From data to probabilistic information and knowledge

Contact: andreas.holzinger@tuwien.ac.at


Warm-up Quiz

00 Reflection

- 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

How to ensure good data quality assessment?

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>The extent to which data is available, easy and quickly accessible.</td>
</tr>
<tr>
<td>Appropriateness of Data</td>
<td>The extent to which the volume of data is appropriate for the task at hand.</td>
</tr>
<tr>
<td>Reliability</td>
<td>The extent to which data is regarded as true.</td>
</tr>
<tr>
<td>Completeness</td>
<td>The extent to which data is not missing and is full enough breadth and depth for the task at hand.</td>
</tr>
<tr>
<td>Usability</td>
<td>The extent to which data is easy to use in different ways and can be applied to different tasks.</td>
</tr>
<tr>
<td>Representational</td>
<td>The extent to which data is represented.</td>
</tr>
<tr>
<td>Consistency</td>
<td>The extent to which data is consistent and reliable.</td>
</tr>
<tr>
<td>Interoperability</td>
<td>The extent to which data is in an appropriate format, accessible, and interoperable.</td>
</tr>
<tr>
<td>Objectivity</td>
<td>The extent to which data is unbiased, unprejudiced, or impartial.</td>
</tr>
<tr>
<td>Relevance</td>
<td>The extent to which data is applicable and helpful for the task at hand.</td>
</tr>
<tr>
<td>Reusability</td>
<td>The extent to which data has multiple uses.</td>
</tr>
<tr>
<td>Security</td>
<td>The extent to which access to data is restricted appropriately to maintain its security.</td>
</tr>
<tr>
<td>Timeliness</td>
<td>The extent to which the data is sufficiently up-to-date for the task at hand.</td>
</tr>
<tr>
<td>Unambiguity</td>
<td>The extent to which data is easily comprehended.</td>
</tr>
<tr>
<td>Value-Added</td>
<td>The extent to which data is beneficial and provides advantages from its use.</td>
</tr>
</tbody>
</table>


What is the FAIR guiding principle for scientific data?

- "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.

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]

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

- Data in traditional Statistics
  - Low-dimensional data ( < ℝ¹⁰⁰)
  - 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 ( ≫ ℝ¹⁰⁰)
  - Problem: not noise, but complexity
  - Much structure, but the structure can not be represented by a simple model


**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.

How can we classify traditional data structures?

- **Data structures**
  - Separable
    - Qualitative
      - Nominal
    - Quantitative
      - Ordinal
      - Interval
      - Ratio
  - Aggregated
    - Mixed
    - Pure

Aggregated attribute $\rightarrow$ a homomorphic map $\mathcal{X}$ from a relational system $\langle \mathcal{A} \rangle \rightarrow \langle \mathcal{B} \rangle$ into a relational system $\langle \mathcal{B} \rangle \rightarrow \langle \mathcal{C} \rangle$, where $\mathcal{A}$ and $\mathcal{B}$ are two different sets of data elements.

This is in contrast with other attributes since the set $\mathcal{C}$ is the set of data elements instead of atomic values.

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

- **On the Theory of Scales of Measurement**
  S. S. Stevens
  Director, Signal-Noise Laboratory, Harvard University

<table>
<thead>
<tr>
<th>Scale</th>
<th>Basic Statistics</th>
<th>Mathematical Operations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Mode, contingency correlation</td>
<td>$=, \neq$</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Median, Percentile</td>
<td>$=, \neq, &gt;, &lt;$</td>
</tr>
<tr>
<td>Interval</td>
<td>Mean, Std. Dev.</td>
<td>$=, &gt;, &lt;, \neq$</td>
</tr>
<tr>
<td>Ratio</td>
<td>Coefficient of variation</td>
<td>$=, &gt;, &lt;, \neq, \neq, &gt;$</td>
</tr>
</tbody>
</table>

What levels of data taxonomy can we identify?

