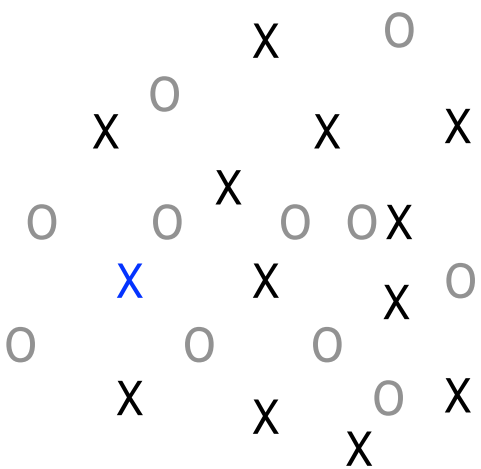
Data Visualization

The

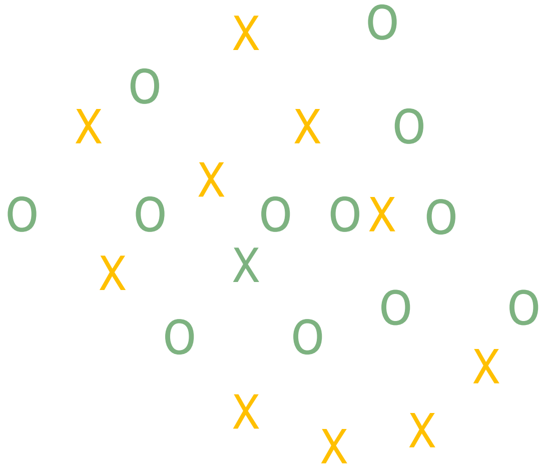
Foundations

The Modes of Perception

Find the outlier



Fast

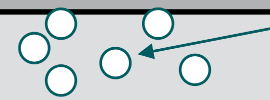
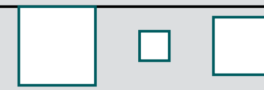
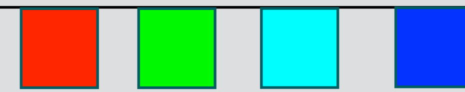
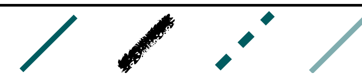
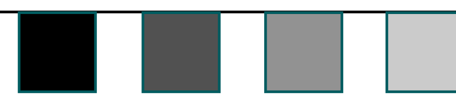


Slow

The Modes of Perception

- Pre-attentive
  - fast
  - parallel processing
  - effortless
- Pattern recognition
  - semi-fast
  - governed by laws of Gestalt
- Attentive
  - slow
  - sequential
  - high effort (attention is a very limited resource)

Main Properties of Graphics

Category	Example
Position	
Shape	
Size	
Color	
Orientation (Line)	
Length (Line)	
Type and Size (Line)	
Brightness	

Main Properties of ~~Graphics~~ Humans

Category	Amount of pre-attentive information
Position	*very high*
Shape	———
Size	*approx. 4*
Color	*approx. 8*
Orientation (Line)	*approx. 4*
Length (Line)	———
Type and Size (Line)	———
Brightness	*approx. 8*

Pre-attentive Perception

- Position
    - fast
    - effective
    - high number of different positions
- Color
    - use with care
- Shape
- Orientation

Pre-attentive perception is effortless.
Exploit this as much as you can.

Pattern Detection

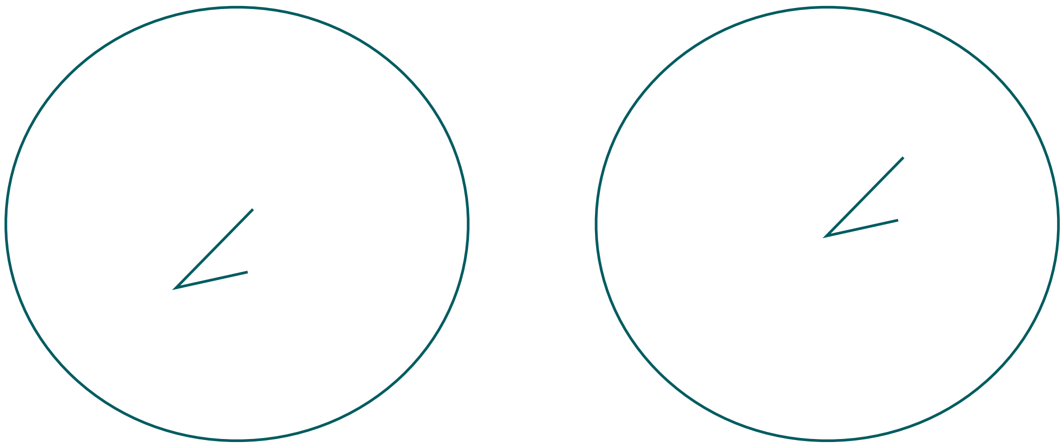
„It is interesting to note that our brain […] subconsciously always prefers meaningful situations and objects.“

- Emergence
- Reiification
- Multi-stability
- Invariance

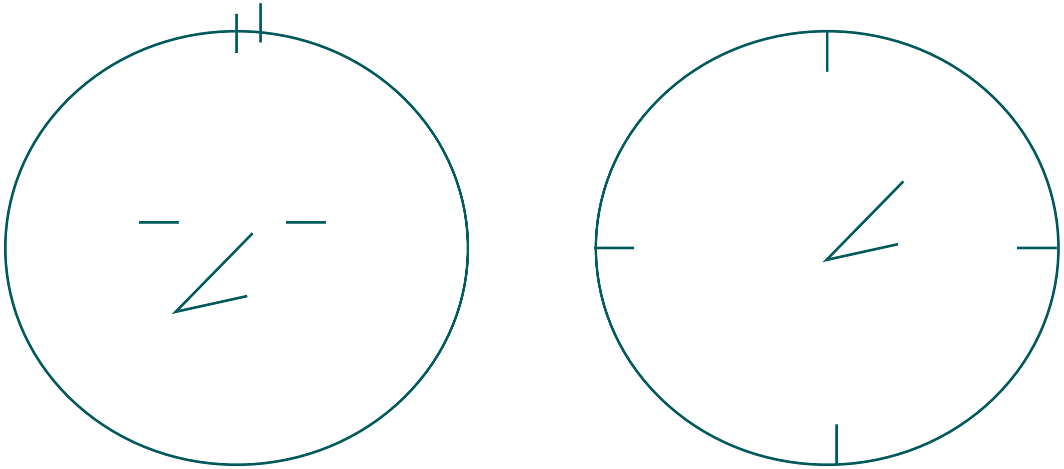
Pattern detection can be trained.
Exploit this for frequent visualizations.

