

AI is Reducing to Data Curation

Bill Howe

UNIVERSITY OF WASHINGTON, USA

Abstract

AI advances have been concentrated in curation-on-read settings: LLMs are trained on massive, weakly curated convenience samples of the internet, while the output is assumed to be subject to careful human review and accountability on a per-instance basis. This regime shifts all responsibility to the end user to identify errors, biases, and compliance issues (e.g., intellectual property violations). There do exist successful AI applications in curation-on-write settings, typically in science: models trained on precise, objectively correct input in order to produce precise, objectively correct output. For example, popular deep learning architectures are poised to outperform physics-based models to predict the weather, simultaneously learning the physics and the parameters directly from observations, without requiring fluid dynamics to be explicitly programmed. We are studying enabling technologies to reduce the cost of developing AI systems in these curation-on-write settings, characterized by limited data, complex multi-modal features, and ambiguous or conflicting labels. As methods continue to be commoditized, costs are driven by finding and organizing training and evaluation data. This perspective of “curation as programming” is an opportunity to design new tooling to empower domain experts to build AI systems for specialized tasks. I will describe some projects my group is pursuing in this space including recovering missing data in urban mobility, identifying speakers and agenda items in local council meetings, and extracting information from legal documents. I will describe the technical questions we encounter in these settings, including how best to use expert-provided ontologies and unlearning specific biases during fine-tuning.


AMW'23: 15th Alberto Mendelzon International Workshop on Foundations of Data Management, May 22–26, 2023, Santiago, Chile

 billhowe@uw.edu (B. Howe)

 <https://faculty.washington.edu/billhowe/> (B. Howe)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)