Universidad Nacional del Altiplano
Escuela de Posgrado
Doctorado en Ciencias de la Computación

Data Mining

Unit 3. Visual Data Mining

Prof. Dr. Ivar Vargas Belizario

ivargasbelizario@gmail.com

2024 - I

Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Visualization in Data Mining

Visual Data Mining: An Introduction and Overview (2008)

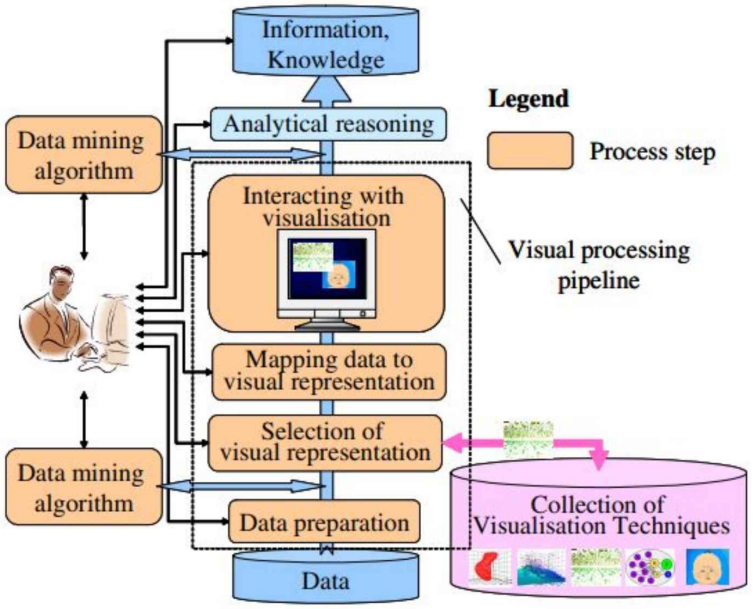
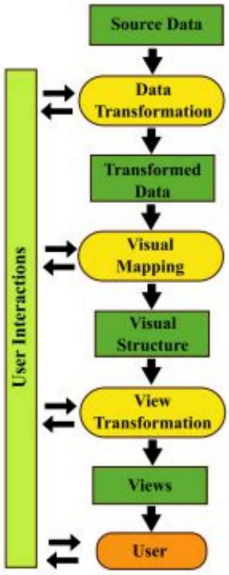


Fig. 1. Visual data mining as a human-centred interactive analytical and discovery process

[20] https://doi.org/10.1007/978-3-540-71080-6_1

Visualization in Data Mining

Visualizing High-Dimensional Data: Advances in the Past Decade (2017)



Data Transformation	**Dimension Reduction** linear projection [21], [22], non-linear DR [23], [24], Control Points Projection [25], [26], Distance Metric [27], [28], Precision Measures [29], [30]	**Subspace Clustering** Dimension Space Exploration [2], [31], [32], Subset of Dimension [33], [34], Non-Axis-Parallel Subspace [35], [36], [37]	**Regression Analysis** Optimization Design Steering [38], [39], [40], Structural Summaries [5], [41]	**Topological Data Analysis** Morse-Smale Complex [42], [43], [44], [45] Reeb Graph [46], [47], [48] Contour Tree [49], [50], Topological Features [51], [52]		
Visual Mapping	**Axis Based** Scatterplot Matrix [53], Parallel Coordinate [54], Radial Layout [55], Hybrid Construction [56], [57], [58], [59]	**Glyphs** Per-Element Glyphs [60], [61] [62], [63], Multi-Object Glyphs [64], [65], [66]	**Pixel-Oriented** Jigsaw Map [67], Pixel Bar Charts [68], Circle Segment [69], Value & Relation Display [70]	**Hierarchy Based** Dimension Hierarchy [71], Topology-based Hierarchy [72], [73], Others [74], [75]	**Animation** GGobi [76], TripAdvisorND [77], Rolling the Dice [78]	**Evaluation** Scatterplot Guideline [79], [80], PCPs Effectiveness [81], Radical Layout [82], Animation [83]
View Transformation	**Illustrative Rendering** Illustrative PCP [84], Illuminated 3D Scatterplot [85], PCP Transfer Function [86] Magic Lens [87], [88]	**Continuous Visual Representation** Continuous Scatterplot [89], [90], Continuous Parallel Coordiante [91], [92], Splatterplots [93] Splatting for PCPs [94]	**Accurate Color Blending** Hue-Preserving Blending [95], Weaving vs. Blending [96]	**Image Space Metrics** Clutter Reduction [97], [98], Pargnostics [99], Pixnostic [100]		

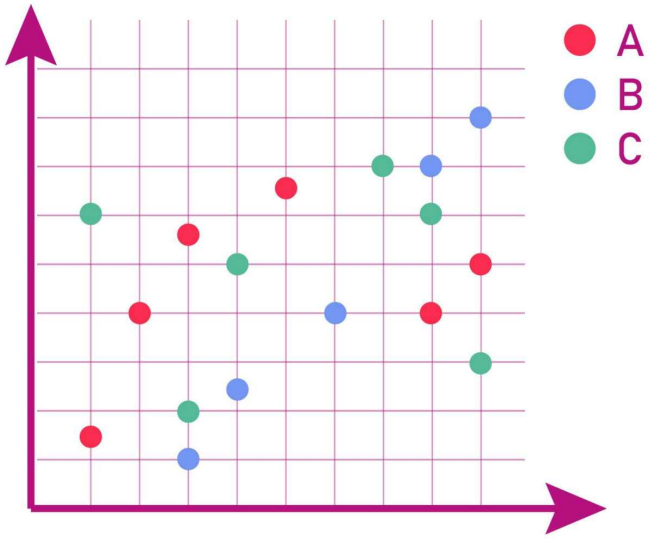
Fig. 1. Research categorization based on different stages of the visualization pipeline, with subcategories that reflect common approaches.

[21]https://doi.org/10.1109/TVCG.2016.2640960

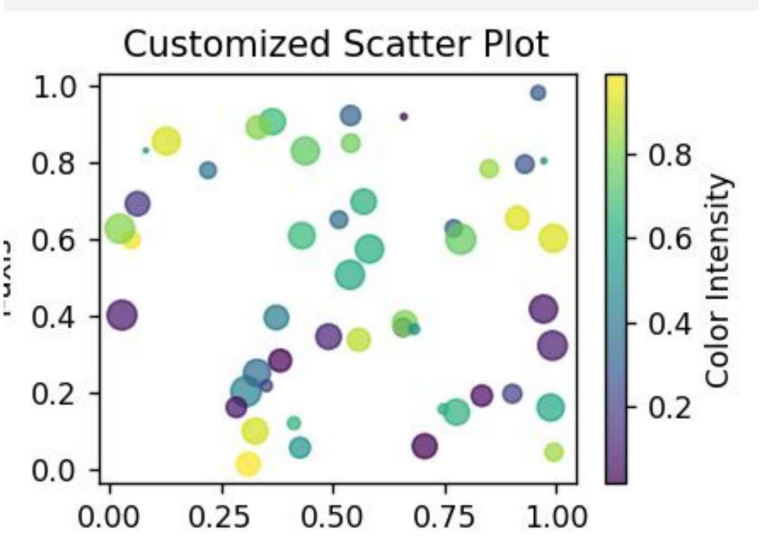
Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

