

Responsible data management, Spring 2026

Class info

INF 385T

In-person. Mondays. 9:00AM-12:00PM UTA 1.204

Instructor

Hanlin Li, PhD, she/her, Assistant Professor

Email: lihanlin@utexas.edu Office: UTA 5.412

Office hour: TBD and by appointment

Course Description

This course will explore common data collection, management, and sharing practices in information technology and emerging technologies. Students will examine the human, social, and ethical impact of these practices and work on group projects to design data systems that are centered around broader impact and social responsibilities.

Prerequisites for the course

None.

Required Materials

All course readings will be available via the course Canvas site.

Long Description

This course will explore common data collection, management, and sharing practices in information technology and emerging technologies, such as search engines and AI systems. Students will read papers and engage in discussions about the pros and cons of established data practices and learn about the three main components of responsible data management: 1) consent and ownership, 2) privacy and anonymity, and 3) broader impact.

Students will also practice how to collect data, make data-driven decisions, and design data-driven products through group projects as UX designers, researchers, and data scientists.

The course will bring in interdisciplinary perspectives with guest speakers from archival science, engineering, and responsible AI, to provide a holistic view of broader data ecosystems and infrastructures.

Learning outcomes

Students will learn the pros and cons of different data collection, management, and sharing practices through readings, discussions, and case studies.

Students will gain hands-on experience with responsible data management or systems as UX designers, researchers, and data scientists by completing a group project.

Students will also be exposed to interdisciplinary research on important ethical considerations about data, e.g. privacy and consent, and learn to apply this knowledge to real world datasets and technologies through assignments.

How will you learn

This course uses a blended strategy of student-led discussions, mini-lectures, and asynchronous assignments. Students will lead discussion of readings every week. In addition to attending and participating in discussions and lectures, students will be expected to contribute to Canvas discussion and complete a semester-long project that can take one of the following forms: a computational investigation of datasets, a systematic literature review, or an evidence-based redesign of an existing data-intensive system.

How will you be evaluated

Students will be evaluated on their completion of assignment and their ability to apply knowledge to several short-term assignments and a group project.

No sharing of course materials

No materials used in this class, including, but not limited to, lecture hand-outs, videos, assessments (quizzes, exams, papers, projects, homework assignments), in-class materials, review sheets, and additional problem sets, may be shared online or with anyone outside of the class without my explicit, written permission. Unauthorized sharing of materials may facilitate cheating. The University is aware of the sites used for sharing materials, and any materials found online that are associated with you, or any suspected unauthorized sharing of materials, will be reported to Student Conduct and Academic Integrity in the Office of the Dean of Students. These reports can result in initiation of the student conduct process and include charge(s) for academic misconduct, potentially resulting in sanctions, including a grade impact.

Reading list and schedule

| Week | Date | Topic | perspectives | readings |
|------|--------|---------------------------------|--------------|---|
| 1 | Jan 12 | Intro | | |
| 2 | Jan 19 | | | |
| 3 | Jan 26 | infrastructures and sovereignty | ALL | <p>11 Indigenous data sovereignty: a Māori health perspective Rawiri Jansen Indigenous Data Sovereignty: Toward an agenda, 2016, pp. 193-212 (20 pages)</p> <p>Yiwei Wu, Leah Ajmani, Nathan TeBlunthuis, and Hanlin Li. 2025. AI Didn't Start the Fire: Examining the Stack Exchange Moderator and Contributor Strike. In Proceedings of Proc. ACM Hum.-Comput. Interact. CSCW (CSCW'26)</p> |
| 4 | Feb 2 | Social and cultural data | producers | <p>Charles Chuankai Zhang, Mo Houtti, C. Estelle Smith, Ruoyan Kong, and Loren Terveen. 2022. Working for the Invisible Machines or Pumping Information into an Empty Void? An Exploration of Wikidata Contributors' Motivations . Proc. ACM Hum.-Comput. Interact. 6, CSCW1, Article 135 (April 2022), 21 pages. https://doi.org/10.1145/3512982</p> <p>Isaac L. Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at Home on the Range: Peer Production and the Urban/Rural Divide. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16). Association for Computing Machinery, New York, NY, USA, 13–25.</p> |