What is this?

What is this?

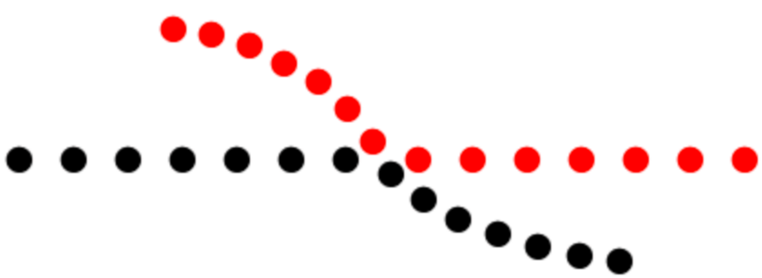


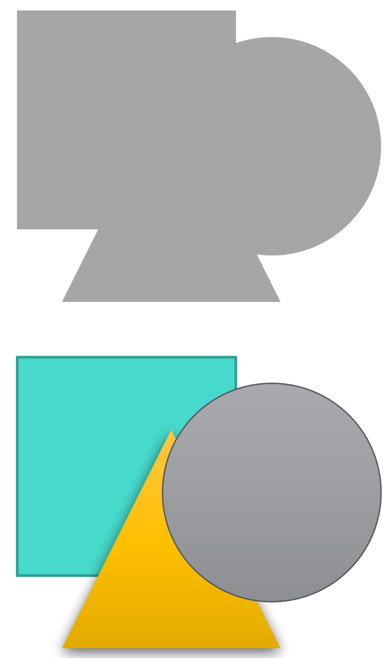
What is this?

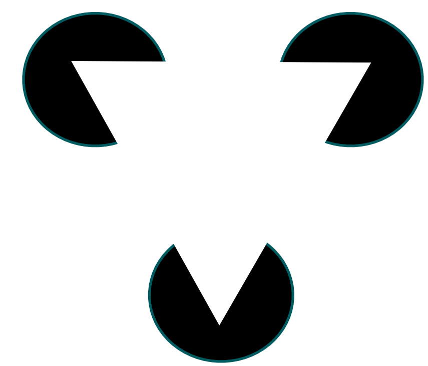


Laws of Gestalt

„It is interesting to note that our brain,
in accordance with the laws of Gestalt,
subconsciously always prefers meaningful situations and objects.“

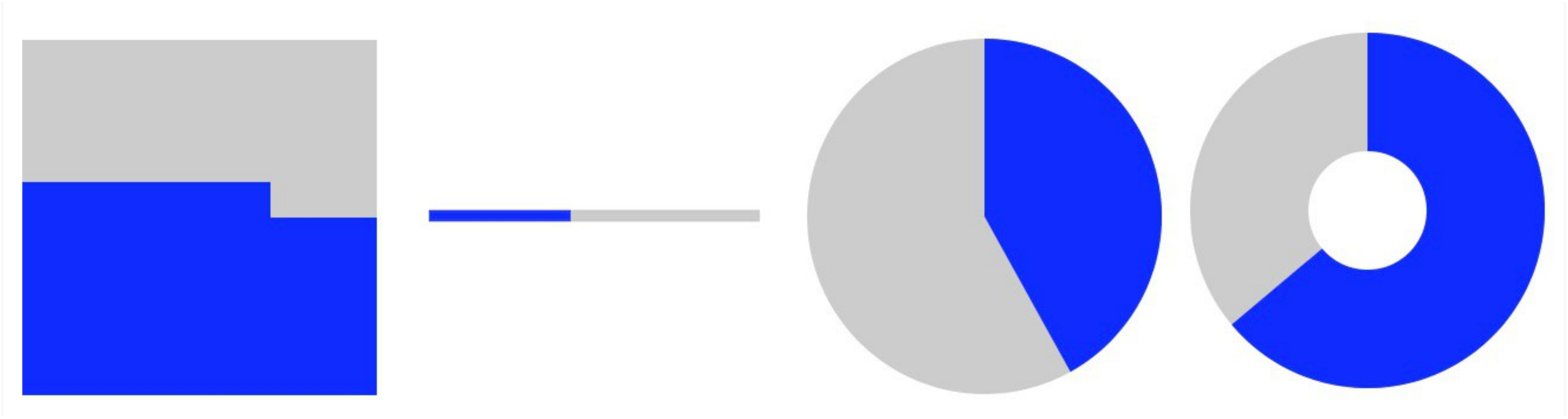






Accuracy of Graphics

Square Pie vs Stacked Bar vs Pie vs Donut



What do you think?

https://eagereyes.org/blog/2016/a-reanalysis-of-a-study-about-square-pie-charts-from-2009

Why do we visualize data?

- Explore
  - use different techniques
  - avoid „construction“ bias
  - be careful with „aesthetics“
  - challenge findings -> use attentive mode

