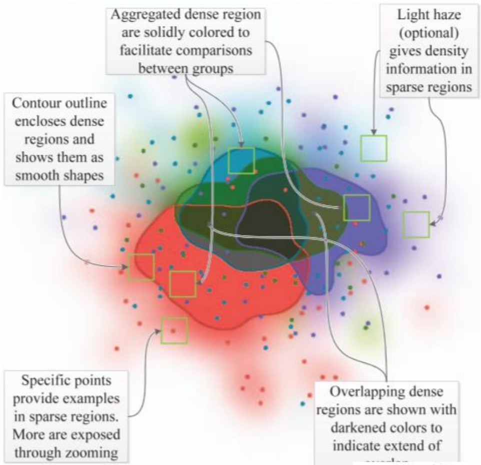
Scatterplot-Based Visual Representations



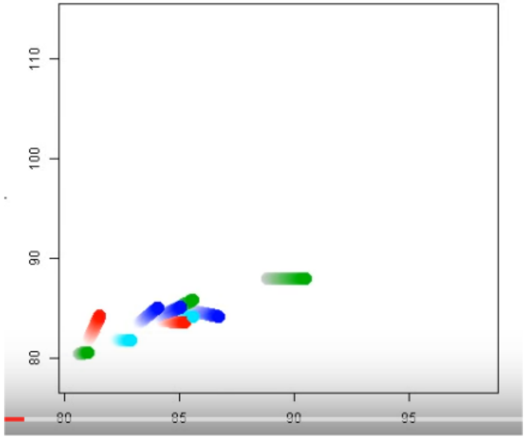
Aggregated dense region are solidly colored to facilitate comparisons between groups

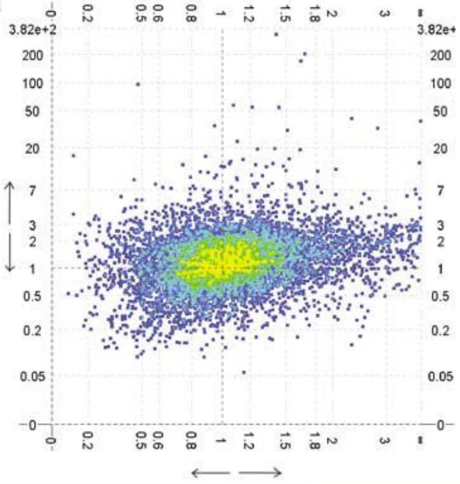
Light haze (optional) gives density information in sparse regions

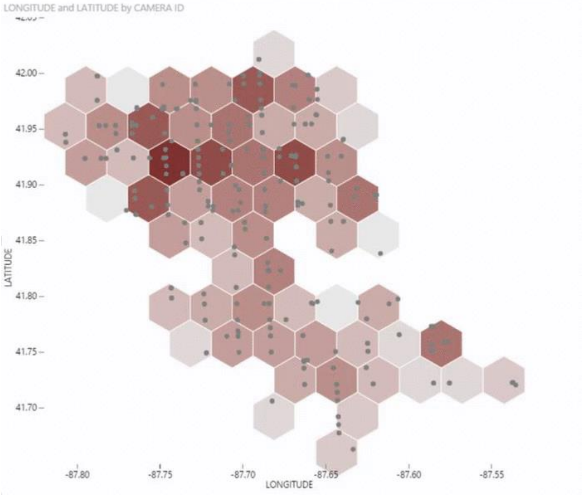
Contour outline encloses dense regions and shows them as smooth shapes

Specific points provide examples in sparse regions. More are exposed through zooming

Overlapping dense regions are shown with darkened colors to indicate extend of







Outline

- Introduction
- 2D and 3D Scatterplots
- Transformation and Distortion
- Multiclass Scatterplots
- Scatterplot Matrices and GPLOMs
- Advanced Scatterplot-Based Representations
  - Binning and Density Plots
  - A scaleable technique: Splatter Plots
  - Scagnostics-Based Exploration
  - Animated Scatterplots
  - Scatterplots on Small Screens

Introduction

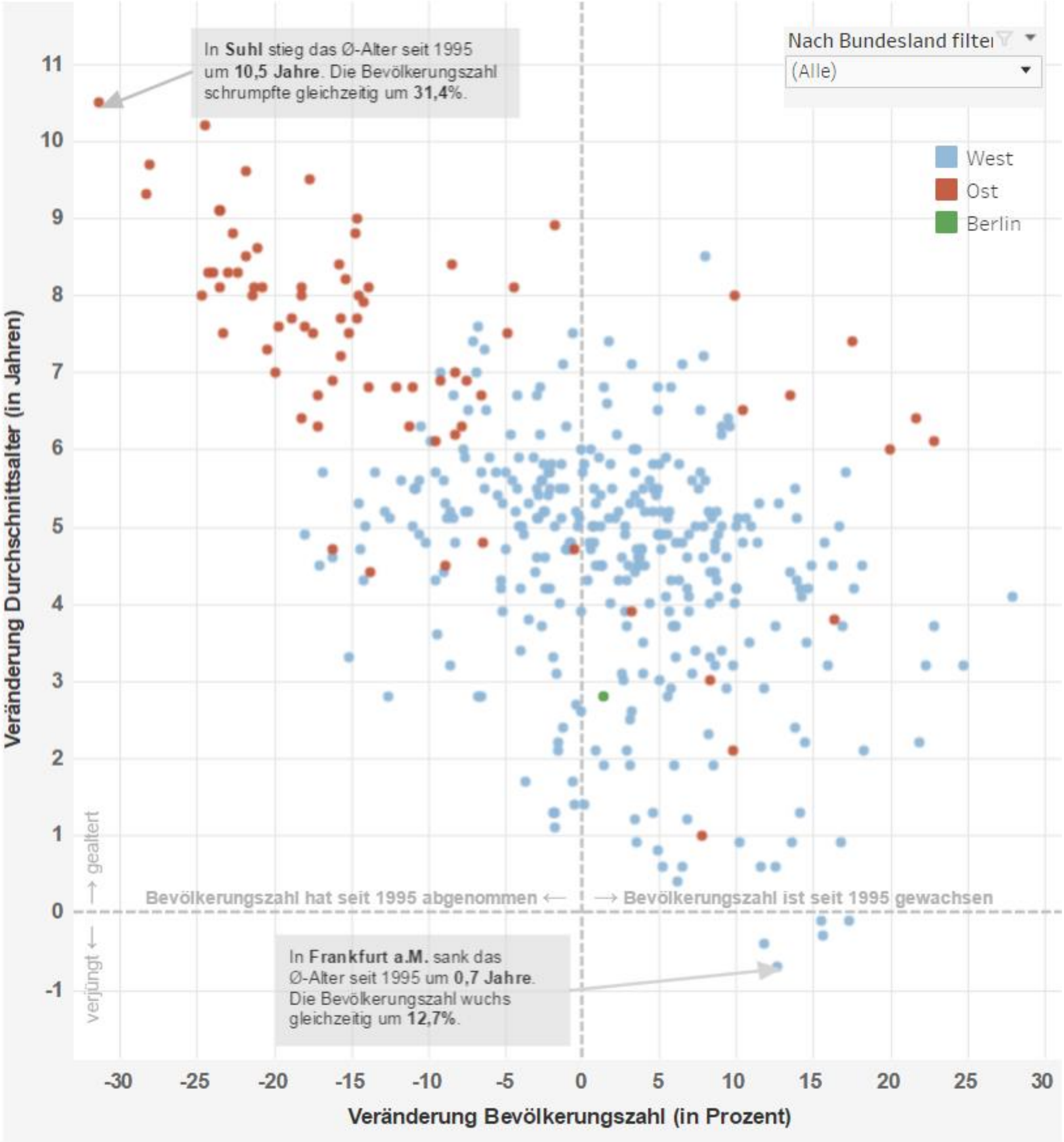
Germany gets older. Only in 5 of 432 districts the average age decreased.

Often, the population decreased and got older.

Interaction:

- Tooltips with precise numbers
- Filtering (Bundesland)

(Screenshot from zeit.de, created with Tableau)



In **Suhl** stieg das Ø-Alter seit 1995 um **10,5 Jahre**. Die Bevölkerungszahl schrumpfte gleichzeitig um **31,4%**.

In **Frankfurt a.M.** sank das Ø-Alter seit 1995 um **0,7 Jahre**. Die Bevölkerungszahl wuchs gleichzeitig um **12,7%**.

Nach Bundesland filtei

(Alle)

Introduction

Scatterplots

- represent distributions of 2 continuous variables *x* and *y* based on a sample (x_i, y_i)
- Serve to study how *y* is influenced by *x*; the type and amount of association
- are based on small glyphs (circles, rectangles) representing the value of objects in 2D
  - Design decisions: glyph shape, size, fill color, border color, transformation, labeling of axes
- represent sparse regions and high-density areas
- are intuitively assessed based on Gestalt laws (perceptual grouping)
- are familiar since the 19th century as a clear overview of data, based on „Small multiples“ (Tufte, 1983)

Introduction

Scatterplots serve a number of purposes (Tufte, 1983):

- Analysis of clusters
- Analysis of correlations
- Identification of outliers
- Comparison of different datasets plotted on the same axes (changes over time, comparison of numbers in different countries, differences man/woman, …)

Introduction

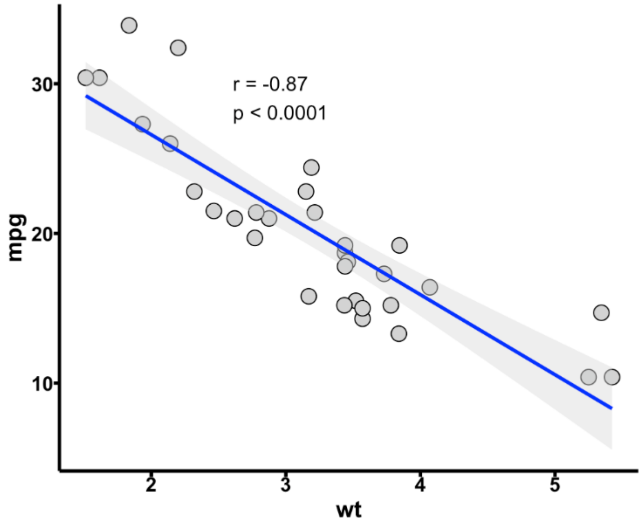
Scaleability:

- Traditional scatterplots: limited to two dimensions and to a few thousand elements
- Beyond the limit: visual clutter and overplotting hamper the perception of groups, correlations, …
- Modern techniques aim at the overplotting problem including
  - adaptations of the basic technique (color mapping, transparency),
  - abstraction from individual points,
  - transformation of axis,
  - distortion

Introduction

Human perception of correlation is not perfect (Rensink, 2010).

- Correlations (r) between 0.2 and 0.6 are underestimated.
- With *r* < 0.2 no correlation is perceived.
- The display of a regression line and a correlation coefficient support a quantitative analysis.
- To consider uncertainty, a confidence interval around the regression line is helpful.

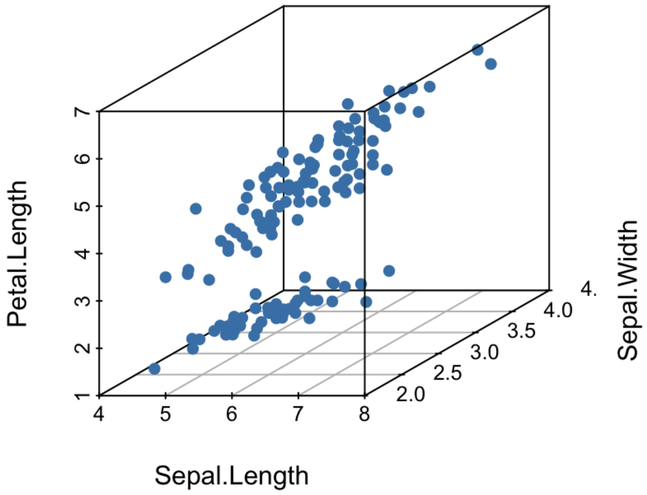


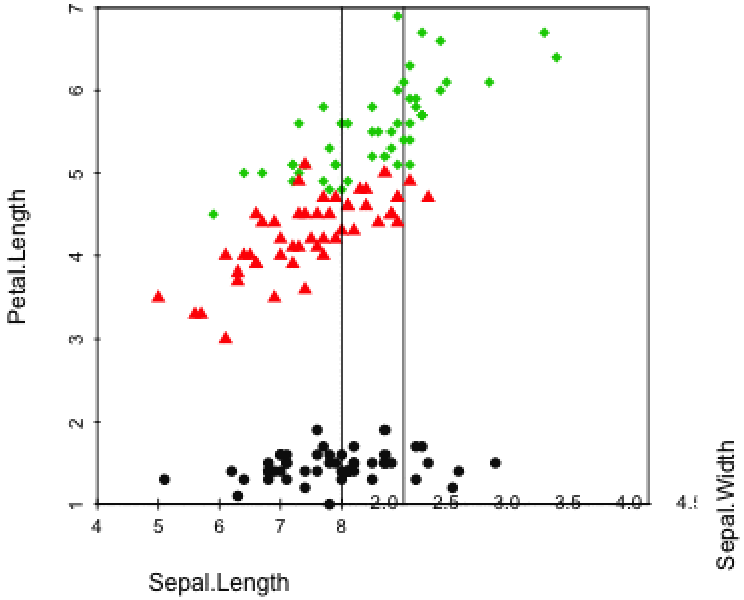
Variables (miles per galon and weight of cars) correlate negatively with a strong coefficient (From: Link)

2D and 3D Scatterplots

Distributions of 3 variables may be depicted with 3D scatterplots

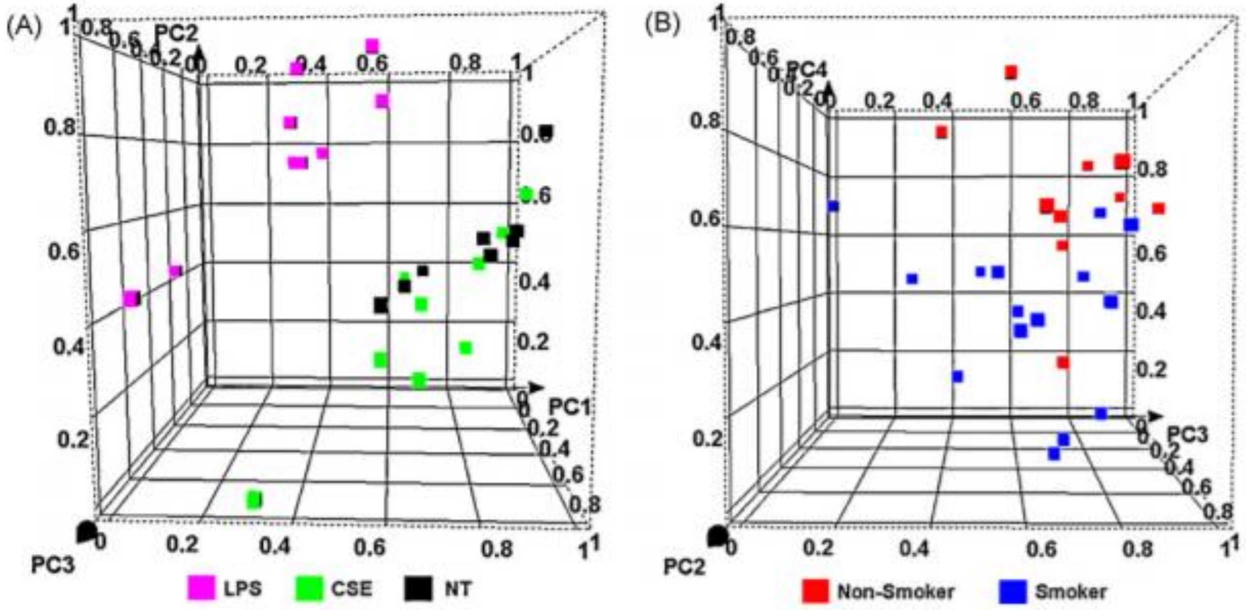
A 3D scatterplot requires a view angle and facilities to change the viewpoint





Screenshot of the R System (From: Link)

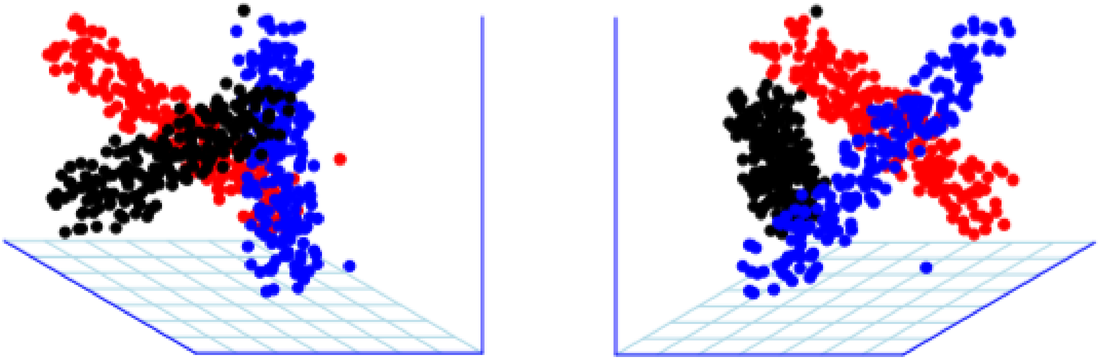
2D and 3D Scatterplots



3D Scatterplots from gene analysis paper
(From: Dolye, 2010)

- „Switching to 3D rarely helps and often hurts, it has higher time costs and prvides less class separability“
- „the time of interaction is high because the user must spend significant time rotating the view to see the structure from different angles“ (Sedlmaier, 2013)

2D and 3D Scatterplots



Three classes shown in 3D scatterplots from different viewing angles. Only by considering a number of view-points the full picture becomes obvious (From: Sedlmaier, 2013).

Most real-world data are much more complex and the classes are not easy to separate.

2D and 3D Scatterplots

3D Scatterplots

- aim to exploit the human capabilities of understanding spatial relations
- suffer from occlusion problems and require mentally demanding interaction

Studies indicate: 3D scatterplots are not favorable. The natural human experience relates to 3D shapes, not to 3D point clouds representing abstract data.

Depth cues, such as shadow and shading, do not make sense for point clouds.