- **Physical level** $\rightarrow$ bit $=$ binary digit $=$ basic indissoluble unit ($=$ Shannon, Sh), $\neq$ Bit ($1$) in Quantum Systems $\rightarrow$ qubit
- **Logical Level** $\rightarrow$ integers, booleans, characters, floating-point numbers, alphanumeric strings, ...
- **Conceptual (Abstract) Level** $\rightarrow$ data-structures, e.g. lists, arrays, trees, graphs, ...
- **Technical Level** $\rightarrow$ Application data, e.g. text, graphics, images, audio, video, multimedia, ...
- **"Hospital Level"** $\rightarrow$ Narrative (textual) data, numerical measurements (physiological data, lab results, vital signs, ...), recorded signals (ECG, EEG, ...), Images (x-ray, MR, CT, PET, ...) ; -omics
Where do data come from at Hospital Level?

- **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, ...

Why is Digital Pathology interesting?

- **How is a WSI produced?**

What is the current state of the art in machine learning for pathology?


What about the ground truth?


Why is Neonatal Screening a good example for data generation?


What is an example for the Data Structure “list”?

Why is the data structure graph so versatile?

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


What is a tree?


What are origins of health-related data?

02 Biomedical data sources: Taxonomy of data
Why is data integration in health an unsolved problem?

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

- Metabolomics
- Chemical/physical processes
- Cellular reactions
- Enzymatic reactions

- Microbiomes
- Microorganisms processes
- Plants, fungi, ...

- Proteomics
- Protein-Protein interactions

- Epigenetics
- Epigenetic modifications

Genomics

- Exposure
- Environmental data
- Air pollution
- Exposure (toxins)

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

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

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

What is a good example for the level “cell”?

- Loose connective tissue with fibroblasts
- Bone tissue with osteocytes
- Smooth muscle cells
- Fat (adipose) cells
- Intestinal epithelial cells
- Bundle of nerve cells
- Red blood cells
- Striated muscle cells


What is life according to Erwin Schrödinger?

- to reproduce ...
- to grow ...
- to evolve ...
- to self-replicate ...
- to generate/utilize energy ...
- to process information ...

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](https://human-centered.ai/open-data-sets)

**What are the most important *omics data?**

- Genomics (sequence annotation)
- Transcriptomics (microarray)
- Proteomics (Proteome Databases)
- Metabolomics (enzyme annotation)
- Fluxomics (isotopic tracing, metabolic pathways)
- Phenomics (biomarkers)
- Epigenomics (epigenetic modifications)
- Microbiomics (microorganisms)
- Lipidomics (pathways of cellular lipids)

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

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


Example: 2.5-D data (structural information & metadata)

http://www.pdb.org

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

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 is the goal of data integration?

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


What is information fusion?

Constructing a multi-modal interaction & correspondence graph (IGC)

Interesting signals from each modality (time-based, image, structured & unstructured) are connected according to pre-defined rules. Each modality’s features lie in their own, un-aligned concept spaces.

Our central hypothesis:
Information may bridge this gap

What is translational health?

Translational Medicine Continuum

Biomedical Informatics Continuum

What are the key problems in medical data management?
What is the combining link?

The combining link is text

Biomedical R&D data (e.g. clinical trial data)
Clinical patient data (e.g. EPR, lab, reports etc.)
Health business data (e.g. costs, utilization, etc.)
Private patient data (e.g. AAL, monitoring, etc.)


Why is medical text important?


Why is medical work relying on team communication?

- ... and requires a lot of open information exchange..

Digression: Medical Communication

Why does Language Understanding require knowledge?

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


Why is text a good example for Non-Standardized Data

What are the typical challenges in data-driven medicine?

- 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]


---

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, ...)

---

How does image augmentation work?


04 Information Theory & Entropy

- Boolean models
- Algebraic models
- Probabilistic models *)

*) Our probabilistic models are based on 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.


Why is life complex information?

What is information?


Why is the work of Bayes, Price, Laplace so important for us?

Bayes' Rule in words

\[ p(h|d) = \frac{p(d|h)p(h)}{\sum_{h'} p(d|h')p(h')} \]

\( \forall h, d \quad \ldots \)

\( p(h|d) \) - Posterior Probability

\( p(d|h)p(h) \) - Likelihood

\( \sum_{h' \in H} p(d|h')p(h') \) - Evidence = marginal likelihood

Prior Probability

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

What are the origins of Entropy?
- Bernoulli (1713)
  - Principle of Insufficient Reason
- Maxwell (1859), Boltzmann (1871), Gibbs (1902)
  - Statistical Modeling of problems in physics
- Bayes (1763), Laplace (1770)
  - How to calculate the state of a system with a limited number of expectation values
- Jeffreys, Cox (1939-1948)
  - Statistical Inference
- Bayesian Statistics
- Entropy Methods

Shannon (1948)
- Information Theory
- Pearson (1900)
  - Goodness of Fit measure
- Fisher (1922)
  - Maximum Likelihood

See next slide

What current Entropy methods can we use?

