2019 ZJU
International Summer School on Visual Analytics

High-Dimensional Data Visualization

Nan Cao

Professor

Intelligent Big Data Visualization Lab

Tongji College of Design and Innovation

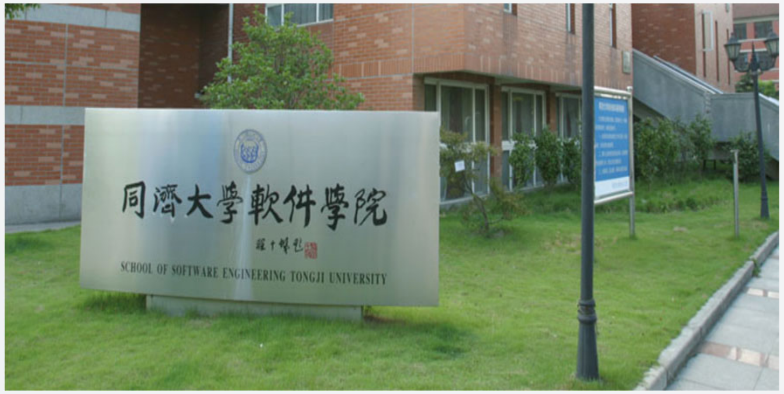
Email: nan.cao@tongji.edu.cn

Nan Cao (曹楠)

http://nancao.org

Tongji University
IBM T.J Watson Research Center
Hong Kong University of Science and Technology

智能大数据可视化实验室



自2016年以来，三年内累计发表国际论文43篇，包括22篇SCI期刊论文，以及21篇国际会议论文

获得ACM IUI 国际会议 最佳论文奖 1 项

在 ACM CHI，IEEE Pacific Vis，ChinaVis 累计获得 3 项最佳论文提名奖

iDVx 实验室

同济大学智能大数据可视化实验室成立于 2016年9月,是同济大学中一个横跨设计创意学院及软件学院的以信息及数据科学为研究方向的创新型试验室,旨在打造数据科学领域中具有世界一流水准的智能化数据分析、可视化、设计、以及人机交互技术,并开展相关技术在 智慧医疗以及智能设计等相关领域的广泛应用。实验室由中组部 "青年千人" 带头,在学术界先后建立了与美国北卡罗来纳州立大学信息学院、美国匹兹堡大学信息学院、亚利桑那州立大学信息系统学院的长期合作及交流访问计划。在工业界,实验室先后与IBM,微软、Adobe、西门子,等国内外大型企业建立合作关系。放眼未来,我们将努力把 iDVx 实验室打造一个成具有国际影响力的创新研究型实验室。

核心技术

实验室正在努力打造并逐步完善以下两项核心技术:

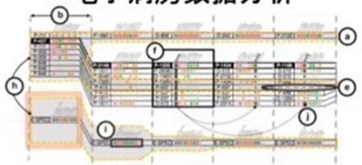
- **针对复杂大规模数据的精准智能化异常检测**。我们从多个领域出发,创造出一系列精准的数据分析算法及模型、直观的可视分析系统、以及高效的人机交互方式,并将相关技术成果应用于疾病检测、互联网信息安全监测、以及城市安全防控等领域。

- **用于辅助视觉传达设计的人工智能算法**。实验室正在开一系列用于辅助视觉传达设计的人工智能算法及相关工具平台。这些技术的综合应用,将大大提高设计过程的自动化程度,简化设计流程,进而辅助设计师从事更有创造力的工作。

智慧医疗

精准医疗(Precision Medicine)是以个体化医疗为基础、随着基因组测序技术快速进步以及生物信息与大数据科学的交叉应用而发展起来的新型医学概念与医疗模式。实验室将核心异常检测算法应用于疾病检测及风险预测,对心脏病、肺癌、等疾病进行辅助诊断,并通过相关电子病历数据对未来风险作出预测。

电子病历数据分析



各大医院所广泛使用的电子病历系统记录了患者的就诊过程及医生制定的诊疗方案。针对该数据,我们分阶段总结历史事件,全面分析病患所面临的潜在风险,预测可能发生的疾病状况,以及为医生提出相关的诊疗建议,并为病患提供及时的报警。

基于人工智能的心电判读

通过人工智能算法,从海量心电数据中检测异常信号,及心率不齐的相关病症,辅助医生进行心电图判读,同时也帮助患者更加及时的了解自身状况。

智能传达设计

传达设计包括图形设计(例如,可视化、信息图、海报、广告等)及字体设计两个部分。传统的视觉传达设计需要耗费设计师的大量精力进行排版、构图。我们利用人工智能技术解决传达设计中的关键问题,发现并总结视觉传达过程的内在规律 及 影响视觉传达力、影响力、和表现力 的根本原因,并据此设计智能算法用以辅助制定最优化的传达设计产品,提高设计自动化程度,从而进一步提高设计师的生产力,让设计作品更加个性化,精准应对用户需求。

招生及实习计划

iDVx 实验室的成员由一群来自同济大学、浙江大学、武汉大学、香港科技大学、复旦大学、上海交通大学、华东师范大学、上海纽约大学的博士、硕士、实习生、及访问学生共同组成。每年一次的暑期实习计划更是吸引了越来越多的优秀学生加入。曾经入选实习计划的学生先后拿到了 Harvard, Yale, UIUC, Georgia Tech, UCSD 等美国名校的录取通知。留在组里继续攻读硕士博士的研究生更是连续发表高水平的国际学术论文。实验室基于同济大学"软件学院"及"设计创意学院"面向 "数学"、"计算机"、"设计学"等相关专业,同时招收"软件工程"及"设计工学" 硕博研究生、实习生、及研究助理。招生详情请关注实验室官方网站。关键时间节点如下:

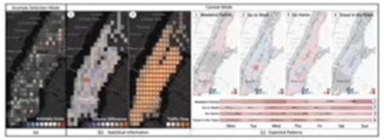
硕博推免:每年9月
实习计划:每年1 - 4月接受申请,6-9月实习
研究助理:全年有效

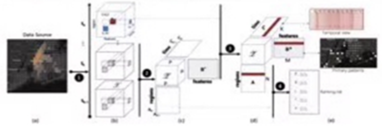
地址:上海市阜新路281号,同济大学设计创意学院,is218
电话:(+86)-21-65986671
电邮:idvx.lab@tongji.edu.cn

其他应用及案例

城市安全

城市安全与我们每个人息息相关。2014年跨年夜上海外滩发生的踩踏事件为我们敲响了警钟。如何在城市中做到防患于未然,打破传统的亡羊补牢安全模式,是智慧城市研究的重点之一。在这个领域,iDVx实验室成功的将异常检测算法应用于城市异常交通检测之上,为城市交通安全保驾护航。

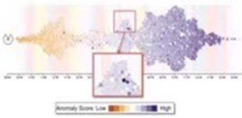




Voila:城市异常交通流量监控及检测系统
(IEEE VAST 2017)

信息安全

信息安全问题与我们的日常生活息息相关,例如,如何识别电信诈骗?如何确保用户公开在网络上的信息不被恶意使用等等,这些问题都是新的网络安全问题,传统的防火墙及杀毒软件都无法解决。在这个领域,iDVx实验室做出了一系列基于用户行为分析的异常检测技术。



FluxFlow:针对社交媒体的谣言检测算法及可视分析系统(IEEE VAST 2014,最佳论文提名奖)

TargetVue: 社交媒体中异常用户行检测算法及可视分析系统 (IEEE VAST 2015)

INDUSTRIAL REVOLUTION OF DATA

Joe Hellerstein, UC Berkley, 2008

Outline

① Data Dimension
- D1
- D2
- D3
- High Dimension

② High-Dimensional Data Visualization
- Dimensionality Reduction
- Scatter-plot Matrix
- Parallel Coordinates
- Glyph-based Methods
- “Small Multiples”
- Interaction: “Dust & Magnet”

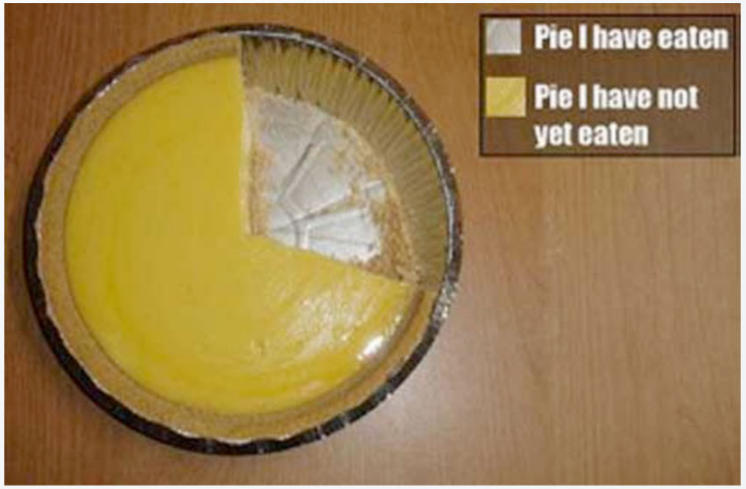
Data Dimension

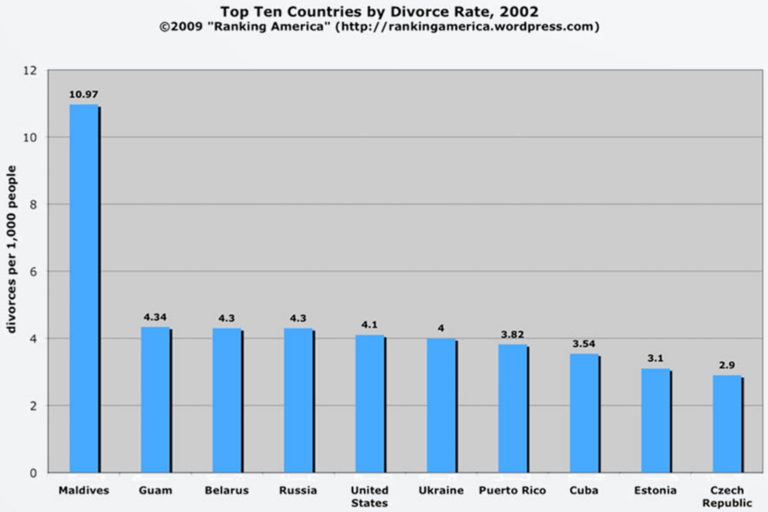
Review: Dimension

- Dimension (Number of attributes):
  - 1-D.
  - 2-D.
  - 3-D.
  - High Dimension.

1-D Data

"eaten or not"

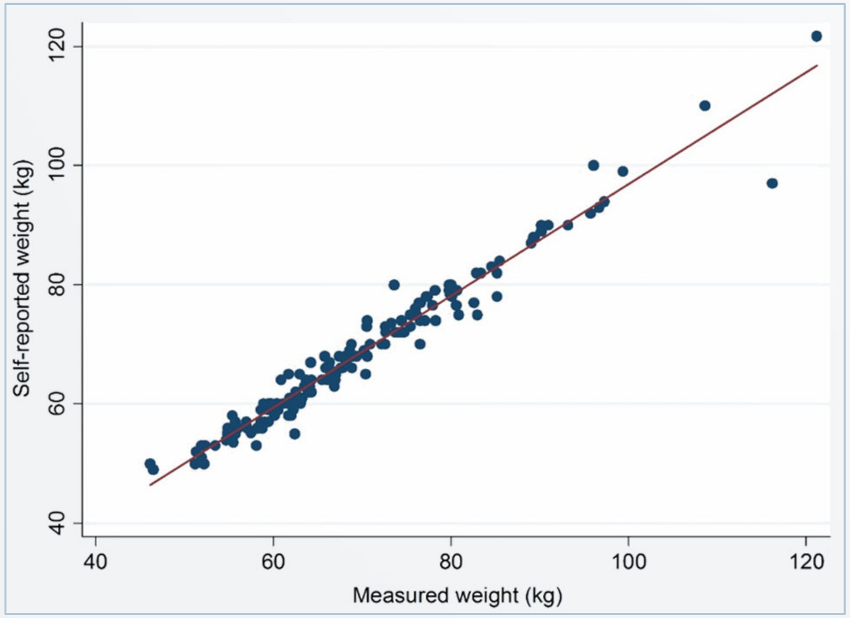


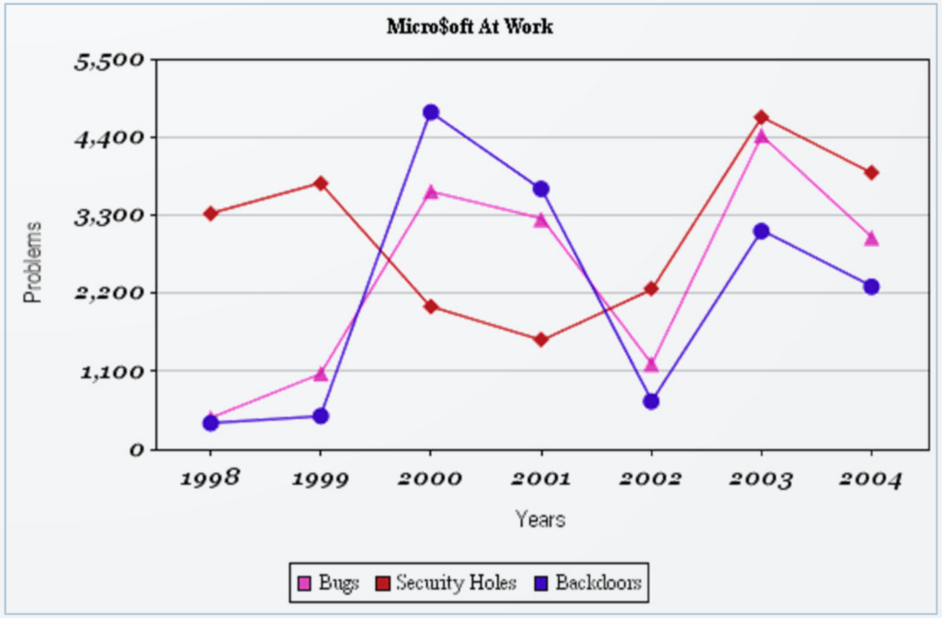


1- D data for each bar

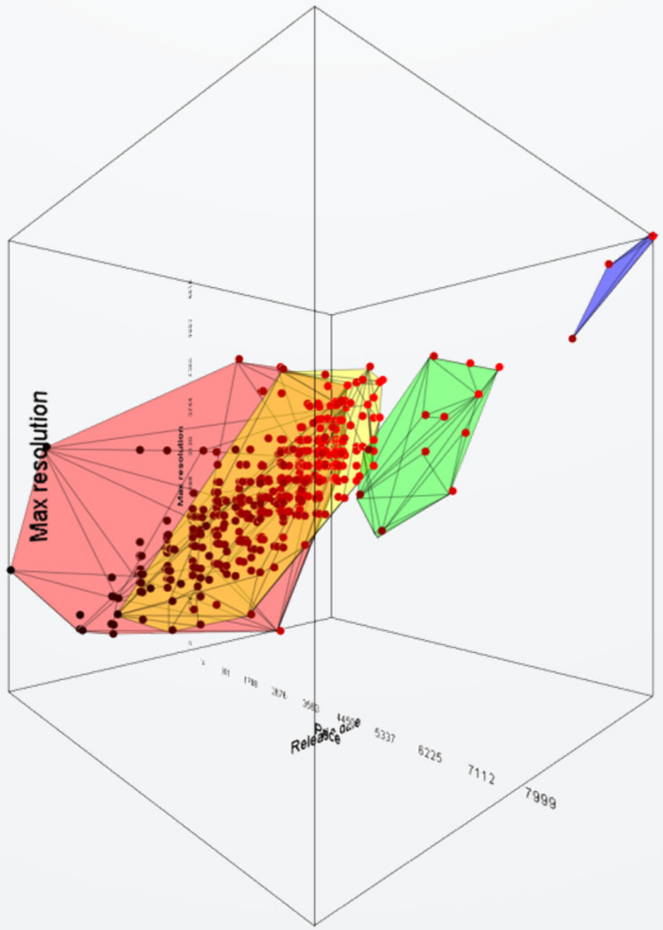
2-D Data

2-D plots





3-D Data



Elmqvist et al. "Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation."
IEEE transactions on Visualization and Computer Graphics 14.6 (2008): 1539-1148.

High-Dim Data

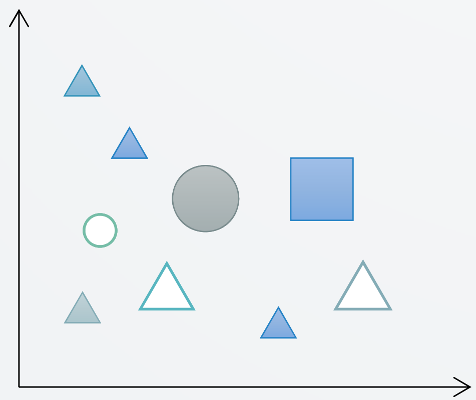
- How to visualize high-dimensional data in visual space(2-D or 3-D)？

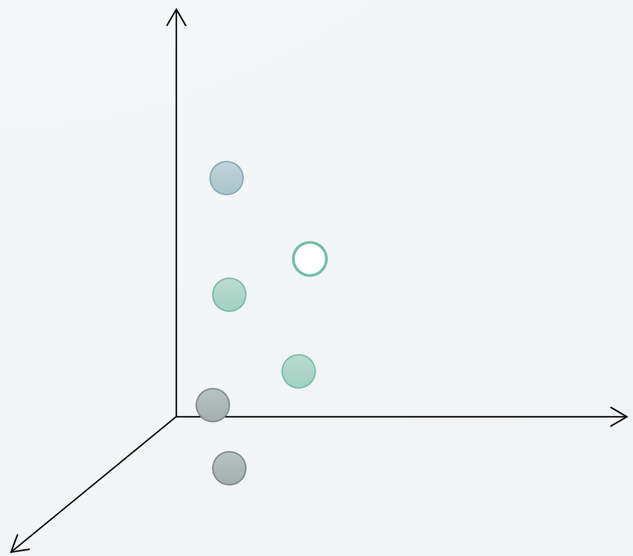
	身高	体重	年龄	性别	教育程度	籍贯
张三	180cm	65kg	23	男	大学	上海
李四	168cm	55kg	18	女	高中	浙江
赵五	175cm	75kg	53	男	初中	广东
…						

Simple Solutions

- Add more channel on 2-D or 3-D plots.

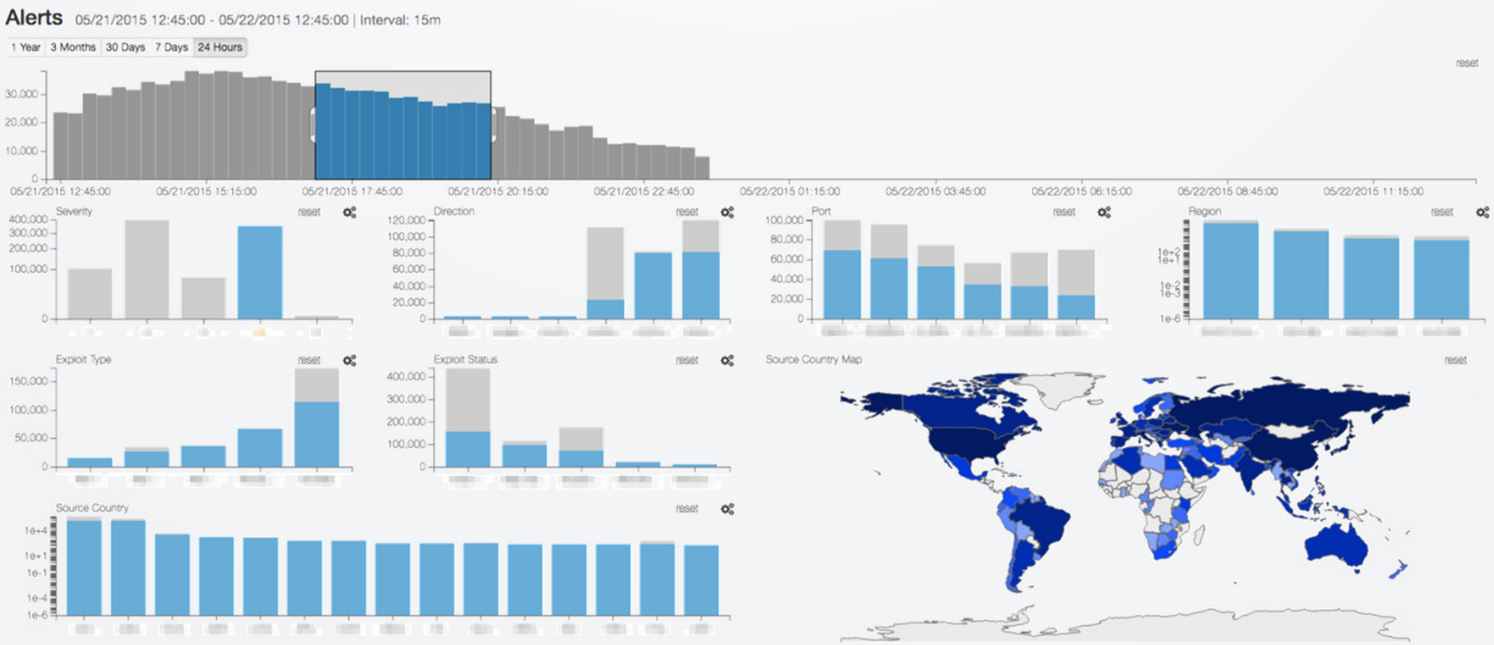
(Shape/fill style/color/size of points)





Simple Solutions

- Multiple coordinated views: present some attributes of objects in a view.



Andrew et al. "Leveraging Interaction History for Intelligent Configuration of Multiple Coordinated Views in Visualization Tools."
LIVVIL: Logging Interactive Visualizations & Visualizing Interaction Logs (2016).

High-Dimensional Data Visualization

- More solutions?

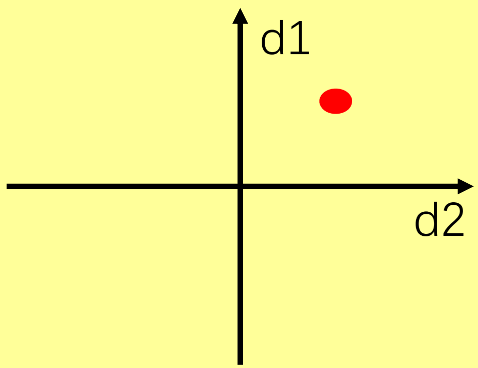
High-Dimensional Data Visualization

OUTLINE

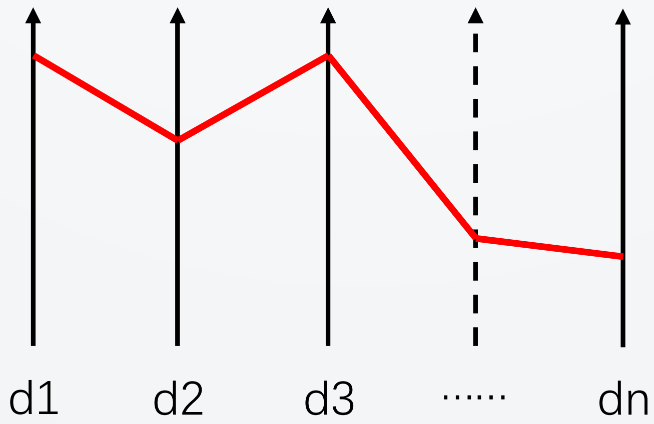
Approaches

- **Coordinate Systems**
  - **Coordinate System**
  - **Scatter-plot Matrix.**
  - **Parallel Coordinates.**
  - **Dimensionality Reduction.**
- Glyph-based Methods.
- Pixel Oriented Techniques
- “Small Multiples”.
- Visual Diagnosis

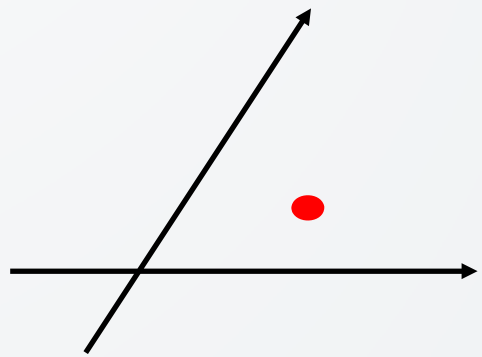
Coordinate Systems



Scatter Plots

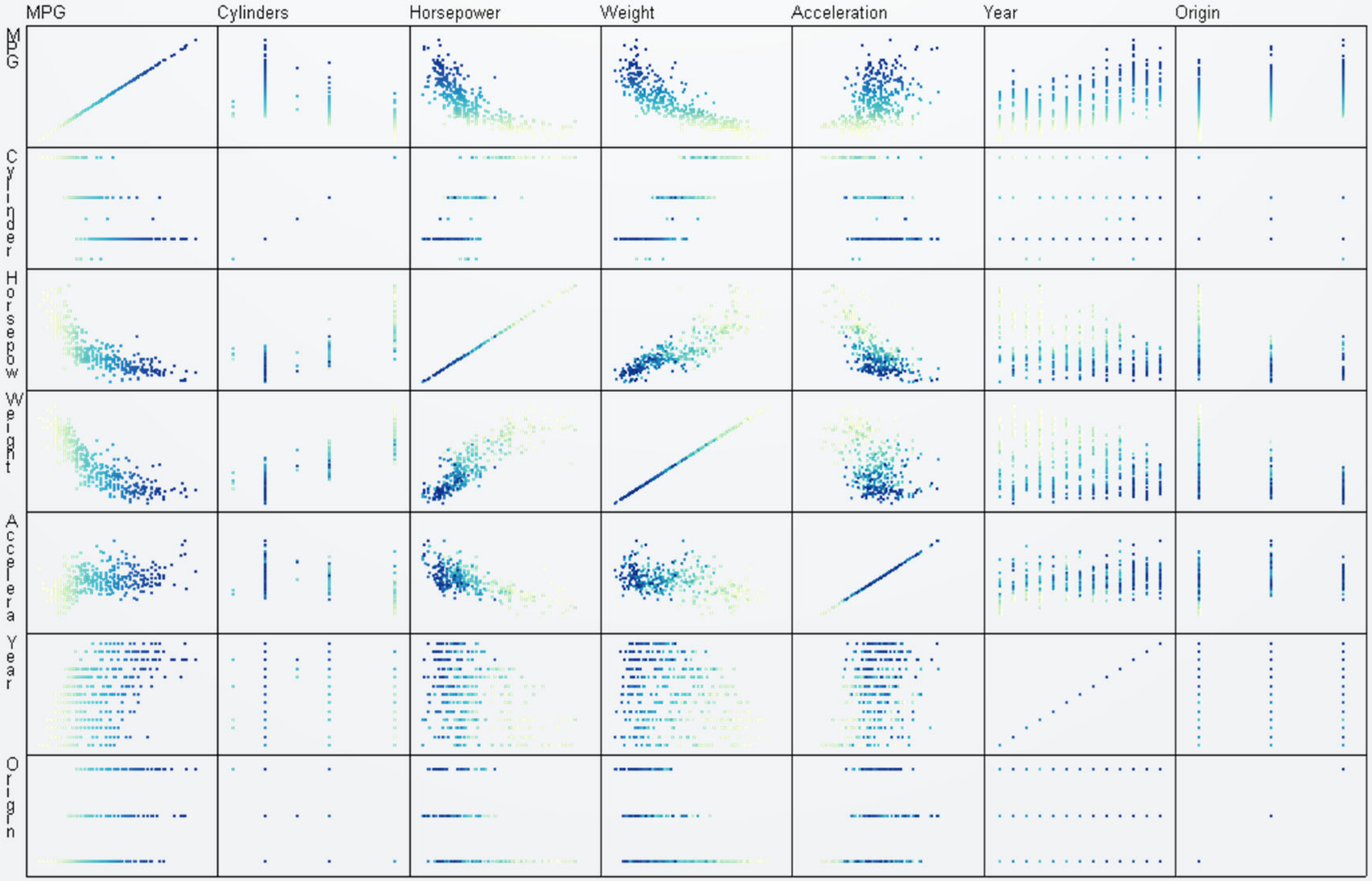


project into the space defined by parallel axes



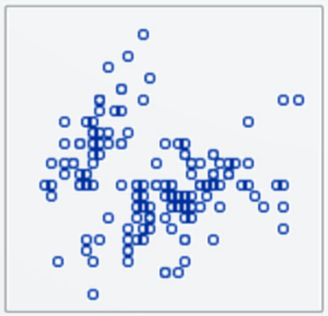
Projection based approaches

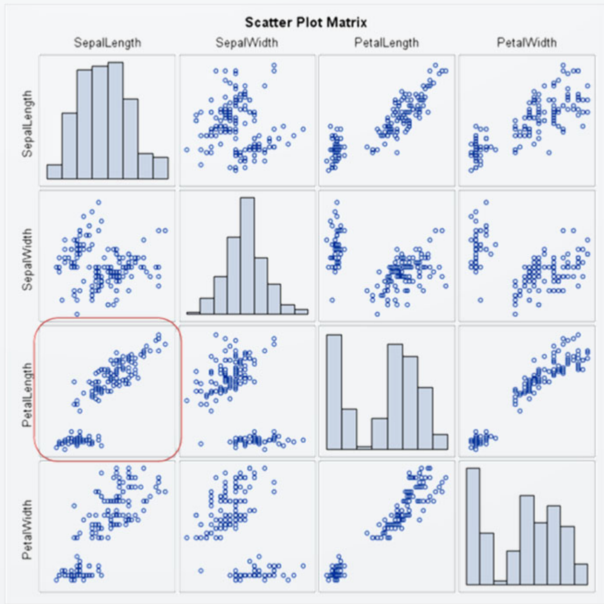
Scatter-plot Matrix



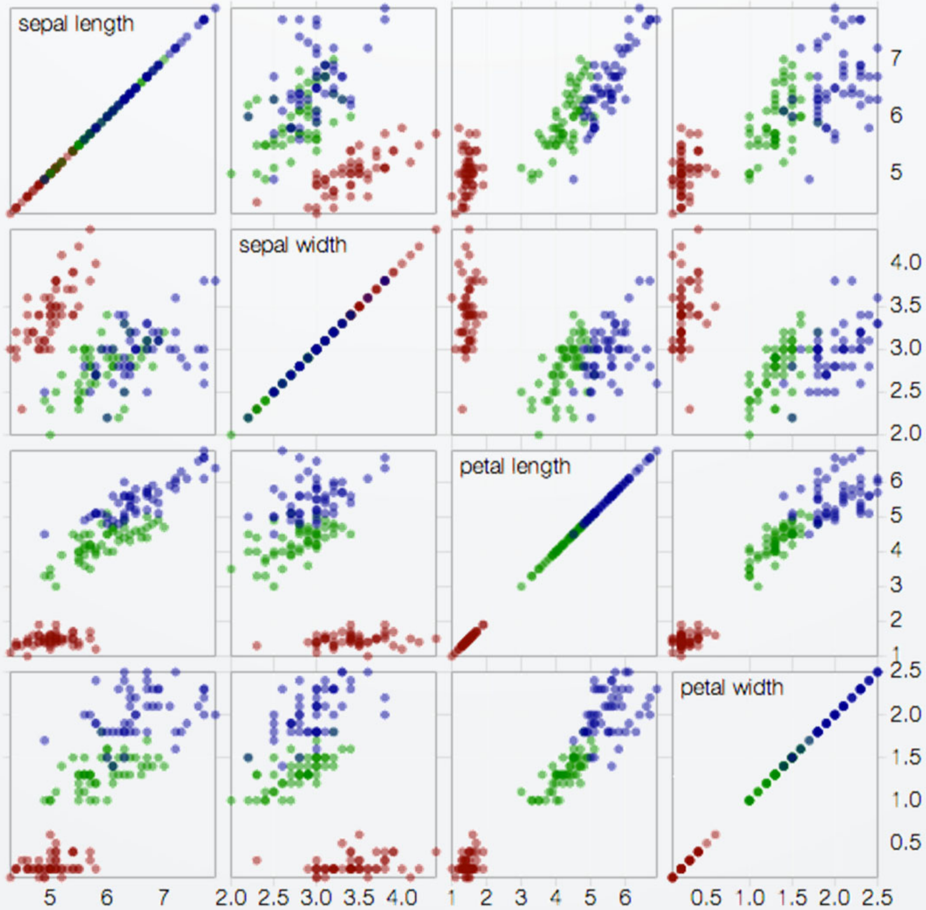
Scatter-plot Matrix

- 2-D plot for each dimension pair.
- Display correlations between dimensions.
- Number of 2-D plots proportional to square of dimensions.





