Instituto de Ciências Matemáticas e de Computação
| Universidade de São Paulo |

DATA VISUALIZATION BASICS

Multidimensional Projections and
Similarity Trees/
Text / other applications

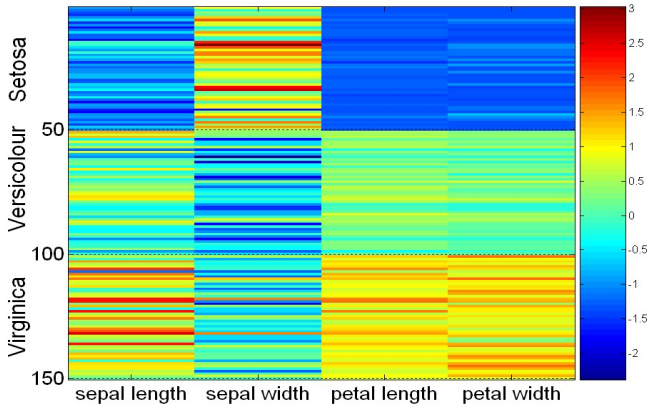
Rosane Minghim
2019-2

Multidimensional Visualization
Projections/Multidimensional Projections
Document Collections
Image Collections

- Visualization
- Visual Mining and Visual Analysis
- Projections
- Examples:
  - Text and Images

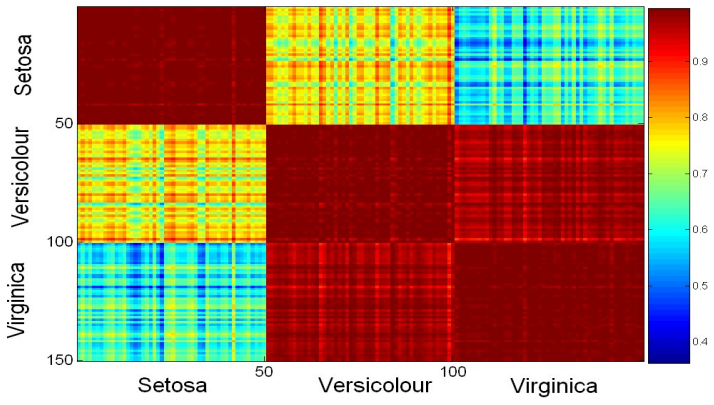
3

Data Matrix



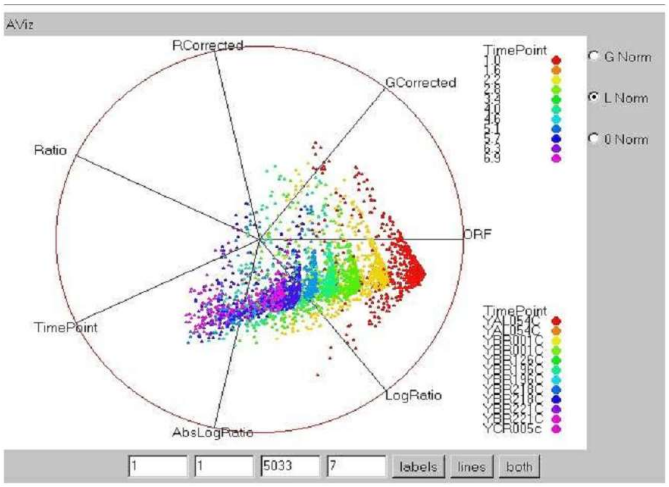
4

Correlation Matrix

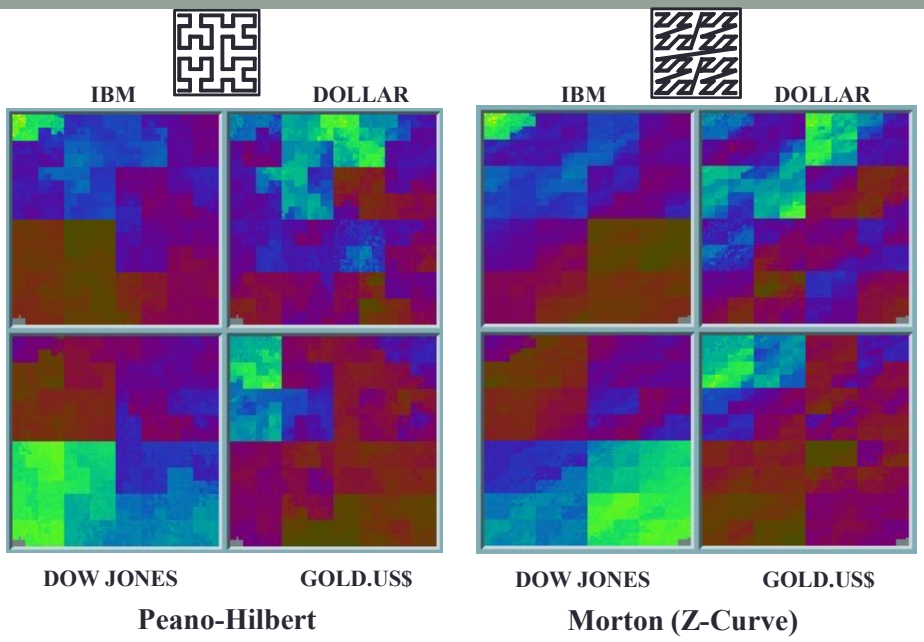


5

RadViz



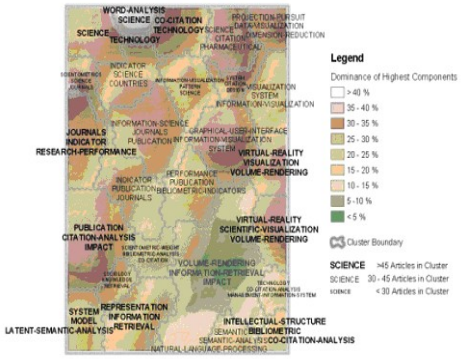
6



Space-Filling Curves

SOM based

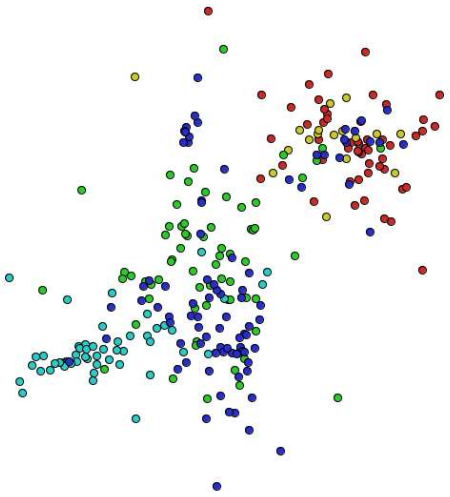
- **Self-Organization Maps (SOMs) cartográficos (ex. Skurpin 2002)**



9

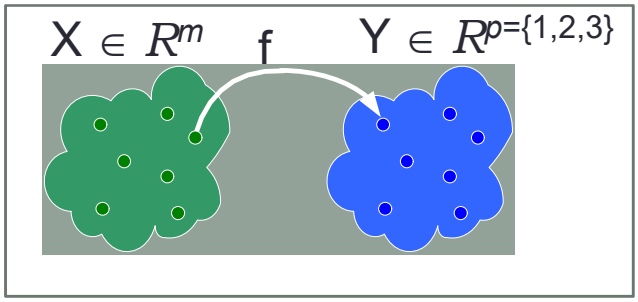
Mapeamento para o plano permitindo a exploração.

