MAKE Decisions: Medical Information Science for Decision Support

WU EXECUTIVE ACADEMY

EFMD EQUIS ACCREDITED

Assoc. Prof. Dr. Andreas HOLZINGER (Med. Uni Graz)

https://hci-kdd.org/mini-course-make-decisions-practice

Day 1 –Part 2 -19.09.2018

Data, Information and Knowledge

Day 1 - Hot Ideas	Day 2 - Cool Practice
01 Information Sciences meets Life Sciences	05 Methods of Explainable-AI
02 Data, Information and Knowledge	Groupwork: Planning of a 500 bed Hospital - Bringing AI into the workflows
03 Decision Making and Decision Support	Plenary: Presenting of the developed concepts
04 DSS: from Expert Systems to explainable AI	

Keywords

- Data
- Information
- Knowledge
- Dimensionality of data
- Biomedical Ontologies
- Standardized Medical Data
- SNOMED
- UMLS

Learning Goals

- … be aware of the types and categories of different data sets in biomedical informatics;
- … know some differences between data, information, and knowledge;
- … be aware of standardized/non-standardized and well-structured/"un-structured" information/data;
- … have a basic overview on some ontological approaches for standardized medicine;
- … have some background on classifications

Advance Organizer (1/2)

- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) for the abnormal state of the patient at hand;
- **Abstraction** = data are filtered according to their relevance for the problem solution and chunked in schemas representing an abstract description of the problem (e.g., abstracting that an adult male with haemoglobin concentration less than 14g/dL is an anaemic patient);
- **Artefact/surrogate** = error or anomaly in the perception or representation of information trough the involved method, equipment or process;
- **Data** = physical entities at the lowest abstraction level which are, e.g. generated by a patient (patient data) or a (biological) process; data contain no meaning;
- **Data quality** = Includes quality parameter such as : Accuracy, Completeness, Update status, Relevance, Consistency, Reliability, Accessibility;
- **Data structure** = way of storing and organizing data to use it efficiently;
- **Deduction** = deriving a particular valid conclusion from a set of general premises;
- **DIK-Model** = Data-Information-Knowledge three level model
- **Disparity** = containing different types of information in different dimensions
- **Heart rate variability (HRV) =** measured by the variation in the beat-to-beat interval;
- **HRV artifact** = noise through errors in the location of the instantaneous heart beat, resulting in errors in the calculation of the HRV, which is highly sensitive to artifact and errors in as low as 2% of the data will result in unwanted biases in HRV calculations;

Advance Organizer (2/2)

- **Induction** = deriving a likely general conclusion from a set of particular statements;
- **Information** = derived from the data by interpretation (with feedback to the clinician);
- **Information Entropy =** a measure for uncertainty: highly structured data contain low entropy, if everything is in order there is no uncertainty, no surprise, ideally H = 0
- **Knowledge** = obtained by inductive reasoning with previously interpreted data, collected from many similar patients or processes, which is added to the "body of knowledge" (explicit knowledge). This knowledge is used for the interpretation of other data and to gain implicit knowledge which guides the clinician in taking further action;
- **Large Data** = consist of at least hundreds of thousands of data points
- **Multi-Dimensionality** = containing more than three dimensions and data are multi-variate
- **Multi-Modality** = a combination of data from different sources
- **Multivariate** = encompassing the simultaneous observation and analysis of more than one statistical variable;
- **Reasoning** = process by which clinicians reach a conclusion after thinking on all facts;
- **Spatiality** = contains at least one (non-scalar) spatial component and non-spatial data
- **Structural Complexity** = ranging from low-structured (simple data structure, but many instances, e.g., flow data, volume data) to high-structured data (complex data structure, but only a few instances, e.g., business data)
- **Time-Dependency** = data is given at several points in time (time series data)
- **Voxel** = volumetric pixel = volumetric picture element

Agenda

- **00 Reflection – follow-up from last lecture**
- **01 What is data?**
- **02 On Standardization**
- **03 Knowledge Representation**
- **04 Biomedical Ontologies**
- **05 Medical Classifications**

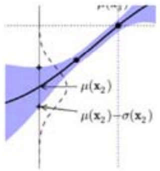
Reflection from last lecture

1

2

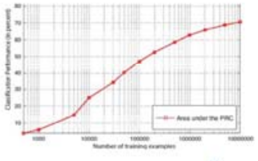
$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta) * p(\theta)}{p(\mathcal{D})}$$

3

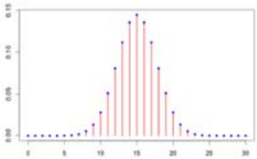


4

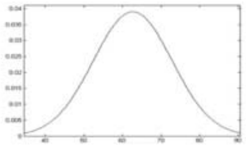
5



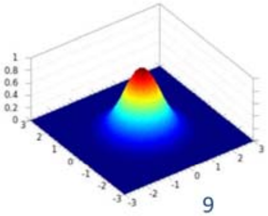
6



7

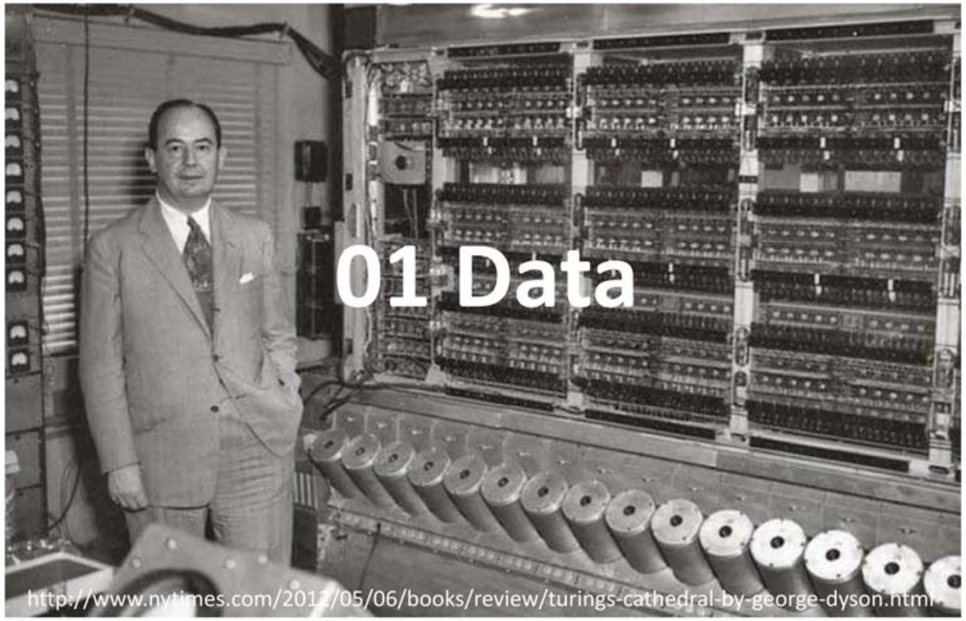


8



9

Institute for Advanced Study, Princeton University



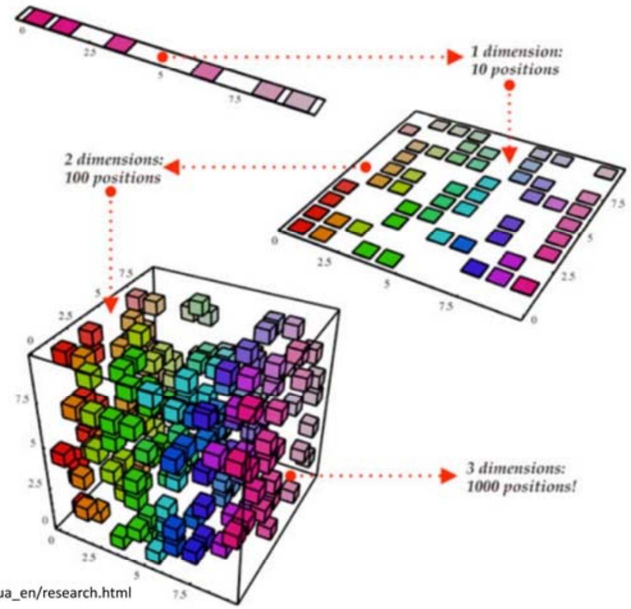
Traditional Statistics versus Machine Learning

- Data in traditional Statistics
- Low-dimensional data ($< \mathbb{R}^{100}$)
- Problem: Much noise in the data
- Not much structure in the data but it can be represented by a simple model

- Data in Machine Learning
- High-dimensional data ($\gg \mathbb{R}^{100}$)
- Problem: not noise , but complexity
- Much structure, but the structure can **not** be represented by a simple model

Lecun, Y., Bengio, Y. & Hinton, G. 2015. Deep learning. Nature, 521, (7553), 436-444.

Note: The curse of dimensionality

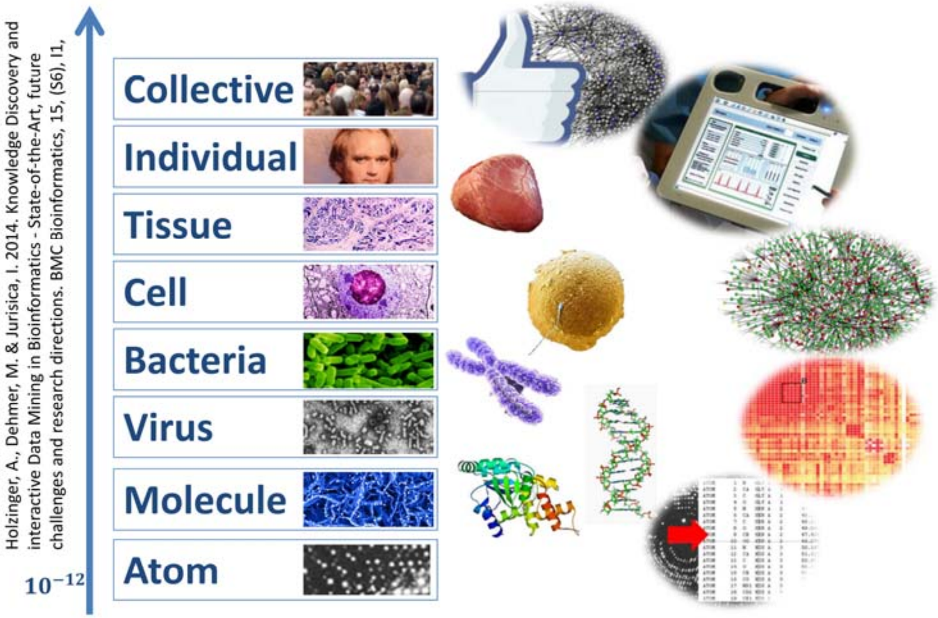


Bengio, S. & Bengio, Y. 2000. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks, 11, (3), 550-557.

http://www.iro.umontreal.ca/~bengioy/yoshua_en/research.html

Biomedical Data Sources

Holzinger, A., Dehmer, M. & Jurisica, I. 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics, 15, (S6), I1,

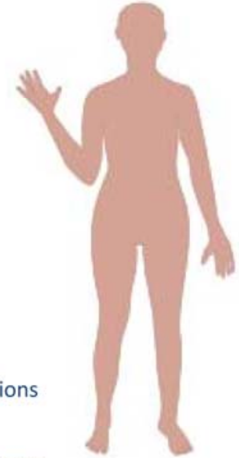


Data for clinical purposes – integration is unsolved!