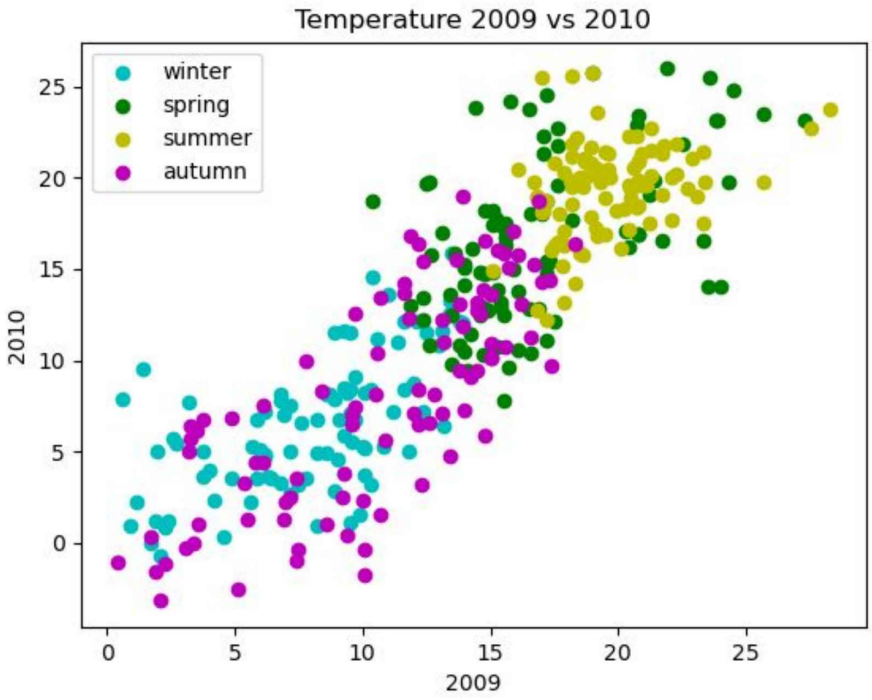
Scatter Plot



Scatter Plot



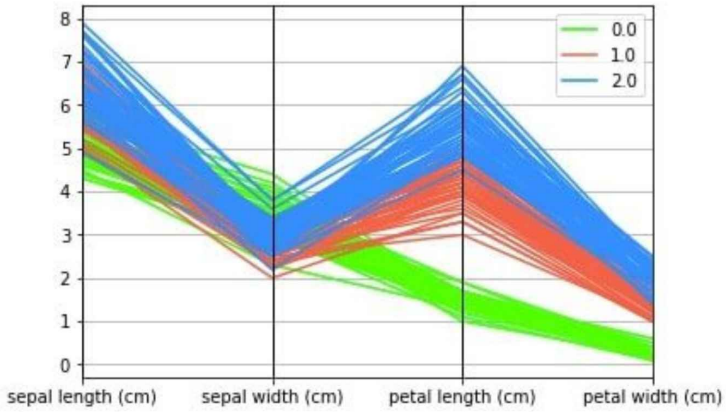
Scatter Plot



Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

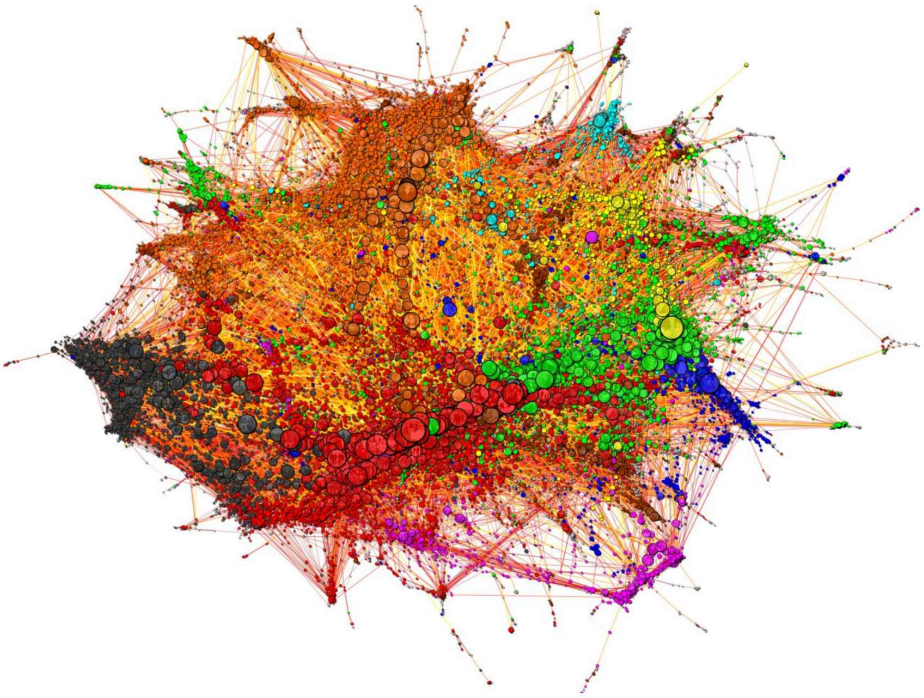
Parallel Coordinates



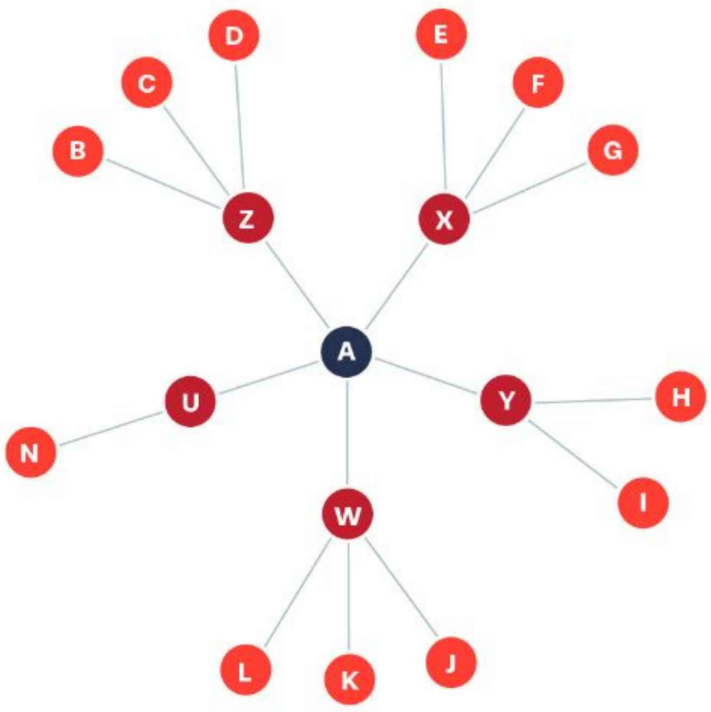
Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Graphs and Trees



