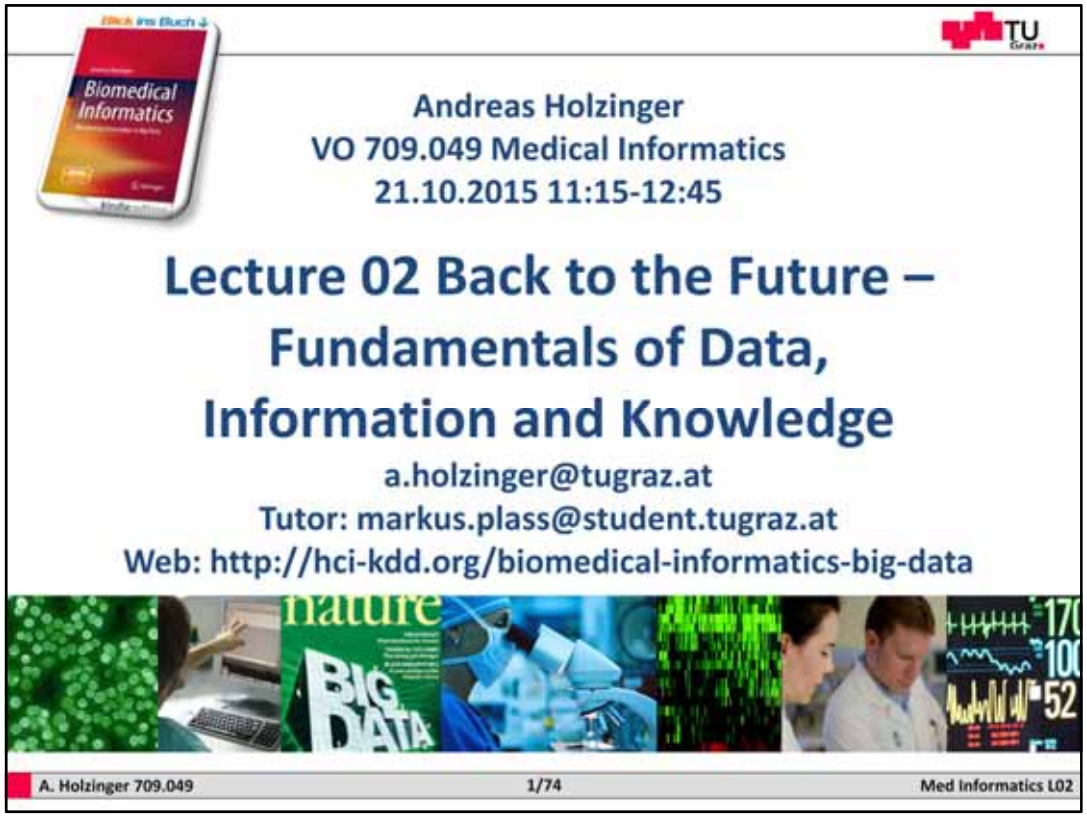
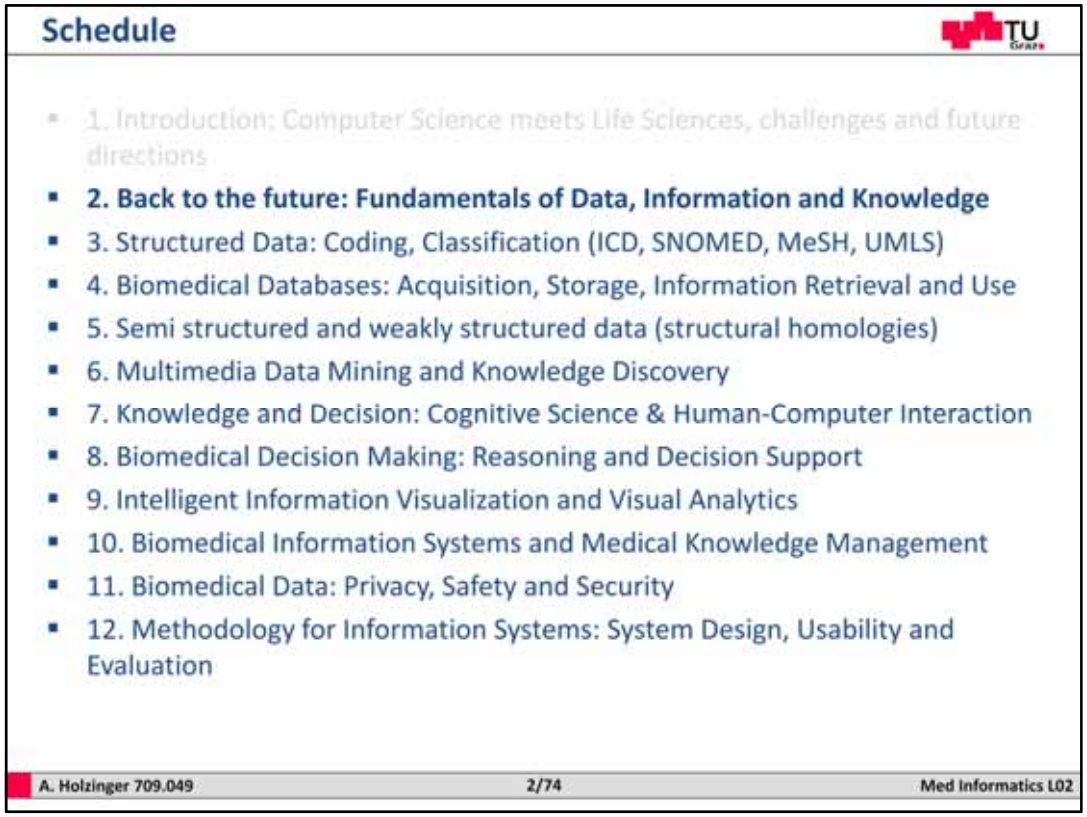
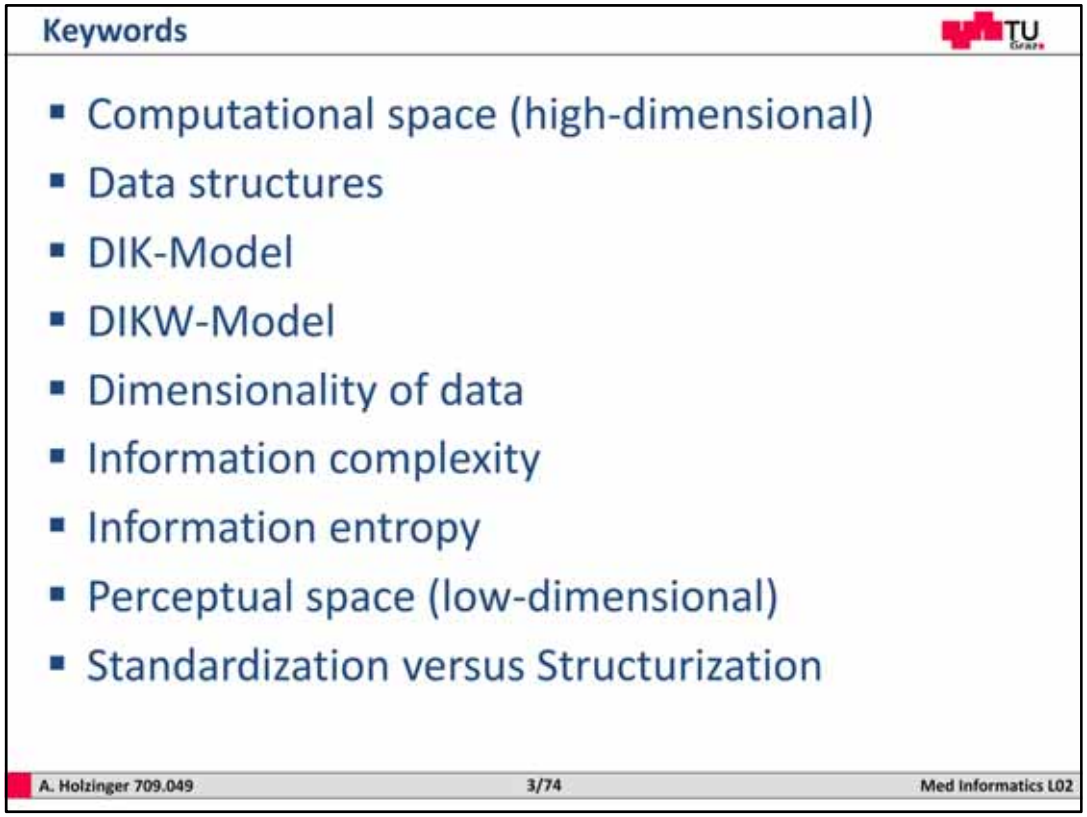


Status as of Mo, 19.10.2015 08:30 Dear Students – welcome to the second lecture of our course "biomedical informatics", please remember from the last lecture the definition: According to the American Association of Medical Informatics (AMIA) the term Medical Informatics has now been expanded to Biomedical Informatics and is defined as "the interdisciplinary field that studies and pursues the effective use of biomedical data, information, and knowledge for scientific inquiry, problem solving, and decision making, motivated by efforts to improve human health". [1]
It is important to know: Bioinformatics + Medical Informatics = Biomedical Informatics (see Slide 1-42).
Note: Computers are just the vehicles to realize the central goals: To harness the power of the machines to support and to amplify human intelligence [2].
[1] Shortliffe, E. H. 2011. Biomedical Informatics: Defining the Science and its Role in Health Professional Education. In: Holzinger, A. & Simonic, K.-M. (eds.) Information Quality in e-Health. Lecture Notes in Computer Science LNCS 7058. Heidelberg, New York: Springer, pp. 711-714.
[2] Holzinger, A. 2013. Human–Computer Interaction and Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? In: Cuzzocrea, A., Kittl, C., Simos, D. E., Weippl, E. & Xu, L. (eds.) Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin, New York: Springer, pp. 319-328.
Regarding the current trend towards personalized medicine have a read of this paper:
Holzinger, A. 2014. Trends in Interactive Knowledge Discovery for Personalized Medicine: Cognitive Science meets Machine Learning. IEEE Intelligent Informatics Bulletin, 15, (1), 6-14.
Online available via:
http://www.comp.hkbu.edu.hk/~cib/2014/Dec/article2/iib_vol15no1_article2.pdf



In this second lecture we start with a look on data sources, review some data structures, discuss standardization versus structurization, review the differences between data, information and knowledge and close with an overview about information entropy.



A central topic is the dimensionality of data and the interrelated (connected) curse of dimensionality which refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of our everyday world.

Learning Goals

- … be aware of the types and categories of different data sets in biomedical informatics;
- … know some differences between data, information, knowledge and wisdom;
- … be aware of standardized/non-standardized and well-structured/un-structured data;
- … have a basic overview on information theory and the concept of information entropy;

Advance Organizer (1/2)

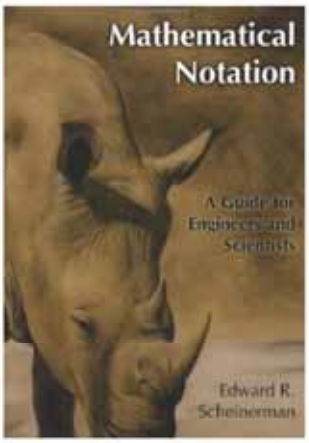
- **Abduction** = cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those (i.e., evaluation of each generated hypothesis on the basis of its expected consequences)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

Advance Organizer (2/2)

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

Common Mathematical Notations with LaTeX commands

"In mathematics you don't understand things. You just get used to them" – John von Neumann



Data

n	Number of samples
d	Number of input variables
$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$	Matrix of input samples
$\mathbf{y} = [y_1, \ldots, y_n]$	Vector of output samples
$\mathbf{Z} = [\mathbf{X}, \mathbf{y}]$	Combined input–output training data or
$\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$	Representation of data points in a feature space

Distribution

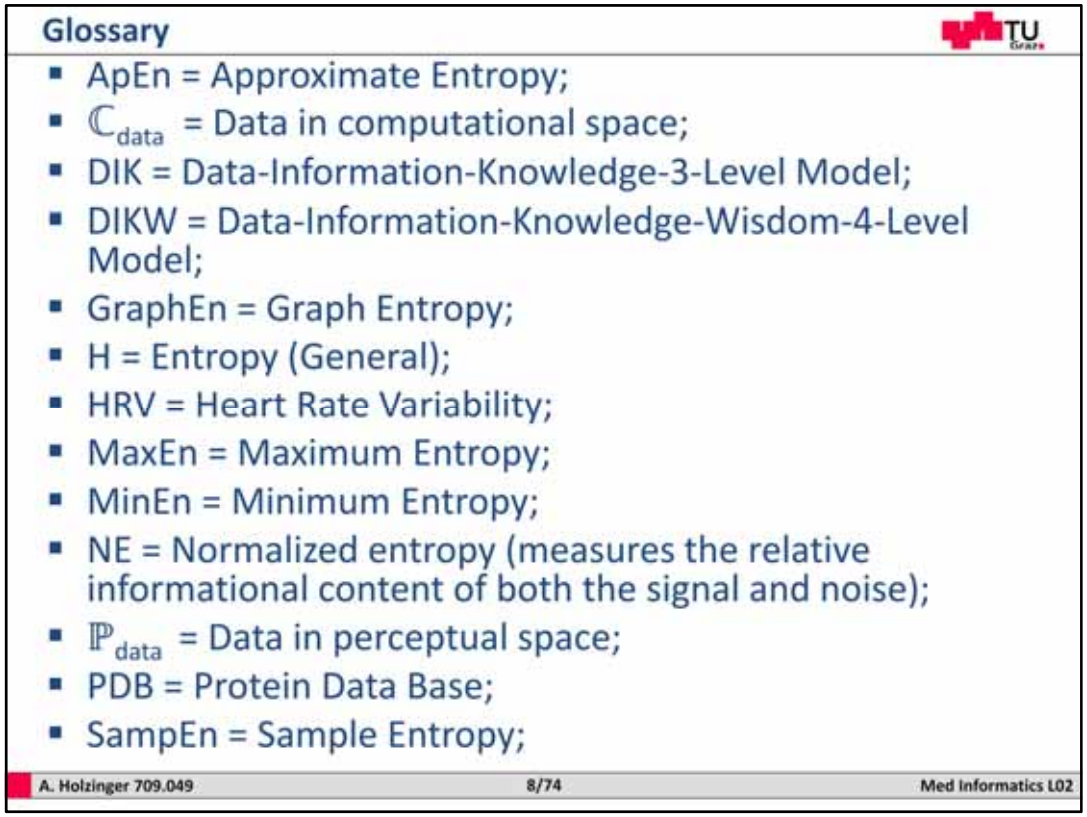
P	Probability
$F(\mathbf{x})$	Cumulative probability distribution function (cdf)
$p(\mathbf{x})$	Probability density function (pdf)
$p(\mathbf{x}, y)$	Joint probability density function
$p(\mathbf{x}; \omega)$	Probability density function, which is parameterized
$p(y\|\mathbf{x})$	Conditional density
$t(\mathbf{x})$	Target function

A. Holzinger 709.049 7/74 Med Informatics L02

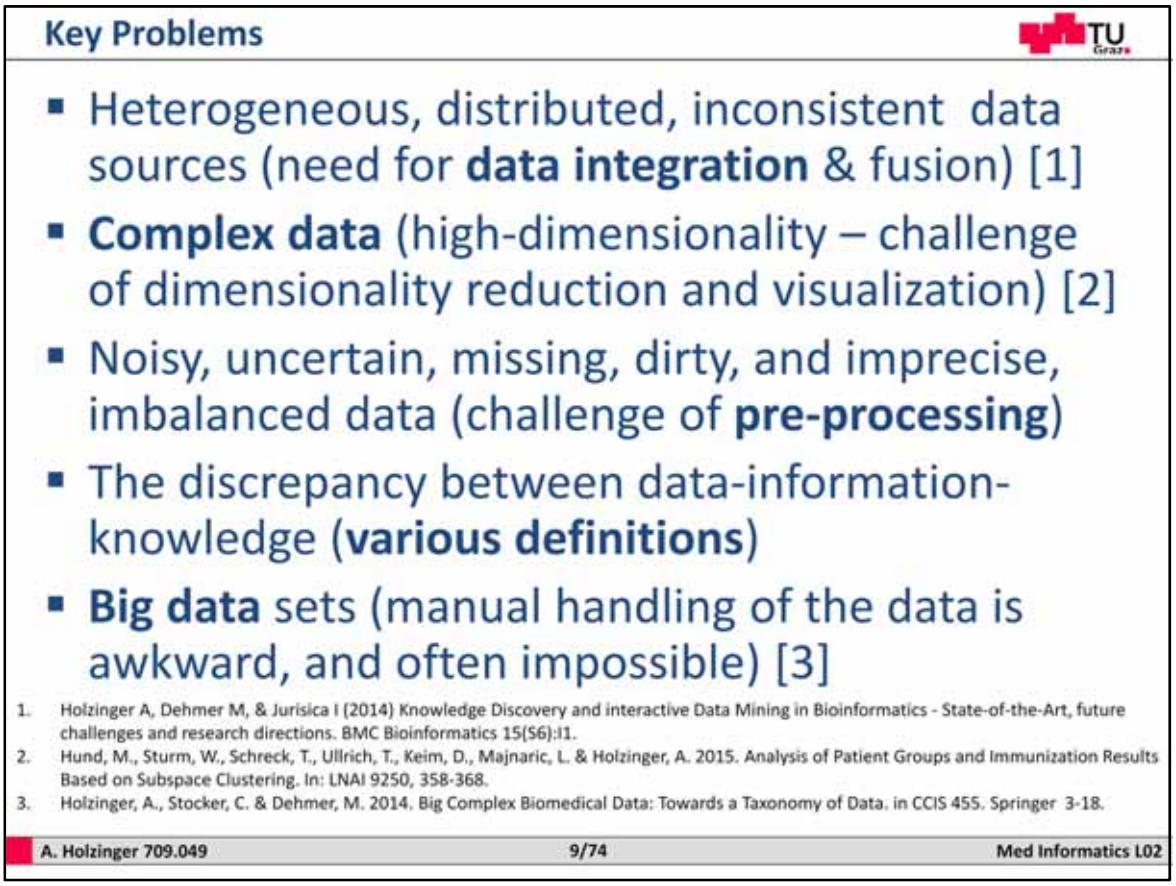
A recommendable small booklet is:

Scheinerman, E. R. 2011. Mathematical Notation: A Guide for Engineers and Scientists, Baltimore (MD), Scheinerman.
Which also includes the most important LATEX commands for producing maths symbols

http://www.ams.jhu.edu/~ers/notation/



components through p and w (Golan, Judge, and Miller; 1996);



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.

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: Guo, Y., Friston, K., Aldo, F., Hill, S. & Peng, H. (eds.) Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Cham: Springer International Publishing, pp. 358-368.

Holzinger, A., Stocker, C. & Dehmer, M. 2014. Big Complex Biomedical Data: Towards a Taxonomy of Data. In: Obaidat, M. S. & Filipe, J. (eds.) Communications in Computer and Information Science CCIS 455. Berlin Heidelberg: Springer pp. 3-18.

Related recommended reading:

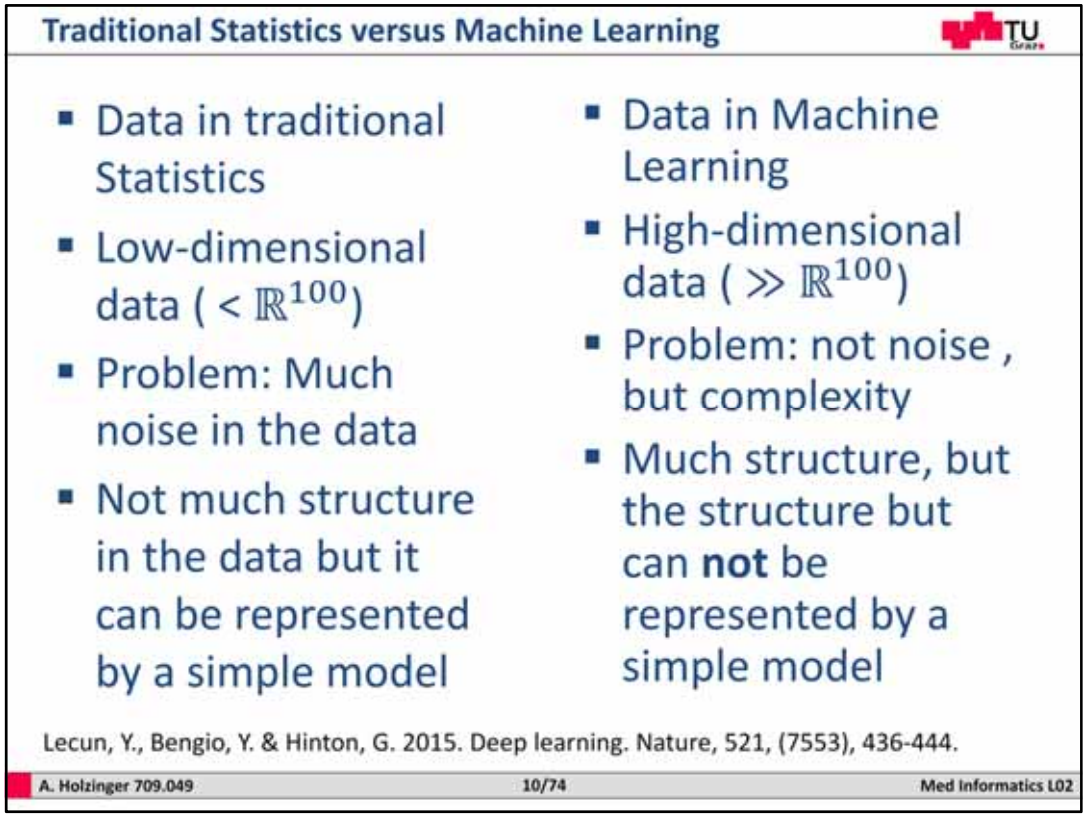
Dong-Hee, S. & Min Jae, C. 2015. Ecological views of big data: Perspectives and issues. Telematics and Informatics, 32, (2), 311-320.

Dong, X. L. & Srivastava, D. 2015. Big Data Integration. Synthesis Lectures on Data Management, 7, (1), 1-198.

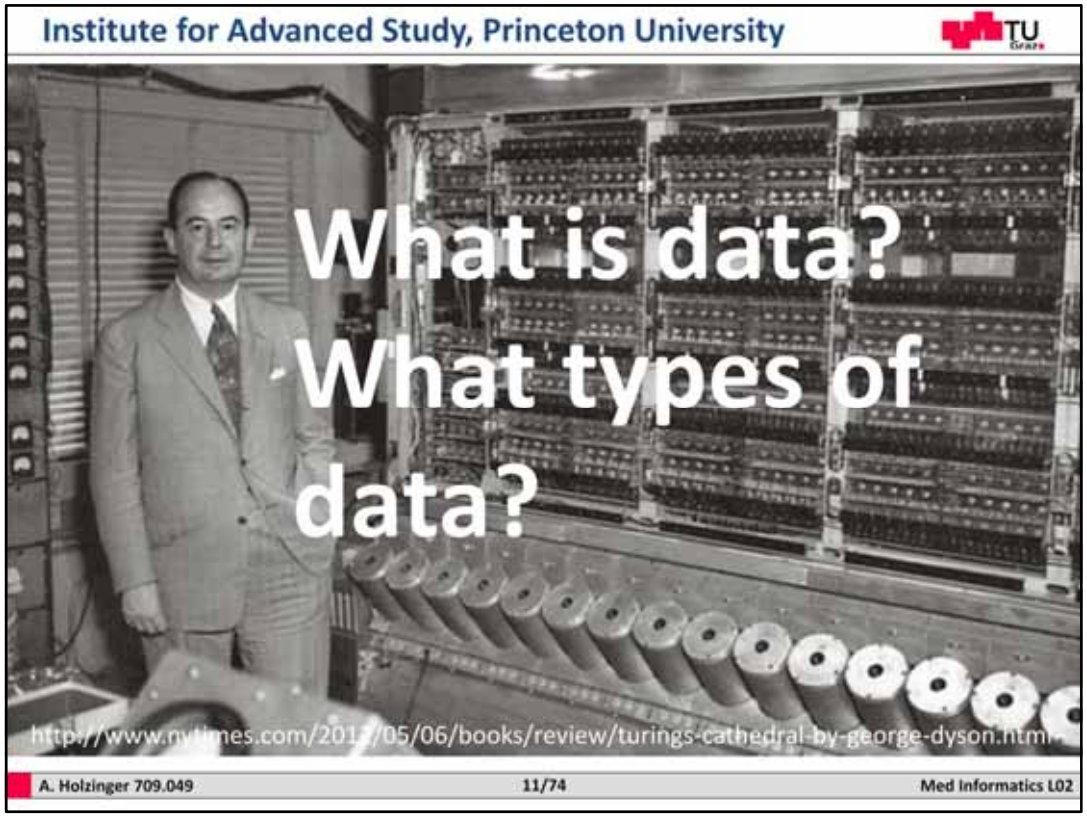
Wu, X. D., Zhu, X. Q., Wu, G. Q. & Ding, W. 2014. Data Mining with Big Data. IEEE Transactions on Knowledge and Data Engineering, 26, (1), 97-107.

Shneiderman, B. 2014. The Big Picture for Big Data: Visualization. Science, 343, (6172), 730-730.

Sackman, J. E. & Kuchenreuther, M. 2014. Marrying Big Data with Personalized Medicine. Biopharm International, 27, (8), 36-38.

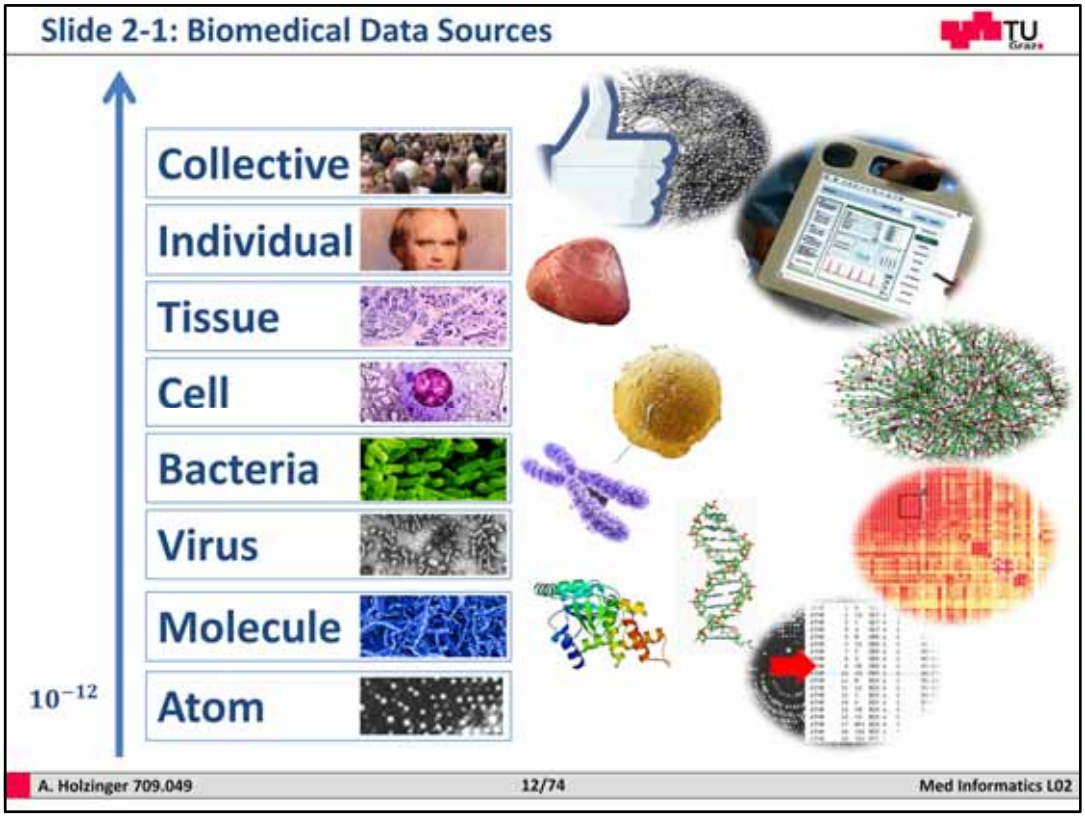


With regard to data, the difference between classical statistics and modern machine learning is that machine learning discovers intricate structures in large data sets to indicate how a machine should change its internal parameters.



