COACTIVE

Tackling Socioeconomic Bias in Machine Learning

Cody Coleman, PhD
CEO & Co-Founder | Coactive AI

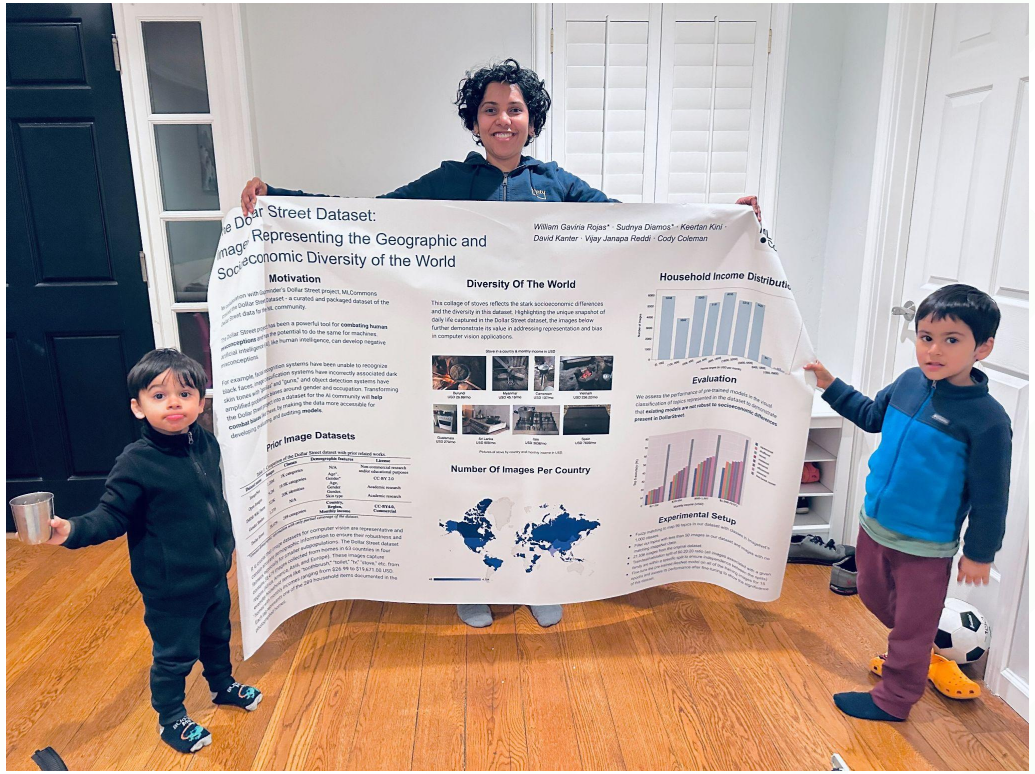
A collaboration between researchers from

COACTIVE

ML Commons

The Team

This work was an amazing collaboration across, academia, and industry. Thank you!



