Mini Course

Fundamentals of Medical AI

Part 02

From Data to Knowledge Representation

Andreas Holzinger

Human-Centered AI Lab (Holzinger Group)

Institute for Medical Informatics/Statistics, Medical University Graz, Austria

and

Explainable AI-Lab, Alberta Machine Intelligence Institute, Edmonton, Canada

Overview

Primer on Probability & Information

Part 1 Theory	Part 2 Practice
01 Introduction to Medical AI and Machine Learning for Health	05 Methods of Explainable AI
02 Data, Information and Knowledge	06 Social, Ethical and Legal Aspects of Medical AI
03 Human Decision Making and AI Decision Support	07 Project: Bringing AI into medical workflows
04 Causal Reasoning and Interpretable AI	08 Presentation of the developed concepts

Written Exam

Agenda

- **00 Reflection – follow up from last lecture**
- **01 Data – the underlying physics of data**
- **02 Biomedical data sources: Taxonomy**
- **03 Data integration, mapping, fusion, augmentation**
- **04 Knowledge Representation**
- **05 Biomedical ontologies**
- **06 Biomedical classifications**
- **Conclusion**

00 Reflection

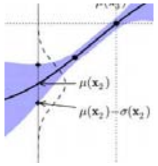
Warm-up Quiz

1

2

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

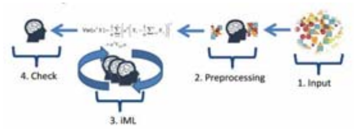
3



4

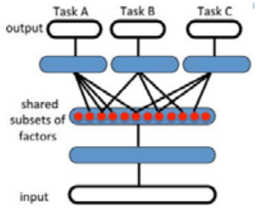
5

6



7

8



9

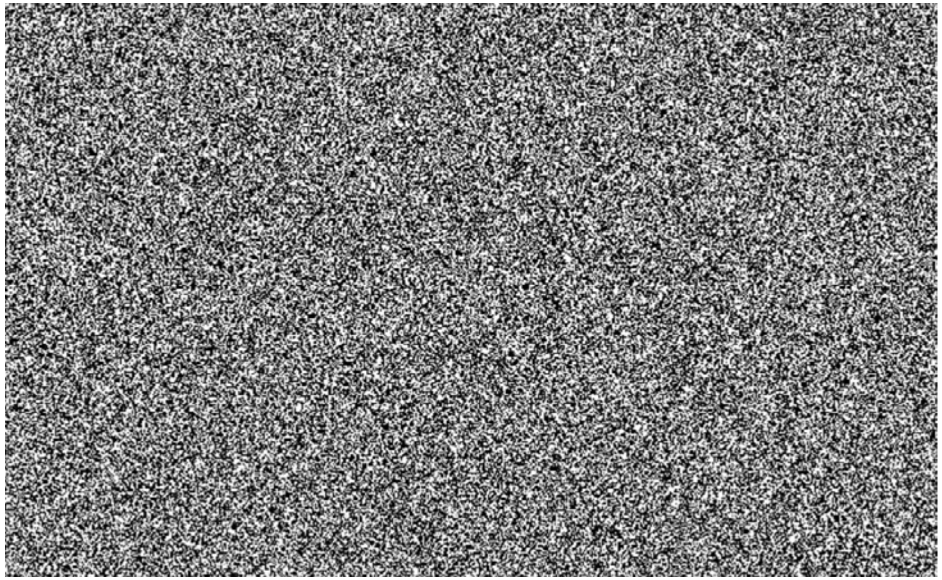
Where is the Biologist in this image ?

Pedro Domingos 2015. The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World, Penguin UK.

What are the underlying physics of data ?

01 Data

What is this ?



What are the key problems in (medical) data science ?

- Heterogeneous, distributed, inconsistent data sources (need for **data integration** & fusion) [1]
- **Complex data** (high-dimensionality – challenge of dimensionality reduction and visualization) [2]
- Noisy, uncertain, missing, dirty, and imprecise, imbalanced data (challenge of **pre-processing**)
- The discrepancy between data-information-knowledge (**various definitions**)
- **Big data** sets in high-dimensions (manual handling of the data is often impossible) [3]

1. 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.
2. Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis of Patient Groups and Immunization Results Based on Subspace Clustering. In: LNAI 9250, 358-368.
3. Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. in CCIS 455. Springer 3-18.

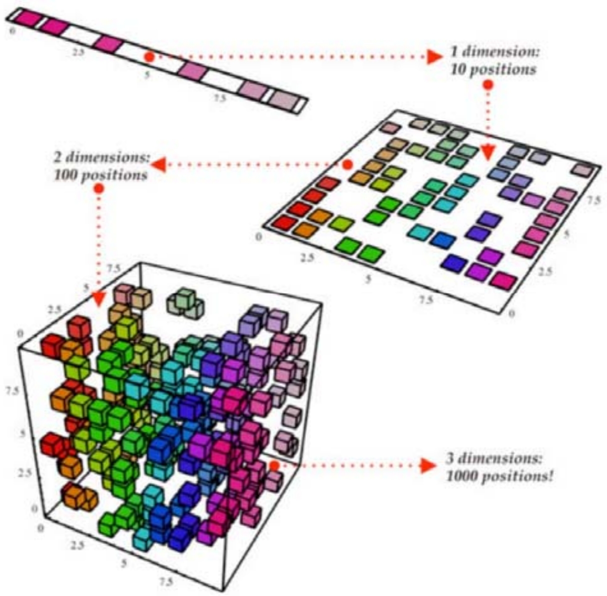
Why can data in ML often not be represented by a simple model ?

- 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

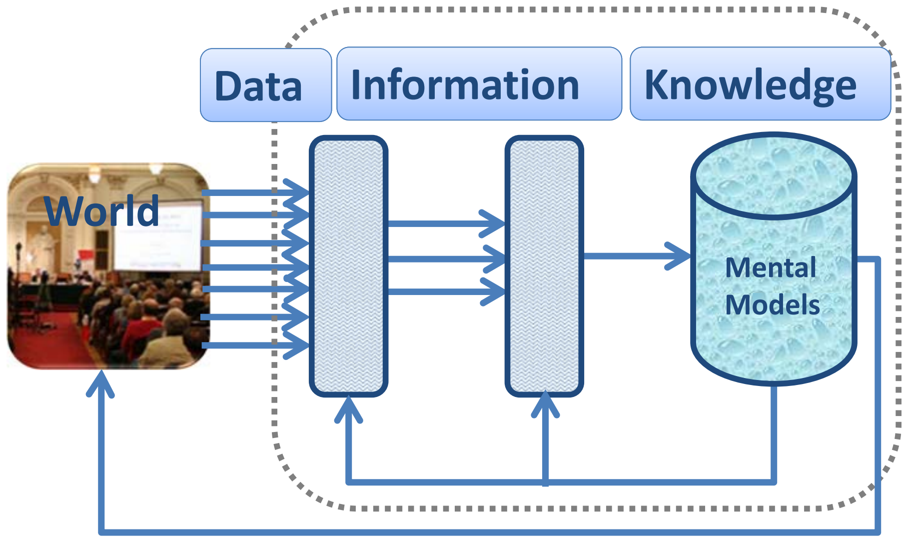
Yann LeCun, Yoshua Bengio & Geoffrey Hinton 2015. Deep learning. Nature, 521, (7553), 436-444, doi:10.1038/nature14539

Why is the curse of dimensionality for us relevant ?



Samy Bengio & Yoshua Bengio 2000. Taking on the curse of dimensionality in joint distributions using neural networks. IEEE Transactions on Neural Networks, 11, (3), 550-557, doi:10.1109/72.846725.

What is the difference between Data – Information – Knowledge ?



Knowledge := a set of expectations

INFORMATICS PROFESSIONALS. LEADING THE WAY.

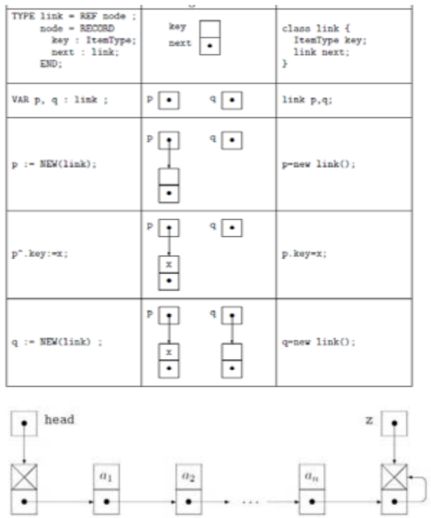
***Biomedical informatics* (BMI) is the interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific problem solving, and decision making, motivated by efforts to improve human health**

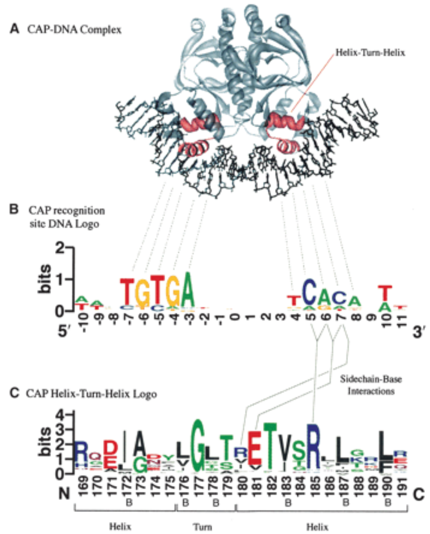
Edward H. Shortliffe 2011. Biomedical Informatics: Defining the Science and its Role in Health Professional Education. In: Holzinger, Andreas & Simonic, Klaus-Martin (eds.) Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058. Heidelberg, New York: Springer, pp. 711-714.

