the field of Natural language processing (NLP) which is highly interdisciplinary, dealing with the interaction between the cognitive space (natural languages) and the computational space (formal languages). As such, NLP is closely related to HCI. Text mining is a subfield of data mining.

The original goal of IR was to find documents which contain answers to questions and not the finding of answers itself (Hearst, 1999). For this purpose statistical measures and methods are used, and we need a formal description first.



This is the general principle: The end user formulates his query via the user interface, in form of a Text Operations ("user need"). The next step is the representation (logical document view D in the formal model in →Slide 4-30) of the documents and the representation of the reasoning strategy, query logical view Q (compare with →Slide 4-30 and →Slide 4-31). The result is a ranking of the retrieved documents, which will be displayed via the user interface.

Slide 4:30: Formal Description of IR Models



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

- **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**");
- \mathcal{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.

Modeling the IR-process is complex, because we are dealing with imprecise, vague and uncertain elements, thus it is difficult to formalize due to high influences of human factors, i.e. relevance and information needs, which are highly subjective and context specific. However, in the definition of any IR-model we can identify some common aspects (Canfora & Cerulo, 2004). The first step is the representation of documents and information needs. From these representations a reasoning strategy can be defined, which solves a representation similarity problem to compute the relevance of documents with respect to the queries. Various strategies have been introduced with the aim of improving the IR-process. We classify these methodologies under two main aspects: Representation (query & document, see Slide →4-33) and Reasoning (application of diverse methods, see →Slide 4-34).

Let the IR Model be a quadruple

$$\text{Eq. 4-1} \quad \text{IR} = \{D, Q, \mathcal{F}, R(q_i, d_j)\}$$

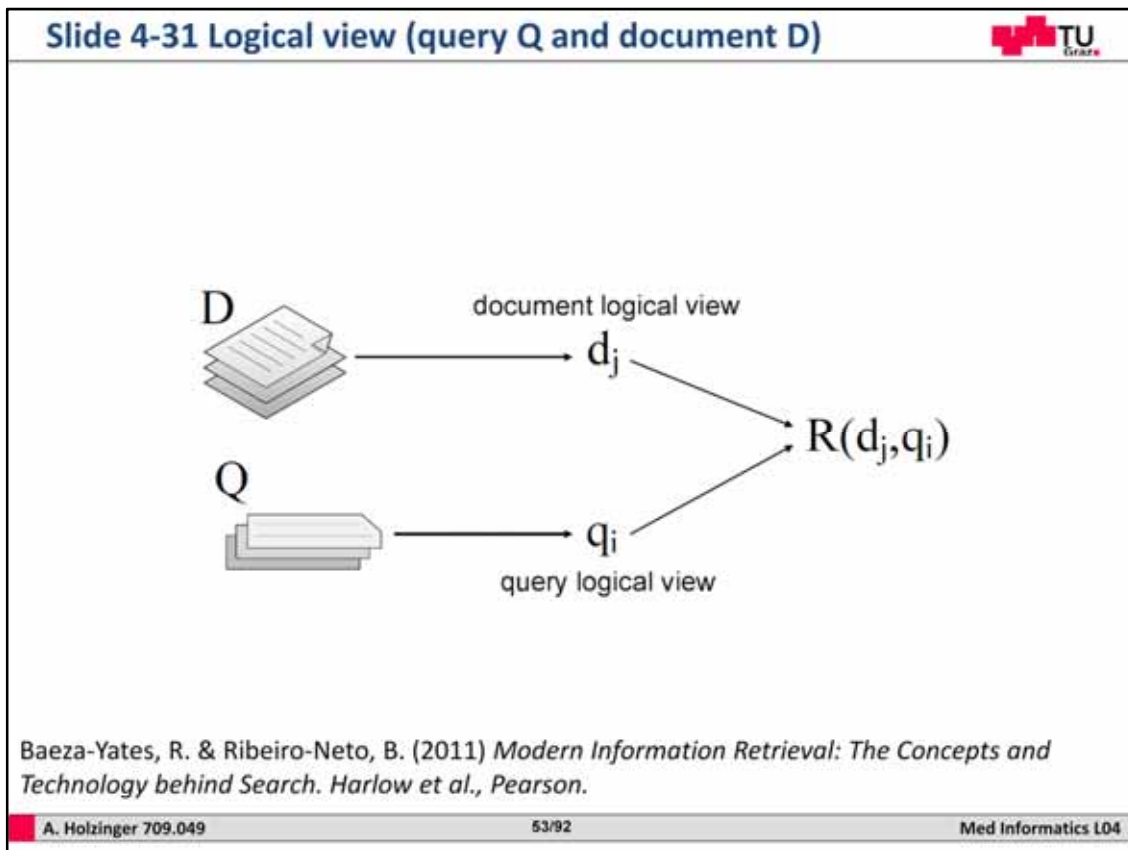
D is a set composed of logical views (representation component) of the documents within a collection;

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;

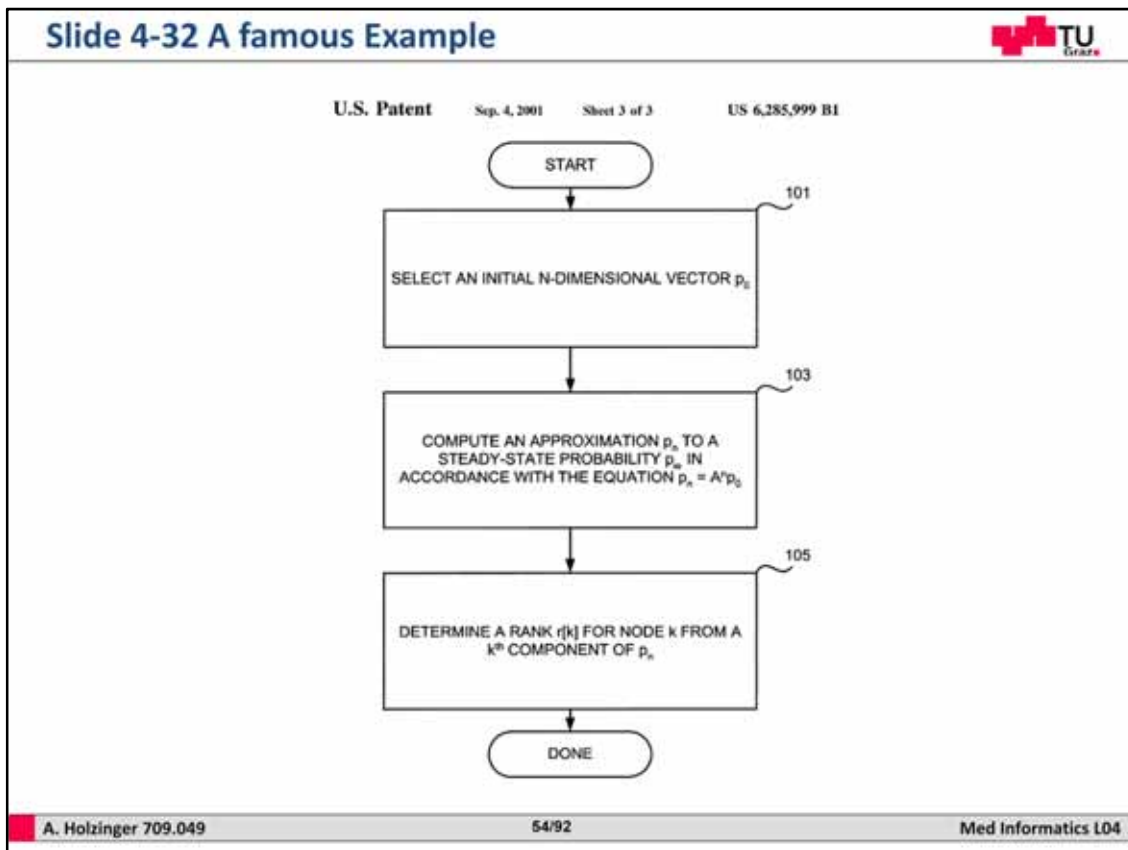
$R(q_i, d_j)$ is a ranking function (→Slide 4-31) that associates a real number with a query representation $q_i \in Q$ and a document representation $d_j \in D$. Such ranking defines an ordering among the docs with regard to the query q_i .

The end user in →Slide 4-29 formulates his query in form of a text operation, the next step is the representation (logical view **D**) of the documents and the representation of the reasoning strategy, both logical views **D** and **Q** (compare with Slide 4-31) result in a ranking of the retrieved documents.



The logical views D and Q result in the ranking function $R(q_i, d_j)$ according to (Baeza-Yates & Ribeiro-Neto, 2011)

