Summary
Surprising knowledge can only be derived from data if we have a mechanism for specifying what is not surprising. One possible approach for specifying expectations of data is the use of probabilistic models. Recently, a large number of languages have been developed for specifying probabilistic models. One such probabilistic programming language will be used in this project to look for surprising patterns.

Background
An important aim of data mining is the discovery of interesting patterns in data. Examples include the discovery of surprising combinations of supermarket purchases, unexpected co-expressed genes in gene expression data, or unexpected connections in social networks.

However, how can we define what is expected? One approach is to use probabilistic models, such as Bayesian networks, Markov networks or Markov models.

In recent years, several languages have been developed that allow for the specification of probabilistic models. This makes it easier to develop probabilistic models. Examples include Markov logic, ProbLog, Factorie, … see http://probabilistic-programming.org/ for a list of such languages.

Till now, these languages have not been used in a data mining context aiming at the discovery of patterns. However, if they could be used successfully, this is expected to make data mining easier and the results of data mining more useful.

The student in this project will chose one of these languages and study its use in a data mining context.

Prerequisites (helpful, not mandatory): Data mining and statistics, Algorithms, Databases, Concepts of programming languages
Summary
Recently a relational constraint algebra has been proposed in which a large number of data mining tasks can be expressed. This language is similar to relational algebras as known from your database classes. However, the equivalent of SQL is still missing for this formal language. This project will develop an SQL-inspired language for relational constraint programming and implement a prototype for it.

Background
A wide range of data mining tasks has been studied in the scientific literature. Implementing a large number of these different data mining tasks is usually not easy. The key idea underlying declarative data mining languages is to create languages in which programmers can express and solve a large number of data mining tasks more easily.

A recent paper has proposed a framework called “relational constraint programming” for the formalization of a large number of different data mining tasks. This framework basically extends relational algebras, as seen in classes on databases. The equivalent of SQL does however not exist yet within this framework.

An SQL-like language is expected to be useful to many programmers, as most programmers are familiar with SQL. Hence, such a language promises to be easy to use.

Prerequisites (helpful, not mandatory):
Data mining and statistics, Algorithms, Databases, Concepts of programming languages
Deep Learning on Molecules

Summary
Deep learning techniques have recently been used in a large number of different problem domains to build accurate classifiers. Based on neural networks, these techniques usually require that examples are described in so-called feature vectors of fixed length. In this project, it will be explored how deep learning techniques can be applied if examples are not of fixed length. As a test case of such a setting, deep learning on molecules will be explored: molecules can be seen as graph structures, and can be arbitrary of size.

Background
Deep learning techniques have quickly gained popularity in recent years. They have been used successfully to learn very accurate classifiers in a wide range of settings, and are used in applications such as Apple's Siri and Google's search engine.

Deep learning techniques essentially learn very deep neural network structures. In their basic setting these networks have an input layer of fixed length.

However, for many types of data the input can not be described using vectors of a fixed length. A good example of this are chemical molecules: molecules can be graphs of an arbitrary size.

It is hence not clear how deep learning techniques can be used on molecules.

In this thesis project, it will be explored how deep learning techniques can nevertheless be applied on such data.

Prerequisites (helpful, not mandatory):
Data mining and statistics, Algorithms
Knowledge of biology or chemistry is not required.