

# Data Mining and Cyber Security

Bhavani Thuraisingham

*Program Director  
Data and Applications Security  
The National Science Foundation  
4201 Wilson, Blvd, Arlington, VA  
On leave from the MITRE Corporation, Bedford, MA*

## **Abstract**

*Data mining is the process of posing queries and extracting patterns, often previously unknown from large quantities of data using pattern matching or other reasoning techniques. Cyber security is the area that deals with protecting from cyber terrorism. Cyber attacks include access control violations, unauthorized intrusions, and denial of service as well as insider threat. We often hear that cyber attacks will cause corporations billions of dollars. For example, one could masquerade as a legitimate user and swindle say a bank of billions of dollars.*

*Data mining and web mining may be used to detect and possibly prevent cyber attacks. For example, anomaly detection techniques could be used to detect unusual patterns and behaviors. Link analysis may be used to trace the viruses to the perpetrators. Classification may be used to group various cyber attacks and then use the profiles to detect an attack when it occurs. Prediction may be used to determine potential future attacks depending in a way on information learnt about terrorists through email and phone conversations. Also, for some threats non real-time data mining may suffice while for certain other threats such as for network intrusions we may need real-time data mining. Many researchers are investigating the use of data mining for intrusion detection. While we*

*need some form of real-time data mining, that is, the results have to be generated in real-time, we also need to build models in real-time. For example, credit card fraud detection is a form of real-time processing. However, here models are usually built ahead of time. Building models in real-time remains a challenge. Data mining can also be used for analyzing web logs as well as analyzing the audit trails. Based on the results of the data mining tool, one can then determine whether any unauthorized intrusions have occurred and/or whether any unauthorized queries have been posed.*

*While data mining can be used to detect and prevent cyber attacks, data mining also exacerbates some security problems such as the inference and privacy problems. With data mining techniques one could infer sensitive associations from the legitimate responses. The presentation will first provide an overview of data mining techniques and cyber threats. Then it will discuss the developments in applying data mining for cyber security. Research challenges will be discussed next. Finally inference and privacy problems that arise due to data mining and potential solutions to these problems will be discussed.*