Advanced Data Visualization

CS 6965

Spring 2018

Prof. Bei Wang Phillips

University of Utah

Lecture 07

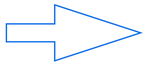
Regression Visual Mapping

HD

Clarification: data vs dim space

Each column
is a “dim” point

Each row
is a “data” point



Customer ID				
101				
102				

Review: Clustering and Vis

- Clustering points in the data space vs in the dim space
- Interplay of data manipulations either in the data space, the dim space or both

Additional Readings

- [WenskovitchCrandellRamakrishnan2017]: Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics Clustering
- [SachaZhangSedlmair2016]: Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis

Regression & Vis

Focus: the interplay between vis and regression analysis

Regression analysis + Vis

- Optimization and design steering (e.g., HyperMoVal)
  - Explore multiple output or response variables
  - The results require a qualitative examination
  - Results are used to inform decisions
- Structural summaries (e.g., HDViz)
  - Using regression to summarize data (e.g., skeleton representations)

HyperMoVal

HyperMoVal: Interactive Visual Validation of Regression Models for Real-Time Simulation

- Validating regression model against actual data
- Uses support vector regression (SVR) to fit a model to high-dim data
- Highlights discrepancies between the data and the model
- Computes sensitivity information on the model

[PiringerBergerKrasser2010]

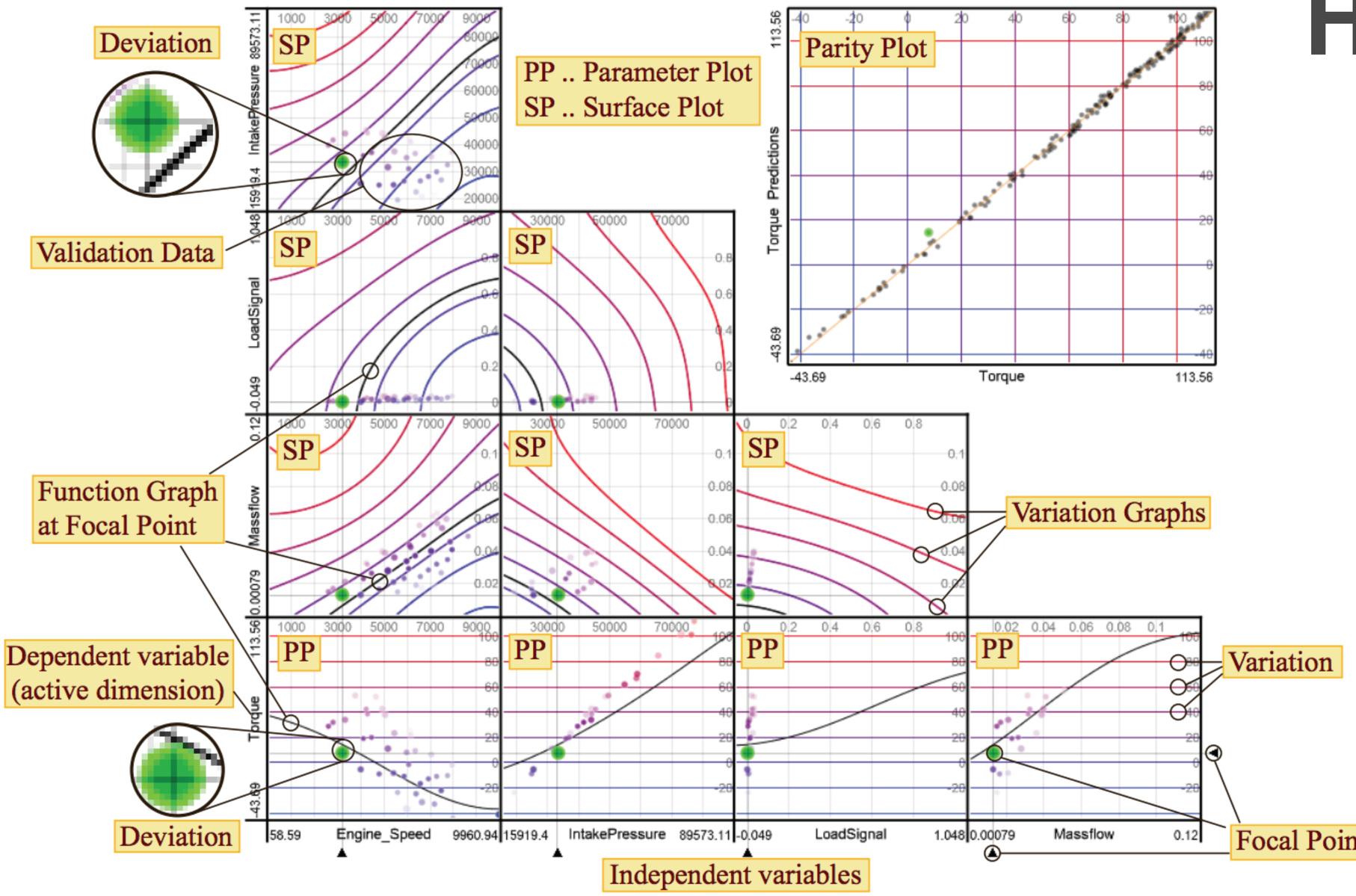
HyperMoVal: Model Validation

1. Comparing known and predicted results
2. Analyzing regions with a bad fit
3. Assessing the physical plausibility of models also outside regions covered by validation data
4. Comparing multiple models

The key idea is to visually relate one or more n-dimensional scalar functions to known validation data within a combined visualization.

[PiringerBergerKrasser2010]

HyperMoVal



[PiringerBergerKrasser2010]

Figure 1: *The layout of HyperMoVal for a real model predicting torque given four parameters. The focal point F is set to a validation data point with a significant deviation. The matrix contains all paraxial 2D slices at F in the 5D model space.*

HDViz

- Approximates a topological clustering (more on this later)
- Construct an inverse linear regression for each cluster of the data
- Regression is used as a post-processing step in order to present summaries of the extracted subsets of data.

[GerberBremerPacucci2010]

HDViz

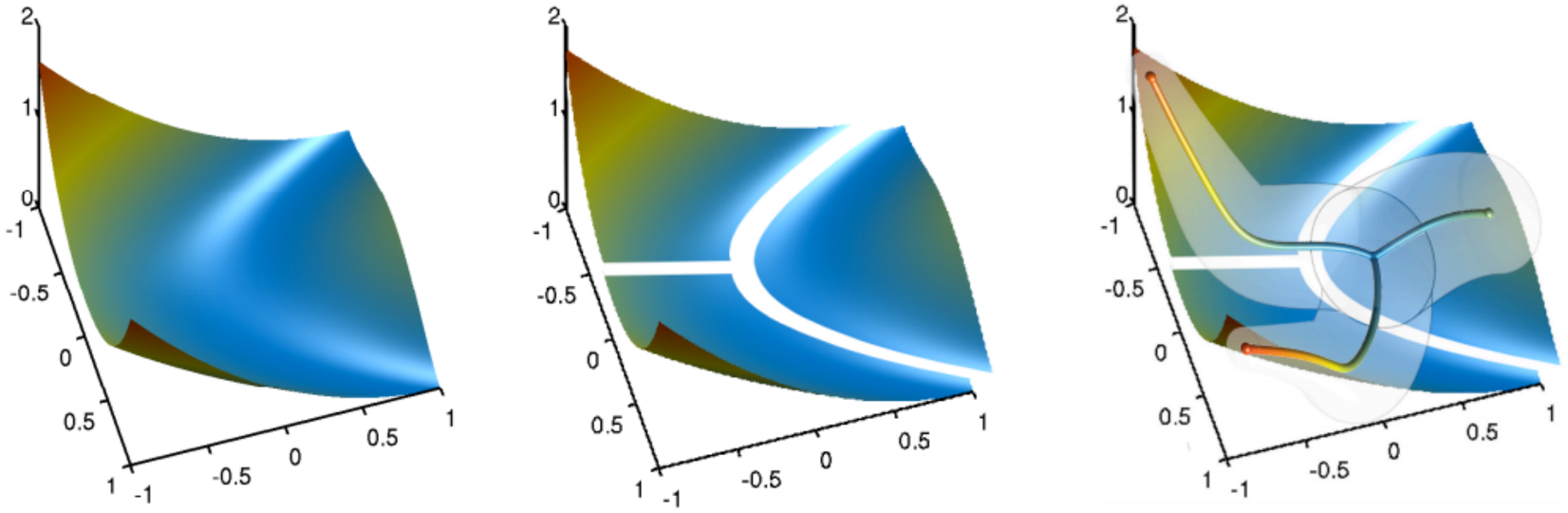
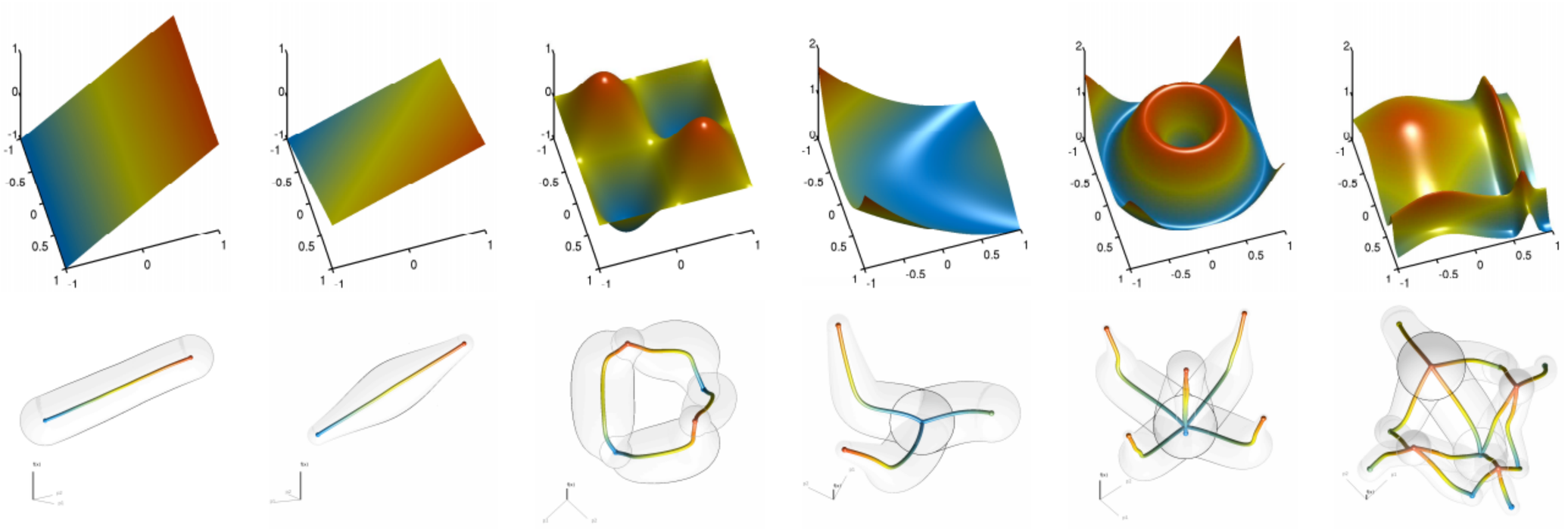


