Visualization
Visual Analytics (VA)

SCHMALKALDEN UNIVERSITY OF APPLIED SCIENCES

Graphics+Interaction Laboratory

Based on Material by Marc Streit and Alexander Lex

VA Motivation

- Possibilities to collect and store data increase
- Faster than ability to use it for decision making

- Danger of getting lost in the data

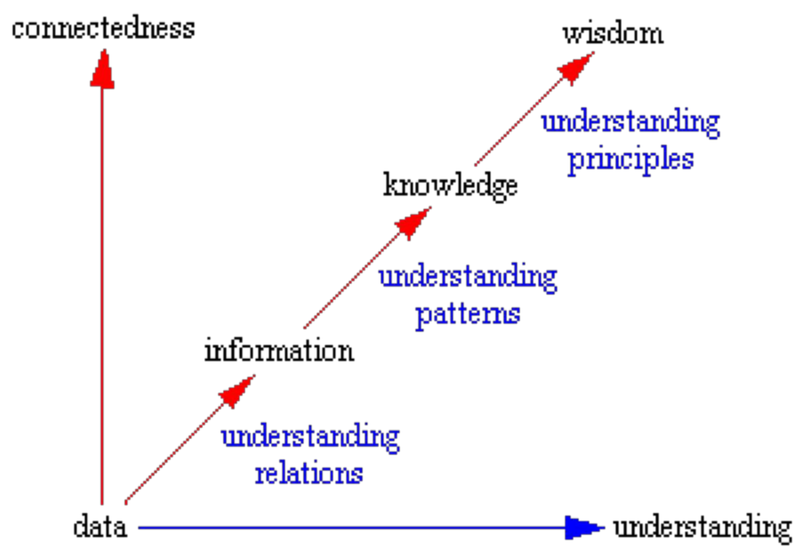
- **Data** has no value in itself
- Extract the **information** contained in it!

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Data → Information → Knowledge → Wisdom

[Bellinger 2004]

- **Data**
  - Symbols
- **Information**
  - Data that are processed to be useful; provides answers to "**who**", "**what**", "**where**", and "**when**" questions
- **Knowledge**
  - Application of data and information; answers "**how**" questions
- **Wisdom**
  - Evaluated understanding



http://www.systems-thinking.org/dikw/dikw.htm

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

History of VA

- Move from confirmatory to exploratory data analysis
  - John W. Tukey 1977 in “Exploratory Data Analysis” book
  - Confirmatory: charts and other visual representations to present data
  - Exploratory: interact with data
- Visual data exploration & visual data mining
- Visual analytics
  - 2004
  - Research and development agenda
    “Illuminating the Path”

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

VA Definition

- "Visual Analytics is the science of analytical reasoning supported by a highly interactive visual interface." [WongThomas 2004]

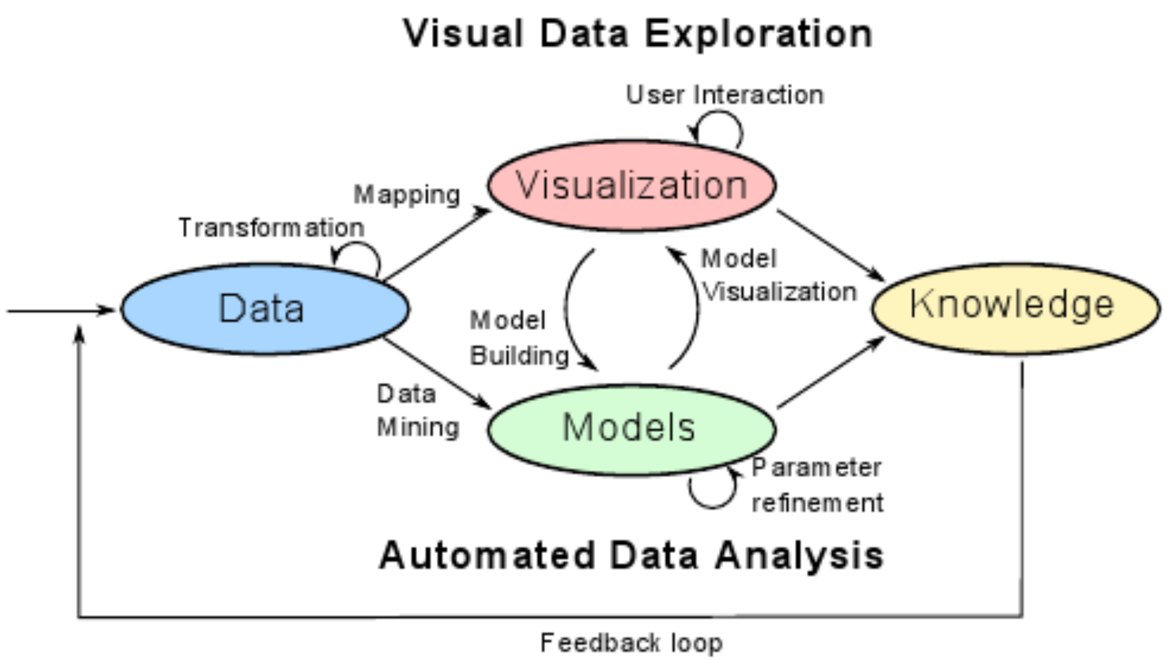
- "Visual Analytics combines **automated analysis** techniques with **interactive visualisations** for an effective **understanding, reasoning and decision making** on the basis of **very large** and **complex datasets**" [Keim 2010]

- Detect the expected and discover the undetected

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Visual Analytics Process

- First step: preprocess and transform data
  - Data cleaning, normalization, grouping, data fusion
- Alternating between visual and automatic methods



Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Application Fields

- Physics
- Astronomy
- Climate and weather
- Biology
- Medicine
- Business Intelligence

