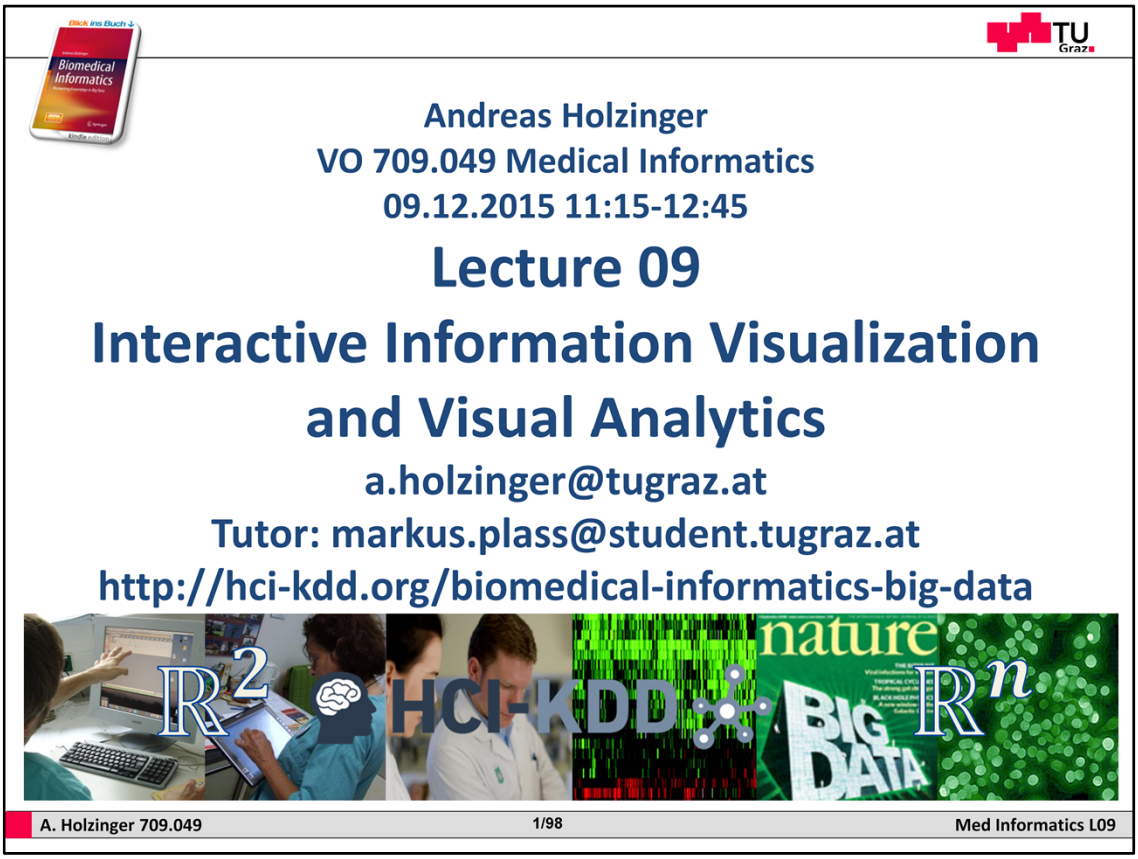


Status as of 09.12.2015 09:30

Dear Students, welcome to the 9th lecture of our course. Please remember from the last lecture: some applications of decision support systems,
Basic architectures of DSS, the certainty factor, the historical roots in DSS, particularly MYCIN, Case-based reasoning systems, and please remember the differences between unsupervised, supervised, semi-supervised and interactive machine learning with the human-in-the-loop to help in solving problems which would otherwise be NP-hard.

Please always be aware of the definition of biomedical informatics (Medizinische Informatik, health informatics):
Biomedical Informatics is the inter-disciplinary 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 (and well-being).

Schedule

- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 2. Back to the future: Fundamentals of Data, Information and Knowledge
- 3. Structured Data: Coding, Classification (ICD, SNOMED, MeSH, UMLS)
- 4. Biomedical Databases: Acquisition, Storage, Information Retrieval and Use
- 5. Semi structured and weakly structured data (structural homologies)
- 6. Multimedia Data Mining and Knowledge Discovery
- 7. Knowledge and Decision: Cognitive Science & Human-Computer Interaction
- 8. Biomedical Decision Making: Reasoning and Decision Support
- **9. Intelligent 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

Within the next 90 minutes you will learn …
a little bit about interactive and hopefully intelligent information visualization and visual analytics

Keywords of the 9th Lecture

- Data visualization
- Flow cytometry
- Human-Computer Interaction (HCI)
- Information visualization
- Interactive information visualization
- k-Anonymization
- Longitudinal data
- Multivariate data
- Parallel coordinates
- RadViz
- Semiotics
- Star plots
- Temporal data analysis
- Visual analytics
- Visual information

Data visualization
Flow cytometry
Human-Computer Interaction (HCI)
Information visualization
Interactive information visualization
k-Anonymization
Longitudinal data
Multivariate data
Parallel coordinates
RadViz
Semiotics
Star plots
Temporal data analysis
Visual analytics
Visual information

Advance Organizer (1/2)

- **Biological data visualization** = as branch of bioinformatics concerned with visualization of sequences, genomes, alignments, phylogenies, macromolecular structures, systems biology, etc.
- **Clustering** = Mapping objects into disjoint subsets to let appear similar objects in the same subset;
- **Data visualization** = visual representation of complex data, to communicate information clearly and effectively, making data useful and usable;
- **Information visualization** = the interdisciplinary study of the visual representation of large-scale collections of non-numerical data, such as files and software, databases, networks etc., to allow users to see, explore, and understand information at once;
- **Multidimensional scaling** = Mapping objects into a low-dimensional space (plane, cube etc.) in order to let appear similar objects close to each other;
- **Multi-Dimensionality** = containing more than three dimensions and data are multivariate;
- **multivariate** = encompassing the simultaneous observation and analysis of more than one statistical variable; (Antonym: univariate = one-dimensional);

A. Holzinger 709.049 4/98 Med Informatics L09

Biological data visualization = in bioinformatics the visualization of sequences, genomes, alignments, phylogenies, macromolecular structures, systems biology, etc.
Business Intelligence (BI) = all issues in a company which provides historical, current and predictive views of business operations; methods include data mining, process mining, content analytics and particularly visual analytics; BI is directly connected with decision support; any BI-system is also a decision support system;
Classification = the problem of identifying to which of a set of categories a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known;
Clustering = Mapping objects into disjoint subsets to let appear similar objects in the same subset; it is a main task in exploratory data mining in bioinformatics;
Content analytics = a general term addressing so-called “unstructured” data – mainly text – by using methods from visual analytics in business intelligence;
Data visualization = visual representation of complex data, to communicate information clearly and effectively, making data useful and usable;
Information visualization = the interdisciplinary study of the visual representation of large-scale collections of non-numerical data, such as files and software, databases, networks etc., to allow users to see, explore, and understand information at once;
Multi-dimensional = containing more than three dimensions and data are multivariate; Multidimensional scaling = Mapping objects into a low-dimensional space (plane, cube etc.) in order to let appear similar objects close to each other;
Multi-variate = encompassing the simultaneous observation and analysis of more

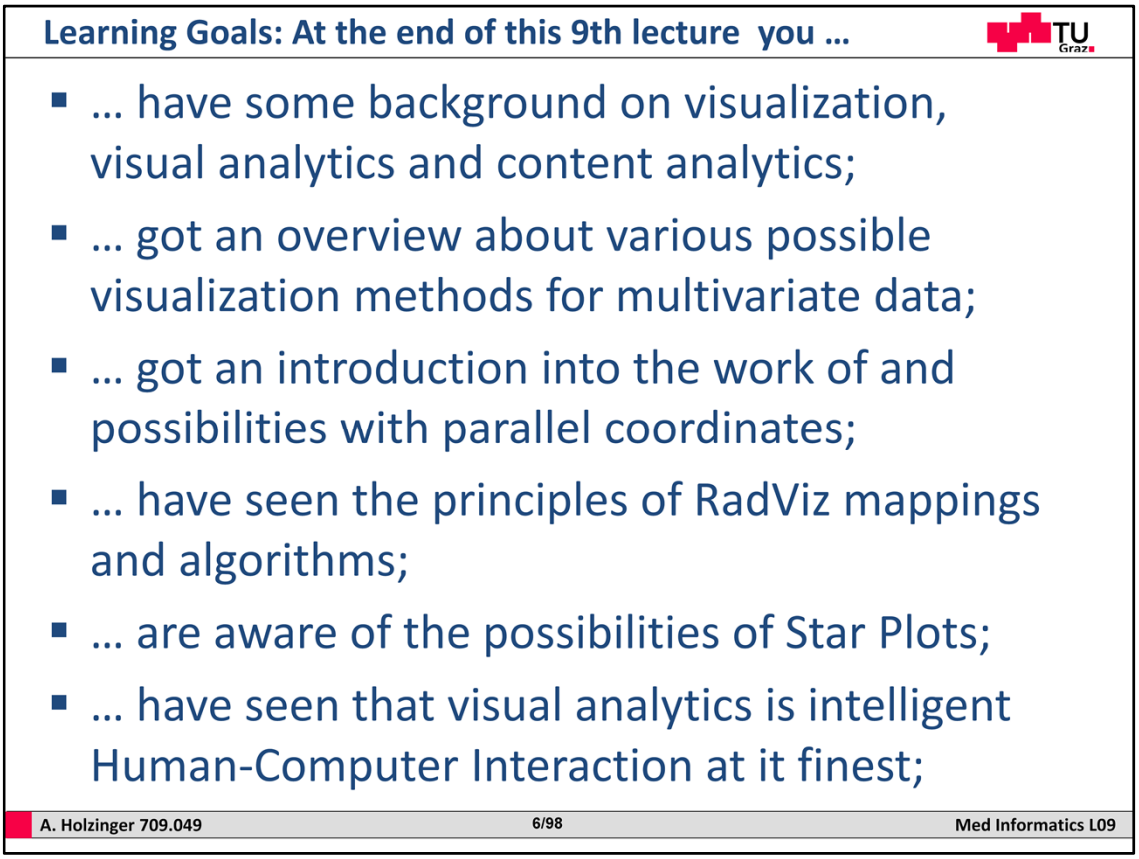
than one statistical variable; (Antonym: univariate = one-dimensional);
Parallel coordinates = for visualizing high-dimensional and multivariate data in the form of N parallel lines, where a data point in the n-dimensional space is transferred to a polyline with vertices on the parallel axes;
RadViz = radial visualization method, which maps a set of m-dimensional points in the 2-D space, similar to Hooke's law in mechanics;
Semiotics = deals with the relationship between symbology and language, pragmatics and linguistics. Information and Communication Technology deals not only in words and pictures but also in ideas and symbology;

Advance Organizer (2/2)

- **Parallel Coordinates** = for visualizing high-dimensional and multivariate data in the form of N parallel lines, where a data point in the n-dimensional space is transferred to a polyline with vertices on the parallel axes;
- **RadViz** = radial visualization method, which maps a set of m-dimensional points in the 2-D space, similar to Hooke's law in mechanics;
- **Semiotics** = deals with the relationship between symbology and language, pragmatics and linguistics. Information and Communication Technology deals not only in words and pictures but also in ideas and symbology;
- **Semiotic engineering** = a process of creating a semiotic system, i.e. a model of human intelligence and knowledge and the logic for communication and cognition;
- **Star Plot** = aka radar chart, spider web diagram, star chart, polygon plot, polar chart, or Kiviat diagram, for displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point;
- **Visual Analytics** = focuses on analytical reasoning of complex data facilitated by interactive visual interfaces;
- **Visualization** = a method of computer science to transform the symbolic into the geometric, to form a mental model and foster unexpected insights;

A. Holzinger 709.049 5/98 Med Informatics L09

Parallel coordinates = for visualizing high-dimensional and multivariate data in the form of N parallel lines, where a data point in the n-dimensional space is transferred to a polyline with vertices on the parallel axes;
RadViz = radial visualization method, which maps a set of m-dimensional points in the 2-D space, similar to Hooke's law in mechanics;
Semiotics = deals with the relationship between symbology and language, pragmatics and linguistics. Information and Communication Technology deals not only in words and pictures but also in ideas and symbology;
Semiotic engineering = a process of creating a semiotic system, i.e. a model of human intelligence and knowledge and the logic for communication and cognition;
Star plot = aka radar chart, spider web diagram, star chart, polygon plot, polar chart, or Kiviat diagram, for displaying multivariate data in the form of a two-dimensional chart of three or more quantitative variables represented on axes starting from the same point;
Visual analytics = focuses on analytical reasoning of complex data facilitated by interactive visual interfaces;
Visualization = a method of computer science to transform the symbolic into the geometric, to form a mental model and foster unexpected insights;
Visualization mantra = "Overview first, zoom & filter on demand" (Shneiderman, 1996);



At the end of this ninth lecture you will …

… have some theoretical background on visualization and visual analytics;
… got an overview about various possible visualization methods for various data;
… got an introduction into the work of and possibilities with parallel coordinates;
… have seen the principles of RadViz mappings and algorithms;
… are aware of the possibilities of Star Plots;
… have seen that visual analytics is intelligent Human-Computer Interaction;

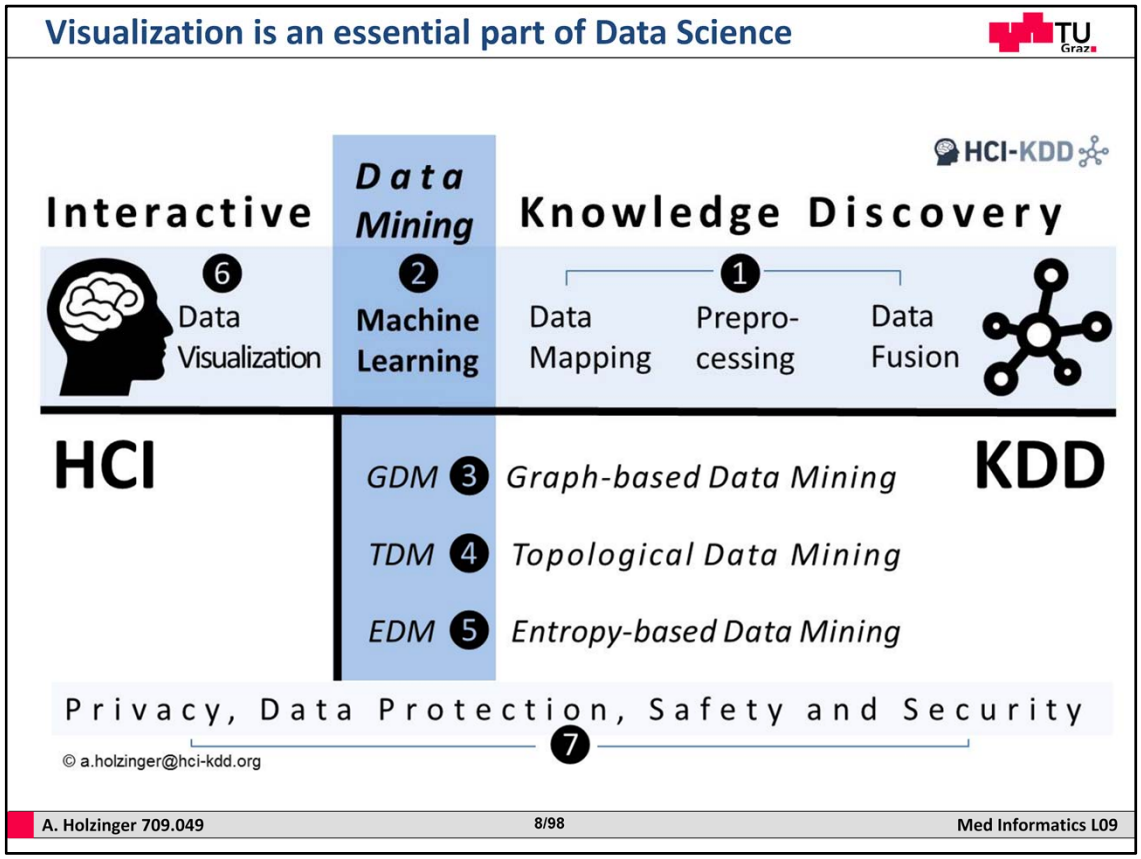
Slide 9-1 Key Challenges

- How to understand high-dimensional spaces?
- The transformation of results from high-dimensional space $\mathbb{R}^N$ into $\mathbb{R}^2$
- From the complex to the simple
- Low integration of visual analytics techniques into the clinical workplace
- Sampling, modelling, rendering, perception, cognition, decision making
- Trade-off between time and accuracy
- How to model uncertainty

A. Holzinger 709.049 7/98 Med Informatics L09

Information visualization is the study of visual representations of abstract data to reinforce human cognition; hence it is very important for decision making. A lot of challenges are involved: The human perceptual system can handle large quantities of data of few dimensions but has great difficulty as the data dimensionality increases (again: the curse of dimensionality (Donoho, 2000)). The grand challenge is to focus not simply on computational methods of displaying large quantities of data but on both perception and cognition of such large amounts of data. One aspect is to focus on how the process of computer visualization can be improved to mirror the process of natural visualization. Our perceptual systems were designed specifically for survival in and understanding of the surrounding external environment, not abstract objects and images (Grinstein, Inselberg & Laskowski, 1998).

Visual analysis is becoming an essential component of medical visualization due to the rapidly growing role and availability of complex multi-dimensional, time-varying, mixed-modality, simulation and multi-subject datasets. The magnitude, complexity and heterogeneity of the data necessitate the use of visual analysis techniques for diagnosis and medical research and, even more importantly, treatment planning and evaluation, e.g. radio therapy planning and post-chemotherapy evaluation (Childs et al., 2013).



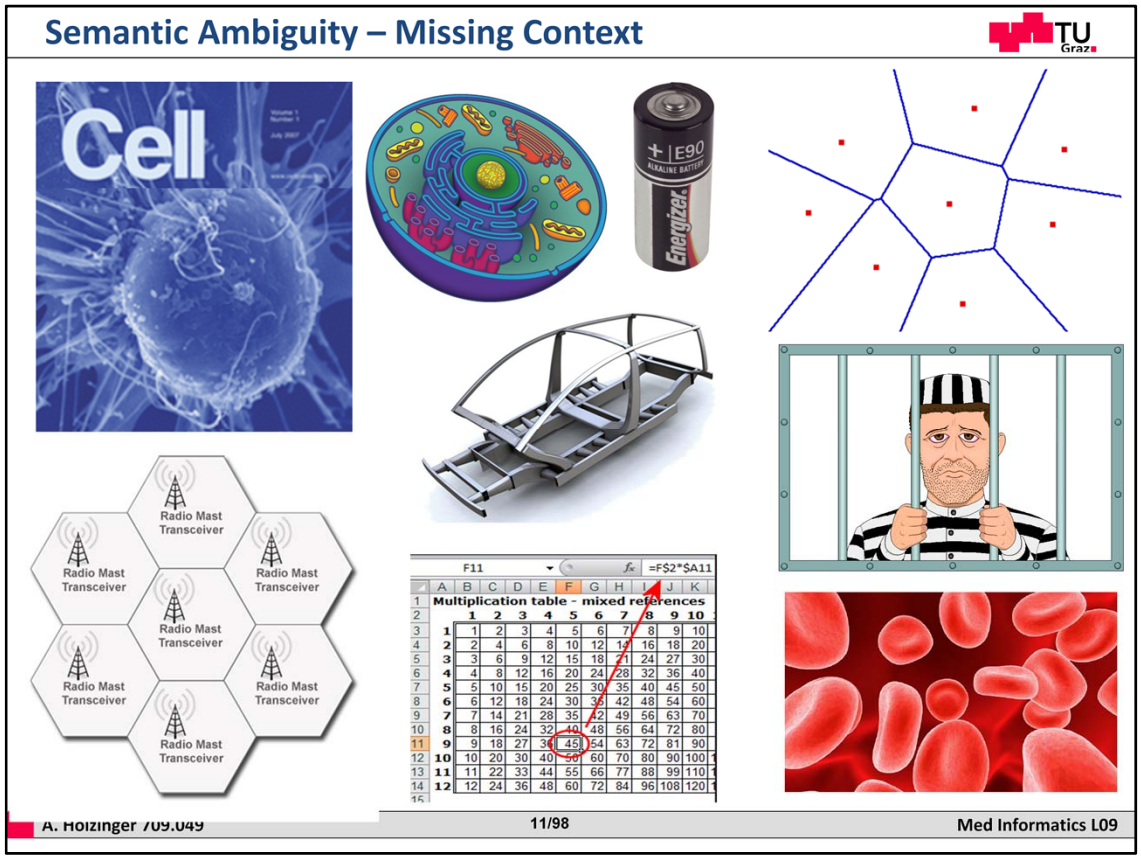
A concerted effort is needed in the horizontal ML-pipeline from data preprocessing to data visualization;
At the end of the pipeline the end-user want to see something …

Verbal Information versus Visual Information

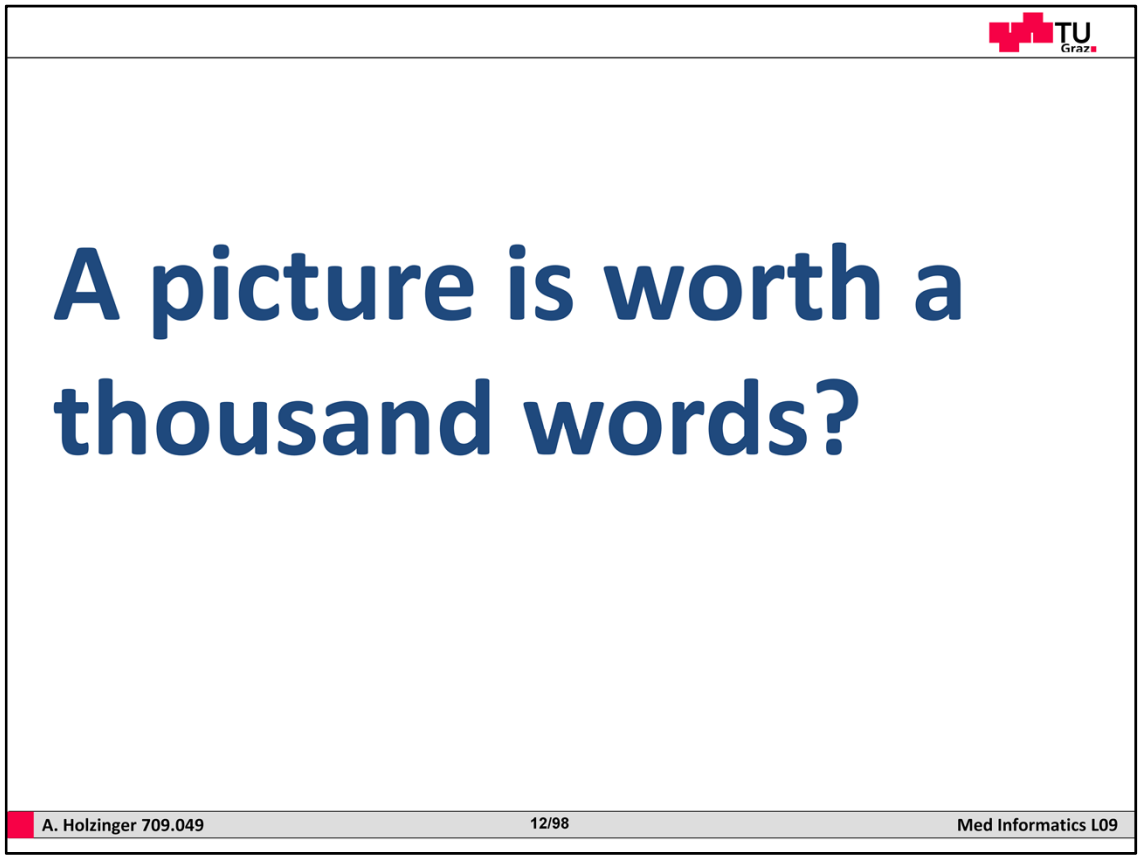
Problem: Context!

A. Holzinger 709.049 10/98 Med Informatics L09

And exactly this poses a grand challenge to computational approaches, because Von-Neumann machines are missing the context! A computer does not know that noses can run – and feet can smell – however, one solution lies in machine learning approaches where we can train the machines to learn the context.

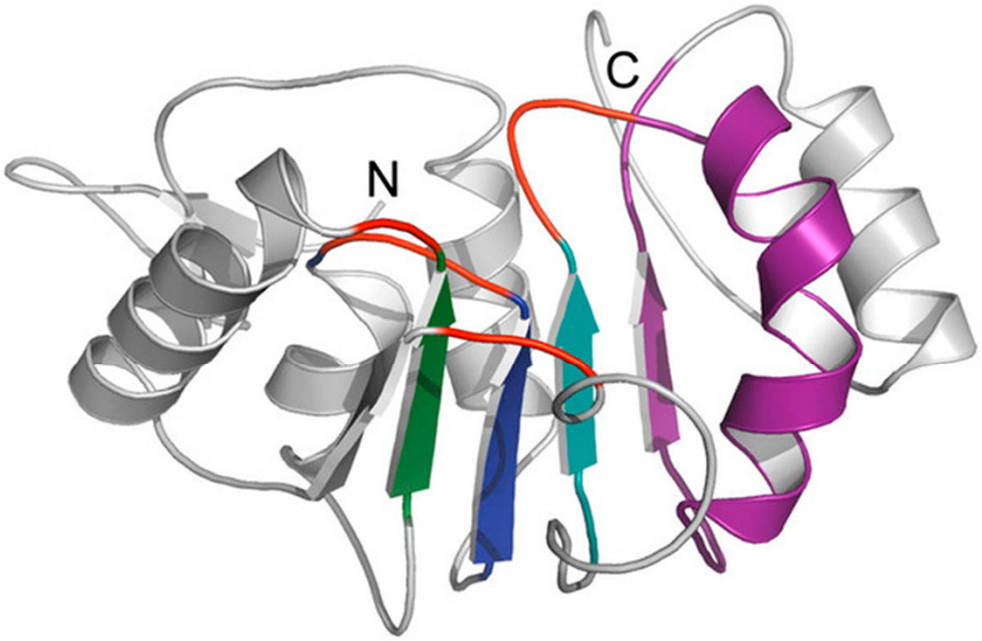


The word "Cell" has a lot of different meanings: the famous Journal, but also the basic building block of life, a battery cell, a Voronoi cell in mathematical topology, a prisoner's cell, a cell of a radio network, a blood cell, a cell of a spreadsheet, an cell in aircrafts or car manufacturing, a foam cell, cellulose (in German: "Zell-Stoff"), etc. This is the most difficult problem: the semantic ambiguity of our natural language (noses can run and feet can smell – without context this is unsolvable for any computer). To better understand these processes let us review more detailed the already learned human information processing.



