Welcome Students!

At first some organizational details:

1) Duration

This course LV 709.049 (formerly LV 444.152) is a one-semester course and consists of **12 lectures**

(see Overview in Slide 0-1) each with a duration of 90 minutes.

2) Topics

This course covers the **computer science aspects of biomedical informatics** (= medical informatics + bioinformatics) with a focus on new topics such as "big data" concentrating on **algorithmic and methodological issues.**

3) Audience

This course is suited for students of Biomedical Engineering (253), students of Telematics (411), students of Software Engineering (524, 924) and students of Informatics (521, 921) with interest in the **computational sciences** with the application area **biomedicine and health.** PhD students and international students are cordially welcome.

4) Language

The language of Science and Engineering is **English,** as it was Greek in ancient times and Latin in mediaeval times, for more information please refer to:

Holzinger, A. 2010. *Process Guide for Students for Interdisciplinary Work in Computer Science/Informatics.* Second Edition, Norderstedt: BoD.
http://www.amazon.de/Process-Students-Interdisciplinary-Computer-Informatics/dp/384232457X
http://castor.tugraz.at/F?func=direct&doc_number=000403422

Accompanying Reading
ALL exam questions with solutions can be found in the Springer textbook available at the Library:
Andreas Holzinger (2014). Biomedical Informatics: Discovering Knowledge in Big Data, New York: Springer.
DOI: 10.1007/978-3-319-04528-3

Holzinger, A. 2012. **Biomedical Informatics: Computational Sciences meets Life Sciences,** Norderstedt, BoD.

The first edition of the lecture notes is available within the university library, see:
http://castor.tugraz.at/F?func=direct&doc_number=000431288

or via Amazon:
http://www.amazon.de/Biomedical-Informatics-Lecture-Notes-444-152/dp/3848222191

2) Alternatively, you can read the Kindle-Edition on a mobile device:
http://www.amazon.de/Biomedical-Informatics-Lecture-444-152-ebook/dp/B009GT0LIM/ref=dp_kinw_strp_1

Slide 0-1: Overview – Roadmap trough this Course

- **01. Intro: Computer Science meets Life Sciences, challenges, future directions**
- 02. Fundamentals of Data, Information and Knowledge
- 03. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)
- 04. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- 05. Semi structured , weakly structured data and unstructured information
- 06. Multimedia Data Mining and Knowledge Discovery
- 07. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 08. Biomedical Decision Making: Reasoning and Decision Support
- 09. Interactive Information Visualization and Visual Analytics
- 10. Biomedical Information Systems and Medical Knowledge Management
- 11. Biomedical Data: Privacy, Safety and Security
- 12. Methodology for Info Systems: System Design, Usability & Evaluation

A. Holzinger 709.049   3/80   Med Informatics L01

The course consists of the following 12 lectures:

1. **Introduction: Computer Science meets Life Sciences.** We start with the basics of life sciences, including biochemical and genetic fundamentals, some cell-physiological basics and a brief overview about the human body; we answer the question "what is biomedical informatics" and we conclude with an outlook into the future.

2. **Fundamentals of Data, Information and Knowledge.** In the 2nd 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.

3. **Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS).** In the 3rd lecture we focus on standardization, ontologies and classifications, in particular on the International Statistical Classification of Diseases, the Systematized Nomenclature of Medicine, Medical Subject Headings and the Unified Medical Language.

4. **Biomedical Databases: Acquisition, Storage, Information Retrieval and Use.** In the 4th lecture we get a first impression of a hospital information system, we discuss some basics of data warehouse systems and biomedical data banks and we concentrate on information retrieval.

5. **Semi structured, weakly structured and unstructured data.** In the 5th lecture we review some basics of XML, before we concentrate on network theory and discuss transcriptional regulatory networks, protein-protein networks and metabolic networks.

6. **Multimedia Data Mining and Knowledge Discovery.** In the 6th lecture we determine types of knowledge, focus on the basics of data mining and close with text mining and semantic methods, such as Latent Semantic Analysis, Latent Dirichlet Allocation and Principal Component Analysis.

7. **Knowledge and Decision: Cognitive Science and Human-Computer Interaction.** In the 7th lecture we review the fundamentals of perception, attention and cognition and discuss the human decision making process, reasoning and problem solving, learn some principles of differential diagnosis and a few basics on human error.

8. **Biomedical Decision Making: Reasoning and Decision Support.** In the 8th lecture we start with the question "Can computers help doctors to make better decisions?", and apply the basics from lecture 7 to the principles of decision support systems and case based reasoning systems.

9. **Interactive Information Visualization and Visual Analytics.** In the 9th lecture we start with the basics of visualization science and review some visualization methods, including Parallel Coordinates, Radial Coordinates, Star Plots and learn a few things about the design of interactive visualizations.

10. **Biomedical Information Systems and Medical Knowledge Management.** In the 10th lecture we discuss workflow modeling, some basics of business enterprise hospital information systems, Picture Archiving and Communication Systems and some standards, including DICOM and HL-7.

11. **Biomedical Data: Privacy, Safety and Security.** In the 11th lecture we start with the famous IOM "Why do accidents happen?" report and its influence on safety engineering, and concentrate on aspects of data protection and privacy issues of medical data.

12. **Methodology for Information Systems: System Design, Usability and Evaluation.** Finally in the 12th lecture we slip into the developer perspective and have a look on design standards, usability engineering methods and on how we evaluate such systems.

Keywords of Lecture 01

- Big Data
- Life
- Proteins – DNA & RNA – Cell – Tissue – Organ – Cardiovascular Systems
- Medicine – Informatics – Computer
- Personalized Medicine
- Translational Informatics – Data Integration
- Open Medical Data
- Biomarker Discovery

A. Holzinger 709.049   4/80   Med Informatics L01

The keywords of the first lecture include:
1) Big Data – Our world in data – from macroscopic data to microscopic data
2) What is Life?
3) Proteins – DNA & RNA – Cell – Tissue – Organ – Cardiovascular Systems
4) Medicine – Informatics – Computer
5) Personalized Medicine (between Standardization and Individualization)
6) Translational Informatics – Data Integration (data fusion)
7) Open Medical Data
8) Biomarker Discovery

Learning Goals

- At the end of this first lecture you will …
- … be fascinated to see our world in data;
- … have a basic understanding of the building blocks of life;
- … be familiar with some differences between Life Sciences and Computer Sciences;
- … be aware of some possibilities and some limits of Biomedical Informatics;
- … have some ideas of some future directions of Biomedical Informatics;

A. Holzinger 709.049   5/80   Med Informatics L01

The first lecture shall provide an insight into the fascinating **world of data** from various dimensions – from the macroscopic to the microscopic. You will be rapidly aware of the fact that two issues are most challenging in science: **time and space.**

Note: The colloquial space most familiar to us is called the **Euclidean vector space** $\mathbb{R}^n$ which is the space of all n-tuples of real numbers $(x1, x2, \dots xn)$. The $\mathbb{R}^2$ is therefore called the Euclidean plane. In Special Relativity Theory of Albert Einstein, this Euclidean three-dimensional space plus the time (often called fourth dimension) are unified into the so-called **Minkowski space**. For us in data mining one of the most important spaces is the **Topological space.**

Advance Organizer (1/2)

- **Bioinformatics** = discipline, as part of biomedical informatics, at the interface between *biol*ogy and *infor*mation science and mathema*tics*; processing of biological data;
- **Biomarker** = a characteristic (e.g. body-temperature (fever) as a biomarker for an infection, or proteins measured in the urine) as an indicator for normal or pathogenic biological processes, or pharmacologic responses to a therapeutic intervention;
- **Biomedical data** = compared with general data, it is characterized by large volumes, complex structures, high dimensionality, evolving biological concepts, and insufficient data modeling practices;
- **Biomedical Informatics** = 2011-definition: similar to medical informatics but including the optimal use of biomedical data, e.g. from genomics, proteomics, metabolomics;
- **Classical Medicine** = is both the science and the art of healing and encompasses a variety of practices to maintain and restore health;
- **Genomics** = branch of molecular biology which is concerned with the structure, function, mapping & evolution of genomes;
- **Medical Informatics** = 1970-definition: "... scientific field that deals with the storage, retrieval, and optimal use of medical information, data, and knowledge for problem solving and decision making";
- **Metabolomics** = study of chemical processes involving metabolites (e.g. enzymes). A challenge is to integrate proteomic, transcriptomic, and metabolomic information to provide a more complete understanding of living organisms;
- **Molecular Medicine** = emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs;

A. Holzinger 709.049   6/80   Med Informatics L01

Ausubel (1960) hypothesized that learning and retention of unfamiliar but meaningful verbal material can be facilitated by the advance introduction of relevant subsuming concepts (organizers). A rapid skimming of the definitions above may help in that respect.

Ausubel, D. P. 1960. The use of advance organizers in the learning and retention of meaningful verbal material. *Journal of Educational Psychology, 51, 267-272.*

Advance Organizer (2/2)

- **Omics data** = data from e.g. genomics, proteomics, metabolomics, etc.
- **Pervasive Computing** = similar to ubiquitous computing (Ubicomp), a post-desktop model of Human-Computer Interaction (HCI) in which information processing is integrated into every-day, miniaturized and embedded objects and activities; having some degree of "intelligence";
- **Pervasive Health** = all unobtrusive, analytical, diagnostic, supportive etc. information functions to improve health care, e.g. remote, automated patient monitoring, diagnosis, home care, self-care, independent living, etc.;
- **Proteome** = the entire complement of proteins that is expressed by a cell, tissue, or organism;
- **Proteomics** = field of molecular biology concerned with determining the proteome;
- **P-Health Model** = Preventive, Participatory, Pre-emptive, Personalized, Predictive, Pervasive (= available to anybody, anytime, anywhere);
- **Space** = a set with some added structure;
- **Technological Performance** = machine "capabilities", e.g. short response time, high throughput, high availability, etc.
- **Time** = a dimension in which events can be ordered along a time line from the past through the present into the future;
- **Translational Medicine** = based on interventional epidemiology; progress of Evidence-Based Medicine (EBM), integrates research from basic science for patient care and prevention;
- **Von-Neumann-Computer** = a 1945 architecture, which still is the predominant machine architecture of today (opp.: Non-Vons, incl. analogue, optical, quantum computers, cell processors, DNA and neural nets (in silico));

A. Holzinger 709.049   7/80   Med Informatics L01

Note: The current and future trend towards personalized medicine has resulted in an explosion in the amount of biomedical data, most of them so-called Omics data, which include but are not limited to data from:

Genomics = study of the genomes of organisms, DNA, genetic mapping, heterosis, epistasis, pleiotropy etc.

Proteomics = study of proteins, especially their structures, functions and interactions etc.

Metabolomics = study of chemical processes involving metabolites, e.g. in the physiology of a cell etc.

Lipidomics = study of pathways and networks of cellular lipids etc.

Transcriptomics = examines the expression level of mRNAs etc.

Epigenetics = study of changes in gene expression or cellular phenotypes etc.

Microbiomics = study of microbiomes of an organism, i.e. the ecological community of commensal, symbiotic, and pathogenic microorganisms that share or body space etc.

Fluxomics = study of the flow of fluid and molecules within the cell

Phenomics = study of the measurement of phenomes e.g. physical and biochemical traits of organisms

For Microbiomics read a current article:

http://www.sciencemag.org/site/products/lst_20130510.xhtml

Cascante, M. & Marin, S. 2008. Metabolomics and fluxomics approaches. Essays Biochemistry, 45, 67-81.
http://www.ncbi.nlm.nih.gov/pubmed/18793124

Acronyms/Abbreviations in Lecture 01

- AI = Artificial Intelligence
- AL = Artificial Life
- CPG = Clinical Practice Guideline
- CPOE = Computerized physician order entry
- CMV = Controlled Medical Vocabulary
- DEC = Digital Equipment Corporation (1957-1998)
- DNA = Deoxyribonucleic Acid
- EBM = Evidence Based Medicine
- EPR = Electronic Patient Record
- GBM = Genome Based Medicine
- GC = Gas Chromatography
- GPM = Genetic Polymorphism
- HCI = Human–Computer Interaction
- LC = Liquid Chromatography
- LNCS = Lecture Notes in Computer Science
- MS = Mass Spectrometry
- mRNA = Messenger RNA
- NGC = New General Catalogue of Nebulae and Star clusters in Astronomy
- NGS = Next Generation Sequencing
- NMR = Nuclear Magnetic Resonance
- PDB = Protein Data Base
- PDP = Programmable Data Processor (mainframe)
- PPI = Protein-Protein Interaction
- RFID = Radio-frequency identification device
- RNA = Ribonucleic Acid
- SNP = Single Nucleotide Polymorphism
- TNF = Tumor Necrosis Factor
- TQM = Total Quality Management

A. Holzinger 709.049   8/80   Med Informatics L01

The abbreviations and acronyms used in this lecture are explained here.

Remark: DEC is short for Digital Equipment Corporation and aka (= also known as) Digital and was a major pioneer in the computer industry between 1957 and 1998, and its PDP mainframe computers and the VAX (short for virtual address extension) were the most widespread of all minicomputers world wide.

Key Problems

- **Zillions** of different biological species (humans, animals, bacteria, virus, plants, ...);
- Enormous **complexity** of the medical domain [1];
- **Complex,** heterogeneous, high-dimensional, big data in the life sciences [2];
- Limited **time,** e.g. a medical doctor in a public hospital has only 5 min. to make a decision [3];
- Limited **computational power** in comparison to the complexity of life (and the natural limitations of the Von-Neumann architecture, ...);

1. Patel VL, Kahol K, & Buchman T (2011) Biomedical Complexity and Error. *J. Biomed. Inform.* 44(3):387-389.
2. 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.
3. Gigerenzer G (2008) Gut Feelings: Short Cuts to Better Decision Making (Penguin, London).

A. Holzinger 709.049   9/80   Med Informatics L01

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.

Patel, V. L., Kahol, K. & Buchman, T. 2011. Biomedical Complexity and Error. Journal of Biomedical Informatics, 44, (3), 387-389.

Gigerenzer, G. 2008. Gut Feelings: Short Cuts to Better Decision Making London, Penguin.

Slide 1-1: Our World in Data (1/2) – Macroscopic Structures

What is the challenge ?

ESO, Atacama, Chile (2011)

A. Holzinger 709.049   10/80   Med Informatics L01

1) What is the challenge?
Let's start with a look at some macroscopic data: Here we see the globular star cluster NGC 5139 Omega Centauri, discovered by Edmund Halley in 1677, with a diameter of about 90 light years, including several millions of stars, and approx. 16,000 light-years away from earth; look at the structure – and consider the aspect of time: when our eyes recognize this structure - it might even no longer exist.

Time and space are the most fascinating principles of our world (Hawking, Penrose & Atiyah, 1996).

2) Our challenge is **Big Data.**

References:

Hawking, S. W., Penrose, R. & Atiyah, M. 1996. The nature of space and time, Princeton University Press Princeton.

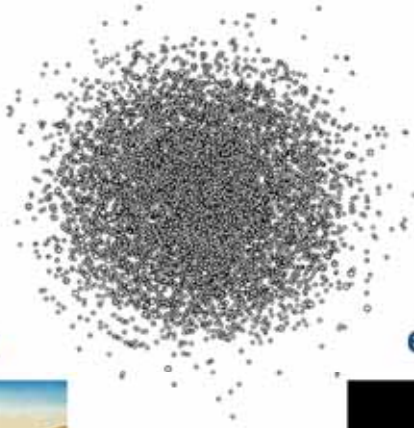Image credit: ESO. Acknowledgement: A. Grado/INAF-Capodimonte Observatory, Online available via: http://www.eso.org/public/images/eso1119b

Excursus: Two thematic mainstreams in dealing with data ...

Time

e.g. Entropy

Space

e.g. Topology

Dali, S. (1931) The persistence of memory

Bagula & Bourke (2012) Klein-Bottle

A. Holzinger 709.049   11/80   Med Informatics L01

The language of nature is mathematics and what you cannot express in mathematical terms you cannot measure (Roger Bacon, 1214-1294)

1) Above: point cloud data set in an arbitrarily high dimensional space

2) Left: The persistence of memory (Time) by Salvador Dali
http://en.wikipedia.org/wiki/The_Persistence_of_Memory

3) Right: A Klein-Bottle – the synonym for "algebraic topology"
http://paulbourke.net/geometry/klein/

Recent example to read on data and time:
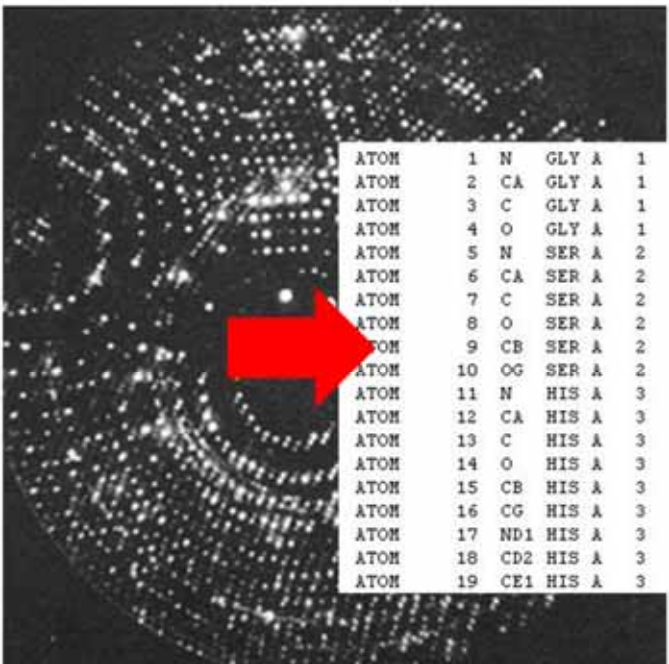
Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J. & Keogh, E. 2013. Addressing Big Data Time Series: Mining Trillions of Time Series Subsequences Under Dynamic Time Warping. ACM Transactions on Knowledge Discovery and Data Mining, 7, (3), 1-31.

Recent example to read on data and space:
Łukasik, S. & Kulczycki, P. 2013. Using Topology Preservation Measures for Multidimensional Intelligent Data Analysis in the Reduced Feature Space. In: Rutkowski, L., Korytkowski, M., Scherer, R., Tadeusiewicz, R., Zadeh, L. & Zurada, J. (eds.) Artificial Intelligence and Soft Computing. Springer Berlin Heidelberg, pp. 184-193.