Iris Dataset

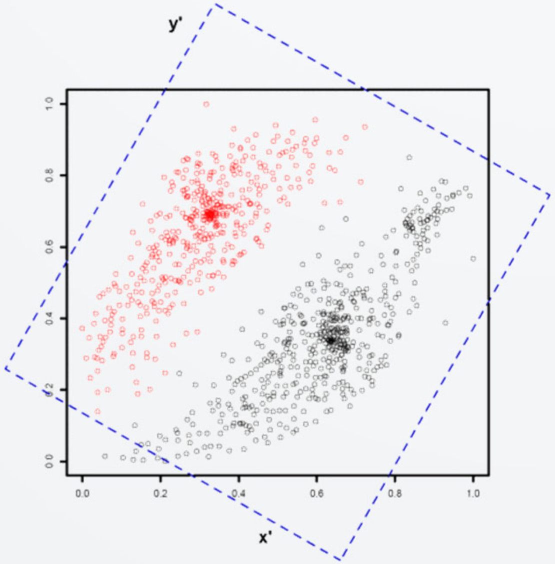


D3.js Example

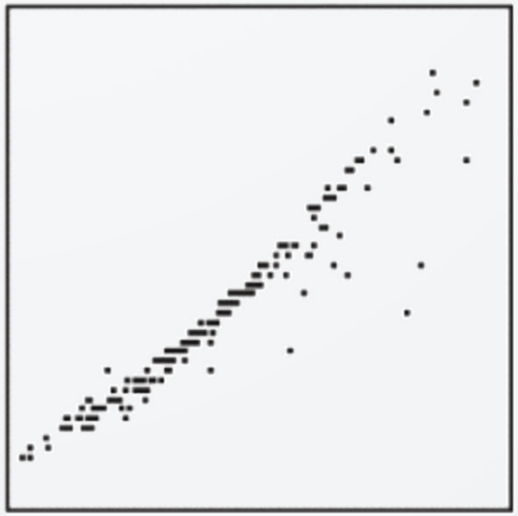
Automatically Exploration

- Recommend scatter-plots with interesting patterns automatically.

Clusters

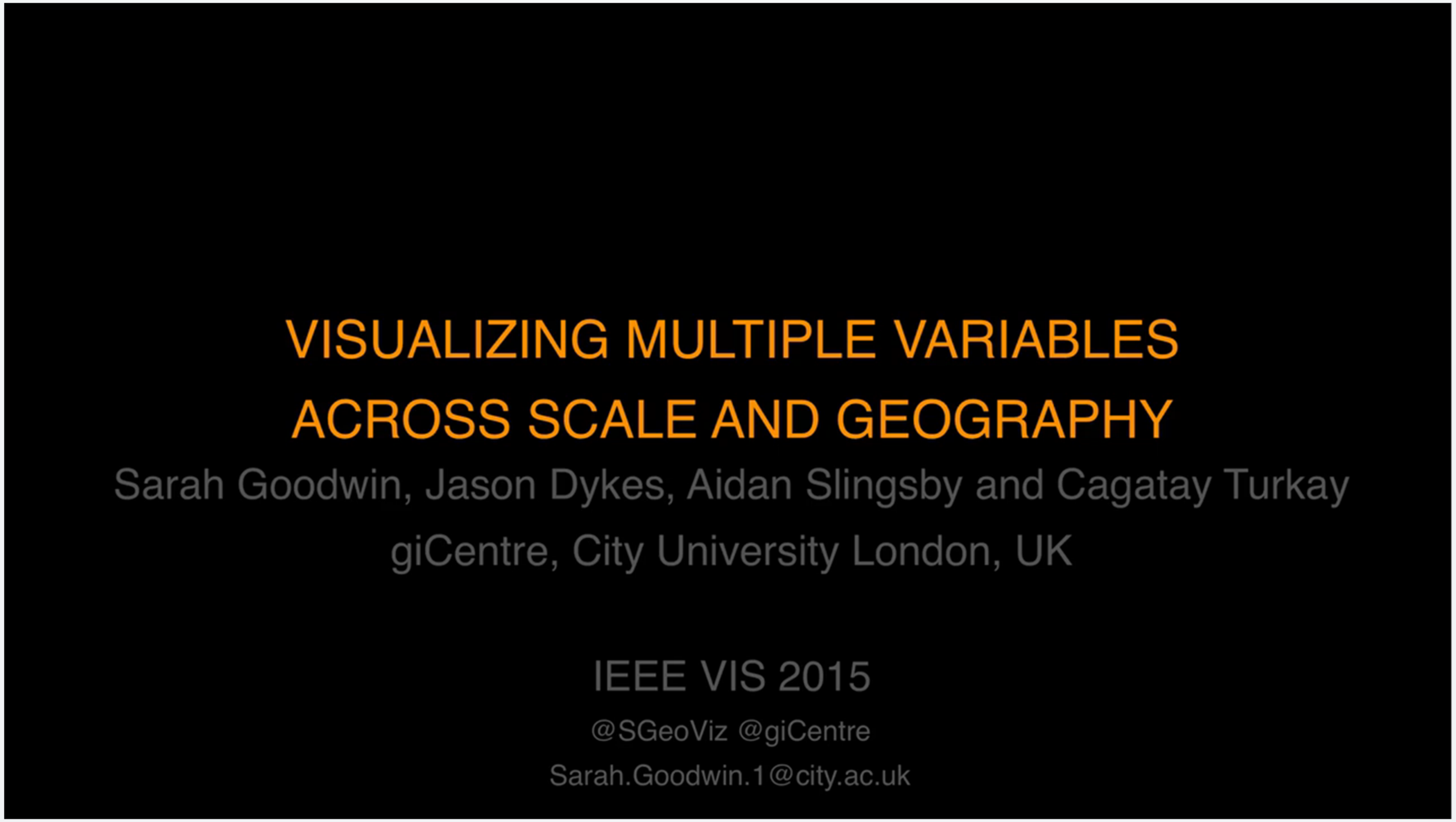


Correlations



Tatu A, Albuquerque G, Eisemann M. “Automated analytical methods to support visual exploration of high-dimensional data.”
IEEE Transactions on Visualization and Computer Graphics (2011).

Across Scale and Geography

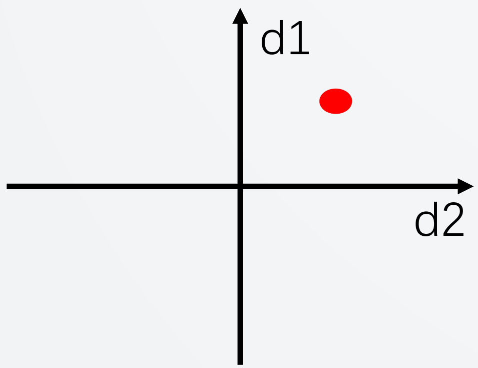


Goodwin S., Dykes J., Slingsby A. "Visualizing multiple variables across scale and geography."
IEEE Transactions on Visualization and Computer Graphics (2016).

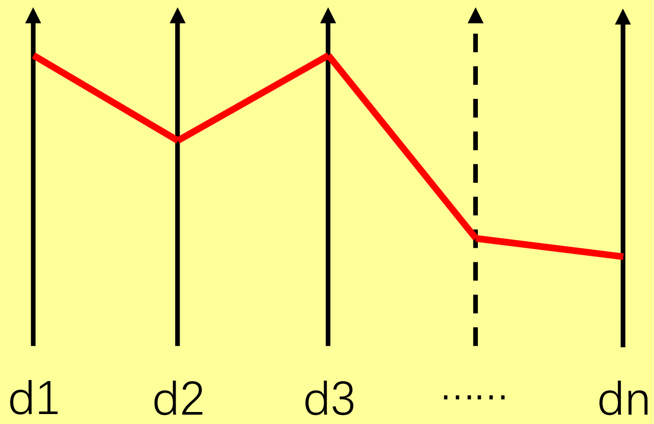
Animated Scatter-plot Matrices

Chen H, Engle S, Joshi A. “Using Animation to Alleviate Overdraw in Multiclass Scatterplot Matrices.”
ACM CHI Conference on Human Factors in Computing Systems (2018).

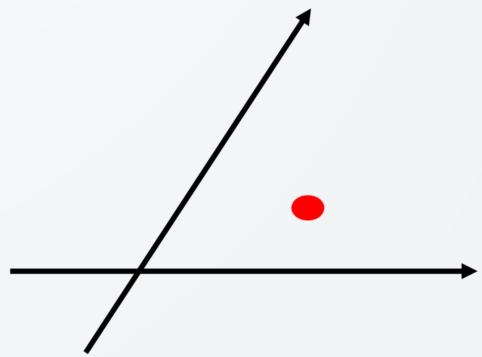
Coordinate Systems



Scatter Plots



project into the space defined by parallel axes

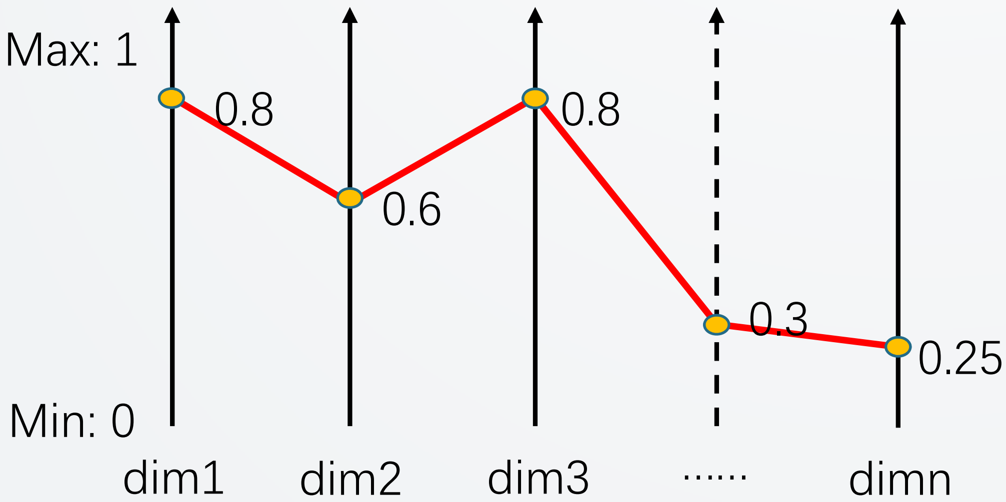


Projection based approaches

Parallel Coordinates

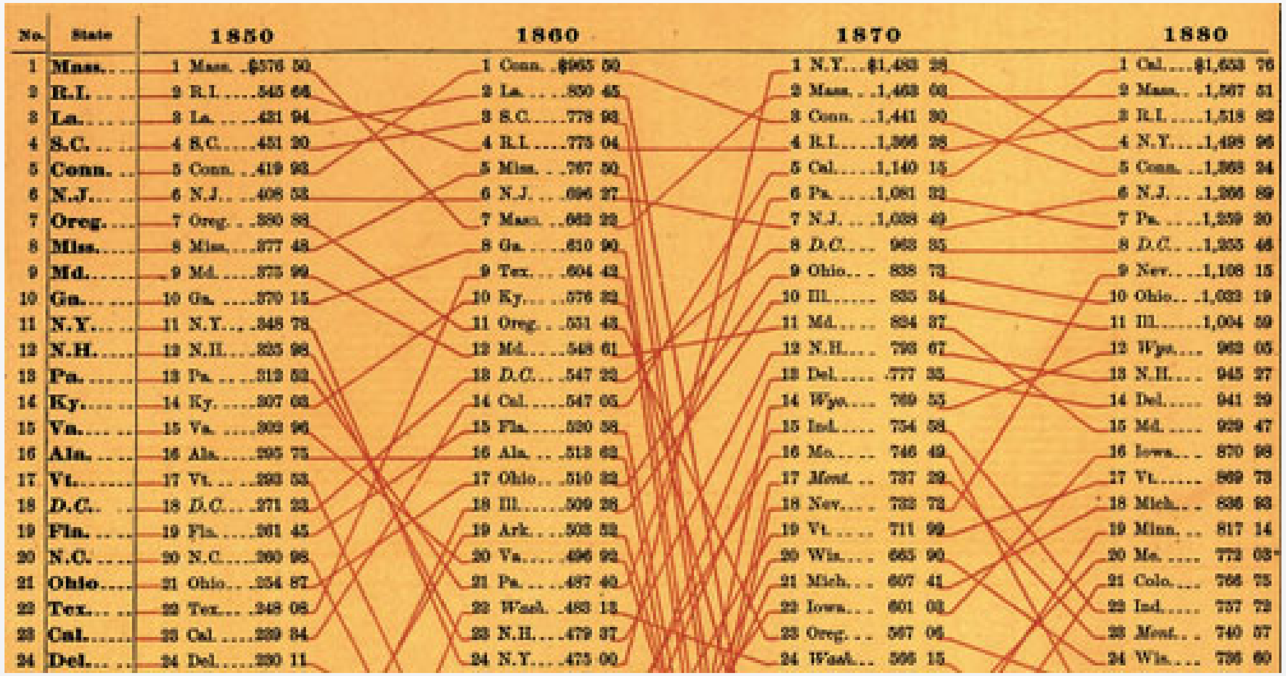
- Presented by Inselberg in 1985 for high-dimensional geometry.
- Parallel axes.
- Data points represented by lines.

(3) Parallel Coordinates – Visual Designs

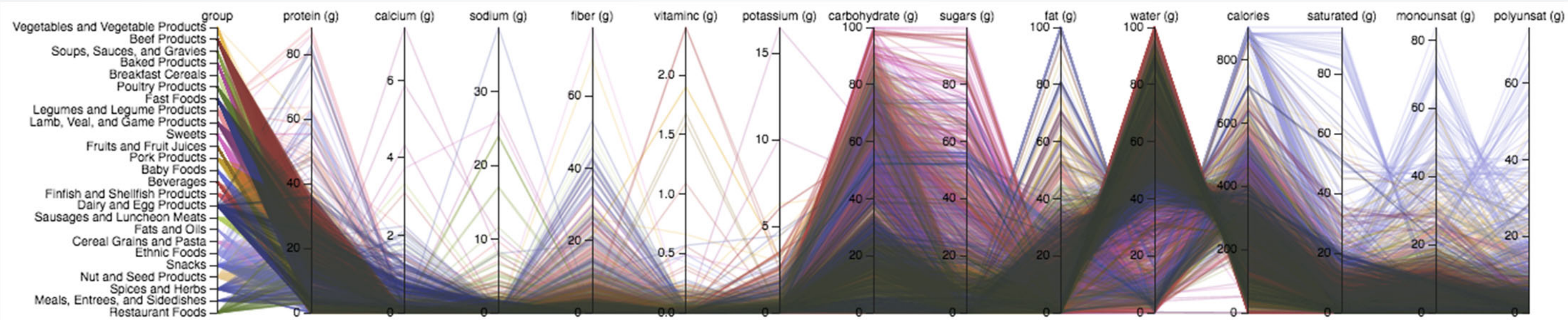


- Dimensions as parallel axes
- Data items as line segments
- Intersections on the axes indicates the values of the corresponding attributes

Parallel Coordinates in 1880



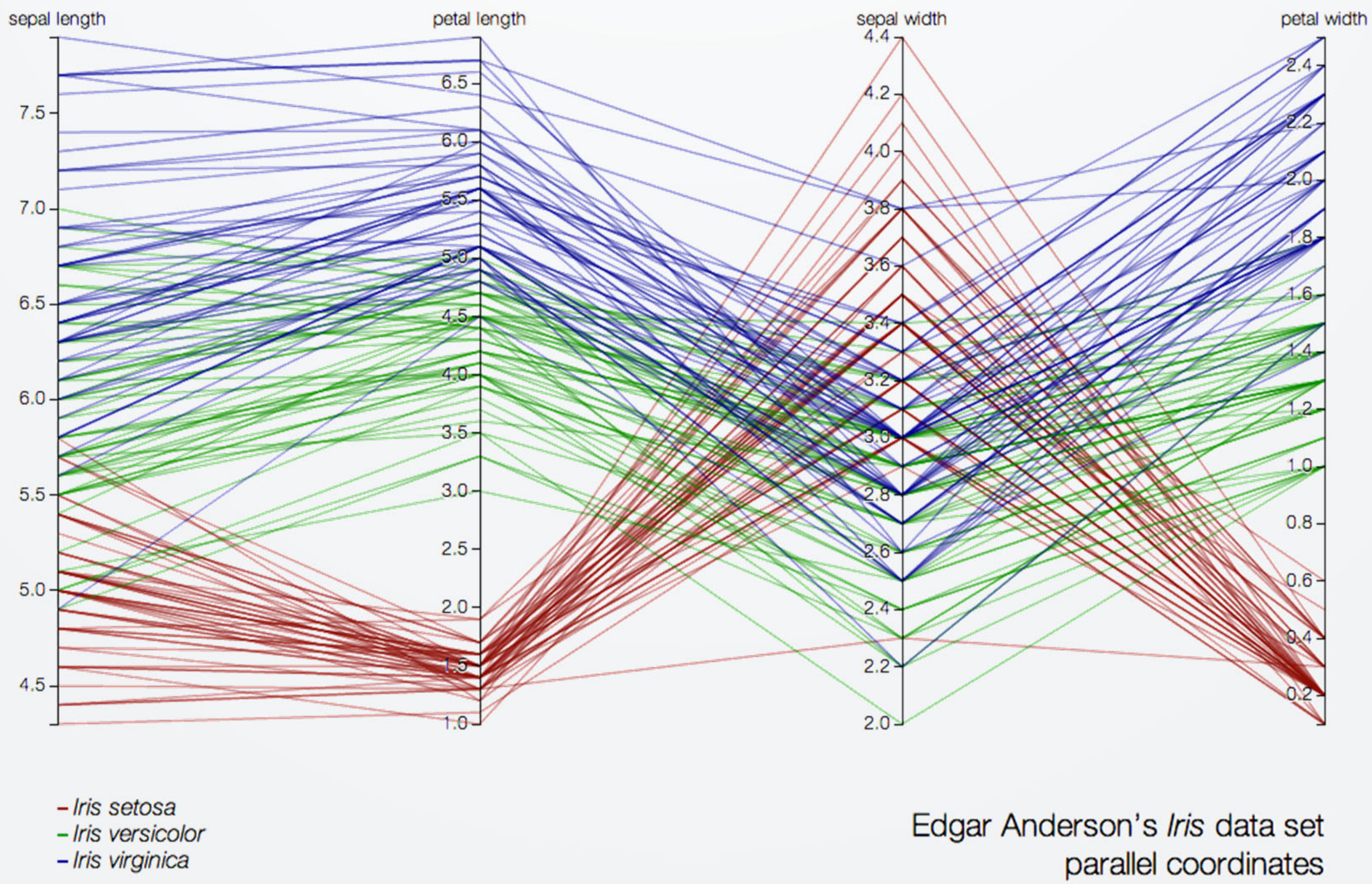
Parallel Coordinates & Interactions



https://youtu.be/ypc7Ul9LkxA

https://syntagmatic.github.io/parallel-coordinates/

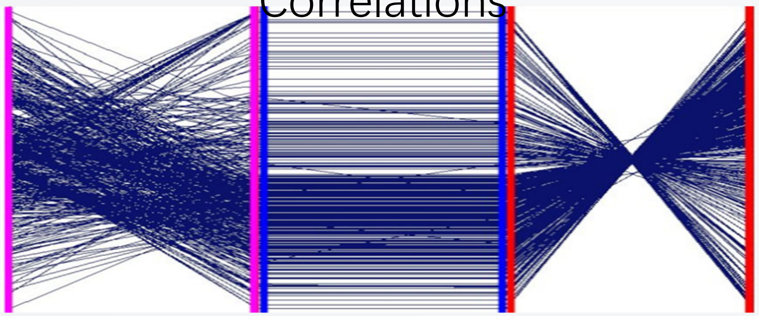
D3.js



Edgar Anderson's *Iris* data set
parallel coordinates

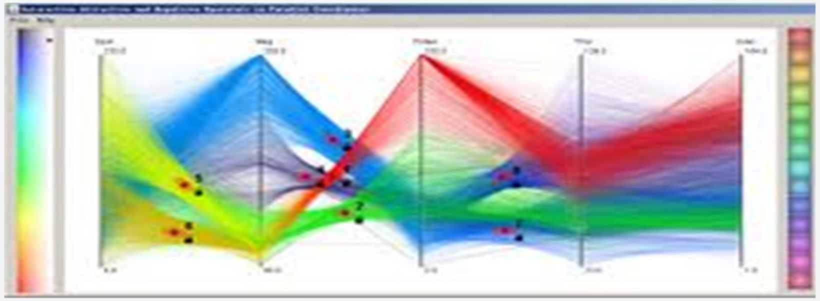
Data Patterns in PCP

Correlations

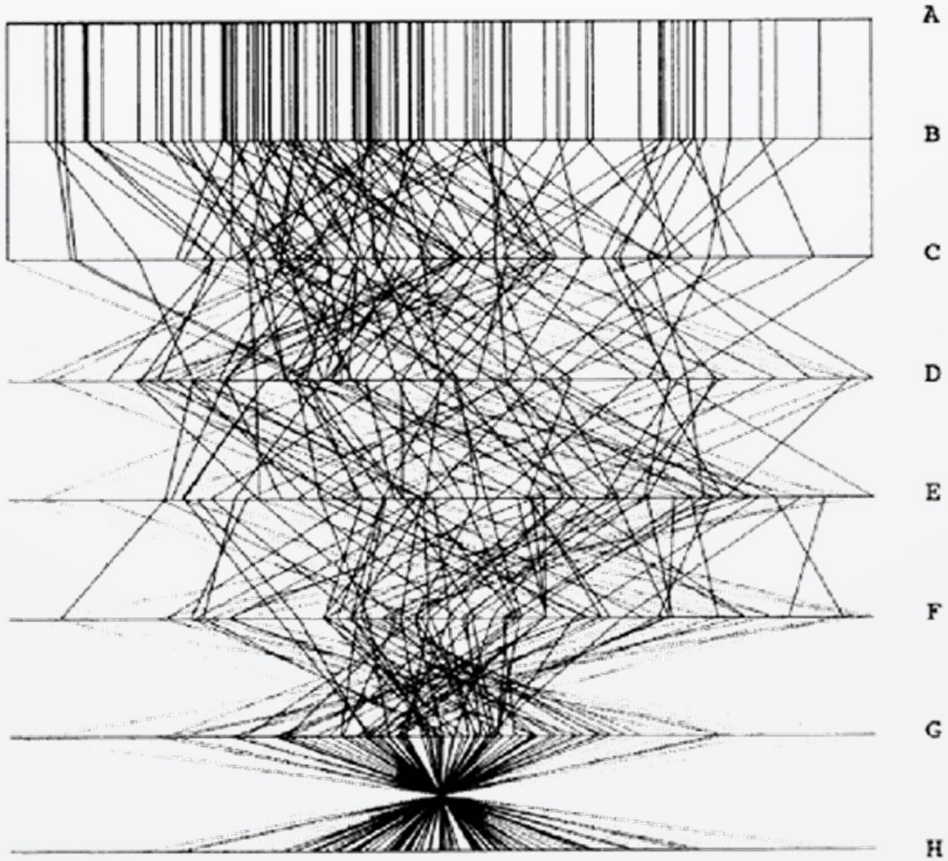


positive negative

Clusters



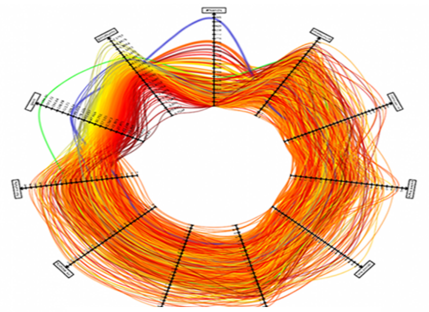
Correlations



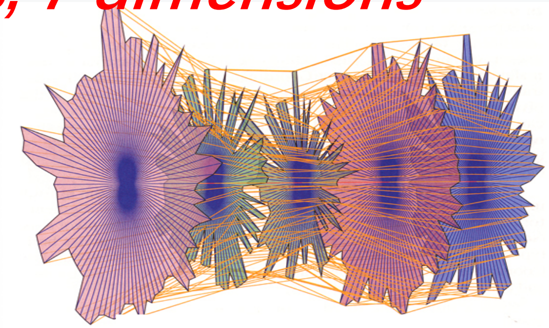
Wegman, Edward J. "Hyperdimensional data analysis using parallel coordinates."
Journal of the American Statistical Association(1990).

