Exam — Answers — Business Intelligence and Process Modelling
Universiteit Leiden — Informatica & Economie, Wednesday May 20, 2015, 10:00–13:00

1. In a good visualization, $x = y$, meaning that all relevant data attributes are mapped to visual attributes, and no important data attributes are left out. In a good visualization, y is not too large, as that would clutter the visualization with too many visual attributes, making the visualization hard to interpret.

2. A two-dimensional image (dollar bill) is used to visualize a one dimensional value (minimum wage in dollars). The 1980 bill looks over 4 times larger than the 1960 bill, but the difference is only a factor 3.

3. The pie charts are misleading: it is not clear how many of each type of employee the company has employed; this could have been visualized using the radius of the pieces of the pie chart. The meaning of two men instead of one is unclear. A pie chart is typically used to visualize portions of some total. In this case, a bar chart with upper and lower bounding boxes would be better.

4. Transactional system vs. data warehouse: volatile data vs. nonvolatile data, oriented on daily operations vs. oriented on data analytics, data storage focus vs. information acquisition focus (and many more).

5. OLAP queries can only answer *descriptive* questions about (and perhaps simple patterns in) the data, whereas data mining techniques can also find *patterns* that are not directly visible from simple querying (such as associations), but moreover using data mining it is possible to make a *predictive* analysis.

6. In *supervised* outlier detection, the task is to distinguish between items that have been labeled as 'normal' or 'abnormal' using some classification technique. *Unsupervised* outlier detection deals with finding instances that are least similar to all the other instances, for example instances that fall into their own cluster after some clustering algorithm is applied to the data. 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 data.

7a. It tells us that all derived features are somewhat unrelated to each other.

7b. No. It might be that the 11th attribute of customer loyalty correlates with none of the attributes, with a combination of some of the attributes, or with all attributes. However, given the fact the features all seem some what independent but were created with the domain knowledge in mind, it may very well be that they work together quite well in a classification algorithm.

8a. The *curse of dimensionality* is generally defined as the problem that when a learning algorithm is trained on a high dimensional feature space, an enormous amount of training data is required to ensure that there are several samples with each combination of possible values for the attributes. With 10 attributes possibly describing customer loyalty, this may very well play a role.

8b. With $n = 10$ variables and $v = 8$ possible values, the number of instances would have to be in the order of $v^m = 8^{10}$ (little over 1 billion). However, seen the fact that the variables are ordinal, the value of v is probably a lot lower in practice (but of course at least $v \geq 2$).

9. Within 3 weeks from now, 10% of all second line calls should be handled by the first line. All support calls should be handled within 5 days by both support lines. The number of support calls handled by a first line support employee should be at least 8 per day on average, over a period of one month.

10. The relevant data can be made available via an API, which can serve as input for a third party's system which takes the data as input and provides the service of computing the KPI's. These KPI's are then output and can again be accessed via an API by the ISSC, for example for visualization in a dashboard. The task of computing the KPI's and outputting the result can be implemented on Algorithmia

11 Collision free *hashing* (either by the ISSC or via a *trusted third party*) can be used for anonymization, ensuring that data can not be traced back to a user, but that two entries regarding the same user can still be identified as such.

12. Because extending a loan from A to B is a directed relation. In an undirected network, there would be no way to distinguish between the bank giving out the loan and the bank receiving the loan.

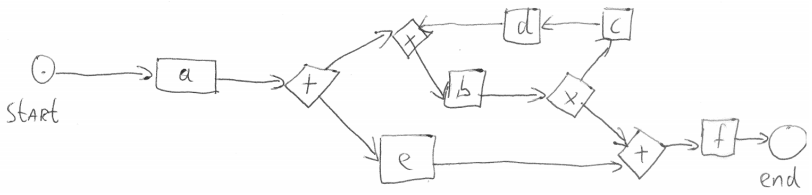
13. Number of edges, density, degree distribution, distance distribution, average clustering coefficient.

14. *Indegree centrality* (the number of nodes pointing to the considered node) and *betweenness centrality* (the relative and normalized number of times the considered node is part of a shortest path).

15. For a local measure such as outdegree centrality, being central means that a bank has outstanding loans to many other banks, indicating a financial dependence of that bank on many other banks, possibly indicating a risk of the banking system if that bank were to go bankrupt.

16.1 A weighted banking network can be constructed by assigning a *weight* to a link which is proportional to the size of the loan.

16.2 Automated process mining reduces the chance of human bias and error, allows for large datasets of event logs to be processed and ensures that modelling is done is a consistent and mathematically verifiable way.

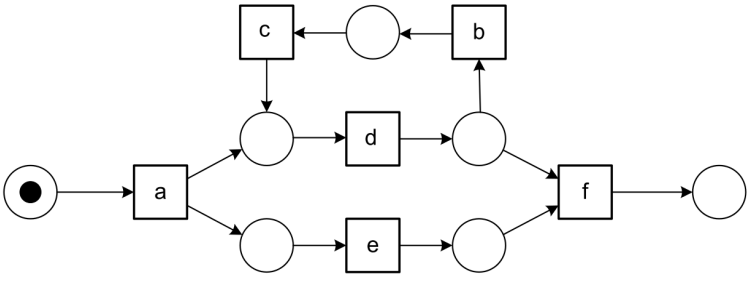


16.

18a. 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 (safeness, proper completion, option to complete and absence of dead parts).

18b. A WF-net is the type of net that fits the BPM life-cycle of cases and is the output of the α-algorithm.

19. $T_L = \{a, b, c, d, e, f\}$
$T_I = \{a\}$
$T_O = \{f\}$
$X_L = \{(\{a\}, \{d\}), (\{a\}, \{e\}), (\{b\}, \{c\}), (\{c\}, \{d\}), (\{d\}, \{b\}), (\{d\}, \{f\}), (\{e\}, \{f\}), (\{a, c\}, \{d\}), (\{d\}, \{b, f\})\}$
$Y_L = \{(\{a\}, \{e\}), (\{b\}, \{c\}), (\{e\}, \{f\}), (\{a, c\}, \{d\}), (\{d\}, \{b, f\})\}$
P_L is exactly the set of places between the transitions in Y_L plus a start and end place. Then deriving F_L is somewhat trivial.
$\alpha(L) = (P_L, T_L, F_L)$



20. *Fitness* (the discovered model should allow for the behavior seen in the event log), *precision* (the discovered model should not allow for behavior completely unrelated to what was seen in the event log, *generalization* (the discovered model should generalize the example behavior seen in the event log) and *simplicity* (the discovered model should be as simple as possible).