Assoc.Prof. Dr. Andreas Holzinger

185.A83 Machine Learning for Health Informatics
2020S, VU, 2.0 h, 3.0 ECTS
Andreas Holzinger, Marcus Bloice, Florian Endel, Anna Saranti
Lecture 02 - Week 13

From data to probabilistic information and knowledge

Contact: andreas.holzinger AT tuwien.ac.at

https://human-centered.ai/machine-learning-for-health-informatics-class-2020

- 00 Reflection
- 01 **Data** – the underlying physics of data
- 02 Biomedical data sources – taxonomy of data
- 03 Data integration, mapping, fusion
- 04 **Information** -Theory – Entropy
- 05 **Knowledge** Representation – Ontologies – Medical Classifications

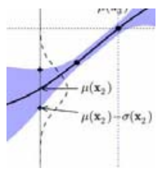
00 Reflection

1

2

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

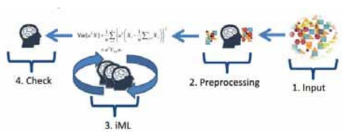
3



4

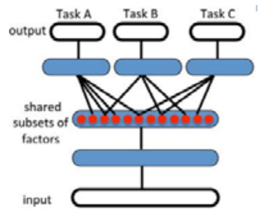
5

6



7

8



9

Image source: http://www.efmc.info/medchemwatch-2014-1/lab.php
This image is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

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

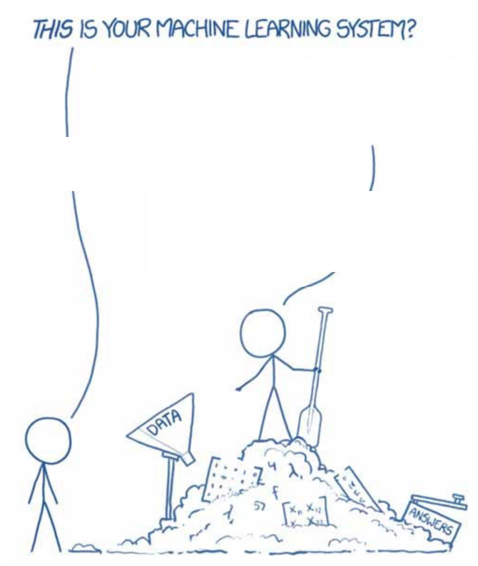


Image Source: Randall Munroe https://xkcd.com
This image is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

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

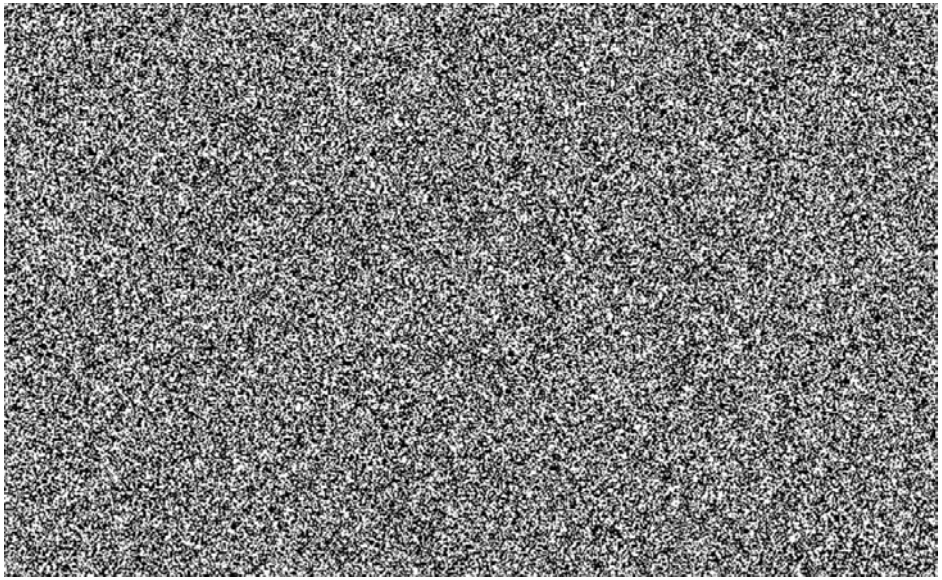
Leo L. Pipino, Yang W. Lee & Richard Y. Wang 2002. Data quality assessment. Communications of the ACM, 45, (4), 211-218.

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

Mark D. Wilkinson, Michel Dumontier, Ijsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J. G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A. C. 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao & Barend Mons 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

01 The underlying physics of data



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]

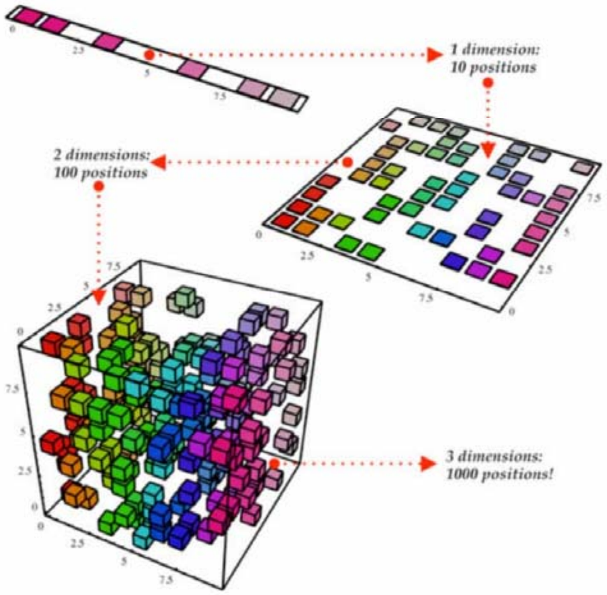
1. Holzinger A, Dehmer M, & Jurisica I (2014) Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions. BMC Bioinformatics 15(S6):I1.
2. Hund, M., Sturm, W., Schreck, T., Ullrich, T., Keim, D., Majnaric, L. & Holzinger, A. 2015. Analysis 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 ?

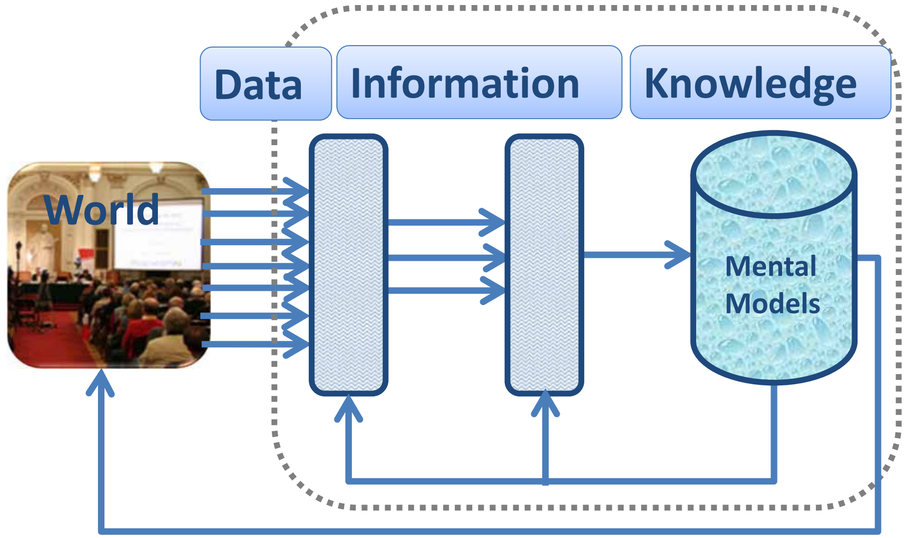
- Data in traditional Statistics
- Low-dimensional data (< $\mathbb{R}^{100}$)
- Problem: Much noise in the data
- Not much structure in the data but it can be represented by a simple model

- 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



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.

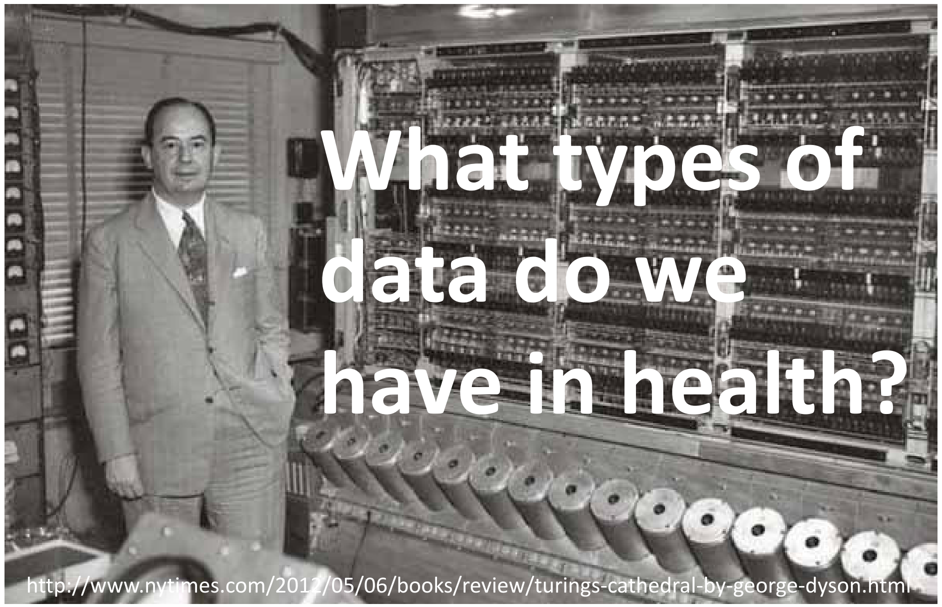


Knowledge := a set of expectations

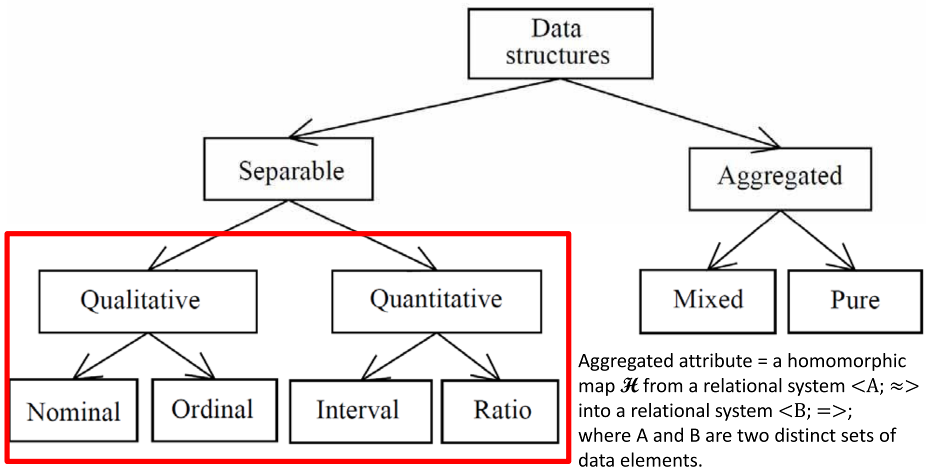
AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

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.



How can we classify traditional data structures ?

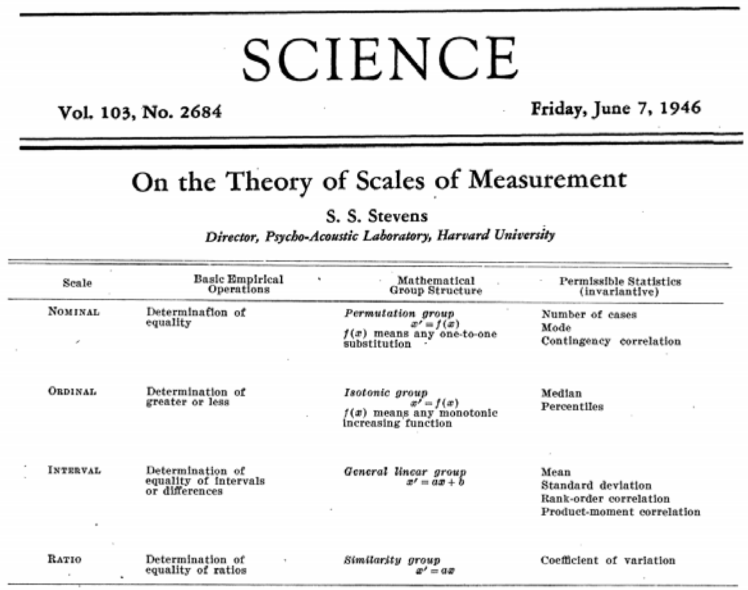


Aggregated attribute = a homomorphic map $\mathcal{H}$ from a relational system <A; ≈> into a relational system <B; =>; where A and B are two distinct sets of data elements.
This is in contrast with other attributes since the set B is the set of data elements instead of atomic values.

Dastani, M. (2002) The Role of Visual Perception in Data Visualization. *Journal of Visual Languages and Computing, 13, 601-622.*

What was the original work on the theory of scales claiming ?

Stanley S. Stevens 1946. On the theory of scales of measurement. Science, 103, (2684), 677-680.



SCIENCE

Vol. 103, No. 2684 Friday, June 7, 1946

On the Theory of Scales of Measurement

S. S. Stevens

Director, Psycho-Acoustic Laboratory, Harvard University

Scale	Basic Empirical Operations	Mathematical Group Structure	Permissible Statistics (invariantive)
NOMINAL	Determination of equality	*Permutation group* $x' = f(x)$ $f(x)$ means any one-to-one substitution	Number of cases Mode Contingency correlation
ORDINAL	Determination of greater or less	*Isotonic group* $x' = f(x)$ $f(x)$ means any monotonic increasing function	Median Percentiles
INTERVAL	Determination of equality of intervals or differences	*General linear group* $x' = ax + b$	Mean Standard deviation Rank-order correlation Product-moment correlation
RATIO	Determination of equality of ratios	*Similarity group* $x' = ax$	Coefficient of variation

What properties do separable data have ?