| | | | | |
|---|--------|-----------------------|----------------|---|
| 5 | Feb 9 | Personal data | producers | <p>Wilcox, Lauren, Robin Brewer, and Fernando Diaz. "AI Consent Futures: A Case Study on Voice Data Collection with Clinicians." Proceedings of the ACM on Human-Computer Interaction 7, no. CSCW2 (2023): 1-30.</p> <p>Understanding Account Deletion and Relevant Dark Patterns on Social Media</p> |
| 6 | Feb 18 | Crowdsourcing | producers | <p>Naja Holten Møller, Claus Bossen, Kathleen H. Pine, Trine Rask Nielsen, and Gina Neff. 2020. Who does the work of data? interactions 27, 3 (May - June 2020), 52–55.</p> <p>A Data-Driven Analysis of Workers' Earnings on Amazon Mechanical Turk</p> |
| 7 | Feb 23 | Research ethics | Intermediaries | <p>Zimmer, M. (2010). "But the data is already public": on the ethics of research in Facebook. Ethics and information technology, 12(4), 313-325.</p> <p>"Participant" Perceptions of Twitter Research Ethics. Casey Fiesler and Nicholas Proferes</p> |
| 8 | Mar 2 | midterm presentations | | |
| 9 | Mar 9 | Collection | Intermediaries | <p>Qisheng Li and Shaomei Wu. 2024. "I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data. Proc. ACM Hum.-Comput. Interact. 8, CSCW2, Article 475 (November 2024), 27 pages. https://doi.org/10.1145/3687014</p> <p>Siobhan Mackenzie Hall, Samantha Dalal, Raesetje Sefala, Foutse Yuehgoh, Aisha Alaagib, Imane Hamzaoui, Shu Ishida, Jabez Magomere, Lauren Crais, Aya Salama, and Tejumade Afonja. 2025. The Human Labour of Data Work: Capturing Cultural Diversity through World Wide Dishes. Proc. ACM Hum.-Comput. Interact. 9, 7, Article CSCW492</p> |

| | | | | |
|----|--------|--------------------------------|------------------------------|---|
| | | | | (November 2025), 43 pages. https://doi.org/10.1145/3757673 |
| 10 | Mar 16 | | | |
| 11 | Mar 23 | Documentation and availability | Intermediaries and consumers | "Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. <i>Communications of the ACM</i> , 64(12), 86-92. Consent in Crisis: The Rapid Decline of the AI Data Commons |
| 12 | Mar 30 | Data science | consumers | Analyzing User Engagement with TikTok's Short Format Video Recommendations using Data Donations Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In <i>Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21)</i> . Association for Computing Machinery, New York, NY, USA, Article 39, 1–15. Amy X. Zhang, Michael Muller, and Dakuo Wang. 2020. How do Data Science Workers Collaborate? Roles, Workflows, and Tools. <i>Proc. ACM Hum.-Comput. Interact.</i> 4, CSCW1, Article 22 (May 2020), 23 pages. https://doi.org/10.1145/3392826 |
| 13 | Apr 6 | Fairness | consumers | Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhur. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. <i>ACL 2020</i> A taxonomy of challenges in curating a fair dataset |
| 14 | Apr 13 | Sharing and deprecation | consumers | A Systematic Review of NeurIPS Dataset Management Practices Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 |

| | | | | |
|----|--------|--------------------------|--|---|
| | | | | Papers" |
| 15 | Apr 20 | Data futures and actions | | 'You are you and the app. There's nobody else.': Building Worker-Designed Data Institutions within Platform Hegemony https://metagov.org/cg-ai/ |
| 16 | Apr 27 | Final presentations | | |

Grading Summary

| | | |
|-------------------------------|---------------------------------------|------------|
| Individual assignments | Class participation | 10% |
| | Leading discussions | 20% |
| | Reading discussions | 20% |
| | Individual assignments | 10% |
| Group Assignments | Project proposal | 5% |
| | Project presentation (midterm) | 10% |
| | Project status report | 5% |
| | Project presentation (final) | 10% |
| | Final paper | 10% |