Parallel Coordinates – Examples

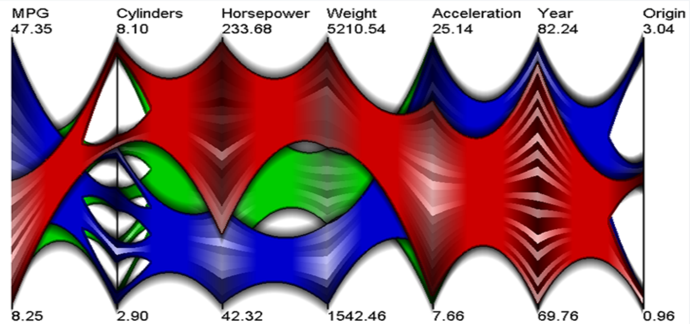
Cars dataset:
392 data items, 7 dimensions



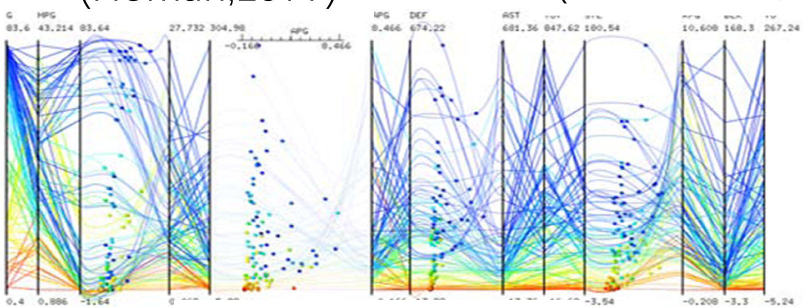
(Homan,1977)



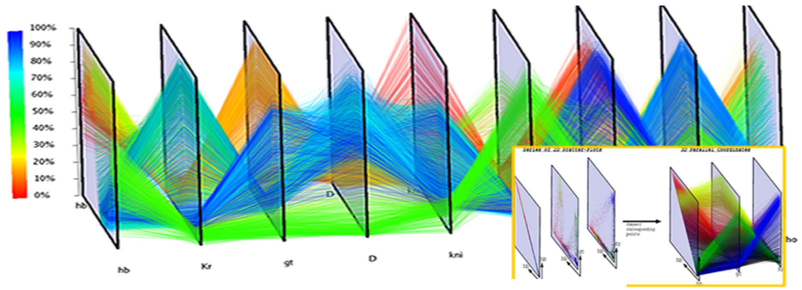
(Fanea et al., 2005)



(McDonnell & Mueller, 2008)

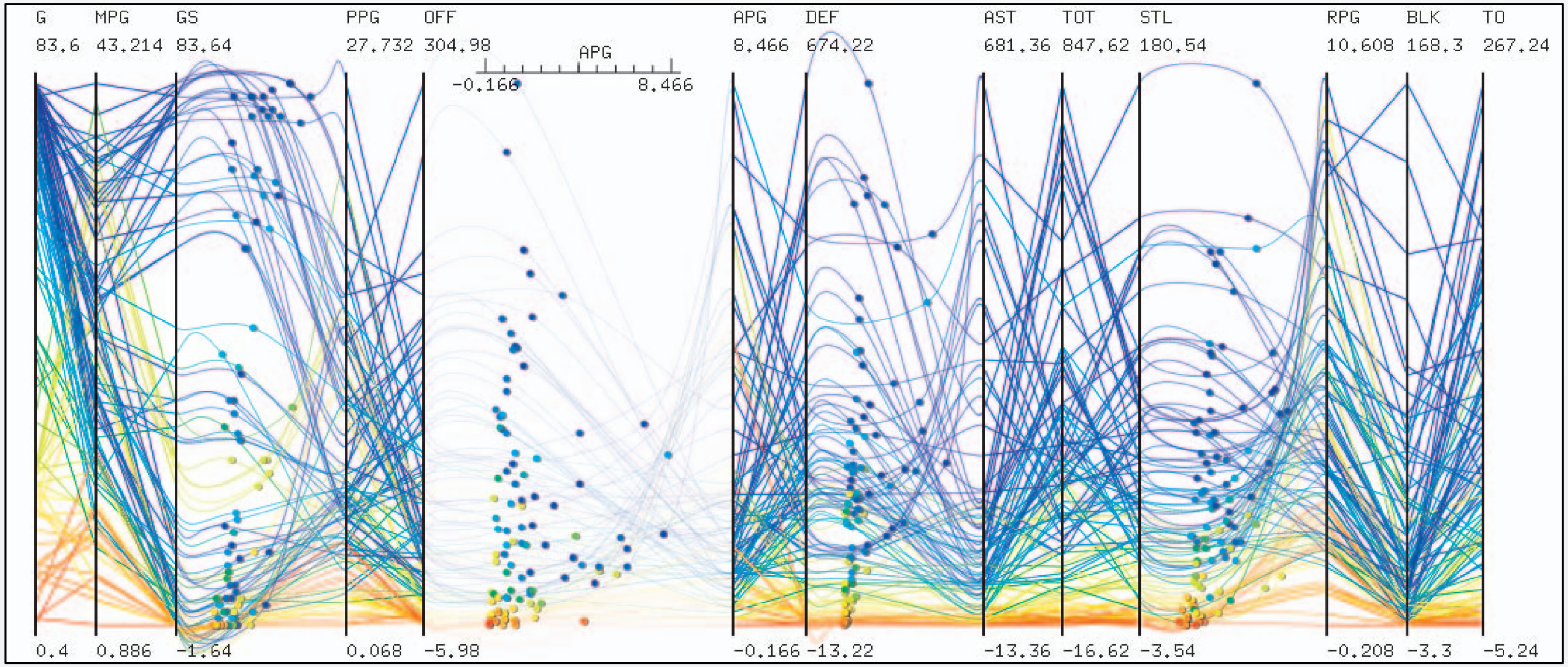


(Yuan et al., 2009)



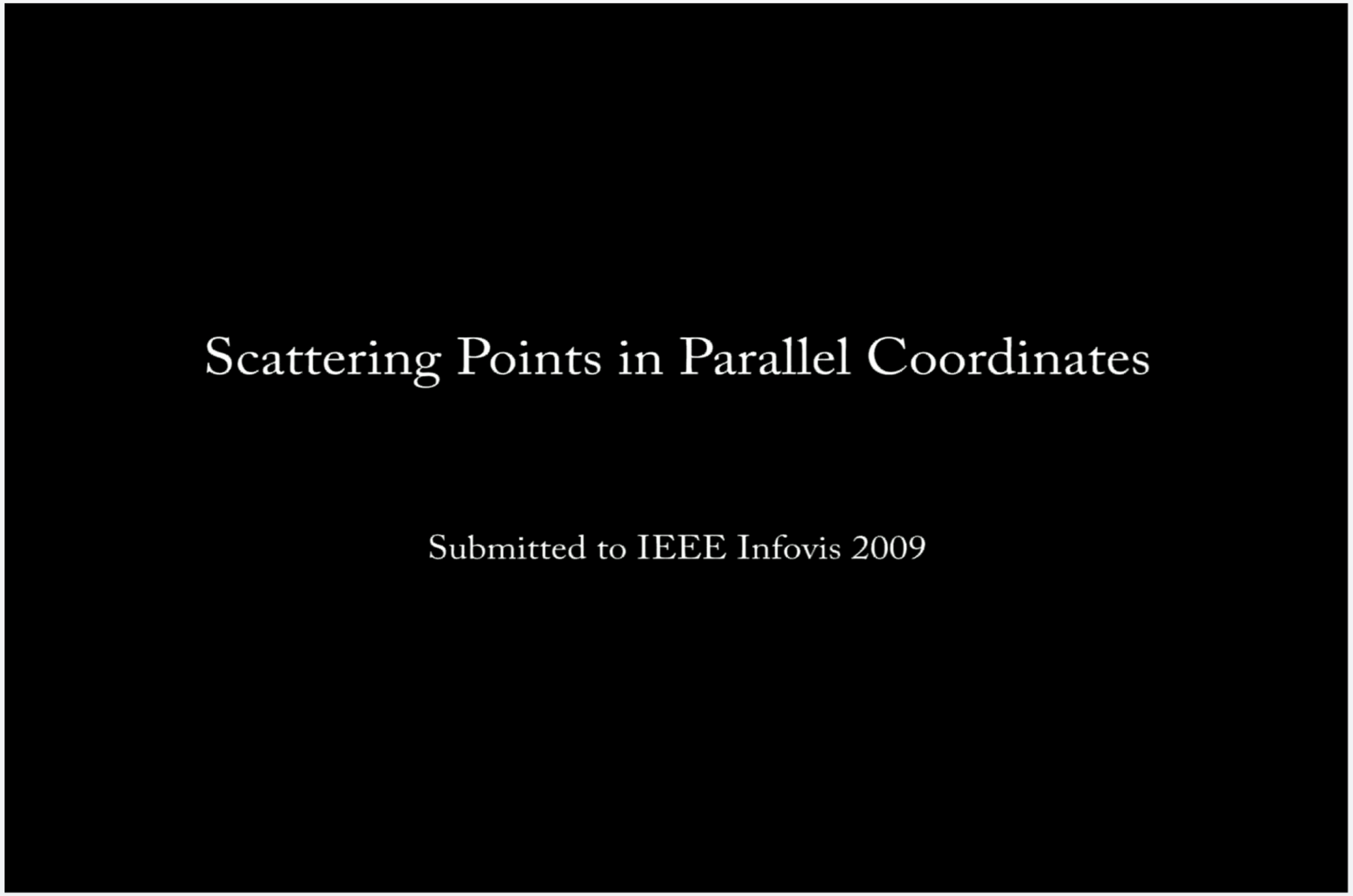
(Rubel et al., 2006)

Parallel Coordinates with Scatter-plots



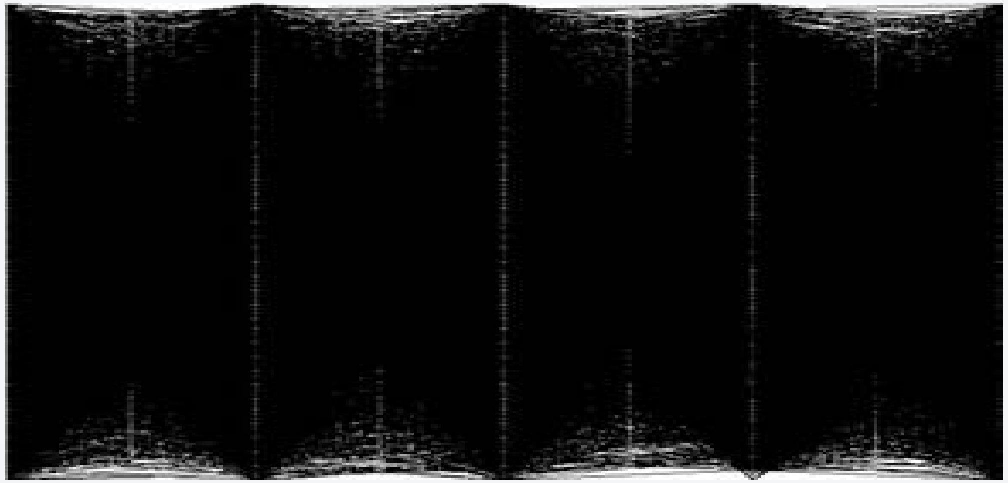
Yuan et al. "Scattering Points in Parallel Coordinates. "
IEEE Transactions on Visualization and Computer Graphics
(2009).

Parallel Coordinates with Scatter-plots



X Yuan, Guo P, H Xiao et al. Scattering points in parallel coordinates.
IEEE Transactions on Visualization and Computer Graphics (2009).

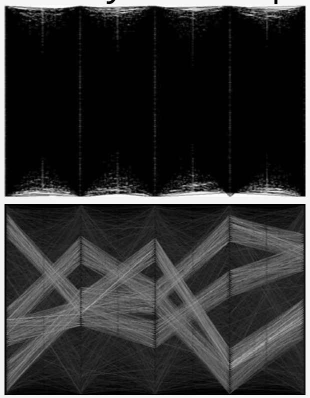
Parallel Coordinates – Clutter Reduction



- Clustering and filtering approaches
- Dimension reordering approaches
- Visual enhancement approaches

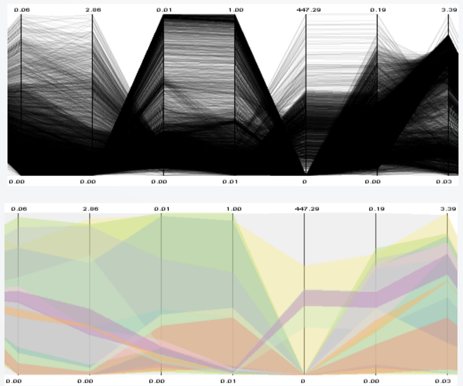
Clutter Reduction – Clustering and Filtering

Filtering on
Density & Frequency



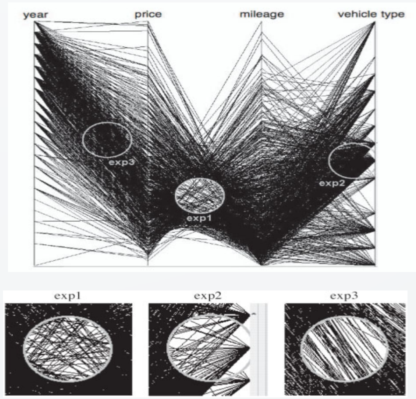
(Artero et al., 2004)

Clustering



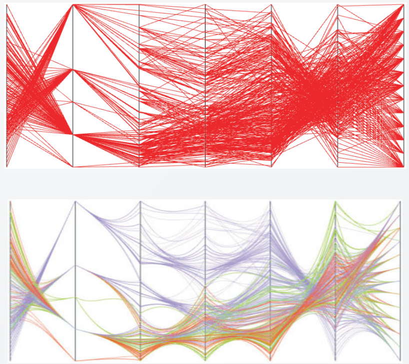
(Novotny et al., 2004)

Sampling



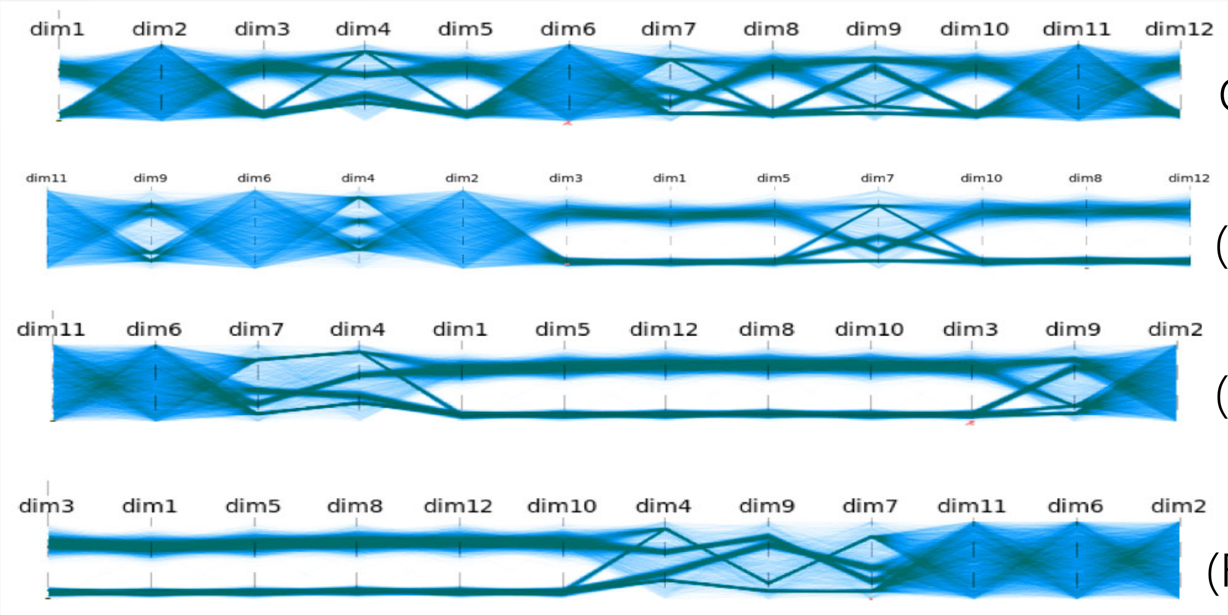
(Ellis & Dix, 2006)

Visual Clustering



(Zhou et al., 2008)

Clutter Reduction – Dimension Reordering



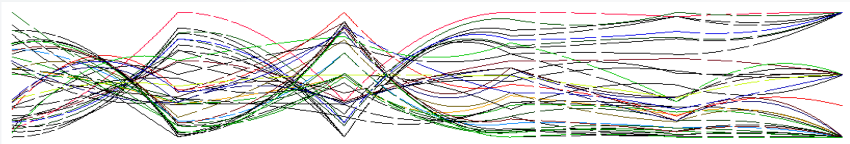
Original order

(Wang et al, 2004)

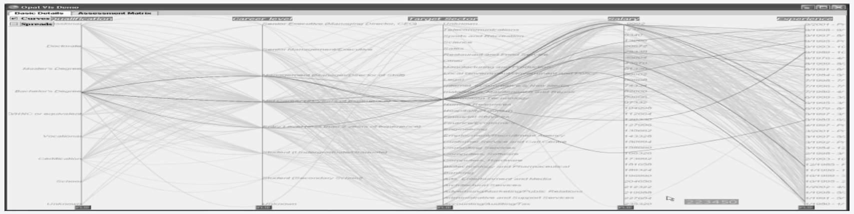
(Ankerst et al, 2004)

(Ferdosi & Roerdink, 2010)

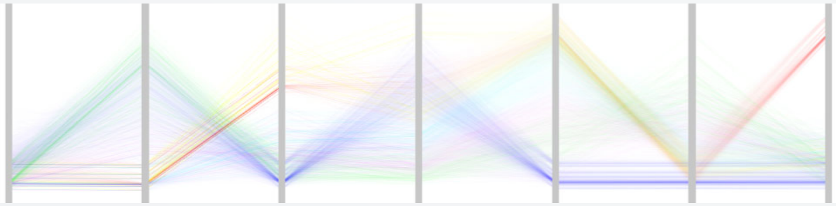
Clutter Reduction – Visual Enhancement



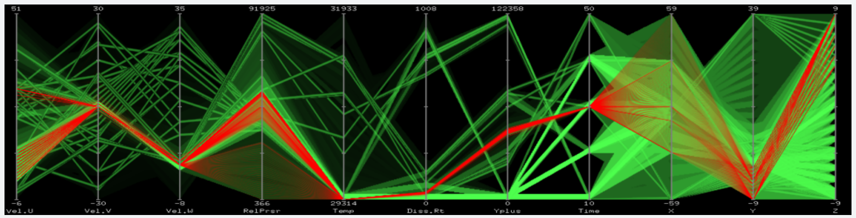
(Theisel, 2000)



(Graham & Kennedy, 2003)



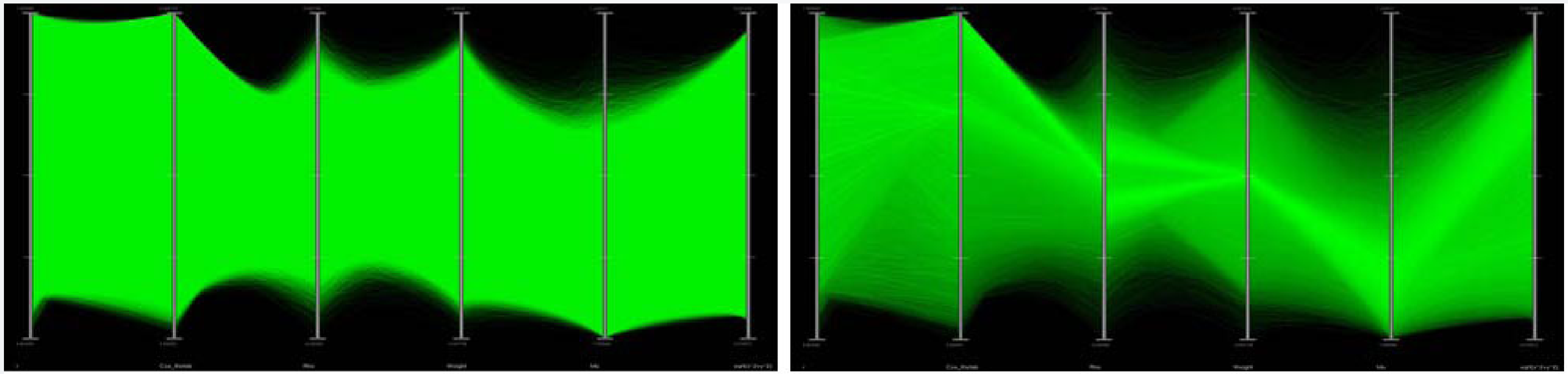
(Johansson et al., 2005)



(Novotny & Hauser, 2006)

Transparent Parallel Coordinates

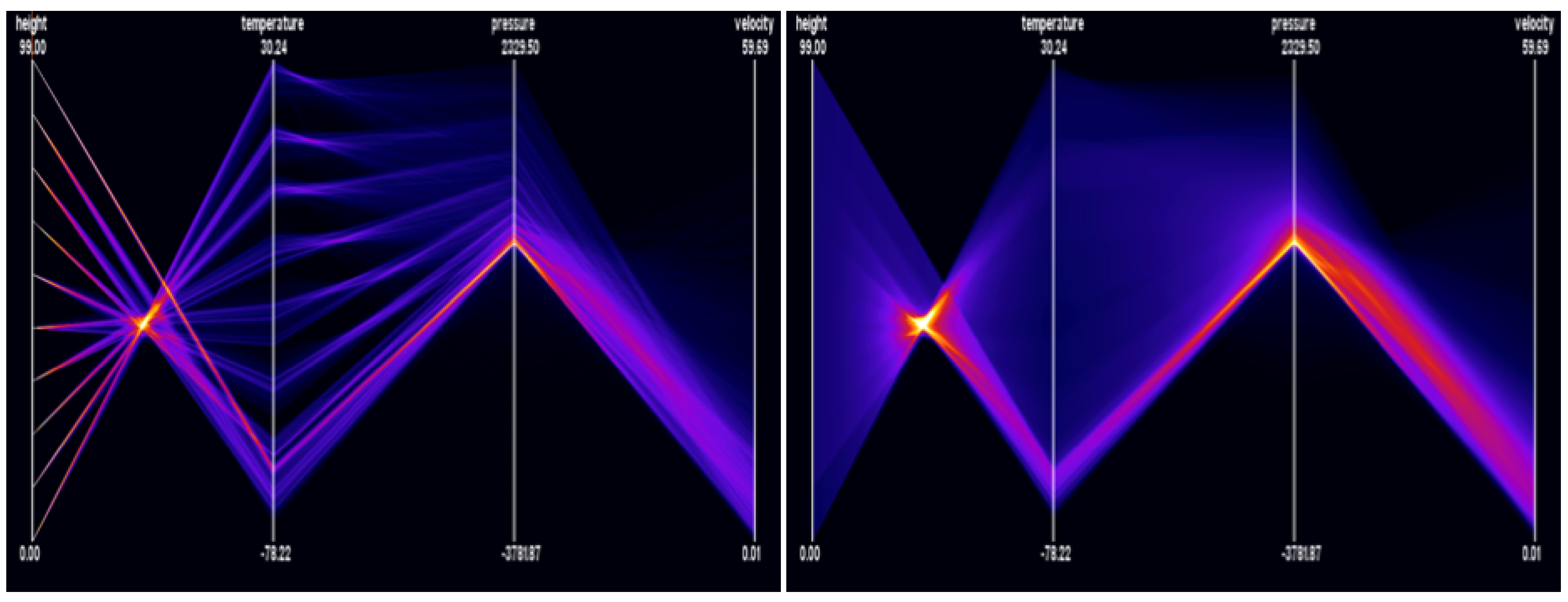
Emphasize main trends



Chad Jones et al. “An Integrated Exploration Approach to Visualizing Multivariate Particle Data. “
Computing in Science & Engineering (2008).

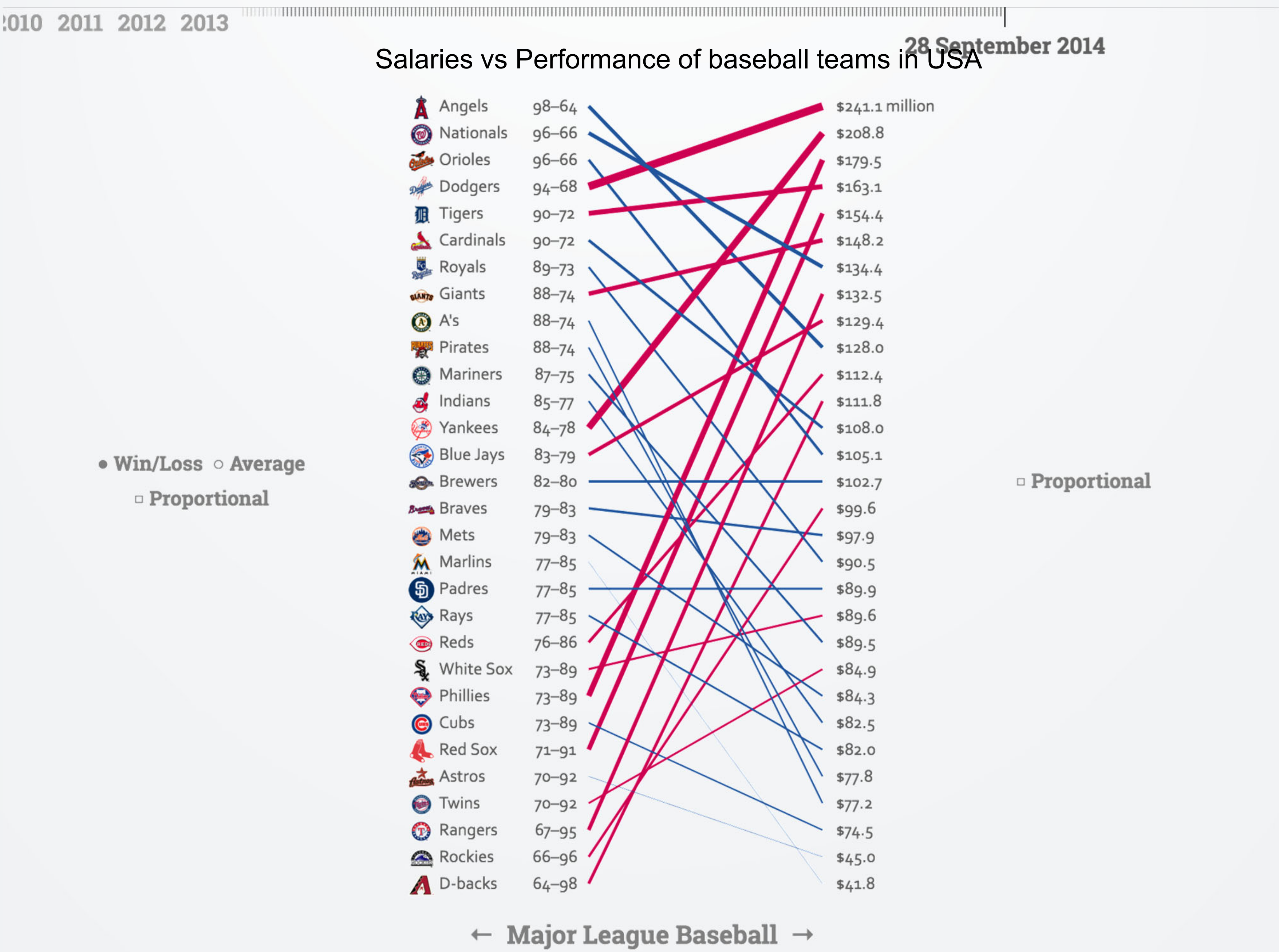
Continuous Parallel Coordinates

Use heatmap to show the trends

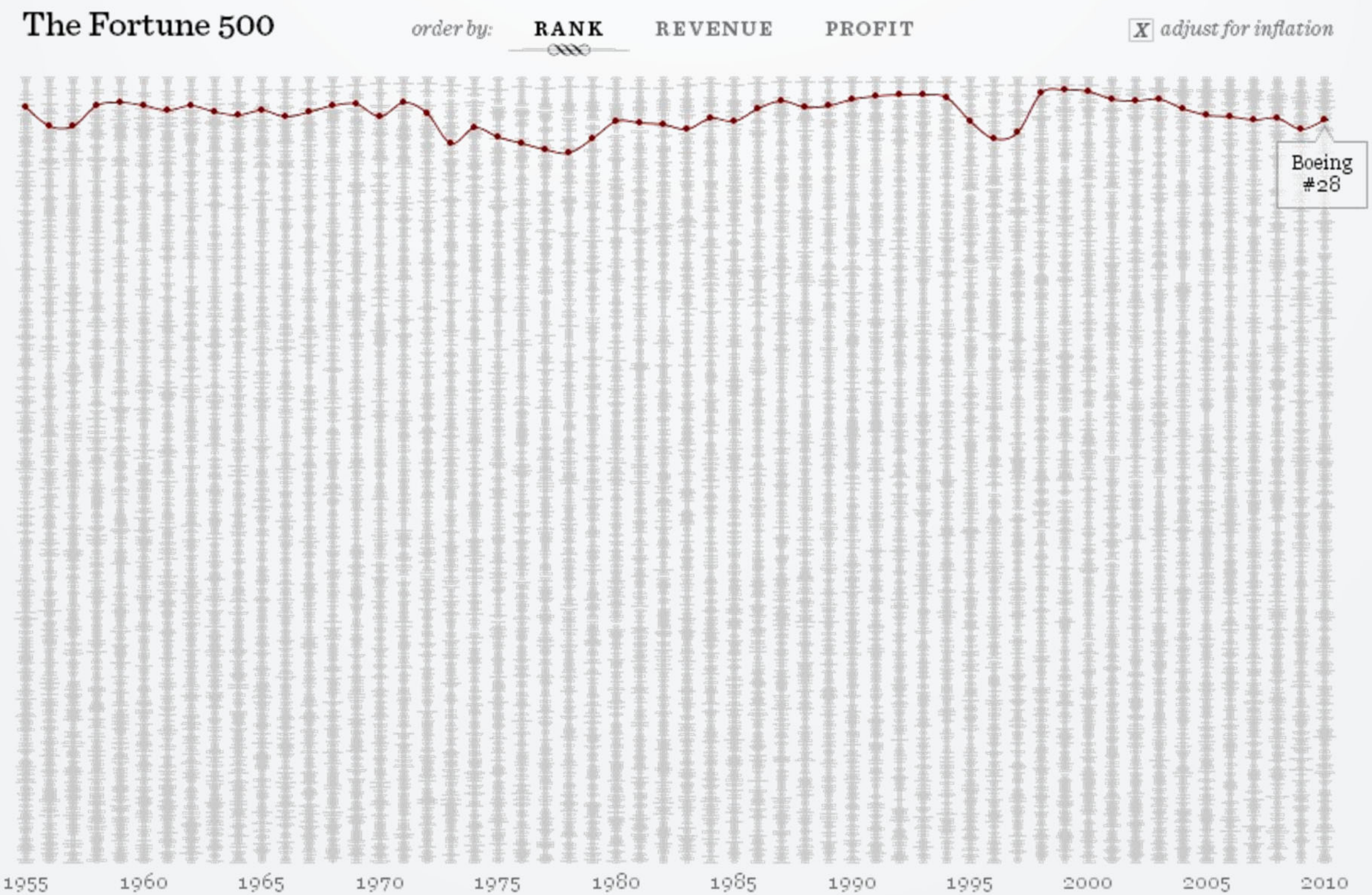


J. Heinrich and D. Weiskopf. “Continuous Parallel Coordinates.”
IEEE Transactions on Visualization and Computer Graphics (2009).