Scale	Empirical Operation	Mathem. Group Structure	Transf. in ℝ	Basic Statistics	Mathematical Operations
NOMINAL	Determination of equality	Permutation *x' = f(x)* *x … 1-to-1*	*x ↦ f(x)*	Mode, contingency correlation	=, ≠
ORDINAL	Determination of more/less	Isotonic *x' = f(x)* *x … mono-tonic incr.*	*x ↦ f(x)*	Median, Percentiles	=, ≠, >, <
INTERVAL	Determination of equality of intervals or differences	General linear *x' = ax + b*	*x ↦ rx+s*	Mean, Std.Dev. Rank-Order Corr., Prod.-Moment Corr.	=, ≠, >, <, -, +
RATIO	Determination of equality or ratios	Similarity *x' = ax*	*x ↦ rx*	Coefficient of variation	=, ≠, >, <, -, +, *, ÷

What levels of data taxonomy can we identify ?

- **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, 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.

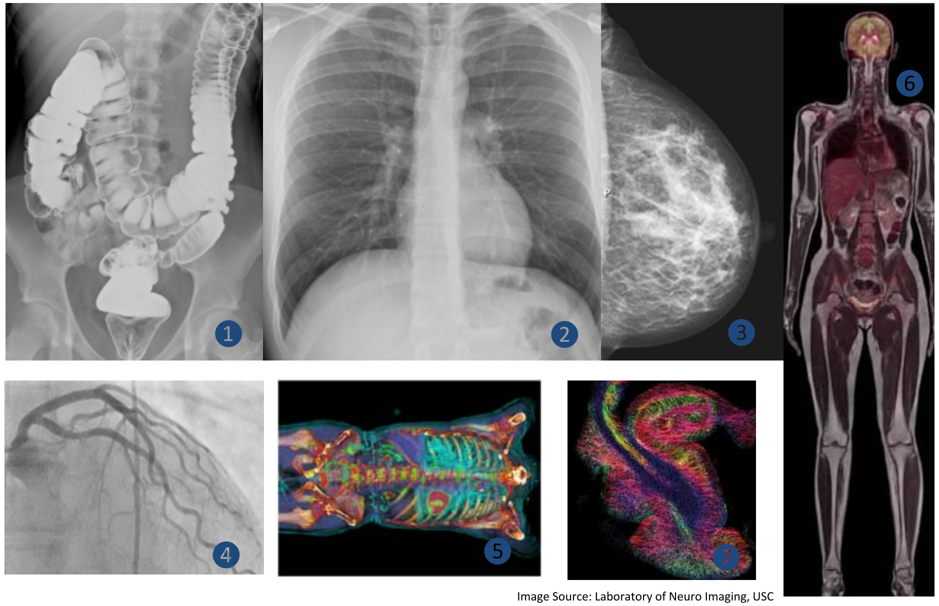
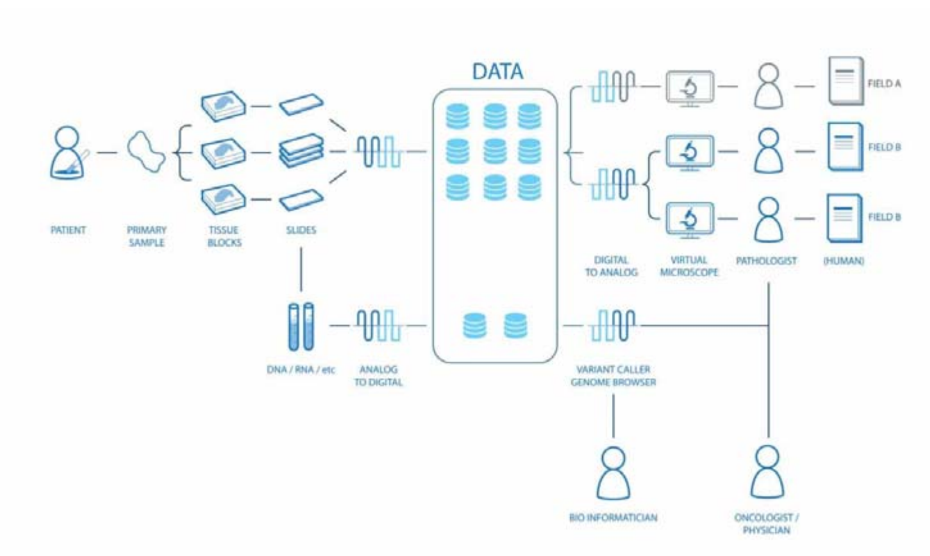
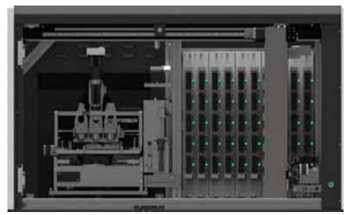
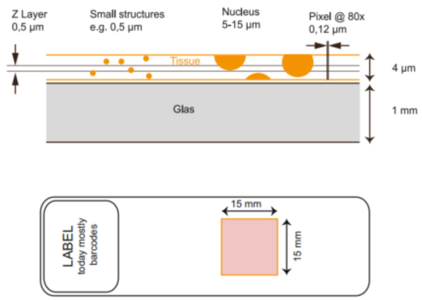


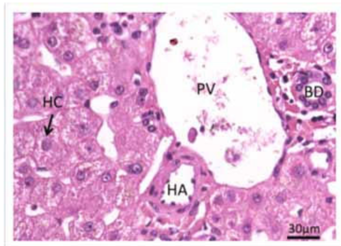
Image Source: Laboratory of Neuro Imaging, USC



Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs & Kurt Zatloukal 2017. Towards the Augmented Pathologist: Challenges of Explainable-AI in Digital Pathology. arXiv:1712.06657.

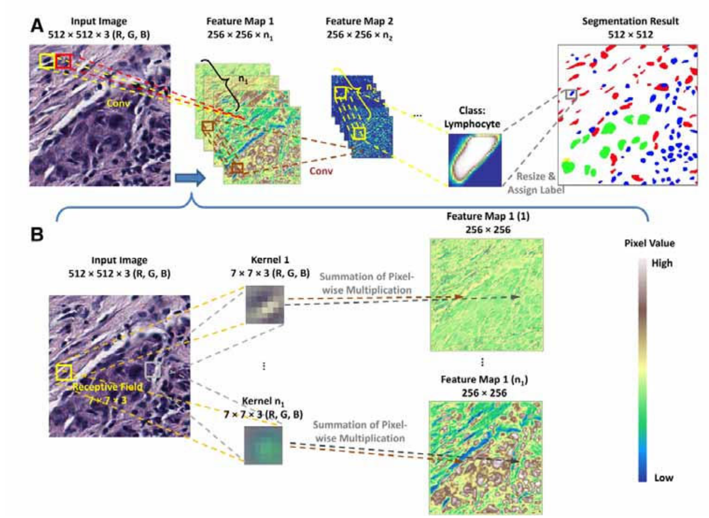




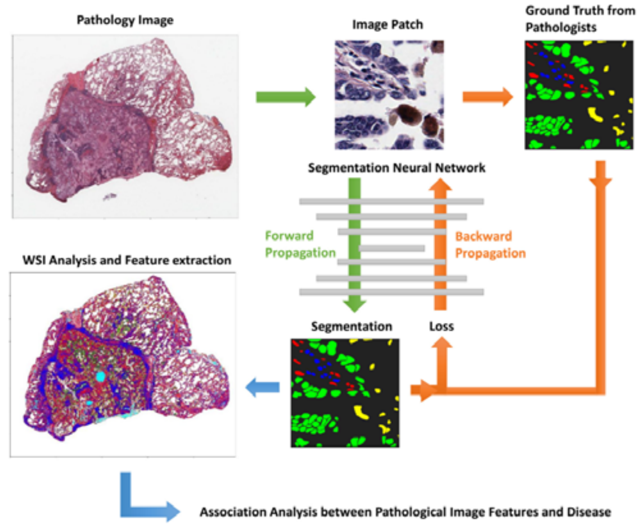


(Image Sources: Pathology Graz)

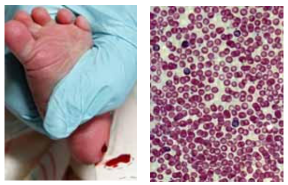
Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Müller, Robert Reihs & Kurt Zatloukal 2017. Machine Learning and Knowledge Extraction in Digital Pathology needs an integrative approach. Towards Integrative Machine Learning and Knowledge Extraction, Springer Lecture Notes in Artificial Intelligence Volume LNAI 10344. Cham: Springer, pp. 13-50, doi:10.1007/978-3-319-69775-8_2.



Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan & Guanghua Xiao 2019. Pathology image analysis using segmentation deep learning algorithms. The American journal of pathology, 189, (9), 1686-1698, doi:10.1016/j.ajpath.2019.05.007

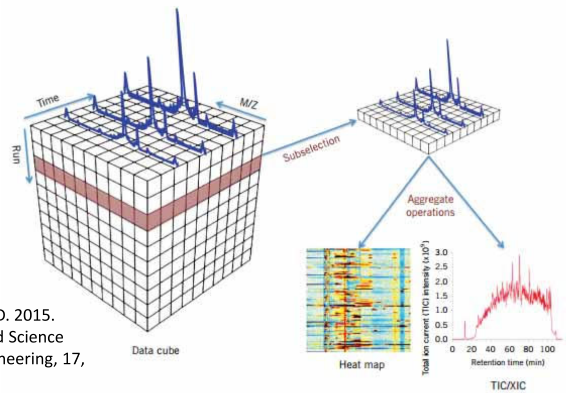


Shidan Wang, Donghan M Yang, Ruichen Rong, Xiaowei Zhan & Guanghua Xiao 2019. Pathology image analysis using segmentation deep learning algorithms. The American journal of pathology, 189, (9), 1686-1698, doi:10.1016/j.ajpath.2019.05.007

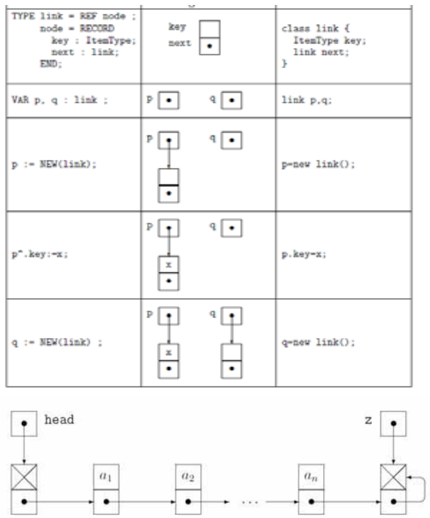


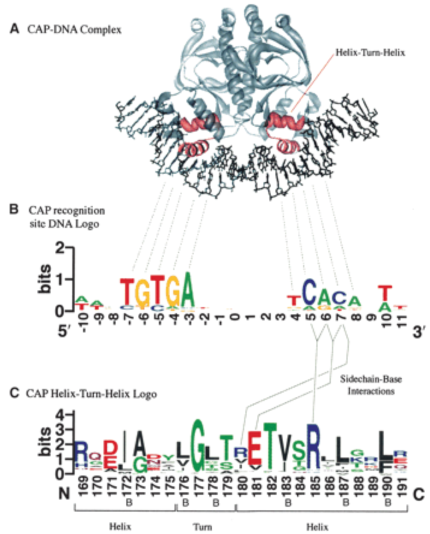
Amino acids (symbols)	Fatty acids (symbols)	Fatty acids (symbols)
Alanine (Ala)	Free carnitine (C0)	Hexadecenoyl-carnitine (C16:1)
Arginine (Arg)	Acetyl-carnitine (C2)	Octadecenoyl-carnitine (C18:1)
Argininosuccinate (Argsuc)	Propionyl-carnitine (C3)	Decenoyl-carnitine (C10:2)
Citrulline (Cit)	Butyryl-carnitine (C4)	Tetradecadienoyl-carnitine (C14:2)
Glutamate (Glu)	Isovaleryl-carnitine (C5)	Octadecadienoyl-carnitine (C18:2)
Glycine (Gly)	Hexanoyl-carnitine (C6)	Hydroxy-isovaleryl-carnitine (C5-OH)
Methionine (Met)	Octanyl-carnitine (C8)	Hydroxytetradecadienoyl-carnitine (C14-OH)
Ornitine (Orn)	Decanoyl-carnitine (C10)	Hydroxypalmitoyl-carnitine (C16-OH)
Phenylalanine (Phe)	Dodecanoyl-carnitine (C12)	Hydroxypalmitoleyl-carnitine (C16:1-OH)
Pyroglutamate (Pyrglt)	Myristoyl-carnitine (C14)	Hydroxyoleyl-carnitine (C18:1-OH)
Serine (Ser)	Hexadecanoyl-carnitine (C16)	Dicarboxyl-butyryl-carnitine (C4-DC)
Tyrosine (Tyr)	Octadecanoyl-carnitine (C18)	Glutaryl-carnitine (C5-DC)
Valine (Val)	Tiglyl-carnitine (C5:1)	Methylglutaryl-carnitine (C6-DC)
Leucine + Isoleucine (Xle)	Decenoyl-carnitine (C10:1)	Methylmalonyl-carnitine (C12-DC)
	Myristoleyl-carnitine (C14:1)	

Fourteen amino acids and 29 fatty acids are analyzed from a single blood spot using MS/MS. The concentrations are given in μmol/L.



Yao, Y., Bowen, B. P., Baron, D. & Poznanski, D. 2015. SciDB for High-Performance Array-Structured Science Data at NERSC. Computing in Science & Engineering, 17, (3), 44-52, doi:10.1109/MCSE.2015.43.

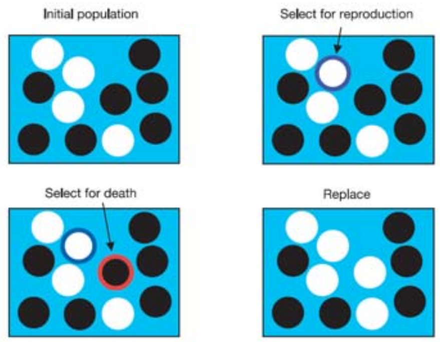


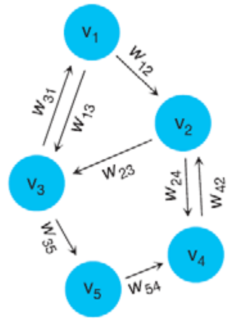


Crooks, G. E., Hon, G., Chandonia, J. M. & Brenner, S. E. (2004) WebLogo: A sequence logo generator. Genome Research, 14, 6, 1188-1190.

Why is the data structure graph so versatile ?

