ICMC SÃO CARLOS USP
Instituto de Ciências Matemáticas e de Computação
| Universidade de São Paulo |

VISUALIZAÇÃO

Módulo IV

Rosane Minghim

Escola Avançada de *Big Data Analysis*, Julho 2017

Outline

- Part I – Concepts and Techniques

Rosane Minghim

- Part II – Applications

Post-grad students CCMC - ICMC

- Part III – Similarity-Based Visualization

Emilio Vital Brazil - IBM

- Part IV – Visualization Case Studies

Vagner Santana – IBM

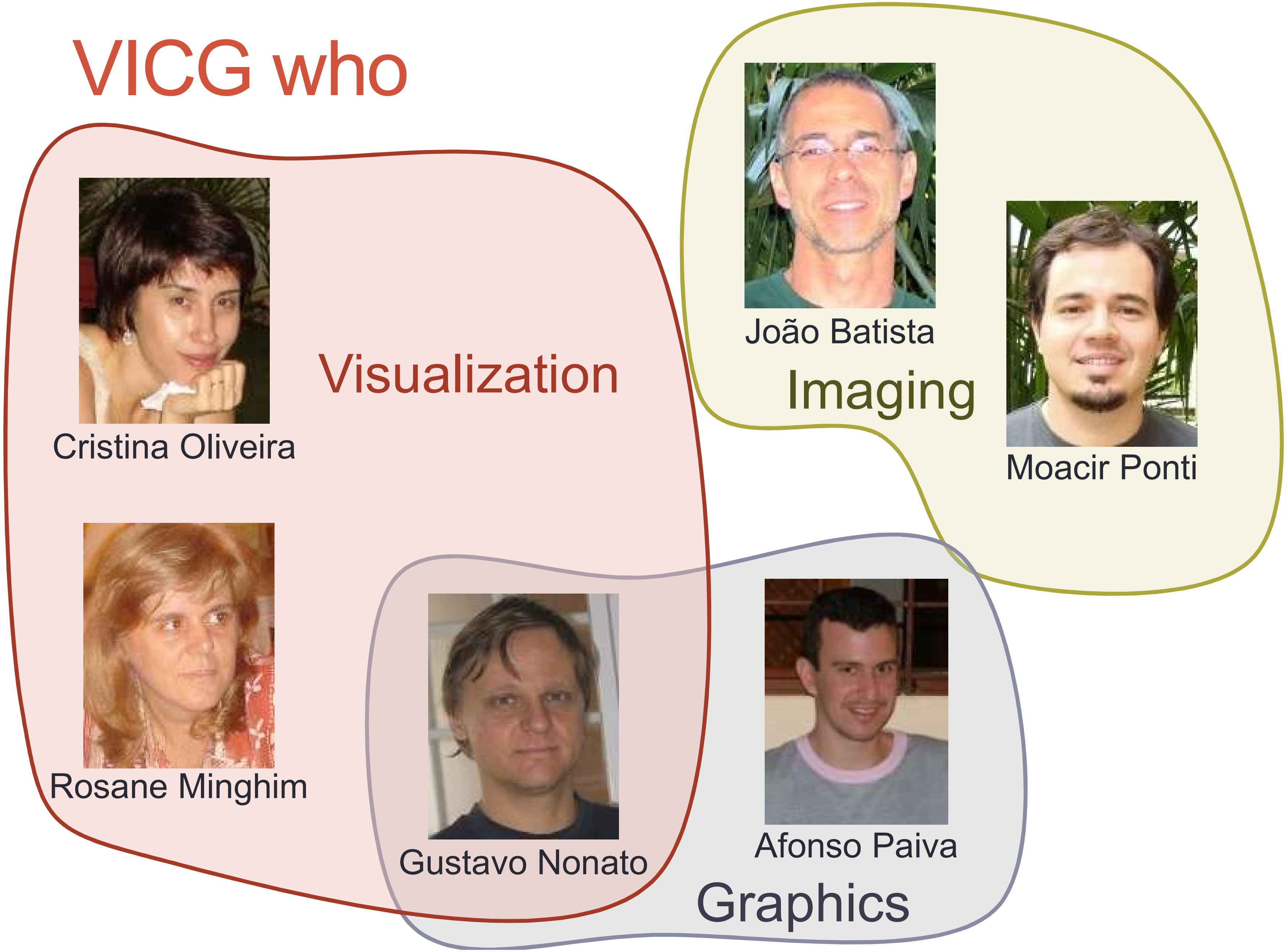
Outline Part I

- Context and Resources
- Data
- Point-centered Visualization
- Attribute-centered Visualization
- Relationship-centered Visualization
- Visualization and Mining

Our research at VICG Visualization, Imaging and Computer Graphics Lab vicg.icmc.usp.br

- Visual Data Mining
- Visual Analytics
- Visualization

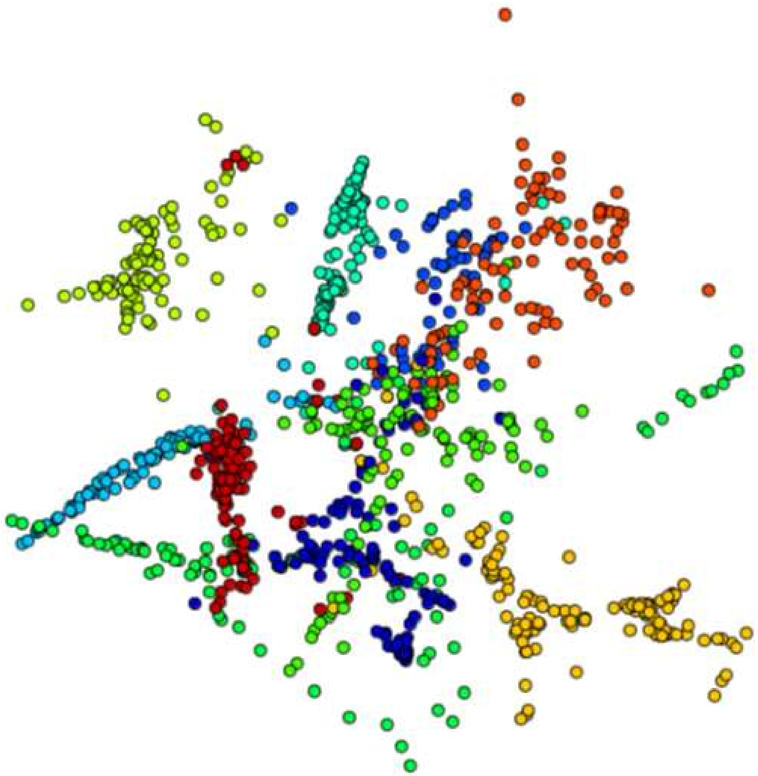
VICG who

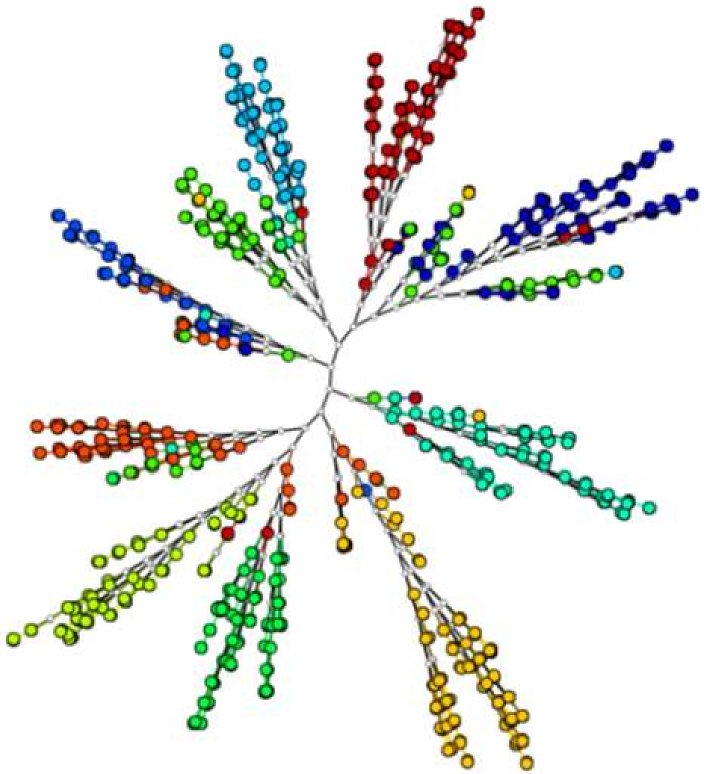

João Batista
Imaging
Moacir Ponti
Visualization
Cristina Oliveira
Rosane Minghim
Gustavo Nonato
Afonso Paiva
Graphics

Motivation

- Context: Abstract Data Analysis
  - Why Visualize?
  - How not visualize?
  - When Visualize?
    - Analysis, Illustration, Demonstration

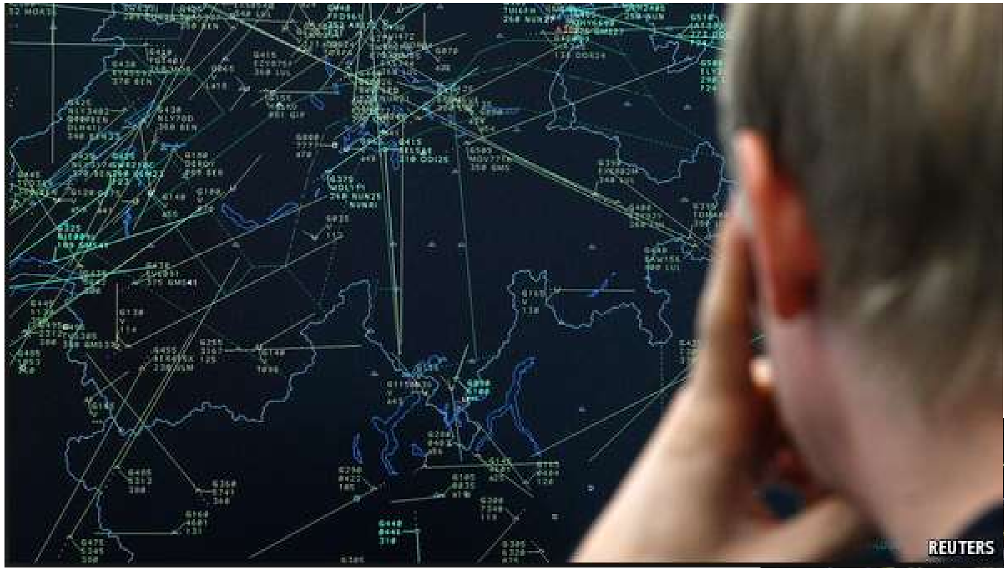
Classification Results - Why





Classification Results - Why

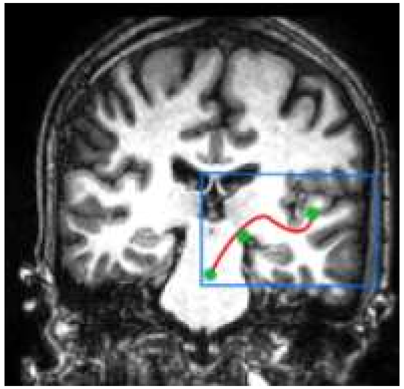
Classification Results – How not

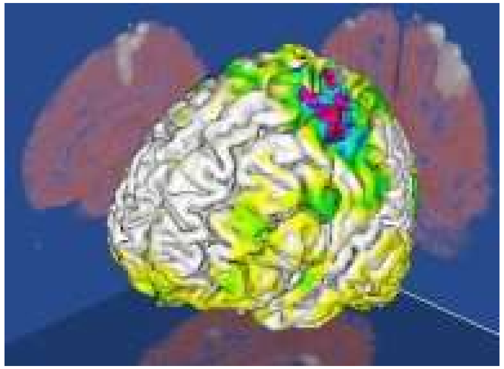


FACET

Before we continue...

Scientific Visualization





http://www.cs.rug.nl/svcg/SciVis/DTIVis

- Medicine
- Simulation
- Engineering
- Etc.

Fonte: Poco et al. 2012

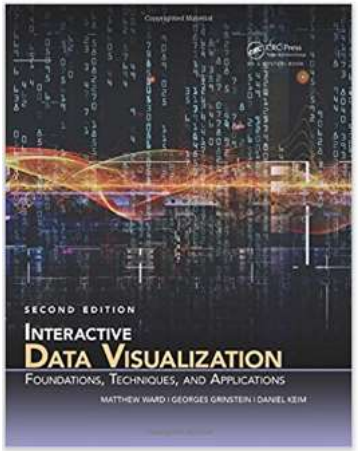
Visualization for Data Science and Big Data What does it take?

- Algorithms
- Statistics – essential
  - Alone will not do the job
- Mining – essential
  - Will not do the whole job, even with statistics
- Visualization – exploratory situations and user centric decision
- Certain skills – from complex reasoning to complete programming to innovative and daring goals. But mostly: Undestand the data

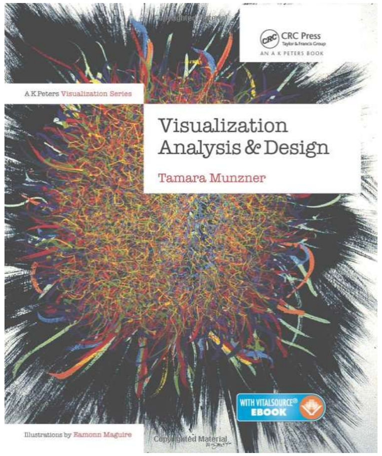
Companies all around the world know the value of Big Data

See, for instance:

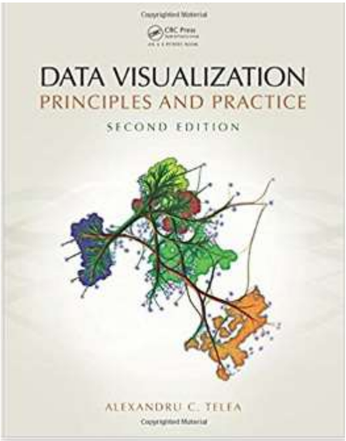
Resources - books



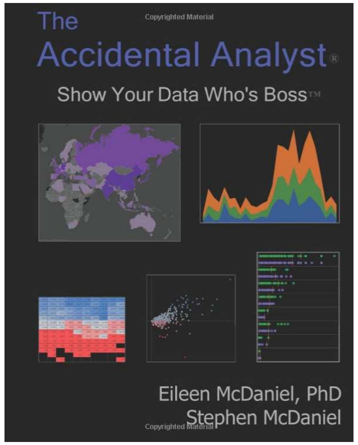
Ward, Grinstein, Keim, 2015



Munzner, 2014



Telea, 2014



McDaniel & McDaniel, 2012

Resources - software

- Most data analytics tools have some sort of visualization capability.
  - Phyton
  - R
  - Javascript (d3.js)
  - VTK + (3D scientific)
  - TensorFlow
  - etc..

  - Watson™
  - Tableau™
  - etc..

Links to sources of data visualization tools and data

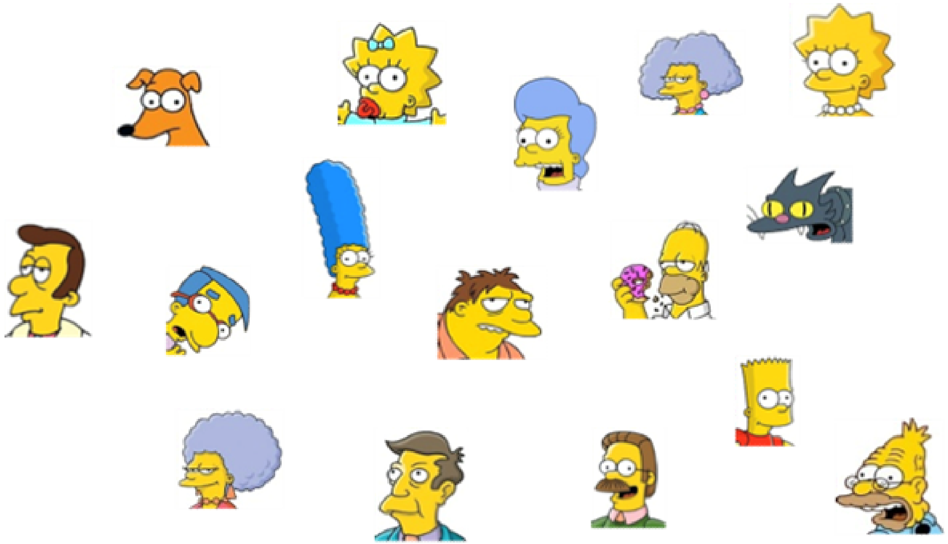
- HDR (ONU):
  - (data) http://hdr.undp.org/en/composite/GII
  - (vis) http://hdr.undp.org/en/data-explorer/

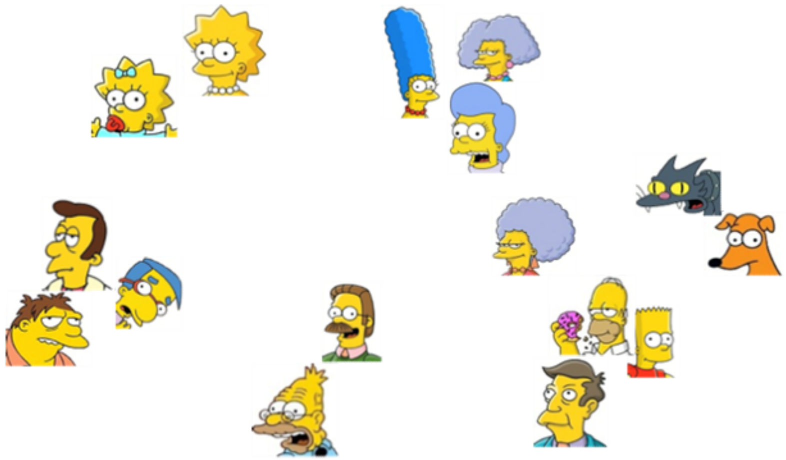
- D3:
  - https://d3js.org/
  - (gallery) https://github.com/mbostock/d3/wiki/Gallery/

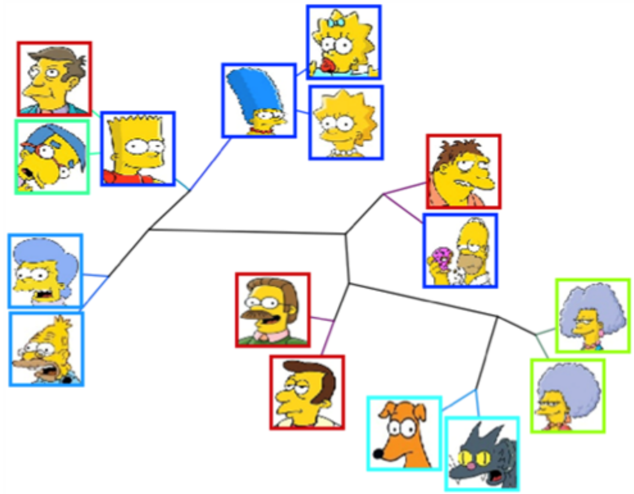
What does your data tell

- People trying to make sense of data

´messy´ data







Data

- Item
- Attribute
- Reference (index, position, location)
- Relationship
- Collection (data set/table, item, network)

Techniques

- Point placement: 2D or 3D similarity-based layouts

15														
6	2													
15	1	2												
15	1	2	1											
15	2	1	2	2										
15	3	10	8	15	15									
15	10	7	10	10	7	15								
15	15	10	15	15	10	15	16							
15	15	10	10	10	10	15	16	16						
15	15	15	15	15	15	15	15	3	15					
15	10	10	10	10	10	15	15	15	15	15				
15	10	10	10	10	10	15	15	15	1	15	15			
15	10	15	12	15	15	10	15	15	15	15	15	15		
15	10	10	10	10	10	15	15	15	15	15	2	8	15	
15	10	10	10	10	10	15	1	15	15	15	15	15	15	15

pairwise distances

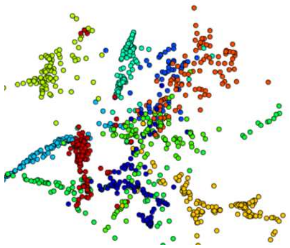
and/or

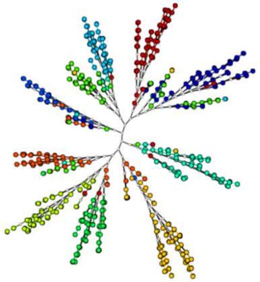
5	12	15	2	7	5	0	12	9	0	8
12	5	0	12	12	12	12	12	18	12	12
0	1	05	10	15	12	8	12	9	11	5
0	12	01	12	9	0	12	10	5	5	12
12	8	05	12	12	12	8	12	9	12	12
10	12	0	11	10	2	7	12	2	16	7
5	6	8	12	12	15	12	6	9	17	0
7	12	05	0	12	12	10	17	9	12	12
2	10	05	15	12	1	12	10	9	8	2
12	12	7	12	0	12	0	12	10	12	12
6	12	05	17	12	10	12	12	9	12	8
12	10	2	12	1	12	12	11	6	0	12
1	12	05	12	12	16	2	12	9	12	0
10	0	12	12	9	12	0	10	12	12	8
0	12	1	12	12	5	1	7	11	12	12
8	2	11	10	7	12	5	12	15	10	0

dimensional embedding
(feature space)

Visualizations

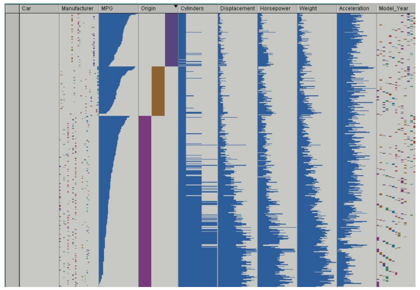
Point-Based

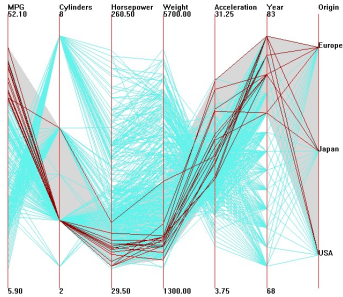




source : Paiva et al.

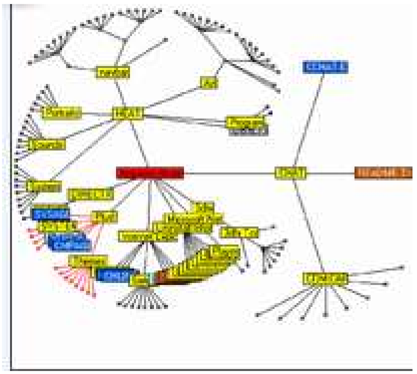
Attribute Based

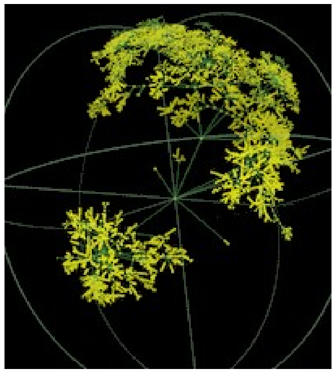




source: Ward et al.

Relationship Based





sources : treevis.net e www.caida.org

Geo

sources: Google Maps e CartoDB

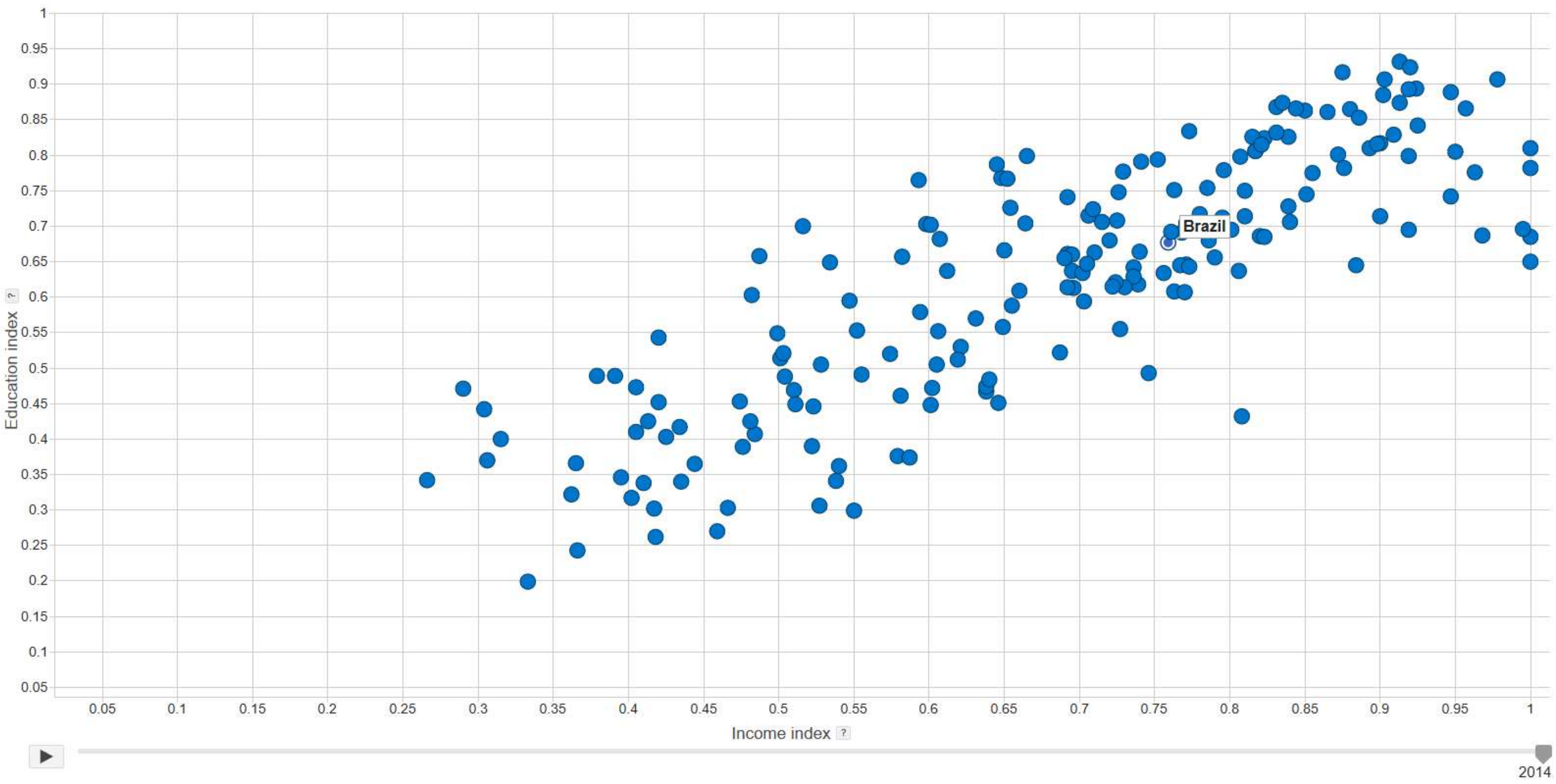
POINT-BASED TECHNIQUES

Scatterplots

Multidimensional projections

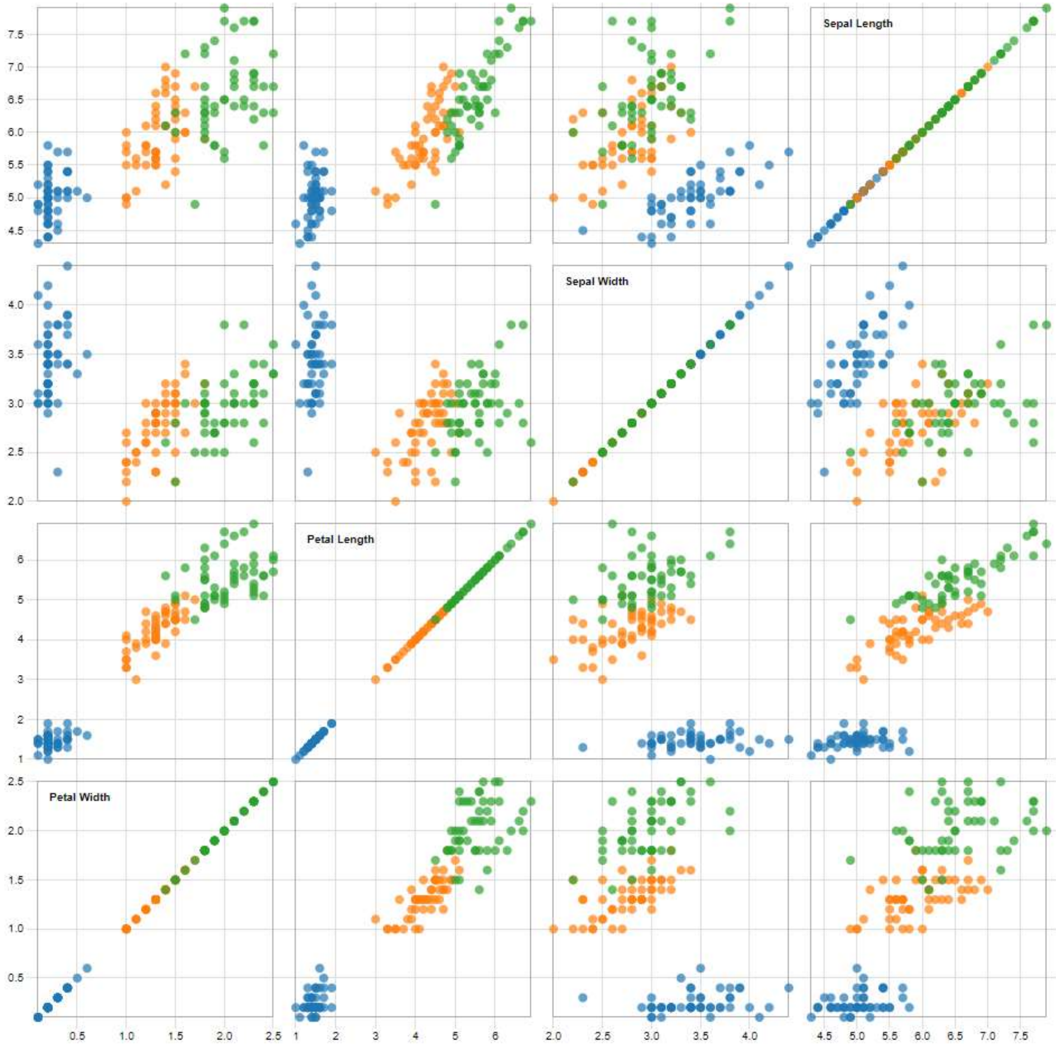
Similarity trees

Scatter Plots



Fonte: HDR (ONU)

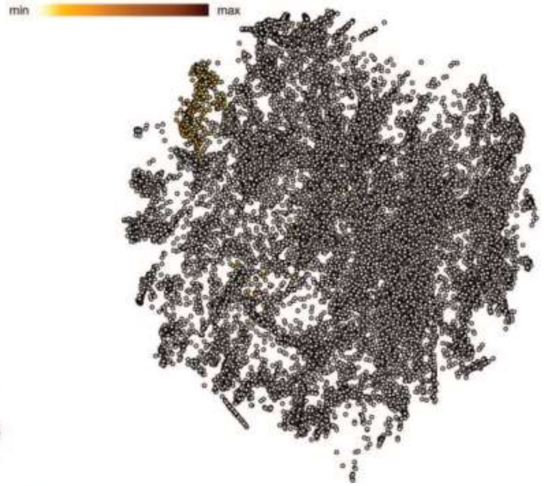
Technique: Scatter Plot Matrix



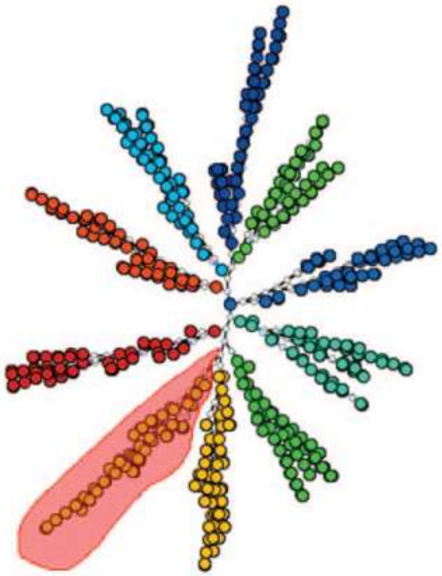
https://bl.ocks.org/mbostock/4063663

Advanced Visual Data Analysis via Point Placement and Dimension Reduction

- Projection-based
  - variations on MDS or other dimension reduction approaches to map data to visual space

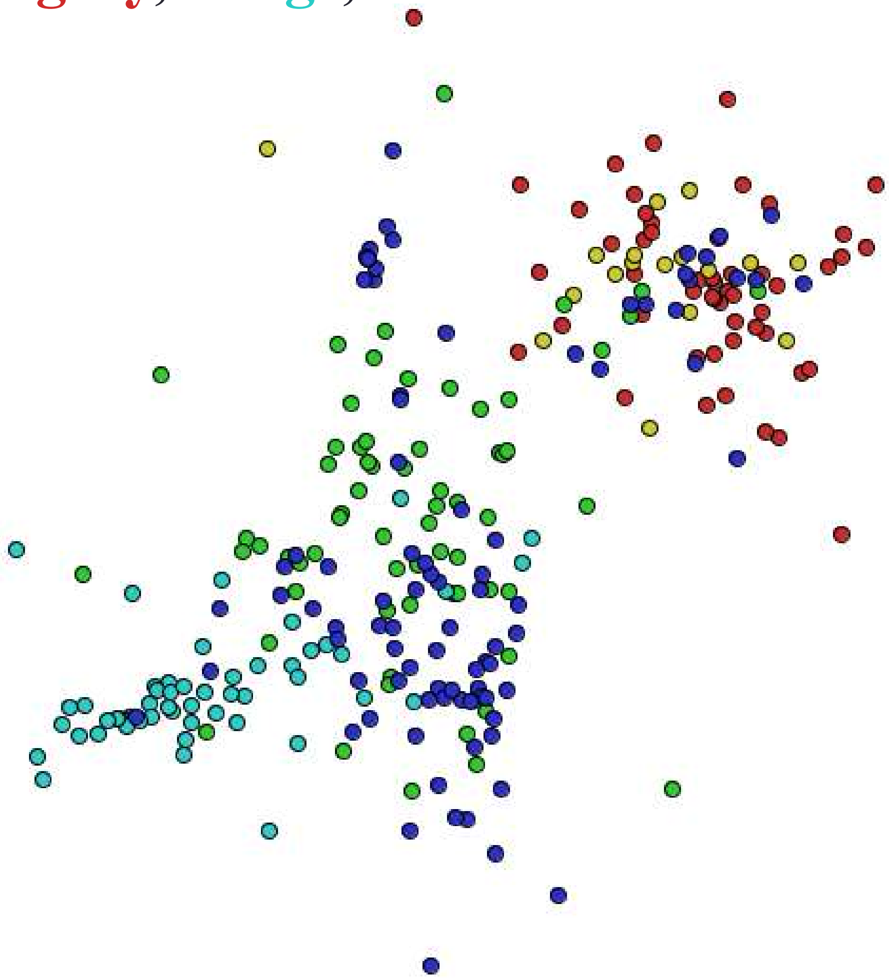


- Tree-based
  - hierarchy of similarity relations
  - variations on tree layouts



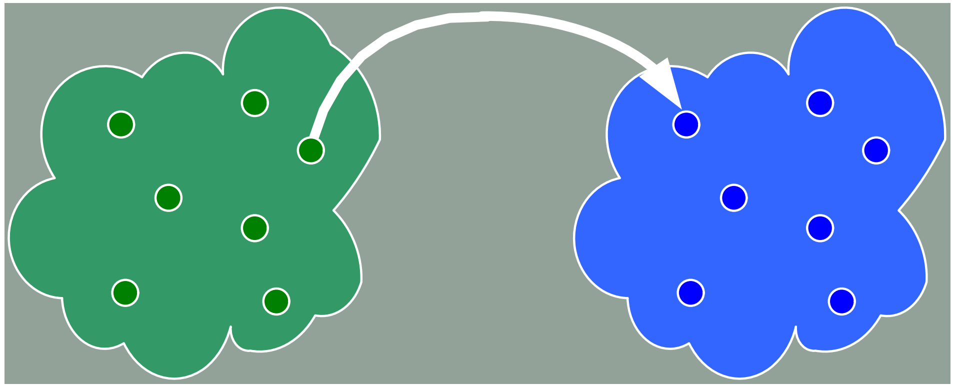
Projection Techniques:
Mapping data set on the plane, allowing direct exploration

Ex: Patents **surgery**, **drugs**, **molecular bio**



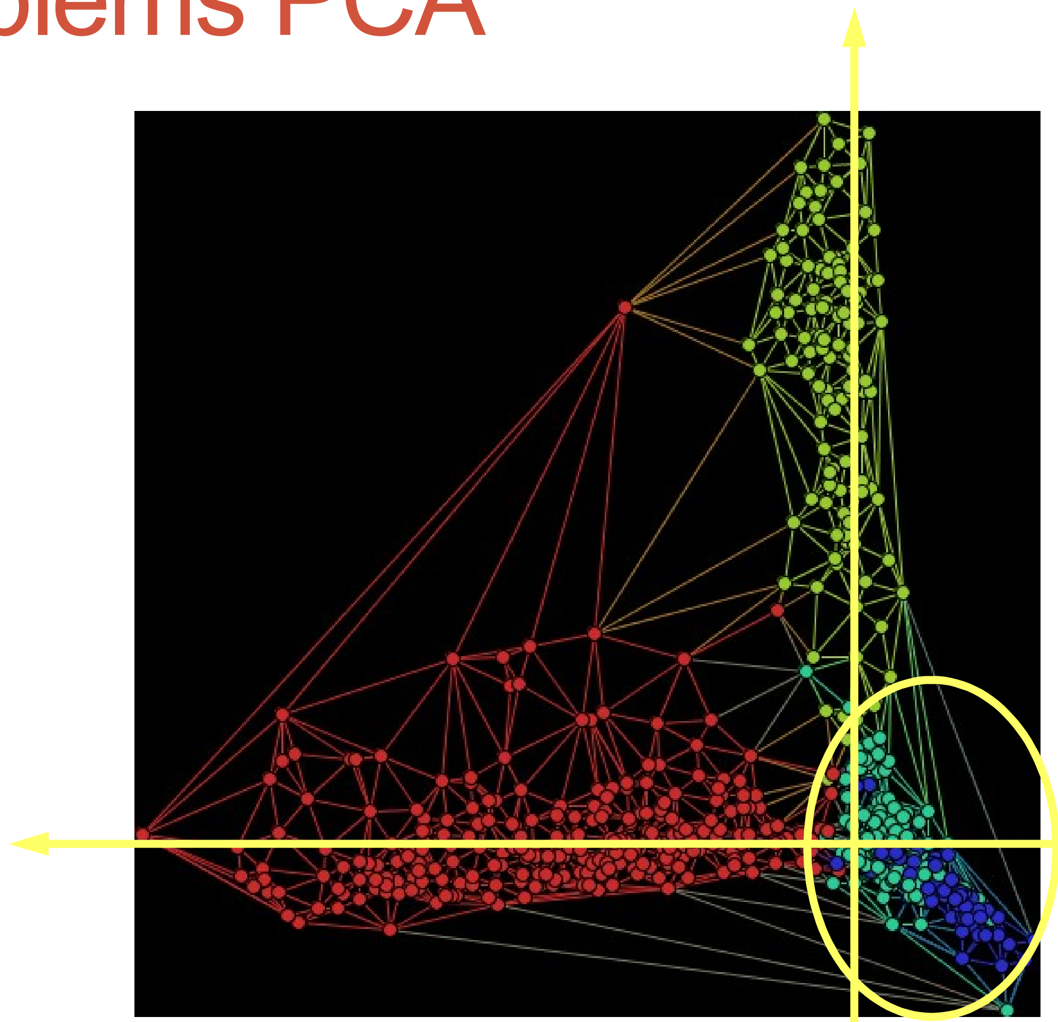
Projection Techniques

$X \in \mathcal{R}^m \quad f \quad Y \in \mathcal{R}^{p=\{1,2,3\}}$



- $\delta: x_i, x_j \rightarrow R,\ x_i, x_j \in X$
- $d: y_i, y_j \rightarrow R,\ y_i, y_j \in Y$
- $f: X \rightarrow Y, |\delta(x_i, x_j) - d(f(x_i), f(x_j))| \approx 0,\ \forall\ x_i, x_j \in X$

