Introduction to Visual Data Science
High-Dimensional Data Visualization & Predictive Analytics

Manuela Waldner

Institute of Visual Computing & Human-Centered Technology, TU Wien, Austria

The Visual Data Science Process

PRESENTATION

Communicate the findings with key stakeholders using plots and interactive visualizations.

DATA EXPLORATION

Form hypotheses about your defined problem by **visually** analyzing the data.

PREDICTIVE MODELING

Train machine learning models, **evaluate** their performance, and **use** them to make predictions.

FEATURE ENGINEERING

Select important features and **construct** more meaningful ones using the raw data that you have.

The Data Science Process

FEATURE ENGINEERING

Select important features and **construct** more meaningful ones using the raw data that you have.

What is big data?

Key	Value
Key 1	Value 1
Key 2	Value 2
Key 3	Value 3
...	...

Key	Variable 1	Variable 2	...
Key 1	Value 1	Value 1	...
Key 2	Value 2	Value 2	...
Key 3	Value 3	Value 3	...
...	...	...	...

.... billions of records

→ **tall** data

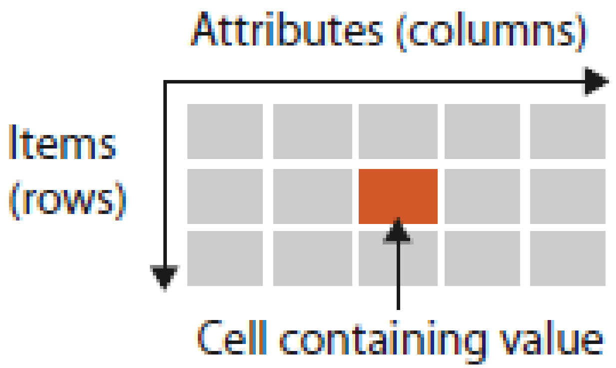
.... thousands of variables

→ **wide** data

[Heer & Kandel, Interactive Analysis of Big Data, ACM XRDS 2012]

Side Note: Features

- From machine learning / pattern recognition:
  - Measurable property of observed phenomenon
  - Vectors (can be high-dimensional!)
- In information visualization:
  - Attributes / variables / (data) dimensions

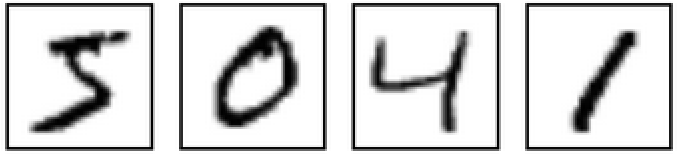


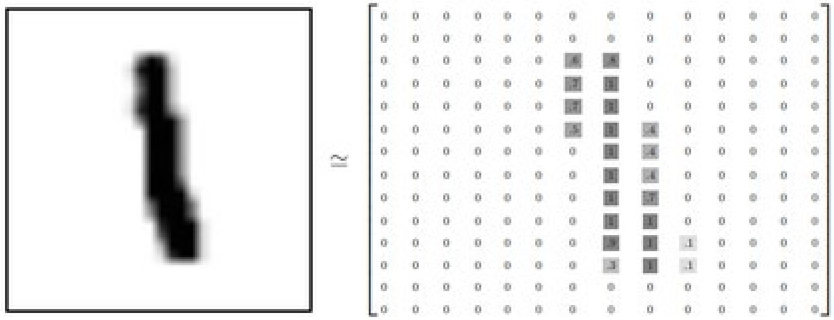
[Munzner 2014]

High-Dimensional Data

- Image features
- Natural language processing
- Gene expression data
- Finance / economy
- ...

- Image features
  - „bag of words“
  - Vocabulary of visual words
  - Example: MNIST
    - 10,000 hand-written digits
    - 28x28 pixels → 784-dimensional feature vector (intensity values) per image

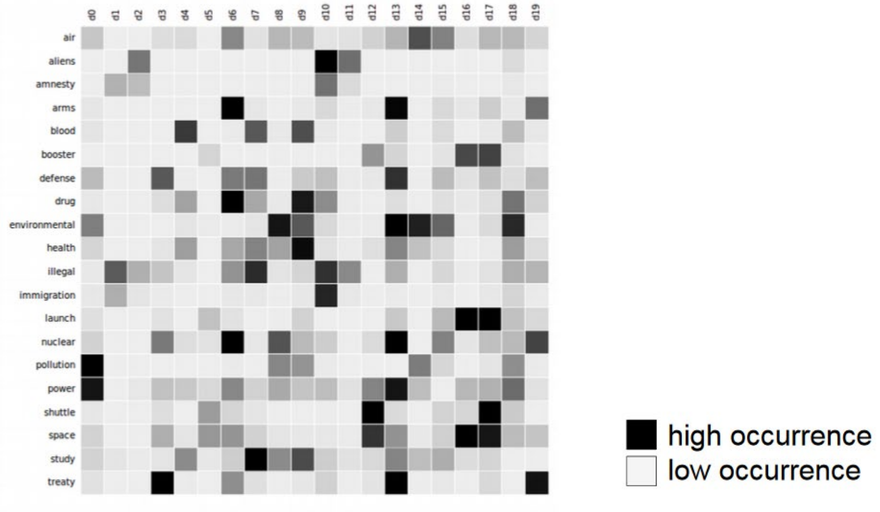




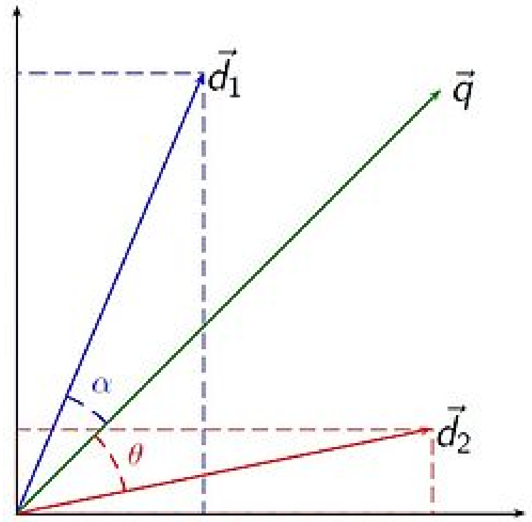
https://www.tensorflow.org

High-Dimensional Data

- Image features
- Natural language processing
  - Vector space model:
    - Dimensions: terms
    - Vectors: documents or queries

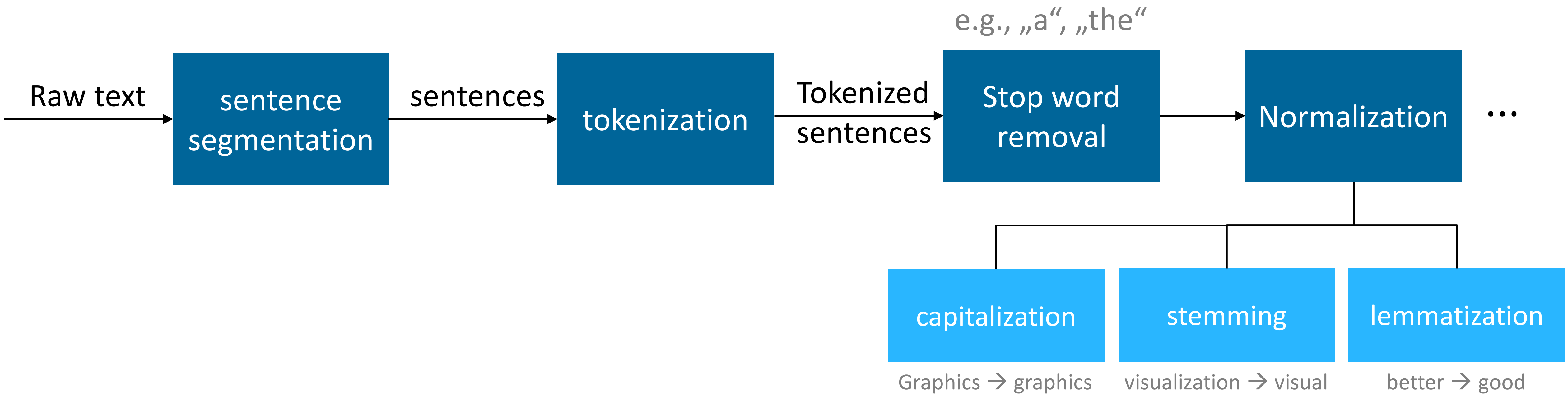


http://topicmodels.west.uni-koblenz.de/ckling/tmt/part1.pdf



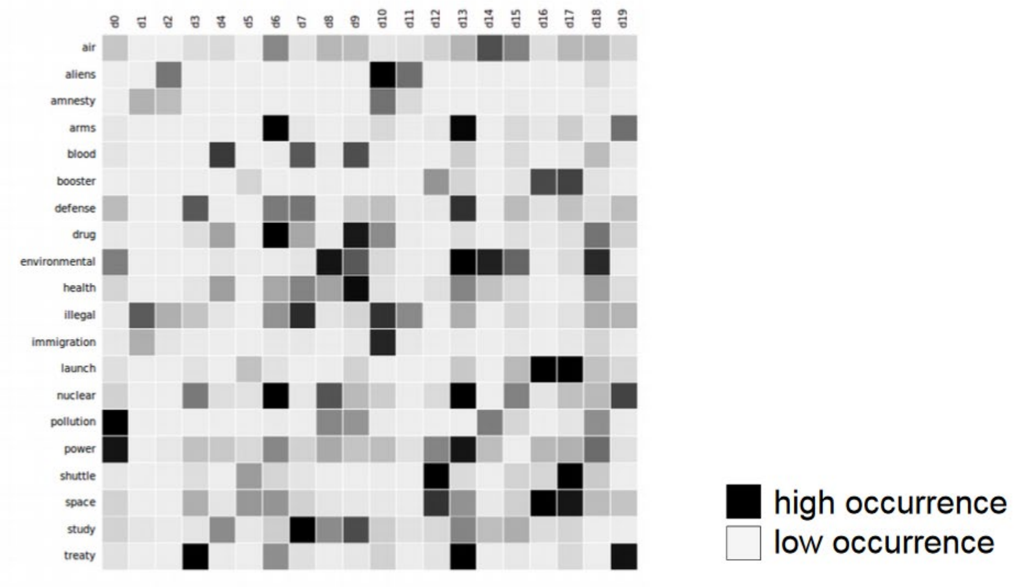
Wikipedia: vector space model

Natual Language Processing Pipeline



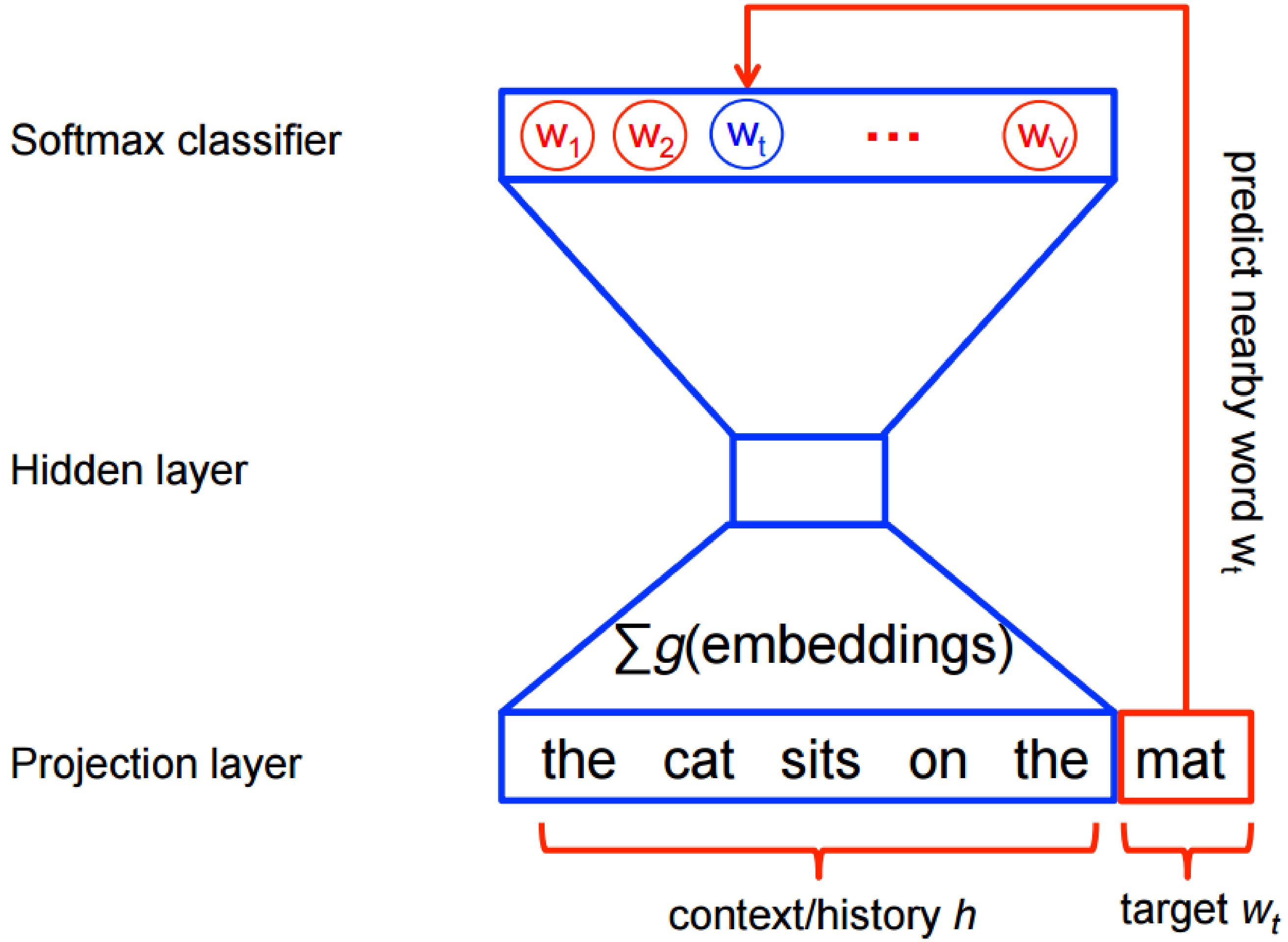
Document-Term Matrix

- Bag of words: orderless representation!
- Document is represented by vector of term weights (e.g., number of term occurrences)
- Word is represented by vector of document weights (e.g., number of occurrences in documents)



Word Embeddings

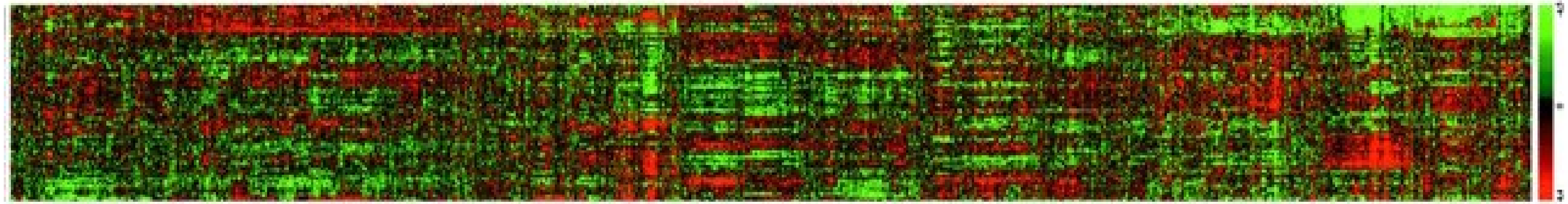
- Word2vec
  - Shallow neural network
  - Input: text window
  - Goal: prediction of nearby words
  - Output: Vector Space Model of words



https://www.tensorflow.org/tutorials/representation/word2vec

High-Dimensional Data

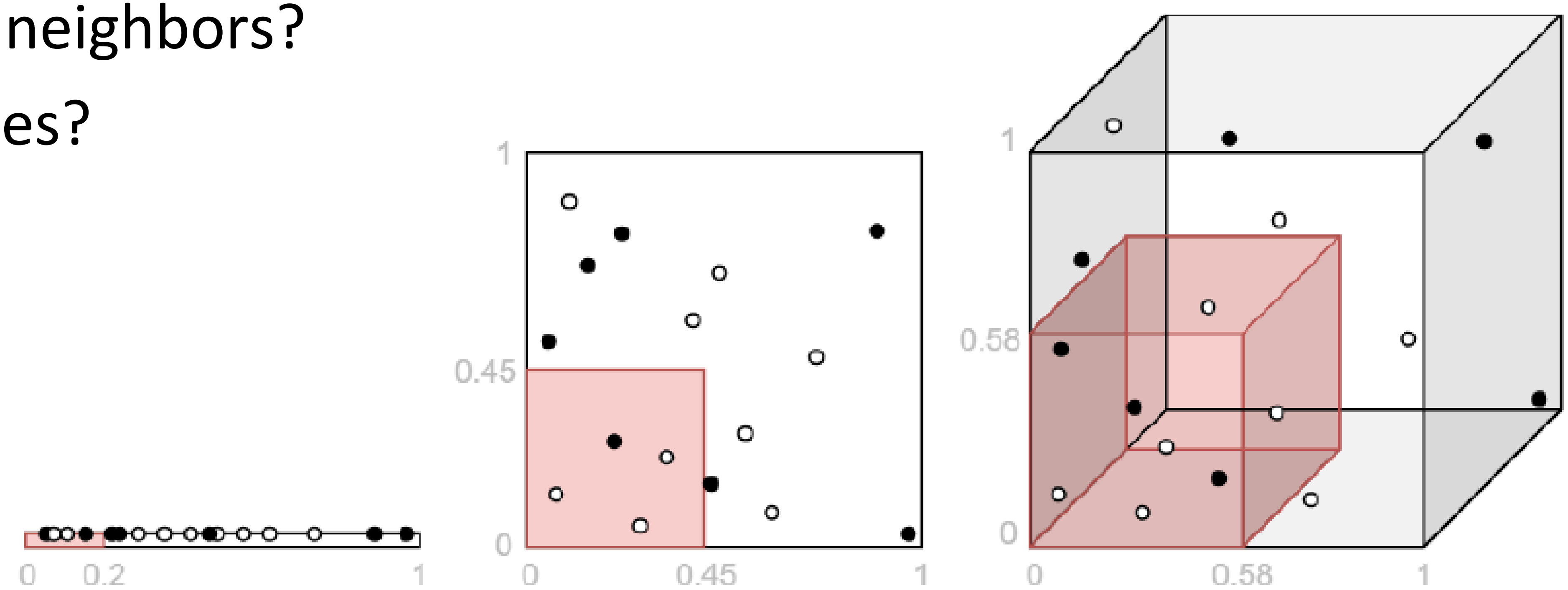
- Image features
- Natural language processing
- Gene expression data
  - Dimensions: genes
  - Samples: experimental conditions / species /...



http://cancerres.aacrjournals.org/content/64/23/8558

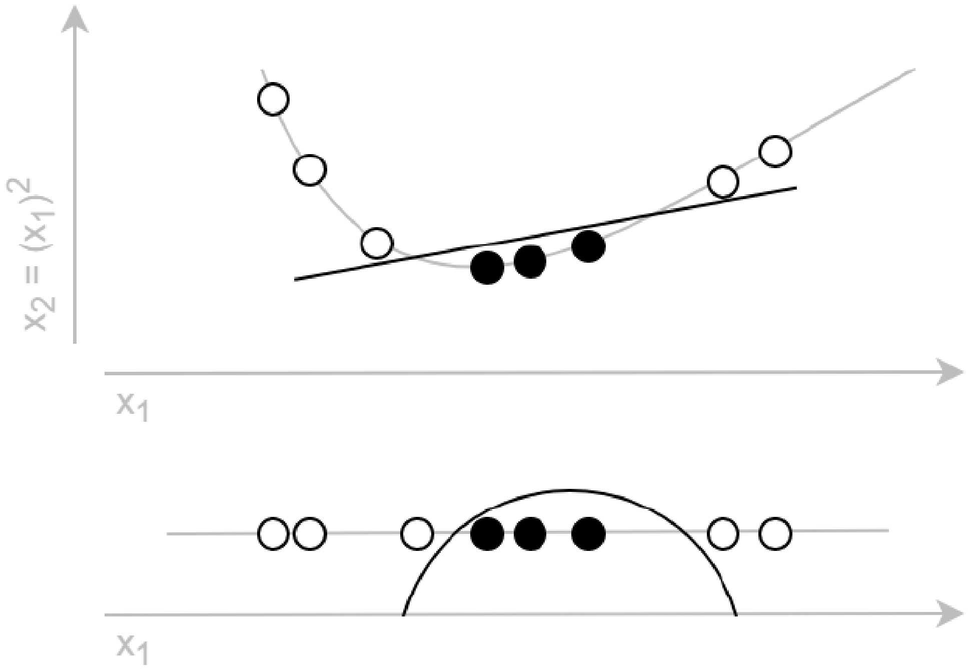
Curse of Dimensionality

- Efficiency of many algorithms depend on the number of dimensions
- With increasing number of dimensions, data becomes sparse
  - Distances increase
  - Nearest neighbors?
  - Anomalies?



Curse of Dimensionality

- Efficiency of many algorithms depend on the number of dimensions
- With increasing number of dimensions, data becomes sparse
- Number of required training samples grows exponentially with the number of dimensions
  - Rule of thumb: 5 samples per dimension minimum



- Visually inspect the features!