- **Physical level** -> bit = binary digit = **b**asic **i**ndissolube 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, numerical measurements (physiological data, lab results, vital signs, …), recorded signals (ECG, EEG, …), Images (x-ray, MR, CT, PET, …) ; -omics

- **Clinical workplace data sources**
  - Medical documents: text (non-standardized ("free-text"), semi-structured, standard terminologies (ICD, SNOMED-CT)
  - Measurements: lab, time series, ECG, EEG, EOG, …
  - Surveys, Clinical study data, trial data
- **Image data sources**
  - Radiology: MRI (256x256, 200 slices, 16 bit per pixel, uncompressed, ~26 MB); CT (512x512, 60 slices, 16 bit per pixel, uncompressed ~32MB; MR, US;
  - Digital Microscopy : WSI (15mm slide, 20x magn., 24 bits per pixel, uncompressed, 2,5 GB, WSI 10 GB; confocal laser scanning, etc.
- **-omics data sources**
  - Sanger sequencing, NGS whole genome sequencing (3 billion reads, read length of 36) ~ 200 GB; NGS exome sequencing ("only" 110,000,000 reads, read length of 75) ~7GB; Microarray, mass-spectrometry, gas chromatography, …

Andreas Holzinger, Christof Stocker & Matthias Dehmer 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. In: Communications in Computer and Information Science CCIS 455. Berlin Heidelberg: Springer pp. 3-18, doi:10.1007/978-3-662-44791-8_1.

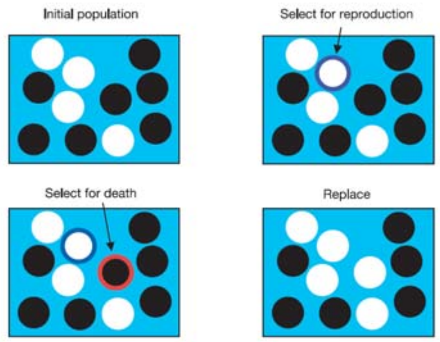


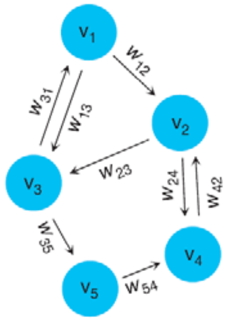


Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

Why is the data structure graph so versatile ?

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



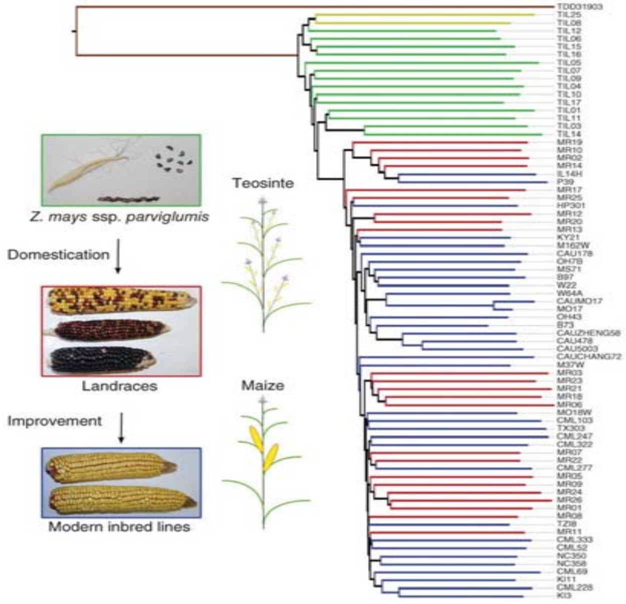


$$W=\begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ 0 & 0 & w_{23} & w_{24} & 0 \\ w_{31} & 0 & 0 & 0 & w_{35} \\ 0 & w_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}$$

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

What is a tree ?

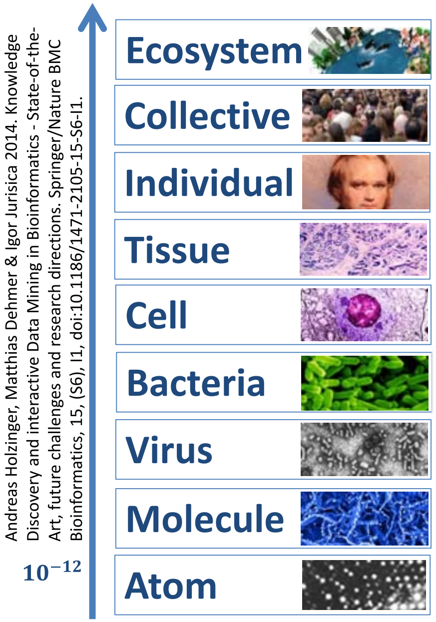
Hufford et. al. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics, 44, (7), 808-811.*

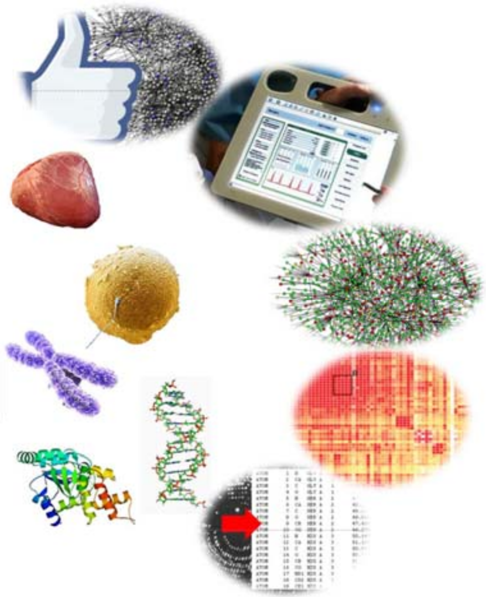


02 Biomedical data sources: Taxonomy of data

What are origins of health-related data ?

Andreas Holzinger, Matthias Dehmer & Igor Jurisica 2014. Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. Springer/Nature BMC Bioinformatics, 15, (S6), I1, doi:10.1186/1471-2105-15-S6-I1.



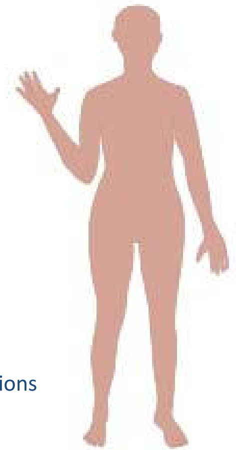


Why is data integration in health an unsolved problem ?

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)

Microbiomes
Microorganisms processes
Plants, Fungi, …

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

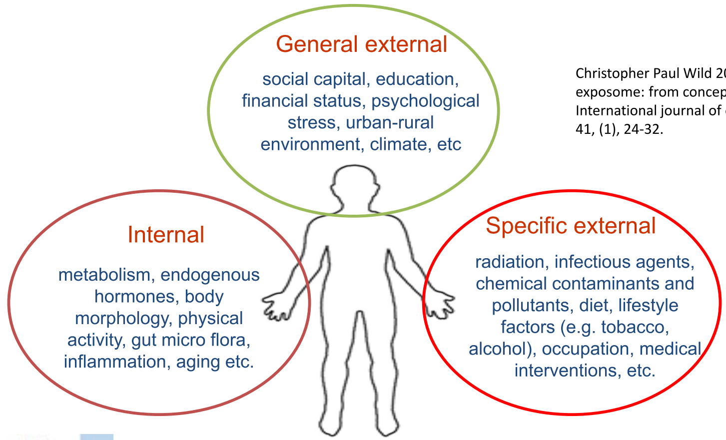
Proteomics
Protein-Protein Interactions

Epigenetics
Epigenetic modifications

Genomics

Transcriptomics
RNA, mRNA, rRNA, tRNA

Why is the human exposome important ?



Christopher Paul Wild 2012. The exposome: from concept to utility. International journal of epidemiology, 41, (1), 24-32.

https://human-centered.ai/project/eu-project-heap-human-exposome-assessment-platform

Where do we get open data sets ?

- Billions of biological data sets are openly available, here only some examples:
- General Repositories:
  - GenBank, EMBL, HMCA, …
- Specialized by data types:
  - UniProt/SwissProt, MMMP, KEGG, PDB, …
- Specialized by organism:
  - WormBase, FlyBase, NeuroMorpho, …
- https://human-centered.ai/open-data-sets

What is *omics data integration ?

