

Metadata for Trustworthy AI

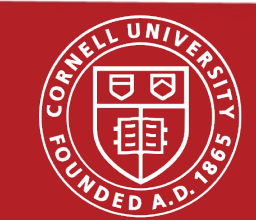
Werner Sun

Cornell High Energy Synchrotron Source (CHESS)

Cornell Laboratory for Accelerator-based Sciences and Education (CLASSE)

CI Compass Workshop: AI Meets CI

13 Jan 2026



Cornell University

CHESS
CORNELL HIGH ENERGY
SYNCHROTRON SOURCE

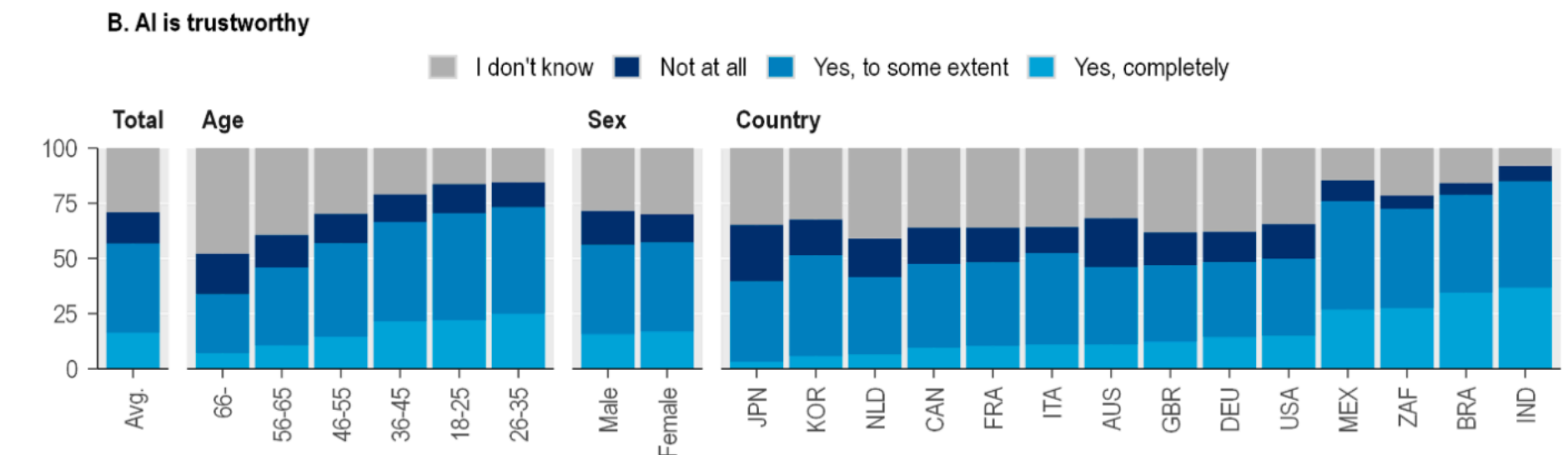


Can We Trust AI?

- AI in science: healthcare, autonomous experiments, chatbots for data analysis

- One barrier to trust: training datasets

- Bias in training datasets → *biased outcomes*
- Recursive training on AI-generated content → *model collapse*



OECD WISE Centre survey, December 2025

- Metadata provides context & provenance that can reduce bias

- Are datasets real or synthetic? From trusted sources? Representative of real world, including long tails?
- AI-ready datasets should include metadata for transparency, necessary (but not sufficient) for trustworthy AI

- CHES strategy: FAIR Open-Science eXtensible Data Exchange Network

- Data services for recording metadata, code, and other research artifacts in real time
- <https://chesscomputing.github.io/FOXDEN/>

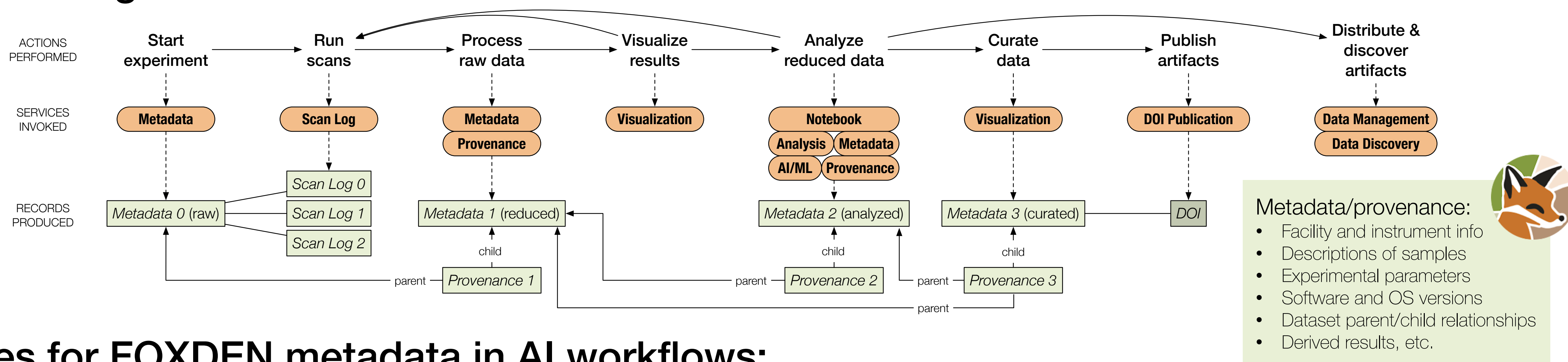
FOXDEN



Generating AI-Ready Datasets with FOXDEN

Contact me: wms8@cornell.edu

- FOXDEN documents research workflows, preserves JSON metadata that goes with data using scientist-defined schemas



- **Uses for FOXDEN metadata in AI workflows:**
 - Curating training datasets, auditing data sources, providing labels for supervised learning or validating GenAI outputs
 - Include metadata directly in model training or in RAG context window (to augment training data itself)
- **In progress: integrate with operations @ CHESS (and MagLab), RAG-based chatbot**
 - Unresolved: how to ensure accuracy & reliability of metadata (governance structure for security and accountability)
- **Summary: FOXDEN helps to build trust in AI**