

# Data Mining/Analysis Checklist

---

Data mining or data analysis involves conducting a project where the student researcher is not the individual who originally collects the data to be used. In other words, the student does not do the 'experimentation' phase of the project him or herself, and someone else did not do the experimentation or data collection specifically for the student. This is a perfectly acceptable method of research and is common in the scientific world. However, if a student does a data mining project, he or she must then conduct an original analysis (or possibly a conformation of previous work necessary to support debated issues) to derive some sort of inference from it. The most common method for this is to use inferential statistics to support/verify conclusions.

Inferential statistics are mathematical principles, based on probability, which can allow researchers to make assertions within a margin of certainty. Common examples are hypothesis testing, confidence intervals, and regression models with correlation coefficients. Some form of these must be used in data mining projects. Descriptive statistics [such as graphs, tables, charts, measures of central tendency (mean, median, mode) and distribution (variance, standard deviation)] are not sufficient by themselves because they are essentially alternate ways of displaying data. Any conclusions that can be drawn from them are subjective and not based on mathematical and other objective principles.

The data used must be public domain data, or written and signed permission from the owner must be included in the proposal. If the data are of a confidential nature (as determined by the SRC and/or IRB), then it becomes a Human Subjects project, which requires informed consent from each subject before the data can be used. **Remember, just because it comes from the internet, or you can otherwise get access to it, doesn't mean it is public domain.**

**EXAMPLES OF PUBLIC DOMAIN DATA:** Information from research journals, magazines, newspapers, TV and radio broadcasts, commercial or government pamphlets – essentially, anything produced for the general public's use, without expressed restrictions.

**EXAMPLES OF NON-PUBLIC DOMAIN DATA:** Medical files, personal information (date of birth, SSN, etc.), school records and information, other people's or institutions' research not published.

## Things to Be Sure You Are Aware of & Include in Your Project Paperwork

- ✓ You need to list all of the specific sources from where the data will be obtained. You can't say you'll find it later. It must be determined beforehand. (List bibliographic info, websites, broadcast info, etc.)
- ✓ If the data is not public domain data, then written and signed permission from the owner needs to be attached to the experimental design.
- ✓ If it is unclear if your data is public domain data or not, it must either be clarified that the data is public domain or written and signed permission from the owner needs to be included.
- ✓ If the data is of a confidential nature, it requires informed consent from each subject before it can be used in the project. Remember, this requires additional paperwork. See the human subjects requirements in the ISEF rules (<http://stemed.unm.edu/PDFs/RESEARCH%20CHALLENGE/2008-2009%20ISEF%20RULES.pdf>) or talk to your sponsor for specifics.
- ✓ Be sure the data analysis is present and valid for this project.
- ✓ The experimental design needs to list the inferential statistics that will be used in supporting/refuting your hypothesis. They must be appropriate for the data and project type.
- ✓ Be sure the conclusions the project wants to make are not commonly accepted everyday knowledge. The project needs to make original inferences or support/refute debated issues.