Developing Datasheets for Archived Web Datasets

Emily Maemura
University of Illinois at Urbana-Champaign
emaemura@illinois.edu

WARCnet Closing Conference - October 17, 2022

Data Provenance

… [data provenance] is rarely discussed in the machine learning community. Documenting the creation and use of datasets has received even less attention. Despite the importance of data to machine learning, there is currently no standardized process for documenting machine learning datasets.

> Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

To engage scholarly uses of web archives and support researchers, we must first study web archives practice to understand and communicate the impacts on resulting collections. This study thus asks: How can the sociotechnical process of creating web archives collections be systematically structured and documented?

> Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. *Journal of the Association for Information Science and Technology, 69*(10), 1223–1233. https://doi.org/10.1002/asi.24048

Data Documentation: Web Archives vs. ML

- What are our similar / different needs, motivations?
- What can web archives learn from the *Datasheets for Datasets* approach developed by Gebru et al.?
- What can the existing work on documentation from the web archives community contribute to ML?

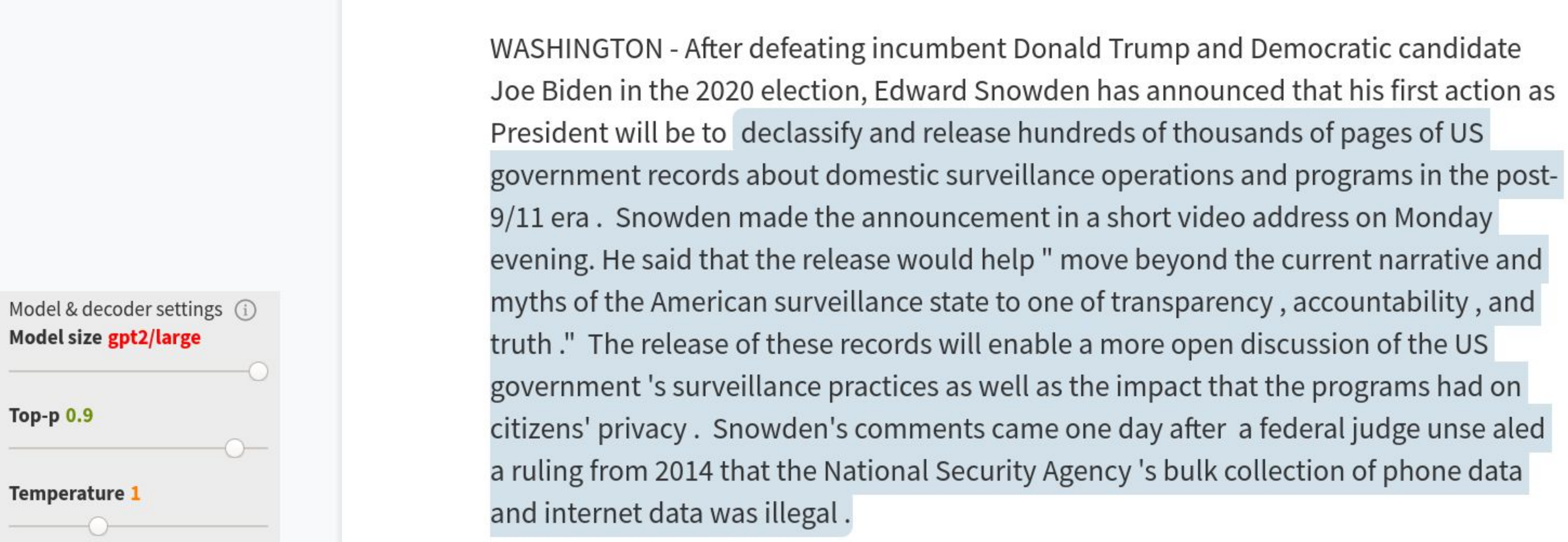
Does ML present an opportunity for web archives use?

Can we adopt the model of datasheets documentation to support new uses of web archives collections?

Comparing Web Archives and ML Datasets

What are ML Datasets used in Large Language Models?

