Supporting Responsible Machine Learning by Improving Data Curation

Eshta Bhardwaj & Christoph Becker
University of Toronto

INTRODUCTION

Prior Work

Bias

- Bias in models can cause discriminatory or unethical judgments
- Biases are attributed to choices made about training datasets

Inappropriate data reuse

- Datasets are often reused outside their original context
- Data work is hidden, tacit, and undervalued which hinders appropriate data reuse

Current Work

- Some studies have started to look at the adoption of principles from archival studies and digital curation into ML

Our Goal

- ML research is currently only looking at the adoption of these concepts in theory, given the challenges in their translation when applied
- We establish how ML dataset development processes can apply data curation in practice

RESEARCH QUESTIONS & METHODS

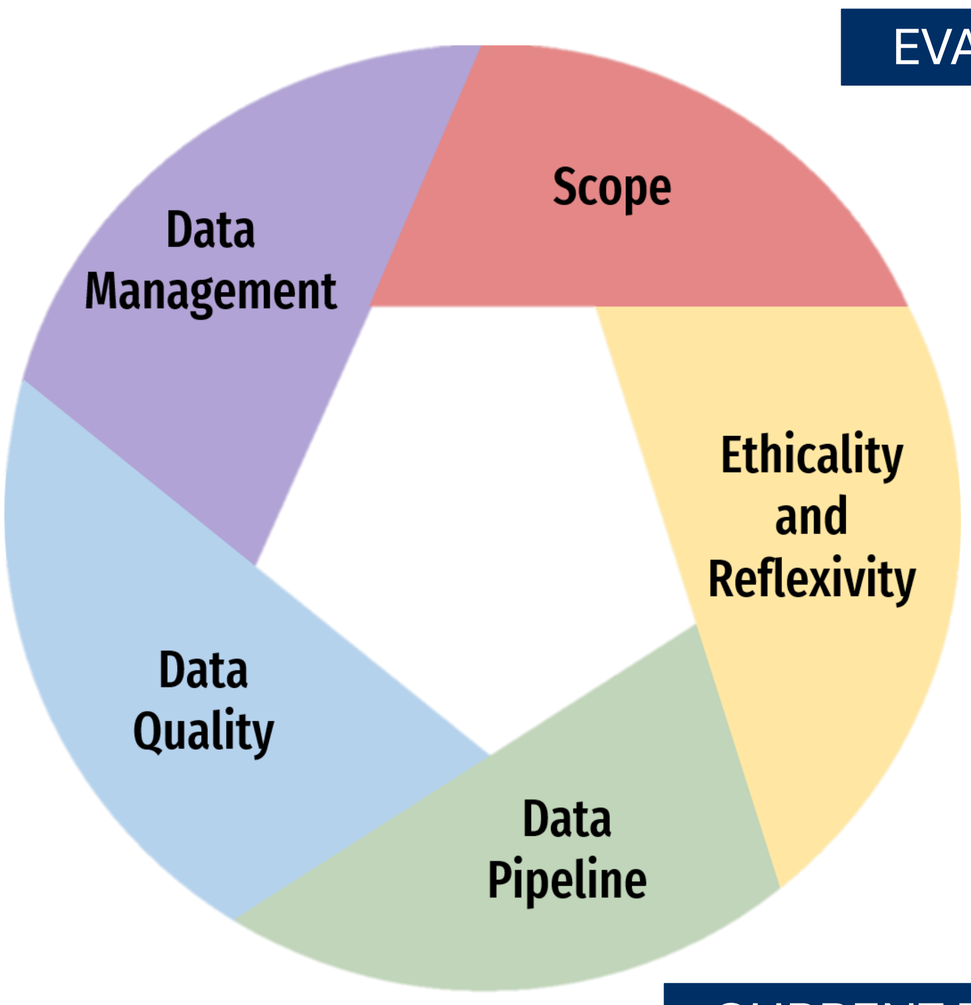
What constitutes a well curated dataset?

- Developed an **evaluation framework** made up of rubric and toolkit
- **Rubric** evaluates dataset contents and dataset design decisions
- **Toolkit** provides application guidance for the rubric

What is the state of data curation at NeurIPS?