Ex: Patents surgery, drugs, molecular bio



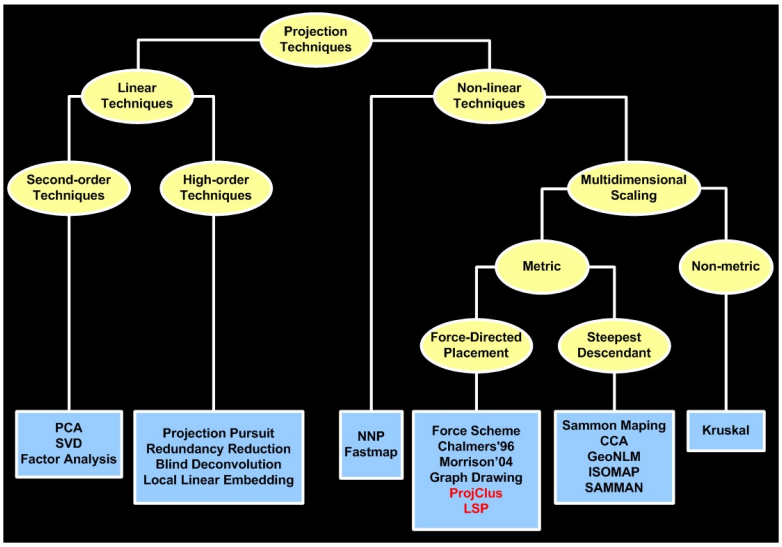
10

Projection Techniques



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

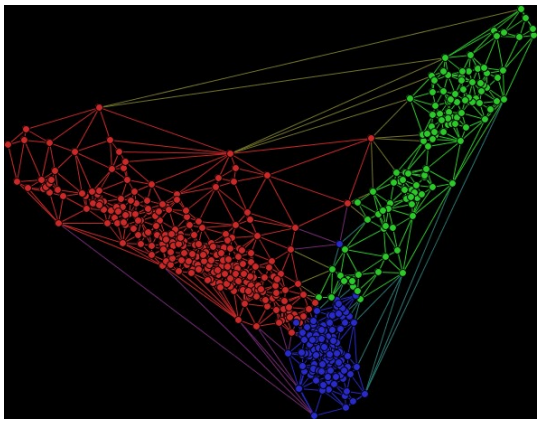
11



12

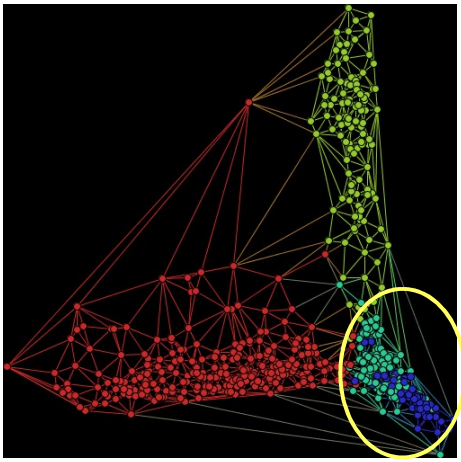
Problems PCA

390 dimensions



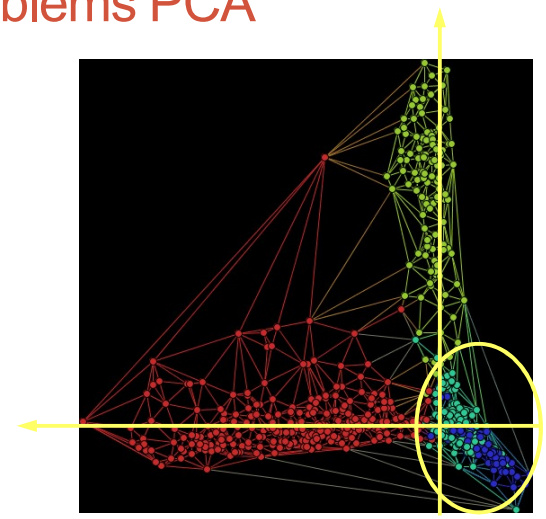
13

Problems PCA



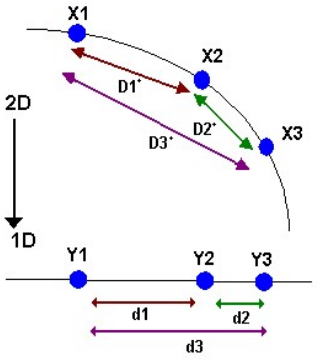
14

Problems PCA



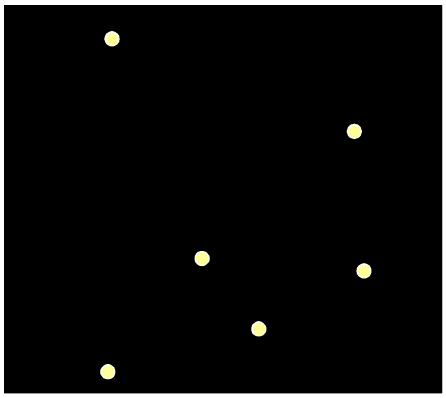
Ex: Sammon Mapping

- Let **X** be the points in the original space R^n, we apply a distance measure d_{ij}* between Xi an Xj., and find **Y, the projected point**, ex. R^2 and d_{ij} the Euclidean distance between them.
- Sammon's method applies an error function to measure the target.



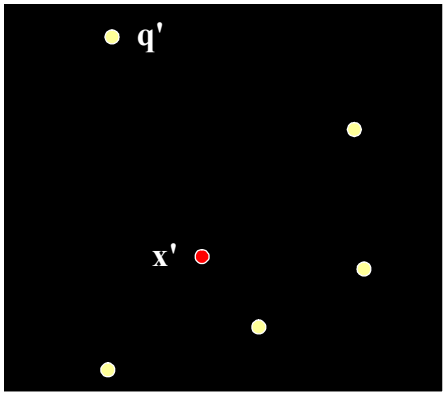
16

Force Based Point Placement



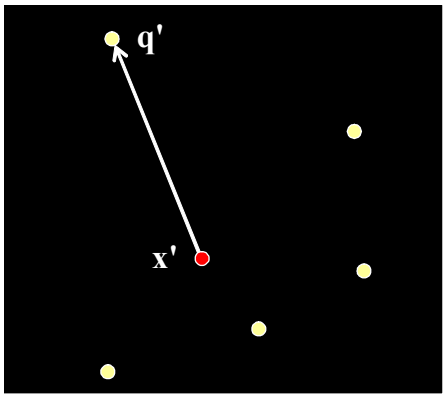
17

Force Scheme [Tejada et al., 2003]



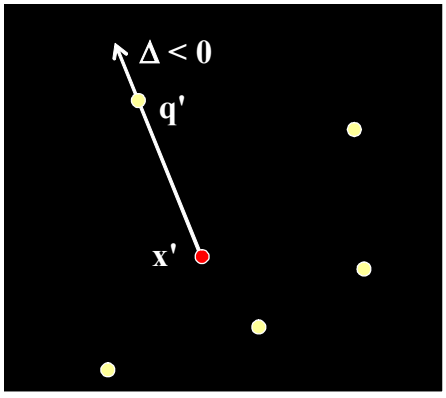
18

Force Scheme [Tejada et al., 2003]

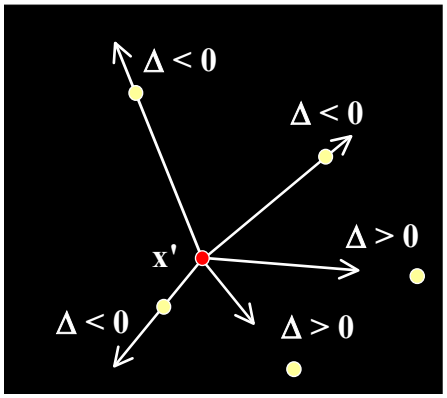


19

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]

22

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.

23

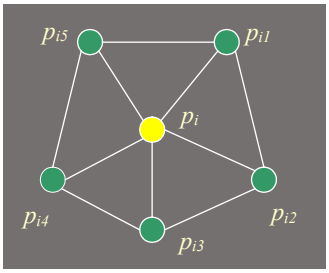
LSP [Paulovich et al., 2006/2008]

- 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,…, x_p are vectors containing the Cartesian coordinates of the points

and L is the matrix defined by:

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

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

27

LSP: Adding control points

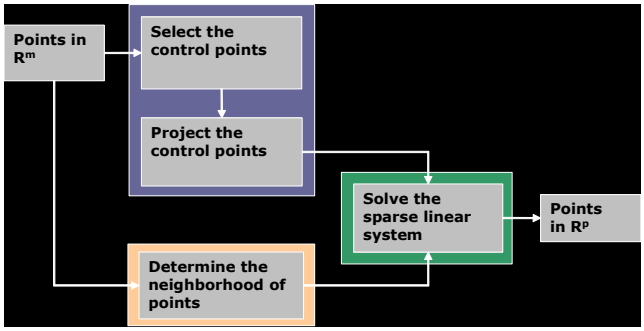
$$A = \begin{pmatrix} L \\ C \end{pmatrix} \qquad Cij = \begin{cases} 1 & p_j \text{ is a control point} \\ 0 & otherwise \end{cases}$$

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

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

28

LSP: Overview



29

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$$L = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \end{bmatrix}$$

30

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$$L = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \end{bmatrix}$$

31

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$$L = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \end{bmatrix}$$

32

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

L

33

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{matrix} \\ \\ \\ \\ \\ \\ \text{C} \\ \end{matrix}$$

L

34

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_{x_3} \\ c_{x_6} \end{bmatrix}$$

35

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_{x_3} \\ c_{x_6} \end{bmatrix}$$

36

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_{y_3} \\ c_{y_6} \end{bmatrix}$$

37

LSP: Exemplo de Sistema

$v_1 = \{p_3 p_4 p_6\}$
$v_2 = \{p_5 p_4 p_6\}$
$v_3 = \{p_1 p_5 p_6\}$
$v_4 = \{p_1 p_6\}$
$v_5 = \{p_3 p_2 p_6\}$
$v_6 = \{p_1 p_2 p_4 p_5\}$

$pc = \{p_3 p_6\}$

$$A = \begin{bmatrix} 1 & 0 & -1/3 & -1/3 & 0 & -1/3 \\ 0 & 1 & 0 & -1/3 & -1/3 & -1/3 \\ -1/3 & 0 & 1 & 0 & -1/3 & -1/3 \\ -1/2 & 0 & 0 & 1 & 0 & -1/2 \\ 0 & -1/3 & -1/3 & 0 & 1 & -1/3 \\ -1/4 & -1/4 & 0 & -1/4 & -1/4 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ c_{y_3} \\ c_{y_6} \end{bmatrix}$$

38

LSP: Solving the system

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

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

- 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

40

Choosing the Control Points

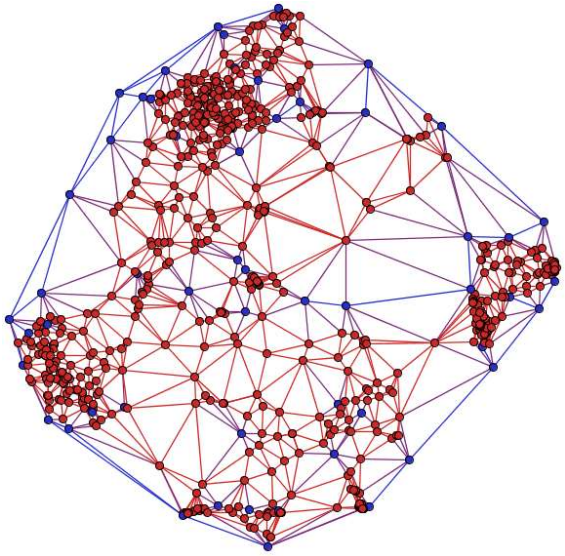
- In order to select the control points
  - the space R^m is split into *nc* clusters using k-medoids.
  - the control points are the medoids of each cluster

