An Introduction to Visual Analytics

Outline

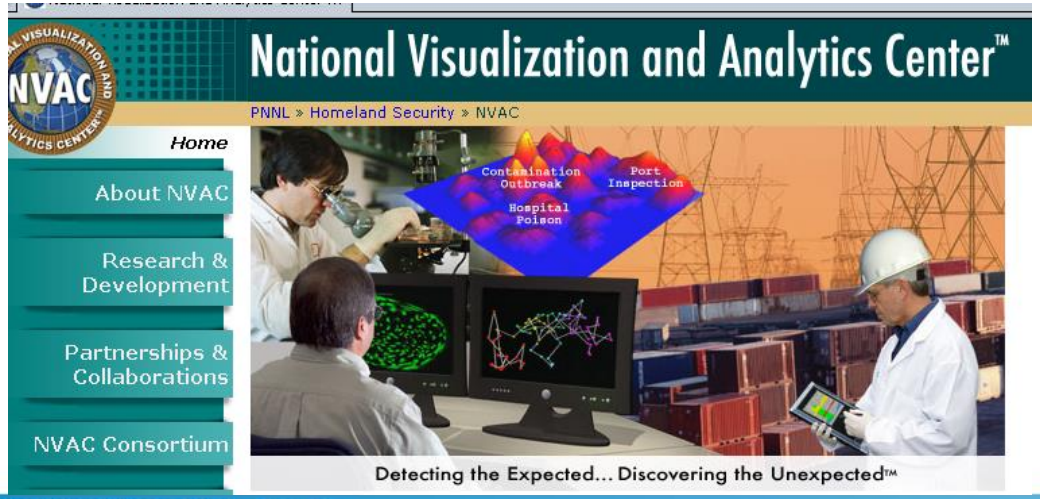
1. Introduction
   - What is „Visual Analytics“?
   - Why „Visual Analytics“
   - Brief History
   - Involved Displicines
2. Scaleability
3. Components of Visual Analytics
4. Strategies and Principles
5. Major Application Areas

What is visual analytics?

- The term „Visual Analytics“ was coined in 2005 by Jim Thomas, the NVAC (National Visualization and Analytics Center):

  „Visual Analytics is the science of analytical reasoning facilitated by interactive visual interfaces”

- Combining approaches from different disciplines to enable the exploration of very large and heterogeneous datasets
- Primarily undirected search for trends and structures
- Major initial focus in the US: Homeland-Secuity



What is visual analytics?

Emerged from information visualization and automatic data analysis. An *integrated* combination of *data analytics* and *interactive visual exploration*.

- **Data analytics**: Machine learning techniques, such as pattern mining, classifiers, clustering, supervised learning techniques (Neural networks, support vector machines), dimension reduction
- **Interactive visual exploration**: Visual filters, multiple coordinated views, brushing and linking, parallel coordinates, scatter plots, glyph-based displays, time-lines, graphs and treemaps.
- **Integrated**: Analytics is not just a preprocess, but carefully integrated in the exploration, thus loops between infovis. techniques and analytics are supported.

What is visual analytics?

- Beyond algorithms for analyzing and displaying data, visual analytics aims at supporting analysts.
- This requires an understanding of reasoning processes, decision making and common sense.
- This understanding should be reflected in interaction techniques to explore the data with carefully chosen methods and parameters and in layouts that present results adequately.
- For complex reasoning processes, guidance may be required to support analytical workflows.

Organisatorisches

Hörerkreis: WB CV-Bachelor ab 6; WB INF-Bachelor ab 6; WB IngINF-Bachelor ab 6; WB WIF-Bachelor ab 6; WPF DKE-Master 1-3; WPF DigiEng-Master 1-3
Abschluss: Prüfung (mündlich)
ECTS-Credits: 5

Prüfungsvoraussetzungen:

- Mindestens 2/3-Anwesenheit in den Übungen
- Mindestens 2/3 aller Punkte in den Übungen
- Präsentation von mindestens 2 Hausaufgaben in den Übungen
- Rechtzeitige Prüfungsanmeldung (ca. vier Wochen vorher!)

Organisatorisches

- Slides: vismd.de → Teaching → Visual Analytics (Link)
- User: lsg_student, Password: analytics@isg
- **Übung 1:**
  Ort: G29-426
  Zeit: Mi., 13:00 bis 15:00
  1. Übung: 12.04.2017
  Anmeldung zur Übung über LSF
- **Übung 2:**
  Ort: G29-426
  Zeit: Do., 13:00 bis 15:00
  1. Übung: 13.04.2017
  Anmeldung zur Übung über LSF

Übung

- Die Fähigkeit, durch analytisches Denken und intelligente Kombination von Algorithmen des Machine Learnings mit Visualisierung aus großen Datenmengen für eine bestimmte Problemstellung nützliches Wissen, Handlungsempfehlungen abzuleiten, erlernt man durch praktische Erfahrung.
- **Ziel**: Handwerkszeug zur Realisierung von datengetriebenen Lösungen zu vermitteln.
- Ausgewählte Konzepte und Algorithmen aus der Vorlesung vertiefen und an interessanten Anwendungsbeispielen visualisieren. Modelle, die versuchen zu erklären,
  - warum die wenigsten Umfrageinstitute auf einen Sieg von Donald Trump gesetzt haben,
  - ob es einen Zusammenhang zwischen Bruttoinlandsprodukt und Lebenserwartung der Bevölkerung eines Landes gibt und ob sich der Zusammenhang verändert hat.

Literature and Links

J. J. Thomas and K. A. Cook. Illuminating the Path: The Research and Development Agenda for Visual Analytics. IEEE Press, 2005.

D. Keim, J. Kohlhammer, G. Ellis, F. Mansmann (eds): Mastering the Information age - solving problems with Visual Analytics (Vismaster Roadmap) , Eurographics Association, 2010, http://www.vismaster.eu/book/

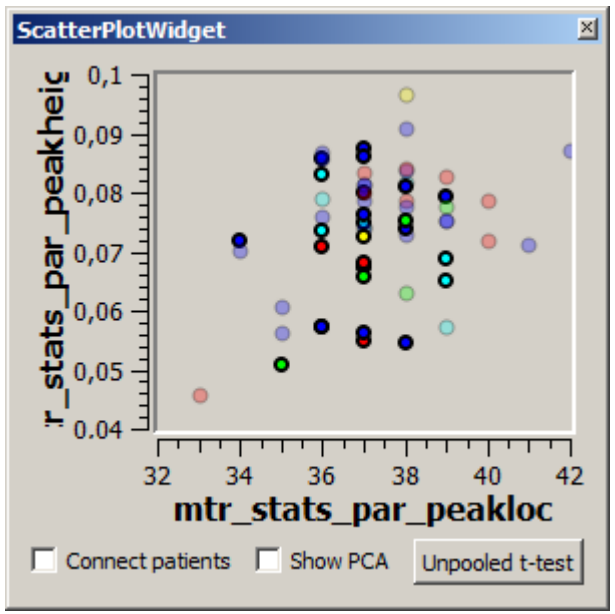
VisMaster Project: http://www.vismaster.eu/

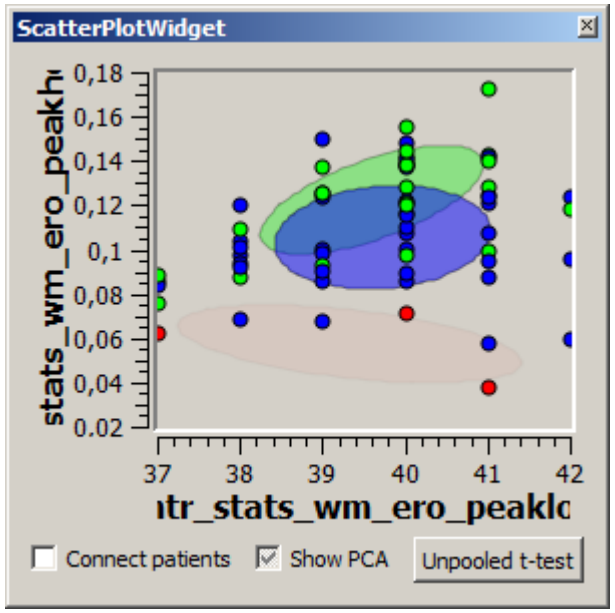
DFG Schwerpunktprogramm Scalable Visual Analytics, http://www.visualanalytics.de/

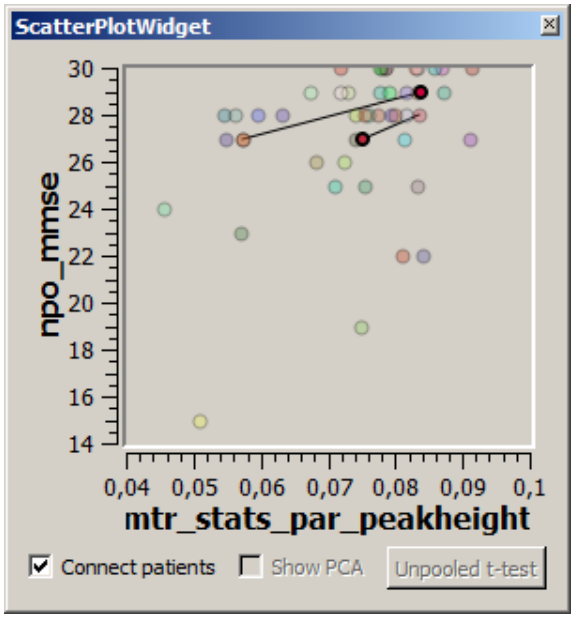
Visual Analytics

Scatterplots with overlaid clusters and (local) regression lines.