Speak: R indexed d subscript j and q subscript i



Guess which algorithm this is? ☺

A short description can be found in

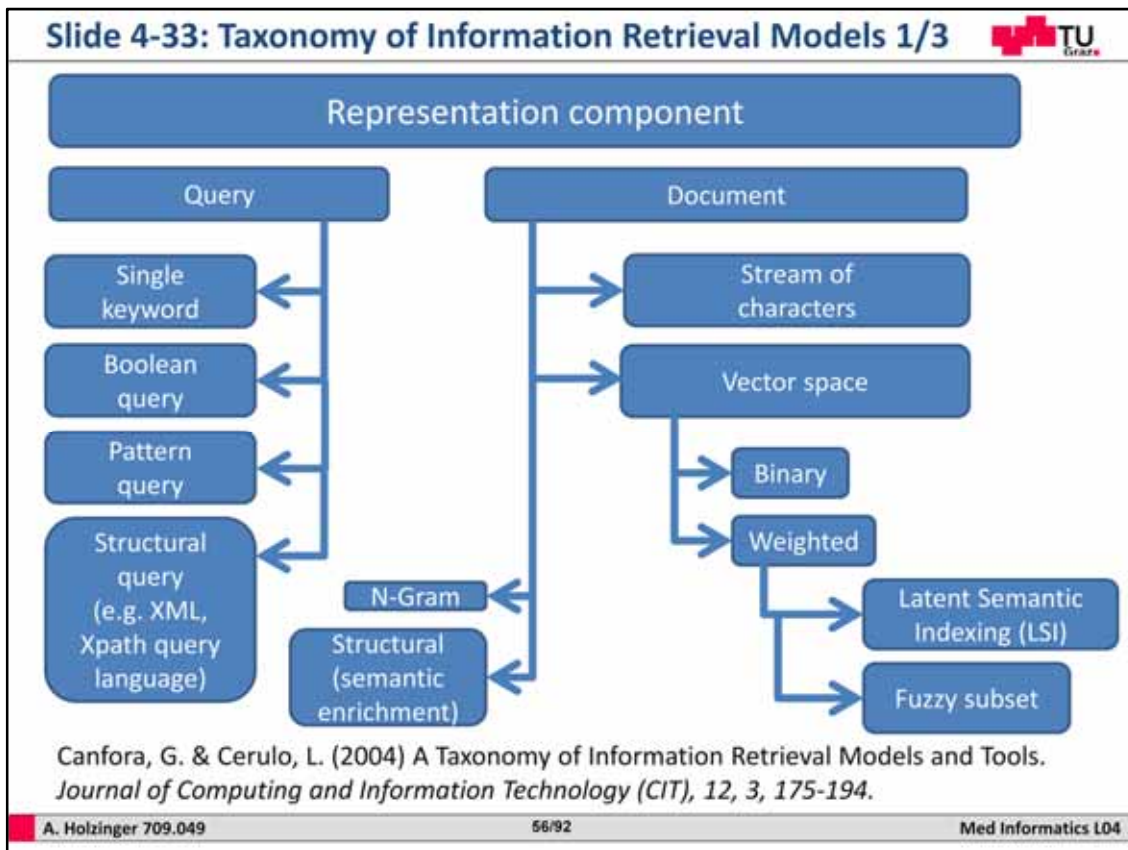
Hastie, T., Tibshirani, R. & Friedman, J. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition, New York, Springer.



Remember: We have two components: Representation and Reasoning component

A. Holzinger 709.049 55/92 Med Informatics L04

Yes! A lot different methods – every method having particular advantages and disadvantages – we cannot discuss much here, but we can get a rough overview.



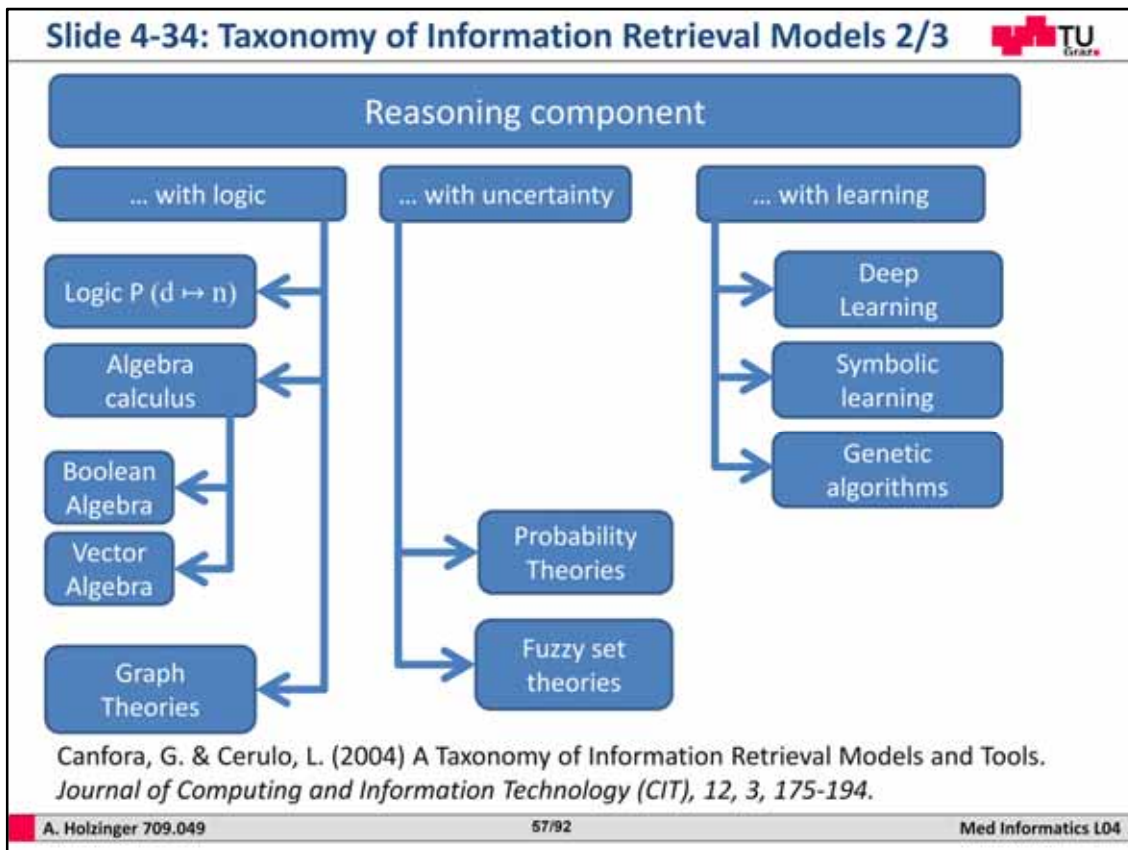
The representation component is an essential part of every IR system, as it is the representation of the information itself (visible to the user): information can be processed if it is represented in an appropriate way. Queries are the representation of information needs of a user.

Note: A text can be characterized by using four attributes: syntax, structure, semantics, and style. A text has a given syntax and a structure, which are usually dictated by the application or by the person who created it. Text also has semantics, specified by the author of the document. Additionally, a document may have a presentation style associated with it, which specifies how it should be displayed or printed. In many approaches to text representation the style is coupled with the document syntax and structure (LaTeX). XML separates the representation of syntax and structures, defined either by a DTD or an XSD, and style, which is captured by XSL (Canfora & Cerulo, 2004).

Note: An n-gram is a subsequence of n items from a given sequence. The items in question can be phonemes, syllables, letters, words or base pairs according to the application.

An n-gram of size 1 is referred to as a "unigram"; size 2 is a "bi-gram" (or, less commonly, a "di-gram"); size 3 is a "tri-gram"; size 4 is a "four-gram" and size 5 or more is simply called an "n-gram". Some language models built from n-grams are "(n - 1)-order Markov models".

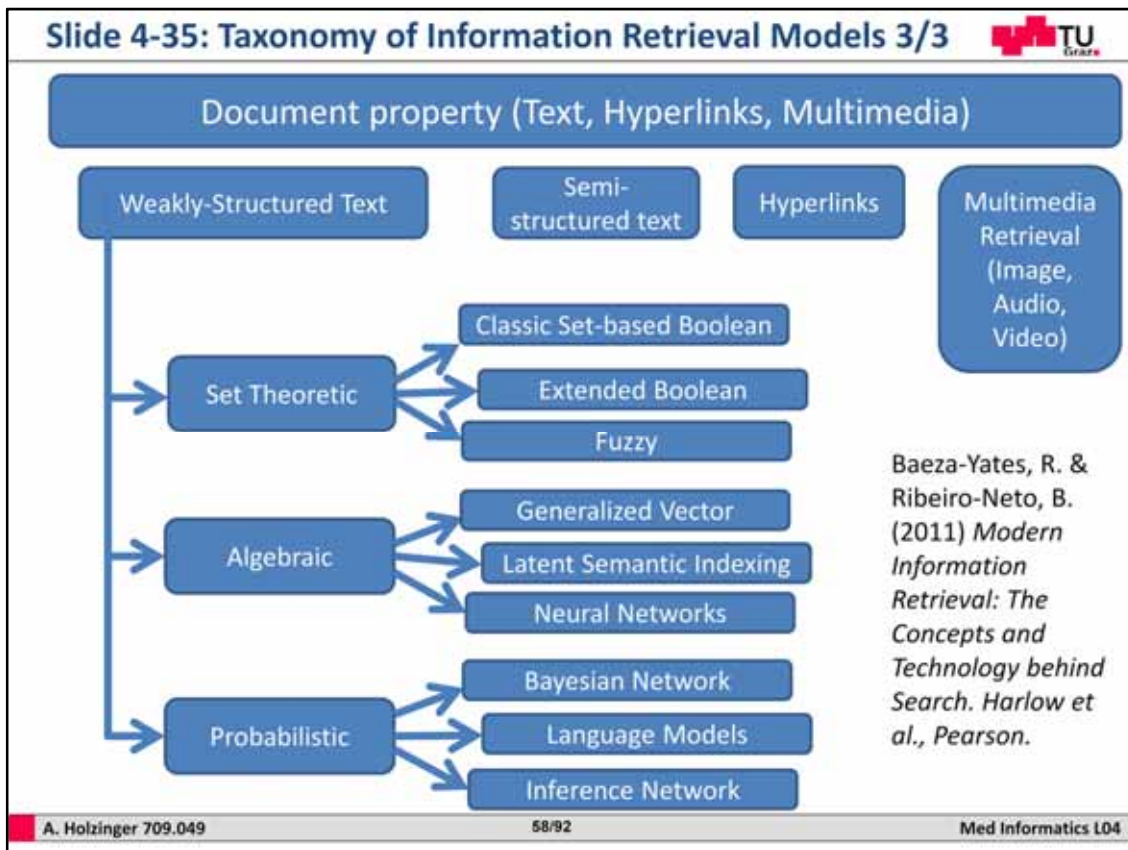
An n-gram model is a type of probabilistic model for predicting the next item in such a sequence. n-gram models are used in various areas of statistical natural language processing and genetic sequence analysis.



Deep learning algorithms are based on distributed representations, with the assumption that observed data is generated by the interactions of many different factors on different levels. Deep learning adds the assumption that these factors are organized into multiple levels, corresponding to different levels of abstraction or composition and various numbers of layers and layer sizes can be used to provide different amounts of abstraction. Bengio, Y.; Courville, A.; Vincent, P. (2013). "Representation Learning: A Review and New Perspectives". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35 (8): 1798–1828

Reasoning refers to the set of methods, models, and technologies used to match document and query representations in the retrieval task. Strictly related with the reasoning component is the concept of relevance. The primary goal of an IR system is to retrieve the documents relevant to a query. The reasoning component defines the framework to measure the relevance between documents and queries using their representations (Canfora & Cerulo, 2004).

