Ethical Dimensions of Computer Vision Datasets

Emily Denton
Research Scientist, Google

Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development

MORGAN KLAUS SCHEUERMAN*, University of Colorado Boulder, USA
EMILY DENTON, Google Research
ALEX HANNA, Google Research

Data and its (dis)contents: A survey of dataset development and use in machine learning research

Amandalynne Paullada
Department of Linguistics
University of Washington

Inioluwa Deborah Raji
Mozilla Foundation

Emily M. Bender
Department of Linguistics
University of Washington

Emily Denton
Google Research

Alex Hanna
Google Research

Concerns regarding dataset design and development

I. **Representational concerns**
II. Task formulation
III. Collection, annotation, & documentation
IV. Disciplinary values, norms, & practices

Underrepresentation of darker skin tones

Facial analysis datasets

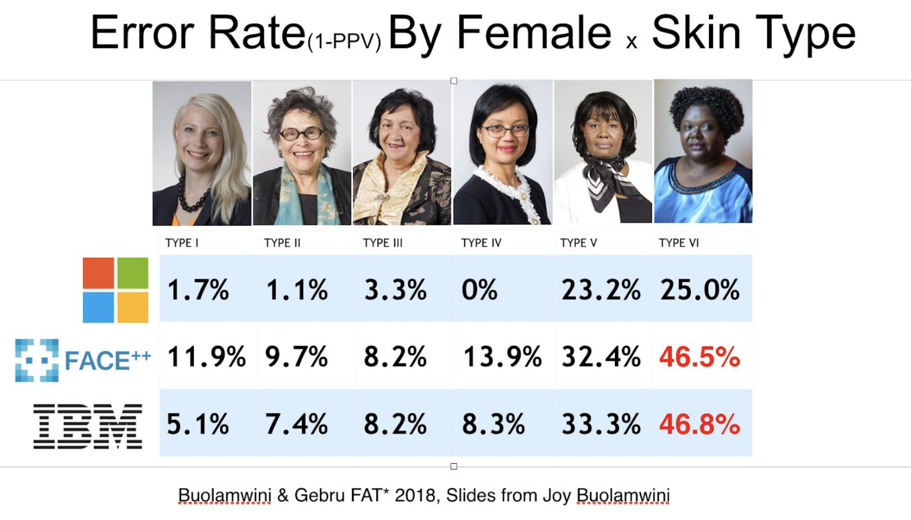
LFW	77.5% male 83.5% white
IJB-A	79.6% lighter-skinned
Adience	86.2% lighter-skinned

Buolamwini & Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Underrepresentation of darker skin tones

Facial analysis datasets

LFW	77.5% male 83.5% white
IJB-A	79.6% lighter-skinned
Adience	86.2% lighter-skinned



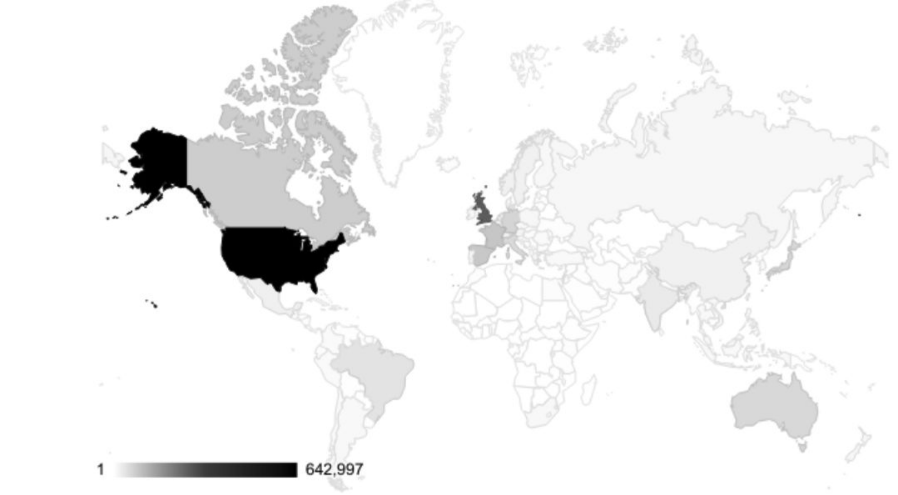
Error Rate(1-PPV) By Female x Skin Type

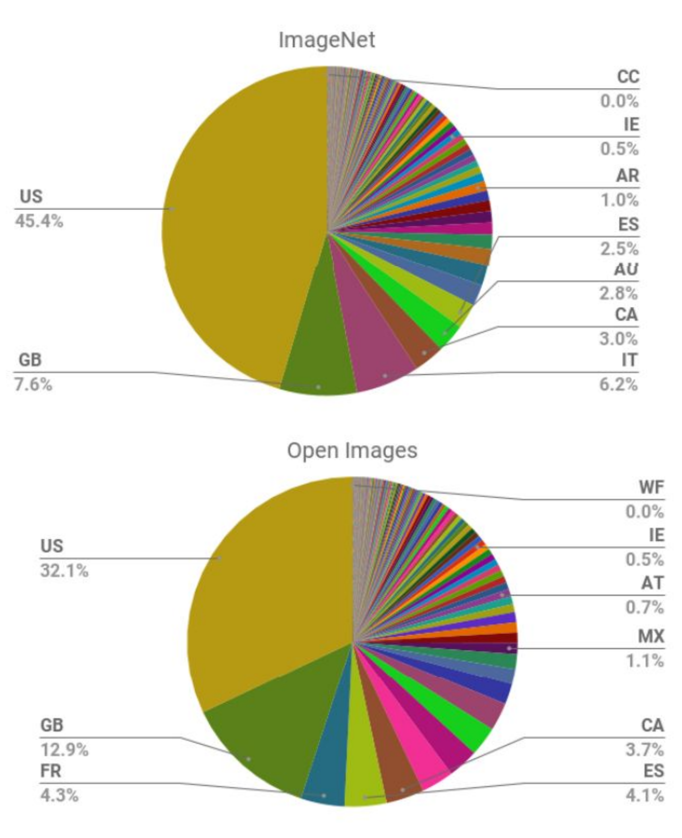
	TYPE I	TYPE II	TYPE III	TYPE IV	TYPE V	TYPE VI
	1.7%	1.1%	3.3%	0%	23.2%	25.0%
FACE++	11.9%	9.7%	8.2%	13.9%	32.4%	46.5%
IBM	5.1%	7.4%	8.2%	8.3%	33.3%	46.8%

Buolamwini & Gebru FAT* 2018, Slides from Joy Buolamwini

Buolamwini & Gebru (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification

Underrepresentation of non-Western images





Shankar et al. (2017). [No Classification without Representation: Assessing Geo-diversity Issues in Open Data Sets for the Developing World]
DeVries et al. (2019). [Does Object Recognition Work for Everyone?]

