Data Exploration & Visualization

Module 11

High-Dimensional Data Visualization

Dr. ZENG Wei

DSAA 5024

The Hong Kong University of Science and Technology (Guangzhou)

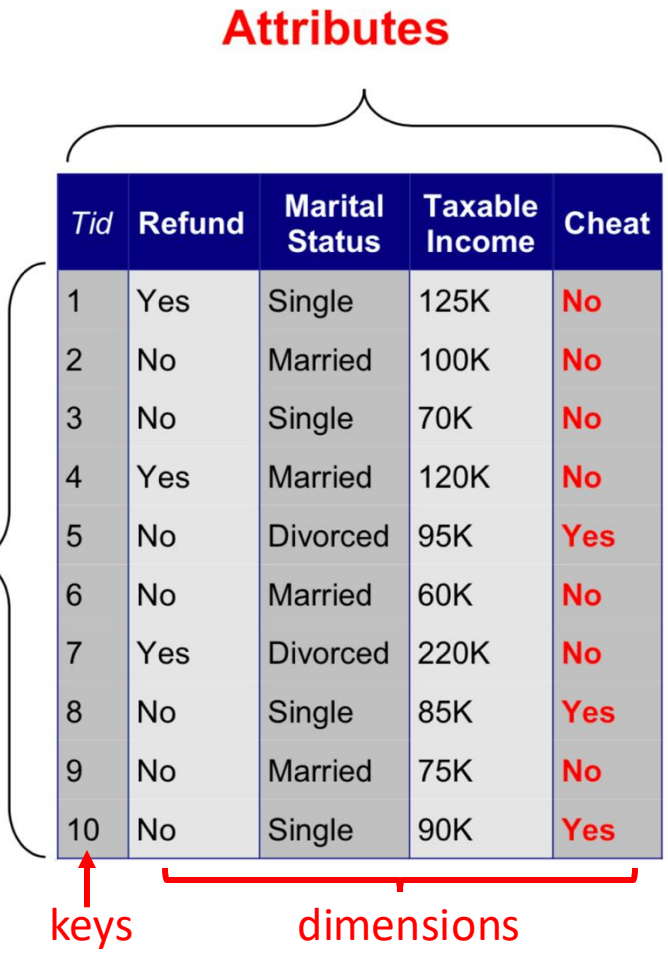
Data Exploration & Visualization

Module 11: High-Dimensional Data Visualization

- Data dimension
  - univariate, bivariate, trivariate, high-dimensional
- High-dimensional data visualization
  - Visual mapping: multiple views, scatterplot matrix, iconic representation, table lens, parallel coordinates
  - Dimension reduction:
    - Linear: PCA, MDS, LDA
    - Non-linear: t-SNE, UMAP

Data dimension

- An attribute is a property or characteristic of an object
  - Examples: eye color of a person, temperature, etc.
  - Attribute is also known as variable, field, characteristic, or feature
- A collection of attributes describe an object
  - Object is also known as record, point, case, sample, entity, or instance

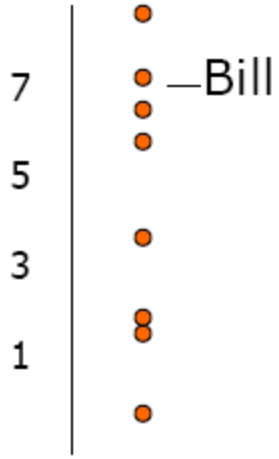


Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

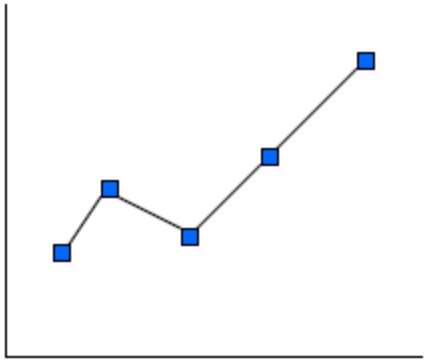
Data dimension

- Data sets of dimensions 1, 2, 3 are common
- Number of dimensions per class
  - 1 - Univariate data
  - 2 - Bivariate data
  - 3 - Trivariate data
  - > 3 – High dimensional data

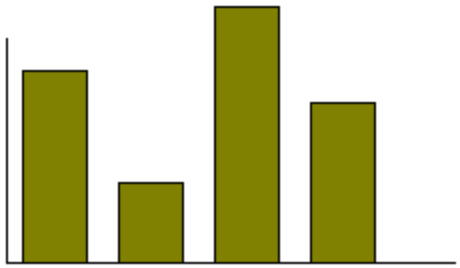
Univariate data representation



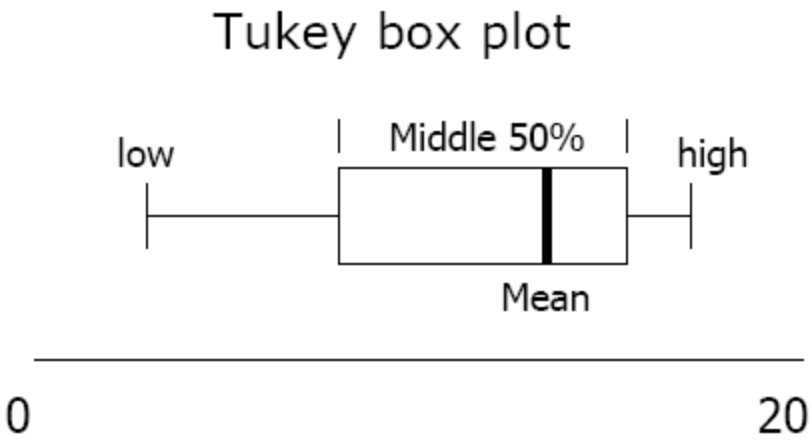
Strip plot



Univariate line plot

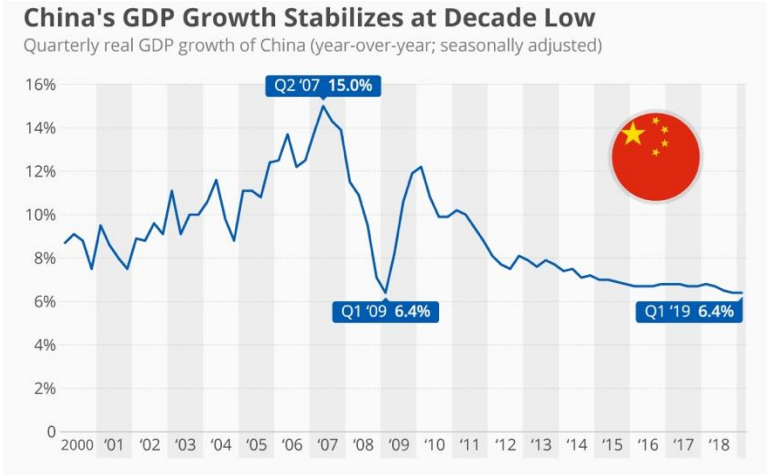


histogram



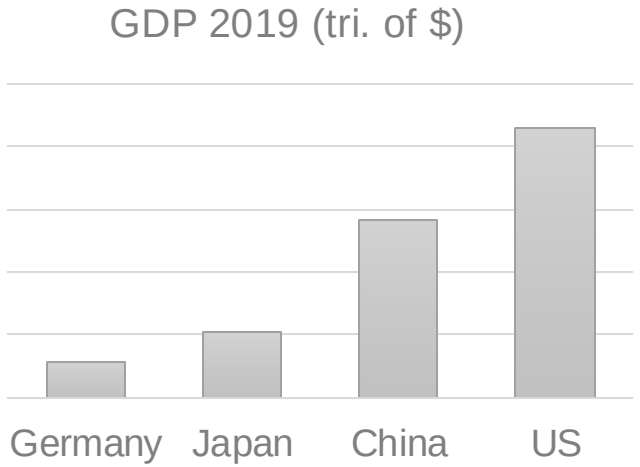
Univariate data representation

- In univariate representations, we often think of the data case as being shown along one dimension, and the value in another



Y Axis is quantitative

Graph shows change in Y over continuous range X



Y Axis is quantitative

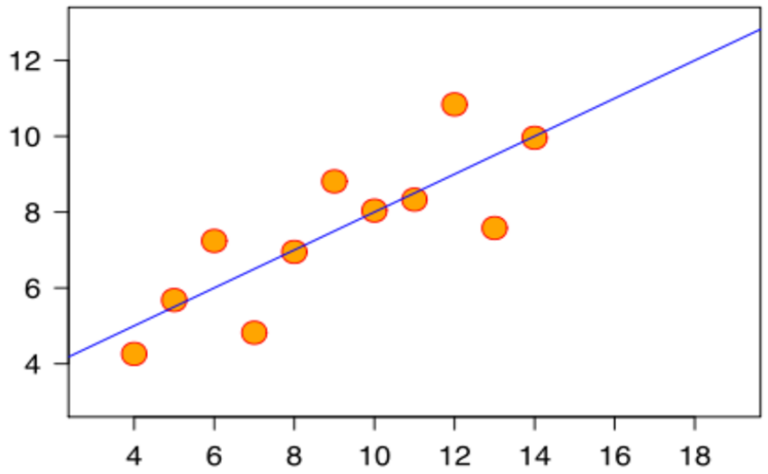
Graph shows value of Y for 4 cases

Univariate data representation

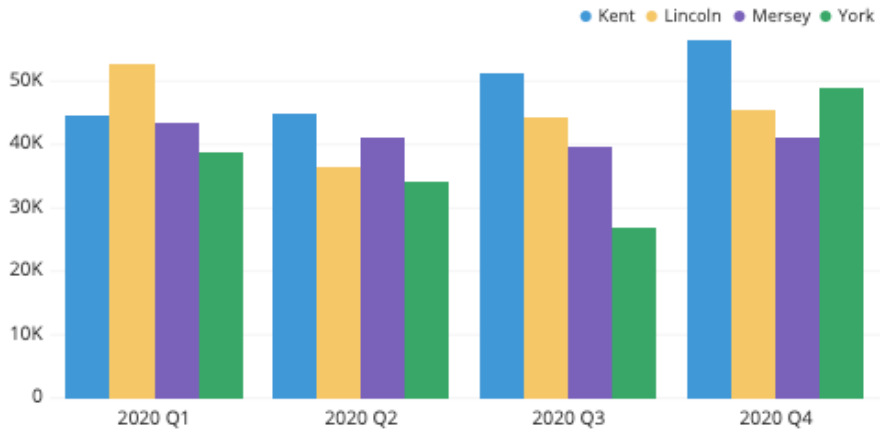
- We may think of graph as representing *independent* (data case) and *dependent* (value) variables

- Guideline:
  - Independent vs. dependent variables
    - Put independent on x-axis
    - See resultant dependent variables along y-axis

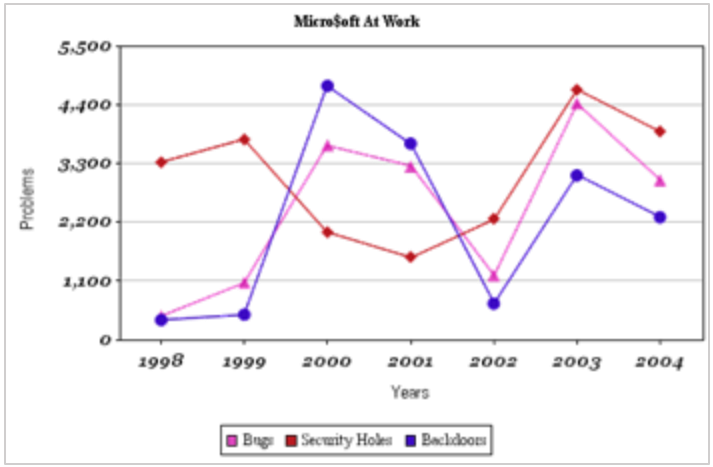
Bivariate data representation



Scatterplot



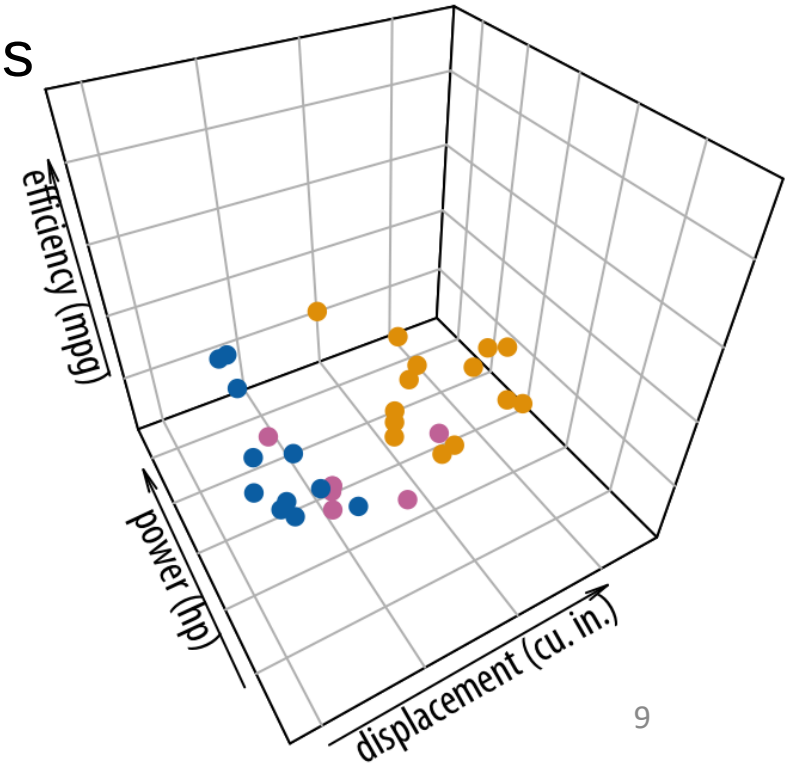
Grouped bar chart



Grouped line chart

Triviarate data representation

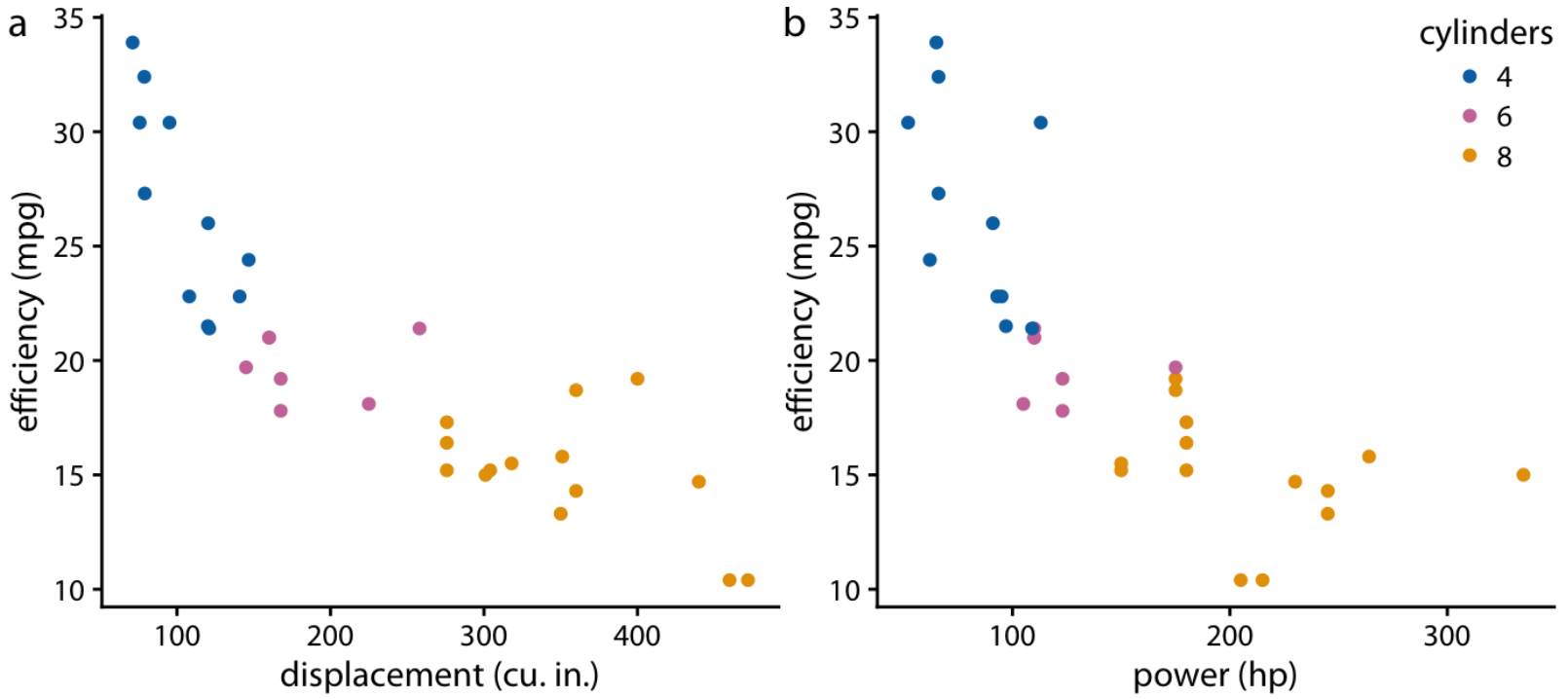
- 3D scatter plot
  - plot displacement along the *x* axis
  - power along the *y* axis
  - fuel efficiency along the *z* axis