Genomics	Transcriptomics	Proteomics	Metabolomics	Protein–DNA interactions	Protein–protein interactions	Fluxomics	Phenomics
Genomics (sequence annotation)	• ORF validation • Regulatory element identification[74]	• SNP effect on protein activity or abundance	• Enzyme annotation	• Binding-site identification[75]	• Functional annotation[79]	• Functional annotation	• Functional annotation[71,103] • Biomarkers[125]
	Transcriptomics (microarray, SAGE)	• Protein: transcript correlation[20]	• Enzyme annotation[108]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[89]	• Regulatory complex identification	• Differential complex formation	• Enzyme capacity	• Functional annotation
			Metabolomics (metabolite abundance)	• Metabolic-transcriptional response		• Metabolic pathway bottlenecks	• Metabolic flexibility • Metabolic engineering[109]
				Protein–DNA interactions (ChIP–chip)	• Signalling cascades[89,102]		• Dynamic network responses[84]
					Protein–protein interactions (yeast 2H, coAP–MS)		• Pathway identification activity[89]
						Fluxomics (isotopic tracing)	• Metabolic engineering
							Phenomics (phenotype arrays, RNAi screens, synthetic lethals)

Joyce, A. R. & Palsson, B. Ø. 2006. The model organism as a system: integrating'omics' data sets. *Nature Reviews Molecular Cell Biology, 7*, ***198-210.***

Examples for lower dimensional data ?

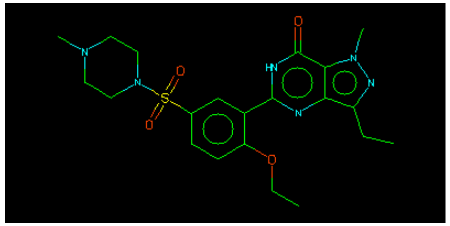
- 0-D data = a data point existing isolated from other data, e.g. integers, letters, Booleans, etc.
- 1-D data = consist of a string of 0-D data, e.g. Sequences representing nucleotide bases and amino acids, SMILES etc.
- 2-D data = having spatial component, such as images, NMR-spectra etc.
- 2.5-D data = 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

What is an example for 1-D data (univariate sequential data)?

SMILES (Simplified Molecular Input Line Entry Specification)

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

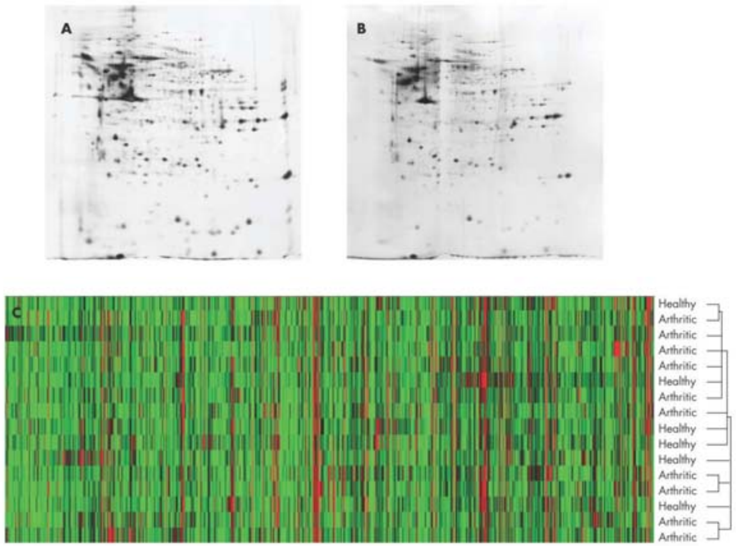
e.g. Viagra:
CCc1nn(C)c2c(=O)[nH]c(nc12)c3cc(ccc3OCC)S(=O)(=O)N4CCN(C)CC4



...is Canonicalizable
...is Comprehensive
...is Well Documented

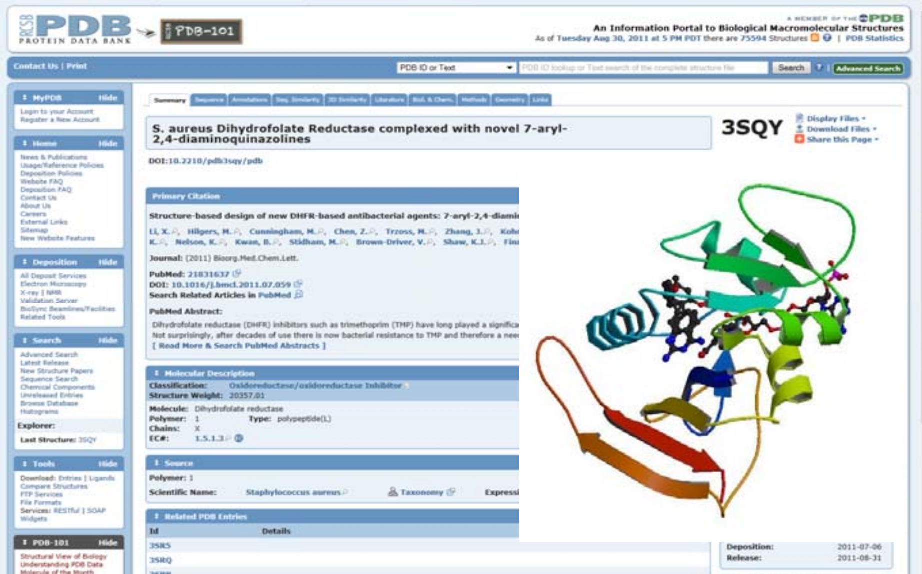
http://www.daylight.com/dayhtml_tutorials/languages/smiles/index.html

What is an example for 2-D data (bivariate data) ?



Kastrinaki et al. (2008) Functional, molecular & proteomic characterisation of bone marrow mesenchymal stem cells in rheumatoid arthritis. *Annals of Rheumatic Diseases, 67, 6, 741-749.*

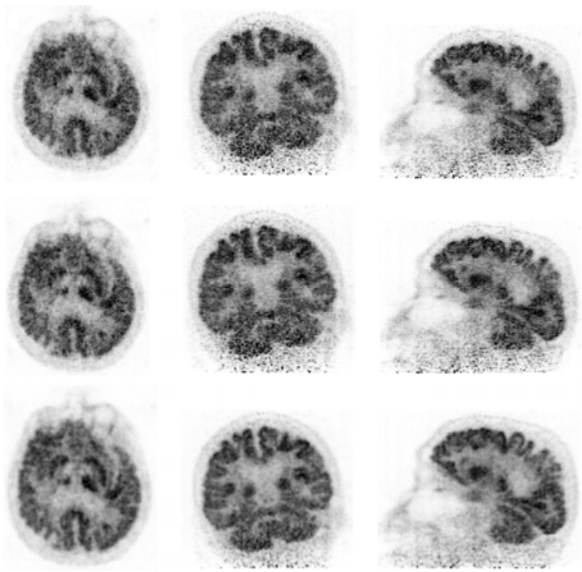
What is an example for 2.5-D data (structural information & metadata)?



http://www.pdb.org

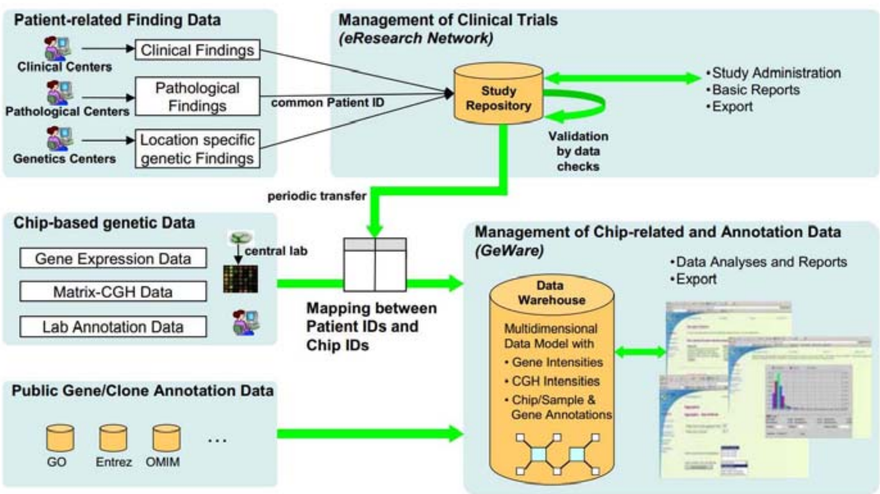
What are 3-D Voxel data (volumetric picture elements) ?

Scheins, J. J., Herzog, H. & Shah, N. J. (2011) Fully-3D PET Image Reconstruction Using Scanner-Independent, Adaptive Projection Data and Highly Rotation-Symmetric Voxel Assemblies. *Medical Imaging, IEEE Transactions on, 30, 3, 879-892.*



03 Data Integration, mapping, fusion

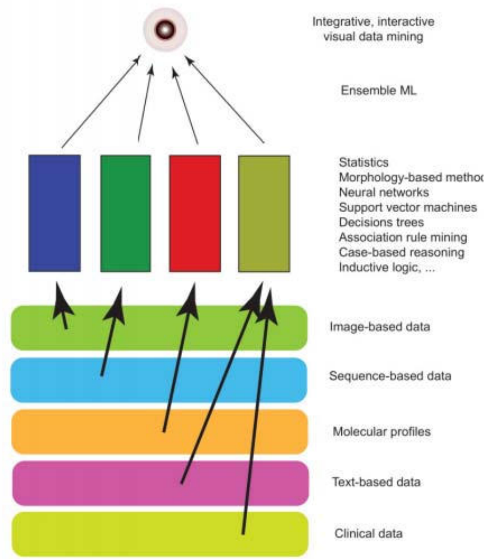
What do we mean with data integration ?