Slide 1-2: Our World in Data (2/2) – Microscopic Structures

ATOM	1	N	GLY	A	1	44.842	51.034	101.284	0.01	27.20
ATOM	2	CA	GLY	A	1	45.640	50.230	100.389	0.01	26.99
ATOM	3	C	GLY	A	1	46.692	49.648	101.308	0.01	26.80
ATOM	4	O	GLY	A	1	46.895	50.222	102.381	0.01	26.91
ATOM	5	N	SER	A	2	47.283	48.516	100.951	1.00	26.26
ATOM	6	CA	SER	A	2	48.277	47.866	101.761	1.00	26.17
ATOM	7	C	SER	A	2	49.212	47.031	100.845	1.00	24.21
ATOM	8	O	SER	A	2	49.060	47.195	99.630	1.00	19.77
ATOM	9	CB	SER	A	2	47.438	47.091	102.800	1.00	26.31
ATOM	10	OG	SER	A	2	46.276	46.356	102.404	1.00	27.99
ATOM	11	N	HIS	A	3	50.147	46.186	101.370	1.00	23.93
ATOM	12	CA	HIS	A	3	51.129	45.389	100.609	1.00	21.44
ATOM	13	C	HIS	A	3	50.953	43.905	100.849	1.00	20.32
ATOM	14	O	HIS	A	3	50.530	43.595	101.950	1.00	22.00
ATOM	15	CB	HIS	A	3	52.555	45.674	100.990	1.00	19.69
ATOM	16	CG	HIS	A	3	52.940	47.090	100.611	1.00	21.44
ATOM	17	ND1	HIS	A	3	53.371	47.470	99.422	1.00	20.87
ATOM	18	CD2	HIS	A	3	52.956	48.175	101.433	1.00	21.69
ATOM	19	CE1	HIS	A	3	53.676	48.730	99.476	1.00	20.57

Wiltgen, M. & Holzinger, A. (2005) Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: *Central European Multimedia and Virtual Reality Conference. Prague, Czech Technical University (CTU), 69-74*

A. Holzinger 709.049   12/80   Med Informatics L01
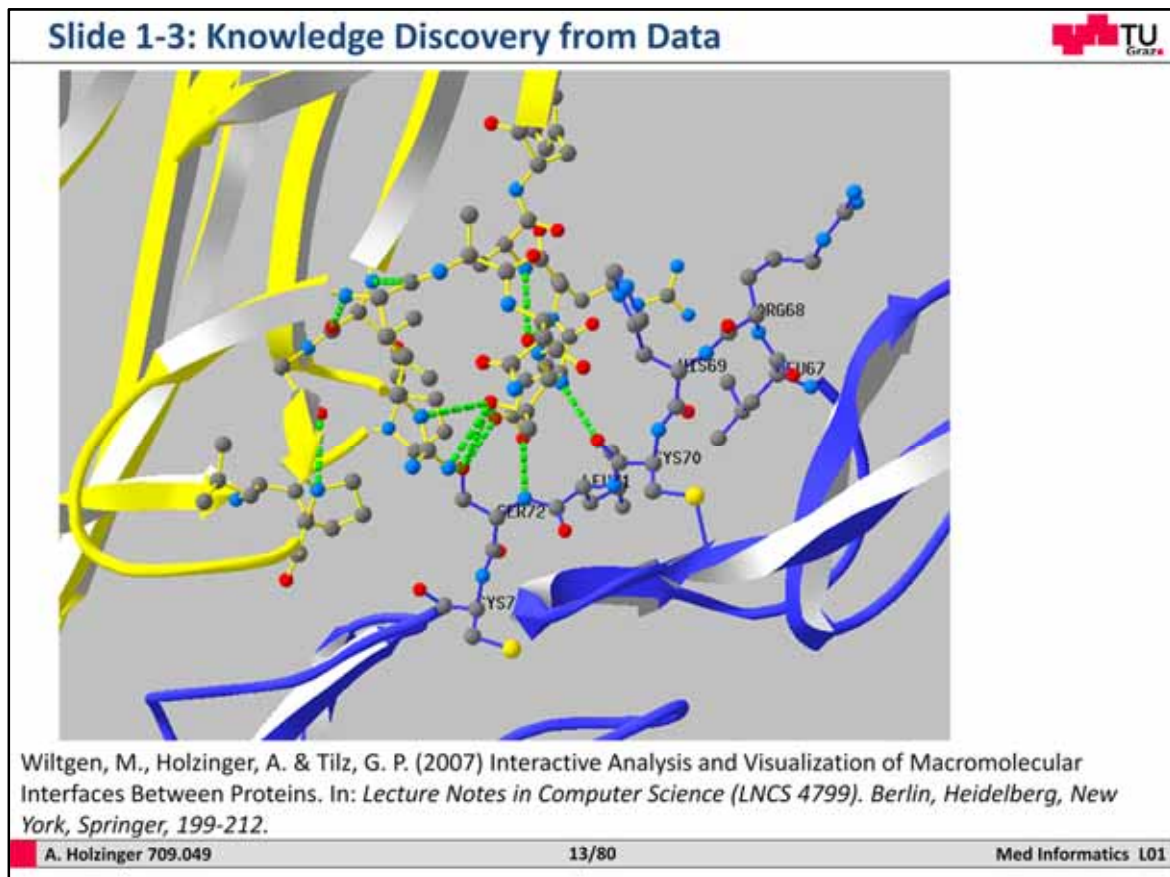
From these large macroscopic structures let us switch to tiny microscopic structures: **Proteins.** These are organic compounds having large molecules composed of long chains of amino acids. Proteins are essential for all living organisms and are used by cells for performing and controlling cellular processes, including degradation and biosynthesis of molecules, physiological signaling, energy storage and conversion, formation of cellular structures etc. Protein structures are determined with crystallographic methods or by nuclear magnetic resonance spectroscopy. Once the atomic coordinates of the protein structure have been determined, a table of these coordinates is deposited into a **protein database (PDB),** an international repository for 3D structure files (see →Lecture 4). In Slide 1-3 we see such a structure and the data, representing the mean positions of the entities within the substance, their chemical relationship etc. (Wiltgen & Holzinger, 2005).

Structures of protein complexes, determined by X-ray crystallography, and the data stored in the PDB database. X-ray crystallography is a standard method to analyze the arrangement of objects (atoms, molecules) within a crystal structure. This data contains the mean positions of the entities within the substance, their chemical relationship, and various others. If a medical professional looks at the data, he or she sees only lengthy tables of numbers, the quest is now to get knowledge out of this data (see Slide 1-4).

Reference:
Wiltgen, M. & Holzinger, A. 2005. Visualization in Bioinformatics: Protein Structures with Physicochemical and Biological Annotations. In: Zara, J. & Sloup, J. (eds.) Central European Multimedia and Virtual Reality Conference (available in EG Eurographics Library). Prague: Czech Technical University (CTU), pp. 69-74.

Slide 1-3: Knowledge Discovery from Data

Wiltgen, M., Holzinger, A. & Tilz, G. P. (2007) Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins. In: *Lecture Notes in Computer Science (LNCS 4799). Berlin, Heidelberg, New York, Springer, 199-212.*

A. Holzinger 709.049 — 13/80 — Med Informatics L01

It is essential to make such structures visible to the domain experts, so that they can understand and gain out knowledge – for instance, it may lead to the discovery of new, unknown structures in order to modify drugs; the transformation of such information into knowledge is vital for prevention and treatment of diseases, consequently a contribution towards personalized medicine. This picture shows the 3D structure of the numbers seen in Slide 1-3. The Tumor necrosis factor TNF (upper part - which causes the death of a cell) is "interacting" with the receptor (lower part). The residues at the macromolecular interface are visualized in a "ball-and-stick" representation. The covalent bonds are represented as sticks between atoms. The atoms are the balls. The rest of the two chains are represented as ribbons. Residue names and numbers of the TNF receptor are labeled. The hydrogen bonds are represented by these yellow dotted lines (Wiltgen, Holzinger & Tilz, 2007).

References:

Wiltgen, M., Holzinger, A. & Tilz, G. P. 2007. Interactive Analysis and Visualization of Macromolecular Interfaces Between Proteins. In: Lecture Notes in Computer Science (LNCS 4799). Berlin, Heidelberg, New York: Springer, pp. 199-212.
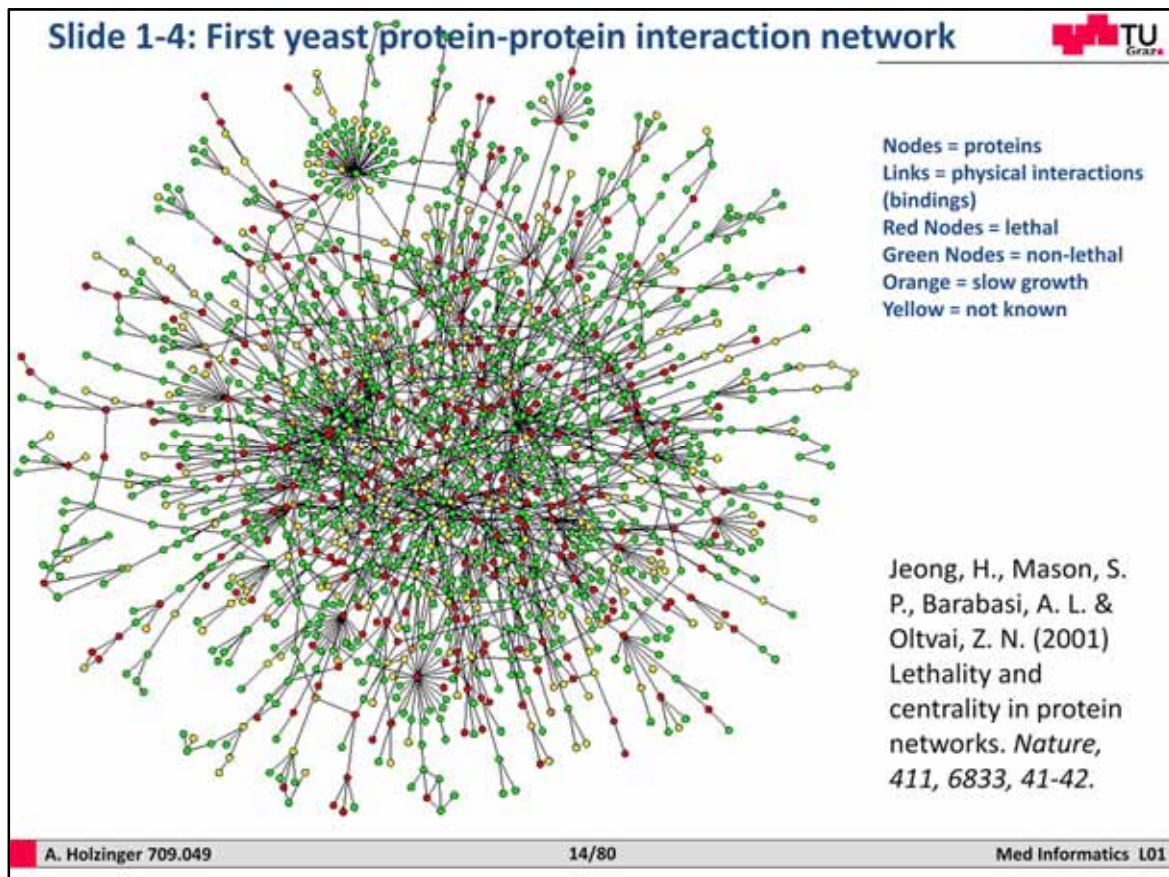
Good examples for data intensive, highly complex microscopic structures are yeast protein networks. Yeast is a eukaryotic micro-organism (fungus), consisting of single oval cells, which asexually reproduces by budding, capable of converting sugar into alcohol and carbon dioxide. There are 1,500 known species currently, estimated to be only 1% of all yeast species. Yeasts are unicellular, typically measuring 4 μm in diameter. In this slide we see the first protein-protein interaction (PPI) network (Jeong et al., 2001): Saccharomyces cerevisiae. It is perhaps the most useful yeast, used for brewing, winemaking, and baking since ancient times. This S. cerevisiae PPI network contains 1,870 proteins as nodes, connected by 2,240 identified direct physical interactions, and is derived from combined, non-overlapping data, obtained mostly by systematic two-hybrid analyses. The nodes are the proteins; the links between them are the physical interactions (bindings), red nodes are lethal, green nodes are non-lethal, orange nodes are slow growing and yellow are not yet known.

References:

Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. 2001. Lethality and centrality in protein networks. Nature, 411, (6833), 41-42.
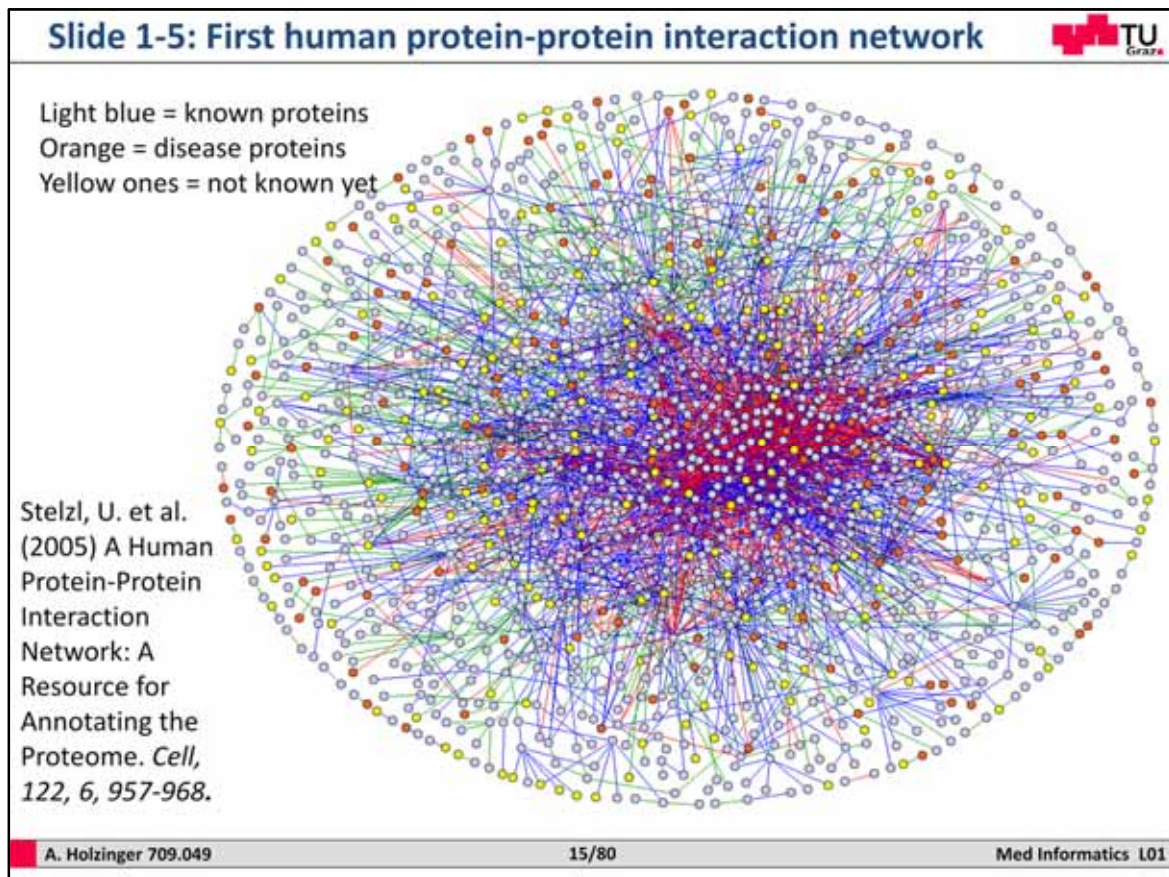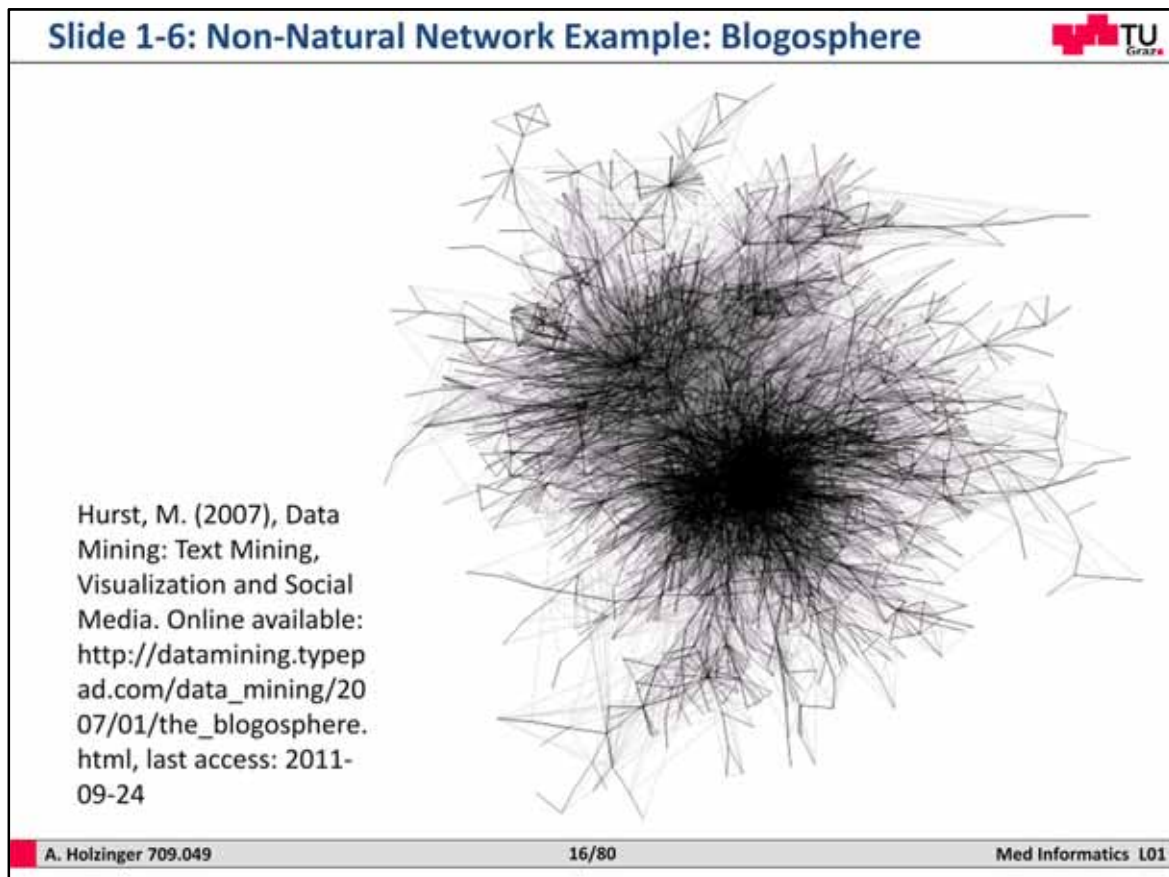
PPIs are essential for all biological processes, and compiling them provides insights into protein functions. Networks are relevant from a systems biology point of view, as they help to uncover the generic organization principles of functional cellular networks, when both spatial and temporal aspects of interactions are considered (Ge, Walhout & Vidal, 2003). In Slide 1-6 we see the first visualization of a human PPI (Stelzl et al., 2005). Most cellular processes rely on such networks (Barabási & Oltvai, 2004), and a breakdown of such networks is responsible for most human diseases (Barabási, Gulbahce & Loscalzo, 2011). **Light blue nodes are known proteins,** orange nodes are disease proteins, **yellow nodes are not known yet.** Contrast this image to Slide 1-2 and look at the similarities.

