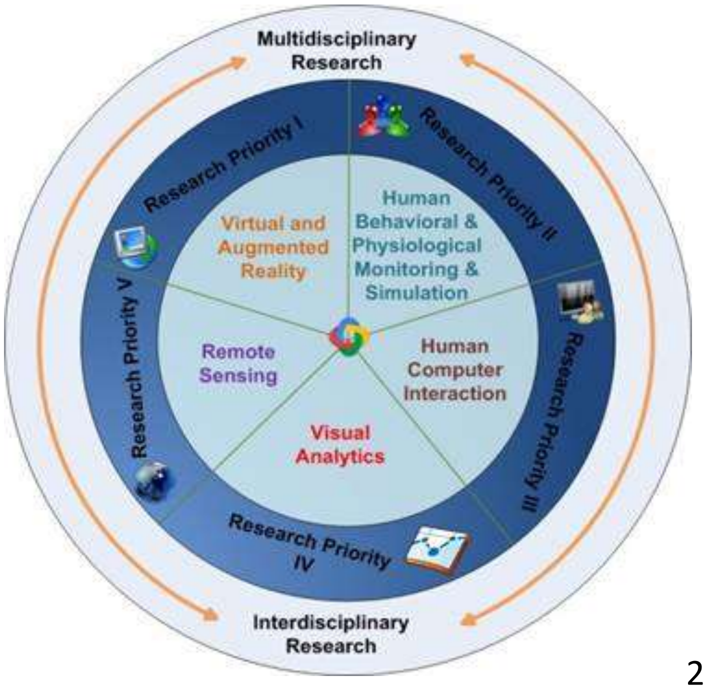
Visual Analytics for Efficient Processing & Analysis of Big Data

Dr. Dimitrios Tzovaras

Director of the Information Technologies Institute
(Researcher A’)

Research areas of the
Centre of Research & Technology – Hellas (CERTH) / Information Technologies Institute (ITI)

- Virtual and augmented reality
- Behavioral, physical and affective observation, modeling and simulation of persons/groups of people
- Human Computer Interaction (HCI)
- **Big data**
- **Visual analytics**



Presentation outline

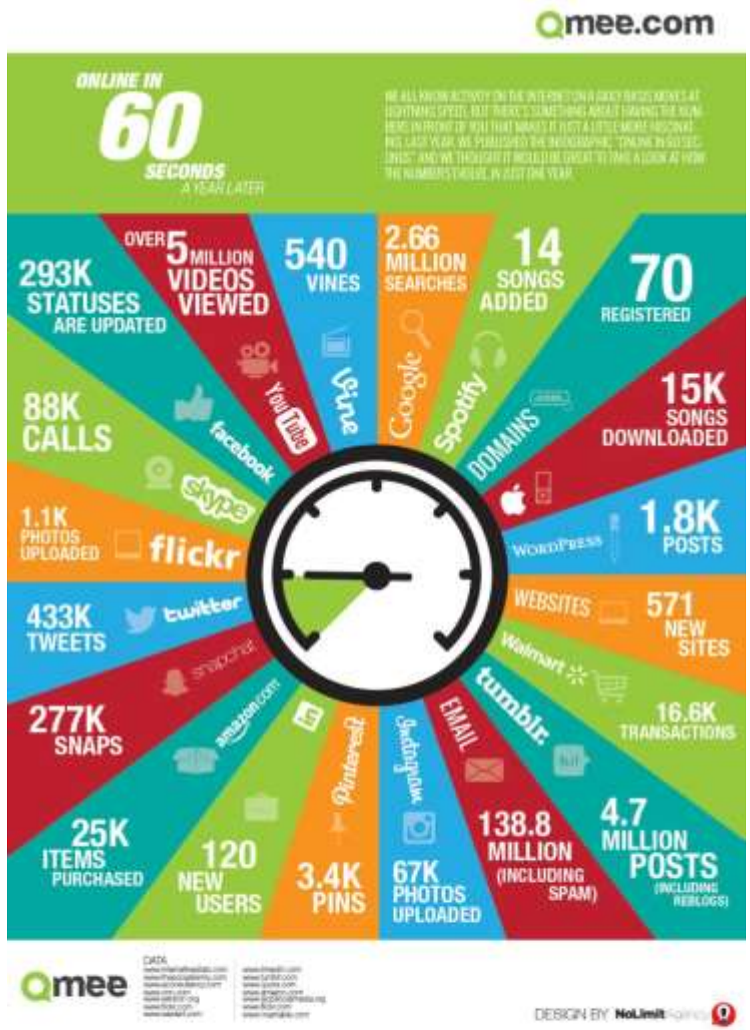
1. Introduction
   - What is Big Data
   - Motivation
   - Visual analytics for Big Data
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Data sources

Social media and networks

Sensor technology

Stock exchange



Mobile Devices

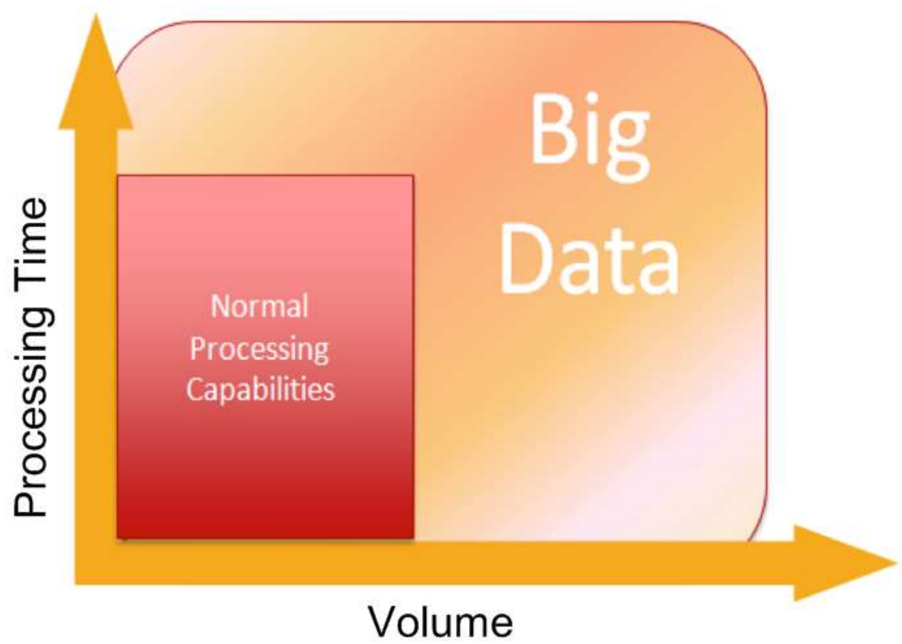
Scientific instruments

Wired and wireless Networks

http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/

Big Data definition

“Big Data” is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it.

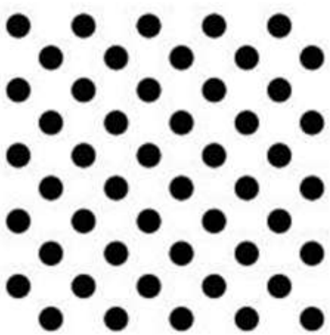


http://www.esg-global.com/blogs/big-data-a-better-definition/

Information Technologies Institute

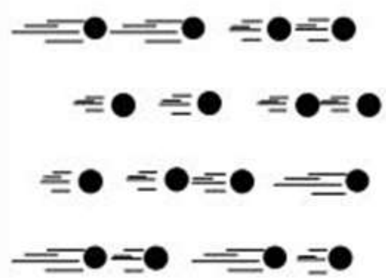
The four V's of Big Data

Volume



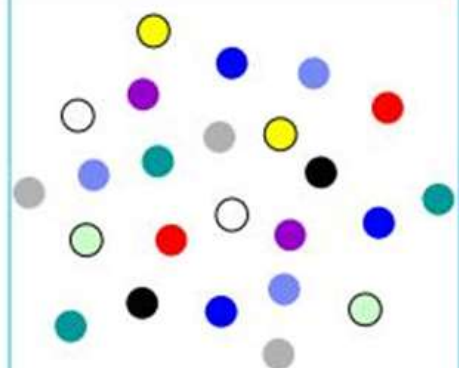
- Large volumes of data produced daily (~10^{18} bytes).
- 43 trillion gigabytes of data until 2020.

Velocity



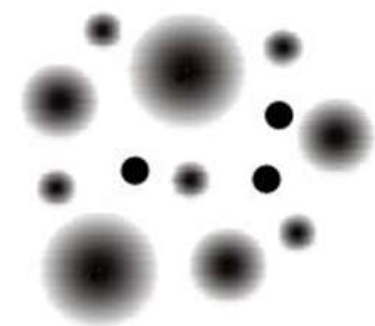
- Analysis of streaming data from networks and sensors

Variety



- Different forms of data
- Data from social media, video, sensors etc.

Veracity



- Uncertainty of data
- Large volume of data is not precise/valid

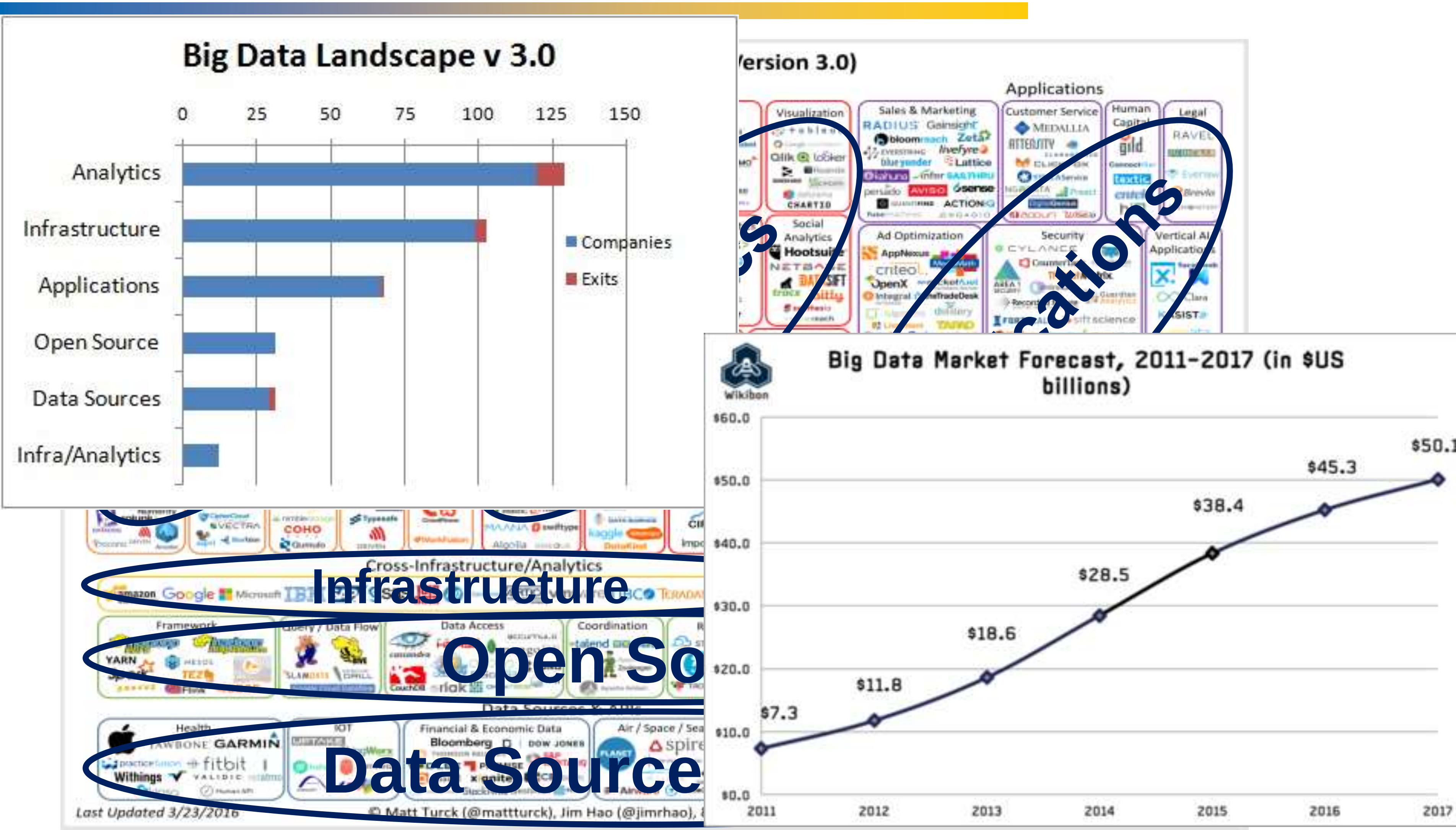
Big Data Technologies

- **Infrastructure**
  - Provide software/hardware for the fast and efficient storage, retrieval, processing and monitoring of Big Data
- **Analysis**
  - Information Visualization
  - Automated Data Analysis (e.g. machine learning, statistical analysis)
  - Visual analytics (Combination of Information Visualization and Automated Data Analysis)
- **Applications**
  - Solutions to specific fields (e.g. finance, health etc.)
- **Some Technologies are open source**
- **Some deal only with data collection (data sources)**

Presentation outline

1. Introduction
   - What is Big Data?
   - Motivation - Big Data Landscape
   - Visual analytics for Big Data
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Big Data Technologies Landscape



http://mattturck.com/2016/02/01/big-data-landscape/

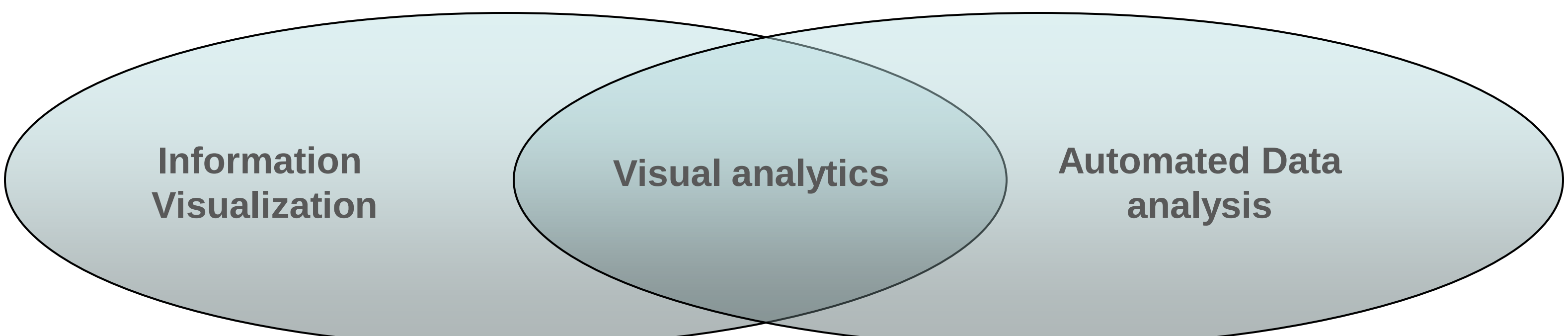
Presentation outline

1. Introduction
   - Big Data
   - Motivation - Big Data Landscape
   - Visual analytics for Big Data
     - Definition
     - Visualization Taxonomy
     - Visual Analytics Challenges & SoA
     - Visual Analytics Application Fields
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Information Visualization vs Automated Data analysis

- **Information Visualization**
  - \+ uses power of human visual system
  - \+ user-guided analysis possible
  - \+ detect interesting features and parameter selections
  - \+ understand results in context
  - – limited dimensionality
  - – often only qualitative results

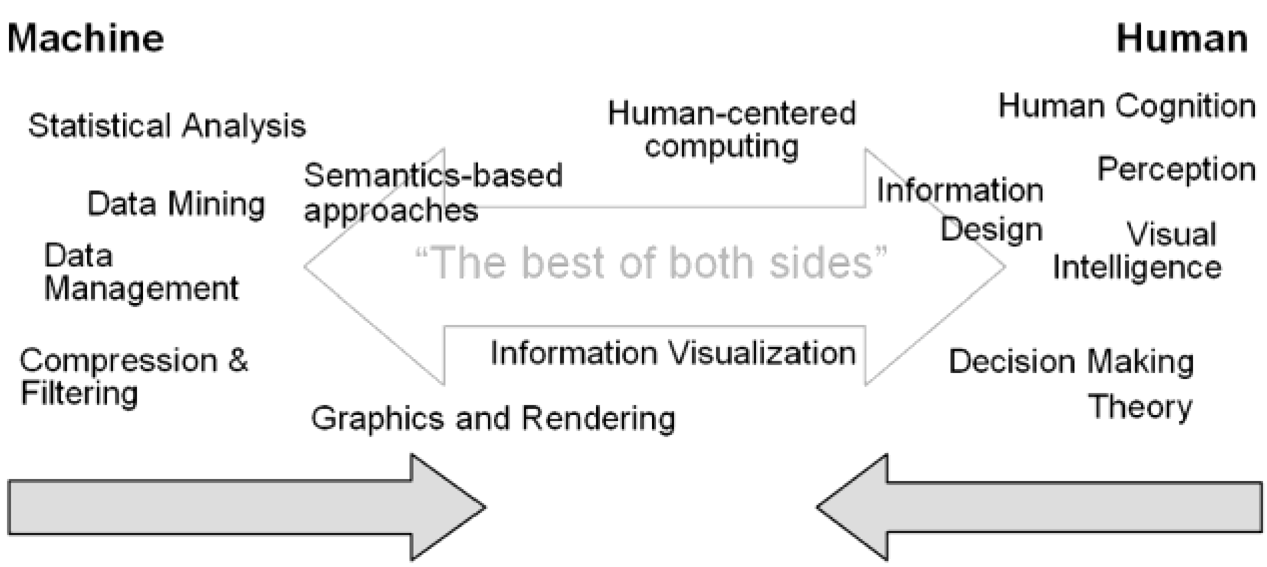
- **Automated Data analysis**
  - \+ hardly any interaction required (after setup)
  - \+ scales better in many dimensions
  - \+ precise results
  - \- needs precise definition of goals
  - \- limited tolerance of data artifacts
  - \- result without explanation
  - \- computationally expensive



Visual Analytics: The best of both Worlds

"Visual analytics is the science of analytical reasoning supported by interactive visual interfaces" *[Thomas et al. "Illuminating the Path", 2005]*

"Visual analytics combines automated analysis techniques with interactive visualizations for an effective understanding, reasoning and decision making on the basis of very large and complex datasets" *[D. Keim et al. "Visual Analytics: Definition, Process, and Challenges", 2008]*



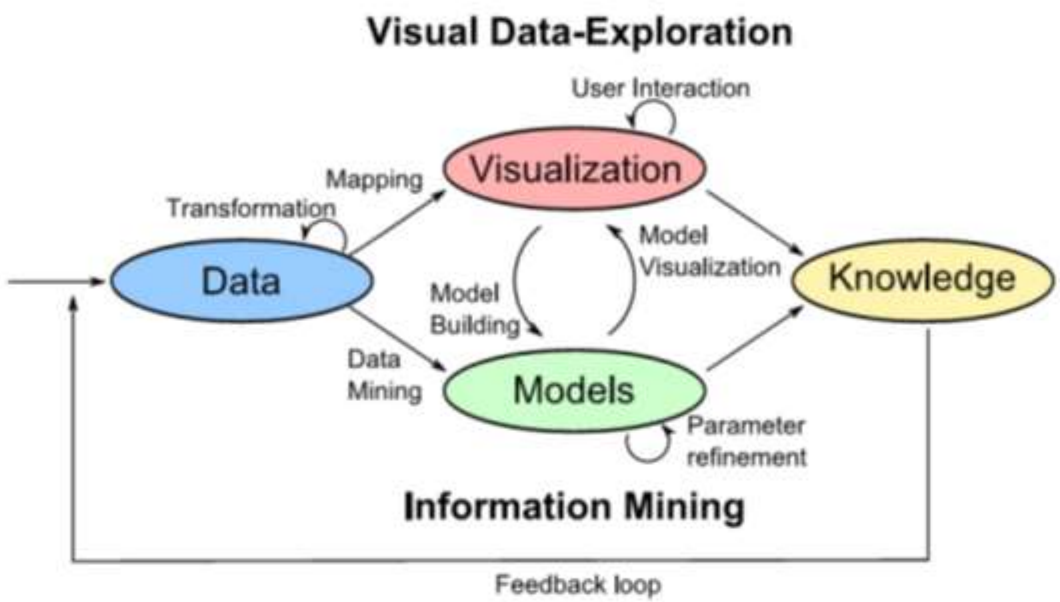
[Keim et al. "Visual Analytics: Definition, Process, and Challenges", 2008]

Visual Analytics process

Is an iterative process involving:

- Information gathering
- Data pre-processing
- Knowledge representation
- Interaction and decision making

Leading to user insight / solution



[Keim et al. “Visual Analytics: Definition, Process, and Challenges”, 2008]

Presentation outline

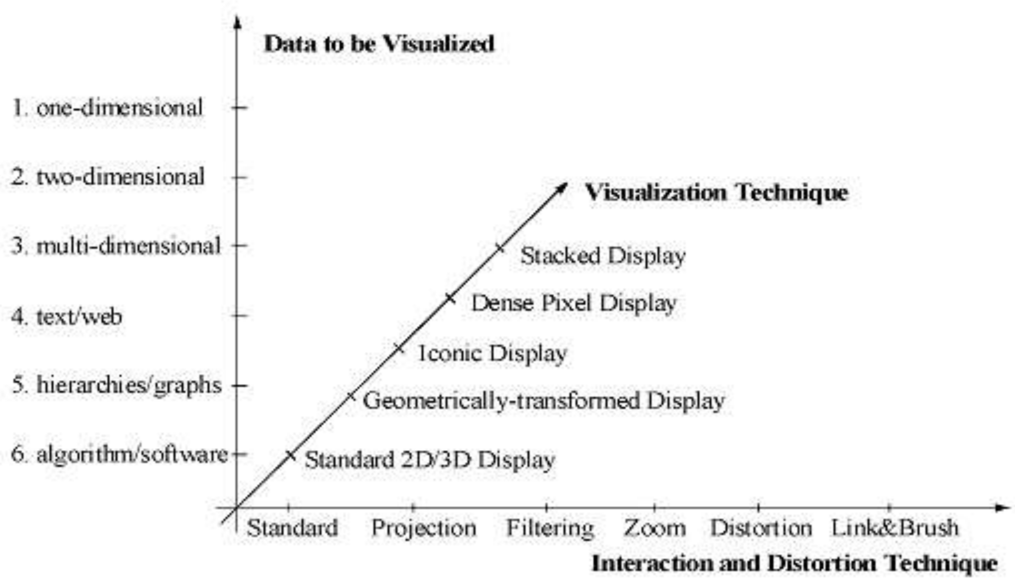
1. Introduction
   - Big Data
   - Motivation - Big Data Landscape
   - Visual analytics for Big Data
     - Definition
     - Visualization Taxonomy
     - Visual Analytics Challenges
     - Visual Analytics Application Fields
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Visual Analytics taxonomy

According to Keim et al.

Three dimensional taxonomy according to Keim et al.:

- Data type
- Visualization technique
- Interaction & distortion technique

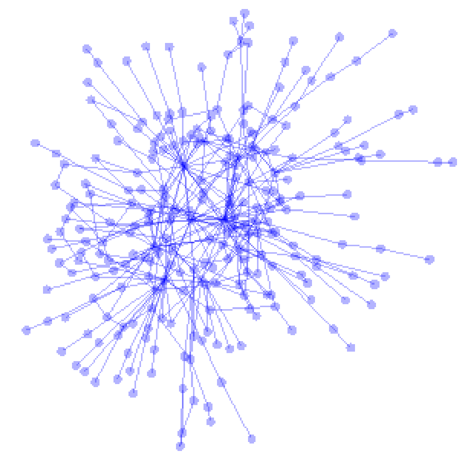


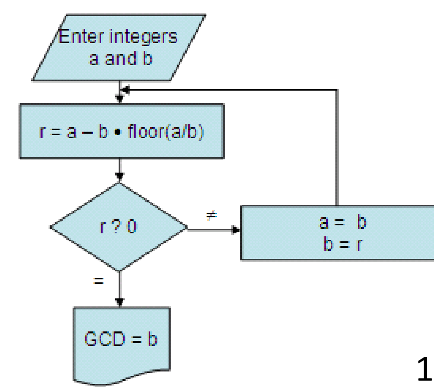
[Keim et al. "Information visualization and visual data mining.“, 2002]

Taxonomy 1/3

According to Data type

- 1D data, e.g. temporal data
- 2D data, e.g. geographical maps
- Multi-dimensional data, e.g. relational tables
- Hierarchies & graphs, e.g. telephone calls
- Text & hypertext, e.g. news articles and Web documents
- Algorithms & software, e.g. debugging operations



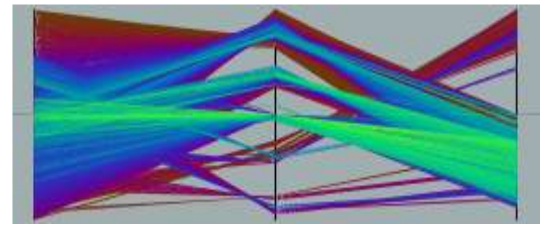


Information Technologies Institute

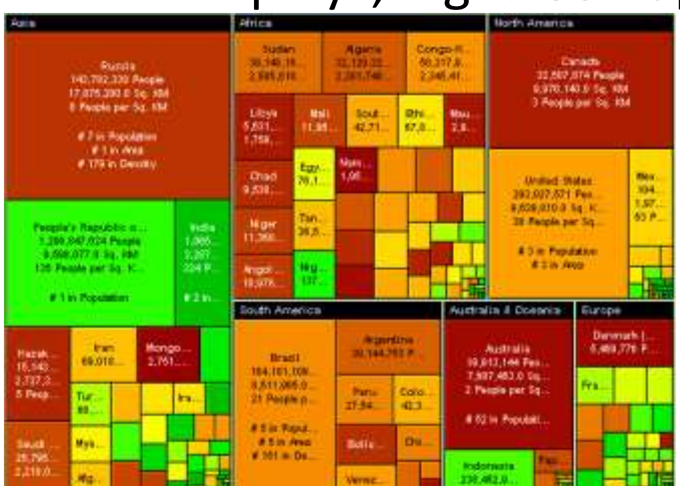
Taxonomy 2/3

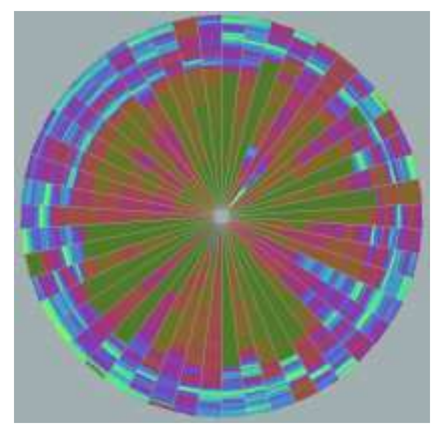
According to Visualization technique

- Standard 2D/3D displays, e.g **bar charts** & **x-y plots**
- Geometrically transformed displays, e.g. landscapes & **parallel coordinates**



- Icon-based displays, e.g **stick figures** & **star icons**
- Dense pixel displays, e.g. **recursive pattern** & **circle segments techniques**
- Stacked displays, e.g. **treemaps** & **dimensional stacking**





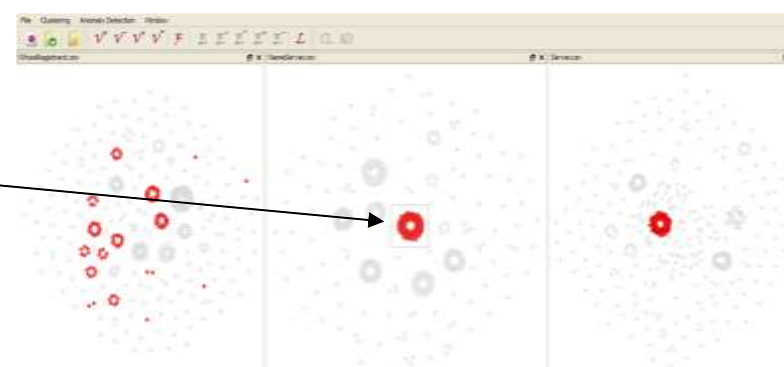
Taxonomy 3/3

According to Interaction technique

- Interactive **Projection**
  - dynamically change the projections → explore multidimensional datasets
- Interactive **Filtering**
  - focus on interesting subsets
- Interactive **Zooming**
- Interactive **Distortion**
  - hyperbolic, spherical

- Interactive **Linking** & **Brushing**
  - combine different visualization methods →overcome the shortcomings of single techniques

User selects this cluster



Presentation outline

1. Introduction
   - Big Data
   - Motivation - Big Data Landscape
   - Visual analytics for Big Data
     - Definition
     - Visualization Taxonomy
     - Visual Analytics Challenges & SoA
     - Visual Analytics Application Fields
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Challenges in Visual Analytics

1. ***Quality of Data & Graphical Representation***: Present the notion of data quality, and the confidence of the analysis algorithm

2. ***Visual Representation & Level of Detail***: Find a balance between overview and detailed views

3. ***Infrastructure:*** Special data structures and mechanisms for handling large amounts of data

4. ***User Interaction Styles & Metaphors***: Development of novel and intuitive interaction techniques to simplify the whole analysis process

5. ***Display Devices:*** Adapt to the constantly evolving display devices

6. ***Scalability with Data Volumes & Data Dimensionality***: Scale with the size and dimensionality of the input data space.