Google, for example, uses a keyword based vector space model (see →Slide 4-38) along with graph-based probability theories and Fuzzy set theories. Slide 4-35 shows a concise overview of some selected methods, according to various document properties.



There are many methods of IR, for details consult a standard reference e.g. Baeza-Yates & Ribeiro-Neto (2011). Set theoretic approaches include the Classic Set-based Boolean, the Extended Boolean and the Fuzzy Approach; Algebraic approaches include the Generalized Vector Model, Latent Semantic Indexing (LSI), Neural Networks; and the Probabilistic approach includes Bayesian Networks, Language Models and Inference Networks. We will discuss only a few and these very briefly, so that you have a quick overview: The set theoretic approach: Boolean Model in Slide 4-36 and Slide 4-37; the Vector Space Model in Slide 4-38 to Slide 4-42; and the Probabilistic Model in Slide 4-43 to Slide 4-44.

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


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

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)

Documents/queries are represented as a set of index terms; queries are Boolean expressions (AND, OR, NOT); For the Boolean model, the index term weight variables are binary, i.e. $w_{(i,j)} \in \{0 | 1\}$. A query q is a conventional Boolean expression. Let \vec{q}_{dnf} be the disjunctive normal form of 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

If $sim(d_j, q) = 1$ then the Boolean model predicts that the document d_j is relevant to query q . Otherwise the prediction is that the document is not relevant. For details please refer to (Baeza-Yates & Ribeiro-Neto, 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.

A. Holzinger 709.049 60/92 Med Informatics L04

The Boolean Model has several advantages, including easy to understand, exact formalism and the query language is expressive; however, serious disadvantages, e.g. no partial matches, the “bag-of-words” representation does not accurately consider the semantics of documents (Vallet, Fernández & Castells, 2005), and the query language is complicated, finally the retrieved documents can not be ranked.

The Extended Boolean Model (EBM) by (Salton, Fox & Wu, 1983) overcomes some disadvantages by making use of partial matching and term weights, similar as in the vector space model. Moreover, as the vector-processing system suffers from one major disadvantage: the structure inherent in the standard Boolean query formulation is absent, the EBM combines the characteristics of the Vector Space Model with the properties of Boolean algebra. Hence, the EBM can also be applied, when the initial query statements are available as natural language formulations of user needs, rather than as conventional Boolean formulations.

Slide 4-38 Example Algebraic Model: Vector Space Model


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

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

Now we carry out a document transformation and get vectors:

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

Now we count the frequency of the terms and get:

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

A. Holzinger 709.049
61/92
Med Informatics L04

The vector space model (VSM) represents documents as vectors in the m-dimensional space (Salton, Wong & Yang, 1975).

Thus, documents can be compared by vector operations and queries can be performed by encoding the query terms similar to the documents in a query vector. This query vector can be compared to each document, which returns a result list by ordering the documents according to the computed similarity. The main task of the vector space representation of documents is to find an appropriate encoding of the feature vector. Each element of a vector usually represents a word (see →Slide 4-40) of the document collection. The size of the vector is defined by the number of words of the complete document collection. The easiest way of document encoding is to use binary term vectors, that means a vector element is set to 1 if the corresponding word is used in the document and to 0 if the word is not (Equation 4-4). This encoding results in a simple Boolean comparison. To improve the performance usually term weighting schemes are used, where the weights reflect the importance of a word in a specific document of the considered collection. Large weights are assigned to terms that are used frequently in relevant documents but rarely in the whole document collection (Salton & Buckley, 1988). Thus a weight $w(d; t)$ for a term t in document d is computed by term frequency $tf(d; t)$ times inverse document frequency $idf(t)$, which describes the term specificity within the document collection. The ranking can be made by using the Cosines Similarity (see →Slide 4-41). The cosine of the angle between two vectors is a measure of how “similar” they are, which in turn, is a measure of the similarity of these strings. If the vectors are of unit length, the cosine of the angle between them is simply the dot product of the vectors (Tata & Patel, 2007).

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




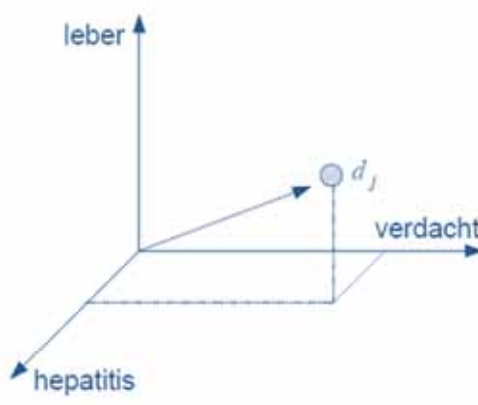
$$D_{m \times n} = \begin{Bmatrix} w_{1,1} & w_{1,2} & \cdots & w_{1,n-1} & w_{1,n} \\ w_{2,1} & w_{2,2} & & w_{2,n-1} & w_{2,n} \\ \vdots & & \ddots & & \vdots \\ w_{m-1,1} & w_{m-1,2} & & w_{m-1,n-1} & w_{m-1,n} \\ w_{m,1} & w_{m,2} & \cdots & w_{m,n-1} & w_{m,n} \end{Bmatrix}$$

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

As a result we get a matrix representation, and now we can apply vector algebra, or particular linear algebra – here still in R3. Mathematically, we can work in arbitrarily high dimensional spaces. The major problem involved is the mapping back into R2.

One very positive aspect is that we can look for getting sparse matrices, i.e. we save a lot of computational power.

Slide 4-40: d_j 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.

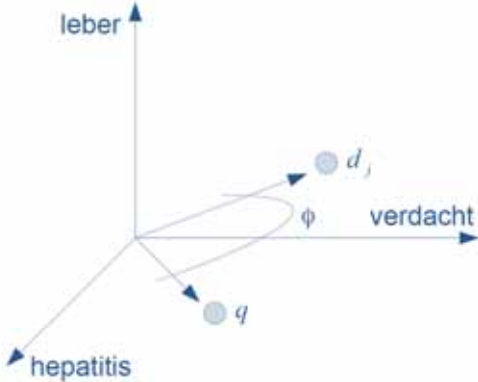
Turney, P. D. & Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, (1), 141-188.
Survey article 922 citations yet ...

A. Holzinger 709.049 63/92 Med Informatics L04

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

Computers understand very little of the meaning of human language. This profoundly limits our ability to give instructions to computers, the ability of computers to explain their actions to us, and the ability of computers to analyse and process text. Vector space models (VSMs) of semantics are beginning to address these limits. Turney et al. (2010) surveys the use of VSMs for semantic processing of text. They organize the literature on VSMs according to the structure of the matrix in a VSM. There are currently three broad classes of VSMs, based on term-document, word-context, and pair-pattern matrices, yielding three classes of applications. They survey a broad range of applications in these three categories and we take a detailed look at a specific open source project in each category. Their goal in this survey is to show the breadth of applications of VSMs for semantics, to provide a new perspective on VSMs for those who are already familiar with the area, and to provide pointers into the literature for those who are less familiar with the field.

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



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

$$\text{sim}(\vec{d}_j, \vec{q}) = \cos(\Phi) = \frac{\vec{d}_j \cdot \vec{q}}{|\vec{d}_j| \times |\vec{q}|}$$

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

A. Holzinger 709.049 64/92 Med Informatics L04

Beim Retrievalverfahren wird ein Ranking ähnlicher Dokumente über die Cosinus Ähnlichkeit im m-Dimensionalen Vektorraum durchgeführt.

Information Need $Q \rightarrow q = (w_1, w_2, \dots, w_m)$

Wird ein Ranking ähnlicher Dokumente über die Cosinus Ähnlichkeit im m dimensionalen Vector Space Model durchgeführt

Der Vorteil dieser Methode ist, dass es ein einfaches mathematisches Modell darstellt,

Die Matrizen sind Sparse (ist also eine günstige Datenstruktur)


Das retrieval kann in $O(n)$ durchgeführt werden, daher gibt es ein relativ schnelles ranking

Nachteile: Die Wortanordnung geht verloren (Bag of Word Ansatz).

Es gibt viele weitere Methoden, wie z.B. Latent Semantic Analysis (LSA) usw.

Probabilistic Latent Semantic Analysis (PLSA)

Latent Dirichlet Allocation (LDA)

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.

A. Holzinger 709.049
65/92
Med Informatics L04

The advantages of the algebraic VSM include that it is easy to understand, partial matches are possible, documents can be sorted by rank, and it uses term-weighting schemes; on the other side there is a higher computational effort to calculate similarity, and the “bag-of-words” representation does not accurately consider the semantics of documents (Vallet, Fernández & Castells, 2005).

