Responsible data management, Spring 2024

Class info
INF 385T
In-person. Wednesdays, 12:00PM- 3:00PM UTA 1.208

Instructor
Hanlin Li, PhD, she/her, Assistant Professor
Email: lihanlin@utexas.edu Office: UTA 5.444
Office hour: Thursdays 10:30-11:30am (Zoom) and by appointment

Academic Assistant: TBD

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 archive 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. 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 folloring forms: a computational investigation, 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

WK1: Introduction	*Complete CITI training.* *Data labor paper*	Class overview: CITI training and the Belmont report + Give an example of how to facilitate a discussion session
Wk 2: Collection	Why paying individual people for their health data is a bad idea https://www.nature.com/articles/s41591-022-01955-4 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. Additional reading: Enhancing the ethics of user-sourced online data collection and sharing https://www.nature.com/articles/s43588-023-00490-7	Lecture: Where does the data come from?
WK3: labor	*Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and*	Literature review + Pick a domain and search for relevant literature

	Collective Identities Underlying Crowdsourced Dataset Annotation. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2342–2351. 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. Additional reading: Milagros Miceli and Julian Posada. 2022. **The Data-Production Dispositif.** Proc. ACM Hum.-Comput. Interact. 6, CSCW2, Article 460 (November 2022), 37 pages. https://doi-org.ezproxy.lib.utexas.edu/10.1145/3555561	
WK4: Access	Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. Ethics and information technology, 12(4), 313-325. “Participant” Perceptions of Twitter Research Ethics. Casey Fiesler and Nicholas Proferes *Additional reading;*	Record management + Identify ethical practices in data sharing

	Freelon, D. (2018). Computational research in the post-API age. Political Communication, 35(4), 665-668. We Research Misinformation on Facebook. It Just Disabled Our Accounts Aug. 2021. New York Times. With Damon McCoy. Linux Foundation: Why Open Data Matters	
WK5: Subjectivity and biases in datasets	Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries, Olteanu et al. 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. Additional reading: Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. Patterns, 2(11).	Algorithm audits + Collect data and run your own audits
WK6: downstream impact	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	Human-centered Machine learning [Stevie Chancellor] +

	High-Stakes AI. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (CHI '21). Association for Computing Machinery, New York, NY, USA, Article 39, 1–15. Mark Diaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing Age-Related Bias in Sentiment Analysis. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 412, 1–14.	Review a model's ethical implications
Wk7: values in data work	Data Feminism - the power chapter, By Catherine D'Ignazio and Lauren F. Klein https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4 *The BigScience ROOTS Corpus: A 1.6TB Composite Multilingual Dataset* *https://arxiv.org/abs/2303.03915*	Designing data-driven products
Wk8: documentation	Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92. Bandy, J., & Vincent, N. (2021, June). Addressing" documentation debt" in machine learning: A retrospective datasheet for bookcorpus. In *Thirty-fifth Conference on Neural Information*	Create a datasheet for these datasets + Group work time

	Processing Systems Datasets and Benchmarks Track (Round 1). Boyd, Karen L. "Datasheets for datasets help ML engineers notice and understand ethical issues in training data." Proceedings of the ACM on Human-Computer Interaction 5.CSCW2 (2021): 1-27. Additional: ArtSheets for Art datasets: https://openreview.net/pdf?id=K7ke_GZ_6N	
Wk9: presentations		
Wk 10: Sharing and deprecation	Peng, K., Mathur, A., & Narayanan, A. (2021). Mitigating dataset harms requires stewardship: Lessons from 1000 papers. ArXiv, abs/2108.02922. Alexandra Sasha Luccioni, Frances Corry, Hamsini Sridharan, Mike Ananny, Jason Schultz, and Kate Crawford. 2022. A Framework for Deprecating Datasets: Standardizing Documentation, Identification, and Communication. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 199–212.	Identifying methods + Group work time
Wk11: Governance	Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., ... &	Group work time

	Hudson, M. (2020). The CARE principles for indigenous data governance. *Data Science Journal*, *19*, 43-43. Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data Governance in the Age of Large-Scale Data-Driven Language Technology. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). Association for Computing Machinery, New York, NY, USA, 2206–2222.	

Wk12: AI, LLMs, Computer vision	Liang, W., Tadesse, G.A., Ho, D. *et al.* Advances, challenges and opportunities in creating data for trustworthy AI. *Nat Mach Intell* 4, 669–677 (2022). Morgan Klaus Scheuerman, Katy Weathington, Tarun Mugunthan, Emily Denton, and Casey Fiesler. 2023. From Human to Data to Dataset: Mapping the Traceability of Human Subjects in Computer Vision Datasets. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 55 (April 2023), 33 pages.	Group work time
Wk13: Pricing and protests	J. Pei, "A Survey on Data Pricing: From Economics to Data Science," in IEEE Transactions on Knowledge and Data Engineering, vol. 34, no. 10, pp. 4586-4608, 1 Oct. 2022, doi: 10.1109/TKDE.2020.3045927. Nicholas Vincent, Hanlin Li, Nicole Tilly, Stevie Chancellor, and Brent Hecht. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 215–227.	Group work time
Wk14: final presentation		

Grading Summary

Class participation	**15%**
Leading discussions	**20%**
Reading discussions	**20%**
Project presentation	**10%**
Project proposal	**10%**
Final presentation	**10%**
Final paper	**15%**

Assignments

- Weekly readings
- A short (2-paragraph) written response to each reading to be posted by midnight the Monday before class. These responses should not summarize the reading, but instead raise questions that would be appropriate for discussion, or propose ideas to think about
- Participation in discussion
- Everyone will be a discussion leader 1-2 times over the quarter in their discussion groups. Discussion leaders should have read all the response paragraphs and prepared to organize the discussion in their group, synthesize for reporting back to the whole class, and prepare a slide deck to facilitate discussion.
- A final project done in groups of 2-5. 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
    - Project Presentation
    - Proposal
    - 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 either because of the grading deadline. If you need to submit your project proposal late, your work will be accepted, subject to 5% grade reduction per day of delay.

Use of Generative AI

All *writing* assignments (including reading discussions, project proposals, and final papers) should be fully prepared by the student. 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 to complete any aspect of writing assignments for this course are not permitted and 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 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. Bryce’s office is located in FAC18S and she holds drop in Office Hours on Wednesday from 2-3pm. 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.