7. ***Evaluation:*** Provide a theoretically founded evaluation framework for the perception of visualization

Focus of the Visual Analytics research community

[Keim et al. "Visual Analytics: Definition, Process, and Challenges", 2008]

Clutter Reduction & Display devices

Definition

- *… is the process of deforming the original data representation by enlarging/condensing regions of the input space, so as to visualize previously hidden patterns (high cluttering rate, occlusions due to high data density, etc.).*

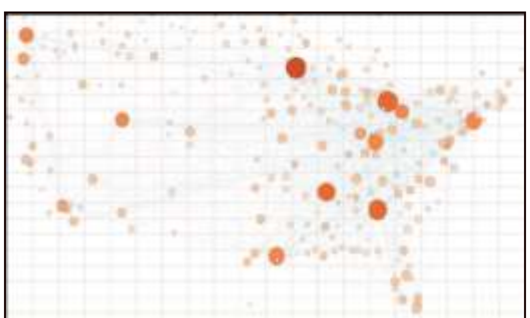
Visual clutter can mislead users into deriving wrong conclusions, and increase the decision condense on erroneous decisions. It can be caused when **large data volumes** are visualized **on small display devices**, which reduce the visualization space and its information capacity.

[Ellis and Dix, 2007]

Clutter Reduction & Display devices

Selected publications in CR problems

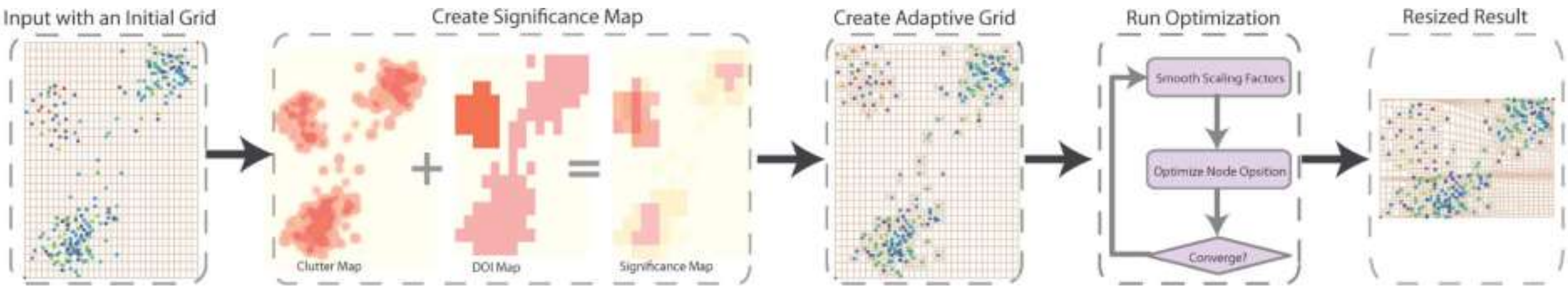
- Several methods have been proposed for clutter reduction, through suggesting…
  - a modifiable point size of the visualized items [Woodruff et al. 1998]
  - a spatially modifiable opacity of the visualization [Fekete 2002]
  - the visual clustering of similar items in order to save space [Bederson et al. 2002]
  - the compression of the visualization via smart sampling [Derthick et al. 2003]
  - interactive exploration of the visualization [Lad et al. 2006]
  - a spatiotemporal animation feature in order to comprehensively visualize more dimensions [Johansson et al. 2006]
  - non-linear deformations for zooming in/out in more/less significant areas [Wu et al. 2013]



CR in Display Devices

SoA Method A

- **ViSizer** method for fitting visualizations on small screens



- Define the Significance Map by combining:
  - Degree Of Interest Map (DOI): The interestingness of the regions in the visualization (e.g. high degree nodes in a graph)
  - Clutter Map: Find crowded regions, with excess/unorganized visual items
- Define a grid *M* = (*V,E,F*), with vertices V, edges E and quad faces F
- Goal is to change the vertex positions and find a new grid M' that fits the new display size, while the distortion of significant regions in minimum

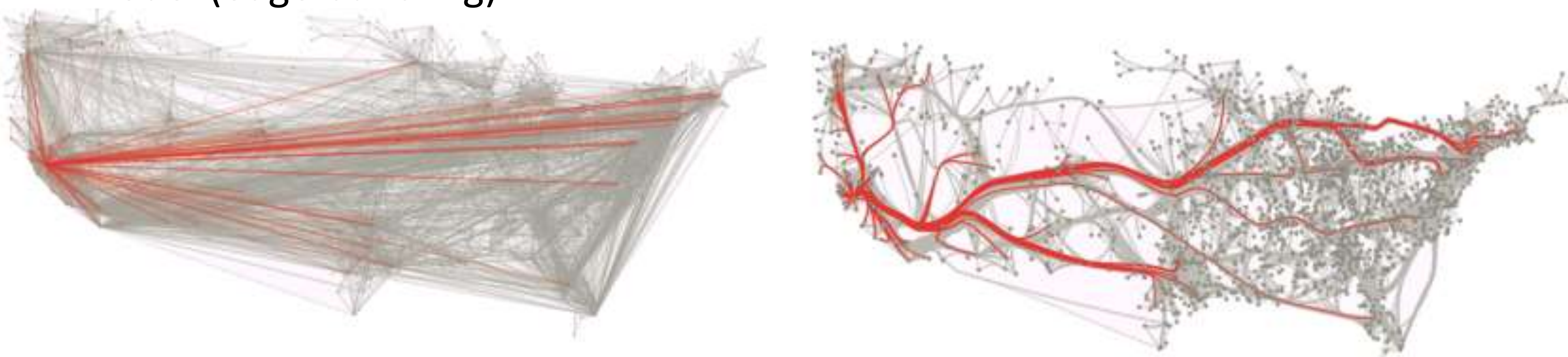
→minimization of the **total grid deformation energy D**, consisting of <u>total quad deformation</u> and <u>total edge deformation</u> $D = D_u + D_l$

[Wu Yingcai, et al. "ViSizer: a visualization resizing framework.", 2013]

Scalability & Data Dimensionality

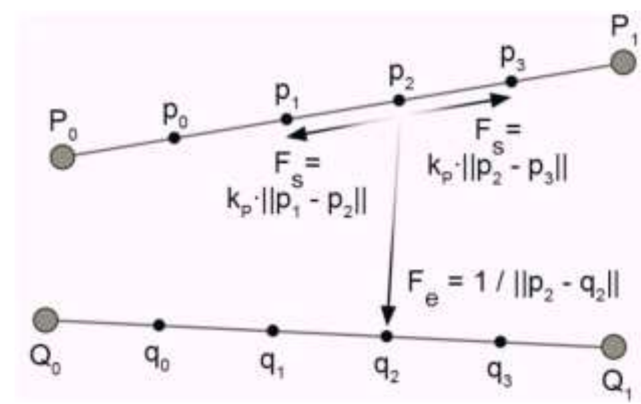
SoA Method B

- Combine edges in a graph in order to reduce visual clutter using a force directed model (edge bundling)



- The force on each segment p_i of edge *P* is defined as follows:

$$F_{p_i} = F_{s_i} + F_{e_i} = k_p\left(\|p_{i-1} - p_i\| + \|p_i - p_{i+1}\|\right) + \sum_{Q \in E} \frac{1}{\|p_i - q_i\|}$$



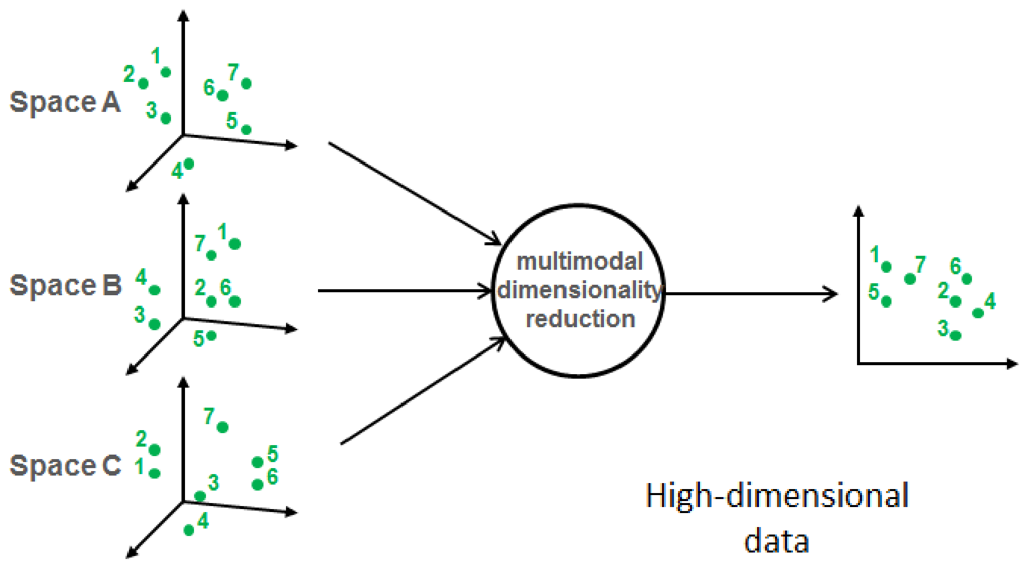
where F_{s_i} is the neighboring spring force, k_p the spring constant, F_{e_i} the electrostatic force applied by all segments except for the ones in *P*, i.e. the set *Q*

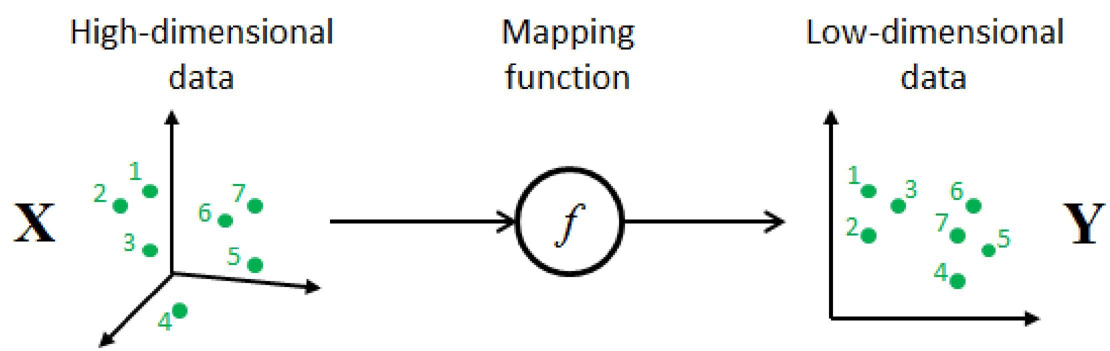
[H. Danny et al. "Force Directed Edge Bundling for Graph Visualization." 2009]

Scalability & Data Dimensionality

Definition

… is the process of mapping high-dimensional data to low-dimensional data, so that data relationships are preserved.





Scalability & Data Dimensionality

Selected publications in mDR problems

- Several methods have been proposed for multimodal Dimensionality Reduction, through suggesting…
  - optimizing the features that form dynamic & high-dimensionality bags of multimodal objects [Zhang and Weng, 2006] [Zhuang et al, 2008]
  - the projection of inter-disciplinary modalities on a common space [Hardoon et al, 2004] [Zhang and Weng, 2006] [Zhang and Meng, 2009] [Rasiwasia et al, 2010]
  - parallel training on each modality type and late-fusion [Nigam and Ghani, 2000] [Brefeld and Scheffer, 2004] [Eaton et al, 2010]
  - pair-wise cross-modal distance fusion [Axenopoulos et al, 2011] [Gonen and Alpaydn, 2011] [Lin et al, 2011]
  - multi-objective optimization frameworks [Ehrgott, 2005] [Coello et al, 2007] [Zitzler et al, 2001]
- while other works have dealt with glyph-based visualizations, co-clustering, projected clustering, multi-task learning, etc.

Scalability & Data Dimensionality

SoA Method A

Graph Embedding framework for dimensionality reduction

- Dimensionality reduction guided by special affinity matrices W.
- Points that are connected in the affinity matrix are close to each other in the reduced space.
- Three types:
  - Direct: $\mathbf{Y} = \arg \min_{\mathbf{Y}^T \mathbf{B} \mathbf{Y} = \mathbf{I}} \sum_{i \neq j} \left\| \mathbf{y}_i - \mathbf{y}_j \right\|^2 W_{ij}$
  - Linearization: $\mathbf{V} = \arg \min_{\mathbf{V}^T \mathbf{X} \mathbf{B} \mathbf{X}^T \mathbf{V} = \mathbf{I}} \sum_{i \neq j} \left\| \mathbf{V}^T \mathbf{x}_i - \mathbf{V}^T \mathbf{x}_j \right\|^2 W_{ij}$
  - Kernelization: $\mathbf{A} = \arg \min_{\mathbf{A}^T \mathbf{K} \mathbf{B} \mathbf{K} \mathbf{A} = \mathbf{I}} \sum_{i \neq j} \left\| \mathbf{A}^T \mathbf{k}_i - \mathbf{A}^T \mathbf{k}_j \right\|^2 W_{ij}$
- Choosing appropriate matrices W, various dimensionality reduction methods can be described by the framework: PCA, LDA, ISOMAP, LLE, LDE, Laplacian Eigenmaps, LPP, etc.

[S. Yan et al. "Graph embedding and extensions: a general framework for dimensionality reduction," 2007]

Scalability & Data Dimensionality

SoA Method B

Multiple Kernel Learning Dimensionality Reduction (MKL-DR)

- Multimodal data are described by multiple kernel matrices $\mathbf{K}_1, \mathbf{K}_2, \ldots, \mathbf{K}_M$
- Modality weights $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_M)$ are introduced.
- A multimodal kernel matrix is formed, using the modality weights:

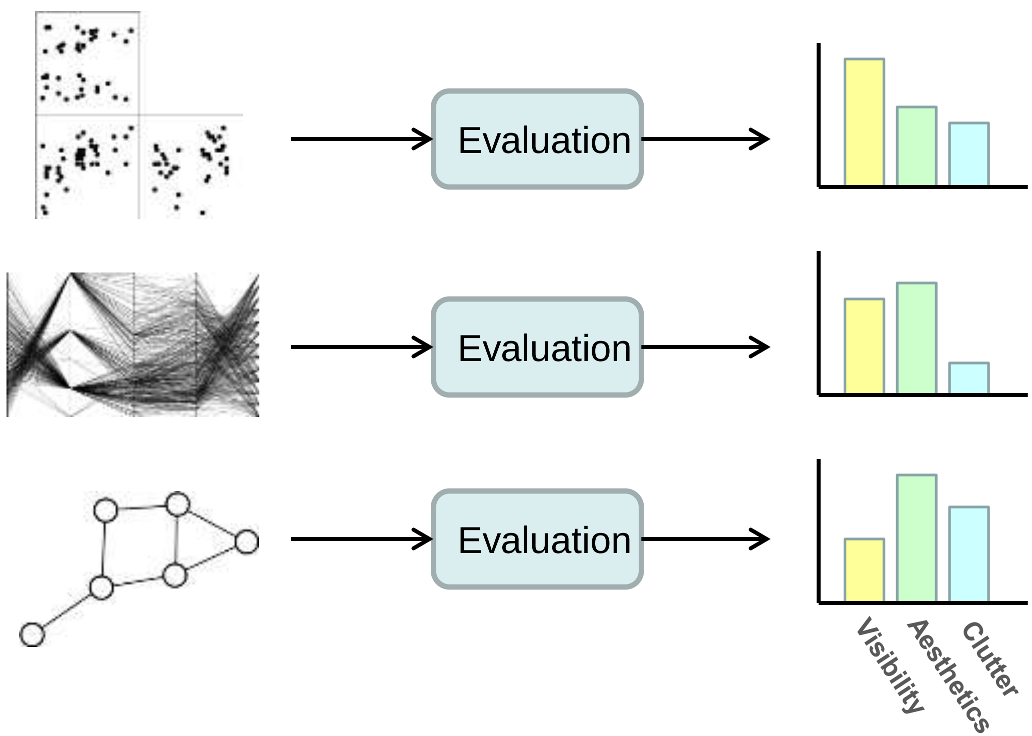
$$\mathbf{K}_{\boldsymbol{\beta}} = \sum_{m=1}^{M} \beta_m \mathbf{K}_m$$

- The multimodal kernel is used with affinity matrices of the Graph Embedding framework, for multimodal dimensionality reduction.
- The output points are calculated as $\mathbf{Y} = \mathbf{A}^T \mathbf{K}_{\boldsymbol{\beta}}$.
- The mapping coefficients A and the modality weights β are calculated through an alternating optimization procedure.

[Y.-Y. Lin, et al. "Multiple kernel learning for dimensionality reduction," 2011]

Evaluation
Definition

… is the process of defining and using quantitative metrics which are able to computationally evaluate a visualization method in terms of information visibility, aesthetics, clutter, etc., aiming at the quantitative comparison of visualization approaches.



Evaluation

Selected publications in Evaluation problems

- Need for evaluation metrics
  - [Tufte and Graves-Morris, 1983][Miller et al., 1997][Chen, 2005]
- Taxonomy of visualization evaluation metrics
  - [Bertini et al., 2011]
- Metrics for specific types of visualizations
  - Scatterplots [Bertini and Santucci, 2004][Urribarri and Castro, 2016]
  - Parallel coordinates [Dasgupta and Kosara, 2010]
  - Graph aesthetic measures [Ware et al., 2002][Dunne et al., 2015]
- Use of perceptual models for metrics definition
  - Perceptual visual quality metrics for images [Lin and Kuo, 2011]
  - Use of computational vision models [Pineo and Ware, 2012]

Evaluation

SoA Method A (1/2)

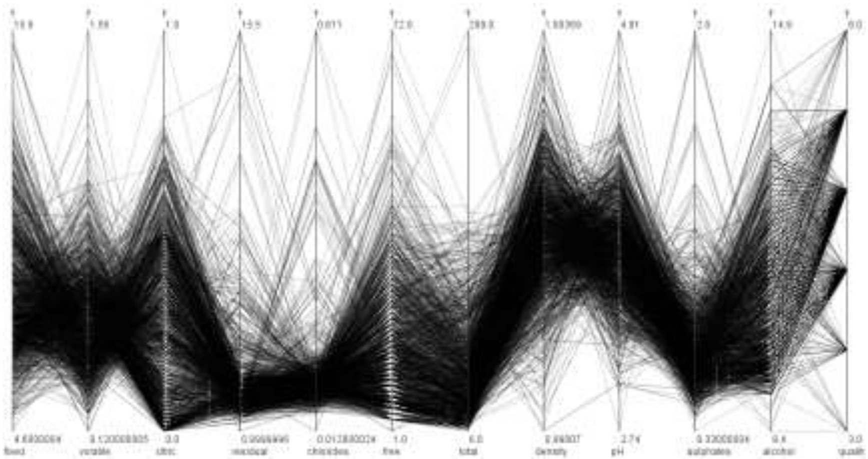
- **Quality metrics** have been proposed for evaluating the effectiveness of visualization.
- Such an example is **Pargnostics**, for the optimization of parallel coordinates visualization
- Quality metrics of Pargnostics:
  - Number of Line Crossings
  - Angles of Crossing
  - Over-plotting $$O = \sum_{i=1}^{h}\sum_{j=1}^{h}\begin{cases} b_{ij} & if\ b_{ij} > 1 \\ 0 & otherwise \end{cases},\ where\ b_{ij}\ a\ bin of\ a\ 2D\ histogram$$
  - Mutual Information $$I = \sum_{i=1}^{h}\sum_{j=1}^{h} p(x_i, y_j)\log\frac{p(x_i, y_j)}{p(x_i)p(y_j)},\ where\ p(x_i) = \frac{b_i}{h},\ and\ p = \frac{b_{ij}}{h}$$
  - Pixel-based entropy $$H = -\sum_{i=0}^{255}\frac{x_i}{n_{pixels}}\log\left(\frac{x_i}{n_{pixels}}\right),\ where\ x_i\ the\ gray\ value\ of\ pixel\ i$$

[D. Aritra et al."Pargnostics: Screen-space metrics for parallel coordinates.“ 2010]

Evaluation

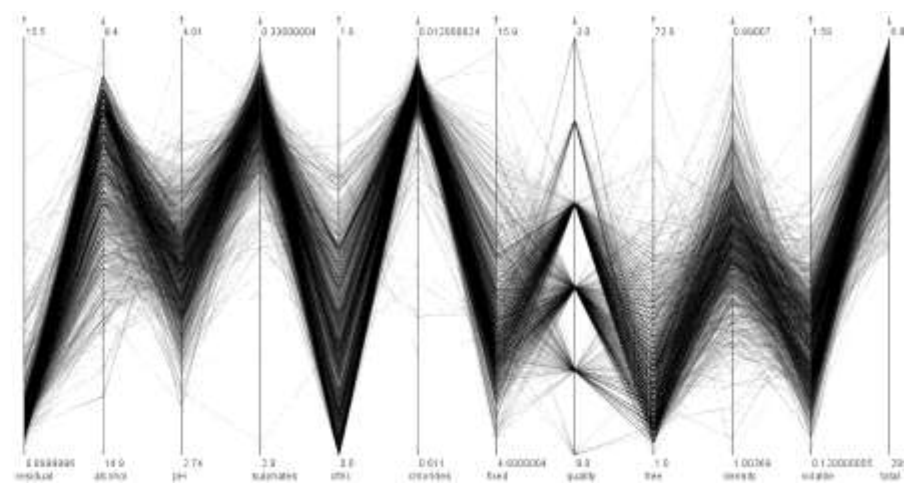
SoA Method A (2/2)

- Problem: Change the **ordering** and/or direction of **parallel coordinates** in order to **maximize/minimize** one or multiple **quality metrics**



Initial layout of the wine dataset

Maximized number of crossings and minimized angles of crossing, including inversions.



[D. Aritra et al."Pargnostics: Screen-space metrics for parallel coordinates.“ 2010]

Evaluation

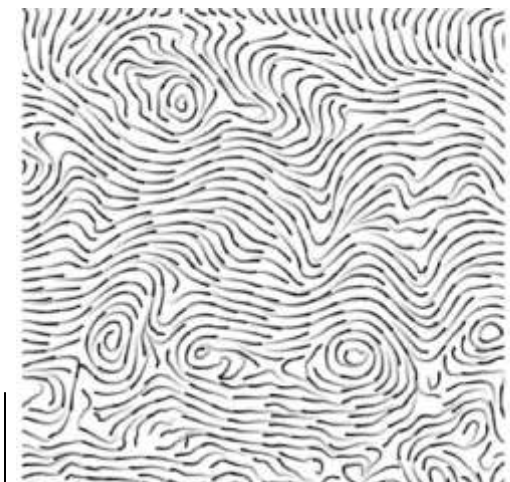
SoA Method B

- *The **Eye Perception Model** is used to generate the most efficient flow visualization*

- Definition of edge detection method based on an eye retina model:

Minimize the following **evaluation metric *O*** computed at different image scales *s*:

$$\boxed{O = \sum_{s} \sum_{i,j} \vec{O}^{'}_{i,j} \frac{Actual_{i,j}}{\left|Actual_{i,j}\right|}}$$



where $\vec{O}^{'}_{i,j}$ is the perceived orientation, and $Actual_{i,j}$ is the actual one.

$$\vec{O}^{'}_{i,j} = \sum_{x,y,\theta} G_{x,y} \begin{bmatrix} V1_{i,j,\theta} \cos(2\theta) \\ V1_{i,j,\theta} \sin(2\theta) \end{bmatrix}, \textit{where } \; V1_{i,j,\theta} = \left| \sum_{x,y} Gabor_{x,y,\theta} R^{w-b}_{i+x,j+y} \right|$$

where $Gabor_{x,y,\theta}$ is a Gabor filter at point (x,y) with angle θ, and $R^{w-b}_{i+x,j+y}$ is the retinal response in the white-black channel.

[P. Daniel et al. "Data visualization optimization via computational modeling of perception."2012]

Presentation outline

1. Introduction
   - Big Data
   - Motivation - Big Data Landscape
   - Visual analytics for Big Data
     - Definition
     - Visualization Taxonomy
     - Visual Analytics Challenges & SoA
     - Visual Analytics Application Fields
2. Visual analytics methods by CERTH/ITI
3. Videos demonstration

Visual Analytics application fields

- Physics and Astronomy
- Business
- Environmental monitoring
- Disaster and Emergency Management
- Software analytics
- Engineering Analytics
- Personal Information Management
- **(Network) Security**
- **Traffic monitoring**
- **Biology, Medicine, and Health**
- **Energy**
- **Accessibility**

CERTH/ITI Fields of Research

Visual Analytics applications

Physics and Astronomy / Business

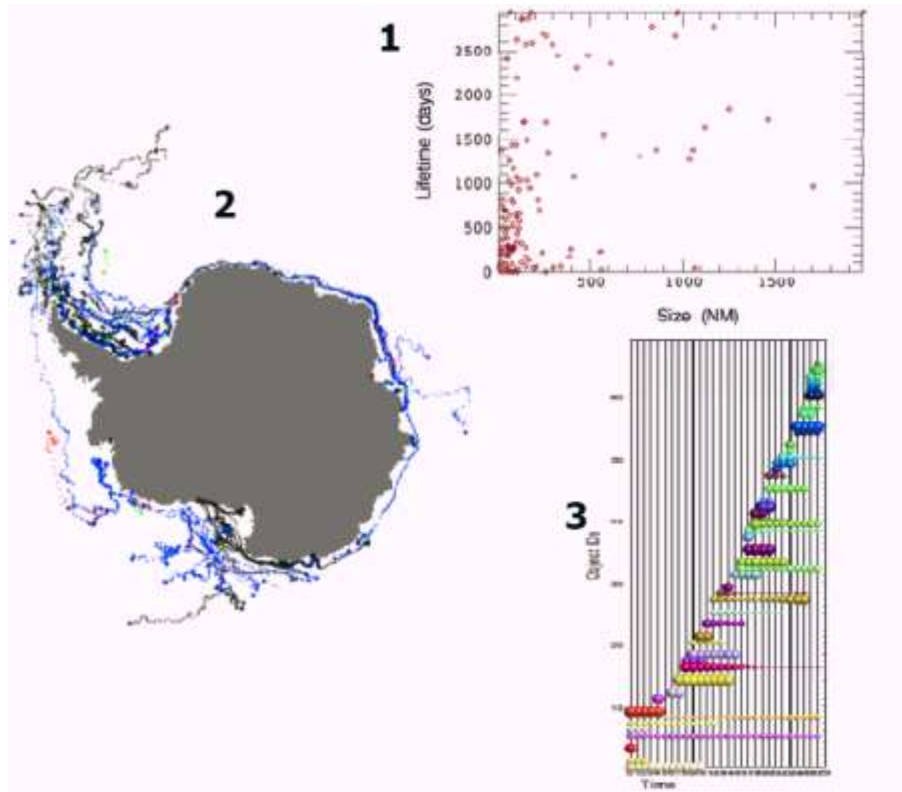
- **Physics and Astronomy:**
  - Flow visualization,
  - Fluid dynamics,
  - Molecular dynamics,
  - Nuclear science
- **Business:**
  - **Understanding** historical and current situations
  - **Predicting** future market trends
  - Need for real-time **monitoring** of the market, which would support the decision making of the users

Visual comparison of the financial market for all assets in 2 countries and 7 market sectors from 01/2006 and 04/2009.

Visual Analytics applications

Environmental monitoring /Disaster & Emergency Management

- **Environmental monitoring**
  - **Measuring** the climate change
  - **Forecasting** the weather
  - **Evaluating** the effects of carbon emission in the atmosphere
- **Disaster & Emergency Management**
  - **Evaluate** the situation
  - **Monitor** the ongoing progress of the emergency
  - **Provide** the people in charge with clues of the kind of immediate action needed
  - Visual Analytics can also help to **prevent** such emergencies



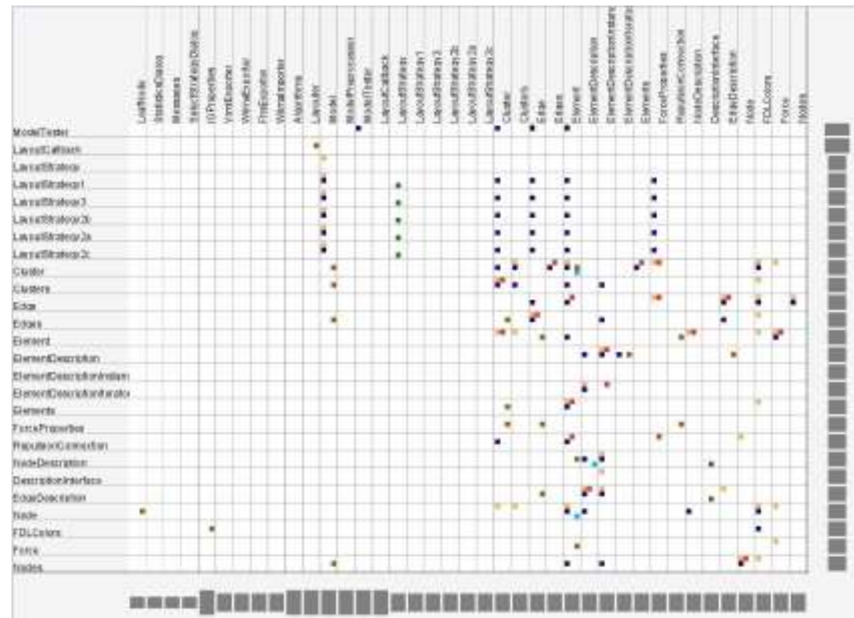
Multiple view environment for visual exploration of iceberg tracks

[Ulanbek et al. "Visual analytics to explore iceberg movement", 2008]

Visual Analytics applications

Software analytics / Engineering Analytics

- Software analytics:
  - **Debug** code
  - **Maintain** code
  - **Restructure** code
  - **Optimize** code
- Engineering Analytics
  - 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



Matrix visualization of relationships between different classes

[Zeckzer et al. "Visualizing software entities using a matrix layout.“, 2010].

