What is Data Mining?

Many researchers like to collect a large amount of data on many possible response variables in the hope to find some explicit relationship amongst different treatments and make informed decisions from this database. When undertaken on a very large scale, this practice is known as *data mining* or sometimes called *data discovery*,

Data mining is a form of secondary data management but is incredibly useful in exploring large relational databases, often collected through observations or aggregated from different sources. It allows researchers to analyze data from many different angles, categorize it and summarize the inferred relationships identified.   At its simplest, the purpose of data mining is to determine correlations between many variables, which might later form the basis for constructing effective experimental designs. In industrial and commercial contexts, data mining may be used in create decision rules in production or service systems based on relationships observed in the data collected. For example, a bank might have noticed that its bank customers with annual income greater than USD 100,000 and living at an address longer than 3 years have never defaulted on their bank loans. Hence, the bank might decide to offer loans to customers who meet these requirements and also who currently do not have a loan. However, generally no causal inference is made, the decision rules are pragmatic in nature.

The data mining approach, however, is a less defensible in traditional experimental contexts. It is not considered appropriate to forgo the process of stating hypotheses and testing them, and, conducting many statistical tests in the hope that something will come up significant is called *fishing* (as in fishing for results). The reason is that *p*-values are valid for a single test, not for a variety of similar tests on the sample data. When multiple tests are performed, the experiment-wise Type I error rate is almost certainly higher than the *p*-value for a single experiment.   The exception, of course, is if all the tests are completely independent.