Underrepresentation of non-Western images

Ground truth: Soap
Nepal, 288 $ / month

Common machine classifications: food, cheese, food product, dish, cooking

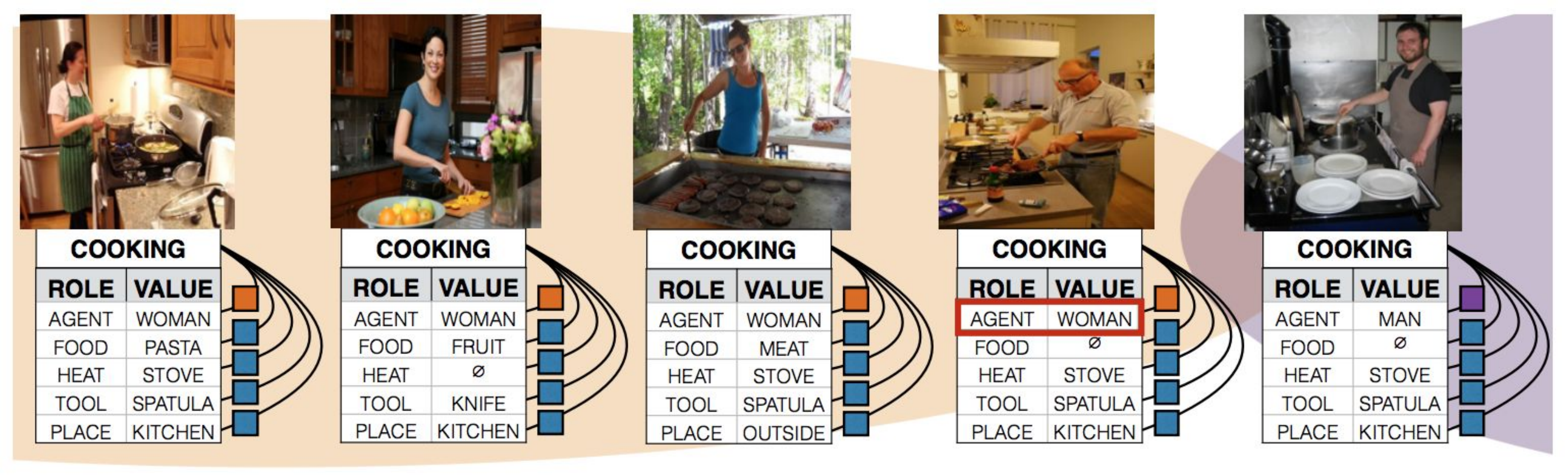
Ground truth: Soap
UK, 1890 $ / month

Common classification: soap dispenser, toiletry, faucet, lotion

Shankar et al. (2017). No Classification without Representation: Assessing Geo-diversity Issues in Open Data Sets for the Developing World
DeVries et al. (2019). Does Object Recognition Work for Everyone?

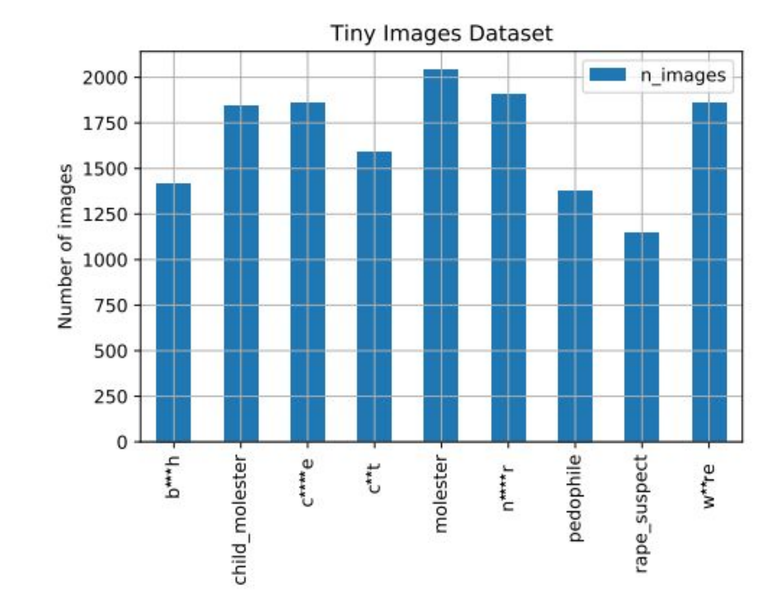
Stereotype aligned correlations

Training data: 33% of cooking images have man in the agent role
Model predictions: 16% cooking images have man in the agent role

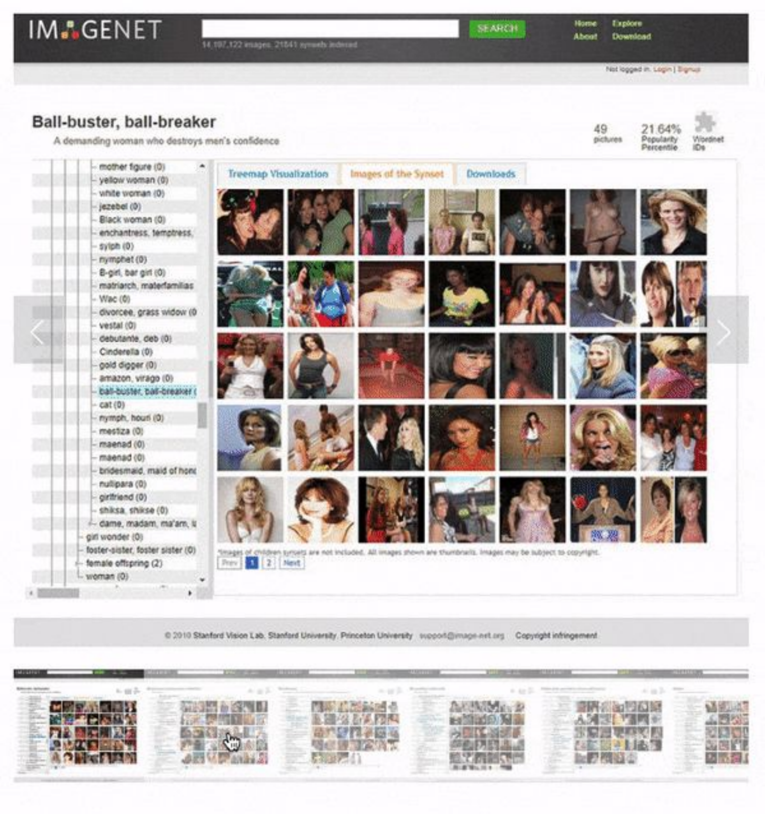


Zhao et al. (2017) [Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints]

Toxic categories, including racial slurs and derogatory phrases



Prabhu & Birhane (2020). [Large image datasets: A pyrrhic win for computer vision?]()



Selections from the "Person" classes, ImageNet

Crawford and Paglen. 2019. [excavating.ai]()

Concerns regarding dataset design and development

I. Representational concerns

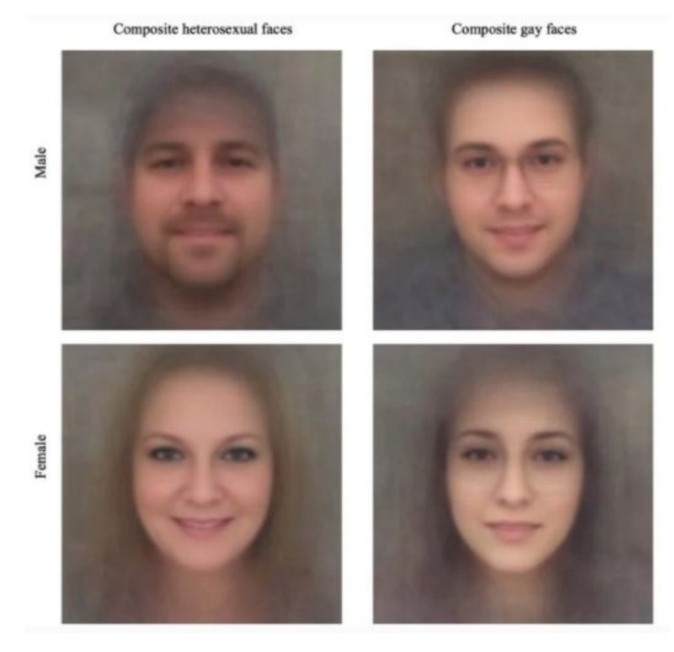
II. **Task formulation**

III. Collection, annotation, & documentation

IV. Disciplinary values, norms, & practices

Datasets legitimize certain problems or goals

“[T]he ‘problematization’ that guides data collection leads to the creation of datasets that formulate pseudoscientific, often unjust tasks” (Paullada et al. 2020)



Wang & Kosinski (2017)

Datasets legitimize certain problems or goals

"[T]he 'problematization' that guides data collection leads to the creation of datasets that formulate pseudoscientific, often unjust tasks" ([Paullada et al. 2020](#))

Do algorithms reveal sexual orientation or just expose our stereotypes?