Visual Analytics applications

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** and their **root causes (**through the appropriate **hypothesis tests**)

Treemap visualization of the spread of botnet computers in China in August 2006

[Mansmann et al. “Visual Analysis of Network Traffic for Resource Planning, Interactive Monitoring, and Interpretation of Security Threats”, 2007]

Visual Analytics applications

Traffic Monitoring

- 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, to identify normal or abnormal patterns of network traffic and to predict imminent states of the network.

An accident risk map of passenger vessels (turquoise), cargo vessels (orange), and tanker vessels (green) in front of Rotterdam harbor.

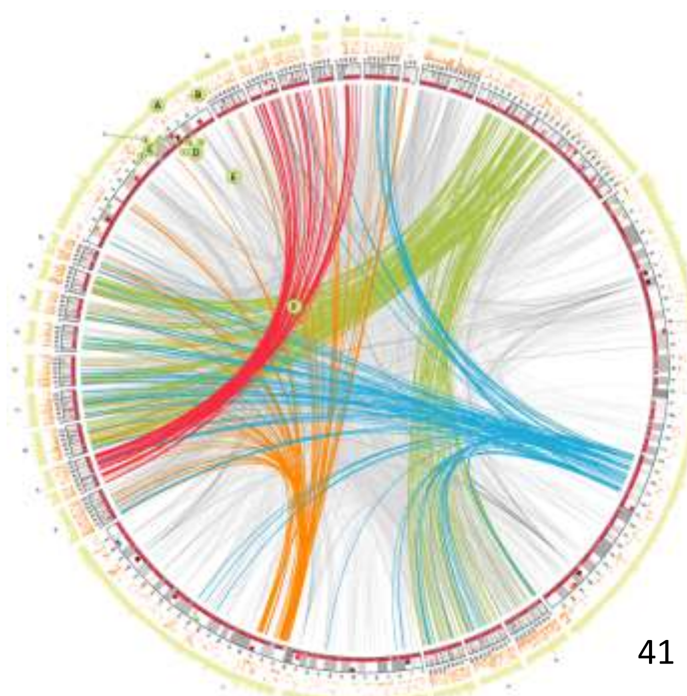
[Scheepens et al. "Composite density maps for multivariate trajectories",2011]

Visual Analytics applications

Biology, Medicine, and Health

- Bio-informatics:
  - Proteomics: Studies of the proteins in a cell
  - Metabolomics: Systematic study of unique chemical fingerprints that specific cellular processes leave behind
  - Combinatorial Chemistry: chemical synthetic methods that make it possible to prepare a large number of compounds in a single process.

- Example data:
  - Human Genome Project, which stores 3 billion base pairs per human



"Circos" visualization the similarities between different genomes

Presentation outline

1. Introduction
   - Big Data
   - Motivation - Big Data Landscape
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - Method 1: Multimodal Minimum Spanning Tree
   - Method 2: Multimodal graph embedding
   - Method 3: Visualization based on multiple criteria optimization
   - Method 4: K-partite graph for the visualization of multidimensional data
   - Method 5: Visualization of streaming in the network using state change graphs
   - …
3. Videos demonstration

Method 1: **Multimodal Minimum Spanning Tree**

Method name:

•*Multimodal Minimum Spanning Tree* for multimodal data visualization.

Research field:

•Multimodal search engines, biomedical research.

Big data issues addressed:

•Variety.

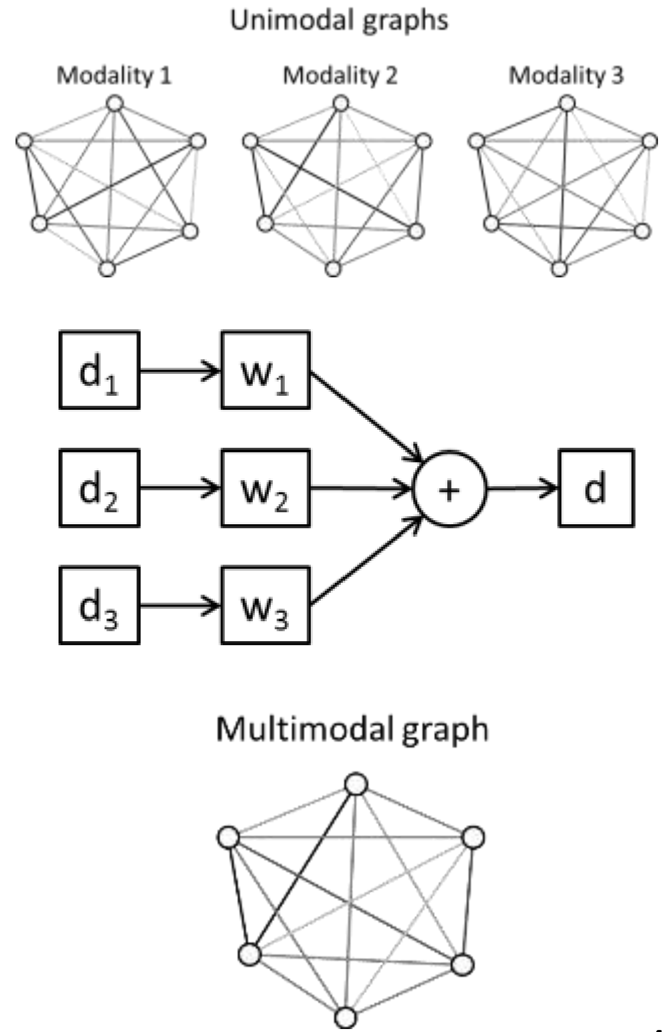
Application areas:

•Visual exploration of multimedia search engine results by end users.

•Visually assisted analysis of biomedical data for biology and medicine researchers and analysts.

Method 1: **Multimodal Minimum Spanning Tree 1/3**

- *1^st^ Step*: Calculation of unimodal distances d_i among the multimodal objects.
- *2^nd^ Step*: Construction of unimodal graphs.
- *3^rd^ Step*: Calculation of **multimodal distances** d as weighted sums of unimodal distances.
  - Modality weights are determined through user interaction.
  - The user selects two objects and the weight of the modality for which the objects are most similar is increased.
- *4^th^ Step*: Construction of **multimodal distance graph**.

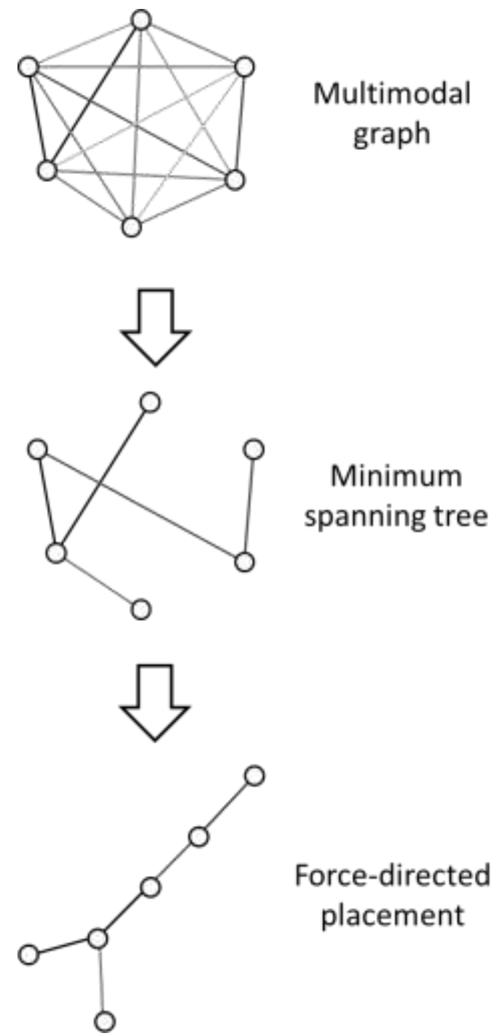


Method 1: **Multimodal Minimum Spanning Tree 2/3**

The multimodal graph is used to visualize the data.

Approach 1:

- *5th Step*: Calculation of the **minimum spanning tree (MST)** for the reduction of the data volume.
  - The MST connects the data that are most similar, with a minimum number of edges.
- *6th Step*: **Force-directed placement** of the MST for embedding in low-dimensional space and for visualization.
  - Vertices are considered as repelling charges, edges as attractive springs.



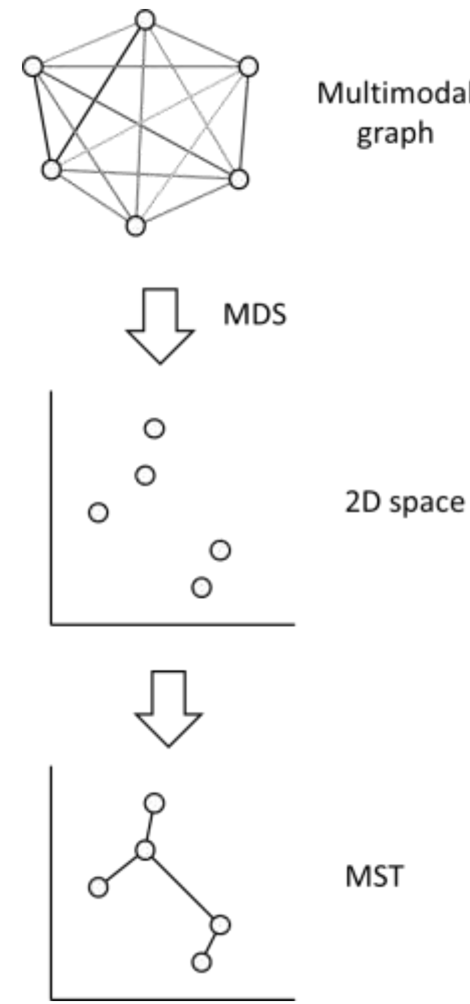
Method 1: **Multimodal Minimum Spanning Tree 3/3**

The multimodal graph is used to visualize the data.

Approach 2:

- *5th Step*: Embedding of the multimodal graph in 2D space, using **Multidimensional Scaling** (MDS).

- *6th Step*: Visualization of the **Minimum Spanning Tree** of the multimodal graph on the 2D space.
  - The MST connects the data that are most similar, with a minimum number of edges.



Application 1: **Visualization of multimodal data for multimedia search engines**

- **Scope/Problem Definition**:
  - Visualization of similarities between different objects
- **Dataset**:
  - Custom multimodal objects of animals consisting of images and sounds.
  - http://160.40.50.78/image-sound-dataset/image_sound_dataset_animals.rar
- **Application**:
  - A multimodal graph is constructed and the Force-Directed MST is presented to the user.

The user selects two objects that should be closer.

The system adjusts the modality weights according to the feedback.

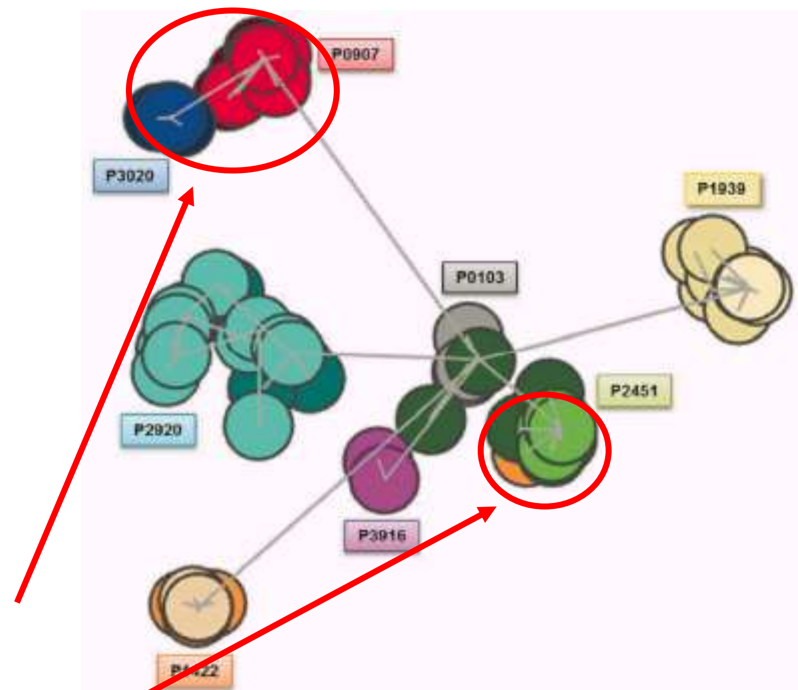
Application 2: **Visualization and analysis of DNA sequences 1/2**

- **Scope/Problem definition**:
  - Identify clusters of **similar sequences**
  - Identify clusters of **similar patients**
  - Identify **mutation paths** and cluster changes over time

- **Dataset:**
  - CLL dataset collected by CERTH/INAB
  - **DNA Sequences** from B-cell receptor immunoglobulins, taken from patients with Chronic Lymphocytic Leukemia (CLL)
    - 781 sequences from 8 patients
    - Data taken in multiple time instances, and from multiple cells of the same patient
    - Each sequence is represented as a string of characters on amino-acid level (21 different characters) and nucleotide level (4 different characters)

Application 2: **Visualization and analysis of DNA sequences 2/2**

- **Application:**
  - **Distance calculation** between all the sequences at different levels, using string distance metrics.
  - **Projection** of the sequences to the 2D plane:
    - Each node represents a unique sequence.
    - Similar sequences are positioned in close proximity on the 2D plane.
  - **Minimum Spanning Tree**
    - Identification of mutation paths.
  - **Results:**
    - Some users have similar disease mutations and are clustered
    - The mutation path of some users (e.g. P1422) terminated in another cluster



Different colors represent different patients. The color intensity represents sequences taken from the same patient at different time instances

Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - Method 1: Multimodal Minimum Spanning Tree
   - Method 2: Multimodal graph embedding
   - Method 3: Visualization based on multiple criteria optimization
   - Method 4: K-partite graph for the visualization of multidimensional data
   - Method 5: Visualization of streaming in the network using state change graphs
   - …
3. Videos demonstration

Method 2: **Multimodal Graph Embedding**

Method name:

• *Multimodal Graph Embedding (MGE)* for dimensionality reduction.

Research field:

• Multimodal search engines, network security.

Big data issues addressed:

• Variety and Volume.

Application areas:

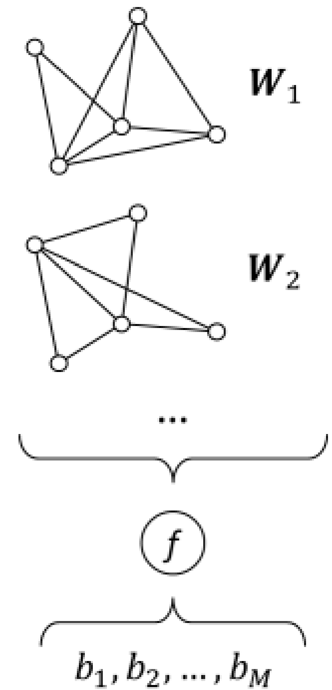
•Visual exploration of large multimedia databases by search engine users.

•Visually assisted analysis of network data by network analysts for threat identification.

Information Technologies Institute

Method 2: **Multimodal Graph Embedding 1/3**

Goal: Construction of a multimodal adjacency graph as a weighted sum of multiple unimodal ones and embedding the multimodal graph on a low-dimensional space.



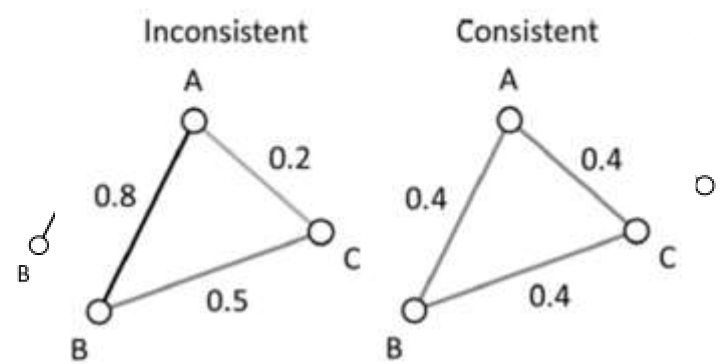
Procedure:

- *1st Step*: Construction of M unimodal affinity matrices $\mathbf{W}_m$.
- *2nd Step*: Automatic calculation of optimal modality weights $\mathbf{b} = (b_1, b_2, \ldots, b_M)$, by solving the optimization problem: $\mathbf{b}_{\text{opt}} = \arg\min_{\mathbf{b}} f(\mathbf{b})$

Graph consistency objective function:

$$f(\mathbf{b}) = \sum_{\{i,j\}\in E^*} \sum_{k=1}^{N} \left(\mathbf{b}^T \mathbf{w}_{ik} - \mathbf{b}^T \mathbf{w}_{jk}\right)^2$$

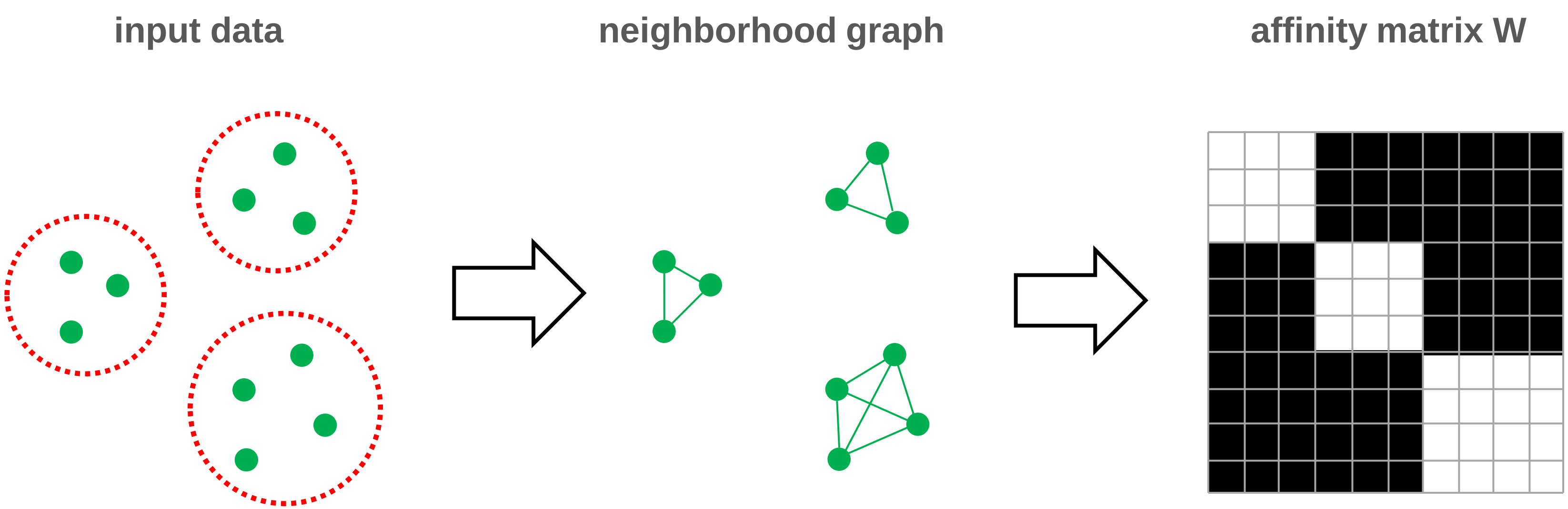
$$\mathbf{w}_{ij} = \left(\mathbf{W}_1(i,j), \mathbf{W}_2(i,j), \ldots, \mathbf{W}_M(i,j)\right)^T$$



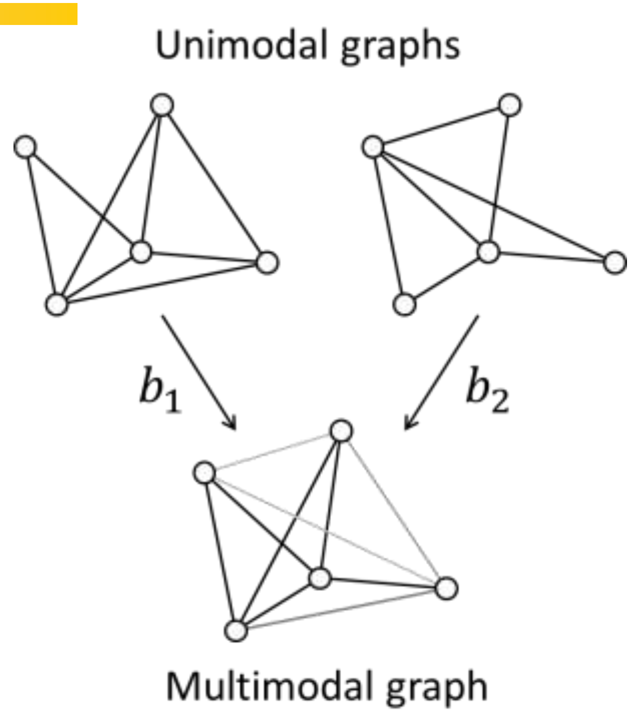
Method 2: **Multimodal Graph Embedding 2/3**

- *3rd Step*: Construction of a **multimodal affinity matrix** W, as a weighted sum of the unimodal matrices, using the optimal modality weights.

- What is the target of neighborhood graph fusion?
  - Data are assumed to be organized in semantic classes.
  - Thus, the ideal affinity matrix would be block-diagonal.



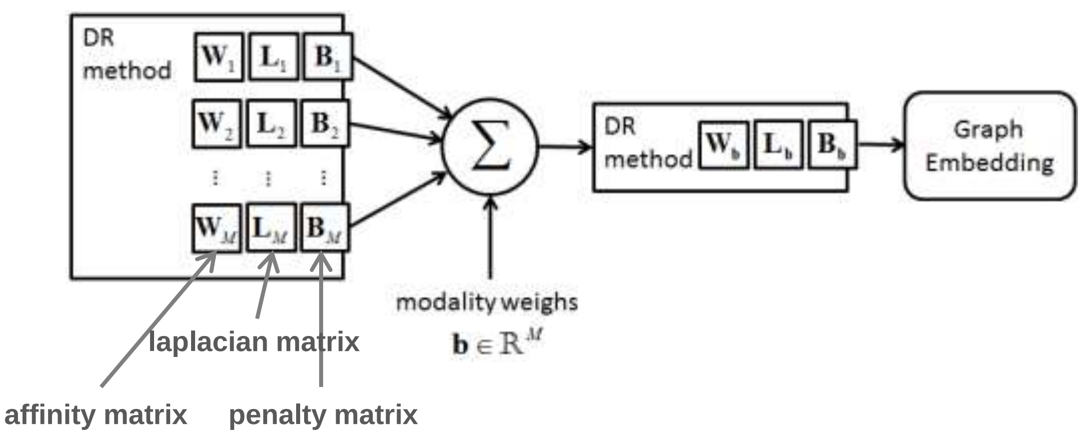
Method 2: **Multimodal Graph Embedding 3/3**

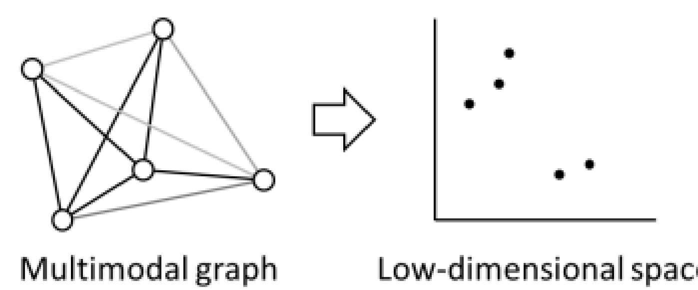


- *4th Step*: State-of-the-art dimensionality reduction methods are used to embed the multimodal graph in a **low-dimensional space**.

$$\mathbf{W} = \sum_{m=1}^{M} b_m \mathbf{W}_m$$

- *5th Step*: The output space can be used for classification, clustering, visualization.





Application 1: Clustering performance in large multimodal image dataset 1/2

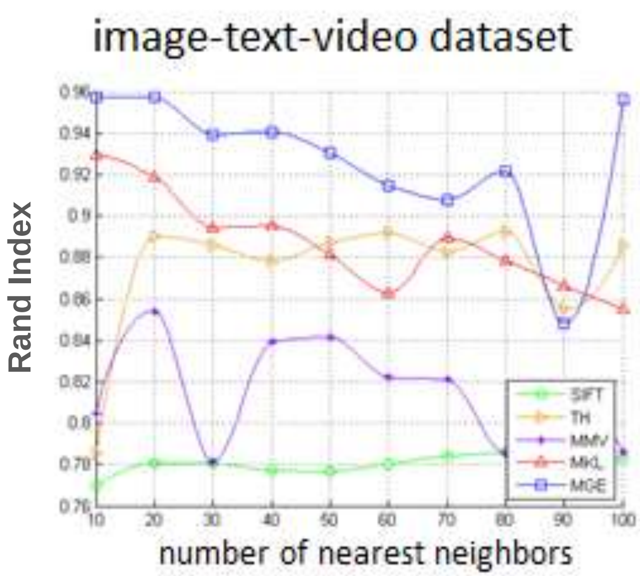
- **Scope/Problem definition**
  - Group semantically similar multimodal objects together

- **Dataset**:
  - Caltech-101 image dataset
  - [L. Fei-Fei et al. "Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories," 2007]
  - Images described by multiple features: SIFT, PHOG, GIST, Geometric Blur

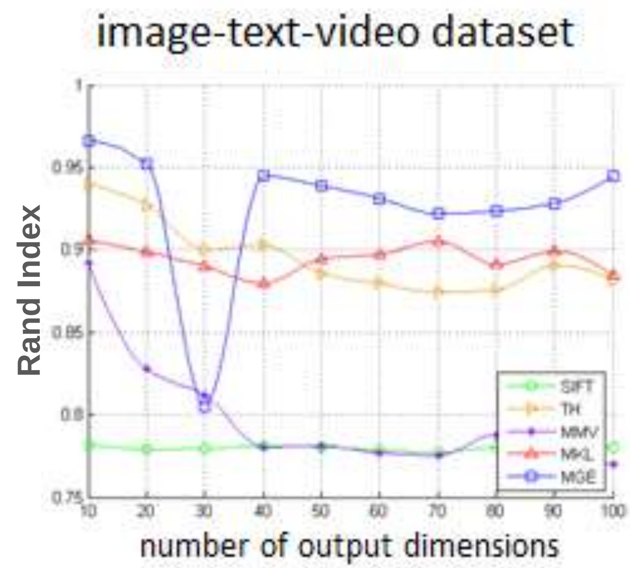
- **Application**:
  - Clustering.
  - Clustering performance measured with the Rand Index, using the ground truth class labels.

Application 1: **Clustering performance in large multimodal image dataset 2/2**

- **Application** (cont.):
  - Comparison with Multiple Kernel Learning dimensionality reduction (MKL-DR)
  - [Y.-Y. Lin, et al. “Multiple kernel learning for dimensionality reduction,” 2011]



The MGE method achieves higher clustering accuracy than SoA methods, for a varying number of nearest neighbors considered for the affinity matrices.

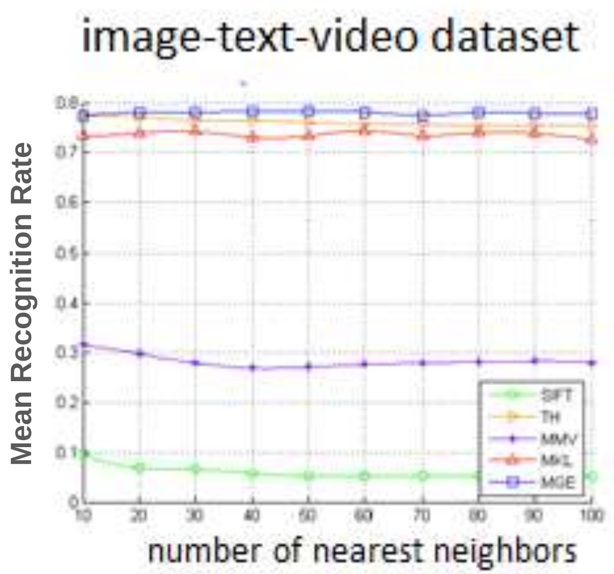


$$Rand\ Index\ (RI) = \frac{TP + TN}{TP + TN + FN + FP}$$

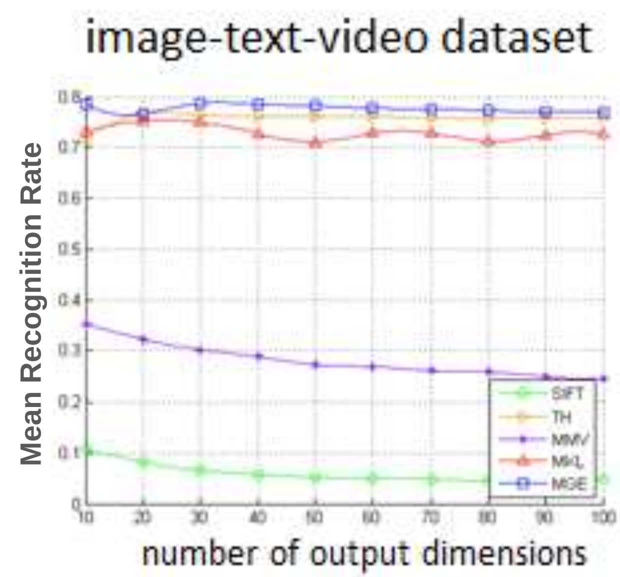
The MGE method achieves higher clustering accuracy than SoA methods, for varying dimensionality of the output space.

