For a while, I have had the idea of some kind of system or algorithm that can correlate weather data (e.g. from weather underground or similar) with traffic data (e.g. from Google maps). Such a system could determine which geographic regions might be particularly affected by different weather conditions (rain, snow, wind, etc.). Ultimately, this could be used to predict traffic delays, perhaps later in the week, given the weather forecast. I could even be used to try to reduce traffic incidents at the most critical locations, perhaps by sending additional police personnel, use of variable-text traffic signals/warnings, etc. Furthermore, this could also be extended to predict the effects of phenomena other than weather (concerts, sporting events, and other events that are not every-day occurrences) on traffic patterns.

The large public datasets for this project are the weather and the traffic information. Analysis will include cross-correlation between different weather patterns and traffic patterns (e.g. snowfall with traffic slowdown) in a given area, which might vary by highway, backroad, etc. It could even be feasible to use radar data to determine the effects of isolated storms on certain areas or roads in a city. This would require spatial correlation between roads and radar data via longitude and latitude coordinates. This information could be very useful to drivers as well as to makers of GPS and navigation hardware and software. Finally, the potential to reduce traffic incidents and increase safety makes this a very attractive project to me.

As an illustration of the feasibility of this project, I have extracted some weather data for a few cities (Columbus, OH; Boston, MA; San Francisco, CA; and Seattle, WA) from Weather Underground via their API with Python. The data I extracted was 36 hour forecast of temperature [F] and rainfall [in]. Similar data is available for historical records and could be scraped from their archive at: <http://www.wunderground.com/history/>. I started working on a script for this using Beautiful Soup but was not able to get it functioning in time for submission.

For traffic information, I found a dataset (22MB) on <http://catalog.data.gov/dataset> for Seattle traffic flow counts in 2011. I have done some very preliminary processing of this data in Python to show some of the trends. The biggest issue that I foresee perhaps being a concern is getting current traffic data from Google. While past traffic and weather data could certainly be used together with a machine learning algorithm (likely multi-variate regression) to predict the likelihood of certain weather events on traffic delays, once a classifier is built it would be much preferable to use current or at least recent traffic data. From what I have read it seems that this can likely be scraped from Google maps, but it may not be trivial. However, I feel that these is a lot of potential in this project, and it would be worthy of in depth study. At the very least, analysis and correlation of historical data could provide some valuable insights into the effects of weather patterns on traffic.

As a side note, the codes and other relevant files from the challenge questions are posted on my GitHub page at <https://github.com/joeplattenburg/data_incubator>.