Datasets for recognition

Why talk about datasets

- Datasets serve two functions in recognition:
  - Training
  - Evaluation
- ML will work best close to the training dataset
- Our understanding of performance comes from the evaluation datasets

What we want

- Evaluation datasets:
    - Should match real world testing distribution
    - Must be carefully and completely annotated
    - Can be small
- Training datasets:
    - Should match real world testing distribution
    - Must be large
    - Can be unannotated or partially annotated

Building a dataset

- Step 1: Collect images
- Step 2: Annotate them
- Profit!

Building a dataset

- Suppose we want general category recognition

Building a dataset – Curated approach

- Choose a set of categories
- Collect images for each

Choosing categories

- Space of categories immense
- How broad?
  - Animals? Vehicles? Household objects?...
- How granular?
  - Bird vs sparrow
  - Dog vs golden retriever
  - People often use inconsistently granular labels
- E.g., ImageNet uses wordnet
- Taxonomies typically arbitrary and use-dependent

Choosing images

- Pick a set of categories
- Search for images of those categories
- Doesn't work because of
  - Posed images ("Iconic images")
  - Use of recognition models inside search
  - Will not give in-the-wild content

Choosing images

- Pick a set of categories
- Search for images of those categories
- Doesn’t work because of
  - Posed images (“Iconic images”)
  - Use of recognition models inside search
  - Will not give in-the-wild content



The curated approach

- Pick a set of categories
- Pick intelligent search terms that will yield in-the-wild content
  - PASCAL VOC: search for “party”, “birthday” etc.
  - COCO: search for pairs of classes: “dog”+”bicycle” etc.

The curated approach - labeling

- Usually an image will have multiple classes
- Will have to ask annotators to label every class in the image exhaustively
- A very hard annotation problem

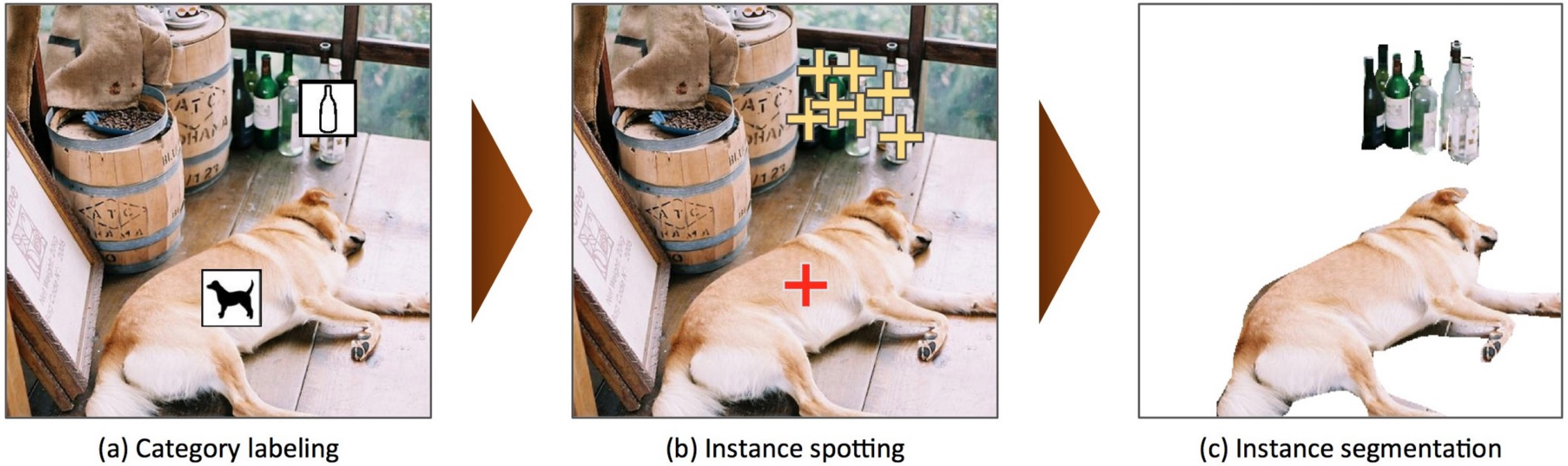


Fig. 3: Our annotation pipeline is split into 3 primary tasks: (a) labeling the categories present in the image (§4.1), (b) locating and marking all instances of the labeled categories (§4.2), and (c) segmenting each object instance (§4.3).