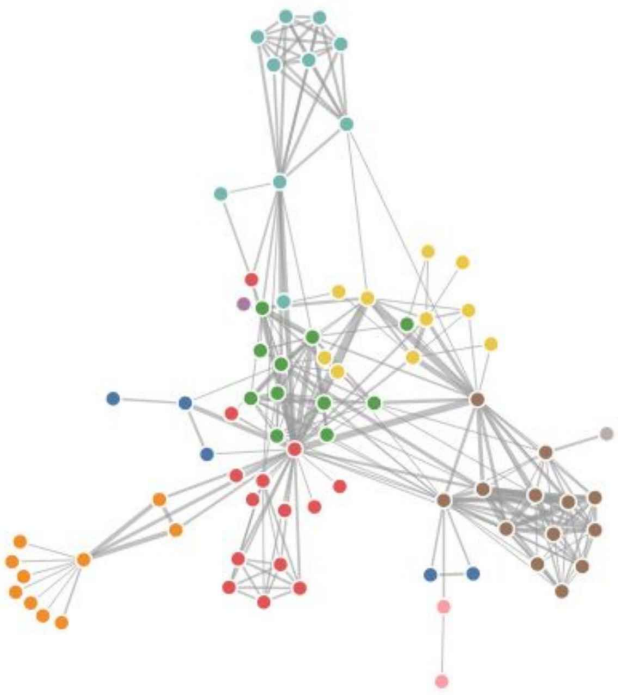
Graphs and Trees



Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Force-directed

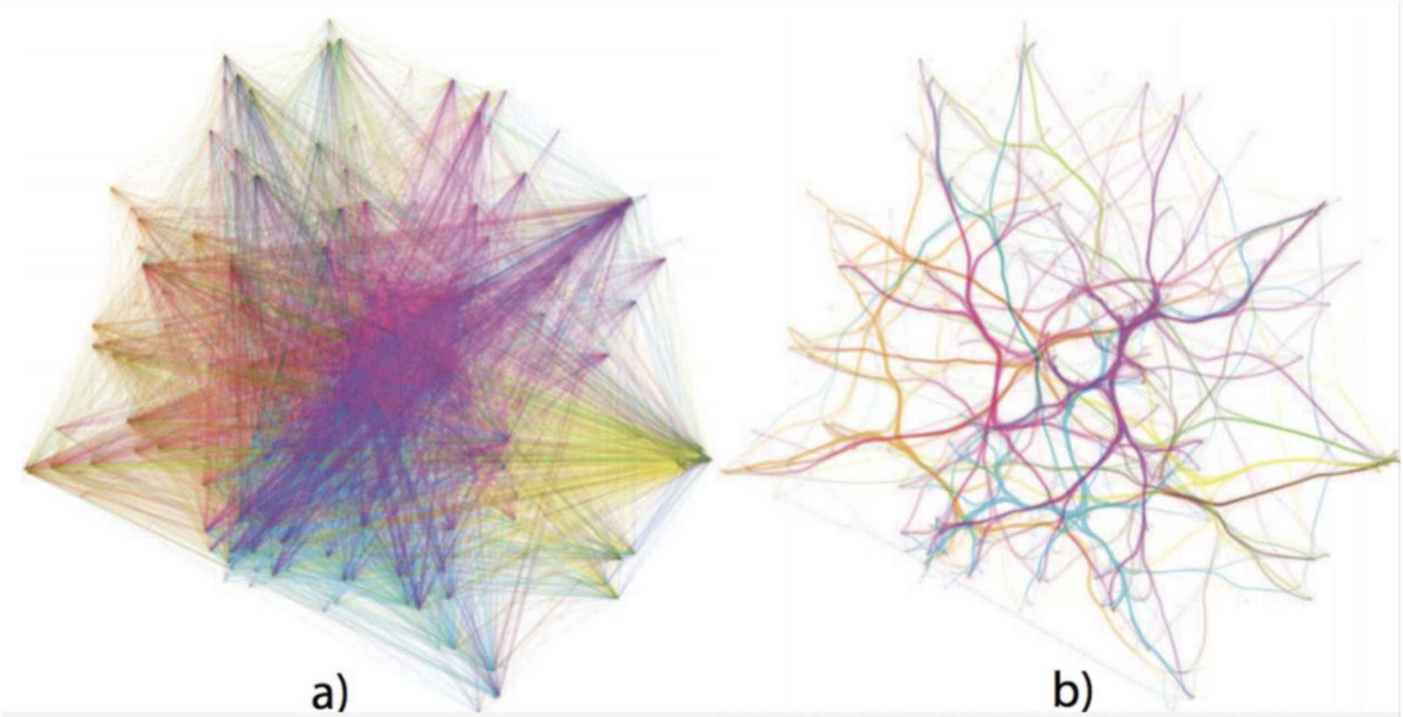


https://observablehq.com/@d3/force-directed-graph-component

Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

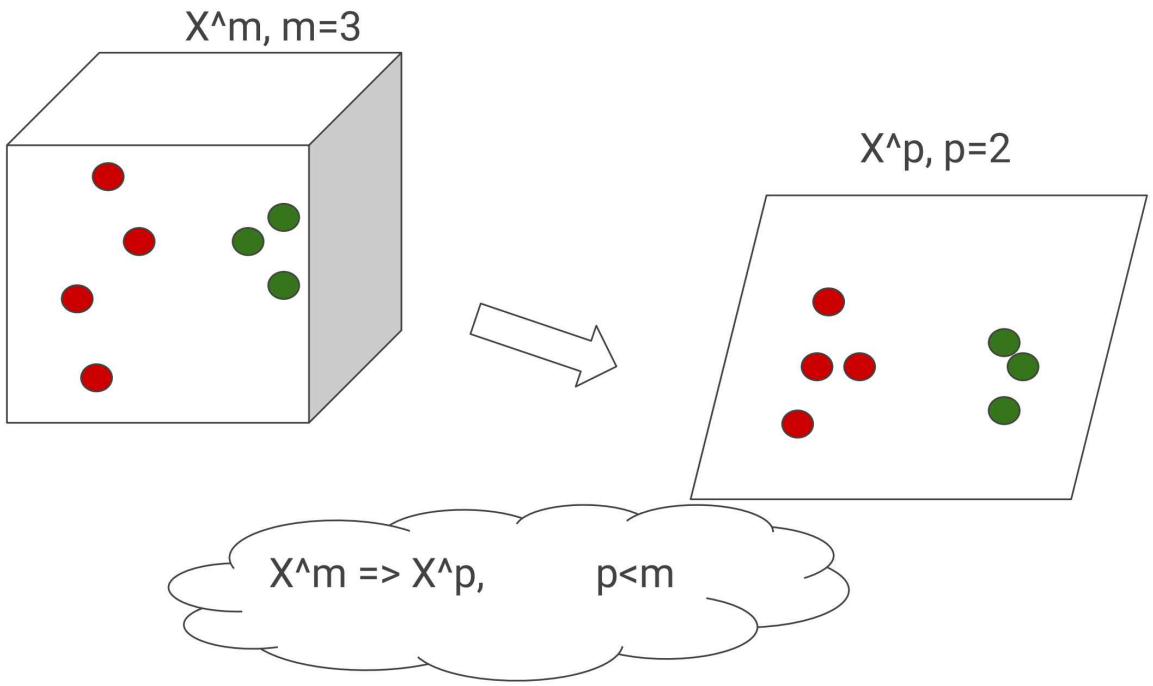
Edge bundling



Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Multidimensional Projections - Dimensionality Reduction



Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Examples

The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization

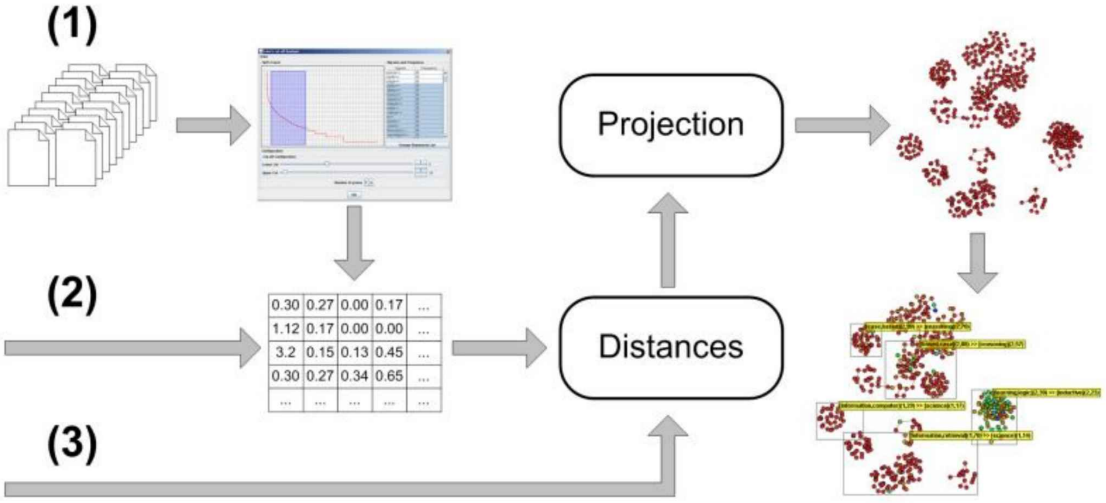


Figure 3. Generating projection maps of multidimensional data with PEx.

[22] https://doi.org/10.1109/SIBGRAPI.2007.21

Examples

The Projection Explorer: A Flexible Tool for Projection-based Multidimensional Visualization

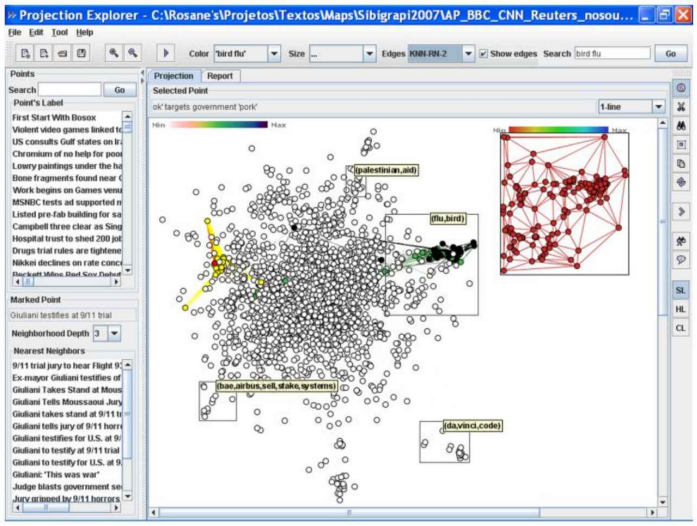


Figure 1. Main window of Projection Explorer (PEx): document map depicts a collection of online news.

