What is the need for data mining?
To extract knowledge to analyze from the large amounts of raw data

What is the need for mining large data sets?
Information is sometimes hidden, and make take human analysts weeks to figure out useful information. Humans also leave much data never analyzed at all.

What is not data mining?
Looking up a phone number in a phone directory.

What are the origins of data mining?
Draws ideas from statistics, artificial intelligence, machine learning, pattern recognition and database systems

What are the two types of data mining tasks?
Prediction Methods
Description Methods

What is Prediction Methods?
Use some variables to predict unknown or future values of other variables.

What are the predictive methods?
Classification
Deviation Detection
Regression

What is Description Methods?
Find human-interpretable patterns that describe the data

What are the descriptive methods?
Clustering
Association Rule Discovery
Sequential Pattern Discovery

What are the six types of data mining tasks?

1. Classification
2. Clustering
3. Deviation Detection
4. Association Rule Discovery
5. Sequential Pattern Discovery
6. Regression

What is the main goal of classification?
Previously unseen records should be assigned a class as accurately as possible.

What is the main goal of deviation detection?
To detect significant deviation from normal behavior.

What is the main goal of clustering?
To find similarities between a set of data points or be able to find less similar data points to separate clusters.

What is the main goal of association rule discovery?
To make recommendations based off of item and user similarity.

What is the main goal of regression?
To predict a value of given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.

What is the main goal of sequential pattern discovery?
When given a set of objects, with each object associated with its own timeline of events, fine rules that predict strong sequential dependencies among different events.

What are the problems of data mining?
Scalability
Dimensionality
Complex and Heterogeneous Data
Data Quality
Data Ownership and Distribution
Privacy Preservation

Data Mining
The principle of sorting through large amounts of data and picking out relevant information.

Data Mining
The nontrivial extraction of implicit, previously unknown, and potentially useful information from data.

Classification, Association Rules and Clustering
Three Data Mining tasks.

Ensemble Method
When multiple techniques are used together.

interesting
The goal of data mining is to discover ___ data patterns hidden in large data sets.

knowledge discovery

Data mining was originally known as ___.
early 90s

When was the term "data mining" coined?
Clients

Data mining tools are often ___ of the data warehouse.
No

Are data mining algorithms build into a data warehouse?

Derived
Mining algorithms require ___ values.
percentages and averages

What types of derived values are often used?

Customer signature
An example data mining structure.

Statistical and Artificial Intelligence
What two camps do the data mining algorithms typically come from.

1960s
Many of the statistical approaches date back to the ___.

early 80s
Artificial intelligence became popular in the ___.

Machine learning
The field in AI that contains Fuzzy Logic, Heuristic reasoning, and Neural Networks.

Machine learning
___ develops algorithms that enable computers to acquire knowledge.

classify, predict
If we can ___ examples with known outcomes then we can use the classifier to ___ unknown outcomes.

directed
Classification is a ___ technique.
supervised learning

Directed learning is also known as ___.
training set

A set of examples with a known outcome.

every outcome
A training set must have examples of ___.

Overfitting
When a set of rules becomes very specific to the training set.

Occam's Razor
Overfitting is a violation of ___.

The law of succinctness
Occam's razor is also known as ___.

Classifiers
Decision trees, neural networks and statistical models are ___.

Accuracy, speed, robustness, scalability and interpretability.
How are classification models evaluated.
Build a model
First step in the decision tree building process.
Use the Model
Second step in the decision tree building process.
Information gain, Gain ratio and Gini index
Three techniques for picking the best split in a Decision Tree.