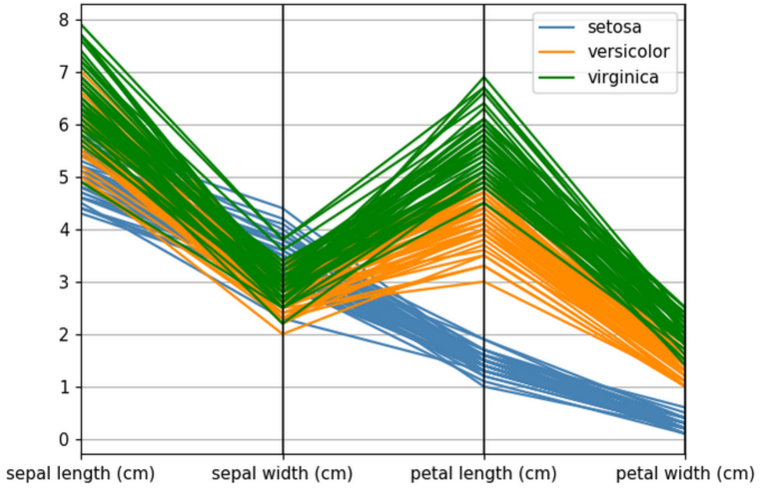
- Example: Iris dataset
  - 3 species:

Wikipedia: Iris flower data set

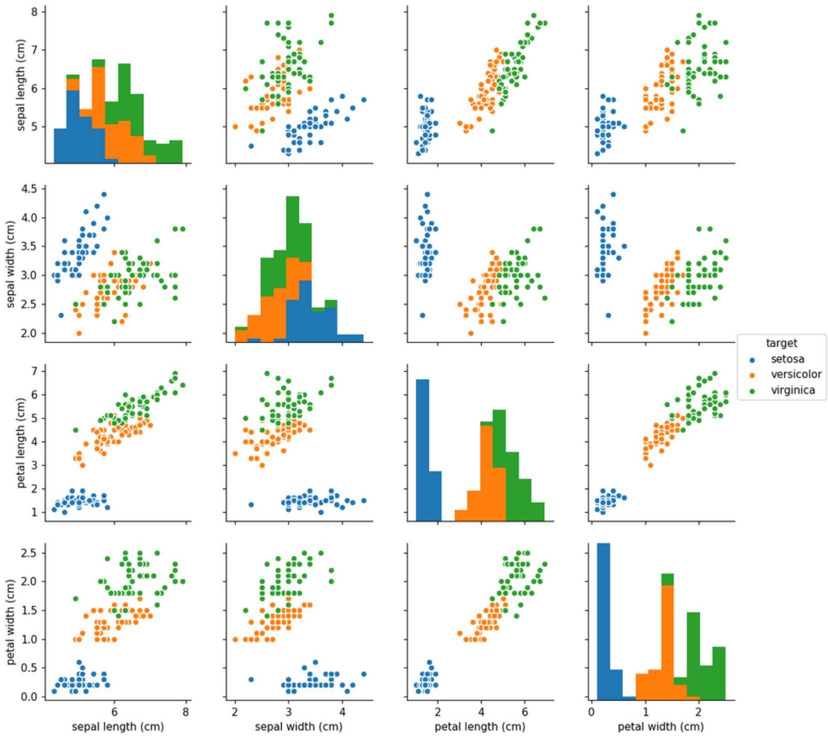
  - 50 samples per species
  - 4 features: length and width of sepals and petals

Multi-Dimensional Data Visualization Techniques

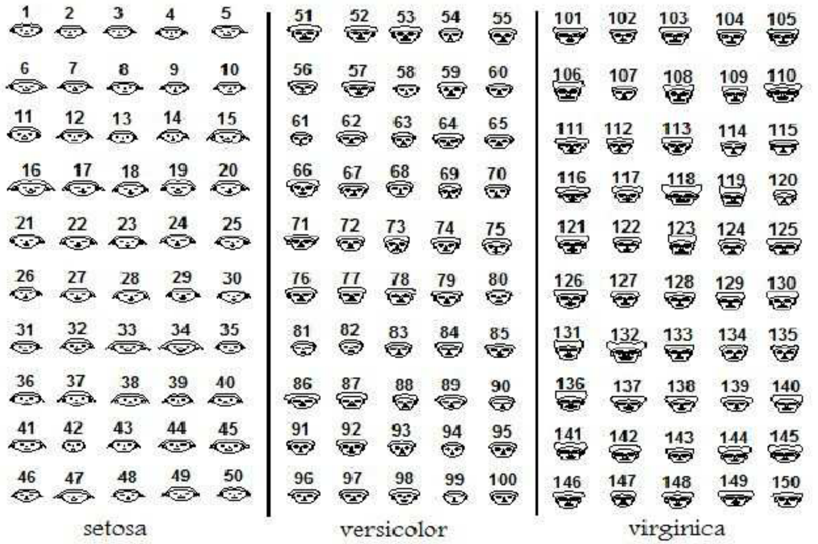
Parallel Coordinates



Scatterplot Matrix

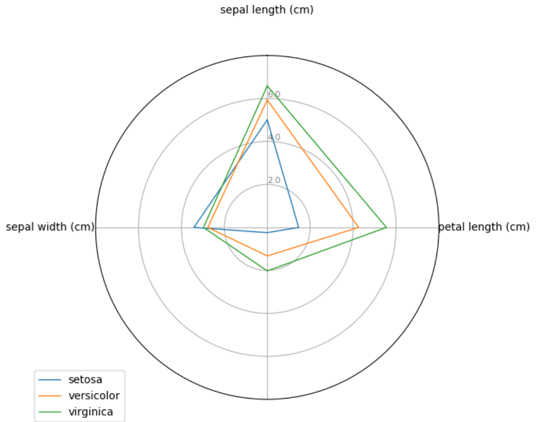


Chernoff Faces



[Icke & Sklar, 2009]

Radar Chart

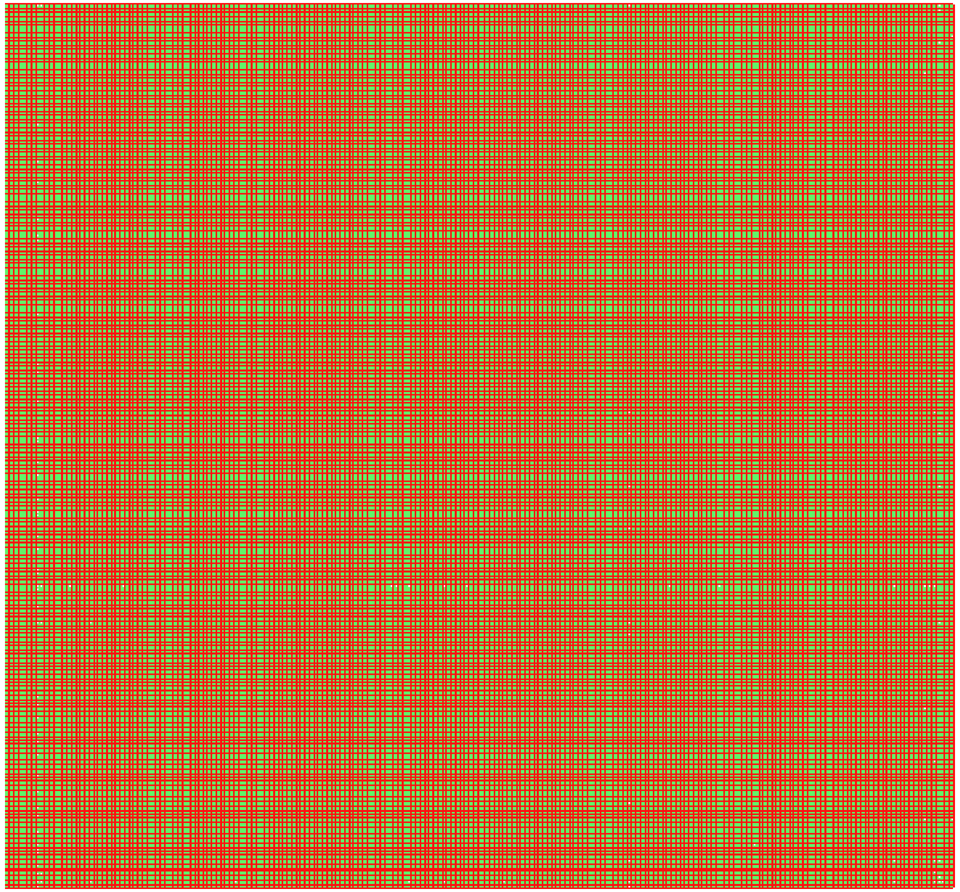


Scalability problems!

Example: Scatterplot Matrix

19 dimensions

~100 dimensions



[Yang et al., 2003]

Approaches

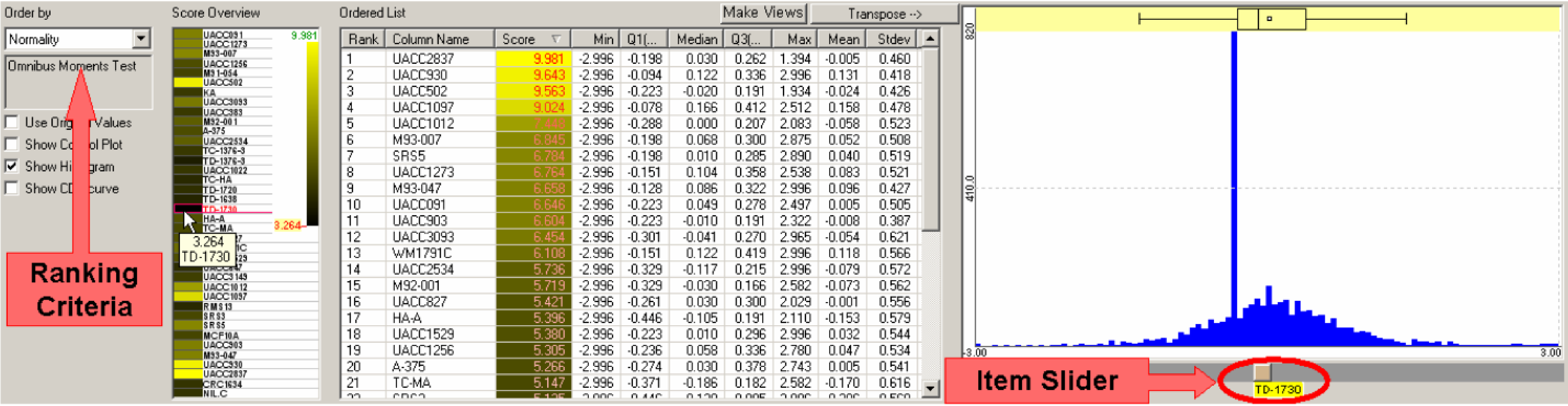
- **Feature selection**
  - Selecting a subset of existing features without a transformation
  - Using multi-dimensional data visualization techniques
- **Feature extraction**
  - Transforming existing features into lower dimensional space
  - Using 1D / 2D (/3D)/nD visualization technique
- **Hybrid approach**
  - Selecting a subset of existing features
  - Transforming feature subset into lower dimensional space

Feature Selection

- Selecting a subset of existing features without a transformation
- Dimensions (or dimension pairs) are ranked based on **quality metric**:
  - Number of outliers
  - Correlation between pair of dimensions
  - Image-based
  - ...
- Quality metrics can be combined
- Visualizing one / two / multiple dimensions of the samples

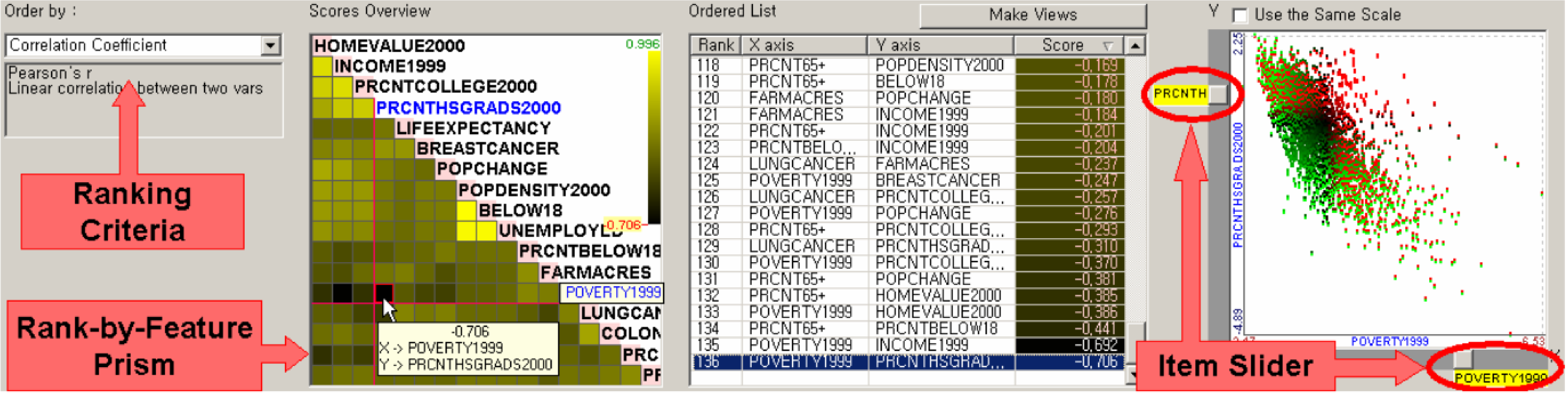
Rank-by-Feature Framework

- Exploratory analysis of multidimensional data
- Based on ranking criteria, axis-parallel projections are ranked
  - 1D ranking criteria: Normality or uniformity (entropy) of distribution, number of potential outliers, number of unique values



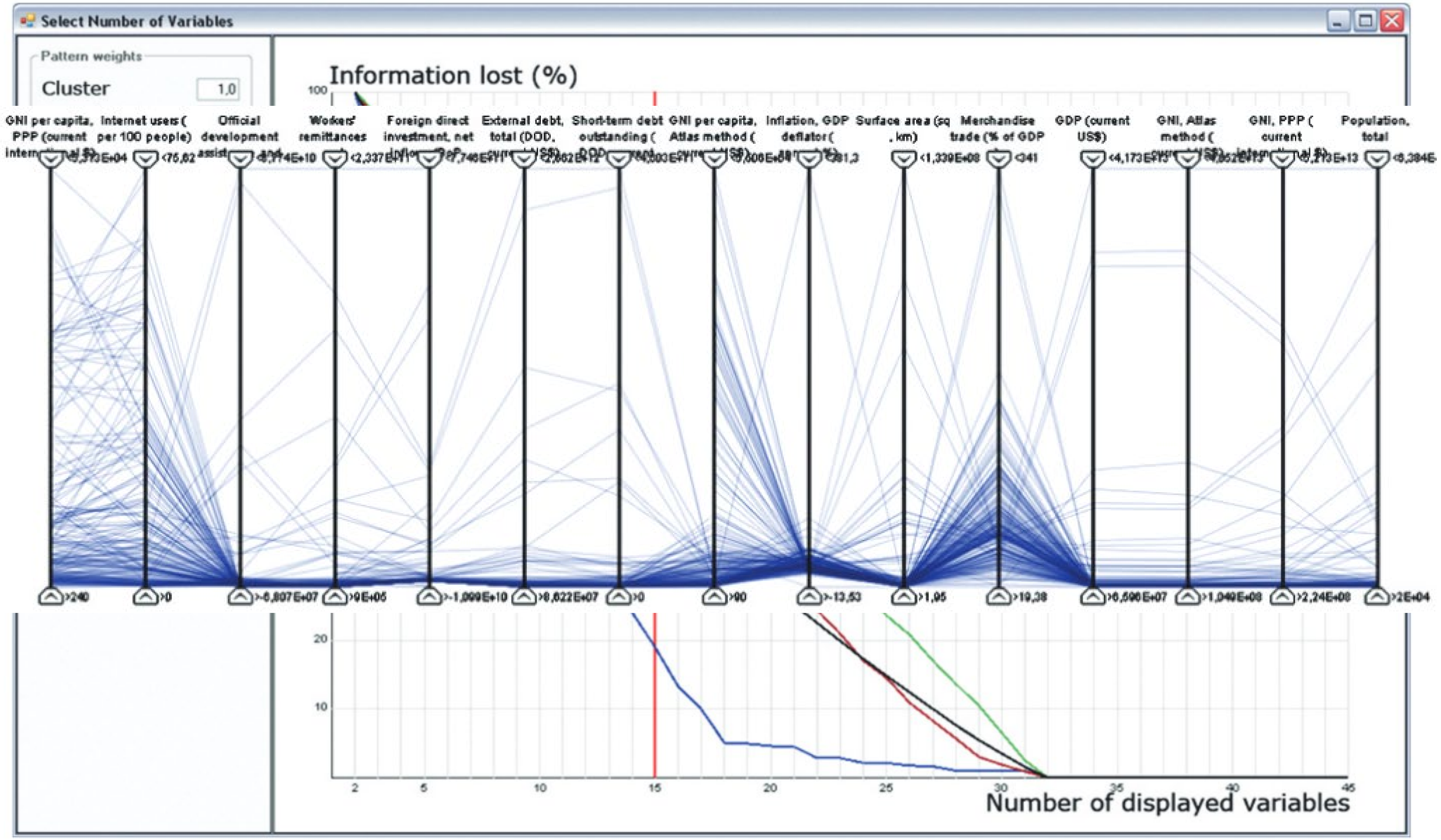
[Seo and Shneiderman, 2004]

- 2D ranking criteria:
  - Correlation coefficient, least squares error for linear regression / curvilinear regression, number of items in region of interest, uniformity of scatterplots



[Seo and Shneiderman, 2004]

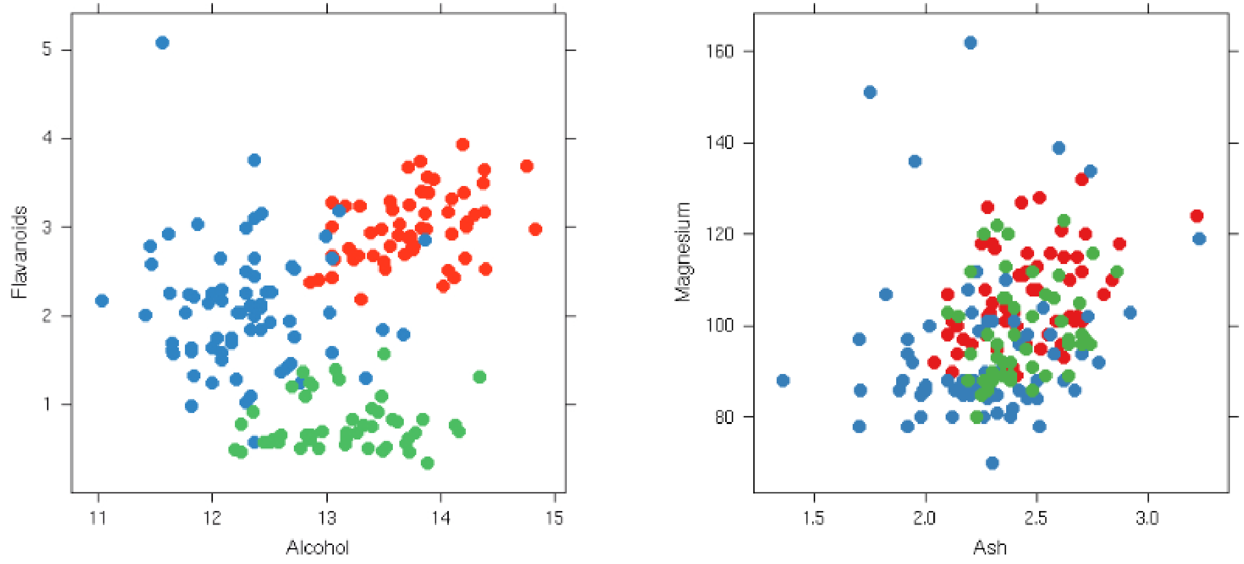
- Predefine number of dimensions to be visualized
- Based on quality metrics
  - Correlation between dimensions
  - Preservation of outliers
  - Cluster quality
- Assigns importance to each dimension



[Johansson and Johansson, 2009]

Class Consistency

- Given: points in high-dimensional space with external class labels
- Class consistency: classes are mapped to regions that are visually separable (~ ratio of data points closest to their class centroid)
- Example:
  - 3 classes of wine (color)
  - 13 attributes describing chemical properties



(a) **DSC**=90   (b) **DSC**=49

[Sips et al., 2009]

Feature Extraction

- Transforming existing features into lower dimensional space
- Dimensionality reduction
  - Linear
  - Non-linear
- Using 1D / 2D (/3D)/nD visualization technique
- Interactive visualizations can be used to steer feature extraction