Remark: There is a nice Video available, called Powers-of-Ten, which demonstrate the dimensions of both worlds very good, it is online available via:
http://www.powersof10.com

The 1977 production by Charles and Ray Eames takes us on an adventure in magnitudes: Starting at a picnic by the lakeside in Chicago, this famous film transports us to the outer edges of the universe. Every ten seconds we view the starting point from ten times farther out until our own galaxy is visible only as a speck of light among many others. Returning - we move inward- into the hand of the picnicker - with ten times more magnification every ten seconds. Our journey ends inside a proton of a carbon atom within a DNA molecule in a white blood cell.
http://www.youtube.com/watch?v=0fKBhvDjuy0

A similar interactive visualization, focused on biological structures can be found here: http://learn.genetics.utah.edu/content/begin/cells/scale
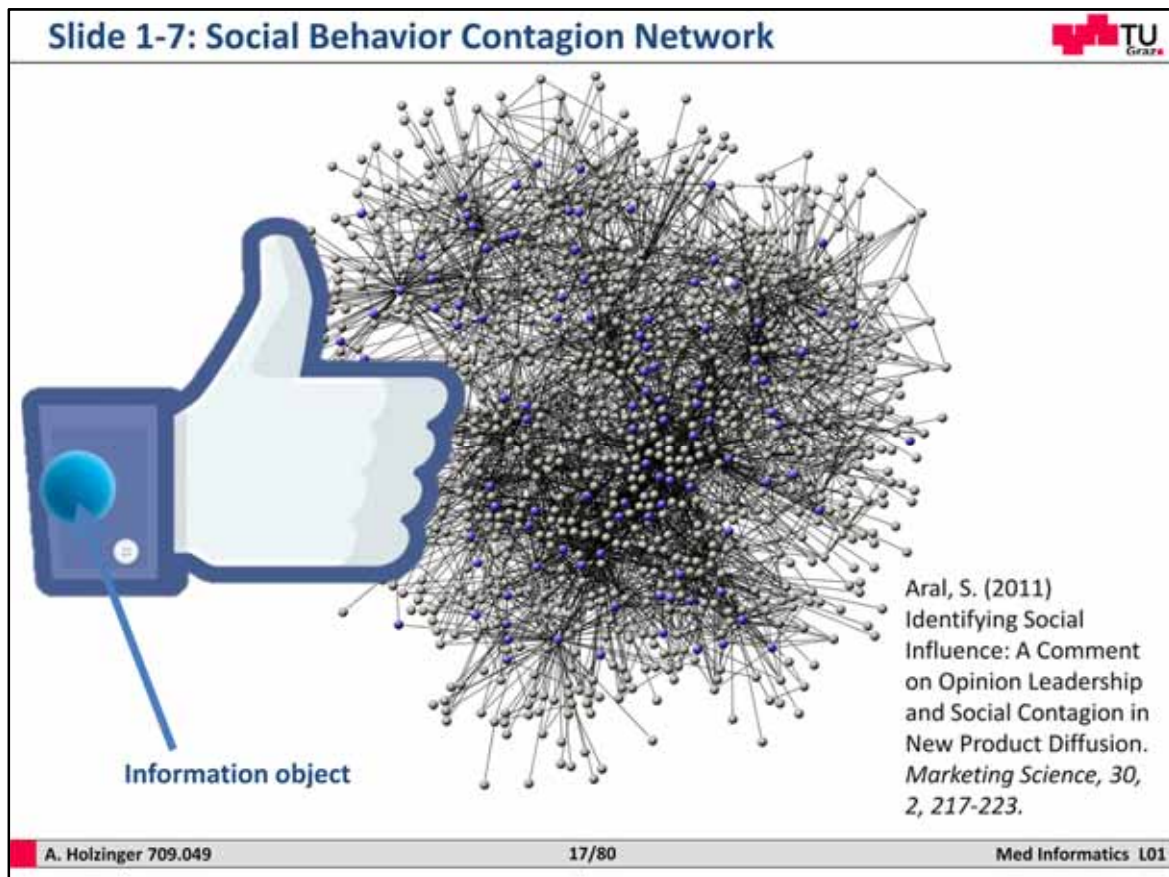
Network biology aims to understand the behavior of cellular networks; a parallel field within medicine is called network medicine and aims to uncover the role of such networks in human disease. To demonstrate the similarity to non-natural structures, we see in Slide 1-7 the visualization of the Blogosphere.

Hurst (2007) mapped this image as a result of six weeks of observation: densely populated areas represent the most active portions of the blogosphere. By showing only the links in the graph, we can get a far better look at the structure than if we include all the nodes. In this image, we are looking at the core of the Blogosphere: The dark edges show the reciprocal links (where A has cited B and vice versa), the lighter edges indicate a-reciprocal links. The larger, denser area of the graph is that part of the Blogosphere generally characterized by socio-political discussion and the periphery contains some topical groupings.

http://datamining.typepad.com/gallery/blogosphere-sketch.png
http://datamining.typepad.com/gallery/blog-map-gallery.html

Slide 1-7: Social Behavior Contagion Network

Information object

Aral, S. (2011) Identifying Social Influence: A Comment on Opinion Leadership and Social Contagion in New Product Diffusion. *Marketing Science, 30, 2, 217-223.*

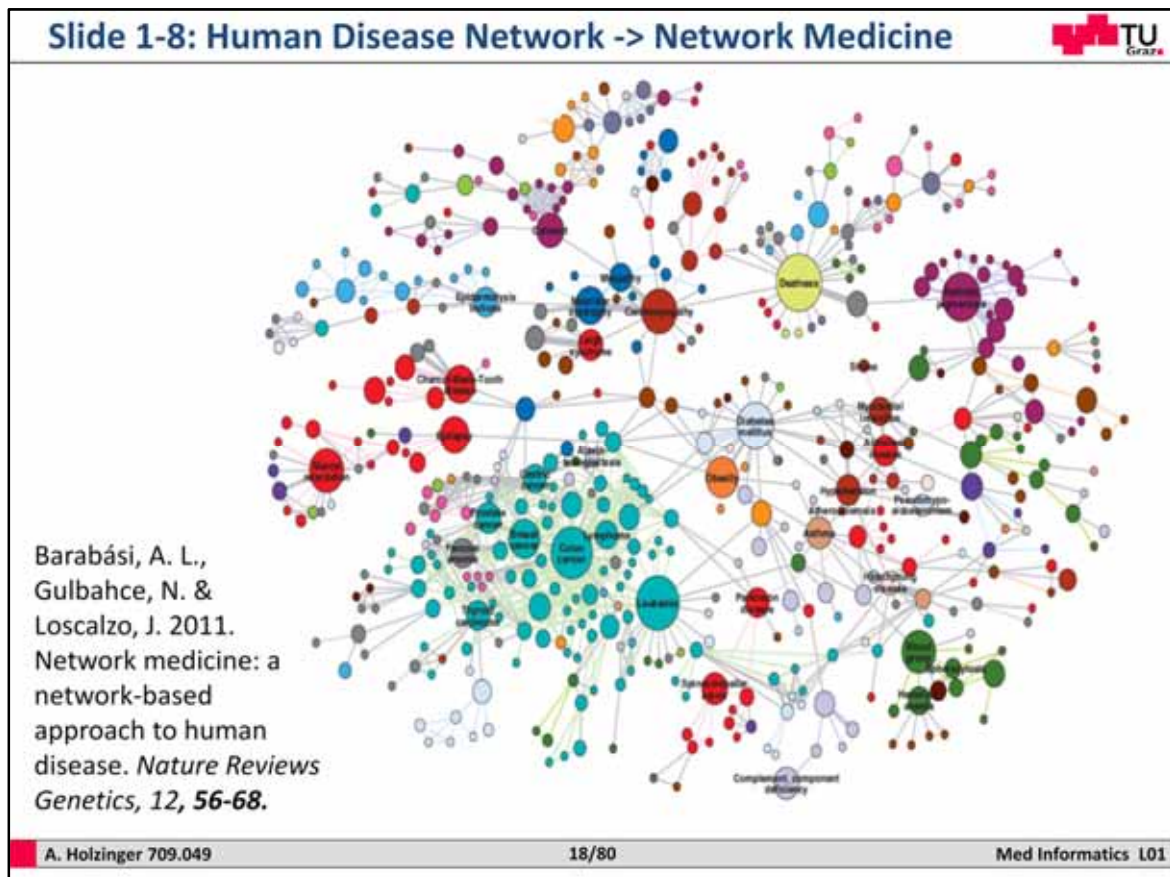A. Holzinger 709.049   17/80   Med Informatics L01

A final example in Slide 1-8 shows the principle of viral marketing. The idea is to spread indirect messages which in return suggest spreading them further. If you press the Like-button in Facebook, a similar process starts, similar to an epidemic in public health.

Aral (2011) calls this behavior contagion and it is of much importance for research to know how human behavior can spread. We can mine masses of social network data in order to gain knowledge about the contagion of information, which is of interest for the health area, in particular for public health. A current trend of research is Opinion Mining, where you analyze such data sets (Petz et al., 2012), (Petz et al., 2013).

References:

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Winkler, S., Schaller, S. & Holzinger, A. 2012. On Text Preprocessing for Opinion Mining Outside of Laboratory Environments. In: Lecture Notes in Computer Science, LNCS 7669. Berlin Heidelberg: Springer, pp. 618-629.

Petz, G., Karpowicz, M., Fürschuß, H., Auinger, A., Stříteský, V. & Holzinger, A. 2013. Opinion Mining on the Web 2.0 – Characteristics of User Generated Content and Their Impacts. In: Lecture Notes in Computer Science LNCS 7947. Heidelberg, Berlin: Springer, pp. 35-46.

Slide 1-8: Human Disease Network -> Network Medicine

Barabási, A. L., Gulbahce, N. & Loscalzo, J. 2011. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12, **56-68.**

A. Holzinger 709.049   18/80   Med Informatics L01

A disease is rarely a consequence of an abnormality in a single gene, but reflects the perturbations of the **complex intracellular network.**

The emerging tools of network medicine offer a platform to explore systematically the molecular complexity of a particular disease. This can lead to the identification of disease pathways, but also the molecular relationships between apparently distinct pathological phenotypes. Advances in this direction are essential to identify new disease-genes and to identify drug targets and biomarkers for complex diseases, see →Slide 1-9.

In this slide we see a **human disease network:** the nodes are diseases; two diseases are linked if they share one or several disease-associated genes. Not shown are small clusters of isolated diseases. The node color reflects the disease class of the corresponding diseases to which they belong, cancers appearing as blue nodes and neurological diseases as red nodes. The node size correlates with the number of genes known to be associated with the corresponding disease (Barabási, Gulbahce & Loscalzo, 2011).

Erwin Schrödinger (1887-1961), Nobel Prize in Physics in 1933. Schrödinger provided in 1943 a series of lectures entitled **"What Is Life? The Physical Aspect of the Living Cell and Mind"** (Schrödinger, 1944). He described some fundamental differences between animate and inanimate matter, and raised some hypotheses about **the nature and molecular structure of genes – ten years before the discoveries of Crick & Watson (1953).** The rules of life seemed to violate fundamental interactions between physical particles such as electrons and protons. It is as if the organic molecules in the cell have a kind of "knowledge" that they are living (Westra et al., 2007). It is both interesting and important to accept the fact that despite all external influences, this "machinery" is working now **for more than 3.8 billion years** (Schidlowski, 1988), (Mojzsis et al., 1996).

Remark: Just some "important figures":
1) Age of our Solar System = 4,5 Billion years (= 4,500,000,000 = 4,5 x 10^9)
2) Oldest Evidence of life = 3,8 Billion years
3) First larger animals on earth = 0,5 Billion years
4) First mammals on earth = 0,2 Billion years
5) Humans on earth = 0,002 Billion years (=2,000,000 years)
6) Modern humans = 0,0002 Billion years (= 200,000 years)
7) Oldest found human (Ötzi) = 5,300 years

A timeline of life's evolution can be found here:
http://exploringorigins.org/timeline.html
Reference:
Westra, R., Tuyls, K., Saeys, Y. & Nowé, A. 2007. Knowledge Discovery and Emergent Complexity in Bioinformatics. In: Tuyls, K., Westra, R., Saeys, Y. & Nowé, A. (eds.) Knowledge Discovery and Emergent Complexity in Bioinformatics. Springer Berlin Heidelberg, pp. 1-9.

Schrödinger's early ideas encouraged many scientists to investigate the molecular basis of life and he stated that information (negative entropy) is the abstract concept that quantifies the notion of this order amongst the building blocks of life.

Life is a fantastic interplay of matter, energy, and information, and essential functions of living beings correspond to the generation, consumption, processing, preservation and duplication for information. Scientists in Artificial Intelligence (AI) and Artificial Life (AL) are interested in understanding the properties of living organisms to build artificial systems that exhibit these properties for useful purposes. AI researchers are interested mostly in perception, cognition and generation of action, whereas AL focuses on evolution, reproduction, morphogenesis and metabolism (Brooks, 2001).

Reference:
Brooks, R. 2001. The relationship between matter and life. Nature, 409, (6818), 409-411.

Slide 1-10: Life is complex information

5 µm

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

A. Holzinger 709.049 | 21/80 | Med Informatics L01

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.

It is also very interesting that in contrast to a few laws that govern the interactions between the few really elementary physical particles, there are at least tens of thousands of different genes and proteins, with millions of possible interactions, and each of these interactions obeys its own peculiarities. Consequently, in the life sciences are different processes involved including transcription, translation and subsequent folding (Hunter, 2009). Advances in bioinformatics generate masses of biological data, increasing the discrepancy between what is observed and what is actually known about life's organization at the molecular level. Knowledge Discovery plays an important role for the understanding, for getting insights and for sensemaking of the masses of observed data (Holzinger, 2013).

References:

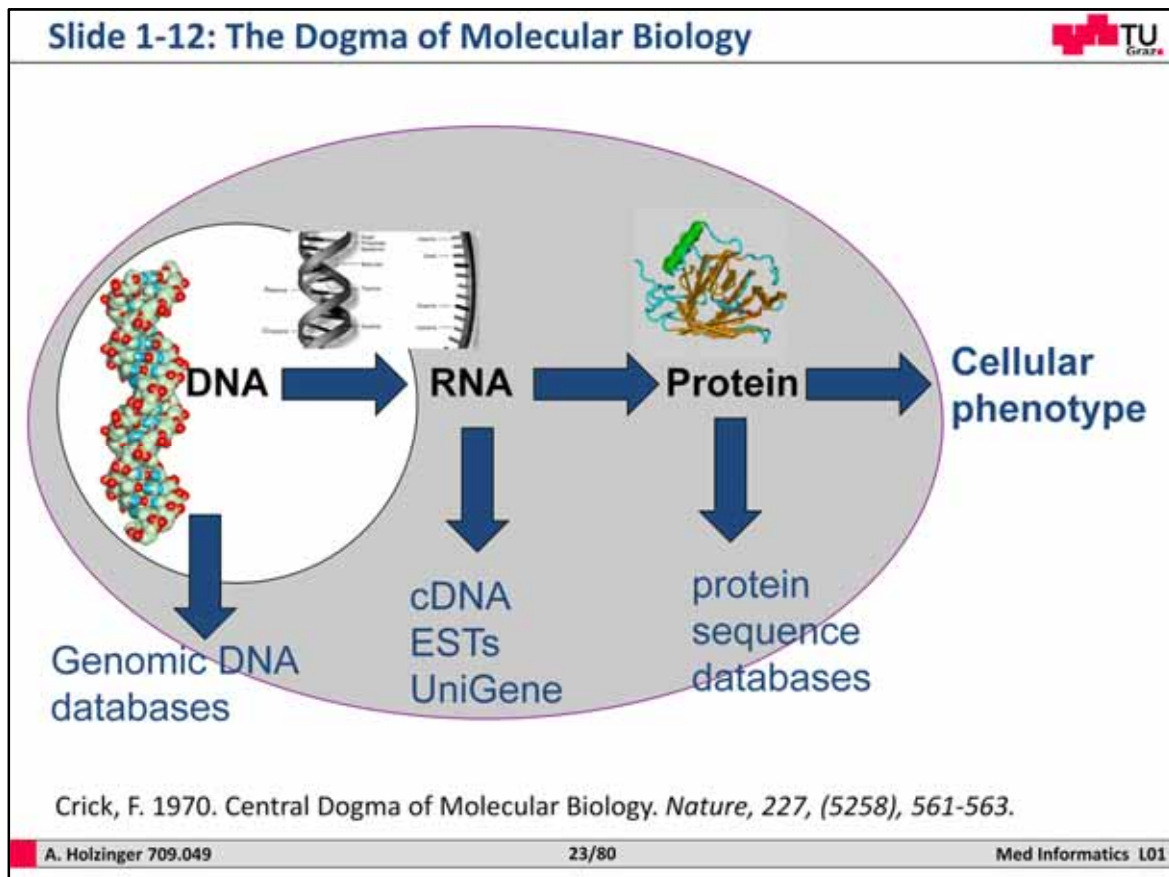Hunter, L. 2009. The Processes of Life: An introduction to molecular biology, Cambridge (MA), MIT Press.

