Visual Analytics Research

Dr. Dimitrios Tzovaras
Researcher A'
CERTH/ITI

Overview

- 1. Definition – Historical Background
- 2. Taxonomy
- 3. Applications
- 4. Visualization of Search Results
- 5. Visualization of Network Traffic/Security
- 6. VisAn focal points
- 7. Visual Analytics work at ITI
- 8. Future Challenges

Definition

- "Visual Analytics is the science of Analytical Reasoning supported by interactive visual interfaces" *Illuminating the Path – 2005*

- The best of both worlds

- Aim: turn the information overload into an opportunity

Iterative process:
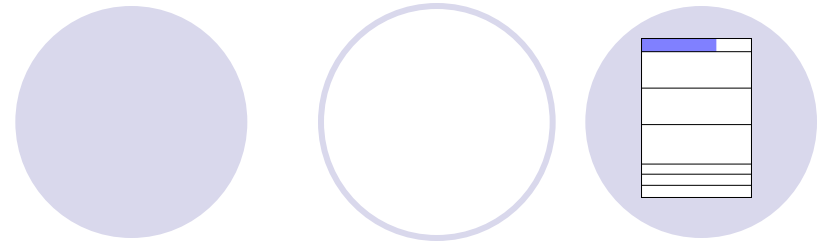
- Information gathering
- Data pre-processing
- Knowledge representation
- Interaction and decision making
- =>user insight / solution

Visual Data-Exploration
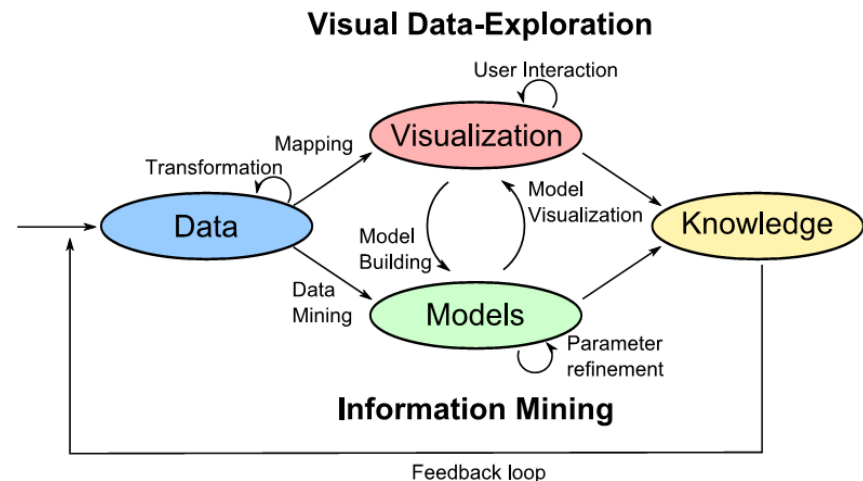
Historical Background

- 1962 – *"The purpose of computing is insight not numbers",* Richard Hamming
- 1962 – One of the first graphics algorithms (Bresenham)
- 1967 – Founding of ACM SIGGRAPH
- 1987 – NSF Panel Report on *"Visualization in Scientific Computing"*
- 1990 – Publication of first book on visualization *"Visualization in Scientific Computing"*
- 2004 – Formation of National Visualization and Analytics Center
- 2005 – *"Illuminating the Path – the R&D Agenda for Visual Analytics"*
- 2006 – NIF/NSF Report on *"Visualization Research Challenges".*
- 2006 – First IEEE Visual Analytics Symposium
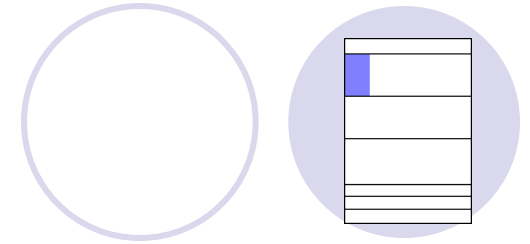
Taxonomy

- As the domain matures, order is needed
- Not a straightforward task
- The last 15 years various classification approaches according to

data-type (text, 2D, 3D, trees, networks etc)

visualization technique (geometric, pixel-oriented, hierarchical etc)

interaction technique (filter, selection, relate etc)

Taxonomy – VAC Views

- December 2008 workshop discussion

- Aim: to give a definition of the study of visual analytics in order to more clearly describe the scope and intent of impact for this research area.

Dimensions of Taxonomy for Information and Visual Analytics

- Domain / Applications

- Analytic Method / Goals

- Science / Technology

- Data Types / Structures

Any application can be defined by mapping across these four dimensions

Taxonomy – VAC Views

Domain/Applications
- Health (human, animal)
- Energy
- Environment
- Science
- Security (military, intelligence, law enforcement)
- Commerce
- Transportation
- Food/Agriculture
- Finance
- Information Assurance
- Web/Cyber
- Individual Info Management
- Organization Info Management

Analytic Methods/Goals
- Confirmatory
- Exploratory
- Predictive
- Science
- Surveillance
- Watch /Warn/ Alert
- Relationship Mapping
- Structure
- Rare Event Identification
- Affect
- Fraud
- Deception
- Alternative Competing Hypothesis

Science /Technology
- Analytic Reasoning and Human Processes
- Interactive Visualization
- Data Representations and Theory of Knowledge
- Theory of Communication
- Systems and Evaluations

Data Types / Structures
- Text (structured/unstructured)
- Complex Document
- Image
- Video
- Audio (voice, sound)
- Transaction (Cyber, Finance, Shipping)
- Numeric
- Graph Structures
- Databases/Files
- Digital Libraries
- Models/Simulations
- Geospatial Coordinates
- Time
- Name/Entity

Taxonomy – VAC Views

Analytic Reasoning & Human Processes

- Modes of Inference
- Knowledge Creation
- Data-Information-Knowledge
- Confidence Development
- Modeling
- Critical Thinking
- Hypothesis Refinement
- Human Processes (Cognition, Perception, Collaboration, Decision Making)

Data Representations & Theory of Knowledge

- Data Sourcing (Search/Gather, Flow Management)
- Item Type
- Representation (Statistical, Semantic)
- Logic Representations
- Knowledge Representations and Abstractions
- Scale and Complexity
- Precision within Context
- Aggregation
- Synthesis
- Ontology
- Transformations
- Model / Predictions Representations

Systems and Evaluations

- Application Programming Interface
- Lightweight Standards
- Service-oriented Architecture
- Evaluation: Test Data Spaces/Flows
- Privacy
- Confidentiality
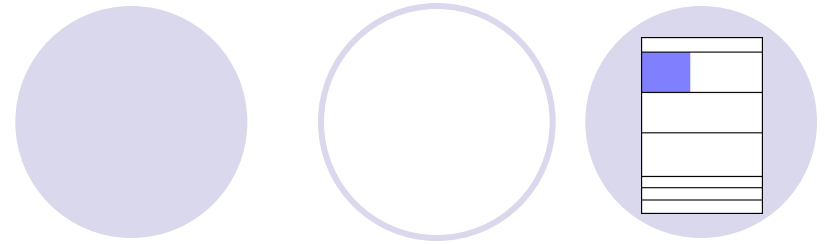
Interactive Visualization

- Science of Visualization

(Icons, Labels, Positioning, Motion, Color/ Presentation, Abstraction, Visual Representations, Connections/ Positioning, Geospatial, Temporal, Graphic, Graphic, Proximity, Linear/Horizontal, Spherical, Design and Art

- Science of Interaction

(Language of Discourse, Dynamics, Concept Representations, Change Representations, Logic Argument, Interactions / Segments, Pace and Tempo, Design and Art, Logic and Reasoning Refinement, Uncertainty Refinement, User-Tailored Interaction, Application/ Situation Interaction, Systems/ Architecture

Theory of Communication

- Story Creation
- Story Segment Identification
- Story Synthesis
- Story Communication
- Theme Identification
- Theme Flow/ Dynamics
- Temporal/Pace Dynamics
- Design and Art
- Active and Dynamic Communications
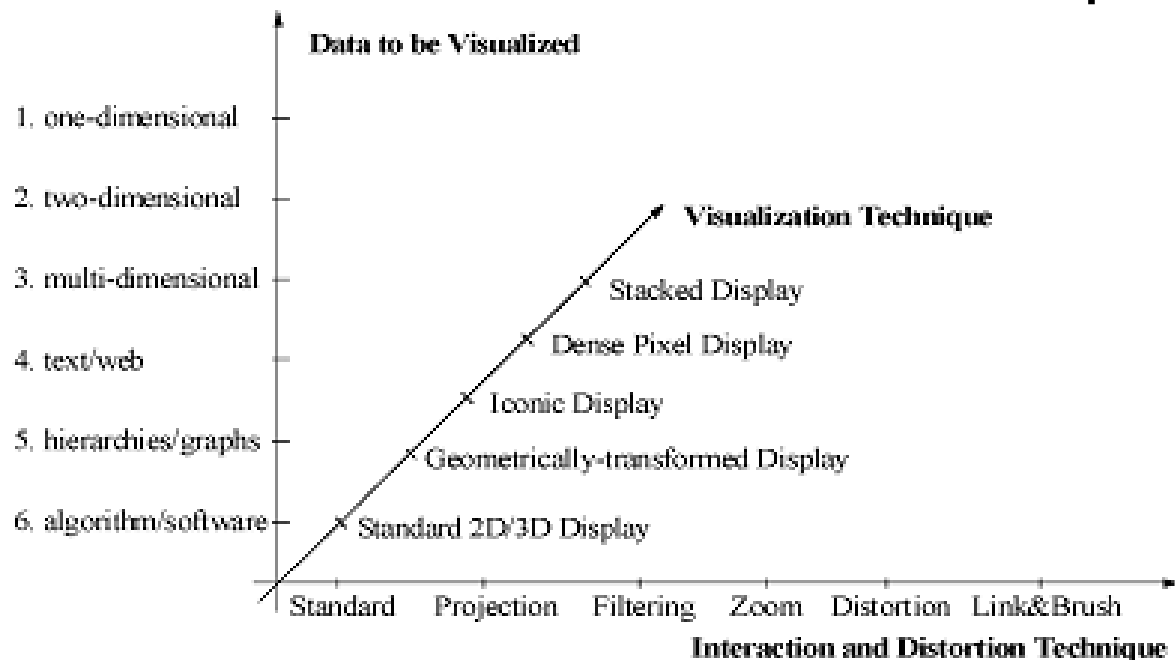- Summarization
- Logic/Reasoning Representations
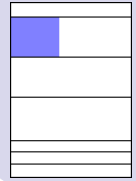
# Taxonomy - Keim

3 dimensional, according to:

- Data type

- Visualization technique

- Interaction and distortion technique

According to Data Type

- One-dimensional data, such as temporal data
- Two-dimensional data, such as geographical maps
- Multidimensional data, such as relational tables
- Text and hypertext, such as news articles and Web documents
- Hierarchies and graphs, such as telephone calls
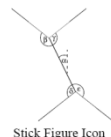- Algorithms and software, such as debugging operations