John von Neumann and his high-speed computer, approx. 1952

Our first question is: Where does the data come from? The second question: What kind of data is this? The third question: How big is this data? So, let us look at some biomedical data sources (see Slide 2-1):

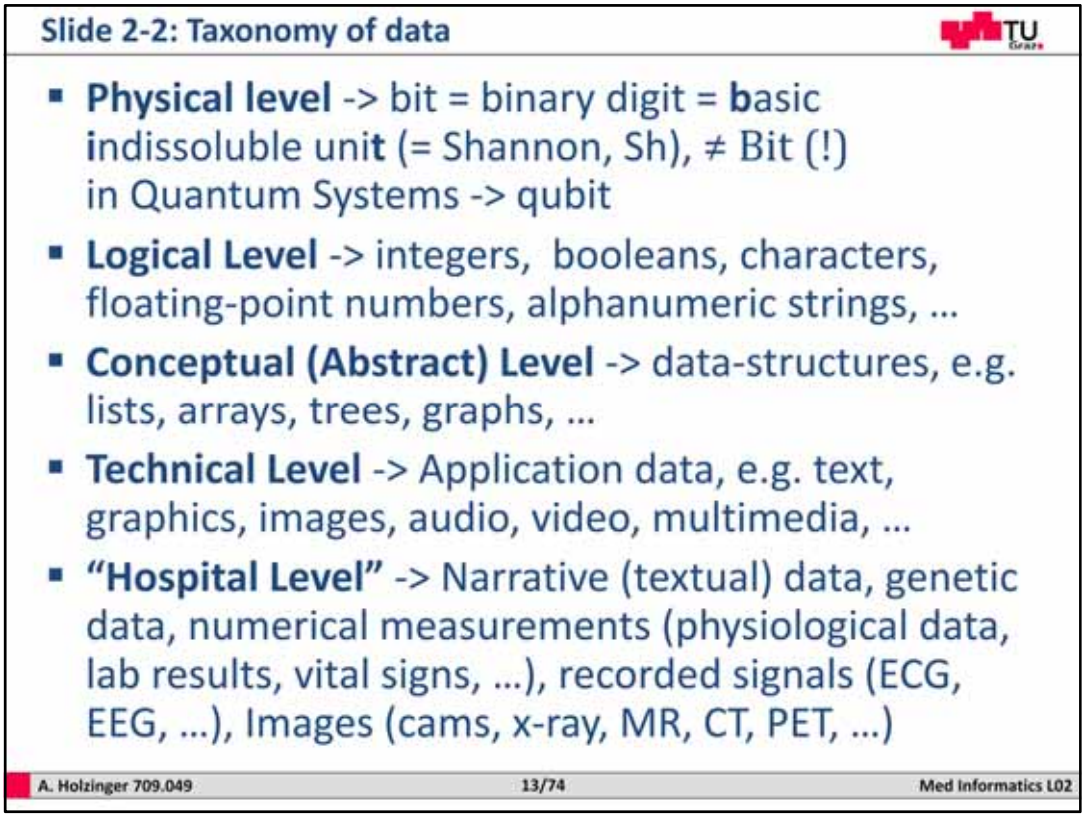


Due to the increasing trend towards personalized and molecular medicine, biomedical data results from various sources in different structural dimensions, ranging from the microscopic world (e.g. genomics, epigenomics, metagenomics, proteomics, metabolomics) to the macroscopic world (e.g. disease spreading data of populations in public health informatics). Just for orientation: the Glucose molecule has a size of 900 pm = 900 × $10^{-12} m$ and the Carbon atom approx. 300 pm. A hepatitis virus is relatively large with 45 nm = 45 × 10^{-9} and the X-Chromosome much bigger with 7 μm = 7 × 10^{-6} m.
Here a lot of "big data" is produced, e.g. genomics, metabolomics and proteomics data. This is really "big data" – the data sets enormously large – whereas in each individual we estimate many Terabytes (1 TB = 1 × 10^{12} Byte = 1000 GByte) of genomics data, we are confronted with Petabytes of proteomics data and the fusion of those for personalized medicine results in Exabytes of data (1 EB = 1 × 10^{18} Byte).

Of course these amounts are for each human individual, however, we have a current world population of 7 Billion (1 Billion in English language is 1 Milliard in European language) people (= 7 × 10^{9} people). So you can see that this is really "big data". This "natural" data is then fused with "produced" data, e.g. the unstructured data (text) in the patient records, or data from physiological sensors etc. – these data is also rapidly increasing in size and complexity. You can imagine that without computational intelligence we have no chance to survive in this complex big data sets.

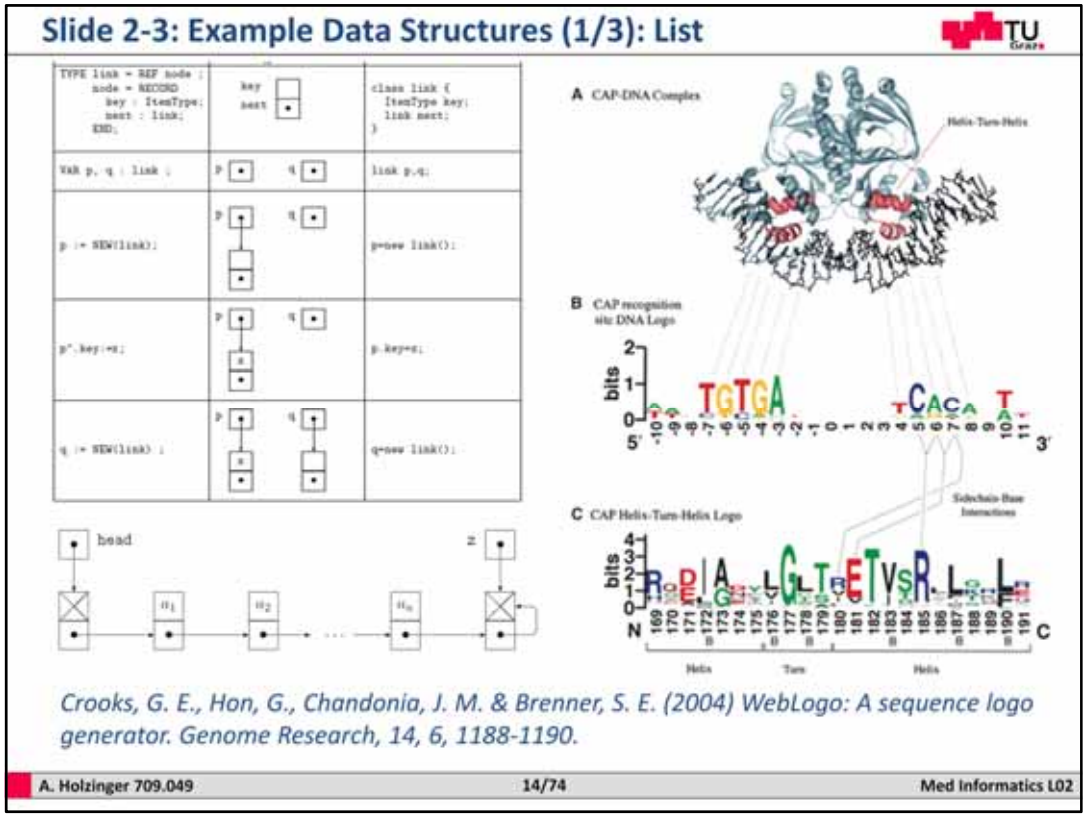
http://learn.genetics.utah.edu/content/begin/cells/scale/
C-Atom 340 pm = 340 . 10-12 m
Molecule Glucose 900 pm
Virus Hepatitis Virus 45 nm = 45. 10-9 m
Microscope 200.10-9 m
Confocalmicroscopy 20.10-6 m
Electron-Microscopy 0,1.10-9 m
X-Chromosome 7.10-6 m
DNA 2.10-9 m
Encyme = Metabolomics

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.



Most of our computers are Von-Neumann machines (see chapter 1), consequently at the lowest physical layer, data is represented as patterns of electrical on/off states (1/0, H/L, high/low); we speak of a **bit,** which is also known as Bit, the **B**asic indissoluble **i**nformation uni**t** (Shannon, 1948). Do not confuse this Bit with the IEC 60027-2 symbol bit – in small letters – which is used as an SI dimension prefix (e.g. 1 Kbit = 1024 bit, 1 Byte = 8 bit). Beginning with the physical level of data we can determine various levels of data structures (see Slide 2-2):
Refer to: http://physics.nist.gov/cuu/Units/binary.html
1) Physical level: in a Von-Neumann system: bit; in a Quantum system: qubit
Note: Regardless of its physical realization (e.g. voltage, or mechanical state, or black/white etc.), a bit is always logically either 0 or 1 (analog to a light-switch). A qubit has similarities to a classical bit, but is overall very different: A classical bit is a scalar variable with the single value of either 0 or 1, so the value is unique, deterministic and unambiguous. A qubit is more general in the sense that it represents a state defined by a pair of complex numbers (a, b), which express the probability that a reading of the value of the qubit will give a value of 0 or 1. Thus, a qubit can be in the state of 0, 1, or some mixture - referred to as a superposition - of the 0 and 1 states. The weights of 0 and 1 in this superposition are determined by (a, b) in the following way: qubit $\triangleq (a, b) \triangleq a \cdot 0_{bit} + b \cdot 1_{bit}$. Please be aware that this model of quantum computation is not the only one (Lanzagorta & Uhlmann, 2008).

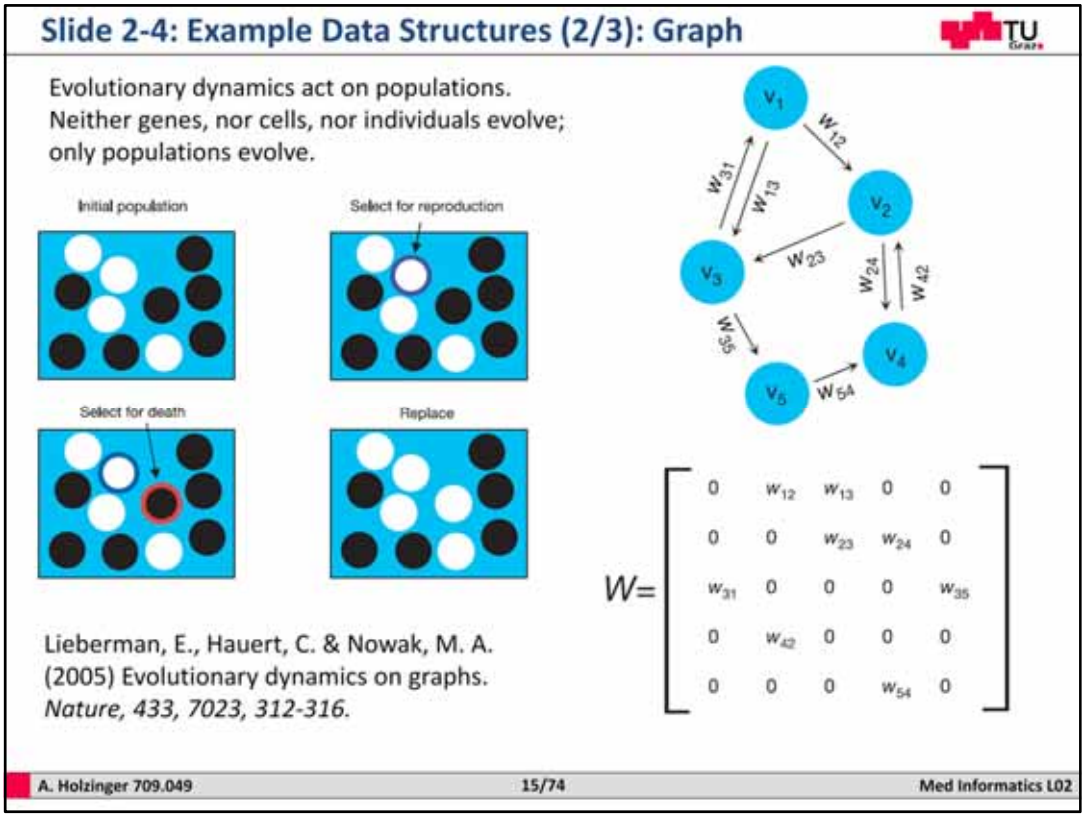
For a recent overview on quantum computation please refer to: http://peterwittek.com/book.html
2) Logical Level:
1) Primitive data types, including:
a) Boolean data type (true/false);
b) numerical data type (e.g. integer ($\mathbb{Z}$), floating-point numbers ("reals"), etc.);
2) composite data types, including: a) array, b) record, c) union, d) set (stores values without any particular order, and no repeated values), e) object (contains others);
3) String and text types, including:
a) alphanumeric characters,
b) alphanumeric strings (= sequence of characters to represent words and text)
3) Abstract Level: including abstract data structures, e.g. queue (FIFO), stack (LIFO), set (no order, no repeated values), lists, hash table, arrays, trees, graphs, …
4) Technical Level: Application data formats, e.g. text, vector graphics, pixel images, audio signals, video sequences, multimedia, …
5) Hospital Level: Narrative (textual, natural language) patient record data (structured/unstructured and standardized/non-standardized), Omics data (genomics, proteomics, metabolomics, microarray data, fluxomics, phenomics), numerical measurements (physiological data, time series, lab results, vital signs, blood pressure, CO_2 partial pressure, temperature, …), recorded signals (ECG, EEG, ENG, EMG, EOG, EP …), graphics (sketches, drawings, handwriting, …); audio signals, images (cams, x-ray, MR, CT, PET, …), etc.



In biomedical informatics we have a lot to do with abstract data types (ADT), consequently we briefly review the most important ones here. For details please refer to a course on Algorithm & Data structures, or to a classic textbook such as (Aho, Hopcroft & Ullman, 1983), (Cormen et al., 2009), or in German (Ottmann & Widmayer, 2012), (Holzinger, 2003) and please take into consideration that data structures and algorithms go hand in hand, so a must-have-on-the-desk of every computer scientist is: Cormen, T. H., Leiserson, C. E., Rivest, R. L. & Stein, C. 2009. Introduction to Algorithms (3rd edition), Cambridge (MA), The MIT Press.

List is a sequential collection of items $a_1, a_2, \dots, a_n$ accessible one after another, beginning at the head and ending at the tail z. In a Von-Neumann machine it is a widely used data structure for applications which do not need random access. It differs from the stack (last-in-first-out, LIFO) and queue (first-in-first-out, FIFO) data structures insofar, that additions and removals can be made at any position in the list. In contrast to a simple set S the order is important. A typical example for the use of a list is a DNA sequence. The combination of GGGTTTAAA is such a list, the elements of the list are the nucleotide bases.

Nucleotides are the joined molecules which form the structural units of the RNA and the DNA and play the central role in metabolism.

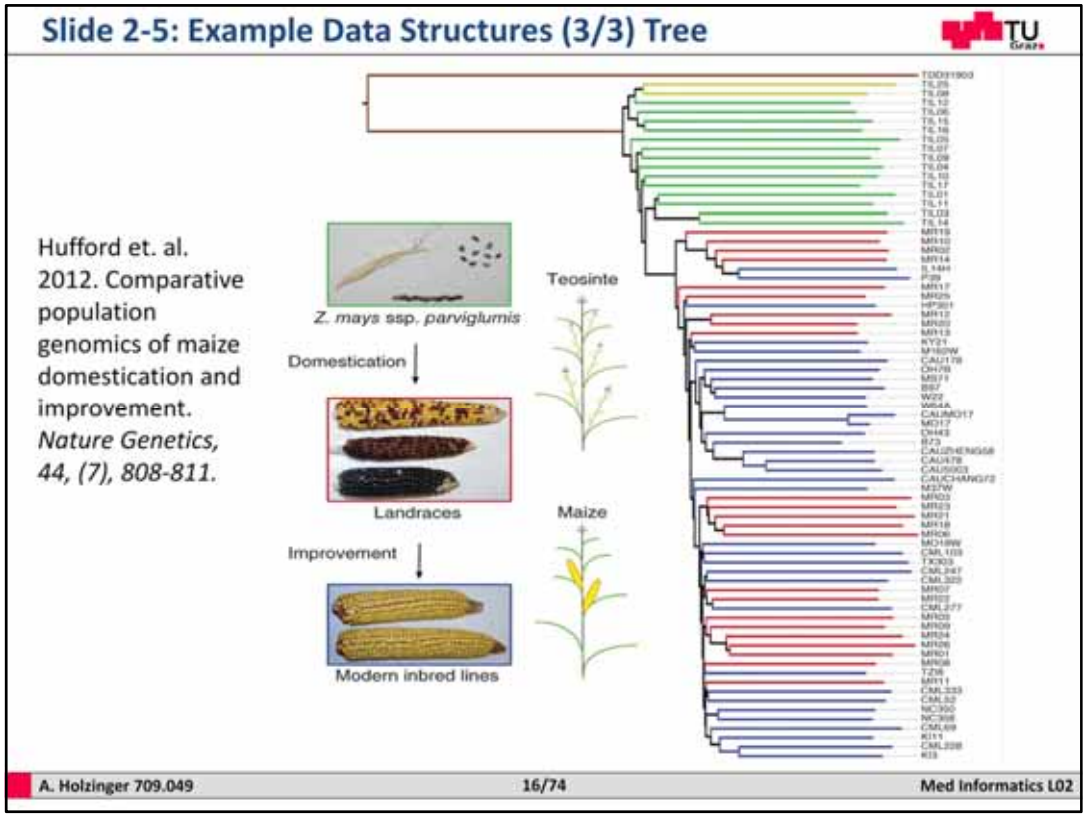


Graph is a pair $G = (V, E)$, where $V(G)$ is a set of finite, non-empty vertices (nodes) and $E(G)$ is a set of edges (lines, arcs), which are 2-element subsets of V. If E is a set of ordered pairs of vertices (arcs, directed edges, arrows), then it is a **directed graph** (digraph). The distances between the edges can be represented within a distance-matric (two dimensional array).

The edges in a graph can be *multidimensional objects,* e.g. vectors containing the results of multiple Gen-expression measures. For this purpose the distance of two edges can be measured by various distance metrics.
Graphs are ideally suited for representing networks in medicine and biology, e.g. metabolism pathways, etc.
In bioinformatics, distance matrices are used to represent protein structures in a coordinate-independent manner, as well as the pairwise distances between two sequences in sequence space. They are used in structural and sequential alignment, and for the determination of protein structures from NMR or X-ray crystallography. Evolutionary dynamics act on populations. Neither genes, nor cells, nor individuals evolve; only *populations evolve.*
This so called **Moran process** describes the stochastic evolution of a finite population of constant size: In each time step, an individual is chosen for reproduction with a probability proportional to its fitness; a second individual is chosen for death. The offspring of the first individual replaces the second and individuals occupy the vertices of a graph. In each time step, an individual is selected with a probability proportional to its fitness; the weights of the outgoing edges determine the probabilities that the corresponding neighbor will be replaced by the offspring. The process is described by a stochastic matrix W, where w_{ij} denotes the probability that an offspring of individual i will replace individual j. At each time step, an edge ij is selected with a probability proportional to its weight and the fitness of the individual at its tail. The Moran process is a complete graph with identical weights (Lieberman, Hauert & Nowak, 2005).

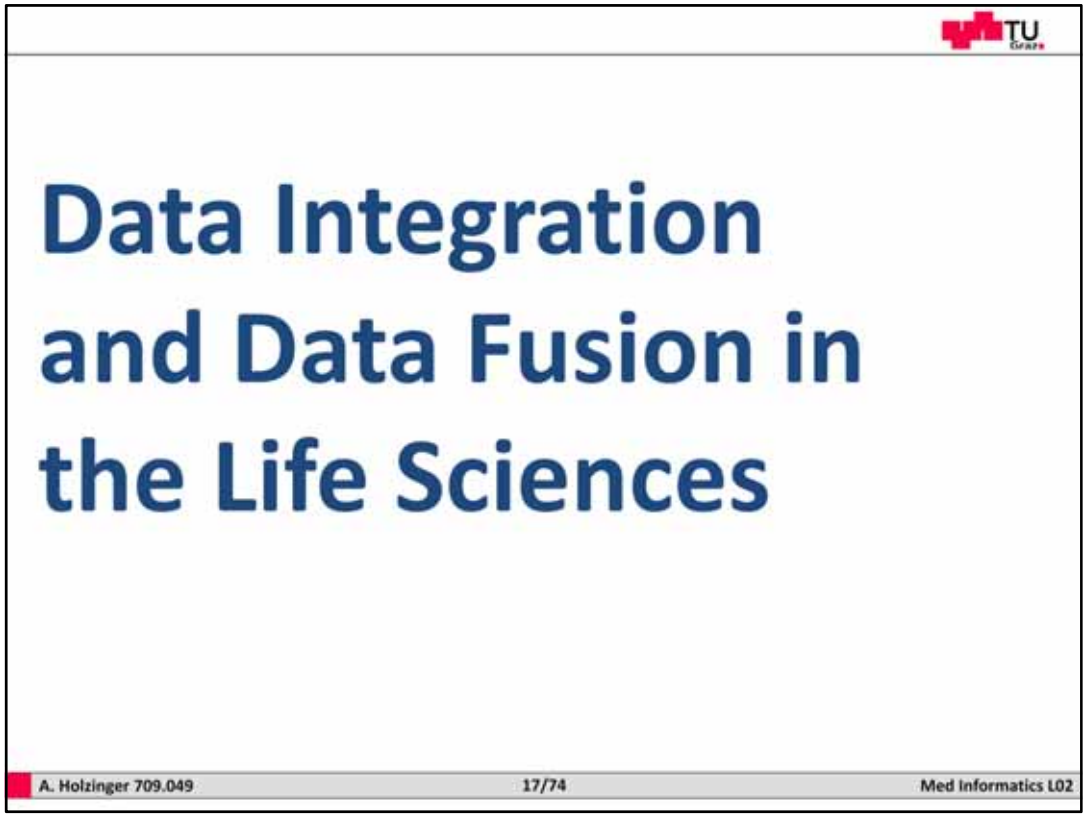
Graphs can be represented computationally by an Adjacency list, Adjacency matrix and an Incidence matrix. The first pre-processing step is to produce point cloud data sets from raw data, see:
Holzinger, A., Malle, B., Bloice, M., Wiltgen, M., Ferri, M., Stanganelli, I. & Hofmann-Wellenhof, R. 2014. On the Generation of Point Cloud Data Sets: Step One in the Knowledge Discovery Process. 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. 57-80.
https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=974579&pCurrPk=83005

For the specific task of getting graphs from image data have a look at: Holzinger, A., Malle, B. & Giuliani, N. 2014. On Graph Extraction from Image Data. In: Slezak, D., Peters, J. F., Tan, A.-H. & Schwabe, L. (eds.) Lecture Notes in Artificial Intelligence, LNAI 8609. Heidelberg, Berlin: Springer, pp. 552-563.
https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=868952&pCurrPk=80830

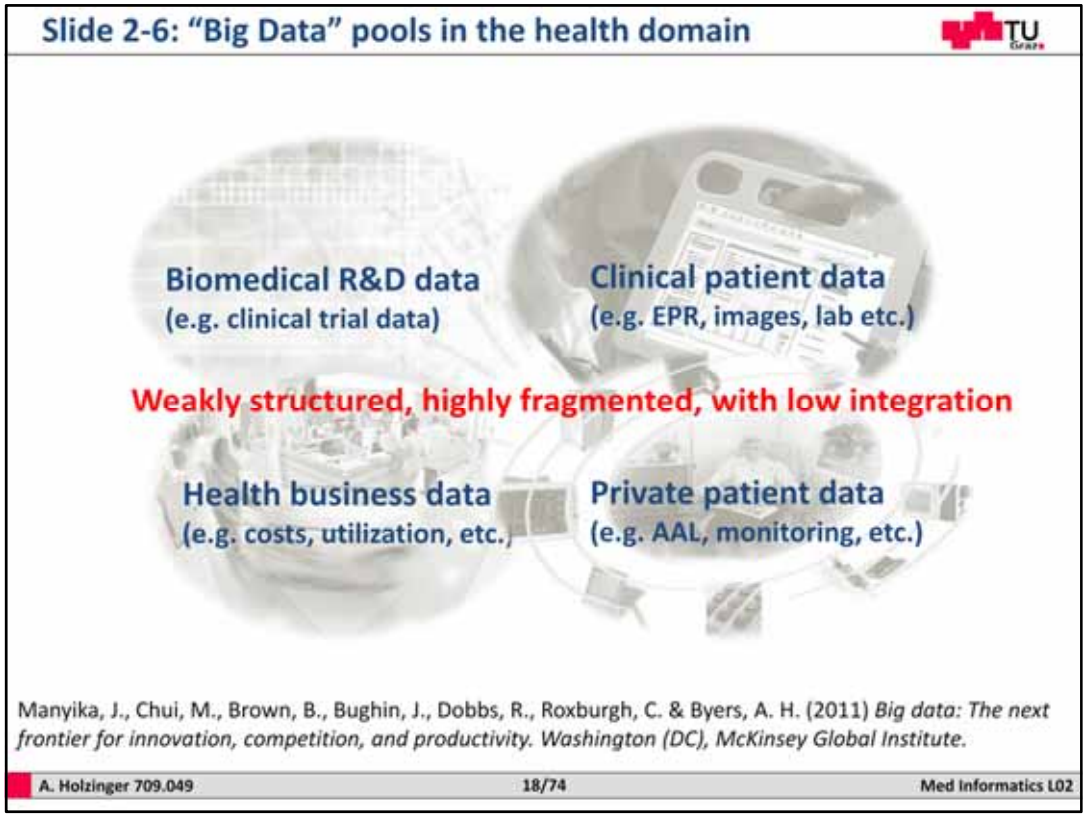


Tree is a collection of elements called nodes, one of which is distinguished as a root, along with a relation ("parenthood") that places a hierarchical structure on the nodes. A node, like an element of a list, can be of whatever type we wish. We often depict a node as a letter, a string, or a number with a circle around it. Formally, a tree can be defined recursively in the following manner:
1. A single node by itself is a tree. This node is also the root of the tree.
2. Suppose n is a node and $T1, T2, \ldots, Tk$ are trees with roots $n1, n2, \ldots, nk$, respectively. We can construct a new tree by making n be the parent of nodes $n1, n2, \ldots, nk$. In this tree n is the root and $T1, T2, \ldots, Tk$ are the subtrees of the root. Nodes $n1, n2, \ldots, nk$ are called the children of node n.