Private Health vault data
Electronic health record data
Physiological data
Laboratory results

Exposome
Environmental data
Air pollution
Exposure (toxicants)

Collective data
Social data
Fitness, Wellness data
Ambient Assisted Living data
(Non-medical) personal data



Metabolomics
Chemical processes
Cellular reactions
Enzymatic reactions

Foodomics, Lipidomics
Nutrition data (Nutrigenomics)
Diet data (allergenics)

Metabolomics
Chemical processes
Cellular reactions
Enzymatic reactions

Imaging data
X-Ray, ultrasound, MR, CT, PET, cams, observation (e.g. sleep laboratory), gait (child walking)

Proteomics
Protein-Protein Interactions

Epigenetics
Epigenetic modifications

Genomics

Transcriptomics
RNA, mRNA, rRNA, tRNA

Taxonomy of data

- **Physical level** -> bit = binary digit = **b**asic **i**ndissoluble uni**t** (= Shannon, Sh), ≠ Bit (!) in Quantum Systems -> qubit
- **Logical Level** -> integers, booleans, characters, floating-point numbers, alphanumeric strings, …
- **Conceptual (Abstract) Level** -> data-structures, e.g. lists, arrays, trees, graphs, …
- **Technical Level** -> Application data, e.g. text, graphics, images, audio, video, multimedia, …
- **"Hospital Level"** -> Narrative (textual) data, genetic data, numerical measurements (physiological data, lab results, vital signs, …), recorded signals (ECG, EEG, …), Images (cams, x-ray, MR, CT, PET, …)

Examples: Imaging Data

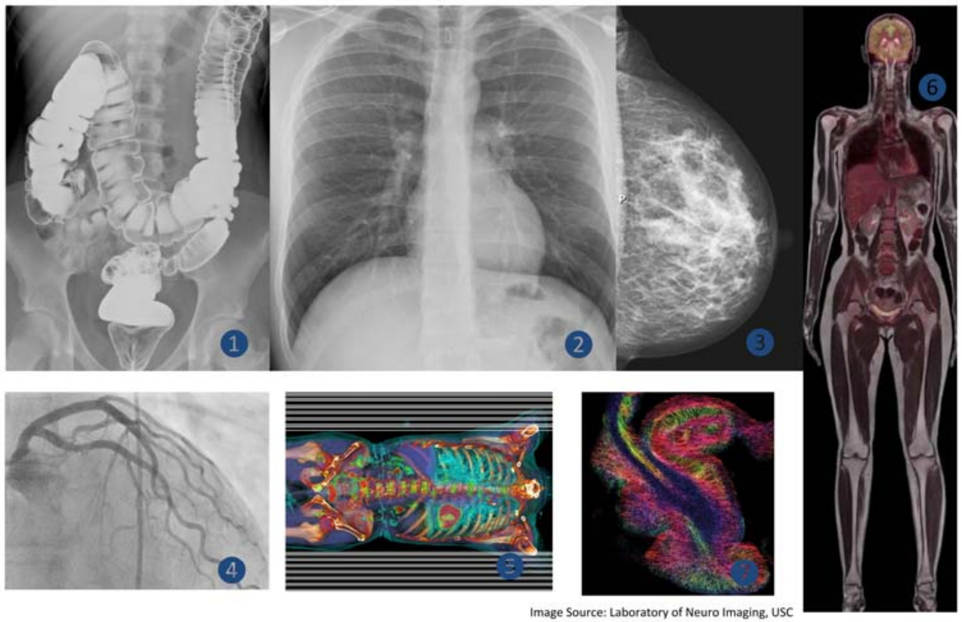
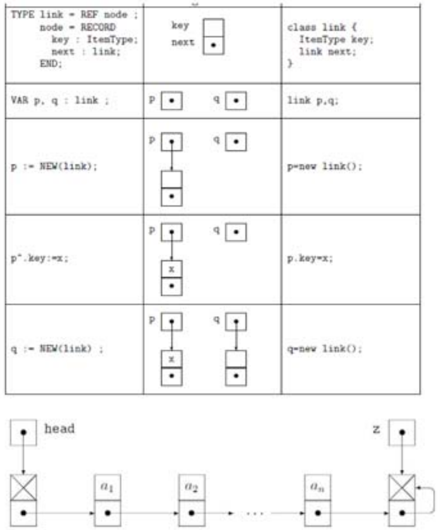
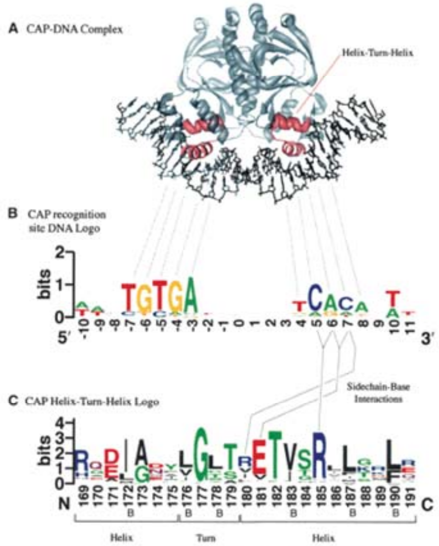


Image Source: Laboratory of Neuro Imaging, USC

Example Data Structures (1/3): List

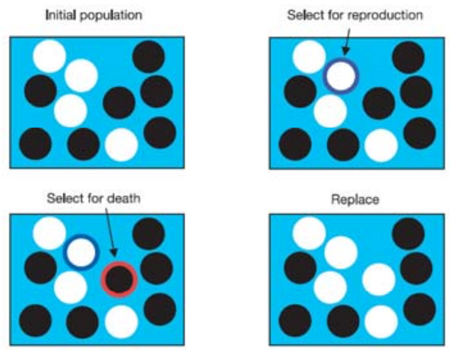


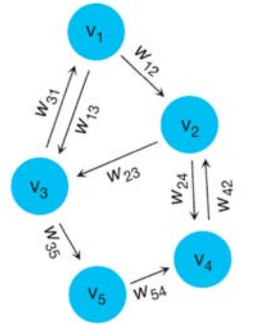


Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

Example Data Structures (2/3): Graph

Evolutionary dynamics act on populations.
Neither genes, nor cells, nor individuals evolve;
only populations evolve.



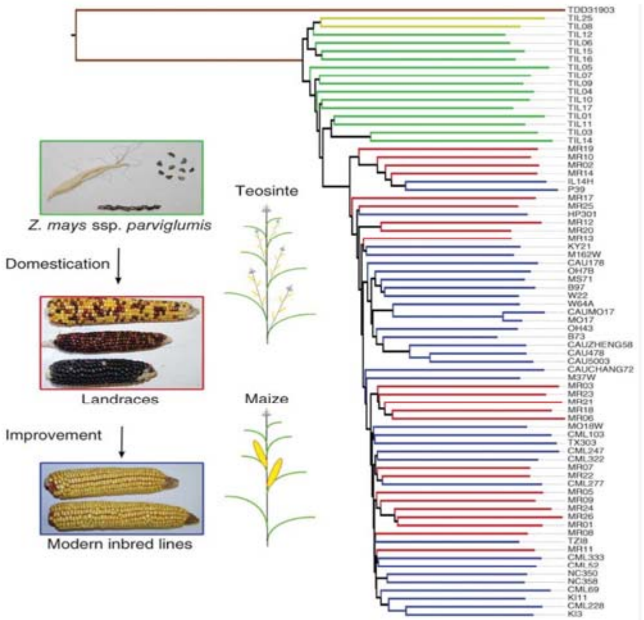


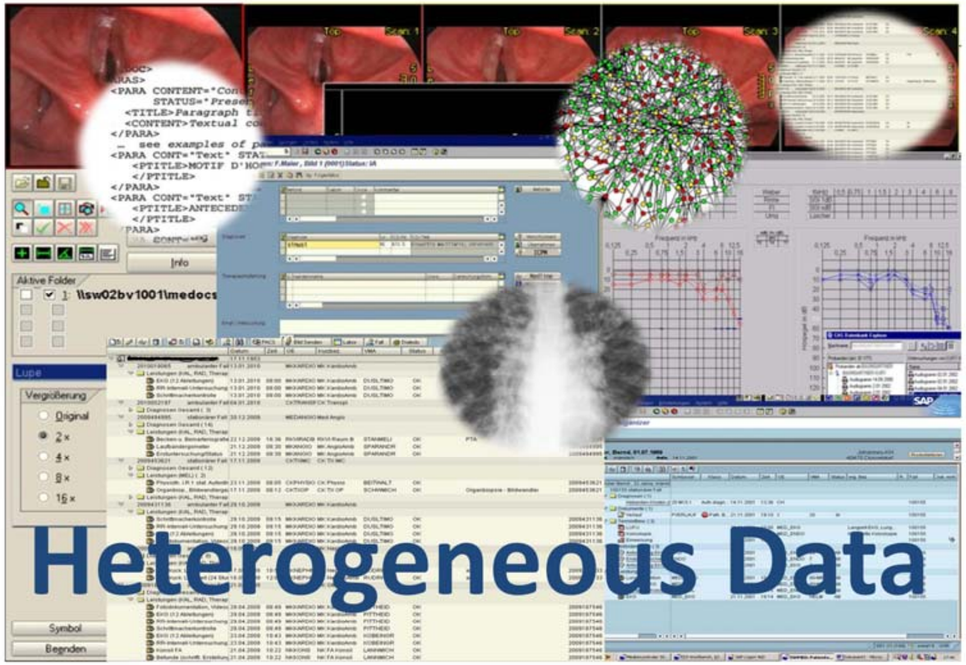
$$W=\begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ 0 & 0 & w_{23} & w_{24} & 0 \\ w_{31} & 0 & 0 & 0 & w_{35} \\ 0 & w_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}$$

Lieberman, E., Hauert, C. & Nowak, M. A. (2005) Evolutionary dynamics on graphs. *Nature, 433, 7023, 312-316.*

Example Data Structures (3/3) Tree

Hufford et. al. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics, 44, (7), 808-811.*





Biomedical R&D data
(e.g. clinical trial data)

Clinical patient data
(e.g. EPR, lab, reports etc.)

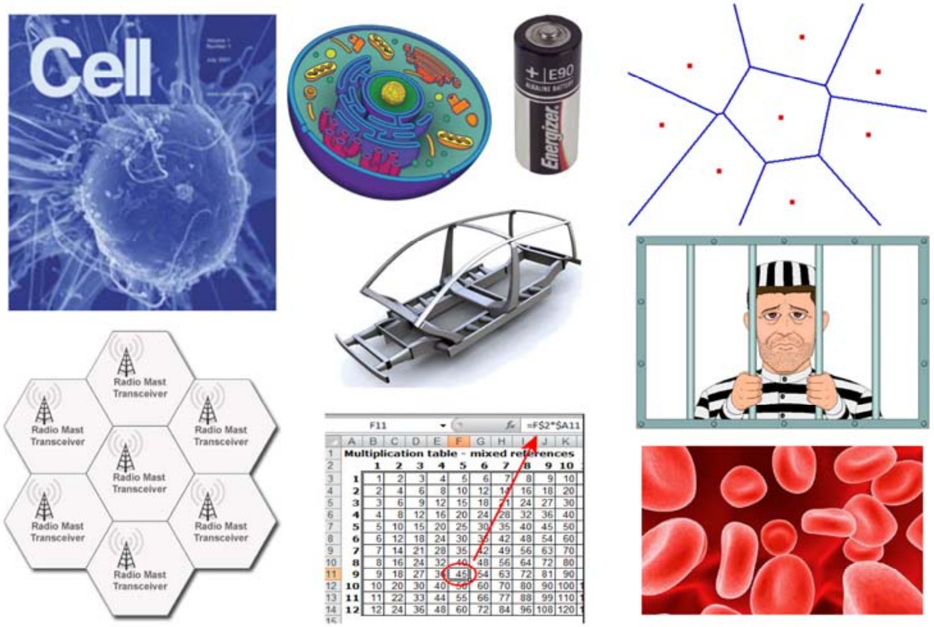
The combining link is text

Health business data
(e.g. costs, utilization, etc.)

Private patient data
(e.g. AAL, monitoring, etc.)

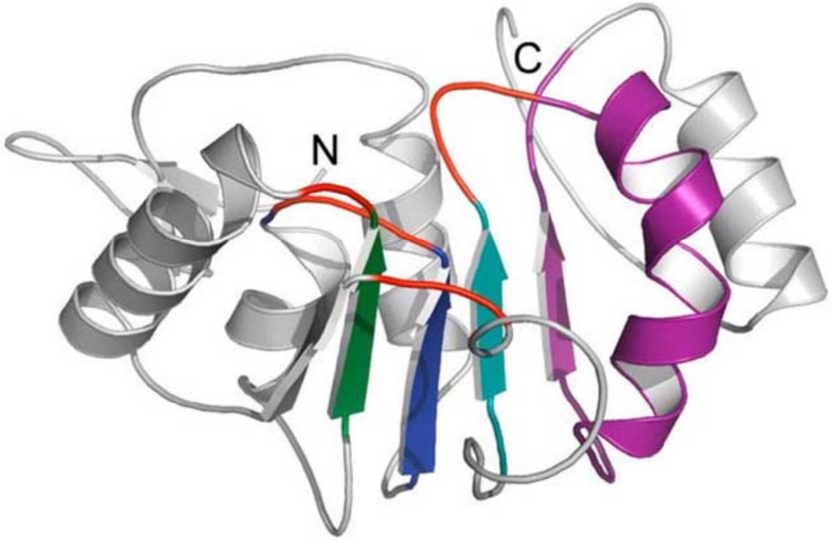
Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C. & Byers, A. H. (2011) *Big data: The next frontier for innovation, competition, and productivity. Washington (DC), McKinsey Global Institute.*

Problem: Context!



Is a picture really worth a thousand words?