According to Visualization Technique
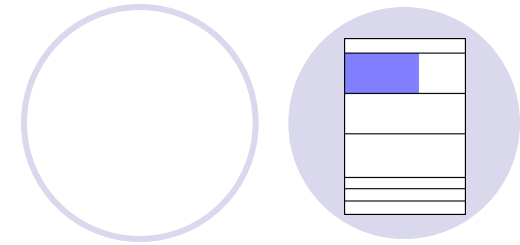
- Standard 2D/3D displays, such as bar charts and x-y plots

- Geometrically transformed displays, such as landscapes and parallel coordinates
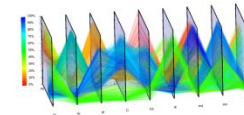
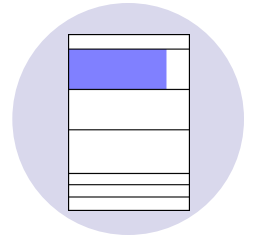- Icon-based displays, such as stick figures and star icons

- Dense pixel displays, such as the recursive pattern and circle segments techniques
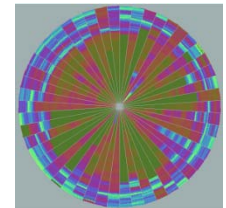
- Stacked displays, such as treemaps or dimensional stacking
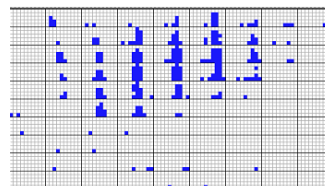
According to interaction and distortion technique

- **Interactive Projection** (dynamically change the projections in order to explore a multidimensional data set)

- **Interactive Filtering** (focus on interesting subsets)

- **Interactive Zooming**

- **Interactive Distortion** (hyperbolic, spherical)

- **Interactive Linking and Brushing** (combine different visualization methods to overcome the shortcomings of single techniques)

Taxonomy –
Periodic Table of Visualization Methods
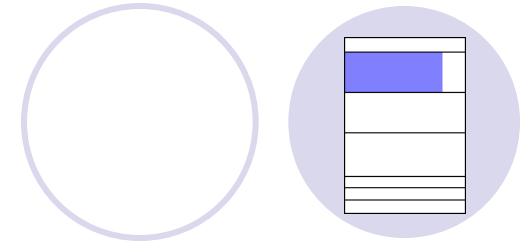
- Around 100 methods at a glance
  - collected from books, scientific articles &websites
- Main categorization principles:
  1. Complexity of Visualization
  2. Main Application or Content Area
  3. Type of Representation
  4. Point of View
  5. Type of Thinking Aid

Taxonomy –
Periodic Table of Visualization Methods

Data Visualization — Visual representations of quantitative data in schematic form (either with or without axes)

Information Visualization — 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

Concept Visualization — Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.

Strategy Visualization — The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.

Metaphor Visualization — Visual Metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed

Compound Visualization — The complementary use of different graphic representation formats in one single schema or frame

Legend:
- **Cy** Process Visualization
- **Hy** Structure Visualization
- ☼ Overview
- ◎ Detail AND Overview
- ¤ Detail
- < > Divergent thinking
- > < Convergent thinking

Elements (selected): C continuum, G graphic facilitation, Tb table, Ca cartesian coordinates, Cs concept skeleton, Mm metro map, Tm temple, St story template, Tr tree, Ct cartoon, Pi pie chart, L line chart, Me meeting trace, Fp flight plan, Cf concept fan, Br bridge, Fu funnel, Ri rich picture, B bar chart, Hi histogram, T timeline, Pa parallel coordinates, Hy hyperbolic tree, Cy cycle diagram, Sa sankey diagram, Ve venn/euler diagram, Mi mindmap, Sq square of oppositions, Co concentric circles, Ar argument slide, Co communication diagram, Gc gantt chart, Pe perspectives diagram, D dilemma diagram, Pr parameter ruler, Kn knowledge map, Ar area chart, Sc scatterplot, R radar chart cobweb, Ch chernoff faces, E entity relationship diagram, Fb feedback cycle diagram, Pa pareto chart, Cl clustering, L layer chart, Py minto pyramid technique, Ca cause-effect chains, Tl toulmin map, Dt decision tree, Cp cpm critical path method, Ev evocative knowledge maps, Co concept map, Ic iceberg, Cm cognitive mapping, Tk tukey box plot, Sp spectogram, Te tensor diagram, Tr treemaps, N naasi shneiderman diagram, Se semantic network, Fl flow chart, Sy system dyn./loop diagrams, So soft system modeling, Sm synergy map, Fo force field diagram, Ib ibis argumentation map, Pr process event chains, Pe pert chart, Sw swim lane diagram, V tree diagram, Hh heaven 'n' hell chart, I infomural, Sd supply demand chain, Pr performance charting, St strategy map, Oc organisation chart, Ho house of quality, Fd feedback diagram, Ft failure tree, Mq magic quadrant, Sr stakeholder rating map, Po porter's five forces, S s-cycle, Sm stakeholder map, Ld life-cycle diagram, Tc technology roadmap, Ed edgeworth box, Pf portfolio diagram, Sg strategic game board, Mz mintzberg's organigraph, Z zwicky's morphological box, Ad affinity diagram, De decision discovery diagram, Bm bcg matrix, Stc strategy canvas, Vc value chain, Hy hype-cycle, Is ishikawa diagram, Ta taps, Sd spray diagram

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

Applications of Visual Analytics
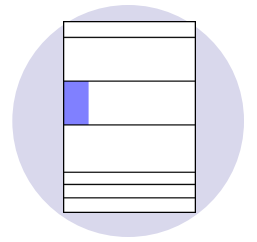
- **Physics and Astronomy**
- **Business**
- **Environmental monitoring**
- **Disaster and Emergency Management**
- **Security**
- **Software analytics**
- **Biology, Medicine and Health**
- **Engineering Analytics**
- **Personal Information Management**
- **Mobile Graphics and Traffic**

Physics and Astronomy
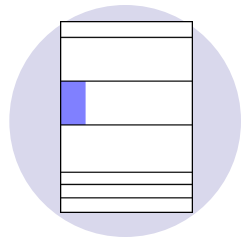
- Flow visualization, fluid dynamics, molecular dynamics, nuclear science
- Astrophysics: separate relevant data from garbage
- Sloan Digital Sky Survey
- COMPLETE project
- CERN's LHC
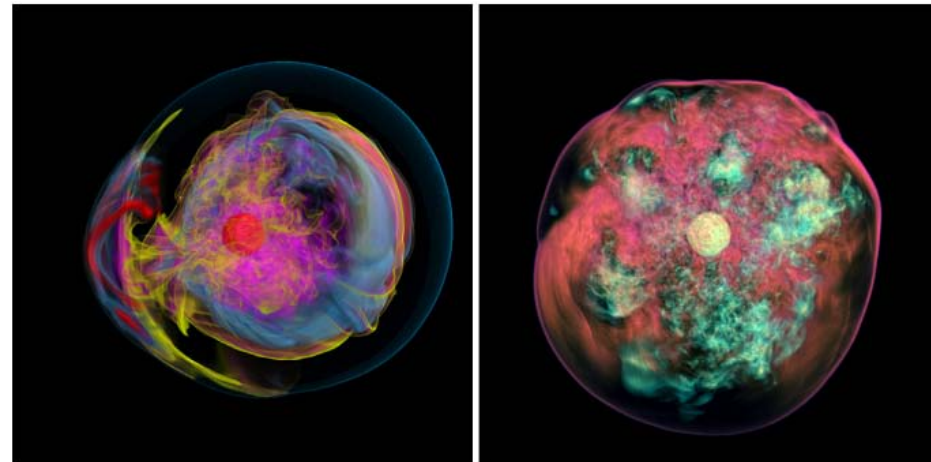
Business

- Stock market around the world, millions of transactions between stocks, bonds, futures, commodities, market indices, currencies etc
- Understanding historical and current situations
- Predicting future market trends
- Need for real time monitoring of the market, which would support the decision making of the users
- Map of the Market, FinDEx
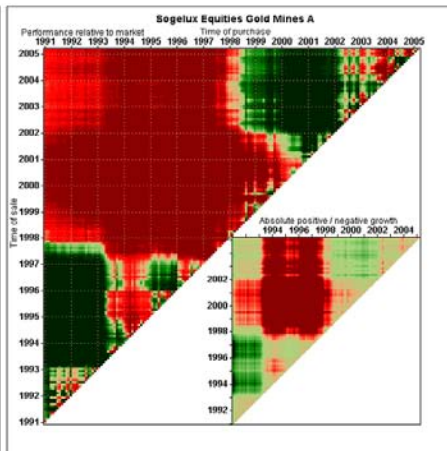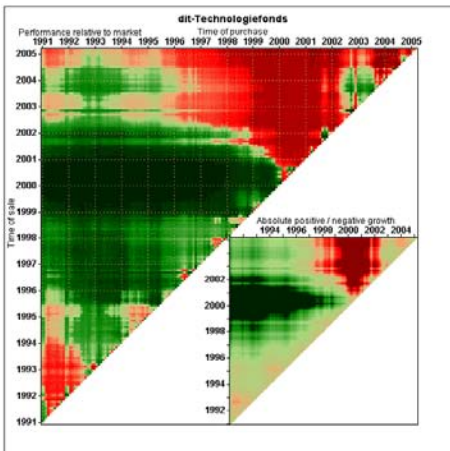
Environmental monitoring
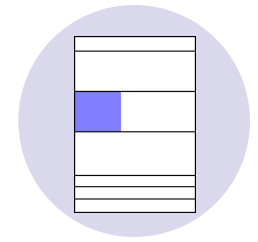
- Huge amounts of data has been gathered and continues to be   gathered for:
  - measuring the climate change
  - forecasting the weather
  - evaluating the effects of carbon emission in the atmosphere
  - melting of the poles
  - the stratospheric ozone depletion