41

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

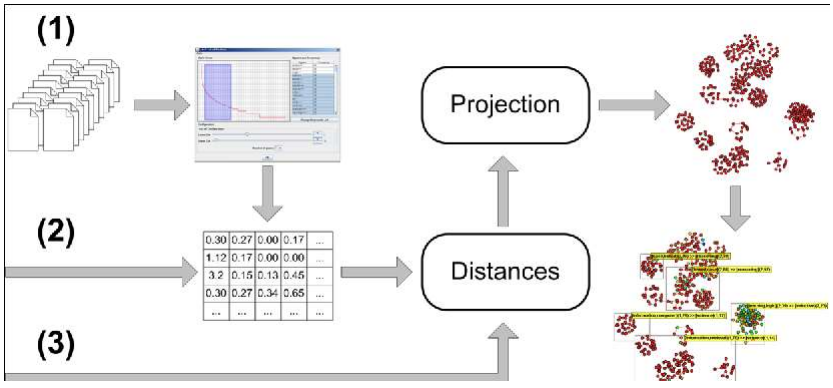
42



Control points in blue

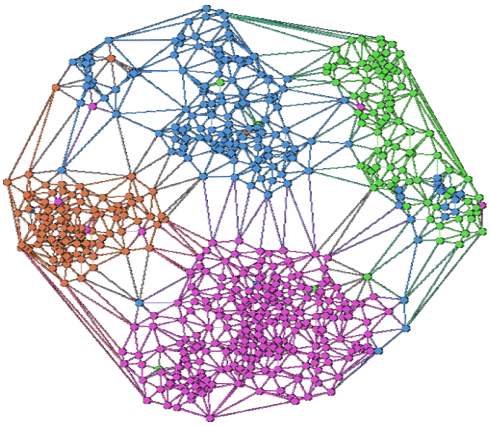
43

Content – based by Projections



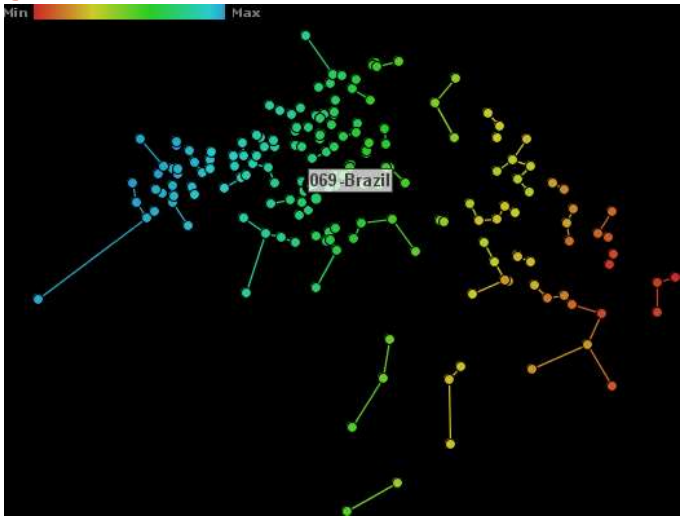
44

Projection



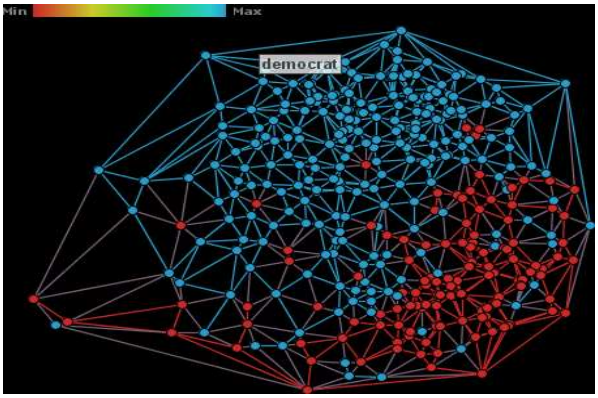
45

Projection: HDI



46

Projection: Voting



47

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

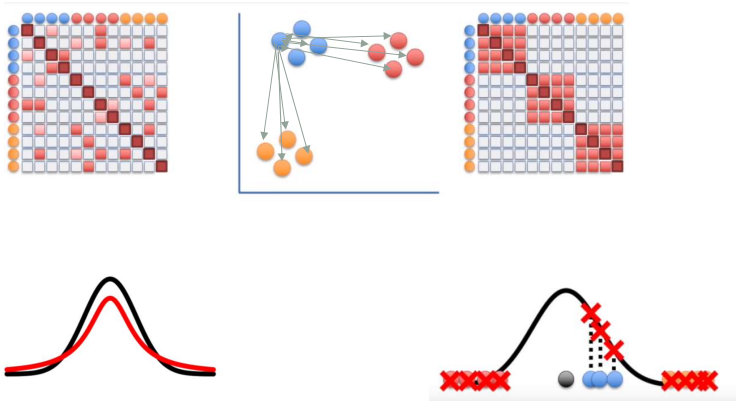
$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

- 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

48

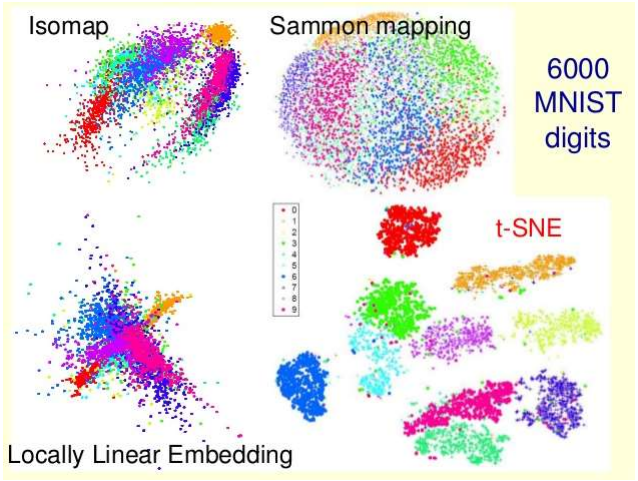
Stochastic Neighborhood Embedding sne and t-sne



Source: StatQuest (adapted)

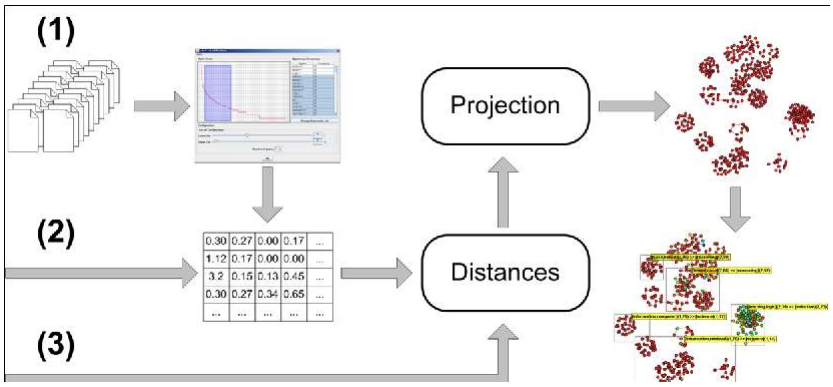
49

T-sne Examples



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

Visualization by Projections



The case of document collections

- Applications
  - Teaching/Research
  - Search
  - Investigation

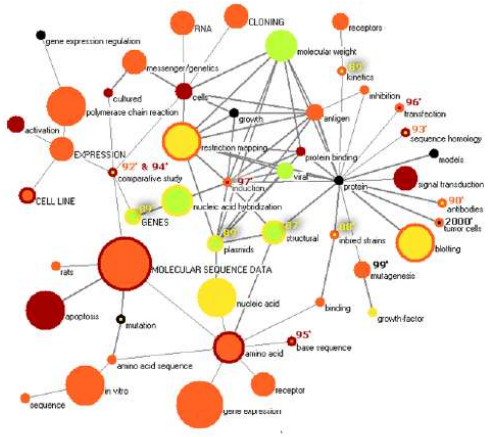
  - Patents
  - Medical reports
  - News

53

- Maps of text Collections
  - Based on Relationships (Borner & Chen)
    - Co-authorship, co-citation
  - Based on Content
    - Similarity and Grouping
    - Common underlying subject
    - ➔Topics

54

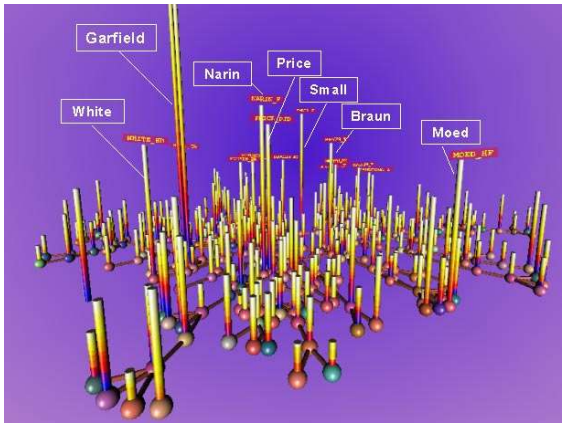
Relationships : Topic Busts and co-word



(Mane and Borner) 2004

55

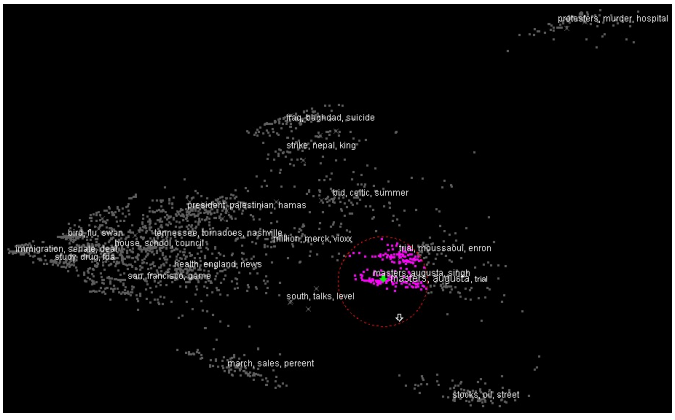
Relationships : Citation and Co-citation



(Borner) (2003)

59

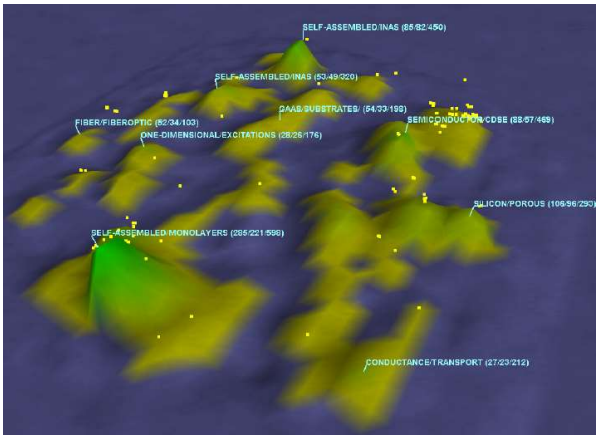
Content - based



(Dimensional Reduction) News flash IN-SPIRE (PNL)

VxInsight

- Sandia National Laboratories, mountain metaphor (Boyack et al., 2002).



