CS 5306 / INFO 5306
Fall 2017
Project 1

The goal of this project is for you to ask a question about how some crowdsourcing effort works and try to answer it using data about how it operates. The paper we discussed in class, "Crowd Diversity and Performance in Wikipedia: The Mediating Effects of Task Conflict and Communication" by Ren and Yan gives an example of a study (far more elaborate than you will conduct) that gives an example of the kind of thing this project is seeking. They identified a project (Wikipedia), asked some questions about it (how worker diversity impacts article quality, and how that might be mediated by editor conflict and communication), identified some data that allowed them to quantify such factors as diversity, article quality, conflict, etc., and then used that data to answer their question.

Although Wikipedia is well-aligned for a project like this because so much of what happens is recorded by the system, you get to pick the crowdsourcing effort you're interested in. Examples include (but are not limited to) Amazon reviews, Github, the Polymath Project, Wikipedia, question-answering sites like Quora and StackOverflow, citizen science projects like GalaxyZoo (or its siblings at the Zooniverse platform), prediction markets, crowdfunding sites, or any of the many open-source software that you can find on the Internet. If you have a project you want to use that doesn't match these examples and you're unsure if it qualifies please send me an email. Keep your aim modest – I'm not looking for a massive study connecting the latest results in behavioral science with multiple detailed analysis of user logs. I just want you to be curious about some facet of a crowdsourcing system and connect that up with some data that might (probably only partially) answer it. Furthermore, it's ok if you find that you either were wrong about what you expected or the results were simply equivocal or otherwise unable to answer the question. The point is to ask a clean question, figure out how to frame it in terms of data, and then attempt to answer it.

In addition to the crowd diversity paper discussed in class, I'm including here other examples of studies that people have conducted about various systems, to help give you ideas. It is even fine for your project to duplicate an existing study, although I would generally advise against it only because the examples in the research literature are often more elaborate than what you need to do for your project.

The timeline for your project is the following:

- Due Thursday, November 2: Project proposal
  This should be at most one page, and should have the following sections:
  - Your name and your partner's – whoever submits this in Gradescope should link in the other project member
  - The crowdsourcing effort you are targeting
  - The question you are asking
  - How you are answering your question – most prominently, what is the data resource that you will use to answer your question
- Due Tuesday, November 21: Project report
  This should be 3-5 pages (3 is totally fine!) not including graphs, charts, or tables. It should include the same sections as your project proposal, but should have one additional section:
  - Were you successful in answering your question?
    - If yes, critique it. Under what conditions might your conclusion be wrong?

- It no, what happened? If you were to follow up on what you did, what would you now do?

Your report should also include a brief appendix outline the contributions each partner made and the division of labor for the project.

It is ok if your project builds off of something else you are already doing, but the work for this project must be separate and new. If you do work on something related to an outside effort – whether for a job, another course, etc., – you must divulge this in your proposal and project report, and everyone involved must be aware of this and ok with it.

There are no constraints on what programming languages you use – use whatever suits you and the task best. I do not need – and do not want – to see any programs you might have written to do your project. The project is about the question you are asking and the answer you get, and I assume you have the capacity to do what you need computationally to accomplish this.

Here are some papers you can review if you need some inspiration – in addition to the papers themselves, the papers they cite may similarly be of value.

- Citizen science/Zooniverse:
  - Luczak-Roesch M, Tinati R, Simperl E, Van Kleek M, Shadbolt N, Simpson RJ. "Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science". In *Proceedings of the International Conference on Weblogs and Social Media* 2014 Jun. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8092/8136
- Discussion forums:
  - Huang J, Dasgupta A, Ghosh A, Manning J, Sanders M. "Superposter behavior in MOOC forums". In *Proceedings of the first ACM conference on learning@ scale conference* 2014 Mar 4 (pp. 117-126). ACM. http://jonathan-huang.org/research/pubs/las14/hdgms_las14.pdf
- Github:
  - Gousios G, Pinzger M, Deursen AV. "An exploratory study of the pull-based software development model". In *Proceedings of the 36th International Conference on Software Engineering* 2014 May 31 (pp. 345-355). ACM. http://www.gousios.gr/pub/exploration-pullreqs.pdf
- Polymath Project:
  - Cranshaw J, Kittur A. "The polymath project: lessons from a successful online collaboration in mathematics". In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* 2011 May 7 (pp. 1865-1874). ACM. http://www.cs.cmu.edu/afs/cs/Web/People/jcransh/papers/cranshaw_kittur.pdf
- Quora:
  - Wang G, Gill K, Mohanlal M, Zheng H, Zhao BY. "Wisdom in the social crowd: an analysis of Quora". In *Proceedings of the 22nd international conference on World Wide Web* 2013 May 13 (pp. 1341-1352). ACM. http://people.cs.uchicago.edu/~ravenben/publications/pdf/quora-www13.pdf
- Stack Overflow:

    - Anderson A, Huttenlocher D, Kleinberg J, Leskovec J. “Discovering value from community activity on focused question answering sites: a case study of stack overflow”. In *Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining* 2012 Aug 12 (pp. 850-858). ACM. http://www-cs.stanford.edu/people/jure/pubs/sof-kdd12.pdf
    - Movshovitz-Attias D, Movshovitz-Attias Y, Steenkiste P, Faloutsos C. “Analysis of the reputation system and user contributions on a question answering website: Stackoverflow”. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* 2013 Aug 25 (pp. 886-893). ACM. http://www.cs.cmu.edu/~dmovshov/papers/asonam_2013.pdf
    - Tausczik YR, Kittur A, Kraut RE. “Collaborative problem solving: A study of MathOverflow”. In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing* 2014 Feb 15 (pp. 355-367). ACM. http://www.cs.cmu.edu/~ylataus/files/TausczikKitturKraut2014.pdf
- Wikipedia:
    - Yu B, Ren Y, Terveen LG, Zhu H. Predicting Member Productivity and Withdrawal from Pre-Joining Attachments in Online Production Groups. In *Proceedings of the ACM conference on Computer Supported Cooperative Work & Social Computing* 2017 Feb 25 (pp. 1775-1784). http://haiyizhu.com/wp-content/uploads/2017/01/predicting-member-productivity-1.pdf