Interactive Information Visualization and Visual Analytics

Andreas Holzinger
VO 709.049 Medical Informatics
09.12.2015 11:15-12:45

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

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; (Anonym: univariate = one-dimensional);

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 semantic 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;

Schedule

- 1. Intro: Computer Science meets Life Sciences, challenges, future directions
- 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
- 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;
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

Visualization is an essential part of Data Science

Verbal Information versus Visual Information

Problem: Context!

Semantic Ambiguity – Missing Context

A picture is worth a thousand words?
Informatics as Semiotics Engineering

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

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

Slide 9-9 Three examples for Visual Languages

Slide 9-10 Informatics as Semiotics Engineering

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

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;
Visualization is a typical HCI topic!


Usefulness of Visualization Science

What do you see in this picture?

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

Slide 9-19 Systematic Visual Analytics > Content Analytics

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) 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;

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

- **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 $X_i$ 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 $\mathcal{C}$.

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

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

---

A polygonal line $\mathcal{P}$ 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 $p_{i-1}, p_i$ marked on the $X_{i-1}$ and $X_i$ 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.

---

Heavier polygonal lines represent end-points.

---

Example: Par Coords in a Vis Software in R

---

Par Coords -> Knowledge Discovery in big data

---

A. Holzinger 709.049

WS 2015
Why are such approaches not used in enterprise hospital information systems?
- 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)

---

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


Source: http://orange.biolab.si/
1. Normalize the data to the interval (0, 1)
\[ x_j = \frac{x_j - \min_j}{\max_j - \min_j} \]
2. Now place the dimensional anchors
3. Now calculate the point to place each record and to draw it:
\[ y_j = \sum \frac{x_j}{\sum_i x_{ij}} \]
\[ s_i = \sum x_{ij} \]


Slide 9-35 RadViz for showing the existence of clusters


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


Slide 9-37 Star Plot production

- Arrange \( N \) axes on a circle in \( \mathbb{R}^2 \)
- \( 3 \leq N \leq N_{\text{max}} \)
  - \( N_{\text{max}} \leq 20 \) 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 points on the axes

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

angle = \( 2 \pi N \) / for each \( a \), from axes[]
angle = \( i \) * angle
\( x = \text{mid}_x + r \times \cos(\text{angle}) \)
\( y = \text{mid}_y + r \times \sin(\text{angle}) \)
\( \text{DrawLine(midpoint.x, midpoint.y, x, y)} \)

maxi = ai.upperBound()
scaled_vali = ai.value() * r / maxi
\( x_{\text{vali}} = \text{mid}_x + \text{scaled_vali} \times \cos(\text{angle}) \)
\( y_{\text{vali}} = \text{mid}_y + \text{scaled_vali} \times \sin(\text{angle}) \)
\( \text{DrawLine(x_{\text{vali}}, y_{\text{vali}}, x_{\text{vali-1}}, y_{\text{vali-1}})} \)

Slide 9-39 Visual Analytics is intelligent HCI
Slide 9-40 Design of Interactive Information Visualization

- What facets of the target information should be visualized?
- What data source should each facet be linked to and what relationships these facets have?
- What layout algorithm should be used to visualize each facet?
- What interactive techniques should be used for each facet and for which infobeats?

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

- Overview: Gain an overview about the entire data set (know your data!)
- Zoom: Zoom in on items of interest
- Filter: Filter out uninteresting items – get rid of distractors – eliminate irrelevant information
- Details-on-demand: Select an item or group and provide details when needed
- Relate: View relationships among items
- History: Keep a history of actions to support undo, replay, and progressive refinement
- 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)

Slide 9-43 Rapid Graphical Summary of Patient Status

- Data of Admission: 1-yr, 2-yr, X-yr
- Glucose: 237 mg/dl

Slide 6-44 Example Project LifeLines

- 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.)


### Example: Subspace Clustering

- **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


### Similar: Principal Component Analysis (PCA)

*Curve of Dimensionality (Bellman, 1977)*: As more dimensions are available, data becomes more sparse and distance measures become meaningless.
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?

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

Visual Multidimensional Geometry and its Applications (1)

Appendix: Node-link graphs to visualize biological networks

Appendix: Deep View Working Environment - Swiss PDB
Appendix: Visual Analytics for Epidemiologists

![Visual Analytics](image1)


Appendix: Motion Analysis & Visualization of Elastic Models

![Motion Analysis](image2)


Typical direct image

![Direct Image](image3)


Repetition: From Physics of Light to Cognition of Thought

![Repetition Diagram](image4)


Remember

![Remember Icon](image5)
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 (C→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

Example: Star Plot Diagram - Radar Chart


The Noisy Channel


Slide 9-46 40 sec of neonatal Polysomnographic recording

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, ECG, EMG, PNG

Gerla et al. (2009)

Using a unique colour for each cluster segment