The answer is: it depends!

Slide 9-7: Example: Ribbon Diagram of a Protein Structure



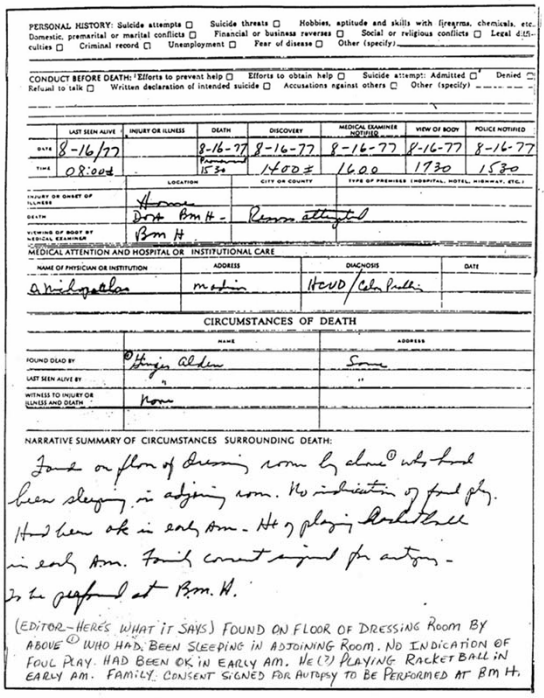
Magnani, R., et al. 2010. Calmodulin methyltransferase is an evolutionarily conserved enzyme that trimethylates Lys-115 in calmodulin. *Nature Communications, 1, 43.*

A. Holzinger 709.049 13/98 Med Informatics L09

The famous proverb "a picture is worth a thousand words" refers to the concept that a complex idea can be conveyed with just one single image and infers a central goal of visualization: to make it possible to perceive and cognitively process large amounts of data quickly. Look at this image. It is a good example on how a picture can explain a complex idea: A **ribbon diagram aka Richardson diagram,** (Richardson, 2000), is a standard method of schematic protein representation. The ribbon shows the overall path and organization of the protein backbone and is generated by interpolating a smooth curve through the polypeptide backbone. So-called α-helices are shown as curly ribbons, β-strands as arrows, and thin lines for non-repetitive coils or loops. The direction of the polypeptide chain is shown locally by the arrows, and may be indicated overall by a color ramp along the length of the ribbon. Such diagrams are useful for expressing the molecular structure (twist, fold and unfold). Remember: A protein is a single chain of amino acids, which folds into a globular structure and the Thermodynamics Hypothesis states that a protein always folds into a state of minimum energy. Computationally, 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. First ribbon diagrams were hand drawn (Richardson, 2000), (Magnani et al., 2010).

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.
http://t3.gstatic.com/images?q=tbn:ANd9GcSGC3eO60_EdByeIfZEVGdNeAWXsQ4JtEOdEBvQ7DkbdUl_AmQe

Slide 9-8 "Is a picture really worth a thousand words?"



PERSONAL HISTORY: Suicide attempts ☐ Suicide threats ☐ Hobbies, aptitude and skills with firearms, chemicals, etc. Domestic, premarital or marital conflicts ☐ Financial or business reverses ☐ Social or religious conflicts ☐ Legal difficulties ☐ Criminal record ☐ Unemployment ☐ Fear of disease ☐ Other (specify) _____

CONDUCT BEFORE DEATH: Efforts to prevent help ☐ Efforts to obtain help ☐ Suicide attempt: Admitted ☐ Denied ☐ Refusal to talk ☐ Written declaration of intended suicide ☐ Accusations against others ☐ Other (specify) _____

	LAST SEEN ALIVE	INJURY OR ILLNESS	DEATH	DISCOVERY	MEDICAL EXAMINER NOTIFIED	VIEW OF BODY	POLICE NOTIFIED
DATE	8-16-77		8-16-77	8-16-77	8-16-77	8-16-77	8-16-77
TIME	08:00±		1530	1400±	1600	1730	1530

	LOCATION	CITY OR COUNTY	TYPE OF PREMISES (HOSPITAL, HOTEL, HIGHWAY, ETC.)
INJURY OR ONSET OF ILLNESS	Home		
DEATH	Brm H.	Bronx	Hospital
VIEWING OF BODY BY MEDICAL EXAMINER	Brm H		

MEDICAL ATTENTION AND HOSPITAL OR INSTITUTIONAL CARE

NAME OF PHYSICIAN OR INSTITUTION	ADDRESS	DIAGNOSIS	DATE
Dr Michaelakos	Bronx	HCVD / Coronary	

CIRCUMSTANCES OF DEATH

	NAME	ADDRESS
FOUND DEAD BY	Othmar Aldinger	Same
LAST SEEN ALIVE BY	"	"
WITNESS TO INJURY OR ILLNESS AND DEATH	None	

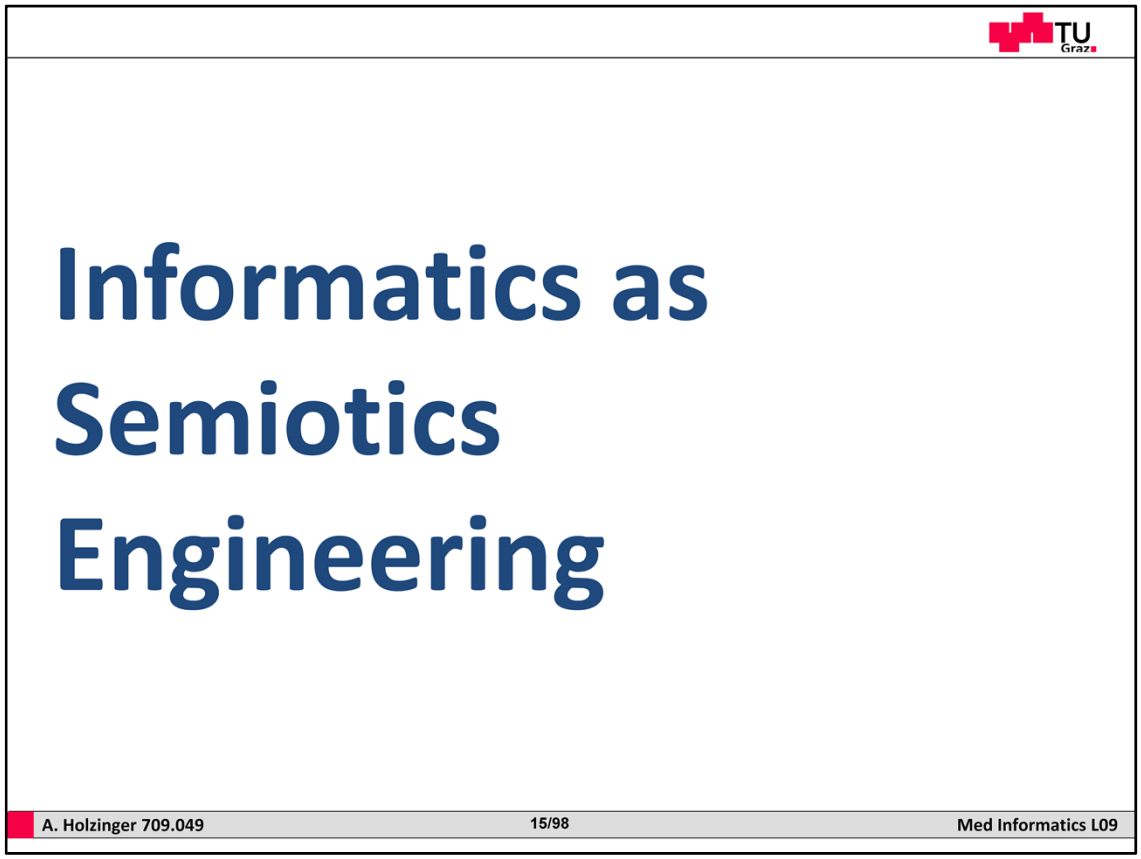
NARRATIVE SUMMARY OF CIRCUMSTANCES SURROUNDING DEATH:

Found on floor of dressing room by cleaner who had been sleeping in adjoining room. No indication of foul play. Had been OK in early Am. He (?) playing Racketball in early Am. Family consent signed for autopsy to be performed at Brm H.

(EDITOR – HERE'S WHAT IT SAYS) FOUND ON FLOOR OF DRESSING ROOM BY ABOVE ⓒ WHO HAD BEEN SLEEPING IN ADJOINING ROOM. NO INDICATION OF FOUL PLAY. HAD BEEN OK IN EARLY AM. HE (?) PLAYING RACKETBALL IN EARLY AM. FAMILY CONSENT SIGNED FOR AUTOPSY TO BE PERFORMED AT BM H.

A. Holzinger 709.049 14/98 Med Informatics L09

Whether and to what extent the proverb above is true is a long debate and there is no clear evidence to date. The best answer to the question "Is a picture worth a thousand words?" is: **"it depends!".** Some researchers are arguing that sometimes a picture might be worth a billion words (Michel et al., 2011), whereas others are arguing that sometimes text is better than an image.
A more profane example: what is the difference between tortoise and turtle? If you are not good in animal biology, you maybe have problems with the words, but if you look at a picture showing a turtle and at one showing a tortoise, you immediately understand that one is a land animal and the other a sea animal.
Be aware, that most information in Hospitals is only available in text format, and that text is the communication media in patient findings, and the amount of this unstructured data is immensely increasing (Holzinger et al., 2008), (Holzinger et al., 2013). Consequently, text mining is a huge area of biomedical informatics (refer to →Lecture 6).



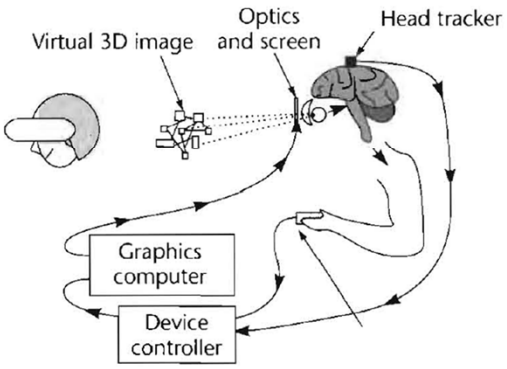
Semiotics := the study of signs and symbols as a significant part of communication. As different from linguistics, however, semiotics also studies non-linguistic sign systems. Semiotics is often divided into three branches:
Semantics: relation between signs and the things to which they refer; their signified denotata, or meaning
Syntactics: relations among or between signs in formal structures
Pragmatics: relation between signs and sign-using agents or interpreters

Slide 9-9 Three examples for Visual Languages

Ware, C. (2004) *Information Visualization: Perception for Design (Interactive Technologies) 2nd Edition. San Francisco, Morgan Kaufmann.*

$$W = \{w : w \in T(v)\}$$

$$H = -\sum_{w \in W} p(w) \log_2 p(w)$$



Holzinger, A., Searle, G., Auinger, A. & Ziefle, M. (2011) Informatics as Semiotics Engineering: Lessons learned from Design, Development and Evaluation of Ambient Assisted Living Applications for Elderly People. *Universal Access in Human-Computer Interaction. Context Diversity. Lecture Notes in Computer Science (LNCS 6767). Berlin, Heidelberg, New York, Springer, 183-192.*

A. Holzinger 709.049 16/98 Med Informatics L09

Computer Science lacks a reliable concept of the human mind, whereas the psychological science lacks solid concepts for algorithms and data structures; consequently, there is a need for a theory in which both domains find a place (Andersen, 2001). A sign ("Zeichen") integrates two sides: physical (=signifier) and psychological (=signified).
Semiotics is the study of signs and therefore can talk about representations (algorithms and data structures as signifiers) and the interpretation by the end user (domain concepts as the signified). However, only those parts of the computational processes that influence the interpretation, and only those parts of the interpretations that are influenced by the computation, may be analyzed by semiotic methods (Holzinger et al., 2011b).

In Figure 9-4 we see three examples of languages, which claim to be visual:
1) Cave paintings (= images), which can be directly interpreted;
2) Schematic diagram, showing a virtual environment and the human-computer interaction on a fairly abstract level;
3) Expression of a mathematical equation, that is on a highly abstract level;

Slide 9-10 Informatics as Semiotics Engineering

- 1. Physical: is it present?
  - Signals, traces, components, points, …
- 2. Empirical: can it be seen?
  - Patterns, entropy, codes, …
- 3. Syntactic: can it be read?
  - Formal structure, logic, deduction, …
- 4. Semantic: can it be understood?
  - Meaning, proposition, truth, …
- 5. Pragmatic: is it useful?
  - Intentions, negotiations, communications, …
- 6. Social: can it be trusted?
  - Beliefs, expectations, culture, …

Burton-Jones, A., Storey, V. C., Sugumaran, V. & Ahluwalia, P. 2005. A semiotic metrics suite for assessing the quality of ontologies. *Data & Knowledge Engineering, 55, (1), 84-102.*

A. Holzinger 709.049 17/98 Med Informatics L09

Semiotic framework stamper 1973, 1996
Semiotics is the study of signs and therefore can describe representations (algorithms and data structures as signifiers) and the interpretation by the end user (domain concepts as the signified). However, only those parts of the computational process that influence the interpretation, and only those parts of the interpretations that are influenced by the computation, can be analyzed by semiotic methods (Holzinger et al., 2011), (Holzinger, 2002a). Semiotics can be divided into three branches:
Syntactics: Relations among signs in formal structures;
Semantics: Relations between signs and their meaning; and
Pragmatics: Relations between signs, and the effects these may have on the end users who use them.
A relatively new field is Biosemiotics, which is a synthesis of biology and semiotics, and studying the origins, action and interpretation of signs and biological codes (Barbieri, 2008): Life is essentially about creating new organic codes and conserving those which have been created (macroevolution). For example, biosemiotics claims that language has biological roots and must be studied as a natural phenomenon, not following the divide between nature and culture. Or another example: The study of protein synthesis has revealed that genes and proteins are not formed spontaneously in the cell but are manufactured by a system of molecular machines based on RNAs. In 1981, the components of this manufacturing system were called ribosoids and the system itself was given the collective name of ribotype. The cell was described in this way as a structure made of genes, proteins and ribosoids, i.e., as a trinity of genotype, phenotype and ribotype (Barbieri, 2010).
A different example: Burton-Jones et al. (2005) proposed metrics to assess the quality of an ontology by drawing upon semiotic theory; their metrics assess the syntactic, semantic, pragmatic, and social aspects of the quality of an ontology.

Slide 9-11 Definitions of the term "Visualization"

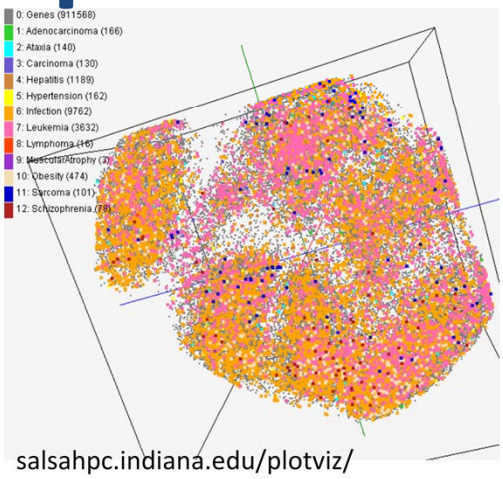
- **Visualization** = generally a method of computer science to **transform the symbolic into the geometric**, to form a mental model and foster unexpected insights;
- **Information visualization** = the interdisciplinary study of the visual representation of large-scale collections of non-numerical data, such as files and software, databases, networks etc., to allow users to see, explore, and understand information at once;
- **Data visualization** = visual representation of complex data, to communicate information clearly and effectively, making data useful and usable;
- **Visual Analytics** = focuses on analytical reasoning of complex data facilitated by **interactive** visual interfaces;
- **Content Analytics** = a general term addressing so-called "unstructured" information – mainly text – by using mixed methods from visual analytics and business intelligence;

A. Holzinger 709.049 18/98 Med Informatics L09

Visualization is a method of computer science to transform the symbolic into the geometric, to support the formation of a mental model and foster insights; as such it is an essential component of the knowledge discovery process (refer to →Lecture 6, Slide 6-3), (2007) .
Information visualization is the interdisciplinary study of the visual representation of large-scale collections of non-numerical data, such as files and software, databases, networks etc., to allow users to see, explore, and understand information at once (Ware, 2004), (Ware, 2012);
Data visualization is the visual representation of complex data, to communicate information clearly and effectively, making data useful and usable (Ward, Grinstein & Keim, 2010);
Visual Analytics focuses on analytical reasoning of complex data facilitated by interactive visual interfaces (Aigner, Bertone & Miksch, 2007);
Content Analytics is a general term addressing so-called "unstructured" data – mainly text – by using mixed methods from visual analytics and business intelligence (Holzinger et al., 2013);

Visualization is a typical HCI topic !

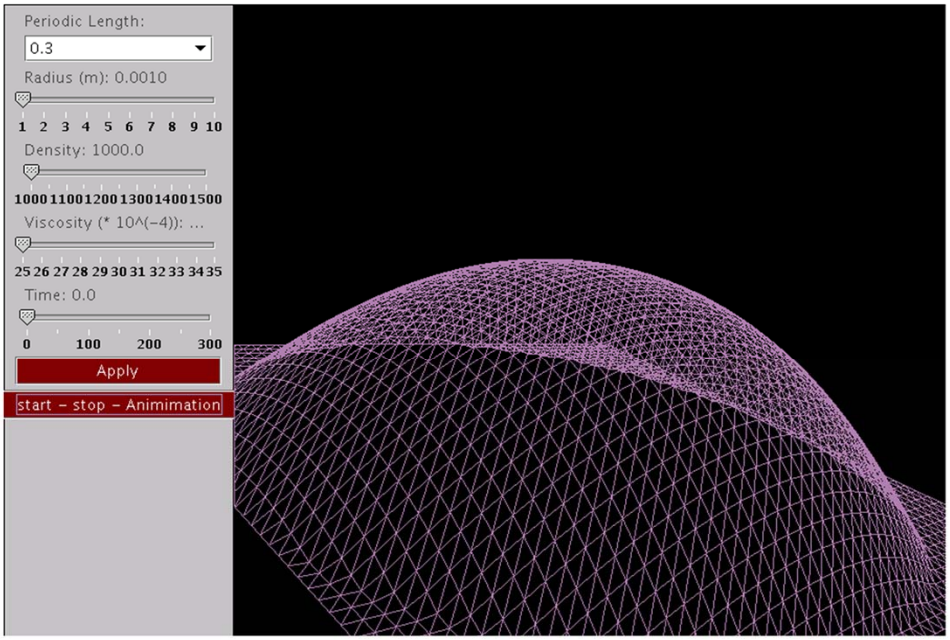
Jong Youl Choi, Seung-Hee Bae, Judy Qiu, Geoffrey Fox, Bin Chen, and David Wild, "Browsing Large Scale Cheminformatics Data with Dimension Reduction," Proceedings of Emerging Computational Methods for the Life Sciences Workshop of ACM HPDC 2010 conference, Chicago, Illinois, June 20-25, 2010.



A. Holzinger 709.049 19/98 Med Informatics L09

Large-scale high dimensional data visualization is highly valuable for scientific discovery in many fields of data mining and information retrieval. PlotViz is a 3D data point browser that visualizes large volume of 2- or 3-dimensional data as points in a virtual space on a computer screen and enable users to explore the virtual space interactively.

Slide 9-12 Process of interactive (data) visualization



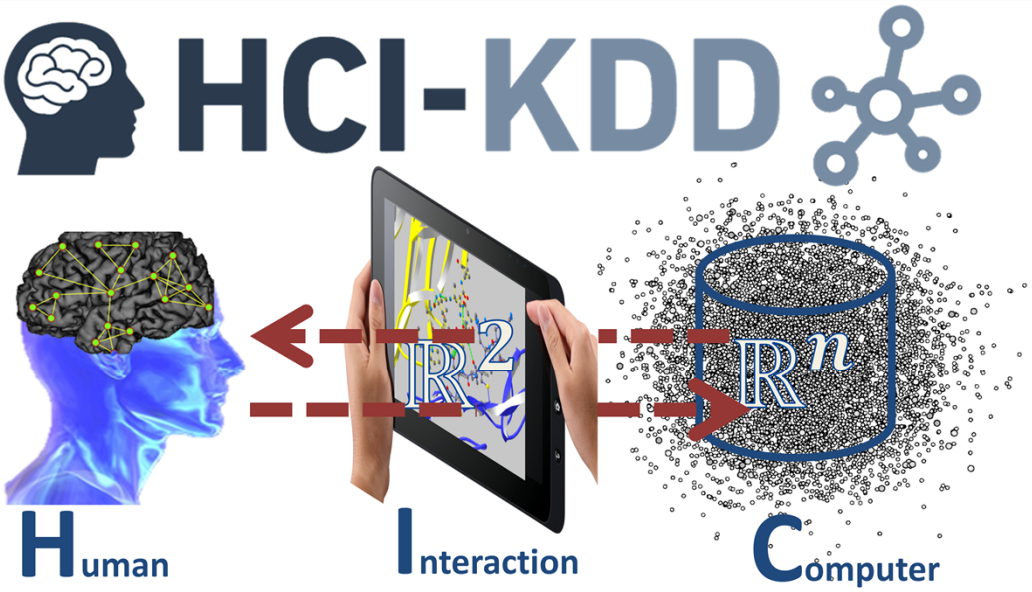
Holzinger, A., Kickmeier-Rust, M. D., Wassertheurer, S. & Hessinger, M. (2009) Learning performance with interactive simulations in medical education: Lessons learned from results of learning complex physiological models with the HAEMOdynamics SIMulator. *Computers & Education, 52, 2, 292-301.*

A. Holzinger 709.049 20/98 Med Informatics L09

Making issues visible which otherwise would be unaccessible.
Interactive Visualizations provide the ability to comprehend data and to interactively analyze information properties. The process of data visualization includes four steps
1) the data itself and interactive data exploration;
2) the preprocessing and transformation of the data;
3) the graphical algorithms to produce the corresponding image on a screen
4) the human perceptual and cognitive system;

McCormick (1987) defined the science of visualization by a taxonomy diagram, wherein he stated that images and signals (captured from cameras, sensors etc.) are transformed by image processing and presented pictorially. Abstractions of these visual representations can then be transformed by computer vision to create symbolic representations in the form of symbols and structures. Finally, by using computer graphics the symbols or structures can be synthesized into visual representations. McCormick concluded that the common denominator of the computational sciences is visualization and indeed the research opportunities and engineering applications of visualization are sheer endless.
In this example we see, how an interactive visualization can enhance student understanding of complex data (Holzinger et al., 2009): Generally, learning in the area of physiology is difficult for medical students for several reasons. Medical students often do not have sufficient mathematical background for understanding physiological models and the dynamics of complex mathematical rules related to these models. Moreover, learning is often without recourse to patients due to ethical restrictions (Simon, 1972). Simulations are assumed to offer various benefits, especially to novice medical students learning theoretical concepts, processes, relationships, as well as invasive procedural skills, which is extremely important within decreasing clinical exposure. Consequently, students can acquire knowledge in a safe environment (Kneebone, 2005) and apply the new knowledge in practice (Weller, 2004).

Slide 9-13 Visualization is a typical HCI topic!



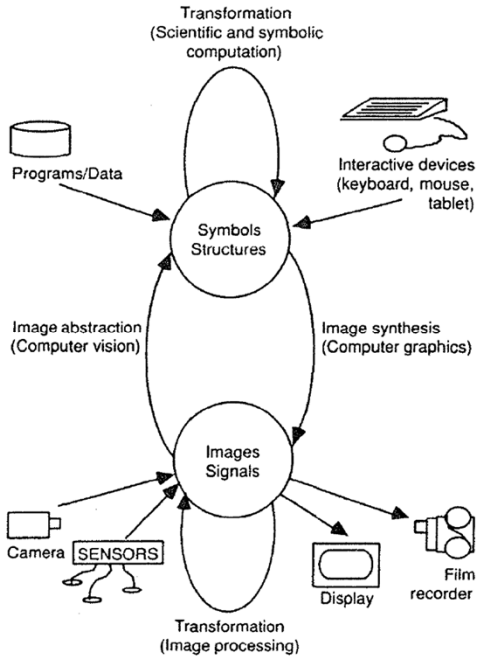
Holzinger, A. 2013. Human–Computer Interaction & Knowledge Discovery (HCI-KDD): What is the benefit of bringing those two fields to work together? *In: Alfredo Cuzzocrea, C. K., Dimitris E. Simos, Edgar Weippl, Lida Xu (ed.) Multidisciplinary Research and Practice for Information Systems, Springer Lecture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

A. Holzinger 709.049 21/98 Med Informatics L09

Taking all these considerations into account interactive visualization is a typical human-computer interaction task.
However, to solve problems in the complex medical domain we need to cover the whole pipeline from data preprocessing to visualization.
Our group logo derives from this horizontal concerted and integrated approach.

