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* ***Amazon Mechanical Turk, a popular crowdsourcing platform.*** *Look at the available tasks (HITS), and see what is asked to earn $0.01, $0.05, or $0.25. Do a few tasks if it lets you.*

This assignment's overall goal is to develop a listing of computer science professors at top US universities, along with metadata like their degrees, research area, and the year they joined. This is a hard task because the information is spread across department websites, professor 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. When data is collected, it might not be correct and needs to be verified, which is itself a challenge.

Assignment Timelines:

April 5 to April 15 - Testing Phase - Complete warm-up on **Amazon Mechanical Turk (MTurk)**, then as a test run, hire crowdworkers through MTurk to collect and verify data about your 8 assigned CS Professors from Carnegie Mellon University.

April 5 to April 15 4:30pm Eastern Time - Collect Ground Truth - You (not crowdworkers) will collect the ground truth data for at least 50 professors in your 3 assigned universities randomly spread among your universities. This is a hard deadline because the data will be saved and cleared after this.

April 15 to April 29 - Main Data Collection and Verification - Time to crowdsource the data! Create tasks and hire crowdworkers to collect and verify data for your 3 assigned universities.

To view your 8 assigned professors for the testing phase, and 3 assigned universities to collect the ground truth for and to do the main assignment **please visit this spreadsheet.** You're encouraged to vary all possible factors: price, instructions, configurations, validation techniques, methods of qualifying workers, etc. Getting the ground truth will take some time so do allocate enough time for it -- the accuracy of your crowdworkers will be computed by being compared with your ground truth.

The assignment will be conducted using **Drafty, a research system.** See an example on the website, e.g. visit Drafty and search for "Yulia Tsvetkov" from Carnegie Mellon University. Visitors to Drafty are required to sign-up. **It is important that you ask your crowdworkers to use their Mechanical Turk ID to sign-up, not their email.** This way we can track work to ensure fair compensation; the csv of edits will be made available to you. When you visit, please sign-up using your Brown email address. Copy and paste this URL in a browser to download worker edit history.

Come up with and test 3 different crowdsourcing strategies, relating them to the recent readings and in-class discussion about crowdsourcing. You will receive $55 to spend on Amazon Mechanical Turk. $5 of this money should be spent for initial testing (test by collecting and verifying data from a set of 8 professors you were assigned from CMU), this should be done by the midpoint on April 15. The remaining $50 should be divided among the strategies you try to employ. Once you create a Requester account on Mechanical Turk you can add $55 to your balance, then submit the receipt from Amazon to

Dawn Reed (dawn reed@brown.edu) who will reimburse you after the assignment. Please let Jeff know if you are unable to use a credit card to create a balance. Make sure that your receipt shows that the funds were used for Mechanical Turk, and do not spend more than $55 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 will consist of these columns:

- **Name:** the full name of the professor.
- **University:** the university the professor is affiliated with.
- **JoinYear:** the year the professor joined the university they are in now.
- **Subfield:** the main research field of the professor.
- **Bachelors:** where the professor received their undergraduate degree.
- **Doctorate:** where the professor received their PhD degree.

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 tasks is up to you. You will want to test many levels of payments and bonuses in the testing phase. One tip is to not limit to "workers with Masters Qualifications", this has been ineffective in the past due to a lack of workers with that qualification, but limiting to workers with a high HIT acceptance rate can be a good idea. When a worker is not sure or cannot find a field they should leave it empty.

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, the total number of unique workers per task and in total, a screenshot of the tasks you released. 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 verification when it was spent, and bonuses and incentives you used. You can then compute the payment per professor 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. The data entered into Drafty will be automatically checked for accuracy by comparing it to your ground truth data, and will become part of your assignment grade. Do not fix by hand any of the data generated by workers.

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 an 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. Put HITs out there as early as you can, as the less time you can afford to wait, the more you have to pay for the same work to get a faster response. On April 15 (check-in), come with a plan for the collection and verification of your data, and your ground truth data collected. You will share some thoughts of what is working or not with the class.