  - Iceberg movement
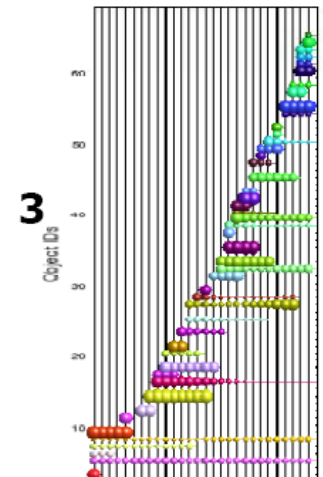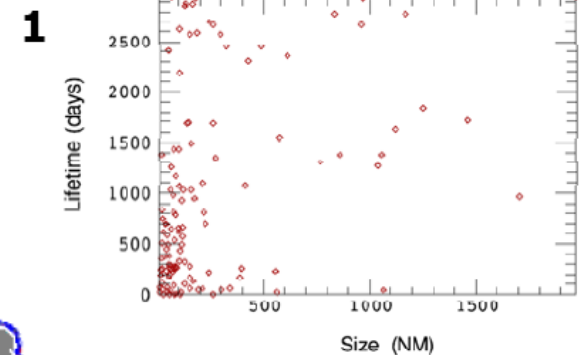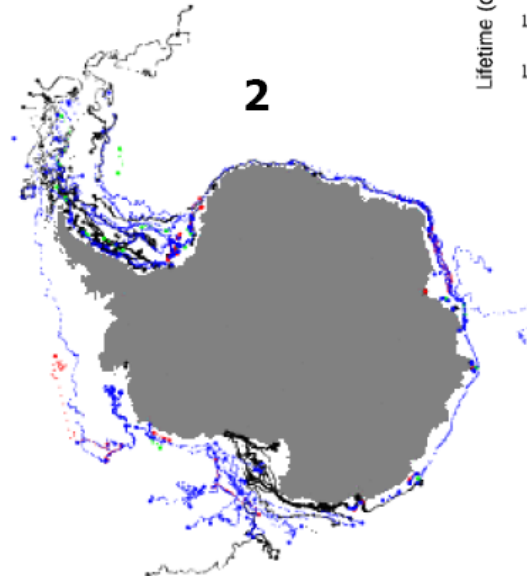
Disaster & Emergency Management
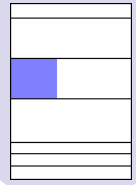
- When emergency situations or disasters occur, the response time is usually crucial to limit the effects of the damage
- Natural catastrophes like floods, earthquakes, tsunamis, storms, fires, epidemic growth of diseases
- Human caused catastrophes like industrial accidents, transport accidents, pollution etc
- Visual Analytics can help to:
  - Evaluate the situation
  - Monitor the on-going progress of the emergency
  - Provide the people in charge with clues of the kind of immediate action needed
  - VisAn can also help to prevent such emergencies
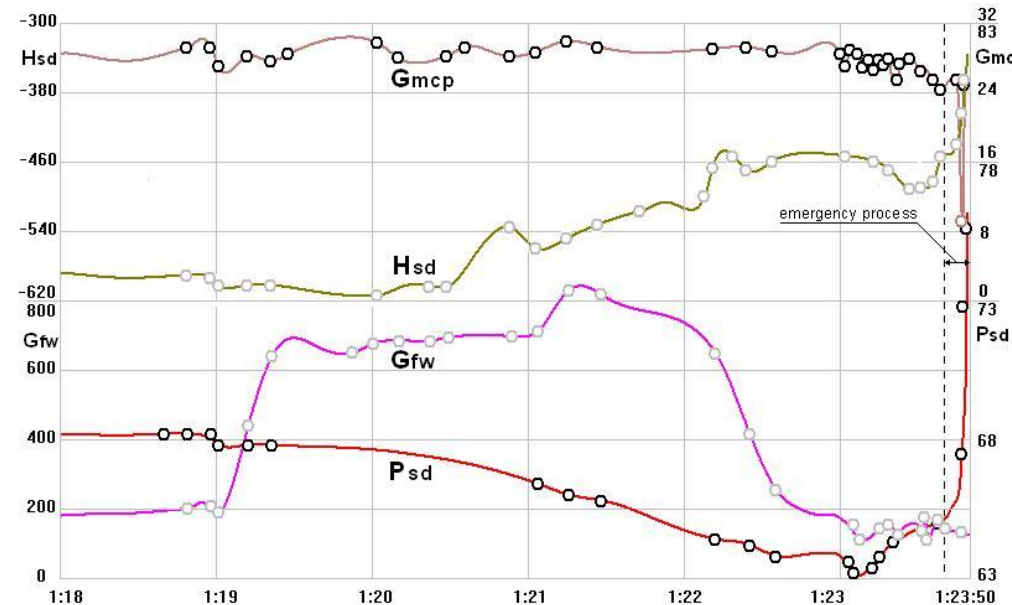
Pic 5. Heathidraulic parameters for last 5 minutes (left side)

G_{mcp} – summary flow rate of coolant (1000 m^3/hour) G_{fw} – flow rate of feed water (ton/hour)
H_{sd} – water level in separator drum (mm) P_{sd} – pressure in separator drum (kG/sm^2)

Security

- Development of applications in the security domain was the main motivation behind the writing of the "Illuminating the Path" agenda

- Wide application field, ranging from terrorism informatics over border protection to Network security

- The focal point in these fields is to bring together bits of information from various sources and relate them, in order to identify potential threats.

- VisAware, JigSaw
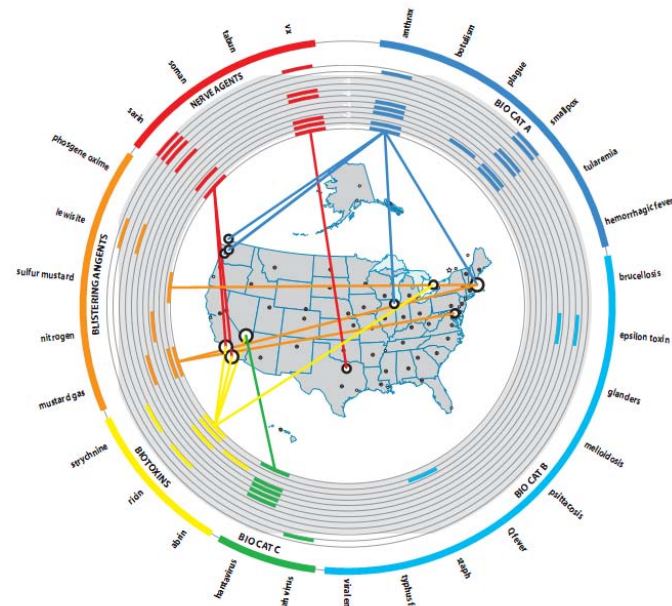
Biology, Medicine and Health
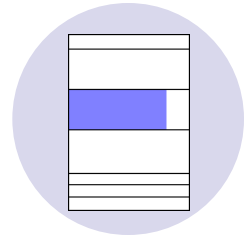
- Human Genome Project, which stores 3 billion base pairs per human
- Proteomics
- Combinatorial Chemistry
- Tools:BLAST and blast2 which are used to search for homologous sequences in nucleotide and protein databases

Distribution of 41 Blast Hits on the Query Sequence

Mouse-over to show defline and scores. Click to show alignments

Color Key for Alignment Scores

| <40 | 40-50 | 50-80 | 80-200 | >=200 |
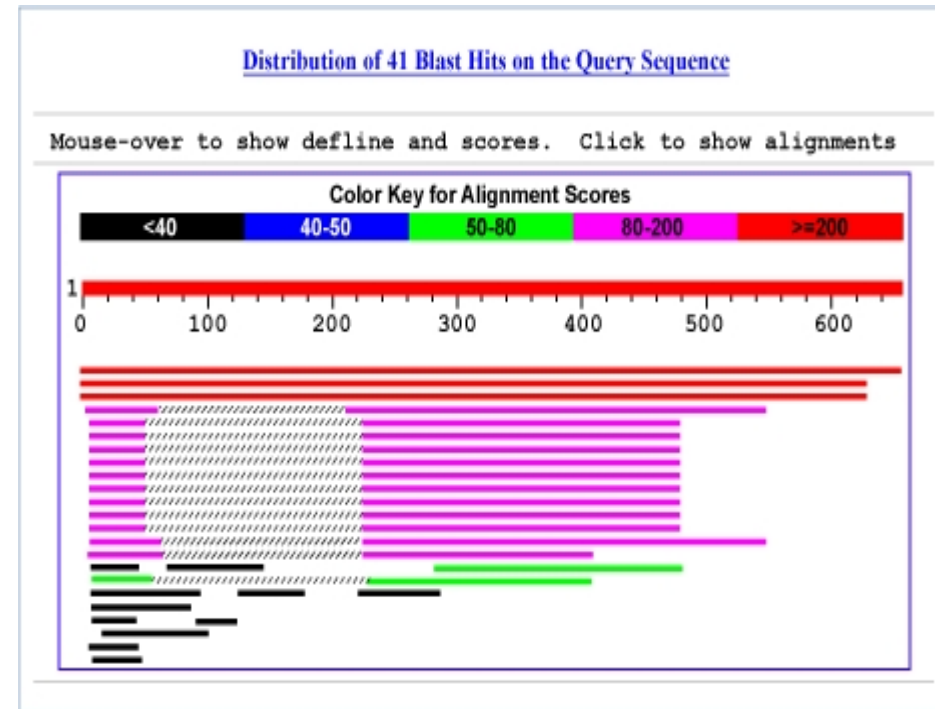
Engineering Analytics

Most usual application: flow visualization for the automotive industry:

○ optimization of the air resistance of vehicles

○ optimization of the flows inside a catalytic converter or a diesel particle filter

○ computation of optimal air flows inside an engine

Instead of solving these problems algorithmically, Visual Analytics can contribute to the understanding of the flows and to interactively change the construction parameters to optimize them

Also used for car crash simulations

Figure: cooling jacket simulation
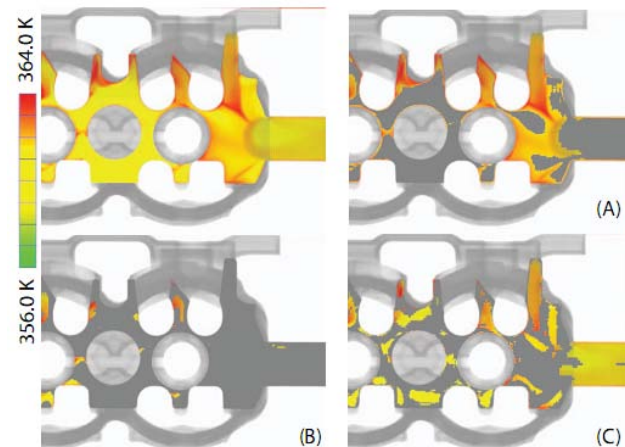
Personal Information Management