66

Text Preprocessing

1. Stopwords elimination
2. Extraction of words radicals (stemming)
3. Creation of n-grams
4. Frequency count and Luhn's lower cut (n-grams appearing less then x times are ignored)
5. Weighting process (*term-frequency inverse document-frequency - (tfidf)*)

67

Result is a Vector Model

- Attributes: terms (n-grams)
- Value: term weight
- Table Data

68

Vector Representation – term weighting

- tf – term frequency
- tfidf – tf x idf = tf x inverse document frequency

$$w_{ik} = tf_{ik} \times log\left(\frac{N}{n_k}\right)$$

69

Vector Representation

	$term_1$	$term_2$	$term_3$	$term_4$	...	$term_m$
Doc_1	0.92	0.62	0.92	0.10	...	0.67
Doc_2	0.13	0.11	1.00	0.34	...	0.33
Doc_3	0.52	0.00	0.00	0.44	...	0.77
...	...	...	...	...	...	...
Doc_n	0.02	0.12	0.22	0.92	...	0.00

70

Vector Representation – Similarity calculatin

EUCLIDEAN

$$sim_{i,j} = \sqrt{(w_{i,1} - w_{j,1})^2 + \ldots + (w_{i,k} - w_{j,k})^2}$$

MANHATAN

$$sim_{i,j} = |w_{i,1} - w_{j,1}| + \ldots + |w_{i,1} - w_{j,1}|$$

COSINE

$$sim_{i,j} = \frac{(w_{i,1} \times w_{j,1}) + \ldots + (w_{i,k} \times w_{j,k})}{(w_{i,1}^2 + \ldots + w_{i,k}^2) \times (w_{j,1}^2 + \ldots + w_{j,k}^2)}$$

71

Vector Representation – distance calculation

$$dis(doc_i, doc_j) = \sqrt{2 * (1 - sim(doc_i, doc_j))}$$

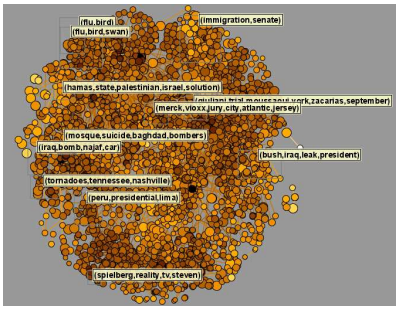
$$sim(doc_i, doc_j) = \frac{doc_i \times doc_j}{\|doc_i\| * \|doc_j\|}$$

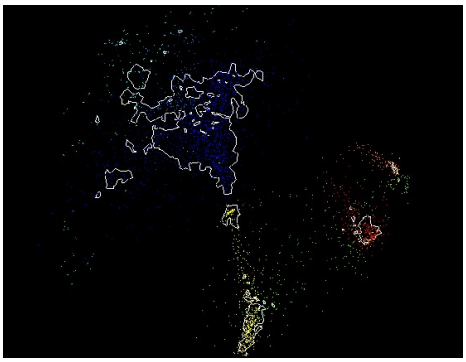
73

Alternatives to Vector Representation

- Similarity Calculation text against text
  - Word2Vec, Doc2Vec (see https://www.tensorflow.org/tutorials/representation/word2vec)
  - Ex: NCD Normalized Compression Distance
    - Approximation of Kolmogorov Complexity
    - Ver: G. P. Telles, R. Minghim, and F. V. Paulovich. 2007. Visual Analytics: Normalized compression distance for visual analysis of document collections. *Comput. Graph.* 31, 3 (June 2007), 327-337. DOI=http://dx.doi.org/10.1016/j.cag.2007.01.024
  - Editing distance
    - Dice's coefficient
    - Matching's coefficient
    - Overlap's coefficient
    - Qgram Distance
    - Ver: Frizzi San Roman Salazar. Um estudo sobre o papel de medidas de similaridade na visualização de coleções de documentos. 2012. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Fundação de Amparo à Pesquisa do Estado de São Paulo. Orientador: Maria Cristina Ferreira de Oliveira.

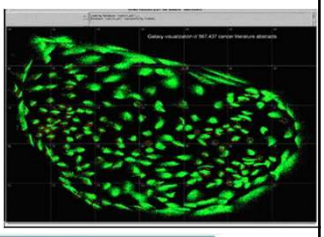
Visual representations: graphs, surfaces, volumes, triangulations

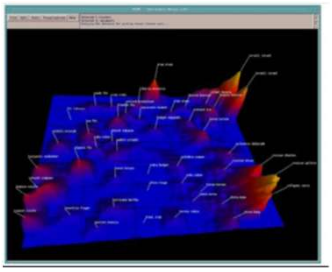




IN-SPIRE

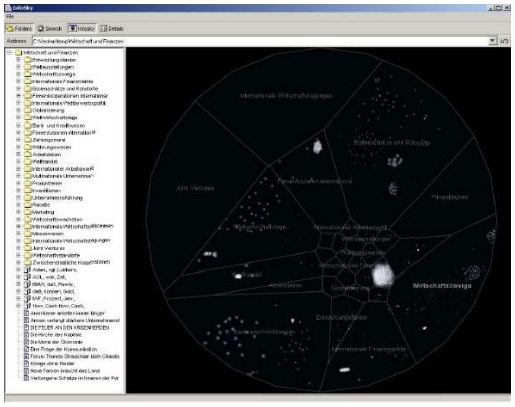
- Spatial Paradigm for Information Retrieval - Pacific Nortwest National Laboratories
- Two Visualization Metaphors:
  - Galaxies – dimensional reduction
  - Themescape



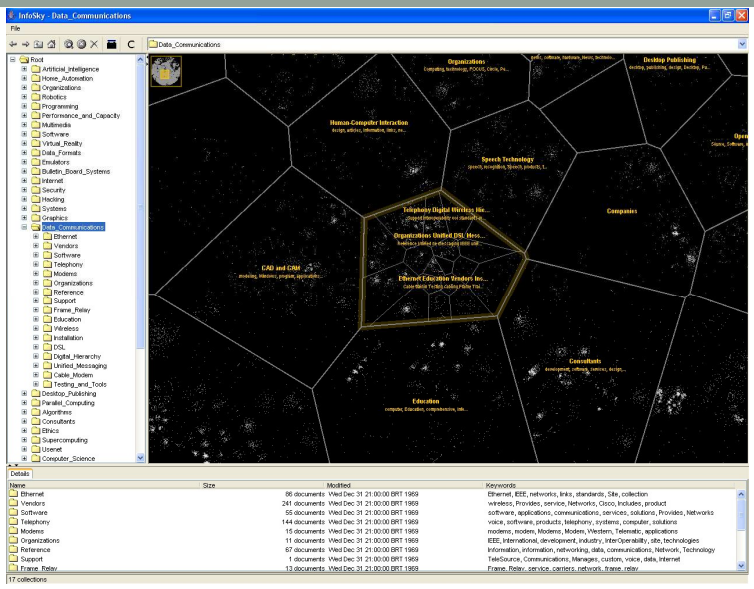


InfoSky

Granitzer (Granitzer et al., 2004) also employs galaxy metaphor



80

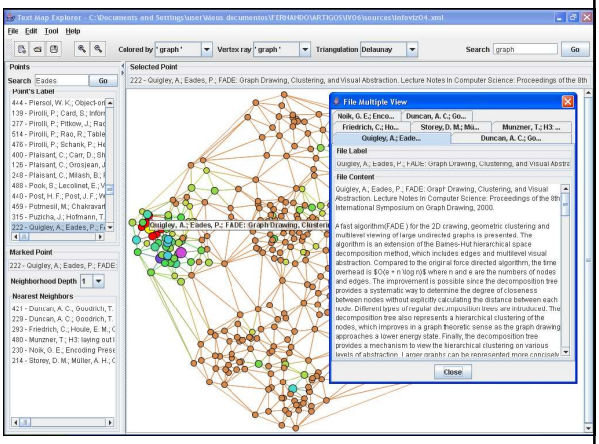


http://en.know-center.at/forschung/wissenserschliessung/downloads_demos/infosky_demo

82

Projection Explorer (PEx)

- Projection and Point placement
- Precision
- Graphs and surfaces (Super Spider)