- Explain
  - focus on message or „story“
  - use pre-attentive mode

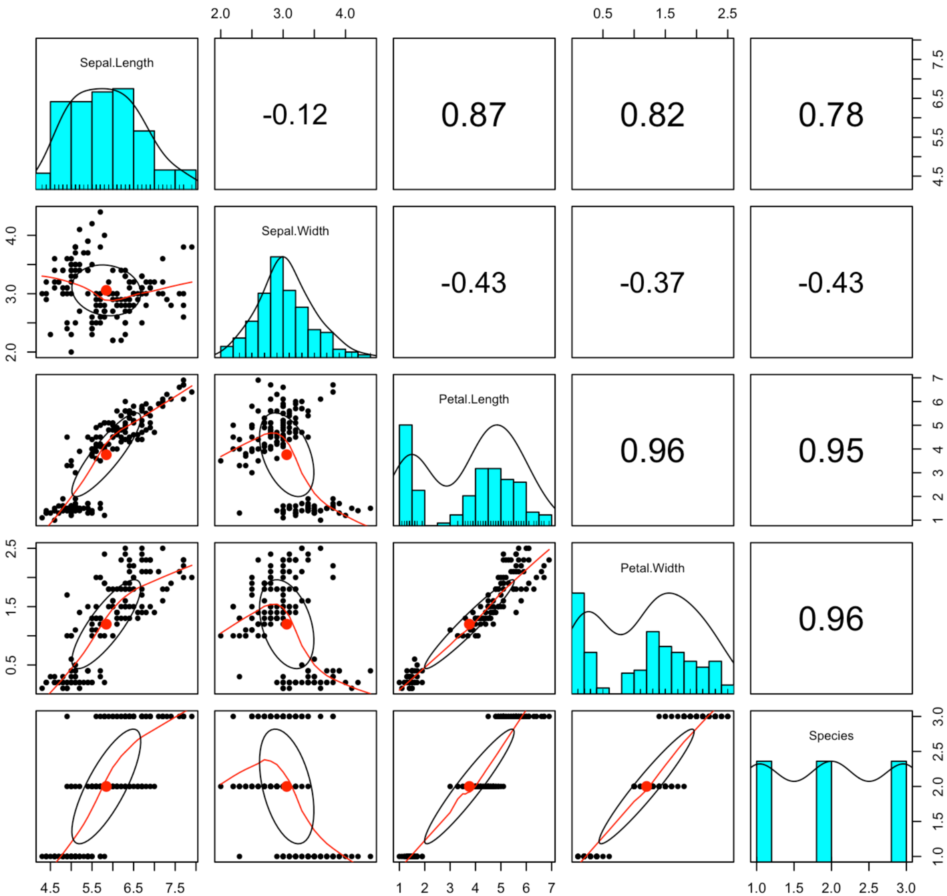
Don’t trust your perception ;-)