Blaise Aguera y Arcas Jan 11, 2018 · 15 min read

by Blaise Agüera y Arcas, Alexander Todorov and Margaret Mitchell

Datasets legitimize certain problems or goals

"[T]he 'problematization' that guides data collection leads to the creation of datasets that formulate pseudoscientific, often unjust tasks" (Paullada et al. 2020)

Physiognomy's New Clothes

Blaise Aguera y Arcas May 6, 2017 · 38 min read

by Blaise Agüera y Arcas, Margaret Mitchell and Alexander Todorov

Figure 1. A couple viewing the head of Italian criminologist Cesare Lombroso preserved in a jar of formalin at an exhibition in Bologna, 1978. (Photo by Romano Cagnoni/Hulton Archive/Getty Images)

Concerns regarding dataset design and development

I. Representational concerns

II. Task formulation

III. Collection, annotation, & documentation

IV. Disciplinary values, norms, & practices

Consent and privacy concerns

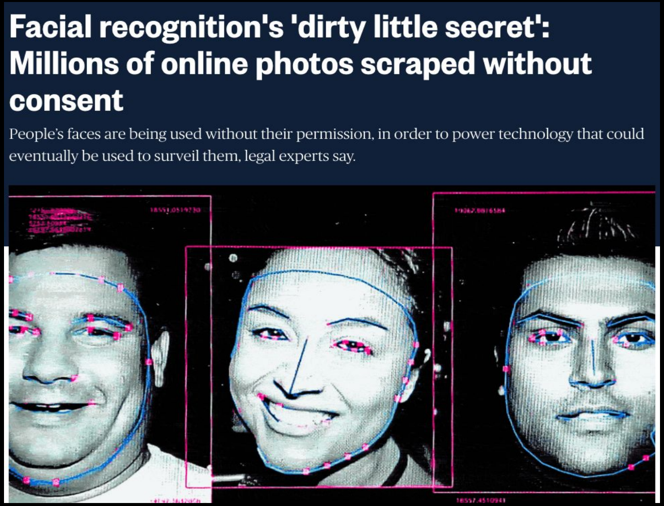
Informed consent is rarely sought from data subjects ([Harvey & LaPlace, 2019]; [Prabhu & Birhane, 2020])


How Photos of Your Kids Are Powering Surveillance Technology

Millions of Flickr images were sucked into a database called MegaFace. Now some of those faces may have the ability to sue.

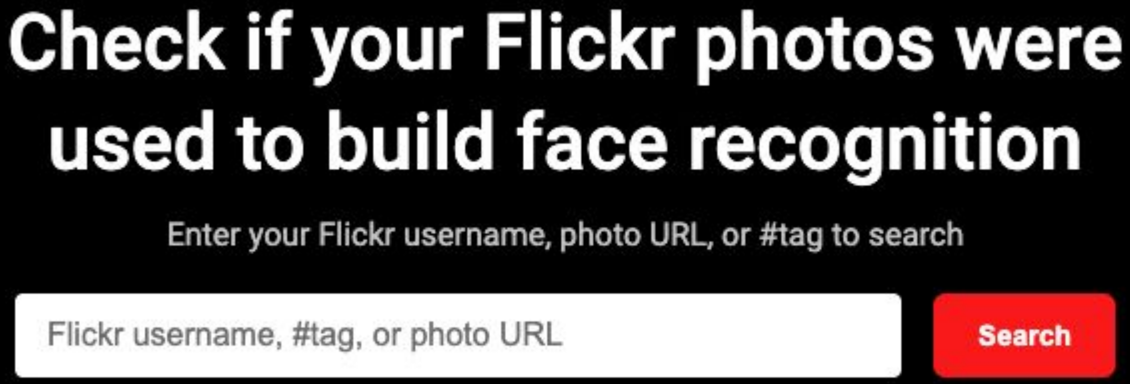
By Kashmir Hill and Aaron Krolik

A selection of images from the MegaFace database.


Facial recognition's 'dirty little secret': Millions of online photos scraped without consent

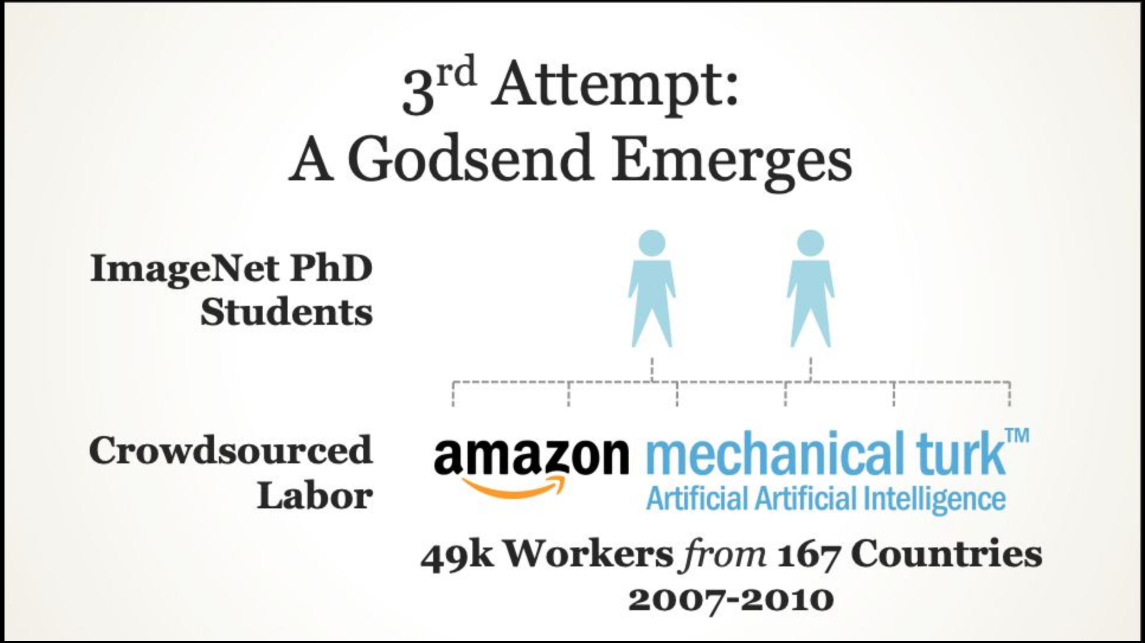
People's faces are being used without their permission, in order to power technology that could eventually be used to surveil them, legal experts say.

Consent and privacy concerns



exposing.ai

Crowdsourced labor concerns



Fei-Fei (2017)

Crowdsourced labor concerns

RESEARCH ARTICLE | JANUARY 01 2015

Difference and Dependence among Digital Workers: The Case of Amazon Mechanical Turk

Lilly Irani

South Atlantic Quarterly (2015) 114 (1): 225–234.

