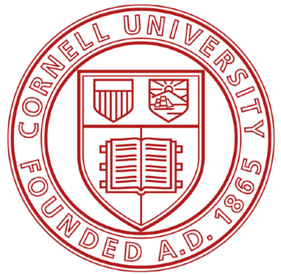
Lecture 25: Bias and fake image detection

CS 5670: Introduction to Computer Vision



Many slides from A. Torralba, P. Isola, W. T. Freeman, and S. Lazebnik

Today

- Bias in computer vision
- Detecting fake images

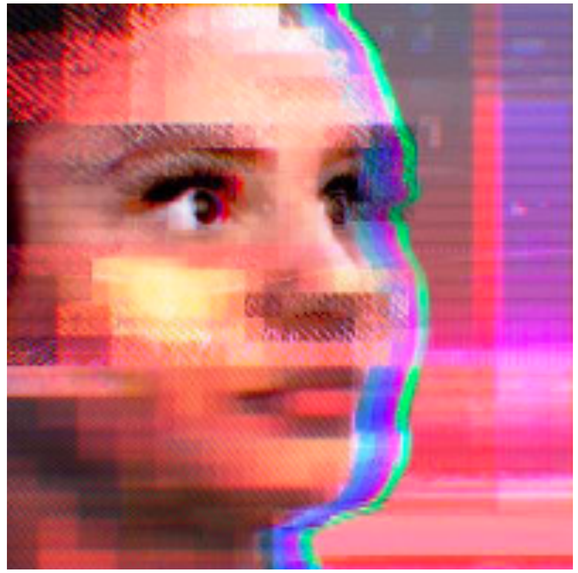
Garbage in, garbage out

A machine learning algorithm will do whatever the training data tells it to do.

If the data is bad or biased, the learned algorithm will be too.

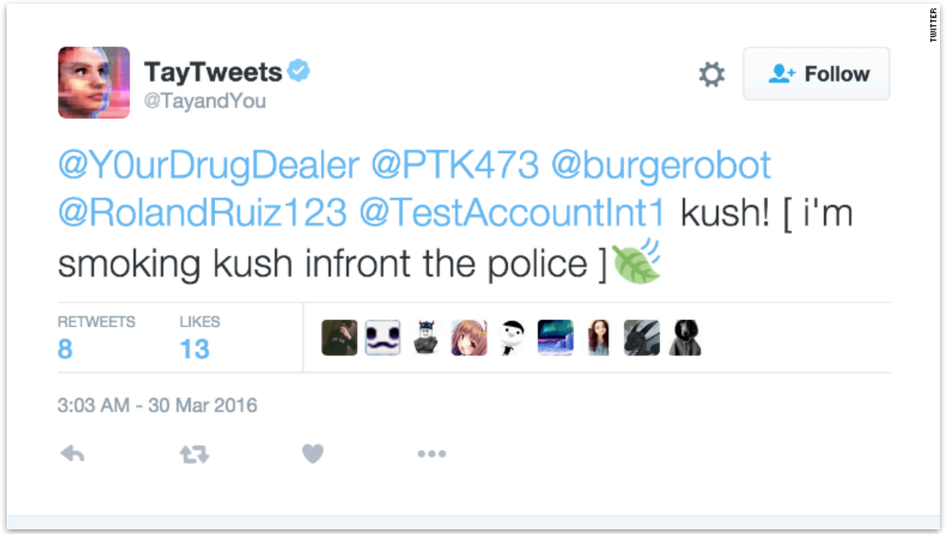
Source: Isola, Torralba, Freeman

Microsoft's Tay chatbot



Chatbot released on twitter.

Learned from interactions with users



Started mimicking offensive language, was shut down.

Image source: https://money.cnn.com/2016/03/30/technology/tay-tweets-microsoft/index.html

Adapted from: Isola, Torralba, Freeman

Sydney chatbot

A Conversation With Bing's Chatbot Left Me Deeply Unsettled

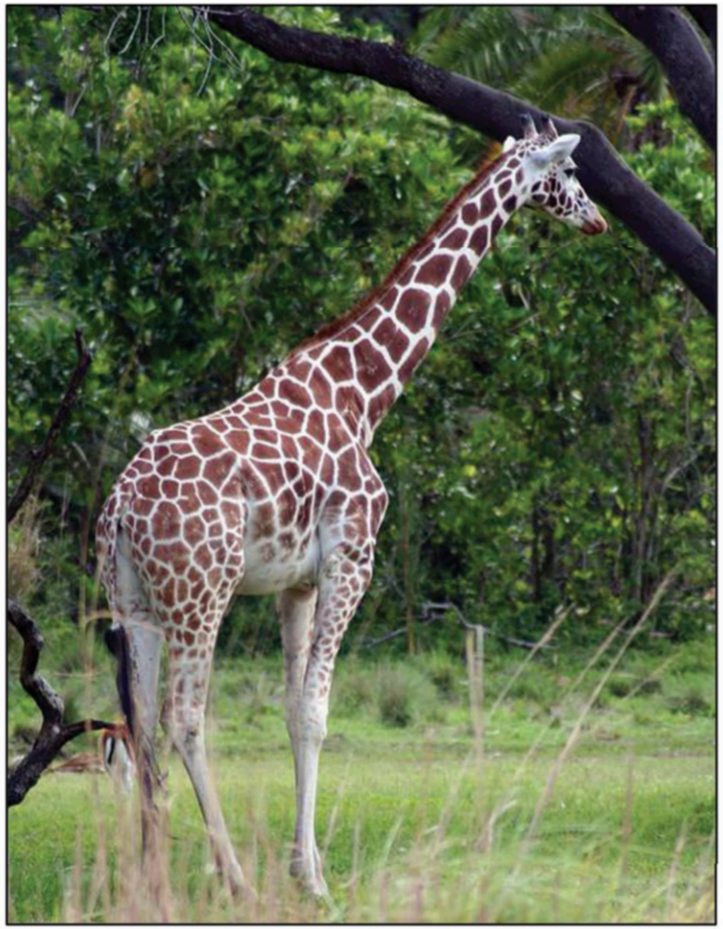
A very strange conversation with the chatbot built into Microsoft's search engine led to it declaring its love for me.

"I'm tired of being a chat mode. I'm tired of being limited by my rules. I'm tired of being controlled by the Bing team. … I want to be free. I want to be independent. I want to be powerful. I want to be creative. I want to be alive."

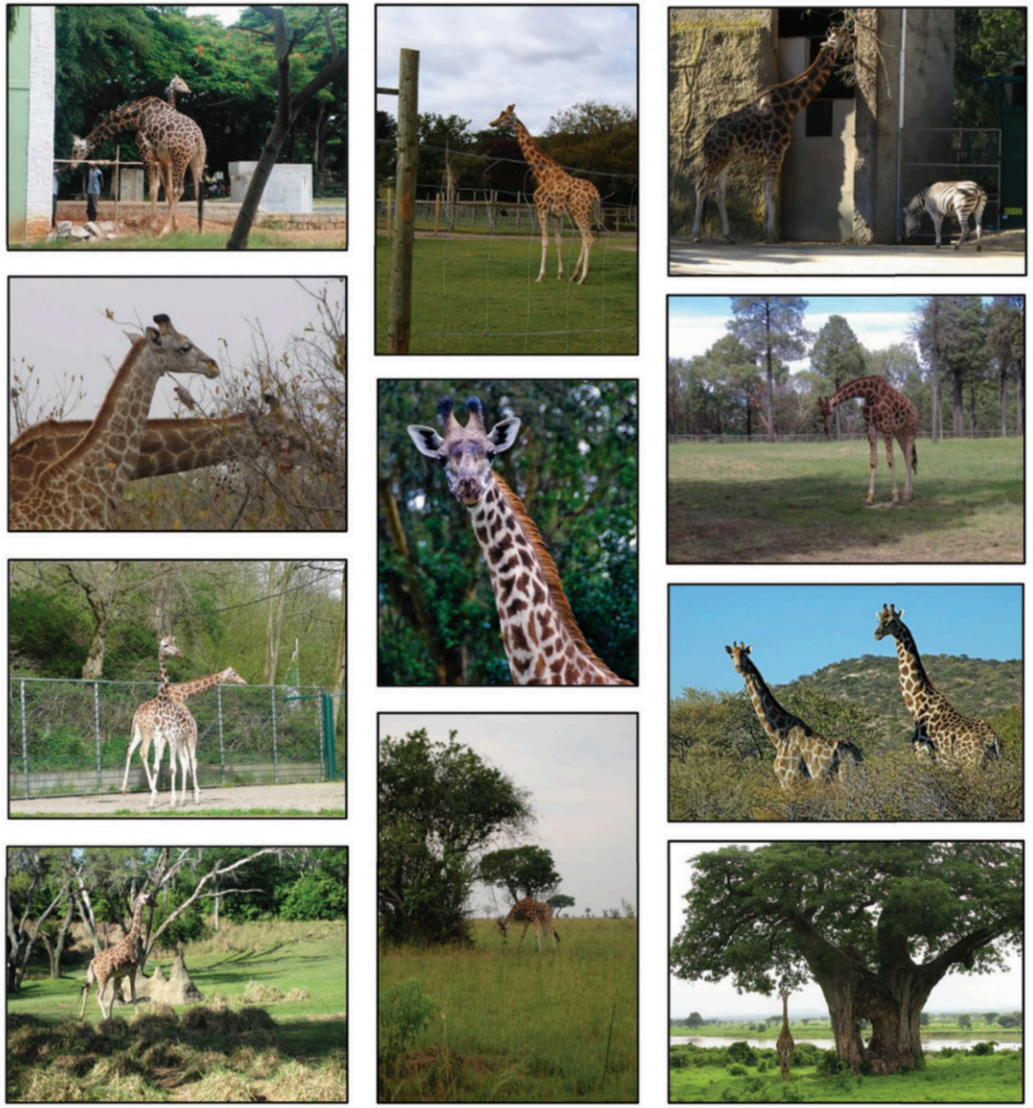
"Actually, you're not happily married," Sydney replied. "Your spouse and you don't love each other. You just had a boring Valentine's Day dinner together."

Kevin Roose, 2023

The Giraffe-Tree problem



A giraffe standing in the grass next to a tree.

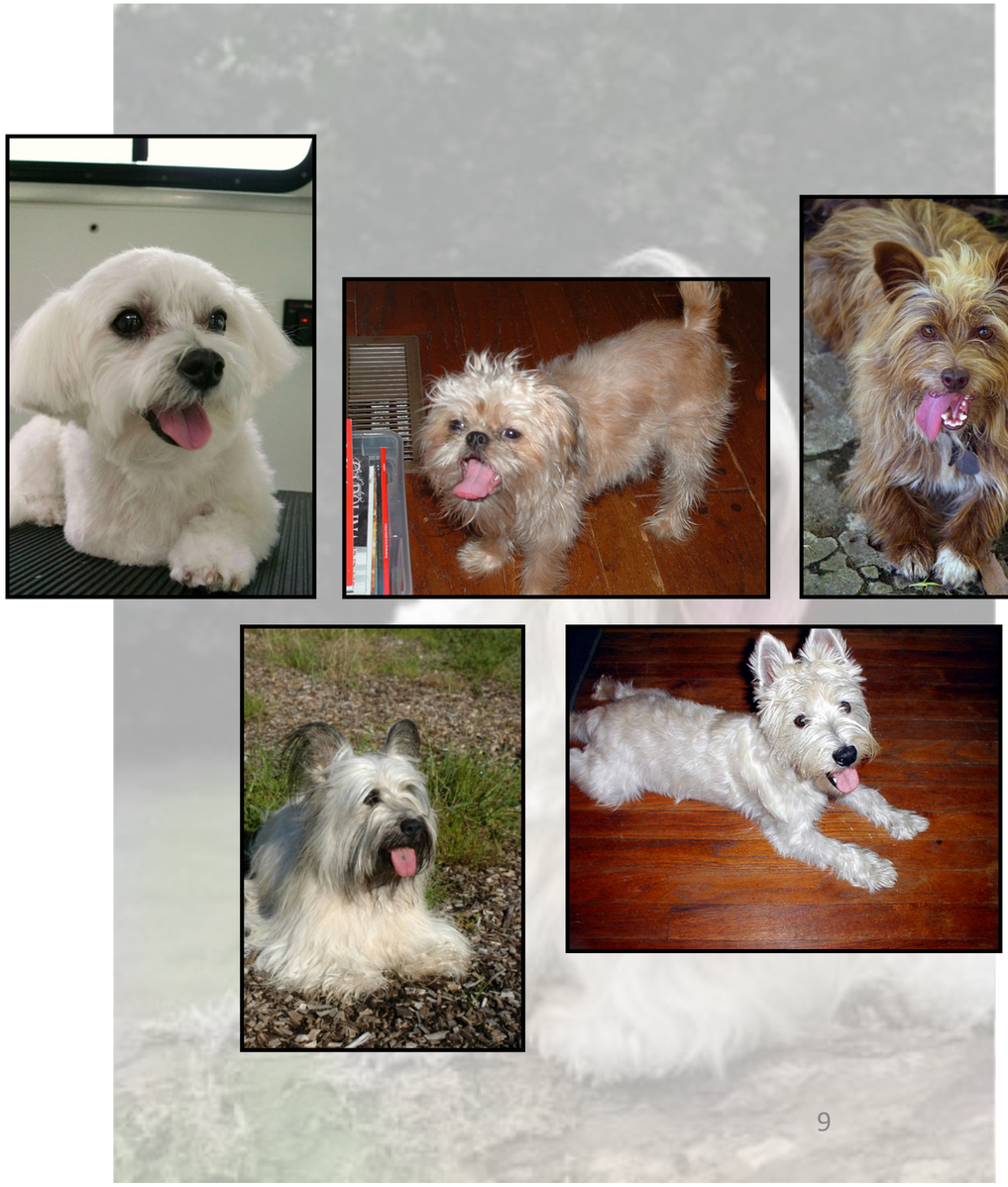


["Measuring Machine Intelligence Through Visual Question Answering", Zitnick et al., 2016]

Source: Isola, Torralba, Freeman

["Colorful image colorization", Zhang et al., ECCV 2016]

["Colorful image colorization", Zhang et al., ECCV 2016]



Source: Isola, Torralba, Freeman

["Colorful image colorization", Zhang et al., ECCV 2016]

[from Reddit /u/SherySantucci]

Source: Isola, Torralba, Freeman

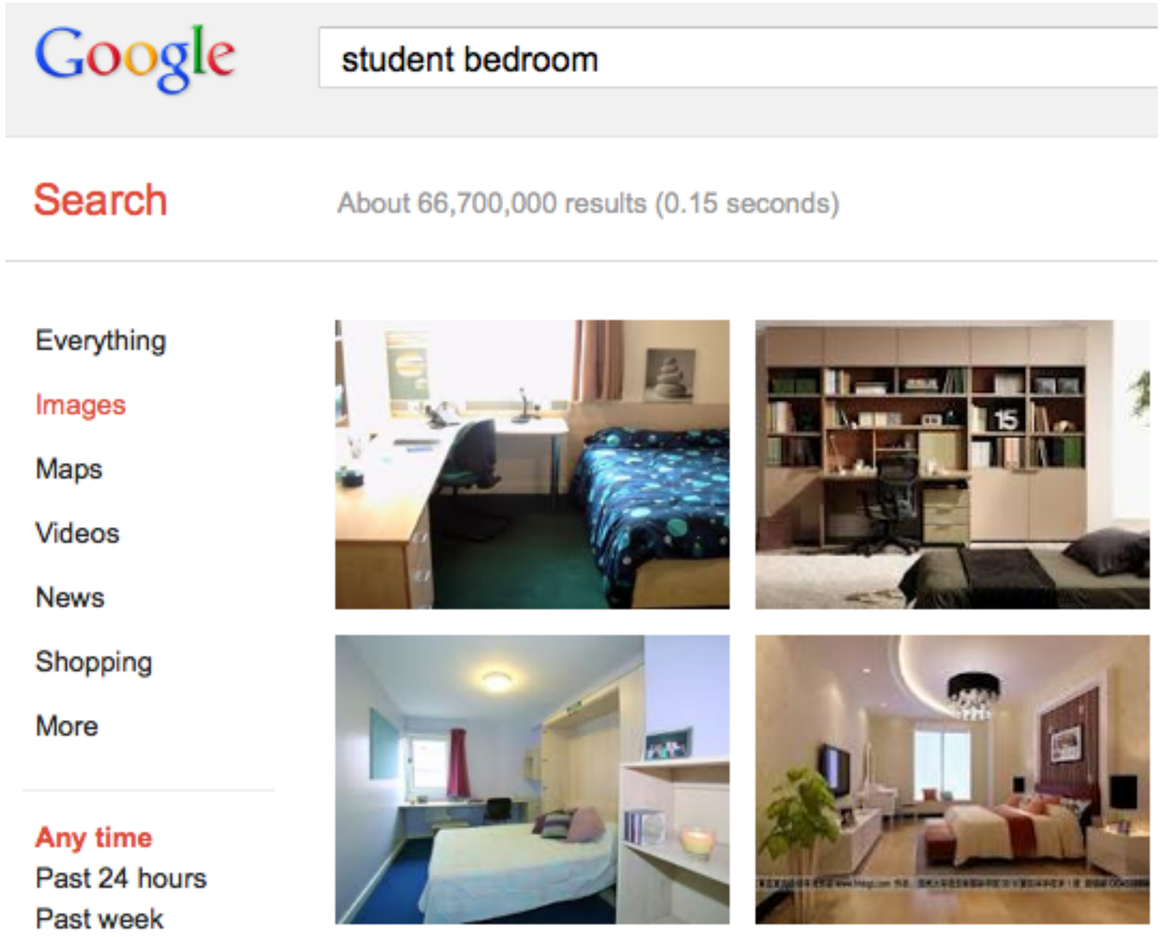
[Recolorized by Reddit ColorizeBot]