Data Visualization

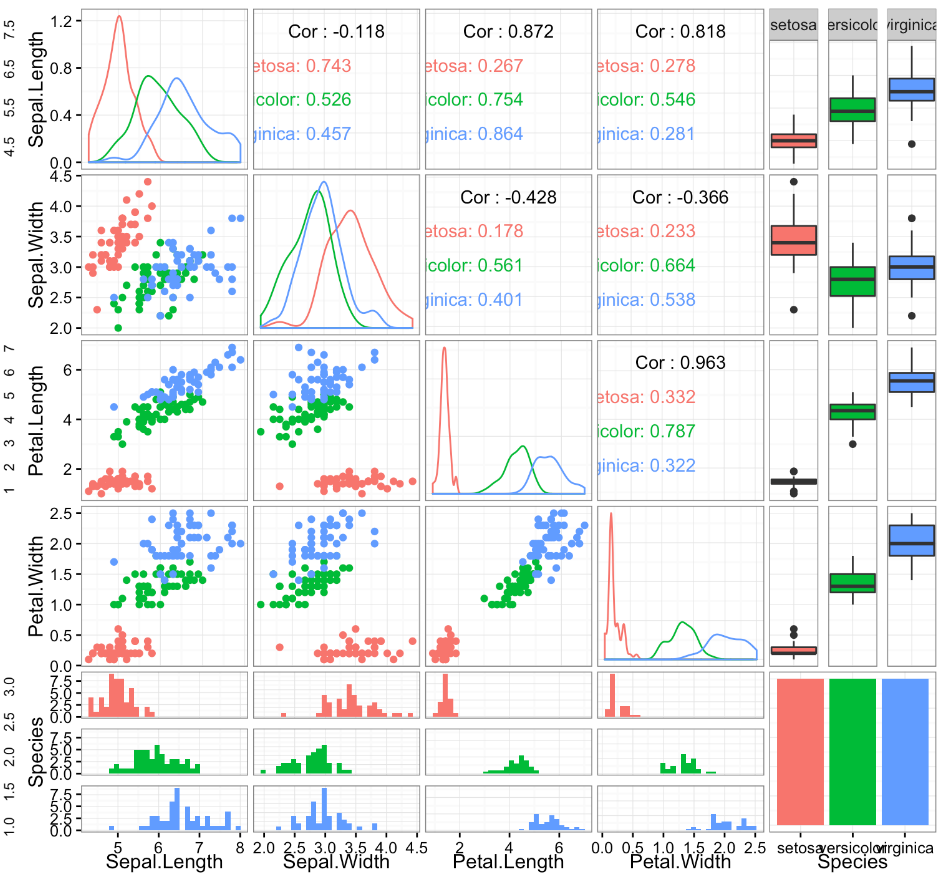
In

Higher Dimensions

Pair Plots

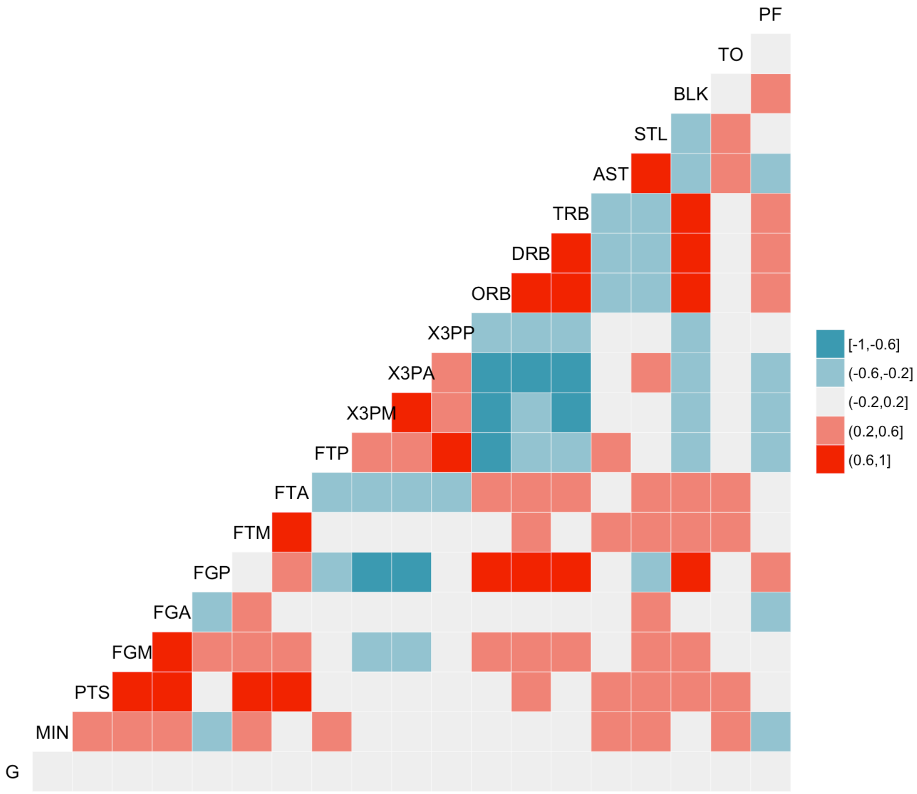


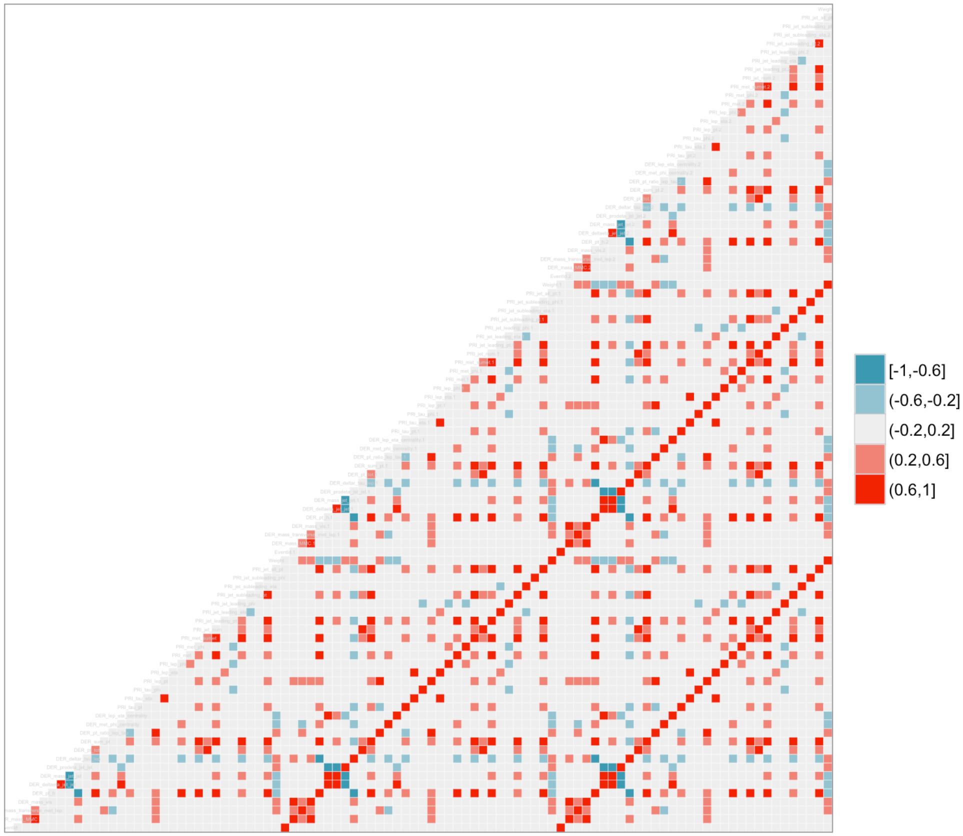
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species
-0.12
0.87
0.82
0.78
-0.43
-0.37
-0.43
0.96
0.95
0.96



Cor : -0.118
Cor : 0.872
Cor : 0.818
Cor : -0.428
Cor : -0.366
Cor : 0.963
setosa
ersicolo
irginica
Sepal.Length
Sepal.Width
Petal.Length
Petal.Width
Species

Correlation Plots





Fundamental Problems

- No accurate method in higher dimensions
  - Approximation methods
  - „Simulated“ dimensions (color, size, shape)
  - Animations?
- No notion of quality or accuracy for visualizations
  - Information Theory?
  - „Stability“?

All visualizations are wrong, but some are useful.