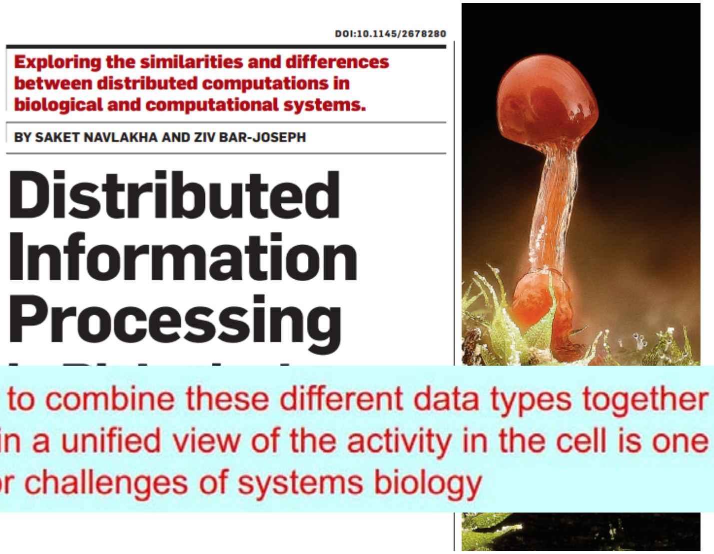
Kirsten, T., Lange, J. & Rahm, E. 2006. An integrated platform for analyzing molecular-biological data within clinical studies. Current Trends in Database Technology–EDBT 2006. Heidelberg: Springer, pp. 399-410, doi:10.1007/11896548_31.

What is the goal of data integration ?

Goal:
Unified View for decision support ("what is relevant?")



Holzinger, A. & Jurisica, I. 2014. Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions In: Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 1-18, doi:10.1007/978-3-662-43968-5_1.



DOI:10.1145/2678280

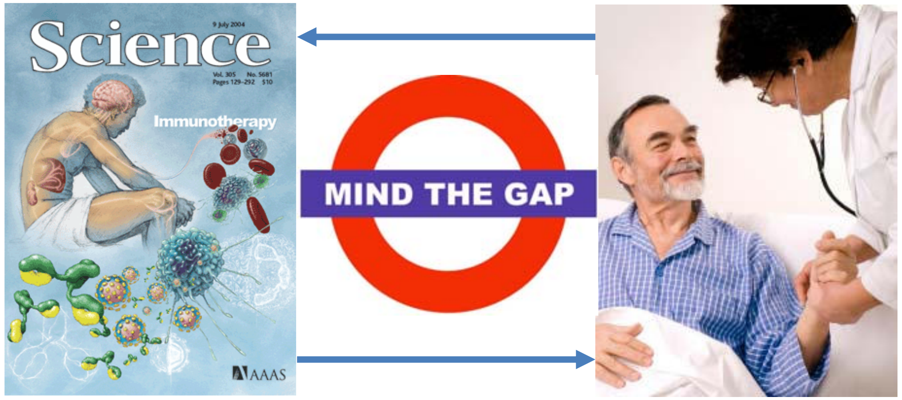
Exploring the similarities and differences between distributed computations in biological and computational systems.

BY SAKET NAVLAKHA AND ZIV BAR-JOSEPH

Distributed Information Processing

How to combine these different data types together to obtain a unified view of the activity in the cell is one of the major challenges of systems biology

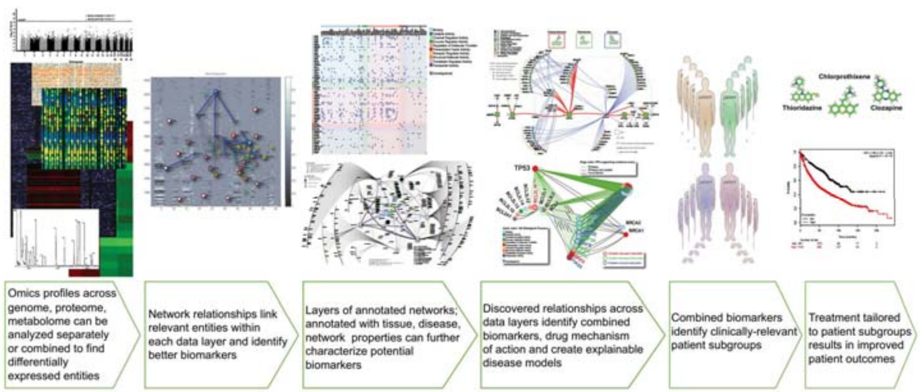
Navlakha, S. & Bar-Joseph, Z. 2014. Distributed information processing in biological and computational systems. *Commun. ACM*, 58, (1), 94-102, doi:10.1145/2678280.



Our central hypothesis: Information may bridge this gap

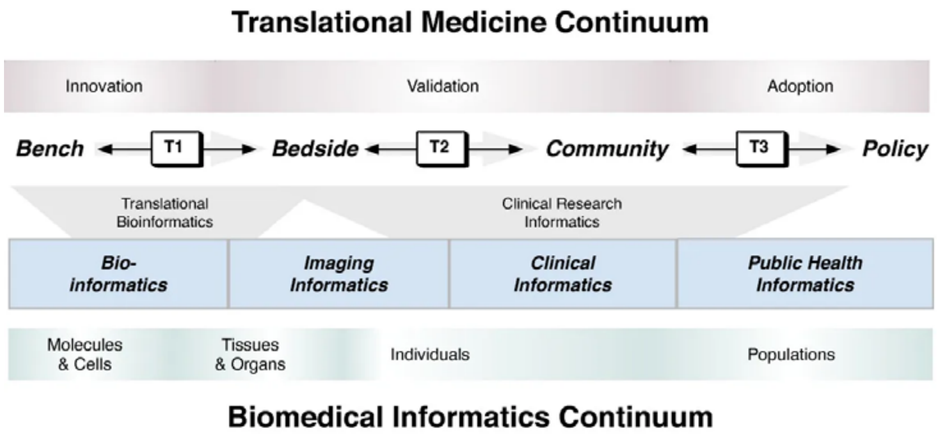
Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*

Why is imaging data along not enough ?



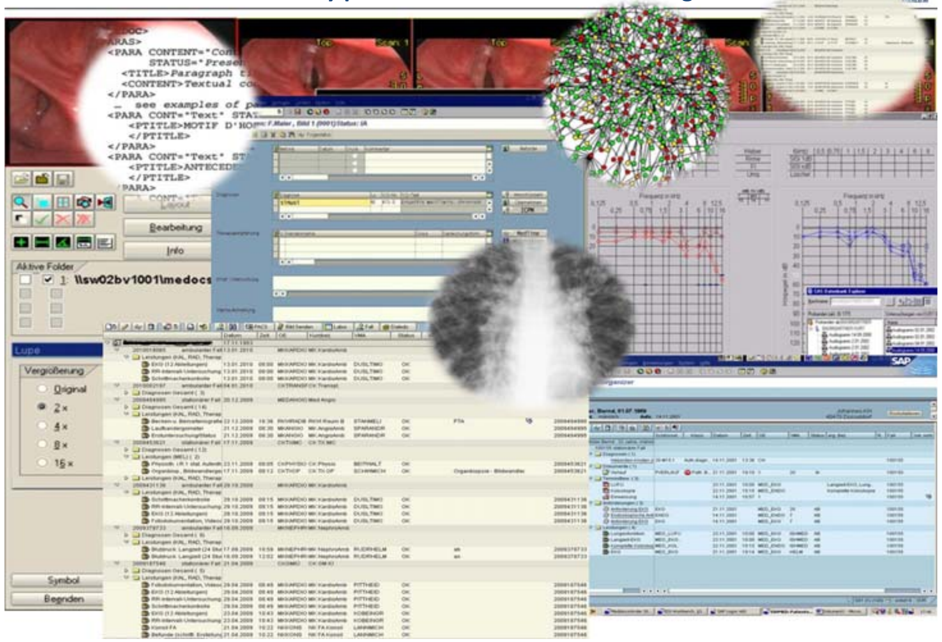
Andreas Holzinger, Benjamin Haibe-Kains & Igor Jurisica 2019. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. European Journal of Nuclear Medicine and Molecular Imaging, 46, (13), 2722-2730, doi:10.1007/s00259-019-04382-9.

What is translational health ?