- Entropic Methods
  - Jaynes (1957)
    - Maximum Entropy (MaxEn)
  - Adler et al. (1965)
    - Topology Entropy (TopEn)
  - Mowshowitz (1968)
    - Graph Entropy (MinEn)
  - Pincus (1991)
    - Approximate Entropy (ApEn)
  - Richman (2000)
    - Sample Entropy (SampEn)

- Generalized Entropy
  - Rényi (1961)
    - Rényi-Entropy
  - Tsalis (1980)
    - Tsalis-Entropy
  - Rubinstein (1997)
    - Cross Entropy (CE)


What can we measure with entropy?

Let: \( \{x_n\} = \{x_1, x_2, \ldots, x_N\} \)

\[ \tilde{X}_t = (x_t, x_{t+1}, \ldots, x_{t+m-1}) \]

\[ ||\tilde{X}_i, \tilde{X}_j|| = \max_{k=1,2,..,m} (|x_{i+k-1} - x_{j+k-1}|) \]

\[ \overline{H}(m,r) = \lim_{N \to \infty} [\phi^m(r) - \phi^{m+1}(r)] \]

\[ C_r^m(i) = \frac{N^m(i)}{N - m + 1} \quad \phi^m(r) = \frac{1}{N - m + 1} \sum_{i=1}^{N-m+1} \ln C_r^m(i) \]


What do we have to consider when measuring entropy?

What is the main advantage of entropy measures?

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

What is the difference between Entropy – KL-divergence and MI?

Solomon Kullback & Richard Leibler (1951)

ON INFORMATION AND SUFFICIENCY
BY S. KULLBACK AND R. A. LEIBLER
The George Washington University and Washington, D. C.

1. Introduction. This note generalizes the abstract case: Shannon’s definition of information [16], [18]. We consider the information (p. 79 of 189) to be essentially the same as Shannon’s although their motivation was different (cf. footnote 1, p. 78 of 189) and Shannon apparently has investigated the concept more completely. H. A. Fisher’s definition of information (maximum accuracy) is well known (p. 799 of 209). However, his concept is quite different from that of Shannon and Wiener, and hence even, although the two are not unrelated as is shown in paragraph 2.

H. A. Fisher’s definition of information is of the following form: “the statistic chosen should summarize the whole of the relevant information supplied by the sample.” (p. 799 of 209). For example, 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, concludes: “We think that confusion has 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. 799 of 209). It is shown in this note that the information in a sample or defined as that in the Shannon-Wiener sense cannot be increased by any statistical operation and is invariant under these if and only if sufficient statistics are employed. For a similar property of Fisher’s information see p. 712 of 206, Dvoretz [16].

We are also concerned with the statistical problem of discrimination [16], [17], by considering a measure of the “uniform” or “expected” between statistical populations [15], [20], [186] in terms of our measure of information. For the statistic of two populations differ more or less according to the difficulty it is to discriminate between them with the best test [14]. The particular measure of information we have been considering by Fisher’s [187], [181] in another connection. He is primarily concerned with the use in providing an invariant density of a priori probability. A special case of this divergence is Mahalanobis’ generalized distance [13].

Why should we remember Shannon Entropy?

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


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

Kullback, S. & Leibler, R. A.
1951. On information and sufficiency. The annals of mathematical statistics, 22, (1), 79-86,
www.jstor.org/stable/223703
What is Conditional Entropy?

\[ H[y|x] = - \int \int p(y, x) \ln p(y|x) \, dy \, dx \]

\[ H[x, y] = H[y|x] + H[x] \]

When do we need the Kullback-Leibler Divergence?

\[
\begin{align*}
KL(p||q) &= - \int p(x) \ln q(x) \, dx - \left( - \int p(x) \ln p(x) \, dx \right) \\
&= - \int p(x) \ln \left( \frac{q(x)}{p(x)} \right) \, dx \\
KL(p||q) &\simeq \frac{1}{N} \sum_{n=1}^{N} \left\{ - \ln q(x_n|\theta) + \ln p(x_n) \right\}
\end{align*}
\]

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

What is important to note when using KL divergence?

\[ q^* = \arg \min_q D_{KL}(p||q) \]

\[ p(x) \quad q^*(x) \]

In summary: Why do we use Entropy measures generally?