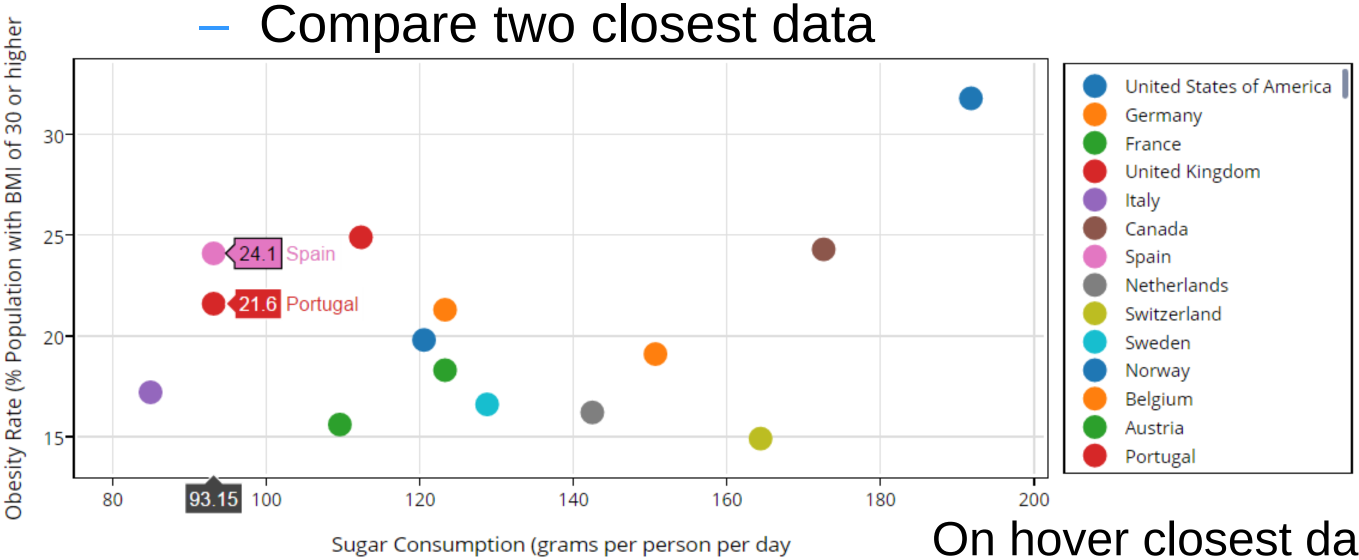
2D and 3D Scatterplots

- In case of three dimensions (*x*,*y*,*z*), a layout with a scatterplot matrix (*x*,*y*), (*y*,*z*), (*x*,*z*) and a 3D scatterplot may be favorable.
- The only advantage of a 3D scatterplot is a more integrated view of three dimensions (Sedlmaier, 2013).
- This does not hold when the data is a result of dimension reduction.

2D Scatterplots

Interaction:

- Zoom In/Zoom Out
- Autoscale
- Select (rectangle/lasso) for a detail view
- Hover
  - Show closest data
  - Compare two closest data



On hover closest data are labeled for comparison (Screenshot from Plotly, Link)

2D Scatterplots: Fight Overplotting

Major goal of most advanced techniques is **scaleabilty** w.r.t. element number and/or dimensions and classes.

- Metrics of visual **clutter** serve as benchmarks:
  - Metrics consider screen-space statistics (number of used pixels, number of free pixels, „collisions“ where 2 elements are mapped to the same pixel, collision ratio) (Bertini, 2006)
  - Ideally, SP-based representation has the same complexity metrics over a wide range of data including very large data.

Clutter reduction in information visualization is a general issue, related to parallel coordinates, tables, … (see Ellis and Dix, 2007 for a taxonomy)

2D Scatterplots: Fight Overplotting

Clutter reduction (Ellis and Dix, 2007):

Appearance:	Change colors, transparency, add blur, …
Geometric:	Filter, adjust sampling, distortion
Re-ordering:	relevant for tables, parallel coordinates; not for scatterplots

For scatterplots, ideally a technique:

- Avoids overlap, or
- Indicates the amount of overlap
- Is scaleable, and
- Is adjustable (control clutter interactively)

2D Scatterplots: Fight Overplotting

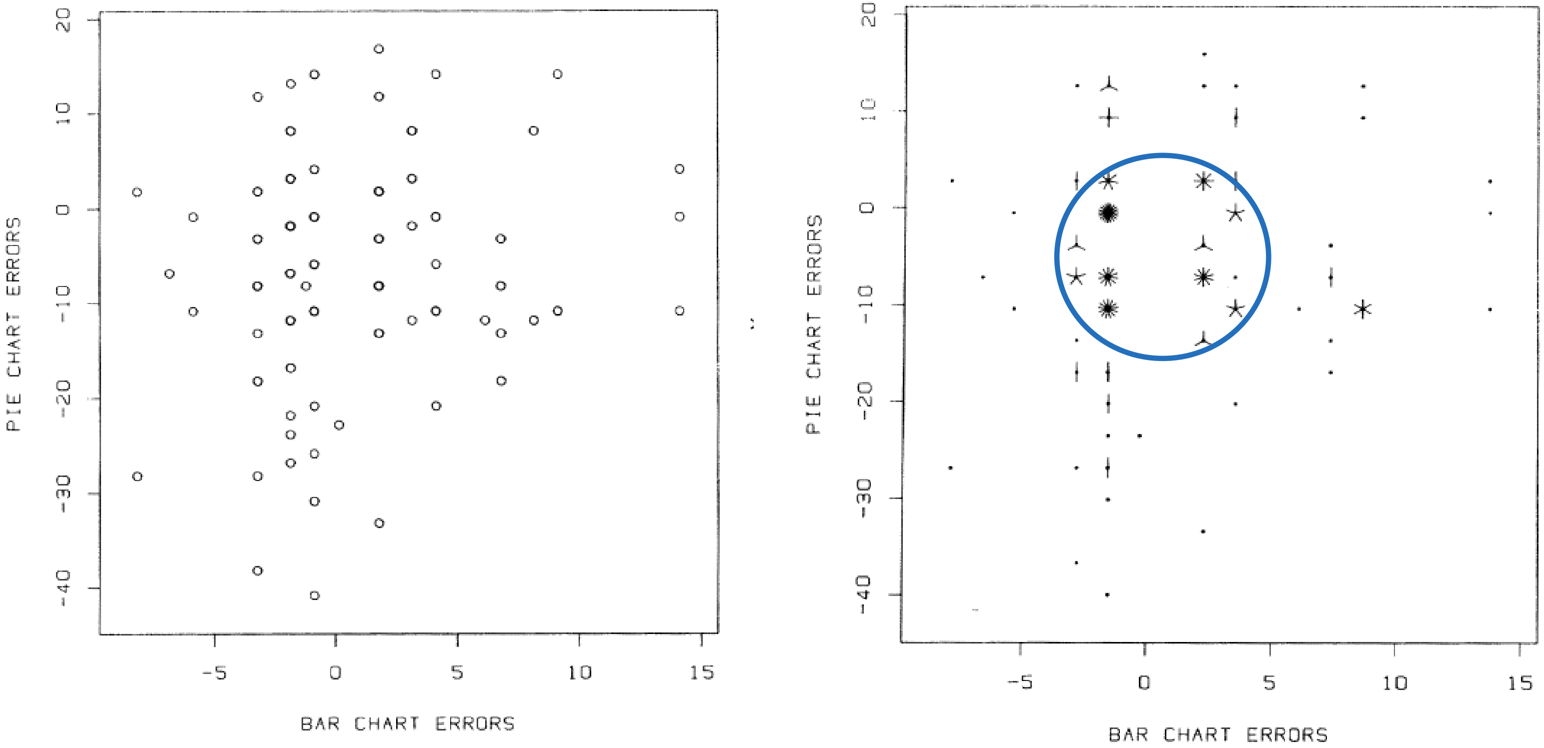
Overplotting distorts the perception of correlations and groupings.

Solutions for the Overplotting Problem (Few, 2008):

- Appearance: Adjust Transparency to the number of elements
- Appearance: Adjust Darkness to the number of elements
- Geometric: Reduce the glyph size
- Geometric: Change glyph shape, e.g. from circle to + shape
- Geometric: Jittering (slightly displace points by adding noise, often considered the best method)
  - Requires an automatic solution
  - Often justified by rounded data (70 kg; actually [69.5, 70.5]

All methods are widely available in vis. analytics products

2D Scatterplots: Fight Overplotting



A classic example: A conventional SP and a SP with „sunflower glyphs“ to indicate overplotting.

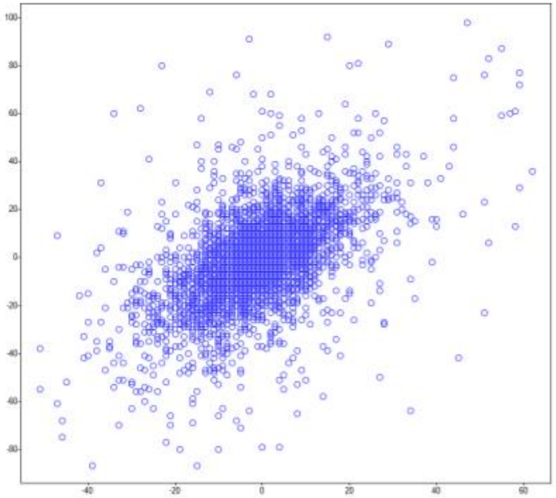
Each bar in the glyph represents one instance indicating up to 15 overlaps (From: Cleveland, 1984).

+ The cluster around (0,0) becomes obvious.

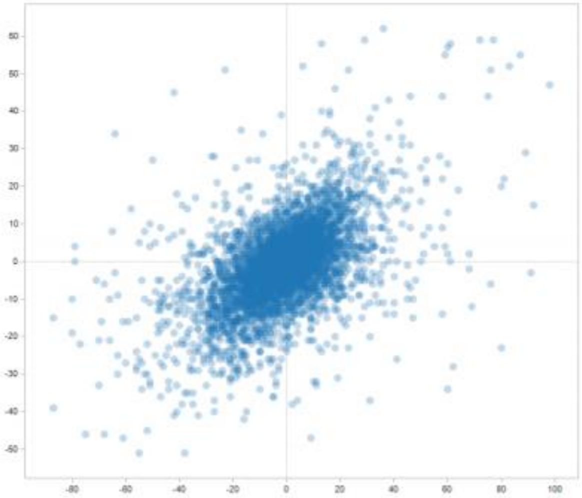
- Each glyph consumes at least 49 pixels.

2D Scatterplots: Fight Overplotting

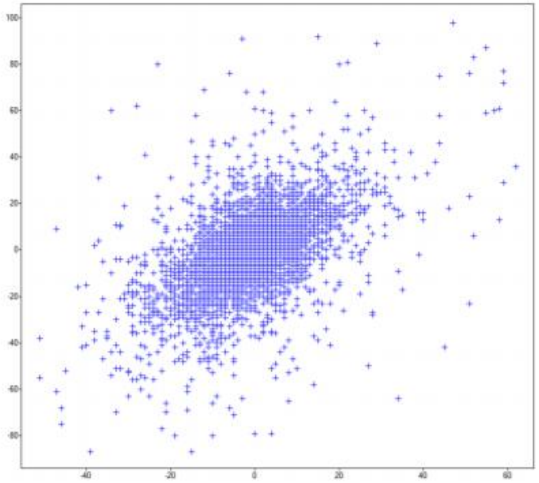
Typical solutions to the overplotting problem



[Created using Spotfire.]



[Created using Tableau.]

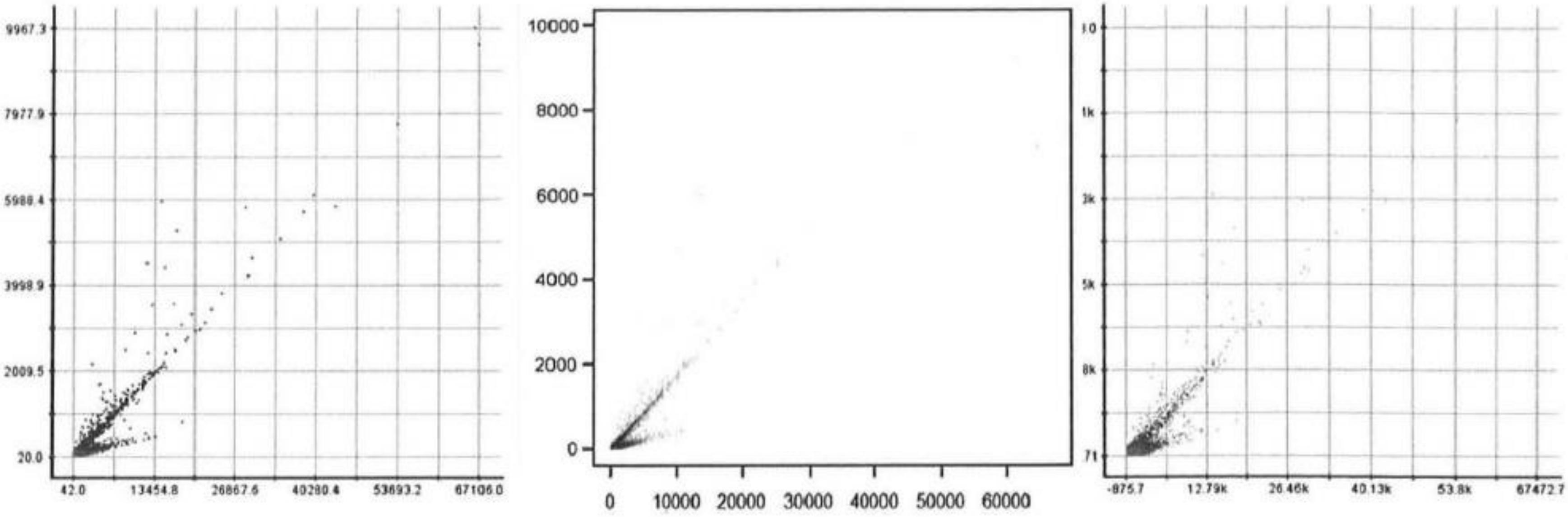


[Created using Spotfire.]

Left: no adjustment, **middle**:Transparency adapted to point number; **right**: glyph shape adjusted (From: Few, 2008)

Could be further enhanced by supporting lens-based interaction: Show more details in a lens region controlled by the user.

2D Scatterplots: Fight Overplotting

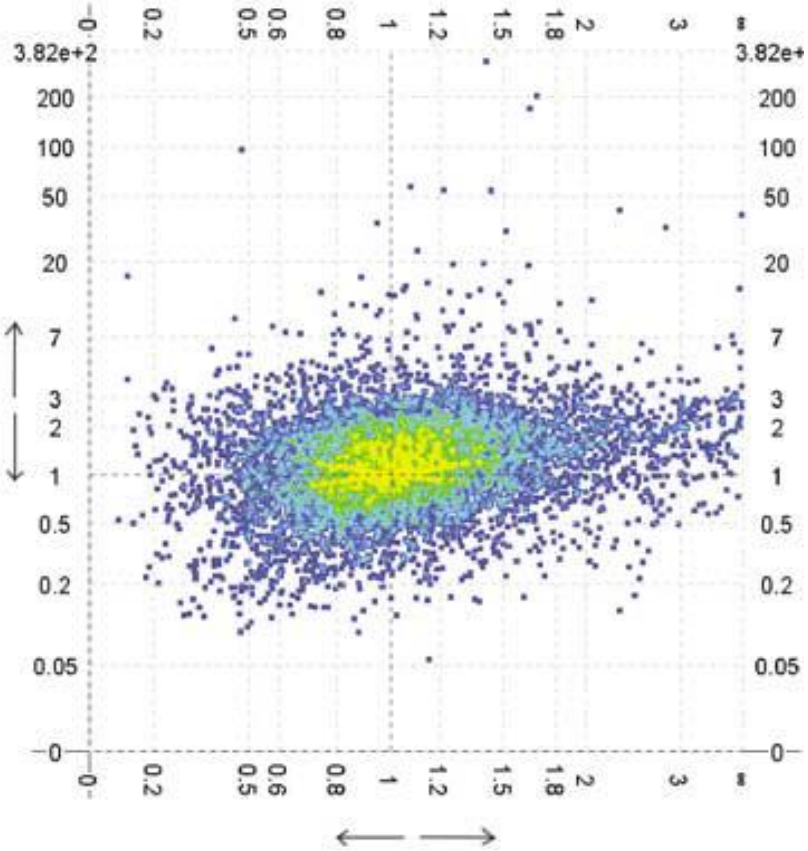


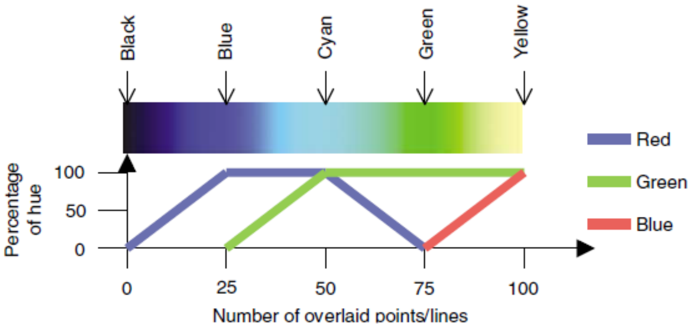
Left: Classic scatterplot.

Middle: Opacity modulation indicates how few points are outside the two diagonal lines.

Right: Jittering cannot fully solve the overlap problem (From: Keim, 2010).

2D Scatterplots: Fight Overplotting

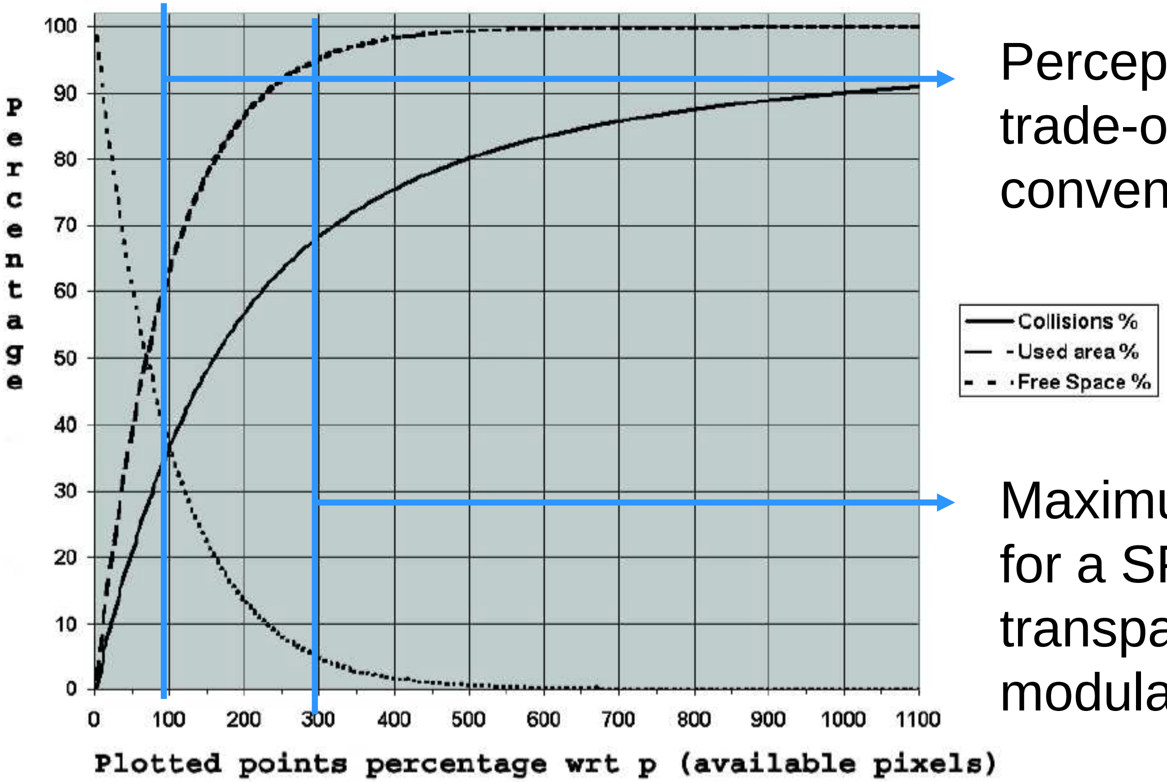




Color indicates the amount of overplotting and thus the central cluster (From: Craig, 2005).

