Exam
Business Intelligence and Process Modelling

Universiteit Leiden — Informatica & Economie

Friday June 10, 2016, 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. (2p) Explain the term "codeless reporting" in the context of Business Intelligence.
2. (6p) When data is visualized, a mapping from a total of x data properties to y visual attributes is made (with finite $x, y > 0$). Assume that all x attributes are meaningful and relevant. Now consider three cases: $x < y$, $x = y$ and $x > y$. What can you say about the quality of the visualization in each of these cases?
3. (4p) Explain at least two things that are wrong with the visualization shown in Figure 1.

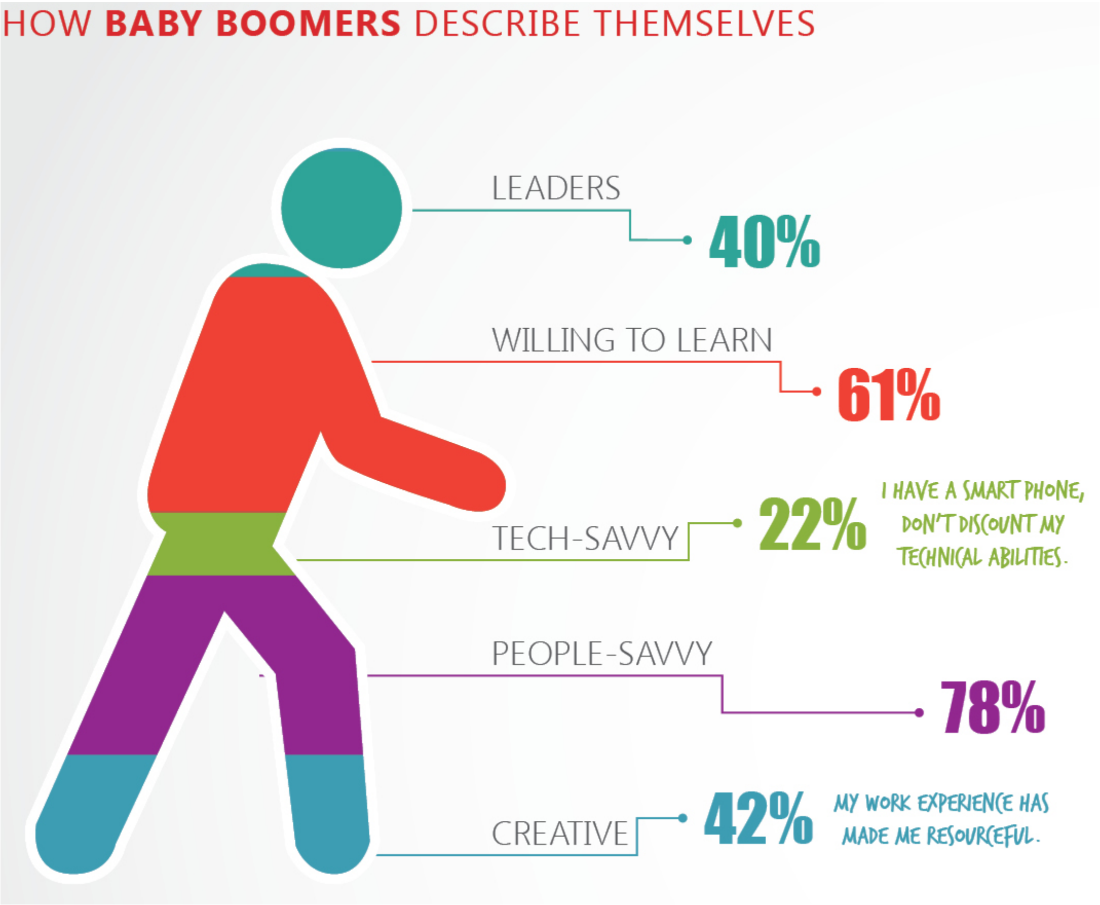

Figure 1: How baby boomers describe themselves.

(38p) Business Intelligence

Theory

4. (3p) Name at least three differences between data in a transactional system and data that is stored in a data warehouse.

5. (3p) In the context of data quality and data mining, researchers often speak of "garbage in, garbage out" Explain this motto.

6. (4p) Explain underfitting and overfitting and relate your explanation to decision trees of depth d that model a dataset with n binary attributes, using d and n in your answer.

7. (6p) Explain how outlier detection can be done in a supervised, unsupervised, and semi-supervised context. Use examples.