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



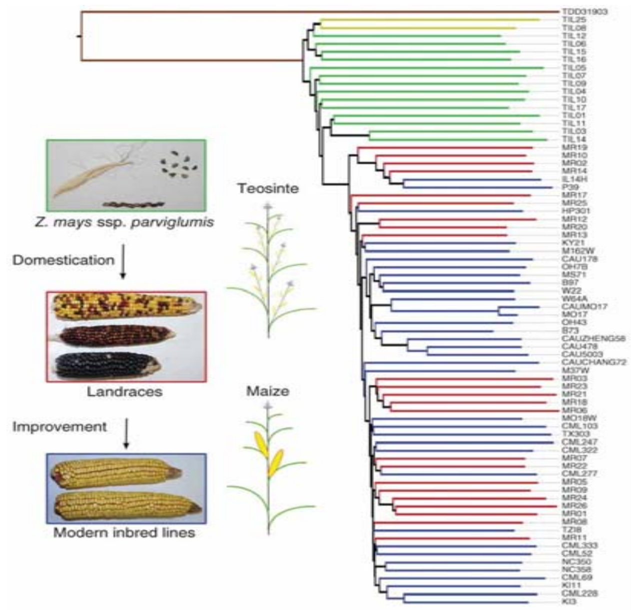


$$W=\begin{bmatrix} 0 & w_{12} & w_{13} & 0 & 0 \\ 0 & 0 & w_{23} & w_{24} & 0 \\ w_{31} & 0 & 0 & 0 & w_{35} \\ 0 & w_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & w_{54} & 0 \end{bmatrix}$$

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

What is a tree ?

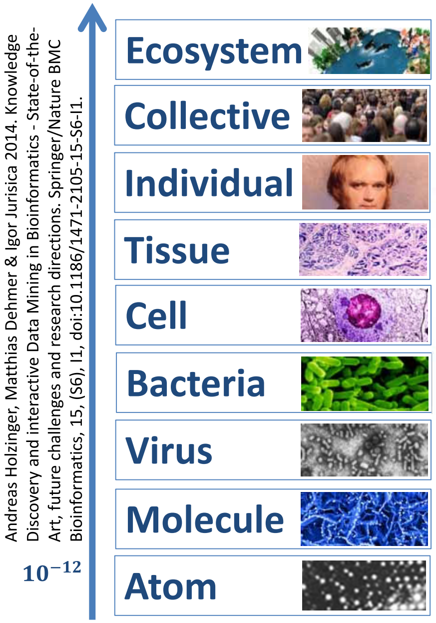
Hufford et. al. 2012. Comparative population genomics of maize domestication and improvement. *Nature Genetics, 44, (7), 808-811.*

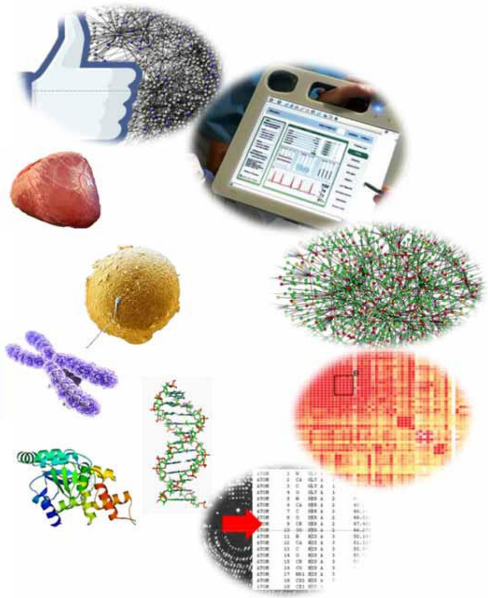


02 Biomedical data sources: Taxonomy of data

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.



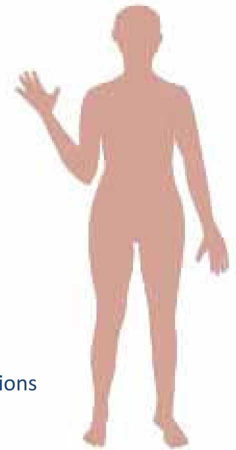


Why is data integration in health an unsolved problem ?

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)

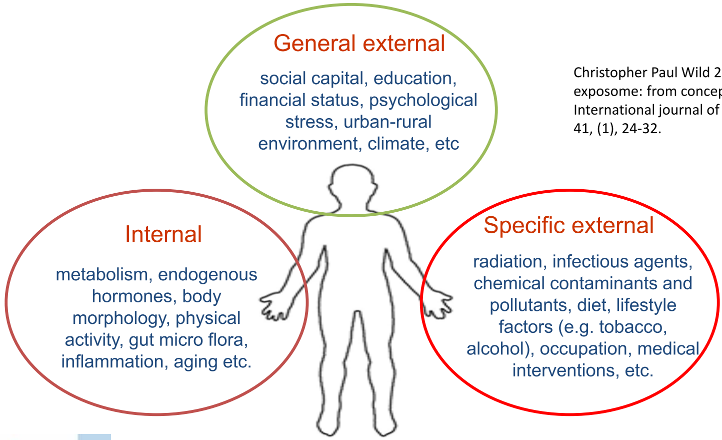
Proteomics
Protein-Protein Interactions

Epigenetics
Epigenetic modifications

Genomics

Transcriptomics
RNA, mRNA, rRNA, tRNA

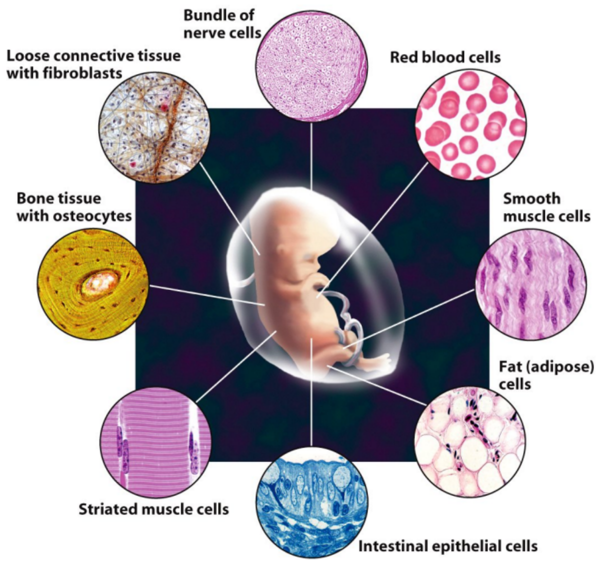
Why is the human exposome important ?



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

What is a good example for the level "cell" ?



Karp, G. 2010. Cell and Molecular Biology: Concepts and Experiments, Gainesville, John Wiley.

What is life according to Erwin Schrödinger ?

to reproduce …

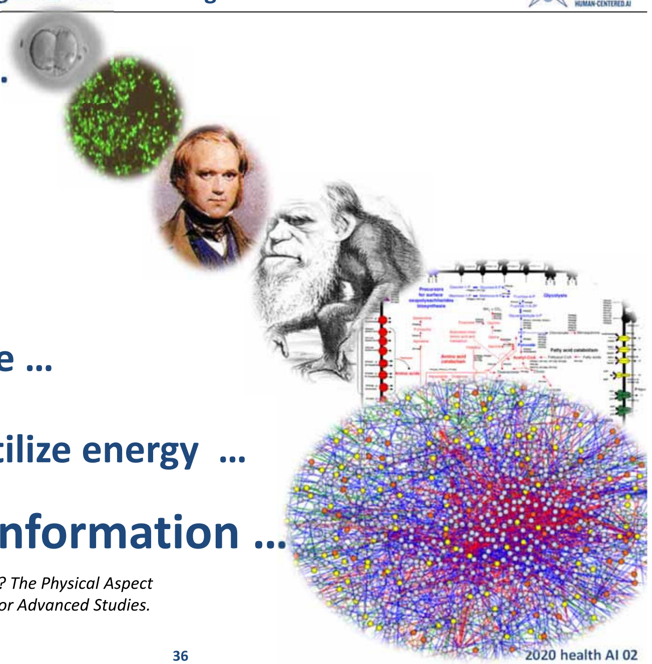
to grow …

to evolve …

to self-replicate …

to generate/utilize energy …

to process information …



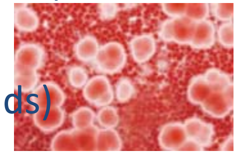
Schrödinger, E. (1944) *What Is Life? The Physical Aspect of the Living Cell. Dublin Institute for Advanced Studies.*

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 are the most important *omics data ?

- **Genom**ics (sequence annotation)
- **Transcript**omics (microarray)
- **Prote**omics (Proteome Databases)
- **Metabol**omics (enzyme annotation)
- **Flux**omics (isotopic tracing, metabolic pathways)
- **Phen**omics (biomarkers)
- **Epigen**omics (epigenetic modifications)
- **Microbi**omics (microorganisms)
- **Lipid**omics (pathways of cellular lipids)



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[109]	• Gene-regulatory networks[76]	• Functional annotation[89] • Protein complex identification[82]		• Functional annotation[102]
		Proteomics (abundance, post-translational modification)	• Enzyme annotation[99]	• 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.***

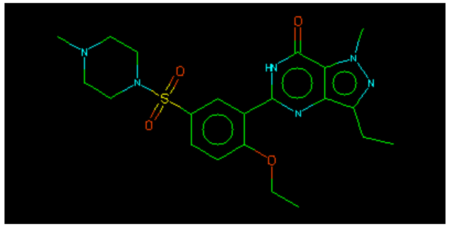
Examples for lower dimensional data ?

- 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

SMILES (Simplified Molecular Input Line Entry Specification)

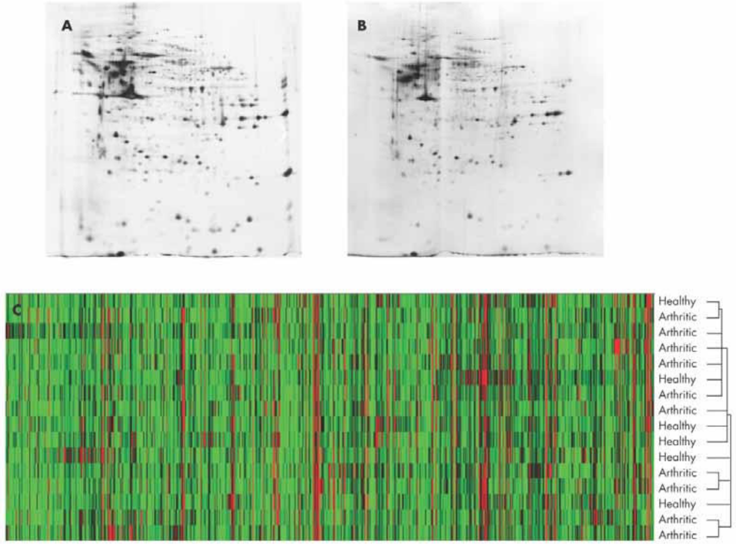
... is a compact machine and human-readable chemical nomenclature:

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

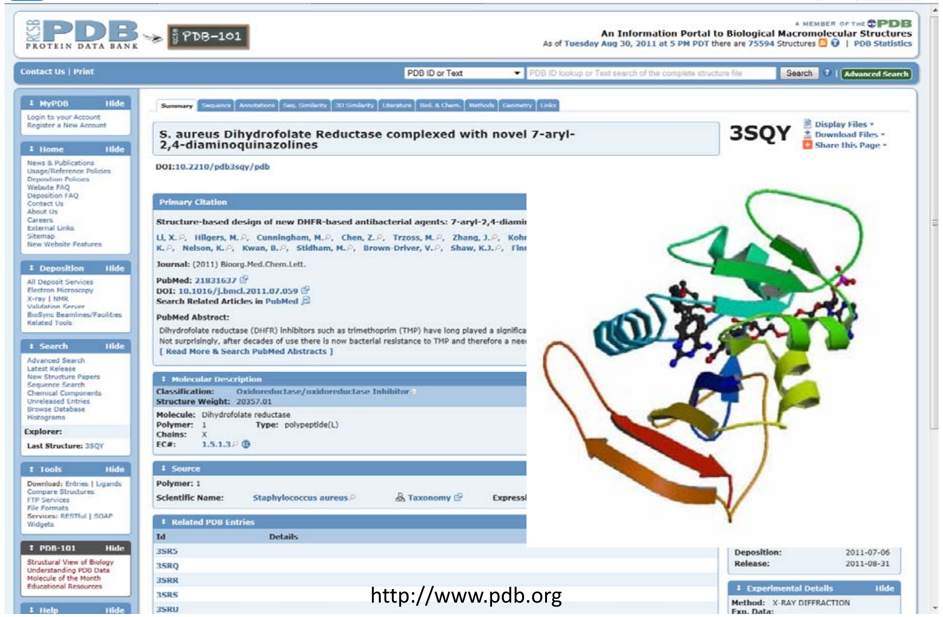


...is Canonicalizable
...is Comprehensive
...is Well Documented

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

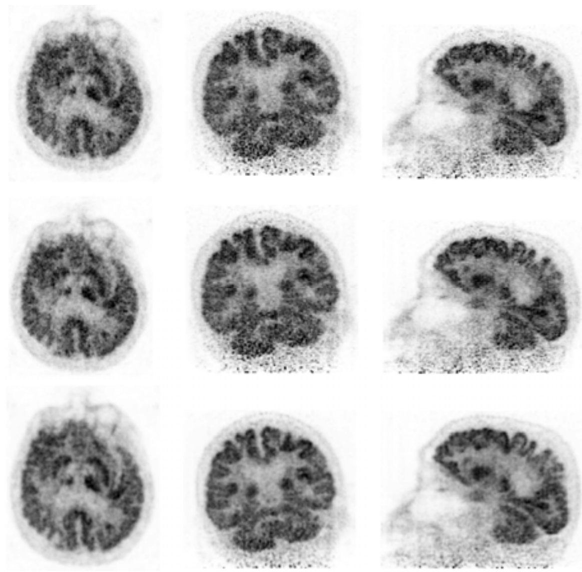


Kastrinaki et al. (2008) Functional, molecular & proteomic characterisation of bone marrow mesenchymal stem cells in rheumatoid arthritis. *Annals of Rheumatic Diseases, 67, 6, 741-749.*



