Andreas Holzinger

Lecture 9 - Version WS 2013/14
Intelligent Information Visualization and Visual Analytics

VO 444.152 Medical Informatics

a.holzinger@tugraz.at
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
Learning Goals: At the end of this 9th lecture you ...

- ... have some background on visualization, visual analytics and content analytics;
- ... got an overview about various possible visualization methods for multivariate 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 at it finest;
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
Advancing 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);
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;
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
Verbal Information versus Visual Information
Slide 9-2 Verbal information versus Visual information
Slide 9-3 Semantic Ambiguity of Verbal Information
Multimedia Presentation

- Words
- Pictures

Sensory Register

- Ears
- Eyes

Working Memory

- Sounds
- Verbal
- Images
- Pictorial

Long-Term Memory

- Prior Knowledge

Integrating

Verbal Information Processing: Written Text

Multimedia Presentation
- Words
- Pictures

Sensory Register
- Ears
- Eyes

Working Memory
- Sounds
- Verbal
- Images
- Pictorial

Long-Term Memory
- Prior Knowledge

Integration of sensory and working memory with prior knowledge.

Verbal Information Processing: Spoken Text

A picture is worth a thousand words?
"Is a picture really worth a thousand words?"
Informatics as Semiotics Engineering
Slide 9-9 Three examples for Visual Languages

\[
W = \{w : w \in T(v)\}
\]
\[
H = - \sum_{w \in W} p(w) \log_2 p(w)
\]


1. Physical: is it present?
   - Signals, traces, components, ...
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, ...

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” data – mainly text – by using mixed methods from visual analytics and business intelligence;
Visualization is a typical HCI topic
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 knowledge discovery in data

Usefulness of Visualization Science
A look back into history...
What do you see in this picture?

Florence Nightingale – first medical quality manager

How many visualization methods do exist?
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, ...)  
4) Metaphor Visualization (Metro maps, story template, iceberg, ...)  
5) Strategy Visualization (Strategy Canvas, roadmap, morpho box, ...)  
6) Compound Visualization
**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;
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
Parallel Coordinates
On the plane with Cartesian-coords, a vertical line, labeled $\bar{X}_i$ 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 $\bar{C}$.

The $N$-vertices with $xy$-coords $(i - 1, c_i)$ are now on the parallel axes.

In $\bar{C}$ the full lines and not only the segments between the axes are included.

Slide 9-25 Polygonal line $\vec{C}$ is representing a single point

$\mathbb{R}^5: \ C = (c_1, c_2, c_3, c_4, c_5)$

Inselberg (2005)
A polygonal line $\overline{P}$ on the $N-1$ points represents a point

$P = (p_1, \ldots, p_{i-1}, p_i \ldots p_N) \in \ell$

since the pair of values $\ldots p_{i-1}, p_i$ 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.
Slide 9-27 Line Interval in $\mathbb{R}^{{10}}$
Example: Par Coords in a Vis Software in R

http://datamining.togaware.com
Ensuring Data Protection with k-Anonymization

Why are such approaches not used in enterprise hospital information systems?
Practical Example: Big data from Flow Cytometry (1)

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

Fulwyler, M. J. (1968) US Patent 3380584 A
Particle Separator, 1965 applied, 1968 published
Practical Example: Flow Cytometry (3) Immunophenotyping

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

Example: Limitations of 2D Parallel Coordinates

Streit et al. (2006)
Parallel Coordinates in 3D

Streit et al. (2006)

Source: http://orange.biolab.si/
Slide 9-33 RadViz Principle

1) Let us consider a point \( y_i = (y_1, y_2, \ldots, y_n) \) 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:

\[
\sum_{j=1}^{n} (\hat{S}_j - \hat{u})y_i = 0 \\
\sum_{j=1}^{n} \hat{S}_j y_j = \hat{u} \sum_{j=1}^{n} y_j
\]

\[
\hat{u} = \frac{\sum_{j=1}^{n} \hat{S}_j y_j}{\sum_{j=1}^{n} y_j} \\
u_1 = \frac{\sum_{j=1}^{n} y_j \cos(\alpha_j)}{\sum_{j=1}^{n} y_j} \\
u_2 = \frac{\sum_{j=1}^{n} y_j \sin(\alpha_j)}{\sum_{j=1}^{n} y_j}
\]

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

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

2. Now place the dimensional anchors

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

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

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

- Arrange $N$ axes on a circle in $\mathbb{R}^2$
- $3 \leq N \leq N_{\text{max}}$
  
  *Note:* An amount of $N_{\text{max}} \leq 20$ 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 just points on the axes
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)
Visual Analytics is intelligent HCI

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?

3) What layout algorithm should be used to visualize each facet?

4) What interactive techniques should be used for each facet and for which infovis tasks?

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;

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)

What are temporal analysis tasks?
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 preprocessing;

**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.)

How do you evaluate visualizations in the real-world?
- 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
- 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?
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://people.cs.uchicago.edu/~wiseman/chernoff](http://people.cs.uchicago.edu/~wiseman/chernoff) (Chernoff Faces in Java)
- [http://lib.stat.emu.edu](http://lib.stat.emu.edu) (Iris sample data set)
- [http://graphics.stanford.edu/data/voldata](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.\(^1\)

**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

Appendix: Deep View Working Environment - Swiss PDB

http://www.expasy.org
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

Repetition: From Physics of Light to Cognition of Thought

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

\[ x_i = [x_{i1}, \ldots, x_{in}] \]

- “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
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.

Slide 9-45 Example Algorithms for Selection

- Scatterplot-Select (xDim, yDim, xMin, xMax, yMin, yMax
- 1 $s \leftarrow 0 \triangleright$ Initialize the set of records
- 2 for each record $i \triangleright$ For each record,
- 3 do $x \leftarrow \text{NORMALIZE}(i, xDim) \triangleright$ derive the location,
- 4 $y \leftarrow \text{NORMALIZE}(i, yDim)$
- 5 if $x_{\text{Min}} < x < x_{\text{Max}}$ and $y_{\text{Min}} < x < y_{\text{Max}}$
- 6 do $s \leftarrow s \cup I \triangleright$ select points within rectangle
- 7 return $s$

Point-in-Point-Polygon(xs, ys, numPoints, x, y)
- 1 $j \leftarrow \text{numPoints} - 1$
- 2 oddNodes $\leftarrow$ false
- 3 for $i \leftarrow 0$ 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 $\leftarrow$ not oddNodes
- 7 $j \leftarrow I$
- 8 return oddNodes

40 sec of neonatal Polysomnographic recording

EEG signal from 8 ref. derivations: FP1, FP2, T3, T4, C3, C4, O1, O2

- EOG = Electrooculogram
- EMG = Electromyogram
- PNG = Pneumogram
- ECG = Electrocardiogram

Visual comparison of clustering results

Expert classification:
AS - active sleep,
QS - quiet sleep,
WK - wakefulness

Representation of final clusters:
clustering into 9 groups, displayed channels:
EEG, EOG, EMG, ECG and PNG

Gerla et al. (2009)
Using a unique colour for each cluster segment