William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, **Cody Coleman**

COACTIVE

CAPMINDER

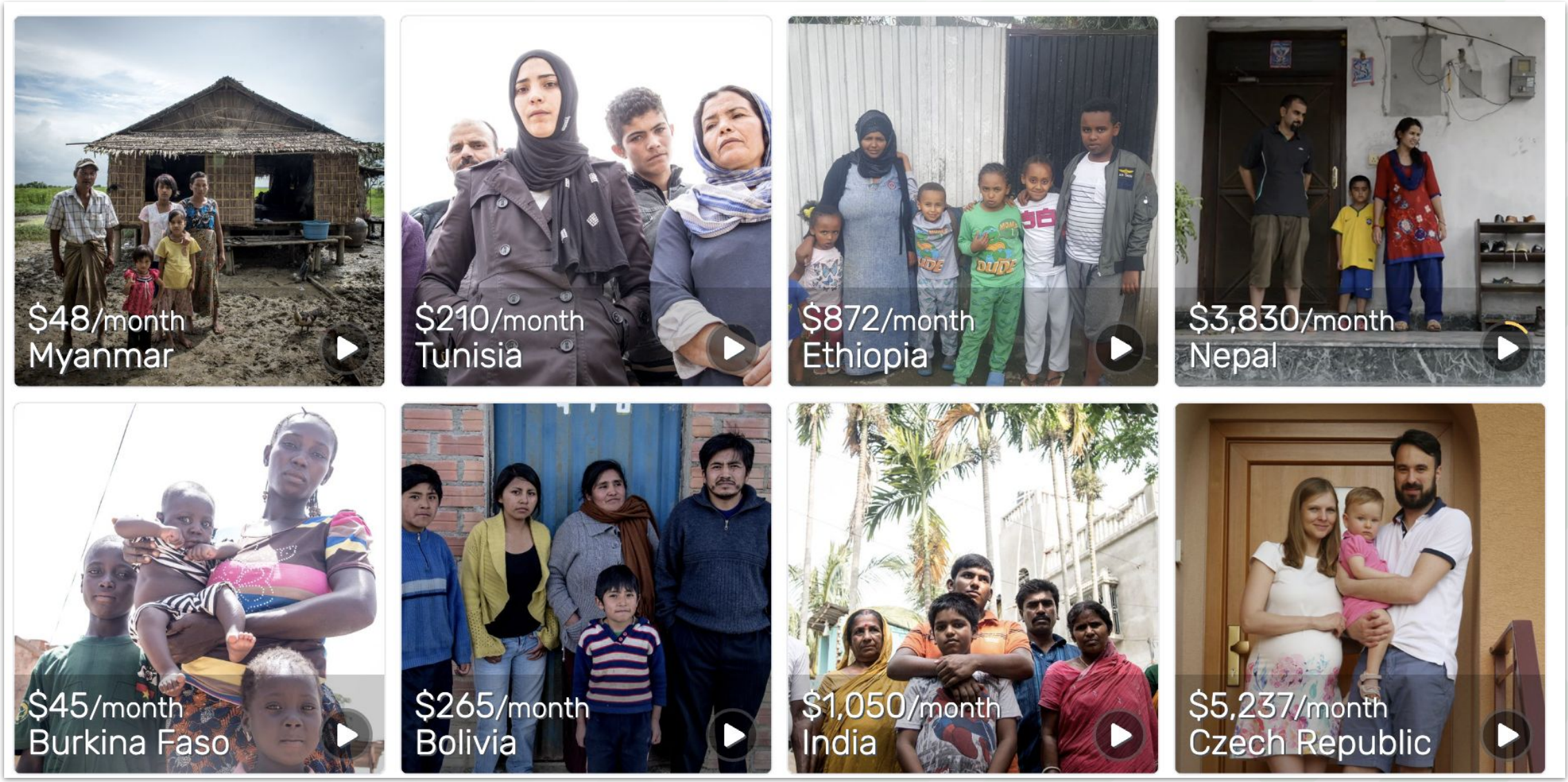
ML Commons

HARVARD UNIVERSITY

COACTIVE

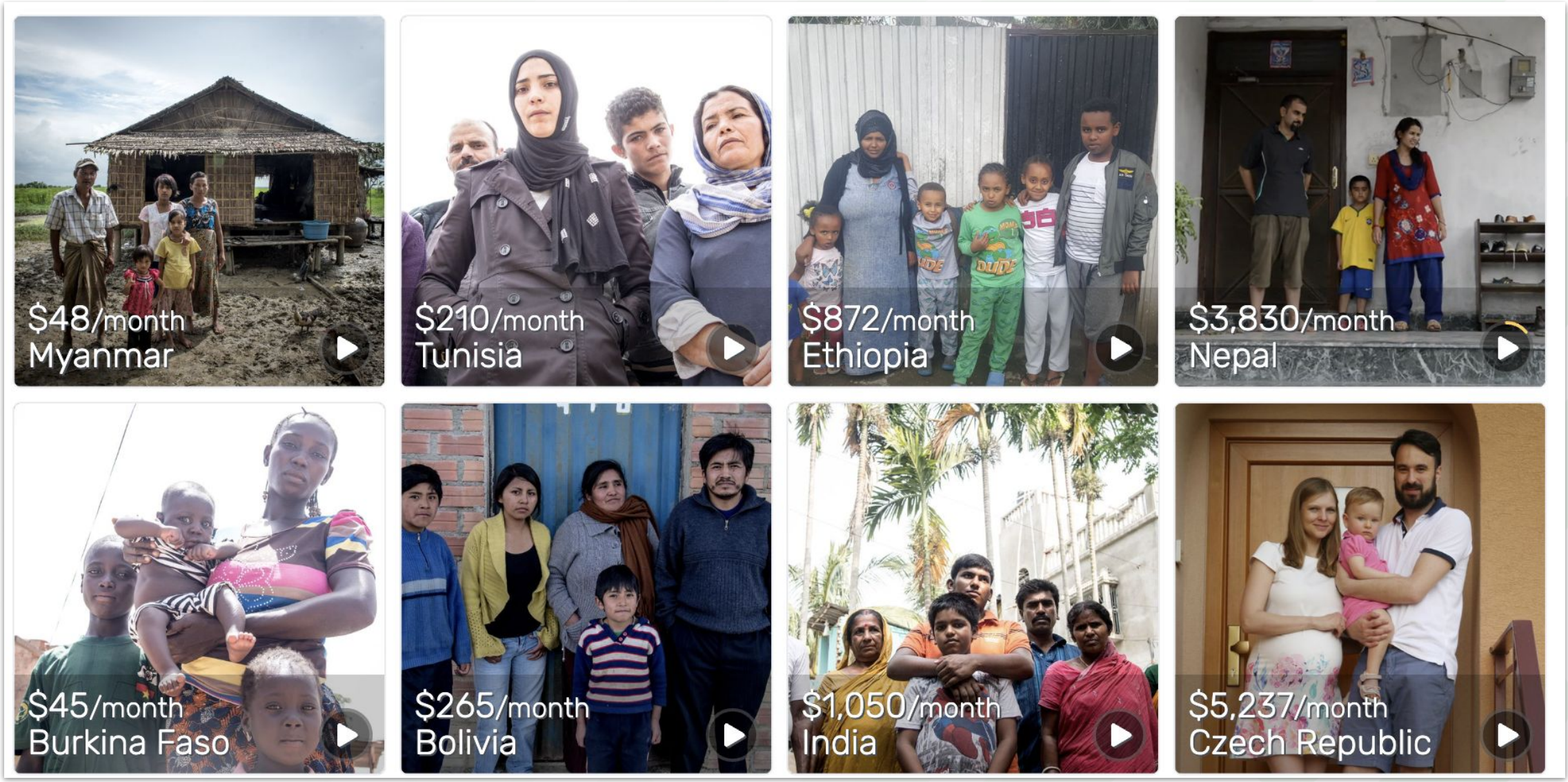
The Dollar Street Project

Combating human bias through photos as data across incomes and borders



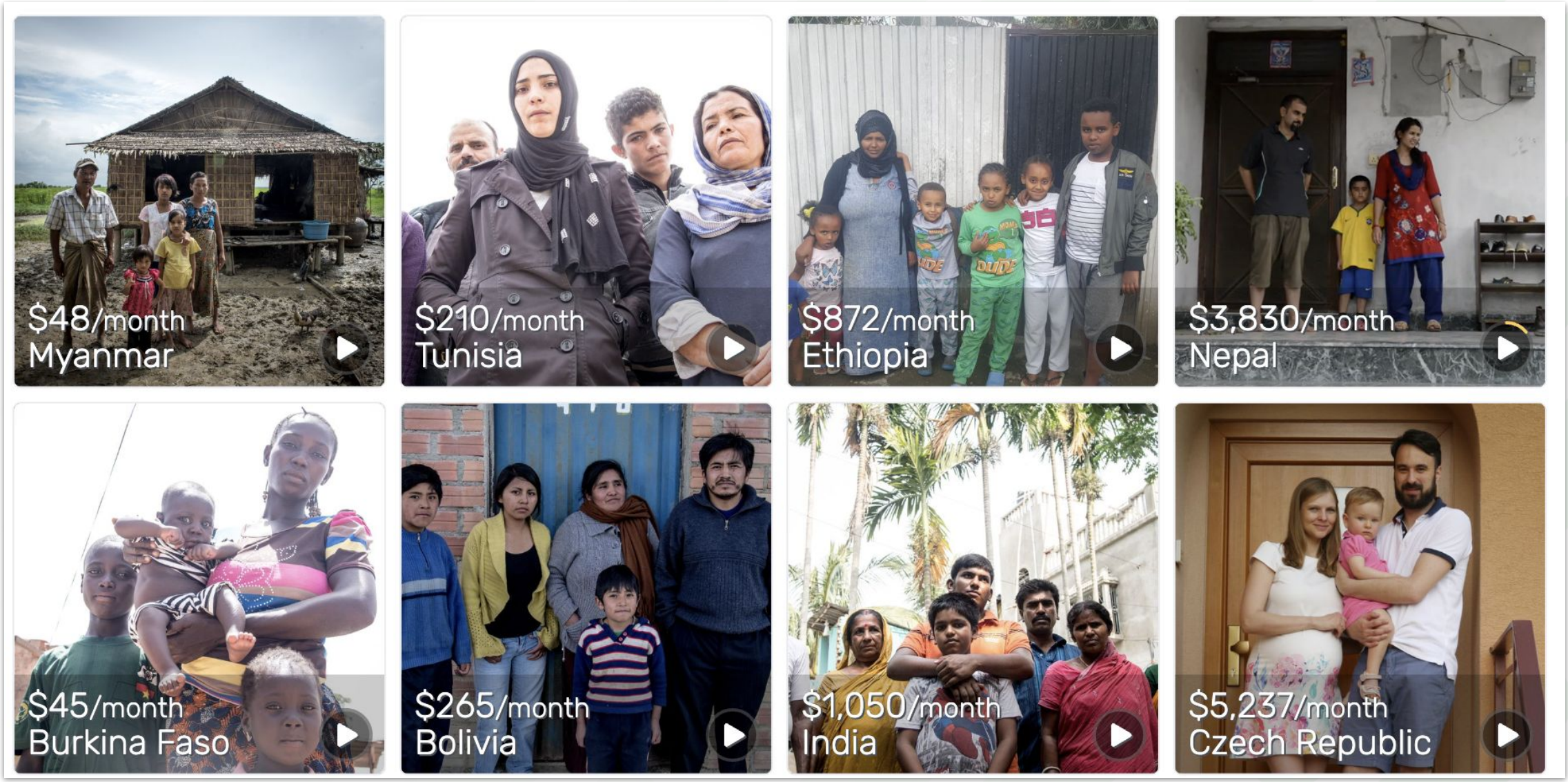