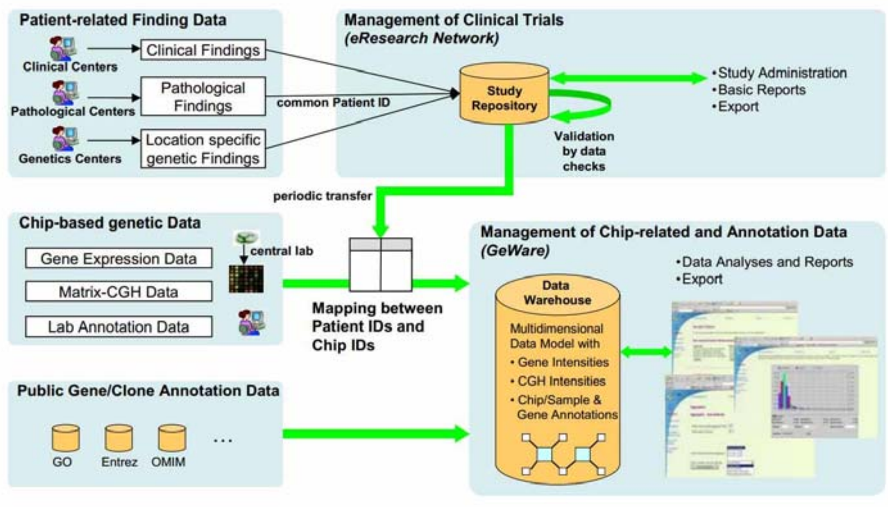
http://www.pdb.org

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

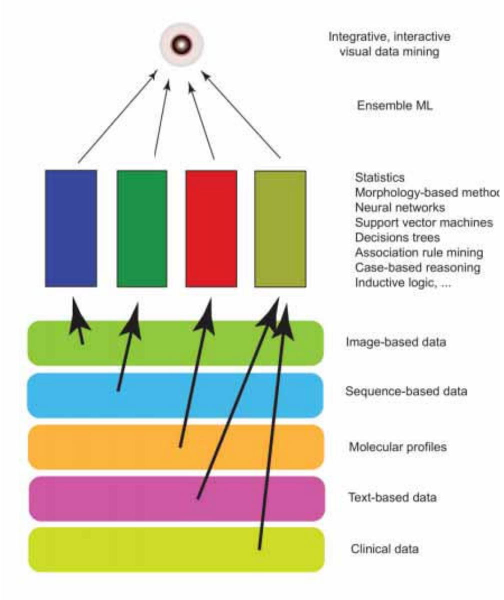
What do we mean with data integration ?



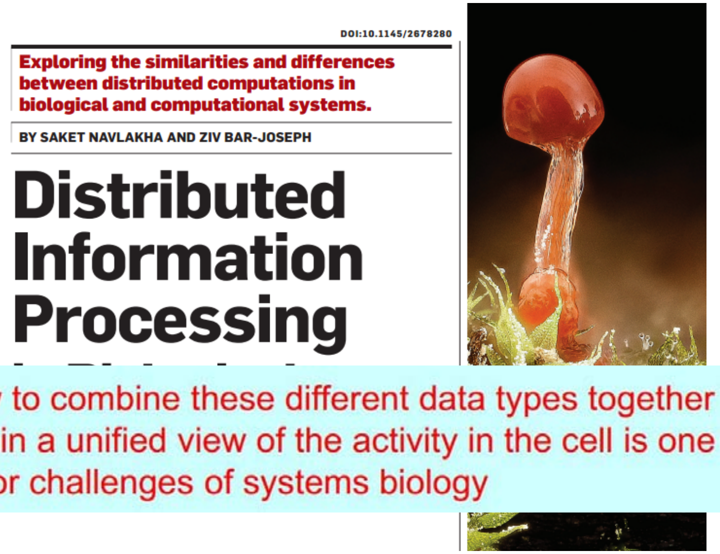
Kirsten, T., Lange, J. & Rahm, E. 2006. An integrated platform for analyzing molecular-biological data within clinical studies. Current Trends in Database Technology–EDBT 2006. Heidelberg: Springer, pp. 399-410, doi:10.1007/11896548_31.

What is the goal of data integration ?

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



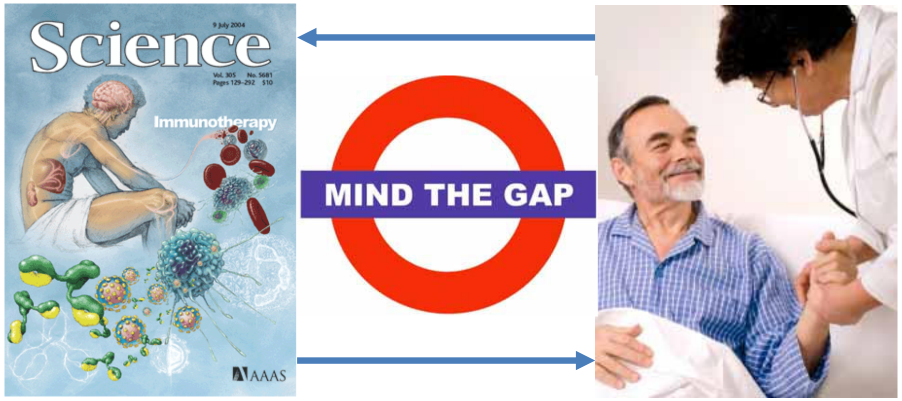
Holzinger, A. & Jurisica, I. 2014. Knowledge Discovery and Data Mining in Biomedical Informatics: The future is in Integrative, Interactive Machine Learning Solutions In: Lecture Notes in Computer Science LNCS 8401. Heidelberg, Berlin: Springer, pp. 1-18, doi:10.1007/978-3-662-43968-5_1.



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.

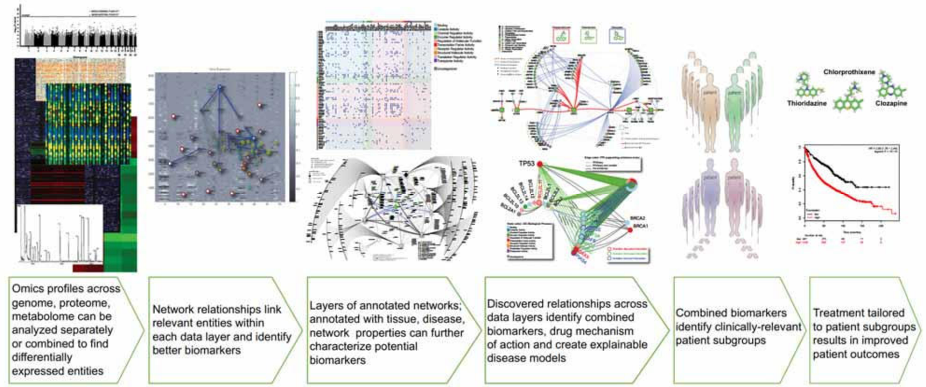
Why may information bridge the gap between both worlds ?



Our central hypothesis: Information may bridge this gap

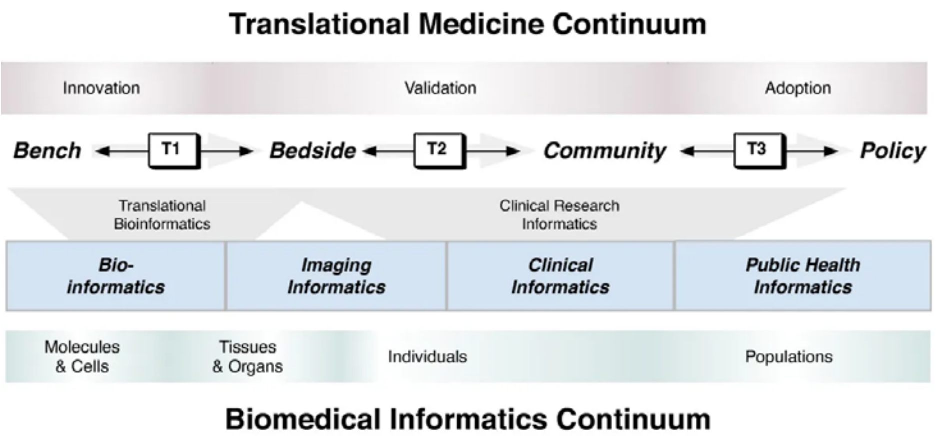
Holzinger, A. & Simonic, K.-M. (eds.) 2011. *Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058, Heidelberg, Berlin, New York: Springer.*

Why is imaging data along not enough ?



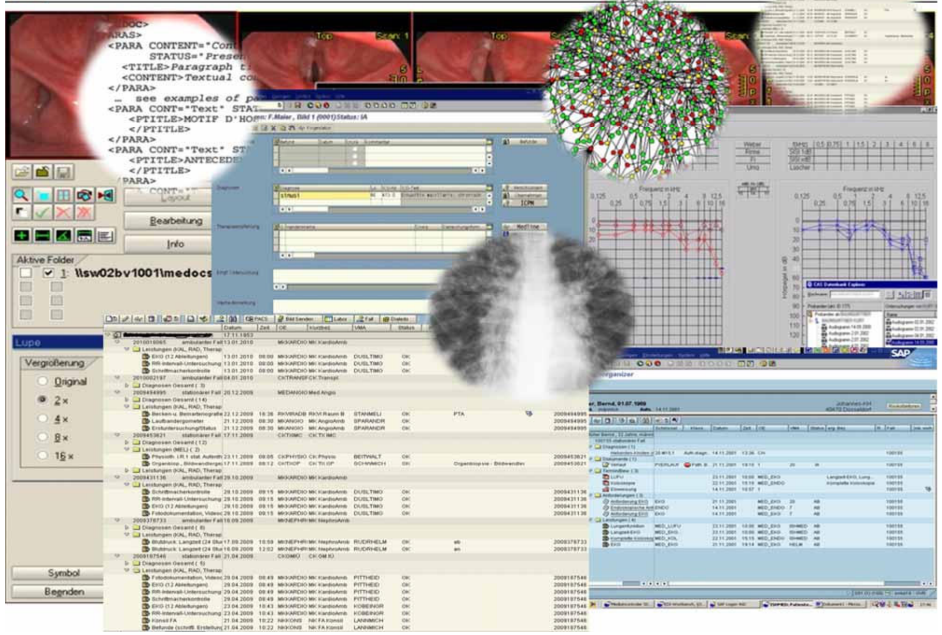
Andreas Holzinger, Benjamin Haibe-Kains & Igor Jurisica 2019. Why imaging data alone is not enough: AI-based integration of imaging, omics, and clinical data. European Journal of Nuclear Medicine and Molecular Imaging, 46, (13), 2722-2730, doi:10.1007/s00259-019-04382-9.

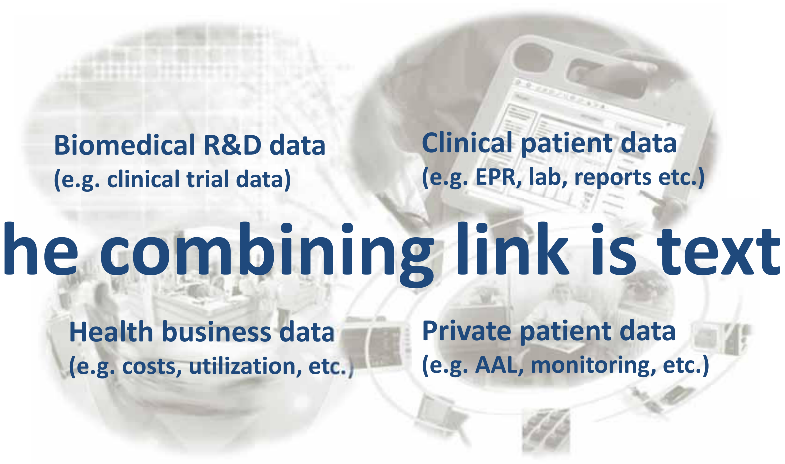
What is translational health ?



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 ?





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.*

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.*

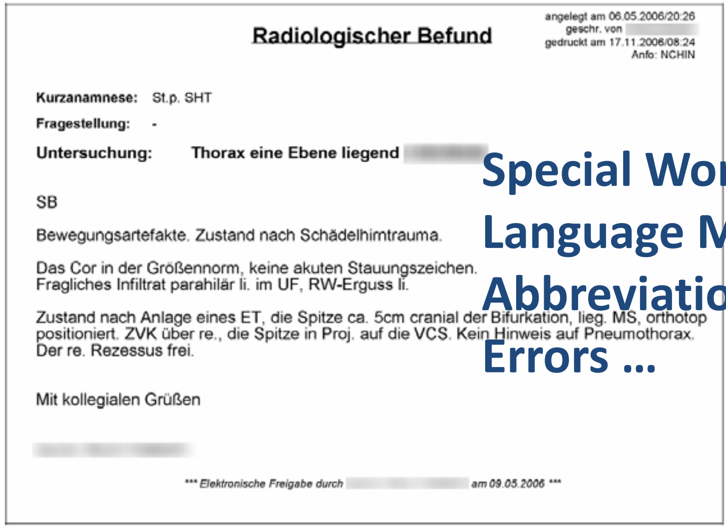
Digression: Medical Communication

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

Image Source: Andreas Holzinger

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.*

What are problems of the medical report ?



Radiologischer Befund

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

Kurzanamnese: St.p. SHT
Fragestellung: -
Untersuchung: **Thorax eine Ebene liegend**

SB

Bewegungsartefakte. Zustand nach Schädelhirntrauma.

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

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

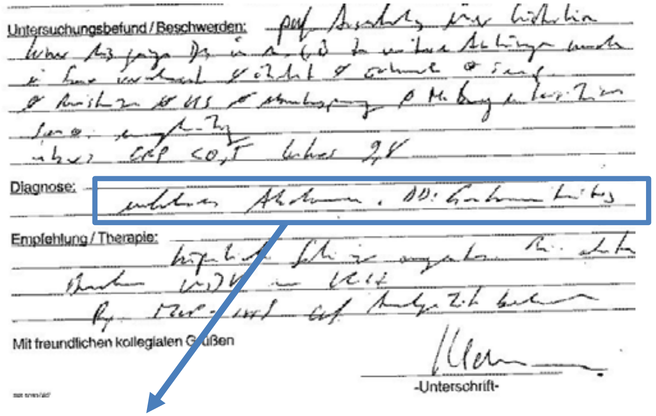
Mit kollegialen Grüßen

*** Elektronische Freigabe durch am 09.05.2006 ***

Special Words
Language Mix
Abbreviations
Errors …

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

Why is Synonymity and ambiguity such a huge problem ?