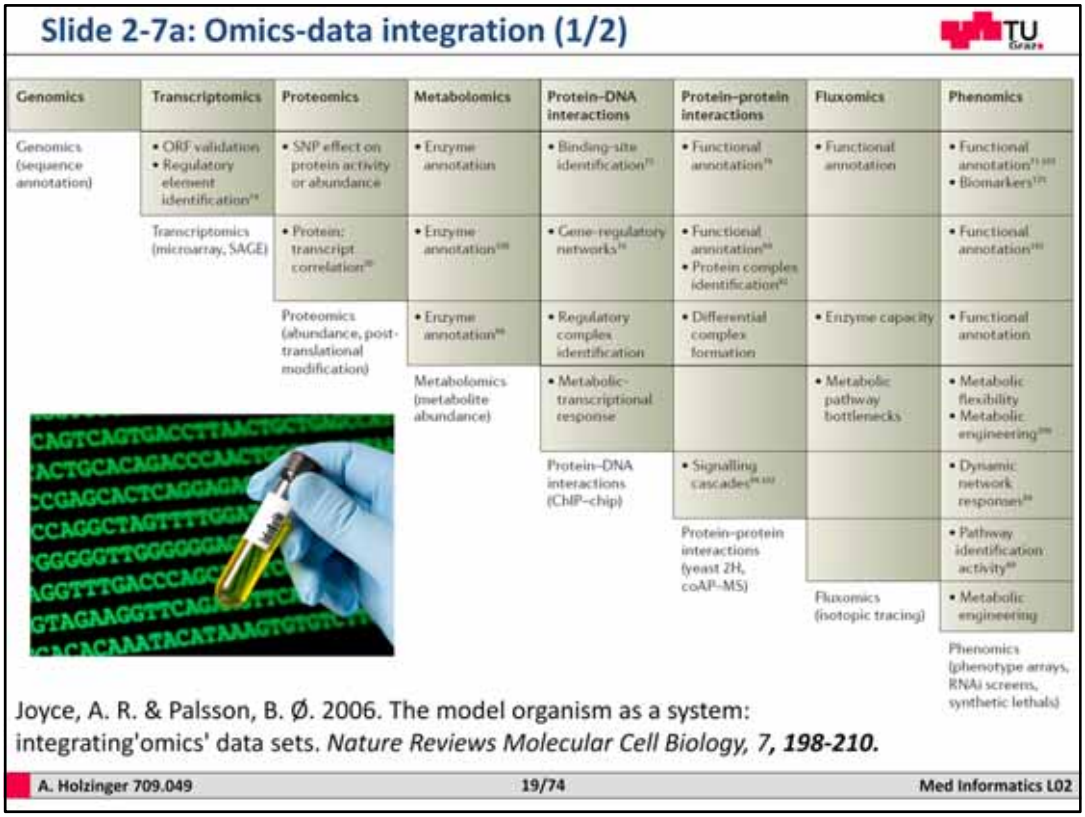
Dendrogram (from Greek dendron "tree", -gramma "drawing") is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering. Dendrograms are often used in computational biology to illustrate the clustering of genes or samples. The origin of such dendrograms can be found in (Darwin, 1859).
The example by (Hufford et al., 2012) shows a neighbor-joining tree and the changing morphology of domesticated maize and its wild relatives. Taxa in the neighbor-joining tree are represented by different colors: parviglumis (green), landraces (red), improved lines (blue), mexicana (yellow) and Tripsacum (brown). The morphological changes are shown for female inflorescences and plant architecture during domestication and improvement.



Please remember the key problems in dealing with data include:
1) Heterogeneous data sources (need for data fusion and data integration)
2) Complexity of the data (high-dimensionality)
3) Noisy, uncertain data (challenge of pre-processing)
4) The discrepancy between data-information-knowledge (various definitions)
5) Big data sets (manual handling of the data is impossible)

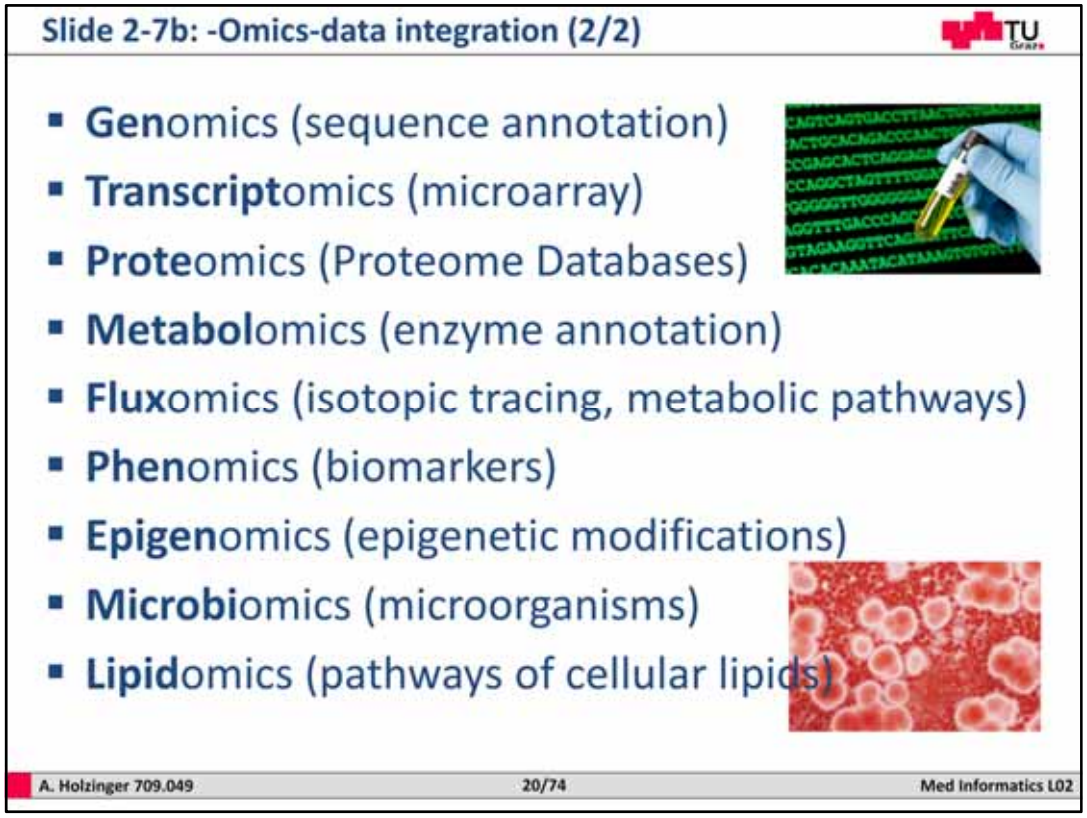


Now that we have seen some examples of data from the biomedical domain, we can look at the "big picture". Manyika et al. (2011) localized four major data pools in the US health care and describe that the data are highly fragmented, with little overlap and low integration. Moreover, they report that approx. 30 % of clinical text/numerical data in the United States, including medical records, bills, laboratory and surgery reports, is still not generated electronically. Even when clinical data are in digital form, they are usually held by an individual provider and rarely shared (see Slide 2-4).
Biomedical research data, e.g. clinical trials, predictive modeling etc., is produced by academia and pharmaceutical companies and stored in data bases and libraries. Clinical data is produced in the hospital and are stored in hospital information systems (HIS), picture archiving and communication systems (PACS) or in laboratory data bases, etc. Much data is health business data produced by payors, providers, insurances, etc. Finally, there is an increasing pool of patient behavior and sentiment data, produced by various customers and stakeholders, outside the typical clinical context, including the growing data from the wellness and ambient assisted living domain.



A major challenge in our networked world is the increasing amount of data – today called "big data". The trend towards personalized medicine has resulted in a sheer mass of the generated (-omics) data, (see Slide 2-7). In the life sciences domain, most data models are characterized by complexity, which makes manual analysis very time-consuming and frequently practically impossible (Holzinger, 2013).

More and more Omics-data are generated, including:
1) Genomics data (e.g. sequence annotation),
2) Transcriptomics data (e.g. microarray data); the **transcriptome** is the set of all RNA molecules, including mRNA, rRNA, tRNA and non-coding RNA produced in the cells.
3) Proteomics data: Proteomic studies generate large volumes of raw experimental data and inferred biological results stored in data repositories, mostly openly available; an overview can be found here: (Riffle & Eng, 2009). The outcome of proteomics experiments is a list of proteins differentially modified or abundant in a certain phenotype. The large size of proteomics datasets requires specialized analytical tools, which deal with large lists of objects 4) Metabolomics (e.g. enzyme annotation), the **metabolome** represents the collection of all metabolites in a cell, tissue, organ or organism.
5) Protein-DNA interactions,
6) Protein-protein interactions; PPI are at the core of the entire **interactomics** system of any living cell.
7) Fluxomics (isotopic tracing, metabolic pathways),
8) Phenomics (biomarkers),
9) Epigenetics, is the study of the changes in gene expression – others than the DNA sequence, therefore the prefix "epi-"
10) Microbiomics
11) Lipidomics
Omics-data integration helps to address interesting biological questions on the biological systems level towards personalized medicine (Joyce & Palsson, 2006).



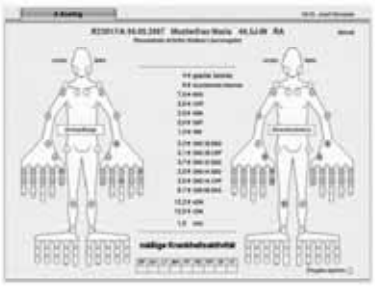
More and more Omics-data are generated, including:
1) Genomics data (e.g. sequence annotation),
2) Transcriptomics data (e.g. microarray data); the transcriptome is the set of all RNA molecules, including mRNA, rRNA, tRNA and non-coding RNA produced in the cells.
3) Proteomics data: Proteomic studies generate large volumes of raw experimental data and inferred biological results stored in data repositories, mostly openly available; an overview can be found here: (Riffle & Eng, 2009). The outcome of proteomics experiments is a list of proteins differentially modified or abundant in a certain phenotype. The large size of proteomics datasets requires specialized analytical tools, which deal with large lists of objects (Bessarabova et al., 2012).
4) Metabolomics (e.g. enzyme annotation), the metabolome represents the collection of all metabolites in a cell, tissue, organ or organism.
5) Protein-DNA interactions,
6) Protein-protein interactions; PPI are at the core of the entire interactomics system of any living cell.
7) Fluxomics (isotopic tracing, metabolic pathways),
8) Phenomics (biomarkers),
9) Epigenetics, is the study of the changes in gene expression – others than the DNA sequence, therefore the prefix "epi-"
10) Microbiomics
11) Lipidomics
Omics-data integration helps to address interesting biological questions on the biological systems level towards personalized medicine (Joyce & Palsson, 2006).

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2908408/

For more information please refer to: Gomez-Cabrero, D., Abugessaisa, I., Maier, D., Teschendorff, A., Merkenschlager, M., Gisel, A., Ballestar, E., Bongcam-Rudloff, E., Conesa, A. & Tegner, J. 2014. Data integration in the era of omics: current and future challenges. BMC Systems Biology, 8, (Suppl 2), I1.
http://www.biomedcentral.com/1752-0509/8/S2/I1

Slide 2-8: Example of typical clinical data sets

- 50+ Patients per day ~ 5000 data points per day …
- Aggregated with specific scores (Disease Activity Score, DAS)
- Current patient status is related to previous data
- = convolution over time
- ⇒ **time-series data**

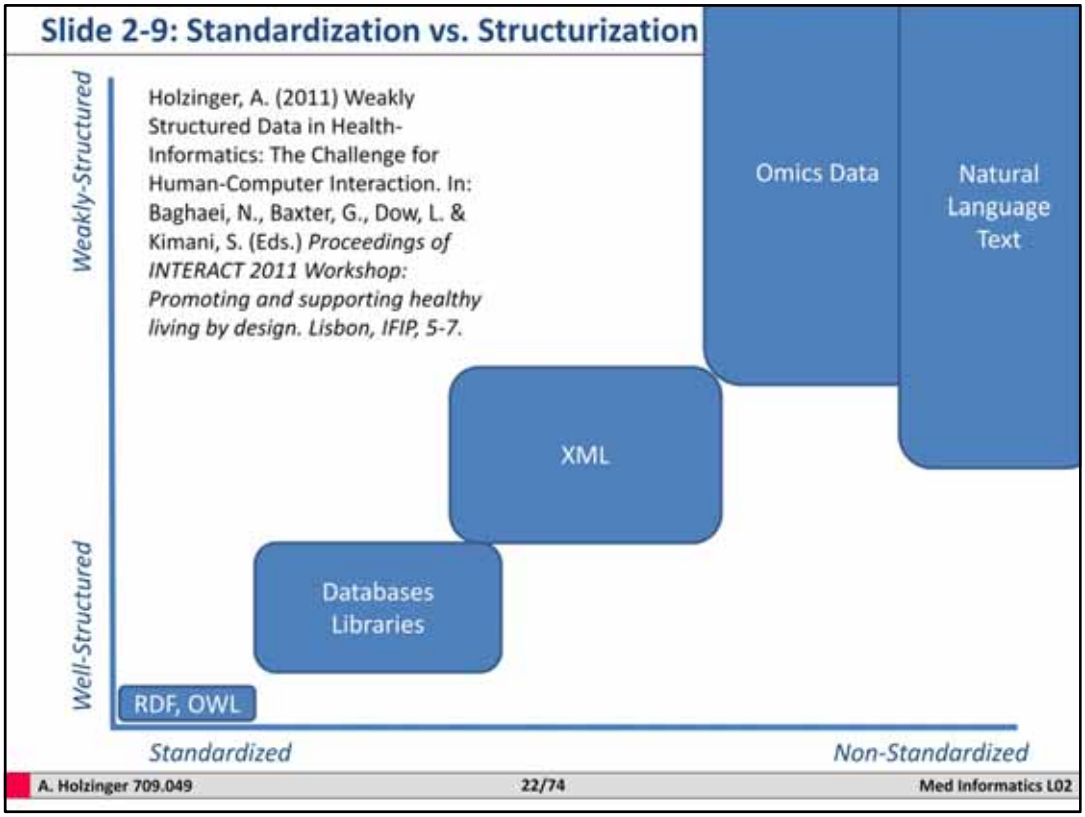


Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. (2011). *Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, Dublin, IEEE, 550-554.*

A. Holzinger 709.049 21/74 Med Informatics L02

A further challenge is to integrate the data and to make it accessible to the clinician. While there is much research on the integration of heterogeneous information systems, a shortcoming is in the integration of available data. Data fusion is the process of merging multiple records representing the same real-world object into a single, consistent, accurate, and useful representation (Bleiholder & Naumann, 2008).
An example for the mix of different data for solving a medical problem can be seen in Slide 2-8.

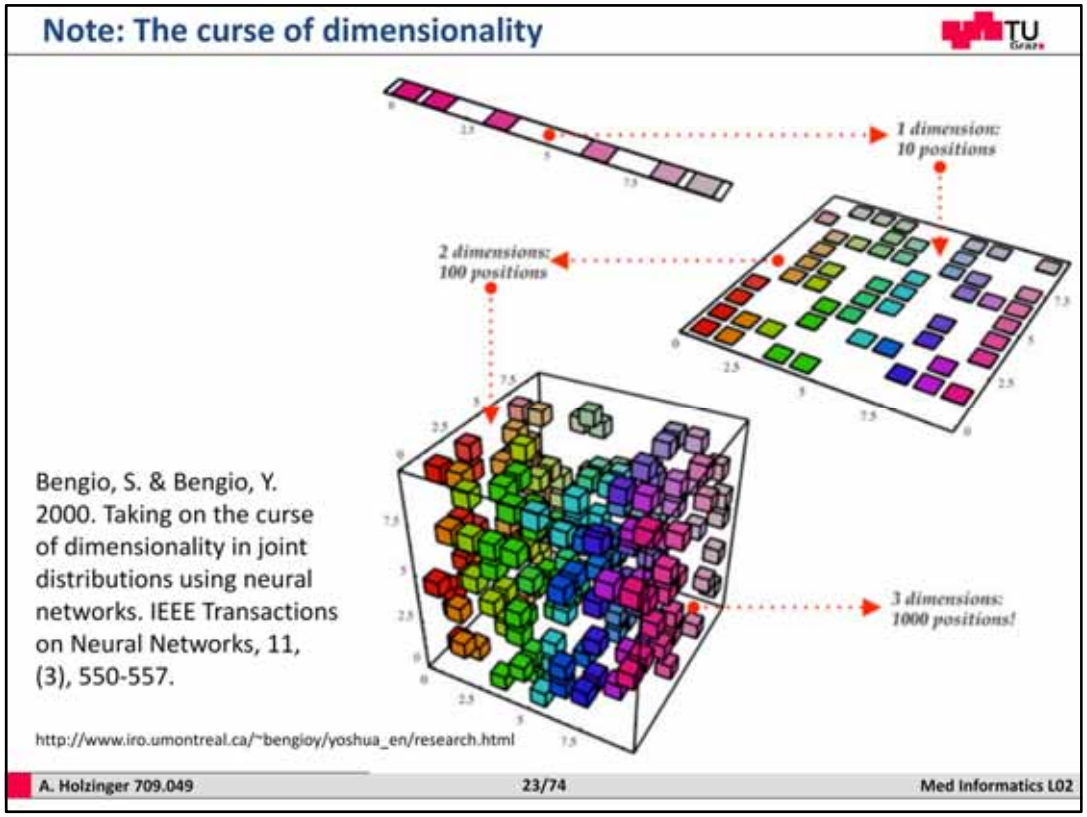
A good example for complex medical data is RCQM, which is an application that manages the flow of data and information in the rheumatology outpatient clinic (50 patients per day, 5 days per week) of Graz University Hospital, on the basis of a quality management process model. Each examination produces 100+ clinical and functional parameters per patient. This amassed data are morphed into better useable information by applying scoring algorithms (e.g. Disease Activity Score, DAS) and are convoluted over time. Together with previous findings, physiological laboratory data, patient record data and Omics data from the Pathology department, these data constitute the information basis for analysis and evaluation of the disease activity. The challenge is in the increasing quantities of such highly complex, multi-dimensional and time series data, see an example here: Simonic, K. M., Holzinger, A., Bloice, M. & Hermann, J. Optimizing Long-Term Treatment of Rheumatoid Arthritis with Systematic Documentation. Proceedings of Pervasive Health - 5th International Conference on Pervasive Computing Technologies for Healthcare, 2011 Dublin. IEEE, 550-554.
http://www.biomedcentral.com/1472-6947/13/103



Do not confuse structure with standardization (see Slide 2-9). Data can be standardized (e.g. numerical entries in laboratory reports) and non-standardized. A typical example is non-standardized text – imprecisely called "Free-Text" or "unstructured data" in an electronic patient record (Kreuzthaler et al., 2011).

Standardized data is *the* basis for accurate communication. In the medical domain, many different people work at different times in various locations. **Data standards** can ensure that information is interpreted by all users with the same understanding. Moreover, standardized data facilitate comparability of data and interoperability of systems. It supports the reusability of the data, improves the efficiency of healthcare services and avoids errors by reducing duplicated efforts in data entry.

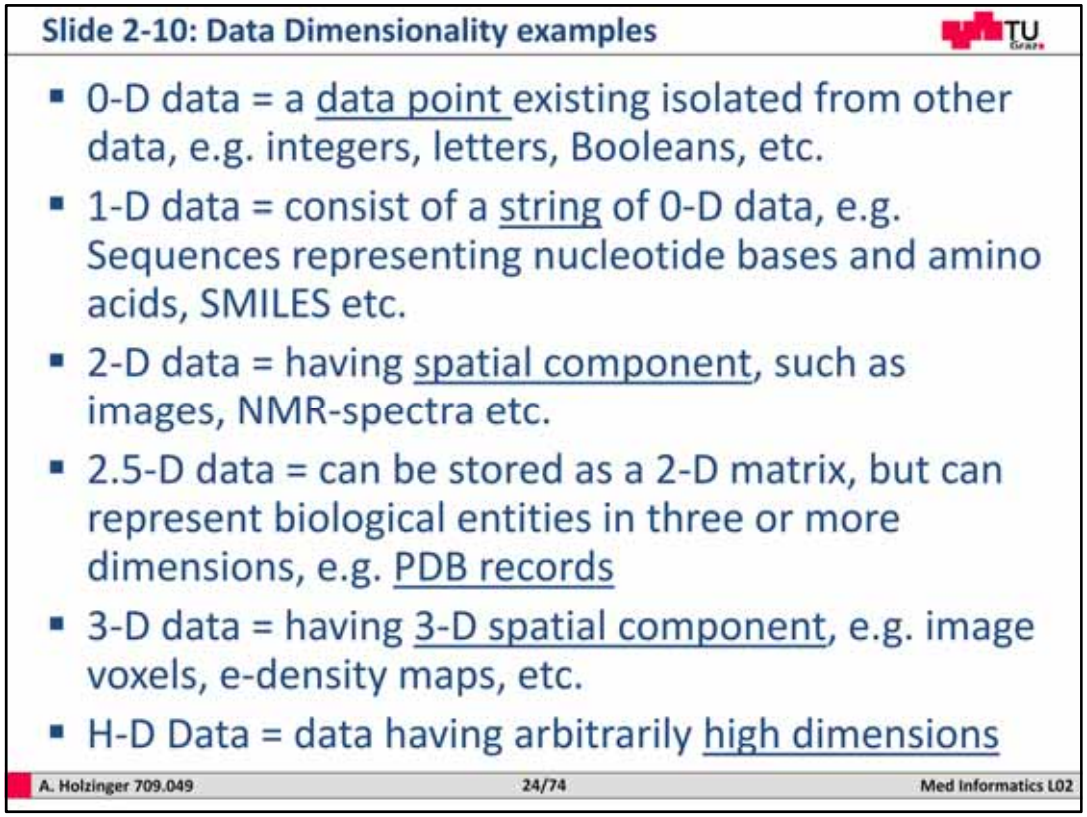
Data standardization refers to
a) the data content;
b) the terminologies that are used to represent the data;
c) how data is exchanged; and
iv) how knowledge, e.g. clinical guidelines, protocols, decision support rules, checklists, standard operating procedures are represented in the health information system (refer to IOM).
Technical elements for data sharing require standardization of identification, record structure, terminology, messaging, privacy etc. The most used standardized data set to date is the international Classification of Diseases (ICD), which was first adopted in 1900 for collecting statistics (Ahmadian et al., 2011), which we will discuss in →Lecture 3.
Non-standardized data is the majority of data and inhibit data quality, data exchange and interoperability.
Well-structured data is the minority of data and an idealistic case when each data element has an associated defined structure, relational tables, or the resource description framework RDF, or the Web Ontology Language OWL (see →Lecture 3).
Note: **Ill-structured** is a term often used for the opposite of well-structured, although this term originally was used in the context of problem solving (Simon, 1973).
Semi-structured is a form of structured data that does not conform with the strict formal structure of tables and data models associated with relational databases but contains tags or markers to separate structure and content, i.e. are schema-less or self-describing; a typical example is a markup-language such as XML (see →Lecture 3 and 4).
Weakly-Structured data is the most of our data in the whole universe, whether it is in macroscopic (astronomy) or microscopic structures (biology) – see →Lecture 5.
Non-structured data or *unstructured data* is an imprecise definition used for *information* expressed in natural language, when no specific structure has been defined. This is an issue for debate: Text has also some structure: words, sentences, paragraphs. If we are very precise, unstructured data would meant that the data is complete randomized – which is usually called noise and is defined by (Duda, Hart & Stork, 2000) as any property of data which is not due to the underlying model but instead to randomness (either in the real world, from the sensors or the measurement procedure).



"Multivariate" and "multidimensional" are modern words and consequently overused in literature. Each item of data is composed of **variables,** and if such a data item is defined by more than one variable it is called a **multivariable data item.**
Variables are frequently classified into two categories: dependent or independent.