[22] https://doi.org/10.1109/SIBGRAPI.2007.21

Examples

Point Placement by Phylogenetic Trees and its Application to Visual Analysis of Document Collections

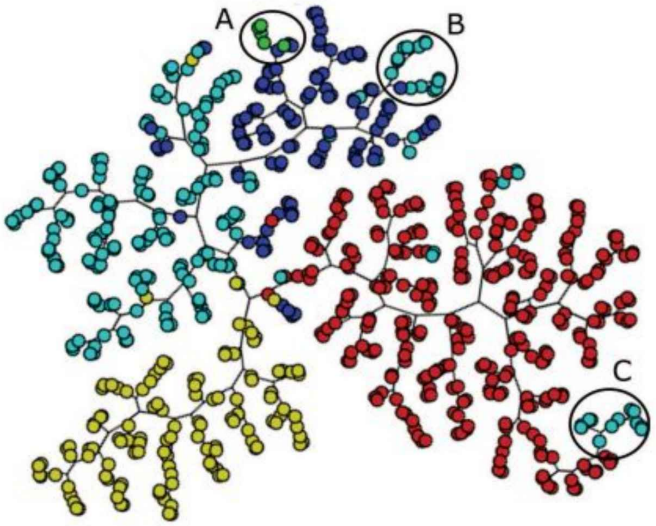


Figure 1: CBR+ILP+IR+SON document map. Points represent the documents, colored according to the area they belong to.

[23] https://doi.org/10.1109/VAST.2007.4389002

Examples

Multidimensional Visualization to Support Analysis of Image Collections

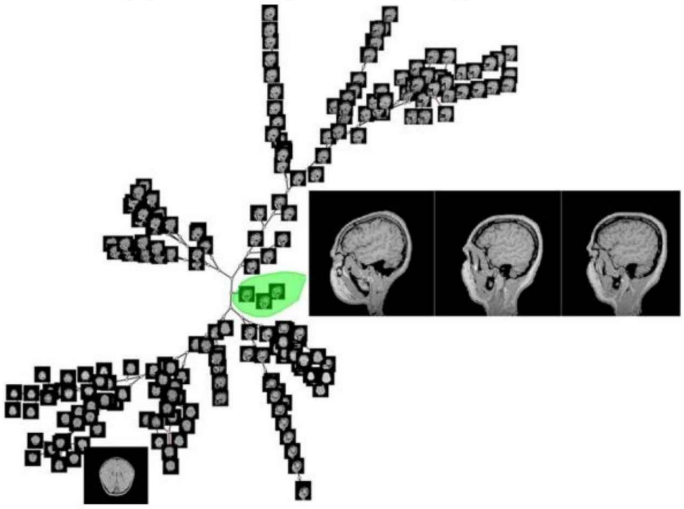
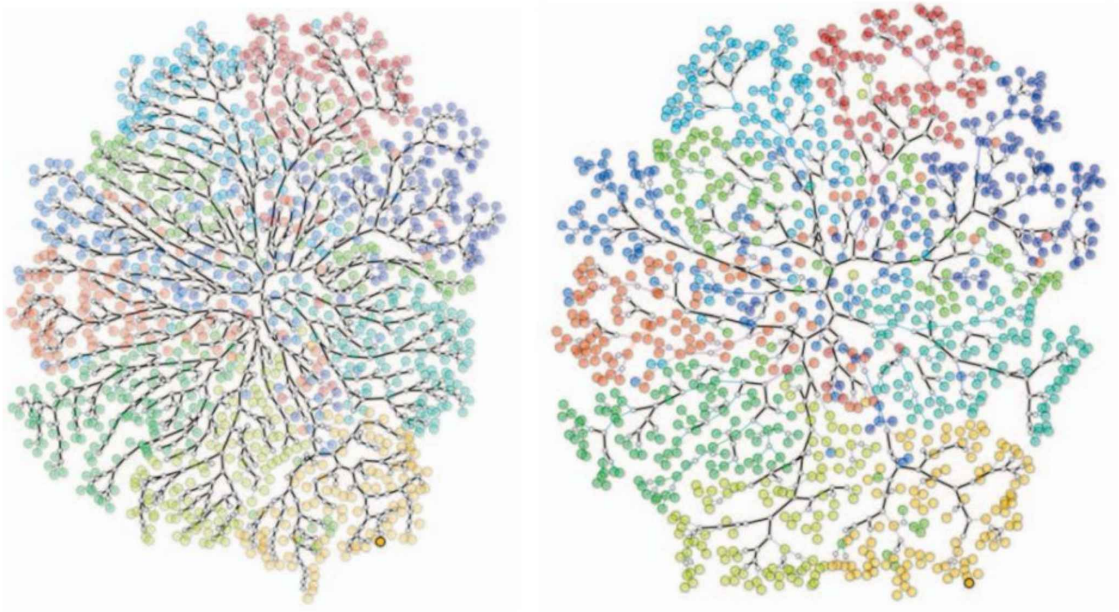


Figure 1. 2D projection of a medical image data set and details. Layout generated with the Neighbor-Joining technique.

[24] https://doi.org/10.1109/SIBGRAPI.2008.30

Examples

Improved Similarity Trees and their Application to Visual Data Classification

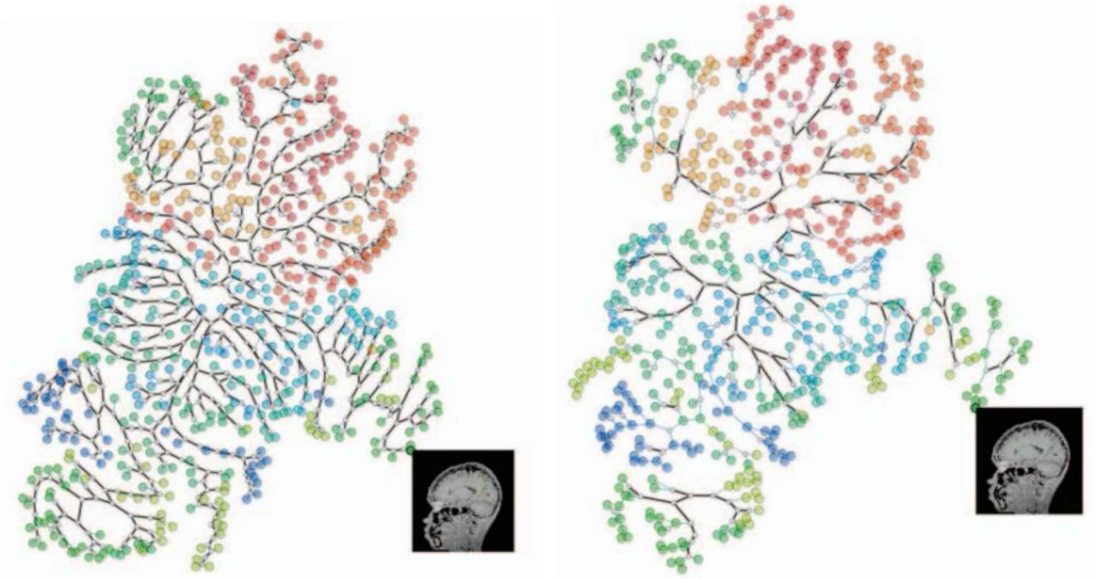


(a) NJ for the COREL data set. (b) PNJ for the COREL data set.

[25] https://doi.org/10.1109/TVCG.2011.212

Examples

Improved Similarity Trees and their Application to Visual Data Classification



(c) NJ for the MEDICAL data set. (d) PNJ for the MEDICAL data set.