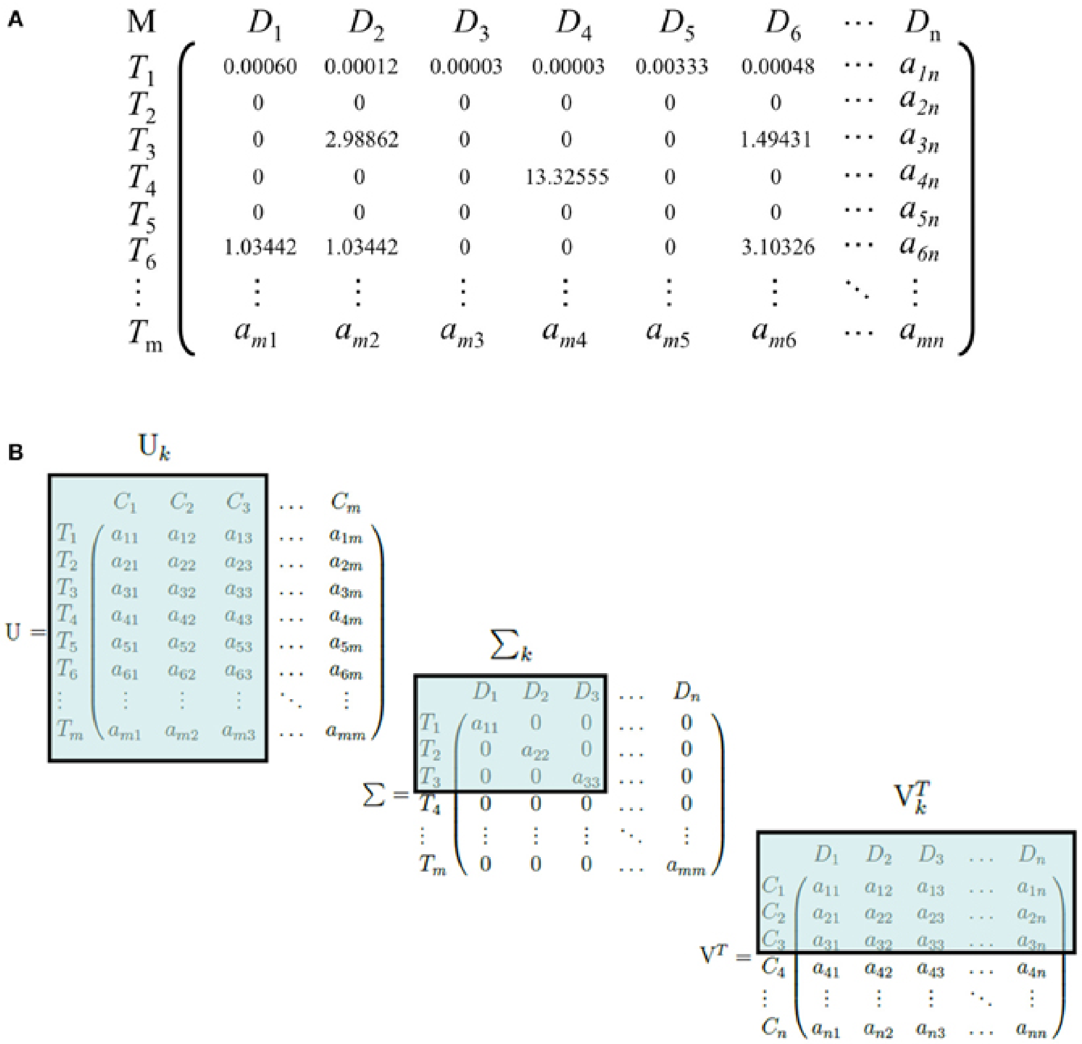
Dimensionality Reduction

- **Linear projection**
  - Linear transformation projecting data from high-dimensional space to low-dimensional space
  - Example: find subset of terms accurately clustering documents
  - Techniques:
    - Principal component analysis (PCA)
    - (metric) multi-dimensional scaling (MDS)
    - …

Singular Value Decomposition (SVD)

$$X = U\Sigma V^T$$

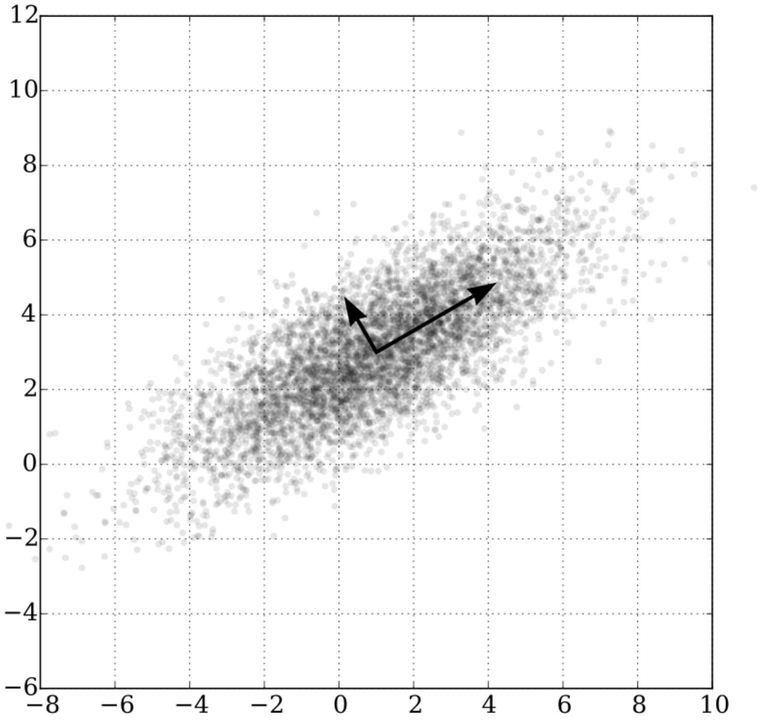
- U: term-concept matrix
- V^T: concept-document matrix
- k largest singular values and corresponding singular vectors from U and V:
- Concepts are base vectors of semantic space
- **Latent semantic indexing** = dimensionality reduction by SVD



Chen et al., Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications, 2013

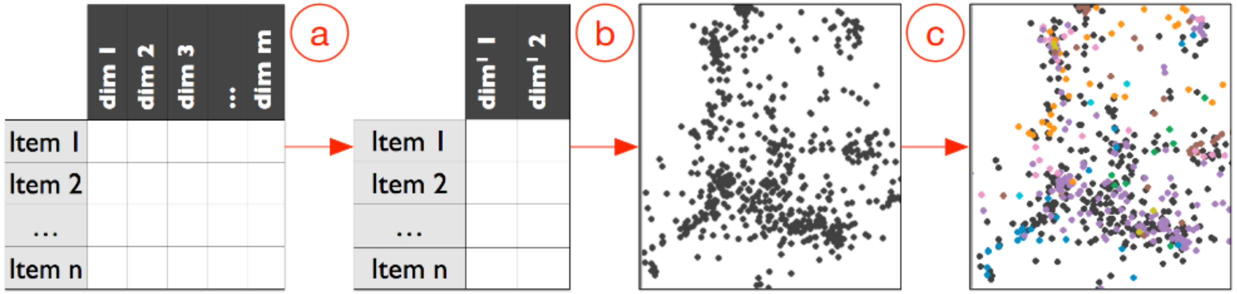
Principal Component Analysis

- SVD on centered data
- Projecting data onto lower dimensions (= principal components)
- First principal component: as much variability of the data as possible
- Principal components are orthogonal



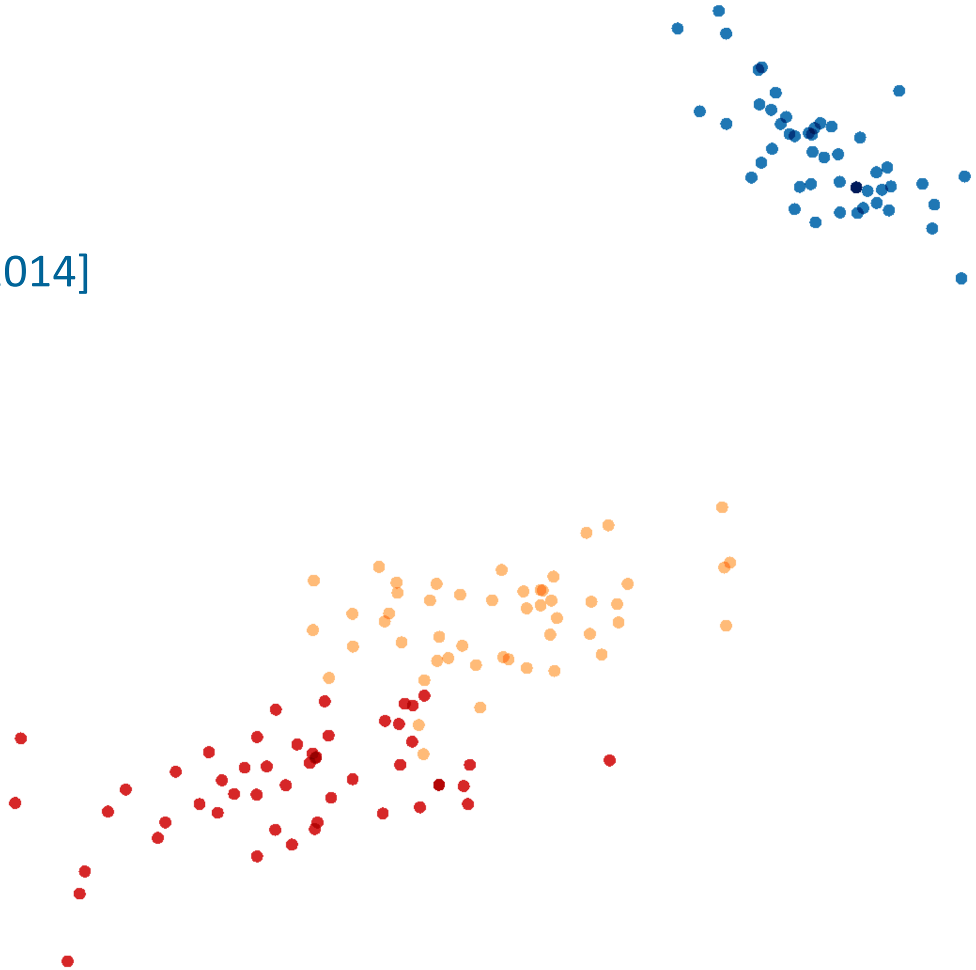
Wikipedia: principal component analysis

Visualization of Projected Data



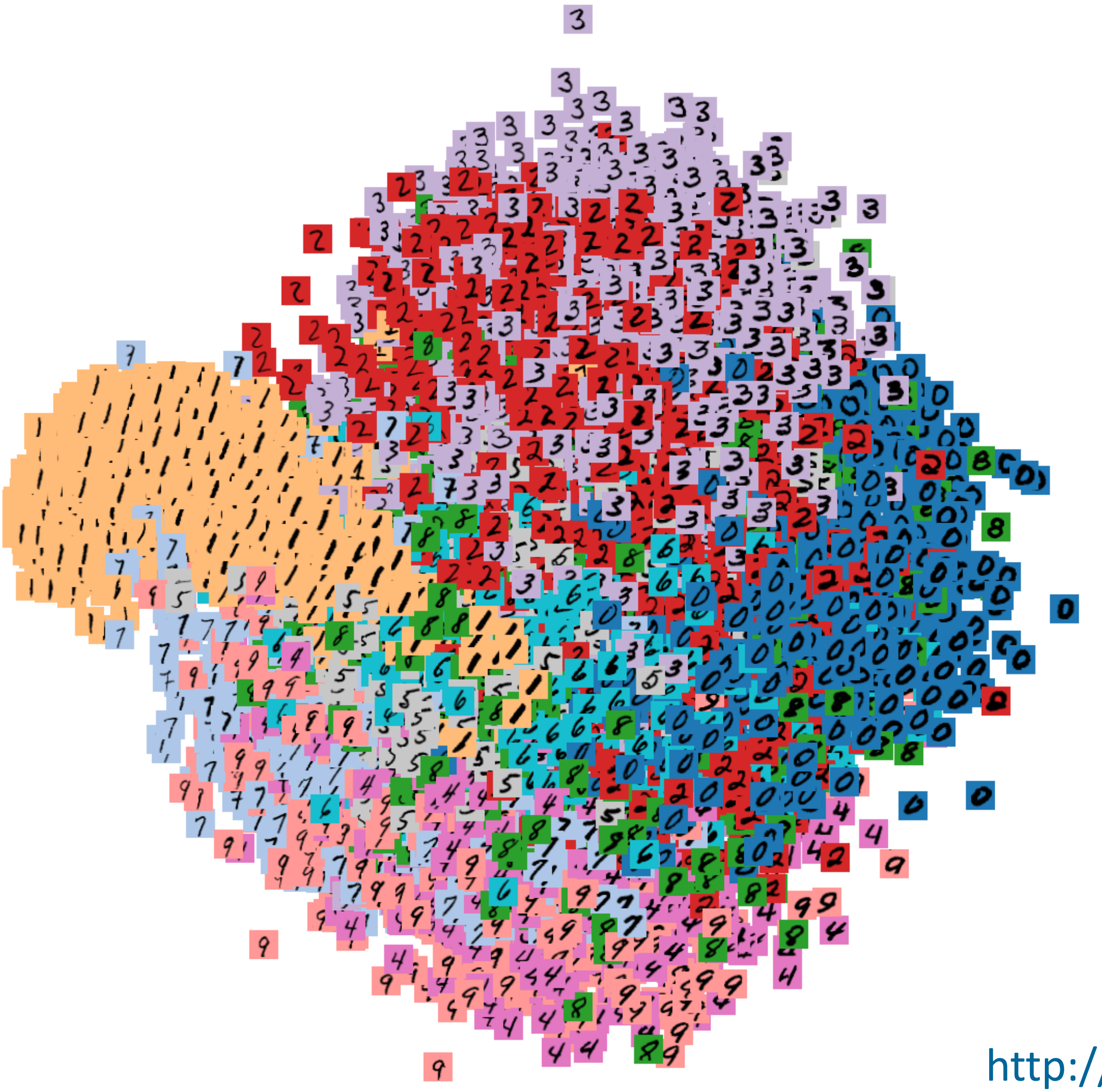
[Brehmer et al., 2014]

- Scatterplot visualization:
  - Color-coded according to classes (if available)
  - Well suited to:
    - Detect / verify / name clusters
    - Detect outliers
    - Match clusters and classes
- Example: Iris data set (PCA)



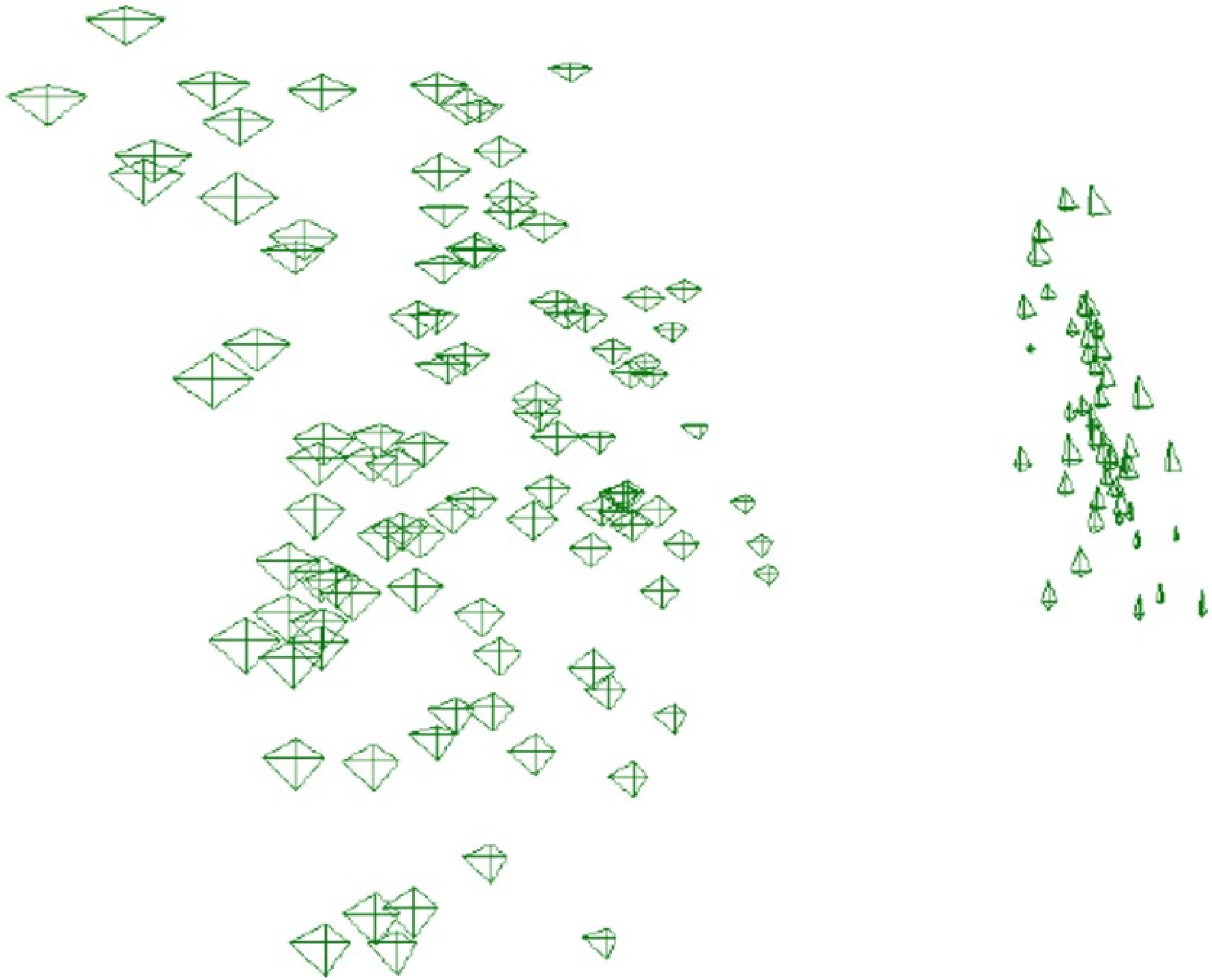
http://projector.tensorflow.org/

MNIST PCA Example



http://projector.tensorflow.org/

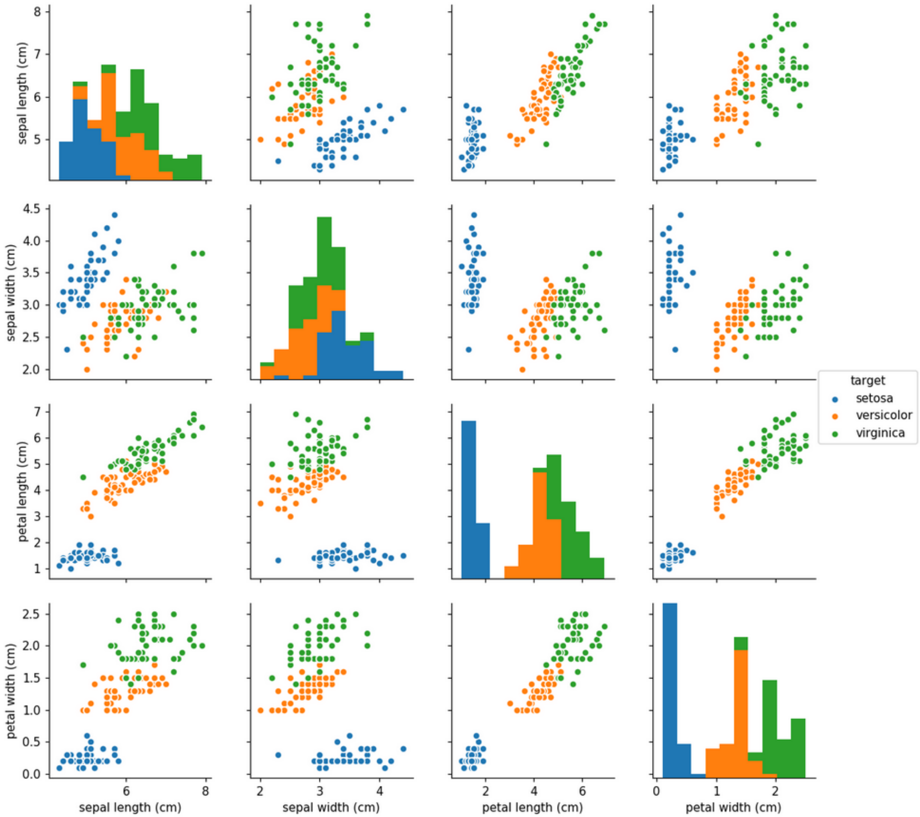
PCA with Star Glyphs (Iris Dataset)

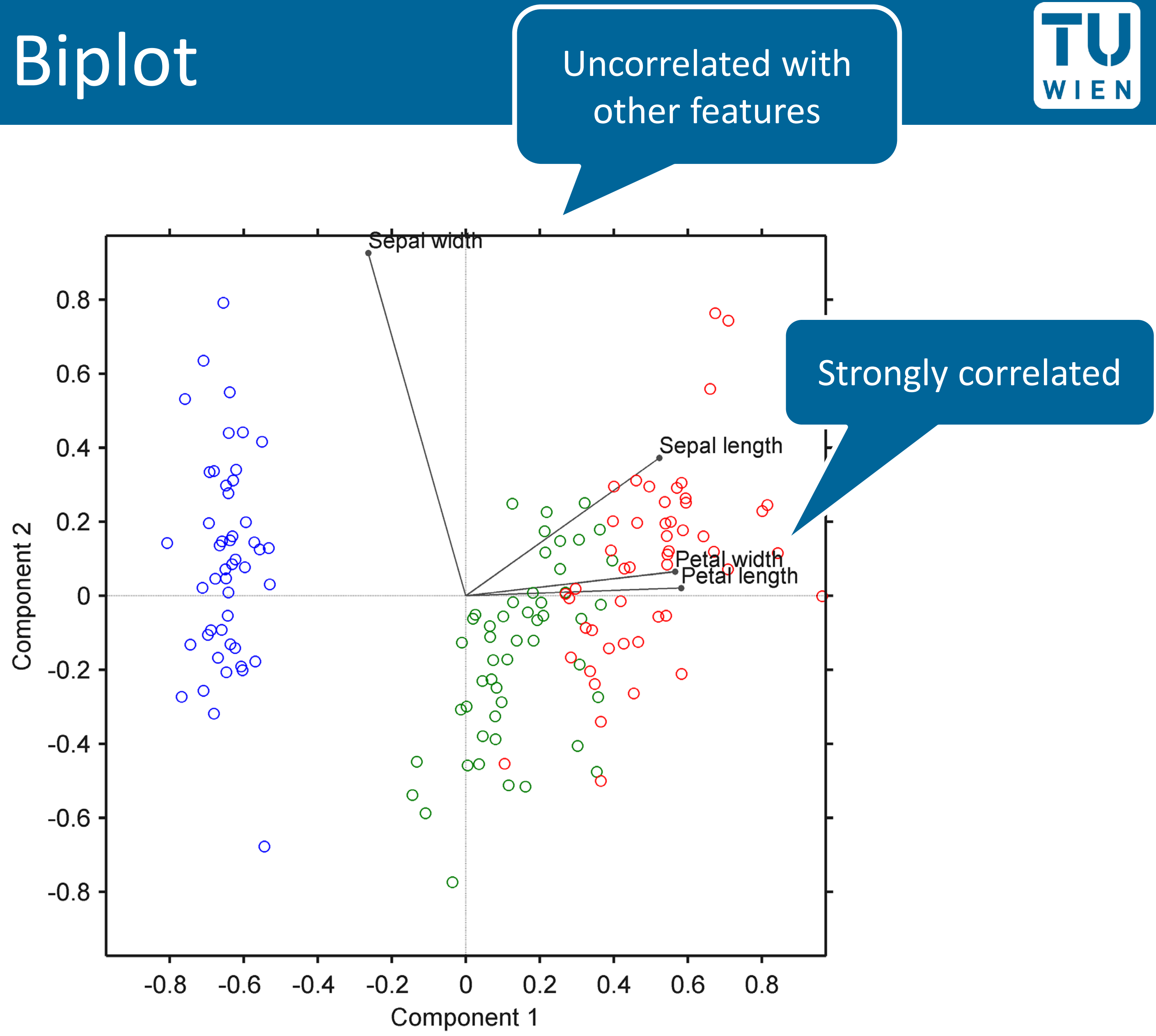


[Keim et al., 2005]