Fig. 3. Schematic illustration of the proposed method. The scalar function (a) is decomposed into piecewise monotonic regions (b) and each region is approximated by a regression curve (c).

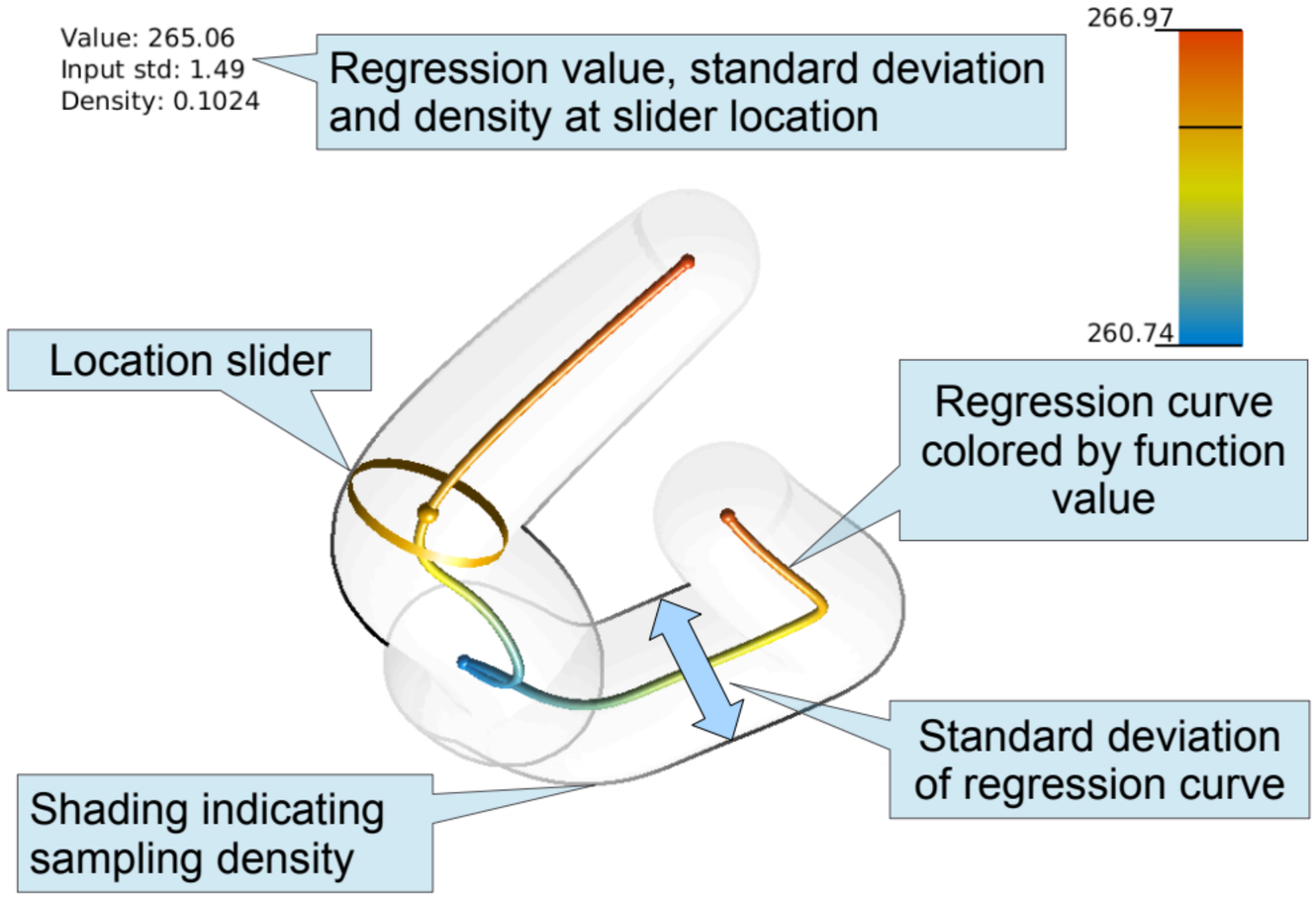
https://www.sci.utah.edu/software/hdvis.html [GerberBremerPacucci2010]

HDViz



https://www.sci.utah.edu/software/hdvis.html [GerberBremerPacucci2010]

HDViz



https://www.sci.utah.edu/software/hdvis.html [GerberBremerPacucci2010]

HDViz: Case Study Combustion

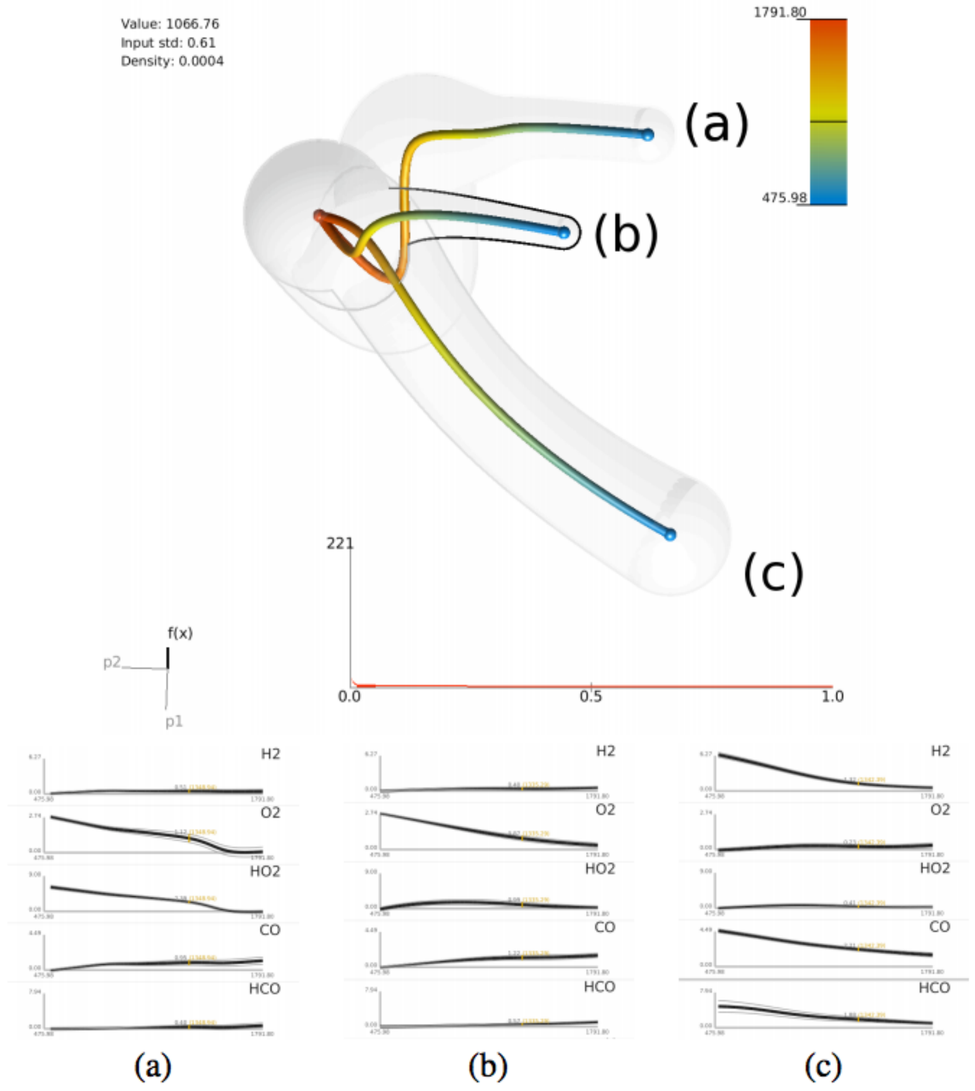
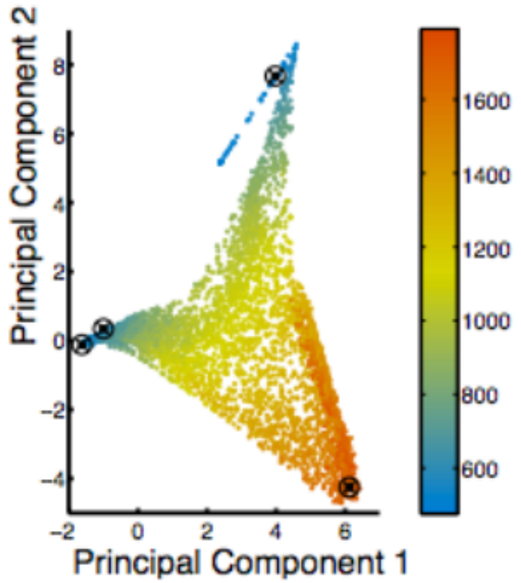


