Instructors:

Dr. Jiawei Han is a highly acclaimed professor in the field of Data Mining. His explanation of the material ensures a very thorough understanding and his slides include a lot of detail which is very helpful when reviewing for the exam.

Prerequisites:

CS 225 (Data Structures) is the only listed prerequisite for this course, and it is definitely required to understand the many algorithms covered in this course; a good understanding of data structures is important to be able to do well in this class. Any probability/statistics courses would be a great help too - usually the ones required by the curriculum are more than enough - MATH 461 (Probability Theory), STAT 400 (Statistics and Probability I) or ECE 313 (Probability with Engineering Applications). These skills help to keep up with the course content.

When to Take It:

This course is offered in the Fall Semester only. Typically, it is taken during one's Junior/Senior year, particularly when exploring the various fields of Computer Science to figure out what is of interest.

Class Content:

This class starts off with basic data management/preprocessing techniques and builds into Data Warehousing. Warehousing techniques are covered in some detail, and the advantages of setting these up are also compared to other ways of storing data. Then the course digs deeper into OLAP technology such as data cubes, where the different ways of parsing the data cube to use stored information, as well as optimizations to these algorithms, are covered in great detail. It then proceeds with frequent pattern mining, classification and clustering. Many algorithms for each of these are covered in great detail, and they are compared to understand exactly why some algorithms are better suited for some data sets. These are all widely used techniques beneficial to anyone dealing with data analysis. The latter half of the class, which includes pattern mining, classification and clustering, definitely seemed more interesting, but the level of difficulty was also much higher.

Work:

The workload really depends on how interested you are in the course. If you like the material and the field, you would probably spend a lot more time on it, but if not, a minimum weekly requirement lies between 4-5 hours of keeping up-to-date with the readings & doing homework. There is a lot of reading involved, and not keeping up can really take its toll during the exams. There is only one MP, typically in the student's choice between C++ or Java, so knowing how to program in one of those languages is essential. There are three problem sets, each spanning about 3 weeks. These also require reading the textbook and keeping up with lectures.

Life After:

This class leads the way into Data analytics and Data Mining. Those interested can consider going on to take CS 512 which requires a project as well. This class is considered part of both the Databases track and the Artificial Intelligence track, so students who enjoy it may want to consider taking other database classes such as CS 411 (Database Systems) and CS 410 (Text Information Systems) or other AI classes such as CS 440 (Artificial Intelligence) and CS 446 (Machine Learning).