Datasheets for Datasets

Presented By: Pritish Mishra

Source: Gebru, Timnit, et al. "Datasheets for datasets." arXiv preprint arXiv:1803.09010 (2018)

01

Motivation

BBC | Sign in | Home | News | Sport | Reel | Worklife | Travel

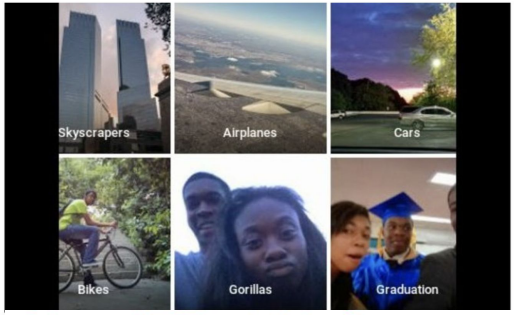
NEWS

Home | Coronavirus | Climate | Video | World | Asia | UK | Business | Tech | Science | Stories

Tech

Google apologises for Photos app's racist blunder

1 July 2015

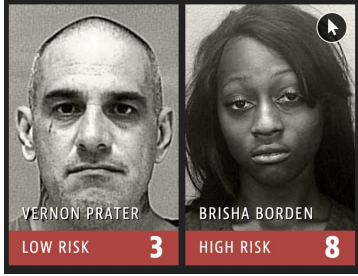


Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

Two Petty Theft Arrests



REUTERS

World | Business | Markets | Breakingviews | Video | More

RETAIL OCTOBER 11, 2018 / 4:34 AM / UPDATED 3 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

By Jeffrey Dastin

8 MIN READ

Conferences > 2021 IEEE International Confe...

Skin tone, Confidence, and Data Quality of Heart Rate Sensing in WearOS Smartwatches

Publisher: IEEE | Cite This | PDF

Ishita Ray ; Daniyal Liaqat ; Moshe Gabel ; Eyal de Lara All Authors

"Inaccurate heart rate readings occur more frequently in dark skin as compared to light skin. Significantly fewer data points for people with darker skin tones, can bias downstream applications."

Uninformed choices can lead to mistakes

A model is unlikely to perform well in the wild if:

- Deployment context does not match training or evaluation dataset
- Datasets reflect unwanted biases
- No account of underlying assumptions made during dataset creation
- Lack of transparency and accountability

02
Proposed Solution

Datasheets for Datasets

Every dataset should be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on.

Stakeholders

Dataset Creators

- Careful reflection on the process of creating, distributing, and maintaining a dataset.
- Declaring underlying assumptions, potential risks, and implications of use.

Dataset Consumers

- Getting sufficient information to make informed decisions about using a dataset.
- Selecting appropriate datasets for the tasks.
- Avoiding unintentional misuse.

Datasheet Workflow

01 **Motivation**
Why a dataset was created?

02 **Composition**
What's in a dataset?

03 **Collection Process**
How was the data collected?

04 **Preprocessing**
How was data pre-processed and cleaned? (if applicable)

05 **Uses**
How the dataset should and should not be used?

06 **Distribution**
How the dataset will be shared?

07 **Maintenance**
Who supports, hosts, or maintains the dataset?

Datasheet Workflow

01. Motivation

Why a dataset was created?

02. Composition

What's in a dataset?

03. Collection Process

How was the data collected?

04. Pre-processing

How was data pre-processed and cleaned? (if applicable)

05. Uses

How the dataset should and should not be used?

06. Distribution

How the dataset will be shared?

07. Maintenance

Who supports, hosts, or maintains the dataset?

03

Case Study

Example Dataset

Labeled faces in the wild:

- Contains labeled face photographs spanning the range of conditions typically encountered in everyday life.
- Exhibits “natural” variability in factors such as pose, lighting, race, accessories, occlusions, and background.
- Created in contrast to controlled environments to estimate better real-world performance.