[25] https://doi.org/10.1109/TVCG.2011.212

Examples

Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping

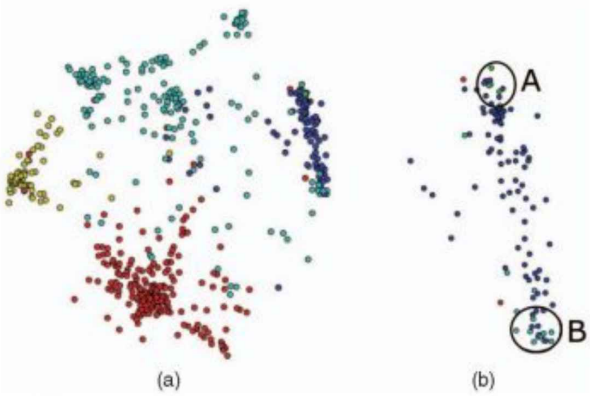


Fig. 2. Projection of a document collection composed of scientific papers in four different areas (colors indicate the areas). (a) Whole map. (b) Zoomed part.

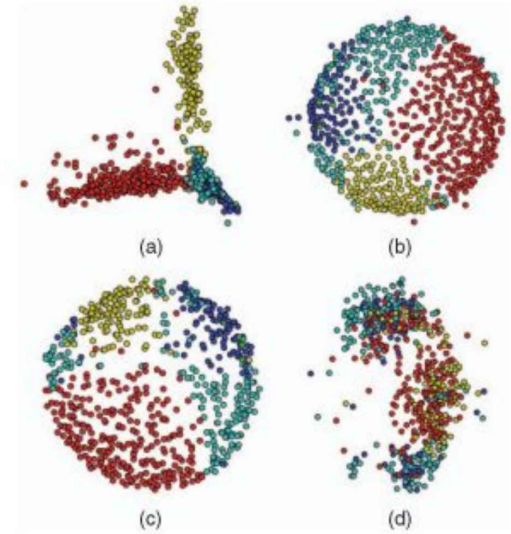


Fig. 11. Examples of projections generated using different techniques for the same data set used in the LSP projection presented in Fig. 2a. (a) PCA. (b) Sammon's mapping. (c) Original FDP model. (d) Approximated FDP model.

[26] https://doi.org/10.1109/TVCG.2007.70443

Examples

Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping

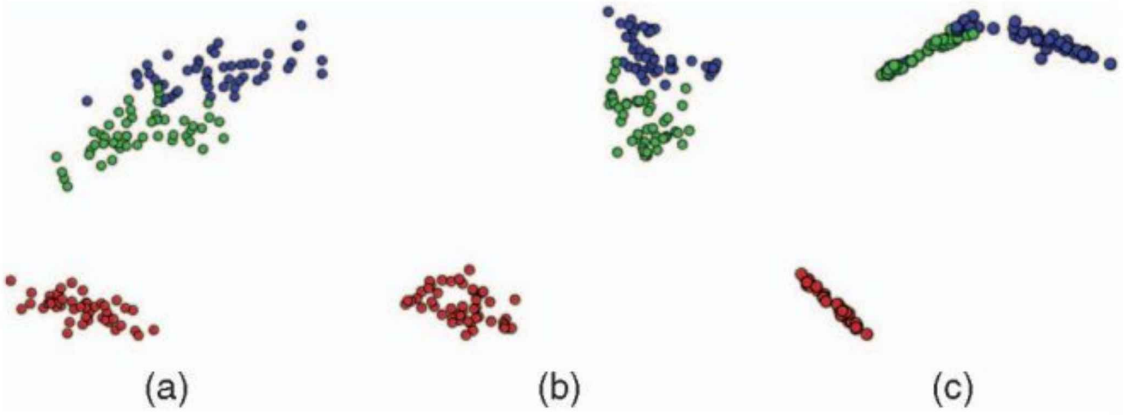
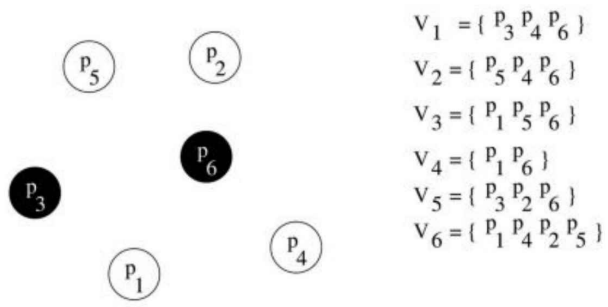


Fig. 7. Comparing projections of the Iris data set using the Force Scheme on all points and the LSP with different numbers of control points. (a) Force Algorithm applied to all points. (b) LSP with 50 percent of control points. (c) LSP with 10 percent of control points.

[26] https://doi.org/10.1109/TVCG.2007.70443

Examples

Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping



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

Fig. 1. Matrix A with control points p_3 and p_6.

[26] https://doi.org/10.1109/TVCG.2007.70443

Examples

Least Square Projection: A Fast High-Precision Multidimensional Projection Technique and Its Application to Document Mapping

4.2 Evaluating the Results

A common way of evaluating the quality of a projection in order to compare then analytically is known as *stress* [36]. Stress aims at measuring the amount of information lost during projection as the difference between the dissimilarities in the m-dimensional space and the distances in the d-dimensional space. The stress function defined by Kruskal [36] is presented as follows:

$$stress = \sqrt{\frac{\sum_{i<j}(d(f(x_i), f(x_j)) - \delta(x_i, x_j))^2}{\sum_{i<j} d(f(x_i), f(x_j))^2}}. \quad (5)$$

[26] https://doi.org/10.1109/TVCG.2007.70443

Examples

Visual Data Exploration to Feature Space Definition

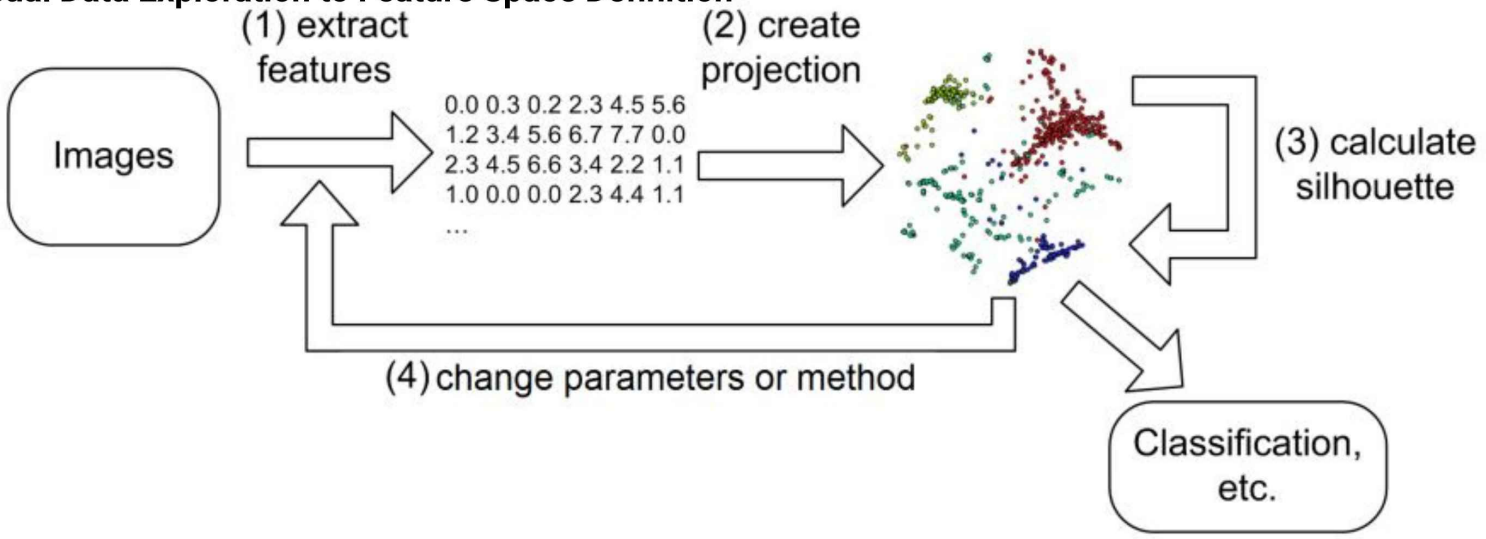
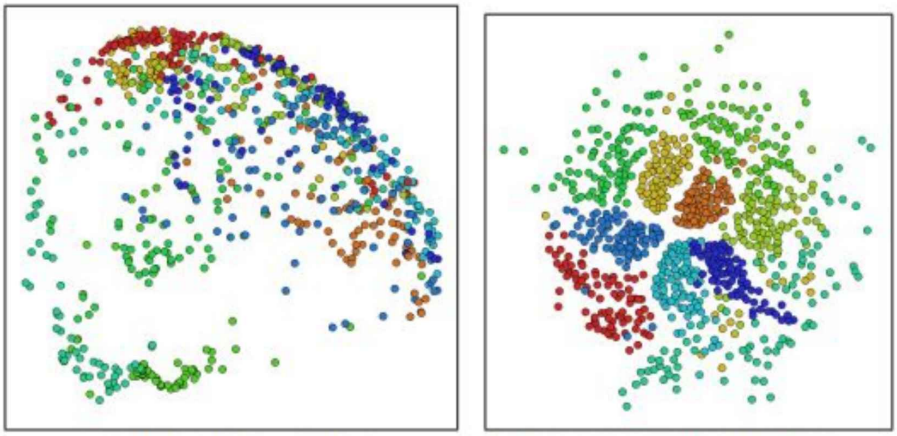


