Multidimensional Projections and Similarity Trees

Rosane Minghim
rminghim@icmc.usp.br

2008-2017

Part I
Principles, Techniques and Applications

- **Techniques and applications**
- **Visual strategies to support data analysis/mining tasks**
- **Problems regarding scale of data sets**

Outline

- About...
- Data Science
- Big Data for productivity
- Visualization
- Visual Mining
- Applications

Problem

- People trying to make sense of data

´messy´ data

Data is...

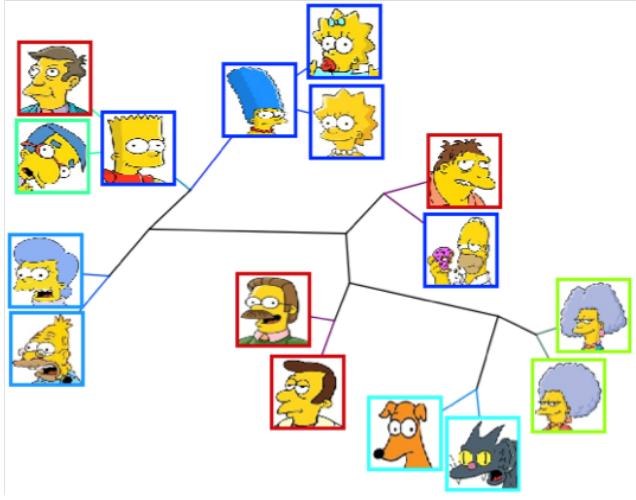
- Far too complex... (many dimensions, many types)
- Far too big... (´easy´ to collect)
- Far too varied... (images, videos, documents, news, networks)
- Never ending... (data streams)
- Much redundancy...
- Many relationships...
- Pieces missing...

- Studying natural & artificial systems and phenomena implies in handling lots of data...

What does your data tell???

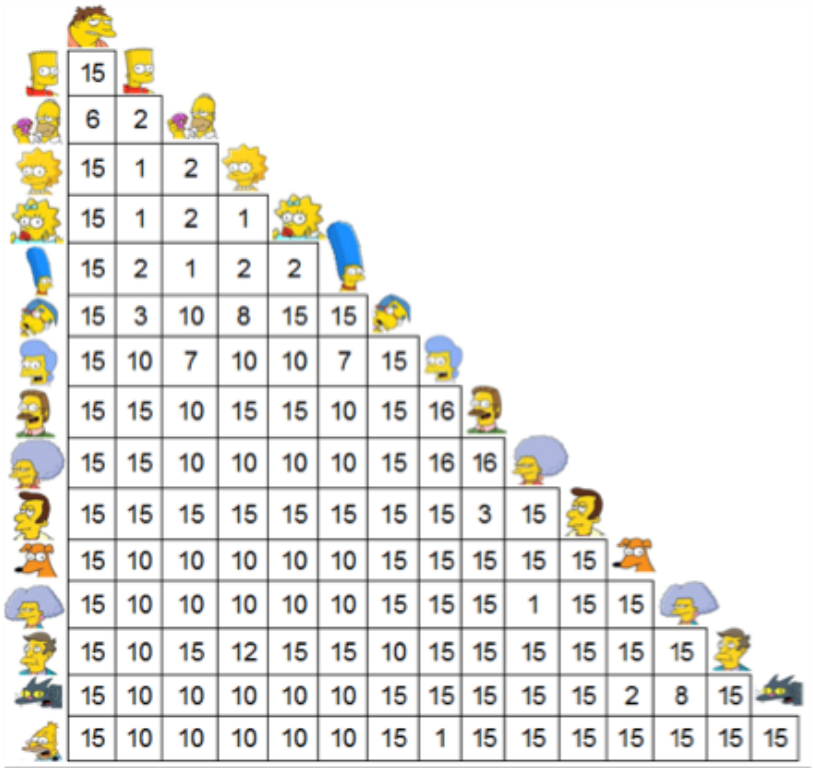
- People trying to make sense of data

´messy´ data



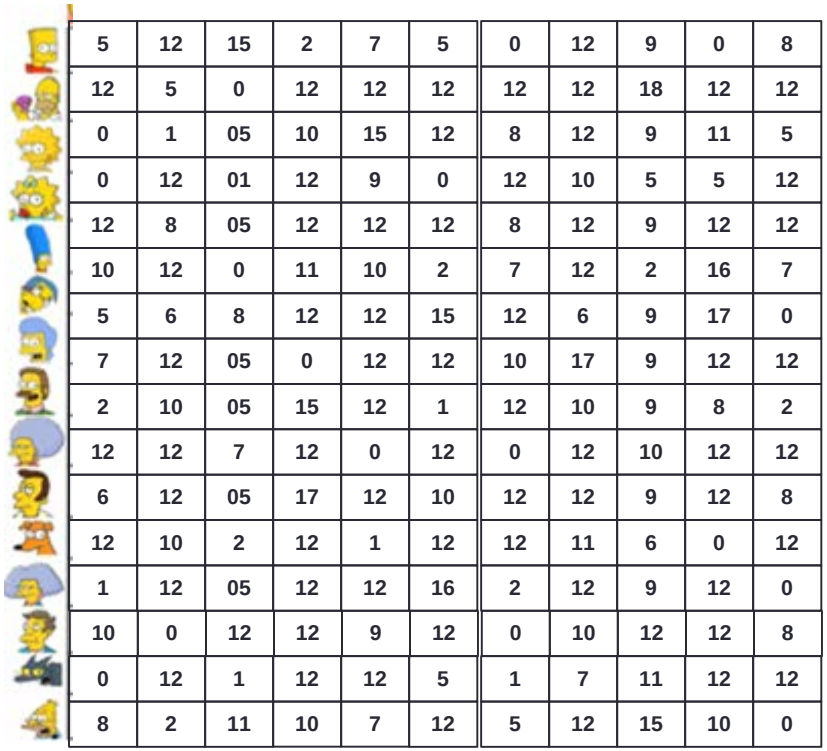
Techniques

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



pairwise distances

and/or

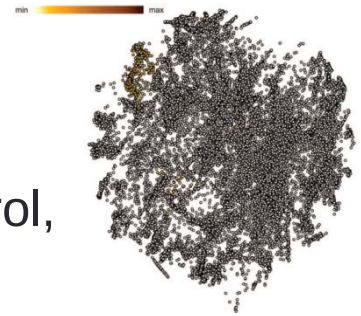


5	12	15	2	7	5	0	12	9	0	8
12	5	0	12	12	12	12	12	18	12	12
0	1	05	10	15	12	8	12	9	11	5
0	12	01	12	9	0	12	10	5	5	12
12	8	05	12	12	12	8	12	9	12	12
10	12	0	11	10	2	7	12	2	16	7
5	6	8	12	12	15	12	6	9	17	0
7	12	05	0	12	12	10	17	9	12	12
2	10	05	15	12	1	12	10	9	8	2
12	12	7	12	0	12	0	12	10	12	12
6	12	05	17	12	10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

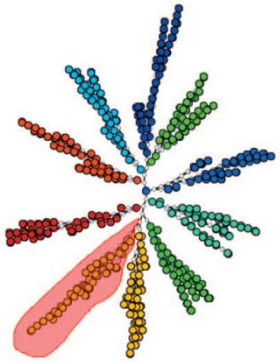
Techniques

- Projection-based
  - variations on MDS or other dimension reduction approaches
  - data mapped to low-dimensional visual space
  - preserving distances vs neighborhoods, global vs. local control, segregation



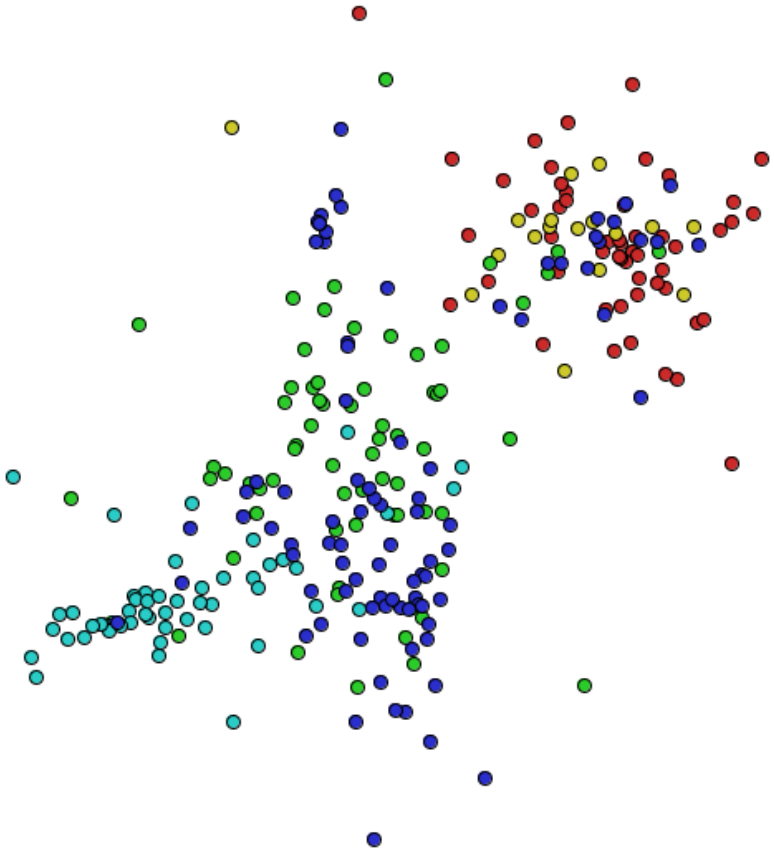
- fully interactive manipulation, dynamically adapting to user feedback
- massive data, sparse high-dimensional data, streaming data

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



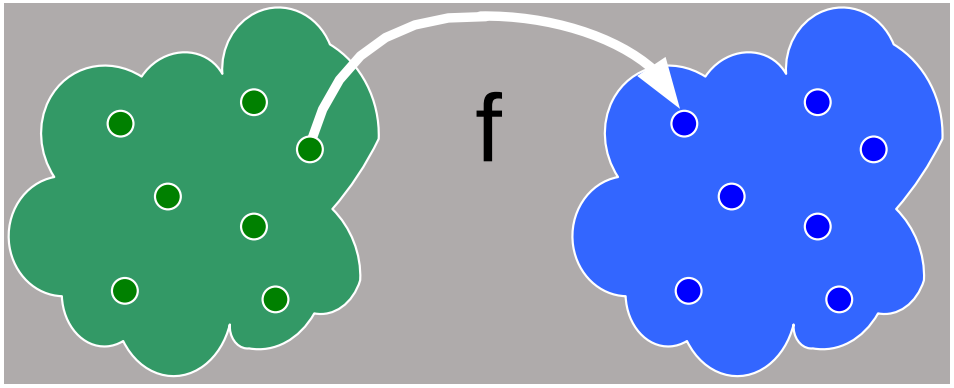
Mapping to Visual Spaces (2D-3D) allowing data exploration.

Ex: Patents surgery, drugs, molecular bio



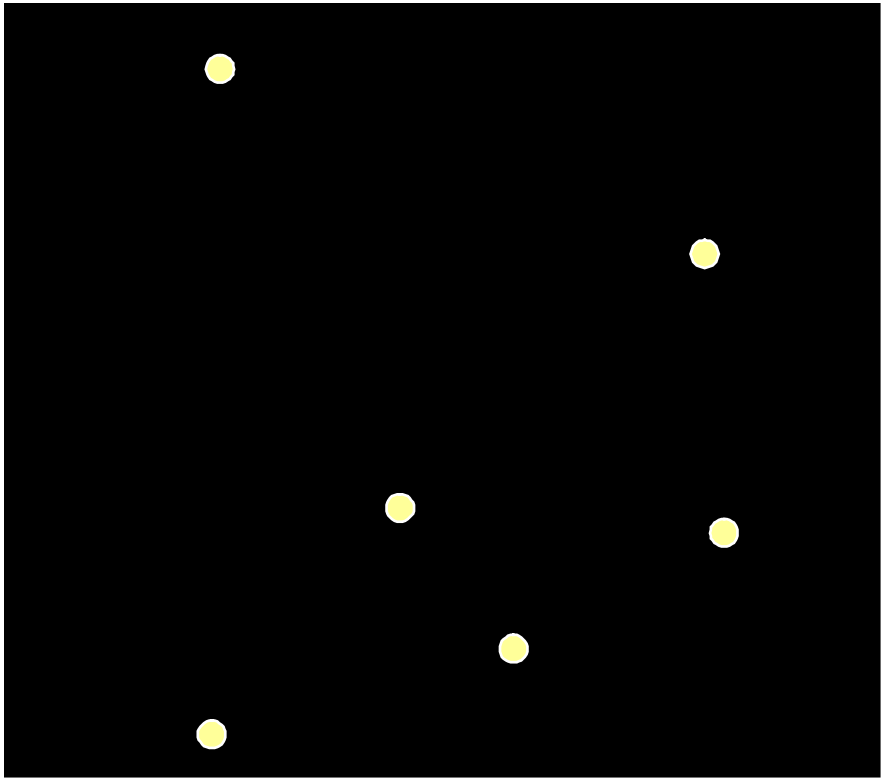
Projection Techniques

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

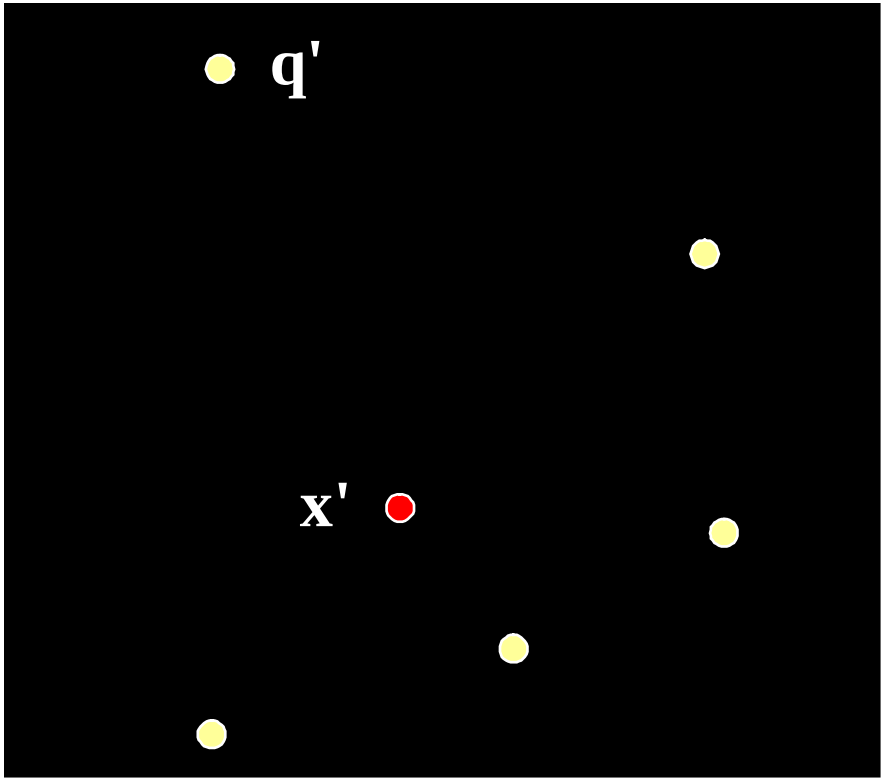


- δ: $x_i, x_j \rightarrow R$, $x_i, x_j \in X$
- d: $y_i, y_j \rightarrow R$, $y_i, y_j \in Y$
- f: $X \rightarrow Y, |\delta(x_i, x_j) - d(f(x_i), f(x_j))| \approx 0, \forall\ x_i, x_j \in X$

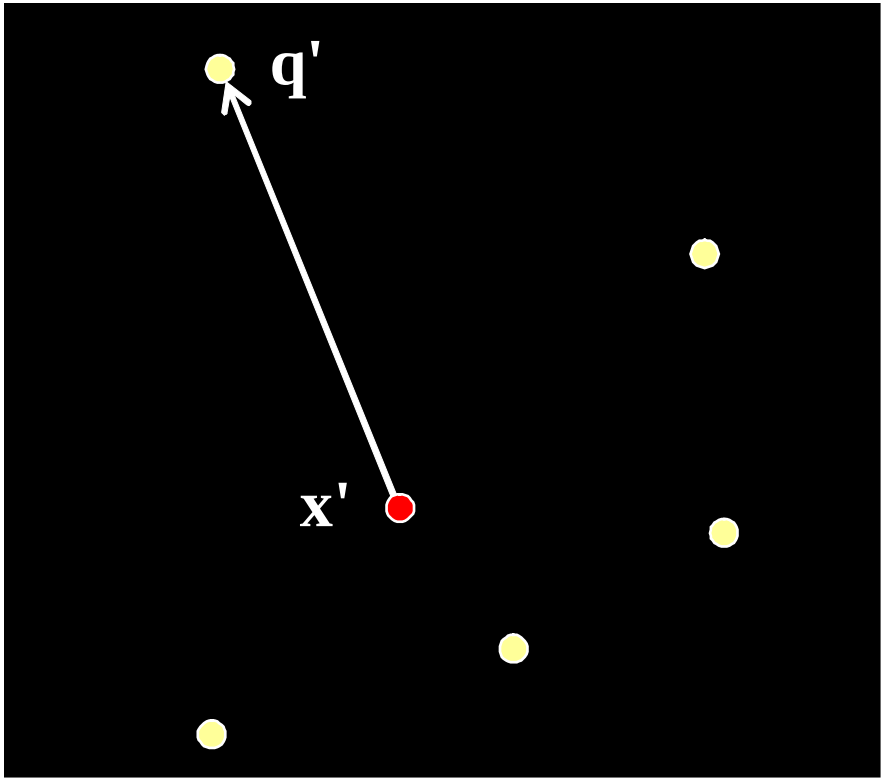
Force Based Point Placement



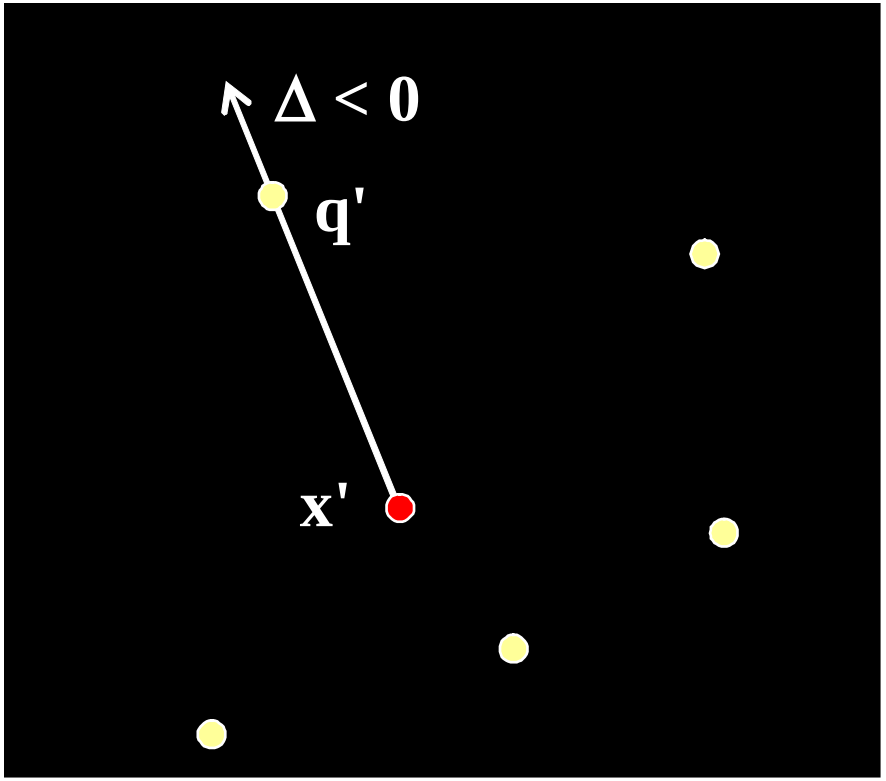
Force Scheme [Tejada et al., 2003]



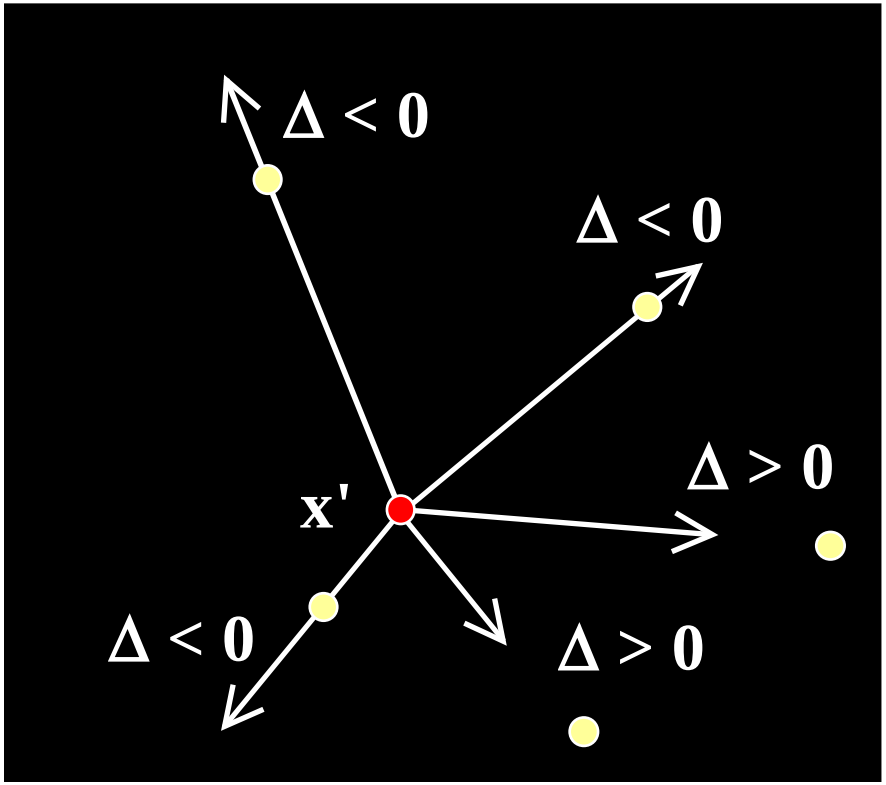
Force Scheme [Tejada et al., 2003]



Force Scheme [Tejada et al., 2003]



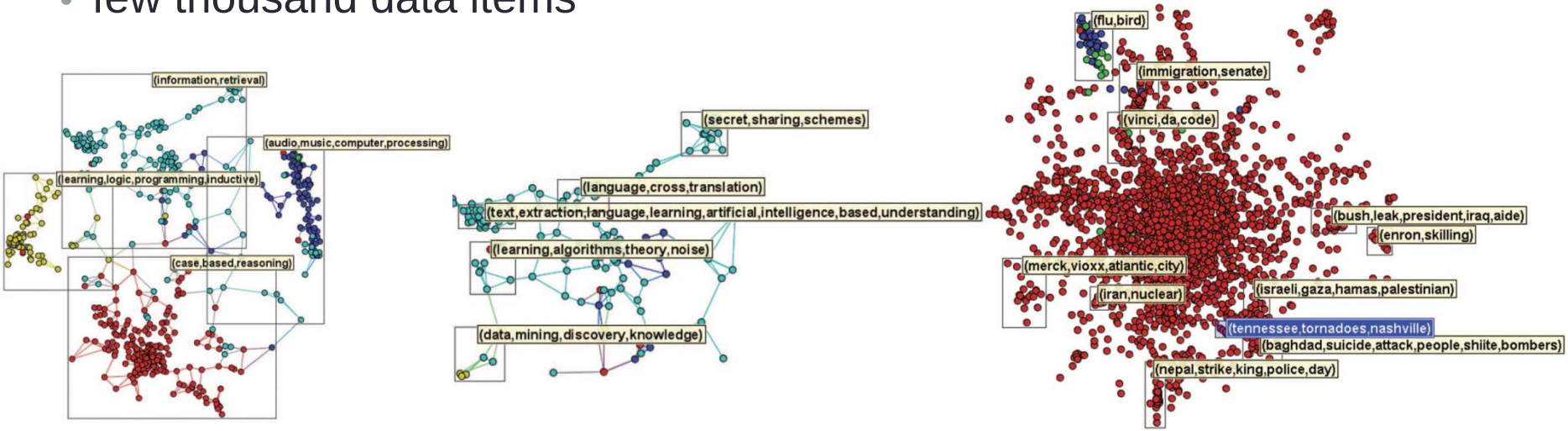
Force Scheme [Tejada et al., 2003]



LSP

- Paulovich, Nonato, Minghim, Levkowitz, Least square projection: A fast high-precision multidimensional projection technique and its application to document mapping, ***IEEE Trans. Visualization and Computer Graphics 2008***

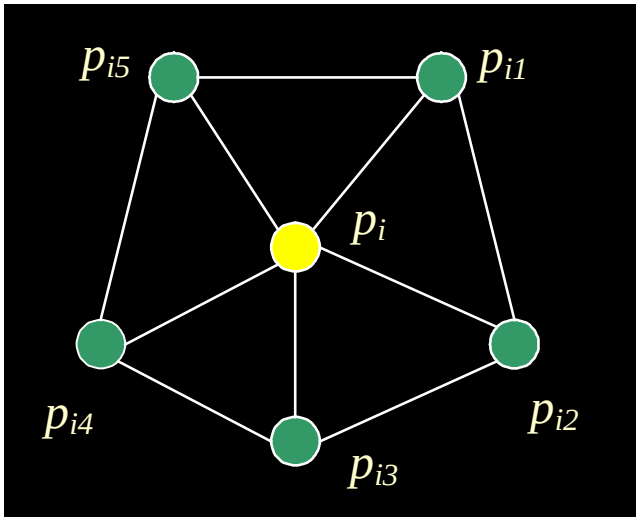
- based on identifying samples (control points) and their neighborhoods
- distance matrices & spatially embedded data
- preserves data neighborhoods
- few thousand data items



LSP: Matriz Laplaciana

- Seja $V_i = \{p_{i1},\ldots,p_{iki}\}$ a vizinhança de um ponto p_i e seja c_i as coordenadas de p_i em R^p

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



- Cada p_i será o centróide dos pontos em V_i

LSP: Matriz Laplaciana

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

onde x_1, x_2,…, x_p são vetores contendo as coordenadas cartesianas dos pontos e *L* é a matriz dada por

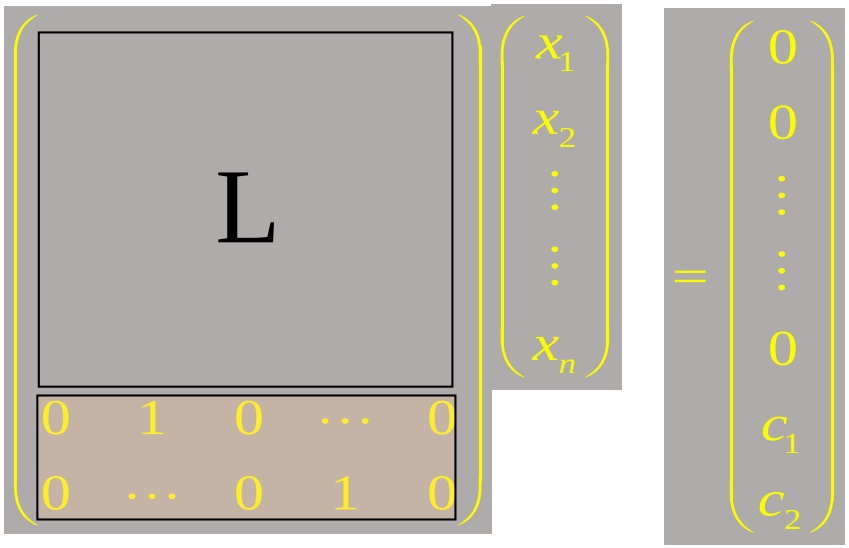
$$Lij = \begin{cases} 1 & i = j \\ -\dfrac{1}{ki} & p_j \in V_i \\ 0 & \text{caso contrário} \end{cases}$$

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

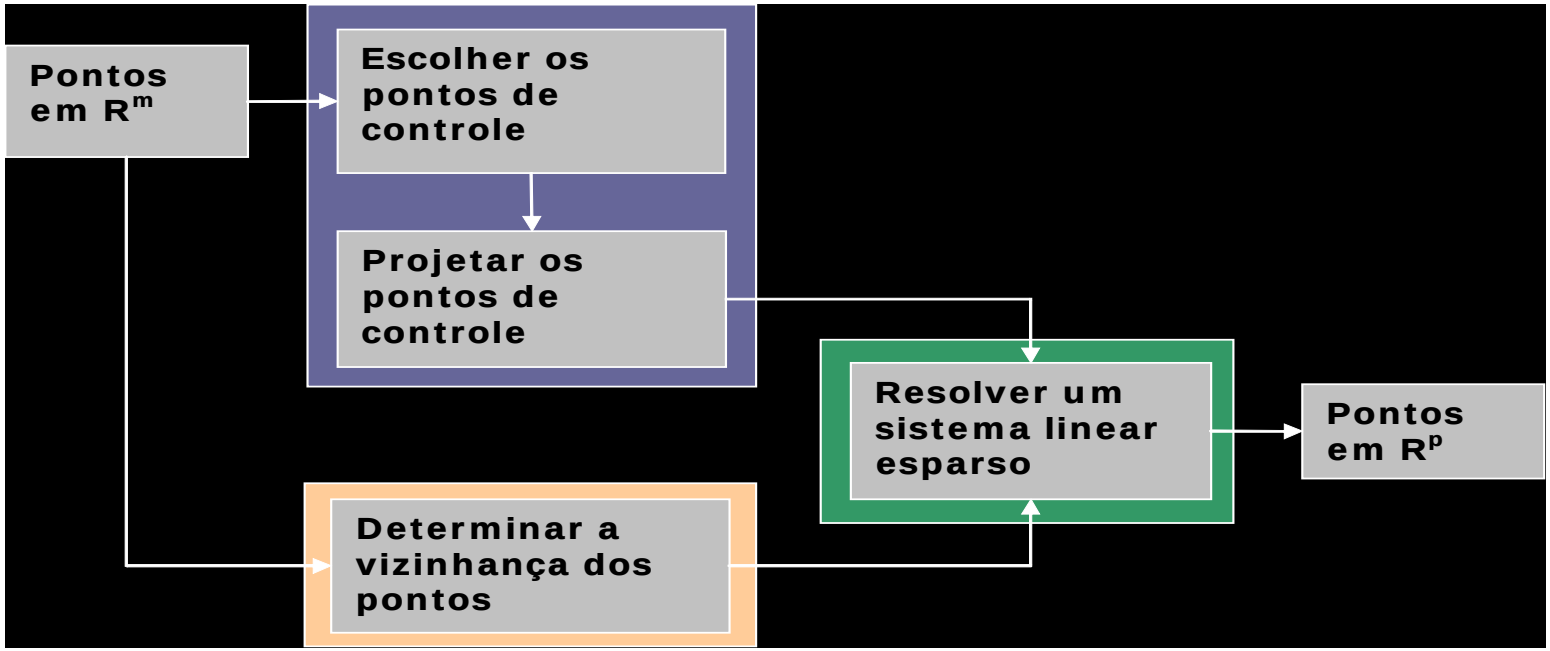
LSP: Adicionando os Pontos de Controle

$$A = \begin{pmatrix} L \\ C \end{pmatrix} \qquad Cij = \begin{cases} 1 & p_j \text{ é um ponto de controle} \\ 0 & \text{caso contrário} \end{cases}$$

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



LSP: Visão Geral

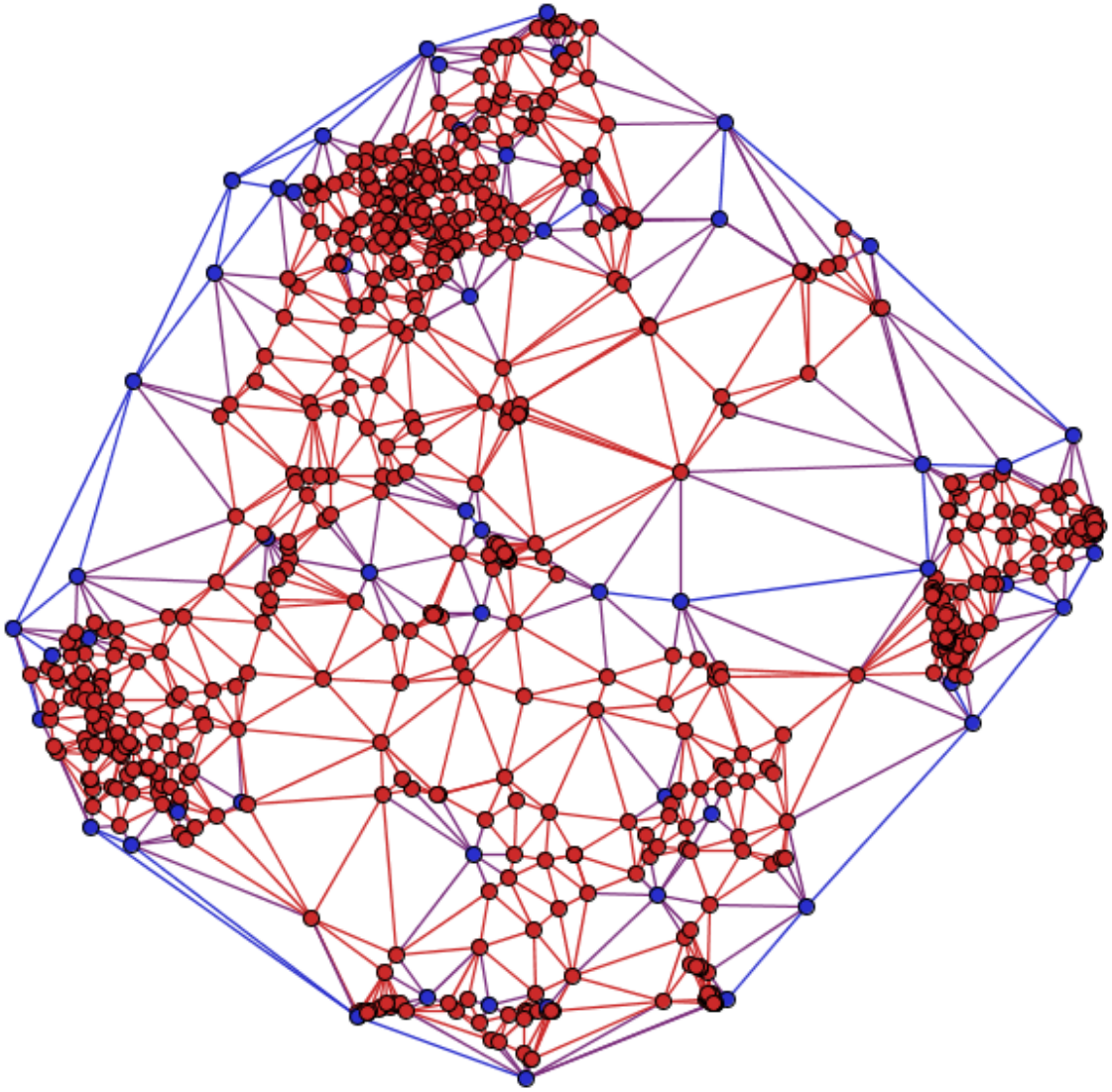


Choosing the Control Points

- In order to select the control points
  - the space R^m is split into *nc* clusters using k-medoids.
  - the control points are the medoids of each cluster

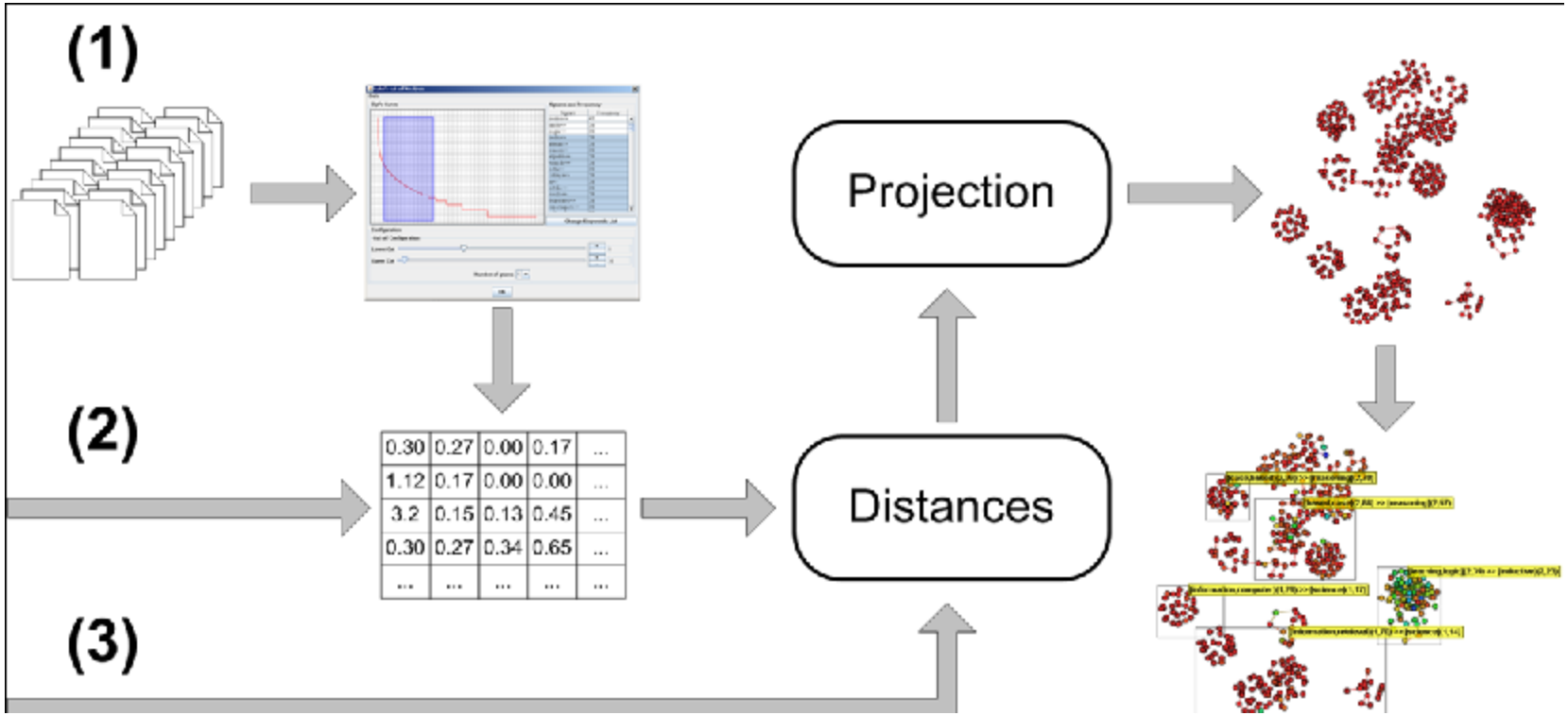
Choosing the Control Points

- Once the control points are chosen, these points are projected onto R^d through a fast dimensionality reduction method
  - Fast Projection (Fastmap or NNP)
  - Force Placement