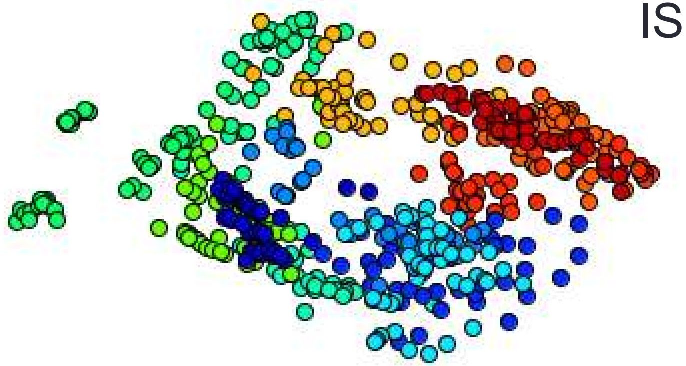
Problems PCA

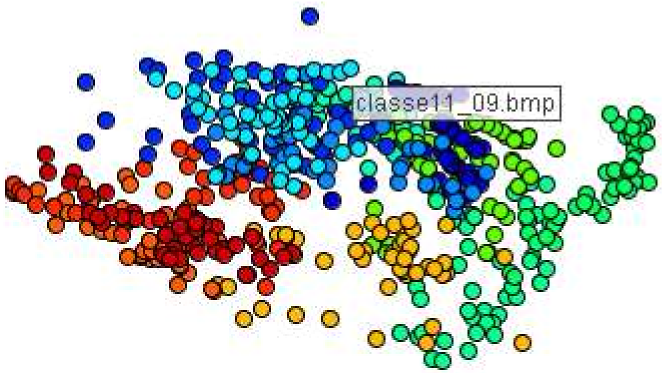


Projection as dimension reduction

- Classical
  - PCA (Principal Component Analysis)
  - Classical Scaling
  - LLE (Local Linear Embedding)
  - ISOMAP
  - Sammon´s mapping

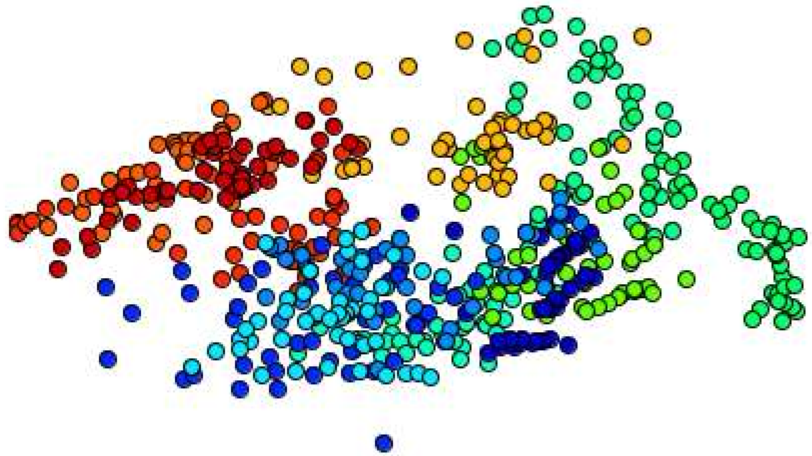
ISOMAP



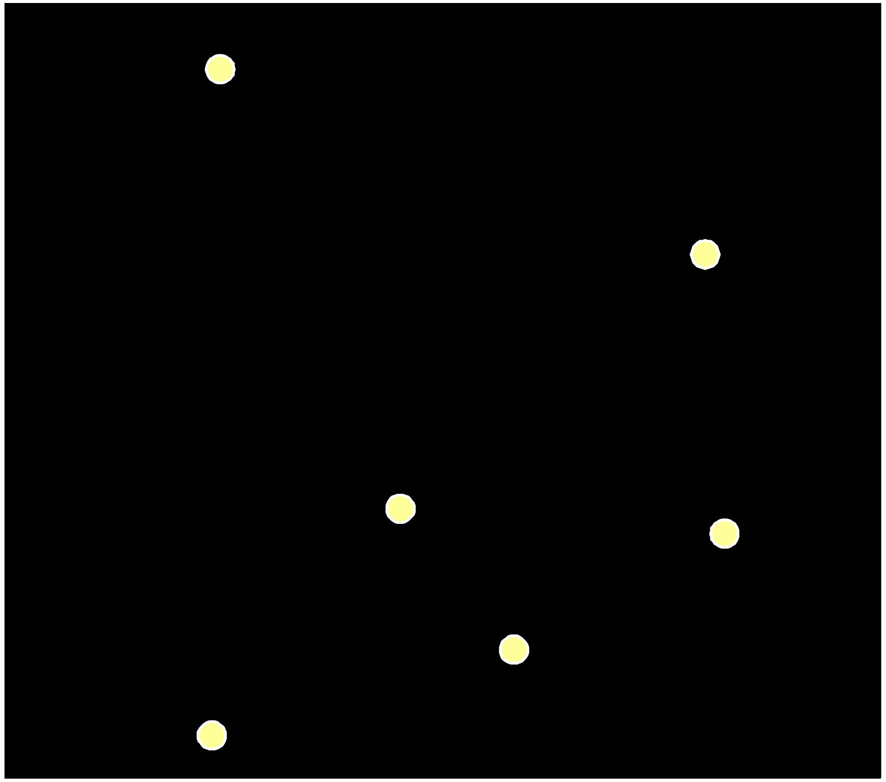


Classical Scaling

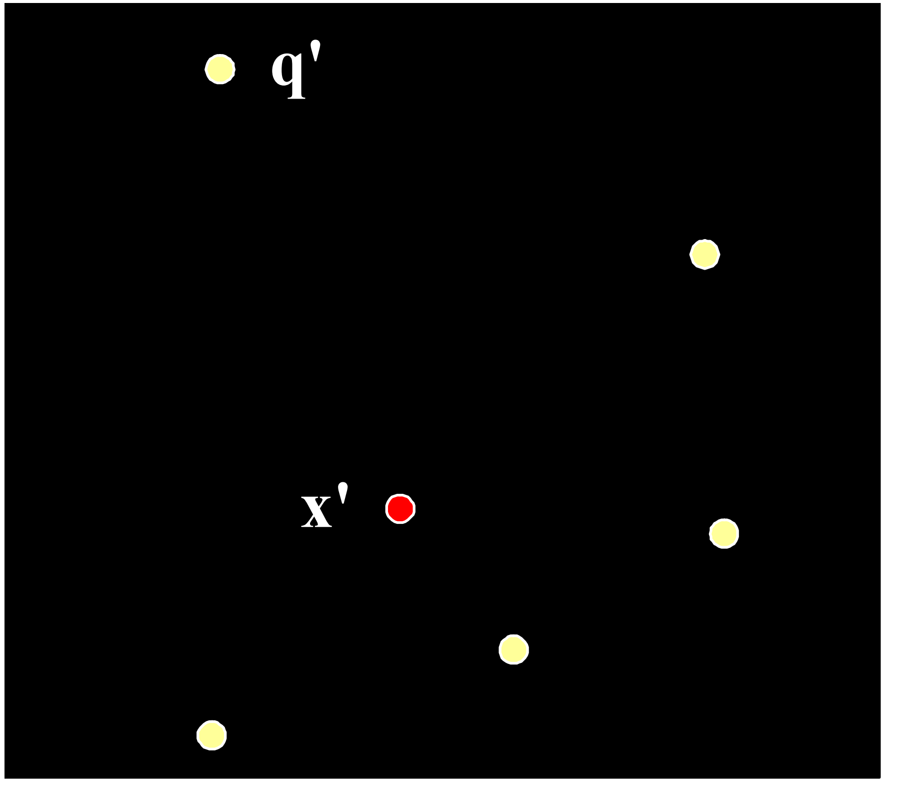
PCA



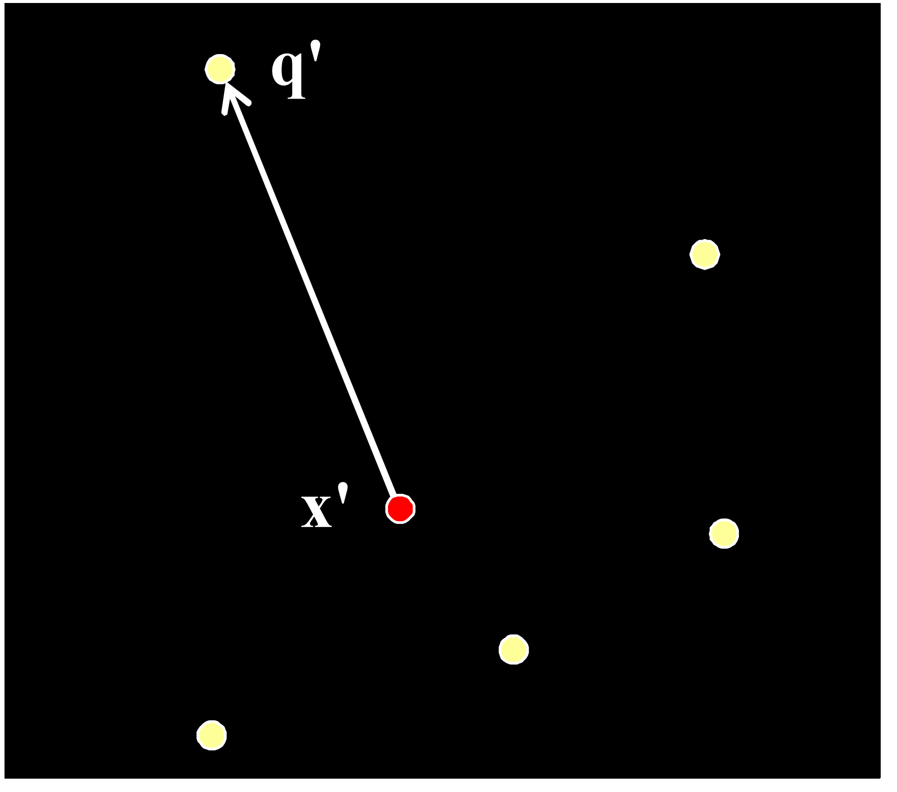
Force Based Point Placement



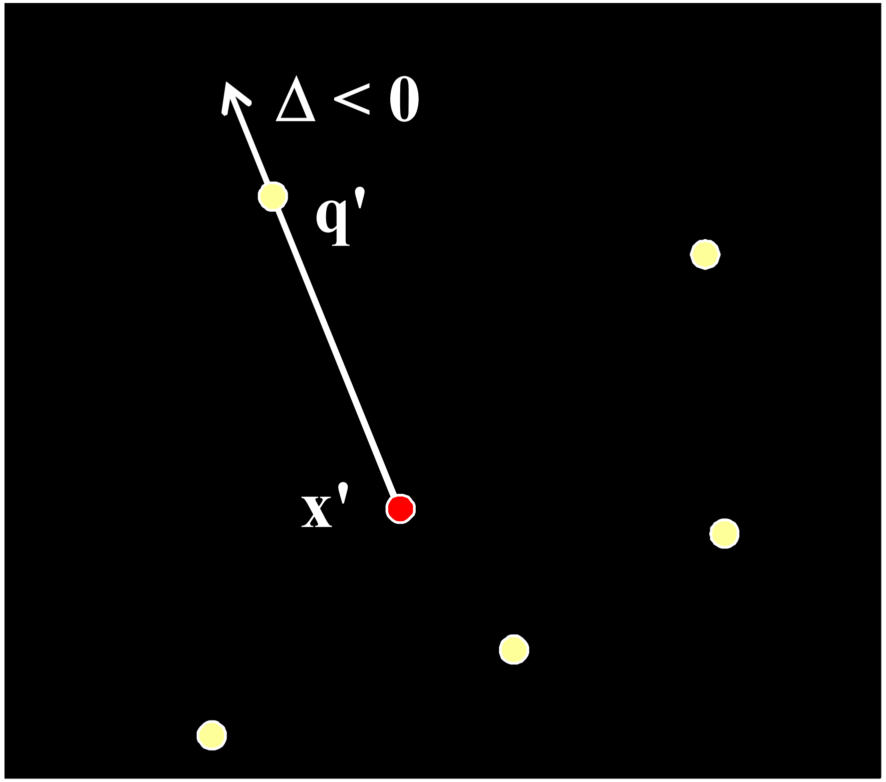
Force Scheme [Tejada et al., 2003]



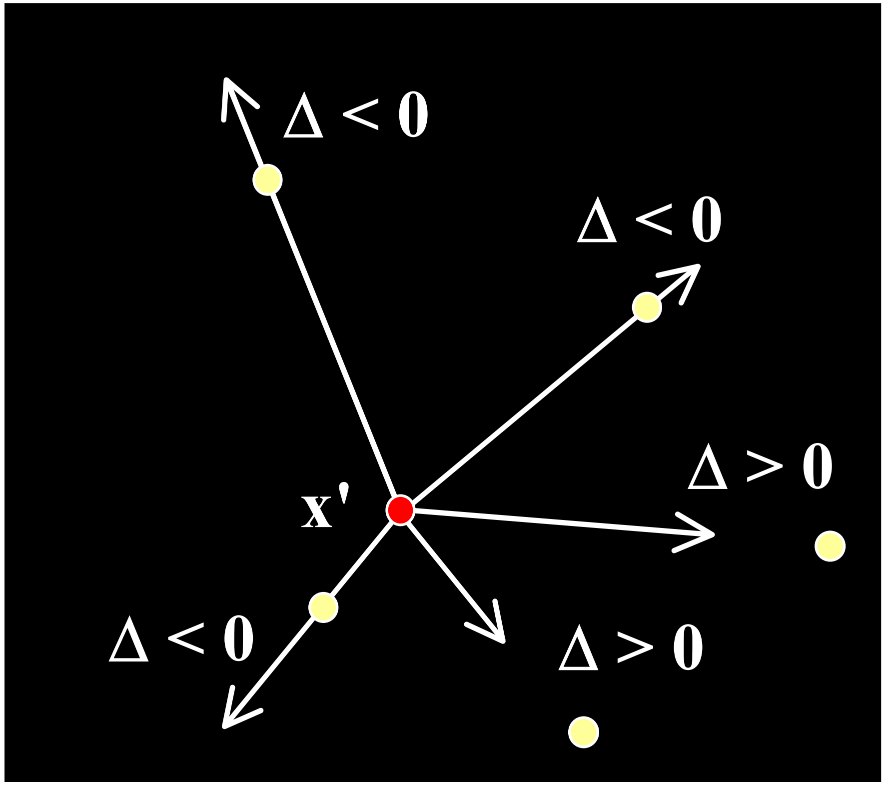
Force Scheme [Tejada et al., 2003]



Force Scheme [Tejada et al., 2003]



Force Scheme [Tejada et al., 2003]



Force Scheme [Tejada et al., 2003]

1. Map each point X to the plane (fastmap, nnp, etc.)
2. For each projected point x
    1. For each projected point q’ ≠ x’
        1. Compute the vector **v** of <x’ to q’>
        2. Move q’ in direction of **v**, one fraction of Δ

$$\Delta = \frac{\delta(x,q) - \delta_{\min}}{\delta_{\max} - \delta_{\min}} - d(x',q')$$

3. Normalize the coordinates between [0,1]

LSP [Paulovich et al., 2006/2008]

- Least-Square Projection (LSP)
- Core idea: project a sub-set of points and interpolate the rest.
- Interpolation seeks to preserve the neighborhood between points.
- Each point is mapped within the convex hull of its neighbors.

LSP [Paulovich et al., 2006/2008]

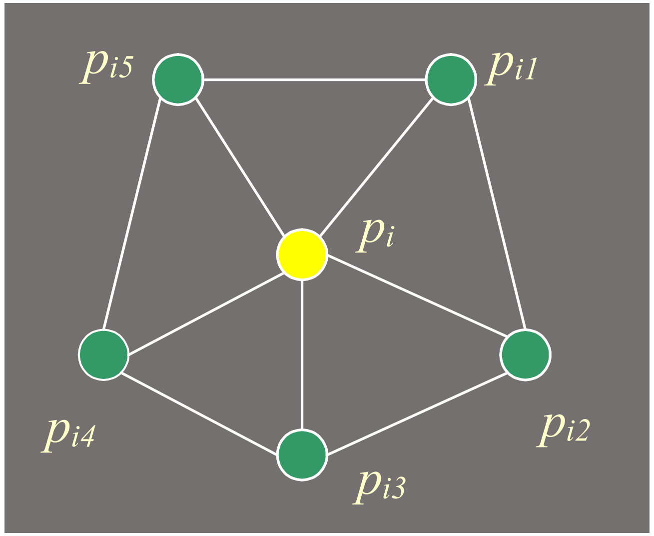
- Three main steps:

1. Select a subset of points(control points) and Project these in R^p
2. Determine the neighborhood of points
3. Create a linear system whose answers are the Cartesian coordinates of points p_i in R^p

LSP: Laplacian Matrix

- Let $V_i = \{p_{i1},\ldots,p_{iki}\}$ *be the neighborhood of a point* p_i *and* c_i the coodinates of p_i in R^p

$$c_i - \frac{1}{ki}\sum_{p_j \in V_i} c_j = 0$$



- Each p_i will be the centroid of points in V_i

LSP: Laplacian Matrix

$L\mathbf{x}_1=0$, $L\mathbf{x}_2=0$, …, $L\mathbf{x}_p=0$

where $x_1, x_2, \ldots, x_p$ are vectors containing the Cartesian coordinates of the points

and L is the matrix defined by:

$$Lij = \begin{cases} 1 & i = j \\ -\dfrac{1}{ki} & p_j \in V_i \\ 0 & otherwise \end{cases}$$

$$\begin{pmatrix} \boxed{\quad L \quad} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \ldots \\ 0 \\ 0 \end{pmatrix}$$

LSP: Adding control points

$$A = \begin{pmatrix} L \\ C \end{pmatrix} \qquad Cij = \begin{cases} 1 & p_j \text{ is a control point} \\ 0 & otherwise \end{cases}$$

$$b_i = \begin{cases} 0 & i \le n \\ x_{p_{c_i}} & n < i \le n + nc \end{cases}$$

$$\begin{pmatrix} & & \mathrm{L} & & \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ c_1 \\ c_2 \end{pmatrix}$$

LSP: Solving the system

- It is necessary to solve $A\mathbf{x} = \mathbf{b}$
- The system is solved by using least squares

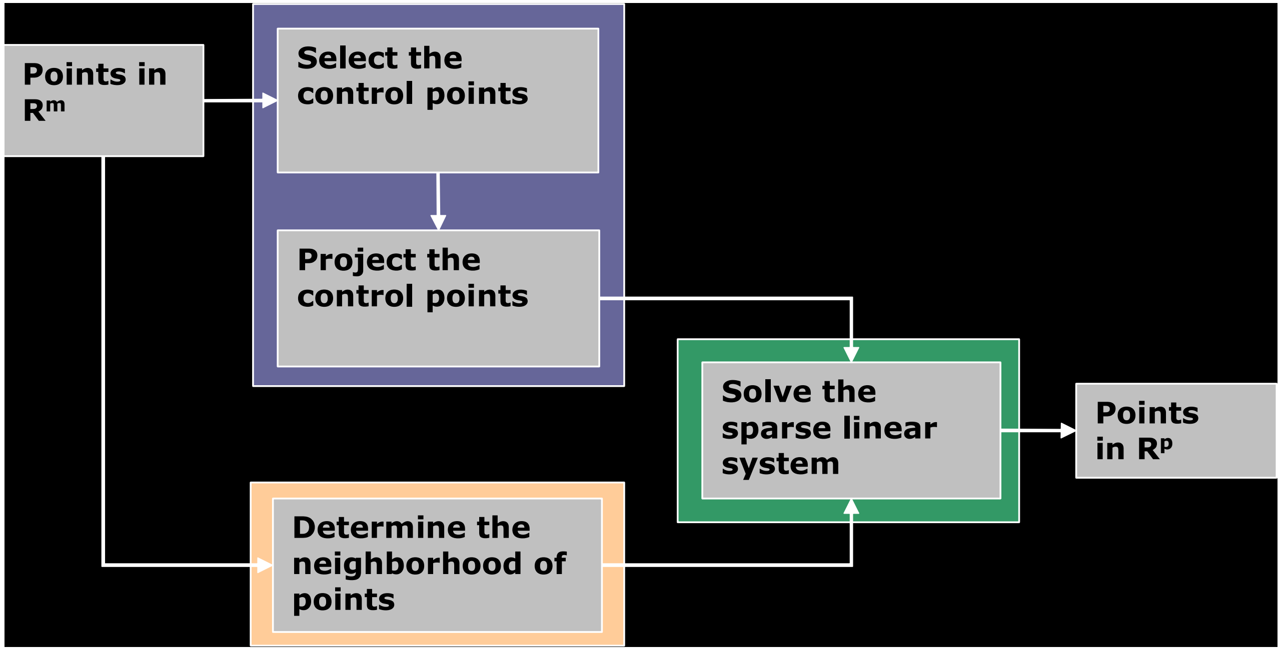
$$\left\| Ax - b \right\|^2$$

- The analytical solution is

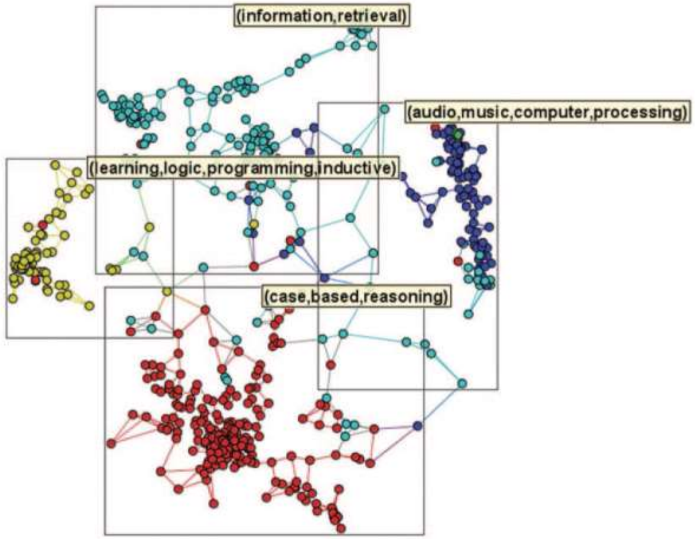
$$A^T A\mathbf{x} = A^T\mathbf{b} \quad \Rightarrow \quad \mathbf{x} = (A^T A)^{-1} A^T\mathbf{b}$$

- A^TA is symmetric and sparse and can be solved using the factorization of Cholesky

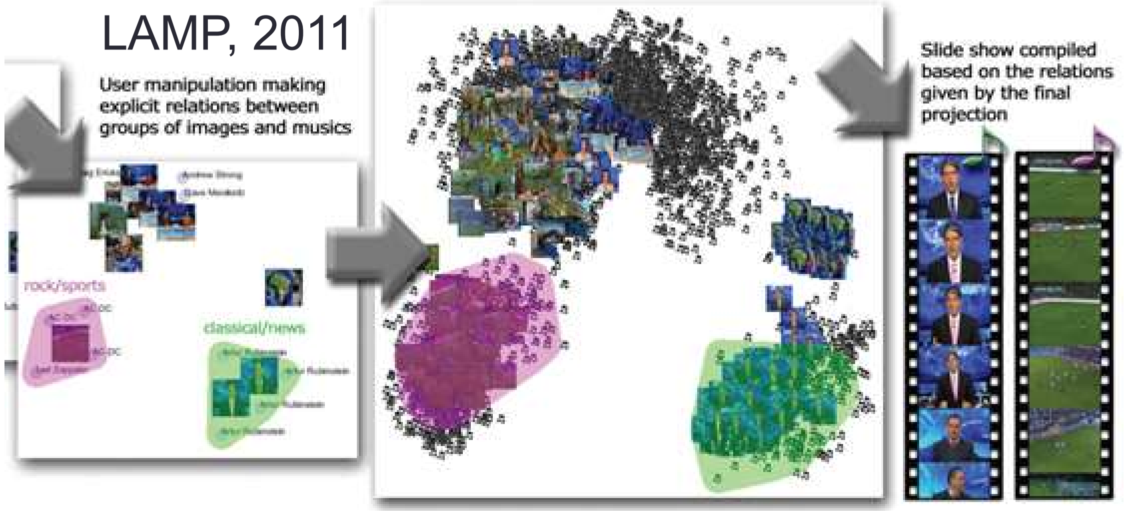
LSP: Overview



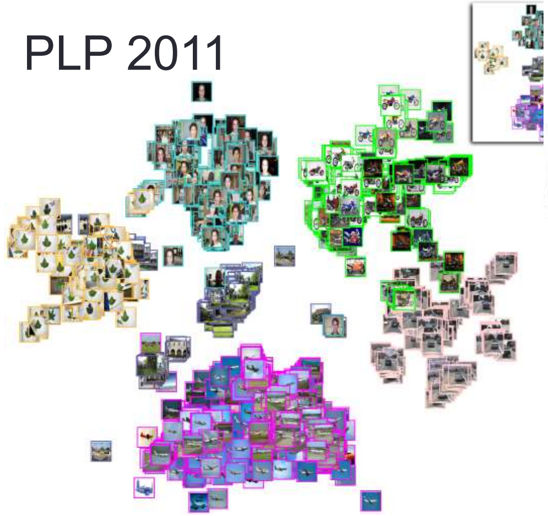
LSP, 2008



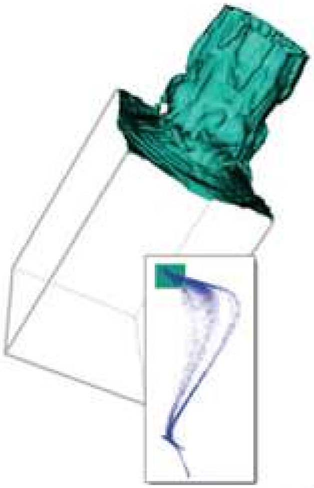
LAMP, 2011



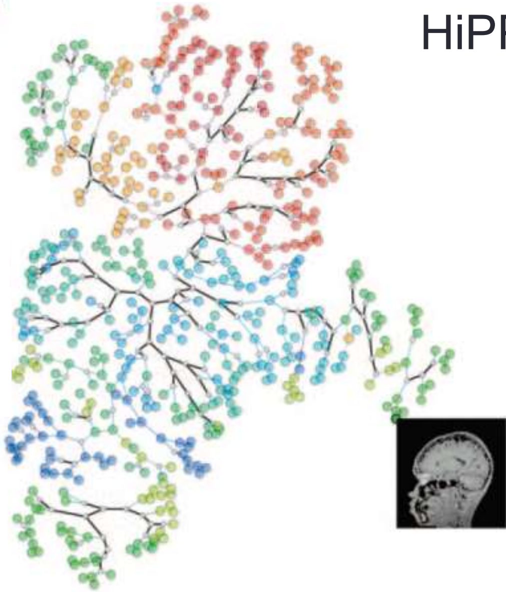
PLP 2011



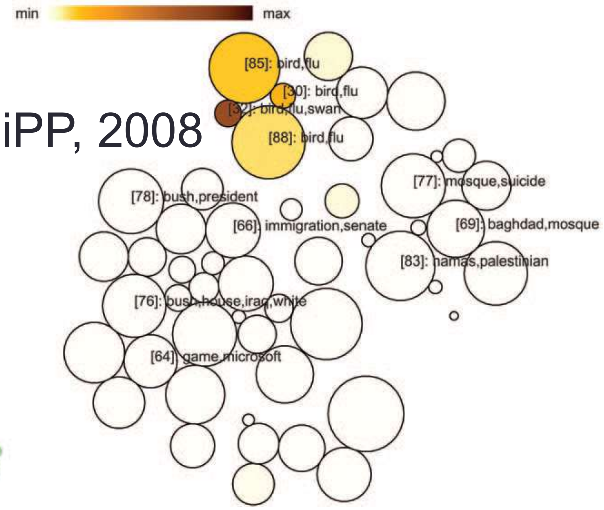
PLMP, 2010



NJ trees, 2007,2011



HiPP, 2008



Stochastic Neighborhood Embedding sne and t-sne

- Distance in original space

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)},$$

- Distance in projected space

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

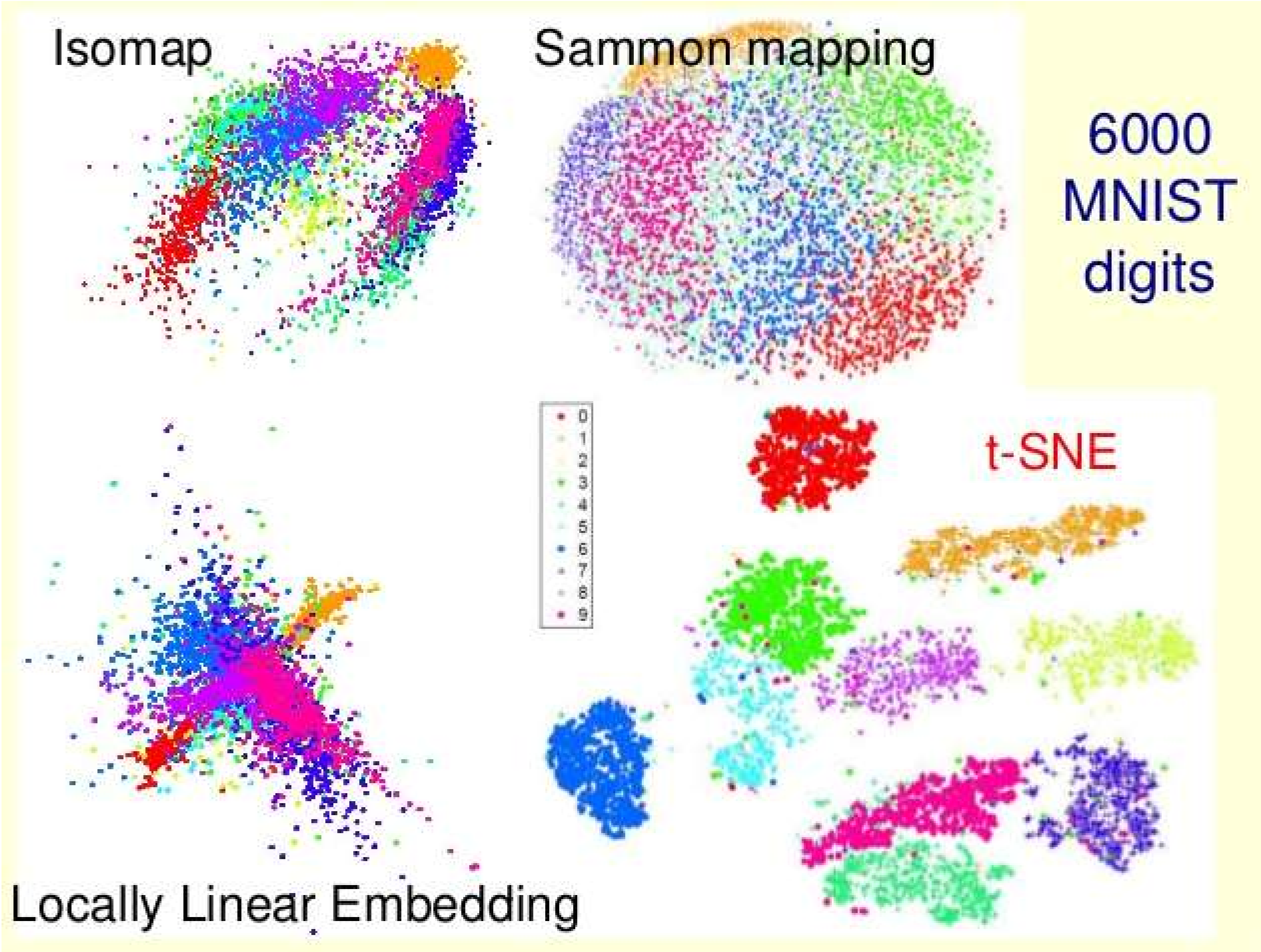
- Cost function

$$C = \sum_i KL(P_i || Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

- Non-gaussian neighborhoods: t-sne

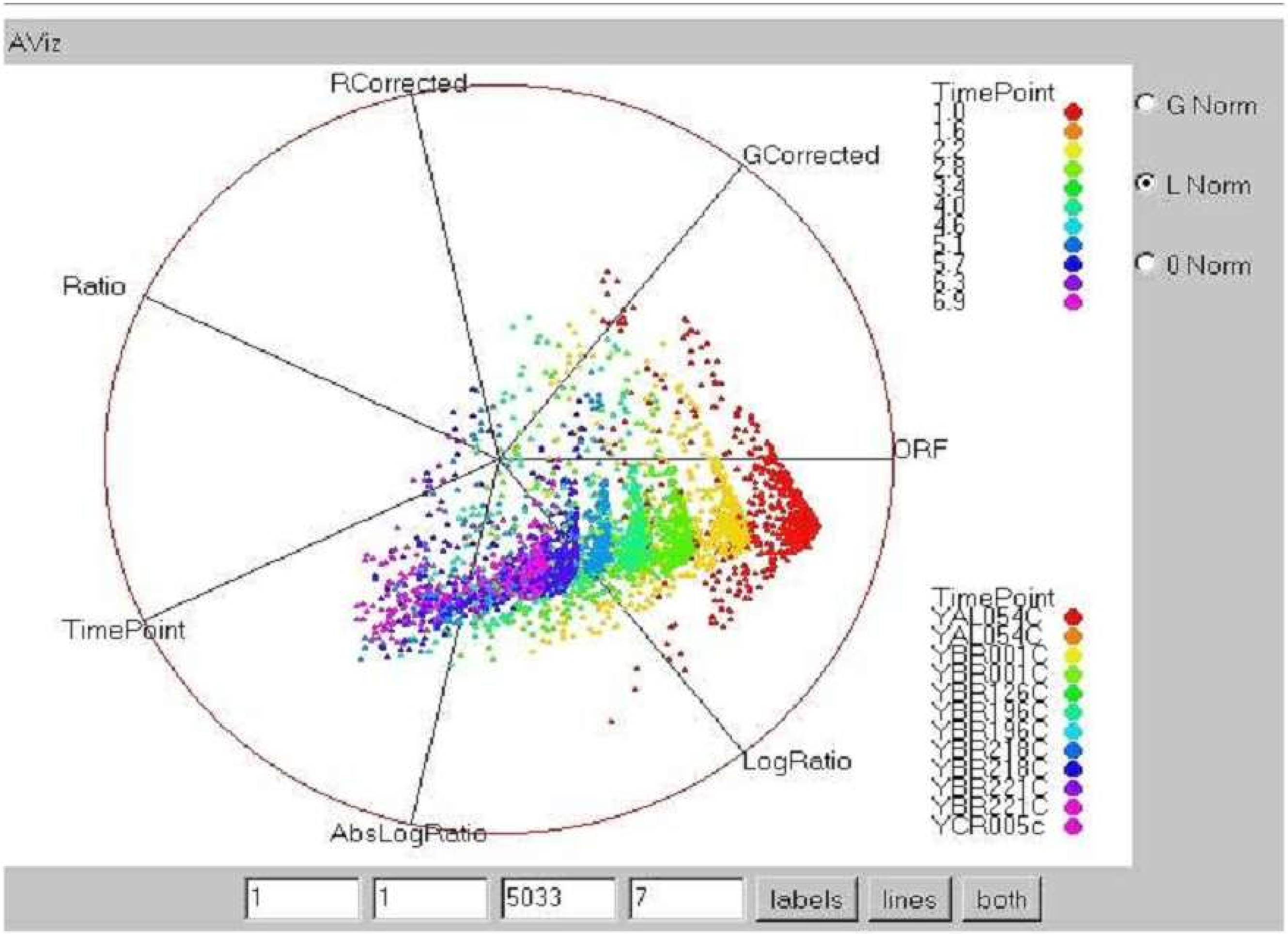
L.J.P. van der Maaten and G.E. Hinton. **Visualizing High-Dimensional Data Using t-SNE**. *Journal of Machine Learning Research* 9(Nov):2579-2605, 2008

T-sne Examples



Source: https://www.slideshare.net/xuyangela/an-introduction-to-tsne

RadViz

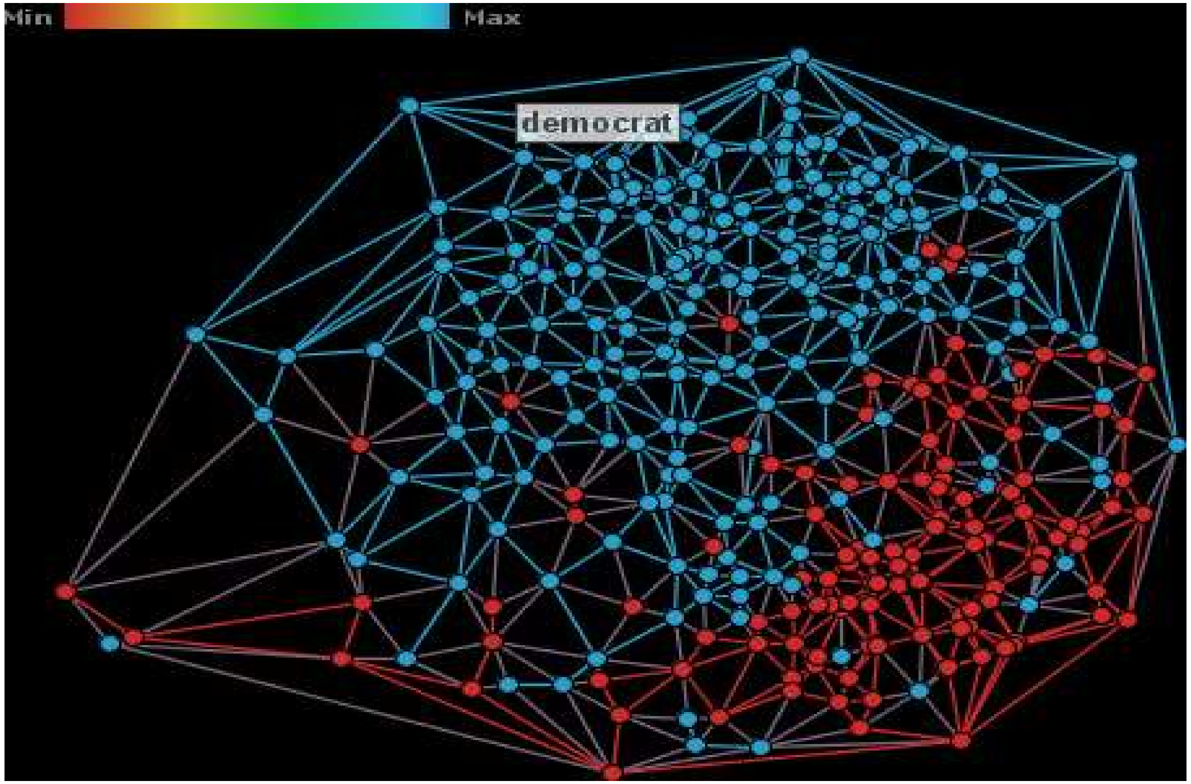


P. Hoffman, G. Grinstein, K. Marx, I. Grosse, and E. Stanley. DNA visual and analytic data mining. In Proceedings of the 8th conference on Visualization '97, VIS '97, pages 437–441, Los Alamitos, CA, USA, 1997. IEEE Computer Society Press.