The Dollar Street Project

Combating **human** bias through photos as data across incomes and borders

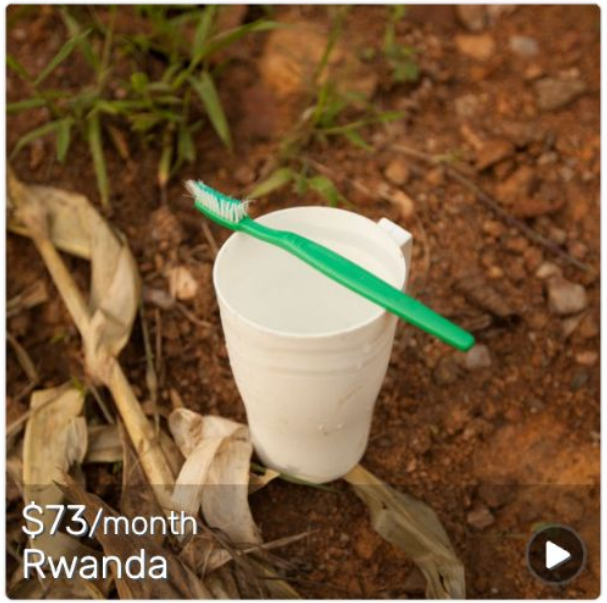


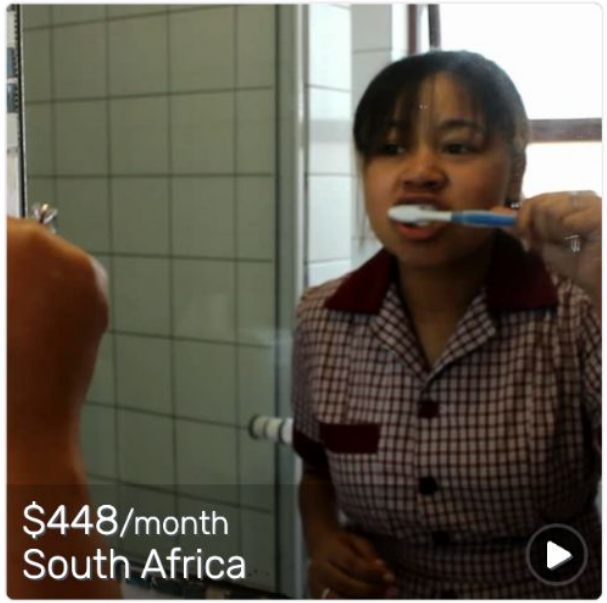
The Dollar Street Project

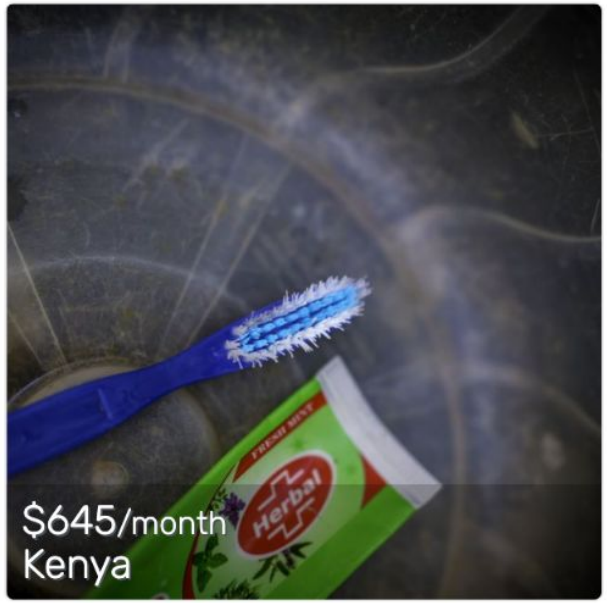
Combating **human** bias through photos as data across incomes and borders