Salaries vs Performance of baseball teams in USA

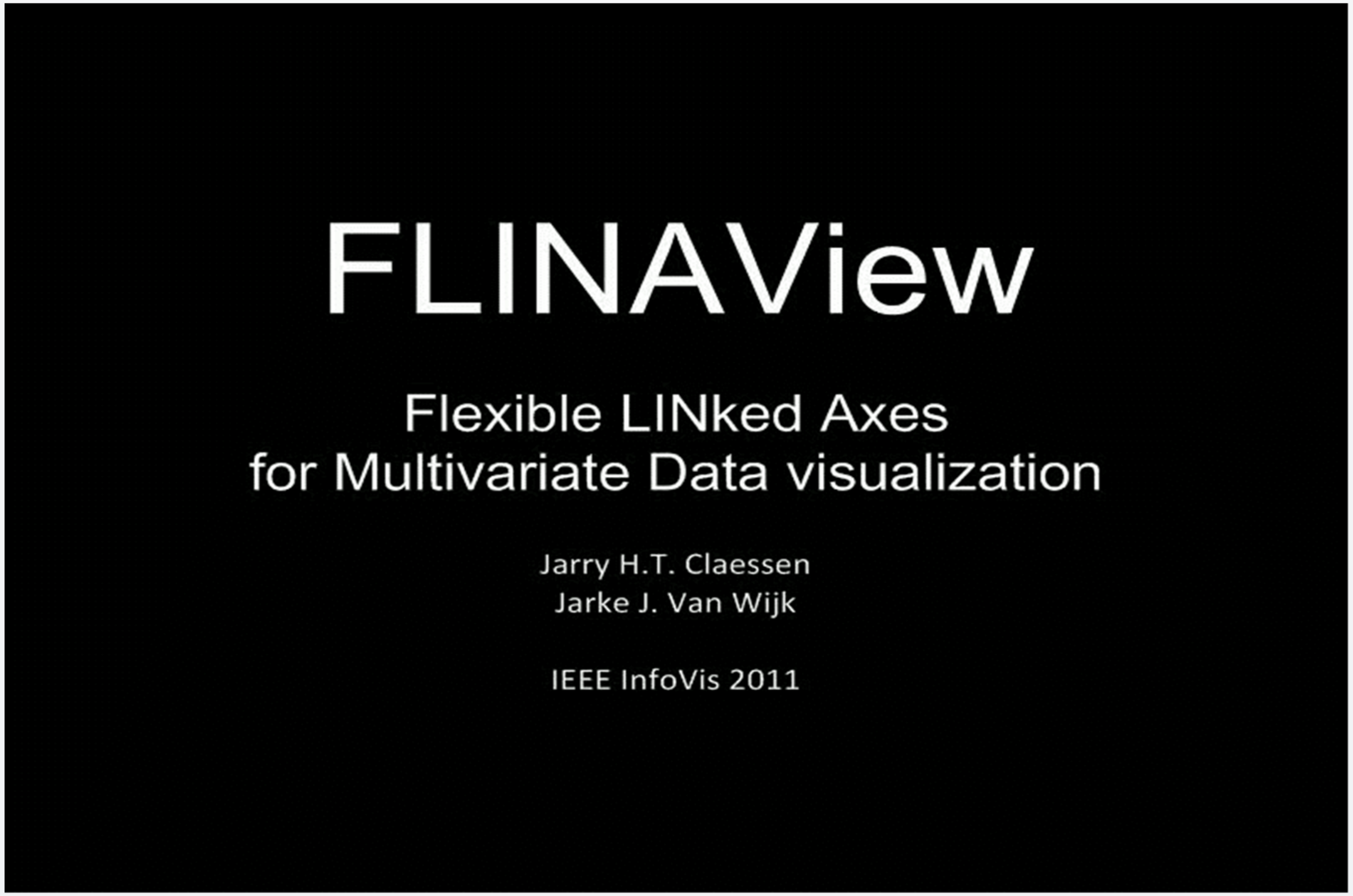


https://fathom.info/salaryper/



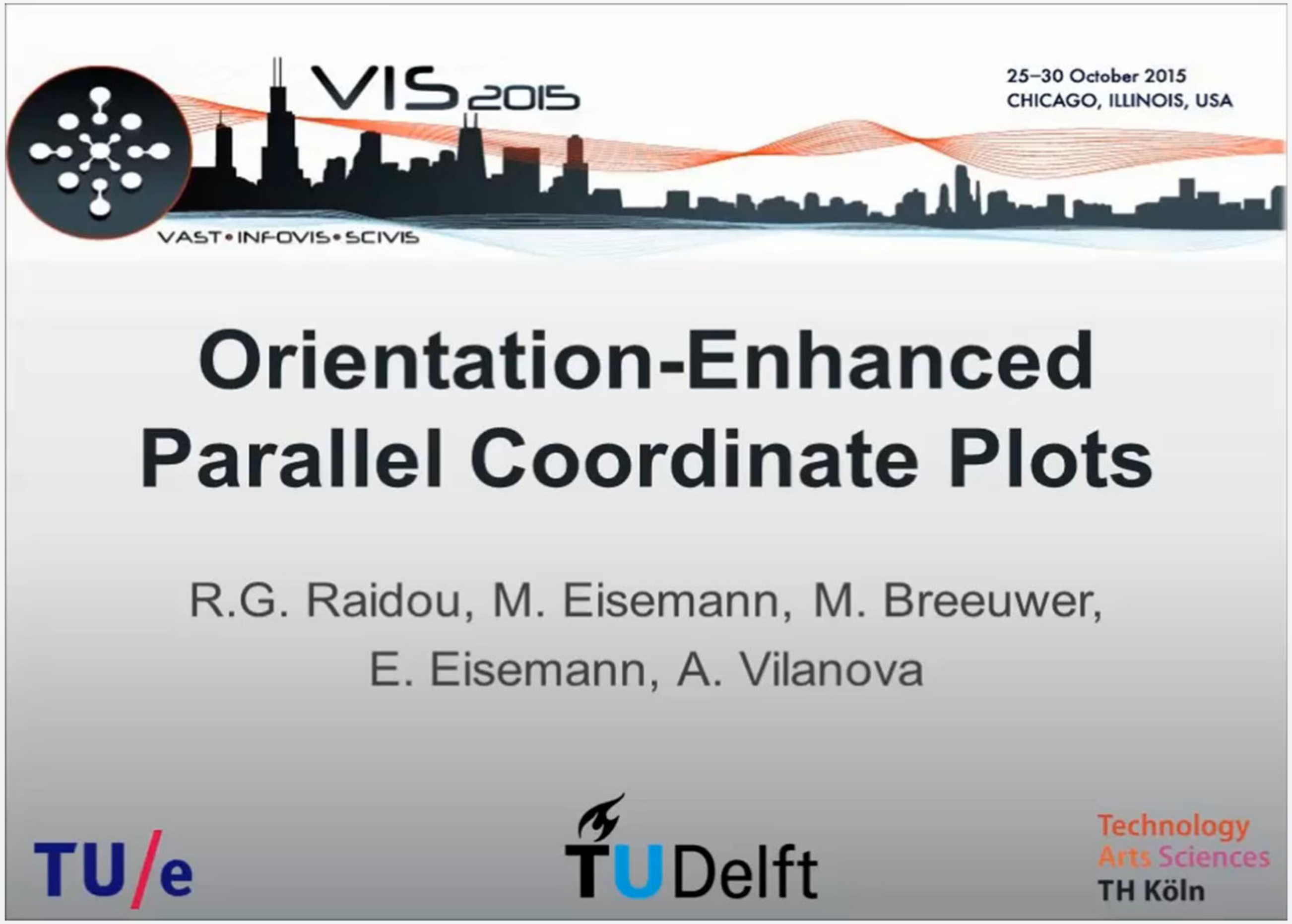
World Top 500 Enterprises Ranking

FLINAView



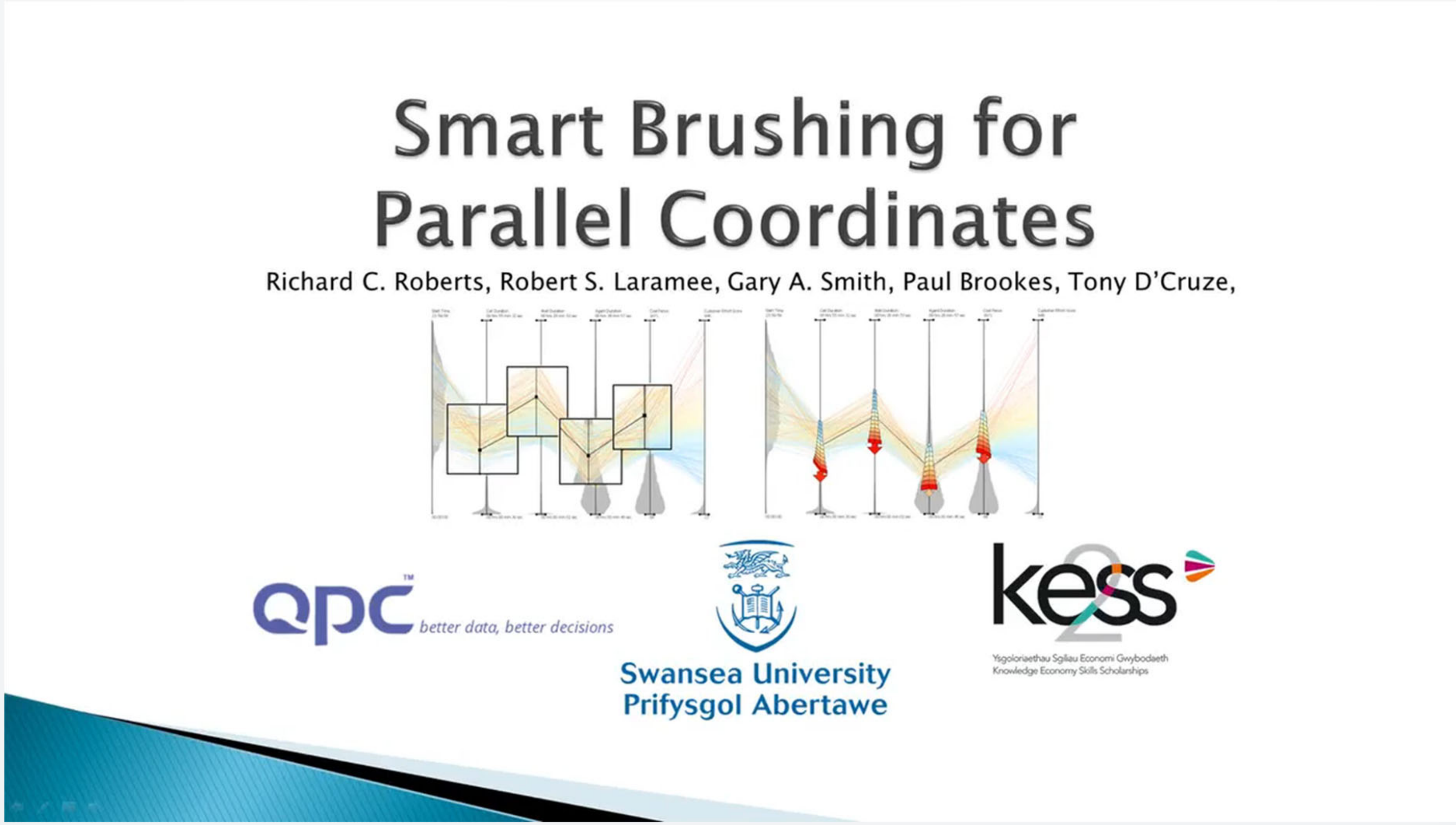
Claessen J H T, Van Wijk J J. Flexible linked axes for multivariate data visualization[J].
IEEE Transactions on Visualization and Computer Graphics (2011)

Orientation-Enhanced Parallel Coordinate Plots



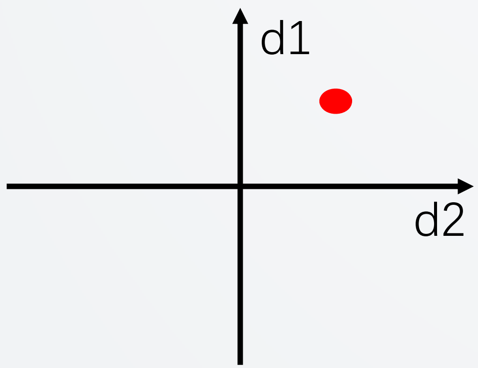
Raidou R G, Eisemann M, Breeuwer M, et al. Orientation-enhanced parallel coordinate plots[J].
IEEE transactions on visualization and computer graphics (2016)

Smart Brushing for Parallel Coordinates

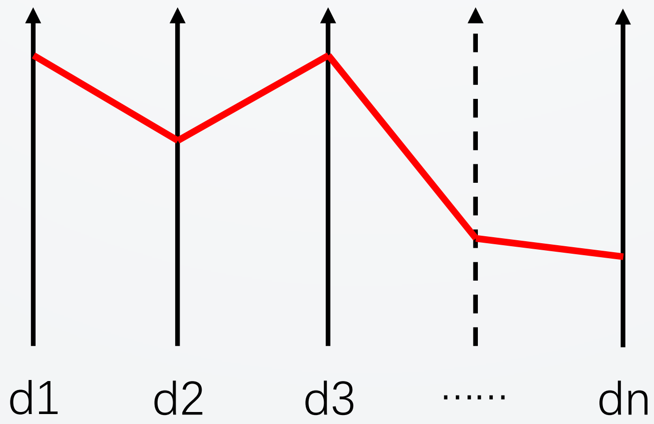


Roberts R, Laramee R S, Smith G A, et al. Smart Brushing for Parallel Coordinates[J]. IEEE Transactions on Visualization and Computer Graphics (2018)

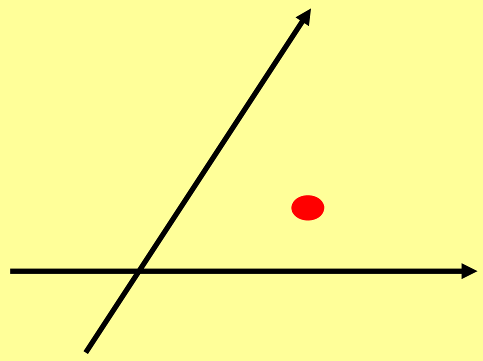
Coordinate Systems



Scatter Plots



project into the space defined by parallel axes



Projection based approaches

Dimensionality Reduction

- Project the high-dimensional data onto a lower-dimensional subspace using linear or non-linear transformations.
- Projection preserves important relations (e.g., no information loss, data discrimination).

$$x = \begin{pmatrix} a_1 \\ a_2 \\ \cdots \\ a_N \end{pmatrix} \rightarrow reduce\ dimensionality \rightarrow \hat{x} = \begin{pmatrix} b_1 \\ b_2 \\ \cdots \\ b_K \end{pmatrix} (K \ll N)$$

Methods

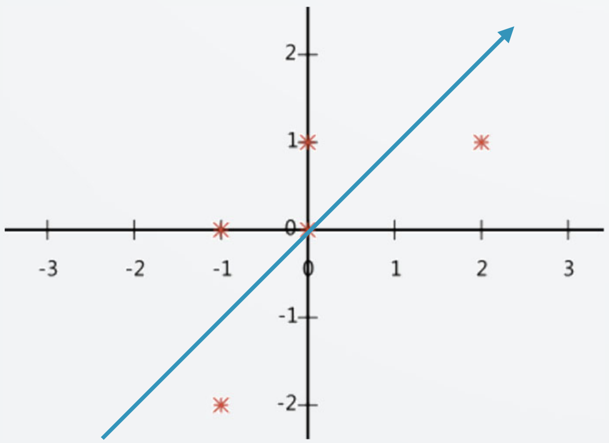
- Linear methods:
  - Principal Component Analysis (PCA).
  - Multidimensional Scaling (MDS).
- Nonlinear methods:
  - * ISOMAP.
  - * Local Linear Embedding (LLE).

PCA Motivation I

- Data set has two dimensions.

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix} \xrightarrow{\text{Subtract the average}} \begin{pmatrix} -1 & -1 & 0 & 2 & 0 \\ -2 & 0 & 0 & 1 & 1 \end{pmatrix}$$

- Projections should spread as much as possible.



How can we use one dimension to represent the data with most information preserved?

Variance

- Variance represents the spread of data items.

$$\frac{1}{m}\sum_{i=1}^{m}(a_i - \bar{a})^2$$

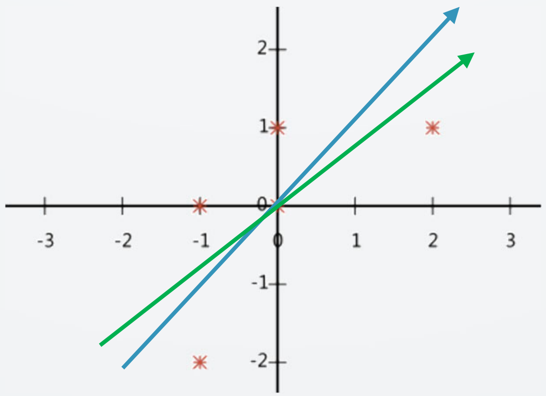
- Let the average in each dimension be 0, i.e., $\bar{a}$=0.

$$\frac{1}{m}\sum_{i=1}^{m}(a_i)^2$$

- Question： How to find one coordinate (projection), such that the data items are projected to the coordinate with the maximum variance?

PCA Motivation II

- How to choose more coordinates?
  - Shall we consider only the variance?
    - Coordinates may overlap.
- Coordinates should be linearly uncorrelated to preserve more information.
  - Correlations mean two dimensions are dependent.



Covariance

- The correlation between dimensions a and b can be represented by their covariance:

$$\frac{1}{m}\sum_{i=1}^{m}(a_i-\bar{a})(b_i-\bar{b})^T$$

- We make $\bar{a}=0,\ \ \bar{b}=0$, so we have $\frac{1}{m}\sum_{i=1}^{m}a_i b_i$
- Covariance $=\ 0$ means a and b are uncorrelated.
  - The second coordinate must be orthogonal to the first one.
  - The two projection coordinates must be orthogonal.

Covariance Matrix

- Given two dimensions a and b, we have matrix X:

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_m \\ b_1 & b_2 & \cdots & b_m \end{pmatrix}$$

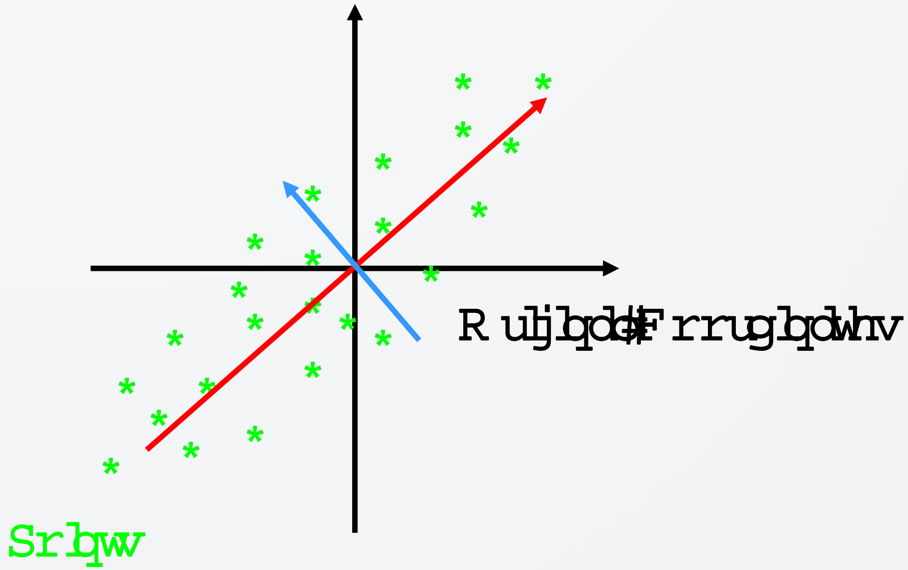
- The covariance matrix can be obtained by

$$S = \frac{1}{m} X X^T = \frac{1}{m} \sum_{i=1}^{m} X_i X_i^T = \begin{pmatrix} \frac{1}{m} \sum_{i=1}^{m} a_i^2 & \frac{1}{m} \sum_{i=1}^{m} a_i b_i \\ \frac{1}{m} \sum_{i=1}^{m} a_i b_i & \frac{1}{m} \sum_{i=1}^{m} b_i^2 \end{pmatrix}$$

Principal Component Analysis (PCA)

- Maximize the variance in each projected dimension.
- Minimize the covariance in each pair of projected dimensions.

X : n x m matrix.
n dimensions, m point.
X_i : the i_{th} data item.



Math Derivation (1/2)

1. Compute central point:

$$\bar{X} = \frac{1}{m}\sum_{i=1}^{m} X_i$$

X : *n* x *m* matrix.
n dimensions, *m* points.
X_i: the i_{th} data item.

2. Subtract the central point:

$$X_i = X_i - \bar{X}$$

3. Compute variance after projection:

$$\frac{1}{m}\sum_{i=1}^{m}(PX_i)^2 = \frac{1}{m}\sum_{i=1}^{m} PX_iX_i^{\ T}P^T$$

$\rightarrow$ PSP^T

$$S = \frac{1}{m}\sum_{i=1}^{m} X_iX_i^{\ T}$$ Co-variance

Math Derivation (2/2)

4. Variance after projection PSP^T

5. Maximize PSP^T, while satisfying $PP^T = 1$

6. Apply lagrangian multiplier method $PSP^T + \lambda(1 - PP^T)$

7. Take its derivative and make it ZERO, we get $SP^T = \lambda P^T$

Eigenvectors

$$SP^T = \lambda P^T$$

That is

$$\boxed{PSP^T} = \lambda$$

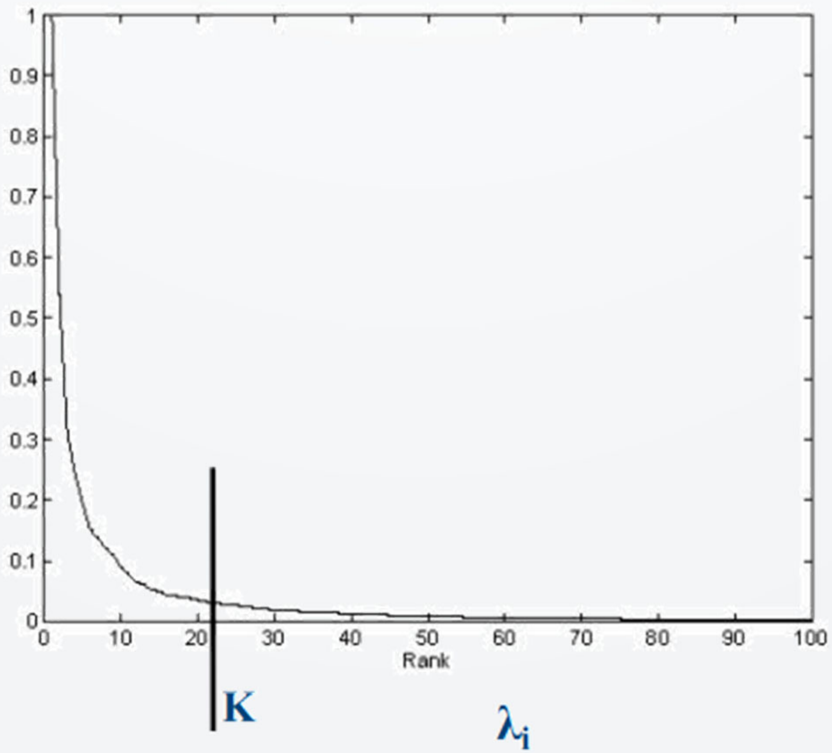
- The variance after projection is the eigenvalue of the covariance matrix.
  - To maximize the variance, choose the largest eigenvalue.
  - The largest eigenvalue is the best coordinates.

PCA Mechanics

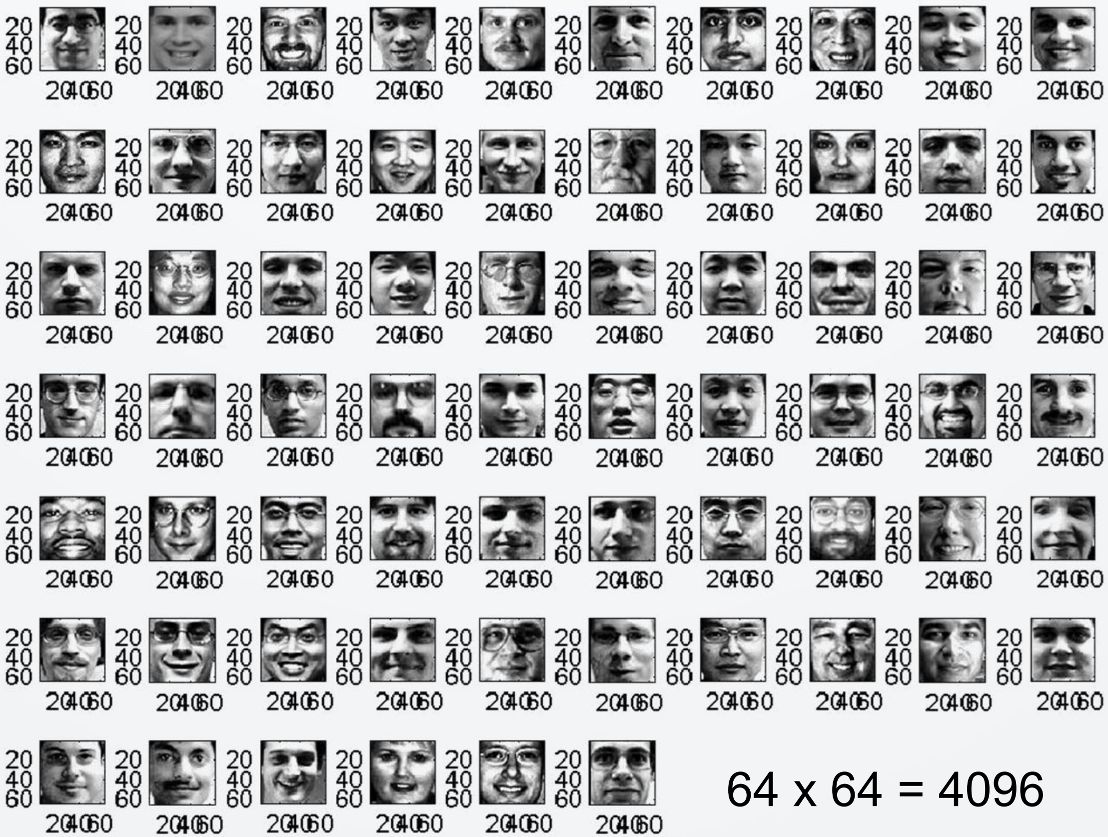
Suppose $x_1, x_2, \ldots, x_M$ are $H \times 1$ vectors:

1. $\bar{x} = \frac{1}{M}\sum_{i=1}^{M} x_i$.
2. Subtract the mean $\Phi_i = x_i - \bar{x}$.
3. Form $H \times M$ matrix $A = [\Phi_1 \Phi_2 \cdots \Phi_\mathrm{M}]$.
4. Compute covariance matrix $C = \frac{1}{M}\sum_{i=1}^{M} \Phi_n \Phi_n^T = AA^T$.
5. Compute eigenvalues of C: $\lambda_1 > \lambda_2 > \cdots > \lambda_N$.
6. Compute eigenvectors of C: $u_1, u_2, \cdots, u_N$.