Some more readings on the homepage of Yosuhua Bengio, University of Montreal:
http://www.iro.umontreal.ca/~bengioy/yoshua_en/research.html

And the MILA Lab – Montreal Institute for Learning Algorithms
http://www.mila.umontreal.ca/



In Physics, Engineering and Statistics a variable is a physical property of a subject, whose quantity can be measured, e.g. mass, length, time, temperature, etc.
In mathematics a 0-dimensional space (nil-dimensional) is a topological space that has dimension zero – which is an infinitesimal small point.
a 1-dimensional space is a line in R1
a 2-dimensional space is the plane in R2
A 3-dimensional space is a sphere (or cube, cylinder etc.) in R3

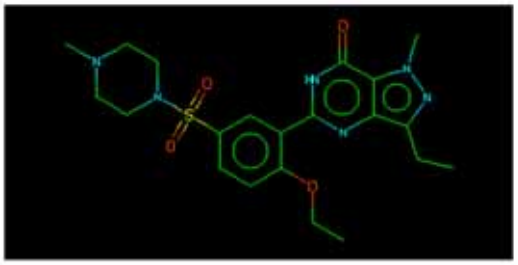
Example: 1-D data (univariate sequential data objects)

SMILES (Simplified Molecular Input Line Entry Specification)

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

e.g. Viagra:

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



...is Canonicalizable

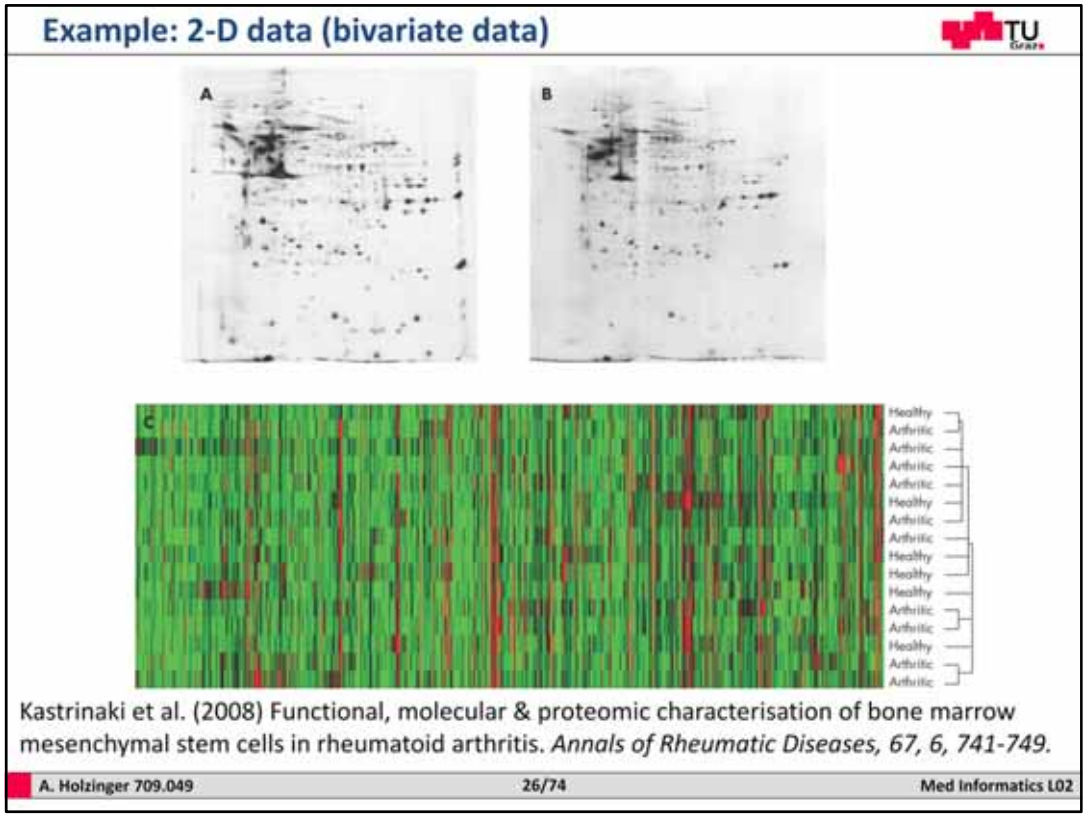
...is Comprehensive

...is Well Documented

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

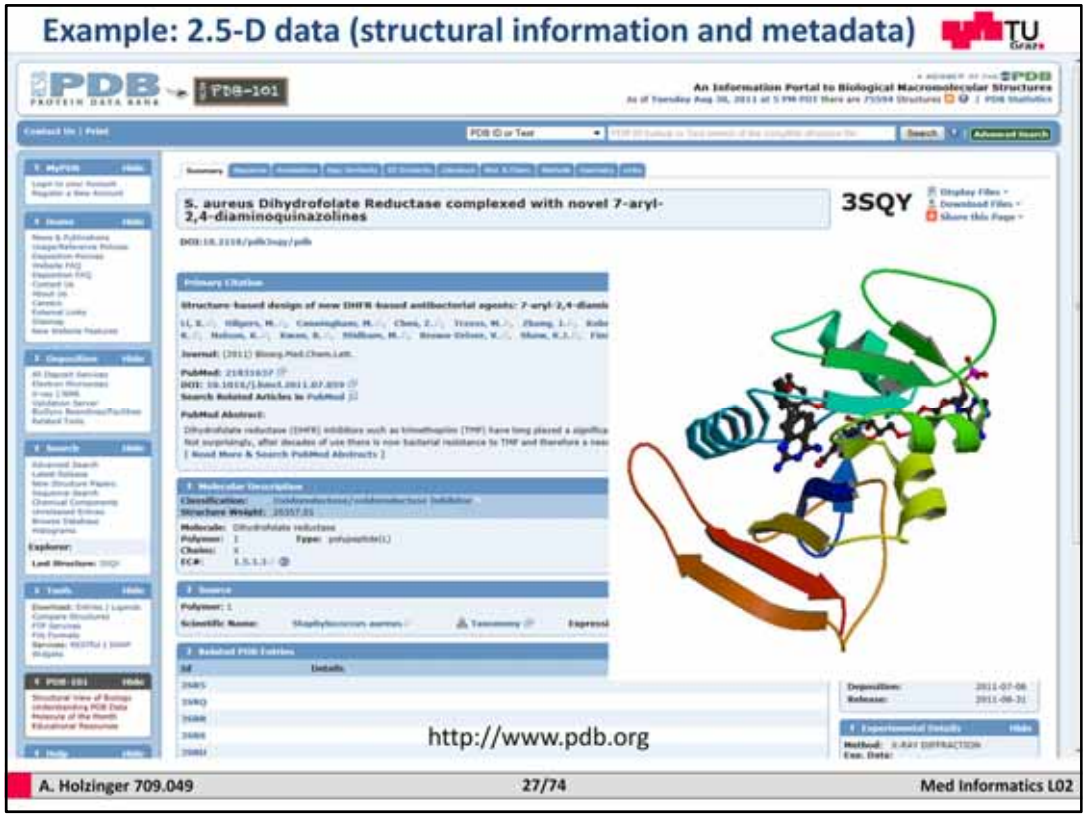
A. Holzinger 709.049 25/74 Med Informatics L02

SMILES data (.smi) consists of a string obtained by the symbol nodes encountered in a depth-first tree traversal of a chemical graph, which is first trimmed to remove hydrogen atoms and cycles are broken to turn it into a spanning tree. Where cycles have been broken, numeric suffix labels are included to indicate the connected nodes.

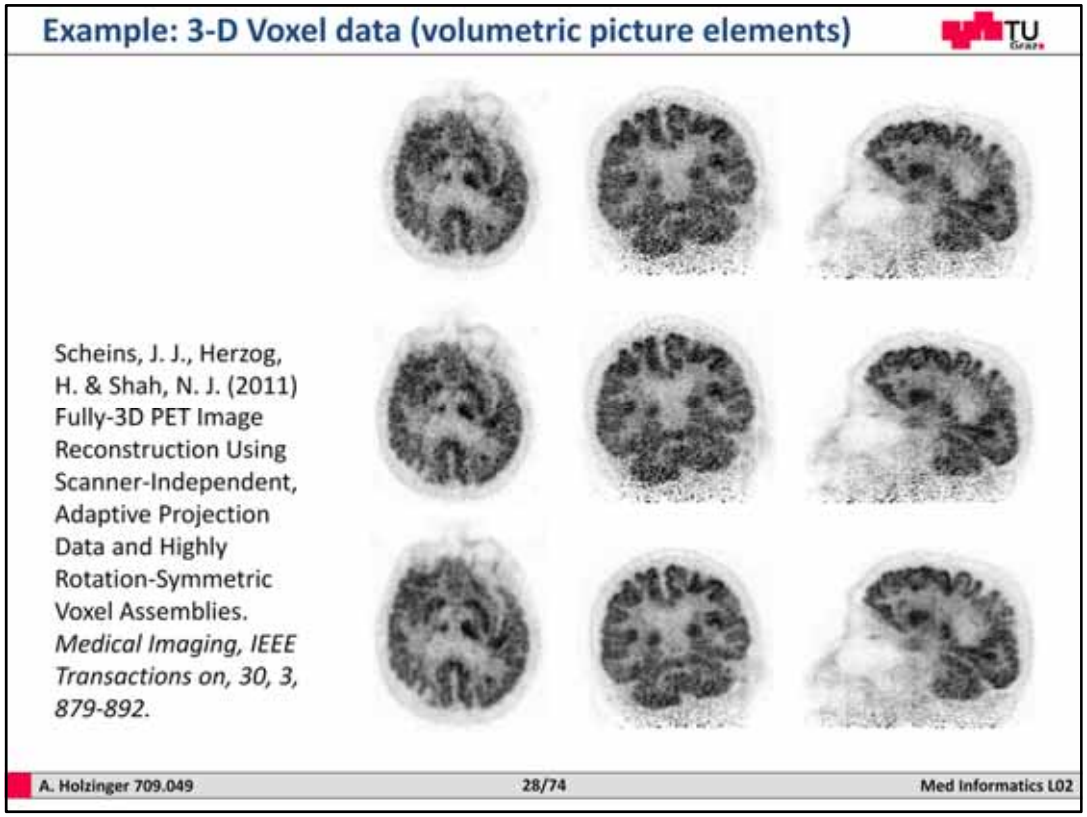


Proteomic analysis of mesenchymal stem cells (MSCs). Two-dimensional gel electrophoresis was performed using whole protein cell extracts from P2 MSC cultures of patients with rheumatoid arthritis (RA) (n = 10) (A) and healthy controls (n = 6) (B). After scanning, spot detection, quantification and normalisation, gels were compared using Hierarchical Clustering Software and Pearson test (C). No cluster could be detected using these proteomic profiles.

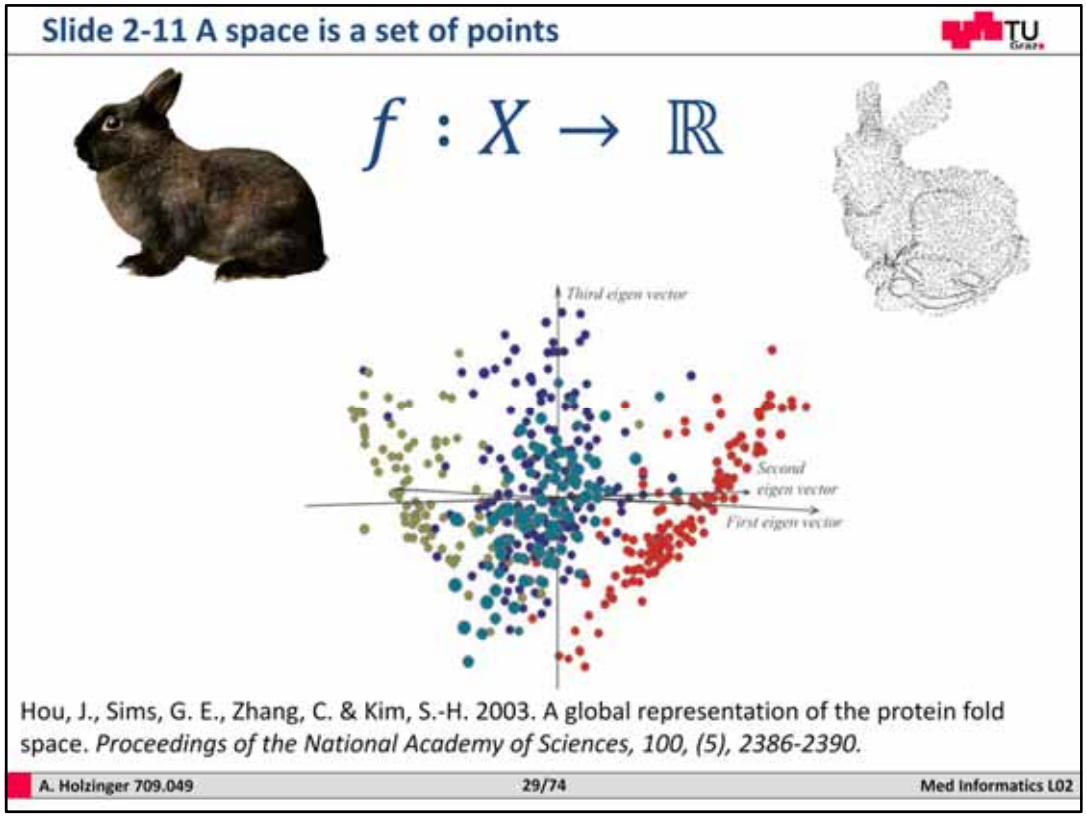
Proteomic analysis: Two-dimensional electrophoresis was performed using P2 MSCs in patients with RA (n = 10) and healthy controls (n = 6) (fig 4A,B). By using the Hierarchical Clustering method, we could not define any cluster that might discriminate patient and control cells (fig 4C). The Pearson correlation coefficient was not significantly different between patient and control cells (r = 0.933 (0.022) and r = 0.929 (0.020), respectively). These data corroborate the lack of significant changes in cytokine production between patients and controls.



http://www.rcsb.org/pdb/images/3ond_bio_r_500.jpg
The PDB is a large repository containing 3-D structural information, established in 1971
Data a stored in 2D but can in fact represent biological entities in three or more dimensions



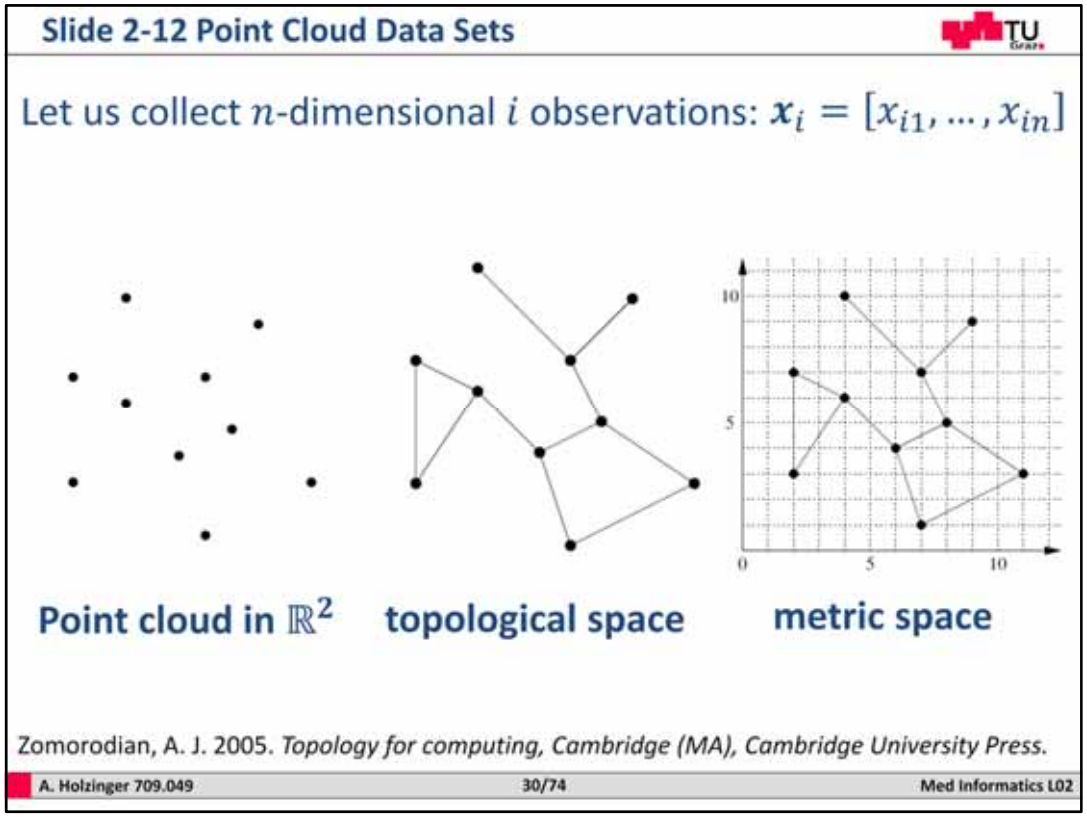
Transaxial (left), coronal (middle), and sagittal (right) images of a patient who was scanned for 30 min in list-mode with the BrainPET scanner; the recording was started 20 min after injection of about 300 MBq fluor-deoxy-glucose.



In Mathematics, hence in Informatics, however, a variable is associated with a *space* – often an n-dimensional Euclidean space $\mathbb{R}_n$ – in which an entity (e.g. a function) or a phenomenon of continuous nature is defined. The data location within this space can be referenced by using a range of coordinate systems (e.g. Cartesian, Polar-coordinates, etc.): The dependent variables are those used to describe the entity (for example the function value) whilst the independent variables are those that represent the coordinate system used to describe the space in which the entity is defined. If a dataset is composed of variables whose interpretation fits this definition our goal is to understand how the 'entity' is defined within the n-dimensional Euclidean space $\mathbb{R}_n$. Sometimes we may distinguish between variables meaning measurement of property, from variables meaning a coordinate system, by referring to the former as **variate**, and referring to the latter as **dimension** (Dos Santos & Brodlie, 2002), (dos Santos & Brodlie, 2004).

A space is a set of points. A metric space has an associated metric, which enables us to measure distances between points in that space and, in turn, implicitly define their neighborhoods. Consequently, a metric provides a space with a topology, and a metric space is a topological one. Topological spaces feel alien to us because we are accustomed to having a metric.

Biomedical Example: A protein is a single chain of amino acids, which folds into a globular structure. The Thermodynamics Hypothesis states that a protein always folds into a state of minimum energy. To predict protein structure, we would like to model the folding of a protein computationally. As such, the protein folding problem becomes an optimization problem: We are looking for a path to the global minimum in a very high-dimensional energy landscape;



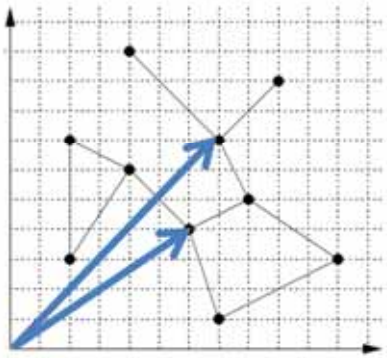
Let us collect n-dimensional i observations in the Euclidean vector space $\mathbb{R}^n$ and we get:
Eq. 2-1

$$\boldsymbol{x}_i = [x_{i1}, \dots, x_{in}]$$

A cloud of points sampled from any source (e.g. medical data, sensor network data, a solid 3-D object, surface etc.). Those data points can be coordinated as an unordered sequence in an arbitrarily high dimensional Euclidean space, where methods of algebraic topology can be applied. The main challenge is in mapping the data back into $\mathbb{R}^3$ or to be more precise into $\mathbb{R}^2$, because our retina is inherently perceiving data in $\mathbb{R}^2$. The cloud of such data points can be used as a computational representation of the respective data object. A temporal version can be found in motion-capture data, where geometric points are recorded as time series. Now you will ask an obvious question: "How do we visualize a four-dimensional object?" The obvious answer is: "How do we visualize a three dimensional object?" Humans do not see in three spatial dimensions directly, but via sequences of planar projections integrated in a manner that is sensed if not comprehended. Little children spend a significant time of their first year of life learning how to infer three-dimensional spatial data from paired planar projections, and many years of practice have tuned a remarkable ability to extract global structure from representations in a strictly lower dimension (Ghrist, 2008). Because we have the same problem here in this book, we must stay in $\mathbb{R}^2$ and therefore the example in Slide 2-12 (Zomorodian, 2005).
In Einstein's theory of Special Relativity, Euclidean 3-space plus time (the "4th-dimension") are unified into the Minkowski space

Slide 2-13: Example Metric Space

A set S with a metric function d is a metric space



$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

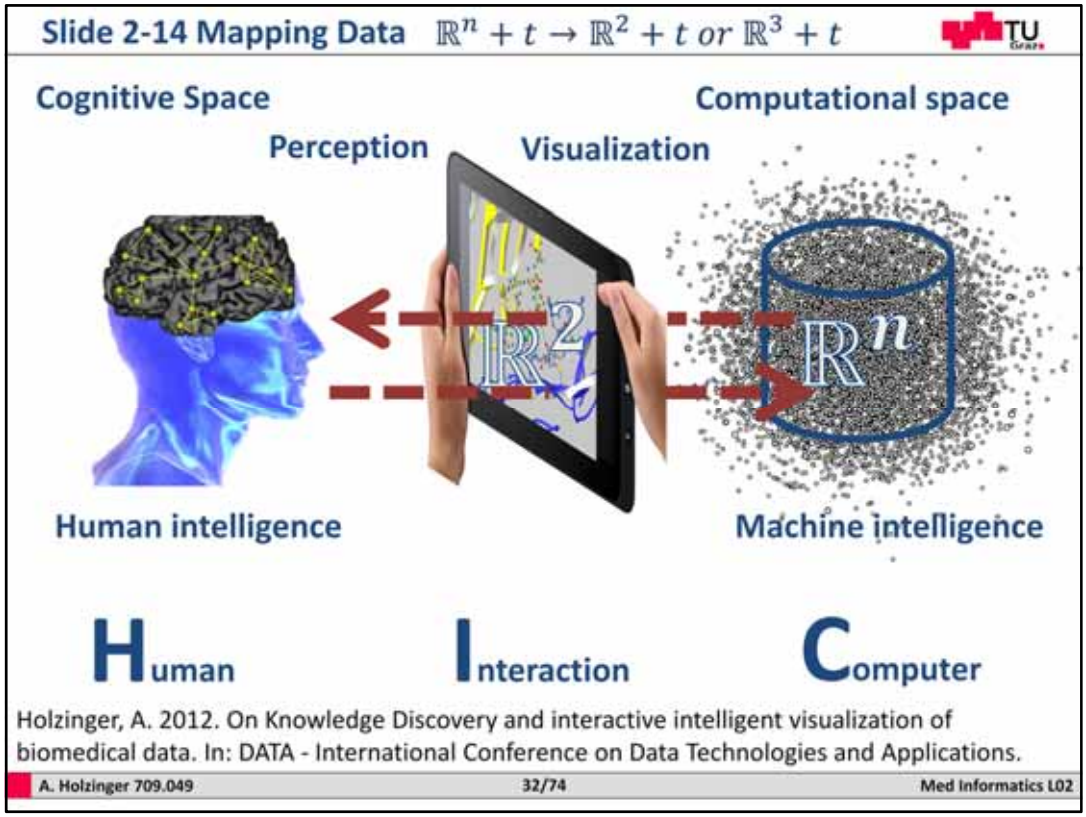
Doob, J. L. 1994. *Measure theory, Springer New York.*

A. Holzinger 709.049 31/74 Med Informatics L02

A metric space has an associated metric, which enables to measure the distances between points in that space and, implicitly define their neighborhoods. Consequently, a metric provides a space with a topology, hence a metric space is a topological space.
A set X with a metric function d is called a metric space. We give it the metric topology of d, where the set of open balls
Most of our "natural" spaces are a particular type of metric spaces: the Euclidean spaces: The Cartesian product of n copies of $\mathbb{R}$, the set of real numbers, along with the Euclidean metric:
Eq. 2-2

$$d(i,j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

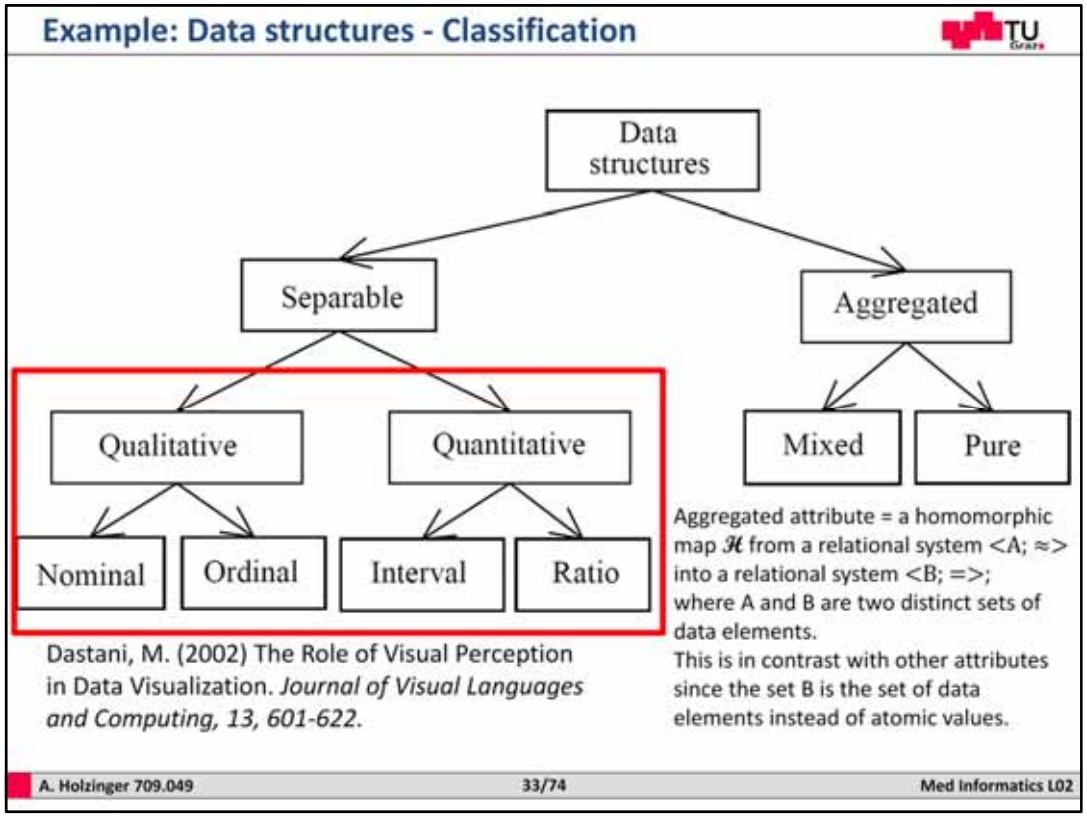
is the n-dimensional Euclidean space $\mathbb{R}^n$.
We may induce a topology on subsets of metric spaces as follows:
If $A \subseteq X$ with topology T, then we get the relative or induced topology T_A by defining T_A
For more information refer to (Zomorodian, 2005) or (Edelsbrunner & Harer, 2010).



Knowledge Discovery from Data: By getting insight into the data; the gained information can be used to build up knowledge. The grand challenge is to map higher dimensional data into lower dimensions, hence make it interactively accessible to the end-user (Holzinger, 2012), (Holzinger, 2013).
This mapping from $\mathbb{R}^n \rightarrow \mathbb{R}^2$ is the core task of visualization and a major component for knowledge discovery: Enabling effective interactive human control over powerful machine algorithms to support human sensemaking (Holzinger, 2012), (Holzinger, 2013).

Holzinger, A. 2013. Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? *In: Alfredo Cuzzocrea, C. K., Dimitris E. Simos, Edgar Weippl, Lida Xu (ed.) Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin, New York: Springer, pp. 319-328.*

An important topic is subspace clustering: 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: Guo, Y., Friston, K., Aldo, F., Hill, S. & Peng, H. (eds.) Brain Informatics and Health, Lecture Notes in Artificial Intelligence LNAI 9250. Cham: Springer International Publishing, pp. 358-368. https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=1198810&pCurrPk=85960



Multivariate dataset is a dataset that has *many dependent variables* and they might be correlated to each other to varying degrees. Usually this type of dataset is associated with *discrete data models.*

Multidimensional dataset is a dataset that has *many independent variables* clearly identified, and one or more dependent variables associated to them. Usually this type of dataset is associated with *continuous data models.*

In other words, every data item (or object) in a computer is represented (stored) as a set of features. Instead of the term features we may use the term dimensions, because an object with n-features can also be represented as a multidimensional point in an n-dimensional space. Dimensionality reduction is the process of mapping an n-dimensional point, into a lower k-dimensional space – this is the main challenge in visualization see →Lecture 9.

The number of dimensions can sometimes be small, e.g. simple 1D-data such as temperature measured at different times, to 3D applications such as medical imaging, where data is captured within a volume. Standard techniques—contouring in 2D; isosurfacing and volume rendering in 3D—have emerged over the years to handle this sort of data. There is no dimension reduction issue in these applications, since the data and display dimensions essentially match.