Example: Ribbon Diagram of a Protein Structure



Magnani, R., et al. 2010. Calmodulin methyltransferase is an evolutionarily conserved enzyme that trimethylates Lys-115 in calmodulin. *Nature Communications, 1, 43.*

Radiologischer Befund

angelegt am 06.05.2006/20:2
geschr. von
gedruckt am 17.11.2006/08:2
Anfo: NCHI

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

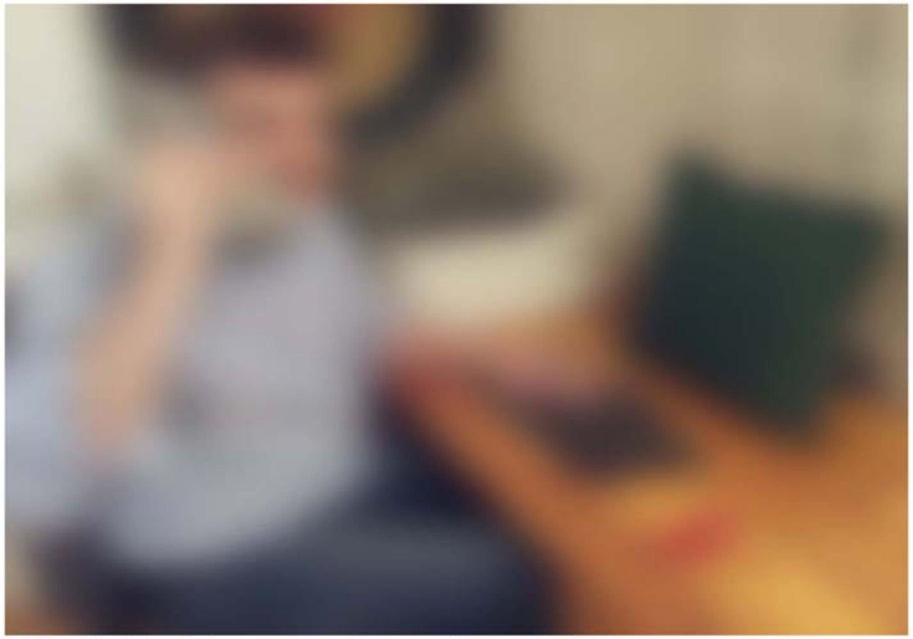
Zustand nach Anlage eines ET, die Spitze ca. 5cm cranial der Bifurkation, lieg. MS, orthot
positioniert. ZVK über re., die Spitze in Proj. auf die VCS. Kein Hinweis auf Pneumothorax
Der re. Rezessus frei.

Mit kollegialen Grüßen

*** Elektronische Freigabe durch am 09.05.2006 ***

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

A picture is worth a thousand words …



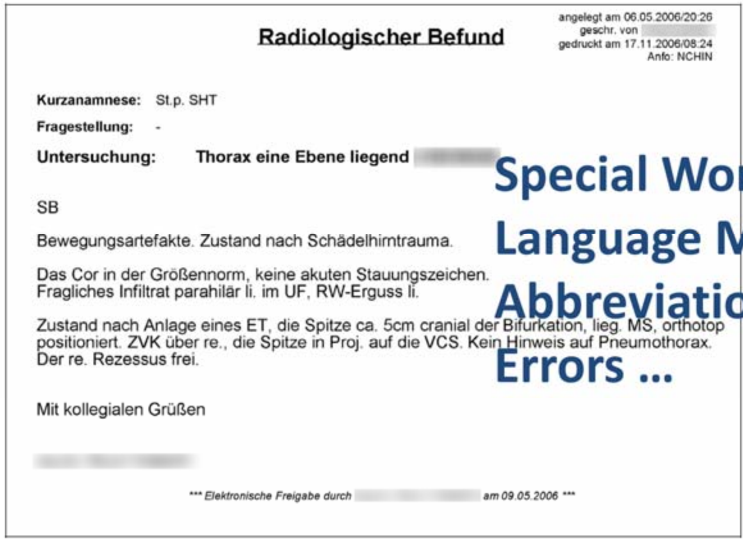
02 Standards

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.*

The medical report is the most important medium



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

Radiologischer Befund

Kurzanamnese: St.p. SHT

Fragestellung: -

Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

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

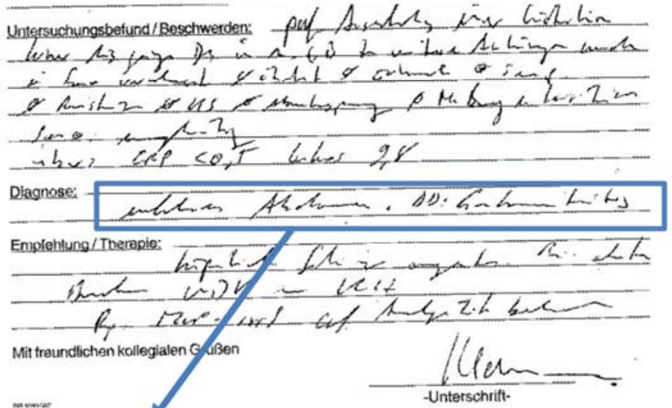
Mit kollegialen Grüßen

*** Elektronische Freigabe durch am 09.05.2006 ***

Special Words
Language Mix
Abbreviations
Errors …

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

German Example: Synonymity and Ambiguity



„die Antrumschleimhaut ist durch Lymphozyten infiltriert“

„lymphozytäre Infiltration der Antrummukosa“

„Lymphoyteninfiltration der Magenschleimhaut im Antrumbereich“

German Local Hospital Abbreviations … (example)

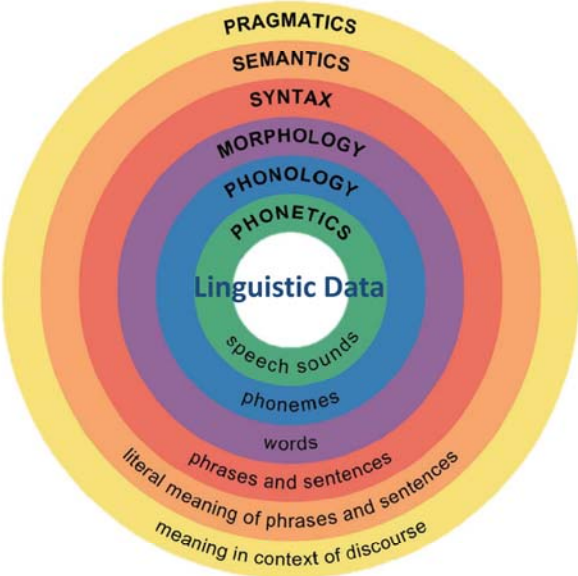
- HWI =
  - Harnwegsinfekt
  - Hinterwandinfarkt
  - Hinterwandischämie
  - Hakenwurminfektion
  - Halswirbelimmobilisation
  - Hip Waist Index
  - Height-Width Index
  - Heart-Work Index
  - Hemodynamically weighted imaging
  - High Water Intake
  - Hot water irrigation
  - Hepatitic weight index
  - Häufig wechselnder Intimpartner
- Leitung = Nervenleitung, Abteilungsleitung, Stromleitung, Wasserleitung, Harnleitung, Ableitung, Vereinsleitung ☺…

Language Understanding requires knowledge

- Syntax
- Semantics
- Pragmatics
- **Context**
- (Emotion)

Andrej Karpathy & Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. 3128-3137.

Text = Good example for Non-Standardized Data

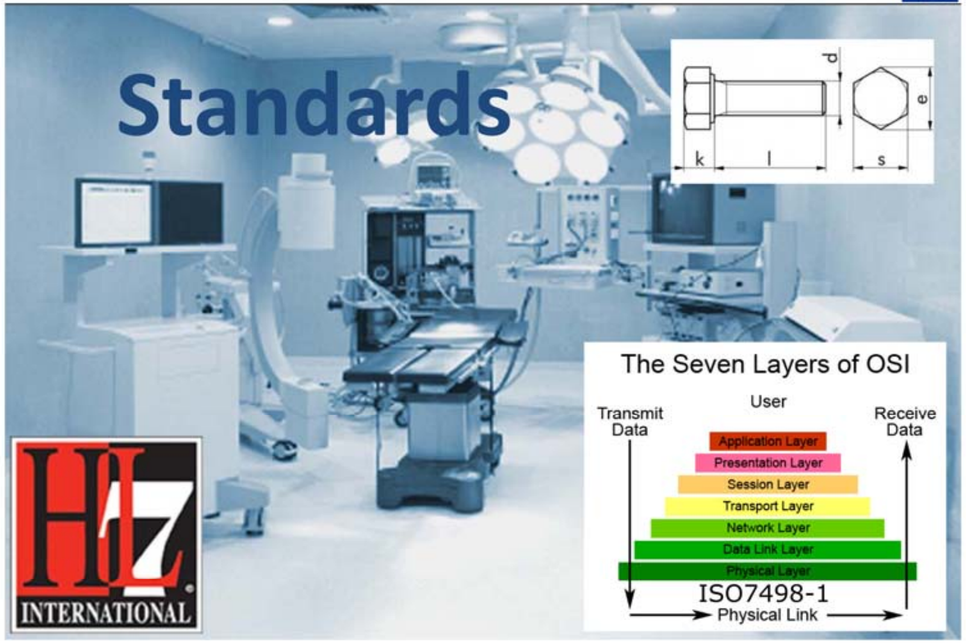


Thomas, J. J. & Cook, K. A. 2005. *Illuminating the path: The research and development agenda for visual analytics, New York, IEEE Computer Society Press.*

Key Challenges

- Increasingly large data sets due to **data-driven medicine** [1]
- Increasing amounts of **non-standardized** data 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 [2]
- **Time** aspects in databases [3]

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



Quest for standardization as old as med. informatics

IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. BME-19, NO. 5, SEPTEMBER 1972 331

Standardization and Health Care

J. H. U. BROWN, SENIOR MEMBER, IEEE, AND DEWITT JAMES LOWELL

***Abstract*—In order to deliver reasonable health care to all people, it is essential that standards be established. Standards vary with the type of control and with the approach desired in determining the quality of care. This paper discusses various kinds of standards and their application in the health care field. Standards may be determined as a process or as a direct regulation. It is probable that regulation of standards by process is the most satisfactory method.**

INTRODUCTION

SOCIETY cannot exist without a yardstick by which its accomplishments or failures are measured. Such yardsticks are called *standards*. They are created by the need for regulation and control as an escape from anarchy or to motivate towards greater achievement. In the ultimate, society dictates these limits by the demands it places upon itself. Standards provide opportunities for security and augmentation of process and output by virtue of the goal and process structure that they provide.

arbiter may be the market place or agencies that rely on expertise from many sources to set acceptable standards of quality or performance. For these reasons, the final moderator may be found in a governmental authority, and its delegation into a system of regulation, law, and judicial action, so that an established code can become the focal point of resolution.

THE OBJECTIVES OF STANDARDIZATION

Standards have value within themselves in that they help establish quality. However, they accomplish more for society than the mere establishment of a level of quality and performance. A standard allows coordination of effort between producers so that like products can be produced. It permits the reproduction of similar units in mass quantity and permits the consumer to judge one product or service against another by performance. It establishes *freedom* of *interchange* of material and ideas, and permits the activity in one part of society

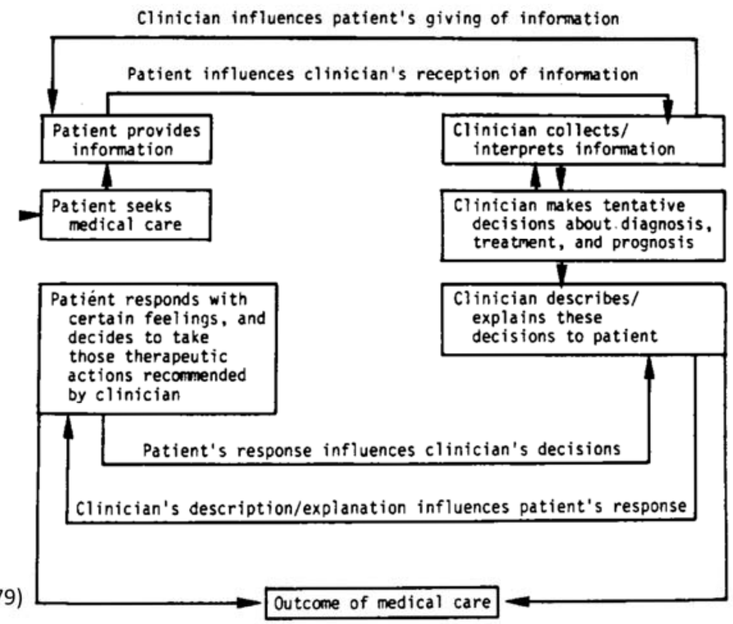
Brown, J. H. U. & Loweli, D. J. (1972) Standardization and Health Care.
IEEE Transactions on Biomedical Engineering, BME-19, 5, 331-334.