Problem: Visualization of categorical data







From: [Steenwijk et al., 2010]: Visual Analysis of Cohort Study Datasets

Why Visual Analytics?

- For *knowledge discovery*. In earlier times, knowledge was acquired by interviewing experts – a tedious, incomplete and error-prone process.
- Instead of interviewing experts, knowledge is inferred from analyzing data (machine learning, since the 1990s).
- Many developments in computing (performance, storage capacity) and other digital devices (variety of sensors, digital cameras and devices that acquire process states) lead to an ever increasing amount of information that can be used to identify frequent patterns, correlations, …

„The „data scientist“ will be one of the most important jobs in the US in the next decade.“ (McKinsey, 2011)

Why Visual Analytics?

- Information is no more the bottleneck.
- Instead, analytical capabilities are essential.
- Information is often huge, high-dimensional, inconsistent and dynamic
  - Not clean data perfectly described in a data dictionary of a data base with full information on attributes, consistency rules, …
- Visual analytics provides flexible means to explore such data, to derive information in a user-steered process that creates trust in the results.

Why Visual Analytics?

- „To gain insight into today‘s large data sources, data mining extracts interesting patterns. To generate knowledge from patterns and benefit from human cognitive capabilities, meaningful visualizations of patterns are crucial.“
- „Human experts are able to quickly identify both correlations and irregularities if data or results are appropriately visualized.“
- „For data analysis or mining tasks, visualization is the central step from patterns to knowledge in a discovery process.“ (Assent, 2007)

Why Visual Analytics?

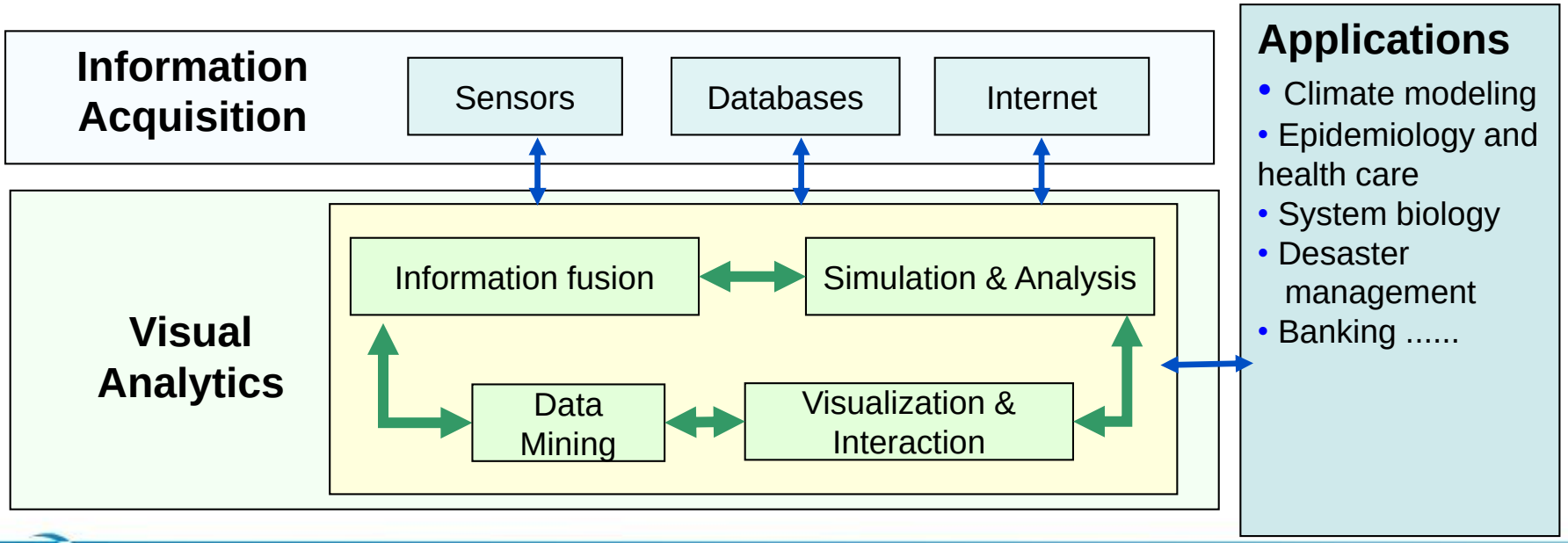
- The promise of visual analytics is to explore the challenging data *under different perspectives* leading to new insights about
  - *trends*,
  - *patterns* and
  - *relations* as well as
  - new hypothesis

  to be confirmed with either statistics or new targeted data acquisition.
- The evaluation of visual analytics systems, thus, is *focussed on insights and hypothesis* generation, not on usability, performance.
- Listen to the experts: Conversation with J. Thomas and P. Hanrahan

Introduction

What is the motivation for„Visual Analytics“?

- New applications and questions, in particular in science (data science) → analysis of data as a scientific method in addition to experiments, reasoning, …



When NOT to use Visual Analytics?

- When datasize is moderate and data is of high quality and just needs to be visualized, e.g. a weather map
  - Information visualization mantra: Overview first, zoom and filter, then details on demand (Shneiderman, 1996)
- When a well-defined routine problem can be solved by non-interactive analytical methods, such as statistics (correlation, regression, ANOVA) or optimization (finding optimal parameters)
- Most problems do not fall in the above-described categories and require iterative trial-and-error strategies to be solved.

Introduction

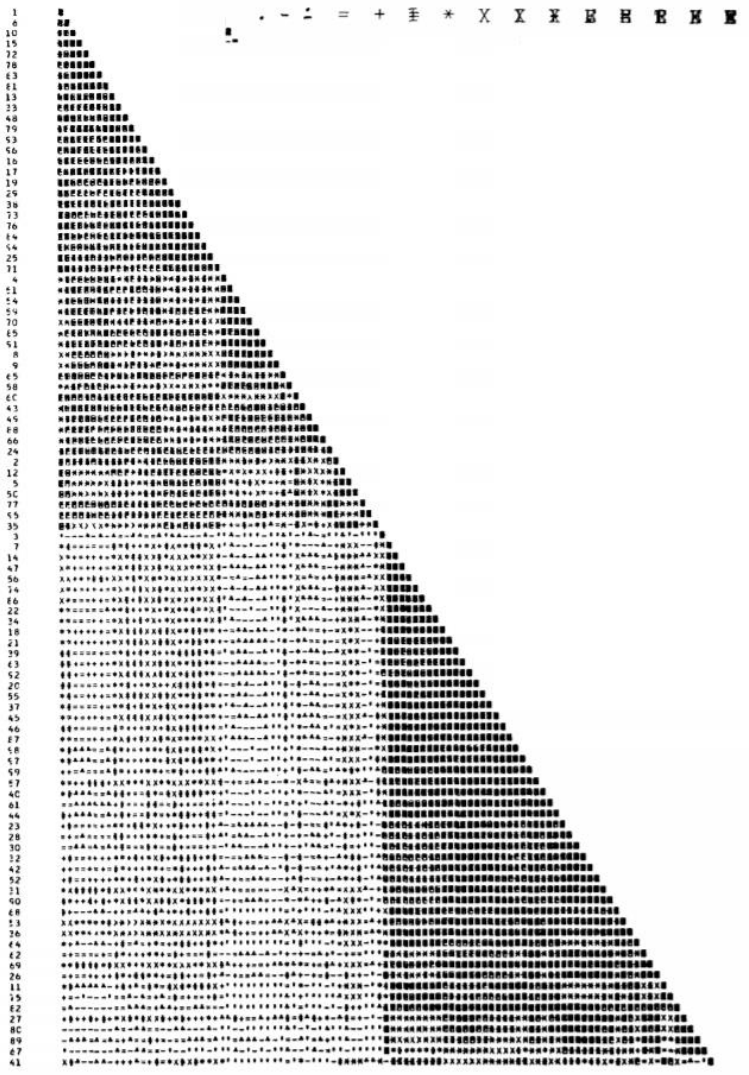
- The orign of different fields leads to inconsistent terminology, e.g. in database technology there are *attributes*, in statistics there are *variates* (multivariate statistics) and in visulization the same is referred to as *dimensions*.
- We use primarily the visualization terminology.