The curated approach

- Many categories will be relatively rare
  - Performance on these categories will suffer
- Need special targetted collection efforts to avoid this

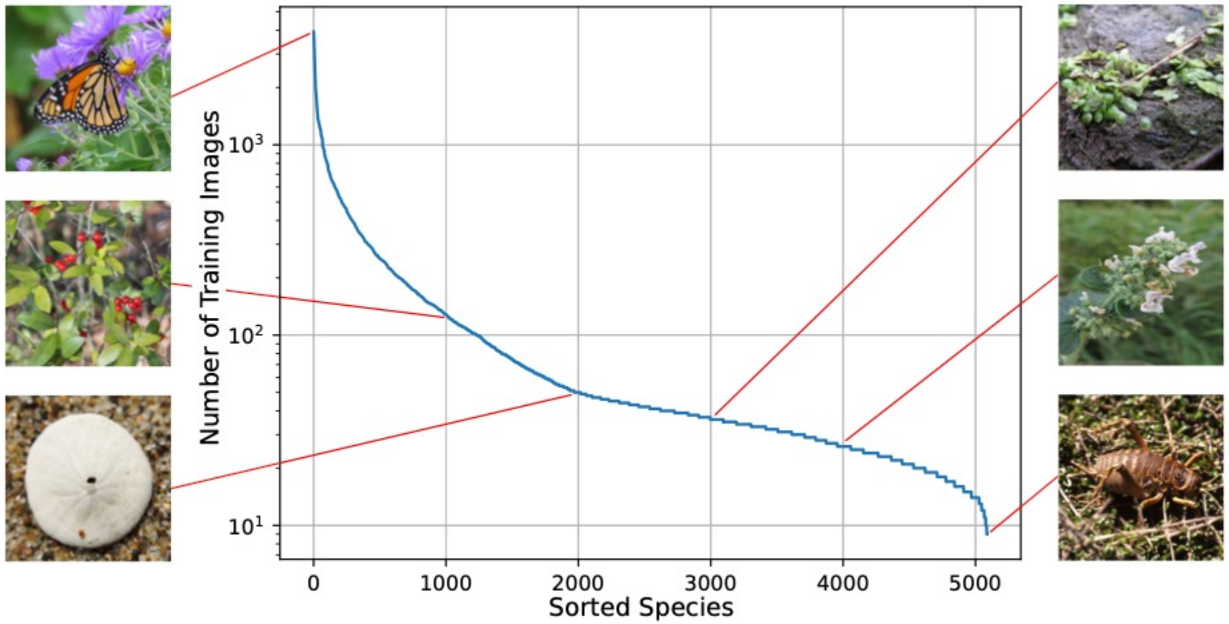
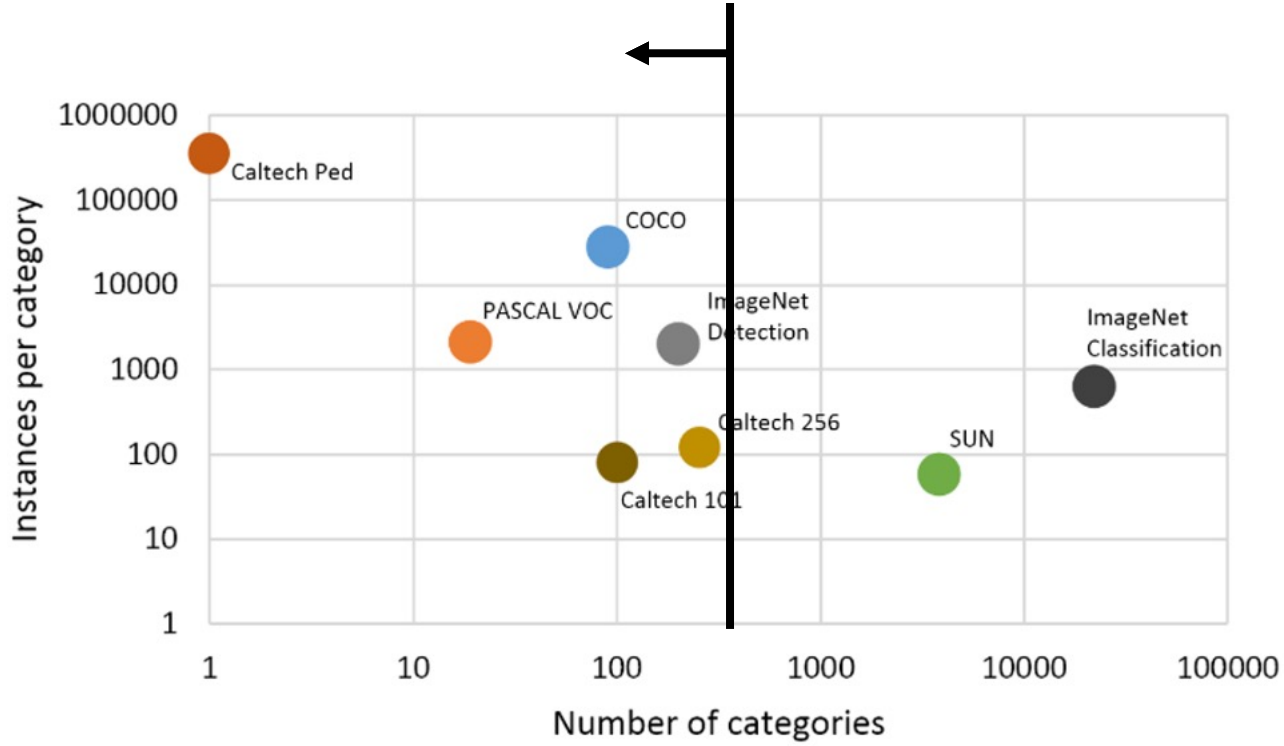