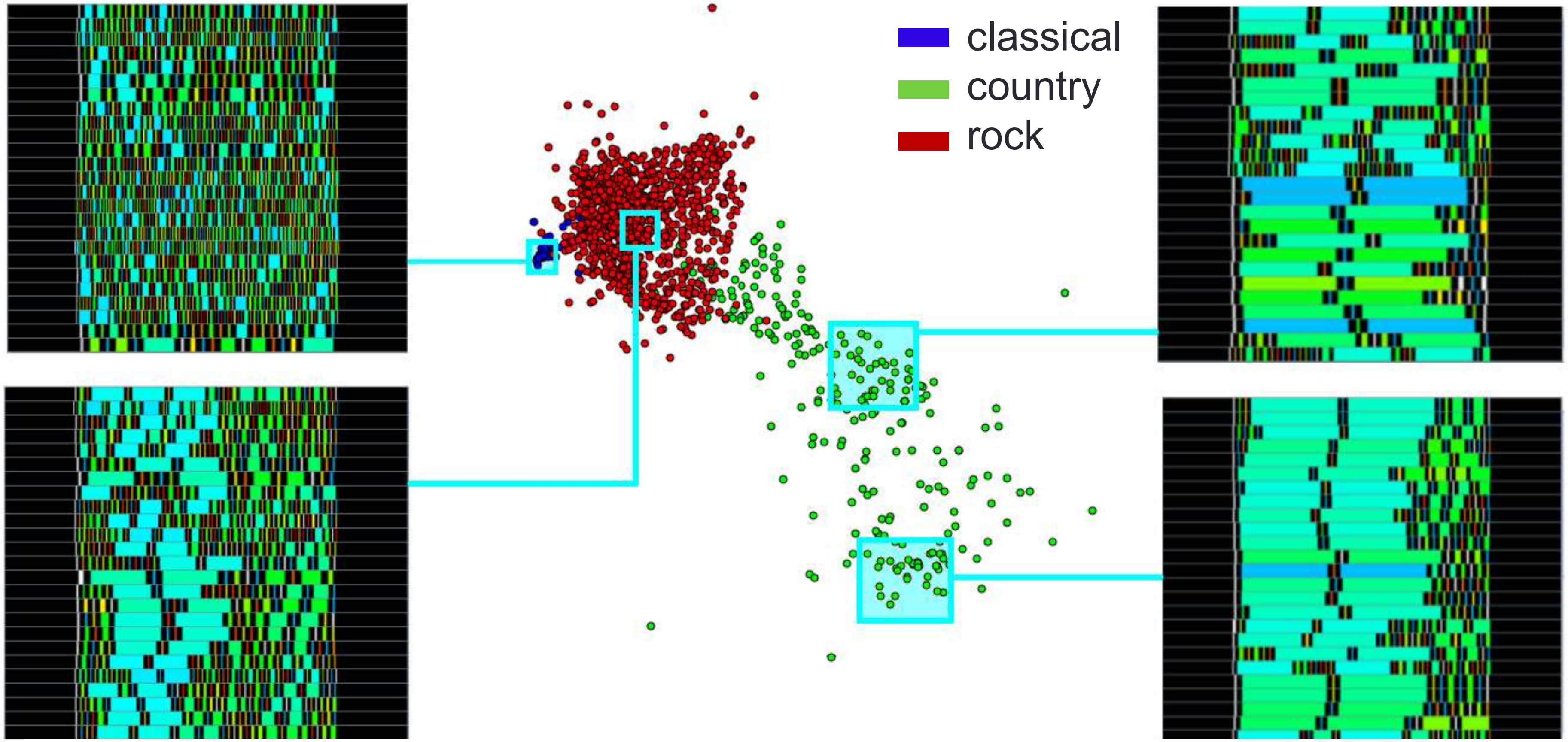
Analysis Examples

- Fluid Samples
- Text
  - Papers
  - News feeds
- Biomarkers
- GPS samples
- Images

Projection: Voting



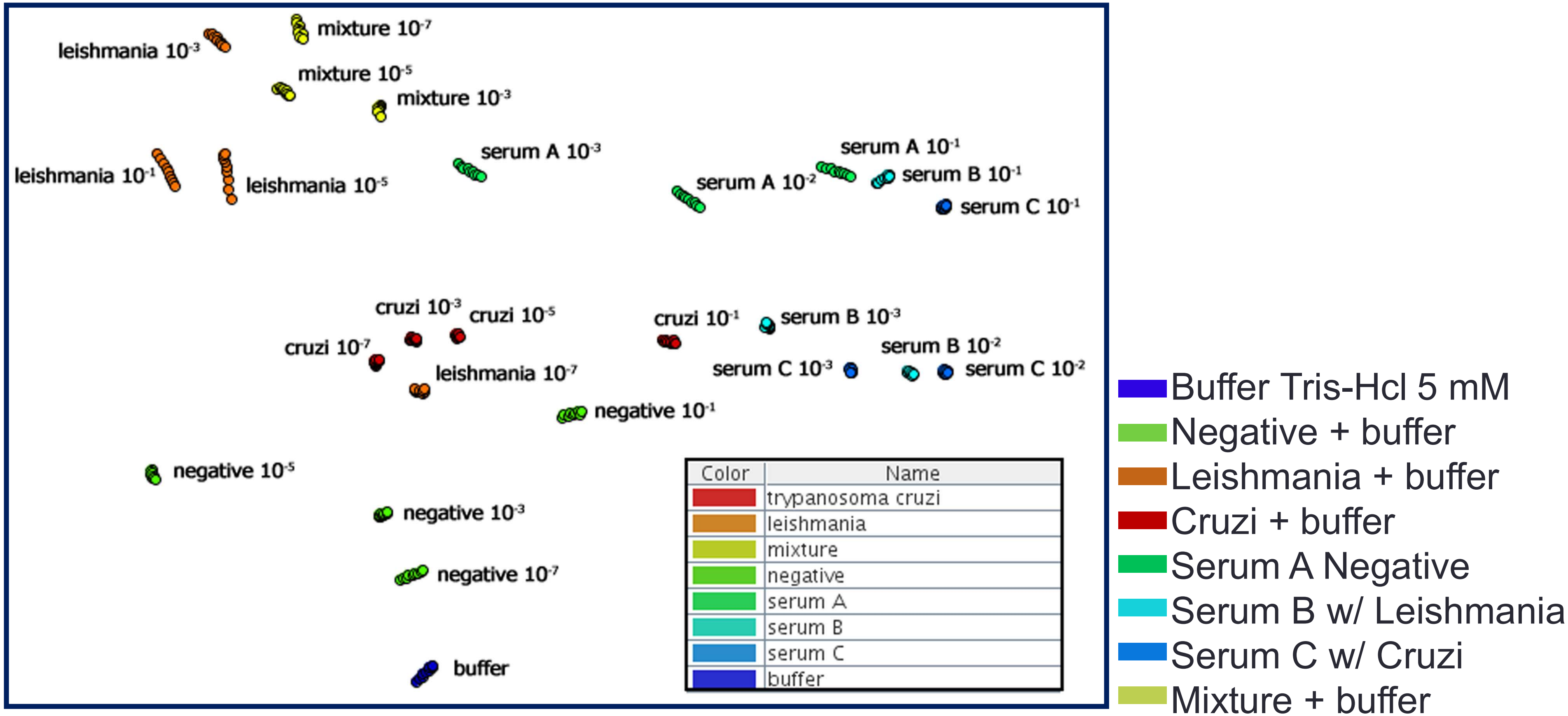
LSP mapping of music features



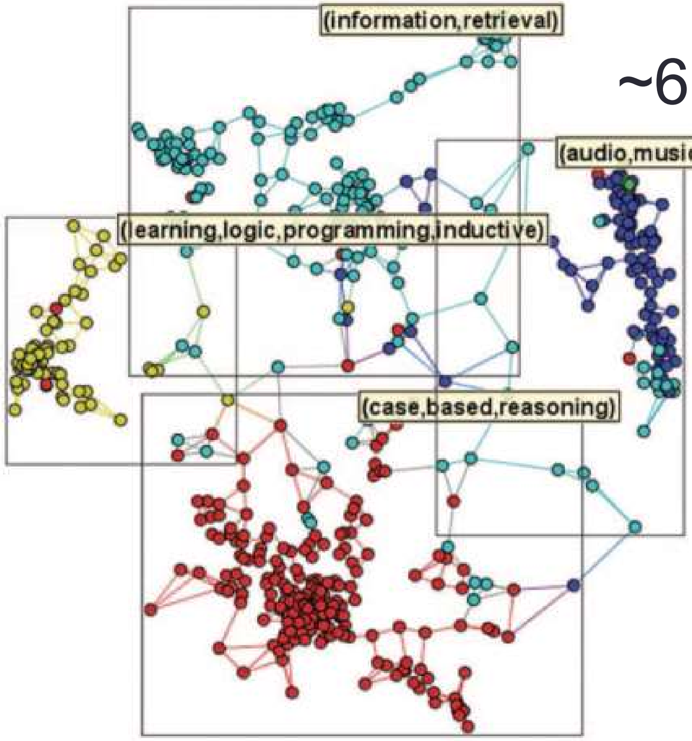
Similarity map of 1,300 songs of 3 genres plus music icons relative to the selections (LSP + DTW, features extracted from MIDI)

Vargas et al. Visualizing music collections based on structural similarity. SIBGRAPI Conf. Graphics, Patterns and Images 2014

Sammon's Mapping of combined responses of four nanosensors

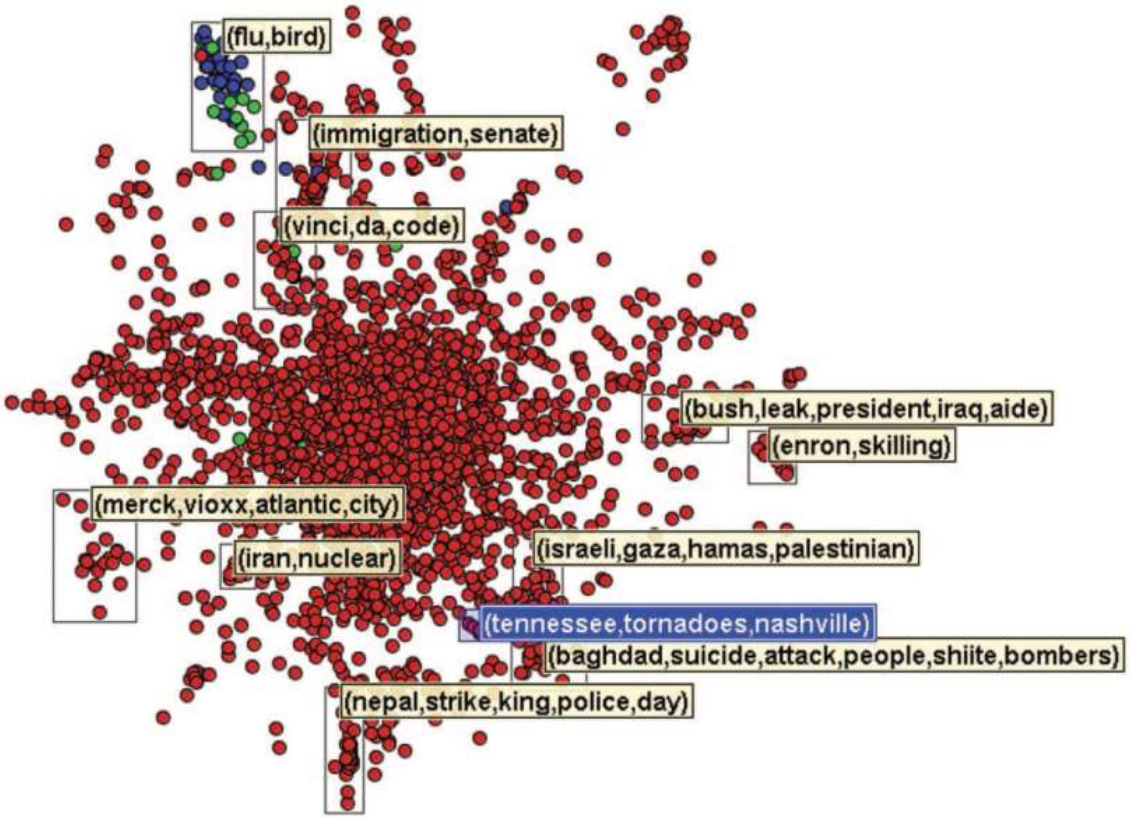


Perinotto et al., *Anal. Chem.* 2010
Paulovich et al., *Anal. Bioanal. Chem. 2011*



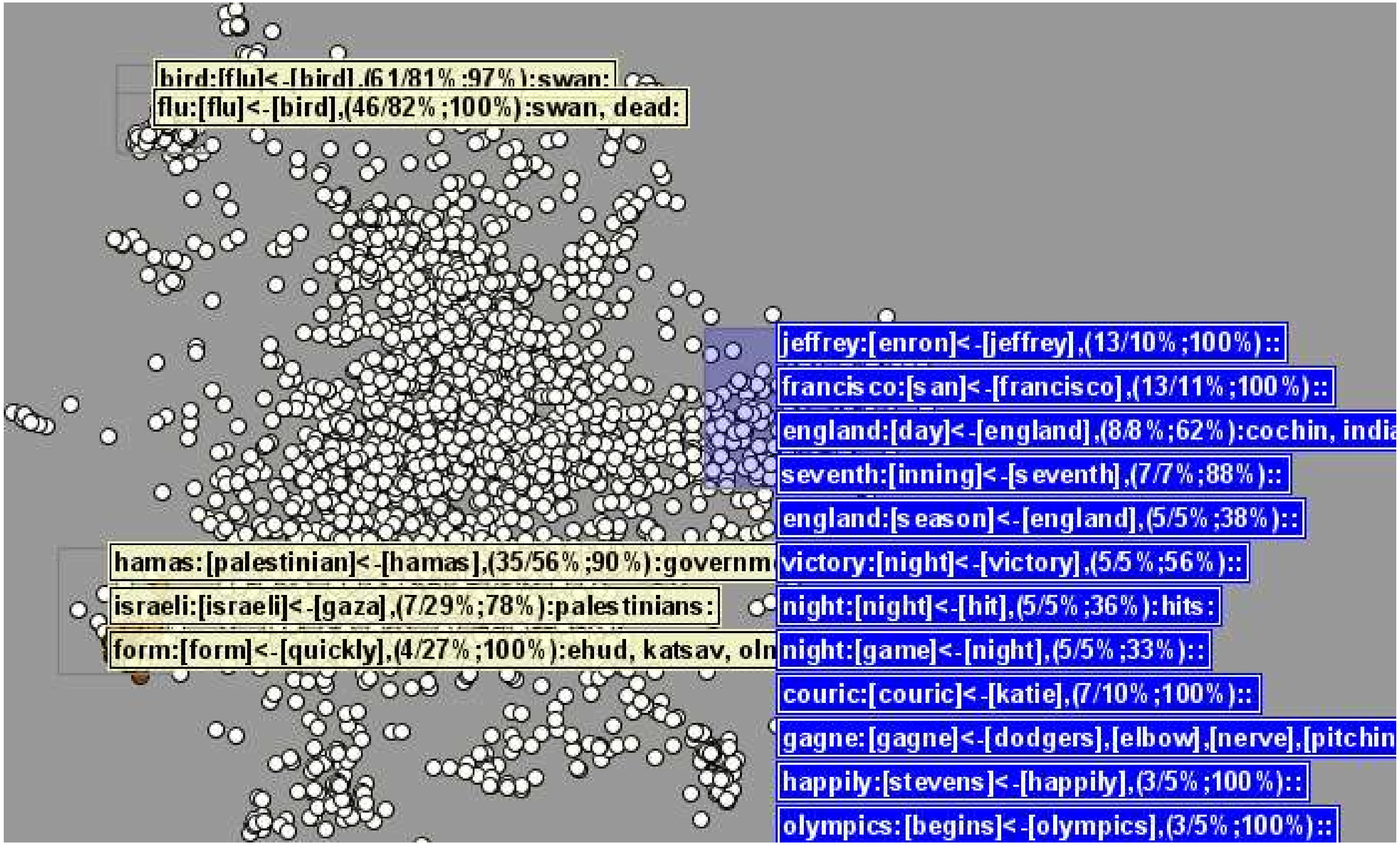
~600 scientific papers

~2,000 RSS news feeds (2006)

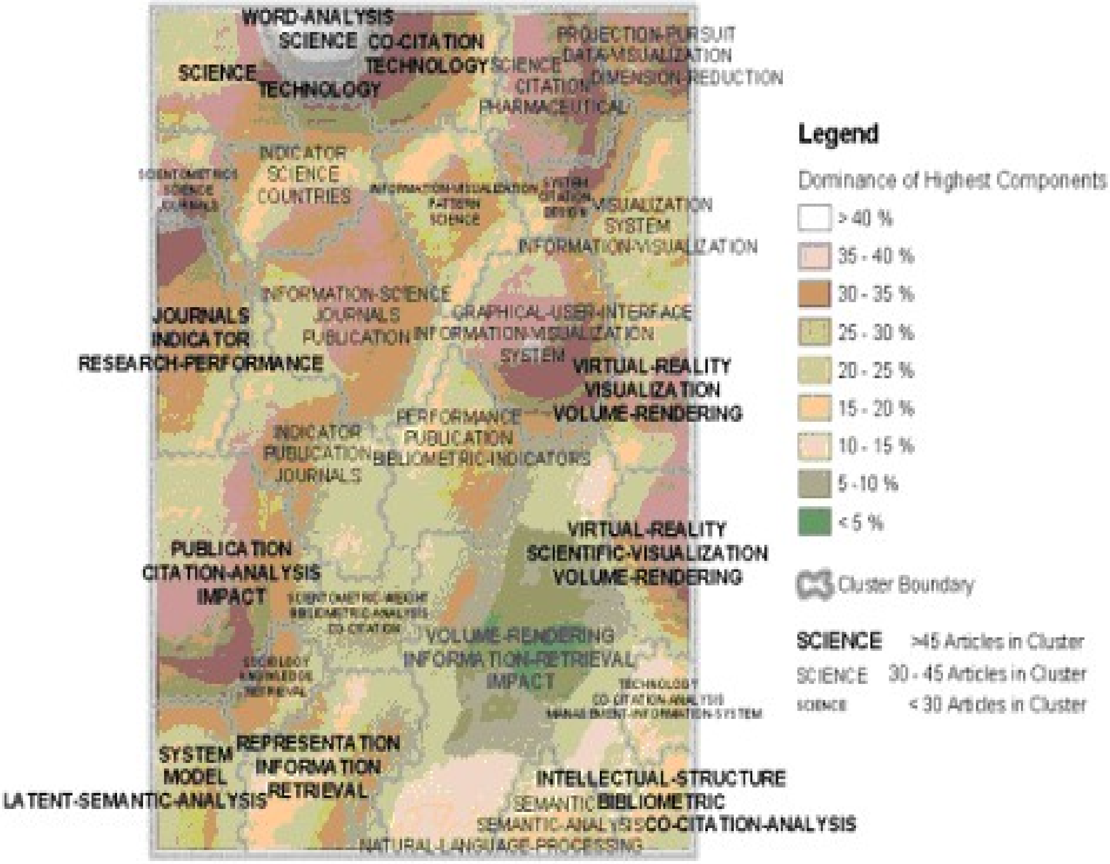


LSP mappings of text collections

Topic detailing

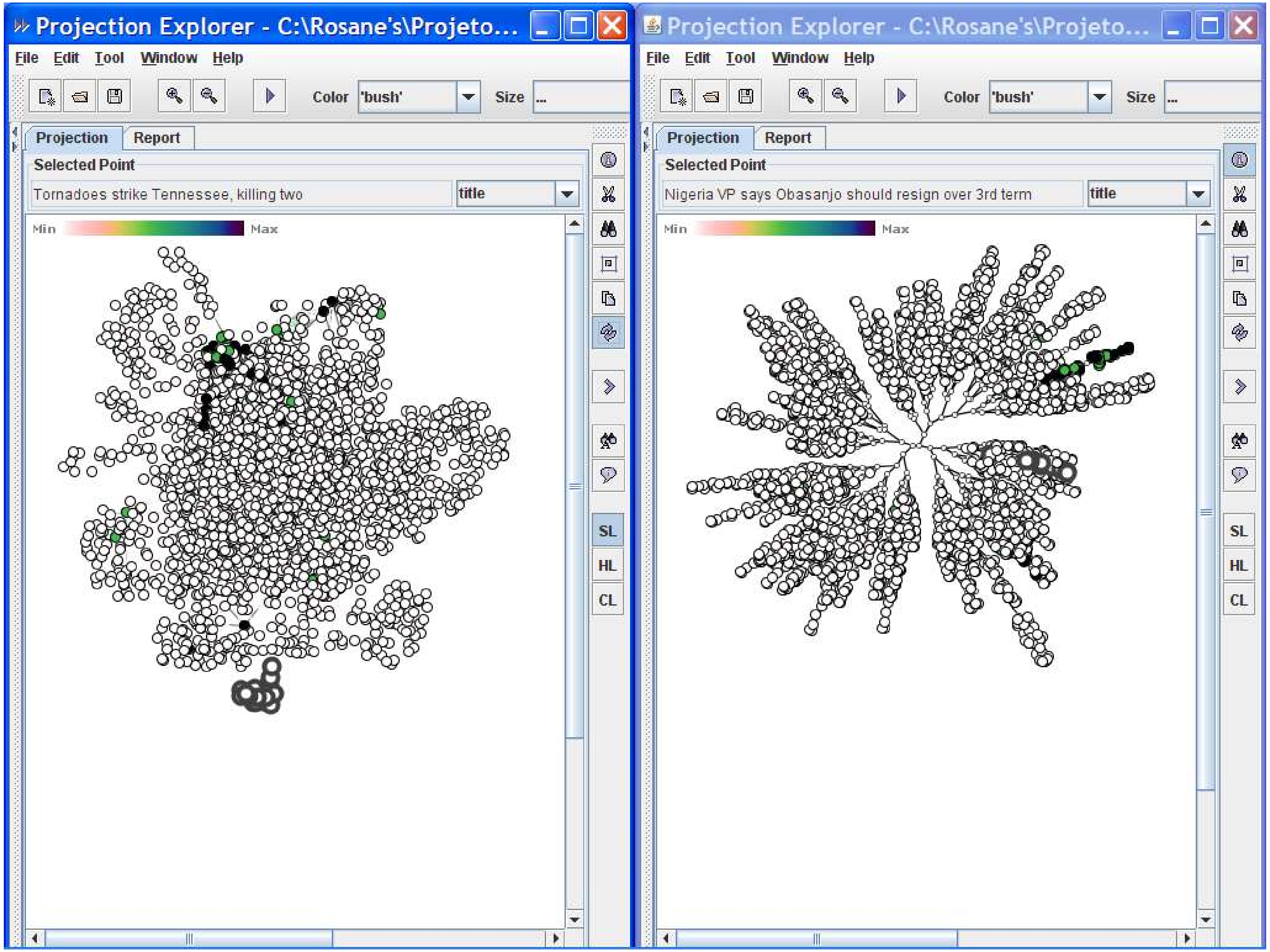


SOM-based projection



Skupin, A. 2004 - The world of geography: Visualizing a knowledge domain with cartographic means, PNAS, vo. 101.

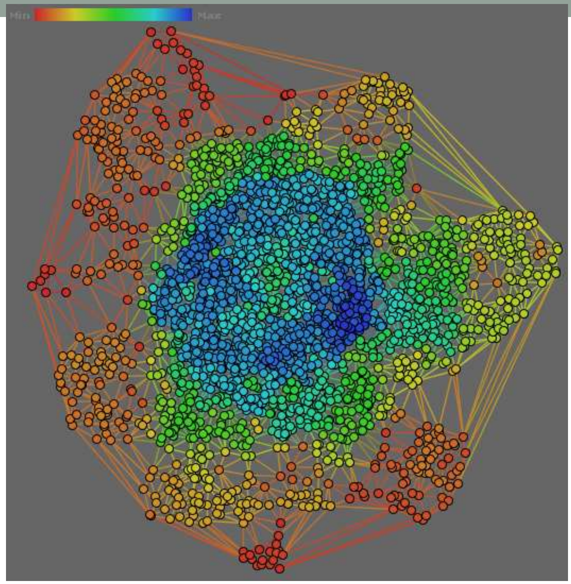
- Coordinating

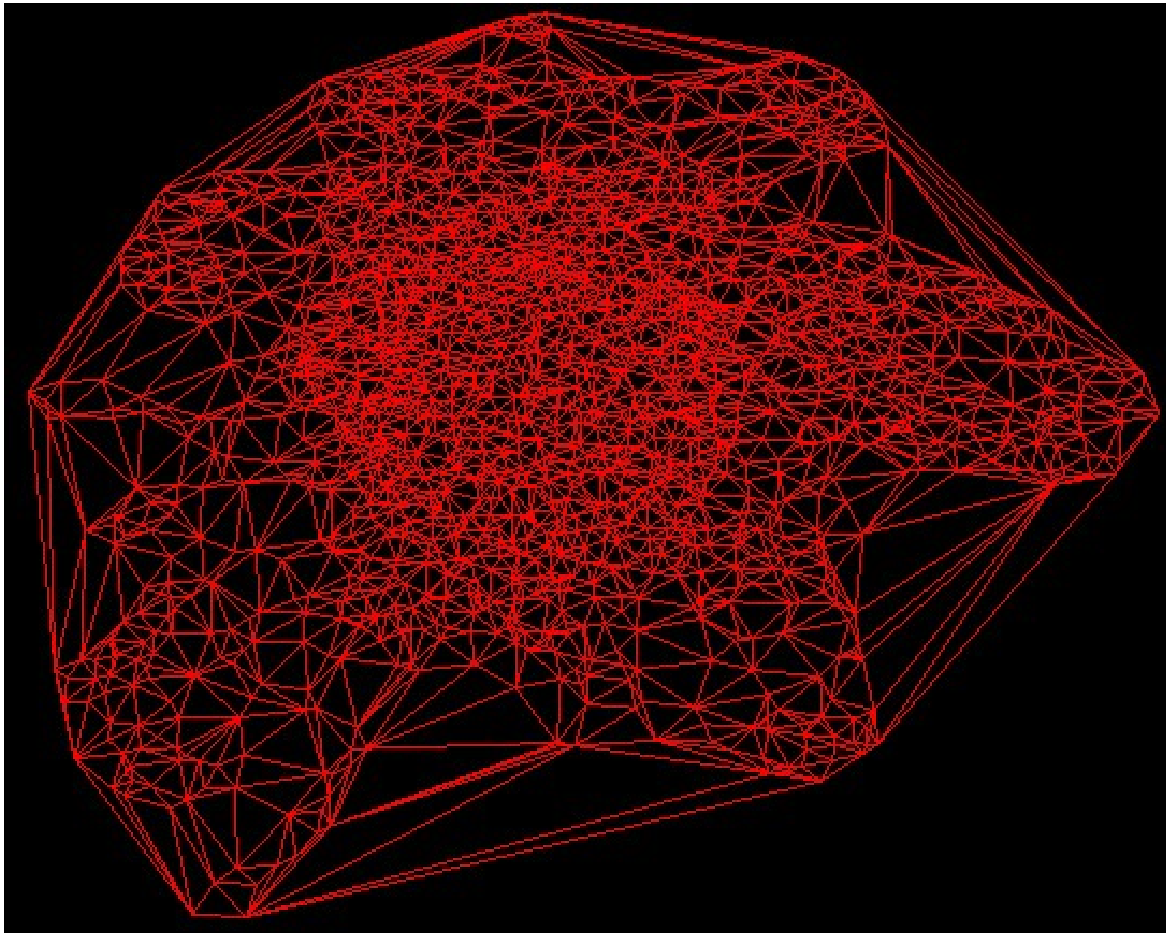


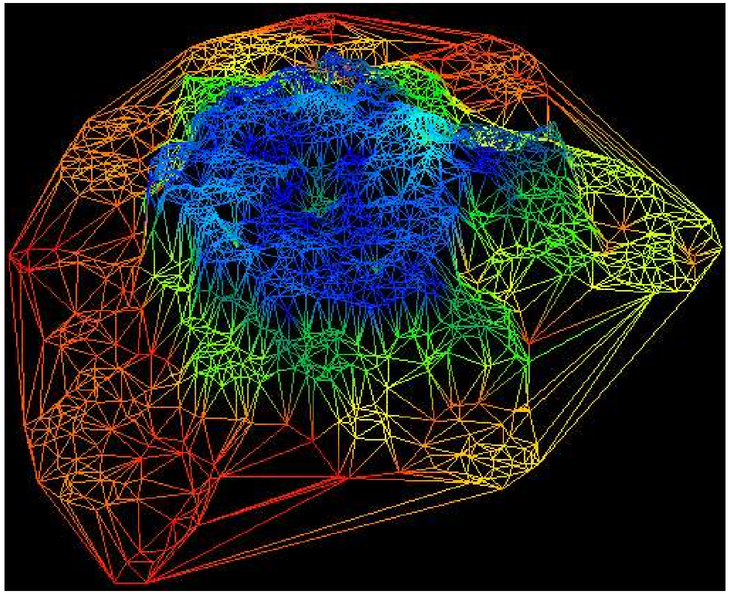
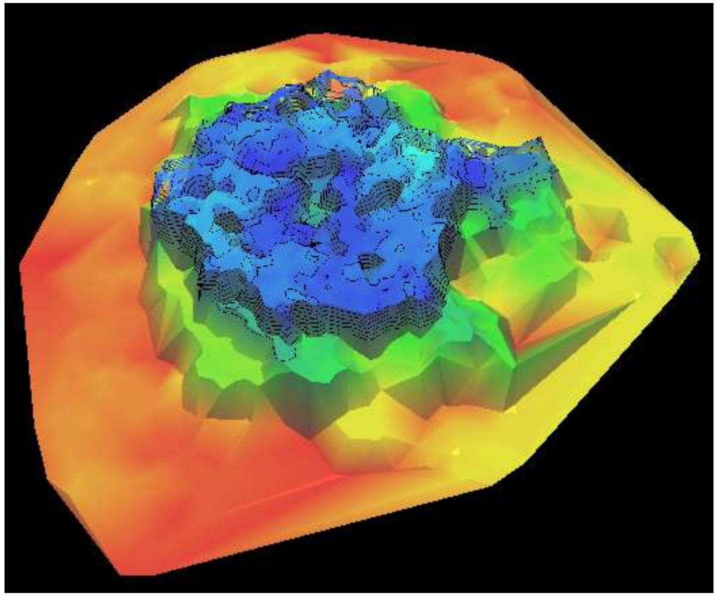
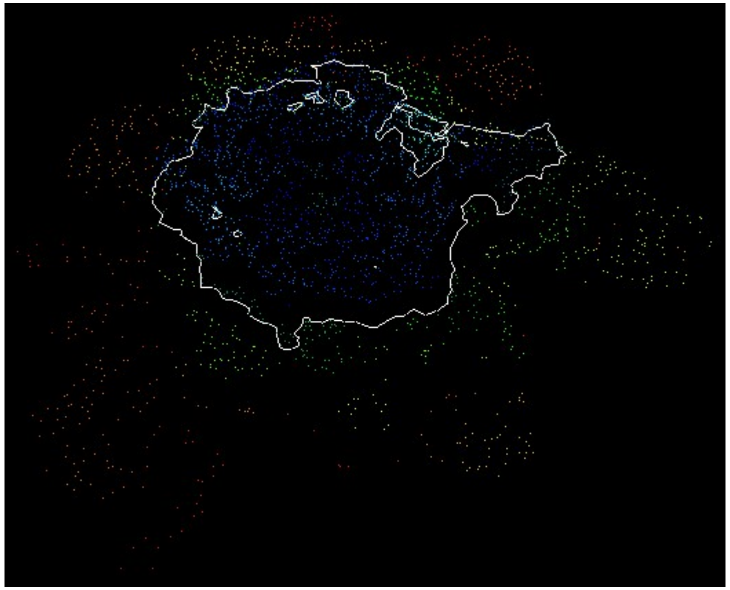
Projection Explorer - C:\Rosane's\Projeto...
File Edit Tool Window Help
Color 'bush' Size ...
Projection Report
Selected Point
Tornadoes strike Tennessee, killing two
title
Min Max
SL
HL
CL
Projection Explorer - C:\Rosane's\Projeto...
File Edit Tool Window Help
Color 'bush' Size ...
Projection Report
Selected Point
Nigeria VP says Obasanjo should resign over 3rd term
title
Min Max
SL
HL
CL

