Exam

Business Intelligence and Process Modelling

Universiteit Leiden — Informatica & Economie

Friday June 2, 2017, 14:00–17:00

This exam consists of **20 questions** divided over four sections. Your answer can be in **Dutch or English**. Always give a precise, to-the-point and well-motivated answer. Write down any non-trivial assumptions. The number of points awarded for each perfectly answered question is listed in front of the question, and sums to **100 points**. Your grade is computed by dividing the number of points by 10. Good luck!

(12p) Visual Analytics

1. (3p) Name three business benefits of using data visualization techniques.

2a. (3p) Explain Simpson's Paradox.

2b. (2p) Hoes does visual analytics help counter Simpson's Paradox?

3. (4p) Describe at least two things that are wrong with the visualization shown in Figure 1.

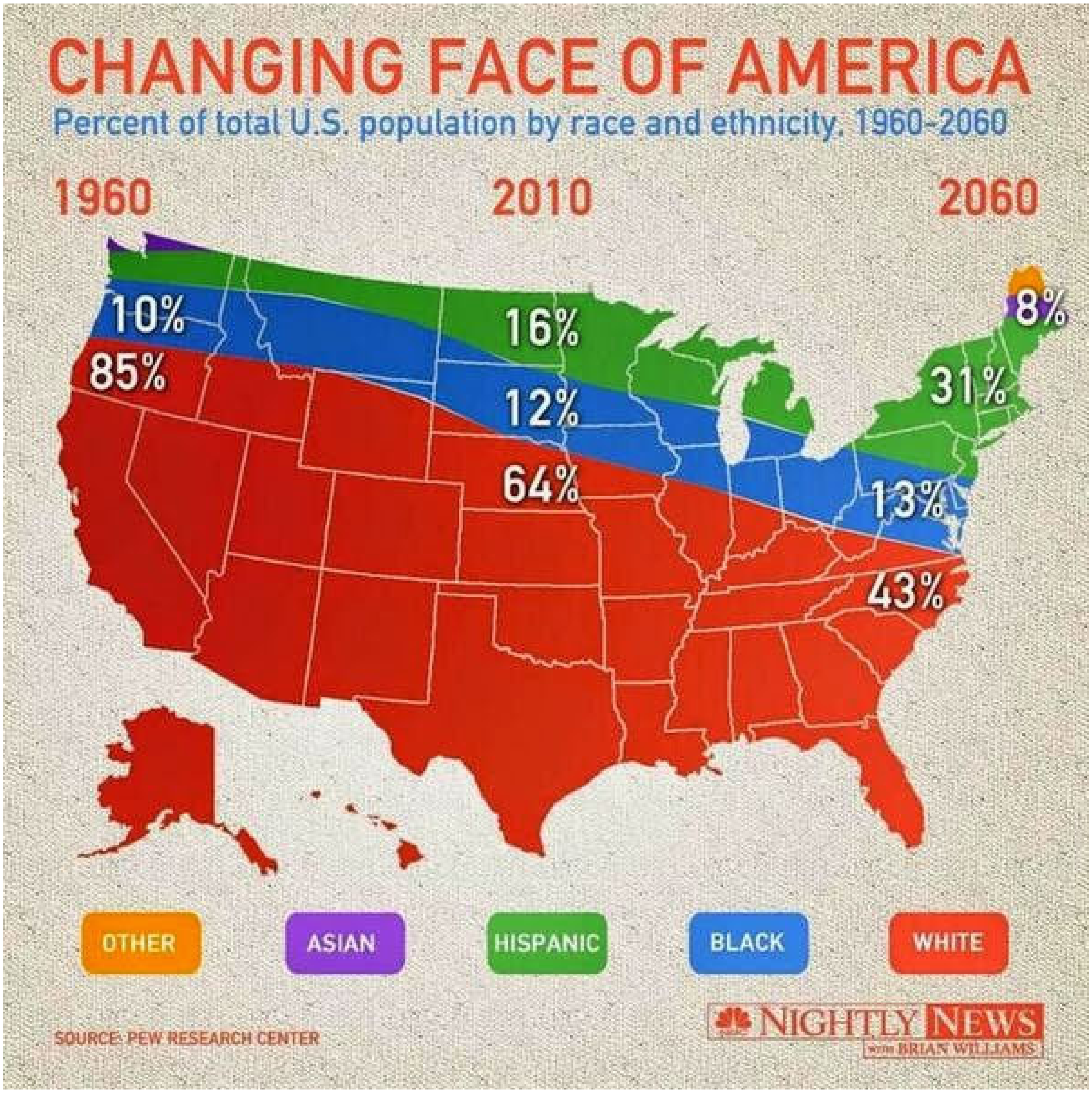

Figure 1: Data visualization.

