Advanced Data Visualization

CS 6965

Spring 2018

Prof. Bei Wang Phillips

University of Utah

Lecture 06

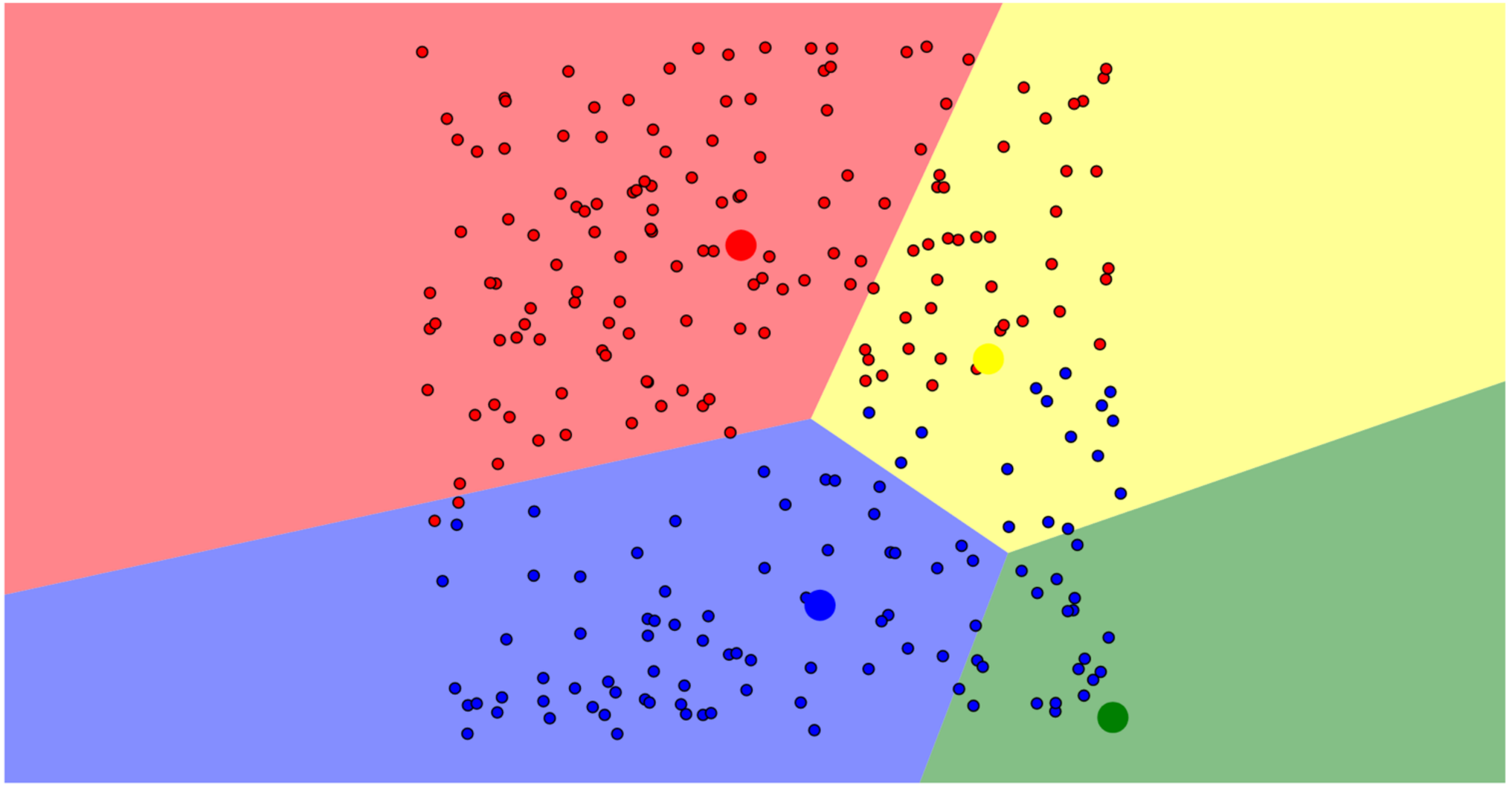
Clustering, Regression and Vis

HD

Clustering & Vis

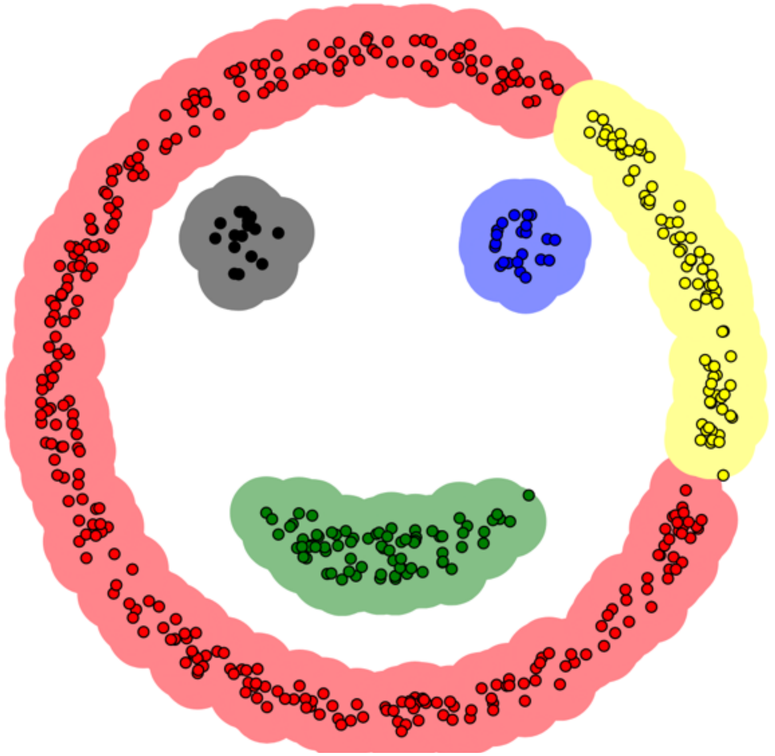
Visualizing Clustering Process

- Visualizing the algorithmic process for clustering (especially iterative ones)



https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Visualizing DBSCAN



The DBSCAN algorithm can be abstracted into the following steps:[4]

1. Find the ε (eps) neighbors of every point, and identify the core points with more than minPts neighbors.
2. Find the connected components of *core* points on the neighbor graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbor, otherwise assign it to noise.

https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/
https://en.wikipedia.org/wiki/DBSCAN

Subspace clustering (SC) & Vis

Subspace clustering vs DR

- Clustering: widely used data-driven analysis methods
- DR: compute one single embedding that best describes the structure of data
- Subspace clustering
  - Identify multiple embeddings, each capturing a different aspect of the data
  - Clustering either the dimensions or the data points