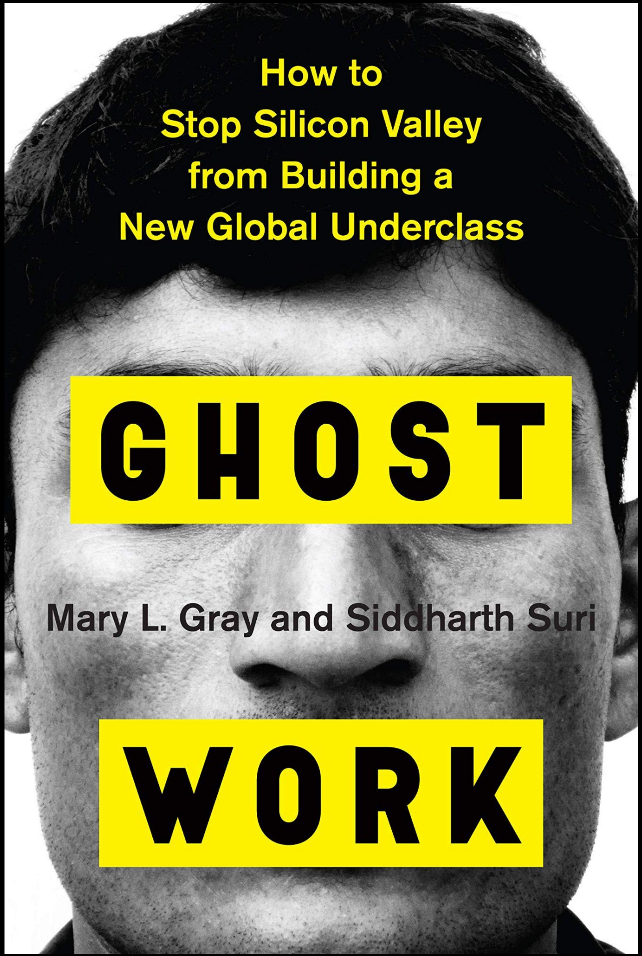
Article

new media & society
0(0) 1–21

DOI: 10.1177/1461444813511926
nms.sagepub.com
SAGE

The cultural work of microwork

Lilly Irani
UC San Diego, USA



Crowdsourced labor concerns

Findings from Scheuerman et al. (2021):

4% of papers presenting new computer vision datasets mentioned if annotators were compensated

"A major focus in discussing human annotation was the time and monetary cost of annotation, particularly as a ***barrier*** to annotating large-scale datasets"

"There is also the goal of minimizing human labor costs, suggesting a ***devaluing*** of labor that is otherwise valuable to the process of dataset curation"

Annotator subjectivities

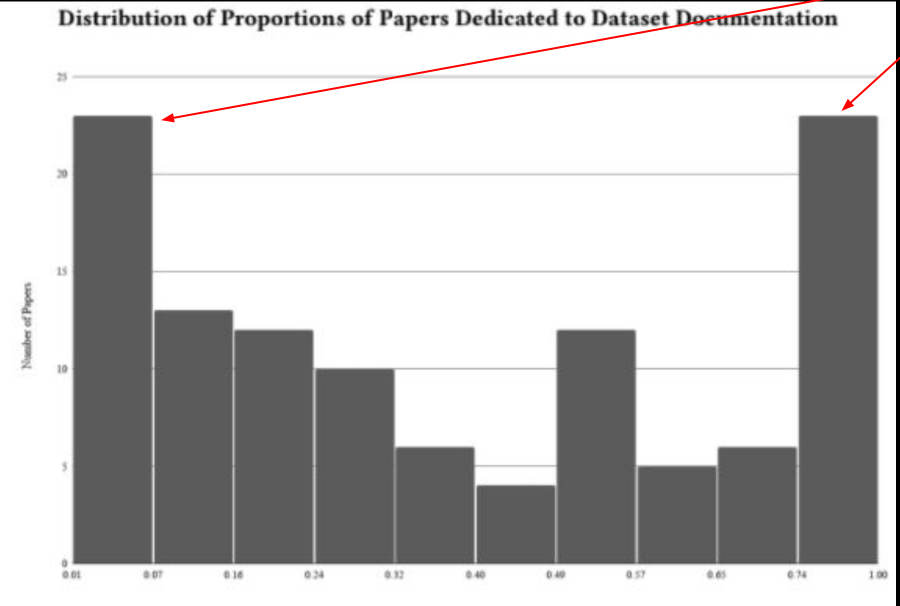
- Annotation discrepancies often attributed to human error rather than differences in perspective, unclear task specifications, subjective interpretation (Scheuerman et al. 2021)
- Annotation and labelling is rarely viewed as interpretive work (Miceli et al. 2020)
    - Annotation demographics often underspecified -- annotators presumed interchangeable (Scheuerman et al. 2021)
- Ground truth often presumed to be fact (Aroyo & Welty, 2015; Muller et al. 2019)

Minimal dataset documentation

- Inconsistent and minimal dataset documentation across ML datasets generally (Geiger et al. 2020; Scheuerman et al. 2020; Gebru, et al. 2018; Holland et al. 2018; Bender and Friedman, 2018; Hutchinson et al., 2020)

Minimal dataset documentation

Bi-model distribution with the majority of papers having either near-0% or near-100% of the paper about the dataset.



Scheuerman et al. (2021) Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development

Minimal dataset documentation

- Inconsistent and minimal dataset documentation across ML (Geiger et al. 2020; Scheuerman et al. 2020; Gebru, et al. 2018; Holland et al. 2018; Bender and Friedman, 2018; Hutchinson et al., 2020)

- Categories tend to be presented as natural
  - Even highly political categories such as race and gender tend to be presented as indisputable and natural (Scheuerman et al. 2020)

- Annotation demographics often underspecified (Scheuerman et al. 2021)

Concerns regarding dataset design and development

I. Representational concerns

II. Task formulation

III. Collection, annotation, & documentation

IV. Disciplinary values, norms, & practices

Devaluation of careful data work

"Publications that report solely on datasets are typically not published. If they are published without a corresponding model or technical development, they are typically relegated to a non-archival technical report, rather than published in a top-tier venue. For this matter, reporting and evaluation of the model work is what is typically incentivized, rather than the careful, slow data work."
(Scheuerman et al. 2021)

Devaluation of careful data work

“*Everyone wants to do the model work, not the data work*”: Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo

Lack of investment in careful dataset maintenance

Dataset Availability via Publication Documentation			
Dataset Property	***k***	***N***	***%***
URL in Paper	69	114	60.5
Any Website (In Paper + Discovered through Search)	97	114	85
Website in Paper Still Available	46	69	66.7
Data Still Downloadable	59	80[a]	73.8
DOI	3	114	2.6
Hosted on Personal/Lab Website	102	114	89.5
Hosted on Institutional Repository	1	114	0.9

Scheuerman et al. (2021) [Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development]

Lack of investment in careful dataset maintenance

Datasets are often not maintained or distributed with care

Dataset

The dataset is no longer publicly available due to copyright issues. For thoese who have already downloaded the dataset, please note that using or distributing it is illegal!

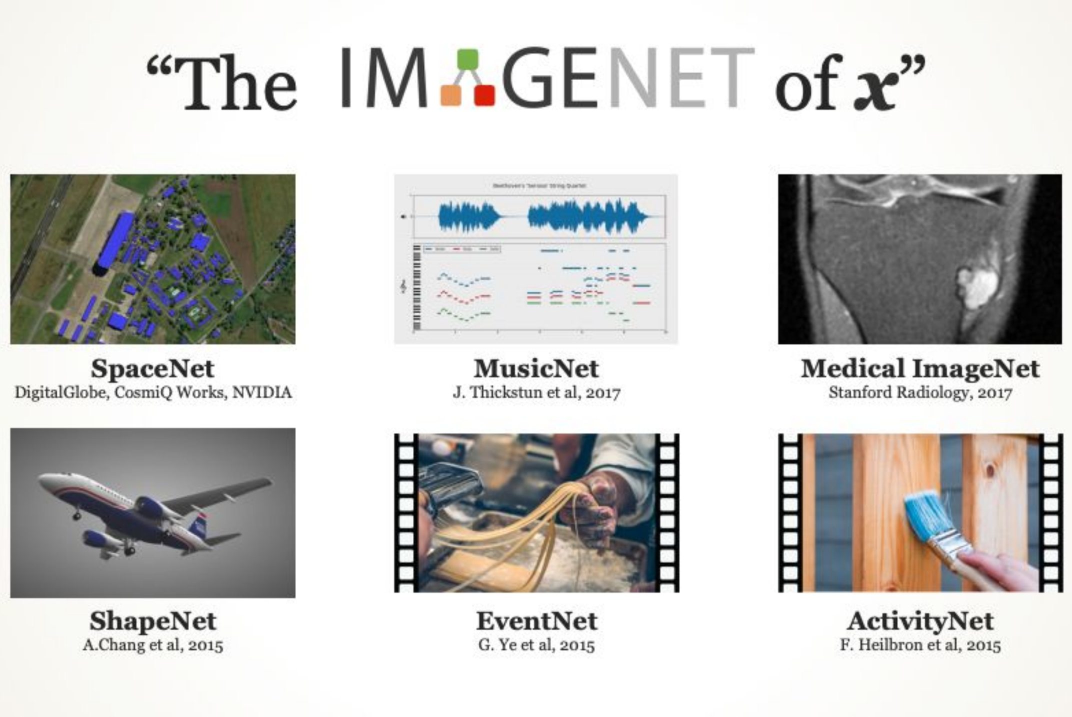
Disclaimer

THIS DATABASE IS PROVIDED "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. The images provided were produced by third-parties, who may have retained copyrights. They are provided strictly for non-profit research purposes, and limited, controlled distributed, intended to fall under the fair-use limitation. We take no guarantees or responsibilities, whatsoever, arising out of any copyright issue. Use at your own risk.