The specific color scale matters: when mapped to a blue-yellow scale, higher numbers are perceived compared to a „red-green“ or gray value scale (Bertini, 2006).

2D Scatterplots: Fight Overplotting



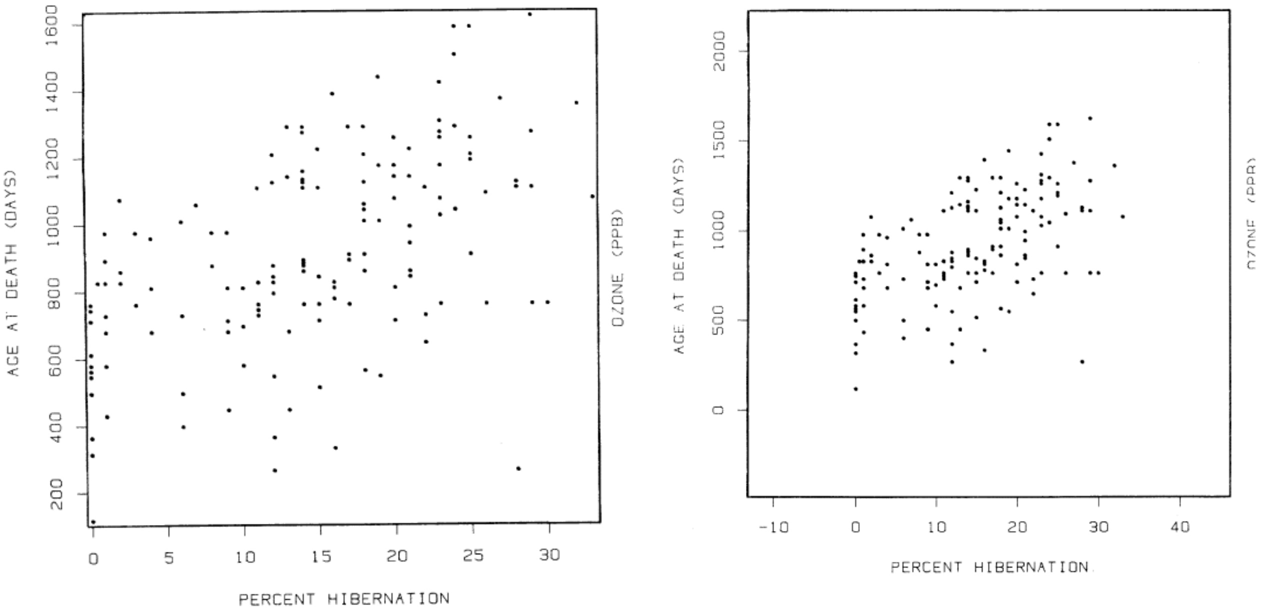
Perceptual maximum trade-off for a conventional SP

Maximum trade-off for a SP with transparency/color modulation

For data that follows a normal distribution, the % screen space and the # collisions increase and the % free space decreases when more data are shown.

With random sampling, a tradeoff between accuracy and overlapping points can be adjusted (From: Bertini, 2006).

Transformation and Distortion



Scale matters: **Left**: SP is scaled to exactly the extent of the data (x_{min}, y_{min}) (x_{max}, y_{max}). **Right**: The SP is scaled twiced that range (From Cleveland, 1984).

The perception of the amount of correlation is increased by 40% on average from left to right.

Transformation and Distortion

What is the right scale?

The frame should be

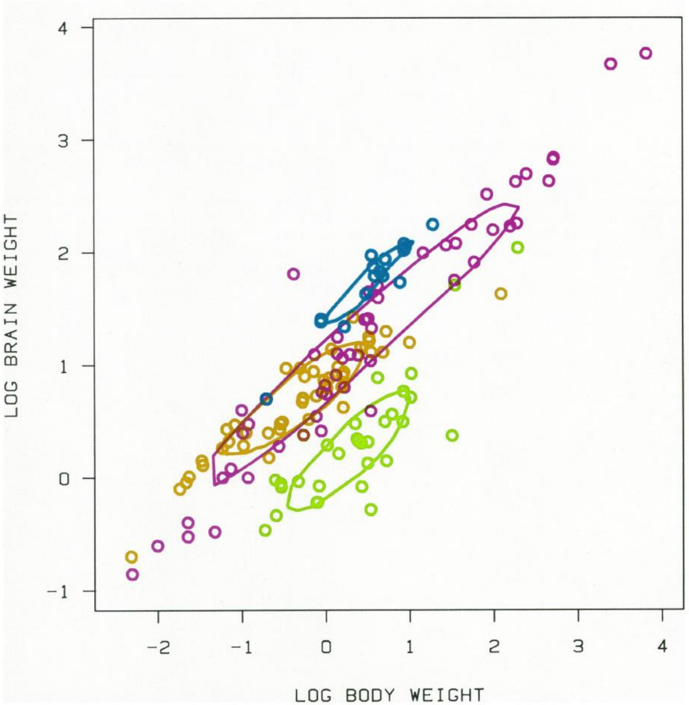
- not too close to the data,
- not too far away from the data

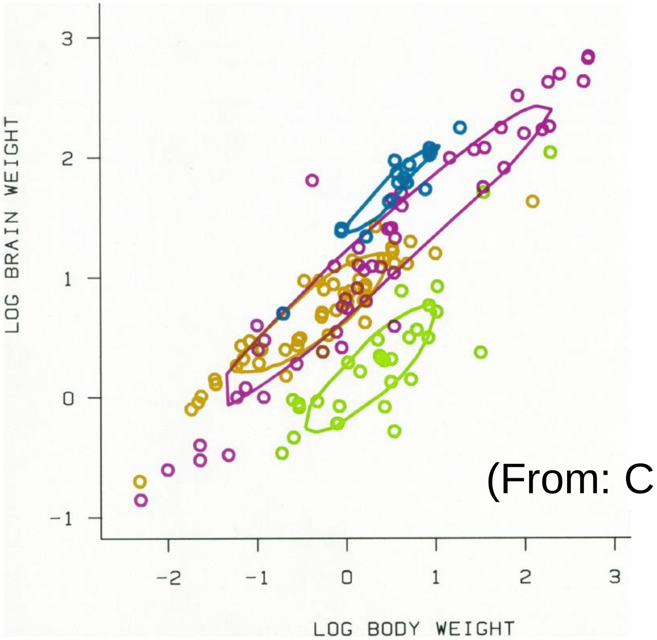
When the frame has the coordinates (0,0) (1,1); the minimum value (x_{min}, y_{min}) should be mapped to a point between 0.05 and 0.10 (in x and y) and the maximum value to 0.90-0.95 (in x and y) (Cleveland, 1984).

Transformation and Distortion

Outliers matter:

- Lead to reductions in point cloud size since the majority of the points are forced in a small region.
- Making 2 SPs with and without outliers is a sensible practice (Cleveland, 1984).

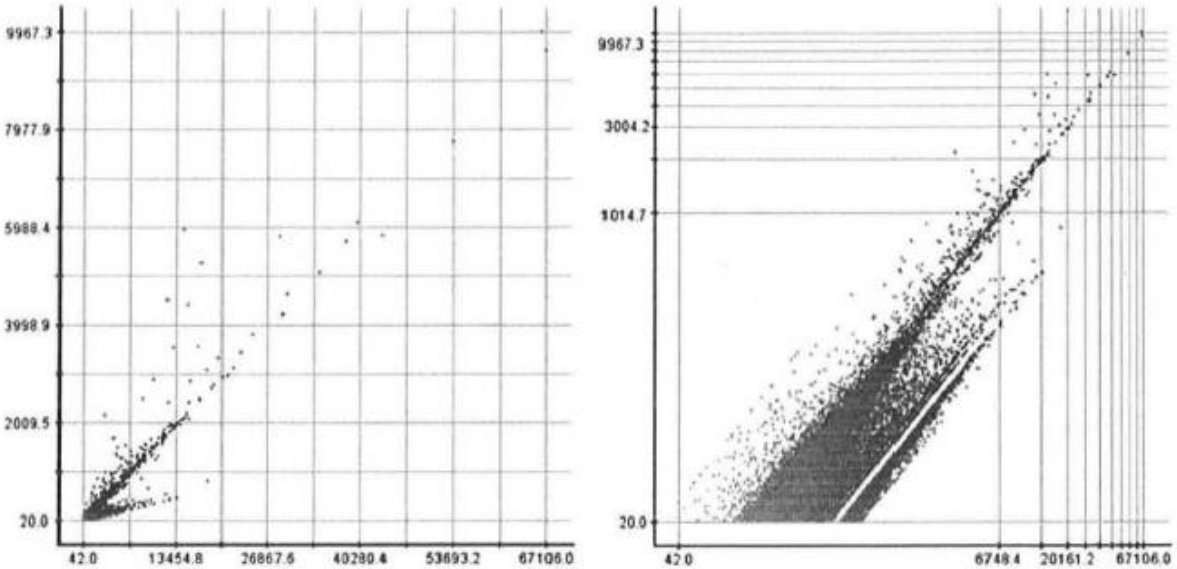




(From: Cleveland, 1984)

Transformation and Distortion

Often, a linear scale leads to large sparse regions. Logarithmic or square-root scaling for one or both axes may improve the display.



Linear and logarithmic scale of data presented in scatterplots. The grid indicates the stretched and compressed regions (From: Keim, 2010).

Transformation and Distortion

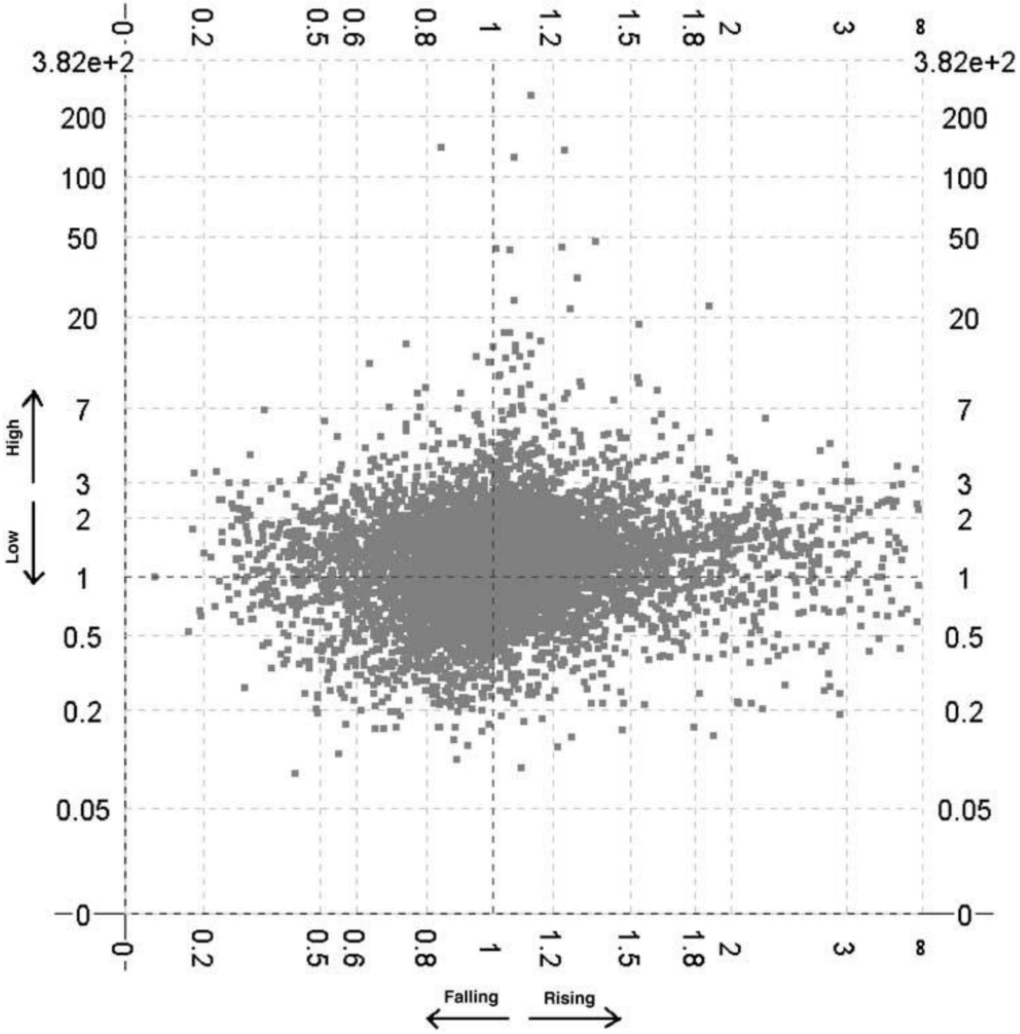
Distortion-based techniques

- increase the space for high-density areas and
- compress the space for low-density (sparse) regions

→ they reduce overlaps and overplotting at the expense of a more difficult-to-interpret visualization

- Users should be
  - aware of the distortion, e.g. by showing grid lines
  - able to control the amount of distortion, e.g. with a slider or just with mouse movement

Transformation and Distortion



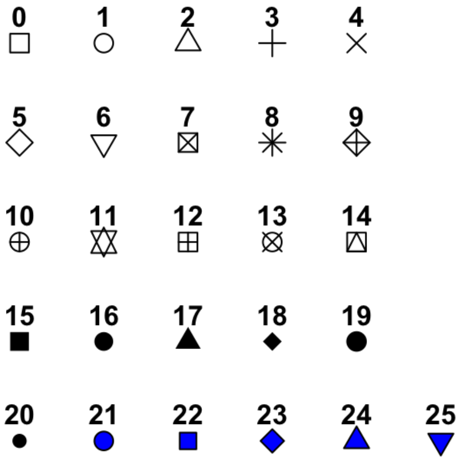
Gene expression data.

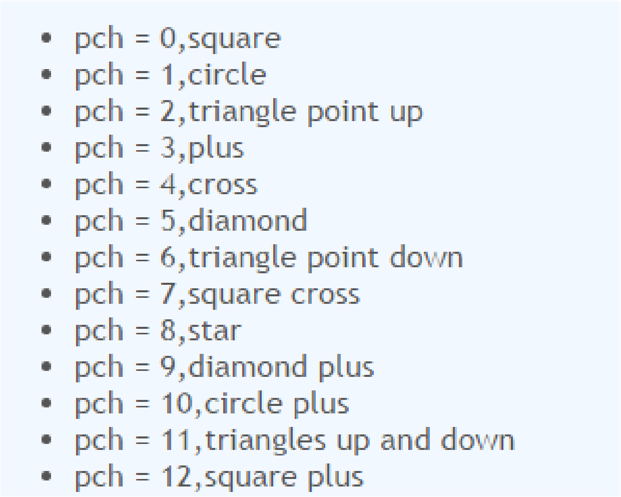
- Changes of gene expression are relevant → relative values indicate falling/rising values.
- Logarithmic scale.
- A displacement algorithm reduces overlaps

(From: Craig, 2005)

Multiclass Scatterplots

- Often, the entries in a scatterplot are discriminated by different classes, e.g. the relation between two variables is assessed to understand differences healthy and ill persons, woman and men, …
- Different classes are represented by
  - different colors, using the hue component for discrimination, or
  - different glyph shapes

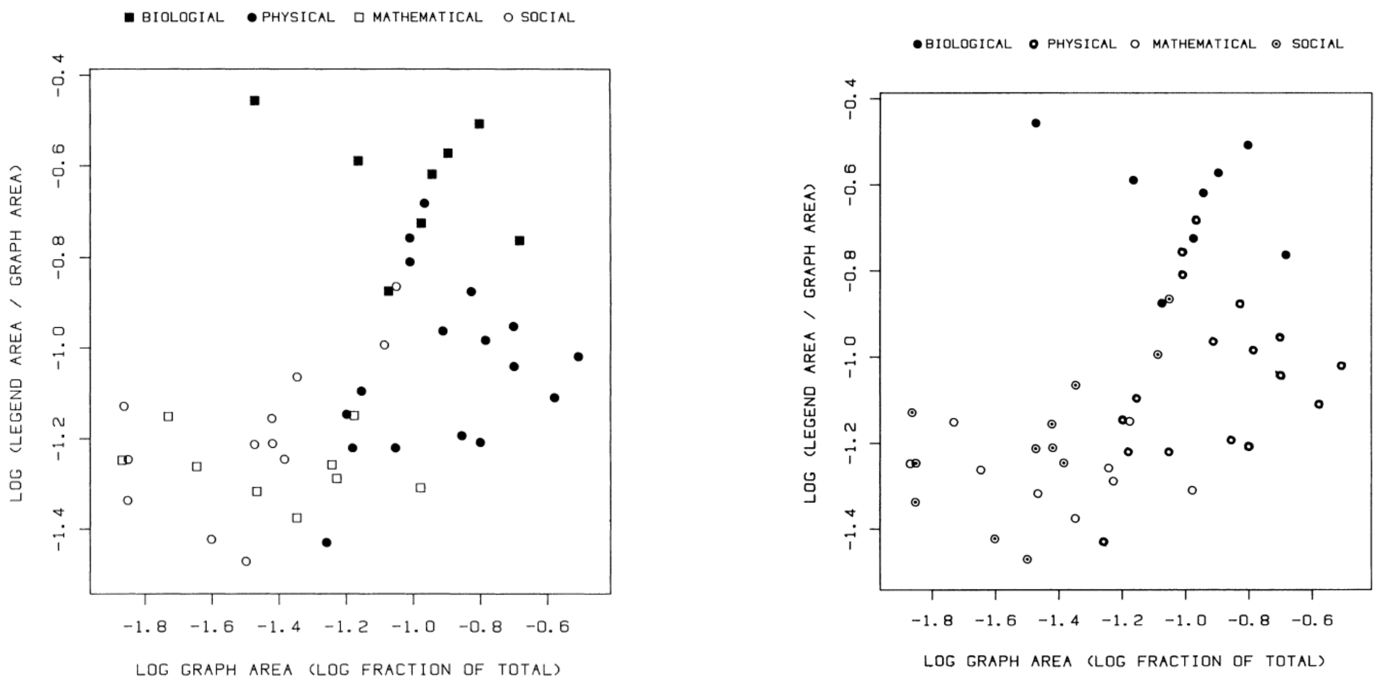




- pch = 0,square
- pch = 1,circle
- pch = 2,triangle point up
- pch = 3,plus
- pch = 4,cross
- pch = 5,diamond
- pch = 6,triangle point down
- pch = 7,square cross
- pch = 8,star
- pch = 9,diamond plus
- pch = 10,circle plus
- pch = 11,triangles up and down
- pch = 12,square plus

Glyph types of the R System (From: Link)

Multiclass Scatterplots



(From: Cleveland, 1984).

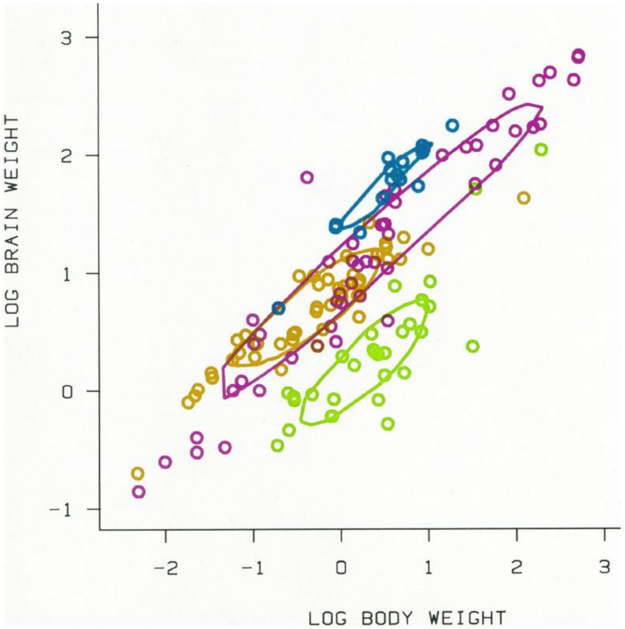
Four classes are shown: The left set of glyphs enables easy discrimination between the two filled and the two enclosed glyphs, but to distinguish between the two different filled/enclosed glyphs is difficult.