Eigenvalue Spectrum



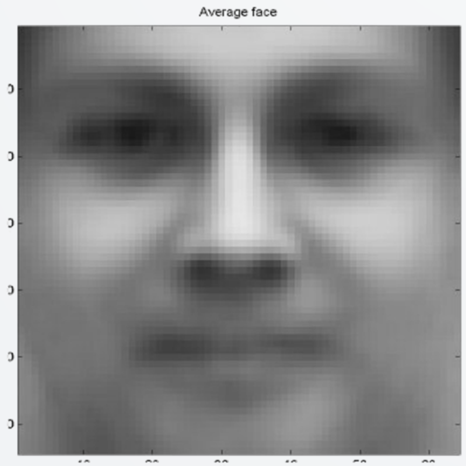
PCA Applied to Faces

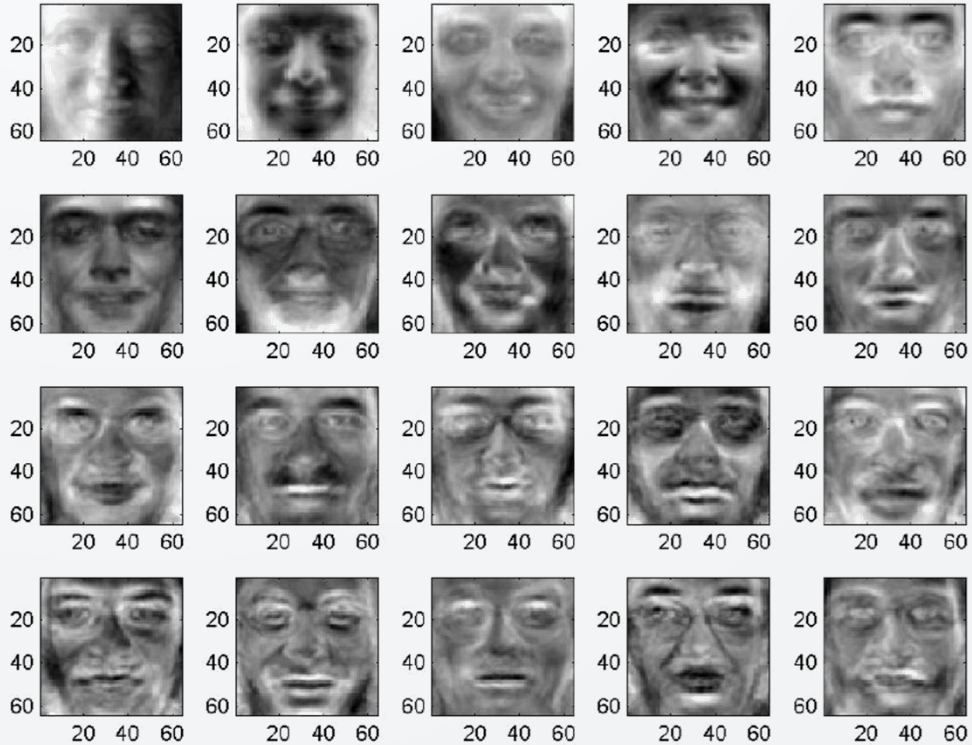


64 x 64 = 4096

PCA Applied to Faces

Reconstruct each face as a linear combination of "basis faces", or Eigenfaces.

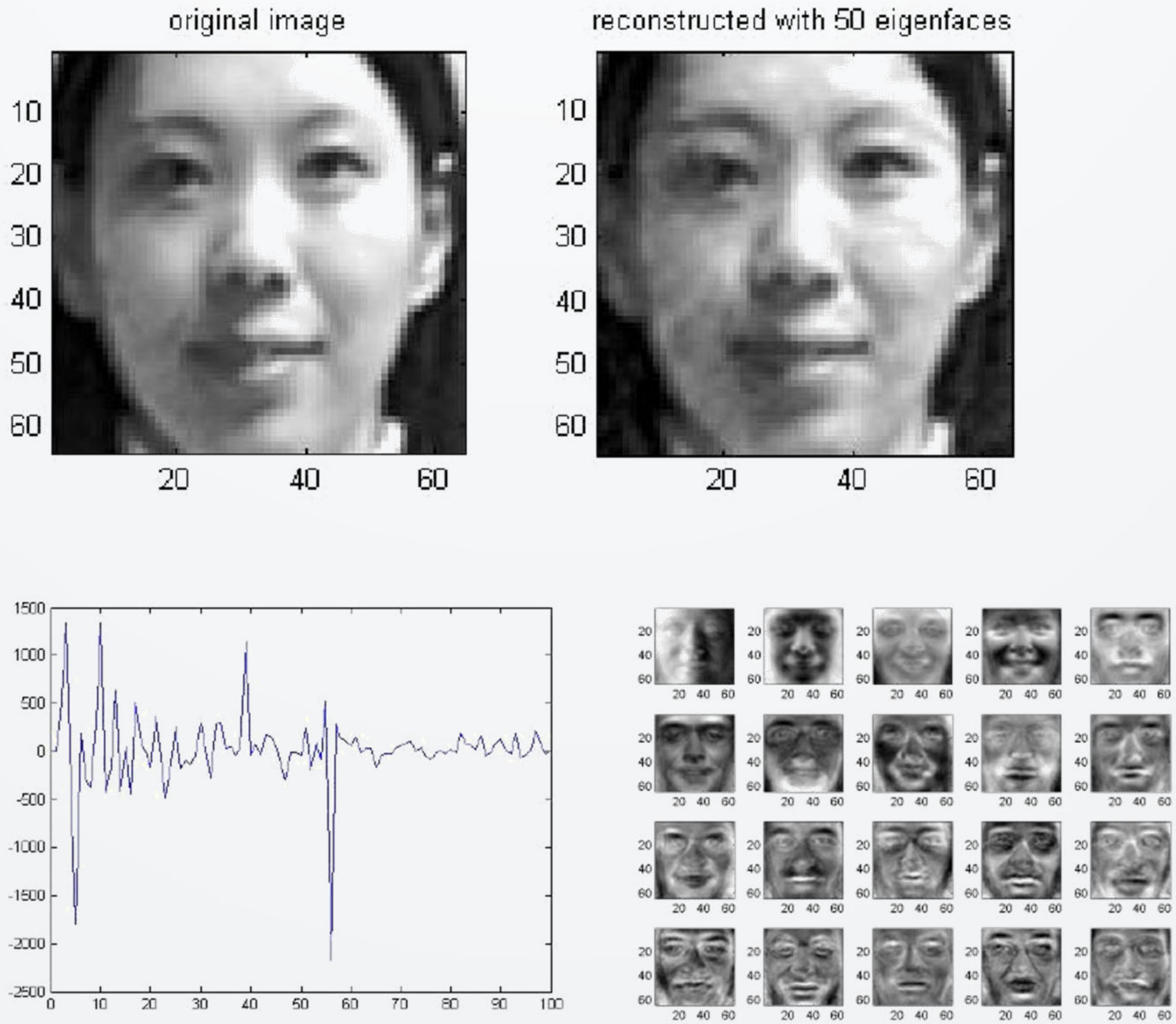




Eigenfaces

Reconstruction

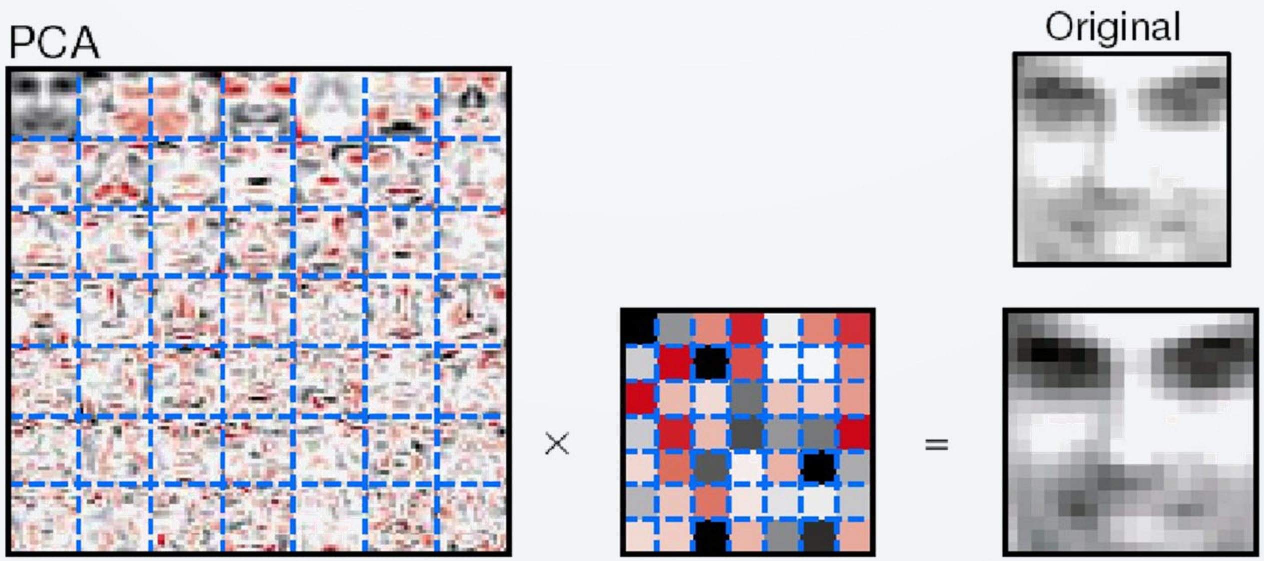
- 90% variance is captured by the first 50 eigenvectors.
- Reconstruct existing faces using only 50 basis images.



original image
reconstructed with 50 eigenfaces

Issues

- PCA involves adding up some basis images and subtracting others.
- The basis images are not physically intuitive.



Multidimensional Scaling (MDS)

- Takes as input a matrix M containing pairwise distances between P -dimensional data points.
- Outputs a projection of data in m-dimensional space where the pairwise distances match the original distances as faithfully as possible.

$$min \sum_{i<j} \mu_{ij}(d_p(x_i, x_j) - d_m(f(x_i), f(x_j)))^2$$

An Example: US Map

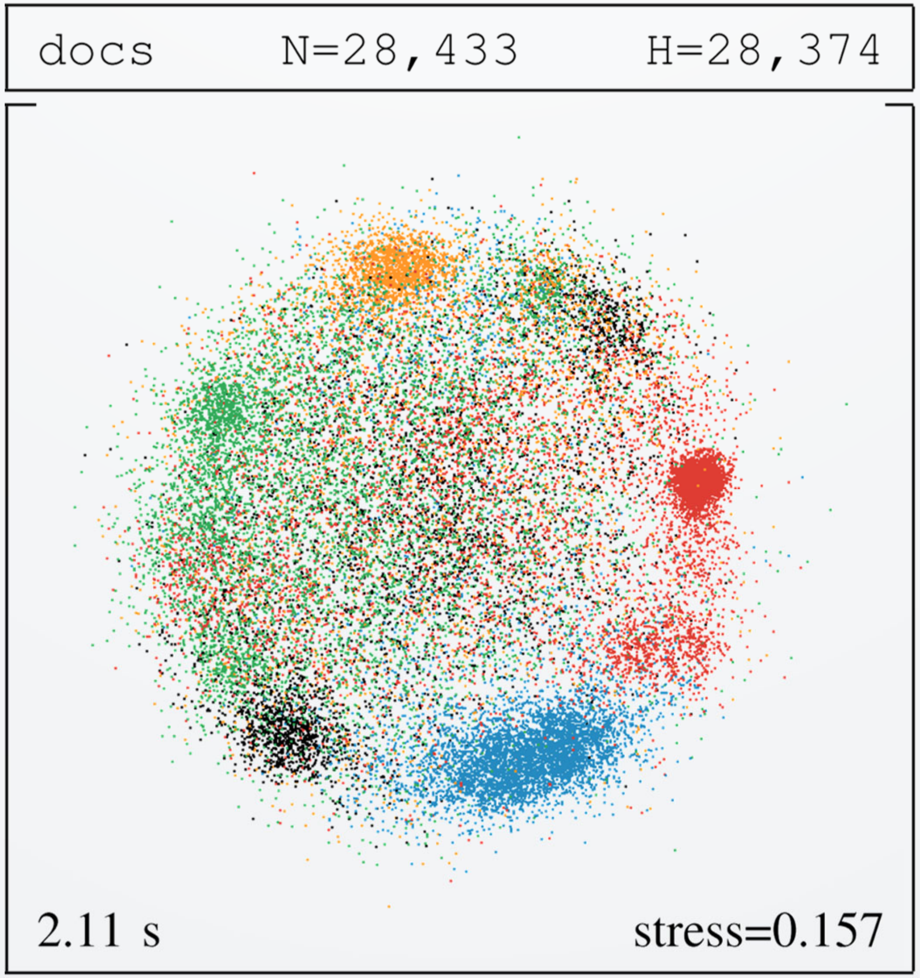
- Suppose you know the distances between a bunch of cities…

	Chicago	Raleigh	Boston	Seattle	S.F.	Austin	Orlando
Chicago	0						
Raleigh	641	0					
Boston	851	608	0				
Seattle	1733	2363	2488	0			
S.F.	1855	2406	2696	684	0		
Austin	972	1167	1691	1764	1495	0	
Orlando	994	520	1105	2565	2458	1015	0

Result of MDS

Actual Plot of Cities

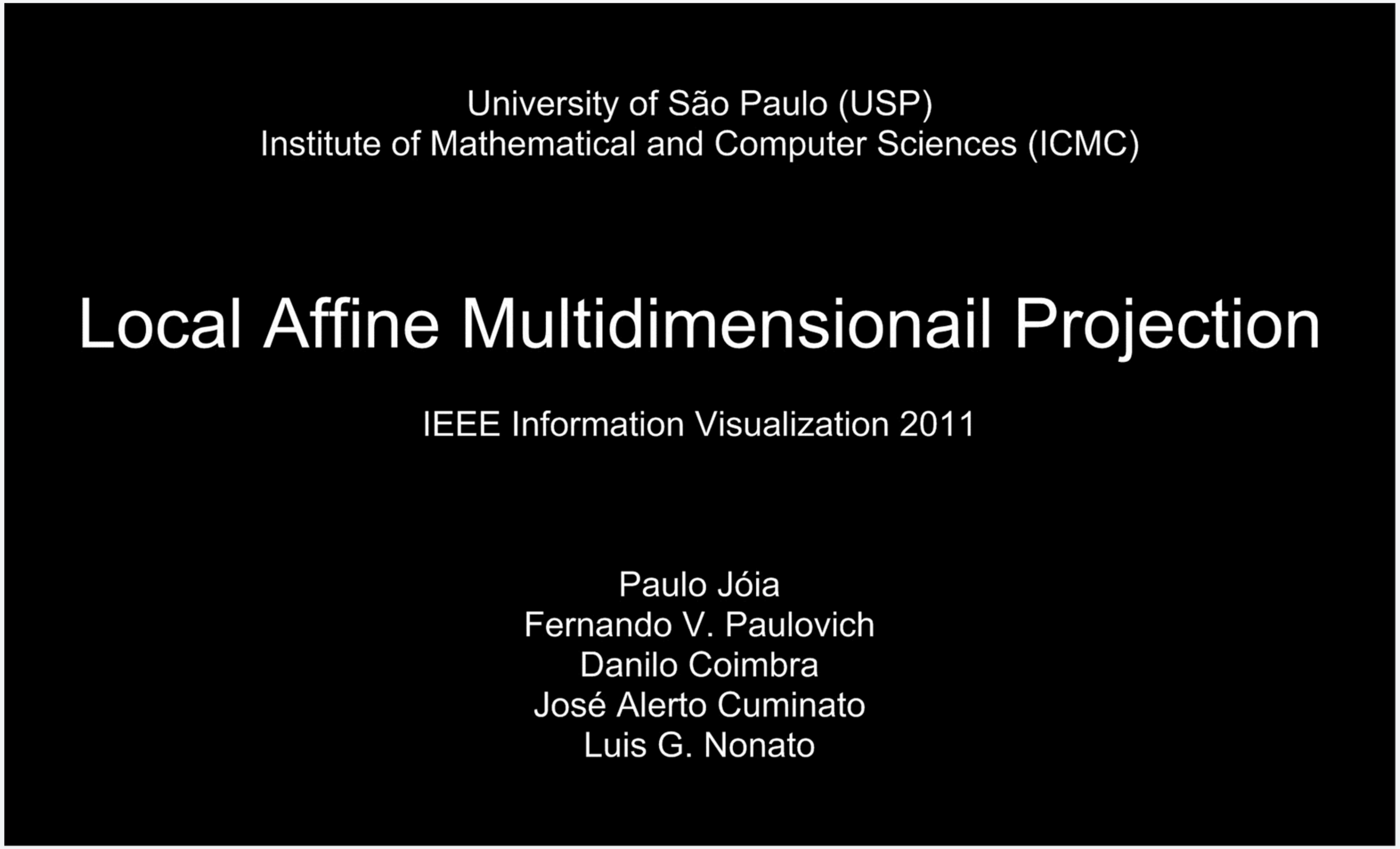
Docs Dataset



Result of MDS on docs dataset by GlimmerJS

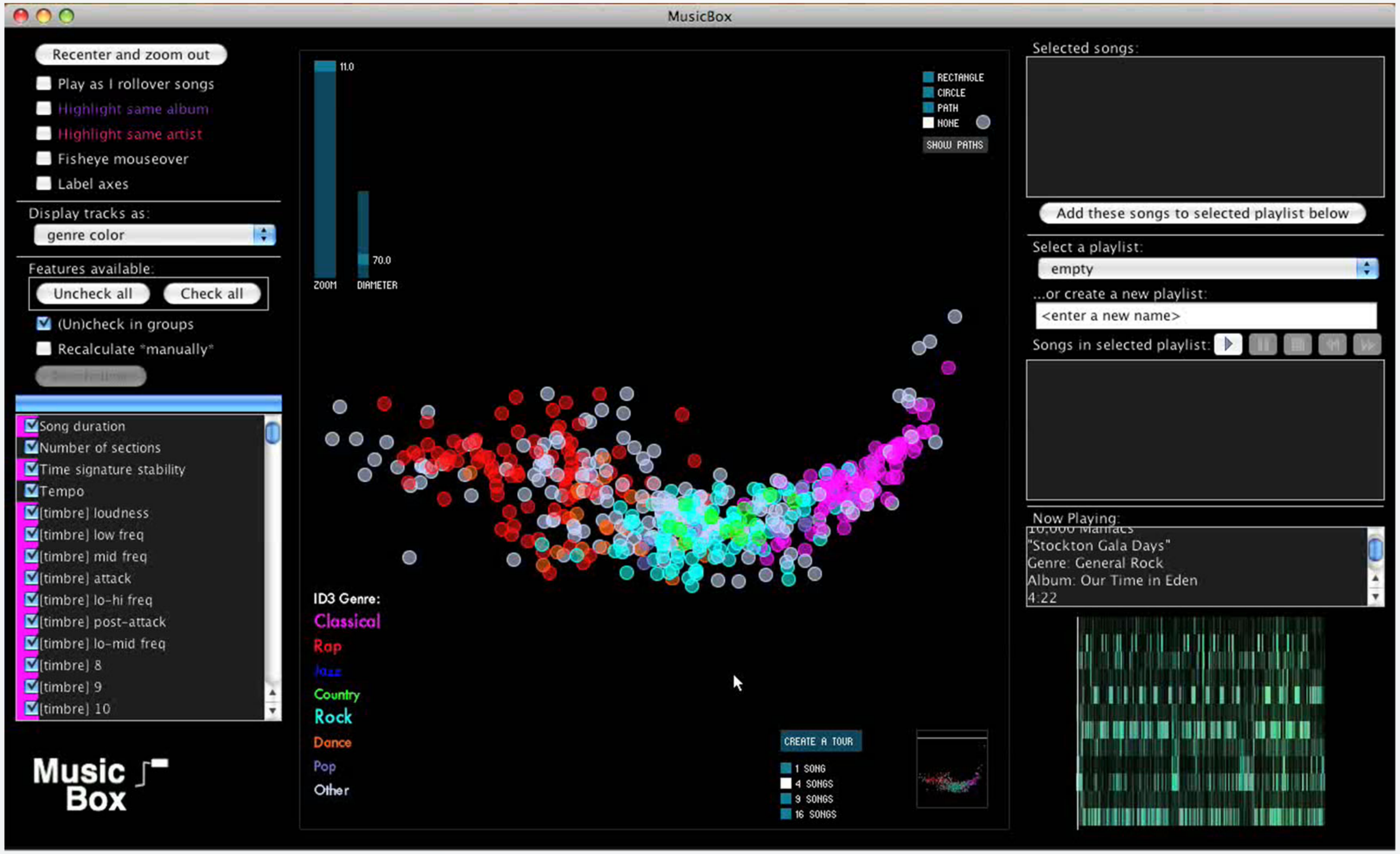
Ingram S, Munzner T, Olano M. Glimmer: Multilevel MDS on the GPU.
IEEE Transactions on Visualization and Computer Graphics

Local Affine Multidimensional Projection

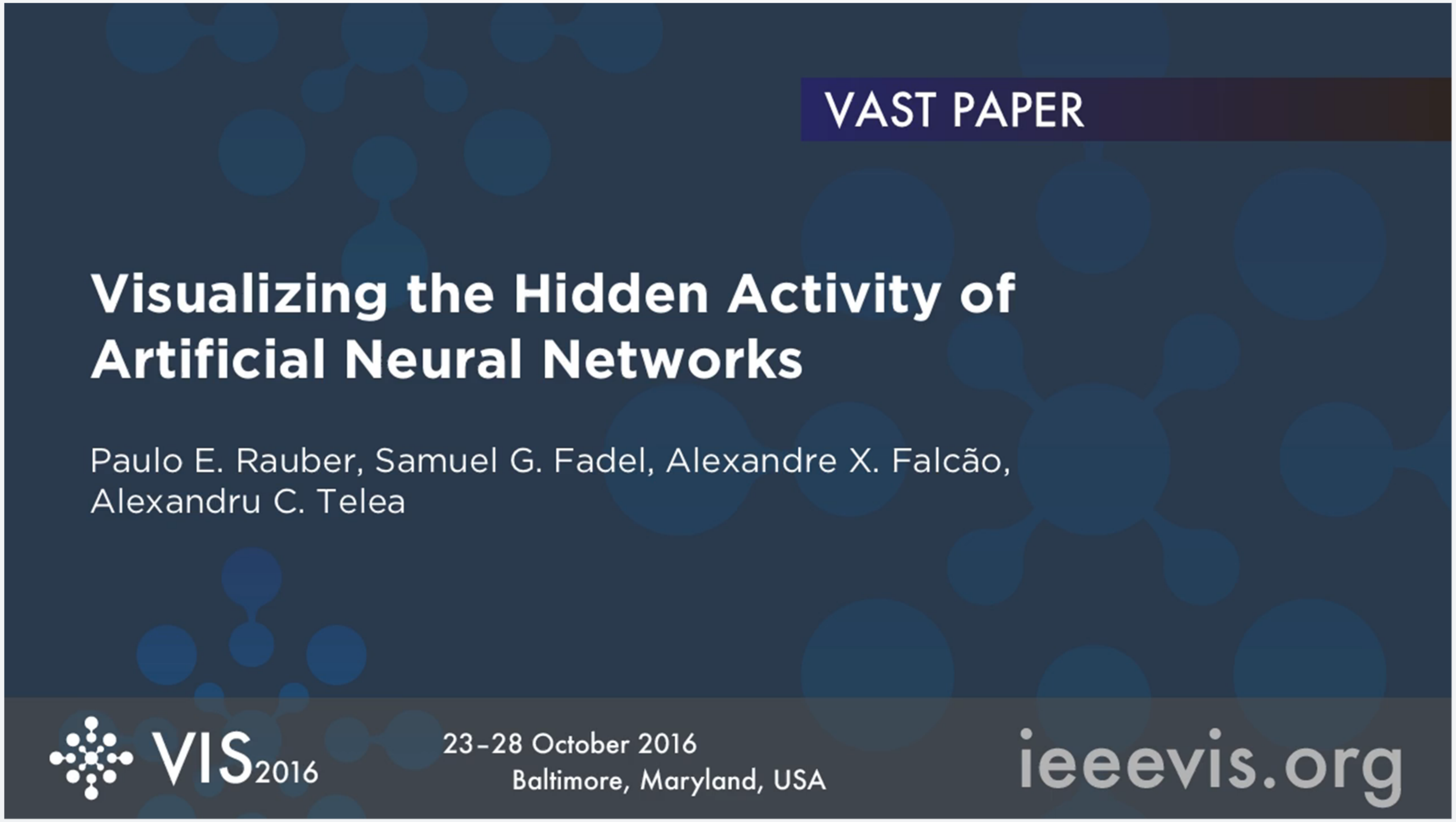


Joia P, Coimbra D, Cuminato J A, et al. “Local affine multidimensional projection”.

MusicBox


MusicBox
Recenter and zoom out
Play as I rollover songs
Highlight same album
Highlight same artist
Fisheye mouseover
Label axes
Display tracks as:
genre color
Features available:
Uncheck all
Check all
(Un)check in groups
Recalculate "manually"
Song duration
Number of sections
Time signature stability
Tempo
[timbre] loudness
[timbre] low freq
[timbre] mid freq
[timbre] attack
[timbre] lo-hi freq
[timbre] post-attack
[timbre] lo-mid freq
[timbre] 8
[timbre] 9
[timbre] 10
Music Box
11.0
70.0
ZOOM
DIAMETER
RECTANGLE
CIRCLE
PATH
NONE
SHOW PATHS
ID3 Genre:
Classical
Rap
Jazz
Country
Rock
Dance
Pop
Other
CREATE A TOUR
1 SONG
4 SONGS
9 SONGS
16 SONGS
Selected songs:
Add these songs to selected playlist below
Select a playlist:
empty
...or create a new playlist:
<enter a new name>
Songs in selected playlist:
Now Playing:
"Stockton Gala Days"
Genre: General Rock
Album: Our Time in Eden
4:22

Neural Networks Visualization



Rauber P E, Fadel S G, Falcao A X, et al. “Visualizing the hidden activity of artificial neural networks”. IEEE transactions on visualization and computer graphics (2017).

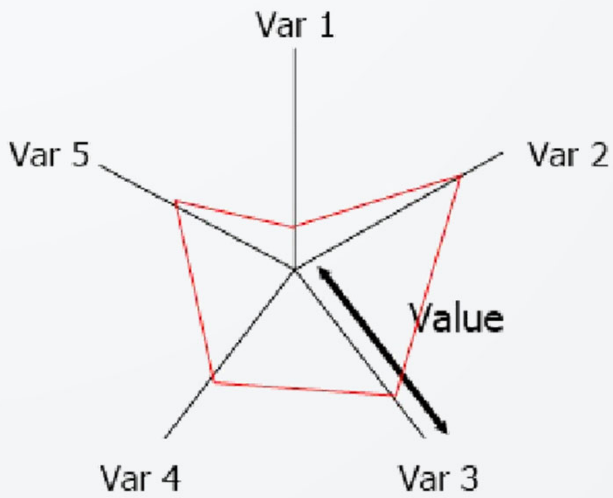
Î ị İ IHÍ Ĝ

Approaches

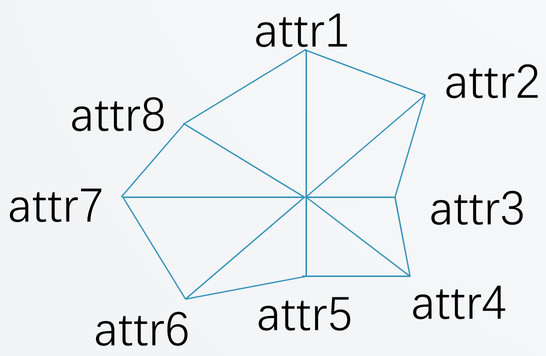
- Coordinate Systems
  - Scatter-plot Matrix.
  - Parallel Coordinates.
  - Dimensionality Reduction.
- **Glyph-based Methods.**
- Pixel Oriented Techniques
- “Small Multiples”.
- Visual Diagnosis

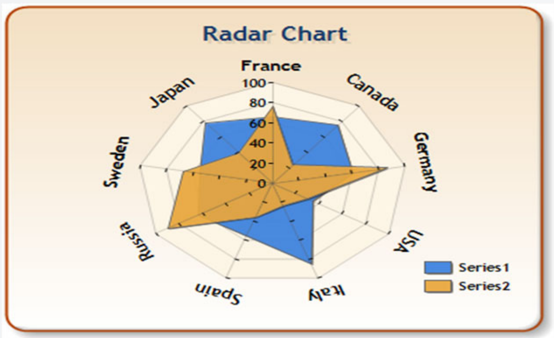
Star Plots

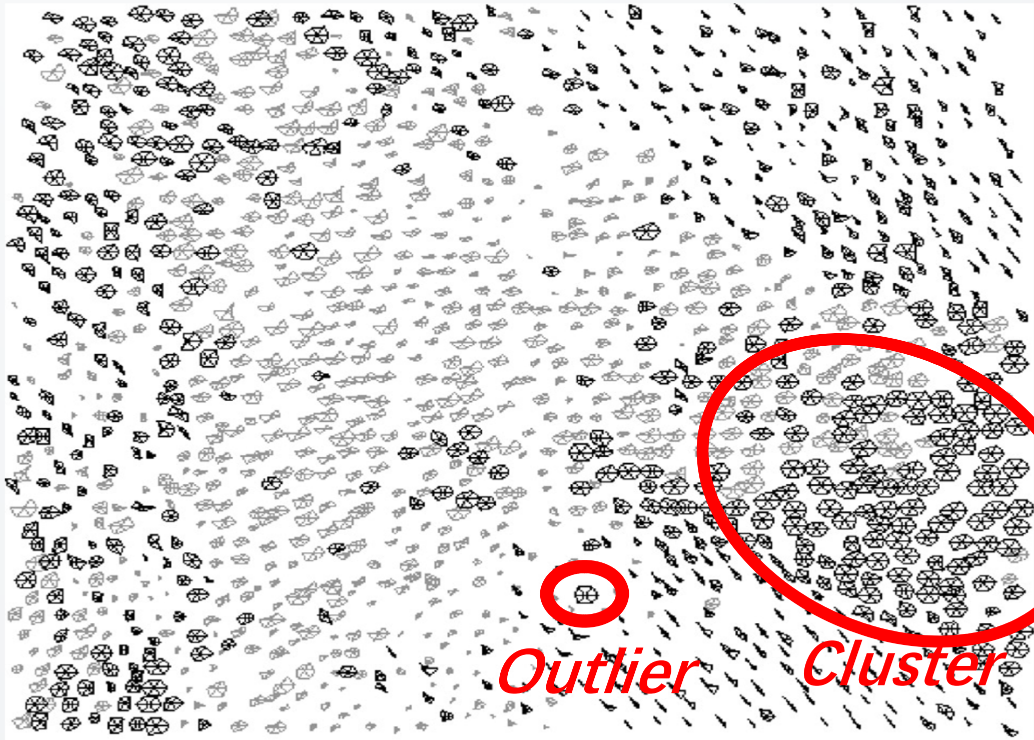
- Space variables around a circle.
- Encode values on “spokes”.
- Data point is now a shape.