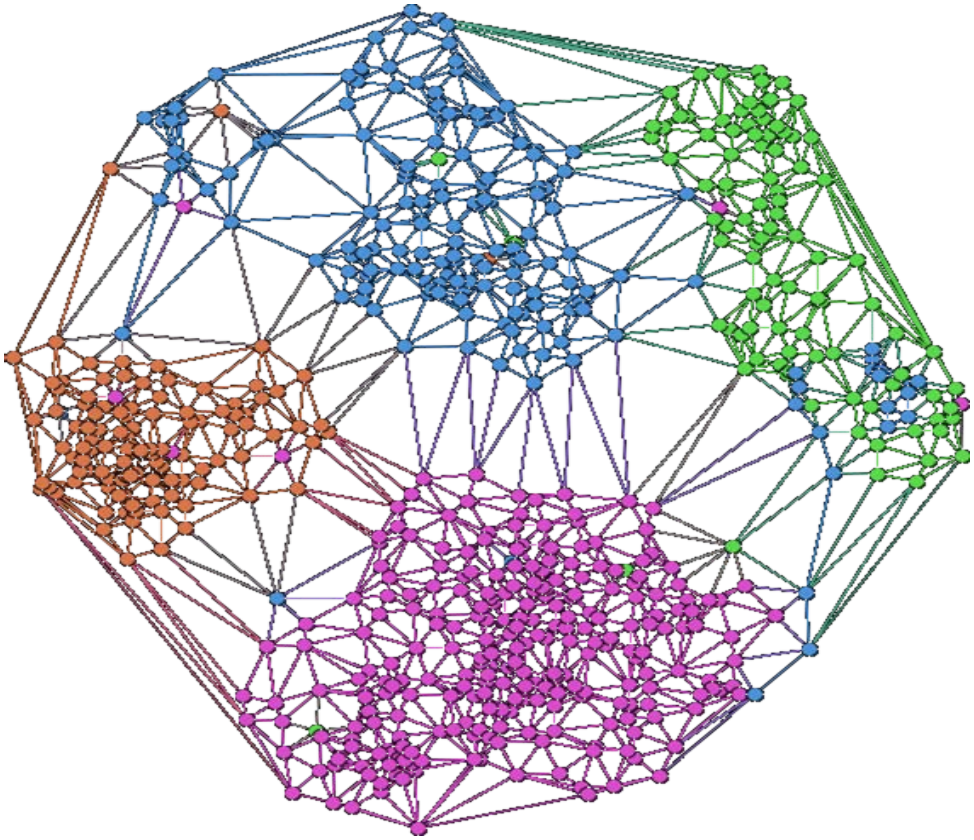


Control points in blue

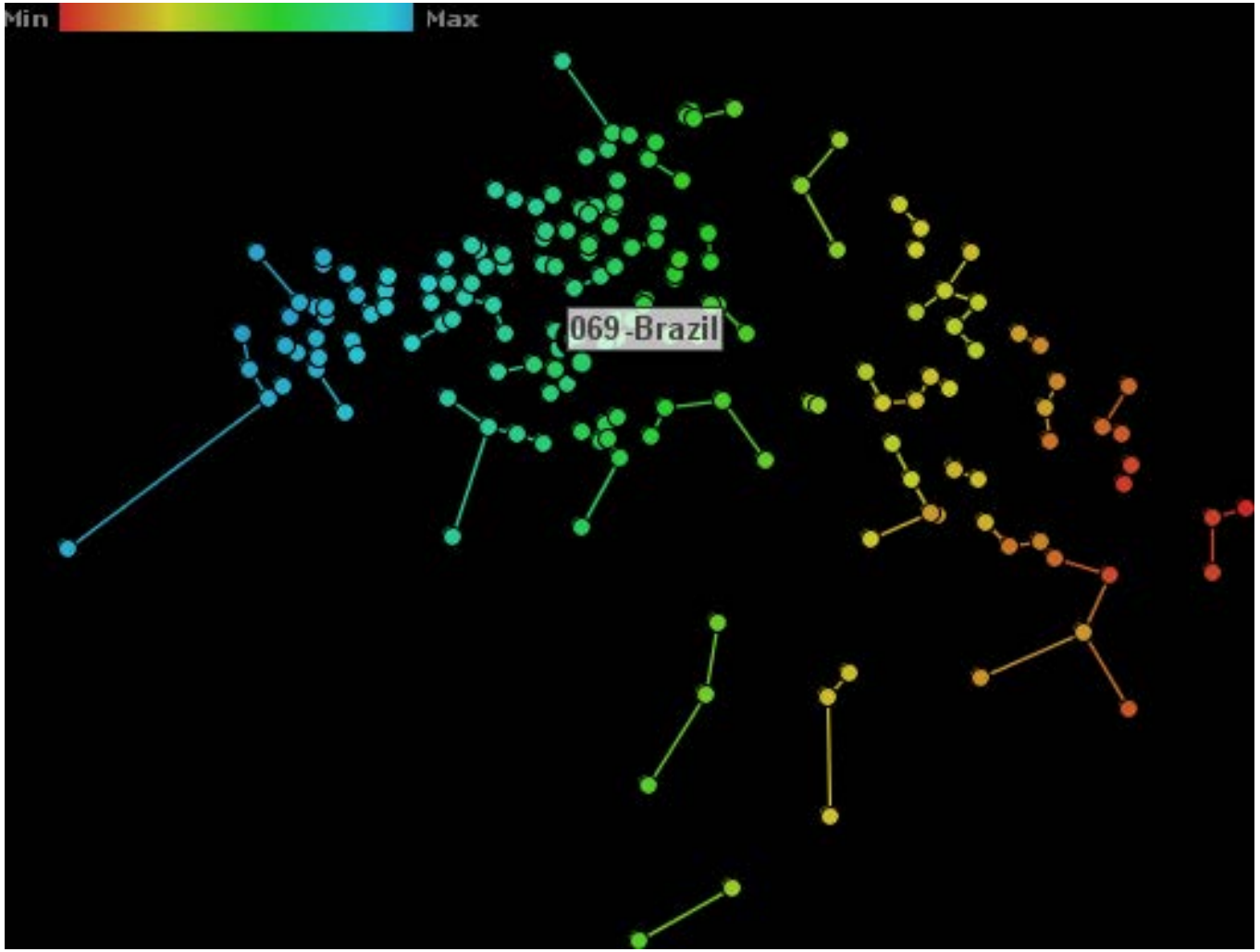
Content – based by Projections



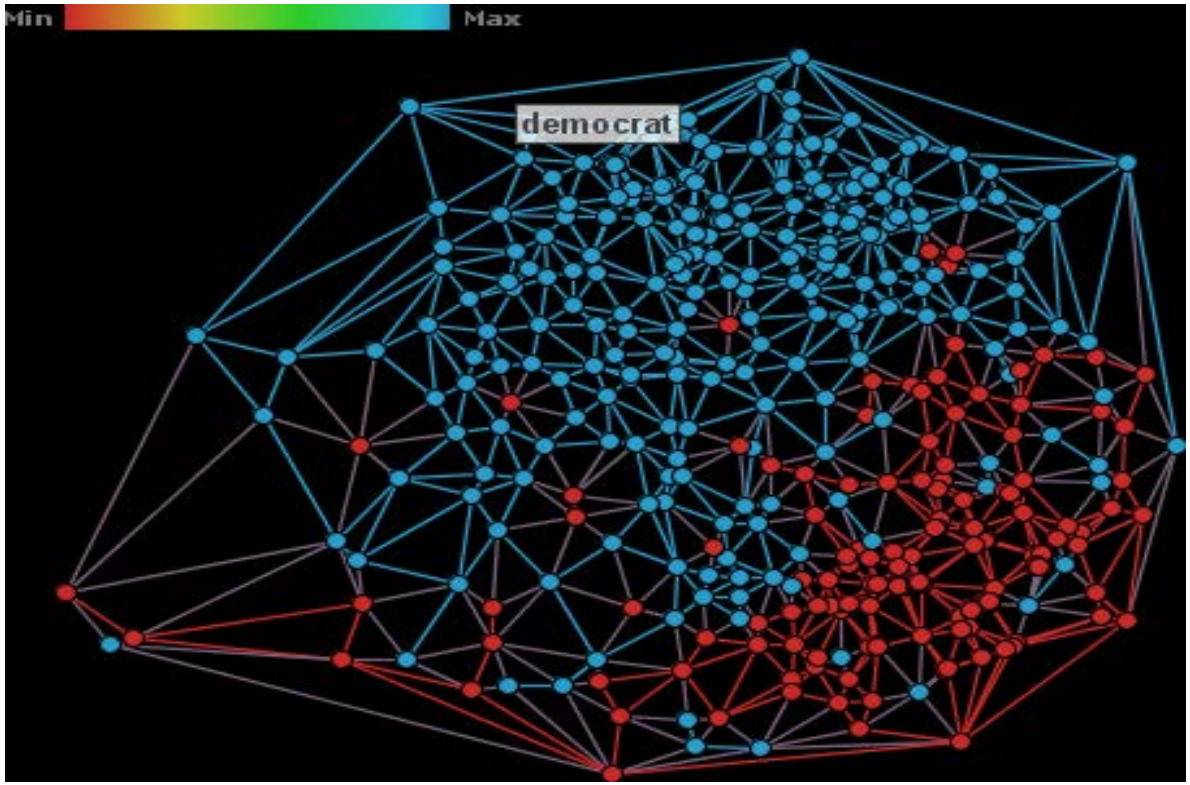
Example



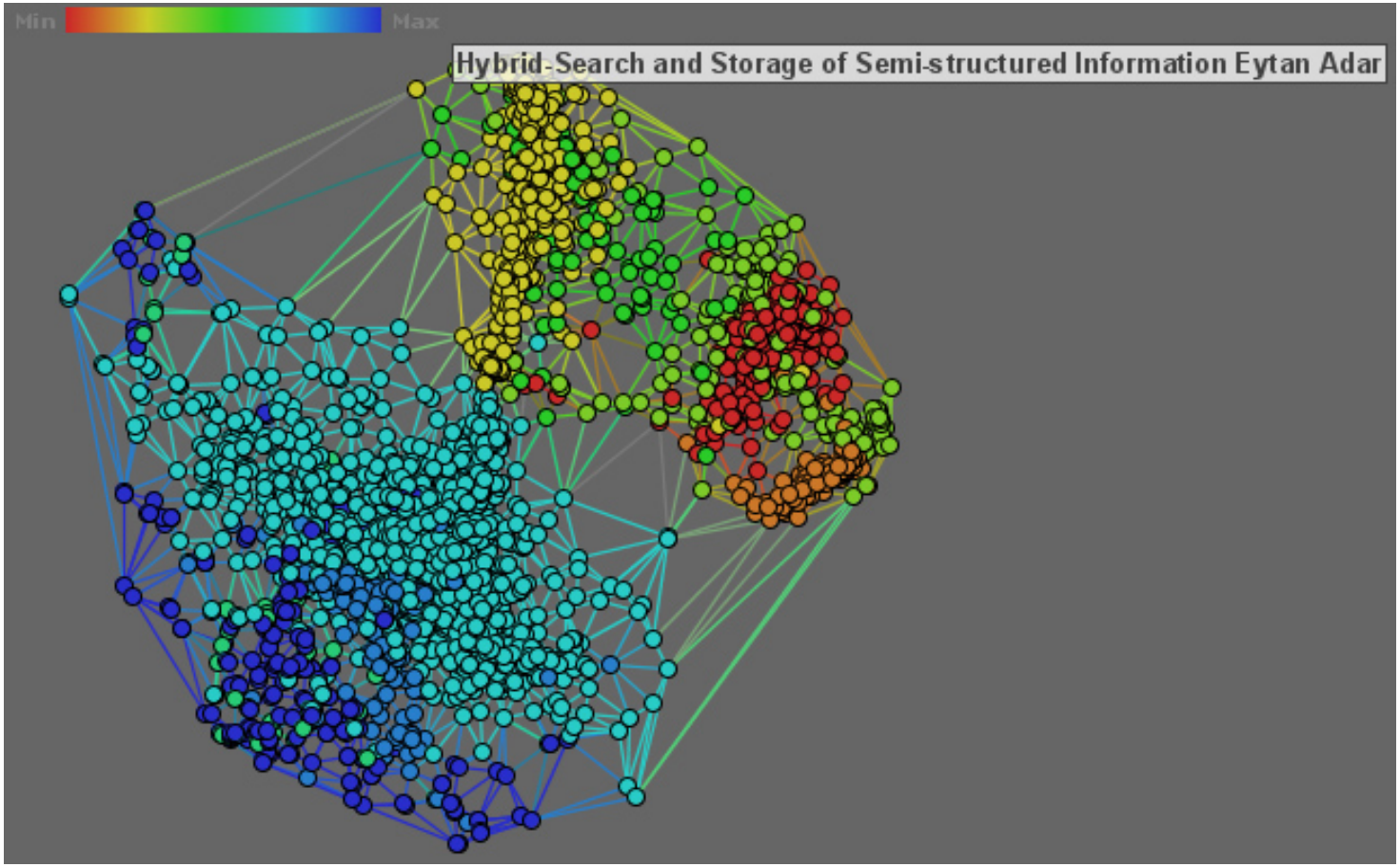
Exemplo de Projeção: IDH

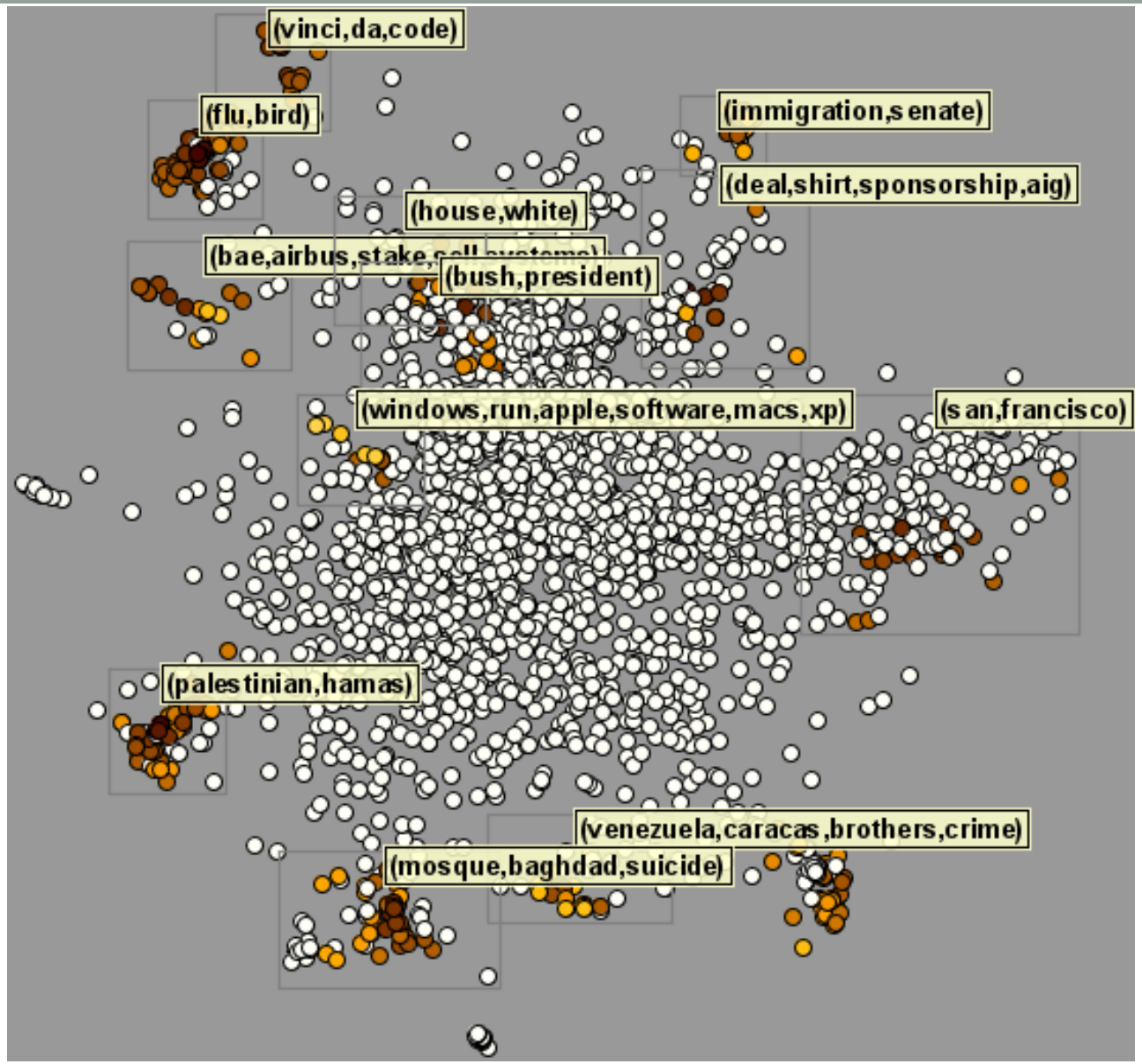


Exemplo de Projeção: Votação

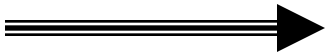


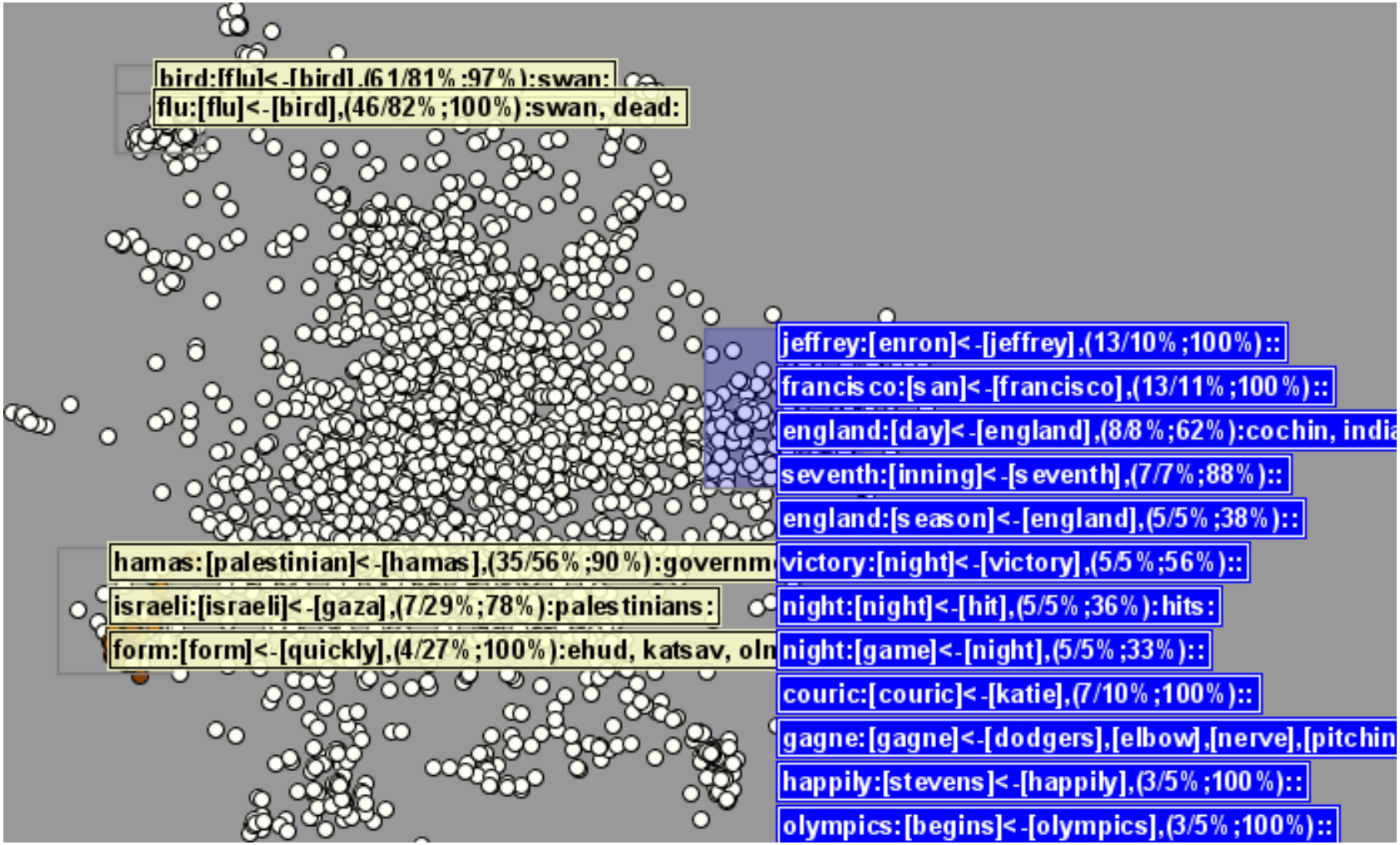
Exploration



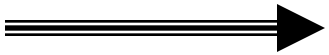

(vinci,da,code)
(flu,bird)
(immigration,senate)
(deal,shirt,sponsorship,aig)
(house,white)
(bae,airbus,stake,
(bush,president)
(windows,run,apple,software,macs,xp)
(san,francisco)
(palestinian,hamas)
(venezuela,caracas,brothers,crime)
(mosque,baghdad,suicide)

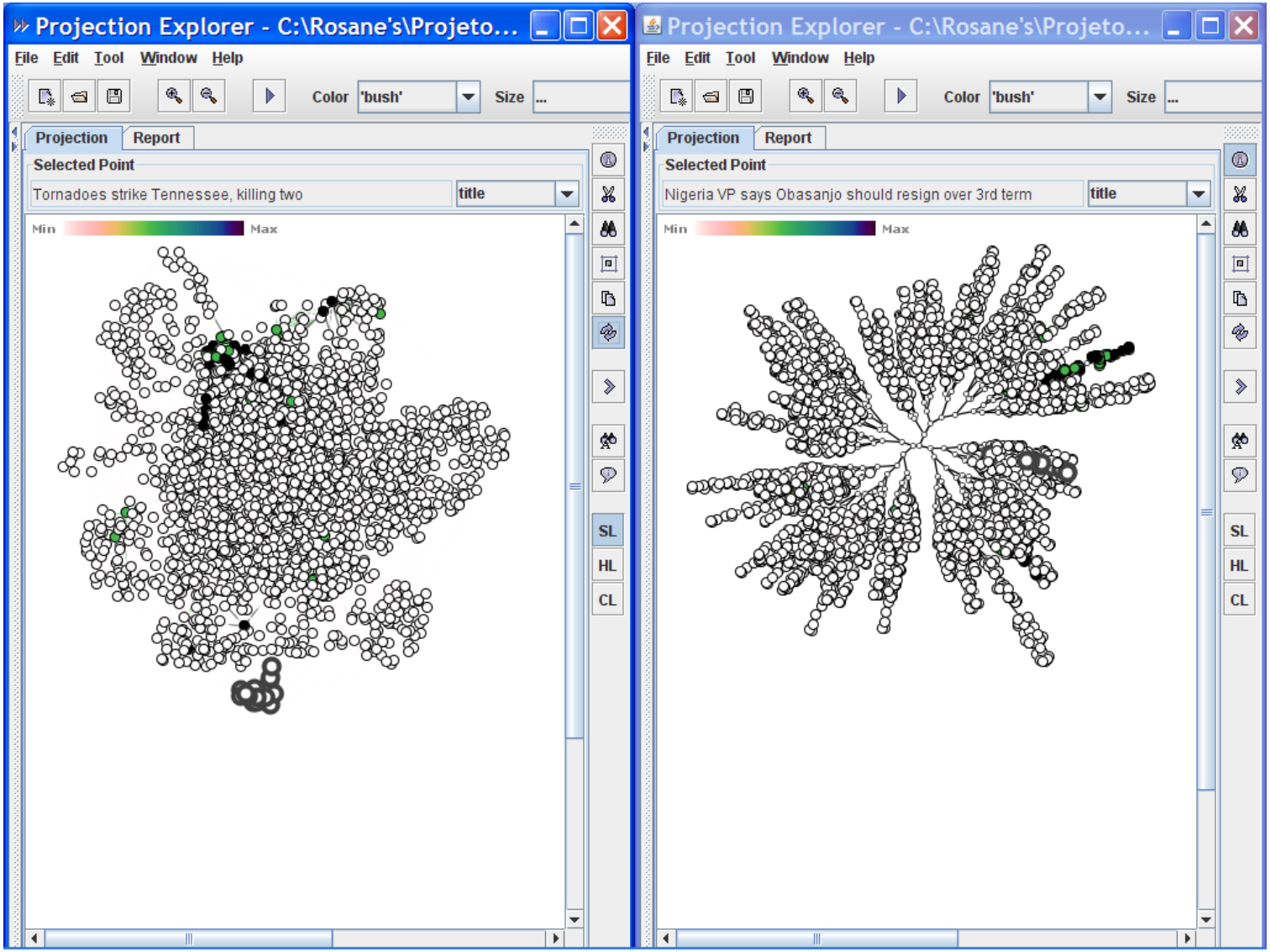
- Detailing topics



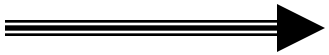

bird:[flu]<-[bird],(61/81%;97%):swan:
flu:[flu]<-[bird],(46/82%;100%):swan, dead:
jeffrey:[enron]<-[jeffrey],(13/10%;100%)::
francisco:[san]<-[francisco],(13/11%;100%)::
england:[day]<-[england],(8/8%;62%):cochin, india
seventh:[inning]<-[seventh],(7/7%;88%)::
england:[season]<-[england],(5/5%;38%)::
hamas:[palestinian]<-[hamas],(35/56%;90%):governm
victory:[night]<-[victory],(5/5%;56%)::
israeli:[israeli]<-[gaza],(7/29%;78%):palestinians:
night:[night]<-[hit],(5/5%;36%):hits:
form:[form]<-[quickly],(4/27%;100%):ehud, katsav, oln
night:[game]<-[night],(5/5%;33%)::
couric:[couric]<-[katie],(7/10%;100%)::
gagne:[gagne]<-[dodgers],[elbow],[nerve],[pitchin
happily:[stevens]<-[happily],(3/5%;100%)::
olympics:[begins]<-[olympics],(3/5%;100%)::

- Finding Relationships

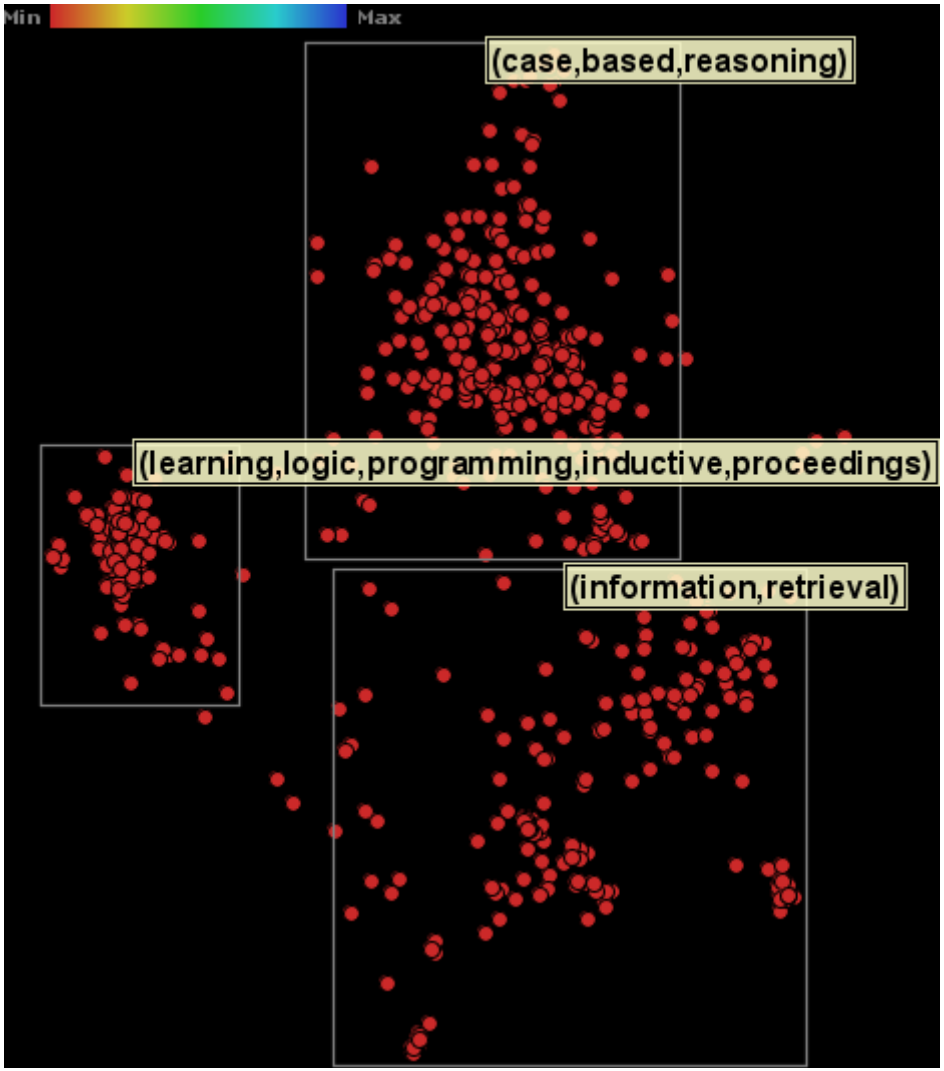



Projection Explorer - C:\Rosane's\Projeto...
File Edit Tool Window Help
Color 'bush'
Size ...
Projection Report
Selected Point
Tornadoes strike Tennessee, killing two
title
Min Max
SL HL CL
Projection Explorer - C:\Rosane's\Projeto...
File Edit Tool Window Help
Color 'bush'
Size ...
Projection Report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