Fig. 15. Chemical composition in relation to heat released during a jet flame combustion simulation. The three distinct minima correspond to pure fuel, pure oxidizer and extinction/reignition. Graphs of chemical composition plotted against temperature for the crystals corresponding to extinction (a), pure oxidizer (b) and pure fuel (c) minima compositions.



[GerberBremerPacucci2010]

HDViz: Case Study Nuclear Simulation

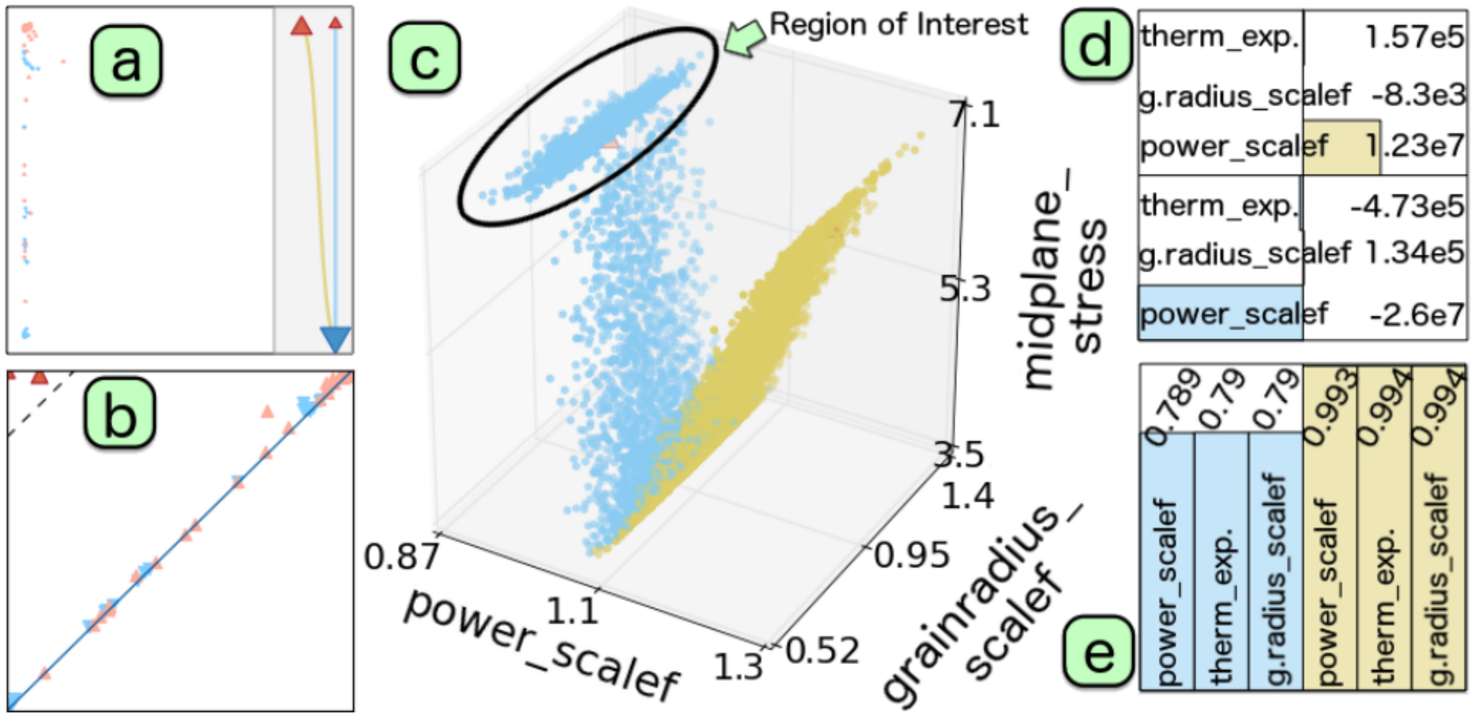
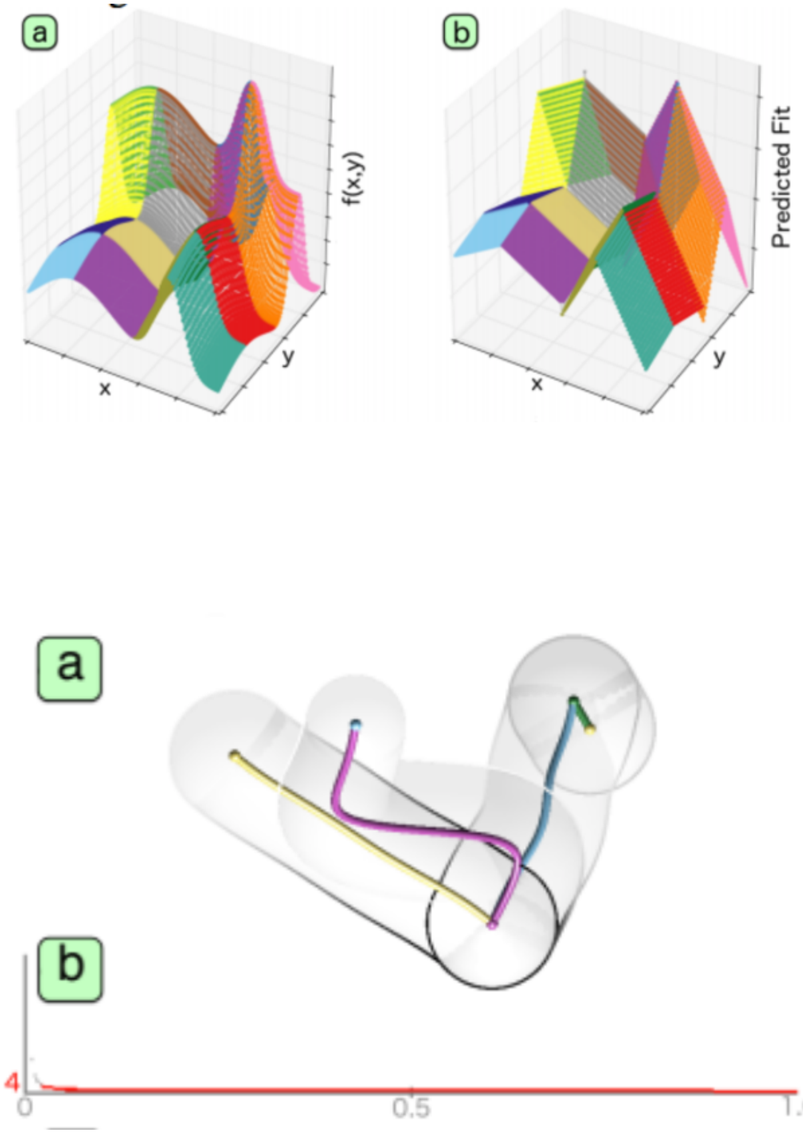


Figure 5: SA of the new nuclear fuel dataset: (a) topology map, (b) persistence diagram, (c) linked scatter plot projection, (d) linear coefficients, and (e) fitness view with stepwise R^2 scores.