Subspace clustering & Vis

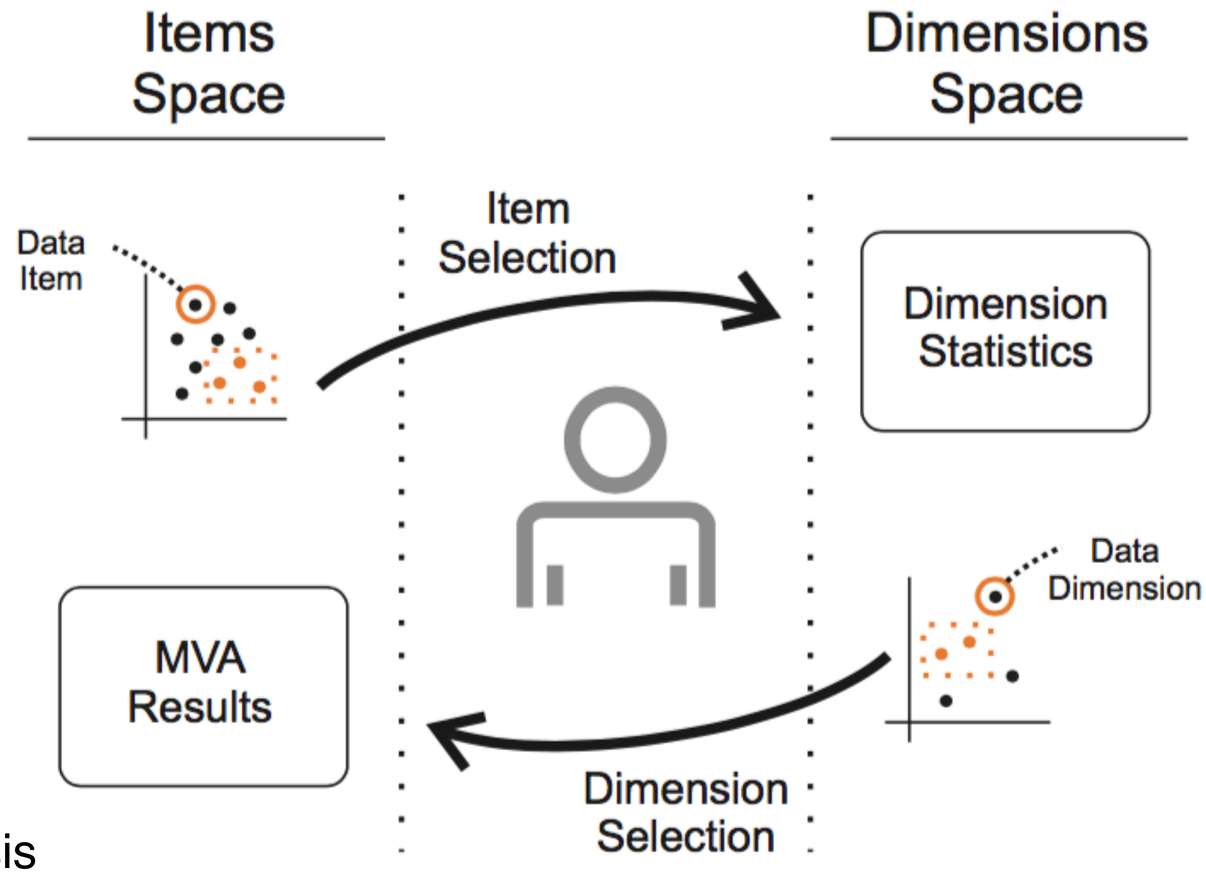
- Explore dimension space
- Explore subsets of dimensions
- Non-Axis-Aligned subspaces

SC: Dimension Space Exploration

Dimension space exploration

- Guided by the user
- Interactively group relevant dimensions into subsets

Dual Analysis Model



MVA: Multivariate analysis

https://fmfatore.files.wordpress.com/2012/09/pres.pdf

[TurkayFilzmoserHauser2011]

Representative factor generation

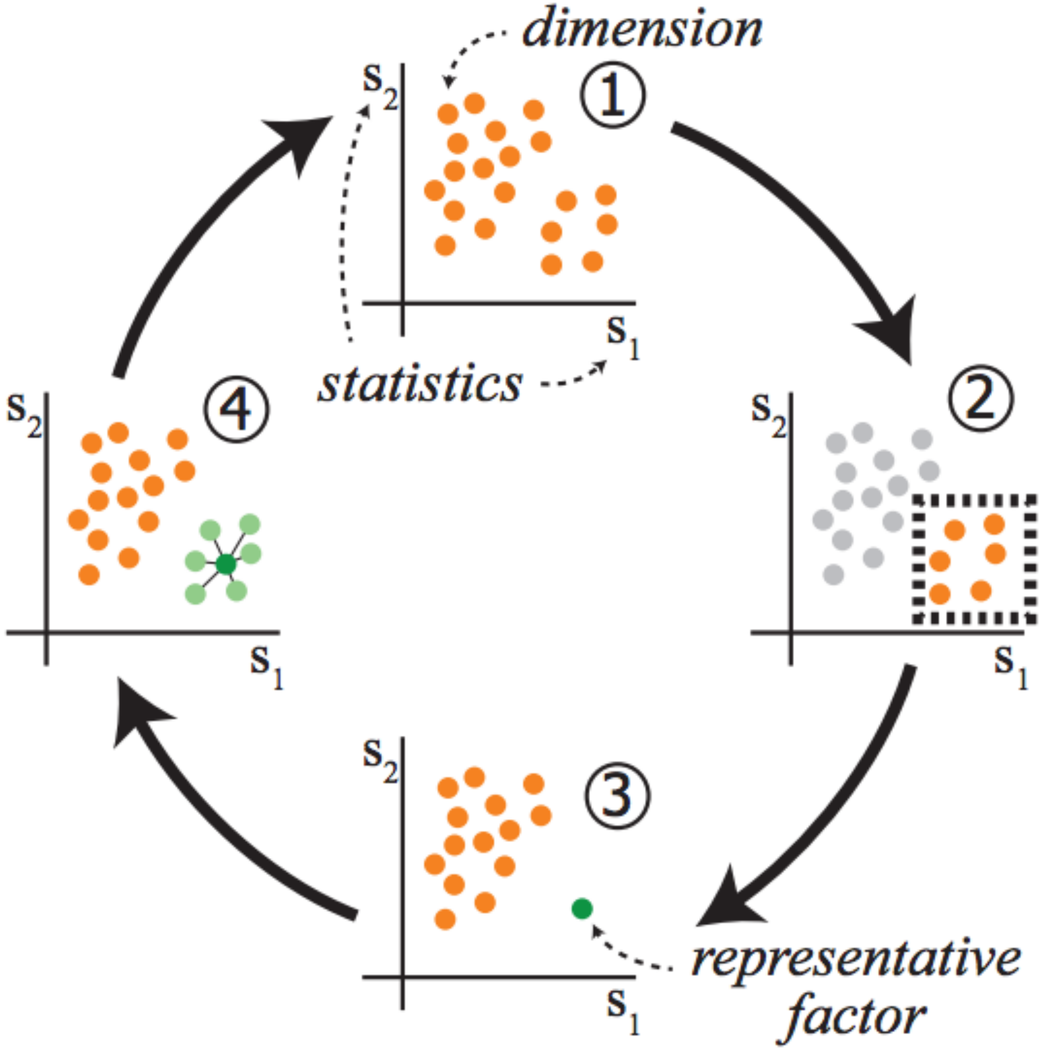


Fig. 1. An illustration of our representative factor generation method. Two statistics s_1 and s_2 are computed for all the dimensions and dimensions are plotted against these two values (1). This view reveals a group that shares similar values of s_1 and s_2 (2) and this group is selected to be represented by a factor. We generate a representative factor for this group and compute the s_1 and s_2 values for the factor (3). We observe the relation of the factor to the represented dimensions and the other dimensions (4). The analysis continues iteratively to refine and compare other structures in the data.

Grouping a collection of dimensions as a factor

[TurkayLundervoldLundervold2012]

