Putting it all together

If you need help at any time, put your **red** sticky note on the back of your laptop. When you've finished the steps on the *front* of this page, put your **green** sticky note on the back of your laptop. Then, you can turn the page over and try out some of the more advanced tricks on the back while you wait for the rest of the group to be ready.

Your task

In this exercise, you'll try and put together everything you've learned so far today, to achieve a practical task.

Your goal is to retrieve and process a set of data files on bike trips taken using the Citi Bike bike sharing program. You are trying to answer the question: which are the most popular stations in the Citi Bike program? Specifically, you want to know which stations appear most often as either the "Start Station" or "End Station" in a dataset of bike trips from January through September 2016. Your output should be a file in which each row is the name of a station and the number of times it appears as either the "Start Station" or "End Station" in the dataset, and the file should be sorted so that the most popular stations are at the top.

You want to include your entire data analysis workflow in a single shell script, to make it easier for others to reproduce. This will also help you if, for example, you decide to repeat this analysis with some small changes (such as running it for 2015, too.)

The data files are available at http://witestlab.poly.edu/bikes/

Hints

- You can find information on many useful Bash shell utilities online. Search for " man page" to find complete usage information; or for " examples" to find usage examples.
- Create a new directory, and make sure all your work happens in there. If at any point you mess up and want to start "fresh", you can just delete it and start again in another new directory.
- Use `unzip` to extract files from a ZIP archive.
- Remember that you can use >, >>, and | to chain together multiple tools!
- `awk` is a useful tool for getting a particular column (or set of columns) from a CSV file. To print the first column of a CSV file `file.csv` to your terminal output, run

```
awk -F "\"*,\"*" '{print $1}' file.csv
```

- The `wc` utility can be used to count the number of words, lines, or characters in a file (or other input).
- The `sort` and `uniq` utilities can be used in combination to get the subset of unique values from a long list of values. (For example, to get the complete list of stations... hint, hint.) First `sort` the values, then pass them through `uniq` to filter out multiple adjacent copies.