Slide 9-14 We can conclude that Visualization is …

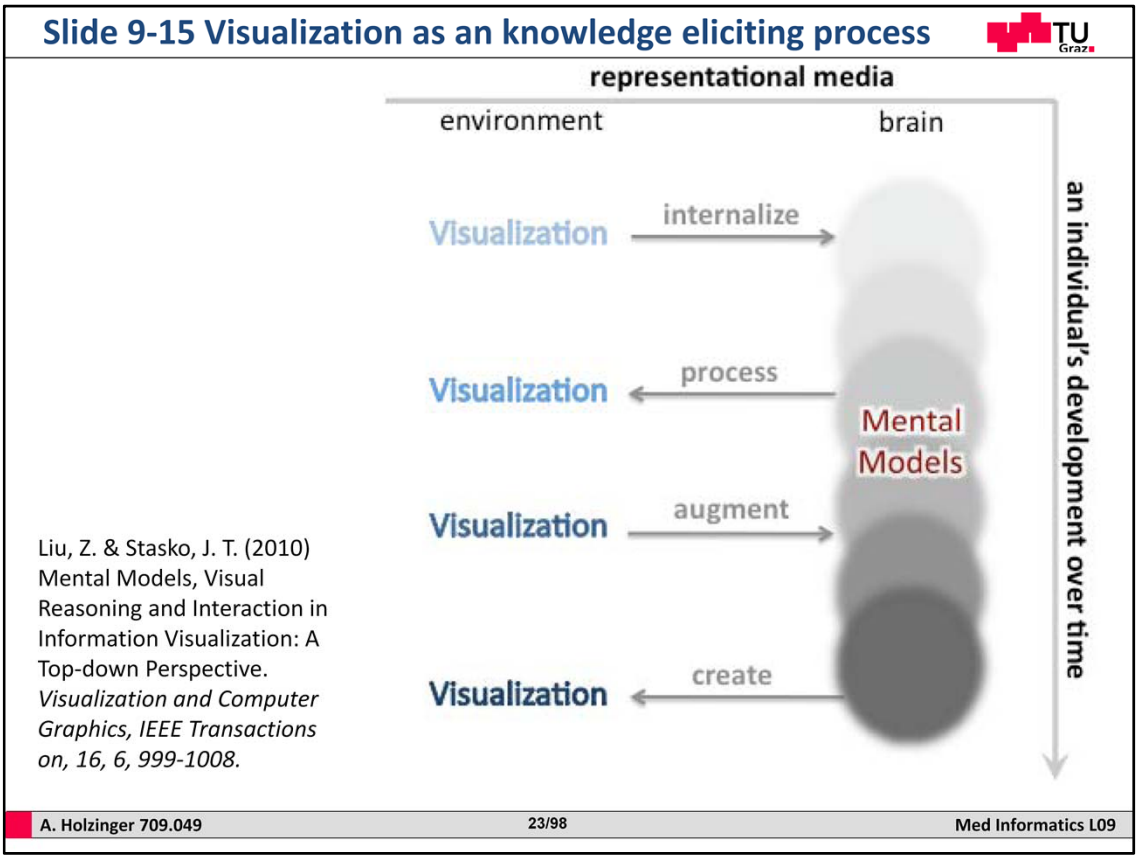
- … the common denominator of Computational sciences
- … the transformation of the symbolic into the geometric
- … the support of human perception
- … facilitating know-ledge discovery in data



McCormick, B. (1987) Scientific and Engineering Research Opportunities. *Computer graphics, 21, 6.*

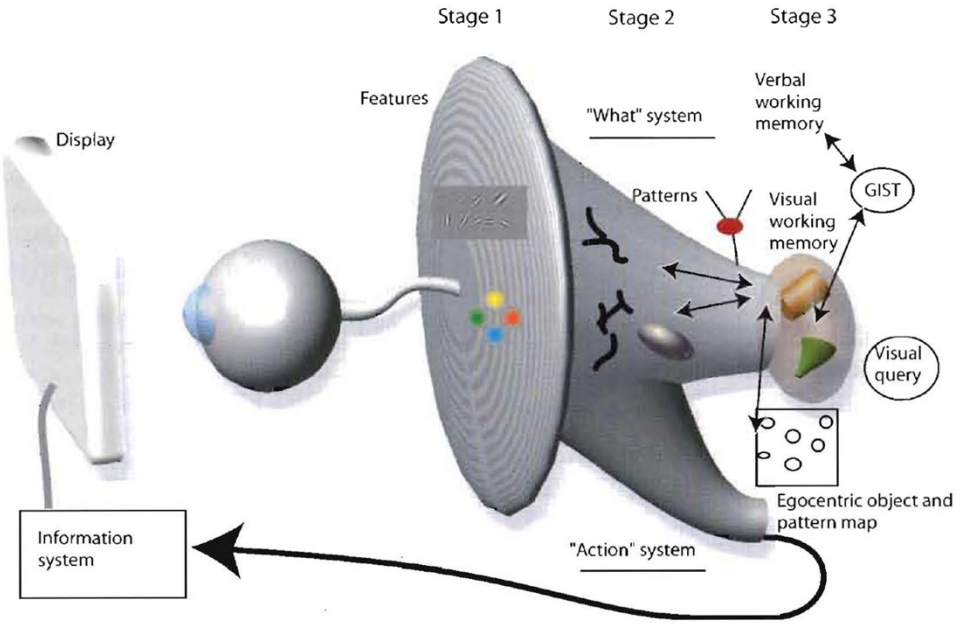
A. Holzinger 709.049 22/98 Med Informatics L09

McCormick, 1987 defined the science of visualization by a taxonomy diagram (see in the slide right), wherein he stated that images and signals (captured from cameras, sensors etc.) are transformed by image processing and presented pictorially. Abstractions of these visual representations can then be transformed by computer vision to create symbolic representations in the form of symbols and structures. Finally, by using computer graphics the symbols or structures can be synthesized into visual representations. McCormick concluded that the common denominator of the computational sciences is visualization and indeed the research opportunities and engineering applications of visualization are sheer endless.



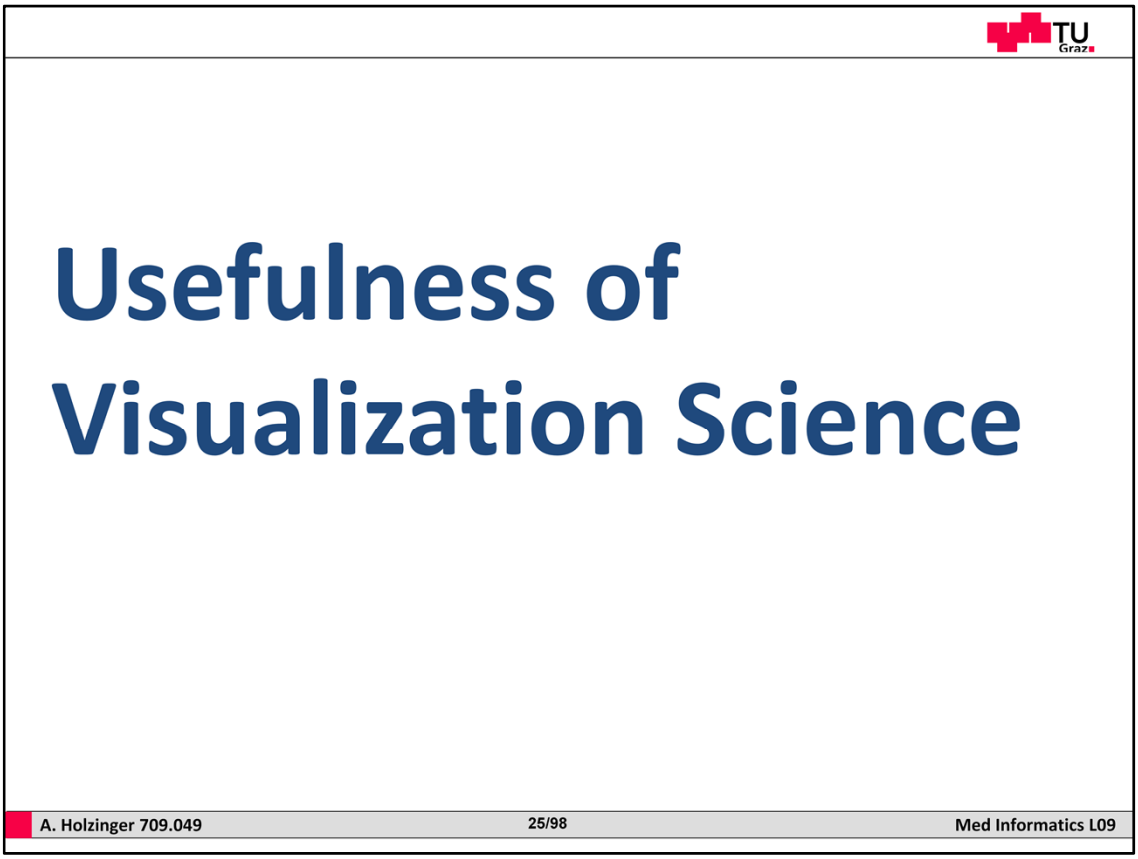
Mental Models can be seen as abstractions of visualizations. The first step between mental models and external representations is internalization. The formation of a mental model happens ontogenetically after the appearance of the original external phenomenon. This insight has been formulated by Vygotsky, who argues that in each individual's development, every higher order cognitive function appears twice: first between people, as an inter-psychological process, then inside an individual, as an intra-psychological process. In information visualization it makes sense to understand the role of visualization in human cognitive activities from a developmental perspective. Mental models can serve as the cognitive basis of creativity and innovation. The construction and simulation of mental models can give rise to new concepts and designs including novel visual representations. Some visual representations such as the scatter plot and line graph have existed for centuries. In the slide we see that visualization is an interactive process (internalize, process, augment, create) across representational media (Liu & Stasko, 2010).

Slide 9-16 Model of Perceptual Visual Processing



Ware, C. (2004) *Information Visualization: Perception for Design (Interactive Technologies) 2nd Edition. San Francisco, Morgan Kaufmann.*

Stage 1: Parallel processing to extract low-level properties of the visual scene, i.e. billions of neurons in the eye and visual cortex work in parallel, extracting features from every part of the visual field simultaneously. The information at this stage is of transitory nature, briefly held in an iconic store (refer to memory models).
Stage 2: Pattern perception, the visual field is divided into regions and patterns (e.g. simple contours, regions of same colour, patterns of motion, etc.). The information at this stage is slowly serially processed in a state of flux.
Stage 3: Sequential goal-directed processing, here objects are held in the visual working memory by demands of active attention.
GIST =

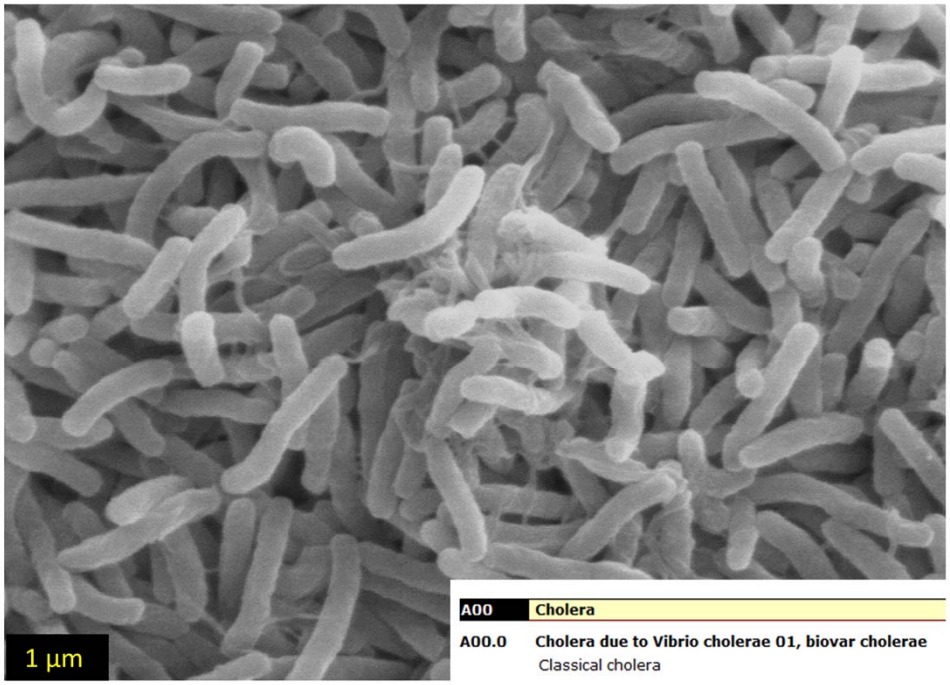


We demonstrate the usefulness of the visualization sciences on some examples.

Slide 9-17 A look back into history …

Next time you are in London – visit Broad Street
There are many famous visualization examples from the past, and one important for medicine is the example on the work done by John Snow in 1854 during the Cholera epidemic in London. He identified the Broad Street pump as the source of cholera by plotting the location of cholera deaths on a map (Figure 9-10). After he removed the pump handle the epidemic ended (McLeod, 2000).

What do you see in this picture?



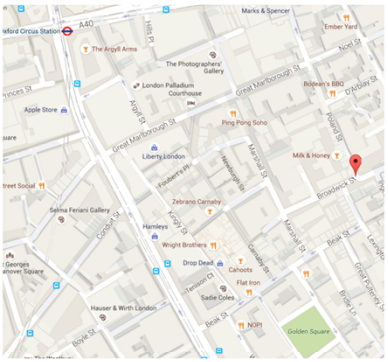
T.J. Kirn, M.J. Lafferty, C.M.P Sandoe and R.K. Taylor (2000) Delineation of pilin domains required for bacterial association into microcolonies and intestinal colonization, Molecular Microbiology, Vol. 35, 896-910

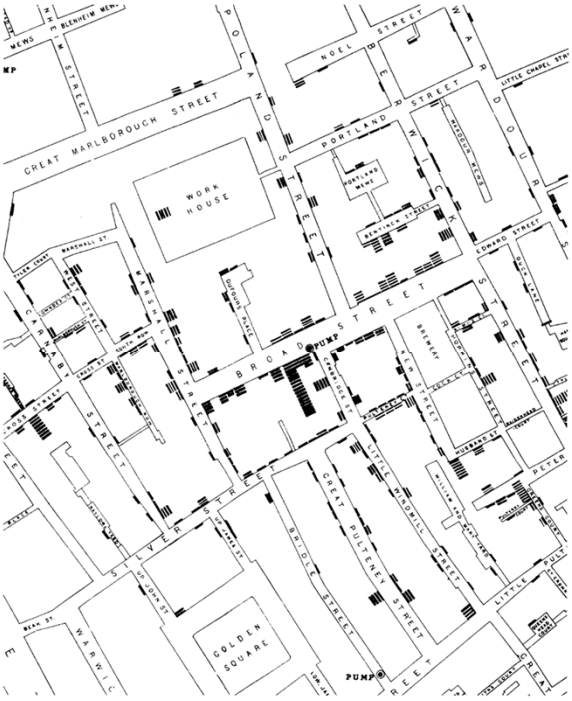
Cholera is an infection of the small intestine that is caused by the bacterium Vibrio cholerae. The main symptoms are profuse watery diarrhea and vomiting. Transmission occurs primarily by drinking or eating water or food that has been contaminated by the diarrhea of an infected person or the feces of an infected but asymptomatic person. The severity of the diarrhea and vomiting can lead to rapid dehydration and electrolyte imbalance and death in some cases. The primary treatment is with oral rehydration solution (ORS) to replace water and electrolytes, and if this is not tolerated or doesn't provide quick enough treatment, intravenous fluids can also be used. Antibiotics are beneficial in those with severe disease to shorten the duration and severity. Worldwide it affects 3–5 million people and causes 100,000–130,000 deaths a year as of 2010. Cholera was one of the earliest infections to be studied by epidemiological methods.

http://www.ahooy.net/un-says-cholera-epidemic-in-somalia/
Scanning electron microscope image of Vibrio cholerae bacteria, which infect the digestive system.

Zeiss DSM 962 SEM

Slide 9-18 Medical Visualization by John Snow (1854)





McLeod, K. S. (2000) Our sense of Snow: the myth of John Snow in medical geography. *Social Science & Medicine, 50, 7-8, 923-935.*

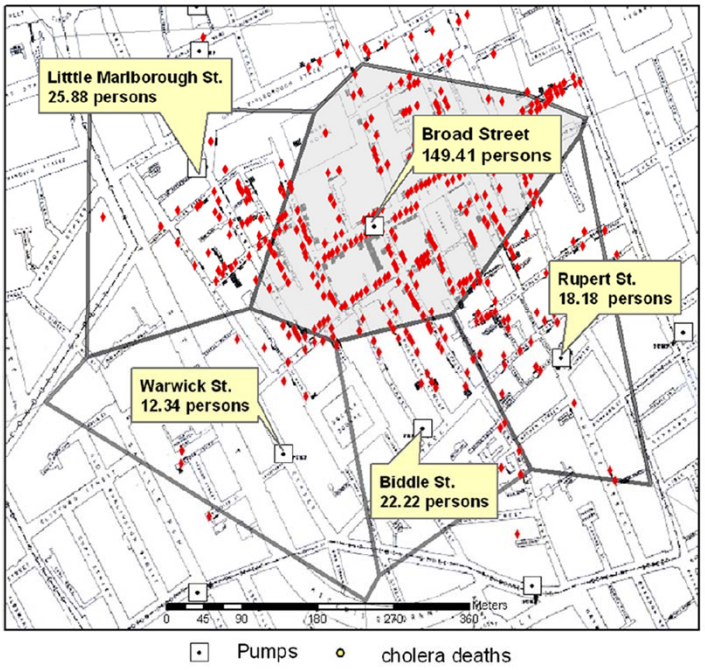
A. Holzinger 709.049 28/98 Med Informatics L09

In Sep 1854, a cholera epidemic hit an area of London around Broad (now Broadwick) Street – near Oxford Circus Underground

- Up to this time, cholera was thought to be an airborne disease.
- Dr. John Snow plotted the deaths on a map and noticed a higher clustering around the Broad Street water pump.
- Workers in the nearby brewery, which had its own water (and beer) supply, were largely unaffected.
- The handle on the Broad Street water pump was then removed.
- Snow had shown that cholera is in fact a waterborne disease.
- Tufte [1997b, pages 27–37] tells the story in more detail.
- Not really infovis, more geovis, since it is based on an underlying map.

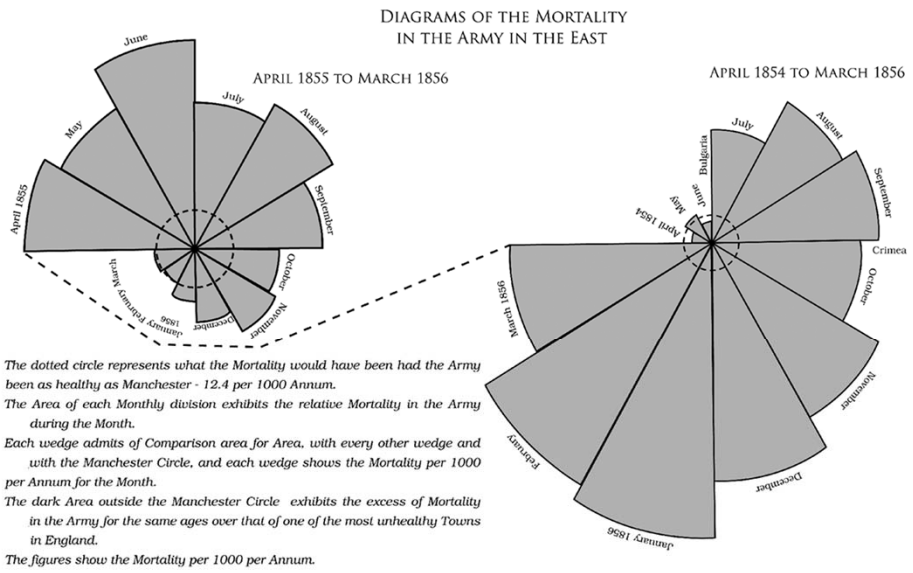
Slide 9-19 Systematic Visual Analytics > Content Analytics

Cholera Mortality per 1,000 persons



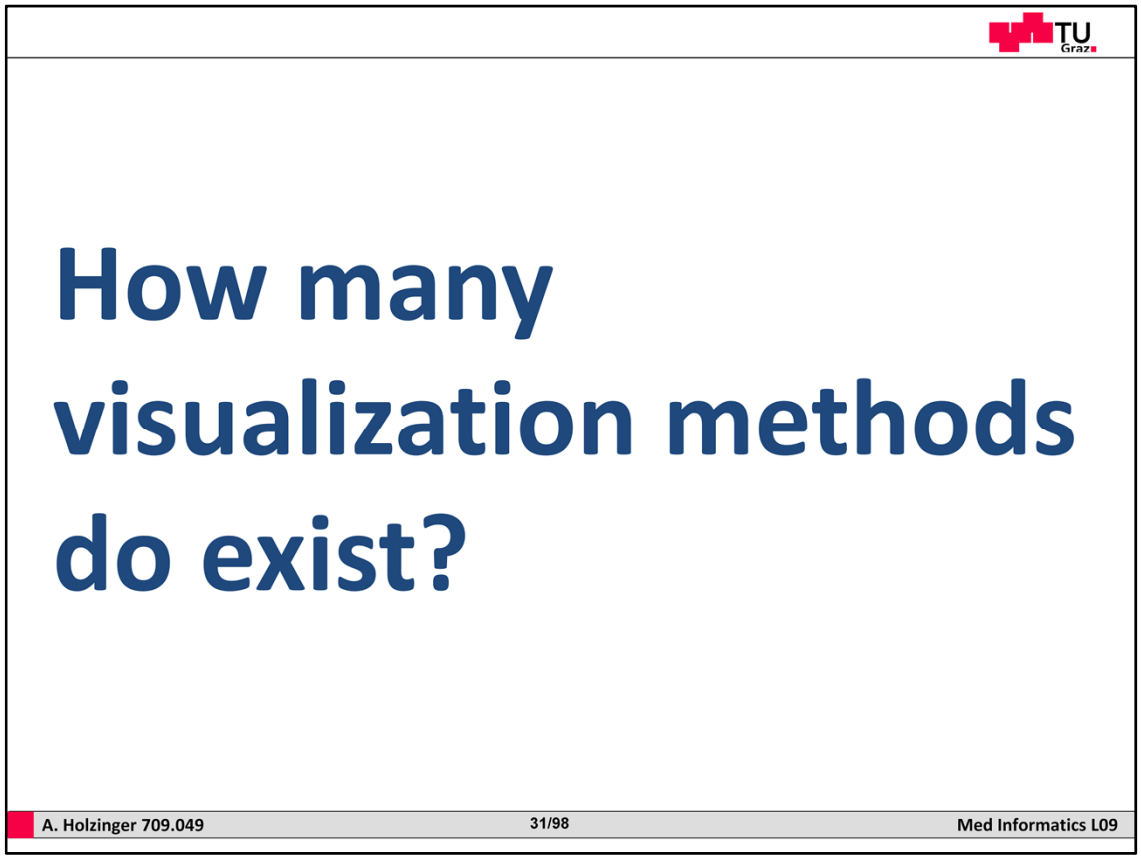
Koch, T. & Denike, K. (2009) Crediting his critics' concerns: Remaking John Snow's map of Broad Street cholera, 1854. *Social Science & Medicine, 69,* ***8, 1246-1251.***

Florence Nightingale – first medical quality manager



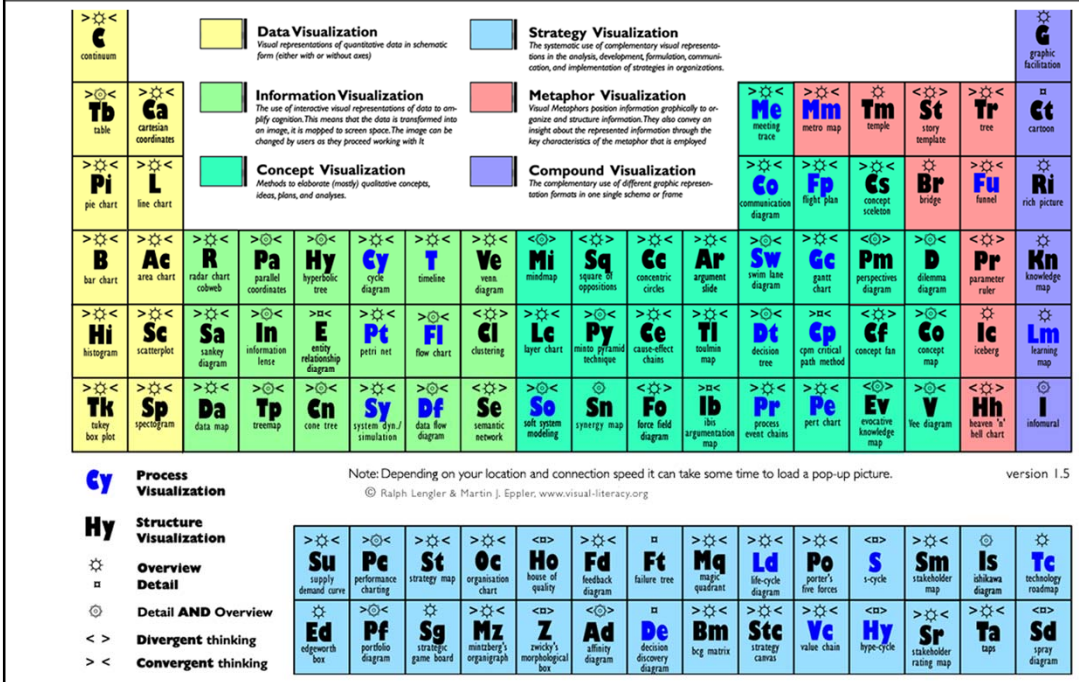
Meyer, B. C. & Bishop, D. S. (2007) Florence Nightingale: nineteenth century apostle of quality. *Journal of Management History, 13, 3, 240-254.*

W. Edwards Deming was once referred to as the "Apostle of Quality" for his role in promoting a philosophy of quality (Gabor, 1990) and he is widely credited with being a major reason for the popularity of the total quality management movement of the 1980s and 1990s. One biography titles Deming "The Man Who Discovered Quality" (Gabor, 1990). If that title is apt, then perhaps we could label Florence Nightingale "The woman who discovered quality." One hundred years before Deming went to Japan, Nightingale went to the Crimean peninsula. In her subsequent work and writings she proved to be a relentless advocate of quality in the operations function. Because her context was the medical profession, she has become well-known in that arena, but her objectives and methods align in many ways with those of the modern quality movement. We believe that Florence was, like Deming, an apostle of quality and deserves a prominent place in quality history.


How many visualization methods do exist?
A. Holzinger 709.049
31/98
Med Informatics L09

Many ☺

Slide 9-20 A periodic table of visualization methods



Lengler, R. & Eppler, M. J. (2007) Towards a periodic table of visualization methods for management. *Proceedings of Graphics and Visualization in Engineering (GVE 2007);* Online: www.visual-literacy.org

Many concepts do exist!

The periodic table of the chemical elements is a tabular form of displaying the chemical elements in categories, first devised in 1869 by the Russian chemist Dmitri Mendeleev; a similar approach to categorize visualization methods was done by (Lengler & Eppler, 2007),
accessible online: www.visual-literacy.org/periodic_table/periodic_table.html
They have subdivided the application area dimension ("groups") into the following categories and distinguished them by background color – see next slide.