Dimension projection matrix/tree

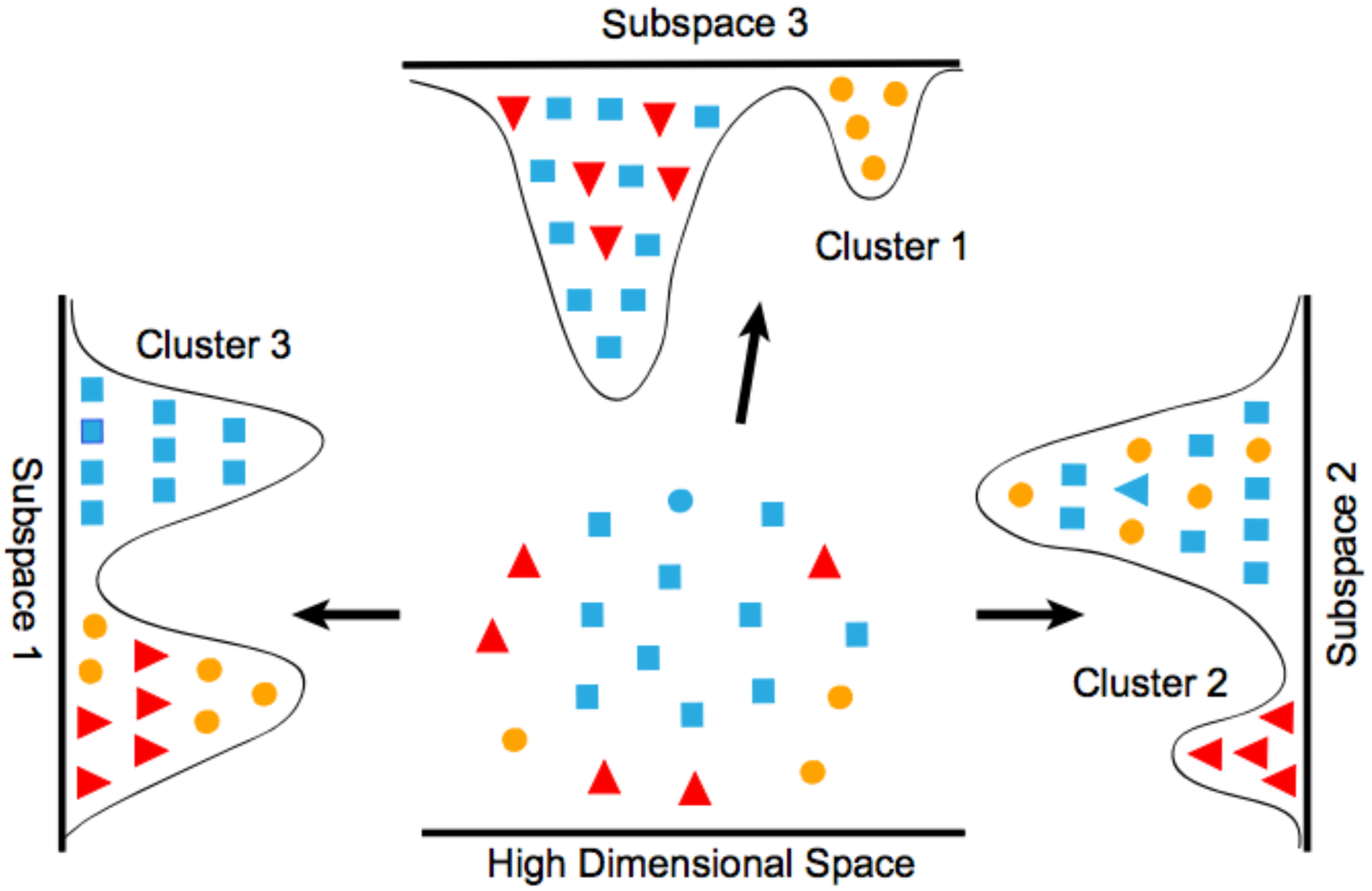
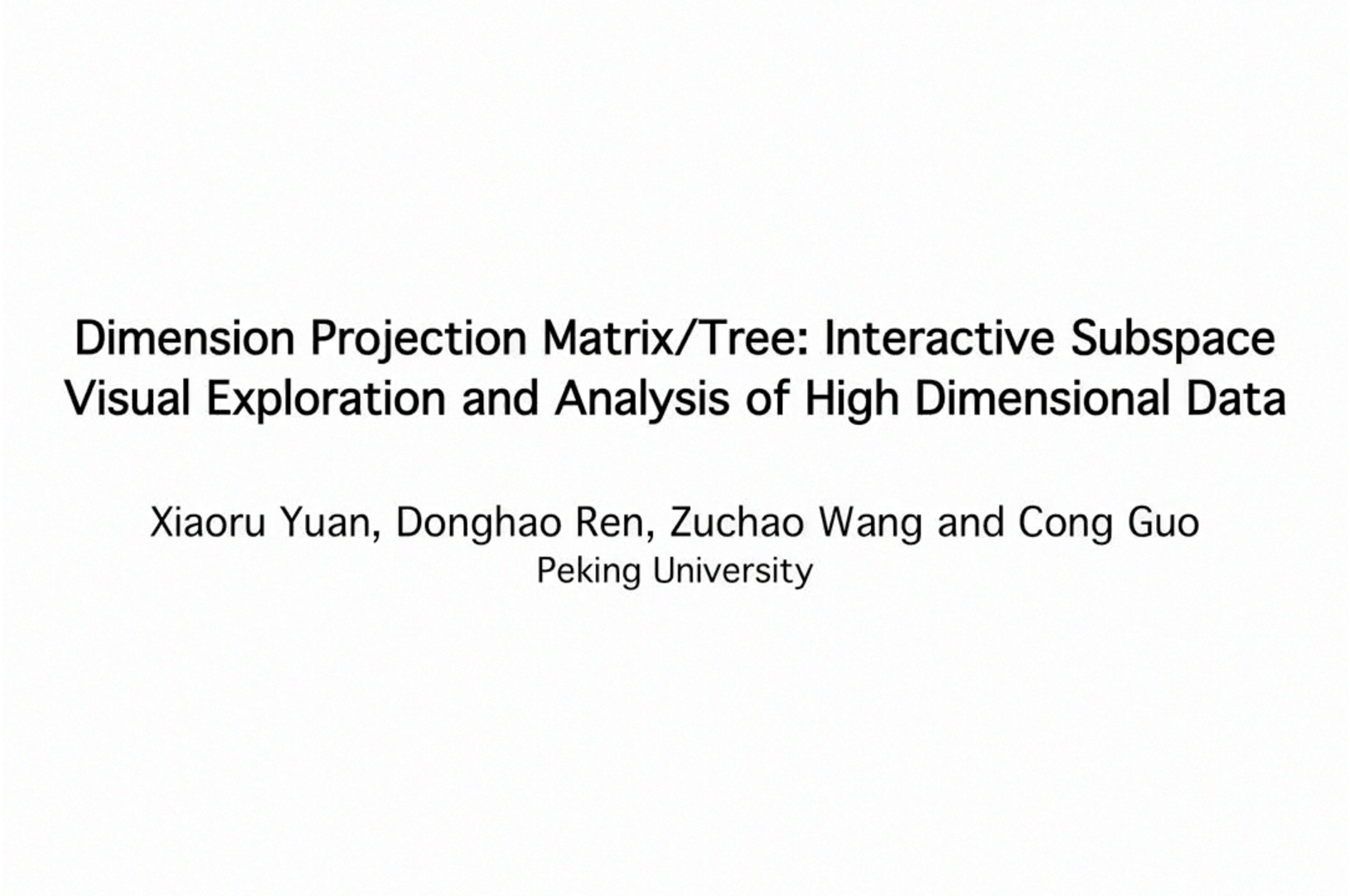


Fig. 2. Illustration of clustering in subspaces. Separation of clusters in appropriate selection of dimension subspaces can be much easier than that in the original high dimensional space.

[YuanRenWang2013]