- No unjustified 3D
  - disparity of depth
  - occlusion
  - interaction complexity
  - perspective distortion
  - text legibility

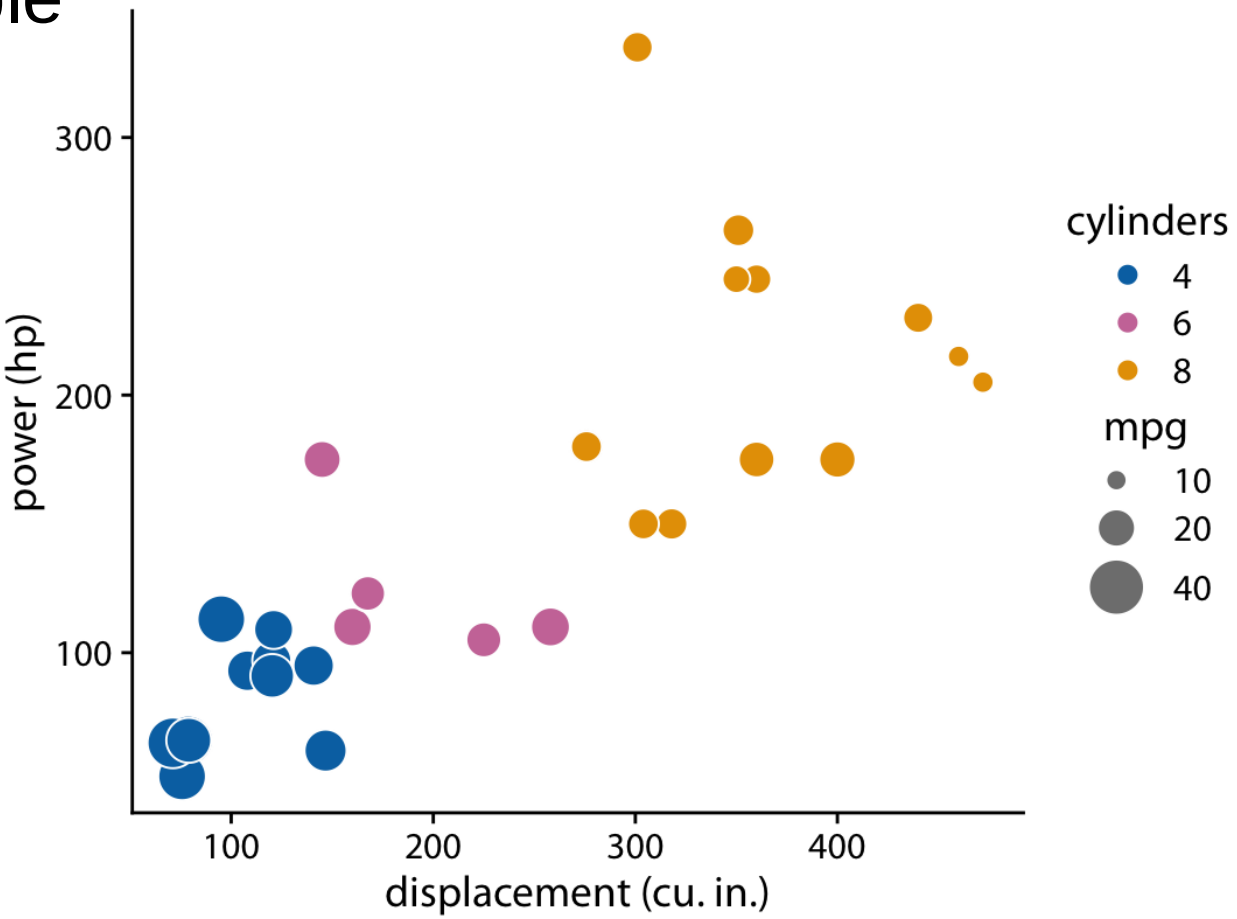
Triviarate data representation

- Alternative: small multiples



Triviarate data representation

- Alternative: Use mark attribute for another variable



High dimensional data

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

Sepal Length	Sepal Width	Petal Length	Petal Width	Species
	…	….	…..	
4.9	3.0	1.4	0.2	I. setosa
4.7	3.2	1.3	0.2	I. setosa
	…	….	…..	
5.7	2.8	4.5	1.3	I. versicolor
6.3	3.3	4.7	1.6	I. versicolor
	…	….	…..	
6.1	2.6	5.6	1.4	I. virginica
6.3	3.4	5.6	2.4	I. virginica
	…	….	…..	

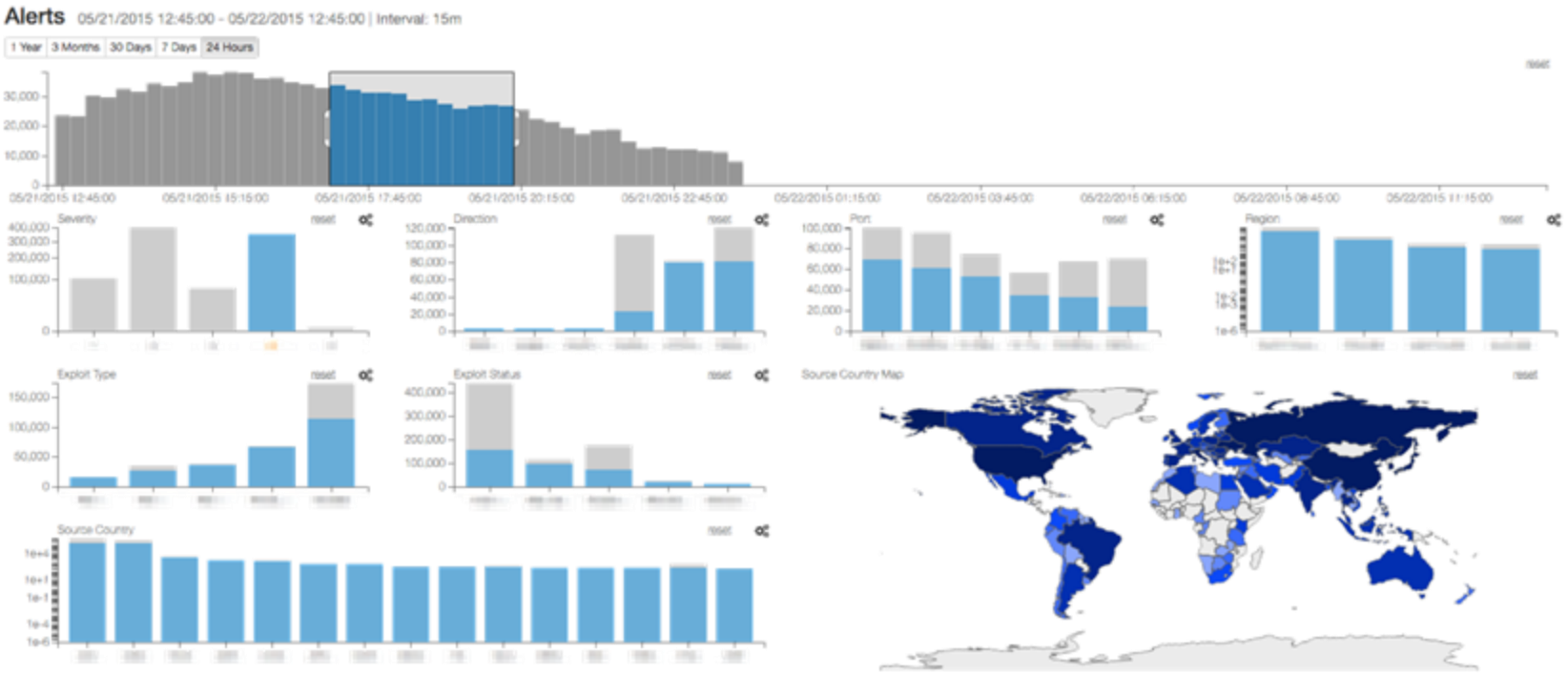
Data Exploration & Visualization

Module 11: High-Dimensional Data Visualization

- Data dimension
  - univariate, bivariate, trivariate, high-dimensional
- High-dimensional data visualization
  - Visual mapping: multiple views, scatterplot matrix, iconic representation, table lens, parallel coordinates
  - Dimension reduction:
    - Linear: PCA, MDS, LDA
    - Non-linear: t-SNE, UMAP

Simple solution

- Stay with standard views, but use lots and lots of them – multiple views.



Andrew et al., 2016

Multiple views

- Design guideline: consistency

Same Rule

If two views contain the **same** data field, that field should be encoded in the **same** way

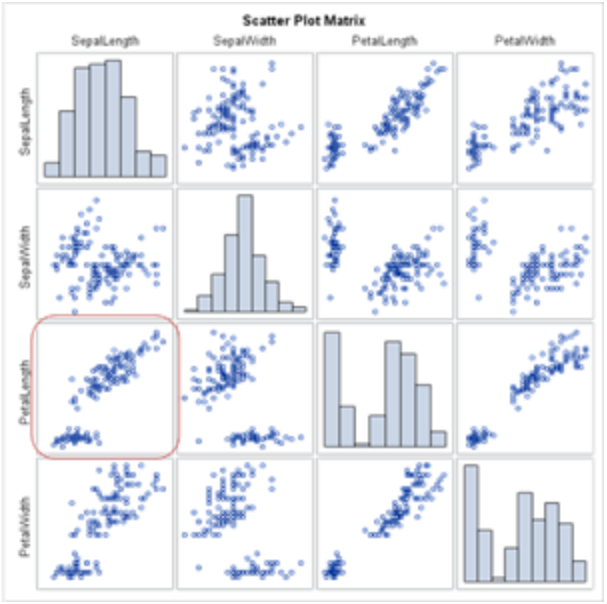
Different Rule

If two views contain **different** data fields, the two fields should be encoded **differently**

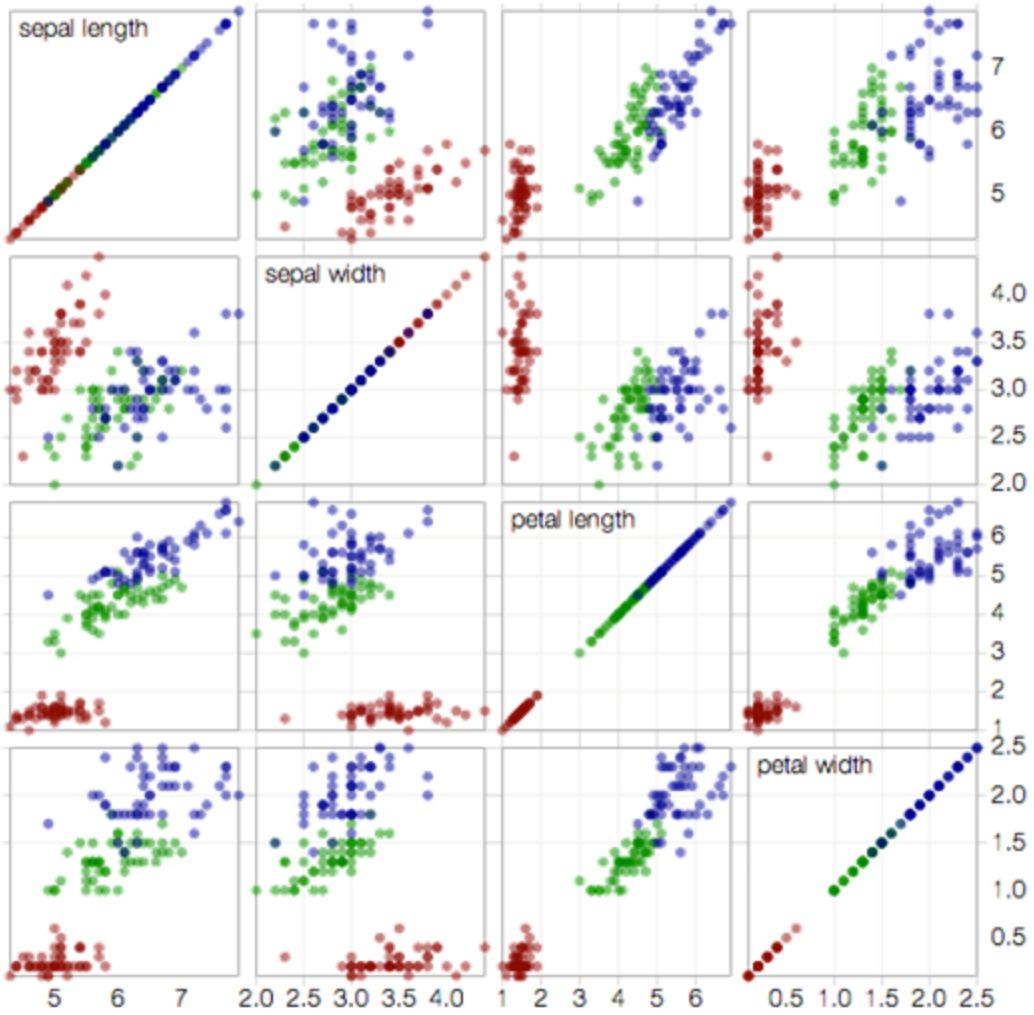
Qu and Hullman, 2018.

Scatterplot matrix

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



Scatterplot matrix

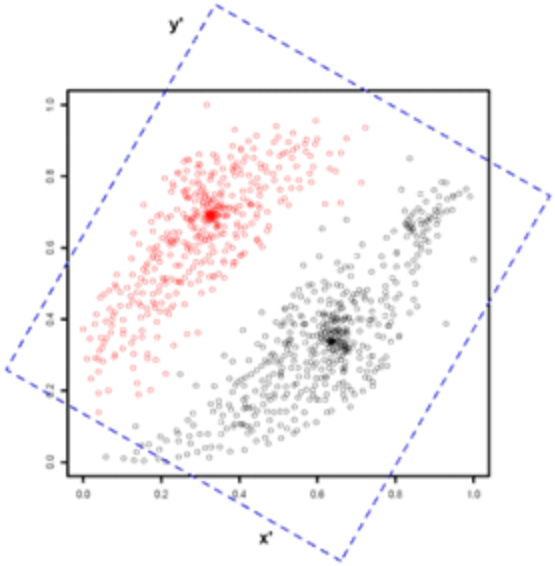


D3.js Example

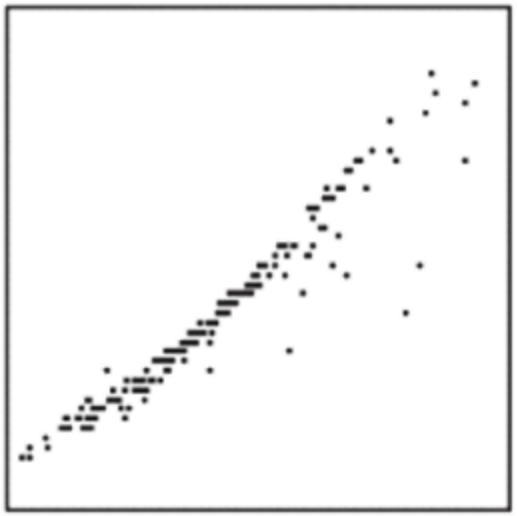
Scatterplot matrix

- Recommend scatter-plots with interesting patterns automatically.