- Personal Information spread upon various digital devices like PCs, netbooks, PDAs, mobile phones etc
- Order needed
- Tools: Remail,
  Oxygen
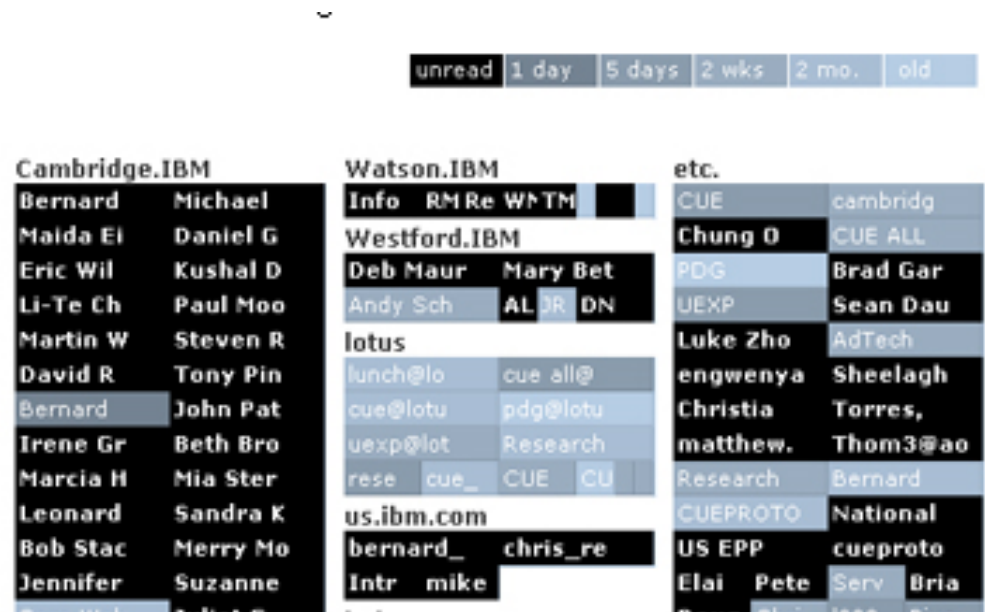
Mobile Graphics and Traffic
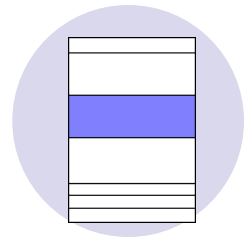
- A lot of information gathered on the road network daily:
  - vehicles' flow
  - accidents
  - weather conditions
  - data from cameras
  - GPS information for targeted vehicles

- Data integrated and presented in a meaningful way in order to give an overview of the current situation of the whole network and easily identify normal or abnormal patterns of network traffic
- VisTracks
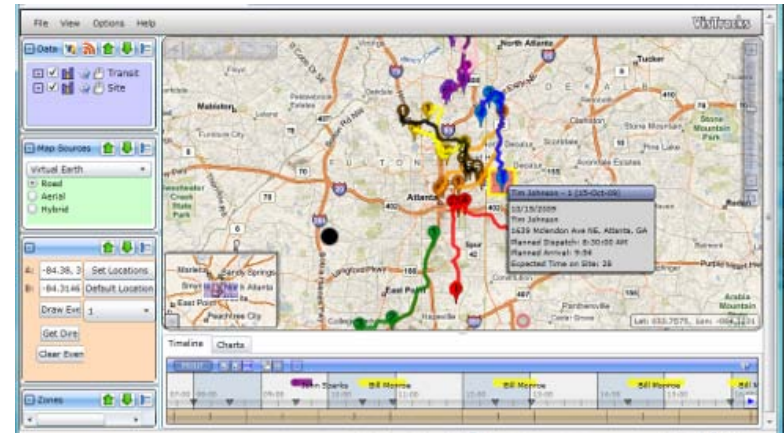
Visualization of Search results
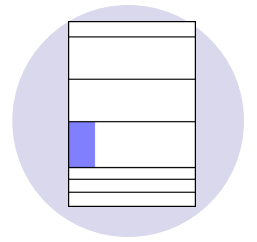
- Rapid evaluation of the result series
- Usually there are more than one search criteria => one dimensional relevance listing not enough
- Help get the overview of the search space

Search Crystal *A. Spoerri*
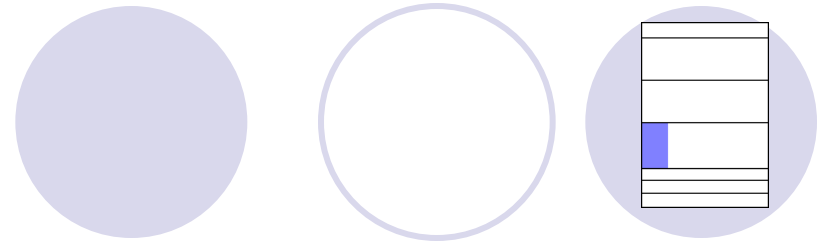
- Visual mash up of text + media search results

- Integration of many search engines

- Radial layout : distance from center indicates relevance

- Color/position indicates type

- Focus and context interaction
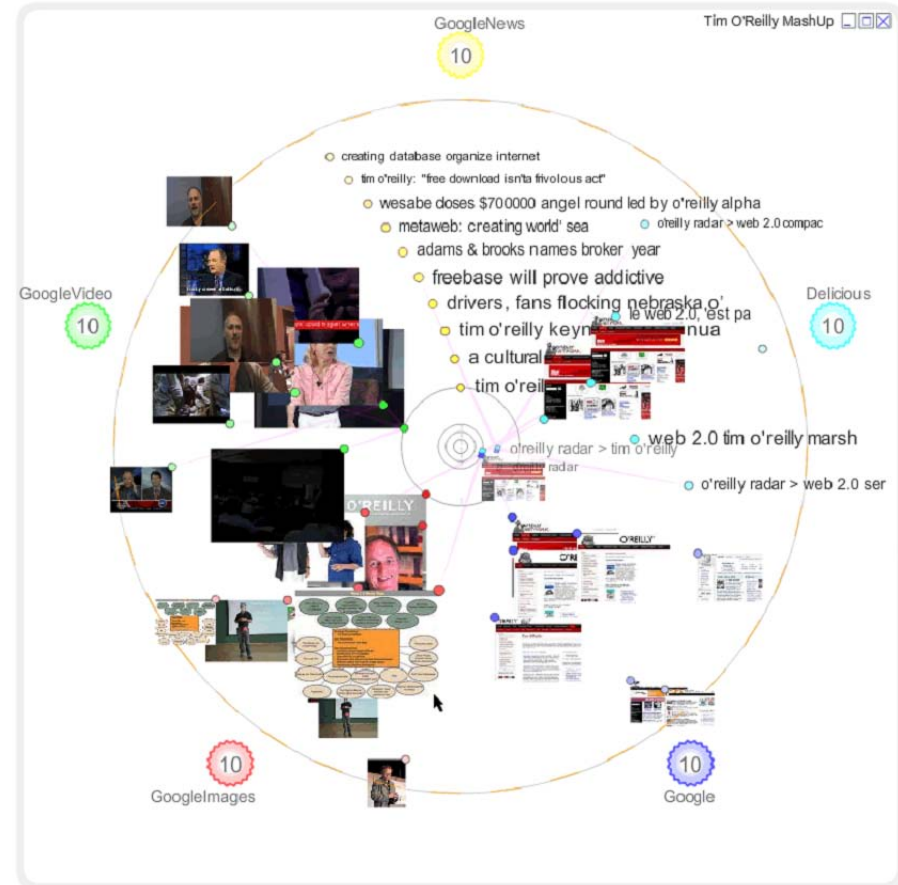
Quintura "see and find"

- Combination of text listing + cloud of words close to the ones in the query
- Helps users to manage their query and navigate through the relevant field by including or excluding the terms from the cloud.

TouchGraph

- Web search tool
  - provides the user with an overview of the search field
  - displays the connections among the search results and provides the context of each one of them
  - groups the results into categories of similar items easily identified by the different color
  - supports a zoom-in functionality
- Text listing with similarity weights
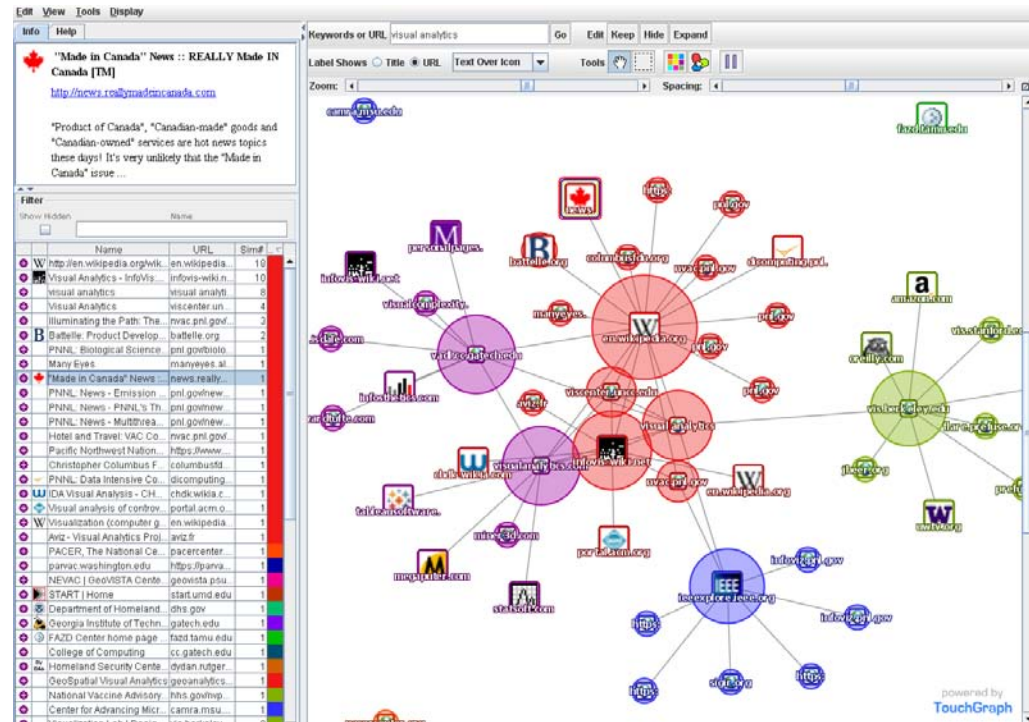
Resultmaps *Clarkson, Desai, Foley*

- A treemap that:
    - encodes a digital repository's full contents according to a hierarchical metadata attribute
    - accentuates certain nodes indicated by a query engine
    - interactively links to a text listing of query responses
- A tool more appropriate for use when there is more interest in metadata and not so much in data itself.
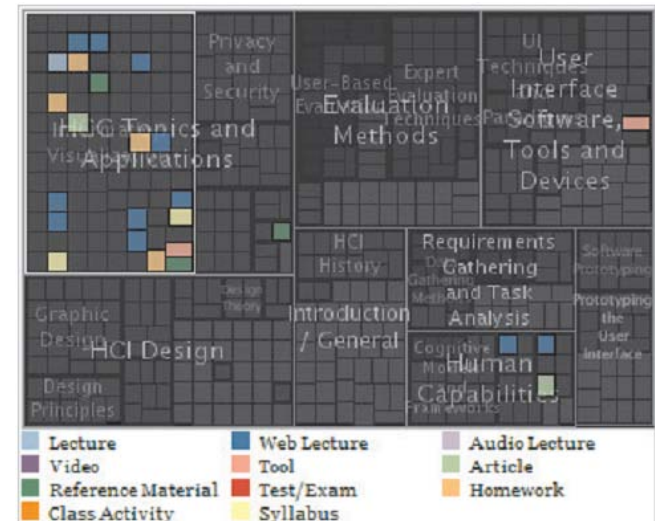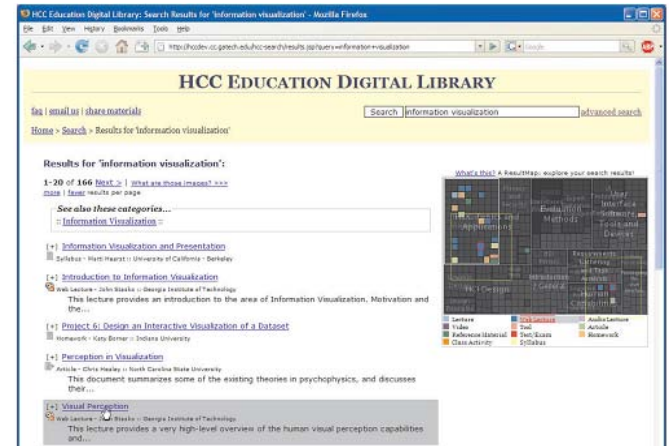
Treemaps for multicriterial search