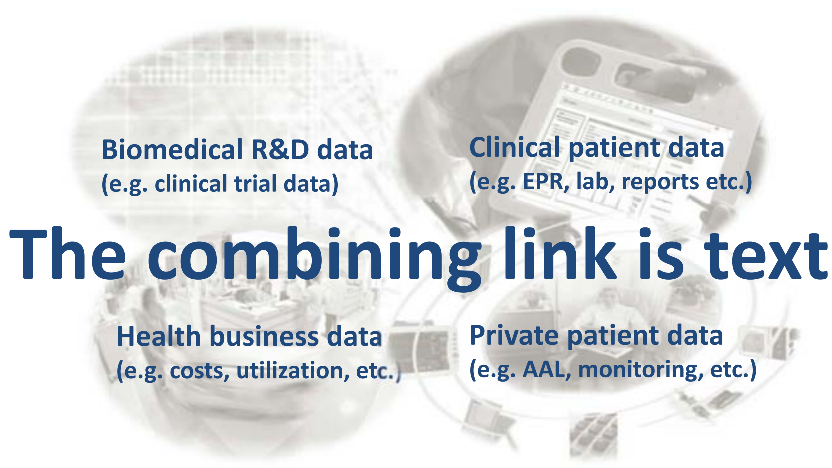


Indra N. Sarkar 2010. Biomedical informatics and translational medicine. Journal of Translational Medicine, 8, (1), 2-12, doi:10.1186/1479-5876-8-22

What are the key problems in medical data management ?



What is the combining link ?



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.*

What are the typical challenges in data-driven medicine ?

- Increasingly large data sets ("big data") due to **data-driven medicine** [1] (which is good for learning!)
- 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.

How to ensure good data quality assessment ?

Leo L. Pipino, Yang W. Lee & Richard Y. Wang 2002. Data quality assessment. Communications of the ACM, 45, (4), 211-218.

Dimensions	Definitions
Accessibility	the extent to which data is available, or easily and quickly retrievable
Appropriate Amount of Data	the extent to which the volume of data is appropriate for the task at hand
Believability	the extent to which data is regarded as true and credible
Completeness	the extent to which data is not missing and is of sufficient breadth and depth for the task at hand
Concise Representation	the extent to which data is compactly represented
Consistent Representation	the extent to which data is presented in the same format
Ease of Manipulation	the extent to which data is easy to manipulate and apply to different tasks
Free-of-Error	the extent to which data is correct and reliable
Interpretability	the extent to which data is in appropriate languages, symbols, and units, and the definitions are clear
Objectivity	the extent to which data is unbiased, unprejudiced, and impartial
Relevancy	the extent to which data is applicable and helpful for the task at hand
Reputation	the extent to which data is highly regarded in terms of its source or content
Security	the extent to which access to data is restricted appropriately to maintain its security
Timeliness	the extent to which the data is sufficiently up-to-date for the task at hand
Understandability	the extent to which data is easily comprehended
Value-Added	the extent to which data is beneficial and provides advantages from its use

- "The value of data lies in reusability".
- What are the attributes that make data reusable?
- **F**indable: metadata -persistent identifier
- **A**ccessible: retrievable by humans and machines through standards, open and free by default; authentication and authorization where necessary
- **I**nteroperable: metadata use a 'formal, accessible, shared, and broadly applicable language for knowledge representation'.
- **R**eusable: metadata provide rich and accurate information; clear usage license; provenance.

Mark D. Wilkinson et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3, 160018, doi:10.1038/sdata.2016.18.

https://www.go-fair.org/fair-principles

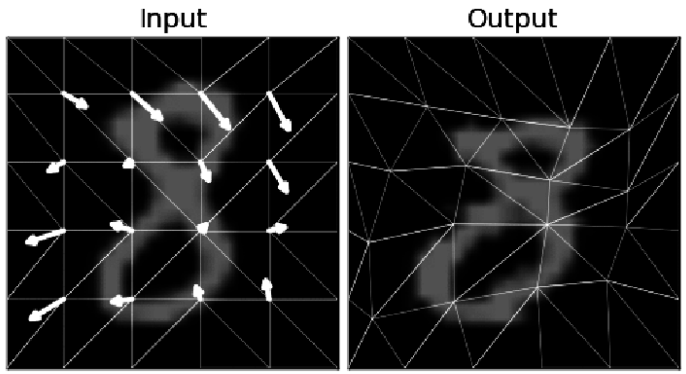
Digression: Data Augmentation

What is data augmentation ?

- Generation of artificial data via expansion of your dataset
- Why ?
- Neural networks require "big data" so augmentation is now basically part of most all deep learning projects
- It is also used to address issues with class imbalance
- It is a cheap and relatively easy way to get more data, which will almost certainly improve the accuracy of a trained model
- It improves model generalisation, model accuracy, and can control overfitting
- Image augmentation is most common, because text augmentation is much harder, and DL is applied to images
- done by making label-preserving transformations to the original images (e.g. rotation, zooming, cropping, …)

Marcus D. Bloice, Peter M. Roth & Andreas Holzinger 2019. Biomedical image augmentation using Augmentor. Oxford Bioinformatics, 35, (1), 4522-4524, doi:10.1093/bioinformatics/btz259.

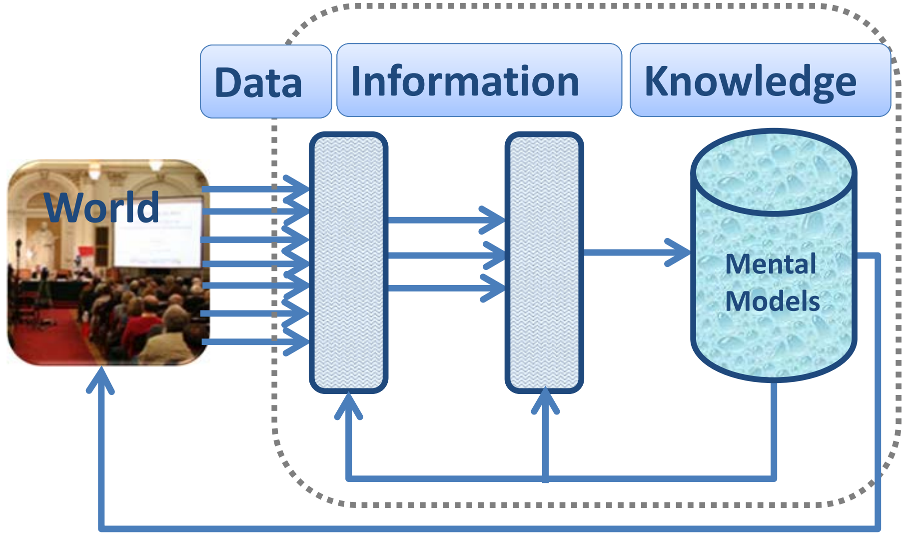
How does image augmentation work ?



Marcus D Bloice, Christof Stocker & Andreas Holzinger 2017. Augmentor: an image augmentation library for machine learning. arXiv preprint arXiv:1708.04680.

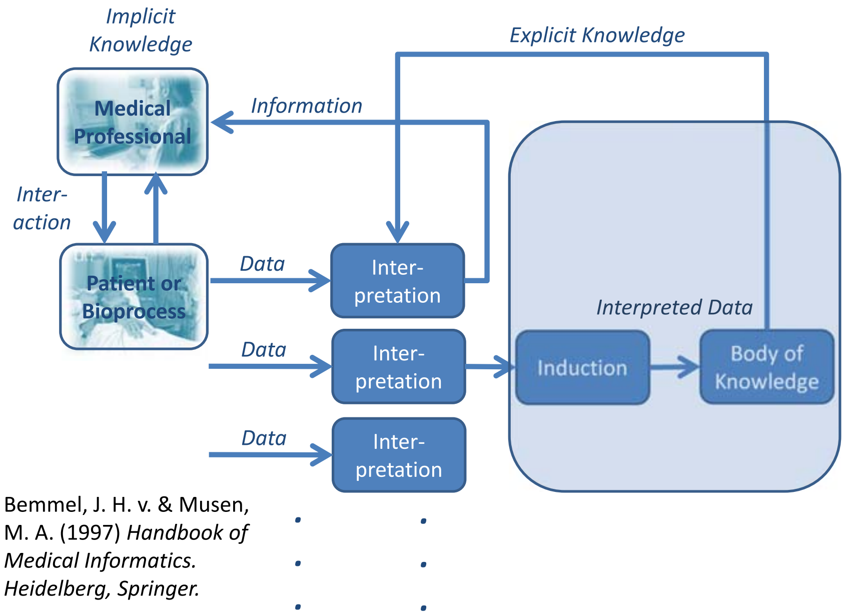
04 Knowledge Representation

What is the difference between Data – Information – Knowledge ?



Knowledge := a set of expectations

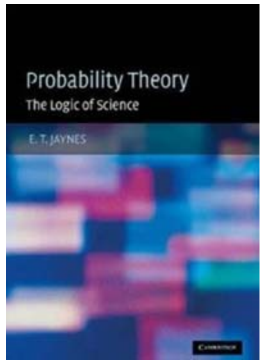
What is medical knowledge ? Where does the ground truth come ?



Bemmel, J. H. v. & Musen, M. A. (1997) *Handbook of Medical Informatics. Heidelberg, Springer.*

Why is logic insufficient for solving complex real-world problems ?