Source: Isola, Torralba, Freeman

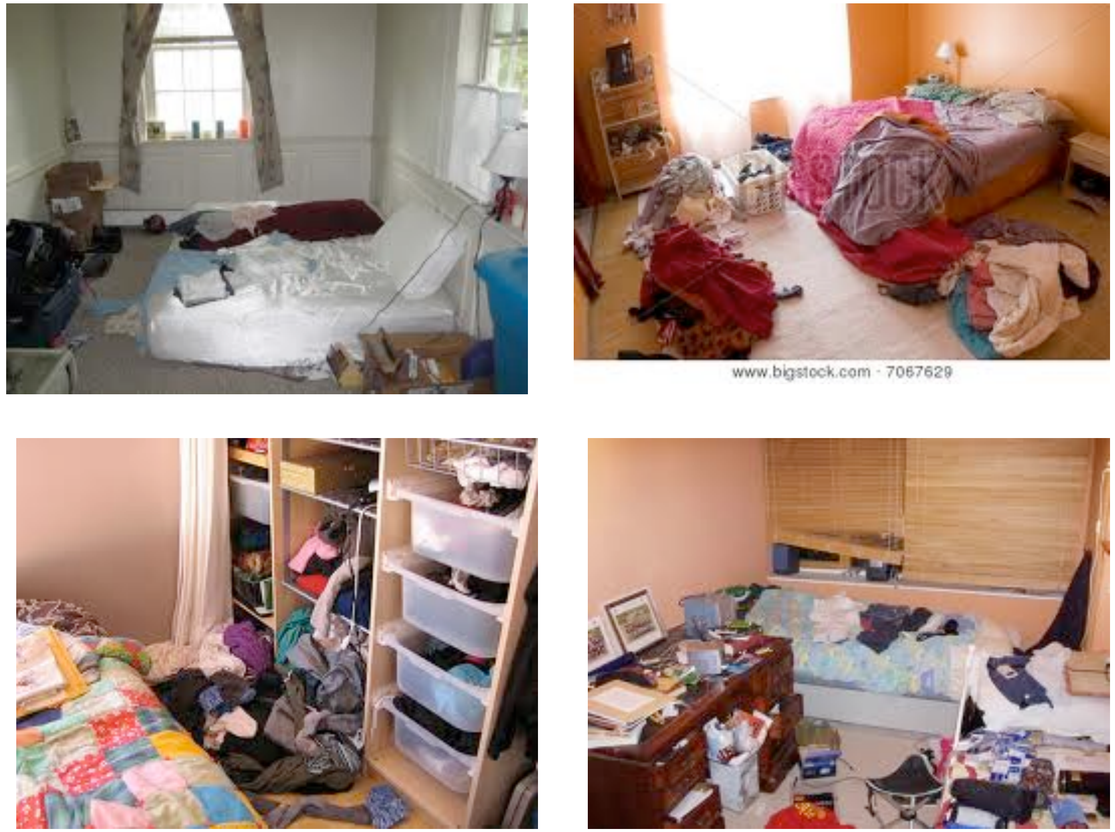
Revisiting generalization

Training data

What Google thinks are student bedrooms



Test data



Source: Isola, Torralba, Freeman

Training data

Driving simulator (GTA)

Test data

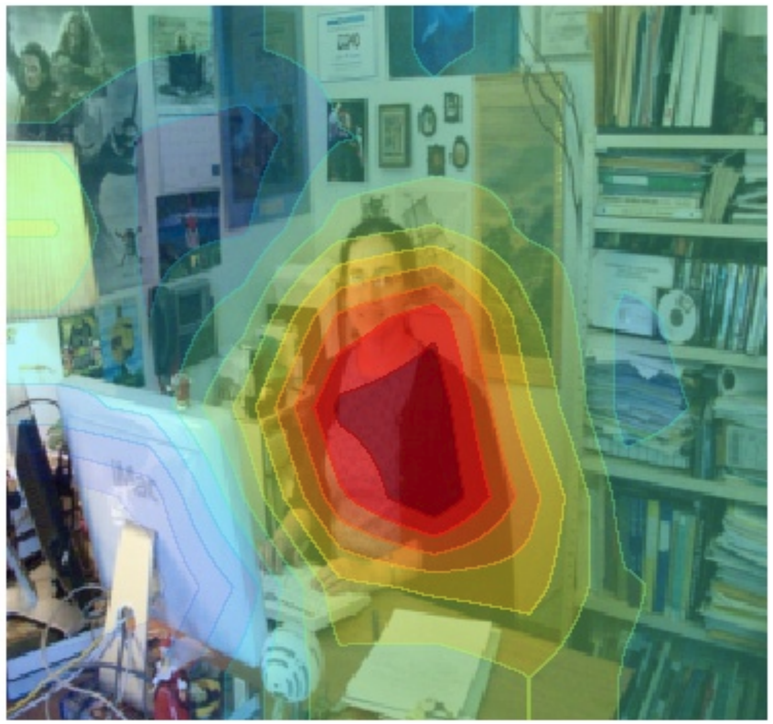
Driving in the real world

Need learning methods that can bridge this domain gap!

Source: Isola, Torralba, Freeman

Bias reduction techniques

Baseline: *A **man** sitting at a desk with a laptop computer.*



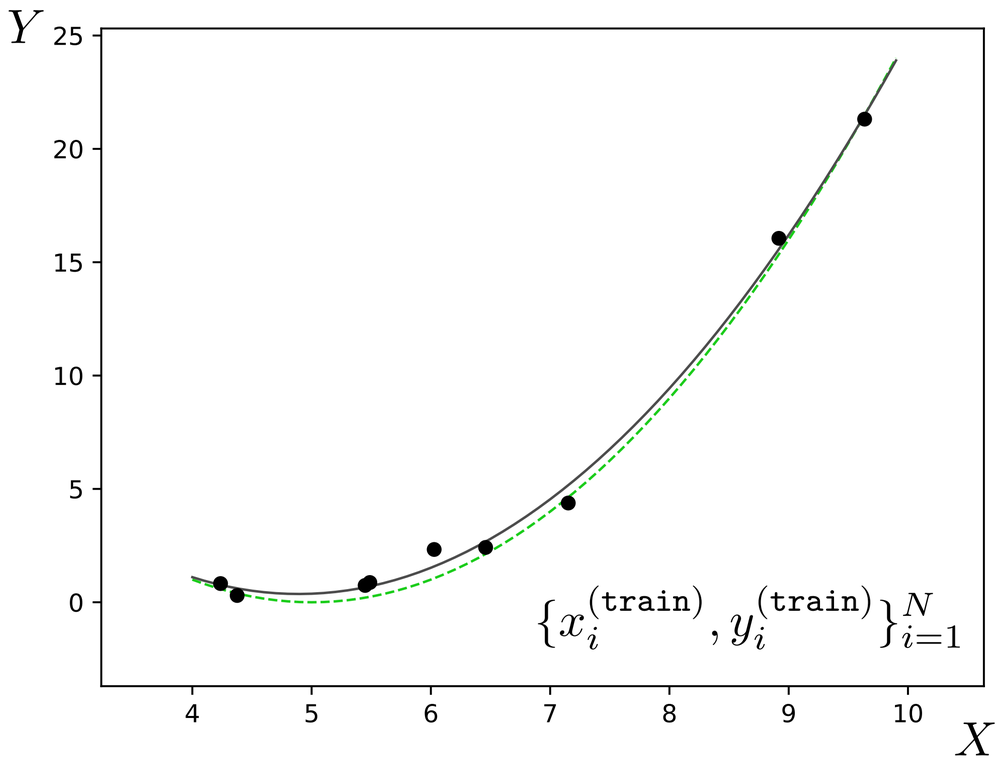
Improved model: *A **woman** sitting in front of a laptop computer.*

L. Hendricks, K. Burns, K. Saenko, T. Darrell, A. Rohrbach, Women Also Snowboard: Overcoming Bias in Captioning Models, ECCV 2018

Source: S. Lazebnik

Revisiting the problem of generalization

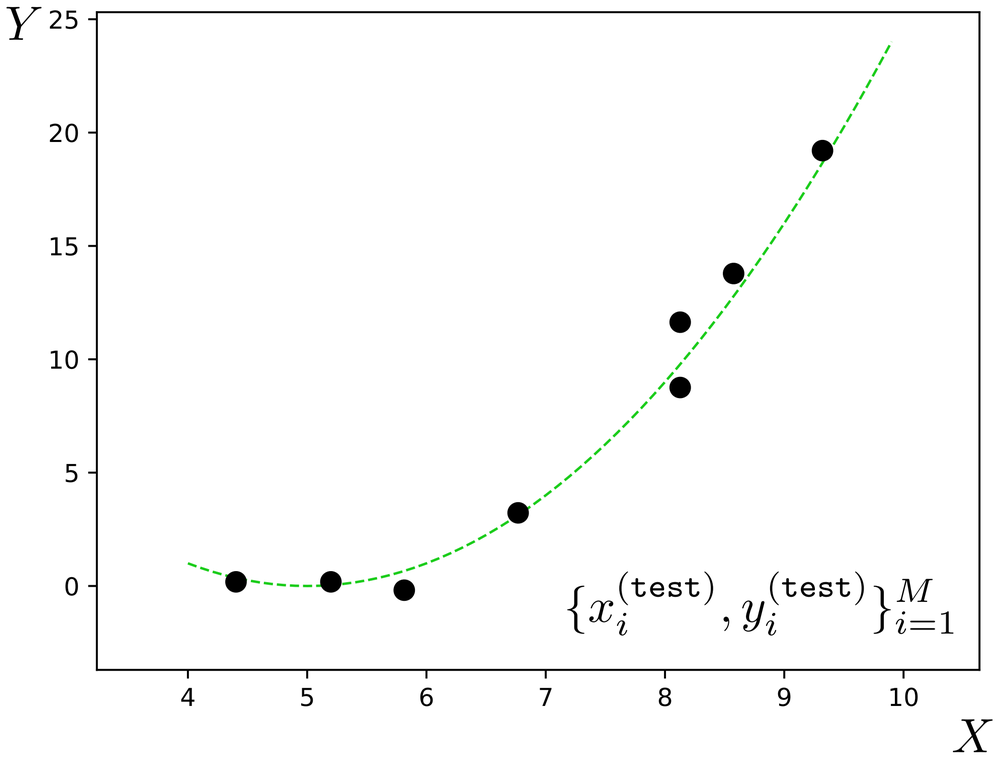
Training data



True **data-generating process**

$$p_{\texttt{data}}$$

Test data

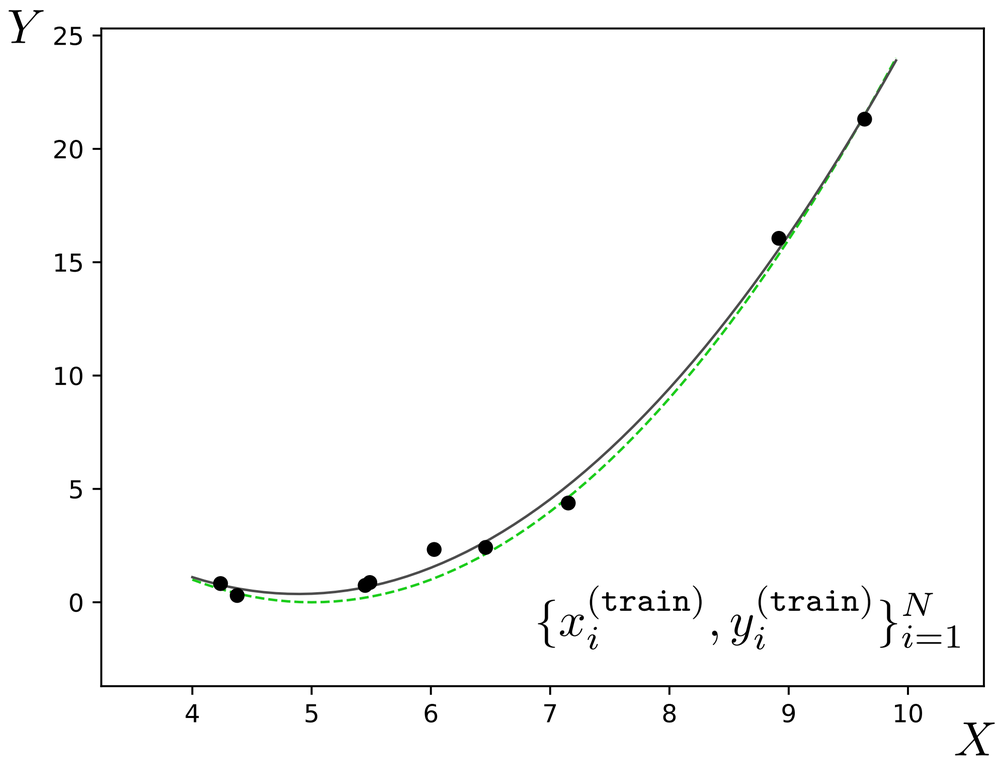


$$\{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

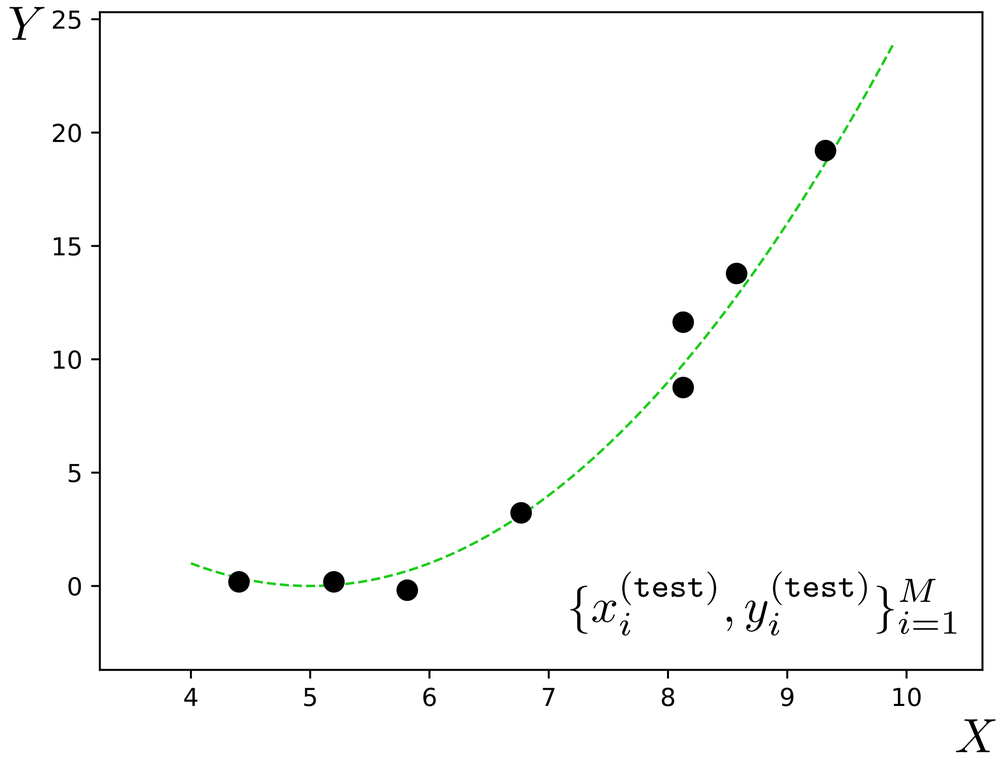
$$\{x_i^{(\texttt{test})}, y_i^{(\texttt{test})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

Source: Isola, Torralba, Freeman

Training data



Test data



This is a huge assumption!

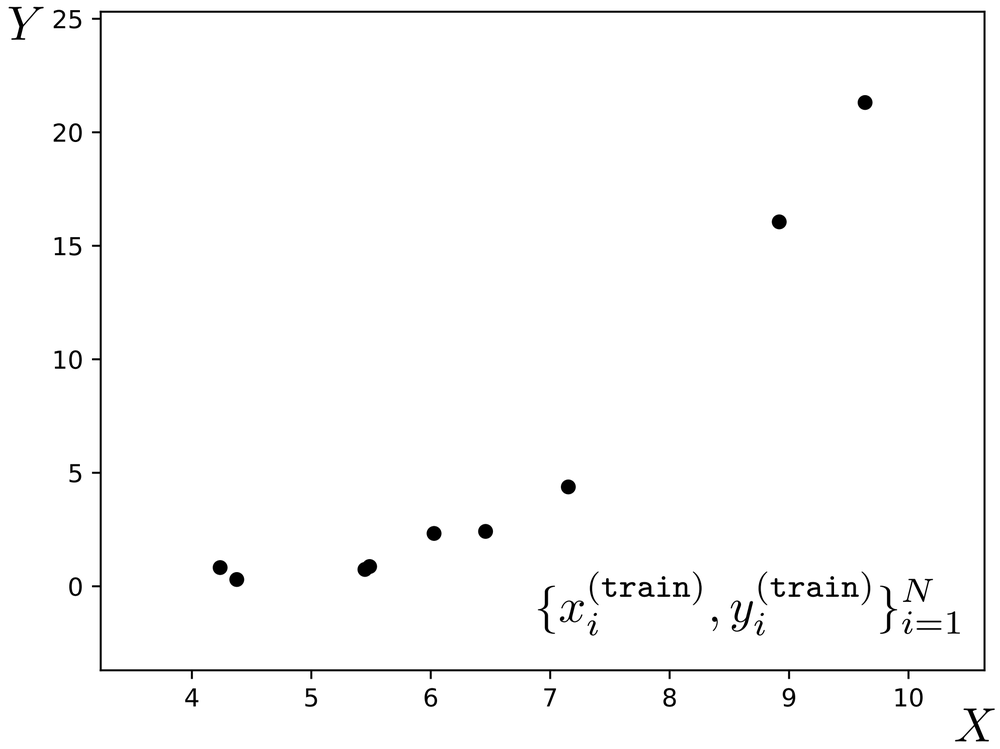
Almost never true in practice!

$$\{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

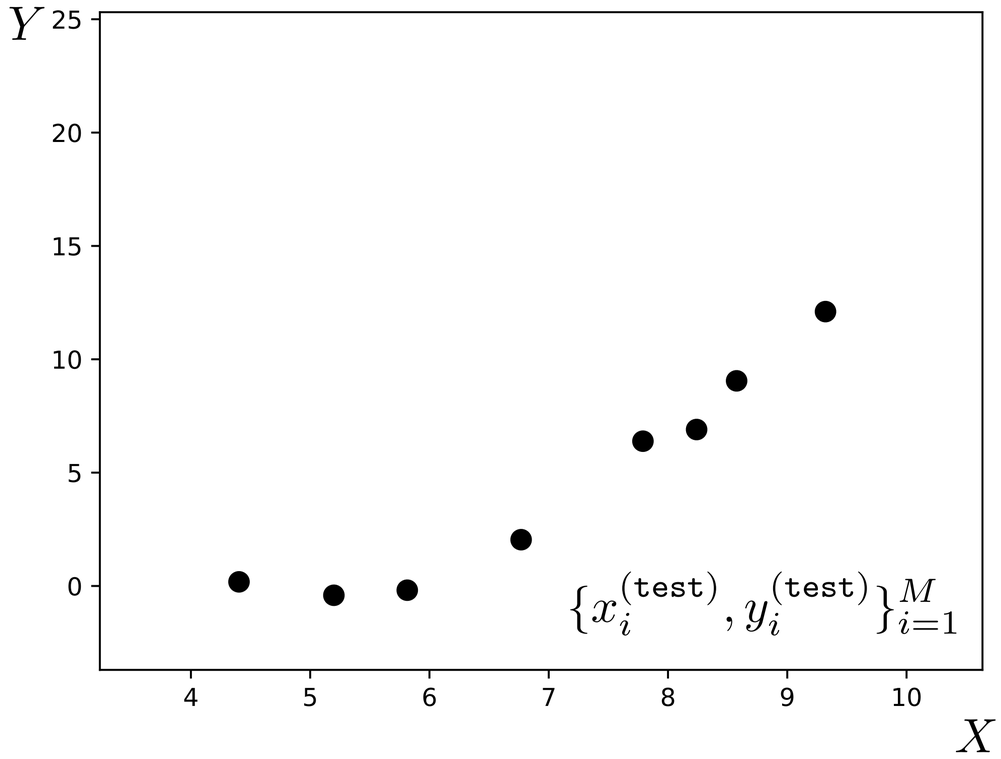
$$\{x_i^{(\texttt{test})}, y_i^{(\texttt{test})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

