Exercise Sheet 8

1. What is Biclustering?

- Data mining technique that searches for subsets in bi-partite graphs
- Bi-partite graphs represent relations between two types of nodes, e.g. associations of people to organizations (parties, sport clubs, …) or researchers and publications
- Examples:
    - A set of students that take the same courses
    - A set of students coming from the same city
- 2D biclusters are linked to related by the students involved → biclusters reveal relations

2. Briefly describe visualization techniques for biclusters.

- Graph visualizations
    - Edge bundling for reducing clutter
- Table visualizations
    - Emphasis of biclusters within tables
- Parallel coordinates
- Node-link diagrams for displaying chained biclusters
- Set-based visualizations

3. Explain 4 applications for scatterplots?

- Represent distributions of two continuous variables
- serve to study how *y* is influenced by *x*
- 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, …)
    - Visualize population change

4. Describe at 5 general solutions to solve issues like visual clutter and over plotting.

- Appearance: Change colors, transparency,
- Geometric: Filter, adjust sampling, distortion
    - 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)

5. What is a {concentration ellipse, discrete density plot, hexbin scatterplot, density contour plot, variable binned scatterplot, splatter plot, animated scatter plot}?
    - **Concentration ellipses:** Abstract depiction of (multiclass) distributions
        - Filled (transparent) ellipse with solid outline surrounding data points (of a class).
        - Midpoint may be explicitly shown.
        - Outliers (mild or severe) may be excluded. Often ellipses surround 95% of the data
        - Contours are easily perceived if they are locally linear and globally circular or elliptical
        - 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

    - **Discrete density plot:** Binning based on a grid (quadrilateral, hexagonal) and indicates frequency/density.
        - Bin size as major parameter
        - Density plots
            - Avoid clutter and over plotting in dense regions
            - do not convey distribution in sparsely populated regions well

    - **Hexabin 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

    - **Density contour plot:** Each contour encloses a certain percentage of the data
    - **Variable binned scatterplot:**
        - Data is binned in *adaptive* ranges
        - Density in these bins determined and color-coded
        - **Interaction**: zoom in and out to adjust level of detail
        - **Extension**: binned scatterplot matrices
    - **Splatter plot:** Improving class separability with smoothly overlapping shapes
        - Involves: Kernel density estimation (KDE) and Color blending
        - Smoothing supports interpretation of (overlapping) shapes
    - **Animated scatter plot:** to display dynamic data
6. Describe measures to compare two or more scatterplot-based representations.
    - **Feature Congestion measure** of visual clutter: clutter in a local part of a display is related to the local variability in certain key features
        - Implementation of the Feature Congestion clutter measure involves four stages: (1) compute local feature (co)variance at multiple scales and compute the volume of the local covariance ellipsoid, (2) combine

clutter across scale, (3) combine clutter across feature types, and (4) pool over space to get a single measure of clutter for each input image
    - use color, orientation, and luminance contrast as features
- **Feature Congestion measure:** For a given number of objects in a scene, the scene will appear less cluttered the more “organized” it is
    - Organization may involve grouping similar objects together; aligning them; and making many of the objects a similar hue, luminance, size
    - The degree of organization of a scene can be thought of in terms of the extent to which each part of the scene is predictable from the rest of the scene, or in terms of the amount of redundancy in the scene
    - The Feature Congestion measure of clutter captures this concept of “organization” to some extent implicitly; by looking at feature covariance, it essentially captures some measure of the grouping by similarity + proximity in the display.