Petkos, Darlagiannis, Moustakas, Tzovaras
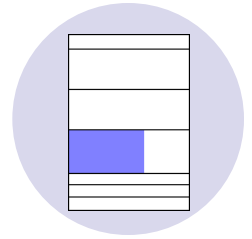
- Gives an overview or the search space of a 3D object database
- 4 shape descriptors for assessing similarity: user can tune each of them according to his search needs
- The top results for each shape descriptor are highlighted with different color in the treemap
- The relevance of items according to different search parameters can instantly be visually observed
- Important groups of items can be instantly determined due to the hierarchical structure of the treemap
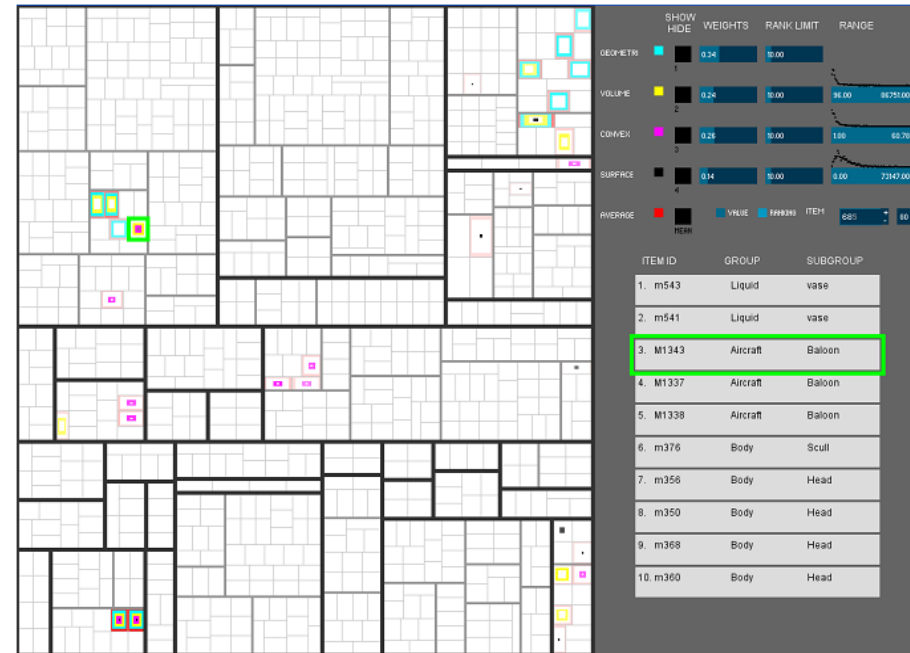
# "Search, Show Context, Expand on Demand" *van Ham, Perer*

- Search in large legal document databases with 300k interconnected items
- Use of the concept "Degree of Interest" (DOI) to tag the items with a value
- Show only the nodes over a DOI threshold after the user chooses a document from the result list
- Then the user can interactively explore the neighborhood of the selected document
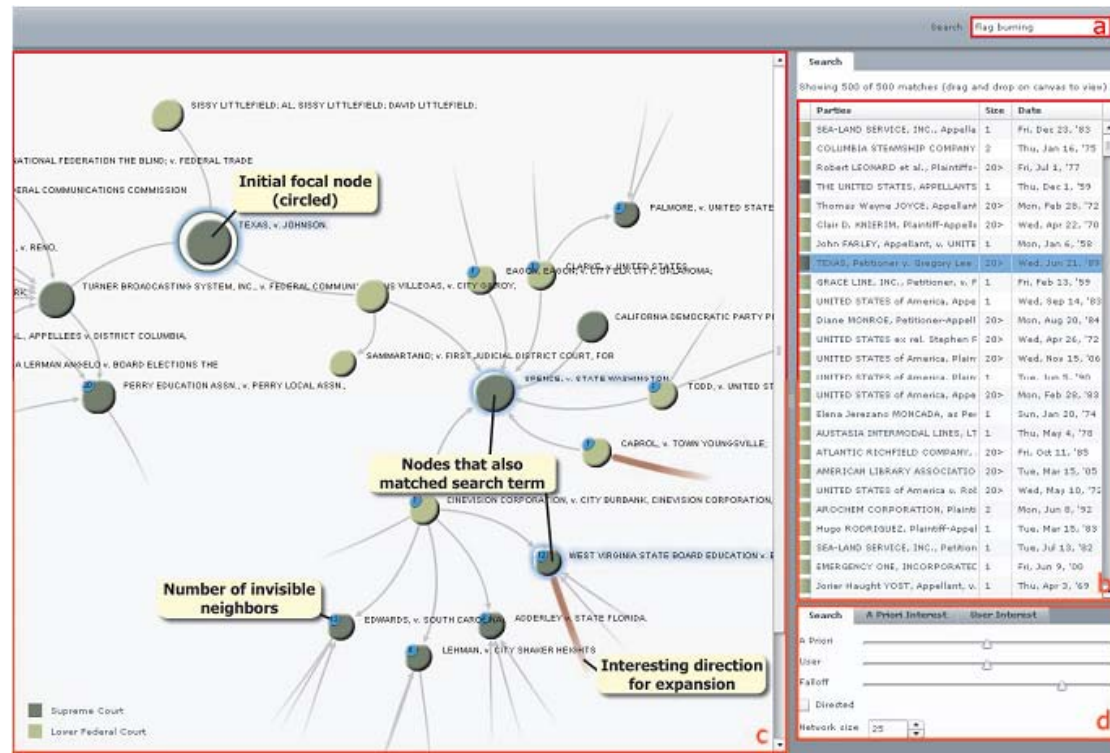
JustClick *Fan, Keim, Gao, Luo, Li*
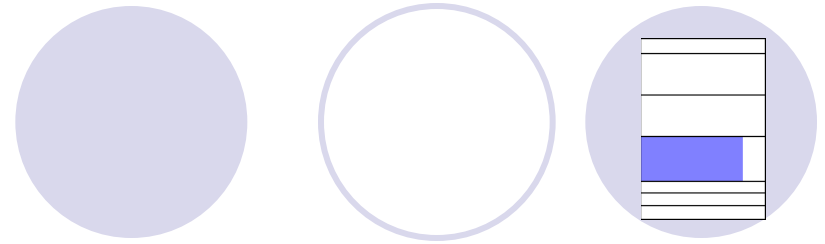
- Tool for image retrieval from the Flickr database via exploratory search
- Automatically generates the topic network
- Hyperbolic visualization – useful for avoiding visual cluttering
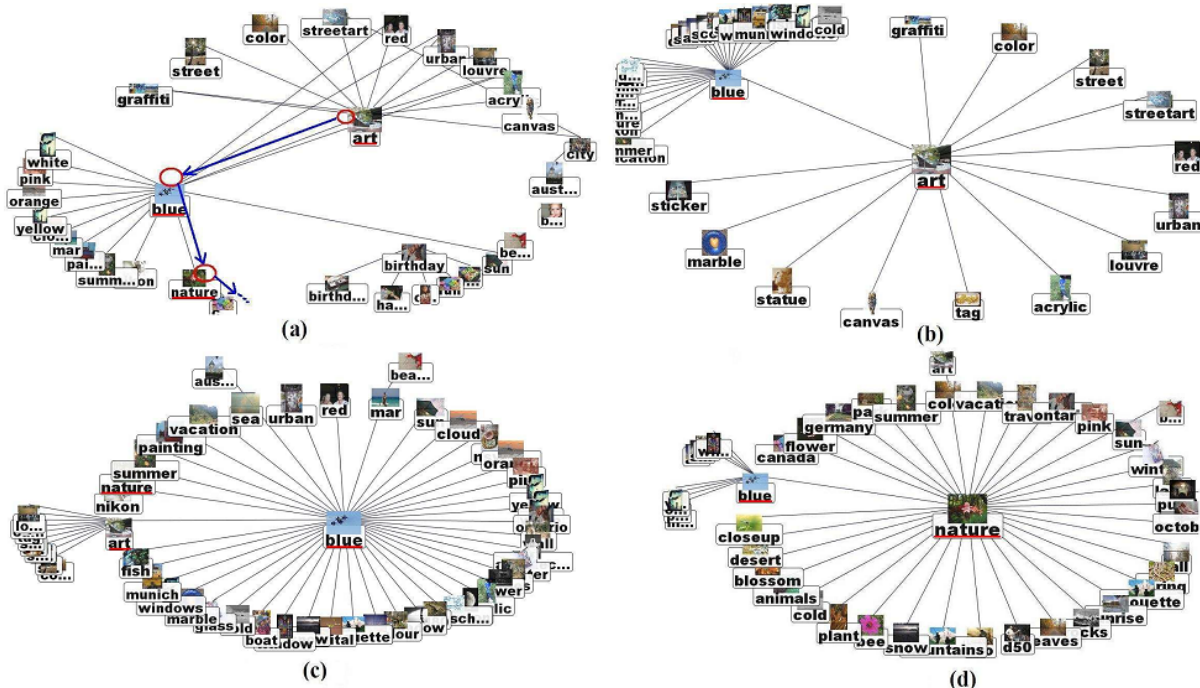
JigSaw *Stasko, Görg, Liu*