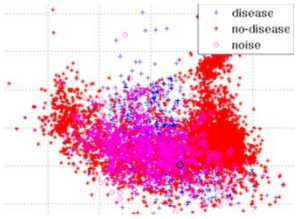
http://www.visual-literacy.org/periodic_table/periodic_table.html

Slide 9-21: A taxonomy of Visualization Methods

- 1) Data Visualization (Pie Charts, Area Charts or Line Graphs, …
- 2) Information Visualization (Semantic networks, tree-maps, radar-chart, …)
- 3) Concept Visualization (Concept map, Gantt chart, PERT diagram, …)
- 3) Metaphor Visualization (Metro maps, story template, iceberg, …)
- 4) Strategy Visualization (Strategy Canvas, roadmap, morpho box,…)
- 5) Compound Visualization

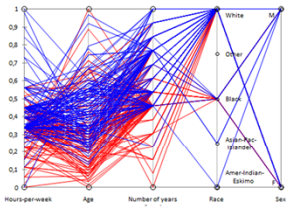
1) Data Visualization includes standard quantitative formats such as Pie Charts, Area Charts or Line Graphs. They are visual representations of quantitative data in schematic form (either with or without axes), they are all-purpose, mainly used for getting an overview of data. They have mapped them to the Alkali Metals which most easily form bonds with non-metals, a correspondence might be the combination between data visualization (answering "how much" questions) and visual metaphors (answering how and why questions).
2) Information Visualization, such as semantic networks or tree-maps, is defined as the use of interactive visual representations of data to amplify cognition. This means that the data is transformed into an image; it is mapped to screen space. The image can be changed by users as they proceed working with it.
3) Concept Visualization, such as a concept map or a Gantt chart; these are methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses through the help of rule-guided mapping procedures. In Concept Visualization knowledge is usually presented in a 2-D graphical display where concepts (usually represented within boxes or circles), connected by directed arcs encoding brief relationships (linking phrases) between pairs of concepts. These relationships usually consist of verbs, forming propositions or phrases for each pair of concepts.
3) Metaphor Visualization, such as metro maps or story template can be used as effective and simple templates to convey complex insights. Visual Metaphors fulfill a dual function, first they position information graphically to organize and structure it. Second they convey an insight about the represented information through the key characteristics of the metaphor that is employed.
4) Strategy Visualization, such as a Strategy Canvas or technology roadmap is defined "as the systematic use of complementary visual representations to improve the analysis, development, formulation, communication, and implementation of strategies in organizations." This is the most specific of all groups, as it has achieved great relevance in management.
5) Compound Visualization consists of several of the aforementioned formats. They can be complex knowledge maps that contain diagrammatic and metaphoric elements, conceptual cartoons with quantitative charts, or wall sized infomurals. This label thus typically designates the complementary use of different graphic representation formats in one single schema or frame. According to Tufte they result from two (or more) spatially distinct different data representations, each of which can operate independently, but can be used together to correlate information in one representation with that in another.

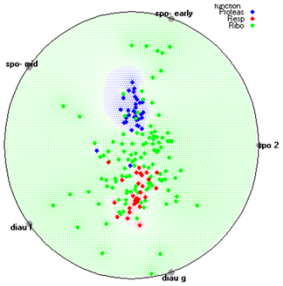
Slide 9-22 Visualizations for multivariate data Overview 1/2



Scatterplot = oldest, point-based technique, projects data from n-dim space to an arbitrary k-dim display space;

Parallel coordinates = (PCP), originally for the study of high-dimensional geometry, data point plotted as polyline;



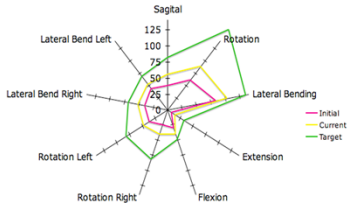


RadViz = Radial Coordinate visualization, is a “force-driven” point layout technique, based on Hooke’s law for equilibrium;

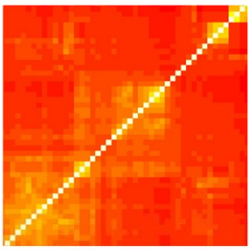
A. Holzinger 709.049 34/98 Med Informatics L09

Slide 9-22 Visualizations for multivariate data Overview 1/2
For a first overview let us summarize some important visualization methods: Scatterplots (SP) are the oldest, point-based techniques, and projects (maps) data from an n-dimensional space into an arbitrary k-dimensional display space (usually it will be the 2-dimensional space ;-). To verify cluster separation in high-dimensional data, analysts often reduce the data with a dimension reduction technique, and then visualize it with 2D Scatterplots, interactive 3D Scatterplots, or Scatterplot Matrices (SPLOMs) (Sedlmair, Munzner & Tory, 2013).
Parallel Coordinates (PCP) is best suited for the study of high-dimensional geometry, where each data point is plotted as a polyline.
Radial Coordinate Visualization (RadViz) is a “force-driven” point layout technique, based on Hooke’s law for equilibrium.

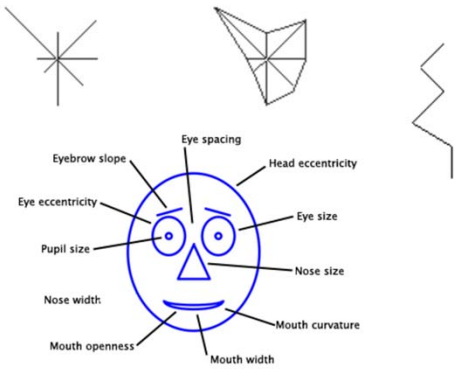
Slide 9-23 Visualizations for multivariate data Overview 2/2



Radar chart (star plot, spider web, polar graph, polygon plot) = radial axis technique;



Heatmap = a tabular display technique using color instead of figures for the entries;



Glyph = a visual representation of the entity, where its attributes are controlled by data attributes;

Chernoff face = a face glyph which displays multivariate data in the shape of a human face

A. Holzinger 709.049 35/98 Med Informatics L09

Slide 9-23 Visualizations for multivariate data Overview 2/2
Radar Chart aka star plot, spider web, polar graph, polygon plot is a radial axis technique.
Heatmap is a tabular display technique using color instead of figures for the entities.
Glyph is a visual representation of the entity, where its attributes are controlled by data attributes.
Chernoff face is a face glyph which displays multivariate data in the shape of a human face.
http://www.mathworks.de/de/help/stats/glyphplot.html

Biology heat maps are typically used in molecular biology to represent the level of expression of many genes across a number of comparable samples (e.g. cells in different states, samples from different patients) as they are obtained from DNA microarrays

Parallel Coordinates

The so called ||-coords have been developed in the context of modern visualization by Alfred Inselberg in the 1950ies and are excellent for visualizing high-dimensional and multivariate data in the form of N parallel lines, where a data point in the N-dimensional space is represented as a polyline with vertices on the parallel axes.

Slide 9-24 Parallel Coordinates – multidim. Visualization

- On the plane with Cartesian-coords, a vertical line, labeled $\overline{X}_i$ is placed at each $x = i - 1$ for $i = 1, 2, \ldots N$.
- These are the axes of the parallel coordinate system for $\mathbb{R}^N$.
- A point $C = (c_1, c_2, \ldots c_N) \in \mathbb{R}^N$ is mapped into the polygonal line $\overline{C}$
- the *N*-vertices with *xy-coords* $(i - 1, c_i)$ are now on the parallel axes.
- In $\overline{C}$ the full lines and not only the segments between the axes are included.

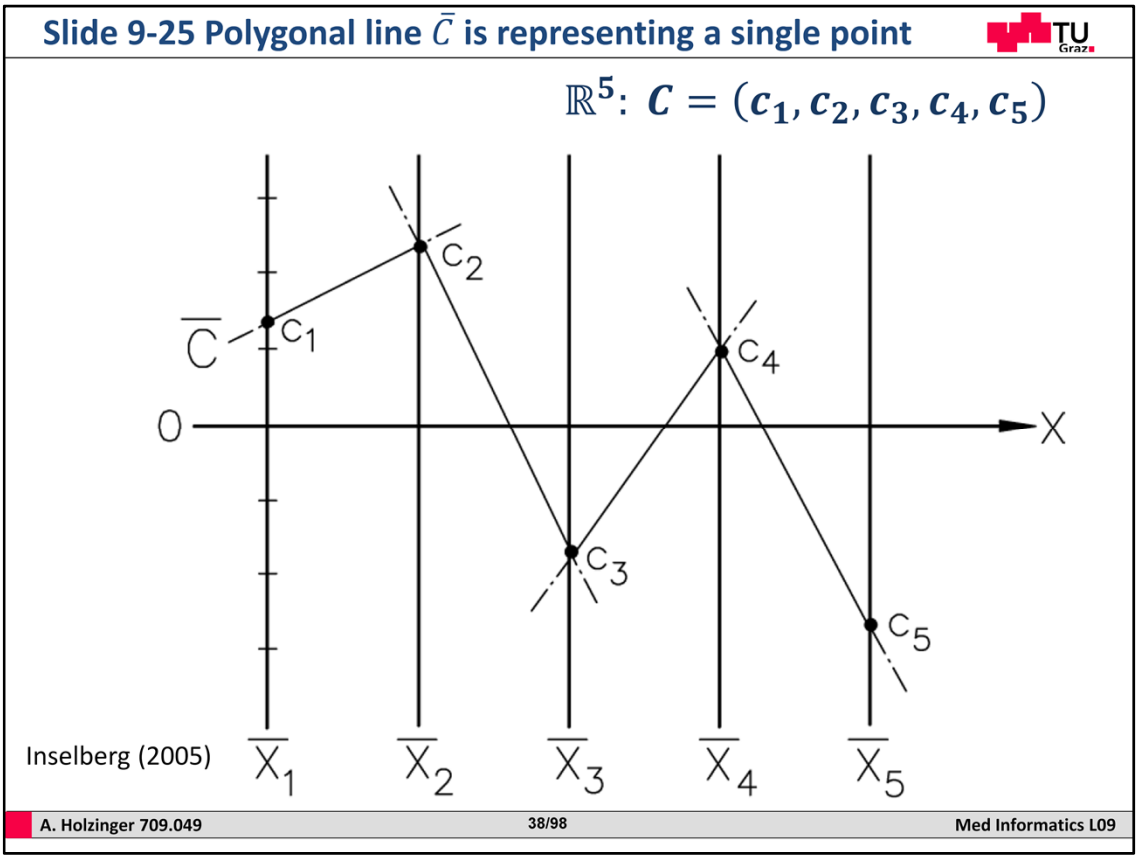
Inselberg, A. (2005) Visualization of concept formation and learning. *Kybernetes: The International Journal of Systems and Cybernetics, 34, 1/2, 151-166.*

A. Holzinger 709.049 37/98 Med Informatics L09

The so called ||-coords have been developed in the context of modern visualization by Alfred Inselberg in the 1950ies and are excellent for visualizing high-dimensional and multivariate data in the form of N parallel lines, where a data point in the N-dimensional space is represented as a polyline with vertices on the parallel axes. We follow the paper of (Inselberg, 2005) and the book of (Inselberg, 2009), which contains an excellent compact disc providing a lot of interactive material.

On the plane with xy-Cartesian coordinates a vertical line, labeled X¯_i is placed at each x=i-1 for i=1,2,…N. These are the axes of the parallel coordinate system for R^N. A point C=(c_1,c_2,… c_N)∈R^N is now mapped into the polygonal line C¯ whose N-vertices with xy-coords (i-1 , c_i) are on the parallel axes. In C¯ the full lines and not only the segments between the axes are included.

||-coords are constructed by placing axes in parallel with respect to the embedding 2D Cartesian coordinate system in the plane (the parallel-coordinates domain). While the orientation of axes can be chosen freely, the most common use is either horizontal (parallel to the x-axis) or vertical (parallel to the y-axis), see next slide.



On the plane with xy-Cartesian coordinates a vertical line, labeled X¯_i is placed at each x=i-1 for i=1,2,…N. These are the axes of the parallel coordinate system for R^N. A point C=(c_1,c_2,… c_N)∈R^N is mapped into the polygonal line C¯ whose N-vertices with xy-coords (i-1 , c_i) are on the parallel axes. In C¯ the full lines and not only the segments between the axes are included:
Note that each point in the n-dimensional space is represented as a polyline with vertices on the parallel axes; the position of the vertex on the i-th axis corresponds to the i-th coordinate of the point (Inselberg, 2005).

Slide 9-26 Heavier polygonal lines represent end-points

- A polygonal line $\overline{P}$ on the $N-1$ points represents a point
- $P = (p_1, \dots p_{i-1}, p_i \dots p_N) \in \ell$
- since the pair of values $\dots p_{i-1}, p_i$ marked on the $\overline{X}_{i-1}$ and $\overline{X}_i$ axes.
- In the following slide we see several polygonal lines, intersecting at $\ell_{(i-1),i}$
- representing data points on a line $\ell \subset \mathbb{R}^{10}$.
- Note: The indexing is essential and is important for the visualization of proximity properties such as the minimum distance between a pair of lines.

A. Holzinger 709.049 39/98 Med Informatics L09

A polygonal line P¯ on the N-1 points represents a point P=(p_1,⟦…p⟧_(i-1),p_i… p_N)∈l since the pair of values ⟦…p⟧_(i-1),p_i marked on the X¯_(i-1) and X¯_i axes. We can see several polygonal lines, intersecting at l_((i-1),i) representing data points on a line ⟦l⊂R⟧^10. If we plot this result we get the line interval in R^10 as can be seen in Slide 9-27.

Note: When thinking of visualizing data with ||-coords, especially in the biomedical domain, one is immediately confronted with some challenges (Heinrich & Weiskopf, 2012):

Overplotting occurs in parallel coordinates if lines potentially occlude patterns in the data.

The order of axes implicitly defines which patterns emerge between adjacent axes.

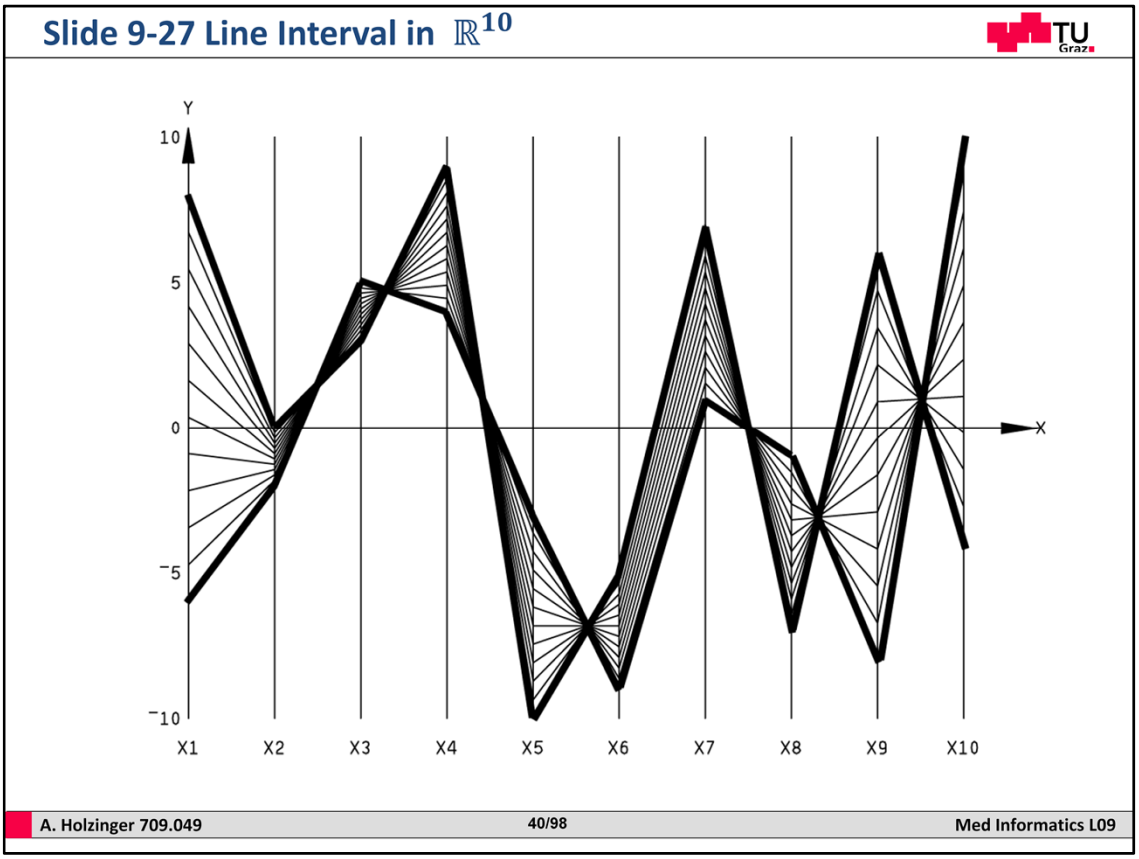
The line-tracing problem occurs if two or more lines intersect an axis at the same position.

Nominal and ordinal data such as sets and clusters have to be mapped to a metric scale before it can be visualized in parallel coordinates.

Time series are special in that time points, if interpreted as dimensions, have a fixed order.

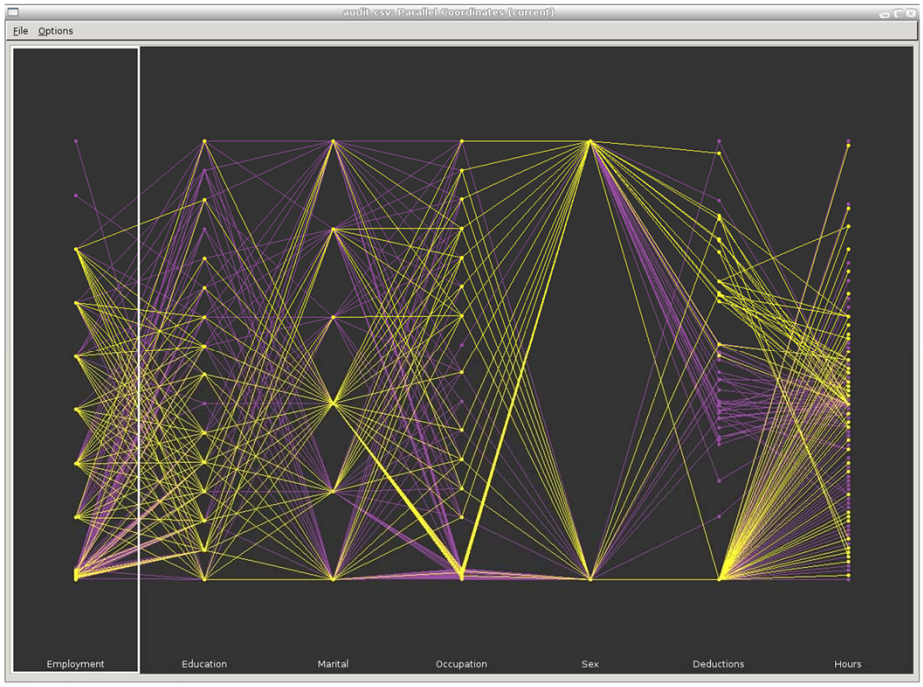
Uncertain data is another challenge for visualization, and there are approaches for the visualization of uncertainty in parallel coordinates.

The indexing is essential and is important for the visualization of proximity properties such as the minimum distance between a pair of lines, see next slide.



Again: The indexing is essential and is important for the visualization of proximity properties such as the minimum distance between a pair of lines.

Slide 9-28 Example: Par Coords in a Vis Software in R

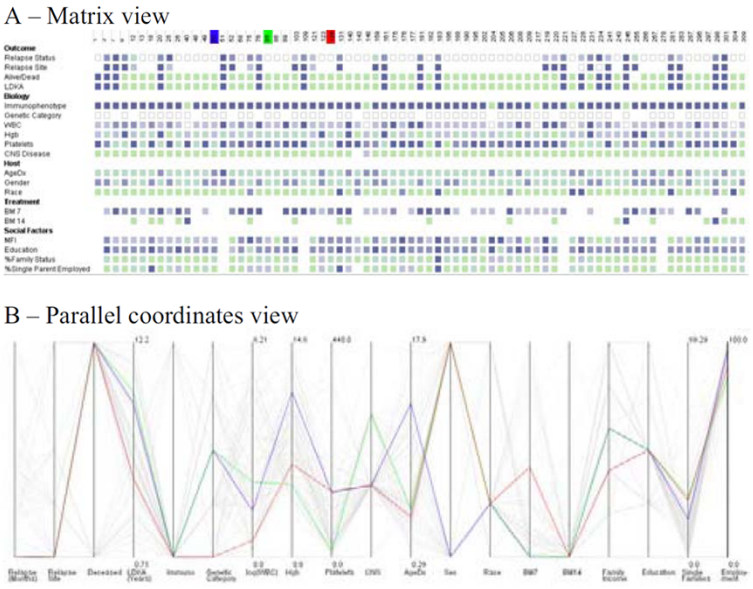


http://datamining.togaware.com

A. Holzinger 709.049 41/98 Med Informatics L09

This is an example of an implementation in R by Graham Williams, see:
http://datamining.togaware.com/survivor/Scatterplot.html

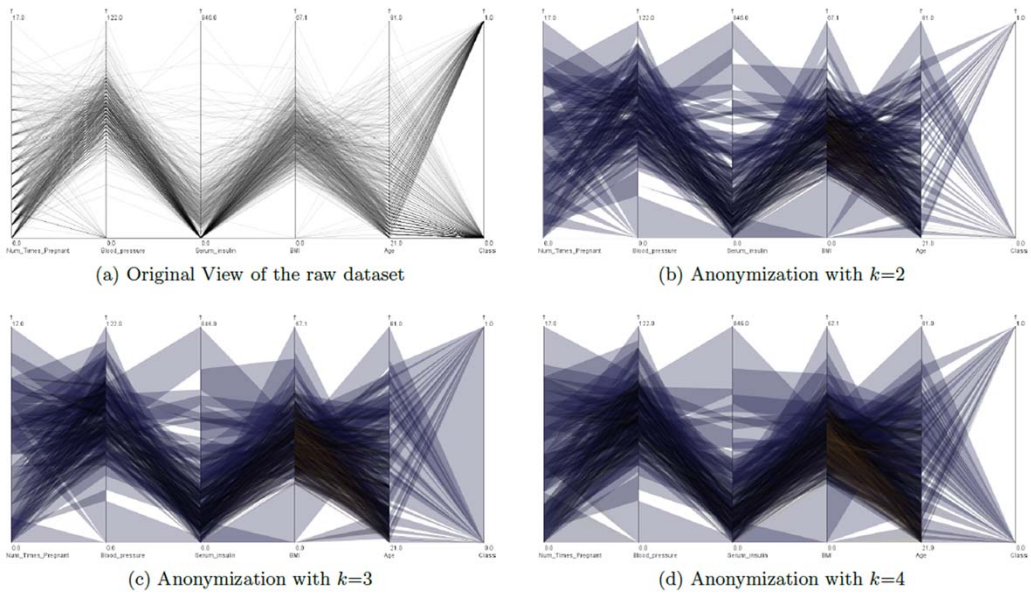
Slide 9-29 Par Coords -> Knowledge Discovery in big data



Mane, K. K. & Börner, K. (2007) Computational Diagnostic: A Novel Approach to View Medical Data. Los Alamos National Laboratory.

This shows an interesting example of (Mane & Börner, 2007): The multiple coordinated views help a medical practitioner to gain a good insight about the medical data variations among selected patients. The matrix view (A) helps to quickly identify similar patterns and worst case conditions. The parallel coordinates view (B) helps to quickly identify and compare trends shared by groups of patients. So the matrix view and parallel coordinate view complement each other to help the medical practitioners gain an understanding of the data.