Slide 2-15: Categorization of Data (Classic "scales")

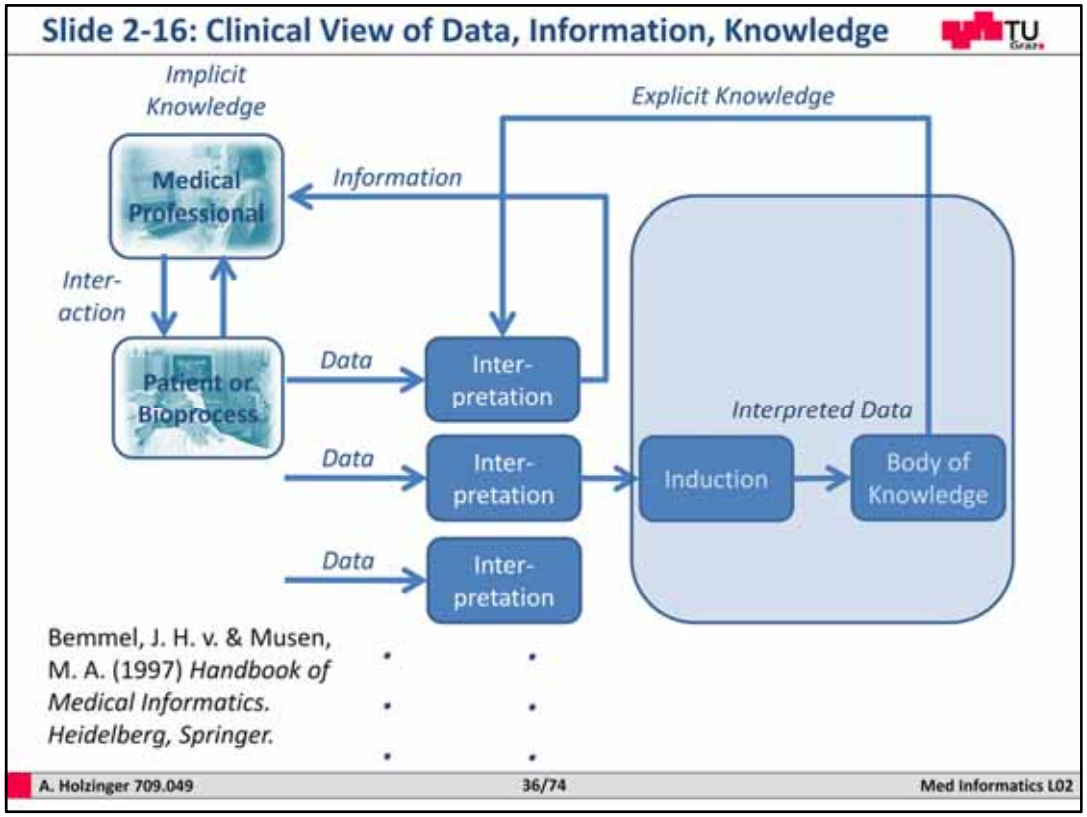
Scale	Empirical Operation	Mathem. Group Structure	Transf. in $\mathbb{R}$	Basic Statistics	Mathematical Operations
NOMINAL	Determination of equality	Permutation $x' = f(x)$ x ... 1-to-1	$x \mapsto f(x)$	Mode, contingency correlation	$=, \neq$
ORDINAL	Determination of more/less	Isotonic $x' = f(x)$ x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

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

A. Holzinger 709.049 34/74 Med Informatics L02

Data can be categorized into qualitative (nominal and ordinal) and quantitative (interval and ratio): Interval and ratio data are parametric, and are used with parametric tools in which distributions are predictable (and often Normal).
Nominal and ordinal data are non-parametric, and do not assume any particular distribution. They are used with non-parametric tools such as the Histogram.
The classic paper on the theory of scales of measurement is (Stevens, 1946).

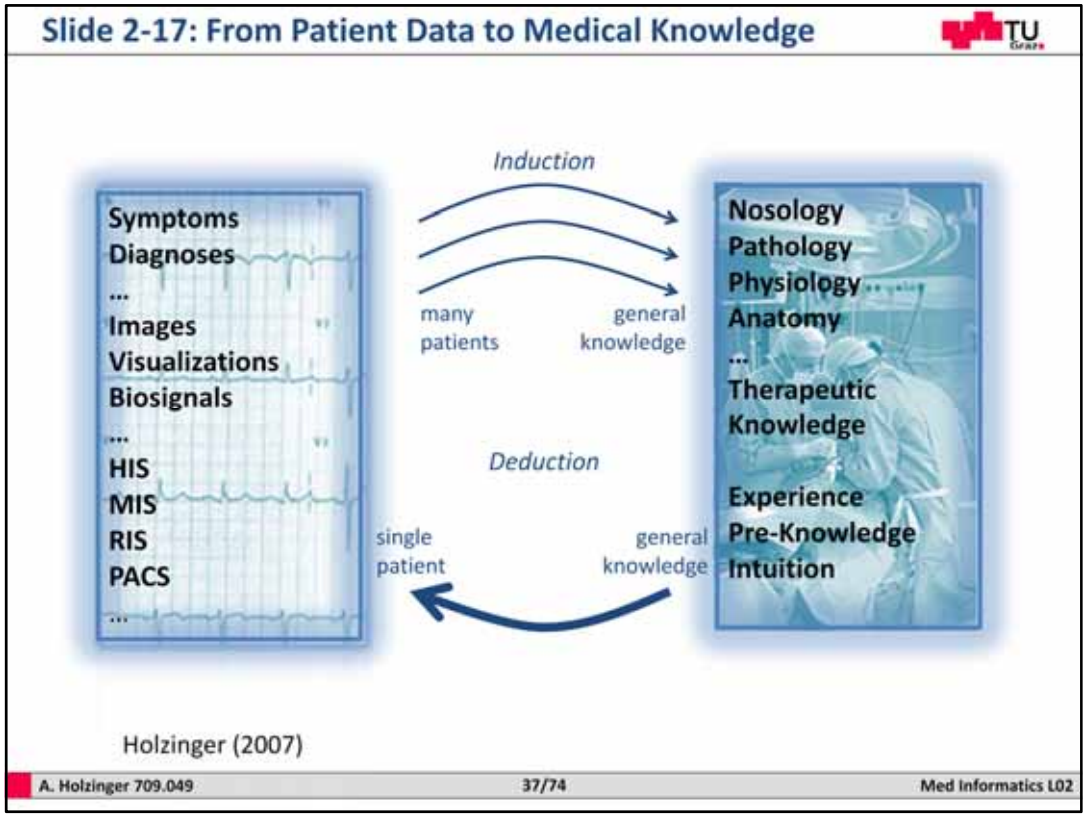
We can summarize what we learned so far about data: Data can be numeric, non-numeric, or both. Non-numeric data can include anything from language data (text) to categorical, image, or video data. Data may range from completely structured, such as categorical data, to semi-structured, such as an XML File containing meta information, to unstructured, such as a narrative "free-text". Note, that term unstructured does not mean that the data are without any pattern, which would mean complete randomness and uncertainty, but rather that "unstructured data" are expressed so, that only humans can meaningfully interpret it. Structure provides information that can be interpreted to determine data organization and meaning, hence it provides a **context** for the information. The inherent structure in the data can form a basis for data representation. An important, yet often neglected issue are **temporal characteristics of data:** Data of all types may have a temporal (time) association, and this association may be either discrete or continuous (Thomas & Cook, 2005).
In Medical Informatics we have a permanent interaction between data, information and knowledge, with different definitions (Bemmel & Musen, 1997), see Slide 2-16:



Data are the physical entities at the lowest abstraction level which are, e.g. generated by a patient (patient data) or a biological process (e.g. Omics data). According to (Bemmel & Musen, 1997) data contain no meaning.

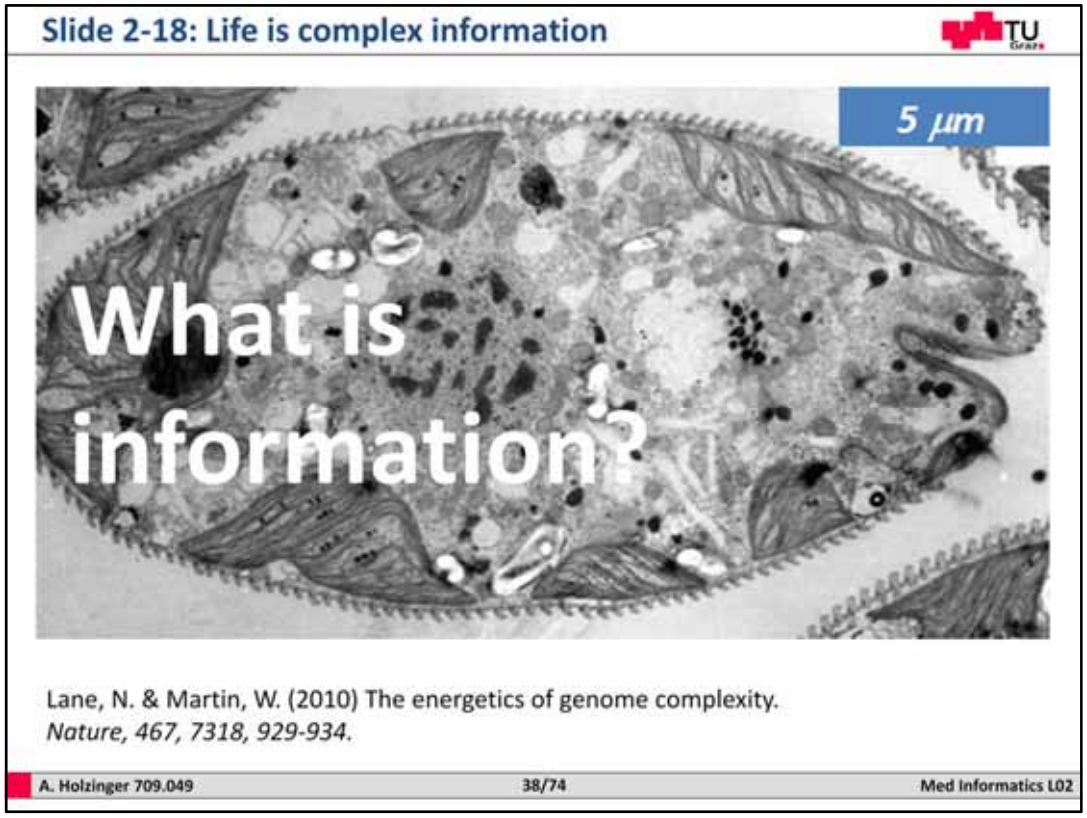
Information is derived by interpretation of the data by a clinician (human intelligence).

Knowledge is obtained by inductive reasoning with previously interpreted data, collected from many similar patients or processes, which is added to the so called body of knowledge in medicine, the **explicit knowledge.** This knowledge is used for the interpretation of other data and to gain **implicit knowledge** which guides the clinician in taking further action.



For hypothesis generation and testing, four types of inferences exist (Peirce, 1955): abstraction, abduction, deduction, and induction. The first two drive hypothesis generation while the latter drive hypothesis testing, see Slide 2-17:
Abstraction means that data are filtered according to their relevance for the problem solution and chunked in schemas representing an *abstract* description of the problem (e.g., abstracting that an adult male with haemoglobin concentration less than 14g/dL is an anaemic patient). Following this, hypotheses that could account for the current situation are related through a process of abduction, characterized by a "backward flow" of inferences across a chain of directed relations which identify those initial conditions from which the current abstract representation of the problem originates. This provides tentative solutions to the problem at hand by way of hypotheses. For example, knowing that disease A will cause symptom B, abduction will try to identify the explanation for B, while deduction will forecast that a patient affected by disease A will manifest symptom B: both inferences are using the same relation along two different directions (Patel & Ramoni, 1997).
Abduction is characterized by a cyclical process of generating possible explanations (i.e., identification of a set of hypotheses that are able to account for the clinical case on the basis of the available data) and testing those explanations (i.e., evaluation of each generated hypothesis on the basis of its expected consequences) for the abnormal state of the patient at hand (Patel, Arocha & Zhang, 2004).

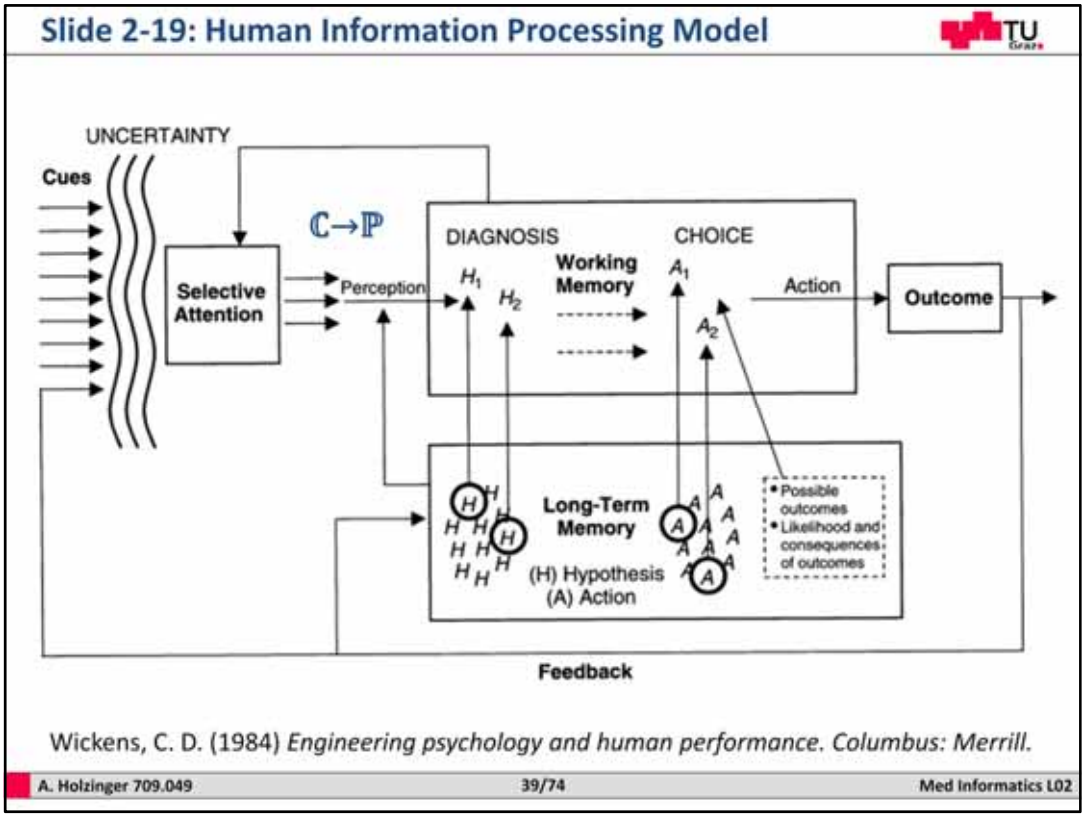
The hypothesis testing procedures can be inferred from Slide 2-17:
General knowledge is gained from many patients, and this general knowledge is then applied to an individual patient. We have to determine between:
Reasoning is the process by which a clinician reaches a conclusion after thinking about all the facts;
Deduction consists of deriving a particular valid conclusion from a set of *general* premises;
Induction consists of deriving a likely general conclusion from a set of particular statements.
Reasoning in the "real world" does not appear to fit neatly into any of these basic types. Therefore, a third form of reasoning has been recognized by Peirce (1955), where deduction and induction are inter-mixed;



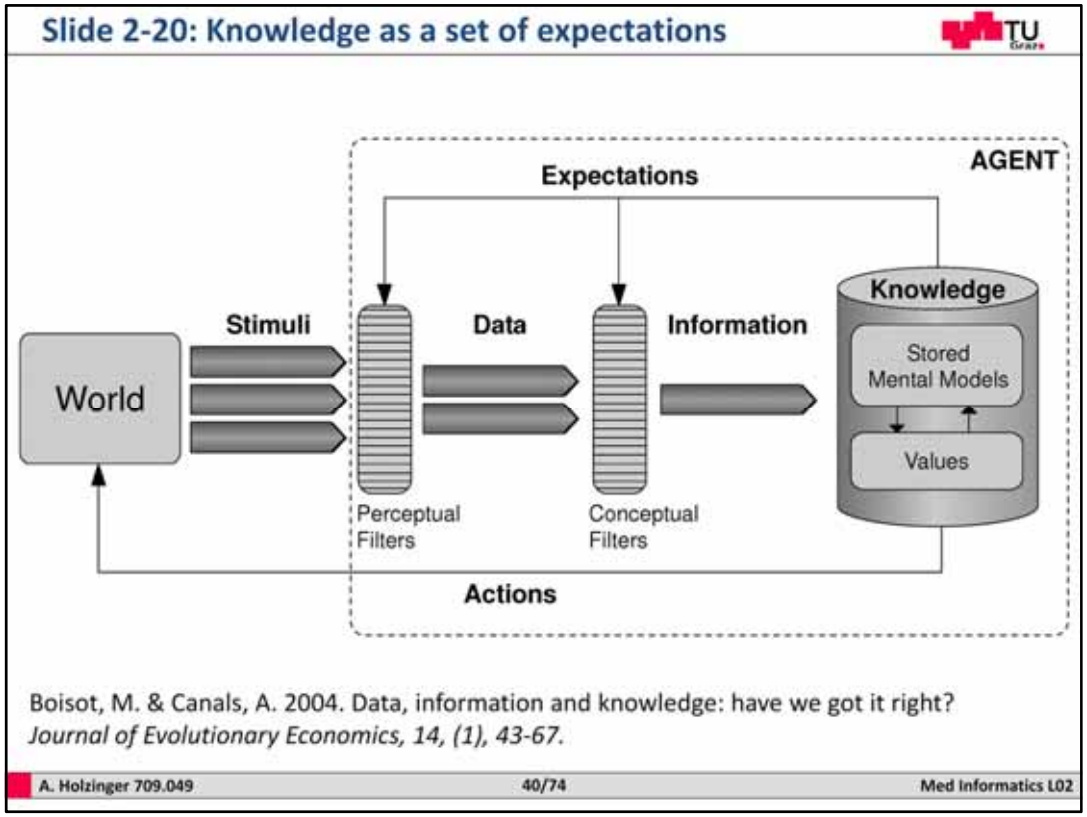
The question "what is information?" is still an open question in basic research, and any definition is depending on the view taken. For example, the definition given by Carl-Friedrich von Weizsäcker: "Information is what is understood," implies that information has both a sender and a receiver who have a common understanding of the representation and the means to convey information using some properties of the physical systems, and his addendum: "Information has no absolute meaning; it exists relatively between two semantic levels" implies the importance of context (Marinescu, 2011). Without doubt information is a fundamentally important concept within our world and life is complex information, see Slide 2-14:

Many systems, e.g. in the quantum world to not obey the classical view of information. In the quantum world and in the life sciences traditional information theory often fails to accurately describe reality … for example in the complexity of a living cell: All complex life is composed of eukaryotic (nucleated) cells (Lane & Martin, 2010). A good example of such a cell is the protist Euglena Gracilis (in German "Augentierchen") with a length of approx. 30 μm. Life can be seen as a delicate interplay of energy, entropy and information, essential functions of living beings correspond to the generation, consumption, processing, preservation and duplication of information.

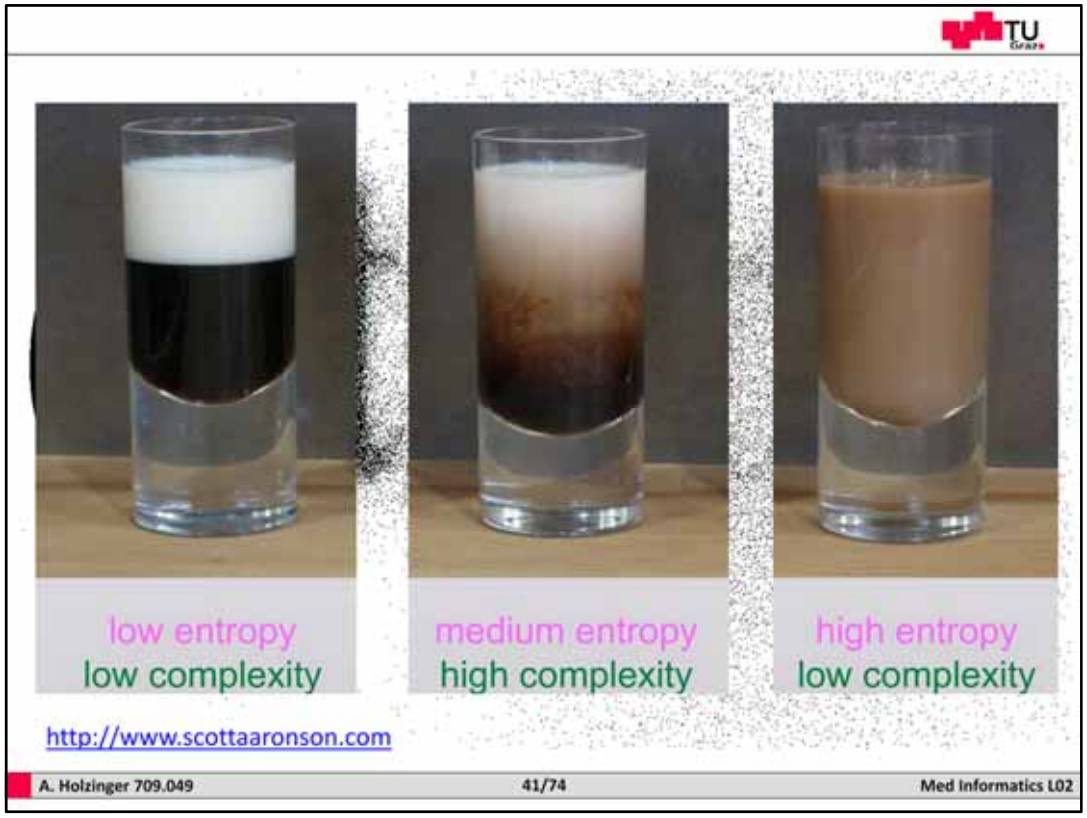
P: Complexity <> Information <> Energy <> Entropy



The etymological origin of the word information can be traced back to the Greek "forma" and the Latin "information" and "informare", to bring something into a shape ("in-a-form"). Consequently, the naive definition in computer science is *"information is data in context"* and therefore different than data or knowledge.
However, we follow the notion of (Boisot & Canals, 2004) and define that information is an extraction from data that, by modifying the relevant probability distributions, has direct influence on an agent's knowledge base. For a better understanding of this concept, we first review the model of human information processing by Wickens (1984): The model by Wickens (1984) beautifully emphasizes our view on data, information and knowledge: the physical data from the real-world are perceived as information through perceptual filters, controlled by selective attention and form hypotheses within the working memory. These hypotheses are the expectations depending on our previous knowledge available in our mental model, stored in the long-term memory. The subjectively best alternative hypothesis will be selected and processed further and may be taken as outcome for an action. Due to the fact that this system is a closed loop, we get feedback through new data perceived as new information and the process goes on.



The incoming stimuli from the physical world must pass both a perceptual filter and a conceptual filter.
The **perceptual filter** orientates the senses (e.g. visual sense) to certain types of stimuli within a certain physical range (e.g. visual signal range, pre-knowledge, attention etc.). Only the stimuli which pass through this filter get registered as *incoming data* – everything else is filtered out. At this point it is important to follow our physical principle of data: to differentiate between two notions that are frequently confused: an experiment's (raw, hard, measured, factual) data and its (meaningful, subjective) interpreted information results. Data are properties concerning only the instrument; it is the **expression of a fact.** The result concerns a property of the world. The following conceptual filters extract information-bearing data from what has been previously registered.
Both types of filters are influenced by the agents' cognitive and affective expectations, stored in their mental models. The enormous utility of data resides in the fact that it can carry information about the physical world. This information may modify set expectations or the state-of-knowledge. These principles allow an **agent** to act in adaptive ways in the physical world (Boisot & Canals, 2004).
Confer this process with the human information processing model by (Wickens, 1984), seen in Slide 2-19 and discussed in →Lecture 7.



Entropy has many different definitions and applications, originally in statistical physics and most often it is used as a **measure for disorder.**
In information theory, **entropy can be used as a measure for the uncertainty in a data set.**

To demonstrate how useful entropy can be - you can have a look at this paper:
Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350.
http://www.mdpi.com/1099-4300/14/11/2324

Σλιδε 2–21: εντροπια

My greatest concern was what to call it. I thought of calling it "information", but the word was overly used, so I decided to call it "uncertainty". When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, "You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, nobody knows what entropy really is, so in a debate you will always have the advantage."

Tribus, M. & McIrvine, E. C. (1971) Energy and Information. *Scientific American, 225, 3, 179-184.*

A. Holzinger 709.049 42/74 Med Informatics L02

The concept of entropy was first introduced in thermodynamics (Clausius, 1850), where it was used to provide a statement of the second law of thermodynamics. Later, statistical mechanics provided a connection between the macroscopic property of entropy and the microscopic state of a system by Boltzmann. Shannon was the first to define entropy and mutual information.