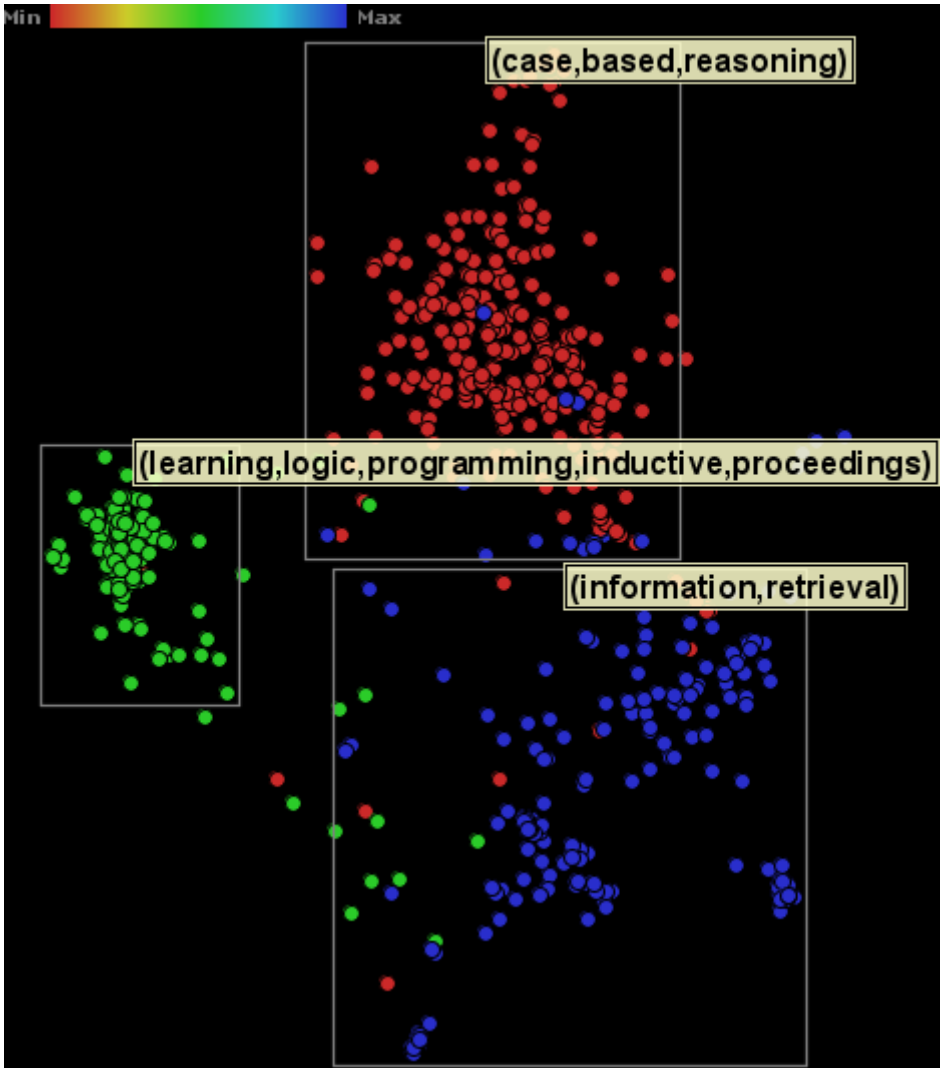
- Building a Surface



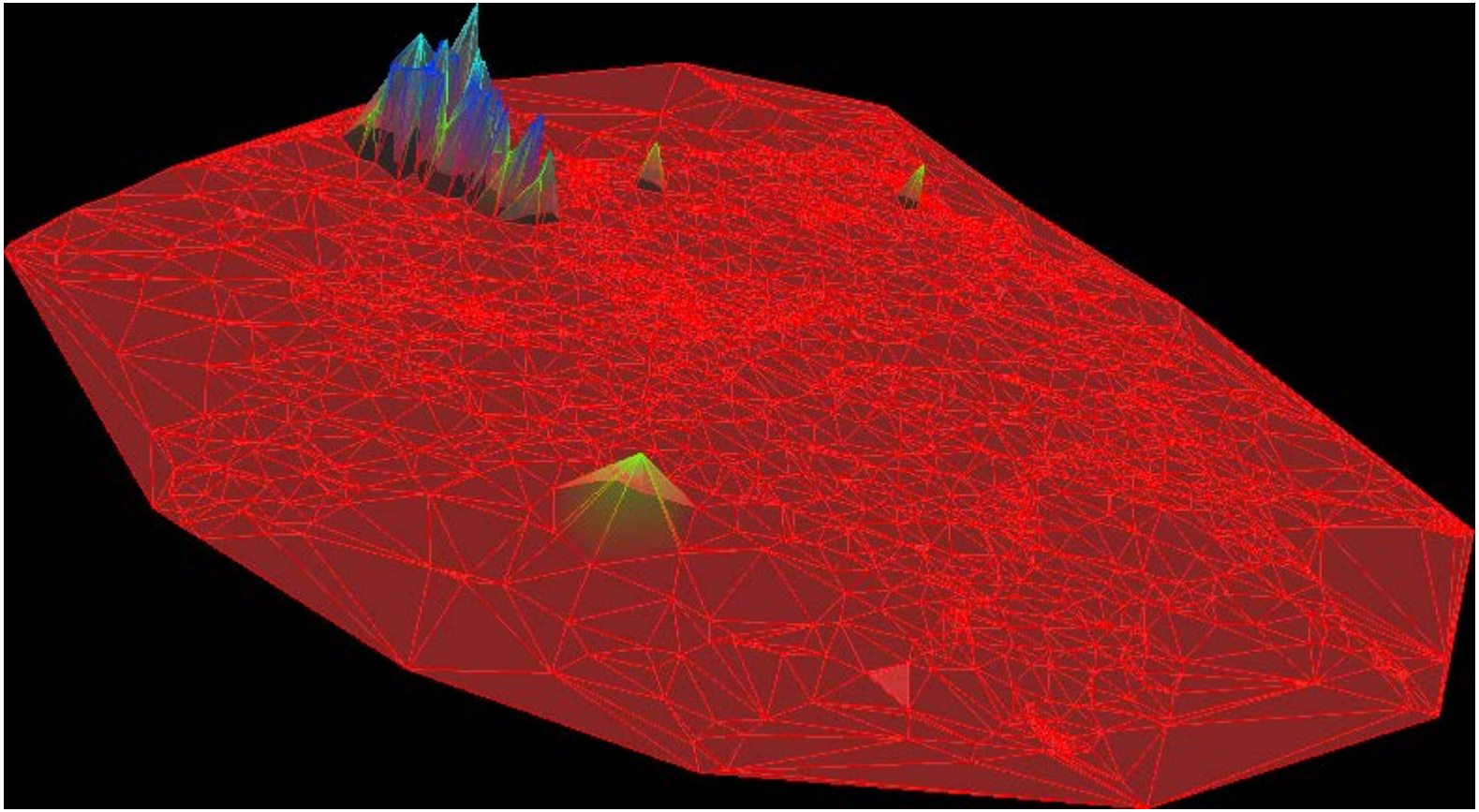
Exemplos de Mapas



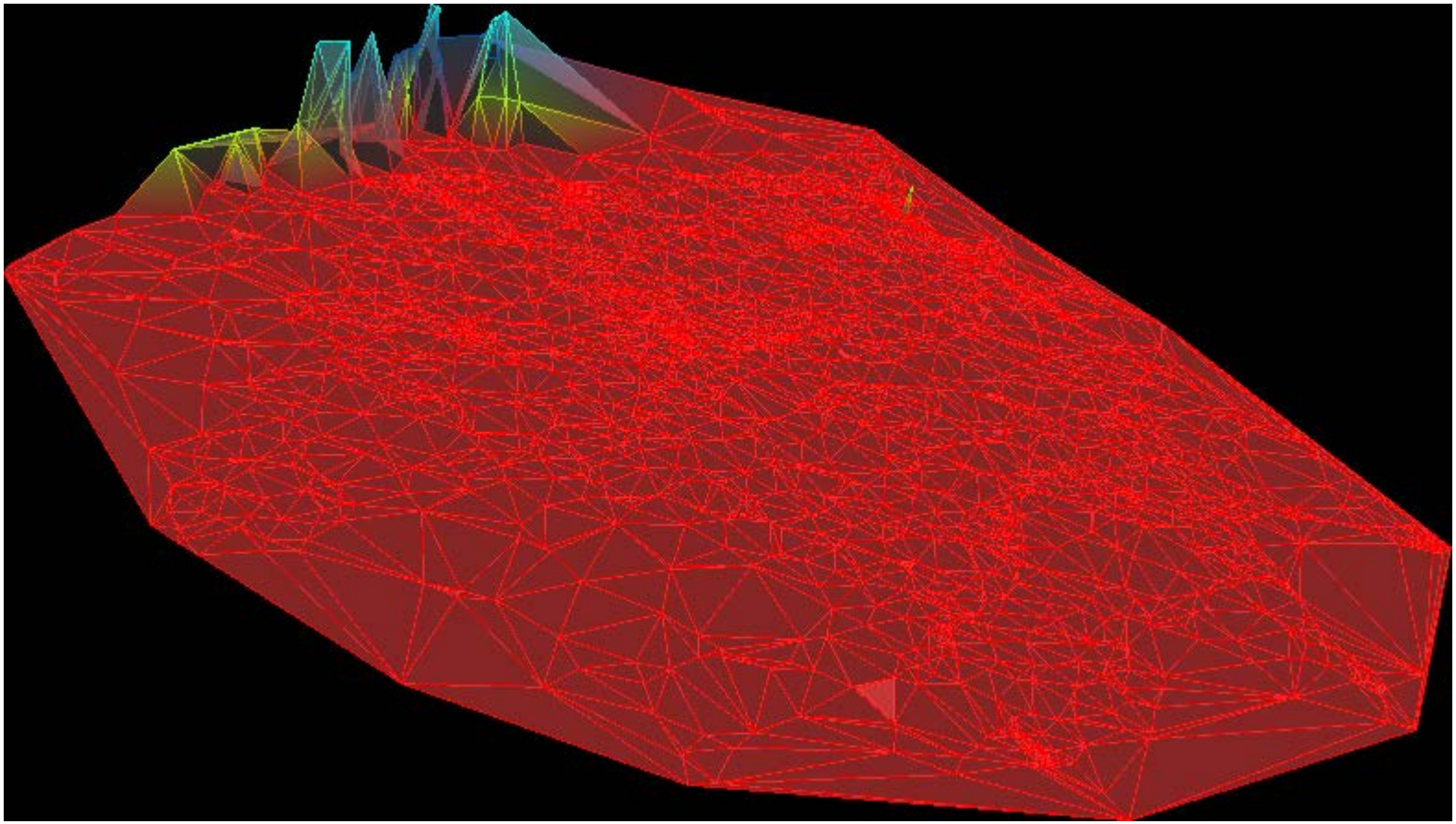
Exemplos de Mapas



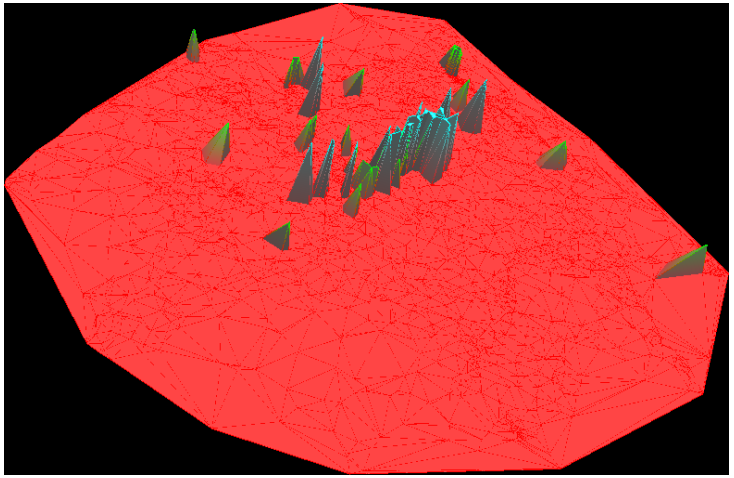
RSS News Flash



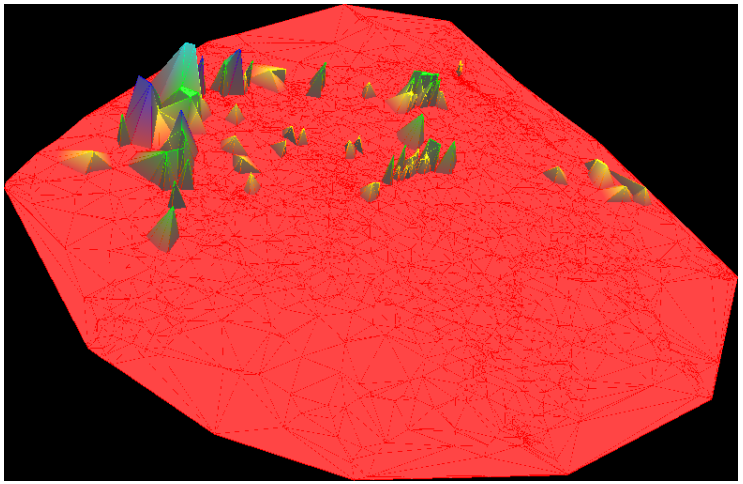
Bird and Flu



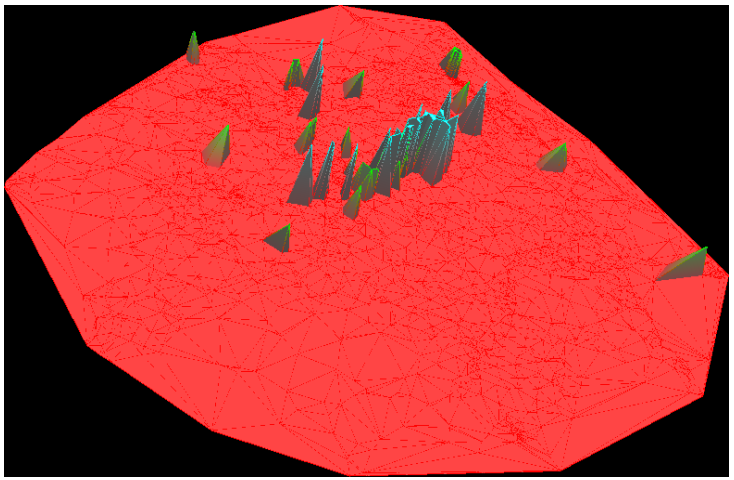
Palestinian



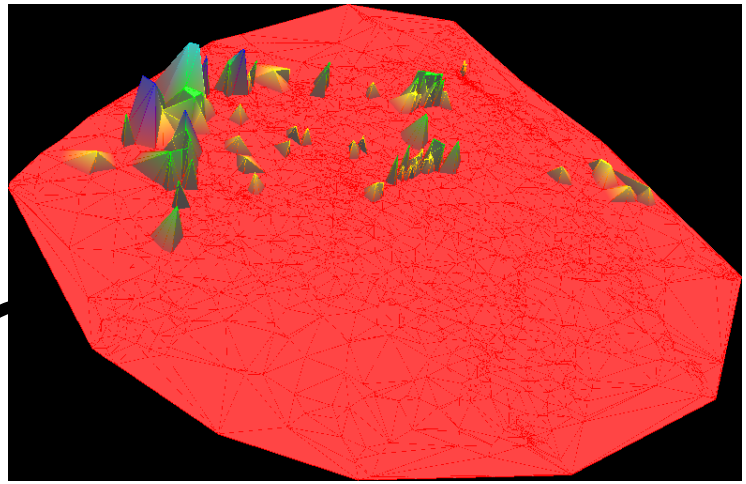
Bush



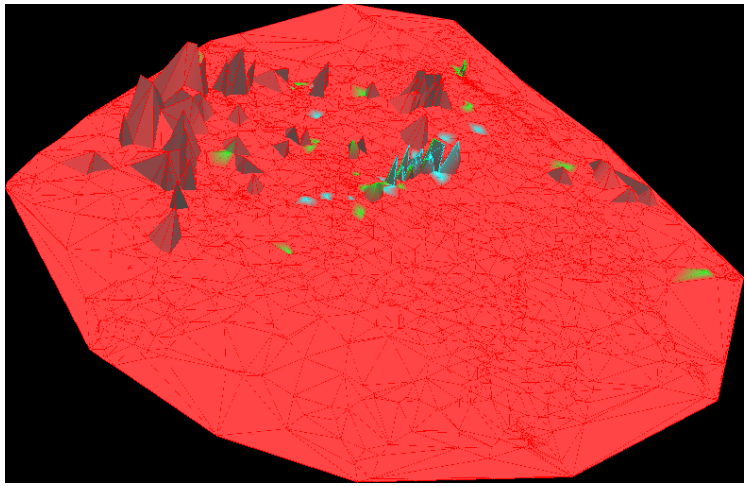
Iraq



Bush



Iraq



Generalizando o processo:
Séries Temporais – Vazão em Hidrelétricas

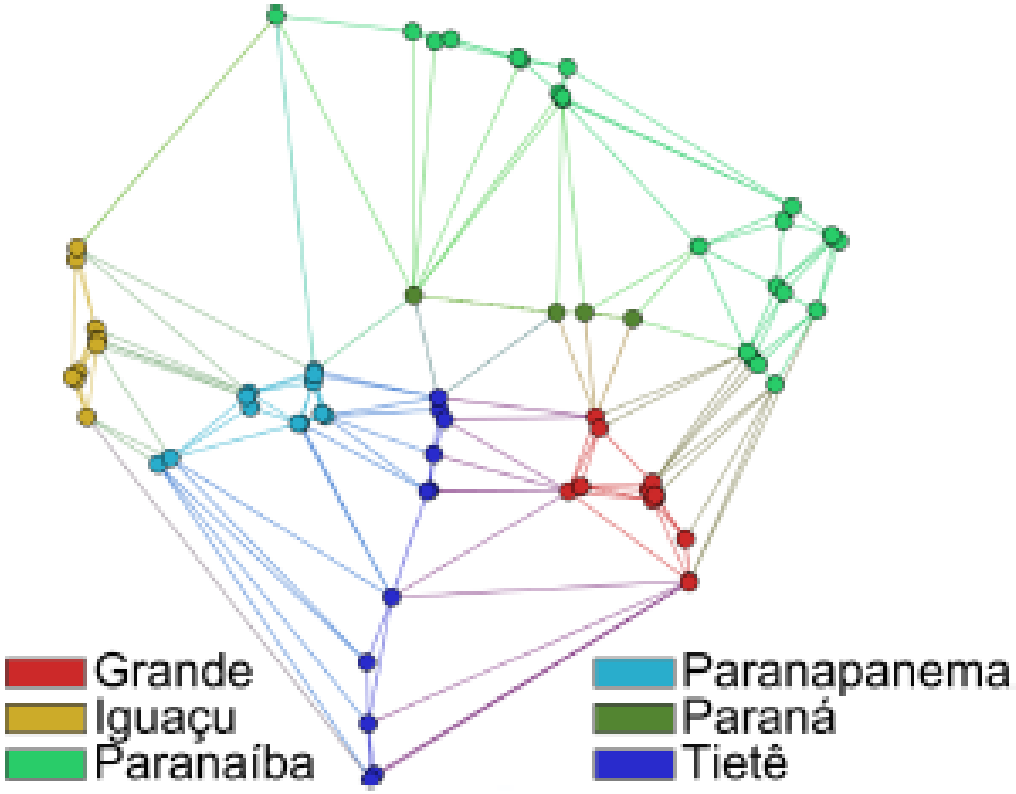


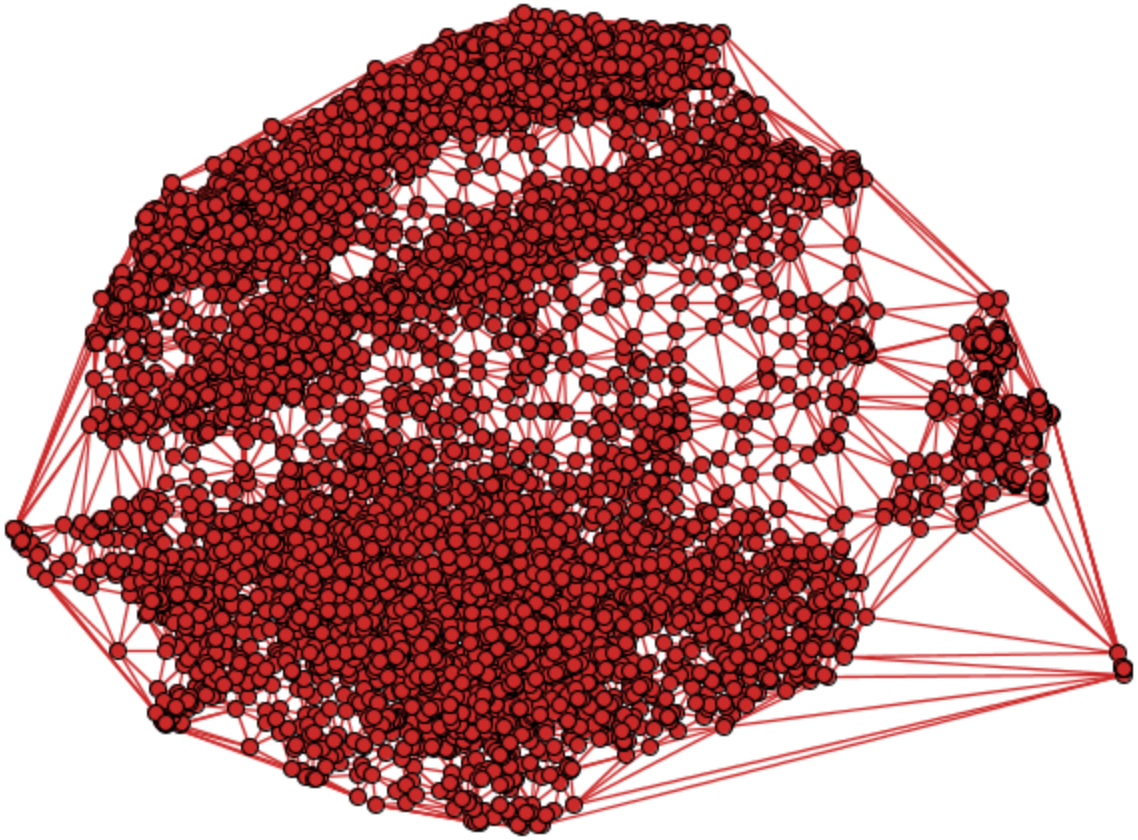
Figure 2. Power plants of the basin Paraná

Text from attributes

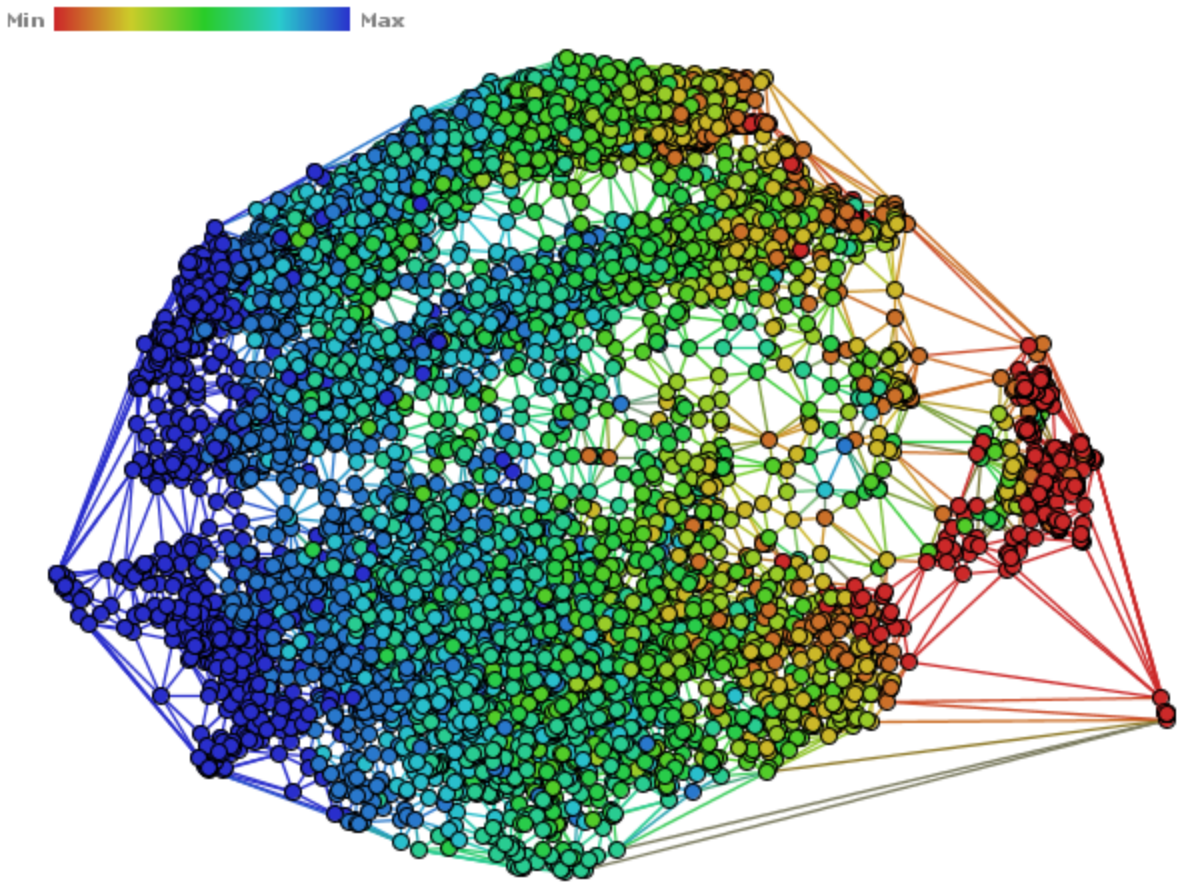
- Cattle performance data
  - Translated to text from categorical information, e.g.,
    - Ranges of weight to words such as:

      {weight_below_fifty_percent; weight_between_fifty_seventy_five; etc..}
  - 9135 individuals

Cattle performance data

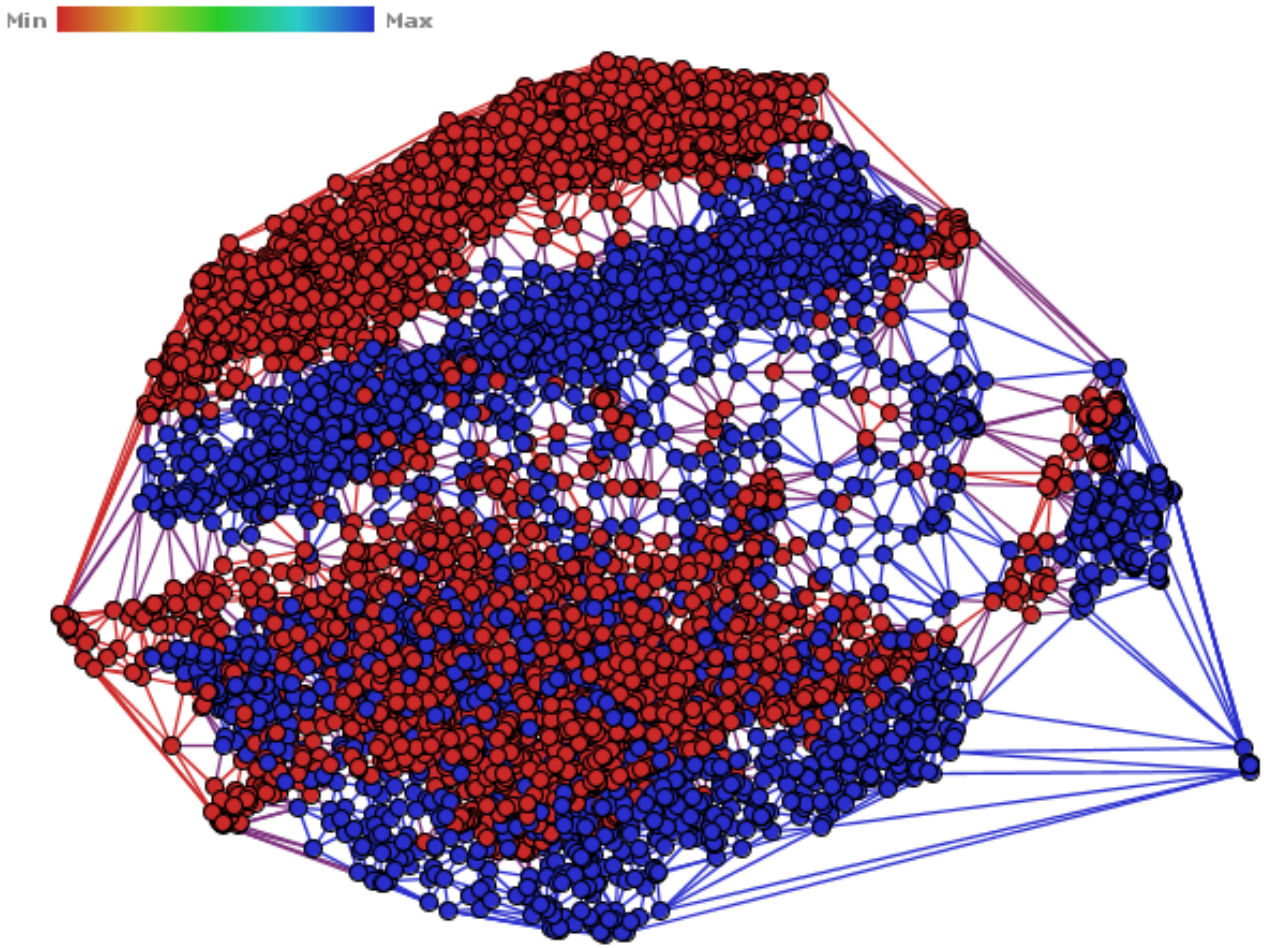


Cattle performance data



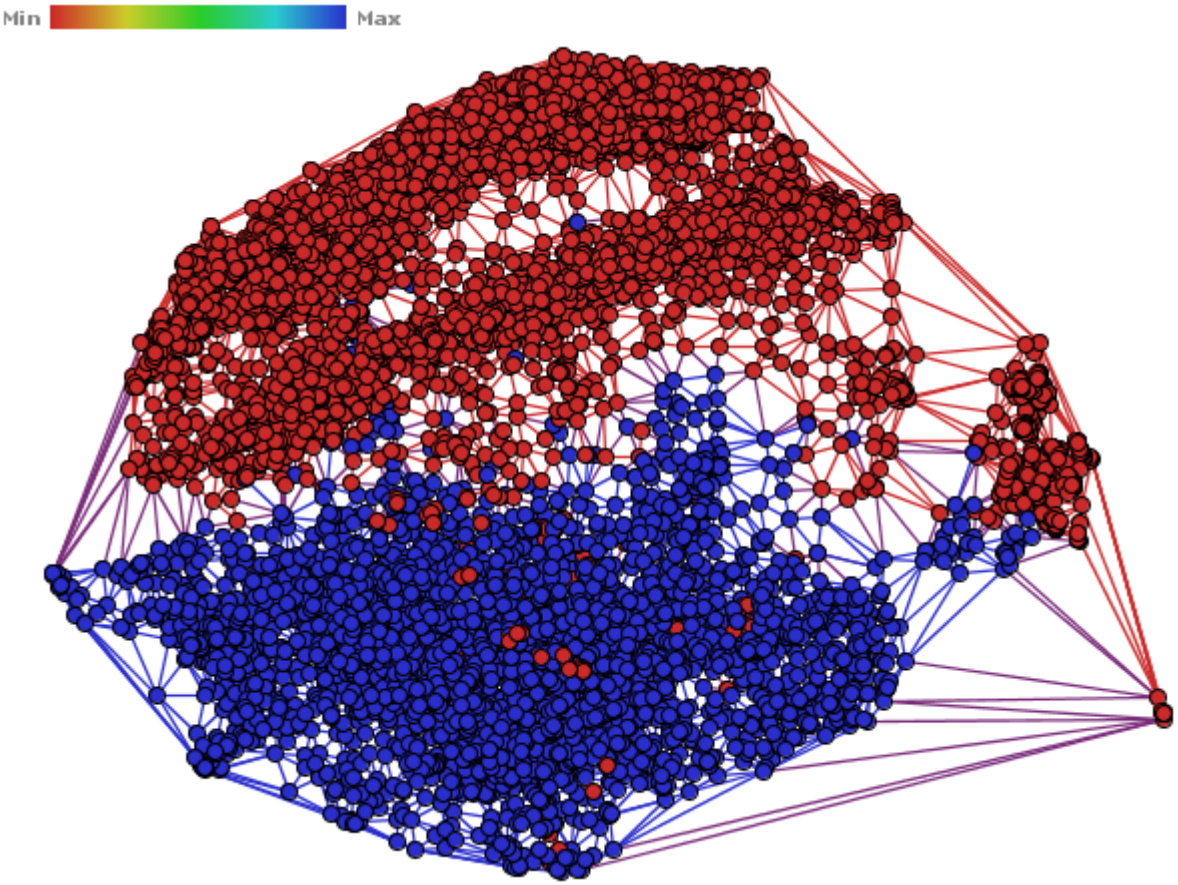
Colored by word ‘top’

Cattle performance data



Colored by female

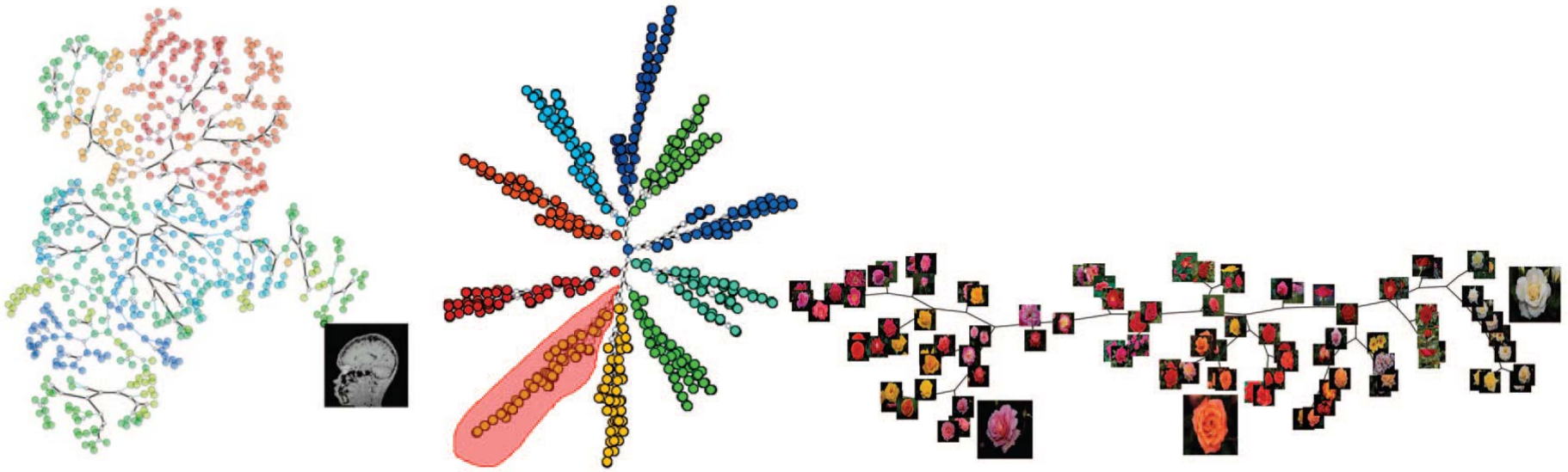
Cattle performance data



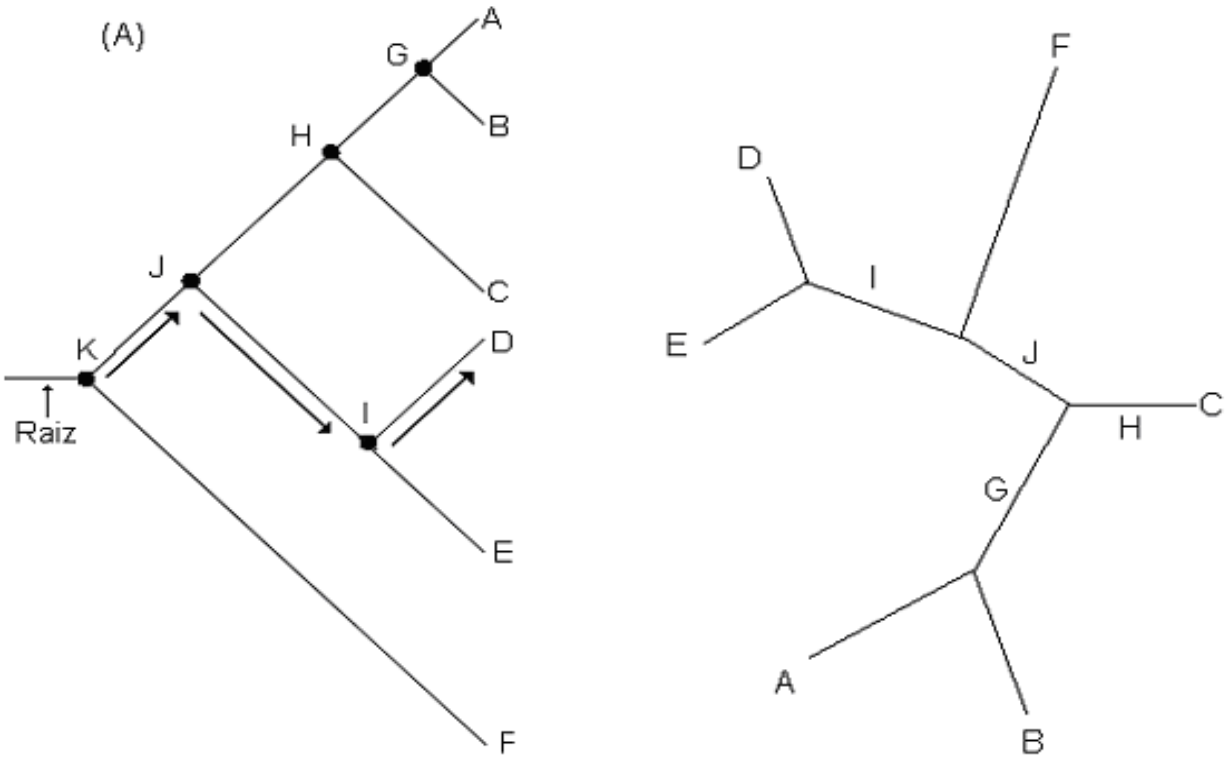
Colored by farm

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.*

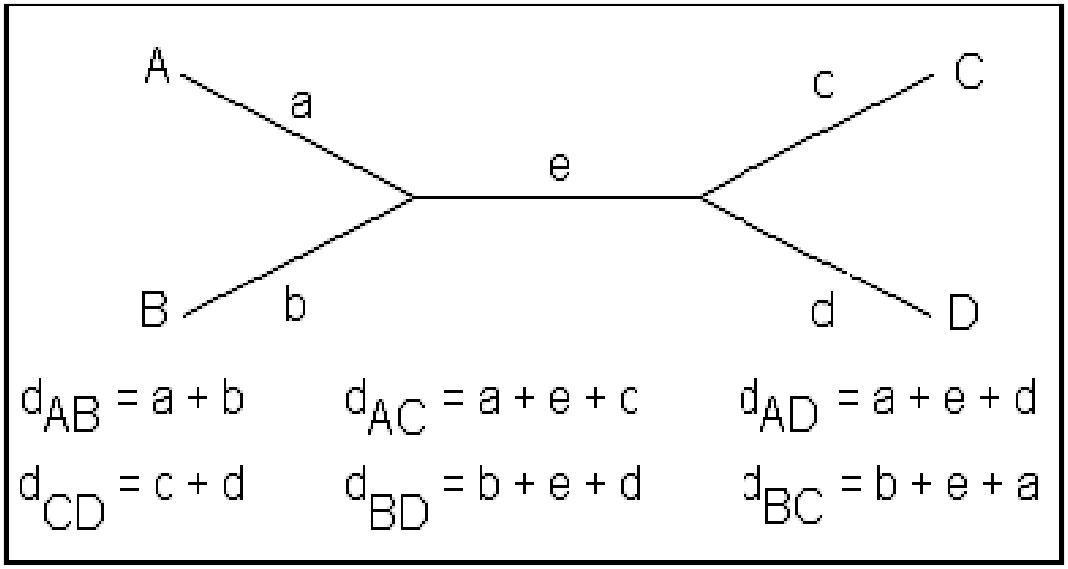


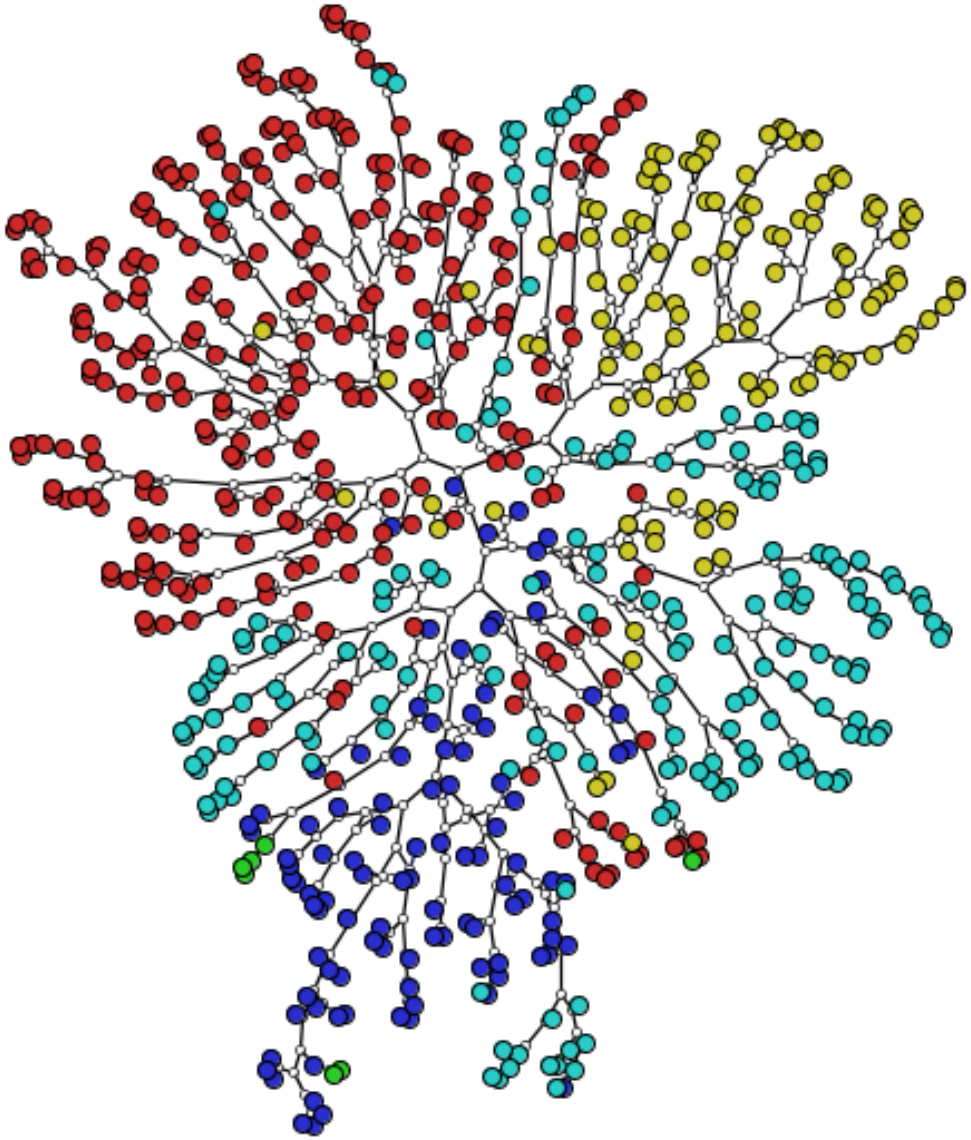
Point Placement by Phylogenetic Tree Construction Algorithms (N-J Trees)



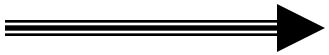
Point Placement by Phylogenetic Tree Construction Algorithms (N-J Trees)

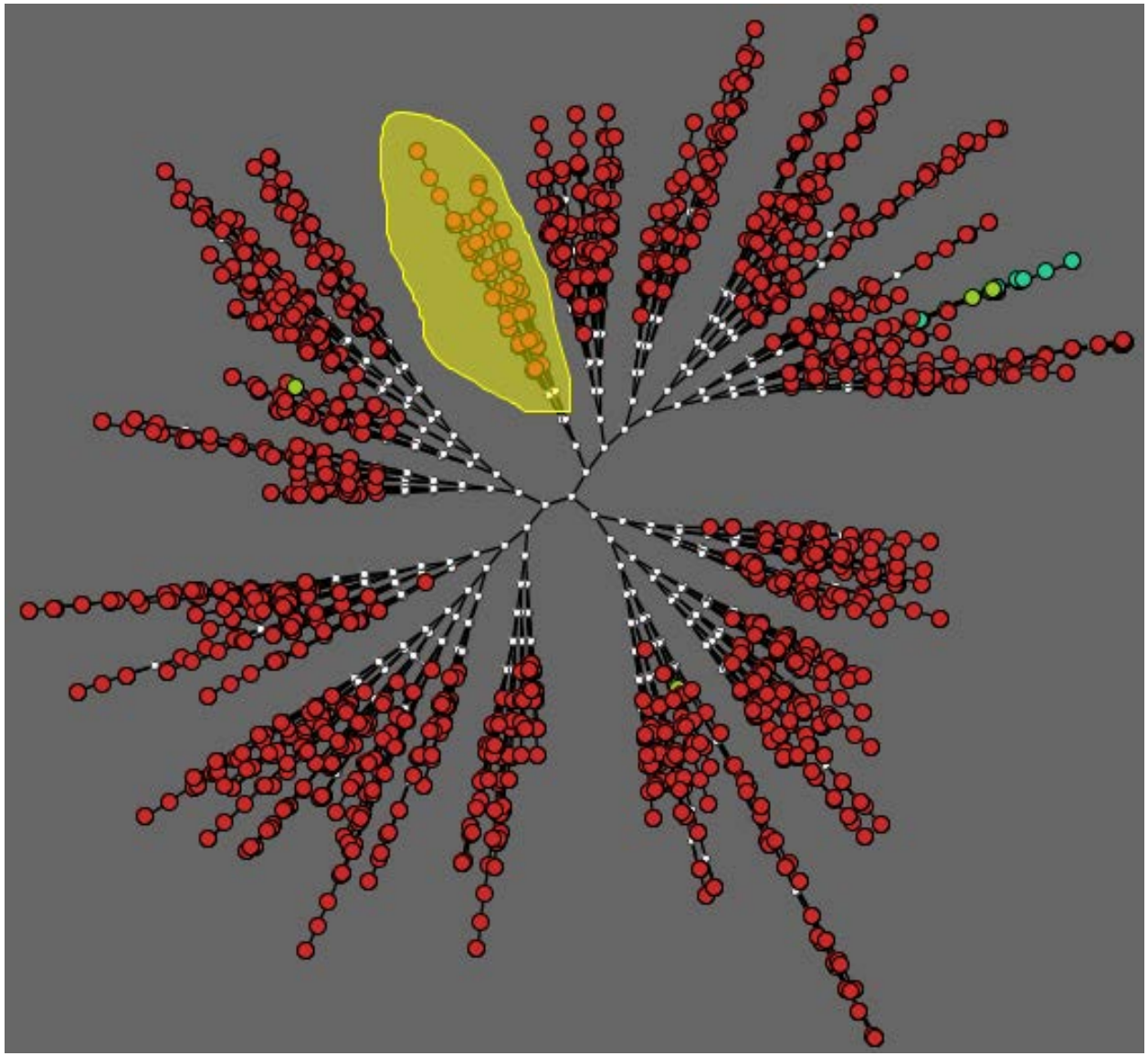
$$d_{AB} + d_{CD} \leq \max\left(d_{AC} + d_{BD}, d_{AD} + d_{BC}\right)$$





- Alternate view (N-J Tree)





Images?