8. (4p) Rather than doing everything in-house, a company could use an existing third party service (for example, using an API in some service-oriented architecture) to analyze the company's business data. Discuss one advantage and one disadvantage of this approach from the perspective of the company.

9. (5p) You followed a course on Business Intelligence and Process Modelling. How would you logically define the research area "Business Process Intelligence", which aims to integrate the two topics? How does it differ from traditional Business Intelligence and traditional Process Modelling?

Case: Student performance

Grading exams is a laborous and sometimes frustrating process for university professors. Their life would be much easier if it were possible to already predict the grade of a student in advance, based on the behavior of the student throughout the semester. For a period of 10 years, for a particular course taken by a large enough number of students, a professor has gathered data on every student's behavior, extracting a total of 20 features describing each student. These features include for example student's attendance, assignment grades and participation in discussions, but of course none of the features is the actual exam grade. The features will be used to train a supervised learning algorithm that predicts the exam grade of the students.

10. (5p) When analyzing the correlation matrix of the 20 features, the professor finds that the correlations between the different features are all between -0.3 and +0.3. What does this tell him about the 20 derived features? Does it say anything about the usefulness of the features for accurately predicting the grades of the students?

11. (4p) Explain how in this particular case the professor can use the features and the actual exam grades to train a supervised learning algorithm, in particular without overfitting the model.

12. (4p) A neural network turns out to perform really well on this data. One of the professor's student assistants argues that in light of the concept "Minimal Description Length", it may be better to use a perceptron rather than a full neural network, because it is then much easier to exactly pinpoint which student behavior feature results in a higher grade. Explain at least two things that are wrong with this line of reasoning.

N.B.: You may assume that this exam will be graded based solely on your answers to the exam questions.

(15p) Ownership Networks

Larger companies are typically not individualistic market actors, but also participate in other companies, for example by owning shares in these other companies. This behavior can be modelled using a *directed network*, where a *node* represents a company/firm and a *link* from firm A to firm B means that A owns a substantial percentage (say, more than 5%) of firm B. We will call this type of network an *ownership network*. For simplicity, we do not consider weights on the links.

13. (6p) Say that we want to compare the ownership network network of firms in Italy ($23,000$ nodes) and the network of firms in Germany ($81,000$ nodes). We are interested in comparing the structure of these two networks. Other than simply comparing the number of nodes, name and explain two interesting network properties that can be used to further compare the structure of these two networks.
14. (4p) To find the important firms in this network, we can use centrality measures. Name two of these measures and explain what they measure in the particular case of the ownership network. Take the directed aspect into account.
15. (5p) Assume that we can observe the ownership network at time τ and then again at some time $\tau + 1$, one year in the future. What kind of information does an analysis of the changes in the structure of these two networks at different points in time give us, and what types of company activities could this possibly reveal?

(35p) Process Modelling

16. (3p) Explain how process modelling can serve a descriptive, prescriptive and explanatory role.
17. (7p) In an appartment building with 5 floors, an elevator moves up and down an elevator shaft to bring people to the right floor. Draw a Petri net modelling the behavior of this elevator. Hint: you can model this system using 5 places and 8 transitions.
18. (7p) Draw a Petri net from the model in Business Process Modelling Notation (BPMN) in Figure 2.

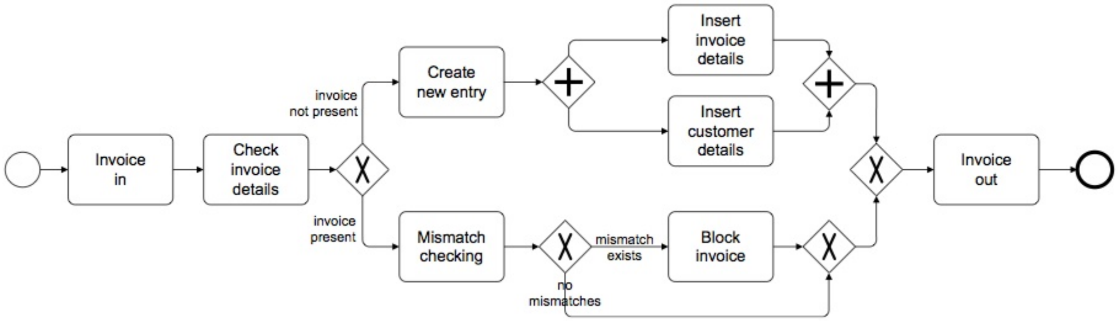

Figure 2: A model in BPMN for a business process model.

19. (4p) Name and explain four ways of judging the quality of a discovered process model.

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