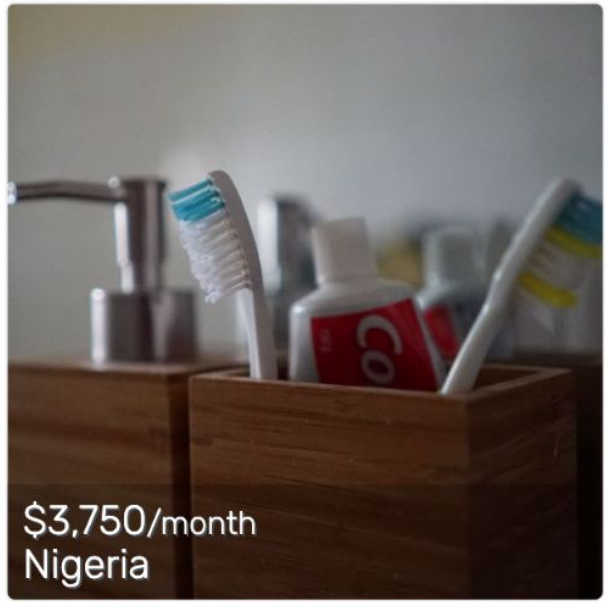


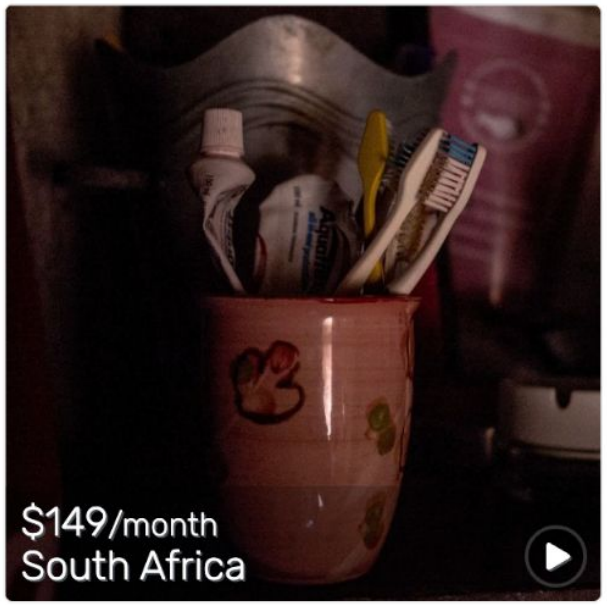
COACTIVE


$73/month
Rwanda

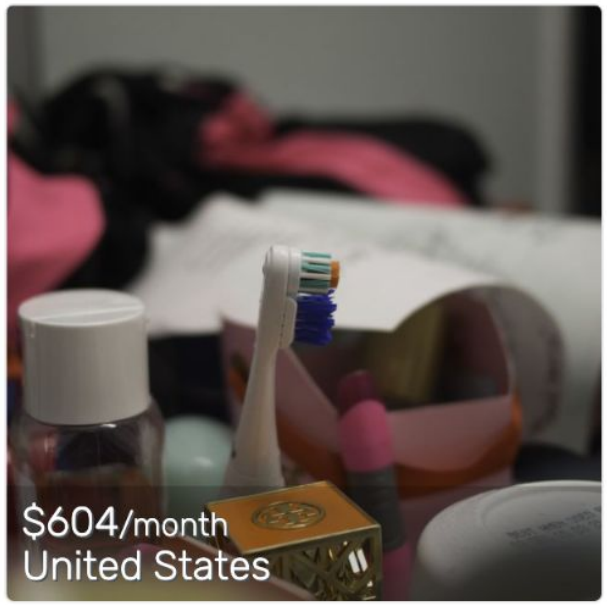

$448/month
South Africa

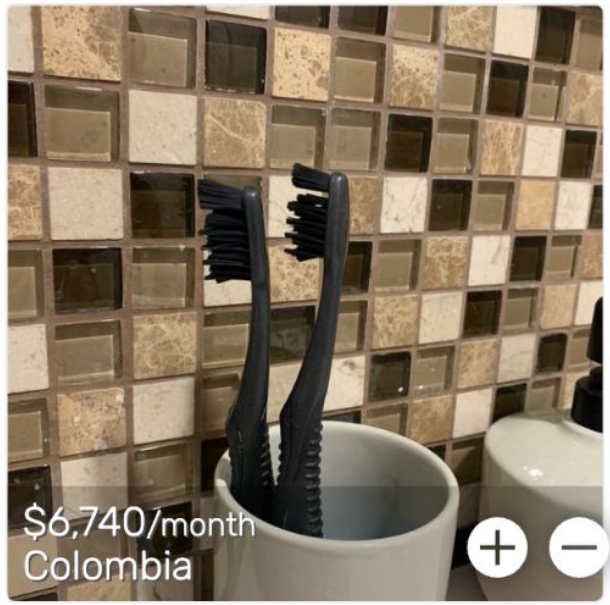

$645/month
Kenya


$3,750/month
Nigeria


$149/month
South Africa

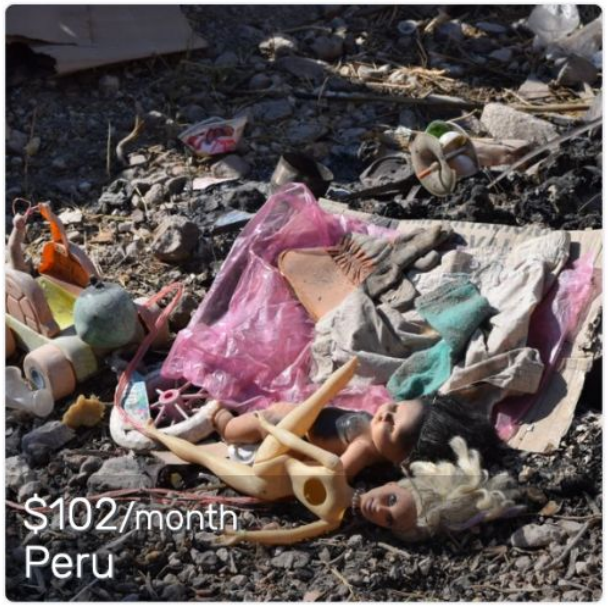

$556/month
South Africa


$604/month
United States


$6,740/month
Colombia

COACTIVE

COACTIVE

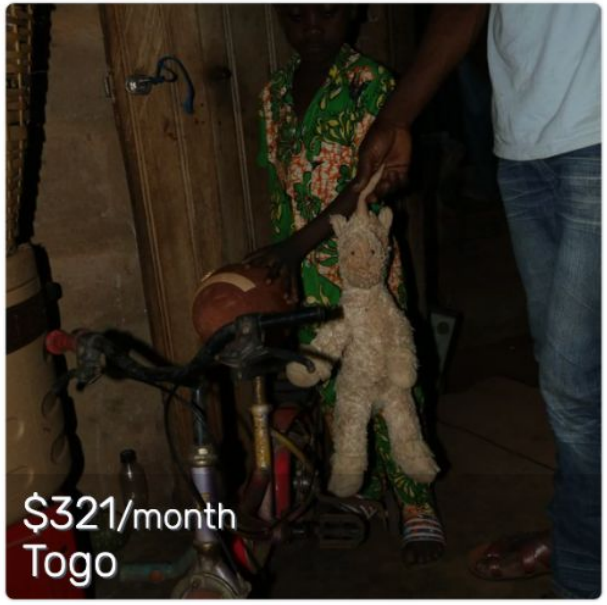

$102/month
Peru


$430/month
Serbia

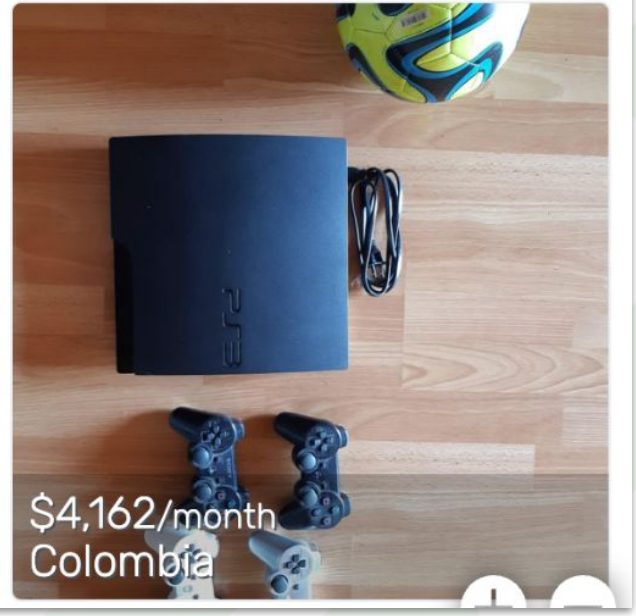