- Tool targeted to investigative analysts
- The system's main focus is to search for and display connections between entities across text documents
- Presents information about documents and entities through multiple distinct visualizations, called views
  - a List View
  - a Graph View
  - a Scatter Plot View
  - a Document View
  - a Calendar View
  - a Document Cluster View
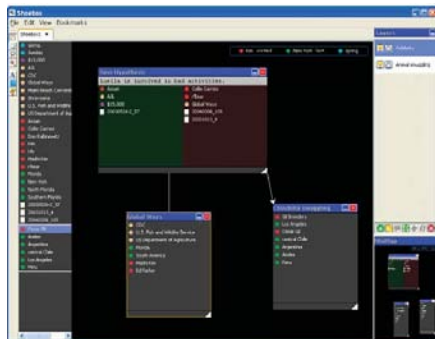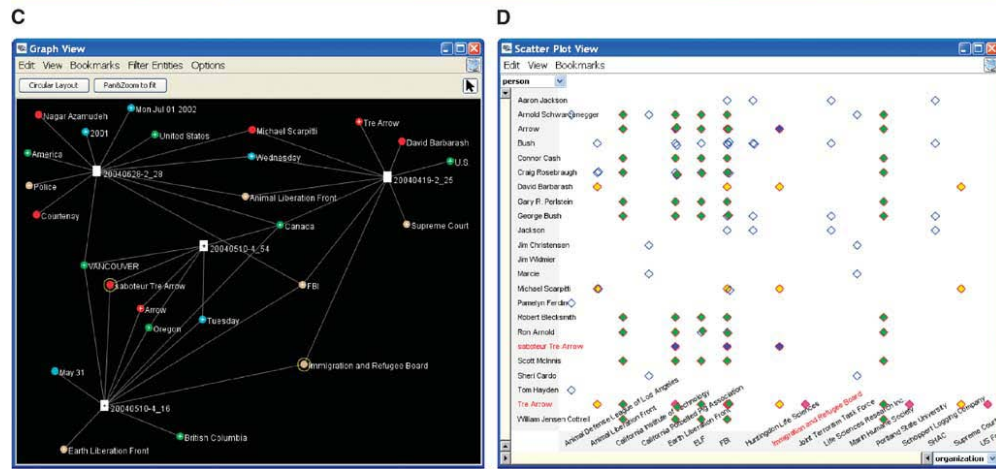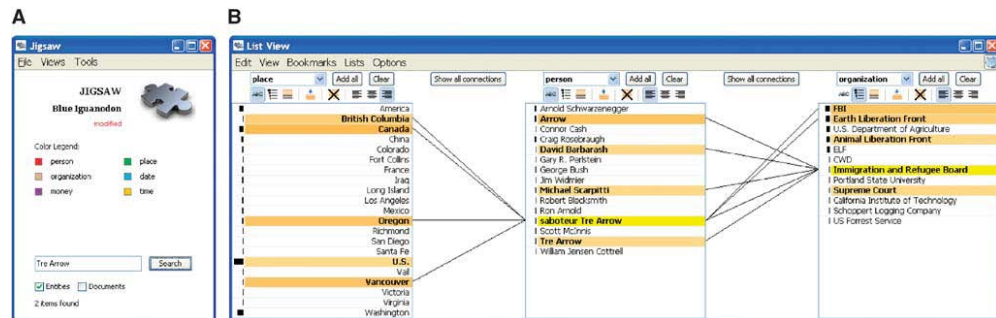
oScope visual Search

- Searches within 6 websites amazon, ebay, flickr, yahoo images, youtube and fotolia
- More dimension when applicable (eg price, salesrank)
- Interaction: change the arrangement, details on demand

Visualization of Network Flow/Security

- Various techniques including:
- Parallel coordinates:

- VisFlowConnect: detect and investigate anomalous Traffic between a local network and external domains
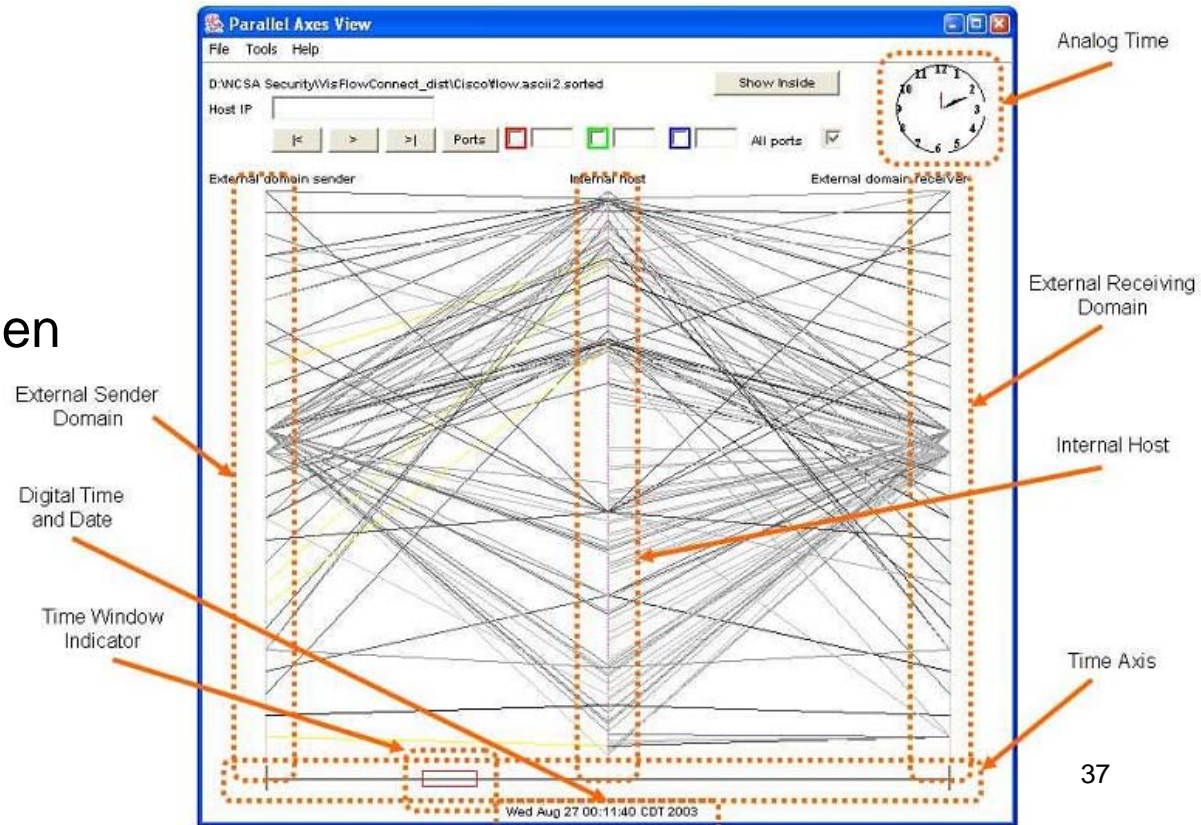
Treemaps++ *Mansmann, Keim et al*

- Visualization of Host Behavior for Network Security
- Interactive exploration of Data Traffic with Hierarchical Network maps
- Visualizing large-scale IP traffic flows
- Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring and Interpretation of Security Threats
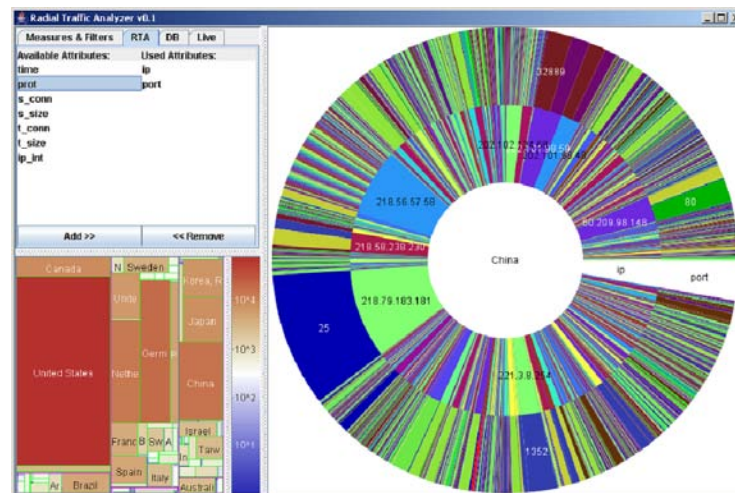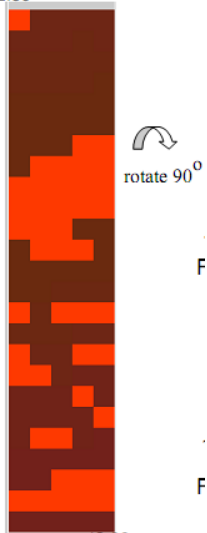
Dense Pixel Display++

Keim et al

- Visual Analytics of Anomaly Detection in Large Data Streams
- Monitoring Network Traffic with Radial Traffic Analyzer
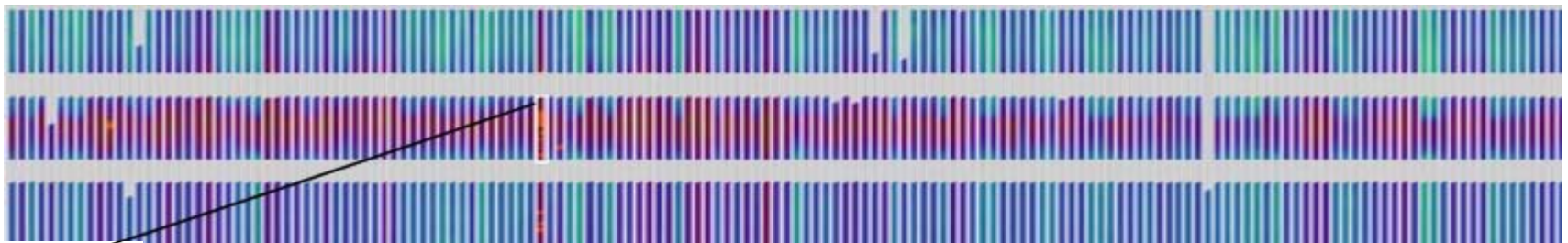
VAST challenge 3:
Genetic Sequences – Tracing the Mutations of a Disease

- Provided: Set of genetic sequences (AGTC...) of a virus during a disease outbreak. Genetic sequences are affected by mutations, random faults during the replication process, resulting in new diseases with new characteristics. Additionally, for each genetic sequence, a set of resulting disease characteristics are provided (symptom severity, mortality, complications, drug resistance, at risk vulnerability).

- Required: Trace likely mutation paths and detect most important mutations.

VAST challenge 3:
Genetic Sequences – Tracing the Mutations of a Disease

- Trace mutations: Treat sequences as nodes of a fully connected graph, compute distances between all pairs of sequences and compute the minimum spanning tree to identify likely mutation paths.

- Visualization, first option: Display tree with nodes placed regularly.

- Visualization, second option: Display tree with nodes placed using Multidimensional Scaling and using a user selected combination of disease characteristics to (re)compute dissimilarity (including sequence similarity).

VAST challenge 3:
Genetic Sequences – Tracing the Mutations of a Disease

Rich set of interaction mechanisms allows the user to perform a variety of functions, such as:

- Select features to be displayed and features to be used for dissimilarity computation.
- Select between regular and MDS based placement of sequences.
- Show or not show tree links.
- Marking of sequences, in order to track them in different visualizations.
- Linked highlighting of sequences between main visualization and list of marked sequences.

