Human AI Interaction

Lecture 13: Thinking about data

aidesignclass.org

Thinking about data

- A few different terms to think about data
- How does data actually look?
- What problems does it have and how do these matter?
- What can we do about it?

A few terms to know

- Supervised data: <input> -> <output>
  - E.g. Instruction -> Results of instruction
  - “Write a poem” -> “<poem>”
- Unsupervised data: only <input>
  - E.g. a collection of images, webpages etc (but see semisupervised)
- Semi-supervised data: <input> -> <output mechanically found from input>
  - E.g. fill in the blank (covering consecutive words)
  - Eg. breaking a sentence “X because Y” to “Why X?” -> “Because Y” (input -> output)
- Synthetic data: artificially created data, to serve a particular purpose

How much data do we have?

- Unsupervised data is “free” – you can find it on the internet.
  - A huge amount of unsupervised data
- Semi-supervised data is “almost free”:
  - Large amount of semi-supervised data, based on how we mechanically translate input to out
- Supervised data is expensive, created often by hand:
  - Small amounts of supervised data

Data “quality”

Let’s look at some data

- Unsupervised: https://huggingface.co/datasets/c4/viewer/en/train

What are the main things you observe?

- How long is each data row?
- What kinds of topics?
- How “good” is the text?

C4 dataset is used in almost all major LLMs today

Data “quality”

Let’s look at some data

- Supervised: https://huggingface.co/datasets/openai/summarize_from_feedback/viewer/axis/validation?row=17

What are the main things you observe?

- How long is each data row?
- What kinds of topics?
- How “good” is the data?

Data “quality”

Let’s look at some data

- Supervised: https://huggingface.co/datasets/gsm8k?row=2

What are the main things you observe?

- How long is each data row?
- What kinds of topics?
- How “good” is the data?

Data “quality”

Let’s look at some data

- Synthetic: https://huggingface.co/datasets/SirNeural/flan_v2/viewer/default/train?row=4

What are the main things you observe?

- How long is each data row?
- What kinds of topics?
- How “good” is the data?

Data quality – what do the errors mean?

- What are the sources of errors?
  - For human created data
  - For semi-supervised data
  - For synthetic data
- What do you do in case of errors?

Data quality – what do the errors mean?

- What are the sources of errors?
  - For human created data: unclear instructions, task is naturally variable, humans did not put in effort,...
  - For semi-supervised data: source of data is not very high-quality, transformations are not high quality
  - For synthetic data: model generating data is not good, instructions are not good
- What do you do in case of errors?

“Humans did not put in enough effort”

“Three employees told TIME they were expected to read and label between 150 and 250 passages of text per nine-hour shift. Those snippets could range from around 100 words to well over 1,000. All of the four employees interviewed by TIME described being mentally scarred by the work. Although they were entitled to attend sessions with “wellness” counselors, all four said these sessions were unhelpful and rare due to high demands to be more productive at work.”

BUSINESS • TECHNOLOGY

Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic

This image was generated by OpenAI's image-generation software, Dall-E 2. The prompt was: "A seemingly endless view of African workers at desks in front of computer screens in a printmaking style." TIME does not typically use AI-generated art to illustrate its stories, but chose to in this instance in order to draw attention to the power of OpenAI's technology and shed light on the labor that makes it possible. Image generated by Dall-E 2/OpenAI

What can we do to improve data practices?

1. Reduce the need for labeled/supervised data; rely on semi-supervised data instead

Benefit: you only need a little bit of very good labeled data

Challenge: very hard to improve quality by adding more good labeled data

LIMA: Less Is More for Alignment

Chunting Zhou[μ*] Pengfei Liu[π*] Puxin Xu[μ] Srini Iyer[μ] Jiao Sun[λ]
Yuning Mao[μ] Xuezhe Ma[λ] Avia Efrat[τ] Ping Yu[μ] Lili Yu[μ] Susan Zhang[μ]
Gargi Ghosh[μ] Mike Lewis[μ] Luke Zettlemoyer[μ] Omer Levy[μ]

[μ] Meta AI
[π] Carnegie Mellon University
[λ] University of Southern California
[τ] Tel Aviv University

Abstract

Large language models are trained in two stages: (1) unsupervised pretraining from raw text, to learn general-purpose representations, and (2) large scale instruction tuning and reinforcement learning, to better align to end tasks and user preferences. We measure the relative importance of these two stages by training LIMA, a 65B parameter LLaMa language model fine-tuned with the standard supervised loss on only 1,000 carefully curated prompts and responses, without any reinforcement learning or human preference modeling. LIMA demonstrates remarkably strong performance, learning to follow specific response formats from only a handful of examples in the training data, including complex queries that range from planning trip itineraries to speculating about alternate history. Moreover, the model tends to generalize well to unseen tasks that did not appear in the training data. In a

What can we do to improve data practices?