Introduction

Brief history:

- Early papers on combining data analysis and visualization for evaluation, e.g. Ling, 1973 „A computer generated aid for cluster analysis“
- Contribution from the statistics community: Exploratory data analysis (Tukey, 1977)
- Keim D. A., Kriegel H.-P.: “VisDB: Database Exploration Using Multidimensional Visualization*”, IEEE Computer Graphics and Applications*, 1994
- Ankerst M.: ”Visual Data Mining”,Ph.D. thesis, Univ. of Munich, 2000
- Niggemann, Oliver: Visual Data Mining of Graph-Based Data. University of Paderborn, 2001
- Workshops on Visual Data Mining (2001-2003)
- Essential articles by Keim, Schumann 2002 on Visual Data Mining

Introduction



Dissimilarity matrix after clustering

(From: Ling, 1973 „A computer generated aid for cluster analysis“)

Introduction

Brief history (continued):

- Many discussions between the US and Germany on Visual Analytics 2004/05 led by Jim Thomas
- Since 2006, regular IEEE Symposium on Visual Analytics Science and Technology (VAST)
- EU VisMaster Project to support the connection between visual analytics researchers (2008-2010)
- In Germany: DFG Priority Programme „Scaleable Visual Analytics“ (2008-2015)
  - Series of projects related to text documents, patents, video data, epidemiological data. Project teams involved data mining and database experts, simulation and image analysis experts, as well as HCI and visualization researchers

Involved Disciplines

Visualization

- InfoVis
- SciVis
- Graph Drawing
- Rendering

Analysis

- Graph Mining
- Image Mining
- Data Mining
- Model generation & Knowledge Discovery

Interaction

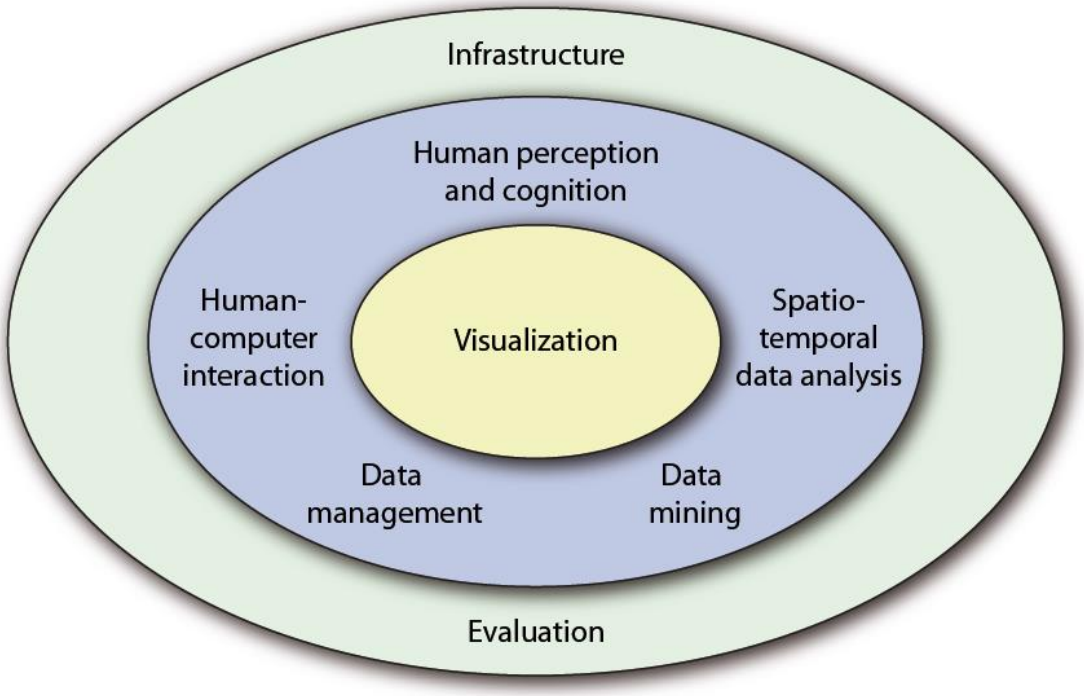
- Cooperative User Interfaces
- Model-based User Interfaces
- Usability and User Experience

Further Disciplines

- Cognition Science (Decision making)
- Visual perception

Involved Disciplines

A visualization-centered view on the relation between these disciplines



Outline of the Lecture

- Scatterplot-Based Visual Representations
- Dimension reduction & Projection
- Cluster Analysis: Methods
- Cluster and Outlier Visualization
- Biclusters and Subspace Clustering
- Visualization of Association Rules
- Visualization of Decision Trees
- Visual Analytics with Regression Models
- Spatio-Temporal Analysis
- Cooperative Visual Analytics
- Visual Analytics in Epidemiology (Statistics)
- Visual Analytics in Public Health (Temporal Data)

Scaleability

The Scalabality Challenge (according to Thomas, Cook):

- Information Scalabilty

  Selection of relevant data and levels of detail from very large data

  – What should be displayed? (information space)

- Visual Scalability (recall Eick, Karr, InfoVis III)

  Selection and parameterization of visualization techniques, design of Visual Cues

  – How to display data (space of presentation variables)