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.
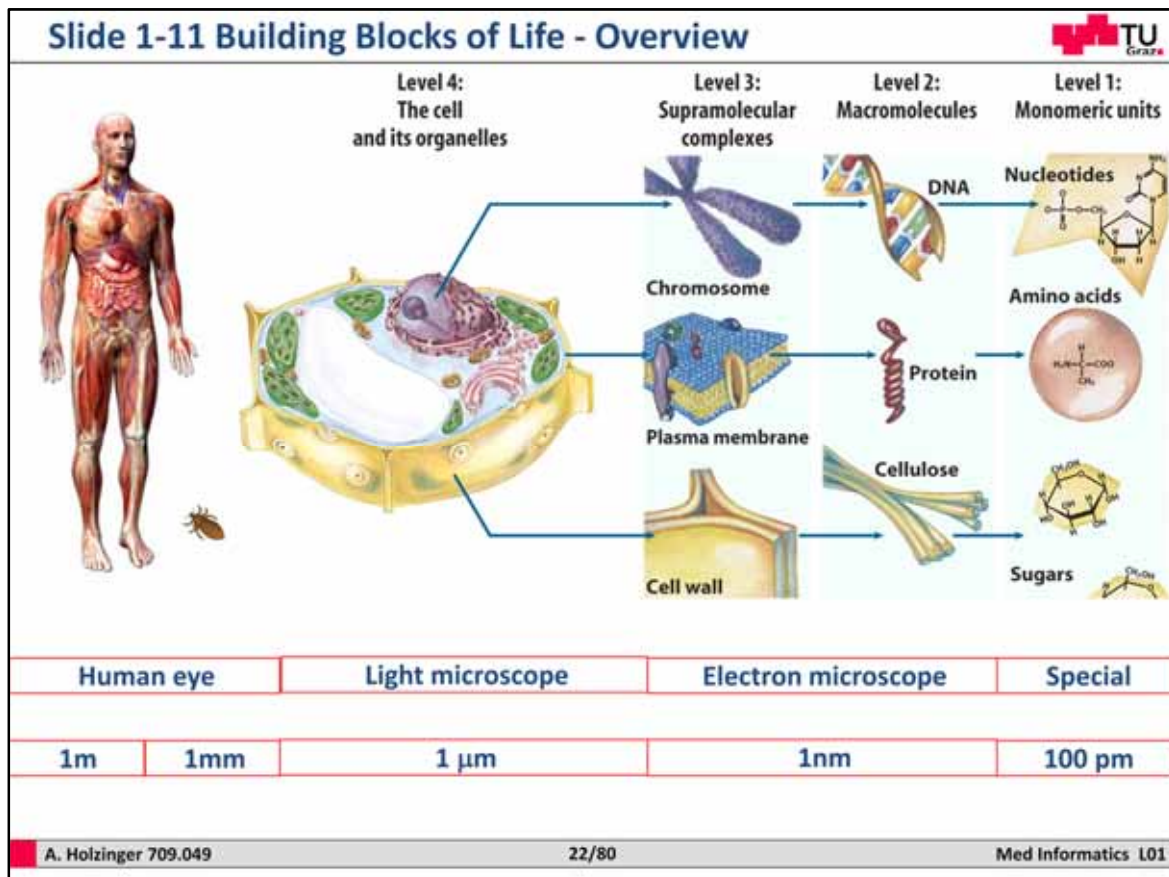
The human body is made up of trillions of cell (this is big data!). To get a "big picture" we have a first look on the whole body: The average 70 kg adult contains approximately $3 \cdot 10^{27}$ atoms and contains about 60 different chemical elements and is build up by approximately 10^{14} cells. The cell is the basic building block and consists of supramolecular complexes, chromosomes, plasma membrans etc., consisting of macromolecules (DNA, proteins, cellulose) and monomeric units, such as nucleotides, amino acid, etc. Very interesting is the large amount of human microbiome, which include microorganisms (bacteria, fungi and archaea) that resides on the surface and in deep layers of skin, in the saliva and oral mucosa, in the conjunctiva, and in the gastrointestinal tracts. Studies of the human microbiome have revealed that healthy individuals differ remarkably. Much of this diversity remains unexplained, although diet, environment, host genetics and early microbial exposure have all been implicated. Accordingly, to characterize the ecology of human-associated microbial communities, the Human Microbiome Project has analysed the largest cohort and set of distinct, clinically relevant body habitats so far (Mitreva, 2012). From the discovery of DNA to the sequencing of the human genome, the formation of biological molecules from gene to RNA and protein has been the central tenet of biology. Yet the origins of many diseases, including allergy, Alzheimer's disease, asthma, autism, diabetes, inflammatory bowel disease, multiple sclerosis, Parkinson's disease and rheumatoid arthritis, continue to evade our understanding (Marth, 2008).

References:
Mitreva, M. 2012. Structure, function and diversity of the healthy human microbiome. Nature, 486, 207-214.
Marth, J. D. 2008. A unified vision of the building blocks of life. Nat Cell Biol, 10, (9), 1015-1015.

Slide 1-12: The Dogma of Molecular Biology

Crick, F. 1970. Central Dogma of Molecular Biology. *Nature, 227, (5258), 561-563.*

A. Holzinger 709.049   23/80   Med Informatics L01

To understand the big picture in slide 1-12, we follow the central dogma of molecular biology, which states that DNA is transcribed into RNA and translated into protein (Crick, 1970): DNA → RNA → Protein → Cellular Phenotype

Similarly, there is a central dogma of genomics, which states (Pevsner, 2009): Genome → Transcriptome → Proteome → Cellular Phenotype

Three perspectives arise from those fundamentals:

1) The cell, 2) the organism, and 3) The tree of life (evolution).

The cell is the basic building block of all organisms and forms organs and tissue. Before we look at this fantastic building block of life, let us first look at the fundamental building blocks of the cell.
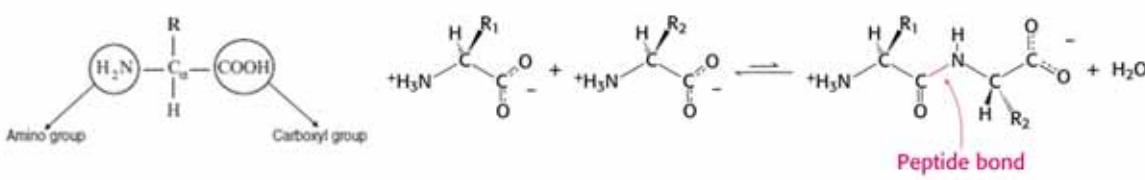
References:

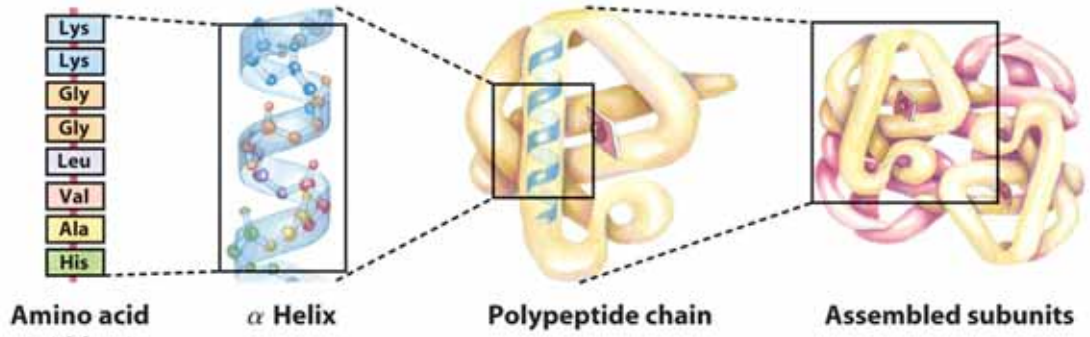Crick, F. 1970. Central Dogma of Molecular Biology. Nature, 227, (5258), 561-563.

Pevsner, J. 2009. Bioinformatics and functional genomics, Hoboken (NJ), John Wiley & Sons.

Note: Proteins are large biological molecules consisting of one or more chains of amino acids and they vary from one to another mainly in their sequence of amino acids, which is dictated by the nucleotide sequence of their genes, and which usually results in folding of the protein into a specific three-dimensional structure that determines its activity. Proteins perform a variety of functions and they regulate cellular and physiological activities. The functional properties of proteins depend on their three-dimensional structures. The native structure of a protein can be experimentally determined using X-ray crystallography, nuclear magnetic resonance (NMR) spectroscopy or electron microscopy. Over the past 40 years, the structures of 55,000+ proteins have been determined. On the other hand, the amino acid sequences are determined for more than eight million proteins. The specific sequence of amino acids in a polypeptide chain folds to generate compact domains with a particular three-dimensional structure. The polypeptide chain itself contains all the necessary information to specify its three-dimensional structure. Deciphering the three-dimensional structure of a protein from its amino acid sequence is a long-standing goal in molecular and computational biology (Gromiha, 2010).

Slide 1-13 Amino-acid > Protein-chain > Protein-structure

Peptide bond

Primary structure — Secondary structure — Tertiary structure — Quaternary structure

Lys / Lys / Gly / Gly / Leu / Val / Ala / His

Amino acid residues — α Helix — Polypeptide chain — Assembled subunits

Gromiha, M. 2010. *Protein Bioinformatics, Amsterdam, Elsevier.*

A. Holzinger 709.049   24/80   Med Informatics L01

From Amino-acid to Proteinstructures in seven steps:

1) **Aminoacid.** Protein sequences consist of 20 different amino acids serving as building blocks of proteins. Amino acids contain a central carbon atom, called Alpha Carbon (C_α) which is attached to a hydrogen atom (H), an amino group (NH_2), and a carboxyl group (COOH). The letter R indicates the presence of a side chain, which distinguishes each amino acid (Hunter, 2009).

2) **Protein chain.** Several amino acids form a protein chain, in which the amino group of the first amino acid and the carboxyl group of the last amino acid remain intact, and the chain is said to extend from the amino (N) to the carboxyl (C) terminus. This chain of amino acids is called a polypeptide chain, main chain, or backbone. Amino acids in a polypeptide chain lack a hydrogen atom at the amino terminal and an OH group at the carboxyl terminal (except at the ends), and hence amino acids are also called amino acid residues (simply residues). Nature selects the combination of amino acid residues to form polypeptide chains for their function, similar to the combination of alphabets to form meaningful words and sentences. These polypeptide chains that have specific functions are called proteins.

3) **Protein structure.** Depending on their complexity, protein molecules may be described by four levels of structure: primary, secondary, tertiary, and quaternary. Because of the advancements in the understanding of protein structures, two additional levels such as super-secondary and domain have been proposed between secondary and tertiary structures. A stable clustering of several elements of secondary structures is referred to as a super-secondary structure. A somewhat higher level of structure is the domain, which refers to a compact region and distinct structural unit within a large polypeptide chain.

4) **3a) Primary Structure.** The linear sequence of amino acid residues in a protein is described by the primary structure: It includes all the covalent bonds between amino acids. The relative spatial arrangement of the linked amino acids is unspecified.

5) **3b) Secondary Structure.** Regular, recurring arrangements in space of adjacent amino acid residues in a polypeptide chain are described by the secondary structure. It is maintained by hydrogen bonds (see the dotted lines in Slide 1-4) between amide hydrogens and carbonyl oxygens of the peptide backbone. The main secondary structures are α-helices and β-folding-structures (β-sheets). The polypeptide backbone is tightly coiled around the long axis of the molecule, and R groups of the amino acid residues protrude outward from the helical backbone.

6) **3c) Tertiary structure.** It refers to the spatial relationship among all amino acids in a polypeptide; it is the complete three-dimensional structure of the polypeptide with atomic details. Tertiary structures are stabilized by interactions of side chains of non-neighboring amino acid residues and primarily by non-covalent interactions.

7) **3d) Quaternary structure.** This refers to the spatial relationship of the polypeptides or subunits within a protein and is the association of two or more polypeptide chains into a multi-subunit or oligomeric protein. The polypeptide chains of an oligomeric protein may be identical or different. The quaternary structure also includes the cofactor and other metals, which form the catalytic unit and functional proteins (Gromiha, 2010).

Reference: Gromiha, M. M. 2010. Protein Bioinformatics, Amsterdam, Elsevier.

Slide 1-14 Tertiary Structure of a Protein

Shehu, A. & Kavraki, L. E. 2012. Modeling structures and motions of loops in protein molecules. *Entropy, 14, (2), 252-290.*

Alpha Helix — Antiparallel beta sheet — Beta strand

Tertiary structure

Domain

A. Holzinger 709.049   25/80   Med Informatics L01

All biological mechanisms in the living cell involve protein molecules, consequently proteins are central components of cellular organization and function. The mechanistic view that the structure (the "shape" in the ribbon diagrams, see Slide 1-15), manages the biological function in proteins has been confirmed in wet laboratory experiments. A unique set of atoms compose a protein molecule and determines to a great extent the spatial arrangement by these atoms for biological function. The state in which a protein carries its biological activity is called the protein native state. Microscopically, this macro-state is an ensemble of native conformations also referred to as the native state ensemble. Proteins fold from a highly disordered state into a highly ordered one. The **folding problem** has been stated as **predicting the tertiary structure from sequential information.** The ensemble of unfolded forms may not be as disordered as believed, and the native form of many proteins may not be described by a single conformation, but rather an ensemble. For the quantification of the relative disorder in the folded and unfolded ensembles entropy (see →Lecture 2) measures are suitable. The tertiary structure of a protein is basically the overall, unique, three dimensional folding of a protein. In a protein folding diagram (ribbon diagram) we can recognize the beta pleated sheets (ribbons with arrows) and the alpha helical regions (barrel shaped structures).

References:

Anfinsen, C. B. 1973. Principles that Govern the Folding of Protein Chains. Science, 181, (4096), 223-230.

Dill, K. A., Bromberg, S., Yue, K., Chan, H. S., Ftebig, K. M., Yee, D. P. & Thomas, P. D. 1995. Principles of protein folding—a perspective from simple exact models. Protein Science, 4, (4), 561-602.

Slide 1-15 Protein Analytics

Rabilloud, et al. 2010. Two-dimensional gel electrophoresis in proteomics: past, present and future. *Journal of proteomics*, 73, (11), 2064-2077.

Xiao, W. Z. & Oefner, P. J. 2001. Denaturing high-performance liquid chromatography: A review. *Human Mutation*, 17, (6), 439-474.

A. Holzinger 709.049   26/80   Med Informatics L01

For Protein analytics a lot of different methods are known, for example:

1) **X-ray crystallography** is the primary method and used for determining the atomic and molecular structures with the output of a three-dimensional picture of the density of electrons within a crystal structure. From this electron density, the mean positions of the atoms (see in Slide 1-3) in the crystal can be determined, as well as their chemical bonds, their disorder and various other information (Schotte et al., 2003).

2) **Gel-Electrophoresis** (2D-Gel-Electrophoresis) had a major impact on the developments of proteomics (Southern, 1975), although it is no longer the exclusive tool (See Mass-Spectrometry). Electrophoresis separates molecules according to their charge (mass ratio) and 2D-electrophoresis separates molecules according to both their charge and their mass. The possibility of analyzing spots of interest coming from 2D gels was the real start of proteomics. At those times where no complete genome was published yet, such techniques provided enough information to look for homologs, or to devise oligonucleotides for screening DNA libraries (Rabilloud et al., 2010).

3) **Chromatography** is the collective term for the separation of mixtures, typ. Gas chromatography (GC) and Liquid chromatography (LC), the latter specifically a separation of proteins by size, i.e. gel filtration chromatography (Xiao & Oefner, 2001). Gas chromatography is an excellent separation technique and detects the ions and generates mass spectrum for each analyte, and this structural information aids in its identification. The synergistic coupling of GC and MS (see below) renders the tandem technique a major analytical workhorse in metabolomics (Yip & Yong Chan, 2013).

4) **Mass spectrometry (MS)** involve either detection of intact proteins, referred to as top-down proteomics, or identification of protein cleavage products, referred to as bottom-up or shotgun proteomics. MS-based proteomics is important for molecular, cellular biology and systems biology. These include the study of PPI via affinity-based isolations on a small and proteome-wide scale, the mapping of numerous organelles, and the generation of quantitative protein profiles from diverse species. The ability of mass spectrometry to identify and, increasingly, to precisely quantify thousands of proteins from complex samples can be expected to impact broadly on biology and medicine (Aebersold & Mann, 2003). **Top-down proteomic** strategies retain a lot of information about protein sequence (protein isoforms). Recent advances in top-down proteomics allow for identification of hundreds of intact proteins in yeast and mammalian cells, however, clinical applications of top-down proteomics are still limited. **Bottom-up proteomic** approaches suffer from a loss of information about protein isoforms and post-translational modification (PTM), especially for low-abundance proteins. Bottom-up proteomics greatly benefits from superior Liquid Chromatography (LC) separation of peptides prior to mass spectrometry, requires lower amounts of material, and provides better peptide fragmentation and higher sensitivity. Due to the very high number of routine protein identifications in biological samples, bottom-up proteomics remains the platform of choice **for biomarker discovery.** (Drabovich et al., 2013).

5) **Nuclear magnetic resonance spectroscopy (NMR)** is a research technique that exploits the magnetic properties of certain atomic nuclei and determines the physical and chemical properties of atoms or the molecules in which they are contained. A goal is to obtain 3-dimensional structures of the protein, similar to what can be achieved by X-ray crystallography. In contrast to X-ray crystallography, NMR spectroscopy is usually limited to proteins smaller than 35 kDa (Note: Dalton is the standard unit used for indicating the atomic mass: 1 Da = 1 g/mol = $1{,}6 \cdot 10^{-27}$ kg) although larger structures have been solved. **NMR spectroscopy is often the only way to obtain high resolution information on partially or wholly intrinsically unstructured proteins.**

Slide 1-16: Comparison of some current Methods

Technology	Sensitivity	Subcellular resolution	Cellular resolution	Minimally invasive?	Live cells?	Real time?
Genetically encoded nanosensors	Nanomolar to millimolar	Nanometer to millimeter	Yes	Yes	Yes	Yes
MRI	Mid-micromolar to millimolar (213)	No	Yes	Yes	Yes	Yes
PET	1–40 Bq mm^{-2} (18)	No	No	No	Yes	Yes
X-ray synchrotron	<1 mg kg^{-1} tissue (transit metals) (204)	No	Yes	No	No	No
SIMS	<1 fmol (67)	Yes	Yes	No	No	No
MALDI or TOF imaging	<1 ppm	Yes	50–300 μm (MALDI) 1–2 μm (TOF)	No	No	No
NIMS imaging	Yoctomolar (85)	No	50–300 μm	No	No	No
Mass spectrometry	Yoctomolar	No?	Yes	No	No	No
Raman	50 μM (70)	Yes	Yes	Yes	Yes	Yes

Okumoto, S., Jones, A. & Frommer, W. B. 2012. Quantitative imaging with fluorescent biosensors. *Annual review of plant biology, 63, 663-706.*

This table shows some current Methods and some properties e.g. whether they are minimally invasive, usable on live cells, in real time etc. (Okumoto, Jones & Frommer, 2012).
Abbreviations: Bq = Becquerel; MALDI = matrix-assisted laser desorption ionization; MRI = magnetic resonance imaging; NIMS = nanostructure initiator mass spectrometry; PET = positron emission tomography; SIMS = secondary ion mass spectrometry; TOF = time-of-flight mass spectrometry.

Slide 1-17 Enzymes

Klibanov, A. M. 2001. Improving enzymes by using them in organic solvents. *Nature, 409, (6817), 241-246.*

A. Holzinger 709.049 — 28/80 — Med Informatics L01

Enzymes are large biological molecules responsible for the thousands of chemical inter-conversions that sustain life. They are highly selective catalysts, greatly accelerating both the rate and specificity of metabolic reactions, from the digestion of food to the synthesis of DNA. Most enzymes are proteins, although some catalytic RNA molecules have been identified.

Enzymes adopt a specific three-dimensional structure, and may employ organic (e.g. biotin) and inorganic (e.g. magnesium ion) cofactors to assist catalysis.

The tremendous potential of enzymes as practical catalysts is well recognized. The scope of industrial bioconversions, especially for the production of specialty chemicals and polymers, is necessarily limited by a variety of considerations. Most such compounds are insoluble in water, and water frequently gives rise to unwanted side reactions and degrades common organic reagents. The thermodynamic equilibrium of many processes are unfavorable in water, and product recovery is sometimes difficult from this medium (Klibanov, 2001).

Slide 1-18 DNA-RNA-Proteins

The DNA, the RNA and the proteins are the three major macromolecules essential for all known forms of life.

Manca, V. (2013). Infobiotics. Springer.

A. Holzinger 709.049   29/80   Med Informatics L01

The Genome (Chromosomes) consists of Deoxyribonucleic acid (DNA), which is a molecule that encodes the genetic instructions used in the development and functioning of all known living organisms and many viruses.

The DNA, the RNA and the proteins are the three major macromolecules essential for all known forms of life.

The Ribonucleic acid (RNA) is a molecule that performs the coding, decoding, regulation, and expression of genes. The RNA is assembled as a chain of nucleotides, but is usually single-stranded. Cellular organisms use messenger RNA (mRNA) to convey genetic information (often notated using the letters G, A, U, and C for the nucleotides guanine, adenine, uracil and cytosine) that directs synthesis of specific proteins, while many viruses encode their genetic information using an RNA genome.