Surface-based – Landscape Views

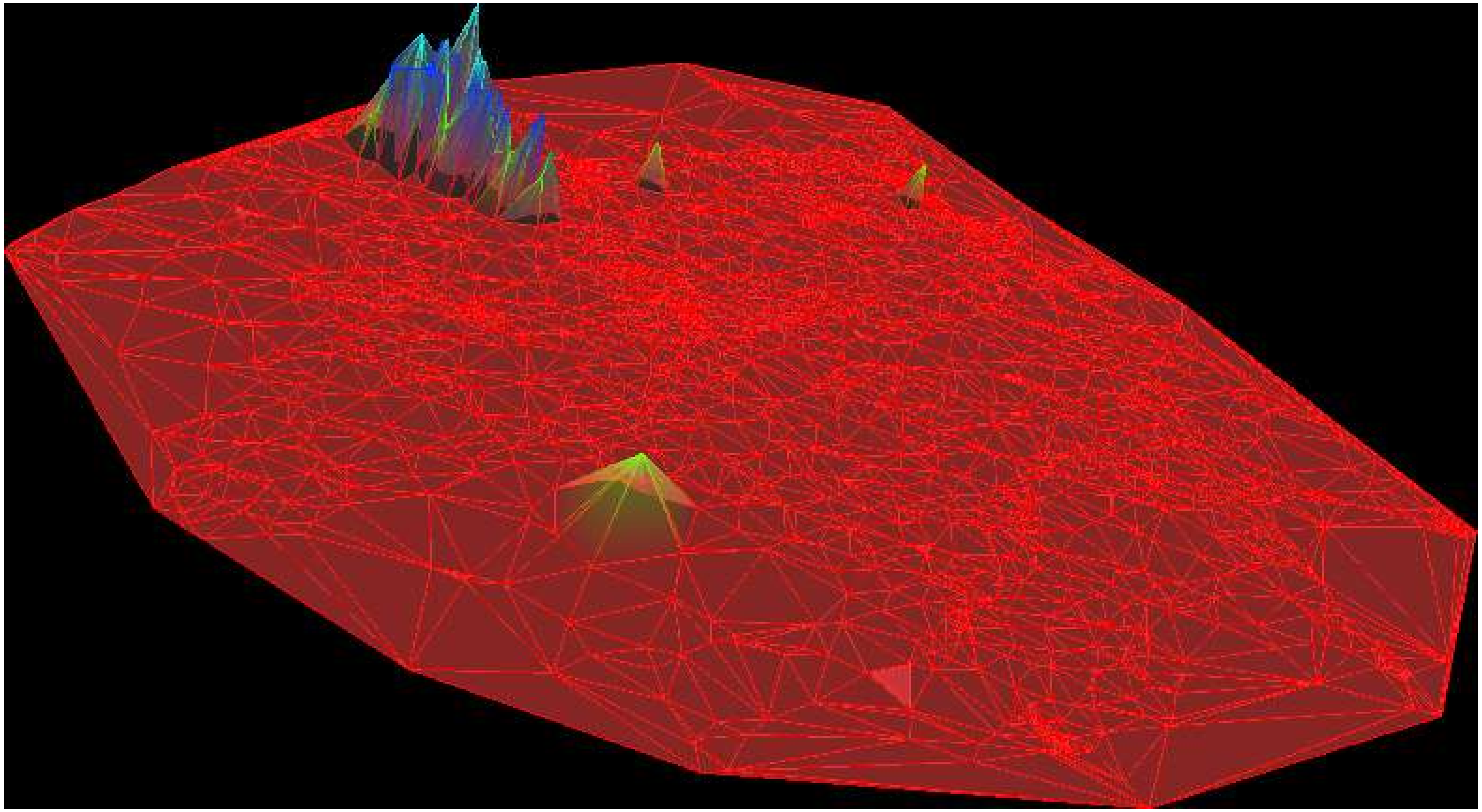
- Building a Surface


Min
Max

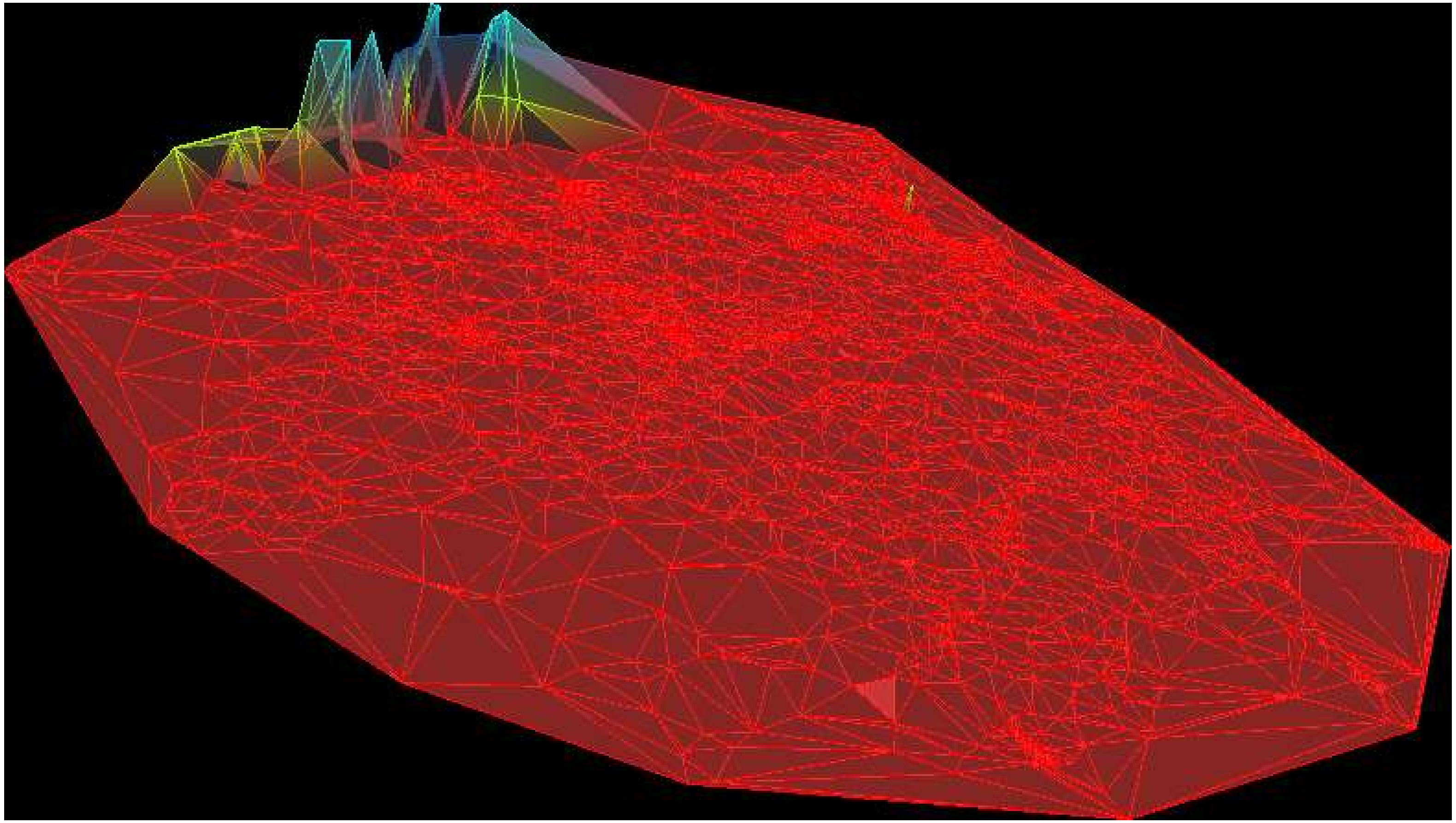


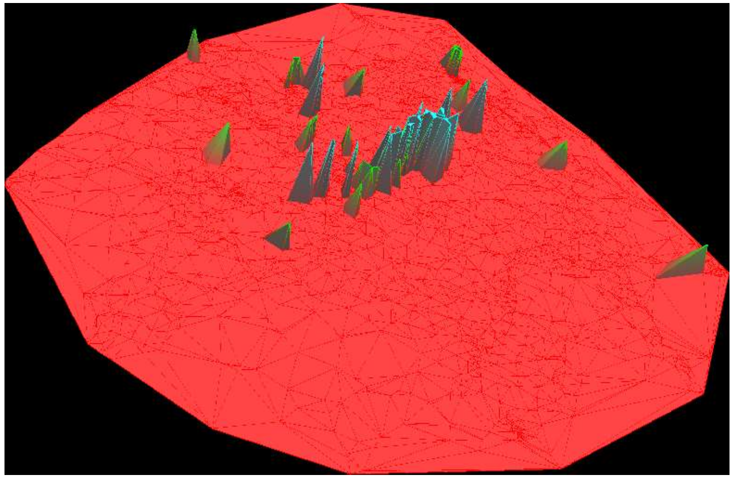
RSS News Flash



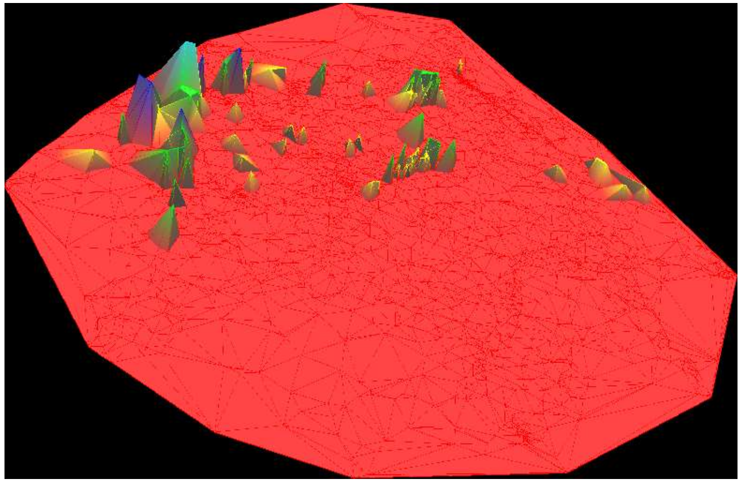
Bird and Flu



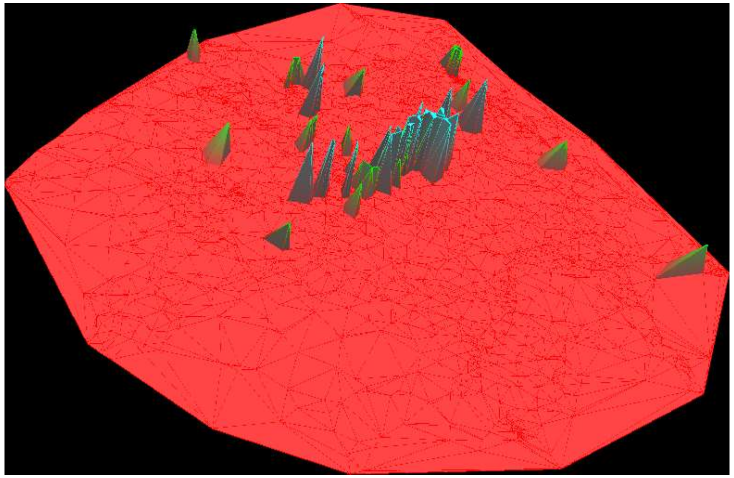
Palestinian



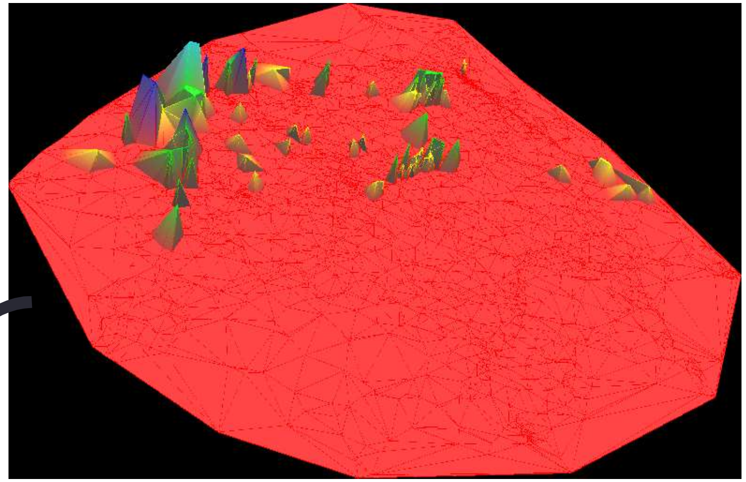
Bush



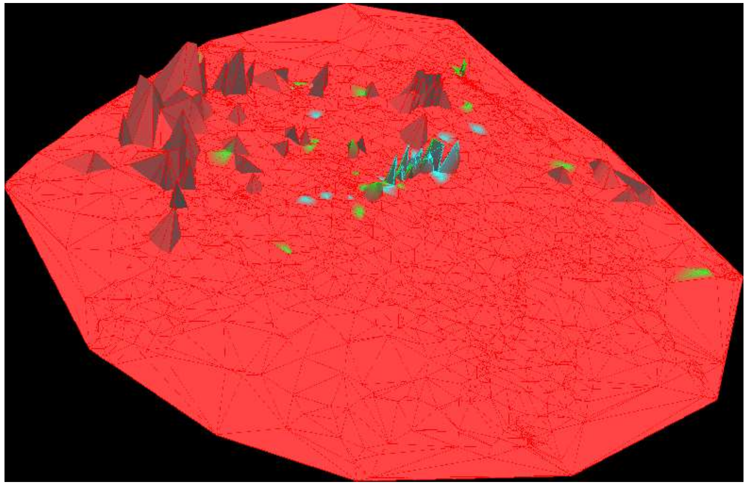
Iraq

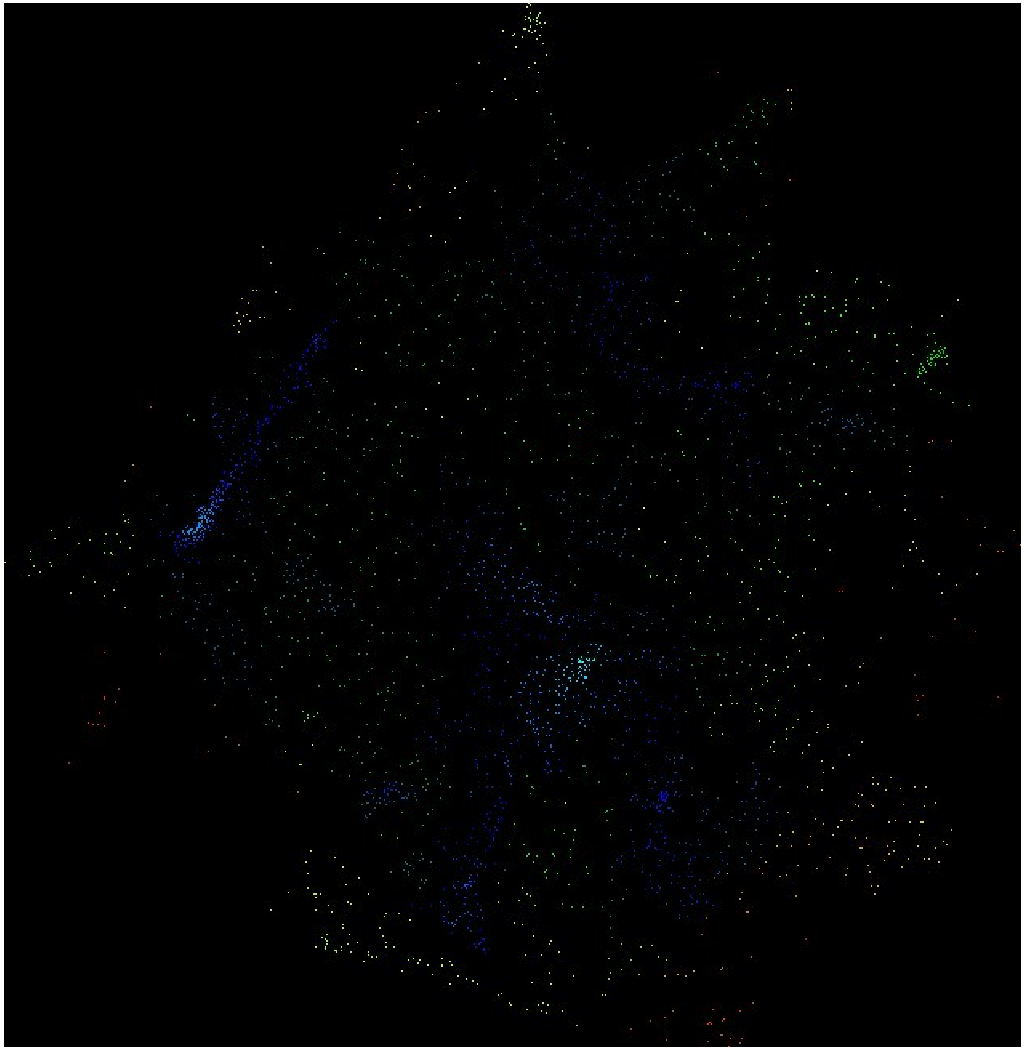


Bush

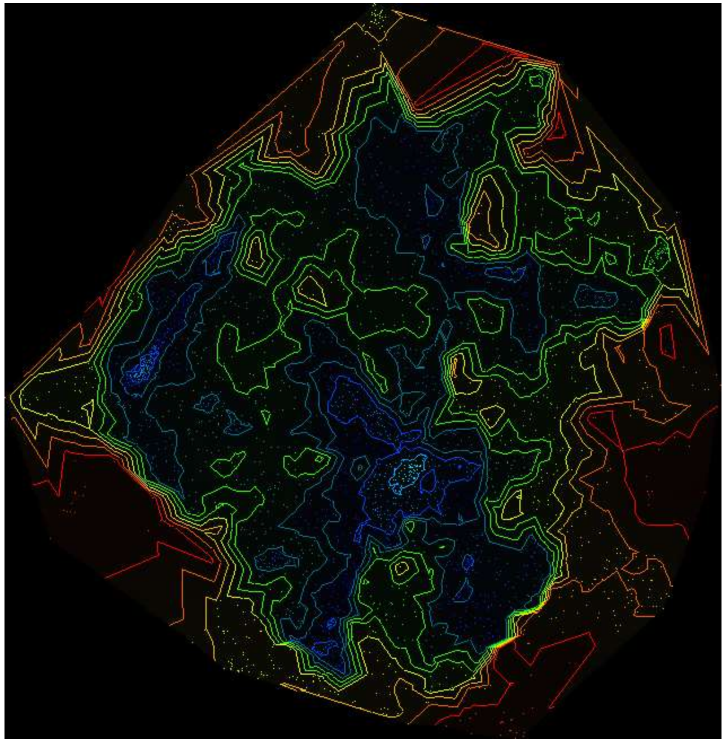


Iraq





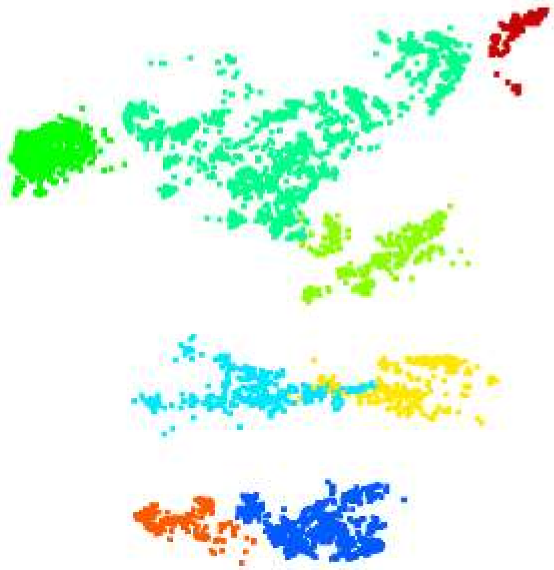
Curvas de Nível



Siqueira, P. H. ; Telles, G. P. ; **MINGHIM, R.** . Revisiting Landscape views in Information Visualization. In: Bruno Lopes; Talita Perciano. (Org.). Learning and Infering. Festschrift for Alejandro C. Frery on the Occasion of his 55th Birthday. 1ed.Londres: College Publications, 2015, p. 131-152.

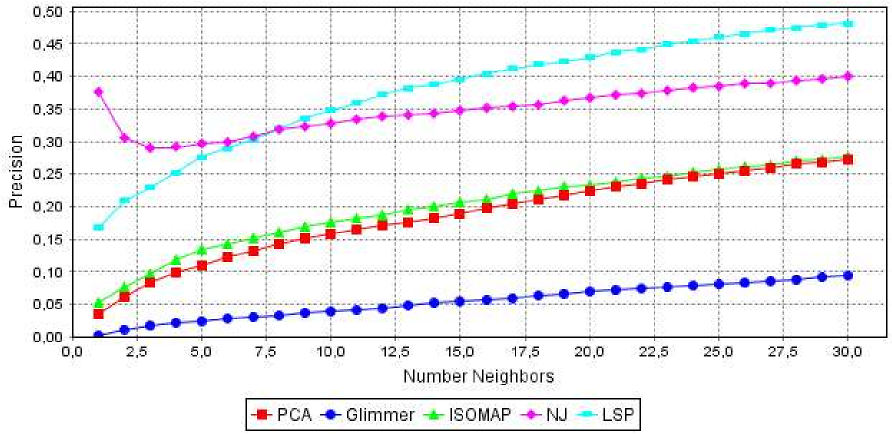
More Applications – Fiber Tracking

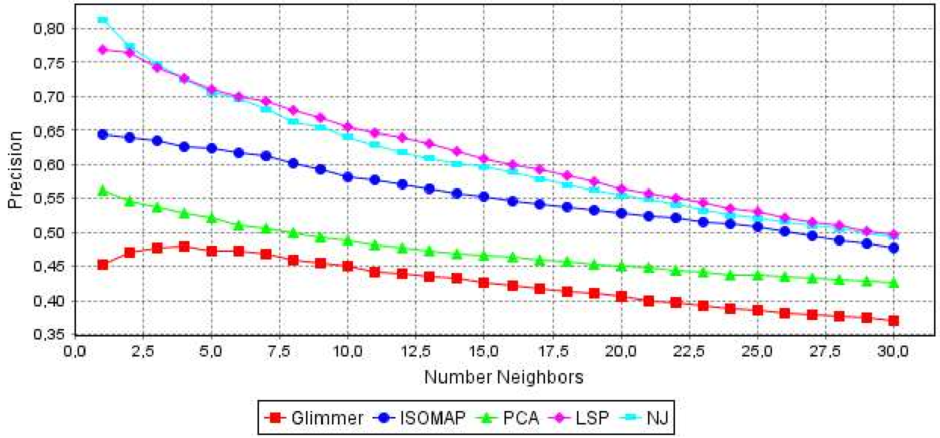
- Projection from fiber features
- Interaction through fast and reconfigurable projections (LAMP)
- Lines, Tubes and Surface Views

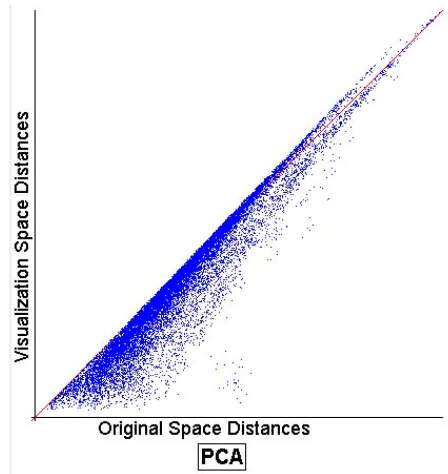


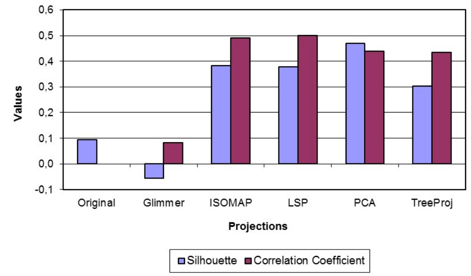
Poco, Eler, Paulovich, Minghim - Employing 2D projections for fast visual exploration of large fiber tracking data, **Computer Graphics Forum**, **Eurovis** 2012.

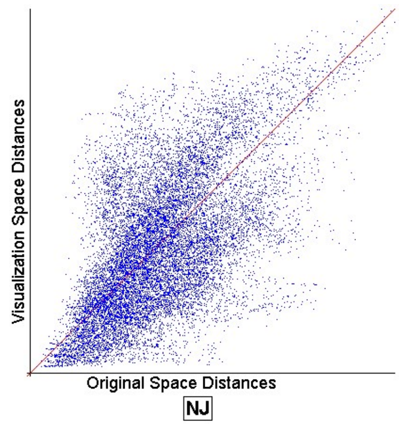
Evaluation


Precision
Number Neighbors
PCA
Glimmer
ISOMAP
NJ
LSP


Precision
Number Neighbors
Glimmer
ISOMAP
PCA
LSP
NJ


Visualization Space Distances
Original Space Distances
PCA


Values
Original
Glimmer
ISOMAP
LSP
PCA
TreeProj
Projections
Silhouette
Correlation Coefficient


Visualization Space Distances
Original Space Distances
NJ

Evaluation

- Specific issues
  - How do users perceive point-placement layouts?
  - What are such layouts good for?
  - Which techniques do best in which situations? How do they compare?
    - Measures from a controlled user study
    - Numerical measures

Evaluation

- Study with 61 subjects aimed at comparing how different layouts are perceived
- 5 point-placement techniques (NJ tree, Glimmer, LSP, ISOMAP and PCA) compared for segregation, precision and clutter avoidance capabilities
- Hypotheses
  - H1 Different projections perform better on different tasks
  - H2 Performance of projections is task dependent
  - H3 Performance of projections depends on data characteristics
  - H4 User preferences for projections are governed by good segregation capability
- Tasks: cluster and outlier perception, neighborhood perception, density perception
- Data sets: image and text collections

Evaluation

- Hypotheses
  - H1 Different projections perform better on different tasks

    Yes!
  - H2 Performance is task dependent

    Partly!
  - H3 Performance depends on data characteristics

    Yes!
  - H4 User preference is governed by good segregation

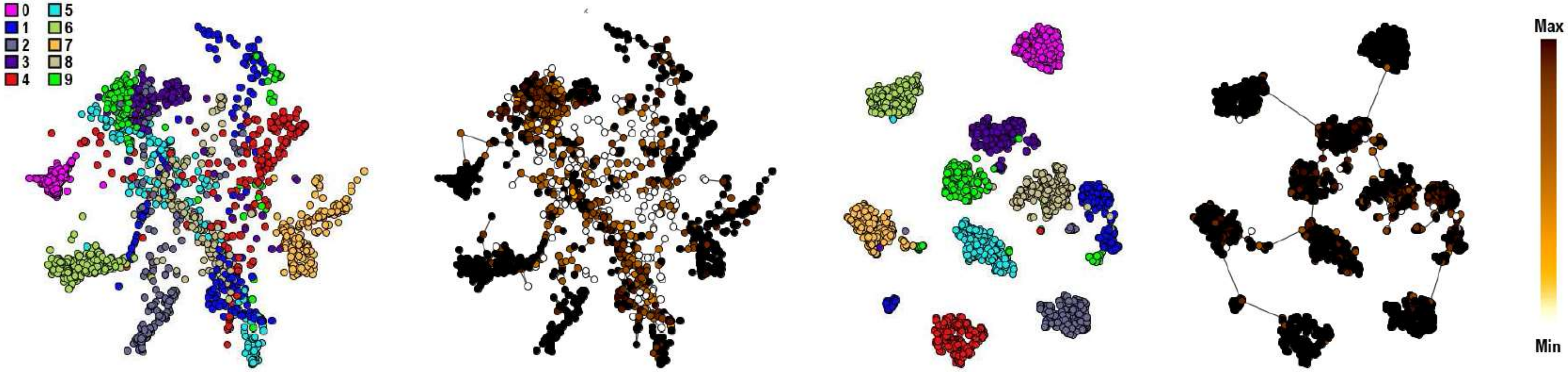
    No!

Etemadpour, R. ; Motta, R; Paiva, J.G.S; Minghim, R.; Oliveira, M. C. F., Linsen, L. Perception-Based Evaluation of Projection Methods for Multidimensional Data Visualization. IEEE Transactions on Visualization and Computer Graphics , v. 21, p. 81-94, 2015

Evaluation

- Numerical measures + contrasting with results from user study
  - Previous: Neighborhood Hit, Silhouette Coefficient, Distance Plots
  - New: topological measures defined on a similarity graph built from the projections
    - Class segregation
    - Group formation
    - Cluttering

Assessing projection mappings

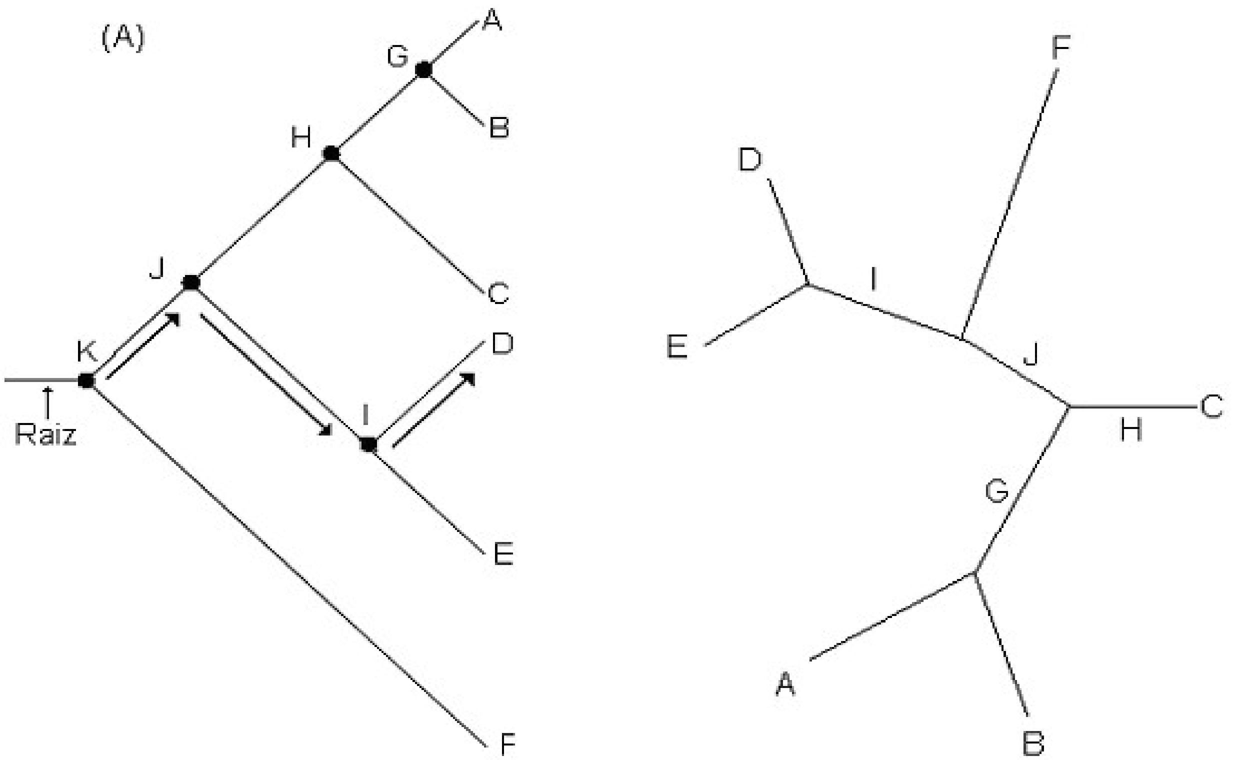


(a) LSP (b) LSP (0.88) (c) t-SNE (d) t-SNE (0.98)

Figure 10: LSP and t-SNE projections of *Optdigits*: (a) LSP with classes; (b) LSP mapping *Class Separation Validation*; (c) t-SNE with classes; (d) t-SNE mapping *Class Separation Validation* (darker is better). Summary measure for the projection is shown in parentheses.

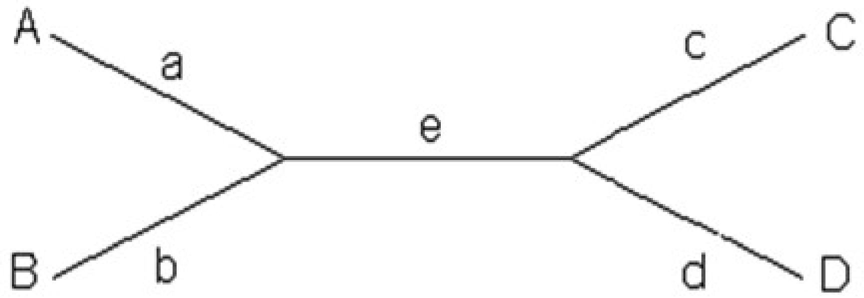
R. Motta, A.A. Lopes, R. Minghim, M.C.F. Oliveira, Graph-based measures to assist user assessment of multidimensional projections. *Neurocomputing 150*: 583-598 (2015)

Point Placement by Phylogenetic Tree Construction Algorithms (N-J Trees)



Point Placement by Phylogenetic Tree Construction Algorithms (N-J Trees)

$$d_{AB} + d_{CD} \leq \max(d_{AC} + d_{BD},\ d_{AD} + d_{BC})$$



$d_{AB} = a + b$ $\quad$ $d_{AC} = a + e + c$ $\quad$ $d_{AD} = a + e + d$

$d_{CD} = c + d$ $\quad$ $d_{BD} = b + e + d$ $\quad$ $d_{BC} = b + e + a$

Algorithm Neighbor-joining

Input: distance matrix

1. Criate a star tree for n objects.
2. Iteration
   1. Select a node pair (i,j) with smaller Sij (branch size)

$$S_{ij} = \frac{1}{2(n-2)} \sum_{k=3}^{N} (D_{ik} + D_{jk}) + \frac{1}{2} D_{ij} + \frac{1}{n-2} \sum_{3 \le m < n} D_{ij}$$

   2. Combine nodes i and j in a new node and calculate the branch size of the new node.

$$L_{ix} = \frac{D_{ij} + D_{iz} - D_{jz}}{2} \qquad L_{jx} = \frac{D_{ij} + D_{jz} - D_{iz}}{2}$$

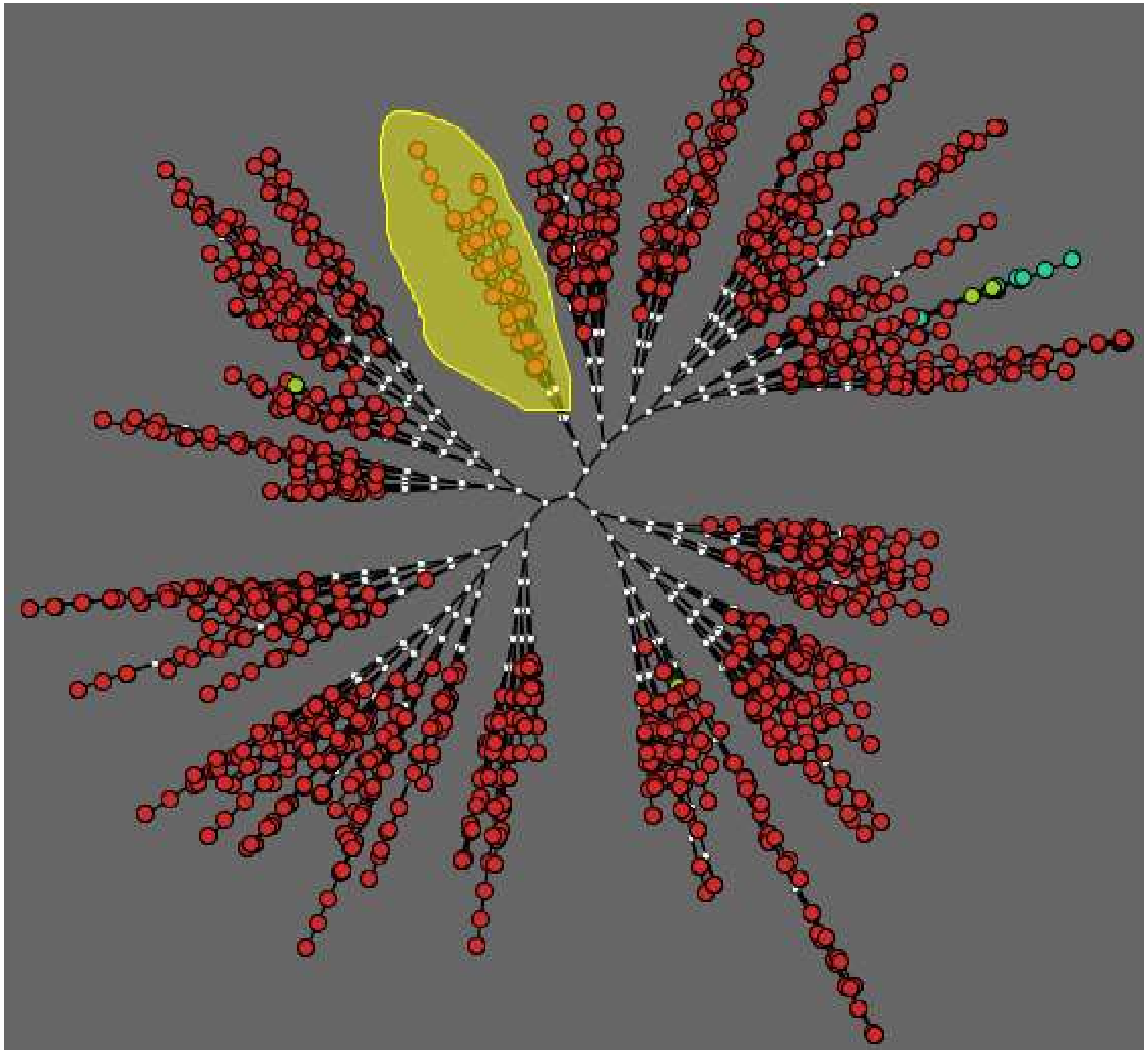
Algorithm Neighbor-joining

3. Calculate new distance matrix, computing the new distances from the new node to the remaining nodes.

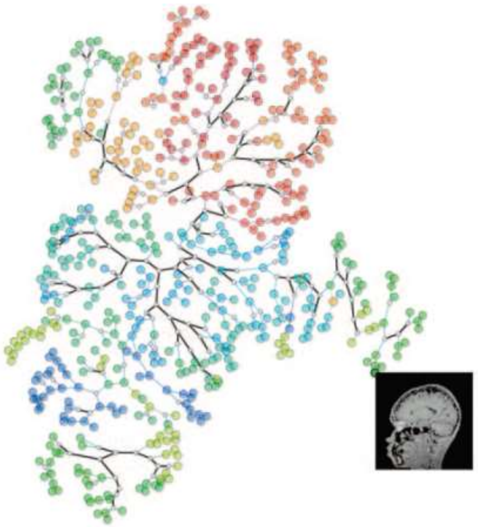
$$D_{(i-j),k} = \frac{(D_{ik} + D_{jk})}{2} \qquad (3 \leq k \leq N)$$

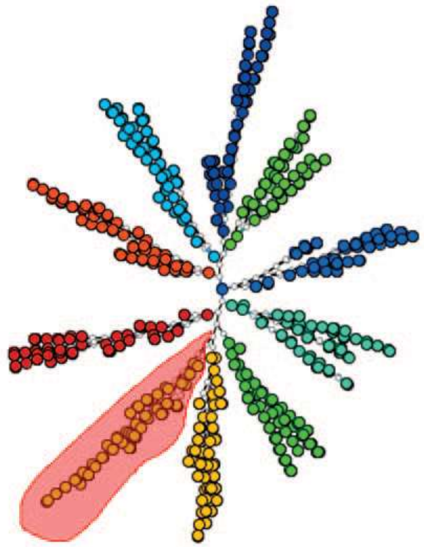
4. Eliminate previous nodes i and j
5. If n>2 then iterate again.