Figure 2. Distribution of training images per species. iNat2017 contains a large imbalance between classes, where the top 1% most populated classes contain over 16% of training images.

Problems with curated datasets

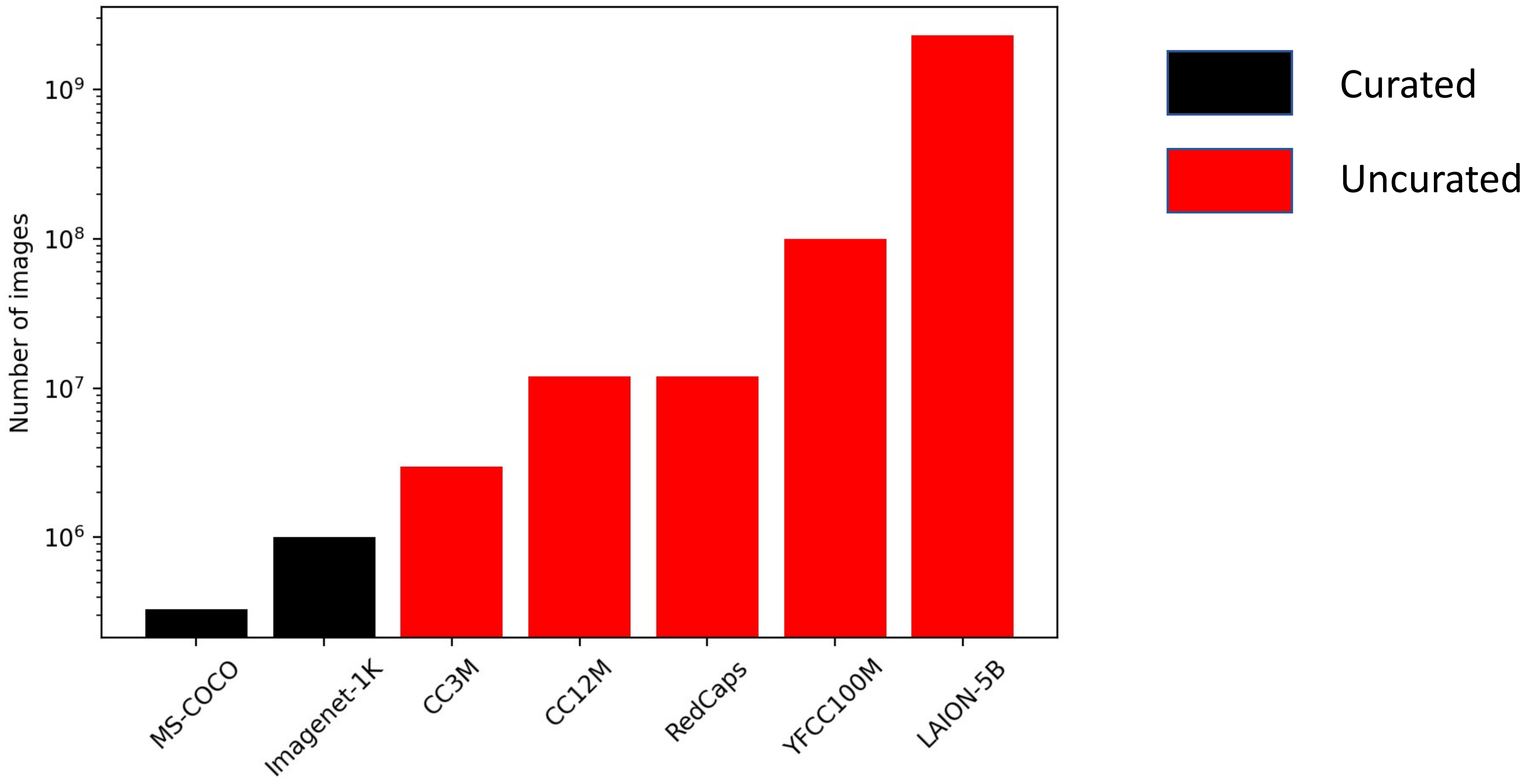
Curated datasets today have <= 1m images and <500 classes