Still a big problem: Inaccuracy of medical data

- Medical (clinical) data are defined and detected disturbingly "soft" …
- … having an obvious degree of **variability** and **inaccuracy**.
- Taking a medical history, the performance of a physical examination, the interpretation of laboratory tests, even the definition of diseases … are surprisingly **inexact**.
- Data is defined, collected, and interpreted with a degree of variability and inaccuracy which falls far short of the standards **which engineers do expect from most data**.
- Moreover, standards might be **interpreted variably** by different medical doctors, different hospitals, different medical schools, different medical cultures, …

Komaroff, A. L. (1979) The variability and inaccuracy of medical data.
Proceedings of the IEEE, 67, 9, 1196-1207.

The patient-clinician dialogue (from 1979)



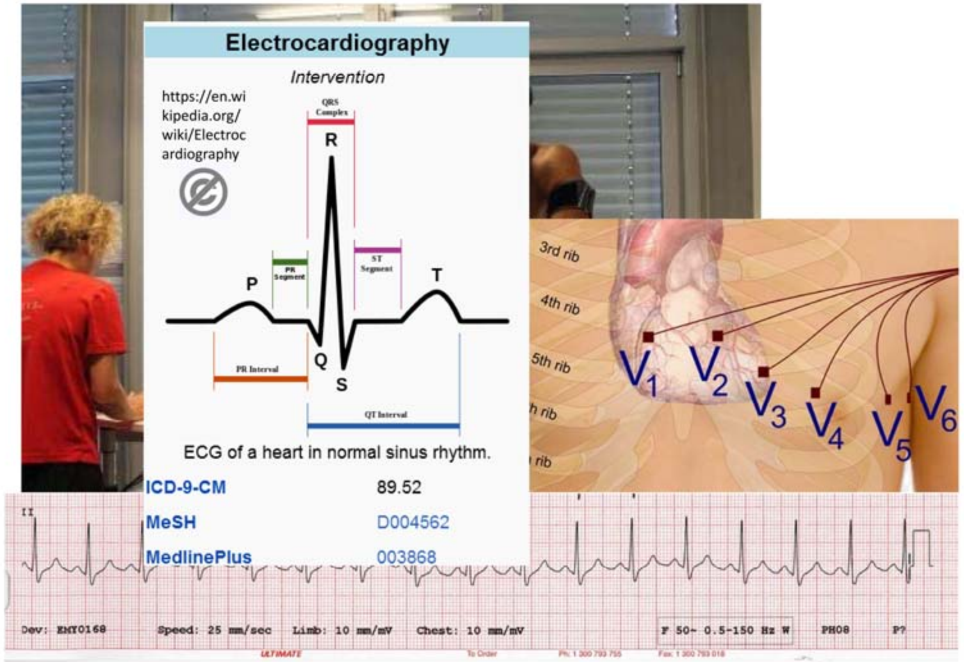
Komaroff (1979)

Standardized data …

- … ensures that information is interpreted by all users with the same understanding;
  - supports the reusability of the data,
  - improves the efficiency of healthcare services and
  - avoids errors by reducing duplicated efforts in data entry;

- Data standardization refers to
  - a) the data content;
  - b) the terminologies that are used to represent the data;
  - c) how data is exchanged; and
  - iv) how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system (refer to IOM).
- Elements for sharing require standardization of identification, record structure, terminology, messaging, privacy etc.
- The most used standardized data set to date is the **International Classification of Diseases (ICD),** which was first adopted in 1900 for collecting statistics (Ahmadian et al. 2011)

Example: ECG



Standardization of ECG data (1/2)

- There has been a large number of ECG storage formats proclaiming to promote interoperability.
- There are three predominant ECG formats:
  - SCP-ECG (1993, European Standard, Binary data)
  - DICOM-ECG (2000, European Standard, Binary data)
  - HL7 aECG (2001, ANSI Standard, XML data)
- A mass of researchers have been proposing their own ECG storage formats to be considered for implementation (= proprietary formats).
- Binary has been the predominant method for storing ECG data

Bond, R. R., Finlay, D. D., Nugent, C. D. & Moore, G. (2011) A review of ECG storage formats. *International Journal of Medical Informatics, 80, 10, 681-697.*

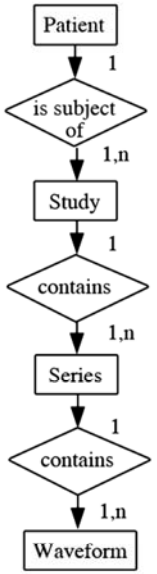
Standardization of ECG (2/2)

- Overview on current ECG storage formats

ECG format	Year	Method of implementation	Specification	Viewers
SCP-ECG	1993	BINARY	Can be freely downloaded from the Internet [7].	Freely available SCP-ECG Viewer made by EcgSoft [8].
DICOM-WS 30	2000	BINARY	Can be freely downloaded from the Internet [5].	Freely available DICOM-ECG viewer made by Charruasoft [9].
HL7 aECG	2001	XML	The XML Schema can be used as the specification or the implementation guide by AMPS [6].	Freely available aECG viewer by AMPS [10].
ecgML	2003	XML	Can be freely downloaded from the Internet [11].	None currently exist. Under development.
MFER	2003	BINARY	Can be freely downloaded from the Internet [12].	Freely available MFER viewer [13].
Philips XML	2004	XML	The specification is packaged with the actual product.	Philips viewer. Not freely available.
XML-ECG	2007	XML	Can be freely downloaded from the Internet [14].	XML-ECG viewer [14]. Not freely available.
mECGml	2008	XML	Can be freely downloaded from the Internet [15].	mECGml mobile viewer [15]. Not freely available.
ecgAware	2008	XML	Can be freely downloaded from the Internet [16].	TeleCardio viewer [16]. Not freely available.

Bond, R. R., Finlay, D. D., Nugent, C. D. & Moore, G. (2011) A review of ECG storage formats. *International Journal of Medical Informatics, 80, 10, 681-697.*

Example of a Binary ECG file




```
31 20 08 20  18 20 55 49  2a 20 31 2e  32 2e 38 32   1 . . UI* 1.2.82
36 2e 30 2e  31 2e 33 34  34 37 31 2e  32 2e 34 34   6.0.1.34471.2.44
2e 36 2e 32  30 30 32 31  31 32 32 30  39 31 30 30   .6.2002112209100
30 2e 2e 31  08 20 20 20  44 41 08 20  32 30 30 32   0..1.   DA. 2002
31 31 32 32  08 20 23 20  44 41 08 20  32 30 30 32   1122. # DA. 2002
31 31 32 32  08 20 2a 20  44 54 0e 20  32 30 30 32   1122. * DT. 2002
31 31 32 32  30 39 31 30  30 30 08 20  30 20 54 4d   1122091000. 0 TM
06 20 30 39  31 30 30 30  08 20 33 20  54 4d 06 20   . 091000. 3 TM.
30 39 31 30  30 30 08 20  50 20 53 48  20 20 08 20   091000. P SH  .
60 20 43 53  04 20 45 43  47 20 08 20  70 20 4c 4f   ` CS. ECG . p LO
08 20 55 6e  6b 6e 6f 77  6e 20 08 20  90 20 50 4e   . Unknown . □ PN
20 20 08 20  60 10 50 4e  20 20 08 20  70 10 50 4e     . `.PN  . p.PN
20 20 08 20  90 10 4c 4f  06 20 45 4c  49 32 35 30     . □.LO. ELI250
10 20 10 20  50 4e 06 20  73 6d 69 74  68 20 10 20   . . PN. smith
20 20 4c 4f  08 20 53 42  4a 2d 31 32  33 20 10 20     LO. SBJ-123 .
30 20 44 41  08 20 31 39  35 33 30 35  30 38 10 20   0 DA. 19530508.
40 20 43 53  02 20 4d 20  10 20 20 10  4c 4f 20 20   @ CS. M .  .LO
10 20 10 10  41 53 20 20  10 20 20 10  44 53 20 20   . ..AS  .  .DS
10 20 30 10  44 53 20 20  18 20 20 10  4c 4f 20 20   . 0.DS  .  .LO
```


Bond et al. (2011)

Example of a XML ECG file

```
<sequenceSet>

  <component>
    <sequence>
      <code code="TIME_ABSOLUTE" codeSystem="2.16.840.1.113883.5.4"
          codeSystemName="ActCode" displayName="Aboslute Time"/>

      <value xsi:type="GLIST_TS">

          <head value="20021122091000.000"/>

          <increment value="0.002" unit="s"/>

      </value>
    </sequence>

  </component>

  <component>
```

Bond et al. (2011)

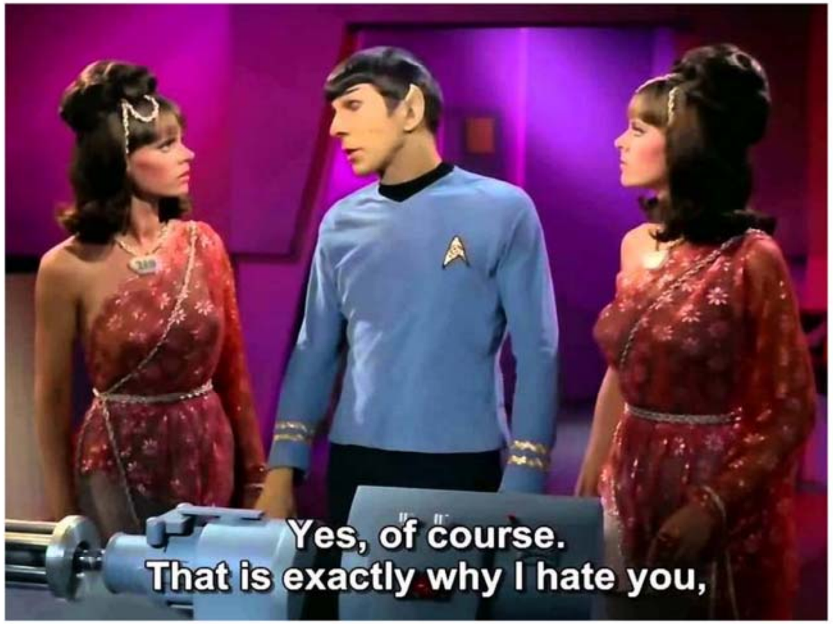
03 Knowledge Representation

Examples for famous knowledge representations

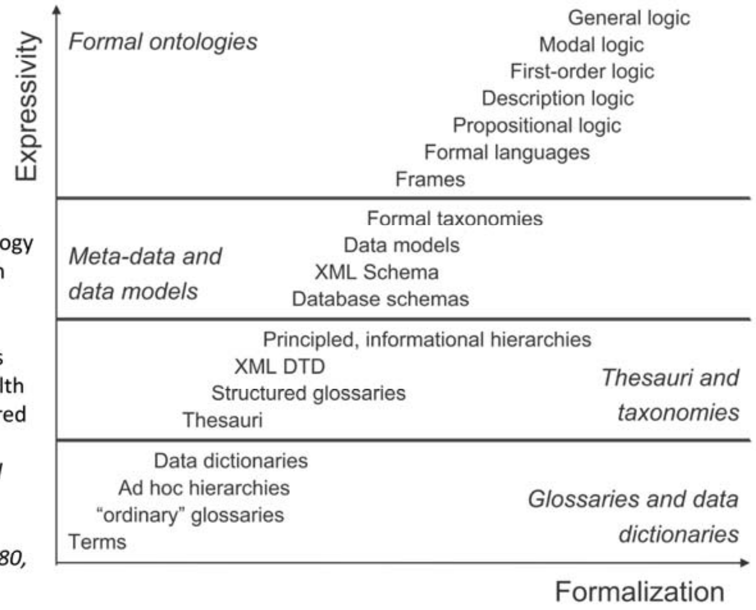
Mathematical Logic	Psychology	Biology	Statistics	Economics
Aristotle				
Descartes				
Boole	James		Laplace	Bentham Pareto
Frege Peano			Bernoullii	Friedman
	Hebb	Lashley	Bayes	
Goedel Post Church Turing Davis Putnam Robinson	Bruner Miller Newell, Simon	Rosenblatt Ashby Lettvin McCulloch, Pitts Heubel, Weisel	Tversky, Kahneman	Von Neumann Simon Raiffa
Logic PROLOG	SOAR **KBS, Frames**	**Connectionism**	**Causal Networks**	**Rational Agents**

Davis, R., Shrobe, H. , Szolovits, P. 1993 What is a knowledge representation? AI Magazine, 14, 1, 17-33.