Pipeline

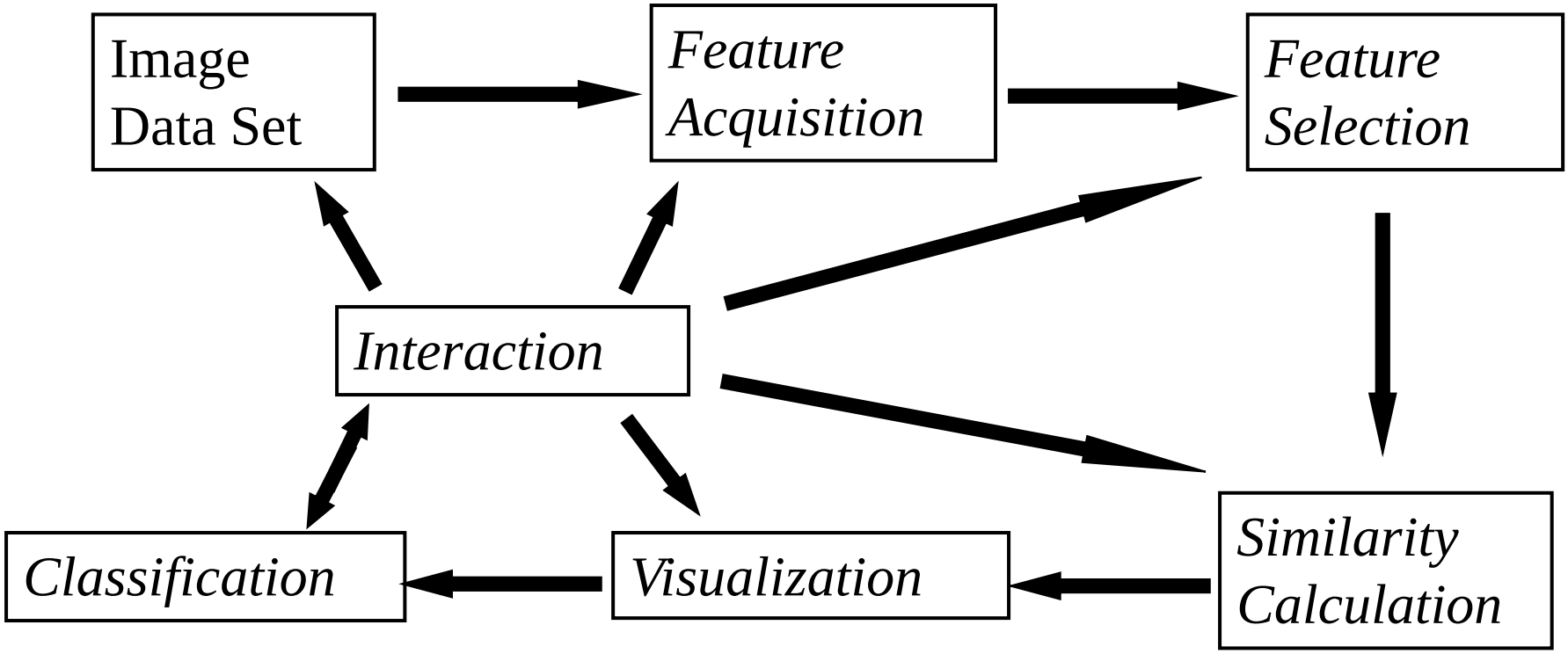
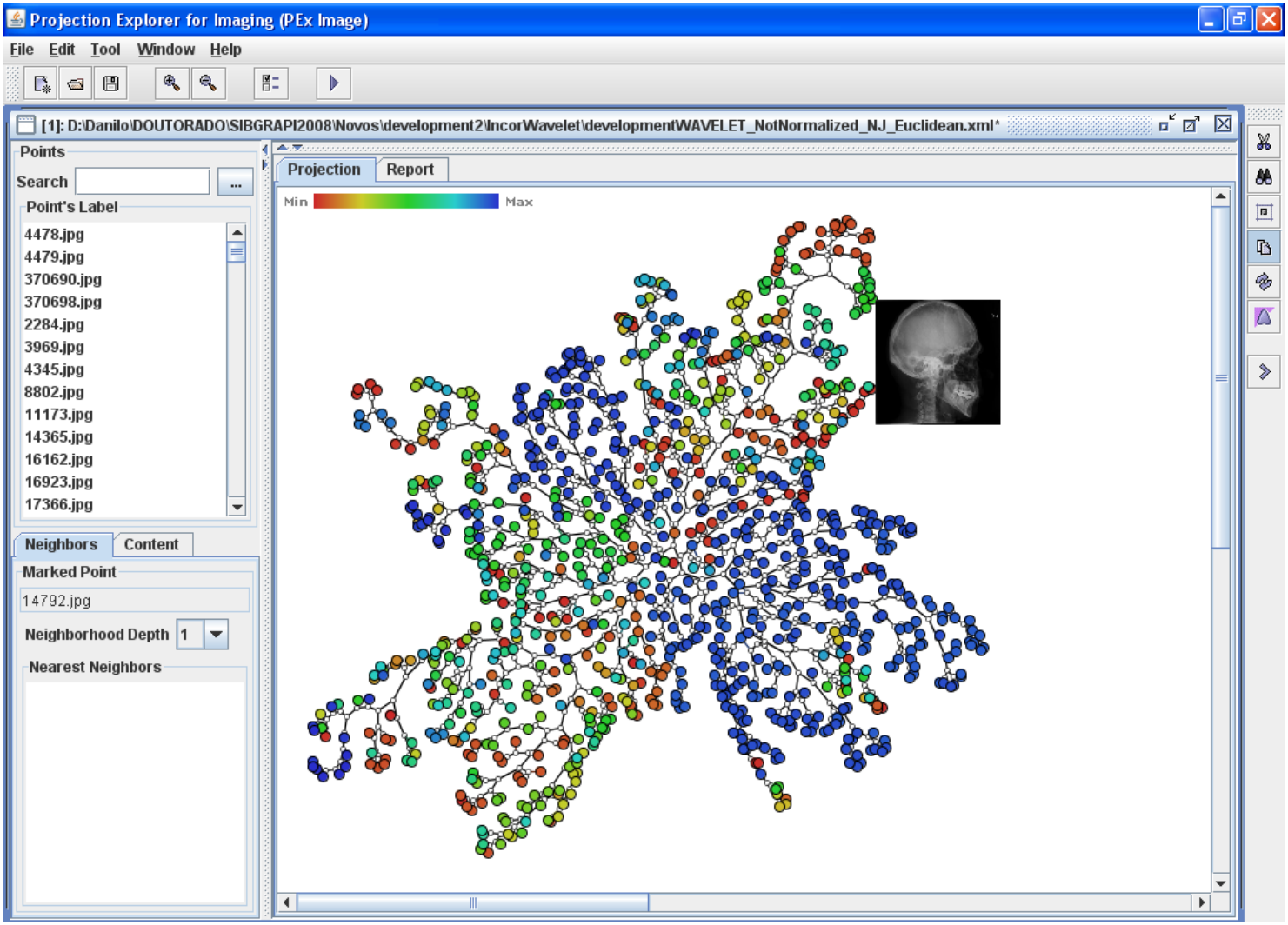
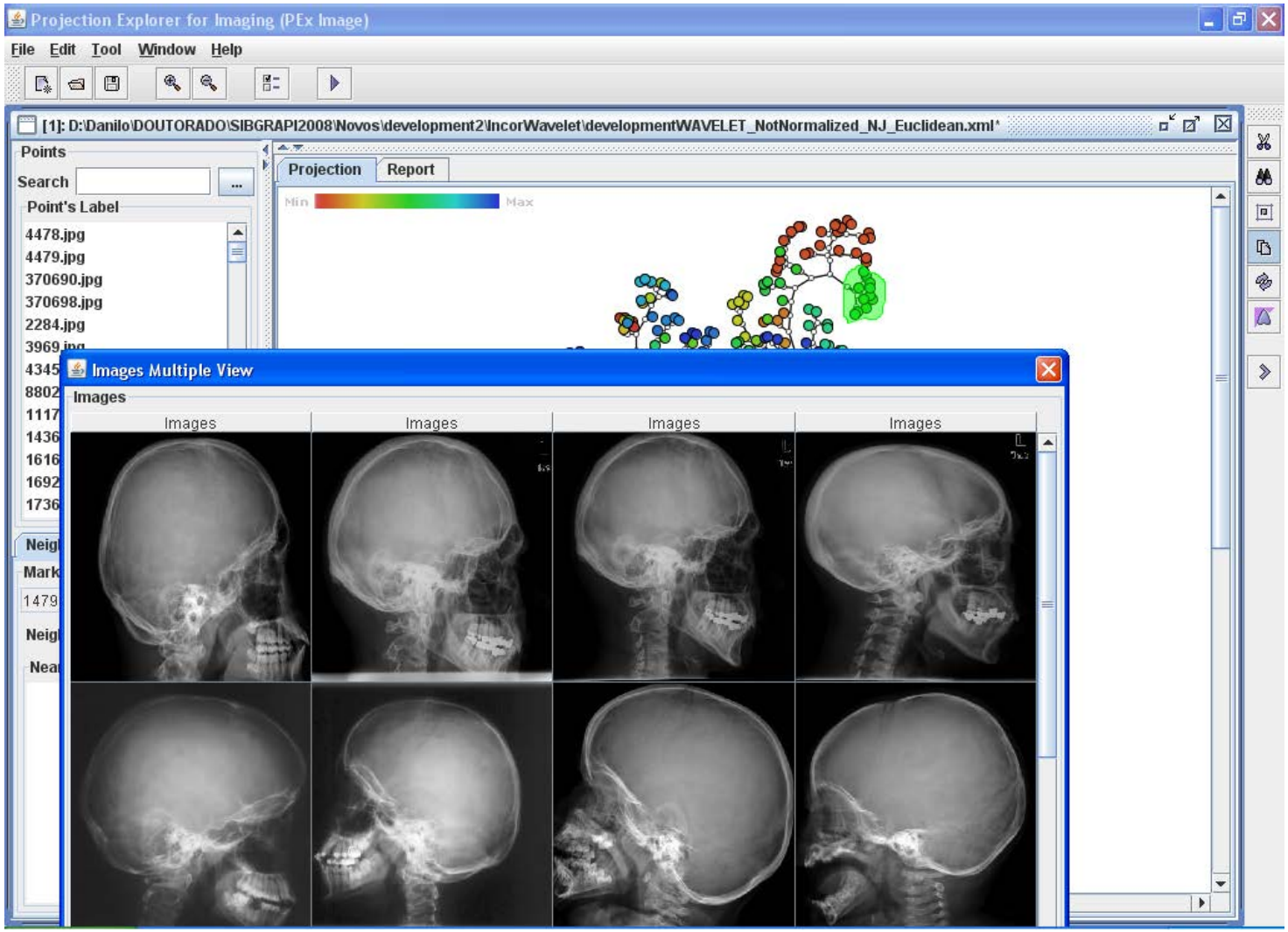


Image
Data Set
Feature
Acquisition
Feature
Selection
Interaction
Classification
Visualization
Similarity
Calculation

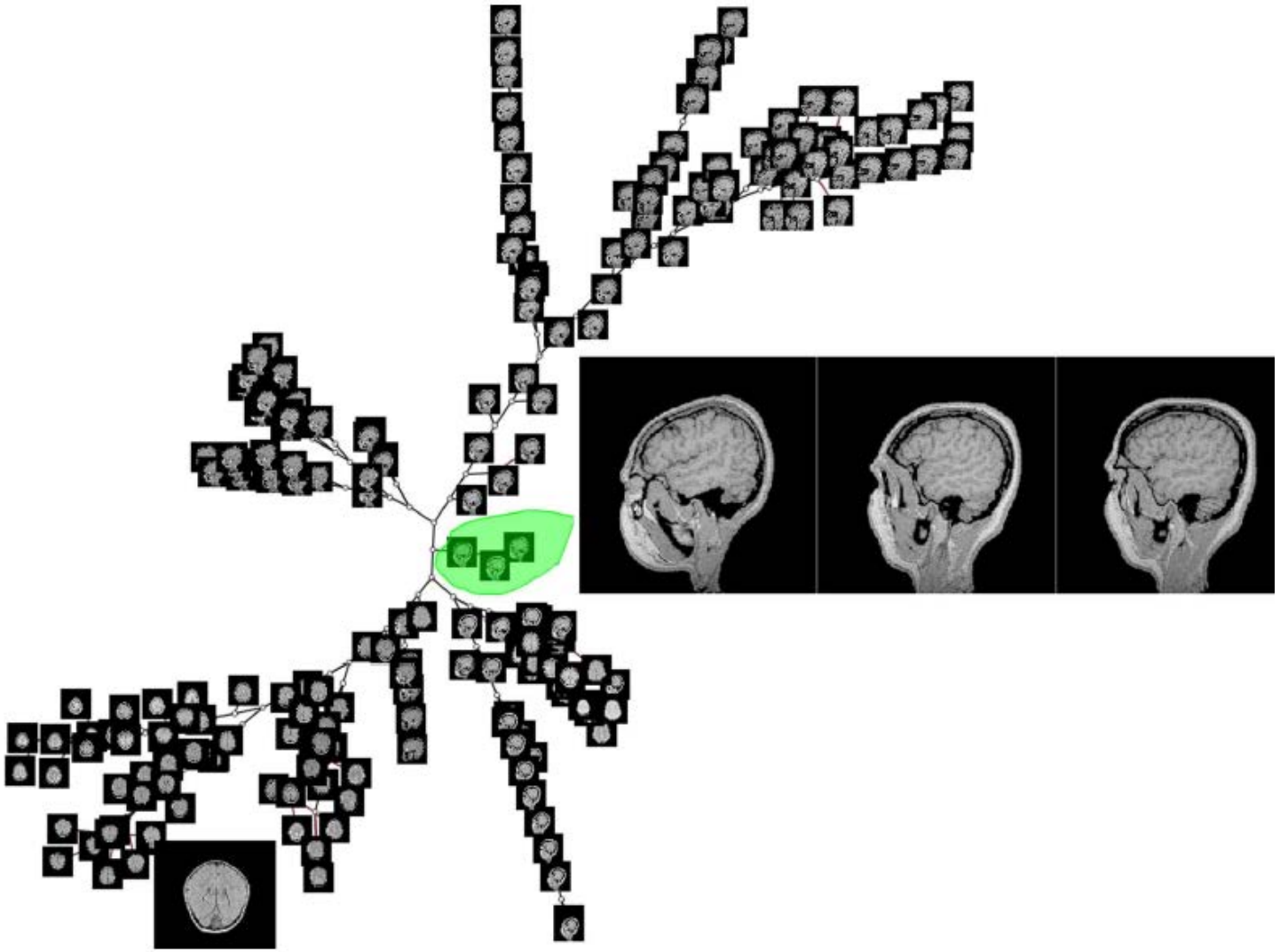
PEx-Image – Sample Content



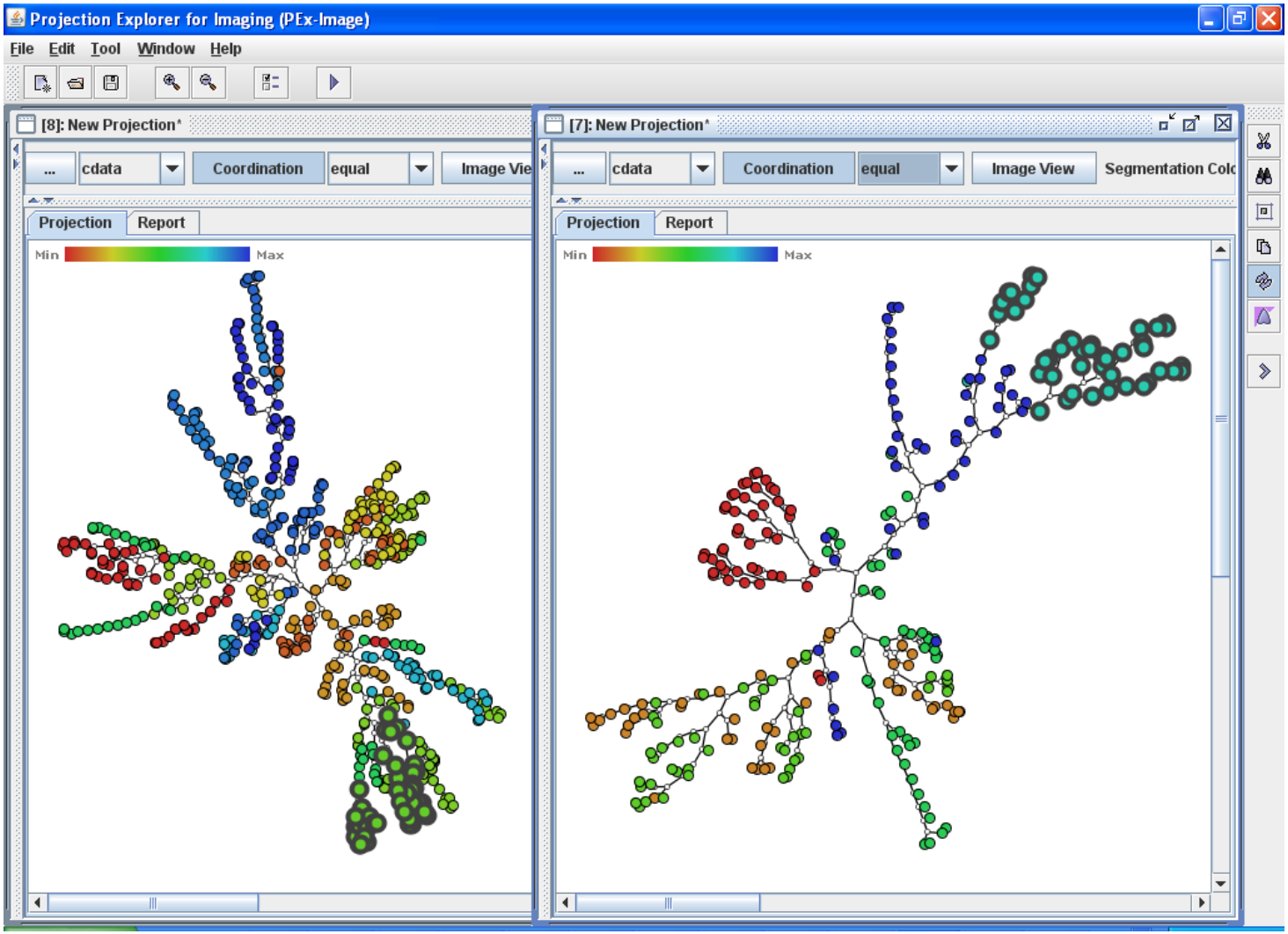
PEx-Image – Group Content



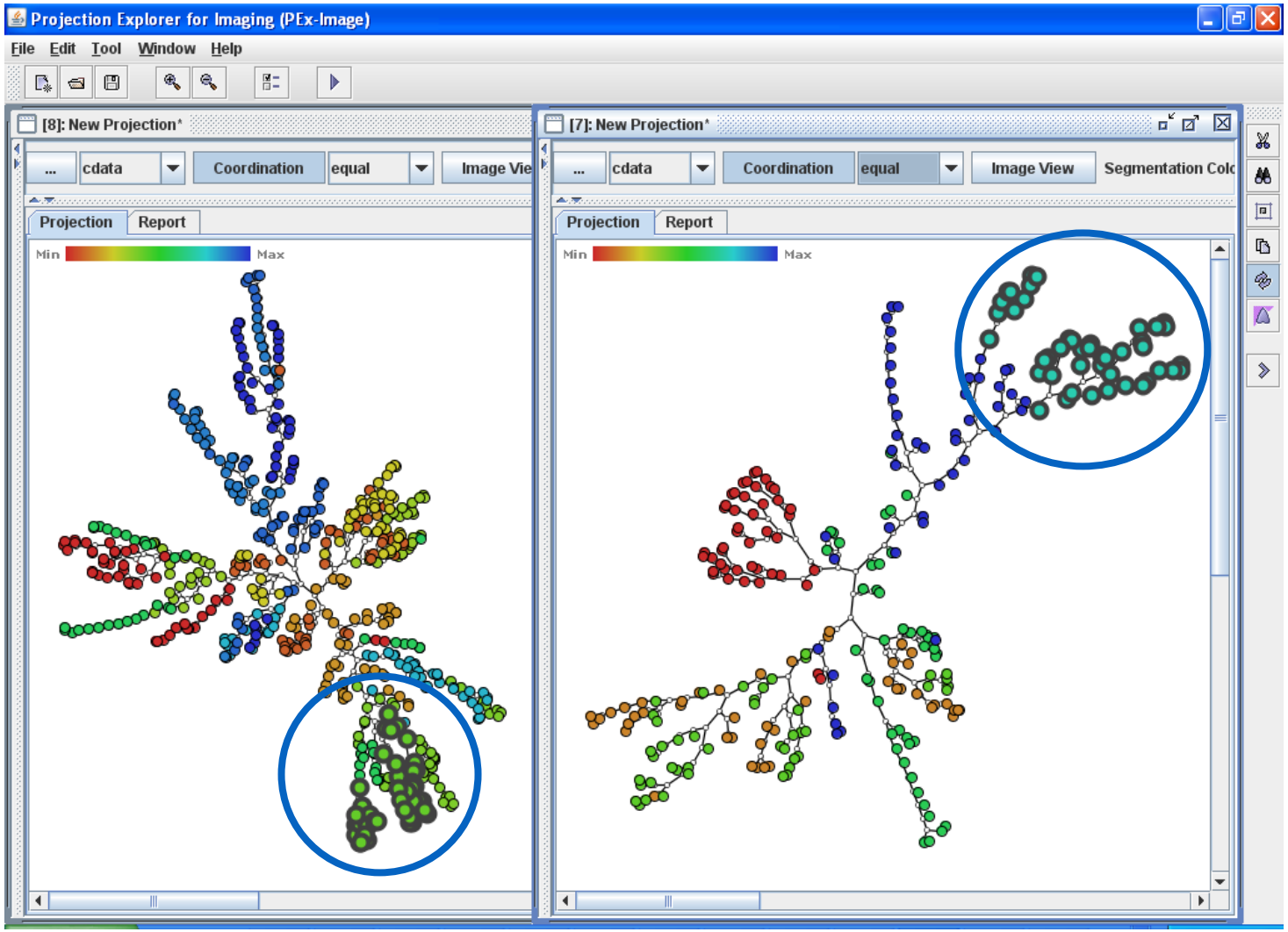
PEx-Image – Image as Visual Mark



PEx-Image – Coordination

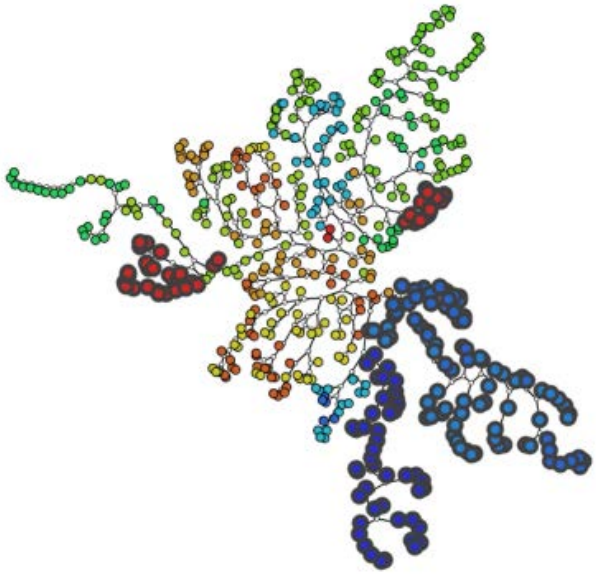


PEx-Image – Coordination

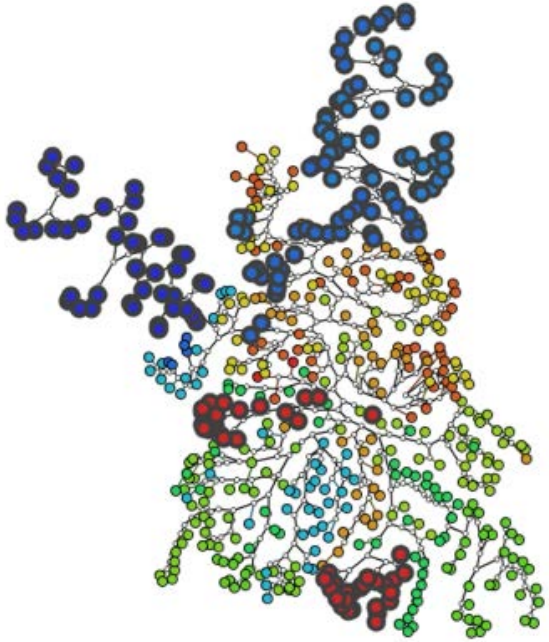


Comparison of Distance Metrics

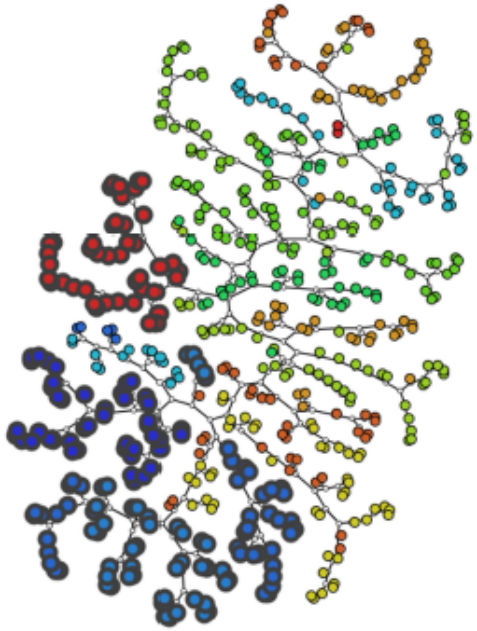
Euclidean



City Block



Cosine



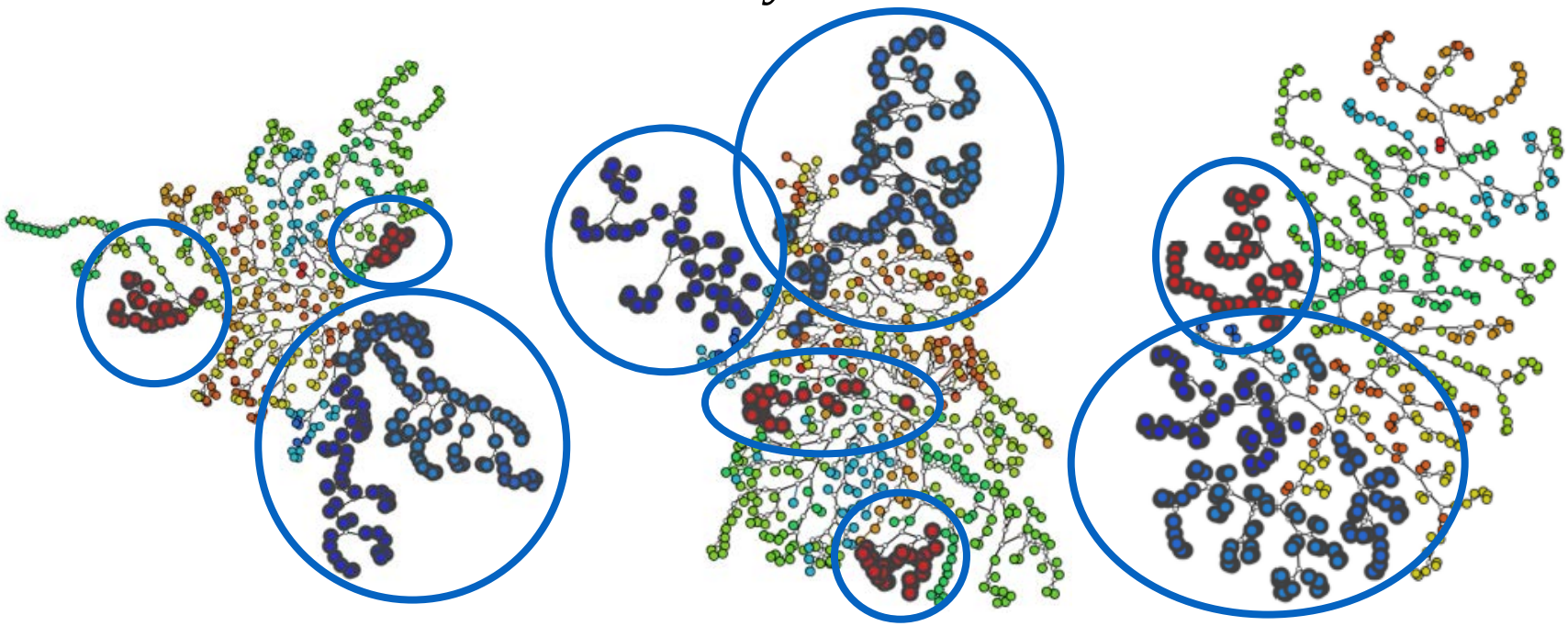
512 MRI medical images
12 classes

Comparison of Distance Metrics

Euclidean

City Block

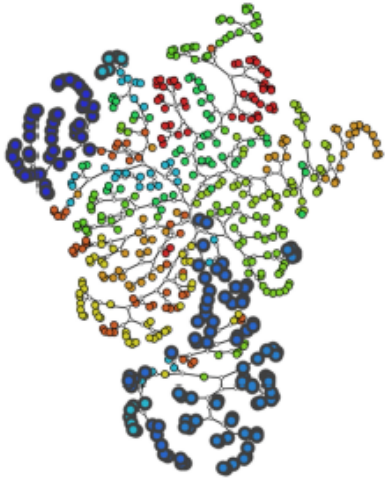
Cosine



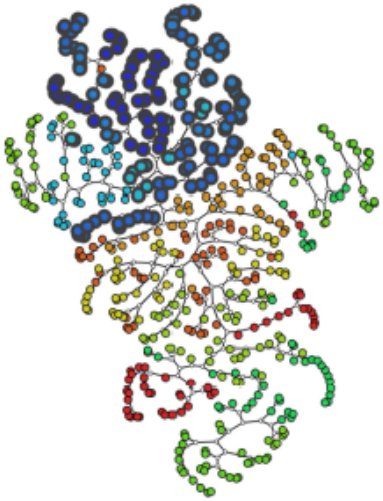
512 MRI medical images
12 classes

Comparison of Feature Space (1)

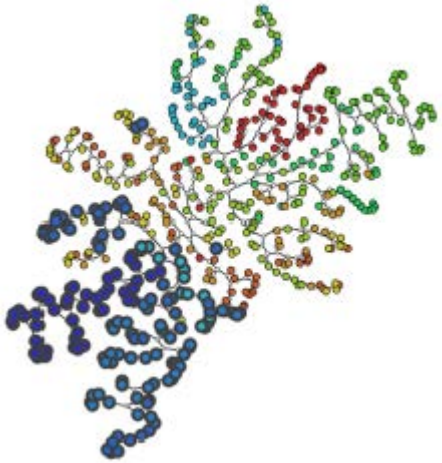
16 Gabor Filters



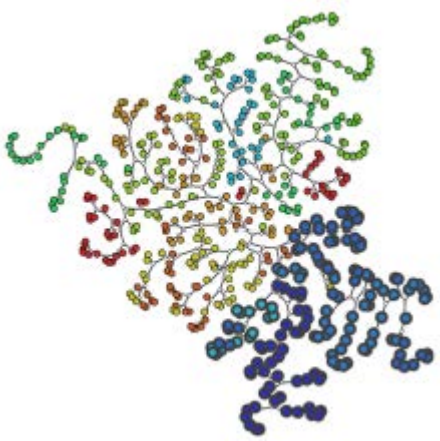
Fourier, Mean and Deviation



72 co-ocurrence matrices



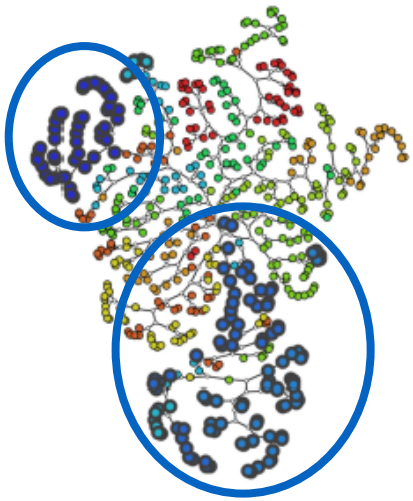
All combined



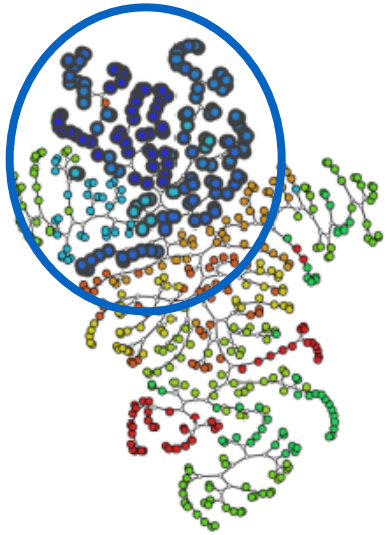
512 MRI medical images
12 classes

Comparison of Feature Space (1)

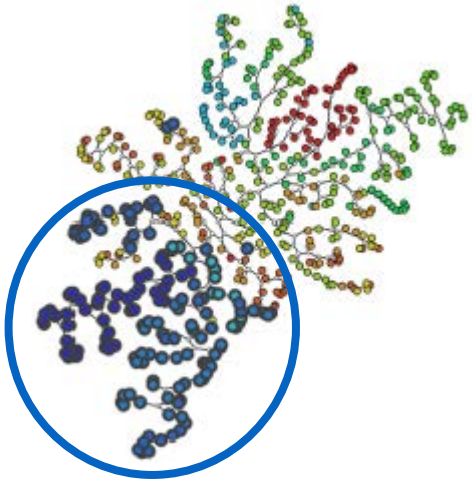
16 Gabor Filters



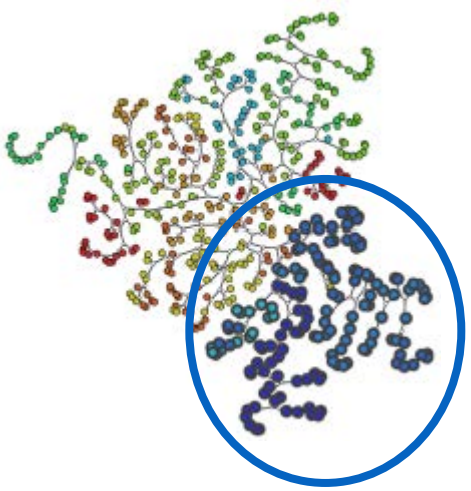
Fourier, Mean and Deviation



72 co-ocurrence matrices



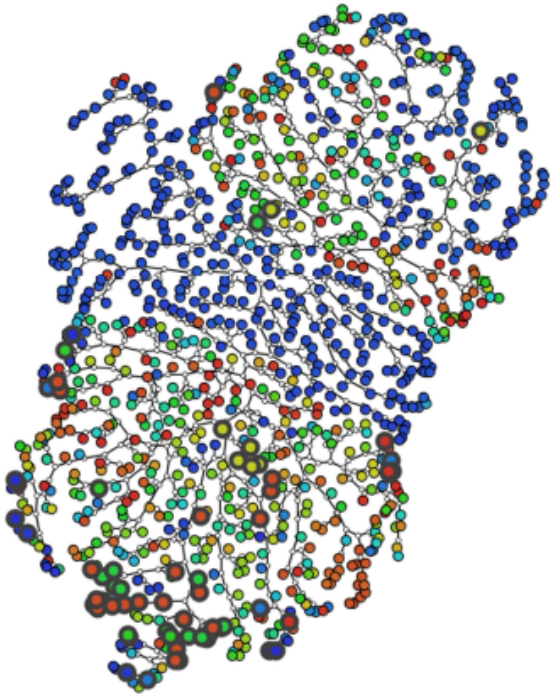
All combined



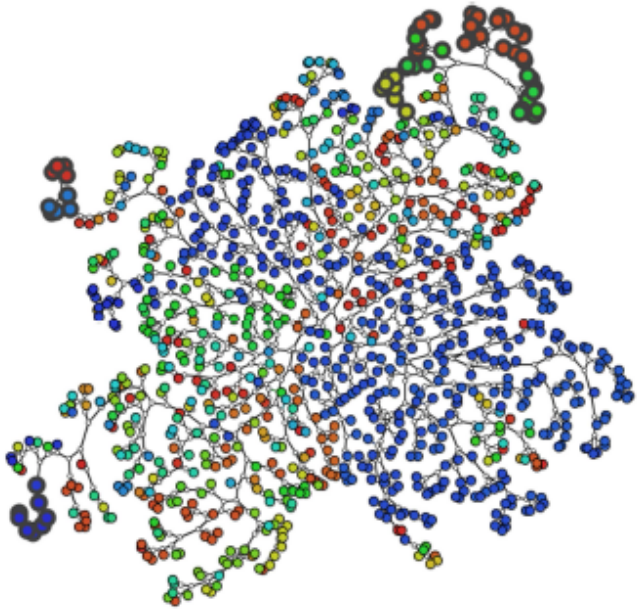
512 MRI medical images
12 classes

Comparison of Feature Space (2)