The right set of glyphs better support visual discrimination

Multiclass Scatterplots

Color supports visual discrimination (for healthy persons) better than shape.

However, saturated colors with maximum differences are perceived as unesthetic and clashing (Las Vegas method of color coding, Cleveland, 1984).

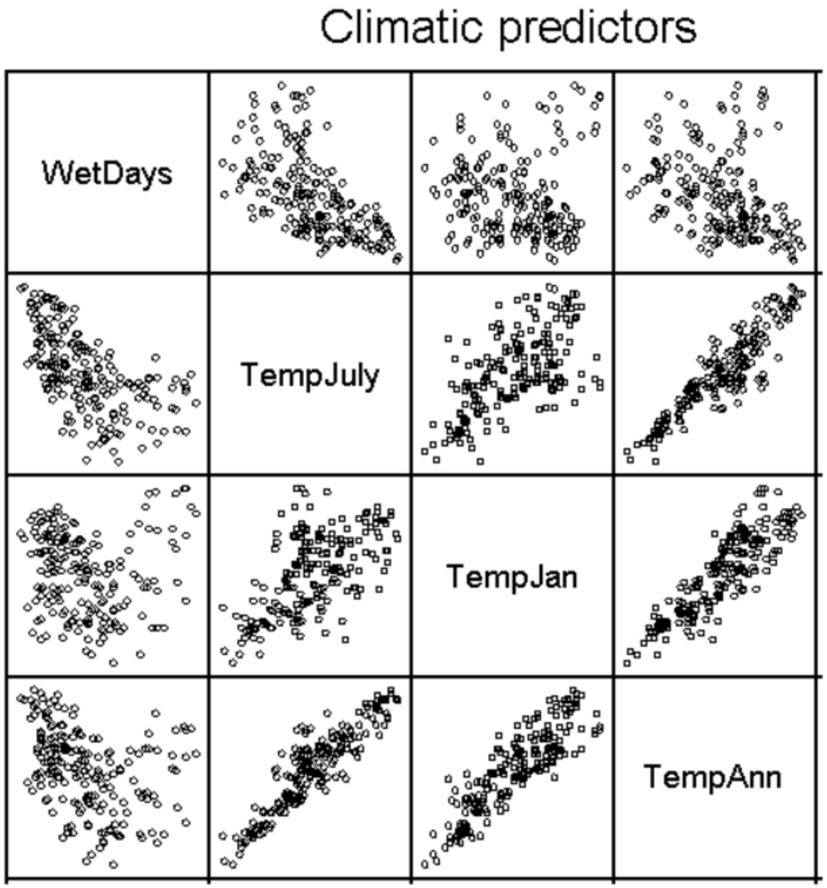


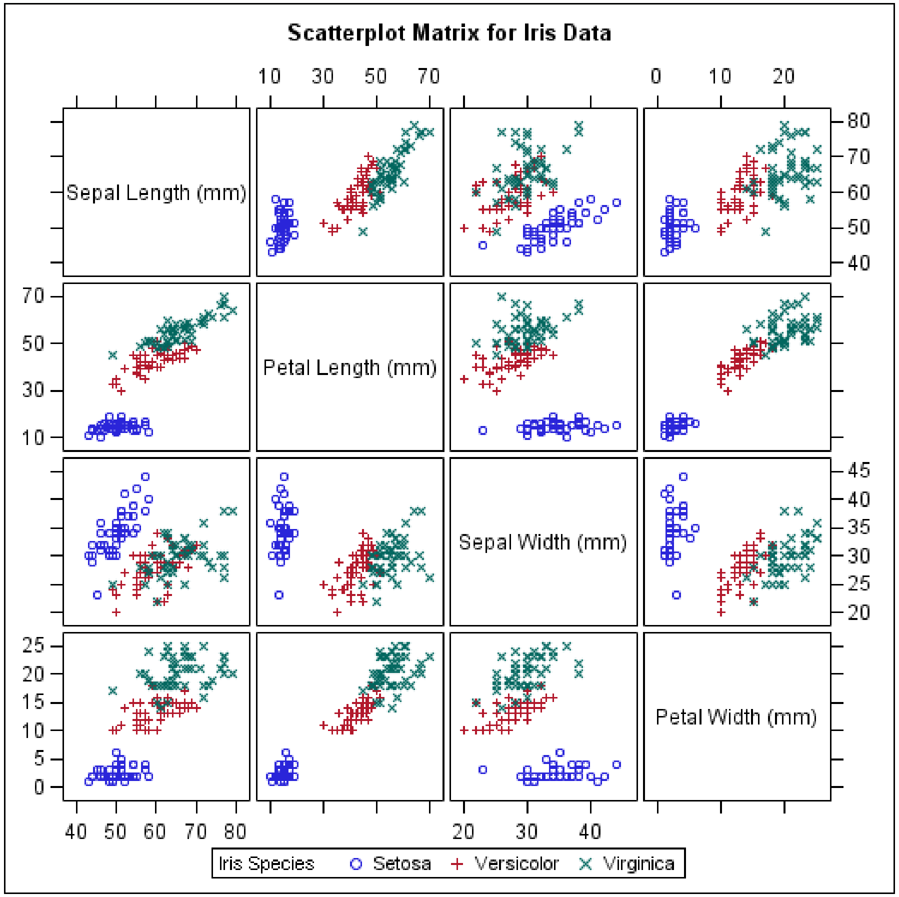
Essential goal: users perceive the distribution of each class separetely (as if other classes are not there) and in relation to the overall distribution. How does a subpopulation differs from the global mean?

Scatterplot Matrices and GPLOMs

- To visualize higher dimensional data, create a matrix where each non-diagonal element is a scatterplot showing the relation between a pair of variables.
- The number of scatterplots grows quadratically.
- Thus, SPLOMS are appropriate for rather low-dimensional data (following examples, n=4, n=5)

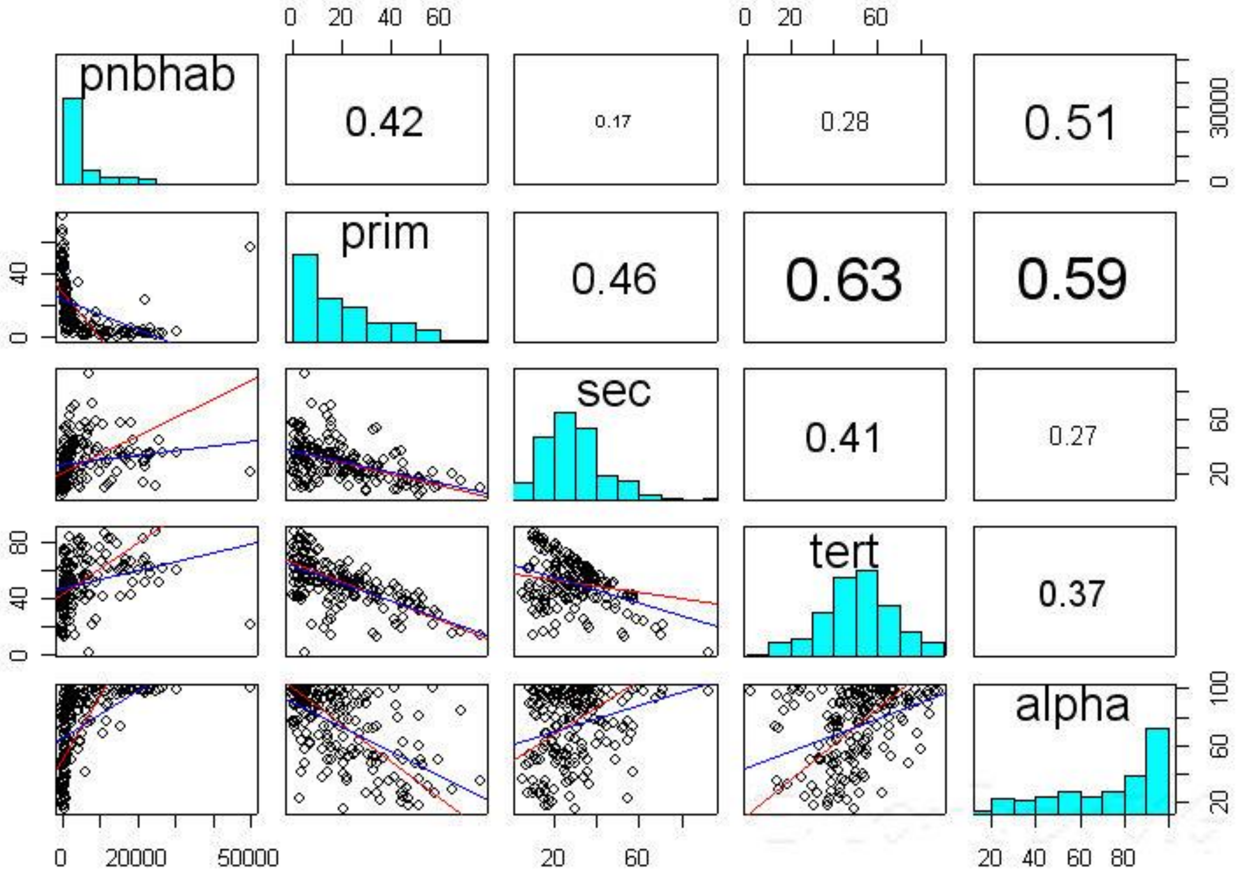
Scatterplot Matrices and GPLOMs





Single class and multiclass scatterplots matrices (also called *Trellis display*). High redundancy. Diagonal elements only used for a label (From: Link1 and Link2)

Scatterplot Matrices and GPLOMs



Better use of the space (but not optimal). Diagonal elements contains a histogramm and elements above the diagonal correlation coefficients (From: Link)

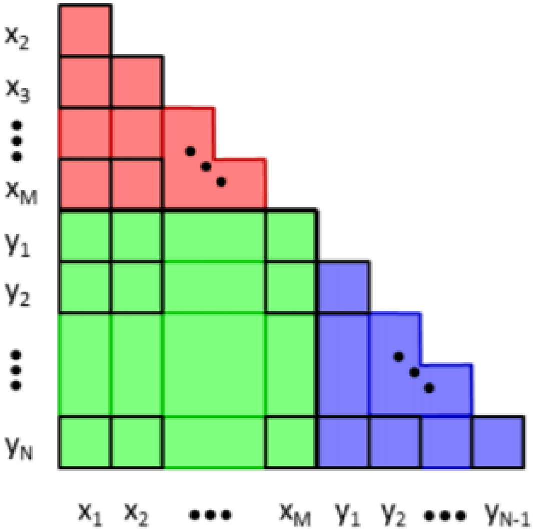
Scatterplot Matrices and GPLOMs

- Scaleing is a crucial issue. If possible, all elements of a SPLOM should have the same scale. Disadvantage: In some scatterplots, elements may be concentrated in a small area.
- Completely different scales and transformations within one SPLOM, e.g. linear, log-linear, log-log, … are very hard to interpret

Scatterplot Matrices and GPLOMs

Generalized Pair Plot Matrix enables also the display of categorical variables. Pairs may represent

- two scalar values,
- a scalar and a categorical value and
- two categorical values (Im, 2013).



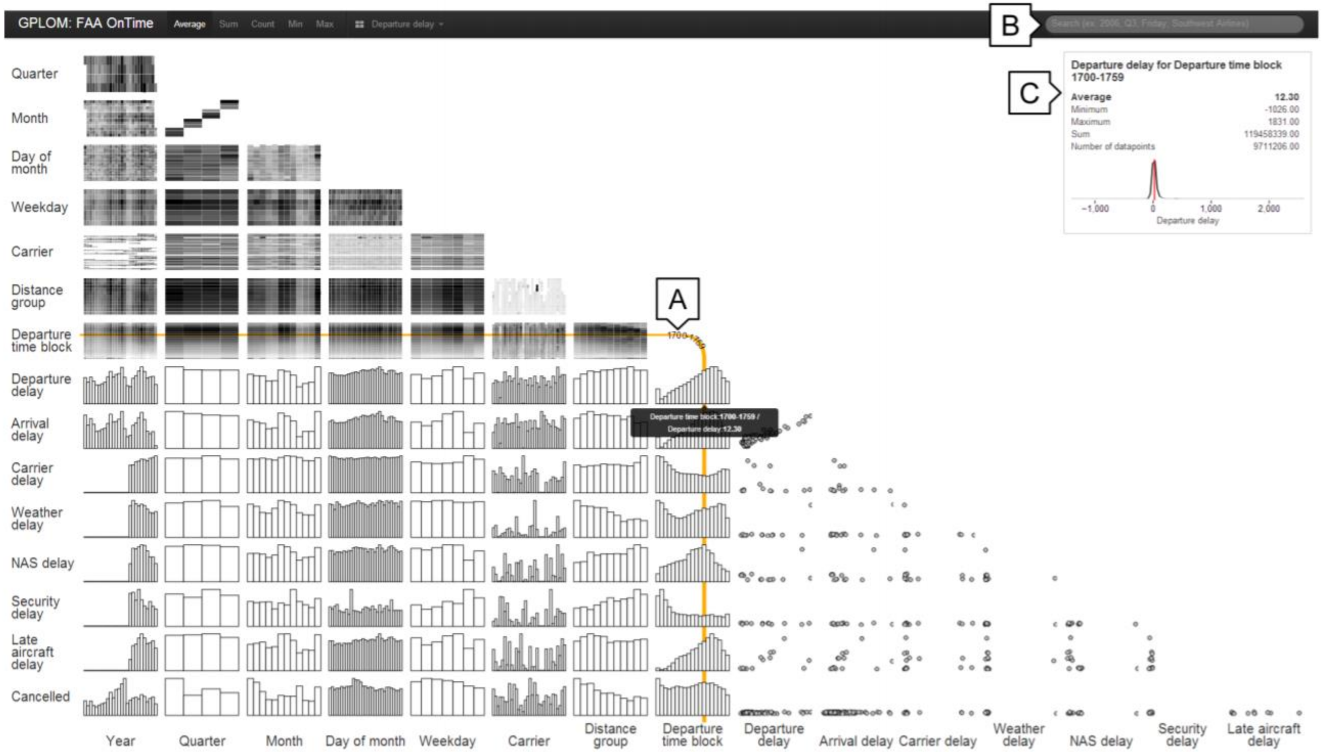
Categorical variables x_1, x_m and scalar variables y_1, y_m (From: Im, 2013).

Blue area → scatter plots are used.

Green area → bar charts (each bar represents one value of the categorical variable).

Red area → heatmaps are used.

Scatterplot Matrices and GPLOMs



Departure time block:1700-1759 /
Departure delay:12.30

Departure delay for Departure time block 1700-1759

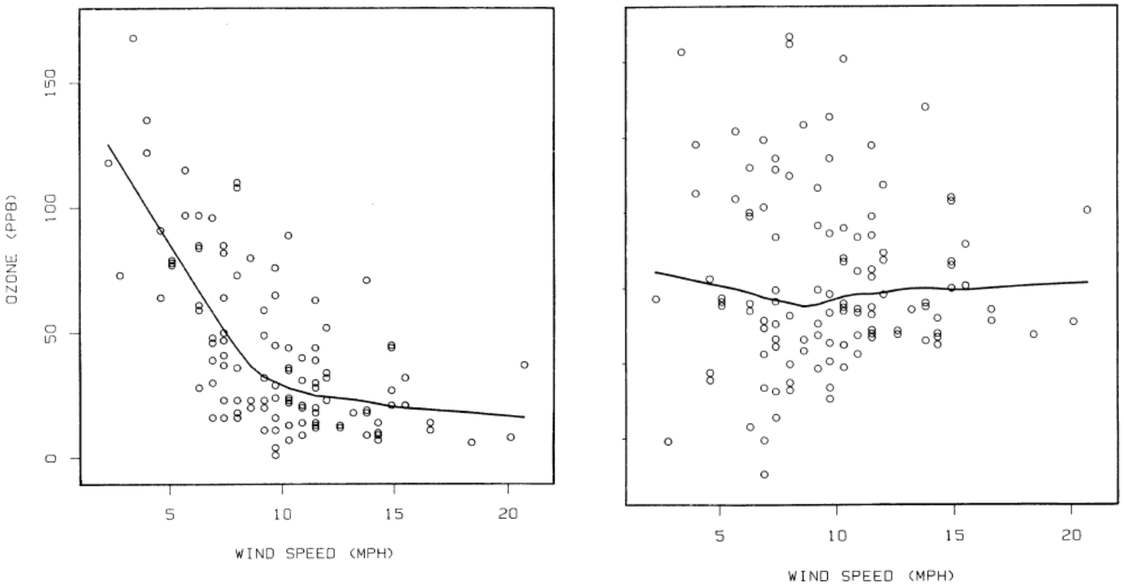
Average	**12.30**
Minimum	-1026.00
Maximum	1831.00
Sum	119458339.00
Number of datapoints	9711206.00

A generalized pair plot matrix supporting comparisons of scalar, categorical and mixed scalar/categorical distributions (From: Im, 2013).

Scatterplots and Regression

- Since SPs serve to assess the influence of x on *y*, frequently regression lines are computed.
- These lines reflect a linear global correlation.
- Correlations are often only locally linear, strongly affected by outliers or non-linear.
- Thus, local regression may be computed.
- Robust local regression restricts the influence of outliers.
- If no strong local correlation exists, a quadratic correlation may be analyzed.

Scatterplots and Regression

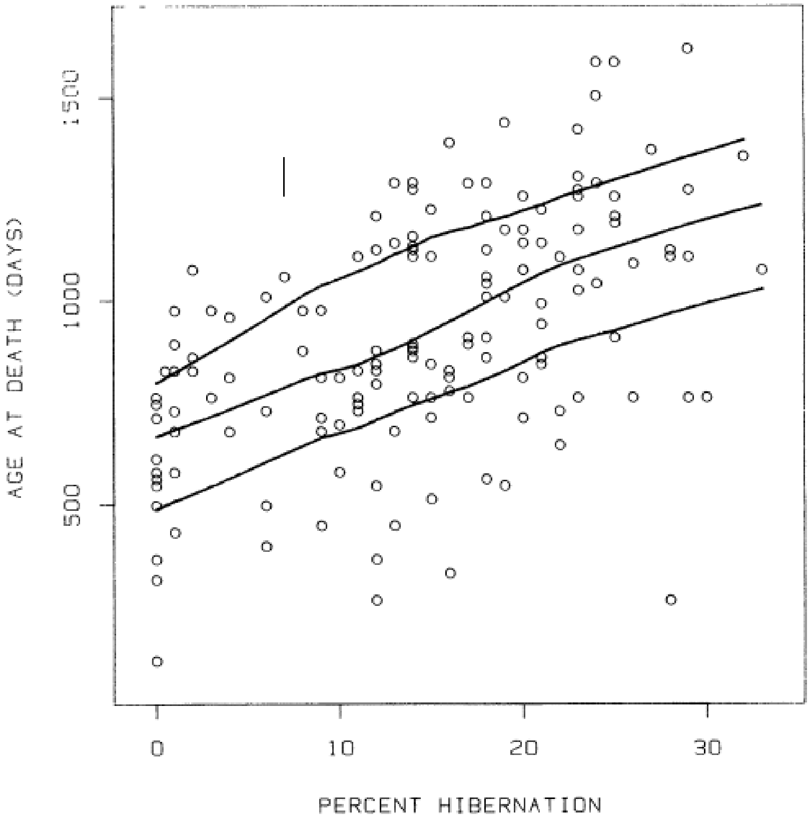


(From: Cleveland, 1984).

