Exam
Business Intelligence and Process Modelling

Universiteit Leiden — Informatica & Economie

Wednesday May 20, 2015, 10:00–13:00

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!

(10p) Visual Analytics

1. (4p) When data is visualized, a mapping from x data properties to y visual attributes is made (with finite $x, y > 0$). In the context of this particular definition, give two possible objective ways to judge the quality of a data visualization (other than aesthetics). Use x and y in your answer.
2. (3p) What is misleading about the visualization shown in Figure 1?
3. (3p) Give your professional judgement of the quality of the visualization shown in Figure 2.

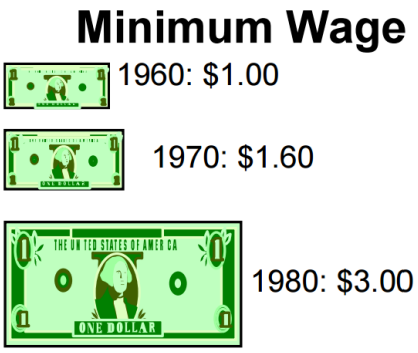

Figure 1: Minimum wage in the United States visualized over time.

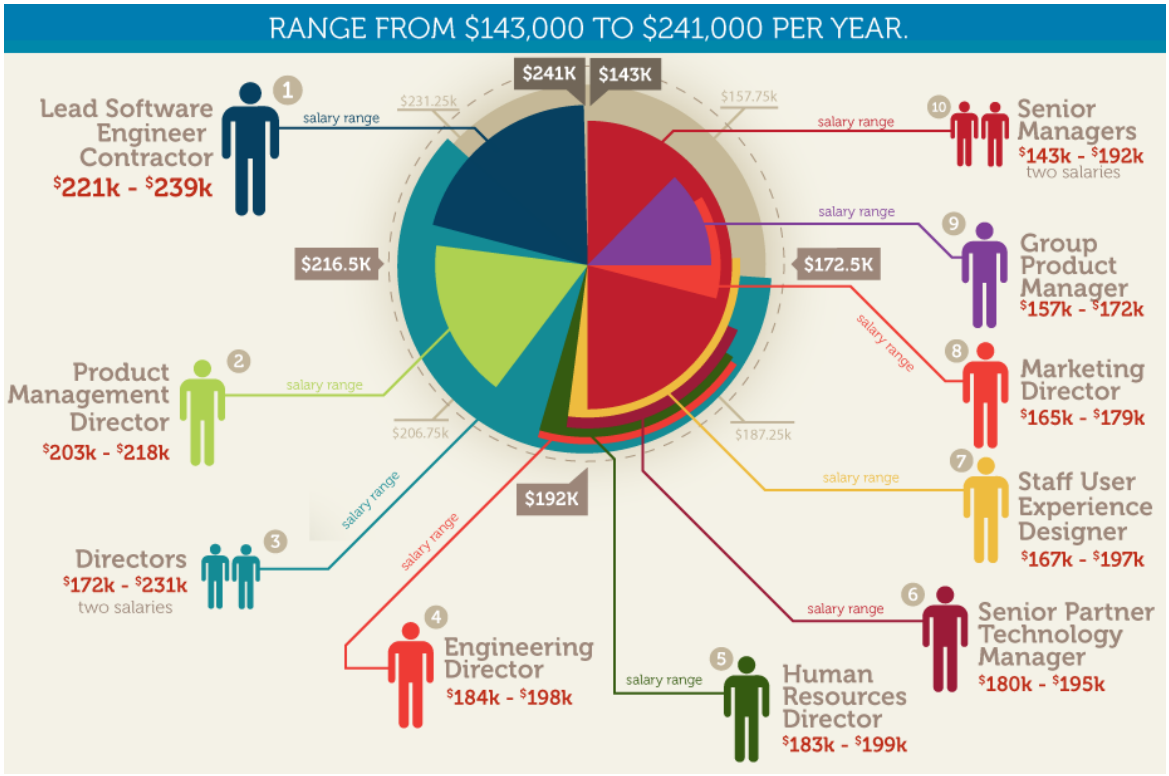

Figure 2: Top-10 salaries at a large tech company.

(35p) Business Intelligence

Theory

4. (3p) Give three typical differences between data in a transactional system and data stored in a data warehouse.

5. (4p) How does the field of business intelligence benefit from using data mining techniques instead of simple OLAP queries?