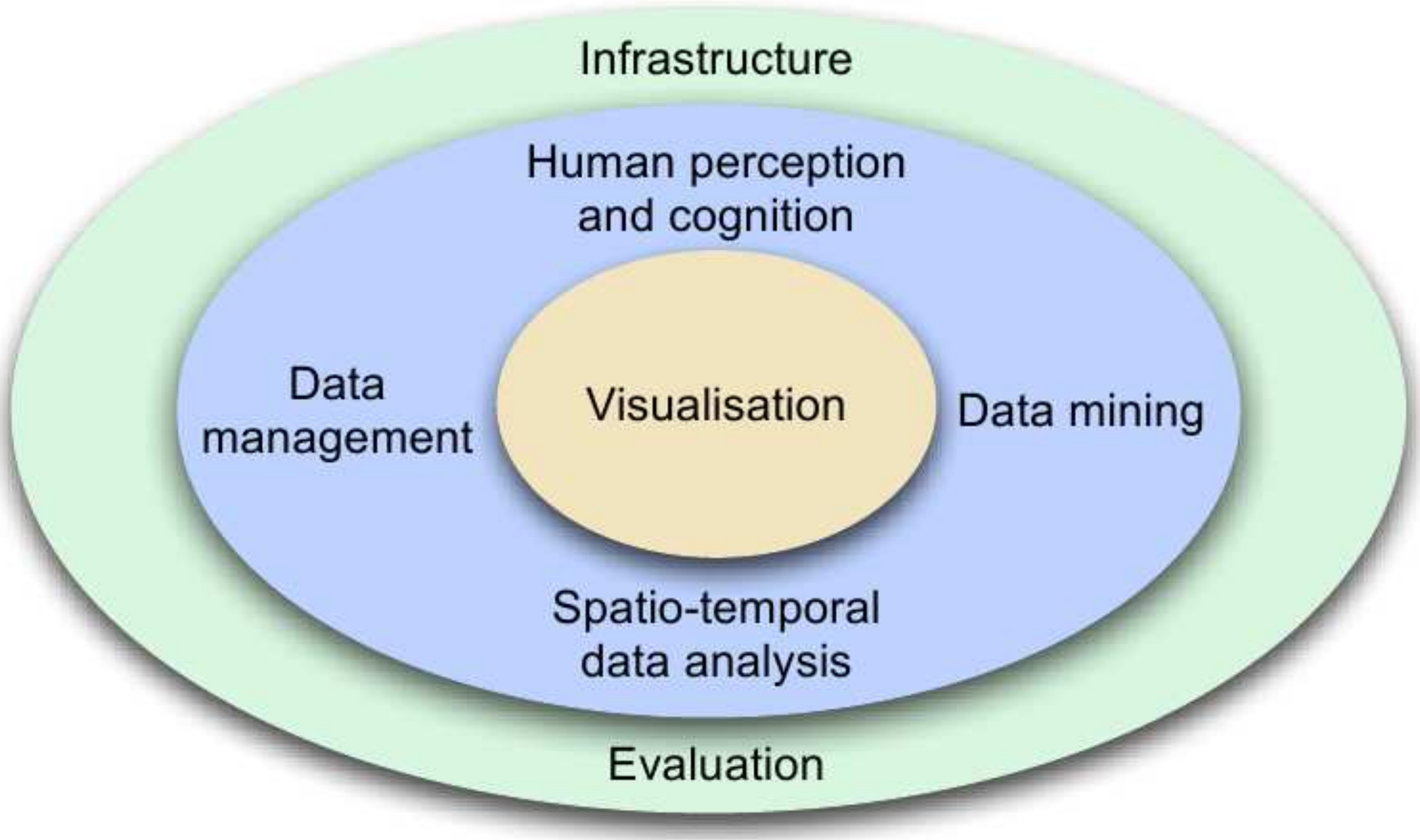
Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

VAST

- IEEE Conference on Visual Analytics Science and Technology
- Founded 2006
- Co-located with IEEE VisWeek (Vis, InfoVis)

- New: EuroVA
  - Co-located with EuroVis

Interdisciplinary!



Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

VISUALIZATION

Already covered in other lectures!

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

DATA MANAGEMENT

Heterogeneous Data

- Until last decade
  - Focus on efficiency and scalability
  - Uniform, structured data
- Numeric data, graphs, text, audio, video, etc.+
- Different formats
- Different sources
- Dealing with missing and inaccurate data values

- Users get overwhelmed
  - Data/information overload problem!

Data Types

- Numeric Data
- Text
- Graphs
- Audio
- Video signals
- etc.

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Data Management

- Data Management is a well understood field
  - Research over past 30 years
- Dynamicity problem
  - Data Management: Static two step interaction
    1. Query formulation
    2. Result collection
  - Interactive analysis
    - Response in < 100 msec necessary
- User interaction life-cycle
  - Data Management: Single user, one shot
  - Interactive analysis:
    Long-term activities and collaborative tasks

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Ways to manage data in VA

- Flat files
  - Lack of typing and metadata
  - E.g., spreadsheets, CSV
- Structured file formats
  - Adds typing
  - E.g., XML
- Traditional (relational) databases
  - Row-based
  - Robust / mature

Ways to manage data in VA (2)

- Analytical databases
  - Column-based architecture
- NoSQL systems
  - Cloud Storage
- Workflow and dataflow systems
  - Apply a previous or well-known process repeatedly

- Interactive analysis needs in-memory storage!

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Data Cleaning

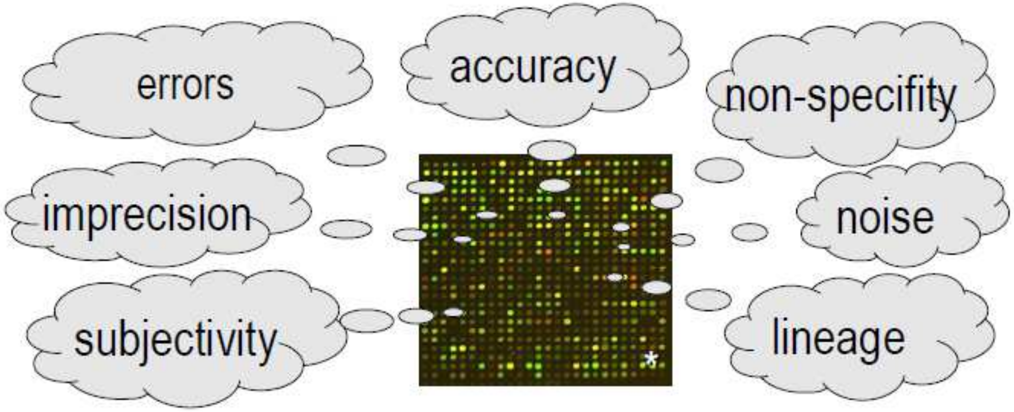
- Missing values
- Inaccurate values
- Null values

- Curative algorithms: Providing an alternative
  - Interpolation
  - Statistically computed
- Visualization