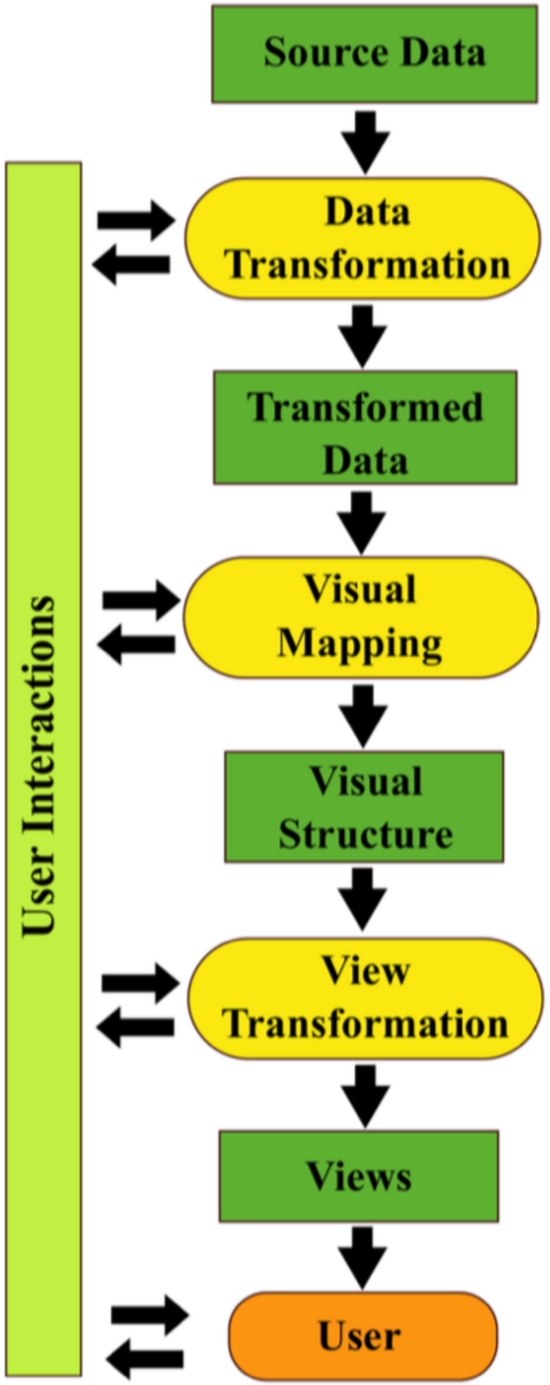


[MaljovecWangRosen2016]

Take home message...

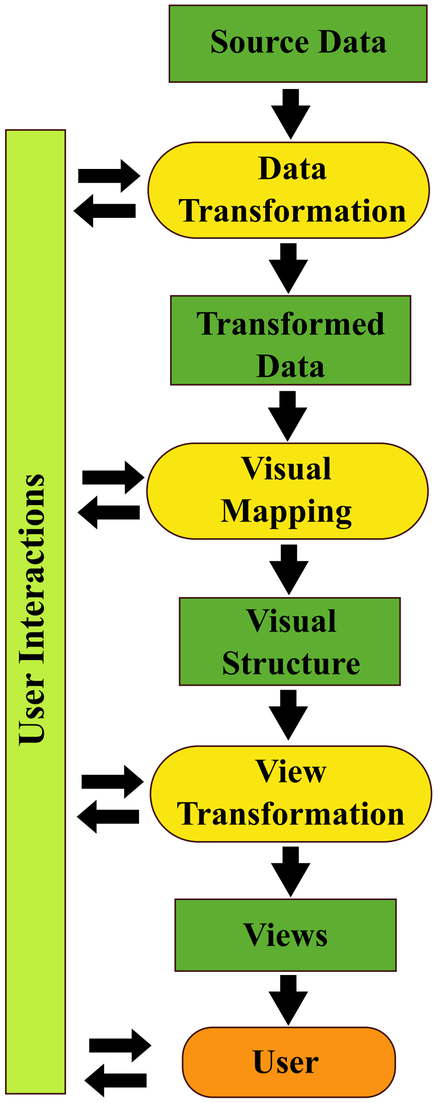
- Subspace clusterings + visualization
- Clustering + regression
- Partition-based regression + visualization

Visual Mapping of high-dim data



Review: Visualization pipeline for high-dim data

[LiuMaljovecWang2017]



Data Transformation

Dimension Reduction
Linear Projection [23], [25],
Nonlinear DR [26], [30],
Control Points Projection [34], [37]
Distance Metric [38, 39],
Precision Measures [42], [44]

Subspace Clustering
Dimension Space Exploration
[47], [48], [49],
Subset of Dimension [51], [53],
Non-Axis-Parallel Subspace
[56], [57], [58]

Regression Analysis
Optimization &
Design Steering
[61], [62], [63],
Structural Summaries
[67], [68]

Topological Data Analysis
Morse-Smale Complex
[166], [168], [169], [170],
Reeb Graph [174], [175], [181]
Contour Tree [179, 180],
Topological Features [191], [192]

Visual Mapping

Axis-Based
Scatterplot Matrix [70],
Parallel Coordinate [77],
Radial Layout [89], [90],
Hybrid Construction
[93], [94], [95], [96]

Glyphs
Per-Element Glyphs
[99], [100],
[101], [102],
Multi-Object Glyphs
[103], [104], [105]

Pixel-Oriented
Jigsaw Map [109],
Pixel Bar Charts [108],
Circle Segment [107]
Value & Relation
Dispaly [110]

Hierarchy-Based
Dimension
Hierarchy [113],
Topology-Based
Hierarchy [197], [198],
Others [115], [117]

Animation
GGob i[119],
TripAdvisorND
[52],
Rolling the
Dice [120]

Evaluation
Scatterplot Guideline
[122], [123]
Parallel Coordinates
Effectiveness [124],
Animation [127]

View Transformation

Illustrative Rendering
Illustrative PCP [128],
Illuminated 3D Scatterplot [129],
PCP Transfer Function [130],
Magic Lens [132, 133]

Continuous Visual Representation
Continuous Scatterplot [134], [135]
Continuous Parallel Coordinates [136],
Splatterplots [138],
Splatting in Parallel Coordinates [136]

Accurate Color Blending
Hue-Preserving
Blending [140],
Weaving vs. Blending
[141]

Image Space Metrics
Clutter Reduction
[142], [143],
Pargnostics [144],
Pixnostic [145]

Visualization pipeline for HD data

[LiuMaljovecWang2017]

Visual Mapper

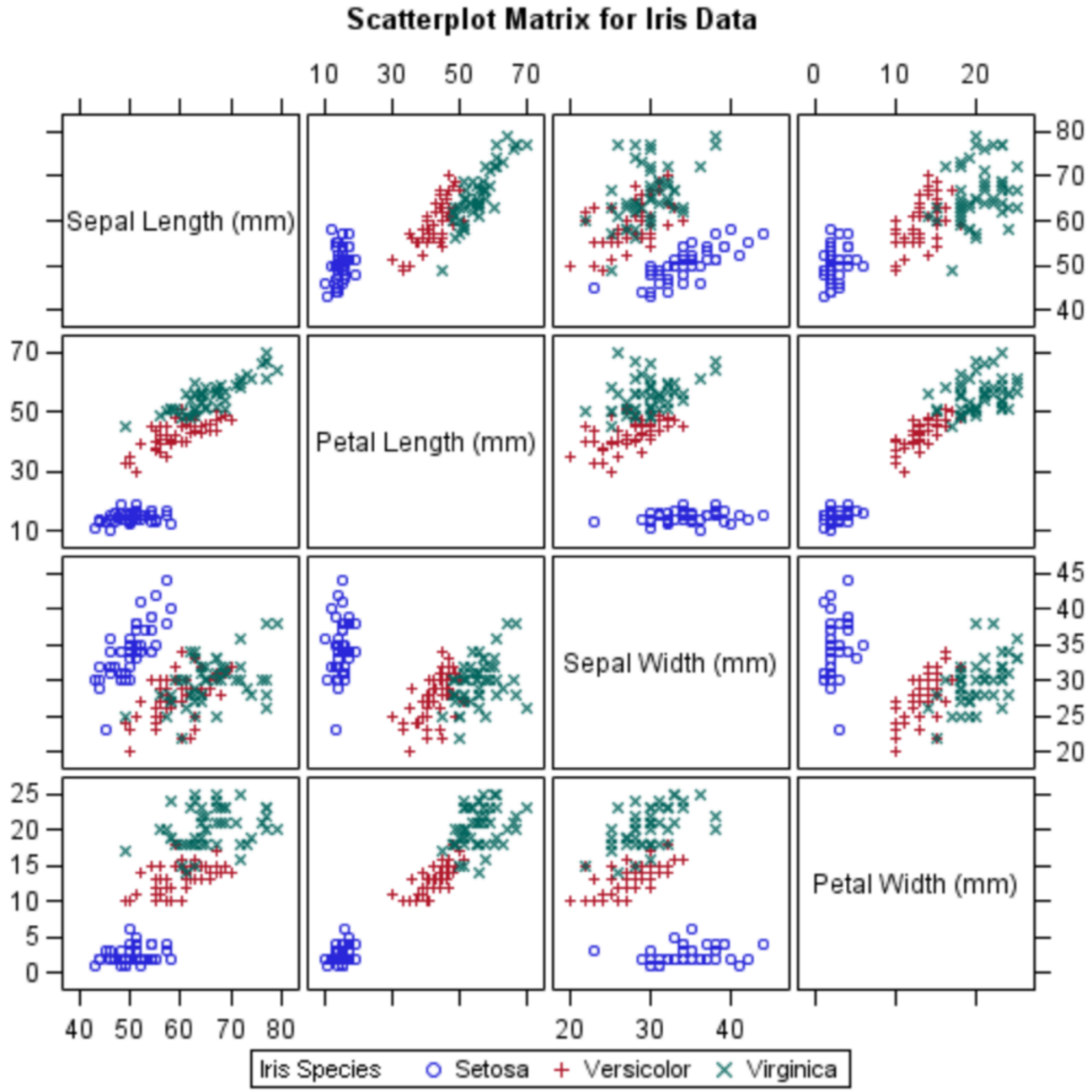
- Plays an essential role in converting analysis results from the data transformation stage into visual structures for rendering in the view transformation stage
- Several approaches based on differences in their structural patterns and visual compositions:
  - Axis-based
  - Glyphs
  - Pixel-oriented
  - Hierarchical-based
  - Animation

Overview

- Axis-based: contain axes corresponding to the original data dimensions, projected dimensions, or combinations thereof.
- Glyphs: encode information into the size, color, shape, and arrangement of small graphical symbols.
- Pixel-oriented: encode individual data values as pixels and focus on arranging the pixels in meaningful ways.
- Hierarchical-based: visualize nesting relationships in multi-resolution and tree-like data.
- Animation: include a temporal element to convey information in the changing of visual elements.