- Assessed datasets to **evaluate current practices of data curation** in ML dataset development
- Analyzed areas in which **improvement** was needed

EVALUATION FRAMEWORK



- Context, purpose, motivation
  Requirements
- Ethicality
  Domain knowledge & data practices
  Context awareness
  Environmental footprint
- Data collection
  Data processing
  Data annotation
- Suitability
  Representativeness
  Authenticity
  Reliability
  Structured documentation
- Findability
  Accessibility
  Interoperability
  Reusability

Context awareness

Context awareness demonstrates an understanding of the subjective, non-neutral nature, and situatedness of data.

Criteria to meet minimum standard

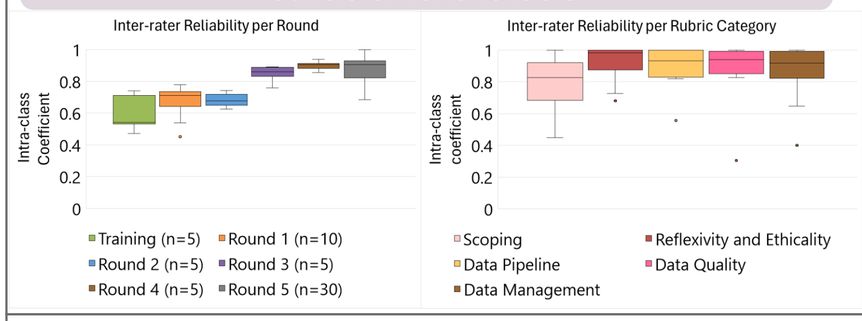
Documentation includes a positionality statement.

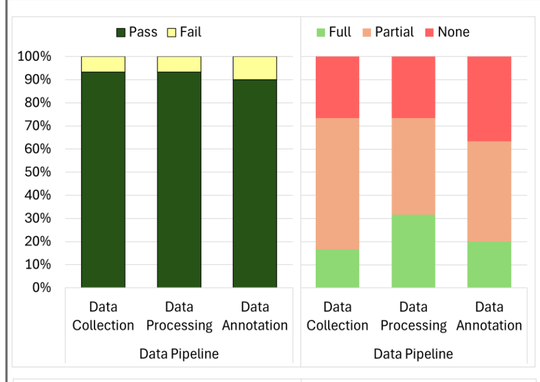
Criteria to meet standard of excellence

Documentation adopts a reflexive approach to dataset development. For example, documentation discusses how field epistemologies impact assumptions, methods, or framings.

CURRENT PRACTICES OF DATA CURATION

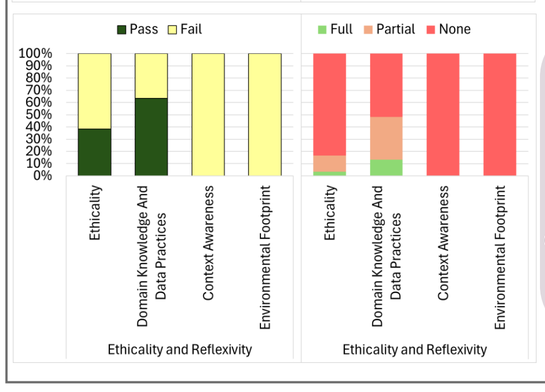
Finding 1 Inter-rater reliability suggests the evaluations are consistent and reliable





Finding 2

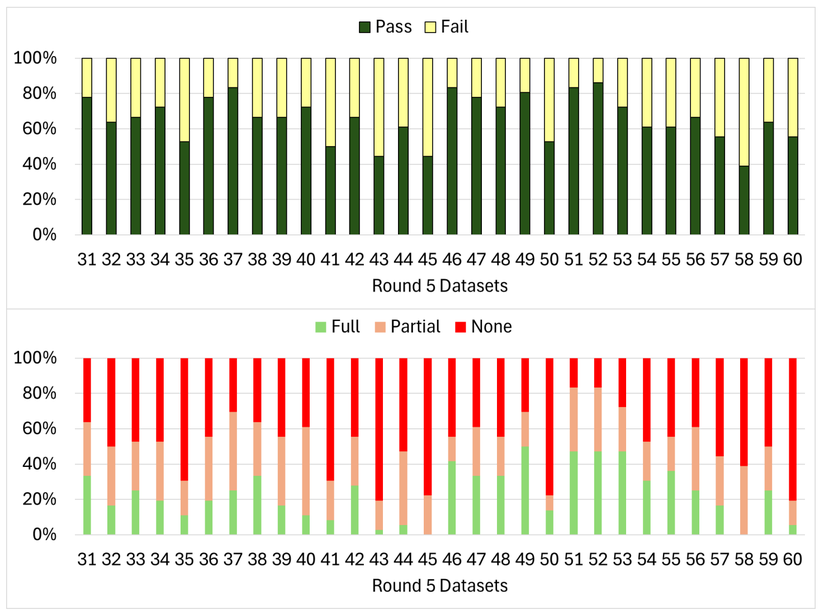
NeurIPS prioritizes model-work adjacent documentation