Figure 1. Proposed visual exploration of the feature space.

[27] https://doi.org/10.1109/SIBGRAPI.2010.13

Examples

Visual Data Exploration to Feature Space Definition



(f) Combined (ALL) (g) Dimensionality Reduction

Figure 8. Projections of different feature spaces extracted from **KTH-TIPS** database. Applying the PCA to reduce the dimensionality results in better classification rates and silhouette values, matching with class separability observed on the produced projection (colors indicates the classes).

[27] https://doi.org/10.1109/SIBGRAPI.2010.13

Examples

Part-Linear Multidimensional Projection (PLMP)

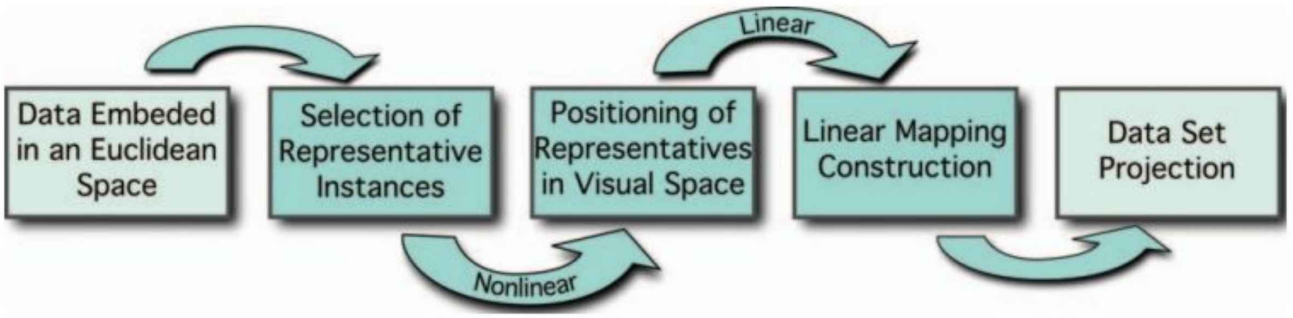


Fig. 2. PLMP pipeline: darker boxes represent the three main steps of the proposed technique.

[28] https://doi.org/10.1109/TVCG.2010.207

Examples

Part-Linear Multidimensional Projection (PLMP)

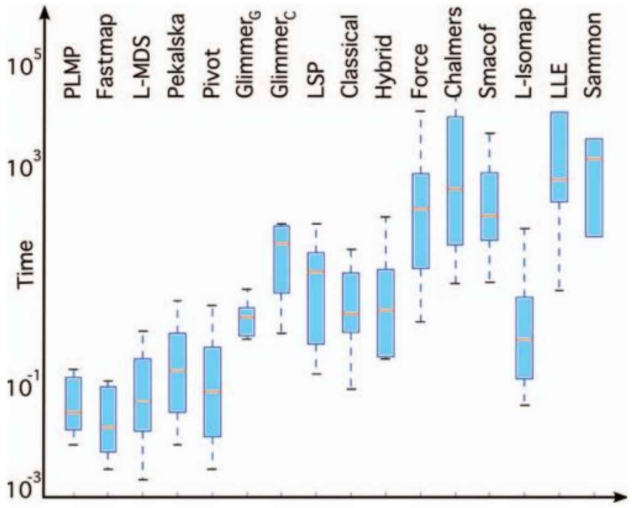


Fig. 4. Boxplot of times (log scale) shown in Table 2

[28] https://doi.org/10.1109/TVCG.2010.207

Examples

Part-Linear Multidimensional Projection (PLMP)

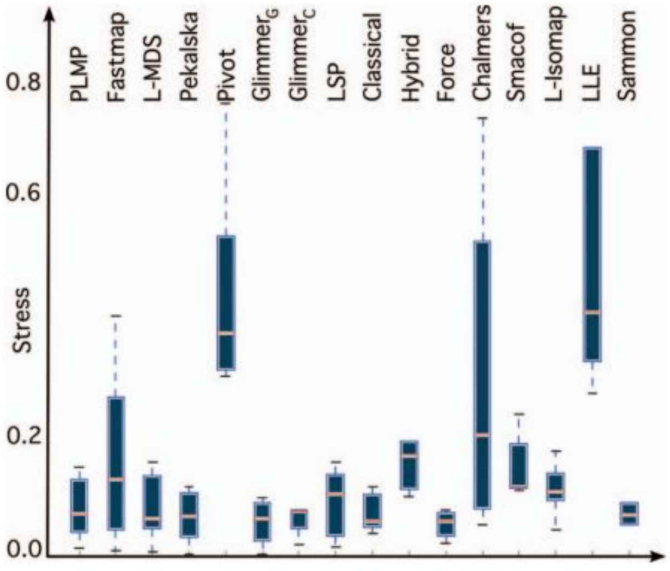


Fig. 5. Boxplot of stress shown in Table 2

[28] https://doi.org/10.1109/TVCG.2010.207

Examples

Part-Linear Multidimensional Projection (PLMP)

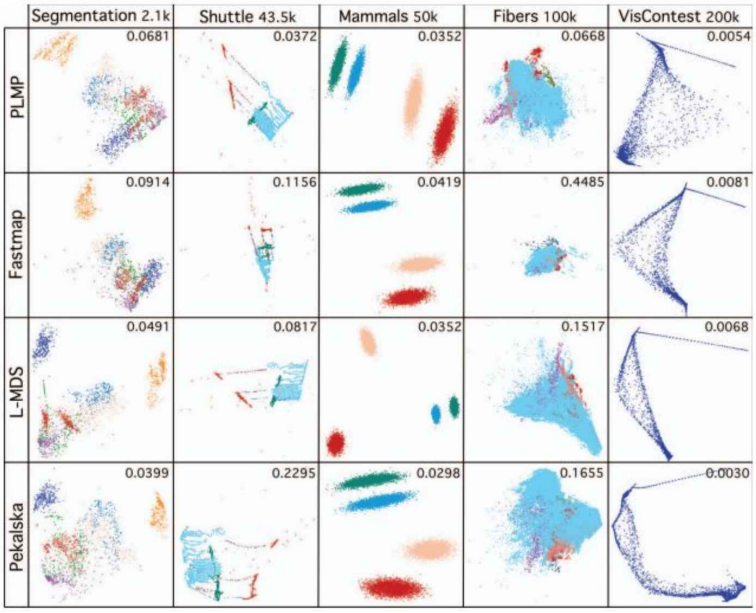


Fig. 8. Projection layouts for five different data sets. The first four data sets are endowed with labels which allow identification of similar instances. Labels are mapped as colors in the plots, providing a qualitative analysis of the projections. Stress is shown in the top right.