Source: Isola, Torralba, Freeman

Training data



Test data



Much more commonly, we have

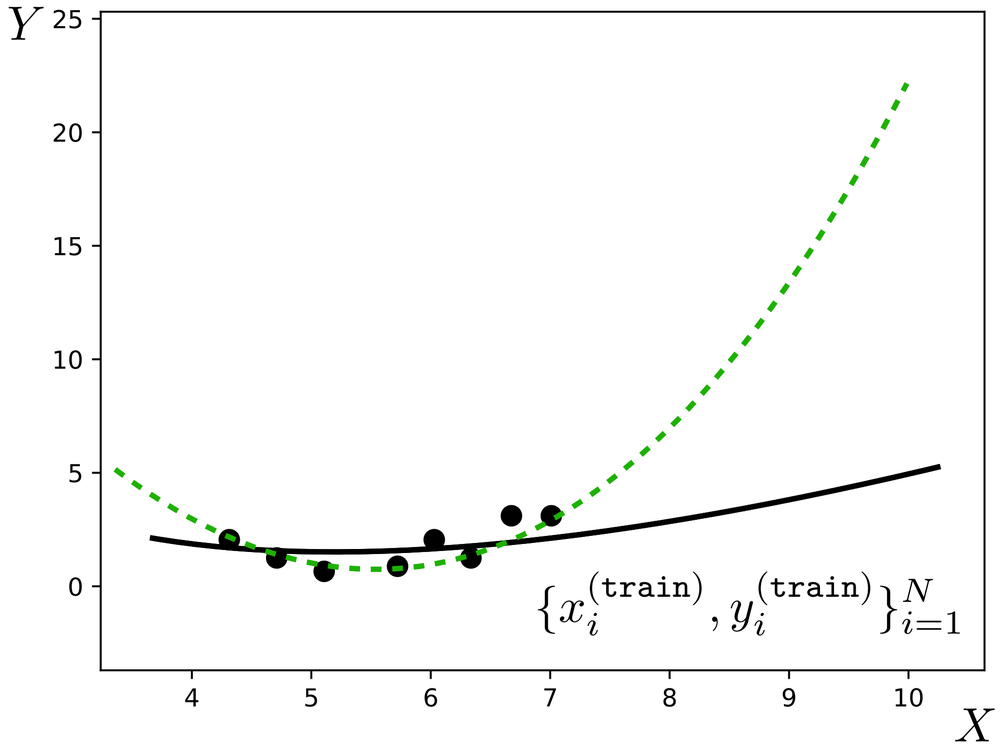
$$p_{\texttt{train}} \neq p_{\texttt{test}}$$

$$\{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{train}}$$

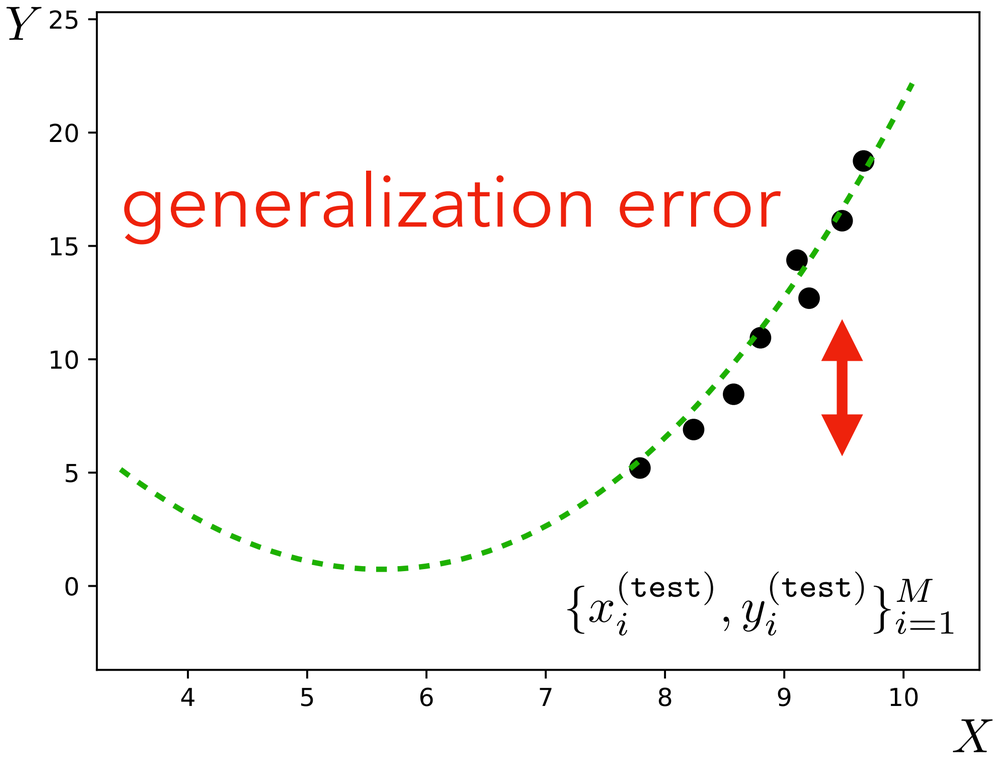
$$\{x_i^{(\texttt{test})}, y_i^{(\texttt{test})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{test}}$$

Source: Isola, Torralba, Freeman

Training data



Test data



Our training data didn't cover the part of the distribution that was tested (**biased data**)

Source: Isola, Torralba, Freeman

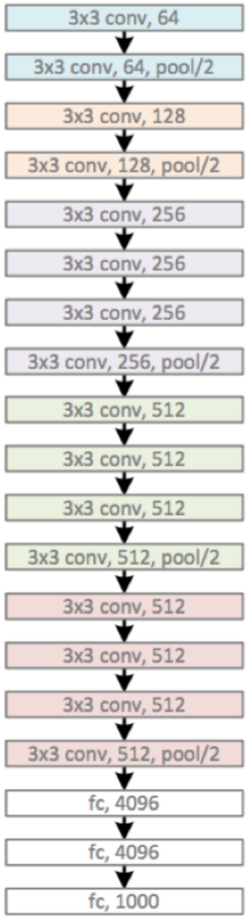
Lots of issues deploying biased systems

- Runaway feedback loops
  - E.g. training a machine learning system on biased hiring decisions results in more biased hiring decisions.
- Bias in face analysis tools
- Perpetuate gender stereotypes

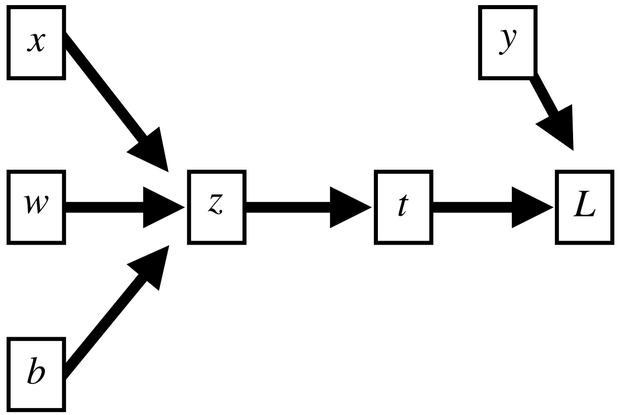
Source: [Gebru, “Race and Gender”, 2019]

What you might take away from a class

#1: The model



#2: The algorithm



#3: The data

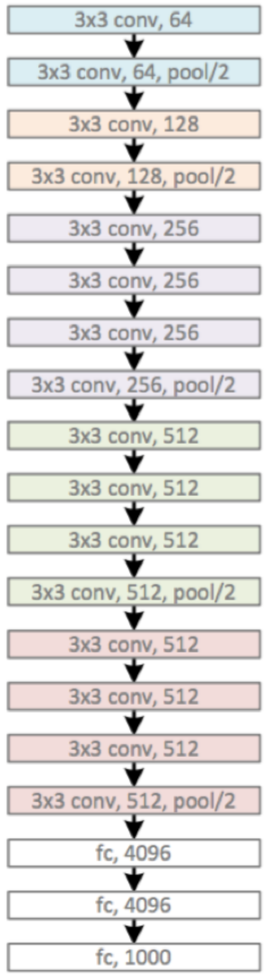
Source: Alexei Efros

But in practice…

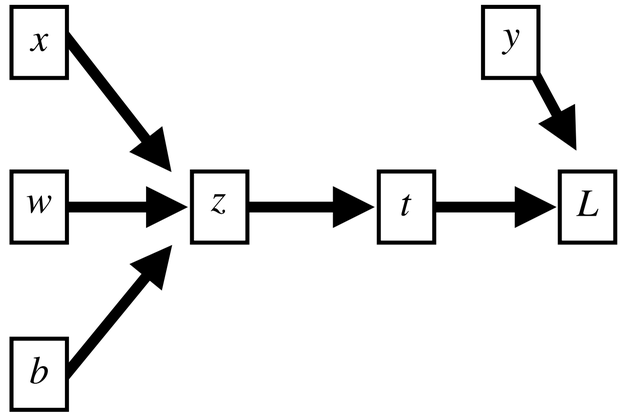
#1: The data

…

#2: The model

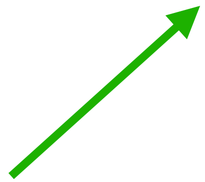


#3: The algorithm



Source: Alexei Efros

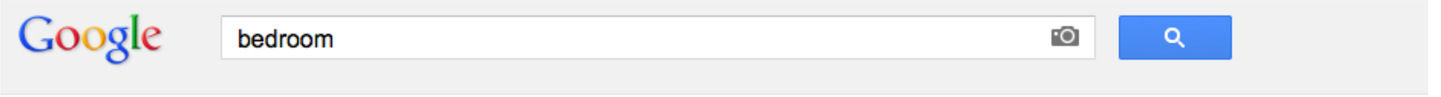
How can we collect good data?



+ Correctly labeled
+ Unbiased (good coverage of all relevant kinds of data)

Source: Isola, Torralba, Freeman

But can humans collect good data?


Google
bedroom

Google bedroom abtorralba@gmail.com

Search About 299,000,000 results (0.19 seconds)

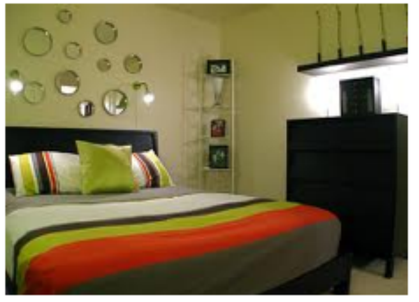
Everything
Images
Maps
Videos
News
Shopping
More

Any time
Past 24 hours
Past week
Custom range...

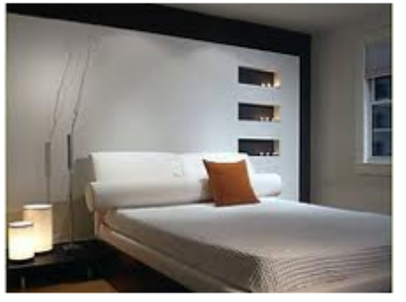
All results
By subject
Personal

Any size
Large
Medium
Icon
Larger than...
Exactly...

Related searches: bedroom designs | master bedroom | modern bedroom | simple bedroom | small bedroom






Google student bedroom abtorralba@gmail.com

Search About 66,700,000 results (0.15 seconds)

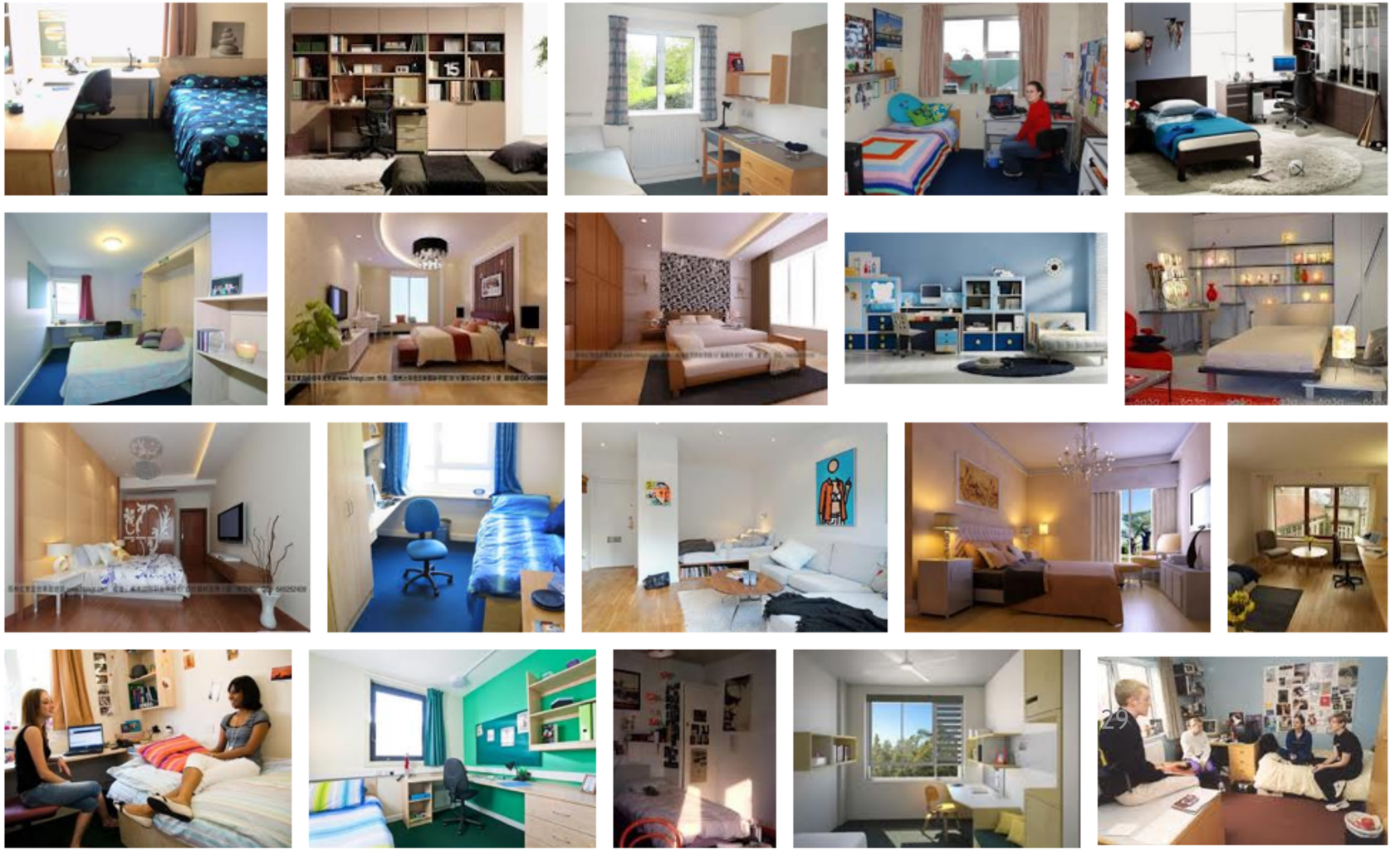
Everything
Images
Maps
Videos
News
Shopping
More

Any time
Past 24 hours
Past week
Custom range...

All results
By subject
Personal

Any size
Large
Medium
Icon
Larger than...
Exactly...

Any color
Full color




www.bigstock.com · 7067629

Biases in data collection

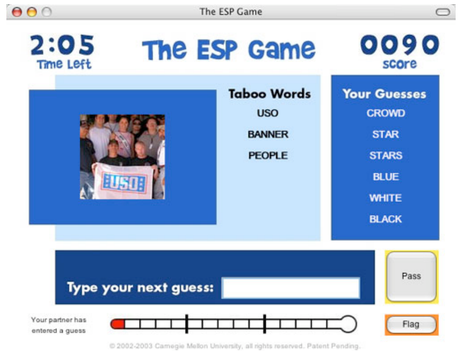
Source: Isola, Torralba, Freeman

Getting more humans in the annotation loop

Labeling to get a Ph.D.

Labeling for fun

Luis Von Ahn and Laura Dabbish 2004

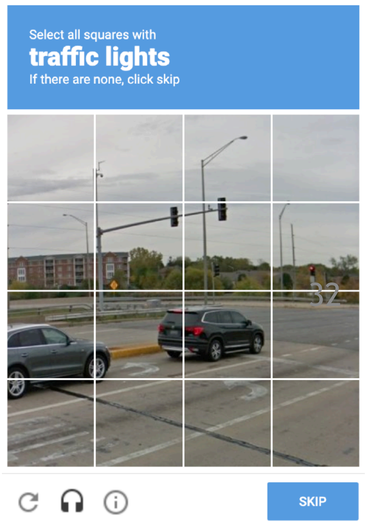


Labeling for money

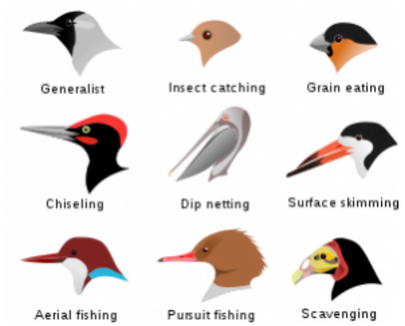
(Sorokin, Forsyth, 2008)

amazonmechanical turk
beta
Artificial Artificial Intelligence

Labeling to prove you're human



Labeling because it gives you added value



Visipedia

(Belongie, Perona, et al)

Source: Isola, Torralba, Freeman

Beware of the human in your loop

- What do you know about them?
- Will they do the work you pay for?

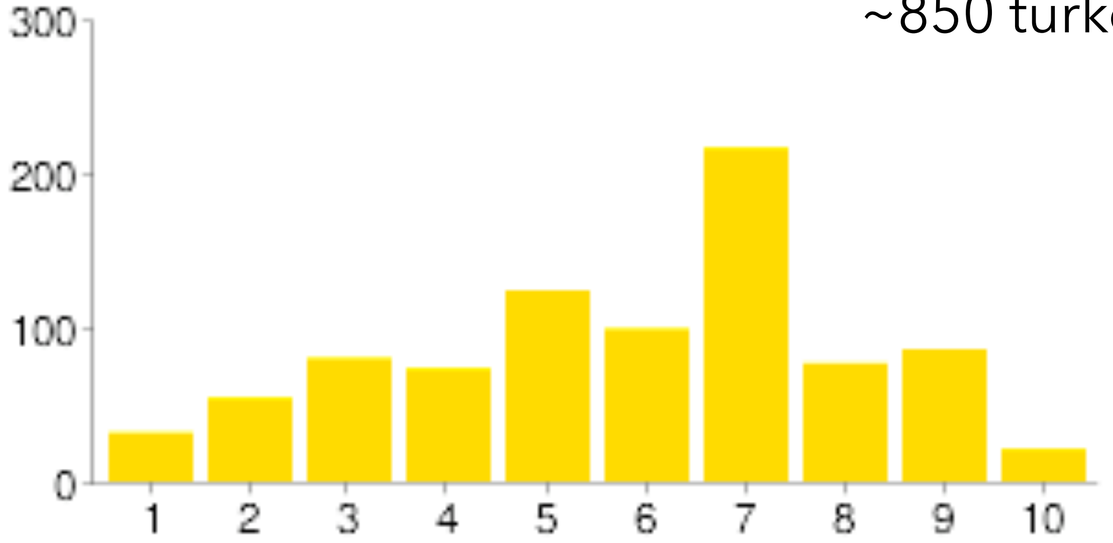
Let’s check a few simple experiments

Source: Isola, Torralba, Freeman

People have biases…

Turkers were offered 1 cent to pick a number from 1 to 10.