Slide 9-30 Ensuring Data Protection with k-Anonymization



(a) Original View of the raw dataset

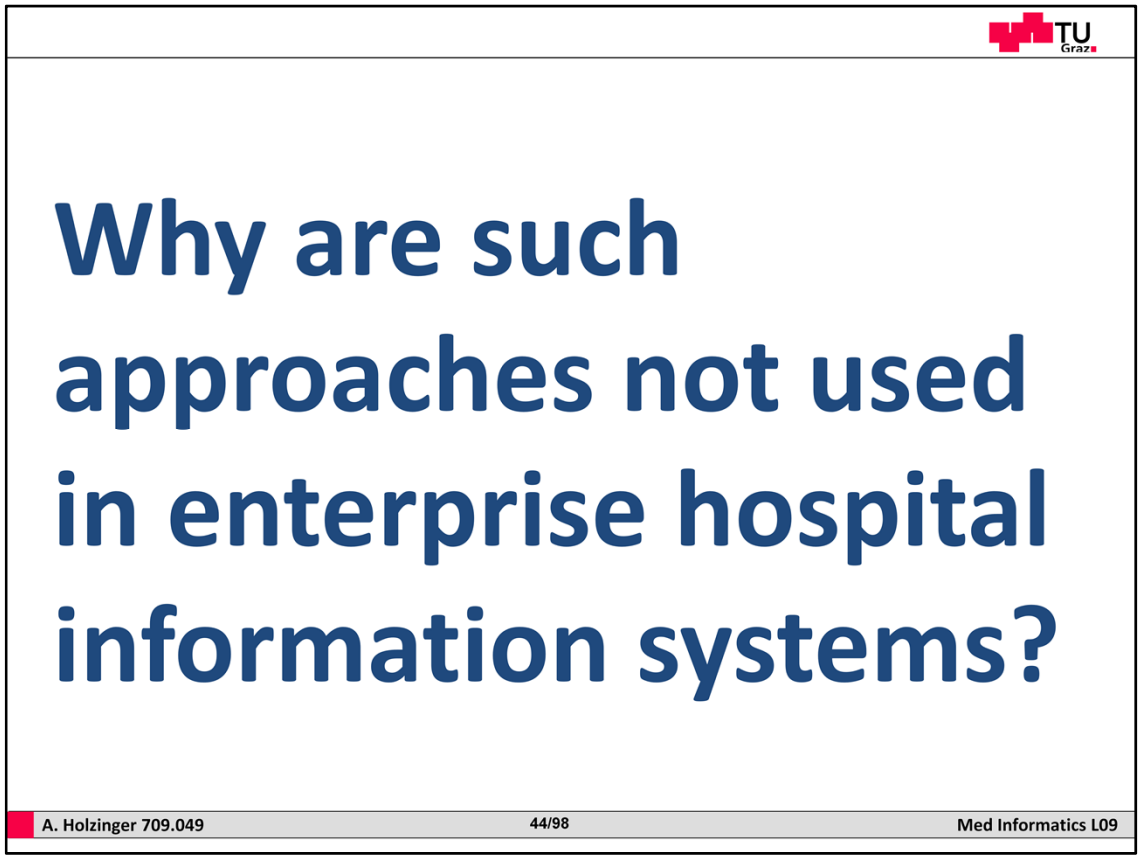
(b) Anonymization with k=2

(c) Anonymization with k=3

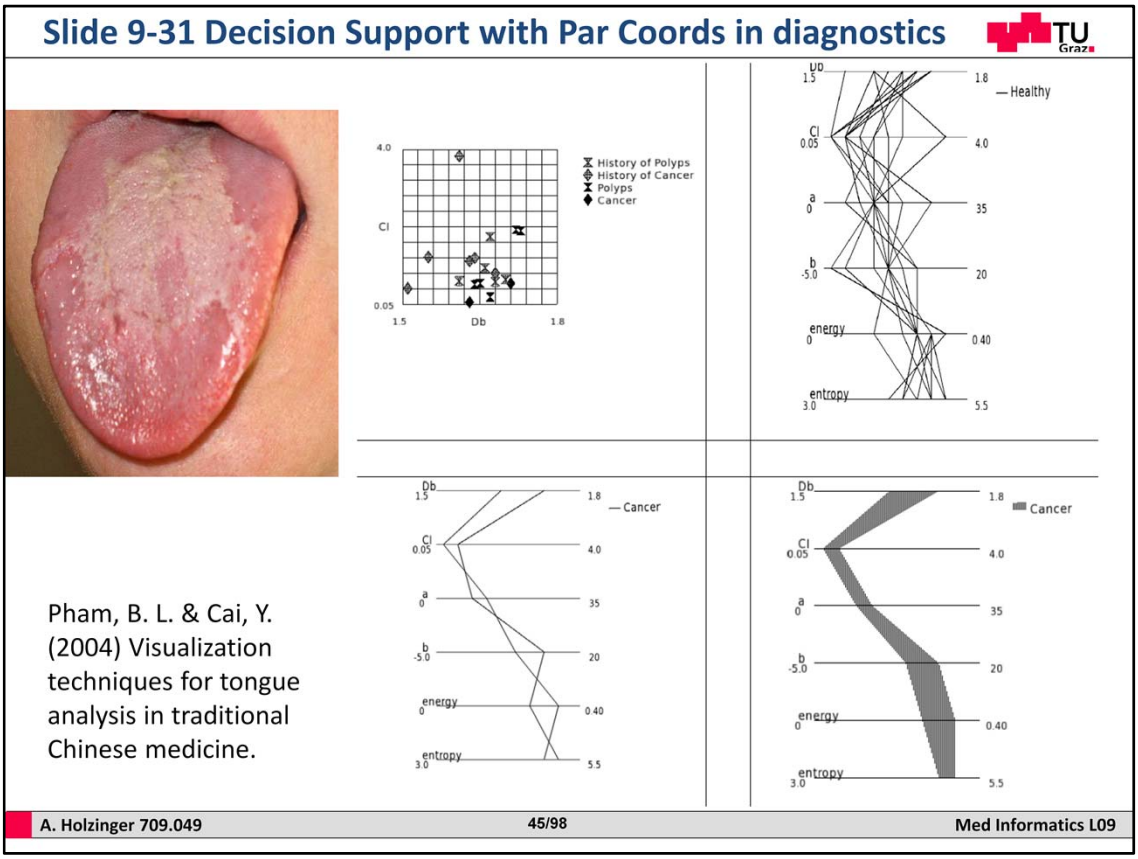
(d) Anonymization with k=4

Dasgupta, A. & Kosara, R. (2011). *Privacy-preserving data visualization using parallel coordinates. Visualization and Data Analysis 2011, San Francisco, SPIE.*

An important issue of all real-world datasets is, that many of their attributes can identify individuals, or the data are proprietary and valuable. The field of data mining has developed a variety of ways for dealing with such data, and has established an entire subfield for privacy-preserving data mining.
The k-anonymity model is an approach (NP-hard) to protect individual records from re-identification. It works by ensuring that each data record in the table is indistinguishable from k -1 other records with respect to the quasi-identifiers in the table (El Emam & Dankar, 2008).
Visualization has seen little work on handling sensitive data. With the growing applicability of data visualization in real-world scenarios, the handling of sensitive data has become a non-trivial issue we need to address in developing visualization tools. Figure 203 shows the work of (Dasgupta & Kosara, 2011): a) shows the parallel coordinates display of the original data, and b) to d) show the anonymized image for different values of k. Since we see aggregated values instead of single ones, represented by polygons instead of clusters, it is not possible to point to particular values on the axes. At the same time, much of the overall structure is still visible in the visualization, even though individual records cannot be identified anymore.



To date rarely such sophisticated visualization methods are used in the context of enterprise business hospital information systems. The question remains open: why?



Decision Support with Parallel Coordinates in Diagnostics. Visual inspection of the tongue is an important diagnostic method of Traditional Chinese Medicine (TCM). Observing the abnormal changes in the tongue and the tongue coating can support in diagnosing diseases. (Pham & Cai, 2004) demonstrate it nicely, how proper visualization can contribute to medical decision making. Also no interest – would be a nice student work – but nobody so far was interested – why?

Practical Example: Big data from Flow Cytometry (1)

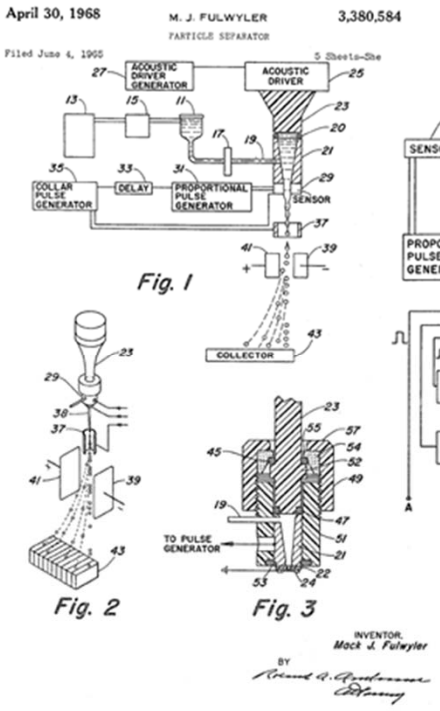
Source: Stem Cell Insititute, Online: http://www.cellmedicine.com

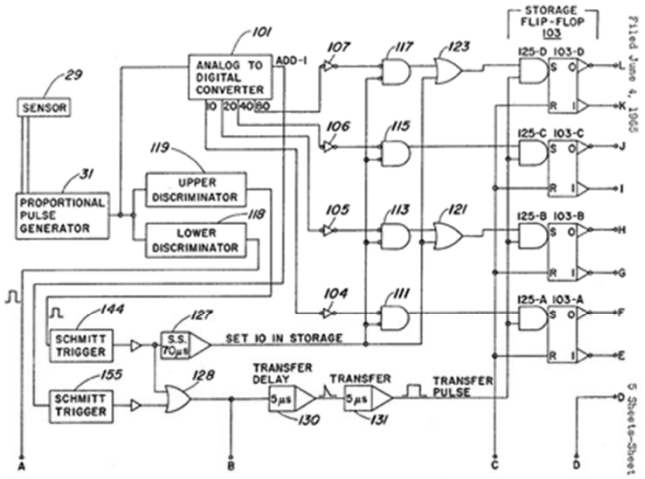
A. Holzinger 709.049 46/98 Med Informatics L09

Flow cytometry is used in diagnosis, especially blood cancers, but has many other applications in both research and clinical practice.

Flow cytometry is a laser based, biophysical technology employed in cell counting, sorting, biomarker detection and protein engineering, by suspending cells in a stream of fluid and passing them by an electronic detection apparatus. It allows simultaneous multiparametric analysis of the physical and/or chemical characteristics of up to thousands of particles per second.

Practical Example: Foundation of Flow Cytometry (2)





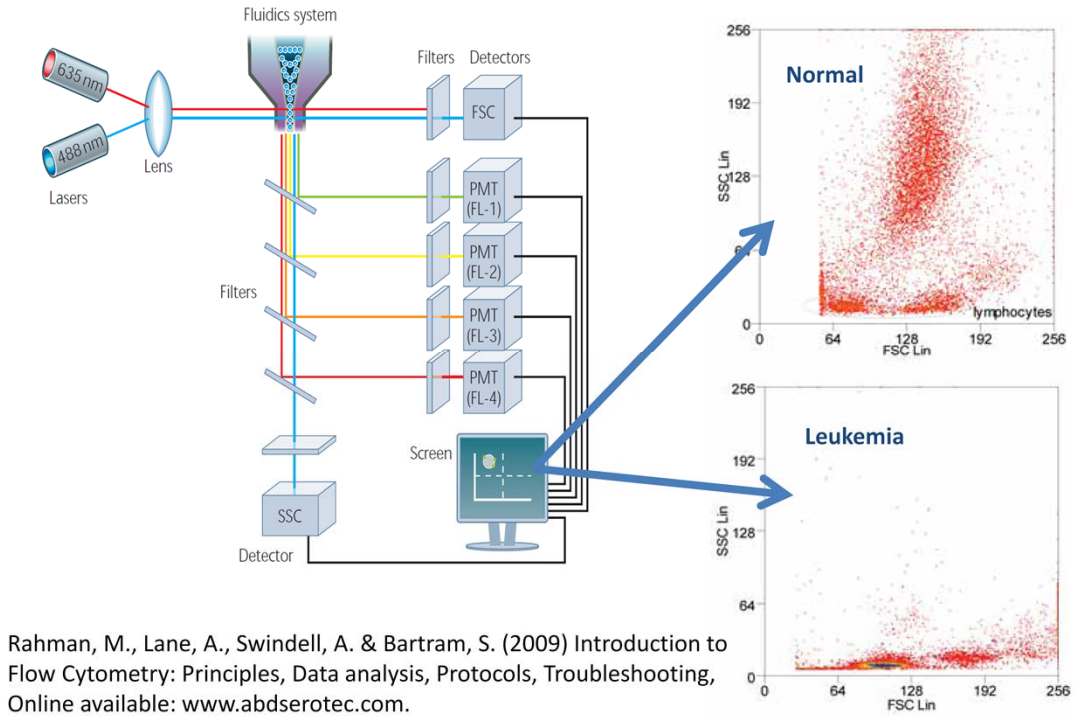
Fulwyler, M. J. (1968) US Patent 3380584 A
Particle Separator, 1965 applied, 1968 published

Fulwyler, M. J. (1965) Electronic Separation of Biological Cells by Volume. *Science, 150, 3698, 910-911.*

A. Holzinger 709.049 47/98 Med Informatics L09

Flow cytometry is routinely used in the diagnosis of health disorders, especially blood cancers, but has many other applications in basic research, clinical practice and clinical trials. A common variation is to physically sort particles based on their properties, so as to purify populations of interest.

Practical Example: Flow Cytometry (3) Immunophenotyping



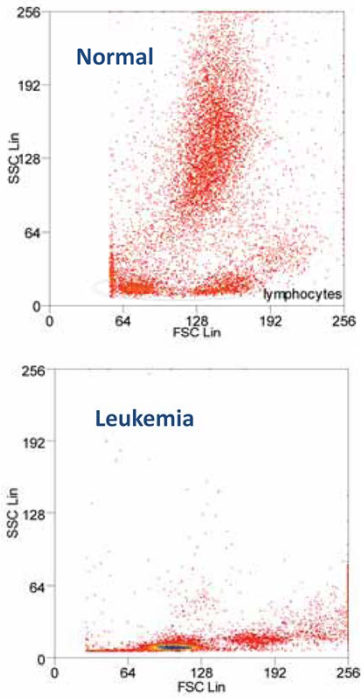
Rahman, M., Lane, A., Swindell, A. & Bartram, S. (2009) Introduction to Flow Cytometry: Principles, Data analysis, Protocols, Troubleshooting, Online available: www.abdserotec.com.

A. Holzinger 709.049 48/98 Med Informatics L09

Here you see typical scatterplot

Practical Example: Flow Cytometry (4) Immunophenotyping

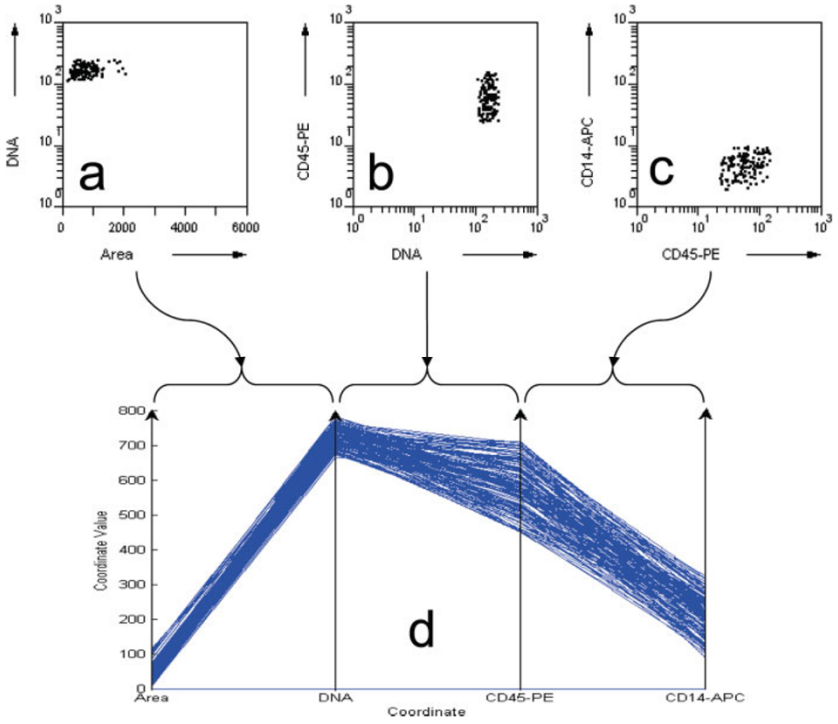
- Forward scatter channel (FSC) intensity equates to the particle's size and can also be used to distinguish between cellular debris and living cells.
- Side scatter channel (SSC) provides information about the granular content within a particle.
- Both FSC and SSC are unique for every particle, and a combination of the two may be used to differentiate different cell types in a heterogeneous sample.



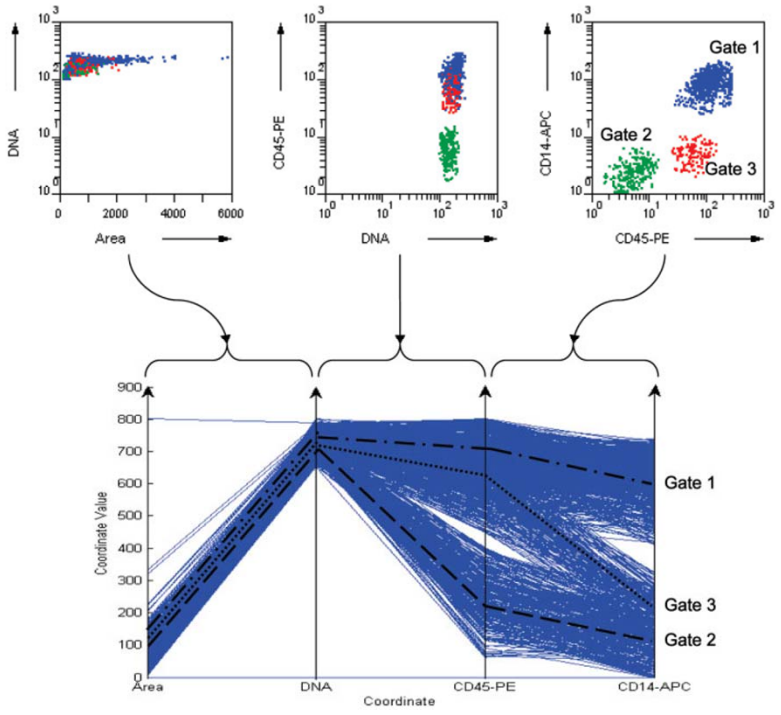
Rahman et al. (2009)

Example: 2D Parallel Coordinates in Cytometry

Streit, M., Ecker, R. C., Österreicher, K., Steiner, G. E., Bischof, H., Bangert, C., Kopp, T. & Rogojanu, R. (2006) 3D parallel coordinate systems—A new data visualization method in the context of microscopy-based multicolor tissue cytometry. *Cytometry Part A, 69A, 7, 601-611.*

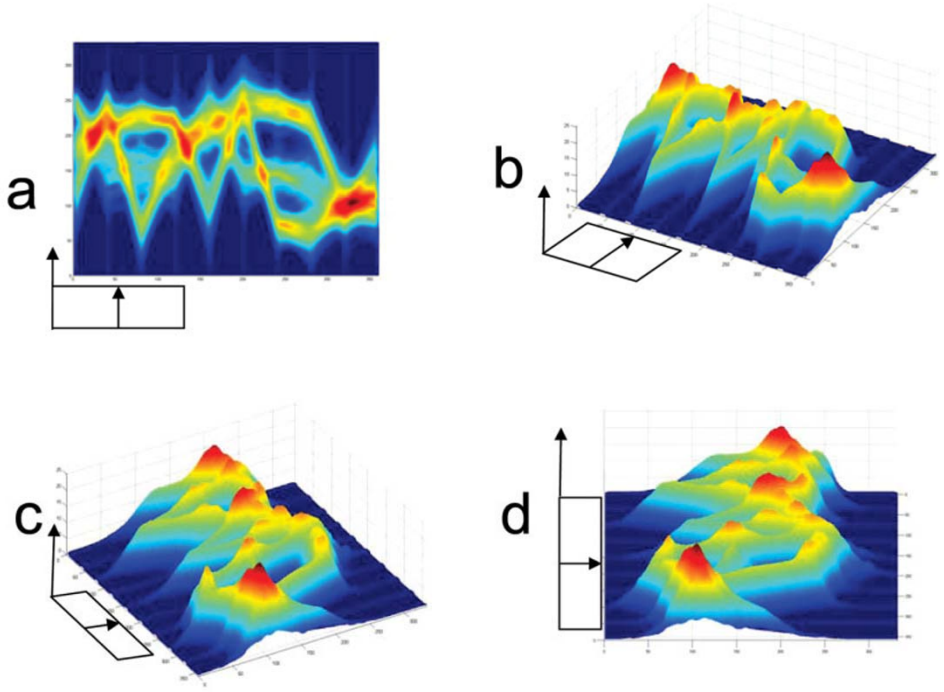


Example: Limitations of 2D Parallel Coordinates



Streit et al. (2006)

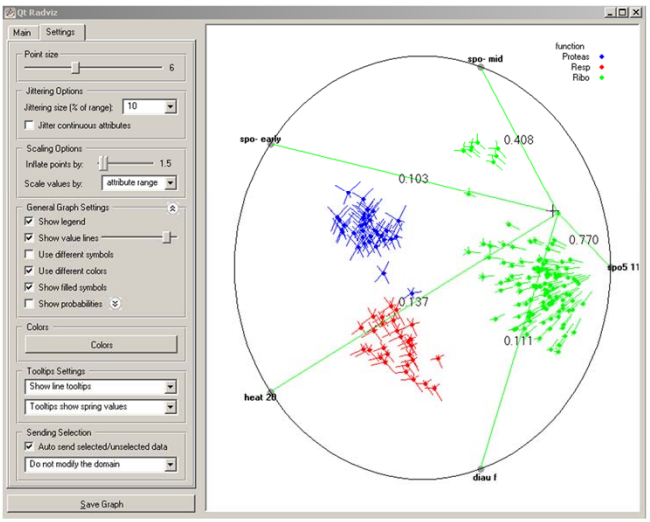
Parallel Coordinates in 3D

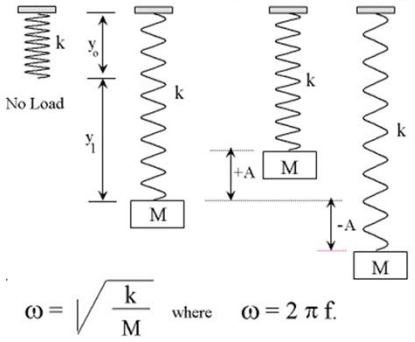


Streit et al. (2006)

Slide 9-32 RadViz – Idea based on Hooke's Law

Demšar, J., Curk, T., & Erjavec, A. Orange: Data Mining Toolbox in Python; Journal of Machine Learning Research 14:2349–2353, 2013.





Source: http://orange.biolab.si/

A. Holzinger 709.049 53/98 Med Informatics L09

http://www.scielo.br/scielo.php?pid=S1678-31662010000300004&script=sci_arttext

RadViz is a method for mapping a set of n-dimensional points into a plane and to identify relations among data. Its main advantage is that it needs no projections and provides a global view on the multidimensional data. This method is following Hooke's law from classical mechanics (spring laws).
A nice tool can be found at: http://orange.biolab.si/
This is an open source data visualization and analysis tool for novice and experts for data mining through visual programming and Python scripting including components for machine learning and add-ons for bioinformatics and text mining (Demšar et al., 2013).

Slide 9-33 RadViz Principle

1) Let us consider a point $y_i = (y_1, y_2, \dots y_n)$ from the *n*-dimensional space

2) This point is now mapped into a single point u in the plane of anchors: for each anchor j the stiffness of its spring is set to y_j

3) Now the Hooke's law is used to find the point u, where all the spring forces reach equilibrium (means they sum to 0). The position of $u = [u_1, u_2]$ is now derived by:

$$\sum_{j=1}^{n} (\vec{S}_j - \vec{u}) y_i = 0 \qquad \sum_{j=1}^{n} \vec{S}_j \, y_j = \vec{u} \sum_{j=1}^{n} y_j$$

$$\vec{u} = \frac{\sum_{j=1}^{n} \vec{S}_j \, y_j}{\sum_{j=1}^{n} y_j} \qquad u_1 = \frac{\sum_{j=1}^{n} y_j \cos(\alpha_j)}{\sum_{j=1}^{n} y_j} \qquad u_2 = \frac{\sum_{j=1}^{n} y_j \sin(\alpha_j)}{\sum_{j=1}^{n} y_j}$$

Novakova, L. & Stepankova, O. (2009). *RadViz and Identification of Clusters in Multidimensional Data. 13th International Conference on Information Visualisation, 104-109.*

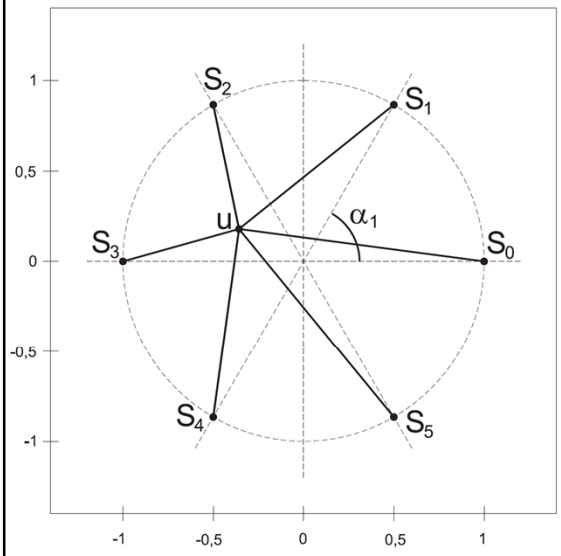
A. Holzinger 709.049 54/98 Med Informatics L09

Each RadViz mapping of points from *n*-dimensional space into a plane is uniquely defined by position of the corresponding *n* anchors (points S_j), which are placed in a single plane. Let us consider a point $y_i = (y_1, y_2, \dots y_n)$ from the *n*-dimensional space; This point is now mapped into a single point u in the plane of anchors: for each anchor j the stiffness of its spring is set to y_j and the Hooke's law is used to find the point u, where all the spring forces reach equilibrium (means they sum to 0). The position of $u = [u_1, u_2]$ is now derived by:

$$\sum_{j=1}^{n} (\vec{S}_j - \vec{u}) y_i = 0 \qquad \sum_{j=1}^{n} \vec{S}_j \, y_j = \vec{u} \sum_{j=1}^{n} y_j$$

$$\vec{u} = \frac{\sum_{j=1}^{n} \vec{S}_j \, y_j}{\sum_{j=1}^{n} y_j} \qquad u_1 = \frac{\sum_{j=1}^{n} y_j \cos(\alpha_j)}{\sum_{j=1}^{n} y_j} \qquad u_2 = \frac{\sum_{j=1}^{n} y_j \sin(\alpha_j)}{\sum_{j=1}^{n} y_j}$$

Slide 9-34 RadViz mapping principle and algorithm



1. Normalize the data to the interval $\langle 0, 1\rangle$

$$\bar{x}_{ij} = \frac{x_{ij} - min_j}{max_j - min_j}$$

2. Now place the dimensional anchors

3. Now calculate the point to place each record and to draw it:

$$y_i = \sum_{j=1}^{n} \bar{x}_{ij}$$

$$\vec{u}_i = \frac{\sum_{j=1}^{n} \vec{S}_j \bar{x}_{ij}}{y_i}$$

Novakova, L. & Stepankova, O. (2009). *RadViz and Identification of Clusters in Multidimensional Data. 13th International Conference on Information Visualisation, 104-109.*

A. Holzinger 709.049 55/98 Med Informatics L09

The algorithm using the RadViz process (Novakova & Stepankova, 2009) is:

1. Normalize the data to the interval $\langle 0, 1\rangle$

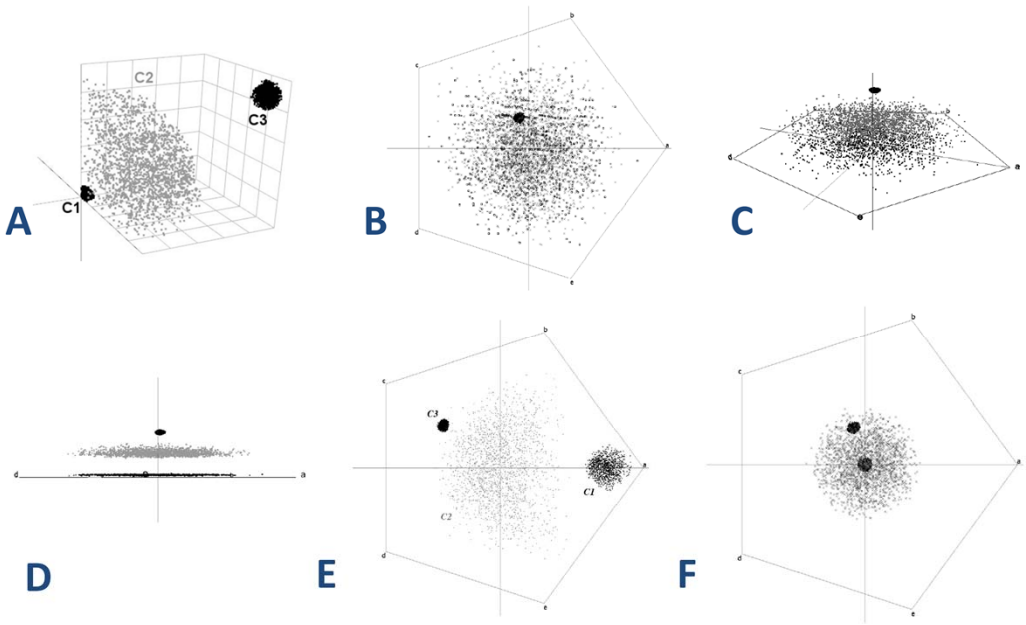
$$\bar{x}_{ij} = \frac{x_{ij} - min_j}{max_j - min_j}$$