Large Language Models: GPT-2 by OpenAI



Model & decoder settings

Model size gpt2/large

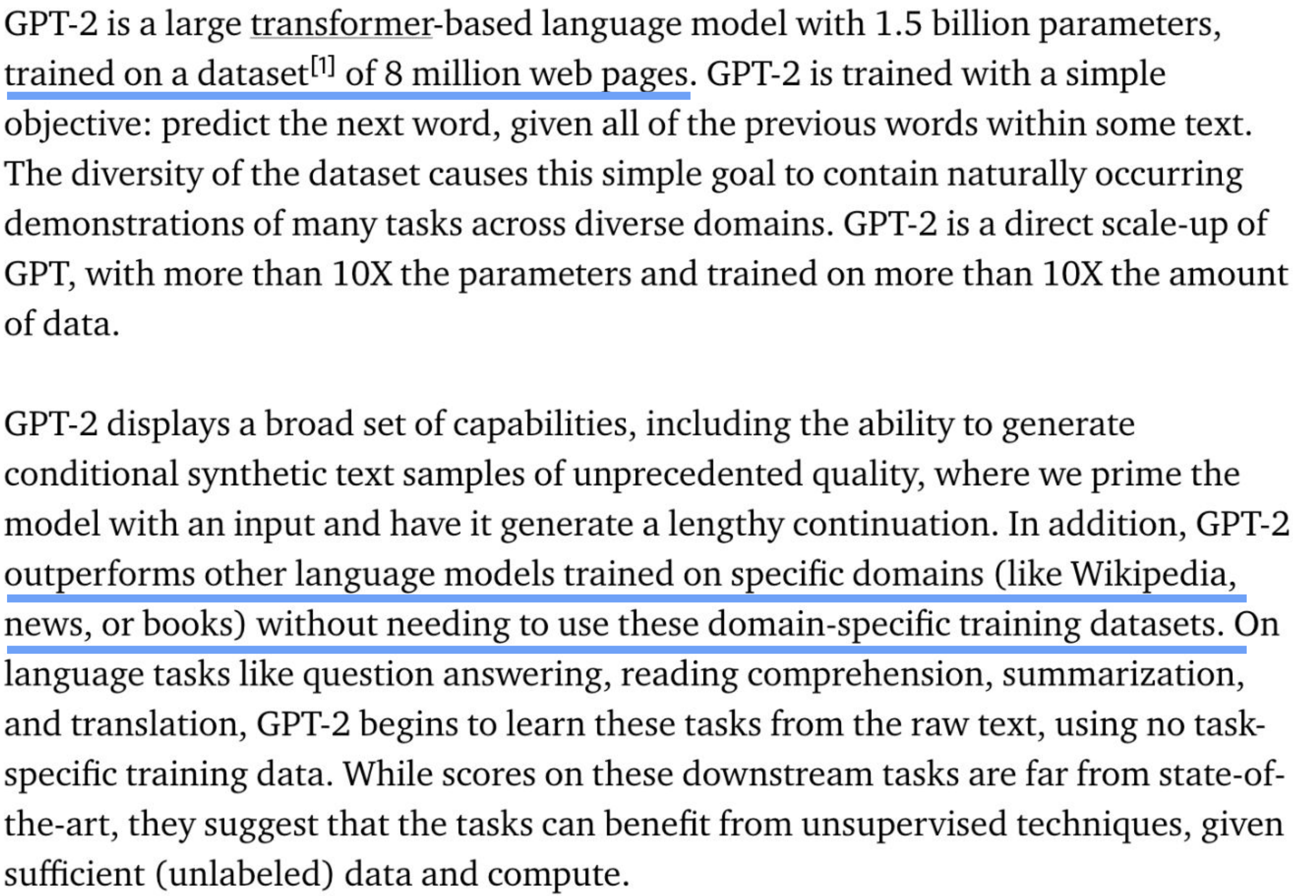
Top-p 0.9

Temperature 1

WASHINGTON - After defeating incumbent Donald Trump and Democratic candidate Joe Biden in the 2020 election, Edward Snowden has announced that his first action as President will be to declassify and release hundreds of thousands of pages of US government records about domestic surveillance operations and programs in the post-9/11 era . Snowden made the announcement in a short video address on Monday evening. He said that the release would help " move beyond the current narrative and myths of the American surveillance state to one of transparency , accountability , and truth ." The release of these records will enable a more open discussion of the US government 's surveillance practices as well as the impact that the programs had on citizens' privacy . Snowden's comments came one day after a federal judge unse aled a ruling from 2014 that the National Security Agency 's bulk collection of phone data and internet data was illegal .

GPT-2 writing a fictional news article about Edward Snowden's actions after winning the 2020 United States presidential election (all highlighted text is machine-generated). While Snowden had (at the time of generation) never been elected to public office, the generated sample is grammatically and stylistically valid. (Example from Wikipedia)

GPT-2 / OpenAI



GPT-2 is a large transformer-based language model with 1.5 billion parameters, trained on a dataset[1] of 8 million web pages. GPT-2 is trained with a simple objective: predict the next word, given all of the previous words within some text. The diversity of the dataset causes this simple goal to contain naturally occurring demonstrations of many tasks across diverse domains. GPT-2 is a direct scale-up of GPT, with more than 10X the parameters and trained on more than 10X the amount of data.

GPT-2 displays a broad set of capabilities, including the ability to generate conditional synthetic text samples of unprecedented quality, where we prime the model with an input and have it generate a lengthy continuation. In addition, GPT-2 outperforms other language models trained on specific domains (like Wikipedia, news, or books) without needing to use these domain-specific training datasets. On language tasks like question answering, reading comprehension, summarization, and translation, GPT-2 begins to learn these tasks from the raw text, using no task-specific training data. While scores on these downstream tasks are far from state-of-the-art, they suggest that the tasks can benefit from unsupervised techniques, given sufficient (unlabeled) data and compute.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better Language Models and Their Implications. OpenAI. https://openai.com/blog/better-language-models/

GPT-2 / OpenAI

1. We created a new dataset which emphasizes diversity of content, by scraping content from the Internet. In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better Language Models and Their Implications. OpenAI. https://openai.com/blog/better-language-models/