Dimension projection matrix/tree

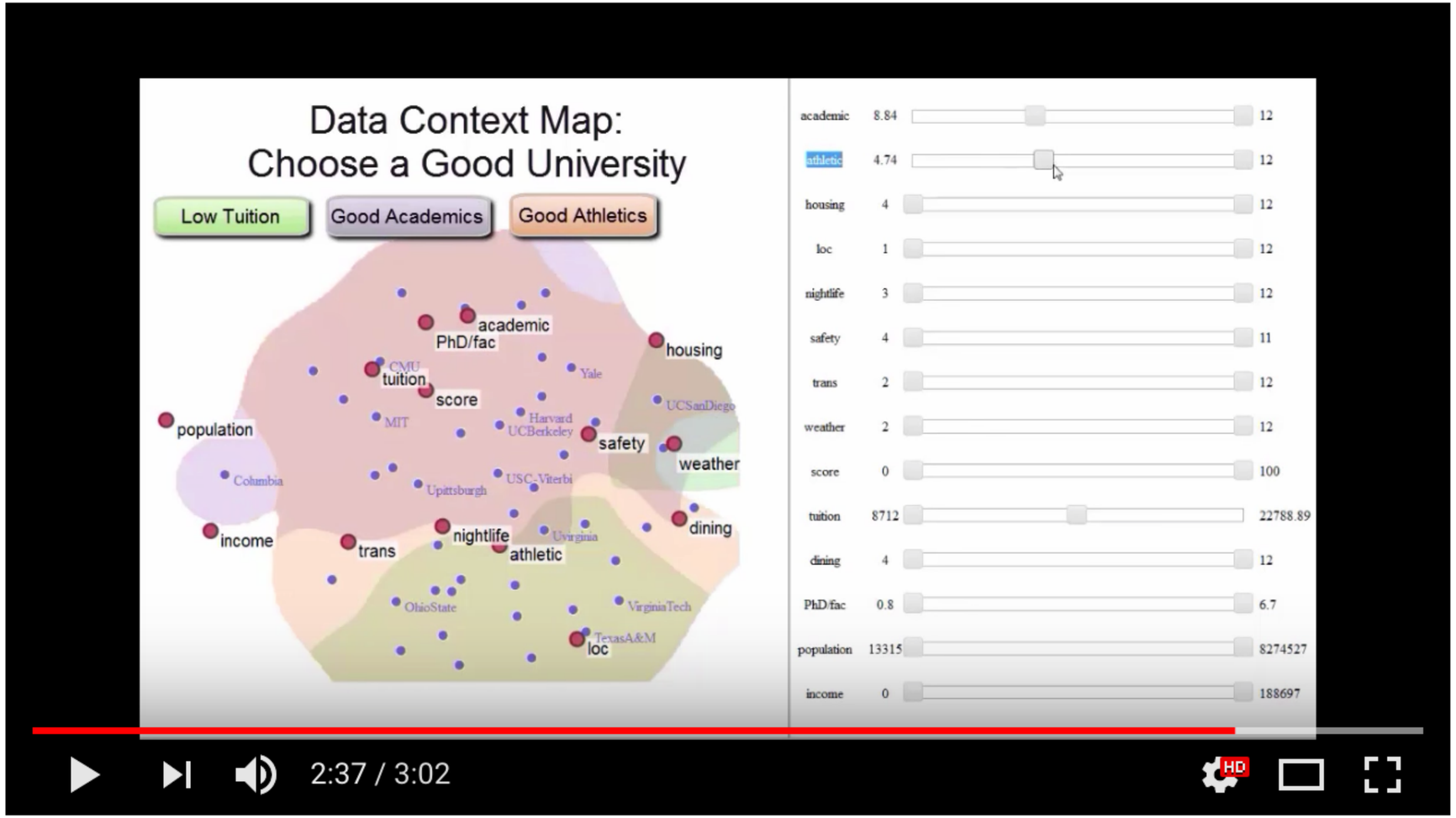


Video: http://vis.pku.edu.cn/wiki/doku.php?id=publication:start

[YuanRenWang2013]

Data Context Map

Observing the data points in the context of the attributes



Combining two similarity matrices typically used in isolation – the matrix encoding the similarity of the attributes and the matrix encoding the similarity of the data points.

https://www.youtube.com/watch?v=nnjkHA8xvbI&feature=youtu.be

http://www3.cs.stonybrook.edu/~mueller/research/pages/dataContextMap/ [ChengMuller2016]

SC:Subsets of Dimensions

Subspace clustering & finding

- Different from dimension space exploration, which relies on users to identify patterns
- Automatically group related dimensions into clusters
- Filter out interferences from irrelevant dimensions
- EUCLUS, TripAdvisor, etc.

CLIQUE

- Discretize the data space into non-overlapping rectangular units (reminder: mapper?)
  - Partitioning every dimension into intervals of equal length.
  - A unit is dense if the fraction of total data points contained in the unit is greater than a threshold.
- Clusters are unions of connected dense units within a subspace.

[AgrawalGehrkeGunopulos1998]

ENCLUS

- Entropy based subspace clustering
- Subspaces are formed by subsets of dimensions (attributes)
- Identify meaningful criteria of high density and correlation of dimensions for goodness of clustering in subspaces
- Criteria of subspace clustering
  - High coverage (same as CLIQUE)
  - High density (cluster can have the same coverage but different density)
- Correlation of dimensions (want the dimensions of the subspace to be correlated)

[ChengFuZhang1999]

ENCLUS: Entropy based metric

- Given a set of criteria for clustering, finding a metric that measures all criteria simultaneously.
- A subspace which has good clustering by the criteria will have high score in this metric.
- Entropy is a measure of uncertainty of a random variable.

$$H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x)$$

X: a variable representing a cell.
$\mathcal{X}$: the set of possible outcomes of X.
p(x): the probability mass function of the random variable X.

H(X): the entropy.

[ChengFuZhang1999]

ENCLUS: Entropy vs Clustering

- The entropy decreases as the coverage increases.
- As the density of the dense units increases, the entropy decreases.
- The problem of correlated variables can be handled by entropy because the independence and dependence of the variables can be detected using the following relationships in entropy:

$$H(X_1, \ldots, X_n) = H(X_1) + \ldots + H(X_n)$$
$$\text{iff } X_1, \ldots, X_n \text{ are independent} \qquad (1)$$

$$H(X_1, \ldots, X_n, Y) = H(X_1, \ldots, X_n)$$
$$\text{iff } Y \text{ is a function of } X_1, \ldots, X_n \qquad (2)$$

[ChengFuZhang1999]

ENCLUS: Computing entropy

- Divide each dimension into intervals of equal length, so the high-dimensional space is partitioned to form a grid.
- Suppose the data set is scanned once to count the number of points contained in each cell of the grid.
- The density of each cell can thus be found.

$$H(X) = -\sum_{x \in \mathcal{X}} d(x) \log d(x)$$

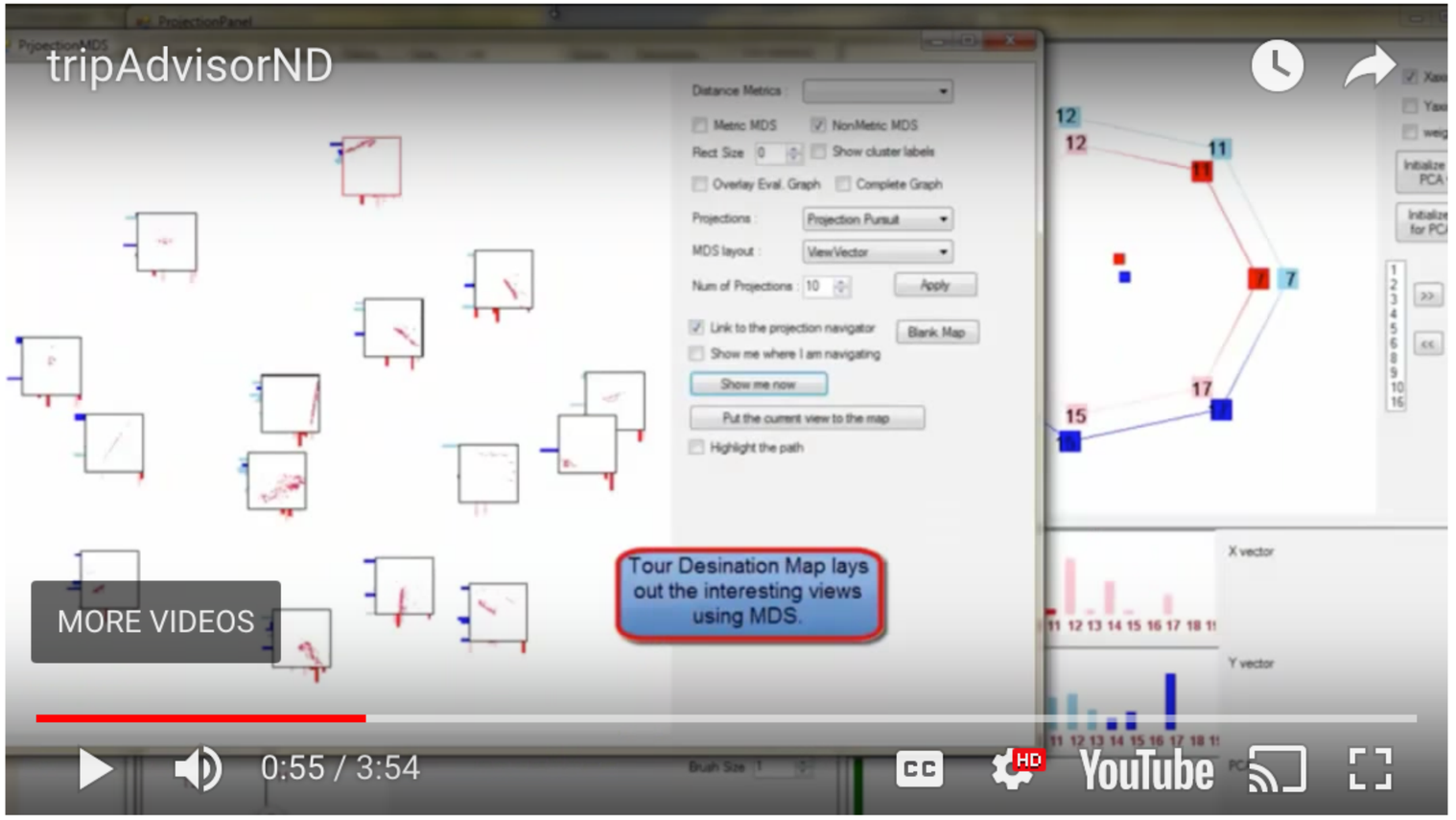
X: a set of all cells.
$\mathcal{X}$: a set of all cells.
d(x): the density of a cell x (in terms of the percentage of data contained in x).