$1,019/month
Romania


$4,553/month
Spain


$96/month
Nepal


$321/month
Togo

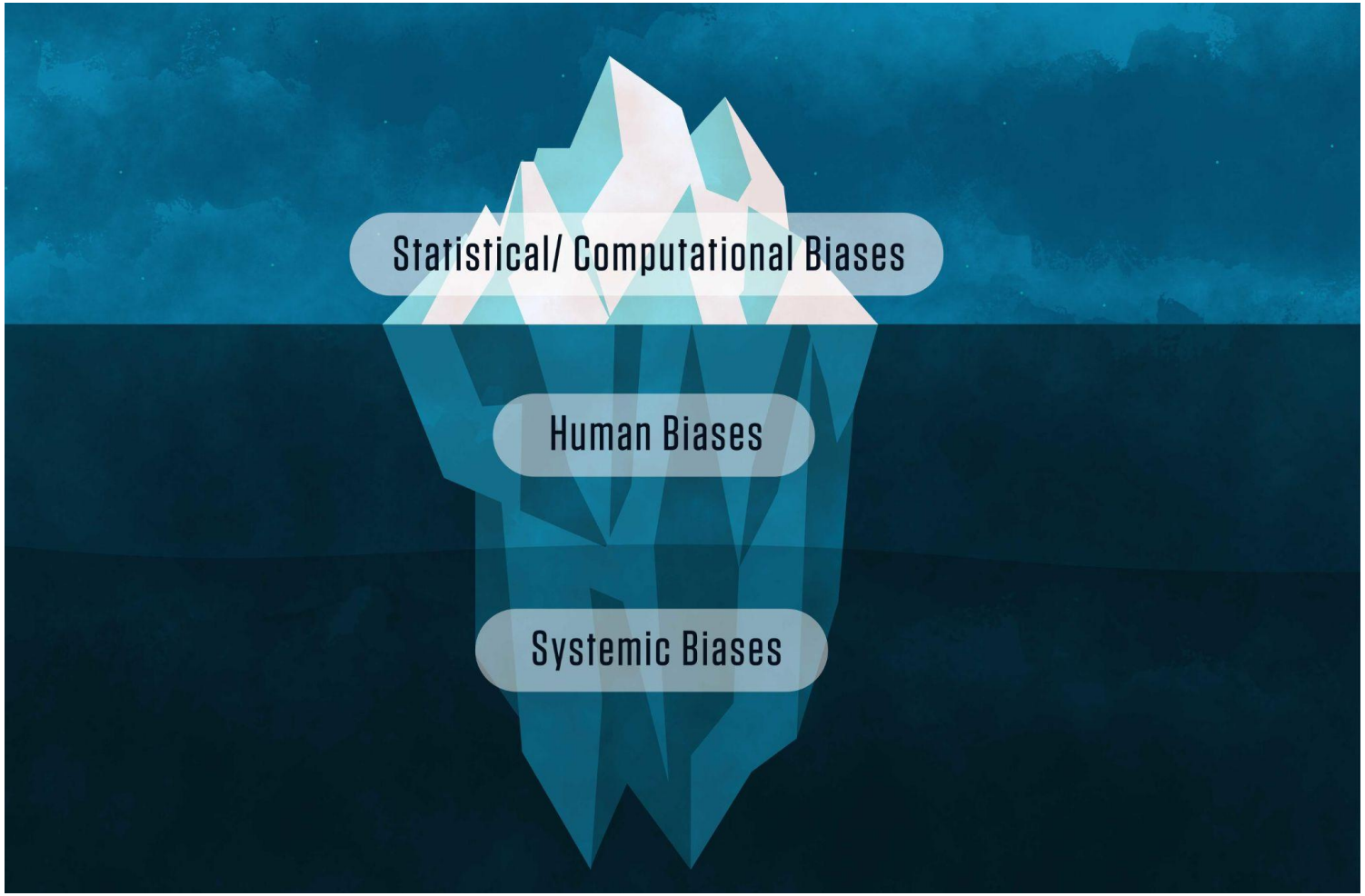

$719/month
Ethiopia


$4,162/month
Colombia

COACTIVE

Human and algorithmic bias are connected

Human biases can get absorbed in AI models and systems

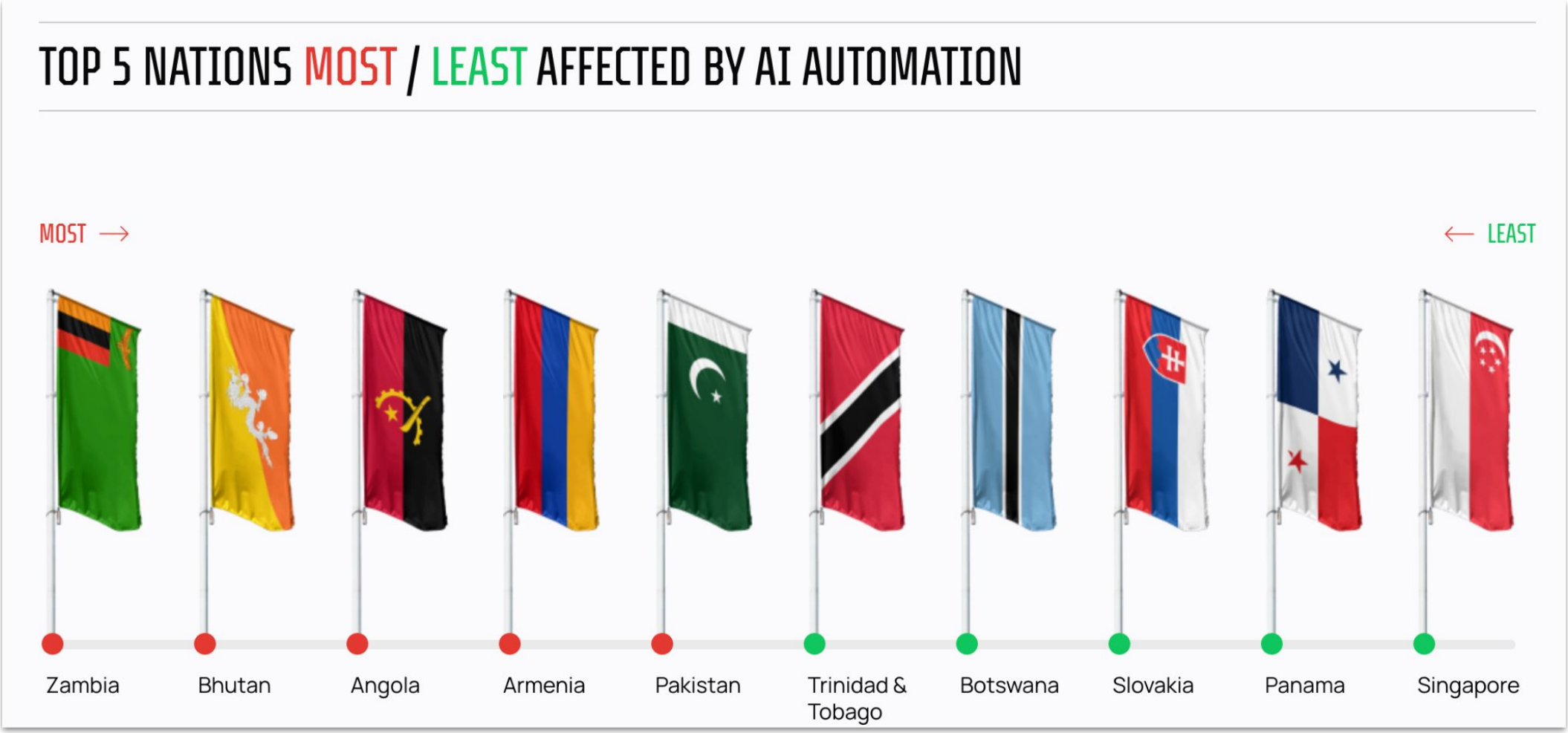


COACTIVE

N. Hanacek/NIST. "There's More to AI Bias Than Biased Data, NIST Report Highlights." https://www.nist.gov/news-events/news/2022/03/theres-more-ai-bias-biased-data-nist-report-highlights (2022).