- Complex and time consuming!
  - Even for small data sets

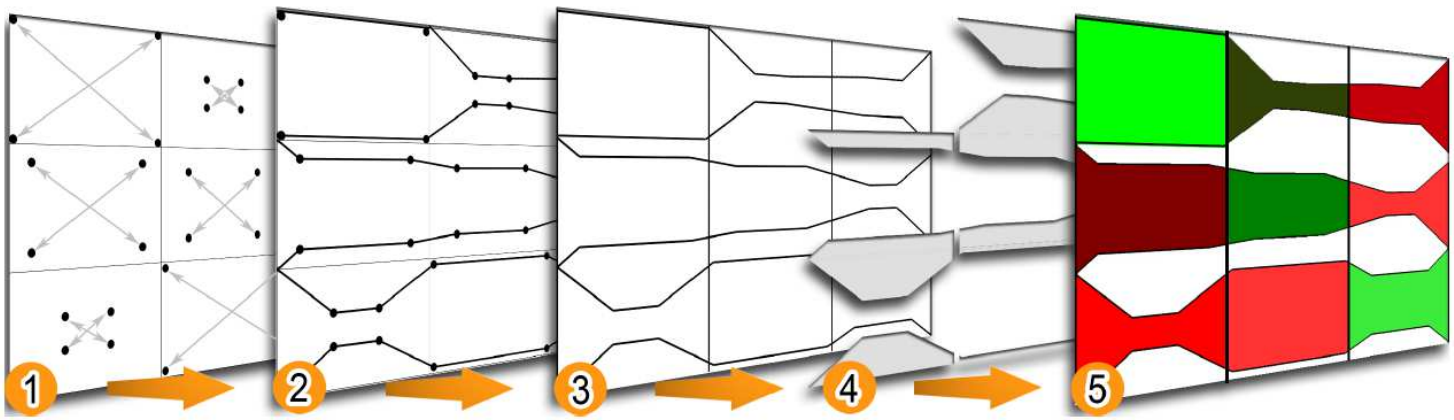
Challenge: Uncertainty

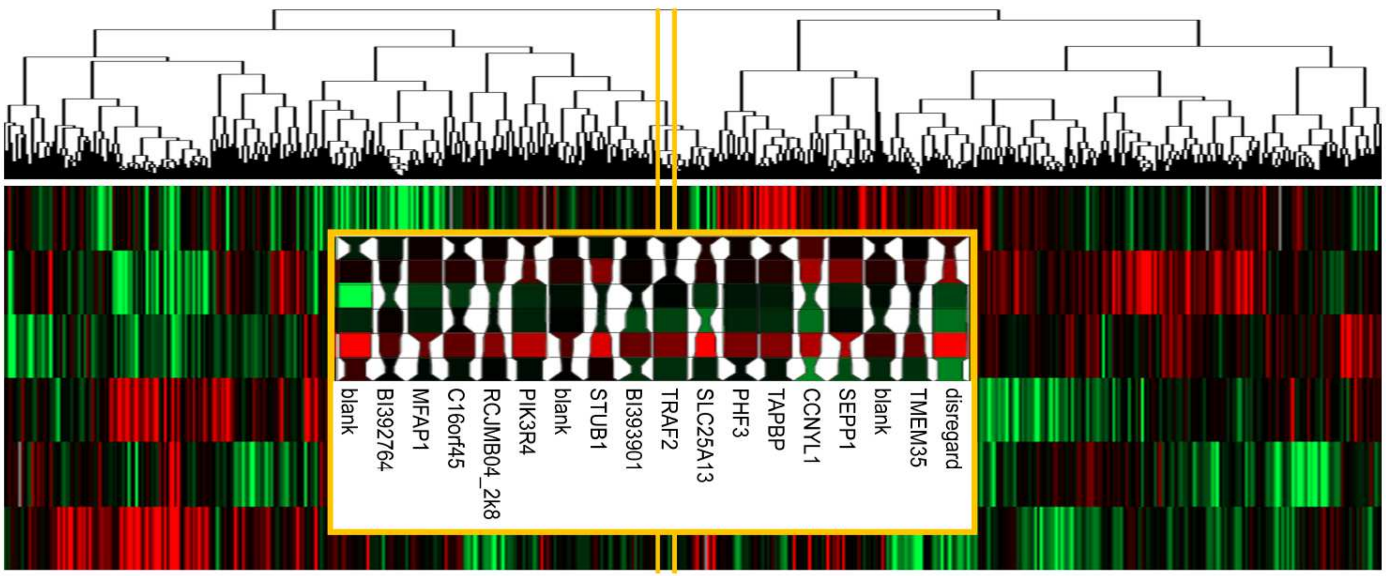
- Definition
  - “Degree to which the lack of knowledge about the amount of error is responsible for hesitancy in accepting results and observations with caution” [Hunter 1993]
- Measurement data
  - E.g., DNA microarray expression data



[Holzhüter 2010]

Uncertainty Visualization Example





[Holzhüter 2010]

Challenge: Semantics Management

- Manage not only data itself
- But also
  - Meta data
  - Abstraction levels
  - Hierarchical structures

- Needed for automatic and semi-automatic analysis

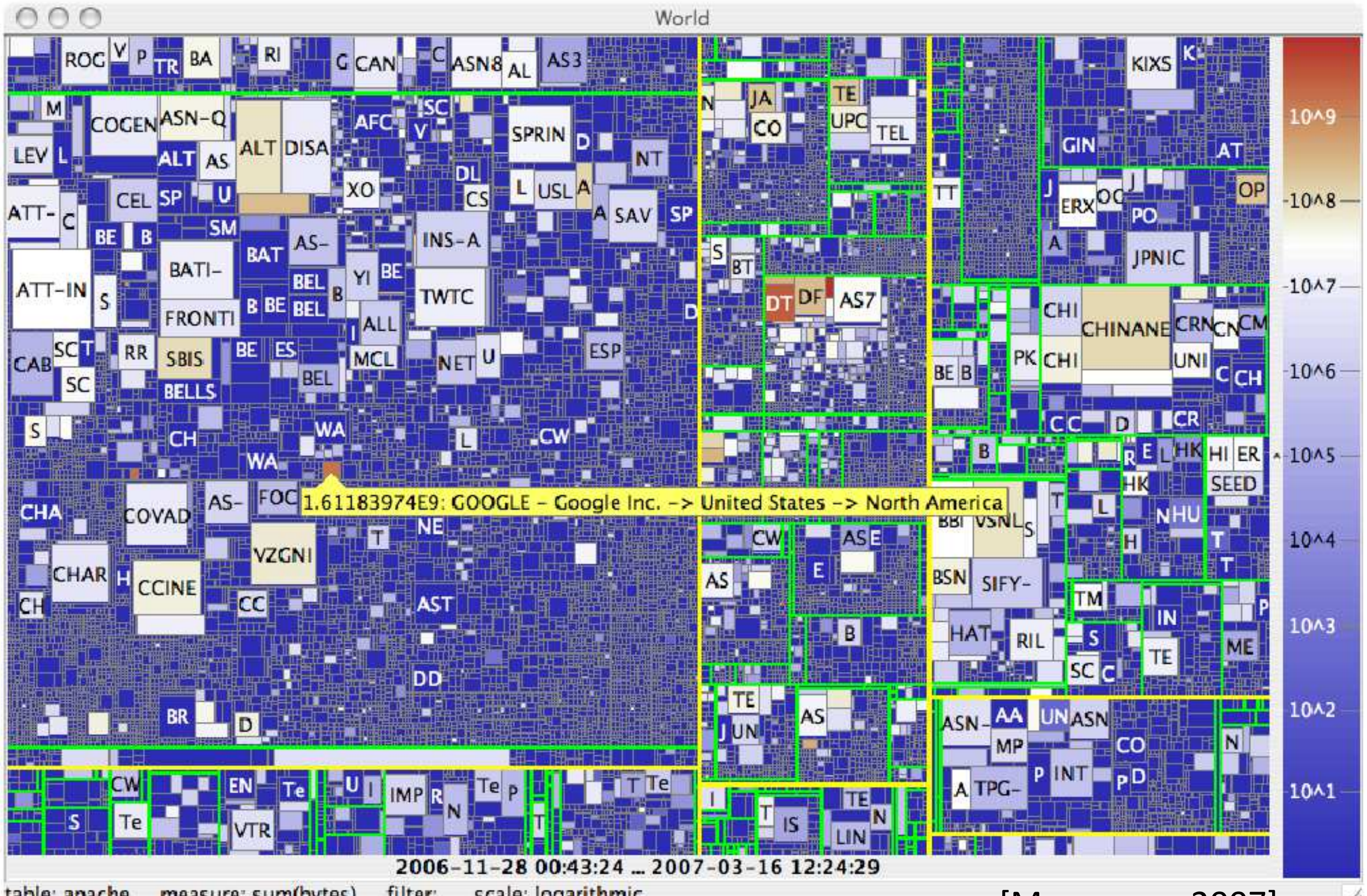
Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Challenge: Data Streaming