H(X): the entropy.

[ChengFuZhang1999]

TripAdvisor-ND

- Employs a sightseeing metaphor for high-dimensional space navigation and exploration.
- Utilizes subspace clustering to identify the sights for the exploration.



[NamMueller2013]

http://www3.cs.stonybrook.edu/~mueller/research/pages/TripAdvisorND/ https://youtu.be/3_qRwOyTD_w

SC:Non-Axis-Aligned Subspace

Projection pursuit

- Early work: automatically identifying the interesting non-axis-aligned subspaces.
- The projections are considered to be more interesting when they deviate more from a normal distribution.
- Projection pursuit index

[FriedmanTukey1974]

Projection pursuit indices

- Measures how interesting a projection is:
  - PDF-based: require an estimation of the probability density function (pdf) of the projected samples. Characterize what could be considered as an uninteresting projection by means of the pdf shape. Most of the indices try to diverge from the normal distribution (considered uninteresting).
  - Moment-based: make use of the sample central moments.
  - Classic-information-based: make use of labeled data to measure the distance among different classes.

[Rodriguez-MartinezGoulermasMu2010]

Subspace analysis & dynamic proj.

- GGobi: randomly selected subspaces, exploratory (implement projection pursuit)
- Pre-determined subspaces

Visual Exploration of High-Dimensional Data through Subspace Analysis and Dynamic Projection

Shusen Liu[1], Bei Wang[1], Jayaraman J. Thiagarajan[2], Peer-Timo Bremer[2,] Valerio Pascucci[1]
[1]SCI Institute, University of Utah
[2]Lawrence Livermore National Laboratory

(Narrated by John Edwards)

http://www.sci.utah.edu/~beiwang/publications/Subspace_Dynamic_Projection_BeiWang_2015_Supp_video.mp4

Random projections

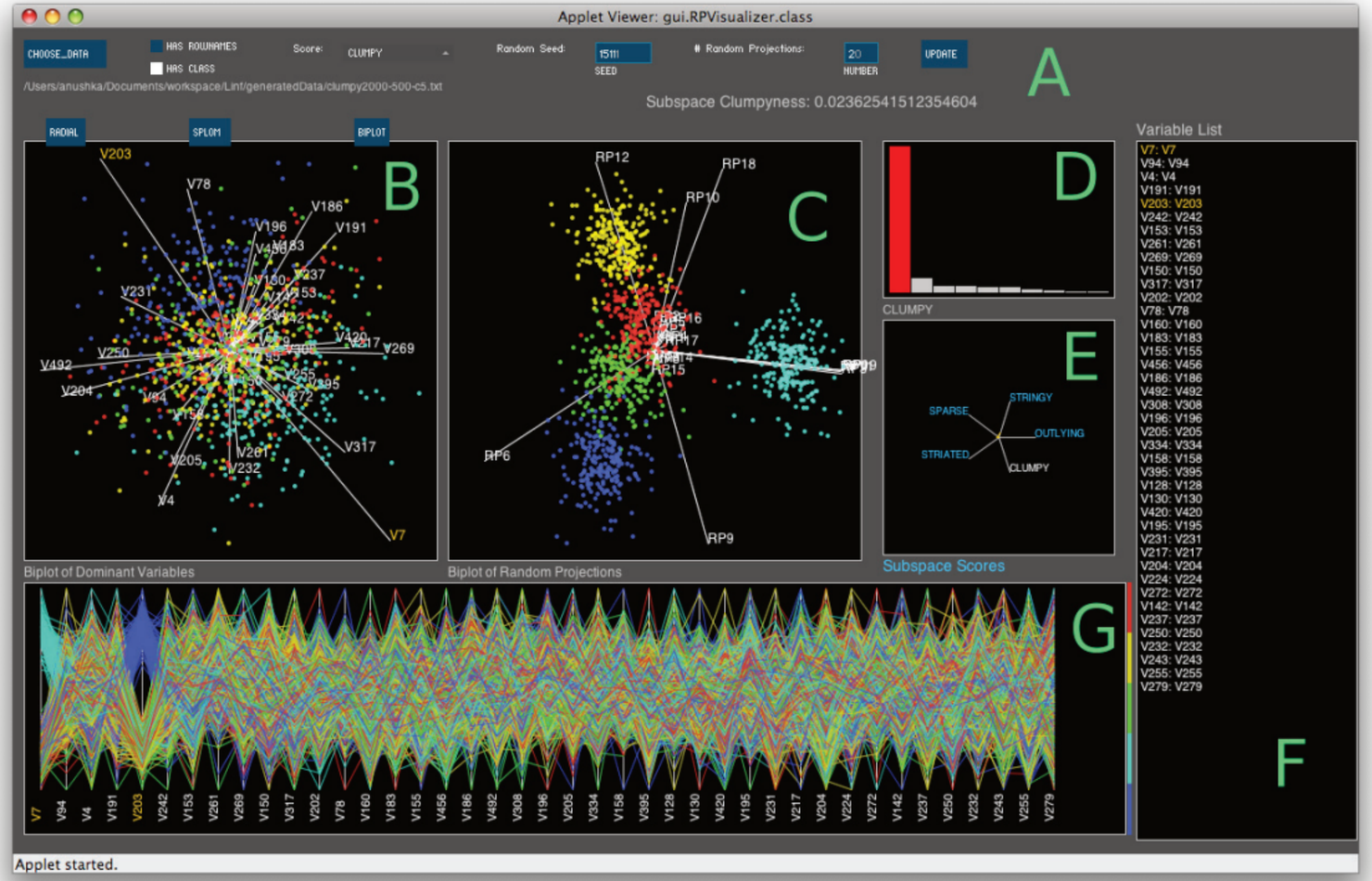


Figure 2: The Subspace Explorer showing a highly Clumpy 20-D random projection for a dataset with 2000 rows, 500 dimensions and clusters embedded in 5 dimensions. The detected known embedding dimensions are shown as orange text. (A) Control options. (B) Biplot View of the data subspace. (C) Random Projection View of the biplot of the projected data space. (D) ScoreView of the top 10 scoring random projections. The selected (red) bar represents one 20-D random projection which is shown in different perspectives in the other plots. (E) Radar plot Icon summary of all scores for the selected random projection. (F) Variable List (same set as B). (G) Parallel Coordinate View of the top used dimensions (same set as B)

[AnandWilkinsonDang2012]

Random projections

- Visual pattern discovery using random projections.
- Define score functions, akin to projection pursuit indices, that characterize visual patterns of the low-dimensional projections that constitute feature subspaces.
- Scoring based on visual pattern features:
  - Outlying: proportion of the total edge length due to edges connected to detected outliers.
  - Clumpy: emphasizes clusters with small intra-cluster distances relative to the length of their connecting edge and ignores clusters with relatively small size.
  - Sparse: measures whether points are confined to a lattice or a small number of locations in the space.
  - etc.