„die Antrumschleimhaut ist durch Lymphozyten infiltriert"
„lymphozytäre Infiltration der Antrummukosa"
„Lymphoyteninfiltration der Magenschleimhaut im Antrumbereich"

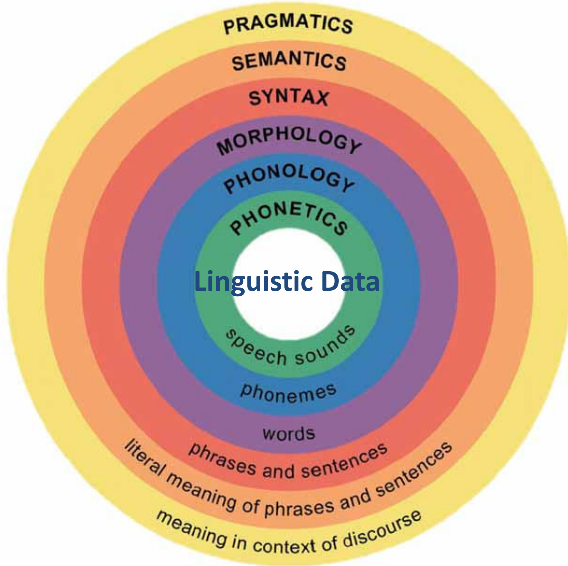
Why does Language Understanding require knowledge ?

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

"a young boy is holding a baseball bat."

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.
Image Source: https://cs.stanford.edu/people/karpathy/deepimagesent/

Why is text a 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.*

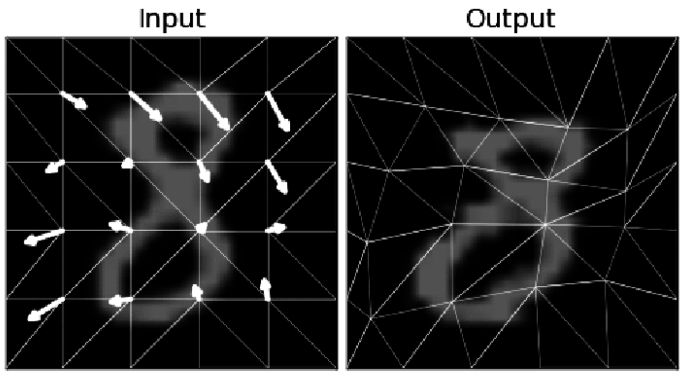
- Increasingly large data sets due to **data-driven medicine** [1]
- Increasing amounts of **non-standardized** data and **un-structured information** (e.g. “free text”)
- Data **quality,** data **integration**, universal **access**
- **Privacy,** security, safety, data protection, data ownership, fair 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.

Digression: 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.



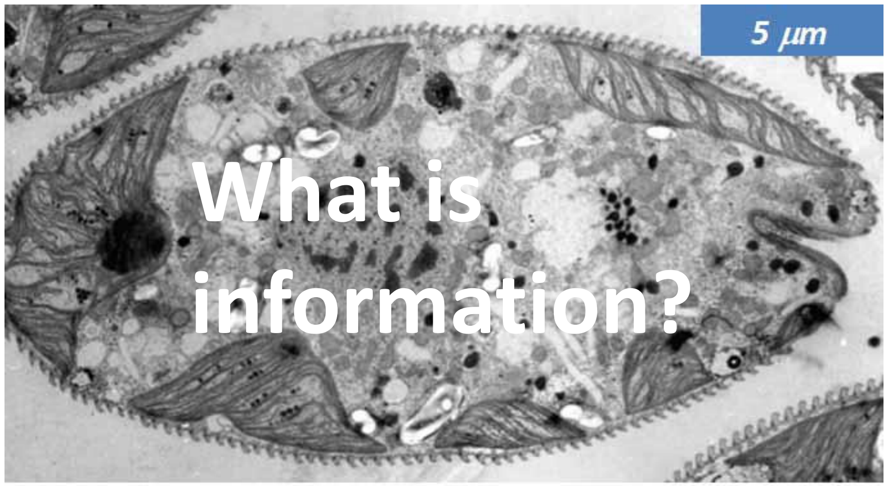
Marcus D Bloice, Christof Stocker & Andreas Holzinger 2017. Augmentor: an image augmentation library for machine learning. arXiv preprint arXiv:1708.04680.

04 Information Theory & Entropy

- Boolean models
- Algebraic models
- Probabilistic models *)

*) Our probabilistic models describes data which we can observe from our environment – and if we use the mathematics of probability theory , in order to express the uncertainties around our model then the inverse probability allows us to infer unknown unknowns … learning from data and making predictions – the core essence of machine learning and of vital importance for health informatics

Ghahramani, Z. 2015. Probabilistic machine learning and artificial intelligence. Nature, 521, (7553), 452-459, doi:10.1038/nature14541.



Lane, N. & Martin, W. (2010) The energetics of genome complexity. *Nature, 467, 7318, 929-934.*

Bayes' Rule in words
d **… data; *h* … hypothesis**
$H = \{H_1, H_2, \ldots, H_n\}$ … Hypothesis space

$$\forall h, d \quad \ldots$$

$$p(h|d) = \frac{p(d|h)p(h)}{\sum_{h \in H} p(d|h')\, p(h')}$$

Posterior Probability

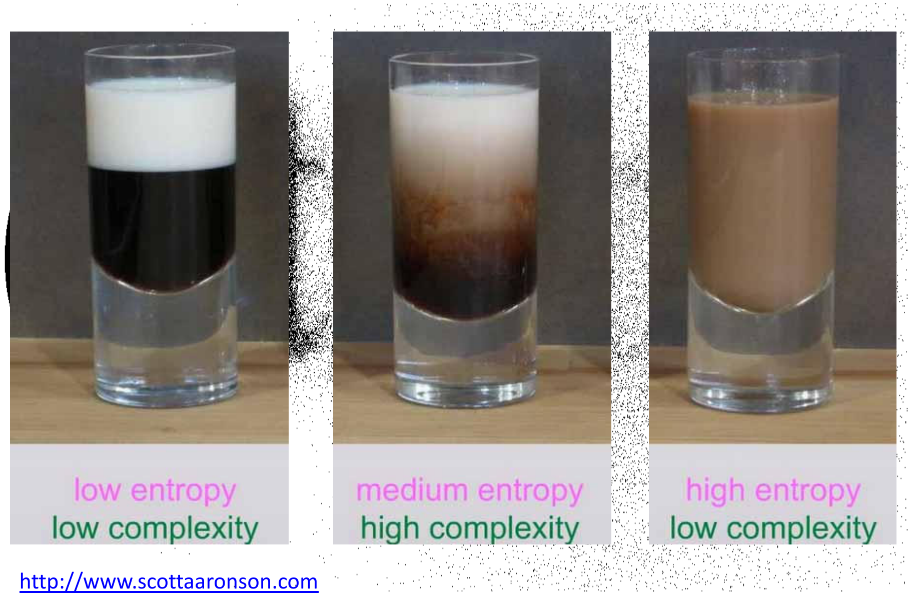
Likelihood

Prior Probability

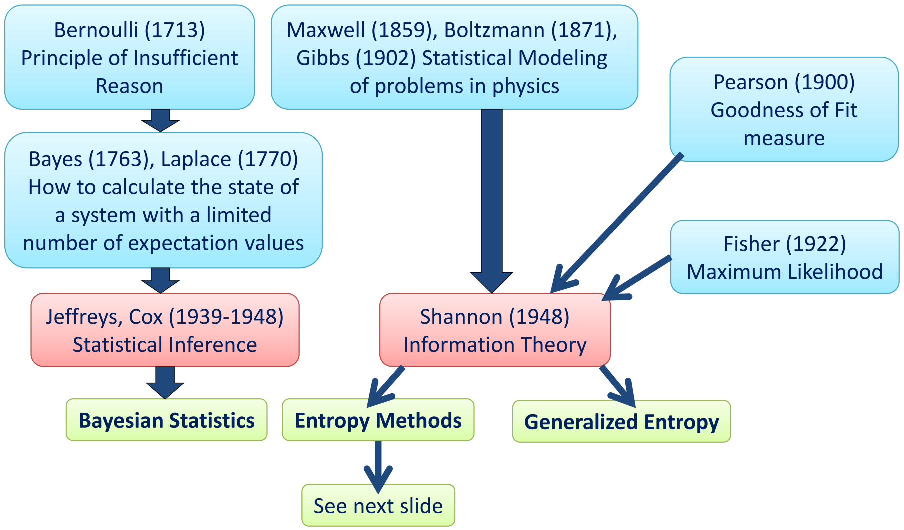
Evidence = marginal likelihood

Sum over space of alternative hypotheses

- Information is the reduction of uncertainty
- If something is 100 % certain its uncertainty = 0
- Uncertainty is max. if all choices are equally probable (I.I.D)
- Uncertainty (as information) sums up for independent sources



http://www.scottaaronson.com



confer also with: Golan, A. (2008) Information and Entropy Econometric: A Review and Synthesis. *Foundations and Trends in Econometrics, 2, 1-2, 1-145.*

Entropic Methods

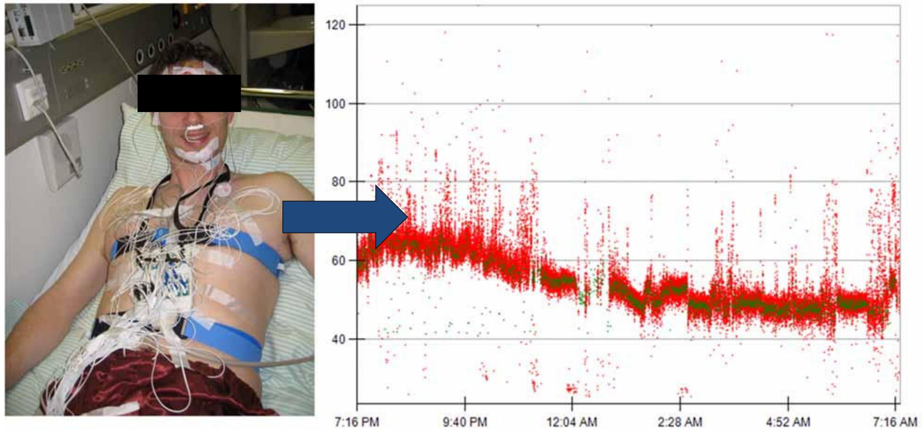
- Jaynes (1957) **Maximum Entropy (MaxEn)**
- Adler et al. (1965) **Topology Entropy (TopEn)**
- Mowshowitz (1968) **Graph Entropy (MinEn)**
- Posner (1975) **Minimum Entropy (MinEn)**
- Pincus (1991) **Approximate Entropy (ApEn)**
- Rubinstein (1997) **Cross Entropy (CE)**
- Richman (2000) **Sample Entropy (SampEn)**

Generalized Entropy

- Renyi (1961) **Renyi-Entropy**
- Tsallis (1980) **Tsallis-Entropy**

Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.

What can we measure with entropy ?



Holzinger, A., Stocker, C., Bruschi, M., Auinger, A., Silva, H., Gamboa, H. & Fred, A. 2012. On Applying Approximate Entropy to ECG Signals for Knowledge Discovery on the Example of Big Sensor Data. *In: Huang, R., Ghorbani, A., Pasi, G., Yamaguchi, T., Yen, N. & Jin, B. (eds.) Active Media Technology, Lecture Notes in Computer Science, LNCS 7669. Berlin Heidelberg: Springer, pp. 646-657.*

EU Project EMERGE (2007-2010)

How does Approximate Entropy work ?

$$Let:\ \langle x_n \rangle = \{x_1, x_2, \dots, x_N\}$$

$$\vec{X}_i = (x_i, x_{(i+1)}, \dots, x_{(i+m-1)})$$

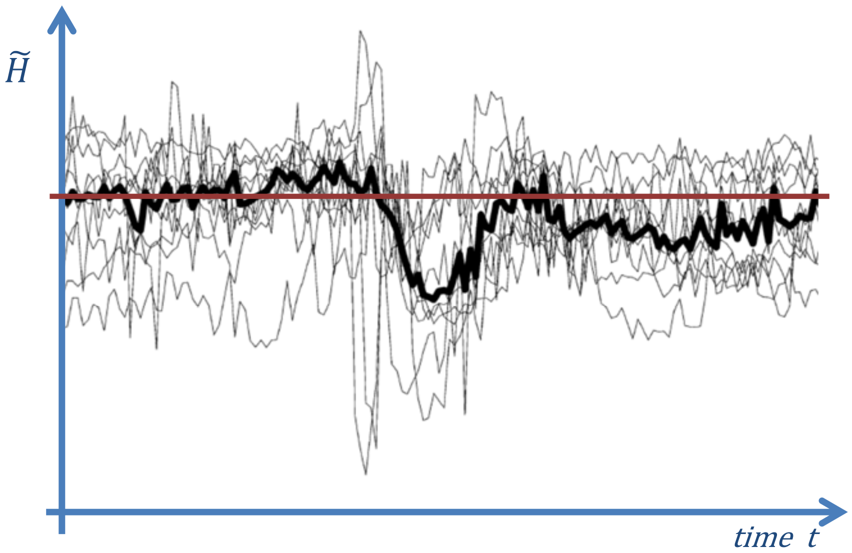
$$\left\| \vec{X}_i , \vec{X}_j \right\| = \max_{k=1,2,\dots,m} \left(\left| x_{(i+k-1)} - x_{(j+k-1)} \right| \right)$$

$$\widetilde{\mathrm{H}}(m,r) = \lim_{N \to \infty} \left[\phi^m(r) - \phi^{m+1}(r) \right]$$

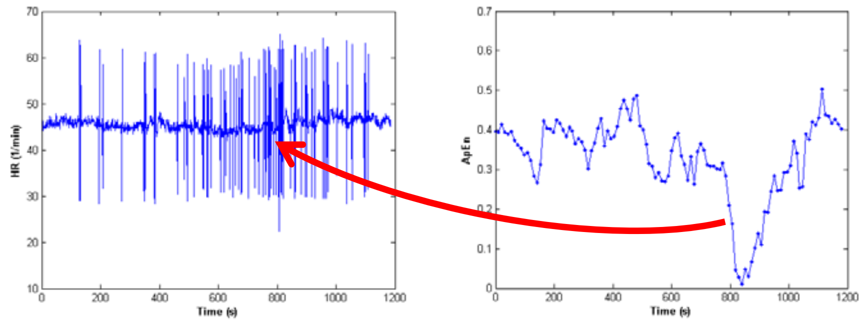
$$C_r^m(i) = \frac{N^m(i)}{N-m+1} \qquad \phi^m(r) = \frac{1}{N-m+1} \sum_{t=1}^{N-m+1} \ln C_r^m(i)$$

Pincus, S. M. (1991) Approximate Entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences of the United States of America, 88, 6, 2297-2301.*

What do we have to consider when measuring entropy ?



