Module 6: The Future of Data Related Work

Background

This module explores what data processing and data work looks like in the context of newer tools, like ChatGPT and other AI systems.

Additional Reading for Facilitators

- Ajmani, Leah, et al. "Data Agency Theory: A Precise Theory of Justice for AI Applications." *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, 2024, pp. 631–41, https://doi.org/10.1145/3630106.3658930.
- Cantú, Aaron. "As California AI Data Centers Grow, So Does Dirty Energy." *Capital & Main*, 21 Jan. 2025, https://capitalandmain.com/as-california-ai-data-centers-grow-so-does-dirty-energy.
- Kaisler, Stephen, et al. "Big Data: Issues and Challenges Moving Forward." *2013 46th Hawaii International Conference on System Sciences*, 2013, pp. 995–1004, https://doi.org/10.1109/HICSS.2013.645.
- Longpre, Shayne, et al. "Consent in Crisis: The Rapid Decline of the AI Data Commons." *Advances in Neural Information Processing Systems*, vol. 37, Dec. 2024, pp. 108042–87.
- Shumailov, Ilia, et al. "AI Models Collapse When Trained on Recursively Generated Data." *Nature*, vol. 631, no. 8022, July 2024, pp. 755–59, https://doi.org/10.1038/s41586-024-07566-y.
- Zhao, Dora, et al. "A Taxonomy of Challenges to Curating Fair Datasets." *arXiv:2406.06407*, arXiv, 31 Oct. 2024, https://doi.org/10.48550/arXiv.2406.06407.

Module Motivation

The goal is to present both the pros and cons of AI tools in data work. While these tools may help automate certain processes, they can also introduce significant errors and unpredictability into the data work pipeline.

Critical Considerations:

- **Data recall functionality**: Many AI systems have a kind of "recall" of data they've seen before
- **Inappropriate use cases**: AI tools are not appropriate when dealing with data we:
    - Do not have the right to share
    - Consider to be in any way sensitive or identifying of individuals

Learning Goals

Students will be able to:

1. Identify whether or not a dataset can be safely input into tools like ChatGPT
2. Identify potential pros and cons of working with AI tools in a given data context, and determine whether or not to proceed with use of those tools

Warm-Up: What Do You Know About ChatGPT?

Discussion Topics:

Check in with learners about their experience with AI tools such as ChatGPT, Claude, Gemini, etc.

Key Questions:

- **Usage experience**: How have learners used these tools?
- **Utility assessment**: What kinds of utility did they find with these tools?
- **Effectiveness**: Were the tools useful?
- **Quality control**: What kinds of corrections have they had to make to tool output?
- **Concerns and optimism**: Do they have concerns about using these tools generally, or are they optimistic about their use?

Facilitation Notes:

- Create a safe space for both positive and negative experiences
- Acknowledge varying levels of familiarity with AI tools
- Use this as an opportunity to gauge the group's baseline knowledge and attitudes

Discussion 1: From Decision Trees to ChatGPT (Slides 4-6)

Purpose:

Walk learners through the differences (but ultimate congruency) between prior work with decision trees and interactions with ChatGPT.

Key Concept:

ChatGPT has the same conceptual basis as a decision tree—it's a product of automated pattern recognition trained on a huge bank (or "corpus") of data.

Demonstration Activity: ChatGPT with "Ask a Manager" Dataset

Setup:

Use the "Ask a Manager" dataset or another dataset of your choosing for a live demonstration.

Critical Discussion Points:

Data Input Implications:

- **What it means**: Discuss putting this dataset into ChatGPT
- **Recall functionality**: ChatGPT "saves" information put into it (if indirectly)
- **Dataset status**: The dataset is open (anyone can see it)
- **Original purpose**: It was never collected with the explicit purpose of being put into ChatGPT or any other AI tool

Privacy Assessment:

At DataWorks, the discussion included:

- Personally identifying characteristics present in the dataset
- Difficulty of identifying any one individual based solely on the information presented
- Exception: Someone who already knew a contributor might be able to identify them

Important facilitation note: The goal is not to convey that using ChatGPT is necessarily correct or incorrect, but that whether it's appropriate to input data is an ongoing, open-ended discussion that should be continually revisited.

Interactive Prompt Engineering Activity:

Sample Prompts:

- "Standardize all the job titles in this list"

Discussion Questions:

- What does the outcome look like?
- Does it fit our notion of standardization?
- What are the strengths and limitations of the AI's approach?

Learner Participation:

- Learners suggest more specific prompts or alternative prompts
- Small exploration of prompt engineering techniques
- Discussion of results and effectiveness

Learning Outcomes:

- Understanding AI tools as extensions of pattern recognition concepts
- Practical experience with prompt engineering
- Critical evaluation of AI tool outputs
- Awareness of data privacy and consent issues

Reflection & Journal Entry

Assignment:

Consider all the tools available, including various LLMs and other AI, as well as those you can imagine developing in the future.

Reflection Prompts:

1. **Role evolution**: How might your professional role change with these tools?
2. **Personal response**: How do you feel about those potential changes?
3. **Future preparation**: What skills or knowledge might you need to develop?
4. **Ethical considerations**: What concerns do you have about the integration of AI in data work?
5. **Opportunities**: What positive possibilities do you see for AI in your field?

Discussion Framework:

Encourage learners to think broadly about:

- **Technical changes**: New workflows, automation possibilities
- **Professional development**: Evolving skill requirements
- **Ethical responsibilities**: Maintaining data integrity and privacy
- **Collaborative dynamics**: How AI tools might change teamwork
- **Quality assurance**: Maintaining accuracy and reliability in AI-assisted work

Learning Outcomes:

- Personal reflection on professional development needs
- Critical thinking about AI integration in data work
- Preparation for ongoing technological change
- Development of ethical framework for AI tool usage