Application 2: **Object Classification performance in large multimodal dataset**

- **Application** (cont.):
  - Comparison with Multiple Kernel Learning dimensionality reduction (MKL-DR)
  - [Y.-Y. Lin, et al. “Multiple kernel learning for dimensionality reduction,” 2011]



The MGE method achieves higher classification performance than SoA methods, for a varying number of nearest neighbors considered for the affinity matrices.



The MGE method achieves higher classification performance than SoA methods, for varying dimensionality of the output space.

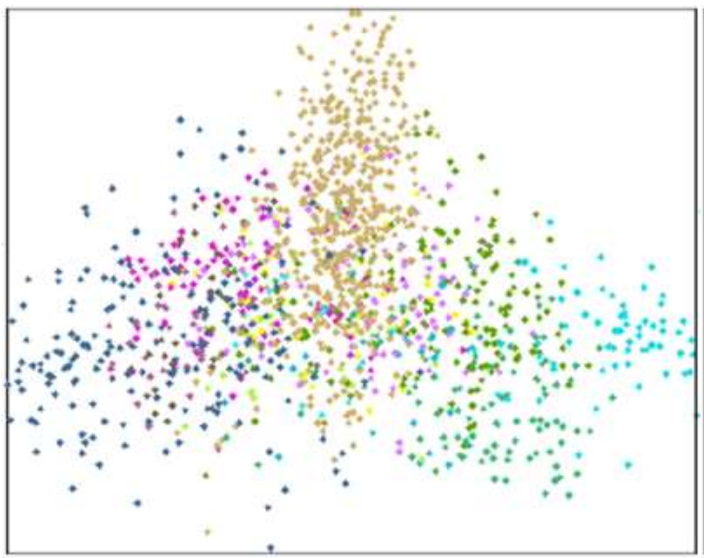
Application 3: **Visualization of large multimodal dataset 1/2**

- **Scope/Problem definition:**
  - Visualization of multimodal objects so that semantically similar ones are close to each other.

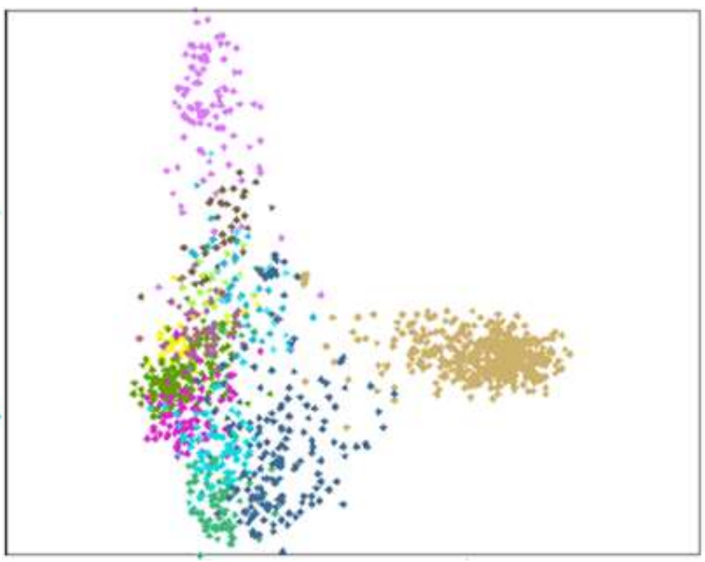
- **Dataset**:
  - EVVE video event dataset
  - [J. Revaud et al. “Event retrieval in large video collections with circulant temporal encoding”, 2013]
  - multimodal objects consisting of multiple media items
    - images
    - text
    - videos

Application 3: **Visualization of large multimodal dataset 2/2**

- **Application**:
  - Use of MGE method for dimensionality reduction to 30 dimensions.
  - Visualization by using Multidimensional Scaling to map the data to 2 dimensions.
  - Comparison with Multiple Kernel Learning dimensionality reduction.
    - [Y.-Y. Lin, et al. “Multiple kernel learning for dimensionality reduction,” 2011]



MKL-DR



MGE

Points represent multimodal objects.

Colors represent ground truth class labels.

The object classes are more apparent when using the MGE method, than when using the MKL-DR method.

Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 3: Visualization based on multiple criteria optimization
   - Method 4: K-partite graph for the visualization of multidimensional data
   - Method 5: Visualization of streaming in the network using state change graphs
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
3. Videos demonstration

Method 3: Visualization based on multiple criteria optimization

Method name:

- *Multi-objective visualization* for multimodal data visualization

Research field:

- Multimodal search engines, traffic monitoring.

Big data issues addressed:

- Variety and Volume.

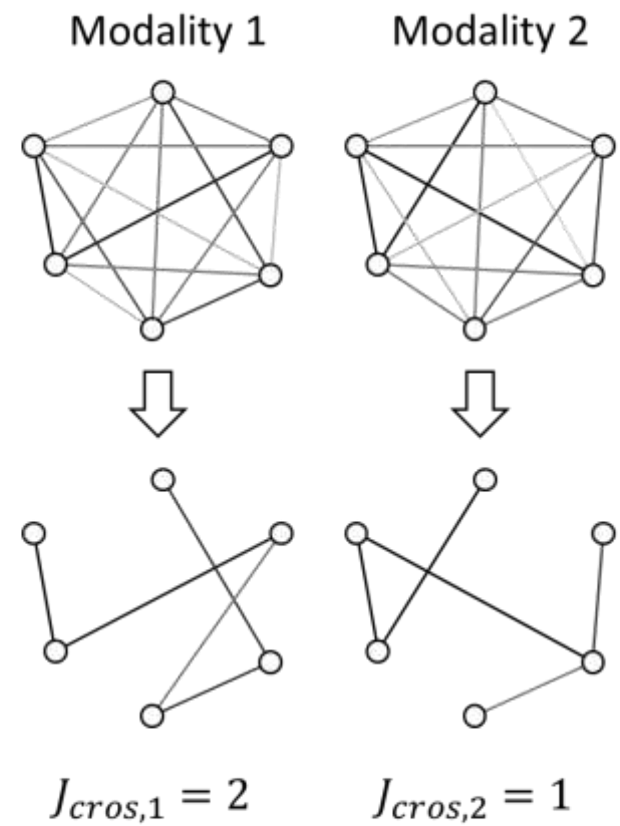
Application areas:

- Visual exploration of large multimedia databases by search engine users.
- Visually-assisted analysis of road traffic for traffic monitoring operators.

Method 3: **Visualization based on multiple criteria optimization 1/3**

Goal: The optimization of unimodal clustering objectives simultaneously for all modalities.

- *1st Step*: Unimodal graphs are constructed and **minimum spanning trees** are extracted.
- *2nd Step*: Unimodal visualization is formulated as an optimization problem, whose solution is the positioning of the data on the plane so that a proper objective function J_m of each unimodal graph is minimized.
- *3rd Step*: Various **graph aesthetic measures** are used as objective functions:
  - Number of edge crossings (minimize)
  - Average angle among neighboring edges (maximize)
  - Minimum potential energy of graph, if seen as a set of charges and springs (minimize)



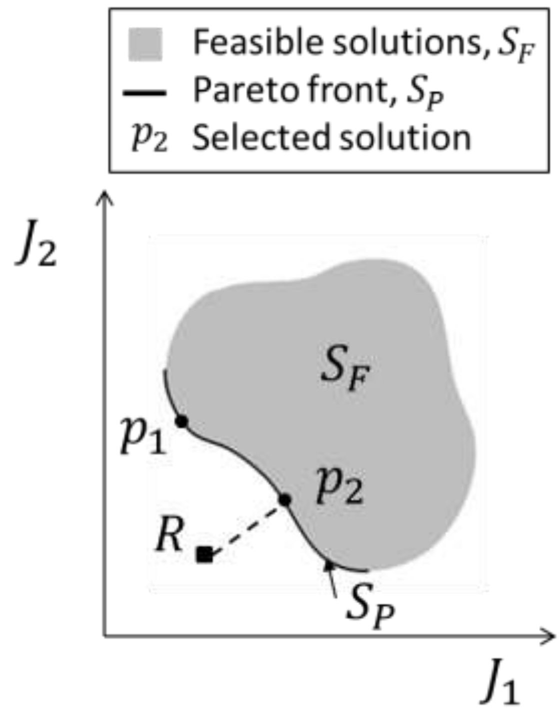
Method 3: **Visualization based on multiple criteria optimization 2/3**

- *4th Step*: **Multi-objective optimization**
  - Multiple modalities → **multiple objective functions** which need to be minimized simultaneously.

$$p_{\text{opt, multimodal}} = \arg\min_{p \in P} \mathbf{J}(p)$$

$$\mathbf{J}(p) = \left(J_1(p), J_2(p), \ldots, J_M(p)\right)$$

  - Multi-objective optimization → **Pareto front** of multiple optimal solutions.
  - Significant reduction of the full feasible solution domain S_F to the much smaller domain of the Pareto-optimal solutions S_P, $S_P << S_F$.

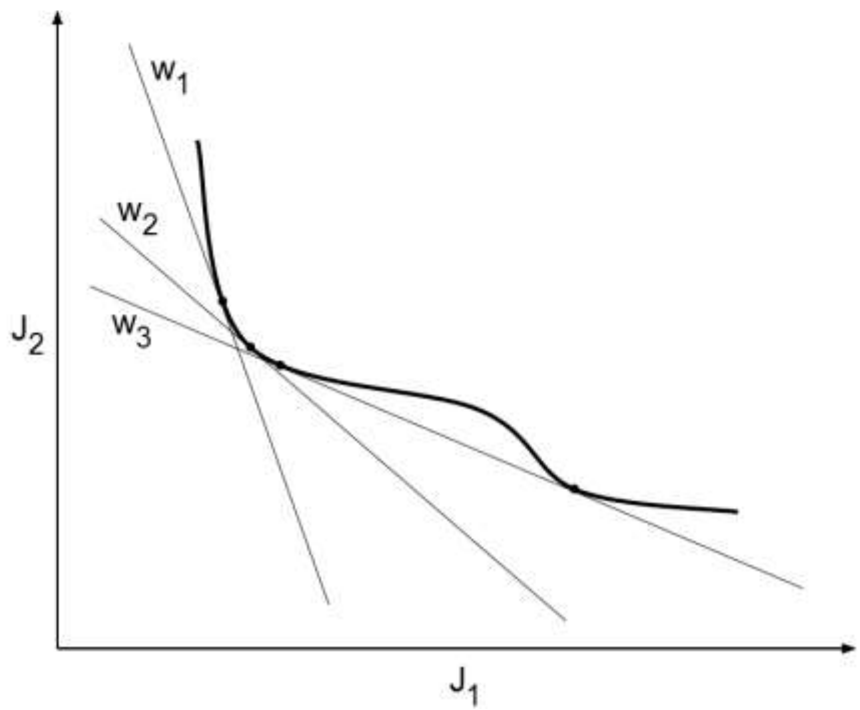


Results:

- Faster convergence to the optimal solution.
- Selection of the optimal solution based on the current user profile of preferences R (i.e. automatic or interactive function).

Method 3: **Visualization based on multiple criteria optimization 3/3**

- Why not combine the objectives in some manner?
  - Weighted-sum-based methods fail to discover solutions in the non-convex part of the Pareto front.

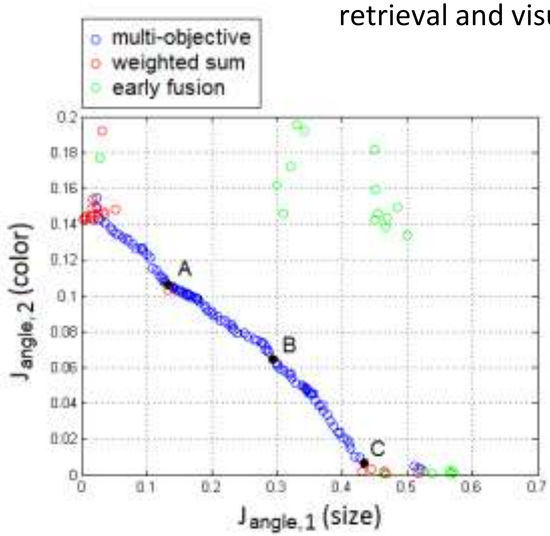


Application 1: Visualization performance in large multimodal datasets 1/2

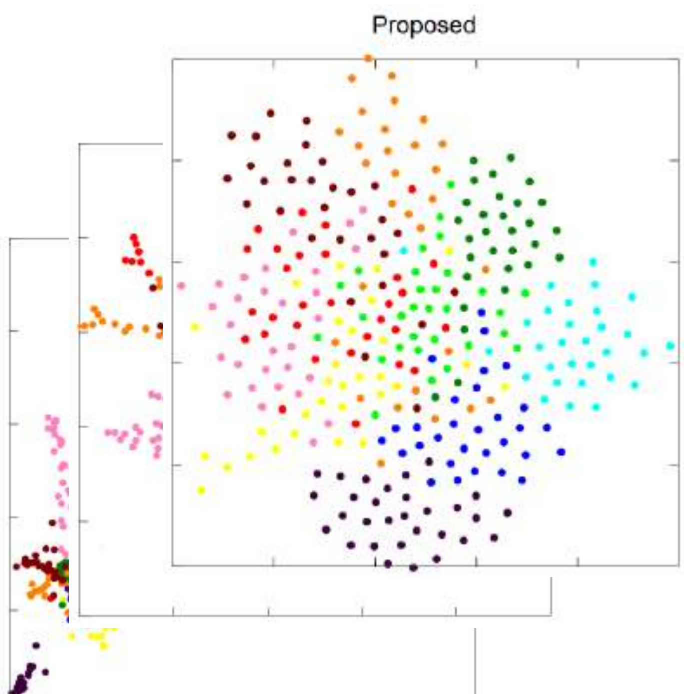
- **Scope/Problem definition**:
  - Interactive visualization and exploration of big datasets of multimodal objects, e.g. for multimedia search engines.
  - Clustering multimodal datasets, so that semantic entities are separated.
- **Dataset**:
  - Caltech-101 image dataset
  - [L. Fei-Fei et al. “Learning generative visual models from few training examples: An incremental Bayesian approach tested on 101 object categories,” 2007]
  - images described by multiple features: SIFT, PHOG, GIST, Geometric Blur
- **Application**:
  - Potential energy of minimum spanning tree as an objective function.
  - Optimization via multiple criteria → Pareto front of optimal solutions.
  - Selection of one of the solutions, based on user profile or interactively, and presentation of the visualization.

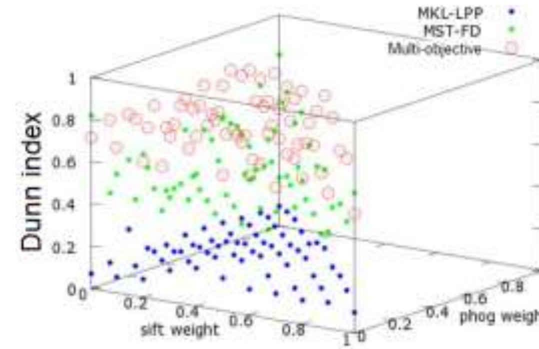
Application 1: **Visualization performance in large multimodal datasets 2/2**

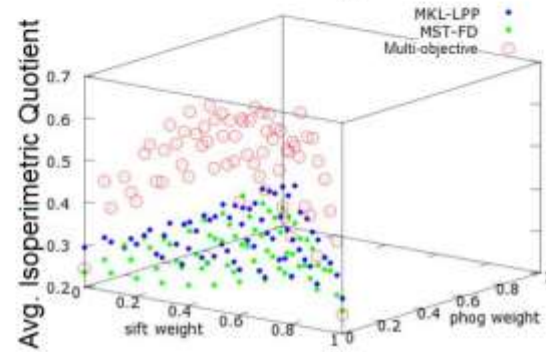
- **Application** (cont.):
  - Comparison via **Dunn Index** & **Avg. Isoperimetric Quotient** with:
    - MKL-DR: [Y.-Y. Lin, et al. “Multiple kernel learning for dimensionality reduction,” 2011]
    - MST-FD: [I. Kalamaras, et al. “A novel framework for multimodal retrieval and visualization of multimedia data”, 2012]



The multi-objective method manages to find solutions in the concave part of the Pareto front, which are not found by other methods.





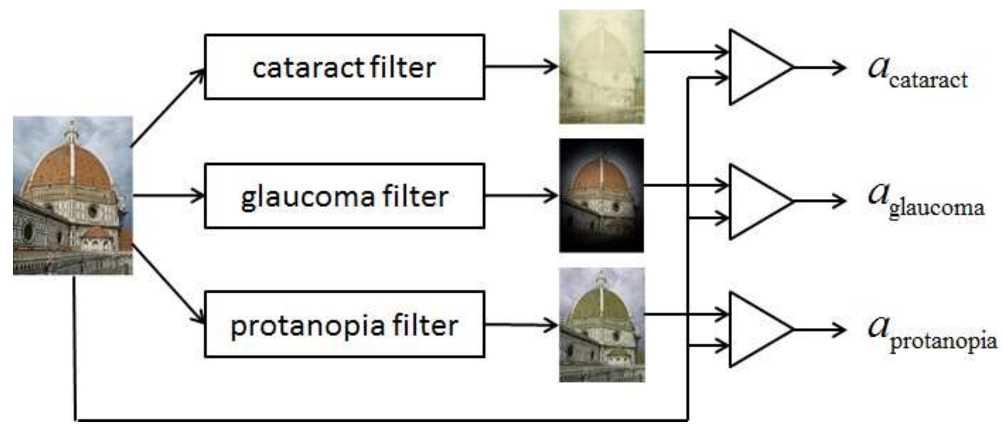


The multi-objective method achieves higher values for both the Dunn index and the AIQ measures, than the MKL-DR and the MST-FD methods, even for various modality weights.

Application 2: **Visualization & accessibility enhancements in search engine applications**

The image classes are apparent in the resulting clustering result.

Images are reranked by the search engine so that the accessible ones to visually-impaired users are promoted.



relevance-based ranking

accessibility-based multimodal ranking

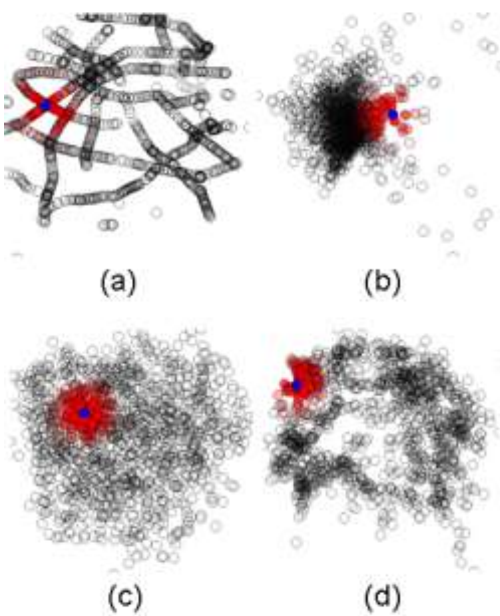
vision simulation

Application 2: **Road clustering for traffic prediction 1/2**

- **Scope/Problem definition**:
  - Visualization of road correlations, based on all available attributes.
  - Prediction of traffic in future time intervals, using the multiple attributes.
- **Dataset**:
  - Berlin roads datastet, from the e-COMPASS European project.
  - Road traffic data for a large number of road segments.
  - Multiple attributes available for each road segment:
    - Geographical position
    - Average vehicle speeds for five-minute time intervals.
    - Time series features extracted from the raw data.
- **Application**:
  - Multiple notions of distances between roads/streets (modalities):
    - Geographical distance
    - Time series (e.g. velocities) correlation
    - Time series phase difference
    - Time series difference estimated via dynamic time warping

Application 2: **Road clustering for traffic prediction 2/2**

- **Application** (cont.):
  - Mapping of inter-roads differences in the 2D space for **clustering**.
  - One optimization criterion/constraint per distance type.
  - Multiple criteria → Pareto front → custom selection of the solution.
  - The operator can select from the various Pareto solutions to view different aspects of traffic.

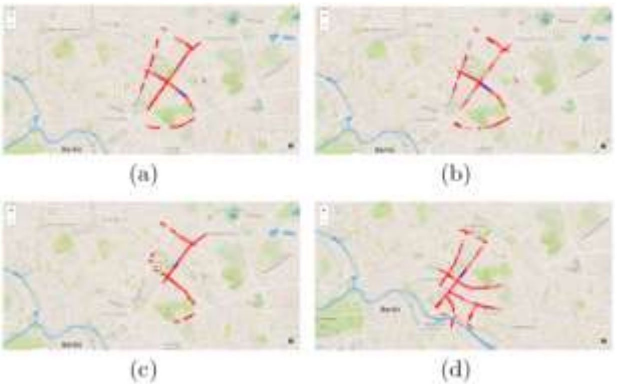


Points represent road segments.

Using the different notions of distances, various clusterings of the roads are produced.

The nearest neighbors of a selected road segment are other segments with similar properties.

Drawing the nearest neighbors of a road segment on the map, results in visualization of different aspects of traffic, for inspection by the analyst. These views are combined via the multi-objective method, for further analysis.



Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 4: K-partite graph for the visualization of multidimensional data
   - Method 5: Visualization of streaming in the network using state change graphs
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings’ energy sustainability
3. Videos demonstration

Method 4: k-partite graph for the visualization of multidimensional data 1/4

Method name:

- K-partite graph for Attack attribution on multi-dimentional datasets

Research field:

- Network Security

Big data issues addressed:

- Variety

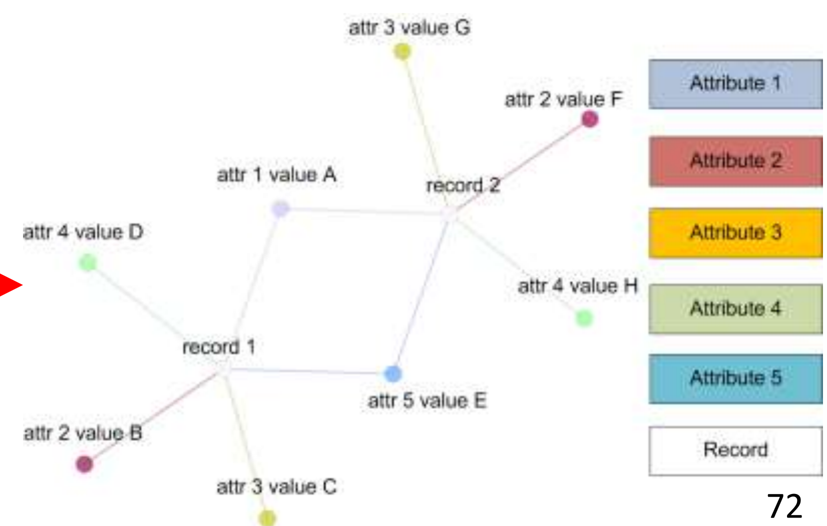
Application areas:

- Network operators

Method 4: **k-partite graph for the visualization of multidimensional data 2/4**

- *1st Step:* Creation of **k-partite** Graph
  - K-partite graph definition: nodes can be divided in k disjoint groups $(V_0, \ldots, V_{k-1})$ such that the graph $G = \langle V_0 \cup \ldots \cup V_{k-1}, E \rangle$ has edges in $E \subset \bigcup_{l=1}^{k-1} \{V_0 \times V_l\}$
  - Record → White vertex
  - Attribute → Colored Vertex
  - Edge →Relationships between various attributes and the corresponding records

Rec	Attr 1	Attr 2	Attr 3	Attr 4	Attr 5
1	value A	value B	value C	value D	value E
2	value A	Value F	value G	value H	value E

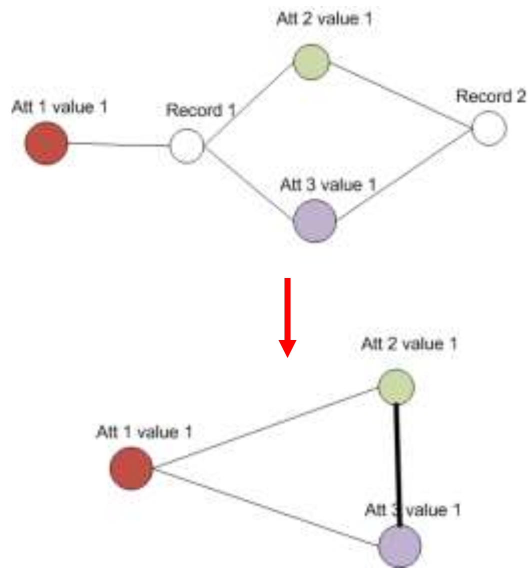


Method 4: **k-partite graph for the visualization of multidimensional data 3/4**

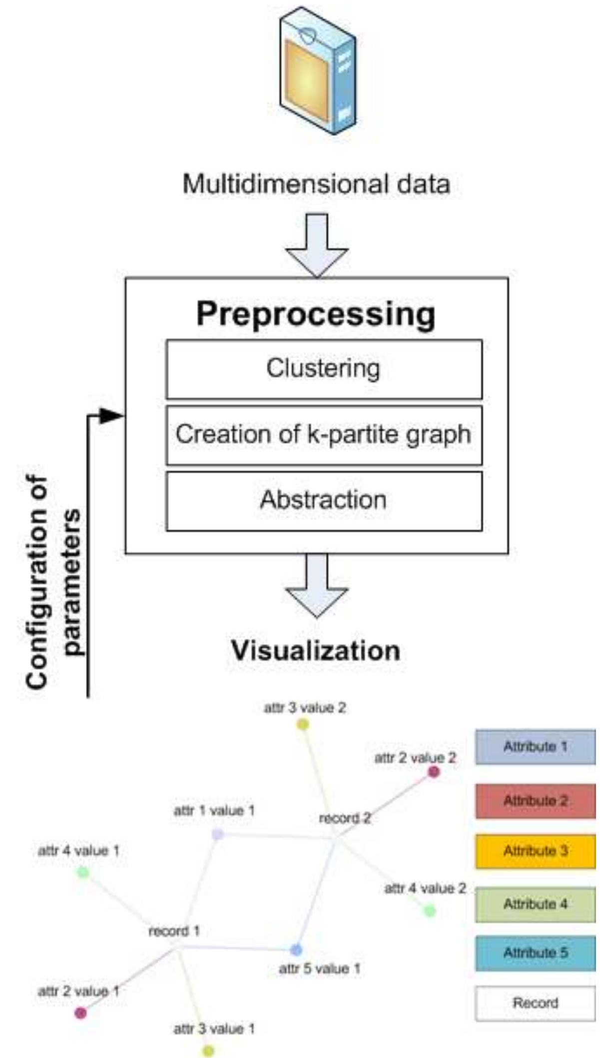
- *2[nd] Step:* **Reduction of the size** of the graph (abstraction)

K-partite graph

Graph Abstraction



- 3[rd] Step**: Clustering of similar vertices** (many common neighbors in the graph)
  - Random walks
- 4[th] Step: User interaction for parameter configuration



K-partite graph

Method 4: k-partite graph for the visualization of multidimensional data 4/4

3rd Step: **Clustering of similar vertices** through random walks

- Setting as ***P*** the transition matrix of the k-partite graph, perform the next three steps until convergence:
  - *Expansion* $\mathbf{C} = \mathbf{P} \cdot \mathbf{C}$ where ***C*** is an expansion matrix
  - *Inflation*, which raises each entry in the matrix ***C*** to the power *r* and then normalizes the rows to sum to 1:

$$C(i,j) = \frac{C(i,j)^r}{\sum_{k=1}^{N} C(i,k)^r}$$

  - *Prune*, which removes entries which have values below a threshold:

$$C(i,j) = \begin{cases} 0 & \text{, if } C(i,j) \leq q \cdot \max_{j=1}^{n}\{C(i,j)\} \\ C(i,j) & \text{, otherwise} \end{cases}$$

- Finally, the expansion matrix ***C*** holds the attractor nodes (clusters) for each node

Application: k-partite graph based attack attribution of malicious URLs 1/2

- **Scope/Problem definition:**
  - Perform **attack attribution**, i.e. identify which URLs where created by the same attacker by examining common attributes

- **Harmur Dataset:**
  - **Malicious URLs:** Contain malicious code (e.g. virus, trojans, etc.)
  - Example **attributes** collected for each malicious URLs
    - web servers, DNS information, geographical location of the servers (and hosting Autonomous System (AS))
    - Sample:

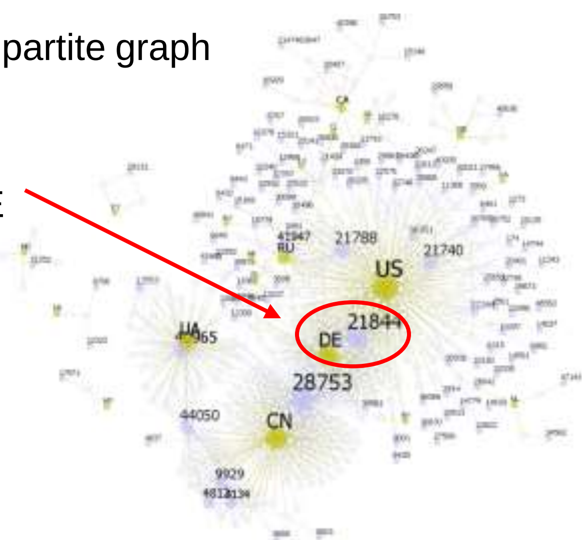
AS number	Location	Domain	Creation Date
24940	DE	pricelessfinish.cn	2009-03-04