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 \bar{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 $P(\bar{R}|\vec{d}_j)$ be the probability that d_j is non-relevant to q . The similarity $\text{sim}(d_j, q)$ of the document d_j to the query q is defined as the ratio

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


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



Rev. Thomas Bayes
(1702-1761)

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

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


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	

A. Holzinger 709.049

67/92

Med Informatics L04

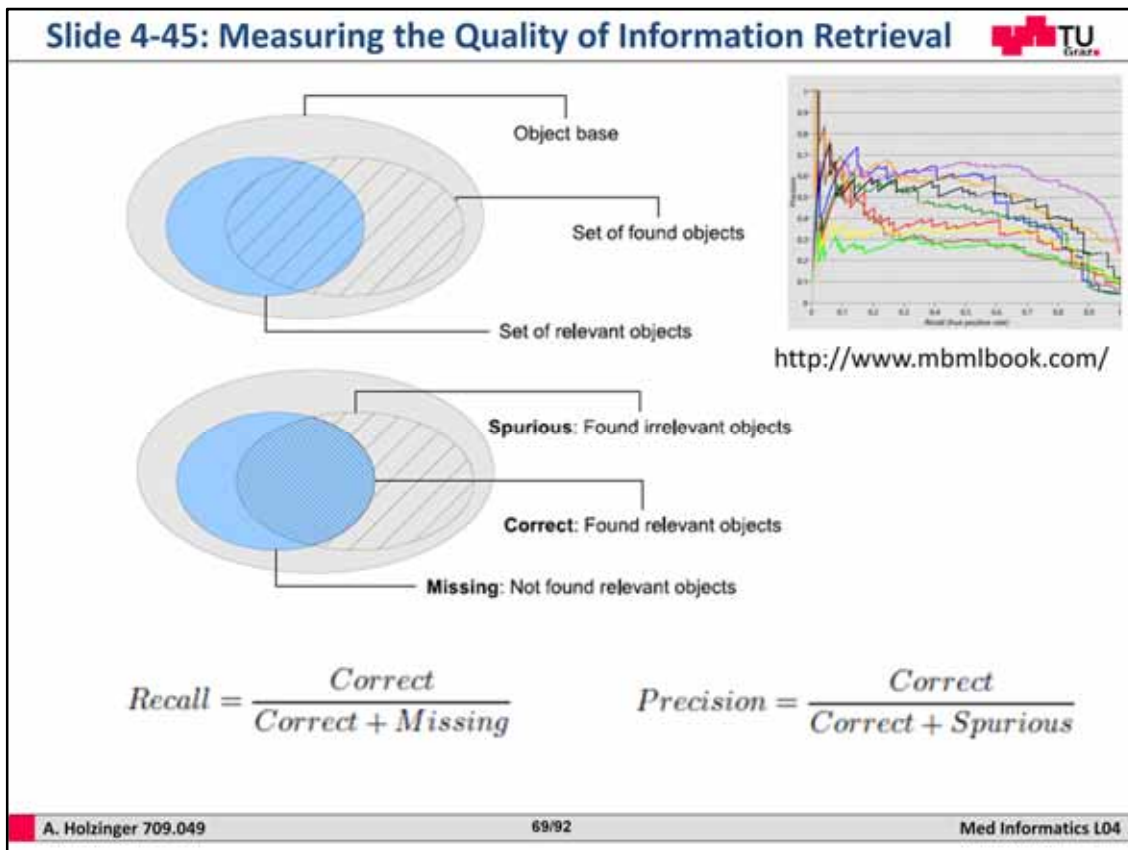
As in all models we have certain pros and cons, the probabilistic model has a big advantage: the documents can be ranked by relevance; however, on the disadvantageous side it is a binary model (binary weights), the index terms are assumed to be independent and lack of document normalization and there 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?

A. Holzinger 709.049 68/92 Med Informatics L04

Well, there are two main measurements



Recall and Precision – hard as a bone

Following this definition: Recall = Correct / (Correct + Missing) and Precision = Correct / (Correct + Spurious)

Precision P is the fraction of retrieved documents that are relevant to the search:

$$P = \frac{|\{\text{set of relevant docs}\} \cap \{\text{set of found docs}\}|}{|\{\text{set of found docs}\}|}$$

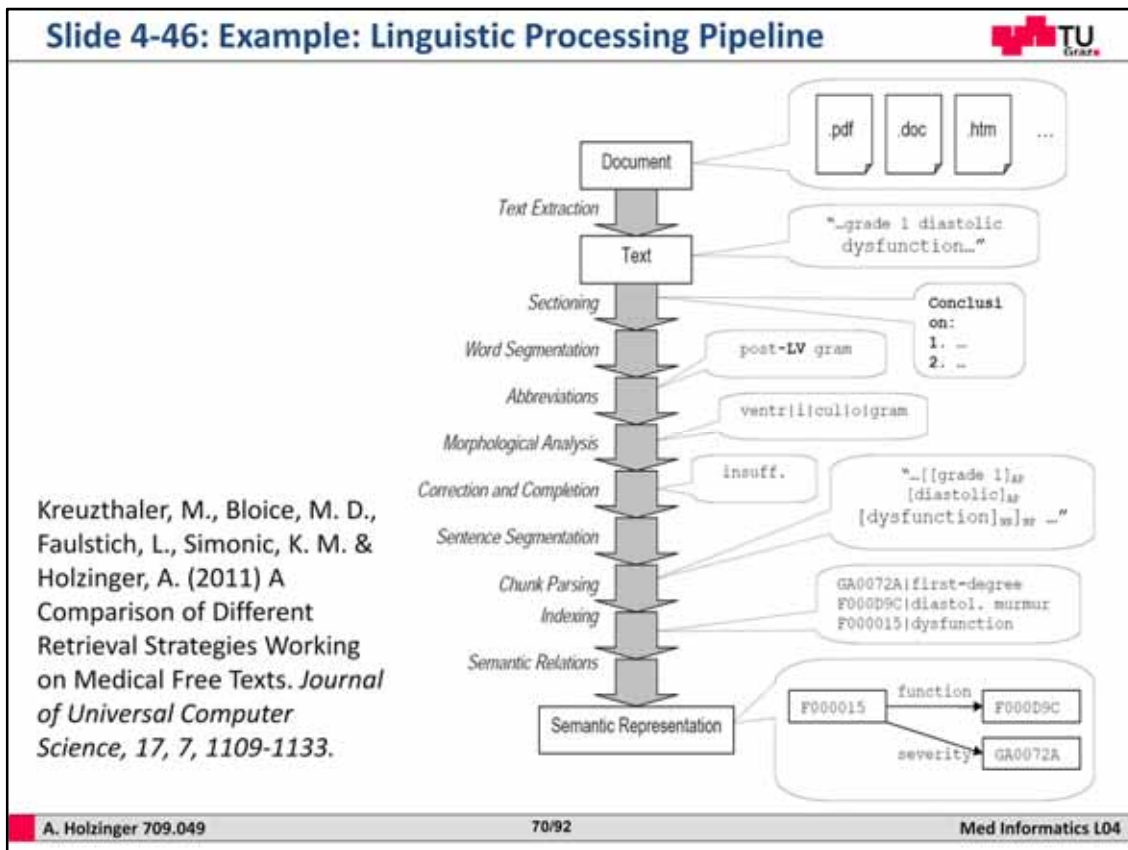
Recall R is the fraction of the documents that are relevant to the query that are successfully retrieved:

$$R = \frac{|\{\text{set of relevant docs}\} \cap \{\text{set of found docs}\}|}{|\{\text{set of relevant docs}\}|}$$

A combination of precision and recall is the harmonic mean of both, which is called F-measure:

$$F = \frac{2 \cdot (P \cdot R)}{(P + R)}$$

In classification 5 terms are used: true positives (=correct); true negatives (=correct); false positives (=spurious); false negatives (=spurious); not detected (=missing).



In this slide we see an overview of the linguistic processing pipeline that describes the steps that are performed from the document to its semantic representation. The domain knowledge used in the semantic retrieval system is modeled in the form of the medical semantic network ID MACSR (MSN). It uses the Wingert Nomenclature (WNC) as its medical terminology. The WNC is based on the German version of SNOMED developed by Friedrich Wingert. Although its main focus is on German, it, to a lesser extent, supports several other languages including English and French. The MSN forms a simple ontology whose concepts are organized in a taxonomy (isA-hierarchy) and a merology (anatomical partOf-hierarchy). Further relations between concepts are modeled by labeled edges. The MSN is divided into several subdomains, including:

- topography (i.e., anatomical concepts)
- morphology (e.g., fracture, fever)
- function (e.g., respiration)
- diseases (e.g., glaucoma)
- agents (e.g., pathogens, pharmaceutical substances)

Currently, the MSN contains more than 90,000 terms and 300,000 unique relations.

The query language follows a simple grammar, namely:

Query ::= Disjunction

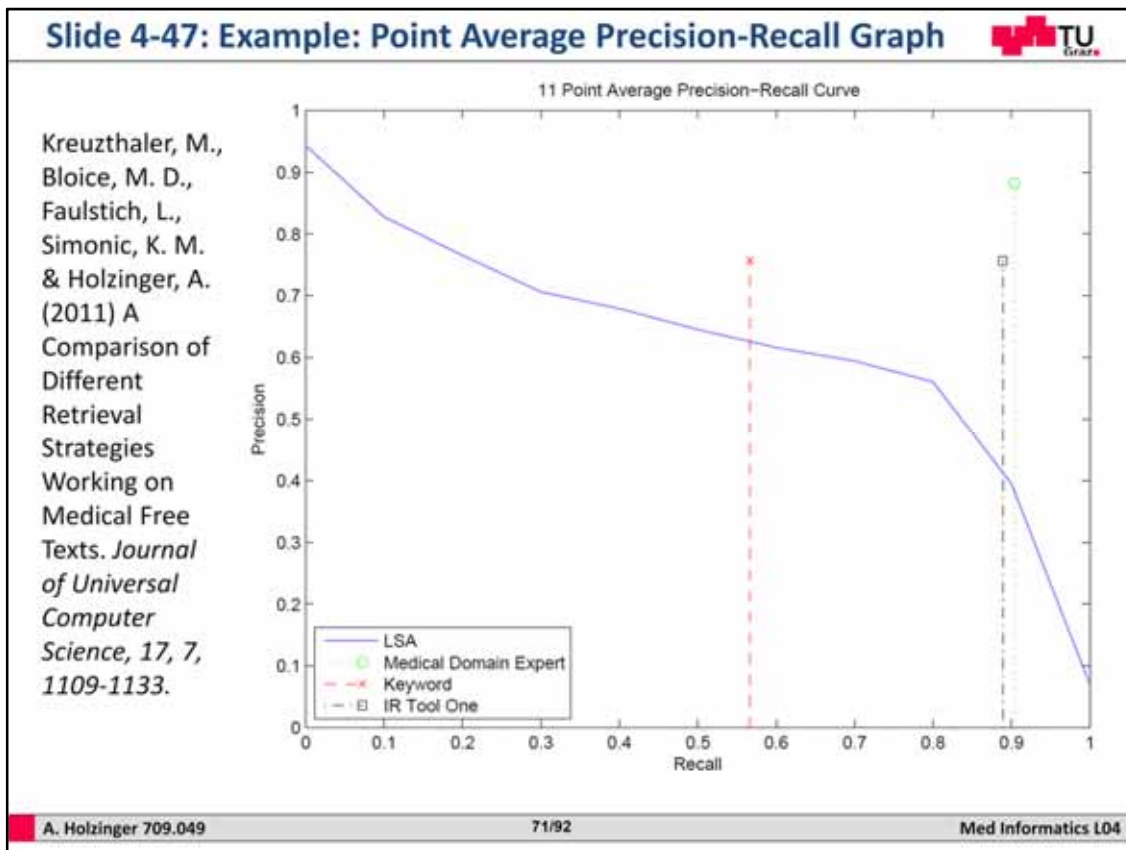
Disjunction ::= Conjunction | Conjunction ";" Disjunction

Conjunction ::= Atom | Atom "," Conjunction

Atom ::= Term | "!" Term

Thus a query forms a Boolean expression in disjunctive form over search

terms. Semantic query expansion has been discussed in several previous work (Kingsland, Harbourt, Syed & Schuyler, 1993), (Aronson, Rindfleisch & Browne, 1994) (Efthimiadis, 1996). The approach is as follows: each search term is indexed (using the linguistic processing methods described above) and replaced by the identifier of the WNC concept matching the term. These concept identifiers are called WNC indices. If the search term refers to a combination of several concepts in the WNC (e.g., Gastroparesis = Stomach + Paresis), the search term is replaced by a conjunction of the WNC (Kreuzthaler et al., 2011).