All combined



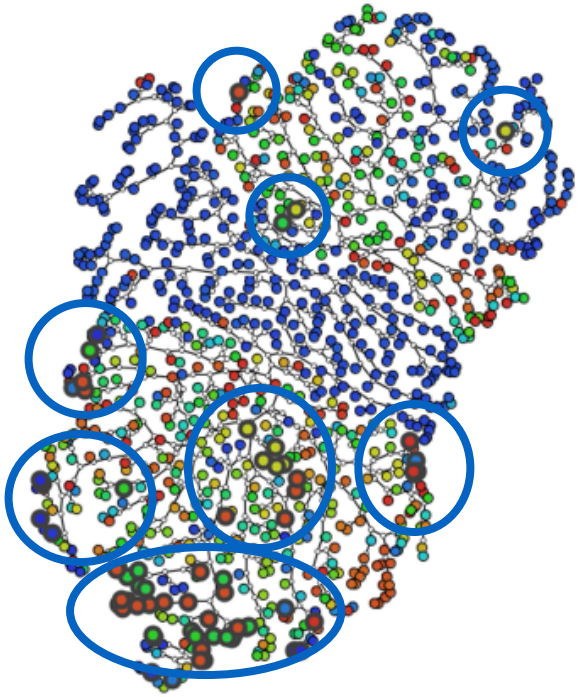
1024 Wavelet Features



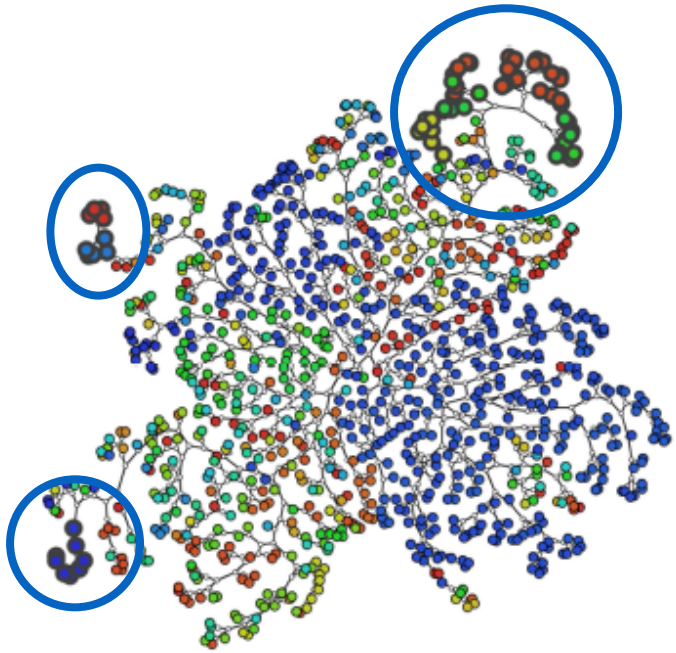
1000 X-Ray images from ImageCLEF
116 classes

Comparison of Feature Space (2)

All combined

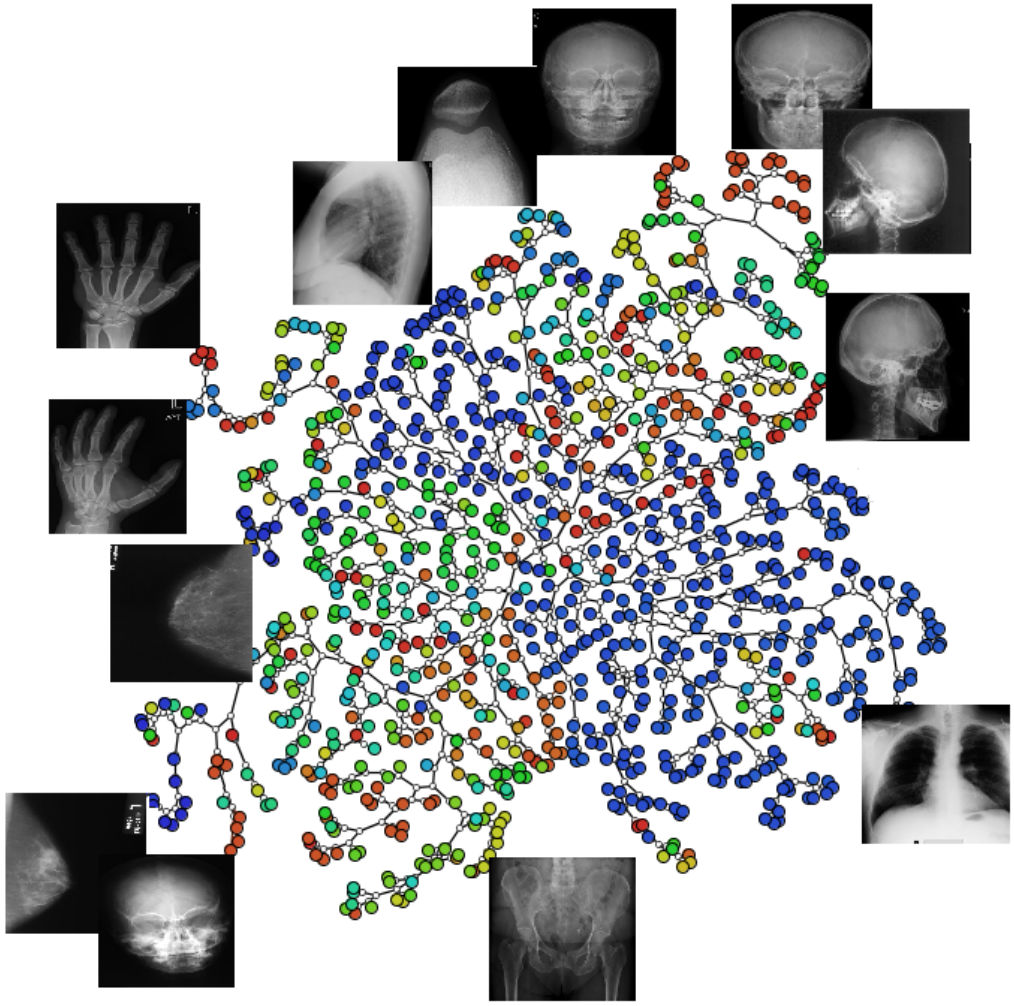


1024 Wavelet Features

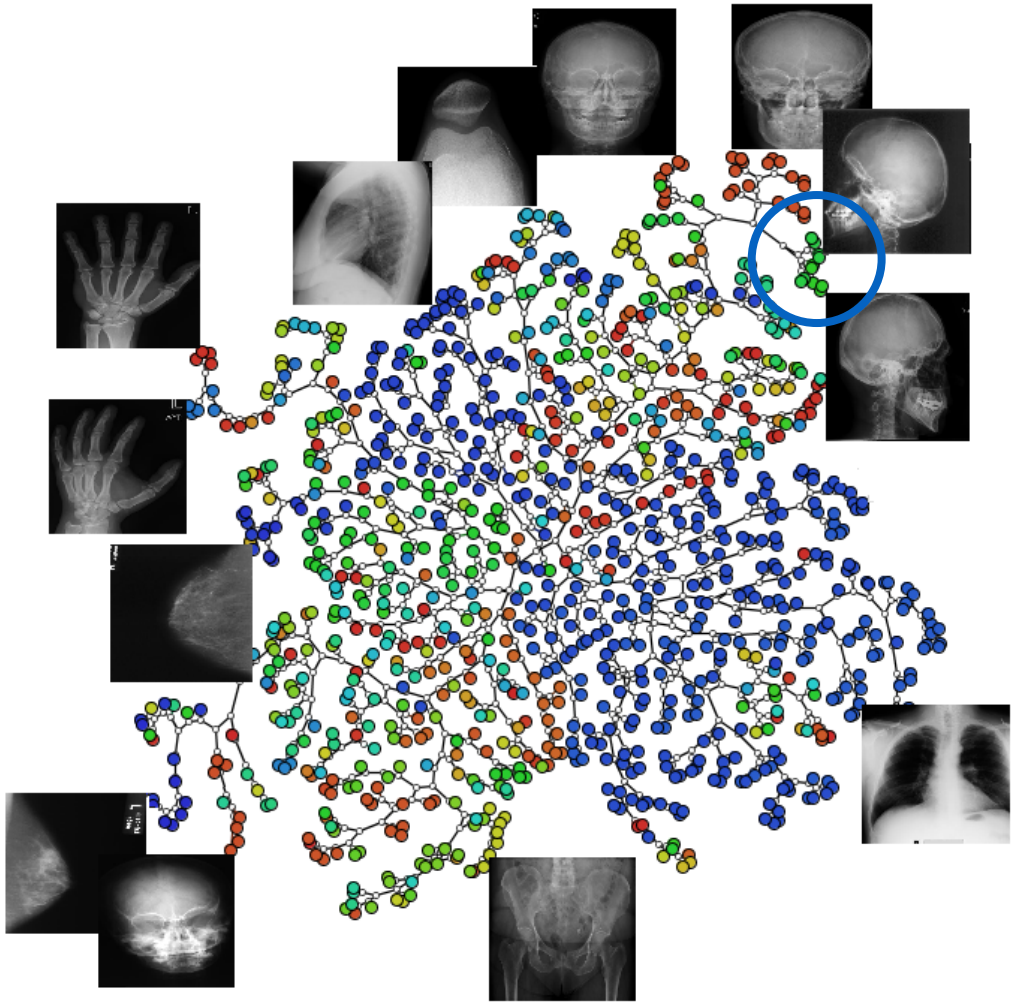


1000 X-Ray images from ImageCLEF
116 classes

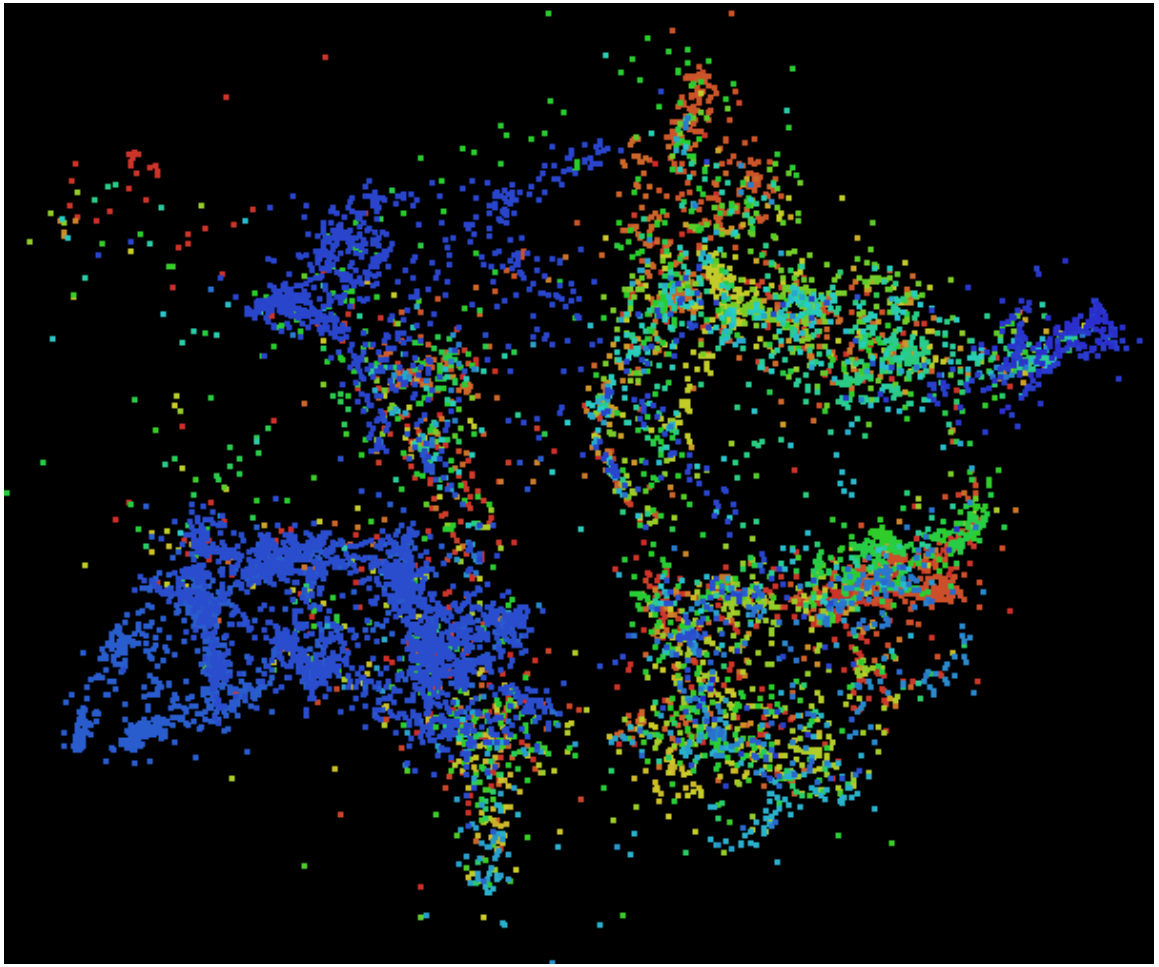
Detailed Inspection



Detailed Inspection



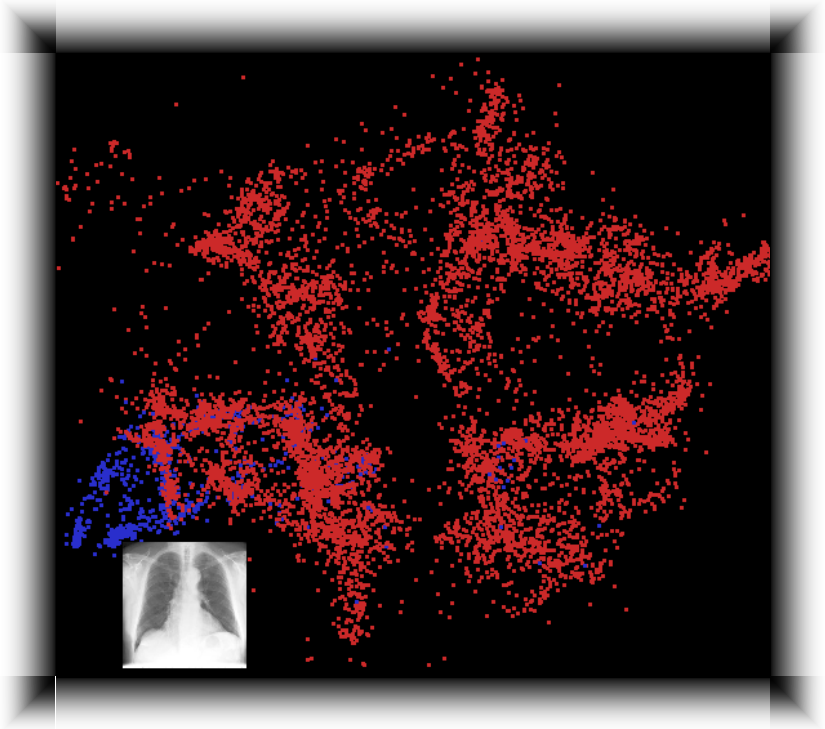
ImageCLEF Training Data Set (1)



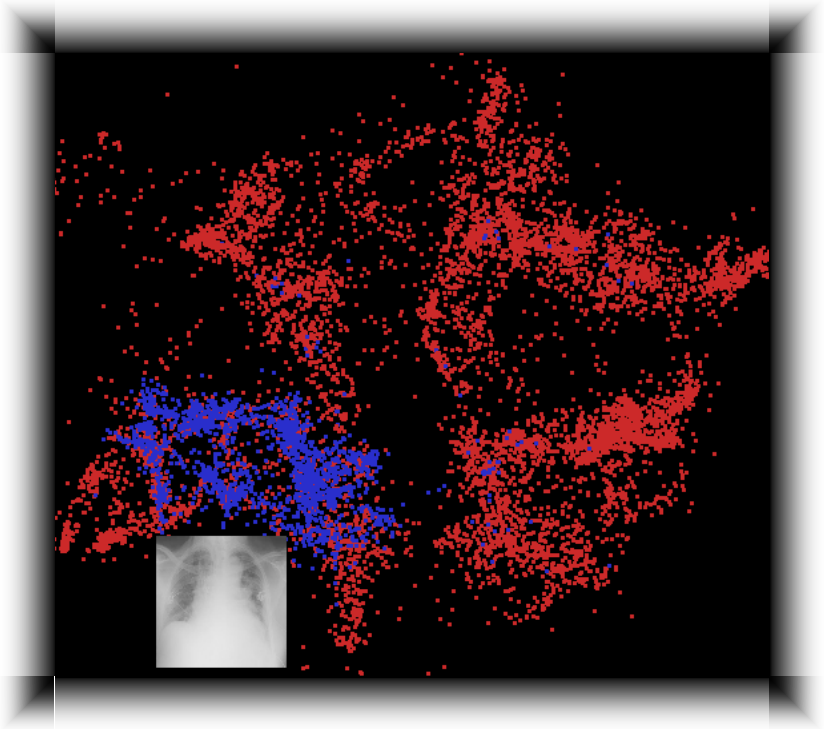
9000 X-Ray images
116 classes

ImageCLEF Training Data Set (2)

Class 108



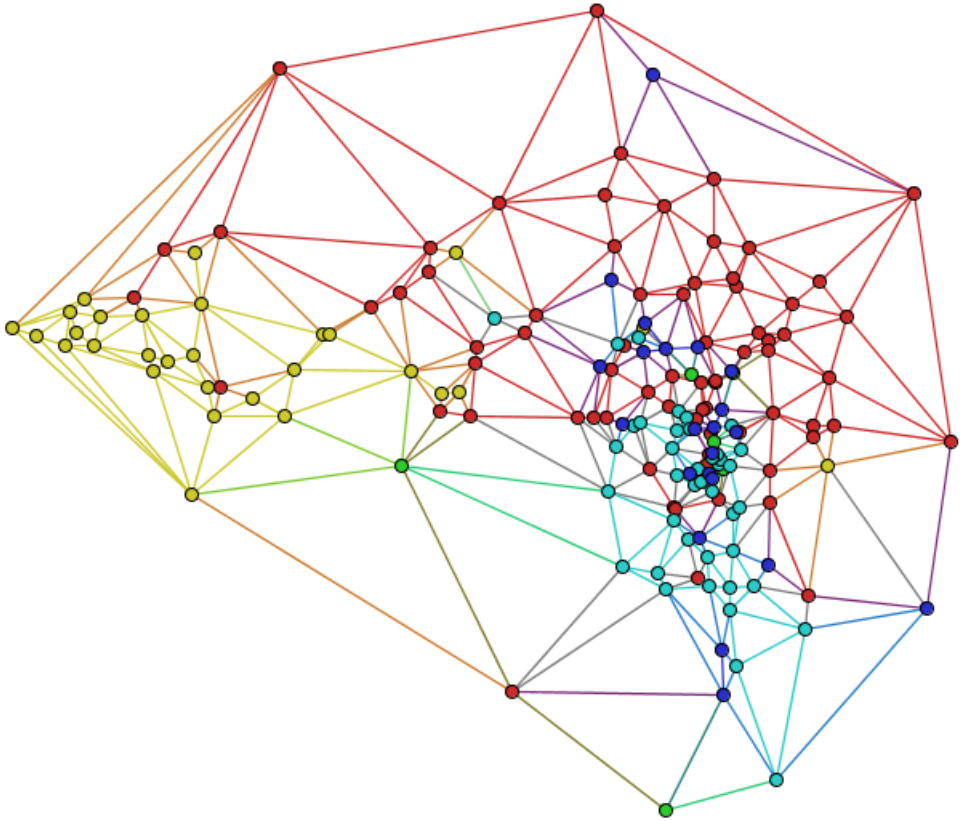
Class 111

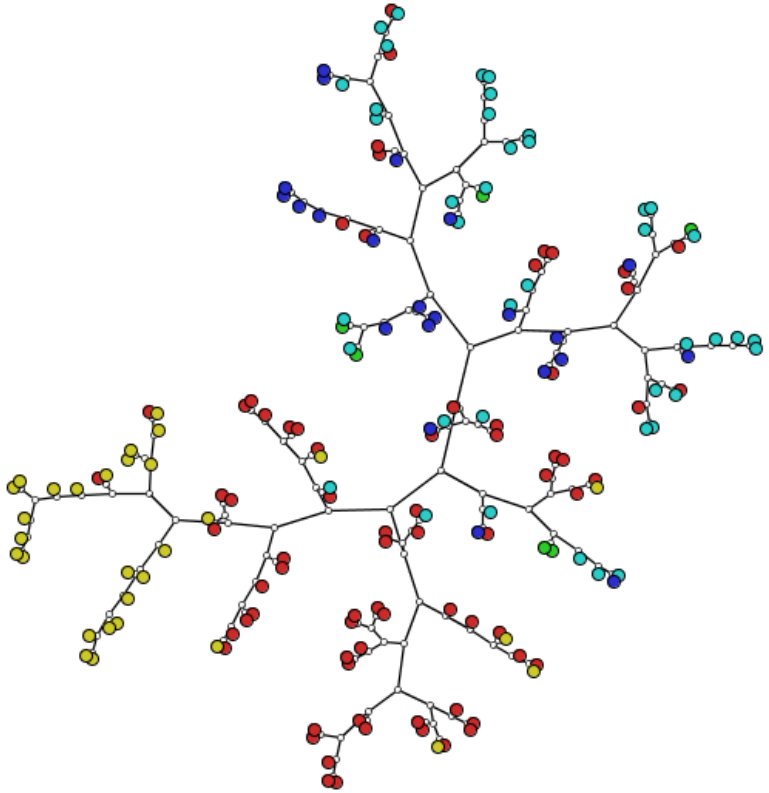


Further Examples on Text

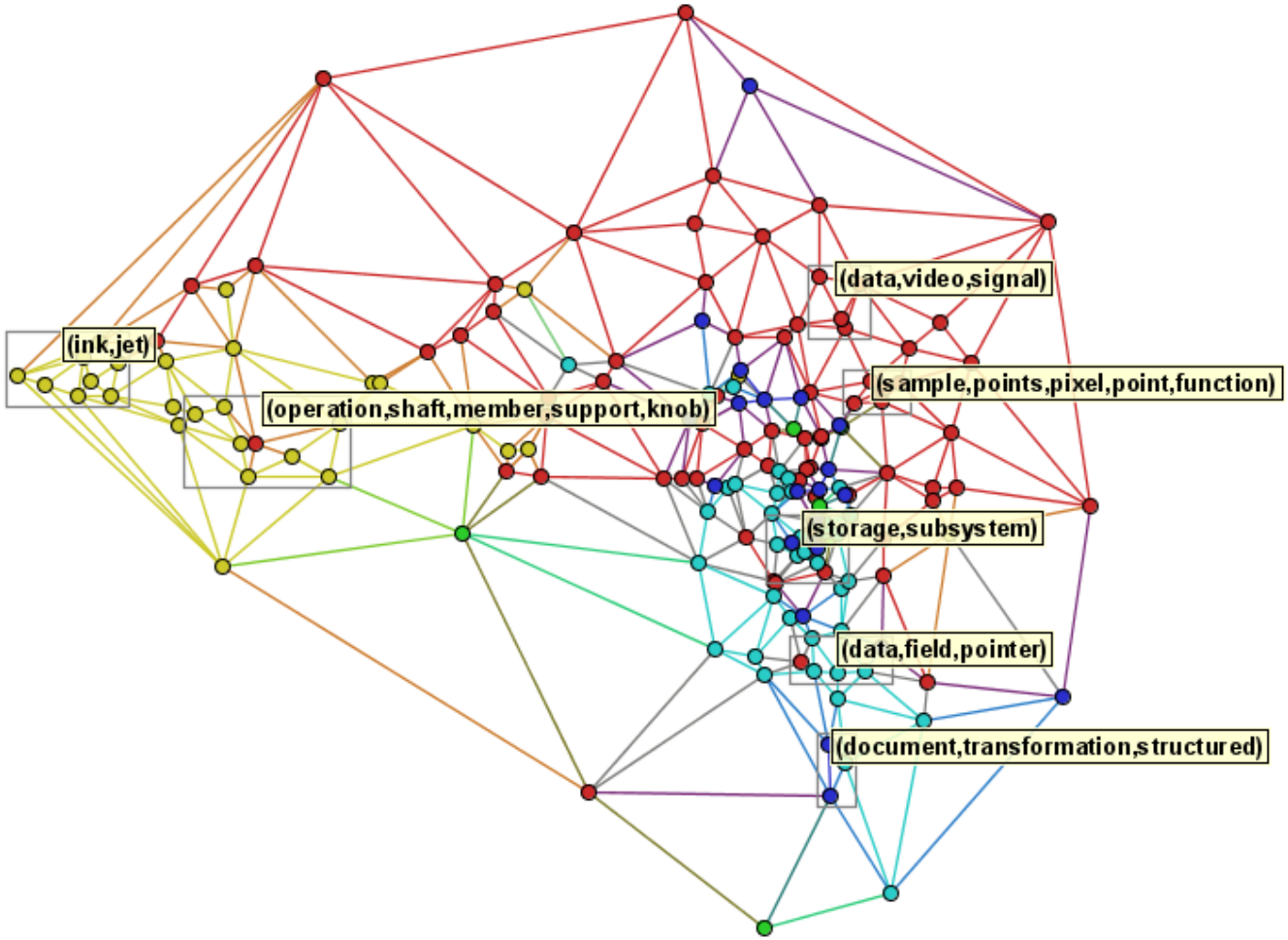
- RSS Patent Data, recovered from the Web http://www.freepatentsonline.com/
- Case 1:
  - 170 files
  - Graphics processing, printer, database, document, ai

Further Examples

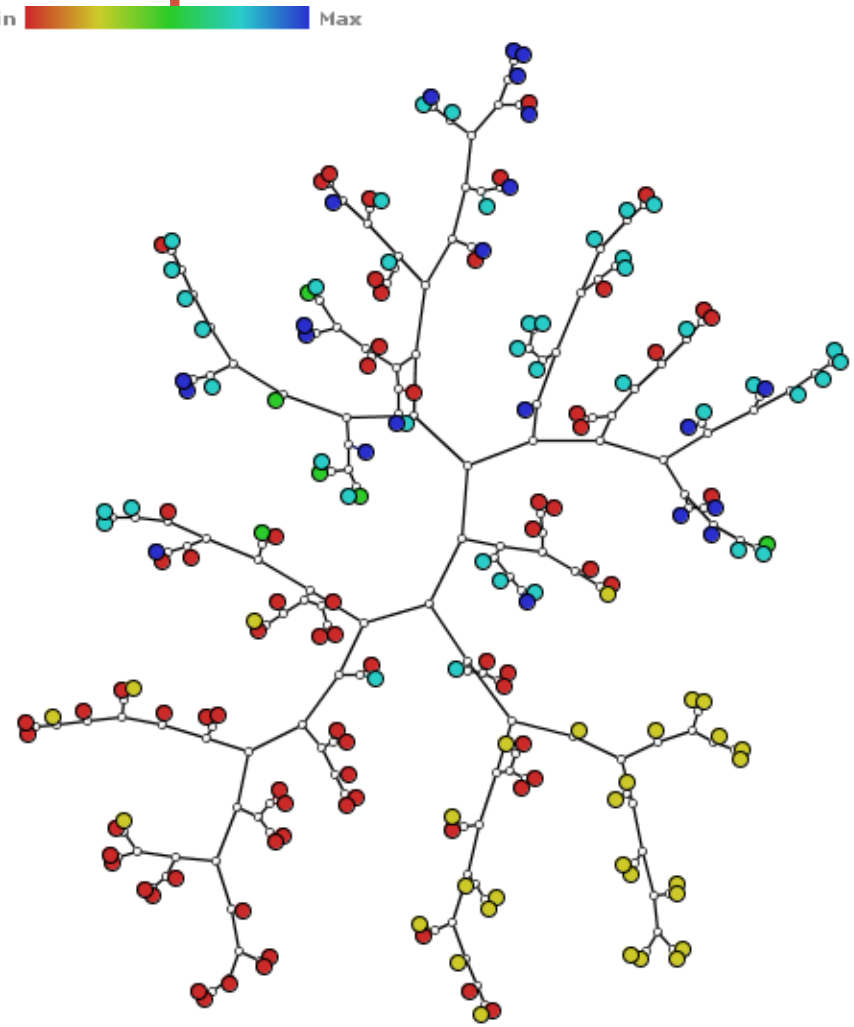


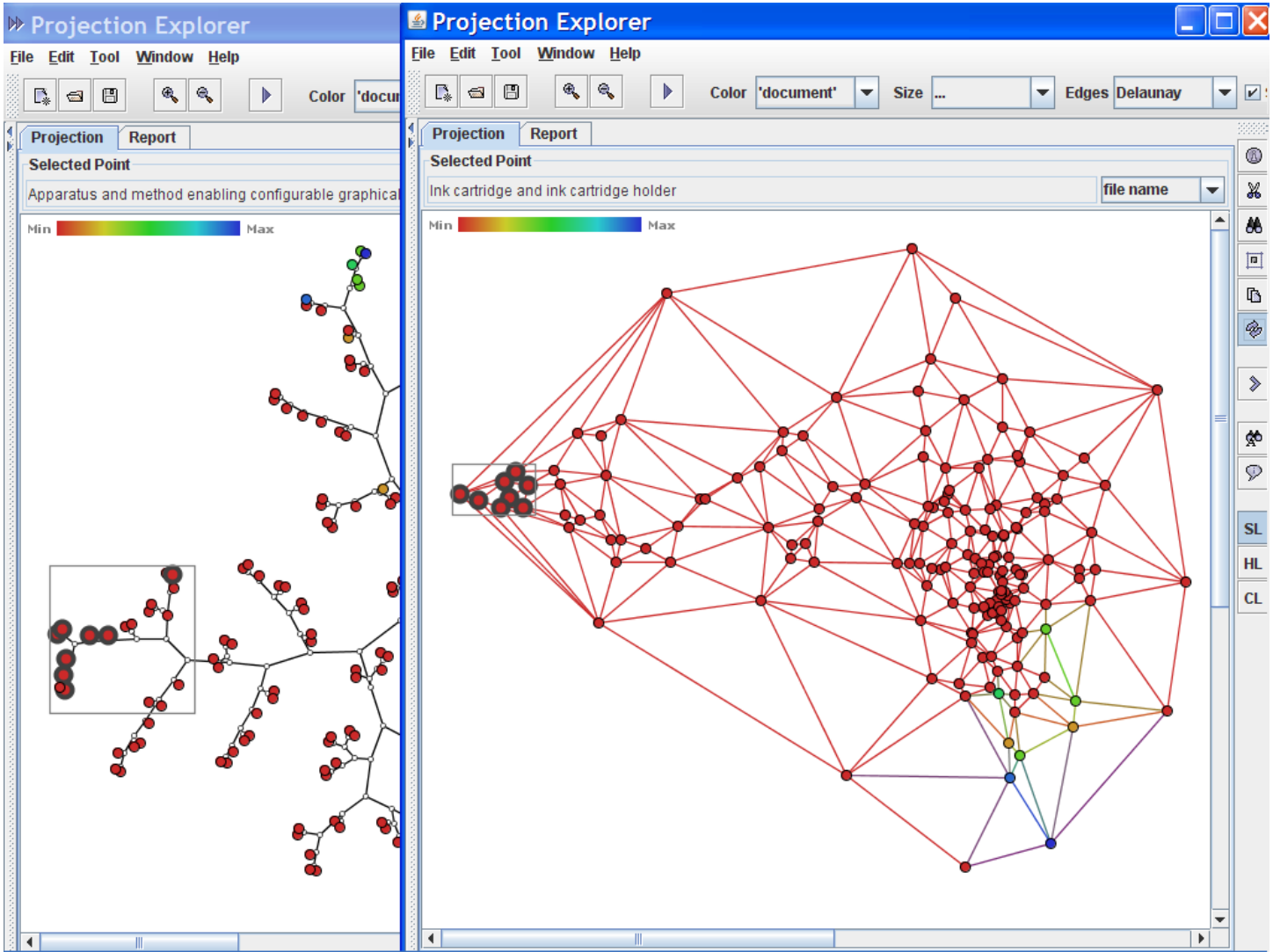


Further Examples

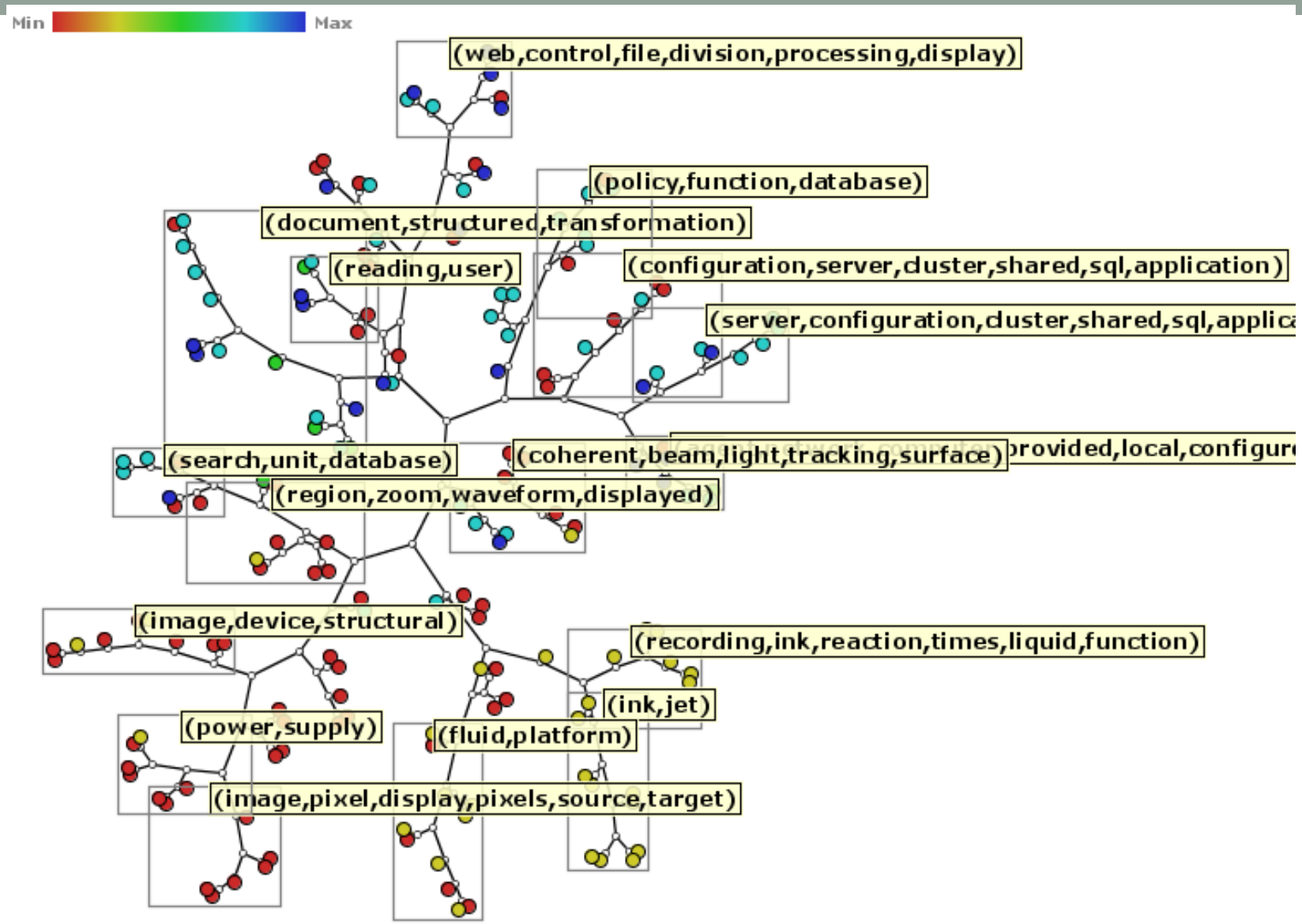


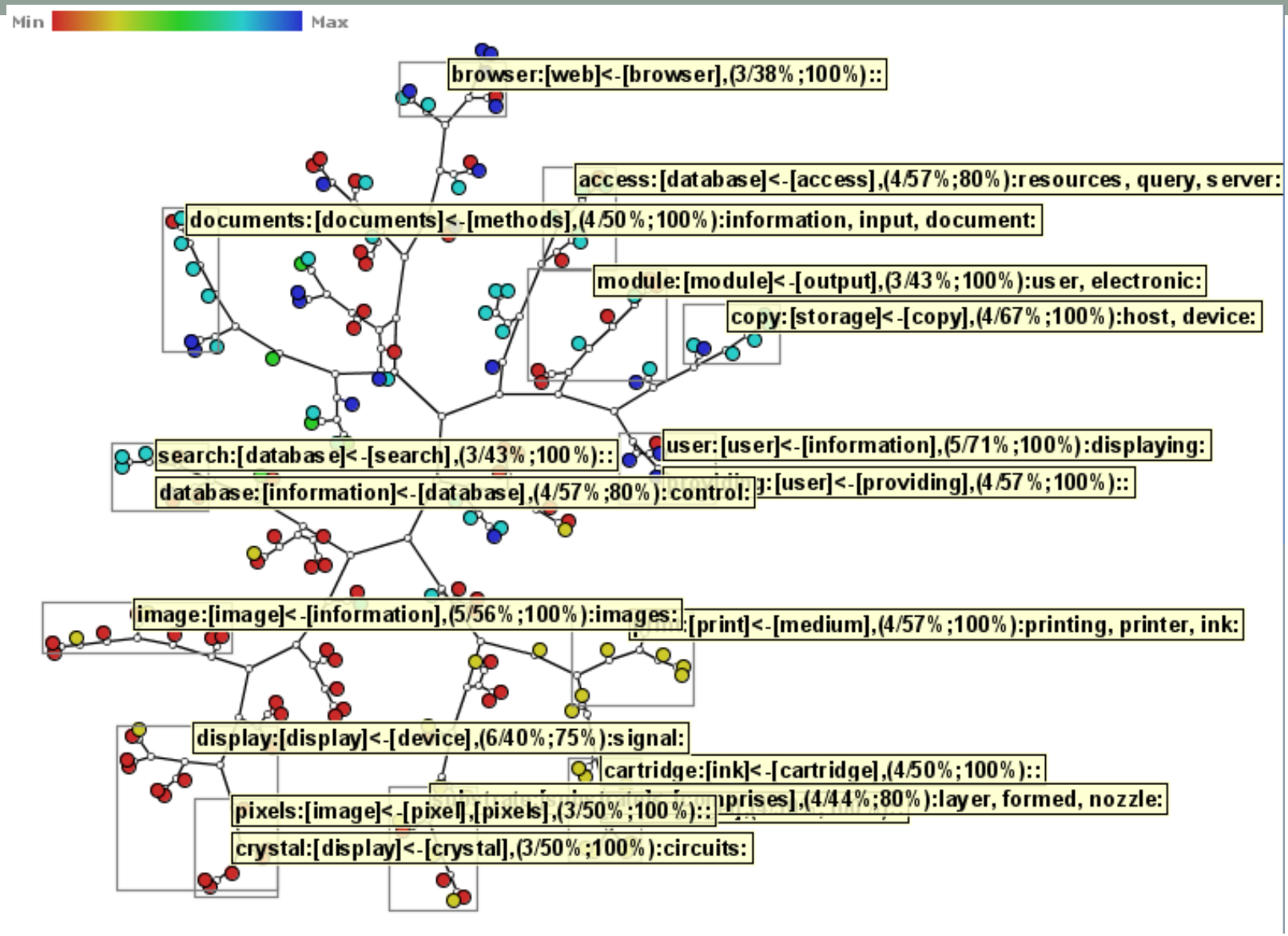
Further Examples





(ink jet, document)