- Dynamic data
- Example
  - VA of social network with life feed data
- Re-calculating everything is not a solution

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Network Traffic Analysis



[Mansmann 2007]

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

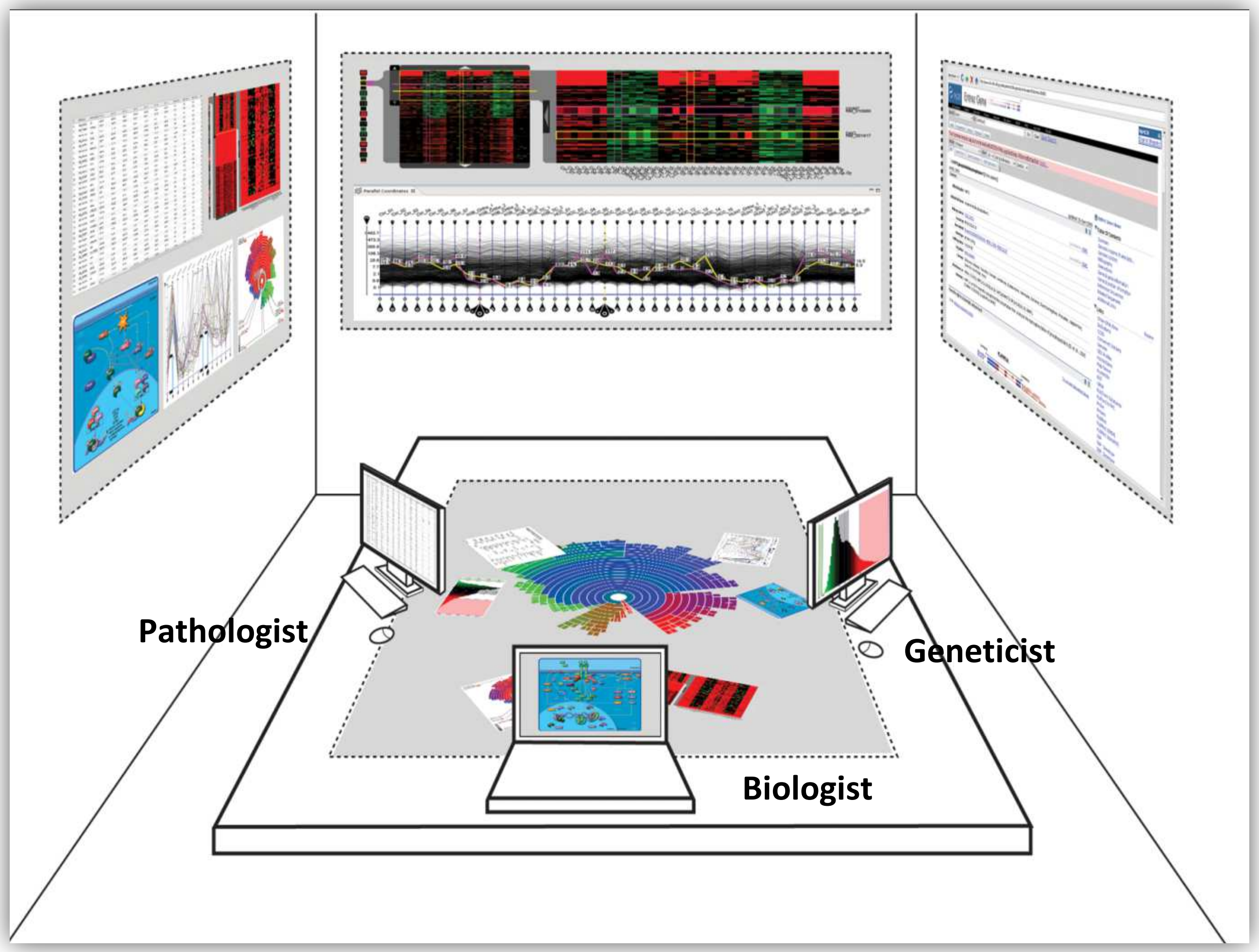
Challenge: Distributed and Collaborative VA

- Interdisciplinary analysis problems
- Single domain expert may not be enough
  → Need for collaboration
- Annotating data and insights
- Share findings with different users

- Co-located vs. distributed

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Co-Located Visual Analytics



[Streit, CoVis 2009]

Deskotheque Lab at ICG

[Waldner, CoVis 2009]

Challenge: VA for the Masses

Web-based system

+

Integrated data management

+

Interactive visualization

=

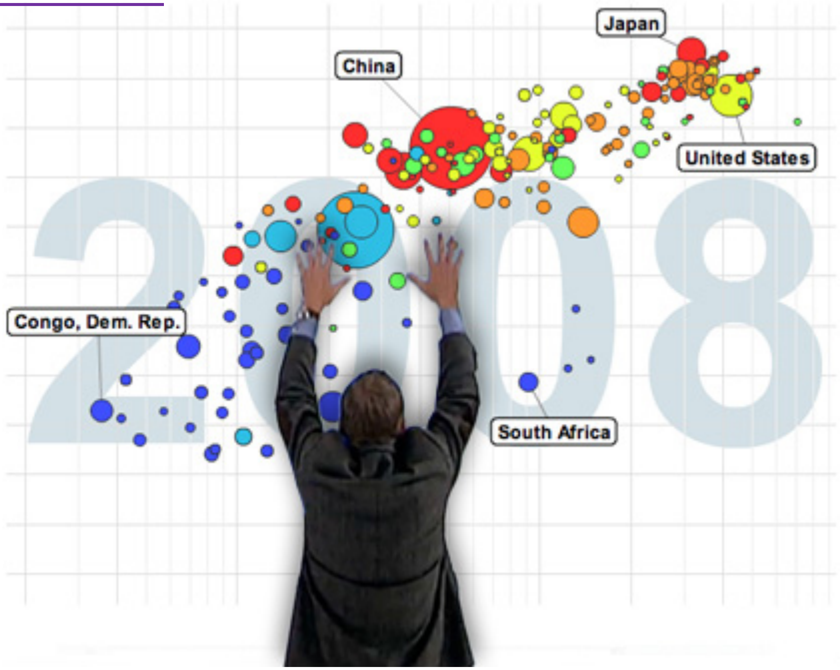
Visual Analytics for the Masses

- Home user becomes naive analyst
- Challenges
  - Raises heterogeneity (data sources and devices)
  - User Acceptance Issues
  - Scalability