[AnandWilkinsonDang2012]

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]

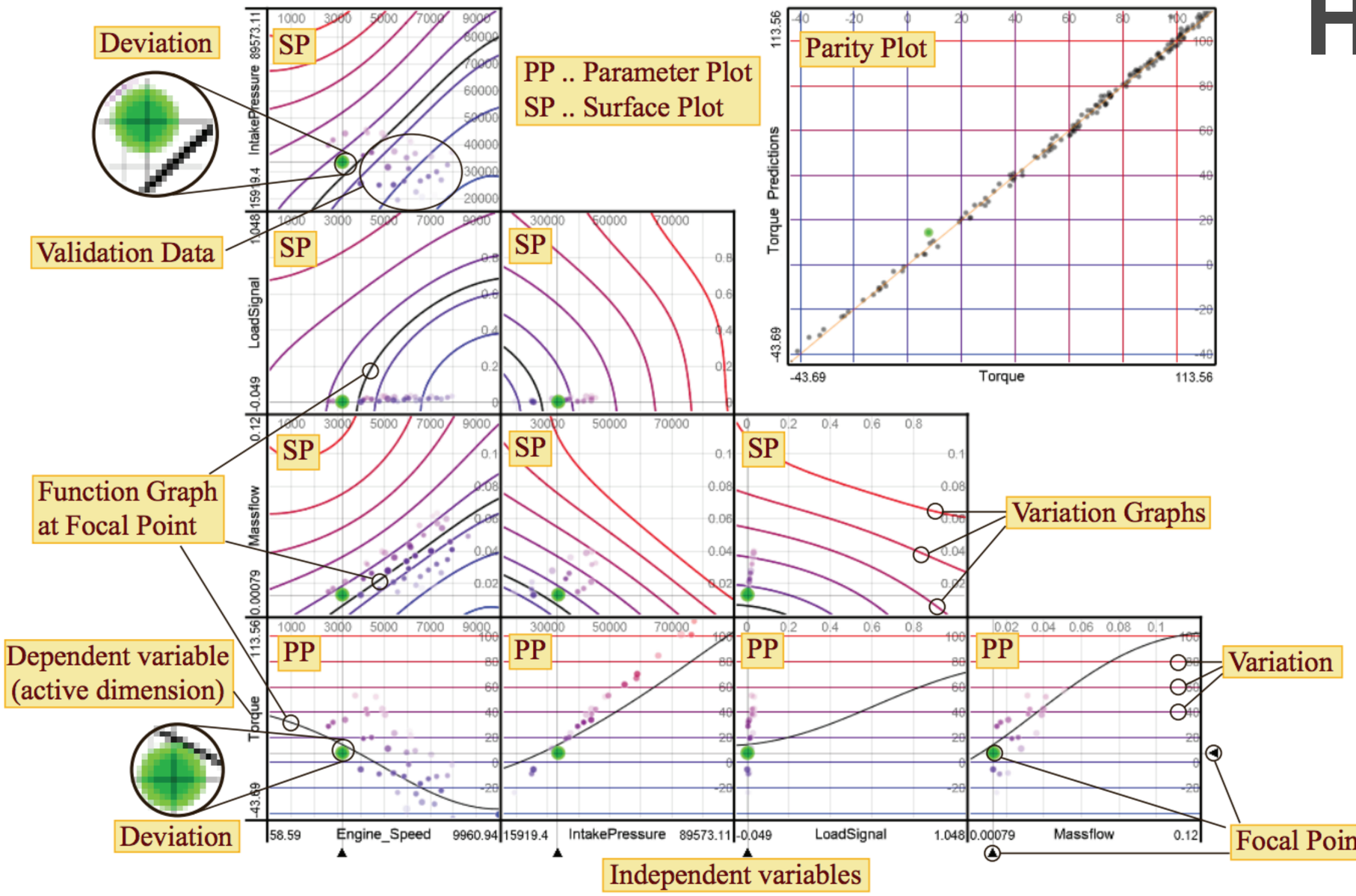
HyperMoVal: Model Validation

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

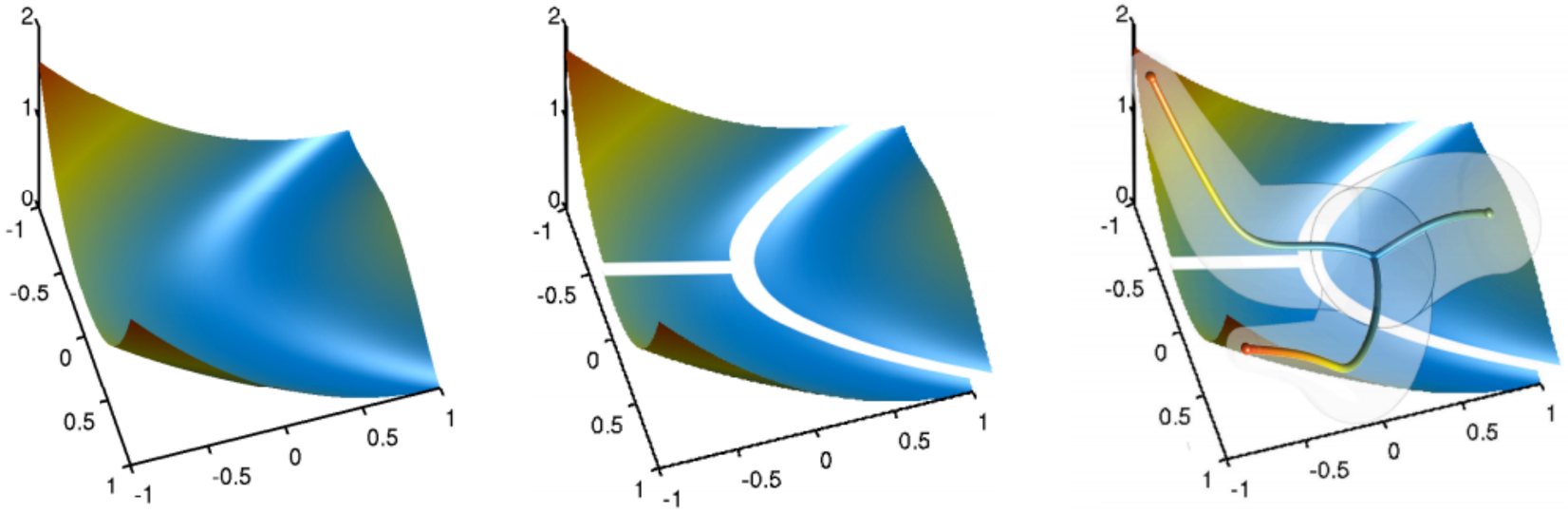
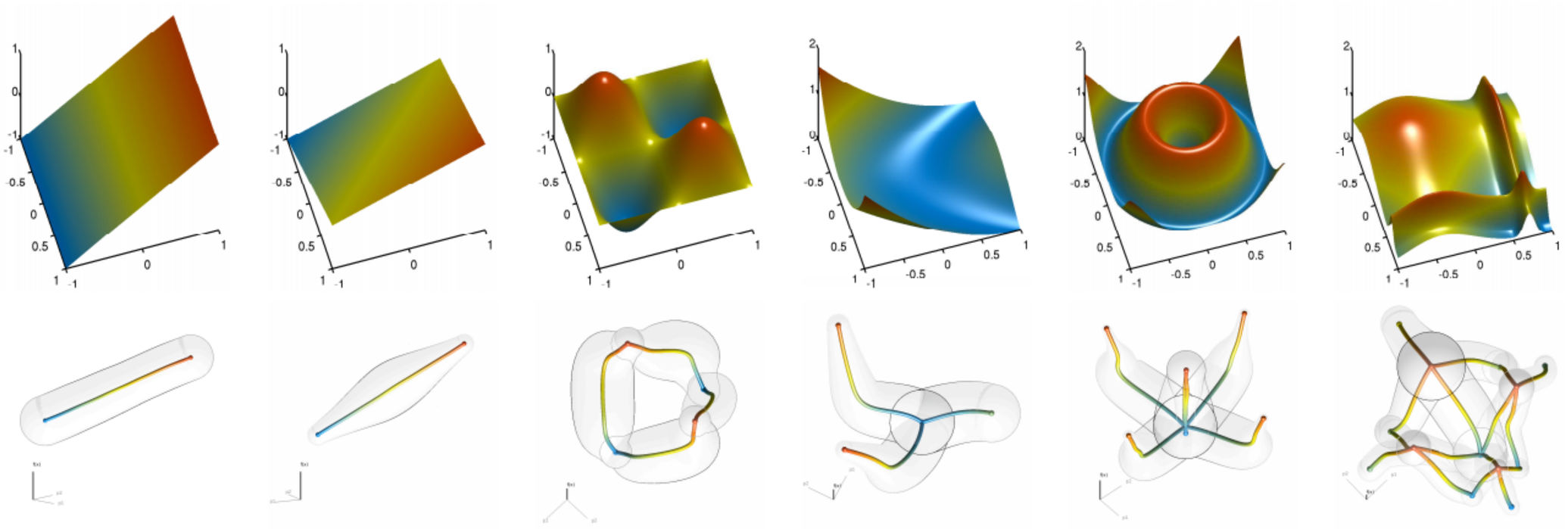


Fig. 3. Schematic illustration of the proposed method. The scalar function (a) is decomposed into piecewise monotonic regions (b) and each region is approximated by a regression curve (c).