Logical Representation as a basis for logical reasoning

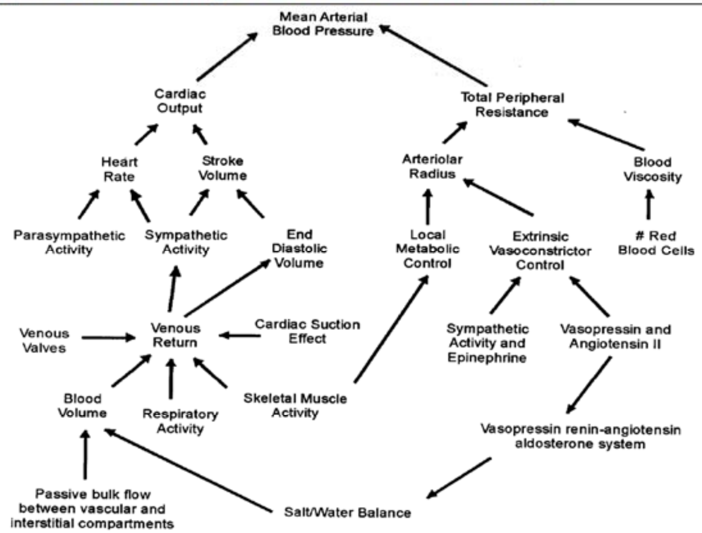


Formalization versus Expressivity



Blobel, B. (2011) Ontology driven health information systems architectures enable pHealth for empowered patients. *International Journal of Medical Informatics, 80, 2, e17-e25.*

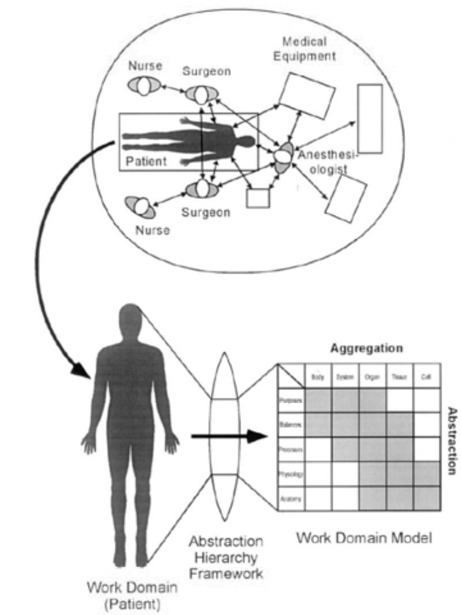
Example for Modeling of biomedical knowledge



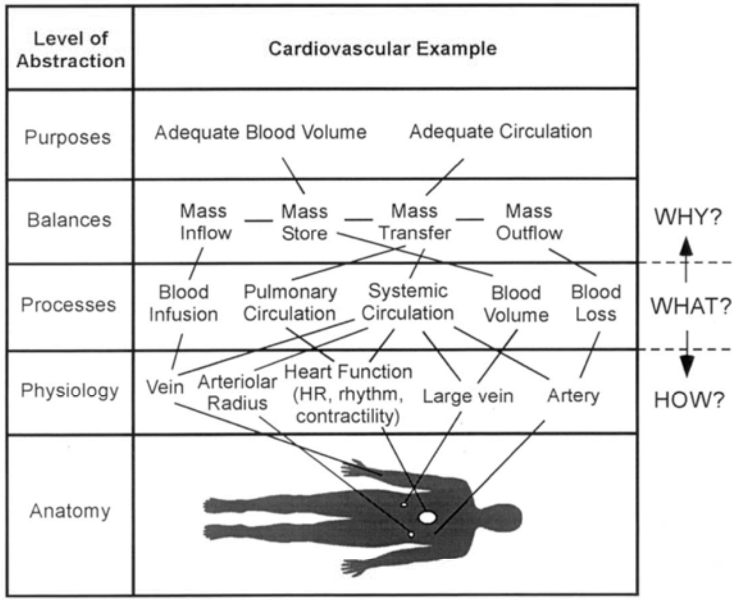
Hajdukiewicz, J. R., Vicente, K. J., Doyle, D. J., Milgram, P. & Burns, C. M. (2001) Modeling a medical environment: an ontology for integrated medical informatics design. *International Journal of Medical Informatics, 62, 1, 79-99.*

Building and Creating a work domain model (WDM)

Hajdukiewicz, J. R., Vicente, K. J., Doyle, D. J., Milgram, P. & Burns, C. M. (2001) Modeling a medical environment: an ontology for integrated medical informatics design. *International Journal of Medical Informatics, 62, 1, 79-99.*



Partial abstraction of the cardiovascular system



Hajdukiewicz et al. (2001)

WDM of: (a) the human body

Level of Aggregation

a)

Level of Abstraction	Body	System	Organ	Tissue	Cell
Purposes	Homeostasis (Maintenance of Internal Environment)	Adequate Circulation, Blood Volume, Oxygenation, Ventilation	Adequate Organ Perfusion, Blood Flow	Adequate Tissue Oxygenation and Perfusion	Adequate Cellular Oxygenation and Perfusion
Balances	Balances: Mass and Energy Inflow, Storage, and Outflow *	System Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Organ Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Tissue Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *	Cellular Balances: Mass and Energy Inflow, Storage, Outflow, and Transfer *
Processes	Total Volume of Body Fluid, Temperature, Supply: O_2, Fluids, Nutrients, Sink: CO_2, Fluids, Wastes	Circulation, Oxygenation, Ventilation, Circulating Volume	Perfusion Pressure, Organ Blood Flow, Vascular Resistance	Tissue Oxygenation, Respiration, Metabolism	Cell Metabolism, Chemical Reaction, Binding, Inflow, Outflow
Physiology		System Function	Organ Function	Tissue Function	Cellular Function
Anatomy			Organ Anatomy	Tissue Anatomy	Cellular Anatomy

* Balances include: Water, Salt, Electrolytes, pH, O_2, CO_2

Hajdukiewicz et al. (2001)

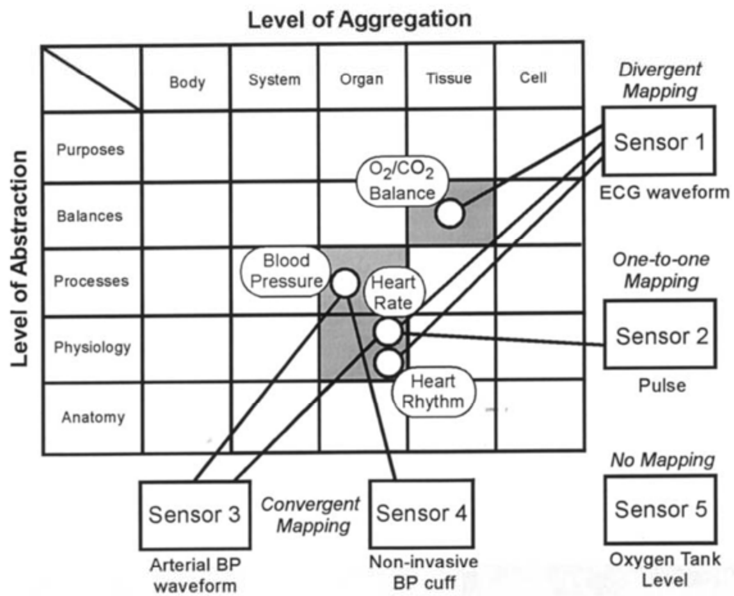
WDM of: (b) the cardiovascular system

b)

Level of Abstraction	System	Subsystem	Organ	Component
Purposes	Adequate Circulation and Blood Volume			
Balances	Cardiovascular System: Mass Inflow, Storage, and Outflow	Pulmonary and Systemic Systems: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer	Organ Vascular Network: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer	Vascular Components: Balance Mass Flows; Mass Inflow, Storage, Outflow, and Transfer
Processes	Circulation, Volume, Fluid Supply and Sink	Pulmonary and Systemic Circulation (Pressure, Flow, Resistance) and Volume, Fluid Supply and Sink	Cardiac Output, Organ Circulation (Pressure, Flow, Resistance), Fluid Supply and Sink from each Vascular Network	Circulation through Vascular Components (Pressure, Flow, Resistance), Vascular Blood Volume, Fluid Supply and Sink
Physiology	Cardiovascular System Function	Pulmonary and Systemic System Function	Cardiac Function (Heart Rate, Rhythm)	Atrial and Ventricular Function; Arterial, Arteriolar, Capillary, Venule, Venous Function
Anatomy			Cardiac Anatomy	Atrial, Ventricular, and Vascular Anatomy

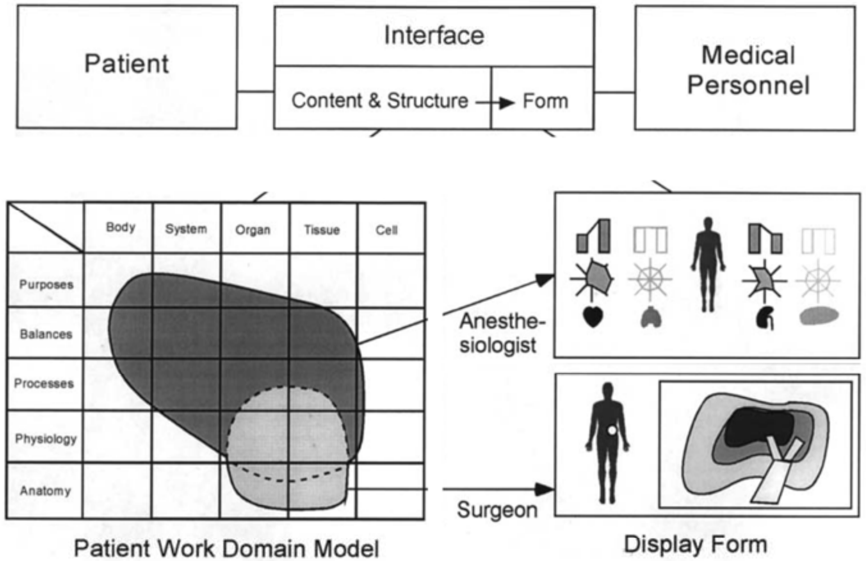
Hajdukiewicz et al. (2001)

Example: Mapping OR sensors onto the WDM



Hajdukiewicz et al. (2001)

Integrated medical informatics design for HCI

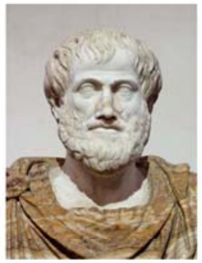


Hajdukiewicz et al. (2001)

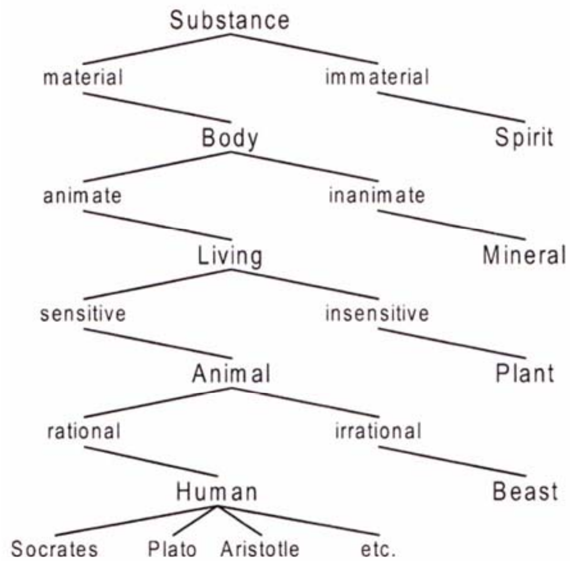
04 Ontologies

A simple question: What is a Jaguar?