AI affects all of us

Automation and technology proliferation concerns worldwide

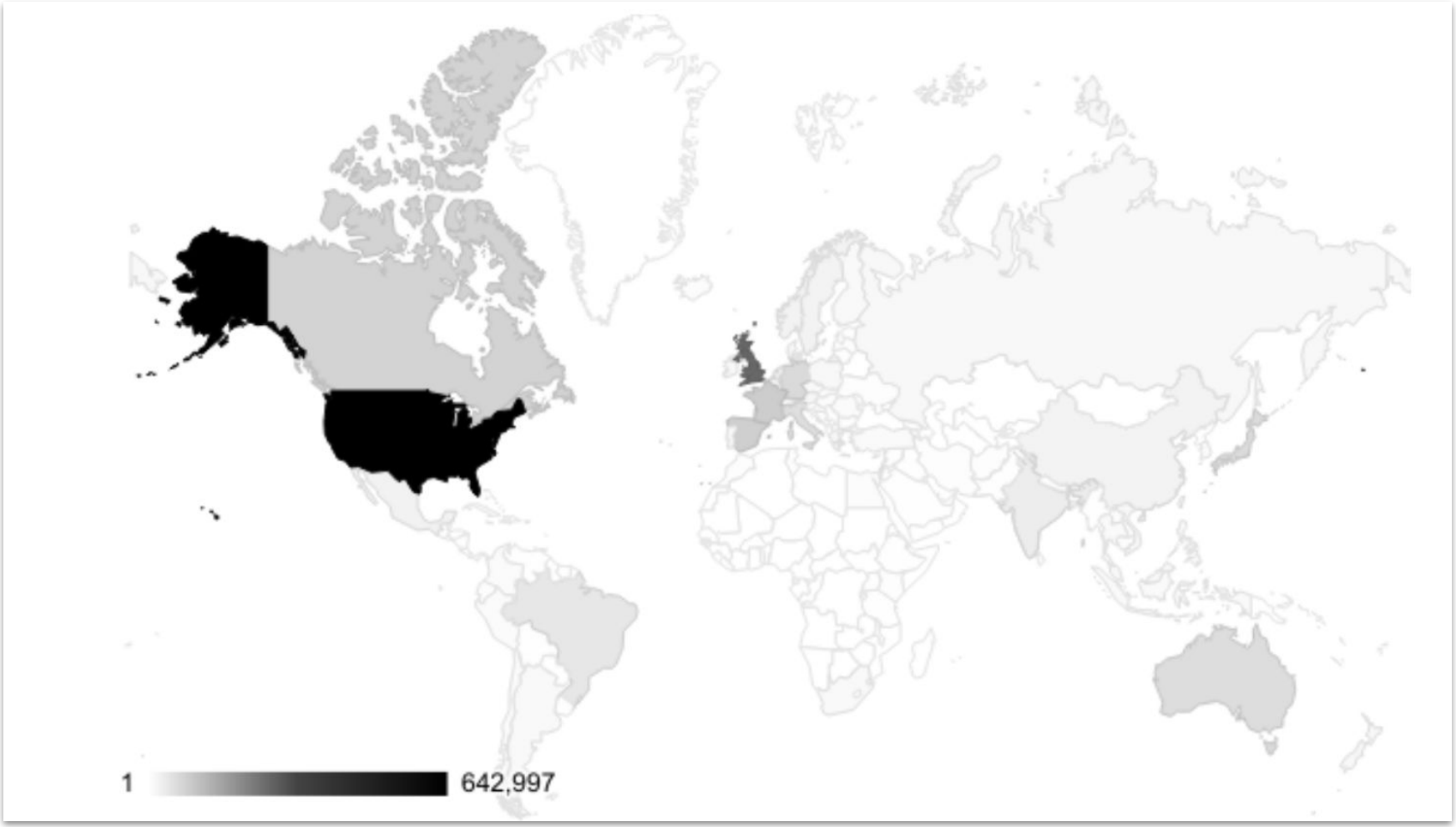


COACTIVE

https://www.bizreport.com/business/countries-most-affected-by-ai-replace-jobs

Yet AI training data is not often representative

Training data sourced from the internet: 33% from US, 60% from six countries

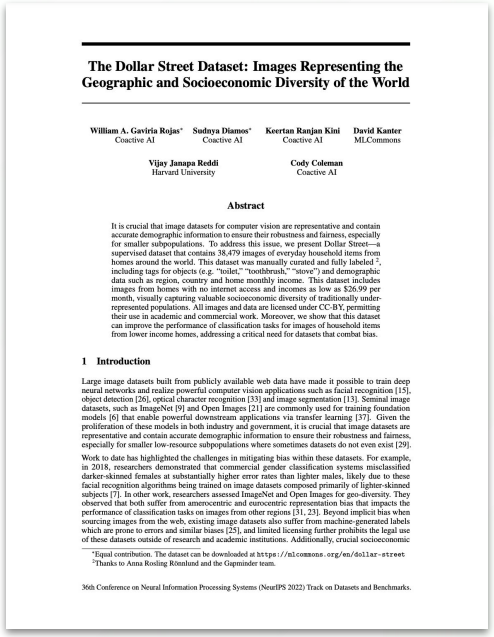


COACTIVE

Shankar, Shreya, et al. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world." arXiv preprint arXiv:1711.08536 (2017).

The Dollar Street dataset for AI training

Combating AI training bias through photos as data across incomes and borders



The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World

William A. Gaviria Rojas* Coactive AI
Sudnya Diamos* Coactive AI
Keertan Ranjan Kini Coactive AI
David Kanter MLCommons
Vijay Janapa Reddi Harvard University
Cody Coleman Coactive AI

Abstract

It is crucial that image datasets for computer vision are representative and contain accurate demographic information to ensure their robustness and fairness, especially for smaller subpopulations. To address this issue, we present Dollar Street—a supervised dataset that contains 38,479 images of everyday household items from homes around the world. This dataset was manually curated and fully labeled [2], including tags for objects (e.g. "toilet," "toothbrush," "stove") and demographic data such as region, country and home monthly income. This dataset includes images from homes with no internet access and incomes as low as $26.99 per month, visually capturing valuable socioeconomic diversity of traditionally under-represented populations. All images and data are licensed under CC-BY, permitting their use in academic and commercial work. Moreover, we show that this dataset can improve the performance of classification tasks for images of household items from lower income homes, addressing a critical need for datasets that combat bias.

1 Introduction

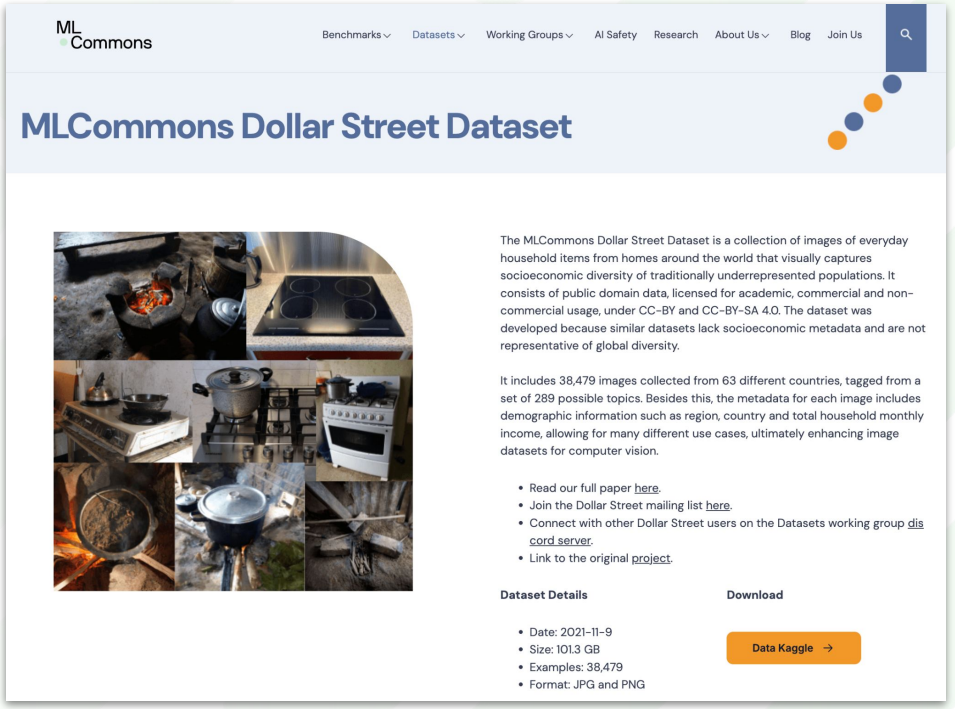
Large image datasets built from publicly available web data have made it possible to train deep neural networks and realize powerful computer vision applications such as facial recognition [15], object detection [26], optical character recognition [33] and image segmentation [13]. Seminal image datasets, such as ImageNet [9] and Open Images [21] are commonly used for training foundation models [6] that enable powerful downstream applications via transfer learning [37]. Given the proliferation of these models in both industry and government, it is crucial that image datasets are representative and contain accurate demographic information to ensure their robustness and fairness, especially for smaller low-resource subpopulations where sometimes datasets do not even exist [29].

Work to date has highlighted the challenges in mitigating bias within these datasets. For example, in 2018, researchers demonstrated that commercial gender classification systems misclassified darker-skinned females at substantially higher error rates than lighter males, likely due to these facial recognition algorithms being trained on image datasets composed primarily of lighter-skinned subjects [7]. In other work, researchers assessed ImageNet and Open Images for geo-diversity. They observed that both suffer from americentric and eurocentric representation bias that impacts the performance of classification tasks on images from other regions [31, 23]. Beyond implicit bias when sourcing images from the web, existing image datasets also suffer from machine-generated labels which are prone to errors and similar biases [25], and limited licensing further prohibits the legal use of these datasets outside of research and academic institutions. Additionally, crucial socioeconomic

*Equal contribution. The dataset can be downloaded at https://mlcommons.org/en/dollar-street
[2]Thanks to Anna Rosling Rönnlund and the Gapminder team.

36th Conference on Neural Information Processing Systems (NeurIPS 2022) Track on Datasets and Benchmarks.