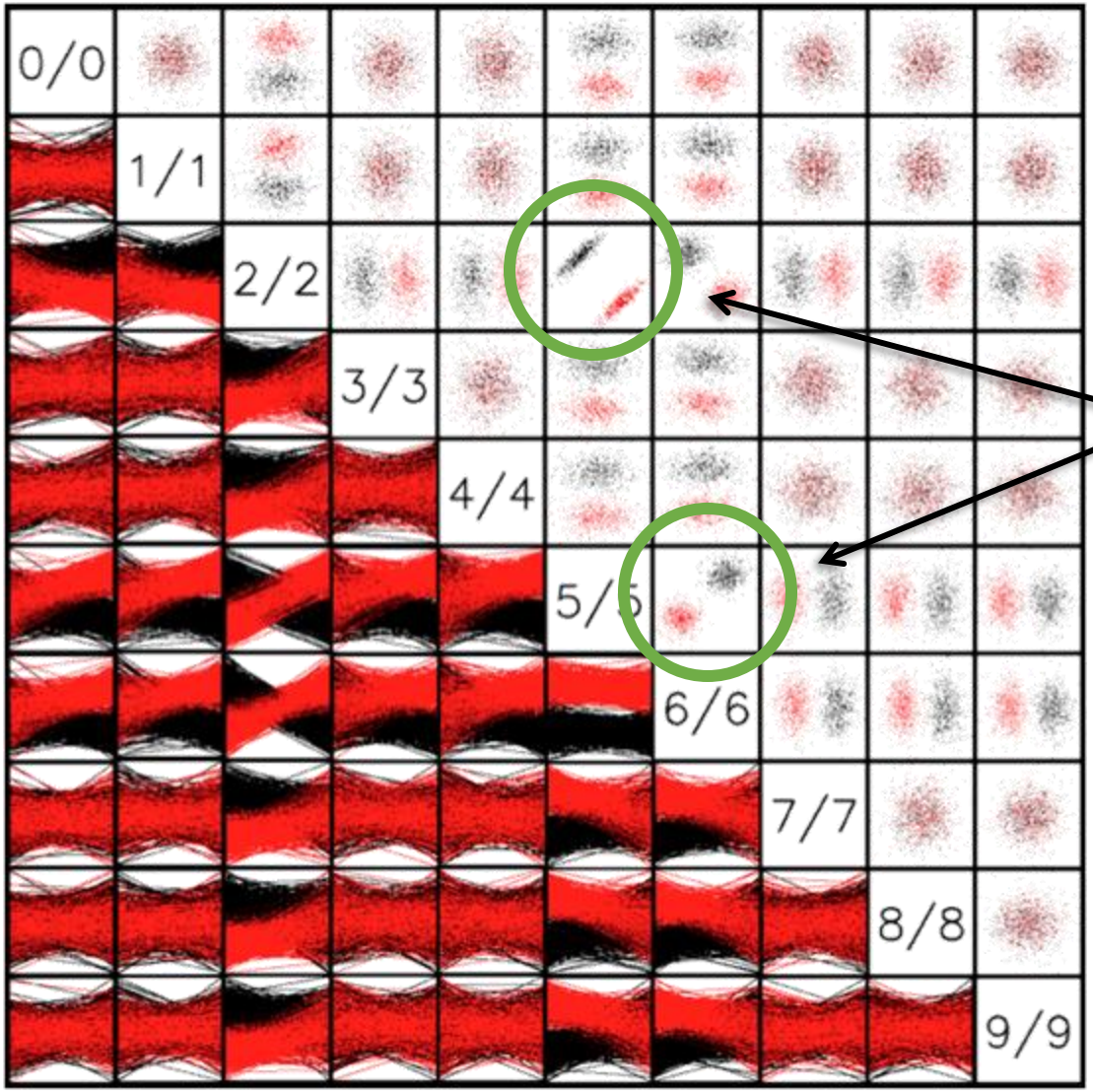
Clusters



Correlations



Tatu et al., 2011

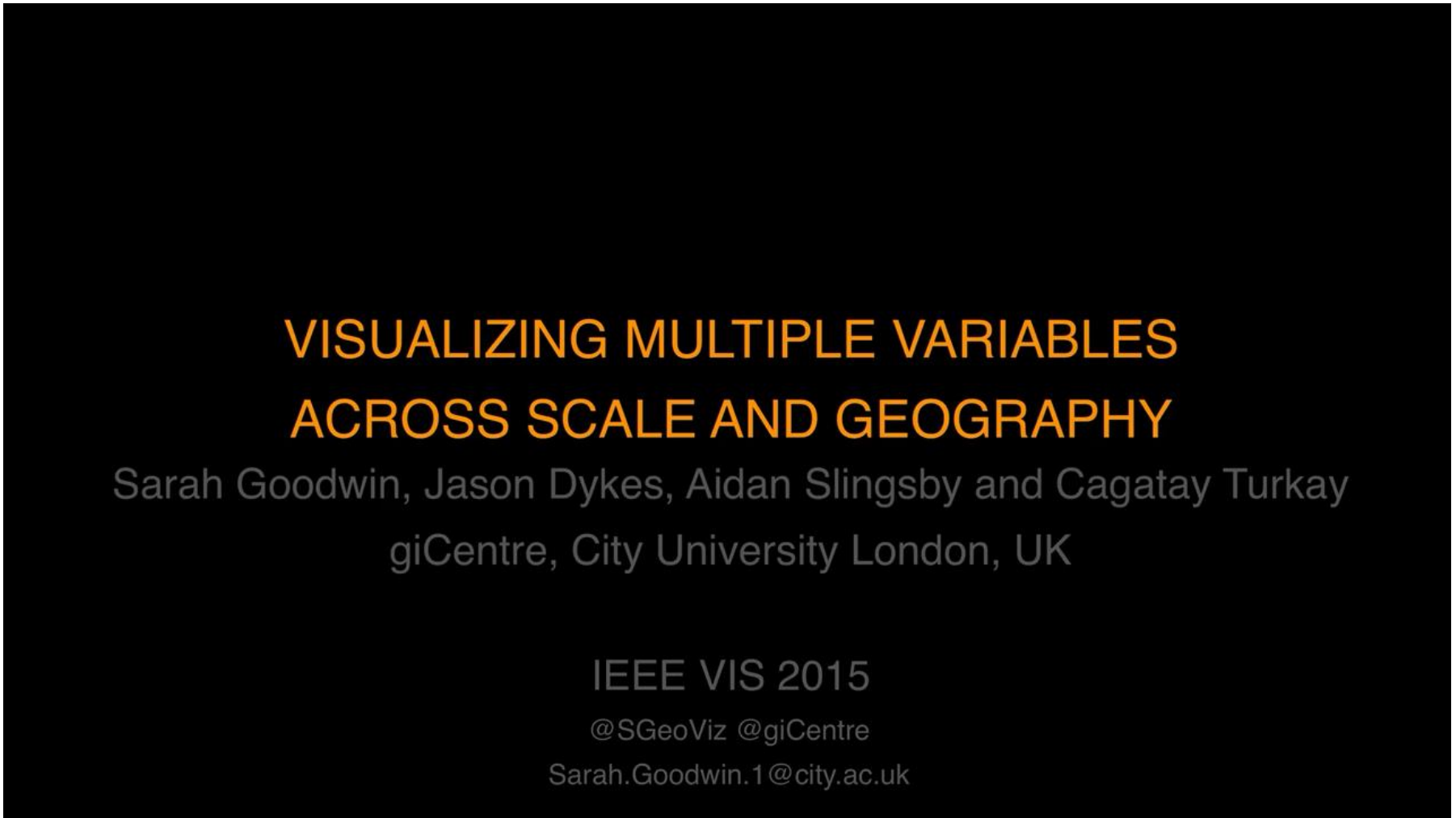


Clusters in scatter-plots

Tatu et al., 2011

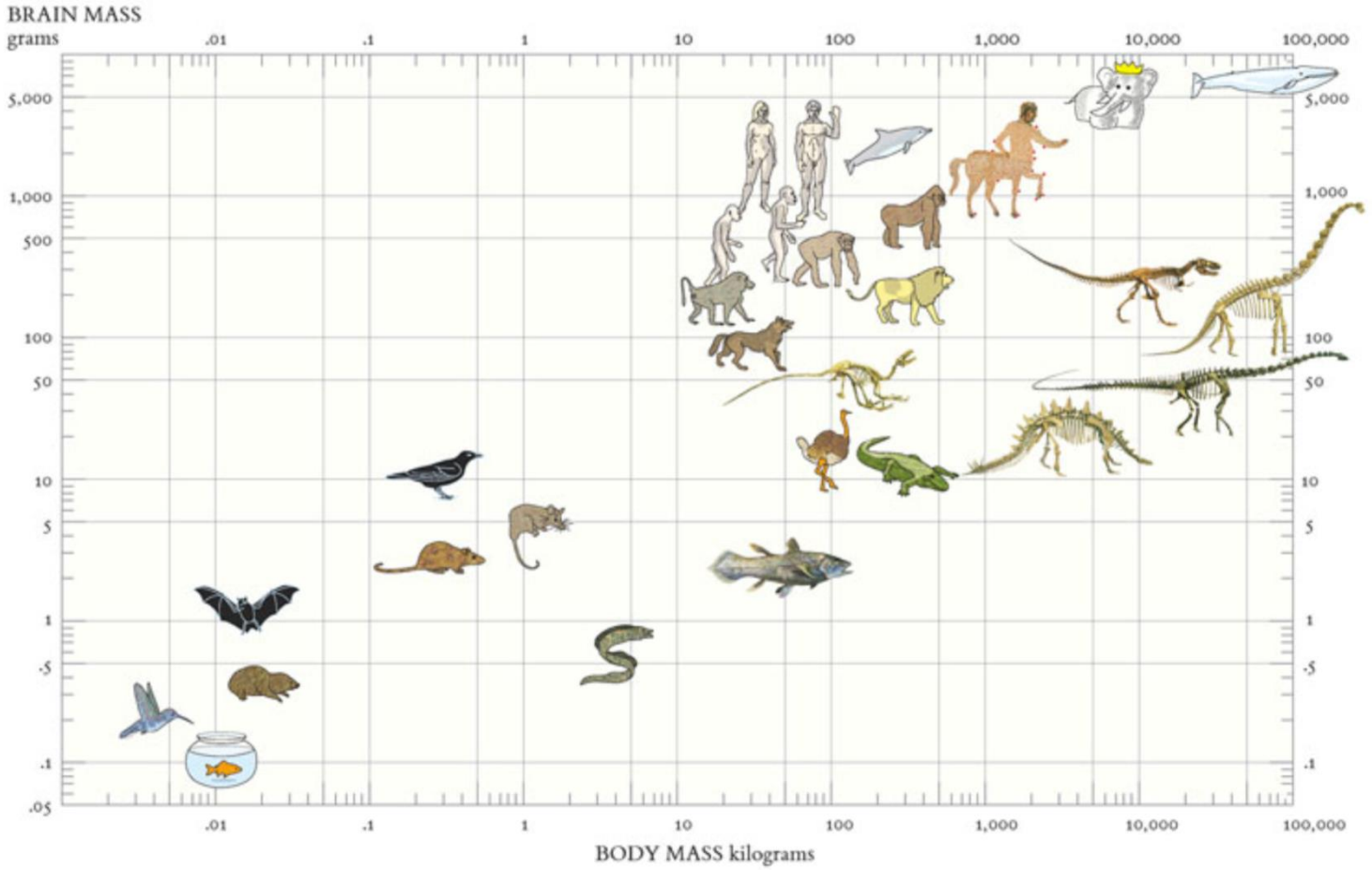
Scatterplot matrix

- Across scale and geography

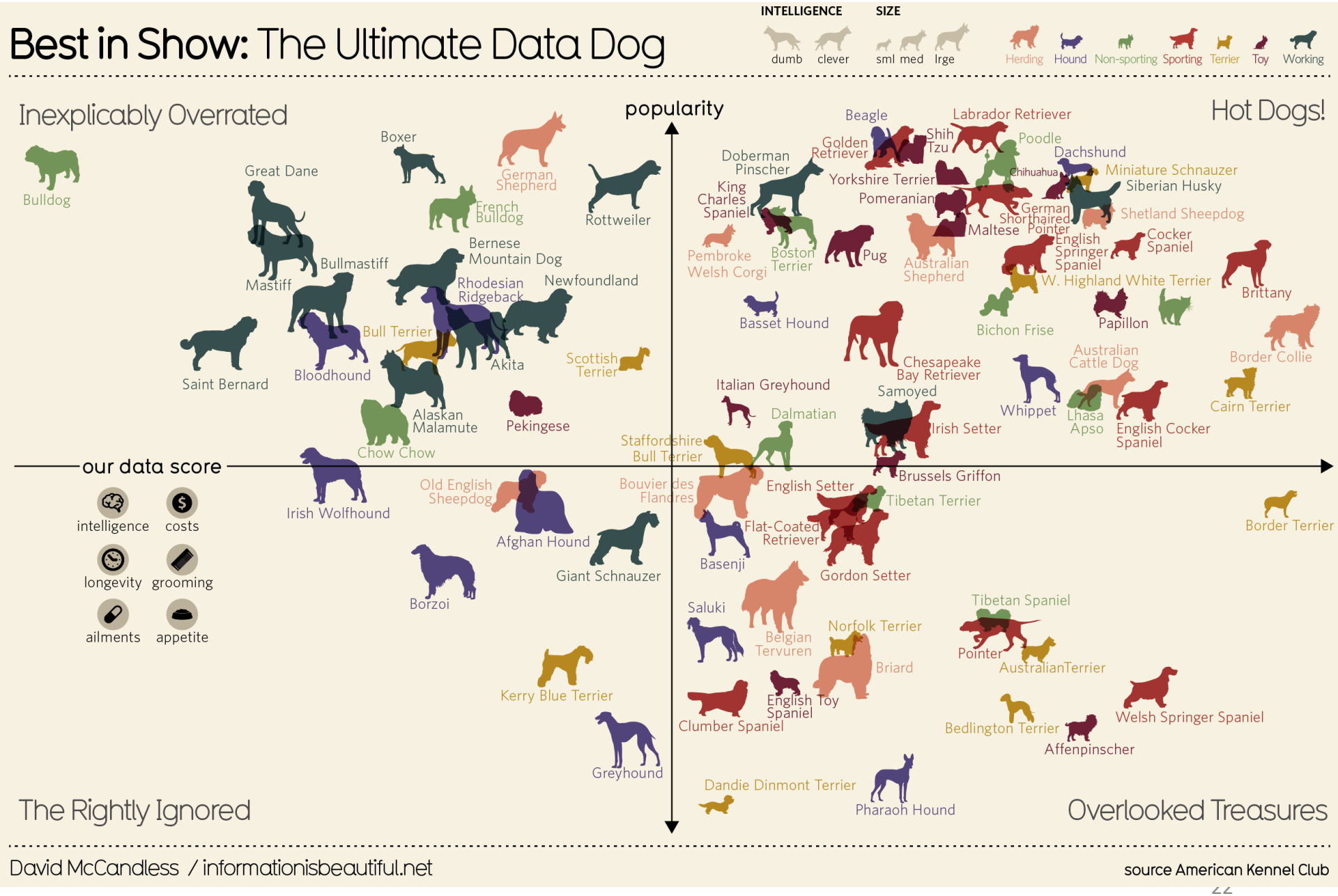


Iconic representation

- Glyph (graphical object) represents a data case
- Visual properties of glyph represent different variables