Biplot

- Axes: principal components
- Vectors: features
- Points: items



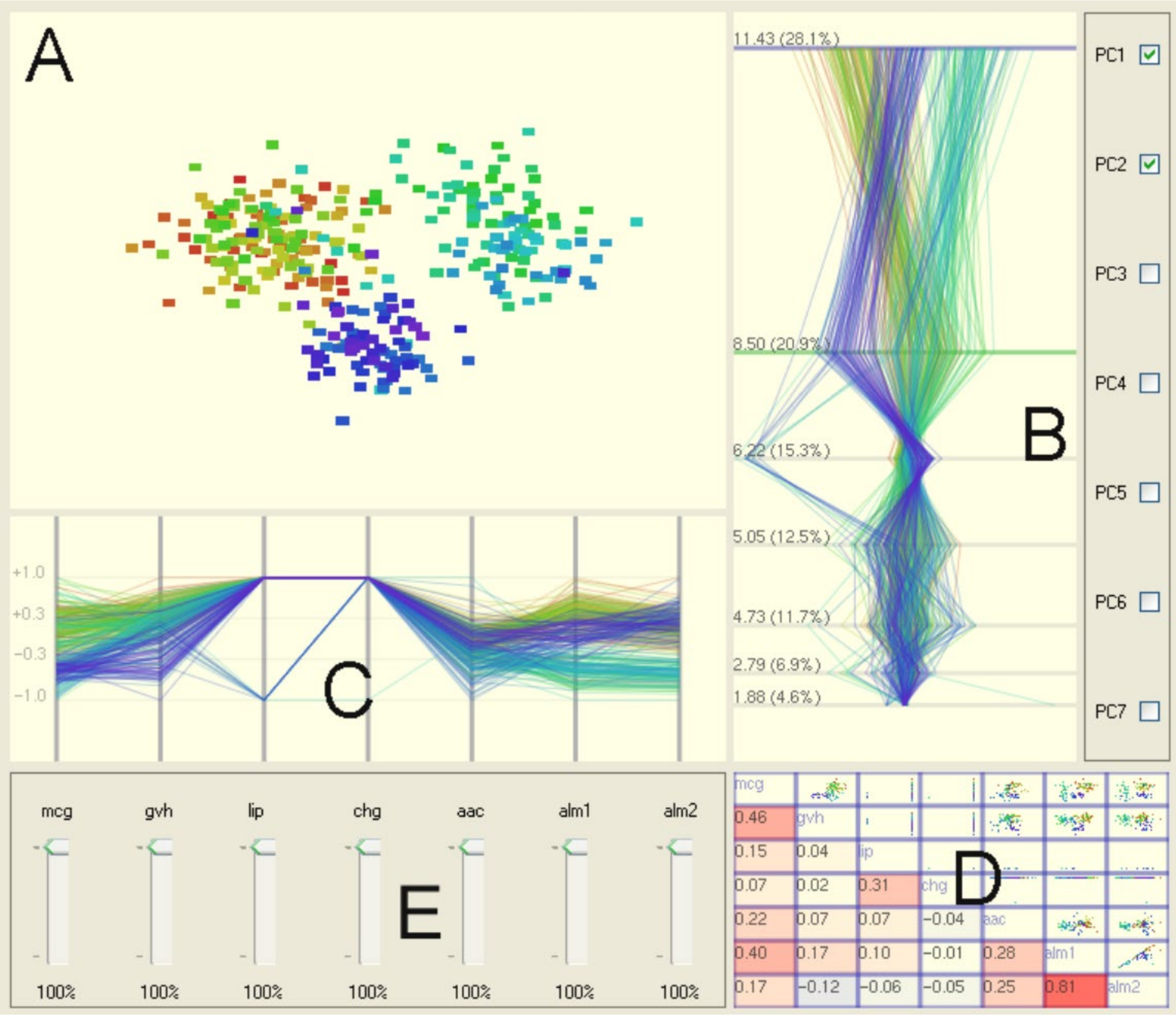


https://www.mathworks.com/matlabcentral/fileexchange/53438-biplot-by-groups

Interactive PCA-based Visual Analytics

Projection onto first two eigenvectors

Dimensions of the original data



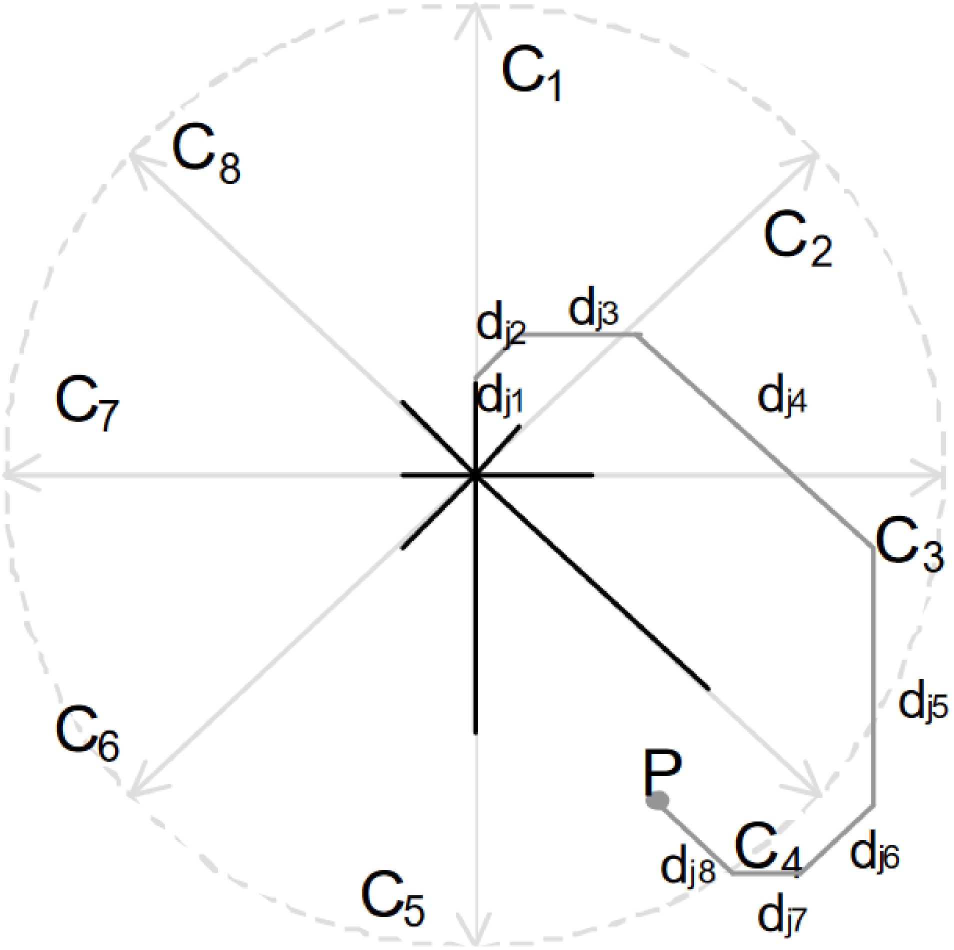
Eigenvectors as parallel coordinate axes

Pearson correlation of variable pairs

[Jeong et al., EuroVis 2009]

Star Coordinates

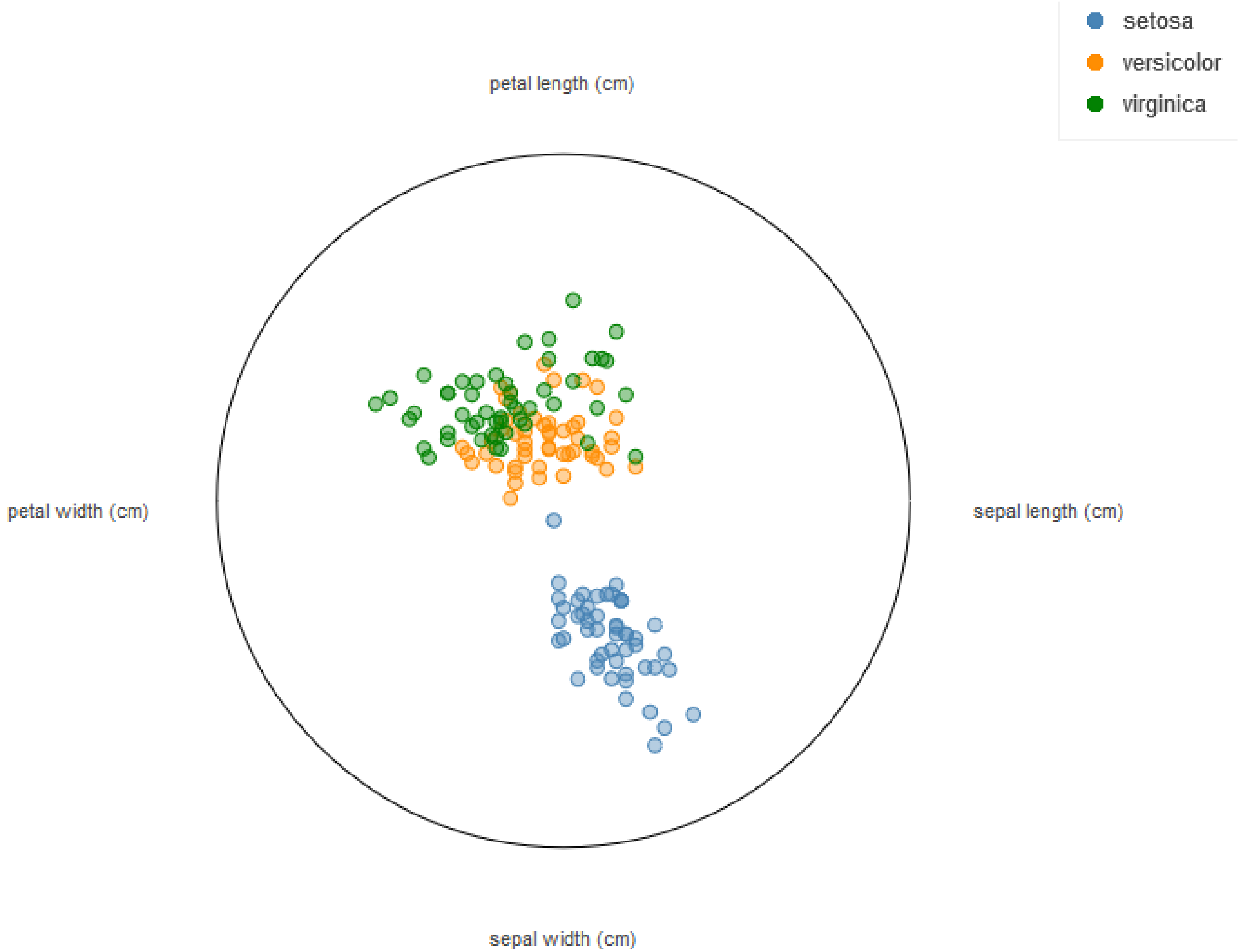
- Curvilinear coordinate system
- Items represented as points:
  - Sum of all unit vectors on each coordinate $u_i = (u_{xi}, u_{yi})$
  - Multiplied by value of data element d_j for that coordinate
  - $P_j(x, y) = \left[\begin{array}{l} o_x + \sum_{i=1}^{n} u_{xi}(d_{ji} - min_i), \\ o_y + \sum_{i=1}^{n} u_{yi}(d_{ji} - min_i) \end{array} \right]$



[Kandogan, 2000]

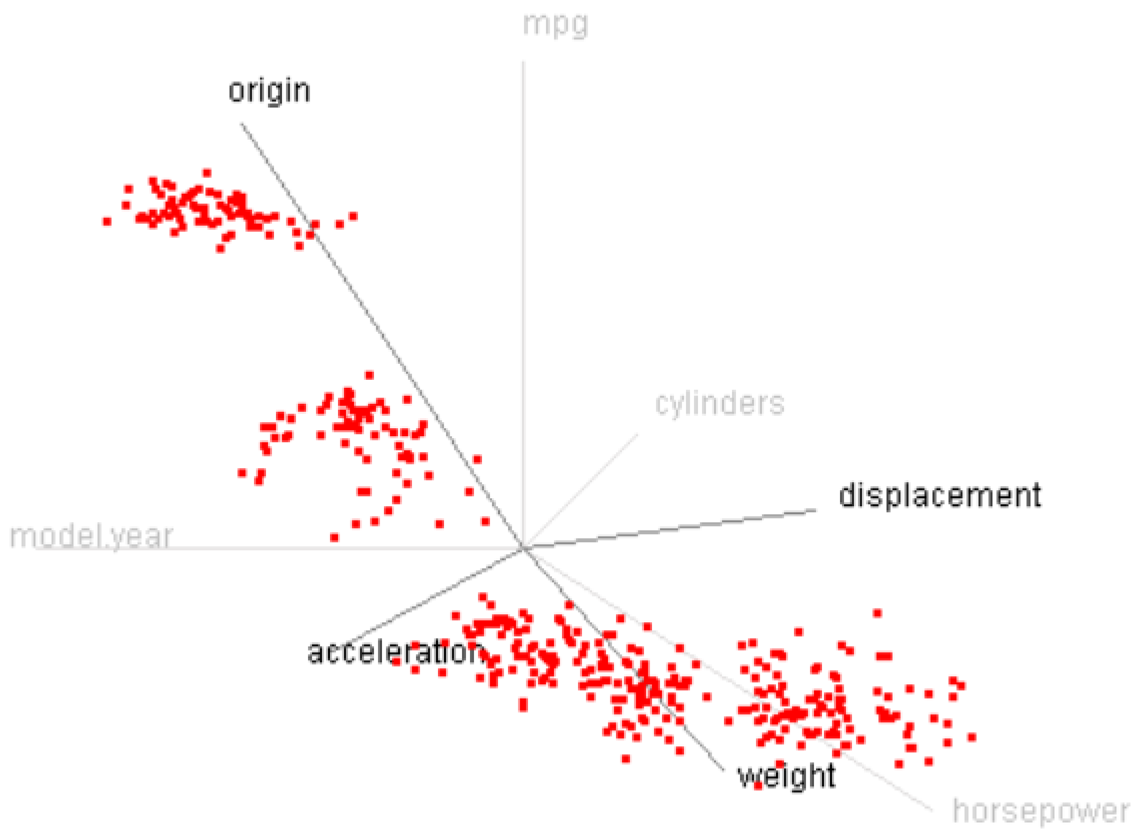
Star Coordinates

- Iris dataset:



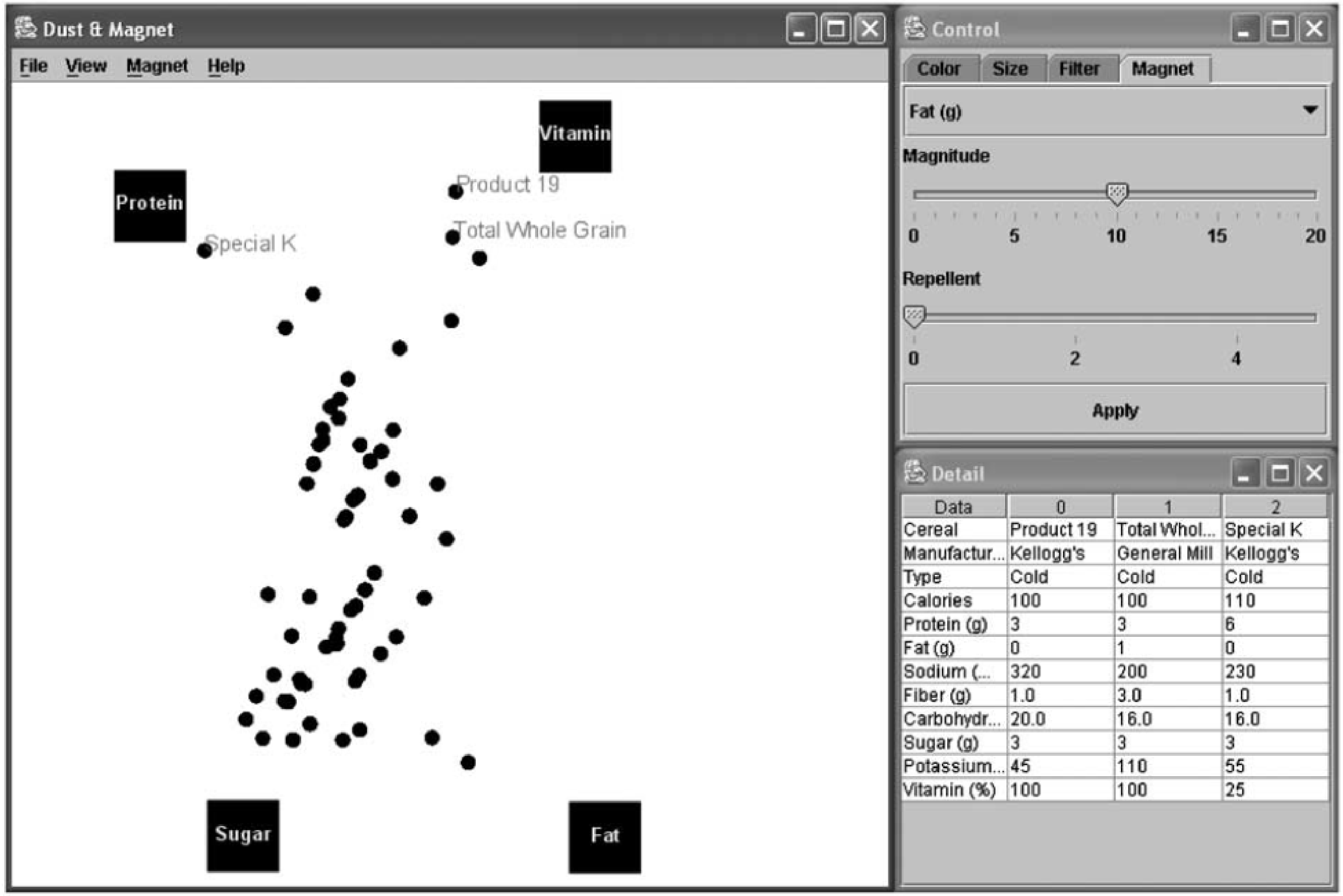
Star Coordinates

- Transformations of axes:
  - Scaling length of axis
    → changing contribution of dimension
  - Rotation of axis vector
    → change correlation with other columns
  - Switching off coordinates
    → “feature selection”



[Kandogan, 2000]

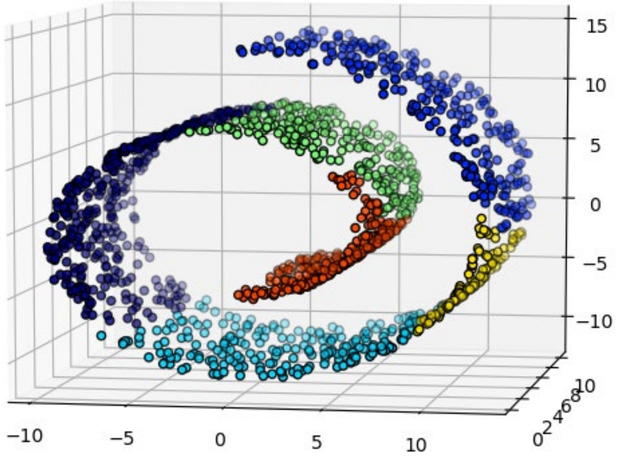
- Dimensions: magnets
- Items: dust particles
- Based on attraction forces



[Ji Soo Yi et al., 2005]

Dimensionality Reduction

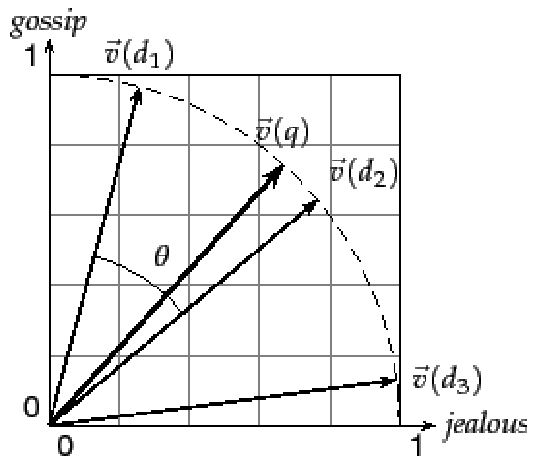
- Linear dimensionality reduction
  - Assumes that there is a lower dimensional linear subspace
  - Finds a linear projection of the data
- **Non-linear dimensionality reduction**
  - Low-dimensional surface embedded non-linearly in high-dimensional space („manifold“)
  - Preserves the neighborhood information
    - Locally linear
    - Pairwise distances