- Logical representations are based on
  - Facts about the world (true or false)
  - These facts can be combined with logical operators
  - Logical inference is based on certainty



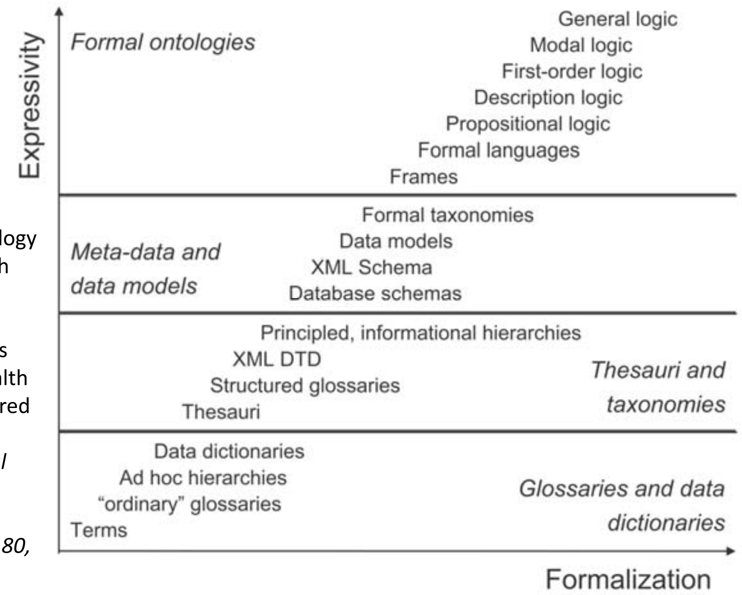
Edwin T. Jaynes 2003. Probability theory: The logic of science, Cambridge, Cambridge University Press.

What are 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

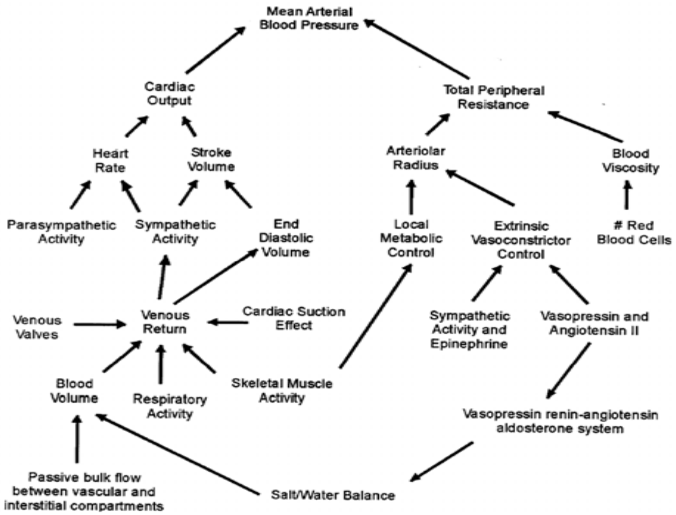
Davis, R., Shrobe, H. , Szolovits, P. 1993 What is a knowledge representation? AI Magazine, 14, 1, 17-33.

What does Formalization versus Expressivity mean ?



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

What do you need for developing clinical decision support systems?



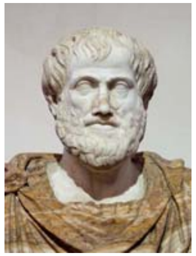
Hajdukiewicz, J. R., Vicente, K. J., Doyle, D. J., Milgram, P. & Burns, C. M. (2001) Modeling a medical environment: an ontology for integrated medical informatics design. *International Journal of Medical Informatics, 62, 1, 79-99.*

06 Ontologies

What happens if you put the word “Jaguar” into a search engine ?

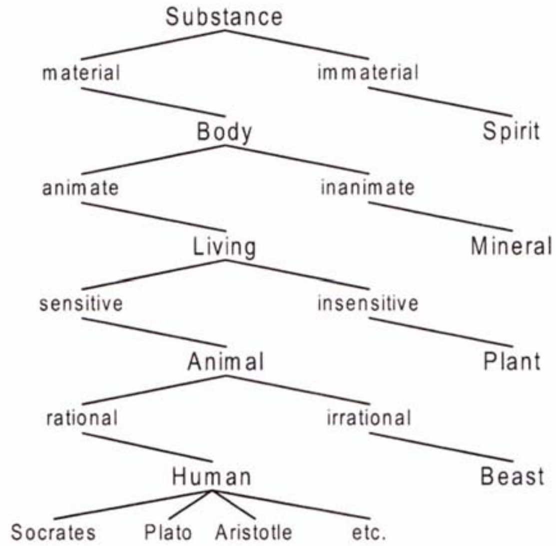
Image Sources: The images are in the public domain and are used according UrhG §42 lit. f Abs 1 as “Belegfunktion” for discussion with students

Who created the first ontology ?



* 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.*



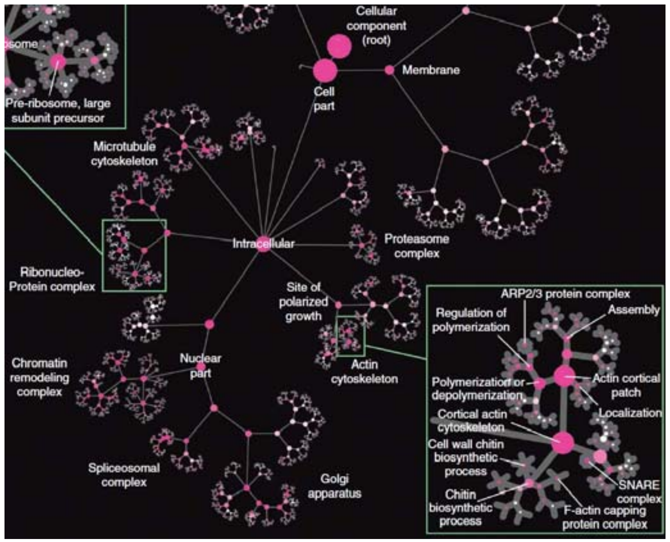
Later: Porphyry (≈ 234-305) tree

What is the classic definition of an ontology ?

- 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.**

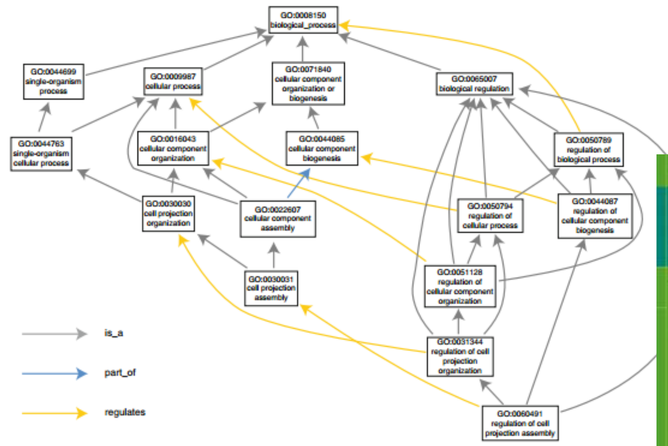
Studer, R., Benjamins, V. R. & Fensel, D. (1998) Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering, 25, 1-2, 161-197.*

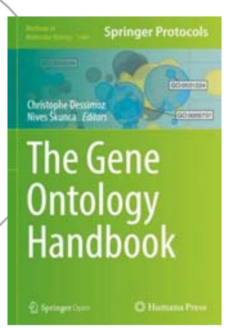
Where are ontologies used today ?



Janusz Dutkowski, Michael Kramer, Michal A Surma, Rama Balakrishnan, J Michael Cherry, Nevan J Krogan & Trey Ideker 2013. A gene ontology inferred from molecular networks. Nature biotechnology, 31, (1), 38.

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 = 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** ...

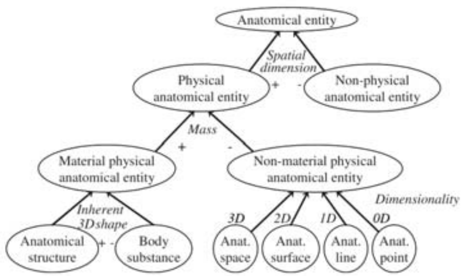
(1) In addition to the IS-A relationship, partitive (meronomic) relationships may hold between concepts, denoted by PART-OF. Every PART-OF relationship is irreflexive, asymmetric and transitive. IS-A and PART-OF are also called hierarchical relationships.

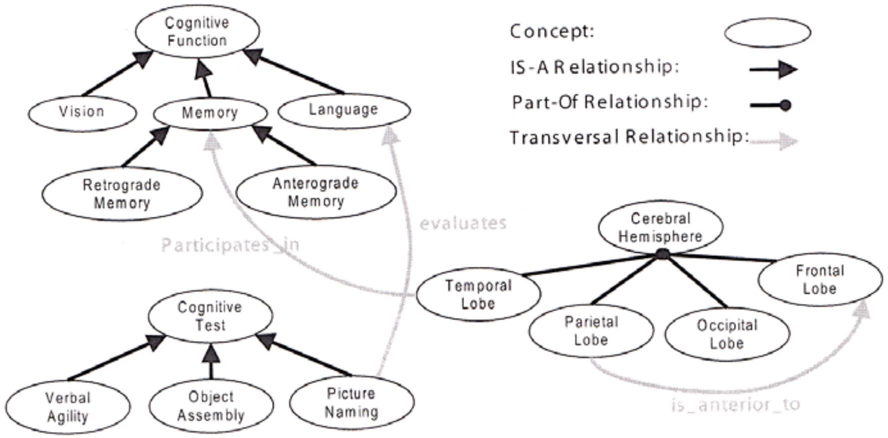
(2) In addition to hierarchical relationships, associative relationships may hold between concepts. Some associative relationships are domain-specific (e.g., the branching relationship between arteries in anatomy and rivers in geography).