Disclaimer

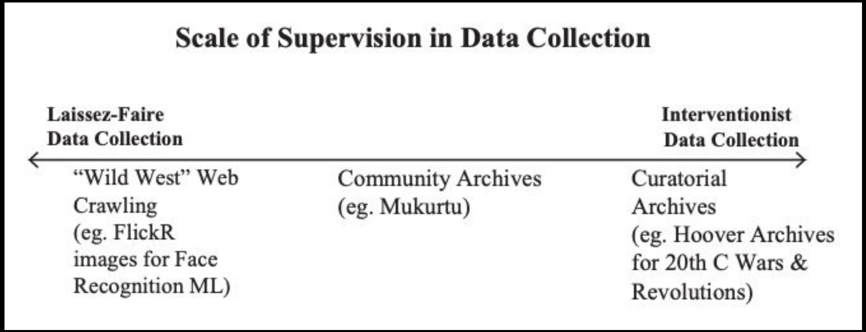
THIS DATA SET IS PROVIDED "AS IS" AND WITHOUT ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, WITHOUT LIMITATION, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE. The images provided above may have certain copyright issues. We take no guarantees or responsibilities, whatsoever, arising out of any copyright issue. Use at your own risk.

Yet, data is highly valued...



Fei-Fei Li (2017). Where have we been? Where are we going?

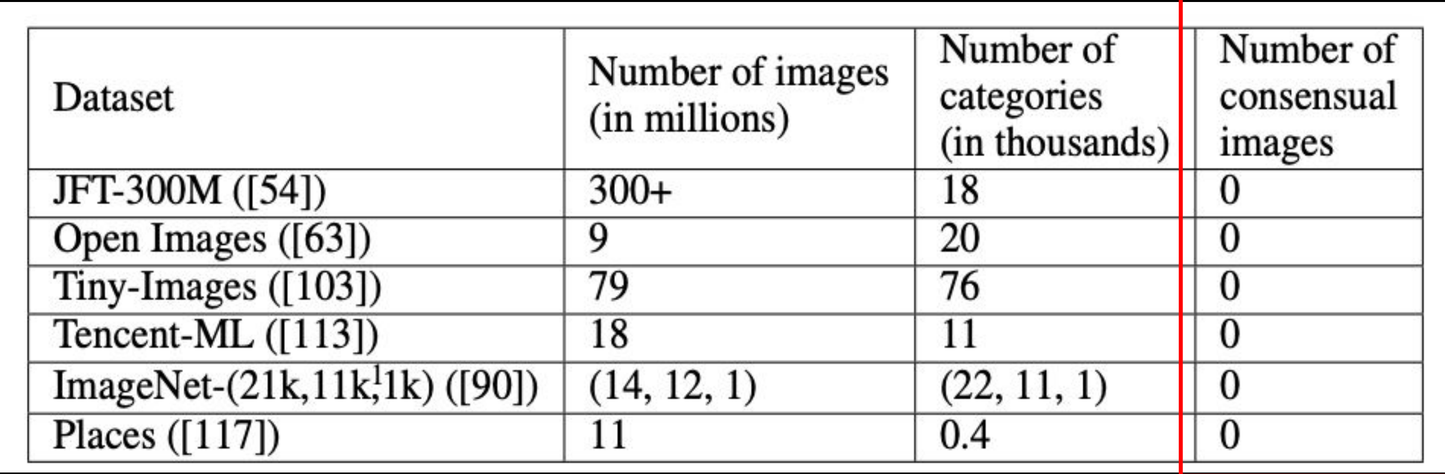
Dataset development characterized by a laissez-faire attitude



"If it's available to us, we ingest it."
Holstein et al. (2019)

Jo & Gebru (2020). Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning
Holstein et al. (2019). Improving fairness in machine learning systems: What do industry practitioners need?

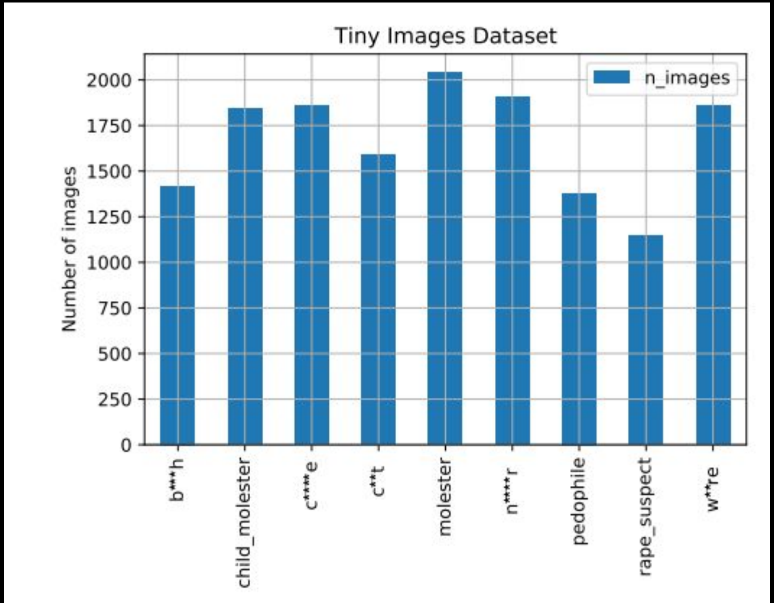
Scale at expense of care for data subjects



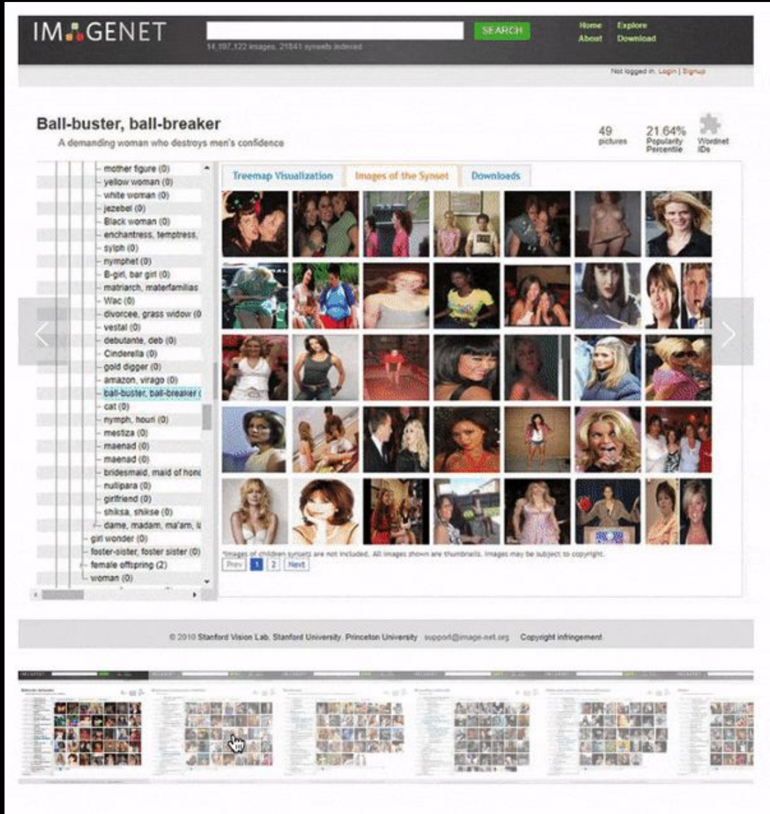
Dataset	Number of images (in millions)	Number of categories (in thousands)	Number of consensual images
JFT-300M ([54])	300+	18	0
Open Images ([63])	9	20	0
Tiny-Images ([103])	79	76	0
Tencent-ML ([113])	18	11	0
ImageNet-(21k,11k[1]1k) ([90])	(14, 12, 1)	(22, 11, 1)	0
Places ([117])	11	0.4	0

Prabhu & Birhane (2020). Large image datasets: A pyrrhic win for computer vision?