What is the main advantage of entropy measures ?



Holzinger, A., Hörtenhuber, M., Mayer, C., Bachler, M., Wassertheurer, S., Pinho, A. & Koslicki, D. 2014. On Entropy-Based Data Mining. In: Holzinger, A. & Jurisica, I. (eds.) Interactive Knowledge Discovery and Data Mining in Biomedical Informatics, Lecture Notes in Computer Science, LNCS 8401. Berlin Heidelberg: Springer, pp. 209-226.

Cross-Entropy Kullback-Leibler Divergence

- Entropy:
  - Measure for the **uncertainty** of random variables
- Kullback-Leibler divergence:
  - **comparing two distributions**
- Mutual Information:
  - measuring the **correlation** of two random variables

ON INFORMATION AND SUFFICIENCY

BY S. KULLBACK AND R. A. LEIBLER

The George Washington University and Washington, D. C.

1. Introduction. This note generalizes to the abstract case Shannon's definition of information [15], [16]. Wiener's information (p. 75 of [18]) is essentially the same as Shannon's although their motivation was different (cf. footnote 1, p. 95 of [16]) and Shannon apparently has investigated the concept more completely. R. A. Fisher's definition of information (intrinsic accuracy) is well known (p. 709 of [6]). However, his concept is quite different from that of Shannon and Wiener, and hence ours, although the two are not unrelated as is shown in paragraph 2.

R. A. Fisher, in his original introduction of the *criterion of sufficiency*, required "that the statistic chosen should summarize the whole of the relevant information supplied by the sample," (p. 316 of [5]). Halmos and Savage in a recent paper, one of the main results of which is a generalization of the well known Fisher-Neyman theorem on sufficient statistics to the abstract case, conclude, "We think that confusion has from time to time been thrown on the subject by . . . , and (c) the assumption that a sufficient statistic contains all the information in only the technical sense of 'information' as measured by variance," (p. 241 of [8]). It is shown in this note that the information in a sample as defined herein, that is, in the Shannon-Wiener sense cannot be increased by any statistical operations and is invariant (not decreased) if and only if sufficient statistics are employed. For a similar property of Fisher's information see p. 717 of [6], Doob [19].

We are also concerned with the statistical problem of discrimination ([3], [17]), by considering a measure of the "distance" or "divergence" between statistical populations ([1], [2], [13]) in terms of our measure of information. For the statistician two populations differ more or less according as to how difficult it is to discriminate between them with the best test [14]. The particular measure of divergence we use has been considered by Jeffreys ([10], [11]) in another connection. He is primarily concerned with its use in providing an invariant density of *a priori* probability. A special case of this divergence is Mahalanobis' generalized distance [13].

Solomon Kullback
1907-1994

Richard Leibler
1914-2003

Kullback, S. & Leibler, R. A. 1951. On information and sufficiency. The annals of mathematical statistics, 22, (1), 79-86, www.jstor.org/stable/2236703

$$\mathrm{H}[x] = -\sum_{x} p(x) \log_2 p(x)$$

Shannon, C. E. 1948. A Mathematical Theory of Communication. Bell System Technical Journal, 27, 379-423.

Important quantity in
- coding theory
- statistical physics
- machine learning

$$\mathrm{H}[\mathbf{y}|\mathbf{x}] = -\iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, \mathrm{d}\mathbf{y} \, \mathrm{d}\mathbf{x}$$

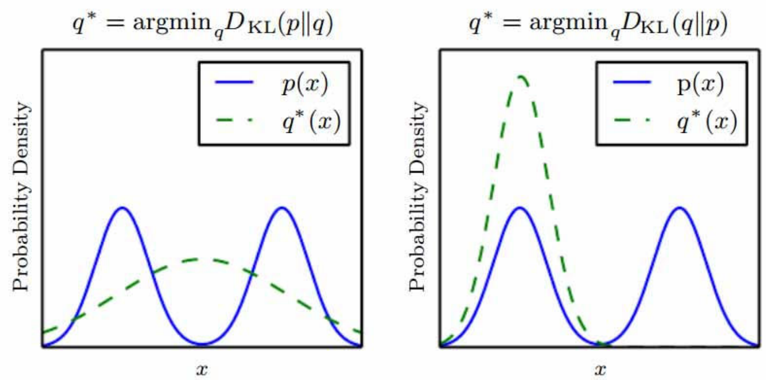
$$\mathrm{H}[\mathbf{x}, \mathbf{y}] = \mathrm{H}[\mathbf{y}|\mathbf{x}] + \mathrm{H}[\mathbf{x}]$$

$$\begin{aligned} \mathrm{KL}(p\|q) &= -\int p(\mathbf{x}) \ln q(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \left(-\int p(\mathbf{x}) \ln p(\mathbf{x}) \, \mathrm{d}\mathbf{x}\right) \\ &= -\int p(\mathbf{x}) \ln \left\{\frac{q(\mathbf{x})}{p(\mathbf{x})}\right\} \mathrm{d}\mathbf{x} \end{aligned}$$

$$\mathrm{KL}(p\|q) \simeq \frac{1}{N} \sum_{n=1}^{N} \{-\ln q(\mathbf{x}_n|\boldsymbol{\theta}) + \ln p(\mathbf{x}_n)\}$$

$$\mathrm{KL}(p\|q) \geqslant 0$$

KL-divergence is often used to measure the distance between two distributions

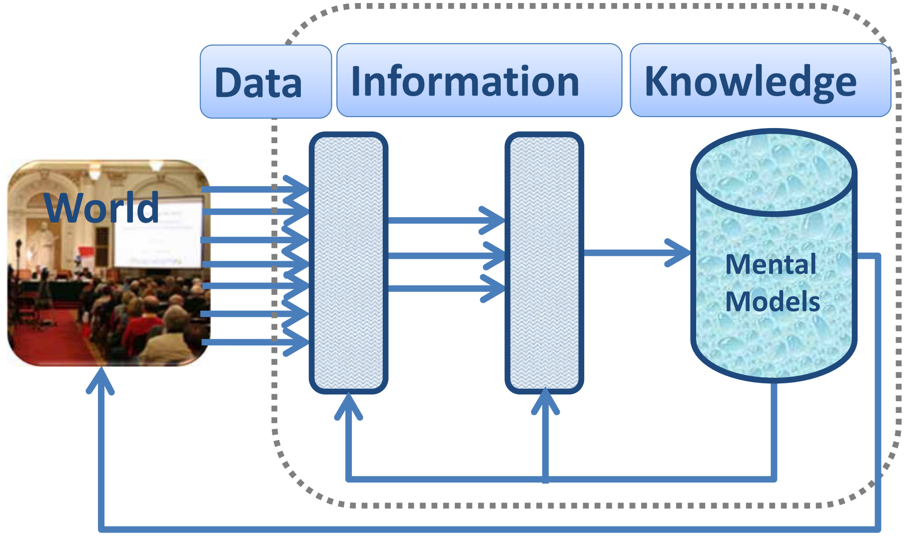


$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$$

Goodfellow, I., Bengio, Y. & Courville, A. 2016. Deep Learning, Cambridge (MA), MIT Press.

- … are **robust** against noise;
- … can be applied to **complex time series** with good replication;
- … is **finite** for stochastic, noisy, composite processes;
- … the values correspond directly to irregularities – good for detecting **anomalies**

05 Knowledge Representation

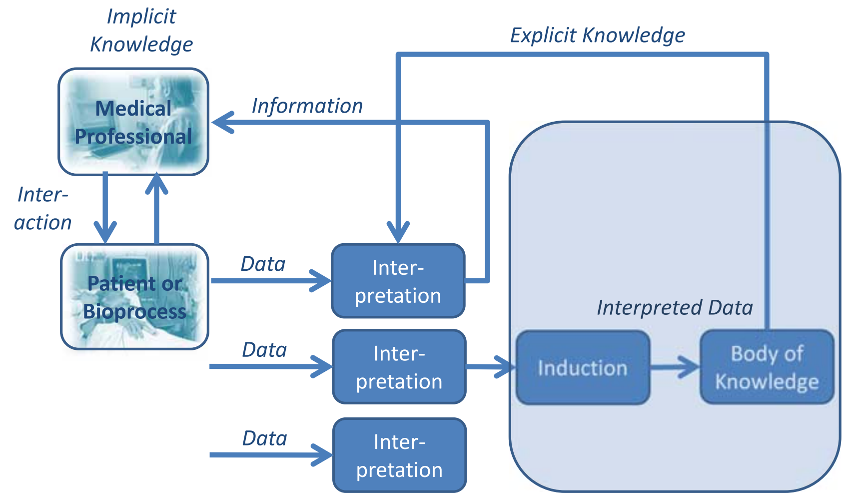


Knowledge := a set of expectations

AMIA 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**

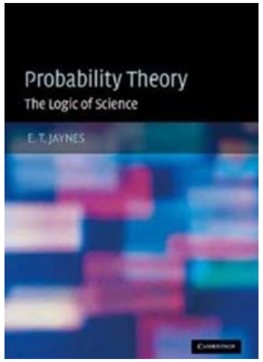
Edward H. Shortliffe 2011. Biomedical Informatics: Defining the Science and its Role in Health Professional Education. In: Holzinger, Andreas & Simonic, Klaus-Martin (eds.) Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058. Heidelberg, New York: Springer, pp. 711-714.



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



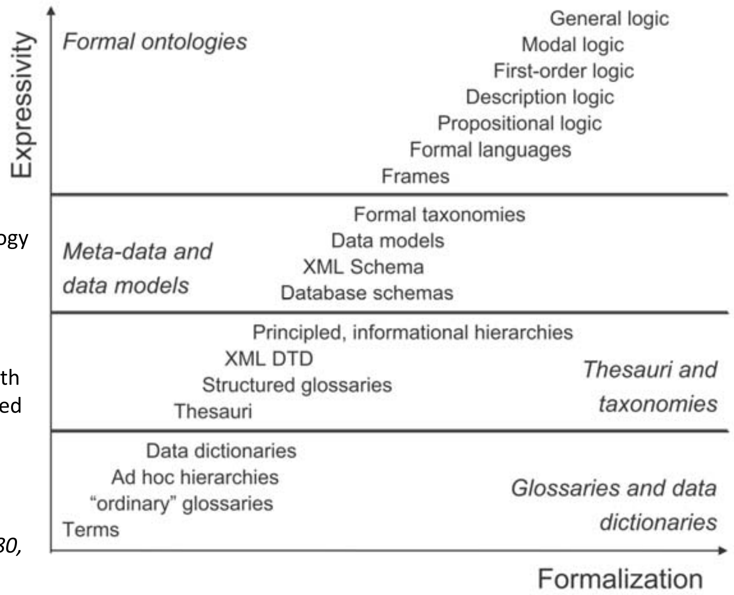
Edwin T. Jaynes 2003. Probability theory: The logic of science, Cambridge, Cambridge University Press.

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**

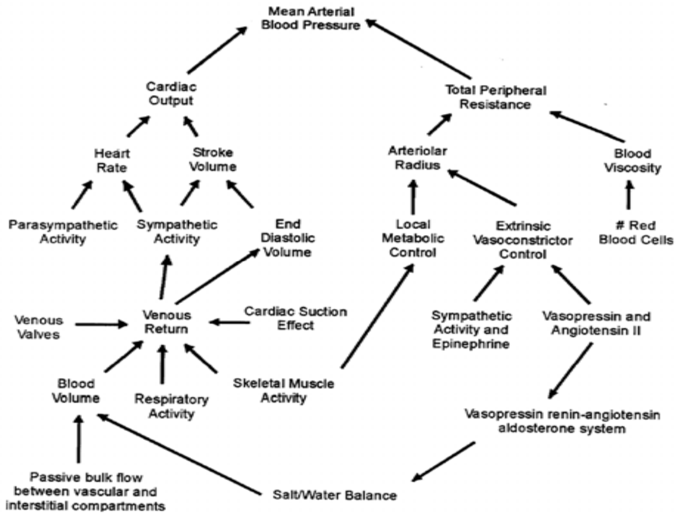
Davis, R., Shrobe, H. , Szolovits, P. 1993 What is a knowledge representation? AI Magazine, 14, 1, 17-33.

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.*

Why is the history of "Deep Learning" interesting for us ?