Best in Show: The Ultimate Data Dog

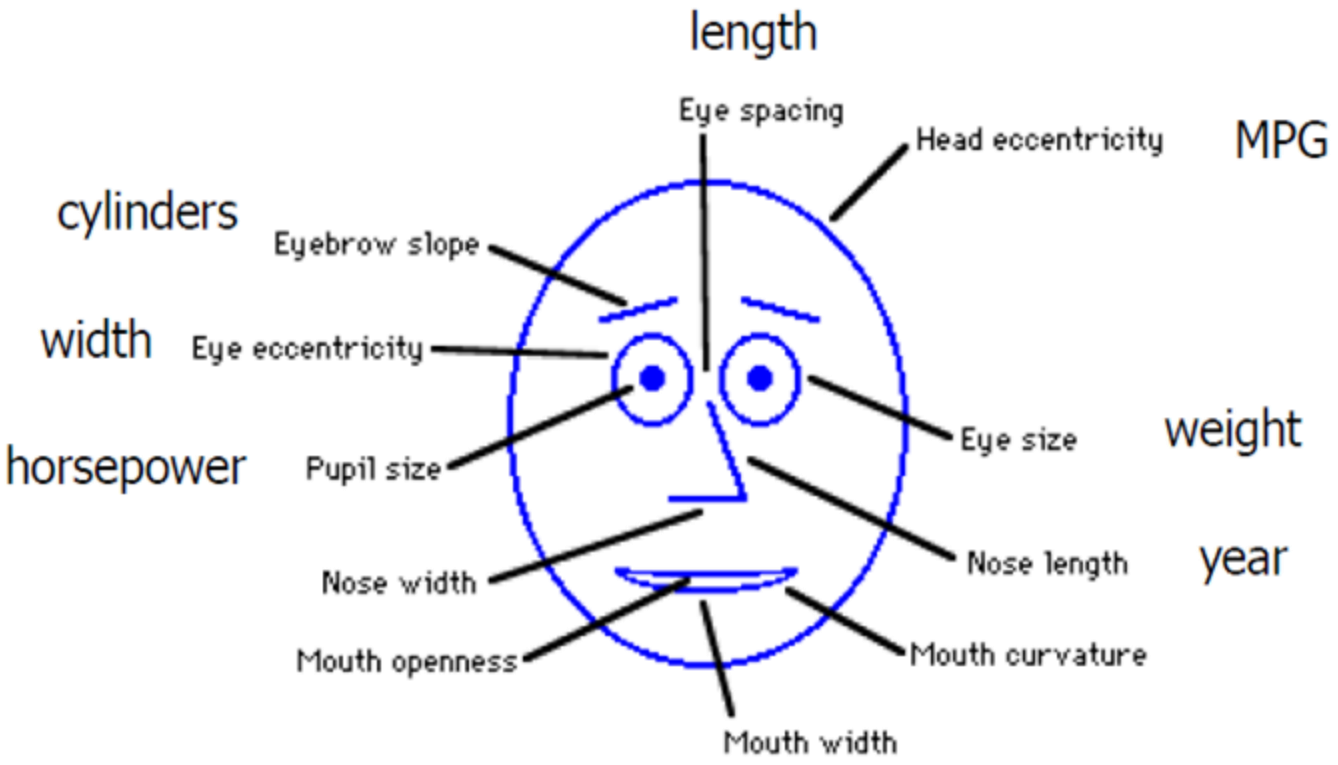


David McCandless / informationisbeautiful.net

source American Kennel Club

Iconic representation

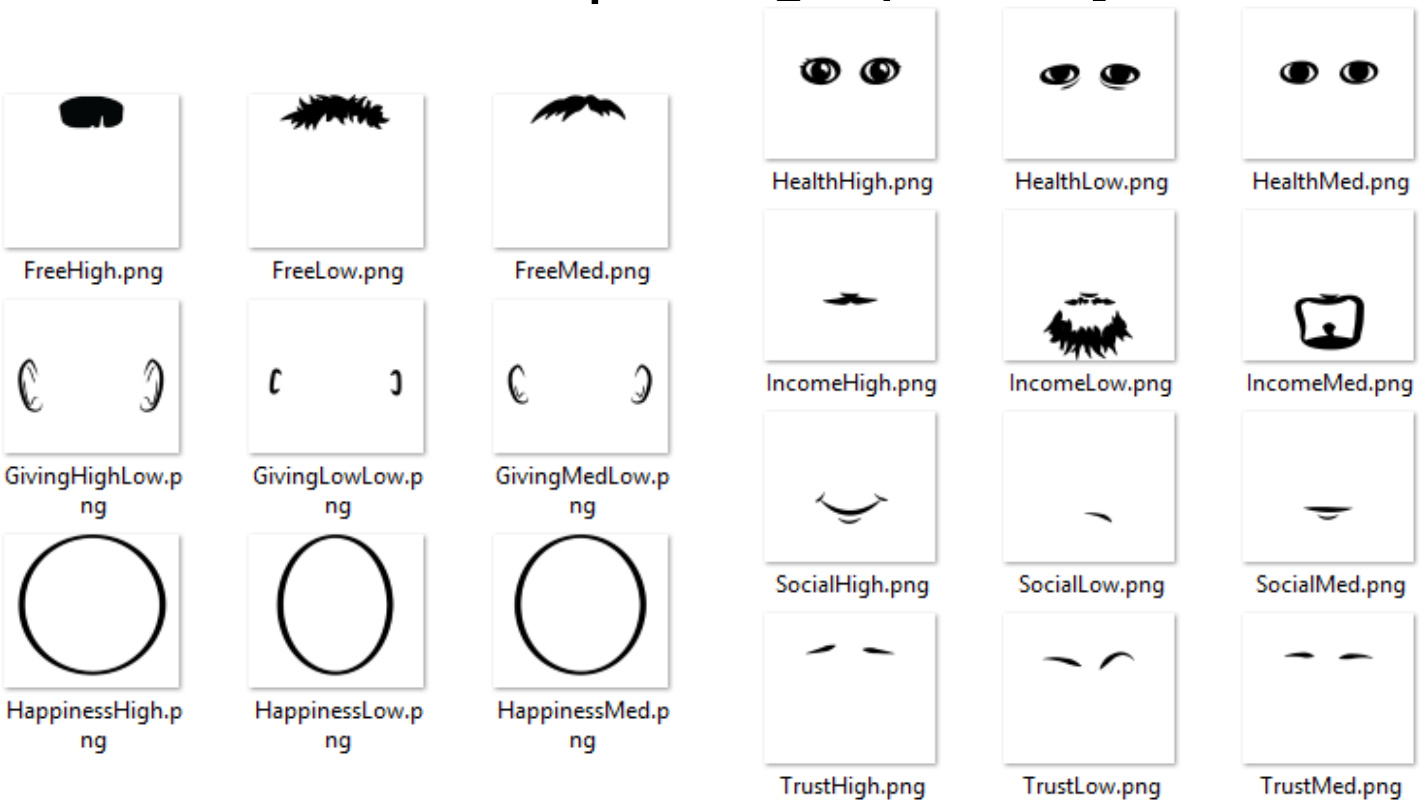
- Chernoff Faces: use face to represent points in K-dimensional space graphically



Chernoff, 1973

Iconic representation

- Chernoff Faces: use face to represent points in K-dimensional space graphically



Credit: https://www.esri.com/arcgis-blog/products/arcgis-pro/mapping/chernoff-faces/

Iconic representation

https://truth-and-beauty.net/projects/oecd-better-life-index

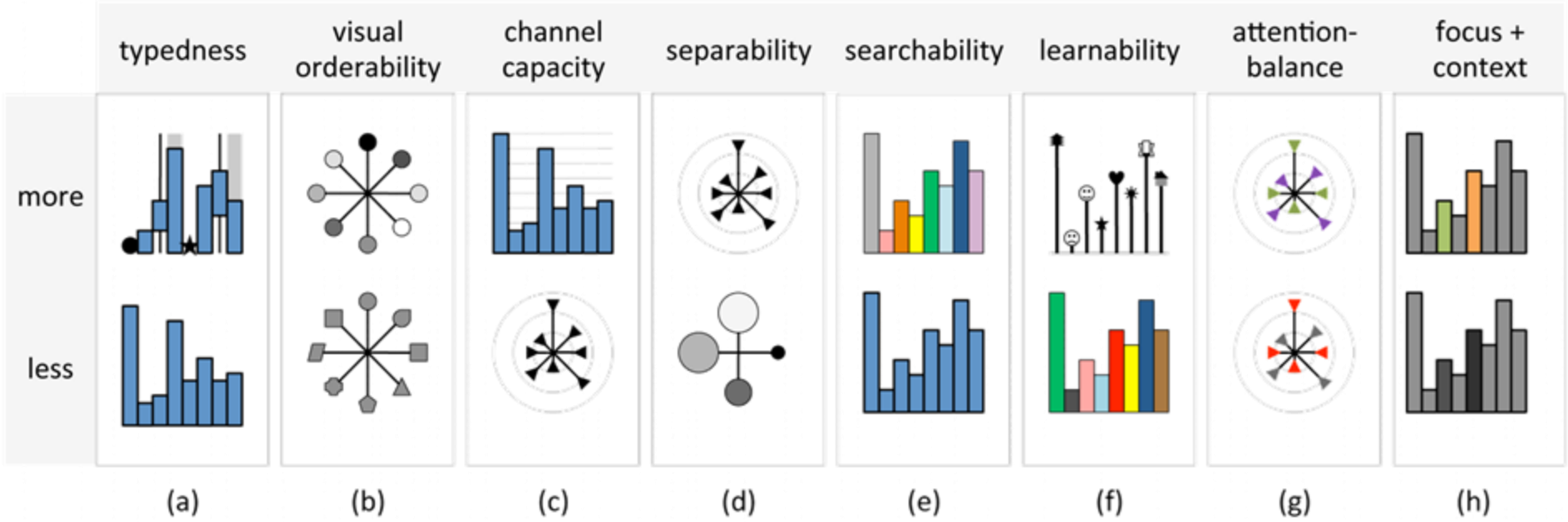
Iconic representation

- Space variables around a circle.
- Encode values on “spokes”.

https://truth-and-beauty.net/projects/oecd-better-life-index

Iconic representation

- Glyph design criteria



Borgo et al., 2013

Table lens

- Spreadsheet is certainly one high-dimensional data presentation
- Idea:
  - Make the text more visual and symbolic
  - Just leverage basic bar chart idea

Expense	Jan-18	Feb-18	Mar-18	Apr-18	May-18
Phone	$ 46.0	$ 47.0	$ 56.0	$ 65.0	$ 58.0
Insurance	$ 80.0	$ 80.0	$ 80.0	$ 80.0	$ 80.0
Rent	$ 900.0	$ 900.0	$ 900.0	$ 900.0	$ 900.0
Medicine	$ 120.0	$ 60.0	$ 87.0	$ 90.0	$ 55.0
Electric Bill	$ 200.0	$ 180.0	$ 145.0	$ 170.0	$ 140.0
Water Bill	$ 120.0	$ 100.0	$ 99.0	$ 110.0	$ 120.0
Total	**$ 1,466.0**	**$ 1,367.0**	**$ 1,3[illegible]7.0**	**$ 1,415.0**	**$ 1,353.0**

Change quantitative values to bars

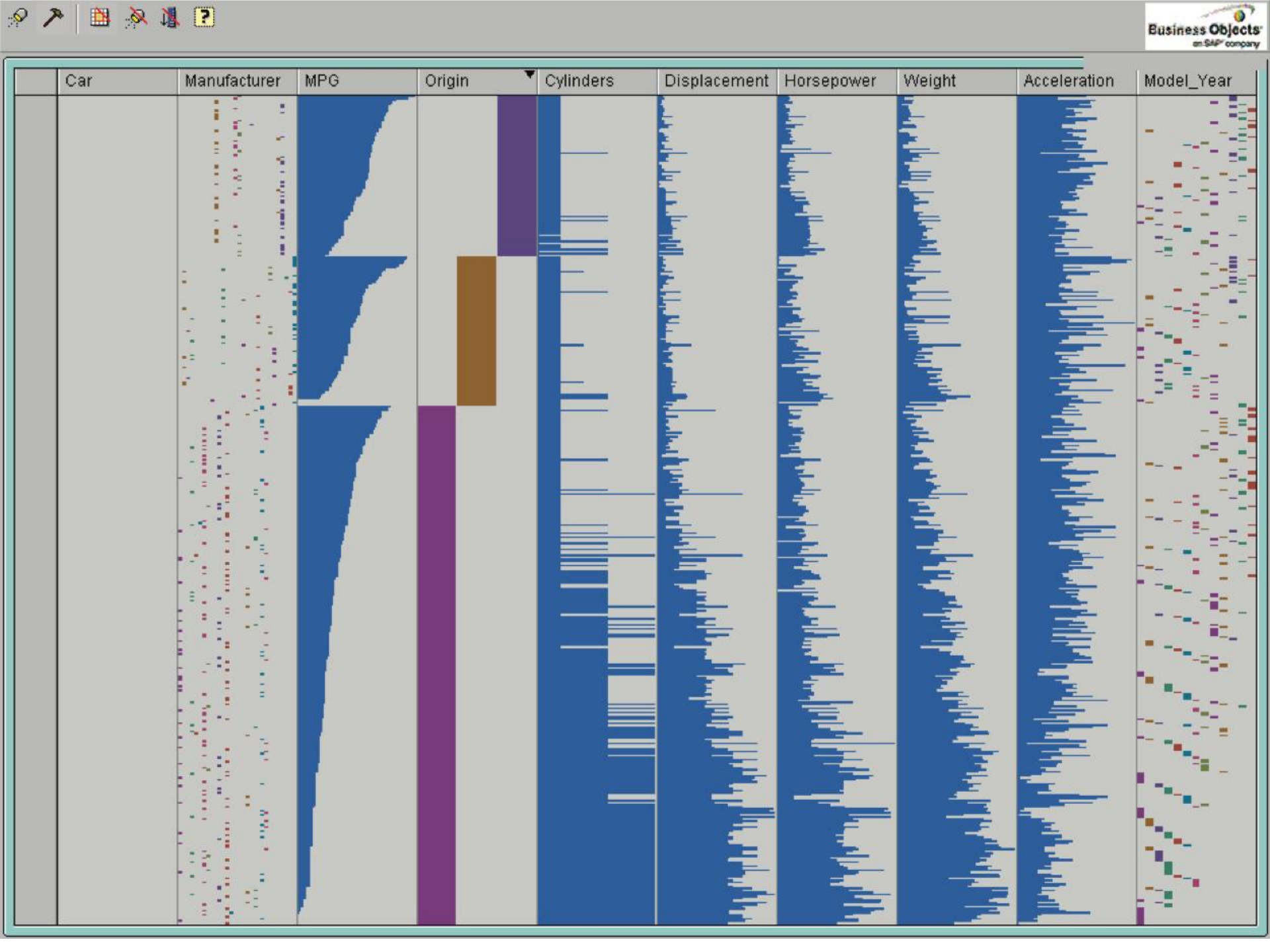
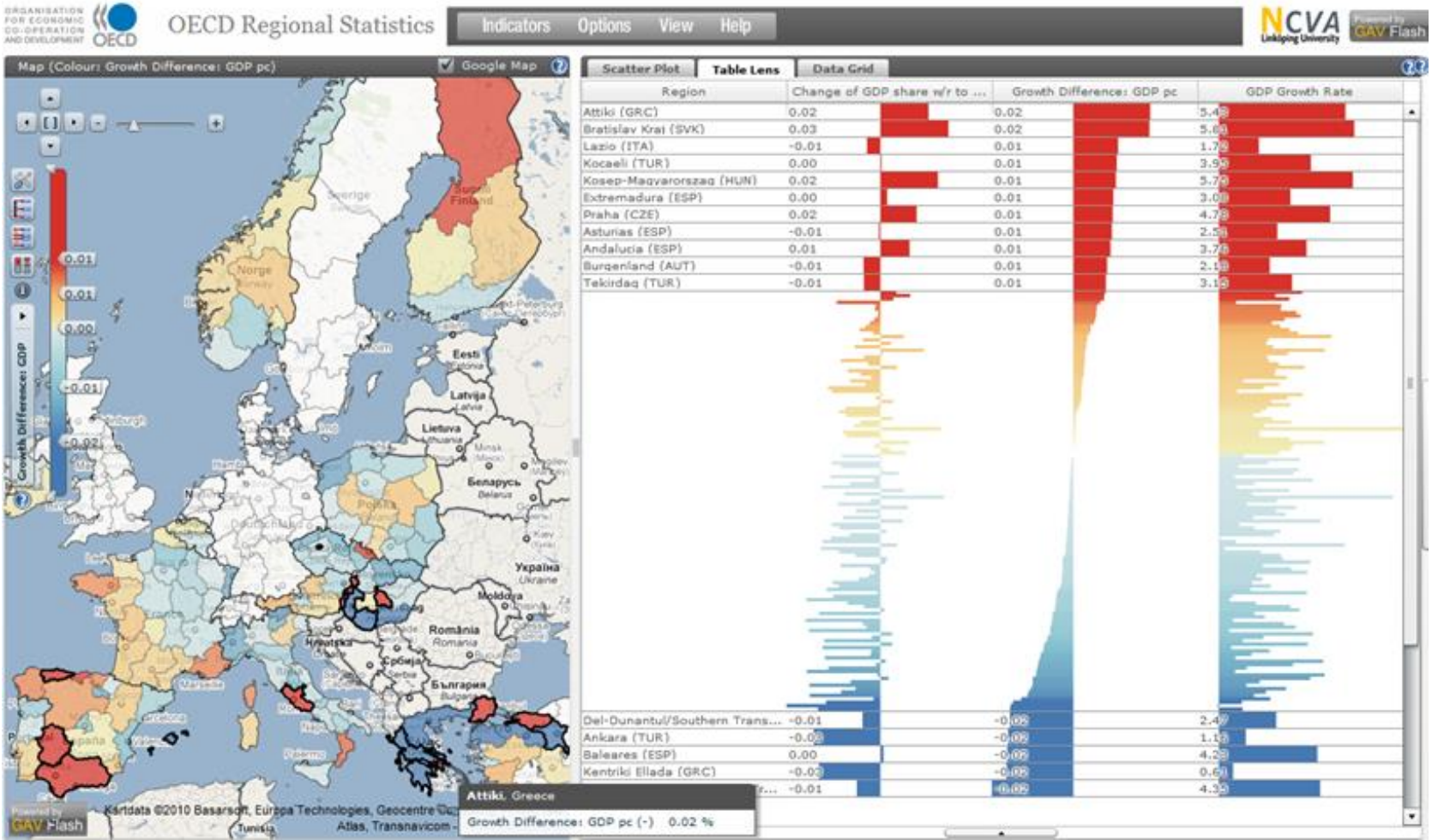
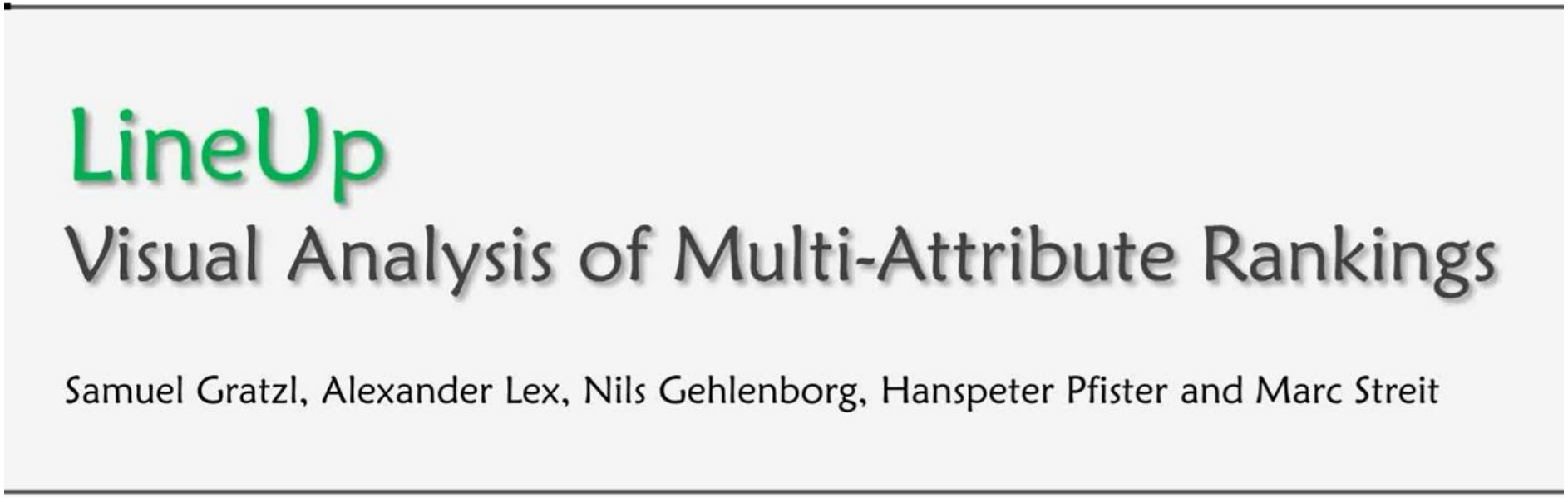