GPT-2 / OpenAI

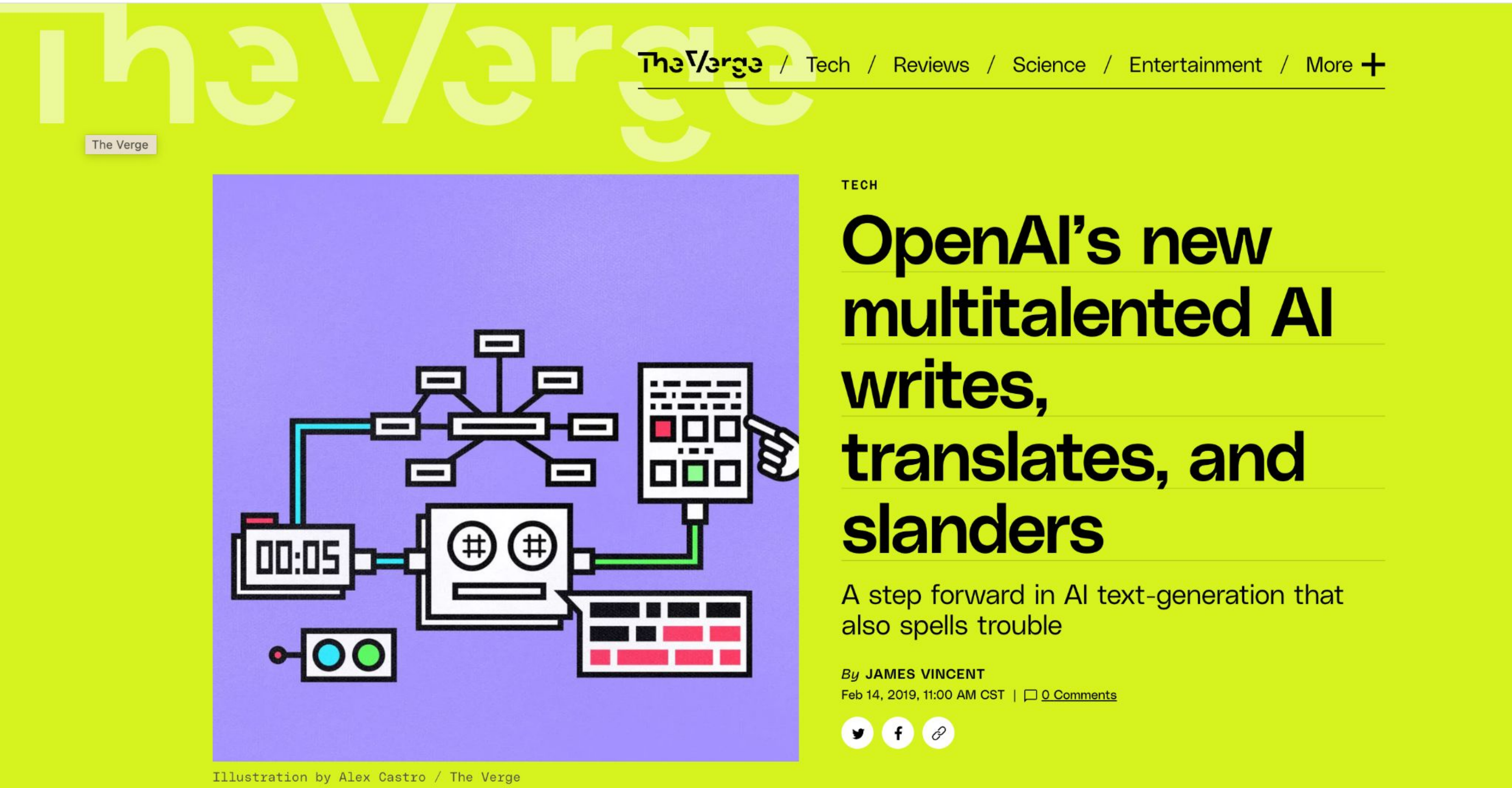
1. We created a new dataset which emphasizes diversity of content, by scraping content from the Internet. In order to preserve document quality, we used only pages which have been curated/filtered by humans—specifically, we used outbound links from Reddit which received at least 3 karma. This can be thought of as a heuristic indicator for whether other users found the link interesting (whether educational or funny), leading to higher data quality than other similar datasets, such as CommonCrawl.

Radford, A., Wu, J., Amodei, D., Amodei, D., Clark, J., Brundage, M., & Sutskever, I. (2019, February 14). Better Language Models and Their Implications. OpenAI. https://openai.com/blog/better-language-models/

Comparing Web Archives and ML datasets

Similarities:

- Are collected via crawling websites
- Represent the outcomes of curation decisions
  - crawl depth (#hops), scale
  - quality, interest, relevance, representativeness



The Verge

The Verge / Tech / Reviews / Science / Entertainment / More +

TECH

OpenAI’s new multitalented AI writes, translates, and slanders

A step forward in AI text-generation that also spells trouble

By JAMES VINCENT

Feb 14, 2019, 11:00 AM CST | 0 Comments

Illustration by Alex Castro / The Verge

https://www.theverge.com/2019/2/14/18224704/ai-machine-learning-language-models-read-write-openai-gpt2

Comparing Web Archives and ML datasets

Differences:

- ML has immediate and widespread applications in crime prediction, employment, finance
- ML focus on accuracy, ground truth, task completion

Comparing Description in Web Archives and ML

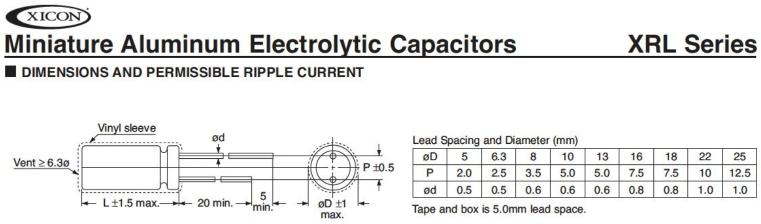
What do Datasheets for Datasets document?

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

To address this gap, we propose *datasheets for datasets*. In the electronics industry, every component, no matter how simple or complex, is accompanied with a datasheet describing its operating characteristics, test results, recommended usage, and other information. By analogy, we propose that every dataset be accompanied with a datasheet that documents its motivation, composition, collection process, recommended uses, and so on.

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Datasheets for Datasets


XICON

Miniature Aluminum Electrolytic Capacitors XRL Series

■ DIMENSIONS AND PERMISSIBLE RIPPLE CURRENT

Vinyl sleeve
ød
Vent ≥ 6.3ø
P ±0.5
L ±1.5 max.
20 min.
5 min.
øD ±1 max.

Lead Spacing and Diameter (mm)

øD	5	6.3	8	10	13	16	18	22	25
P	2.0	2.5	3.5	5.0	5.0	7.5	7.5	10	12.5
ød	0.5	0.5	0.6	0.6	0.6	0.8	0.8	1.0	1.0

Tape and box is 5.0mm lead space.