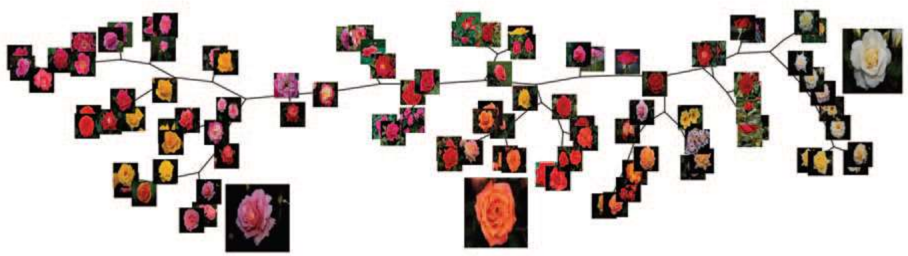
- Initial view (N-J Tree)



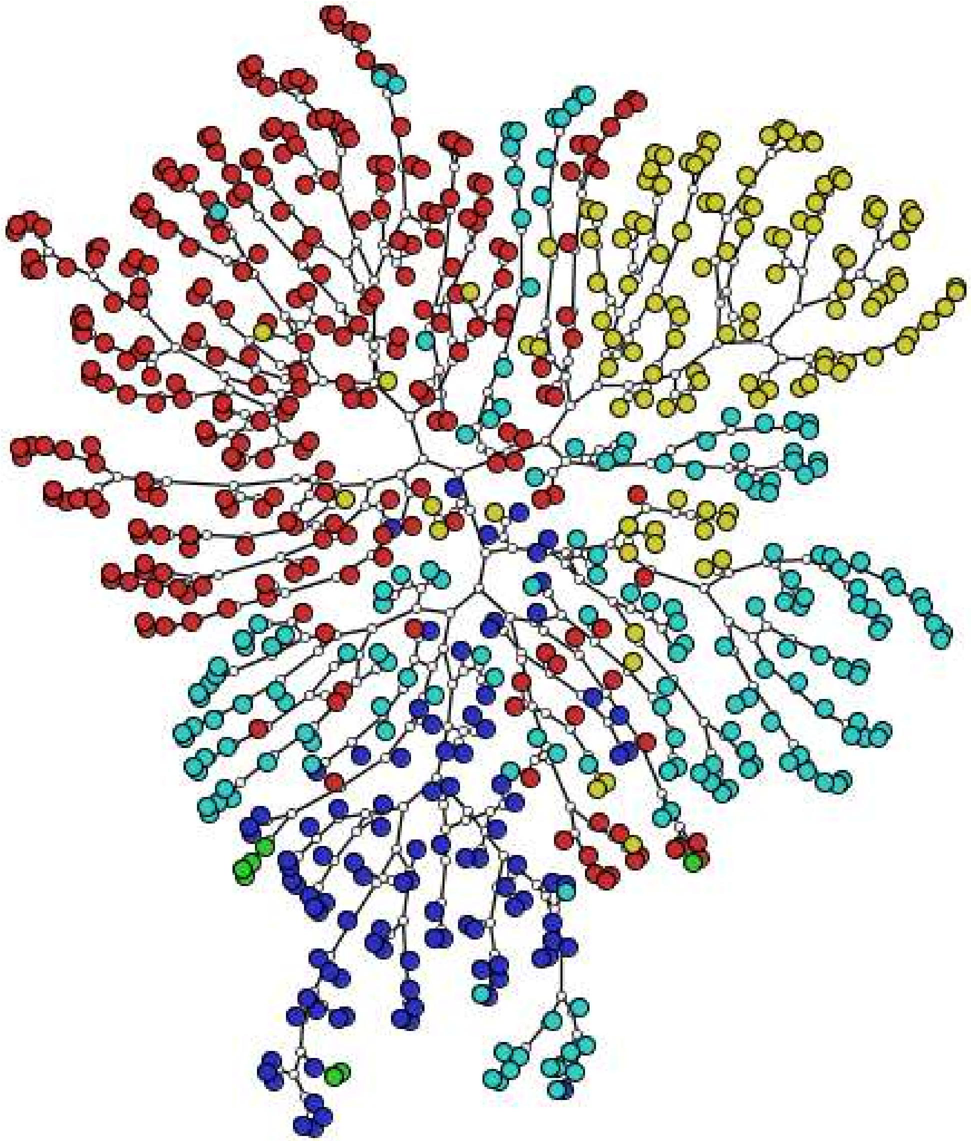
NJ & PNJ Trees





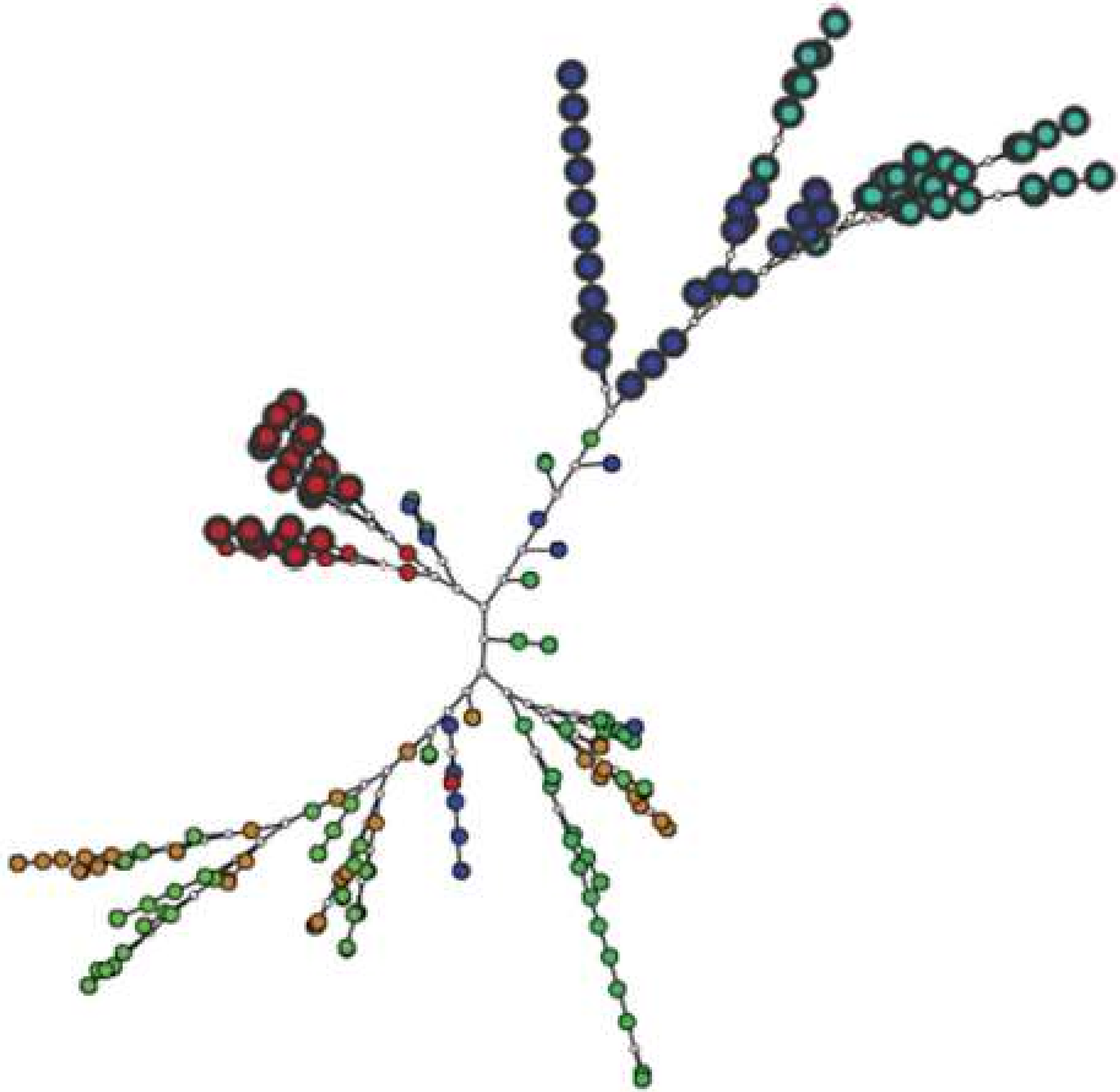


- Cuadros, Paulovich, Minghim, Telles, Point placement by phylogenetic trees and its application to visual analysis of document collections, ***IEEE VAST*** *2007.*
- Paiva, Florian-Cruz, Pedrini, Telles, Minghim, Improved Similarity Trees and their Application to Visual Data Classification, ***IEEE Trans. Visualization and Computer Graphics****, 2011.*

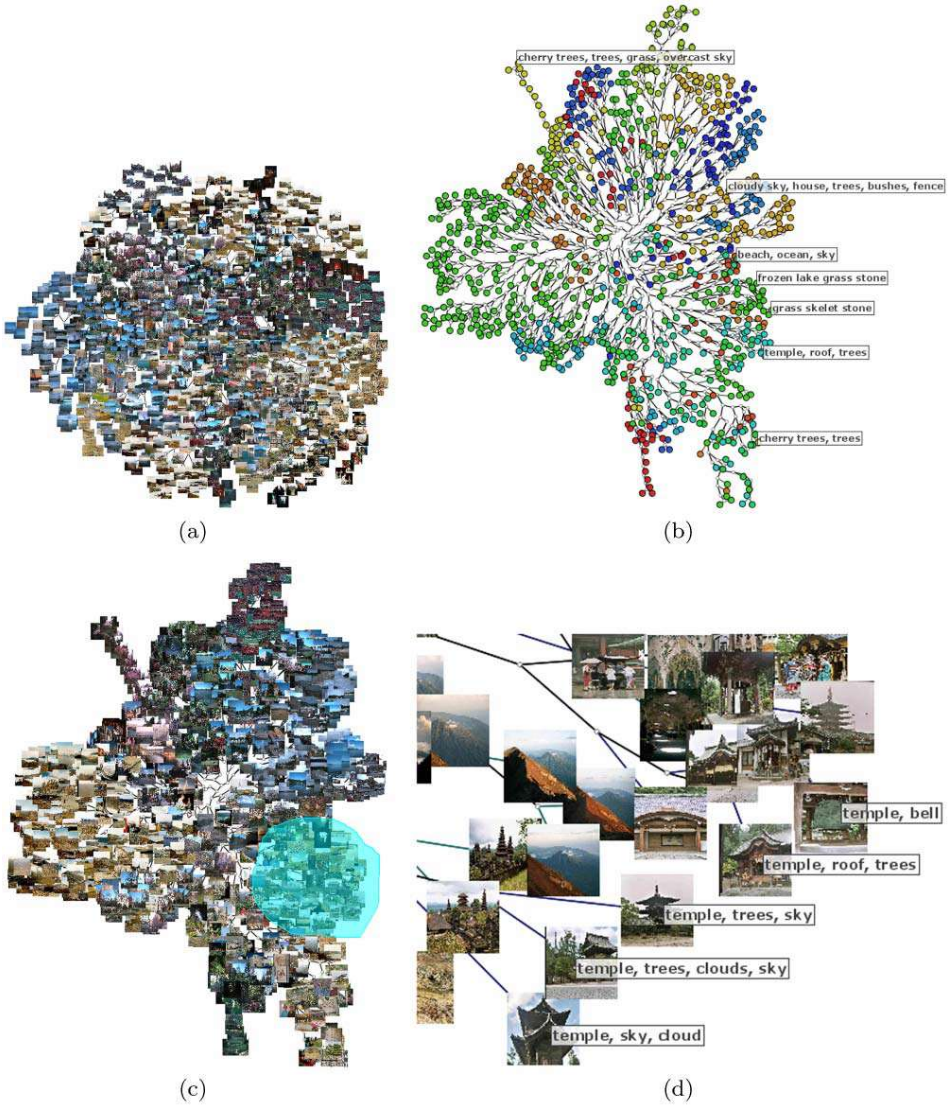


Example: Images

Example: Images



Applications

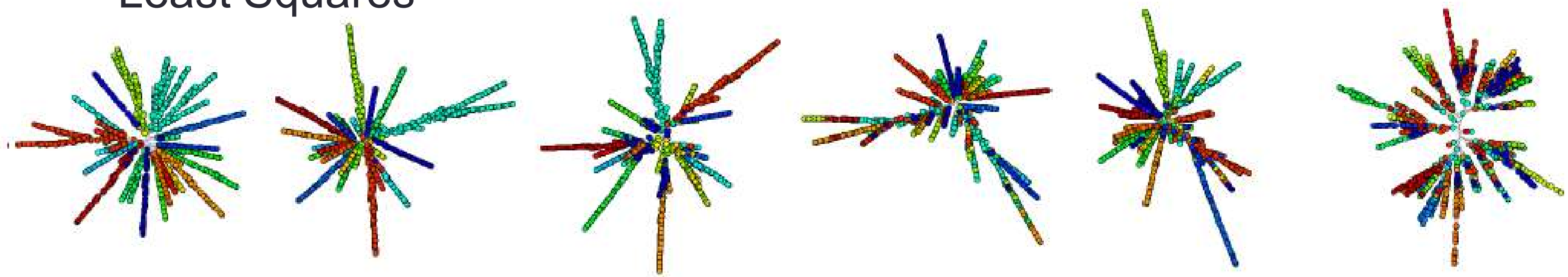


Exploratory visualization of

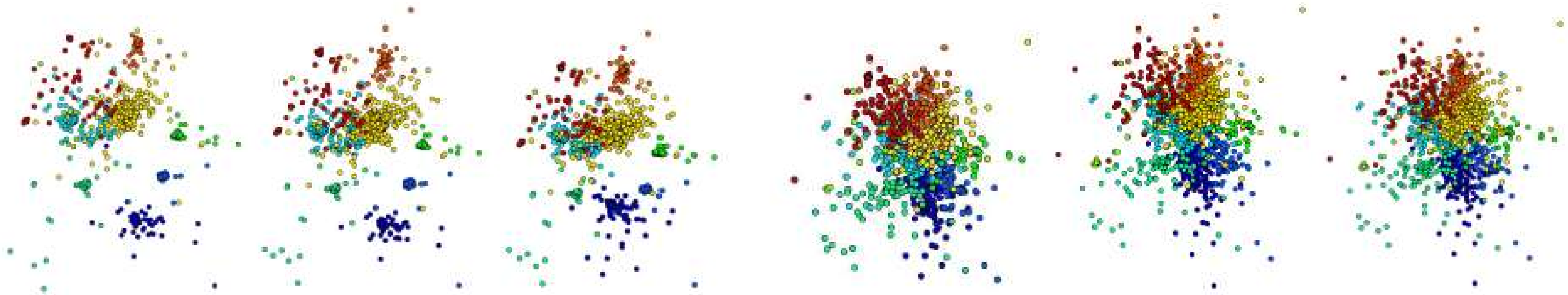
- images
- text: news, scientific papers, web search results
- sensor measurements
- volumetric data: vector, scalar
- social networks
- neural fibers
- particle trajectories
- time series

VDM – Supervision in DR and Projections

- User feeds samples in small amounts
- Larger data sets are reduced in dimensions or projected via Partial Least Squares



(a) PCA (b) PLSoaa (c) PLSmulti (d) PivotMDS (e) LLE (f) ISOMAP



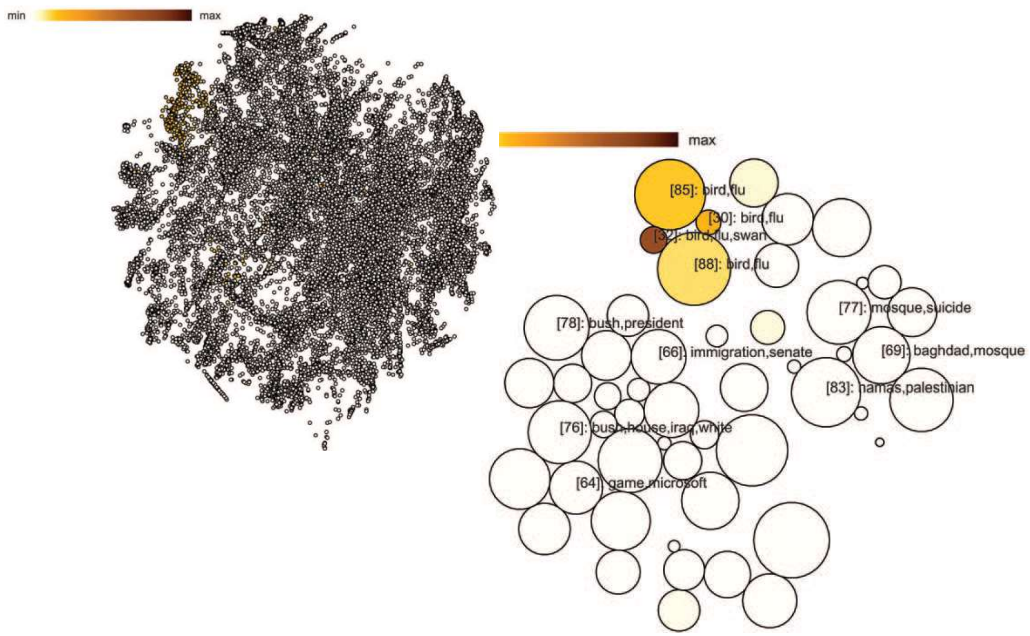
(a) 796 points (b) 1237 points (c) 1520 points (d) 1968 points (e) 2402 points (f) all 2814 points

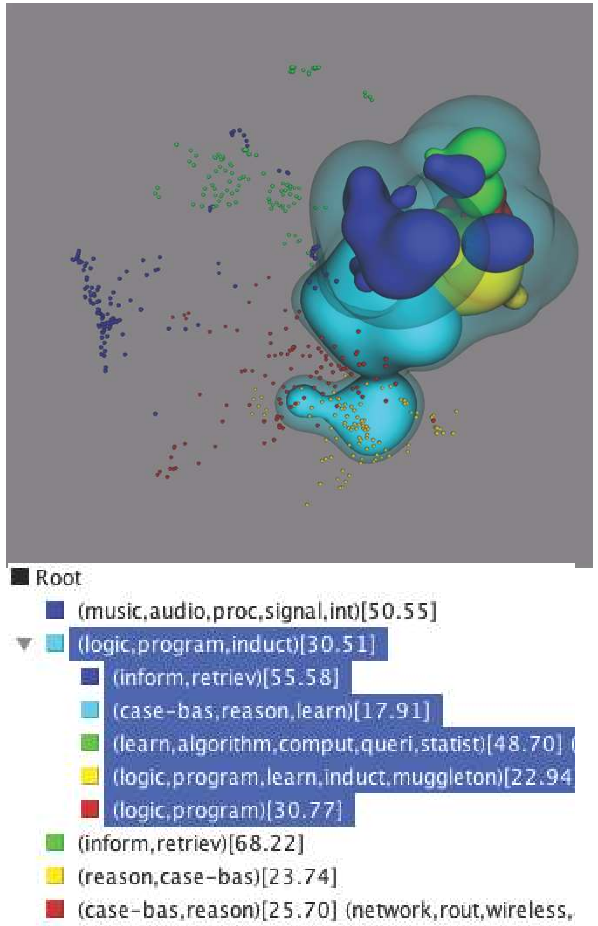
Paiva, Schwartz, Pedrini, Minghim, Semi-Supervised Dimensionality Reduction based on Partial Least Squares for Visual Analysis of High Dimensional Data**, Computer Graphics Forum, Eurovis 2012.**

Challenges

- Sheer volume
- Data transformation/formatting/structuring
- Ownership of the data
- Different types
- Spurious correlations
- Inespecificity of questions

Visualization examples: clutter

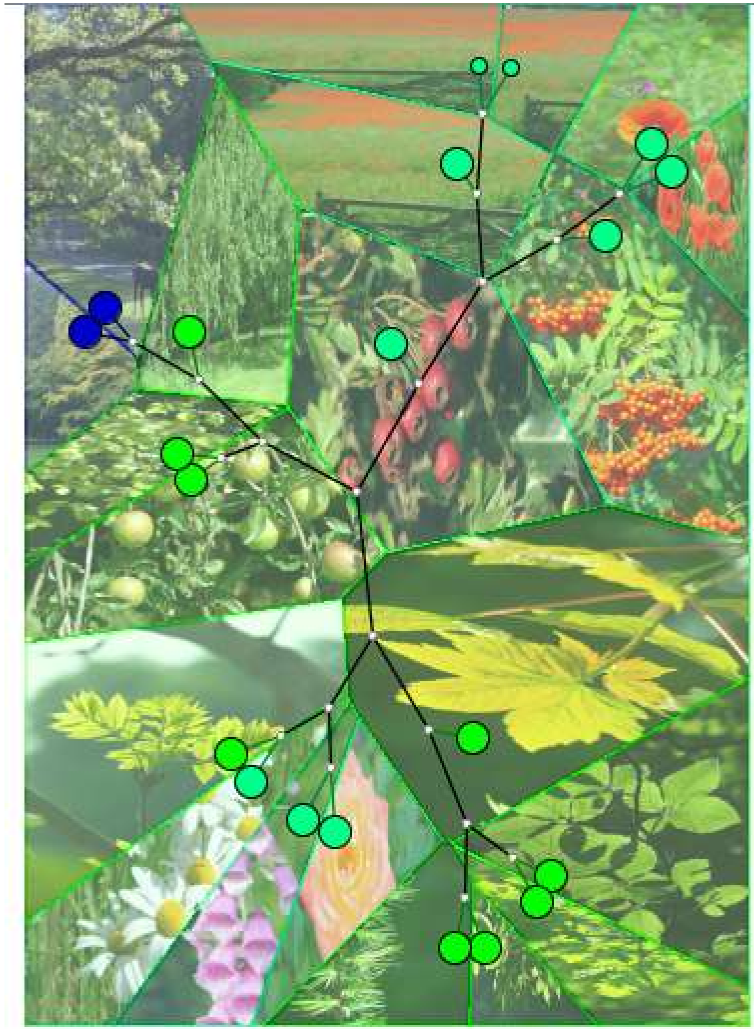


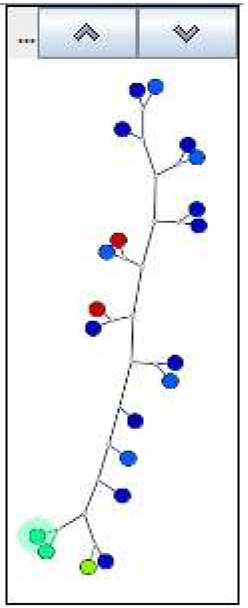


Paulovich and Minghim, HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections, ***IEEE Trans. Visualization & Computer Graphics***, *2008*

Poco; Etedmapour, Paulovich, Long, Rosenthal, Oliveira, Linsen, Minghim. A framework for exploring multidimensional data with 3D projections, ***Computer Graphics Forum***, Eurovis 2011.

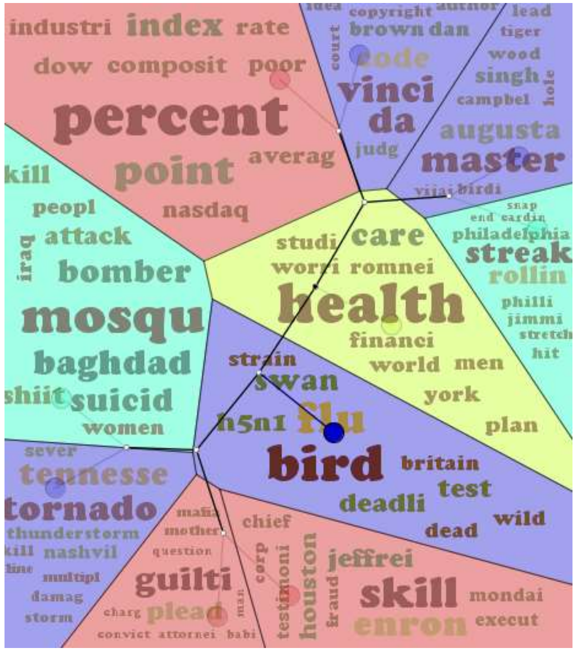
Handling scalability?
The Visual Super Tree





More Data – Summarization

- Wordclouds
- Representative Images



Multi-level text

Multi-level images

ATTRIBUTE BASED VISUALIZATION

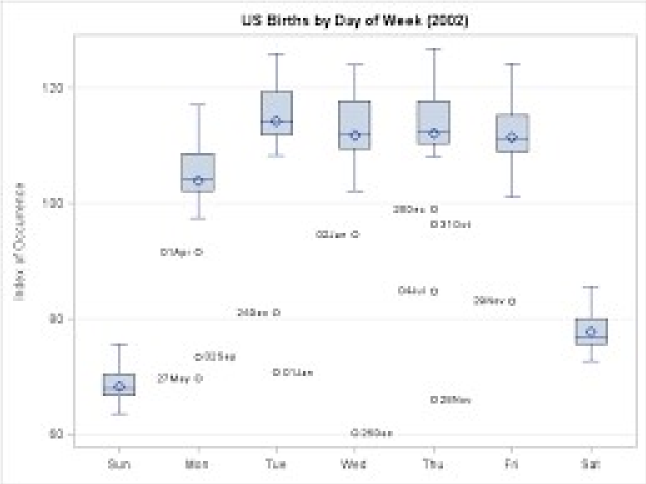
Table Views

Parallel Coordinates

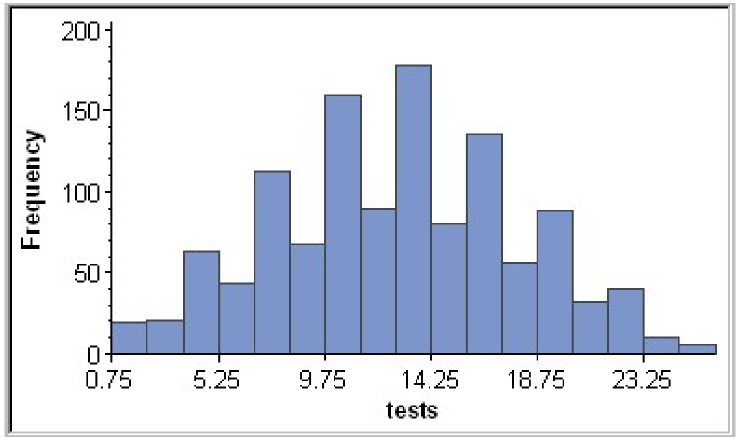
Tag Clouds

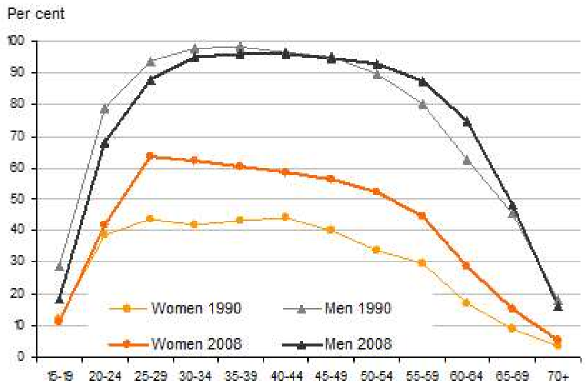
Time dependente and text

Stats

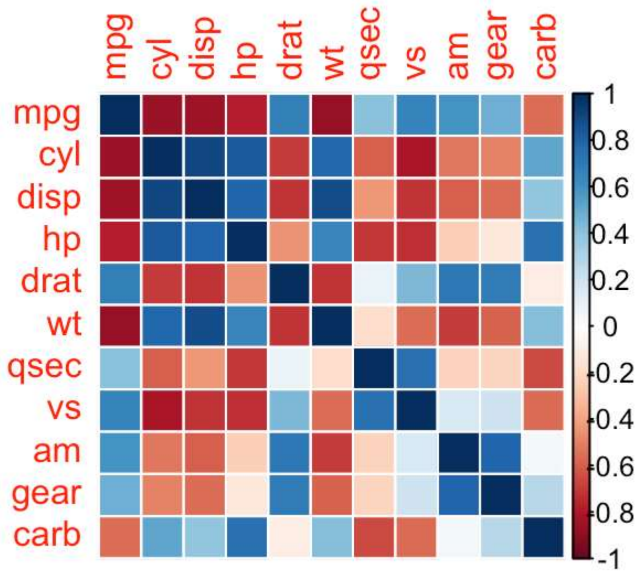

US Births by Day of Week (2002)
Index of Occurrence
120
100
80
60
Sun
Mon
Tue
Wed
Thu
Fri
Sat


Product Y
$10,000
$32,000
$25,000
$15,000
$24,000
$27,000

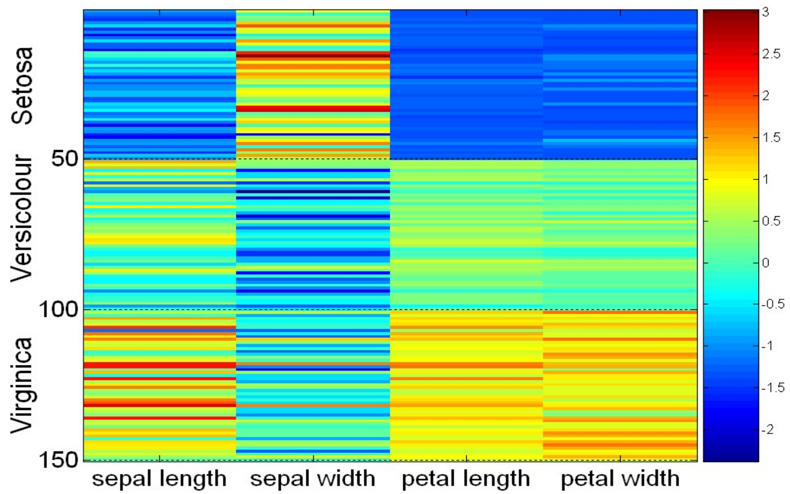

Frequency
200
150
100
50
0
0.75
5.25
9.75
14.25
18.75
23.25
tests

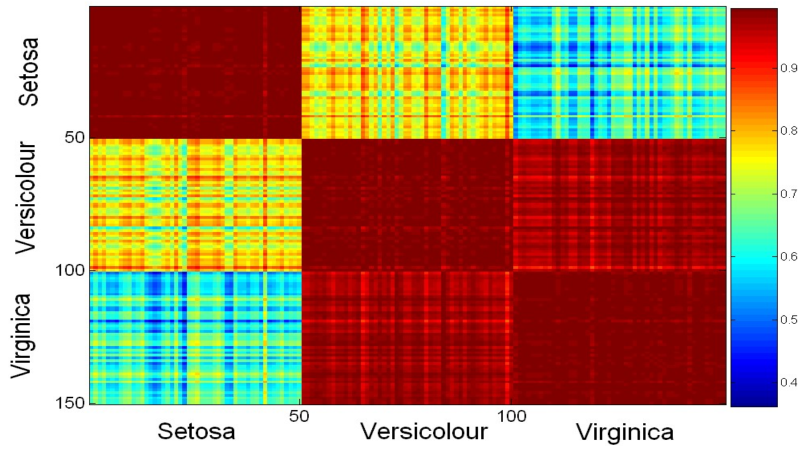

Per cent
100
90
80
70
60
50
40
30
20
10
0
15-19
20-24
25-29
30-34
35-39
40-44
45-49
50-54
55-59
60-64
65-69
70+
Women 1990
Men 1990
Women 2008
Men 2008

Data Matrix and Correlation Matrix view by heatmap

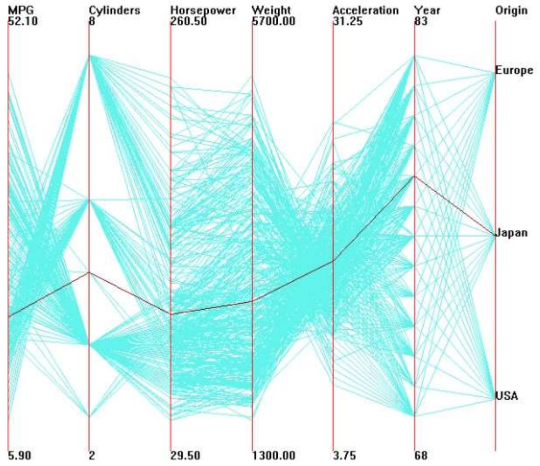


http://www.sthda.com/english/wiki/visualize-correlation-matrix-using-correlogram

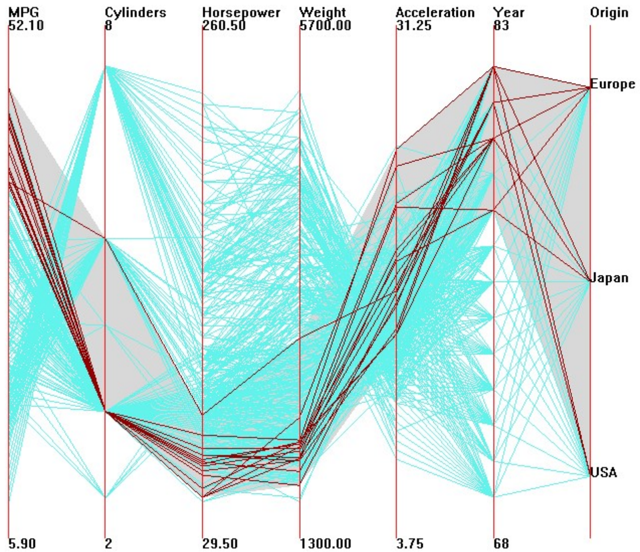


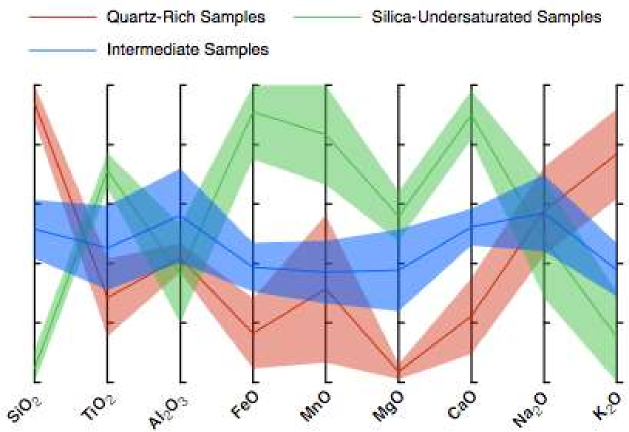


Parallel Coordinates

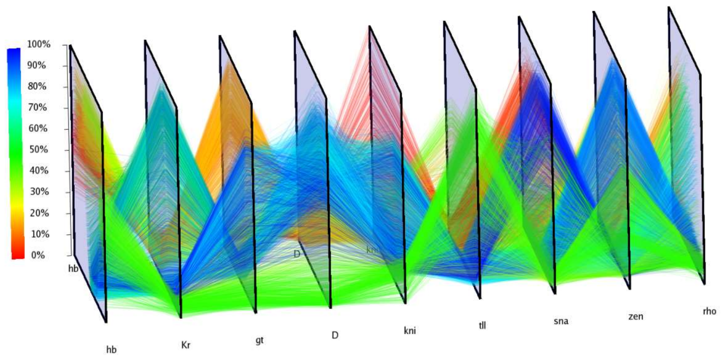


Ward, Grinstein, Keim, 2015





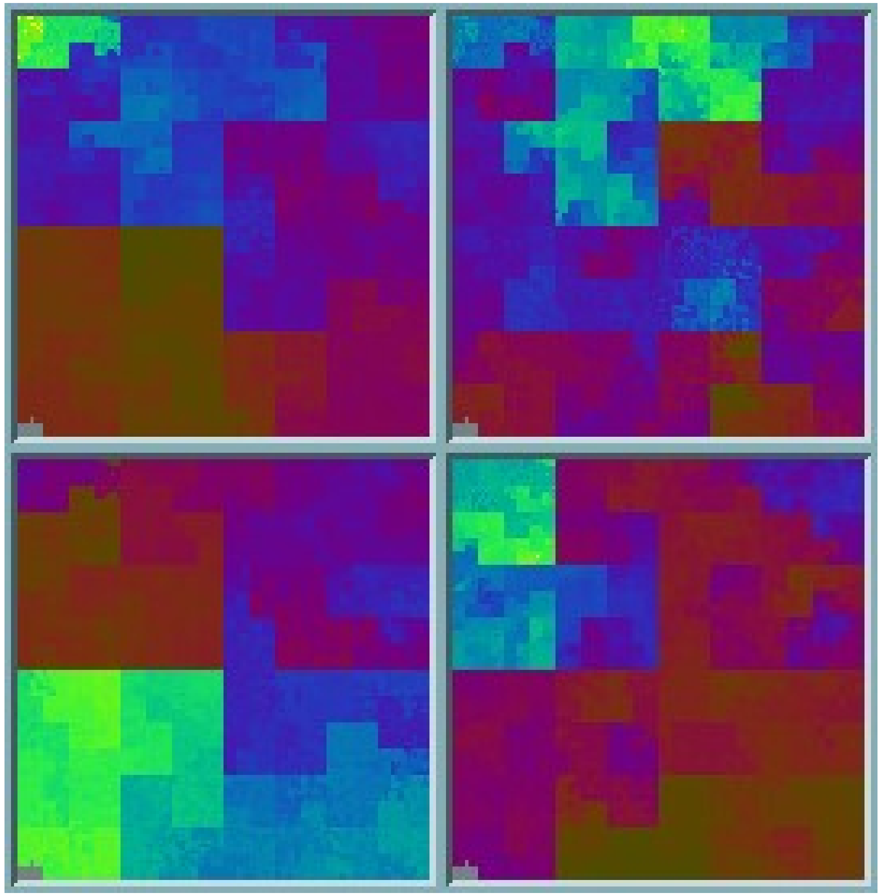
https://www.gigawiz.com/parallelco3.html



http://www-vis.lbl.gov/Events/SC07/Drosophila/3DParallelCoordinates.png

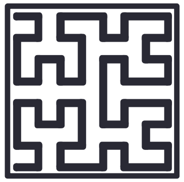
Pixel-based techniques

IBM DOLLAR

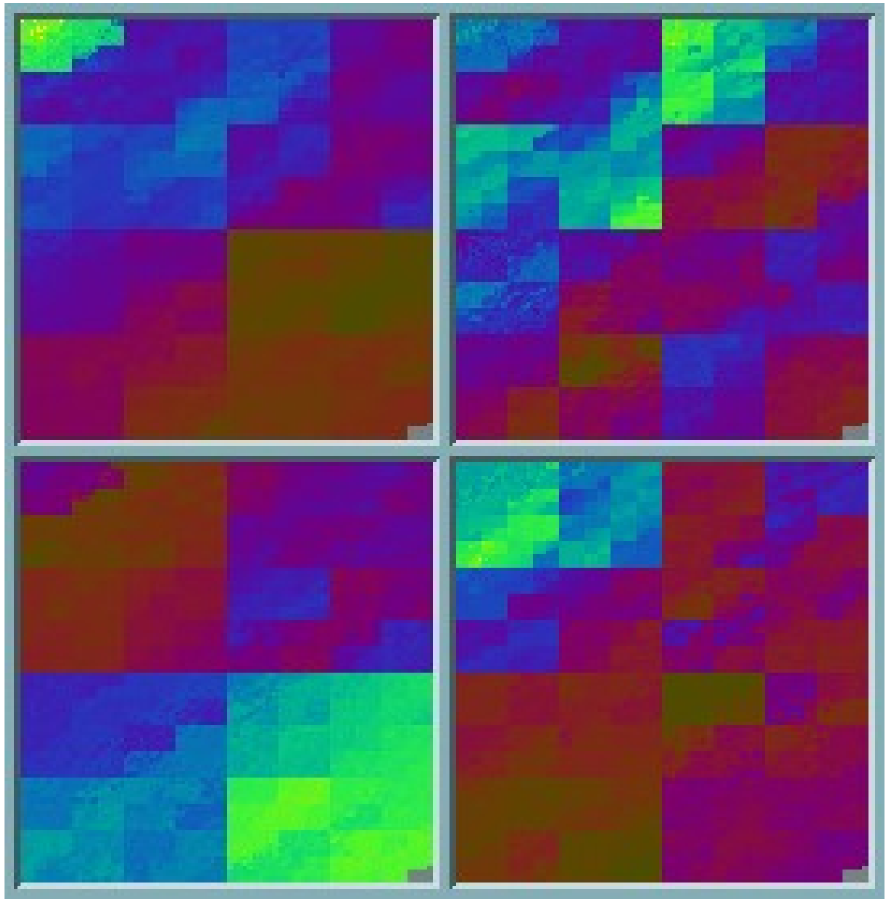


DOW JONES GOLD.US$

Peano-Hilbert

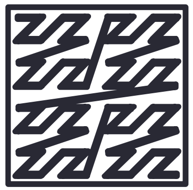


IBM DOLLAR



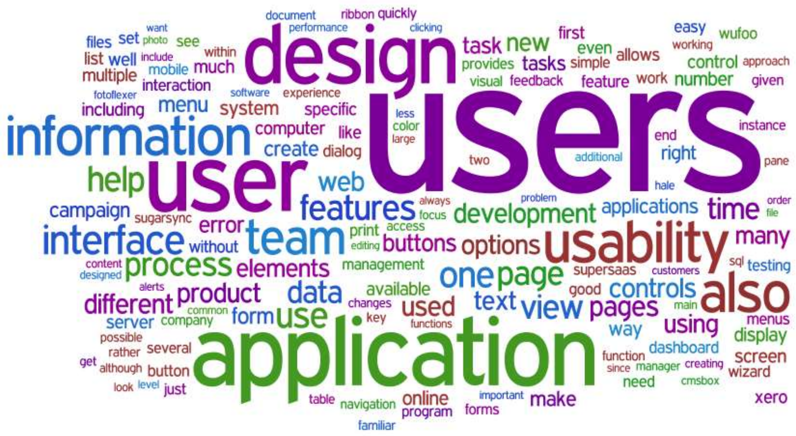
DOW JONES GOLD.US$

Morton (Z-Curve)