Robust local regression lines („smoothings“) are computed. Right: To indicate the deviation, the display is adapted (deviations to the line are shown).

Local regression: for each x-value, compute a weighted average of *n* points with the most similar x-values (least square fitting) (From: Cleveland, 1984).

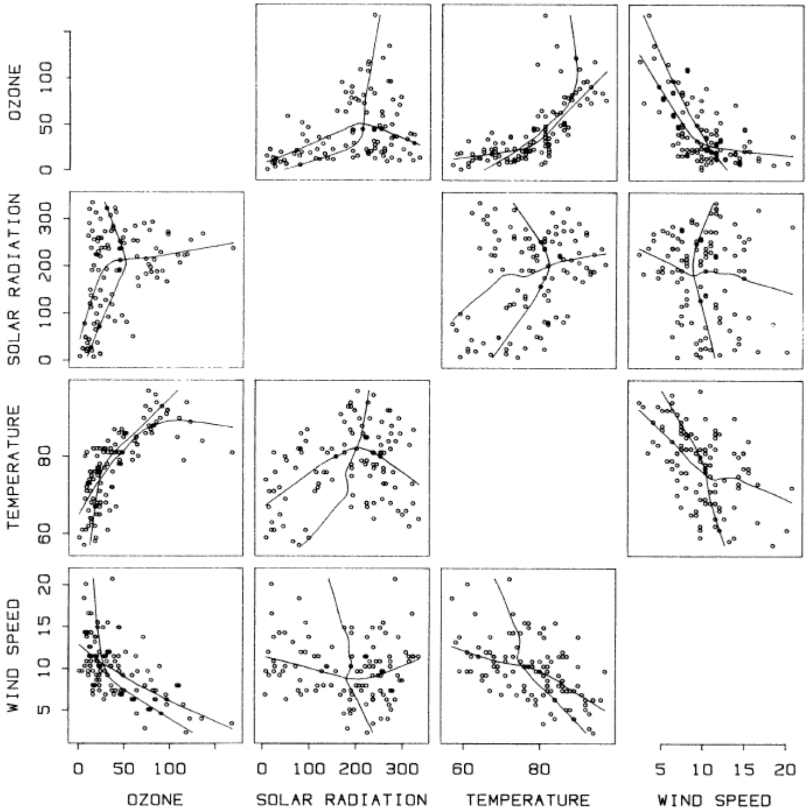
Scatterplots and Regression



(From: Cleveland, 1984).

Low, middle and high “smoothings” are determined. For each x-value, the same values are considered and split into 1/3 with the highest, middle and lowest y-values. A linear correlation in all regions results.

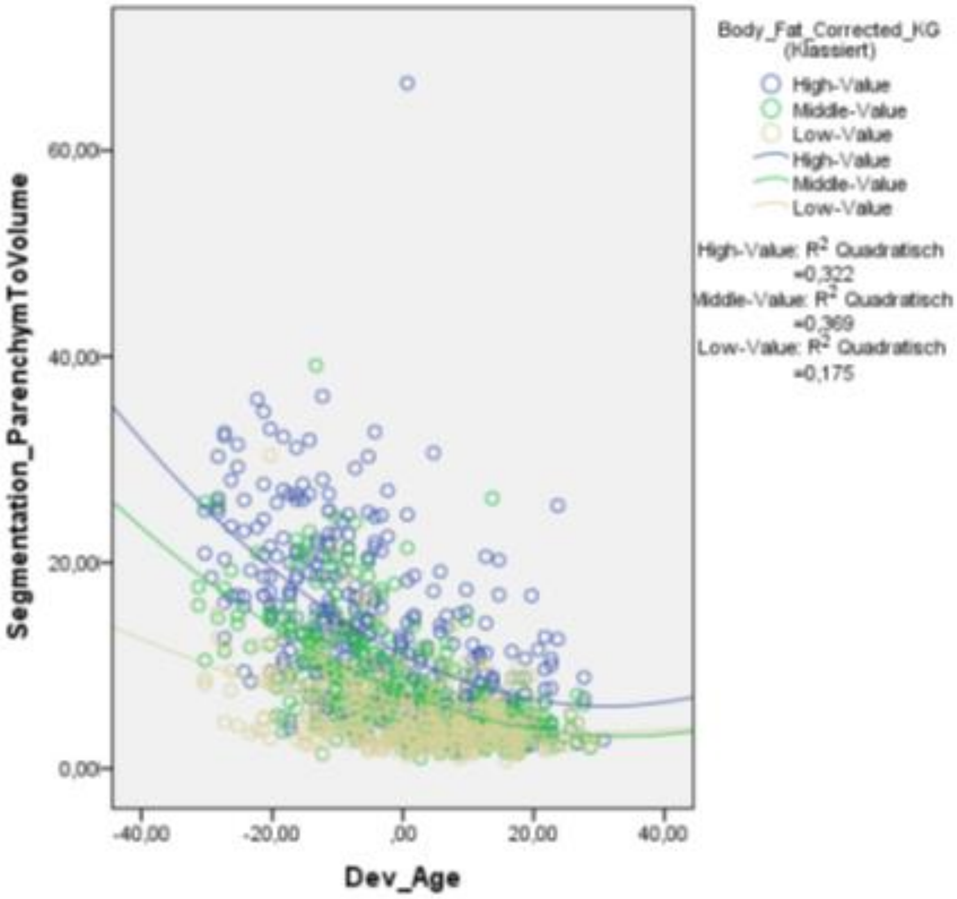
Scatterplots and Regression



(From: Cleveland, 1984).

For multiclass data, a comparison of „smoothings“ supports interpretation. Technique may be extended to SPLOMs.

Scatterplots and Regression

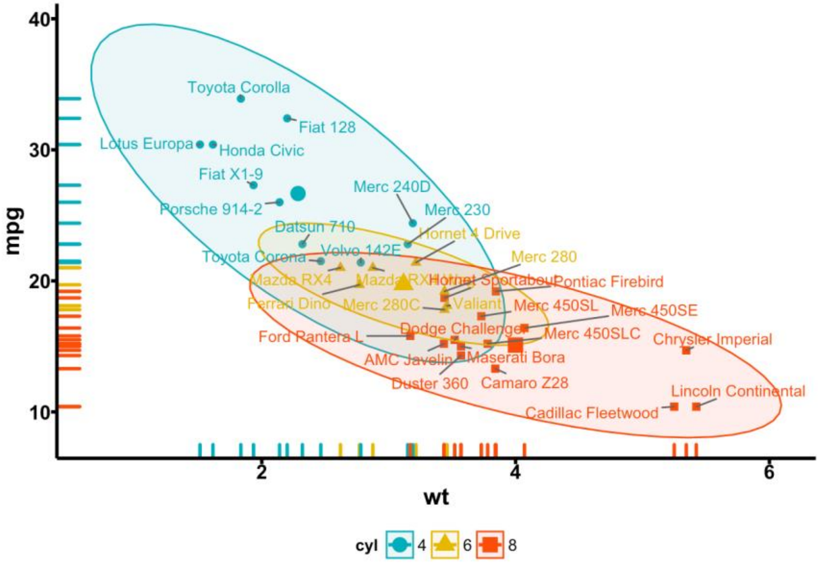


Relation between two variables and a discrete moderator variable (three categories) is displayed along with best fitting quadaratic regression (Courtesy of Mehdi Alborzi)

Advanced 2D Scatterplots

Concentration ellipses: A more abstract depiction of (multiclass) distributions.

- Filled (transparent) ellipse with solid outline surrounding data points (of a class).
- Midpoint may be explitly shown.
- Outliers (mild or severe) may be excluded. Often ellipses surround 95% of the data.



Scatter plot with concentration ellipses and midpoint. Classes represent cars with 4, 6 and 8 cylinders. Low element number.

Advanced 2D Scatterplots

Discussion: Concentration ellipses:

- Good perceptual motivation: Contours are easily perceived if they are „locally linear and globally circular or elliptical“ (Kuai, 2006).
- Smooth contours with solid fill are beneficial.
- Abstracted visualization. For larger number of datasets, individual elements are not visible.
- Interaction required to explore data locally.

Advanced 2D Scatterplots

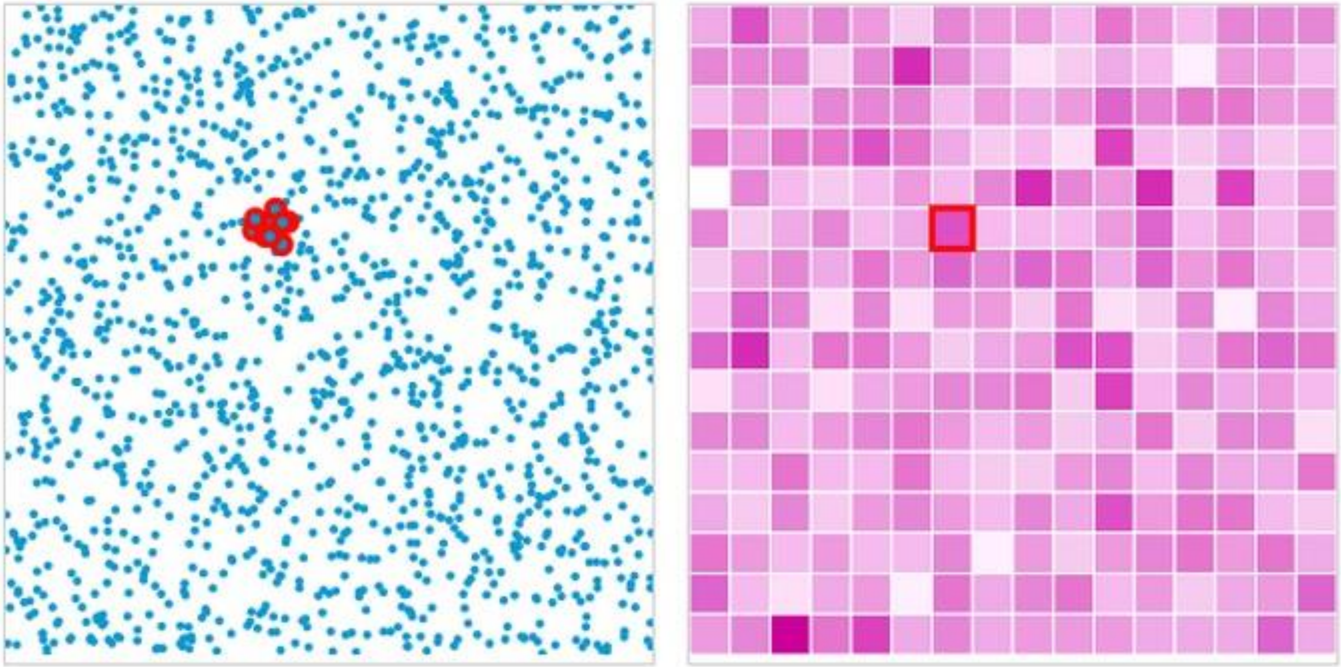
Binning

- technique of data aggregation used for grouping a dataset (abstraction from individual elements)
- Used, e.g. in histogram generation

Density plots:

- Binning is based on a grid (quadrilateral, hexagonal, …) and indicates frequency/density.
- Bin size is the major parameter.
- Density plots
  - + Avoid clutter and overplotting in dense regions
  - - do not convey distribution in sparsely populated regions well.

Advanced 2D Scatterplots: Binning

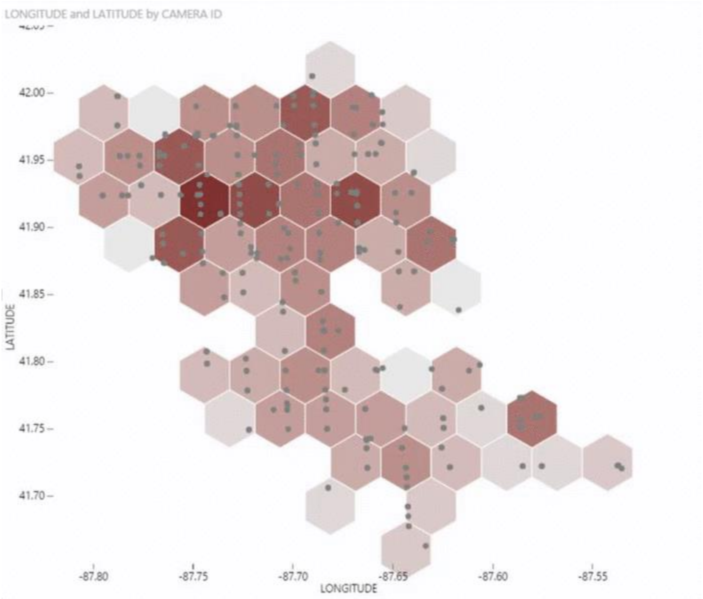


A scatterplot and its rectangular binned approximation (**discrete** density plot). The highlighted points (left) correspond to the selected cell right. Darkness indicates frequency (From: Link).

Advanced 2D Scatterplots: Binning

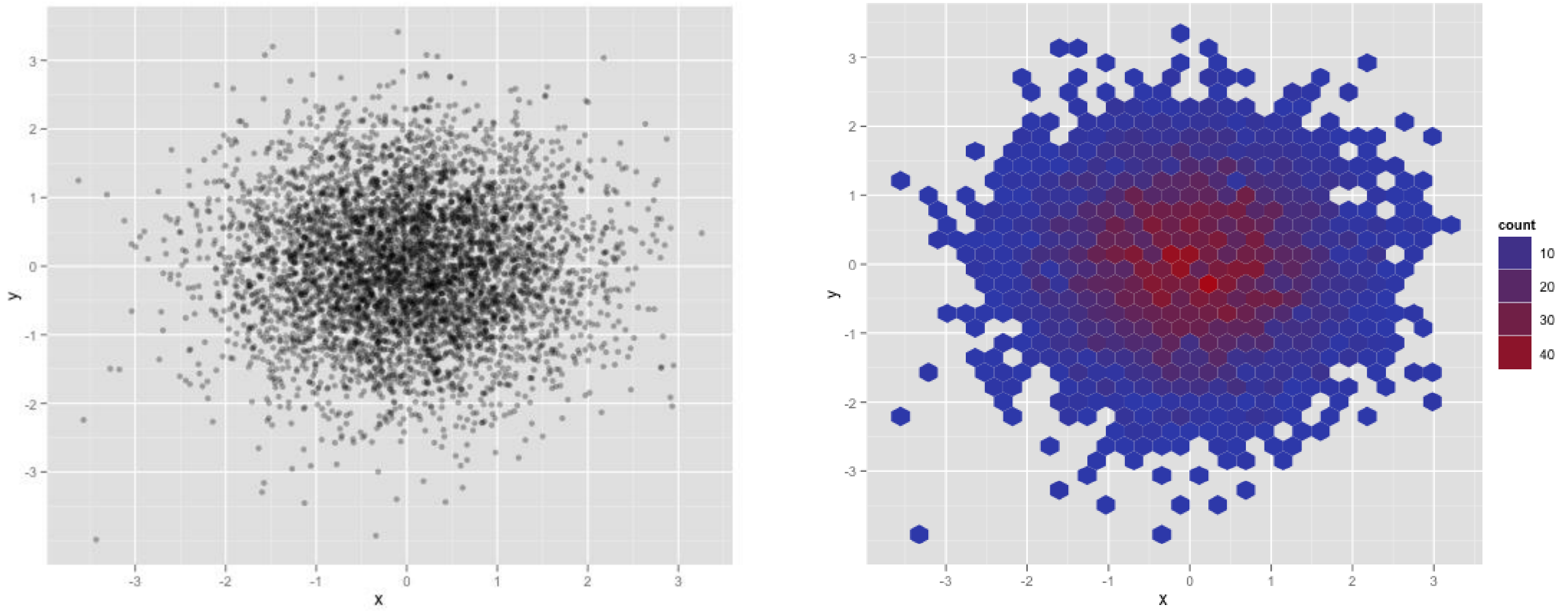
Hexbin scatterplot:

- clustering points onto a uniform grid of hexagons.
- uses variation in bin color similar to a heat map.
- higher saturation or darker color indicates higher density



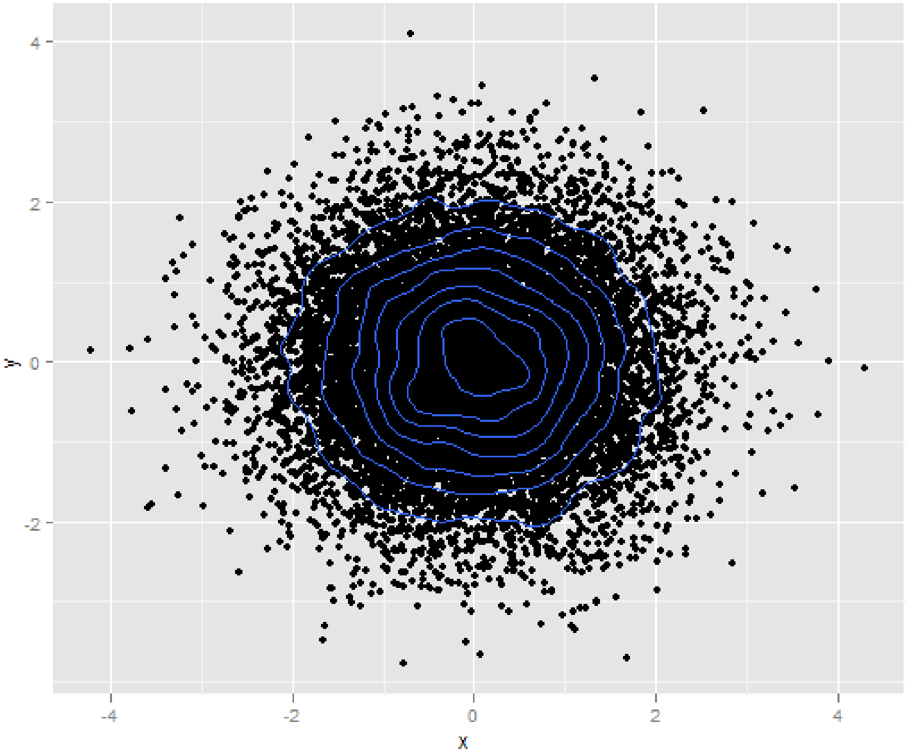
HexBin Scatterplot (From: Link, see Also the video)

Advanced 2D Scatterplots: Binning



Comparison of a transparency-adjusted scatterplot and a hexbin plot. Hexbin plots enable faster interpretation but do not indicate all details (From: Link)