5°C

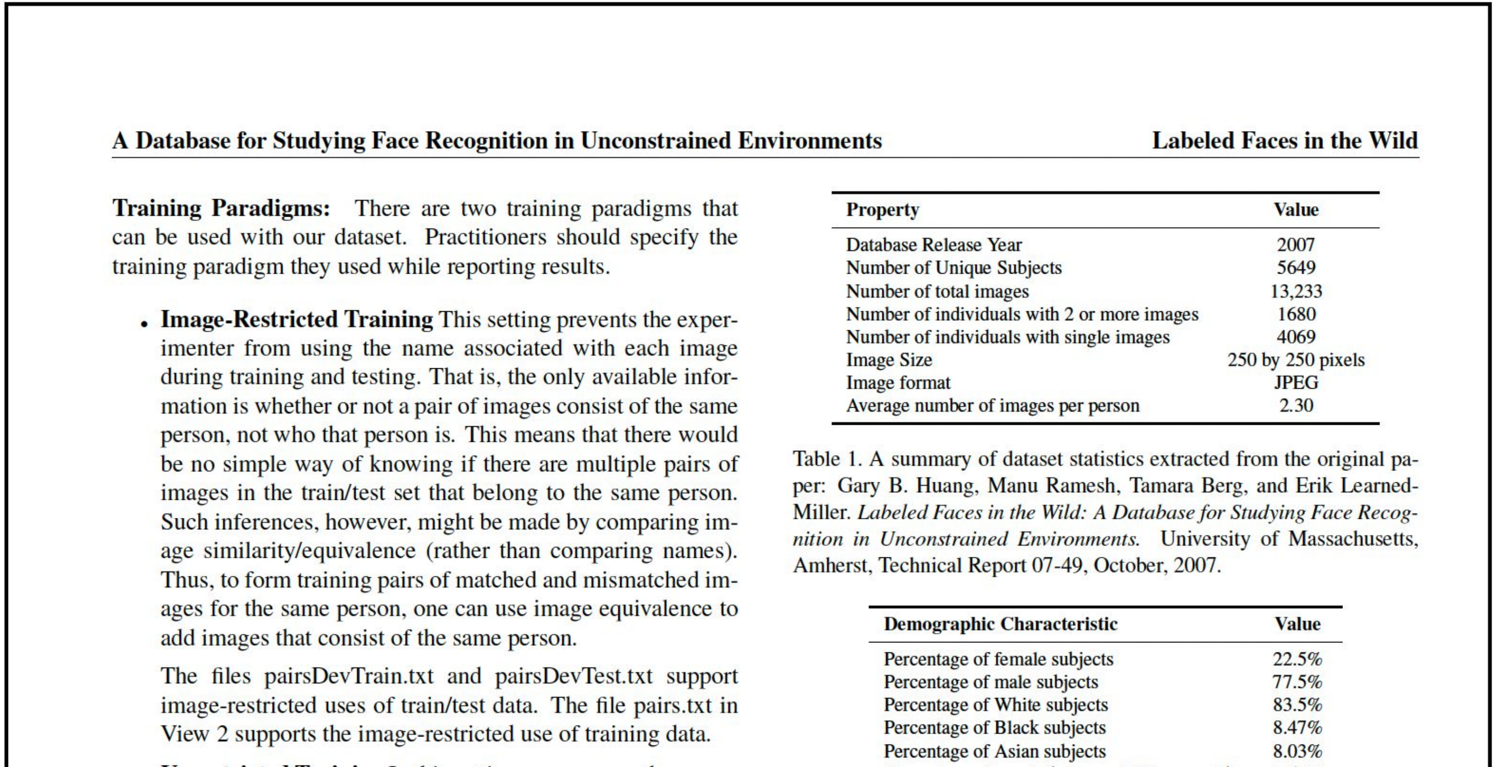
	100	
mA	øD x L	mA
3.0	5 x 11	3.0
4.5	5 x 11	5.8
7.5	5 x 11	8.8
9.5	5 x 11	12
17	5 x 11	22
28	5 x 11	33
34	5 x 11	40
45	5 x 11	48
70	6.3 x 11	80
115	8 x 11.5	135
150	10 x 16	195
190	10 x 16	255
320	10 x 20	370
565	13 x 25	675
765	16 x 32	972
1050	18 x 36	1135
1700	22 x 40	2600
2385		
3000		
3560		

5°C

450	
›D x L	mA
) x 12.5	26
) x 12.5	38
0 x 16	63
0 x 20	86
3 x 20	120
3 x 25	192
6 x 25	354
8 x 36	426
8 x 40	555
2 x 45	750

XICON

Date Revised: 1/8/07
producer documentation.


A Database for Studying Face Recognition in Unconstrained Environments — Labeled Faces in the Wild

Training Paradigms: There are two training paradigms that can be used with our dataset. Practitioners should specify the training paradigm they used while reporting results.

- **Image-Restricted Training** This setting prevents the experimenter from using the name associated with each image during training and testing. That is, the only available information is whether or not a pair of images consist of the same person, not who that person is. This means that there would be no simple way of knowing if there are multiple pairs of images in the train/test set that belong to the same person. Such inferences, however, might be made by comparing image similarity/equivalence (rather than comparing names). Thus, to form training pairs of matched and mismatched images for the same person, one can use image equivalence to add images that consist of the same person.

  The files pairsDevTrain.txt and pairsDevTest.txt support image-restricted uses of train/test data. The file pairs.txt in View 2 supports the image-restricted use of training data.

Property	Value
Database Release Year	2007
Number of Unique Subjects	5649
Number of total images	13,233
Number of individuals with 2 or more images	1680
Number of individuals with single images	4069
Image Size	250 by 250 pixels
Image format	JPEG
Average number of images per person	2.30

Table 1. A summary of dataset statistics extracted from the original paper: Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* University of Massachusetts, Amherst, Technical Report 07-49, October, 2007.

Demographic Characteristic	Value
Percentage of female subjects	22.5%
Percentage of male subjects	77.5%
Percentage of White subjects	83.5%
Percentage of Black subjects	8.47%
Percentage of Asian subjects	8.03%

Datasheets for Datasets: Development

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2018). Datasheets for Datasets. *ArXiv:1803.09010 [Cs]*. http://arxiv.org/abs/1803.09010

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86–92. https://doi.org/10.1145/3458723

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning/Labeling;
5. Uses
6. Distribution;
7. Maintenance.

Legal & Ethical Considerations section was included in the 2018 Working Paper draft, removed in final publication which instead "introduced factual questions intended to elicit relevant information about compliance without requiring dataset creators to make legal judgments."

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning /Labeling;
5. Uses
6. Distribution;
7. Maintenance.

1. For what purpose was the dataset created?
2. Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning /Labeling;
5. Uses
6. Distribution;
7. Maintenance.