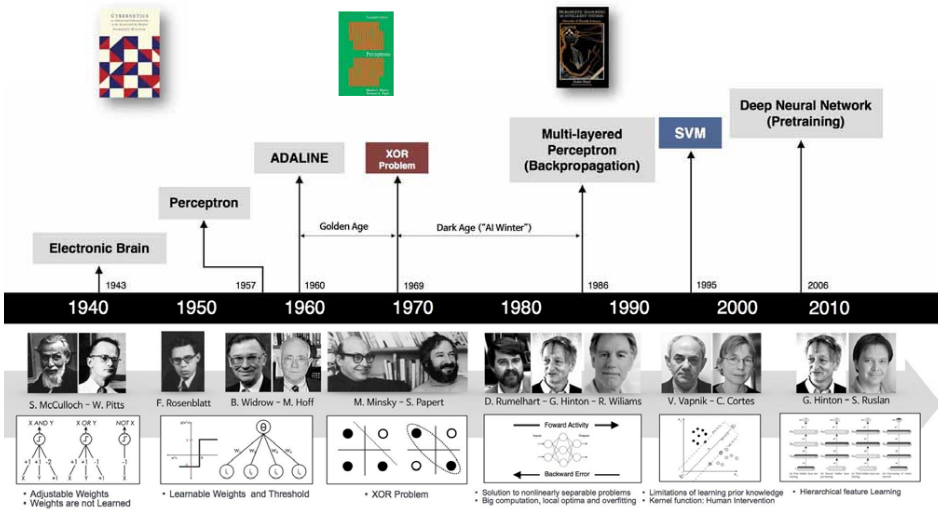
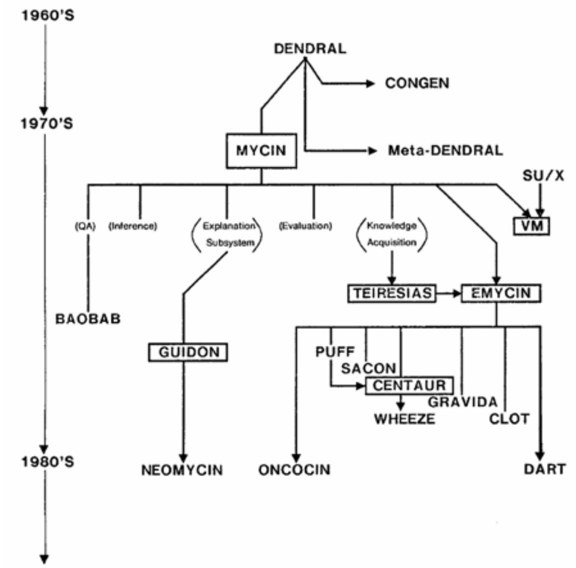


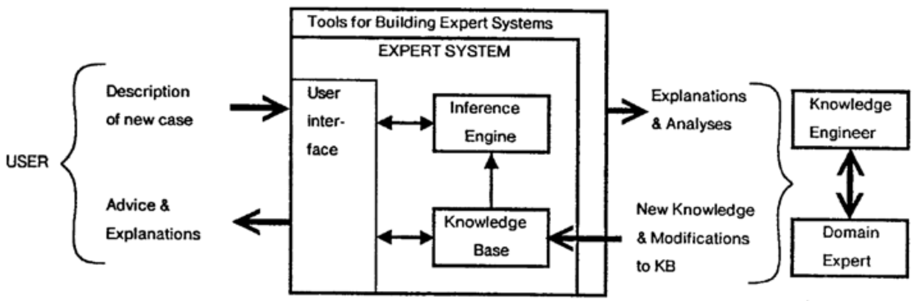
Image source: Andrew Beam, Department of Biomedical Informatics, Harvard Medical School
https://slides.com/beamandrew/deep-learning-101/#/12
This image is used according UrhG §42 lit. f Abs 1 as "Belegfunktion" for discussion with students

Why were early Decision Support Systems no success ?

Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

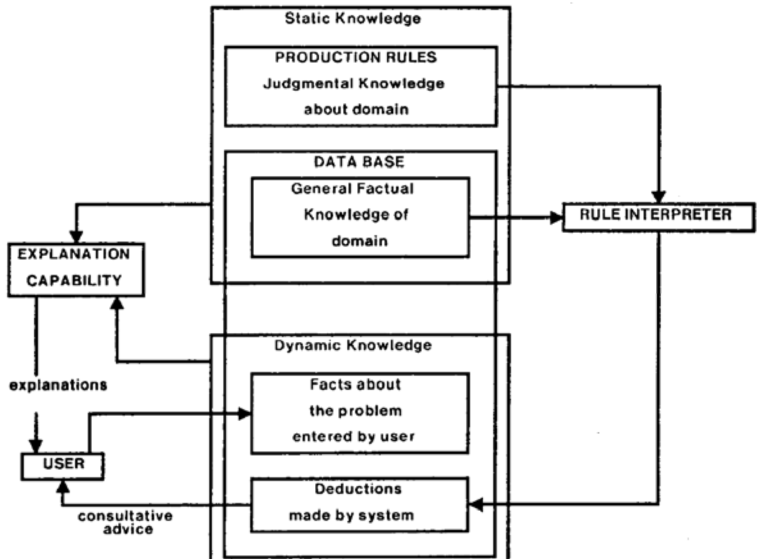


What is the connection of early expert systems with current xAI ?



Shortliffe, T. & Davis, R. (1975) Some considerations for the implementation of knowledge-based expert systems *ACM SIGART Bulletin, 55, 9-12.*

How did an expert system work ?



Shortliffe & Buchanan (1984)

What was the certainty factor in the MYCIN System ?

- MYCIN is a rule-based Expert System, which is used for therapy planning for patients with bacterial infections
- Goal oriented strategy (“Rückwärtsverkettung”)
- To every rule and every entry a certainty factor (CF) is assigned, which is between 0 und 1
- Two measures are derived:
- MB: measure of belief
- MD: measure of disbelief
- Certainty factor – CF of an element is calculated by:
  CF[h] = MB[h] – MD[h]
- CF is positive, if more evidence is given for a hypothesis, otherwise CF is negative
- CF[h] = +1 -> h is 100 % true
- CF[h] = –1 -> h is 100% false

How does an example of the Certainty Factor look like ?

h_1 = The identity of ORGANISM-1 is streptococcus
h_2 = PATIENT-1 is febrile
h_3 = The name of PATIENT-1 is John Jones

$CF[h_1,E] = .8$	:	There is strongly suggestive evidence (.8) that the identity of ORGANISM-1 is streptococcus
$CF[h_2,E] = -.3$	:	There is weakly suggestive evidence (.3) that PATIENT-1 is not febrile
$CF[h_3,E] = +1$	:	It is definite (1) that the name of PATIENT-1 is John Jones

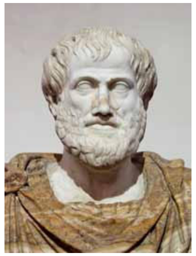
Shortliffe, E. H. & Buchanan, B. G. (1984) *Rule-based expert systems: the MYCIN experiments of the Stanford Heuristic Programming Project. Addison-Wesley.*

Ontologies

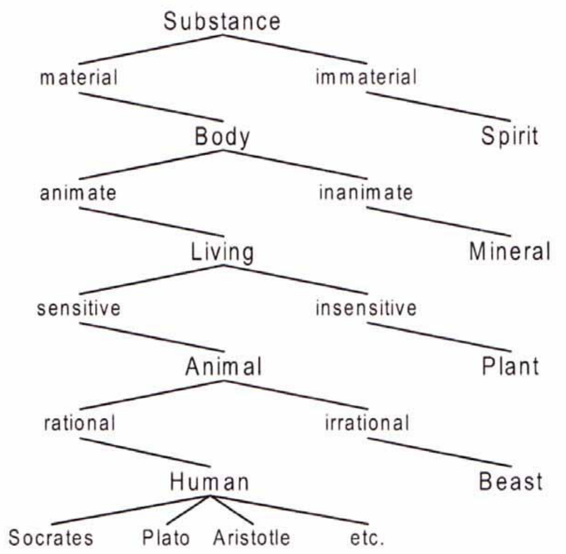
What happens if you put the word “Jaguar” into a search engine ?



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



* 384 BC † 322 BC

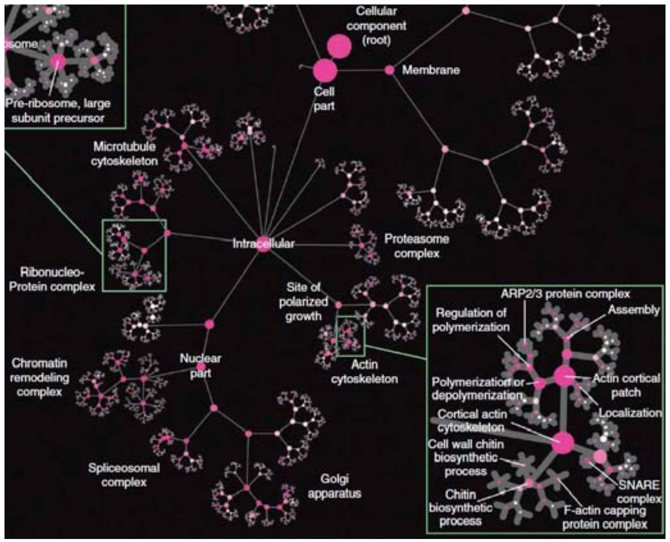


Simonet, M., Messai, R., Diallo, G. & Simonet, A. (2009) Ontologies in the Health Field. In: Berka, P., Rauch, J. & Zighed, D. A. (Eds.) *Data Mining and Medical Knowledge Management: Cases and Applications. New York, Medical Information Science Reference, 37-56.*

Later: Porphyry (≈ 234-305) tree

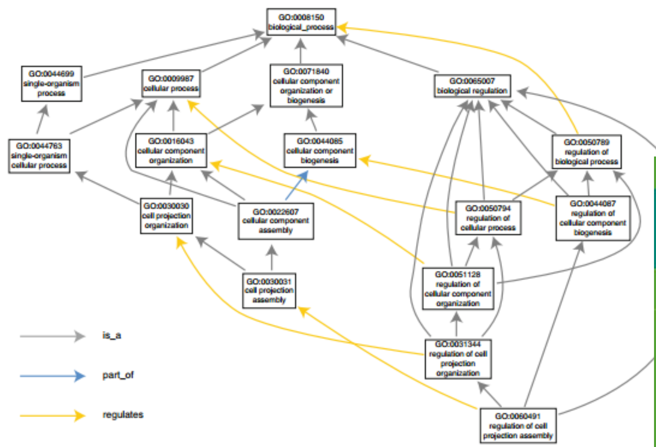
- Aristotle attempted to **classify the things in the world** - where it is employed to describe the existence of beings in the world;
- Artificial Intelligence and Knowledge Engineering deals also with **reasoning about models of the world**.
- Therefore, AI researchers adopted the term 'ontology' to 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.**

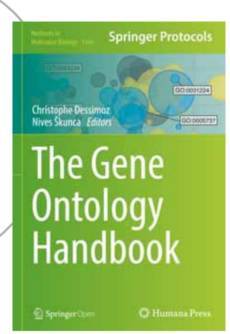
Studer, R., Benjamins, V. R. & Fensel, D. (1998) Knowledge Engineering: Principles and methods. *Data & Knowledge Engineering, 25, 1-2, 161-197.*



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.

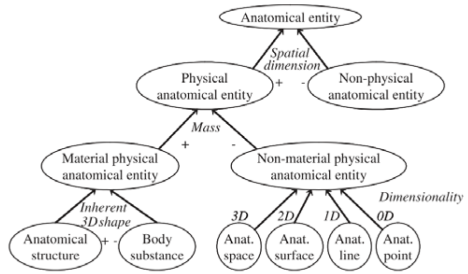
What is the difference between an ontology and a terminology ?

- 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** …

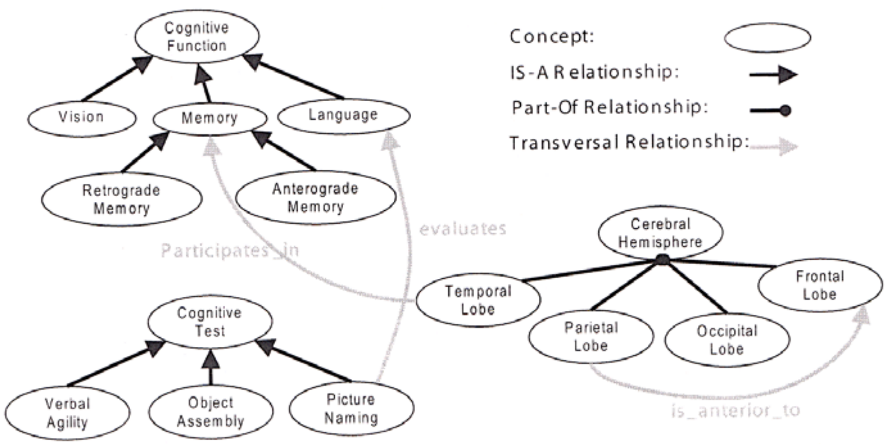
What are the conditions an ontology may satisfy ?

(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.*



What is a semantic relationship ?



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				Subs. Hier.	Version / Notes
				Min	Max	Med	Avg		
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.*

- **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**

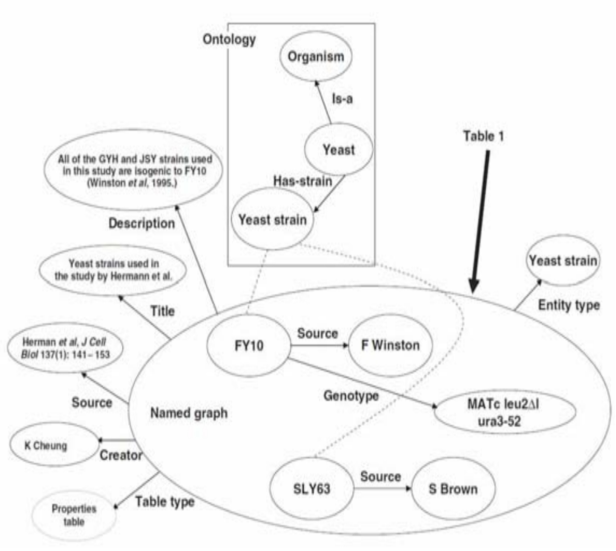


Table I Yeast strains used in the study by Hermann et al (1997)

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*.

DL = Description Logic

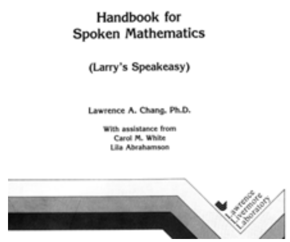
Concept inclusion, Speak: All C1 are C2

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	$\top \sqsubseteq \leq 1P$	$\top \sqsubseteq \leq$ 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.*

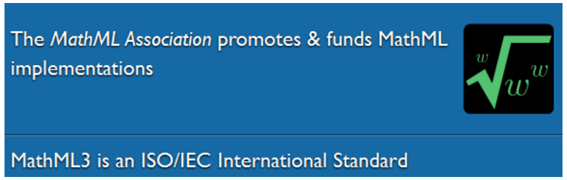
web.efzg.hr/dok/MAT/vkojic/Larrys_speakeasy.pdf



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 ?