Source: Huang, Gary B., et al. "Labeled faces in the wild: A database for studying face recognition in unconstrained environments." Workshop on faces in 'Real-Life' Images: detection, alignment, and recognition. 2008.

Composition of Dataset

What do the instances that comprise the dataset represent?
Each instance is a pair of images labeled with the name of the person in the image.

Are relationships between individual instances made explicit?
No known relationships between instances except that they are all individuals who appeared in news sources online.

Are there recommended data splits (e.g., training, development/validation, testing)?
Extensive details including recommended data splits, validation methods and training paradigms.

Better Understanding

Composition - Ethical Considerations

Does the dataset contain data that might be considered confidential?
No. All data was derived from publicly available news sources.

Ethical Reflection

Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
No. The dataset only consists of faces and associated names.

Does the dataset identify any subpopulations (e.g., by age, gender)?
Provides distribution of images by age, race and gender.

Fairness

Collection Process

How was the data associated with each instance acquired?
Manually - names in the dataset determined by an operator by looking at the caption associated with the person's photograph

What mechanisms were used to collect the data?
Images in this database were gathered from the web using software to crawl news articles.

If the dataset is a sample from a larger set, what was the sampling strategy?
Sampled from original Faces in the Wild database. No specific sampling strategy, many groups have few instances (e.g. only 1.57% of dataset consists of individuals under 20 years)

Better Understanding

Reproducibility

Fairness

Collection - Ethical/Legal Considerations

Were any ethical review processes conducted?
Unknown

Did the individuals in question consent to the collection and use of their data?
No. All subjects in the dataset appeared in news sources so the images that we used along with the captions are already public.

Were the individuals in question notified about the data collection?
Unknown

Ethical Reflection
Legal Declaration

Preprocessing

Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data?
The raw unprocessed data is saved.

Is the software used to preprocess/clean/label the instances available?
All software used to process the data is open source and has been specified.

Reproducibility

Uses

Has the dataset been used for any tasks already?
Papers using this dataset are listed here.

What (other) tasks could the dataset be used for?
It can be used for the face identification problem. Some researchers have also used it for developing protocols for face identification.

Are there tasks for which the dataset should not be used?
The dataset should not be used for tasks that are high stakes (e.g. law enforcement).

Task Suitability

04

Discussion

Benefits

- Ethical considerations during creation of dataset
- (Legal) Declarations by dataset creators
- Recommendations about dataset usage
- Statistical information - data distribution, etc.
- Information about biases, assumptions, potential limitations, etc.
- Ease of reproducibility
- Able to compare two separate datasets
- Useful links - download location, prior usages, contact information, etc.

Benefits for MLHC

- Medical data results can be wildly different for various factors - age, gender, race, pre-conditions, etc.
- Data distribution statistics will help to understand fairness.
- Medical is highly sensitive, could cause ethical, legal and privacy issues.
- Datasheets provide upfront declaration from creators and cautions for consumers.
- Improves reproducibility. Model results from one hospital environment is often different in other.

Impact

- How practically useful is a datasheet for a consumer?
- Study involving 23 ML engineers provided with a datasheet and followed by an Ethical Sensitivity evaluation.
- All but one participant who were given a Datasheet did open and refer to it.
- Participants were ethically sensitive enough to notice ethical problems in the dataset.
- Participants who were given Datasheets were more likely to mention potential ethical problems in the dataset while they were working.
- Participants with Datasheets heavily relied on them while particularizing.

Source: “Datasheets for Datasets help ML engineers notice and understand ethical issues in training data”, Medium Blog, to be presented at ACM CSCW 2021.

Limitations

- Datasets - only part of the problem. ML models using the dataset could use/ignore certain features.
  Follow-up work - Model cards.
- Needs effort to standardise process. No quantitative measurement to judge the quality of datasheets.
  e.g. What is the incentive/repercussion if a dataset doesn't have good data distribution?
- Retrospective updates (audits) by other researchers to original datasheets

Thank You!