2. Question the assumption of <input> -> <output>

Multiple outputs may often be correct! (And it depends on the human labeling too!)

Challenge: hard to decide how many different perspectives to include

Jury Learning: Integrating Dissenting Voices into Machine Learning Models

Mitchell L. Gordon
Stanford University
Stanford, USA
mgord@cs.stanford.edu

Michelle S. Lam
Stanford University
Stanford, USA
mlam4@stanford.edu

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Kayur Patel
Apple Inc.
Seattle, USA
kayur@apple.com

Jeffrey T. Hancock
Stanford University
Stanford, USA
hancockj@stanford.edu

Tatsunori Hashimoto
Stanford University
Stanford, USA
tatsu@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

arXiv:2202.02950v1 [cs.HC] 7 Feb 2022

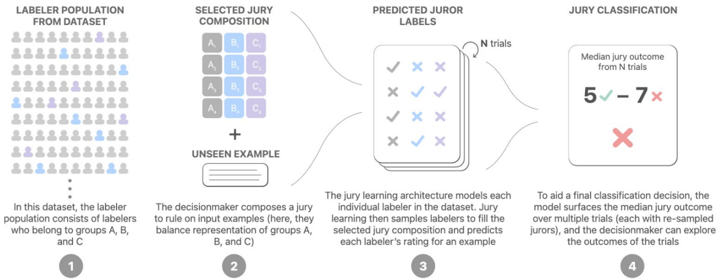

Figure 1: An overview of jury learning. (1) Given a dataset annotated by labelers from different groups, (2) the machine learning practitioner can compose a jury to rule on an unseen input example by allocating seats to labelers from the dataset with specified characteristics. (3) Then, the jury learning architecture models each individual labeler in the dataset, and performs *N* trials in which it samples labelers as jurors to populate the specified jury composition and predicts each juror's decision for the example. (4) The system then outputs a median-of-means jury outcome alongside jury outcome exploration visualizations that the decisionmaker can use to reach a classification decision.

What can we do to improve data practices?

3. Turn data collection into a consultative, conversational process

Benefit: Turns the process of collecting data into a more scientific hypothesis-driven process. (“Can we collect this?”, “what if we did something else?”)

Challenge: Needs a broader change to data collection processes, change in power structures.

From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work

CINDY KAIYING LIN, Pennsylvania State University, USA
STEVEN J. JACKSON, Cornell University, USA

Managing error has become an increasingly central and contested arena within data science work. While recent scholarship in artificial intelligence and machine learning has focused on limiting and eliminating error, practitioners have long used error as a site of collaboration and learning vis-à-vis labelers, domain experts, and the worlds data scientists seek to model and understand. Drawing from work in CSCW, STS, HCML, and repair studies, as well as from multi-sited ethnographic fieldwork within a government institution and a non-profit organization, we move beyond the notion of error as an edge case or anomaly to make three basic arguments. First, error discloses or calls to attention existing structures of collaboration unseen or underappreciated under ‘working’ systems. Second, error calls into being new forms and sites of collaboration (including, sometimes, new actors). Third, error redeploys old sites and actors in new ways, often through restructuring relations of hierarchy and expertise which recenter or devalue the position of different actors. We conclude by discussing how an artful living with error can better support the creative strategies of negotiation and adjustment which data scientists and their collaborators engage in when faced with disruption, breakdown, and friction in their work.

CCS Concepts: • **Human-centered computing → Collaborative and social computing → Collaborative and social computing design and evaluation methods → Ethnographic studies**

Additional Key Words and Phrases: Error; Data Science; Machine Learning; Critical Data Studies; Repair; AI ethics

ACM Reference format:
Cindy Kaiying Lin and Steven J. Jackson. 2023. From Bias to Repair: Error as a Site of Collaboration and Negotiation in Applied Data Science Work. *Proc. ACM Hum.-Comput. Interact*, 7, CSCW1, Article 131 (April 2023), 32 pages, https://doi.org/10.1145/3579607.