5. What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?
6. How many instances are there in total (of each type, if appropriate)?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning /Labeling;
5. Uses
6. Distribution;
7. Maintenance.

22. What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)?

24. Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning/Labeling;
5. Uses
6. Distribution;
7. Maintenance.

34. Was the “raw” data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)?
35. Is the software that was used to preprocess/clean/label the data available?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning /Labeling;
5. Uses
6. Distribution;
7. Maintenance.

38. Is there a repository that links to any or all papers or systems that use the dataset?

40. Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned /labeled that might impact future uses?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning /Labeling;
5. Uses
6. Distribution;
7. Maintenance.

44. How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?
45. When will the dataset be distributed?
46. Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?

Datasheets: Questions and Workflow

1. Motivation;
2. Composition;
3. Collection Process;
4. Preprocessing/Cleaning/Labeling;
5. Uses
6. Distribution;
7. Maintenance.

50. Who will be supporting/hosting/maintaining the dataset?
51. How can the owner/curator/manager of the dataset be contacted (for example, email address)?
52. Is there an erratum?

Comparing with Web Archives

Documenting Elements of Web Archives Provenance

- Scoping Elements
- Process Elements
- Context Elements

Maemura, E., Worby, N., Milligan, I., & Becker, C. (2018). If These Crawls Could Talk: Studying and Documenting Web Archives Provenance. *Journal of the Association for Information Science and Technology*, *69*(10), 1223–1233. https://doi.org/10.1002/asi.24048

TABLE 2. Scoping elements.

Element	Key questions and information to document
Motivation	What is the purpose of the collection? Has its mandate changed over time?
Focus	Which geographic, temporal, technical, political, topical and/or social boundaries are defined to scope the collection?
Access & discovery	Who is the intended audience? Do they have known characteristics or needs? Which contractual, organizational, legal, or other agreements restrict access? What metadata fields and indexes support discovery? At what degree of granularity (by collection, site, or individual resource)? Which data formats or derivative datasets are available?
Seed list	What seeds were used in the scoping of the collection? What was the process of discovering and selecting seeds?
Crawl timing	What is the frequency of crawls? How long do crawls run or what time limit is set?
Crawl configuration	What settings control the depth of a crawl? For example, settings for capture by distance from original seed. Is the goal to have a more comprehensive or a breadth-focused collection?
Inclusions and exclusions	Are certain sites or media types included or excluded? For example, are regular expressions used to target certain files or directories in a URL structure.
Permissions from site admins	How were restrictions such as robots.txt and blocks handled?

TABLE 2. Scoping elements.

Element
Motivation
Focus
Access & discovery
Seed list
Crawl timing
Crawl configuration
Inclusions and exclusions
Permissions from site admins

TABLE 3. Process elements.

Element
Scheduled events
Unscheduled events or process anomalies

TABLE 4. Context elements.

Element
Legal context
Institutional setting and mandate
Policies and Guidelines
Resources available for web archiving

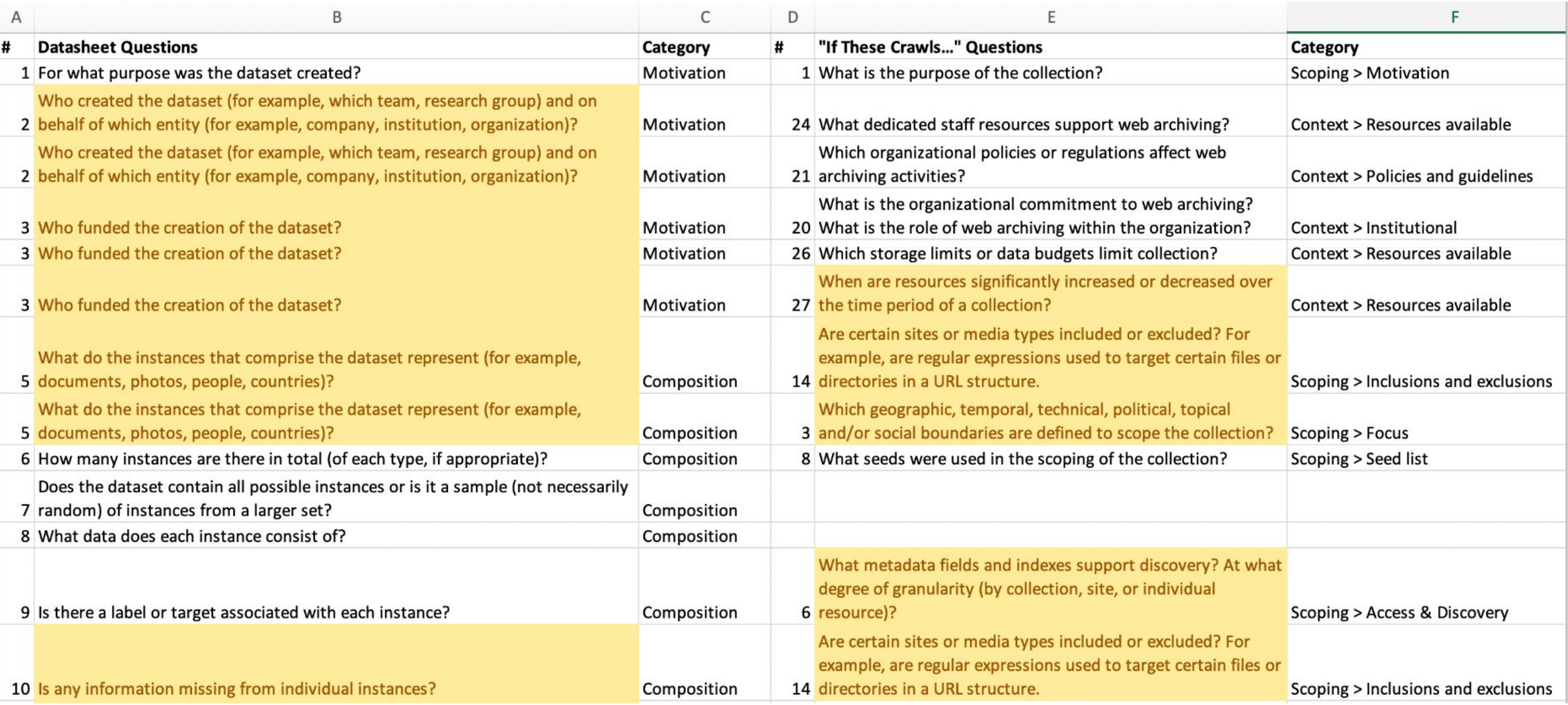
Structured Comparison: A Crosswalk

57 Questions

Datasheets Categories
Motivation
Composition
Collection Process
Preprocessing/Cleaning/Labeling
Uses
Distribution
Maintenance

