Module 2: Documentation and the Data Lifecycles

Background

This module introduces the concept of documentation for datasets and data work through an introduction to READMEs (the standard data documentation tool at DataWorks) as well as mechanisms for standardization and visualization to reflect and audit completed work.

Session Structure:

- Review of prior week's journal entry (with time for learners to review their work thus far)
- Open-ended discussion of what "good" data documentation is and what it looks like
- Practice analyzing and developing appropriate dataset documentation
- Formal introduction of READMEs and data lifecycle concepts
- Discussion of how data flows through the organization (tailored to DataWorks—adapt for your organization)

Additional Reading for Facilitators

- Bandy, Jack, and Nicholas Vincent. "Addressing 'Documentation Debt' in Machine Learning Research: A Retrospective Datasheet for BookCorpus." *arXiv:2105.05241*, arXiv, 11 May 2021, https://doi.org/10.48550/arXiv.2105.05241.
- Geiger, R. Stuart, et al. "The Types, Roles, and Practices of Documentation in Data Analytics Open Source Software Libraries." *Computer Supported Cooperative Work (CSCW)*, vol. 27, no. 3, Dec. 2018, pp. 767–802, https://doi.org/10.1007/s10606-018-9333-1.
- Heger, Amy K., et al. "Understanding Machine Learning Practitioners' Data Documentation Perceptions, Needs, Challenges, and Desiderata." *Proc. ACM Hum.-Comput. Interact.*, vol. 6, no. CSCW2, Nov. 2022, p. 340:1-340:29, https://doi.org/10.1145/3555760.
- Trace, Ciaran B., and James A. Hodges. "The Role of Paradata in Algorithmic Accountability." *Perspectives on Paradata: Research and Practice of Documenting*

Process Knowledge, edited by Isto Huvila et al., Springer International Publishing, 2024, pp. 197–213, https://doi.org/10.1007/978-3-031-53946-6_11.
- Wang, April Yi, et al. "Documentation Matters: Human-Centered AI System to Assist Data Science Code Documentation in Computational Notebooks." *ACM Transactions on Computer-Human Interaction*, vol. 29, no. 2, Apr. 2022, pp. 1–33, https://doi.org/10.1145/3489465.

Module Motivation

At DataWorks, ensuring everyone is on the same page about data documentation helps the team become more organized and unified in practices. Data documentation is crucial for critical data practices—if you aren't sure how a dataset was developed or how work was completed on it, it's much harder to make sense of the stories the dataset is trying to tell.

Learning Goals

Students will be able to:

1. Identify key parts of a README and create their own for a dataset they've worked on
2. Analyze what standardization means in the context of a particular dataset and execute the appropriate steps to standardize that dataset
3. Understand their role in a data lifecycle within a particular organization and the responsibilities that come with that placement

Discussion 1: Introducing Good Documentation (Slide 3)

Framework: The Six W's and H

Note: This will look different for every organization based on the kinds of data being worked on.

At DataWorks, documentation focuses on **"who, what, where, when, why, and how"**:

Who

- Who requested the data work?

What

- What is the scope of the data work to be completed?

Where

- Where was the work done?

When

- When was it done?
- Range from project tracking of individual hours to large-scale representation of projects moving through DataWorks

Why

- Why was it done?
- Why did we take on the project?
- What perspective do we—or the Data Fellows—bring to the work?

How (Most Critical)

- **Software systems used**: Tools, desktop environments (important for potential downstream incompatibility)
- **Major data moves and key transformations**: How addresses were changed, what APIs or lookups were used

Facilitation Notes:

- Adapt this framework to your organization's specific data work
- Emphasize the critical importance of documenting the "how"
- Encourage discussion about organization-specific documentation needs