2. Now place the dimensional anchors

3. Now calculate the point to place each record and to draw it:

$$y_i = \sum_{j=1}^{n} \bar{x}_{ij} \qquad \vec{u}_i = \frac{\sum_{j=1}^{n} \vec{S}_j \bar{x}_{ij}}{y_i}$$

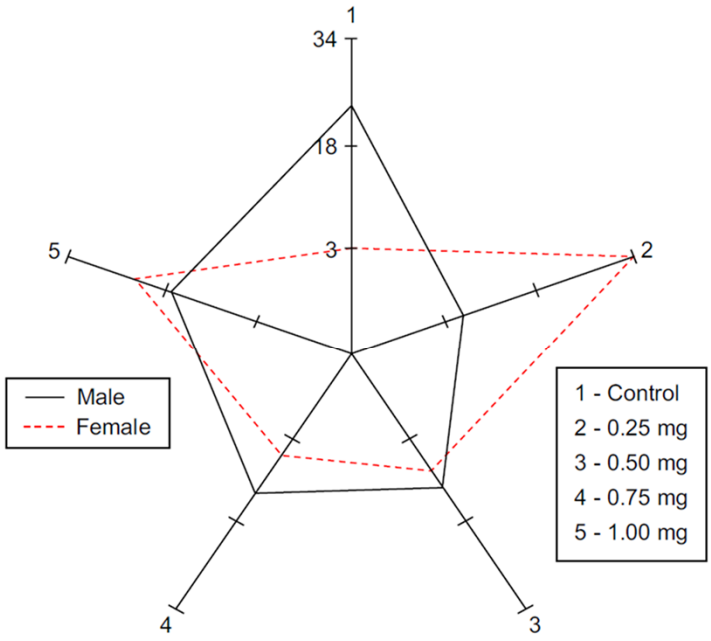
Slide 9-35 RadViz for showing the existence of clusters



Novakova, L. & Stepankova, O. (2009). *RadViz and Identification of Clusters in Multidimensional Data. 13th International Conference on Information Visualisation, 104-109.*

Here we see some RadViz Examples from (Novakova & Stepankova, 2009):
A = data projection in 3D space; B = the example from Slide 9-34; C= RadVizS clustering; D = same from another angle; E = Mirror projection; F= after normalization

Slide 9-36 Star plots/Radar chart/Spider-web/Polygon plot



Saary, M. J. (2008) Radar plots: a useful way for presenting multivariate health care data. *Journal Of Clinical Epidemiology, 61, 4, 311-317.*

Slide 9-36 Star plots/Radar chart/Spider-web/Polygon plot
StarPlots aka radar chart, spider web diagram, polygon plot, polar chart, or Kiviat diagram, are graphical methods for displaying multivariate data in the form of a 2D chart of three or more quantitative variables represented on axes starting from the same point.
Despite of their usefulness, such diagrams have not been widely used in the biomedical domain. One example of their use include distinguishing metabolic profiles of different cancer classes with star plots by the work of (Vion-Dury et al., 1993).
Vion-Dury, J., Favre, R., Sciaky, M., Kriat, M., Confort-Gouny, S., Harle, J., Grazziani, N., Viout, P., Grisoli, F. & Cozzone, P. 1993. Graphic-aided study of metabolic modifications of plasma in cancer using proton magnetic resonance spectroscopy. NMR in Biomedicine, 6, (1), 58-65.
In this slide we see an example of gender differences in death rate by treatment overlaid (Saary, 2008).

Slide 9-37 Star Plot production

- Arrange N axes on a circle in $\mathbb{R}^2$
- $3 \leq N \leq N_{max}$
  Note: An amount of $N_{max} \leq 20$ is just useful, according to Lanzenberger et al. (2005)
- Map coordinate vectors $P \in \mathbb{R}^N$ from $\mathbb{R}^N \rightarrow \mathbb{R}^2$
- $P = \{p_1, p_2, \ldots, p_N\} \in \mathbb{R}^N$ where each p_i represents a different attribute with a different physical unit
- Each axis represents one attribute of data
- Each data record, or data point P is visualized by a line along the data points
- A Line is perceived better than points on the axes

A. Holzinger 709.049 58/98 Med Informatics L09

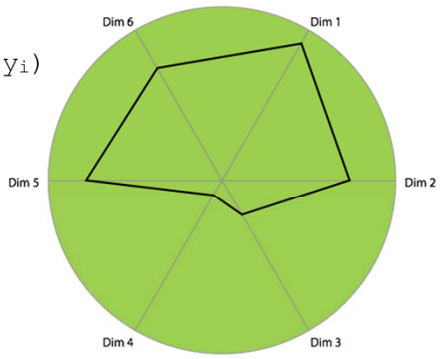
Each multivariate observation can be seen as a data point in an n-dimensional vector space:

Arrange N axes on a circle in R^2
3 ≤ N ≤ Nmax
Map coordinate vectors P ∈ ℝN from ℝN→ℝ2
P={p_1,p_2,…,p_N }∈ R^N where each pi represents a different attribute with a different physical unit
Each axis represents one attribute of data
Each data record, or data point P is visualized by a line along the data points
A line is perceived better than just points on the axes

There are commercial software tools which can help, e.g. the SAS statistical software package (http://www.sas.com) and Microsoft Excel (http://office.microsoft.com/en-us/excel-help/present-your-data-in-a-radar-chart-HA010218672.aspx). In SAS you can use the GRADAR procedure, the data can consist of one or more group variables and one or more outcome variables for which there is a count or frequency for each level of the group variable. The vertices of the radar plot are determined by the levels of a single variable that is given in the CHART statement. The spokes in the chart are positioned much like a clock starting at the 12-o'clock position and moving in a clockwise direction.
In Excel, the radar plot is generated by using the Insert function that allows the option to insert one of a variety of charts. Radar plots are among a number of other less commonly used graphing styles available on this menu including surface, doughnut, bubble, and stock plots. In this example, Excel does not prepare the chart by performing calculations on the raw data; hence, the sums (or means) of the raw-data groups must be represented in a separate new table (Saary, 2008).

Slide 9-38 Algorithm for drawing the axes and the lines

```
angle_sector = 2 * π / N
for each a_i from axes[]
{
angle_i = i * angle_sector
x_i = mid.x + r * cos(angle_i)
y_i = mid.y + r * sin(angle_i)
DrawLine(midpoint.x, midpoint.y, x_i, y_i)
```

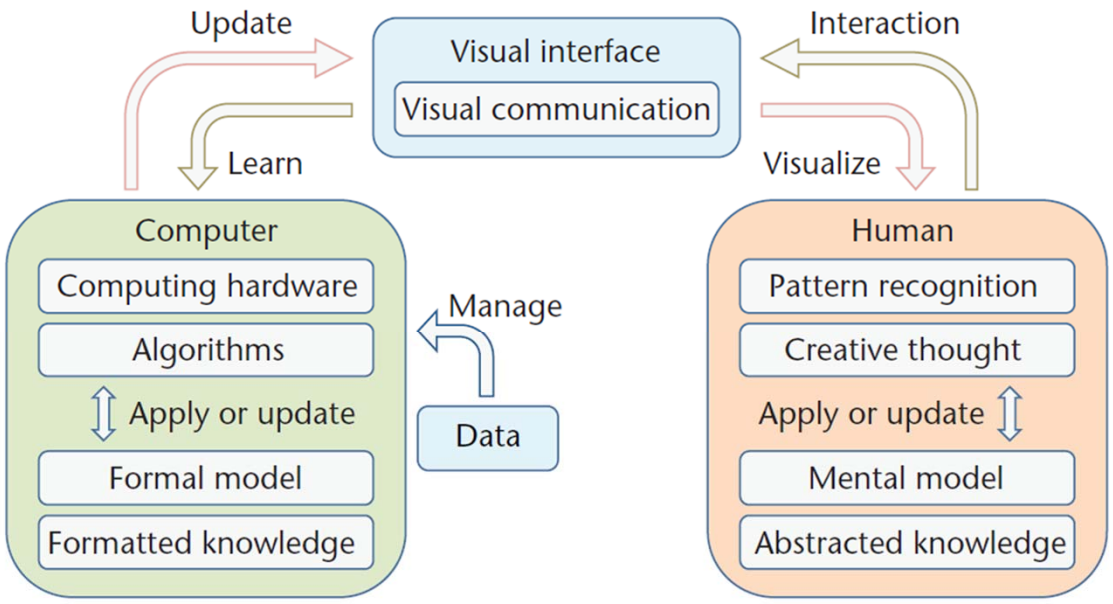



```
max_i = a_i.upperBound()

scaled_val_i = a_i.value() * r / max_i
x_val_i = mid.x + scaled_val_i * cos(angle_i)

y_val_i = mid.y + scaled_val_i * sin(angle_i)
DrawLine(x_val_i, y_val_i, x_val_i-1, y_val_i-1)
}
```

A. Holzinger 709.049 59/98 Med Informatics L09

Here we see an simple algorithm for drawing the axes and lines of a star plot diagram.

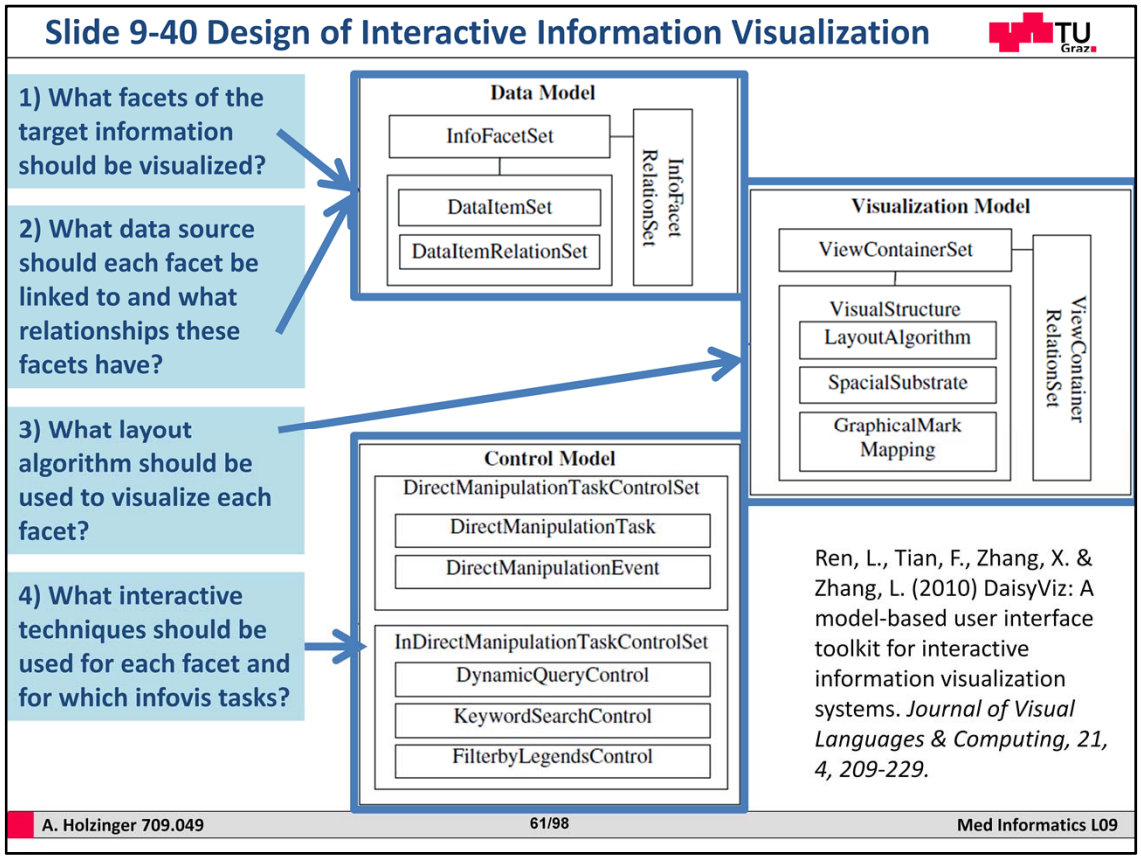
Slide 9-39 Visual Analytics is intelligent HCI



Mueller, K., Garg, S., Nam, J. E., Berg, T. & McDonnell, K. T. (2011) Can Computers Master the Art of Communication?: A Focus on Visual Analytics. *Computer Graphics and Applications, IEEE, 31, 3, 14-21.*

A. Holzinger 709.049 60/98 Med Informatics L09

Visual analytics seeks to support an "intelligent" interaction discourse with the end user through images, to stimulate curiosity and a penchant to decipher the unknown. In this slide we see a representative visual analytics process (Mueller et al., 2011) – which is also a nice example for Human–Computer Interaction: The computer supports the user in analytical reasoning, constructing a formal model of the given data and enabling insight. Validation and refinement of this computational model of insight can occur only in the human domain expert's mind. The human user must guide the computer in the formalization (learning process) of sophisticated models that capture what the human desires, cf. with (Holzinger, 2012).



A recent example is from (Ren et al., 2010), who developed DaisyViz, which is based on the idea of model-based interface development, which uses a declarative model of user interfaces to enhance the development process. Basically, a user interface model abstracts the features of a user interface and represents all the relevant aspects of the user interface in a formal language. The user interface model, the core of development process, is then parsed according to knowledge bases to generate applications. From the perspective of end-users, their only design concern is to construct an interface model. The slide shows the user interface model for Infovis (UIMI), developed by Ren et al. (2010). In this framework, one can construct a model by simply answering several questions:
1) What facets of the target information should be visualized?
2) What data source should each facet be linked to and what relationships these facets have?
The answers to these two questions result in the data model.
3) What layout algorithm should be used to visualize each facet?
The answer to this question results in the visualization model.
4) What interactive techniques should be used for each facet and for which infovis tasks?
The answer to this question results in the control model.

Slide 9-41 Overview first - then zoom and filter on Demand

- 1) Overview: Gain an overview about the entire data set (know your data!);
- 2) Zoom : Zoom in on items of interest;
- 3) Filter: filter out uninteresting items – get rid of distractors – eliminate irrelevant information;
- 4) Details-on-demand: Select an item or group and provide details when needed;
- 5) Relate: View relationships among items;
- 6) History: Keep a history of actions to support undo, replay, and progressive refinement;
- 7) Extract: Allow extraction of sub-collections and of the query parameters;

*) Shneiderman, B. (1996). *The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. Proceedings of the 1996 IEEE Symposium on Visual Languages, 336-343.*

A. Holzinger 709.049 62/98 Med Informatics L09

There are many visual design guidelines but the basic principle might be summarized as the Visual Information Seeking Mantra by (Shneiderman, 1996): Overview first, zoom and filter, then details-on-demand, which has been further elaborated by (Keim et al., 2008):
1) Overview: Gain an overview about the entire data set (know your data!);
2) Zoom: Zoom in on items of interest; (Remember: the question "what is interesting?" is a very hard one.
3) Filter: filter out uninteresting items – get rid of distractors – eliminate irrelevant information;
4) Details-on-demand: Select an item or group and provide details when needed;
5) Relate: View relationships among items;
6) History: Keep a history of actions to support undo, replay, and progressive refinement;
7) Extract: Allow extraction of sub-collections and of the query parameters;

Slide 9-42 Letting the user interactively manipulate the data

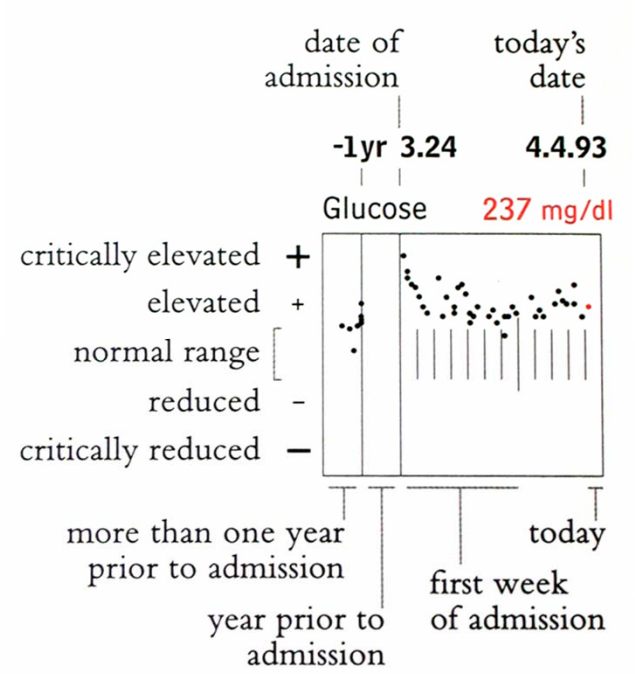
- Focus Selection = via direct manipulation and selection tools, e.g. multi-touch (in data space a n-dim location might be indicated);
- Extent Selection = specifying extents for an interaction, e.g. via a vector of values (a range for each data dimension or a set of constraints;
- Interaction type selection = e.g. a pair of menus: one to select the space, and the other to specify the general class of the interaction;
- Interaction level selection = e.g. the magnitude of scaling that will occur at the focal point (via a slider, along with a reset button)

Ward, M., Grinstein, G. & Keim, D. (2010) *Interactive Data Visualization: Foundations, Techniques and Applications. Natick (MA), Peters.*

A. Holzinger 709.049 63/98 Med Informatics L09

Ward, Grinstein & Keim (2010) summarize how a developer can let the user interactively manipulate the data:
Focus Selection = via direct manipulation and selection tools, e.g. multi-touch (in data space a n-dim location might be indicated);
Extent Selection = specifying extents for an interaction, e.g. via a vector of values (a range for each data dimension or a set of constraints;
Interaction type selection = e.g. a pair of menus: one to select the space, and the other to specify the general class of the interaction;
Interaction level selection = e.g. the magnitude of scaling that will occur at the focal point (via a slider, along with a reset button).

Slide 9-43 Rapid Graphical Summary of Patient Status

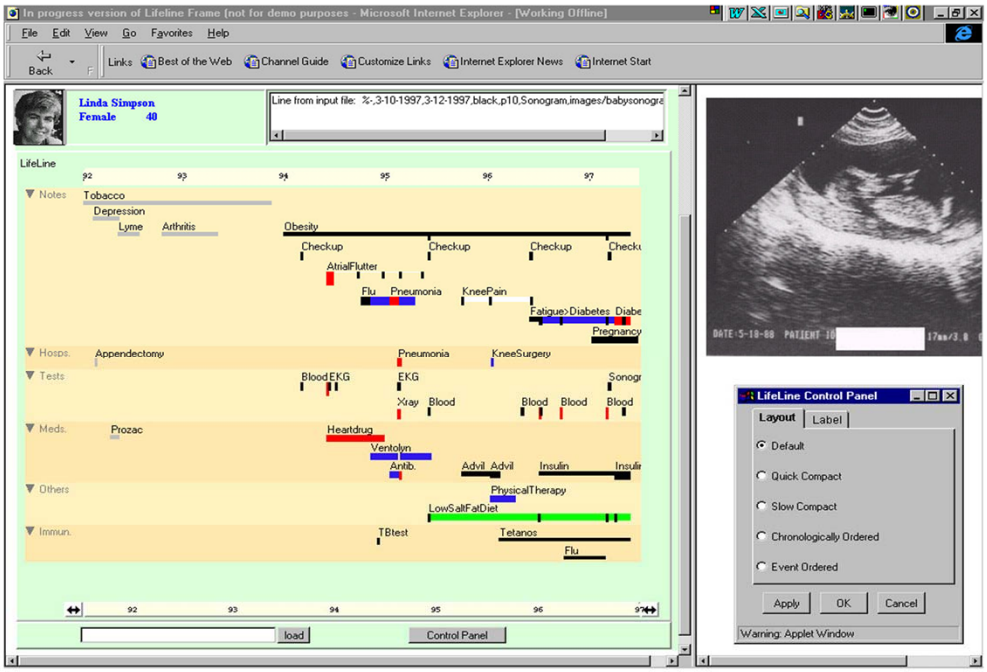


Powsner, S. M. & Tufte, E. R. (1994) Graphical Summary of Patient status. *The Lancet, 344,* ***8919, 386-389.***

A. Holzinger 709.049 64/98 Med Informatics L09

(Powsner & Tufte, 1994) presented a graphical summary of the patient status, which maps findings and treatments over time. The example shows that no tests of serum glucose were done during the 12 months before admission, although many were made more than 1 year earlier. The non-linear time-scale compresses years of data into a context for assessing recent trends.

Slide 6-44 Example Project LifeLines

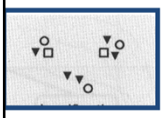


Plaisant, C., Milash, B., Rose, A., Widoff, S. & Shneiderman, B. (1996). *Life Lines: Visualizing Personal Histories. ACM CHI '96, Vancouver, BC, Canada, April 13-18, 1996.*

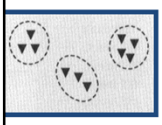
Life Lines was a very early project on a general visualization environment for personal patient histories. A Java user interface presents a one-screen overview of a computerized patient record using timelines. Problems, diagnoses, test results or medications can be represented as dots or horizontal lines. Zooming provides more details; line color and thickness illustrate relationships or significance. The visual display acts as a giant menu, giving direct access to the data (Plaisant et al., 1996).

What are temporal analysis tasks?

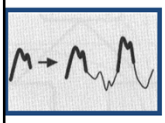
Slide 6-45 Temporal analysis tasks



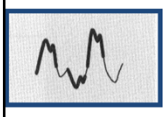
Classification = given a set of classes: the aim is to determine which class the dataset belongs to; a classification is often necessary as pre-processing;



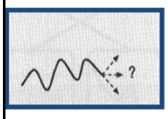
Clustering = grouping data into clusters based on similarity; the similarity measure is the key aspect of the clustering process;



Search/Retrieval = look for a priori specified queries in large data sets (query-by-example), can be exact matched or approximate matched (similarity measures are needed that define the degree of exactness);



Pattern discovery = automatically discovering relevant patterns in the data, e.g. local structures in the data or combinations thereof;



Prediction = foresee likely future behaviour of data – to infer from the data collected in the past and present how the data will evolve in the future (e.g. autoregressive models, rule-based models etc.)

Aigner, W., Miksch, S., Schumann, H. & Tominski, C. (2011) *Visualization of Time-Oriented Data. Human-Computer Interaction Series. London, Springer.*

A. Holzinger 709.049 67/98 Med Informatics L09

Temporal means of, relating to, or limited by time

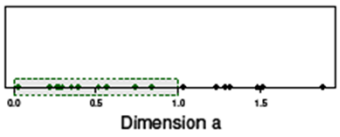
Temporal analysis and temporal data mining are especially concerned with extracting useful information from time-oriented data

Classification – given a set of classes, the goal of classification is to determine which class a dataset, sequence.

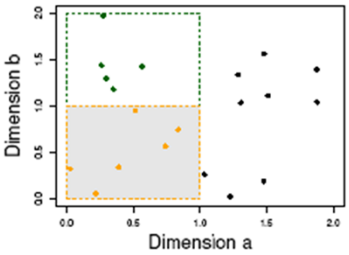
The analysis of time-oriented data is very important in the biomedical domain. In recent years, a variety of techniques for visualizing such data are available, e.g.:
Classification = given a set of classes: the aim is to determine which class the dataset belongs to; a classification is often necessary as pre-processing;
Clustering = grouping data into clusters based on similarity; the similarity measure is the key aspect of the clustering process;
Search/Retrieval = look for a priori specified queries in large data sets (query-by-example), can be exact matched or approximate matched (similarity measures are needed that define the degree of exactness);
Pattern discovery = automatically discovering relevant patterns in the data, e.g. local structures in the data or combinations thereof;
Prediction = foresee likely future behaviour of data – to infer from the data collected in the past and present how the data will evolve in the future (e.g. autoregressive models, rule-based models etc.)
For an excellent overview consult the book by (Aigner et al., 2011).