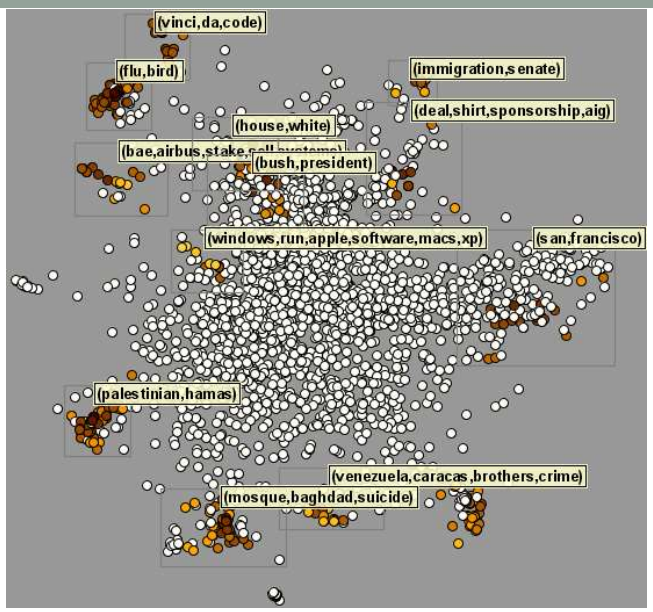


83

Mapping Text Collections via Projections and Point Placement

- Positioning and labeling

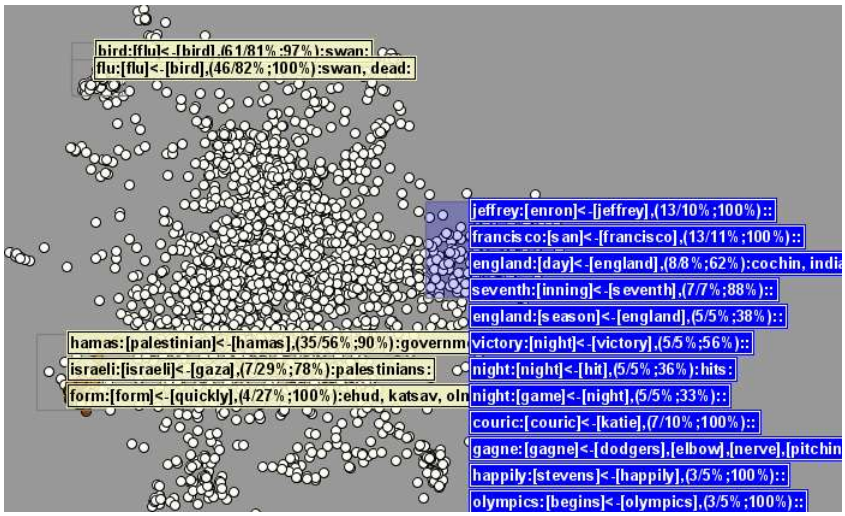
84



85

- Detailing topics

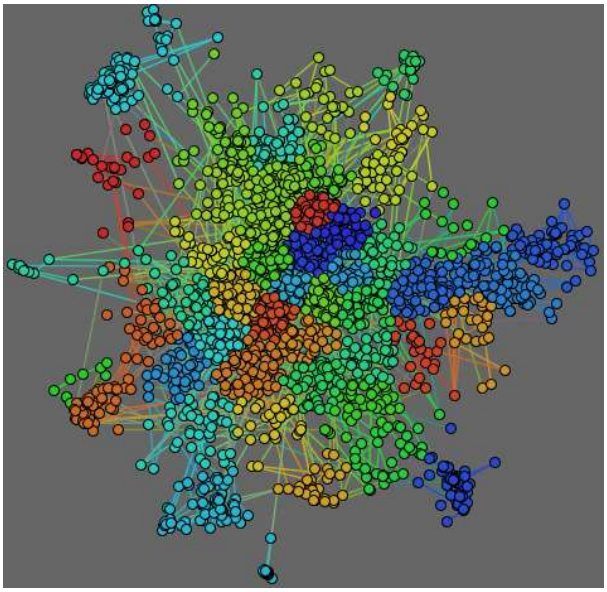
86



87

- Finding Relationships

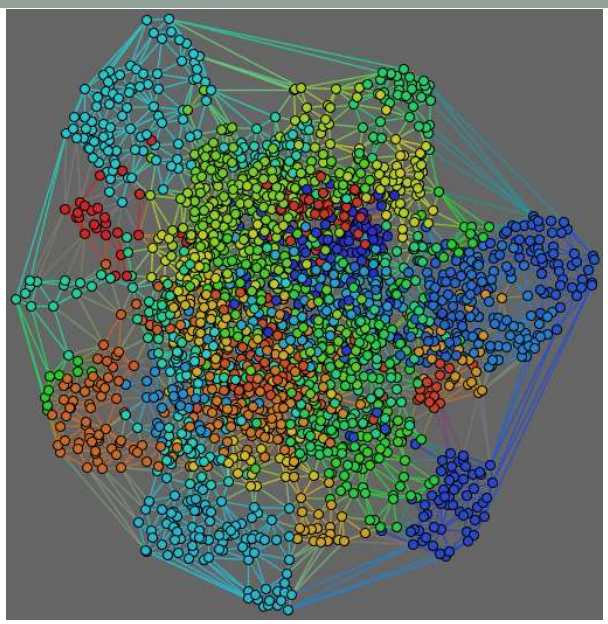
88



91

- Building a mesh

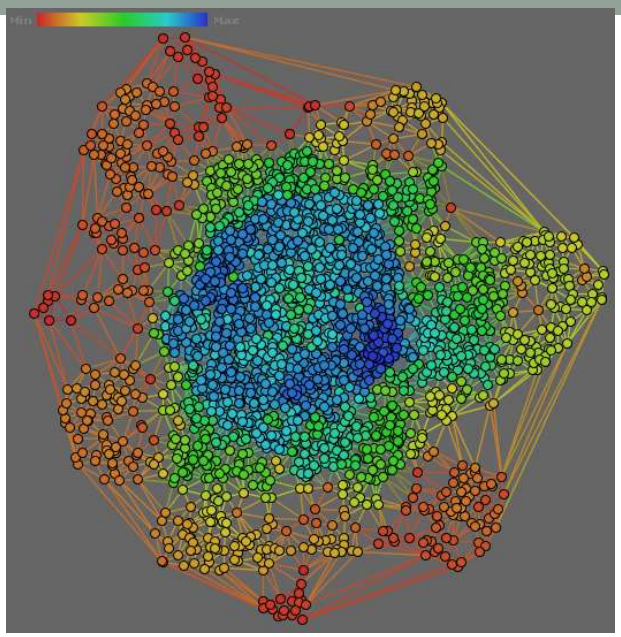
92



93

- Coloring by degree of proximity

94



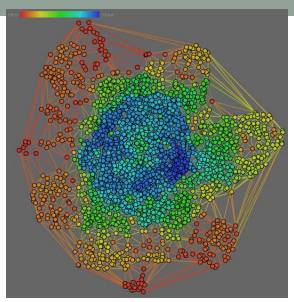
95

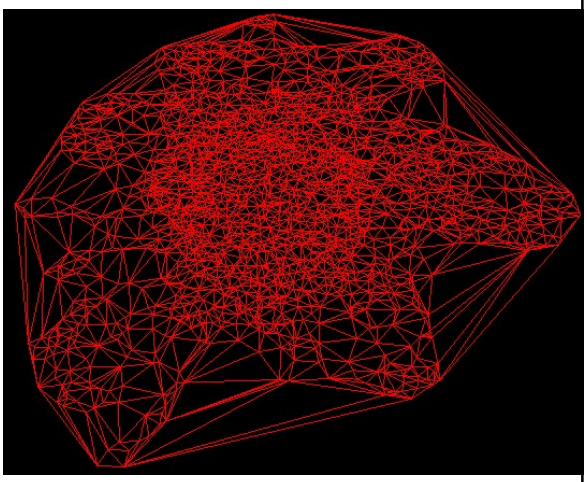
- Coordinating

97

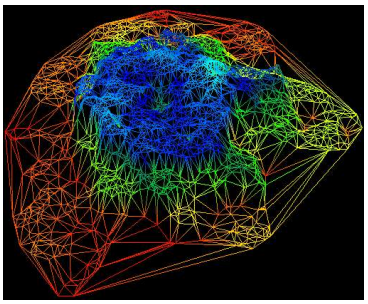
- Building a Surface

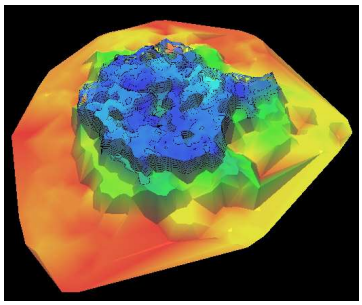
98

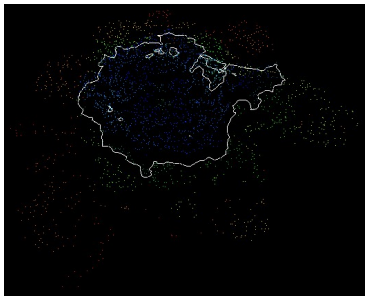




99

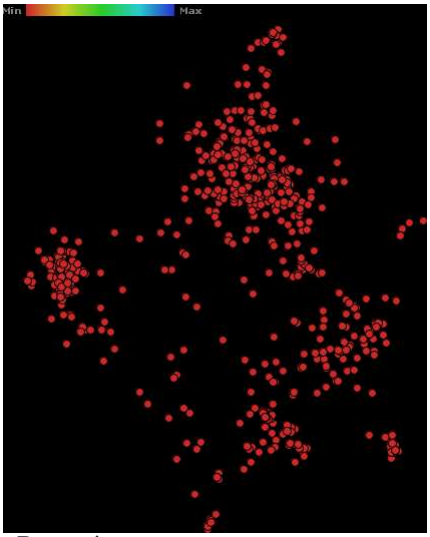






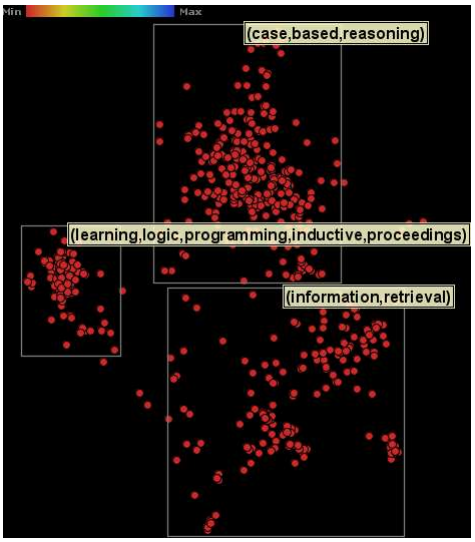
100

Explorando

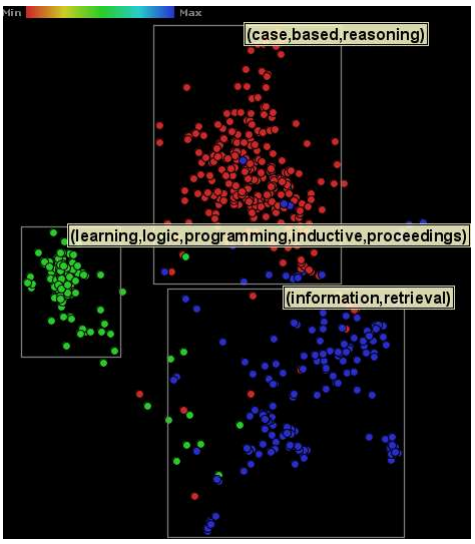


