Responsible data management, Spring 2024 [DRAFT]

Hanlin Li, PhD, lihanlin@utexas.edu

Short Intro
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.

Long Intro
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 design ethical 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.

Objective
Students will learn the pros and cons of different data collection, management, and sharing practices.

Students will gain hands-on experience with designing data-driven products or systems as UX designers, researchers, and data scientists.

Students will also be exposed to interdisciplinary research on important ethical considerations about data, e.g. privacy and consent.

Format
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 complete a semester-long project that takes one of the folloring forms: a computational investigation, a systematic literature review, or a evidence-based redesign of an existing data-intensive system. [More forms may be added]

Reading list [PRELIMINARY DRAFT]

WK1: Data and crowdwork	Introduction *Mark Díaz, Ian Kivlichan, Rachel Rosen, Dylan Baker, Razvan Amironesei, Vinodkumar Prabhakaran, and Emily Denton. 2022. CrowdWorkSheets: Accounting for Individual and 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.*
WK2: Labor	Yuling Sun, Xiaojuan Ma, Silvia Lindtner, and Liang He. 2023. Data Work of Frontline Care Workers: Practices, Problems, and Opportunities in the Context of Data-Driven Long-Term Care. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 42 (April 2023), 28 pages. 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.
WK3: Collection	Freelon, D. (2018). Computational research in the post-API age. Political Communication, 35(4), 665-668. Zimmer, M. (2010). “But the data is already public”: on the ethics of research in Facebook. Ethics and information technology, 12(4), 313-325.
WK4: Subjectivity and biases	Teanna Barrett, Quanze Chen, and Amy Zhang. 2023. Skin Deep: Investigating Subjectivity in Skin Tone Annotations for Computer Vision Benchmark Datasets. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1757–1771. 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.
WK5: 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 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.
Wk6: values	Data Feminism - the power chapter, By Catherine D'Ignazio and Lauren F. Klein https://data-feminism.mitpress.mit.edu/pub/vi8obxh7/release/4 Guest lecture
Wk7: 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 Processing Systems Datasets and Benchmarks Track (Round 1)*. ArtSheets for Art datasets: https://openreview.net/pdf?id=K7ke_GZ_6N
Wk8: midterm presentations	
Wk9: 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.
Wk10: Governance	Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., ... & 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.
Wk11: 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.
Wk12: Pricing	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.

	Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2021. On the Value of Wikipedia as a Gateway to the Web. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 249–260.
Wk13: protests and legal issues	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. Min, S., Gururangan, S., Wallace, E., Hajishirzi, H., Smith, N. A., & Zettlemoyer, L. (2023). SILO Language Models: Isolating Legal Risk In a Nonparametric Datastore. arXiv preprint arXiv:2308.04430.
Wk14: final presentation	

Grading Summary [Subject to change]

Attendance and participation	**10%**
Leading discussions	**20%**
Reading reflections	**20%**
Semester-long project	**50%**

Other readings:

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

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).

Hanlin Li, Nicholas Vincent, Stevie Chancellor, and Brent Hecht. 2023. The Dimensions of Data Labor: A Road Map for Researchers, Activists, and Policymakers to Empower Data Producers. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23). Association for Computing Machinery, New York, NY, USA, 1151–1161