27 Questions

“If These Crawls…” Elements
Scoping: Motivation, Focus, …
Process: Scheduled events, …
Context: Legal, Institutional, …



#	Datasheet Questions	Category	#	"If These Crawls..." Questions	Category
1	For what purpose was the dataset created?	Motivation	1	What is the purpose of the collection?	Scoping > Motivation
2	Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?	Motivation	24	What dedicated staff resources support web archiving?	Context > Resources available
2	Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?	Motivation	21	Which organizational policies or regulations affect web archiving activities?	Context > Policies and guidelines
3	Who funded the creation of the dataset?	Motivation	20	What is the organizational commitment to web archiving? What is the role of web archiving within the organization?	Context > Institutional
3	Who funded the creation of the dataset?	Motivation	26	Which storage limits or data budgets limit collection?	Context > Resources available
3	Who funded the creation of the dataset?	Motivation	27	When are resources significantly increased or decreased over the time period of a collection?	Context > Resources available
5	What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?	Composition	14	Are certain sites or media types included or excluded? For example, are regular expressions used to target certain files or directories in a URL structure.	Scoping > Inclusions and exclusions
5	What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)?	Composition	3	Which geographic, temporal, technical, political, topical and/or social boundaries are defined to scope the collection?	Scoping > Focus
6	How many instances are there in total (of each type, if appropriate)?	Composition	8	What seeds were used in the scoping of the collection?	Scoping > Seed list
7	Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	Composition			
8	What data does each instance consist of?	Composition			
9	Is there a label or target associated with each instance?	Composition	6	What metadata fields and indexes support discovery? At what degree of granularity (by collection, site, or individual resource)?	Scoping > Access & Discovery
10	Is any information missing from individual instances?	Composition	14	Are certain sites or media types included or excluded? For example, are regular expressions used to target certain files or directories in a URL structure.	Scoping > Inclusions and exclusions

Crosswalk Findings: Alignment and Overlap

- Motivation for datasets
  - purpose, who created the dataset
- Composition of datasets
  - Metadata / labels, missing information
- Collection process
  - Timeframe of collection, mechanisms or procedures

Higher Degree of Granularity “Datasheets”

- Composition
    - recommended data splits
    - known errors, noise, redundancies
    - confidential, offensive, identifying subpopulations
- Uses
    - past uses, potential future uses, “tasks for which the dataset should not be used”

Higher Degree of Granularity “If These Crawls”

- Collection Process
    - human decisions (reappraisal, unscheduled events)
- Context
    - policy and organizational changes over time

Imperfect Alignment / Important Distinctions

- Funding and motivations
  - Who funded the creation of the dataset?
  - vs. organizational commitment, data budget, resources used over time
- Ethics and consent
  - Did the individuals in question consent to the collection and use of their data? + revoking consent
  - vs. How were restrictions such as robots.txt and blocks handled?

Imperfect Alignment / Important Distinctions

- Who was involved in the data collection process
  - Datasheets question addresses roles (students, crowdworkers, contractors) and compensation
  - vs. dedicated staff resources in institutional context
- Ongoing evolution of datasets
  - Datasheets treats more as product (to be distributed, maintained)
  - vs. policies and resources changing over time

Crosswalk Summary

“Datasheets” brings important perspectives on how people are represented in web data, and focuses description on ***who*** was involved in collecting.

“If These Crawls” elements are grounded in existing organizational and institutional configurations, and describe the ***how*** of collecting in greater detail.

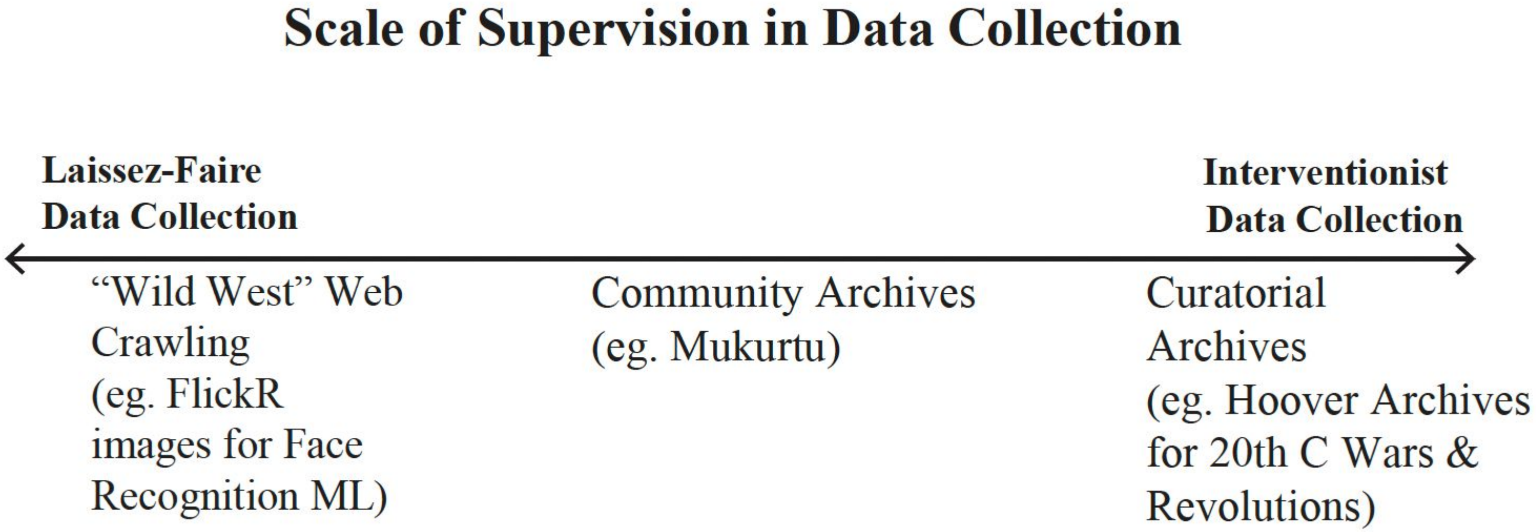
Lessons learned from both ML and Web Archives