Shannon (1948) used a Gedankenexperiment (thought experiment) to propose a measure of uncertainty in a discrete distribution based on the Boltzmann entropy of classical statistical mechanics, see Slide 2-22:

Slide 2-22: Entropy H as a measure for uncertainty (1/3)

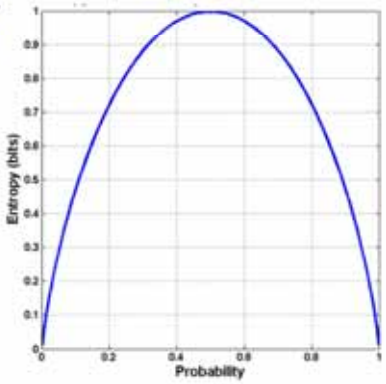
$$Q \ldots P = \{p_1, \ldots, p_n\} \qquad H(Q) = -\sum_{i=1}^{n} (p_i * \log p_i)$$

$$Qb = \{a_1, a_2\} \text{ with } P = \{p, 1-p\}$$

$$H(Qb) = p * \log\frac{1}{p} + p * \log\frac{1}{1-p}$$

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

Shannon, C. E. & Weaver, W. (1949) *The Mathematical Theory of Communication. Urbana (IL), University of Illinois Press.*



A. Holzinger 709.049 43/74 Med Informatics L02

An example shall demonstrate the usefulness of this approach:
1) Let Q be a discrete data set with associated probabilities p_i:
Eq. 2-5

$$Q \ \ldots \ P = \{p_1, \ldots, p_n\}$$

2) Now we apply Shannon's equation Eq. 2-4:
Eq. 2-6

$$H(Q) = -\sum_{i=1}^{n} p_i log_2(p_i)$$

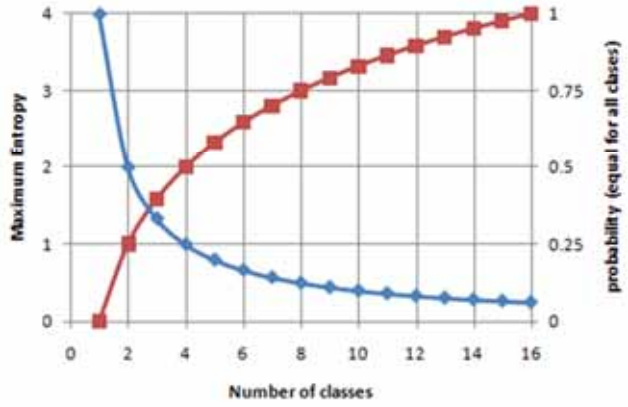
3) We assume that our source has two values (ball = white, ball = black)
Let us do the famous simple Gedankenexperiment (thought experiment): Imagine a box which can contain two colored balls: black and white. This is our set of discrete symbols with associated probabilities. If we grab blindly into this box to get a ball, we are dealing with uncertainty, because we do not know which ball we touch. We can ask: Is the ball black? NO. THEN it must be white, so we need one question to surely provide the right answer. Because it is a binary decision (YES/NO) the maximum number of (binary) questions required to reduce the uncertainty is: $\log_2(N)$, where N is the number of the possible outcomes. If there are N events with equal probability p then $N = 1/p$. If you have only 1 black ball, then $\log_2(1) = 0$, which means there is no uncertainty.
Eq. 2-7
$Qb = \{a_1, a_2\}$ with $P = \{p, 1-p\}$
4) Now we solve numerically Eq. 2-6:
Eq. 2-8

$$H(Q_b) = p * \log\frac{1}{p} + p * \log\frac{1}{1-p}$$

Since p ranges from 0 (for impossible events) to 1 (for certain events), the entropy value ranges from infinity (for impossible events) to 0 (for certain events). So, we can summarize that the entropy is the weighted average of the surprise for all possible outcomes. For our example with the two balls we can draw the following function:
The entropy value is 1 for p=0,5 and it is both 0 for either p=0 or p=1. This example might seem trivial, but the entropy principle has been developed a lot since Shannon and there are many different methods, which are very useful for dealing with data.

Slide 2-23: A measure for uncertainty (2/3)

$$\log_2 \frac{1}{p} = -\log_2 p \qquad H = -\sum_{i=1}^{N} p_i log_2(p_i)$$



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

A. Holzinger 709.049 44/74 Med Informatics L02

Shannon called it the **information entropy** (aka Shannon entropy) and defined:
Eq. 2-9

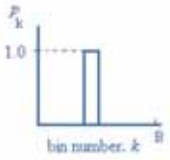
$$\log_2 \frac{1}{p} = -\log_2 p$$

where p is the probability of the event occurring. If p is not identical for all events then the entropy H is a weighted average of all probabilities, which Shannon defined as:
Eq. 2-10

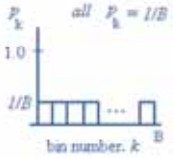
$$H = -\sum_{i=1}^{N} p_i log2\ (p_i)$$

Basically, the entropy p(x) approaches zero if we have a maximum of structure – and opposite, the entropy p(x) reaches high values if there is no structure – hence, ideally, if the entropy is a maximum, we have complete randomness, total uncertainty.
Low Entropy means differences, structure, individuality – high Entropy means no differences, no structure, no individuality. Consequently, life needs low entropy.

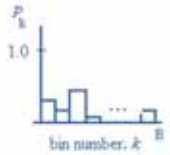
Slide 2-24: Entropy H as a measure for uncertainty (3/3)



$$H_B = -\sum_{k=1} p_k \log_2 p_k = -1 * \log_2(1) = 0$$



$$H_B = -\sum_{k=1}^{B} \frac{1}{B} \log_2 \frac{1}{B} = \log_2(B)$$



$$H = H_{min} = 0 \qquad H = H_{max} = \log_2 N$$

A. Holzinger 709.049 45/74 Med Informatics L02

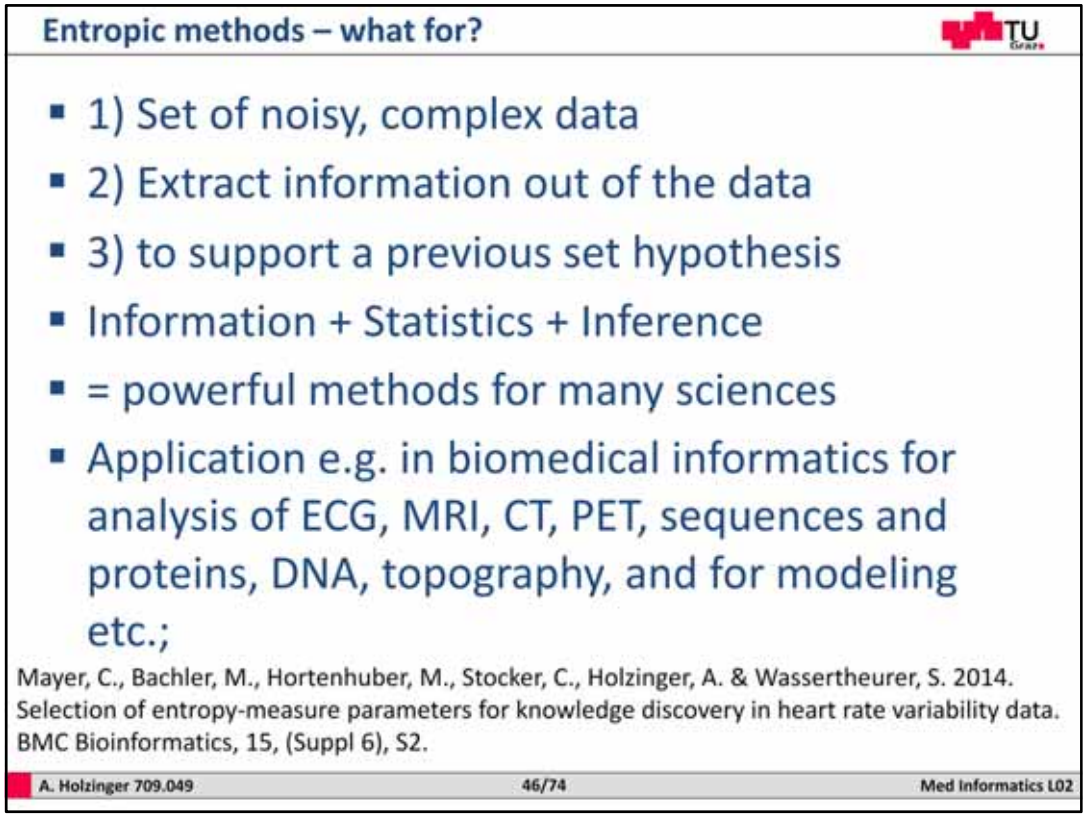
The principle what we can infer from entropy values is:
1) Low entropy values mean high probability, **high certainty,** hence a high degree of structurization in the data.
2) High entropy values mean low probability, **low certainty** (≅ high uncertainty ;-), hence a low degree of structurization in the data.
Maximum entropy would mean complete randomness and total uncertainty.
Highly structured data contain low entropy; ideally if everything is in order and there is no surprise (no uncertainty) the entropy is low:
Eq. 2-11

$$H = H_{min} = 0$$

Eq. 2-12

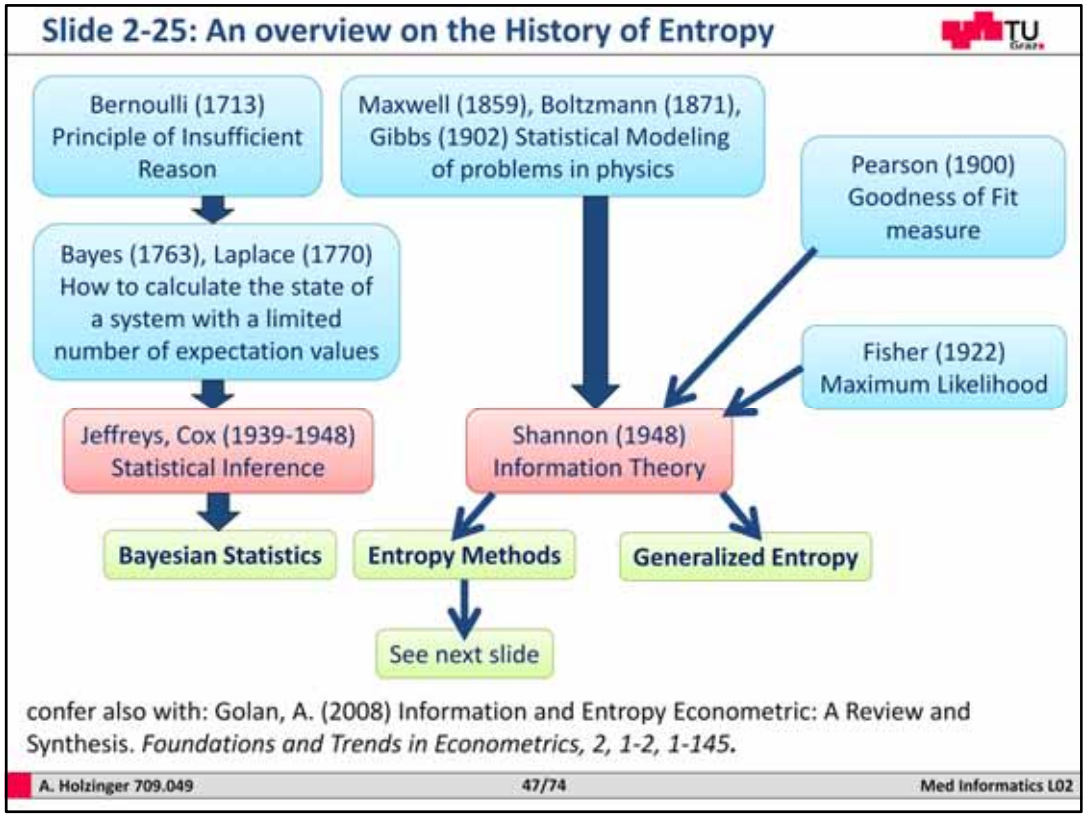
$$H = H_{max} = \log_2 N .$$

On the other hand if the data are weakly structured – as for example in biological data – and there is no ability to guess (all data is equally likely) the entropy is high:
If we follow this approach, "unstructured data" would mean complete randomness. Let us look on the history of entropy to understand what we can do in future, see Slide 2-25.

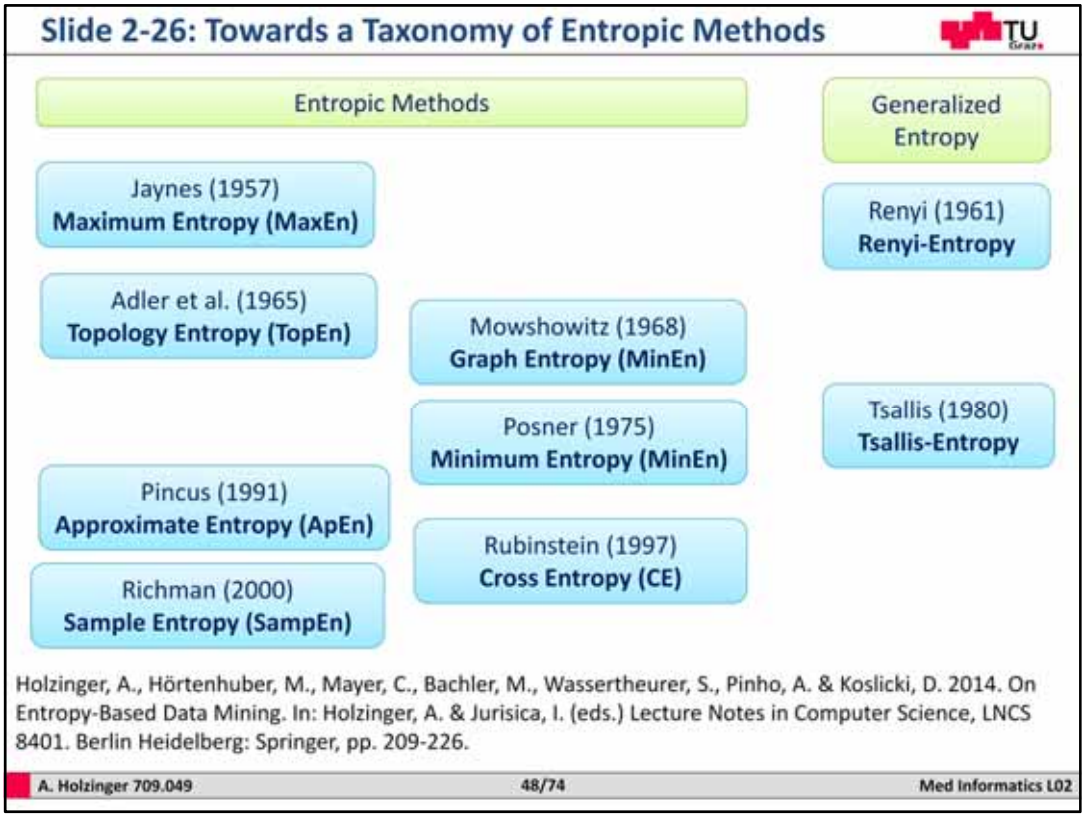


You might argue what the practical purpose of this approach is – manifold applications!

Example: Heart rate variability is the variation of the time interval between consecutive heartbeats. Entropy is a commonly used tool to describe the regularity of data sets. Entropy functions are defined using multiple parameters, the selection of which is controversial and depends on the intended purpose. Mayer et al. (2014) describe the results of tests conducted to support parameter selection, towards the goal of enabling further biomarker discovery. They dealt with approximate, sample, fuzzy, and fuzzy measure entropies. All data were obtained from PhysioNet https://www.physionet.org, a free-access, on-line archive of physiological signals, and represent various medical conditions. Five tests were defined and conducted to examine the influence of: varying the threshold value r (as multiples of the sample standard deviation ?, or the entropy-maximizing rChon), the data length N, the weighting factors n for fuzzy and fuzzy measure entropies, and the thresholds rF and rL for fuzzy measure entropy. The results were tested for normality using Lilliefors' composite goodness-of-fit test. Consequently, the p-value was calculated with either a two sample t-test or a Wilcoxon rank sum test. The first test shows a cross-over of entropy values with regard to a change of r. Thus, a clear statement that a higher entropy corresponds to a high irregularity is not possible, but is rather an indicator of differences in regularity. N should be at least 200 data points for r = 0.2 ? and should even exceed a length of 1000 for r = rChon. The results for the weighting parameters n for the fuzzy membership function show different behavior when coupled with different r values, therefore the weighting parameters have been chosen independently for the different threshold values. The tests concerning rF and rL showed that there is no optimal choice, but r = rF = rL is reasonable with r = rChon or r = 0.2?. CONCLUSIONS: Some of the tests showed a dependency of the test significance on the data at hand. Nevertheless, as the medical conditions are unknown beforehand, compromises had to be made. Optimal parameter combinations are suggested for the methods considered. Yet, due to the high number of potential parameter combinations, further investigations of entropy for heart rate variability data will be necessary.
Mayer, C., Bachler, M., Hortenhuber, M., Stocker, C., Holzinger, A. & Wassertheurer, S. 2014. Selection of entropy-measure parameters for knowledge discovery in heart rate variability data. BMC Bioinformatics, 15, (Suppl 6), S2.
http://www.ncbi.nlm.nih.gov/pubmed/25078574

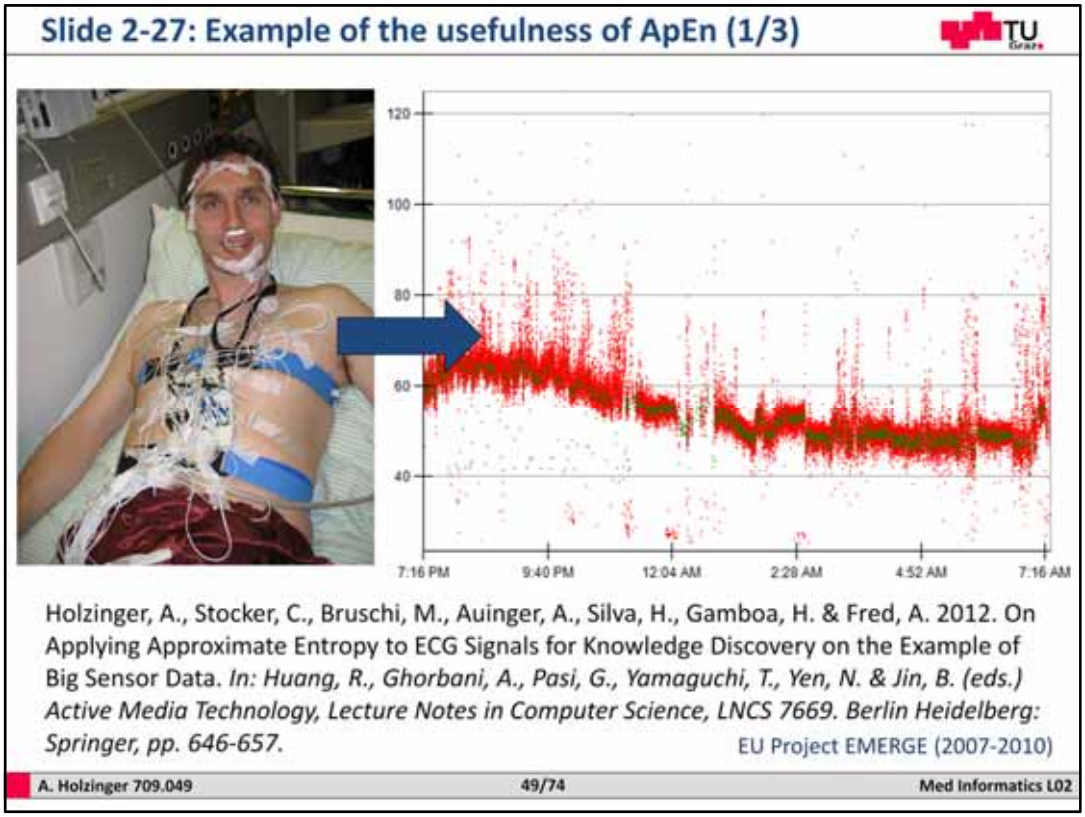


The origin may be found in the work of Jakob Bernoulli, describing the principle of insufficient reason: we are ignorant of the ways an event can occur, the event will occur equally likely in any way. Thomas Bayes (1763) and Pierre-Simon Laplace (1774) carried on and Harold Jeffreys and David Cox solidified it in the Bayesian Statistics, aka statistical inference. The second path leading to the classical Maximum Entropy, en-route with the Shannon Entropy, can be identified with the work of James Clerk Maxwell and Ludwig Boltzmann, continued by Willard Gibbs and finally Claude Elwood Shannon. This work is geared toward developing the mathematical tools for statistical modeling of problems in information. These two independent lines of research are very similar. The objective of the first line of research is to formulate a theory/methodology that allows understanding of the *general characteristics* (distribution) of a system from partial and incomplete information. In the second route of research, the same objective is expressed as determining how to assign (initial) numerical values of probabilities when only some (theoretical) limited global quantities of the investigated system are known. Recognizing the common basic objectives of these two lines of research aided Jaynes in the development of his classical work, the Maximum Entropy formalism. This formalism is based on the first line of research and the mathematics of the second line of research. The interrelationship between Information Theory, statistics and inference, and the Maximum Entropy (MaxEnt) principle became clear in 1950ies, and many different methods arose from these principles (Golan, 2008), see next Slide

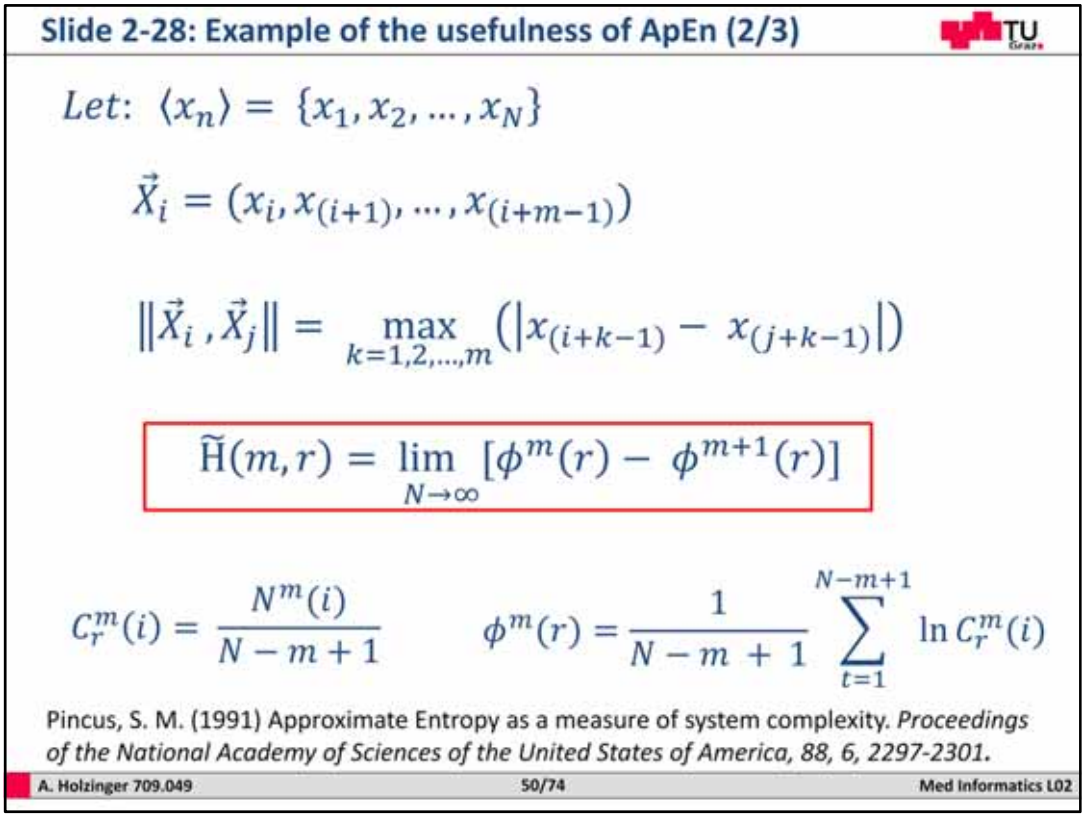


Maximum Entropy (MaxEn), described by (Jaynes, 1957), is used to estimate unknown parameters of a multinomial discrete choice problem, whereas the Generalized Maximum Entropy (GME) model includes noise terms in the multinomial information constraints. Each noise term is modeled as the mean of a finite set of a priori known points in the interval $[-1,1]$ with unknown probabilities where no parametric assumptions about the error distribution are made. A GME model for the multinomial probabilities and for the distributions, associated with the noise terms is derived by maximizing the joint entropy of multinomial and noise distributions, under the assumption of independence (Jaynes, 1957).
Topological Entropy (TopEn), was introduced by (Adler, Konheim & McAndrew, 1965) with the purpose to introduce the notion of entropy as an invariant for continuous mappings: Let (X, T) be a topological dynamical system, i.e., let X be a nonempty compact Hausdorff space and $T: X \to X$ a continuous map; the TopEn is a nonnegative number which measures the complexity of the system (Adler, Downarowicz & Misiurewicz, 2008).
Graph Entropy was described by (Mowshowitz, 1968) to measure structural information content of graphs, and a different definition, more focused on problems in information and coding theory, was introduced by (Körner, 1973). Graph entropy is often used for the characterization of the the structure of graph-based systems, e.g. in mathematical biochemistry. In these applications the entropy of a graph is interpreted as its structural information content and serves as a complexity measure, and such a measure is associated with an equivalence relation defined on a finite graph; by application of Shannon's Eq. 2.4 with the probability distribution we get a numerical value that serves as an index of the structural feature captured by the equivalence relation (Dehmer & Mowshowitz, 2011).

Minimum Entropy (MinEn), described by (Posner, 1975), provides us the least random, and the least uniform probability distribution of a data set, i.e. the minimum uncertainty, which is the limit of our knowledge and of the structure of the system. Often, the classical pattern recognition is described as a quest for minimum entropy. Mathematically, it is more difficult to determine a minimum entropy probability distribution than a maximum entropy probability distribution; while the latter has a global maximum due to the concavity of the entropy, the former has to be obtained by calculating all local minima, consequently the minimum entropy probability distribution may not exist in many cases (Yuan & Kesavan, 1998).
Cross Entropy (CE), discussed by (Rubinstein, 1997), was motivated by an adaptive algorithm for estimating probabilities of rare events in complex stochastic networks, which involves variance minimization. CE can also be used for combinatorial optimization problems (COP). This is done by translating the "deterministic" optimization problem into a related "stochastic" optimization problem and then using rare event simulation techniques (De Boer et al., 2005).
Rényi entropy is a generalization of the Shannon entropy (information theory), and **Tsallis entropy** is a generalization of the standard Boltzmann–Gibbs entropy (statistical physics).
For us more important are:
Approximate Entropy (ApEn), described by (Pincus, 1991), is useable to quantify regularity in data without any a priori knowledge about the system, see an example in Slide 2-20.
Sample Entropy (SampEn), was used by (Richman & Moorman, 2000) for a new related measure of time series regularity. SampEn was designed to reduce the bias of ApEn and is better suited for data sets with known probabilistic content.



Problem: Monitoring body movements along with vital parameters during sleep provides important medical information regarding the general health, and can therefore be used to detect trends (large epidemiology studies) to discover severe illnesses including hypertension (which is enormously increasing in our society).
This seemingly simple data – only from one night period – demonstrates the complexity and the boundaries of standard methods (for example Fast Fourier Transformation) to discover knowledge (for example deviations, similarities etc.).
Due to the complexity and uncertainty of such data sets, standard methods (such as FFT) comprise the danger of modeling artifacts. Since the knowledge of interest for medical purposes is in anomalies (alterations, differences, a-typicalities, irregularities), the application of entropic methods provides benefits.
Photograph taken during the EU Project EMERGE and used with permission.



