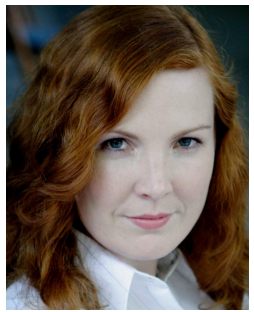
Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research









Premissas

1. Data plays a critical role in machine learning.

2. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets.

3. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases.

Problem

Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

Solutions

“every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on. (...) Increase transparency and accountability within the machine learning community, mitigate unwanted societal biases in machine learning models, facilitate greater reproducibility of machine learning results, and help researchers and practitioners to select more appropriate datasets for their chosen tasks”


who?

datasets creators	dataset consumers
encourage careful reflection on the process of creating, distributing, and maintaining a dataset, including any underlying assumptions, potential risks or harms, and implications of use.	ensure they have the information they need to make informed decisions about using a dataset

Transparency

greater reproducibility

+

create alternative datasets

Development Process

our experiences as researchers with diverse backgrounds

AND

team of lawyers

Questions and Workflow

1. motivation
2. composition
3. collection process
4. preprocessing/cleaning/labeling
5. uses
6. distribution
7. maintenance

1. motivation

For what purpose was the dataset created?

Who funded the creation of the dataset?

Who created the dataset?

Any other comments?

2. composition

What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?

- How many instances are there in total (of each type, if appropriate)?

- What data does each instance consist of?

- Is there a label or target associated with each instance?

- Are relationships between individual instances made explicit?

2. composition

- Are there any errors, sources of noise, or redundancies in the dataset?
- Does the dataset contain data that might be considered confidential?
- Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
- Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly?

3. collection process

- How was the data associated with each instance acquired? (*remember: readymade x handmade -Salganik*)
- What mechanisms or procedures were used to collect the data?
- Who was involved in the data collection process and how were they compensated?
- Over what timeframe was the data collected?
- Were any ethical review processes conducted?
- Has an analysis of the potential impact of the dataset and its use on data subjects been conducted?

4. preprocessing/cleaning/labeling

- Was any preprocessing/cleaning/labeling of the data done?
- Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?
- Is the software that was used to preprocess/clean/label the data available?

5. uses

- Has the dataset been used for any tasks already?
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
- Are there tasks for which the dataset should not be used?

6. distribution

- Has the dataset been used for any tasks already?
- Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?
- Are there tasks for which the dataset should not be used?

7. maintenance

- Who will be supporting/hosting/maintaining the dataset?
- How can the owner/curator/manager of the dataset be contacted?
- Is there an erratum?
- Will the dataset be updated?
- If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?

Impact and Challenges

1. Datasheets for datasets do not provide a complete solution to mitigating unwanted societal biases or potential risks or harms.

2. Dataset creators cannot anticipate every possible use of a dataset

4. When creating datasets that relate to people, and hence their accompanying datasheets, it may be necessary for dataset creators to work with experts in other domains such as anthropology, sociology, and science and technology studies. There are complex and contextual social, historical, and geographical factors that influence how best to collect data from individuals in a manner that is respectful.

4. overhead on dataset creators: time and money


Thank
You