Introduction: Programs that have the potential to violate the privacy and security of a system can be labeled as Privacy Invasive Software (PIS). These programs include: Spyware which may collect personal information of user and relay it to third party with/without user knowledge; adware which automatically shows the advertisements to users as per their personal preferences; Trojans which are harmful programs and backdoors which can help an intruder to gain remote access of the system. Originally, since the seventies viruses represented the only major malicious threats to computer users and since then much research has been carried out in order to successfully detect and remove viruses from computer systems. However, a more recent type of malicious threat is represented by Spyware. Its presence was first reported in 1999 and this threat has not been extensively studied. Spyware may compromise confidentiality, integrity, and availability of the system. They may obtain sensitive information such as credit card numbers, logins and their passwords, shopping habits, bank account information and personal preferences with or without user consent. In 2003, FTC released the report that 27.3 million Americans are victims of different Spyware in the last 5 years and in 2007 report states about 30.2 millions American as victims in one year. From this increasing trend, it can be estimated that in the coming years the problem of Spyware will grow. The increase of the amount of Spyware released is due to commercial motivation in the form of revenue for different actors in commission or fees paid to them or profit by sale of personal information to companies. There are many big companies such as BMG, CheapTickets, Citi, Howard Johnson, Netflix, NetZero, Orchard Bank, Sage Software, Sprint, T-Mobile, and United Airlines, who use the services of Spyware vendors thus giving boost to this industry.

Unlike viruses which are always unwanted, Spyware can sometimes be installed with the user's expressed consent, since it may provide some useful functionality either on its own or by an accompanying software application or may be mentioned in End User License Agreement (EULA) in tricky way. Due to this reason Spyware overlaps the boundaries of what is considered legal and illegal software. Its status may depend upon which user is being asked. However in most cases, the Spyware vendors do not seem to provide the user any realistic opportunity to give an informed consent or to reject the installation of a software application in order to prevent Spyware from being installed.

Problem: Knowledge about Spyware is generally perceived as low among common users and the process of Spyware identification and removal is often considered as outside their competence. Users may have anti-virus software installed but it may not be helpful against Spyware unless it is designed particularly for Spyware, as Spyware differs from regular viruses e.g., in that they use a different infection technique and stealth technique.

> *"ILOVEYOU" virus, appeared in May 2000, spread itself as attachment with self generated emails. When penetrated in computer, it attached its copy with all the files on computer and also started sending infected emails to all the contacts in user's address book. While a famous Spyware "Gator" was installed by users themselves or without their knowledge when they visited a website, clicked on an advertising link or installed files obtained from file sharing software e.g. KaZaA, eMule,and iMesh."Gator" provided useful functionality of storing personal information such as username, passwords, and credit card information for visited website to the users but also installed a program "OfferCompanion" which tracked visited website and provided information to advertisers to show their pop-up ads.*

Specific anti-Spyware tools have been developed as countermeasures but there seem to be no single anti-Spyware tool that can prevent all existing and future Spyware. Current anti-Spyware tools make use of: signature-based methods which uses specific patterns / information, called signature, extracted from Spyware and match them in any file for detection or heuristic-based methods which uses the rules made by human experts to detect new Spyware, as approaches against Spyware. Signature-based systems demonstrates good detection results for known Spyware but often lacks the capability of identifying new and unseen instances. Heuristic-based systems try to detect known and unknown Spyware on the basis of rules. The heuristic method is considered costly, time consuming and often ineffective against new Spyware. So it is needed to apply some other existing technologies which can help in detecting both known and new Spyware.

Aim: To find a viable solution for distinguishing between Spyware and benign software based on data mining. Data mining will help in finding and describing structural patterns in data of Spyware which will be used to make predictions about known and new instances.

Objectives: I try to find a solution for detection of both known and unknown Spyware by using data mining. Data mining is used for detection of pattern and/or finding correlations between data and using them for unseen scenarios / situation. For data mining problems, learning algorithms are applied to detect patterns and to find correlations between data instances and attributes. On the basis of this detection, applications can be classified as PIS or not. During this process it will be investigated that which features, such as n-grams of byte sequences i.e. specific length string of hexadecimal dump of program, instruction sequences, calls to API or DLL, text strings, of binary/executable files can be used for distinguishing between legitimate software and Spyware. I try to find suitable approaches for reducing the number of features (extracted from the binaries) and determine which learning algorithms and their parameter configurations are suitable for optimized spyware detection rate.

Raja M. Khurram Shahzad
Blekinge Tekniska Högskola