UNIVERSITY OF TORONTO
FACULTY OF INFORMATION
Digital Curation Institute
Just Sustainability Design Lab

The State of Data Curation at NeurIPS: An Assessment of Dataset Development Practices in the Datasets & Benchmarks Track

Eshta Bhardwaj, Harshit Gujral, Siyi Wu, Ciara Zogheib, Tegan Maharaj, & Christoph Becker

NEURAL INFORMATION PROCESSING SYSTEMS

INTRODUCTION

Introduction

- NeurIPS has responded to the rising urgency and recognized impact of data research through the introduction of the **D&B track**
- This track aims to address the issue of datasets being used **outside their original scope**

Background

- Data curation involves "maintaining and adding value to digital research data for current and future use"
- **Field** of data curation has **established methods and discourse** on how to maintain large amounts of data and manage ethical concerns

Motivation

- ML research has turned towards the **improvement of data to improve model results** and fundamental understanding
- Current **gap in recognition and uptake** of data curation concepts in the ML community

Our Goal

- Document and improve the standard of **dataset development in NeurIPS** so that future benchmarks and datasets can be effectively found, easily accessed, ethically used, consistently evaluated, and appropriately reused

RESEARCH QUESTIONS & METHODS

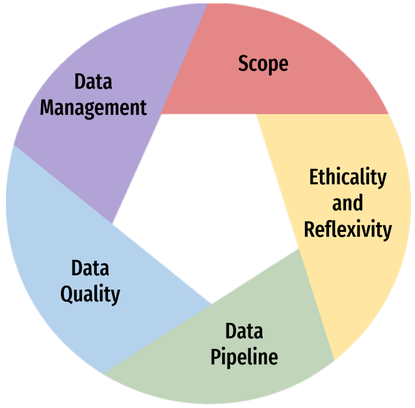
What constitutes a well curated dataset?

- Developed an **evaluation framework** made up 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?

- Applied framework to **evaluate** NeurIPS datasets
- Examined the consistency in application by measuring **inter-rater reliability** (IRR)
- 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.

STRATEGIES TO IMPROVE DATA CURATION IN ML

Requirements

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

Ethicality

- Consider **proportionality principle**
- Reflect on whether and how benefits outweigh the harms

Context awareness

- Include **positionality statements** to increase reflexivity on how identity impacts data-related choices

Environmental Footprint

- **Quantify** the environmental footprint of datasets to improve transparent reporting of resource consumption

Findability

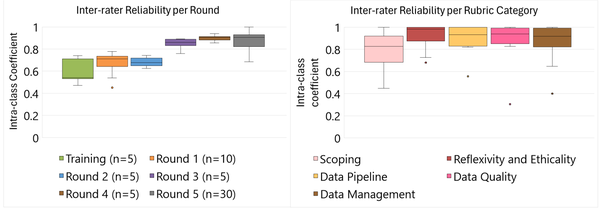
- Assign **persistent identifiers** to metadata and host in a searchable repository

Reusability

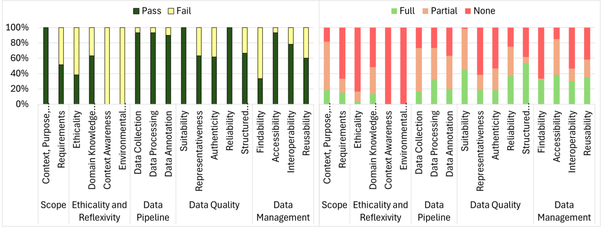
- Include **identifier** information, dataset **characteristics**, and dataset **provenance**

CURRENT PRACTICES OF DATA CURATION

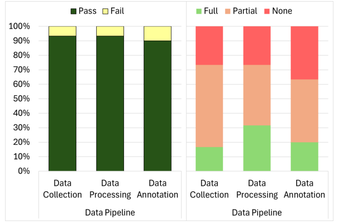
Finding 1 Inter-rater reliability suggests the evaluations are consistent and reliable



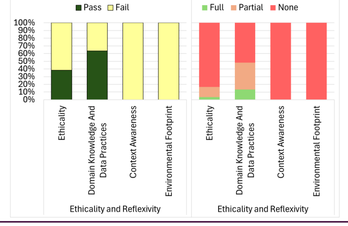
Finding 2 Documentation often remains incomplete



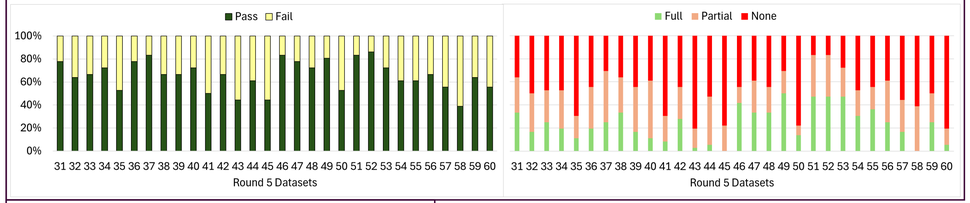
Finding 3 NeurIPS prioritizes model-work adjacent documentation



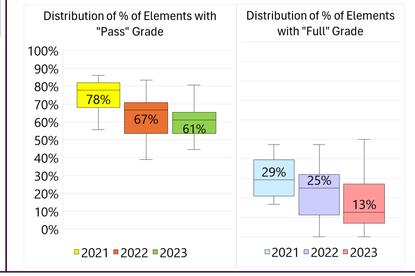
Finding 4 Documentation is rarely context-aware and typically does not quantify environmental footprint



Finding 5 Documentation quality varies widely across datasets



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**.