[28] https://doi.org/10.1109/TVCG.2010.207

Examples

Local Affine Multidimensional Projection

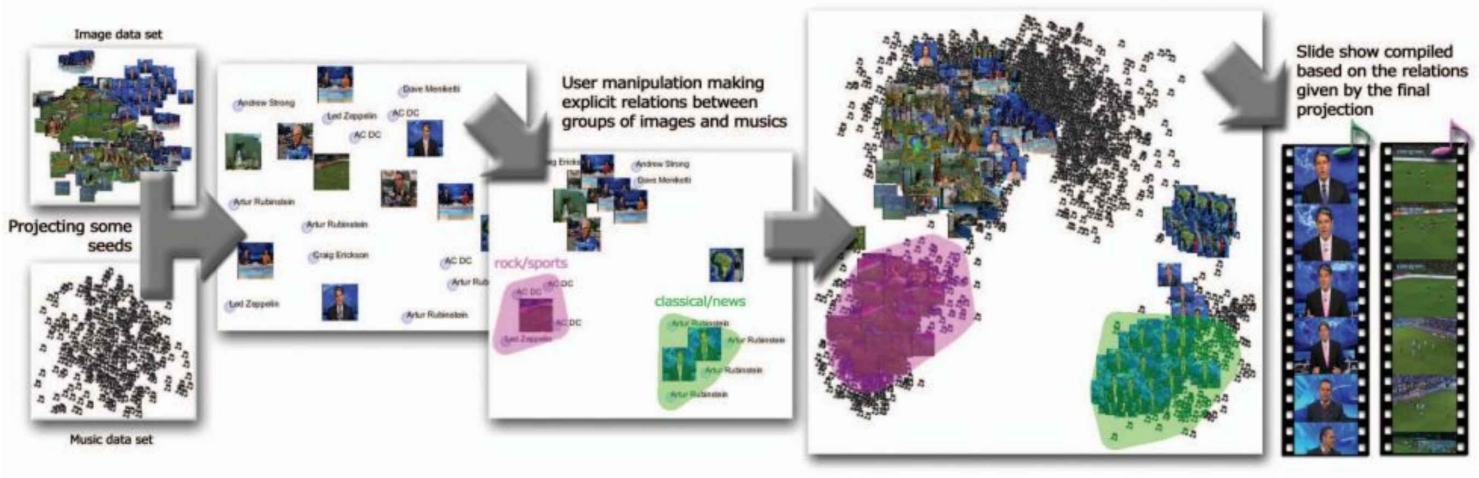
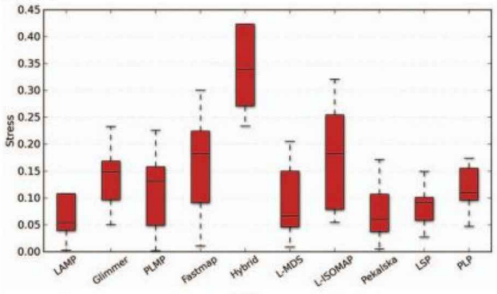


Fig. 1. Using projections to correlate different data sets that do not have explicit relation among instances. An initial projection is created using a few instances of each data set (music and images). Then, the relation amongst selected instances is defined by grouping images and music in the visual space, creating an explicit correlation. Considering this initial manipulation, the projections of the entire data sets are accomplished and the correspondence is settled. Finally, the lists of corresponding elements are used to produce slide shows where the images and related music are played in a synchronized manner.

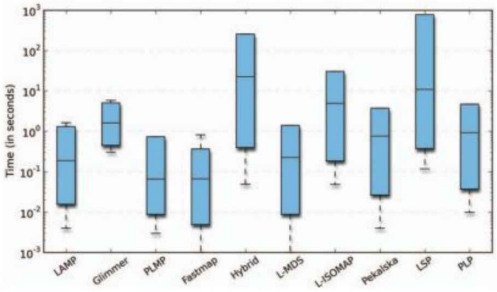
[29] https://doi.org/10.1109/TVCG.2011.220

Examples

Local Affine Multidimensional Projection



(a)



(b)

Fig. 5. Stress and computational times boxplots.

[29] https://doi.org/10.1109/TVCG.2011.220

Examples

Local Affine Multidimensional Projection

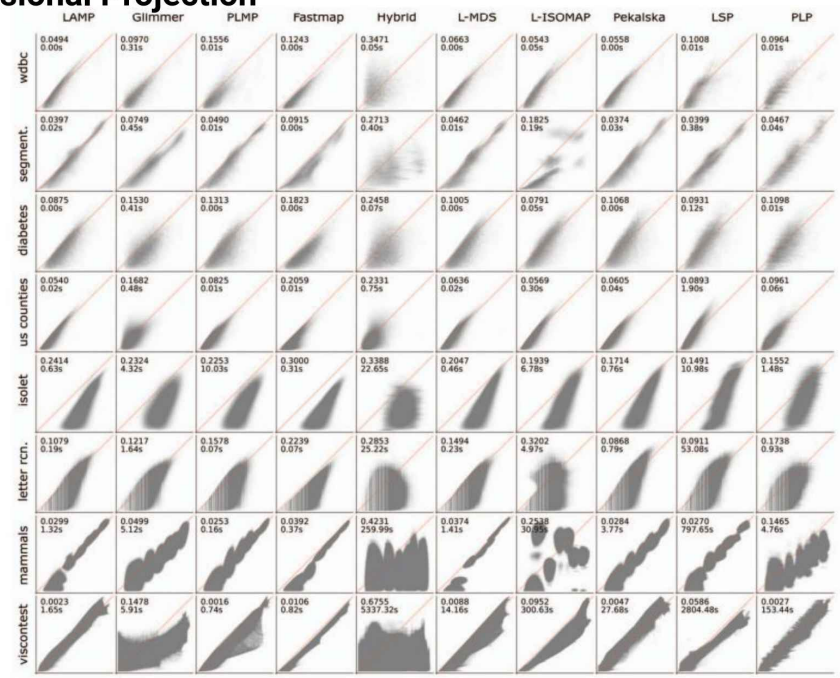


Fig. 6. *original-distance* × *projected-distance* scatter plots. Top-left numbers correspond to normalized stress and computational time (seconds).