As can be seen from this Slide the medical domain expert outperforms the other retrieval methods, achieving high precision at a high recall level. Interestingly, the semantic based information retrieval tool achieves approximately the same recall level as the medical domain expert while having a lower precision value. This performance result is good, remembering the fact what effort the medical domain expert has to make to translate the information need into a query string. In contrast to this, the input for the information retrieval tool is short and clear so therefore less effort has to be made to transform the information need to the query language understood by the information retrieval tool.

Keyword search has a high precision value but a lower recall value. This result is clear when considering the fact that information needs that can be described by using these keyword(s) will achieve a high precision value. So, if documents are found they will be relevant but the


recall level will generally suffer. Looking at the Slide 4-47, keyword search achieves approximately the same precision as IR Tool One but a far worse recall. It is also possible that no search results are found at all when using the keyword search methodology as can be seen for the Neubildung, Darm information need (see Appendix B and Appendix A). In contrast to this, for this information need, IR Tool One has about the same precision recall levels as the medical domain expert, reflecting the semantic processing chain of the tool.

The LSA statistical retrieval method has, when compared to the other methods, a lower precision for all measured recall levels. This result gives the impression that LSA is applicable for getting high precision values for a particular amount of search results but hard

to use to achieve both high precision and high recall values, which is needed for example in clinical studies.


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

McKinsey Global Institute



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.




- \$600 to buy a disk drive that can store all of the world's music
- 5 billion mobile phones in use in 2010
- 30 billion pieces of content shared on Facebook every month
- 40% projected growth in global data generated per year vs. 5% growth in global IT spending
- 235 terabytes data collected by the US Library of Congress by April 2011
- 15 out of 17 sectors in the United States have more data stored per company than the US Library of Congress

A. Holzinger 709.049 72/92 Med Informatics L04

The future of big data is ... big ☺ and there will be many challenges for us to solve!

Not big data is the real challenge ... but



Complex Data!

What is interesting?

What is relevant?

A. Holzinger 709.049

73/92

Med Informatics L04

The grand questions of the future is how to make sense out of the data – mega questions include are: “What is interesting?” – and “What is relevant?”



Thank you!

A. Holzinger 709.049

74/92


Med Informatics L04

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	dosage	bid
PREPULSID SUSP 1MG/ML 100ML		10 mg	(08:00)
cisapride (ala 1 - water)	ORAL	15 mg	(18:00)
N	START: 28-07-04/10:23		
	STOP: 28-07-04/15:15		
6210	enter: Start Medication *33725		
LET OPI - STOPDATUM IS INGEVULD			41537

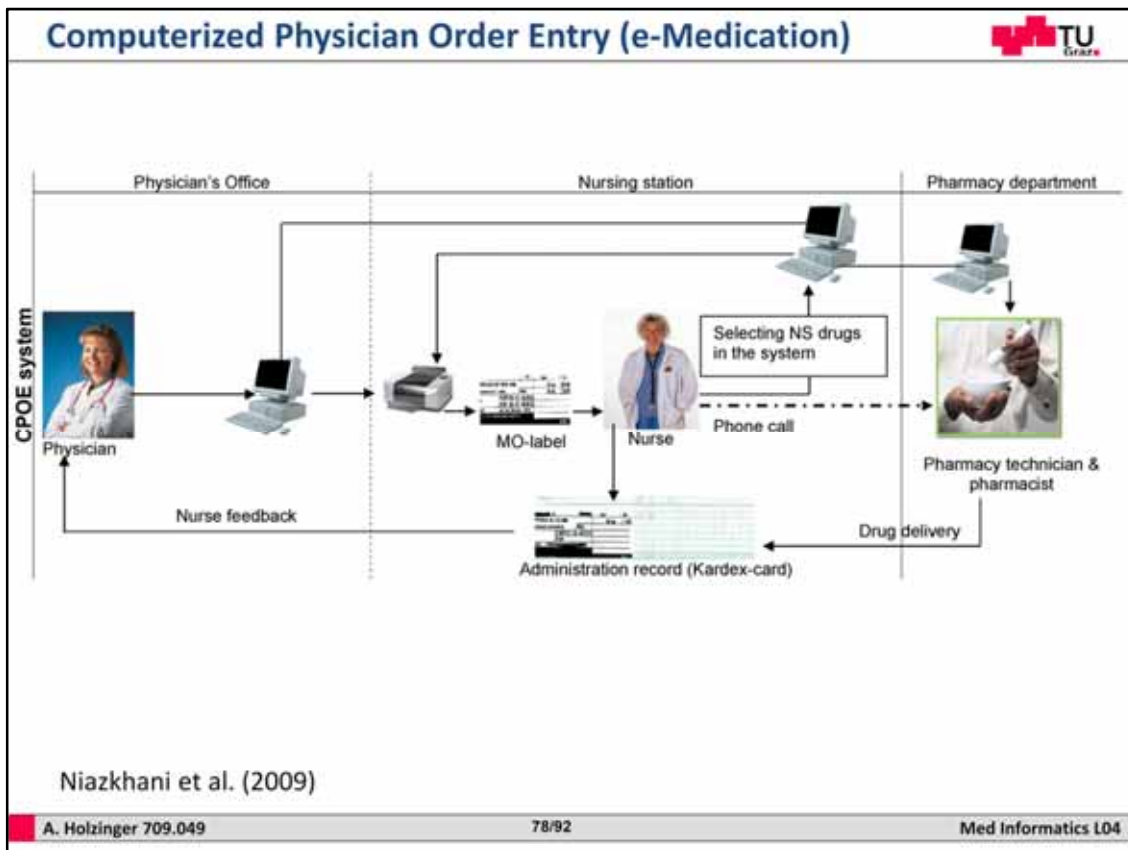
Dr. Info	Regelung	Prescription	Test	Test
05.11	Tramadol caps	50 mg	0	
	Benzhexon	50 mg	14	
	3	50 mg	22	
	9M04			
H	W	Test		
A	B	1233J06		

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.

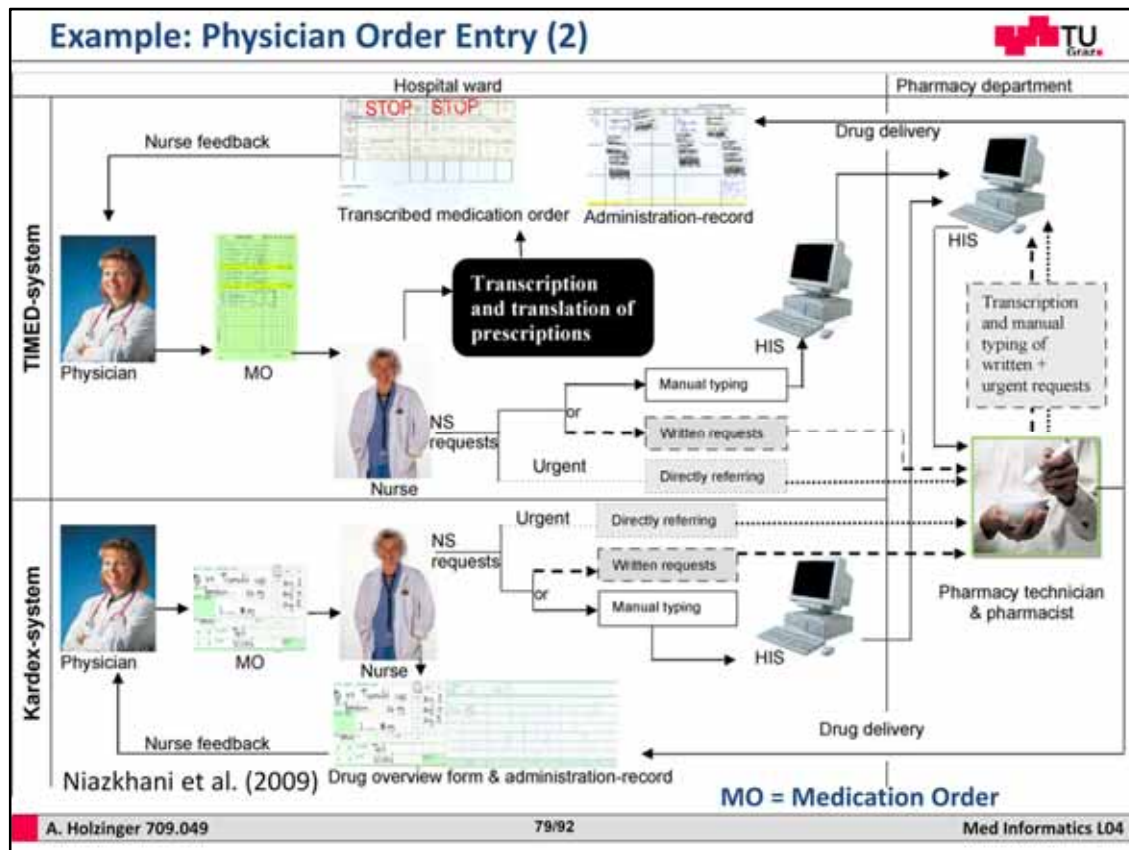


Adverse Drug Events (ADE) are very common and therefore the order entry must be taken special care of. The medication orders in different medication systems. (a) Kardex system, (b) TIMED system, and (c) CPOE system.