1) We have a given data set $\langle x_n \rangle$ where capital N is the number of data points:
Eq. 2-13

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

2) Now we form m-dimensional vectors
Eq. 2-14

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

3) We measure the distance between every component, i.e. the maximum absolute difference between their scalar components
Eq. 2-15

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

4) We look – so to say – in which dimension is the biggest difference; as a result we get the Approximate Entropy (if there is no difference we have zero relative entropy):
Eq. 2-16

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

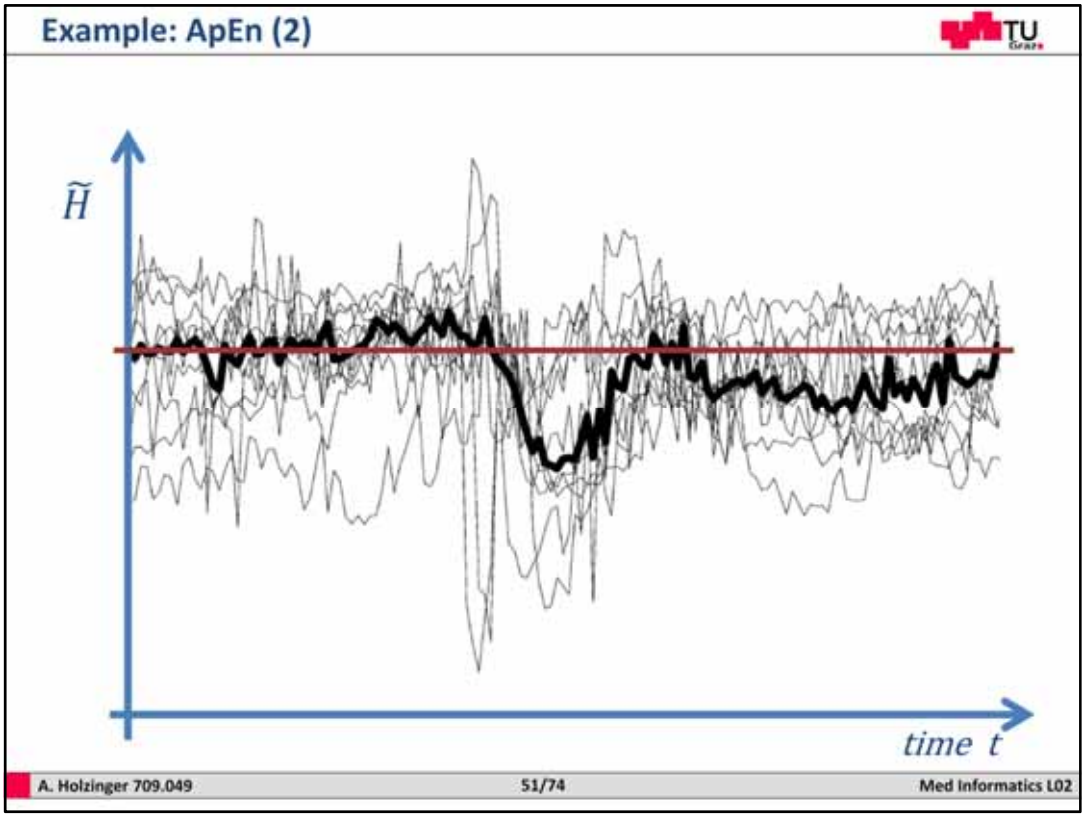
where m is the run length and r is the tolerance window r (let us assume that m is equal to r), ApEn (m,r) could also be written as $\tilde{H}(m,r)$
5) $\phi^m(r)$ is computed by
Eq. 2-17

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

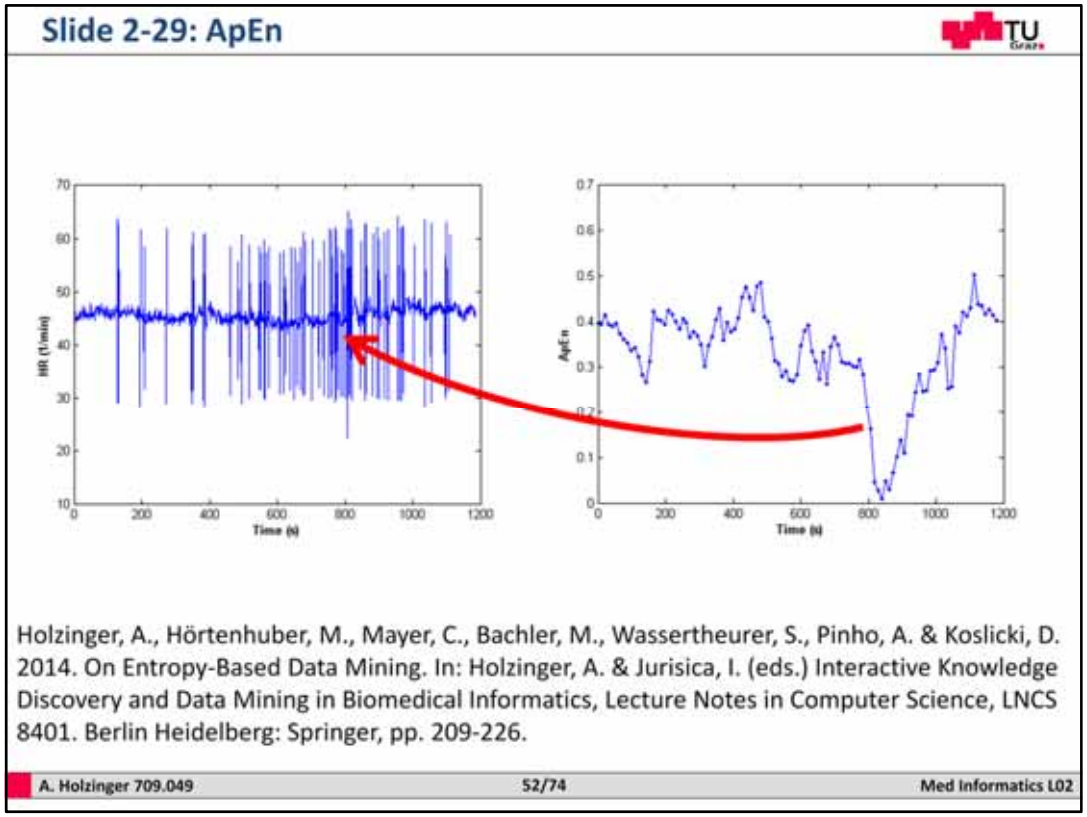
with
Eq. 2-18

$$C_r^m(i) = \frac{N^m(i)}{N-m+1}$$

6) C_r^m measures within the tolerance r the regularity of patterns similar to a given one of window length m
7) Finally we increase the dimension to $m+1$ and repeat the steps before and get as a result the approximate entropy $\mathrm{ApEn}(m,r)$
$\mathrm{ApEn}(m,r,N)$ is approximately the negative natural logarithm of the conditional probability (CP) that a dataset of length N, having repeated itself within a tolerance r for m points, will also repeat itself for $m+1$ points. An important point to keep in mind about the parameter r is that it is commonly expressed as a fraction of the Standard deviation (SD) of the data and in this way makes ApEn a scale-invariant measure. A low value arises from a high probability of repeated template sequences in the data (Hornero et al., 2006).

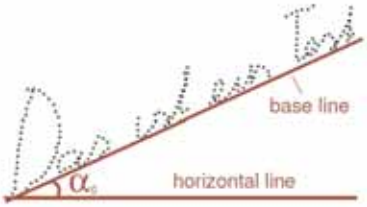


In this slide we can see the plot of the normalized approximate entropy for each of the episodes and the median across all the episodes. From this figure we can see that the entropy is a minimum where we have no alterations and entropy is increasing when having irregularities.
If we have no differences we get zero entropy



A final example should make the advantage of such an entropy method totally clear: In the right diagram it is hard to discover irregularities for a medical professional – especially over a longer period, but an anomaly can easily be detected by displaying the measured relative ApEn.
What can we learn from this experiment? Approximate entropy is relatively unaffected by noise; it can be applied to complex time series with good reproduction; it is finite for stochastic, noisy, composite processes; the values correspond directly to irregularities; and it is applicable to many other areas – for example for the classification of large sets of texts – the ability to guess algorithmically the subject of a text collection without having to read it would permit automated classification.

Example: Skew and Slant correction in Handwriting



$$p_{y,j}(\alpha) = \frac{1}{m+1} \sum_{t=0}^{m} \chi_j(y_\alpha(t))$$

Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350.

A. Holzinger 709.049 53/74 Med Informatics L02

Calculation of Entropy

Algorithm 2 Calculate the entropy $H_{y,\alpha}(X)$ for the projection profiles $p_{y,j}(\alpha)$ for a range of angles α

Require: K = number of data points in X
Require: $\forall k \in \mathbb{N}, 1 \leq k \leq K$: $y_{\min} \leq X[k].y$ **AND** $y_{\max} \geq X[k].y$
Require: $l, \alpha_{\min}, \alpha_{\max} \in \mathbb{N}$ **AND** $-35 \leq \alpha_{\min} < \alpha_{\max} \leq 35$

1: **function** CALCULATEHY($X, K, \alpha_{\min}, \alpha_{\max}, y_{\min}, y_{\max}, l$)
2: $range \leftarrow |\alpha_{\max} - \alpha_{\min}|$
3: $H_y :=$ new vector of size $range$ ▷ Denote the index range from $\alpha_{\min}$ to $\alpha_{\max}$
4: **for** $\alpha = \alpha_{\min} \rightarrow \alpha_{\max}$ **do**
5: $X_\alpha \leftarrow$ ROTATEDATAPOINTS(X, K, α)
6: $w \leftarrow |y_{\max} - y_{\min}| \,/\, l$
7: $p_y \leftarrow$ CALCULATECURRENTPY($X_\alpha, K, y_{\min}, y_{\max}, w$)
8: $H_y[\alpha] \leftarrow 0$
9: **for** $j = 1 \rightarrow l$ **do**
10: $H_y[\alpha] \leftarrow H_y[\alpha] + p_y[j] \cdot \log_2 p_y[j]$
11: **end for**
12: $H_y[\alpha] \leftarrow -1 \cdot H_y[\alpha]$
13: **end for**
14: **return** H_y
15: **end function**

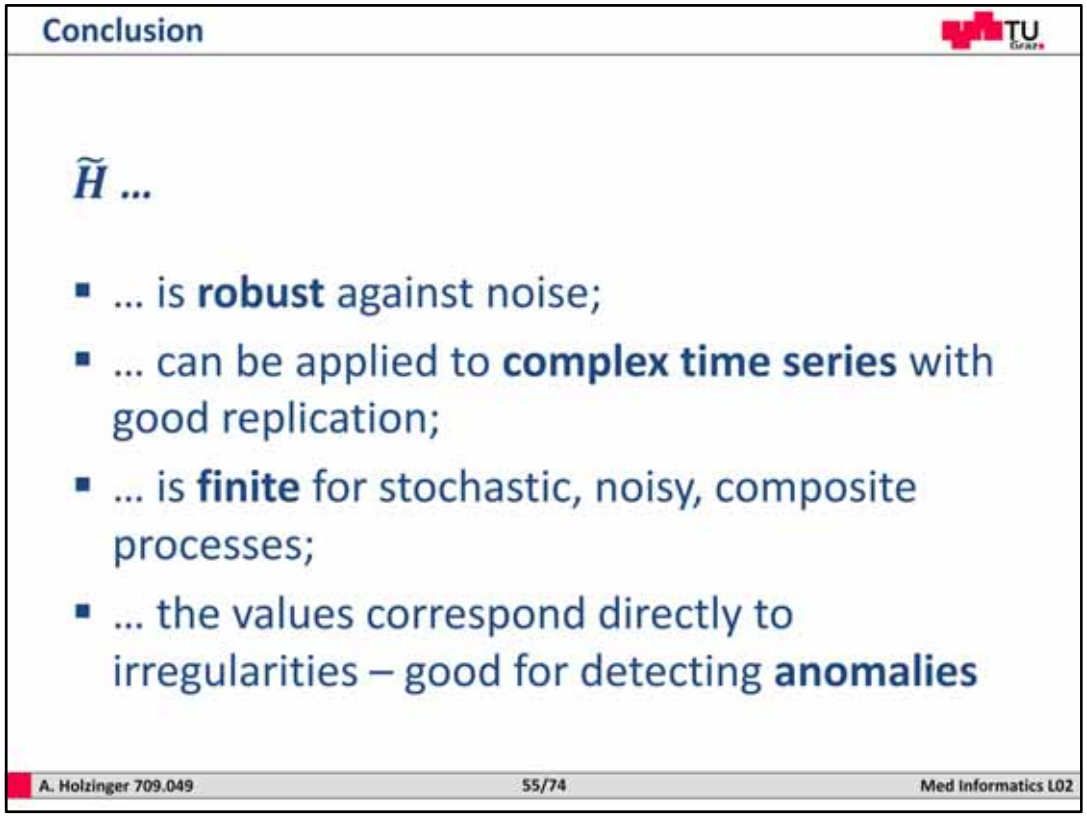
Algorithm 1 Rotating the data points in X for α degree
Require: K = number of data points in X

1: **function** ROTATEDATAPOINTS(X, K, α)
2: $X_\alpha :=$ new vector of size K
3: **for** $k = 1 \rightarrow K$ **do**
4: $X_\alpha[k].x = X[k].x \cdot \cos(\alpha) - X[k].y \cdot \sin(\alpha)$
5: $X_\alpha[k].y = X[k].x \cdot \sin(\alpha) + X[k].y \cdot \cos(\alpha)$
6: $X_\alpha[k].p = X[k].p$
7: **end for**
8: **return** X_α
9: **end function**

Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350.

A. Holzinger 709.049 54/74 Med Informatics L02

Algorithm 1: With rotateDataPoints" defined we can calculate the projection profile p y;j (**α**) for a range of different
angles and with those we can compute the entropy H y,**α** (X) for each angle. Holzinger et al. (2012) implemented this algorithm as it is described in Algorithm 2.
For more information please refer to the paper. This just shall demonstrate the strengths and weaknesses of using entropy for skew- and slant-correction – which is important in handwriting recognition. However, the entropy based skew correction does not outperform older methods like skew correction based on the least squares method: the noise in the drawing distorts the real minima of the entropy distribution. In many cases where the global minimum was the wrong choice, there was a local minimum close to the real error angle. Even though both approaches yield satisfying result for words longer than five letters, we suggest
further investigation into the entropy-based skew correction method, with noise reduction in mind. On the other hand, it shows that entropy is in fact useful when performing slant correction, as it does outperform the window-based approach! The conclusion is, that the window-based
method is too much dependent on a number of factors. Its performance is influenced a lot by the outcome of zone detection and by the writing style of the writer. It is also influenced a lot by window selection.
Holzinger, A., Stocker, C., Peischl, B. & Simonic, K.-M. 2012. On Using Entropy for Enhancing Handwriting Preprocessing. Entropy, 14, (11), 2324-2350



What can we learn from this experiment? Approximate entropy is relatively unaffected by noise; it can be applied to complex time series with good reproducibility; it is finite for stochastic, noisy, composite processes; the values correspond directly to irregularities;

and it is applicable to many other areas – for example for the classification of large sets of texts – the ability to guess algorithmically the subject of a text collection without having to read it would permit automated classification.

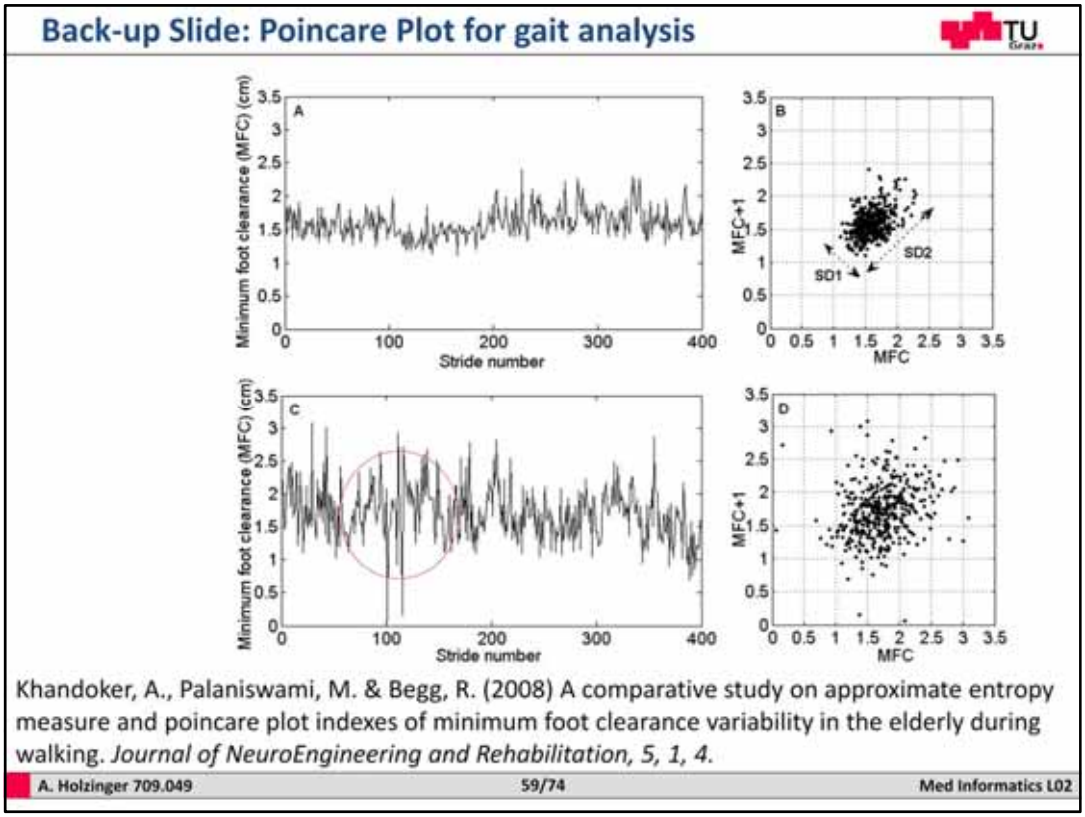
My DEDICATION is to make data valuable … Thank you!

Sample Questions (1)

- Why is modeling of artifacts a huge problem?
- What do we need to transfer information into Knowledge?
- What type of data does the PDB basically store?
- What is the "curse of dimensionality"?
- What type of separable data is blood sedimentation rate?
- Is the mathematical operation "multiplication" allowed with ordinal data?
- What characterizes standardized data?
- Why are structural homologies interesting?
- How did Bemmel & van Musen describe the clinical view on data, information and knowledge?
- Where are the differences between patient data and medical knowledge from a clinical viewpoint?
- Which weaknesses of the DIKW Model do you recognize?
- How do we get theories?
- What is the main limitation of transferring data from the computational space into the perceptual space from the viewpoint of the human information processing model?

Sample Questions (2)

- Why is the knowledge about human information processing necessary for medical informatics?
- What is the difference between the perceptual space and the computational space in terms of data, information and knowledge?
- What does information interaction mean?
- How does knowledge-assisted visualization work in principle?
- Why is non-structured data an rather incorrect term?
- Give an example of the data structure tree in biomedical informatics!
- Why is data quality important? What are the related issues?
- How do you ensure data accessibility?
- What is the main idea of Shannon's Entropy?
- Why is Entropy interesting for medical informatics?
- What are typical entropic methods?
- What is the main purpose of Approximate Entropy?
- What is the big advantage of entropic methods?
- What are the differences of ApEn and SampEn?
- Which possibilities do you have with Graph Entropy Measures?



MFC = Minimum Foot Clearance
Stride = step
You can see brilliantly what you can measure with entropy – you can determine anomalies, i.e. the balance problems of elderly gait

MFC Poincaré plots. Top panels show MFC time series from a healthy elderly subject (A) and its corresponding Poincaré plot (B). Bottom panels show MFC time series from an elderly subject with balance problem (C) and its corresponding Poincaré plot (D).

Significant relationships of mean MFC with Poincaré plot indexes (SD1, SD2) and ApEn ($r = 0.70$, $p < 0.05$; $r = 0.86$, $p < 0.01$; $r = 0.74$, $p < 0.05$) were found in the falls-risk elderly group. On the other hand, such relationships were absent in the healthy elderly group. In contrast, the ApEn values of MFC data series were significantly ($p < 0.05$) correlated with Poincaré plot indexes of MFC in the healthy elderly group, whereas correlations were absent in the falls-risk group. The ApEn values in the falls-risk group (mean ApEn = 0.18 ± 0.03) was significantly ($p < 0.05$) higher than that in the healthy group (mean ApEn = 0.13 ± 0.13). The higher ApEn values in the falls-risk group might indicate increased irregularities and randomness in their gait patterns and an indication of loss of gait control mechanism. ApEn values of randomly shuffled MFC data of falls risk subjects did not show any significant relationship with mean MFC.

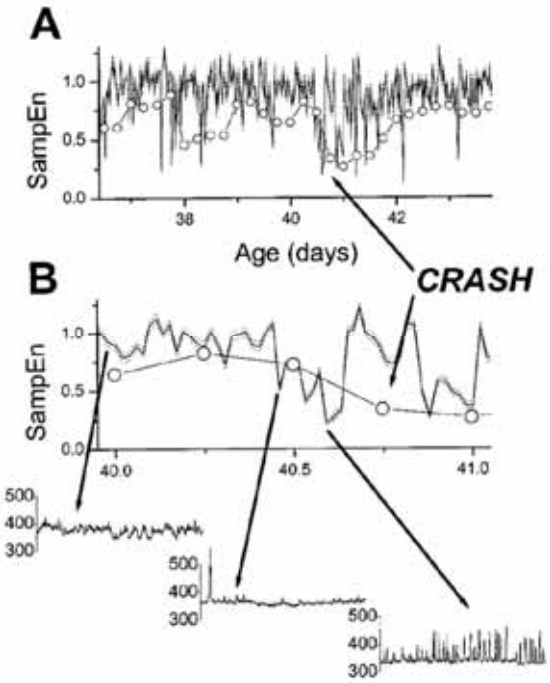
Sample Exam Questions – Yes/No Answers

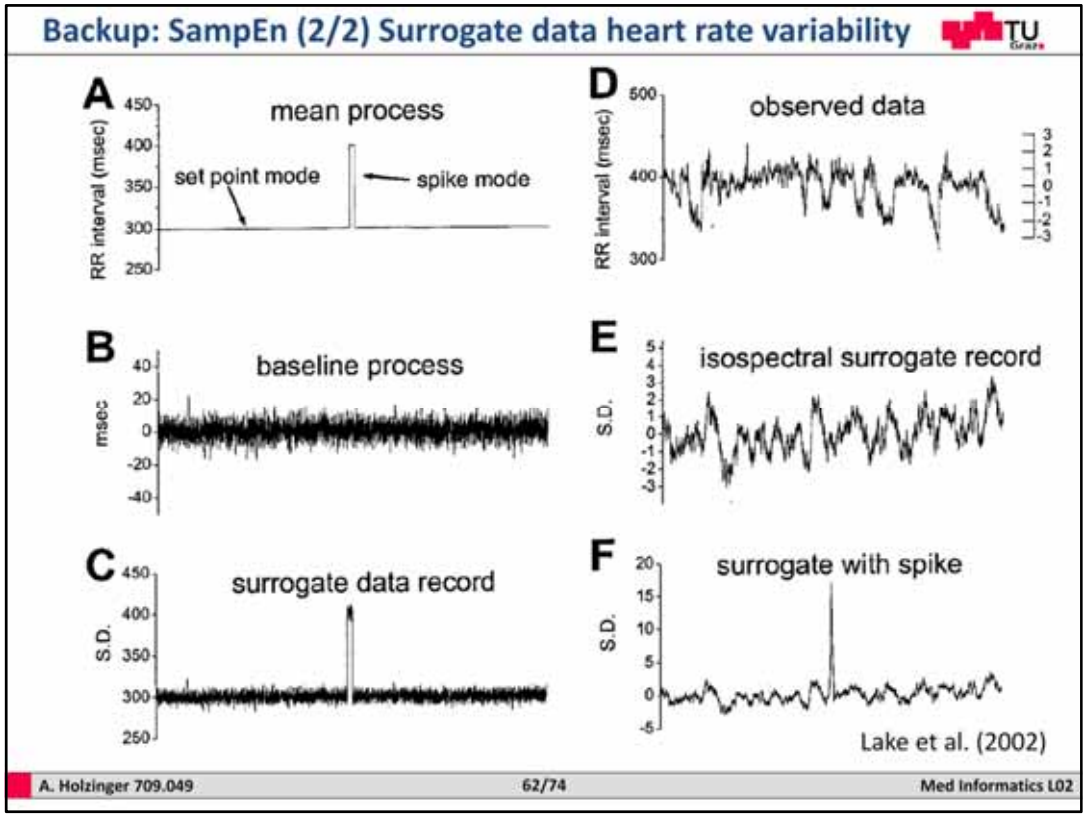
01	An array is a composite data type on physical level.	❒ Yes ❒ No	2 total
02	In a Von-Neumann machine "List" is a widely used data structure for applications which do not need random access.	❒ Yes ❒ No	2 total
03	The edges in a graph can be multidimensional objects, e.g. vectors containing the results of multiple Gen-expression measures.	❒ Yes ❒ No	2 total
04	Each item of data is composed of variables, and if such a data item is defined by more than one variable it is called a multivariable data item	❒ Yes ❒ No	2 total
05	A dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering.	❒ Yes ❒ No	2 total
06	Nominal and ordinal data are parametric, and do assume a particular distribution.	❒ Yes ❒ No	2 total
07	Abstraction is characterized by a cyclical process of generating possible explanations and testing those explanations.	❒ Yes ❒ No	2 total
08	A metric space has an associated metric, which enables us to measure distances between points in that space and, in turn, implicitly define their neighborhoods.	❒ Yes ❒ No	2 total
09	Induction consists of deriving a likely general conclusion from a set of particular statements.	❒ Yes ❒ No	2 total
10	In the model of Boisot & Canals (2004), the perceptual filter orientates the senses (e.g. visual sense) to certain types of stimuli within a certain physical range.	❒ Yes ❒ No	2 total
	Sum of Question Block A (max. 20 points)		

A. Holzinger 709.049 60/74 Med Informatics L02

Back-up Slide: SampEn (1/2)

Lake, D. E., Richman, J. S., Griffin, M. P. & Moorman, J. R. (2002) Sample entropy analysis of neonatal heart rate variability. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology, 283,* ***3, R789-R797.***





Surrogate data records. A and B show the major components. A: the mean process, which has set point and spike modes. B: the baseline process,
here meaning the heart rate variability, modeled as Gaussian random numbers. C: their sum, a surrogate data record. D–F: a more realistic surrogate with the same frequency content as the observed data. D: a clinically observed data record of 4,096 R-R intervals. The lefthand ordinate is labeled in ms and the righthand ordinate in SD. E: a 4,096-point isospectral surrogate dataset formed using the inverse Fourier transform of the periodogram of the data in D. F: the surrogate data after addition of a clinically observed deceleration lasting 50 points and scaled so that the variance of the record is increased from 1 to 2.

Backup Slide: Comparison ApEn - SampEn

ApEn

Given a signal x(n)=x(1), x(2),…, x(N), where N is the total number of data points, ApEn algorithm can be summarized as follows [1]:

1) Form *m*-vectors, *X(1)* to *X(N-m+1)* defined by:

$$X(i) = [x(i), x(i+1), \ldots, X(i+m-1)] \quad i = 1, N-m+1 \quad (1)$$

2) Define the distance *d[X(i),X(j)]* between vectors *X(i)* and *X(j)* as the maximum absolute difference between their respective scalar components:

$$d[X(i), X(j)] = \max_{k=0,m-1} [|x(i+k) - x(j+k)|] \quad (2)$$

3) Define for each i, for i=1, N-m+1, let

$$C_r^m(i) = V^m(i)/(N-m+1) \quad (3)$$

$$\text{where} \quad V^m(i) = no.\,of\, d[X(i), X(j)] \le r$$

4) Take the natural logarithm of each $C_r^m(i)$, and average it over *i* as defined in step 3):

$$\phi^m(r) = \frac{1}{N-m+1} \sum_{i=1}^{N-m+1} \ln(C_r^m(i)) \quad (4)$$

5) Increase the dimension to m+1 and repeat steps 1) to 4).