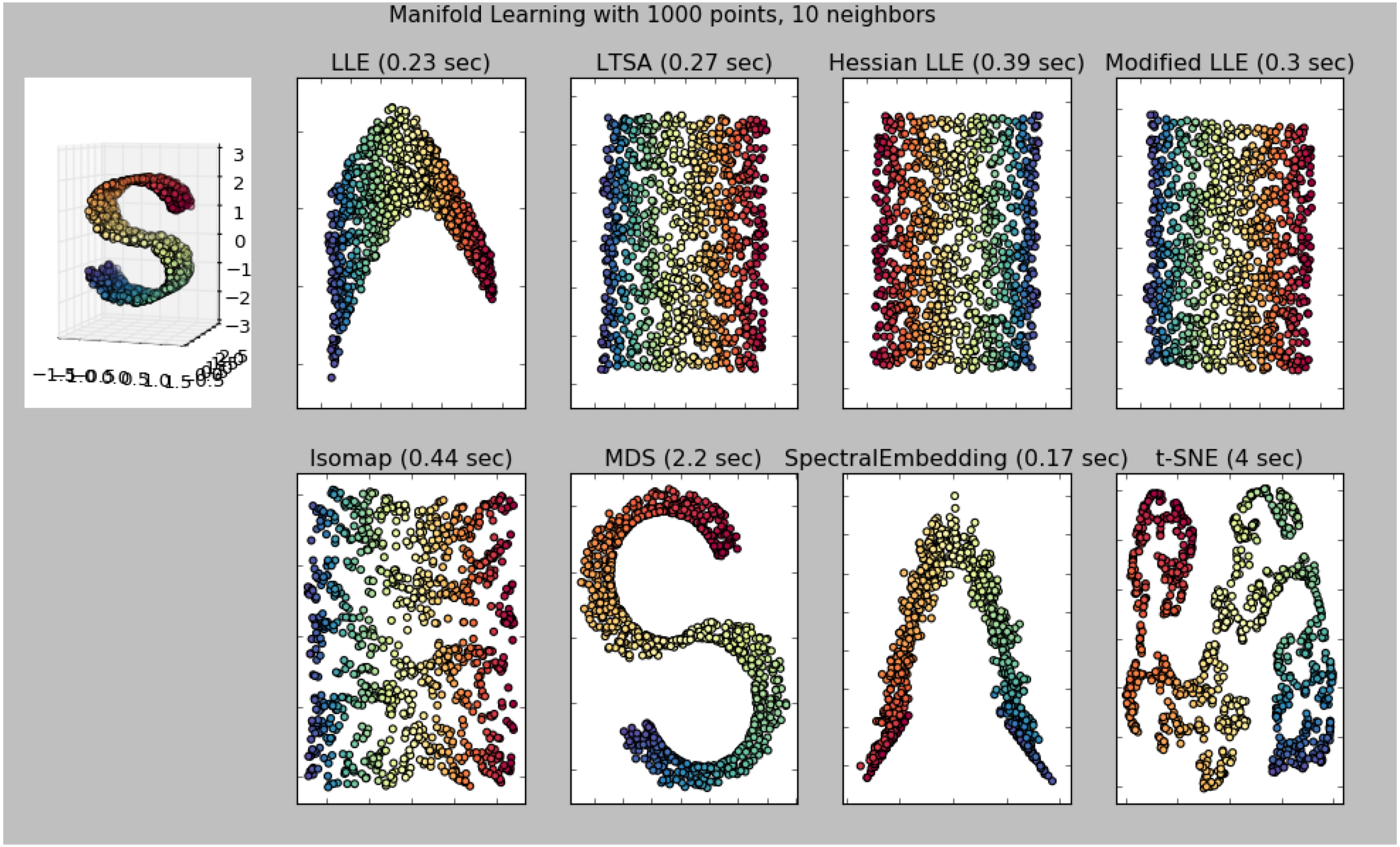
Approximation Methods

- Pair Plots
  - Axis-aligned projections
  - Interpretable in terms of original variables
- Singular Value Decomposition
  - Optimal with respect to 2-norm (Euclidean norm) and supremum norm
  - Comes with an error estimate
- Other methods
  - Stochastic Neighbor Embedding ((t-)SNE)
  - „Manifold Learning“

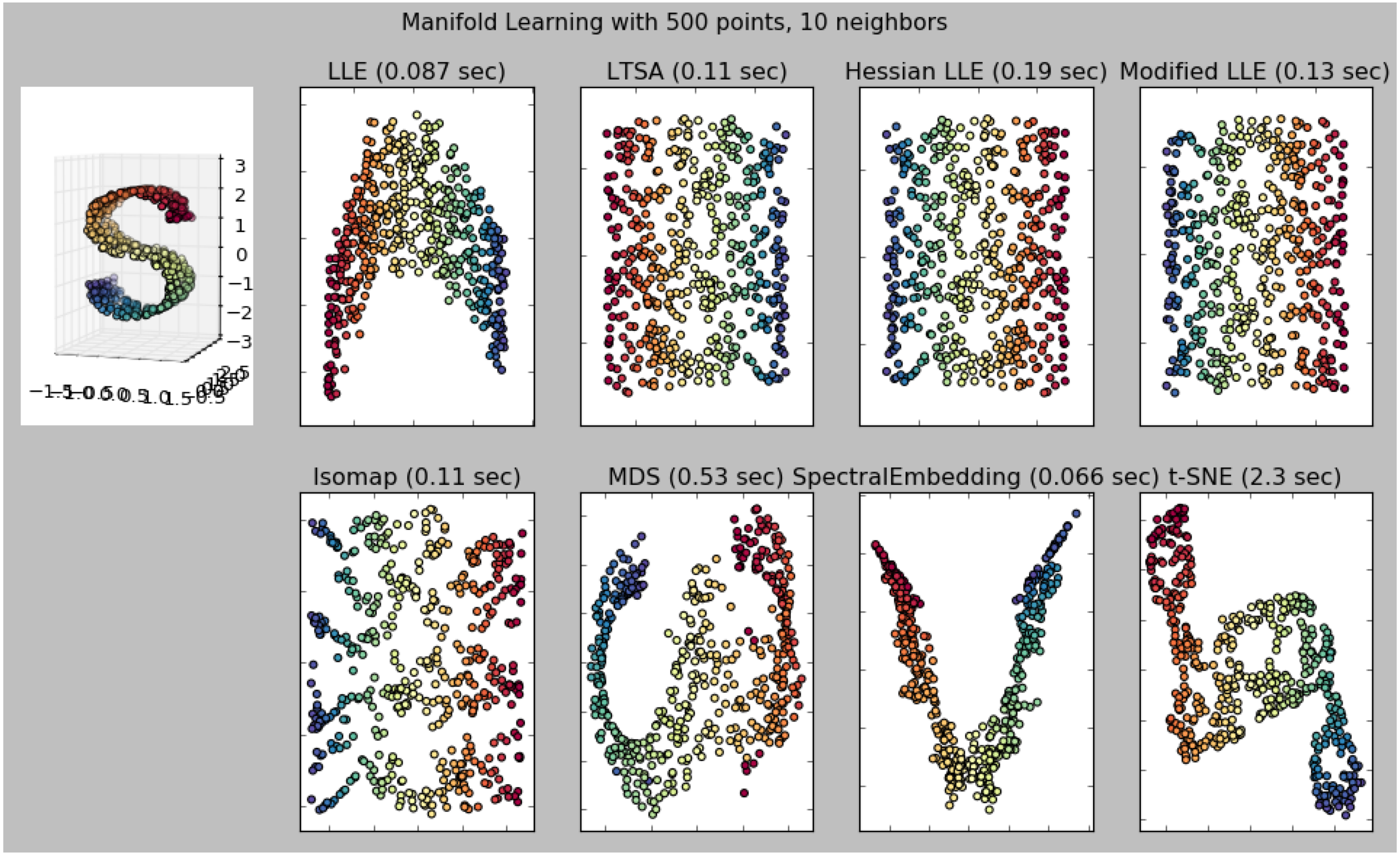
Manifold Learning Methods

- Locally Linear Embedding
  - Neighborhood-preserving embedding
- Isomap
  - quasi-isometric
- Multi-Dimensional Scaling
  - quasi-isometric
- Spectral Embedding
  - Spectral clustering based on similarity
- Stochastic Neighbor Embedding (SNE, t-SNE)
  - preserves conditional probabilities for similarity
- Local Tangent Space Alignment (LTSA)