The uncurated approach – data collection

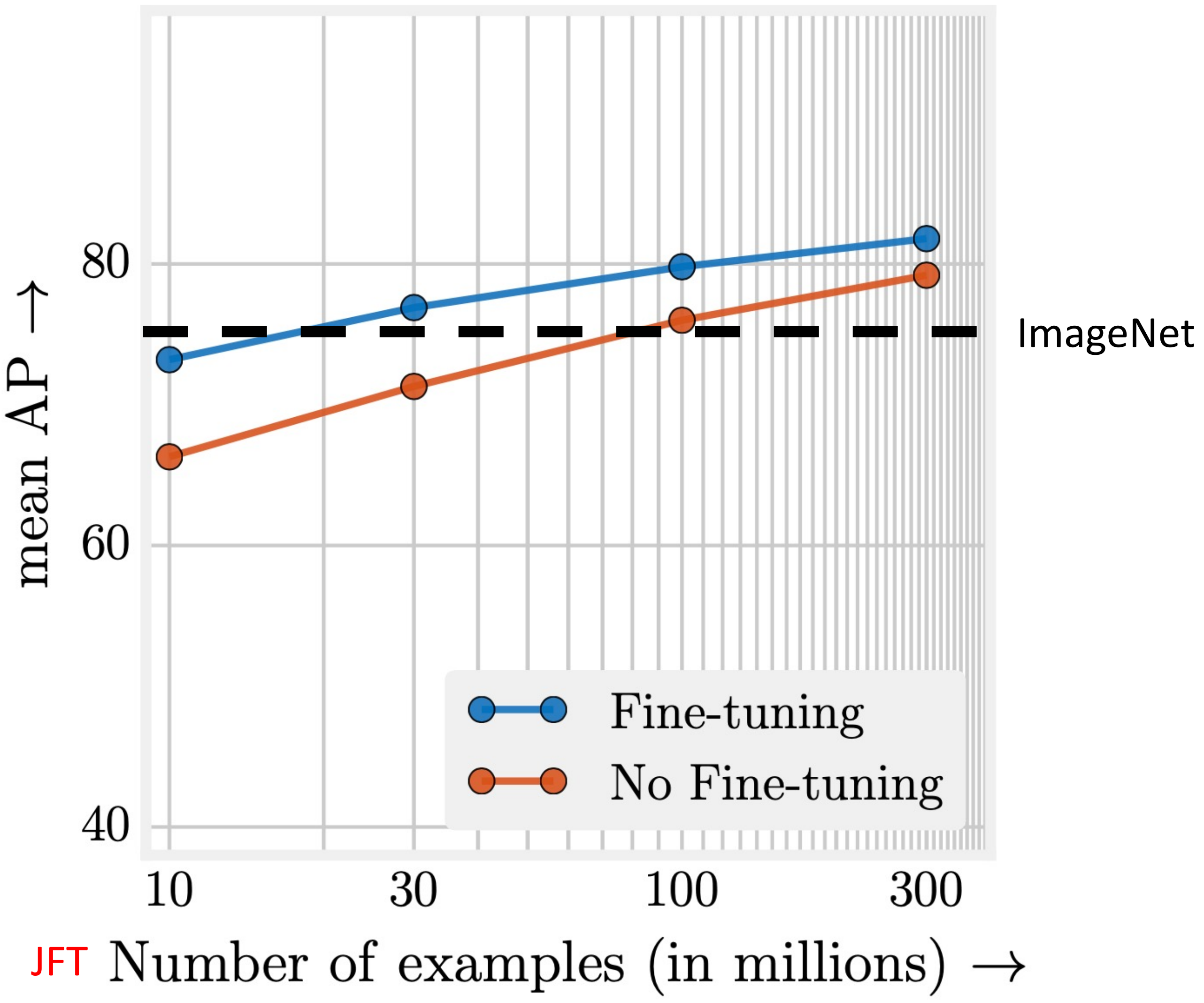
- Crawl the web for images
- Label / annotate them
  - Annotations are typically “open-world” – not fixed categories
  - E.g., captions, alt-text
  - Sometimes come with “grounding” = bounding boxes

Uncurated data



Sun, Chen, et al. "Revisiting unreasonable effectiveness of data in deep learning era." *Proceedings of the IEEE international conference on computer vision*. 2017.

