CSCI 2951L – Crowdsourcing Assignment Final Due Date: February 12, 2014

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

This assignment's overall goal is to *develop a complete listing of computer science professors at the top universities*, along with metadata like their degrees, research area, year joined and rank. This is a hard task since the information is spread across department websites, faculty curriculum vitaes, bios, and other miscellaneous sources. It is also tricky for the workers who probably do not understand computer science or academia, e.g. vocabulary like 'post-doc' or 'machine learning' are not familiar to them.

You will receive $25 to spend on a crowdsourcing service, like Mechanical Turk, Crowdflower, etc. to generate the data. You may use any other crowdsourcing service but I recommend you stick with Mechanical Turk unless you have some prior experience. You can use an additional $5 for initial testing to help you debug your instructions and determine the right incentives. You can either spend and be reimbursed afterwards (easier), or ask for my credit card at the appropriate step when you create your account. My admin, Saara Moskowitz (546 CIT) will help process your reimbursement. Do not spend more than $30 regardless.

Select 5 universities on the first 2 pages of the Computer Science rankings from US News to fill out (sign up in class). At most two students can select the same university.

When designing the crowdsourcing process, you may want to consider:

- What will be the unit size for a task? (micro-tasks vs macro-tasks)
- 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 ground truth?
- Will you have multiple workers collect the same data for redundancy and overlap?
- Will you make the tasks more fun? Provide bonuses? Hold a contest?
- How will you decide how much to pay each task?

Ensure you provide informed consent to the workers that this is part of a class assignment and potential research. You may use a spreadsheet like Google Docs, survey form, wiki, or any other tool for workers to input data. Your data should have these columns in this order:

Name, University, JoinYear, Rank, Subfield, Bachelors, Masters, Doctorate, PostDoc, Sources.

Here is an example of data for five professors from the University of Washington filled in: `http://bit.ly/1eWGZnF`

- **Name**: the full name of the professor.
- **JoinYear**: the year the professor joined the university they are in now.
- **Rank**: typically one of: Assistant, Associate, Full, Adjunct.

- **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.
- **PostDoc** : where the professor did their first PostDoc.
- **Sources**: links to where the information was gathered from, for future reference.

We should use consistent vocabulary for research fields and university names. The research field should be one of these 24: `http://academic.research.microsoft.com/?SearchDomain=2&entitytype=2`. For university names, use either 1) the domain name of the university without the .edu if their domain ends in edu, or 2) its full name (where the title from its page on Wikipedia is the gold standard). For example, Illinois or University of Illinois at Urbana-Champaign (but not UIUC, University of Illinois, University of Illinois at Champaign, etc.), and MSU or Michigan State University (but not Michigan State). We will probably end up with over 1,000 people in the listing from the entire class.

Keep a journal (lab notes) of your work as you go. Report your ideas, the procedure followed, results observed, and whether the results were as expected. Our goal as a class is to find out what works and what doesn't. **Document everything!** Place your journal along with materials created (instructions, data generated by workers, etc.) to a web page as the final report, which you will copy to `/pro/web/web/courses/cs2951-l/[yourcsid]/crowdsourcing.html`. Do not fix by hand any data generated by workers—the only thing you should be doing is moving data around between tasks.

At the end of the assignment, we will review together everyone's work to identify key lessons. 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 published. If at least one student's setup can generate accurate information, I will spend from my personal funds to fill out the listing for the top 100 universities, and this will become a terrific public resource (with your names on it). 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 qualitatively compares the models developed by each student and describing lessons learned. So be rigorous with your work.

This is a big assignment! You have 16 days to do it, but start early. It takes time for workers to do your tasks (usually a few days). On January 31, we will discuss strategies, so come to class with a plan. On February 5 (midpoint), you should have already generated at least some data, and have some thoughts to share with the class. As a class we will see how far along everyone is, whether the allocated money is sufficient, look at the data quality, and discuss ideas for improvements. Your grade will be 70% based on the descriptiveness of your report, and 30% based on the quality of your workers' output.