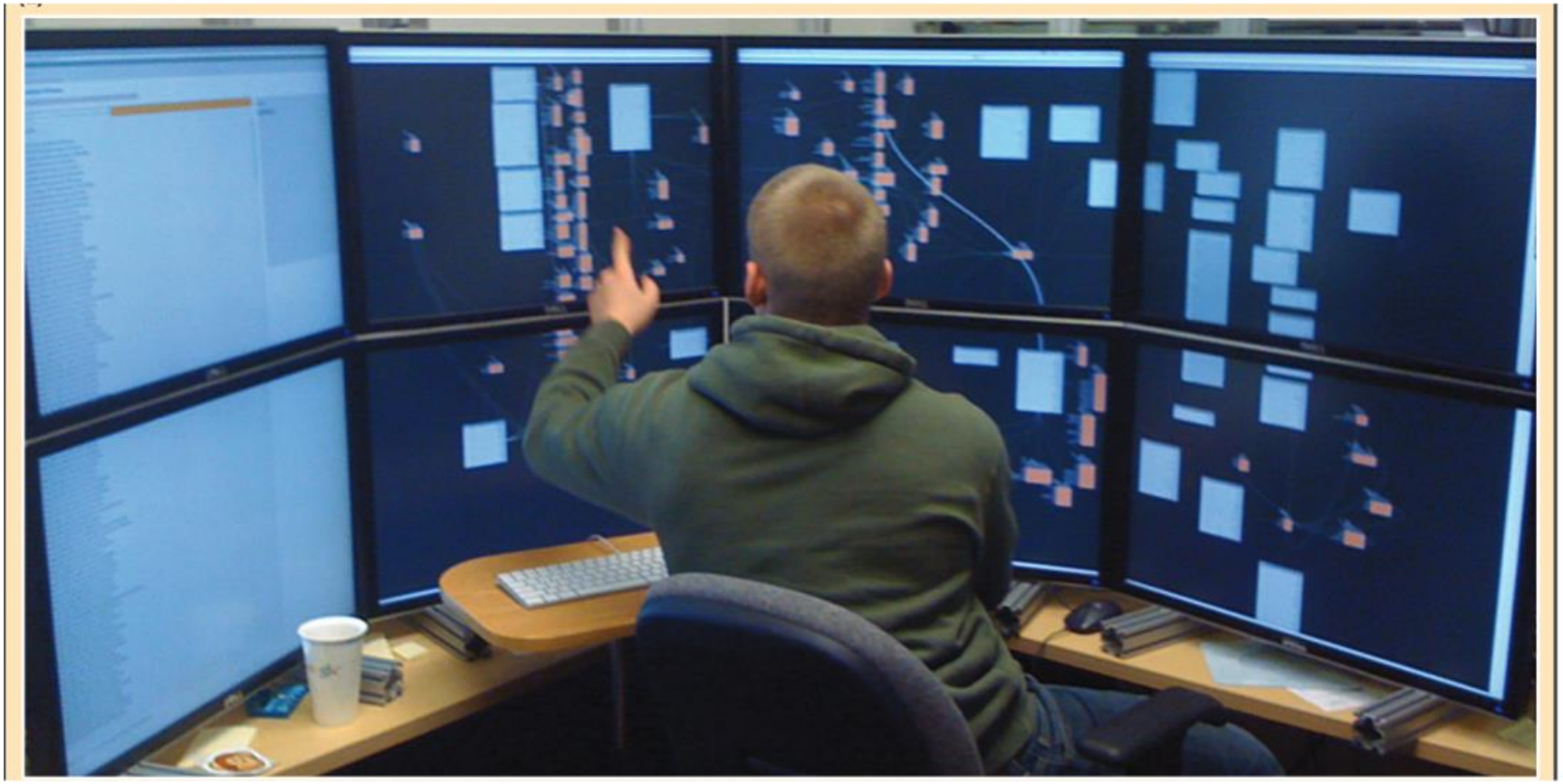
Scaleability

- Display Scalability
  - Adaptation of the visualization to different display sizes

  (Solutions with 8 monitors may have 10.000 x 3.000 pixels)
  - Where something should be displayed? (device space)

- Human Scalability
  - Adaptation of the data and the visual representations to user groups
  - Who belongs to the target user group?

Scaleability



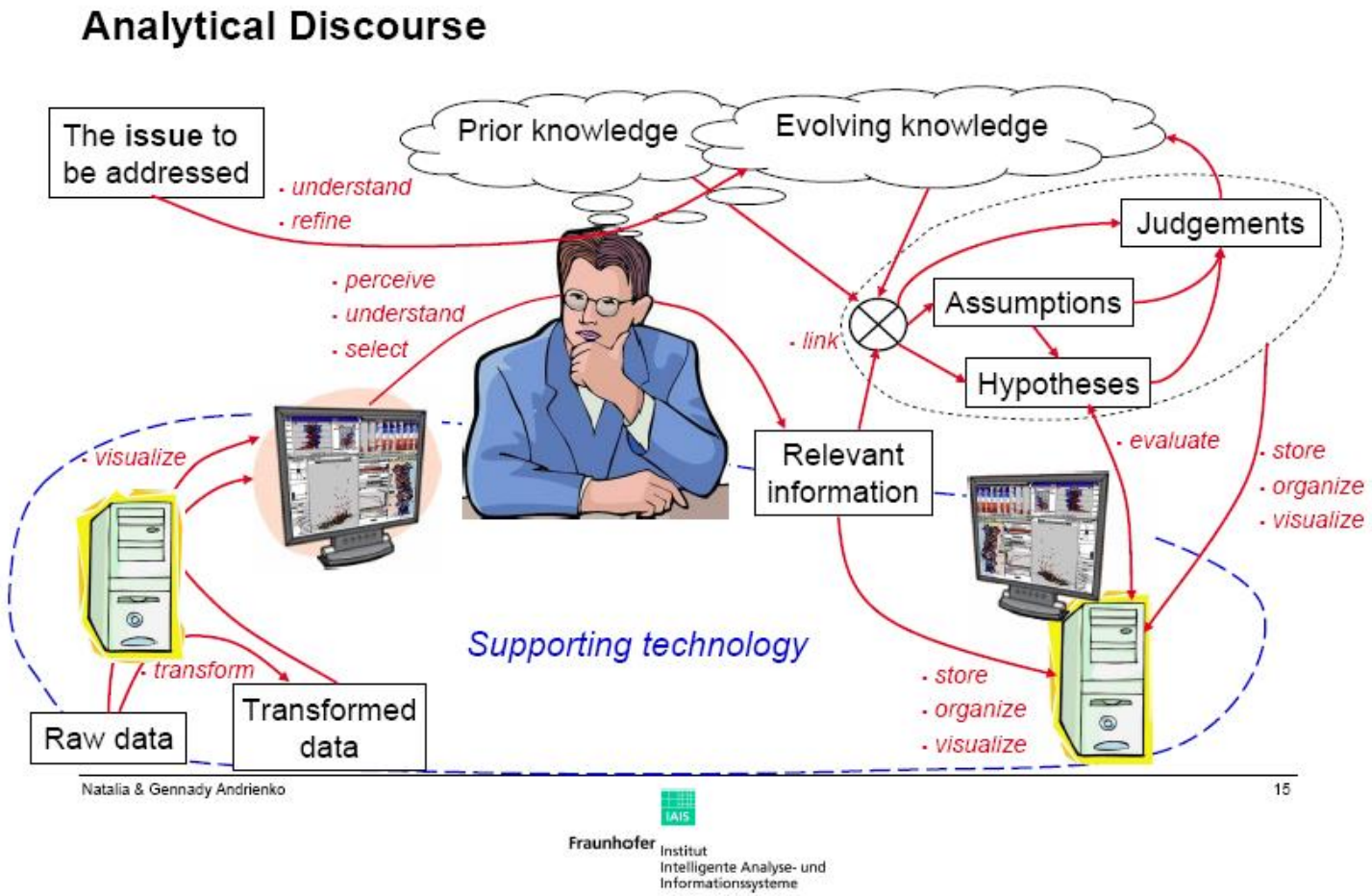
Professional visual analytics systems often employ large workspaces, large clusters of computers for heavy computation including advanced data management (From: Fiaux, 2013)

Components

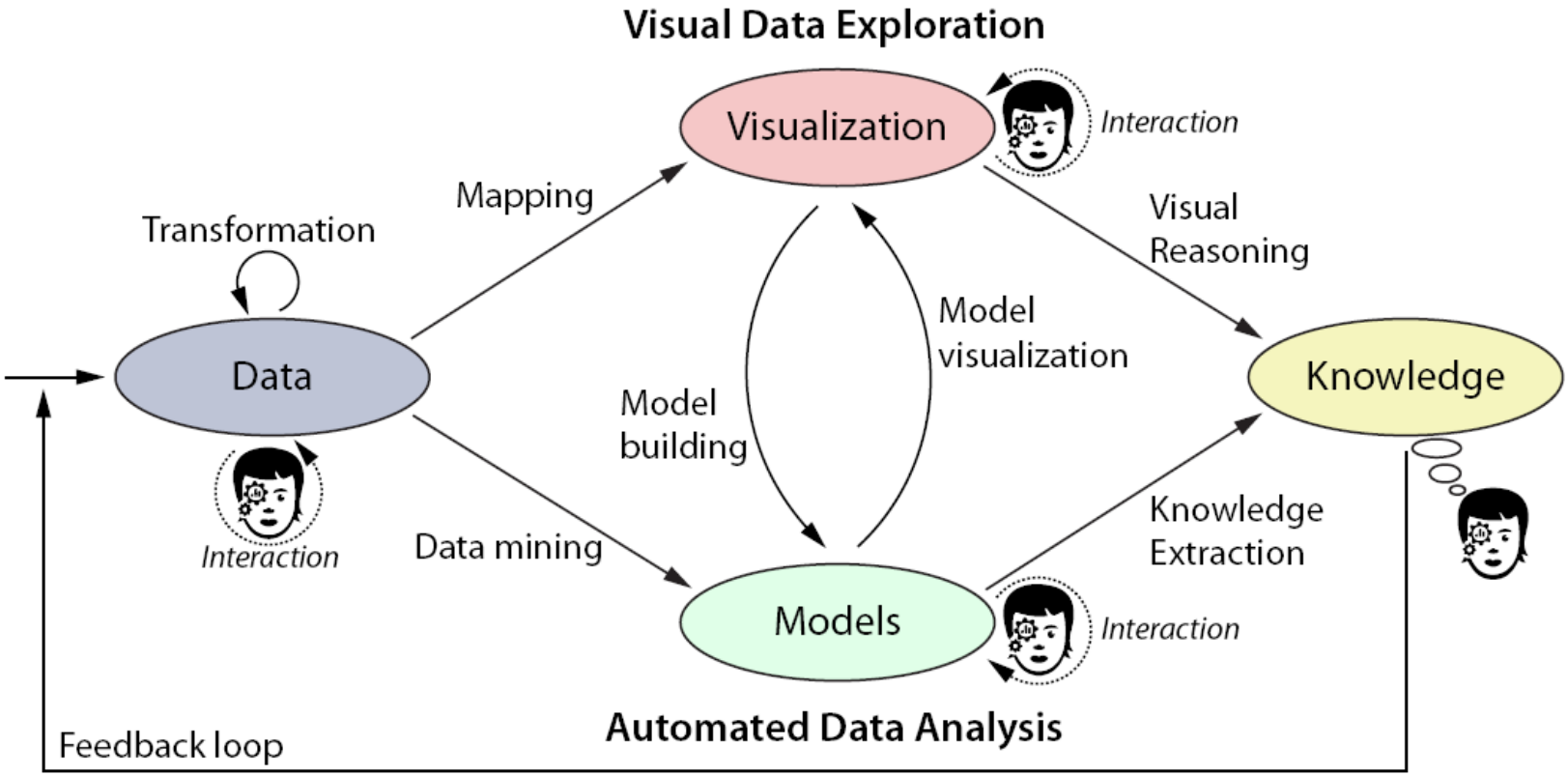
Components of Visual Analytics (according to G. Andrienko)