Example: Subspace Clustering

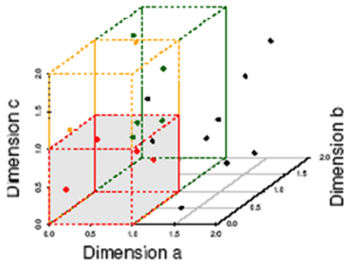
Remember: The curse of dimensionality



(a) 11 Objects in One Unit Bin



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

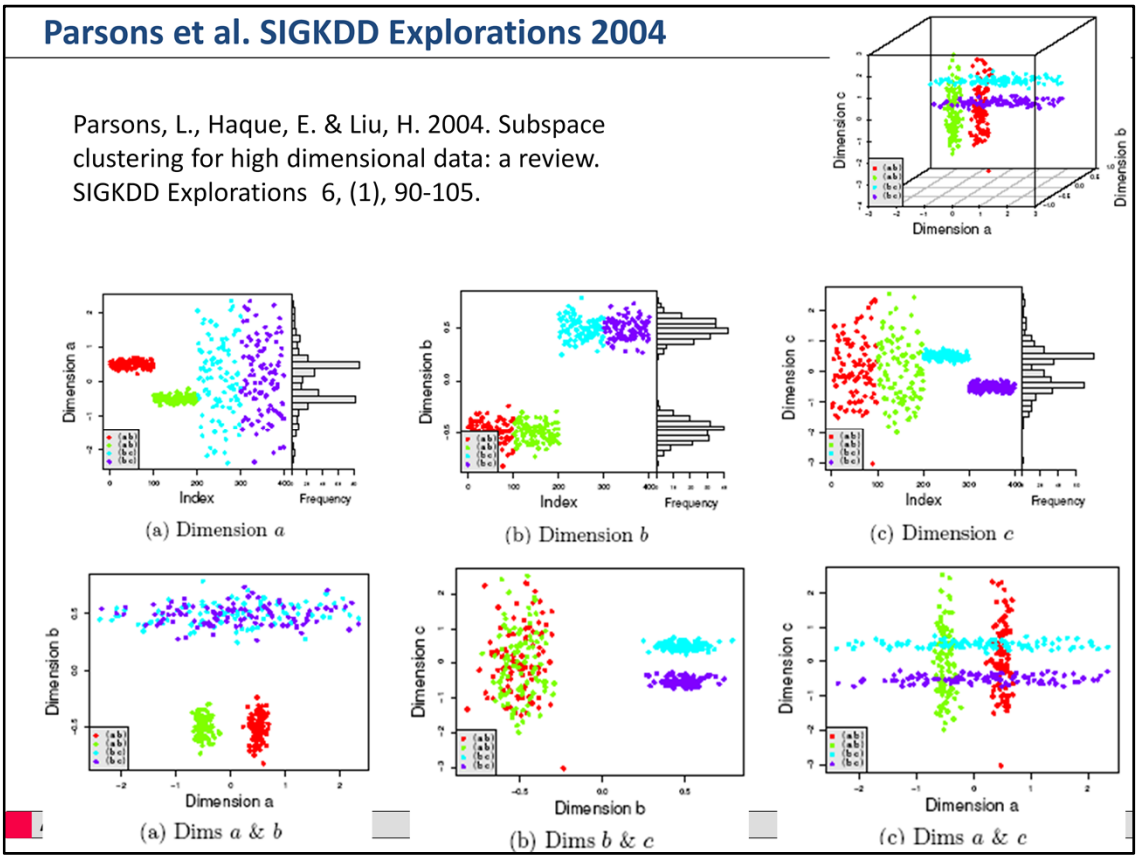
Data in only one dimension is relatively packed
Adding a dimension "stretch" the points across that dimension, making them further apart
Adding more dimensions will make the points further apart—high dimensional data is extremely sparse
Distance measure becomes meaningless—due to equi-distance

Repeat some definitions

- Dataset - consists of a matrix of data values, rows represent individual instances and columns represent dimensions.
- Instance - refers to a vector of *d* measurements.
- Cluster - group of instances in a dataset that are more similar to each other than to other instances. Often, similarity is measured using a distance metric over some or all of the dimensions in the dataset.
- Subspace - is a subset of the *d* dimensions of a given dataset.
- Subspace Clustering – seek to find clusters in a dataset by selecting the most *relevant* dimensions for each cluster separately .
- Feature Selection - process of determining and selecting the dimensions (features) that are most relevant to the data mining task.

Parsons et al. SIGKDD Explorations 2004

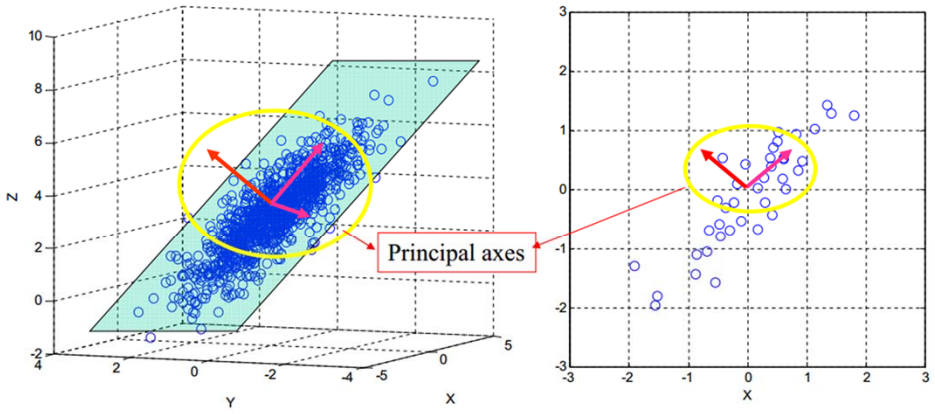
Parsons, L., Haque, E. & Liu, H. 2004. Subspace clustering for high dimensional data: a review. SIGKDD Explorations 6, (1), 90-105.



Interesting clusters may exist only in some subspaces – this helps in reducing the high-dimensionality search space

Often in high dimensional data, many dimensions are irrelev ant and can mask existing clusters in noisy
data. F ature selection removes irrelev ant and redundant dimensions by analyzing the entire dataset. Subspace clustering algorithms localize the search for relev ant dimensions allowing them to find clusters that exist in multiple, possibly overlapping subspaces. There are two major branches
of subspace clustering based on their search strategy . Topdown algorithms ¯nd an initial clustering in the full set of
dimensions and ev aluate the subspaces of each cluster, iteratively improving the results. Bottom-up approaches ¯nd
dense regions in low dimensional spaces and combine them
to form clusters. This paper presents a survey of the various
subspace clustering algorithms along with a hierarchy organizing the algorithms by their de¯ning characteristics. We
then compare the two main approaches to subspace clustering using empirical scalability and accuracy tests and discuss
some potential applications where subspace clustering could
be particularly useful.

Similar: Principal Component Analysis (PCA)



Curse of dimensionality (Bellman, 1957): As more dimensions are available, data becomes more sparse and distance measures are less meaningful.

A. Holzinger 709.049 72/98 Med Informatics L09

n machine learning problems that involve learning a "state-of-nature" (maybe an infinite distribution) from a finite number of data samples in a high-dimensional feature space with each feature having a number of possible values, an enormous amount of training data are required to ensure that there are several samples with each combination of values. With a fixed number of training samples, the predictive power reduces as the dimensionality increases, and this is known as the Hughes effect[3] or Hughes phenomenon (named after Gordon F. Hughes).[4][5]

Slide 6-46 Future Outlook

- Time (e.g. entropy) and Space (e.g. topology)
- Knowledge Discovery from “unstructured” ;-) (Forrester: >80%) data and applications of structured components as methods to index and organize data -> Content Analytics
- Open data, Big data, sometimes: small data
- Integration in “real-world” (e.g. Hospital), mobile
- How can we measure the benefits of visual analysis as compared to traditional methods?
- Can (and how can) we develop powerful visual analytics tools for the non-expert end user?

A. Holzinger 709.049 73/98 Med Informatics L09

Visualization is most important, because this is what the end user “experiences”.
Very important in the future are visualizations of time (e.g. entropy) and space (e.g. topology).
As we know from the famous Forrester Reports, more than 80% of all data contain “unstructured” elements, hence content analytics techniques along with advanced interactive visualizations are essential.
Amazingly, there are only very few of the sophisticated visualization methods integrated in “real-world” (e.g. Hospital Information System) and we are lacking of visualization methods for mobile computers. Finally, two major questions will be important in the future:
How can we measure the benefits of visual analysis as compared to traditional methods?
Can (and how can) we develop powerful visual analytics tools for the non-expert end user?
Future research questions include: finding proper visual encodings, understanding multi-dimensional spaces, facilitating sensitivity, understanding of uncertainty and facilitating trade-offs (Saad, Hamarneh & Möller, 2010), (Tory & Möller, 2004).

My DEDICATION is to make data valuable … Thank you!

Sample Questions (1)

- What is semiotic engineering?
- Please explain the process of intelligent interactive information visualization!
- What is the difference between visualization and visual analytics?
- Explain the model of perceptual visual processing according to Ware (2004)!
- What was the historical start of systematic visual analytics? Why is this an important example?
- Please describe very shortly 6 of the most important visualization techniques!
- Transform five given data points into parallel coordinates!
- How can you ensure data protection in using parallel coordinates?
- What is the basic idea of RadViz?
- For which problem would you use a star-plot visualization?

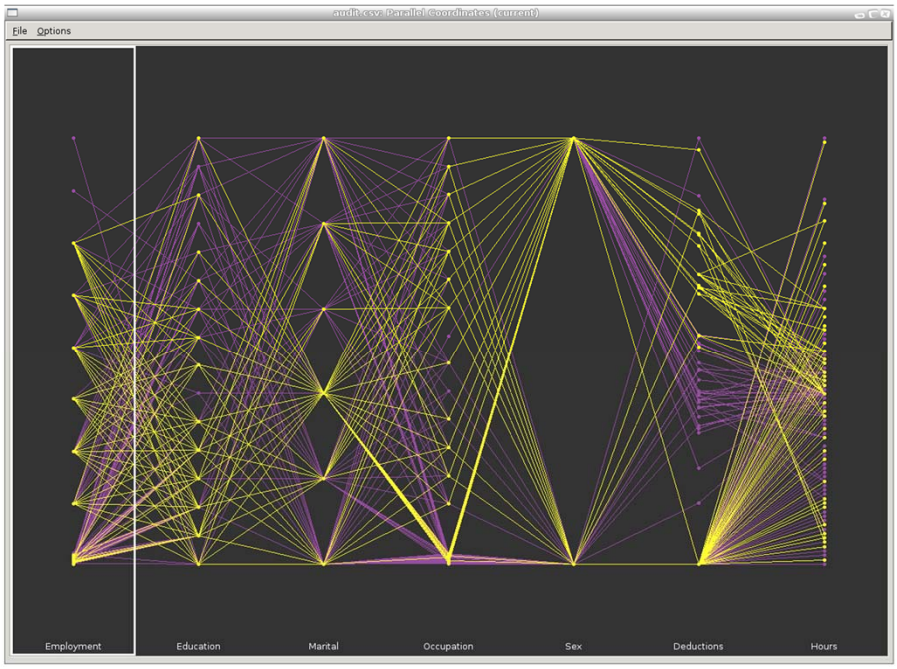
Sample Questions (2)

- What are the basic design principles of interactive intelligent visualization?
- What is the visual information seeking mantra of Shneiderman (1996)?
- Which concepts are important to let the end user interactively manipulate the data?
- What is the problem involved in looking at neonatal polysomnographic recordings?
- Why is time very important in medical informatics?
- What was the goal of LifeLines by Plaisant et al (1996)?
- Which temporal analysis tasks can you determine?
- Why is pattern discovery in medical informatics so important?
- What is the aim of foreseeing the future behaviour of medical data?

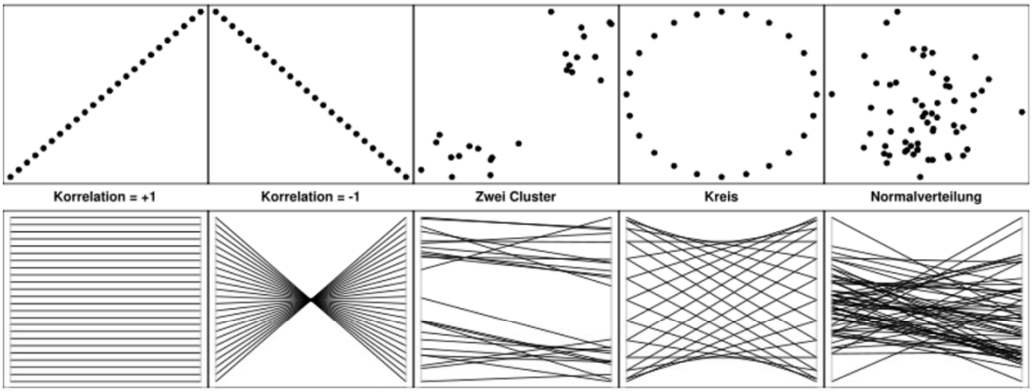
Some useful links

- http://vis.lbl.gov/Events/SC07/Drosophila/ (some really cool examples of high-dimensional data)
- http://people.cs.uchicago.edu/~wiseman/chernoff (Chernoff Faces in Java)
- http://lib.stat.emu.edu (Iris sample data set)
- http://graphics.stanford.edu/data/voldata (113-slice MRI data set of CT studies of cadaver heads)

Appendix: Parallel Coordinates in a Vis Software in R



http://datamining.togaware.com



Zur Visualisierung von hochdimensionalen Daten in der Statistik müssen drei wichtige Aspekte beachtet werden:

die Anordnung der Achsen

Die Anordnung der Achsen ist entscheidend für die Suche nach Strukturen in den Daten. In einer typischen Datenanalyse werden meist viele Anordnungen ausprobiert. Es wurden Anordnungsheuristiken entwickelt, die Einblicke in interessante Strukturen erlauben.[7]

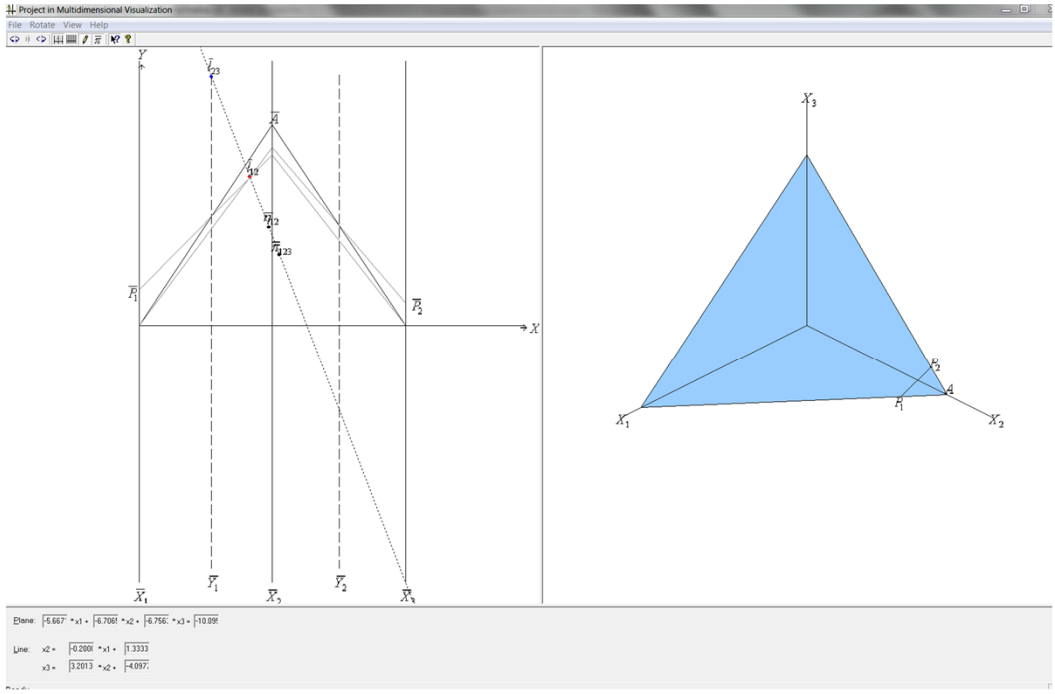
die Rotation der Achsen (Daten)

Da die i-te Koordinate durch die Ecke auf der i-ten Achse bestimmt wird, kann eine Rotation der Achsen (= Rotation der Daten) ein anderes Bild ergeben. Die beiden linken Grafiken können als Rotation der Achsen (oder Daten) um 90 Grad aufgefasst werden. Trotz gleicher Struktur ergeben sich unterschiedliche Strukturen in den parallelen Koordinaten.

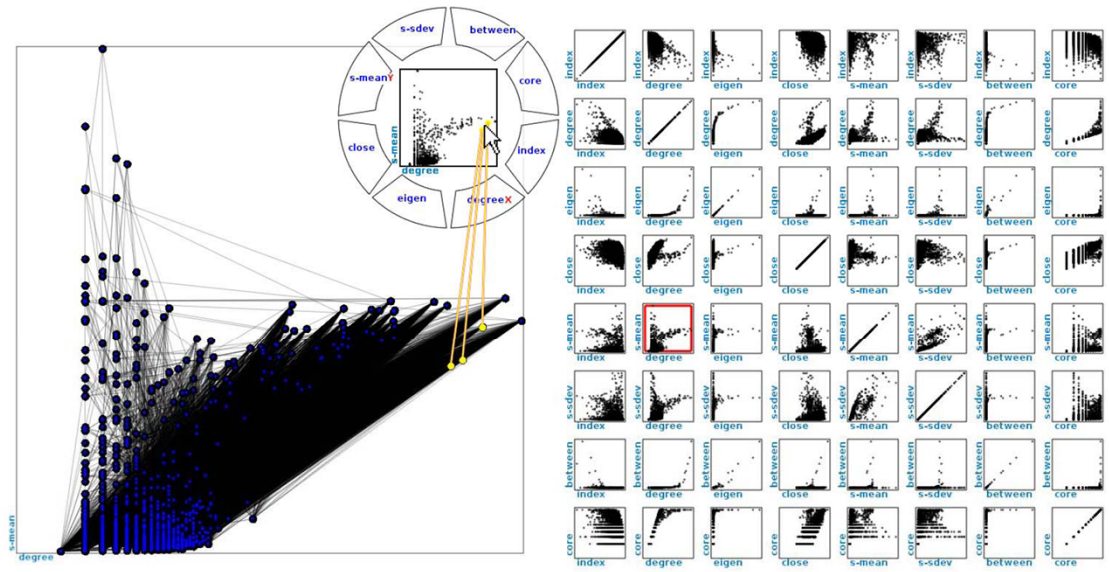
die Skalierung der Achsen

Die parallelen Koordinaten sind im Wesentlichen eine Aneinanderreihung von Linien zwischen Paaren von Koordinatenachsen.[5] Daher sollten die Variablen auf einen ähnlichen Maßstab skaliert sein. Verschiedene Skalierungen können ebenfalls interessante Einsichten in die Daten geben.

Visual Multidimensional Geometry and its Applications (1)



Appendix: Node-link graphs to visualize biological networks



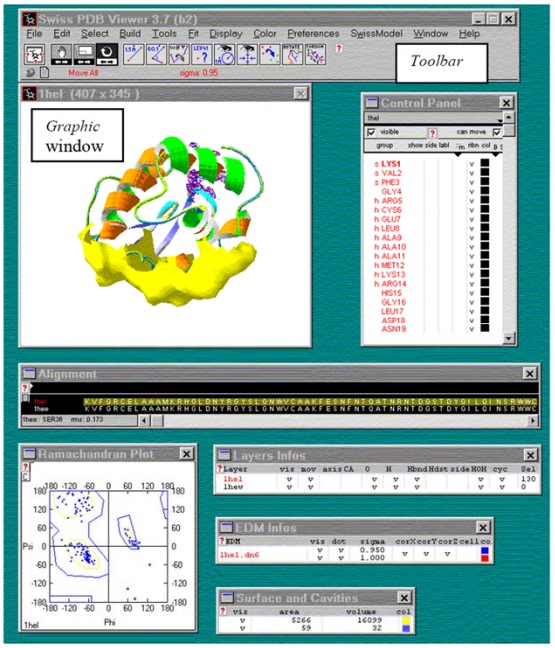
Viau, C., McGuffin, M. J., Chiricota, Y. & Jurisica, I. (2010) The FlowVizMenu and Parallel Scatterplot Matrix: Hybrid Multidimensional Visualizations for Network Exploration. *Visualization and Computer Graphics, IEEE Transactions on, 16,* ***6, 1100-1108.***

The user interface. At left: the node-link diagram, here with nodes positioned according to an attribute-driven layout, i.e., adopting their corresponding positions within a degree × s-mean scatterplot. Top middle: the FlowVizMenu is popped up and
contains the same scatterplot. Fluid gestures within the menu select dimensions to drive the attribute-driven layout with smoothly
animated transitions. At right: the P-SPLOM, here showing a SPLOM of the nodes' metrics.

Abstract—A standard approach for visualizing multivariate networks is to use one or more multidimensional views (for example,
scatterplots) for selecting nodes by various metrics, possibly coordinated with a node-link view of the network. In this paper, we
present three novel approaches for achieving a tighter integration of these views through hybrid techniques for multidimensional
visualization, graph selection and layout. First, we present the FlowVizMenu, a radial menu containing a scatterplot that can be popped
up transiently and manipulated with rapid, fluid gestures to select and modify the axes of its scatterplot. Second, the FlowVizMenu
can be used to steer an attribute-driven layout of the network, causing certain nodes of a node-link diagram to move toward their
corresponding positions in a scatterplot while others can be positioned manually or by force-directed layout. Third, we describe a
novel hybrid approach that combines a scatterplot matrix (SPLOM) and parallel coordinates called the Parallel Scatterplot Matrix (PSPLOM),
which can be used to visualize and select features within the network. We also describe a novel arrangement of scatterplots
called the Scatterplot Staircase (SPLOS) that requires less space than a traditional scatterplot matrix. Initial user feedback is reported.

Index Terms—Interactive graph drawing, network layout, attribute-driven layout, parallel coordinates, scatterplot matrix, radial menu.

Appendix: Deep View Working Environment - Swiss PDB



http://www.expasy.org

A. Holzinger 709.049 82/98 Med Informatics L09

DeepView – the Swiss-PdbViewer (or SPDBV), is an interactive molecular graphics program for viewing and analyzing protein and nucleic acid structures. In combination with Swiss-Model (a server for automated comparative protein modeling maintained at http://www.expasy.org/swissmod) new protein structures can also be modeled.
Annex 5: Glossary provides an extended dictionary for DeepView terminology. To facilitate understanding of the following chapters, some essential terms are introduced here:
A molecular coordinate file (e.g. *.pdb, *.mmCIF, etc.) is a text file containing, amongst other information, the atom coordinates of one or several molecules. It can be opened from a local directory or imported from a remote server by entering its PDB accession code. The content of one coordinate file is loaded in one (or more) layers, the first one will be referred to as the "reference layer".
DeepView can simultaneously display several layers, and this constitutes a project. When working on projects, the layer that is currently governed by the Control Panel is called the currently active layer.
Each molecule is composed of groups, which can be amino acids, hetero-groups, water molecules, etc. and each group is composed of atoms.
Non-coordinate files containin specific information other than atom coordinates. Molecular surfaces, electrostatic potential maps, and electron density maps are examples of non-coordinate files, which can either be computed by DeepView, or loaded from specialized external programs.

Appendix: Visual Analytics for Epidemiologists

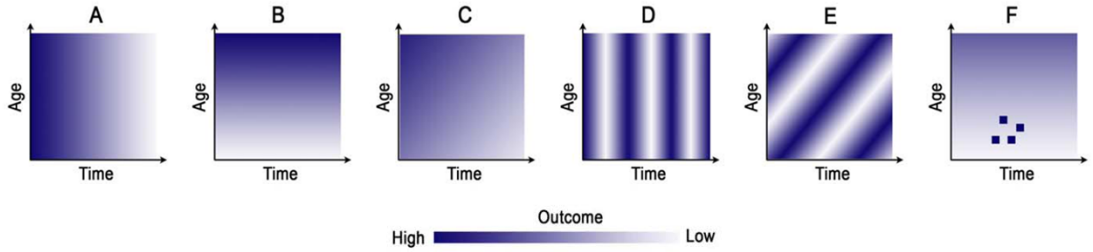


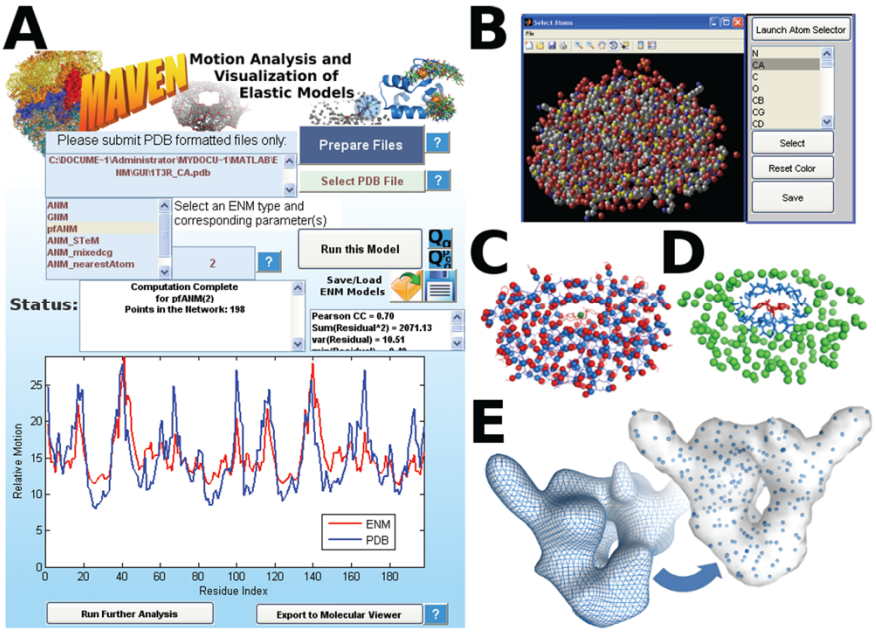
Figure 1. Typical patterns observed in image plots used to study the association between age, time, and the disease of interest.
doi:10.1371/journal.pone.0014683.g001

Chui, K. K. H., Wenger, J. B., Cohen, S. A. & Naumova, E. N. (2011) Visual Analytics for Epidemiologists: Understanding the Interactions Between Age, Time, and Disease with Multi-Panel Graphs. *Plos One, 6,* ***2.***

Image Plots
An image plot is capable of displaying information for at least three variables. The first two variables are shared and represented by the horizontal axis and vertical axis, which form a grid of numerous rectangular tiles. In our example, these two variables are time and age. The third variable is represented by different hues or saturations of colors in the rectangular tiles. In our examples, the third variables are disease rates or counts.
New information previously masked by case pyramids and timeseries plots can be revealed from image plots. However, reading the graph properly requires some training and practice. Figure 1 displays some of the typical patterns. In this figure, a monochromatic color scale is used, with higher saturation, or darker colors, representing higher values of an outcome. Panel A shows a decrease of color saturation from left to right, indicating that the outcome decreases along time, and the observed decrease is somewhat uniform across age. Panel B shows a decrease of color saturation from top to bottom, indicating an increase in an outcome in the older age group irrespective of time. Panel C shows the combined effect of Panels A and B: an outcome increasing with age but decreasing across time. Panel D shows a striated pattern typical for periodic fluctuations in an outcome across time—evidence of seasonality. Panel E also shows a striated pattern, but tilted at an angle. This pattern indicates that at any time point, the outcome across age is uneven and likely to be cohort-specific. This phenomenon, known as the ‘‘age-cohort effect,’’ [21] can be observed if an outcome was measured in the same cohort repeatedly and an outcome remains higher or lower in a group of subjects as they aged. Finally, Panel F shows a set of four distinct clusters combined with Panel B, where substantially higher outcomes in the younger group at the middle of the time duration are observed. These clusters might be indicative of disproportionally high values of an outcome, or an aberration in the expected values. Depending on the context a cluster, the pattern may indicate potential outbreaks.