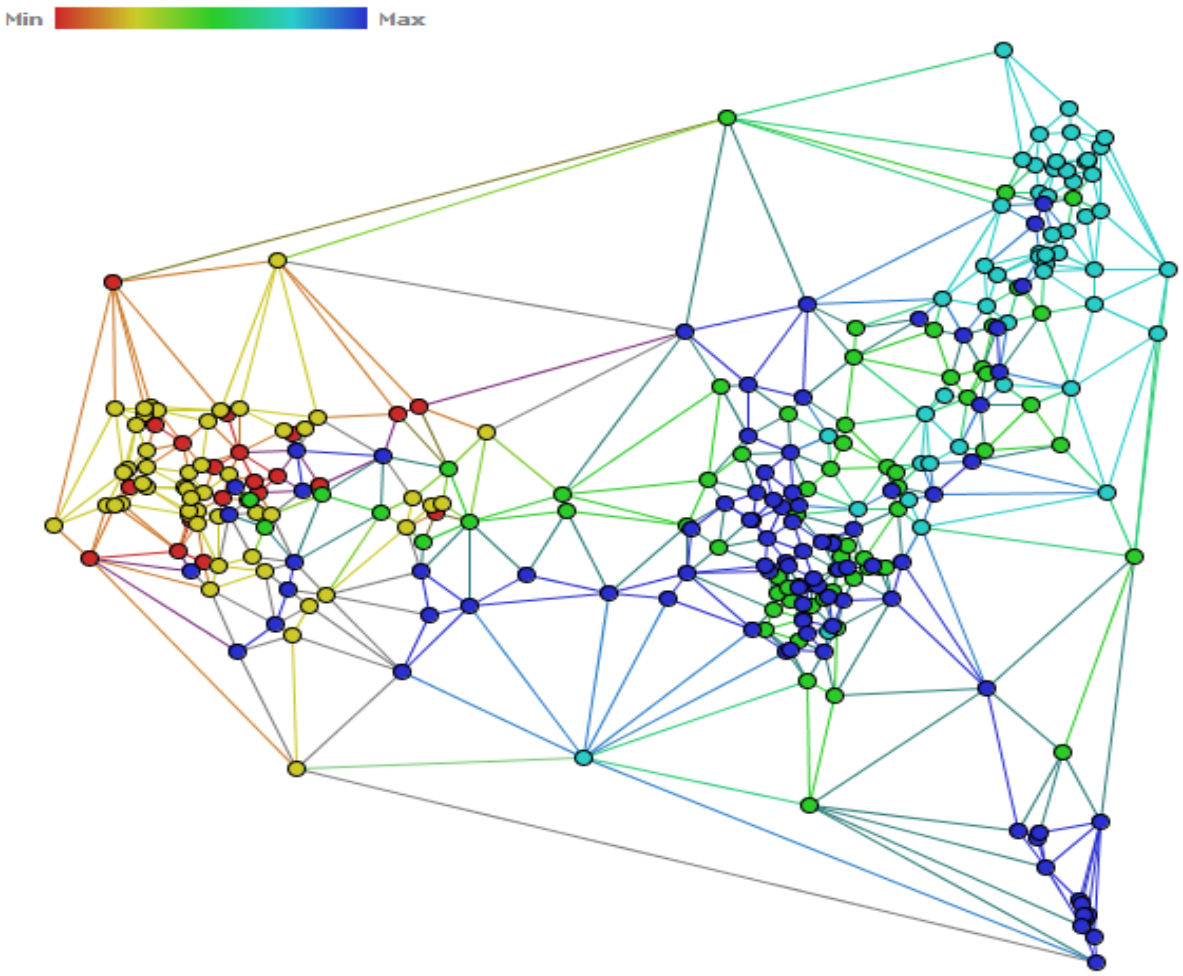

Min
Max
(web,control,file,division,processing,display)
(policy,function,database)
(document,structured,transformation)
(reading,user)
(configuration,server,cluster,shared,sql,application)
(server,configuration,cluster,shared,sql,applica
(search,unit,database)
(coherent,beam,light,tracking,surface)
(region,zoom,waveform,displayed)
(image,device,structural)
(recording,ink,reaction,times,liquid,function)
(ink,jet)
(power,supply)
(fluid,platform)
(image,pixel,display,pixels,source,target)


Min
Max
browser:[web]<-[browser],(3/38%;100%)::
access:[database]<-[access],(4/57%;80%):resources, query, server:
documents:[documents]<-[methods],(4/50%;100%):information, input, document:
module:[module]<-[output],(3/43%;100%):user, electronic:
copy:[storage]<-[copy],(4/67%;100%):host, device:
search:[database]<-[search],(3/43%;100%)::
user:[user]<-[information],(5/71%;100%):displaying:
database:[information]<-[database],(4/57%;80%):control:
image:[image]<-[information],(5/56%;100%):images:
[print]<-[medium],(4/57%;100%):printing, printer, ink:
display:[display]<-[device],(6/40%;75%):signal:
cartridge:[ink]<-[cartridge],(4/50%;100%)::
pixels:[image]<-[pixel],[pixels],(3/50%;100%)::
crystal:[display]<-[crystal],(3/50%;100%):circuits:

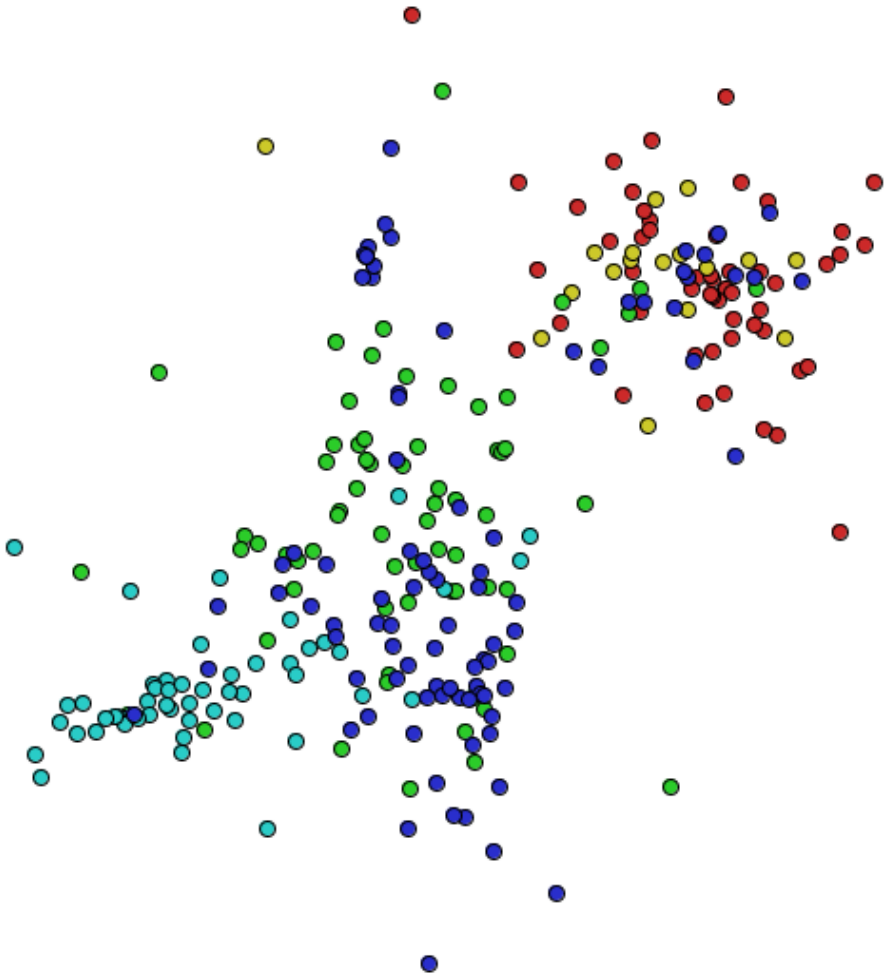
Patents – case 2

- http://www.freepatentsonline.com/
- 172 files
- surgery (2), drugs(2), molecular biology

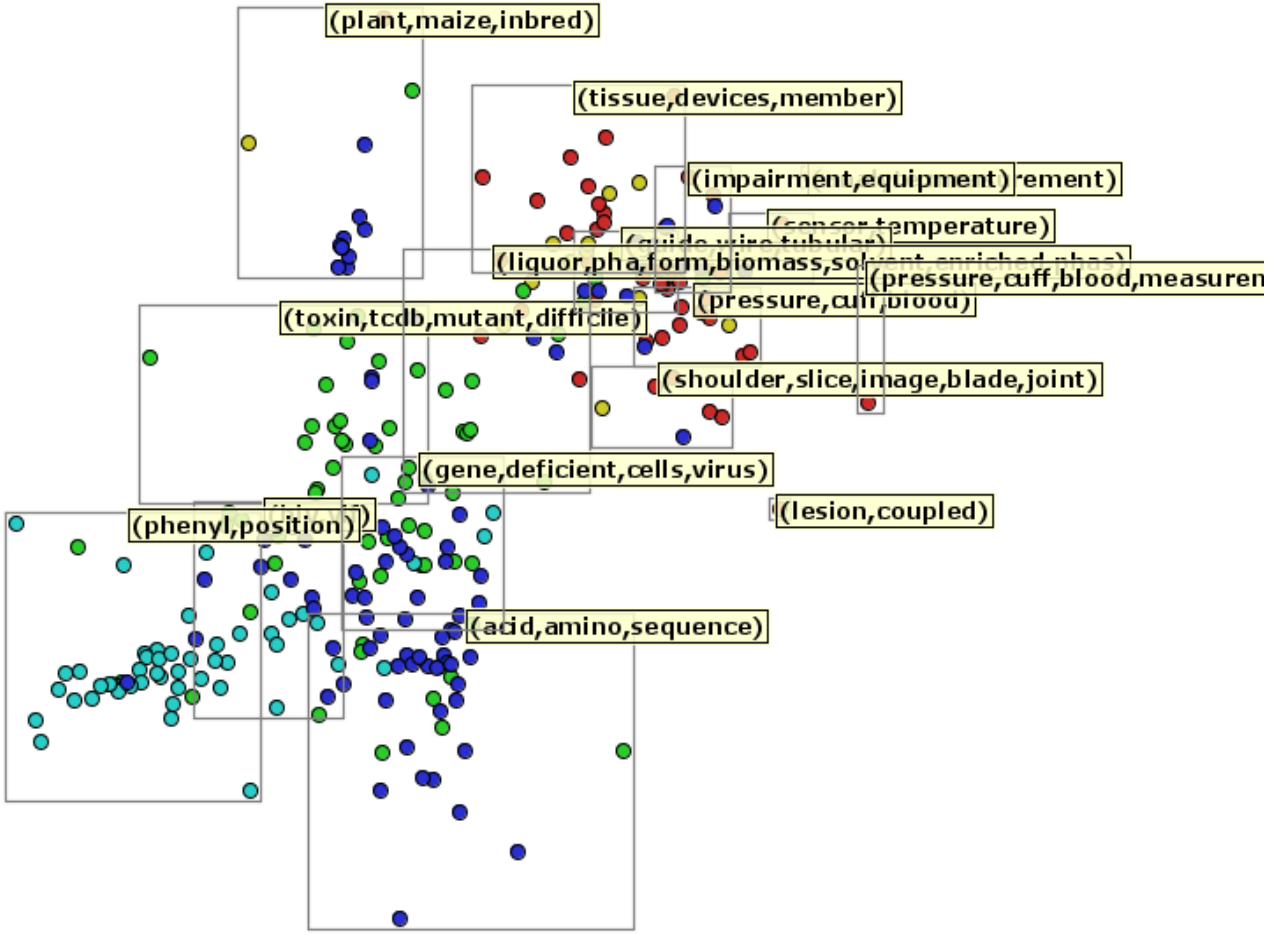
Patents surgery, drugs, molecular bio



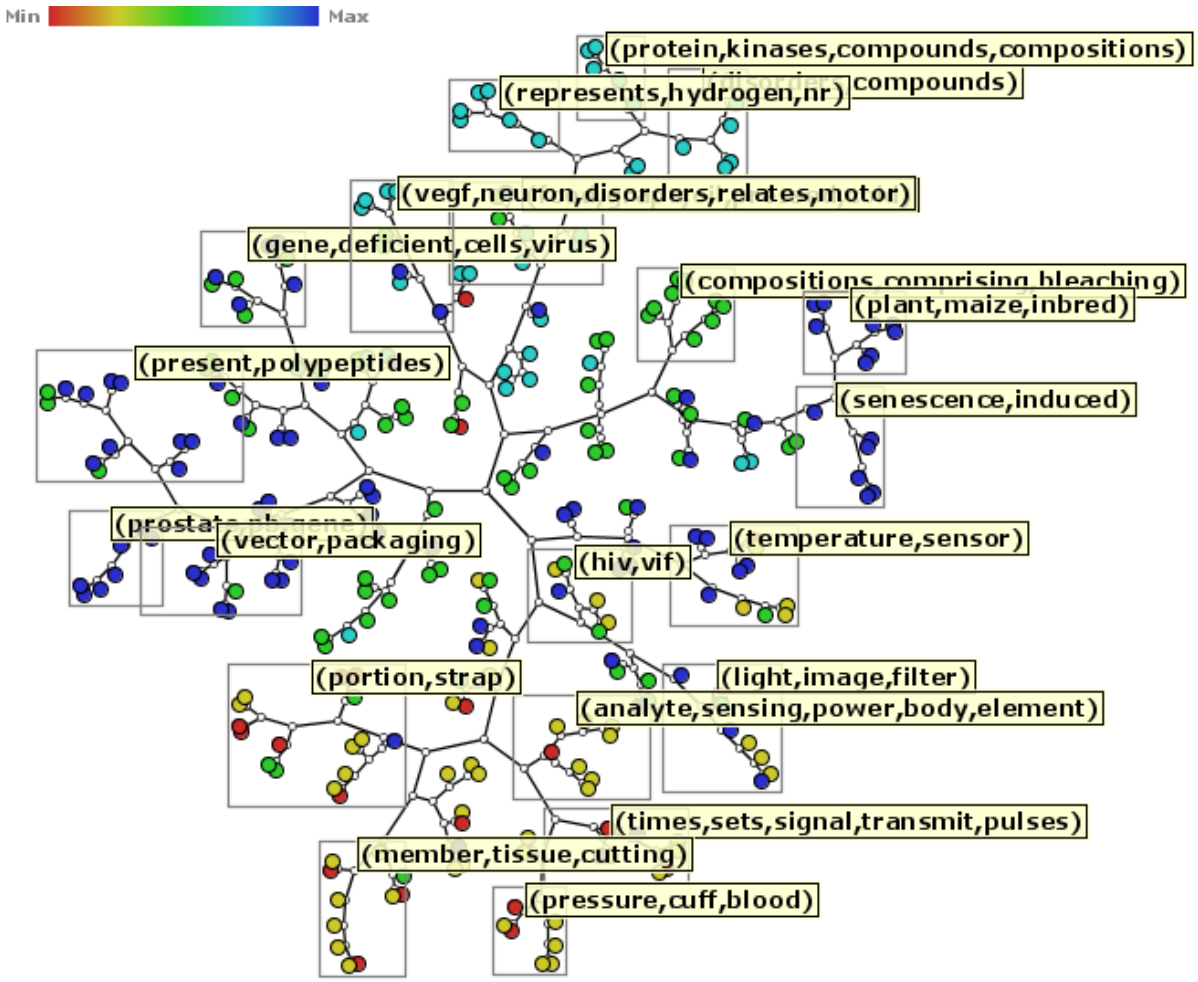
Patents surgery, drugs, molecular bio stopwords selection



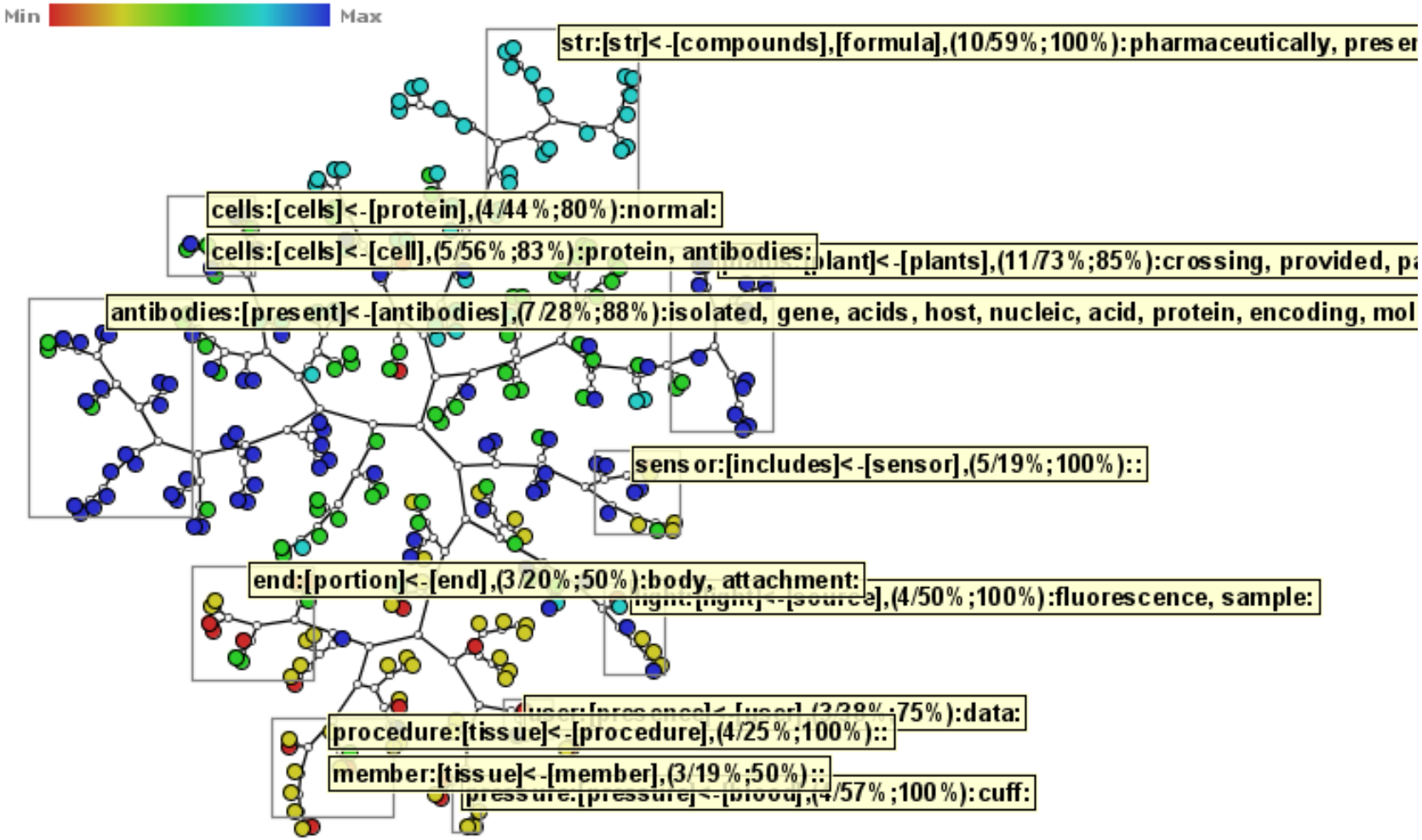
Patents surgery, drugs, molecular bio topics



Patents surgery, drugs, molecular bio



Patents surgery, drugs, molecular bio



More Techniques

- Projection-based
  - PCA
  - MDS
  - Sammon Mapping
  - LSP - like

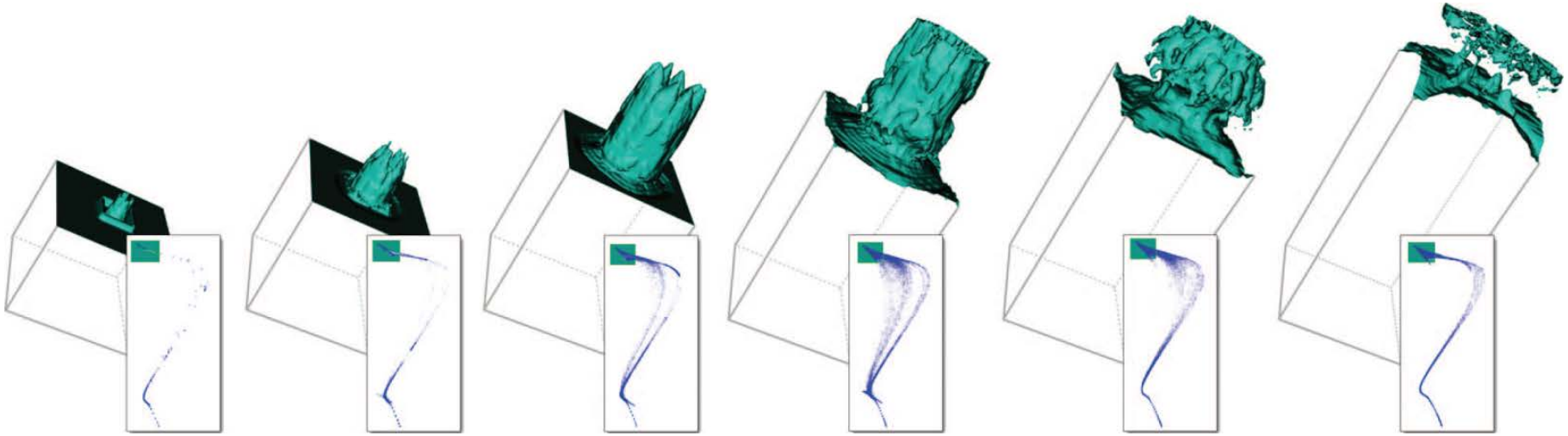
- Glimmer (distance)
  - Stephen Ingram, Tamara Munzner, Marc Olano: Glimmer: Multilevel MDS on the GPU. IEEE Trans. Vis. Comput. Graph. 15(2): 249-261 (2009)

- T-sne (segregation)
  - 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.
  - http://lvdmaaten.github.io/tsne/

PLMP

- Paulovich, Silva, Nonato, Two-Phase Mapping for Projecting Massive Data Sets, ***IEEE Trans. Visualization and Computer Graphics****, 2010*

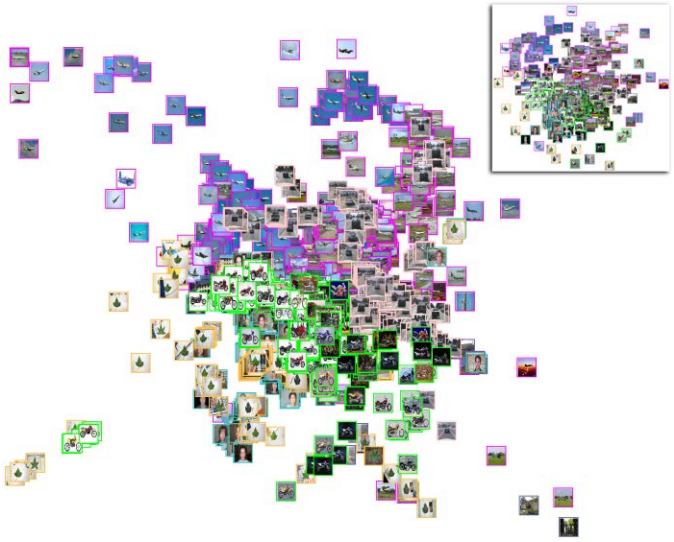
- spatially embedded data, more samples than dimensions
- millions of data items
- time varying and streaming data
- reduced amount of distance information

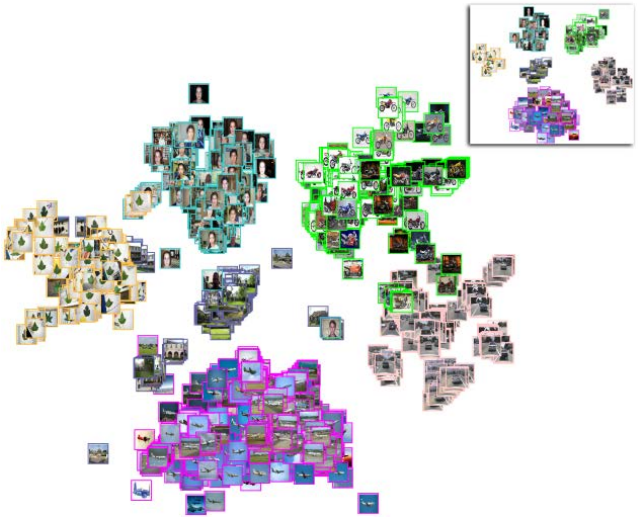


PLP

- Paulovich, Eler, Poco, Botha, Minghim, Nonato, Piecewise Laplacian-based Projection for Interactive Data Exploration and Organization, ***Computer Graphics Forum*** *2011*

- local control
- flexibility in handling user interaction: users may change the layout based on previous knowledge/perception of similarity

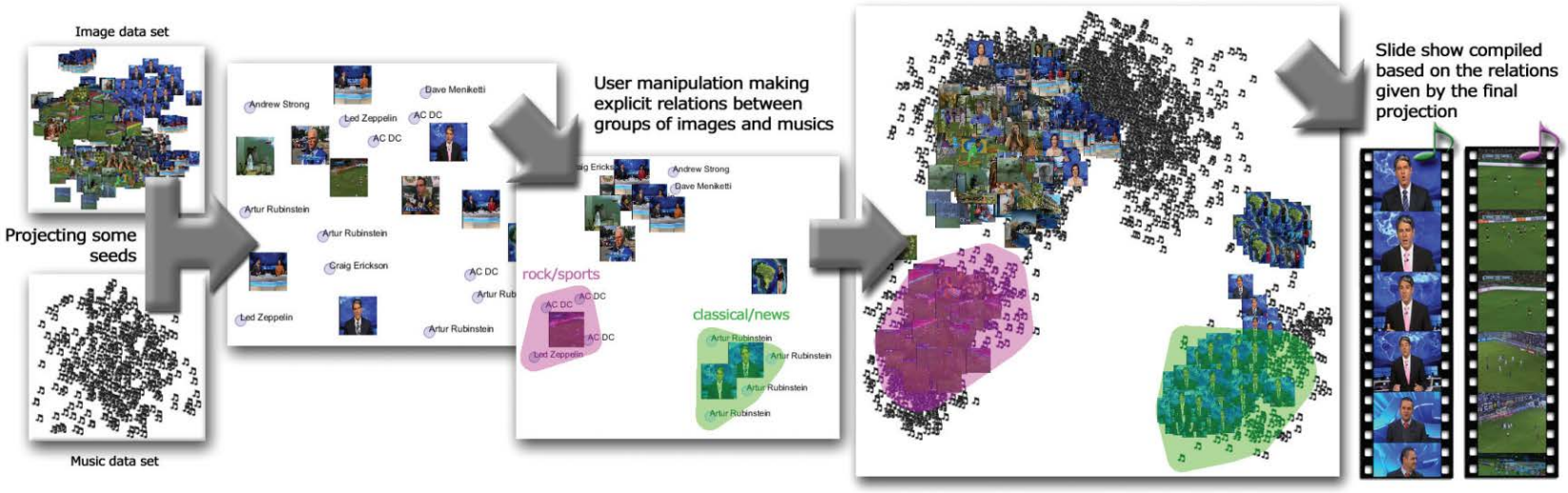




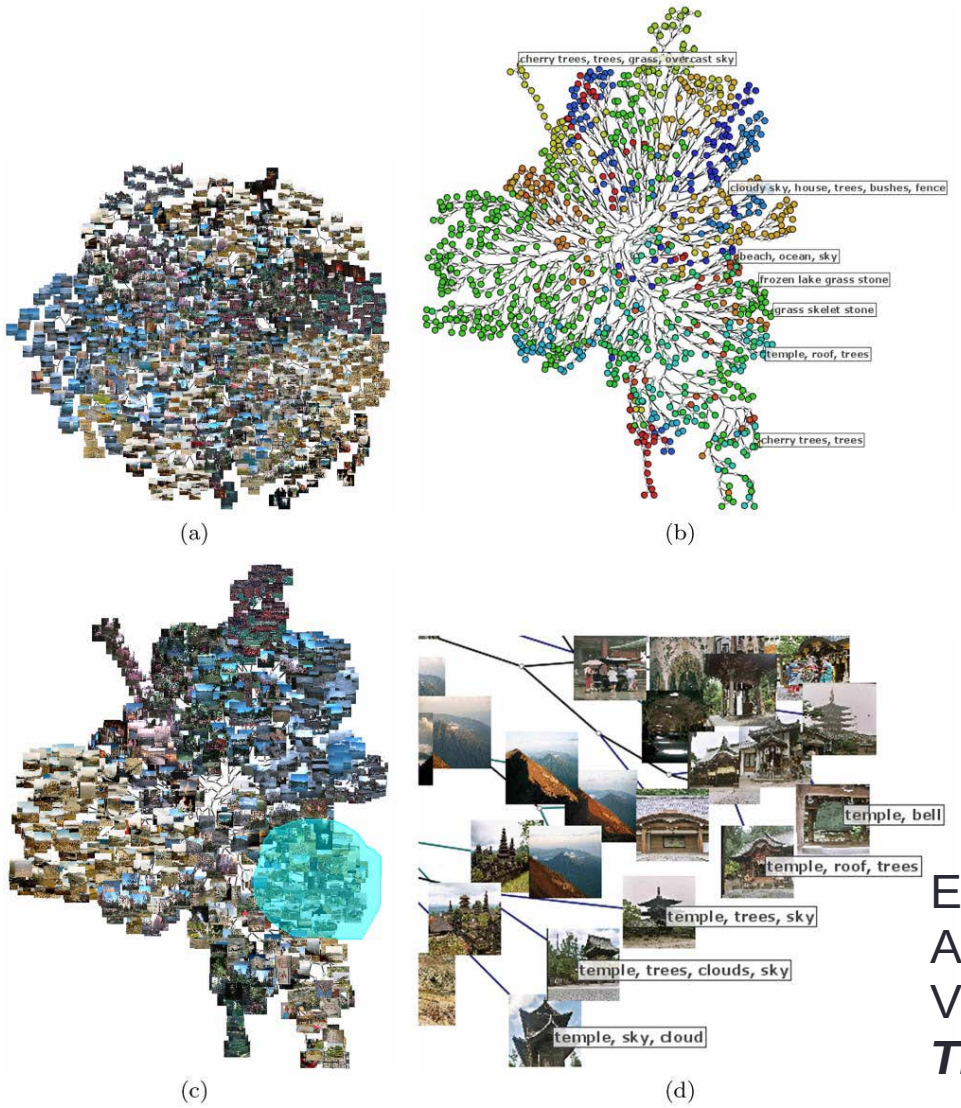
LAMP

- Joia; Paulovich, Coimbra, Cuminato, Nonato, Local Affine Multidimensional Projection, ***IEEE Trans. Visualization and Computer Graphics*** *2011*

- orthogonal mapping theory
- global and local control
- ability to correlate data from unconnected data sets
- cost effective and highly precise



Applications

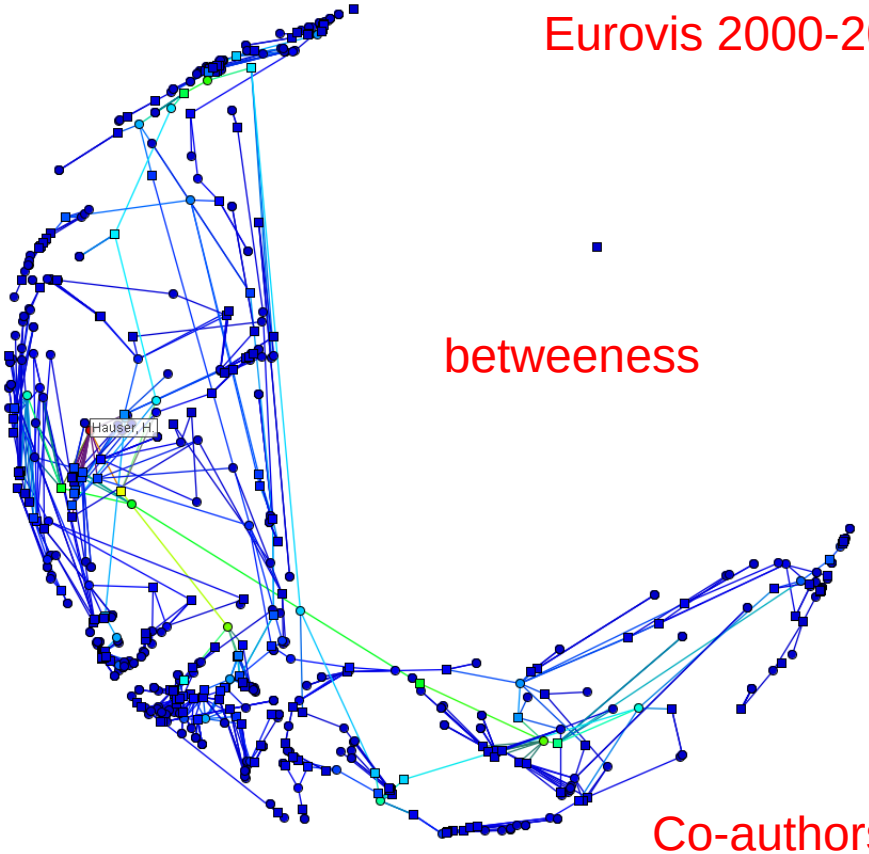


Exploratory visualization of

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

Eler, Nakazaki, Paulovich, Santos, Andery, Oliveira, Batista Neto, Minghim, Visual analysis of image collections ***The Visual Computer***, 2009

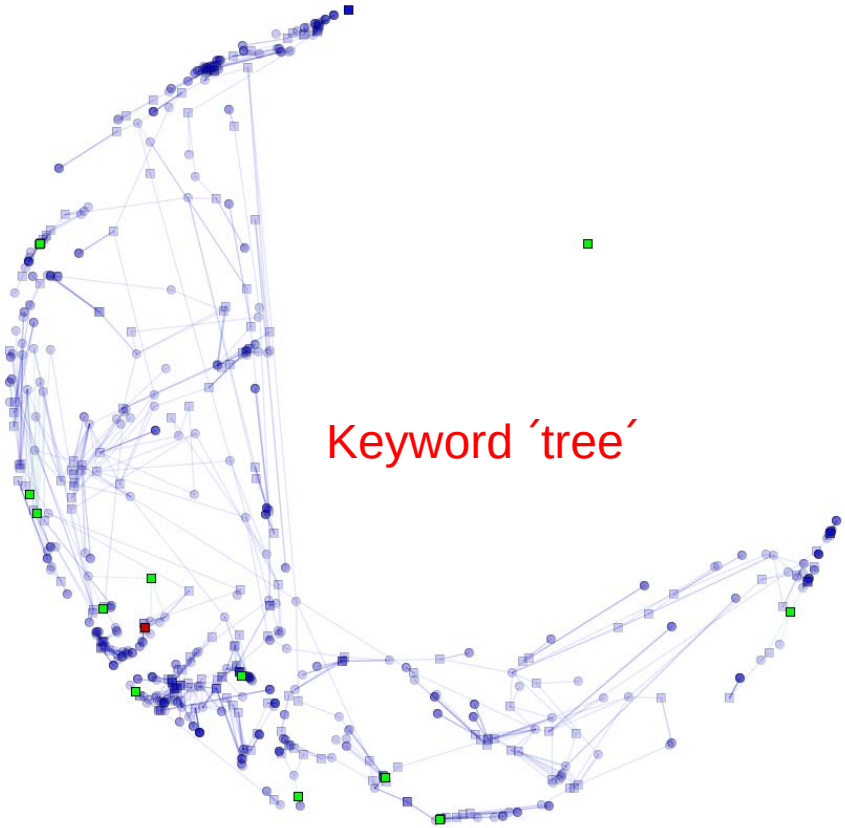
Social Networks



Eurovis 2000-2010

betweeness

Co-authorship



Keyword ´tree´

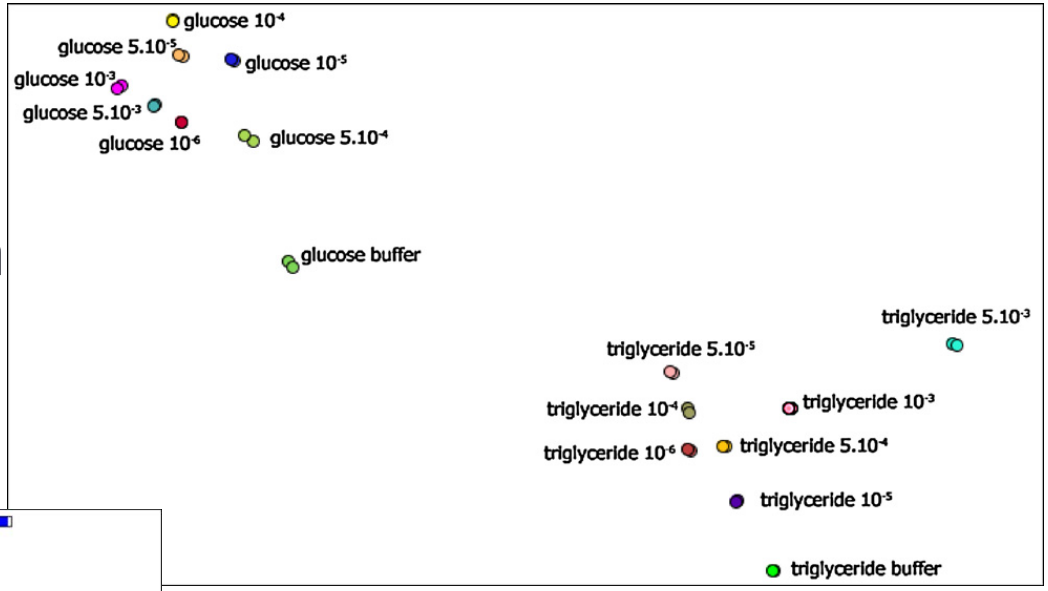
Martins, Andery, Heberle, Lopes, Pedrini, Minghim; Multidimensional Projections for Visual Analysis of Social Networks (to appear), **JCST** 2012