Published in top AI academic conference (NeurIPS 2022)



And made freely available to the world through MLCommons, with a CC-BY-4.0 license

COACTIVE

The Dollar Street captures unique geographic and socioeconomic information

38k	289	404	63
Images	Everyday household items	Homes	Countries

- Data was manually collected and verified
- Includes homes with no internet access
- Includes households with incomes as low as $26.99 per month
- All data is licensed under CC-BY 4.0

COACTIVE

The Dollar Street dataset captures visual diversity

Burundi
$26.99 / month

Myanmar
$45.18 / month

Cameroon
$137.00 / month

Vietnam
$236.22 / month

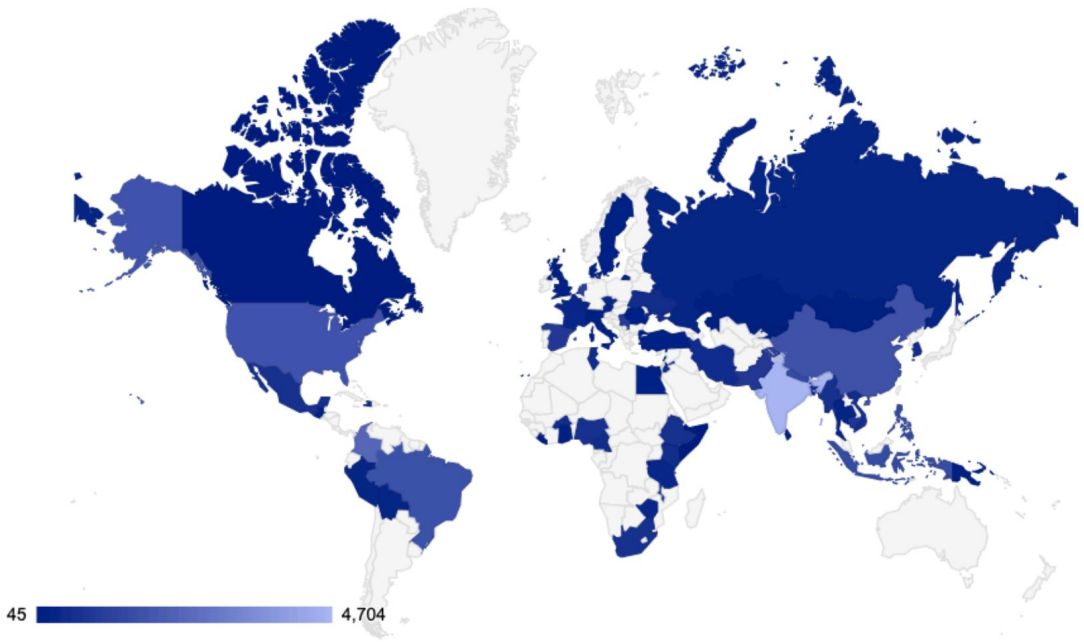
Guatemala
$270.00 / month

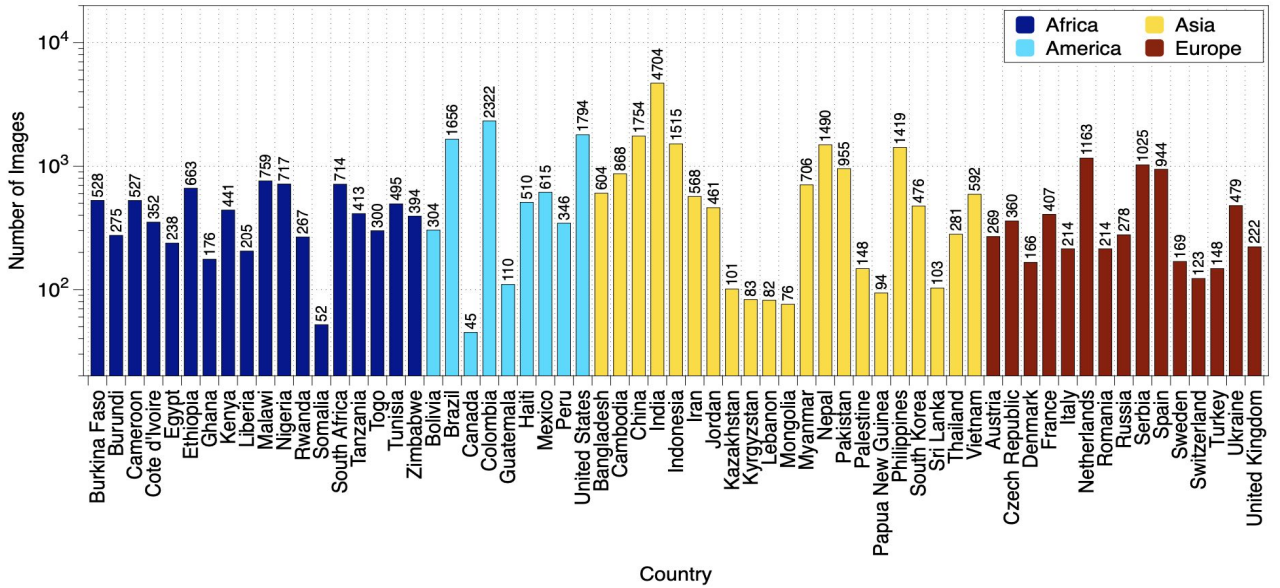
Sri Lanka
$909.00 / month

Italy
$3,636.00 / month

Denmark
$5,343.00 / month

The Dollar Street dataset captures geographical representation

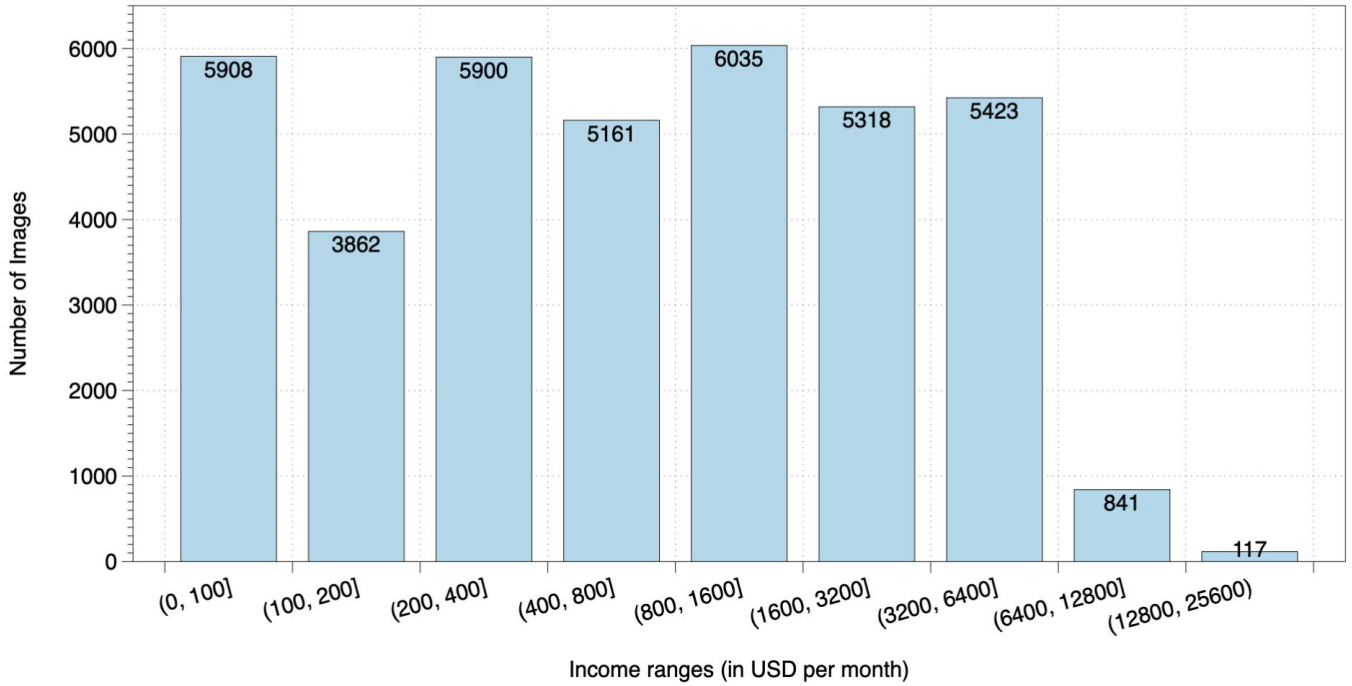




Number of images per country

COACTIVE

The Dollar Street dataset captures socioeconomic representation



Number of images by monthly income

The Dollar Street dataset contains images from homes with a wide range of monthly incomes (in USD).
At a glance: Min = $26.99, Median = $685.00, Max = $19,671.00