Space-Filling Curves

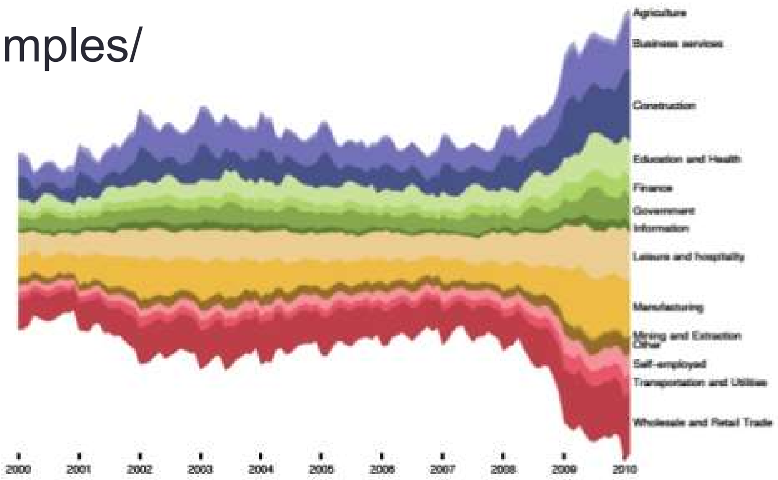
Tag-clouds and Theme River



https://www.nngroup.com/articles/tag-cloud-examples/

Stacked graph

aka Steam graph, Theme river



Heer et al, 2010

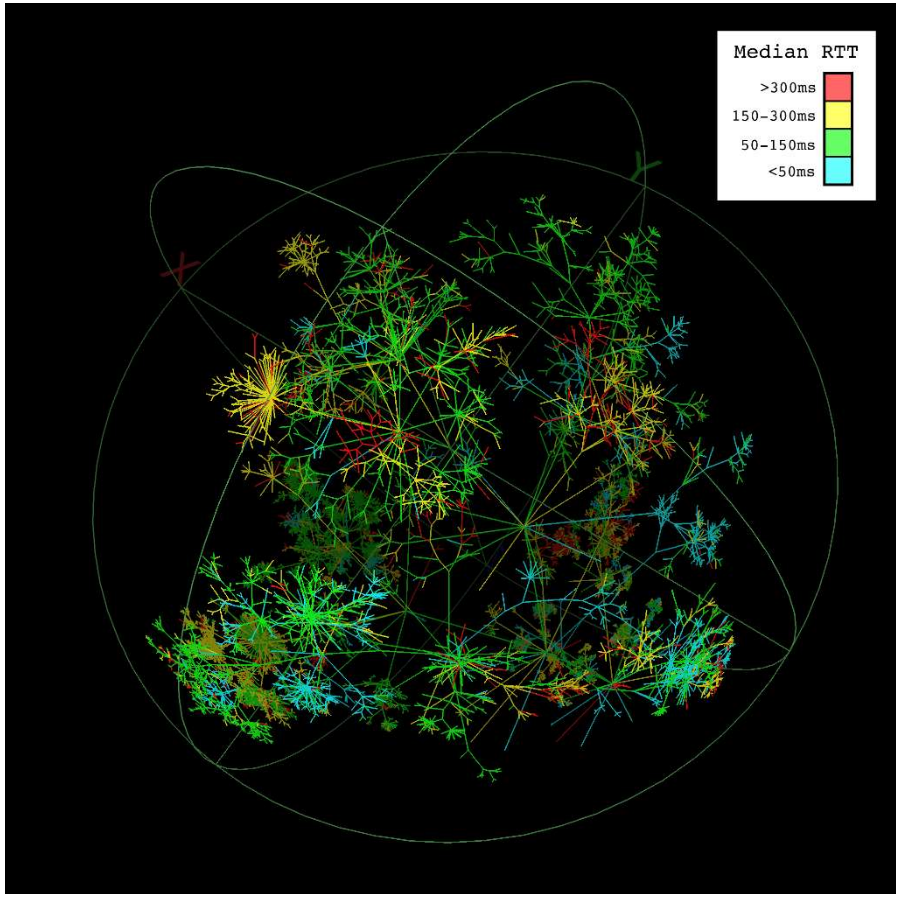
http://complexdatavisualized.com/time-series-visualizations-an-overview/

RELATIONSHIP – BASED VISUALIZATION

Graphs

Trees

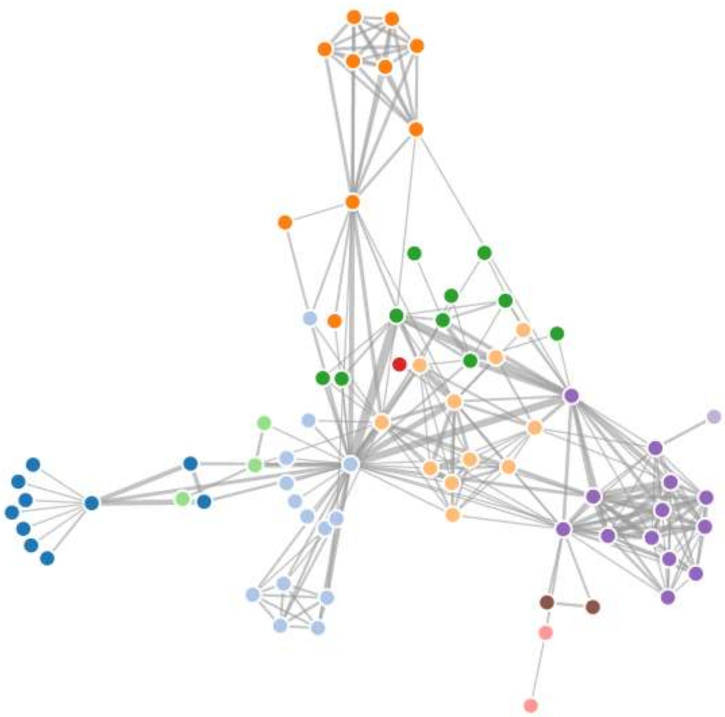
Graphs and trees can be large



https://www.caida.org/research/performance/rtt/walrus0202/a-root-rtt-05-key.png

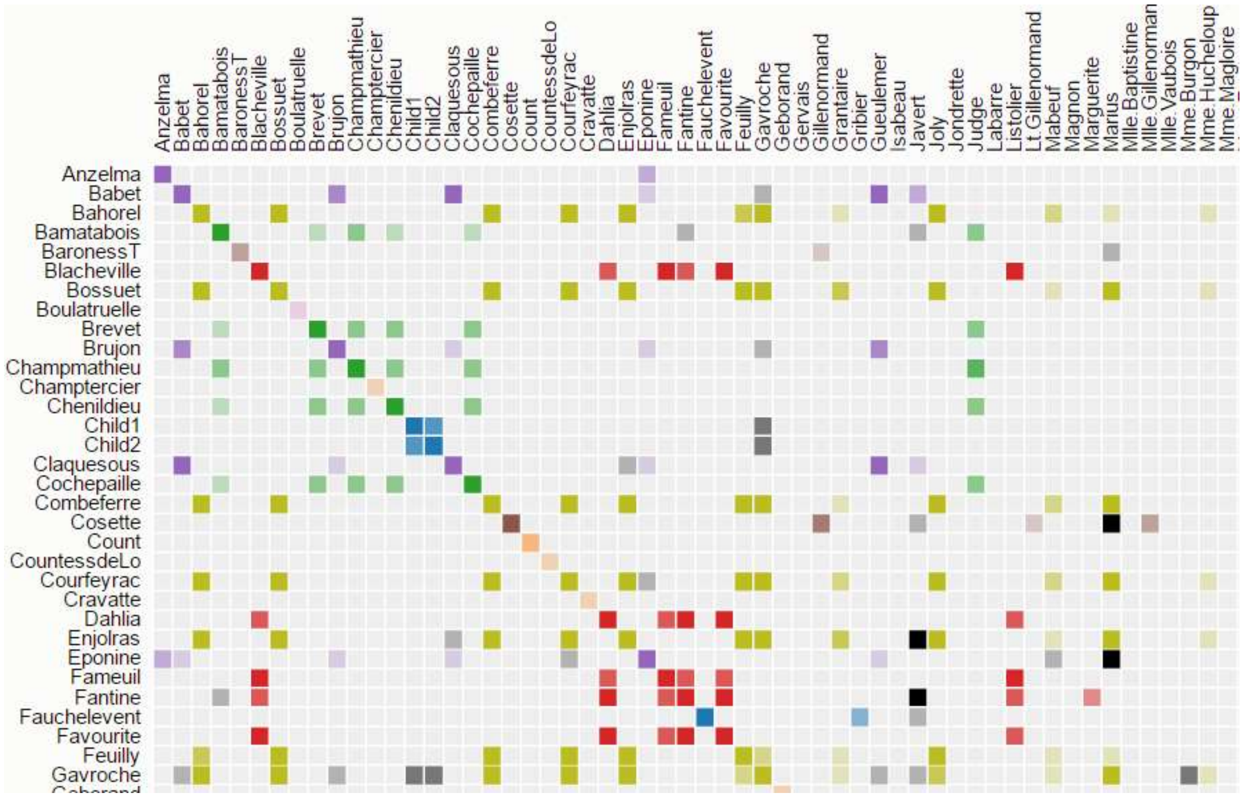
Graphs: Force-directed Graph Layout

- http://bl.ocks.org/mbostock/4062045

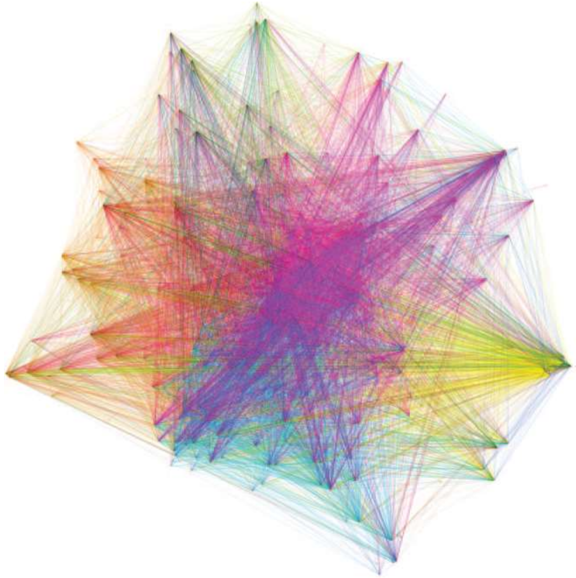


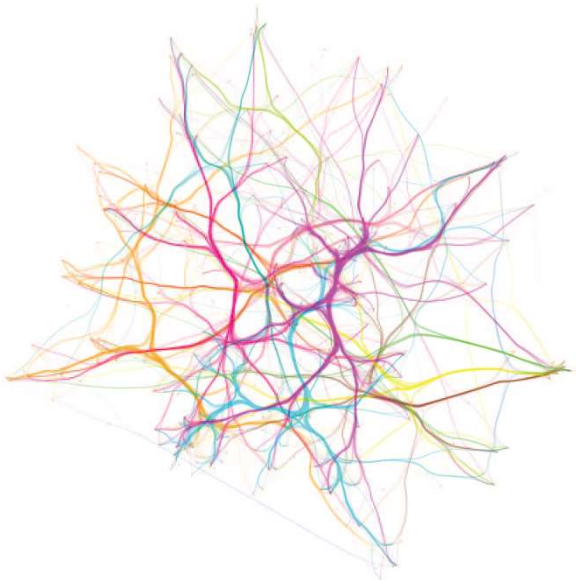
Adjacency Matrix Graph Layout

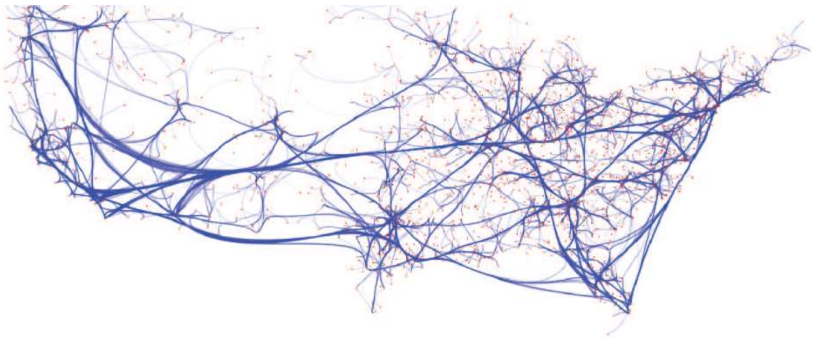
- https://bost.ocks.org/mike/miserables/



Clutter – graph bundling

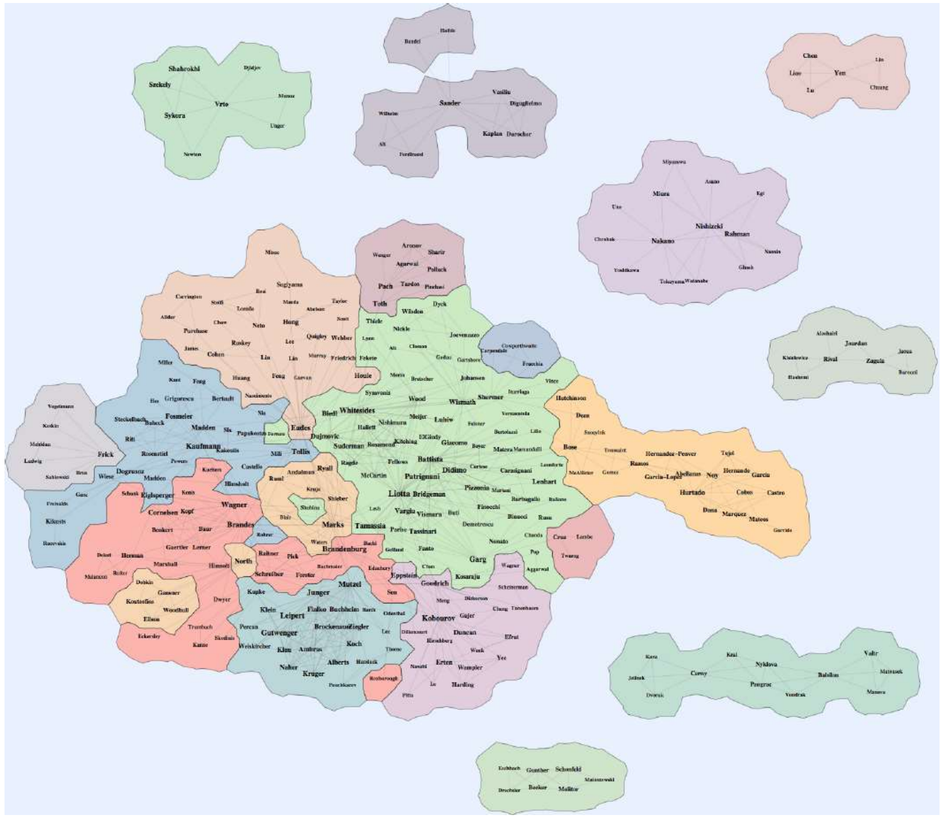


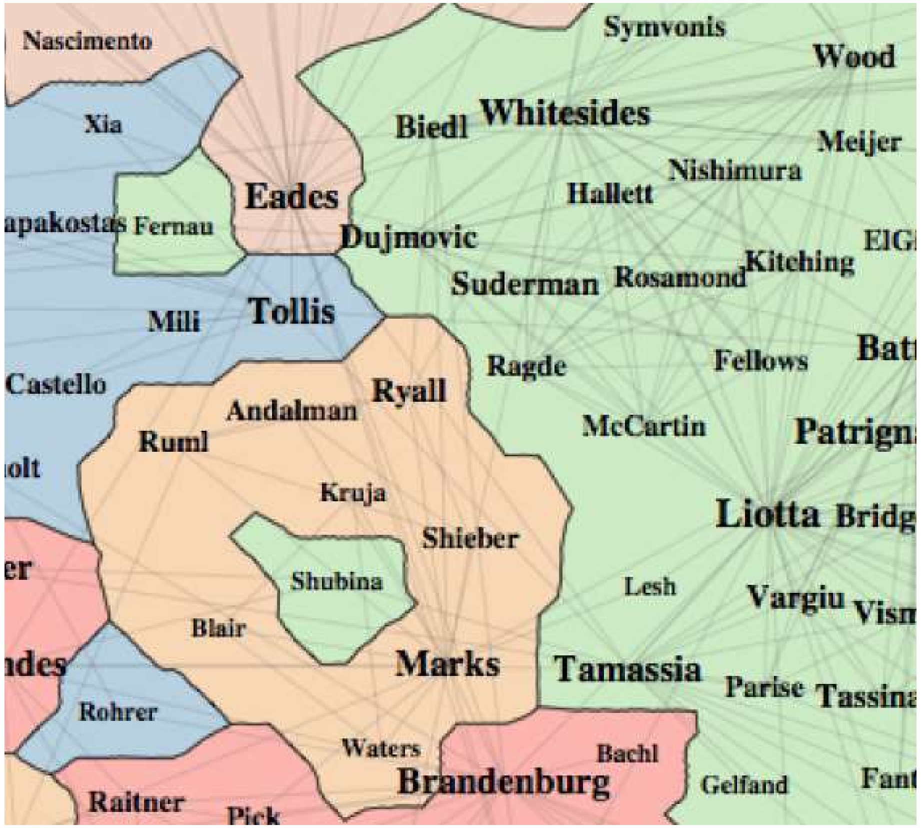




Readjusting Layouts on
Cluster-based Readjustment

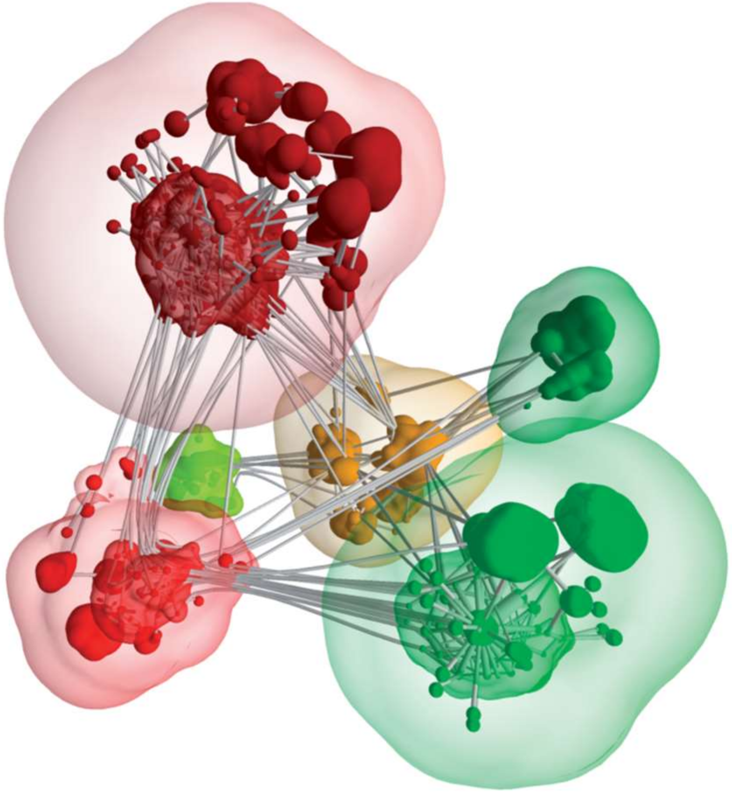
- **2D Abstraction:** Gmap [Gansner et al. 2010]

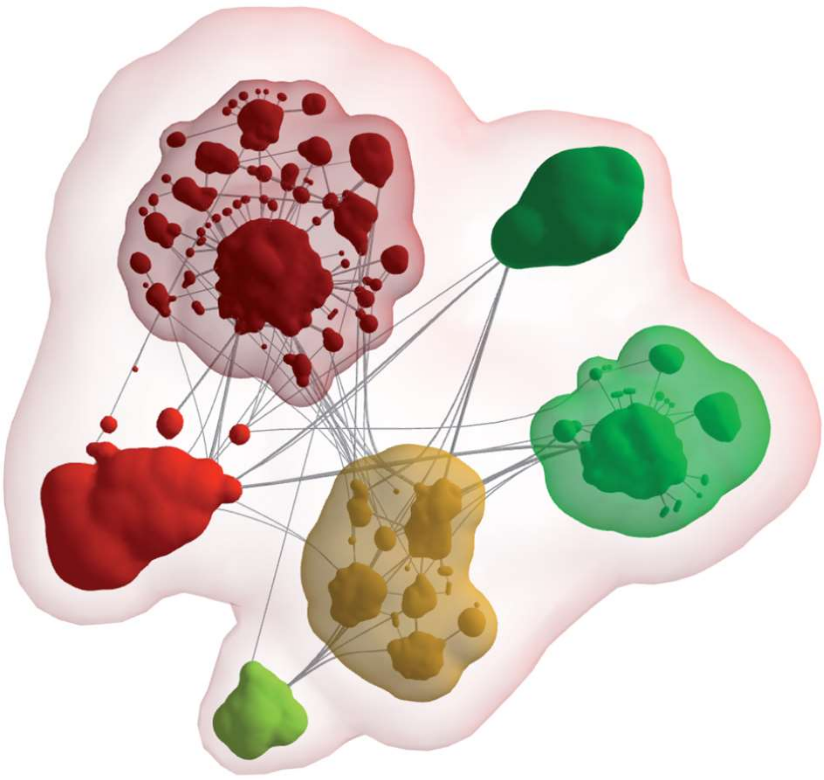




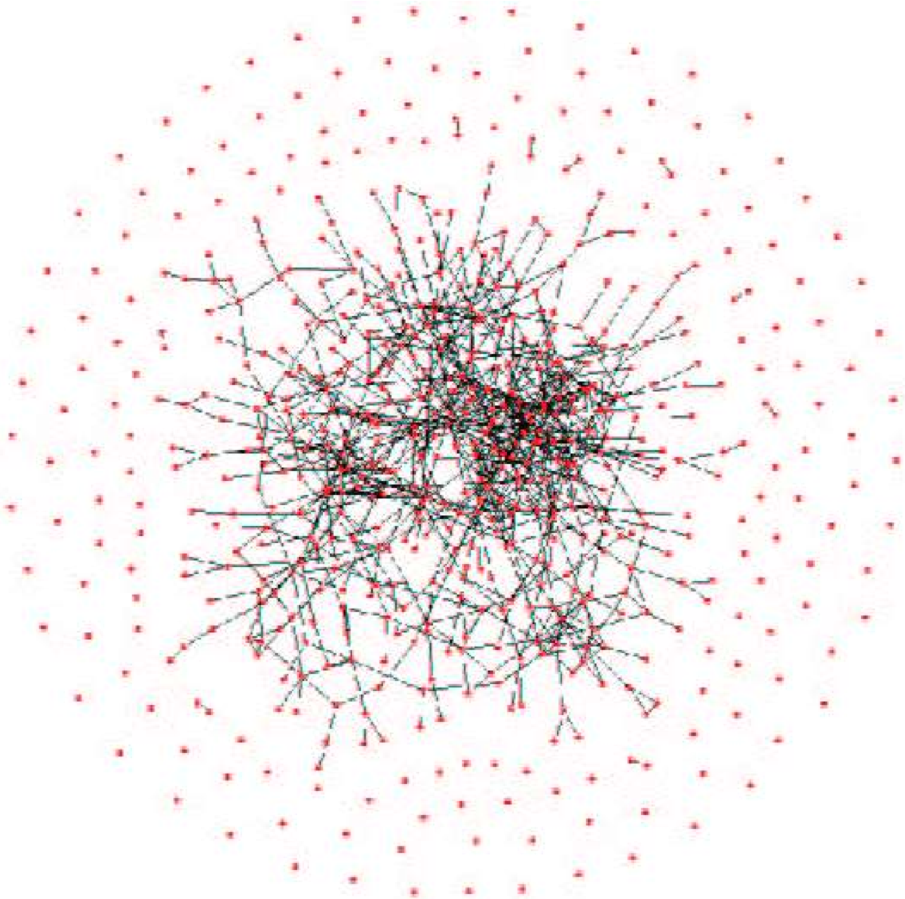
Readjusting Layouts on
Cluster-based Readjustment

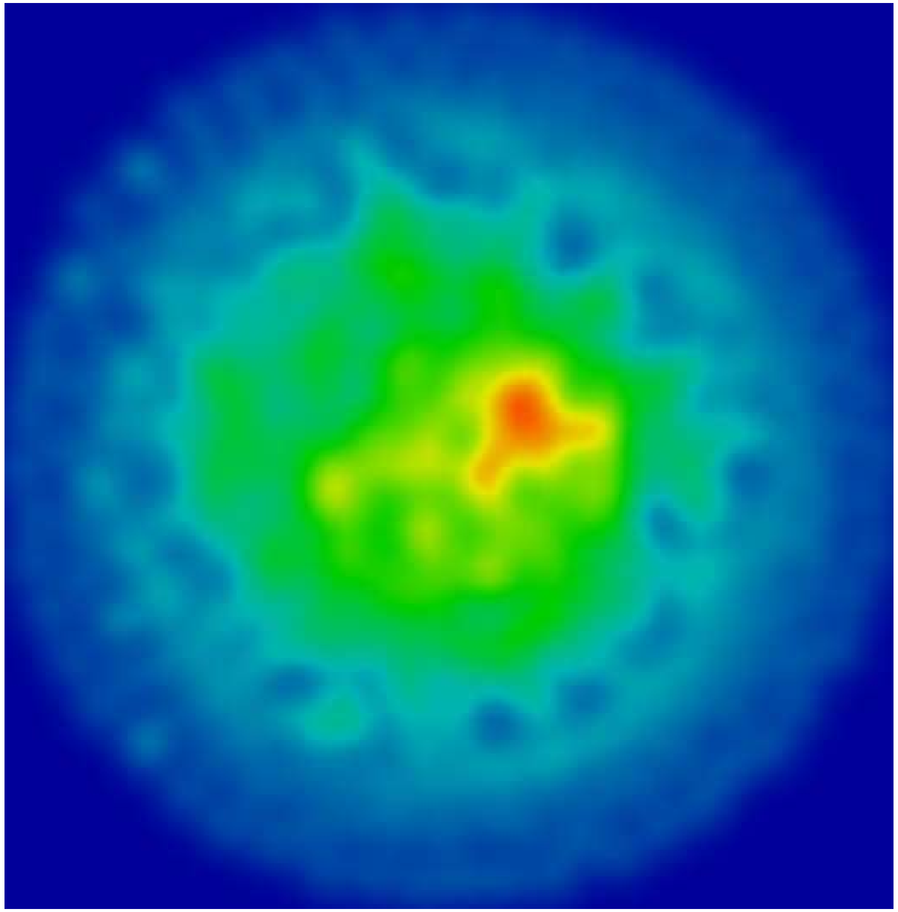
- **3D Abstraction:** Graph Cluster Surfaces
  [Balzer+Deussen 2007]





Density and attribute mapping



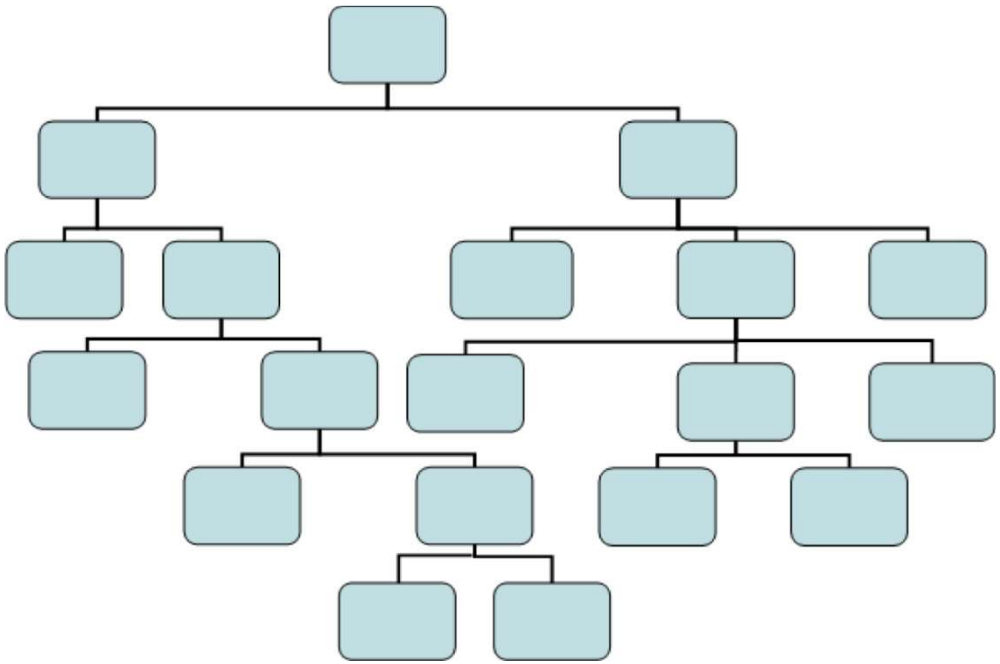


Trees

A hierarchy must be present, detected or imposed

Many layouts and applications: see http://treevis.net

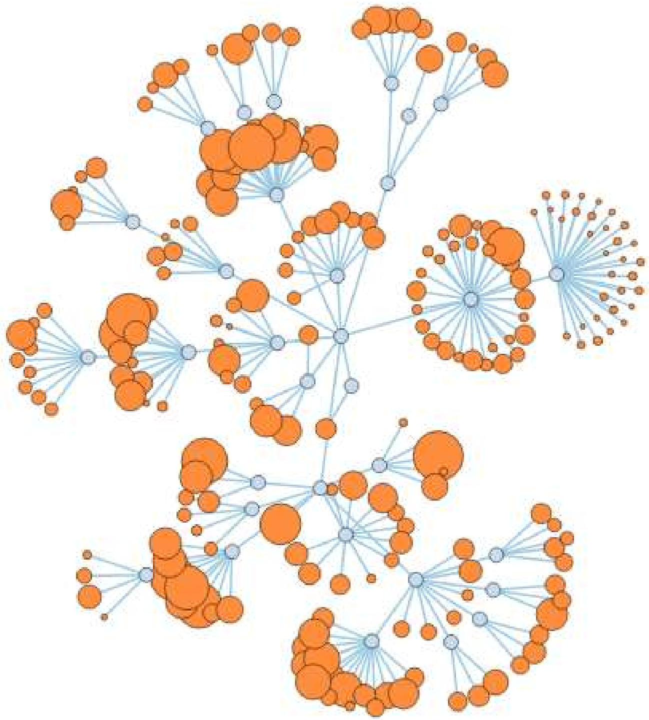
Standard layout – link-node (nó e aresta)

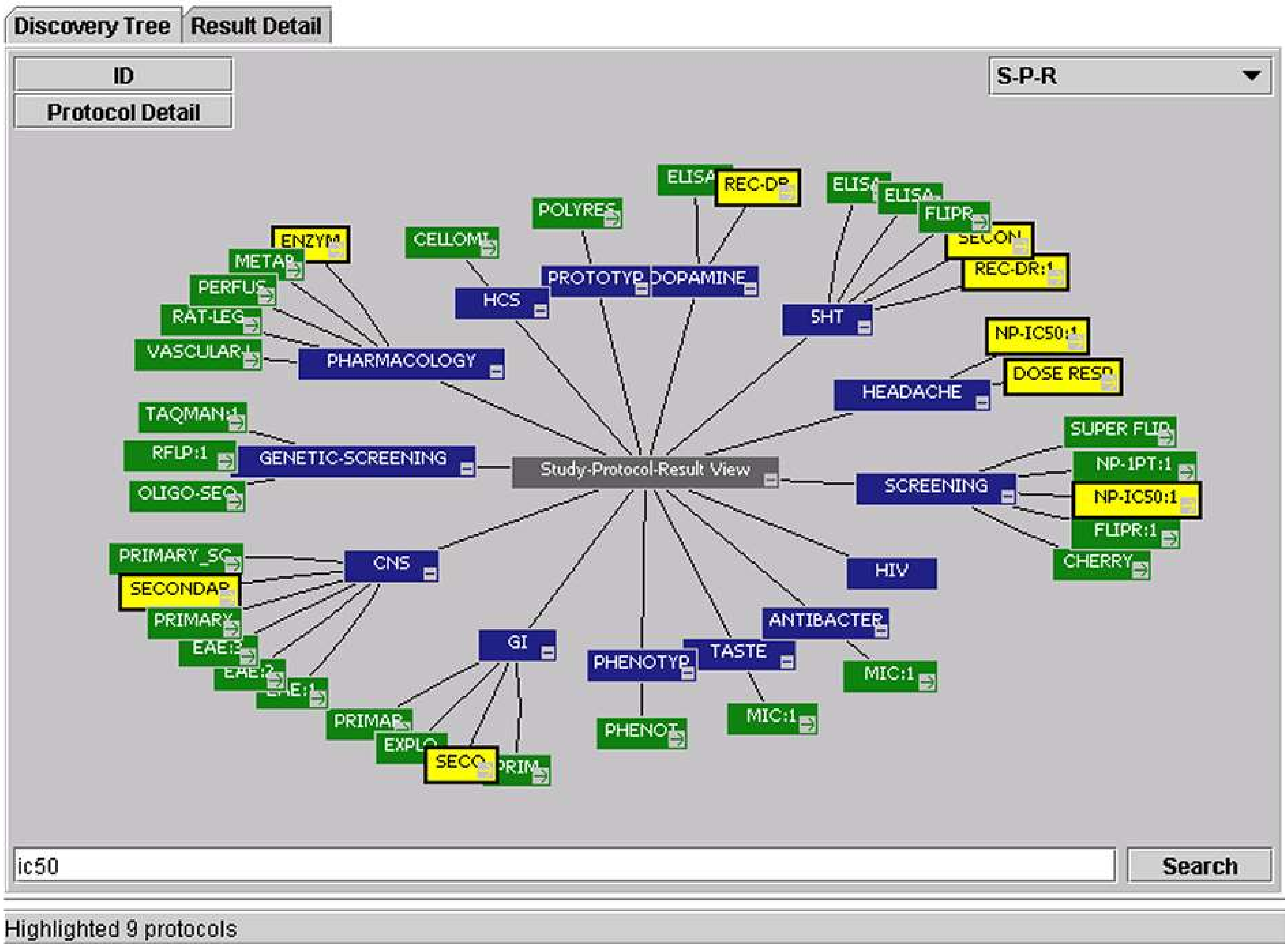


Ward et al. 2015

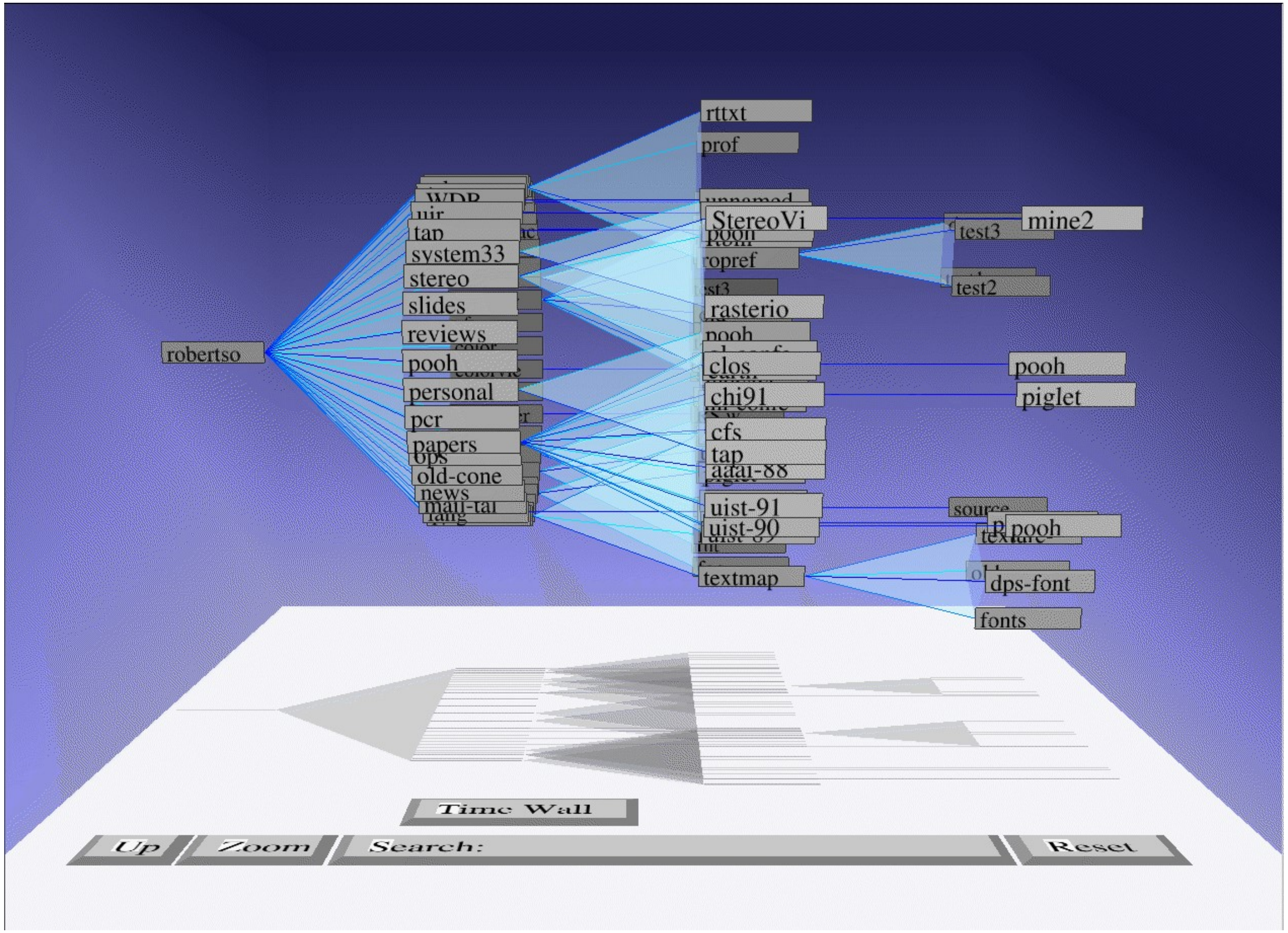
Trees - Force Layout

- http://mbostock.github.io/d3/talk/20111116/force-collapsible.html




Discovery Tree
Result Detail
ID
Protocol Detail
S-P-R
Study-Protocol-Result View
PHARMACOLOGY
GENETIC-SCREENING
CNS
GI
PHENOTYP
TASTE
ANTIBACTER
HIV
SCREENING
HEADACHE
5HT
DOPAMINE
PROTOTYP
HCS
ENZYM
METAB
PERFUS
RAT-LEG
VASCULAR-I
CELLOMI
POLYRES
ELISA
REC-DR
FLIPR
SECON
REC-DR:1
NP-IC50:1
DOSE RESP
SUPER FLIP
NP-1PT:1
NP-IC50:1
FLIPR:1
CHERRY
TAQMAN:1
RFLP:1
OLIGO-SEQ
PRIMARY_SC
SECONDAR
PRIMARY
EAE:3
EAE:2
EAE:1
PRIMAR
EXPLO
SECO
PRIM
PHENOT
MIC:1
MIC:1
ic50
Search
Highlighted 9 protocols