Advanced 2D Scatterplots: Binning



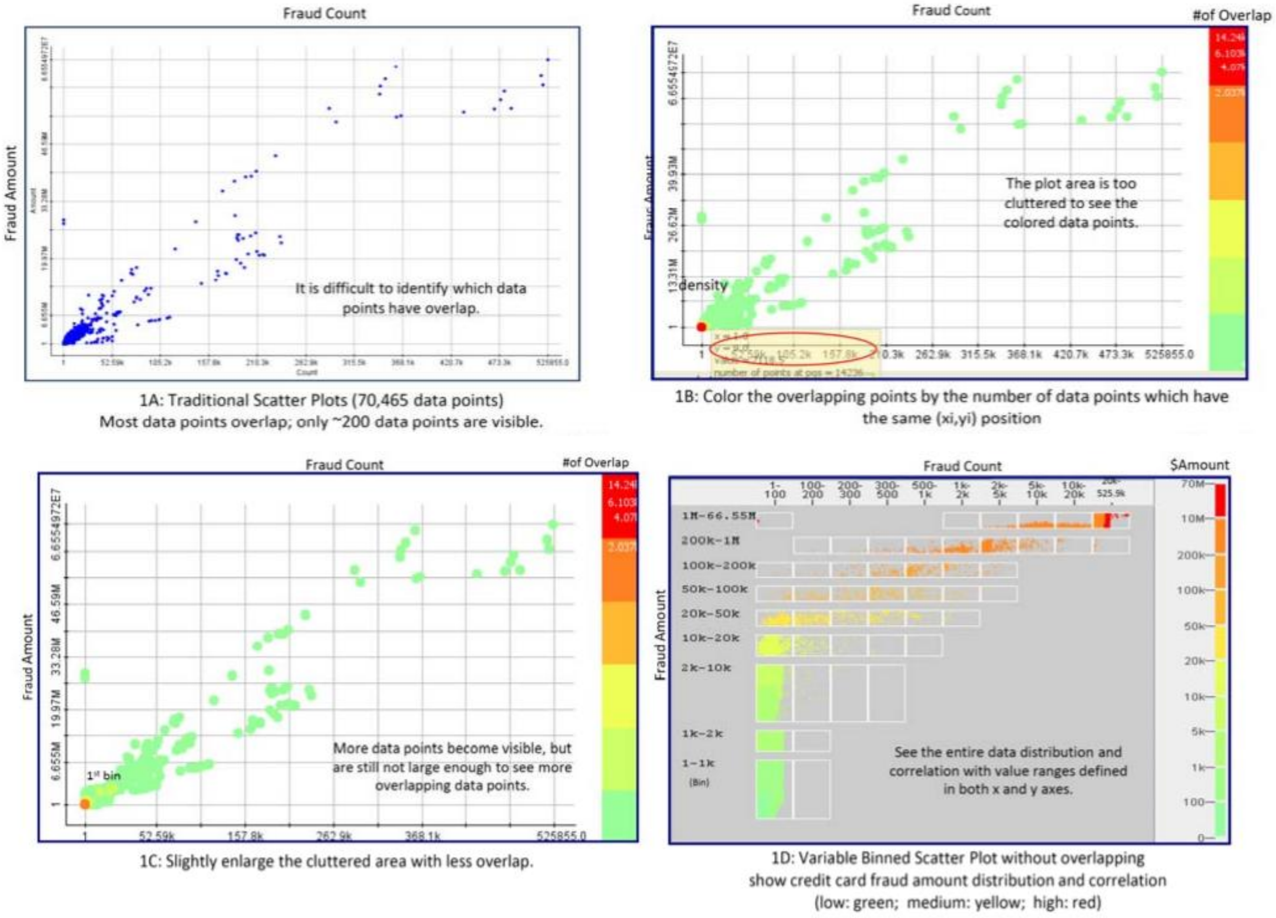
Density contour plot: Each contour encloses a certain percentage of the data, e.g. 10% (From: Link)

Advanced 2D Scatterplots: Binning

Variable Binned Scatterplots (Hao, 2010):

- Data is binned in *adaptive* ranges
- Density in these bins is determined and color-coded
- Interaction: zoom in and out to adjust level of detail
- Extension: binned scatterplot matrices

Advanced 2D Scatterplots: Binning



It is difficult to identify which data points have overlap.

1A: Traditional Scatter Plots (70,465 data points)
Most data points overlap; only ~200 data points are visible.

The plot area is too cluttered to see the colored data points.

1B: Color the overlapping points by the number of data points which have the same (xi,yi) position

More data points become visible, but are still not large enough to see more overlapping data points.

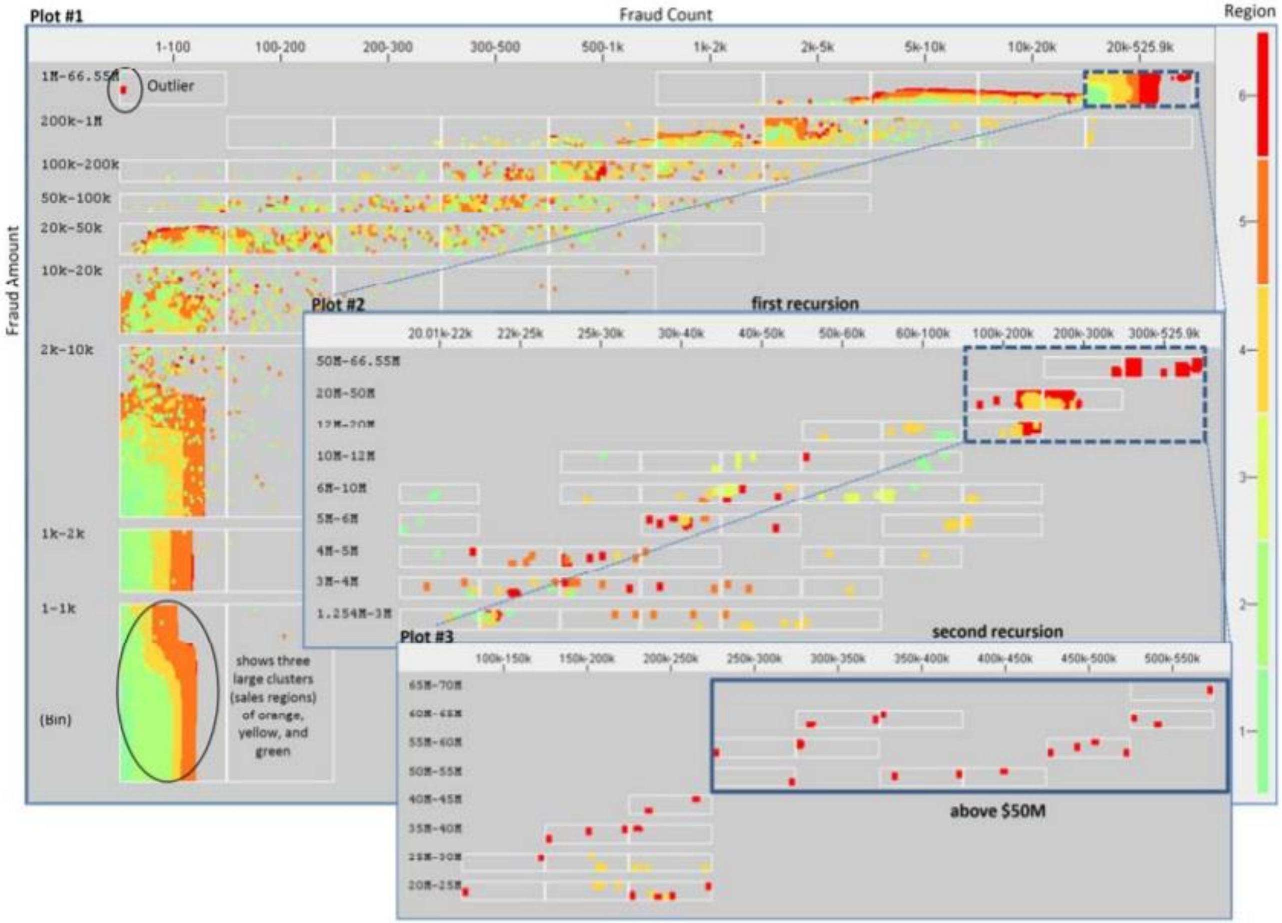
1C: Slightly enlarge the cluttered area with less overlap.

See the entire data distribution and correlation with value ranges defined in both x and y axes.

1D: Variable Binned Scatter Plot without overlapping show credit card fraud amount distribution and correlation (low: green; medium: yellow; high: red)

Construction of variable binned scatterplots from classic scatterplots (From: Hao, 2010).

Advanced 2D Scatterplots: Binning



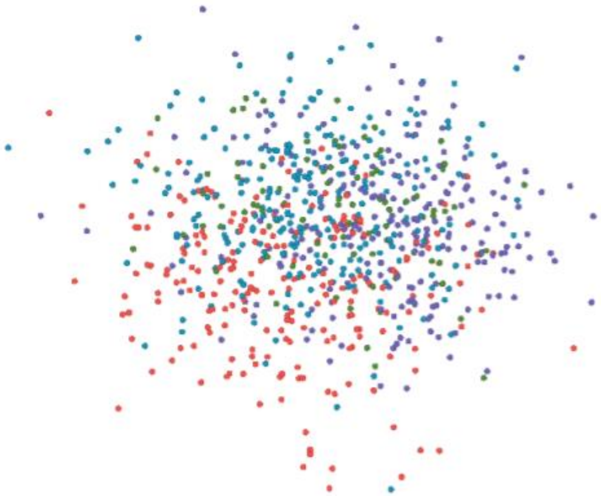
Variable Binned Scatterplot. Fraud count and fraud amount in six different regions (mapped to color). (From: Hao, 2010).

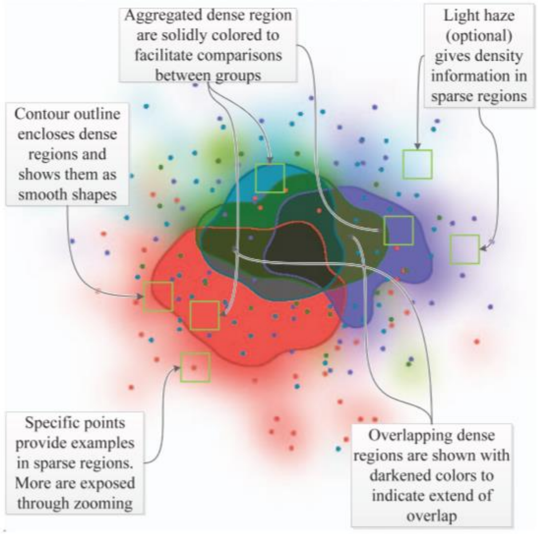
Advanced 2D Scatterplots: Splatter Plots

Splatter plot: Improving class separability with smoothly overlapping shapes (Mayorga, 2013)

Creation of splatter plots involves:

- Kernel density estimation (KDE)
- Color blending





Left: poor class separability with a conventional SP.

Right: based on KDE and a density threshold overlapping shapes are generated (From: Mayorga, 2013).

Advanced 2D Scatterplots: Splatter Plots

Visualization of densely populated regions:

- General concept (Fekete, 2002): Connect densely packed points to shapes
- Density: computed as a smooth function over the whole domain
- In contrast to binned density plots, no (adaptive) grid is involved.
- Regions where the density per class is above a threshold are shown as transparent and colored polygons with contours (**continuous** density)
- Polygons may be smoothed to enhance perception of clustered dense regions.
- Color modulation in overlapping regions

Advanced 2D Scatterplots: Splatter Plots

Visualization of Sparse regions:

- Individual points are shown on top of strongly transparent background
- Points are selected based on a filtering and sampling strategy to avoid too many distracting outliers
- Selection of points is guided by the goals to show representative points and to consider perceptual issues (avoid overload and clutter)

Advanced 2D Scatterplots: Splatter Plots

Kernel density estimation at point *q* considering *n* nearest neighbors

$$\hat{f}(\mathbf{q}) = \frac{1}{n}\sum_{i=1}^{n} K_B(r_i),$$

K – kernel function, *B* - bandwidth parameter,
r_i – distance to *q*

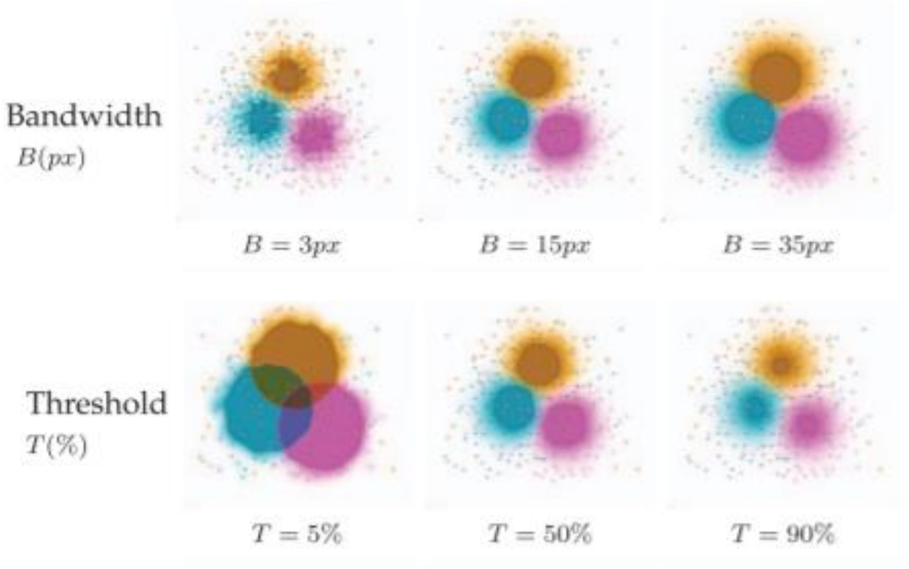
Typical choice for *K*: Gaussian, for *B*: variance of the Gaussian

$$K(r) = \frac{1}{\sqrt{2\pi}} e^{\frac{-r^2}{2\sigma^2}},$$

Advanced 2D Scatterplots: Splatter Plots

Major parameters of splatterplots:

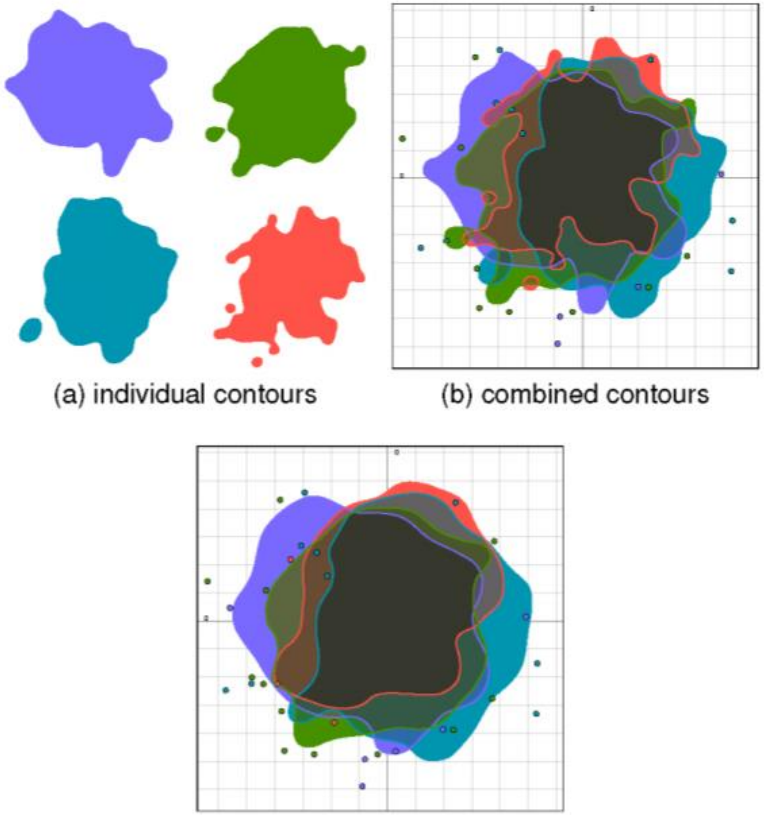
- Density Threshold, bandwidth, point filtering in sparse regions, contour width, smoothing



Influence of different splatter plot parameters for an artificial dataset
(From: Mayorga, 2013).

Advanced 2D Scatterplots: Splatter Plots

Splatterplots: Smoothing


(a) individual contours

(b) combined contours

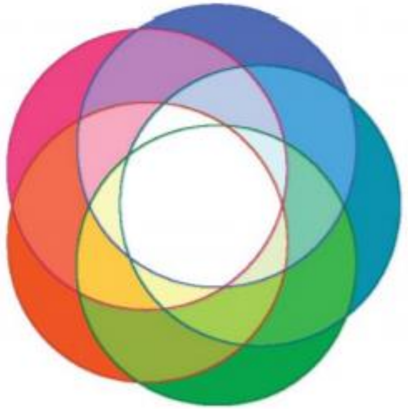
(c) increased smoothing

Smoothing supports interpretation of (overlapping) shapes (recall perceptual discussion of concentration ellipses)
(From: Mayorga, 2013).

Advanced 2D Scatterplots: Splatter Plots

Colors:

- Colors for each class are chosen in LAB space to be as different as possible
- Colors for overlapping regions are weighted interpolations in LAB space
- Overlapping colors are modified to indicate the amount of overlapping (increased lightness and decreased saturation)
- Color scheme is useful for any kind of set-based vis.

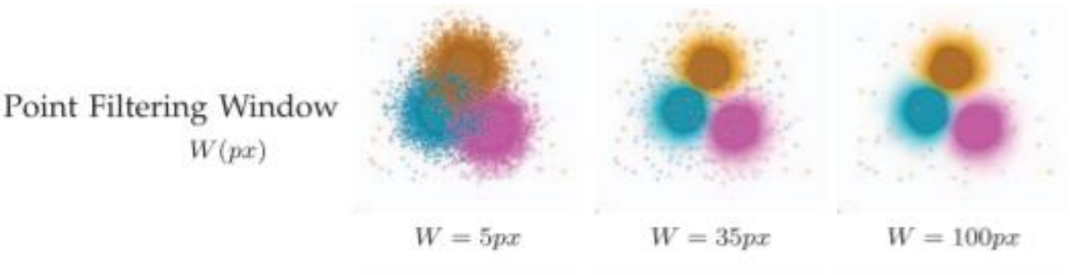


Automatic selection of five colors and up to three overlaping shapes (From: Mayorga, 2013).