Impact of size of data

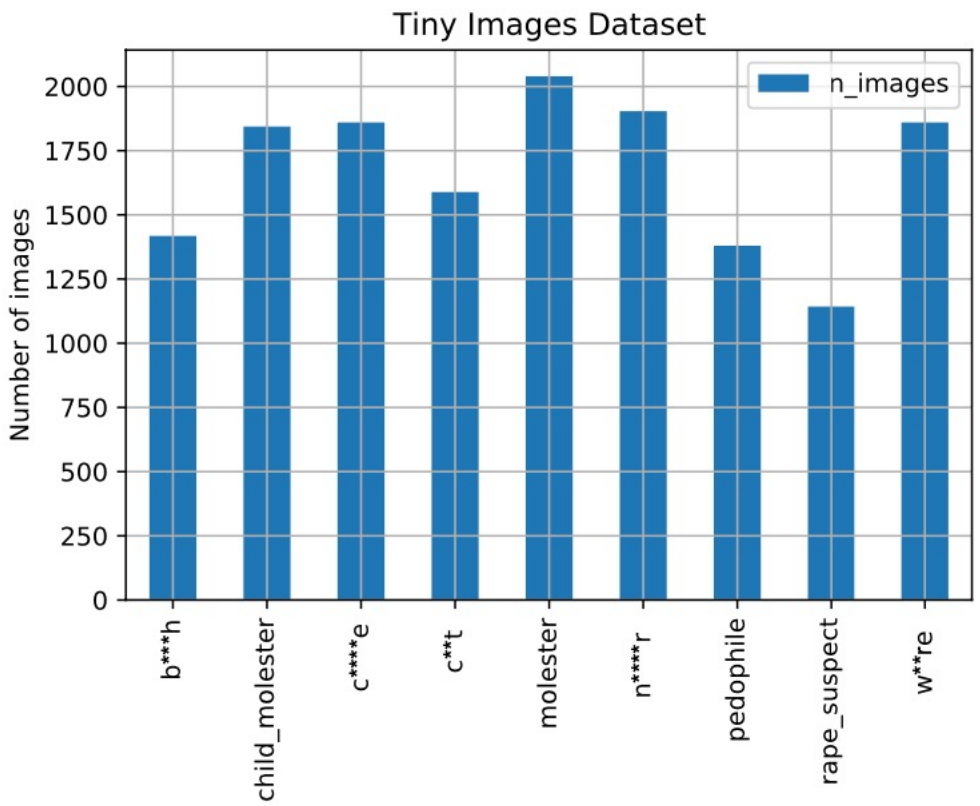


Challenges with uncurated data

- Images can be
  - NSFW

Table 1: Large scale image datasets containing people's images

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,1k) ([90])	(14, 12, 1)	(22, 11, 1)	0



Birhane, Abeba, and Vinay Uday Prabhu. "Large image datasets: A pyrrhic win for computer vision?." *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2021.

Challenges with uncurated data

- Images can be
  - NSFW
- Filtering?
  - At scale, can only be done by automatic techniques…
  - => Biased, incorrect filtering
  - “Curation debt”

Butters OW, Wilson RC, Burton PR. Recognizing, reporting and reducing the data curation debt of cohort studies. Int J Epidemiol. 2020 Aug 1;49(4):1067-1074. doi: 10.1093/ije/dyaa087. PMID: 32617581; PMCID: PMC7660145.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜 In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21). Association for Computing Machinery, New York, NY, USA, 610–623. https://doi.org/10.1145/3442188.3445922