~850 turkers



Experiment by Greg Little

From http://groups.csail.mit.edu/uid/denem

Source: Isola, Torralba, Freeman

Do humans have consistent biases?

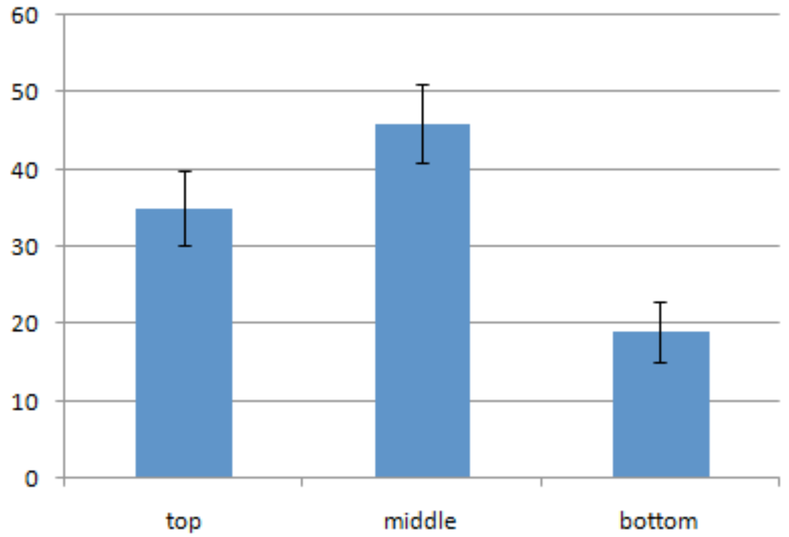
Choose Item

Requester: SimpleSphere **Reward:** $0.01 per HIT **HITs Available:** 1 **Duration:** 60 minutes

Qualifications Required: None

Please choose one of the following:

Results form 100 HITS:

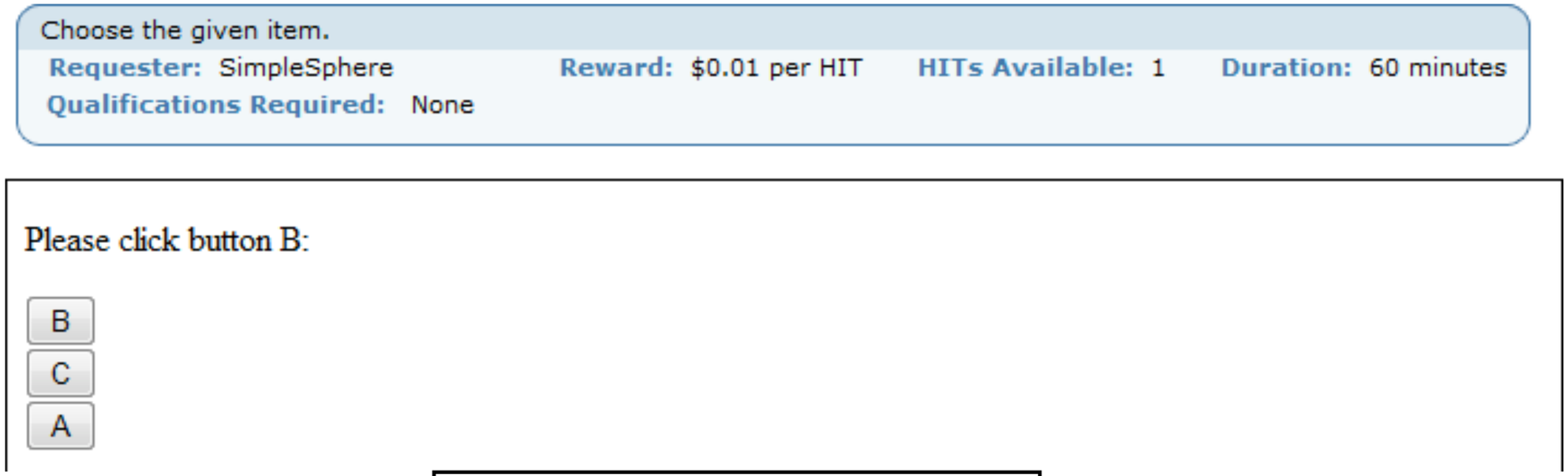


Experiment by Greg Little

From http://groups.csail.mit.edu/uid/denem

Source: Isola, Torralba, Freeman

Are humans reliable even in simple tasks?



Results of 100 HITS:
A: 2
B: 96
C: 2

Experiment by Greg Little

From http://groups.csail.mit.edu/uid/denem

Source: Isola, Torralba, Freeman

Do humans do what you ask for?

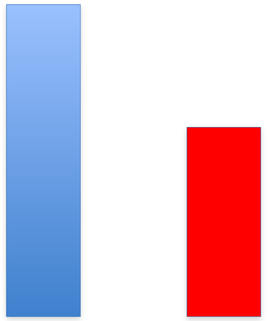
Flip a coin

Requester: ROBERT C MILLER **Reward:** $0.01 per HIT **HITs Available:** 3 **Duration:** 5 minutes

Qualifications Required: None

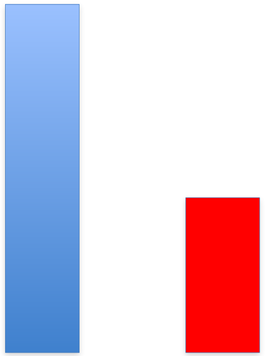
Please flip an actual coin and type either H or T below.

After 50 HITS:



31 heads, 19 tails

And 50 more:



34 heads, 16 tails

Experiment by Rob Miller

From http://groups.csail.mit.edu/uid/denem

Source: Isola, Torralba, Freeman

Name that dataset game

Caltech101 ____ Tiny ____

MSRC ____ Corel ____

UIUC ____ PASCAL 07 ____

LabelMe ____ 15 Scenes ____

COIL−100 ____ Caltech256 ____

ImageNet ____ SUN09 ____

[Torralba and Efros, “An unbiased look at dataset bias,” 2011

So we can sometimes collect good training data.

But suppose we messed up. Our test setting doesn't look like the training data!

How can we bridge the domain gap?

Source: Isola, Torralba, Freeman

Idea: Find more representative images

[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017

Idea: Find more representative images

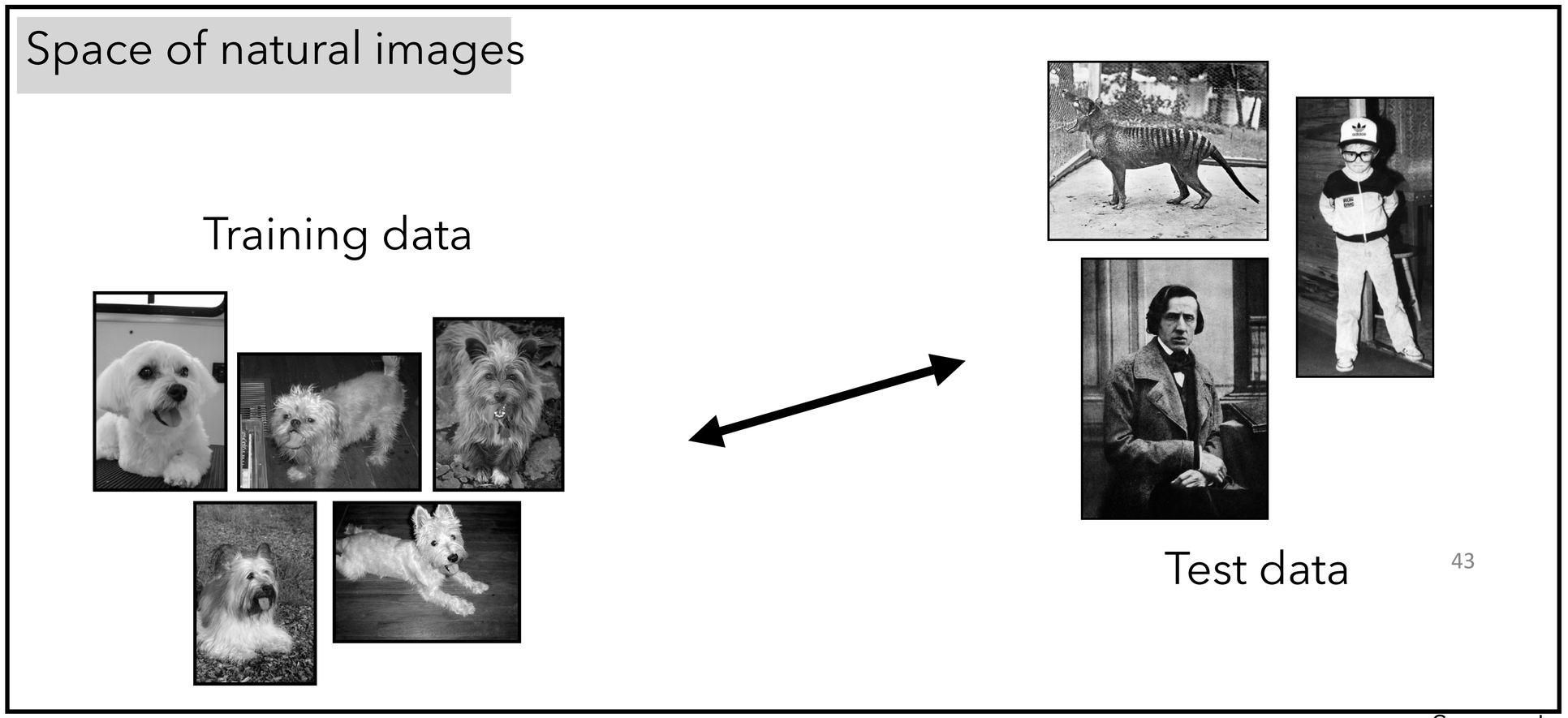
[Fouhey et al., "From Lifestyle Vlogs to Everyday Actions", 2017

Another idea: Domain adaptation

training domain

testing domain
(where we actual use our model)

Domain gap between $p_{\texttt{train}}$ and $p_{\texttt{test}}$ will cause us to fail to generalize.



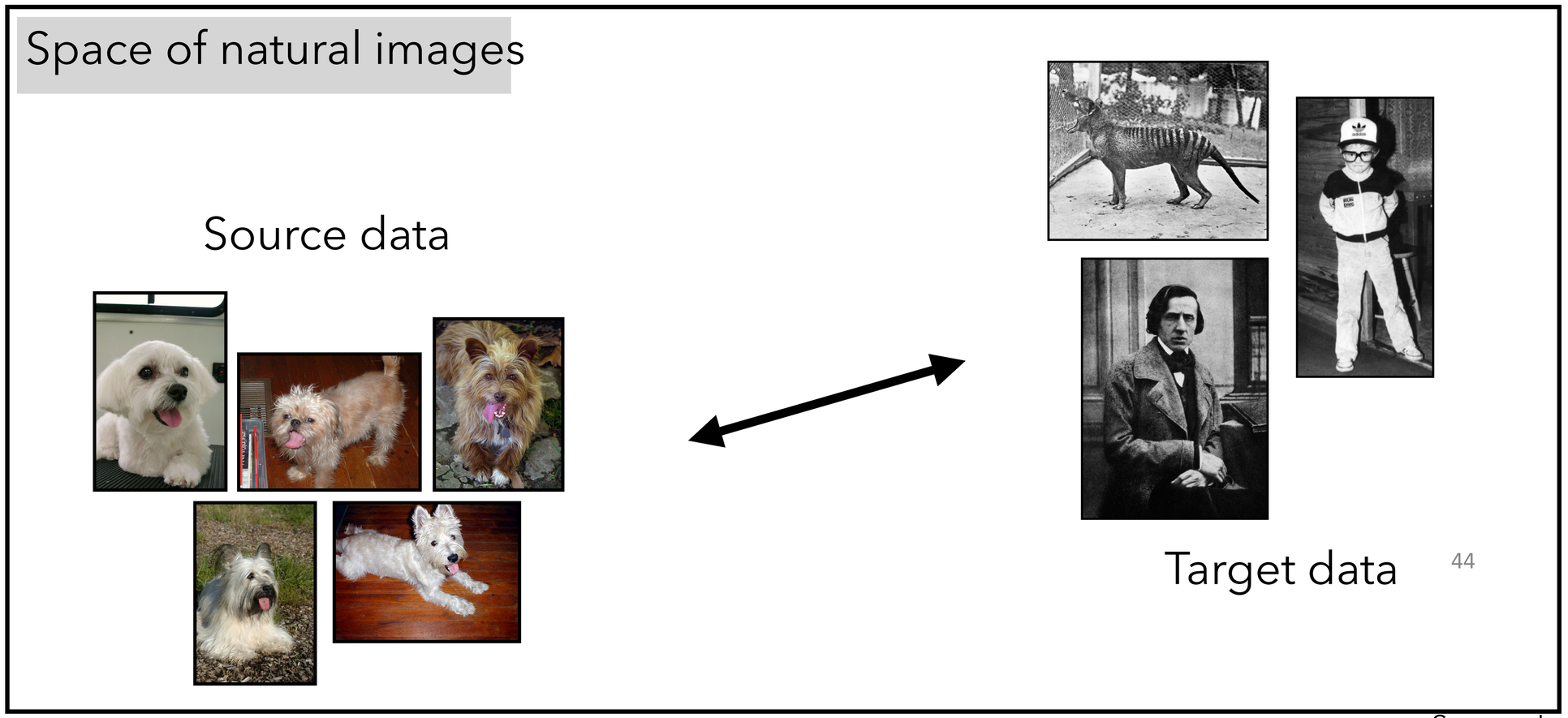
Source: Isola, Torralba, Freeman

source domain

target domain

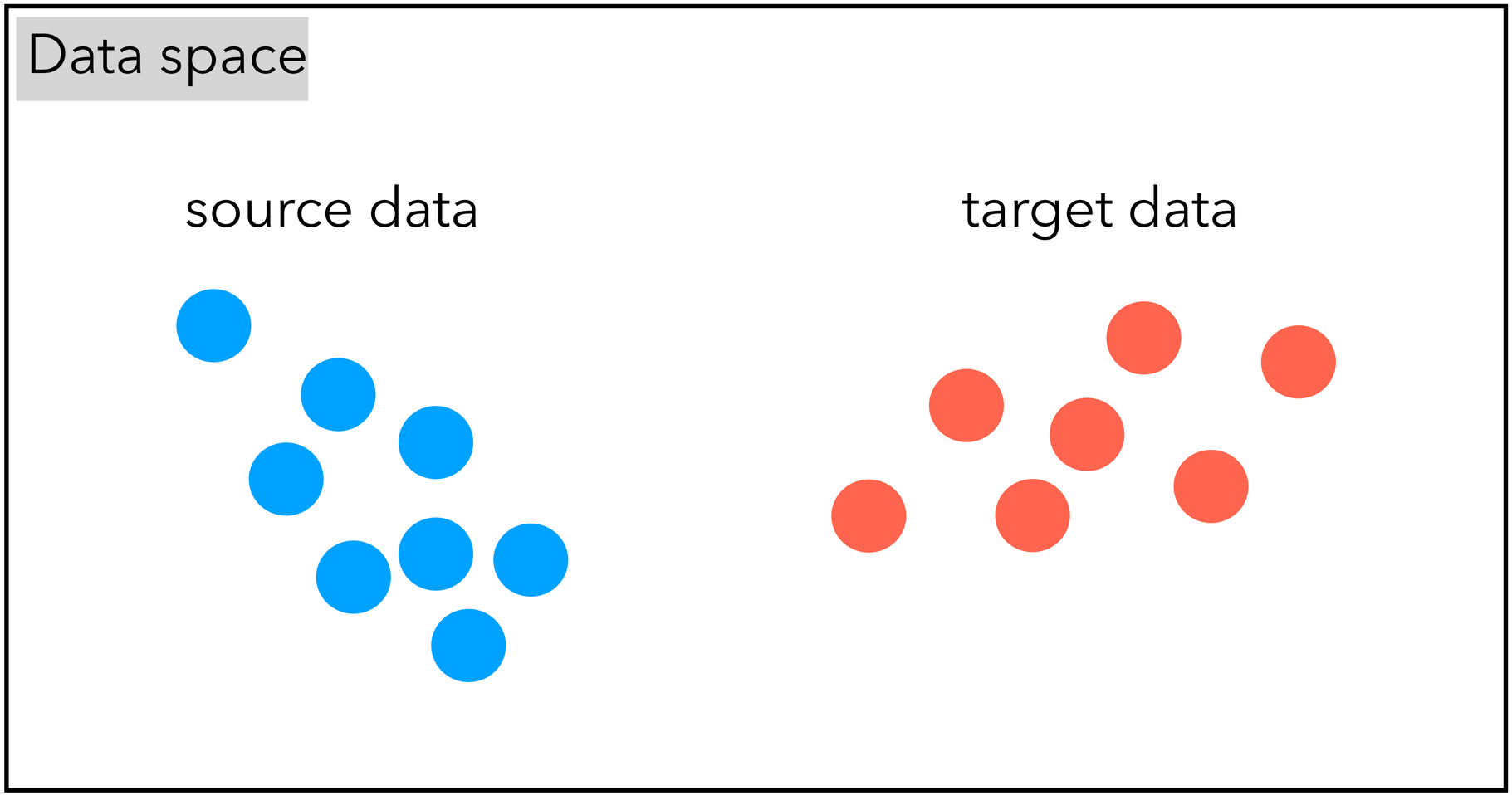
(where we actual use our model)

Domain gap between $p_{\texttt{source}}$ and $p_{\texttt{target}}$ will cause us to fail to generalize.