„swiss roll“
http://scikit-learn.org

Pairwise Similarities

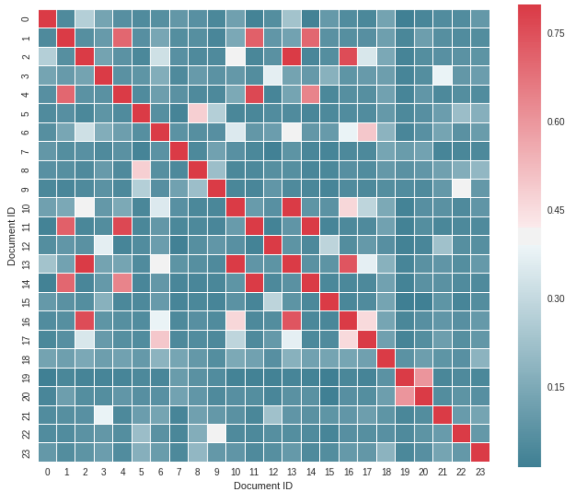
- Cosine similarity:
  - Corpus is represented by a set of vectors in vector space (axes: terms)
  - Document similarity is defined by cosine similarity between the document vectors



Cosine similarity illustrated. $\mathrm{sim}(d_1, d_2) = \cos\theta$.

http://nlp.stanford.edu/IR-book/html/htmledition/dot-products-1.html

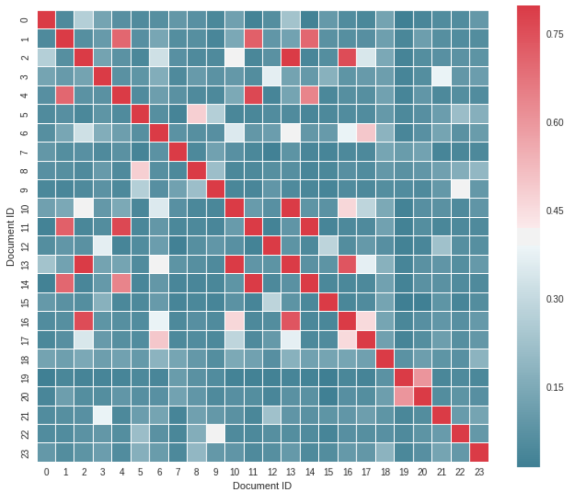
- Document similarity matrix



https://github.com/utkuozbulak/unsupervised-learning-document-clustering/blob/master/README.md

Multi-Dimensional Scaling

- Computation of low-dimensional embedding Y that best preserves pair-wise distances between data points X
  - $Cost = \sum_{i<j}(d_{ij} - \delta_{ij})$
  - $d_{ij} = \left\|x_i - x_j\right\|^2$
  - $\delta_{ij} = \left\|y_i - y_j\right\|^2$

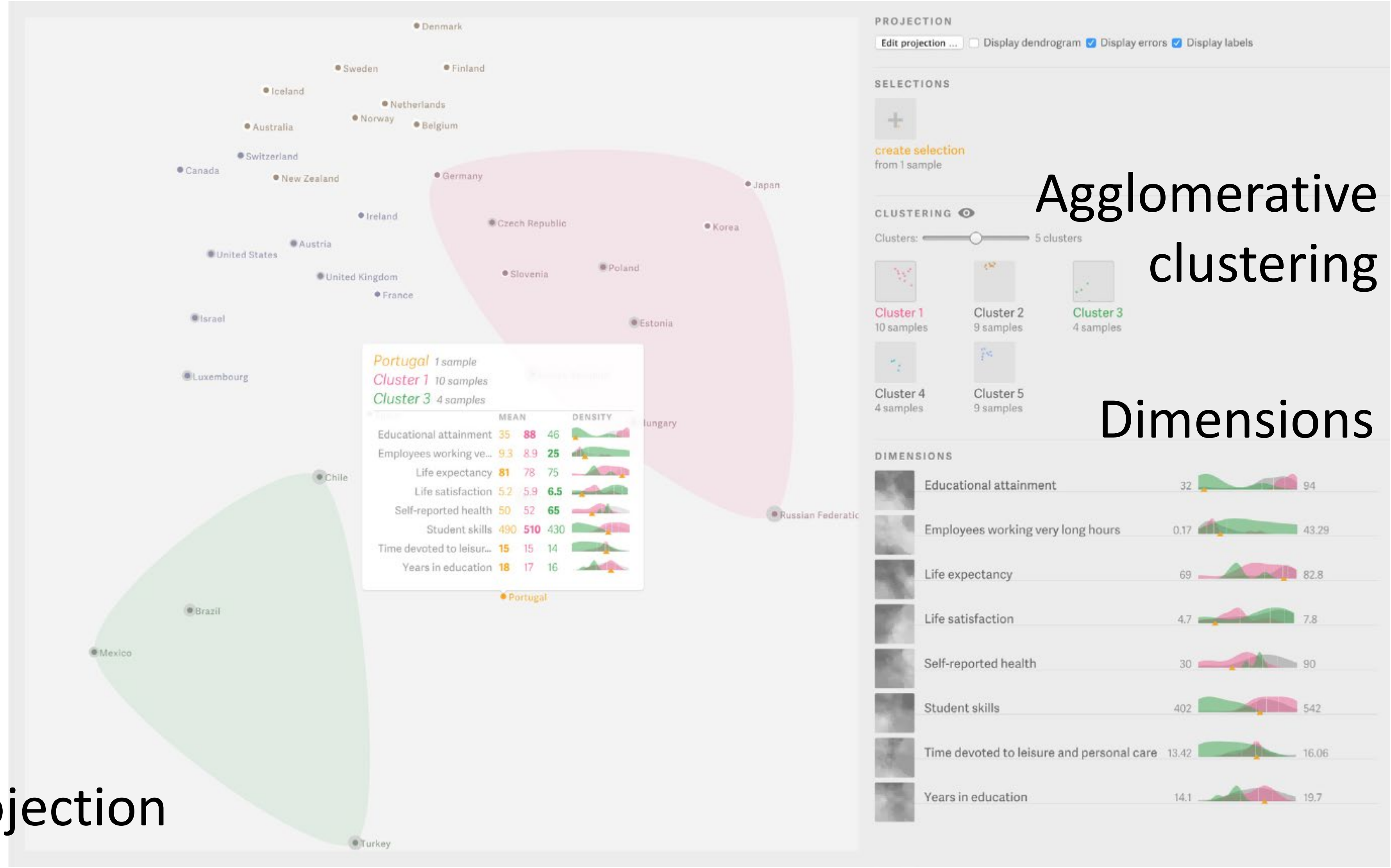


https://github.com/utkuozbulak/unsupervised-learning-document-clustering/blob/master/README.md

- Euclidean distances: MDS equivalent to PCA

Visual Analysis of Dimensionality Reductions

- Example: OECD countries:
  - 36 countries
  - 8 dimensions



Agglomerative clustering

Dimensions

MDS projection

[Stahnke et al., 2016]

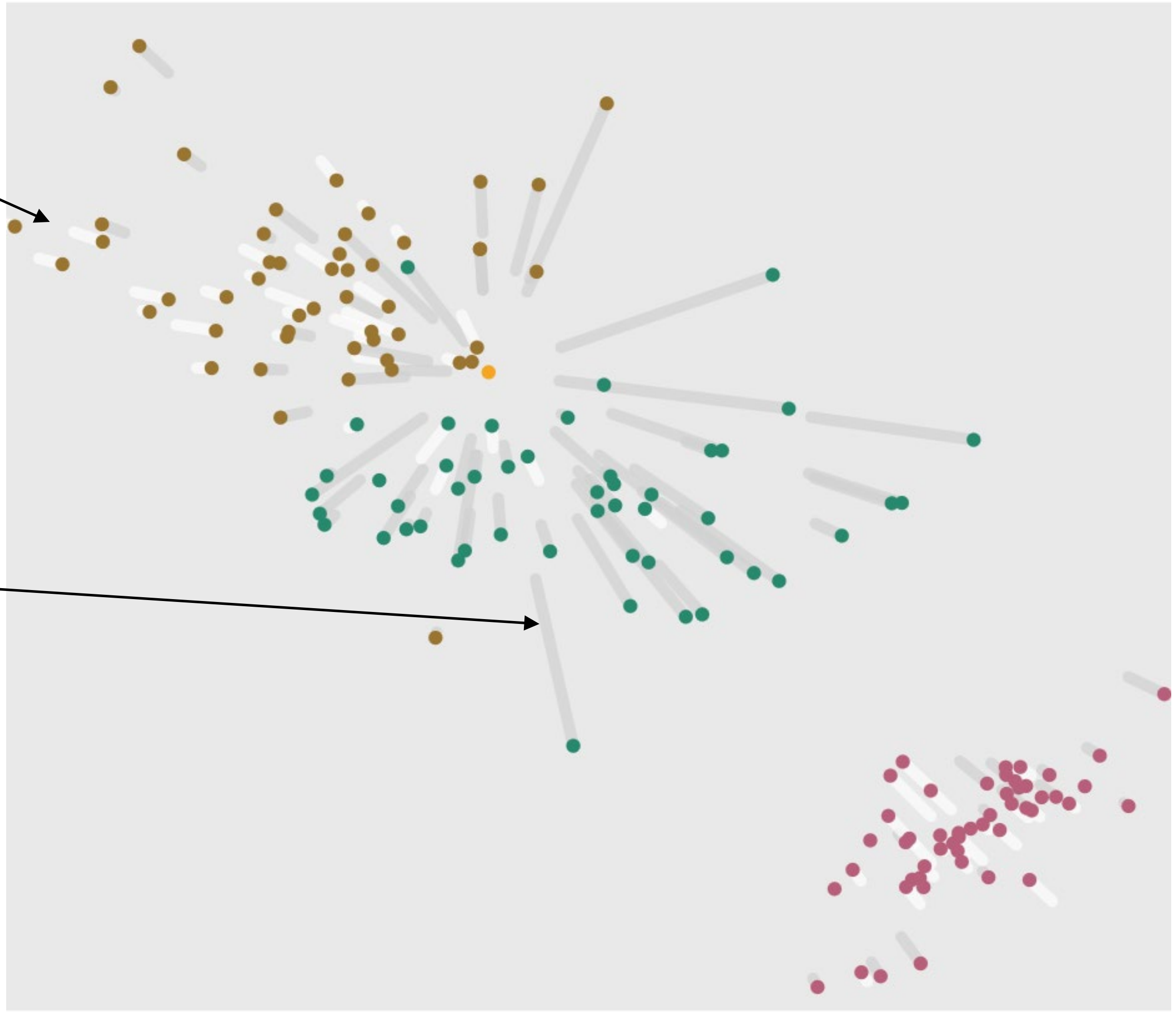
- Inspection techniques: dimension heatmap on projection

[Stahnke et al., 2016]

- Inspection techniques: projection errors

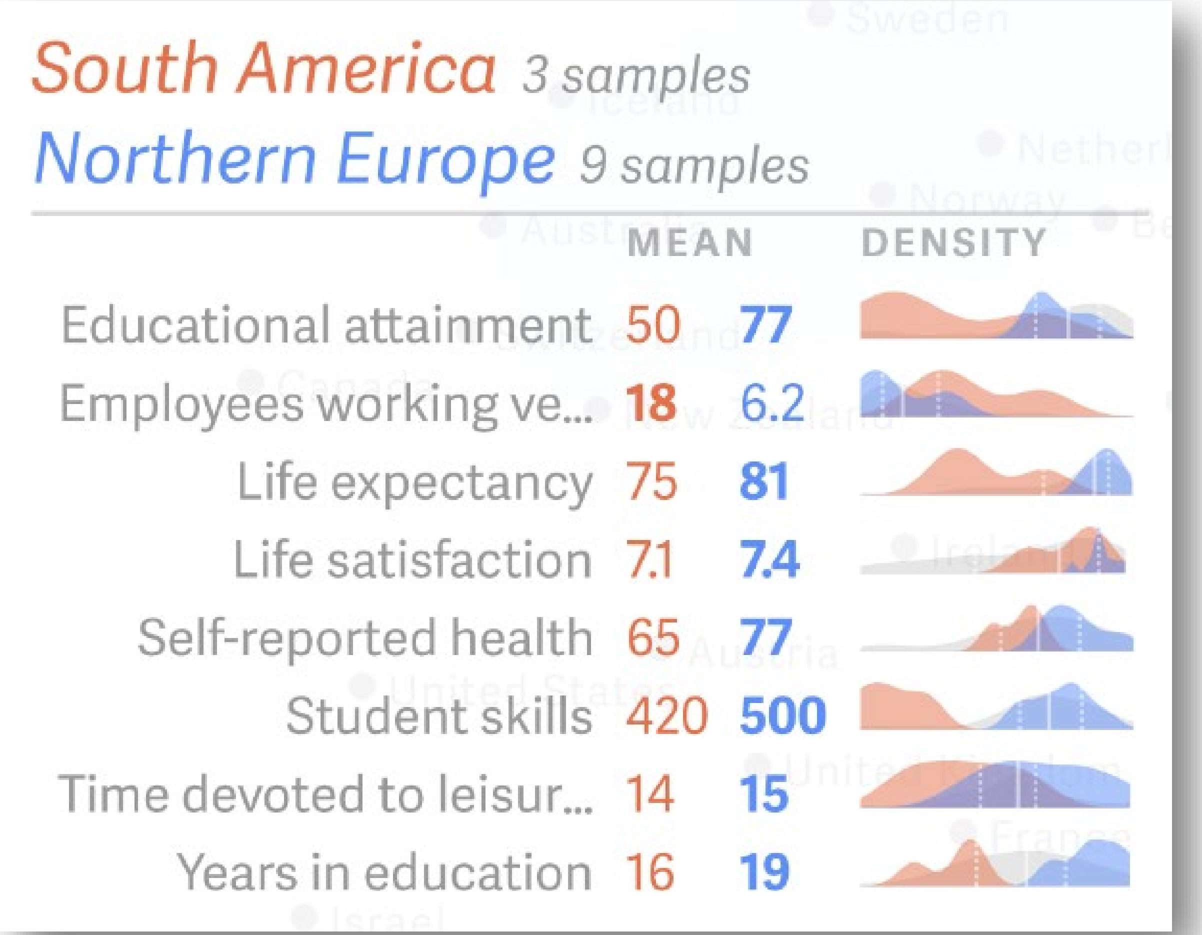
White traces: higher similarity in high-dimensional space

Gray traces: lower similarity in high-dimensional space



[Stahnke et al., 2016]

- Inspection techniques: comparison of group selections



[Stahnke et al., 2016]

t-SNE

- t-Distributed Stochastic Neighbor Embedding
  - Input: matrix of pair-wise similarities
  - Similarities presented as joint probability matrix P:

$$p_{ij} = \frac{\exp(- \| x_i - x_j \|^2 / 2\sigma^2)}{\sum_{k \neq l} \exp(- \| x_k - x_l \|^2 / 2\sigma^2)}.$$

  - Low-dimensional conditional probability matrix Q using Student-t distribution:

$$q_{ij} = \frac{(1 + \| y_i - y_j \|^2)^{-1}}{\sum_{k \neq l} (1 + \| y_k - y_l \|^2)^{-1}}.$$

[van der Maaten and Hinton, 2008]

t-SNE

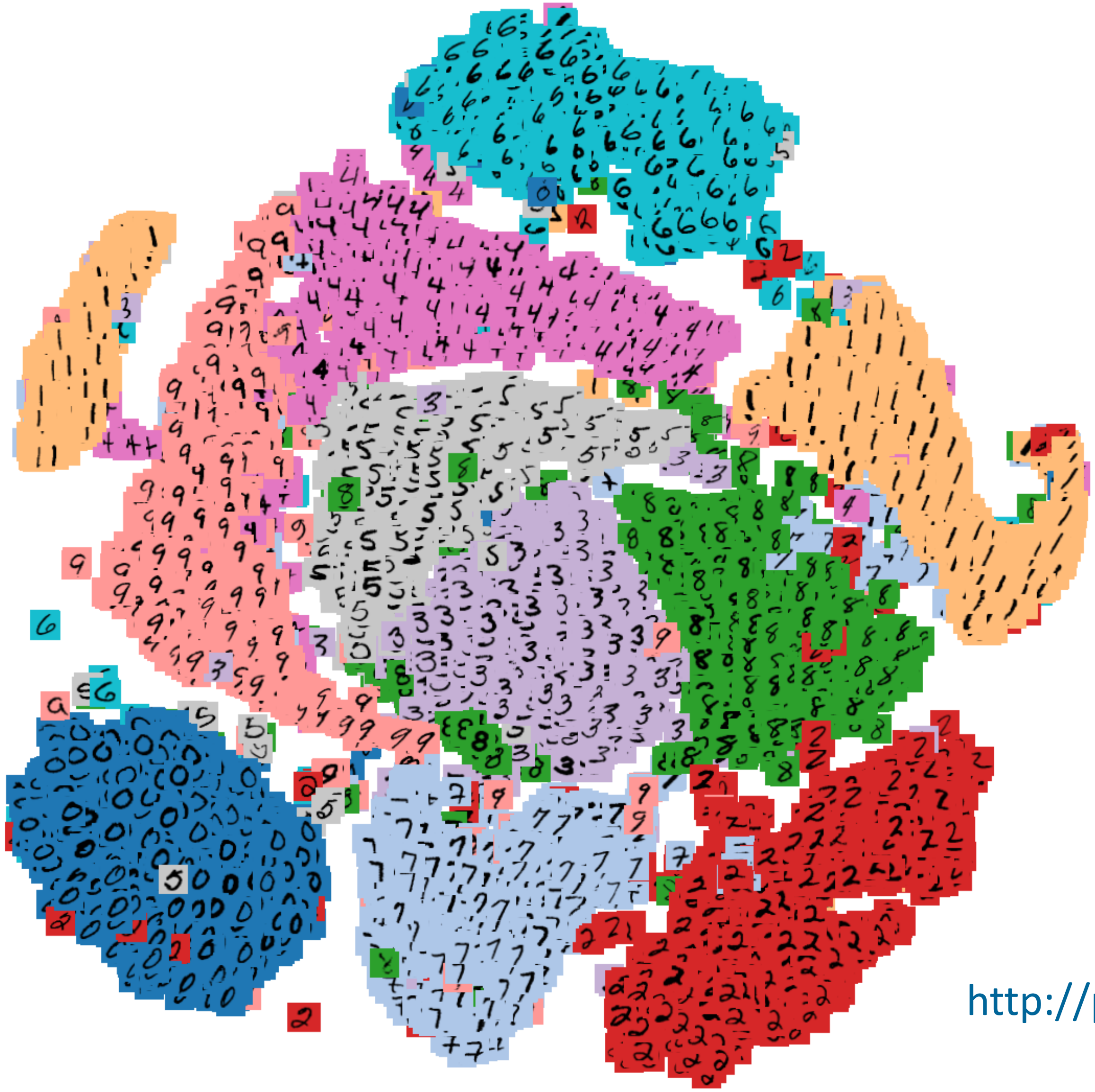
- Goal: find a low-dimensional data representation that minimizes the mismatch between p_{ji} and q_{ji}
- Minimization of sum of Kullback-Leibler divergences over all data points using a gradient descent method:

$$C = KL(P \,\|\, Q) = \sum_i \sum_{j \neq i} p_{ij} \log \frac{p_{ij}}{q_{ij}}.$$

- Can be implemented via Barnes-Hut approximations

[van der Maaten and Hinton, 2008]

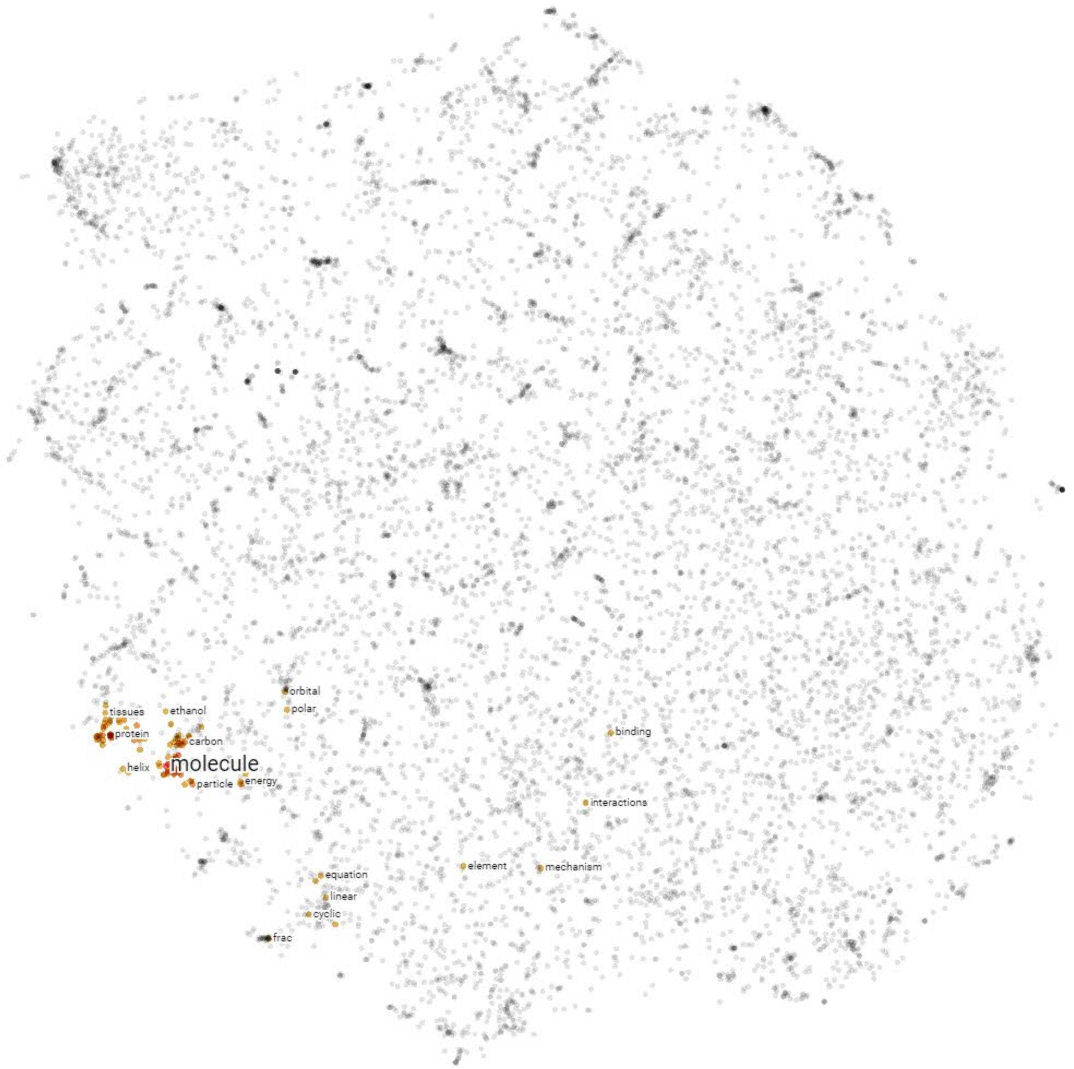
MNIST t-SNE Example



Perplexity: 25
Learning rate: 10
Iterations: 342

http://projector.tensorflow.org/

Tensorflow Embedding Projector: word2vec 10K



molecule

word molecule
count 501

Search: mole — by: word

neighbors 100

distance COSINE EUCLIDEAN

Nearest points in the original space:

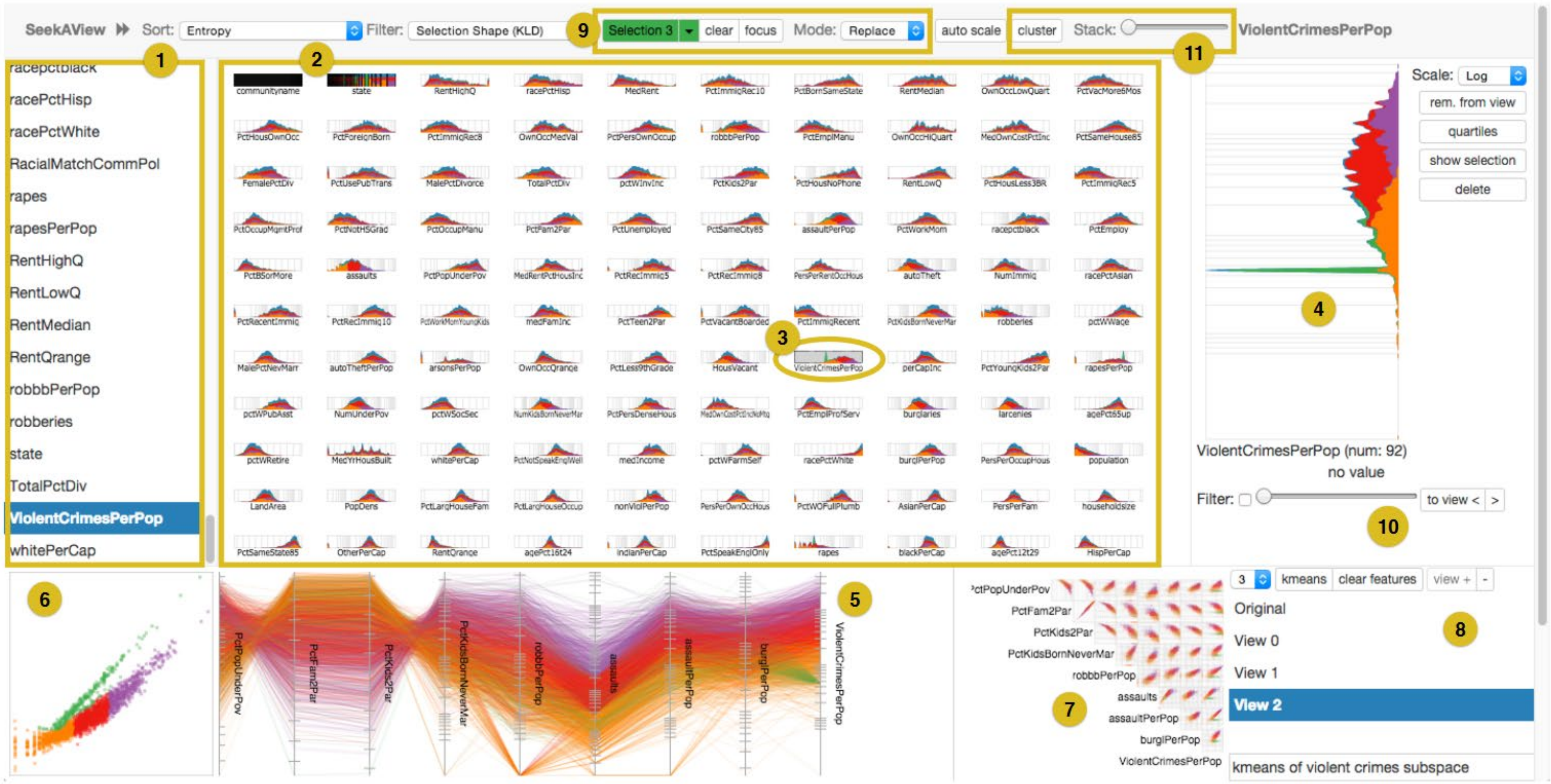
molecules	0.278
atom	0.356
atoms	0.363
hydrogen	0.412
ions	0.441
electrons	0.459
protein	0.459
electron	0.471
protons	0.473
nucleus	0.483
proteins	0.484
dna	0.485
amino	0.490
particle	0.508
ion	0.513
proton	0.515
molecular	0.517
carbon	0.517
acids	0.528
enzymes	0.531
oxygen	0.531
bonding	0.533
orbitals	0.537
membrane	0.544
enzyme	0.546
glucose	0.553
receptor	0.556
particles	0.557
bonds	0.559
nuclei	0.561
dipole	0.561
organism	0.568
cell	0.569
cells	0.570
rna	0.577

Hybrid Approaches

- Dimensionality reduction often unwanted because domain knowledge is required to understand which dimension combinations make sense
- Combination of feature selection and feature extraction
- Feature selection:
  - User selection based on visual analysis
  - Quality metrics
- Feature extraction is performed on selected dimensions
- Using multi-dimensional data visualization techniques

Example: SeekAView

- Example: 1995 US FBI Crime report (147 dimensions, 2000+ items)



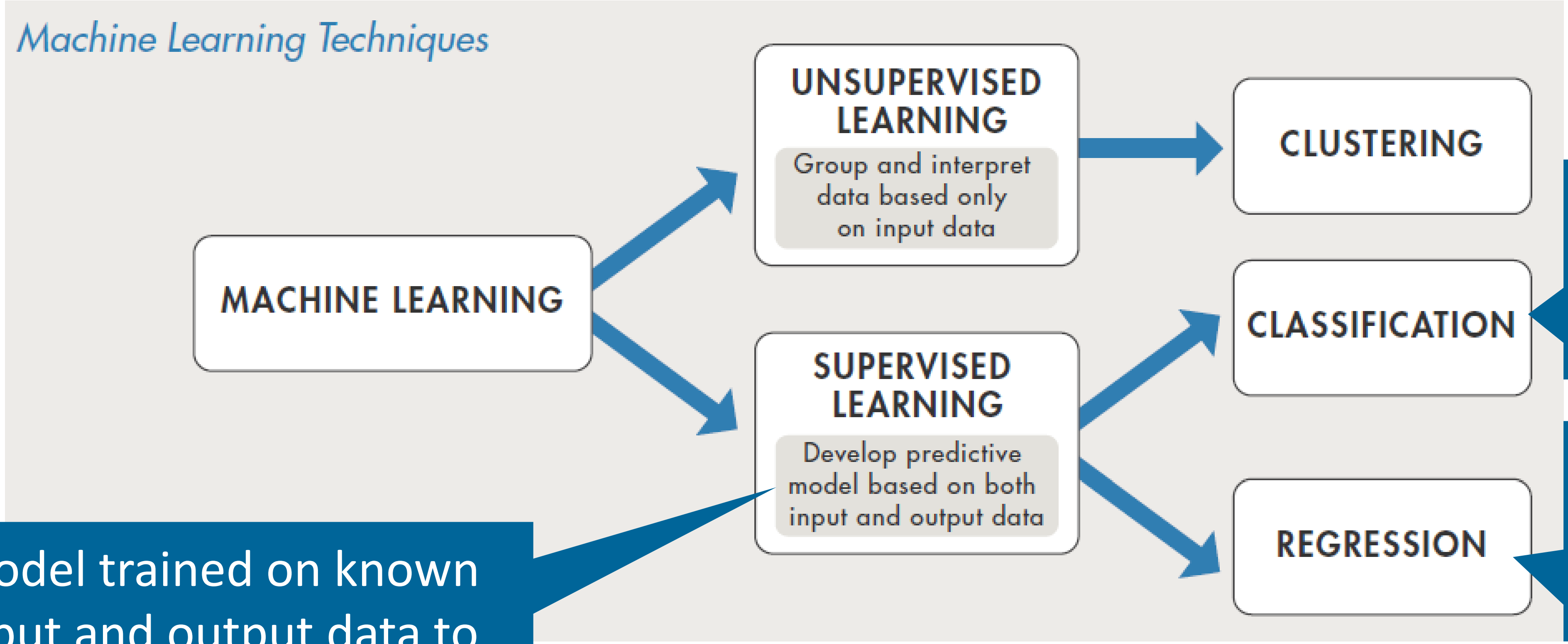
[Krause et al., 2007]

The Data Science Process

PREDICTIVE MODELING

Train machine learning models, **evaluate** their performance, and **use** them to make predictions.

Predictive Models



Prediction of discrete responses

Prediction of continuous responses

Model trained on known input and output data to predict future outputs

MathWorks

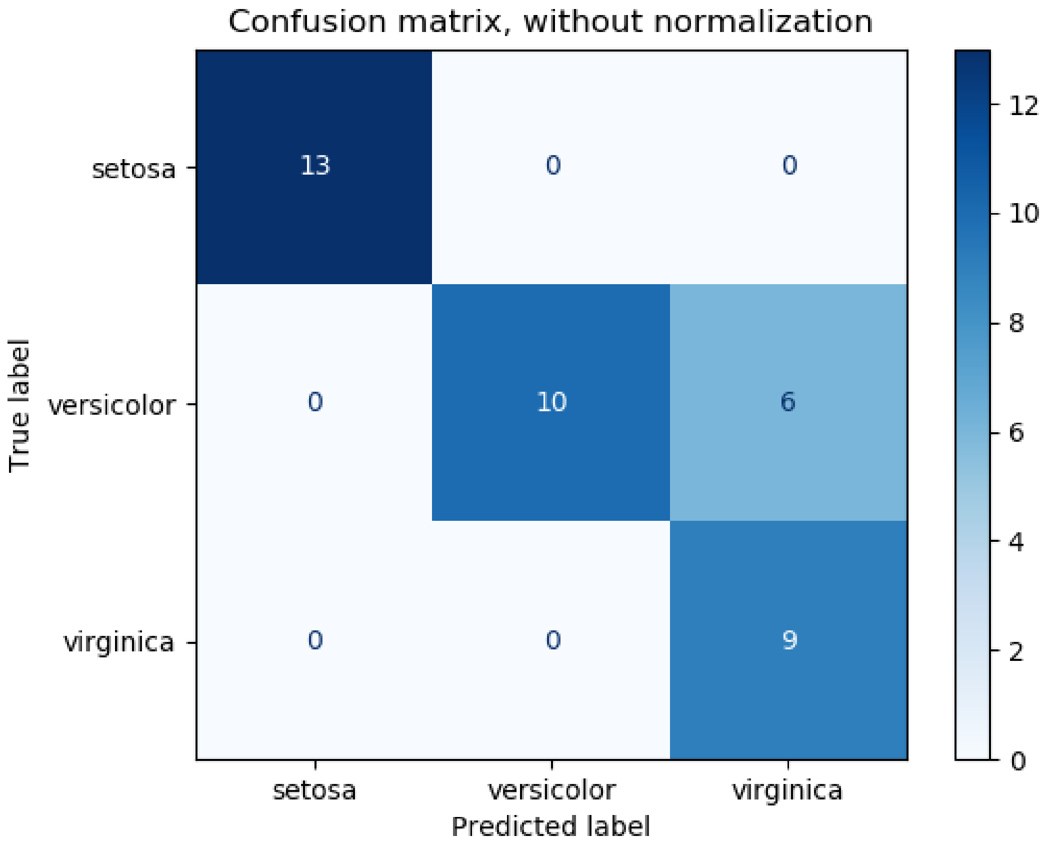
Predictive Visual Analytics

- Why do we need visualization?
  - Evaluate: Validation and comparison
  - Train: Model improvement and training
  - Make predictions
    - AI interpretability and explainability

Predictive Visual Analytics

- Why do we need visualization?
  - **Evaluate: Validation and comparison**
  - Train: Model improvement and training
  - Make predictions
    - AI interpretability and explainability

- Confusion Matrix

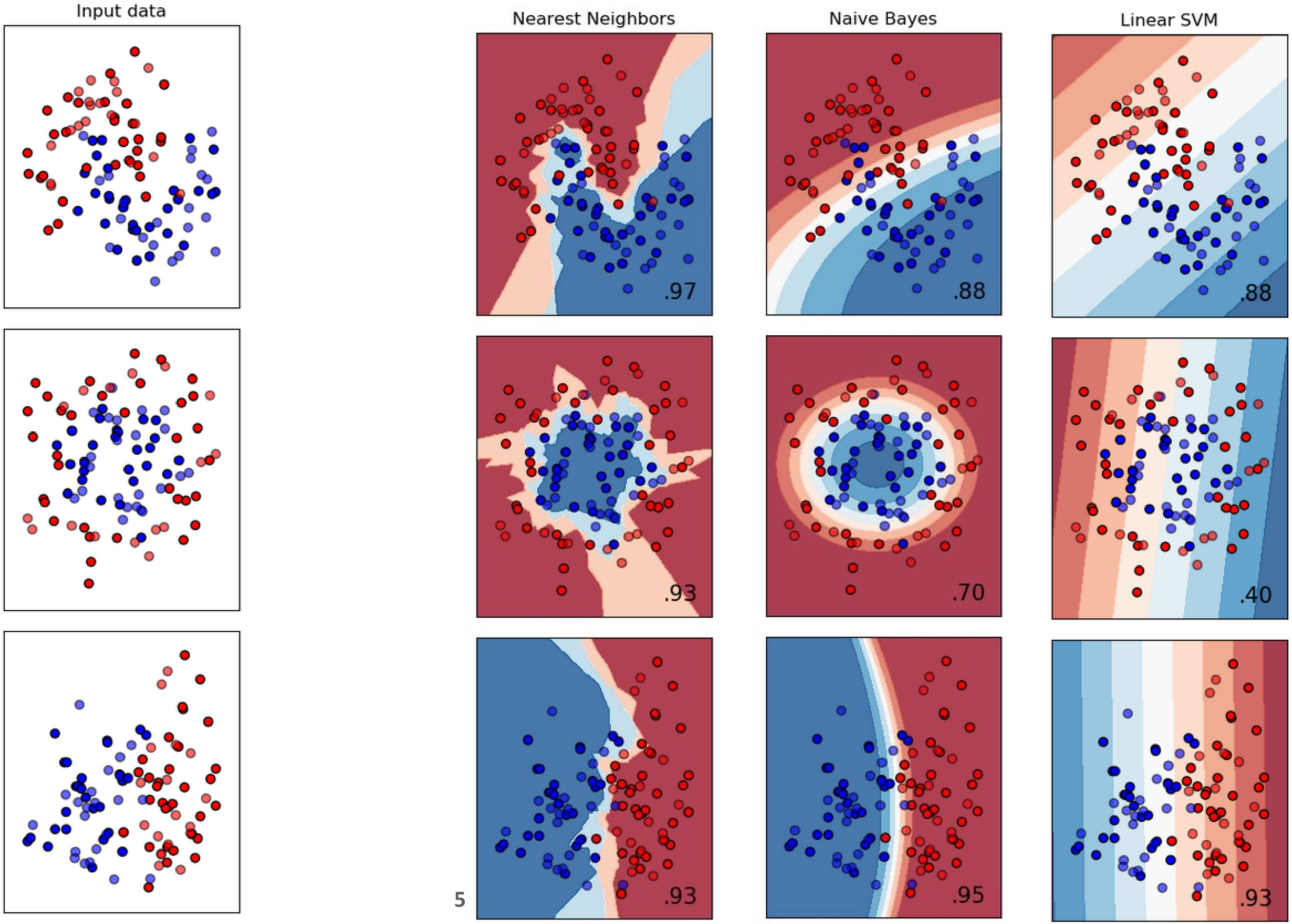


https://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html

Evaluation of Classifier Accuracy: Scatterplots

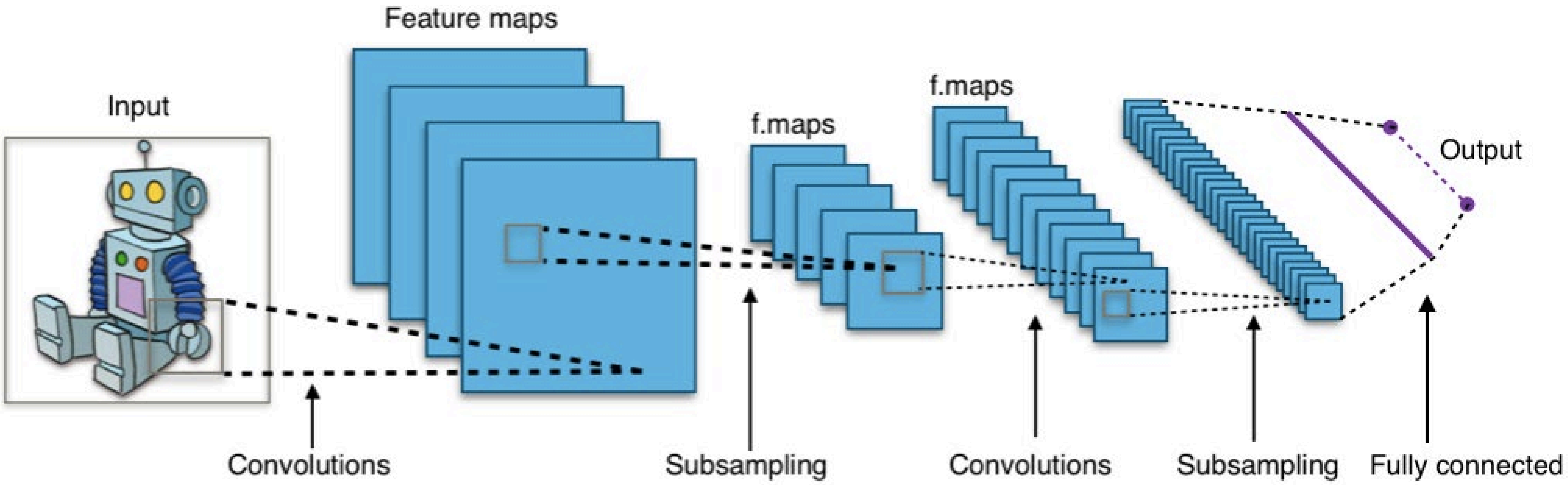
Training point

Testing point

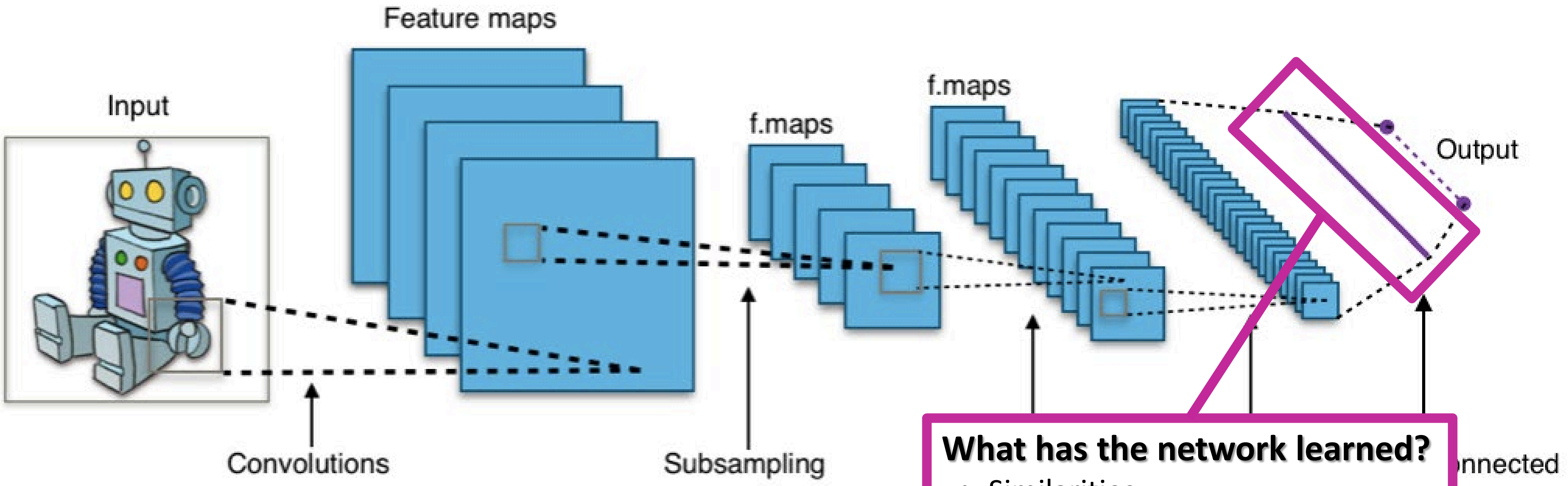


https://scikit-learn.org/stable/auto_examples/classification/plot_classifier_comparison.html