- Analytical reasoning
  How to maximise human capacity to perceive, understand, and reason about complex and dynamic data and situations?
- Visual representations and interaction techniques
  How to augment cognitive reasoning with perceptual reasoning through visual representations and interaction?
- Data representations and transformations
  How to transform data into a representation that is appropriate to the analytical task and effectively conveys the important content?
- Production, presentation, and dissemination
  How to convey analytical results in meaningful ways to various audiences? (Also to justify the efforts of visual analytics)

Components


Analytical Discourse
The issue to be addressed
Prior knowledge
Evolving knowledge
· understand
· refine
· perceive
· understand
· select
· link
Judgements
Assumptions
Hypotheses
· evaluate
· store
· organize
· visualize
Relevant information
· visualize
· transform
Supporting technology
Raw data
Transformed data
· store
· organize
· visualize
Natalia & Gennady Andrienko
15
Fraunhofer Institut Intelligente Analyse- und Informationssysteme

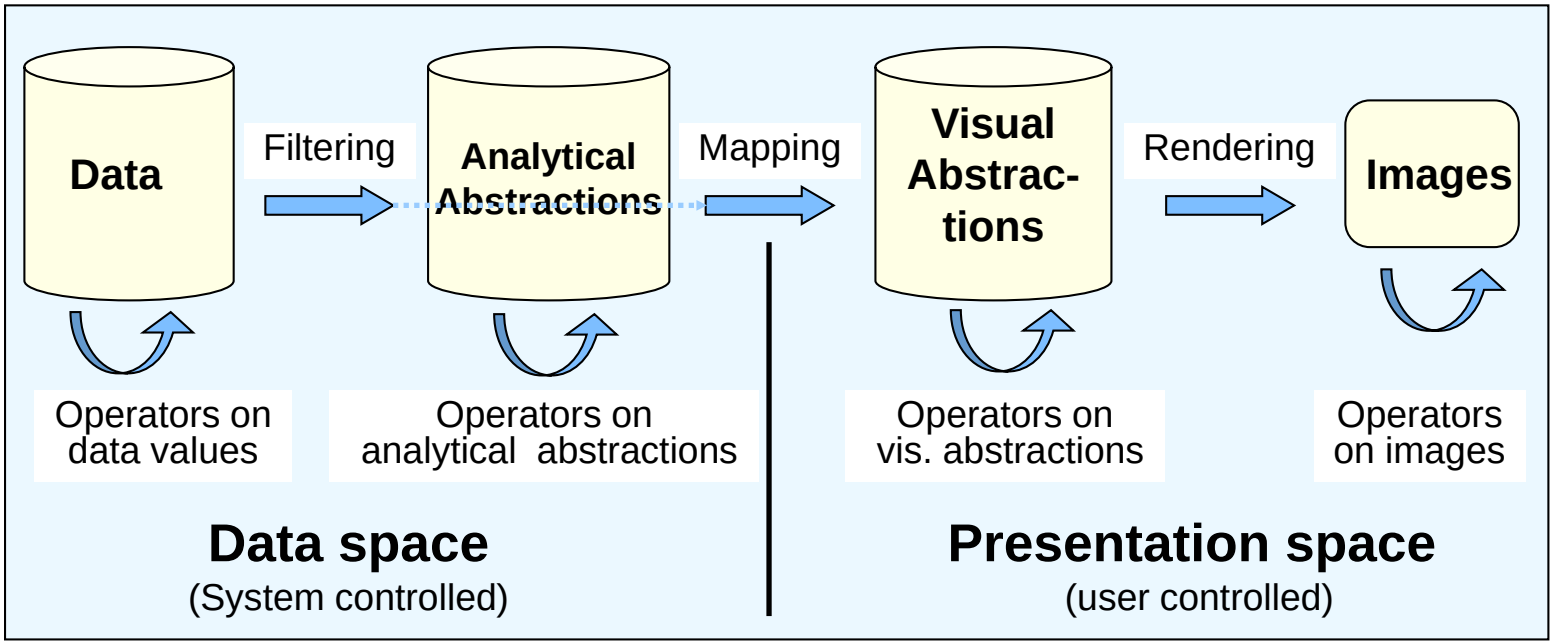
Components



A more concrete sketch of a visual analytics system. Analytics (data analysis) is applied before and after visualization (Courtesy of Steffen Oeltze-Jafra based on Daniel Keim)

Components

- For the combination of different methods to support the analytical discourse the visualization pipeline needs to be extended. Suggestion by H. Schumann (Rostock) based on the Data State Reference Model (Chi, 2000)



Components: Filtering

Filtering operates on data values and includes

- Data cleansing
- Analytical computations, e.g. statistics
  - to generate aggregated data (different levels of data)
  - to generate derived data (e.g. rates of change by gradient computation)
  - to create a hierarchy of the data elements based on similarity for focussing further analyis (e.g. by applying bottom-up hierarchical clustering)

Strategies and Principles

Flexibility vs. Guidance:

- Analysts need a wide set of tools and considerable flexibility to use them (parameterization, combination)
- But: Analysts also benefit from *guidance*, from pre-defined workflows

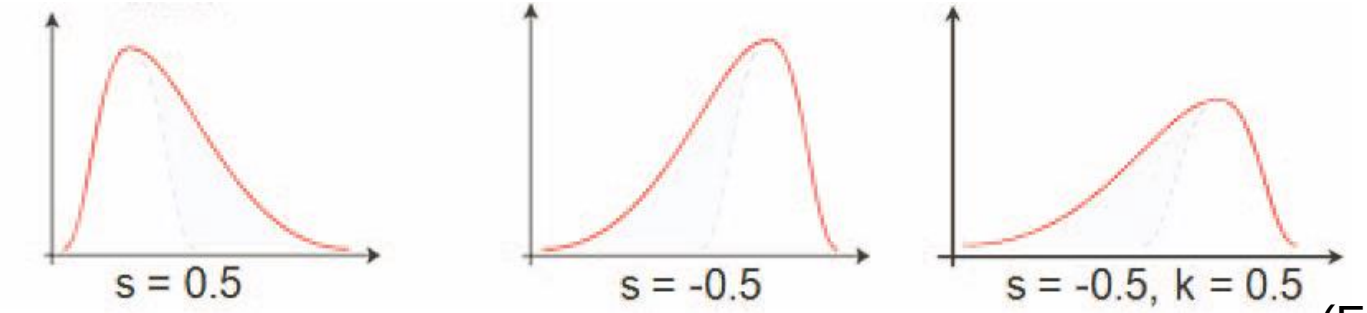
> „An orderly process of exploration is vital, even though there will … be … iterations, and shifts of attention from details to overviews and back“ (Seo, 2004)

- Next slides: GRID Principles: Graphics, Ranking and Interaction for Discovery

Strategies and Principles

Strategies and principles suggested by Moore, 1999 and Seo, 2004

- Enable to analyze dimensions first (distributions)
- Enable exploration of relationships among dimensions
- Offer graphical displays and numerical summaries
- Rank the dimensions according to statistical scores, such as skewness (as a key of asymmetry) and kurtosis (as a key of peakness)
- Get insight and perform statistics to confirm