- Case-Base Reasoning
- Information Retrieval
- Inductive Logic Programming

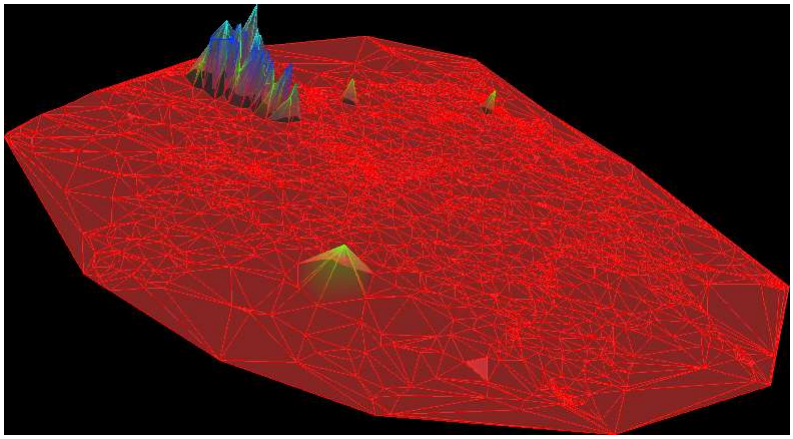
Exemplos de Mapas



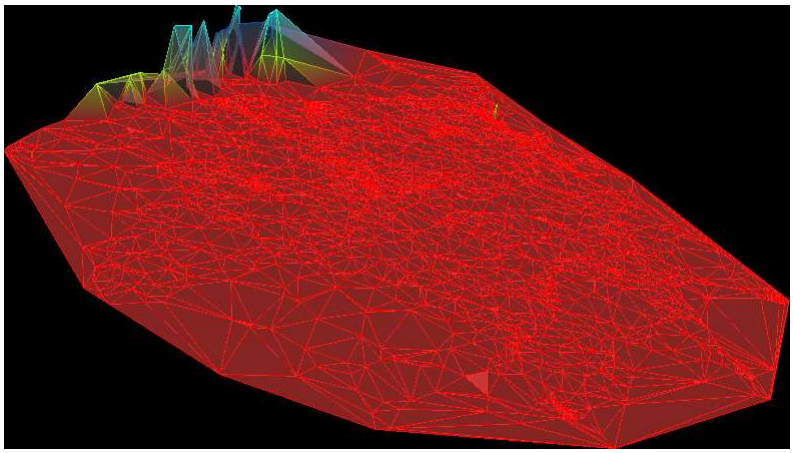
Exemplos de Mapas



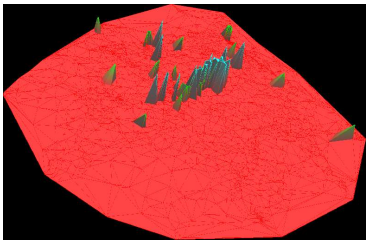
RSS News Flash



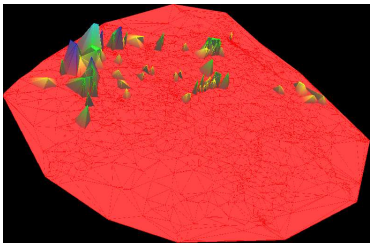
Bird and Flu



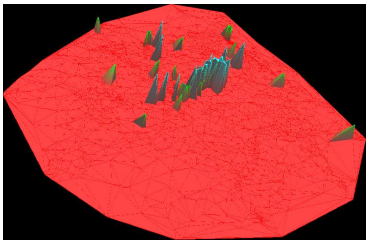
Palestinian



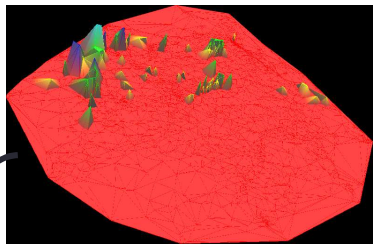
Bush



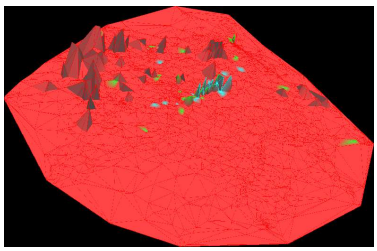
Iraq



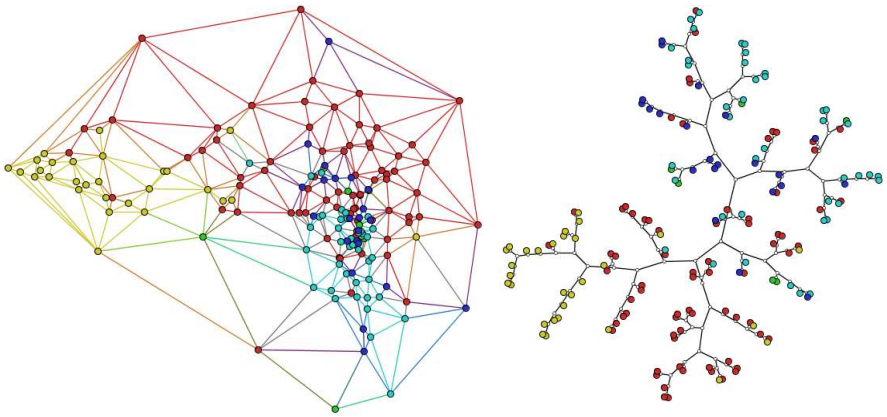
Bush



Iraq



Further Example - patents



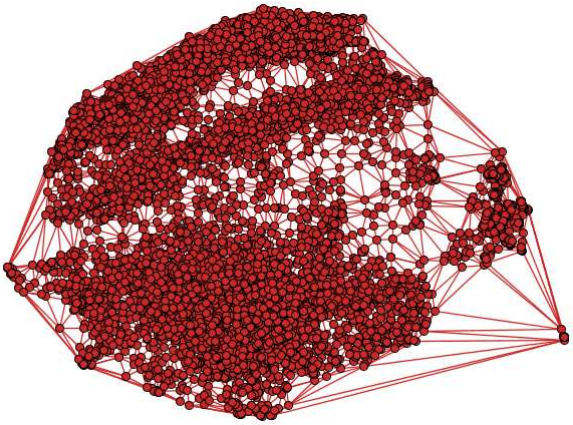
Further Example

- 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

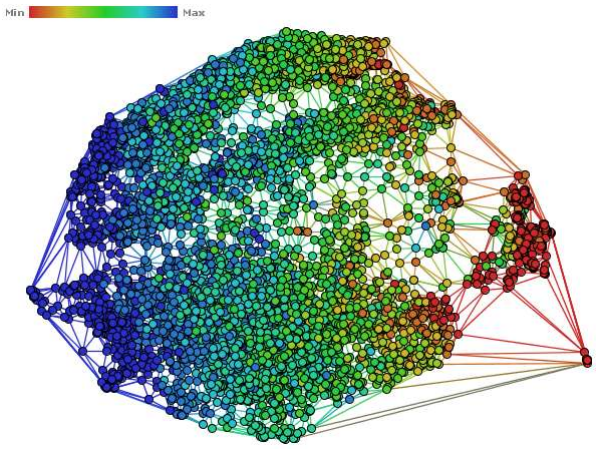
112

Cattle performance data



113

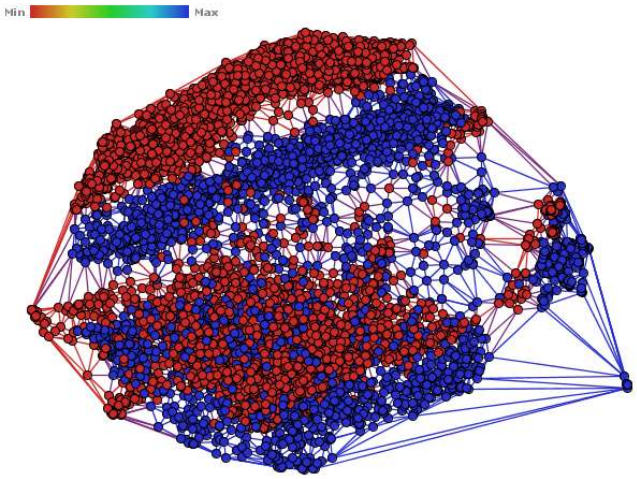
Cattle performance data



Colored by word 'top'