Convolutional Neural Networks

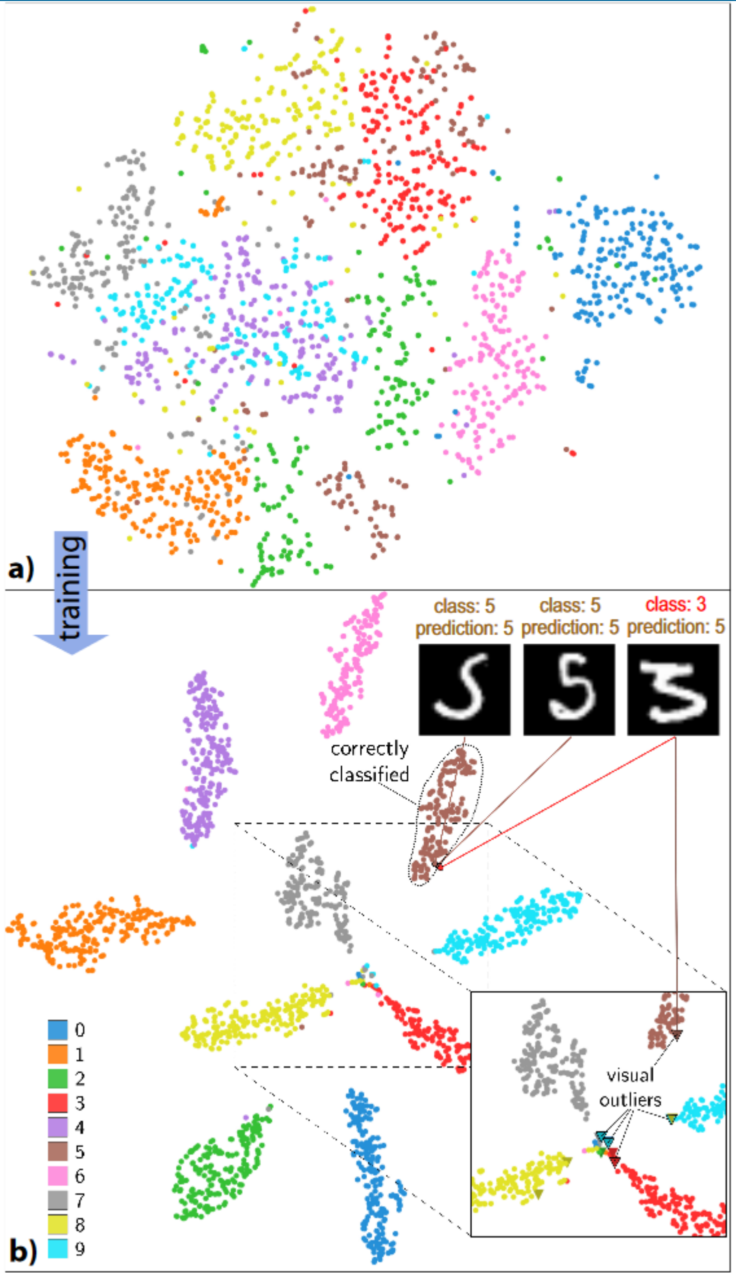


Convolutional Neural Networks

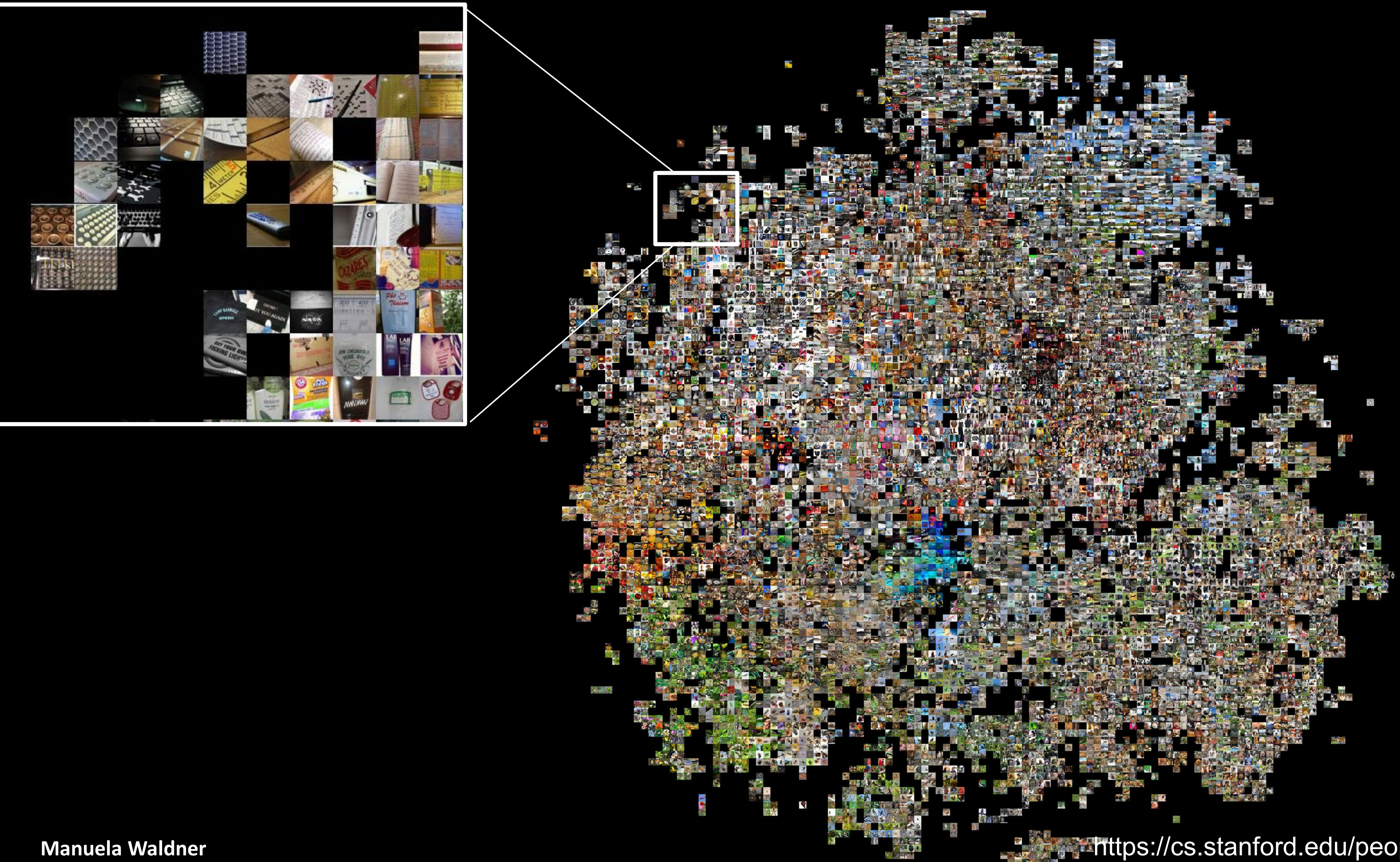


What has the network learned?
- Similarities
- Outliers
- Ambiguities

Inspecting Training Effects [Rauber et al. 2016]



CNN Code Projection



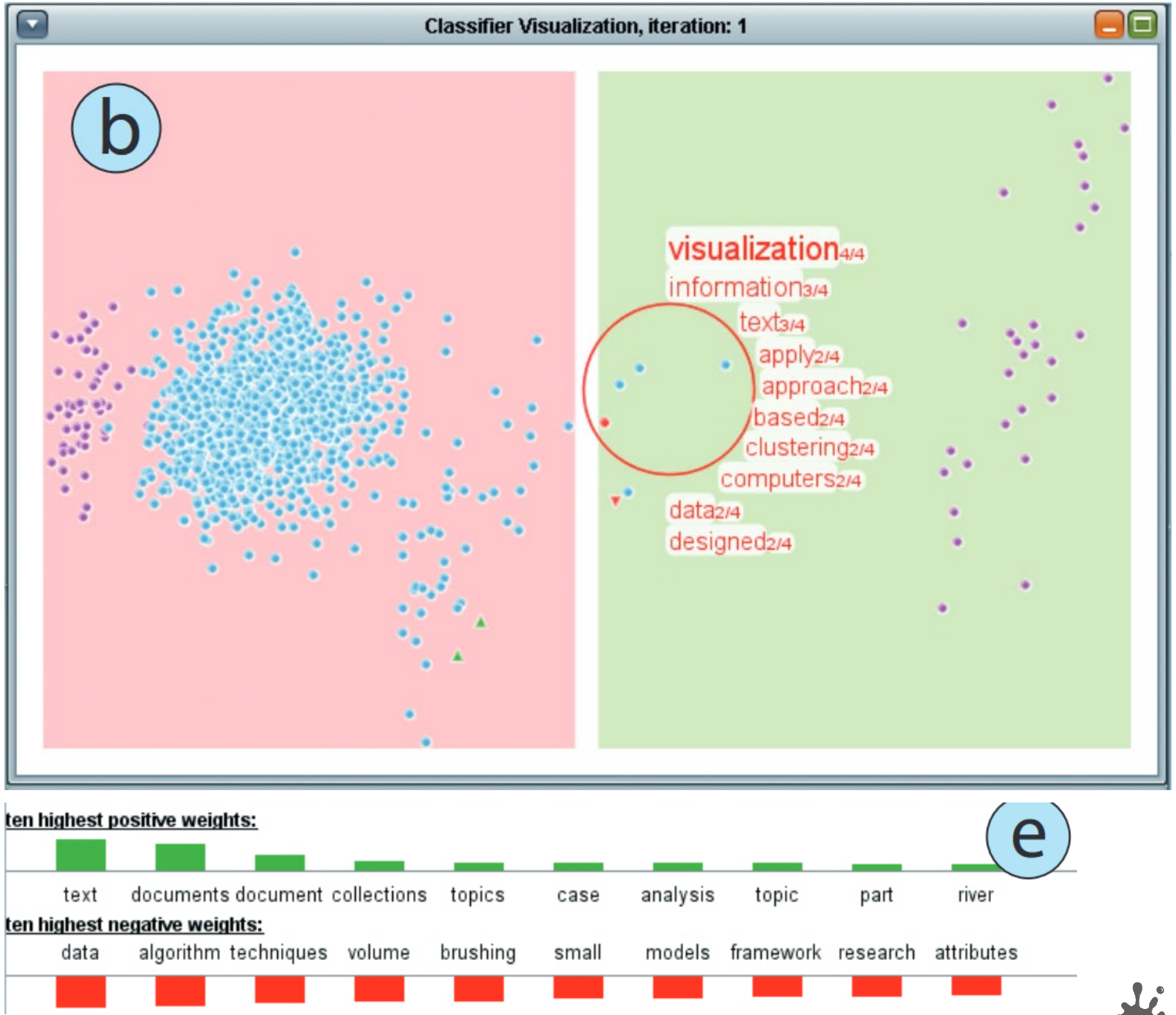
https://cs.stanford.edu/people/karpathy/cnnembed/

Predictive Visual Analytics

- Why do we need visualization?
  - Evaluate: Validation and comparison
  - **Train: Model improvement and training**
  - Make predictions
    - AI interpretability and explainability

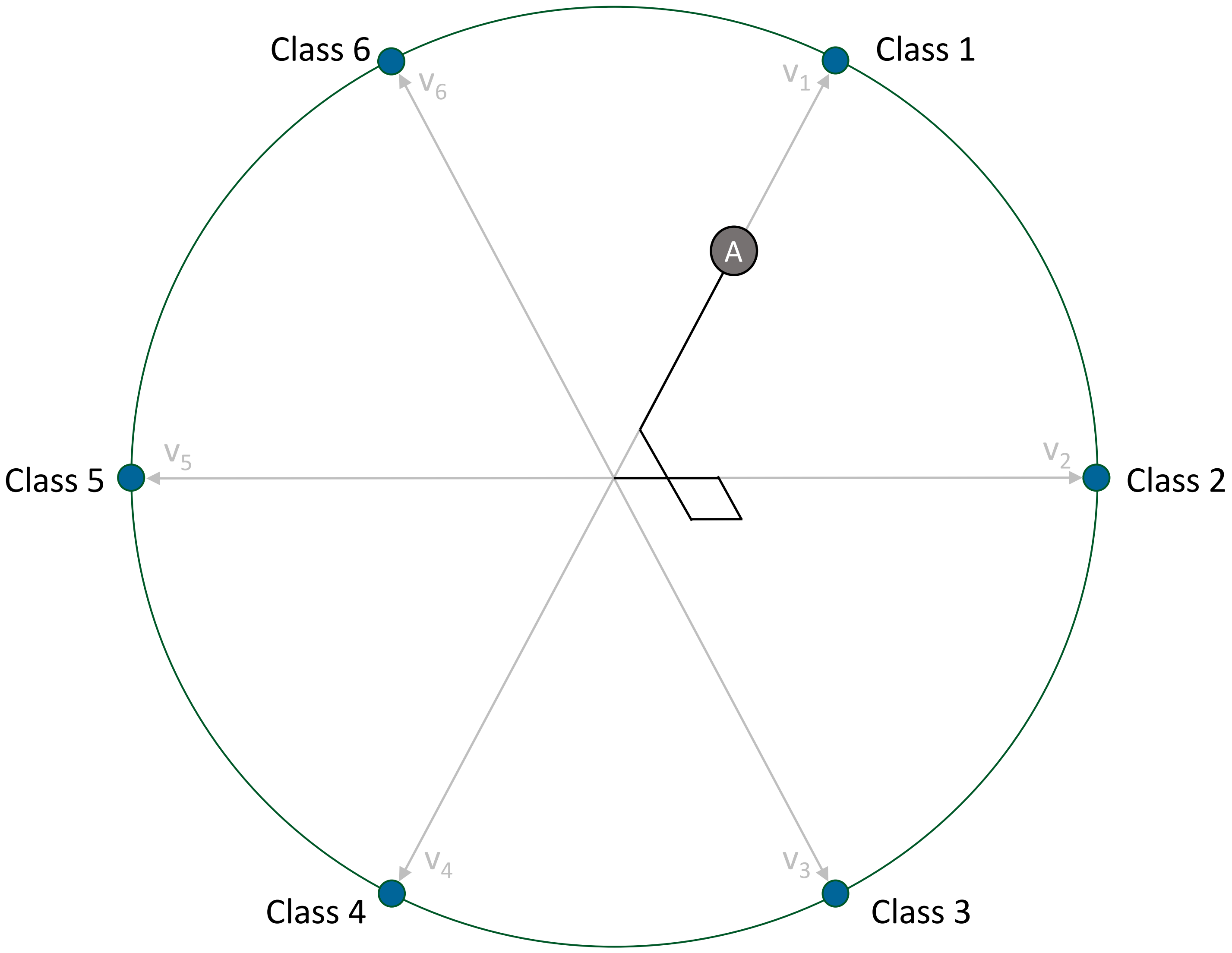
Train SVM

- Document classification for a given query
  - Relevant
  - Irrelevant
- Samples = documents
  - Labeled
  - Unlabeled
- Visualizes SVM decision boundary



[Heimerl et al., TVCG 2012]

Train Naïve Bayes



Class Associations

	1	2	3	4	5	6
A	0.4	0.2	0.1	0	0.1	0.2

Train Naïve Bayes

Train Naïve Bayes

Predictive Visual Analytics

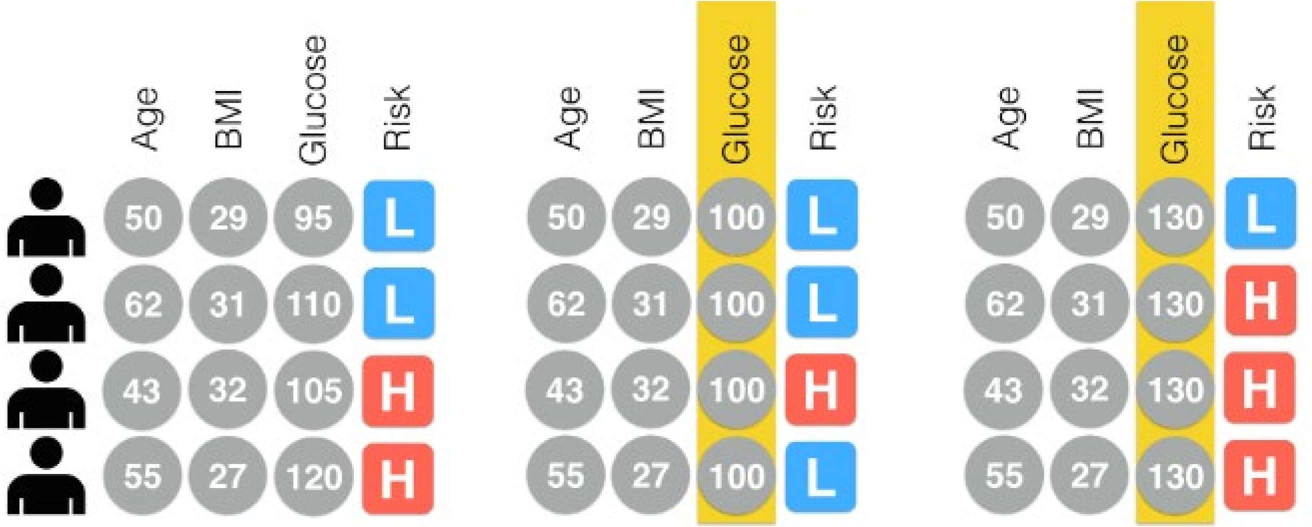
- Why do we need visualization?
  - Evaluate: Validation and comparison
  - Train: Model improvement and training
  - Make predictions
    - **AI interpretability and explainability**

The extent to which a cause and effect can be observed within a system

The extent to which the **internal** mechanics of a machine learning system can be explained in human terms

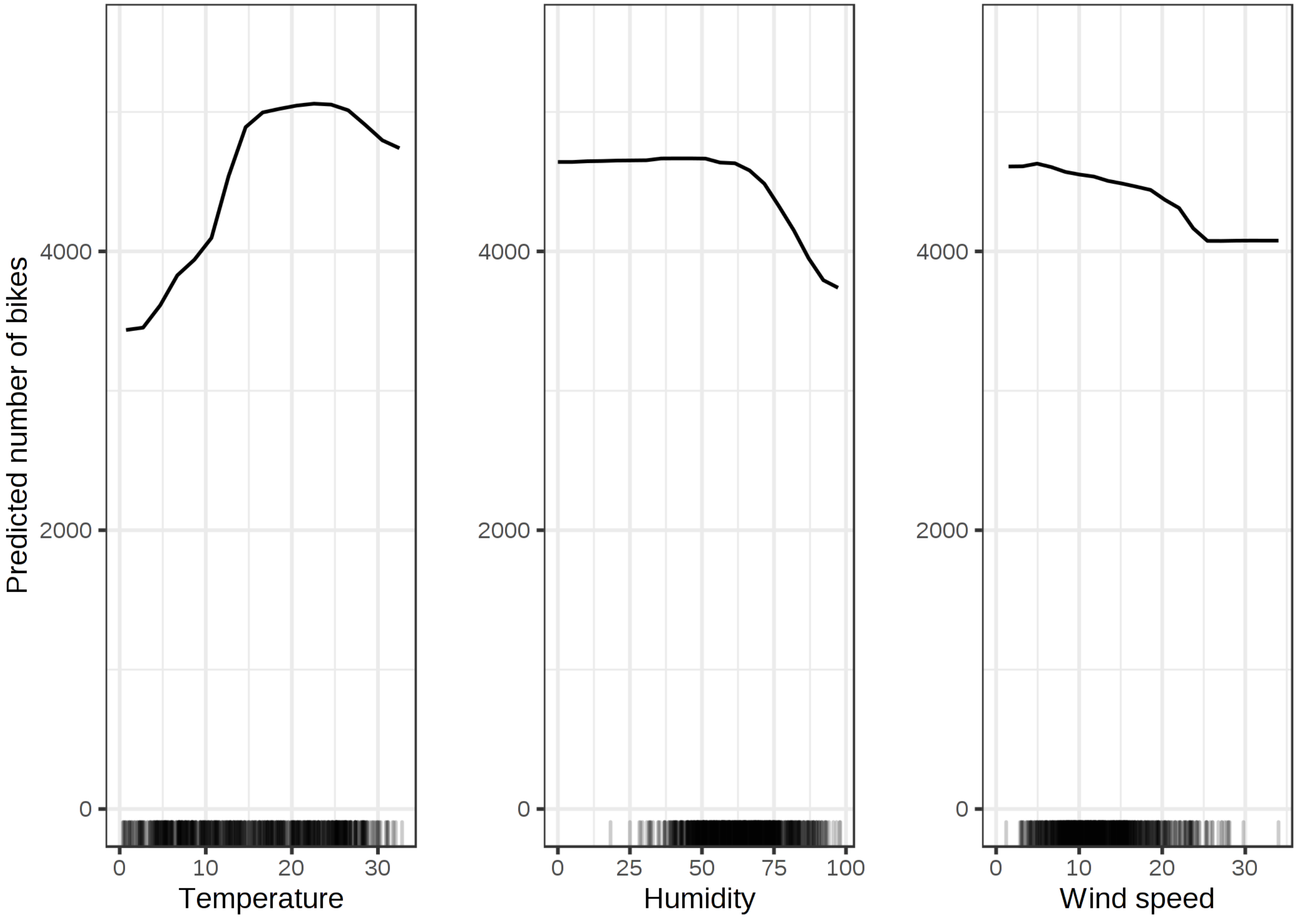
Interpretability

- Partial dependence plot
  - Assessing influence of a feature on the prediction
  - Shows marginal effect a feature has on predicted outcome
  - Based on averages in training data:



[Krause et al. 2016]