Cone Tree

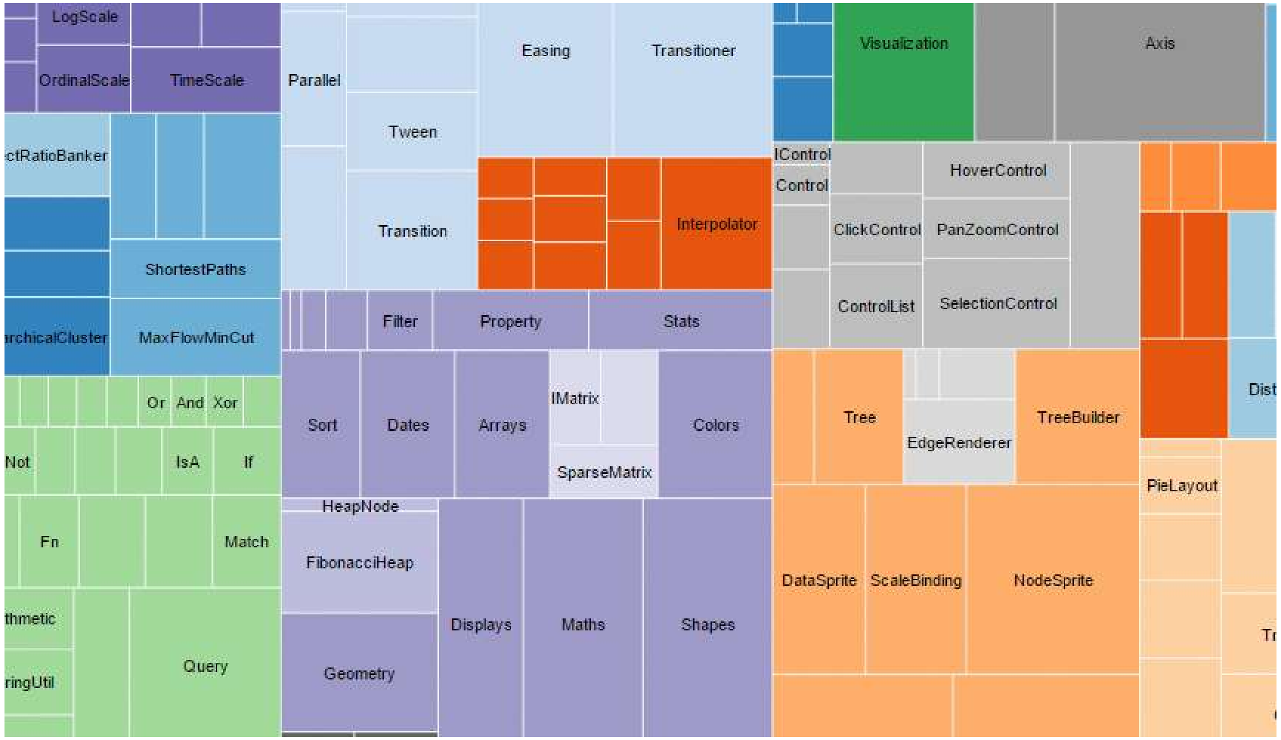


File system structure visualized as a cone tree

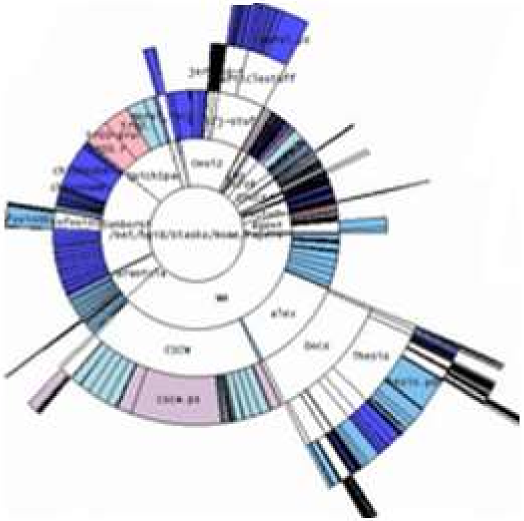
Animated 3D visualization of hierarchical data

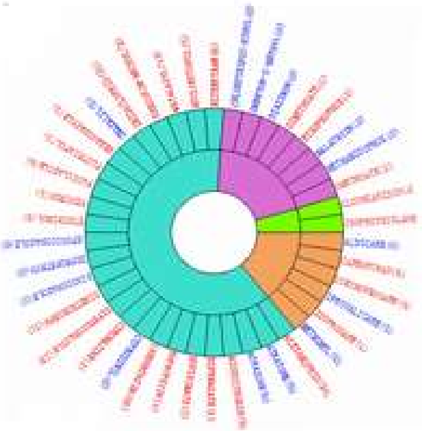
Tree - TreeMap

- http://mbostock.github.io/d3/talk/20111018/treemap.html



Trees – Sunburst and Curved Trees



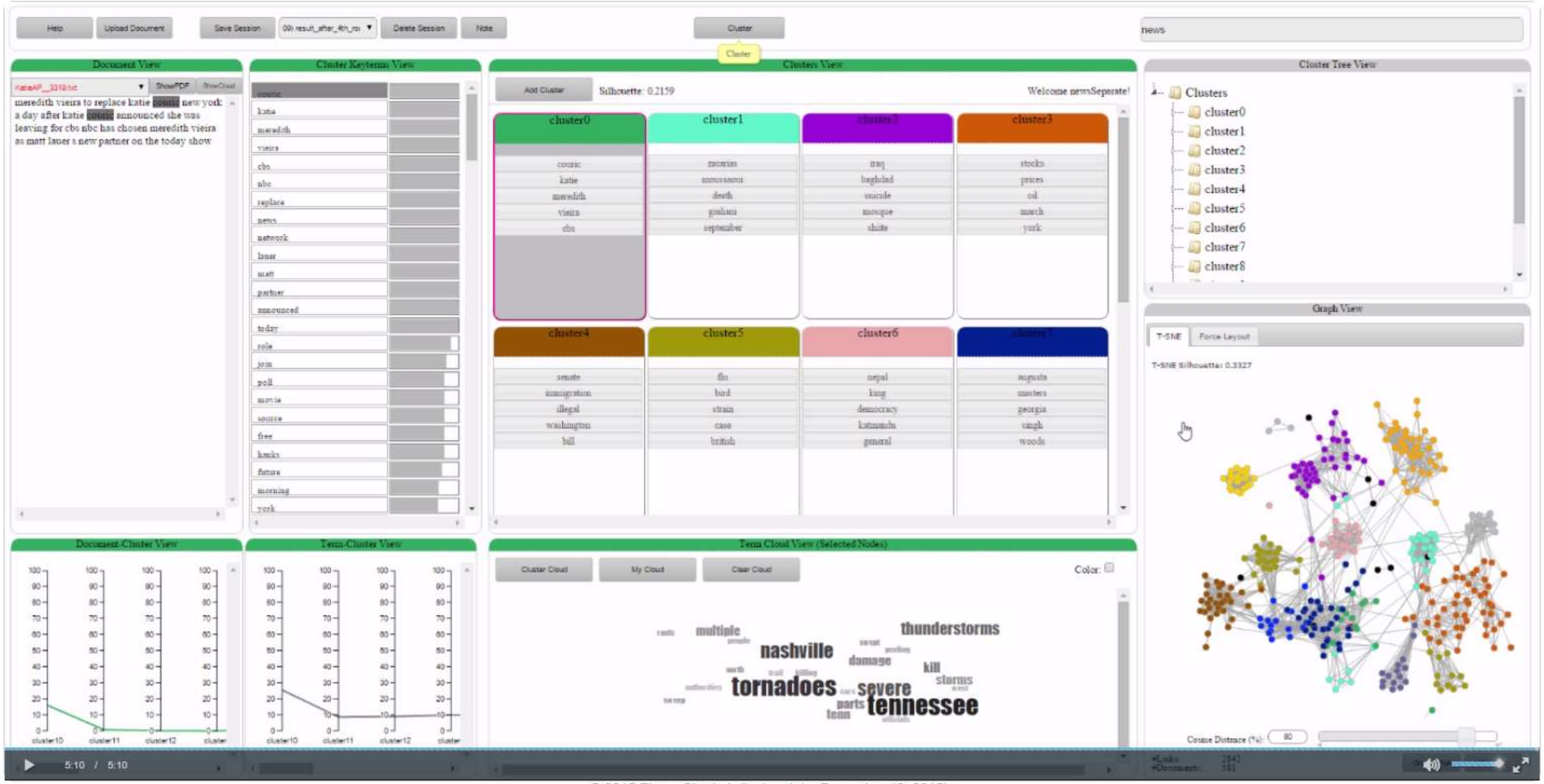


M. Ali Rostami, Azin Azadi and H. Martin Bücker, A new approach to visualizing general trees using thickness-adjustable quadratic curves, GD'14: Proceedings of the International Symposium on Graph Drawing, pages 525-52.
Chen, Y., Zhang, X., Feng, Y. et al. Sunburst with ordered nodes based on hierarchical clustering: a visual analyzing method for associated hierarchical pesticide residue data. J Vis (2015) 18-237.
Stasko, Catrambone, Guzdial, McDonald 2000: An evaluation of space-filling information visualizations for depiction hierarchical structures

Visualization for Clustering

- User: important role in cluster analysis
- User perspective is of importance in many applications
- IvisClustering (LDA-based clustering)
- JigsawCluster
- Kt-vis (keyterm-based clustering)

Visualization for Clustering kt-vis



© 2015 Ehsan Sherkat. (last update: December 10, 2015)

Nourashrafeddin, Sherkat, Minghim, Milios – A Visual Approach for Interactive Keyterm-based Clustering - Submitted ACM TIIS

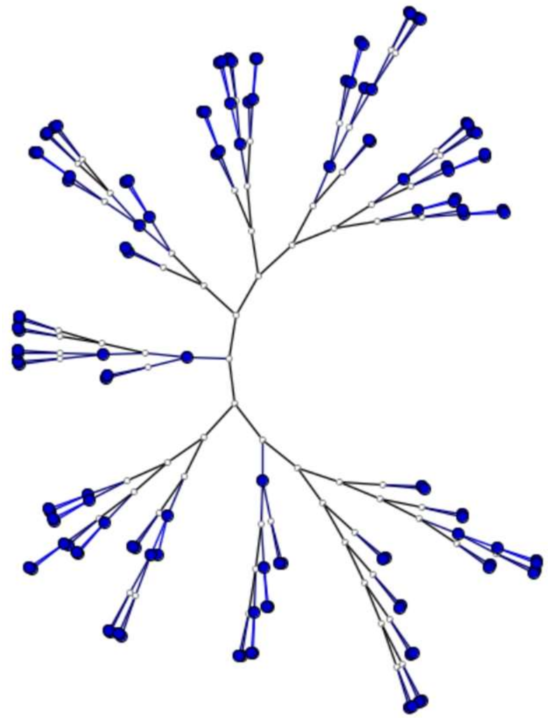
Visualization for Classification

- User: important role in building, applying and adjusting classifiers
  - Knowledge of the problem
  - Insertion of the classification process

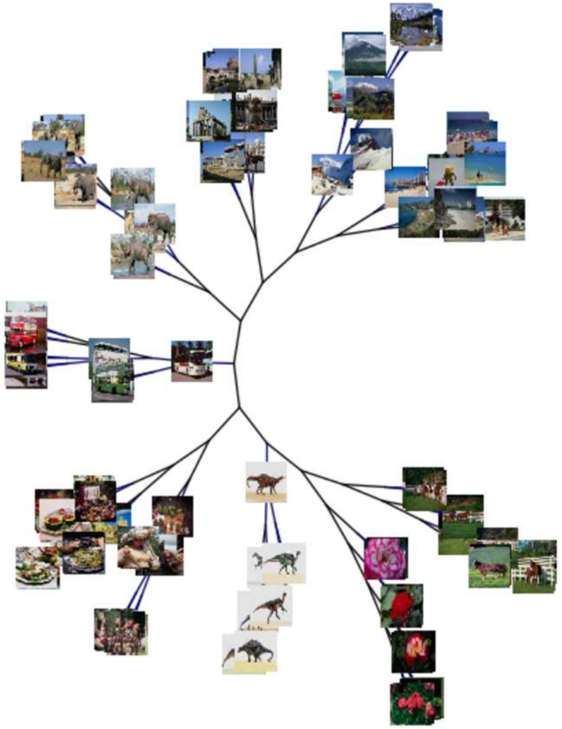
- Insertion may be more effective: better data sets presentation
  - Data set structure and instances relationship understanding
  - Detection of specificities that justify classifiers behaviors

Paiva, J.G.S., Schwartz, W. R., Pedrini, H, Minghim, R. An Approach to Supporting Incremental Visual Data Classification. IEEE Transactions on Visualization and Computer Graphics, v. 21, p. 4-17, 2015

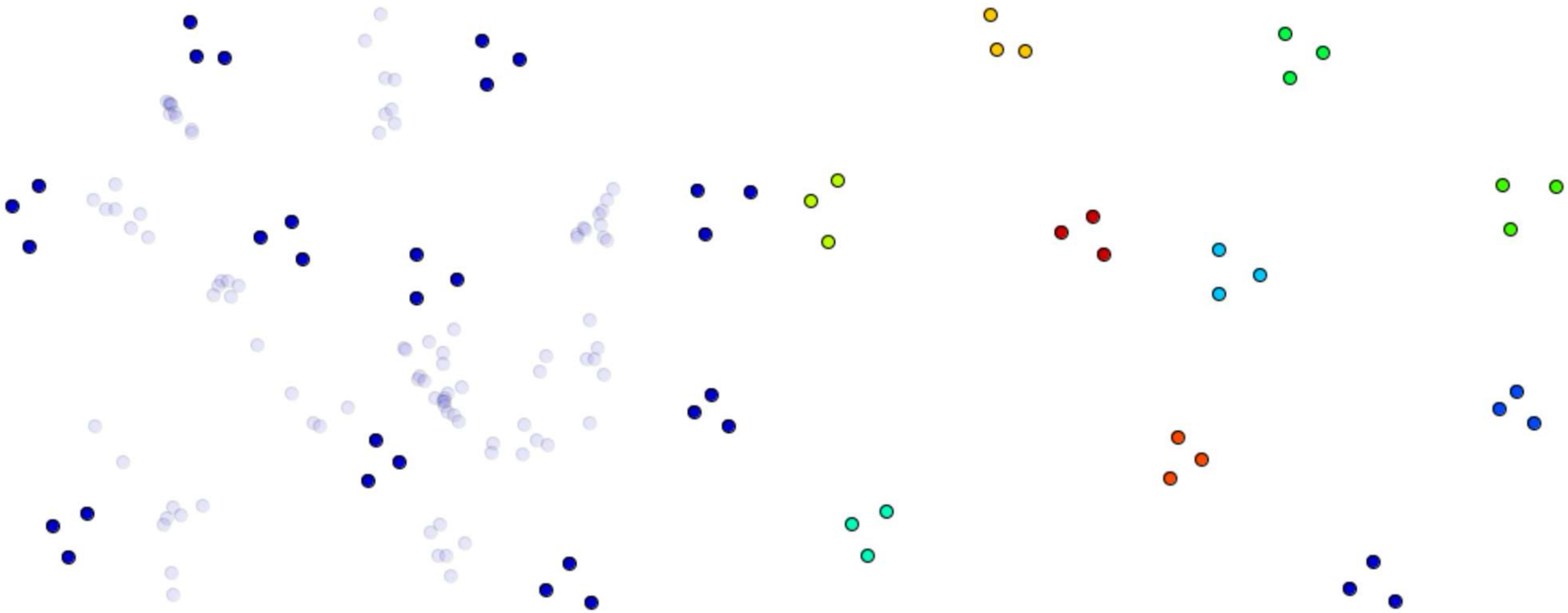
Similarity Organization



Similarity Organization

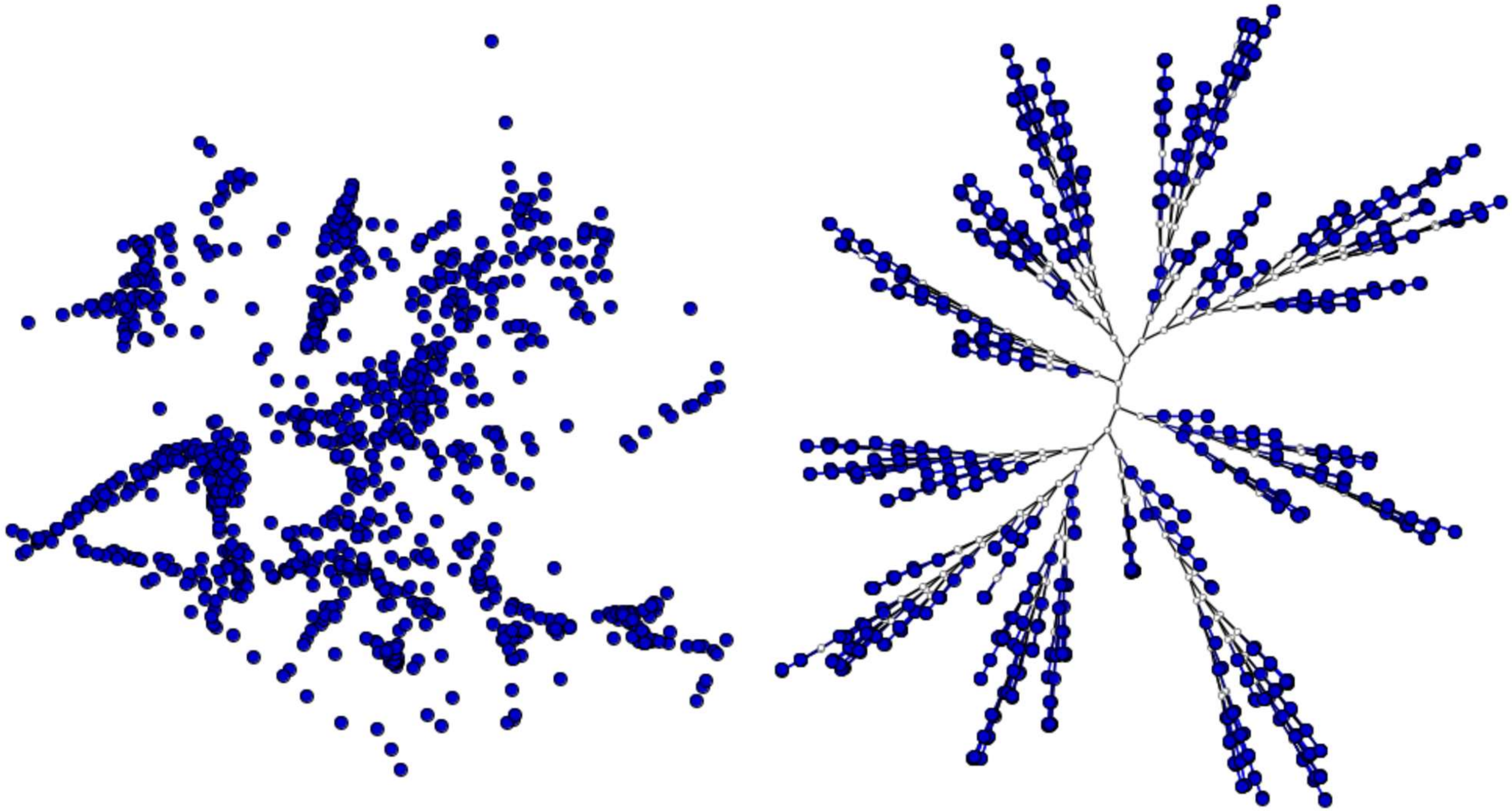


Selection of Representative Instances

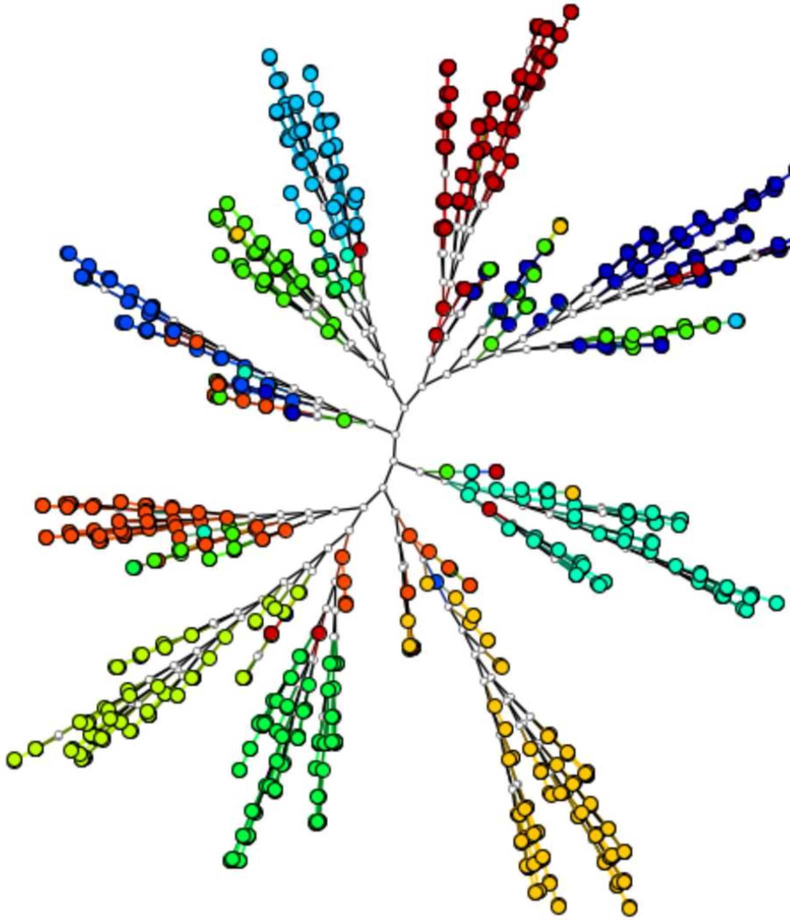


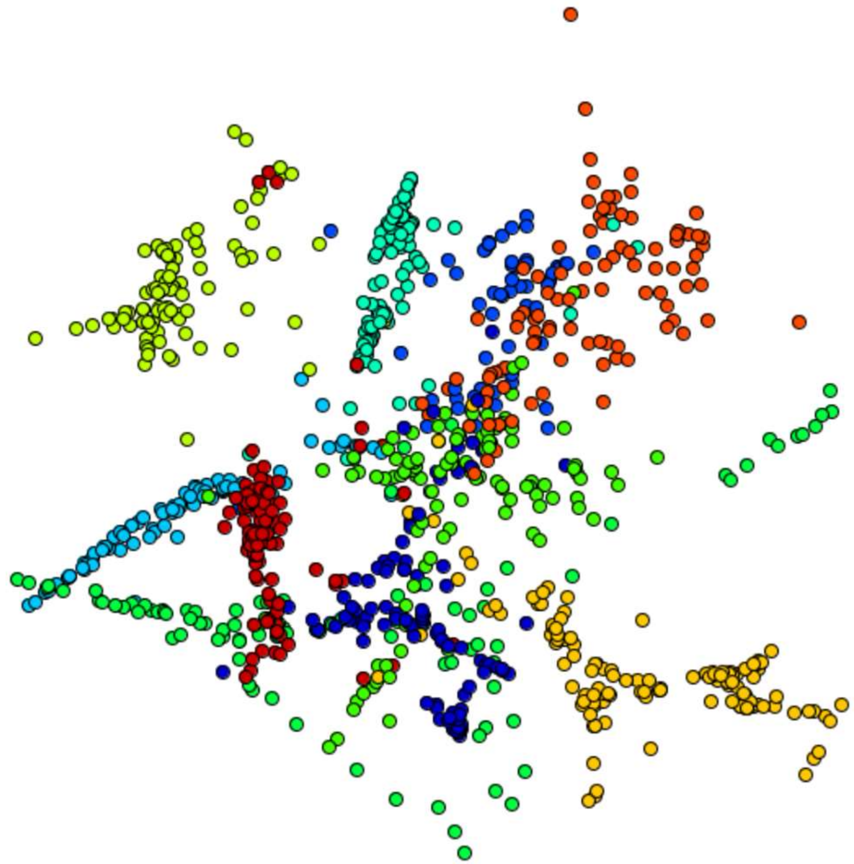
Instances selected to train classification model

Classification using Created Model

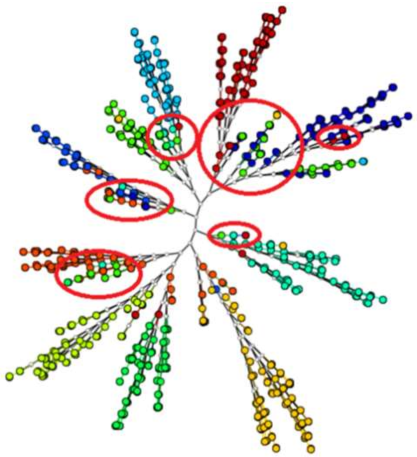


Classification Results

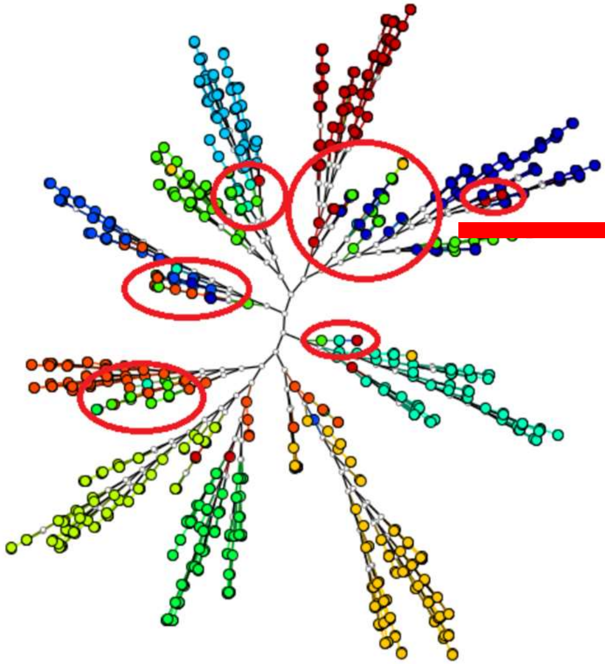




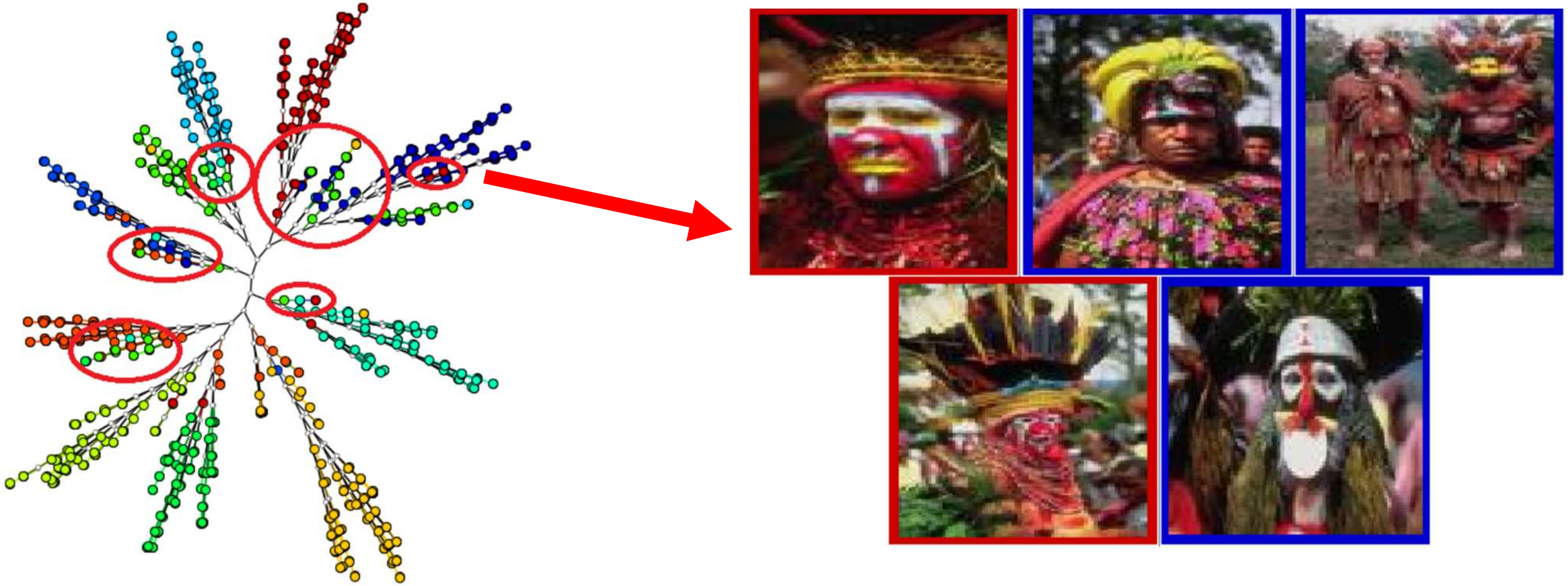
Evaluation: Classification Results



Evaluation: Classification Results



Evaluation: Classification Results



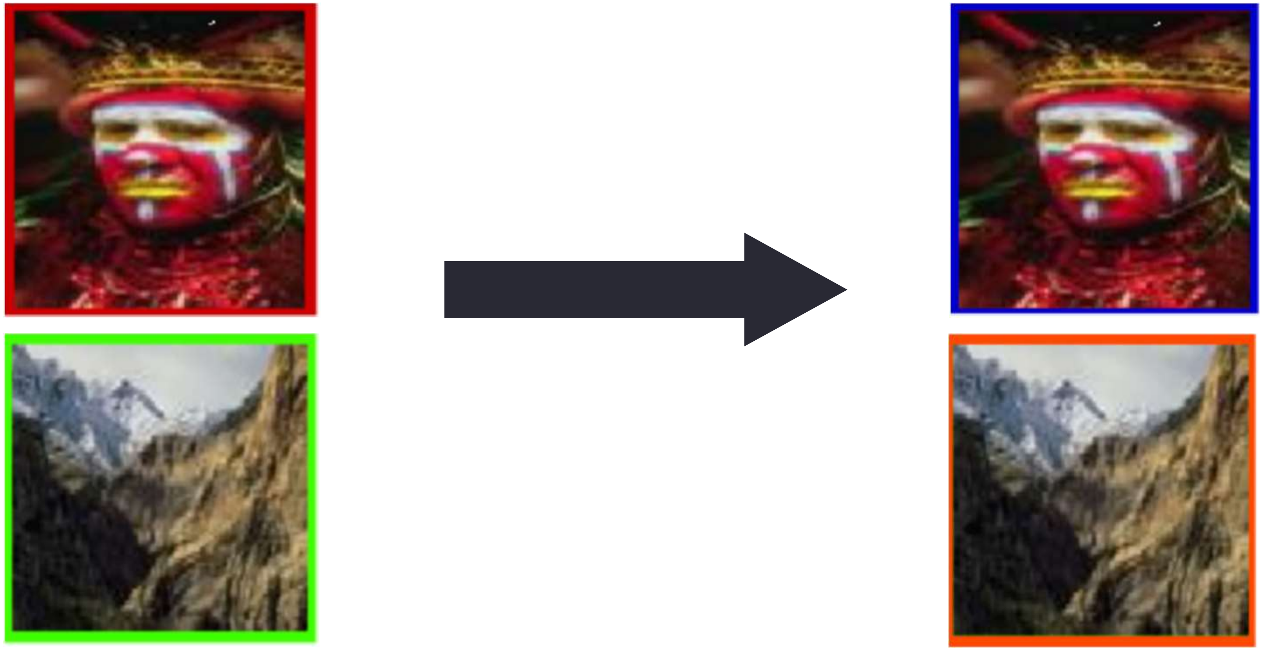
Evaluation: Classification Results

Classification Model Upgrading

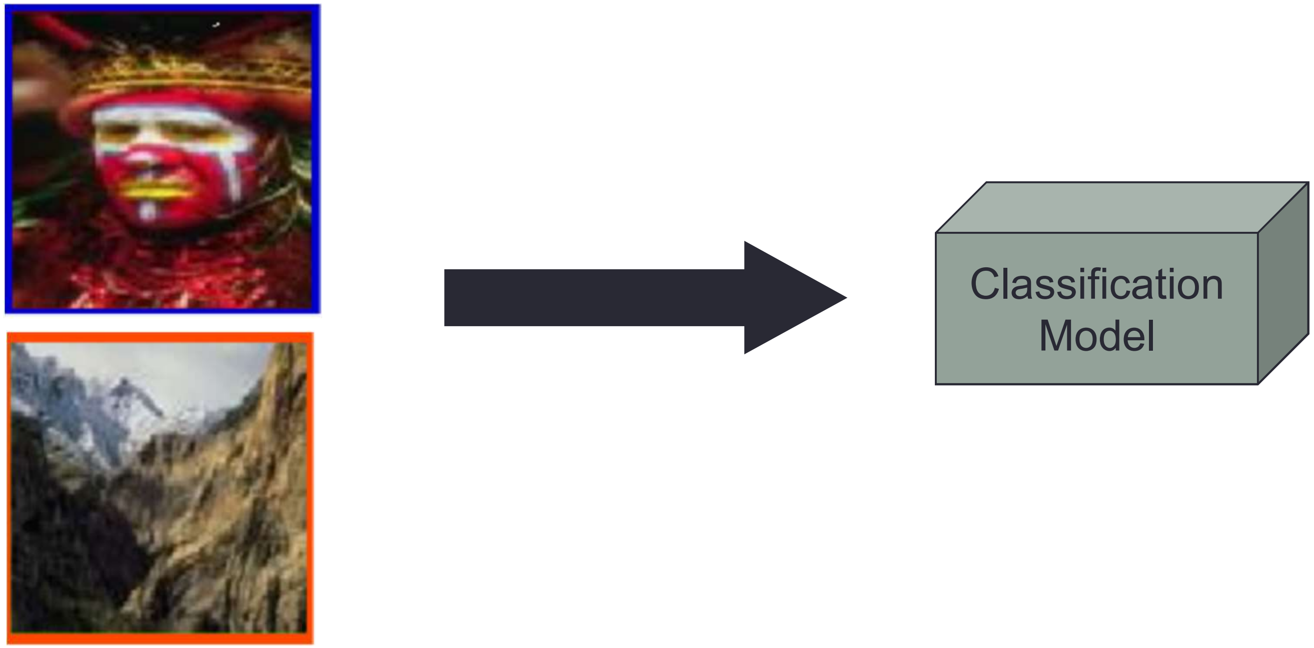
- Several upgrade strategies: Layout also works as a guide
  - Example: relabeling of strategic instances: adjustment to specific scenarios

- Successive iterations: classifiers adaptation
  - Insertion of user knowledge on the classification model
  - Convergence to desired results