114

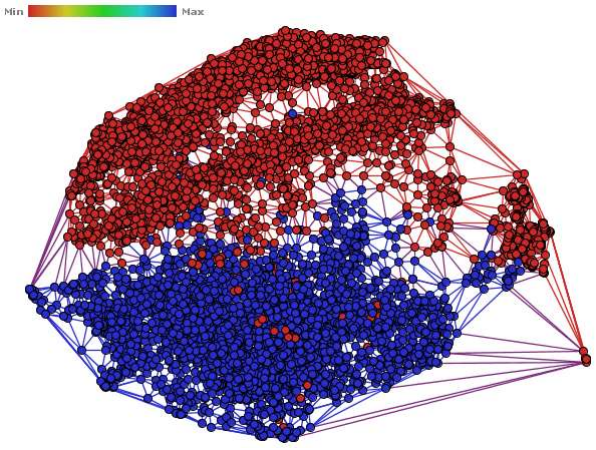
Cattle performance data



Colored by female

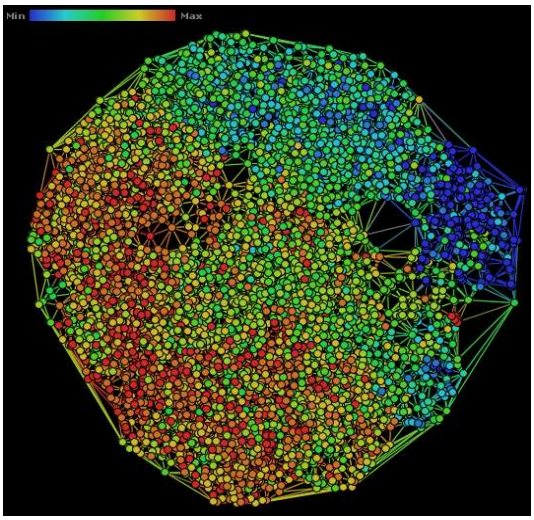
115

Cattle performance data



Colored by farm

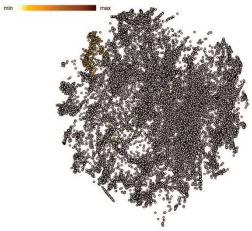
Cattle performance data

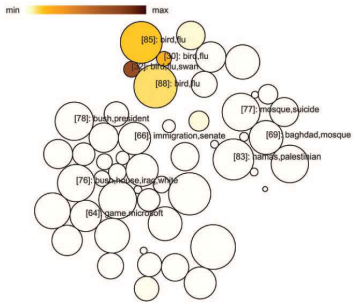


Colored by word 'top'

Handling scalability? Xhipp

Partitioning and Projection

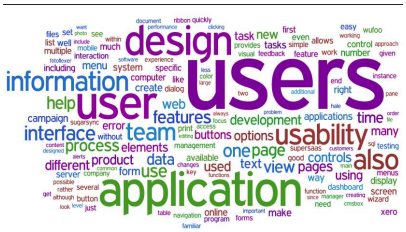




F. V. Paulovich and R. Minghim, "HiPP: A Novel Hierarchical Point Placement Strategy and its Application to the Exploration of Document Collections," in ***IEEE Transactions on Visualization and Computer Graphics***, vol. 14, no. 6, pp. 1229-1236, Nov.-Dec. 2008.

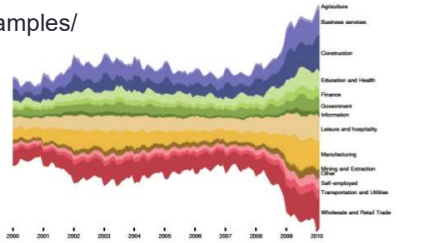
F. Dias and R. Minghim, "xHiPP: eXtended Hierarchical Point Placement Strategy",***31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)***, Parana, Brazil, pp. 361-368, 2018, IEEE CS Press.

Tag-clouds and Theme River



Stacked graph
aka Steam graph, Theme river

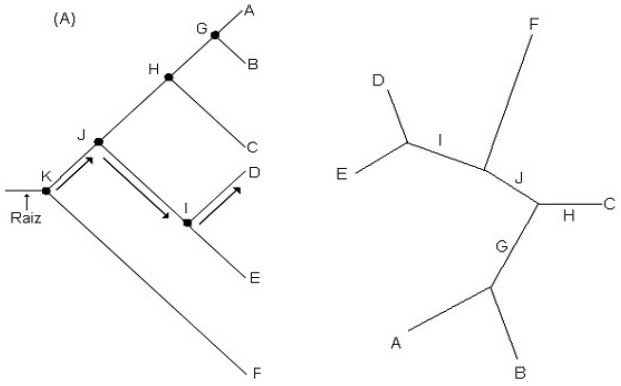
https://www.nngroup.com/articles/tag-cloud-examples/



Heer et al, 2010

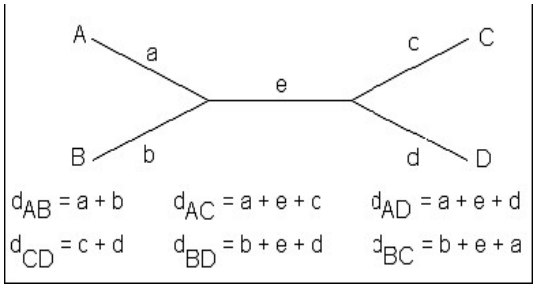
http://complexdatavisualized.com/time-series-visualizations-an-overview/

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



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

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



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)

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

   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

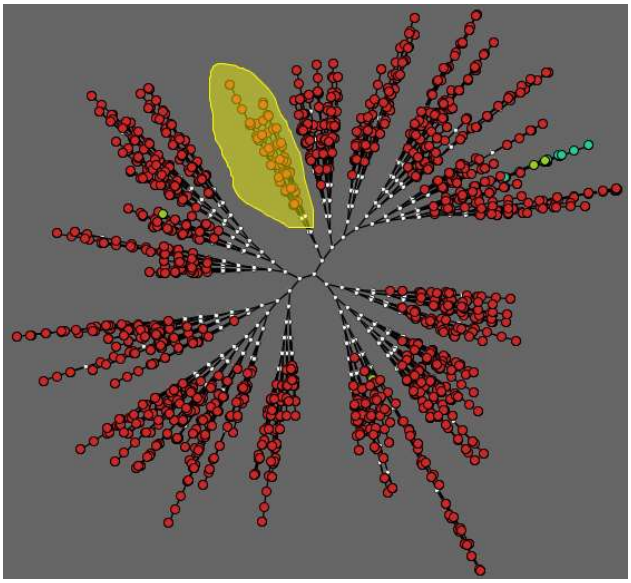
3. Calculate new distance matrix, computing the new distances from the new node to the remaining nodes.

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

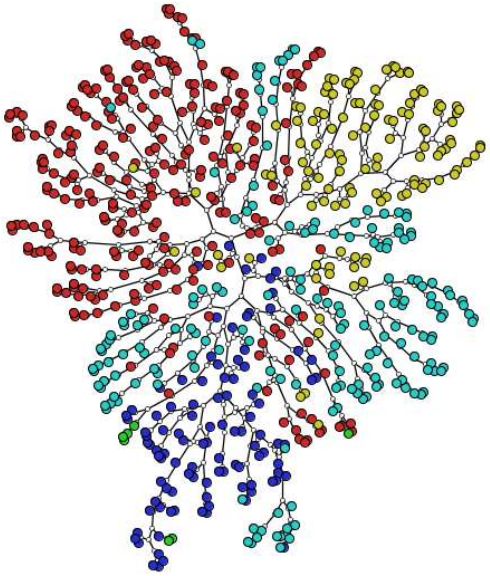
4. Eliminate previous nodes i and j
5. If n>2 then iterate again.

- Alternate view (N-J Tree)