Business Objects
an SAP company
Car
Manufacturer
MPG
Origin
Cylinders
Displacement
Horsepower
Weight
Acceleration
Model_Year

Table lens

- Geo visual analytics



LineUp

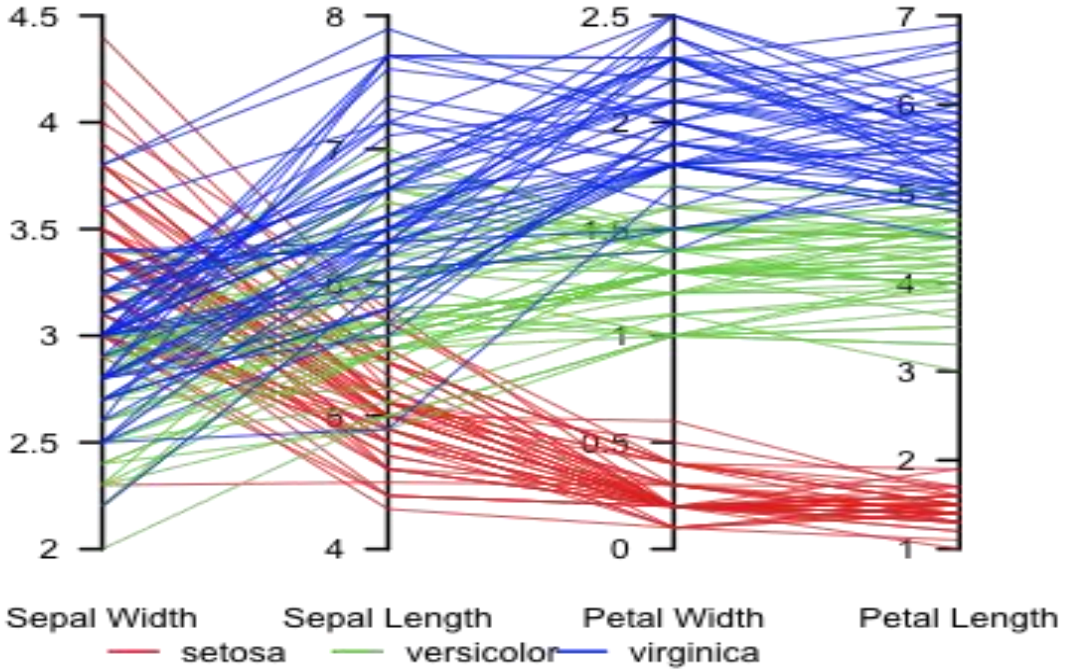

LineUp
Visual Analysis of Multi-Attribute Rankings
Samuel Gratzl, Alexander Lex, Nils Gehlenborg, Hanspeter Pfister and Marc Streit
CALEYDO
JKU
JOHANNES KEPLER
UNIVERSITÄT LINZ
HARVARD
School of Engineering
and Applied Sciences
HARVARD
MEDICAL SCHOOL

Limitations

- How about categorical data attributes?
  - Gender: female, male
  - Country: China, US, UK, France
  - University: HKUST, CUHK, HKU
- Run out of rows/columns for lots of data cases
- How about an alternative generic representation?

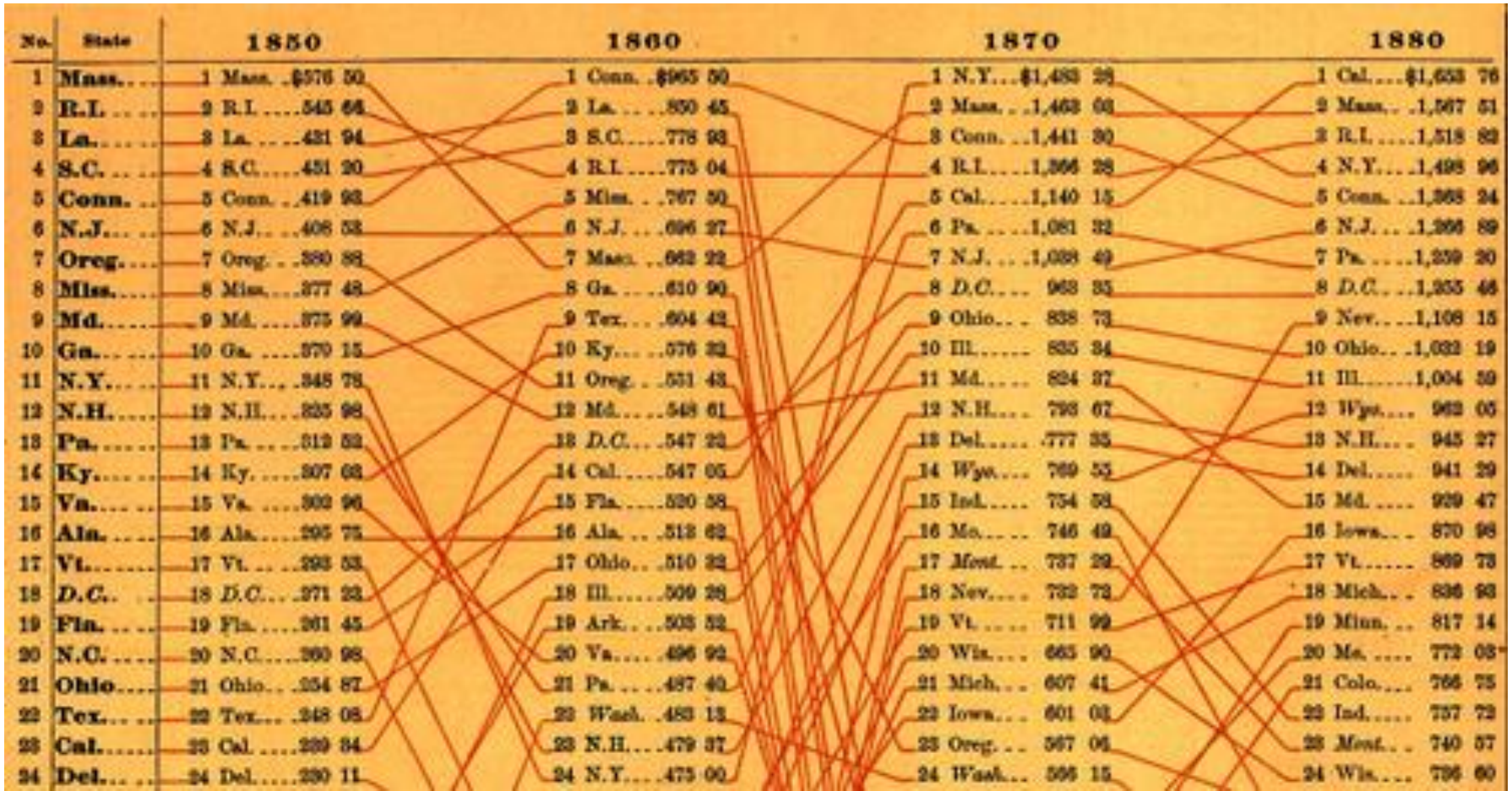
Parallel coordinates

- Each column of space is assigned a variable
- Vertical line specifies different values that variable can take
- Each data case is a polyline that puts a vertex on each column at its corresponding data value