Slide 1-19 DNA & RNA – and the five principal bases

All nucleotides have a common structure
(pentose sugar phosphate + base)

The five principal nucleobases are nitrogen-containing biological compounds cytosine (in DNA and RNA), guanine (in DNA and RNA), adenine (in DNA and RNA), thymine (only in DNA) and uracil (only in RNA), abbreviated as C, G, A, T, and U, respectively. They are usually simply called **bases** in genetics.

Because A, G, C, and T appear in the DNA, these molecules are called **DNA-bases;**
A, G, C, and U are called **RNA-bases.**

Slide 1-20: Nobel Prize 1957
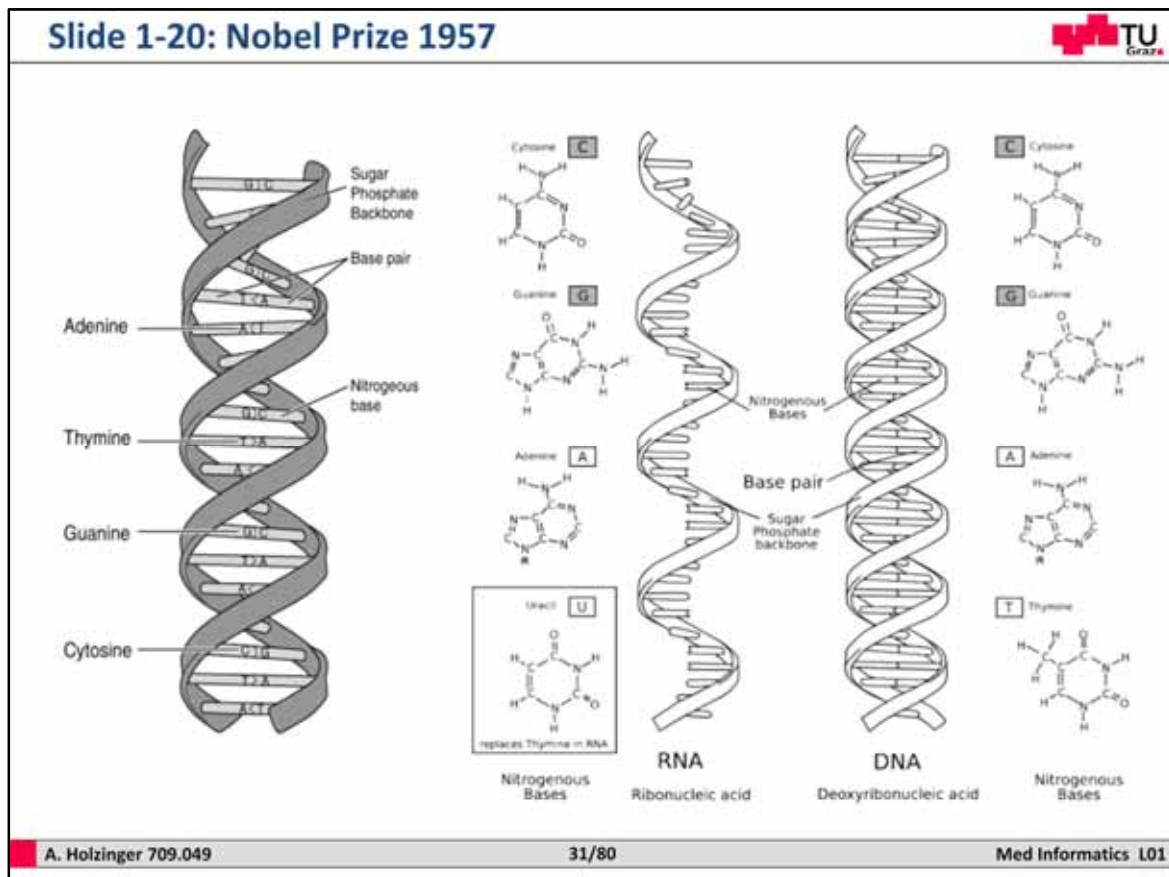
The 1962 Nobel Prize was given to Crick, Watson and Wilkins "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material" – they described in (Crick & Watson, 1953) the famous double helix structure of the DNA; Some further landmarks in DNA (Trent, 2012): 1972 = Recombinant DNA technologies allow DNA to be cloned; 1977 = Sequencing methods (Sanger & Gilbert awarded Nobel Prize); 1995 = First complete bacterial sequence described for *H. influenzae;* 2007 = Complete human diploid genome sequences publicly announced; 2009 = First human genome sequence using single sequencing technique; 2010 = First publication of a human metagenome – an additional layer of complexity for bioinformatics.

Some definitions:

Genetics concerns the process of trait inheritance from parents to progenies, including the molecular structure and function of genes, gene behavior in the context of a cell or organism, gene distribution, and variation and change in populations. Given that genes are universal to living organisms, genetics can be applied to the study of all living systems; including bacteria, plants, animals. The observation that living things inherit traits from their parents has been used since prehistoric times to improve crop plants and animals through selective breeding. The modern science of genetics, seeking to understand this process, began with the work of Gregor Mendel (1822-1884). Genetics today is focused on DNA (Brown, 2012).

Genes are molecular units of heredity of a living organism. It is widely accepted by the scientific community as a name given to some stretches of DNA and RNA that code for a polypeptide or for an RNA chain. Living beings depend on genes, as they specify all proteins and functional RNA chains. Genes hold the information to build and maintain an organism's cells and pass genetic traits to offspring. The Human Genome Project has revealed that there are about 20,000-25,000 haploid protein coding genes. The completed human sequence can now identify their locations. But only about 1.5% of the genome codes for proteins, while the rest consists of non-coding RNA genes, regulatory sequences, introns, and noncoding DNA (junk DNA). Surprisingly, the number of human genes seems to be less than a factor of two greater than that of many much simpler organism (earthworm, fruit fly). An example on how informatics can help in this field can be found in (Wassertheurer et al., 2003), (Holzinger et al., 2008).
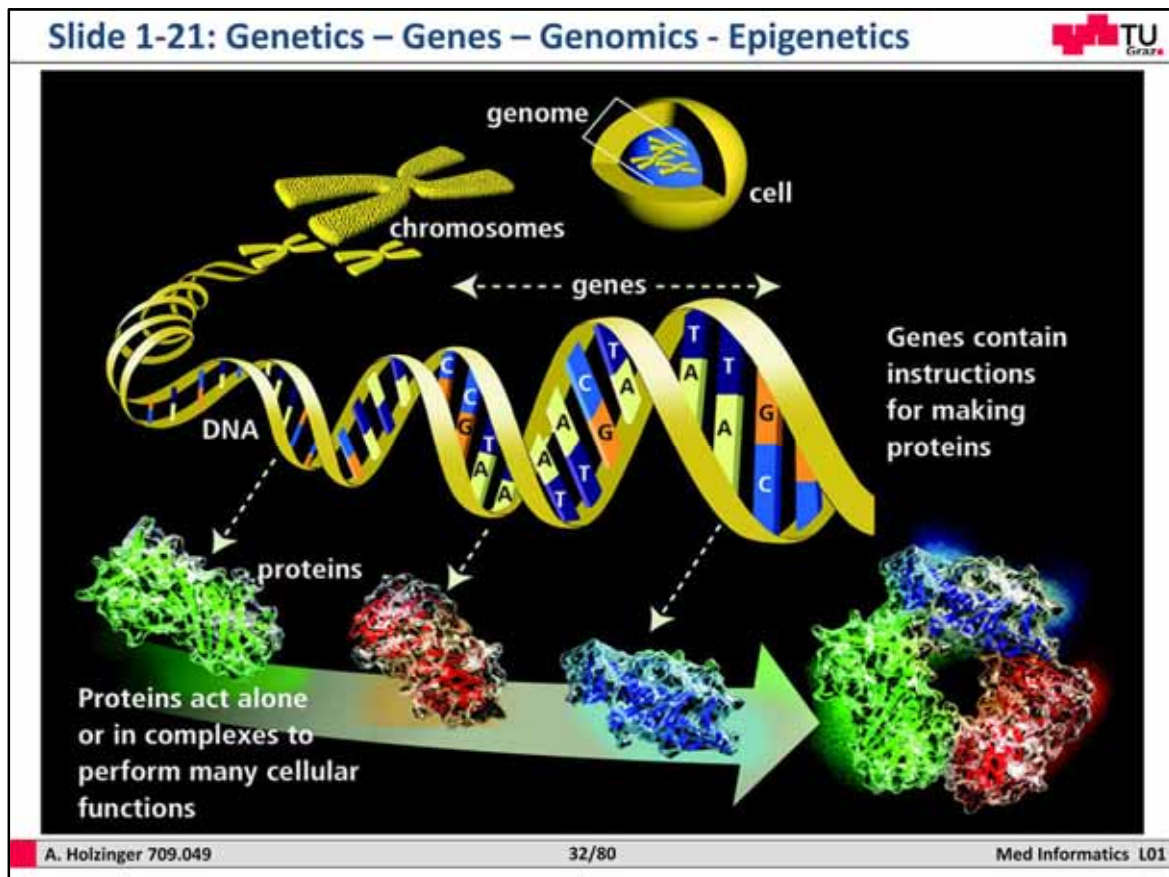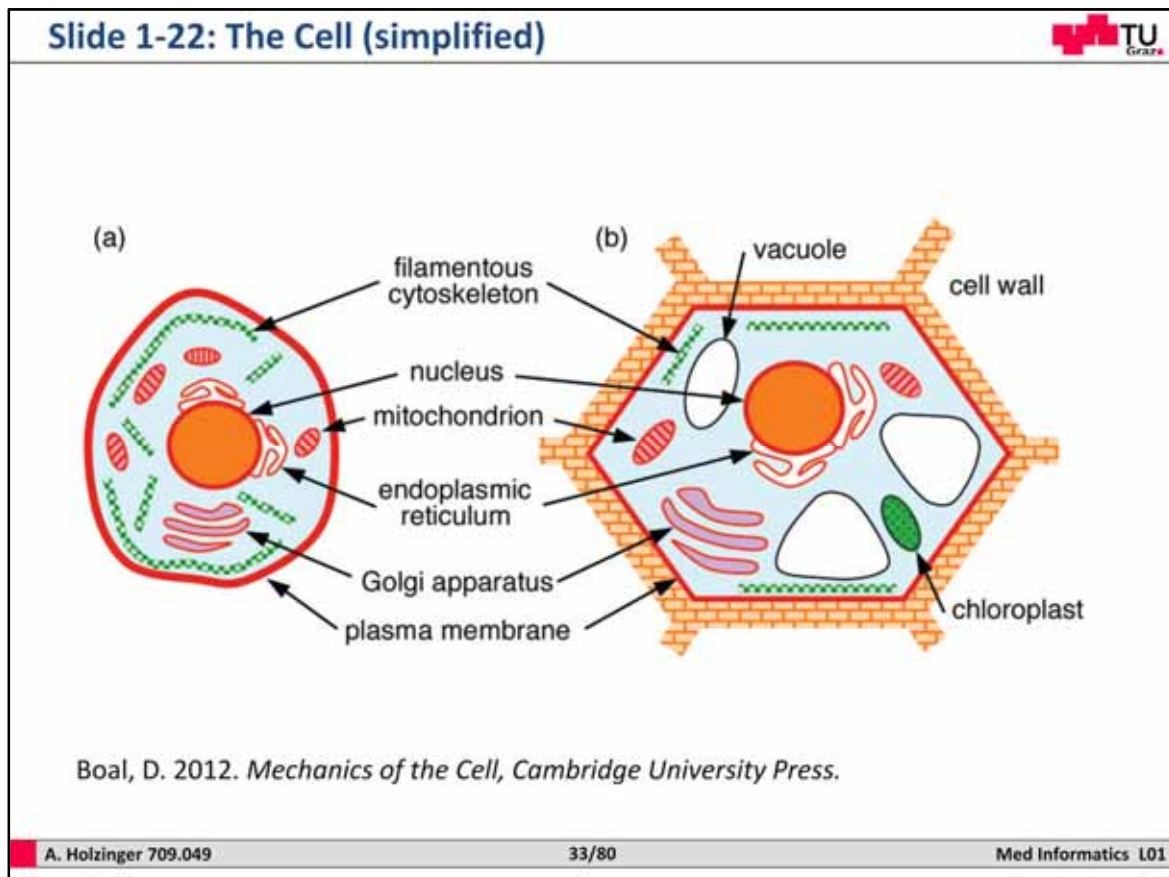
Genomics is a discipline in genetics that applies recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and the structure of genomes (the complete set of DNA within a single cell). It includes topics such as heterosis, epistasis, pleiotropy and other interactions between loci and alleles within the genome (Pevsner, 2009), (Trent, 2012).

Epigenetics is the study of changes in gene expression or cellular phenotypes, caused by mechanisms other than changes in the underlying DNA sequence, therefore it is called: epi-(Greek: επί- over, above, outer) -genetics. Some of these changes have been shown to be heritable. Example: Looking beyond DNA-associated molecules, prions (infectious proteins) are clearly epigenetic, perpetuating themselves through altered folding states. These states can act as sensors of environmental stress and, through the phenotypic changes they promote, potentially drive evolution (Trygve, 2011), (Kiberstis, 2012).

References:

Brown, T. 2012. Introduction to genetics: a molecular approach, New York, Garland.

Wassertheurer, S., Holzinger, A., Emberger, W. & Breitenecker, F. 2003. Medical education via WWW on continuous physiological models. 4th IMACS Symposium on Mathematical Modelling. Vienna University of Technology (Austria). 308–314.

Pevsner, J. 2009. Bioinformatics and functional genomics, Hoboken (NJ), John Wiley & Sons.

Holzinger, A., Emberger, W., Wassertheurer, S. & Neal, L. 2008. Design, Development and Evaluation of Online Interactive Simulation Software for Learning Human Genetics. Elektrotechnik & Informationstechnik (e&i), 125, (5), 190-196.

Trent, R. J. 2012. Molecular Medicine: Genomics to Personalized Healthcare. 4th Edition, Amsterdam, Elsevier.

Trygve, T. 2011. Handbook of Epigenetics, San Diego, Academic Press.

Kiberstis, P. A. 2012. All Eyes on Epigenetics. Science, 335, (6069), 637.

Source: http://www.broadinstitute.org/files/news/stories/full/genome-glossary.jpg

Slide 1-22: The Cell (simplified)

Boal, D. 2012. *Mechanics of the Cell, Cambridge University Press.*

A. Holzinger 709.049   33/80   Med Informatics L01

Cells are the fundamental structural, functional and physiological units of organisms, and they are the smallest unit of life that is classified as a "living thing", and therefore are called the "building blocks of life".

There are two types of cells: eukaryotes, which contain a nucleus; and prokaryotes, without nucleus. Prokaryotic cells are usually single-celled organisms, while eukaryotic cells can be either single-celled or part of multicellular organisms.

Cells consist of protoplasm enclosed within a cell membrane, which contains biomolecules such as proteins and nucleic acids. Organisms can be classified as unicellular (consisting of only one single cell, e.g. bacteria) or multicellular (such as animals (=human) and plants). While the number of cells in plants and animals varies from species to species, humans contain about 100 trillion ($[\![10]\!]^{14}$) cells. A simplified view on the cell let us recognize the basic physiological elements and we also see the similarity between animal cells and plant cells, for details please refer to one of the standard textbooks, e.g. (Boal, 2012). An example how informatics may help in cell biology can be found in (Jeanquartier & Holzinger, 2013).
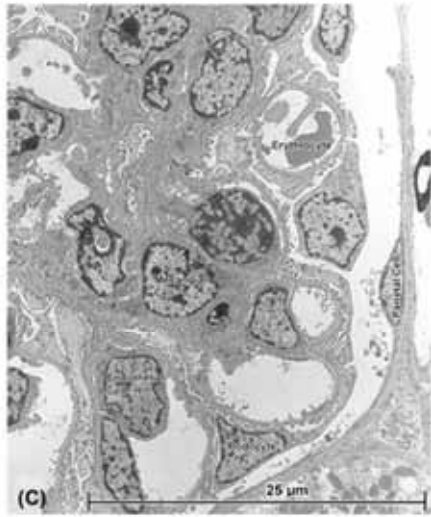
References:
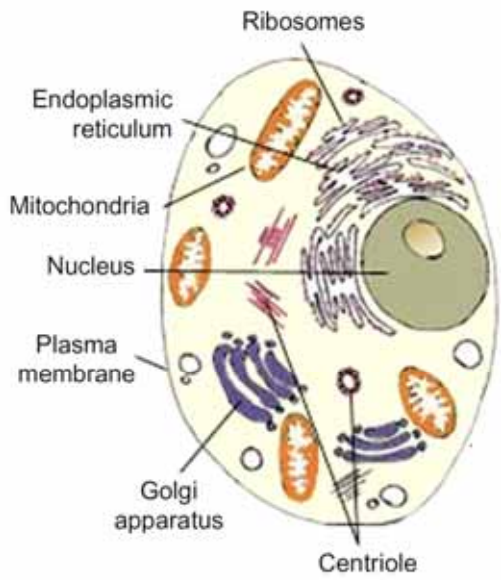
Boal, D. 2012. Mechanics of the Cell, Cambridge University Press.
Jeanquartier, F. & Holzinger, A. 2013. On Visual Analytics And Evaluation In Cell Physiology: A Case Study. In: Cuzzocrea, A., Kittl, C., Simos, D. E., Weippl, E. & Xu, L. (eds.) Multidisciplinary Research and Practice for Information Systems, Springer Lecture Notes in Computer Science LNCS 8127. Heidelberg, Berlin: Springer, pp. 495-502.

Just to get a feeling of the size of the cells, here a look at a microscopic image, the bar representing 25 μm and we are able to identify examples of the various cell types found in glomeruli, which is a part of the renal corpuscle in the kidney, part of the urinary (= renal) organ system (Sperelakis, 2012).