The first "Ontology of what exists"



* 384 BC † 322 BC



Simonet, M., Messai, R., Diallo, G. & Simonet, A. (2009) Ontologies in the Health Field. In: Berka, P., Rauch, J. & Zighed, D. A. (Eds.) *Data Mining and Medical Knowledge Management: Cases and Applications. New York, Medical Information Science Reference, 37-56.*

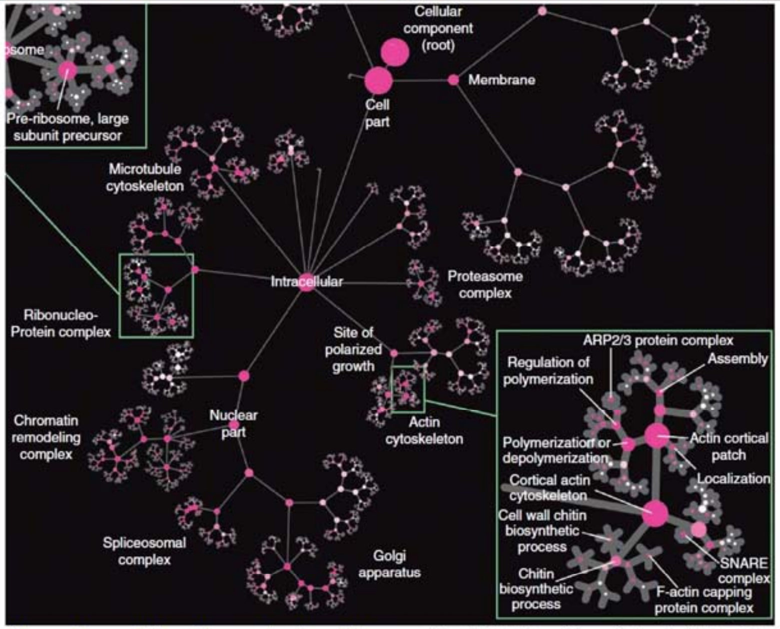
Later: Porphyry (≈ 234-305) ⮕ tree

Ontology: Classic definition

- Aristotle attempted to **classify the things in the world** - where it is employed to describe the existence of beings in the world;
- Artificial Intelligence and Knowledge Engineering deals also with **reasoning about models of the world**.
- Therefore, AI researchers adopted the term 'ontology' to describe **what can be computationally represented** of the world within a program.
- **"An ontology is a formal, explicit specification of a shared conceptualization".**
  - A 'conceptualization' refers to an **abstract model** of some phenomenon in the world by having identified the relevant concepts of that phenomenon.
  - 'Explicit' means that the type of concepts used, and the constraints on their use are **explicitly defined.**

Studer, R., Benjamins, V. R. & Fensel, D. (1998) Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering, 25, 1-2, 161-197.*

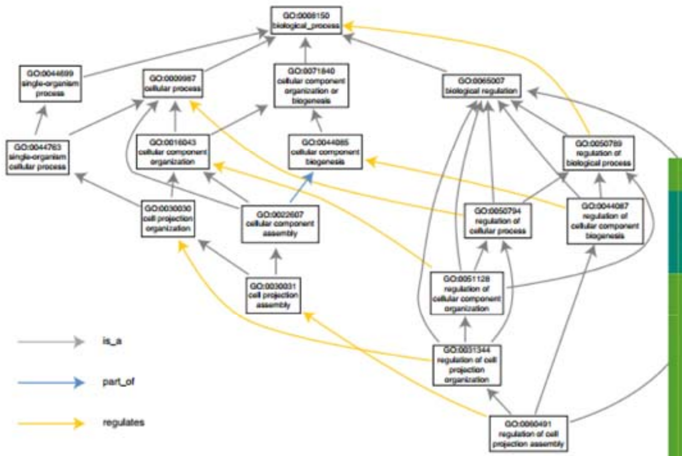
Example: Network-Extracted Ontology of human cell

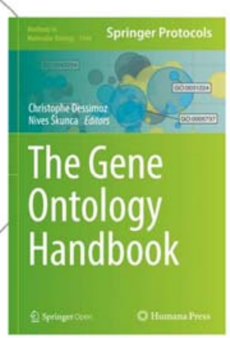


http://www.kurzweilai.net/images/cell-model.png (Credit: UC San Diego School of Medicine)

Example: GO

http://geneontology.org/





Hastings, J. 2017. Primer on Ontologies. In: Dessimoz, C. & Škunca, N. (eds.) The Gene Ontology Handbook. New York, NY: Springer New York, pp. 3-13, doi:10.1007/978-1-4939-3743-1_1.

Ontology: Terminology

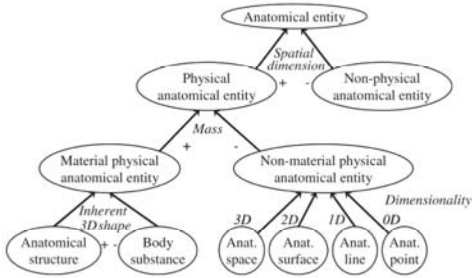
- Ontology = a structured description of a domain in form of **concepts ↔ relations**;
- The **IS-A relation** provides a taxonomic skeleton;
- Other relations reflect the **domain semantics**;
- Formalizes the **terminology** in the domain;
- Terminology = terms definition and usage in the specific **context**;
- Knowledge base = **instance classification** and **concept classification**;
- Classification provides the **domain terminology**

...

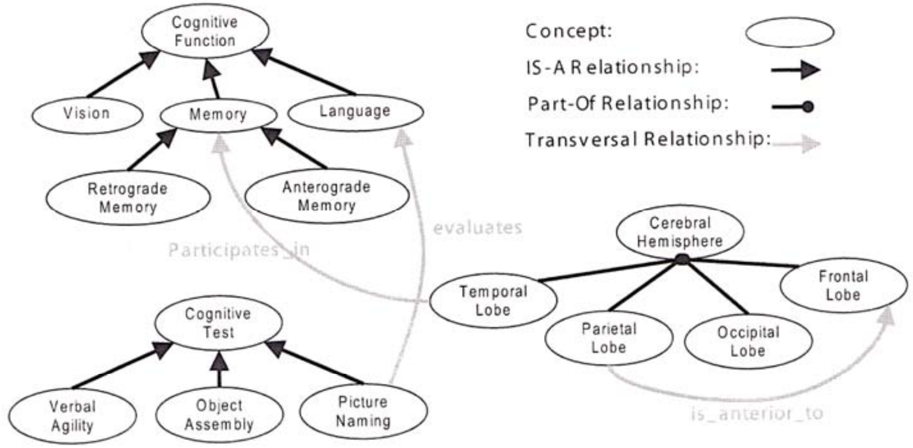
Additionally an ontology may satisfy:

(1) In addition to the IS-A relationship, partitive (meronomic) relationships may hold between concepts, denoted by PART-OF. Every PART-OF relationship is irreflexive, asymmetric and transitive. IS-A and PART-OF are also called hierarchical relationships.
(2) In addition to hierarchical relationships, associative relationships may hold between concepts. Some associative relationships are domain-specific (e.g., the branching relationship between arteries in anatomy and rivers in geography).
(3) Relationships r and r' are inverses if, for every pair of concepts x and y, the relations $\langle x, r, y \rangle$ and $\langle y, r', x \rangle$ hold simultaneously. A symmetric relationship is its own inverse. Inverses of hierarchical relationships are called INVERSE-IS-A and HAS-PART, respectively.
(4) Every non-taxonomic relation of x to z, $\langle x, r, z \rangle$, is either inherited ($\langle y, r, z \rangle$) or refined ($\langle y, r, z' \rangle$ where z' is more specific than z) by every child y of x. In other words, every child y of x has the same properties (z) as it parent or more specific properties (z').

Zhang, S. & Bodenreider, O. 2006. Law and order: Assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy. *Computers in Biology and Medicine, 36, (7-8), 674-693.*



Example of a conceptual structure from CogSci



Simonet, M., Messai, R., Diallo, G. & Simonet, A. (2009) Ontologies in the Health Field. In: Berka, P., Rauch, J. & Zighed, D. A. (Eds.) *Data Mining and Medical Knowledge Management: Cases and Applications. New York, Medical Information Science Reference, 37-56.*

Examples of Biomedical Ontologies

Name	Ref.	Scope	# concepts	# concept names Min	Max	Med	Avg	Subs. Hier.	Version / Notes
SNOMED CT	[21]	Clinical medicine (patient records)	310,314	1	37	2	2.57	yes	July 31, 2007
LOINC	[24]	Clinical observations and laboratory tests	46,406	1	3	3	2.85	no	Version 2.21 (no "natural language" names)
FMA	[25]	Human anatomical structures	~72,000	1	?	?	~1.50	yes	(not yet in the UMLS)
Gene Ontology	[28]	Functional annotation of gene products	22,546	1	24	1	2.15	yes	Jan. 2, 2007
RxNorm	[31]	Standard names for prescription drugs	93,426	1	2	1	1.10	no	Aug. 31, 2007
NCI Thesaurus	[34]	Cancer research, clinical care, public information	58,868	1	100	2	2.68	yes	2007_05E
ICD-10	[36]	Diseases and conditions (health statistics)	12,318	1	1	1	1.00	no	1998 (tabular)
MeSH	[38]	Biomedicine (descriptors for indexing the literature)	24,767	1	208	5	7.47	no	Aug. 27, 2007
UMLS Meta.	[41]	Terminology integration in the life sciences	1,4 M	1	339	2	3.77	n/a	2007AC (English only)

Bodenreider, O. (2008) Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Methods of Information In Medicine, 47, Supplement 1, 67-79.*

Taxonomy of Ontology Languages

- **1) Graph notations**
  - Semantic networks
  - Topic Maps (ISO/IEC 13250)
  - Unified Modeling Language (UML)
  - Resource Description Framework (RDF)
- **2) Logic based**
  - Description Logics (e.g., OIL, DAML+OIL, OWL)
  - Rules (e.g. RuleML, LP/Prolog)
  - First Order Logic (KIF – Knowledge Interchange Format)
  - Conceptual graphs
  - (Syntactically) higher order logics (e.g. LBase)
  - Non-classical logics (e.g. Flogic, Non-Mon, modalities)
- **3) Probabilistic/fuzzy**

Example for (1) Graphical Notation: RDF

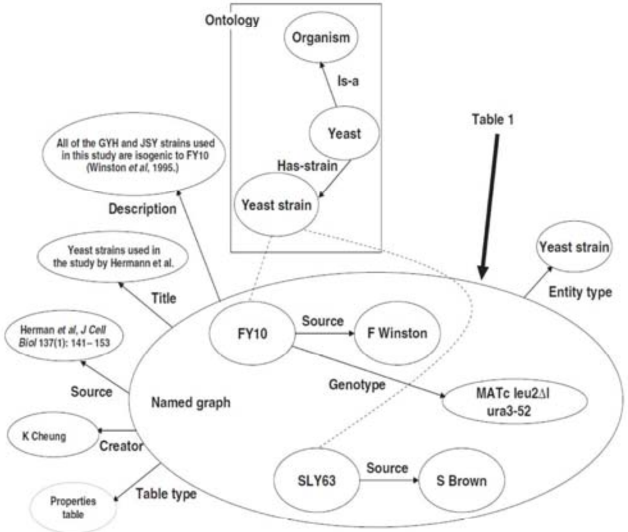


Table 1 Yeast strains used in the study by Hermann et al (1997)

Name	Genotype*	Source
FY10	MATα leu2Δ1 ura3-52	F Winston
FY22	MATα his3Δ200 ura3-52	F Winston
GHY1	MATα leu2Δ1 his3Δ200 ura3-52 mdm20-1	This study
JSY707	MATα his3Δ200 ura3-52 tpm1D::HIS3	This study
JSY948	MATα leu2Δ1/leu2Δ1 ura3-52/ura3-52	This study
JSY999	MATα leu2Δ1 his3Δ200 ura3-52	This study
JSY1065	MATα leu2Δ1 his3Δ200 ura3-52 mdm20D::LEU2	This study
JSY1084	MATα leu2Δ1 his3Δ200 ura3-52 tpm1D::HIS3	This study
JSY1138	MATα leu2Δ1/leu2Δ1 his3Δ200/his3Δ 200 ura3-52/ura3-52 tpm1D::HIS3/ + mdm20D::LEU2/ +	This study
JSY1285	MATα leu2Δ1 his3Δ200 ura3-52 tpm2D::HIS3	This study
JSY1340	MATα leu2Δ1 his3Δ200 ura3-52 mdm20D::LEU2	This study
JSY1374	MATα leu2Δ1/leu2Δ1 his3Δ200/his3Δ200 ura3-52/ura3-52 tpm2D::HIS3/ + mdm20D::LEU2/ +	This study
ABY1249	MATα leu2-3,112 ura3-52 lys2-801 ade2-101 ade3 bem2-10	A Bretscher
IGY4	MATα leu2-3,112 his3Δ200 ura3-52 lys2-801 ade2 sac6D::LEU2	A Adams
SLY63	MATα leu2-3,112 ura3-52 trp1-1 his6 myo2-66	S Brown

Cheung, K.-H., Samwald, M., Auerbach, R. K. & Gerstein, M. B. 2010. Structured digital tables on the Semantic Web: toward a structured digital literature. *Molecular Systems Biology, 6, 403.*

Example for (2) Web Ontology Language OWL

DL = Description Logic

Concept inclusion, Speak: All C1 are C2

Concept equivalence Speak: C1 is equivalent to C2

Axiom	DL syntax	Example
Sub class	$C_1 \sqsubseteq C_2$	Alga ⊑ Plant ⊑ Organism
Equivalent class	$C_1 \equiv C_2$	Cancer ≡ Neoplastic Process
Disjoint with	$C_1 \sqsubseteq \neg C_2$	Vertebrate ⊑ ¬Invertebrate
Same individual	$x_1 \equiv x_2$	Blue_Shark ≡ Prionace_Glauca
Different from	$x_1 \sqsubseteq \neg x_2$	Sea Horse ⊑ ¬Horse
Sub property	$P_1 \sqsubseteq P_2$	has_mother ⊑ has_parent
Equivalent property	$P_1 \equiv P_2$	treated_by ≡ cured_by
Inverse	$P_1 \equiv P_2^-$	location_of ≡ has_location$^-$
Transitive property	$P^+ \sqsubseteq P$	part_of$^+$ ⊑ part_of
Functional property	$\top \sqsubseteq \leq 1P$	$\top \sqsubseteq \leq$ 1has_tributary
Inverse functional property	$\top \sqsubseteq \leq 1P^-$	$\top \sqsubseteq \leq$ 1has_scientific_name$^-$