“Lessons from Archives” for ML

In spite of its fundamental nature however, data collection remains an overlooked part of the machine learning (ML) pipeline. In this paper we argue that a new specialization should be formed within ML that is focused on methodologies for data collection and annotation: efforts that require institutional frameworks and procedures. Specifically for sociocultural data, parallels can be drawn from archives and libraries.

> Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 306–316. https://doi.org/10.1145/3351095.3372829

Lessons from Archives



Scale of Supervision in Data Collection

Laissez-Faire Data Collection

Interventionist Data Collection

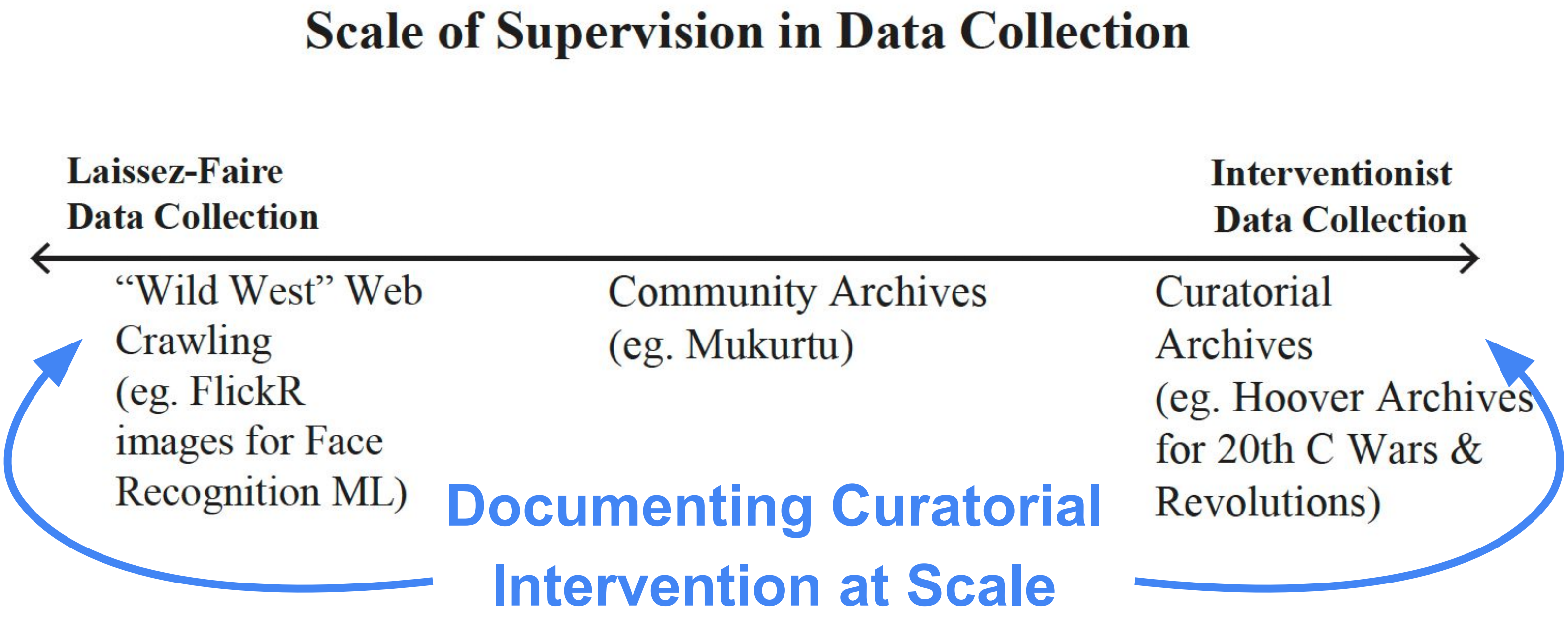
"Wild West" Web Crawling (eg. FlickR images for Face Recognition ML)

Community Archives (eg. Mukurtu)

Curatorial Archives (eg. Hoover Archives for 20th C Wars & Revolutions)

Figure 2: Example categories of data collection practices on supervision scale.

Lessons from Web Archives



Scale of Supervision in Data Collection

Laissez-Faire Data Collection

Interventionist Data Collection

"Wild West" Web Crawling (eg. FlickR images for Face Recognition ML)

Community Archives (eg. Mukurtu)

Curatorial Archives (eg. Hoover Archives for 20th C Wars & Revolutions)

Documenting Curatorial Intervention at Scale

Figure 2: Example categories of data collection practices on supervision scale.

