1.1 (d) Describe the steps involved in data mining when viewed as a process of knowledge discovery.

The steps involved in data mining when viewed as a process of knowledge discovery are as follows:

- **Data cleaning**, a process that removes or transforms noise and inconsistent data
- **Data integration**, where multiple data sources may be combined
- **Data selection**, where data relevant to the analysis task are retrieved from the database
- **Data transformation**, where data are transformed or consolidated into forms appropriate for mining
- **Data mining**, an essential process where intelligent and efficient methods are applied in order to extract patterns
- **Pattern evaluation**, a process that identifies the truly interesting patterns representing knowledge based on some interestingness measures
- **Knowledge presentation**, where visualization and knowledge representation techniques are used to present the mined knowledge to the user

1.12 (a) Present three application examples of spatiotemporal data streams.

i. Sequences of sensor images of a geographical region, for instance, San Francisco, along time.

ii. The climate images transmitted from satellites.

iii. Data that describe the evolution of natural phenomena, such as forest coverage, forest fire, and so on.

(b) Discuss what kind of interesting knowledge can be mined from such data streams, with limited time and resources.

The knowledge that can be mined from spatiotemporal data streams really depends on the application. However, one unique type of knowledge about stream data is the patterns of spatial change with respect to the time. For example, the changing of the traffic status of several highway junctions in a city, from the early morning to rush hours and back to off-peak hours, can show clearly where the traffic comes from and goes to and hence, would help the traffic officers plan effective alternative lanes in order to avoid traffic jam. As another example, a sudden appearance of a point in the spectrum space image may indicate that a new astronomic object is being formed. The changing of humidity, temperature, and pressure in climate data may reveal patterns of how hurricanes are formed.

(c) Identify and discuss the major challenges in spatiotemporal data mining.

One major challenge is how to deal with the continuing large-scale data. Since the data keep flowing in and each snapshot of data is usually huge (e.g., the spectrum image of space), it is often impossible to store all of the data. Some aggregation or compression techniques may have to be applied, and old raw data may have to be dropped. Mining under such aggregated (or lossy) data is challenging. In addition, some patterns may occur with respect to a long time period, but it may not be possible to keep the data for such a long duration. Thus, these patterns may not be uncovered. The spatial data sensed may not be so accurate, so the algorithms must have high tolerance with respect to noise.
2.6. In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

The various methods for handling the problem of missing values in data tuples include:
(a) **Ignoring the tuple**: This is usually done when the class label is missing (assuming the mining task involves classification or description). This method is not very effective unless the tuple contains several attributes with missing values. It is especially poor when the percentage of missing values per attribute varies considerably.
(b) **Manually filling in the missing value**: In general, this approach is time-consuming and may not be a reasonable task for large data sets with many missing values, especially when the value to be filled in is not easily determined.
(c) **Using a global constant to fill in the missing value**: Replace all missing attribute values by the same constant, such as a label like “Unknown,” or -∞. If missing values are replaced by, say, “Unknown,” then the mining program may mistakenly think that they form an interesting concept, since they all have a value in common | that of \Unknown." Hence, although this method is simple, it is not recommended.
(d) **Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values**: For example, suppose that the average income of All Electronics customers is $28,000. Use this value to replace any missing values for income.
(e) **Using the attribute mean for quantitative (numeric) values or attribute mode for categorical (nominal) values, for all samples belonging to the same class as the given tuple**: For example, if classifying customers according to credit risk, replace the missing value with the average income value for customers in the same credit risk category as that of the given tuple.
(f) **Using the most probable value to fill in the missing value**: This may be determined with regression, inference-based tools using Bayesian formalism, or decision tree induction. For example, using the other customer attributes in the data set, we can construct a decision tree to predict the missing values for income.
2.17(a) Consider the data as two-dimensional data points. Given a new data point, \( x = (1.4, 1.6) \) as a query, rank the database points based on similarity with the query using (1) Euclidean distance (Equation 7.5), and (2) cosine similarity (Equation 7.16).

The Euclidean distance of two \( n \)-dimensional vectors, \( x \) and \( y \), is defined as: \( 2 \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \). The cosine similarity of \( x \) and \( y \) is defined as: \( \frac{x^t \cdot y}{||x|| ||y||} \), where \( x^t \) is a transposition of vector \( x \), \( ||x|| \) is the Euclidean norm of vector \( x \), and \( ||y|| \) is the Euclidean norm of vector \( y \). Using these definitions we obtain the distance from each point to the query point.

<table>
<thead>
<tr>
<th></th>
<th>( x_1 )</th>
<th>( x_2 )</th>
<th>( x_3 )</th>
<th>( x_4 )</th>
<th>( x_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean distance</td>
<td>0.14</td>
<td>0.67</td>
<td>0.28</td>
<td>0.22</td>
<td>0.61</td>
</tr>
<tr>
<td>Cosine similarity</td>
<td>0.9999</td>
<td>0.9957</td>
<td>0.9999</td>
<td>0.9990</td>
<td>0.9653</td>
</tr>
</tbody>
</table>

Based on the Euclidean distance, the ranked order is \( x_1, x_4, x_3, x_5, x_2 \). Based on the cosine similarity, the order is \( x_1, x_3, x_4, x_2, x_5 \).