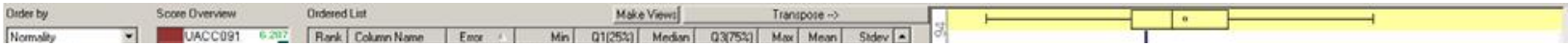


(From: Cao, 2011)

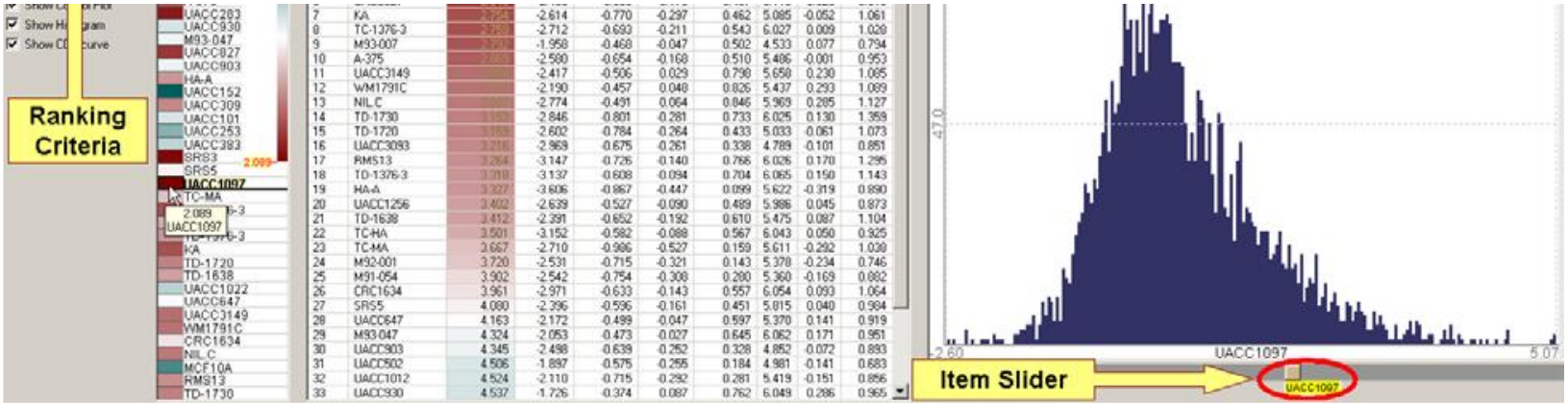
Strategies and Principles

Rank-by-feature framework (Seo, 2004)

- Provide an overview of correlations, e.g. with a scatterplot matrix
- Overlay statistical measures
- Rank the correlations, e.g. by certain types of regression or statistical power



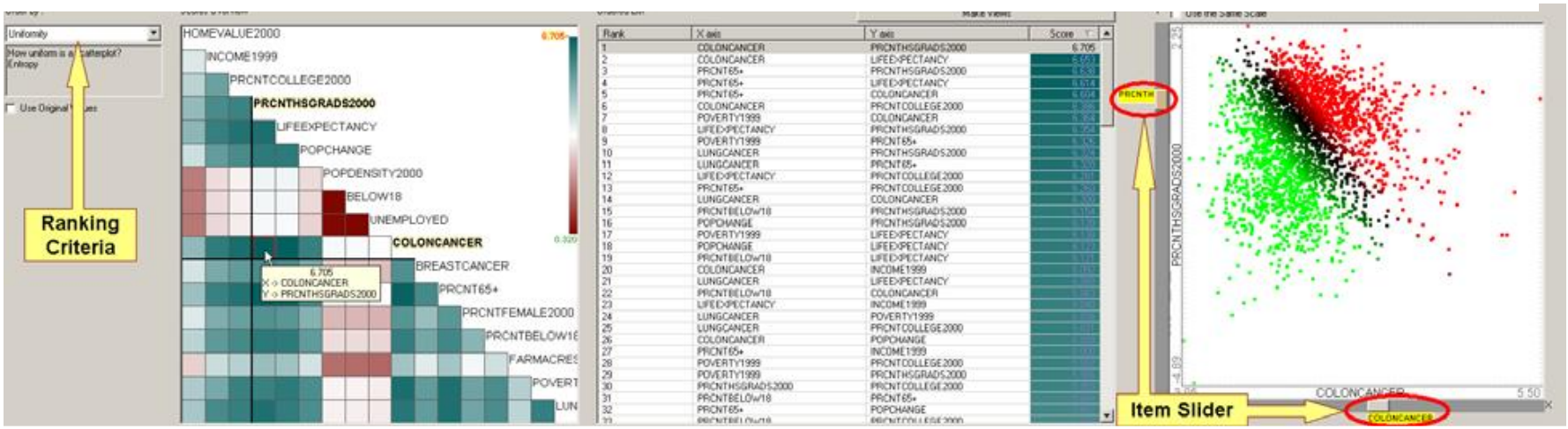
Control panel, score overview, ordered list, histogram browser



Select a ranking criteria for one dimension, sort the dimensions accordingly, provide statistical measures and show the distribution (histogram, boxplot) (From: Seo, 2004).

Strategies and Principles

Control panel, score overview, ordered list, Scatterplot



Explore relations in the rank-by-feature framework. Select a type of relation, present the permutation matrix, provide a list of related attribute pairs and show the selected pair in the scatterplot (From: Seo, 2004)

Rank-by-feature is an example for *metrics-based analysis*.

Drawback of rankings: margin between two items is not conveyed. Are the Top-N almost equal or is there a „boundary“ where the above items have much higher scores?

Reproducibility of Visual Analytics

- A key aspect of an analysts‘ work is to provide *reproducible* results.
- Thus, the interaction and analysis sequence *needs to be stored*.
- Undo of steps and redo must be supported.
- Results should be stored along with the current state of the system to later start interaction based on a previous result.
- Storage of all parameters/intermediate results is challenging → compression may be enabled.
- Interaction history may be visually represented as a tree that enables to go back to interesting points.