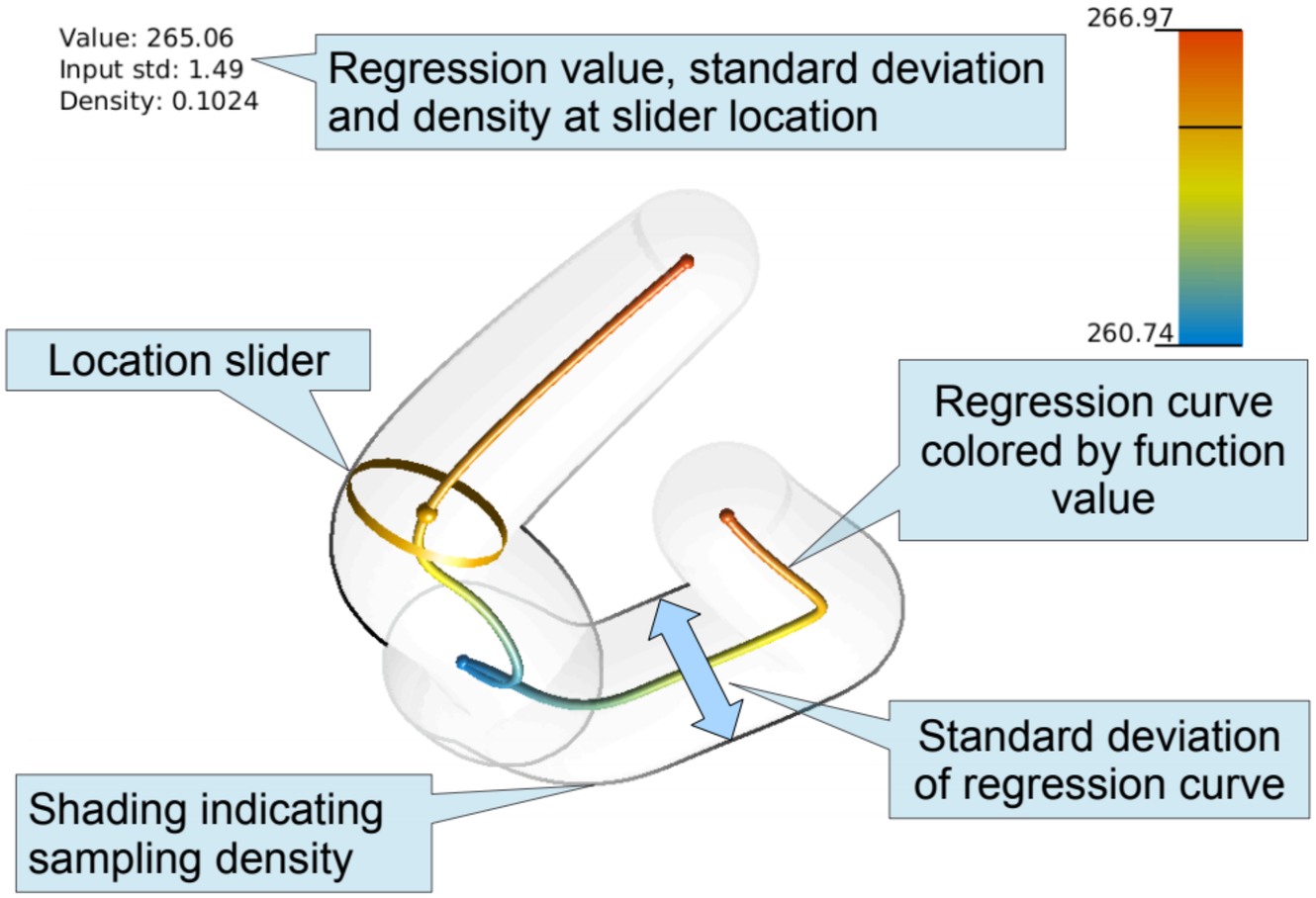
https://www.sci.utah.edu/software/hdvis.html [GerberBremerPacucci2010]

HDViz



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

HDViz



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

HDViz: Case Study Combustion

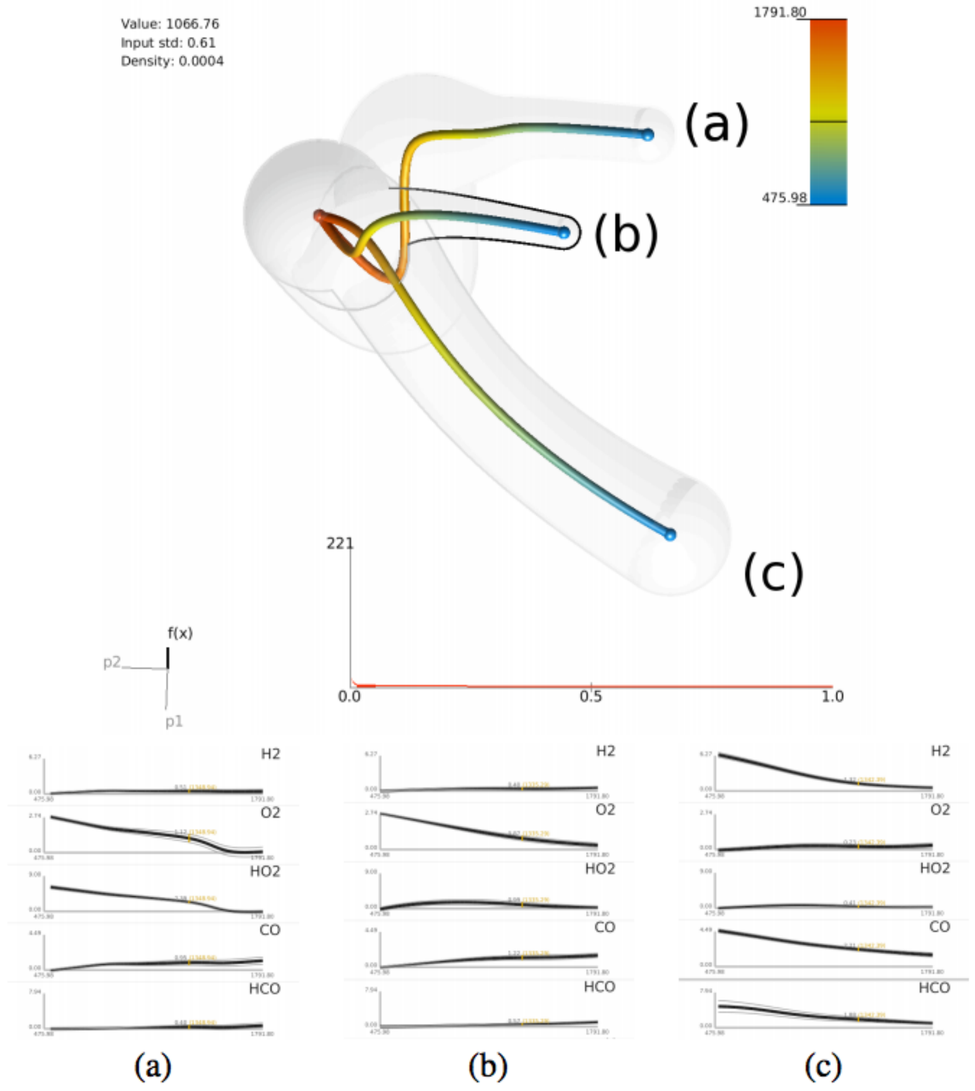
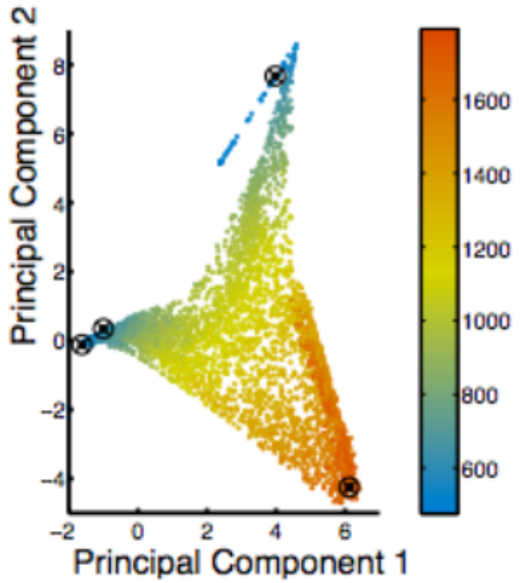