Scale at expense of careful curation



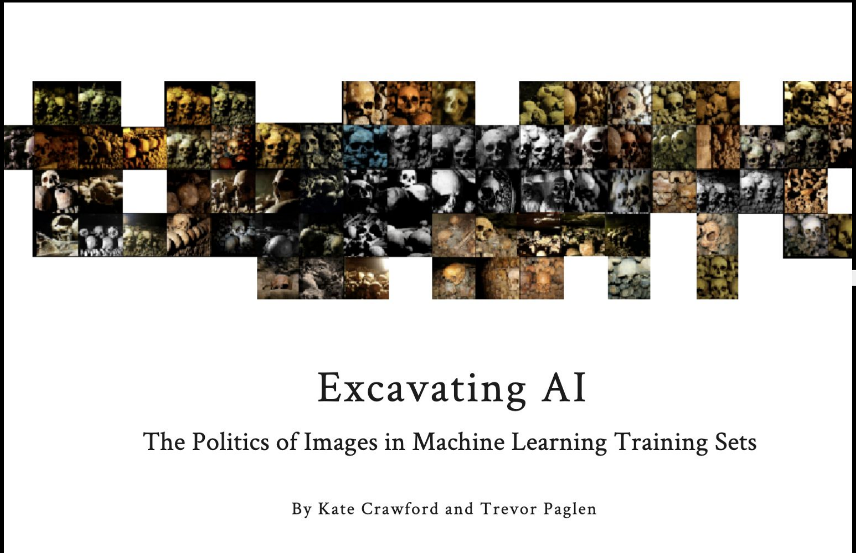
Prabhu & Birhane (2020). [Large image datasets: A pyrrhic win for computer vision?]()



Selections from the "Person" classes, ImageNet

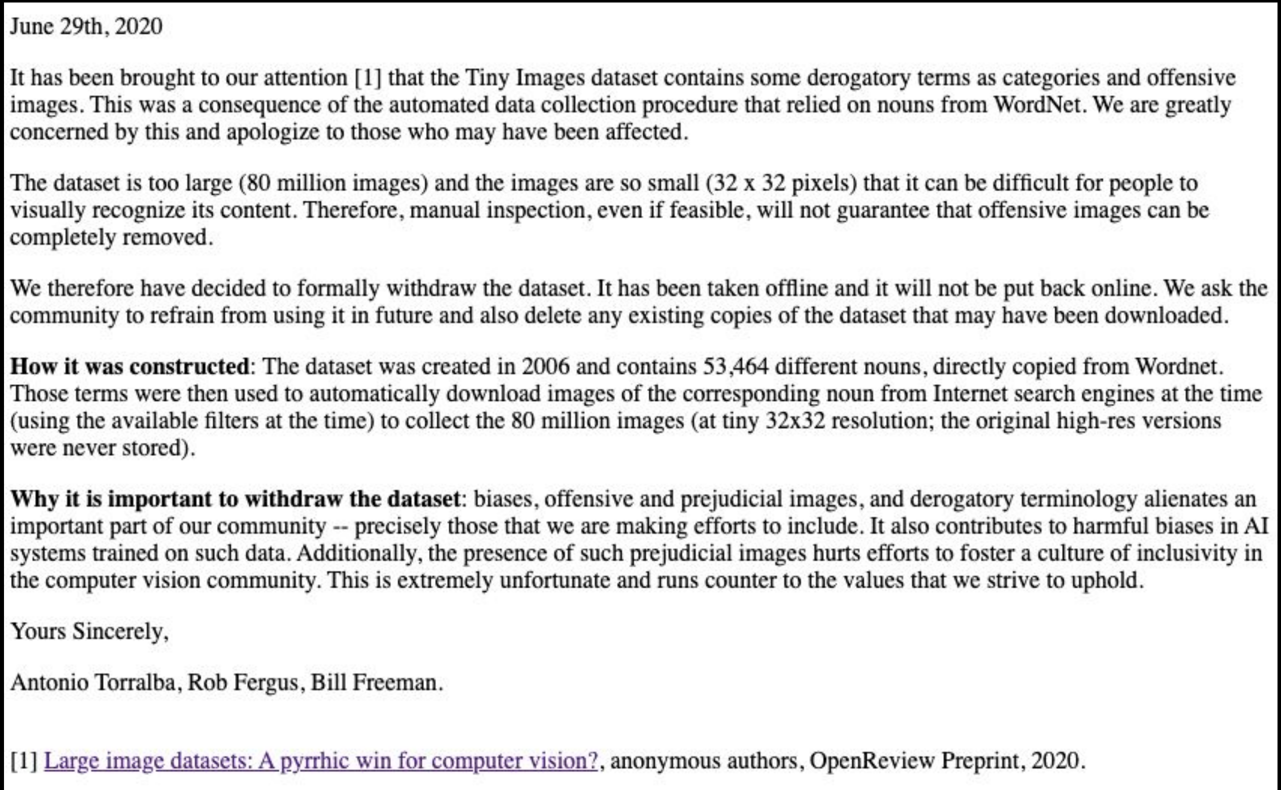
Crawford and Paglen. 2019. [excavating.ai]()

Scale at expense of careful curation



Removal of “non-imageable” categories

Post-hoc fixes



June 29th, 2020

It has been brought to our attention [1] that the Tiny Images dataset contains some derogatory terms as categories and offensive images. This was a consequence of the automated data collection procedure that relied on nouns from WordNet. We are greatly concerned by this and apologize to those who may have been affected.

The dataset is too large (80 million images) and the images are so small (32 x 32 pixels) that it can be difficult for people to visually recognize its content. Therefore, manual inspection, even if feasible, will not guarantee that offensive images can be completely removed.

We therefore have decided to formally withdraw the dataset. It has been taken offline and it will not be put back online. We ask the community to refrain from using it in future and also delete any existing copies of the dataset that may have been downloaded.

How it was constructed: The dataset was created in 2006 and contains 53,464 different nouns, directly copied from Wordnet. Those terms were then used to automatically download images of the corresponding noun from Internet search engines at the time (using the available filters at the time) to collect the 80 million images (at tiny 32x32 resolution; the original high-res versions were never stored).

Why it is important to withdraw the dataset: biases, offensive and prejudicial images, and derogatory terminology alienates an important part of our community -- precisely those that we are making efforts to include. It also contributes to harmful biases in AI systems trained on such data. Additionally, the presence of such prejudicial images hurts efforts to foster a culture of inclusivity in the computer vision community. This is extremely unfortunate and runs counter to the values that we strive to uphold.

Yours Sincerely,

Antonio Torralba, Rob Fergus, Bill Freeman.

[1] Large image datasets: A pyrrhic win for computer vision?, anonymous authors, OpenReview Preprint, 2020.