Assignments

- Weekly readings
- A short written response to **each reading** to be posted before class (200+ words). These reading responses are meant to help students reflect on what they have learned, what they disagree with the readings, and what lingering questions they have. Reading responses should be of reasonable length. Note that there are multiple mandatory readings per week, and you will be asked to submit a response per reading. Your reading responses will be graded for quality and originality. I expect you would spend the first two sentences summarizing the readings and the rest of the response would be to elaborate your reflections and questions. You can relate it to your personal experiences, your own coursework or projects, or other courses you have taken. No late work will be accepted.
- Participation in discussion
- Everyone will be a discussion leader twice. Discussion leaders should have read all the responses and prepared to organize the discussion, synthesize for reporting back to the whole class, and prepare a slide deck to facilitate discussion.
- Individual assignment - Crowdsourcing
- A final project done in groups of 1-2. This can be a computational investigation (describing a new problem you examined, or a replication of someone else's published work, a theoretical result, etc.), a systematic literature review, or an evidence-based redesign of an existing data-intensive system. The project consists of four components/deadlines
 - Proposal
 - Project presentation

- Final presentation
- Final paper

Late assignments

Late work for reading discussions and presentations will not be accepted since they are a prerequisite for coming to class. Late work for final papers will not be accepted due to grading deadlines.

For other assignments, instead of asking for extensions, you are given seven “slip day” credits for the semester; you will be charged a “slip day” for each day an assignment is late. A slip day is accrued starting immediately after the assignment is due (i.e. an assignment which is one hour late will incur a full slip day). This means you can hand in a submission up to seven days after the due date. You can also distribute your slip day credits across multiple assignments, e.g. one assignment 3 days later and the other 4 days late, without penalty. Once all of your slip day credits have been applied, any late submission will result in 5% grade deduction per day.

Use of Generative AI

Developing strong competencies in writing, communicating, brainstorming, and project development, will prepare you for success in your degree pathway and, ultimately, a competitive career. Therefore, the use of generative AI tools should be clearly documented for any writing assignment. If you decide to use generative AI tools for any submission, you should upload two versions of your submission: 1) the raw materials you uploaded to generative AI tools, and 2) the final output you produced with edits by generative AI tools. Note that spellchecking and grammar corrections do NOT fall under generative AI. Undisclosed use of generative AI tools will be treated as plagiarism.

You may use ChatGPT or similar generative AI tools to write code if your project involves coding, e.g. running a descriptive analysis of a dataset. In such a case, you must provide complete logs for any outputs you use directly and any artifacts you submit should indicate the provenance of any unedited generative AI outputs. For example, “This code was generated with the help of ChatGPT, but heavily edited.”

If you have questions about what constitutes a violation of this statement, please contact me.

Wellbeing and Safety

I urge students who are struggling for any reason and who believe that it might impact their performance in the course to reach out to me if they feel comfortable. This will allow me to provide any resources or accommodations that I can. If you are seeking mental health support, call the Counseling and Mental Health Center (CMHC) at 512-471-3515 (8a.m.-5p.m., Monday-Friday), or you may also contact Bryce Moffett, LCSW-S (iSchool CARE counselor) at 512-232-4449. For urgent mental health concerns, please contact the CMHC 24/7 Crisis Line at 512-471-2255.

Disability and Access

The university is committed to creating an accessible and inclusive learning environment consistent with university policy and federal and state law. Please let me know if you experience any barriers to learning so I can work with you to ensure you have equal opportunity to participate fully in this course. If you are a student with a disability, or think you may have a disability, and need accommodations please contact Disability & Access (D&A). Please refer to the D&A website for more information: <http://diversity.utexas.edu/disability/>. If you are already registered with D&A, please deliver your Accommodation Letter to me as early as possible in the semester so we can discuss your approved accommodations and needs in this course.