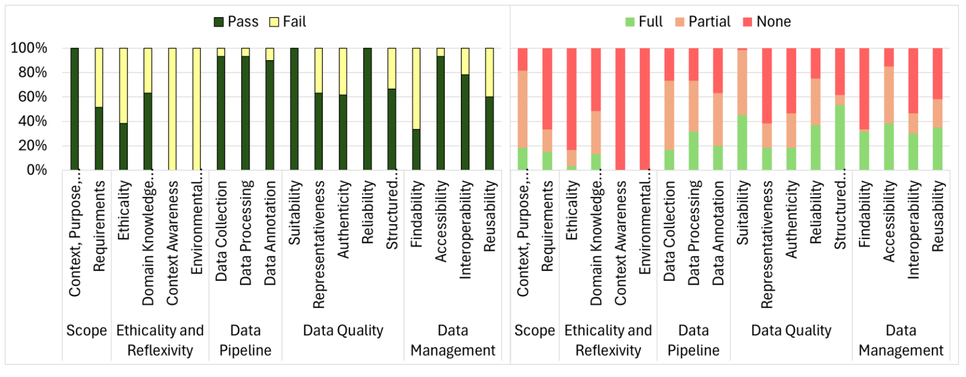
Finding 3

Documentation is rarely context-aware and typically does not quantify environmental footprint

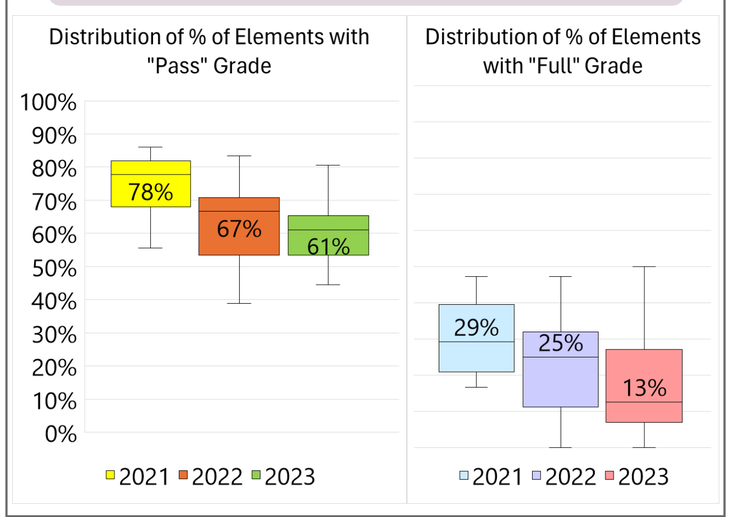
Finding 4 Documentation quality varies widely across datasets



Finding 5 Documentation often remains incomplete



Finding 6 Findings suggest no improvements occurred over time



KEY TAKEAWAY

The creation of the D&B track shows that **dataset quality is the foundation of continued progress in ML** applications. There is no better database of knowledge than data curation to aid in this venture.

Our evaluation framework provides a practical lens on how NeurIPS can spearhead the requirement for **rigorous data curation in ML**.

STRATEGIES TO IMPROVE DATA CURATION IN ML

Requirements

- Create **purpose statements**
- Document **initial formulation** vs. the dataset creation scheme

Ethicality

- Consider **proportionality principle**

Context awareness

- Include **positionality statements** to increase reflexivity

Environmental footprint

- **Quantify** the environmental footprint of datasets

Findability

- Assign **persistent identifiers** to metadata to avoid link rot

Reusability

- Include **identifier** information, dataset **characteristics**, and dataset **provenance**

UNIVERSITY OF TORONTO
FACULTY OF INFORMATION
Digital Curation Institute

Women in Machine Learning

NEURAL INFORMATION PROCESSING SYSTEMS

Just Sustainability Design Lab