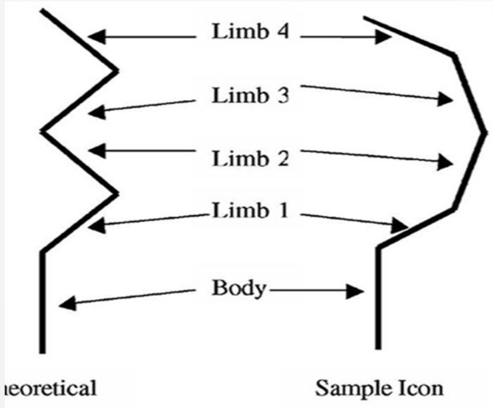
Star Glyph

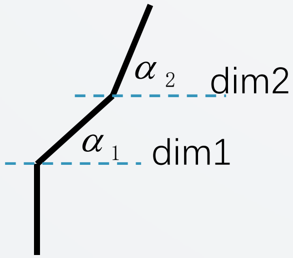

attr1
attr2
attr3
attr4
attr5
attr6
attr7
attr8


Radar Chart
France
Canada
Germany
USA
Italy
Spain
Russia
Sweden
Japan
100
80
60
40
20
0
Series1
Series2

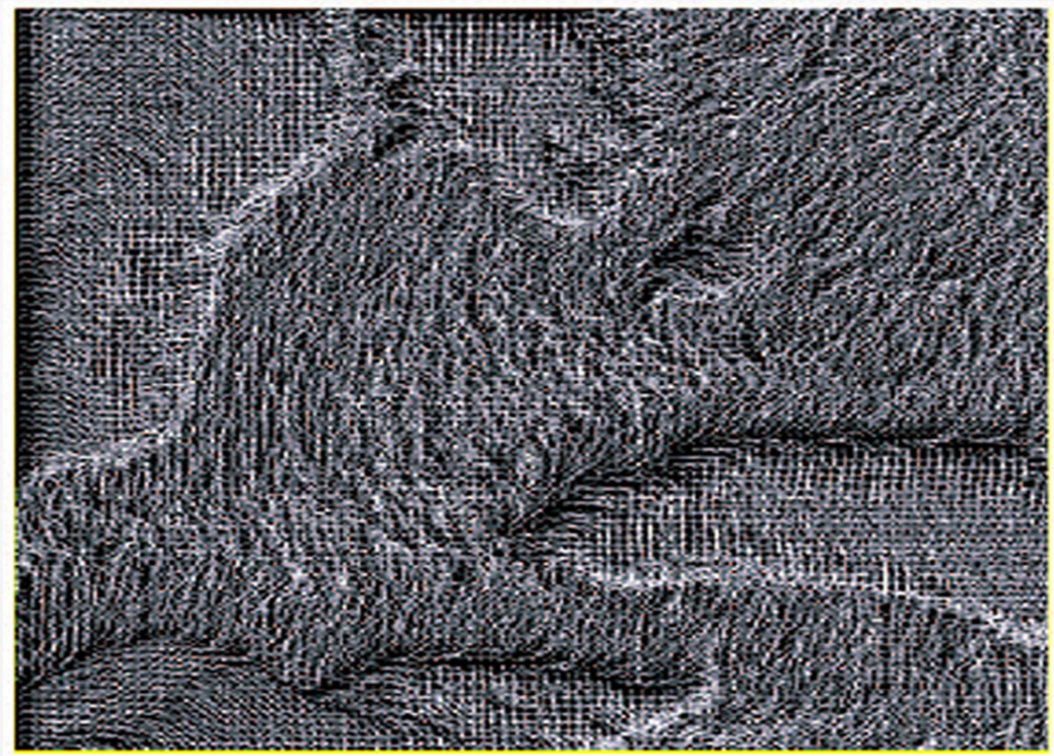

Outlier
Cluster

Stick Figure

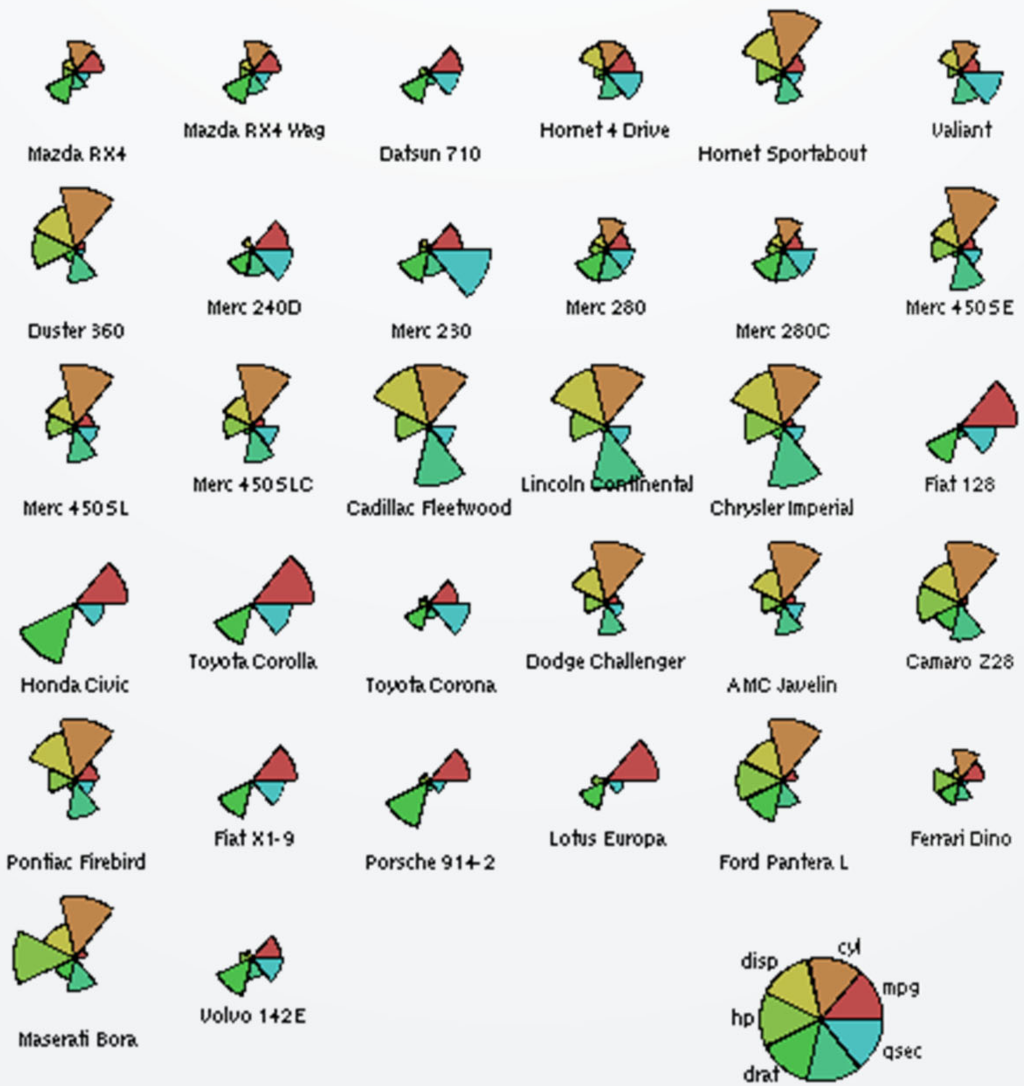




User angles of sticks
to encode dimensions

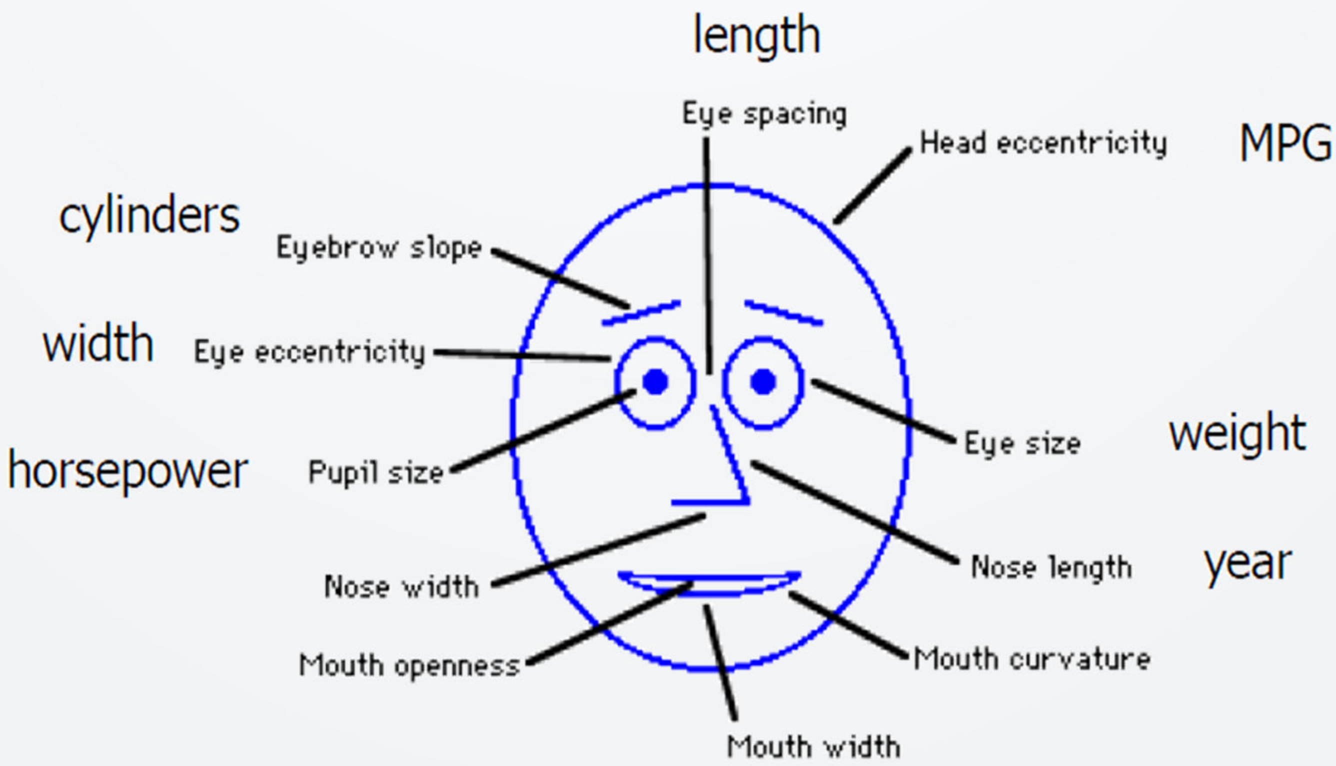


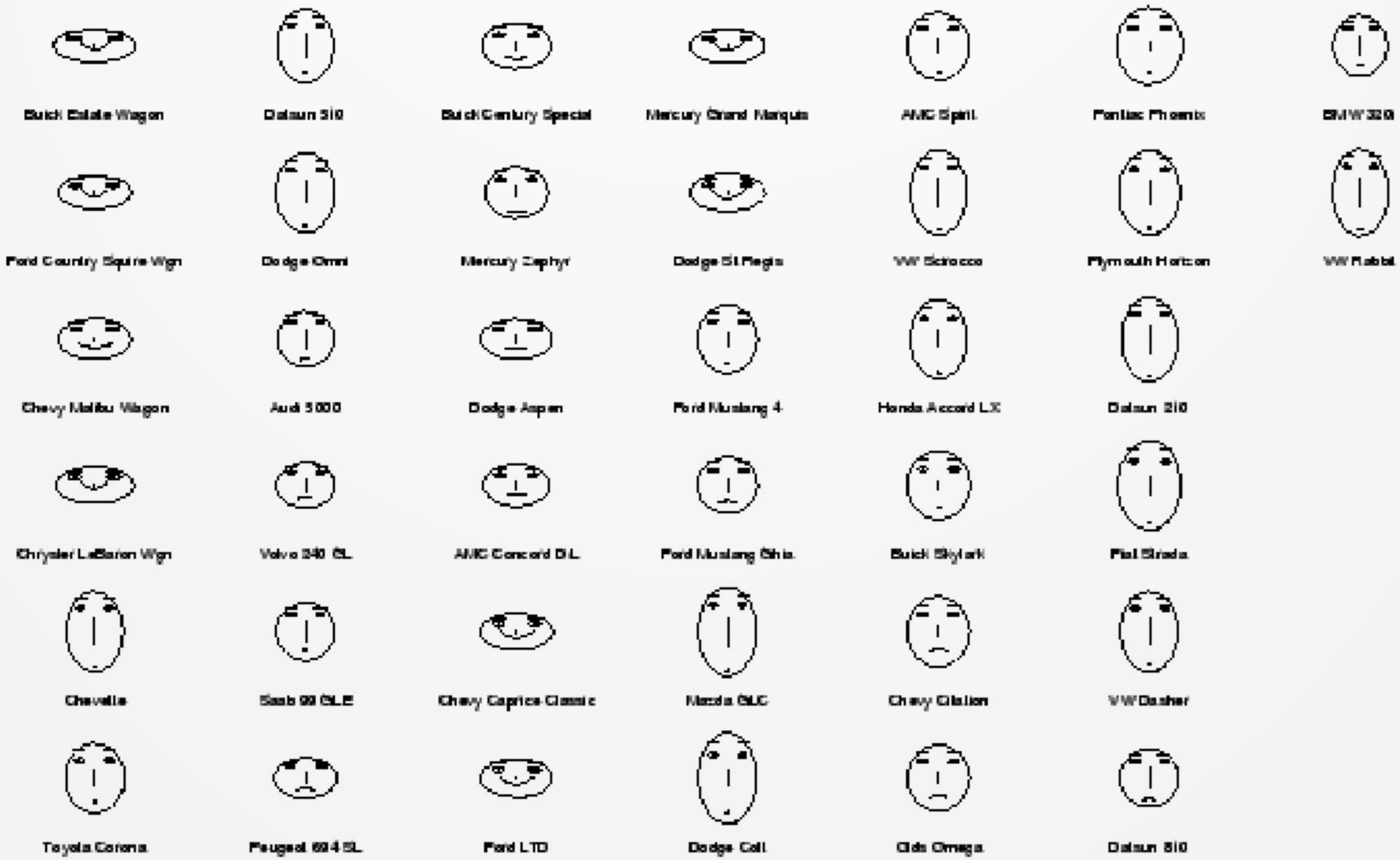
Motor Trend Cars



Chernoff Faces

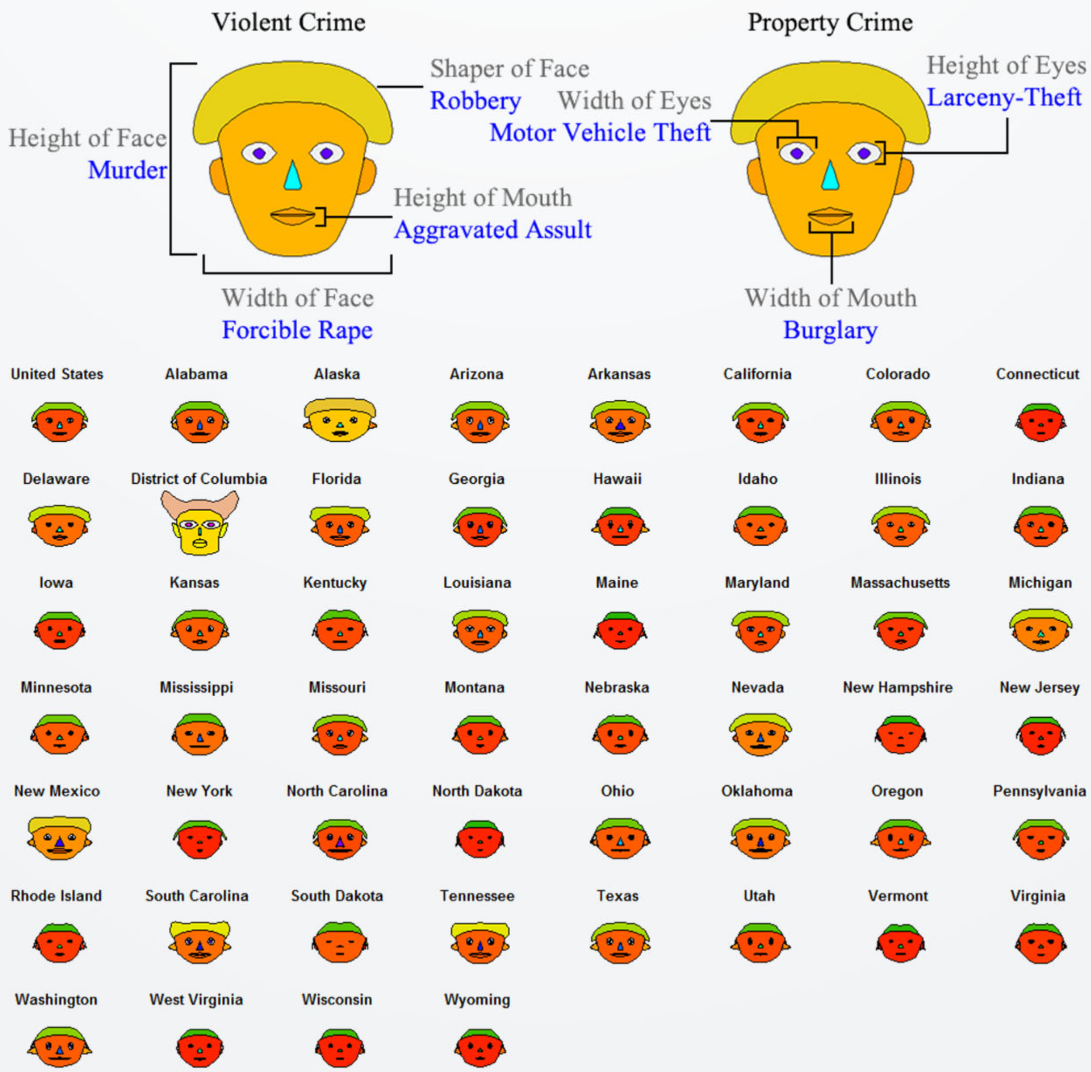
- Use face to encode different attributes.



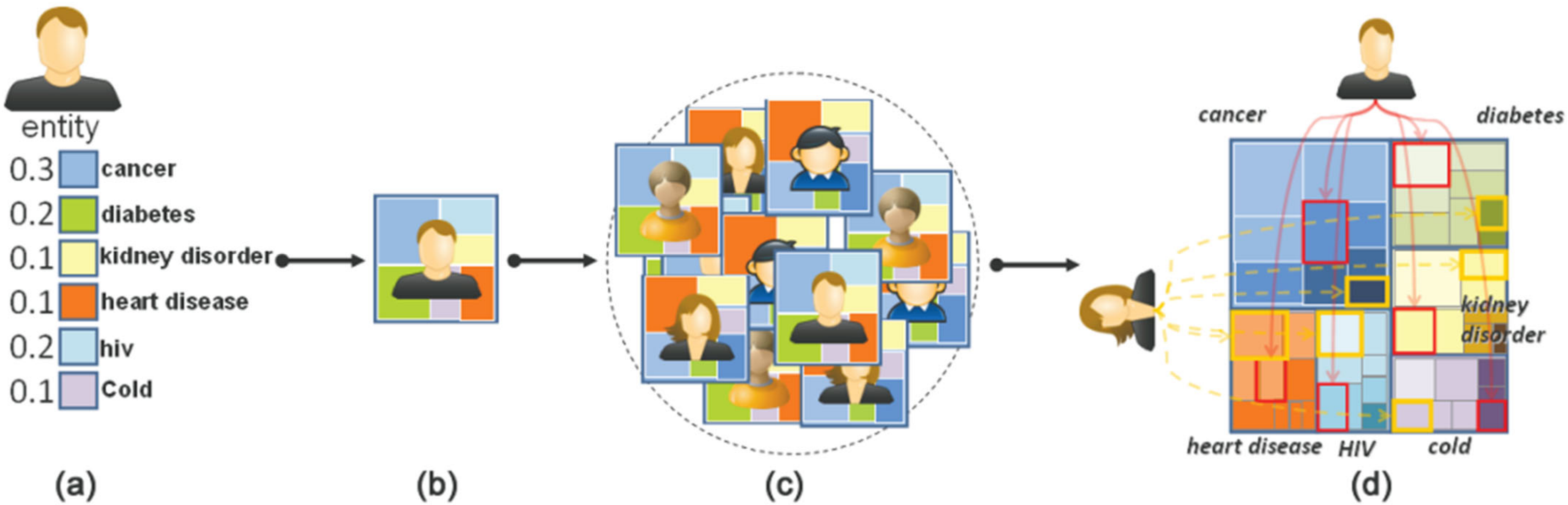


应用实例：http://hesketh.com/schampeo/projects/Faces/chernoff.html

The Face of Crime in the United States

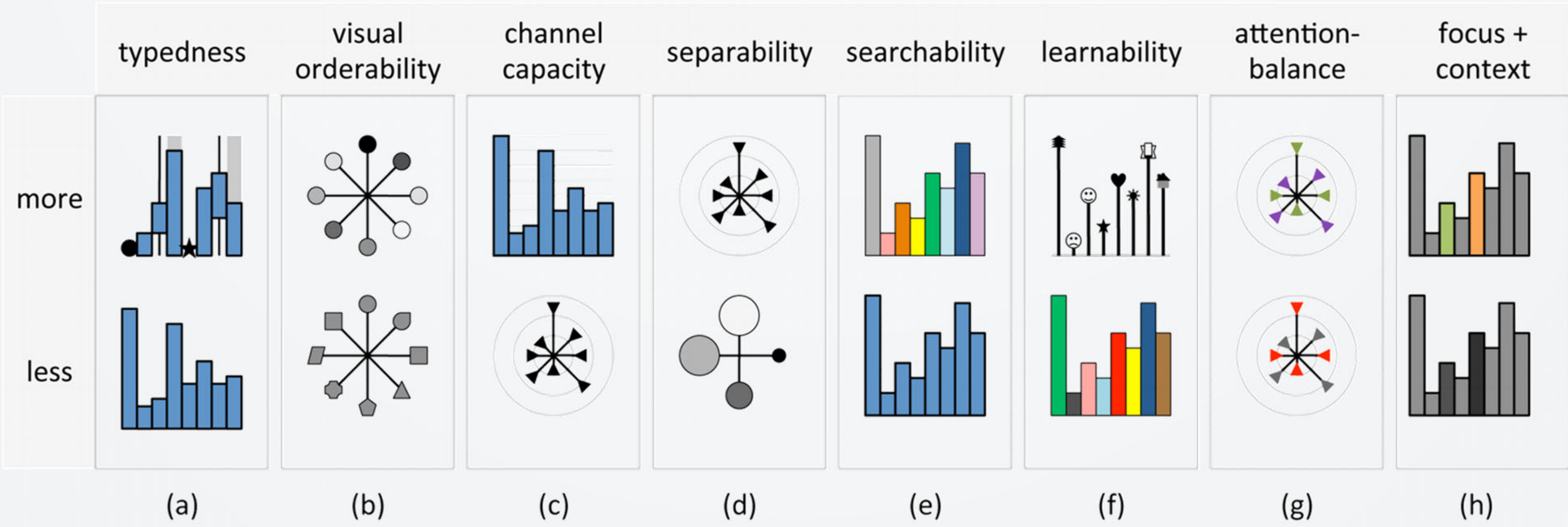


Dynamic ICON (DICON)



https://www.youtube.com/watch?v=1WyFfTENk_I&list=PLVWEtqbKu-5Vs36PA5yaSlsIHalsmVFxx&index=5

Glyph design criteria



Borgo R, Kehrer J et al. Glyph-based Visualization: Foundations, Design Guidelines, Techniques and Applications[C]
Eurographics (STARs) (2013)

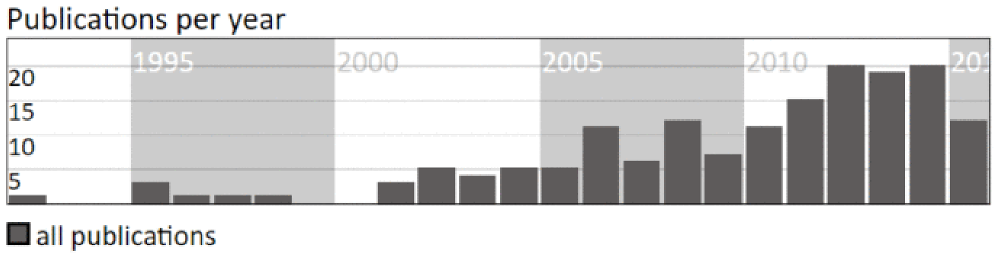
Glyph Sorting

Chung, David HS, et al. "Glyph sorting: Interactive visualization for multi-dimensional data." InfoVis (2015)

Text + Glyph

A Brief History of Dynamic Graph Visualization

The visualization of dynamic graphs is a growing research area. Starting with **first approaches in the 90s**, the field has been steadily growing to around **20 new publications per year recently**. While early publications mainly introduced new **visualization techniques**, currently a well-balanced mix of **technique**, **application**, and **evaluation** papers is published.



First, animation-based approaches showing graph evolution as **animated changes in node-link diagrams** dominated. **Since 2002**, however, alternative **timeline-based approaches** were suggested that provide an overview of the graph evolution in one view without animation. **By 2010 and later**, these techniques are even dominating the newly proposed approaches, some combining animation and timeline in **hybrid techniques**. Also, researchers **recently** explored using **adjacency matrices** instead of **node-link diagrams** to represent the individual graphs.

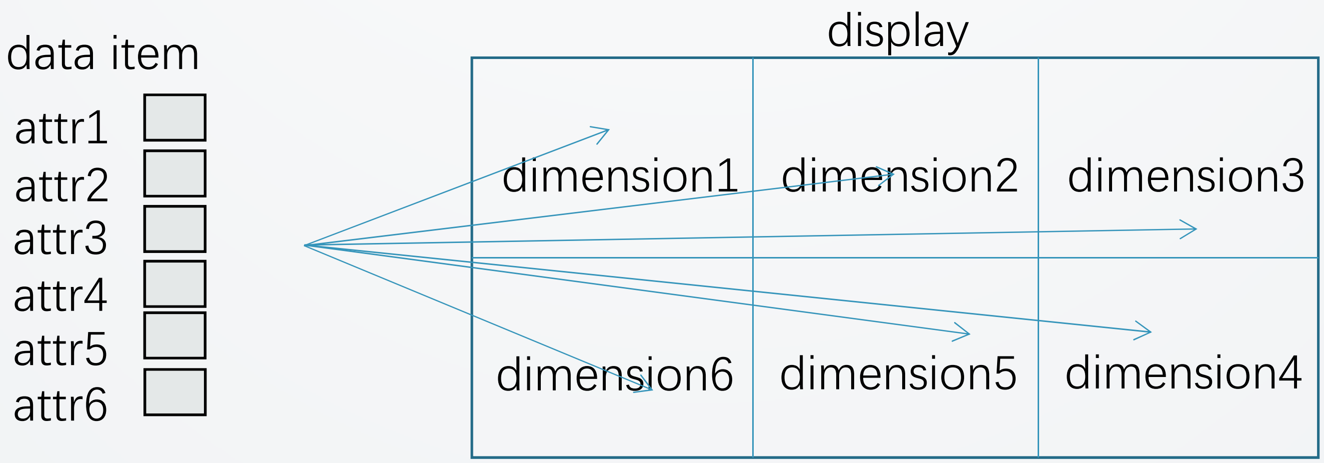
Beck F, Weiskopf D. Word-sized graphics for scientific texts[J]. IEEE transactions on visualization and computer graphics (2017).