124

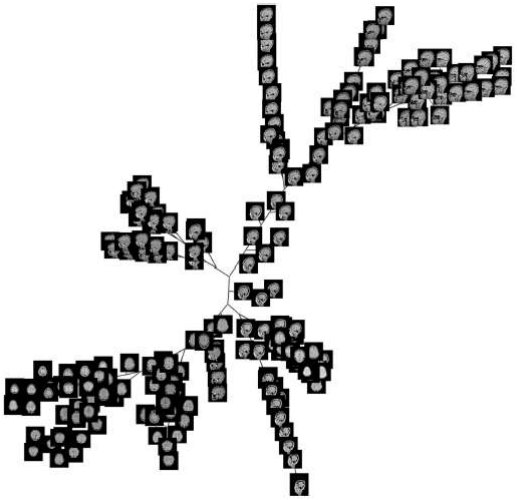


125



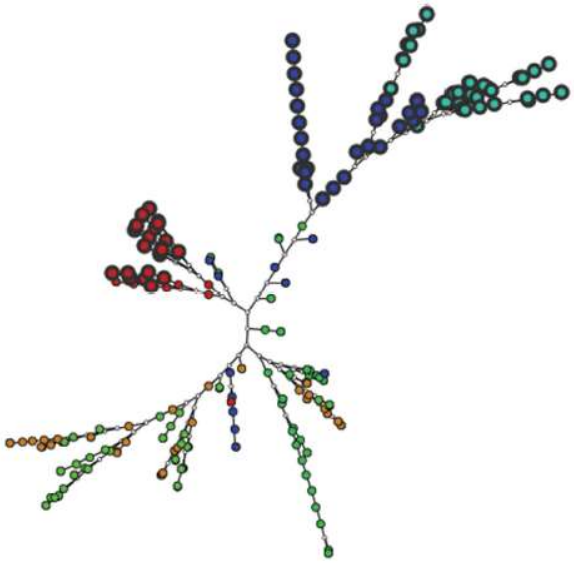
126

Images

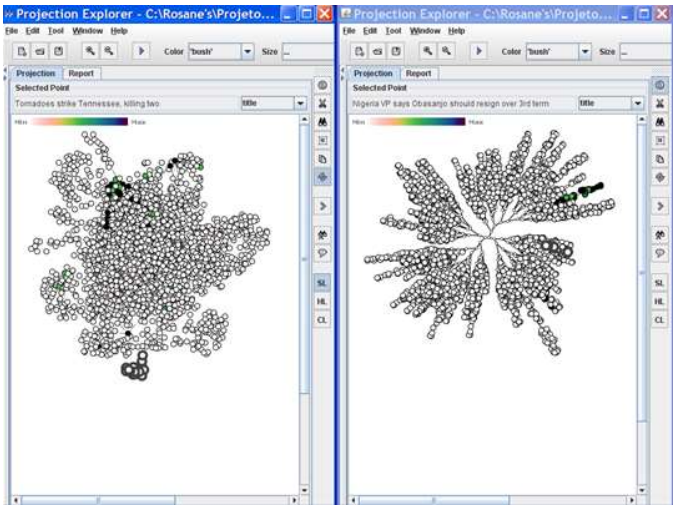


127

Images

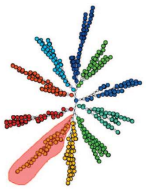


Projections & Trees

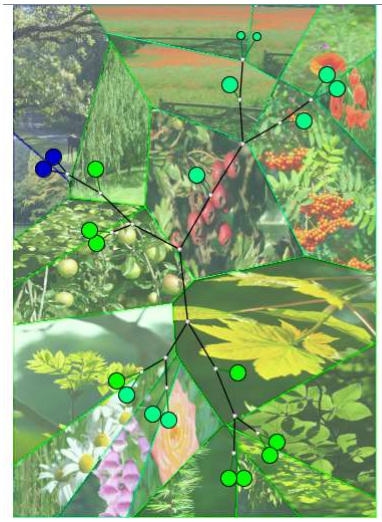


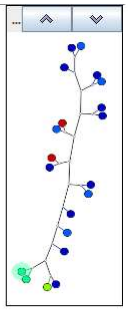
Application to Visual Data Classification

- Sample selection
- Classification
- Evolution of models
- Cooperation IC/UNICAMP
  - Helio Pedrini
  - William Schwartz (now UFMG)
- *Aplications: GPS data, Systems biology Data, data on quality of text*
  - *Cooperations with UNESP / Presidente Prudente, LNBio/CNPEM, NILC/ICMC*



Scalability
The Visual Super Tree



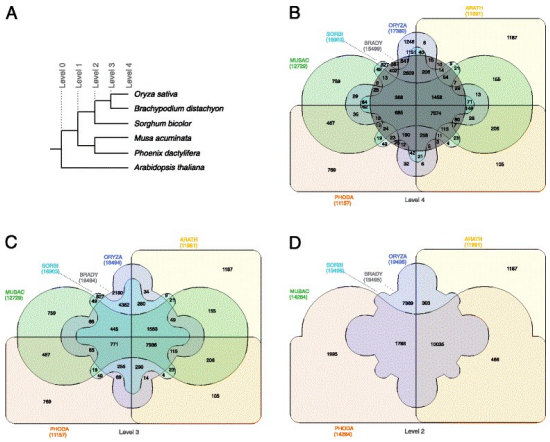


Cooperation:

Guilherme Pimentel Telles
IC/UNICAMP

Application: comparison of sets

- Cooperation ICMC/UNICAMP/LNBio (Campinas)/ Embrapa (Campinas)
- Fig.: Comparison of gene lists from different species
- www.interactiVenn.net



Context: Visual Data Mining

- Definition [Ankerst 2000]
  - step in process of knowledge discovery / extraction (KDD)
  - utilizes visualization as communication channel between computer and user
  - to support identification of new and interpretable patterns

Homework 3

- Explore the data sets left in Moodle-USP using:
  - Vispipeline (After Sept 19th)
  - Any tool available (go fetch!!)
- For the news data set:
  - Mention 5 headlines of importance
  - Describe generally what happened regarding each one.
- Create or obtain a new text or image data set.
  - Format using .data or .dmat (and .zip, if text) for Vispipeline (see pex-manual for that)
  - Explore using both projections and trees.
  - Write and illustrate your findings in two pages.

References

- Cuadros, A. M, Paulovich, F. V., Minghim, R., Telles, G. P - Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections IEEE VAST 2007, Sacramento, CA, USA, IEEE CS Press, pp.99-106.
- Paulovih, F. V., Oliveira, M.C.F., Minghim, R. - The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization, IEEE Sibgrapi 2007, IEEE CS Press, Belo Horizonte, Brazil,pp. 27-34.
- Lopes, A. A., Minghim, R., Melo, V., Paulovich, F.V.; Mapping texts through dimensionality reduction and visualization techniques for interactive exploration of document collections, **SPIE Conference on Visualization and Data Analysis**, San Jose, CA, USA Jan. 2006, 6060T-11.
- Minghim, R., Paulovich, F.V., Lopes, A. A.; Content-based text mapping using multidimensional projections for exploration of document collections, **SPIE Conference on Visualization and Data Analysis,** San Jose, CA, USA Jan. 2006, 6060T-11.

References

- Pinho, R. D. ; Oliveira, M. C. F. ; Minghim, R. ; Andrade, M. G. . Voromap: A Voronoi-based Tool for Visual Exploration of Multidimensional Data. In: **10th International Conference on Information Visualization**, 2006, Londres. Proceedings of Information Visualisation 2006, 2006. v. 1. p. 39-44
- Paulovich, F. V. ; Minghim, R. . Text Map Explorer: a Tool to Create and Explore Document Maps. In: Information Visualisation 2006 (IV06) **10th International Conference on Information Visualisation**, 2006, Londres. Proceedings of Information Visualisation 2006, 2006. v. 1. p. 245-251.
- Paulovich, F. V. ; Nonato, L. G. ; MINGHIM, R. ; Levkowitz, H. . Least Square Projection: a fast high precision multidimensional projection technique and its application to document mapping. IEEE Transactions on Visualization and Computer Graphics, 2008.
- Minghim, R. ; Levkowitz, H. ; Nonato, L. G. ; Watanabe, L. S. ; Salvador, V. C. L. ; Lopes, H. ; Pesco, S. ; Tavares, G. . Spider Cursor: A simple versatile interaction tool for data visualization and exploration. In: **ACM GRAPHITE'05** - 3rd International Conference on Computer Graphics and Interactive Techniques in Australasia and Southeast Asia, 2005, Dunedin. Proceedings of Graphite 2005, 2005. p. 307-314.
- Heberle, H.; Meirelles, G. V.; da Silva, F. R.; Telles, G. P.; Minghim, R. ***InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams***. BMC Bioinformatics 16:169 (2015).
- Carnielli, C.M.M; de Rossi, C.C.S; Granato, T.; Rivera, D. C; Domingues, R. R.; Pauletti, B. B.; Yokoo, S.; Heberle, H.; busso-lopes, ariane fidelis cervigne, nilva karla sawazaki-calone, iris meirelles, gabriela vaz marchi, fábio albuquerque telles, guilherme pimentel **minghim, rosane** ribeiro, ana carolina prado brandão, thaís bianca de castro, gilberto gonzález-arriagada, wilfredo alejandro gomes, alexandre penteado, fabio santos-silva, alan roger lopes, márcio ajudarte rodrigues, priscila campioni , sundquist, elias salo, tuula da silva, sabrina daniela alaoui-jamali, moulay a. graner, edgard fox, jay w. coletta, ricardo della paes leme, adriana franco ; Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nature Communications, v.9, 3598, 2018.

References – Biological Data

- Heberle, H.; Meirelles, G. V.; da Silva, F. R.; Telles, G. P.; Minghim, R. ***InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams***. BMC Bioinformatics 16:169 (2015).
- Kawahara, R. ; Meirelles, G. V. ; Heberle, H. ; DOMINGUES, R. R. ; GRANATO, D. C. ; YOKOO, S. ; CANEVAROLO, R. R. ; WINCK, F. V. ; RIBEIRO, A. C. P. ; BRANDAO, T. B. ; FILGUEIRAS, P. R. ; CRUZ, K. S. P. ; BARBUTO, J. A. ; POPPI, R. J. ; MINGHIM, R. ; Telles, G. P. ; FONSECA, F. P. ; FOX, J. W. ; SANTOS-SILVA, A. R. ; COLETTA, R. D. ; SHERMAN, N. E. ; Leme, A. F. P. . Integrative analysis to select cancer candidate biomarkers to targeted validation. OncoTarget, p.43635-43652, 2017.
- 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, 395, 2017.
- Carnielli, C.M.M; de Rossi, C.C.S; Granato, T.; Rivera, D. C; Domingues, R. R.; Pauletti, B. B.; Yokoo, S.; Heberle, H.; busso-lopes, ariane fidelis cervigne, nilva karla sawazaki-calone, iris meirelles, gabriela vaz marchi, fábio albuquerque telles, guilherme pimentel **minghim, rosane,** ribeiro, ana carolina prado brandão, thaís bianca de castro, gilberto gonzález-arriagada, wilfredo alejandro gomes, alexandre penteado, fabio santos-silva, alan roger lopes, márcio ajudarte rodrigues, priscila campioni , sundquist, elias salo, tuula da silva, sabrina daniela alaoui-jamali, moulay a. graner, edgard fox, jay w. coletta, ricardo della paes leme, adriana franco ; Combining discovery and targeted proteomics reveals a prognostic signature in oral cancer. Nature Communications, v.9, 3598, 2018.