Finally, evaluate the effectiveness of visual encodings.

Axis-Based Methods



Scatterplot Matrix (SPLOM)

http://support.sas.com/documentation/cdl/en/grstatproc/62603/HTML/default/viewer.htm#a003155769.htm

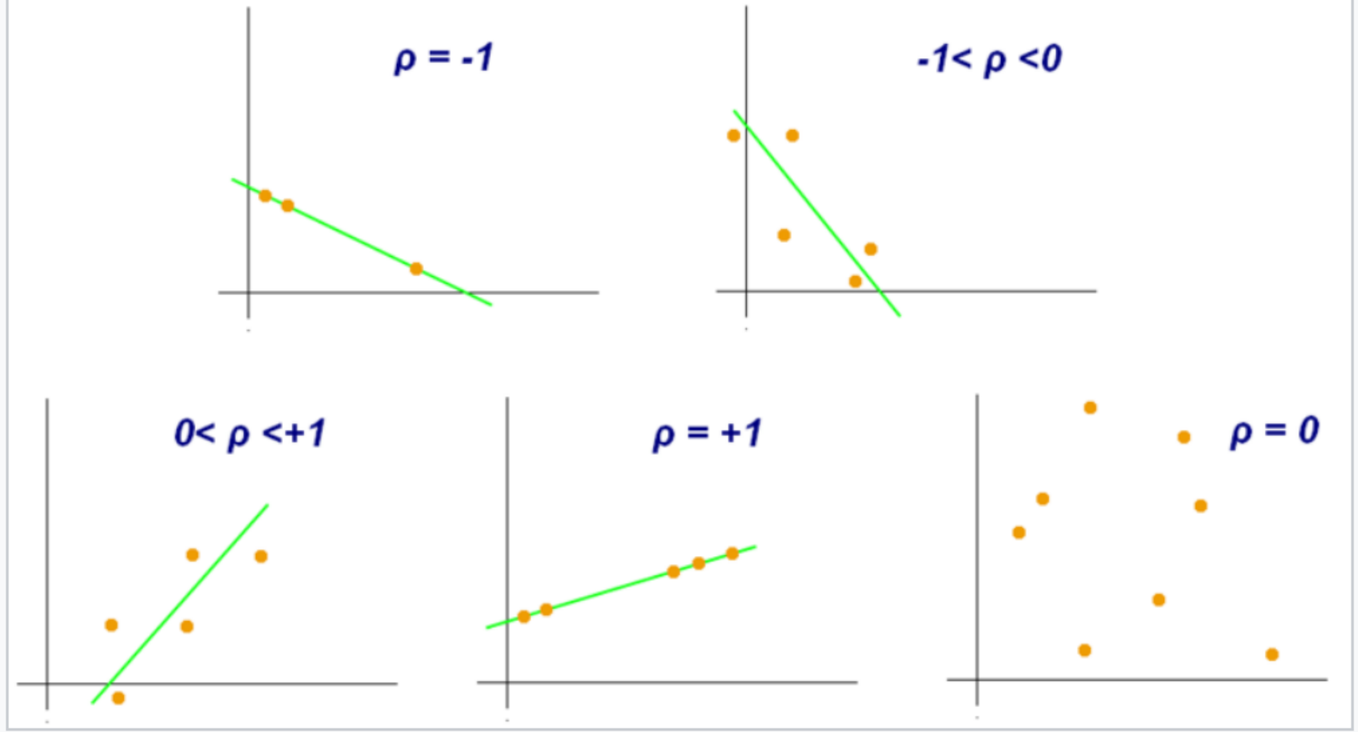
A collection of bivariate scatterplots: view multiple bivariate relationships simultaneously

Review: correlation

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

where:

- cov is the covariance
- σ_X is the standard deviation of X
- σ_Y is the standard deviation of Y

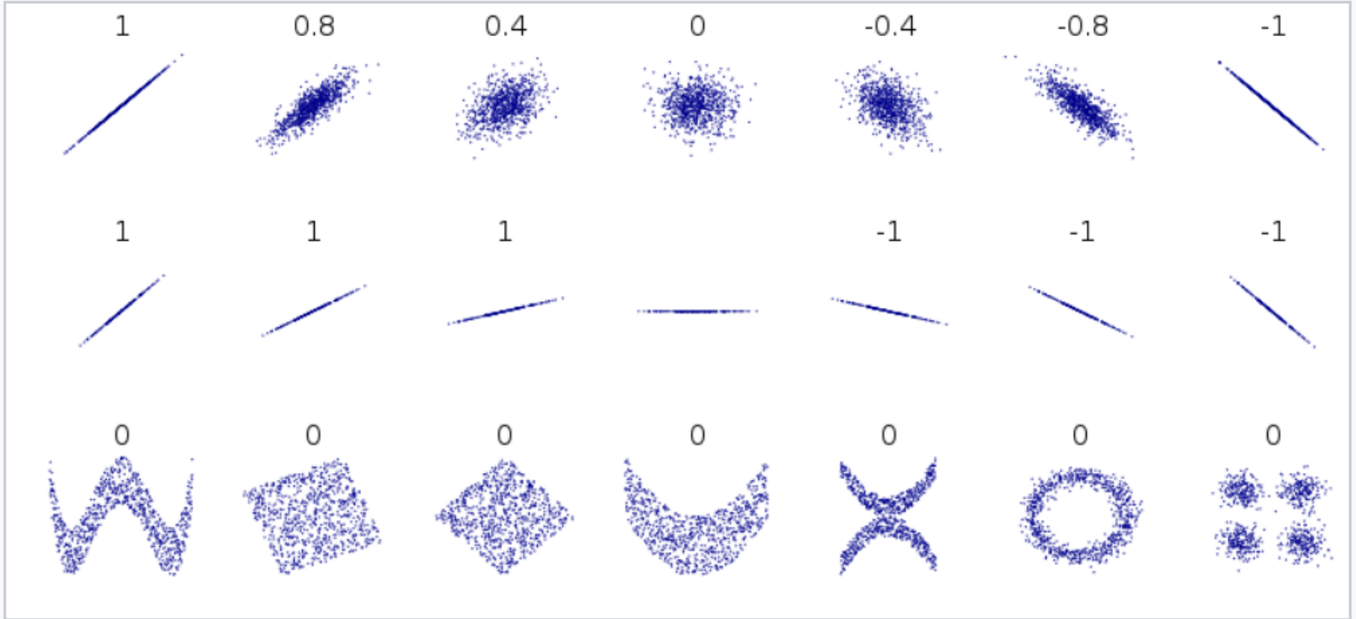


Examples of scatter diagrams with different values of correlation coefficient (ρ)

Pearson correlation coefficient

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Review: correlation



Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the non-linearity and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

Pearson correlation coefficient

SPLOM

- Major drawback: scalability
- How do we improve the scalability by automatically or semi-automatically identifying interesting plots?