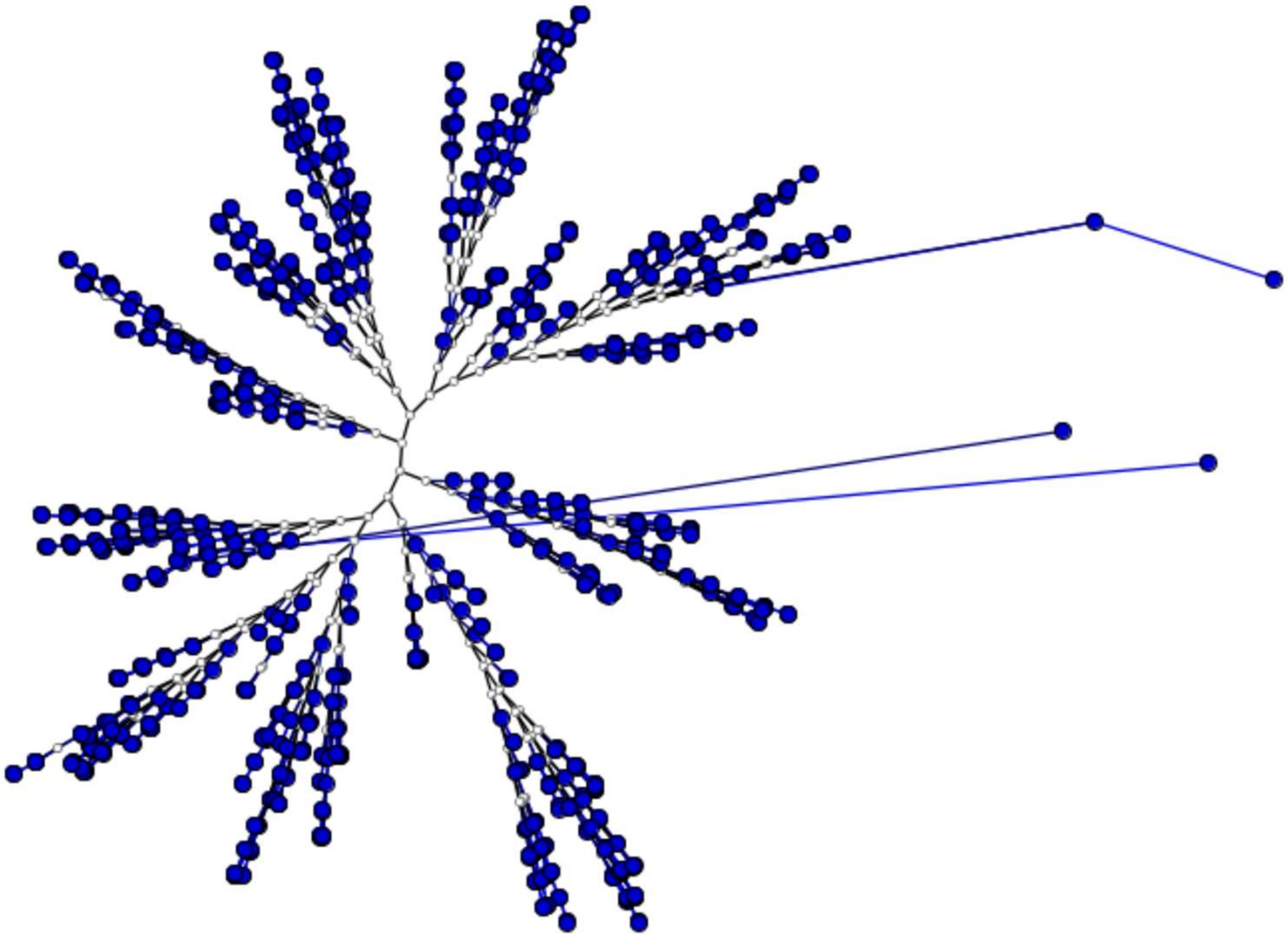
Misclassified Instances Relabeling



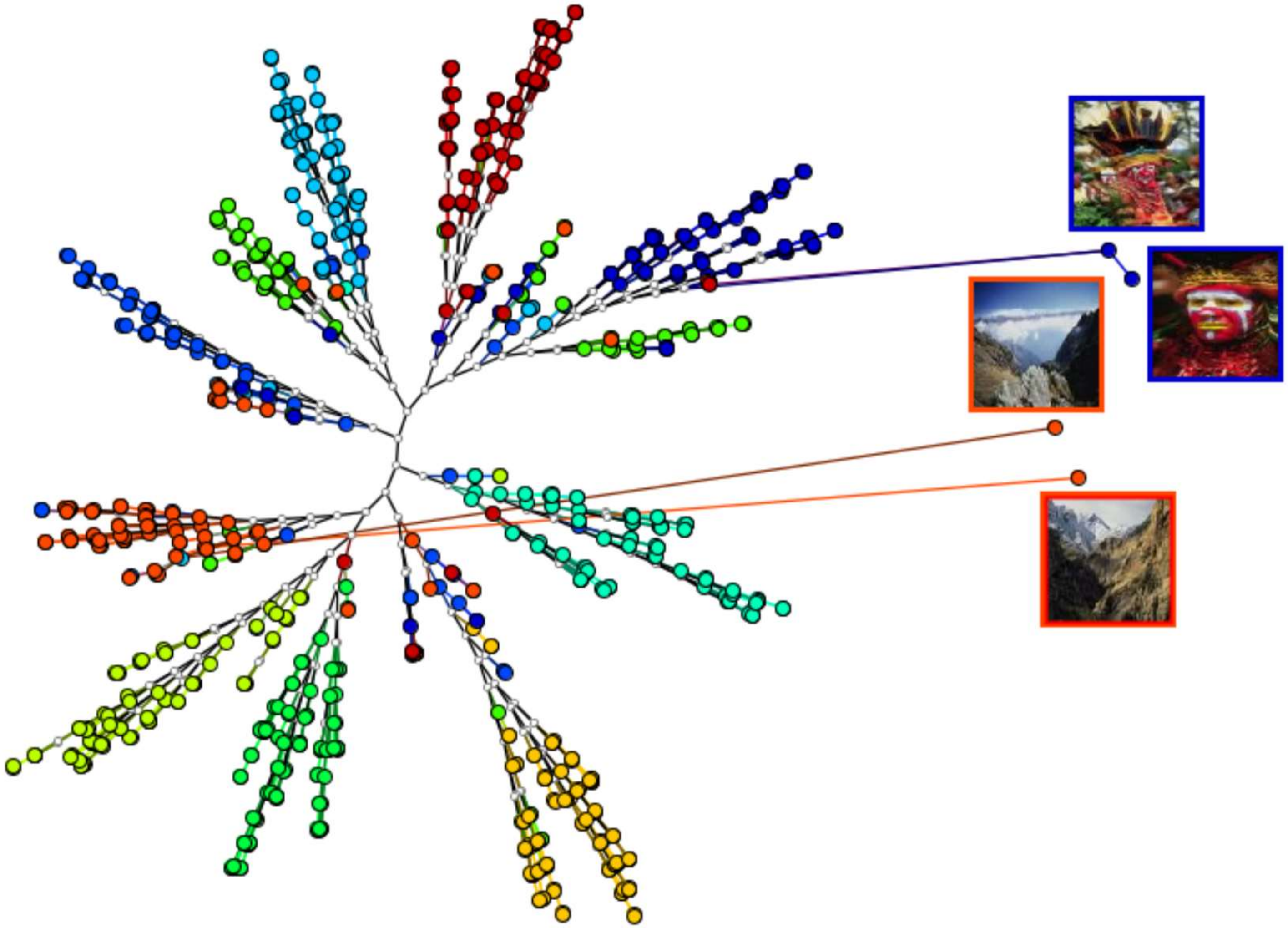
Classification Model Upgrading



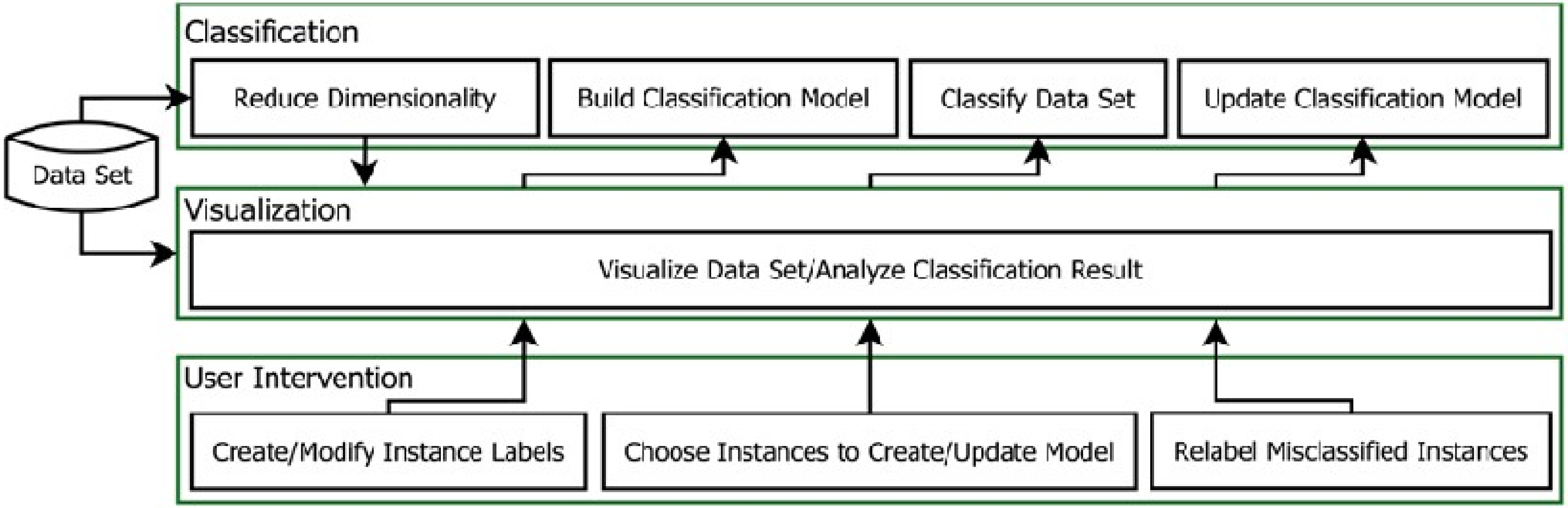
Reclassification - Upgraded Model



Reclassification - Upgraded Model

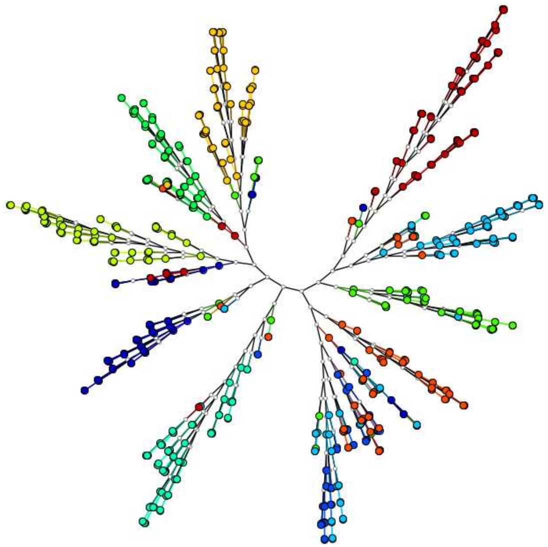


Visual Classification Methodology



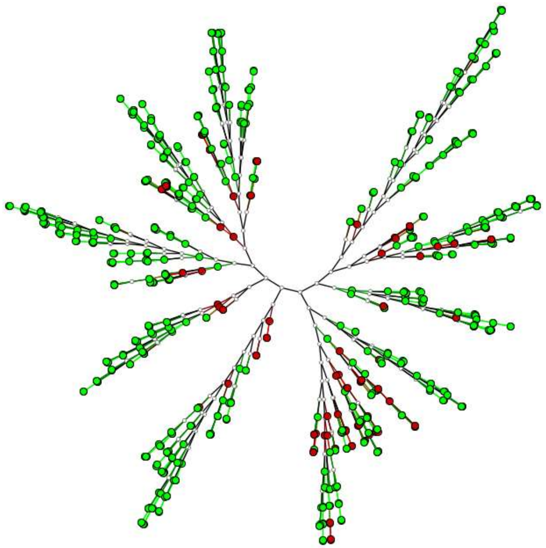
Classifier Validation with Ground Truth

- Analysis of the classification results



Classifier Validation with Ground Truth

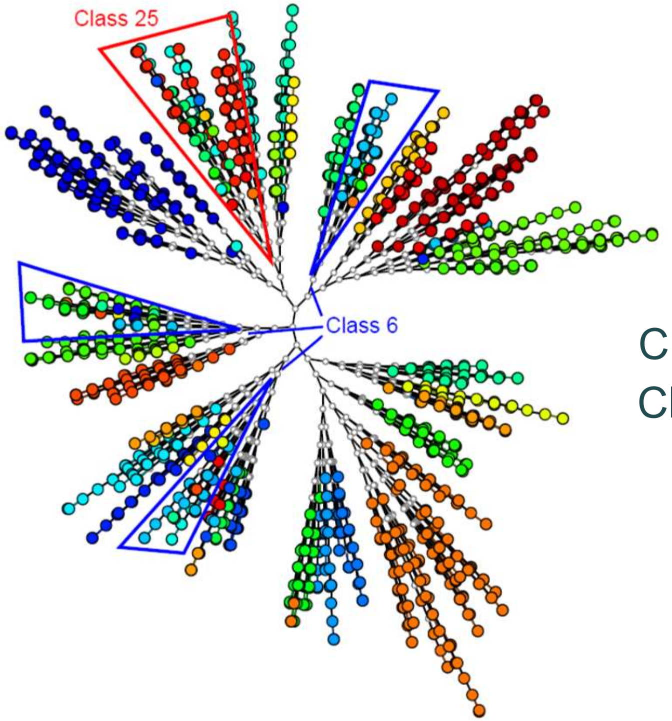
- Analysis of the classification results



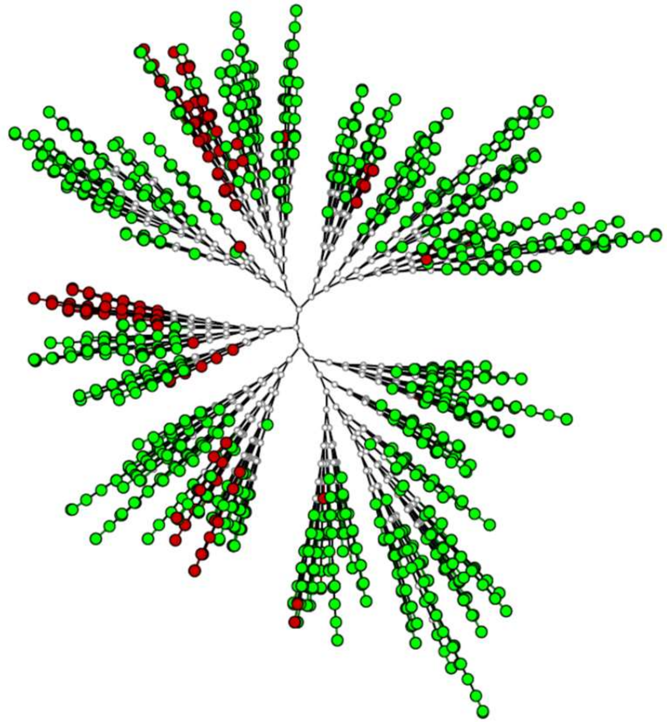
Matching instances :	598 - 85,4%
Non-matching instances :	102 - 14,6%
Non-corresponding instances :	0 - 0%
Accuracy :	97,16%
Precision :	86,55%
Recall :	85,43%
F1 :	0,86

		Classification										
		0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	FNR
Ground Truth	0.0	57	1	6	0	0	2	0	0	0	4	18,57%
	1.0	2	35	21	2	0	3	1	0	6	0	50,00%
	2.0	0	4	63	0	0	0	0	0	2	1	10,00%
	3.0	0	0	1	69	0	0	0	0	0	0	1,43%
	4.0	1	0	0	0	62	0	0	2	3	2	11,43%
	5.0	1	1	9	0	0	52	0	5	0	2	25,71%
	6.0	0	0	0	0	0	0	70	0	0	0	0,00%
	7.0	1	0	2	0	0	0	0	66	0	1	5,71%
	8.0	2	5	3	0	0	4	0	0	56	0	20,00%
	9.0	0	0	1	0	0	0	0	0	1	68	2,86%
FPR		1,10%	1,72%	6,39%	0,32%	0,00%	1,41%	0,16%	1,10%	1,87%	1,56%	

Classifier Validation with Ground Truth



Class 6: 109 miss
Class 25: 65 miss

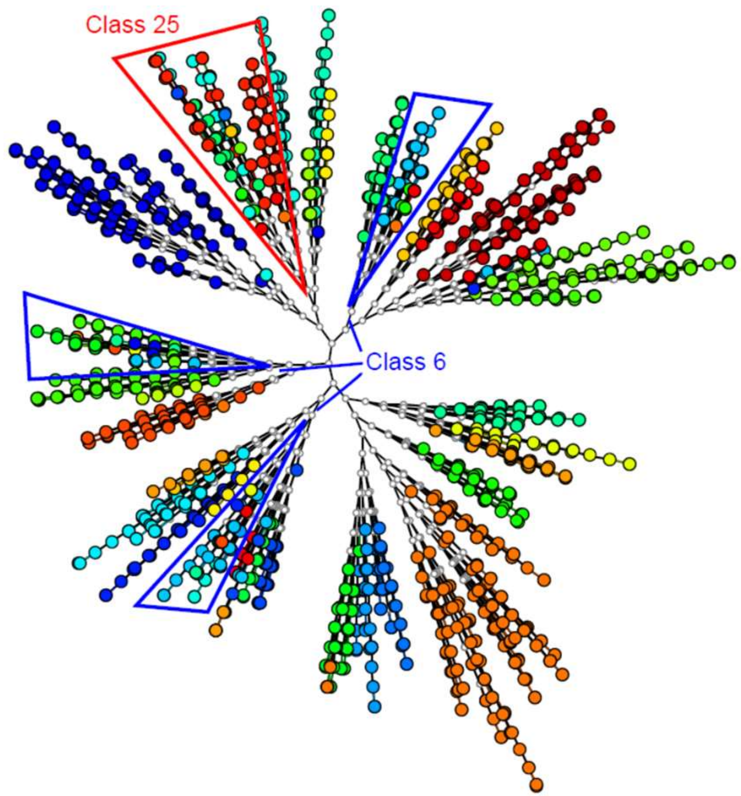


Classmatch tree

Classifier Validation with Ground Truth

- Layout: clues
  - Data set structure
  - Classifier behaviour

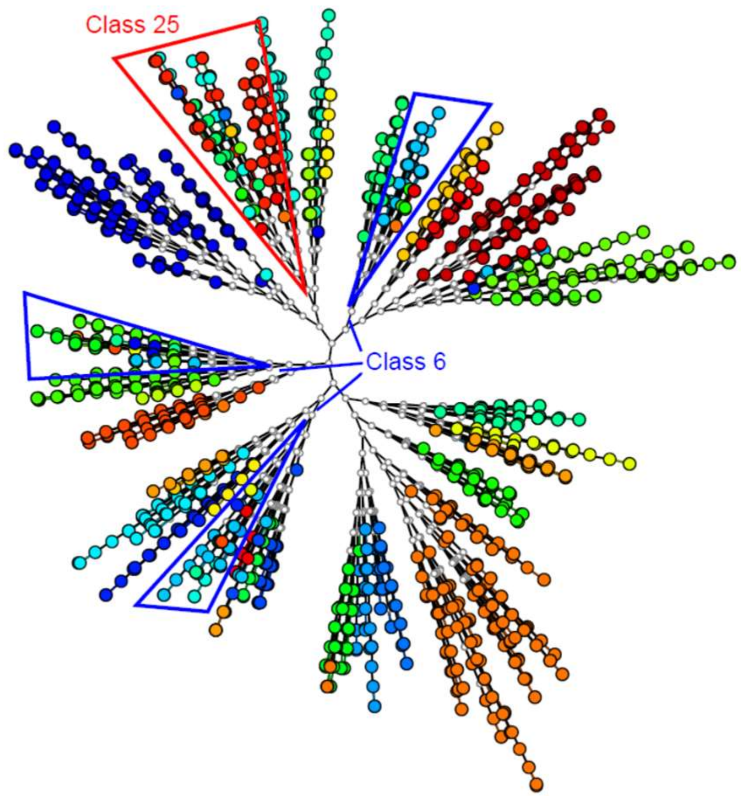
- Classes 6 and 25: heterogeneous branches
  - Classification is mixing instances in these branches



Classifier Validation with Ground Truth

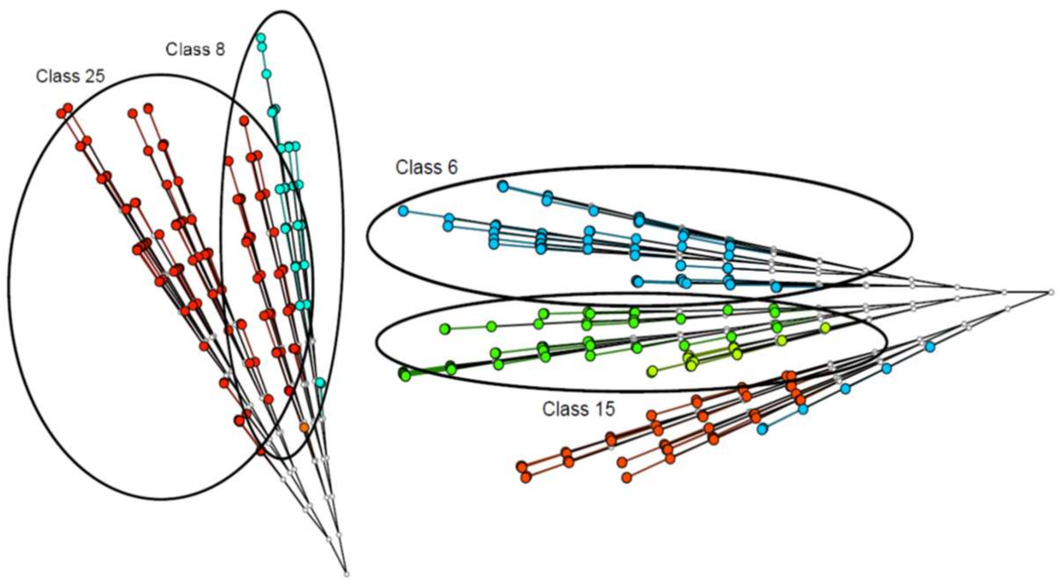
- Layout: clues
  - Data set structure
  - Classifier behaviour

- Classes 6: several branches
  - Class covers a wide range of features
  - Naturally heterogeneous: could be divided into more homogeneous subclasses



Classifier Validation with Ground Truth

- Layout: clues
  - Data set structure
  - Classifier behaviour



- Neighborhood of classes 6 and 25
- Confusion Matrix
  - Class 25 instances → class 8
  - Class 6 instances → class 15

Classifier Validation with Ground Truth

- Several upgrade strategies
  - Verification of branches with high misclassification rates
  - Classmatch tree: 23 representative misclassified instances
    - Class 25: 8 instances
    - Class 6: 15 intances

class 6

class 25

Classifier Validation with Ground Truth

	Initial Model	Upgraded Model
ETHZ-Reduced		
Matching Instances	1704 (88.1%)	1808 (93.4%)
Non-matching Instances	231 (11.9%)	127 (6.6%)
Accuracy	98.47%	99.14%
Precision	89.05%	94.06%
Recall	88.06%	93.44%
ALL-Reduced		
Matching Instances	1875 (67.7%)	1991 (71.9%)
Non-matching Instances	894 (32.3%)	778 (28.1%)
Accuracy	86.61%	88.45%
Precision	71.98%	73.79%
Recall	67.71%	71.90%

Collection Evolution - New Classes

- Dynamic scenarios

Data Set	Content	Classes	Items	Attributes
NEWS2011	News	7	1018	3731

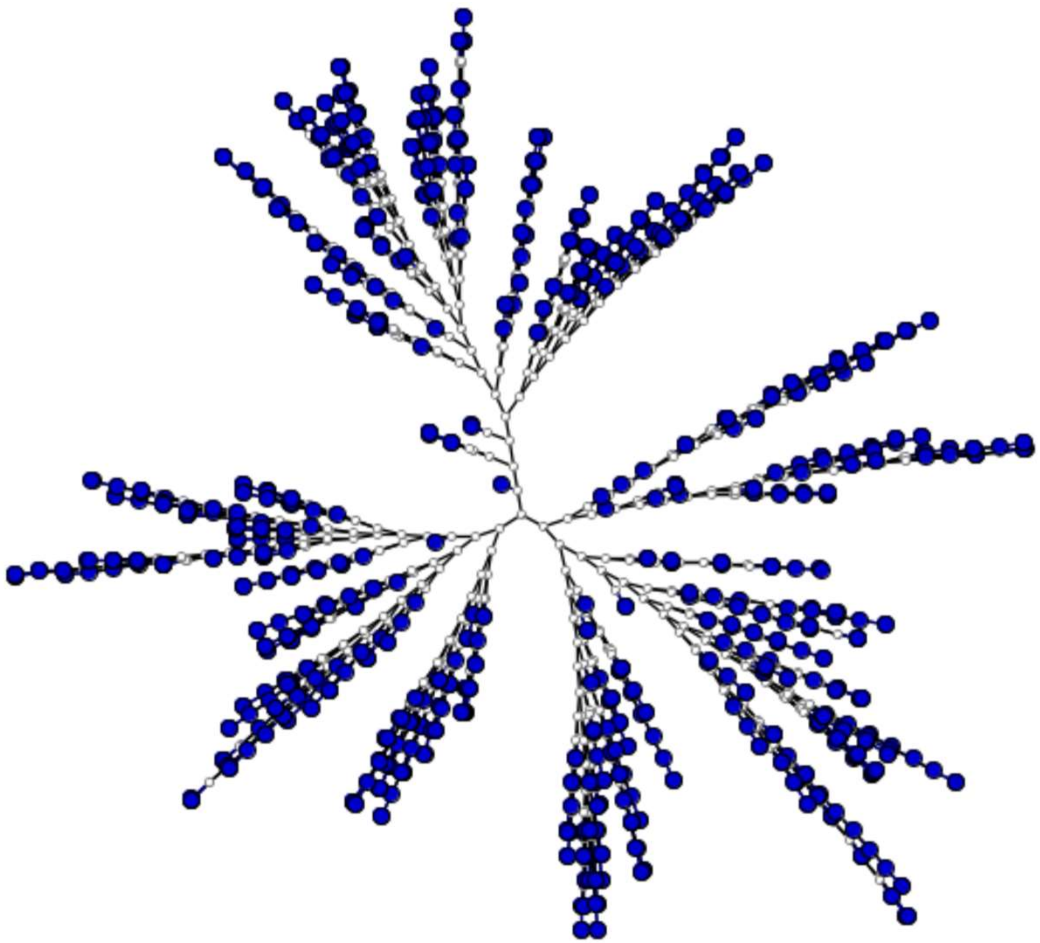
Instances

Training Set	Test Set
42	976

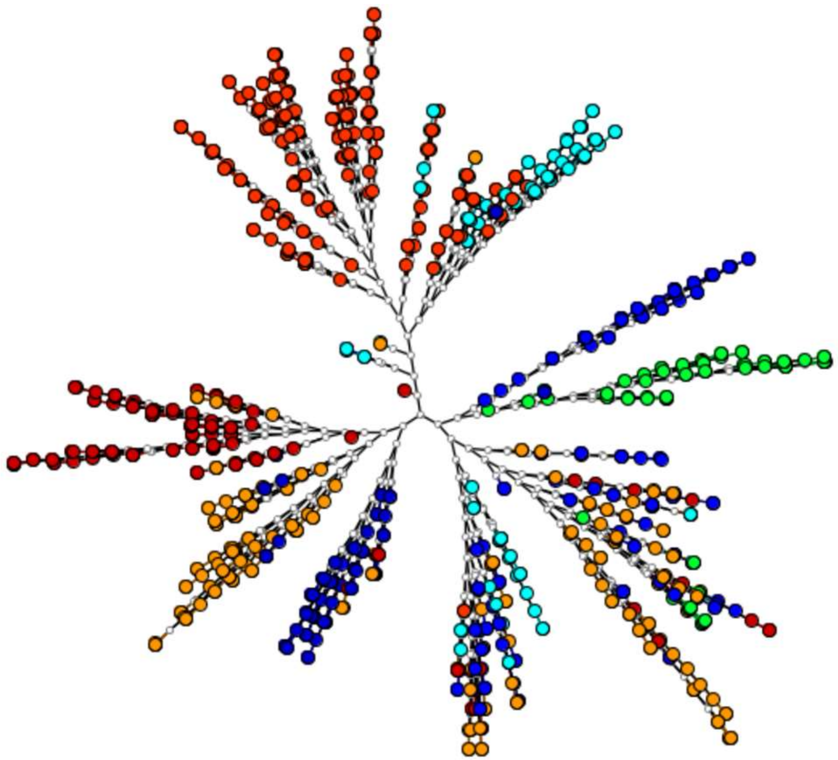
Collection Evolution - New Classes

- Classification model: 7 classes
  - Greek Financial Crisis
  - Ratko Mladic judgment
  - Syria Conflicts
  - USA Crisis
  - Yemen Attacks
  - Afghanistan
  - AmyWinehouse Death

Collection Evolution - New Classes

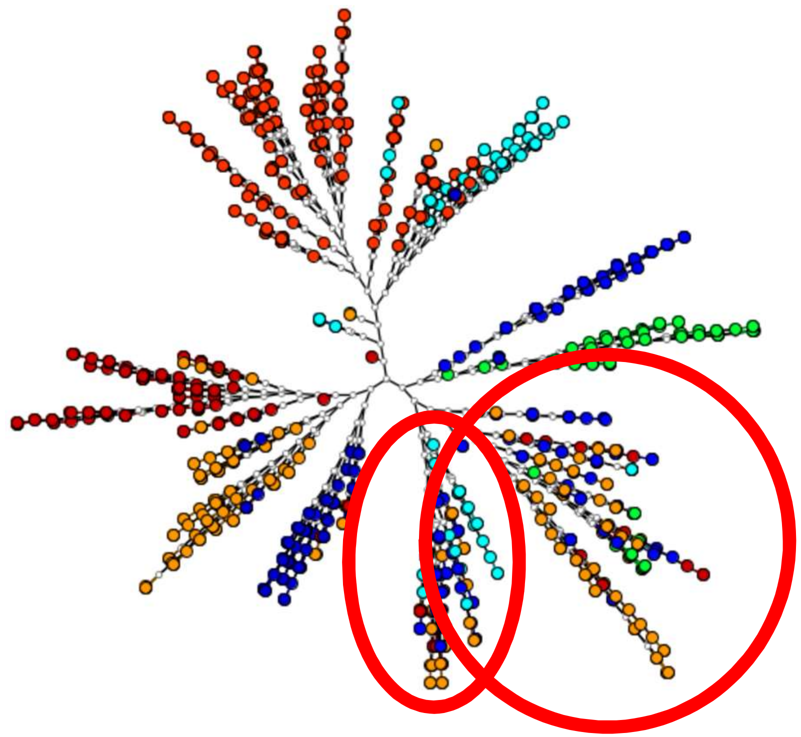


Collection Evolution - New Classes



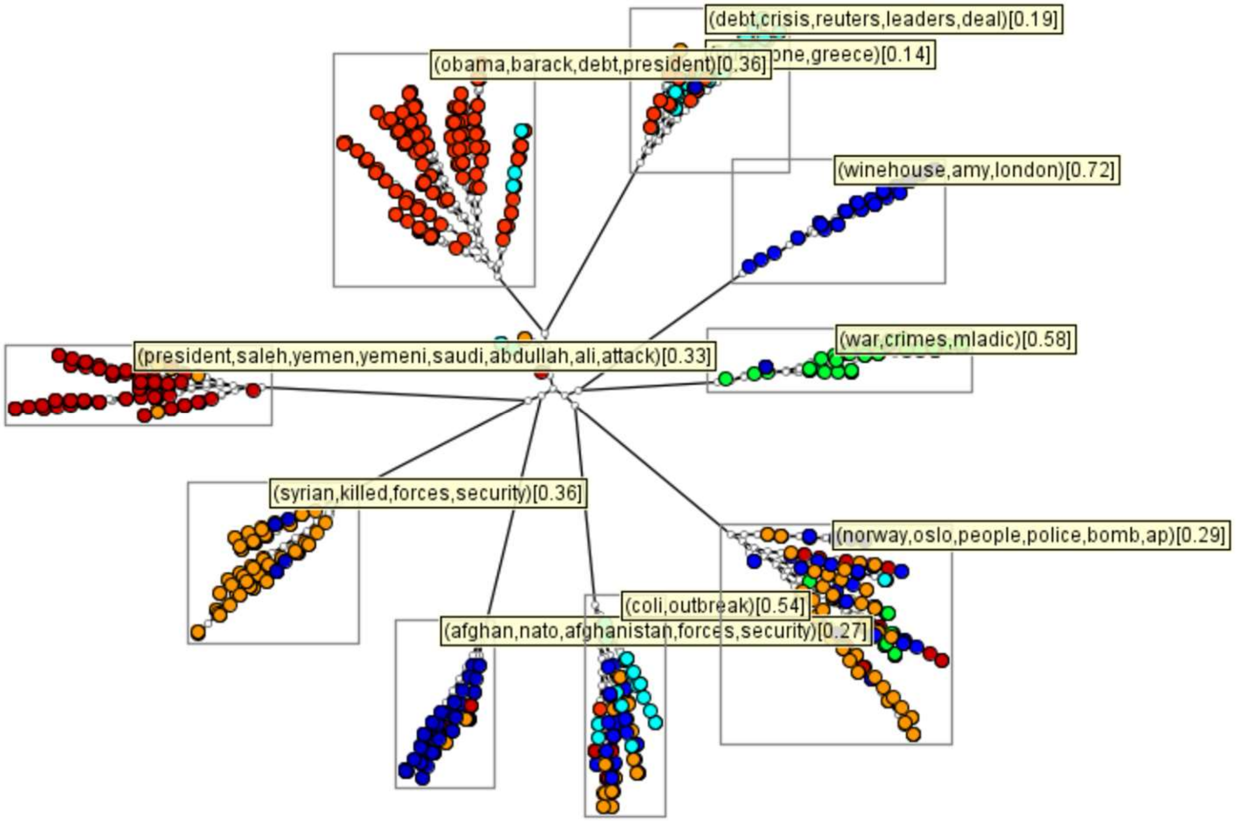
Collection Evolution - New Classes

- Two heterogeneous branches
  - Unknown patterns



Collection Evolution - New Classes

- Main topics per branch
- New topics
  - Norway bomb
  - Escherichia coli outbreak

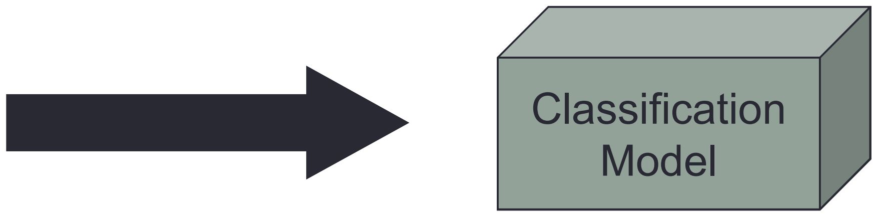


Collection Evolution - New Classes

- Instances of new topics are used to update the classification model

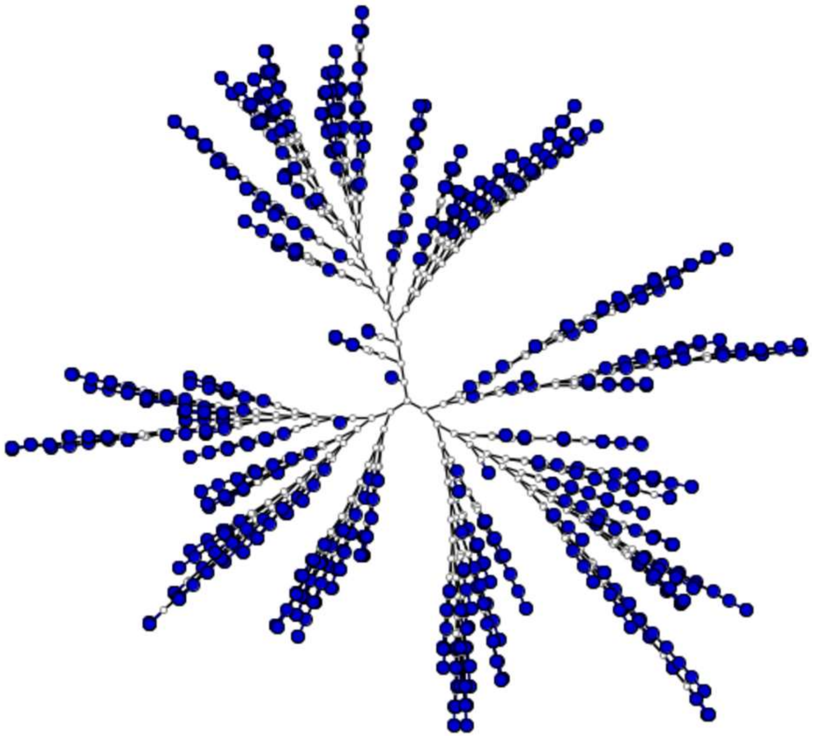
Norway bomb

Escherichia coli outbreak



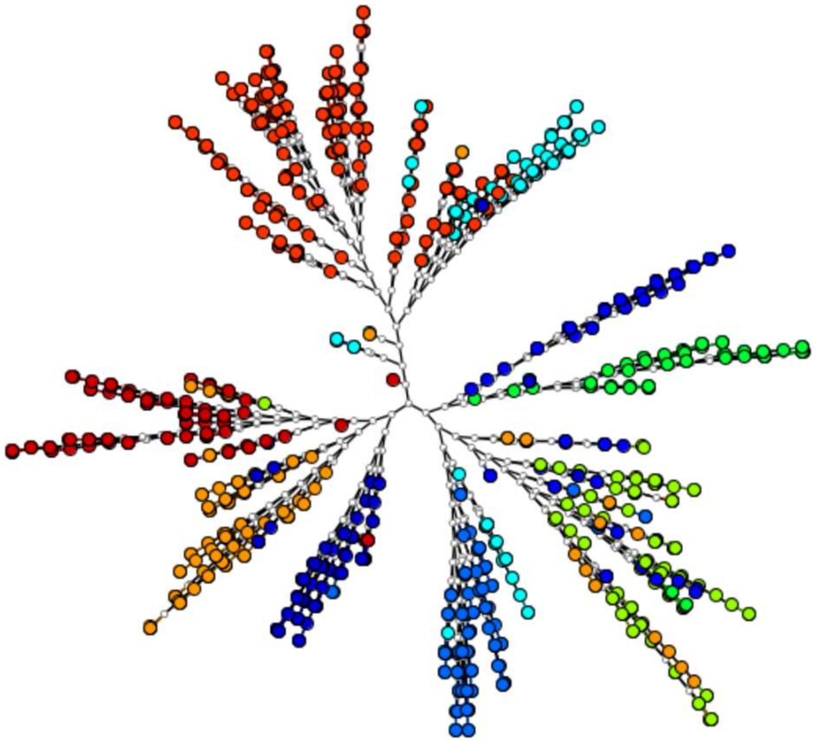
Collection Evolution - New Classes

- Reclassification



Collection Evolution - New Classes

- Reclassification



Point placement Techniques Capabilities

- Projection techniques
  - Group shapes are mantained
  - Outliers are more evident
  - Clutter is higher
- NJ Trees
  - Capable of describing the organization of the similarity
  - Similarity is consistent at the branches' ends
  - Does not require any parameter

Final Remarks

- *Processing scalability*
- *Data access scalability*
- *Visual scalability*
- *Partition Strategies*
- *Sampling Strategies*
- *Data Structures related to Analysis processes and tasks*
- *Integration with Mining, Learning, Data Analysis in General*
- *Applications*

Partners & collaborators

- *In respect to own work included in this presentation we acknowledge the cooperation between ICMC – USP and:*

- *IC – UNICAMP, Campinas – SP, Brazil*
- *LNCC – CNPEM, Campinas – SP, Brazil*
- *Embrapa Campinas – SP, Brazil*
- *Dalhousie University – Halifax, Canada*
- *Jacobs University Bremen – Germany*
- *FMRP – USP*
- *Poli – USP*
- Department of Biology, UFSCar

Funding

- CAPES / DAAD PROBRAL
- CAPES / NUFFIC
- CNPQ personal grants / student grants
- CNPq Universal 2012-2013,2013-2016
- FAPESP – student and postdoc grants
- FAPESP Thematic Research project
- Thanks!!!!