COSC 757 Data Mining Spring 2016
Project Proposal
Kevin Kuo and Mary Snyder

Problem Area for Investigation:

The crime data set contains many attributes and statistics about the communities in which crime data is collected. The crime data set is one of the most extensive data sets by the Federal Bureau of Investigation. We will analyze the data set and determine which indicators are associated with areas of high violent crime and more importantly how changes in indicators precede an increase or decrease in per capita crime levels.

Literature Review:

1) Redmond and Baveja 'A data-driven software tool for enabling cooperative information sharing among police departments' in European Journal of Operational Research 141 (2002) 660-678
2) Law Enforcement Information Sharing | ISE 2016. *https://www.ise.gov/law-enforcement-information-sharing*. Accessed: 2016-02-28
3) It's criminal: Why data sharing lags among law enforcement agencies: 2016. *http://www.computerworld.com/article/2486359/government-it/it-s-criminal--why-data-sharing-lags-among-law-enforcement-agencies.html*. Accessed: 2016-02-28
4) Miller, Patrick. How Can We Improve Information Sharing Among Local Law Enforcement Agencies? September 2005. *http://www.dtic.mil/dtic/tr/fulltext/u2/a439576.pdf* Accessed: 2016-02-28

Purpose of Our Experiment:

If changes in crime rates can be predicted based on changes of highly correlated indicators, resources could be better allocated to help local governments in addressing crime occurring in their jurisdiction.

Dataset Description:

"Abstract: Communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR." The dataset contains much information on variables that may be related to violent crimes, the community in which they took place, as well as the law enforcement presence in the area.

http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime

Data Mining Techniques:

We will begin by taking an exploratory data analysis approach. We will first visualize the data we have gathered and then begin comparing multiple variables. By comparing multiple variables through histograms, scatter plots, and various other visualization trends, we will note any obvious patterns by examining their two-way interactions. We will try to describe this relationship with a type of relationship that best fits. Once we have completed most of our exploratory data analysis, we will begin to perform testing and cross-validation.

COSC 757 Data Mining Spring 2016

Project Proposal

Kevin Kuo and Mary Snyder

Risks:

With a dataset this large, there is a high risk that we will not be able to filter out the "noise" to find a definite answer to our question. There are such a large number of factors contributing to any trends in the data that it may be hard to isolate which is actually the actual cause and what may be red herrings. Another risk we may run into is expectations of trends. While people strive to be bias free in their research there are certain ideas that seem intuitive about crime and criminal activity in general. For example, some people believe there is more crime in areas of lower income, but this may not be true among all races, geographical areas, etc. We will need to be diligent in minimizing assumptions and avoiding trying to get the data to match an assumption when no correlation exists. Another significant risk is the accuracy of the data set we will be using. The FBI uses the Uniform Crime Reporting Statistics; however, that data is first collected by local law enforcement agencies, which may not have the same reporting criteria or standards. This possible discrepancy is a point of discussion among many crime studies.