- ... 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

- 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

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


What are examples for famous knowledge representations?

<table>
<thead>
<tr>
<th>Mathematical Logic</th>
<th>Psychology</th>
<th>Biology</th>
<th>Statistics</th>
<th>Economics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aristotle</td>
<td>Descartes</td>
<td>Boole</td>
<td>James</td>
<td>Laplace</td>
</tr>
<tr>
<td>Frege</td>
<td>Newton</td>
<td>Hebb</td>
<td>Babbage</td>
<td>Berthier</td>
</tr>
<tr>
<td>Goedel</td>
<td>Frege</td>
<td>Poincaré</td>
<td>Laplace</td>
<td>Bernoulli</td>
</tr>
<tr>
<td>Peirce</td>
<td>Boole</td>
<td>Millar</td>
<td>Frege</td>
<td>Friedman</td>
</tr>
<tr>
<td>Church</td>
<td>Poincaré</td>
<td>Newell</td>
<td>Millar</td>
<td>Fermat</td>
</tr>
<tr>
<td>Turing</td>
<td>Poincaré</td>
<td>Simon</td>
<td>Millar</td>
<td>Von Neumann</td>
</tr>
<tr>
<td>Davis</td>
<td>Poincaré</td>
<td>von Neumann</td>
<td>Millar</td>
<td>Simons</td>
</tr>
<tr>
<td>Putnam</td>
<td>Poincaré</td>
<td>von Neumann</td>
<td>Millar</td>
<td>Simons</td>
</tr>
<tr>
<td>Robinson</td>
<td>Poincaré</td>
<td>von Neumann</td>
<td>Millar</td>
<td>Simons</td>
</tr>
</tbody>
</table>

What does Formalization versus Expressivity mean?


What do you need for developing clinical decision support systems?


Why is the history of “Deep Learning” interesting for us?

Image source: Andrew Ng, Department of Biomedical Informatics, Harvard Medical School. https://slickos.com/earn/download/deep-learning-101.html

This image is used according to the terms of the GNU Free Documentation License (GFDL) which is the same license used by Wikipedia. This license allows the use and redistribution of the image as long as the original author is credited and the license is included in the image.

Why were early Decision Support Systems no success?

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

Tools for Building Expert Systems

EXPERT SYSTEM

User Interface

Inference Engine

Knowledge Base

EXPLANATIONS & ANALYSES

Knowledge Engineer

Description of new case

Advice & Explanations

New Knowledge & Modifications to KB

Domain Expert


How did an expert system work?

Static Knowledge

PRODUCTION RULES

Judgmental Knowledge about domain

DATA BASE

General Factual Knowledge of domain

Dynamic Knowledge

Facts about the problem entered by user

Explanation capability

Rule Interpreter

Deductions made by system

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 and 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
  \[ CF[h] = +1 \Rightarrow h \text{ is } 100\% \text{ true} \]
  \[ CF[h] = -1 \Rightarrow h \text{ is } 100\% \text{ false} \]

How does an example of the Certainty Factor look like?

\[ h_1 = \text{The identity of ORGANISM-1 is streptococcus} \]
\[ h_2 = \text{PATIENT-1 is febrile} \]
\[ h_3 = \text{The name of PATIENT-1 is John Jones} \]

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

Ontologies

Who created the first 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.
What are the principles and methods of Knowledge Engineering?


Where are ontologies used today?


What is a Gene Ontology?

http://geneontology.org/


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 PARTOF. Every PARTOF relationship is reflexive, asymmetric and transitive. IS-A and PARTOF 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 \( (x, r, y) \) and \( (y, r', x) \) hold simultaneously. A symmetric relationship is its own inverse. Inverses of hierarchical relationships are called superiors and inferiors, respectively.

(4) Every non-taxonomic relation of \( x \) to \( z \), \( (x, r, z) \), is either inherited \( ((y, r, z)) \) or refined \( ((y, r, z')) \) 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 its parent or more specific properties \( z' \).