The about 100 trillion cells make up the human body, which is the entire structure of a human organism. The anatomy describes the body plan and the physiology is the study of how the body works. So let us now move from the basic building blocks to the organ systems.
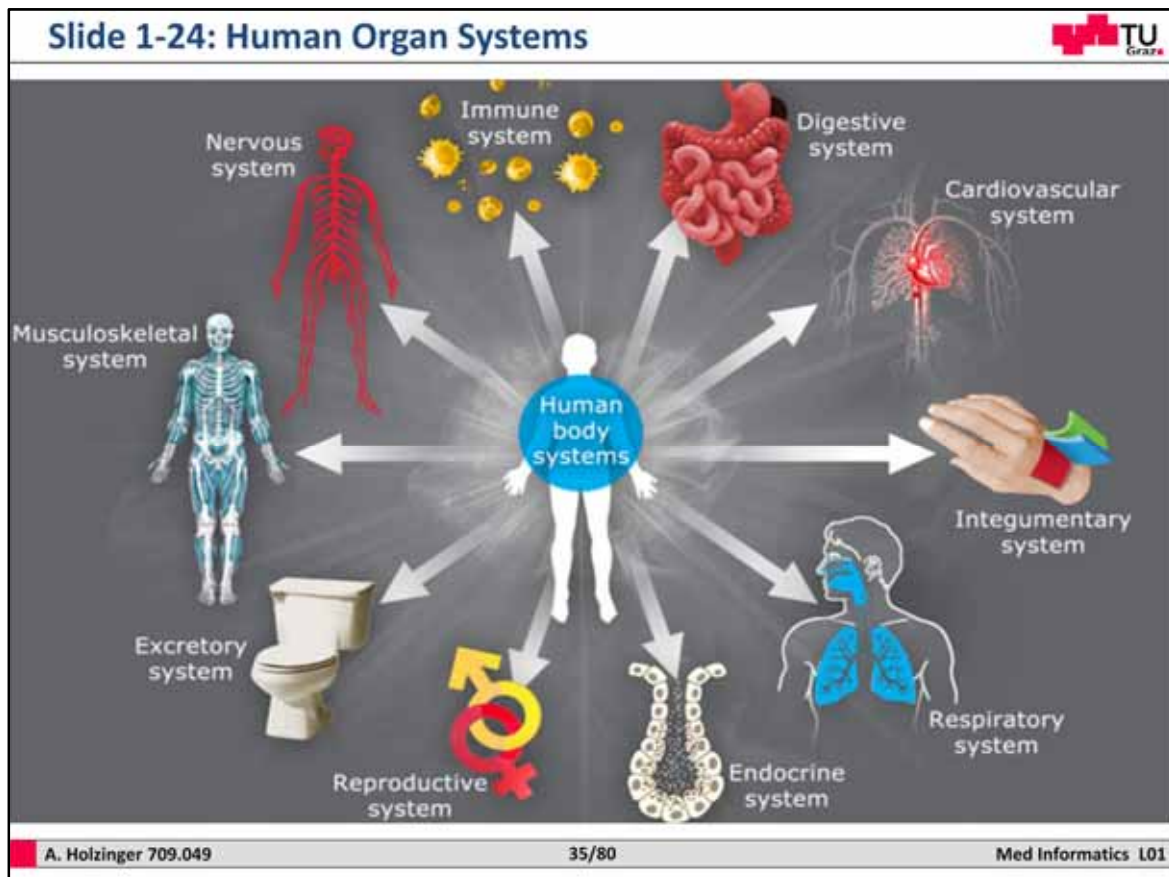
The major systems of the human body include:

Cardiovascular system: the blood circulation with the heart, the arteries and veins;

Respiratory system: the lungs and the trachea;

Endocrine system: communication within the body using hormones;

Digestive system: esophagus, stomach and intestines;

Urinary and excretory system: eliminating wastes from the body;

Immune and Lymph-system: defending against disease-causing agents;

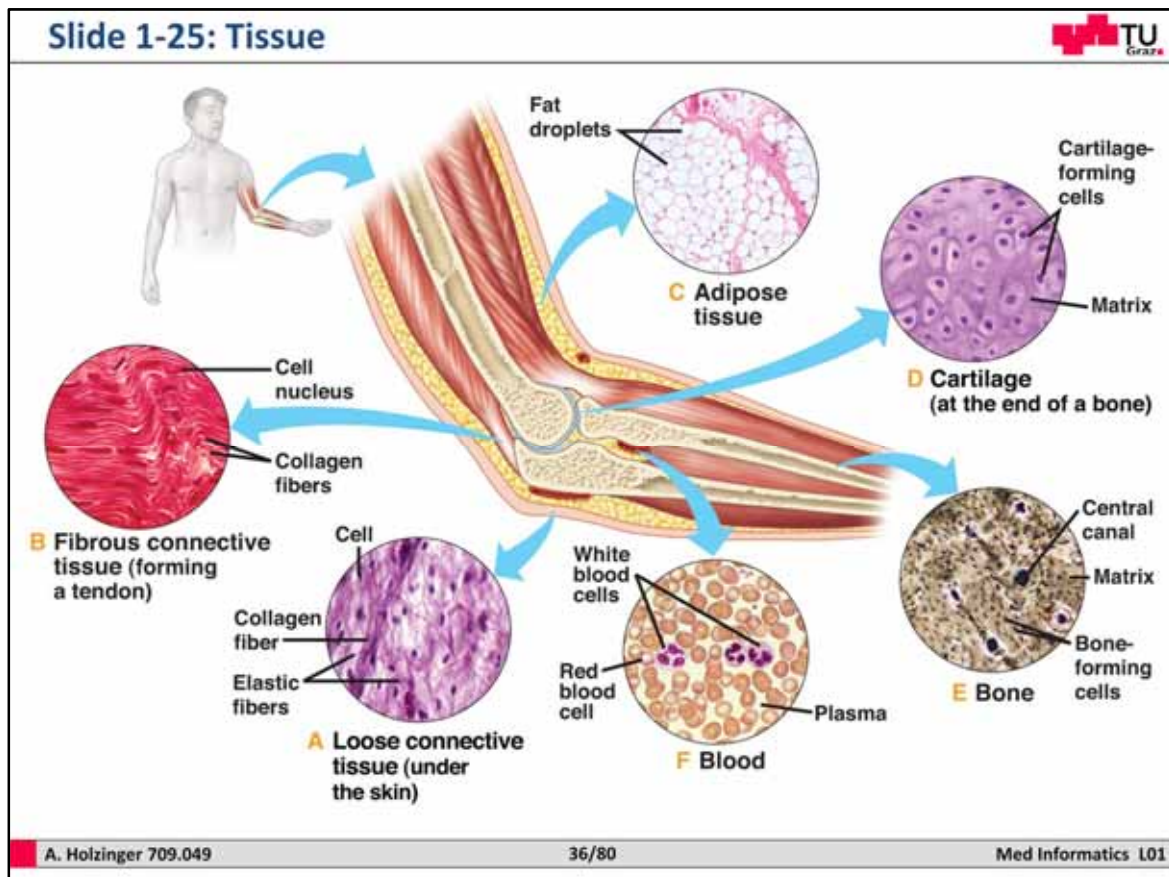Muscosceletal system: stability and moving the body with muscles

Nervous system: collecting, transferring and processing information;

Reproductive system: the sex organs.

For a detailed view on this topics please refer to a classical text book, e.g. the Color Atlas and Textbook of Human Anatomy: three volumes, authored by Werner Kahle und Michael Frotscher (2010, 6th Edition, Stuttgart: Thieme) or "the Sobotta", the Atlas of Human Anatomy: Musculoskelatal system, internal organs, head, neck, neuroanatomy (with Access to www.e-sobotta.com), authored by Friedrich Paulsen and Jens Waschke (2011, 15th Edition, Amsterdam: Elsevier).

Source:
http://www.sciencelearn.org.nz/Science-Stories/Our-Senses/Sci-Media/Images/The-body-s-systems

Tissue is a cellular organization between cells and a complete organism, it can be an ensemble of similar cells from the same origin, which in the collective carry out a specific function. Organs are formed by the functional grouping together of multiple tissues.

The study of tissue is known as **histology** and in connection with disease, we speak about: **histopathology.**

As we can see in the slide, there is a range of different tissues, e.g.
a)   the skin,
b)   fibrous connective tissue, forming a tendon (or sinew, which connects muscles to bones, in German: "Sehne"),
c)   adipose tissue (fat),
d)   cartilage, at the end of a bone (in German: "Knorpel"),
e)   bone,
f)   blood (white cells, red cells and plasma).

An area of research is **tissue engineering,** which involves the use of materials and cells with the goal of trying to understand tissue function and tissue or organ on the body to be made de novo (Rouwkema, Rivron & van Blitterswijk, 2008).

http://asavory.edublogs.org/files/2012/11/20_05ConnectiveTissue-L-10fyvfx.jpg

Slide 1-26: Organ Example: Heart
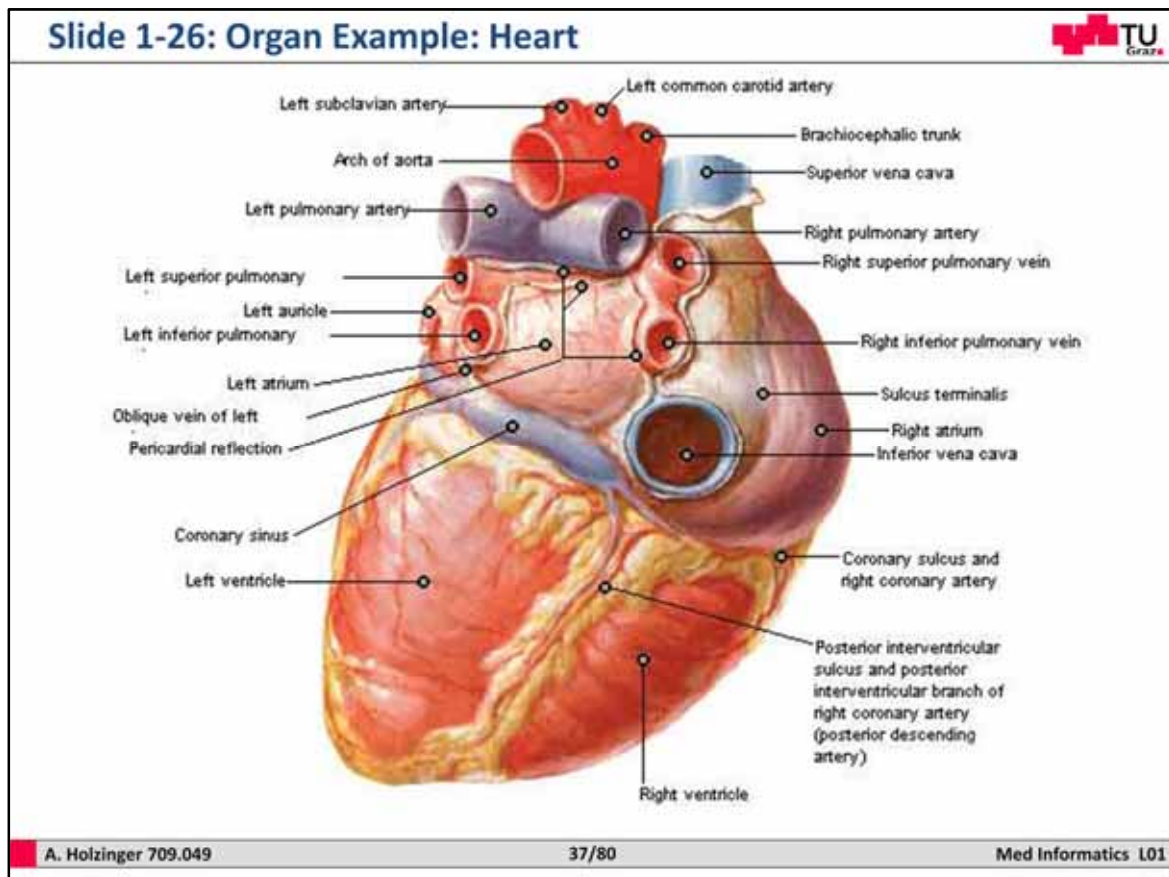
A collection of tissues joined in a structural unit to serve a specific function is called organ, we will just look at one single organ as an example:

The heart is a hollow muscle pumping blood throughout the blood vessels in the circulatory cardiovascular system, by repeated, rhythmic contractions. The adjective cardiac (from Greek: καρδιά = heart) means "related to the heart". Cardiology is the medical subject that deals with cardiac diseases and abnormalities.

Source: http://www.atihealthnet.com/pages/heart2.jpg
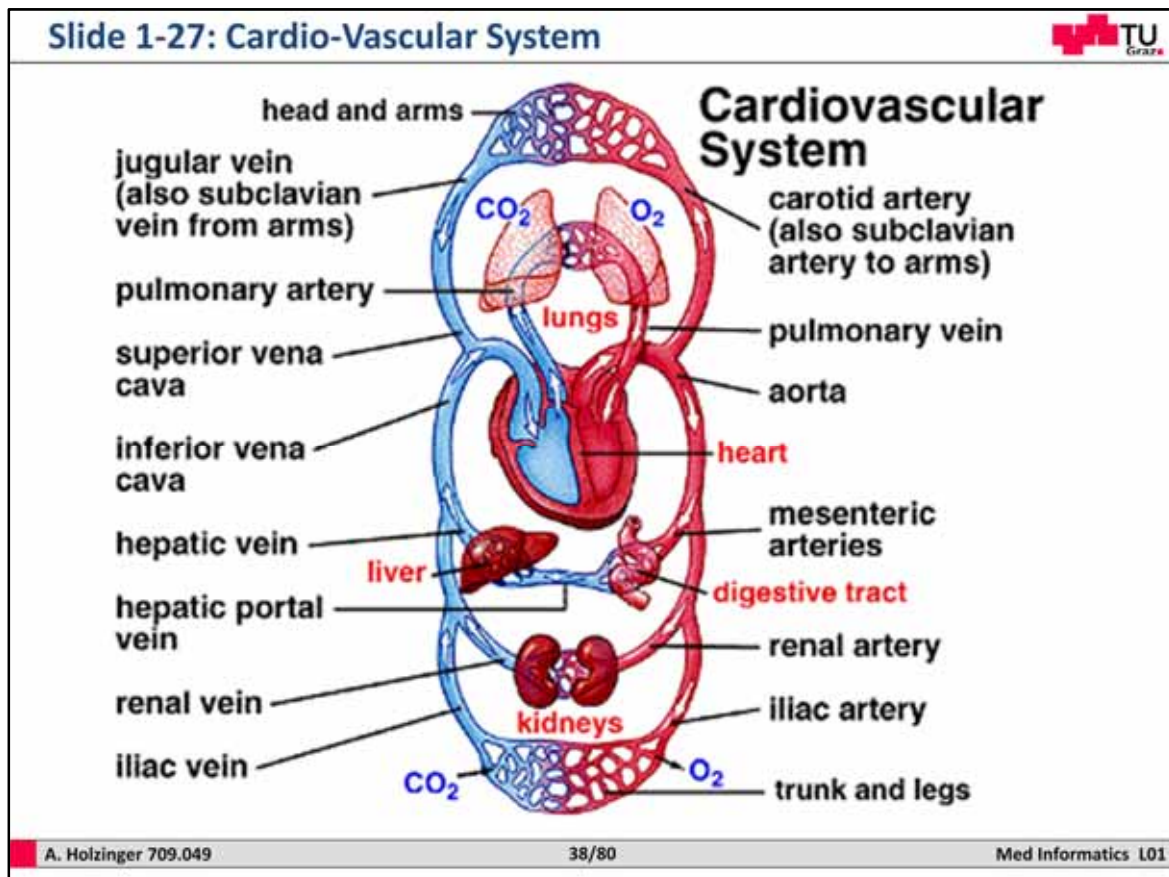
The essential components of the human cardiovascular system include the heart, the blood vessels (arteries and veins) and the blood itself. There are two circulatory systems: the pulmonary circulation, through the lungs where blood is oxygenated; and the systemic circulation, through the remaining body to provide oxygenated blood. An average adult contains five liter of blood.

Source: http://www.scientificpsychic.com/health/cardiovascular-system.gif
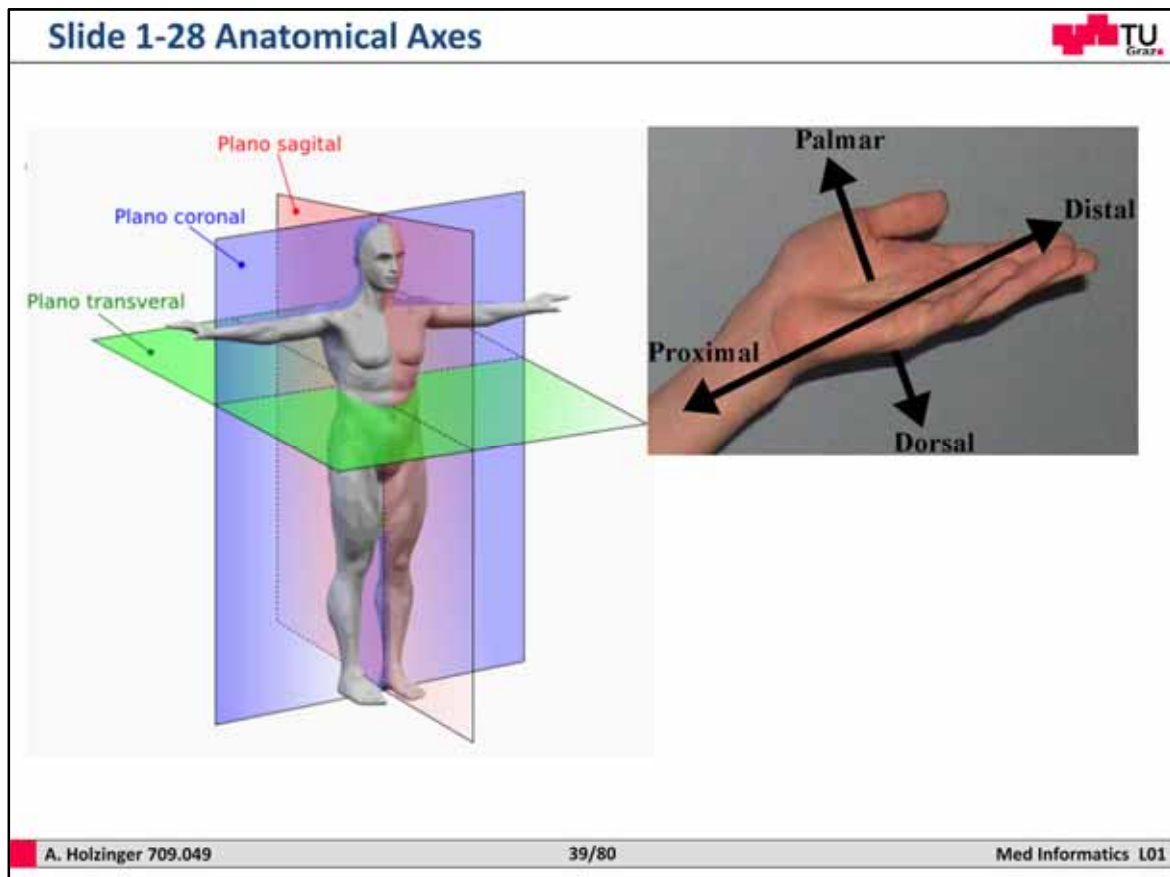
In Anatomy three reference planes are used:
1) a **sagittal** plane, which divides the body into sinister and dexter (left and right) portions;
2) A **coronal** or frontal plane divides the body into dorsal and ventral (back and front, or posterior and anterior); and
3) A **transversal** plane, also known as an axial plane or cross-section, divides the body into cranial and caudal (head and tail) portions.

In standard anatomical position, the palms of the hands point anteriorly, so anterior can be used to describe the palm of the hand, and posterior can be used to describe the back of the hand and arm.