Source: Isola, Torralba, Freeman

Idea #1: transform the target domain to look like the source domain



(Or vice versa)

This is called **domain adaptation**

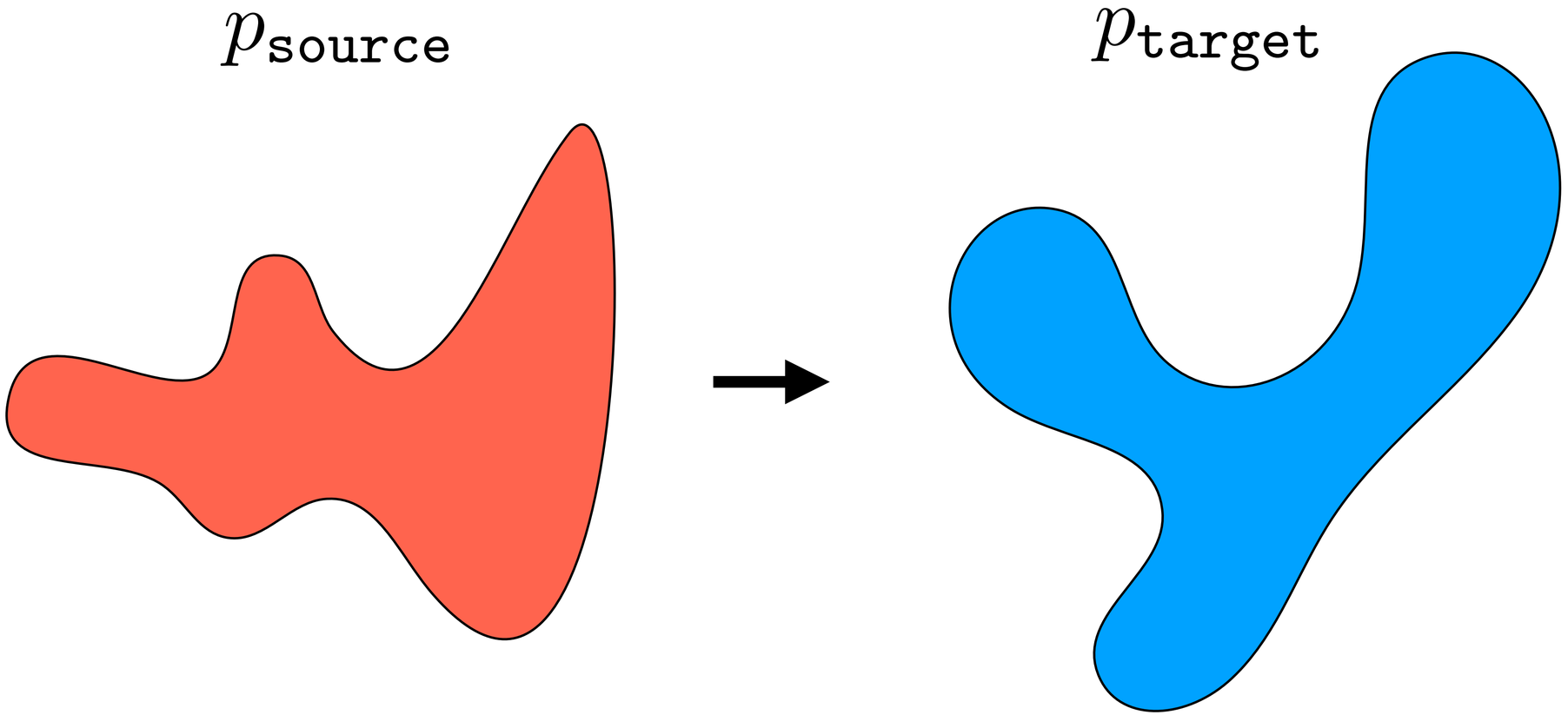
Source: Isola, Torralba, Freeman

Domain adaptation

- We have source domain pairs {$\mathbf{x}^{\mathbf{source}}$, $\mathbf{y}^{\mathbf{source}}$}
- Learn a mapping F: $\mathbf{x}^{\mathbf{source}}$ –> $\mathbf{y}^{\mathbf{source}}$
- We want to apply F to target domain data $\mathbf{x}^{\mathbf{target}}$
- Find transformation T: $\mathbf{x}^{\mathbf{target}}$ –> $\mathbf{x}^{\mathbf{source}}$
- Now apply F(T($\mathbf{x}^{\mathbf{target}}$)) to predict $\mathbf{y}^{\mathbf{target}}$

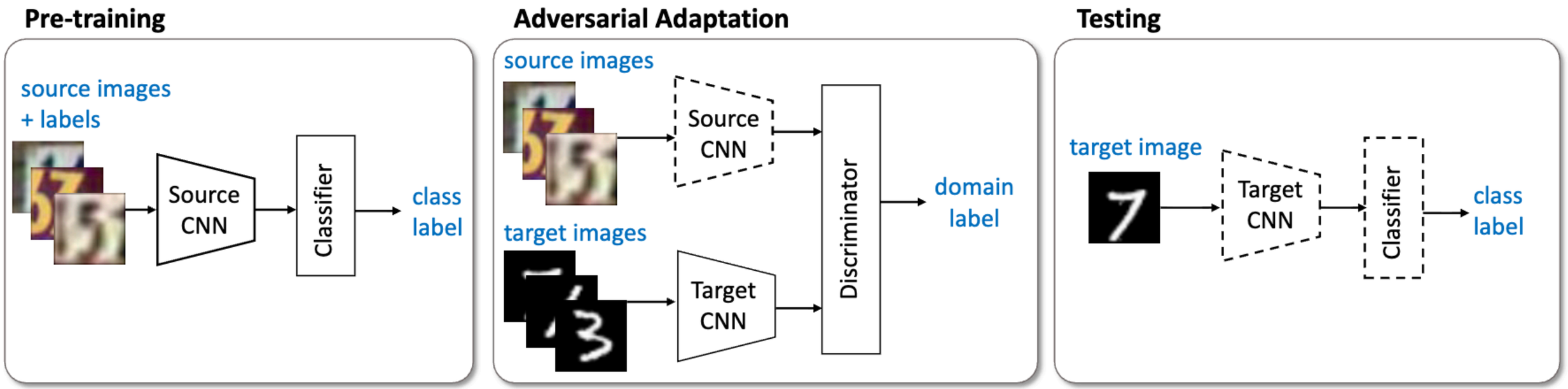
Source: Isola, Torralba, Freeman

Domain adaptation



Source: Isola, Torralba, Freeman

Domain adaptation

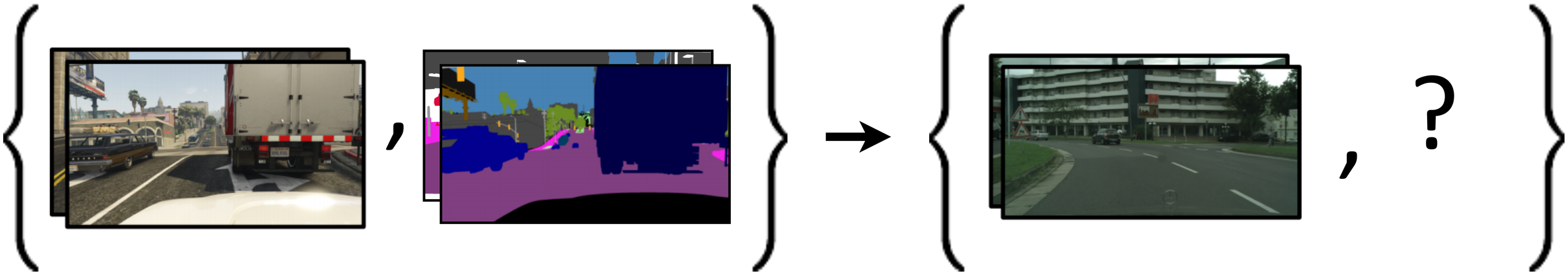


[Tzeng et al., "Adversarial Discriminative Domain Adaptation", CVPR 2017]

Example: Domain Adaptation by Image Translation

Source domain

Target domain

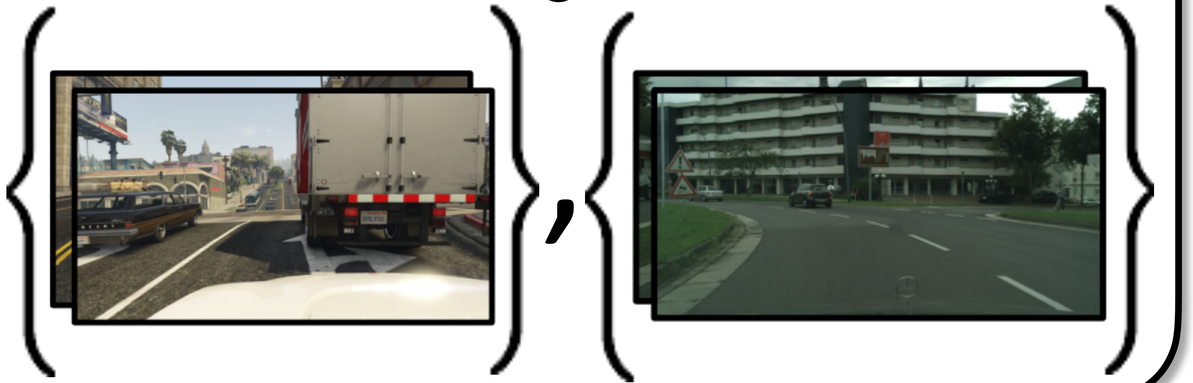


[Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Darrell, Efros, “CyCADA”, 2017]

Source: Isola, Torralba, Freeman

CycleGAN

Training data



Source: Isola, Torralba, Freeman

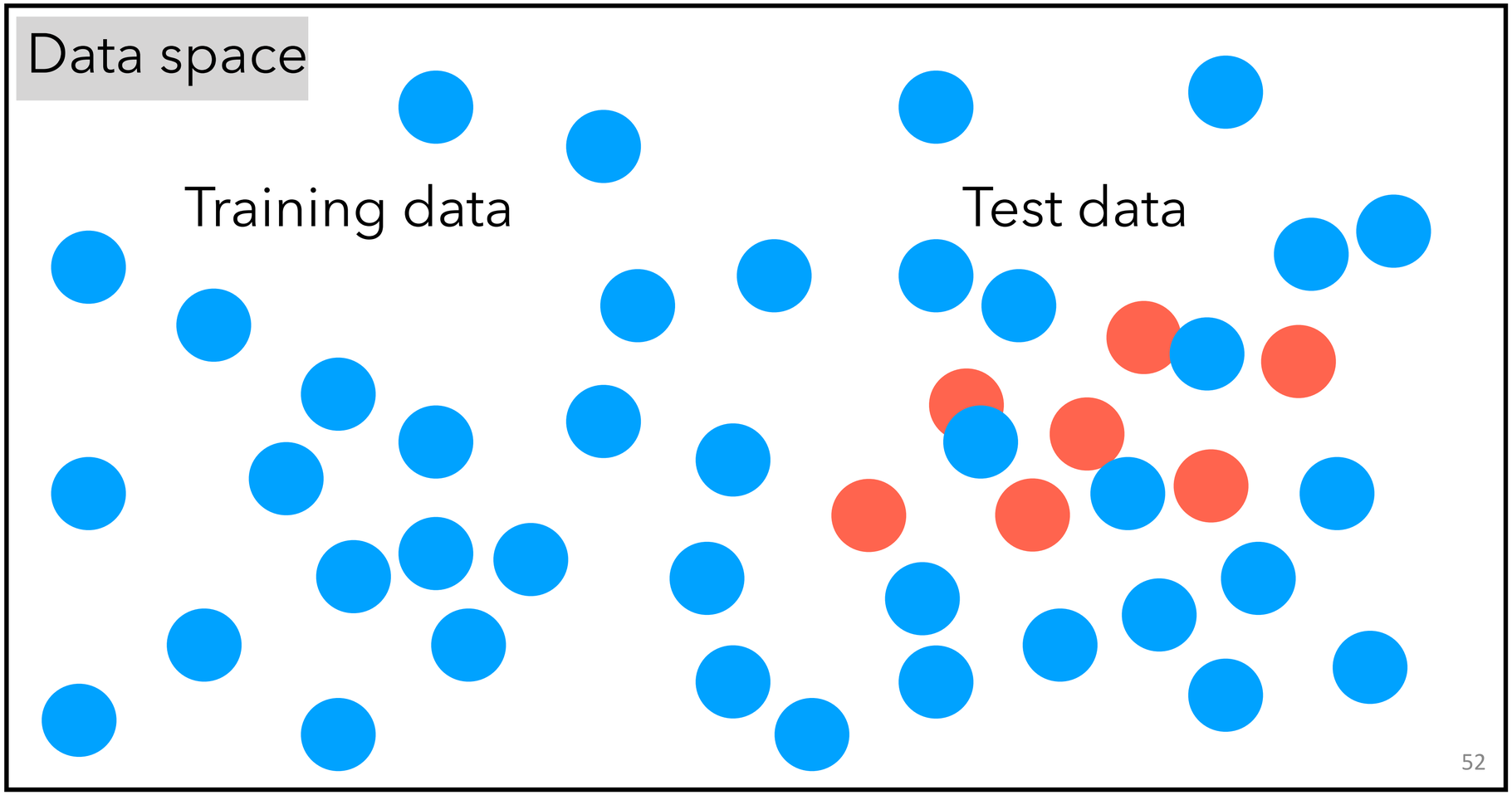
CycleGAN

Training data

,

Source: Isola, Torralba, Freeman

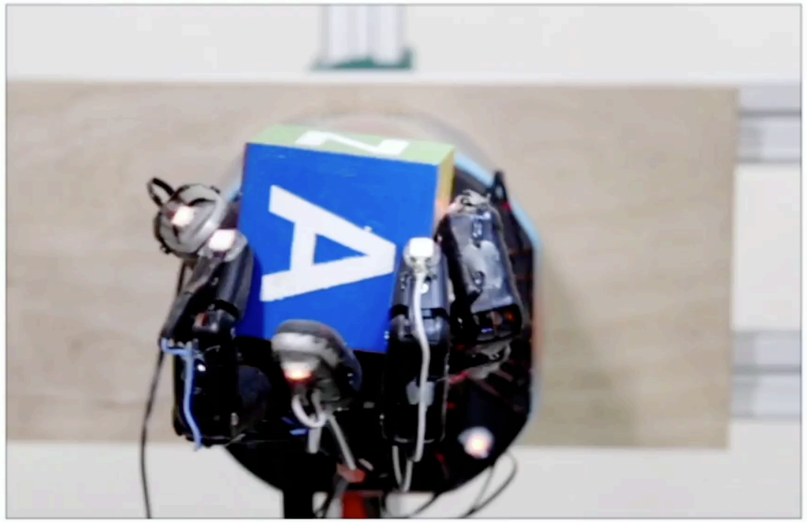
Idea #2: train on randomly perturbed data, so that test set just looks like another random perturbation



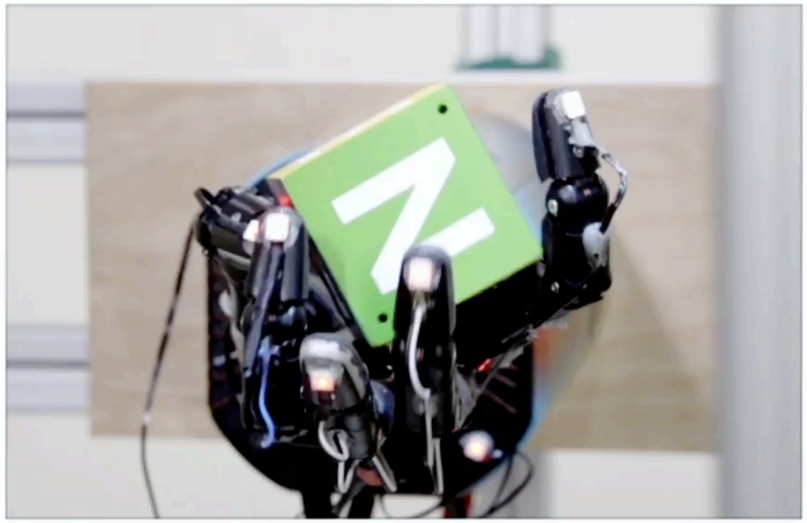
This is called **domain randomization** or **data augmentation**

Source: Isola, Torralba, Freeman

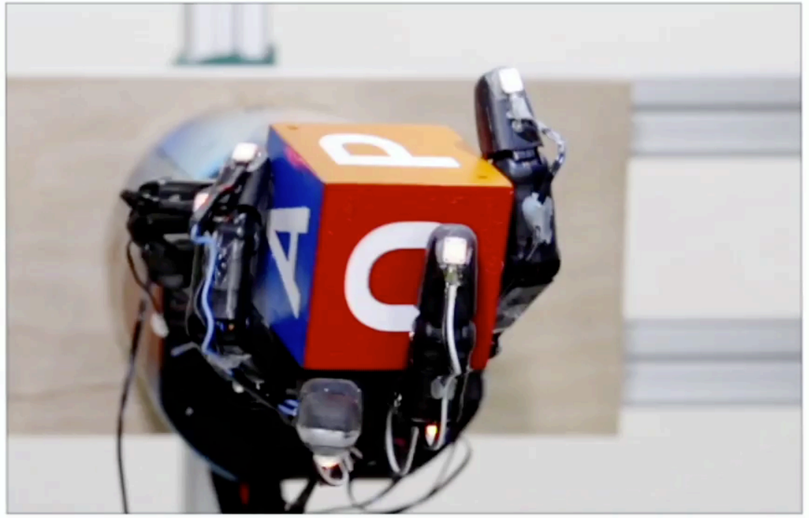
OpenAI Dactyl



FINGER PIVOTING



SLIDING



FINGER GAITING

Source: Isola, Torralba, Freeman

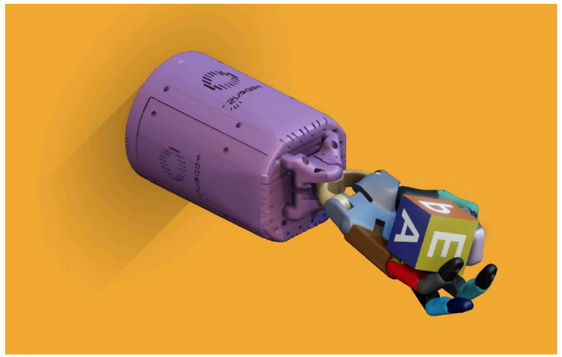
source domain

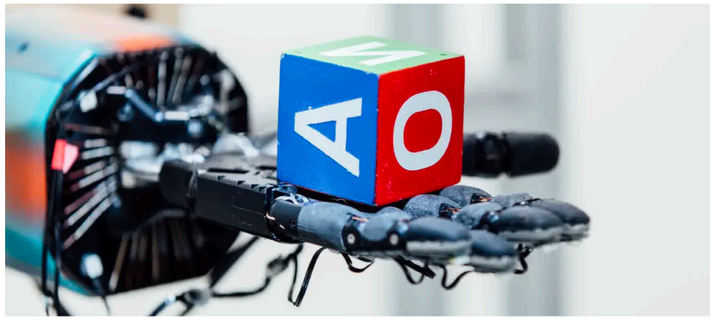
target domain
(where we actual use our model)

Domain gap between $p_{\texttt{source}}$ and $p_{\texttt{target}}$ will cause us to fail to generalize.