Advanced 2D Scatterplots: Splatter Plots

Representation of Sparse Regions:

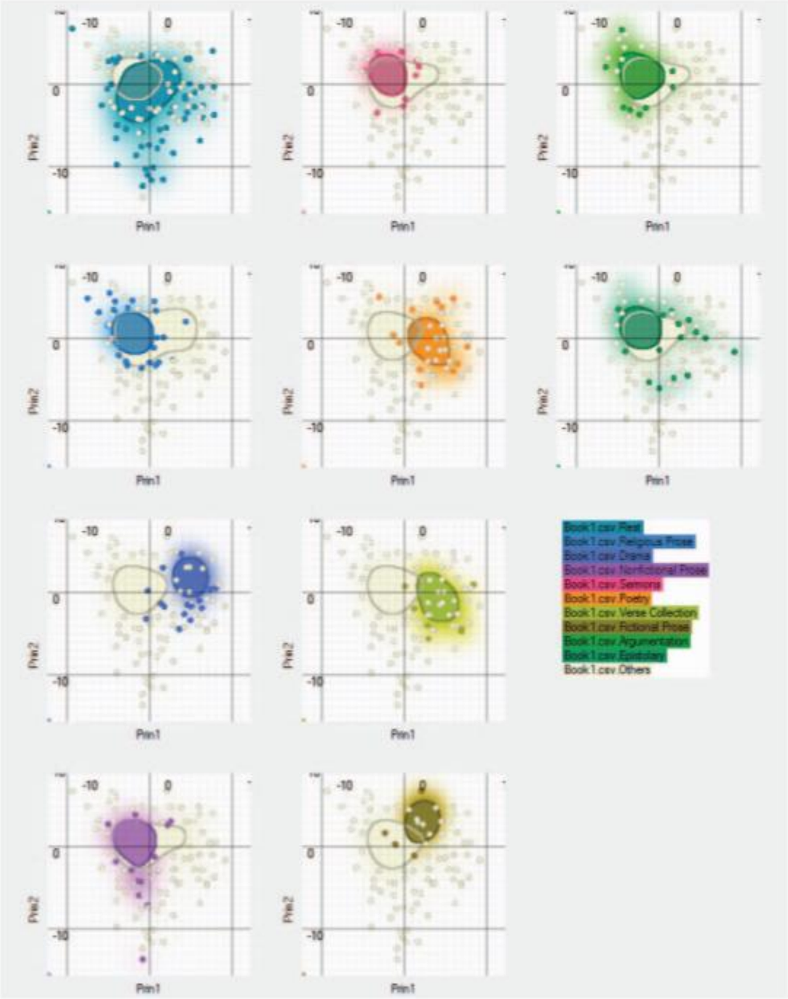
- Control of information density in screen space
- A filter is applied to avoid that
  - points are rendered too closely to each other
  - Points are rendered close to contours



Different amount of point filtering (From: Mayorga, 2013).

Advanced 2D Scatterplots: Splatter Plots

- Splatter plots may also be used to compare one class/cluster against all remaining elements.
- Often, one class/cluster that differs most from the global mean is particularly interesting.



(From: Mayargo, 2013)

Advanced 2D Scatterplots:

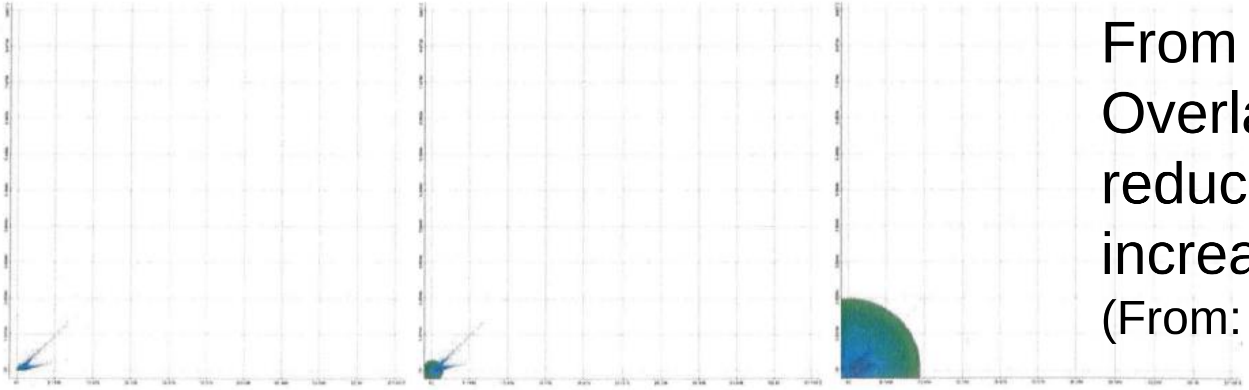
Generalized scatterplots (Keim, 2010):

- Provides an adjustable trade-off between an overlap-free but distorted view and an undistorted view with considerable overlap.
- Users indicate with a slider the amount of acceptable overlap.
- Interaction is essential to understand the data.

Advanced 2D Scatterplots:

Algorithmic realization:

- For any point in the output view it is checked how often it is used so far.
- If this counter is below the adjusted limit, the point is (over)drawn.
- Otherwise, the algorithm searches in the surrounding for a point with a counter below the threshold.
- The 2D representation is updated.



From left to right: Overlapping is reduced; distortion increased (From: Keim, 2010).

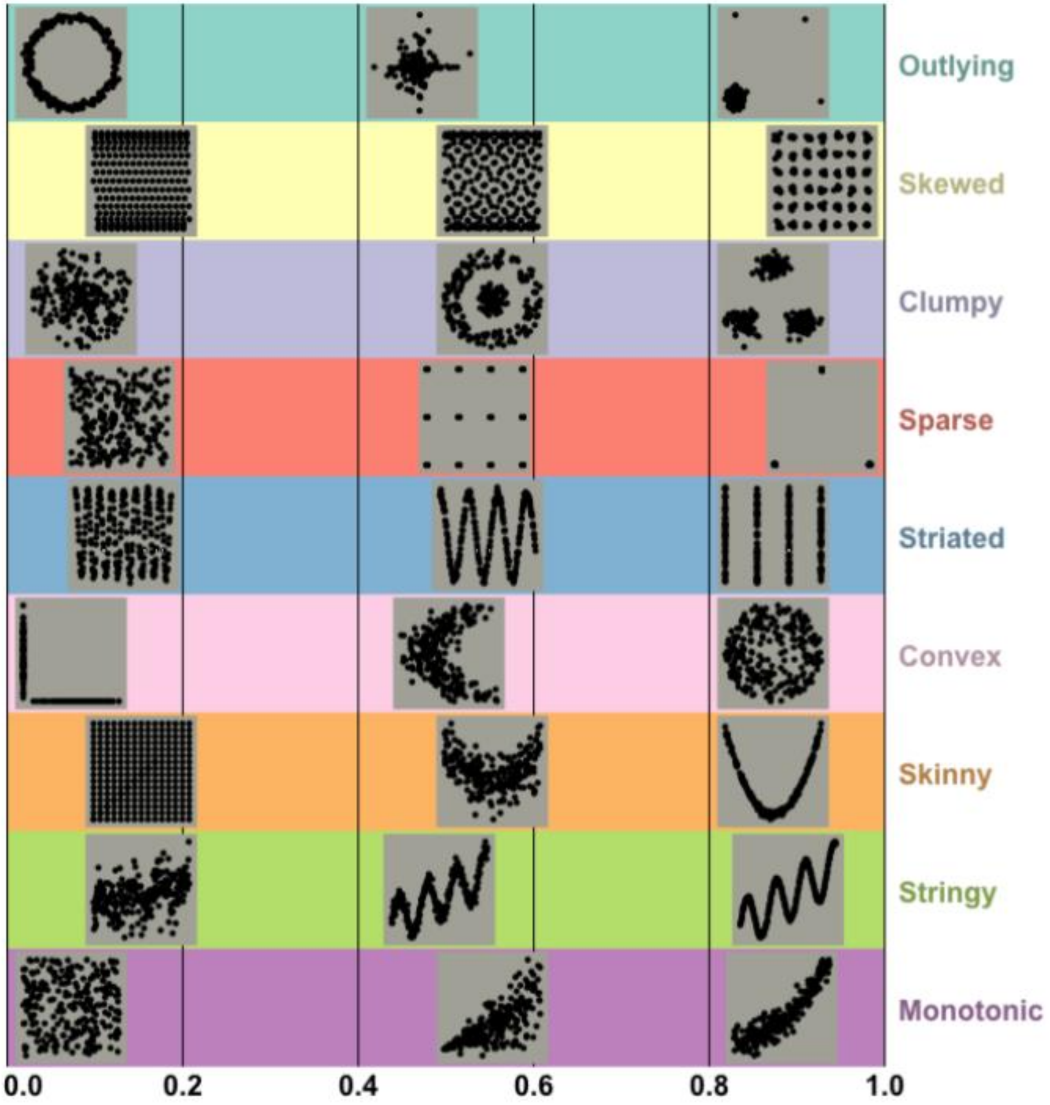
Scagnostics-Based Exploration

- Scagnostics – scatterplot and diagnostics (Wilkinson, 2005)
- Derives 9 attributes from a scatterplot

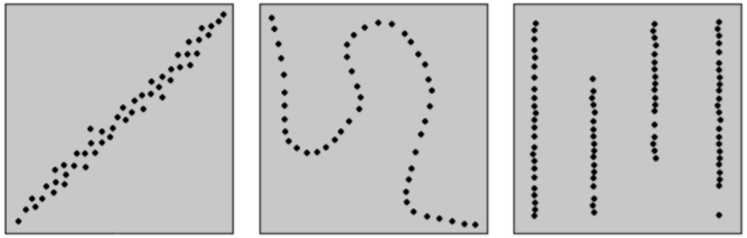
Scagnostic features may be used for (Dang, 2014)

- Exploration of large sets of scatterplots
- filtering, e.g. with monotonicity >0.5 and outlierness <0.5
- Similarity-based search, e.g. show clusters of scatterplots with similar features

Scagnostics-Based Exploration

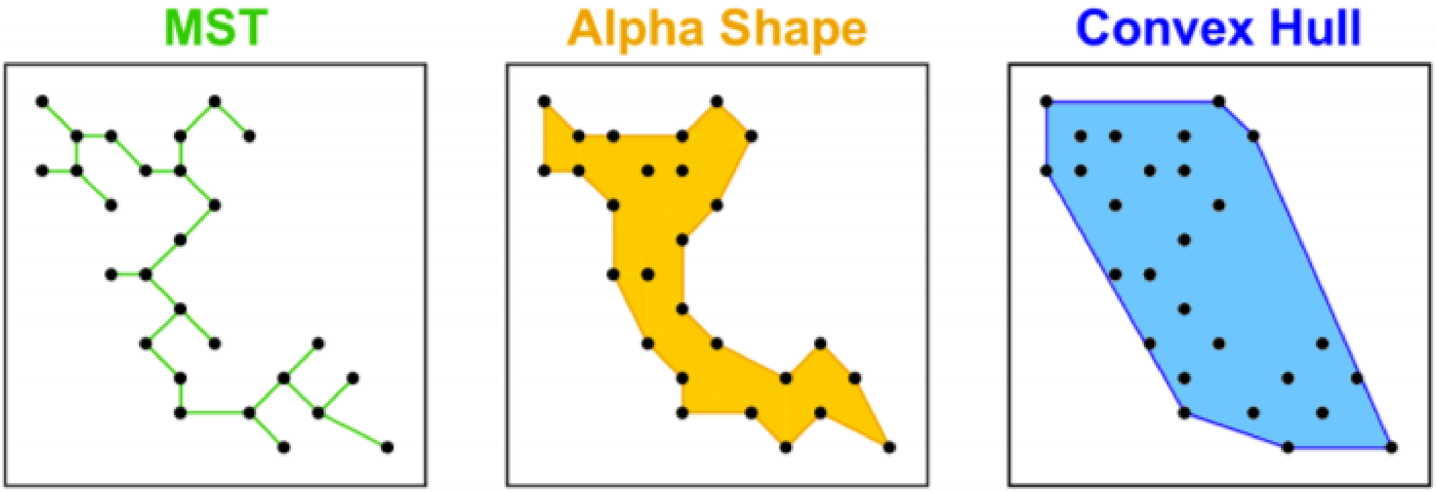


The nine measures with examples for low, moderate and high values (From: Dang, 2014).



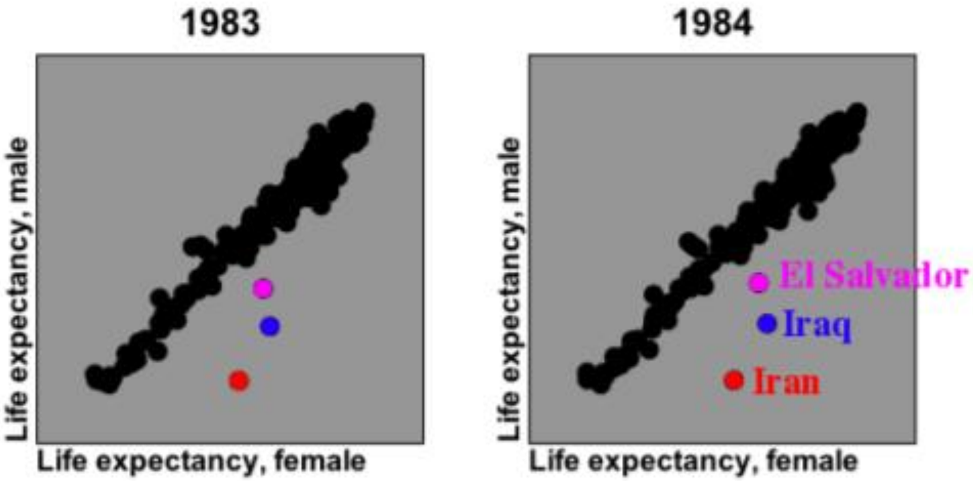
Scatterplots with high values for monotonicity, stringyness and striated distribution (From: Dang, 2014).

Scagnostics-Based Exploration



Scagnostics features are determined based on a geometrical analysis of the point distribution including Minimum Spanning Tree (From: Dang, 2014).

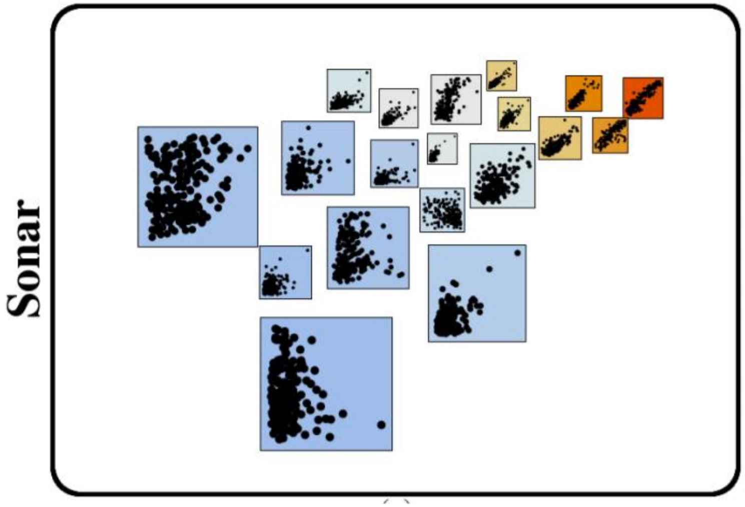
Scagnostics-Based Exploration

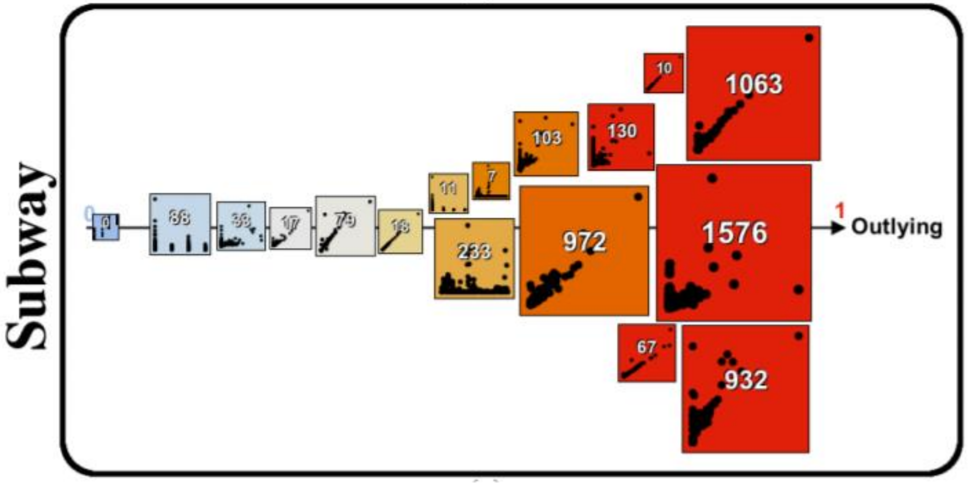


Filtering for high outlierness and high monotonicity. Life expecantancy of woman and men in ~150 countries are highly correlated. Due to war, in 3 countries, the male life expectancy is strongly lower (From: Dang, 2014).

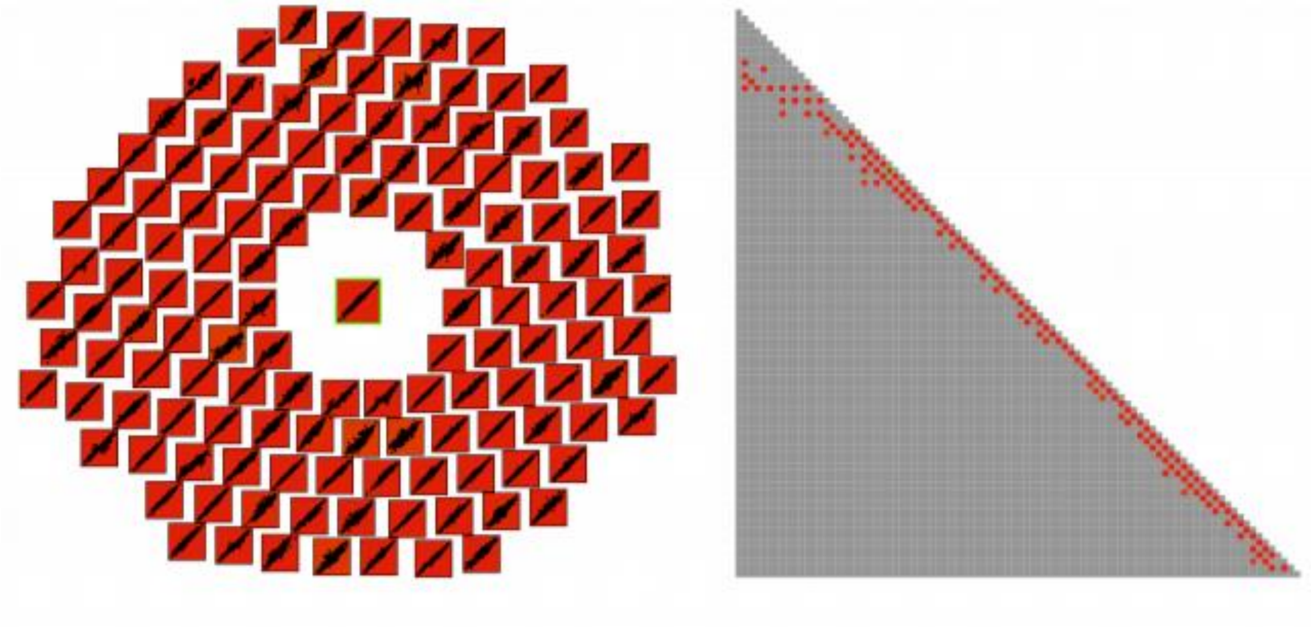
Scagnostics-Based Exploration

For large number of scatterplots: clustering and selection of relevant clusters. Rectangle size encodes cluster size.





Scagnostics-Based Exploration



The current cluster contains a few dozens from several thousand scatterplots (From: Dang, 2014).

Scagnostics-Based Exploration

Discussion:

- Scagnostics-based exploration is scaleable w.r.t. number of dimensions
- Not an end-user tool; requires collaboration between data scientist and domain expert.

