CSCI 2300 – Crowdsourcing Assignment Due: Thursday Feb 19, 11:55pm

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.

This assignment's overall goal is to *develop a complete listing of computer science professors at nearly 100 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." or "machine learning" can be confusing to them. There are two parts to the assignment: first to collect the data using a predefined strategy below (data collection), and second to validate and correct the data using any technique of your choosing (data validation).

To select your university, pick a University Group Number that correspond to 4 universities on this signup spreadsheet `http://tinyurl.com/nf4vjrf`. Once you pick a University Group Number, you will find the strategy you will use to collect data for all four universities at the bottom of the signup spreadsheet. The 5 data collection strategies are:

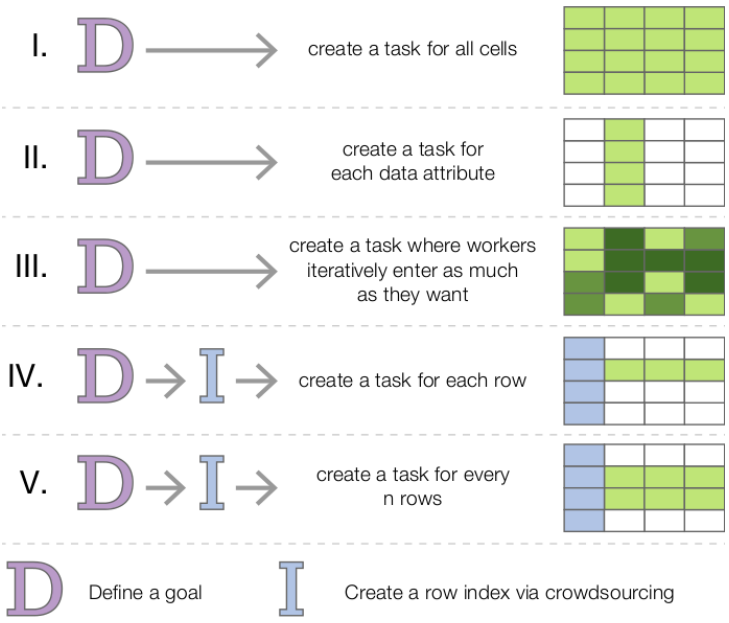
I. One worker enters all the information for a university

II. One worker enters only a specific column e.g. the PhD degrees of all faculty

III. Multiple workers have access to the same spreadsheet, adding to any missing cells

IV. Each worker enters all the information for a single faculty (a single row)

V. Each worker enters all the information for multiple faculty (some but not all rows)



Since the data collection strategies will be straightforward, we ask you to be creative in the data validation strategies you will use. When considering data validation techniques, consider:

- Is it worth identifying the skill level of workers, or filtering out bad workers initially?
- How will you validate the data, and what will you use as the gold standard?
- Will you have multiple workers collect the same data for redundancy and overlap?
- How will you decide how much to pay each task?

You will receive $40 to spend on Amazon Mechanical Turk, a popular crowdsourcing platform. $8 of this money should be spent for initial testing (about $2 for each university). The remaining $32 should be divided equally: $8 for each university; $4 should be assigned for data collection and $4 for data validation. Once you create a Requester account on Mechanical Turk you can add $40 on your balance, then submit the receipt from Amazon to Saara Moskowitz (546 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 $40 regardless.

Ensure you provide informed consent that this is part of a class and potential research. You may use a spreadsheet like Google Docs, survey form, or any other tool for workers to input data. Your data should have these columns in this exact order: Name, University, Gender, JoinYear, Rank, Subfield, Bachelors, Masters, Doctorate, Photo, Sources.

- **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.
- **Photo**: direct link to a photo of the professor.
- **Sources**: links to where the information was gathered from, for future reference.

Here is an example template with 5 professors from the University of Washington - USA filled in: `http://tinyurl.com/lmo6ea5`. Note the dropdown menus that contain all acceptable answers. We should use consistent vocabulary for research subfields and university names, or it will be hard to analyze the data automatically. The subfield should be one of these 24 options:

`http://tinyurl.com/l9ao8p5`, and your workers should select university names from this list `http://tinyurl.com/on8vbdo` (these are already set up in the template above). If a school 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). We will probably end up with around 3,000 professor in the listing from the entire class. Part of this data has been collected by the students who took the same class last year. We will use their data as ground truth to evaluate your submissions.

Document everything! Keep a journal of your work as you go. Report your ideas, what procedure you followed, what results were observed, and whether that was expected. Include at least the following information: payment per faculty, total amount spent to collect all data for each university, payment for index of faculty, bonuses and incentives you used, 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. Include a copy of the data collected before any validation or corrections.

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, and the results have the potential to be widely visible or to be published. 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. There is also some potential for research to come out of this, perhaps as a paper that compares the data collection and validation techniques, and qualitatively describes the lessons learned. So be rigorous with your work so we can include your contribution.

This is a big assignment! You have many days to do it, but start early. It takes time for workers to do your tasks (usually several days). On February 4 (midpoint), we will evaluate everyone's test data and give you feedback. You should also have some thoughts to share with the class. As a class we will see how far along everyone is, look at the data quality, and discuss ideas for improvements. 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 midpoint 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? How will you punish bad workers? Only you can decide!