Bhatt, M., Rahayu, W., Soni, S. P. & Wouters, C. (2009) Ontology driven semantic profiling and retrieval in medical information systems. *Web Semantics: Science, Services and Agents on the World Wide Web, 7, 4, 317-331.*

OWL class constructors

Intersection/conjunction of concepts, Speak: C1 and … Cn

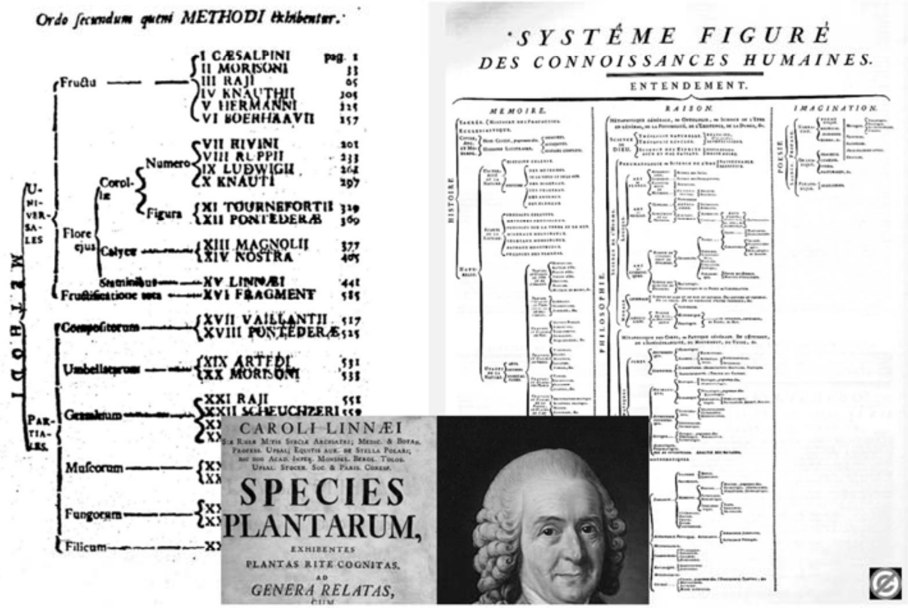
Constructor	DL syntax	Example
Intersection	$C_1 \sqcap \ldots \sqcap C_n$	Anatomical_Abnormality ⊓ Pathological_Function
Union	$C_1 \sqcup \ldots \sqcup C_n$	Body_Substance ⊔ Organic_Chemical
Complement	$\neg C$	¬Invertebrate
One of	$x_1 \sqcup \ldots \sqcup x_n$	Oestrogen ⊔ Progesterone
All values from	$\forall P.C$	∀co_occurs_with.Plant
Some values	$\exists P.C$	∃co_occurs_with.Animal
Max cardinality	$\leq nP$	≤ 1has_ingredient
Min cardinality	$\geq nP$	≥ 2has_ingredient

Universal Restriction Speak: All P-successors are in C

Existential Restriction Speak: An P-successor exists in C

Bhatt et al. (2009)

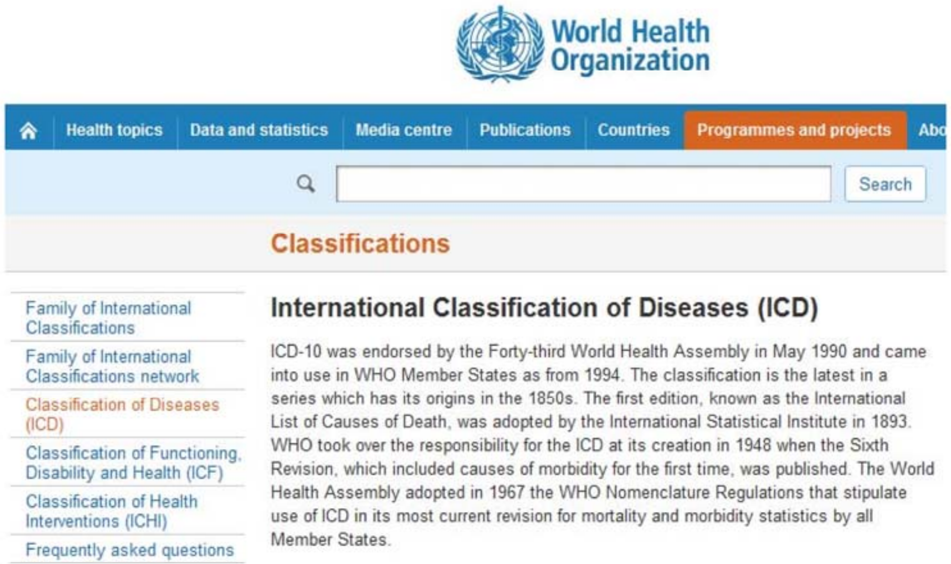
05 Medical Classifications



Medical Classifications – rough overview

- Since the classification by Carl von Linne (1735) approx. 100+ various classifications in use:
  - **I**nternational **C**lassification of **D**iseases (ICD)
  - **S**ystematized **No**menclature of **Med**icine (SNOMED)
  - **Me**dical **S**ubject **H**eadings (MeSH)
  - **F**oundational **M**odel of **A**natomy (FMA)
  - **G**ene **O**ntology (GO)
  - **U**nified **M**edical **L**anguage **S**ystem (UMLS)
  - **L**ogical **O**bservation **I**dentifiers **N**ames & **C**odes (LOINC)
  - **N**ational **C**ancer **I**nstitute Thesaurus (NCI Thesaurus)

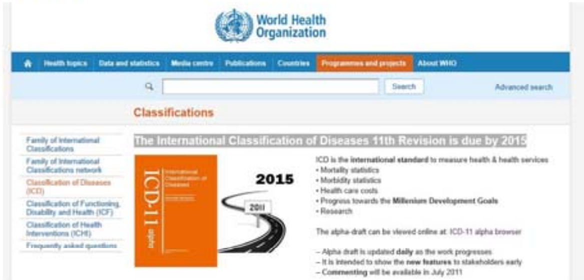
International Classification of Diseases (ICD)



http://www.who.int/classifications/icd/en

International Classification of Diseases (ICD)

- 1629 London Bills of Mortality
- 1855 **William Farr** (London, one founder of medical statistics): List of causes of death, list of diseases
- 1893 von Jacques Bertillot: List of causes of death
- 1900 International Statistical Institute (ISI) accepts Bertillot's list
- 1938 5th Edition
- 1948 WHO
- 1965 ICD-8
- 1989 ICD-10
- 2015 ICD-11 due
- 2018 ICD-11 adopt



Systematized Nomenclature of Medicine SNOMED

- 1965 SNOP, 1974 SNOMED, 1979 SNOMED II
- 1997 (Logical Observation Identifiers Names and Codes (LOINC) integrated into SNOMED
- 2000 SNOMED RT, 2002 SNOMED CT

INTERNATIONAL HEALTH TERMINOLOGY STANDARDS DEVELOPMENT ORGANISATION

239 pages

SNOMED CT® Technical Reference Guide

January 2011 International Release

(US English)

http://www.isb.nhs.uk/documents/isb-0034/amd-26-2006/techrefguid.pdf

SNOMED Example Hypertension

A

24184005|Finding of increased blood pressure (finding) →
38936003|Abnormal blood pressure (finding) AND
roleGroup SOME
(363714003|Interprets (attribute) SOME
75367002|Blood pressure (observable entity))

B

12763006|Finding of decreased blood pressure (finding)→
392570002|Blood pressure finding (finding) AND
roleGroup SOME
(363714003|Interprets (attribute) SOME
75367002|Blood pressure (observable entity))

Rector, A. L. & Brandt, S. (2008) Why Do It the Hard Way? The Case for an Expressive Description Logic for SNOMED. *Journal of the American Medical Informatics Association, 15, 6, 744-751.*

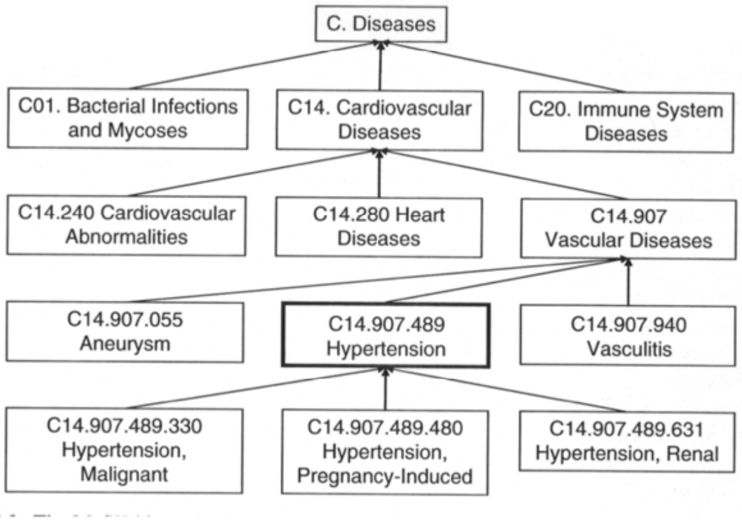
Medical Subject Headings (MeSH)

- MeSH thesaurus is produced by the National Library of Medicine (NLM) since 1960.
- Used for cataloging documents and related media and as an index to search these documents in a database and is part of the metathesaurus of the Unified Medical Language System (UMLS).
- This thesaurus originates from keyword lists of the Index Medicus (today Medline);
- MeSH thesaurus is polyhierarchic, i.e. every concept can occur multiple times. It consists of the three parts:
  - 1. MeSH Tree Structures,
  - 2. MeSH Annotated Alphabetic List and
  - 3. Permuted MeSH.

The 16 trees in MeSH

1. Anatomy [A]
2. Organisms [B]
3. Diseases [C]
4. Chemicals and Drugs [D]
5. Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. Psychiatry and Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
9. Anthropology, Education, Sociology, Social Phenomena [I]
10. Technology, Industry, Agriculture [J]
11. Humanities [K]
12. Information Science [L]
13. Named Groups [M]
14. Health Care [N]
15. Publication Characteristics [V]
16. Geographicals [Z]

MeSH Hierarchy: e.g. heading Hypertension 1/2



Hersh, W. (2010) *Information Retrieval: A Health and Biomedical Perspective. New York, Springer.*

MeSH Example Hypertension 2/2

National Library of Medicine - Medical Subject Headings

2011 MeSH

MeSH Descriptor Data

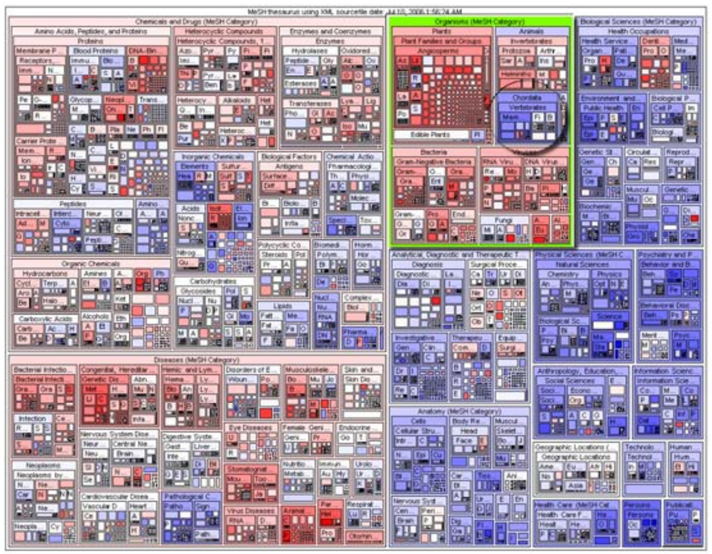
Return to Entry Page

Standard View. Go to Concept View; Go to Expanded Concept View

MeSH Heading	Hypertension
Tree Number	C14.907.489
Annotation	not for intracranial or intraocular pressure; relation to BLOOD PRESSURE: Manual 23.27; Goldblatt kidney is HYPERTENSION, GOLDBLATT see HYPERTENSION, RENOVASCULAR; hypertension with kidney disease is probably HYPERTENSION, RENAL, not HYPERTENSION; venous hypertension: index under VENOUS PRESSURE (IM) & do not coordinate with HYPERTENSION; PREHYPERTENSION is also available
Scope Note	Persistently high systemic arterial BLOOD PRESSURE. Based on multiple readings (BLOOD PRESSURE DETERMINATION), hypertension is currently defined as when SYSTOLIC PRESSURE is consistently greater than 140 mm Hg or when DIASTOLIC PRESSURE is consistently 90 mm Hg or more.
Entry Term	Blood Pressure, High
See Also	Antihypertensive Agents
See Also	Vascular Resistance
Allowable Qualifiers	BL CF CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH RI RT SU TH UR US VE VI
Date of Entry	19990101
Unique ID	D006973