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

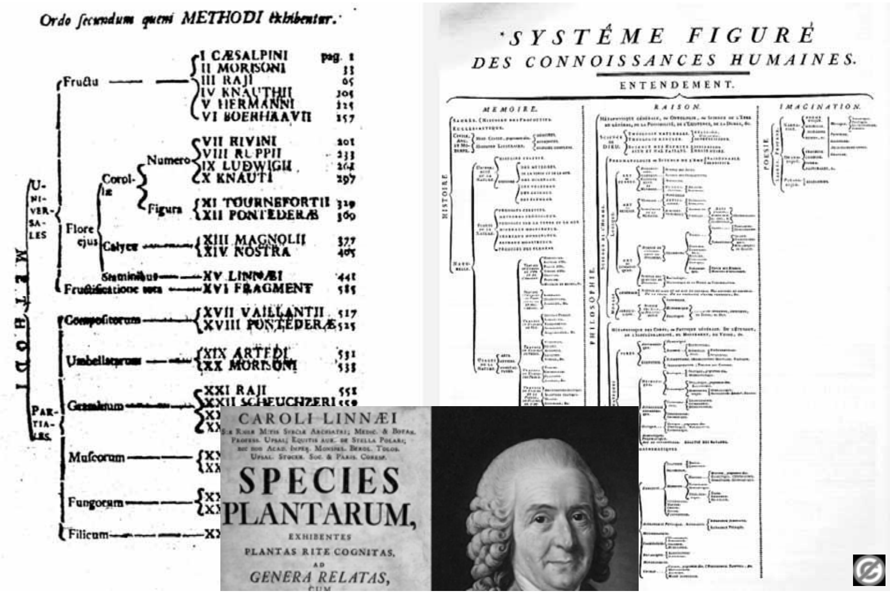
Universal Restriction
Speak: All P-successors are in C

Existential Restriction
Speak: An P-successor exists in C

Bhatt et al. (2009)

Medical Classifications

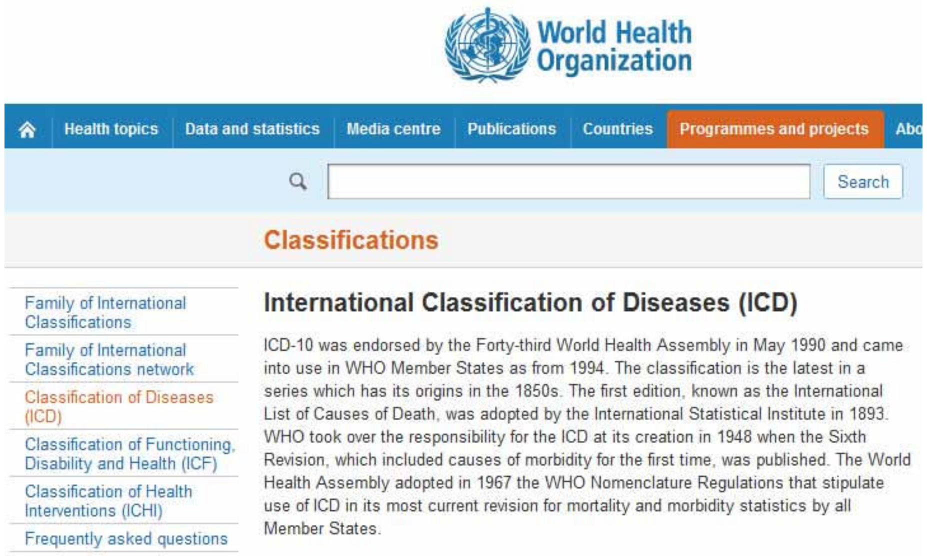
What is classification generally ?



What medical classification systems do we know ?

- Since the classification by Carl von Linne (1735) approx. 100+ various classifications in use:
  - International Classification of Diseases (ICD)
  - Systematized Nomenclature of Medicine (SNOMED)
  - Medical Subject Headings (MeSH)
  - Foundational Model of Anatomy (FMA)
  - Gene Ontology (GO)
  - Unified Medical Language System (UMLS)
  - Logical Observation Identifiers Names & Codes (LOINC)
  - National Cancer Institute Thesaurus (NCI Thesaurus)

What is the International Classification of Diseases (ICD) ?



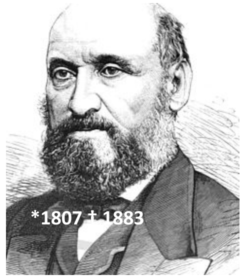
International Classification of Diseases (ICD)

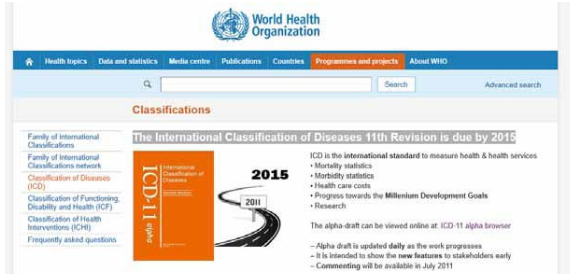
ICD-10 was endorsed by the Forty-third World Health Assembly in May 1990 and came into use in WHO Member States as from 1994. The classification is the latest in a series which has its origins in the 1850s. The first edition, known as the International List of Causes of Death, was adopted by the International Statistical Institute in 1893. WHO took over the responsibility for the ICD at its creation in 1948 when the Sixth Revision, which included causes of morbidity for the first time, was published. The World Health Assembly adopted in 1967 the WHO Nomenclature Regulations that stipulate use of ICD in its most current revision for mortality and morbidity statistics by all Member States.

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 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





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

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))

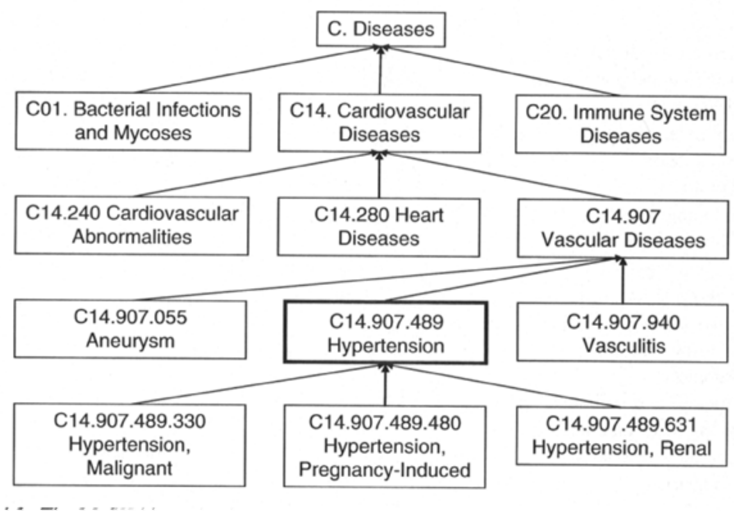
B

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.*

- MeSH thesaurus is produced by the National Library of Medicine (NLM) since 1960.
- Used for cataloging documents and related media and as an <u>index</u> 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.

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]



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

National Library of Medicine - Medical Subject Headings

2011 MeSH

MeSH Descriptor Data

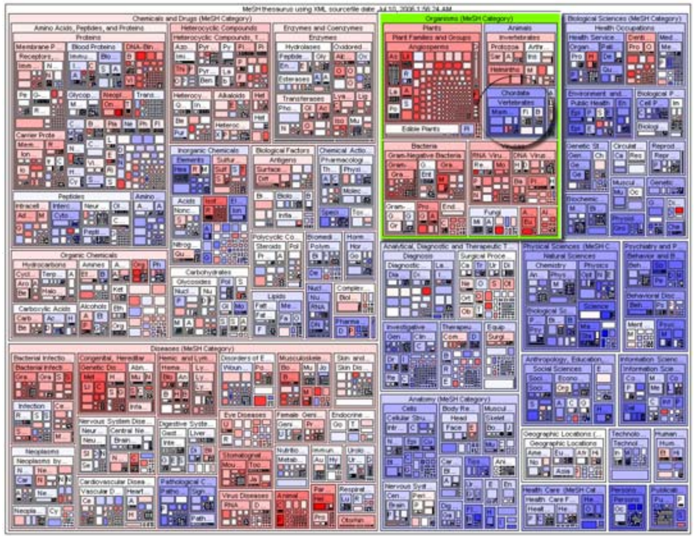
Return to Entry Page

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

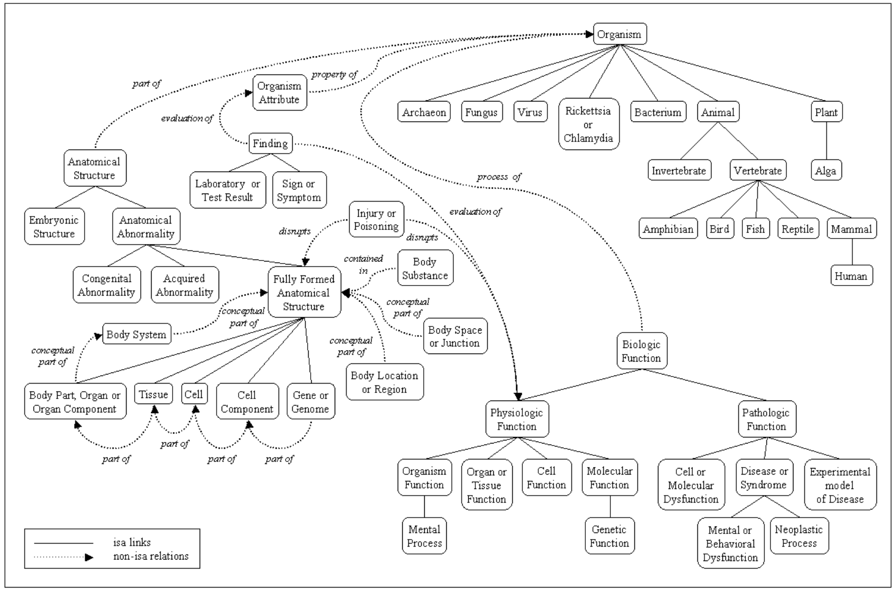
http://www.nlm.nih.gov/mesh/

How does Mesh look in an interactive tree visualization ?



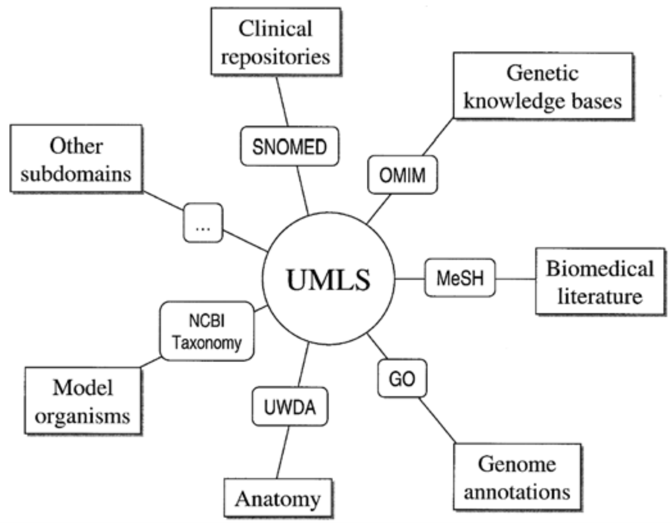
Eckert, K. (2008) A methodology for supervised automatic document annotation. *Bulletin of IEEE Technical Committee on Digital Libraries TCDL, 4, 2.*

What is UMLS – Unified Medical Language System ?



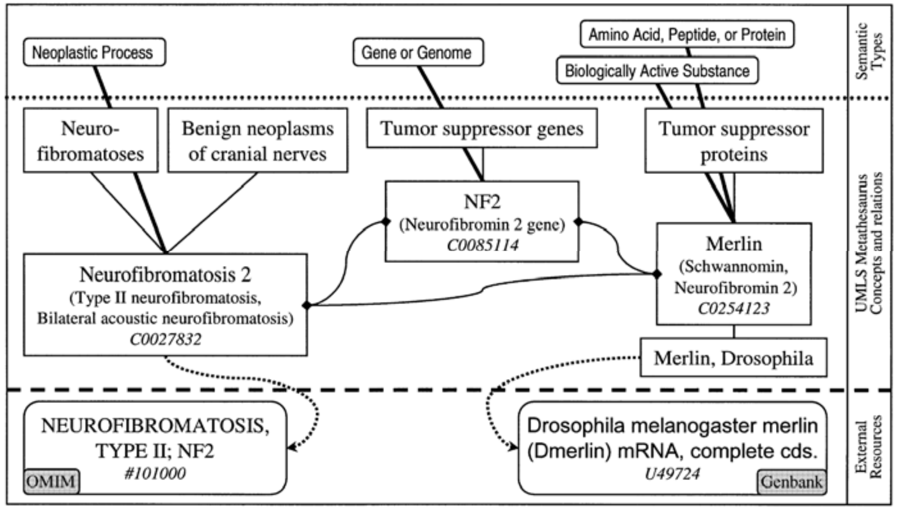
http://www.nlm.nih.gov/research/umls/

What subdomains does UMLS Metathesaurus integrate ?



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

How does an Example of proteins and diseases in the UMLS look ?

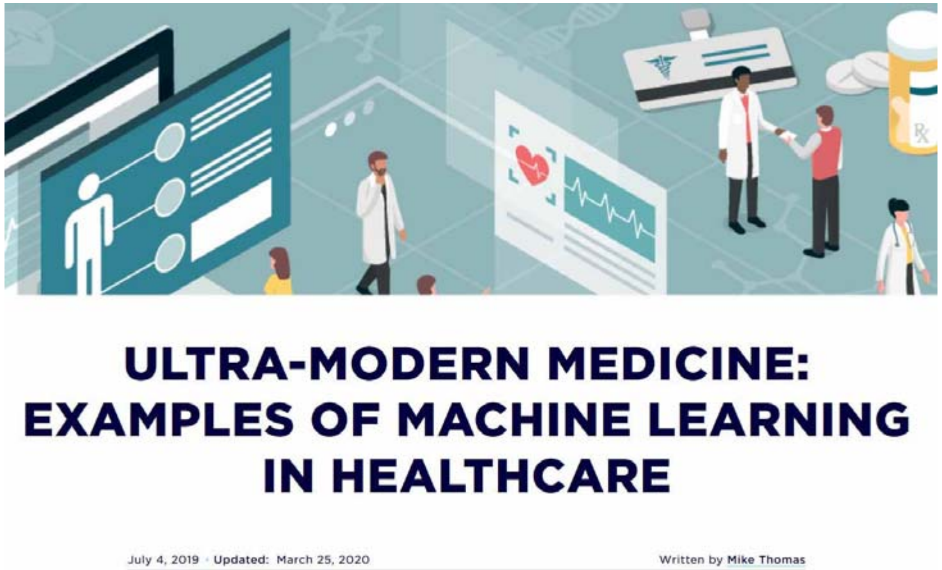


Bodenreider, O. (2004) The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research, 32, D267-D270.*

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.



Thank you!