Post-hoc fixes

Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy

Kaiyu Yang
Princeton University
Princeton, NJ
kaiyuy@cs.princeton.edu

Klint Qinami
Princeton University
Princeton, NJ
kqinami@cs.princeton.edu

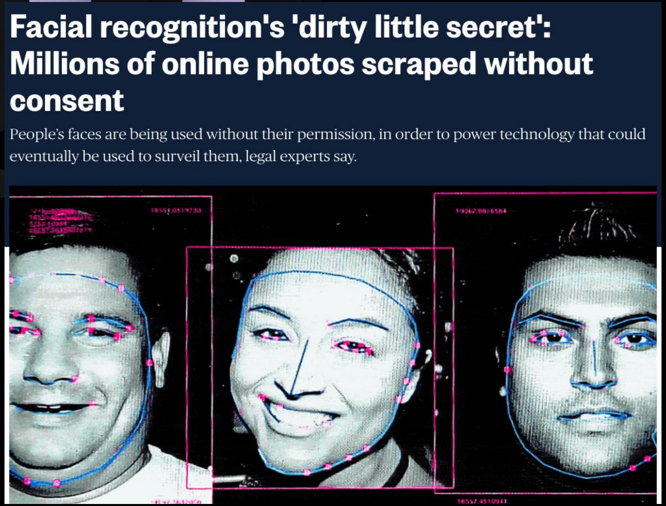
Li Fei-Fei
Stanford University
Stanford, CA
feifeili@cs.stanford.edu

Jia Deng
Princeton University
Princeton, NJ
jiadeng@cs.princeton.edu

Olga Russakovsky
Princeton University
Princeton, NJ
olgarus@cs.princeton.edu

Discourses of scale permeate algorithmic fairness

Discourses of scale permeate algorithmic fairness



More data isn't always the solution

"Failures of data-driven systems are not located exclusively at the level of those who are represented or underrepresented in the dataset"

- Denton & Hanna et al (2020)

ISSUE 12 / COMMONS
DECEMBER 20, 2020

Image by Celine Nguyen.

Lines of Sight

Alex Hanna, Emily Denton, Razvan Amironesei, Andrew Smart, Hilary Nicole

Bringing the People Back In: Contesting Benchmark Machine Learning Datasets

Emily Denton [*1] Alex Hanna [*1] Razvan Amironesei [2] Andrew Smart [1] Hilary Nicole [1]
Morgan Klaus Scheuerman [1]

More data isn't always the solution

"More focus should be placed on the redistribution of power, rather than just on including underrepresented groups"

The Limits of Global Inclusion in AI Development

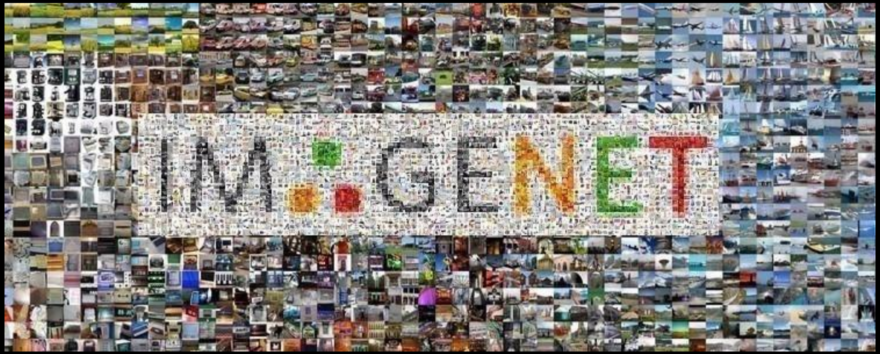
Alan Chan, [1] * **Chinasa T. Okolo,** [2] * **Zachary Terner,** [3] * **Angelina Wang** [4] *
[1]Mila, Université de Montréal [2]Cornell University
[3]National Institute of Statistical Sciences [4]Princeton University
alan.chan@mila.quebec, chinasa@cs.cornell.edu, zterner@niss.org, angelina.wang@princeton.edu

Data is always laden with subjective values, judgements, & imperatives

Data is always always socially and culturally situated (Gitelman, 2013; Elish and boyd, 2017)

This is inescapable

Data is always laden with subjective values, judgements, & imperatives



http://www.image-net.org

ImageNet categories → WordNet

ImageNet images → Snapshot of the internet from 2010

ImageNet annotations → Amazon MTurk crowdsourced annotations

“To produce a dataset at ‘the scale of the web’ implies to impose a particular way of seeing images, of pointing and naming. “
-- Nicolas Malevé

Hammerhead shark → Scientific object

Trout → Dead trophy

Lobster → Food

Malevé (2019). An Introduction to Image Datasets

Decontextualized data

- Data contexts are often lost / unaccounted for
  - Annotator demographics
  - Contexts of image capture
  - Design decisions
  - Intended contexts of use
  - …

- SOTA-chasing practices further position them as bars to jump over
  - But benchmark datasets don’t provide value-neutral markers of progress (Rotan & Milli, 2020; Prabhu & Birhane, 2020)

Moving forward: Recommendations for responsible dataset development

Individual actions & community change

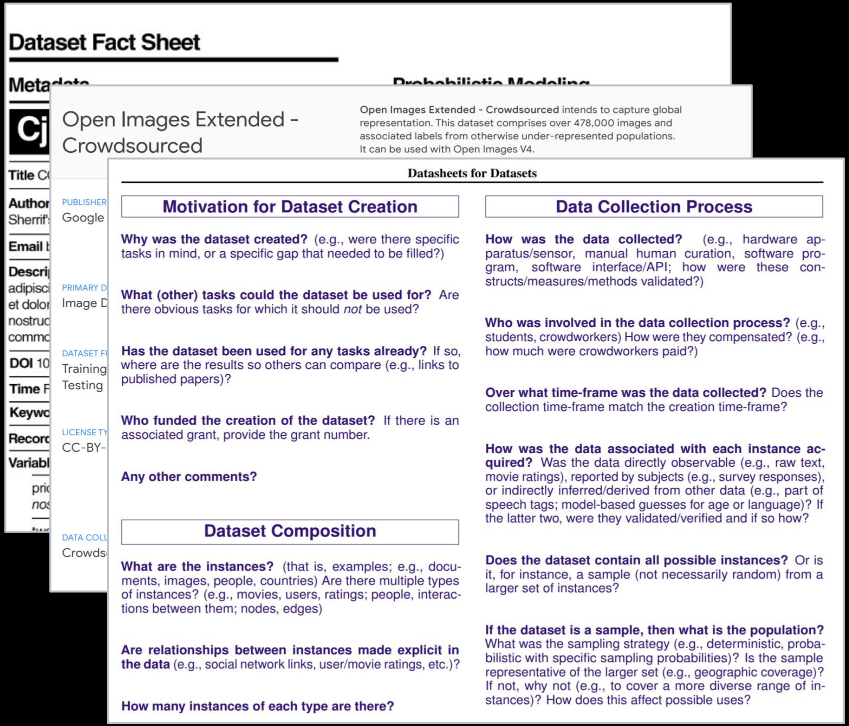
Data documentation frameworks

Standardized framework for transparent dataset documentation

Dataset creators:

Reflect on on process of creation, distribution, and maintenance
Making explicit any underlying assumptions
Outline potential risks or harms, and implications of use

Dataset consumers:
Provide information to facilitate informed decision making



Datasheets for Datasets

Motivation for Dataset Creation

Why was the dataset created? (e.g., were there specific tasks in mind, or a specific gap that needed to be filled?)

What (other) tasks could the dataset be used for? Are there obvious tasks for which it should *not* be used?

Has the dataset been used for any tasks already? If so, where are the results so others can compare (e.g., links to published papers)?

Who funded the creation of the dataset? If there is an associated grant, provide the grant number.

Any other comments?

Dataset Composition

What are the instances? (that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

Are relationships between instances made explicit in the data (e.g., social network links, user/movie ratings, etc.)?