Graph-Theoretic Scagnostics

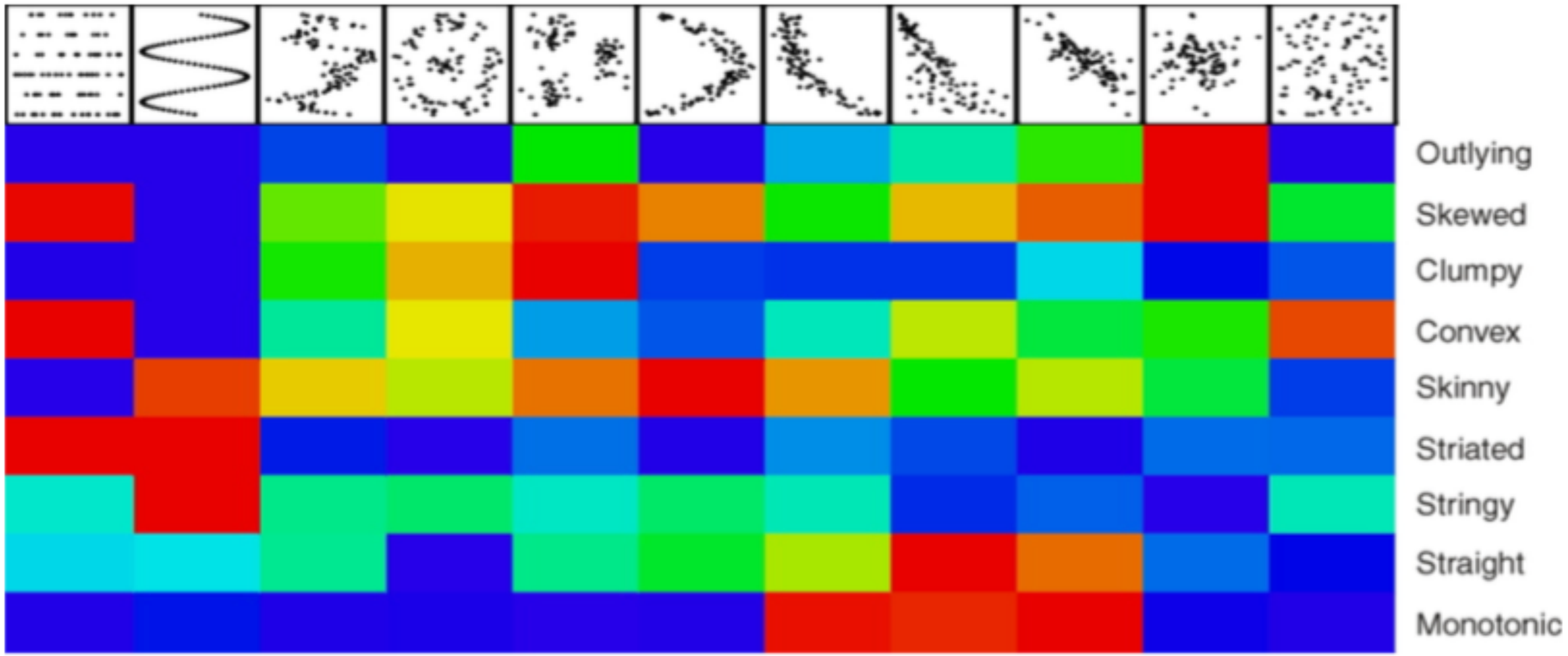
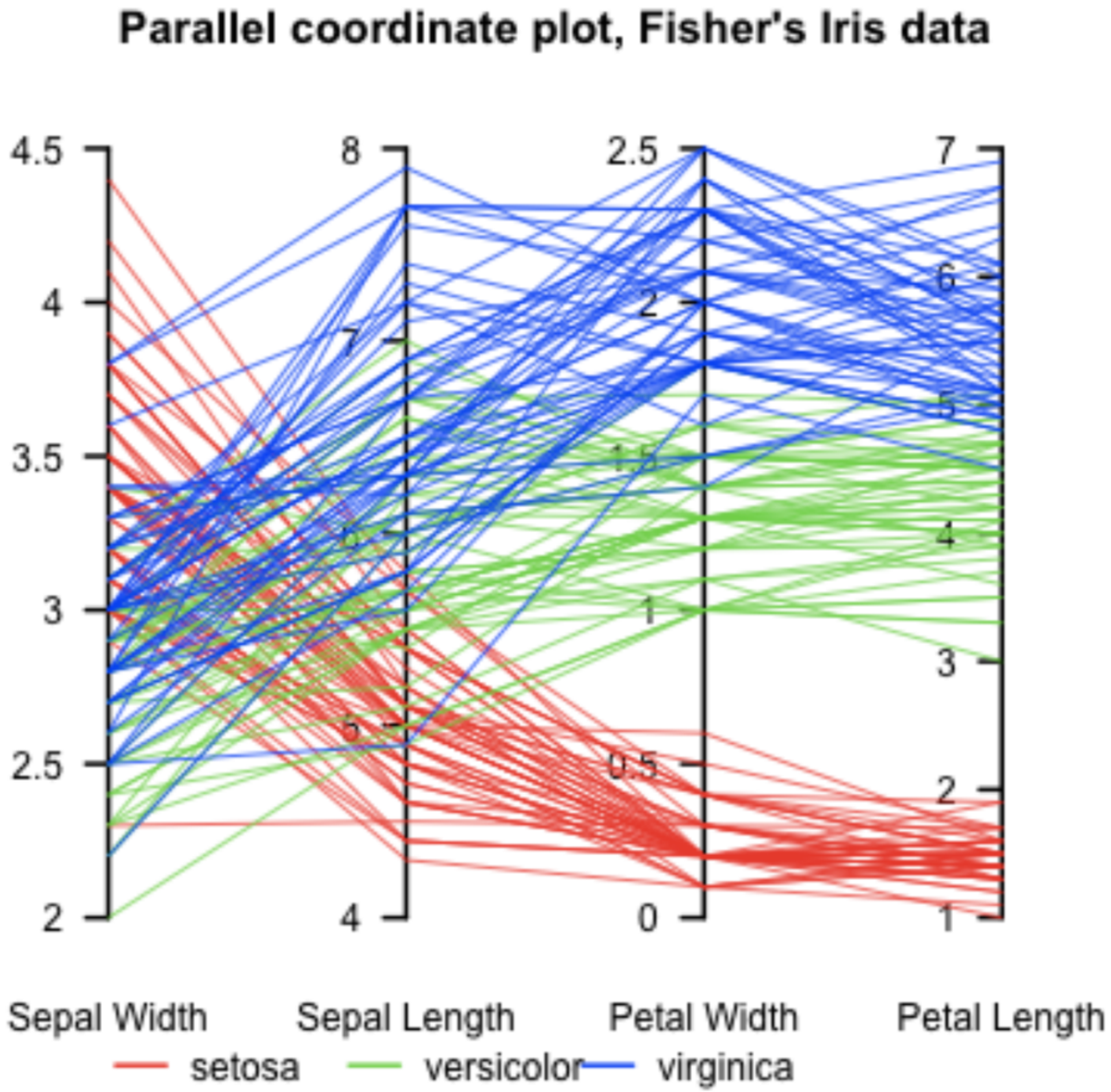


Figure 3: Scaled graph-theoretic measures (blue=low, red=high) for eleven scatter patterns

[WilkinsonAnandGrossman2005]

SPLOM: other considerations

- Rank-by-feature: histogram distribution properties; or correlation coefficients between axes [SeoShneiderman2004]
- Class labels play an important role in identifying interesting plots and ranking order
- Class consistency: distance to the center of the class or entropies of the spatial distributions of classes
- Class density measure or histogram density measure to rank scatterplots



https://en.wikipedia.org/wiki/Parallel_coordinates

Parallel Coordinates (PCP)

PCP

- Instead of directly express bivariate relationships (as in SPLOM), PCP allow patterns that highlight multivariate relations to be revealed by showing all the axes at once
- Key question: determine the appropriate orders of the axes
- Users can only focus on visual patterns of nearby axes
- Reduce search space by focusing on localized axes orders: consecutive dimension triples or pairwise dimensions

PCP: Combing quality metrics

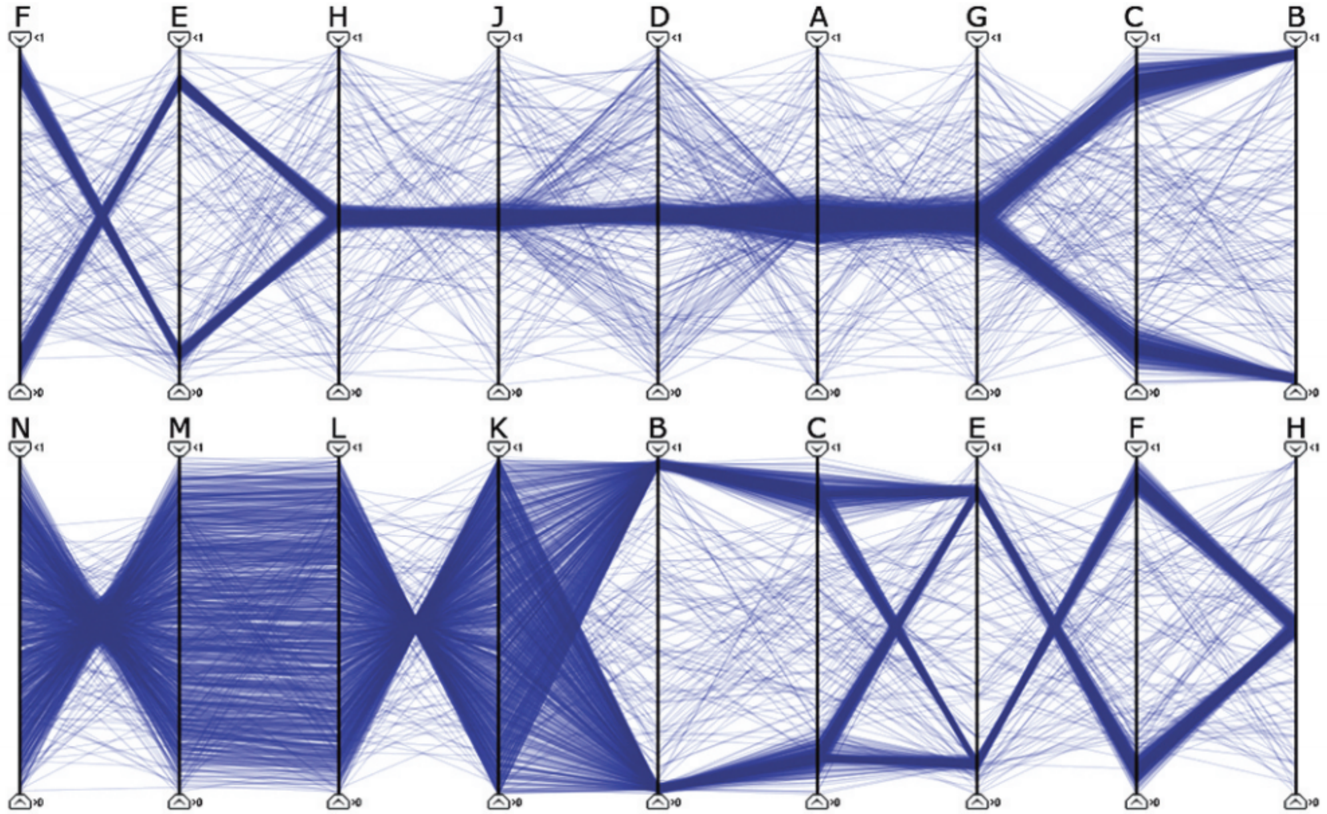


Fig. 2. The synthetic data set reduced to 9 variables using different quality metric weights and variable orders. In the top view clustering is assigned a large weight and the variables are ordered to enhance the cluster structures. In the bottom view a corresponding weighting and ordering is made for correlation structures.