VAST challenge 3:
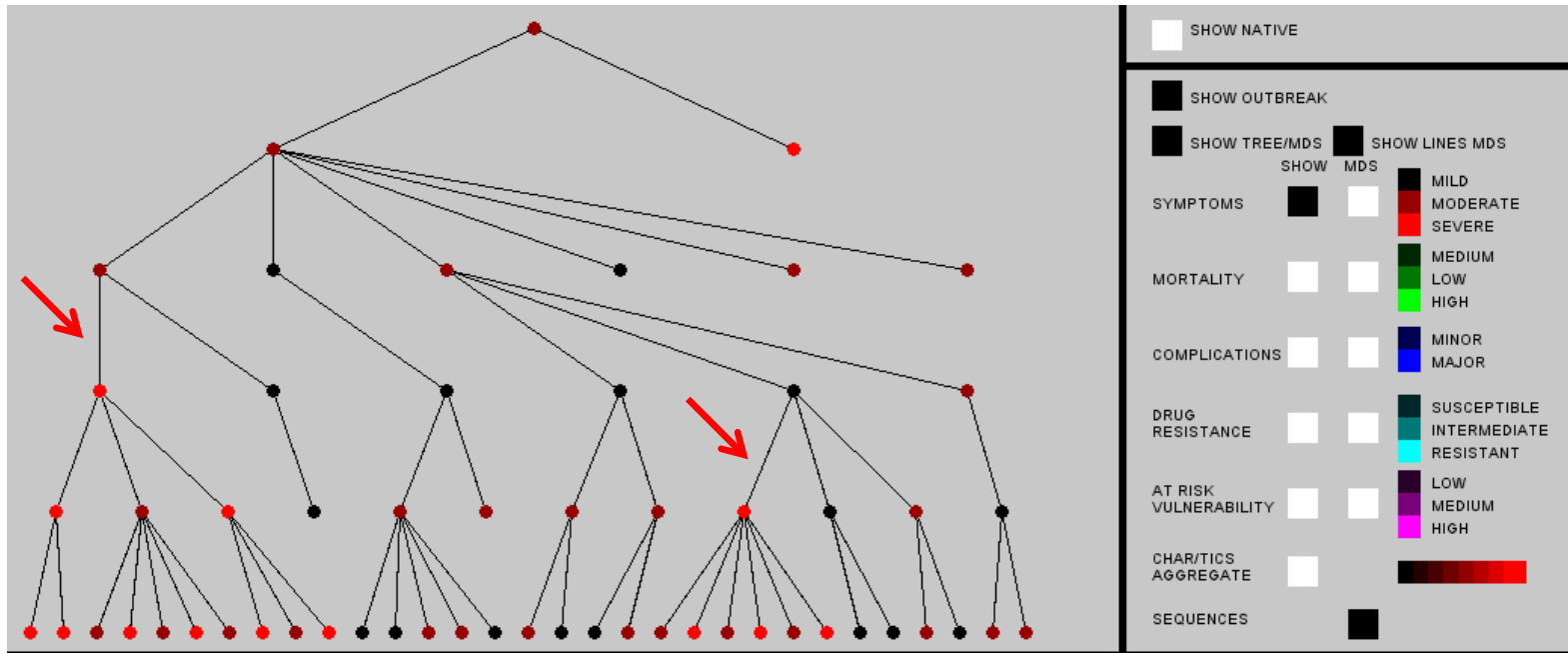# Genetic Sequences – Tracing the Mutations of a Disease
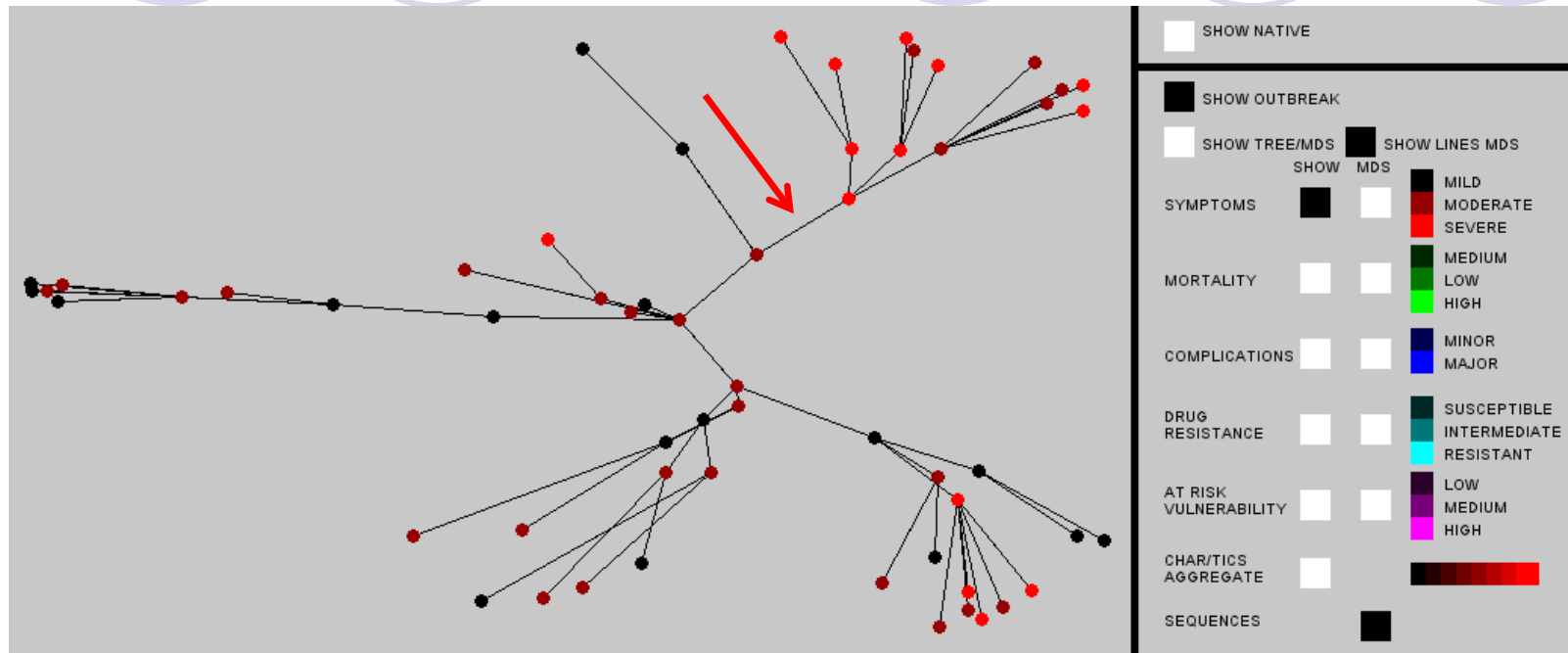
Instance of MDS based representation.

Symptom severity shown.

Looking for mutations that deteriorate symptom severity most.

Looking for subtree with most intense colors.

Regular placement of nodes.

Easy to visually detect subtrees with the most intence colors.

User can change the root node in order to obtain a clearer visualization, if required.

I-SEARCH Project

The aim of I-Search is to provide a novel unified framework for multimodal content indexing, sharing, search and retrieval.

Within I-Search, visual analytics enters within the search interface. Novel interfaces for providence of multimodal search criteria, presentation of results and relevance feedback will be created. Emphasis will be given on:

- The inclusion of multiple modalities at once.
- The presentation of a very large number of results using appropriate visualization and interaction mechanisms.

I-Search:
Treemap for presentation of multicriterial 3D item search
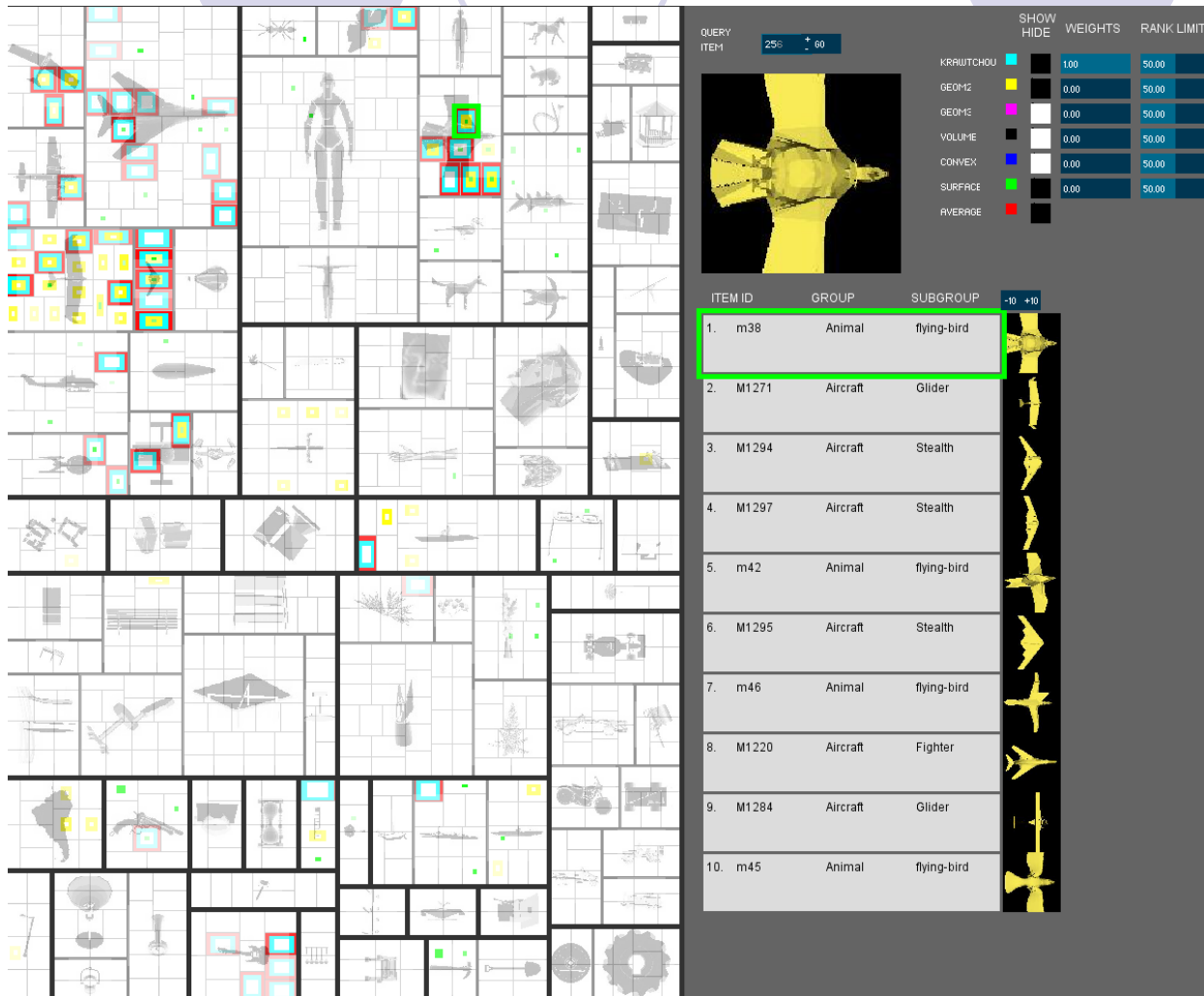
Database organized in a two level hierarchy.

The relevance of iterms according to different search criteria is denoted with different color codings.

User can select how many items to be displayed.

Multiple relevant items and groups of items (according to different criteria can easily be identified)

I-Search: Next steps

Develop methods for interactive and rich presentation of search results according to multiple modalities, not just multiple features.

Allow for huge databases, possibly by utilizing a purely hierarchical approach for presentation of results.