Physicians must enter their medication orders into the system; nurses may not accept any hand-written prescription. A physician enters a medication order by selecting a drug and its dosage form, strength, administration route, dosage regimen, start date and time.



Comparison showed that the medication ordering and administration process after the implementation resembles that of the Kardex-system, while it is completely different from that of the TIMED-system. In both Kardex and TIMED units, we compared nurse attitudes towards the computerized process in the post-implementation phase with their attitudes towards the paper-based process in the preimplementation phase.



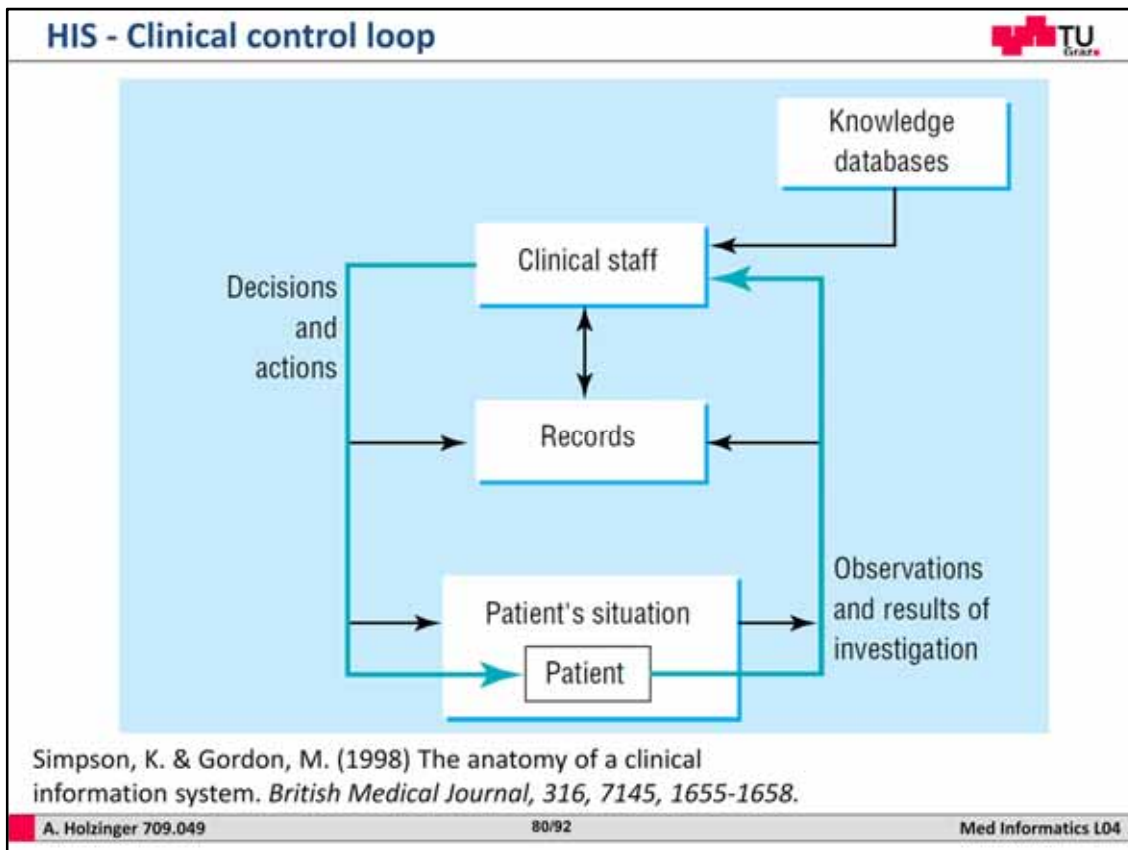
NO=No Stock

The medication ordering and administration processes in Kardex-system and TIMED-system; MO (Medication Order); HIS (Hospital Information System); NS (Non-Stock); for requesting urgent NS drugs, nurses often directly referred to the pharmacy with hand-written requests.

Comparison of Figs. 2 and 3 shows that the medication ordering and administration process after the implementation resembles that of the Kardex-system, while it is completely different from that of the TIMED-system. In both Kardex and TIMED units, we compared nurse attitudes towards the computerized process in the post-implementation phase with their attitudes towards the paper-based process in the preimplementation phase.

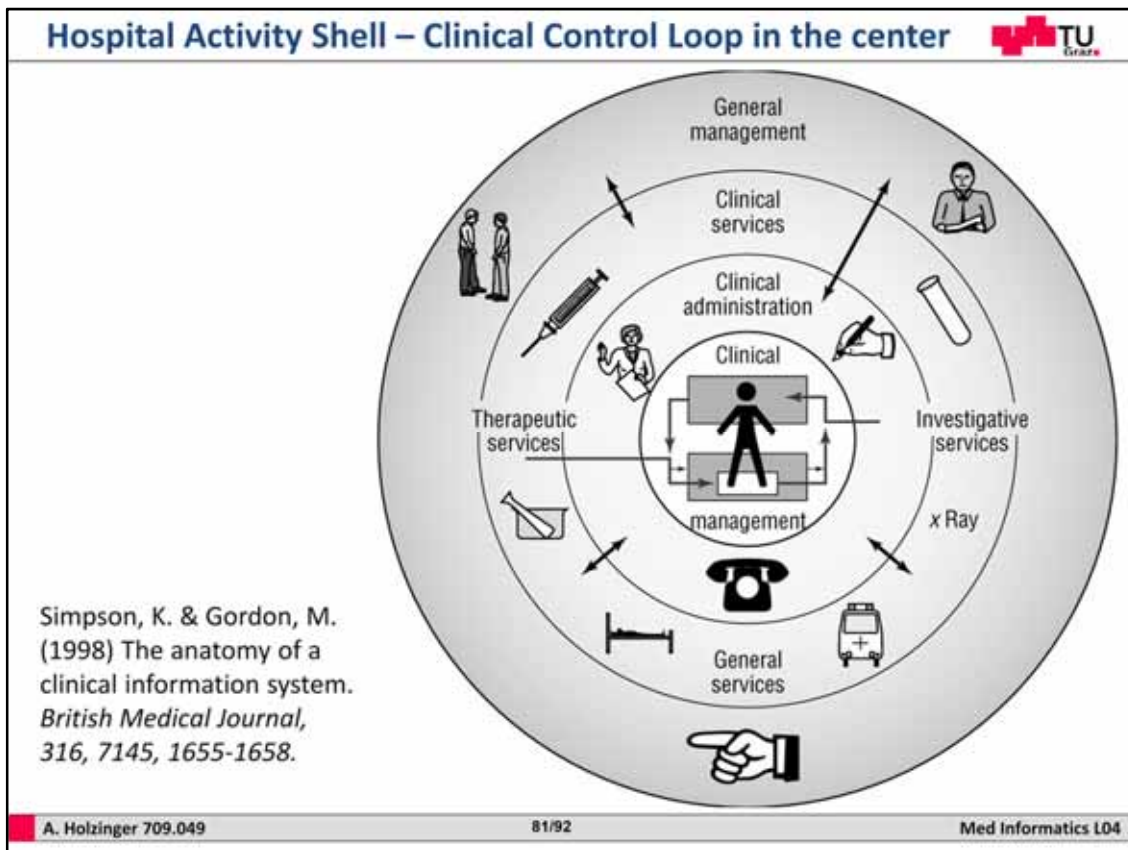
There is no clear definition about this, but it is definitely about management of data, information and knowledge for decision support.

Let us look into a practical example – physician order – where a lot of errors happened in the past due to a mess of paper based orders producing a lot of paper chaos (you all know the post-it syndrome)



Observations and results of investigations—including history, signs, and symptoms—are converted by clinical staff into decisions and appropriate actions. Control usually requires the use of records and external sources of knowledge

The care of each patient can be considered to be a control loop in which data from observations and investigations lead to decisions and actions designed to take care of a patient's problems and their consequences in a safe, effective, and legitimate manner. This loop occurs in all specialties and is the source of all the activities of a healthcare facility such as a hospital. Though complex, these activities can be set out as four concentric shells.



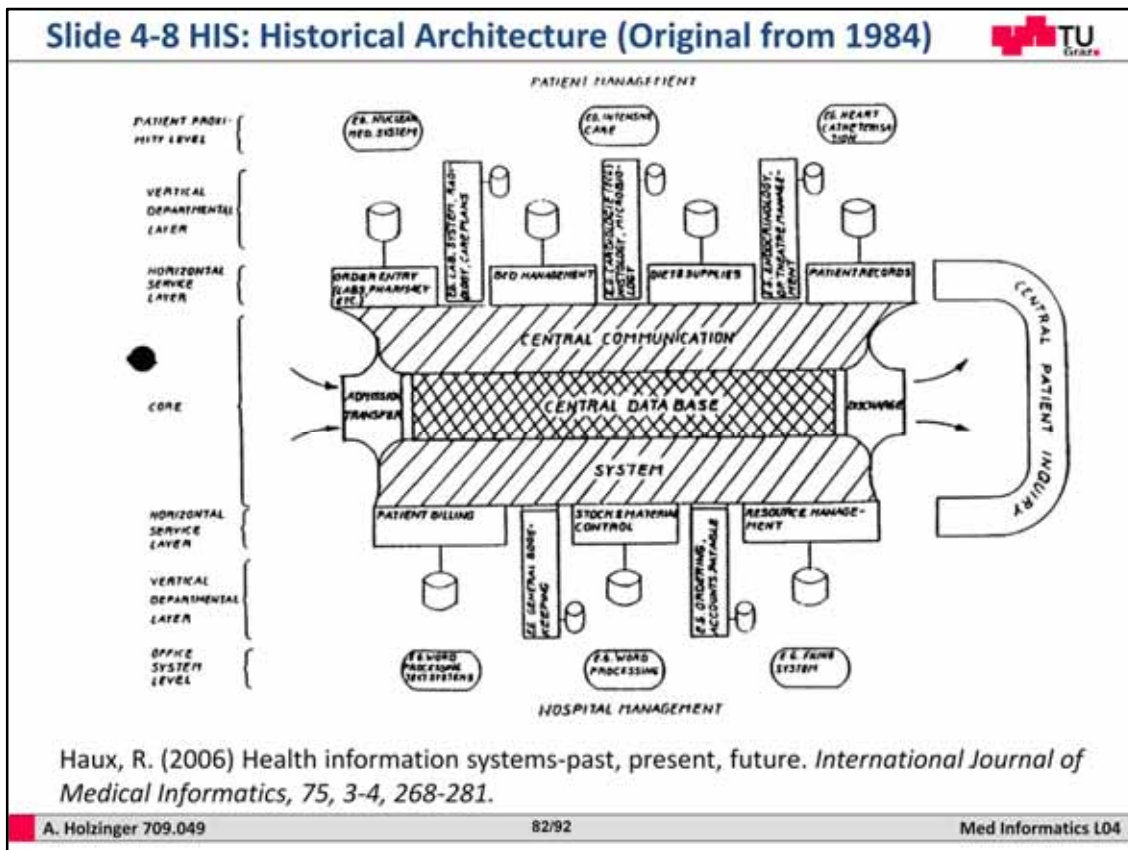
The clinical control loop is at the core of a complex organisation represented by four “shells” that exchange data. Activity shells of clinical control loop