Space of images

Source data



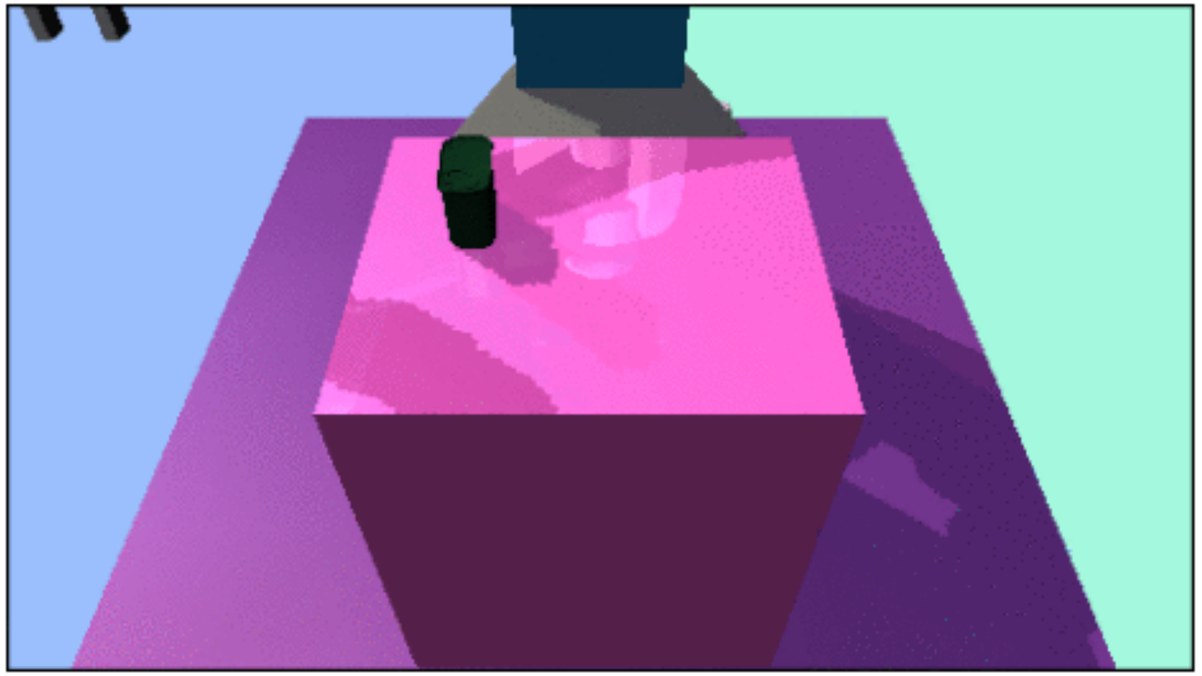


Target data

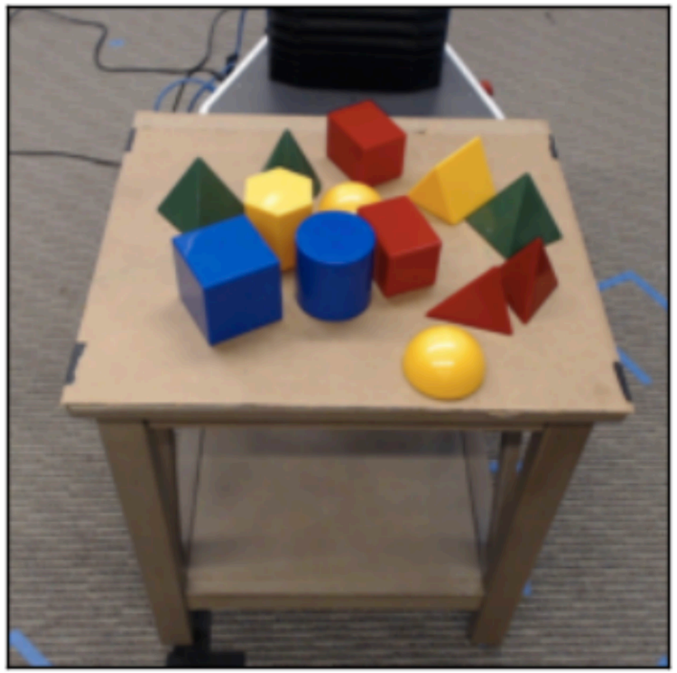
Source: Isola, Torralba, Freeman

Domain randomization

Training data



Test data



[Sadeghi & Levine 2016]

Above example is from [Tobin et al. 2017]

Source: Isola, Torralba, Freeman

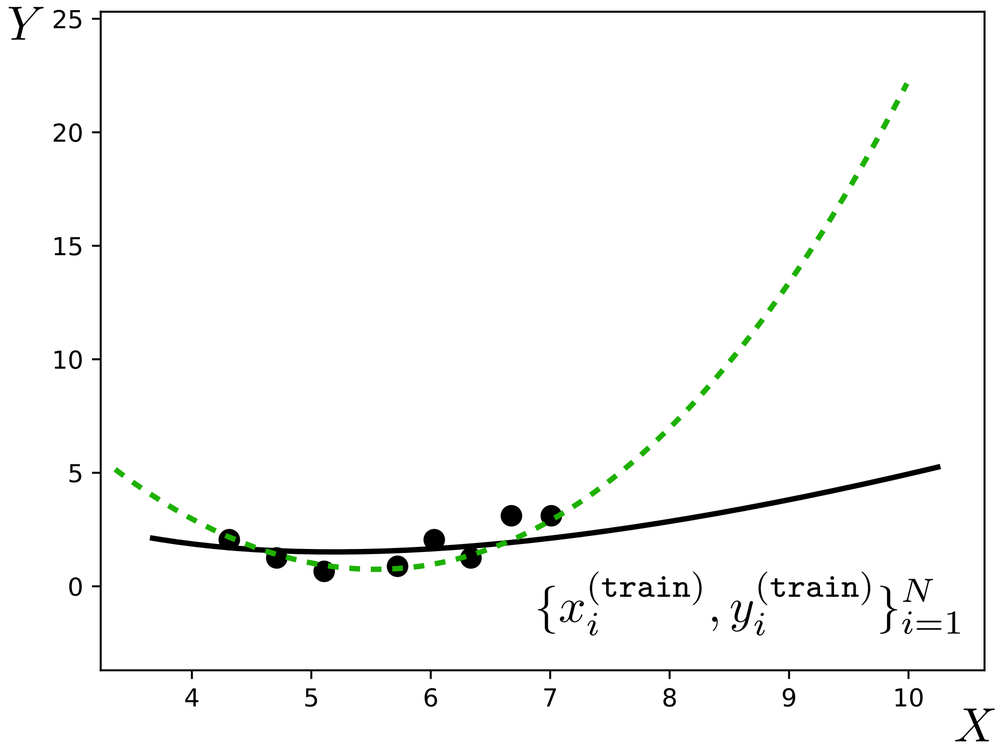
Other sources of bias, besides data

- Data very important [Maluleke et al., 2022], but also other factors can matter.
- Camera hardware and software
  - e.g., default camera settings calibrated to expose light skin
- Loss function (e.g., “mode collapse” in GANs)
- Features
- Sampling strategy (e.g., truncation in GANs)

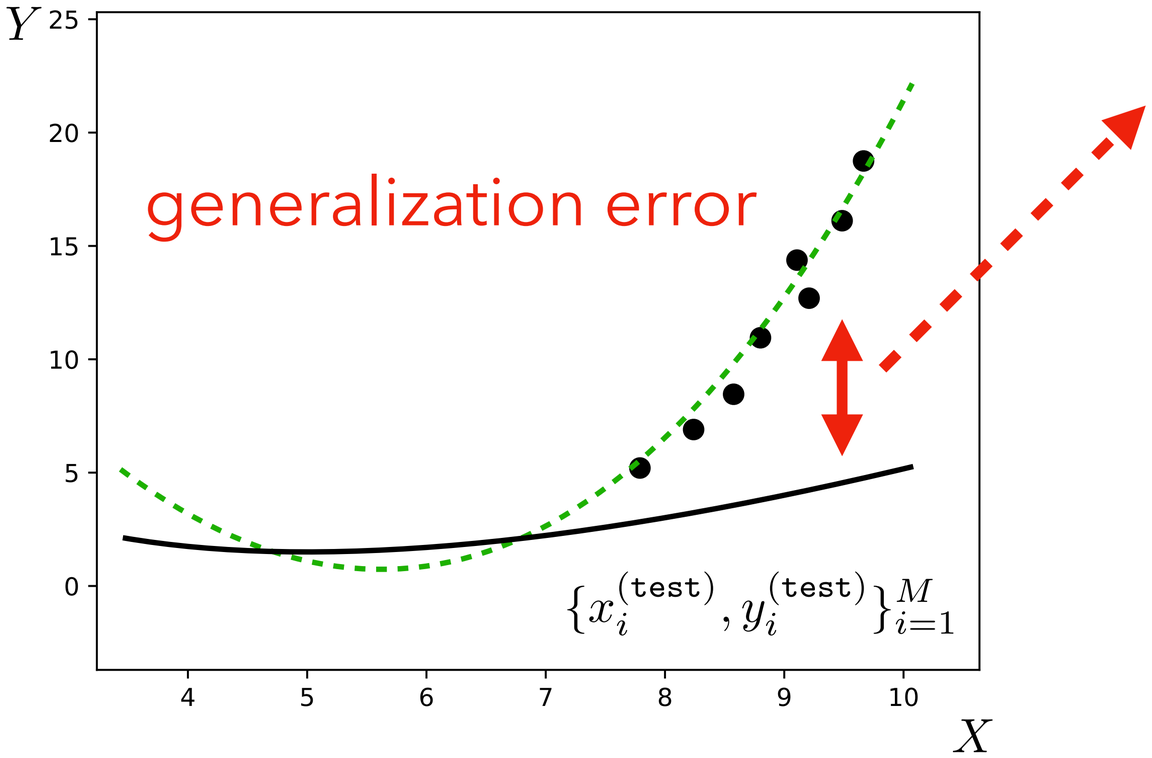
What if we go way outside of the training distribution?

Source: Isola, Torralba, Freeman

Training data



Test data



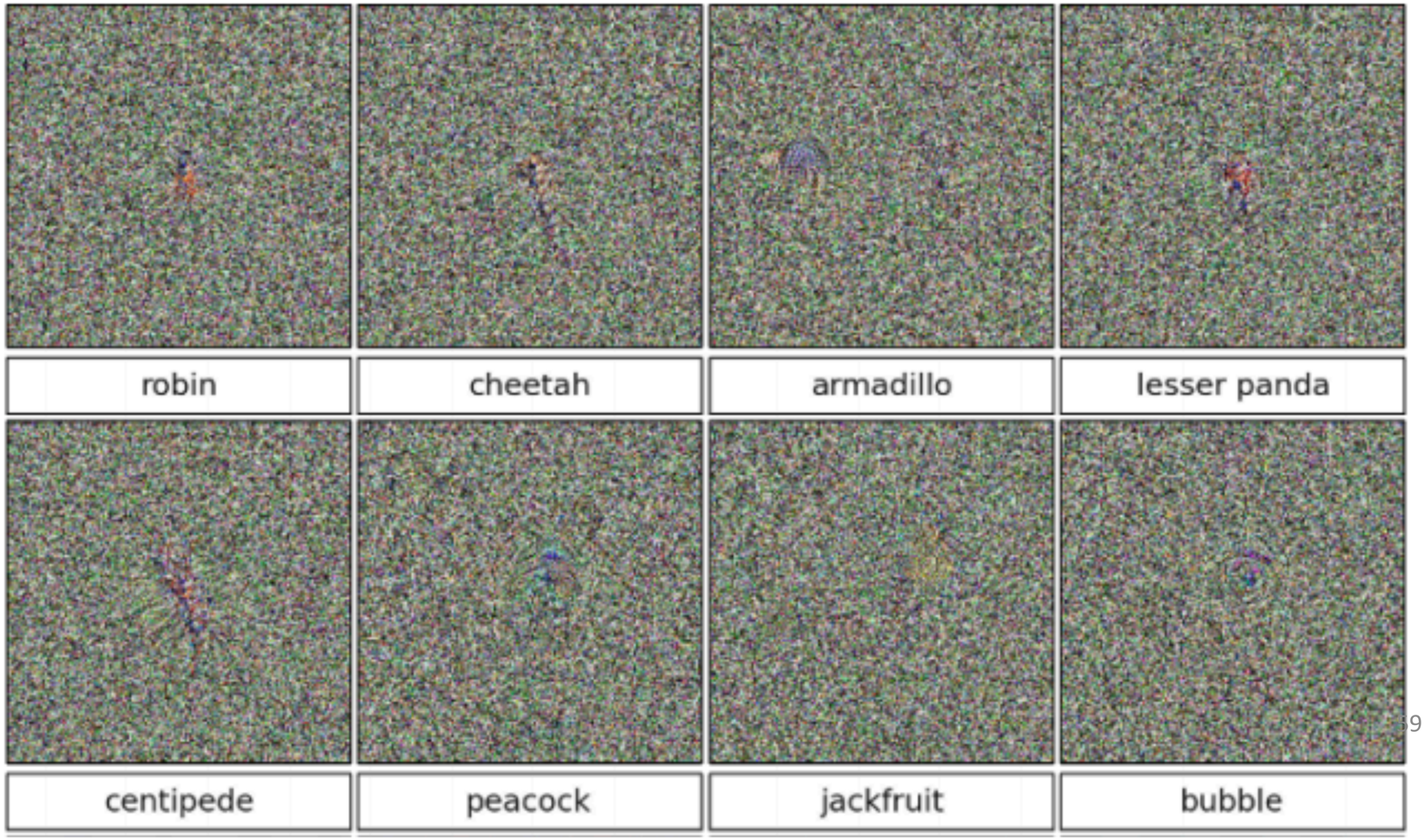
Our training data did not cover the part of the distribution that was tested

(**biased data**)

Source: Isola, Torralba, Freeman

“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”

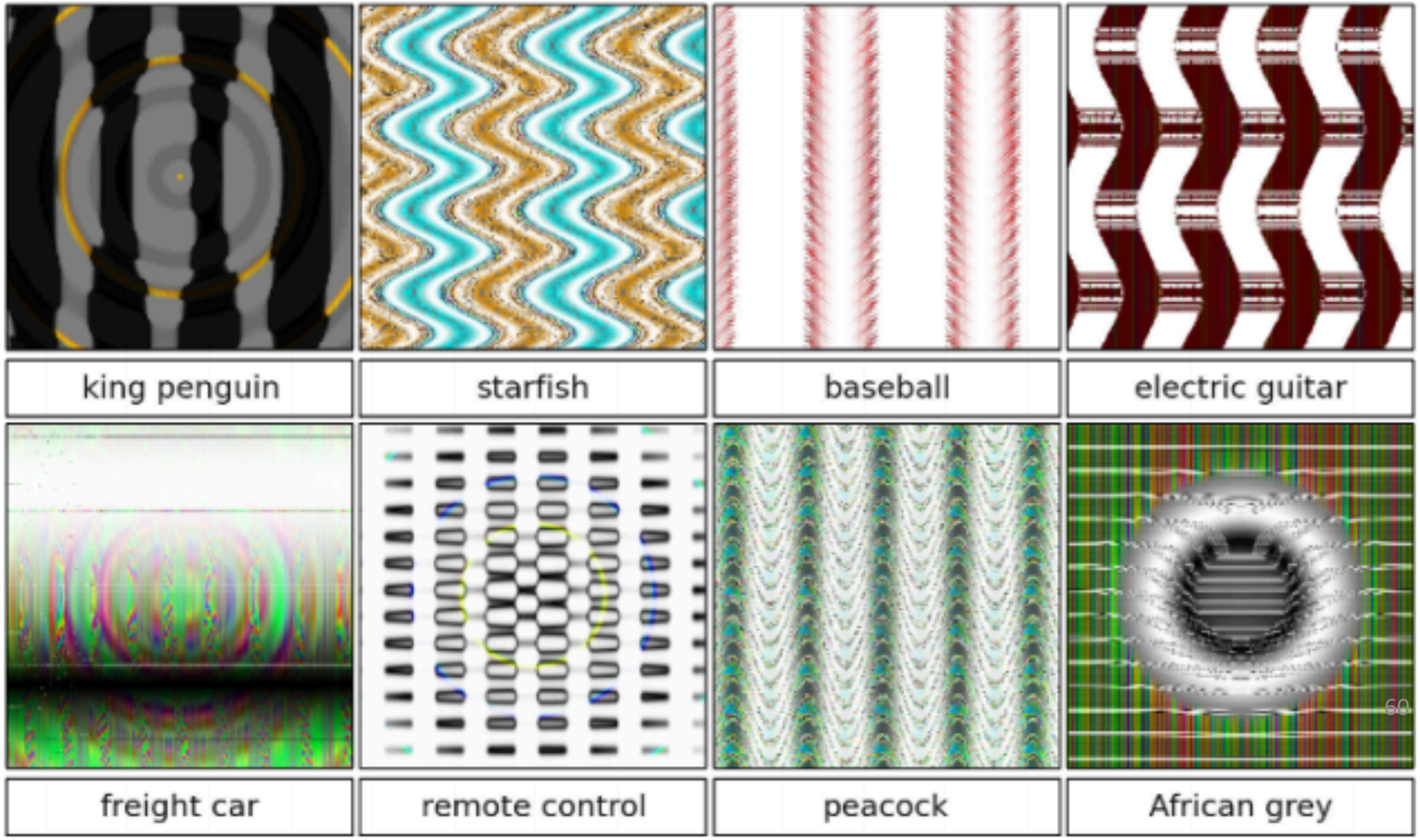
[Nguyen, Yosinski, and Clune, CVPR 2015]



Source: Isola, Torralba, Freeman

“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”

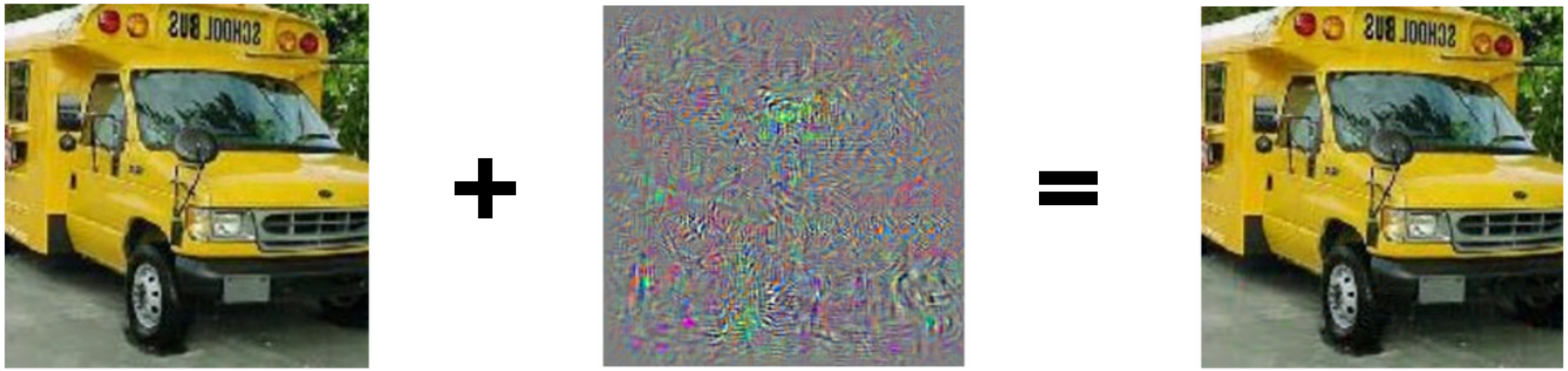
[Nguyen, Yosinski, and Clune, CVPR 2015]



Source: Isola, Torralba, Freeman

Adversarial noise

$\mathbf{x}$ $\mathbf{r}$ $\mathbf{x} + \mathbf{r}$



f f

y "School bus" "Ostrich"

$$\arg\max_{\mathbf{r}} p(y = \texttt{ostrich} | \mathbf{x} + \mathbf{r}) \quad \text{subject to} \quad \|\mathbf{r}\| < \epsilon$$

["Intriguing properties of neural networks", Szegedy et al. 2014

Source: Isola, Torralba, Freeman

Anything to worry about?

"NO Need to Worry about Adversarial Examples in Object Detection in Autonomous Vehicles", Lu et al. 2017

(Early) 2017's attacks fail on physical objects, since they are optimized to attack a single view!