Application: **k-partite graph for the analysis of data from malicious URLs 2/2**

- **Application**
  - **Malicious URLs** attributes selected:
    - AS number, and location of URL
  - K-partite graph based visualization
    - Visualization of correlations between different URLs
  - Clustering
    - Identification of URLs with common characteristics (**attack attribution**)

K-partite graph

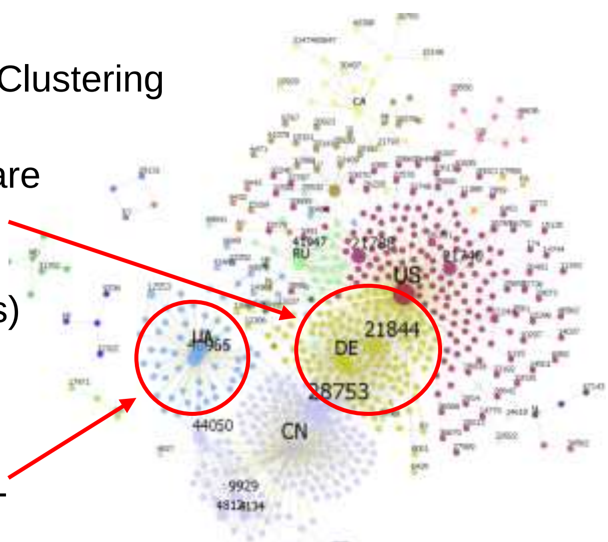
AS-21844 has many URLs in DE (positioned close)



Clustering

AS-21844 and DE are clustered together along with multiple URLs (yellow nodes)

Another cluster comprised of AS-40965 and UA



Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - **Method 5: Visualization of streaming in the network using state change graphs**
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings’ energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Method 5: **Visualization of streaming in the network using state change graphs**

Method name:

- *State change graphs* for attack detection and root cause analysis in networks

Research field:

- Security

Big data issues addressed:

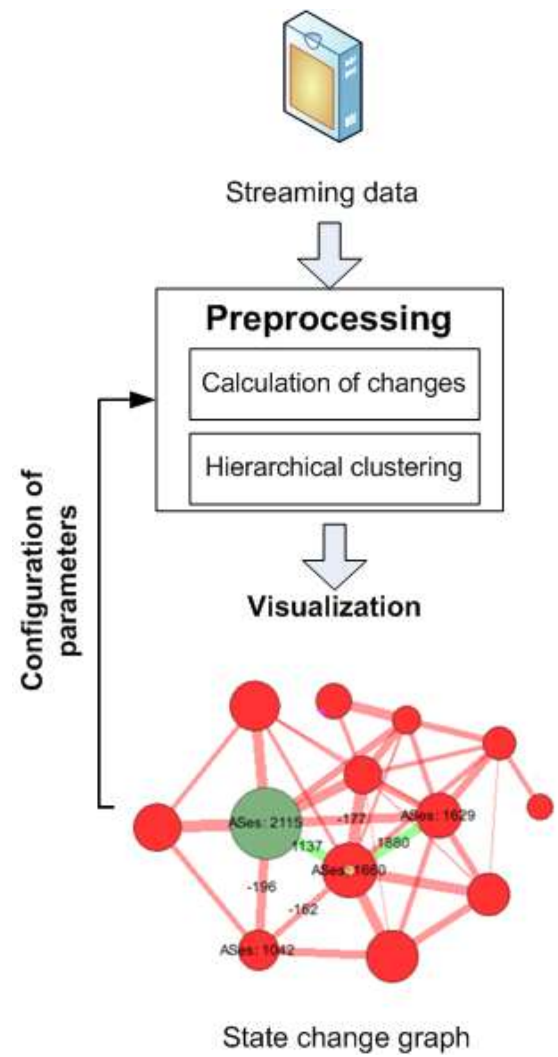
- Velocity

Application areas:

- Network operator

Method 5: **Visualization of streaming in the network using state change graphs 1/3**

- *1st Step: Streaming data*
  - Data that characterize the state of the network in each time instance (e.g. signaling in a mobile network)
- *2nd Step: Calculation of changes in the state of the network*
  - For specific time windows
  - For specific regions in the network
- *3rd Step:* ***State change graph***
  - State changes with respect to the previous time window (e.g. change of traffic in a network)
- *4th Step: Optimization of the graph visualization by* ***maximizing its entropy***
- *5th Step*: Hierarchical clustering for the reduction of the size of the graph
- *6th Step:* User interaction for parameter configuration



Method 5: **Visualization of streaming in the network using state change graphs 2/3**

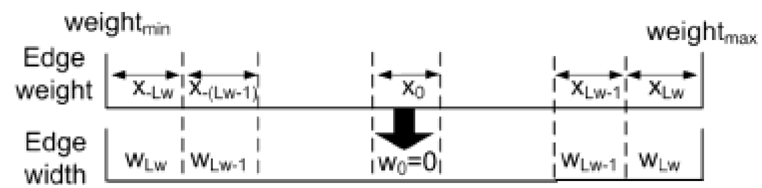
4th Step: **Optimization** of the graph visualization by maximizing its **entropy**

- Define the mapping function:

$$F(e_{weight}) = |i| \;, if\; X_{i-1} \le e_{weight} < X_i$$

where:

$$X_i = \left\{ \begin{array}{cc} 0 & , i < -L_w \\ \sum\limits_{j=-L_w}^{i} x_j & , -L_w \le i \le L_w \\ \sum\limits_{j=-L_w}^{L_w} x_j & , i > L_w \end{array} \right\}$$



The mapping function

where e_{weight} is the corresponding edge weight that is mapped to width w_k, for $k = F(e_{weight})$

- Maximize the following objective, where H_G^{out} is the entropy of the visualized information mapped on the edges E^c:

$$\overline{x'} = \arg\max_{\overline{x'}} \left\{ H_G^{out}\left(E^c,\ F'\left(\overline{x'}\right)\ \right) \right\}$$

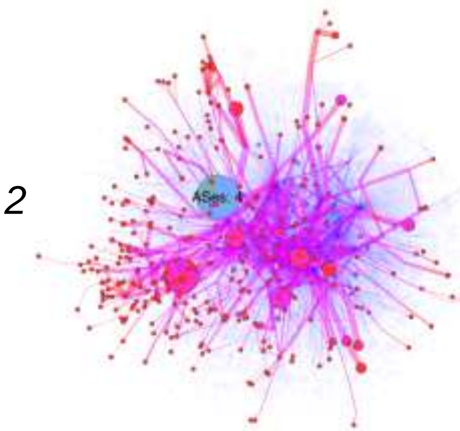
where: $\overline{x'} = (x_1, \ldots .x_{(L_w-1)}, x_{L_w}).$

Method 5: **Visualization of streaming in the network using state change graphs 3/3**

- *5th Step:* **Hierarchical clustering method**
  - Position the nodes using a force directed model
  - Calculate the proximity graph of the nodes based on the following formula for adding edges (relative neighborhood graph):

$$\|v^l_{i-pos} - v^l_{j-pos}\| \le \max\{\, \|v^l_{i-pos} - v^l_{k-pos}\|, \\ \|v^l_{j-pos} - v^l_{k-pos}\| \,\}, \forall v^l_k \in V^l,\ and\ i \neq j$$

Level 2

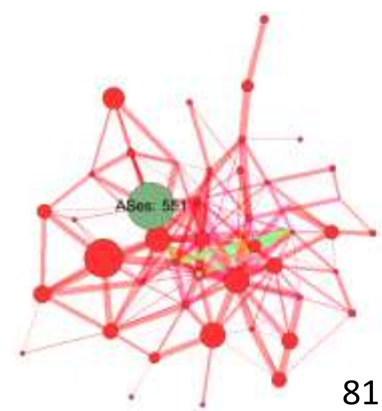


  - Combine pairs of neighboring nodes in order to maximize the weighted sum of the following metrics:

1) Geometric proximity: $\frac{1}{\|v^l_{i-pos} - v^l_{j-pos}\|}$
2) Similarity of neighborhood: $\frac{|N^l_i \cap N^l_j|}{|N^l_i \cup N^l_j|}$
3) Degree: $\frac{1}{\deg^l_i * \deg^l_j}$

  - where *l is the level of clustering hierarchy,* N_i *the neighbors of node* v_i *,and* deg_i *its degree in the proximity graph*

Level 10



Application: Visualization of routing data in the IP network 1/2

- **Scope/Problem definition**:
  - **Identify anomalies**, e.g. Large changes in the routing traffic either due to hardware failure or due to router misconfiguration
  - Perform **root cause analysis**, i.e. identify which ASes are responsible or involved in the detected anomalies

- **Data collected from the RIPE repository**:
  - **BGP (Border Gateway Protocol) messages** (>4,000 messages/min)
    - Contain **reachability information** for a specific prefix, i.e the AS-path followed to reach the owner of the prefix
    - Compared to the previous reachability state, they might contain routing changes, i.e. **changes in the reachability** of specific prefixes.

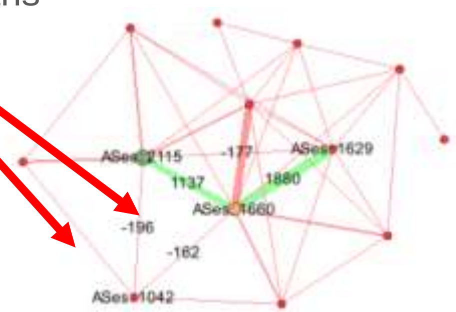
http://www.ripe.net/

Information Technologies Institute

Application: Visualization of routing data in the IP network 2/2

- **Application**
  - Calculation of the routing changes in each time window
  - State change graph
    - The edge size represents the change in the volume of the size of the routing change
    - Red color represents negative and green positive change
  - Optimization of the graph visualization by **maximizing its entropy**

Many of the edge widths look the same

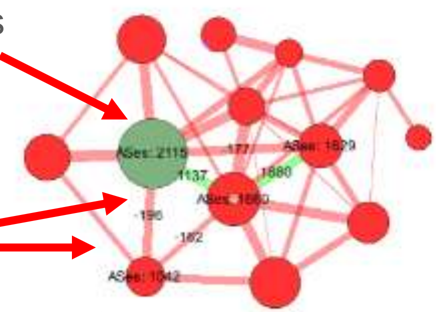


(b) $H_G^{out}(E, F') = 1.48$
$H_G^{out}(V, F') = 0.99$

Low entropy

This AS-cluster hijacks a large number of prefixes

Edge width difference enhanced through entopy



(a) $H_G^{out}(E, F') = 3.22$
$H_G^{out}(V, F') = 3.70$

Maximum entropy

Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings’ energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Method 6: **Graph-based descriptors for network anomalies detection & visualization**

Method name:

- *Graph descriptors for the detection and visualization of network anomalies*

Research field:

- Security

Big data issues addressed:

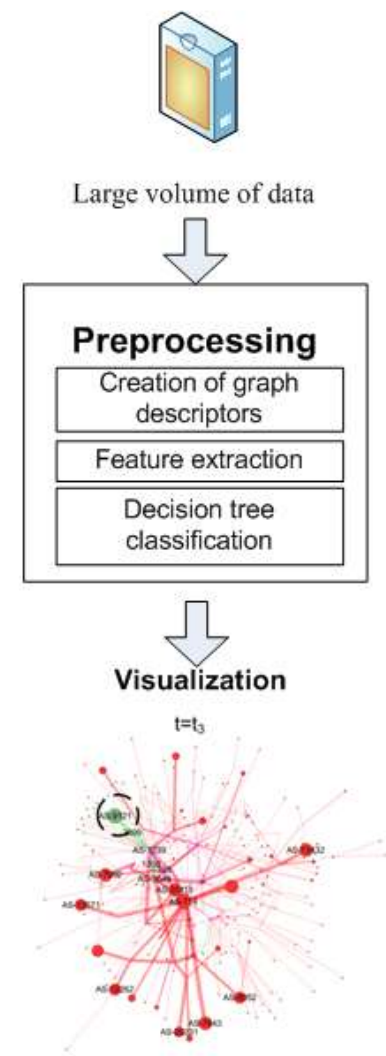
- Volume

Application areas:

- Network operator

Method 6: **Graph-based descriptors for network anomalies detection & visualization 1/3**

- *1st Step: Network data*
- *2nd Step:* For each pair of nodes/objects create multiple attributes
  - e.g. volume of messages between two network components, or the traffic change between ASes
- *3rd Step:* ***Graph descriptors***
  - Add the calculated nodes and edge attribute weights to the graph
- *4th Step:* ***Feature extraction***
  - Graph-based features, e.g. graph entropy
- *5th Step*: Decision tree classification for anomaly detection
- *6th Step:* Visualization of graphs for root cause analysis



Visualization of graph descriptos

Information Technologies Institute

Method 6: Graph-based descriptors for network anomalies detection & visualization 2/3

- *4^{th} Step:* ***Feature extraction***
  - ***Volume:***

$$f_{vol}^{G_i} = \sum_{e_j \in E_i} g\left(W(e_j)\right), g(x) = \begin{cases} 1, & for\ |x| \neq 0 \\ 0, & for\ |x| = 0 \end{cases}$$

  - ***Edge entropy:***

$$f_{ee}^{G_i} = \sum_{j=1}^{Y^i} \frac{y_j^i}{y_{total}^i} \log\left(\frac{y_j^i}{y_{total}^i}\right)$$

  - ***Graph entropy:***

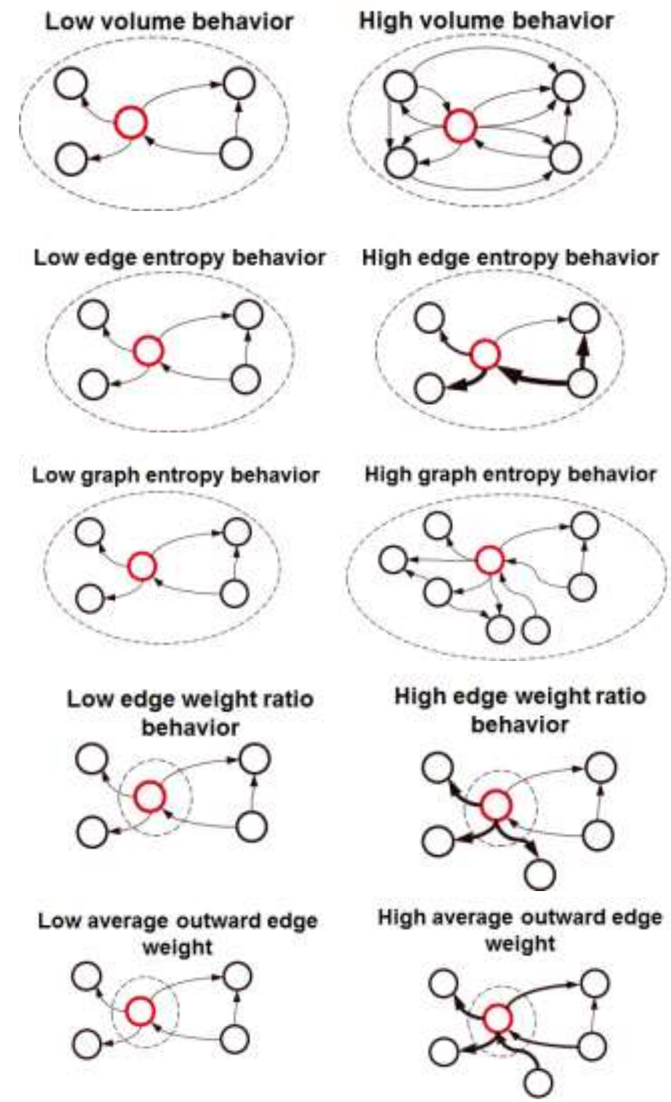
$$f_{ge}^{G_i} = \min_{X,Y} I(X \wedge Y)$$

  - ***Edge weight ratio:***

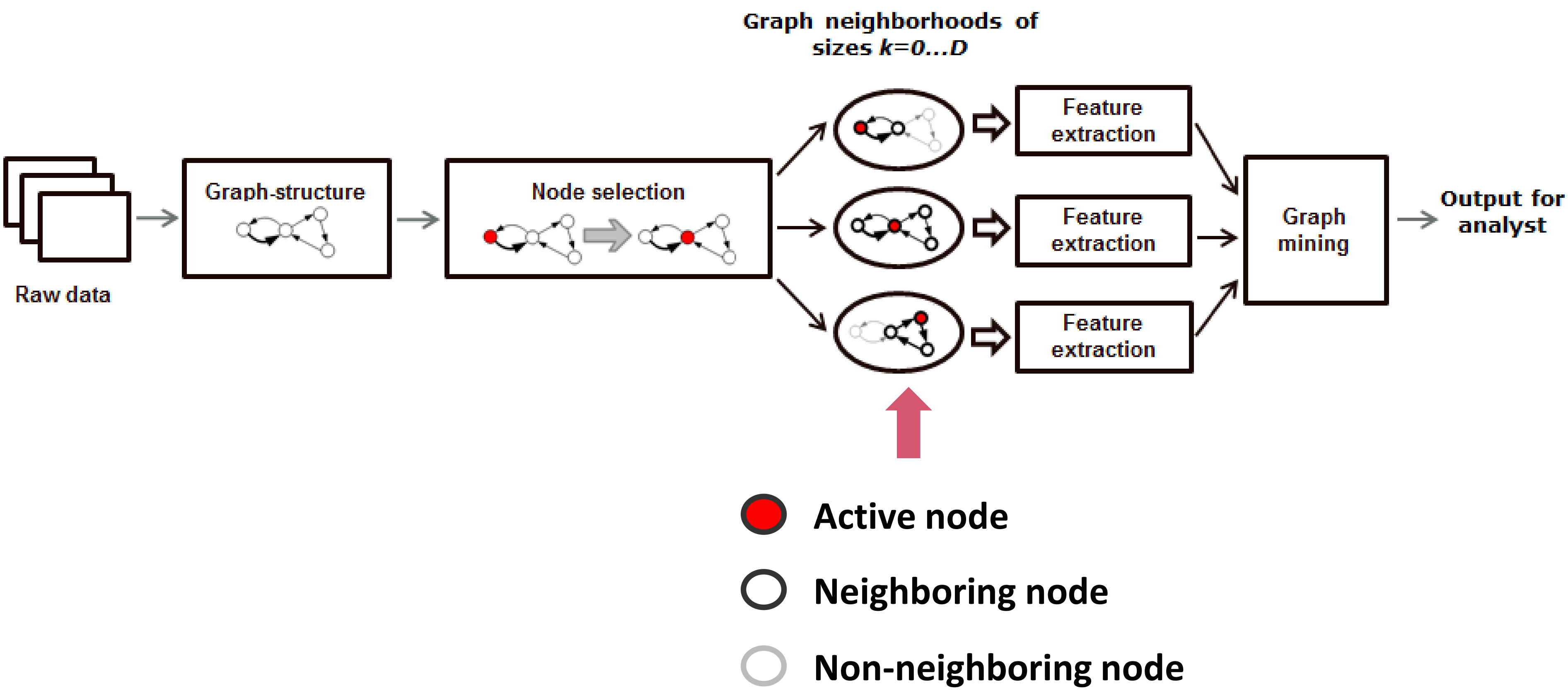
$$f_{wr}^{G_i} = \frac{\sum_{e_j \in E_i^{out}} W(e_j)}{\sum_{e_j \in E_i^{in}} W(e_j)}$$

  - ***Average outward/inward edge weight:***

$$f_{avout}^{G_i} = \frac{\sum_{e_j \in E_i^{out}} W(e_j)}{|E_i^{out}|}$$



Method 6: Graph-based descriptors for network anomalies detection & visualization 3/3



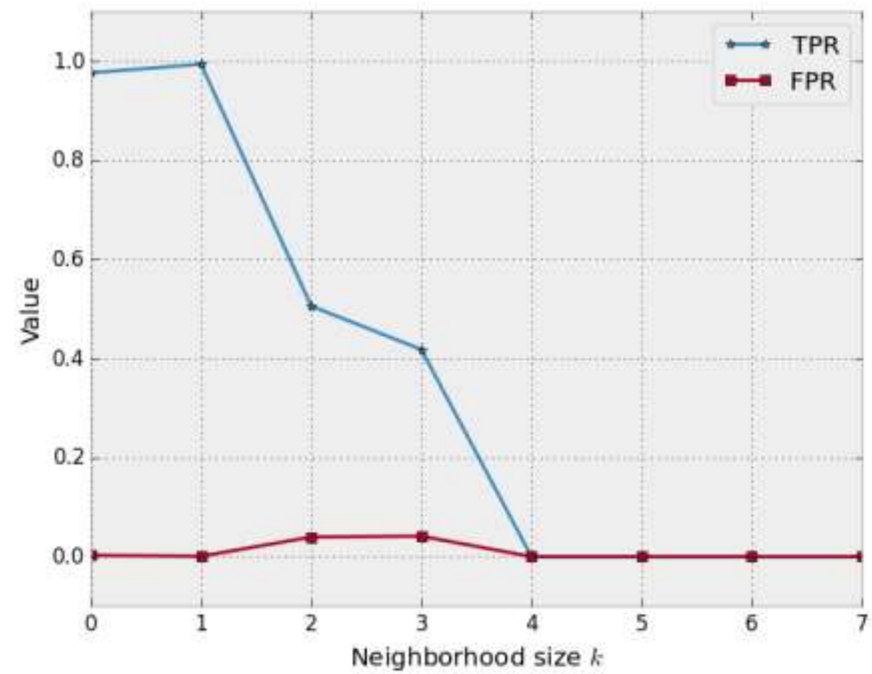
Application 1: Detection performance in SMS flood attack

- DDoS attack
- 4800 mobile devices
- 300 infected devices

Comparison with other anomaly detection methods

Method	TPR	FPR
Graph descriptor	**99.40%**	**0%**
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010	31.58%	2.74%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010 with RF	99.12%	0.07%
K. Henderson et al., SIGKDD 2011	97.66%	0.14%
K. Henderson et al., SIGKDD 2011 with RF	97.66%	0.21%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014	40.06%	0.93%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014 with RF	99.12%	**0%**
Kim et al. Security and Privacy in Communication Networks 2013	93.2%	1.4%
Yan et al. Recent Advances in Intrusion Detection 2009	96.5%	2.1%

Anomaly detection results



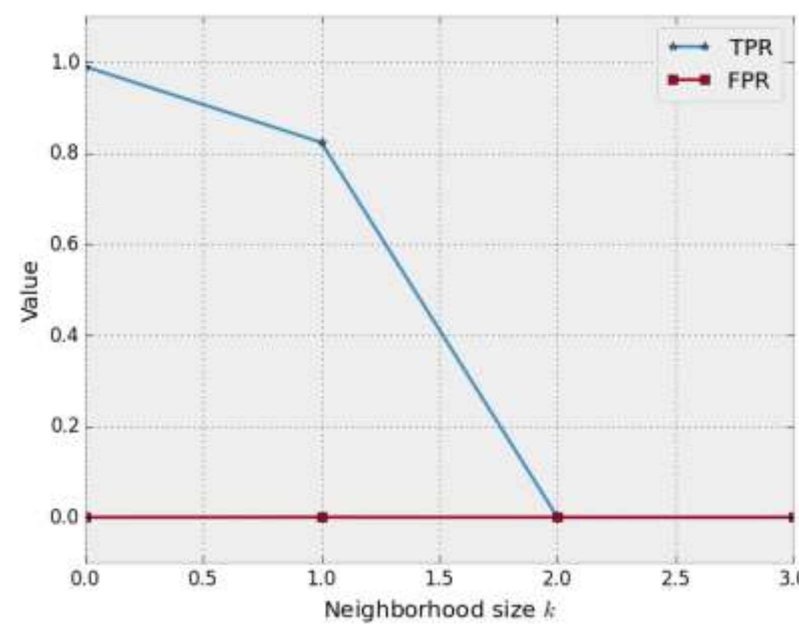
Application 2: Detection performance in Spam SMS attack

- Malware sends spam
- 10000 mobile devices
- 102 infected devices

Comparison with other anomaly detection methods

Method	TPR	FPR
Graph descriptor	**98.05%**	0.01%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010	33.01%	0.11%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010 with RF	87.37%	0.1%
K. Henderson et al., SIGKDD 2011	8.73%	**0%**
K. Henderson et al., SIGKDD 2011 with RF	7.76%	0.03%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014	33.01%	8.86%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014 with RF	38.83%	0.07%
Xu et al., IEEE Intelligent Systems 2012 (PCA)	87.2%	0.03%
Xu et al., IEEE Intelligent Systems 2012 (all features)	79.4%	0.10%

Anomaly detection results



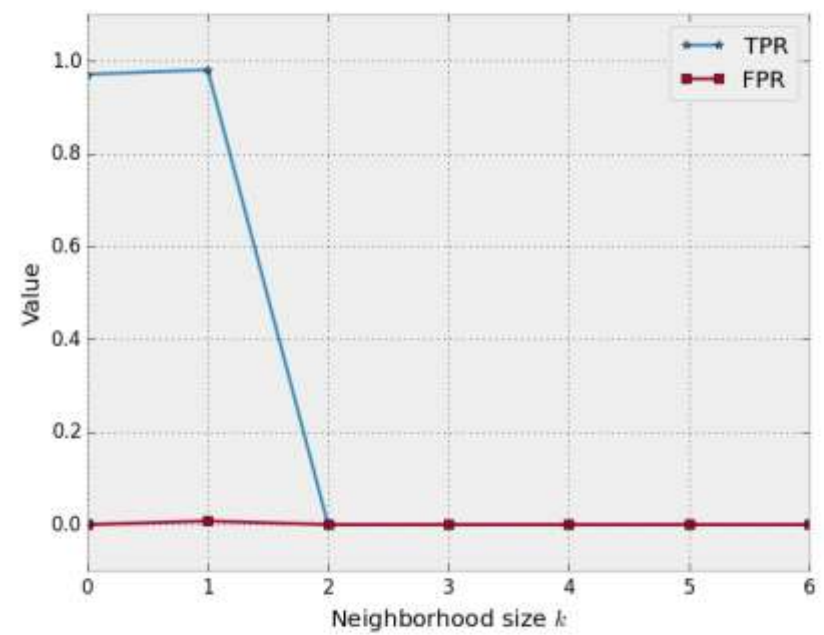
Application 3: Detection performance in RRC attacks

- **DDoS attack**
- **200 mobile devices**
- **100 infected**

Comparison with other anomaly detection methods

Method	TPR	FPR
Graph descriptor	**99%**	**0.74%**
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010	0%	16.29%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010 with RF	93.06%	4.44%
K. Henderson et al., SIGKDD 2011	96.04%	16.29%
K. Henderson et al., SIGKDD 2011 with RF	98.02%	5.18%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014	0.99%	2.74%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014 with RF	95.05%	**0.74%**

Anomaly detection results



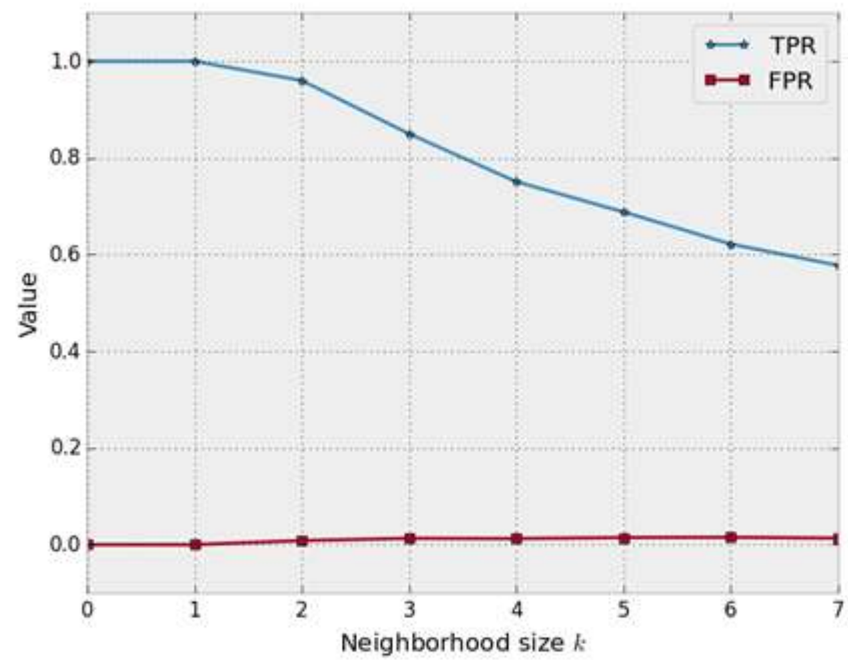
Application 4: Detection performance in Malware infection cases

- **Malware sends spam**
- **Infects new devices**
- **2000 mobile devices**

Comparison with other anomaly detection methods

Method	TPR	FPR
Graph descriptor	**99.82%**	**0.01%**
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010	48.63%	1.17%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010 with RF	99.67%	0.10%
K. Henderson et al., SIGKDD 2011	97.21%	0.86%
K. Henderson et al., SIGKDD 2011 with RF	98.76%	0.61%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014	4.12%	4.77%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014 with RF	69.16%	14.08%