COACTIVE

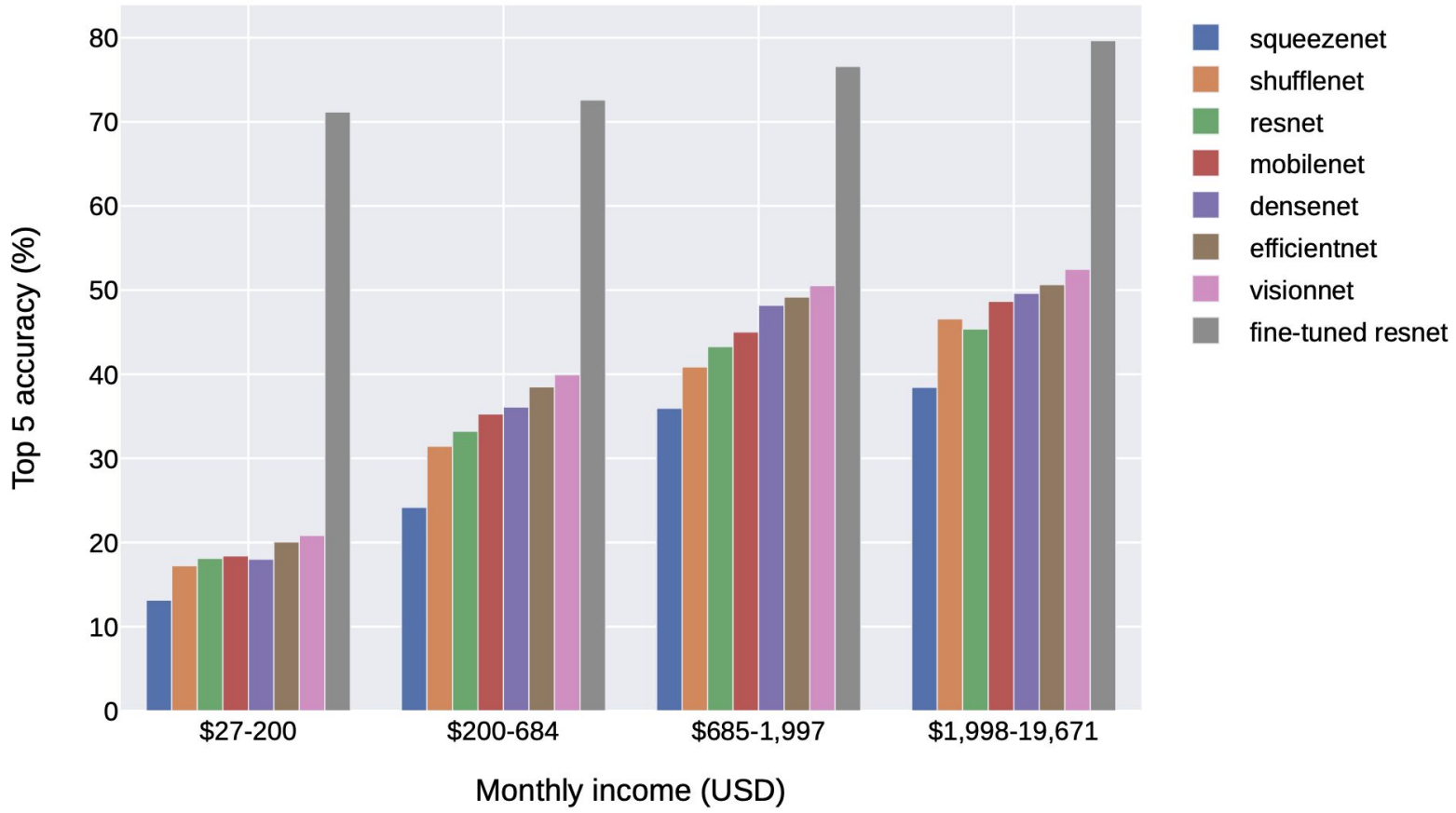
The results: equity in AI performance

Visual classification accuracy improves by +30-50% across various income levels

Key results:
Pre-trained models show **monotonically decreasing classification accuracy** with **decreasing income.**

Fine-tuning using the Dollar Street dataset → substantially performance improvement

The large improvement for the lowest-income quartile highlights the importance of the Dollar Street dataset.



COACTIVE

The results: inspiring others to tackle AI bias

The **dollar street dataset**: Images representing the geographic and socioeconomic diversity of the world

WAG Rojas, S Diamos, KR Kini, D Kanter… - … -sixth Conference on …, 2022 - openreview.net

… **Dollar Street**—a supervised **dataset** that contains 38,479 images of everyday household items from homes around the world. This **dataset** … This **dataset** includes images from homes with …

☆ Save Cite Cited by 54 Related articles All 2 versions »

50+ citations across research teams in academia, industry and startups

Facet: Fairness in computer vision evaluation benchmark

L Gustafson, C Rolland, N Ravi… - Proceedings of the …, 2023 - openaccess.thecvf.com

Computer vision models have known performance disparities across attributes such as gender and skin tone. This means during tasks such as classification and detection, model …

☆ Save Cite Cited by 17 Related articles All 6 versions »

A whac-a-mole dilemma: Shortcuts come in multiples where mitigating one amplifies others

Z Li, I Evtimov, A Gordo, C Hazirbas… - Proceedings of the …, 2023 - openaccess.thecvf.com

Abstract Machine learning models have been found to learn shortcuts---unintended decision rules that are unable to generalize---undermining models' reliability. Previous works address …

☆ Save Cite Cited by 40 Related articles All 9 versions »

Survey of social bias in vision-language models

N Lee, Y Bang, H Lovenia, S Cahyawijaya… - arXiv preprint arXiv …, 2023 - arxiv.org

In recent years, the rapid advancement of machine learning (ML) models, particularly transformer-based pre-trained models, has revolutionized Natural Language Processing …

☆ Save Cite Cited by 5 Related articles All 2 versions »

The casual conversations v2 dataset

B Porgali, V Albiero, J Ryda… - Proceedings of the …, 2023 - openaccess.thecvf.com

This paper introduces a new large consent-driven dataset aimed at assisting in the evaluation of algorithmic bias and robustness of computer vision and audio speech models …

☆ Save Cite Cited by 19 Related articles All 6 versions »

Representation in AI evaluations

AS Bergman, LA Hendricks, M Rauh, B Wu… - Proceedings of the …, 2023 - dl.acm.org

Calls for representation in artificial intelligence (AI) and machine learning (ML) are widespread, with" representation" or" representativeness" generally understood to be both …

☆ Save Cite Cited by 15 Related articles

Overwriting pretrained bias with finetuning data

A Wang, O Russakovsky - Proceedings of the IEEE/CVF …, 2023 - openaccess.thecvf.com

Transfer learning is beneficial by allowing the expressive features of models pretrained on large-scale datasets to be finetuned for the target task of smaller, more domain-specific …

☆ Save Cite Cited by 17 Related articles All 7 versions »

COACTIVE

How you can support this work

- **Download** the Dollar Street dataset from MLcommons at: mlcommons.org/en/dollar-street
- **Contact** us at dollarstreet@mlcommons.org
- **Learn** more about the Dollar Street project by visiting: gapminder.org/dollar-street
- **Spread the word!**

Download the Dollar Street dataset

Visit the Dollar Street project

COACTIVE

We’re hiring!

Build multimodal AI systems that help humans and machines work together.

COACTIVE

https://www.coactive.ai/careers

How you can support this work

- **Download** the Dollar Street dataset from MLcommons at: mlcommons.org/en/dollar-street
- **Contact** us at dollarstreet@mlcommons.org
- **Learn** more about the Dollar Street project by visiting: gapminder.org/dollar-street
- **Spread the word!**

Download the Dollar Street dataset

Visit the Dollar Street project

COACTIVE