Problem Description 1:

According to a March 2009 report put out by the Nielsen Company, two-thirds of the world's internet population visits social networking or blogging sites [1]. These sites have surpassed "personal Email to become the world's fourth most popular online sector after search portals and PC software applications." The study notes that "87.25 million U.S. users visited Facebook from home and work... and each of those people spent an average of 4 hours, 39 minutes and 33 seconds on the site during the month." Time on Facebook is generally spent reading other user's updates; interacting via comments, "likes," or wall posts; posting pictures or video; and providing updates.

Facebook user interactions provide a myriad of challenges to an applied statistician. The first challenge: acquiring the data. As a starting point, I have implemented a python script which searches for new updates every 50 minutes and adds them to an XML file. The XML file holds the name of the person making the update, the update given by the user, the time of the update, the number of comments for the update, and the number of other users who "like" the comment. While this script was written in python, this is not the only way to scrape Facebook updates. Presumably, the RCurl package could be used to write an R script which logs into Facebook, goes to the updates page, scrapes the page for updates, and writes this information to a file.

After the data has been acquired, analyses can be performed on the updates. Updates are generally short, concise, statements which give some indication of a user's emotional status or geographic location. One approach to analyzing these statements, currently being employed by Martin Schultz in the context of analyzing the text in biostatistics journals [2], is to create a matrix where rows correspond to updates and columns correspond to the number of times a word appears in the update. Techniques such as principle component analysis can then be used to reduce the dimension of the data and facilitate further exploration.

One such exploration would be to attempt to understand what makes an update "likable" and what kind of updates generate interaction (comments). It is possible that a user gets high "like" or comment counts simply because she is, in some sense, popular. It is also possible that a user gets higher like or comment counts because of the content of the updates she gives. Theses two alternatives can be explored via regression on the scraped data set along with the dimension reduced data set.

[1] http://blog.nielsen.com/nielsenwire/wp-content/uploads/2009/03/nielsen_globalfaces_mar09.pdf
[2] from a conversation with Martin Schultz, June 2009