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

VA for the Masses: Gapminder

- World census data
- http://graphs.gapminder.org/world



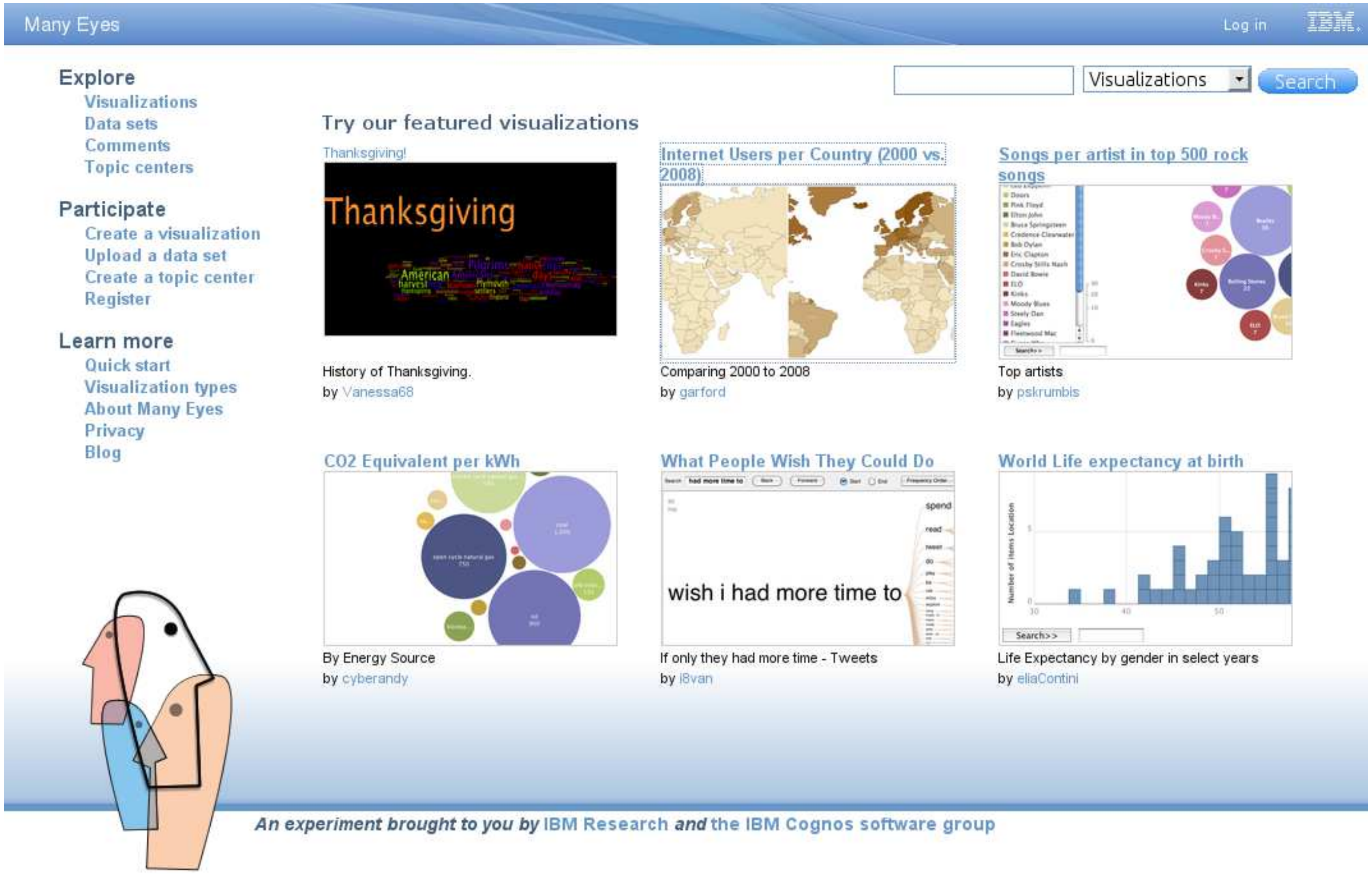
- Software: Trendalyzer
  - Acquired by Google in 2007
  - Interactive 2D-Scatterplot
  - Plus color and size
    for additional attributes
  - Linking and brushing
    - Sliders

VA for the Masses: Gapminder (2)

- Hans Rosling – TED talks
  - http://www.gapminder.org/videos/
  - http://www.youtube.com/watch?v=jbkSRLYSojo

VA for the Masses: ManyEyes

- IBM Research
- http://www-958.ibm.com/software/data/cognos/manyeyes/



DATA MINING

Data Mining Intro

- Definition
    - Automatic algorithmic extraction of valuable information from raw data
- Find interesting facts in large datasets

Statistics vs. Visualization

- Ascombe's quartett
- Statistics profile is the same for all!
  - Mean of x = 9.0
  - Mean of y = 7.5
  - Sums of squared errors = 110
  - Correlation coefficient = 0.82
  - Coefficient of determination = (
  - etc.

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

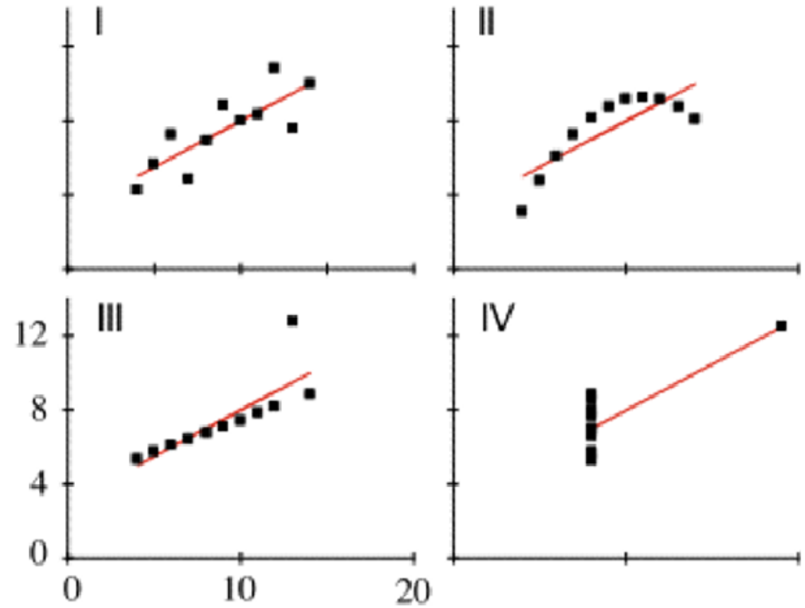
(a) Four datasets with different values and the same statistical profile