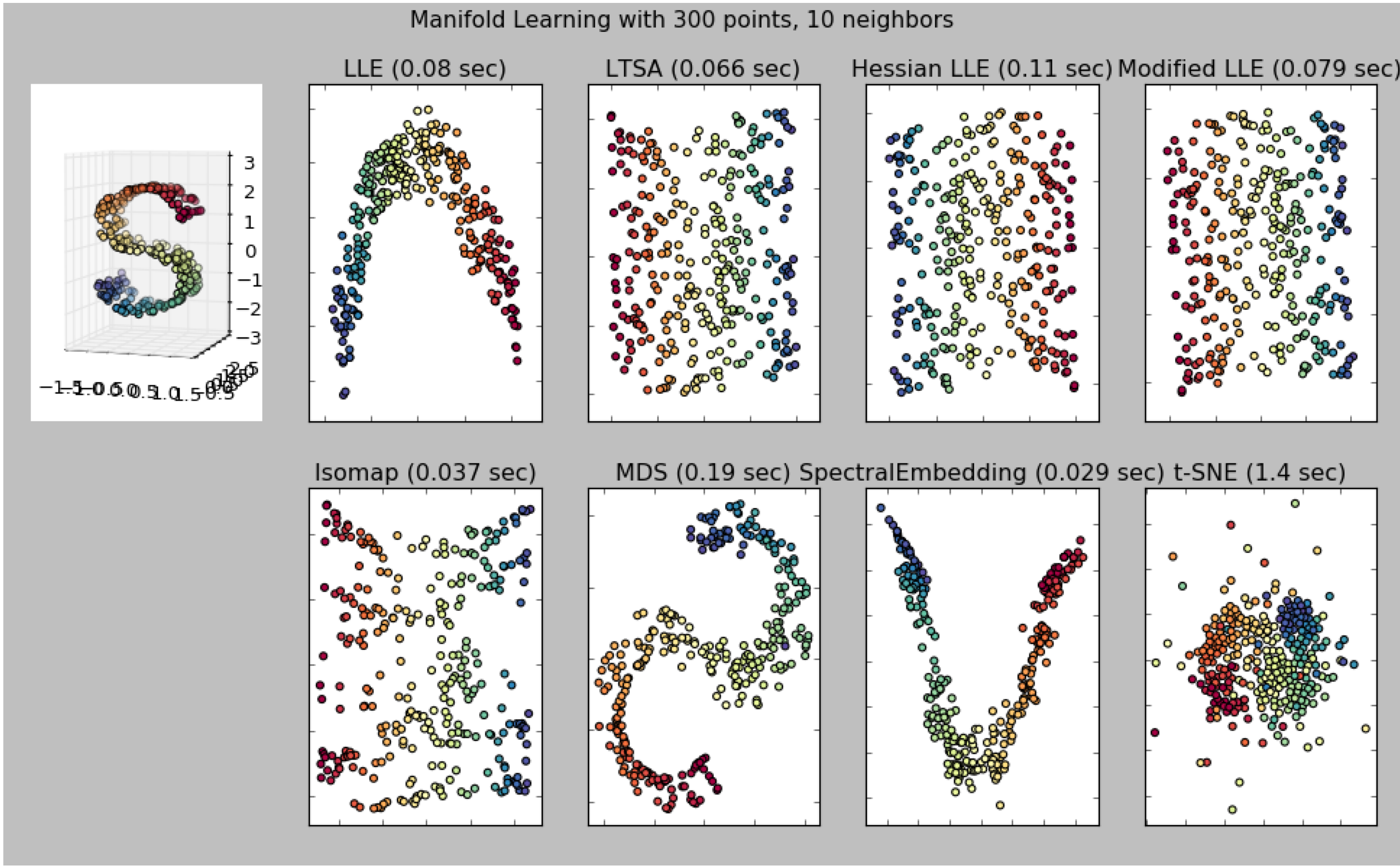
Manifold Learning



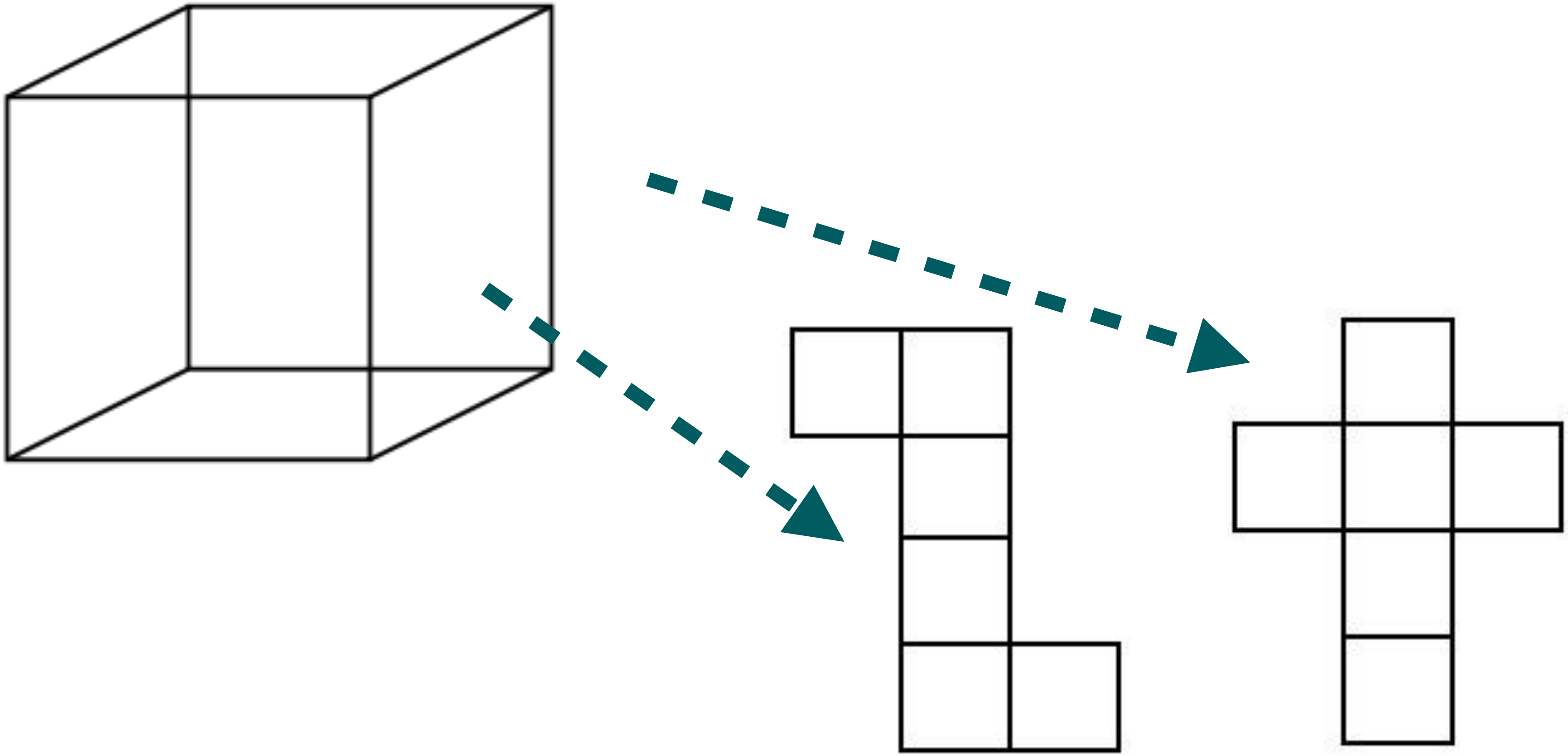
Manifold Learning



Manifold Learning



Local Tangent Space Alignment



Principal Components and Curves

- Principal Component Analysis
  - orthogonal decomposition based on SVD
  - linear in all variables
  - tries to preserve variance
- Principal Curves
  - minimize the Sum of Squared Errors with respect to all variables (as PCA, preserve variance)
  - nonlinear
  - smooth

