Using Data Mining techniques, finding correlations or patterns among large sets of data can be used to predict future behavior.

The ultimate goal of Data Mining is prediction.

The most common type of data mining is the classification of data. This is the process of determining a predefined class, to which an example belongs to.

Data mining is the process of discovering patterns and regularities in large datasets, presenting them in an intelligible form, and using them to make predictions.

Why Data Mining? Cont.

Commercial Applications:
1. Recruiting/Attracting customers.
3. Builds profiles of customers likely to use which services.
4. Web Mining.

Scientific Applications:
1. Network Intrusion Detection.
3. Medicine
4. Identifying Spam
5. Detecting Advertisements on the Web

Classification Tool C5.0/See5
- C5.0 is designed to run on Unix/Linux.
- See5 is designed to run on Windows.
- C5.0/See5 was developed by Ross Quinlan as the successor to ID3 and C4.5 systems.
- It is a tool for discovering patterns and regularities in the database, presenting them in an intelligible form, and using them to make predictions.

Output generated by Classification Tools include:
- Decision Trees – Tree shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset.
- Rulesets – The extraction of useful if-then rules from data based on statistical significance.

Data Mining Techniques

Clustering
- A data mining technique used to discover and explore groupings with data or entities.
- Clustering method allows entities to be partitioned into distinct groups called "segments".
- Example – Segmentation of diseases by its type, such as viral or bacterial.
- Association
- A data mining technique used to find patterns or regularities that are found in transactional-type data.
- Example – customer database in a supermarket.
- 56% of customers who purchased Article 1 also purchased Article 2.
- Sequential
- Involves mining frequently occurring patterns of activity over a period of time.
- Similar to Association Rules, except that Sequential Rules look for patterns across time.
- Example – A sequence of purchases made over a period of time.
- Classification
- A process of determining a predefined class, to which an example belongs to.
- Example – customer database
- Question – Is the customer loyal or disloyal?
- Typical rule formulated – If Purchased = Monthly and Profit >$5000 and Incidents = 0 then Customer_Type = Loyal

Data Mining and CSRL
- The reason we need Data Mining in the CSRL is so we can improve the software JACED_ipables.
- JACED_ipables is the software that use SNORT and IP Tables to improve computer security.
- Using Data Mining, we are able to identify hackers vs. regular wanted traffic.

Acknowledgements

Thanks to Dr. Beheshti, Dr. Kowalski, and thanks to Johnly Tomelden and Joel Ortiz for lending their support in the Computer Science Research Lab.

Future
1. Larose, Daniel T.; Discovering Knowledge in Data: An Introduction to DATA MINING, 2005.