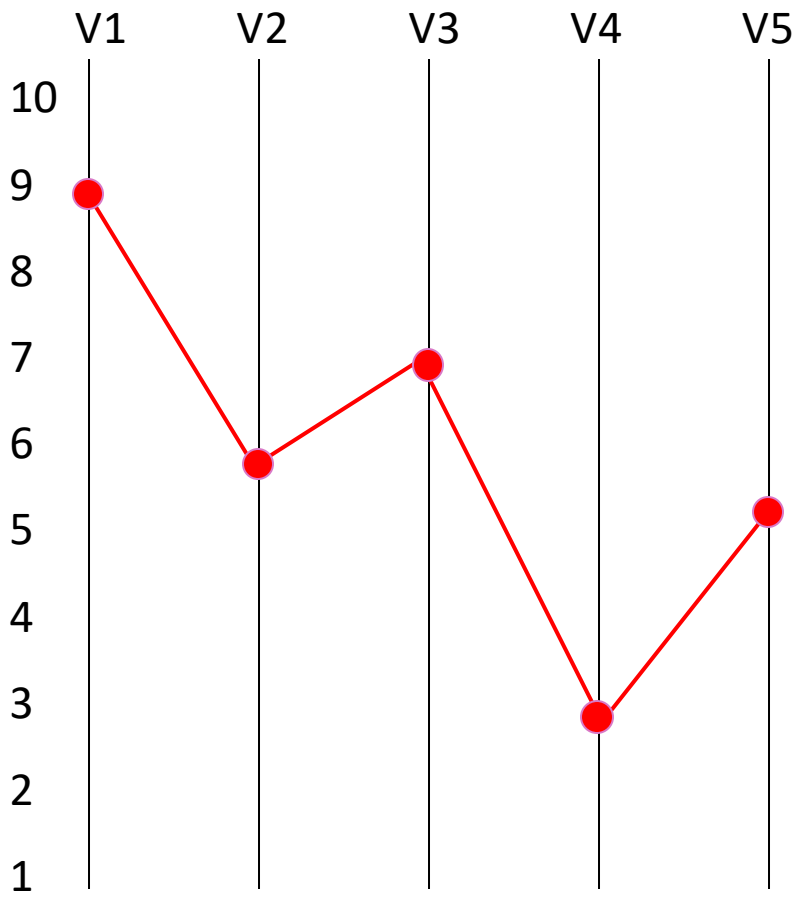


Parallel coordinates

- Early history 1880



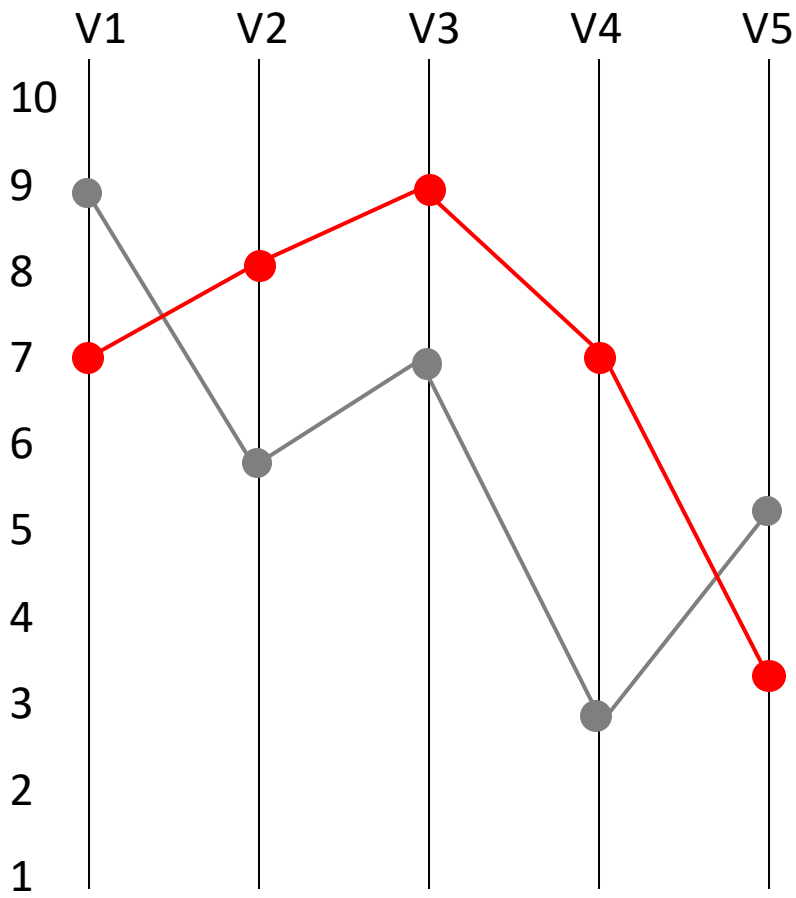
Parallel coordinates



- Case 1

V1	V2	V3	V4	V5
9	6	7	3	5

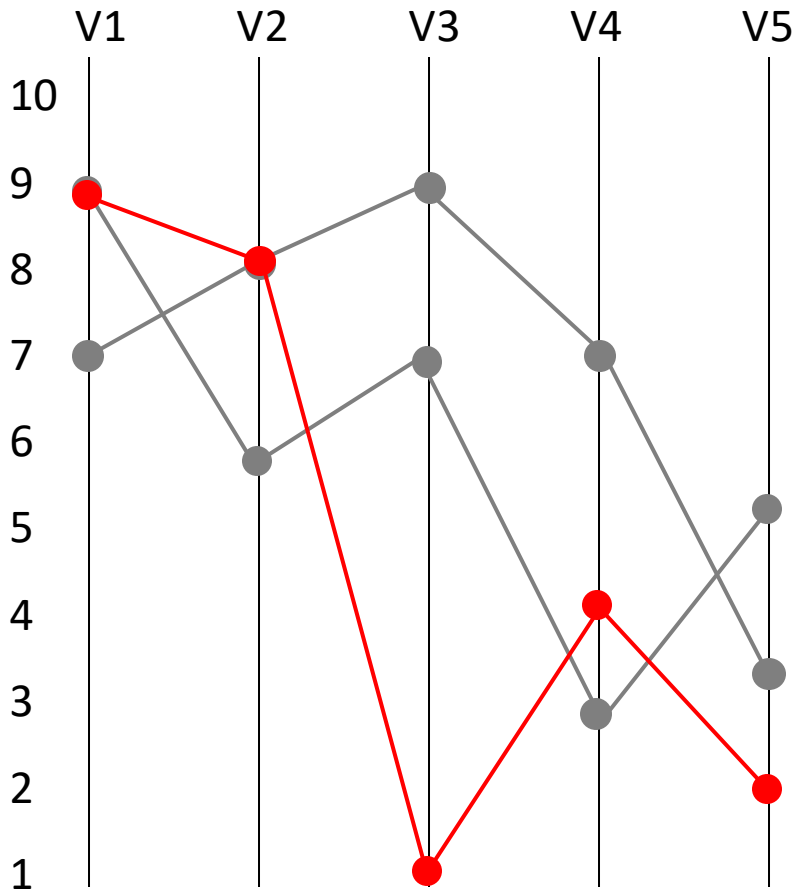
Parallel coordinates



- Case 2

V1	V2	V3	V4	V5
9	6	7	3	5
7	8	9	7	3

Parallel coordinates

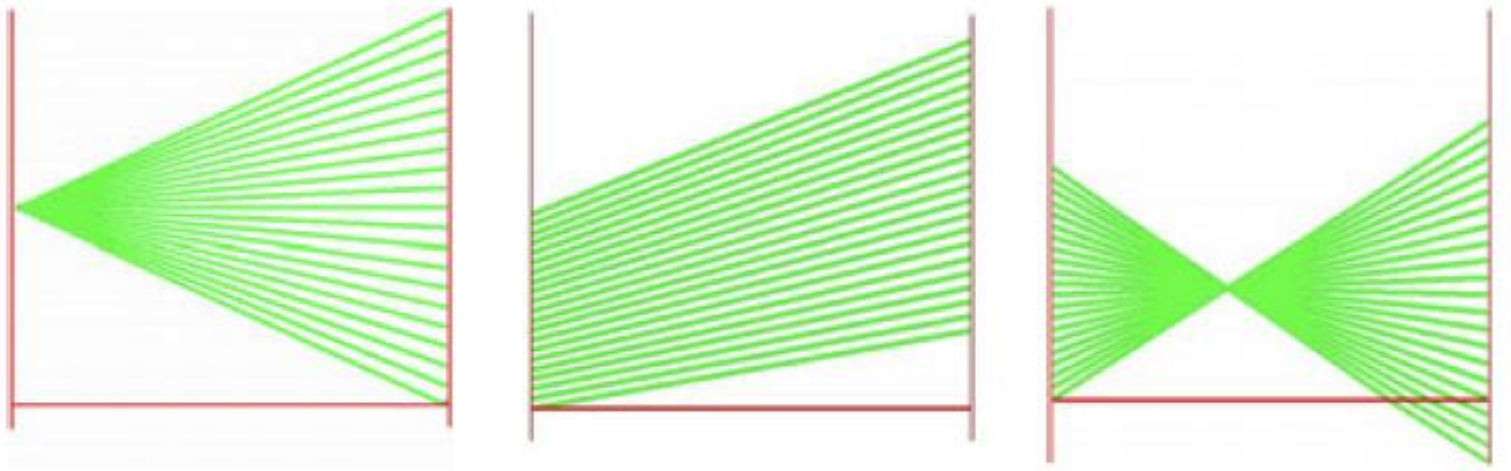


- Case 3

V1	V2	V3	V4	V5
9	6	7	3	5
7	8	9	7	3
9	8	1	4	2

Parallel coordinates

- What do two correlated variables look like?
- What do two inversely correlated variables look like?

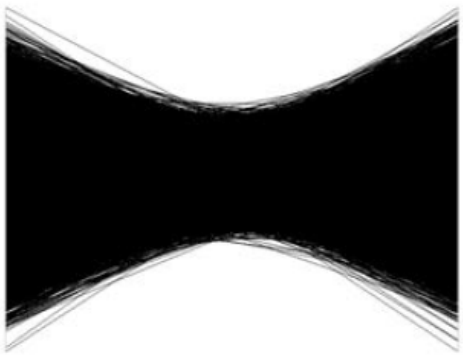


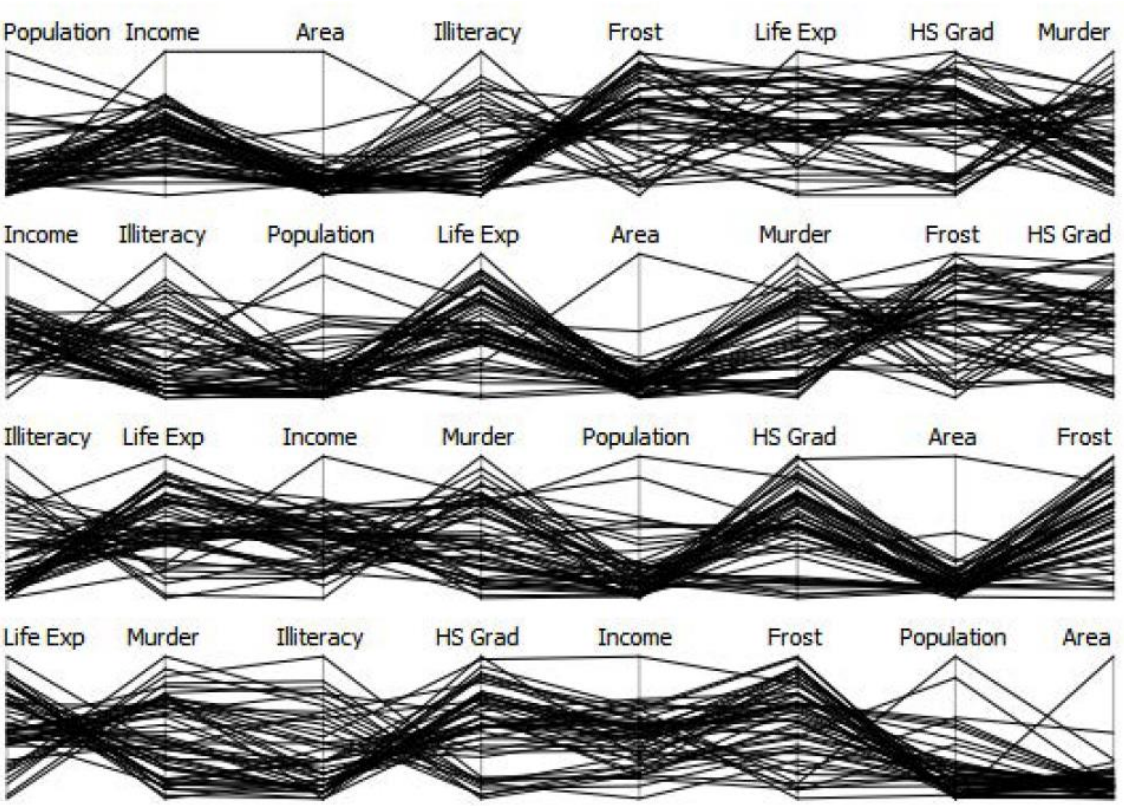
constant positive linear negative linear

http://www.parallelcoordinates.de/blog/tutorial/2015/08/01/point-line-duality/

Parallel coordinates

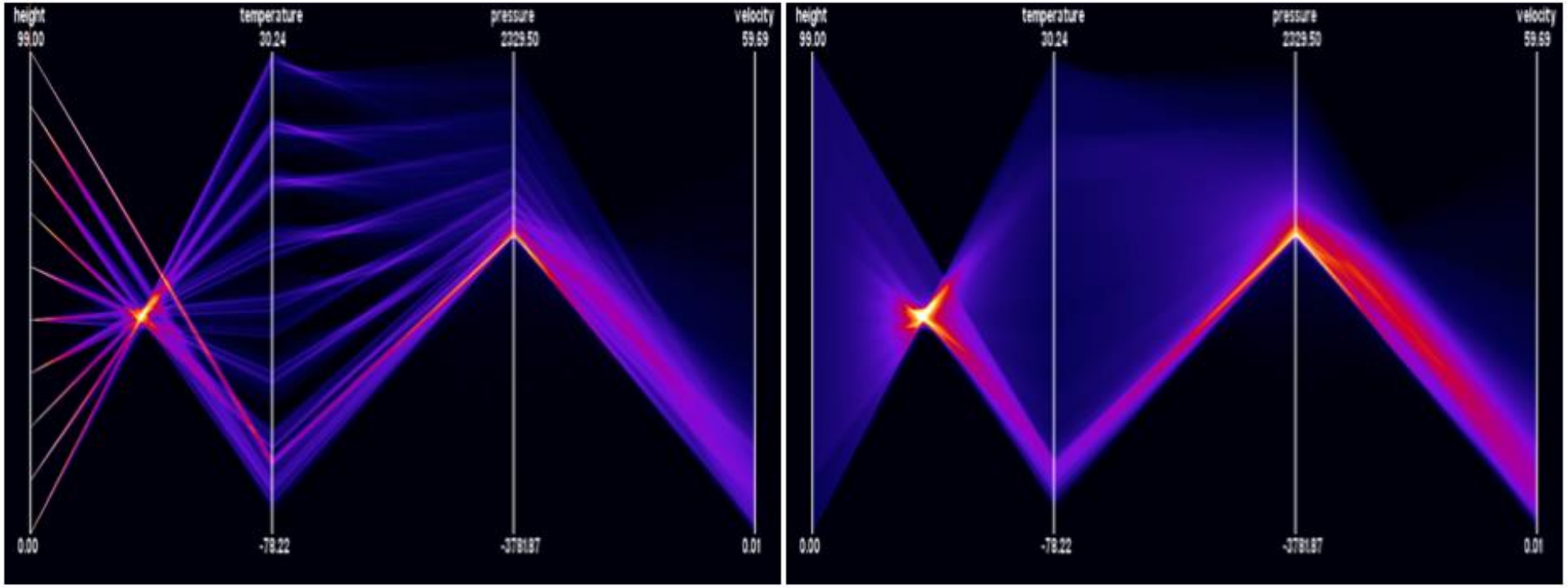
- Challenges
  - Too much data
  - Order of dimensions really matters





Parallel coordinates

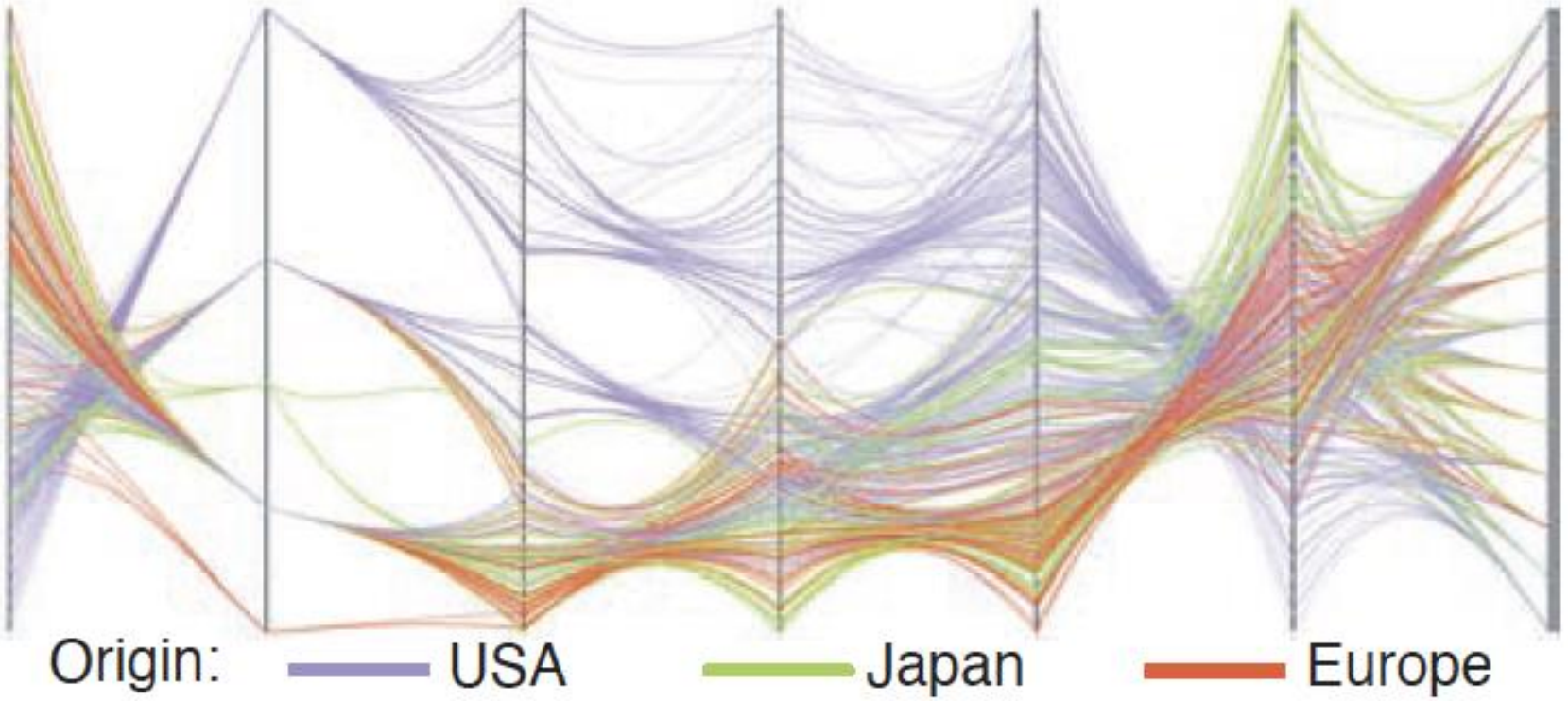
- Density field model for parallel coordinates