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Simple Visualization: Dot plot

I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

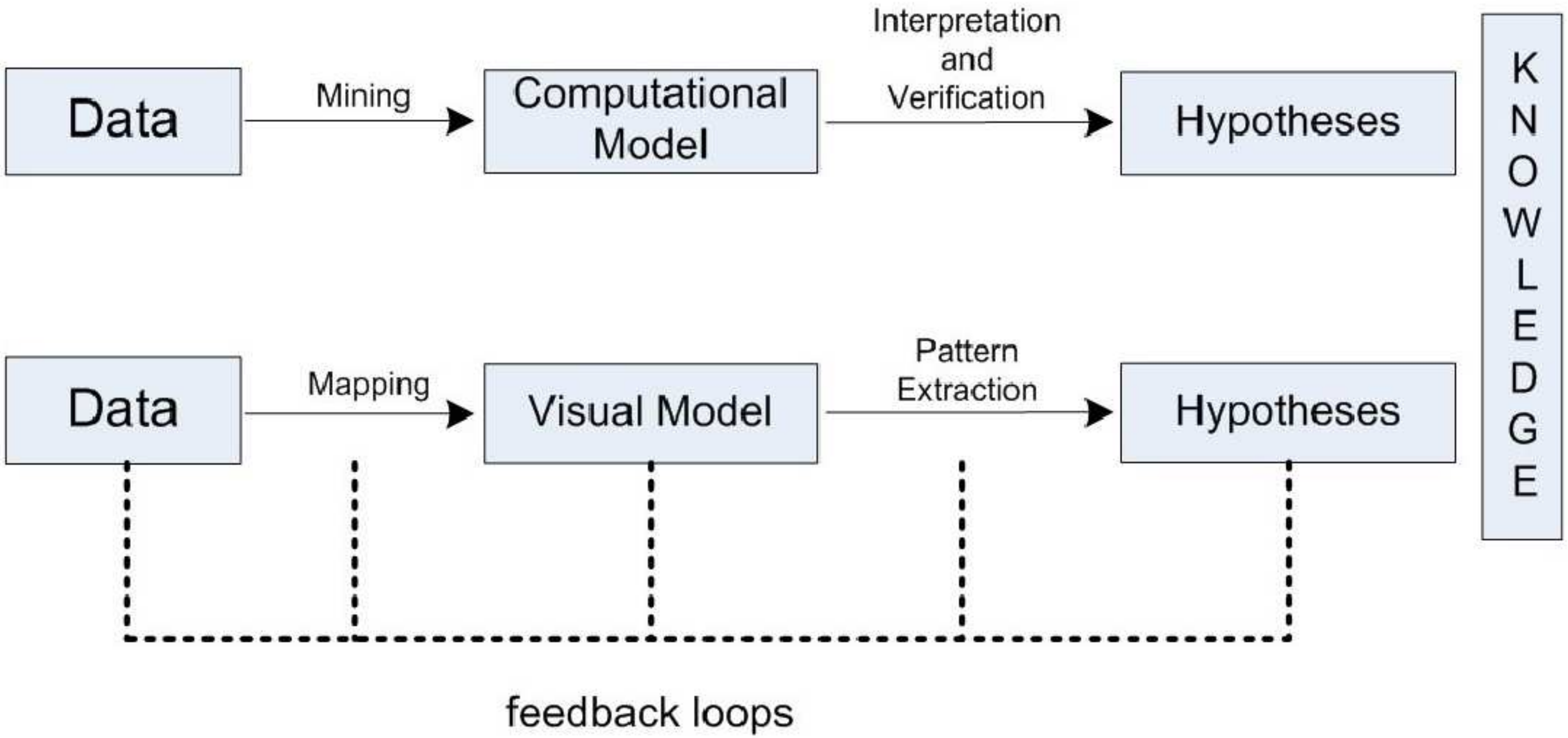
(a) Four datasets with different values and the same statistical profile



(b) Dot Plot of the four datasets

Fig. 6. Anscombe's Quartet

Traditional Data Mining vs. Visual Analysis Processes



Knowledge Discovery and Data Mining (KDD)

- Semi or fully automated analysis of massive data sets
- Contributions are more about general methodologies
- Black-box methods in the hands of end users
  - Users need to understand the algorithms for using them
  - What attributes to use? What similarity measure? etc.
  - Often trial and error

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

In Contrast: Visualization

- Incorporate
  - Experts' background knowledge
  - Creativity
  - Intuition
  - But: only relatively small data sets

- VA has to bridge these two fields!

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Supervised vs. Unsupervised Learning

- Supervised learning
  - Based on set of training samples
  - Learn models for classification of previously unseen data samples
- Unsupervised learning
  - Extract structure form data without prior knowledge
  - Example: Cluster analysis
  - Example: Dimensionality reduction