What is a semantic relationship?

- Cognitive Function
- Visuospatial
- Memory
- Language
- Anomalous


What are typical medical ontologies?

<table>
<thead>
<tr>
<th>Name</th>
<th>Rel.</th>
<th>Scope</th>
<th># of concepts</th>
<th># of concept names</th>
<th>Size Mb</th>
<th>Version / Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>SHOMOMT</td>
<td>[71]</td>
<td>Clinical medicine</td>
<td>110,014</td>
<td>4</td>
<td>2.57</td>
<td>July 31, 2007</td>
</tr>
<tr>
<td>UNIC</td>
<td>[74]</td>
<td>Clinical observations and laboratory tests</td>
<td>40,640</td>
<td>12</td>
<td>2.57</td>
<td>2007-2017</td>
</tr>
<tr>
<td>HLA</td>
<td>[25]</td>
<td>Human relationship studies</td>
<td>72,000</td>
<td>434</td>
<td>1.5</td>
<td>(Refer to the HLA)</td>
</tr>
<tr>
<td>MCR Resources</td>
<td>[54]</td>
<td>Clinical resources</td>
<td>50,000</td>
<td>1</td>
<td>2.48</td>
<td>2007-2008</td>
</tr>
<tr>
<td>KI/IO</td>
<td>[36]</td>
<td>Clinical conditions</td>
<td>12,210</td>
<td>1</td>
<td>1.00</td>
<td>2008 (Biology)</td>
</tr>
<tr>
<td>NCGI</td>
<td>[57]</td>
<td>Biomedical informatics</td>
<td>25,347</td>
<td>120</td>
<td>5.47</td>
<td>Aug 2007</td>
</tr>
<tr>
<td>JBI</td>
<td>[70]</td>
<td>Terminological integration of the life sciences</td>
<td>1,441</td>
<td>1</td>
<td>3.77</td>
<td>2008-2009 (English only)</td>
</tr>
</tbody>
</table>


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?


What is the purpose of the Web Ontology Language OWL?


How do you pronounce all these math expressions?


What are ontological class constructors?

Bhatt et al. (2009)
Medical Classifications

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 mortality 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

239 pages

SNOMED CT® Technical Reference Guide

January 2011 International Release

(US English)


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.

A

24184005|Finding of increased blood pressure (finding) \rightarrow 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) \rightarrow 392570002|Blood pressure finding (finding) AND roleGroup SOME (363714003|Interprets (attribute) SOME 75367002|Blood pressure (observable entity))

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, Psychology [F]
7. Biological Sciences [G]
8. Natural Sciences [H]
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?

How does the MeSH Example Hypertension look?

National Library of Medicine - Medical Subject Headings

2011 MeSH

MeSH Descriptor Data

Standard View: Go to Concept View: Go to Expanded Concept View

Mesh Heading: Hypertension

Tree Number: C14.907.055

Annotation: Not for intracranial or intravascular pressure; relates to BLOOD PRESSURE. Manual (2,3,22); Goldstein, Sidney in HYPERVENTILATION; SUGAR; LACTATE; see HYPERVENTILATION, SUGAR, LACTATE; Hypertension with kidney disease is probably HYPERVENTILATION, SUGAR, LACTATE, and HYPERVENTILATION; venous hypertension. Index under VENOUS PRESSURE. (PM & do not confuse with HYPERVENTILATION, SUGAR, LACTATE, and HYPERVENTILATION; venous hypertension 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 140 mm Hg or more.

Entry Term: Blood Pressure, High

See Also: Perioperative Aspects

Net Also: Venous Resistance

Allowable Qualifiers: BL; OF; CT; DL; ON; CD; ON; DE; ET; EU; OH; ESR; IN; EP; ET; GE; HG; IM; MI; NO; MF; NA; PM; PC; PP; PS; QA; WH; RE; RE; ME; ET; MI; LA; VE; WE

Drug of Choice: [N/A]

Unique ID: D00057

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

How does Mesh look in an interactive tree visualization?

What is UMLS – Unified Medical Language System?

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

What subdomains does UMLS Metathesaurus integrate?

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

Conclusion

- Progress in machine learning is driven by the explosion in the availability of big data and low-cost computation...
- We need top-quality data and/or robust models to deal with the non-iid character of real-world data

Thank you!