(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.*





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.*

What are typical medical 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.*

What notations of ontologies do we use ?

- **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**

How does a graphical notation look like ?

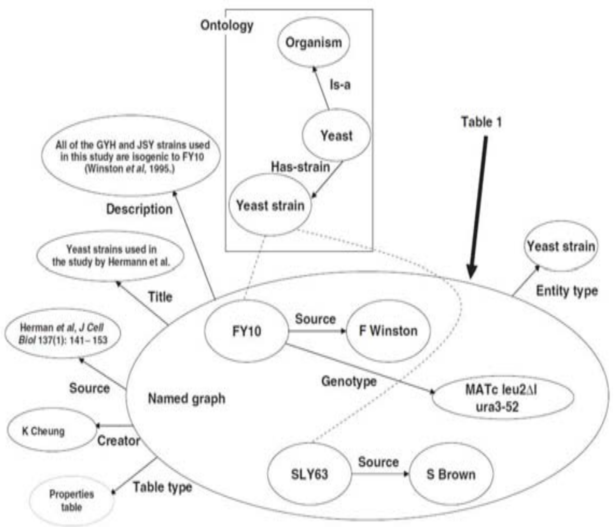


Table 1 Yeast strains used in the study by Hermann et al (1997)

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.*

What is the purpose of the Web Ontology Language OWL ?

DL = Description Logic

Concept inclusion, Speak: All C1 are C2

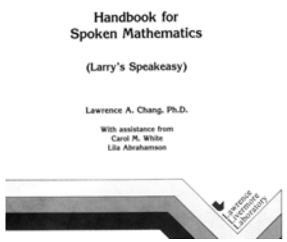
Concept equivalence Speak: C1 is equivalent to C2

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

How do you pronounce all these math expressions ?

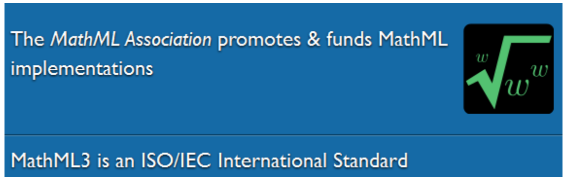
web.efzg.hr/dok/MAT/vkojic/Larrys_speakeasy.pdf



HELPFUL: https://en.wikipedia.org/wiki/List_of_mathematical_symbols

LaTeX Symbols : http://www.artofproblemsolving.com/wiki/index.php/LaTeX:Symbols

Math ML: http://www.robinlionheart.com/stds/html4/entities-mathml



What are ontological class constructors ?

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

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

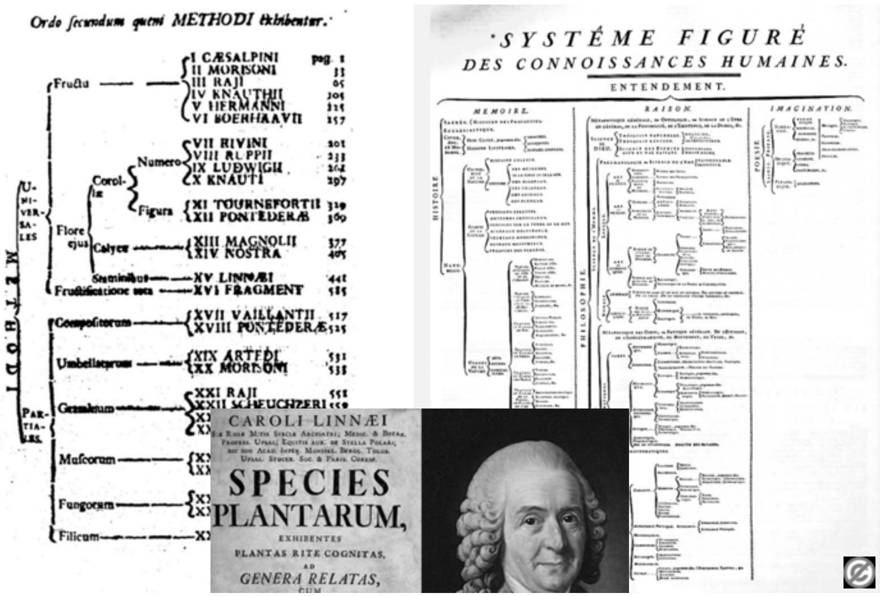
Universal Restriction
Speak: All P-successors are in C

Existential Restriction
Speak: An P-successor exists in C

Bhatt et al. (2009)

07 Medical Classifications

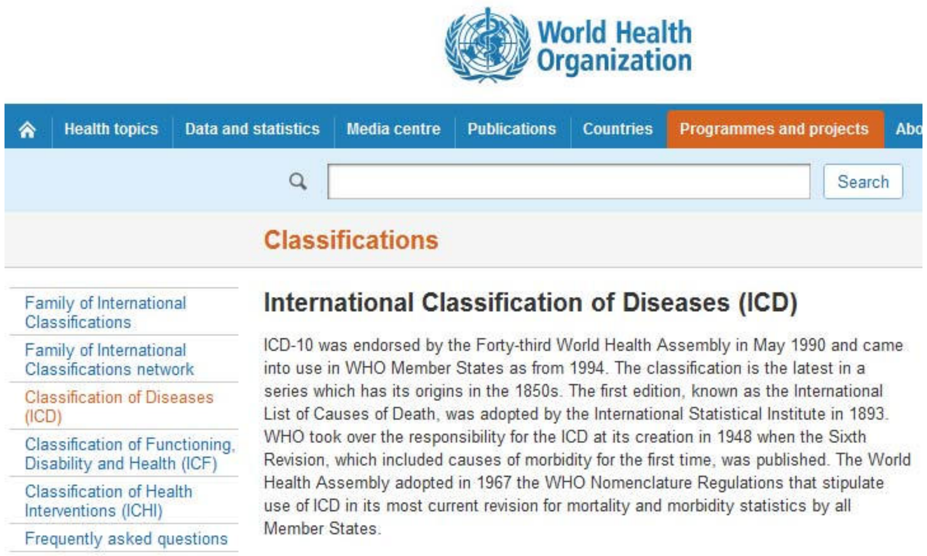
What is classification generally ?



What medical classification systems do we know ?

- 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)

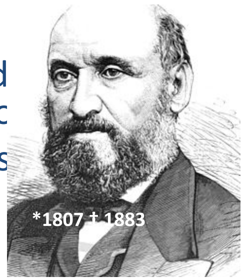
What is the International Classification of Diseases (ICD) ?



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

How did the International Classification of Diseases evolve ?

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



What is 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

How does Hypertension look in SNOMED ?

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.*

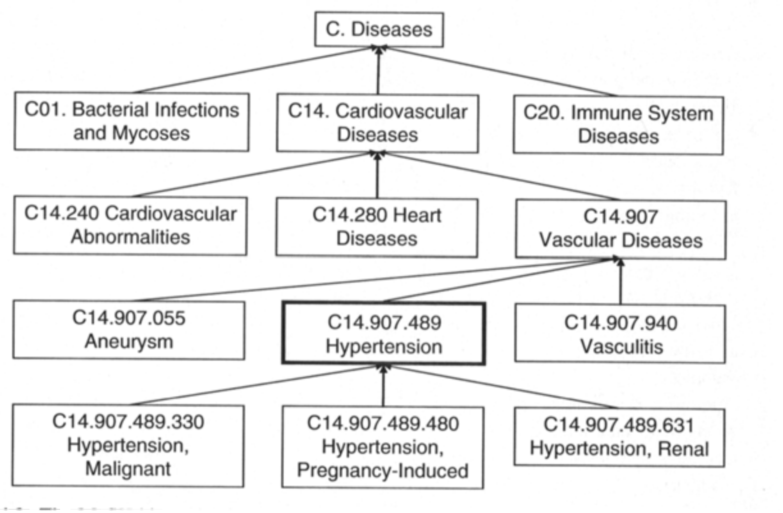
What is MeSH ?

- 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.

What are 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]

How does the MeSH hierarchy look ?



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

How does the MeSH Example Hypertension look ?