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

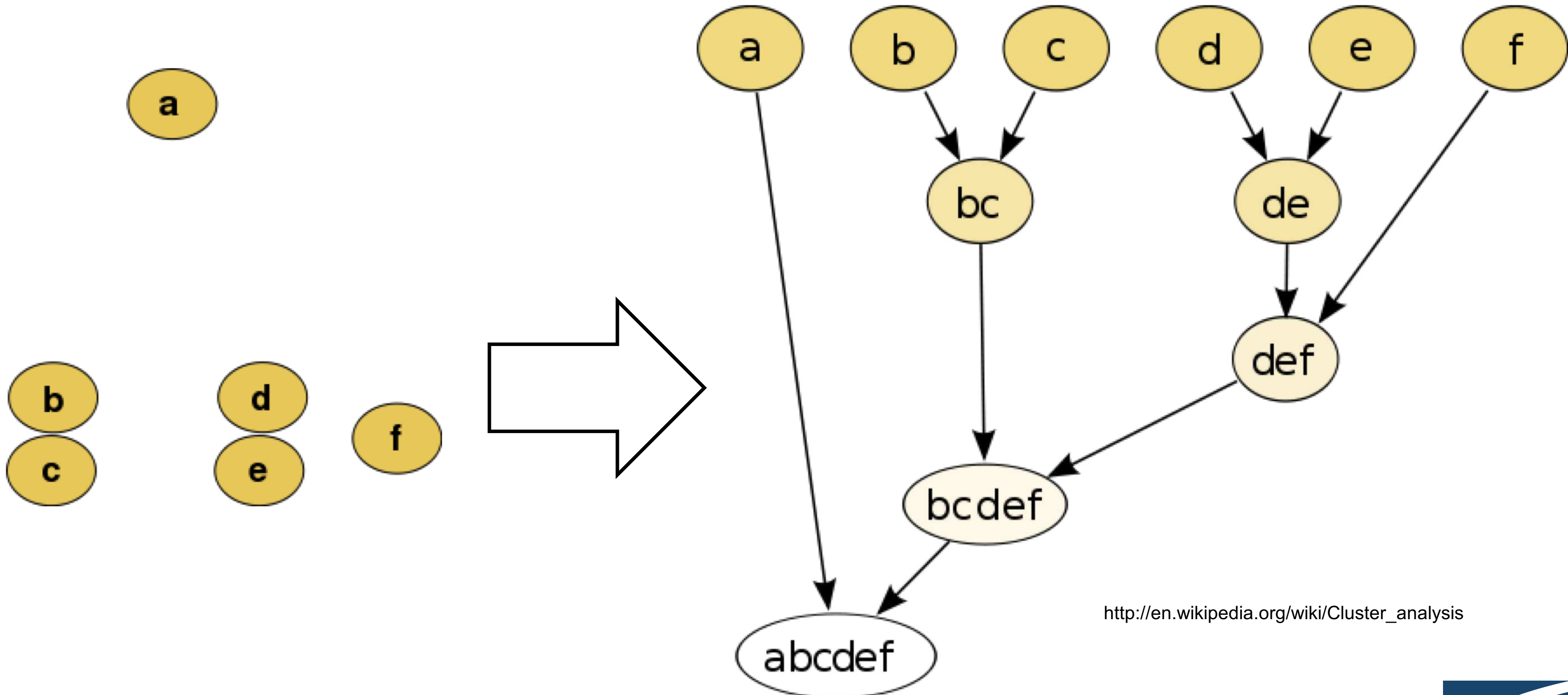
Cluster Analysis

- Automatically group data instances into classed based on mutual similarity
- Distance metric

- Hierarchical
- Partitional
  - K-Means
- Bi-clustering
  - Simultaneous clustering of rows and columns
- Fuzzy

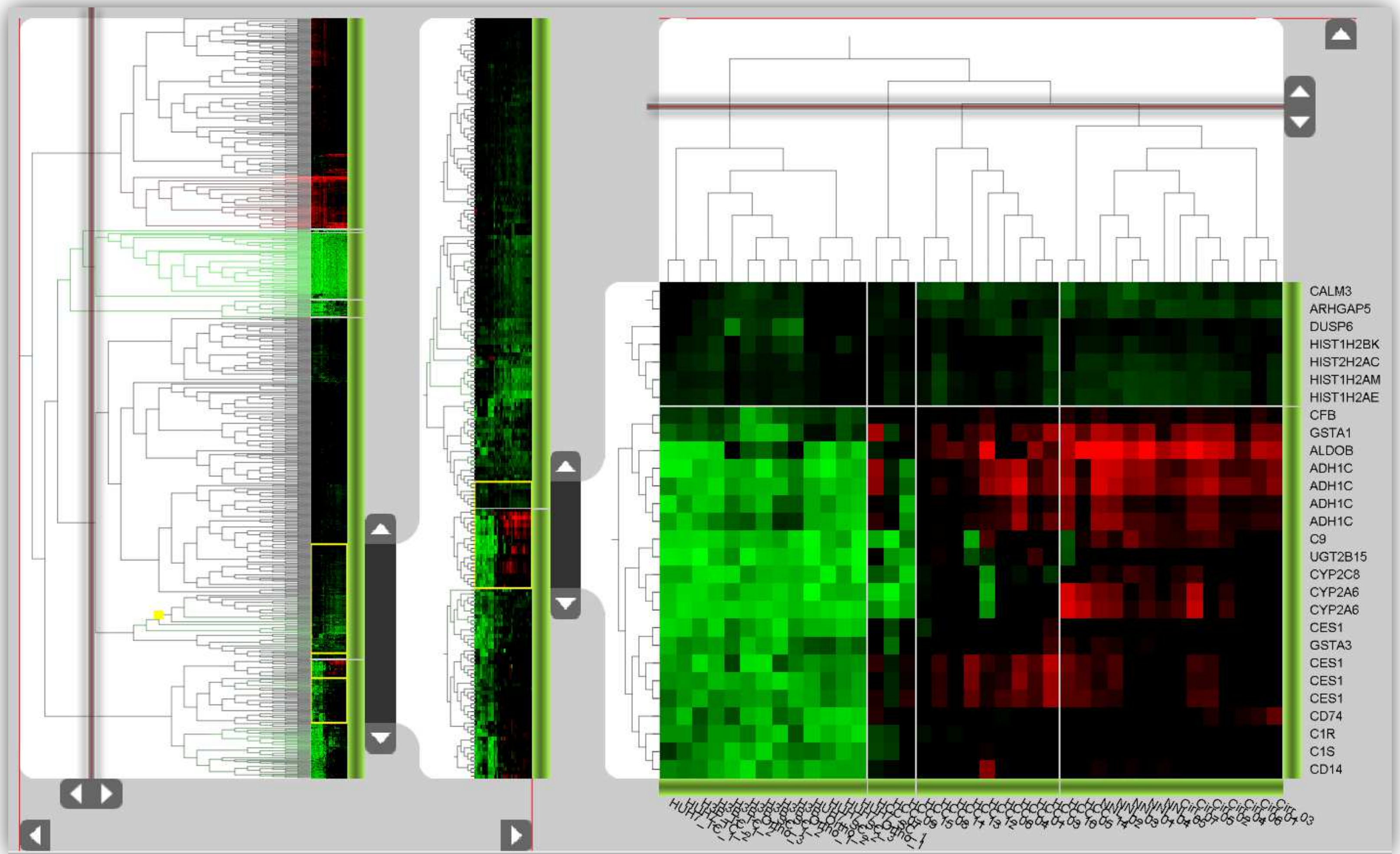
Hierarchical Clustering

- Distance metric: Euclidean distance



http://en.wikipedia.org/wiki/Cluster_analysis

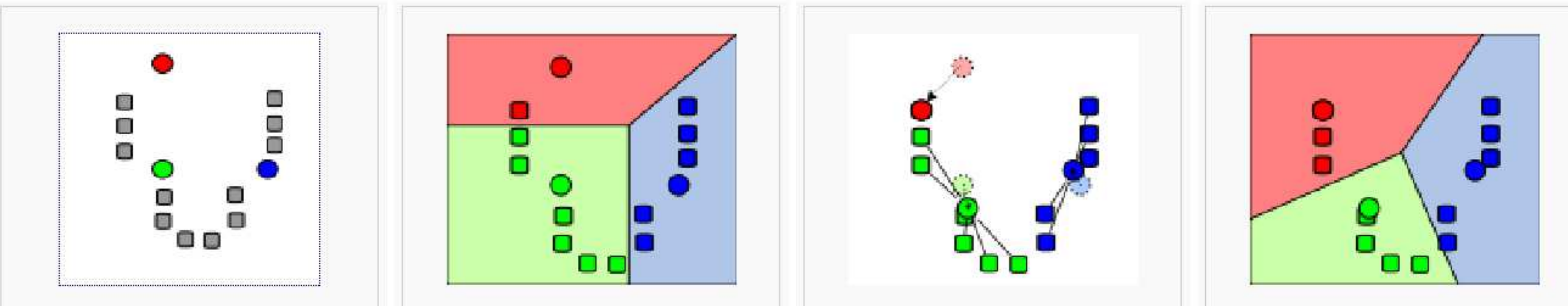
Hierarchical, Clustered Heat Map



[Lex, PacificVis 2010]

K-Means Clustering

- Partition n observations into k clusters
- Each observation belongs to the cluster with the nearest mean



1) *k* initial "means" (in this case *k*=3) are randomly selected from the data set (shown in color).