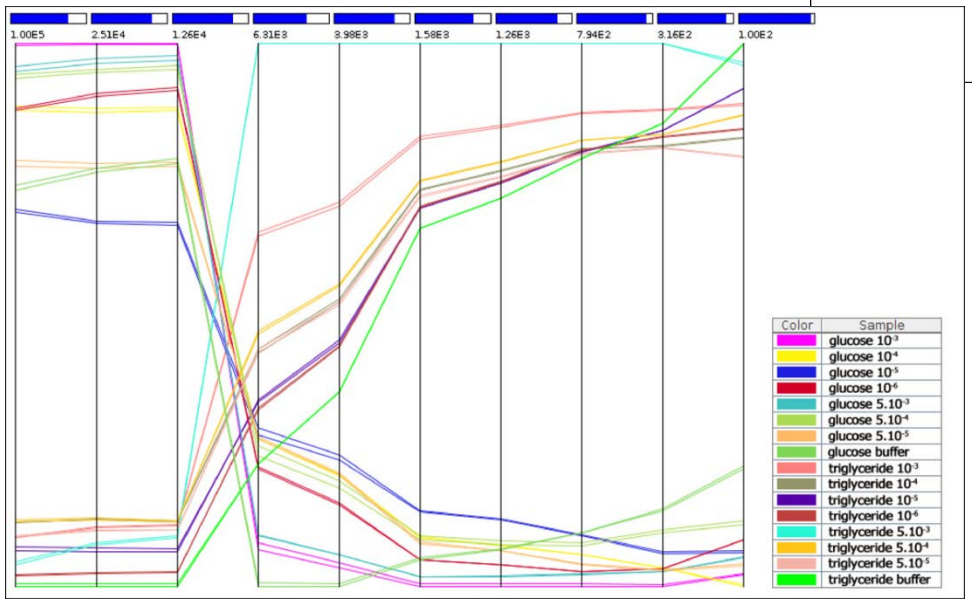
Data from nanotech sensors & biosensors

- Volpati et. al, Toward the optimization of an e-tongue system using information visualization: a case study with perylene tetracarboxylic derivative films in the sensing units, ***Langmuir***, 2012
- Paulovich et al., Information visualization techniques for sensing and biosensing, ***Analyst***, 2011
- Paulovich et al., Using multidimensional projection techniques for reaching a high distinguishing ability in biosensing. ***Analytical and Bioanalytical Chemistry***, 2011
- Siqueira Jr. et al., Strategies to optimize biosensors based on impedance spectroscopy to detect phytic acid using layer-by-layer films, ***Analytical Chemistry***, 2010
- Perinoto et al., Biosensors for efficient diagnosis of leishmaniasis: innovations in bioanalytics for a neglected disease, ***Analytical Chemistry***, 2010

Data from nanotech sensors & biosensors

finding good sensor configurations: segregation tasks on data





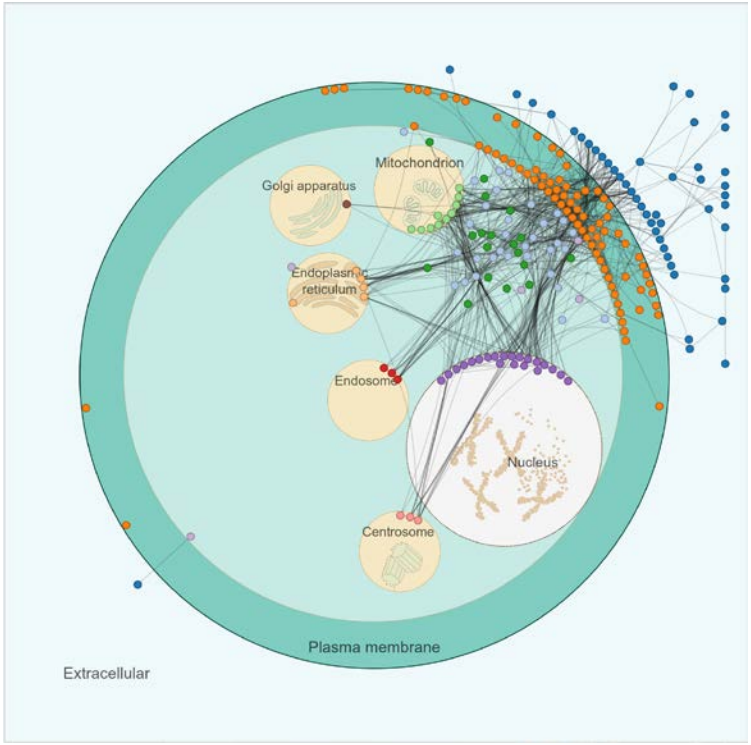
Moraes et. al, Detection of glucose and triglycerides using information visualization methods to process impedance spectroscopy data, ***Sensors & Actuators B***, 2012

On studies on ecology and environment

- D.Sc. project: Visual exploration of feature spaces to support green algae taxonomic classification
- Classification based on features from imagens & other sources
- Collaboration with Dr. Armando Vieira, Department of Biology, UFSCar
- Time-varying images, feature extraction, representation and analysis

Example: Proteomics and Cancer

- PhD project: Protein networks by force based constrained to geometric forms

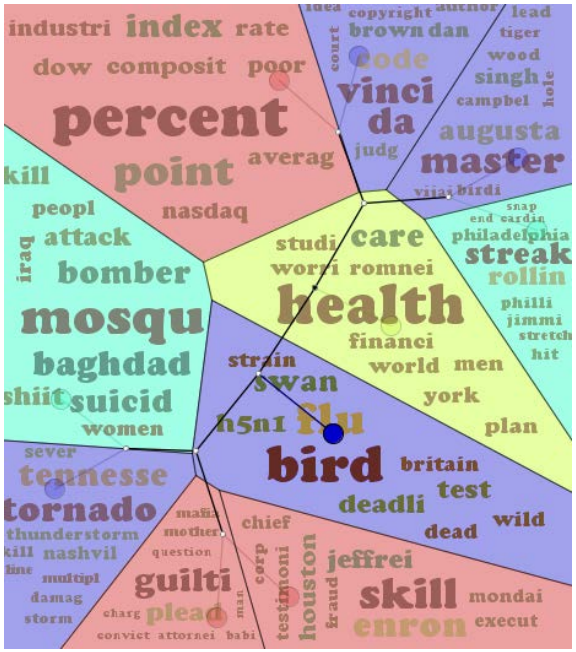


Heberle, Henry ; Carazzolle, Marcelo Falsarella ; Telles, Guilherme P. ; Meirelles, Gabriela Vaz; Minghim**,** Rosane . CellNetVis: a web tool for visualization of biological networks using force-directed layout constrained by cellular components. BMC BIOINFORMATICS, v. 18, p. 395, 2017.

Kawahara, R., Meirelles, G., Heberle, H., Domingues, R., Granato, D., Yokoo, S., Canevarolo, R., Winck, F., Ribeiro, A. C., Brand~ao, T. B., Filgueiras, P., Cruz, K., Barbuto, J. A., Poppi, R., Minghim, R., Telles, G., Fonseca, F. P., Fox, J., Santos-Silva, A., Coletta, R., Sherman, N., and Leme, A. P. Integrative analysis to select cancer candidate biomarkers to targeted validation. **Oncotarget** 6, 41 (2015), 43635-43652.

More Data – Summarization

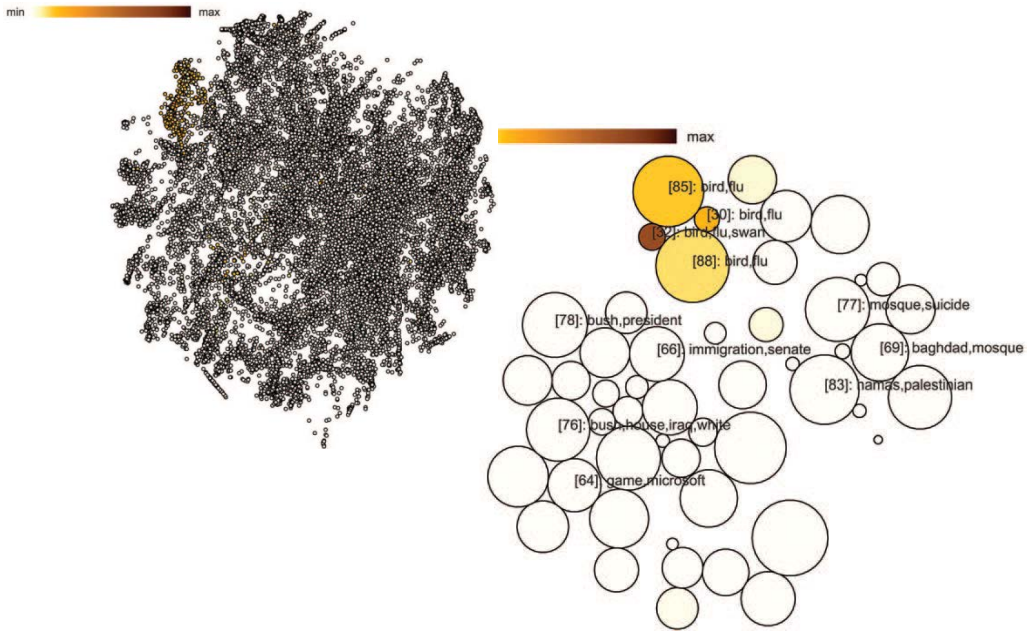
- Wordclouds
- Representative Images



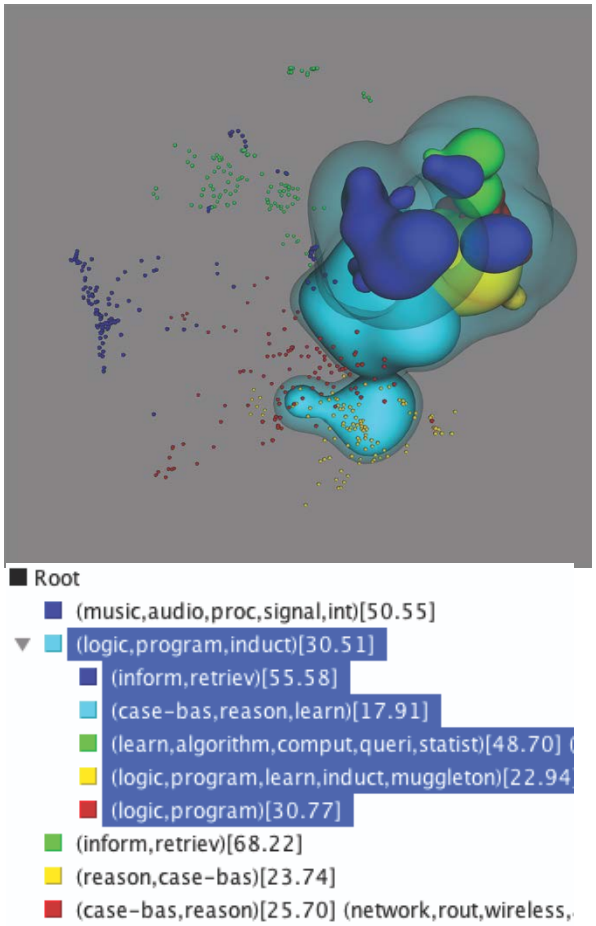
Multi-level text

Multi-level images

Metaphors: clutter

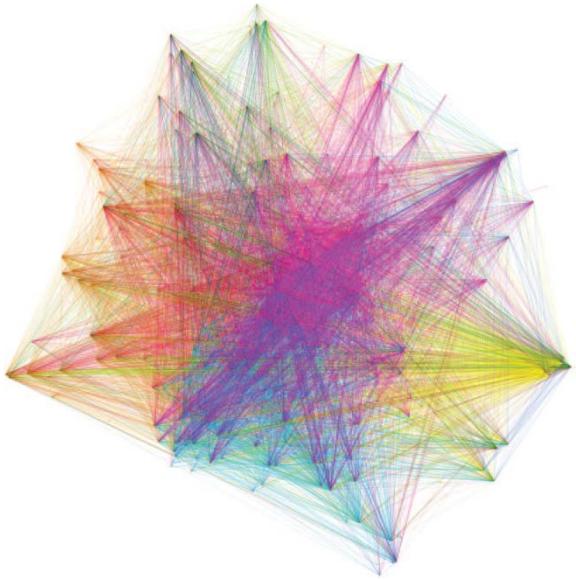


Paulovich and Minghim, HiPP: a novel hierarchical point placement strategy and its application to the exploration of document collections, ***IEEE Trans. Visualization & Computer Graphics***, *2008*

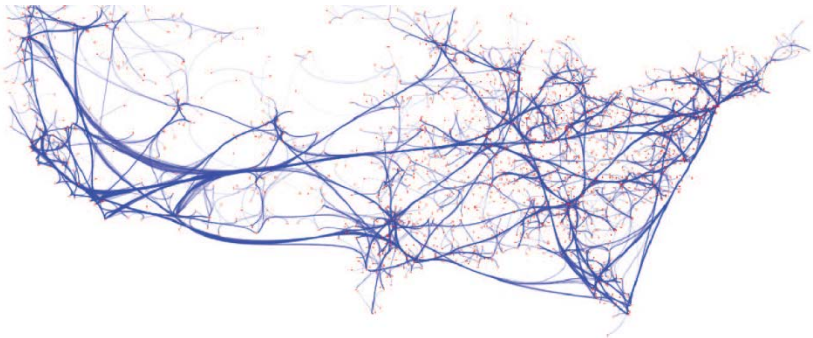


Poco; Etedmapour, Paulovich, Long, Rosenthal, Oliveira, Linsen, Minghim. A framework for exploring multidimensional data with 3D projections, ***Computer Graphics Forum***, Eurovis 2011.

Metaphors: clutter

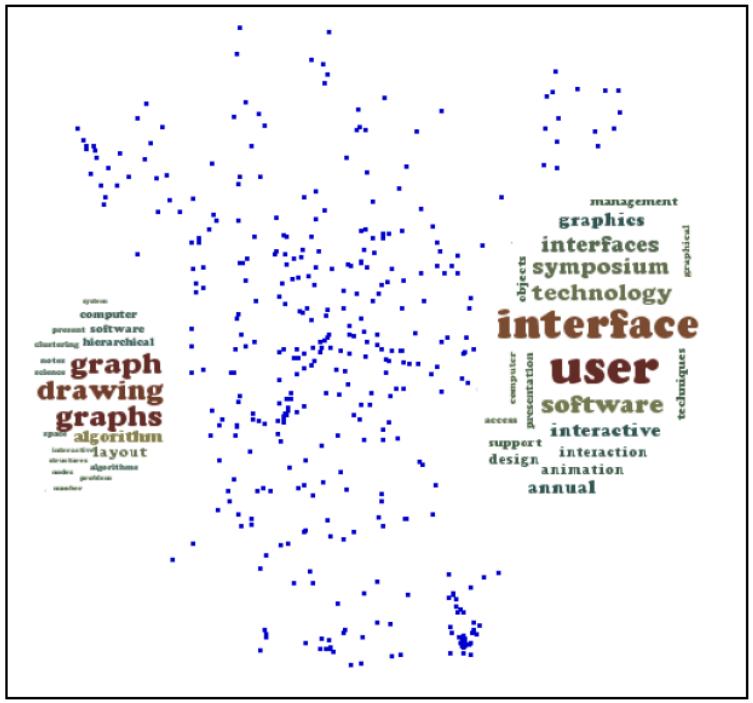


Ersoy, Hurter, Paulovich, Cantareira, Telea, Skeleton-based edge bundling for graph visualization. ***IEEE Trans. Visualization and Computer Graphics,*** Infovis 2011



More Applications – Word clouds

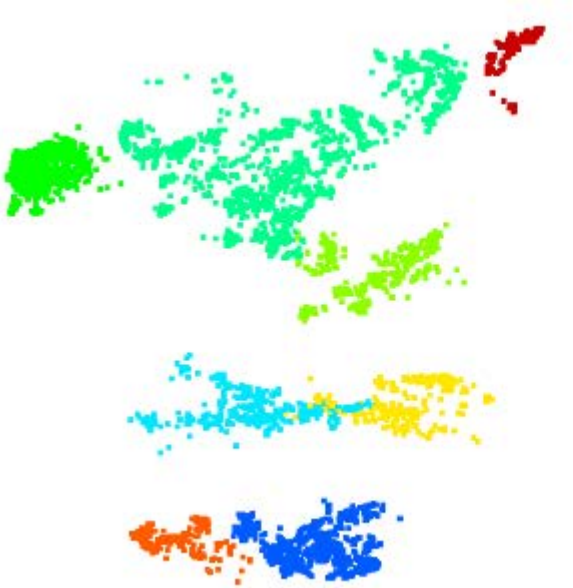
- Semantical ordering of keywords from projected points (new)
- Semantical filling of polygons over projections (new)



Paulovich, Telles, Toledo, Minghim, Nonato – Semantic Wordification of Document Collections, **Computer Graphics Forum, Eurovis 2012**

More Applications – Fiber Tracking

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

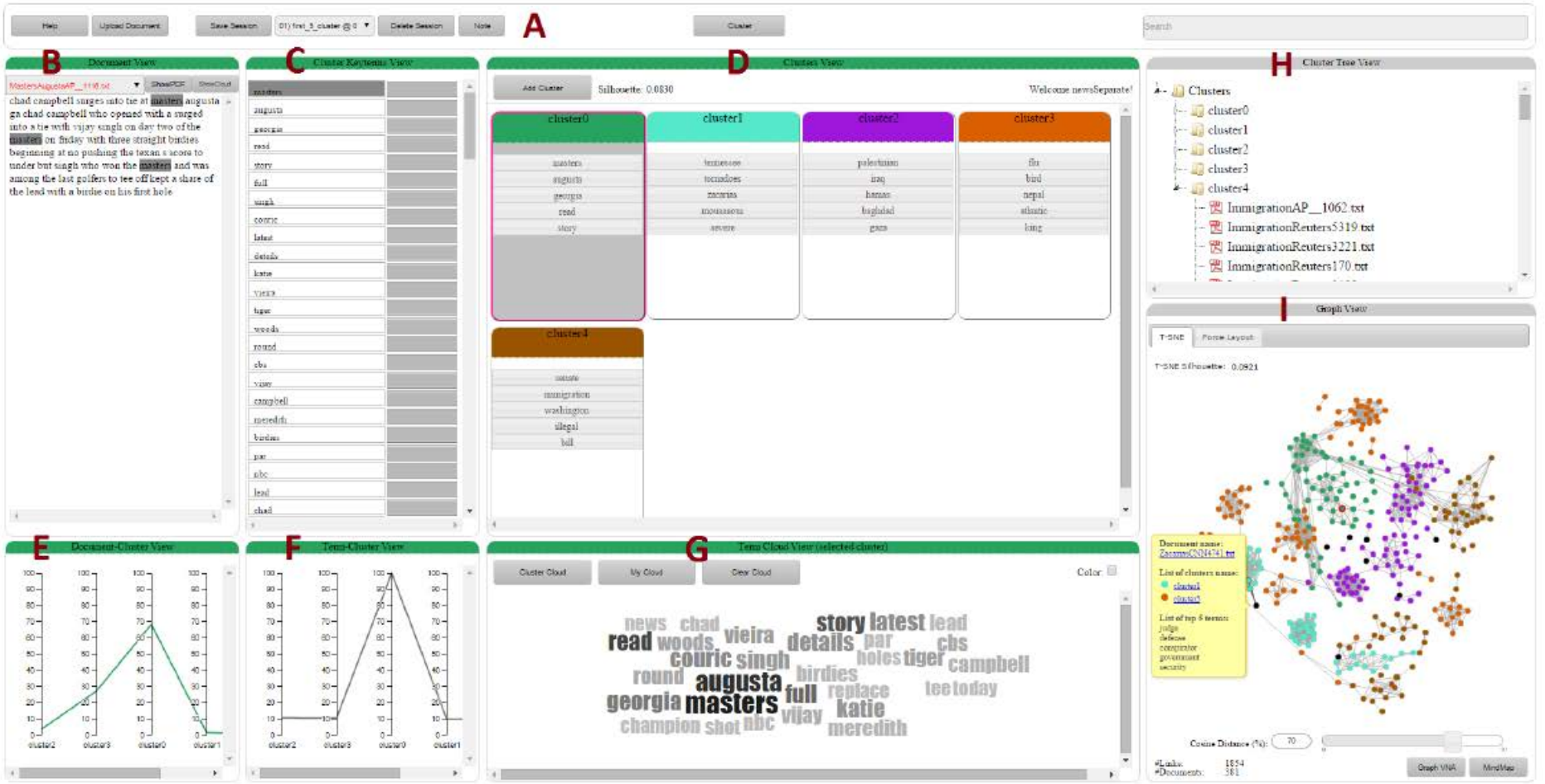


Poco, Eler, Paulovich, Minghim - Employing 2D projections for fast visual exploration of large fiber tracking data, **Computer Graphics Forum**, **Eurovis** 2012.

Open problems

- Metaphors: user interface, scalability, user control...
  - Handling text
  - Handling time-varying data
  - Going small: portable devices
  - Growing large: scalability issues
- Evaluation: user perception, quantitative & qualitative metrics
- Applications, reaching out to users: understanding their needs & tuning to specific profiles and application domains

Visually supported keyterm-based clustering

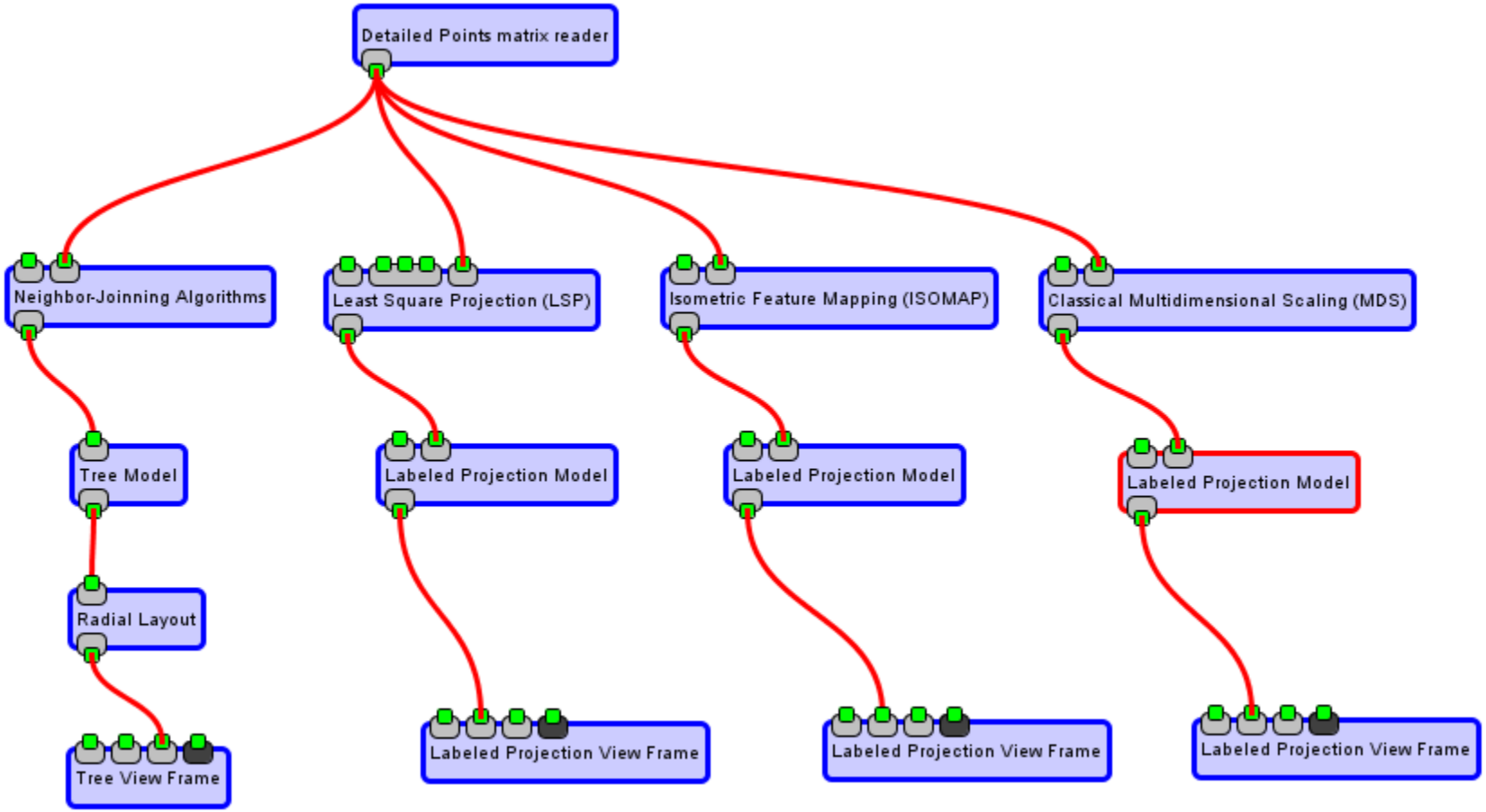


With Dalhousie University, Canada

Visual Analysis of Microblog Data (Rumour)

With Dalhousie University, Canada

Before we continue... Vispipeline



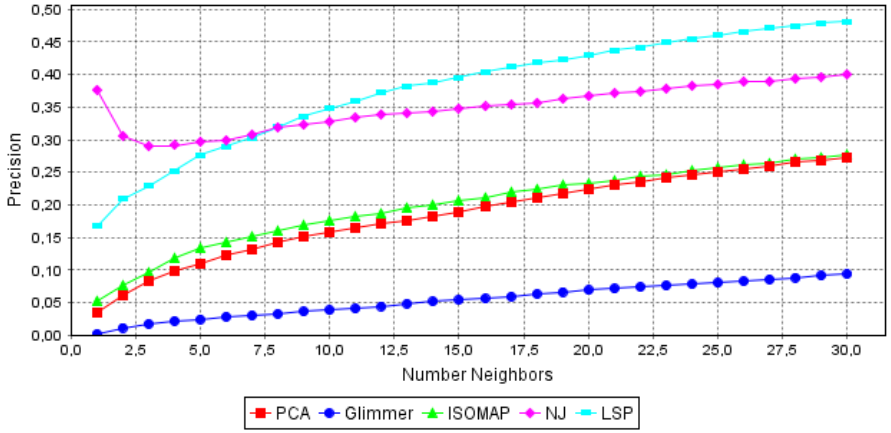
Evaluation

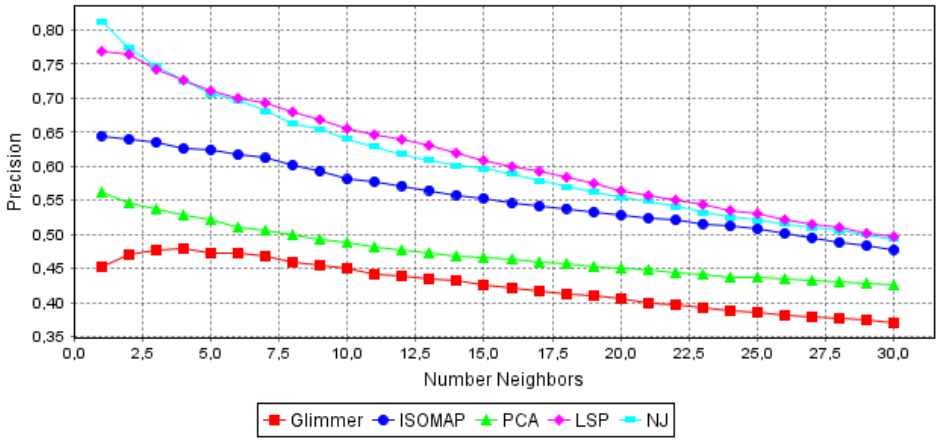
- Contrasting users results with numerical measurements
  - Cluster based
  - Neighborhood
  - Distance based
  - Community based
  - Task

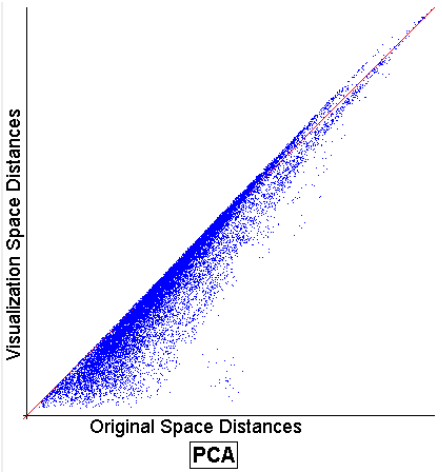
Evaluation

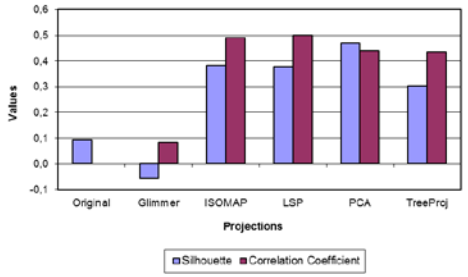
- Numerical Evaluation
- Distance Preservation
- Neighborhood Preservation
- Segregation
- User Understanding
- Visual Explanations

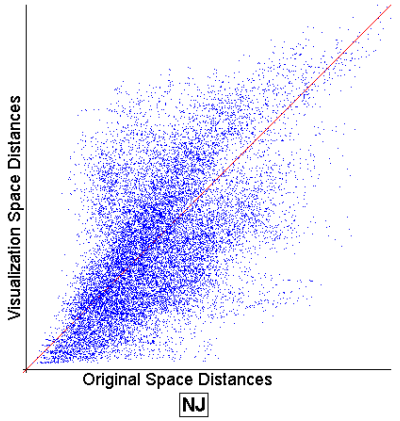
Evaluation


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

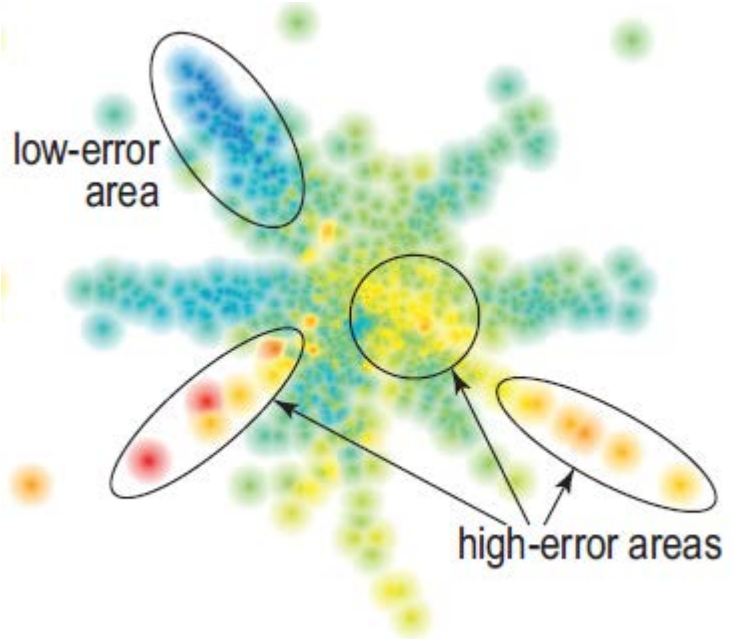
    Yes!
  - H4 User preference is governed by good segregation

    No!

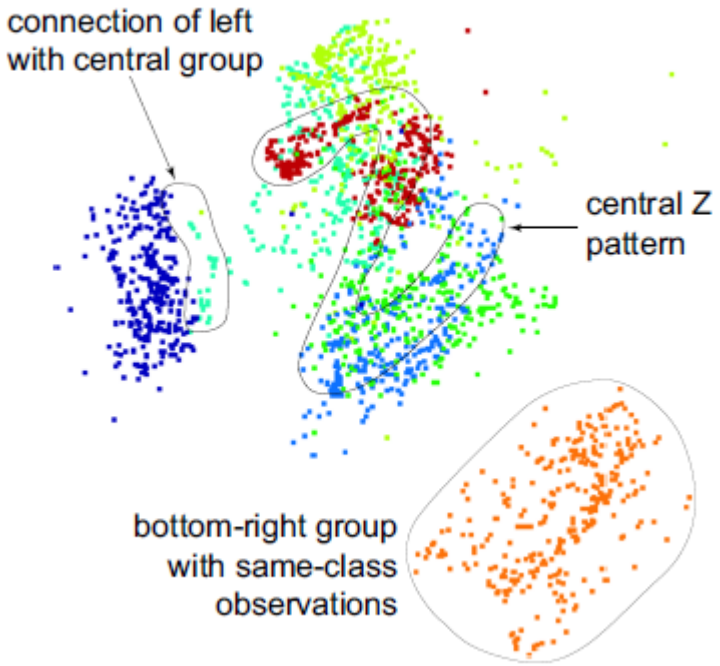
Etemadpour, R. ; Motta, R. ; Paiva, J. G. S. ; MINGHIM, R. ; Oliveira, M. C. F. ; Linsen, L. . Perception-Based Evaluation of Projection Methods for Multidimensional Data Visualization. IEEE Transactions on Visualization and Computer Graphics, v. 21, p. 81-94, 2015.