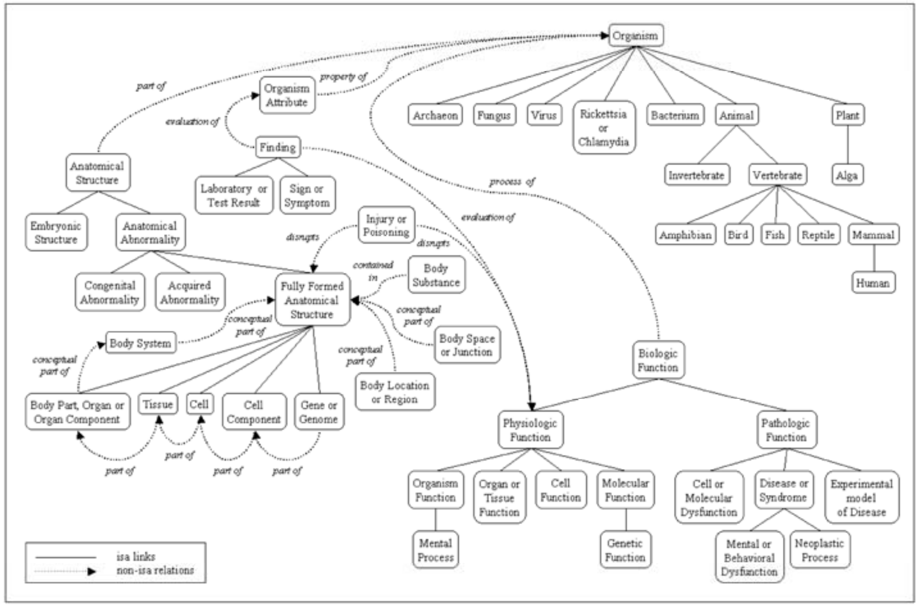
http://www.nlm.nih.gov/mesh/

MeSH Interactive Tree-Map Visualization (see L 9)

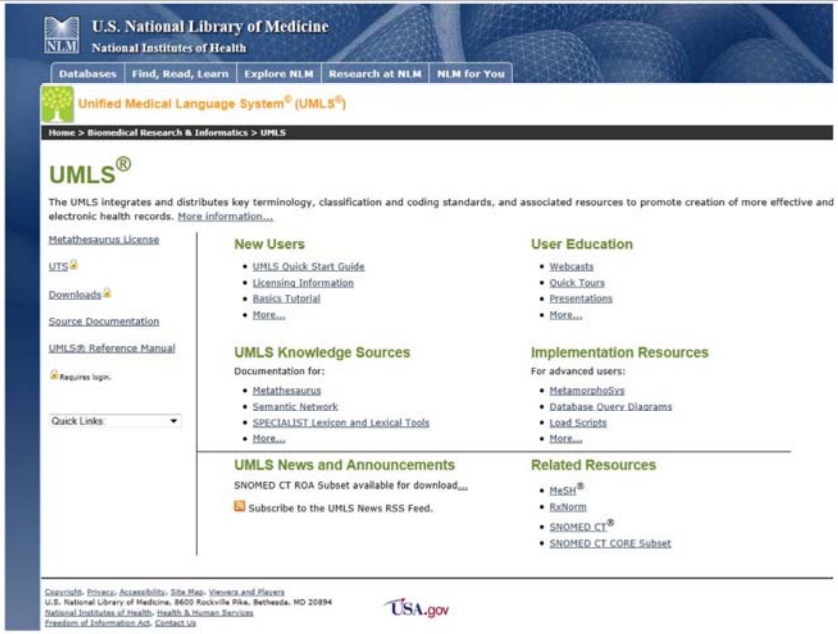


Eckert, K. (2008) A methodology for supervised automatic document annotation. *Bulletin of IEEE Technical Committee on Digital Libraries TCDL, 4, 2.*

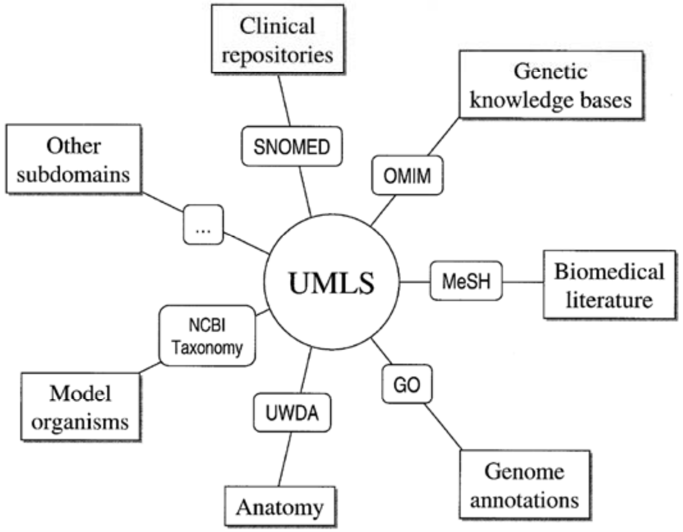
UMLS – Unified Medical Language System



http://www.nlm.nih.gov/research/umls/

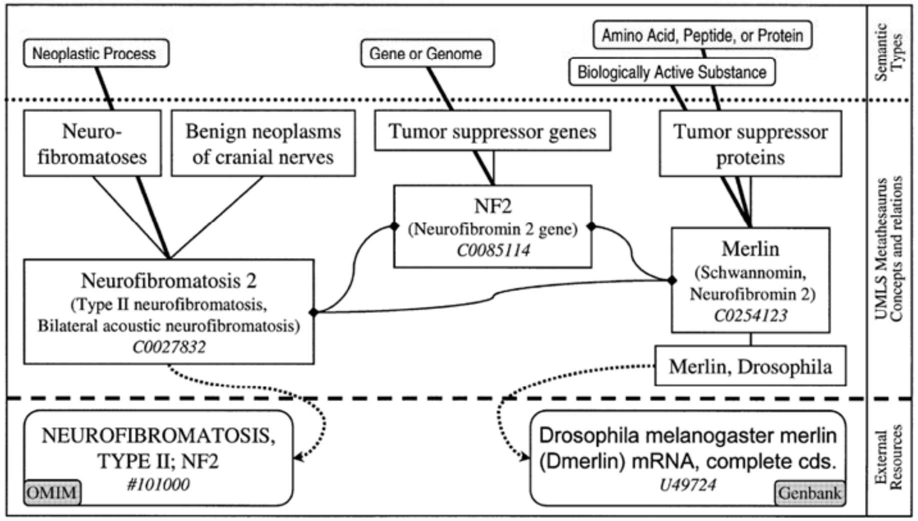


UMLS Metathesaurus integrates sub-domains



Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32, D267-D270.*

Example of proteins and diseases in the UMLS



Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32, D267-D270.*

Conclusion and Future Challenges

Key Problems

- To find a trade-off between standardization and **personalization** [1];
- The large amounts of **non-standardized data** and **unstructured information** (“free text”) [2];
- **Low integration** of standardized terminologies in the daily clinical practice (Who is using e.g. SNOMED, MeSH, UMLS in daily routine?);
- **Low acceptance** of classification codes amongst practitioners;

1. Holmes, C., Mcdonald, F., Jones, M., Ozdemir, V., Graham, J. E. 2010. Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study. Omics-Journal of Integr. Biology, 14, (3), 327-332.
2. Holzinger, A., Schantl, J., Schroettner, M., Seifert, C. & Verspoor, K. 2014. Biomedical Text Mining: State-of-the-Art, Open Problems and Future Challenges. In: LNCS 8401. Berlin Heidelberg: Springer pp. 271-300.

Slide 3-45: Future Challenges

- Data fusion – Data integration in the life sciences
- Self learning stochastic ontologies [1]
- Interactive, integrative machine learning and interactive ontologies - human-in-the-loop
- Never ending learning machines [2] for automatically building knowledge spaces
- Integrating ontologies in daily work
- Knowledge and **context awareness**

[1] Ongenae, F., Claeys, M., Dupont, T., Kerckhove, W., Verhoeve, P., Dhaene, T. & De Turck, F. 2013. A probabilistic ontology-based platform for self-learning context-aware healthcare applications. Expert Systems with Applications, 40, (18), 7629-7646.
[2] Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Hruschka Jr, E. R. & Mitchell, T. M. 2010. Toward an Architecture for Never-Ending Language Learning. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10). Atlanta: AAAI. 1306-1313.

Thank you!

Appendix

Privacy, Security, Safety, Data Protection, Anonymity, Fair Use, …

EBM CPG

Standardized Medicine

Person-alized Medicine

GBM GPM

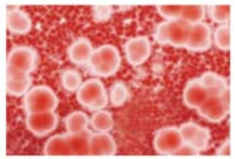
Pervasive Healthcare

Preventive Health Integration

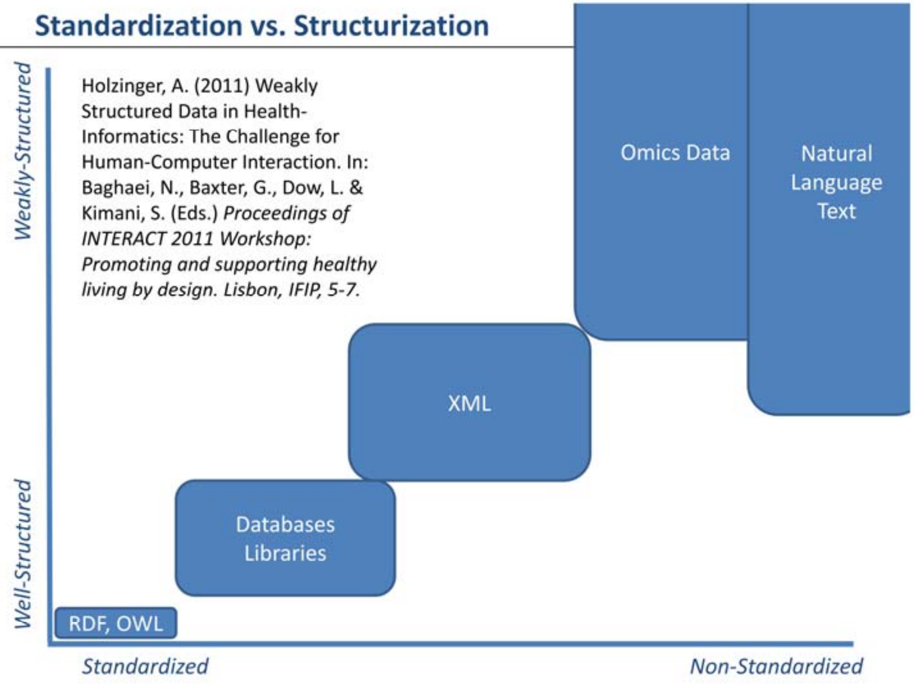
EBM = Evidence Based Medicine
CPG = Clinical Practice Guideline
GBM = Genome Based Medicine
GPM = Genetic Polymorphism

Tanaka, H. (2010)

- **Genomics** (sequence annotation)
- **Transcriptomics** (microarray)
- **Proteomics** (Proteome Databases)
- **Metabolomics** (enzyme annotation)
- **Protein-DNA interactions**
- **Protein-Protein interactions**
- **Fluxomics** (isotopic tracing, metabolic pathways)
- **Phenomics** (biomarkers)
- **Epigenetics**
- **Microbiomics**
- **Lipidomics**



Standardization vs. Structurization



Data Dimensionality

- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data = can be stored as a 2-D matrix, but can represent biological entities in three or more dimensions, e.g. PDB records
- 3-D data = having 3-D spatial component, e.g. image voxels, e-density maps, etc.
- H-D Data = data having arbitrarily high dimensions

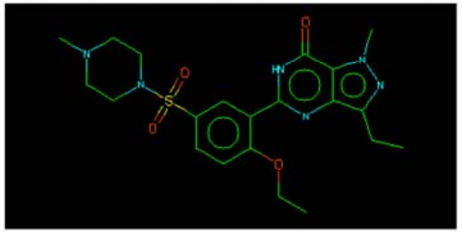
Example: 1-D data (univariate sequential data objects)

SMILES (Simplified Molecular Input Line Entry Specification)

... is a compact machine and human-readable chemical nomenclature:

e.g. Viagra:

CCc1nn(C)c2c(=O)[nH]c(nc12)c3cc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4



...is Canonicalizable

...is Comprehensive

...is Well Documented

http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

Reflection from last lecture

- The Quiz-Slide will be shown during the course