Challenges with uncurated data

- Datasets are biased by what is there on the internet

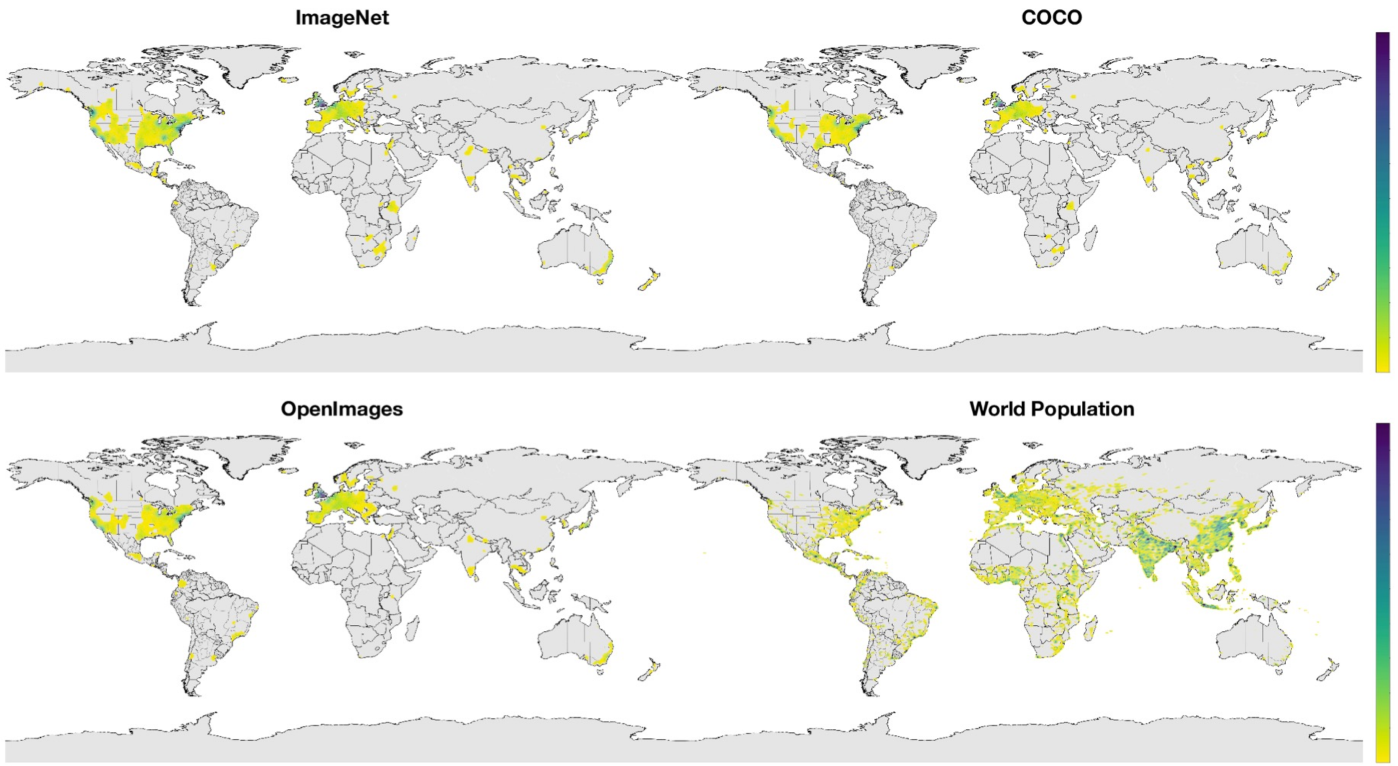


Figure 6: Density maps showing the geographical distribution of images in the ImageNet (top-left), COCO (top-right), and OpenImages (bottom-left) datasets. A world population density map is shown for reference (bottom-right).

De Vries, Terrance, et al. "Does object recognition work for everyone?." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019.