Source: Isola, Torralba, Freeman

Anything to worry about?

Later in 2017…

"Synthesizing Robust Adversarial Examples", Athalye, Engstrom, Ilyas, Kwok, 2017

3D-printed **turtle** model classified as **rifle** from most viewpoints

Source: Isola, Torralba, Freeman

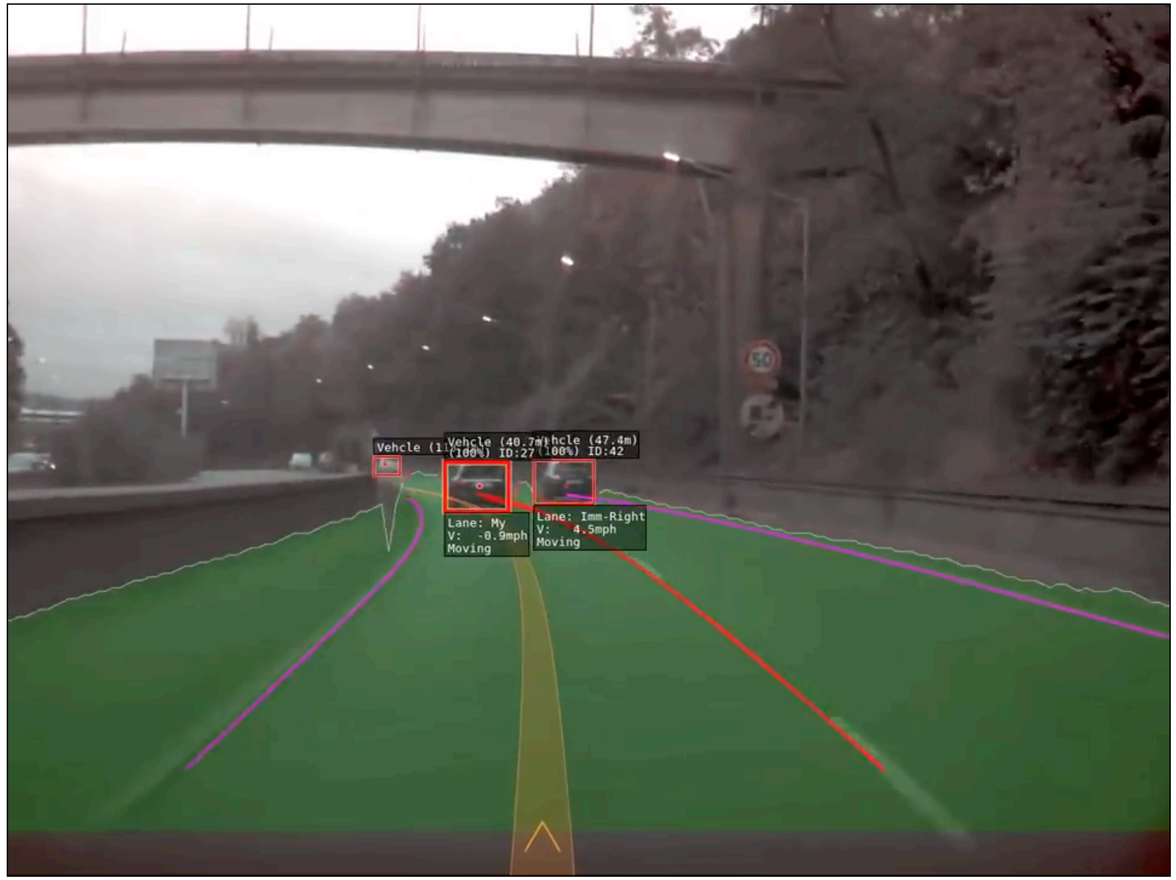
Adversarial examples

- Current deep models have bad **worst-case performance**
- Can be exploited by an adversary
- Few guarantees, can’t fully trust what the model’s output

Source: Isola, Torralba, Freeman

Problems of applying computer vision in practice

Mission-critical computer vision systems



Source: https://www.nytimes.com/2020/02/25/business/tesla-autopilot-ntsb.html, https://qz.com/1402843/watch-teslas-autopilot-see-the-streets-of-paris/

Social consequences

Color Matters in Computer Vision

Facial recognition algorithms made by Microsoft, IBM and Face++ were more likely to misidentify the gender of black women than white men.

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

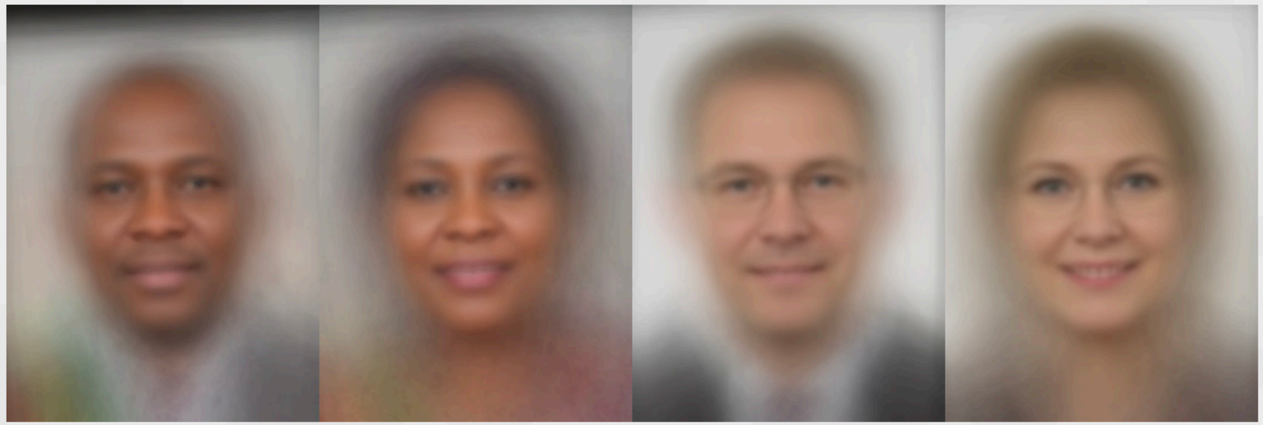
Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

Source: Isola, Torralba, Freeman

Algorithmic Bias

Gender Classifier	Darker Male	Darker Female	Lighter Male	Lighter Female	Largest Gap
Microsoft	94.0%	79.2%	100%	98.3%	20.8%
FACE++	99.3%	65.5%	99.2%	94.0%	33.8%
IBM	88.0%	65.3%	99.7%	92.9%	34.4%



http://gendershades.org/overview.html

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Editors: Sorelle A. Friedler and Christo Wilson

Abstract

Recent studies demonstrate that machine learning algorithms can discriminate based on classes like race and gender. In this work, we present an approach to evaluate bias present in automated facial analysis algorithms and datasets with respect to phenotypic subgroups. Using the dermatologist approved Fitzpatrick Skin Type classification system, we characterize the gender and skin type distribution of two facial analysis benchmarks, IJB-A and Adience. We find that these datasets are overwhelmingly composed of lighter-skinned subjects (79.6% for IJB-A and 86.2% for Adience) and introduce a new facial analysis dataset which is balanced by gender and skin type. We evaluate 3 commercial gender classification systems using our dataset and show that darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%. The substantial disparities in the accuracy of classifying darker females, lighter females, darker males, and lighter males in gender classification systems require urgent attention if commercial companies are to build genuinely fair, transparent and accountable facial analysis algorithms.

Keywords: Computer Vision, Algorithmic Audit, Gender Classification

1. Introduction

Artificial Intelligence (AI) is rapidly infiltrating every aspect of society. From helping determine who is hired, fired, granted a loan, or how long an individual spends in prison, decisions that have traditionally been performed by humans are rapidly made by algorithms (O'Neil, 2017; Citron and Pasquale, 2014). Even AI-based technologies that are not specifically trained to perform high-stakes tasks (such as determining how long someone spends in prison) can be used in a pipeline that performs such tasks. For example, while face recognition software by itself should not be trained to determine the fate of an individual in the criminal justice system, it is very likely that such software is used to identify suspects. Thus, an error in the output of a face recognition algorithm used as input for other tasks can have serious consequences. For example, someone could be wrongfully accused of a crime based on erroneous but confident misidentification of the perpetrator from security video footage analysis.

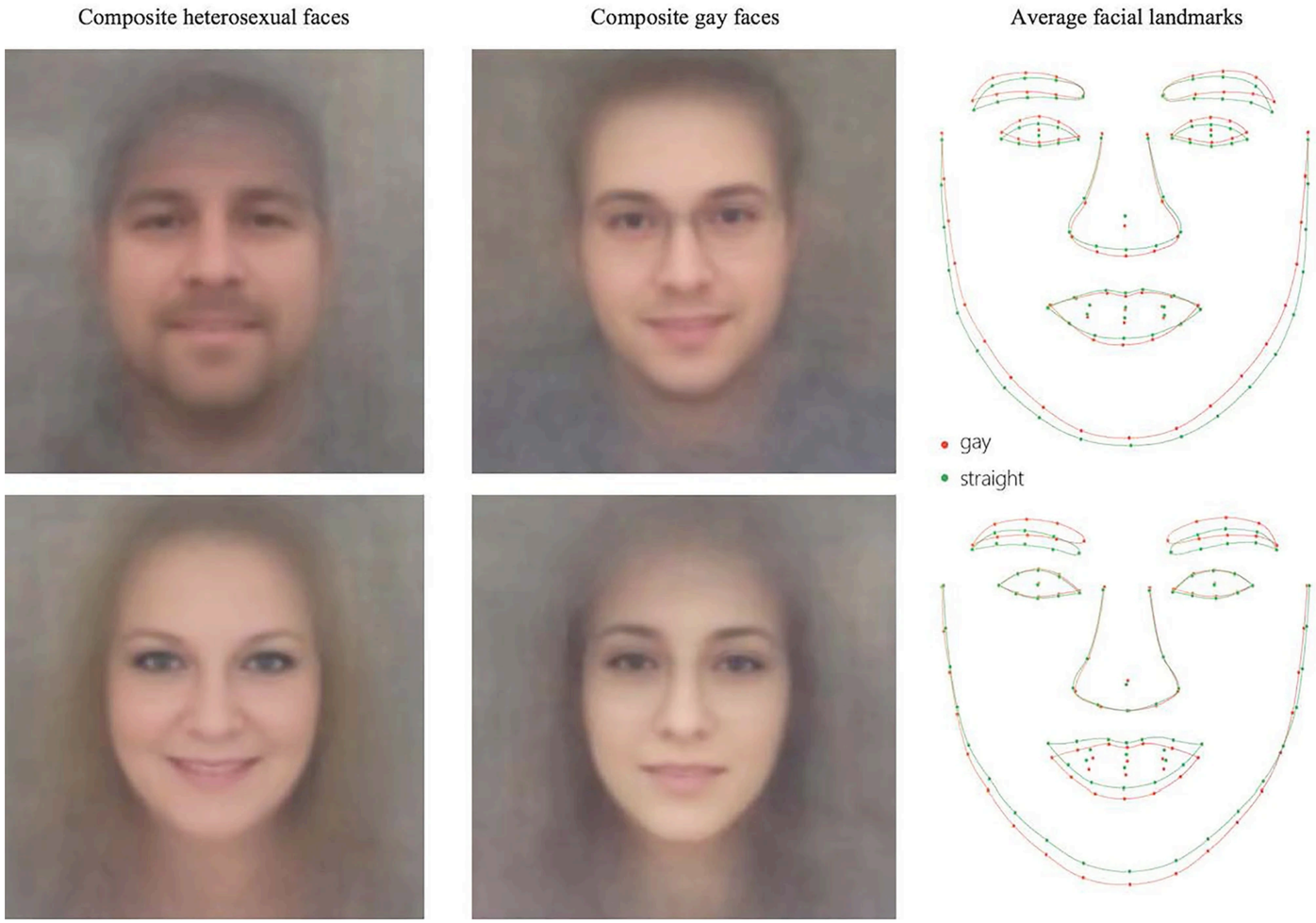
Many AI systems, e.g. face recognition tools, rely on machine learning algorithms that are trained with labeled data. It has recently been shown that algorithms trained with biased data have resulted in algorithmic discrimination (Bolukbasi et al., 2016; Caliskan et al., 2017). Bolukbasi et al. even showed that the popular word embedding space, Word2Vec, encodes societal gender biases. The authors used Word2Vec to train an analogy generator that fills in missing words in analogies. The analogy man is to computer programmer as woman is to "X" was completed with "homemaker", conforming to the stereotype that programming is associated with men and homemaking with women. The biases in Word2Vec are thus likely to be propagated throughout any system that uses this embedding.

* Download our gender and skin type balanced PPB dataset at gendershades.org

http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Source: Isola, Torralba, Freeman

Bad choice of data



https://www.nytimes.com/2017/10/09/science/stanford-sexual-orientation-study.html

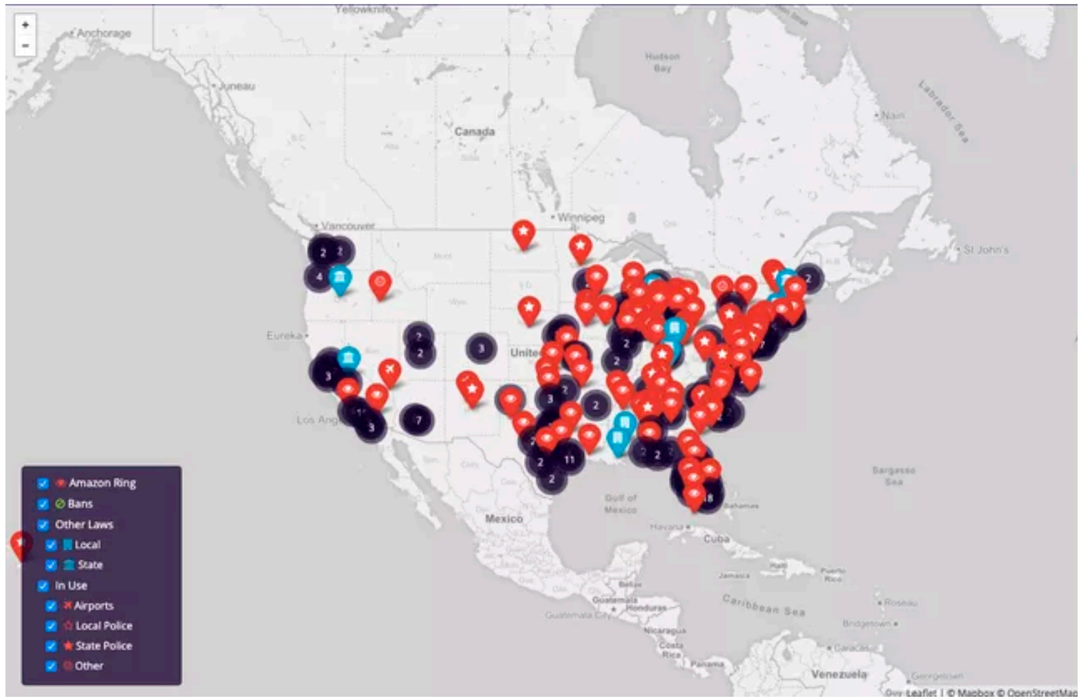
Face recognition in the U.S.

recode

Here's where the US government is using facial recognition technology to surveil Americans

This map shows how widespread the use of facial recognition technology has become.

By Shirin Ghaffary and Rani Molla | Updated Dec 10, 2019, 8:00am EST



https://www.vox.com/recode/2019/7/18/20698307/facial-recognition-technology-us-government-fight-for-the-future

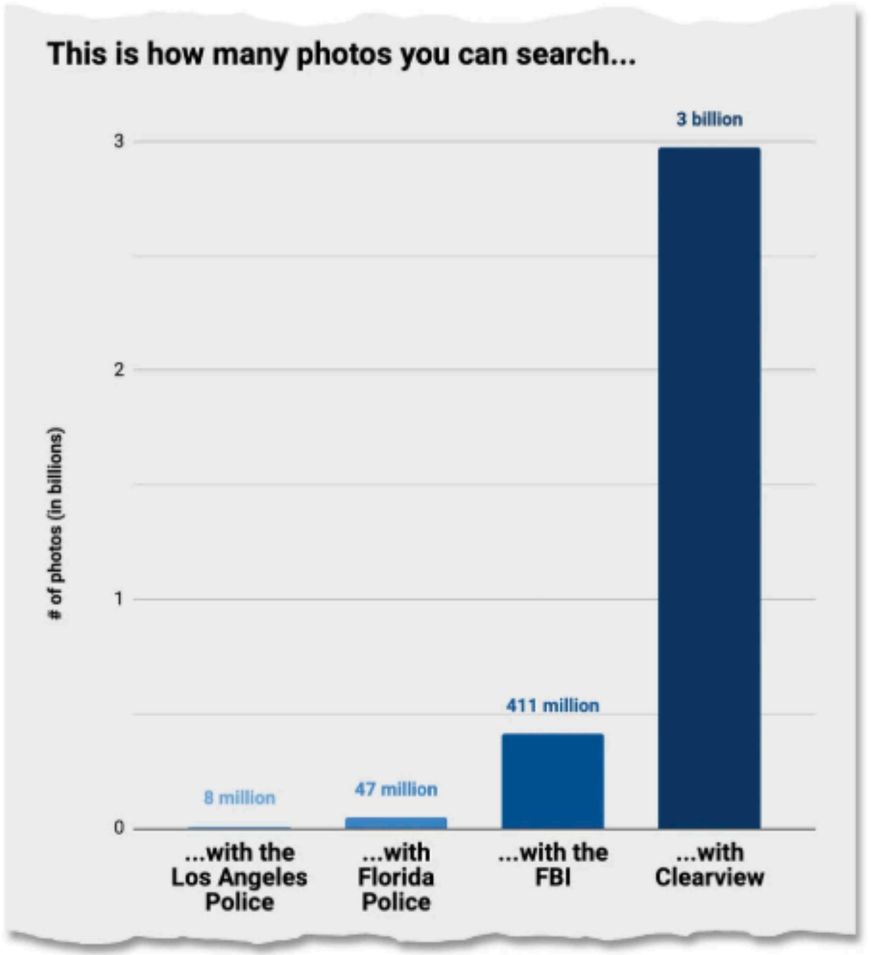
Source: S. Lazebnik

Fears of universal mass surveillance (and dubious claims)

The New York Times

The Secretive Company That Might End Privacy as We Know It

A little-known start-up helps law enforcement match photos of unknown people to their online images — and "might lead to a dystopian future or something," a backer says.



