CS 5306 / INFO 5306
Fall 2025

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. You've now seen some examples of the kind of thing this project is seeking – although you will likely be doing something smaller than these larger research paper efforts. Each of these projects can be characterized by the following:

1. The online system/resource that they considered (Wikipedia, r/Place, the Polymath Project, etc.).
2. One or more questions about how the system operates.
3. Online data that could be used as a window on the question.
4. A report on what the data revealed about the answer to the question.

Your project will do the same. You get to pick the crowdsourcing effort and the question(s) you're interested in answering about it. It need not be one discussed in class. If in doubt, Wikipedia is a good choice to consider, since it has ample data available. (There are some example papers included at the end about Wikipedia and other systems for you to explore if you're looking for ideas.) Your investigation might be totally original, or it might replicate work that has already been done (which can be especially interesting if some time has passed since the earlier study, so the answer might now be different). Your goal is to ask a question and use data that you gather to answer the question as best you can. You will have some belief about what the answer might be, but you will not be judged by whether your guess was right. It might very well be that what you found contradicted what you expected or was equivocal. What's important is being clear about the question you're asking, how you're answering it, and what answer you found.

I'm not looking for a massive study. I do 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. The point is to ask a clean question, figure out how to frame it in terms of data, and then attempt to answer it.

There are three "deliverables" for this assignment:

- Due Friday, September 12, 11:59pm: Project proposal.
  This should be 1-2 pages, and should have the following sections:
  - The names and Net IDs of the students on your team.
  - The crowdsourcing system/project that you are targeting.
  - The question you are asking.
  - How you are answering your question – **most importantly, the data resource(s) that you will use to answer your question**. It's important that you've thought carefully about and investigated this – you don't want to propose a project that you only discover later that you can't do due to lack of data.

  It should be submitted in pdf format on Canvas.
  You will not receive a grade on this (beyond our recording whether it was submitted). Its purpose is to give you feedback on your proposed project.

- Due Tuesday, October 7, 11:59pm: One-slide summary of work.
  You're going to have 1 minute to tell your fellow students what you did. Think of this as meeting

someone in an elevator and giving them a quick description of something you did. It doesn't need to be quantitative or detailed.
It should be submitted in pptx format on Canvas.

- Due Wednesday, October 15, 7:00pm: Project report.

  Describe what you did. Sometimes projects morph a bit along the way. That's ok. You have two goals:

  1. Convey what you asked, how you went about answering it, and what the answer was.
  2. Give some sense of what technical work was involved/what you spent your time working on – make us aware what three people spent four weeks doing.

  It should be submitted in pdf format on Canvas.

A few words of warning:

- Understand the distinction between describing versus understanding. Documenting, for example, the time gaps between user contributions to a system *describes* some element of the system. That is not answering a question, it's describing the system's use. That very same data, however, might be exactly what you need to answer a question that you have, such as observing different gaps for different kinds of users. Make sure you're asking something about how the system works and then using data to answer it, rather than describing some data about the system.
- **Make sure you you have the data that you need for your project before submitting your proposal. This is the most important element of your proposal.**

Here are some papers in addition to those covered in class that you can review if you need some inspiration. Furthermore, many of these papers in turn have citations to yet other papers that might similarly provide inspiration to you.

- Citizen science:
  - "Why Won't Aliens Talk to Us? Content and Community Dynamics in Online Citizen Science", https://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewFile/8092/8136
  - "Citizen science frontiers: Efficiency, engagement, and serendipitous discovery with human–machine systems", https://www.pnas.org/content/pnas/116/6/1902.full.pdf
  - "Collaborative Problem Solving in an Open-Ended Scientific Discovery Game", https://dl.acm.org/doi/pdf/10.1145/3134657
- Open source software:
  - "An Exploratory Study of the Pull-Based Software Development Model", https://www.aau.at/wp-content/uploads/2019/11/Gousios2014-pullbasedmodel.pdf
  - "How do firms influence open source software communities? A framework and empirical analysis of different governance modes", https://www.sciencedirect.com/science/article/abs/pii/S1471772715000111
  - "The Perils and Pitfalls of Mining SourceForge", https://surface.syr.edu/cgi/viewcontent.cgi?article=1109&context=istpub

    - "Herding a Deluge of Good Samaritans: How GitHub Projects Respond to Increased Attention", https://deepblue.lib.umich.edu/bitstream/handle/2027.42/153786/Maldeniya%20et%20al.%202020.pdf
- Online math:
    - "The polymath project: lessons from a successful online collaboration in mathematics", http://www.cs.cmu.edu/afs/cs/Web/People/jcransh/papers/cranshaw_kittur.pdf
    - "Collaborative problem solving: A study of MathOverflow", https://kraut.hciresearch.info/wp-content/uploads/tausczik14-CollaborativeProblemSolving-MathOverflow.pdf
- Wikipedia:
    - "Content is King, Leadership Lags: Effects of Prior Experience on Newcomer Retention and Productivity in Online Production Groups", https://dl.acm.org/doi/pdf/10.1145/3173574.3174080
    - "Participation of New Editors after Times of Shock on Wikipedia", https://ojs.aaai.org/index.php/ICWSM/article/download/3253/3121
    - "Out With The Old, In With The New? Unpacking Member Turnover in Online Production Groups", http://www.xinyiwang.org/uploads/8/6/7/1/86712134/wikiprojectnewcomerturnover.pdf
    - "Predicting Member Productivity and Withdrawal from Pre-Joining Attachments in Online Production Groups", http://haiyizhu.com/wp-content/uploads/2017/01/predicting-member-productivity1.pdf