Heinrich and Weiskopf, 2009

Parallel coordinates

- Edge bundling for parallel coordinates

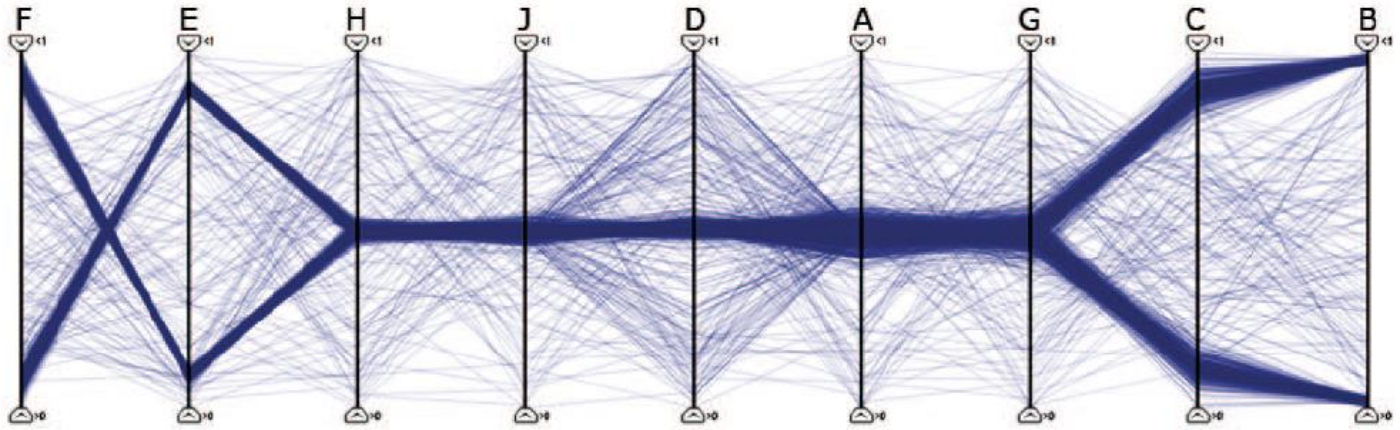


Zhou et al., 2008

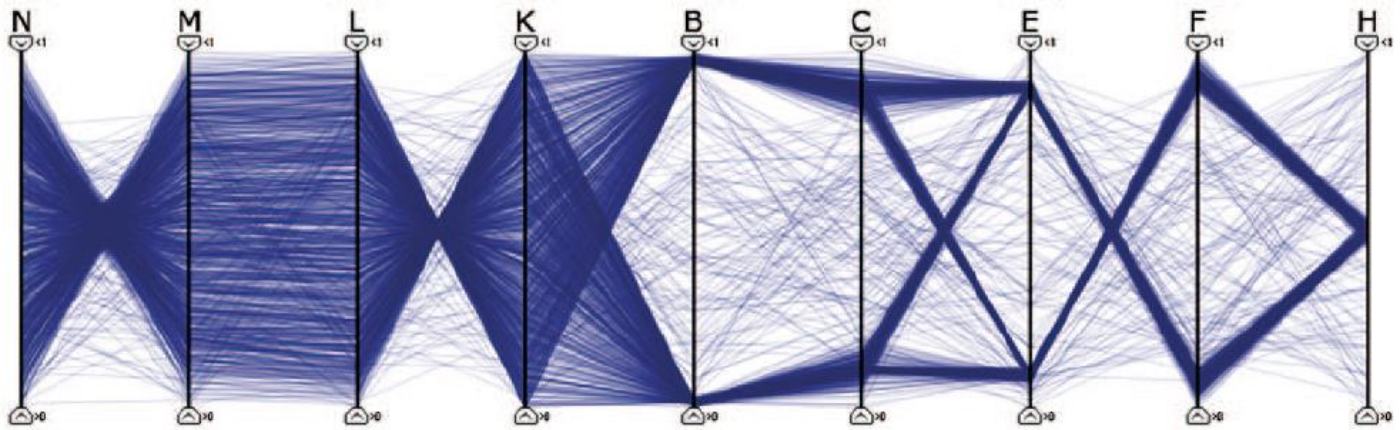
Parallel coordinates

- Sort the axes

Best order
for clustering

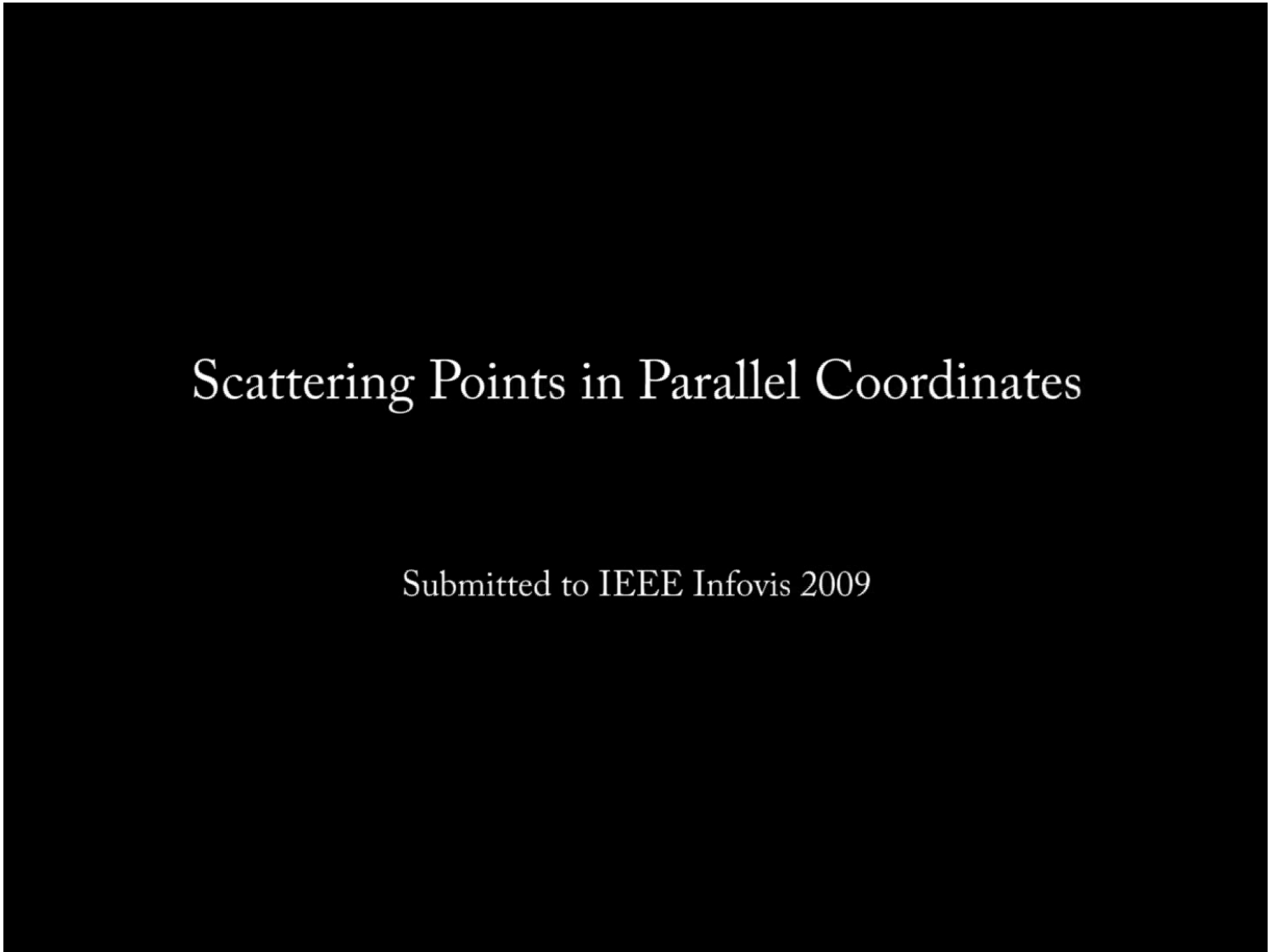


Best order
for correlation



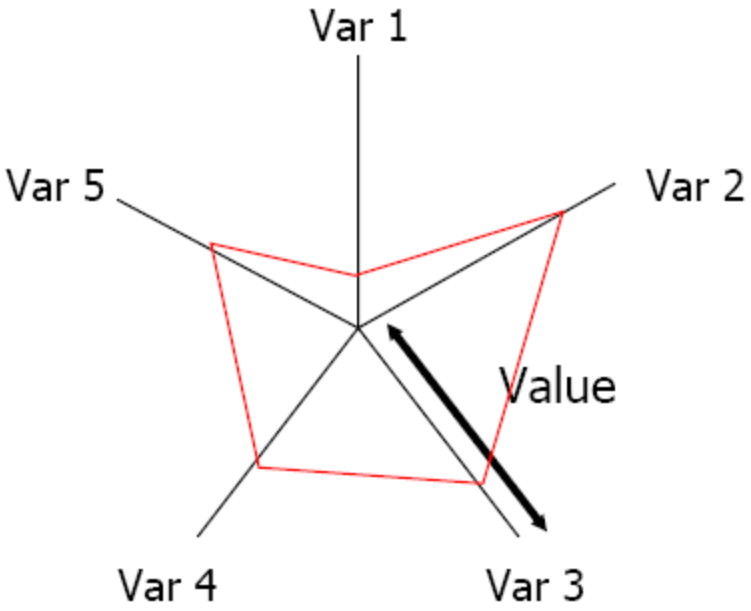
Bertini et al., 2011

Parallel coordinates

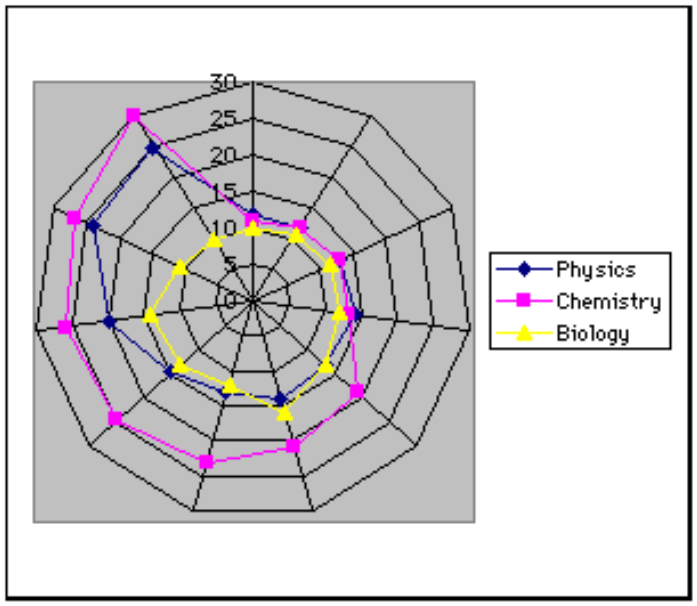


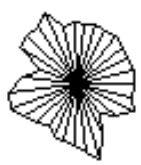
Star plot

- Space out the n variables at equal angles around a circle
- Each “spoke” encodes a variable’s value
- Data point is now a “shape”

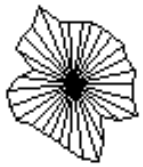


Star plot

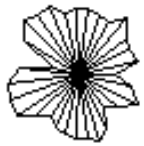




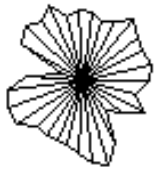
Connecticut



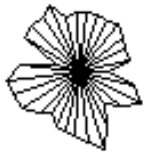
New Hampshire



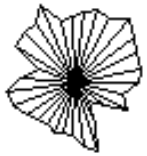
Pennsylvania



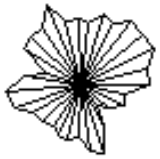
Maine



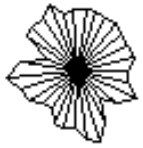
New Jersey



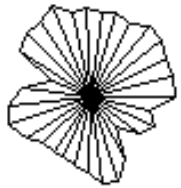
Rhode Island



Massachusetts



New York



Vermont

Data Exploration & Visualization

Module 11: High-Dimensional Data Visualization

- Data dimension
  - univariate, bivariate, trivariate, high-dimensional
- High-dimensional data visualization
  - Visual mapping: multiple views, scatterplot matrix, iconic representation, table lens, parallel coordinates
  - Dimension reduction:
    - Linear: PCA, MDS, LDA
    - Non-linear: t-SNE, UMAP

Dimension 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)

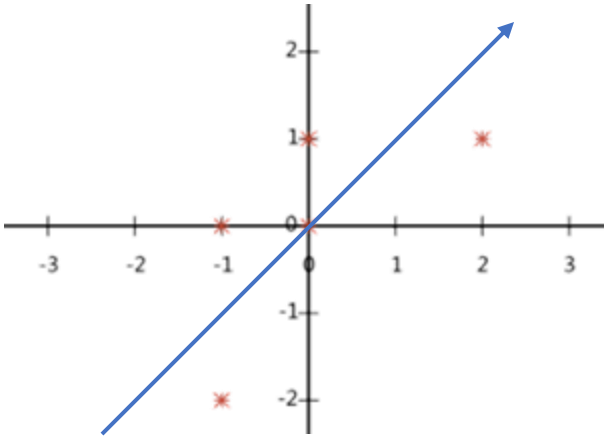
- Nonlinear methods:
  - t-SNE
  - UMAP

PCA Motivation I

- Data set has two dimensions.

$$\begin{pmatrix} 1 & 1 & 2 & 4 & 2 \\ 1 & 3 & 3 & 4 & 4 \end{pmatrix}$$ 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.

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

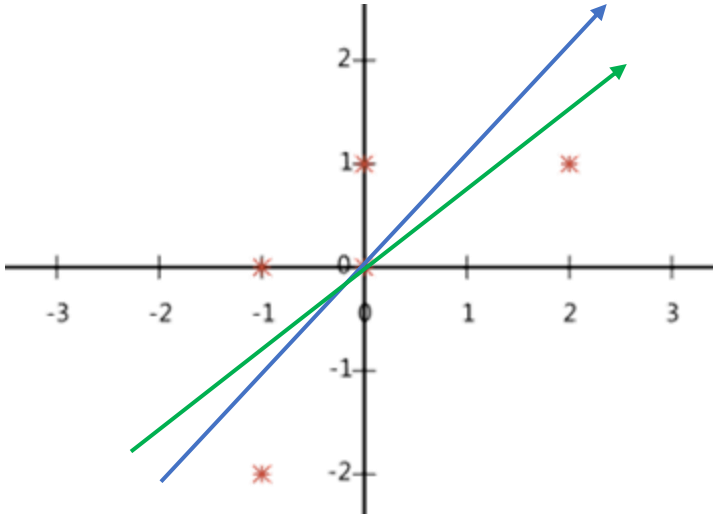
- Let the average in each dimension be 0, i.e., ā = 0.

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

- 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 $\bar{a} = 0, \quad \bar{b} = 0$, 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:

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

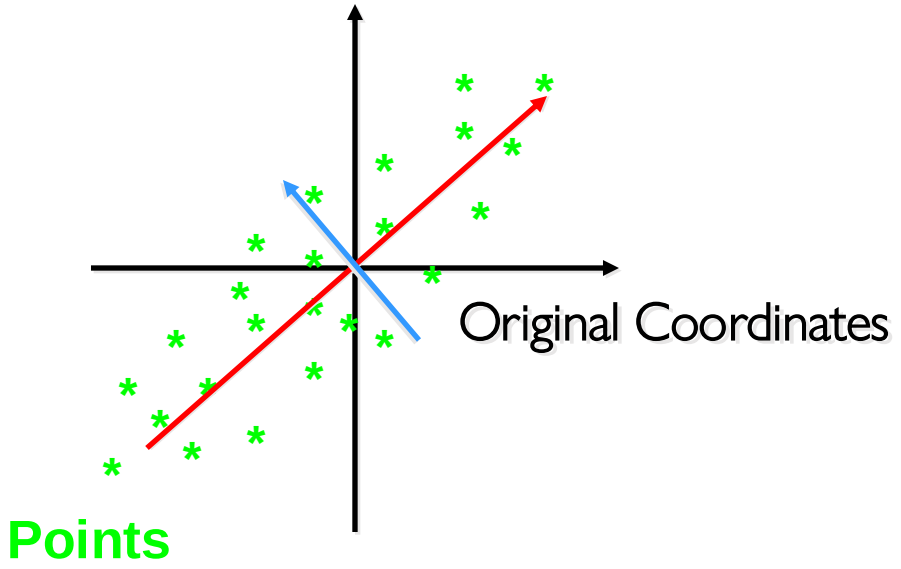
- The covariance matrix can be obtained by

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

Principal Component Analysis

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

X : n x m matrix.
n dimensions, m points.
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:

$$\frac{1}{m}\sum_{i=1}^{m}(PX_i)^2 = \frac{1}{m}\sum_{i=1}^{m} PX_i{X_i}^T P^T \quad \rightarrow \quad PSP^T$$

$$S = \frac{1}{m}\sum_{i=1}^{m} X_i{X_i}^T$$ 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

$$PSP^T = \lambda$$

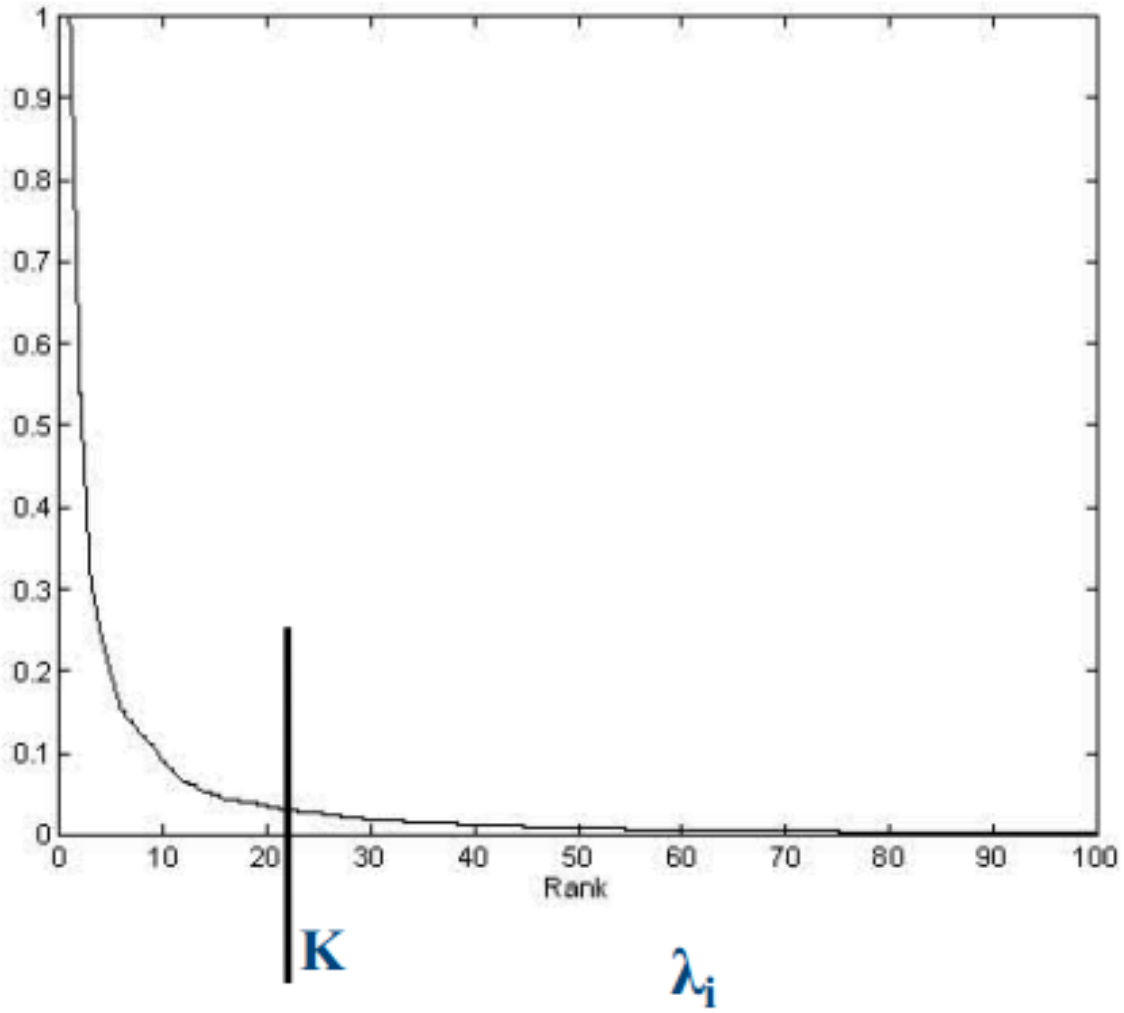
- The variance after projection is the eigenvalue of the covariance matrix.
  - To maximize the variance, choose the largest eigenvalue.
  - The largest eigenvalue is the best coordinates.