Anomaly detection results

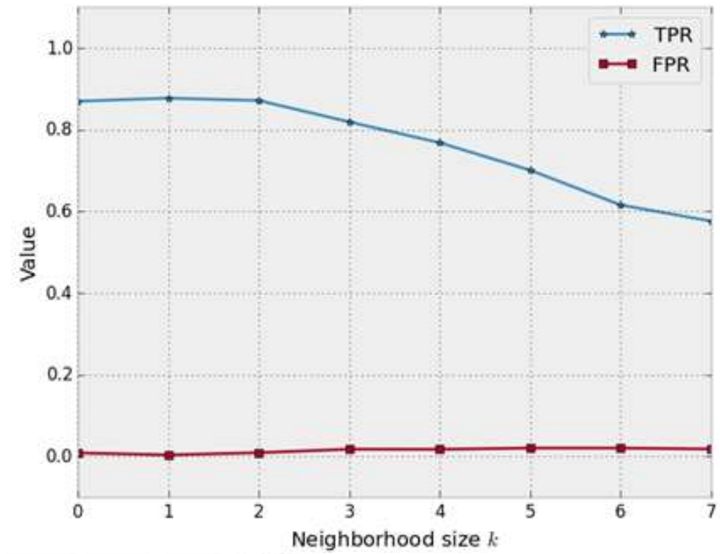


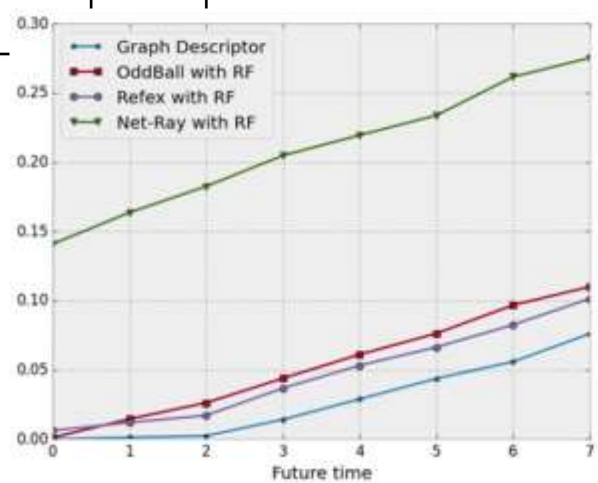
Application 5: Prediction performance in Malware infection cases

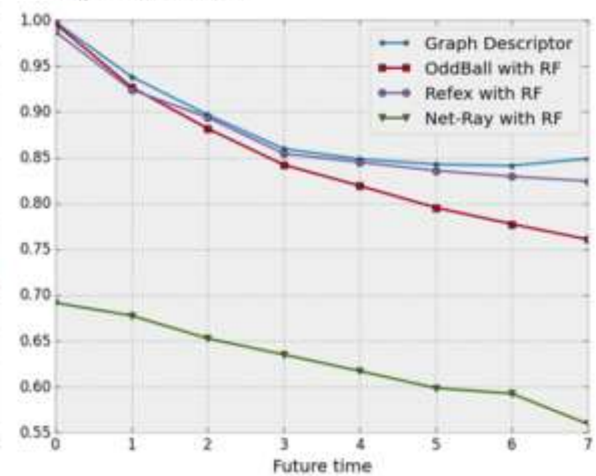
Comparison with other anomaly detection methods for prediction at t+2

Method	TPR	FPR
Graph descriptor	**89.69%**	**0.23%**
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010	38.21%	1.22%
L. Akoglu et al., Advances in Knowledge Discovery and Data Mining 2010 with RF	88.19%	2.61%
K. Henderson et al., SIGKDD 2011	84.42%	1.41%
K. Henderson et al., SIGKDD 2011 with RF	89.40%	1.71%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014	3.57%	4.82%
U. Kang et al., Advances in Knowledge Discovery and Data Mining 2014 with RF	65.26%	18.24%

Anomaly prediction results at t+2





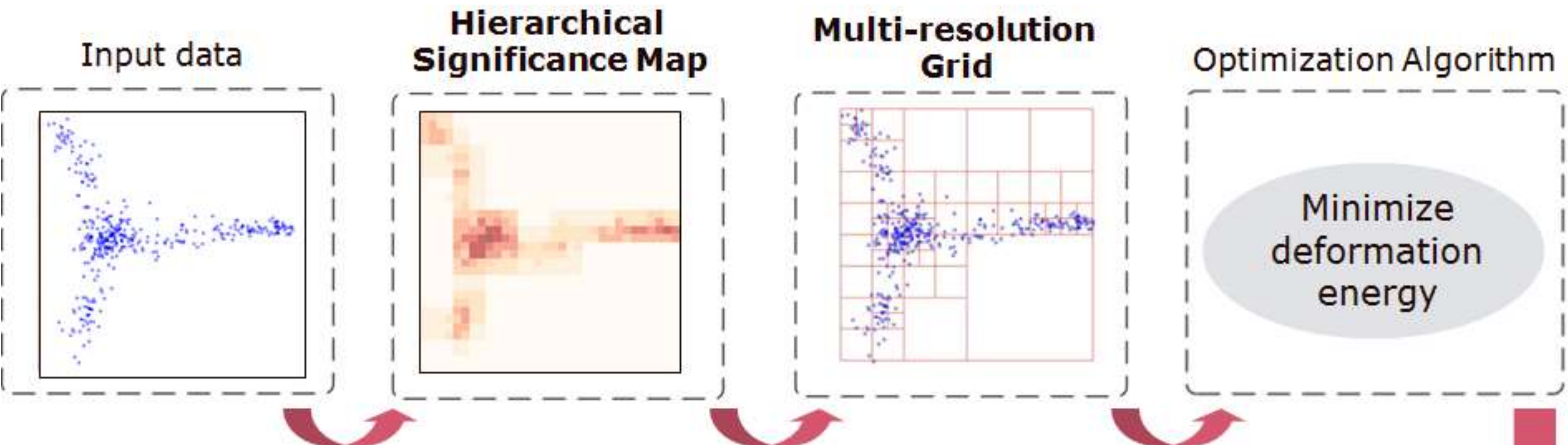


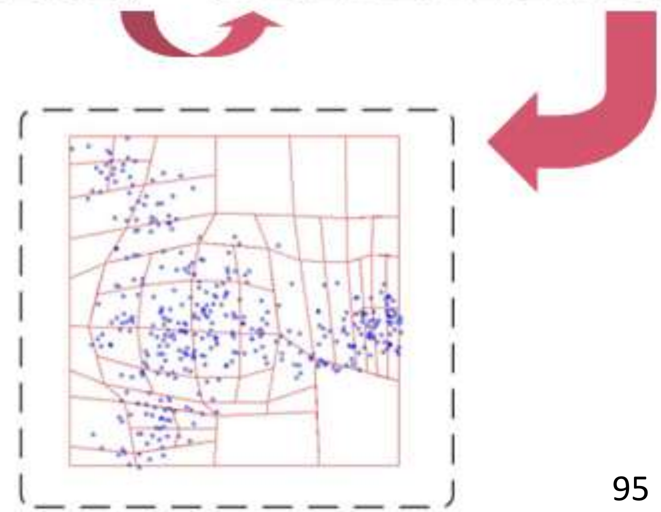
Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings' energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Method 7: **Hierarchical Magnification for insight gain in smaller displays 1/3**

current SoA vS proposed method





Magnified View:

- Preservation of the **original** patterns
- Non-linear magnification

Method 7: Hierarchical Magnification for insight gain in smaller displays 2/3

- **Significance map generation**

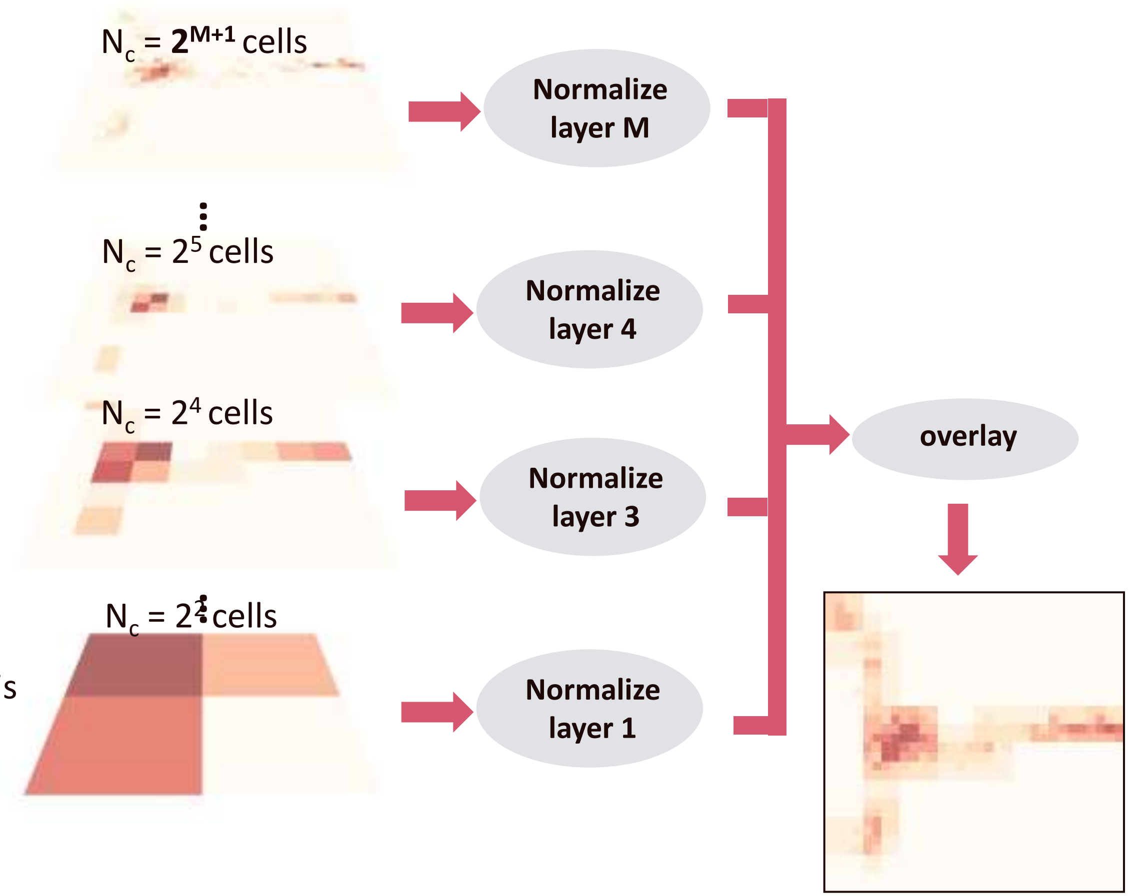
The final multi-resolution **hierarchical significance** map S is defined as

$$S^{hier} = \bigcup s_i^{hier}$$

,where i is the hypercube index

$$s_i^{hier} = \sum_{s_j^l \in Q_s\left(f_i^M\right)} N\left(s_j^l\right)$$

,where j is the hypercube index in the lth layer, Qs is the set of overalpping hypercube , and N is a normalization operator for eliminating the layer-dependent amplitude differences.



Method 7: **Hierarchical Magnification for insight gain in smaller displays 3/3**

- Definition of the total **quad deformation energy**:

$$D_u = \sum_{f \in F} w_f D_u(f), \;\; where \;\; D_u(f) = \sum_{\{i,j\} \in E(f)} \left\| \left(v_i^{'} - v_j^{'}\right) - s_f \left(v_i - v_j\right) \right\|^2$$

where $v_i^{'}$ are the new vertex positions, v_i are the initial vertex positions, s_f is the bin scaling factor, and w_f the **significance** of hyperrectangle *f*

- Definition of the **total edge deformation energy**:

$$D_l = w_e \sum_{\{i,j\} \in E} \left\| \left(v_i^{'} - v_j^{'}\right) - l_{ij} \left(v_i - v_j\right) \right\|^2, \;\; where \; l_{ij} = \frac{\left\| v_i^{'} - v_j^{'} \right\|}{\left\| v_i - v_j \right\|}$$

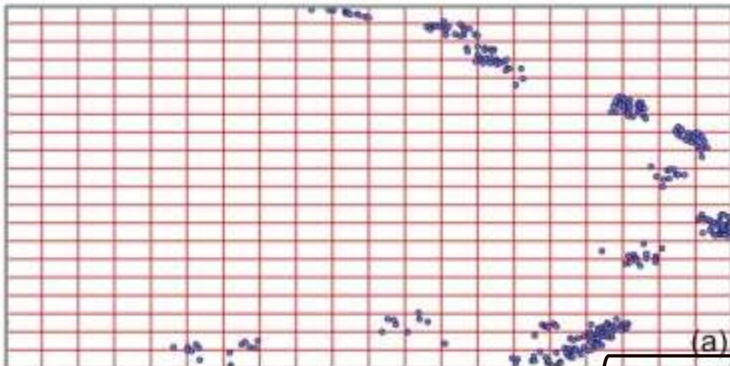
where w_e is the **significance** of each edge e

- **Optimization** (i.e. minimization) of the **total grid deformation energy** (Quadratic Form) *D:*

$$D = D_u + D_l$$

- **Solved iteratively:**
  1. **Find s_f such that $D_u' = 0$**
  2. **Minimize *D***

Application 7.1: **Hierarchical Magnification for 2D scatterplots**

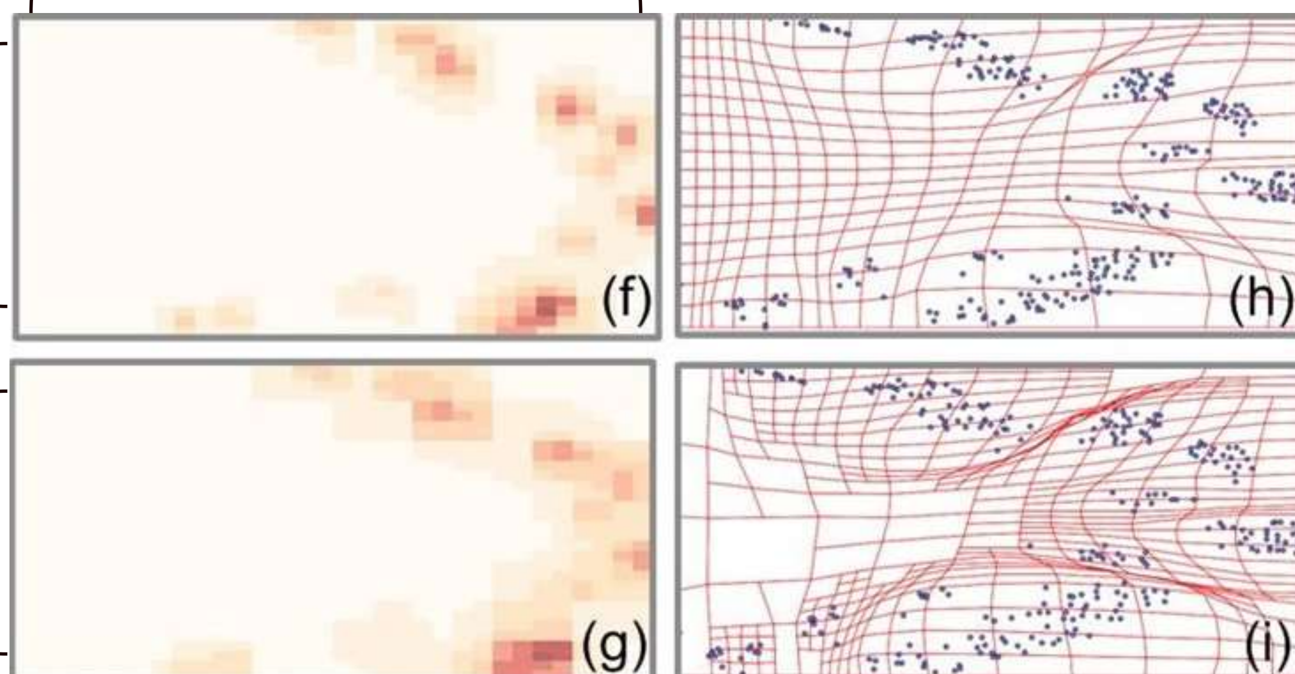


Original 2D scatterplot

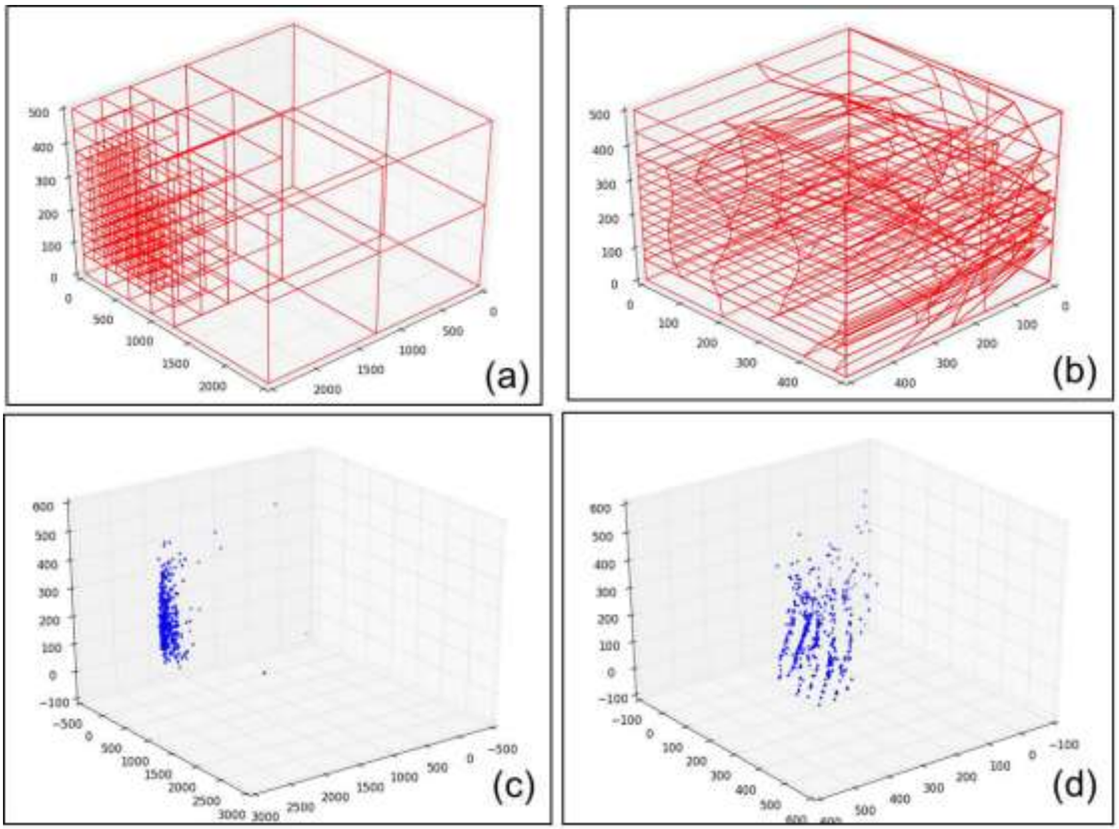
Saliency

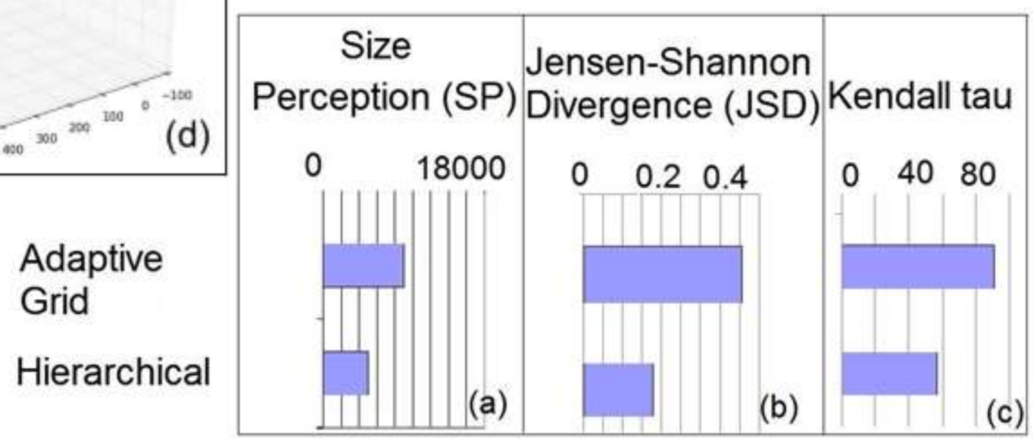
Wu et al.

Proposed approach

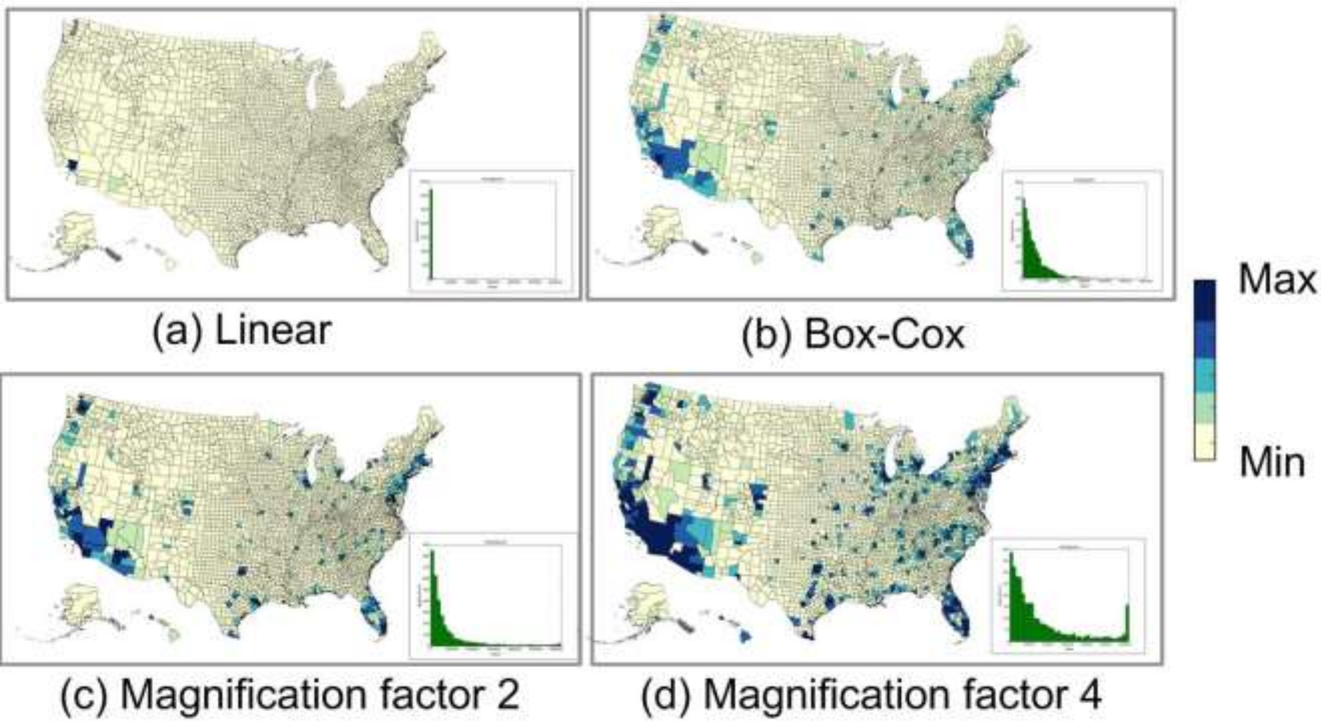


Application 7.2: **Hierarchical Magnification for 3D scatterplots**





Application 7.3: **Hierarchical Magnification for Choropleth maps**

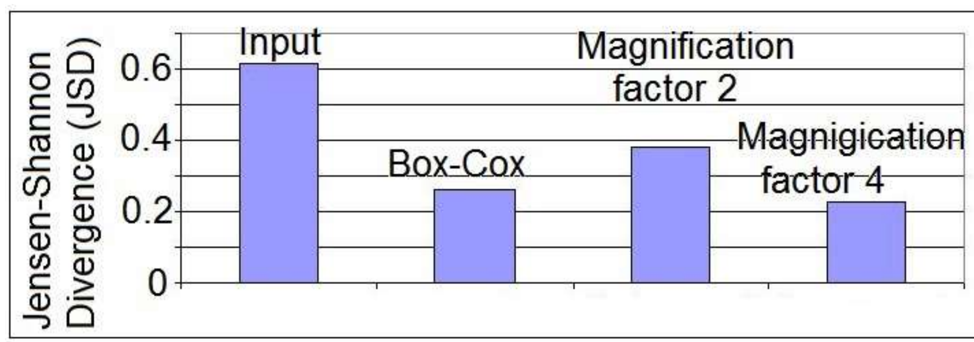


(a) Linear

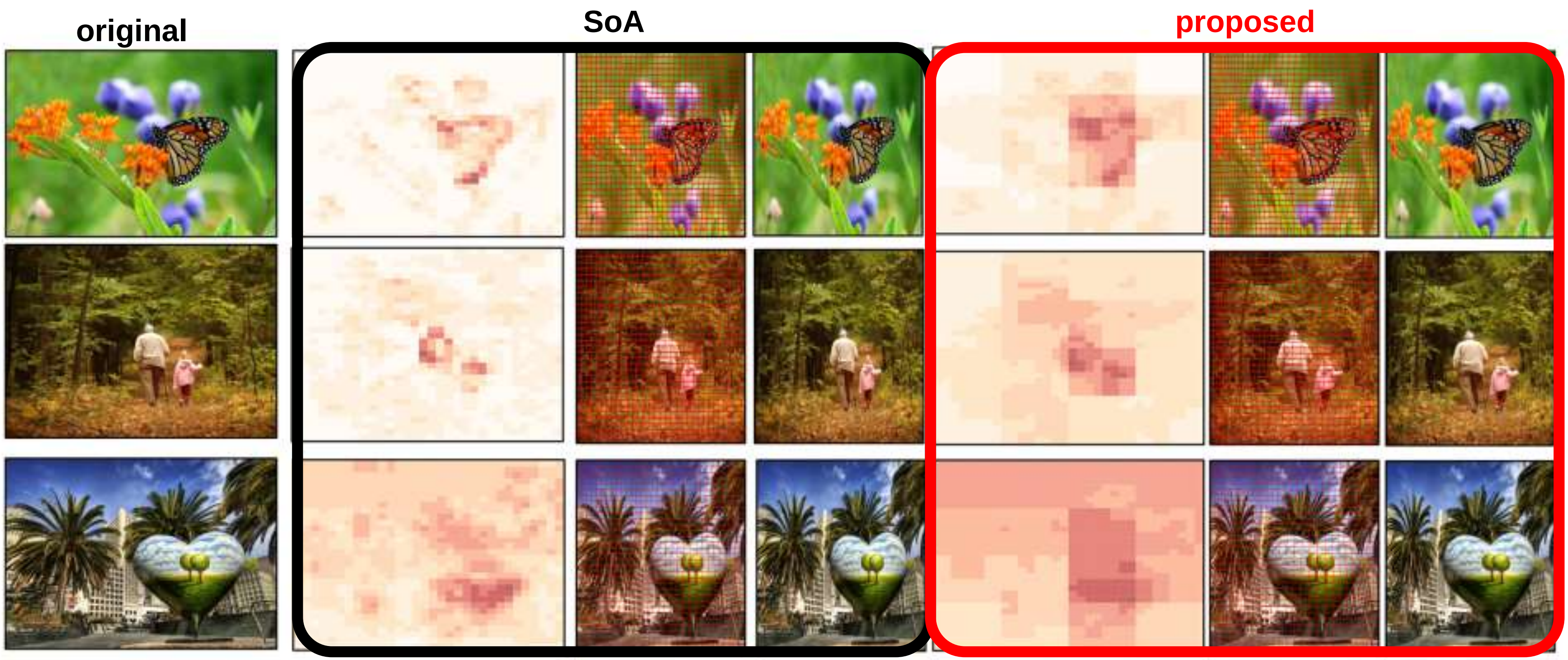
(b) Box-Cox

(c) Magnification factor 2

(d) Magnification factor 4



Application 7.4: Hierarchical Magnification for Image Resizing



Original Image

Single Layer Significance map

Resize results u ing the single La Significance ma

Hierarchical Significance map

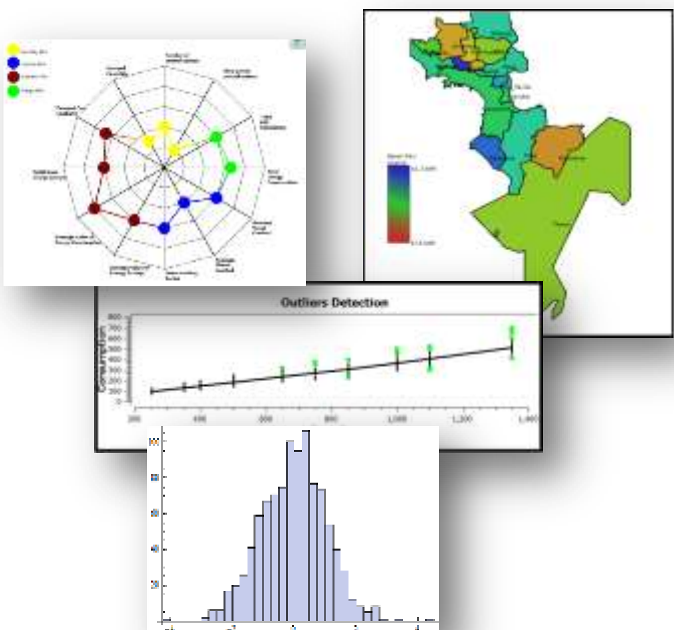
Resize resul sing the Hiera cal Significa e map

