Data Provenance for Transparent and Understandable AI Systems

Abstract

"Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness."[1]

Even though provenance has gained much traction on processes like supply chain management, scientific experimentation, or complex data processing, provenance is still barely used to improve the training data quality of an AI system.

We propose an application of data provenance for the tracing of input data of an AI system. This work examines how data provenance can help make an AI system more transparent and understandable.

Problem

Through large-scale scraping of the Internet, data augmentation techniques and other data retrieval methods, there is abundant data to train AI models. However, the quality of this data is often neglected. Not knowing the origin and processing steps of the data used to train an AI system can have disastrous consequences, especially in critical applications such as medicine. For example, one question is whether qualified workers (doctors, etc.) were responsible for creating the classification labels for medical images, or unqualified ones?

Goal

The goal of this research proposal is to find and evaluate an approach on how to store meta-data about the origin and manipulation of data that will be used for training AI models. Some of the benefits of knowing the origin and processing of training data of an AI system would be:

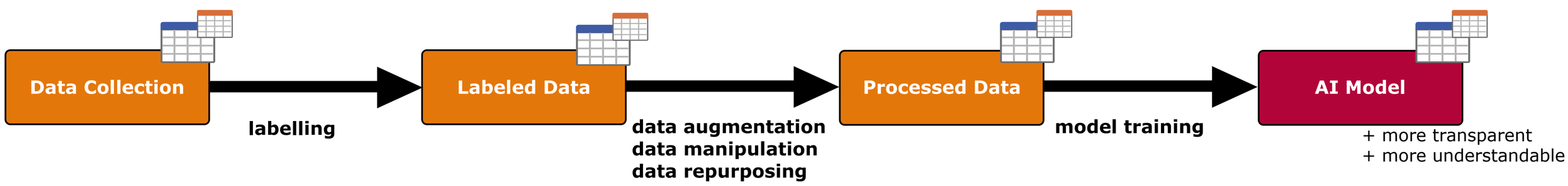
Transparency. For example, you could improve the transparency of an AI model by knowing who created the labels for a dataset or you could release what training data was used for a recommendation system.

Understandability. Provenance could be used to explain a model's output and reduce biases. For example, knowing that an AI model for face recognition has been trained exclusively with images of white males can explain why it performs poorly with other subjects.

Reproducibility. Provenance could also help with the reproducibility of AI experiments, for which it is important to know, for example, how the data was processed, whether pseudo-labels were used, or whether augmentation occurred.

Solution/Approach

The following approach is based on eager computation where we store the origin information as well as any modifications of our data in so called "bread crumbs". With this approach, companies that are training AI models have to store meta-data for every step, as shown in the diagram below. They must also preserve provenance information for data downloaded from the Internet or other data sources to ensure consistent data provenance.



Johannes Hagemann

Master student IT-Systems Engineering
Hasso Plattner Institute, Potsdam, Germany

E-Mail: johannes.hagemann@student.hpi.de

References

[1] https://www.w3.org/TR/prov-dm/

Connection to the Lecture

Professor Melanie Herschel' s talk "Data Provenance" discussed the various benefits of Provenance. In this proposed use case for data provenance, multiple of her stated benefits of data provenance are covered such as the transparency and understandability aspect.

Furthermore, the proposed solution of using data provenance to improve the transparency and understandability of an AI model is based on an adapted approach of the eager computation algorithm presented in her talk.

HPI Hasso Plattner Institut