Exam — Answers — Data Science and Process Modelling
Universiteit Leiden — Informatica & Economie, Wednesday May 29, 2019, 14:00–17:00

1. Proximity, the fact that objects are visualized close together, connectedness, the fact that objects are connected in the visualization, and similarity, the fact that objects look alike in a visualization, should all reflect that the underlying data objects have data attribute values in common and are thus in fact similar or connected in some way.

2. Simpson's Paradox relates to the case when there appears to be a particular (e.g., positive) relation between variables in a dataset, where actually the data contains two or more groups for which an opposite relation (e.g., negative) is observed.

3. In knowledge-driven detection of anomalies, the detection of anomalies by an algorithm or tool is supported by human knowledge to classify regular and anomalous behavior.

4. Too many data attributes have been mapped to visual attributes, cluttering the visualization. The three colors do not each time map to the same type of statistic (author, publication, citation) and it is unclear relative to what other measurement the growth rate is calculated.

5. Veracity; the quality of the data is uncertain, i.e., it can be inaccurate or incomplete (e.g., differently aligned medical measurement devices). Velocity; the data may arrive very quickly and may not be directly analyzable or even storable (e.g., high resolution/framerate 3D medical scan video). Variety; there are different types of data that together result in insights (e.g., using different types of blood measurements and video data to automatically get insights in performance of organs).

6. In data-driven research, data is the starting point, and patterns are discovered (e.g., using visualization or machine learning) that may lead to new hypotheses and with that insights in the domain. In hypothesis-driven research, the starting point is a hypothesis, for which data is collected, which is then used to statistically test the one hypothesis in order to gain insight in that one hypothesis.

7. The Pearson correlation only indicates whether there is a linear relationship between the two variables, while other nonlinear relations may be present. In addition, it only says if there is a correlation between the two variables; not if there is a causal relationship.

8. Hierarchical clustering may reveal a clear hierarchy if the data has one; which is to be expected given the hierarchical "kingdoms" commonly referenced in biology. A disadvantage is that the method is quadratic in the number of species, which may result in a high running time given the data size.

9. The XOR problem is not linearly separable, i.e., the two-dimensional binary classification problem underlying it cannot be solved by drawing a line between two sets of points visualized in the plane based on their binary attribute values.

10. This relates to how the input and output values of the neural network are represented. In local encoding, one node's numeric value represents possible values of the input/output value, whereas in distributed encoding a particular binary value permutation of multiple nodes encodes the input/output value.

11. A node's indegree refers to the number of nodes that have a directed link point at that node, here this is the number of followers of a user. A node's outdegree is the number of outgoing links from that node, here referring to the number of people that the considered user follows.

12. The clustering coefficient measures the connectivity of a node's neighbors, computing the number of triangles in a node's neighborhood relative to the possible number of triangles. Non-celebrities are expected to have a higher value, as their friends are more likely to know each other from real-world interaction, meaning they would have more clustered groups of friends.

13. PageRank, measuring the importance of users based on the number of other important users that follow this user and betweenness centrality, assessing the the extent to which the user connects different groups of users in a broker type of role. Multiple answers possible.

14. Concurrency refers to the fact that certain activities in a process take place at the same time, i.e., not sequentially. Mutual exclusion refers to the fact that certain activities in a process model may never execute certain critical steps concurrently, for example because they both need the same resource.

15. Many correct answers are possible. Remember the initial marking.

16. Encapsulation is realized in BPMN through an explicit construct representing the encapsulated part of the process, whereas in Petri nets, formally the entire encapsulated process should be displayed, or an informal note indicating that the subprocess takes place within one particular activity, must be made. Overall one could say that BPMN has better native support for this concept.

17. Assume that initially $[start]$ is marked. The reachability graph should be drawn with 11 states: $[start], [c_1, c_2], [c_2, c_3], [c_2, c_4], [c_1, c_5], [c_1, c_6], [c_3, c_5], [c_3, c_6], [c_4, c_5], [c_4, c_6]$ and $[end]$. Deriving the edges of the reachability graph is somewhat trivial. Do not forget the cycle from $[c_2, c_4]$ to $[c_4, c_5]$ and a starting arrow pointing at $[start]$.

18. 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).

19. If the event data is incomplete, then certain important parts of the process may not be represented in the process model. If there is noise (incorrect data) in the event data, then certain undesired behavior may end up in the process model.

20.
```
  a  b  c  d  e  f  g
a #  -> -> -> #  #  #
b <- #  #  || #  #  ->
c <- #  #  || -> -> #
d <- || || #  || || ->
e #  #  <- || ## || ->
f #  #  <- || || #  ->
g #  <- #  <- <- <- #
```

$T_L = \{a, b, c, d, e, f, g\}$, $T_I = \{a\}$, $T_O = \{d\}$
$X_L = \{(\{a\}, \{b\}), (\{a\}, \{c\}), (\{a\}, \{d\}), (\{a\}, \{b, c\}), (\{b\}, \{g\}), (\{c\}, \{e\}), (\{c\}, \{f\}), (\{d\}, \{g\}), (\{e\}, \{g\}), (\{f\}, \{g\}), (\{b, e\}, \{g\}), (\{b, f\}, \{g\})\}$
$Y_L = \{(\{a\}, \{d\}), (\{a\}, \{b, c\}), (\{c\}, \{e\}), (\{c\}, \{f\}), (\{d\}, \{g\}), (\{b, e\}, \{g\}), (\{b, f\}, \{g\})\}$
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 $|Y_L| + 2 = 9$ places and $|T_L| = 7$ activities and a correct initial marking.