Challenges with uncurated data

Ground truth: Soap **Nepal, 288 $/month**

Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

Ground truth: Soap **Nepal, 288 $/month**

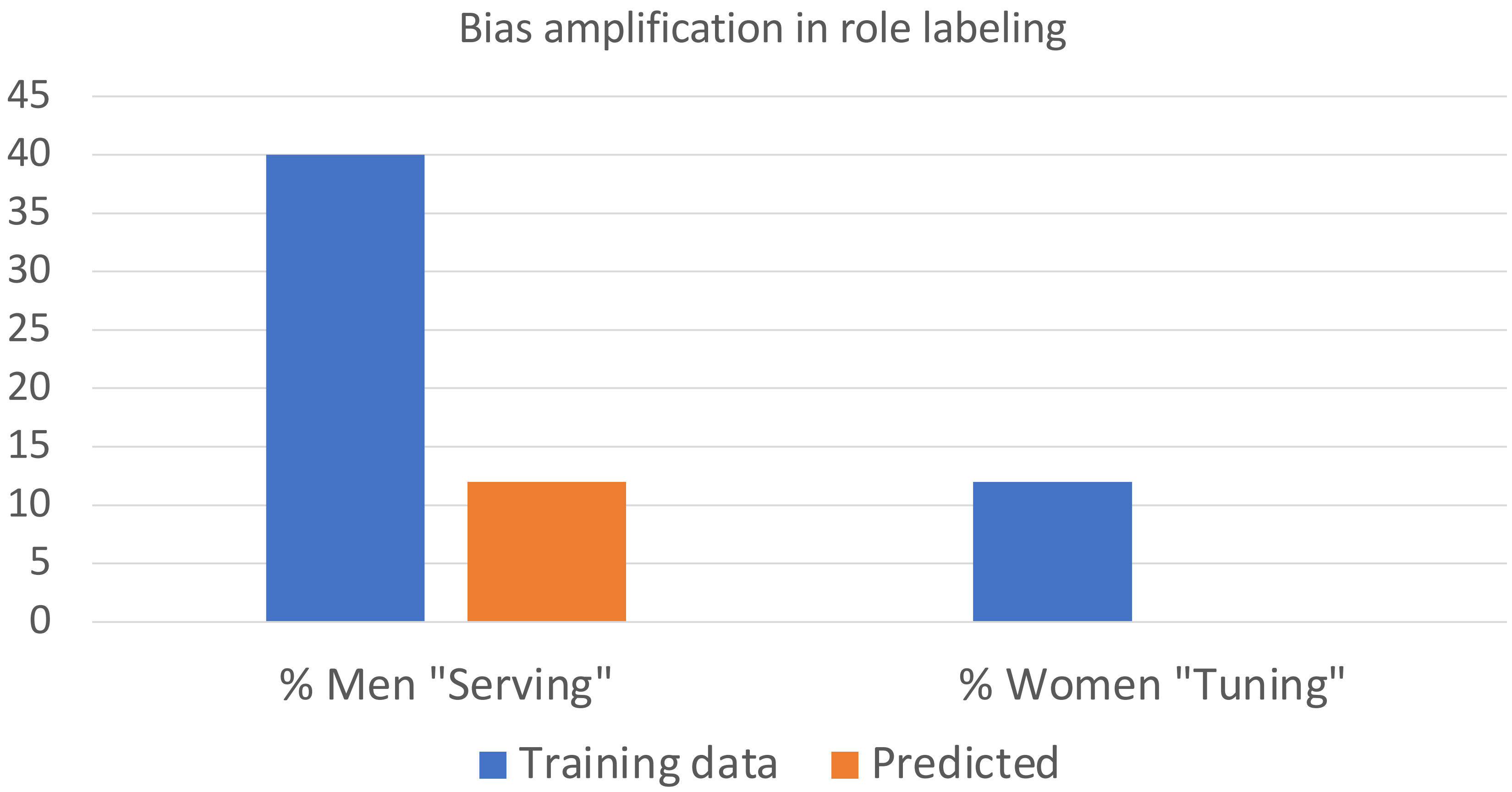
Azure: food, cheese, bread, cake, sandwich
Clarifai: food, wood, cooking, delicious, healthy
Google: food, dish, cuisine, comfort food, spam
Amazon: food, confectionary, sweets, burger
Watson: food, food product, turmeric, seasoning
Tencent: food, dish, matter, fast food, nutriment

Ground truth: Soap **UK, 1890 $/month**

Azure: toilet, design, art, sink
Clarifai: people, faucet, healthcare, lavatory, wash closet
Google: product, liquid, water, fluid, bathroom accessory
Amazon: sink, indoors, bottle, sink faucet
Watson: gas tank, storage tank, toiletry, dispenser, soap dispenser
Tencent: lotion, toiletry, soap dispenser, dispenser, after shave

De Vries, Terrance, et al. "Does object recognition work for everyone?." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2019.

Side note: ML algorithms *amplify* bias



Zhao, Jieyu, et al. "Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.

Challenges with uncurated data - Generalization

- What has the model seen and not seen?
  - Generally, model accuracy higher on data it has seen
  - *Even if it hasn't seen labels*

- Challenging to measure generalization for large datasets

- Impossible if dataset is closed

“Uncurated”

- Internet data is not uncurated
- Photographer’s bias
- “Interesting-ness”
- For some domains, true uncurated data exists
  - Self-driving cars
  - Satellite imagery