Visual Analytics for Big Data – Sample Examination Questions

This course is evaluated by means of a multiple-choice exam. The 10 questions below cover the material mentioned during the course. The study material is given by the slides discussed during the lectures.

The questions are grouped with respect to the theoretical visualization aspects they address. Each question has only **one** correct answer, given the context of the question. To find it, study the formulation of the question carefully. For each question, mark the right answer, by filling in the rectangle next to that answer.

During the exam, students are allowed to browse the course textbook "Data Visualization – Principles and Practice" (1st or 2nd edition). Any other written or electronic material besides this book is not allowed.

A. Aims and scope of visualization

1. Within data visualization, we distinguish several subfields, such as scientific visualization (SciVis) and information visualization (InfoVis). Concerning InfoVis, which is the correct characterization?

- **a.** □ InfoVis focuses on the creation of visualizations of multidimensional data.
- **b.** □ InfoVis focuses on the creation of interactive visualizations for data exploration.
- **c.** □ InfoVis focuses on the creation of visualizations for non-spatial abstract data.
- **d.** □ InfoVis aims at visualizing datasets which are typically larger than SciVis ones ('big data').

2. Resampling is a critical problem for InfoVis applications, as these may involve attributes which are not of the quantitative or integral types. Consider the case where we need to aggregate N values v1….vN of a categorical attribute, which is encoded (stored) as an integer value. The result of the aggregation should be a categorical value in the same domain as the values v1…vN. Which is the best strategy to use to compute the value v of the aggregated attribute?

- **a.** □ Use averaging, i.e. define v as the average of the values v1…vN.
- **b.** □ Define v to the median of the set of values v1…vN.
- **c.** □ Set v to the histogram of the values v1…vN.
- **d.** □ Set v to the categorical value whose frequency, in the above histogram, is largest, if this frequency is above a given threshold. If the frequency is below that threshold, set v to soem default 'unknown' value.

B. Visual encoding

3. Colors are typically represented, in visualization, in the RGB and HSV systems. The two systems are dual, in the sense that a RGB color can be represented in the HSV space and conversely. Given this, which of the following statements is true?

   a. □ The hue of black (in the HSV system) is undefined.
   b. □ The mapping between RGB and HSV can be described by functions that are linear in the R, G, and B coordinates.
   c. □ Not all RGB colors have an HSV equivalent.
   d. □ Not all HSV colors have an RGB equivalent.

4. Different colormaps exist for different attribute types, such as categorical, ordinal, integral, and quantitative. Which of the following statements is true?
   a. □ Grayscale and heat colormaps are good options for all attribute types.
   b. □ No categorical colormap can use a single hue design.
   c. □ Some quantitative color maps do not make sense for integral data.
   d. □ No ordinal colormap can use a multiple hue design.

5. Different visual variables can interfere with each other when used together in a visual encoding scenario. For instance, color and transparency interfere when both used for drawing a scatterplot. Consider such a scatterplot, where we have three quantitative attributes, which we would like to map to the size, color, and transparency of the scatterplot points. Which is the best scenario in terms of diminishing interference?
   a. □ Map each attribute to size, transparency, and color respectively, but use a single hue colormap.
   b. □ Map two attributes to size and color, using any desired colormap; refrain from mapping anything to transparency.
   c. □ As (b), but draw the points from the largest first to the smallest last.
   d. □ As (c), but use a single hue colormap.

C. Visualization techniques

6. Pie charts are an useful tool for displaying a set of values which sum up to one (a so-called partition of unity). However, they have several challenges. Which of the following describes a relevant such challenge for pie charts?
   a. □ A pie chart is visually less scalable than a bar chart; that is, given N data values, we can perceive less well these values if drawn via a pie chart than via a bar chart (both charts have the same screen-space dimension).
   b. □ A pie chart is less effective than a bar chart for displaying a set of ordered data values.
   c. □ Pie charts cannot represent categorical data values.
   d. □ Comparing two (or more) pie charts to see trends in data values is harder than comparing two or more bar charts (for the same set of data values).

7. Table data and tree data can be thought as dual representations. That is, we can encode an attributed tree as a table, and given a table, we can construct a tree representing its data values. Which of the following is true regarding the conversion process of tables to trees?

a. □ A tree can be converted to a table automatically (that is, without any user decision). However, converting a table to a tree requires some user decisions during the conversion process.

b. □ The tree resulting from the conversion of a table depends on the order of the data columns in the table.

c. □ We can convert a table to a tree only if the table contains just categorical values.

d. □ The conversion of a tree to a table is unique. That is, given a tree, there is a single table (having the same order of columns and rows) generated by this process.

8. Parallel coordinate plots (PCPs) are one of the most important tools for visualizing multidimensional data. However, despite their power, they have several limitations. Which of the following limitations is true?

a. □ PCPs cannot show the distribution of values along a specific coordinate axis (dimension).

b. □ PCPs do not allow one to study the precise values of a specific observation (data sample) along all its dimensions.

c. □ The result displayed by a PCP depends on the order in which we draw the observations.

d. □ PCPs require user interaction if we want to clearly compare any pair of dimensions.

9. Projections are another key tool for visualizing multidimensional data. In particular, they are used to detect the presence of groups (clusters) of similar observations in a given dataset. For this task, which of the following is true?

a. □ If we see a clear segregation of points in a 2D projection, then a clear segregation of data values must exist in the original high-dimensional space.

b. □ If points are clearly segregated in the original high-dimensional space, we will see a clear segregation of their projections in a 2D projection.

c. □ Neither (a) nor (b) are true.

d. □ Both (a) and (b) are true.

10. Projections are subject to various errors. Understanding such errors is important for being able to utilize a projection to make inferences about the structure of the high-dimensional data. Consider a projection of such a dataset which delivers an undiferentiated 'blob' of 2D points. If we encounter such a situation, which is the best next step to follow in our visual analysis?

a. □ An undiferentiated blob tells us that the high-dimensional data is not strongly segregated into clusters. We don't need any subsequent analysis to confirm this.

b. □ We should fine-tune the parameters of the projection algorithm, or alternatively change the projection algorithm, until we obtain a clear segregation of the projected points into groups. Only then can we next reason about the structure of the data.

c. □ We should use fewer features (dimensions) to project the data so as to obtain a clearer segregation into clusters.

d. □ We should check the presence of false neighbors and missing neighbors in the projection. If these are numerous, we cannot trust the projection and should perform (b). Else, we can trust the projection and we can follow the outcome of (a).