(30p) Business Intelligence

4. (3p) Give three reasons why it is better to analyze data from a data warehouse than to directly use an organization's transactional systems.

5. (5p) Name five characteristics of a good KPI, and give an example of each in the context of a large online electronics shop.

6. (3p) What is the difference between descriptive, predictive and prescriptive analytics?

7. (3p) What is wrong with blindly using regression to understand the relationship between two variables?

8. (6p) Explain how outlier detection can be done in a supervised, unsupervised, and semi-supervised context. Give examples.

9. (6p) Consider clustering rows in a tabular dataset. Name three criteria to judge whether hierarchical clustering or k-means clustering should be used.

10. (4p) Explain how the principle of Occam's Razor relates to Principal Component Analysis (PCA).

(12p) Network Analytics

Case: Dating networks
Dating apps in which people are rated by their photos have recently increased in popularity. The general idea behind these apps is that the user is repeatedly presented with a photo of a geographically nearby person. The user must then express whether he or she is interested in the person in the photo. If the expression of interest is mutual, users have a match, can start a chat and possibly plan a date.

We consider three years of activity on such a platform, and for simplicity we restrict ourselves to heterosexual activity. We only look at the expression of interest (in some apps associated with "swiping right", or pressing "Like"). This activity can be modeled by a *network*, in which a person/user is a *node*, and the expression of interest of user A in user B is represented by a *directed link* from A to B. Recent research has shown that the resulting network has in fact a giant component, low average pairwise distances (around six) and a power law degree distribution.

11a. (2p) Some (perhaps good-looking) users receive a lot of expressions of interest in dating apps. High values for which node-based network metric characterize this type of user?

11b. (2p) Other users do not have very strict partner preferences and employ an "always right" swiping principle in their photo judging activity. High values for which node-based network metric characterize this type of user?

12. (4p) One distinct property of many real-world networks is a high clustering coefficient. Is this likely also the case in this particular dating network?

13. (4p) Now assume that we look only at the links between people who actually went on at least one date. We consider the spread of a rumor in this network. For simplicity, assume that the rumor always spreads whenever two people meet. In theory, a rumor starting at a random user would on average reach all other users in x steps, where x is the average distance between two users in the network. Explain why it may still take a long time or even never happen that all users have heard the rumor.

(46p) Process Modelling

14. (4p) How would you logically define the research area "Business Process Intelligence", and how does it relate to the topics of "Business Intelligence", "Process Modelling" and "Process Mining"?

15. (6p) Name three key differences in terms of expressiveness between process models drawn using Petri Nets and using Business Process Modelling Notation (BPMN).

16. (4p) Exactly which changes needs to be made to the Petri net in Figure 2 such that for each time that activity 'h' fires, activity 'g' has fired at least three times in the past?

17. (8p) From a bridge, frogs jump into a stream of water, nondeterministically choose one of the two nearby beaches to swim to, and then hop to the bridge to start over again. In search for her prince, a girl picks up every third frog from the stream, kisses the frog, and puts the frog back on the bridge. Model this fairy tale as a Petri net. Suppose that three frogs are initially on the bridge.

18a. (4p) Name and explain four ways of judging the quality of a discovered process model.

18b. (4p) Use these criteria to judge the Petri net in Figure 2.

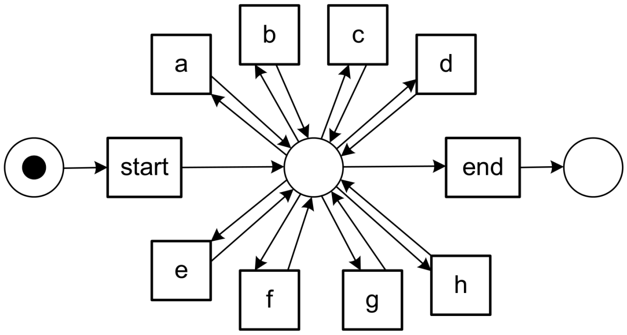

Figure 2: A Petri net of a real-world (non-random) business process.

19. (4p) How can k-fold cross-validation be used to prevent a process discovery algorithm from overfitting a process model?

20. (12p) Apply the α-algorithm to derive the footprint and final corresponding Petri net of event log L.

$$L = [\langle a, d, e, c, b, d, f\rangle, \langle a, d, b, e, c, d, f\rangle, \langle a, b, d, e, c, d, f\rangle \langle a, d, e, b, c, d, f\rangle, \langle a, d, b, f\rangle,]$$

Explain your steps in detail.

End of exam. Please do not forget to fill in the evaluation form!