OUTLINE

Approaches

- Coordinate Systems
  - Scatter-plot Matrix.
  - Parallel Coordinates.
  - Dimensionality Reduction.
- Glyph-based Methods.
- **Pixel Oriented Techniques**
- “Small Multiples”.
- Visual Diagnosis

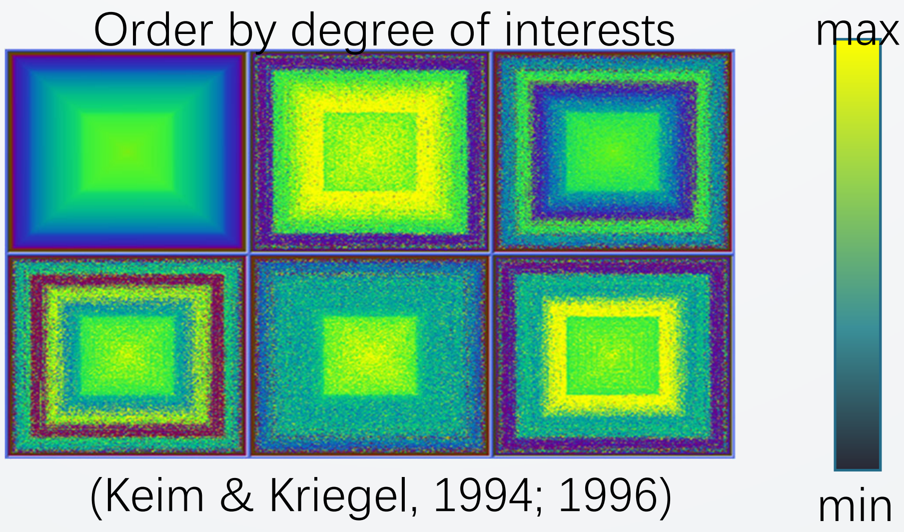
Pixel Oriented Techniques – Encoding



- A multidimensional data item contains 6 attributes
- Splitting the display into 6 regions (one for each dimension)
- An item is represented by multiple pixels in a split manner

Pixel Oriented Techniques

- Database visualization (10,000 items, 6 dimensions)

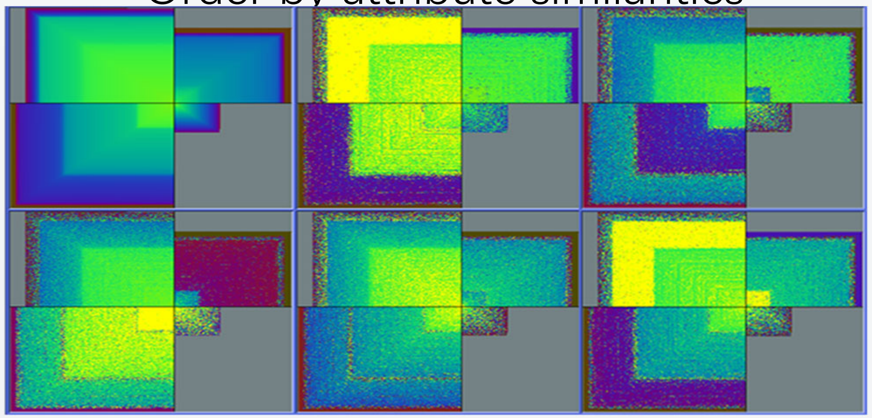


(Keim & Kriegel, 1994; 1996)

Pixel Oriented Techniques – Pixel Ordering

- Database Visualization

Order by attribute similarities

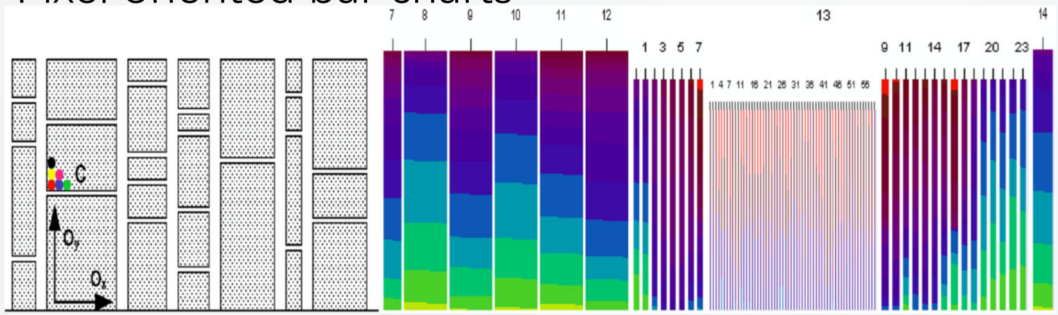


(Keim & Kriegel, 1994; 1996)

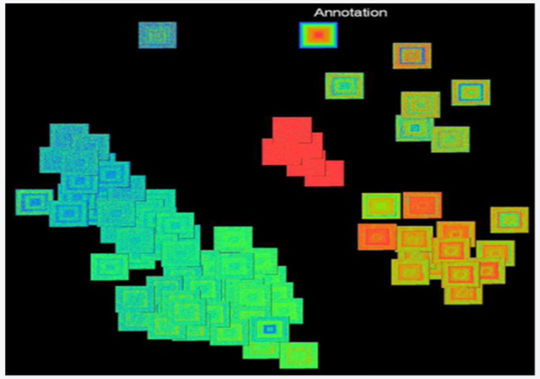
Pixel Oriented Techniques – Display Region

- Ways of splitting the display region

Pixel oriented bar charts



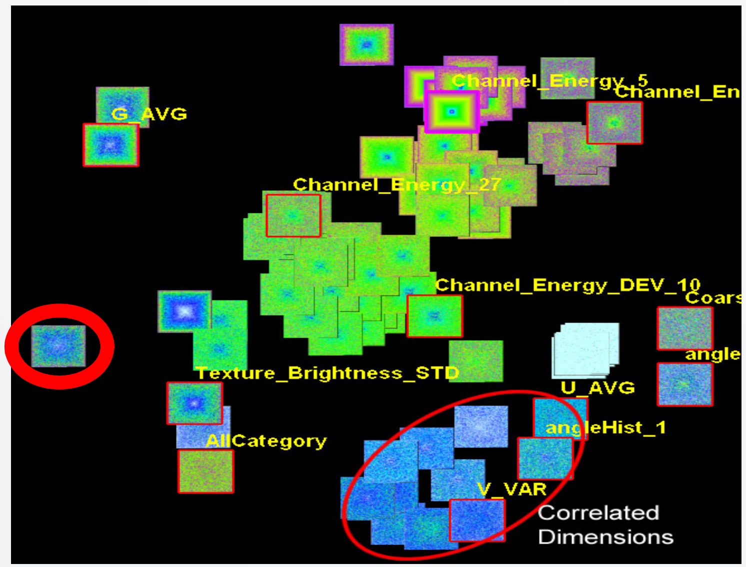
(Keim et al., 2001; 2002)



(Yang et al., 2006)

Pixel Oriented Techniques

Value & Relation View



(Yang et al., 2007)

- Visual attributes in each dimension as an pixel icon
- Project the icons based on MDS
- Similar dimensions are clustered
- Merge the similar dimensions
Delete the outliers

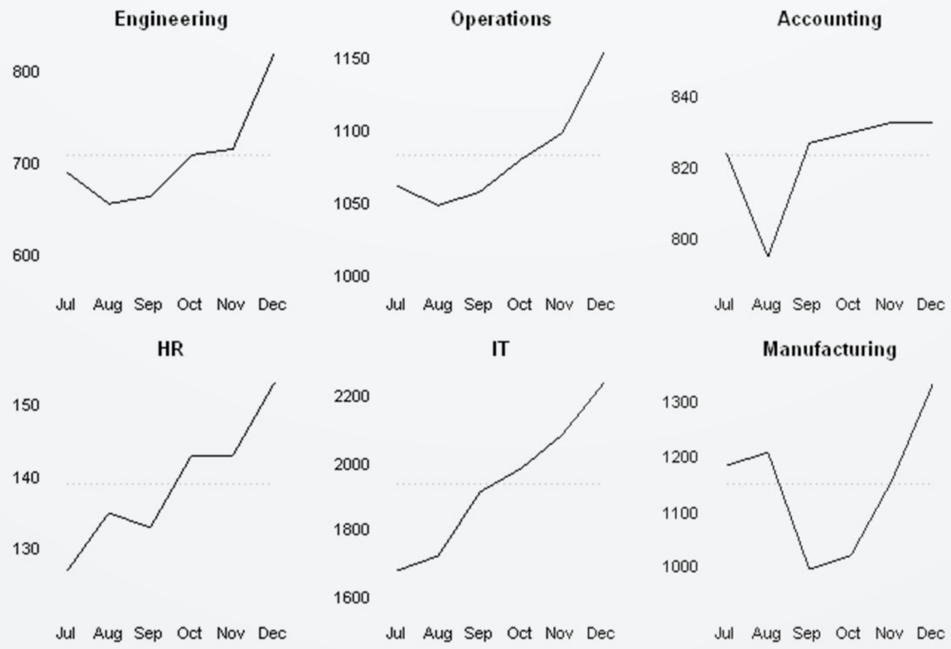
OUTLINE

Approaches

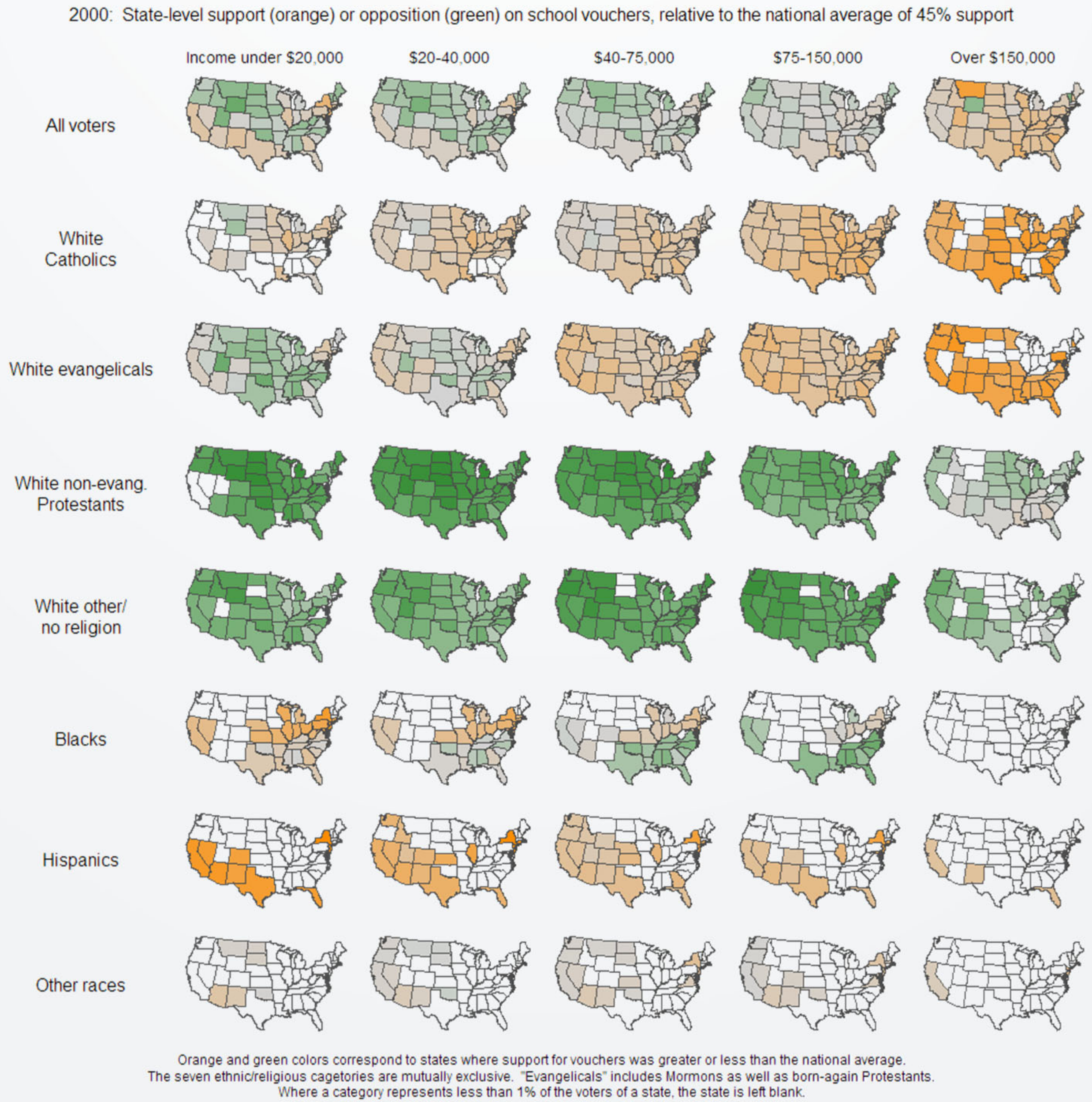
- Coordinate Systems
  - Scatter-plot Matrix.
  - Parallel Coordinates.
  - Dimensionality Reduction.
- Glyph-based Methods.
- Pixel Oriented Techniques
- **“Small Multiples”.**
- Visual Diagnosis

Small Multiples

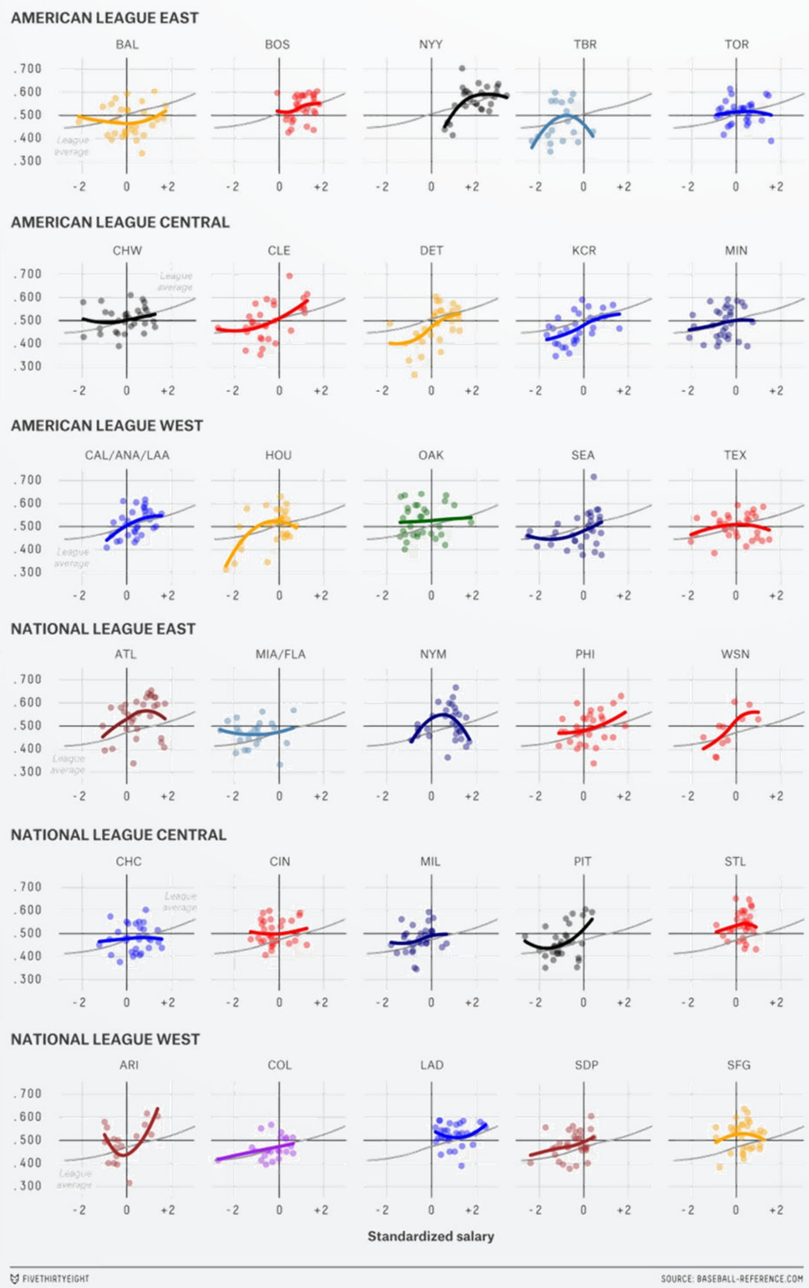
- Popularized by Edward Tufte.
- A series or grid of small similar graphics or charts for comparison.



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



http://andrewgelman.com/2009/07/hard_sell_for_b/



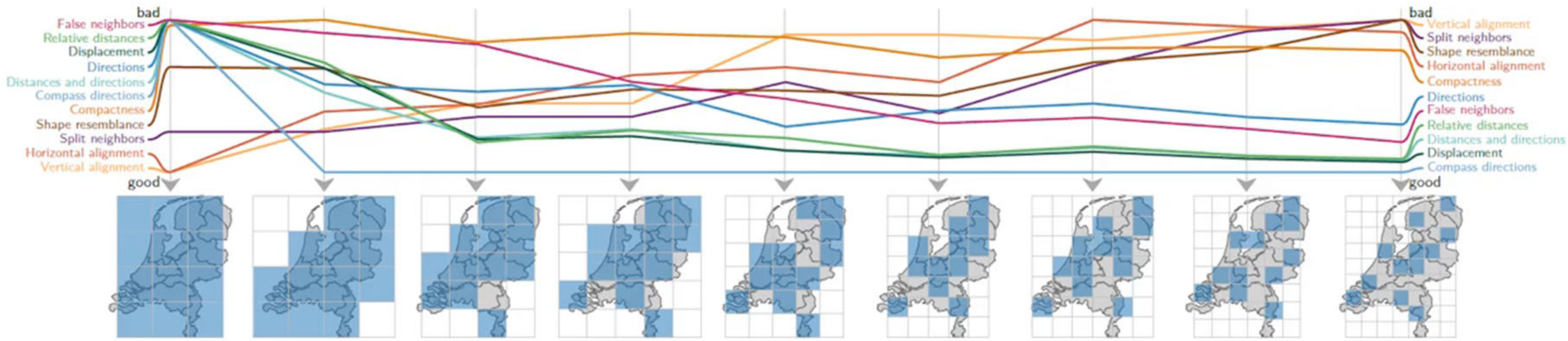
How your favorite baseball team blows-its-money?

https://fivethirtyeight.com/features/how-your-favorite-baseball-team-blows-its-money/Small_multiple

Small Multiples with Gaps

Small Multiples

Wouter Meulemans Jason Dykes Aidan Slingsby
Cagatay Turkay Jo Wood



http://www.gicentre.net/smwg

Meulemans W, Dykes J, Slingsby A, et al. Small multiples with gaps[J].
IEEE transactions on visualization and computer graphics, (2017)

HiPiler

Lekschas F, Bach B et al. HiPiler: visual exploration of large genome interaction matrices with interactive small multiples[J].
IEEE transactions on visualization and computer graphics (2018)

OUTLINE

Approaches

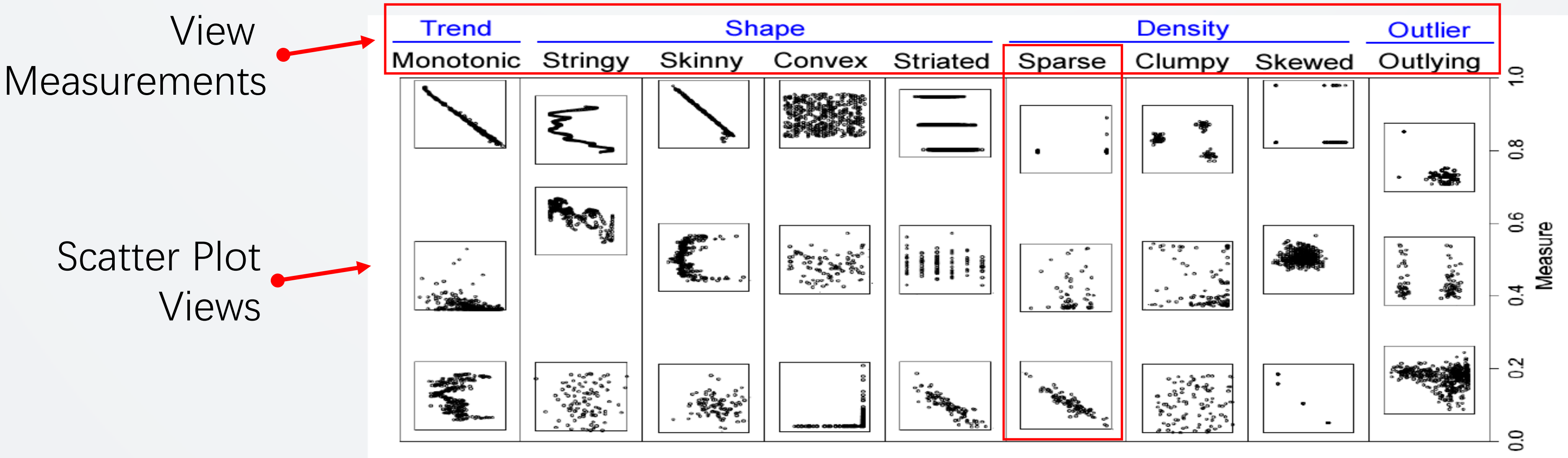
- Coordinate Systems
  - Scatter-plot Matrix.
  - Parallel Coordinates.
  - Dimensionality Reduction.
- Glyph-based Methods.
- Pixel Oriented Techniques
- “Small Multiples”.
- ***Visual Diagnosis***

Visual Diagnostics

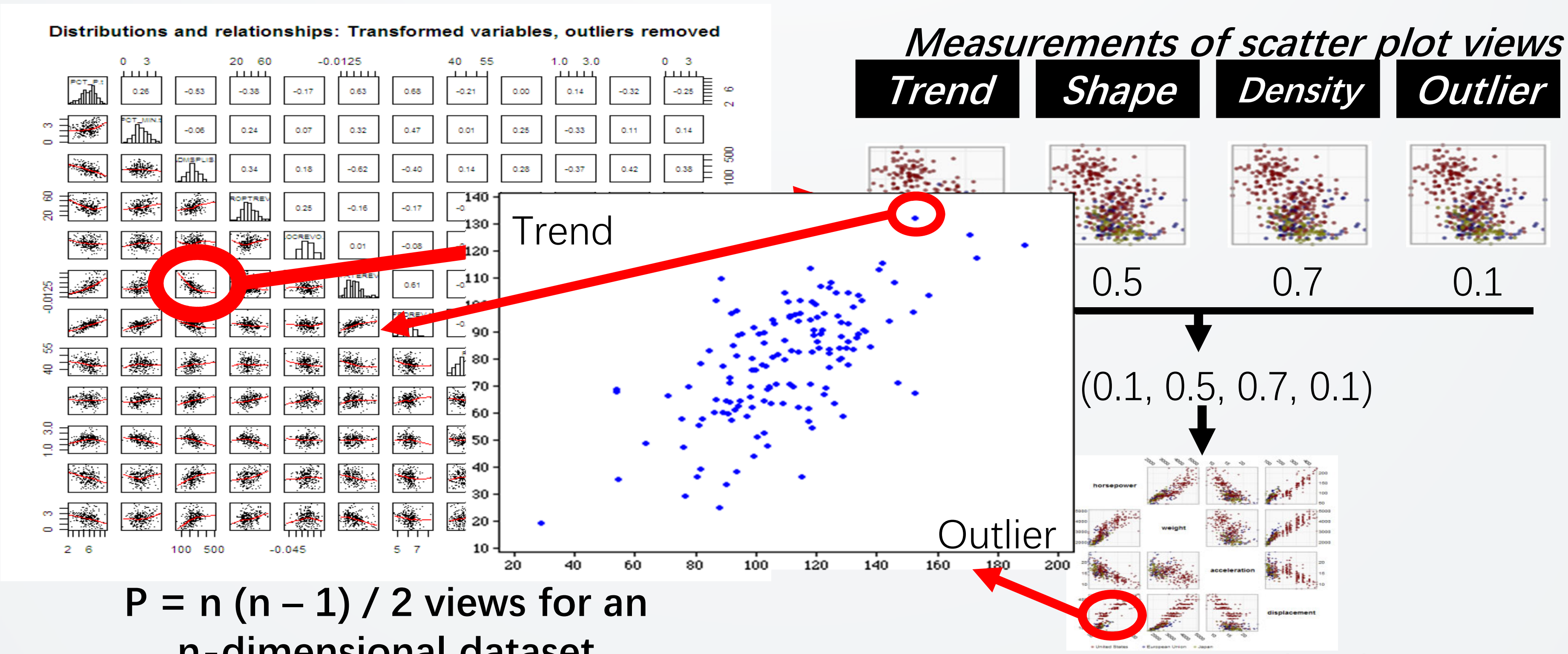
- Visual diagnostics:
  - Estimates multidimensional visualization views based on a set of measures
  - Recommends the views with interesting visual patterns based on their measurements
- Applications:
  - Scagnostics (scatter plots + diagnostics)
  - Pixnostics (pixel + diagnostics)
  - Pragnostics (parallel coordinates + diagnostics)

Scagnostics (Scatter plots + Diagnostics)

- Scatter plots diagnostics estimates scatter plot views based on a set of pre-defined measurements
- Views are ranked based on their measurements



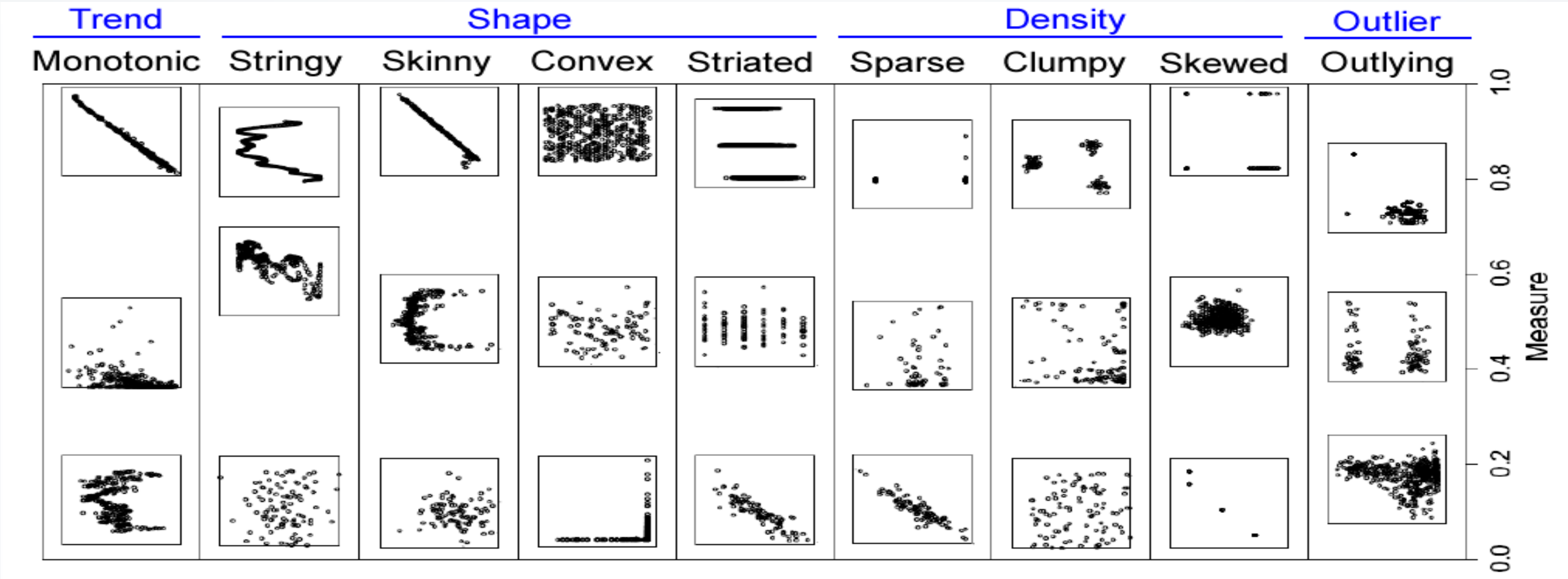
Scagnostics *(**Sca**tter plots + Dia**gnostics**)*



P = n (n – 1) / 2 views for an n-dimensional dataset

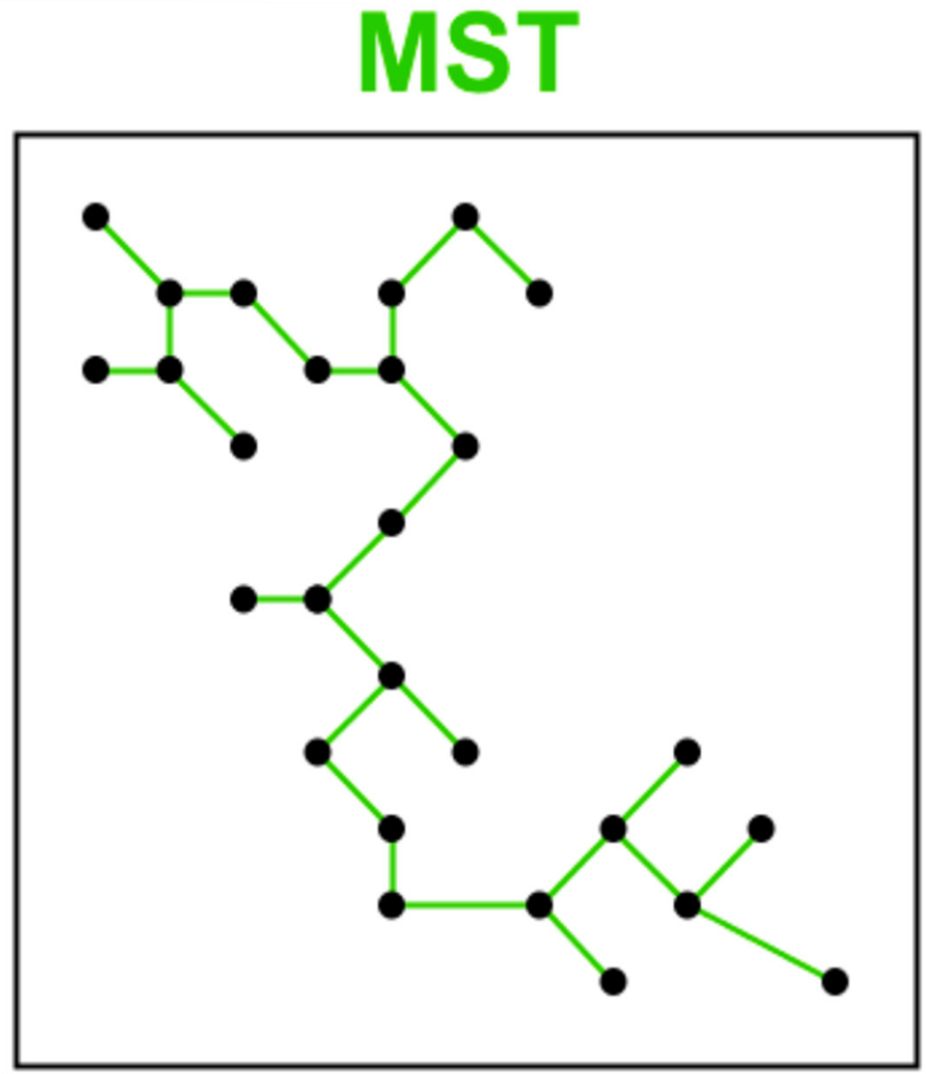
scatter plot matrices in a much smaller scale

Measurements for Scatter Plots



Proximity Graphs

Features are computed based on three types of graphs derived from the scatter plot

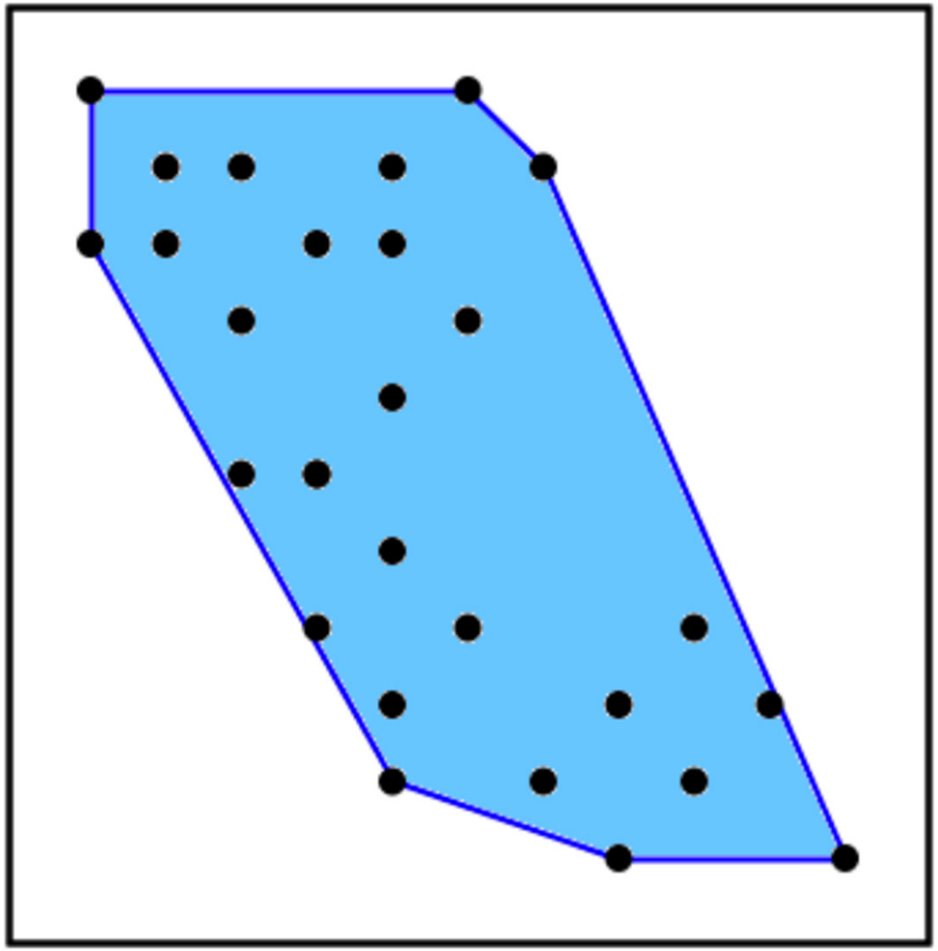


- A path is a list of successively adjacent, distinct edges
- A tree is a graph in which any two nodes are connected by exactly one path
- A spanning tree is an undirected graph whose edges are structured as a tree
- A minimum spanning tree (MST) is a spanning tree whose total length (sum of edge weights) is least of all spanning trees on a given set of points
- The edge weights (edge lengths) of a geometric MST are computed from the distances between its vertices.