[29] https://doi.org/10.1109/TVCG.2011.220

Examples

Colorization by Multidimensional Projection

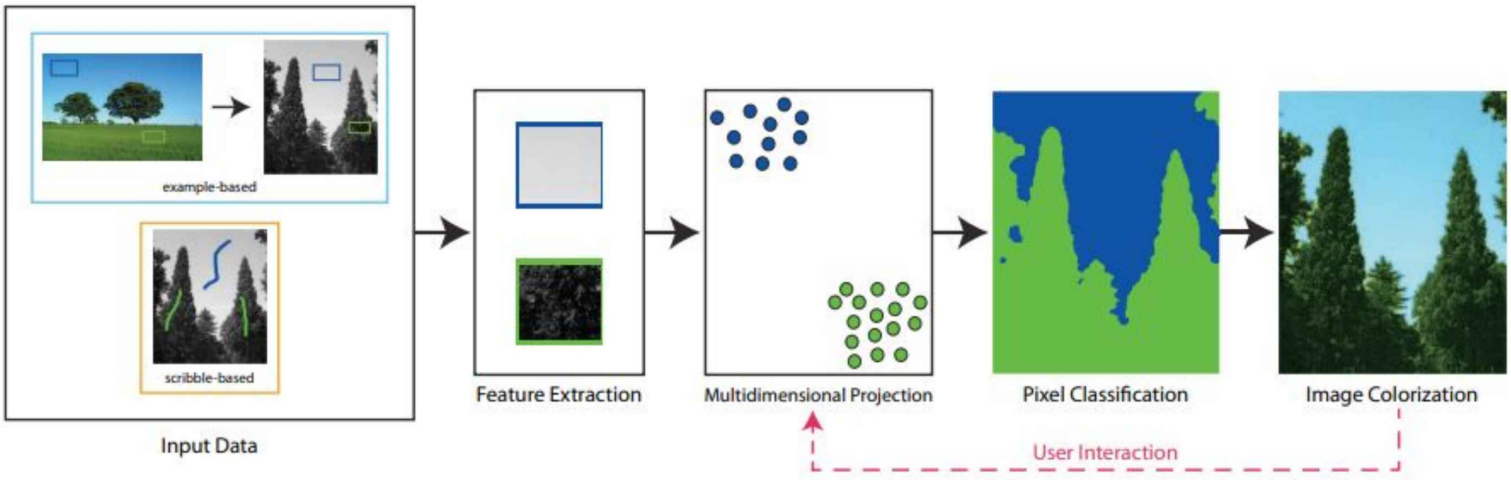
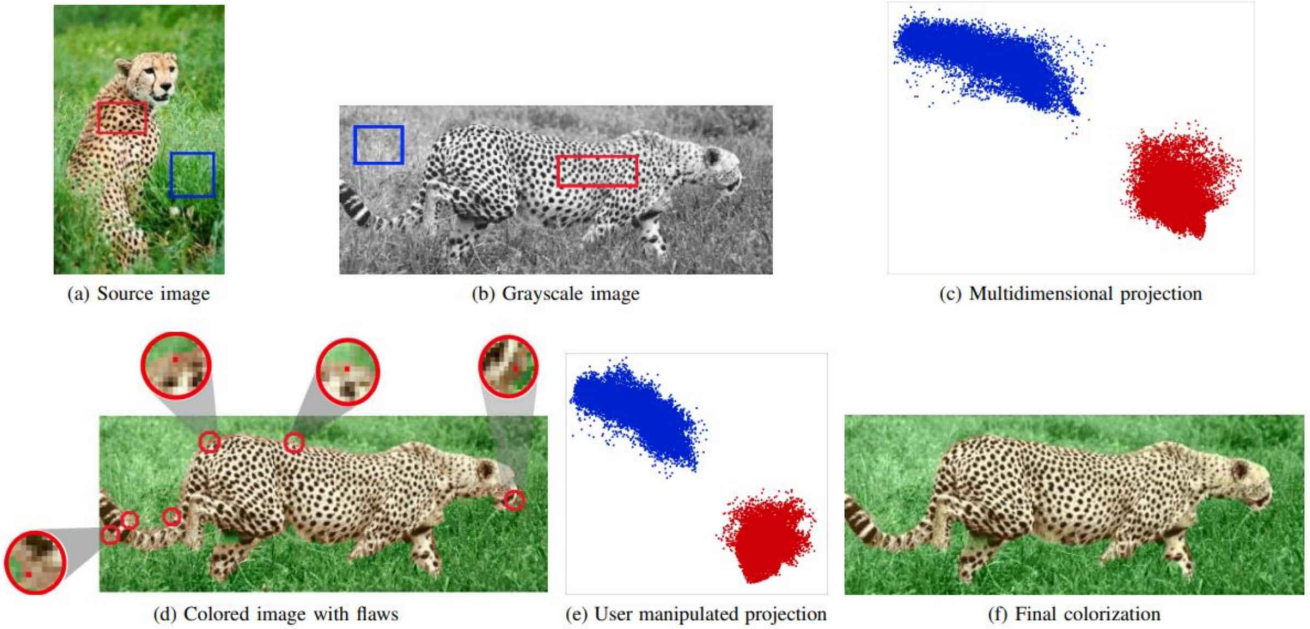


Fig. 2. Projcolor pipeline.

[30] https://doi.org/10.1109/SIBGRAPI.2012.14

Examples

Colorization by Multidimensional Projection



(a) Source image (b) Grayscale image (c) Multidimensional projection

(d) Colored image with flaws (e) User manipulated projection (f) Final colorization

Fig. 5. Neighborhood manipulation to further improving colorization.

[30] https://doi.org/10.1109/SIBGRAPI.2012.14

Examples

Graphs from Features: Tree-Based Graph Layout for Feature Analysis

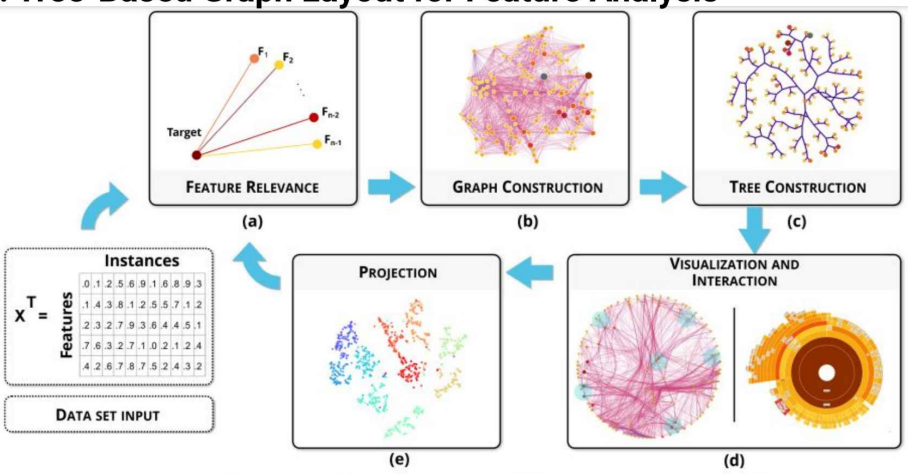


Figure 1. Analysis pipeline of Graphs from Features (GFF). From a dataset X with n data instances having d features each, (**a**) a proximity measure between pairs of features is evaluated, and (**b**) a complete graph is created where vertices represent features and edges represent dissimilarity between features. (**c**) A tree is constructed to summarize the information in the complete graph. (**d**) Visualization and interaction tools that combine information from the tree and from the underlying graph may be used by the user to search for patterns among the features and to select the most discriminating features. (**e**) Data instances may be projected based on the features selected by the user during the interaction.

[31] https://doi.org/10.3390/a13110302

Examples

Graphs from Features: Tree-Based Graph Layout for Feature Analysis

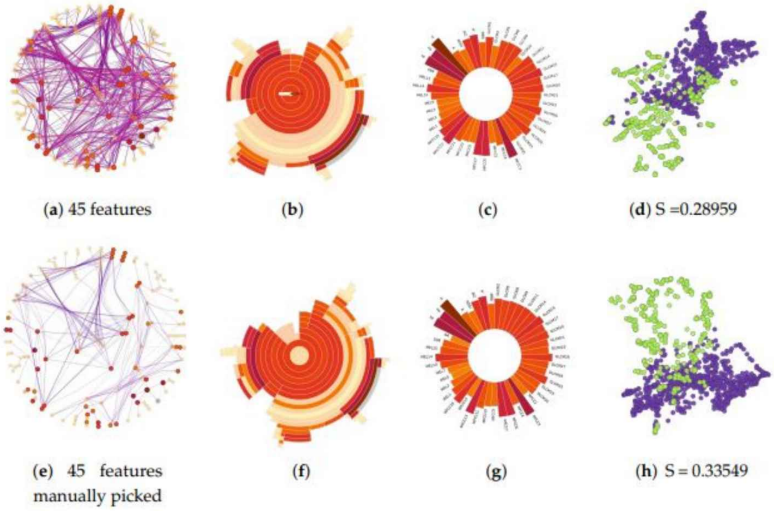


Figure 7. Soundscape ecology data with 141 features from bird and insect classes. (**a**) MST with additional graph edges highlighting 45 features selected automatically by ranked relevance and (**b**) Sunburst visualization of the features. (**c**) Circular histogram of the selected features. (**d**) The LSP projection of the 1662 data points from the space of selected features on 2D and its silhouette coefficient. (**e–h**) The same types of layouts for 45 features selected by a combination of calculated rank and interaction.

[31] https://doi.org/10.3390/a13110302

Contenido

- Visualization in Data Mining
- Visualization Techniques
  - Scatter Plot
  - Parallel Coordinates
  - Graphs and Trees
  - Force-directed
  - Edge bundling
- Multidimensional Projections
- Examples
- Tools

Tools

- D3.js (JavaScript): https://d3js.org/
- Java 2D (Java): https://docs.oracle.com/javase/tutorial/2d/
- Seaborn (Python): https://seaborn.pydata.org/

Universidad Nacional del Altiplano
Escuela de Posgrado
Doctorado en Ciencias de la Computación

Data Mining

Unit 3. Visual Data Mining

Gracias

Prof. Dr. Ivar Vargas Belizario

ivargasbelizario@gmail.com

2024 - I