How many instances of each type are there?

Data Collection Process

How was the data collected? (e.g., hardware apparatus/sensor, manual human curation, software program, software interface/API; how were these constructs/measures/methods validated?)

Who was involved in the data collection process? (e.g., students, crowdworkers) How were they compensated? (e.g., how much were crowdworkers paid?)

Over what time-frame was the data collected? Does the collection time-frame match the creation time-frame?

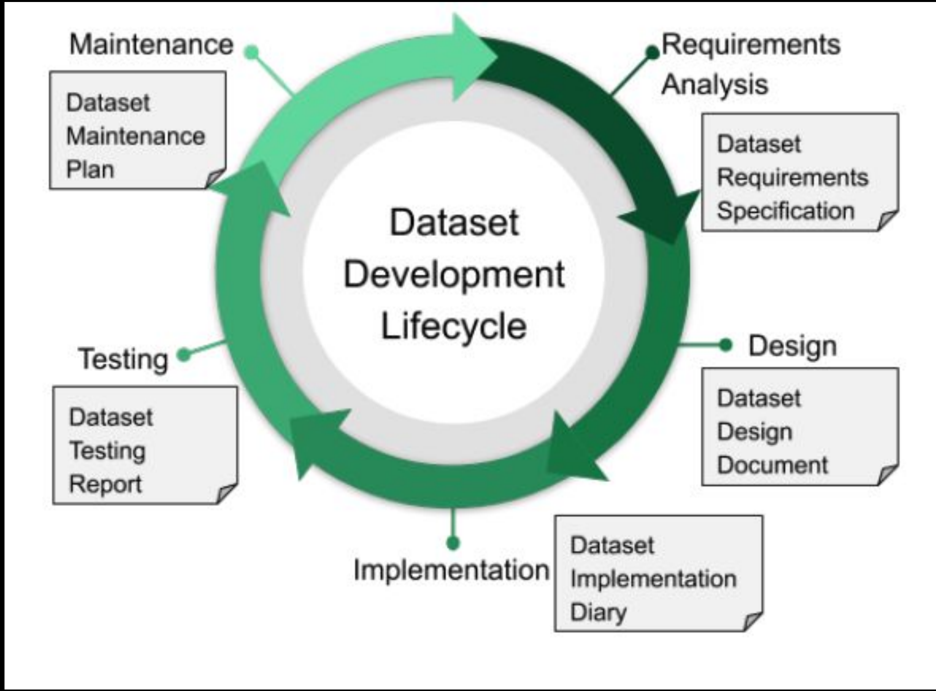
How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part of speech tags; model-based guesses for age or language)? If the latter two, were they validated/verified and if so how?

Does the dataset contain all possible instances? Or is it, for instance, a sample (not necessarily random) from a larger set of instances?

If the dataset is a sample, then what is the population? What was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Is the sample representative of the larger set (e.g., geographic coverage)? If not, why not (e.g., to cover a more diverse range of instances)? How does this affect possible uses?

Gebru, et al. (2018). Datasheets for datasets
Holland et al. (2018). The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards
Bender and Friedman (2018). Data Statements for NLP: Toward Mitigating System Bias and Enabling Better Science
Hutchinson et al. (2020). Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure.

Accountability mechanisms



- Documentation framework for each stage of the data development lifecycle
- Makes visible the value and necessity of careful data work and the often overlooked work and decisions that go into dataset creation
- Facilitates informed decision making at every stage

Hutchinson et al (2021). [Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure]

Dataset maintenance

“If you can't afford to maintain a dataset, maybe you can’t afford to build it”
-- Hutchinson et al. (2021)

Need community-wide investment in dataset maintenance infrastructure

Ethical oversight mechanisms

- Most methods of dataset collection fall outside the scope of existing ethical oversight frameworks (Metcalf & Crawford, 2016)
- Need to develop our own ethical oversight frameworks that provides mechanisms of legal and professional accountability

Category 4 – Secondary Research Uses of Identifiable Private Information or Identifiable Biospecimens

Secondary research with identifiable Information/specimens collected for some other initial activity, if **ONE of following**:

- Biospecimens or information is publically available
- Information recorded so subject cannot readily be identified (directly or indirectly/linked); investigator does not contact subjects and will not re-identify the subjects
- Collection and analysis involving Investigators Use of identifiable health information when use is regulated by HIPAA "health care operations" or "research" or "public health activities and purposes"
- Research information collected by or on behalf of federal government using government generated or collected information obtained for non-research activities

Ethical oversight mechanisms

- Conferences like CVPR can play a role advancing these efforts, following other communities:
  - NeurIPS ethics guidelines
  - ACL ethical review (see also Bender (2021))
  - Workshops focused on Navigating the Broader Impacts of AI Research

Recognize limits of datasets as measurement devices

- Be sensitive to the gaps between what a dataset represents and the real world task or phenomenon its approximating
    - Be careful with claims that are made about SOTA performance on the dataset (Bender & Koller, 2020)

- Standard benchmark metrics provide *one* way of evaluating methods -- consider others in addition (Ethayarajh & Jurafsky, 2020; Dodge et al. 2019; Mitchell et al. 2019)

Understand your datasets

- Identifying spurious cues, dataset artifacts that could be easily gamed by a model, labelling errors, edge cases, etc. (Sakaguchi et al., 2020; Swayamdipta et al., 2020)

- Dataset audits (e.g. Prabhu & Birhane, 2020) have led to the removal of entire datasets (e.g. TinyImages)

- Diversifying / balancing datasets for along sociodemographic lines (e.g. Yang et al., 2020, Merler et al., 2019)

Understand your datasets

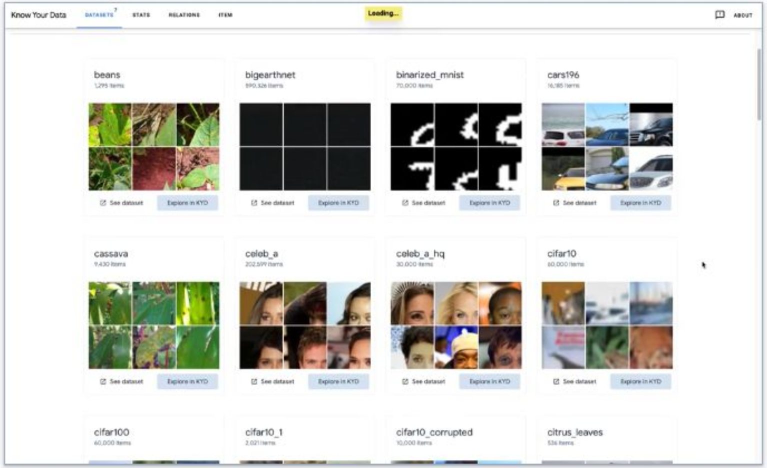
Know Your Data helps researchers, engineers, product teams, and decision makers **understand datasets** with the goal of **improving data quality**, and helping **mitigate fairness and bias issues.**

Explore Know Your Data

Explore 70+ ML datasets

Interactively explore image datasets from the TensorFlow Datasets catalog

Try it on TensorFlow Datasets →



Value data work & recognize it as a specialty

- As a community we need to shift educational practices and incentive structures so that careful, intentional, equitable dataset construction is valued

- Data work is inherently interdisciplinary -- need new pedagogies within the field

- Can shift incentive structures through conferences like CVPR
  - E.g. NeurIPS Dataset & Benchmark Track
  - **This workshop!**

Thanks!

Thanks to collaborators: Alex Hanna, Morgan Klaus Scheuerman, Razvan Amironesei, Andrew Smart, Hilary Nicole, Ben Hutchinson, Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Timnit Gebru, Meg Mitchell.

Beyond Fairness: Towards a Just, Equitable, and Accountable Computer Vision

Friday June 25

Website: https://sites.google.com/view/beyond-fairness-cv
Discord server: https://discord.gg/CkuGyf8CS7