Source: http://www.docstoc.com/docs/93735451/Anatomical-Position-and-Directional-Terms
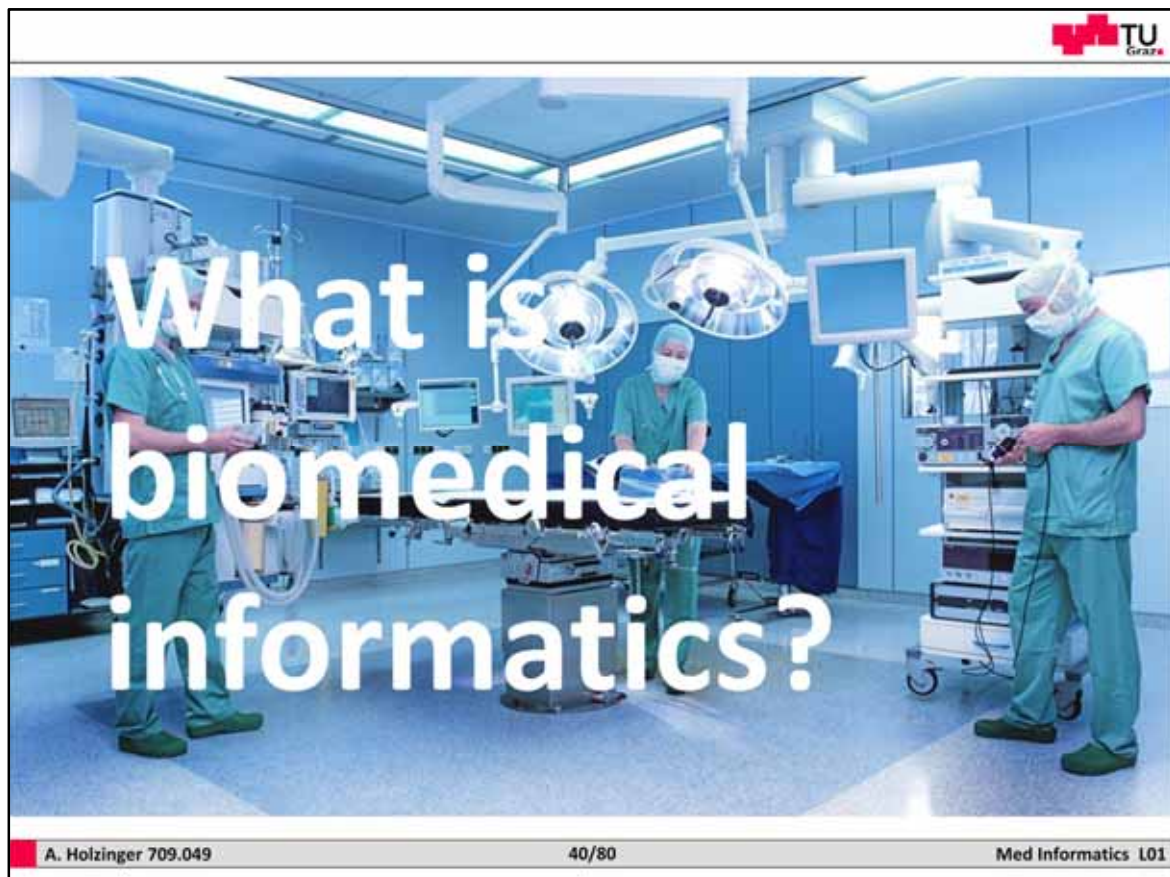
For a first understanding, let us compare the two fields: Medicine & Informatics,
In particular let us start with defining what "medicine" is.

From Clinical Medicine to Molecular Medicine

Yapijakis, C. (2009) Hippocrates of Kos, the Father of Clinical Medicine, and Asclepiades of Bithynia, the Father of Molecular Medicine. *In Vivo, 23, 4, 507-514.*

A. Holzinger 709.049   41/80   Med Informatics L01

Medicine is both the science and the art of healing and encompasses a variety of practices to maintain and restore health. Medicine is a very old discipline, having a more than 3000 year long tradition (Imhotep, 3rd millennium BC, Egypt) and is established since Hippocrates of Kos (460 BC – 370 BC) the founder of western *clinical medicine* at the time of Pericles in Classical Athens and Asclepiades of Bithynia (124-40 BCE) the first physician, who established Greek medicine in Rome. Influenced by the Epicureans, he followed atomic theory, modification and evolution and can be regarded as the founder of *molecular medicine* (Yapijakis, 2009). Clinical medicine can roughly be separated into three huge areas: 1) neoplasm (abnormal growth of cells); 2) inflammation (= part of the biological response of vascular tissues to harmful stimuli, such as pathogens, damaged cells, or irritants, Note: this is not a synonym for infection), and 3) trauma (physiological injuries caused by an external source). Note: You will often hear the term "Evidence based medicine (EBM)". EBM has evolved from clinical epidemiology and includes scientific methods in clinical decision making (Sackett et al., 1996).

Informatics on the other hand is a very young discipline, existing for slightly more than 50 years. The word is stemming from the combination of the words *informat*ion + automa*tics* (Steinbuch, 1957) and was a coined word for the science of automatic *information processing.* Automatic refers to the use of a machine which evolved slightly more than 10 years before that time and was called computer. Informatics is the **science of information,** *not* necessarily of computers, but Informatics is researching in and with computational methods, therefore is using computers – so as Astronomers uses telescopes – but do not necessarily construct telescopes (Holzinger, 2003a).

Holzinger, A. 2003. Basiswissen IT/Informatik. Band 2: Informatik, Wuerzburg, Vogel Buchverlag.

A. Holzinger 709.049   42/80   Med Informatics L01
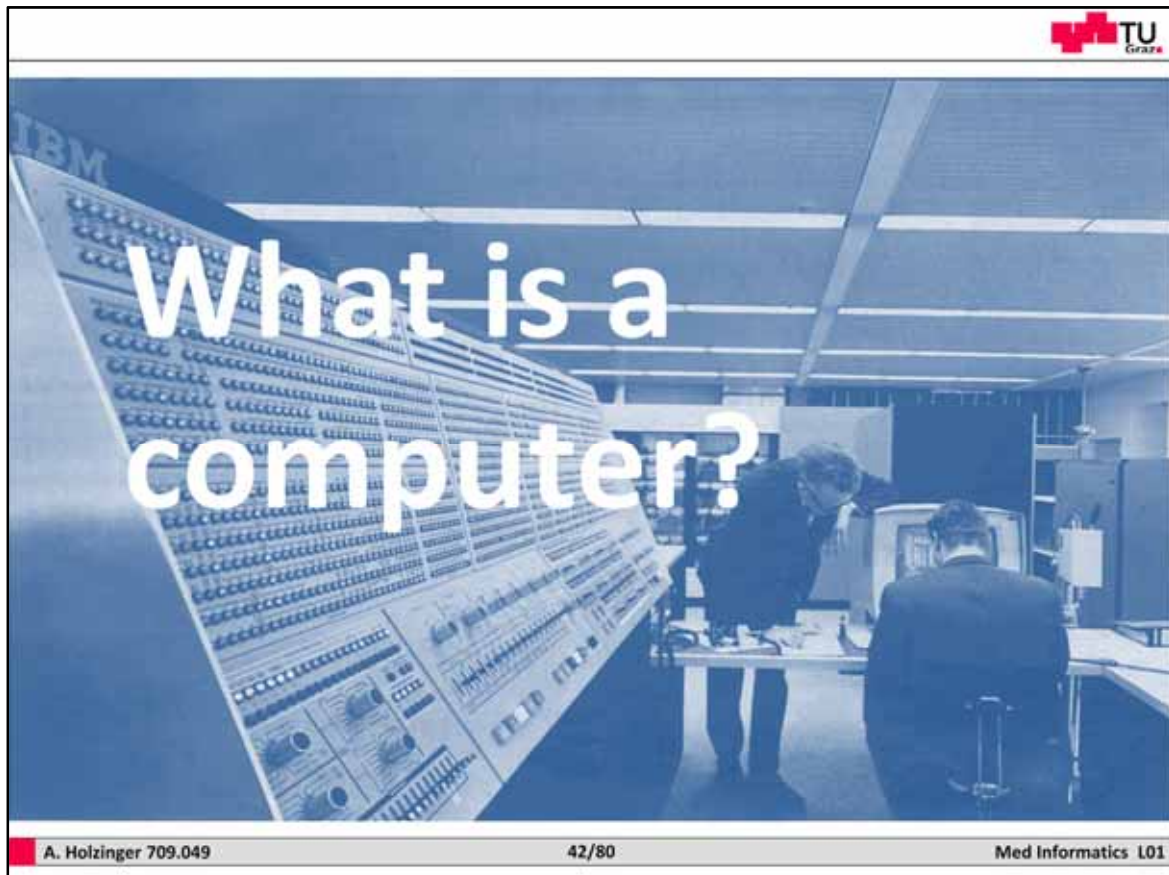
Computers are physical devices; and, actually what we understand as a computer is a general purpose electronic digital programmable machine. This machine (hardware) responds to a specific set of instructions in a well-defined manner and executes pre-recorded lists of instructions (software or program). Up to date in nearly all our everyday computer systems the so-called von Neumann architecture is used (Neumann, 1945). This is a model which uses a central processing unit (CPU) and a single separate memory to hold instructions and data (see Slide 1-28).

Note: A major contribution towards the development of the general purpose digital programmable machine was done by Alan Turing (1912-1954):
Turing's work is based on a more mathematical work by his doctoral advisor: Alonzo Church (1903-1995), whose work on lambda calculus intertwined with Turing's in a formal theory of computation: The Church–Turing thesis, which states that Turing machines capture the informal notion of effective methods in logic and mathematics, and provide a precise definition of an algorithm (=mechanical procedure). All theory development reaches back to this format and is far reaching; here just one example: In 1980 Richard Feynman (Nobel Prize Winner 1965) proposed a quantum computer, actually an analogous version of a quantum system. Contrary to digital computers, an analogue computer works with continuous variables; however, such analogue computers have worked only for special problems (see next Slide).
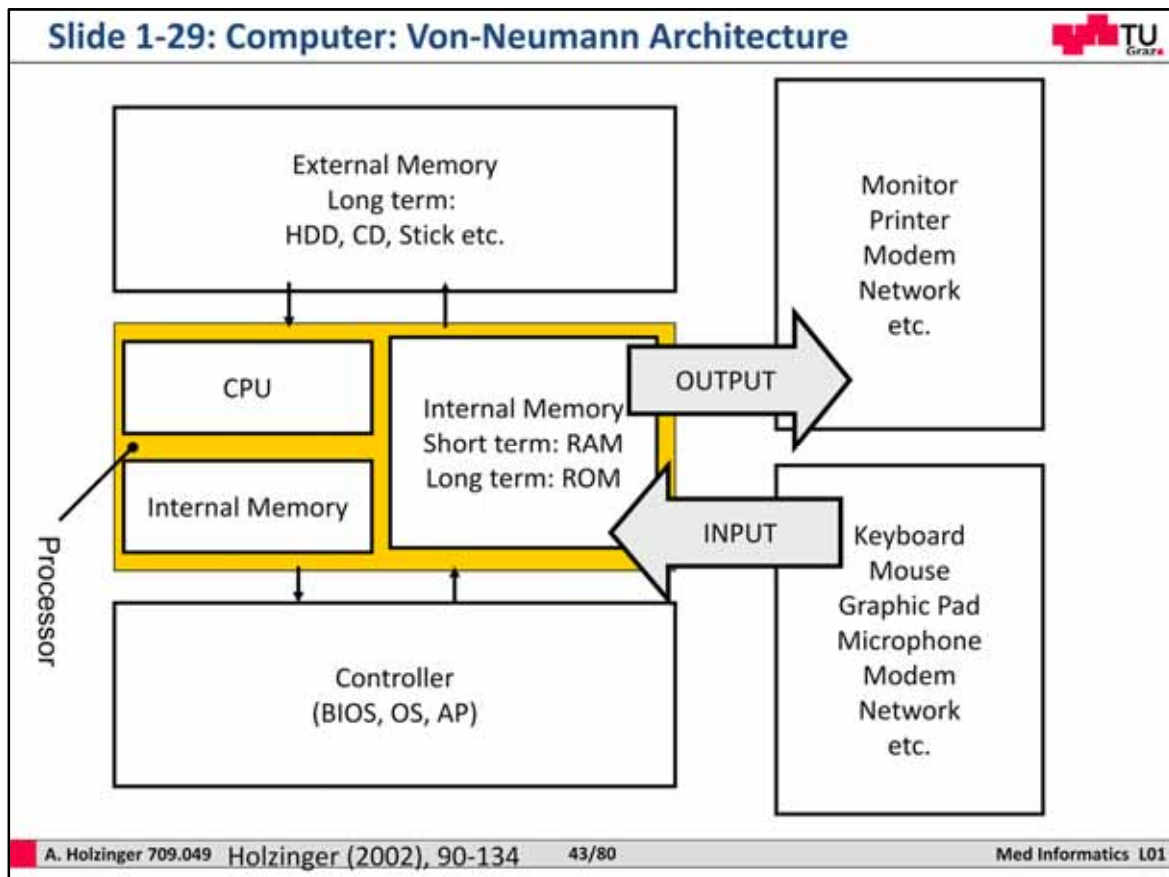
In this image we can see the anatomy of a Von-Neumann machine: The processor executes the instructions with the central processing unit (CPU), the "heart" of the computer. The memory enables a computer to store, data and programs; the external memory are mass storage devices and allows a computer to permanently retain large amounts of data. The input device consists usually of a keyboard and mouse, but nowadays the finger replaces the mouse more and more (Holzinger, 2003b).

Finally the output device displays the results on a screen etc. Actually, the large mainframe computers until the 1970ies had neither keyboard nor screens, they were programmed via punch-cards and the results were obtained via punch-cards or tapes. The Von-Neumann Architecture is theoretically equivalent to a **Universal Turing Machine (UTM).** It was described by Alan Turing (1912-1954) in 1936, who called it an "A(utomatic)-machine", and was not intended as a practical computer – just as a Gedankenexperiment; a Turing machine can be used to simulate the logic of any computer algorithm, and is useful in explaining the functions of a CPU inside a computer (Holzinger, 2002).

This **digital computer,** where everything is expressed in **bits**, has proven to be **universally applicable** and Feynman's innovative idea was without issue until the proposal was put in the form of a **quantum Turing machine,** that is, a **digital quantum computer.**

The Von Neumann machine (single instruction, single data stream) is still the basis for our digital world of today, although there are some other architectures of so called "Non-Vons": Single Instruction, Multiple Data streams (SIMD), Multiple Instruction, Single Data stream (MISD) and Multiple Instruction, Multiple Data streams (MIMD) – apart from different computational paradigms, e.g. evolutionary algorithms, swarm intelligence or cellular automata etc. (Burgin & Eberbach, 2012), (Cooper, Loewe & Sorbi, 2008).
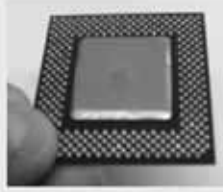
We may ask why digital technology generally is so successful. The answer is in the progress of technological performance, i.e. "Digital Power", which is mainly based on Moore's Law (see the next Slide 1-29).

Slide 1-30: Technological Performance / Digital Power

Gordon E. Moore (1965, 1989, 1997)

Digital Power :=

communication × computing × storage × content

„fiber law"   „Moore's law"   „disk law"   „community law"

⇓   ⇓   ⇓   ⇓

doubles every 9 months   doubles every 18 months   doubles every 12 months 2^n wherin n is the # of people

Holzinger, A. 2002. *Basiswissen IT/Informatik Band 1: Informationstechnik. Das Basiswissen für die Informationsgesellschaft des 21. Jahrhunderts, Wuerzburg, Vogel Buchverlag.*

A. Holzinger 709.049   44/80   Med Informatics L01

In 1965, Intel co-founder Gordon E. Moore (1929–) noted that the number of components in integrated circuits had doubled every year from 1958 to 1965 and he predicted this trend would continue "for at least ten years" (Moore, 1965). In 1975, Moore revised the Law to predict the doubling of computer processing power every two years – and this prediction has come true ever since, often doubling within 18 months. Argued several times, this law shows no sign of slowing down, and the prediction is beyond 2025 (Holzinger, 2002).

Directly connected with the increasing performance of digital technology, another astonishing trend is visible, parallel to the raising processing speed and memory capacity: a vast reduction in cost and size (Slide 1-30).
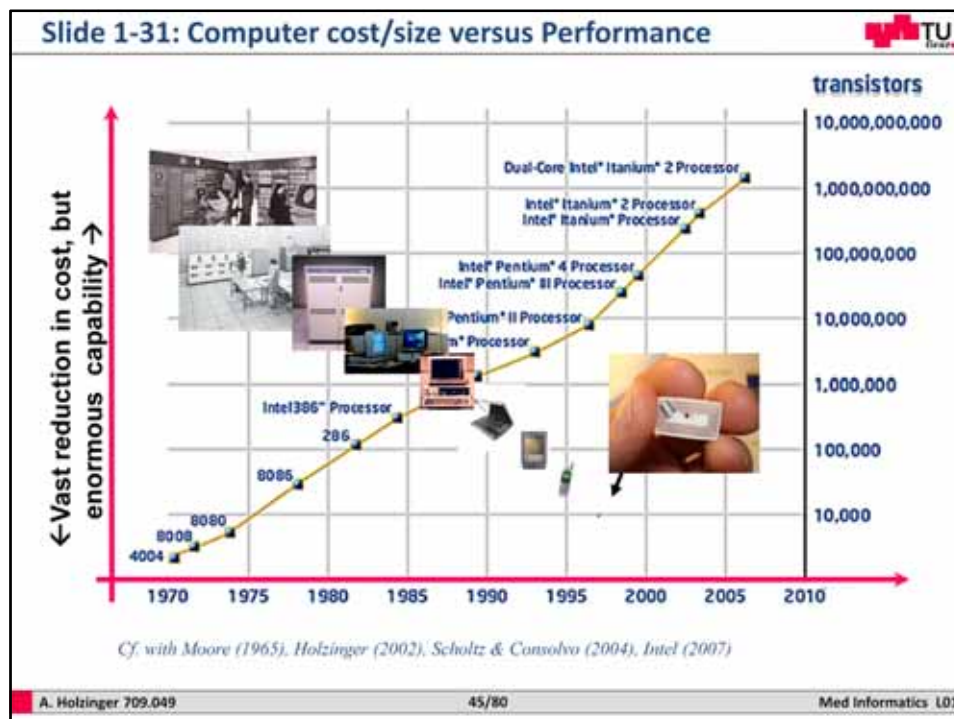
In this slide we can observe a phenomenon which is directly connected with Moore's law: Whilst the digital power is increasing, both cost and size are decreasing, making large computational power available and affordable today at pocket size (Smartphone).
However, the limiting factor for continued miniaturization is that the structures will eventually reach the limits of miniaturization at atomic levels.

