Ten Tasks

1. Retrieve Value

General Description: Given a set of specific cases, find attributes of those cases.

Pro Forma Abstract: What are the values of attributes {X, Y, Z,...} in the data cases {A, B, C, ...}?

Examples:

- What is the mileage per gallon of the Audi TT?

- How long is the movie Gone with the Wind?

2. Filter

General Description: Given some concrete conditions on attribute values, find data cases satisfying those conditions.

(“Filter” doesn’t rely on properties of all other elements in the data set.)

Pro Forma Abstract: Which data cases satisfy conditions {A, B, C...}?

Examples:

- What Kellogg's cereals have high fiber?

- What comedies have won awards?

- Which funds underperformed the SP-500?

3. Compute Derived Value

General Description: Given a set of data cases, compute an aggregate numeric representation of those data cases.

Pro Forma Abstract: What is the value of aggregation function F over a given set S of data cases?

e.g: mode, average, median, variance, standard deviation, count

Examples:

- What is the average calorie content of Post cereals?

- What is the gross income of all stores combined?

- How many manufacturers of cars are there?

4. Find Extremum

General Description: Find data cases possessing an extreme value of an attribute over its range within the data set.

Pro Forma Abstract: What are the top/bottom N data cases with respect to attribute A?

Examples:

- What is the car with the highest MPG?

- What director/film has won the most awards?

- What Robin Williams film has the most recent release date?

5. Sort

General Description: Given a set of data cases, rank them according to some ordinal metric.

Pro Forma Abstract: What is the sorted order of a set S of data cases according to their value of attribute A?

Examples:

- Order the cars by weight.

- Rank the cereals by calories.

6. Determine Range

General Description: Given a set of data cases and an attribute of interest, find the span of values within the set.

Pro Forma Abstract: What is the range of values of attribute A in a set S of data cases?

Examples:

- What is the range of film lengths?

- What is the range of car horsepowers?

- What actresses are in the data set?

7. Characterize Distribution

General Description: Given a set of data cases and a quantitative attribute of interest, characterize the distribution of that attribute’s values over the set.

Pro Forma Abstract: What is the distribution of values of attribute A in a set S of data cases?

Examples:

- What is the distribution of carbohydrates in cereals?

- What is the age distribution of shoppers?

8. Find Anomalies

General Description: Identify any anomalies within a given set of data cases with respect to a given relationship or expectation, e.g. statistical outliers.

Pro Forma Abstract: Which data cases in a set S of data cases have unexpected / exceptional values?

Examples:

- Are there exceptions to the relationship between horsepower and acceleration?

- Are there any outliers in protein?

9. Cluster

General Description: Given a set of data cases, find clusters of similar attribute values.

Pro Forma Abstract: Which data cases in a set S of data cases are similar in value for attributes {X, Y, Z, …}?

Examples:

- Are there groups of cereals similar fat/calories/sugar?

- Is there a cluster of typical film lengths?

10. Correlate

General Description: Given a set of data cases and two attributes, determine useful relationships between the values of those attributes.

Pro Forma Abstract: What is the correlation between attributes X and Y over a given set S of data cases?

Examples:

- Is there a correlation between carbohydrates and fat?

- Is there a correlation between country of origin and MPG?

- Do different genders have a preferred payment method?

- Is there a trend of increasing film length over the years?