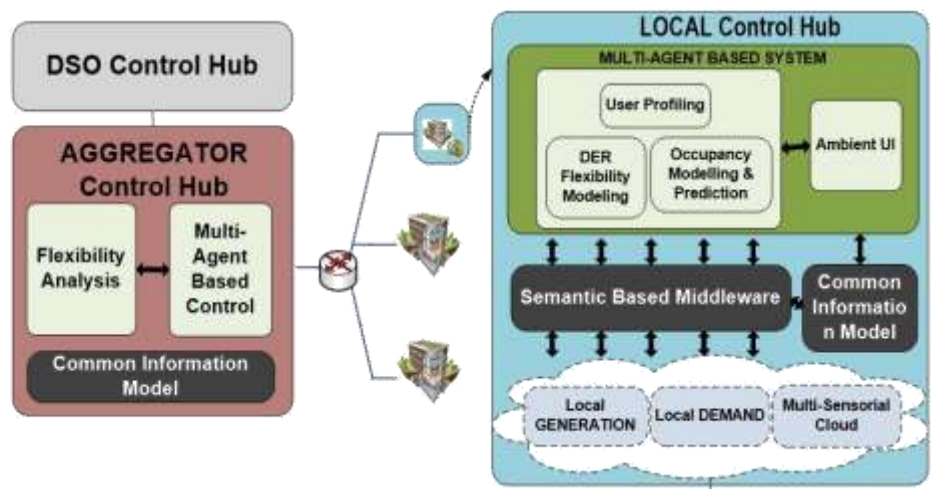
Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings’ energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Method 8: **Energy sustainability of buildings' energy sustainability**

INERTIA VA tool supports the analysis of large volumes of DER related Energy / Flexibility Profile data of Aggregators Portfolios





Method 8: Energy sustainability of buildings' energy sustainability 2/2

Normal Operation Data Analysis (Aggregator as a retailer)

- Clustering/Classification of Local Hubs portfolio based on
  - energy profile (energy consumption/ cost of energy) data
  - flexibility (potential flexibility) data
- Trend Analysis for the extraction of patterns - more precise placement in energy markets (trend analysis towards forecasting operations, what if analysis)
- Outliers Analysis on the available dataset of Local Hub's portfolio

Demand Response Operation Related Analysis (Aggregator as DR services provider)

- Clustering/Classification of Aggregator's portfolio (DR operation)
- Pattern Recognition/Trend Analysis during DR operation
- Outliers analysis during the DR operation

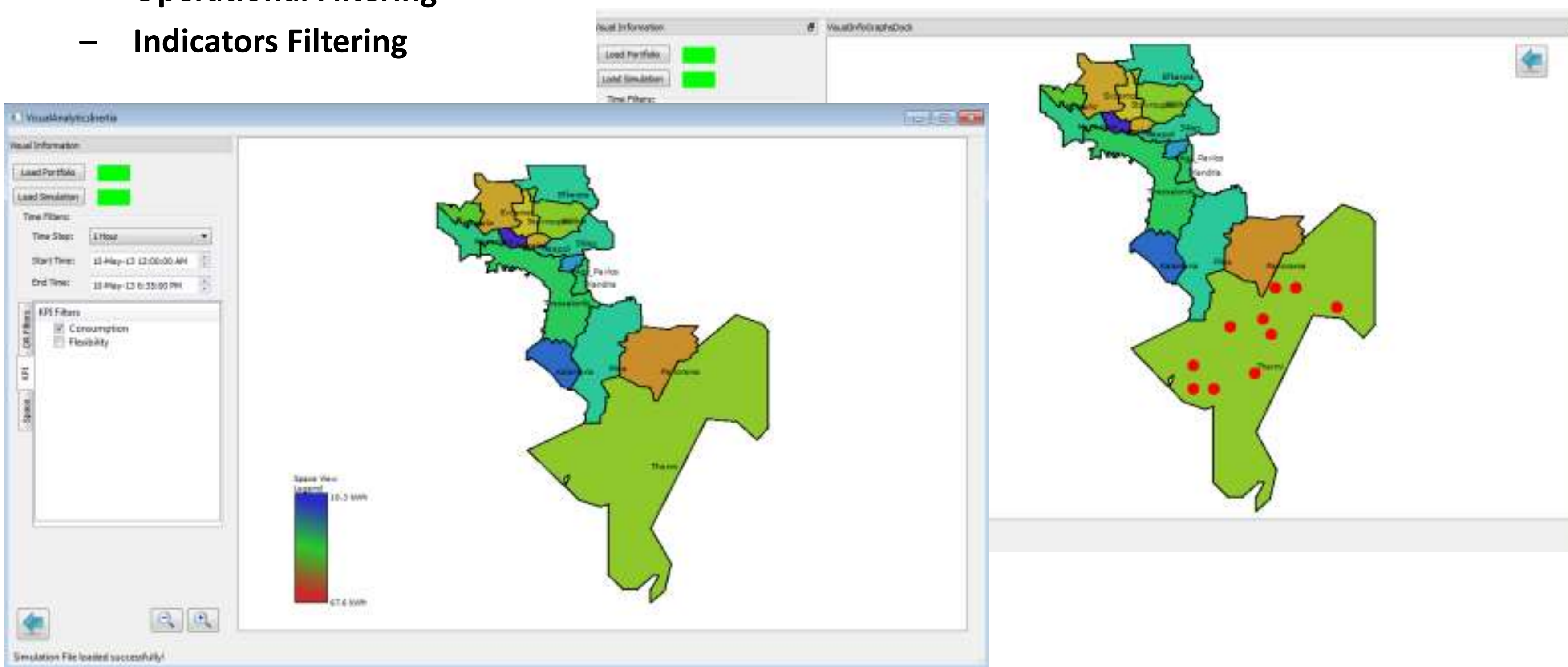
Functions supported by the tool

- Information Visualization Analysis
- Portfolio Scenario Analysis
- Optimization Scenario Analysis

Application 8.1: Information Visualization 1/2

An overview analysis on the portfolio is provided based on:

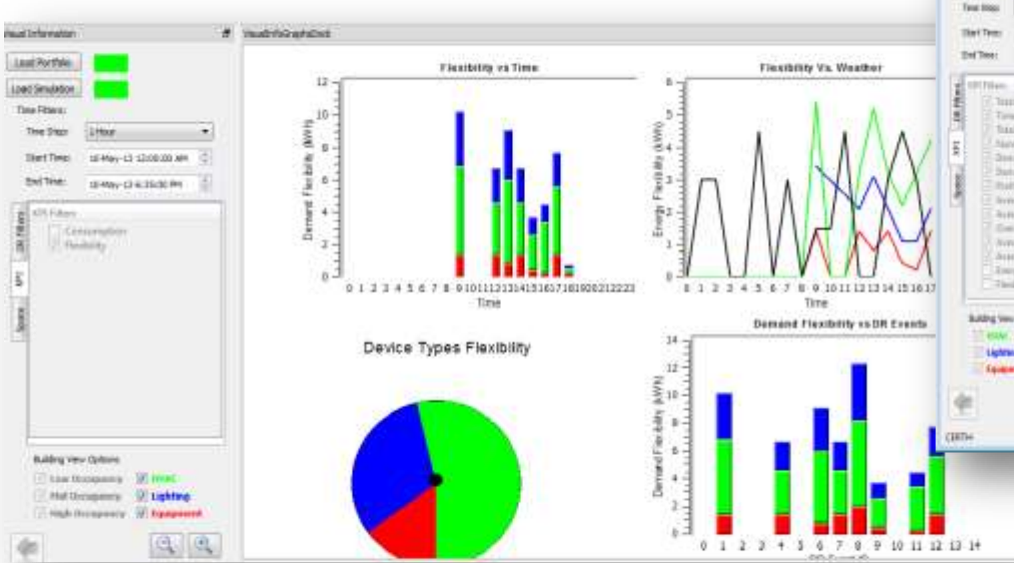
- **Time period Filtering**
- **Spatial/location Filtering**
- **Operational Filtering**
- **Indicators Filtering**

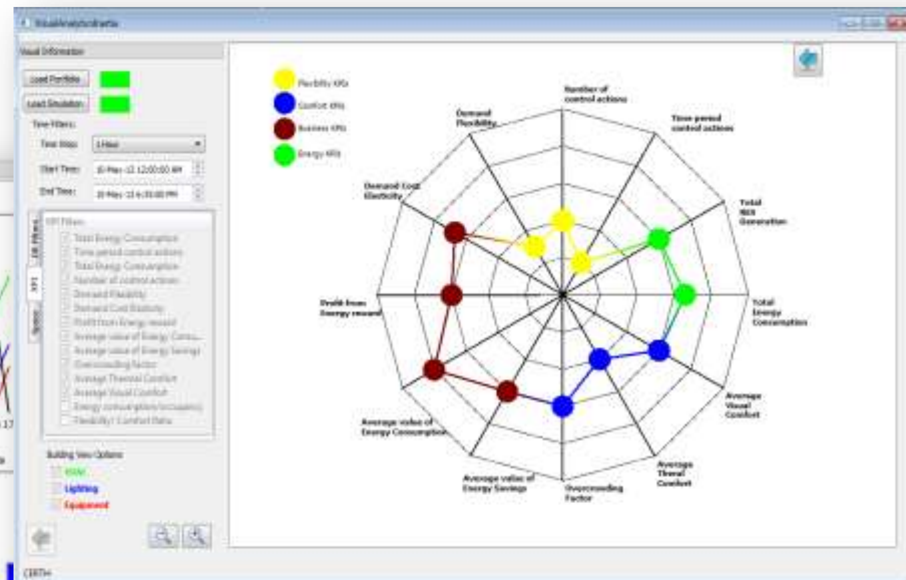


Application 8.1: Information Visualization 2/2

Insights for each Local Hub of the portfolio

- **Overview of KPIs (Energy/ Flexibility/ Business)**
- **Detailed time series presentation of KPIs**



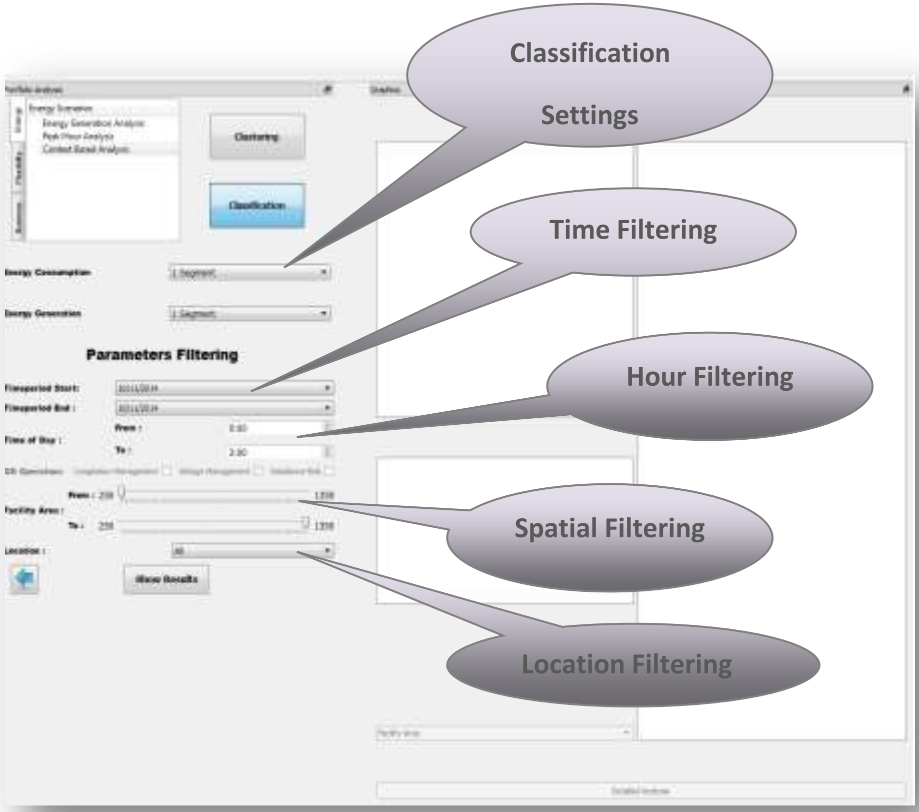


Application 8.2: Portfolio Analysis Scenarios 1/2

Methodology:

- **Clustering techniques for the extraction of energy/ business/ flexibility based clusters**

- **Classification techniques for hierarchical management of portfolio in predefined clusters**

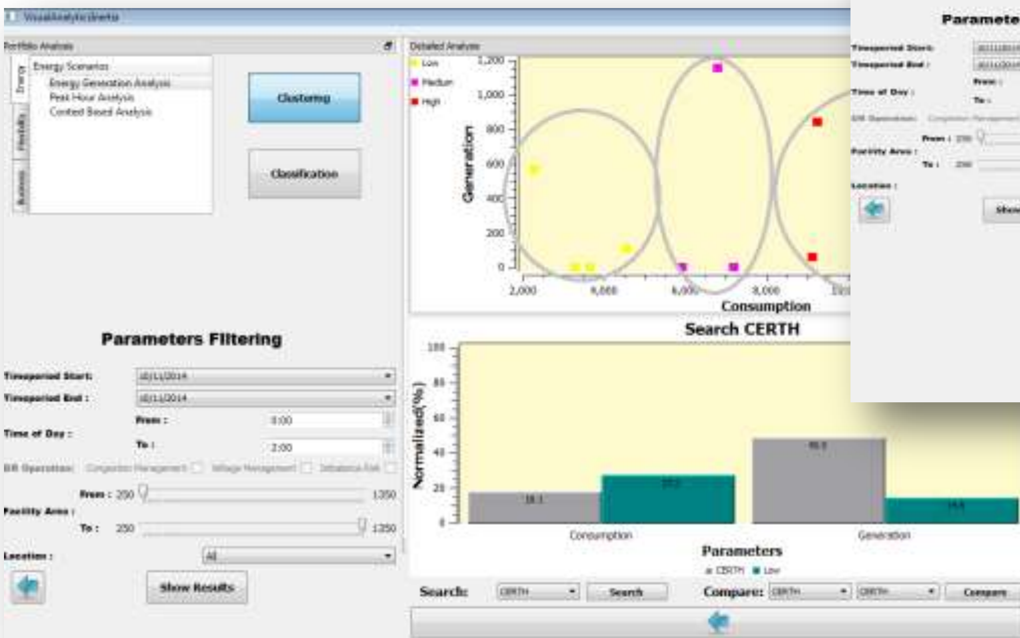


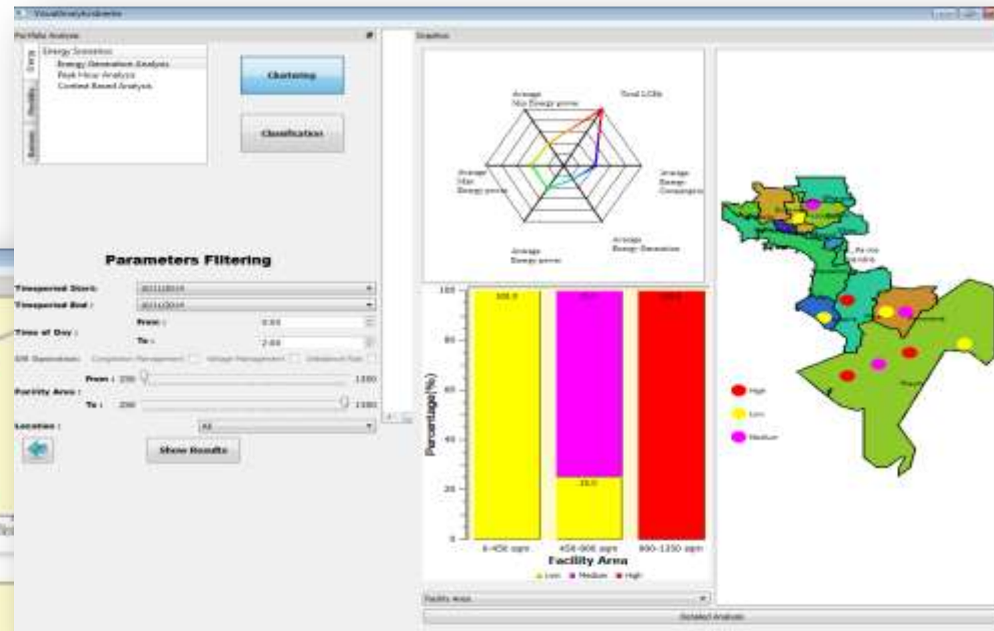
Multiple selection tab for the criteria / parameters of analysis supported

Application 8.2: **Portfolio Analysis Scenarios 2/2**

Alternative views are available for visual presentation:

- **Kiviat Diagram**
- **Map Presentation**
- **Bar Charts & Histograms**
- **Point Charts**

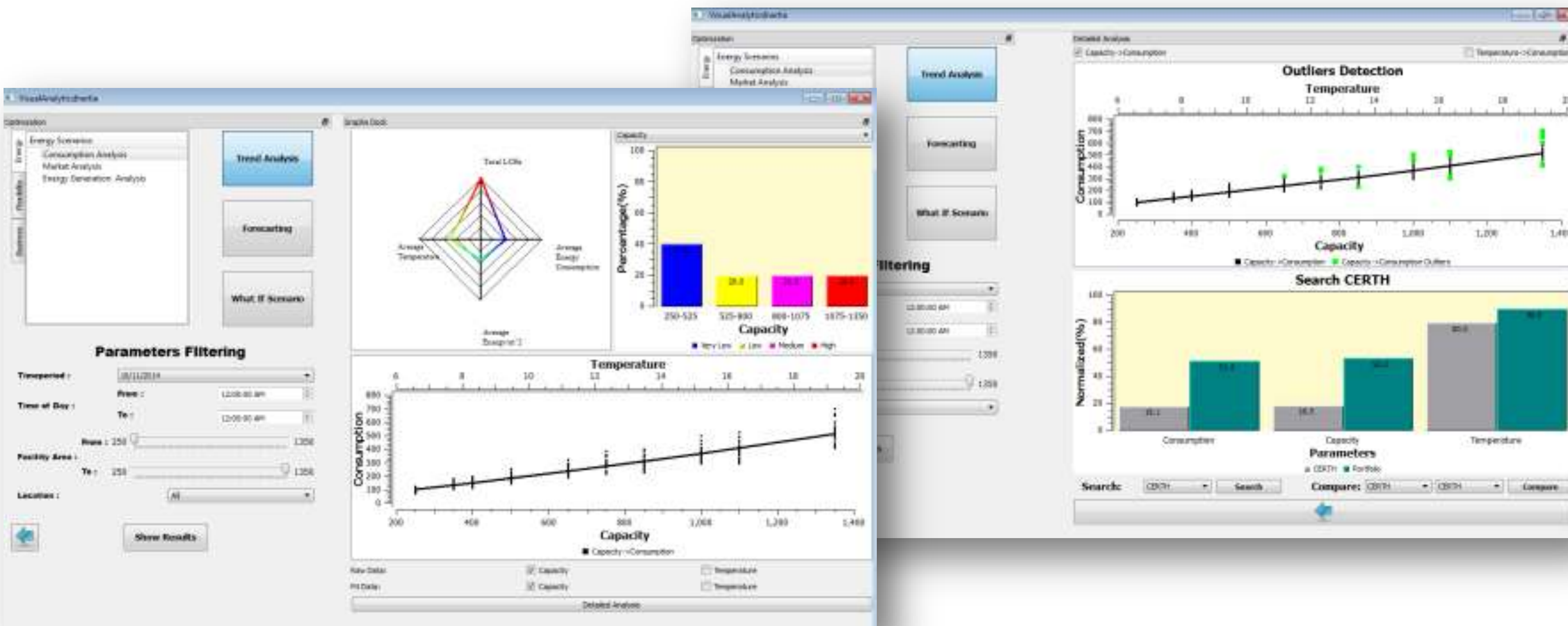




Application 8.3: Optimization Analysis Scenarios 1/2

Multiple Scenarios are examined as part of the Optimization Process

- **Trend Analysis towards the extraction of trends within the portfolio**
- **Anomaly Detection & Outliers Analysis → Deviation from trend line**

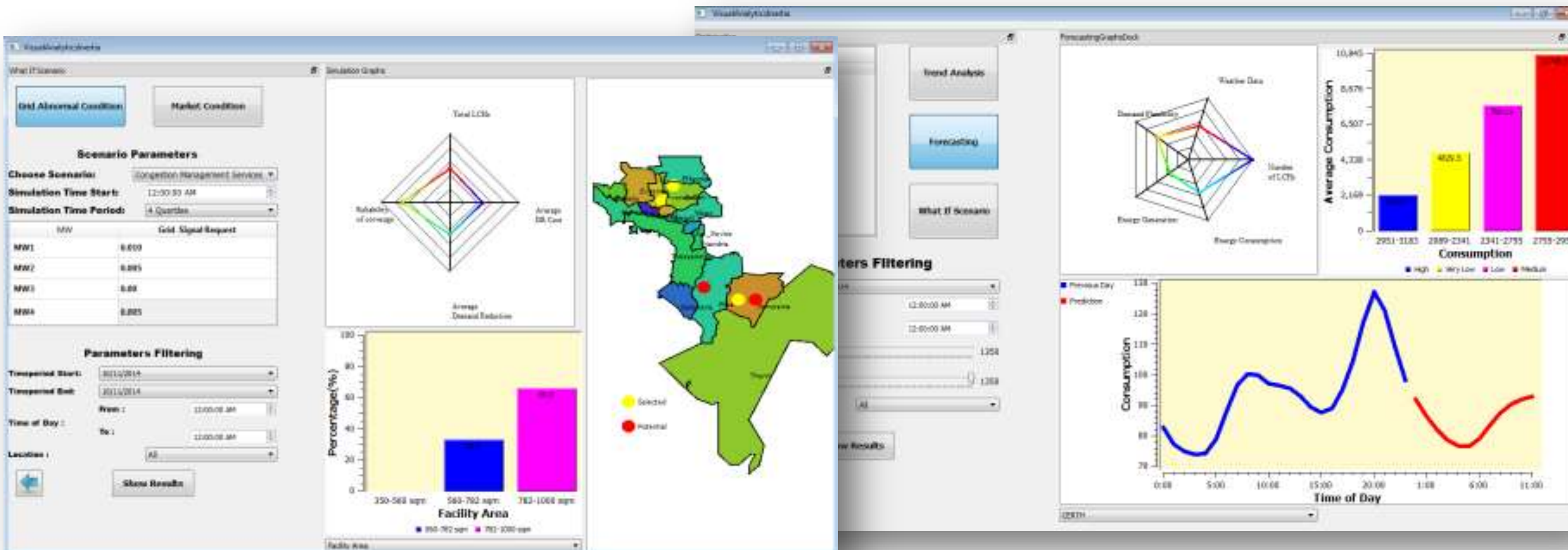


Information Technologies Institute

Application 8.3: Optimization Analysis Scenarios 2/2

Simulation analysis addressing also the DR signals

- "What - if" analytics → Based on historical data during normal conditions
- Simulation analysis → Addressing portfolio performance during DR conditions
- Forecasting Engine → Short term forecasting based on historical data



Presentation outline

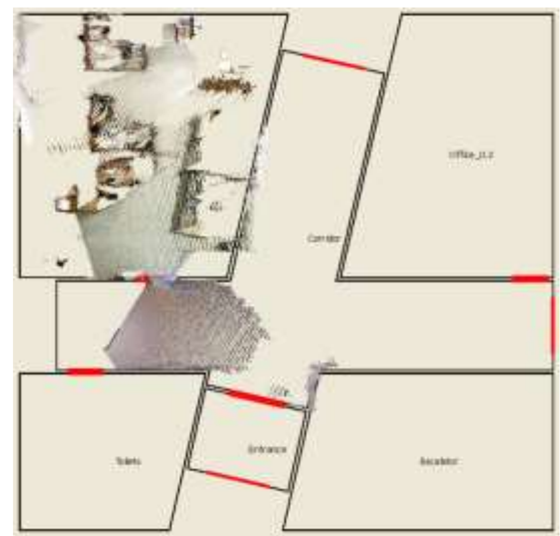
1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings' energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Method 9: **Occupancy tracking in closed spaces**

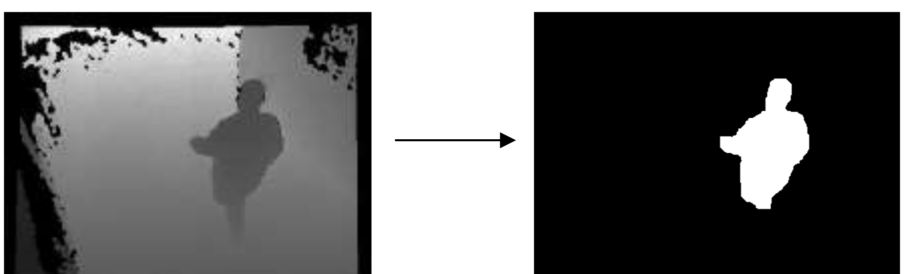
Occupancy tracking in indoor environments:

- **Multi-space**
- **Multi-camera (privacy preserving)**
- **Camera calibration on the architectural map**
- **Multi-occupant tracking**
- **Occupants' tracking**
- **Extraction:**
  - **Occupancy flows**
  - **Occupancy statistics (per occupant, per space, heat maps, etc.)**

Multi-space & multi-camera system



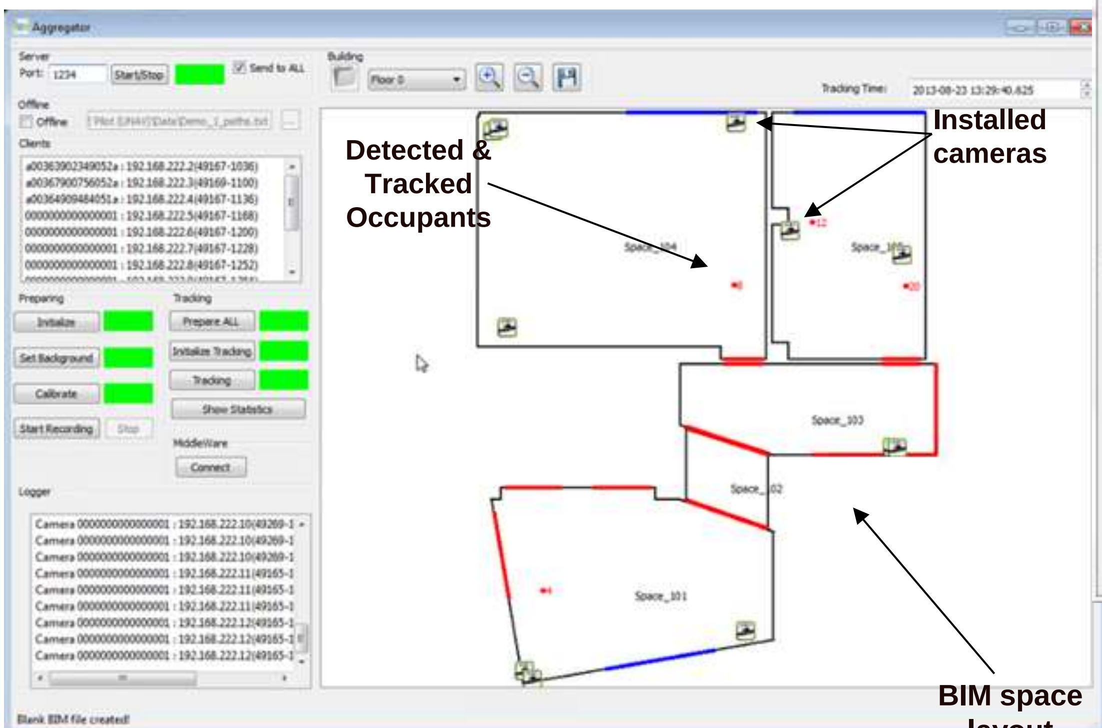
Cameras' view projection on an architectural map



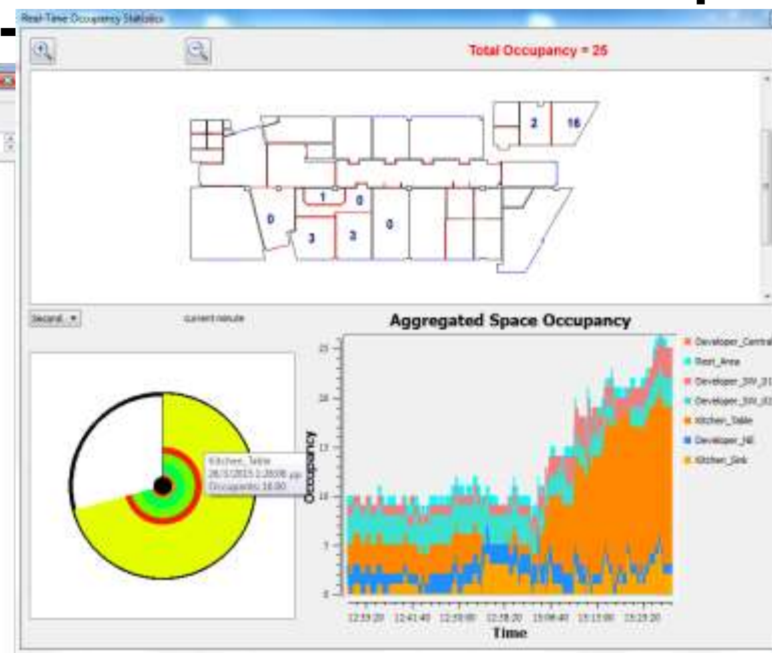
Foreground Extraction

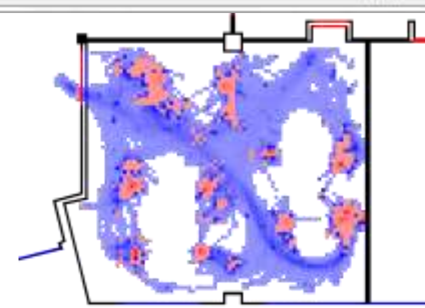
Method 9: Occupancy tracking in closed spaces