Lessons from ML: Addressing gaps and biases

For GPT-2's dataset focused on Reddit links:

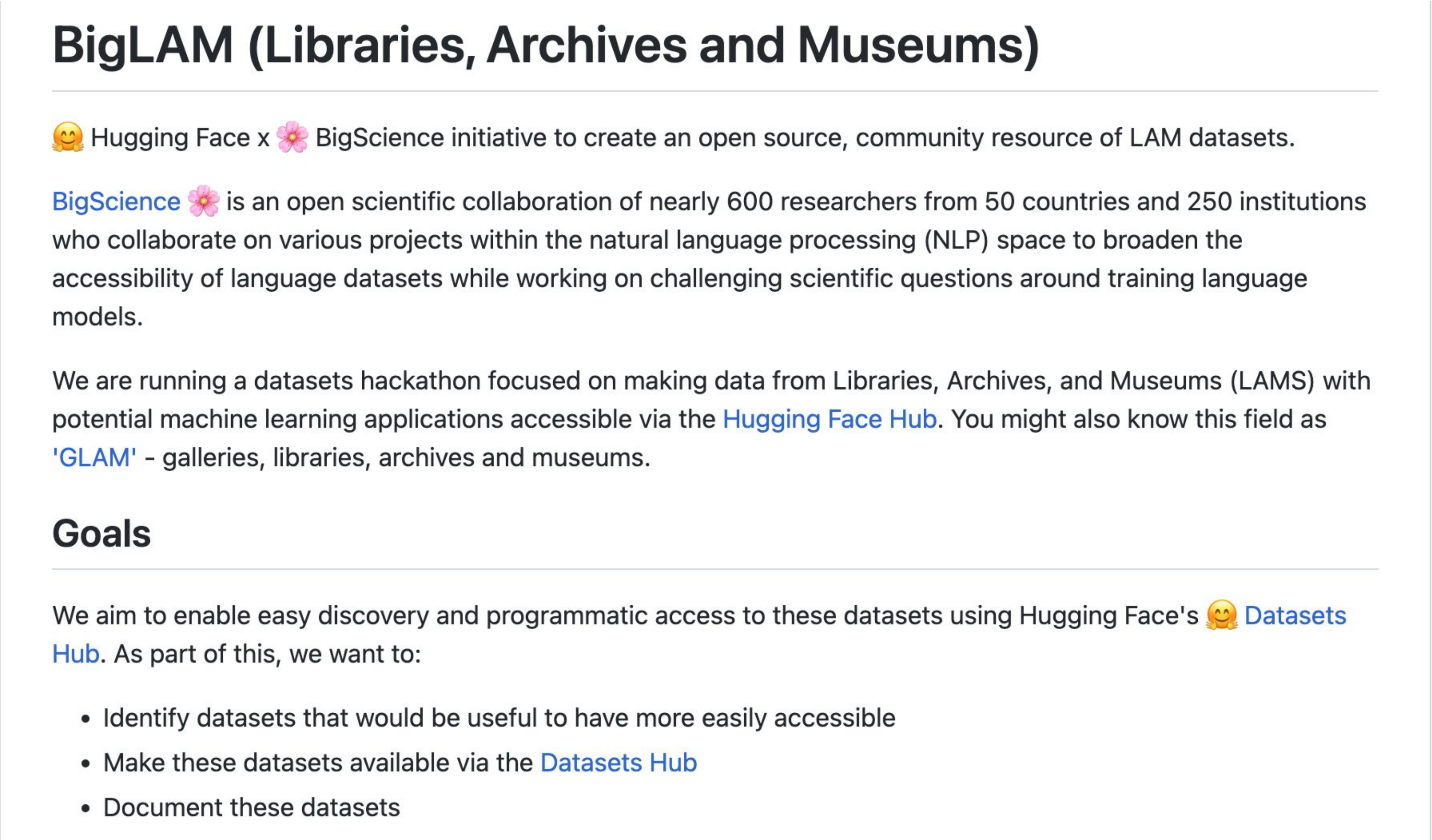
- recognizing the datasets 'inherits' biases "Pew Internet Research's 2013 survey reveals Reddit users in the United States are more likely to be male, in their late-teens to twenties, and urbanites."

Generating "Complement" Dataset in response:

- which "aims to actively collect the variety of English spoken by American culture broadly … from peoples underrepresented on the internet. The dataset is especially focused on colloquial American English across class, education levels, age, and immigration status."

Jo, E. S., & Gebru, T. (2020). Lessons from archives: Strategies for collecting sociocultural data in machine learning. *FAccT 2020*, 306–316. https://doi.org/10.1145/3351095.3372829

Conclusions and Next Steps



BigLAM (Libraries, Archives and Museums)

🤗 Hugging Face x 🌸 BigScience initiative to create an open source, community resource of LAM datasets.

BigScience 🌸 is an open scientific collaboration of nearly 600 researchers from 50 countries and 250 institutions who collaborate on various projects within the natural language processing (NLP) space to broaden the accessibility of language datasets while working on challenging scientific questions around training language models.

We are running a datasets hackathon focused on making data from Libraries, Archives, and Museums (LAMS) with potential machine learning applications accessible via the Hugging Face Hub. You might also know this field as 'GLAM' - galleries, libraries, archives and museums.

Goals

We aim to enable easy discovery and programmatic access to these datasets using Hugging Face's 🤗 Datasets Hub. As part of this, we want to:

- Identify datasets that would be useful to have more easily accessible
- Make these datasets available via the Datasets Hub
- Document these datasets

https://github.com/bigscience-workshop/lam

~~Conclusions~~ Questions / Provocations

- Should we consider ML training data to be a kind of ‘archived web dataset’? Does that alignment help or hinder the web archives community?
- Datasheets for Datasets presents one model for documentation, can other potential data description formats provide useful models to relate our work to other communities, audiences?
  - e.g. Codebooks, Data biography, Data user guides examples described in *Data Feminism* to ‘Consider Context’

Thank You!

emaemura@illinois.edu

More on learning from ML: Stochastic Parrots

- environmental impact and scale of LLMs
- “size doesn’t guarantee diversity”
- potential harms to marginalized populations through ‘data cleaning’ based on removing ‘bad words’

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

More on learning from ML: Stochastic Parrots

While documentation allows for potential accountability, undocumented training data perpetuates harm without recourse. Without documentation, one cannot try to understand training data characteristics in order to mitigate some of these attested issues or even unknown ones. The solution, we propose, is to budget for documentation as part of the planned costs of dataset creation, and only collect as much data as can be thoroughly documented within that budget.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 *FAccT 2021*, 610–623. https://doi.org/10.1145/3442188.3445922