Principal Components and Curves

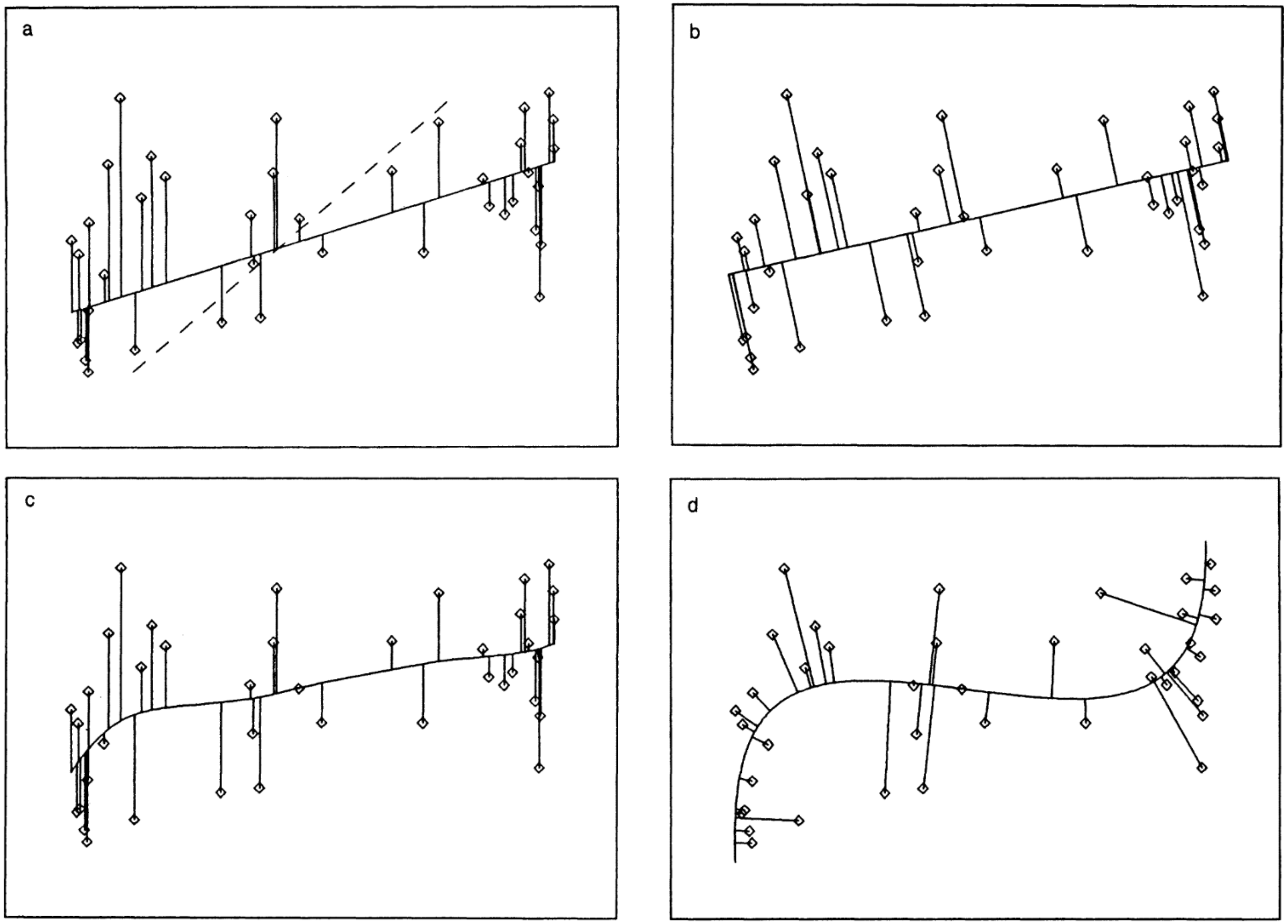
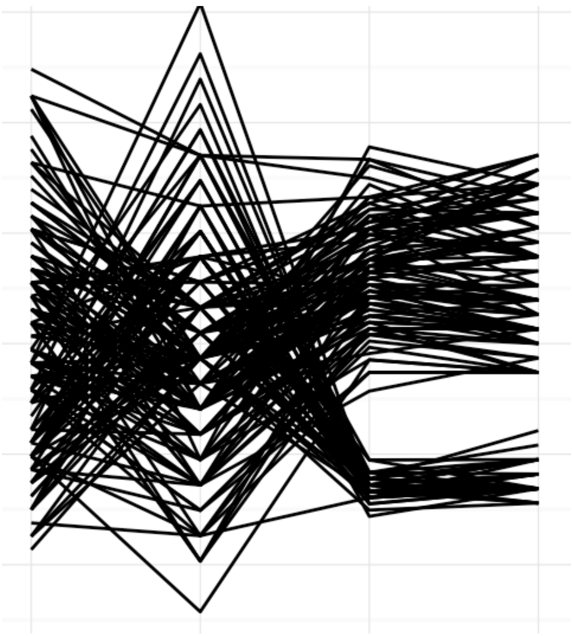
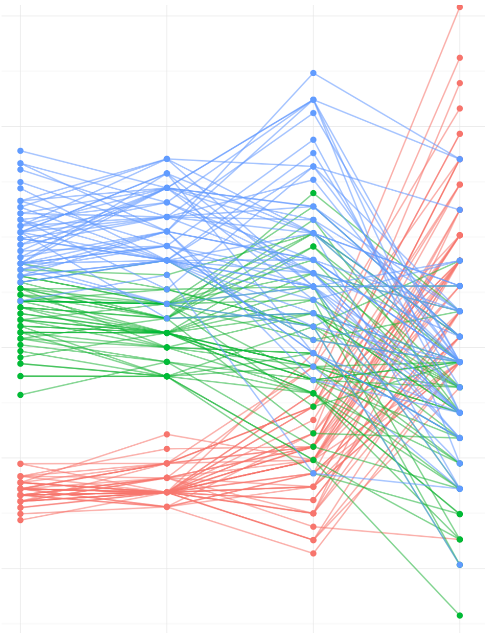