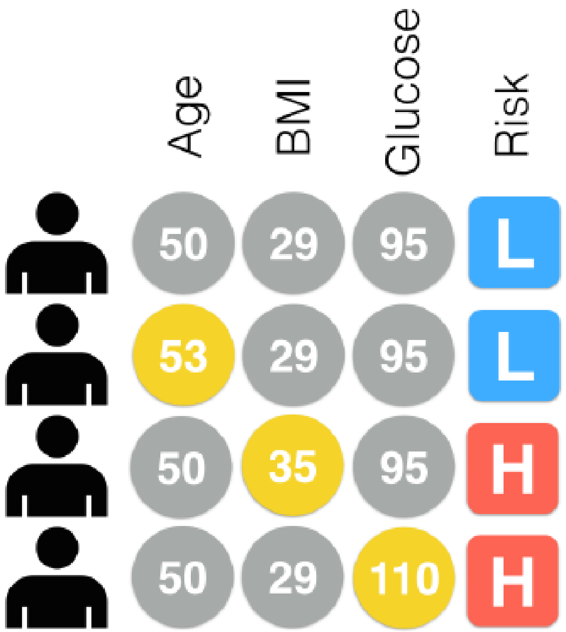
Partial Dependence Plot



https://christophm.github.io/interpretable-ml-book/pdp.html

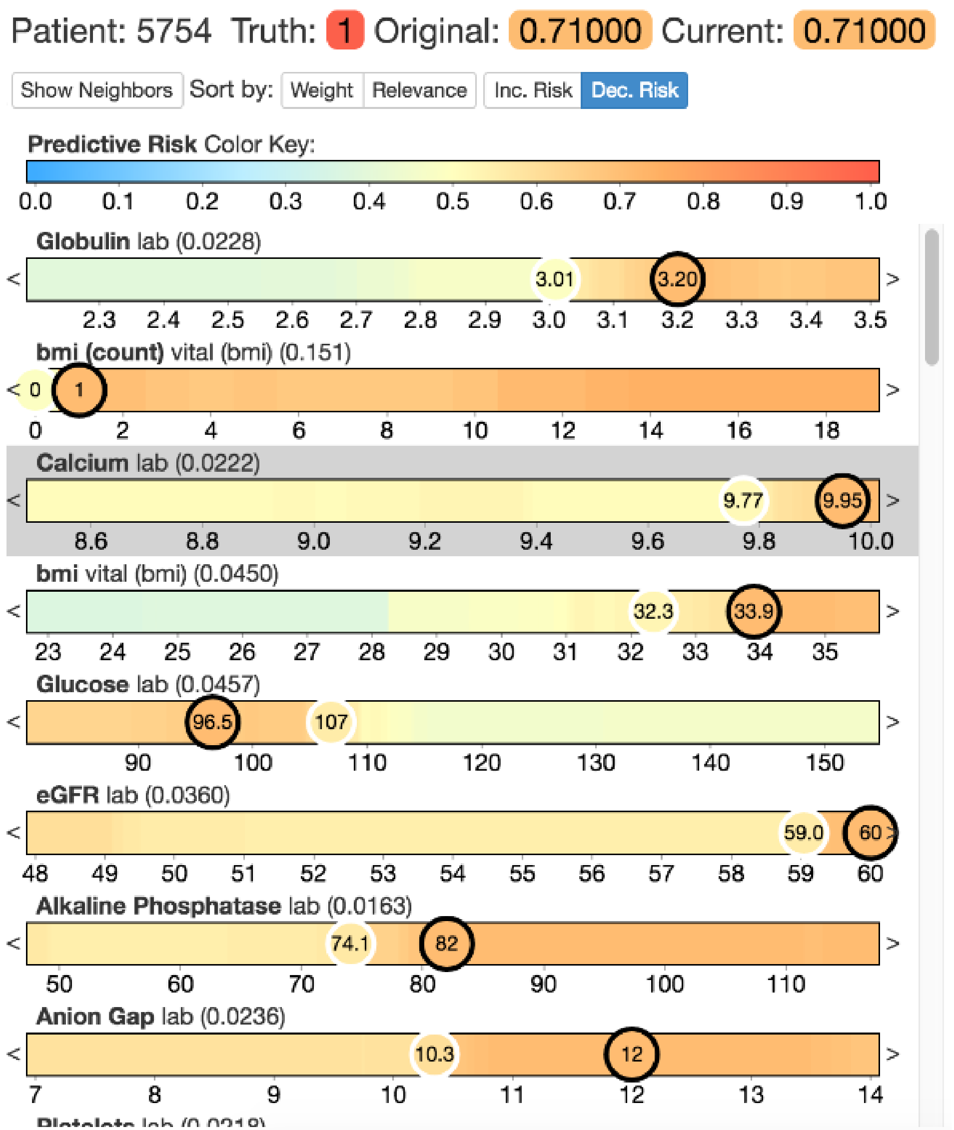
Interactive Partial Dependence

- Interactively testing scenarios:



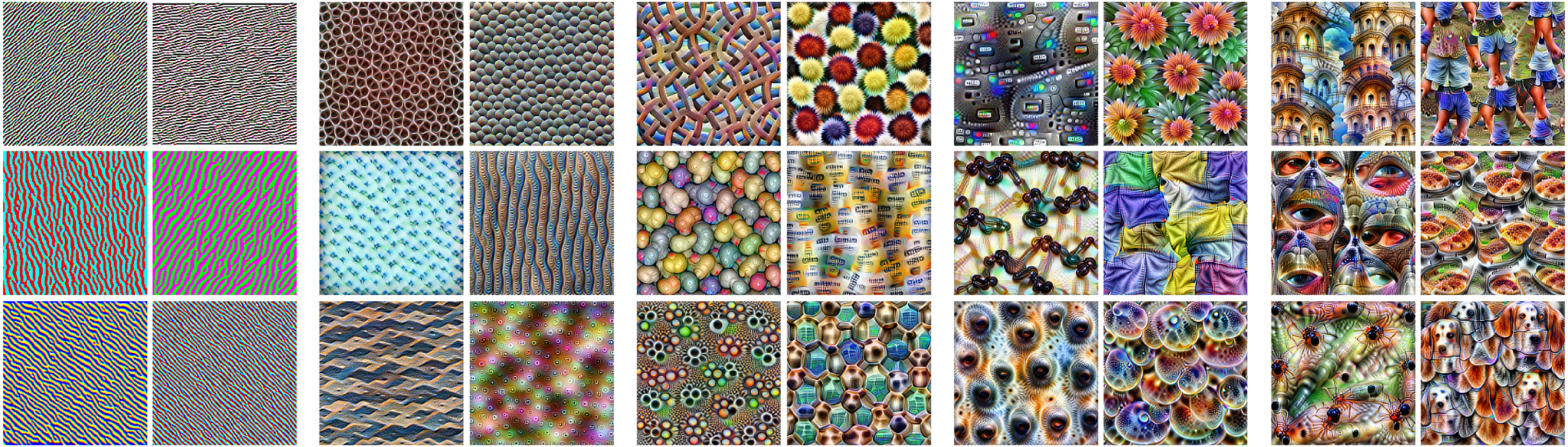
[Krause et al. 2016]

Interactive Partial Dependence



[Krause et al. 2016]

Feature visualization:



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)

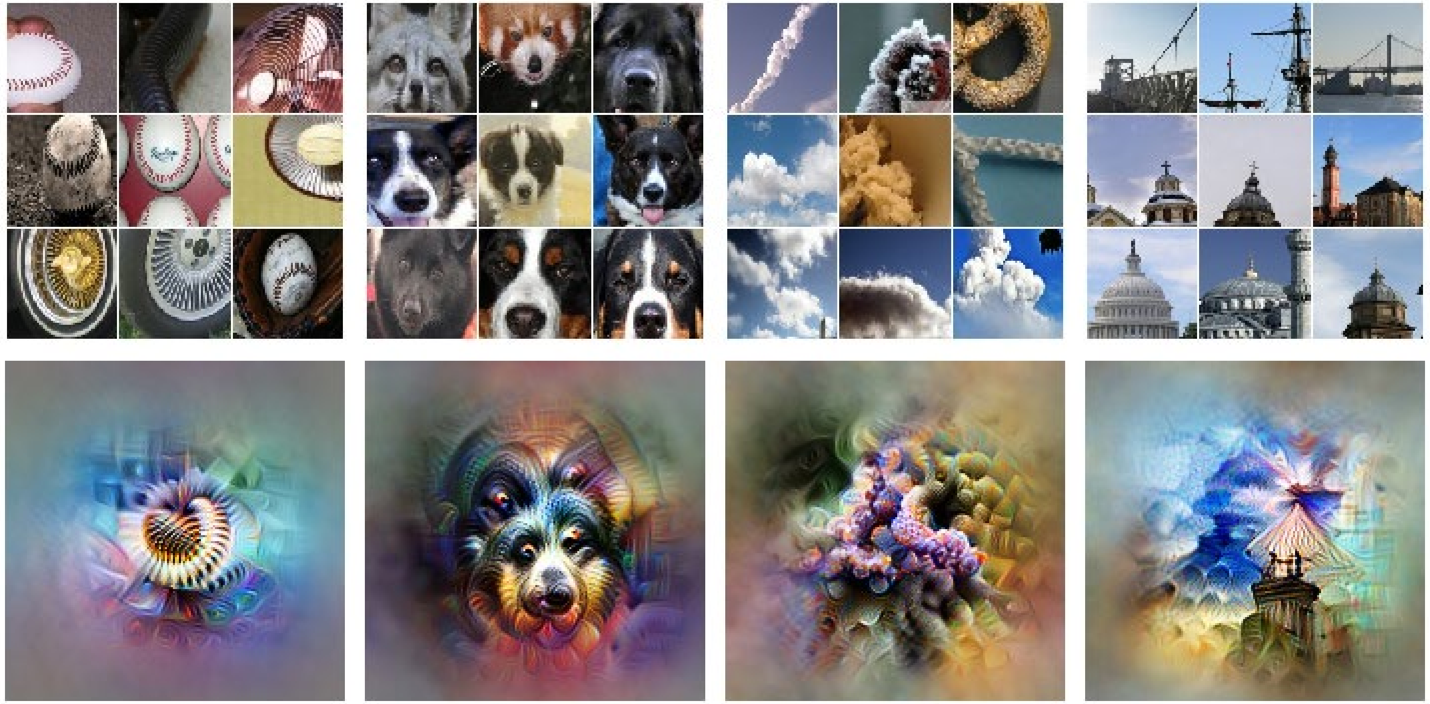
Feature visualization allows us to see how GoogLeNet [1], trained on the ImageNet [2] dataset, builds up its understanding of images over many layers. Visualizations of all channels are available in the appendix.

https://distill.pub/2017/feature-visualization/

Feature visualization:

Dataset Examples show us what neurons respond to in practice

Optimization isolates the causes of behavior from mere correlations. A neuron may not be detecting what you initially thought.



Baseball—or stripes? *mixed4a, Unit 6*

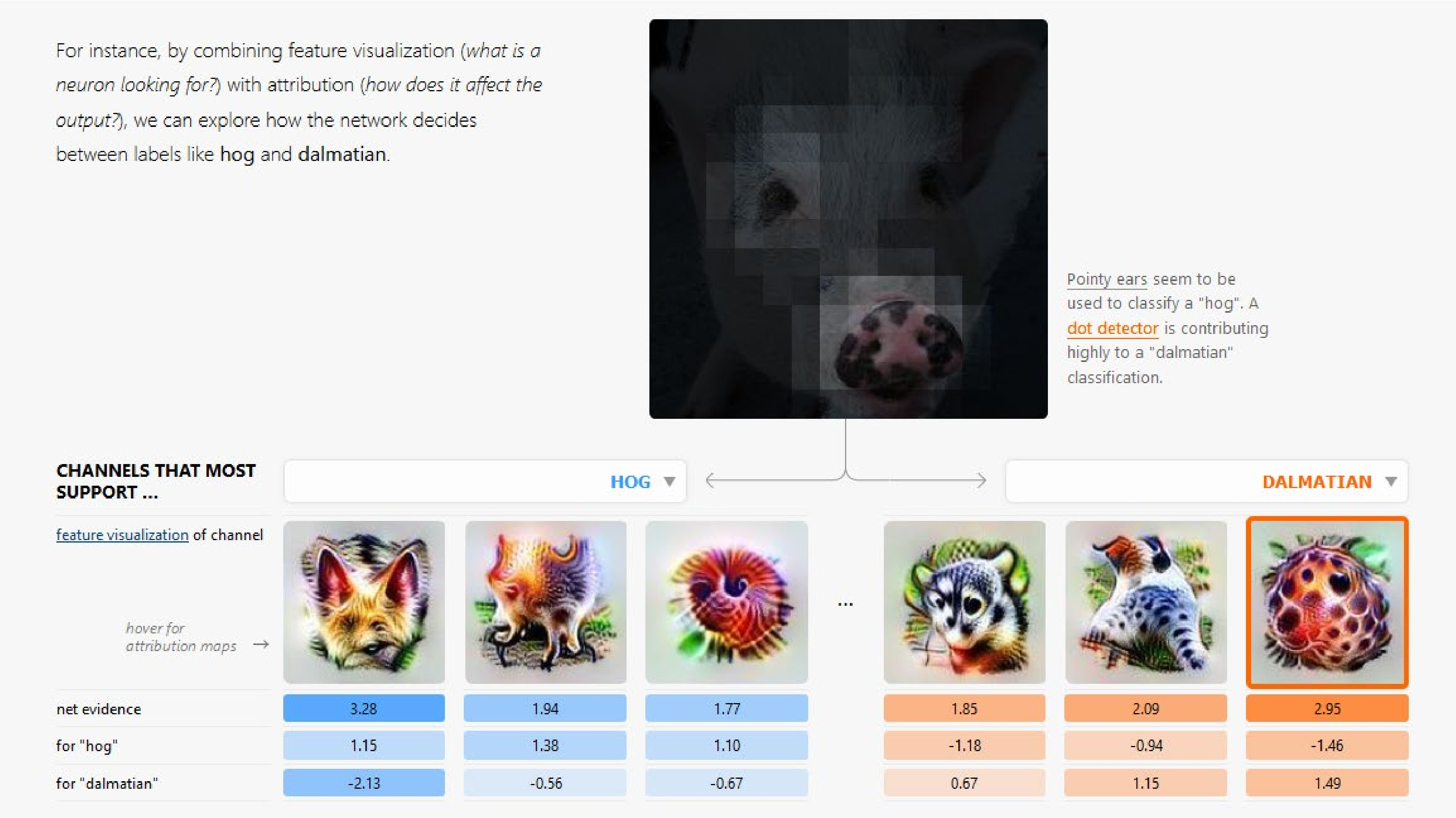
Animal faces—or snouts? *mixed4a, Unit 240*

Clouds—or fluffiness? *mixed4a, Unit 453*

Buildings—or sky? *mixed4a, Unit 492*

https://distill.pub/2018/building-blocks/

Feature visualizations and attribution maps:



For instance, by combining feature visualization (*what is a neuron looking for?*) with attribution (*how does it affect the output?*), we can explore how the network decides between labels like **hog** and **dalmatian**.

Pointy ears seem to be used to classify a "hog". A dot detector is contributing highly to a "dalmatian" classification.

CHANNELS THAT MOST SUPPORT ...	HOG				DALMATIAN		
feature visualization of channel *hover for attribution maps* →				...			
net evidence	3.28	1.94	1.77		1.85	2.09	2.95
for "hog"	1.15	1.38	1.10		-1.18	-0.94	-1.46
for "dalmatian"	-2.13	-0.56	-0.67		0.67	1.15	1.49

https://distill.pub/2018/building-blocks/

Visual Data Science - Summary

- Data exploration / scalable visualization
  - Perceptual scalability: model-based / aggregate visualization
  - Interactive scalability: online aggregation, aggregate queries, data tiles
- Feature engineering / high-dimensional data visualization
  - Feature selection
  - Feature extraction (dimensionality reduction)
  - Hybrid approach
- Predictive visual analytics
  - Supervised machine learning (regression, classification)
  - Evaluation, training, interpretability & explainability

References

- Liu et al., **Visualizing High-Dimensional Data: Advances in the Past Decade**, EuroVis 2015 State of the Art Report
- Brehmer, M., Sedlmair, M., Ingram, S., & Munzner, T. (2014, November). **Visualizing dimensionally-reduced data: Interviews with analysts and a characterization of task sequences**. In *Proceedings of the Fifth Workshop on Beyond Time and Errors: Novel Evaluation Methods for Visualization* (pp. 1-8). ACM.
- Bertini, E., Tatu, A., & Keim, D. (2011). **Quality metrics in high-dimensional data visualization: An overview and systematization.** *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2203-2212.
- Seo, J., & Shneiderman, B. (2005). **A rank-by-feature framework for interactive exploration of multidimensional data**. *Information visualization*, *4*(2), 96-113.
- Sips, M., Neubert, B., Lewis, J. P., & Hanrahan, P. (2009, June). **Selecting good views of high-dimensional data using class consistency**. In *Computer Graphics Forum* (Vol. 28, No. 3, pp. 831-838). Blackwell Publishing Ltd.
- Johansson, S., & Johansson, J. (2009). **Interactive dimensionality reduction through user-defined combinations of quality metrics.** *IEEE transactions on visualization and computer graphics*, *15*(6), 993-1000.

References

- Tatu, A., Albuquerque, G., Eisemann, M., Schneidewind, J., Theisel, H., Magnork, M., & Keim, D. (2009, October). **Combining automated analysis and visualization techniques for effective exploration of high-dimensional data**. In *Visual Analytics Science and Technology, 2009. VAST 2009. IEEE Symposium on* (pp. 59-66). IEEE
- Molchanov, V., & Linsen, L. (2014). **Interactive Design of Multidimensional Data Projection Layout**.
- Jeong, D. H., Ziemkiewicz, C., Fisher, B., Ribarsky, W., & Chang, R. (2009, June). **iPCA: An Interactive System for PCA-based Visual Analytics**. In *Computer Graphics Forum* (Vol. 28, No. 3, pp. 767-774). Blackwell Publishing Ltd.
- Wise, J. A., Thomas, J. J., Pennock, K., Lantrip, D., Pottier, M., Schur, A., & Crow, V. (1995, October). **Visualizing the non-visual: Spatial analysis and interaction with information from text documents**. In *Information Visualization, 1995. Proceedings.* (pp. 51-58). IEEE.
- Fried, D., & Kobourov, S. G. (2014, March). **Maps of computer science**. In *Visualization Symposium (PacificVis), 2014 IEEE Pacific* (pp. 113-120). IEEE.
- Stahnke, J., Dörk, M., Müller, B., & Thom, A. (2016). **Probing projections: Interaction techniques for interpreting arrangements and errors of dimensionality reductions**. *IEEE transactions on visualization and computer graphics*, *22*(1), 629-638.

References

- Maaten, L. V. D., & Hinton, G. (2008). **Visualizing data using t-SNE**. *Journal of machine learning research*, *9*(Nov), 2579-2605.
- Rieck, B., & Leitte, H. (2015, June). **Persistent homology for the evaluation of dimensionality reduction schemes.** In *Computer Graphics Forum* (Vol. 34, No. 3, pp. 431-440).
- Endert, A., Fiaux, P., & North, C. (2012, May). **Semantic interaction for visual text analytics.** In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 473-482). ACM.
- Turkay, C., Filzmoser, P., & Hauser, H. (2011). **Brushing dimensions-a dual visual analysis model for high-dimensional da**ta. *IEEE transactions on visualization and computer graphics*, *17*(12), 2591-2599.
- Yang, J., Ward, M. O., & Rundensteiner, E. A. (2002). **Visual hierarchical dimension reduction for exploration of high dimensional datasets**.
- Krause, J., Dasgupta, A., Fekete, J. D., & Bertini, E. (2016, October). **SeekAView: An intelligent dimensionality reduction strategy for navigating high-dimensional data spaces.** In *Large Data Analysis and Visualization (LDAV), 2016 IEEE 6th Symposium on* (pp. 11-19). IEEE.
- Wilkinson, L., Anand, A., & Grossman, R. (2005). **Graph-theoretic scagnostics**.

References

- Jiang, L., Liu, S., & Chen, C. (2018). **Recent research advances on interactive machine learning.** *Journal of Visualization*, 1-17.
- Sacha, D., Zhang, L., Sedlmair, M., Lee, J. A., Peltonen, J., Weiskopf, D., ... & Keim, D. A. (2017). **Visual interaction with dimensionality reduction: A structured literature analysis**. *IEEE transactions on visualization and computer graphics*, *23*(1), 241-250.
- Endert, A., Ribarsky, W., Turkay, C., Wong, B. W., Nabney, I., Blanco, I. D., & Rossi, F. (2017, December). **The state of the art in integrating machine learning into visual analytics**. In *Computer Graphics Forum* (Vol. 36, No. 8, pp. 458-486).
- Fails, J. A., & Olsen Jr, D. R. (2003, January). **Interactive machine learning**. In *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 39-45). ACM.
- Kandogan, E. (2001, August). **Visualizing multi-dimensional clusters, trends, and outliers using star coordinates**. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 107-116). ACM.
- Zhao, Y., Luo, F., Chen, M., Wang, Y., Xia, J., Zhou, F., ... & Chen, W. (2019). **Evaluating multi-dimensional visualizations for understanding fuzzy clusters.** *IEEE transactions on visualization and computer graphics*, *25*(1), 12-21.

References

- Liu, S., Wang, X., Liu, M., & Zhu, J. (2017). **Towards better analysis of machine learning models: A visual analytics perspective**. *Visual Informatics*, *1*(1), 48-56.
- Kriegel, H. P., Kröger, P., & Zimek, A. (2009). **Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering**. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *3*(1), 1.
- Hu, J., & Pei, J. (2018). **Subspace multi-clustering: a review**. *Knowledge and Information Systems*, *56*(2), 257-284.
- Nam, J. E., & Mueller, K. (2013). **TripAdvisor^{ND}: A Tourism-Inspired High-Dimensional Space Exploration Framework with Overview and Detail.** *IEEE transactions on visualization and computer graphics*, *19*(2), 291-305.
- Friedman, J. H., & Stuetzle, W. (2002). **John W. Tukey's work on interactive graphics**. *The Annals of Statistics*, *30*(6), 1629-1639.