National Library of Medicine - Medical Subject Headings

2011 MeSH

MeSH Descriptor Data

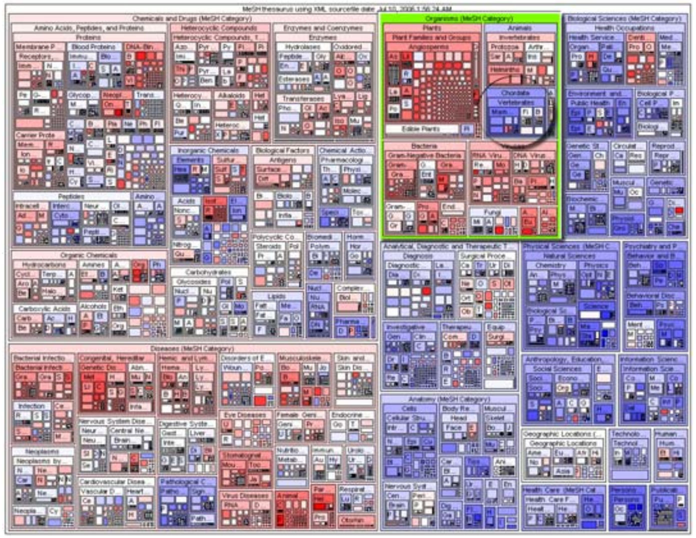
Return to Entry Page

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

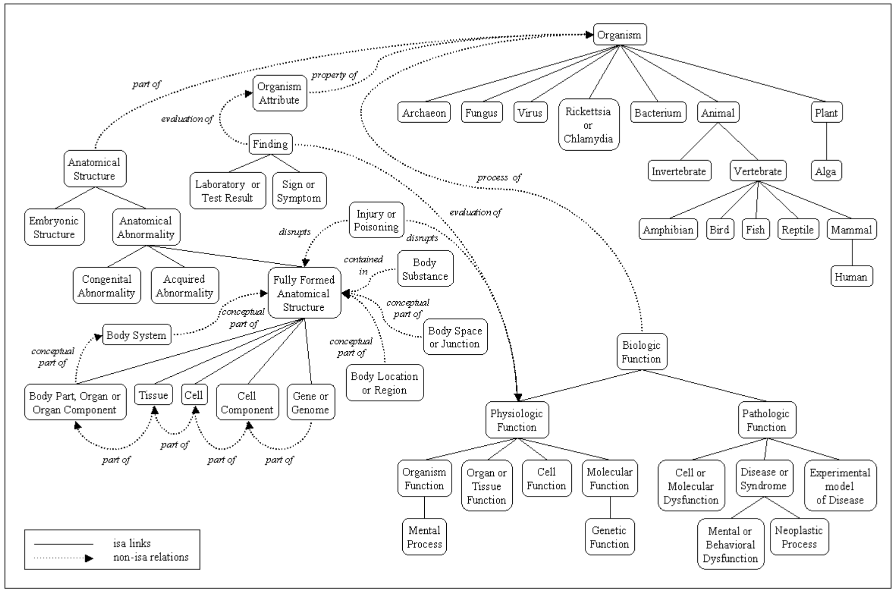
http://www.nlm.nih.gov/mesh/

How does Mesh look in an interactive tree visualization ?

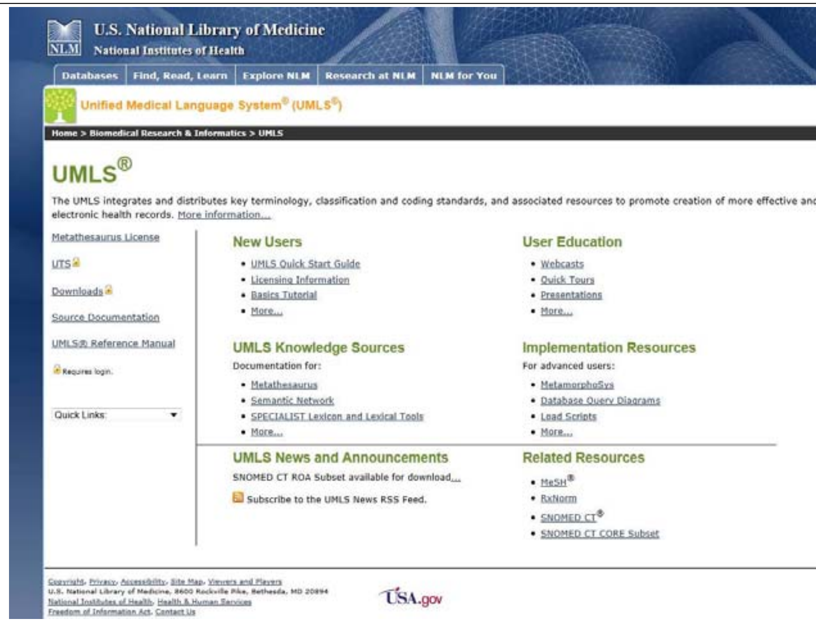


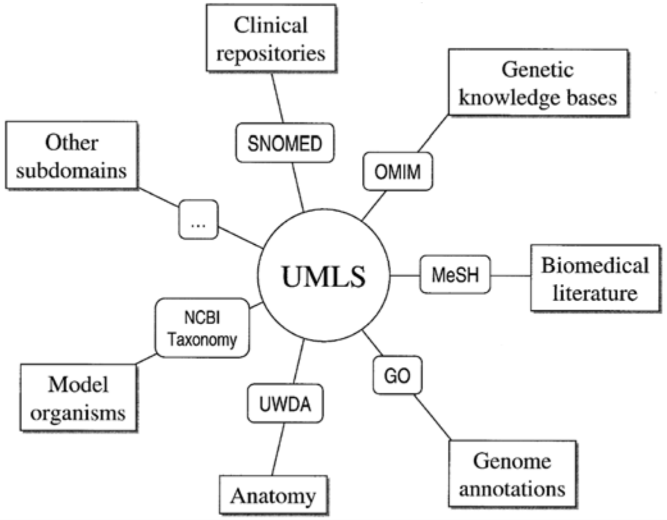
Eckert, K. (2008) A methodology for supervised automatic document annotation. *Bulletin of IEEE Technical Committee on Digital Libraries TCDL, 4, 2.*

What is UMLS – Unified Medical Language System ?

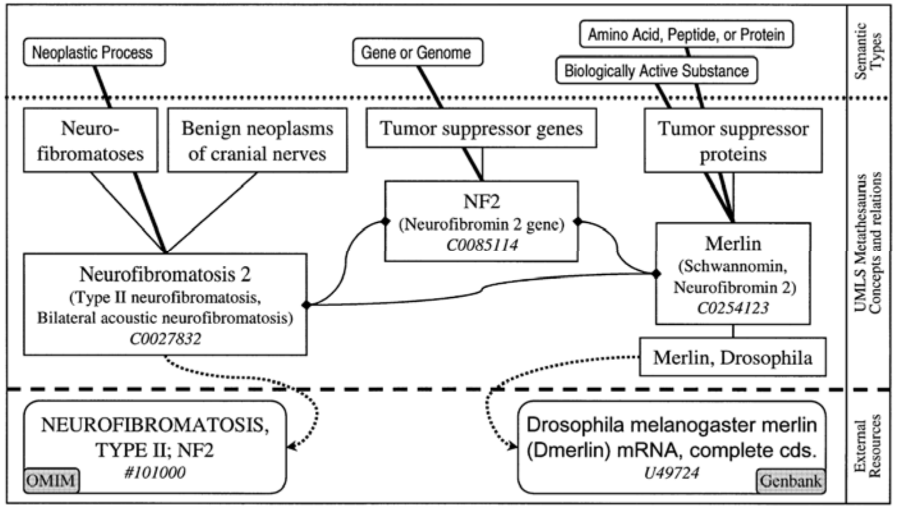


http://www.nlm.nih.gov/research/umls/





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



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

Concluding remark

- Progress in machine learning is driven by the explosion in the availability of **big data** and **low-cost computation …**
- **Health is amongst the biggest challenges**

Jordan, M. I. & Mitchell, T. M. 2015. Machine learning: Trends, perspectives, and prospects. Science, 349, (6245), 255-260.

ULTRA-MODERN MEDICINE: EXAMPLES OF MACHINE LEARNING IN HEALTHCARE

July 4, 2019 · Updated: March 25, 2020 Written by Mike Thomas

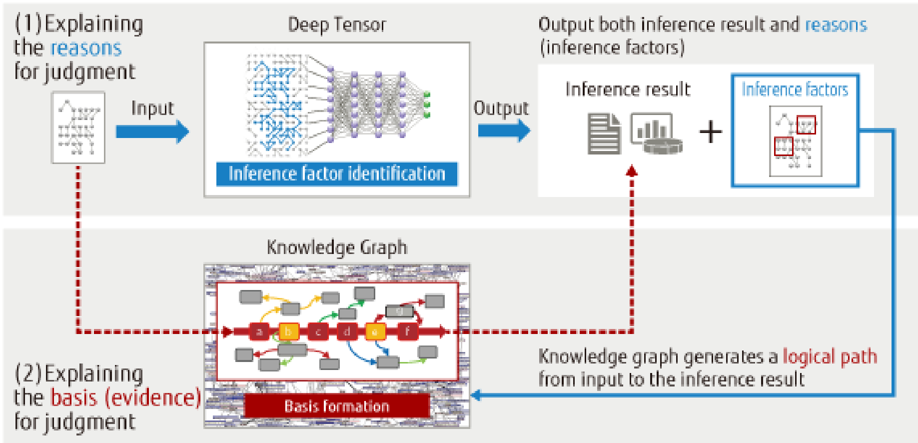
Conclusion

- 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.

- 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.



Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg & Andreas Holzinger 2018. Explainable AI: the new 42? Springer Lecture Notes in Computer Science LNCS 11015

Thank you!