Develop appropriate relevance feedback mechanisms.

VIS-SENSE

- Goal: Research and development of techniques for the identification and prediction of abnormal network events
  - Data mining
  - Visual Analytics
  - Hypothesis formulation and visual validation
- Use cases:
  - Attack attribution: Identification of the "root cause" of the attack by clustering "attack-full" data
  - BGP (Border Gateway Protocol) hijacking: Detect attacks on the control plane of the Internet and try to correlate with well-know attacks, e.g. spamming, DoS
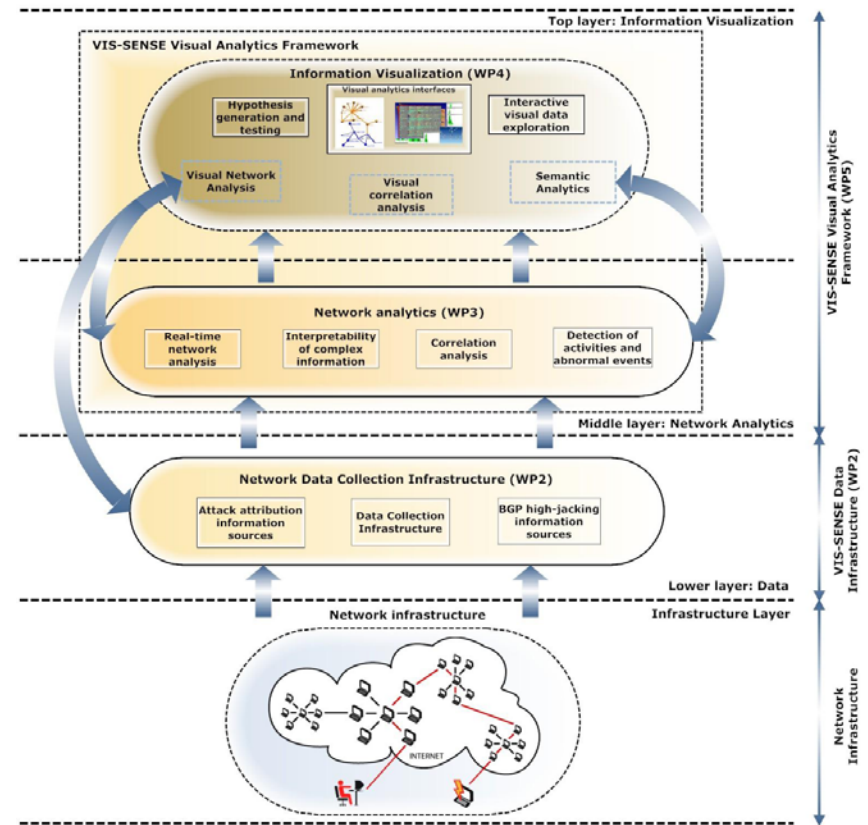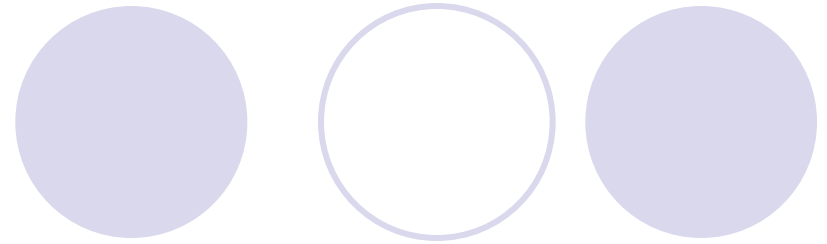
BGP Hijacking

- Prefix hijacking as invalid (a) origin or (b) transit
- Objective: Blackholing, imposture, interception
- Investigate the dynamics of BGP graph for the detection of major BGP events
- Use temporal information by modeling distances of subsequent graph instances
- Investigate the use of bipartite (AS – prefix) models
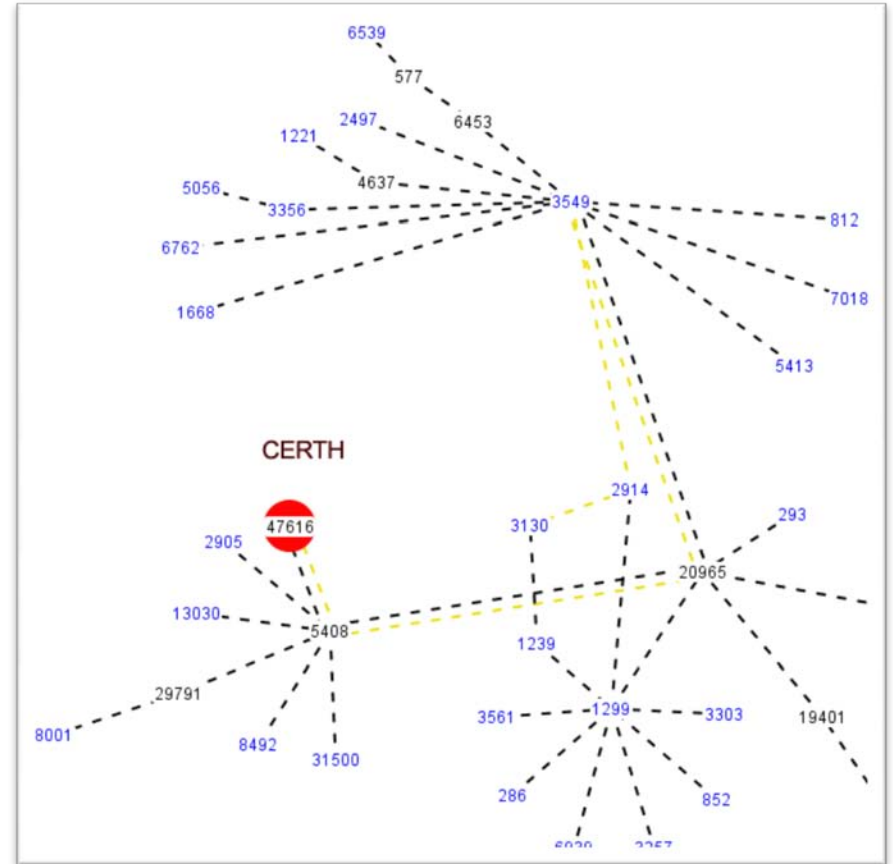
Attack Attribution

- Input: Attack-full data
- Objective: Clustering of attacks that are likely due to the same root cause
- Approach: (a) Initial clustering in attack events (b) cluster attack events using various features (c) data fusion
- Investigate the use of alternative clustering and data fusion techniques
- Use Visual analytics techniques to guide the user in the attack attribution process
  - Visualize the clustering process using graphs
  - Enable user to interact with the framework in order to explore the characteristics of each cluster, using time series plots, histograms, bar plots etc
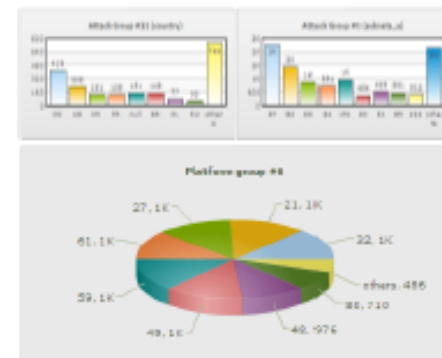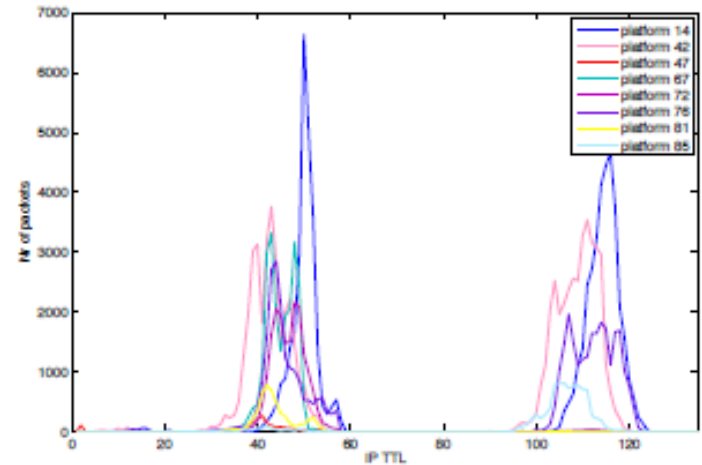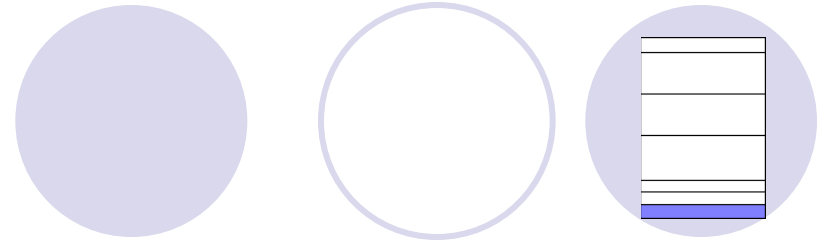
Visual Analytics

- Use multiple views of the data (graphs, plots, MDS plots using various distance measures)
- Allow correlation of multiple data sources
- Allow interactivity by enabling user to supply queries, define his/her own similarity measures and enrich input data by other sources
- Allow hypothesis formulation and visual validation
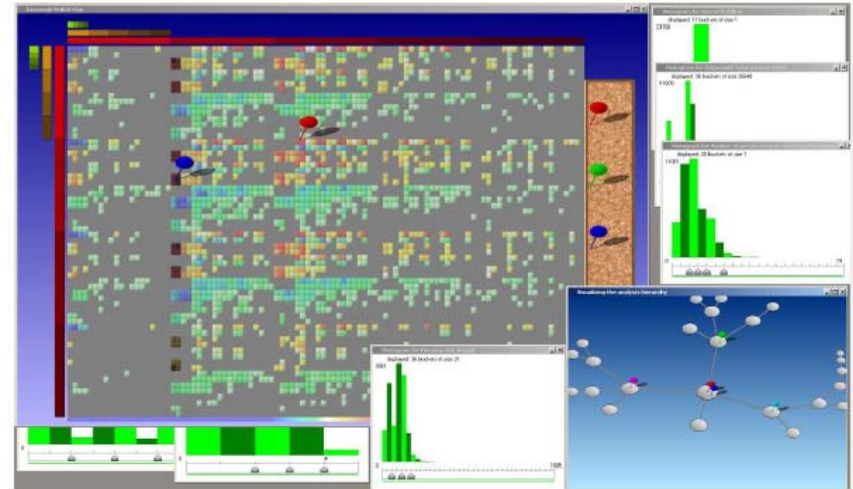
Future Challenges

- Visualization: novel techniques for data pre-processing

- Data Management: Find integrated representations for different data types

- Data Mining and Analysis: Inclusion of semi-structured+complex data

- Perception and Cognition: research of perceptual, cognitive and graphical principles => lead to improved visual communication of data and analysis result

- Spatio-Temporal Data Analysis: scale, data uncertainty

- Evaluation: not yet a common framework for evaluation.