PCA Mechanics

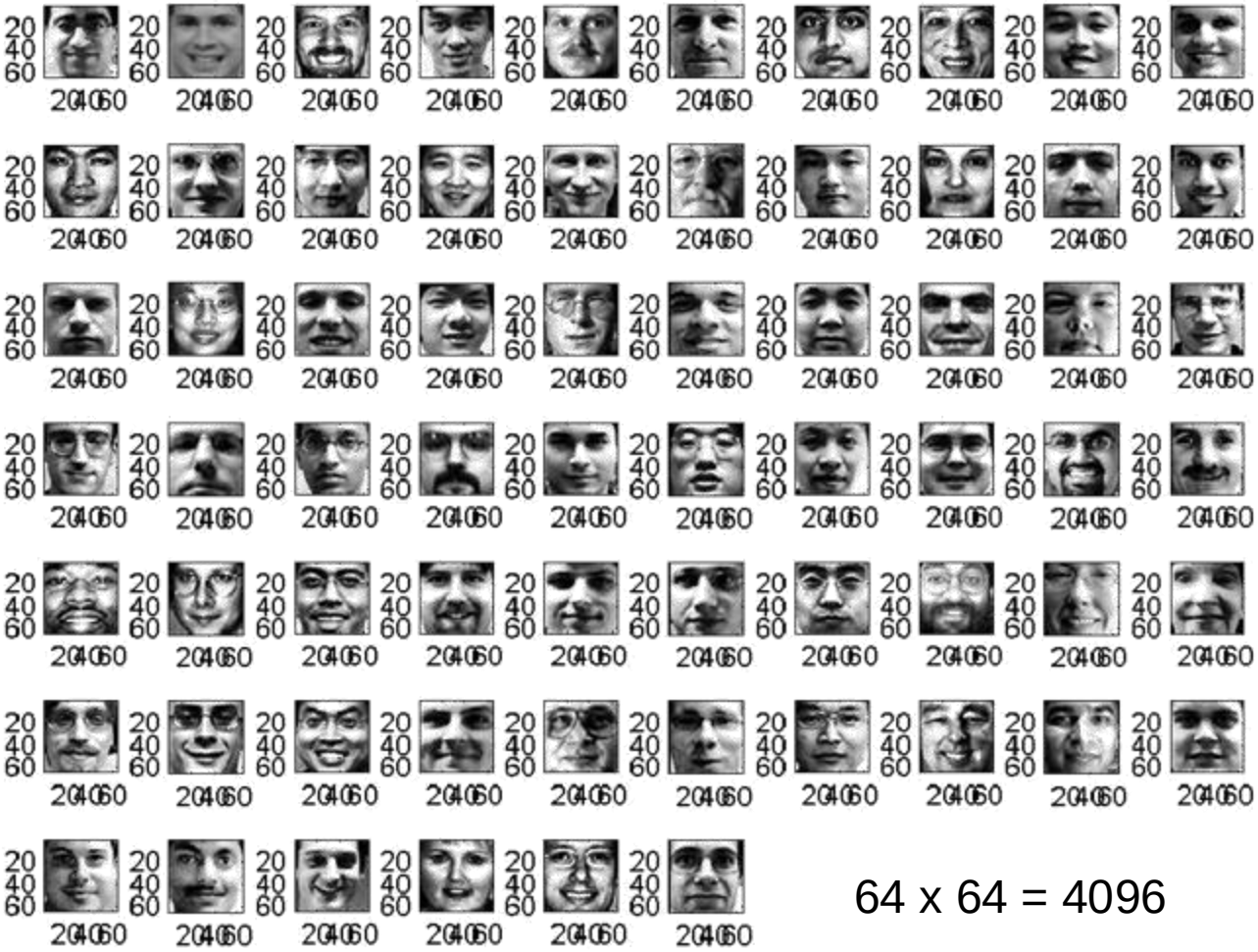
Suppose $x_1, x_2, \ldots, x_M$ are $H \times 1$ vectors:

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 $A = [\Phi_1 \Phi_2 \cdots \Phi_M]$.
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

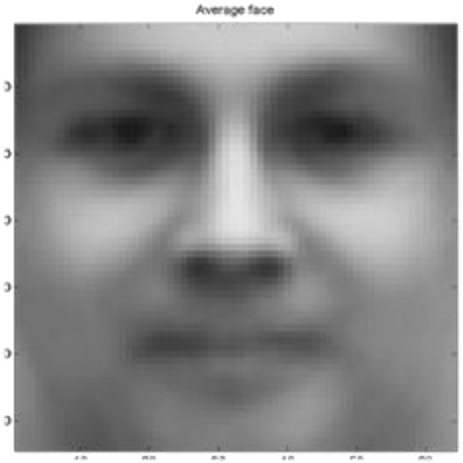


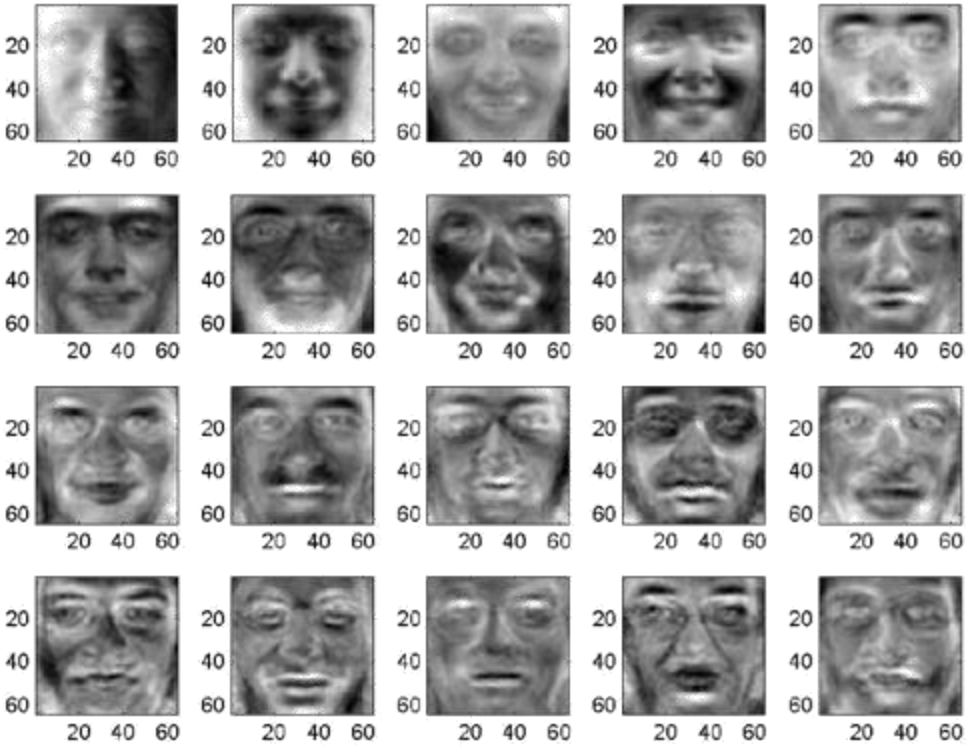
64 x 64 = 4096

Turk et al., 1991

PCA Applied to Faces

- Reconstruct each face as a linear combination of “basis faces”, or Eigenfaces.



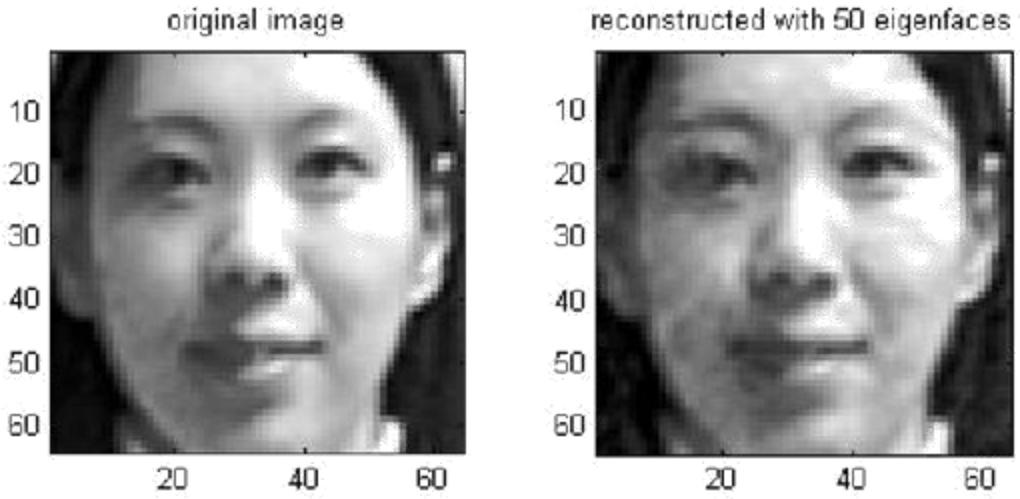


Eigenfaces

Turk et al., 1991

Reconstruction

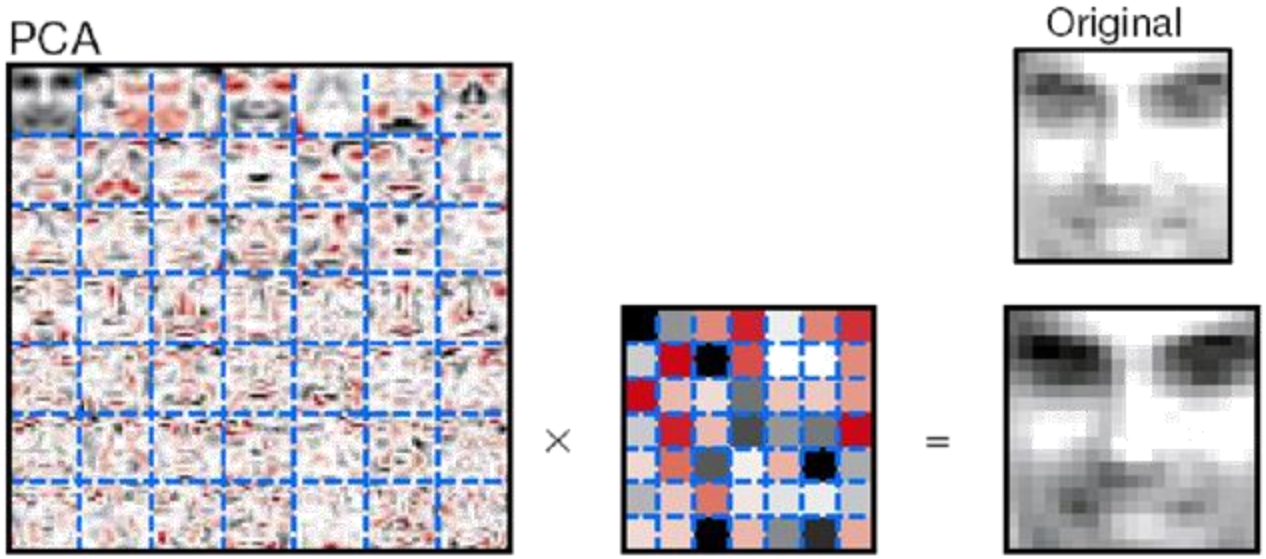
- 90% variance is captured by the first 50 eigenvectors.
- Reconstruct existing faces using only 50 basis images.



Turk et al., 1991

Issues of PCA

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



Turk, Matthew A and Pentland, Alex P. [Face recognition using eigenfaces]. Proceedings CVPR'91.

Multidimensional Scaling (MDS)

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

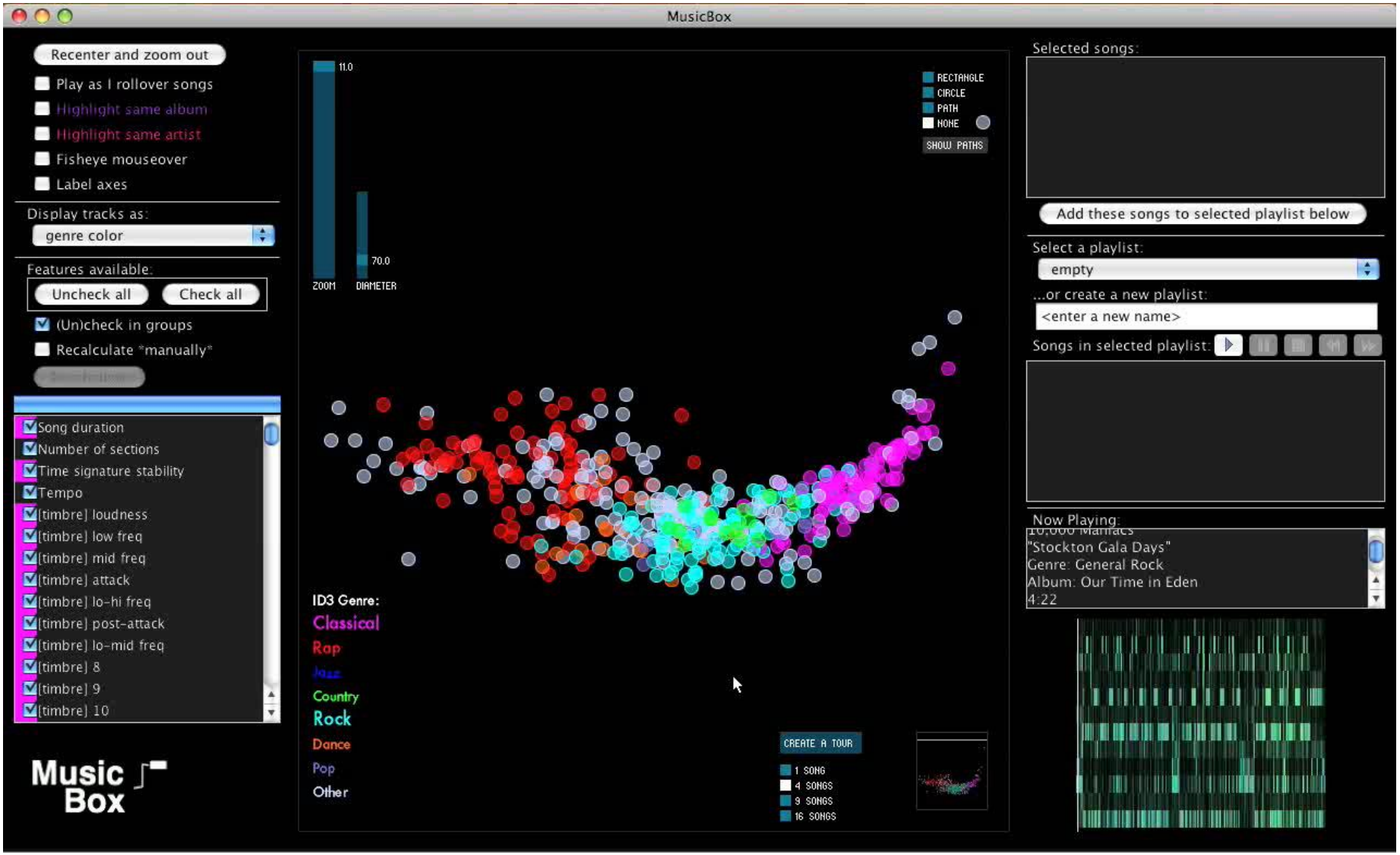
An Example: US Map

- 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

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

Latent Space Cartography