Alpha Shape

- A proximity graph (or neighborhood graph) is a geometric graph whose edges are determined by an indicator function based on distances between a given set of points in a metric space.
- To define this indicator function, we use an open disk D.
  - D touches a point if that point is on the boundary of D
  - D contains a point if that point is in D
  - Denote an open disk of fixed radius D(r)
- In an alpha shape graph, an edge exists between any pair of points that can be touched by an open disk D(r) containing no points

Convex Hull



- A hull of a set of points X in R^2 is a collection of the boundaries of one or more polygons that have a subset of the points in X for their vertices and that collectively contain all the points in X

- A hull is convex if it contains all the straight line segments connecting any pair of points in its interior

Measurements

- The length of an edge, i.e., the Euclidean distance between its vertices
- The length of a graph, i.e., is the sum of the lengths of its edges
- A path is a list of vertices such that all pairs of adjacent vertices in the list are edges
- A path is closed if its first and last vertex are the same
- A closed path is the boundary of a polygon
- The perimeter of a polygon is the length of its boundary
- The area of a polygon is the area of its interior

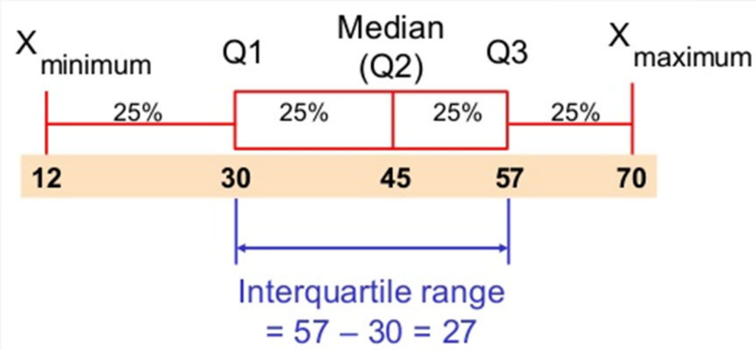
Outlying

- We consider an outlier to be a vertex whose adjacent edges in the MST all have a weight (length) greater than w

$$\omega = q_{75} + 1.5(q_{75} - q_{25})$$

- A measure of the proportion of the total edge length due to extremely long edges connected to points of single degree:

$$c_{outlying} = length(T_{outliers})/length(T)$$



Density

- **Skewed**: relatively robust measure of skewness in the distribution of edge lengths

$$c_{skew} = (q_{90} - q_{50})/(q_{90} - q_{10})$$

- **Clumpy**: clusters with small intracluster distances relative to the length of their connecting edge and ignores runt clusters with relatively small runt size:

$$c_{clumpy} = \max_j \left[1 - \max_k [length(e_k)]/length(e_j) \right]$$

Density

- **Sparse**: sparseness statistic measures whether points in a 2D scatterplot are confined to a lattice or a small number of locations on the plane

$$c_{sparse} = q_{90}$$

- **Striated:** We define coherence in a set of points as the presence of relatively smooth paths in the minimum spanning tree.

$$c_{striate} = \frac{1}{|V|} \sum_{v \in V^{(2)}} I(\cos \theta_{e(v,a)e(v,b)} < -.75)$$

Shape

- **Convex.** Our convexity measure is based on the ratio of the area of the alpha hull and the area of the convex hull. This ratio will be 1 if the nonconvex hull and the convex hull have identical areas:

$$c_{convex} = area(A)/area(H)$$

- **Skinny.** The ratio of perimeter to area of a polygon measures, roughly, how skinny it is.

$$c_{skinny} = 1 - \sqrt{4\pi area(A)}/perimeter(A)$$

- **Stringy**. A s shape with no branches.

$$c_{stringy} = \frac{|V^{(2)}|}{|V| - |V^{(1)}|}$$

# of vertices of degree 2 in MST

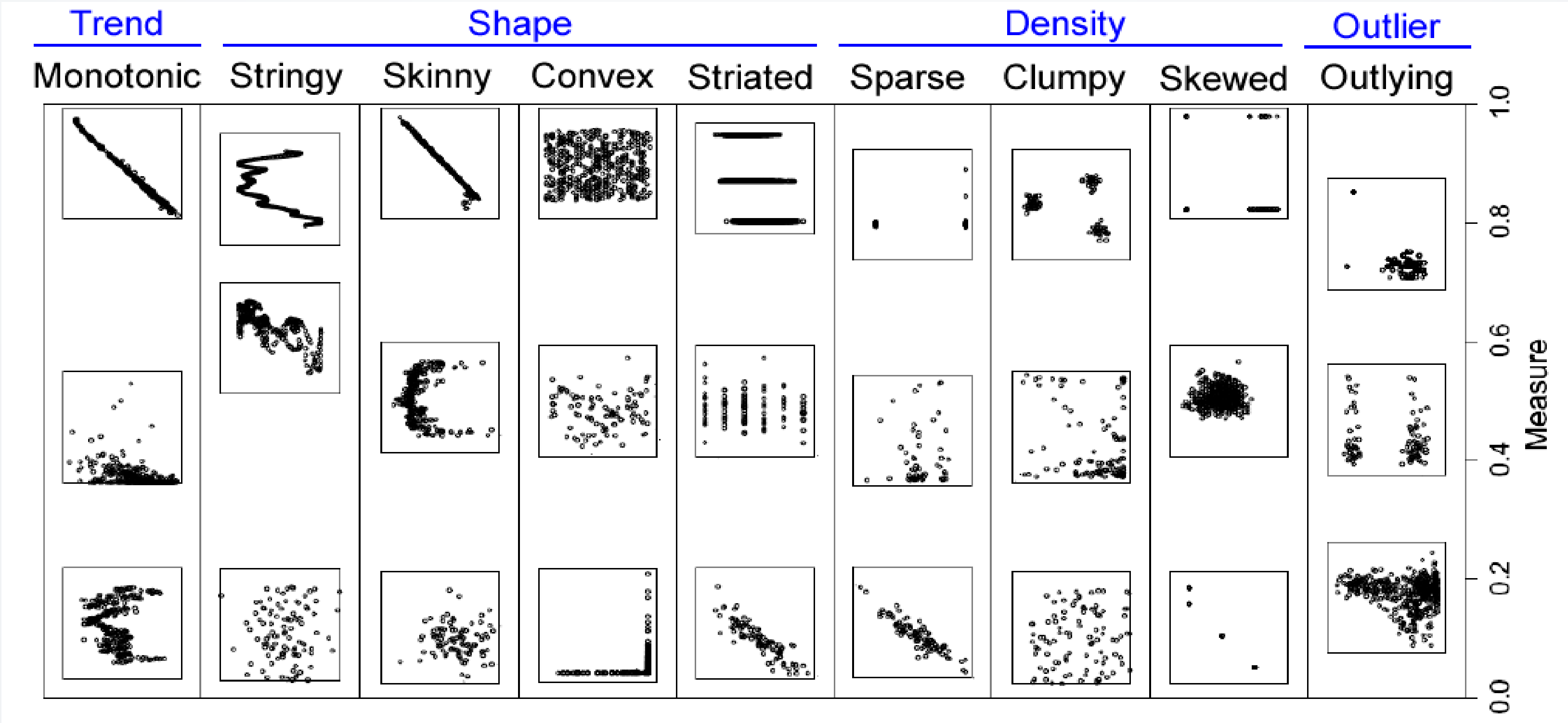
# of vertices - # of single-degree vertices

Association

- Monotonic: the squared Spearman correlation coefficient, which is a Pearson correlation on the ranks of x and y (corrected for ties), to assess monotonicity in a scatterplot

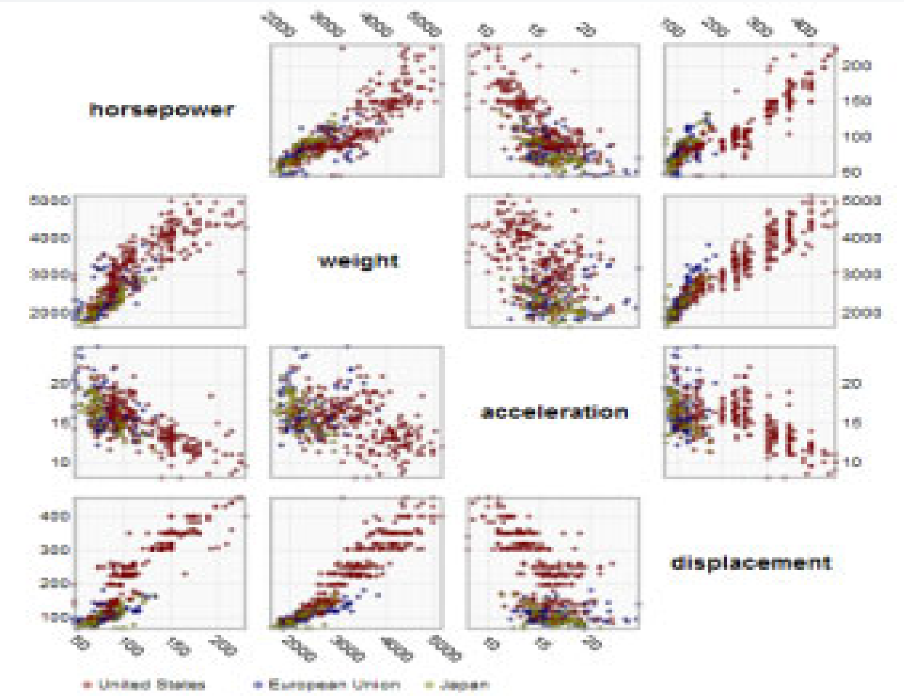
$$c_{monotonic} = {r^2}_{spearman}$$

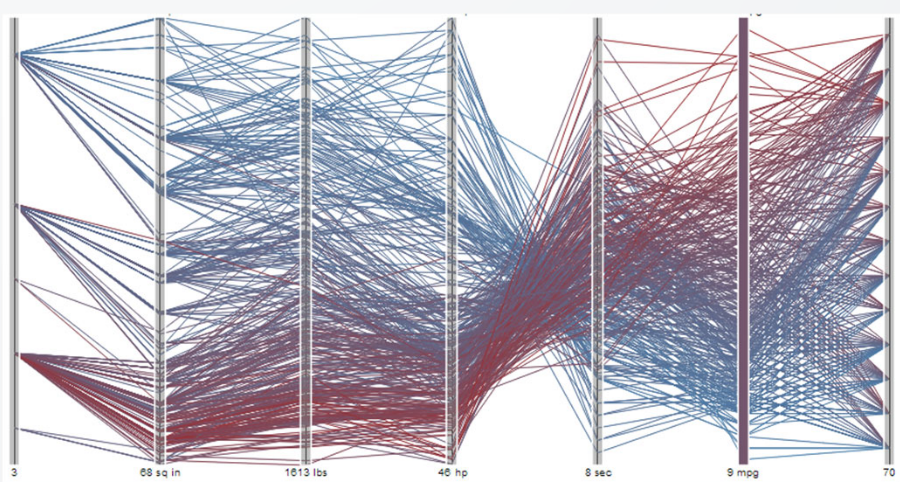
Measurements for Scatter Plots



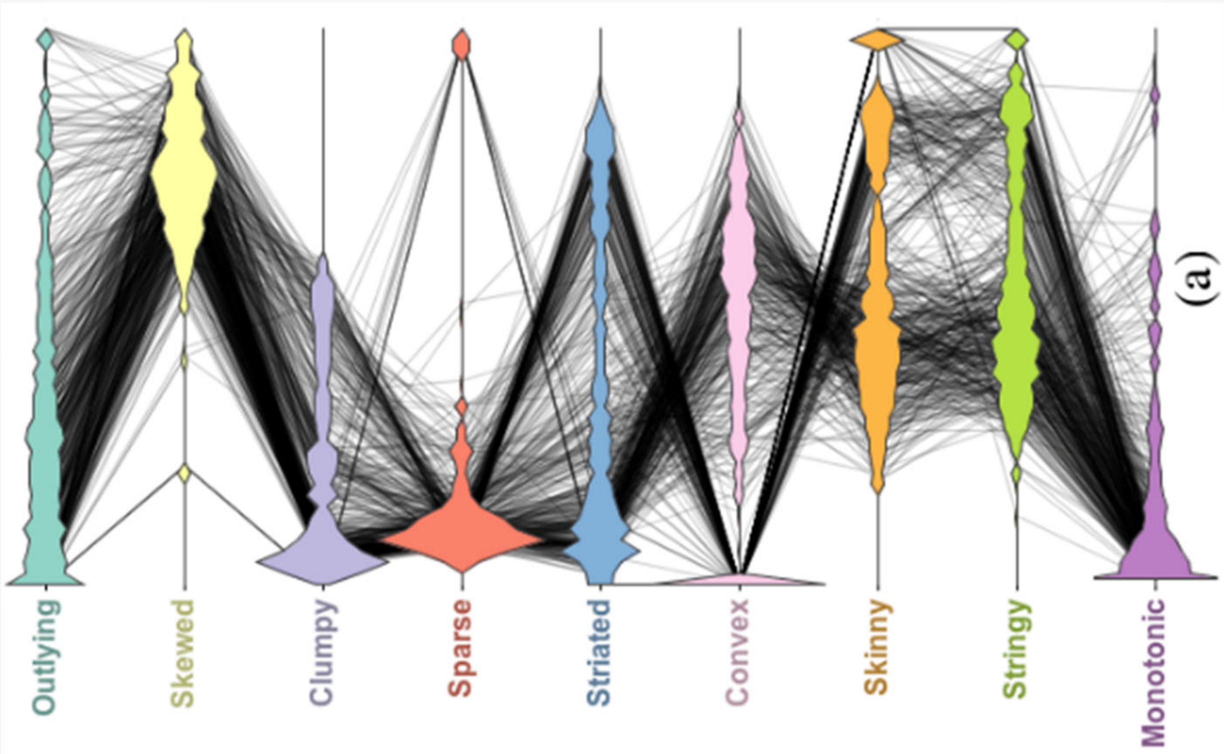
Use 10 minutes to design a visualization to represent the scagnostics results

Visualized in another SPM or PCP

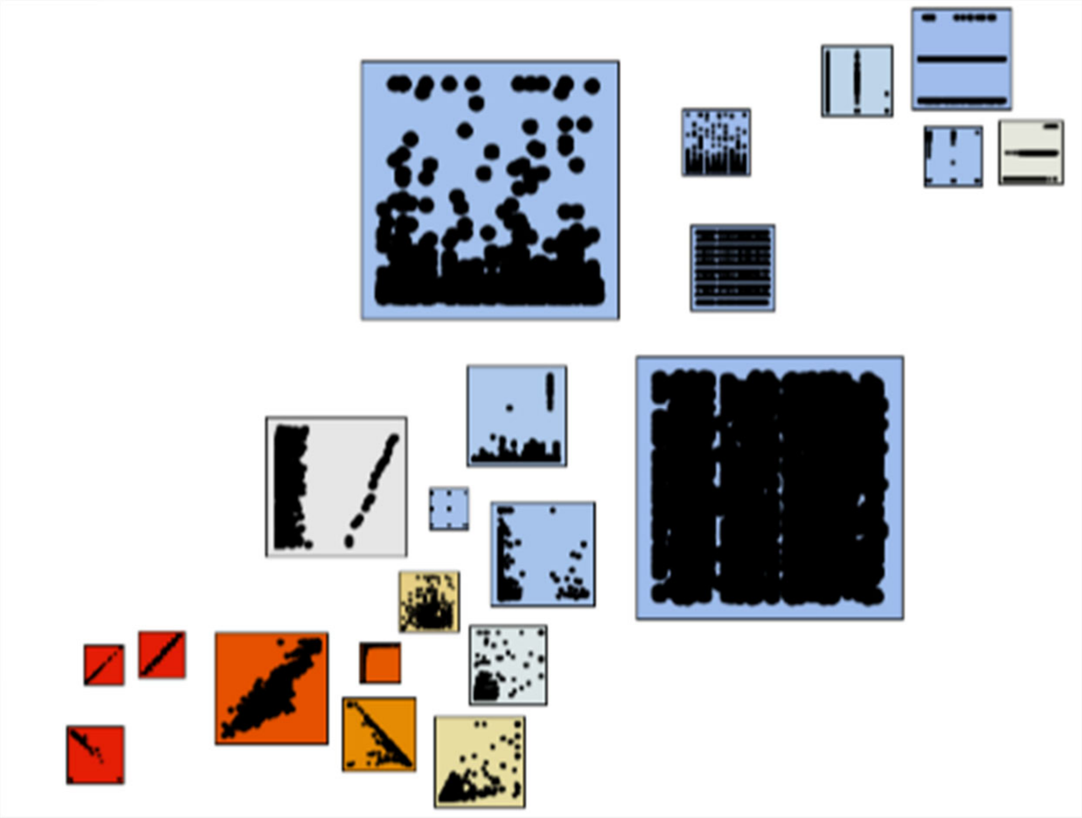




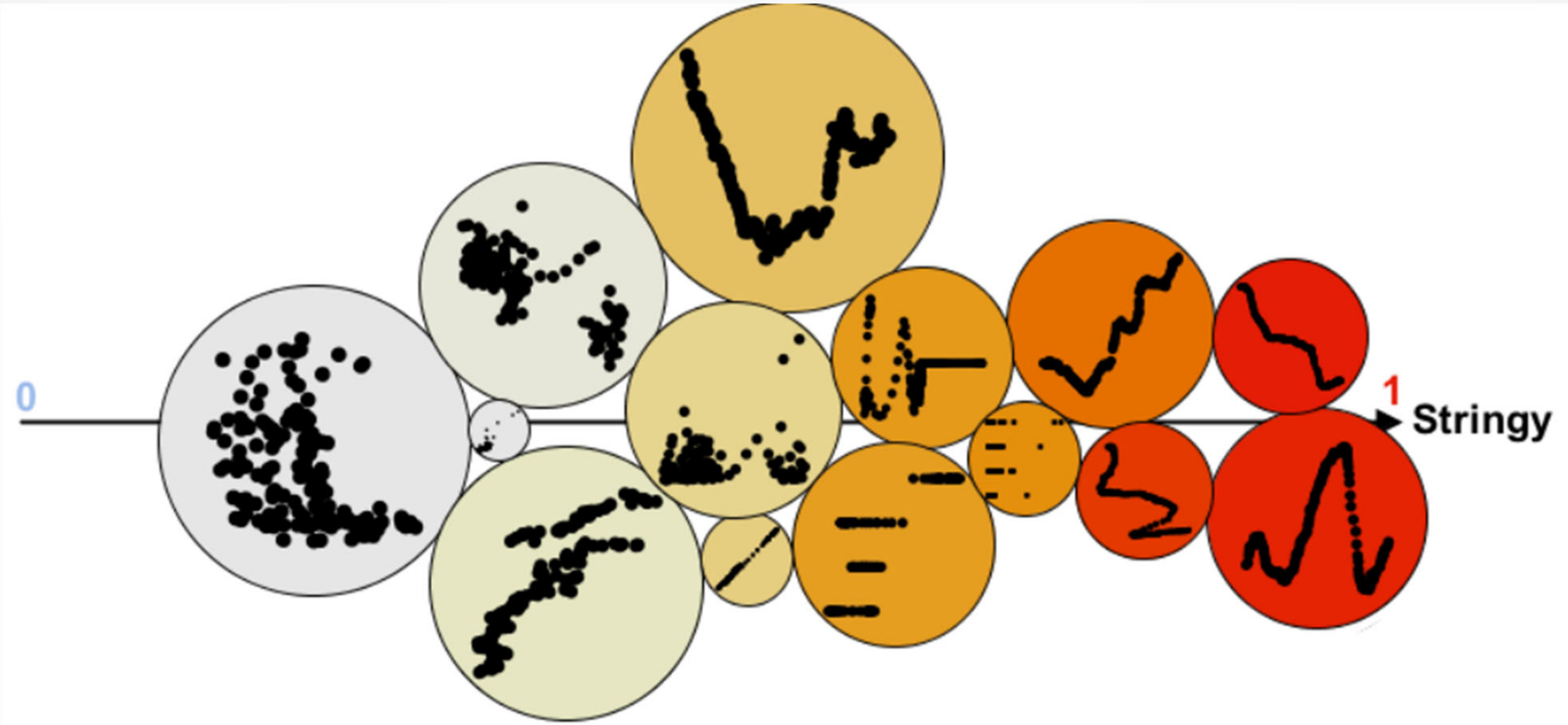
Is there a better approach ?



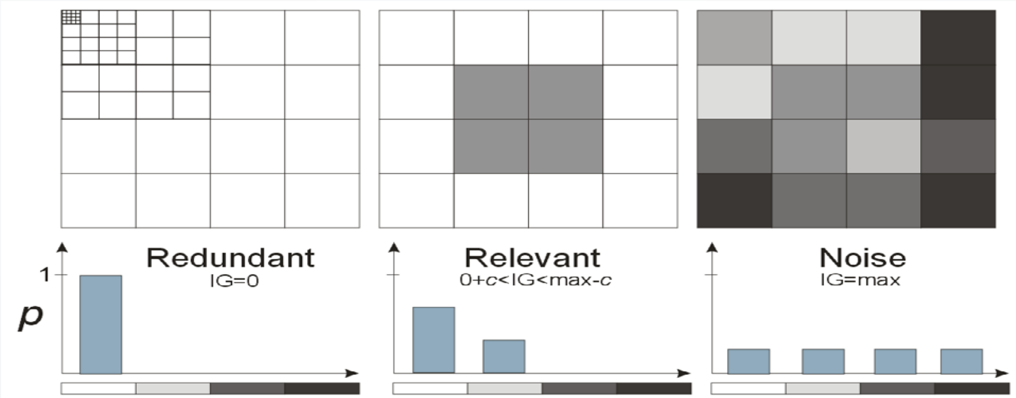
Is there a better approach ?



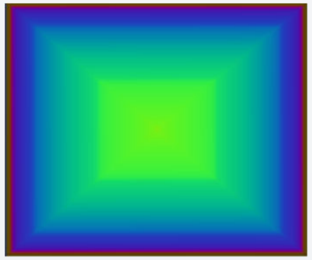
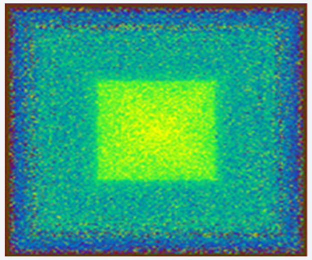
Is there a better approach ?



Pixnostics (pixel + diagnostics)



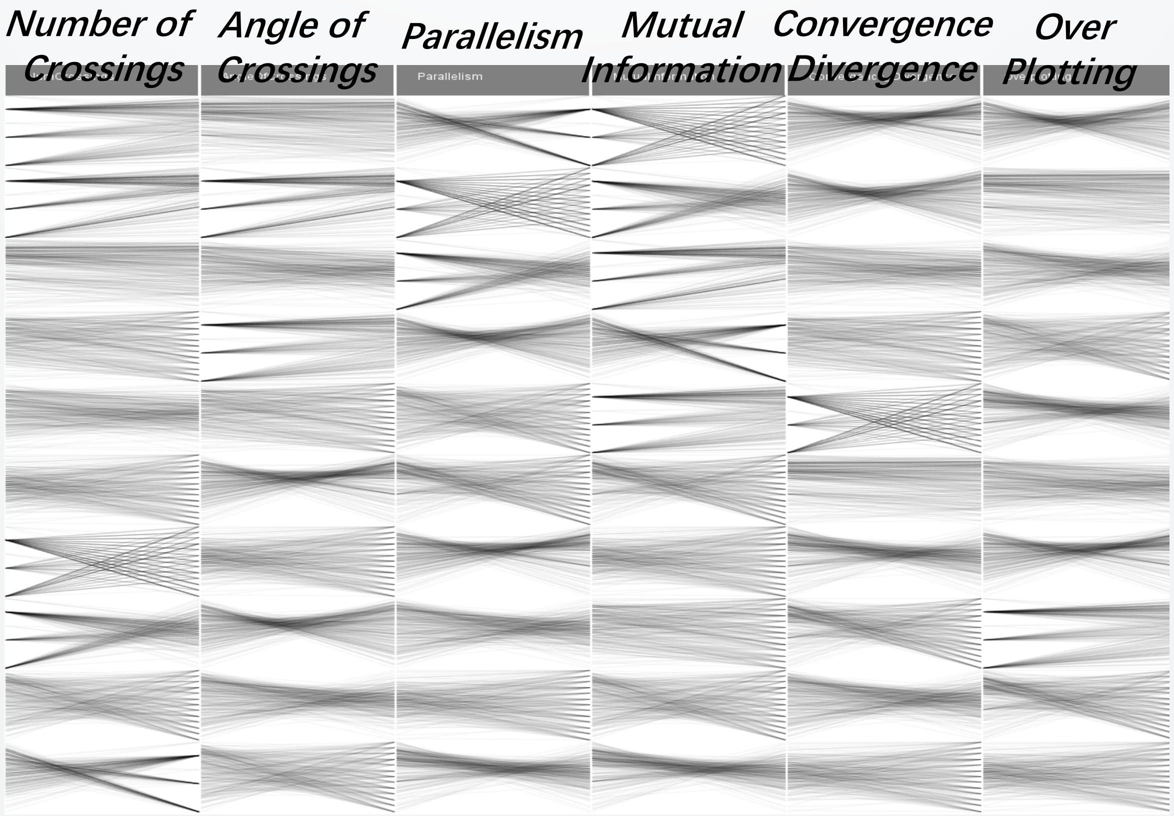
examples of pixel measures

Relevant score:  > 

0.8 0.5

(schneidewind et al., 2006)

Pragnostics (parallel coordinates + diagnostics)



Ranking the order of dimension pairs
(Dasgupta & Kosara, 2010)

Thank You！