Consequently, there are international efforts to find alternatives for Silicon-based semiconductors (see: http://www.itrs.net). The smallest size for both memory and logical devices depends on the mass of the information transporting particles, i.e. the smallest barrier width in Si devices is 5 nm (size of electrons in Si), consequently there is a physical barrier semiconductor designers will face. A big chance is biological computing: Due to the fact that the living cell is an information processor, which is extremely efficient in the execution of its functions, there are huge efforts towards creating of a Bio-cell, which outperforms the Si-cell in every aspect, see Slide 1-31.
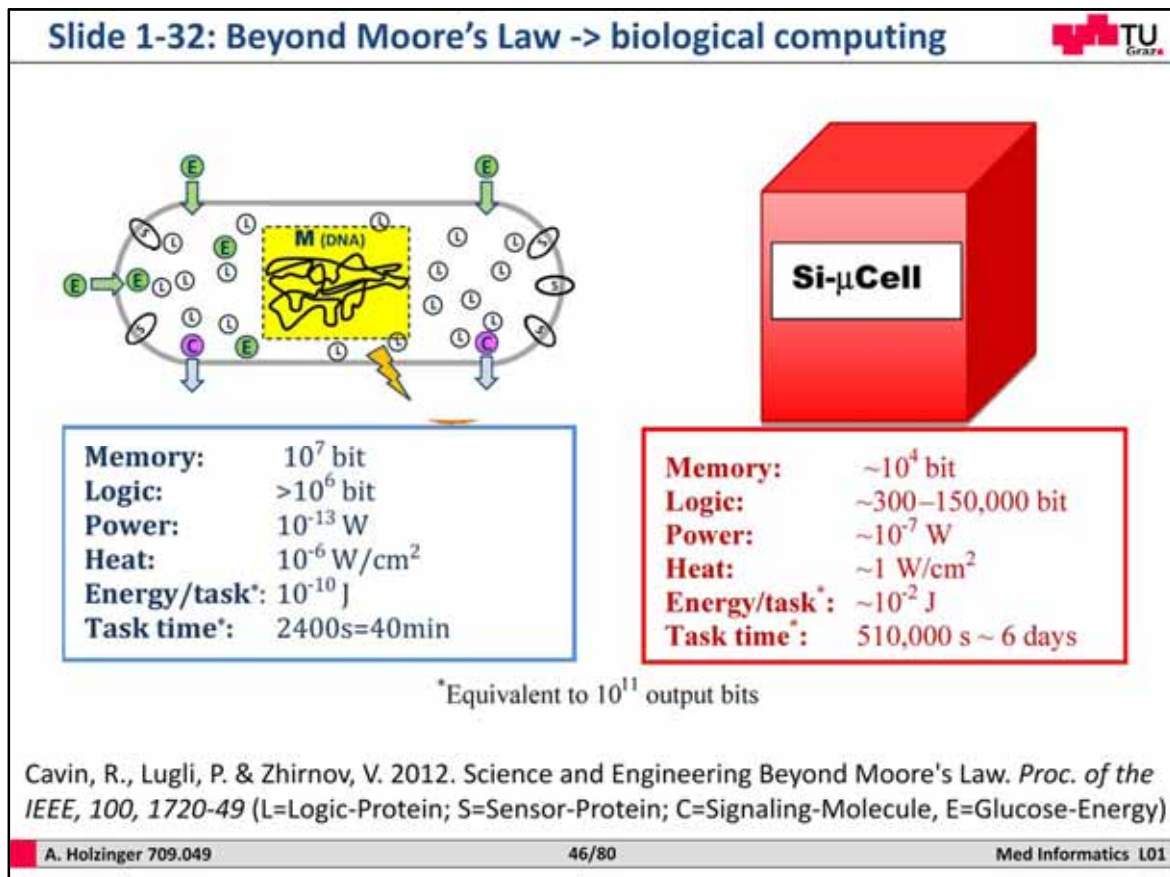
Bio-Cell versus Si-Cell: Comparison between unicellular organism as information processing device with a modern Si-Cell; In the Bio-Cell (left) the components include: L=Logic-Proteins; S=Sensor-Proteins; C=Signaling-Molecules, E=Glucose-Energy (Cavin, Lugli & Zhirnov, 2012).

Reference:
Cavin, R., Lugli, P. & Zhirnov, V. 2012. Science and Engineering Beyond Moore's Law. Proceedings of the IEEE, 100, (13), 1720-1749.

Slide 1-33 From mainframe to Ubiquitous Computing
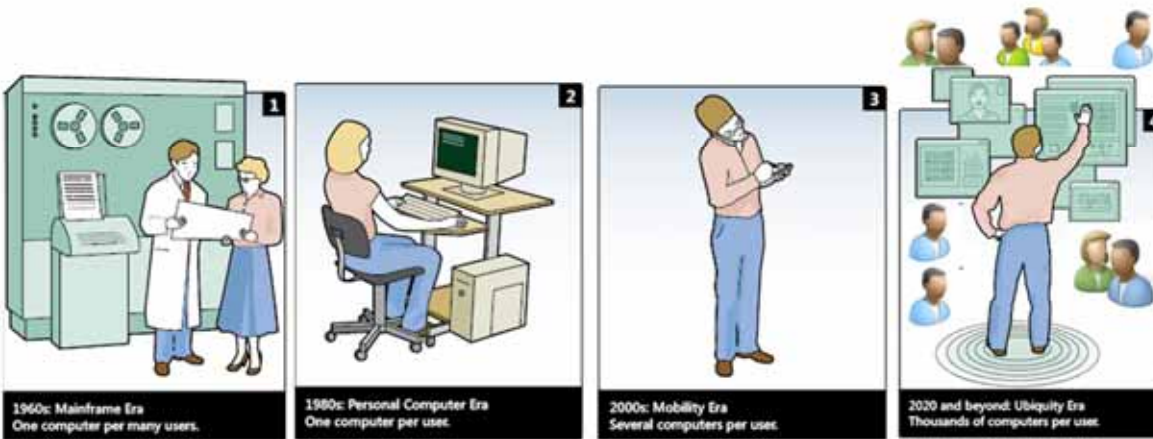
■ … using technology to augment human capabilities for structuring, retrieving and managing information

1960s: Mainframe Era
One computer per many users.

1980s: Personal Computer Era
One computer per user.

2000s: Mobility Era
Several computers per user.

2020 and beyond: Ubiquity Era
Thousands of computers per user.

Harper, R., Rodden, T., Rogers, Y. & Sellen, A. (2008) *Being Human: Human-Computer Interaction in the Year 2020*. Cambridge, Microsoft Research.

A. Holzinger 709.049   47/80   Med Informatics L01

Since the advent of computers, a dream of mankind was, to use such computers to augment human capabilities for structuring, retrieving and managing information – however, this was and it is still not easy, although Human–Computer Interaction (HCI) has dramatically changed:

The effects described in Slide 1-30 are directly connected with a change in Human-Computer Interaction:
1) At the beginning (1960) there were one computer for many users, and the users were experts, who mostly worked in computer science;
2) The next idea was to provide every user a "personal" computer – the PC was born (2);
3) The reduction in size and the progress in wireless networking made computers mobile (3);
4) Finally, computers became pervasive and ubiquitous (4);

Ubiquitous computing (Ubicomp) was proposed by (Weiser, 1991) and is a post-desktop model of Human–Computer Interaction in which the information processing is integrated into everyday objects, see Slide 1-33:

Slide 1-34: Ubiquitous Computing – Smart Objects

Holzinger, A., Nischelwitzer, A., Friedl, S. & Hu, B. (2010) Towards life long learning: three models for ubiquitous applications. *Wireless Communications and Mobile Computing*, 10, 10, 1350-1365.

A. Holzinger 709.049   48/80   Med Informatics L01
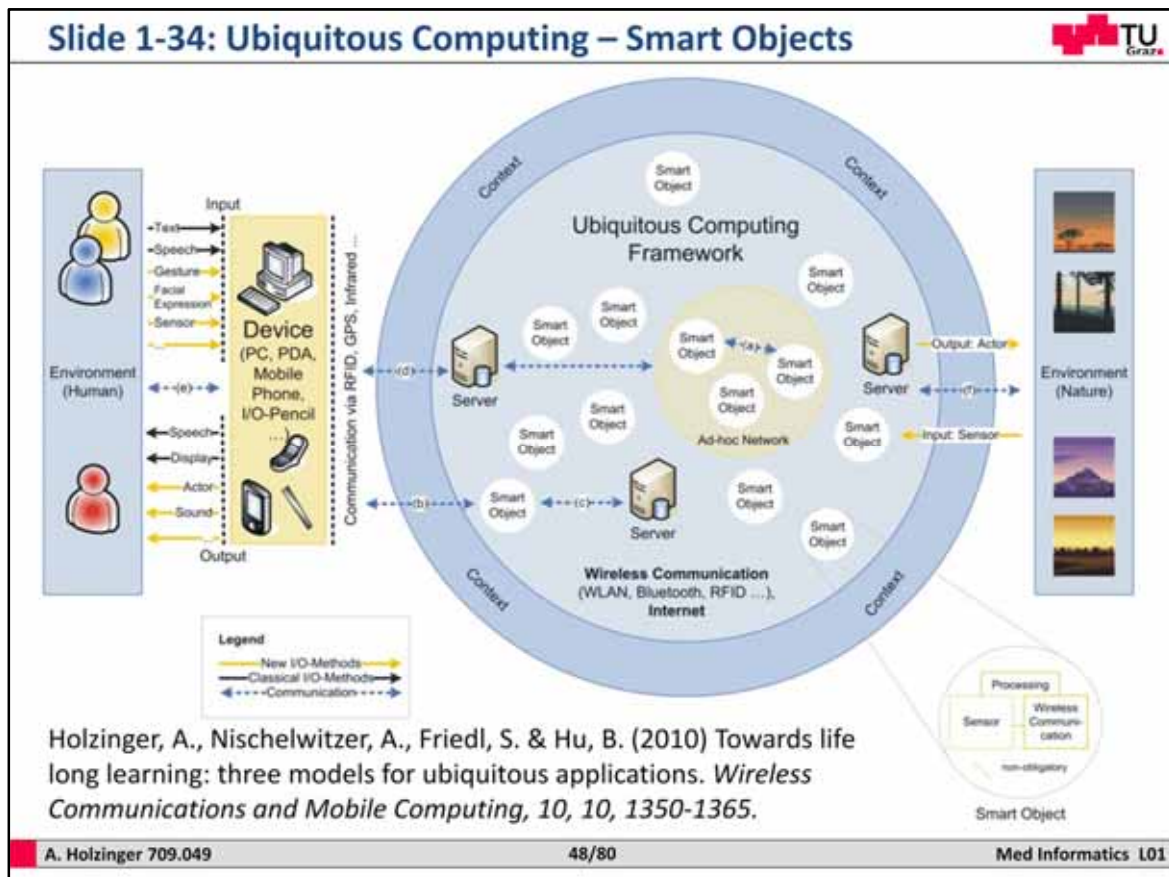
Ubicomp: Smart objects can exchange information, consequently can "talk" to each other (Holzinger et al., 2010b). Usually we speak from ubiquitous and pervasive technology: "ubiquitous" literally means omnipresent (in German "allgegenwärtig"); and the term pervasive implies that the technology is penetrating our daily life (in German "durchdringend"). Such devices are intended:

1) to support end users in their tasks without overwhelming them with complexity of networks, devices, software, databases; (stay connected at every place in the world)

2) to ensure "context awareness" and knowledge about the environment in which the users focus on their tasks; (have access to all your data from everywhere)

3) to provide natural interaction between user and technology e.g. by using gesture, speech, etc., and most intuitively by using multi-touch and advanced stylus technology (Holzinger et al., 2012b), (Holzinger et al., 2012a), (Holzinger et al., 2011a), (Holzinger et al., 2010c).
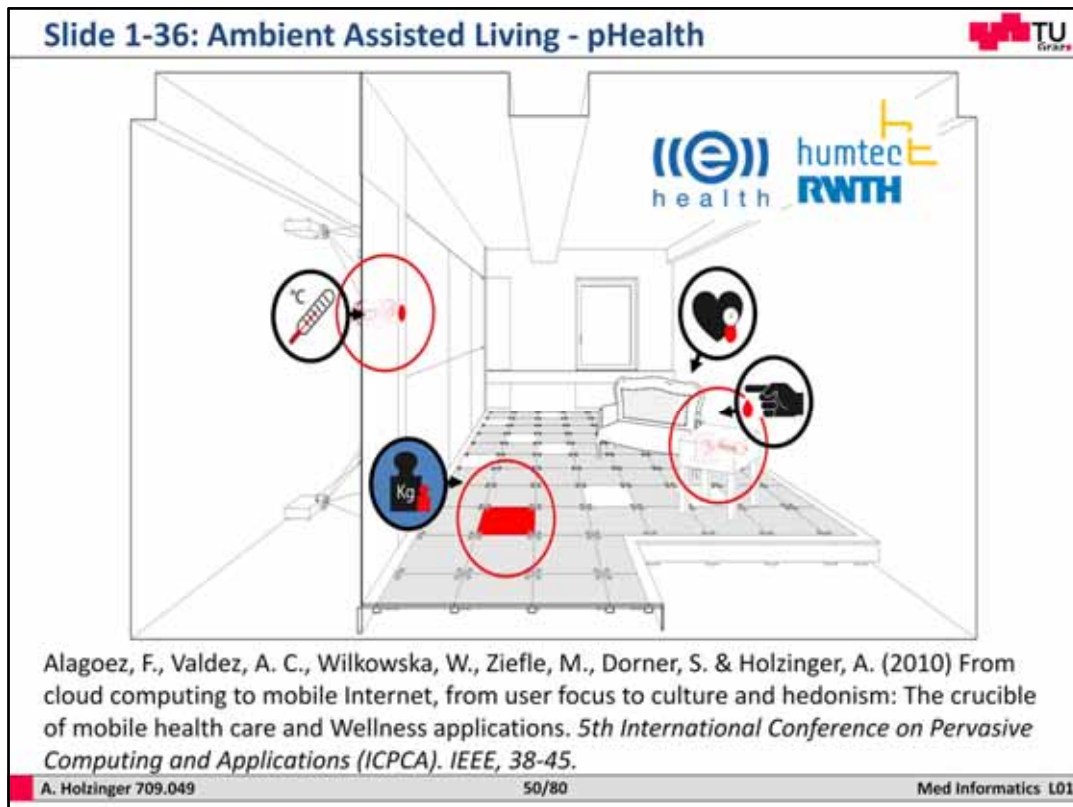
Slide 1-35 Example: Pervasive Health Computing

Holzinger, A., Schaupp, K. & Eder-Halbedl, W. (2008) An Investigation on Acceptance of Ubiquitous Devices for the Elderly in an Geriatric Hospital Environment: using the Example of Person Tracking In: *Lecture Notes in Computer Science (LNCS 5105). Heidelberg, Springer, 22-29.*

A. Holzinger 709.049   49/80   Med Informatics L01

Ubicomp include Radio Frequency Identification (RFID), sensor networks, Augmented Reality, mobile, wearable and implanted devices. Since work in hospitals is formed by many cooperating clinicians having a high degree of mobility, parallel activities and no fixed workplace, existing IT solutions often fail to consider these issues (Bardram, 2004), (Mitchell et al., 2000). Ubicomp in Health is promising (Varshney, 2009) and may address such issues see Slide 1-34:

An example for the application of Ubicomp in the area of Ambient Assisted Living: RFID technology for localizing elderly people, suffering from dementia (Holzinger, Schaupp & Eder-Halbedl, 2008)

Slide 1-36: Ambient Assisted Living - pHealth

Alagoez, F., Valdez, A. C., Wilkowska, W., Ziefle, M., Dorner, S. & Holzinger, A. (2010) From cloud computing to mobile Internet, from user focus to culture and hedonism: The crucible of mobile health care and Wellness applications. *5th International Conference on Pervasive Computing and Applications (ICPCA). IEEE, 38-45.*

A. Holzinger 709.049   50/80   Med Informatics L01

A further example for Ubicomp technologies can be seen in the future care lab of the RWTH Aachen (Slide 1-35): An integrated set of smart sensor technologies provides unobtrusive monitoring of a patient's vital health functions, such as:

a smart floor, providing location tracking of multiple persons, fall detection and weight measurement,

an infrared camera, for non-invasive temperature measurement,

measurement equipment which is naturally integrated into the furniture, such as blood pressure or coagulation measurement devices.

Centerpiece of all human computer interaction inside the future care lab is the 4,8m x 2,4m big multi-touch display wall.

Advances in Biomedical Informatics & Biomedical Engineering provide the foundations for modern patient-centered healthcare solutions, health care systems, technologies and techniques. Ubicomp in the living lab, measures unobtrusively the vital parameters: Blood pressure, Pulse rate, Body temperature, weight (Alagöz et al., 2010) – a further good example for "big data".

Slide 1-37: Pervasive Computing in the Hospital

Midrange Reader Feig MR100

EPR

Wristband including RFID Transponder Infineon my-d (15693)

Compact Flash Slot Module for Tablet PC or Mobile Devices Microsensys

Mid Range Reader PRH100 Serial COM Interface

Pad Antenna Feig 300/300

Holzinger, A., Schwaberger, K. & Weitlaner, M. (2005) Ubiquitous Computing for Hospital Applications: RFID-Applications to enable research in Real-Life environments *29th Annual IEEE International Computer Software & Applications Conference (IEEE COMPSAC), 19-20.*

A. Holzinger 709.049   51/80   Med Informatics L01

Patients check in at the Hospital – in addition to an ordinary wristband an RFID transponder is supplied. Patient data is entered via our application at the check-in-point, any previous patient data can be retrieved from the HIS. From this information, uncritical but important data (such as name, blood type, allergies, vital medication etc.) is transferred to the wristband's RFID transponder. The Electronic Patient Record (EPR) is created and stored at the central server. From this time the patient is easily and unmistakably identifiable. All information can be read from the wristband's transponder or can be easily retrieved from the EPR by identifying the patient with a reader. In contrast to manual identification, automatic processes are less error-prone. Unlike barcodes, RFID transponders can be read without line of sight, through the human body and most other materials. This enables physicians and nurses to retrieve, verify and modify information in the Hospital accurately and instantly. In addition, this system provides patient identification and patient data – even when the network is crashed (Holzinger, Schwaberger & Weitlaner, 2005).

Slide 1-41: 4 decades from Medical to Biomedical Informatics

- **1970+ Begin of Medical Informatics**
  - Focus on data acquisition, storage, accounting (typ. "EDV")
    - The term was first used in 1968 and the first course was set up 1978
- **1985+ Health Telematics**
  - Health care networks, Telemedicine, CPOE-Systems etc.
- **1995+ Web Era**
  - Web based applications, Services, EPR, etc.
- **2005+ Ambient Era**
  - Pervasive & Ubiquitous Computing
- **2010+ Quality Era – Biomedical Informatics**
  - Information Quality, Patient empowerment, individual molecular medicine, End-User Programmable Mashups

A. Holzinger 709.049   55/80   Med Informatics L01

1970+ Begin of **Medical Informatics**
    Focus on data acquisition, storage, accounting (typ. "EDV")
    The term was first used in 1968 and the first course was set up 1978
1985+ Health Telematics
    Health care networks, Telemedicine, CPOE-Systems etc.
1995+ Web Era
    Web based applications, Services, EPR, etc.
2005+ Ambient Era
    Pervasive & Ubiquitous Computing
2010+ Quality Era – **Biomedical Informatics**
Information Quality, Patient empowerment, individual molecular medicine, End-User Programmable Mashups (Auinger et al., 2009).

74