2) *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3) The centroid of each of the *k* clusters becomes the new means.

4) Steps 2 and 3 are repeated until convergence has been reached.

http://en.wikipedia.org/wiki/K-means_clustering

Dimension Reduction

- High-dimensional data
- Transform to space with fewer dimensions
- Linear and non-linear approaches
- Example
  - PCA (Principle Component Analysis)

- Disadvantages
  - Hard to preserve semantics of single dimensions
  - Hard to understand and interpret

INFRASTRUCTURE

Infrastructure

- Linking together all the processes, functions and services required by VA applications
- Current state
  - Custom-built stand-alone applications (ad-hoc systems)
  - In-memory data storage (rather than DBMS)
  - No off-the-shelf systems
  - Need to implement them with limited domain skills
  - No intercompatibility / interoperatibility
- Problematic commercial market

Data Analysis Environments

- Statistical analysis
  - R, SPSS, SAS
- Scientific computation
  - Matlab, Scilab
- Machine learning toolkits
  - WEKA
- Textual Analysis
  - GATE, UIMA, SPSS/Text, SAS Text Miner
- Video/image analysis
  - OpenCV, IRIS Explorer

PERCEPTION AND COGNITION

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Differentiation

- Perception
  - How people interpret the surroundings
- Cognition
  - Ability to understand visual information
  - Largely based on prior learning

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

EVALUATION

Prof. Hartmut Seichter, PhD | seichter@fh-sm.de | Schmalkalden University of Applied Sciences

Evaluation

- Goal
  - Compare approaches
  - Identify problems
- Assess
  - User acceptance
  - Effectiveness
    - doing "right" things, i.e. setting right targets to achieve an overall goal (the effect)
  - Efficiency
    - doing things in the most economical way (good input to output ratio)

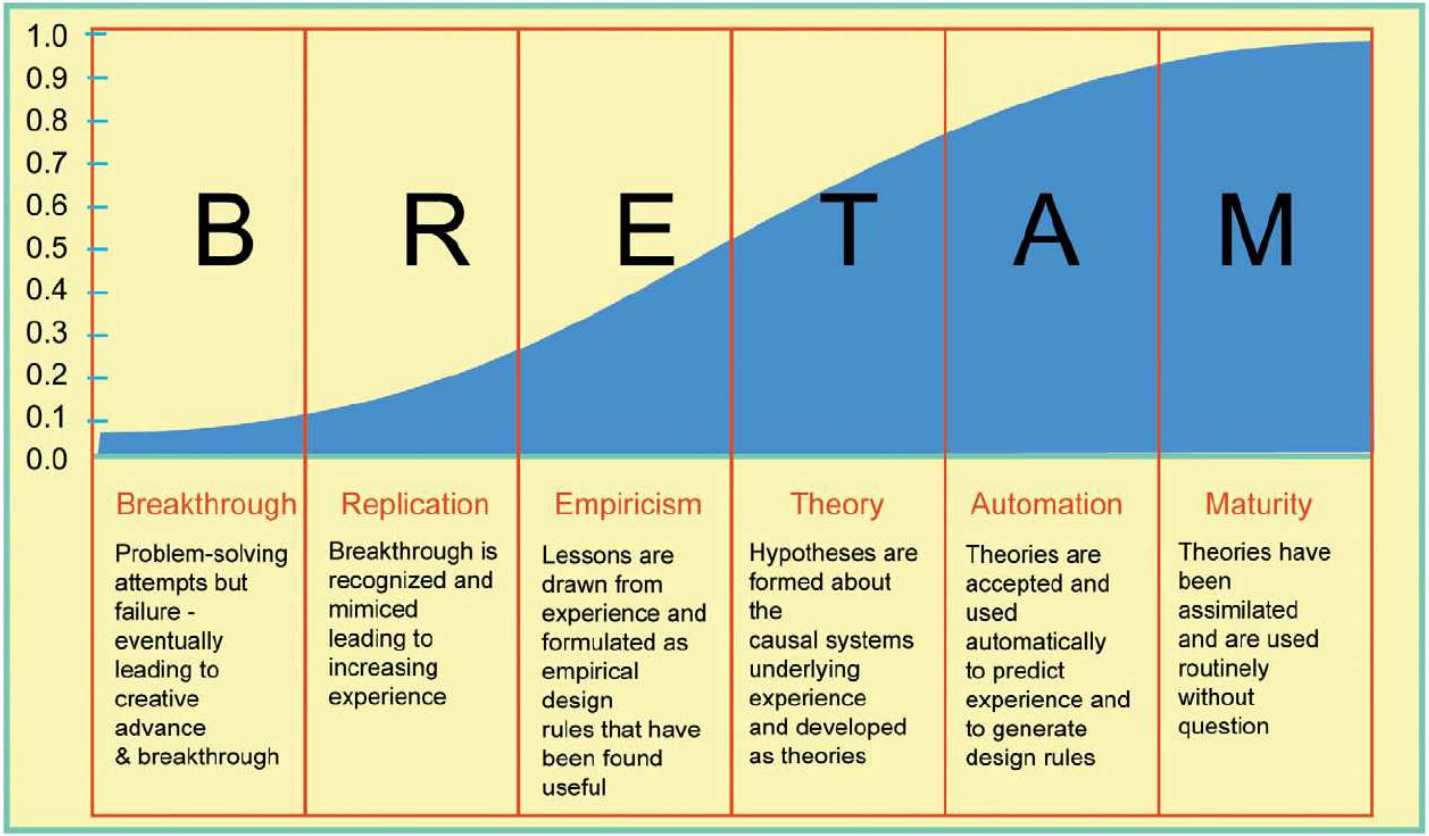
Evaluation (2)

- Quantitative vs. Qualitative methods

- Recently evaluation becomes more prominent
- Challenge
  - How to evaluate interactive, explorative visual data analysis?

VA Conclusion

- Every research field runs through same stages
- BRETAM Model -- VA is only at replication stage!



[Gaines 1991]

The End

QUESTIONS?