Figure 1. (a) The linear regression line minimizes the sum of squared deviations in the response variable. (b) The principal-component line minimizes the sum of squared deviations in all of the variables. (c) The smooth regression curve minimizes the sum of squared deviations in the response variable, subject to smoothness constraints. (d) The principal curve minimizes the sum of squared deviations in all of the variables, subject to smoothness constraints.

Parallel Coordinates

- Parallel Coordinates
  - especially useful for high-dimensional data
  - depends on ordering and scaling

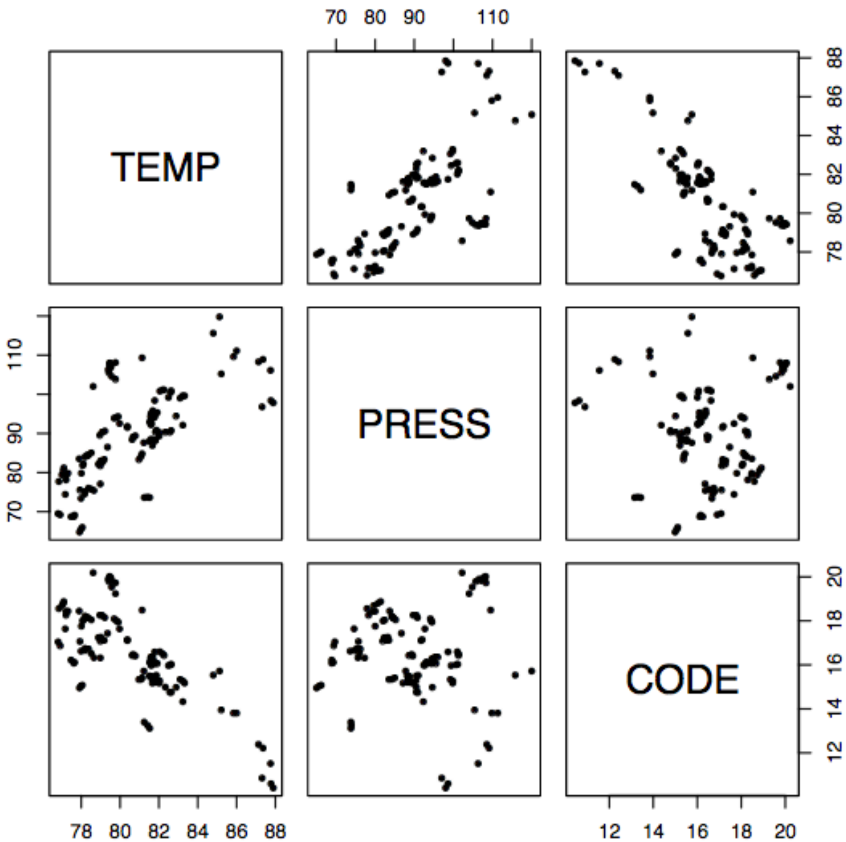


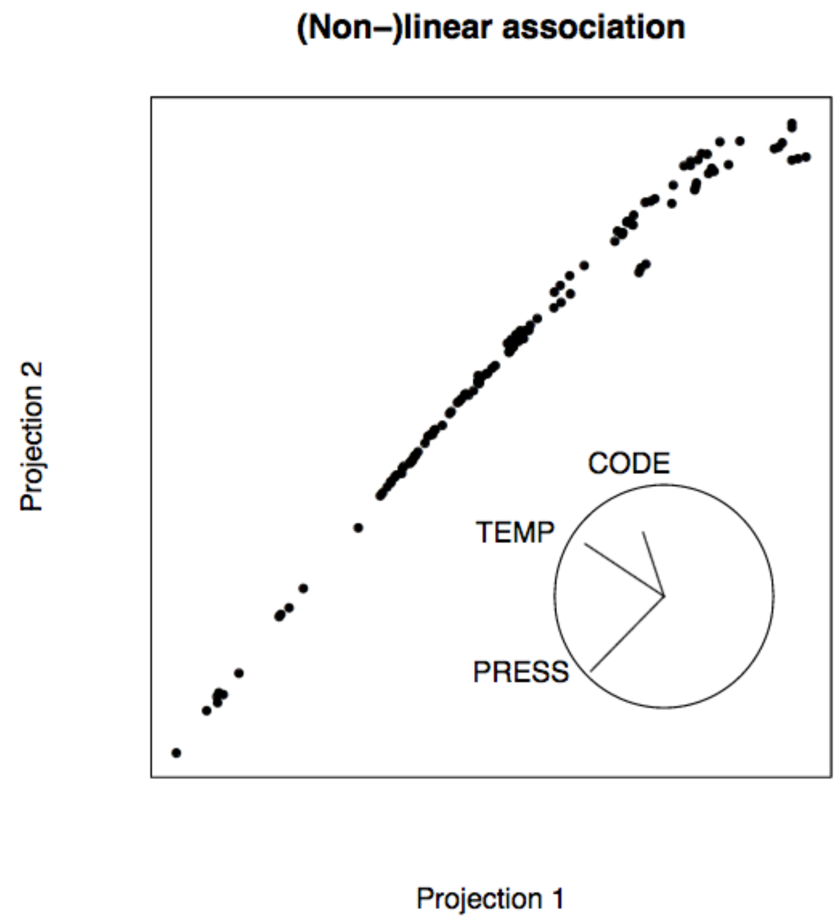


The Grand Tour

- Animated sequence of 2-D projections
  - https://en.wikipedia.org/wiki/Grand_Tour_(data_visualisation)
  - Asimov (1985): The grand tour: a tool for viewing multidimensional data.
- Underlying idea
  - Randomly generate 2-D projections (random walk)
  - Over time generate a dense subset of all possible 2-D projections
  - Optional: Follow a given path / guided tour

The Grand Tour





The Grand Tour