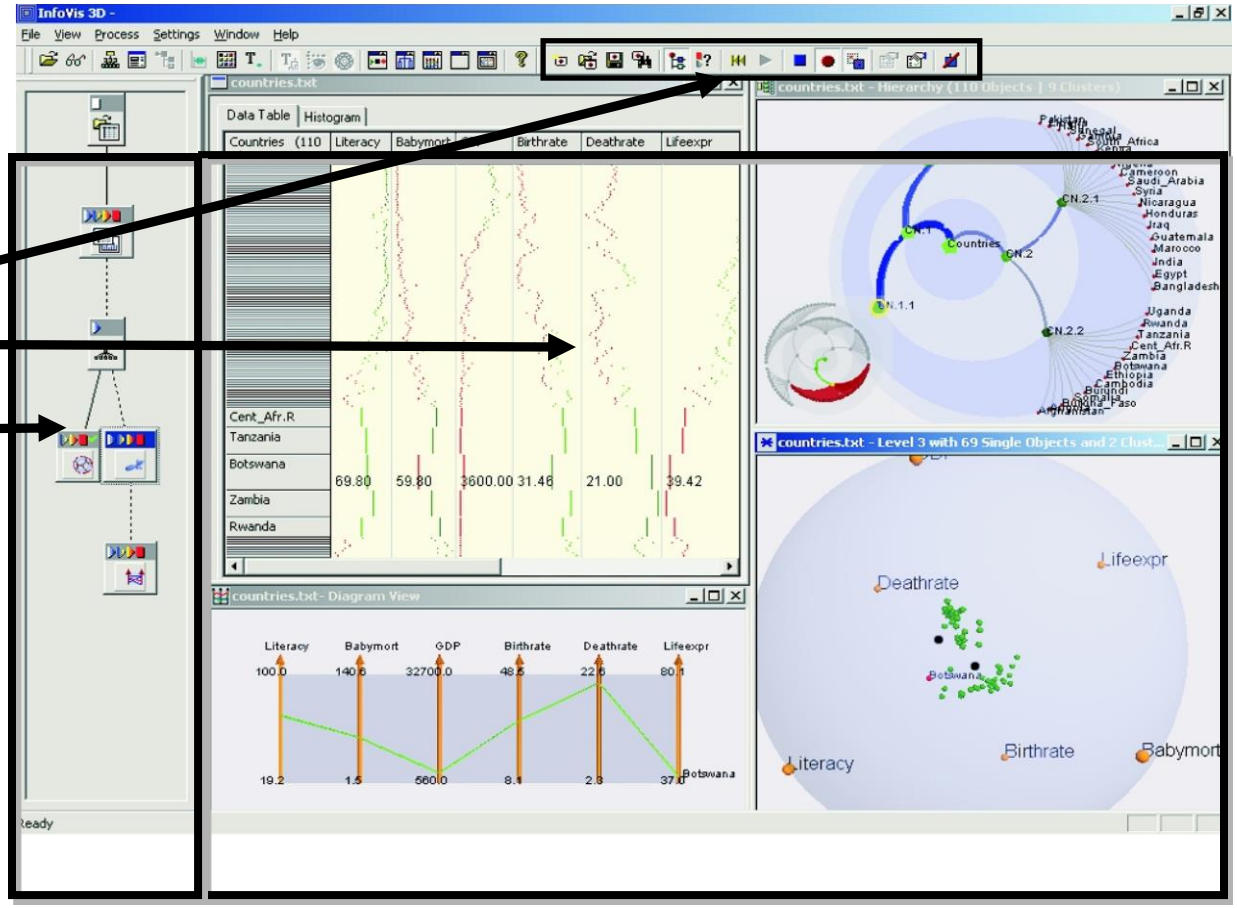
Reproducibility of Visual Analytics

Integration of a history tree in a visual analytics system

History Toolbar

Rendering view

History tree window



(Courtesy of Heidrun Schumann, University of Rostock)

Pair and Cooperative Visual Analytics

- Visual analytics benefits from tools that support different perspectives on the data AND from different analysis strategies from cooperating users.
- Cooperation
  - involves communication, e.g. explicit thoughts on which features are explored for what purpose.
  - may occur between domain scientists (engineers, business analysts, …) or between a computer scientist and a domain scientist (*pair analytics* ~ pair programming, (Ariaz, 2011)
- Cooperative visual analytics benefits from large displays, multiple input devices

Progressive Visual Analytics

- With large data and many dimensions, computations may take a long time.
- Progressive visual analytics (Stolper, 2014) means that reasonable intermediate results representative for the whole data are generated and presented.
- Lengthy computations can then be terminated and parameters adjusted.
- When new results are computed, the user is made aware of them (subtly) and can enable that these results are presented.
- Instead of interacting – waiting – interpreting-cycles, a (more) fluent exploration is supported.
- Algorithms have to be adapted, e.g. by performing a breadth-first instead of a depth-first search.

Major Applications

- Business and finance (Business intelligence)
  - Understand the time-varying data of the stock market, certificates, stock indices, currency relations in relation to business data (transactions, revenue, sales) and political decisions, e.g. related to interest rates (Zinssätze)
- Desaster/Emergency management
  - Support the decision making process in events such as large accidents, heavy pollution, epidemiologic growth of a disease, floodings or heavy storms
  - Decisions should minimize damage and save ressources.
  - Measures include evacuations and creation of physical barriers.
  - What is most urgent and most effective?
- Security
  - Detection of credit card misuse, virus software, …

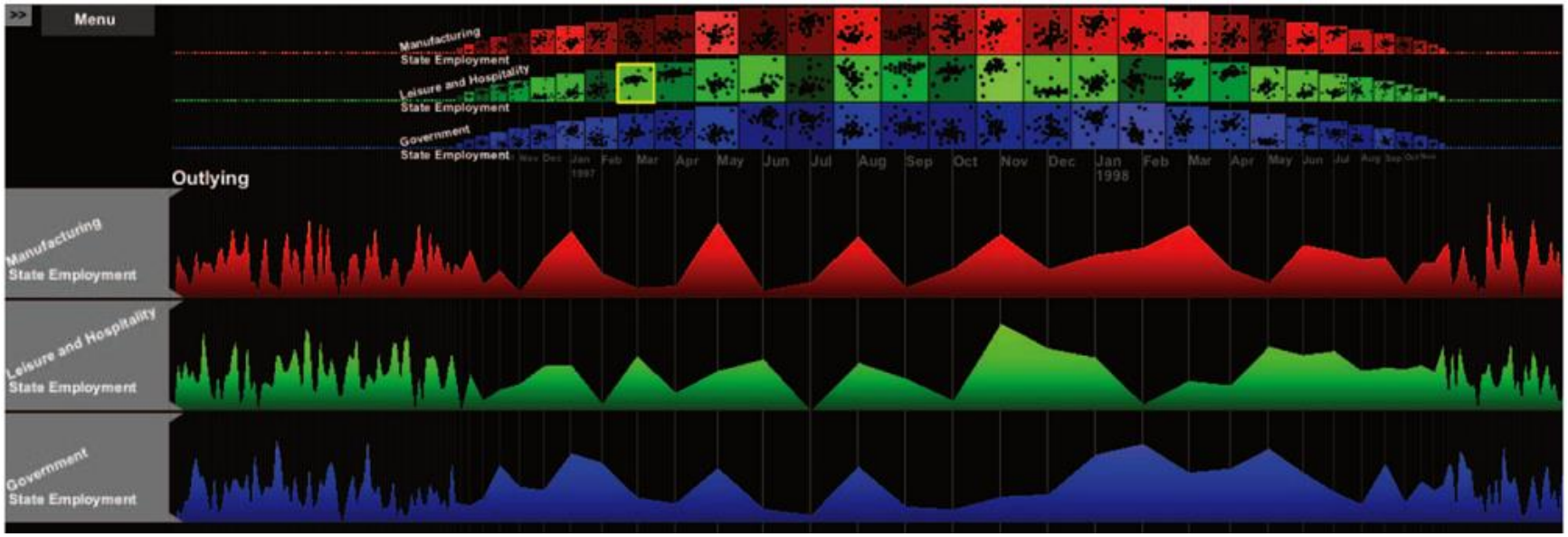
(List is inspired by a discussion in Keim, 2008 and by Sun, 2014)

Major Applications

- Sport analytics
  - Analyze movement patterns, e.g. in football and tennis, performance data, health attributes to provide an objective and high-level description of the performance of individual sportsmen and whole teams
- Astronomy
  - Huge amounts of data about the universe are available. Extract potentially relevant information to characterize developments and identify new relations.
- Network analytics
  - Network traffic is analyzed with various sensors to detect anomalies and potential intruders
- Climate and weather research
  - Based on a multitude of sensors, reflecting temperature, pressure

