Quiz 1. What is data mining?

Which of the following is a data mining task?

1. Search Wikipedia for “Nobel laureates”.

Query for a known attribute value

-> **not data mining**

2. Group news articles by topic.

Finding groups by similarity between articles, the topics are not given, looking for implicit (hidden) relationships between objects

-> **data mining.**

The question is ambiguous, since it is not clear that the topics are not given

3. Confirm that there is indeed a correlation between bread and milk in everyday grocery transactions.

We have a hypothesis that there is a correlation, and we are seeking to confirm our hypothesis using statistical methods – the relationship is not hidden, it is already in our head

-> **not data mining**

4. Find out people opinions about a new i-pod:

We just are sifting through all the opinions, looking into a raw data

-> **not data mining.**

5. Give a 100% correct prognosis of the future success or failure for a new business.

Data mining is not a future teller; it can only predict the probability of an event, since there are too many unknown variables which can influence the outcome of a business

-> **not data mining**

6. Create profile of a drug smuggler based on historical data.

From all the historical records about border crossing extract the model (for example, decision tree) to predict the class label: smuggler/not smuggler. The model is hidden, and so we extract it

-> **data mining**

7. Identify profile of customers who are likely to purchase an extended home insurance.

Extract a classification model from historical data

-> **data mining**

8. Find out which family names are prevalent in different locations.

Neither family names, nor locations are given: we are looking for hidden relationships between names and locations

-> **data mining**

The question can be a data mining task of correlation between family names and areas, but is also can be performed as a ranked query, for each area give me

9. Discover the most important factors of big salaries from a census dataset.

Extract hidden model – the relationship between attribute values and the target class variable – salary

-> **data mining**

10. Look up a phone number for taxicabs.

A query, **not data mining.**