**Exam — Answers — Business Intelligence and Process Modelling**
Universiteit Leiden — Informatica & Economie, Friday June 8, 2018, 14:00–17:00

Grades were compensated for the fact that the exam was longer than usual (110 instead of 100 points).

1. (3p) Codeless reporting refers to making data and insights from this data accessible such that one does not have to do any programming (or SQL-querying) to view and understand this data. An example is visualization through a dashboard.

2. (4p) Anscombe's Quartet refers to four sets of two-dimensional numerical data that according to simple statistics such as the mean, standard deviation and correlation coefficient appear identical, but follow very different distributions. This difference is immediately visible when the data is visualized.

3. (6p) This question was 2018-specific, based on a guest lecture. Two answers are possible. 1) Data-driven, alert-driven and knowledge-driven. 2) Deviating average times, deviating start and end activities and prohibited subsequent actions.

4. (4p) A gradient from blue to white was applied to the colors of the pies, but this color is not related to the data. Countries are sorted alphabetically, rather than based on for example the percentage value.

4. (5p) SMART, see lecture slides. Examples of all five aspects in one example is OK, if made explicit.

5. (2p) Completeness (all data should be present, or it should be known which data is missing) and accuracy (the data should be correct, or the error must be quantifiable).

6. (3p) A data warehouse is centered around subjects instead of applications and as such more focused on information acquisition (1p). It typically also contains (more) historic data to enable analysis over time (1p) and as opposed to a transaction system is nonvolatile, essential for consistent analysis (1p).

7. (2p) The knowledge gap refers to the gap between an organization's available data and insights (knowledge) from that data (data left unanalyzed). The execution gap refers to the difference between knowledge obtained from data and knowledge actually employed and used ("wisdom") in decision making or day-to-day operations.

8. (2p) Hierarchical clustering does not require one to set a parameter for $k$ in advance, allowing multiple clusterings of different sizes to be dervied. It may reveal a clear hierarchy if the data has one. It is less prone to outliers than for example $k$-means clustering.

9. (6p) In supervised outlier detection, we distinguish between items labeled as 'normal' or 'abnormal' using some classifier. Unsupervised outlier detection deals with finding instances that are least similar to all other instances, e.g., instances that fall into their own cluster as a result of applying a clustering algorithm. In semi-supervised outlier detection, first normal instances are characterized using a training data set, and then unlabeled examples are judged according to whether or not they fit the trained model.

10. (4p) Problems that are not linearly separable. An example is the exclusive OR (XOR) problem.

11. (6p) The learning rate $\alpha$, number of epochs, number of layers and the number of nodes per layer.

11. a) (2p) A node's indegree refers to the number of nodes that have a directed link point at that node, whereas a node's outdegree is the number of outgoing links from that node.
    b) (2p) For a given page, indegree refers to the number of pages that point to that page, whereas outdegree indicates how many links a present on a page.

12. (6p) The giant component refers to the fact that there is one large connected component (a subset of the network's nodes that are all connected through paths of edges). The power law degree distribution indicates that there are many nodes with a low degree, and a few nodes with a very high degree. The small-world phenomenon indicates that the average distance between two pages is very low.

13. (3p) It means that the pages on India and China are apparently well-connected locally, but not at a central position in the entire network (so globally). It may indicate that the English version of Wikipedia is perhaps more oriented towards western countries; although locally central, India and China may reside in their own "community". Many other informed answers also OK.

14. (6p) Petri nets have formal verification. Petri nets can better model concurrent processes. BPMN has more primitive building blocks, so that complicated processes can be written in a more compact way. Subprocesses can more easily be modelled in BPMN. BPMN is more readable for non-experts.

15. (8p) See Figure 1. Note that many correct answers are possible. Remember the initial marking.

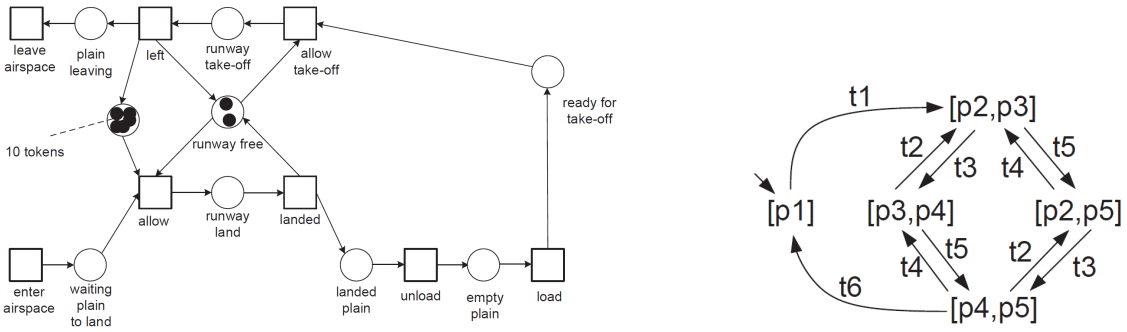


Figure 1: Answers for question 15 (left) and 18 (right).

16. (6p) Discovery refers to automatically deriving a process model from event data. Conformance refers to checking whether the model matches with the reality or a specification. Enhancement refers to how insight from process models can be used to in general improve the performance of the organization.

17. (4p) 1) Fitness (the discovered model should allow for the behavior seen in the event log), 2) Precision (the discovered model should not allow for behavior completely unrelated to what was seen in the event log, 3) Generalization (the discovered model should generalize the example behavior seen in the event log) and 4) Simplicity (the discovered model should be as simple as possible).

18. (7p) See Figure 1. Do not forget to indicate the starting marking using an incoming arrow.

19. a) (4p) A Petri net is sound when it is safe, has proper completion, the option to complete and absence of dead parts.
    b) (3p) A WorkFlow-net (WF-net) is a Petri net with a fixed input place (source), fixed output place (sink) and adherence to the soundness criterion. A WF-net is the type of net that fits the BPM life-cycle of cases and is the output of the $\alpha$-algorithm.

20. ```
    #  -> -> #  #  #
    <- #  || -> -> <-
    <- || #  -> -> <-
    #  <- <- #  #  #
    #  <- <- #  #  ->
    #  -> -> #  <-  #
    ```
    $T_L = \{a, b, c, d, e, f\}$, $T_I = \{a\}$, $T_O = \{d\}$
    $X_L = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{b\}, \{d\}), (\{b\}, \{e\}), (\{b\}, \{d, e\}), (\{c\}, \{d\}), (\{c\}, \{e\}), (\{c\}, \{d, e\}), (\{e\}, \{f\}),$
    $(\{f\}, \{b\}), (\{f\}, \{c\}), (\{a, f\}, \{b\}), (\{a, f\}, \{c\})\}$
    $Y_L = \{(\{b\}, \{d, e\}), (\{c\}, \{d, e\}), (\{e\}, \{f\}), (\{a, f\}, \{b\}), (\{a, f\}, \{c\})\}$
    $P_L$ is exactly the set of places between the transitions in $Y_L$ plus a start place to $a$ and end place from $d$. Then deriving $F_L$ is somewhat trivial. Finally, $\alpha(L) = (P_L, T_L, F_L)$. Petri should be drawn with $|Y_L| + 2 = 7$ places and $|T_L| = 6$ activities and a correct initial marking.
    ```