CSCI 2300 – Crowdsourcing Assignment Due: February 22 at 1pm

Crowdsourcing is a way of solving problems that are hard for computers to do, by systematically subdividing the work to be done by a group of people.

Before you begin, do a little warm up—see what it takes to make money on Mechanical Turk. Look at the available tasks (HITS), and see what is asked to earn $0.01, $0.05, or $0.25.

This assignment's overall goal is to *develop a complete listing of computer science professors at top US and Canadian universities*, along with metadata like their degrees, research area, year they joined, and rank. This is a hard task because the information is spread across department websites, faculty curriculum vitae, bios, and other miscellaneous sources. It is also tricky for the workers who probably do not understand computer science or academia, e.g. terms like "Sc.B." might not make sense, or "machine learning" and "computer education" seem semantically similar. There are two parts to the assignment: first to collect the data (data collection), and second to validate and correct the data (data validation).

To select your university, pick a University Group Number (which correspond to five universities with roughly the same total size) on this signup spreadsheet `http://bit.ly/2Ep1xts`. For each university, you will use a different data validation strategy, but the same data collection strategy.

You will receive $50 to spend on Amazon Mechanical Turk, a popular crowdsourcing platform. $5 of this money should be spent for initial testing (test by collecting data about a department in the School of Computer Science at Carnegie Mellon University). The remaining $45 should be divided equally: $9 for each university with $4 going to data collection and $5 going to data validation. Once you create a Requester account on Mechanical Turk you can add $50 to your balance, then submit the receipt from Amazon to Dawn Reed (4th floor CIT) who will reimburse you after the assignment. Make sure that your receipt shows that the funds were used for Mechanical Turk, and do not spend more than $50 regardless.

Ensure you provide informed consent that this is part of a class and potential research—take a look at example informed consent messages online. Your data should have these columns in this exact order:

- **Name**: the full name of the professor.
- **University**: the year the professor joined the university they are in now.
- **Gender**: male or female.
- **JoinYear**: the year the professor joined the university they are in now.
- **Rank**: one of: Assistant, Associate, Full.
- **Subfield**: the main research field of the professor.
- **Bachelors**: where the professor received their undergraduate degree.
- **Masters**: where the professor received their Masters degree.

- **Doctorate**: where the professor received their PhD degree.
- **Sources**: links to where the information was gathered from, for reference.

Here is an example template with 5 professors from the University of Washington - USA filled in: `http://bit.ly/2EmcRGy`. Note the dropdown menus that contain all acceptable answers. Please use the template (copy it), remove the data, and fill out the name and university of the professors that fit the (rank) criteria in the spreadsheet leaving all other cells blank.

To start, collect data by assigning Amazon Turk workers to collect information about your professors, essentially filling out the row for each professor. This includes the fields above except Name and University (which will be already filled in by you). For data validation, randomly assign each of the universities to one of the five data validation strategies below so that each strategy is uniquely mapped to a university. Use a dice or the random number generator on Google.

- **Find-Fix-Verify:** As in the Bernstein paper.
- **Find-Fix:** Only the Find and Fix steps as in the Bernstein paper.
- **Find and Fix:** The Find and Fix steps done by the same worker.
- **Majority Rule:** Doing redundant data collection until a majority of workers' data (and at least 2) agree.
- **Expert Rule:** Doing data collection twice and asking a third worker to resolve differences.

Mechanical Turk requires a fixed payment per task, and an optional bonus depending on how you feel. How exactly you structure the payment and bonus incentives for the different tasks is up to you. It can be different per university but you should use the same approach within a single university. In other words, don't switch up how you pay workers that are collecting and validating data for the same university. You will want to test many levels of payments and bonuses in the testing phase on the Carnegie Mellon University test case. One tip is to make sure you do not limit to "workers with Masters Qualifications" in Amazon as filtering that way seems to be a poor choice from past experience.

We should use consistent vocabulary for research subfields and university names, or it will be hard to analyze the data automatically. For example, your workers should select university names from the template sheet (these are already set up in the template above). If a university is not included in this list, there should be an option for manual input. When a worker is not sure or cannot find a field they should leave it empty. Make sure your workers are aware that they should only collect professors with ranks of Full, Associate, Assistant (not adjunct, research, visiting, or emeritus faculty). Don't waste your funds on non-relevant professors (who are not Full, Associate, or Assistant professors).

Document everything! Keep a journal of your work as you go. Report your ideas, what procedure you followed, what results were observed, communication with the workers, and whether that was expected. Include at least the following information in your report: completion time for each task, any communication with workers, total number of unique workers per task and in total, screenshot of the tasks you released. Also save the spreadsheets after data collection, but before data validation. Make sure you account for every penny spent in a spreadsheet: what university and what data it was spent for, whether it was for data collection or validation, when it was spent, and bonuses and incentives you used. You can then compute the payment per faculty for each university and the total amount spent to collect all data for each university.

Your journal should eventually be in pdf format and contain all of your work. You will also create and submit a separate .csv file for each of the four universities you crowdsourced. Do not fix by hand any of the data generated by workers; after setup and submitting new tasks, the only manual work involved from you should copy and pasting. In essence, only do tasks by hand that could be automated by a program (otherwise your crowdsourcing method won't scale).

At the end of the assignment, we will review together everyone's work to extract key lessons learned. As a class, we will gain first-hand experience of what works and what doesn't in crowdsourcing. This listing would benefit students applying to graduate schools or academic positions, people doing analytics on computer science trends, journalists and recruiters could use this as a source. Be rigorous with your work so we can analyze it in a group paper later.

This is a assignment that requires waiting patiently! Start early—it takes time for workers to do your tasks (usually several days), and it might take a few tries to get it right. On February 6 (check-in), bring your test data to class (printed) so you can get some feedback and come with a plan for collection and validation. You should also have some thoughts of what is working or not working to share with the class. Your grade will be based on the descriptiveness of your report (6 points), a reasonable data validation technique you created or adopted (4 points), your check-in progress (2 points), thorough instructions given to workers (4 points), and the overall quality of the data output (4 points) for a total of 20 points.

Will you be a benevolent Requester? Will you pay more or reward more in bonuses. How will you deal with those who submit bad data? Only you can decide!