Advanced 2D Scatterplots: Small Screens

Small Screens (Bühring, 2006):

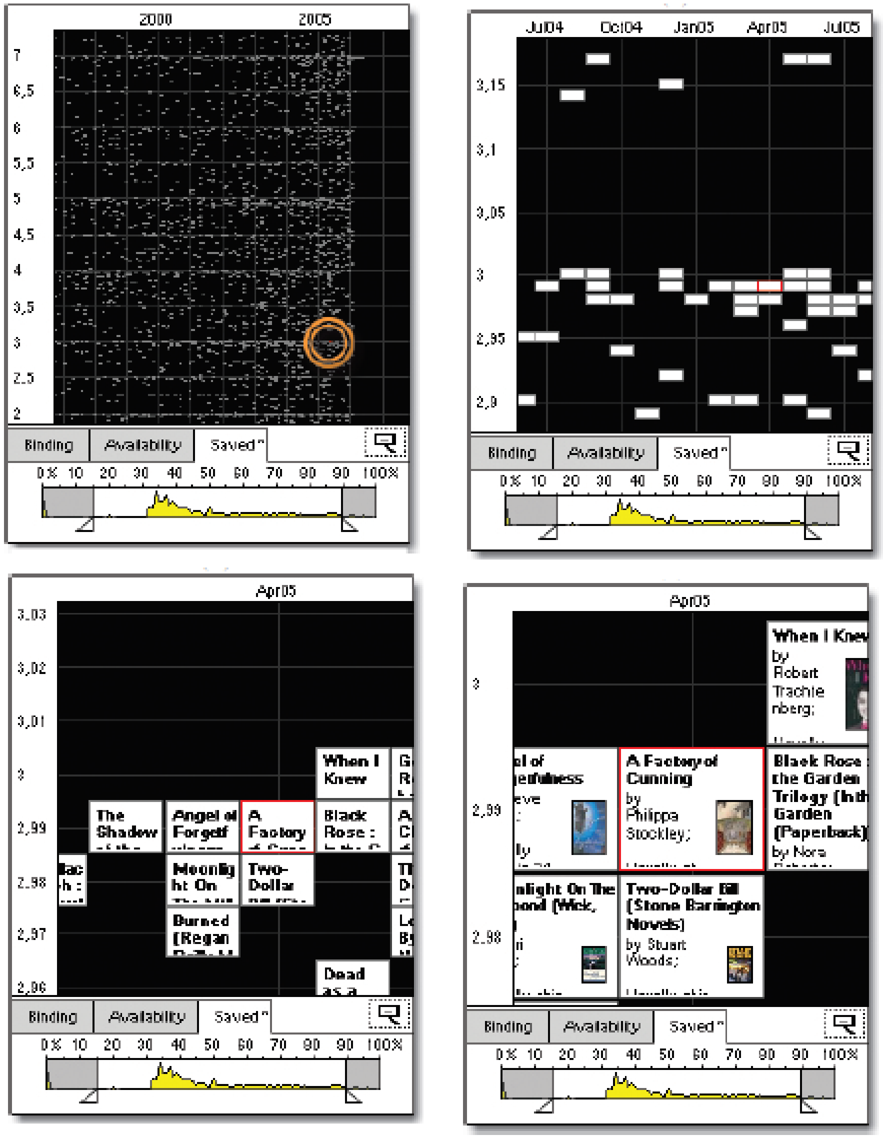
- Relevant for mobile experts, e.g. sales persons and real-estate experts (Immobilienmakler)
- Scatterplots with details-on demand presented in a separate window are not feasible for small screens
  → overview and detail need to be integrated in one screen
  → smooth transition between visualization with and without detail, use of fisheye zoom combining geometric distortions and semantic zoom (change representations)

Advanced 2D Scatterplots: Small Screens

Highlighting (Bühring, 2006):

- Zooming always targets at an item → Avoid that the user gets lost between items (desert fog)
- Continous zoom starts after a delay, e.g. 150 ms

Advanced 2D Scatterplots: Small Screens



A small screen with a scatter-plot.

Zooming - the representation of the elements changes towards rectangles, labeled rectangles and more detailed representations. Elements of a book database are explored (From: Bühring, 2006).

Animated Scatterplots

- Humans are familiar with analyzing animated movements and *observe* changes.
- Human motion perception is restricted by the amount of moving objects and the complexity of movements.
- Fast movements are perceived as urgent; slow movements in the periphery may be overlooked (change blindness)
- Few items, moving linearly or on smooth trajectories, following similar directions are easy to interpret.

Animated Scatterplots

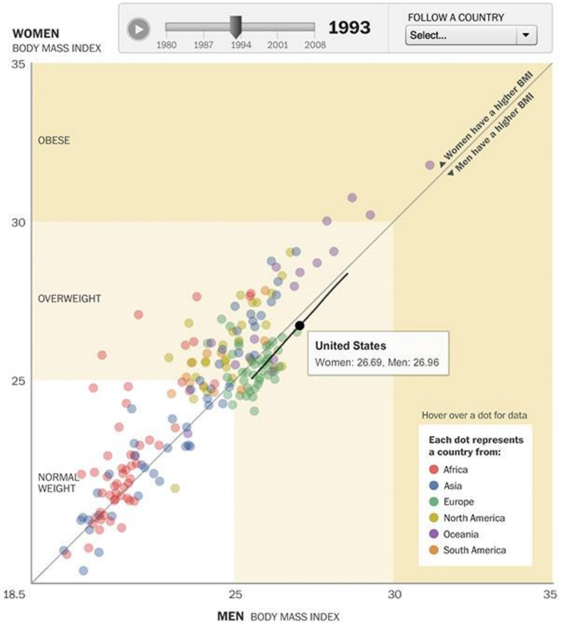
Animation may be incorporated

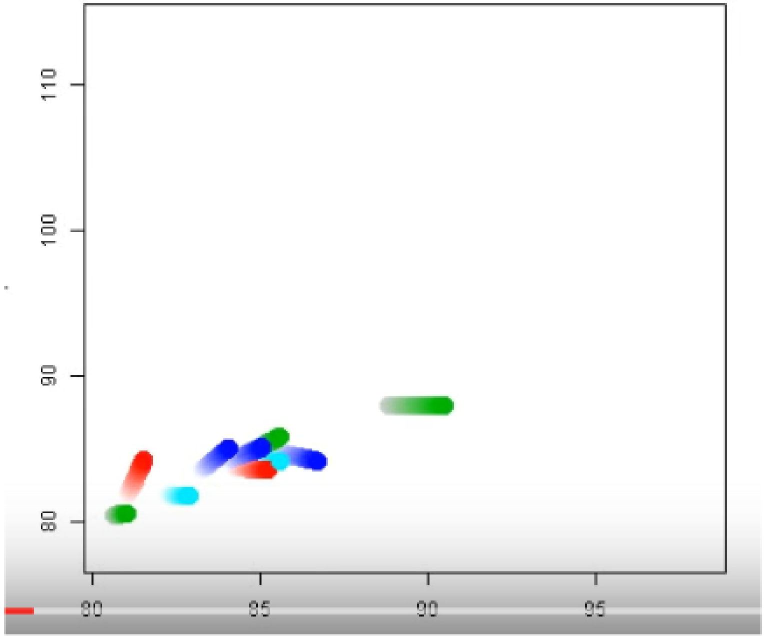
- to support interaction with a *static dataset* (animate the transition to a new adjustment, e.g. w.r.t. filtering, sampling, distortion)
- to support the exploration of *dynamic datasets*, e.g. changes of sugar consumption and BMI over time

Although the latter is more natural, both are useful.

For dynamic data: Users need „tight control over animation speed" to inspect relevant intervals in detail and skip over others (Craig, 2005)

Animated Scatterplots

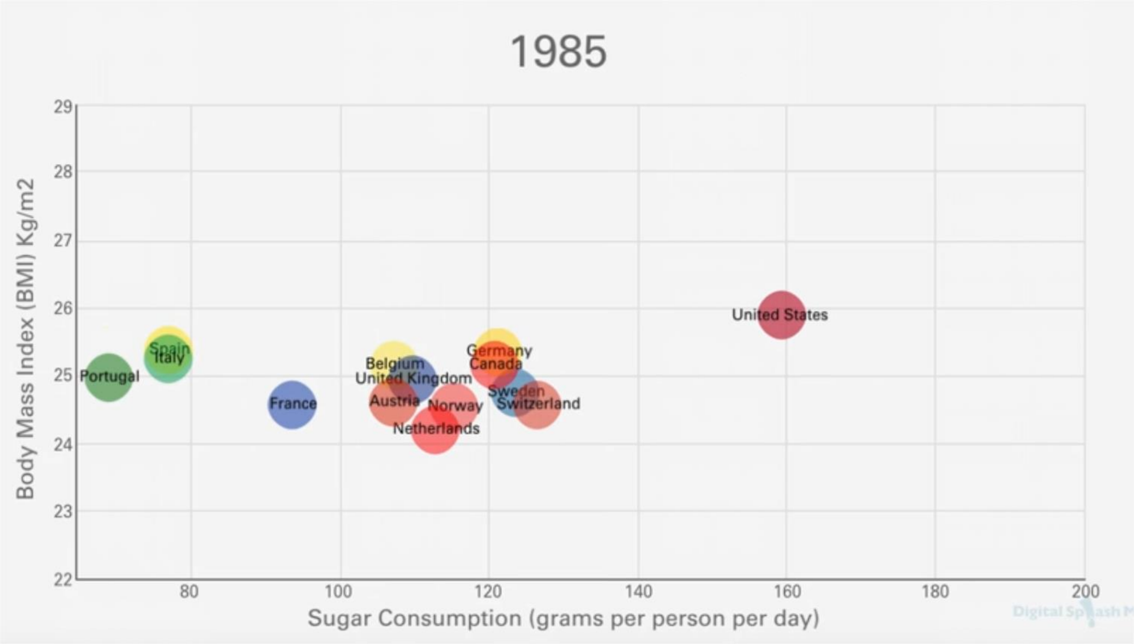




Left: User may select a timepoint or use the „play“ button to start an animation (From: Link)

Trails support perception of movement. Technique for few datasets (From: Link, see also Link)

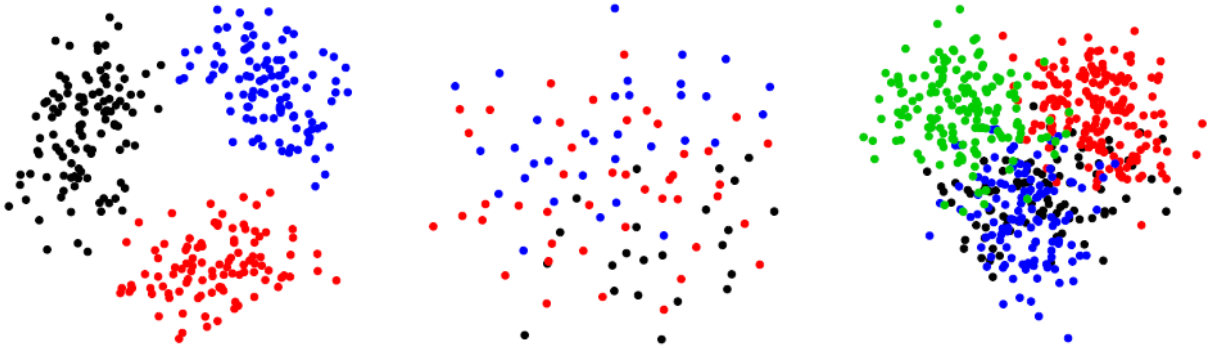
Animated Bubble Charts



Bubble Chart: Instead of a point, a circle represents a scalar value (Sreenshot from Link showing the correlation between 1980 and 2004).

Evaluation

- Comparison of Scatterplot-Based Displays
- Selection of a representative set of test data Menge (with different types and strength of correlations, clusters, outliers)
- Criteria (Sedlmaier, 2013):
  - Visual Separability of Clusters
  - Correct assessment of correlations
  - Trust (certainty) of interpretation



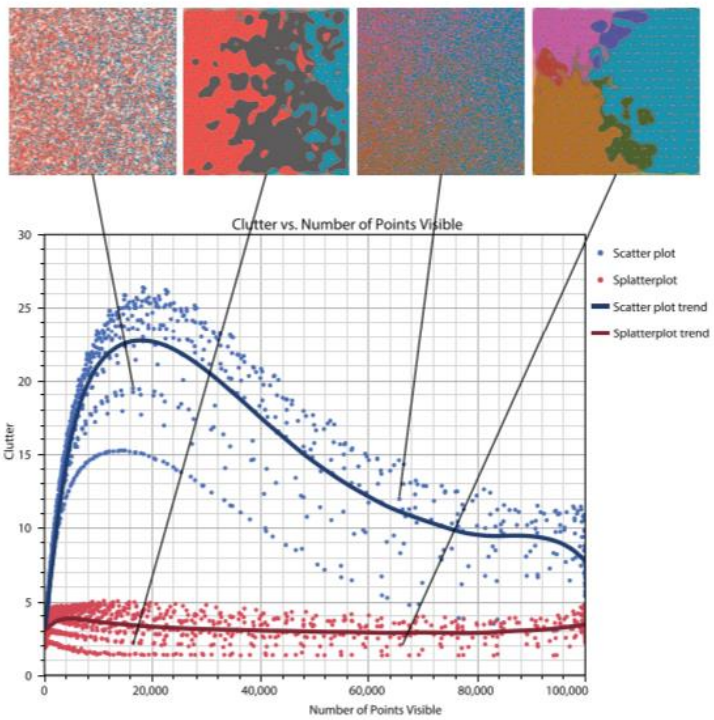
(a) max sep: (5,5,5) (b) no sep: (1,1,1) (c) mixed: (4,3,2,1)

Evaluation w.r.t. separability based on three datasets

(From: Sedlmaier, 2013)

Evaluation

A conventional scatterplot and splatterplots are compared w.r.t. a visual clutter metric (Rosenholtz, 2007) for a wide range of data. Splatterplots have a lower and almost constant trendline.



Red indicates splatterplots; right scatterplots. Solid lines represent the trend
(From: Mayorga, 2013)

Evaluation

Some details on metrics of visual clutter (Rosenholtz, 2007):

- Such metrics try to predict,
  - how complex a number of human observers assess a visualization
  - how long the user needs to find specific items (relation to Visual search theories)
- Metrics include item number, redundancy and grouping of items, contrast and saliency.
- Most metrics are quite general (also applicable to photos)

A major application is the comparative assessment of information visualization techniques and user interface designs (where different layouts with buttons, sliders, etc. are compared).

Summary

- Scatterplot-based displays are a frequent technique for information visualization
- Scatterplots are often enhanced with cluster analysis and regression analysis (visual analytics)
- Overplotting is the essential scaleability problem increased by a large number of datasets and/or a large number of dimensions (SPLOMs)
- Density plots, jittered placement, variable binning are scaleable variants based on aggregation of values or distortion
- Interaction is essential for advanced variants, e.g. to control the trade-off between distortion and overlap

References

E. Bertini, G. Santucci: „Quality Metrics for 2D Scatterplot Graphics: Automatically Reducing Visual Clutter". *Proc. of Smart Graphics* 2004: 77-89

E. Bertini, G. Santucci: „Improving 2D Scatterplots Effectiveness through Sampling, Displacement, and User Perception", *Proc. of Information Visualization* 2005: 826-834

E. Bertini: **"**Quality Metrics in High-Dimensional Data Visualization: An Overview and Systematization", *IEEE Trans. Vis. Comput. Graph.* 17(12): 2203-2212 (2011)

T. Büring, J. Gerken, H. Reiterer: User Interaction with Scatterplots on Small Screens - A Comparative Evaluation of Geometric-Semantic Zoom and Fisheye Distortion. *IEEE Trans. Vis. Comput. Graph.* 12(5): 829-836 (2006)

W. S.CLEVELAND and R. McGILL (1984). "The Many Faces of a Scatterplot", *Journal of the American Statistical Association*, Vol. 79(388)

Craig, P., J. Kennedy, A. Cumming. "Animated interval scatter-plot views for the exploratory analysis of large-scale microarray time-course data." *Information Visualization* 4.3 (2005): 149-163.

Dang Tuan Nhon, L. Wilkinson (2014). "ScagExplorer: Exploring Scatterplots by Their Scagnostics", *Proc. of IEEE PacificVis* 2014: 73-80

Doyle I, Ratcliffe M, Walding A, Vanden Bon E, Dymond et al. „Differential gene expression analysis in human monocyte-derived macrophages: impact of cigarette smoke on host defence", *Mol Immunol.* 2010 Feb;47(5):1058-1065.

JD Fekete, C Plaisant (2002). „Interactive information visualization of a million items", *Proc. of IEEE Symposium on Information Visualization*, pp. 117-124

S. Few (2008). Solutions to the Problem of Over-Plotting in Graphs, Link

References (II)

M. Gleicher, M. Correll, C. Nothelfer, S. Franconeri: „Perception of Average Value in Multiclass Scatterplots“, *IEEE Trans. Vis. Comput. Graph.* 19(12): 2316-2325 (2013)

M. C. Hao, U. Dayal, R. K. Sharma, D. A. Keim, H. Janetzko: „Variable binned scatter plots“, *Information Visualization* 9(3): 194-203 (2010)

S. Haroz, R. Kosara, S. L. Franconeri: The Connected Scatterplot for Presenting Paired Time Series. *IEEE Trans. Vis. Comput. Graph*. 22(9): 2174-2186 (2016)

J.-F. Im, M. J. McGuffin, R. Leung: GPLOM: The Generalized Plot Matrix for Visualizing Multidimensional Multivariate Data. IEEE Trans. Vis. Comput. Graph. 19(12): 2606-2614 (2013)

D. A. Keim, M. C. Hao, U. Dayal, H. Janetzko, P. Bak: „Generalized scatter plots“, *Information Visualization* 9(4): 301-311 (2010)

S. Kuai, C. Yu (2006). „Constant contour integration in peripheral vision for stimuli with good gestalt properties“, *Journal of Vision,* Vol. 9(2)

J. Li, J.-B. Martens, J. J Van Wijk (2010). „Judging correlation from scatterplots and parallel coordinate plots“. *Information Visualization* 9, 1 (2010),13–30

A. Mayorga, M. Gleicher (2013). „Splatterplots: Overcoming Overdraw in Scatter Plots“, *IEEE Trans. Vis. Comput. Graph.* 19(9): 1526-1538

A. V. Pandey, J. Krause, C. Felix, J. Boy, E. Bertini: „Towards Understanding Human Similarity Perception in the Analysis of Large Sets of Scatter Plots“, *Proc. of the ACM SIGCHI Conference on Human Factors in Computing Systems*, 2016, pp. 3659-3669

R. A. Rensink, G. Baldridge (2010). “The perception of correlation in scatterplots”, *Computer Graphics Forum*, Vol. 29, pp. 1203–1210

References (III)

R. Rosenholtz, Y. Li, L. Nakano (2007). “Measuring visual clutter”, *Journal of Vision*, Vol. 7(17)

Rousseeuw, P. J.; Ruts I.; Tukey J.W. (1999). "The Bagplot: A Bivariate Boxplot“, *The American Statistician*. Vol. 53 (4): 382–387

M. Sedlmair, T. Munzner, M. Tory (2013). „Empirical Guidance on Scatterplot and Dimension Reduction Technique Choices“, *IEEE Trans. Vis. Comput. Graph.,* Vol. 19, (12): 2634-2643.

A. Tatu, P. Bak, E. Bertini, D. A. Keim, J. Schneidewind (2010). „Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data“. *Proc. of Advanced Visual Interfaces (AVI), pp.* 49-56 (2010)

E. Tufte (1983). *The Visual Display of Quantitative Information*, Graphics Press

Leland Wilkinson, Anushka Anand, Robert L. Grossman (2005). “Graph-Theoretic Scagnostics“, *Proc. of INFOVIS* 2005, pp. 21-28