Fig. 15. Chemical composition in relation to heat released during a jet flame combustion simulation. The three distinct minima correspond to pure fuel, pure oxidizer and extinction/reignition. Graphs of chemical composition plotted against temperature for the crystals corresponding to extinction (a),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



[GerberBremerPacucci2010]

HDViz: Case Study Nuclear Simulation

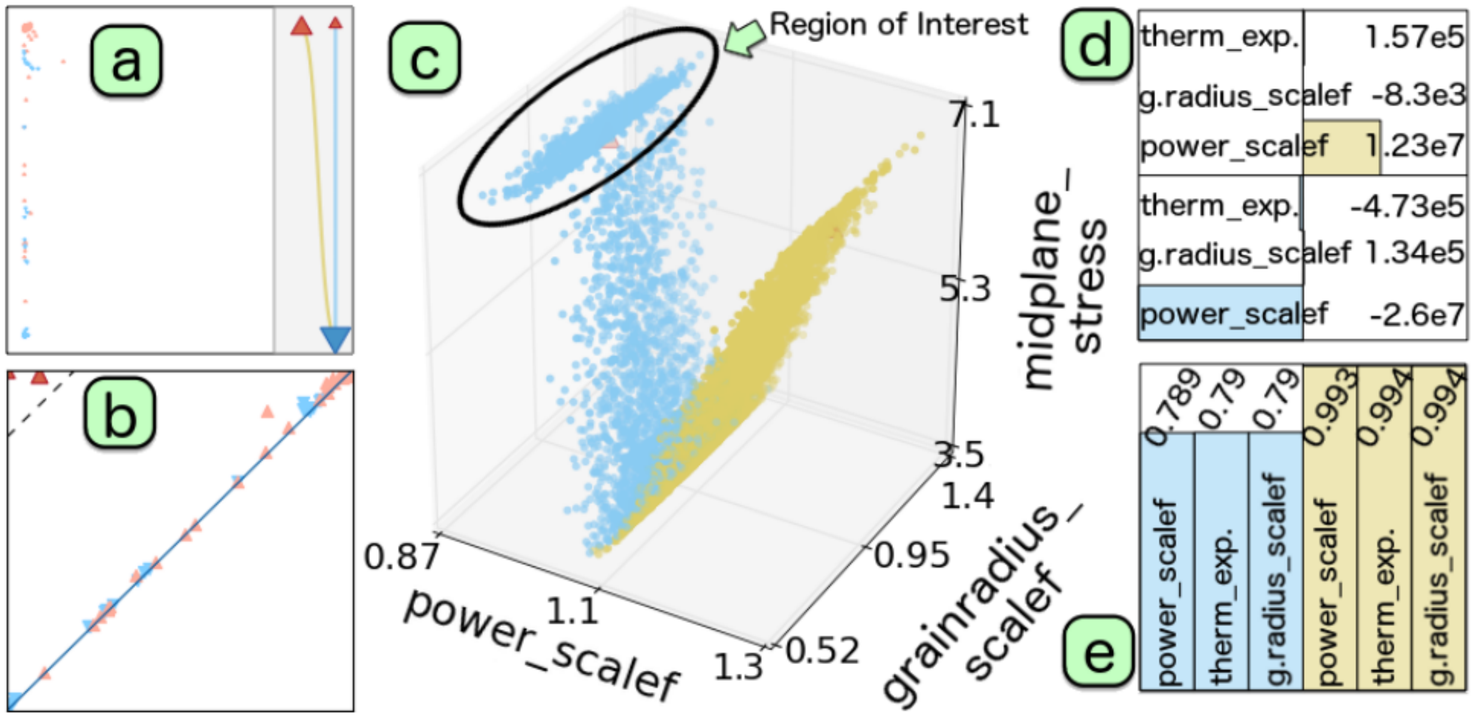
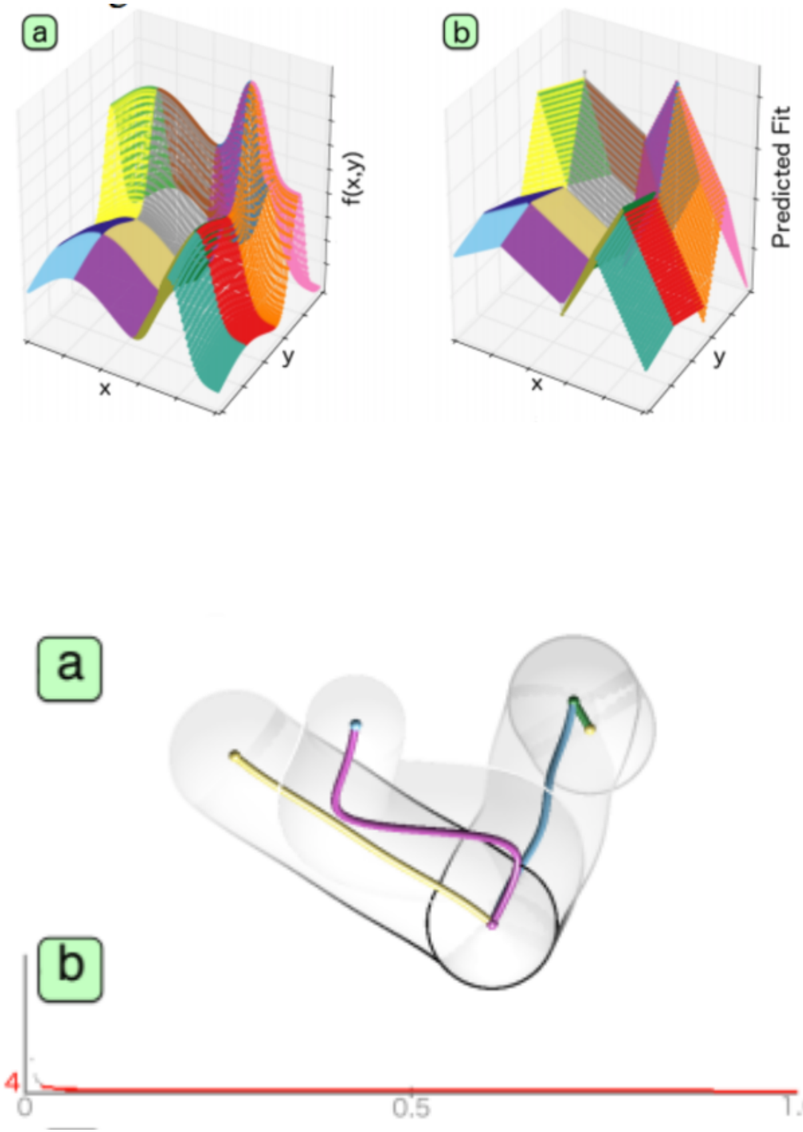


Figure 5: SA of the new nuclear fuel dataset: (a) topology map, (b) persistence diagram, (c) linked scatter plot projection, (d) linear coefficients, and (e) fitness view with stepwise R^2 scores.



[MaljovecWangRosen2016]

Take home message...

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

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