Use a weighted combination of quality metrics for dimension selection and automatic ordering of the axes to enhance visual patterns such as clustering correlation

[JohanssonJohansson2009]

PCP: Combing quality metrics

1. A data set is loaded into the system and the user selects quality metrics to use and sets the parameters for the quality metric analysis.

2. The system performs quality analysis for the selected metrics individually and determines a quality value for each variable and metric.

3. The relationship between the number of variables to keep and loss of information is presented to the user in an interactive display. At this point the user can also modify the importance of individual quality metrics, updating the display accordingly.

4. The user decides on the number of variables to keep in the reduced data set and the system selects the most important variables from the original data set based on quality values and metric importance.

5. In the final step before the reduced data set is displayed, the user selects which visual representations to use and which quality metric the variable ordering should enhance.

6. The reduced data set is displayed using the selected representations and orderings. From here any of the previous steps can be repeated to modify the reduced data set.

- Quality metrics (for variables/ dimensions)
  - Correlation analysis
  - Outliers
  - Cluster detection: uses a clustering algorithm to identify low- dim sub-clusters, which are then the base of computing a cluster quality value for every variable.

[JohanssonJohansson2009]

PCP: other considerations

- Visual clutter due to the number of dimensions and line density
- Clutter reduction via: filtering, aggregation, visual encoding, and dimension reordering
- Example: line bundling

PCP: line bundling

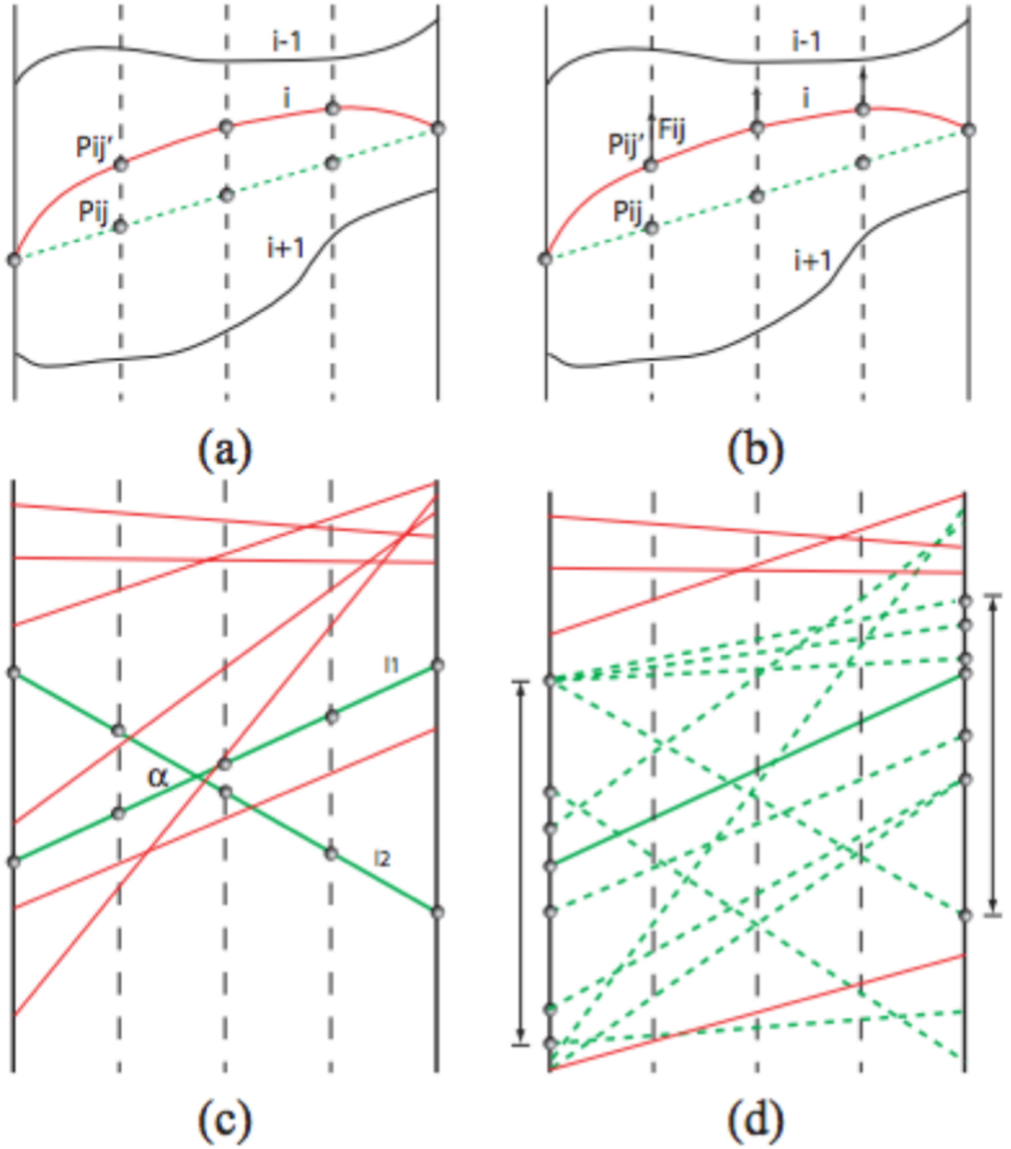


Figure 1: *Energy Terms: (a) Curvature energy term; (b) Gravitation energy term; (c) Computing the force for the gravitation energy term; (d) The range of neighboring lines for gravitation interaction. m is set to be 3 for Eq. 2 and 3.*

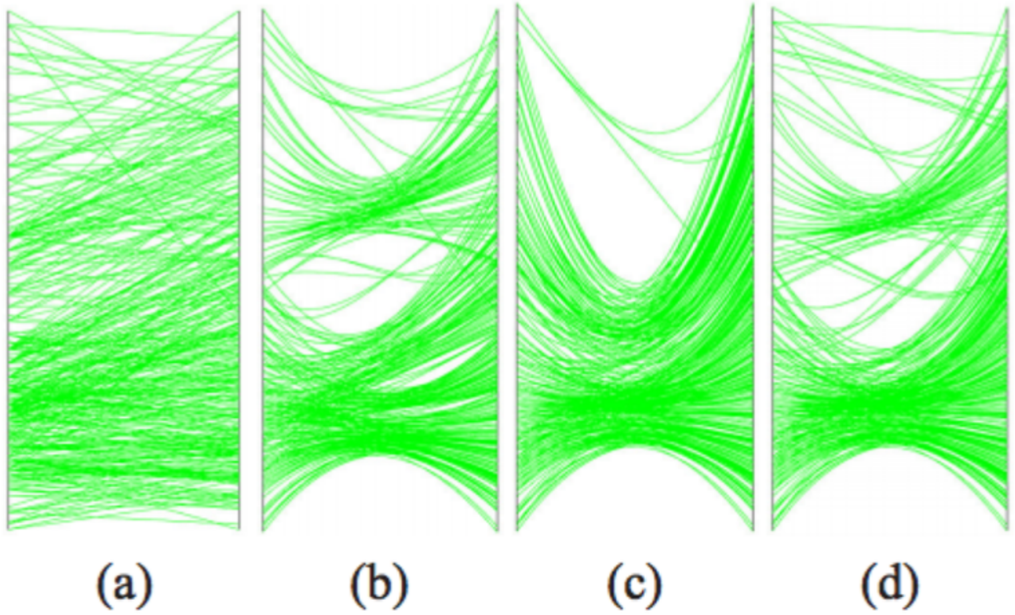
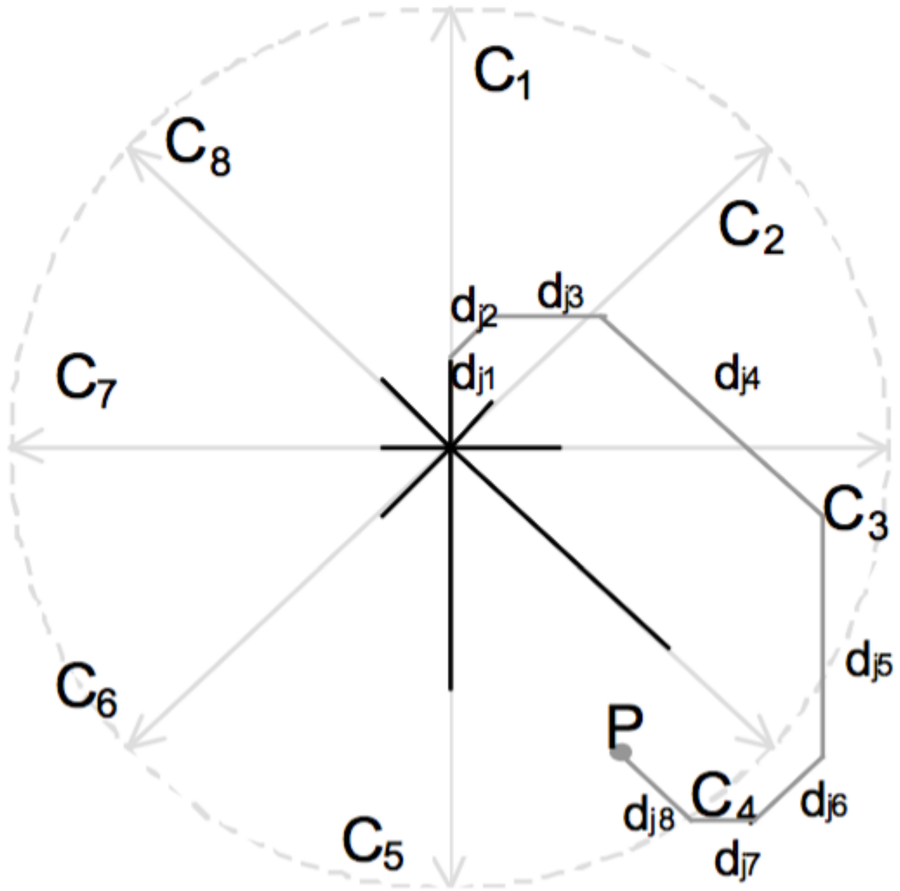