Evaluation
Explaining a projections

Error mapping



Label ckecking



Explaining neighborhood preservation for multidimensional projections

RM Martins, R Minghim, AC Telea - EG UK Computer Graphics and Visual Computing, 2015

Plus

- Scalability
- Multiscale
- Understanting of feature spaces
- Time-varying volumes
- More evaluation
- Many more applications (molecular interactions, genome)
- Change of visual layouts
- Visual classification of images and other data

vicg.icmc.usp.br

Part II
MINING MEETS VISUALIZATION

Techniques and applications

Visual strategies to support data analysis/mining tasks

Problems regarding scale of data sets

Visual Data Mining

- Dimension Reduction
- Clustering Visualization
- Labeling
- Classification: sample selection, model creation and application, evolution of models
- Cooperation UNICAMP (Campinas), UFU (Uberlândia) and UFMG (Belo Horizonte)

Open problems

- Metaphors: user interface, scalability, user control...
  - Handling text
  - Handling time-varying data
  - Going small: portable devices
  - Growing large: scalability issues
- Evaluation: user perception, quantitative & qualitative metrics
- Development software platform
- Recent developments and current work

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

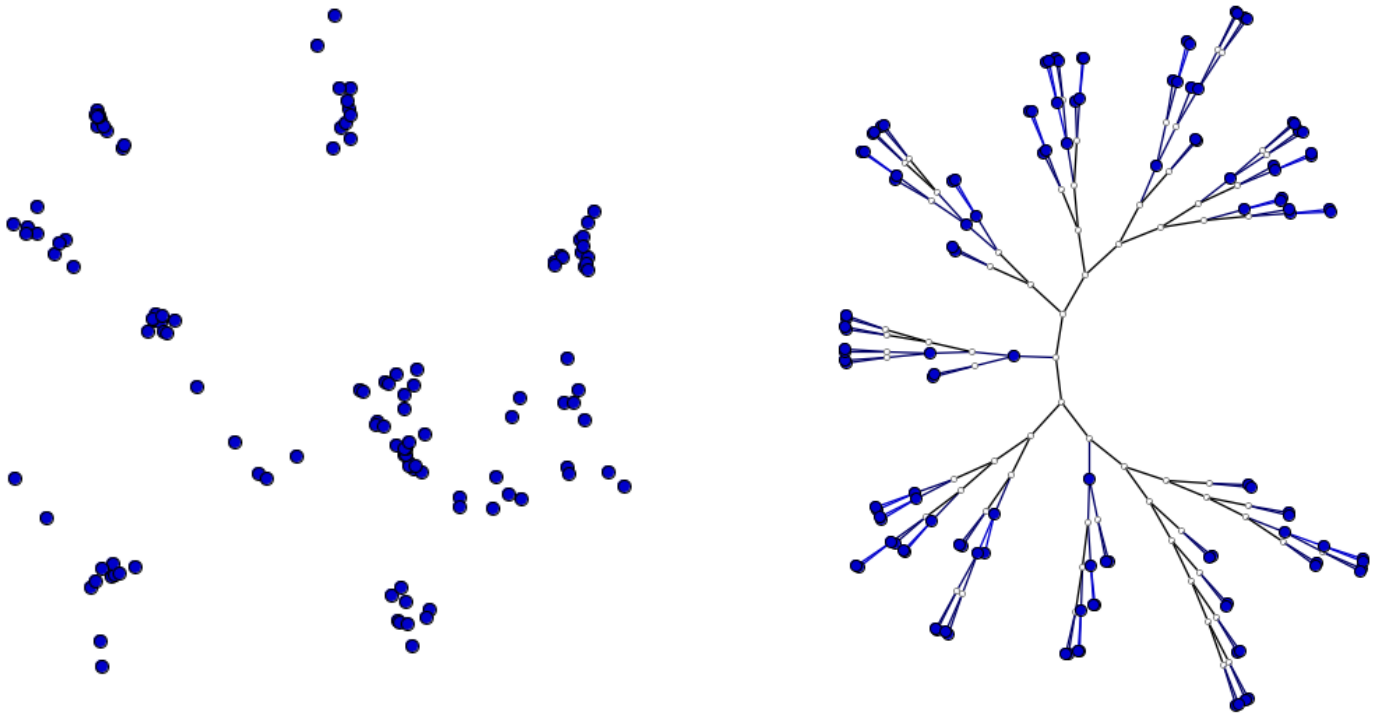
Contribution

- Visual classification methodology (VCM)
- User insertion in the classification process
- Association
- Automatic classifiers
- Similarity and point-based visualization techniques

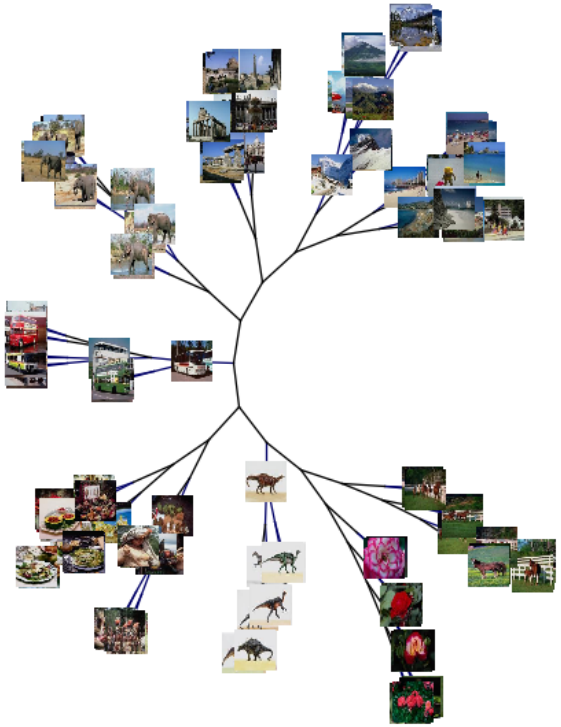
- Possibilities
- Support in labeling
- Model creation for data classification
- Detailed visual analysis of classification results
- Incremental update for results convergence

Task: Classification of Unlabeled Data set

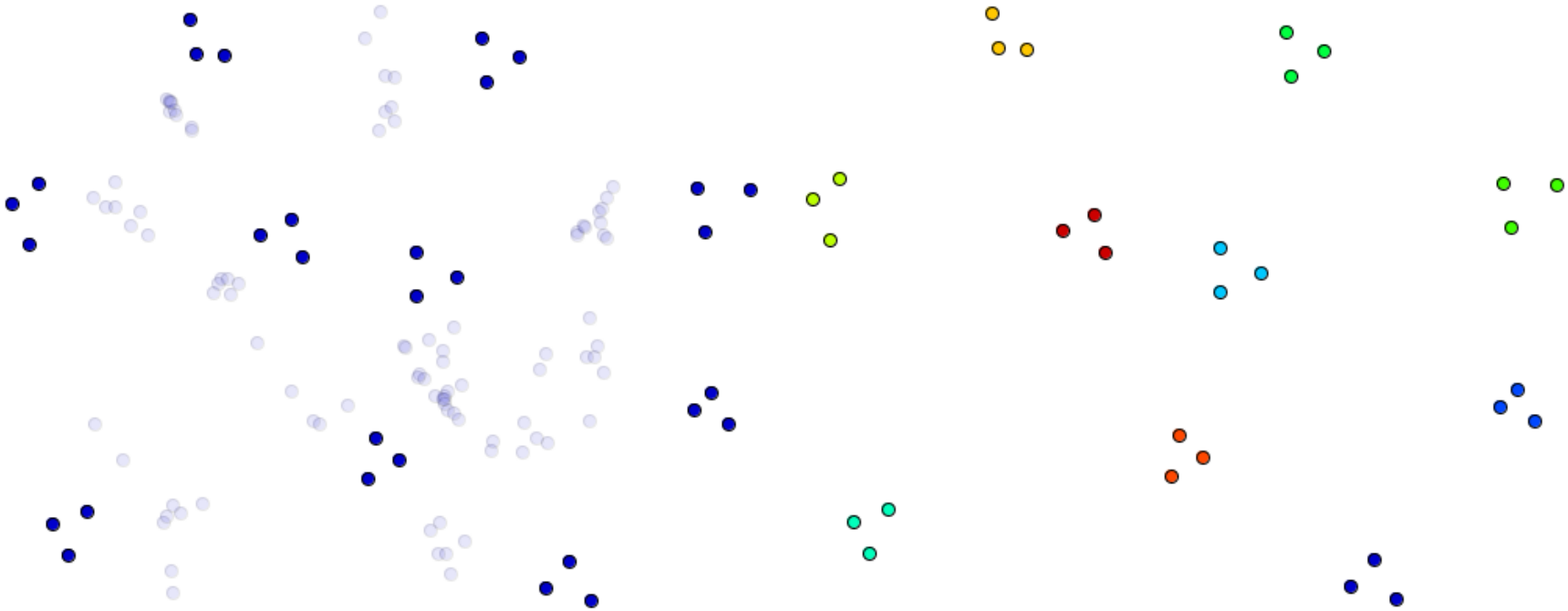
Similarity Organization



Similarity Organization

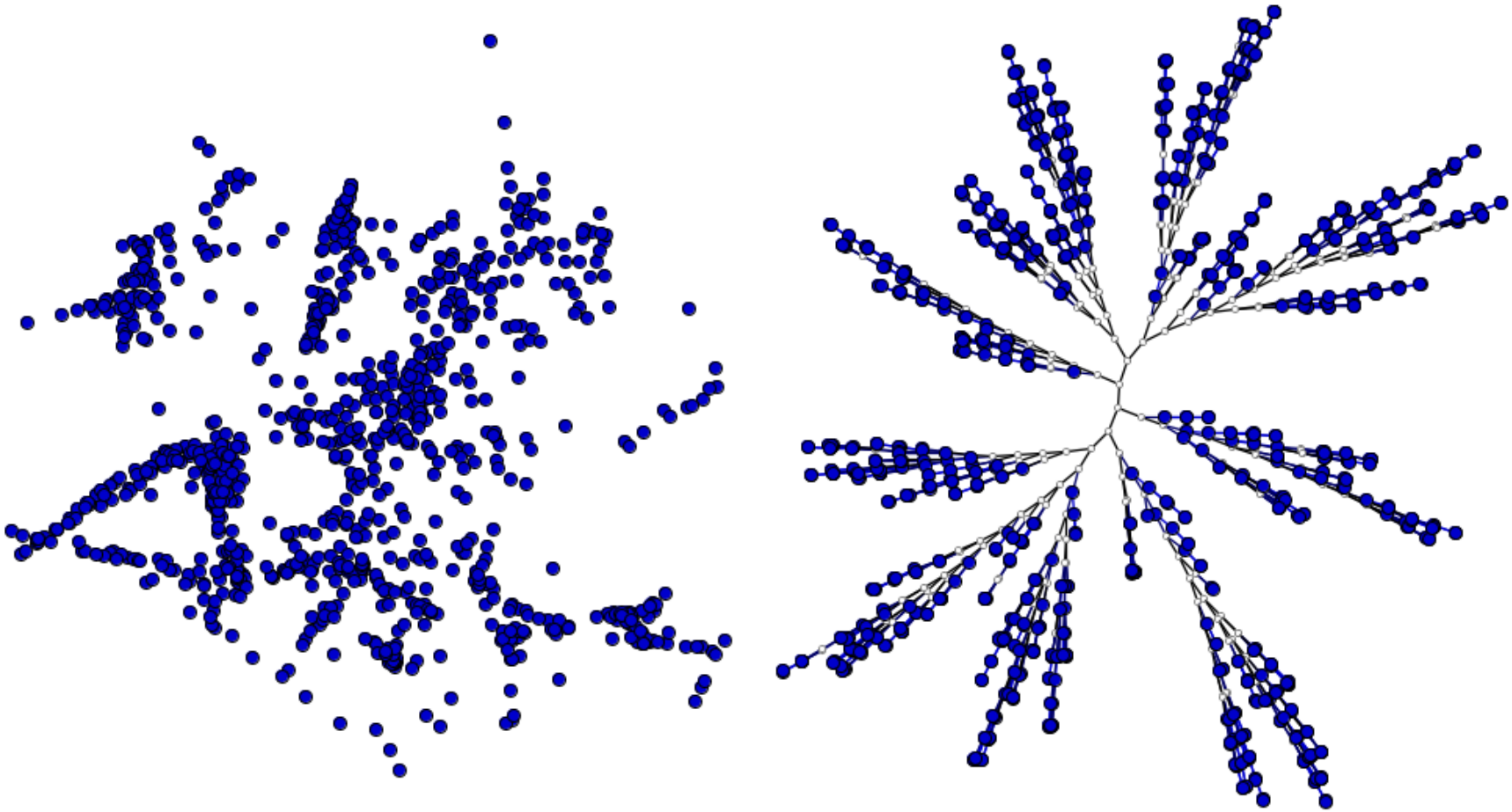


Selection of Representative Instances

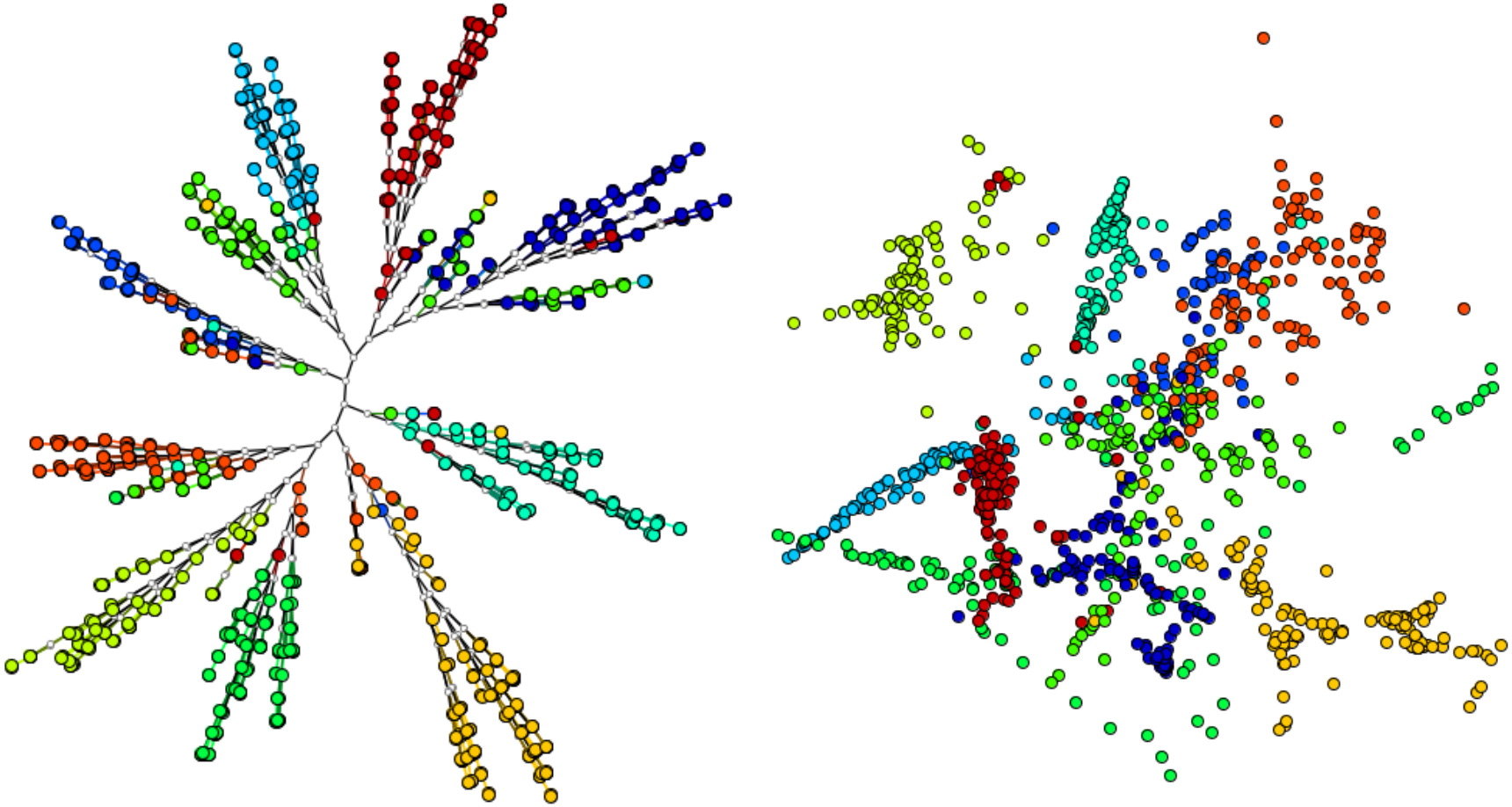


Instances selected to train classification model

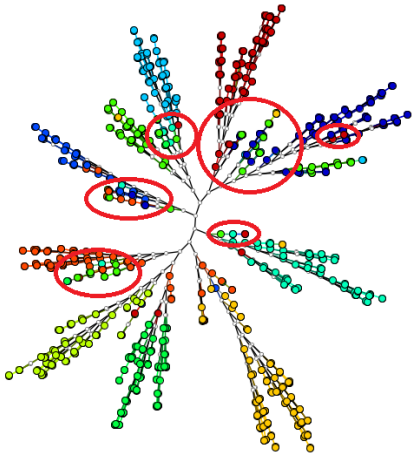
Classification using Created Model



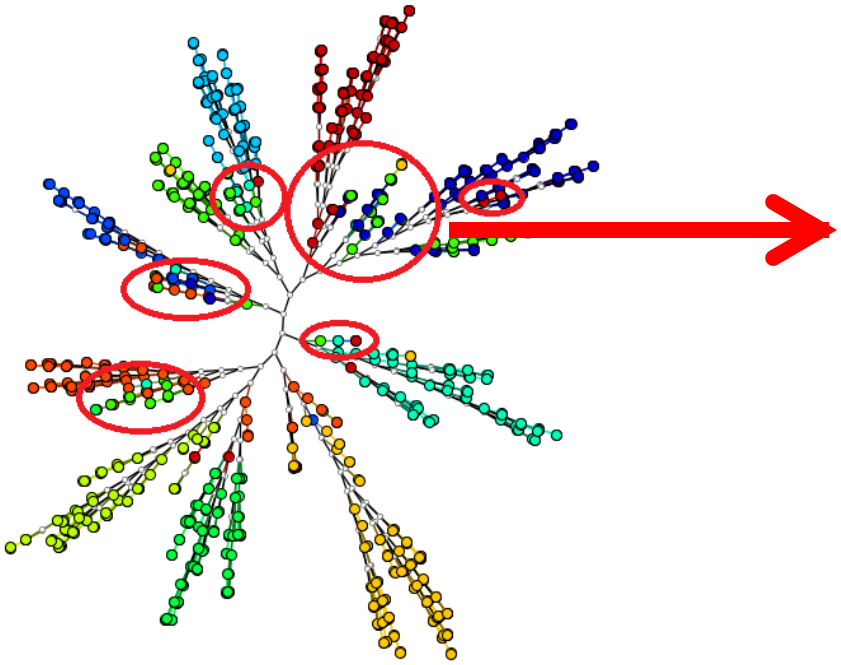
Classification Results



Evaluation: Classification Results

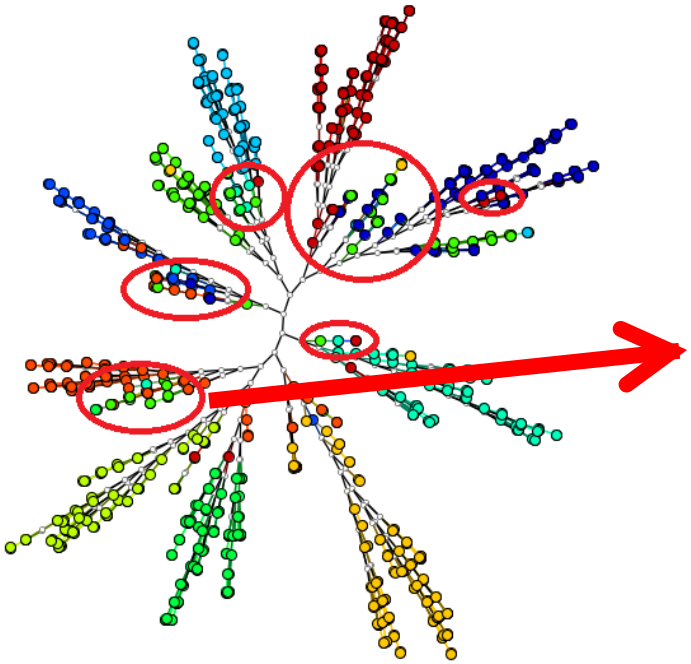


Evaluation: Classification Results



Evaluation: Classification Results

Evaluation: Classification Results

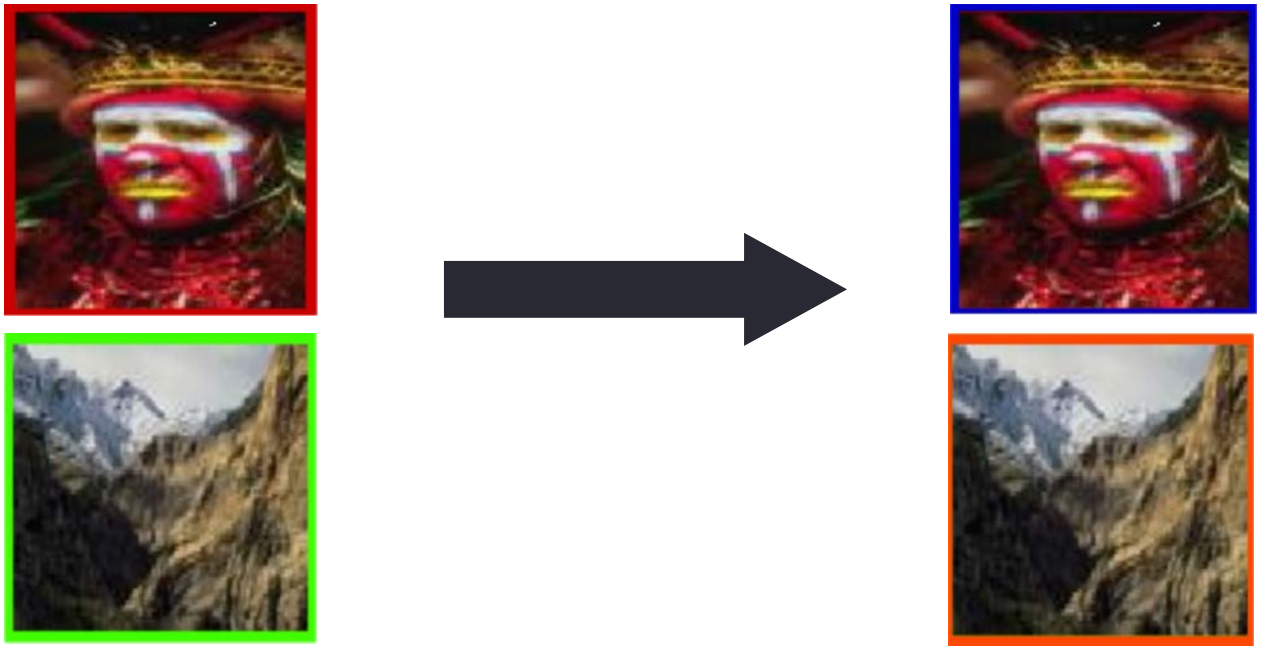


Classification Model Upgrading

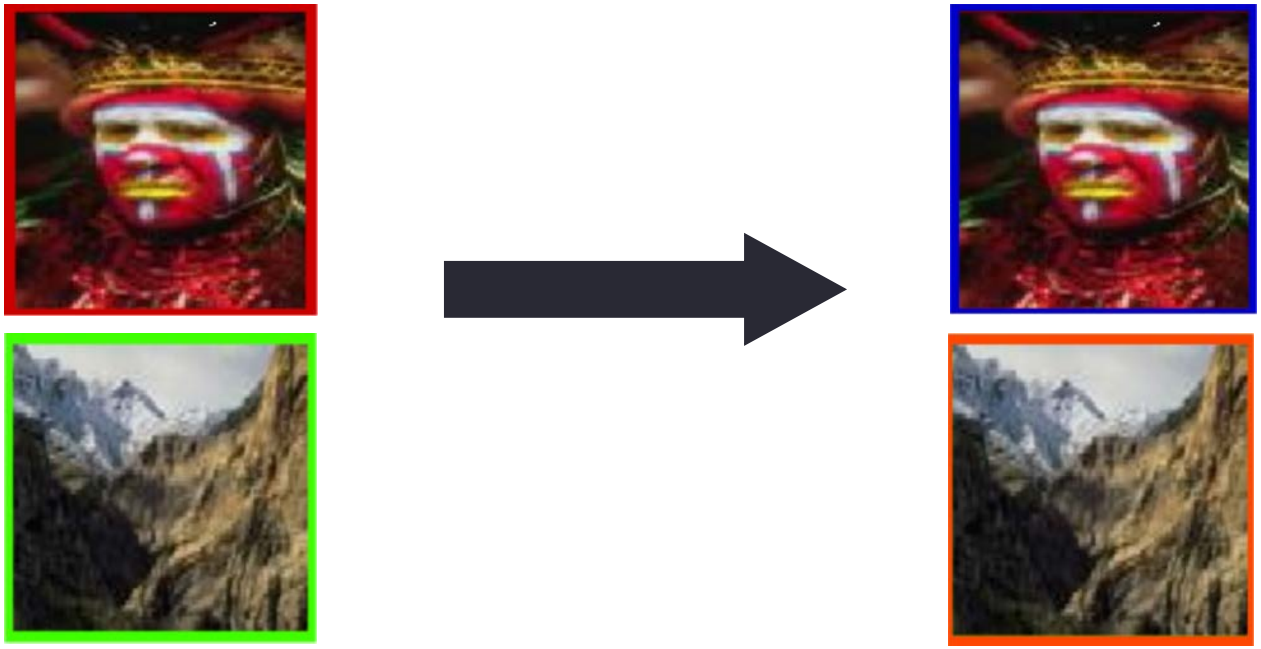
- Several upgrade strategies: Layout also works as a guide
- Example: relabeling of strategic instances: adjustment to specific scenarios

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

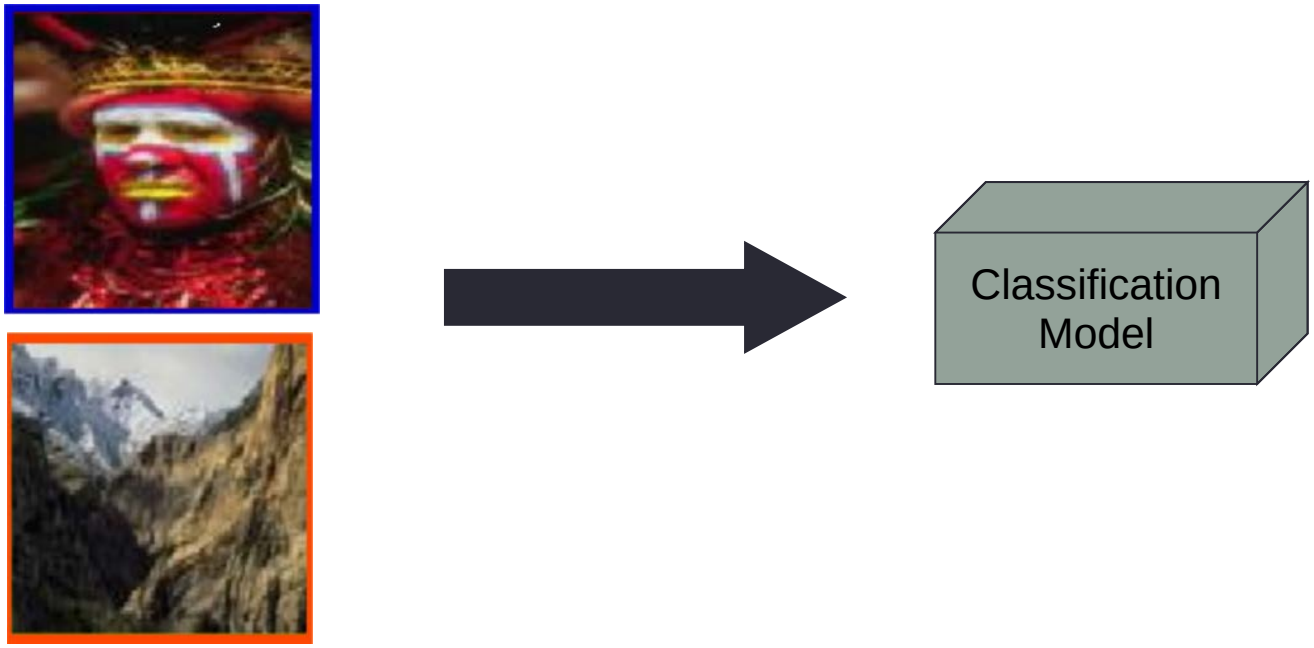
Misclassified Instances Relabeling



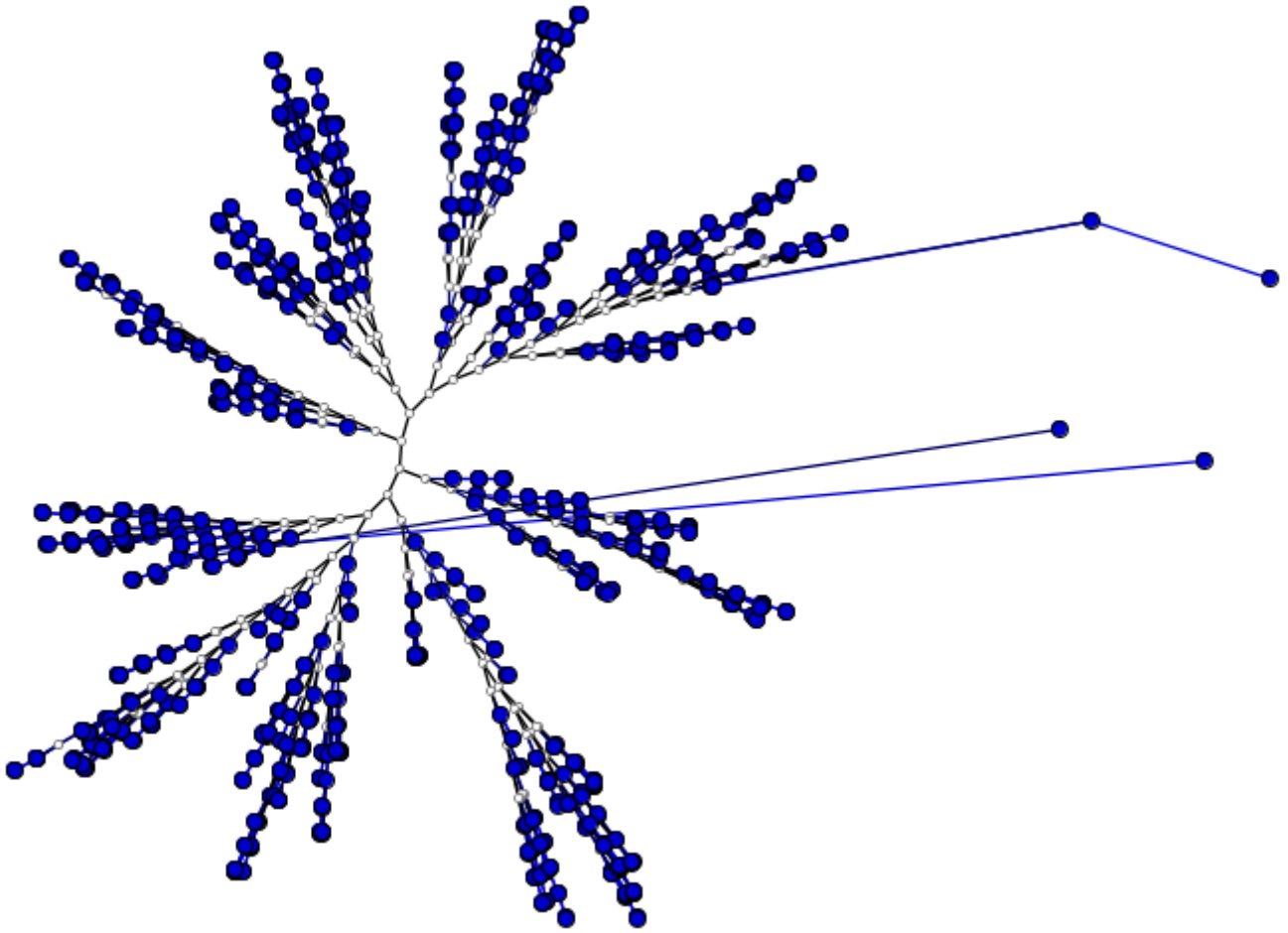
Misclassified Instances Relabeling



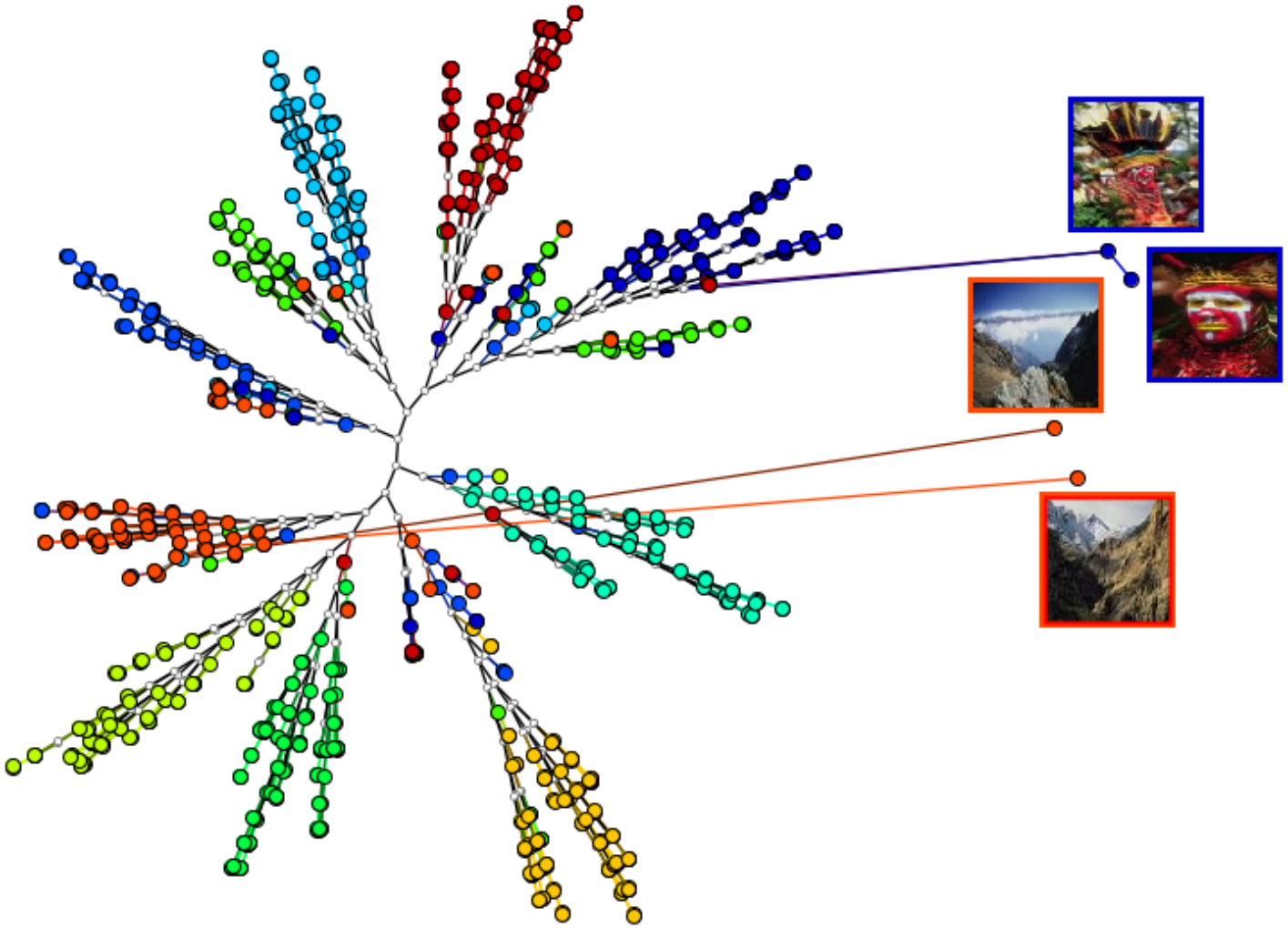
Classification Model Upgrading



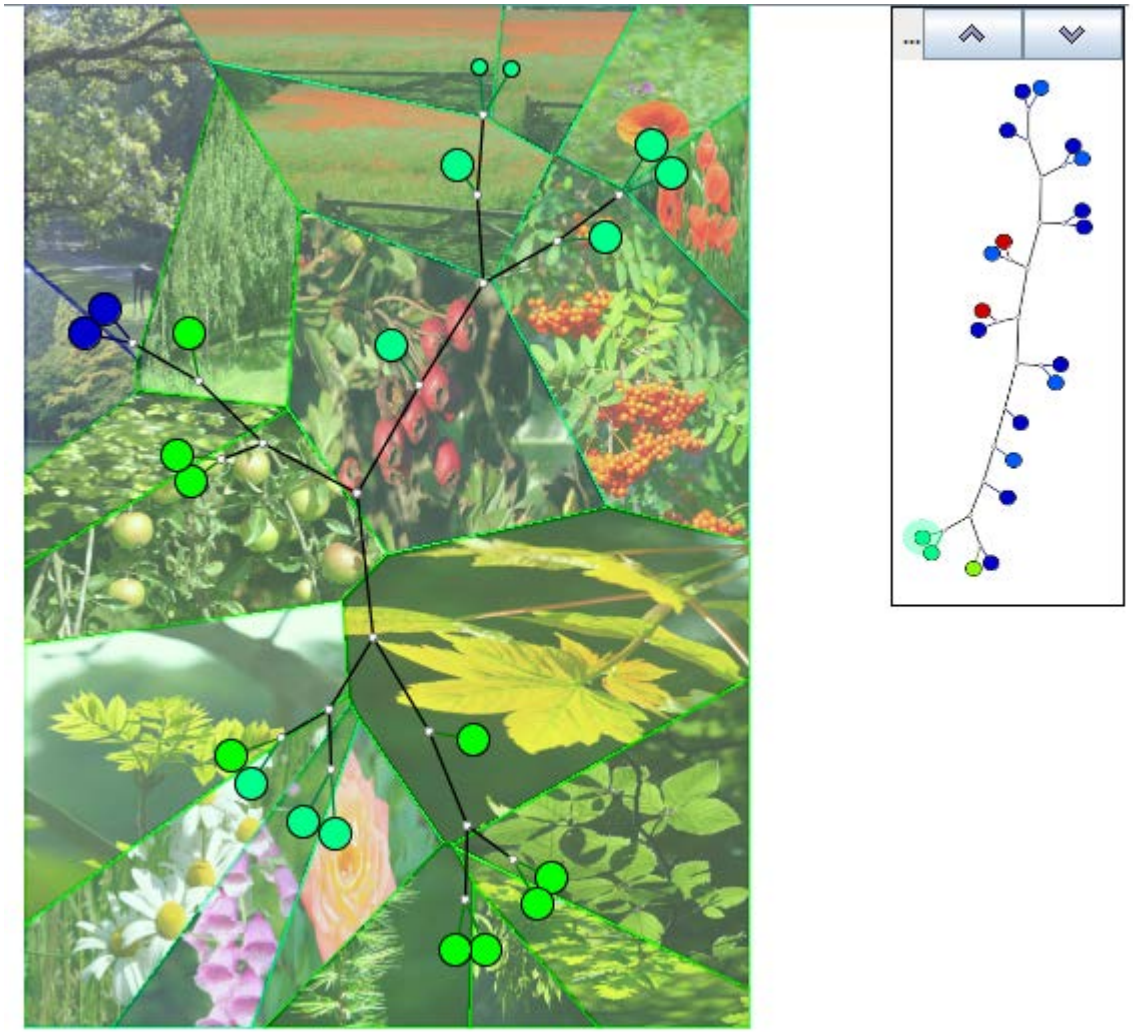
Reclassification - Upgraded Model



Reclassification - Upgraded Model



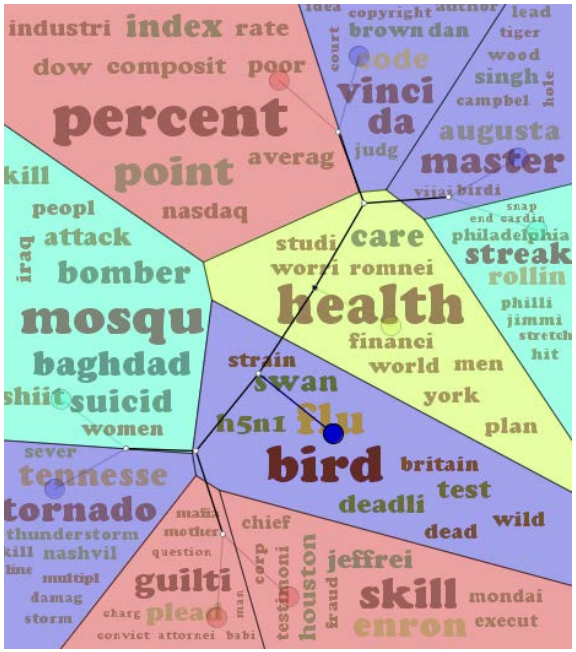
Current Work: handling scalability? The Visual Super Tree



More Data – Summarization

Wordclouds

Representative Images



Multi-level text

Multi-level images

The people

José Gustavo S. Paiva
UFU

Maria Cristina F. Oliveira
ICMC

Fernando V. Paulovich
ICMC

Guilherme P. Telles
UNICAMP

Luis Gustavo Nonato
ICMC

William Robson Schwartz
UFMG

Hélio Pedrini
UNICAMP

Some more people

Renato Oliveira

Henry Heberle

Sonia Castelo

Danilo Eler (UNESP)

Rafael Martins

Thanks!!!!

Fábio Rolli

Partners & collaborators

- Guilherme P. Telles, Hélio Pedrini IC-UNICAMP
- William Schwartz, UFMG
- Danilo Medeiros Eler, UNESP

- Evangelos Milios & team, Dalhousie U., Canada.
- Alexandru Telea, *University of Groningen, the Netherlands*
- Stan Matwin & team – Dalhousie U.

- Osvaldo Novais de Oliveira Jr., IFSC-USP, and nBioNet research network http://www.ifsc.usp.br/nbionet/
- Armando Vieira, Biology Department, UFSCar

- National Laboratory for Biosciences – Campinas – Brazil.

Funding (current)

- CAPES / DAAD PROBRAL
- CAPES / NUFFIC
- CAPES 04/CII-2008, Network NANOBIOTEC-Brasil
- FAPESP – student grants
- CNPQ personal grants / student grants
- CNPq Universal 2012-2013
- CNPq, INCT – MACC Network

Thanks!!!!

CAPES

FAPESP

CNPq

60 ANOS