Clinical management shell—Assessment of observations and results of investigations. Formulations of decisions including those based on observations, investigations, and procedures carried out during a consultation

Clinical administrative shell—Administrative activities which facilitate the clinical management shell and link it to the other shells, such as arranging appointments and investigations, clinical correspondence, filing results, and clinical audit


Clinical services shell—Investigative, therapeutic, and general services provided by laboratories, imaging facilities, therapy units, operating theatres, wards, supplies departments, transport, etc

General management shell—General management of health care, by hospital managers, financial controllers, healthcare purchasers, and statutory authorities




Example of a visualized information system architecture, here of the computer-supported part of the hospital information system of the Medical School Hanover from 1984 ([1], p. 9).

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.

A. Holzinger 709.049

83/92

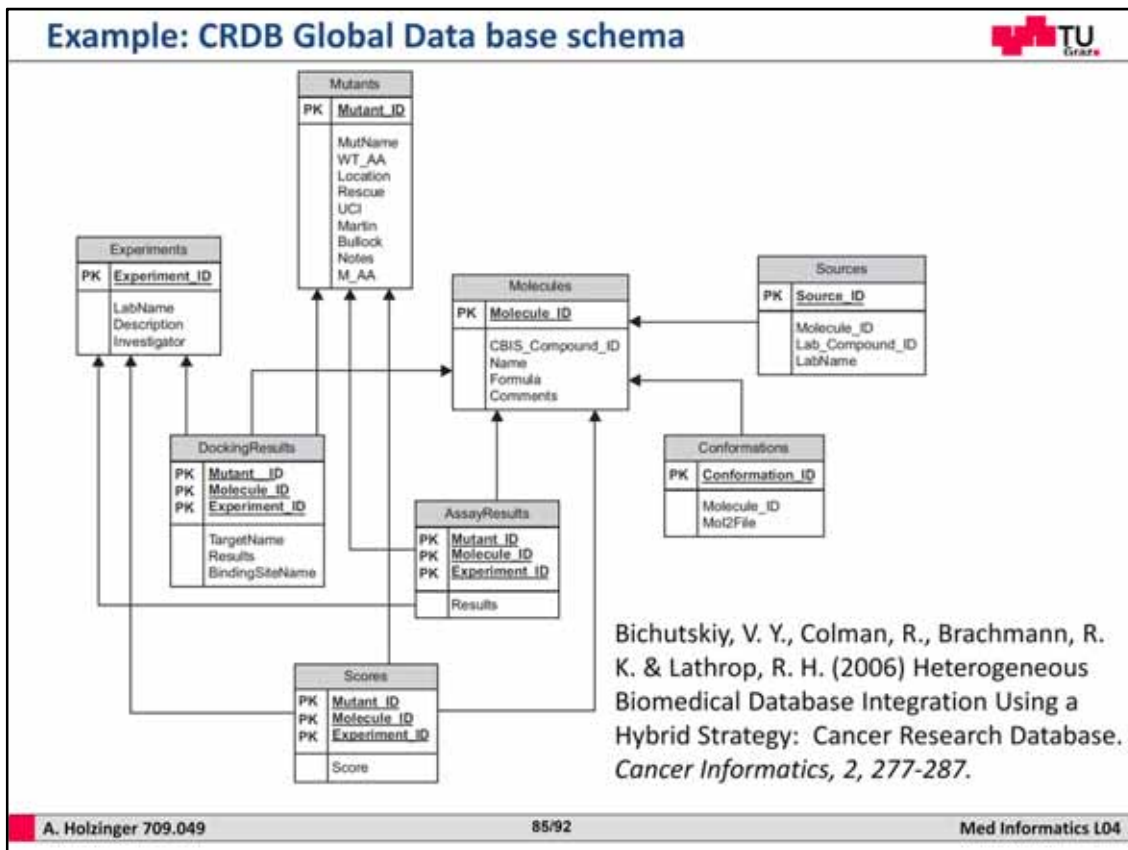
Med Informatics L04

Mayo's Enterprise Data Modeling (EDM) provides a context for Mayo enterprise activities.

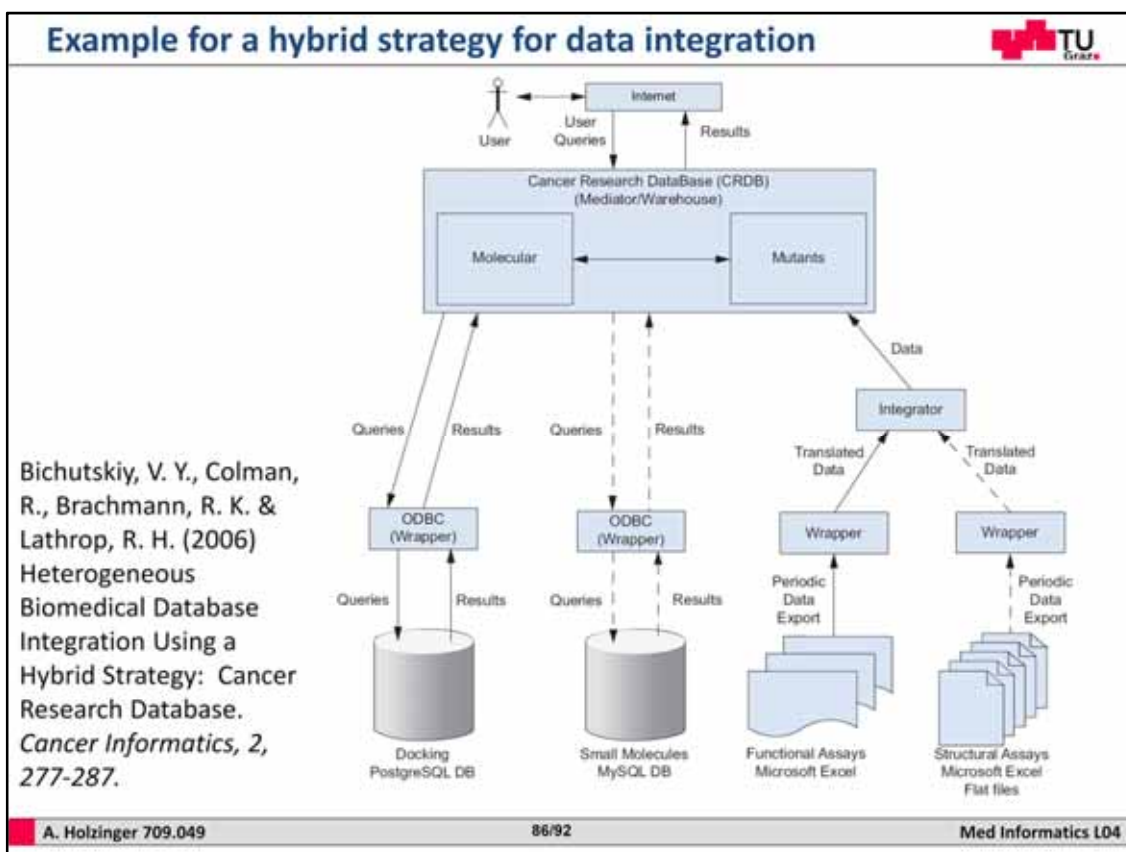
Backup: For your own experiments: www.care2x.org

The screenshot displays the Care2x web-based HIS interface. The main window is titled 'Admission Data (2003800000)'. It features a sidebar on the left with navigation links: Home, Persons, Appointments, Admissions, Ambulatory, Medicals, Dietetics, Radiology, Pharmacy, Medical Dept, Directory, Tech Support, EDP, Internet Email, Special Tools, and Login. The main content area shows patient details for 'Makaria', including admission date (06/10/2003), admission time (12:26:43), family name, given name, date of birth (05/06/2003), sex (male), address, admission class (Outpatient), clinic/department (Cytology), diagnosis (R100.10), referred by (R), therapy (R), patient notes (rm), billing type (Health Fund), insurance AC (678), insurance company (Advance Bank), and admitted by (Eduardo Llanillo). A photo of the patient is shown. On the right, there is a list of actions for this patient: Confirmation of inability to work, Charts folder, Diagnostic Results, Meds, DRG (composite), Prescriptions, Notes & Reports, Immunization, Measurements, Pregnancies, Birth details, Show Person registration, Update Person registration, DR Record's History, and Cancel this admission. At the bottom of the main content area, there are buttons for 'Update Data', 'Barcode labels', 'Make wristbands', and 'Close'. The footer of the screenshot shows 'A. Holzinger 709.049', '84/92', and 'Med Informatics L04'.

Care2x1 is a generic multi-language open-source project that implements a modern Hospital Information System. The project was started in May 2002 with the release of the first beta version of Care2x by a nurse who was dissatisfied with the HIS in the hospital where he was working. Until today the development team has grown to over 100 members from over 20 countries. Care2x is a web-based HIS that is built upon other open-source projects: the Apache web server from the Apache Foundation (<http://www.apache.org/>), the script language PHP (<http://www.php.org/>) and the relational database management system MySQL (<http://www.mysql.com/>). There exist several source code branches that try to integrate the option to choose from other RDBMS like Oracle and PostgreSQL. The latter one is already supported in the current version at the time of writing: "deployment 2.1". For our investigations we have chosen the most feature-rich version that was available from the Care2x webpage in early fall 2004. This release had the version number "pre-deployment 2.0.2". Some minor deficiencies that we report later may already be fixed in the current version "deployment 2.1".