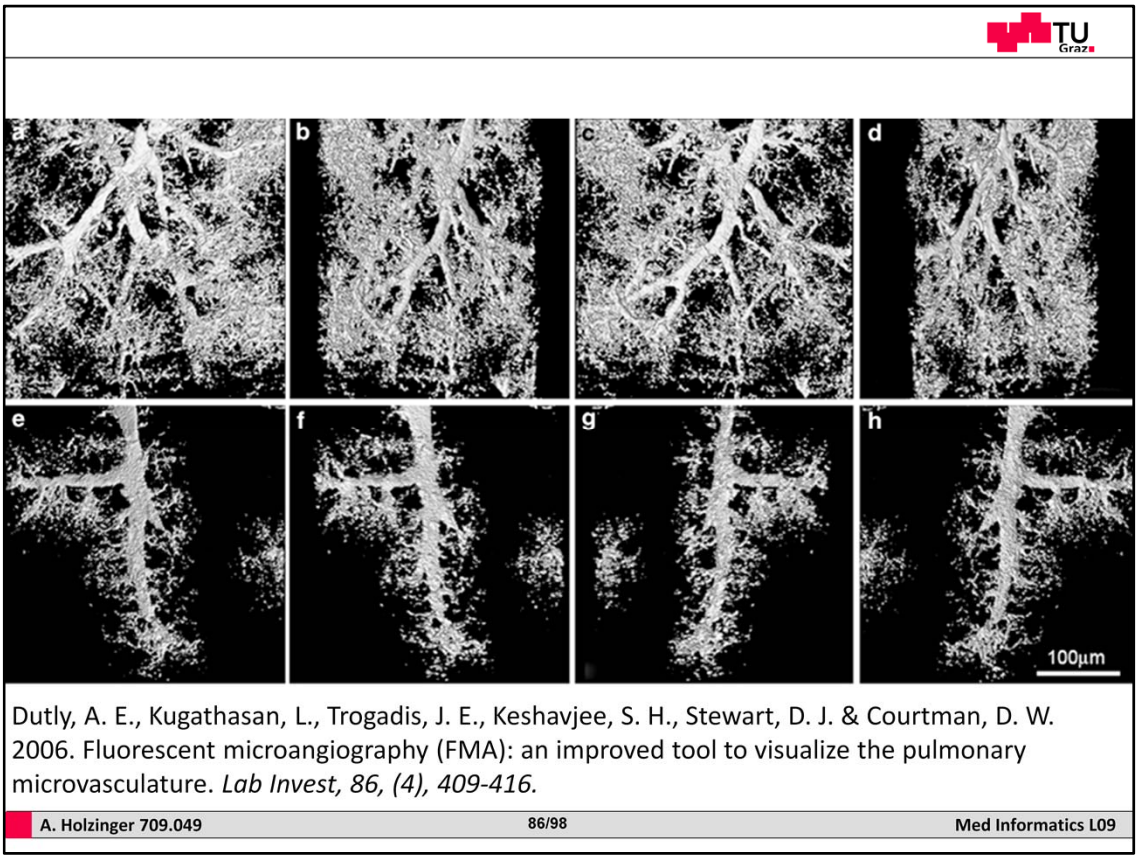
Appendix: Motion Analysis & Visualization of Elastic Models

Zimmermann, M., Kloczkowski, A. & Jernigan, R. (2011) MAVENs: Motion analysis and visualization of elastic networks and structural ensembles. *Bmc Bioinformatics, 12, 1, 264.*



Typical direct image

Erginousakis, D. et al. 2011. Comparative Prospective Randomized Study Comparing Conservative Treatment and Percutaneous Disk Decompression for Treatment of Intervertebral Disk Herniation. *Radiology, 260, (2), 487-493.*

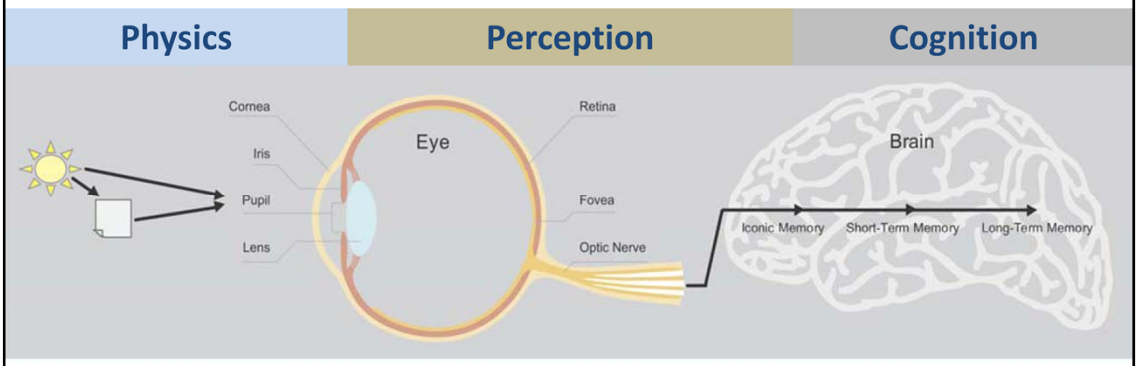


http://www.nature.com/labinvest/journal/v86/n4/images/3700399f4.jpg

Three-dimensional reconstruction. Representative two-dimensional views from 3D reconstruction of the normal (a–d) and MCT treated (e–h) rat pulmonary vasculature are displayed. 75 optical sections (1 mum step size) were used to construct the image views, note the marked loss of vasculature in the pulmonary hypertensive MCT treated lung. Additional 3D animations can be viewed online as Supplements at http://www.nature.com/labinvest.

Visualization of the complex lung microvasculature and resolution of its three-dimensional architecture remains a difficult experimental challenge. We present a novel fluorescent microscopy technique to visualize both the normal and diseased pulmonary microvasculature. Physiologically relevant pulmonary perfusion conditions were applied using a low-viscosity perfusate infused under continuous airway ventilation. Intensely fluorescent polystyrene microspheres, confined to the vascular space, were imaged through confocal optical sectioning of 200 mum-thick lung sections. We applied this technique to rat lungs and the markedly enhanced depth of field in projected images allowed us to follow vascular branching patterns in both normal lungs and lungs from animals with experimentally induced pulmonary arterial hypertension. In addition, this method allowed complementary immunostaining and identification of cellular components surrounding the blood vessels. Fluorescent microangiography is a widely applicable and quantitative tool for the study of vascular changes in animal models of pulmonary disease.

Repetition: From Physics of Light to Cognition of Thought

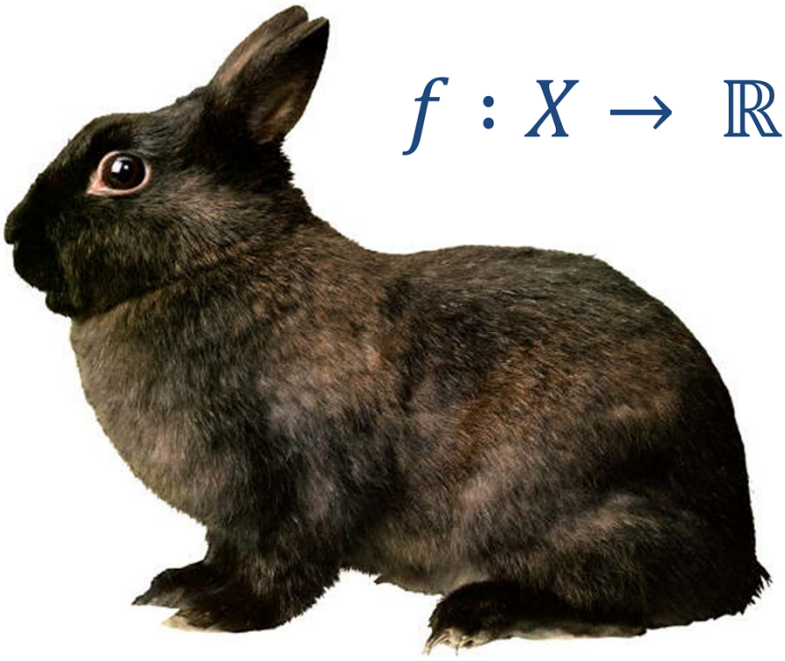


Few, S. (2006) *Information Dashboard Design. Sebastopol (CA), O'Reilly.*

A. Holzinger 709.049 87/98 Med Informatics L09

When compared to our other senses (hearing, smell, taste, and touch), which are like narrow alleyways paved in cobblestones, vision is like a superhighway.
Perceptual Edge Tapping the Power of Visual Perception Page 2
From Light to Thought
Figure 1 provides a visual representation of the primary components of visual perception. We don't actually see physical objects; we see light, either emitted by objects or reflected off of their surfaces. This light enters our eyes through an opening in the iris called the pupil. When we focus directly on objects, the emitted or reflected light shines on a small area on the retina at the back of the eye called the fovea. The retina consists of millions of light receptors, subdivided into two basic types, rods and cones. Rods sense dim light and record what they detect in black and white. Cones sense brighter light and record what they detect in color. Cones are further subdivided into three types, each of which detects a different range of the color spectrum: roughly blue, green, and red. The fovea is simply an area with an extremely dense collection of cones. As a result, light that shines on the fovea can be seen in extremely fine detail. We're capable of seeing up to 625 separate data points in a one-inch square area, such as a dense collection of dots in a scatter plot. Perception of visual stimuli detected by parts of the retina other than the fovea is much less detailed, but it's capable of simultaneously processing vast amounts of information throughout one's span of vision, ready to notice a point of interest that invites greater attention (for example, the peripheral approach of a speeding car), which then leads to a quick shift in one's gaze to that area of interest. Rods and cones translate what they detect into electrochemical signals and pass them on, through the optic nerve, to the brain where they can be processed. Our eyes sense visual stimuli, then our brains perceive that data, making sense of it.

Remember



$$f : X \rightarrow \mathbb{R}$$

Remember: Data – Information (it is a visualization task!)

Each multivariate observation can be seen as a data point in an *n*-dimensional vector space

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

- **"Look at your data"**
- **transfer data into information**
- **By use of human intelligence …**
- **to transfer information into knowledge ($\mathbb{C} \rightarrow \mathbb{P}$)**
- **Challenge: To reduce the dimensionality of the data …**
- **… it is an information retrieval task!**

Remember: The quality can be measured by two measures:

- **Recall**
- **Precision**

A. Holzinger 709.049 89/98 Med Informatics L09

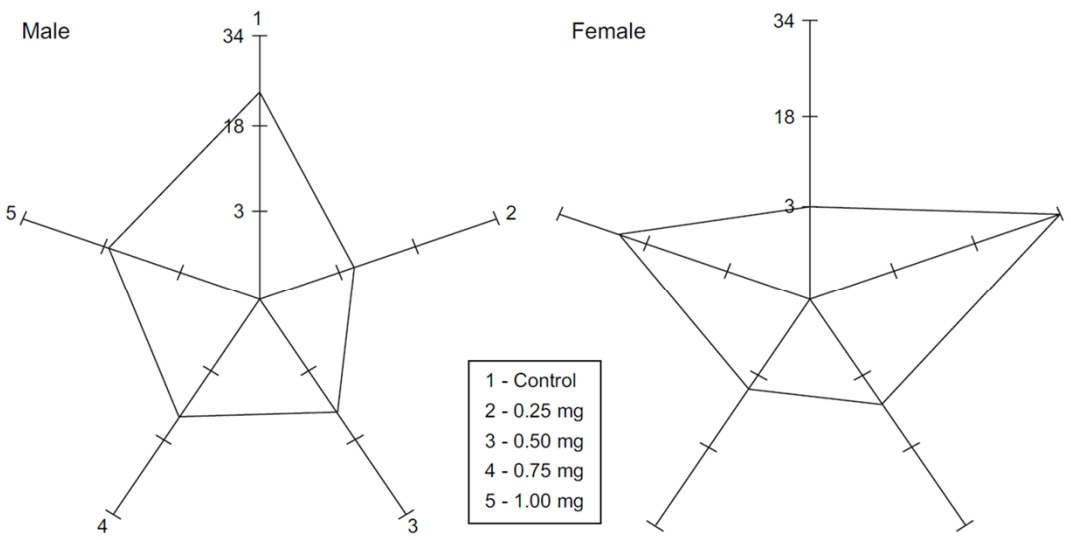
"Looking at the data" is central for exploratory data analysis. Dimensionality reduction for data visualization can be represented as an information retrieval task, where the quality of visualization can be measured by precision and recall measures and their smoothed extensions. Furthermore,
we show that visualization can be optimized to directly maximize the quality for any desired tradeoff between precision and recall, yielding very well-performing visualization methods.

Typical Problems in the Medical Clinical Domain

Holzinger, A., Hoeller, M., Bloice, M. & Urlesberger, B. (2008). *Typical Problems with developing mobile applications for health care: Some lessons learned from developing user-centered mobile applications in a hospital environment. International Conference on E-Business (ICE-B 2008), Porto (PT), IEEE, 235-240.*

Masses of text …

Example: Star Plot Diagram - Radar Chart



Saary, M. J. (2008) Radar plots: a useful way for presenting multivariate health care data. *Journal Of Clinical Epidemiology, 61, 4, 311-317.*

The Noisy Channel

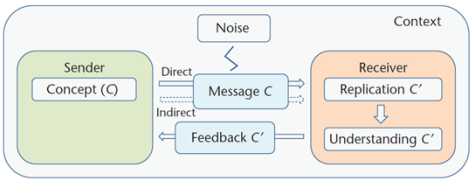
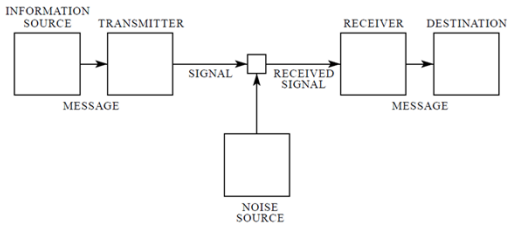


Figure 2. The interpersonal-communication protocol. A sender would like a receiver to comprehend message C, conveyed either straightforwardly or via indirect or subconscious mechanisms. However, noise in the communication channel or the receiver's failure to fully comprehend the message's intended meaning can undermine the sender's objective. An iterative clarification process eventually leads to a mutual understanding of the message.



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

Slide 9-45 Example Algorithms for Selection

```
Scatterplot-Select (xDim, yDim, xMin, xMax, yMin, yMax
1 s ←0⊳ Initialize the set of records
2 for each record i ⊳ For each record,
3   do x ← NORMALIZE(i,xDim) ⊳ derive the location,
4      y ← NORMALIZE(i,yDim)
5      if xMin < x < xMax and yMin < x < yMax
6          do s ← s ∪ I ⊳ select points within rectangle
7 return s
Point-in-Point-Polygon(xs, ys, numPoints, x,y)
1 j ← numPoints -1
2 oddNodes ← false
3 for i←0 to numPoints -1
4   do if ys[i]<y and ys[j]>=y or ys[j]<y and ys[i]>=y
5      do if xs[i]+(y-ys[i]/(ys[j]-ys[i])*(xs[j]-xs[i]<x
6          do oddNodes ← not oddNodes
7      j ← I
8 return oddNodes
```

Ward, M., Grinstein, G. & Keim, D. (2010) *Interactive Data Visualization: Foundations, Techniques and Applications. Natick (MA), Peters.*

Slide 9-46 40 sec of neonatal Polysomnographic recording

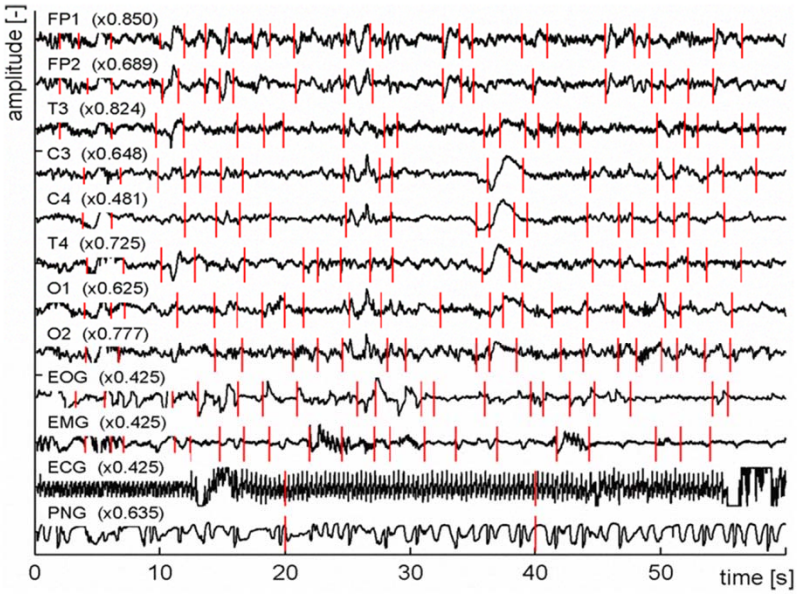
EEG signal from 8 ref. derivations: FP1, FP2, T3, T4, C3, C4, 01, 02

EOG = Electrooculogram

EMG = Electromyogram

PNG = Pneumogram

ECG = Electrocardiogram



Gerla, V., Djordjevic, V., Lhotska, L. & Krajca, V. (2009). *Visualization methods used for evaluation of neonatal polysomnographic data. ITAB 2009, Information Technology and Applications in Biomedicine, Cyprus, IEEE, 1-4.*

A. Holzinger 709.049 94/98 Med Informatics L09

Polysomnographic (PSG) signal processing represents a complex process consisting of several subsequent steps, namely pre-processing, segmentation, extraction of descriptive features, and classification. In this paper we focus on visualization methods that are also unseparable part of the whole process. The aim of these methods is to ease the work of medical doctors and to show trends that are not obvious when performing a manual inspection of the recorded signal.
In this study, the designed methods are applied to neonatal PSG data and enable the enhancement in visual differentiation between three important behavioral states: quiet sleep (QS), active sleep (AS) and wakefulness (WK). The ratio of these states is a significant indicator of the maturity of the newborn brain in clinical practice.

Many data mining techniques applied to neonatal PSG signals need the investigated signal to be fully or at least partially segmented to parts with similar characteristics and interpretation [2]-[8]. A common approach is based on splitting the signal in the time domain into smaller windows, called segments, and describing each one of them by extracted features. There exist several approaches to constant and adaptive signal segmentation, which divide non-stationary signal to quasi-stationary segments.
Fig. 1. The results of the segmentation. The real clinical neonatal polysomnographic data were used. Adaptive segmentation was applied to EEG signals (electrodes Fpl , Fp2, T3, C3, C4, T4, 01 , 02), EOG signal and EMG signal. For ECG and Respiration signal (PNG)a constantsegmentation was used.
III. FEATUREEXTRACTION
The most important step in the process of data evaluation is feature extraction. We used visualization of obtained features as an additional tool that helped us to decide which In this paper we present adaptive segmentation. Our approach involves two parameters: amplitude- and frequency-dependent changes in PSG signals.

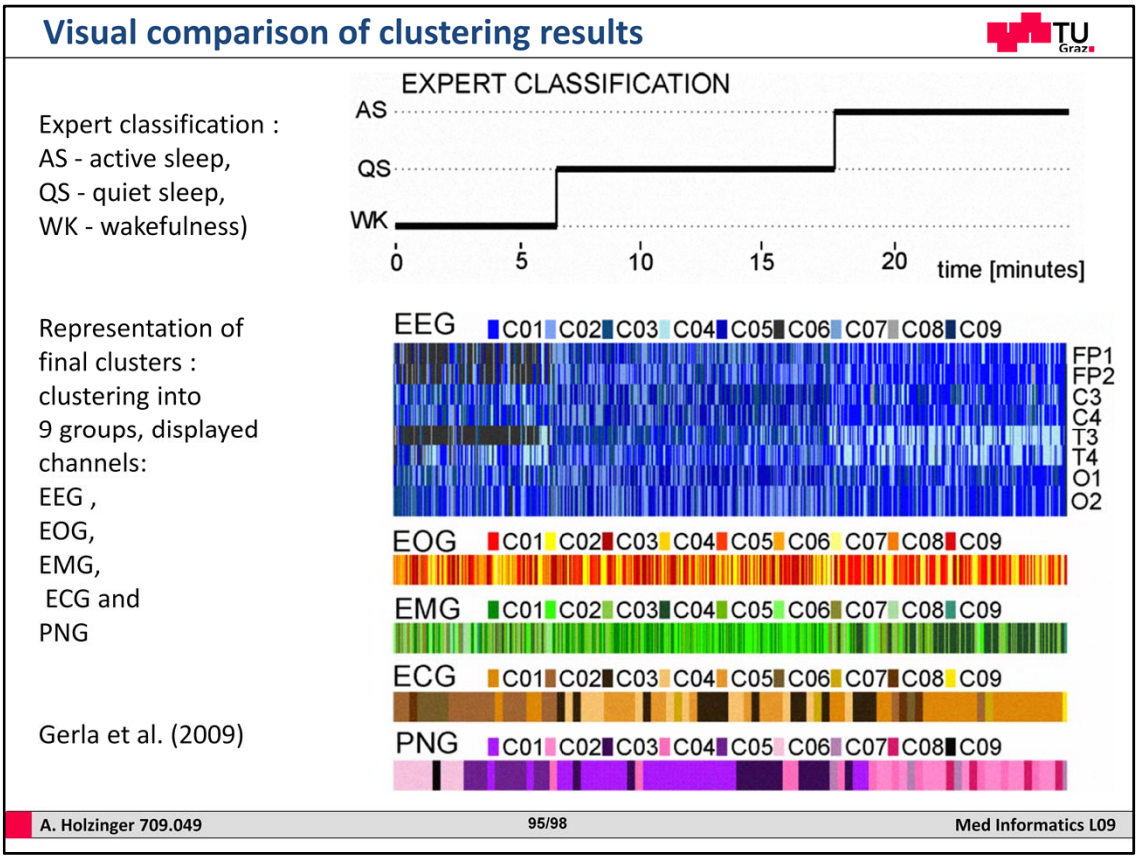
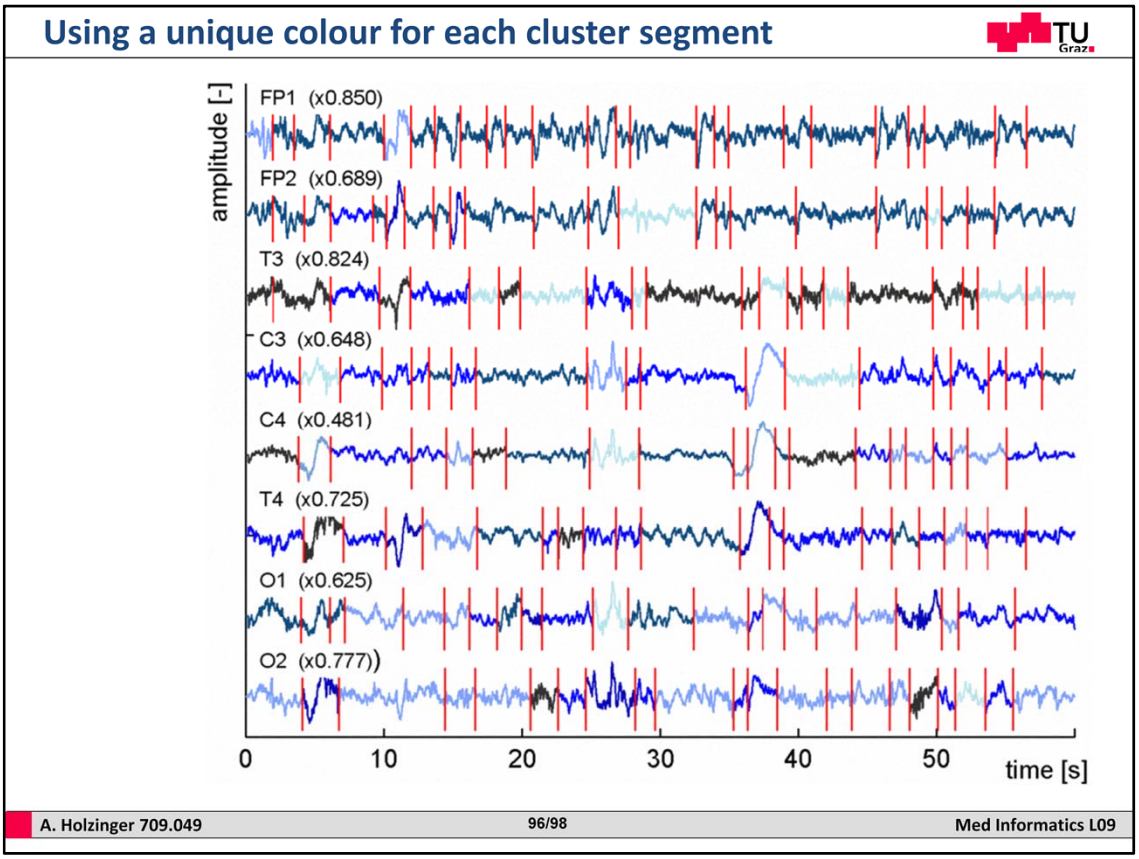


Fig. 2. Visual comparison of clustering results. On the top: expert classification (AS - active sleep, QS - quiet sleep, WK - wakefulness); on the bottom: representation of final clusters (clustering into 9 groups, displayed channels: EEG , EOG, EMG, ECG and PNG).

In this study we focused on hierarchical clustering methods. We used the Ward's method as a specific procedure for hierarchical clustering analysis (Ward's linkage is suitable for decreasing the total within-cluster sum of square error) . Fig. 2. illustrates clustering results for PSG channels. When cluster analysis is done, found clusters may be manually identified by an expert, or an automatic classification method can be performed. Because we aim to make the differentiation between three stages in our dataset (wake, active sleep and quiet sleep) it is necessary to use at least three clusters in the clustering process. But for reliable differentiation of stages, we used greater number of clusters .

The most important step in the process of data evaluation is feature extraction. We used visualization of obtained features as an additional tool that helped us to decide whichfeatures to select. Proper selection of features and visual comparison of the computed features may significantly influence the success rate of the classification. Extracted features were: statistical parameters (mean values, skewness, kurtosis); mean and maximum values of first and second derivation of the samples in segment ; absolute and relative power/energy for important EEG frequency bands derived from Fourier or wavelet transform; statistical values of the wavelet coefficients corresponding to different decomposition scales ; Shannon's entropy of wavelet transform details and approximations; and mean and maximum values of wavelet coefficients of first and second derivation.



Useful additional way of cluster analysis visualization is identifying segments from different clusters in the original signals and representing them by different colours. This kind of visualization is shown in Fig. 3. together with segment borders obtained by adaptive segmentation.
The used clustering approach is also compared with kmeans algorithm [14]. The results for Ward's and k-means clustering algorithms were very similar for used PSG datasets (statistically comparable).

Colored segments ofEEG data (electrodes FPI , FP2, rs, C3, C4, T4, 01 , 02,). For each cluster ofsegments a unique color is used.