Occupancy tracking and analysis in indoor environments

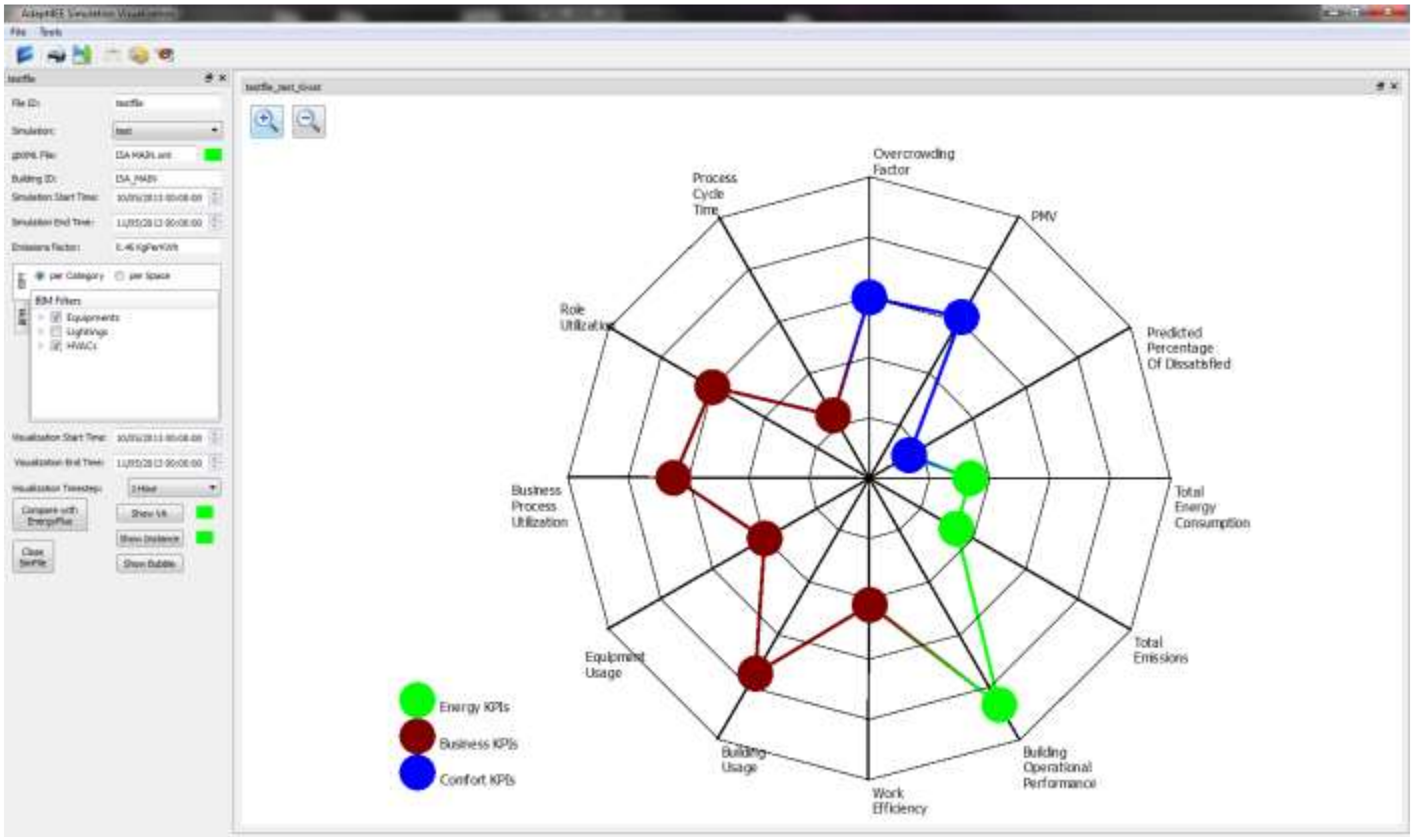


Occupancy Extraction System



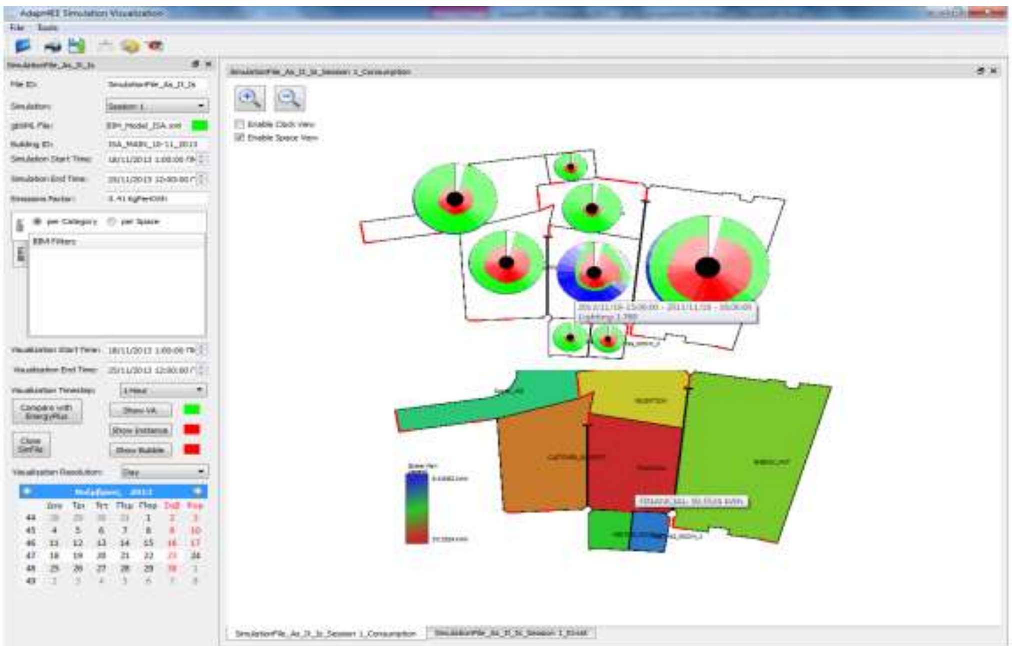


Application 9.1: **Kiviat diagram with (actual & simulated) KPIs**

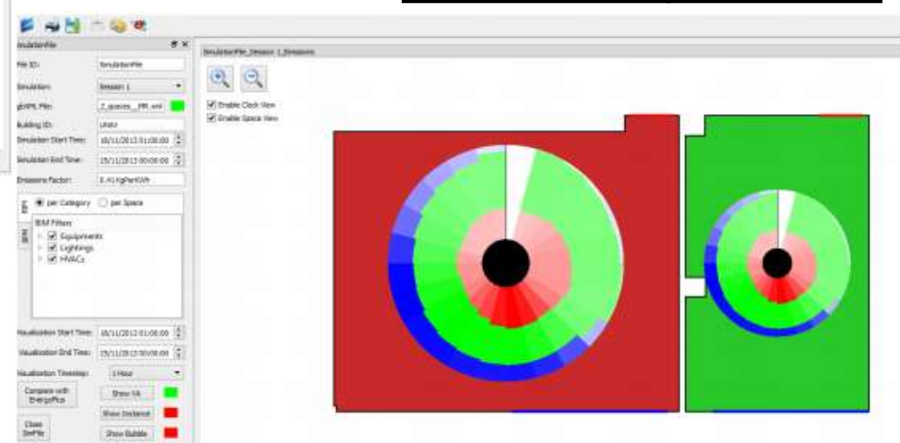


Application 9.2: **Detailed Spatio-temporal analysis of Building Performance**

<u>Detailed spatiotemporal analysis</u> (clock view)



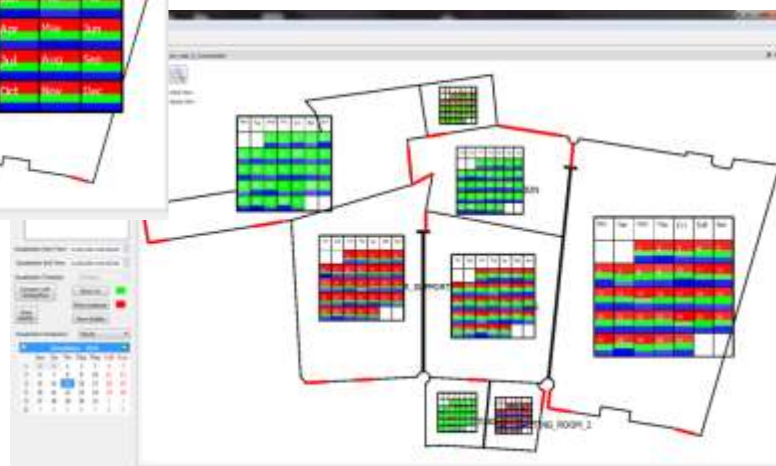
<u>Combined space view</u>

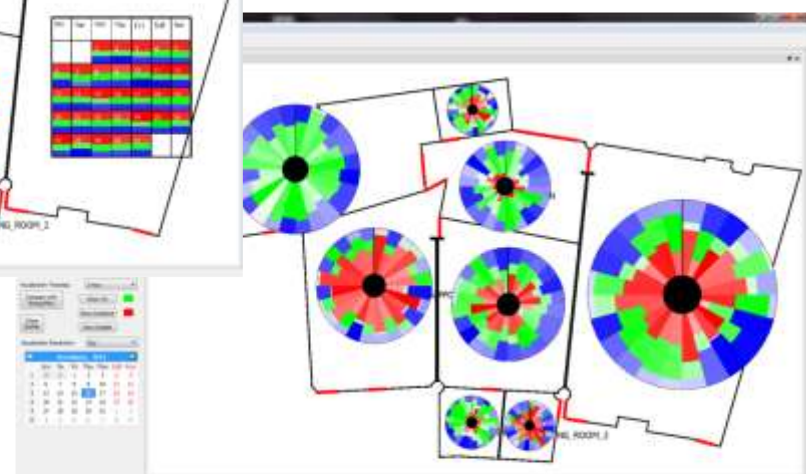


Application 9.3: **KPI drill-in building level**

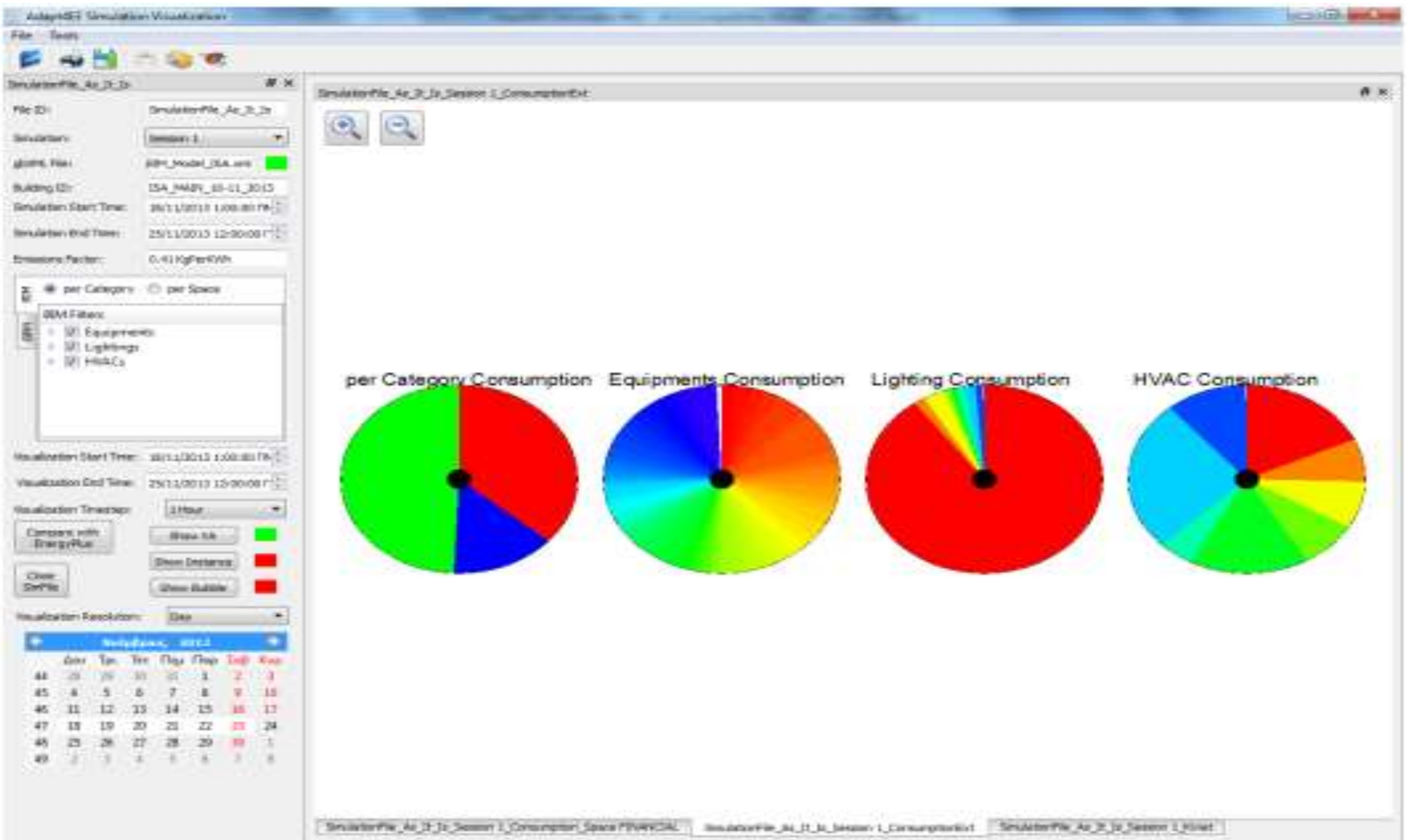
Time Resolution Filters

(year, month & day view)

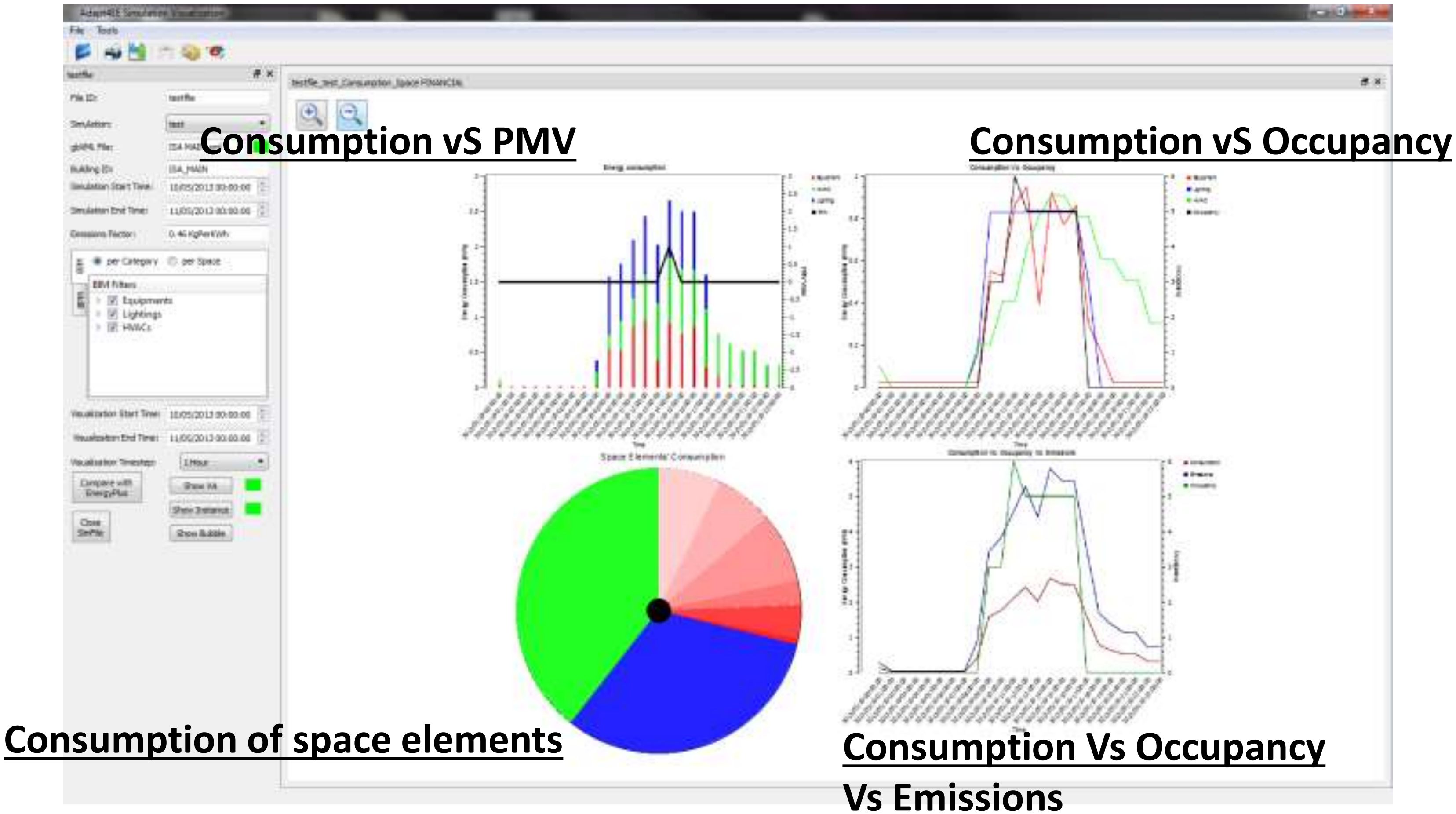




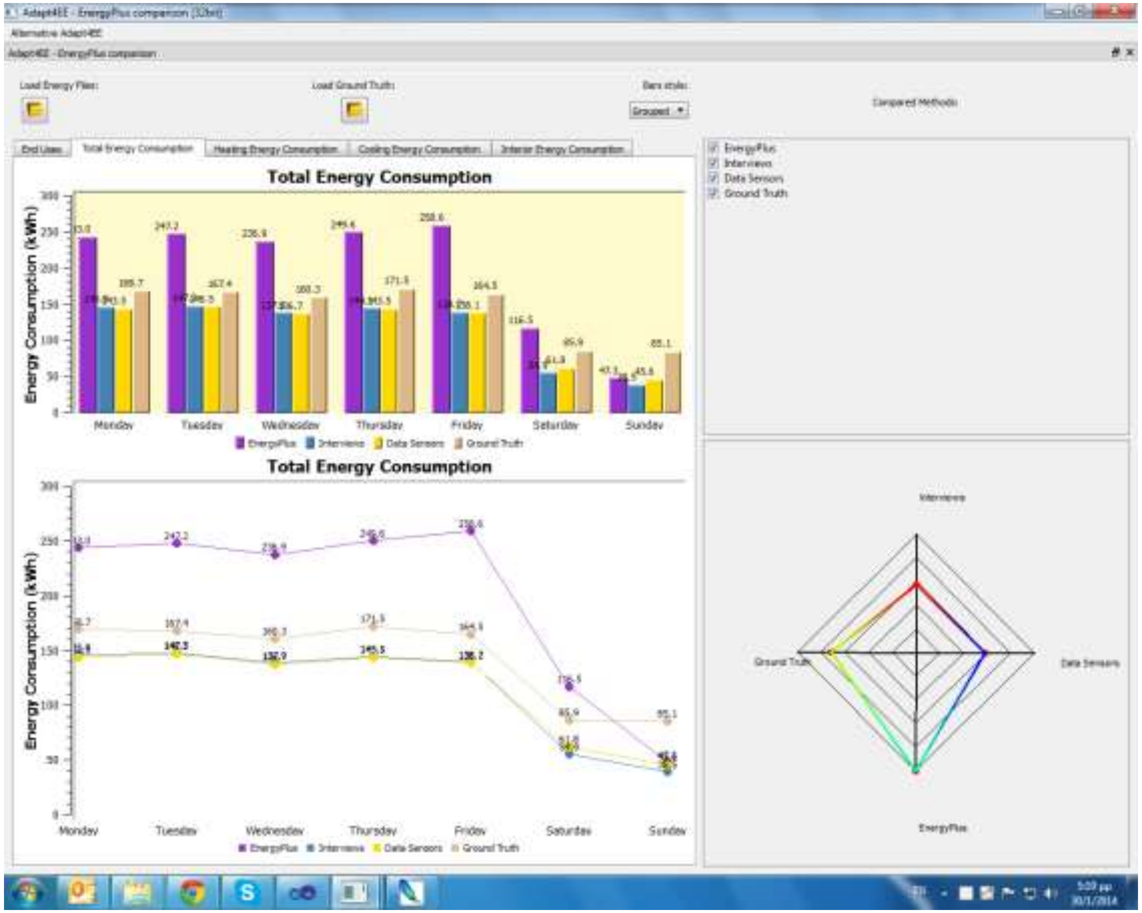
Application 9.4: **Analysis per load category**



Application 9.5: Specific KPI drill-in space level



Application 9.6: **Comparison with EnergyPlus output**



Presentation outline

1. Introduction
   - Big Data
   - Visual analytics for big data
2. Visual analytics methods developed by CERTH/ITI
   - …
   - Method 6: Graph-based descriptors for the detection and visualization of network anomalies
   - Method 7: Hierarchical Magnification for insight gain in smaller displays
   - Method 8: Energy sustainability of buildings' energy sustainability
   - Method 9: Occupancy tracking in closed spaces
3. Videos demonstration

Big Data Group

Director of ITI
(Researcher A')
Dr. D. Tzovaras

Postdoctoral
Research Fellow
Dr. A. Drosou

Research Assistant
Mr. I. Kalamaras

Research Assistant
Mr. S. Papadopoulos

Research Assistant
Mr. P. Moschonas

Relevant Publications
Journals

Published/Accepted for Publication

1. I. Kalamaras, A. Drosou, D. Tzovaras, “Accessibility-based re-ranking in multimedia search engines”, Multimedia Tools and Applications, accepted for publication
2. S. Papadopoulos, A. Drosou, D. Tzovaras “A Novel Graph-based Descriptor for the Detection of Billing-related Anomalies in Cellular Mobile Networks”, IEEE Trans. Mobile Comput., Early Access, 2016, doi: 10.1109/TMC.2016.2518668.
3. S. Papadopoulos, K. Moustakas, A. Drosou, D. Tzovaras “Border gateway protocol graph: detecting and visualising internet routing anomalies”, IET Information Security, vol. 10, no. 3, pp. 125-133, doi:10.1049/iet-ifs.2014.0525.
4. I. Kalamaras, A. Drosou, D. Tzovaras , “Multi-Objective Optimization for Multimodal Visualization”, IEEE Trans. Multimedia, vol.16, no.5, 2014, doi: 10.1109/TMM.2014.2316473.

under Review

1. I. Kalamaras, A. Zamihos, G. Margaritis, A. Drosou, D. Kehagias, A. Salamanis, D. Tzovaras, “An interactive Visual Analytics Platform for smart Intelligent Transportation Systems management”, SI: IEEE Trans. Intell. Transp. Syst., under review.
2. A. Drosou, I. Kalamaras, S. Papadopoulos, D. Tzovaras, “An enhanced Graph Analytics Platform (GAP) providing insight in Big Network Data”, Journal of Innovation in Digital Ecosystems, SI: Digital ecosystem management, under review.
3. I. Kalamaras, A. Drosou, D. Tzovaras, “A Consistency-based Multimodal Graph Embedding Method for Dimensionality Reduction”, IEEE Trans. Multimedia, under review.

Publications

Conferences 1/4

Published/Accepted for Publication

1. V. Bikos, M. Karypidou, E. Stalika, P. Baliakas, ... & P. Algara, “An Immunogenetic Signature of Ongoing Antigen Interactions in Splenic Marginal Zone Lymphoma Expressing IGHV1-2* 04 Receptors”. Clinical Cancer Research, 22(8), 2032-2040, 2016.
2. Polychronidou E., Xochelli A., Moschonas P., Papadopoulos S., Hatzidimitriou A.,, Vlamos P., Stamatopoulos K.,Tzovaras D., “Chronic Lymphocytic Leukemia patient clustering based on mutation analysis”, 2nd World Congress on Genetics, Geriatrics and Neurogenerative Diseases Research (GeNeDis), 2016.
3. A. Drosou, N. Dimitriou, N. Sarris, A. Konstantinidis, D. Tzovaras, “Research directions for harvesting cross-sectorial correlations towards improved policy making”, Data for Policy 2016, to appear.
4. I. Kalamaras, S. Papadopoulos, A. Drosou, D. Tzovaras “MoVA: A Visual Analytics tool providing insight in the Big Mobile Network Data”, The 11th International Conference on Artificial Intelligence Applications and Innovations (AIAI'15), vol. 458, pp. 383-396, doi:10.1007/978-3-319-23868-5_27.
5. S. Papadopoulos, A. Drosou, D. Tzovaras, “Fast Frequent Episode Mining based on Finite-State Machines”, 30th International Symposium on Computer and Information Sciences (ISCIS), Volume 363 of the series Lecture Notes in Electrical Engineering pp. 199-208, 2015, doi:10.1007/978-3-319-22635-4_18.

Information Technologies Institute

Publications

Conferences 2/4

6. S. Papadopoulos, A. Drosou, N. Dimitriou, O. Abdelrahman, G. Gorbil, D. Tzovaras “A BRPCA based approach for anomaly detection in mobile networks”, 30th International Symposium on Computer and Information Sciences (ISCIS), Volume 363 of the series Lecture Notes in Electrical Engineering, pp. 115-125, 2015, doi:10.1007/978-3-319-22635-4_10.
7. I. Kalamaras, A. Drosou, D. Tzovaras, “A multi-objective approach for the clustering of abnormal behaviours in mobile networks”, IEEE International Conference in Communications Workshop (ICCW), pp.1491-1496, 2015, doi: 10.1109/ICCW.2015.7247390.
8. L. Sutton, P. Moschonas, A. Vardi, V. Bikos, X. Yan, M. Chatzouli, A. Anagnostopoulos, C. Belessi, N. Chiorazzi, R. Rosenquist, D. Tzovaras, K. Stamatopoulos, A. Hadzidimitriou, "Matched Pattern Discovery across Paired Immunoglobulin Heavy and Light Chains in CLL Reveals Unique Subset-defining Amino Acid Associations", Immune Profiling in Health and Disease, Nature, Adaptive Biotechnologies, September 9th-11th, 2015, Seattle, WA, USA.
9. E. Polychronidou, A. Xochelli, P. Moschonas, A. Hadzidimitriou, Pa. Vlamos, K. Stamatopoulos, D. Tzovaras, "An informatics probabilistic method for pattern discovery in immunoglobulin amino acid sequences", In Proceedings of the of the 10th Hellenic Society for Computational Biology & Bioinformatics (HSCBB15), Athens, Greece, October 9th-11th, 2015.
10. D. Ioannidis, A. Fotiadou, S. Krinidis, G. Stavropoulos, D. Tzovaras and S. Likothanassis, “Big Data & Visual Analytics for Building Performance Comparison”, 11th International Conference on Artificial Intelligence Applications and Innovations (AIAI'15), Bayonne/Biarritz, France, 14-17 September 2015.

Publications

Conferences 3/4

11. S. Papadopoulos, V.Mavroudis, A. Drosou, D. Tzovaras, “Visual Analytics for enhancing supervised attack attribution in mobile networks”, Information Sciences and Systems, pp 193-203, 2014, doi:10.1007/978-3-319-09465-6_21.
12. G. Stavropoulos, S. Krinidis, D. Ioannidis, K. Moustakas and D. Tzovaras, “A Building Performance Evaluation & Visualization System”, IEEE International Conference on Big Data (BigData’14), pp. 1077-1085, Washington DC, USA, 27-30 October 2014.
13. S. Papadopoulos, K. Moustakas, D. Tzovaras, “BGPViewer: Using Graph representations to explore BGP routing changes”, 18th International Conference on Digital Signal Processing (DSP), 1-3 July 2013.
14. S. Papadopoulos, K. Moustakas, D. Tzovaras, “Hierarchical Visualization of BGP Routing Changes Using Entropy Measures”, 8th International Symposium on Visual Computing, July 16-18, 2012.
15. Kalamaras, I., Mademlis, A., Malassiotis, S., Tzovaras, D., “A novel framework for multimodal retrieval and visualization of multimedia data”, Signal Processing, Pattern Recognition and Applications / 779: Computer Graphics and Imaging (SPPRA), 2012.

Publications

Conferences 4/4

under Review

1. S. Papadopoulos, A. Drosou, D. Tzovaras, “A Hierarchical Magnification Approach for enhancing the Insight in Data Visualizations”, in Proc. of the International Conference on Information Visualization Theory and Applications (IVAP 2016), under review.
2. S. Papadopoulos, A. Drosou, D. Tzovaras, “A Hierarchical Scale-and-Stretch Approach for Image Retargeting”, in Proc. of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP 2016), under review.
3. S. Papadopoulos, A. Drosou, I. Kalamaras, D. Tzovaras, “A Multi-Objective Behavioral Clustering Approach using Graph-based Features”, IEEE ICC Communications and Information Systems Security Symposium (CISS), under review.

Information Technologies Institute

Contact Details: Dr. *Dimitrios Tzovaras*
dimitrios.tzovaras@iti.gr

Centre of Research & Technology - Hellas
Information Technologies Institute
6th km Xarilaou - Thermi, 57001, Thessaloniki, Greece