# STA 141C - HW 1 - Group By

Version 1.1

- See Canvas for due dates and grading rubric.
- You can answer these using whatever programming language you prefer, although we will be teaching and supporting R.
- Turn in a neatly typed report that answers the questions in clear English, using complete sentences. It should be 2-4 pages (single sided, including tables and figures).
- Attach your code in an appendix. This code must run, and support your answers.

## Overview

We're going to look at US federal government spending broken down by agency. In this assignment we'll learn:

- The 'group by' model of computation
- How to process data spread over many files

The goal is to learn about the limits of data processing on a local laptop or desktop. Later, we can compare this to the performance of a remote cluster.

## Data

The data is available at http://anson.ucdavis.edu/~clarkf/sta141c/ see `awards.zip`. Inside this zip file are 720 files. Each file corresponds to one agency. `0.csv` is the exception- it contains those records where the funding agency field was NULL. You do not have to process `0.csv`.

Download `awards.zip` onto your local machine. The data is 1.5 GB compressed, and 7.4 GB uncompressed, which becomes unwieldy on a personal machine that only has 8 GB of memory.

Look up the meanings of the data and fields at https://www.usaspending.gov/#/.

## Hints

You may find these R functions useful as you complete this assignment:

`unzip, unz, list.files, lapply, tapply, median, log, hist, file.info, system.time`

## Questions

### 1 - Computation

Compute the median annual spending for each agency.

1. Which agencies have the highest median annual spending?
2. Qualitatively describe the distribution of median annual spending.
3. Qualitatively describe the distribution of the logarithm of the median annual spending. Plot the histogram.
4. Is there a clear separation between agencies that spend a large amount of money, and those which spend less money?

### 2 - Reflecting

1. Qualitatively describe the distribution of the file sizes.
2. How does the size of the file relate to the number of rows in that file?
3. How long does it take to process all the data?
4. Do you think this same approach you took work for 10 times as many files? What if each file was 10 times larger?
5. How do you imagine you could make it faster?