6) Calculate ApEn value for a finite data length of *N*:

$$ApEn(m, r, N) = \phi^m(r) - \phi^{m+1}(r) \quad (5)$$

Xinnian, C. et al. (2005). *Comparison of the Use of Approximate Entropy and Sample Entropy: Applications to Neural Respiratory Signal. Engineering in Medicine and Biology IEEE-EMBS 2005, 4212-4215.*

SampEn

Given a signal x(n)=x(1), x(2),…, x(N), where N is the total number of data points, SampEn algorithm can be summarized as follows [5]:

1) Form *m*-vectors, *X(1)* to *X(N-m+1)* defined by:

$$X(i) = [x(i), x(i+1), \ldots, X(i+m-1)] \quad i = 1, N-m+1 \quad (6)$$

2) Define the distance $d_m[X(i), X(j)]$ between vectors *X(i)* and *X(j)* as the maximum absolute difference between their respective scalar components:

$$d_m[X(i), X(j)] = \max_{k=0,m-1} [|x(i+k) - x(j+k)|] \quad (7)$$

3) Define for each i, for i=1, N-m, let

$$B_i^m(r) = \frac{1}{N-m-1} \times no.\,of\, d_m[X(i), X(j)] \le r,\ i \ne j \quad (8)$$

4) Similarly, define for each i, for i=1, N-m, let

$$A_i^m(r) = \frac{1}{N-m-1} \times no.\,of\, d_{m+1}[X(i), X(j)] \le r, i \ne j \quad (9)$$

5) Define $B^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} B_i^m(r)$ (10)

$$A^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} A_i^m(r) \quad (11)$$

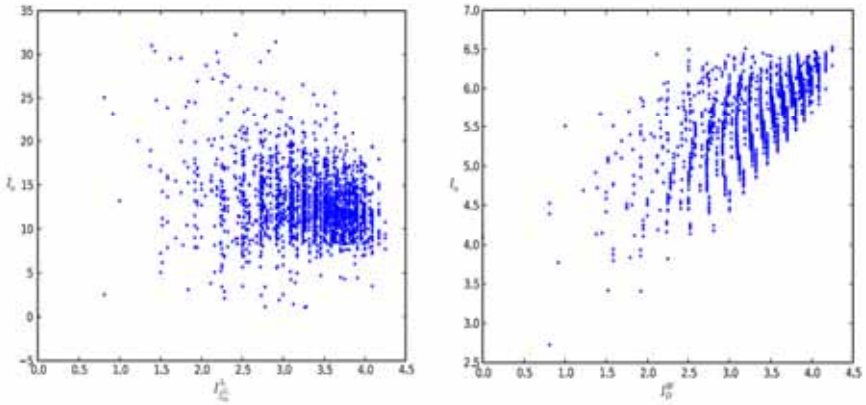
6) SampEn value for a finite data length of *N* can be estimated:

$$SampEn(m, r, N) = -\ln\left(A^m(r) / B^m(r)\right) \quad (12)$$

A. Holzinger 709.049 63/74 Med Informatics L02

Backup Slide: Graph Entropy Measures

- The most important question: Which kind of structural information does the entropy measure detect?
- the topological complexity of a molecular graph is characterized by its number of vertices and edges, branching, cyclicity etc.



Dehmer, M. & Mowshowitz, A. (2011) A history of graph entropy measures. *Information Sciences, 181, 1, 57-78.*

Backup: English/German Subject Codes OEFOS 2012

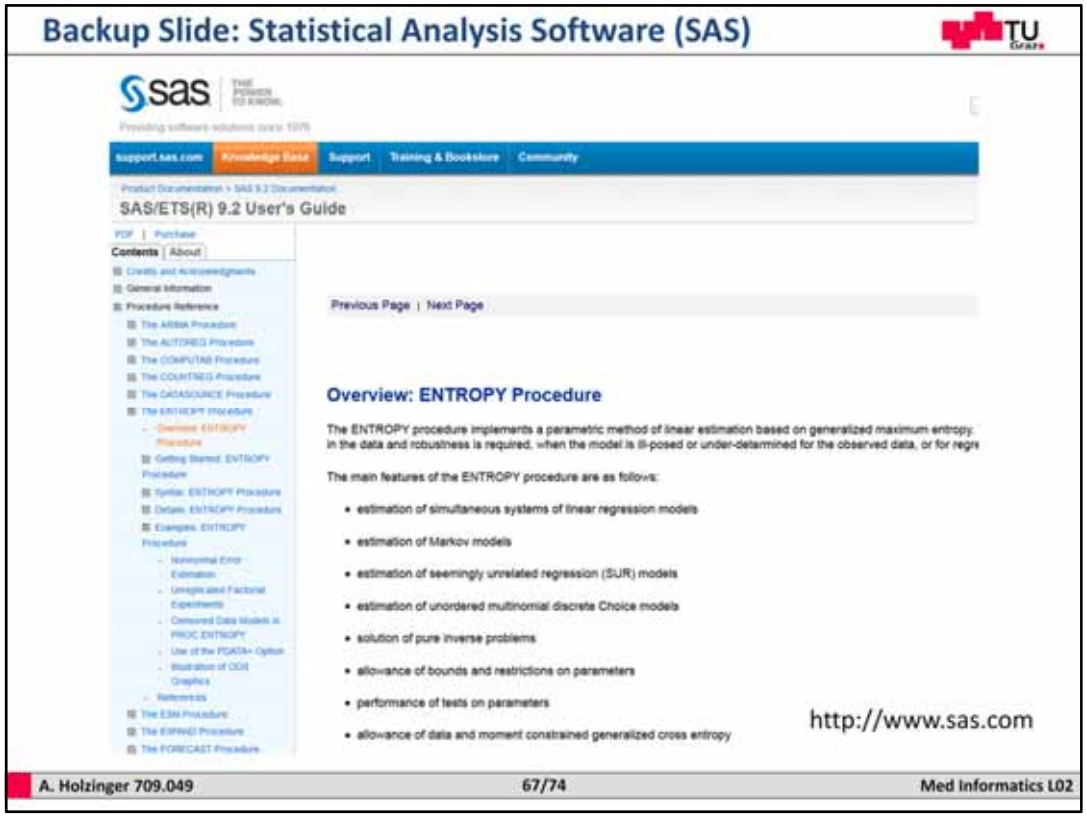
106005	Bioinformatics	Bioinformatik
106007	Biostatistics	Biostatistik
304005	Medical Biotechnology	Medizinische Biotechnologie
305901	Computer-aided diagnosis and therapy	Computerunterstützte Diagnose und Therapie
304003	Genetic engineering, -technology	Gentechnik, -technologie
3906 (old)	Medical computer sciences	Medizinische Computerwissenschaften
305906	Medical cybernetics	Medizinische Kybernetik
305904	Medical documentation	Medizinische Dokumentation
305905	Medical informatics	Medizinische Informatik
305907	Medical statistics	Medizinische Statistik

http://www.statistik.at

Backup: English/German Subject Codes OEFOS 2012

102001	Artificial Intelligence	Künstliche Intelligenz
102032	Computational Intelligence	Computational Intelligence
102033	Data Mining	Data Mining
102013	Human-Computer Interaction	Human-Computer Interaction
102014	Information design	Informationsdesign
102015	Information systems	Informationssysteme
102028	Knowledge engineering	Knowledge Engineering
102019	Machine Learning	Maschinelles Lernen
102020	Medical Informatics	Medizinische Informatik
102021	Pervasive Computing	Pervasive Computing
102022	Software development	Softwarenetwicklung
102027	Web engineering	Web Engineering

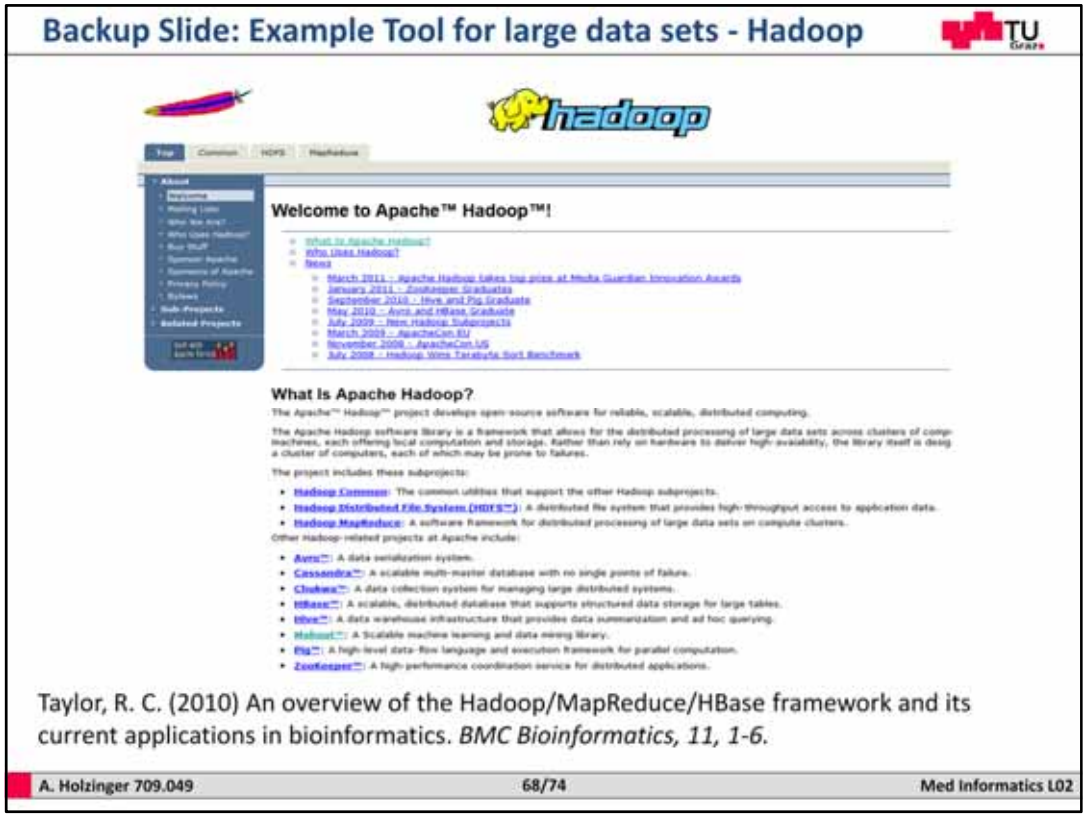
http://www.statistik.at



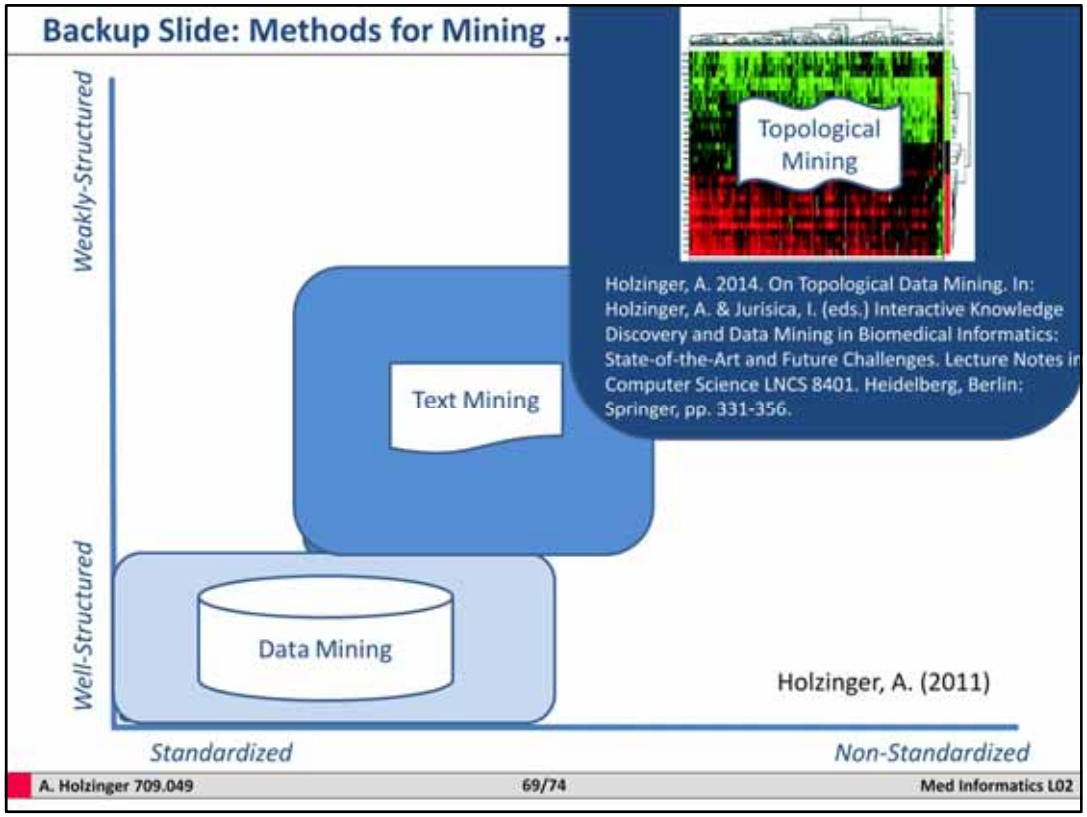
http://support.sas.com/documentation/cdl/en/etsug/60372/HTML/default/viewer.htm#etsug_entropy_sect018.htm

Where many other languages refer to tables, rows, and columns/fields, SAS uses the terms data sets, observations, and variables. There are only two kinds of variables in SAS: numeric and character (string). By default all numeric variables are stored as (8 byte) real. It is possible to reduce precision in external storage only. Date and datetime variables are numeric variables that inherit the C tradition and are stored as either the number of days (for date variables) or seconds (for datetime variables).

http://www.sas.com/technologies/analytics/statistics/stat/index.html

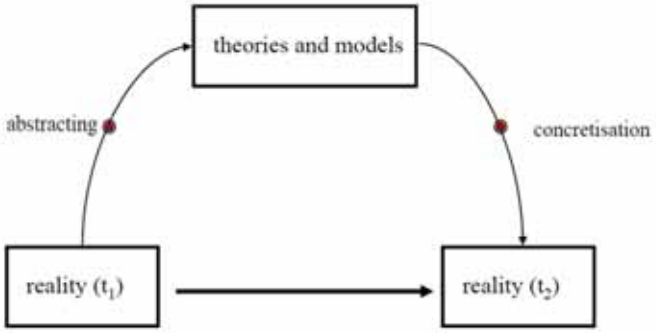


Hadoop and the MapReduce programming paradigm already have a substantial base in the bioinformatics community – in particular in the field of high-throughput next-generation sequencing analysis.
This is due to the cost-effectiveness of Hadoop-based analysis on commodity Linux clusters, and in the cloud via data upload to cloud vendors who have implemented Hadoop/HBase; and due to the effectiveness and ease-of-use of the MapReduce method in parallelization of many data analysis algorithms.



The challenge we face is that an estimated average of 5% of data are structured, the rest is either semi-structured, weakly structured and most of our data is unstructured. Maybe the most important field for the future is data mining – especially novel techniques of data mining, including both time and space (e.g. graph-based, entropy-based, topological-based data mining approaches).
Read more here:
Holzinger, A. 2014. Extravaganza Tutorial on Hot Ideas for Interactive Knowledge Discovery and Data Mining in Biomedical Informatics. In: Slezak, D., Tan, A.-H., Peters, J. F. & Schwabe, L. (eds.) Brain Informatics and Health, BIH 2014, Lecture Notes in Artificial Intelligence, LNAI 8609. Heidelberg, Berlin: Springer, pp. 502-515.
https://online.tugraz.at/tug_online/voe_main2.getVollText?pDocumentNr=764238&pCurrPk=79139

Backup Slide: Excursion: How to get theories?



positivism :

{theory, model} $\notin$ reality

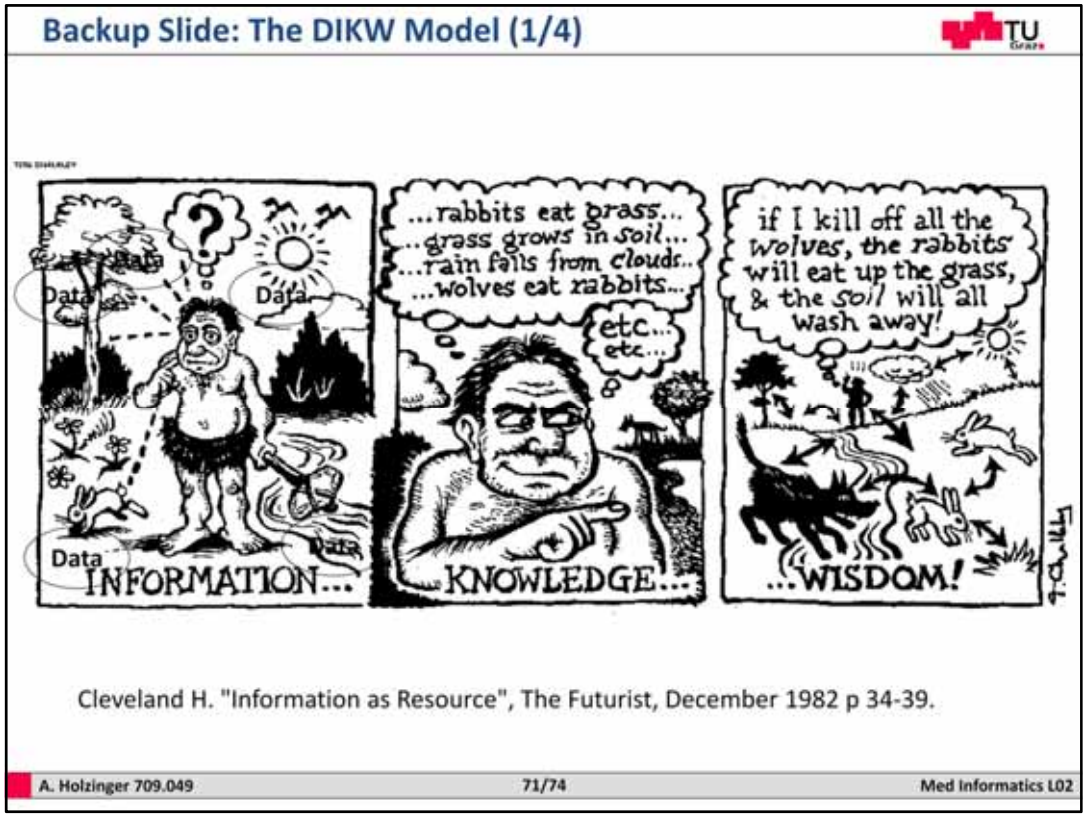
reality (t_1) $\approx$ reality (t_2)

constructionism :

{theory, model} $\in$ reality

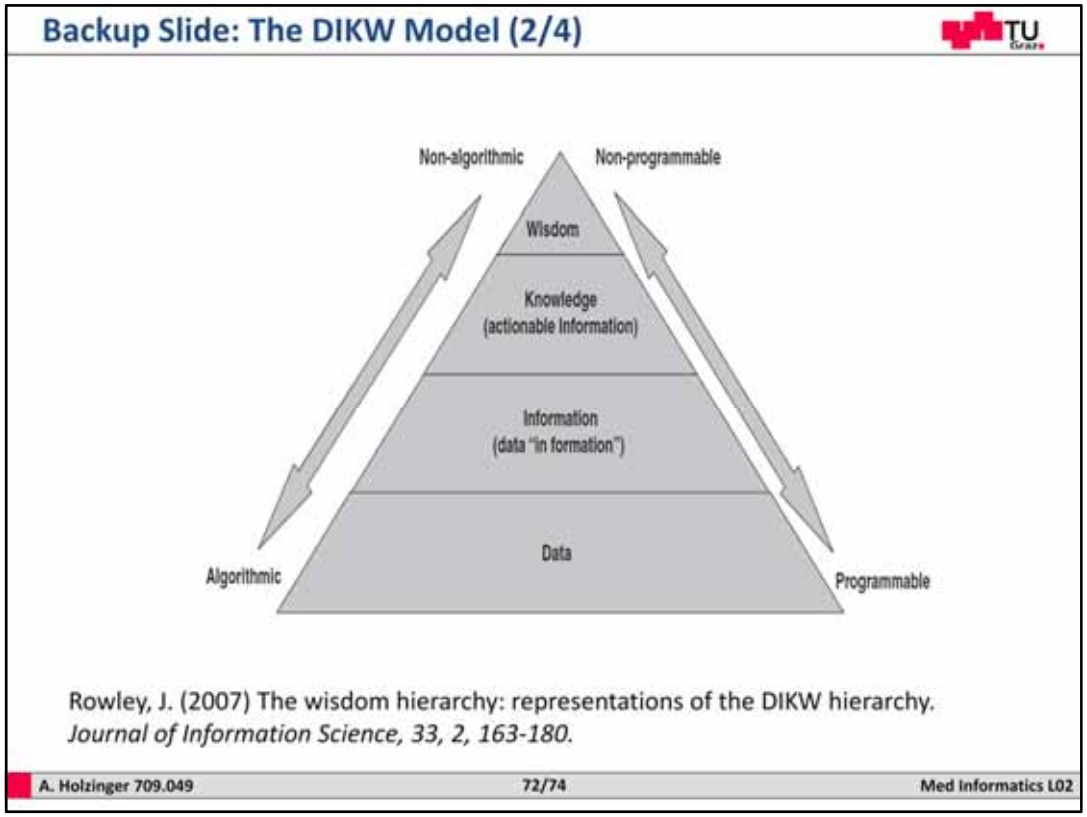
reality (t_1) $\neq$ reality (t_2)

Rauterberg, M. (2006) HCI as an engineering discipline: to be or not to be.
African Journal of Information and Communication Technology, 2, 4, 163-184.



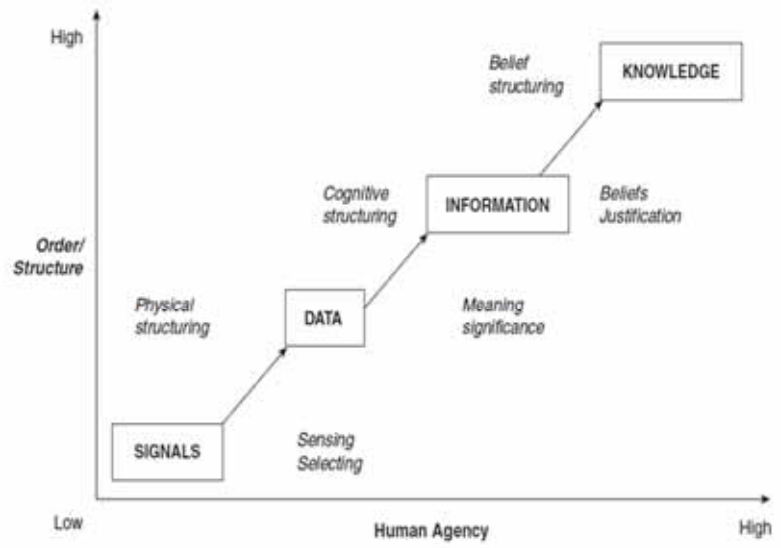
http://minnesotafuturist.pbworks.com/w/page/21441129/DIKW

A funny description of data information knowledge.



A very placative image. Nice to look at – but the usefulness is questionable.

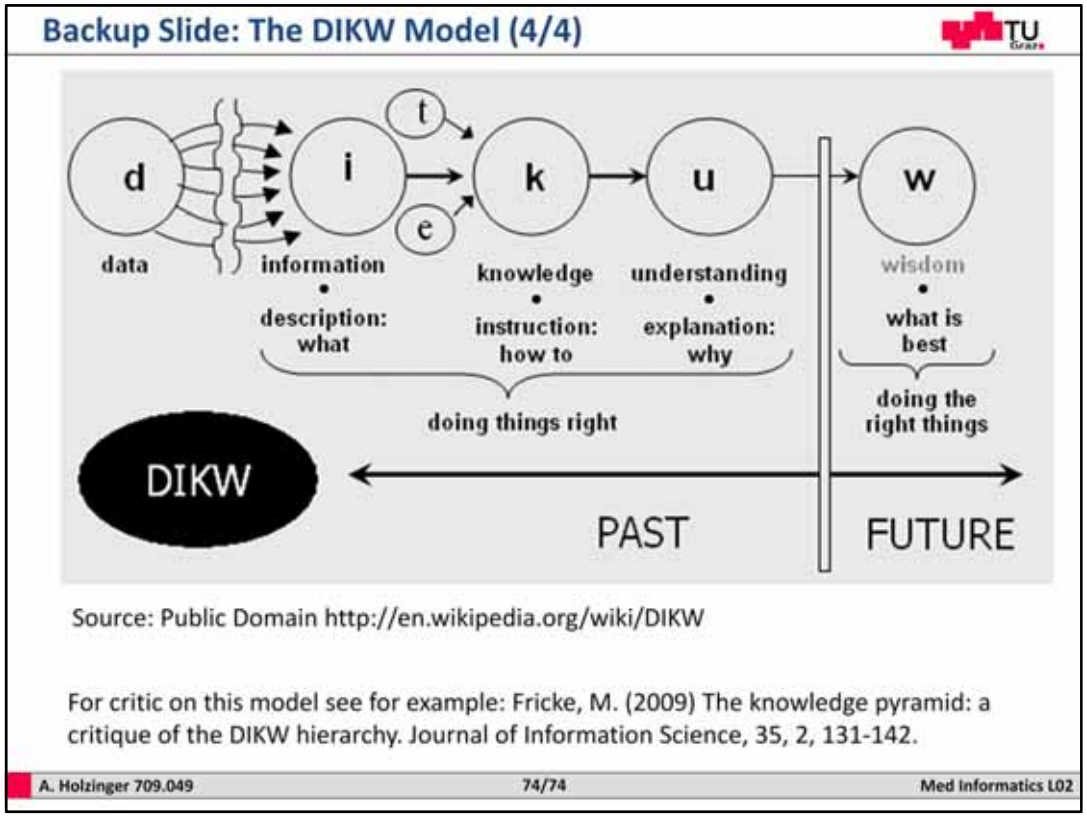
Backup Slide: The DIKW Model (3/4)



Rowley, J. (2007) The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science, 33, 2, 163-180.*

A. Holzinger 709.049 73/74 Med Informatics L02

All this models are very questionable. Please remember that we follow in our lecture the notion of Boisot & Canals.



The interesting issue of this graphic is that it includes a time-axis, which is important for decision making and predictive analytics.
"Past behaviour is a good predictor for future behaviour"
Although this is a oversimplification, scientists who study human behavior agree that past behavior may be a useful marker for future behavior, however, only under certain specific conditions. Read more:
Ajzen, I. 1991. The theory of planned behavior. Organizational behavior and human decision processes, 50, (2), 179-211.