6. (3p) Explain how outlier detection can be done in a supervised, unsupervised, and semi-supervised context. Use examples.

Case: Customer Loyalty
Consider a project in which a dataset of customers of a mobile phone provider has to be analyzed in order to predict customer loyalty. A student has defined 10 different numeric features that he thinks are relevant to predict customer loyalty. The features will be used to train an algorithm to make a decision tree to predict an 11th attribute describing the loyalty of the customer.

7. (5p) When analyzing the *correlation matrix* of the 10 features, the student finds that the correlations between the different features are all between -0.25 and +0.25.

    a. What does this tell us about the the 10 derived features?

    b. Do these values tell us anything about the usefulness of the features to describe customer loyalty?

8. (5p) Now, assume we can convert all 10 features to ordinal variables with at most 8 possible values.

    a. What is the *curse of dimensionality*, and how does it apply to the customer loyalty case described above?

    b. Roughly how many instances/samples do we need so that we can expect that the curse of dimensionality does not play a significant role?

Case: Support Department
One of the main tasks of the ICT Shared Service Center (ISSC) of a university is to adequately handle ICT support requests in a timely manner. Support requests are initially received by a "first line" team through either an e-mail or a telephone call. This team then either addresses the support call directly or assigns it to a "second line" employee who will then address the problem at a later point in time.

9. (6p) Define three SMART Key Performance Indicators (KPIs) for the ISSC given the task description above.

10. (4p) The ISSC has all data on handling their support requests available, but for some reason cannot compute the value of its own KPI's. Explain how a third party can do this using a service-oriented architecture such as Algorithmia.

11. (5p) For privacy reasons, the ISSC is not particularly happy about giving out its customer and employee data to a third party. Explain how data can be exchanged between two parties without giving away the identity of individuals represented in the data.

(20p) Financial Systems and Network Science

During the guest lectures we have looked at financial markets in a business process modelling context. We will now consider financial systems in a network science (social network analysis) context. In particular, we consider a directed network as a model for the banking system. In this network, a node represents a bank and an link from bank A to bank B means that B borrowed money from A (so A gave out a loan to B). We will simply call this network a *banking network*.

12. (3p) Why is the banking system modeled as a directed network and not as an undirected network?
13. (6p) Assume we want to compare the local banking network of Germany (230 banks) and the local banking network of the United States (6800 banks). So, we only consider loans within one country, and ignore international links between banks for now. This means that the banking networks of Germany and the United States are two separate networks. We are interested in comparing the structure of these two networks. Other than simply comparing the number of nodes (230 vs. 6800 banks), name three interesting *network properties* from the field of network science that can be used to further compare the structure of these two networks.
14. (4p) Give two examples of *centrality measures*, and briefly explain both.
15. (4p) What does it mean in the context of financial systems when a bank is *central* (has a high centrality value) in the global (world-wide) banking network?
16. (3p) Explain how the model of the banking system could possibly be extended from a directed network to a directed *weighted* network.

(35p) Process Modelling

16. (6p) Give three reasons why automated process mining is preferred over manual process modelling.
17. (7p) Draw a model in Business Process Modelling Notation (BPMN) from the Petri net in Figure 3.

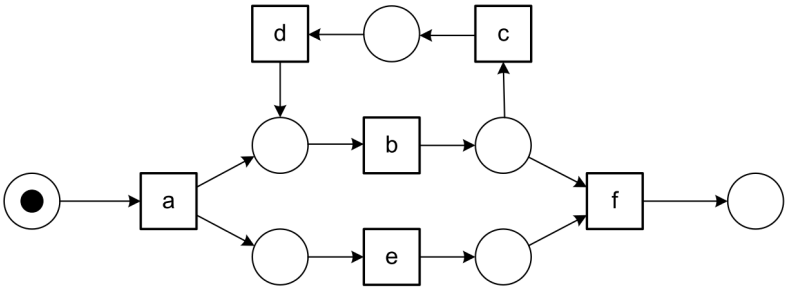

Figure 3: A Petri net for a business process model.

18. (5p) (a) How does a WorkFlow net (WF-net) differ from a Petri net? (b) Why is distinguishing between WF-nets and Petri nets important?
19. (12p) Apply the α-algorithm to derive the footprint and final corresponding Petri net of event log L.

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

Explain your steps in detail.

20. (5p) 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!