Major Application Areas

Visual Analytics of Temporal Data

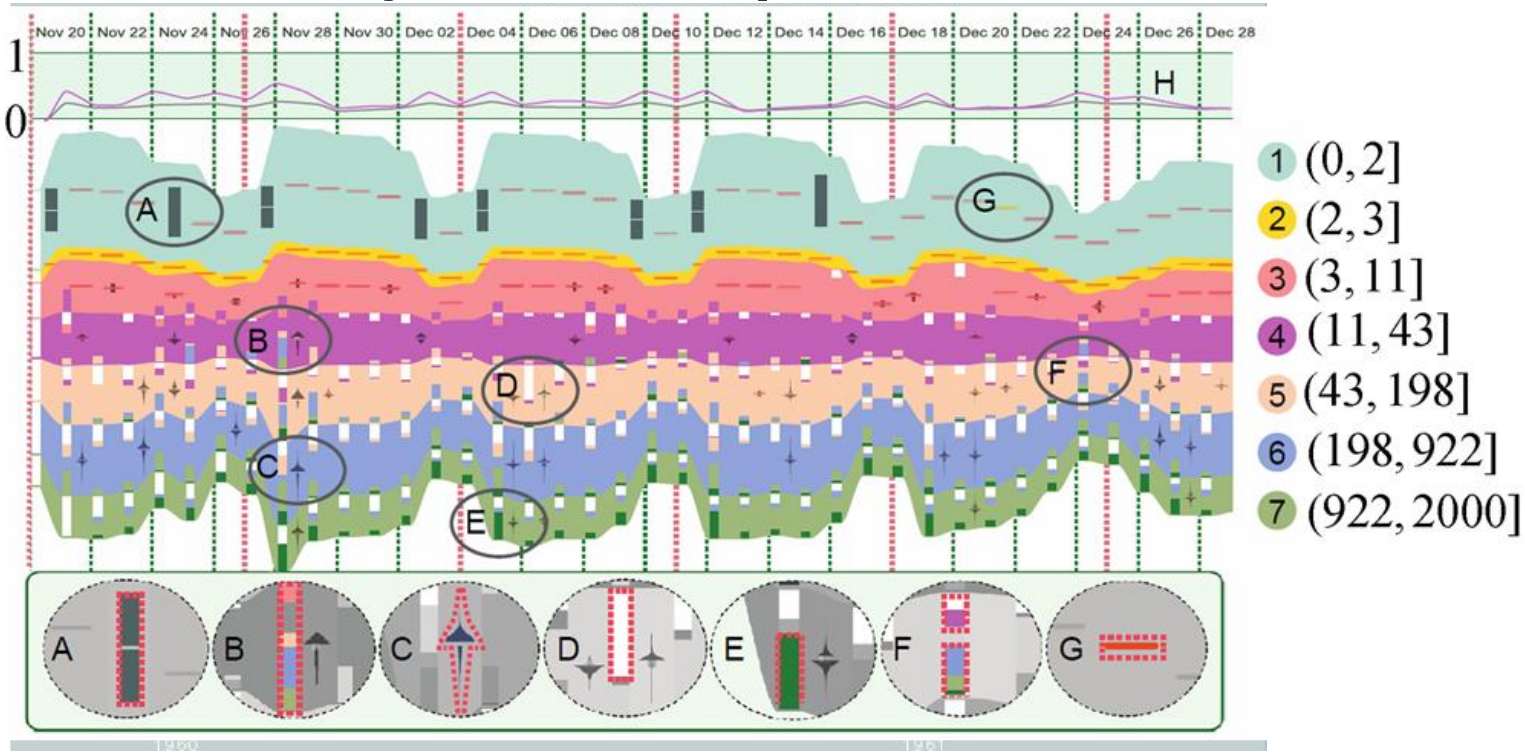


Series of US Employment Data. Visualized and analyzed with TimeSeer (Dang, 2013). Data comprises 10 sectors, 50 US States and 52 weeks of the year. Outlying data are emphasized. Searching, filtering and analyis of different features are supported.

Details are shown in scatterplots (above), binned in 40x40 and adaptively rebinned if too empty. Table lens effect on the SPs.

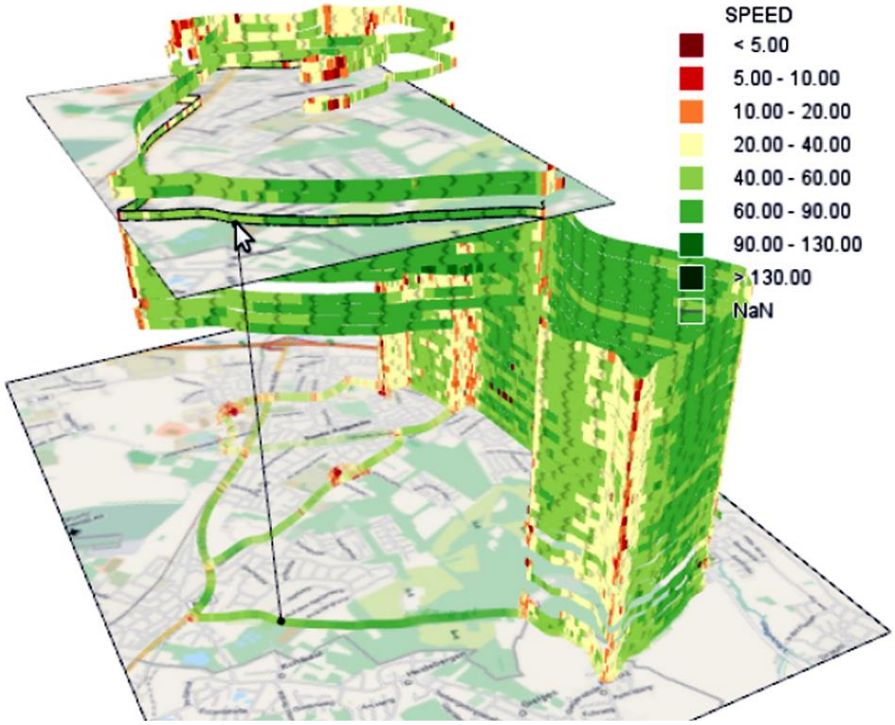
Major Application Areas

Visual Analytics of Temporal Data



RankExplorer visualization of the top 2000 Bing search queries from Nov. 20 to Dec. 29 in 2011 (Shi, 2012). Changes within and across a category are emphasized. Time data is segmented and the frequency and rank of search queries is displayed. Based on the ThemeRiver infovis. Technique.

Visual Analytics of Spatio-Temporal Data



Stacked trajectories are displayed in 3D space to reveal frequent traffic routes and jam (From: Tominski, 2012).

Use of the 3rddimension to incorporate time is a frequent visualization pattern in spatio-temporal data.

GeoVista is a wide-spread tool for visual analytics of GeoScience data (video channell).

Major Application Areas

In most complex situations, one tool is not sufficient to support the whole analytics discourse.

To replicate, e.g. all advanced statistics methods of one tool in another one, is not feasible.

Interoperability needs to be considered → aim at a seamless integration of visual analytics, statistics, presentation … tools

Further Information

Visual Analytics is primarily discussed at the yearly IEEE Vast Conference and the IEEE TVCG journal.

A few important videos:

- David Ebert: Visual Analytics for Global Challenges
- J. D. Fekete Visual Analytics - Mastering the Information Age
- Daniel Keim: Solving Problems with Visual Analytics

References

Richard Arias-Hernández, Linda T. Kaastra, Tera Marie Green, Brian D. Fisher: Pair Analytics: Capturing Reasoning Processes in Collaborative Visual Analytics. *Hawaii International International Conference on Systems Science*, 2011: 1-10

Ira Assent, Ralph Krieger, Emmanuel Müller, Thomas Seidl: VISA: visual subspace clustering analysis. *SIGKDD Explorations* 9(2): 5-12 (2007)

EH Chi. “A taxonomy of visualization techniques using the data state reference model”, *IEEE Symposium on Information Visualization*, 2000, 69-75

N Cao, D Gotz, J Sun, H Qu. Dicon: Interactive visual analysis of multidimen-sional clusters, *IEEE Trans. Vis. Comput. Graph.,* Vol. 17 (12), 2581-2590, 2011

Dang T N, Anand A, Wilkinson L. TimeSeer: Scagnostics for high-dimensional time series. *IEEE Trans. Vis. Comput. Graph.*, 2013, 19(3): 470-483.

Stephen G. Eick and Alan F. Karr. Visual Scalability, *Technical Report 106, National Institute of Statistical Sciences*, 2000

Patrick Fiaux, Maoyuan Sun, Lauren Bradel, Chris North, Naren Ramakrishnan, and Alex Endert. 2013. Bixplorer: Visual Analytics with Biclusters. *Computer* 46, 8 (August 2013), 90-94.

Jeffrey Heer, Maneesh Agrawala: Design considerations for collaborative visual analytics. *Information Visualization* 7(1): 49-62 (2008)

Daniel A. Keim, Florian Mansmann, Jörn Schneidewind, Jim Thomas, Hartmut Ziegler: „Visual Analytics: Scope and Challenges“. *Visual Data Mining 2008*: 76-90

References (II)

M Kreuseler, H Schumann (2002). „A flexible approach for visual data mining“, *IEEE Trans. Vis. Comput. Graph.*, 8 (1), 39-51

Robert L. Ling. 1973. A computer generated aid for cluster analysis. *Commun. ACM* 16, 6 (June 1973), 355-361

J Seo, B Shneiderman. “A rank-by-feature framework for interactive exploration of multidimensional data”, *Information Visualization* 4 (2), 96-113, 2005

Shi C, Cui W, Liu S, Xu P, Chen W, Qu H. RankExplorer: Visualization of ranking changes in large time series data. *IEEE Trans. Vis. Comput. Graph.*, 2012, 18(12): 2669-2678.

Charles D. Stolper, Adam Perer, David Gotz: Progressive Visual Analytics: User-Driven Visual Exploration of In-Progress Analytics. *IEEE Trans. Vis. Comput. Graph.* 20(12): 1653-1662 (2014)

G. D. Sun, Y. C. Wu, R. H. Liang, S. Liu. A Survey of Visual Analytics Techniques and Applications: State-of-the-Art Research and Future Challenges, *Journal of Computer Science and Technology*, 28(5):852-867

Tominski C, Schumann H, Andrienko G, Andrienko N. Stacking-based visualization of trajectory attribute data. *IEEE Trans.Vis. Comput. Graph.*, 2012, 18(12): 2565-2574

Tukey J.W. Exploratory data analysis. Addison-Wesley, Reading, MA, USA, 1977.