This is just to show you an example of a global database schema. Each molecule ("Molecules" table) may have more than one conformation ("Conformations" table) and it may come from more than one source ("Sources" table). There are two types of experiments ("Experiments" table) that are done on molecules: computational docking and biological assays. The results ("DockingResults" and "AssayResults" tables) of these experiments were captured in the database. Each type of experiment is done on a particular p53 mutant ("Mutants" table) and has a score ("Scores" table) associated with it.



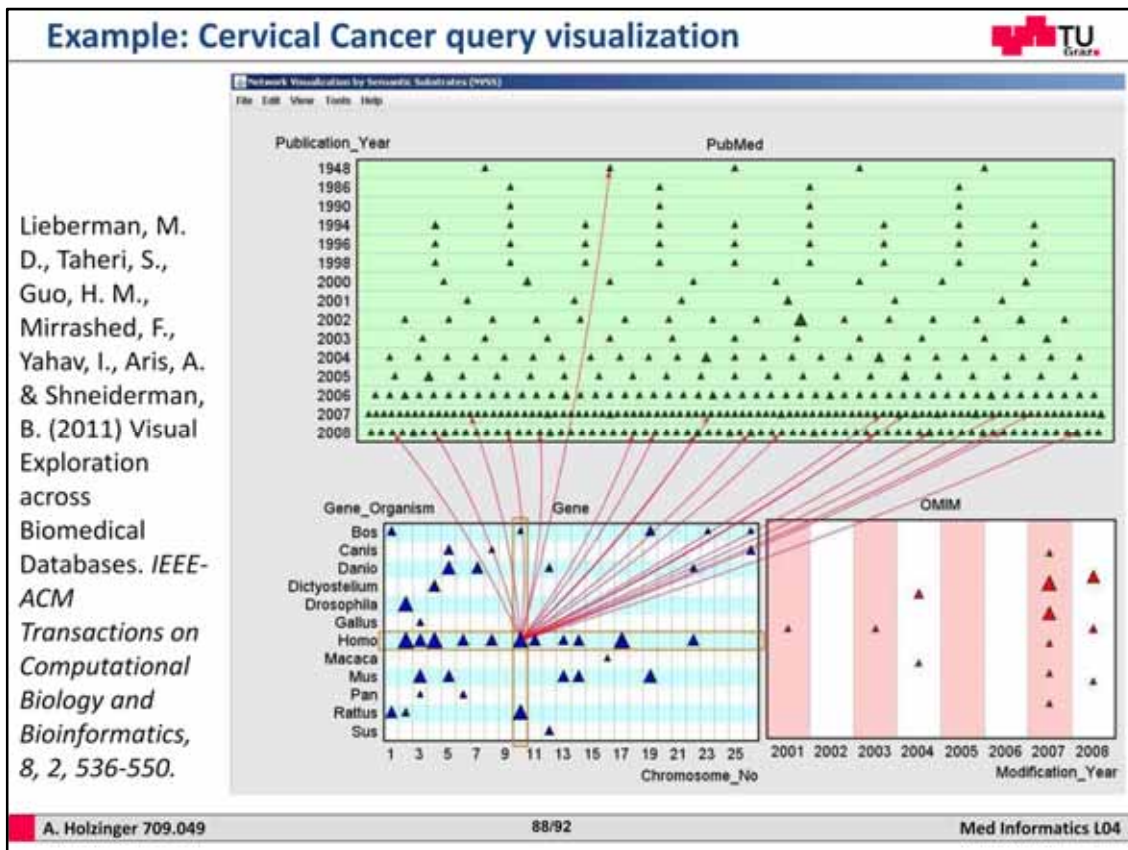
Open Database Connectivity – API in C for accessing DBMS

System architecture and the hybrid strategy to data integration. Docking and small molecule data use the mediation approach, while the functional and structural assay data use the data warehousing approach. The CRDB is both a mediator and a data warehouse. "Mutants" and "Molecular" are data marts of the warehouse. The ODBC drivers are wrappers in the mediation approach. Dashed lines indicate integration planned in the future.

Slide 4.22 Example Database: Protein Structure Data Bank

The screenshot displays the Protein Data Bank (PDB) website. At the top, it says "Slide 4.22 Example Database: Protein Structure Data Bank" and "TU Graz". The PDB logo is prominent, along with the text "An Information Portal to Biological Macromolecular Structures". A search bar is located below the header, with a placeholder text "e.g. PDB ID, molecule name, author". The main content area is titled "Biological Macromolecular Resource" and includes a "Full Description" link. There are several featured sections: "Molecule of the Month" (PDB Pioneers), "Protein Structure Initiative Featured System" (The Perils of Protein Secretion), and "Explore Archive" with filters for Organism, Expression, and Polymer Type. The right sidebar contains links for "New Structures", "New Features", and "PDB40 Symposium". The footer shows the URL "http://www.pdb.org" and a slide number "87/92".

The atomic coordinates of a protein are deposited into the protein database (PDB), an international repository for 3D structure files. At the moment PDB contains more than 26.000 protein structures



We will deal with visualizations in lecture 9 – here just an appetizer what you can display

This shows a cervical cancer query visualization. The Gene nodes are positioned using both chromosome number and organism name. This positioning method allows users to focus on a particular gene and species using NVSS's slider filters. Nodes are size-coded according to their indegree, which provides an additional visual cue about the node's importance.

Example Patent of Scoring Documents in a linked database

US006799176B1

START

101

SELECT AN INITIAL N-DIMENSIONAL VECTOR p_0

102

COMPUTE AN APPROXIMATION p_k TO A STEADY-STATE PROBABILITY p_k IN ACCORDANCE WITH THE EQUATION $p_k = A^k p_0$

105

DETERMINE A RANK $r(k)$ FOR NODE k FROM A k^{th} COMPONENT OF p_k

DONE

(12) **United States Patent**
Page

(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. 30, 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**

(10) Patent No.: **US 6,799,176 B1**

(45) Date of Patent: ***Sep. 28, 2004**

6,914,678 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	707/8
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	Kosle	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, Psychometrika, vol. 18, pp. 39-43.

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

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

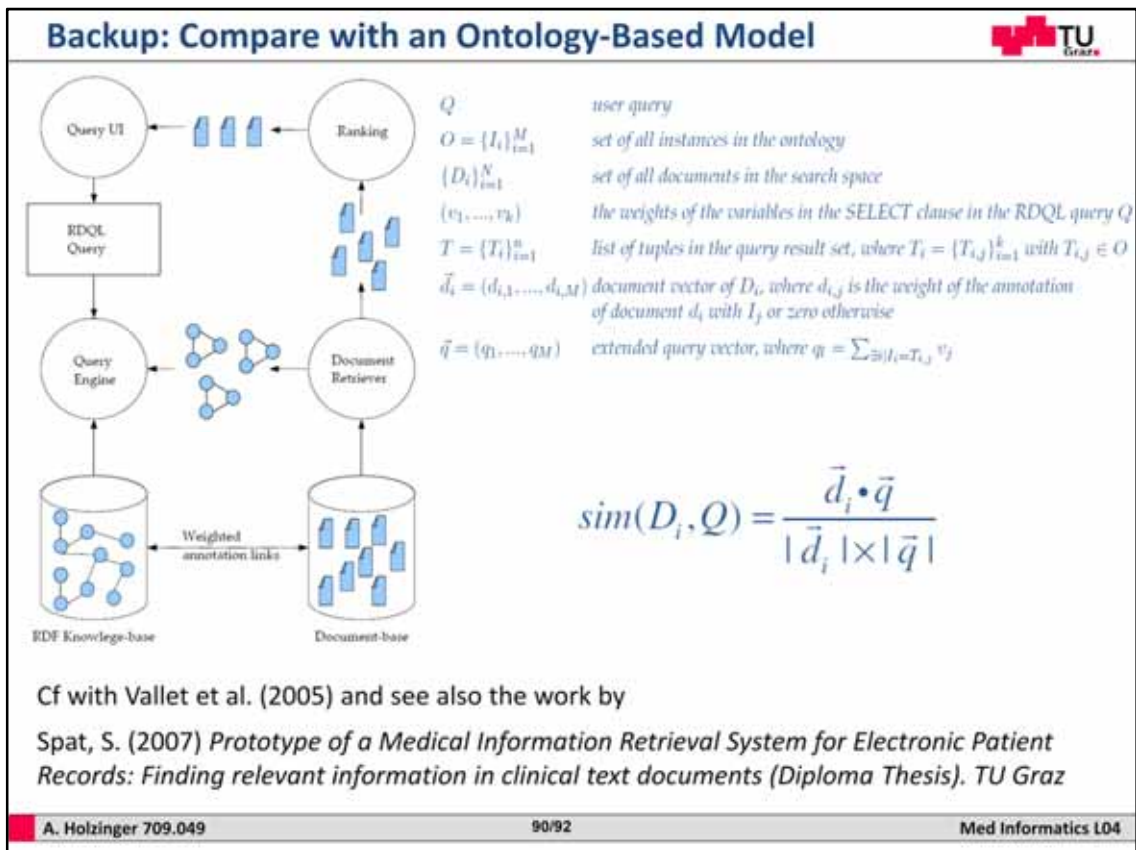
E. Garfield, "Citation analysis as a tool in journal evaluation", 1972, Science, vol. 178, pp. 471-479.


(List continued on next page.)

A. Holzinger 709.049


89/92

Med Informatics L04



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	
A. Holzinger 709.049		91/92 Med Informatics L04

Some Useful Links



- <http://www.library.tufts.edu/hsl/resources/databases.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>

A. Holzinger 709.049

92/92

Med Informatics L04

http://psychology.wikia.com/wiki/Information_retrieval
<http://www.eecs.wsu.edu/mgd/gdb.html>
(Graph Datasets)