Figure 2: *The effect of energy term weighting on visual clustering: (a) No visual clustering; (b)* $\alpha_c = 0$, $q_\alpha = q_d = 15$; *(c)* $\alpha_c = 0$, $q_\alpha = q_d = 30$; *(d)* $\alpha_c = 0.15$, $q_\alpha = q_d = 30$.

[ZhouYuanQu2008]

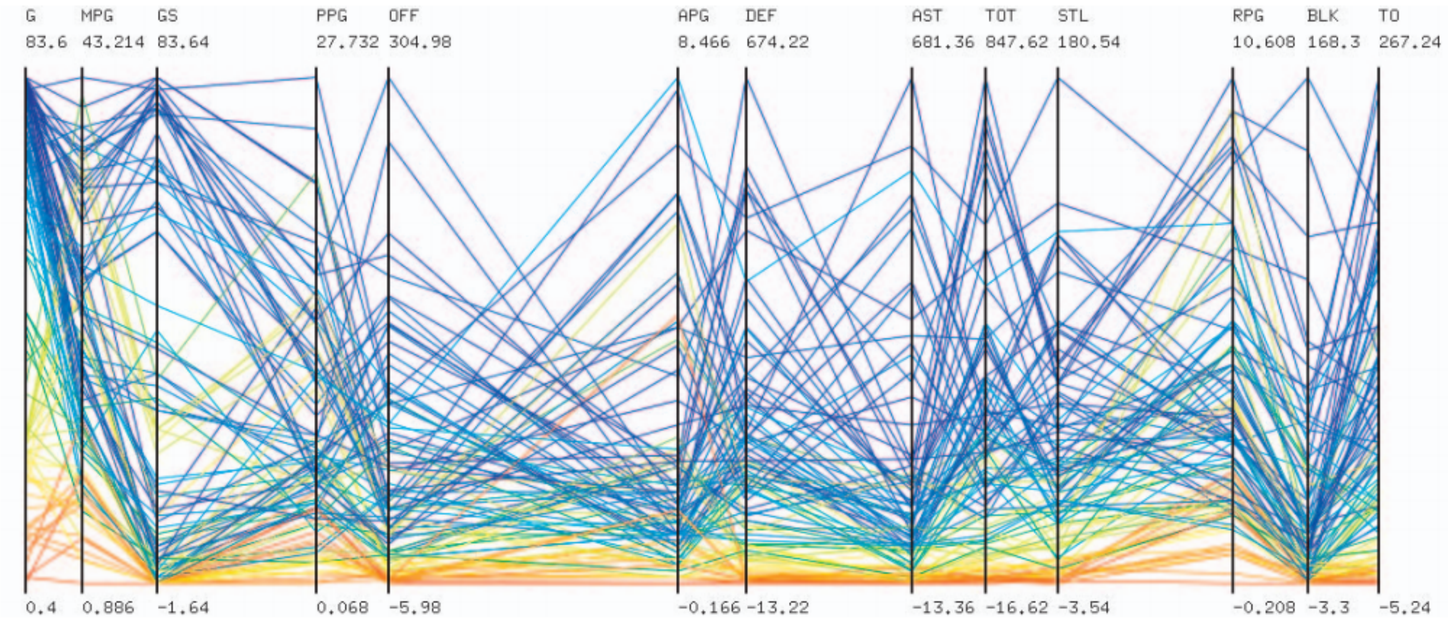


[Kandogan2000]

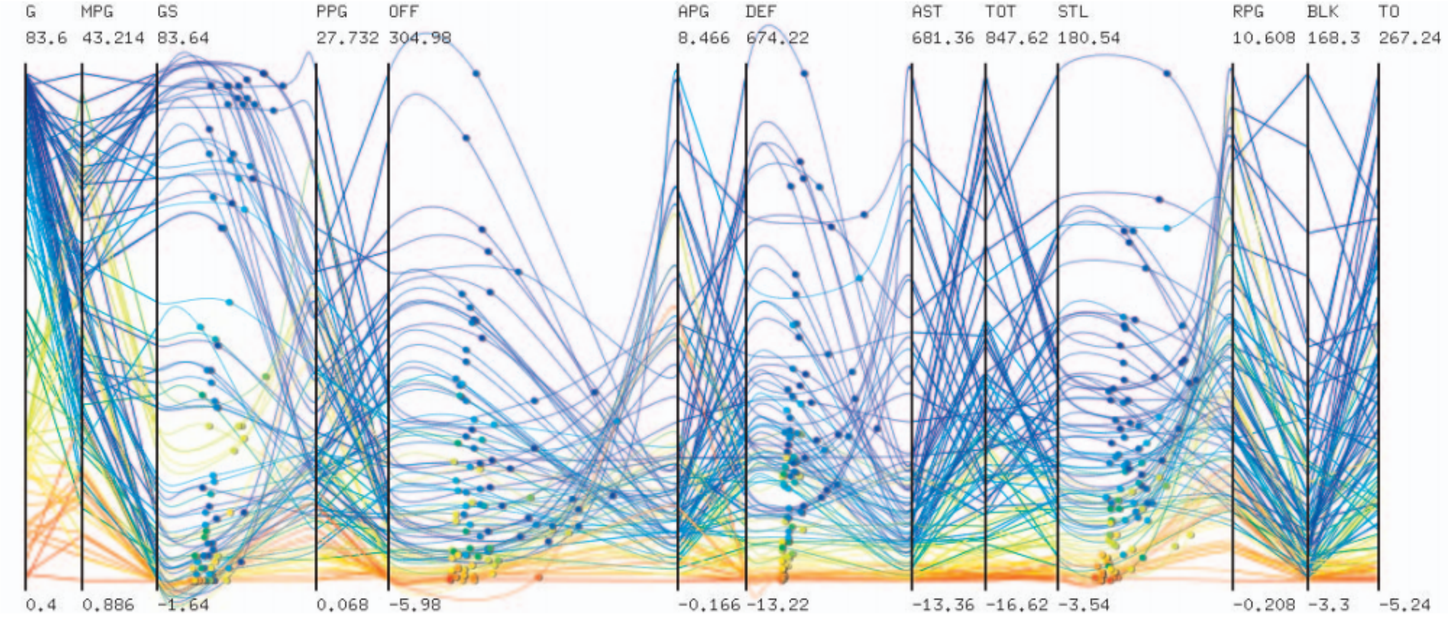
Radial Layout (star coordinate plot, bi-plot)

Radial layout

- An extension of typical 2d and 3d scatter-plots to higher dimensions with normalization.



(a)



(b)

Hybrid Constructions

[YuanGuoXiao2007]

Hybrid Constructions

- Combine axis-based methods to create new visualizations
- In the previous example: embeds an MDS plot between a pair of PCP axes
- Other examples:
  - Generalization of PCP and SPLOM
  - Integrate PCP with glyphs
  - Angular histograms

Thanks!

Any questions?

You can find me at: **beiwang@sci.utah.edu**

CREDITS

Special thanks to all people who made and share these awesome resources for free:

- Presentation template designed by Slidesmash
- Photographs by unsplash.com and pexels.com
- Vector Icons by Matthew Skiles

Presentation Design

This presentation uses the following typographies and colors:

Free Fonts used:

http://www.1001fonts.com/oswald-font.html

https://www.fontsquirrel.com/fonts/open-sans

Colors used