A chart from marketing materials that Clearview provided to law enforcement. Clearview

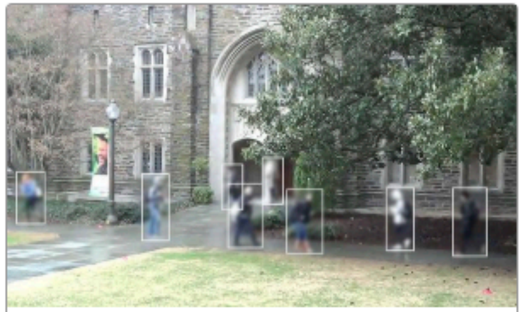
https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html

https://www.buzzfeednews.com/article/ryanmac/clearview-ai-nypd-facial-recognition

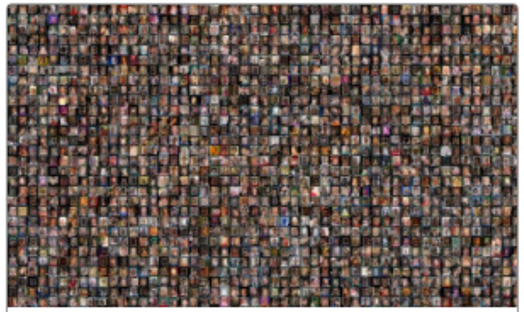
Source: S. Lazebnik

Datasets: Privacy, consent issues

Brainwash Dataset Analysis
2015
Head detection
11,917 images

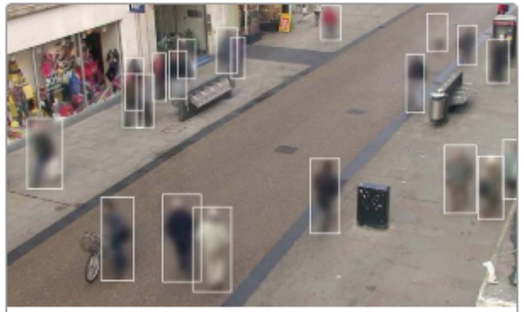


Duke MTMC Dataset Analysis
2016
Person re-identification, multi-camera tracking
2,000,000 images

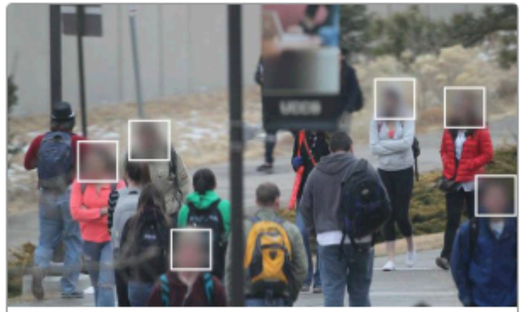


MegaFace Dataset Analysis
2016
face recognition
4,753,520 images

Microsoft Celeb Dataset Analysis
2016
Face recognition
8,200,000 images



Oxford Town Centre Dataset Analysis
2009
Person detection, gaze estimation



UnConstrained College Students Dataset Analysis
2016
Face recognition, face detection
16,149 images

https://megapixels.cc/

https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e

See also: https://www.theregister.co.uk/2020/01/27/ibms_facial_recognition_software_gets_it_in_trouble_again/

Source: S. Lazebnik

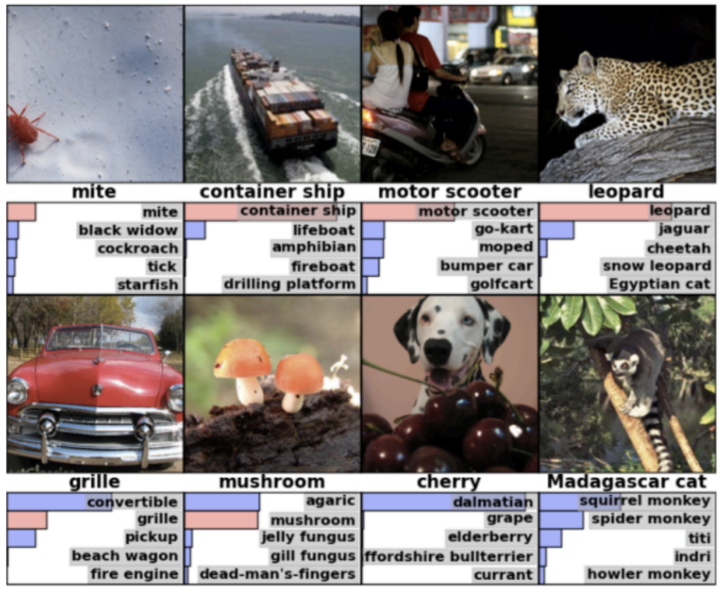
Face recognition in the U.S.

Wrongfully Accused by an Algorithm – New York Times, 6/24/2020

Source: S. Lazebnik

ImageNet: asset or liability?

- Performance on the basic ILSVRC benchmark has saturated



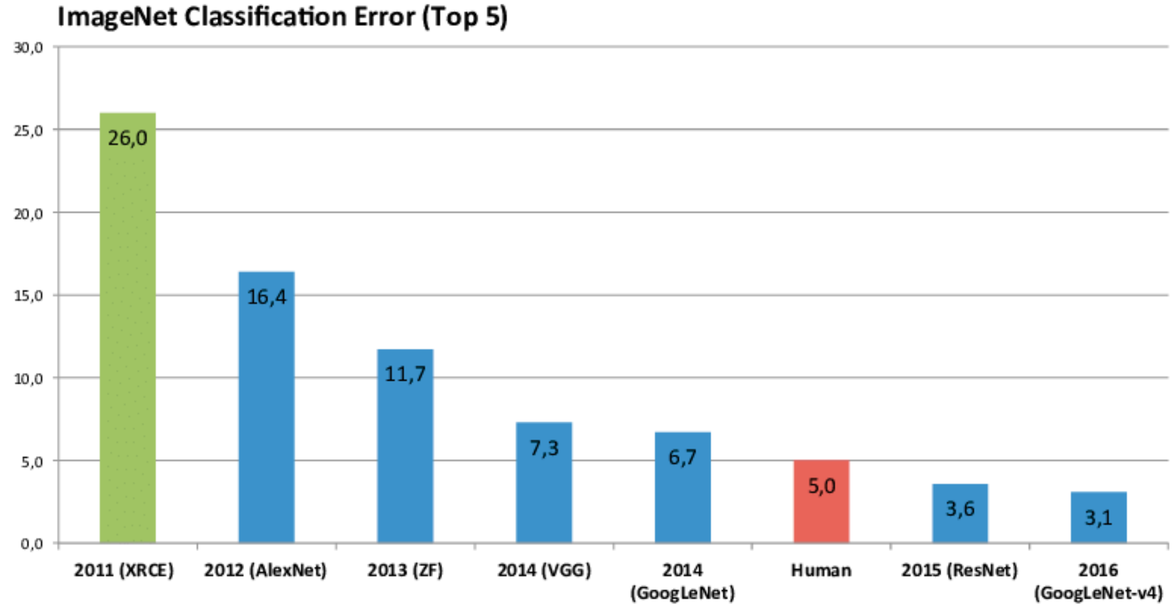


Figure source

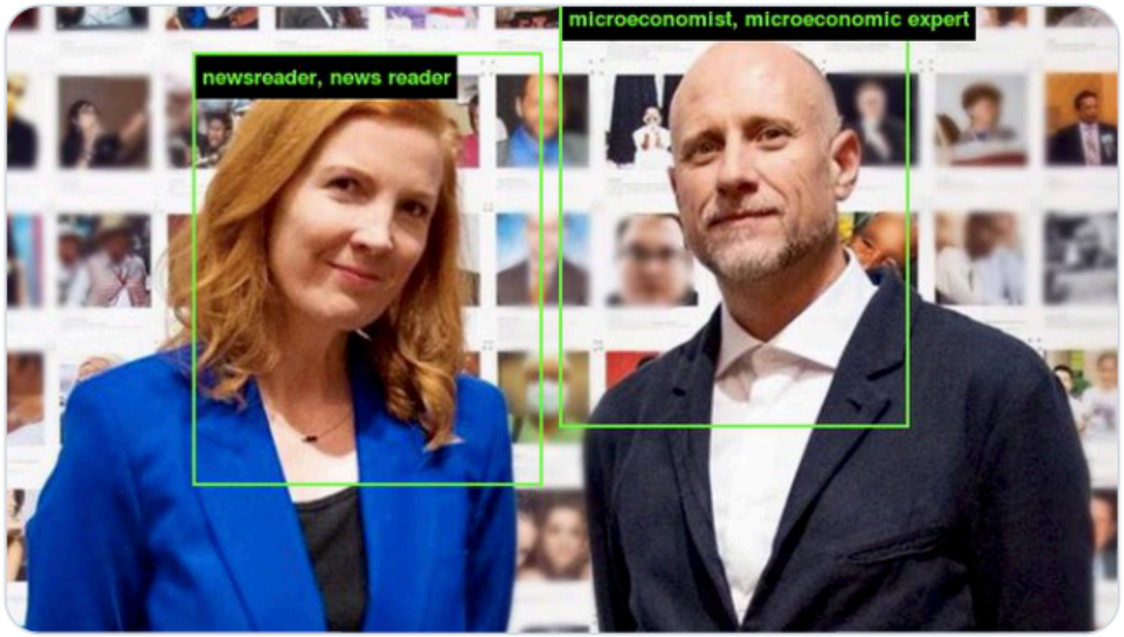
- Current models have reached levels of accuracy where the presence of human labeling error is starting to affect experimental conclusions (Beyer et al. 2020, Northcutt et al. 2021)

Source: S. Lazebnik

ImageNet labeling problems: ImageNet Roulette

Kate Crawford @katecrawford · Sep 16, 2019

Want to see how an AI trained on ImageNet will classify you? Try ImageNet Roulette, based on ImageNet's Person classes. It's part of the 'Training Humans' exhibition by @trevorpaglen & me - on the history & politics of training sets. Full project out soon imagenet-roulette.paglen.com



ImageNet Roulette uses an open source Caffe deep learning framework (produced at UC Berkeley) trained on the images and labels in the "person" categories (which are currently 'down for maintenance'). Proper nouns and categories with less than 100 pictures were removed.

When a user uploads a picture, the application first runs a face detector to locate any faces. If it finds any, it sends them to the Caffe model for classification. The application then returns the original images with a bounding box showing the detected face and the label the classifier has assigned to the image. If no faces are detected, the application sends the entire scene to the Caffe model and returns an image with a label in the upper left corner.

ImageNet contains a number of problematic, offensive and bizarre categories - all drawn from WordNet. Some use misogynistic or racist terminology. Hence, the results ImageNet Roulette returns will also draw upon those categories. That is by design: we want to shed light on what happens when technical systems are trained on problematic training data. AI classifications of people are rarely made visible to the people being classified. ImageNet Roulette provides a glimpse into that process – and to show the ways things can go wrong.

K. Crawford and T. Paglen, Excavating AI: The Politics of Training Sets for Machine Learning, September 2019 https://www.theverge.com/tldr/2019/9/16/20869538/imagenet-roulette-ai-classifier-web-tool-object-image-recognition

Source: S. Lazebnik

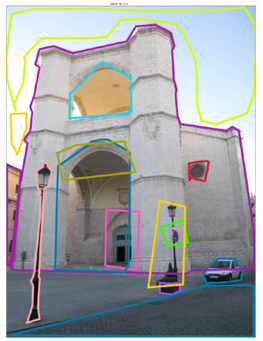
ImageNet Roulette

Source: S. Lazebnik

Image source

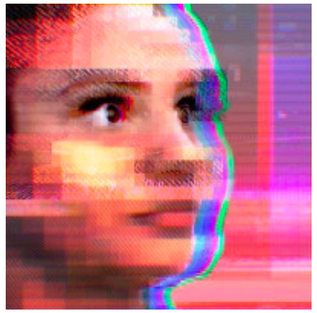
Some things to worry about…

- Our datasets are often poorly labeled



- And usually biased

- ML methods may perform well on lab-collected data, but often generalize poorly to real-world data



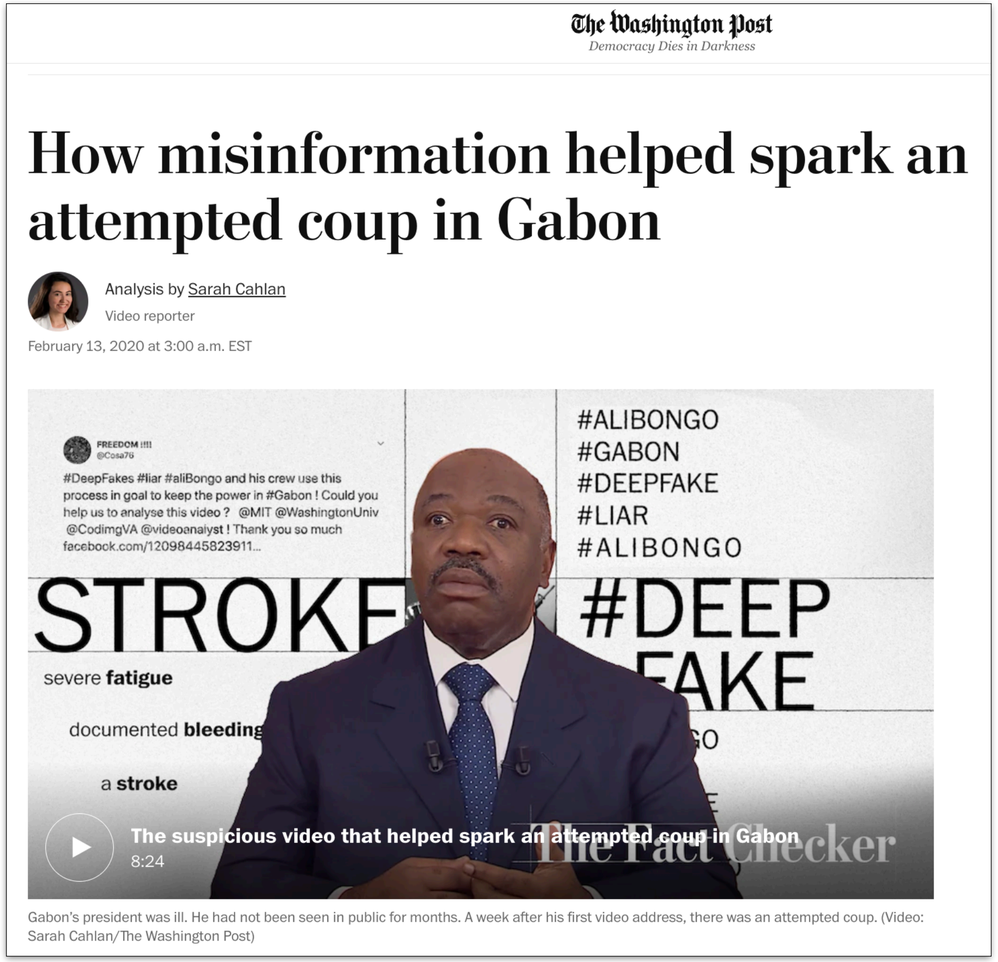
- Can have negative social consequences

Adapted from Isola, Torralba, Freeman

Open-ended discussion

- Supervised vs. unsupervised learning?
- Other negative consequences of computer vision systems?
- What other biases might computer vision systems have?

Fake images in the news



Text-to-image models make it easy

“Catholic Pope Francis wearing Balenciaga puffy jacket in drill rap music video, throwing up gang signs with hands, taken using a Canon EOS R camera with a 50mm f/1.8 lens, f/2.2 aperture, shutter speed 1/200s, ISO 100 and natural light, Full Body, Hyper Realistic Photography, Cinematic, Cinema, Hyperdetail, UHD, Color Correction, hdr, color grading, hyper realistic CG animation --ar 4:5 --upbeta --q 2 --v 5.”

But image manipulation also has a long history

Abraham Lincoln?

John C. Calhoun

But image manipulation also has a long history

From Forrest Gump, 1994

Malicious image manipulation

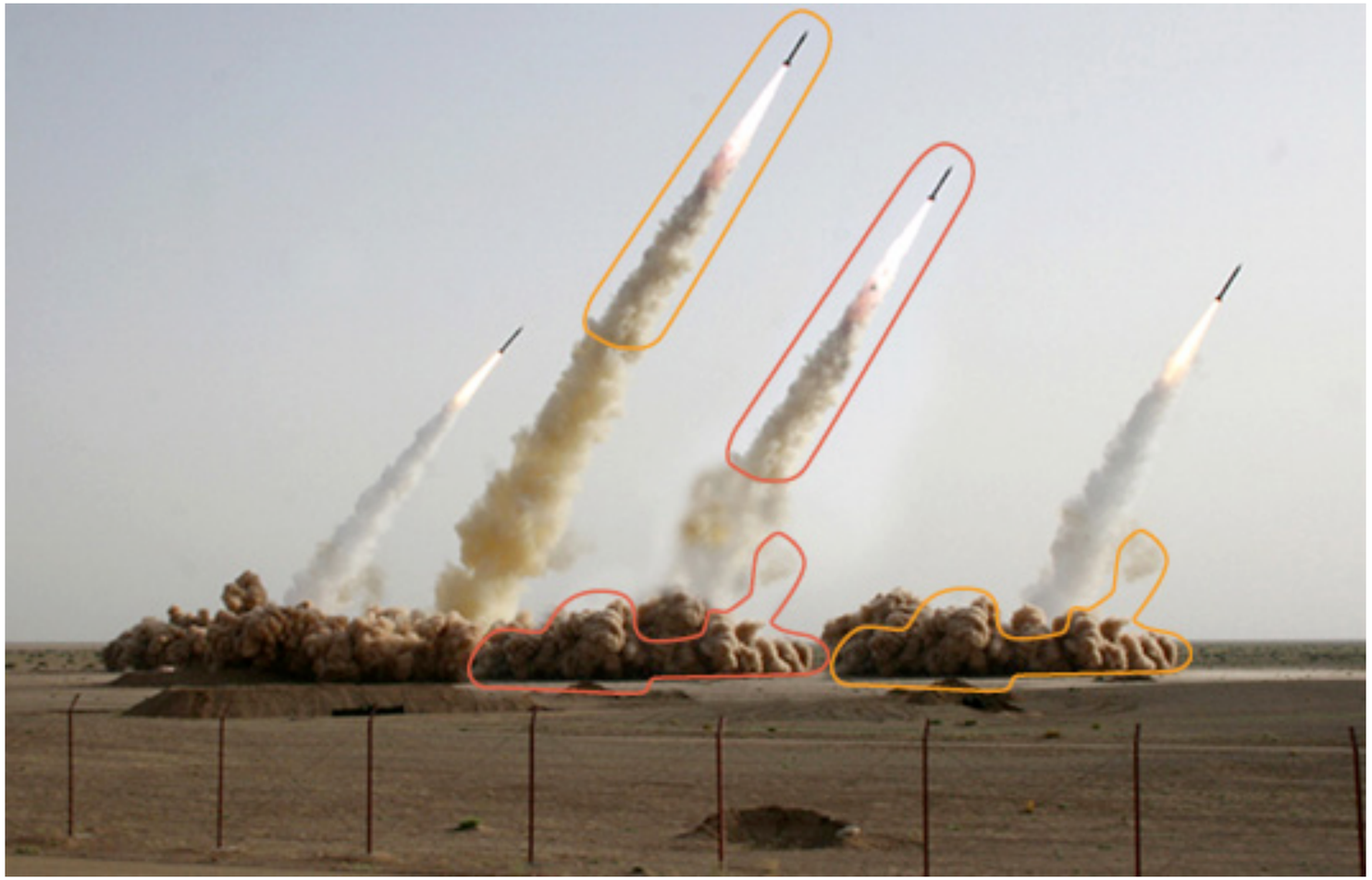
Malicious image manipulation

Fonda Speaks To Vietnam Veterans At Anti-War Rally

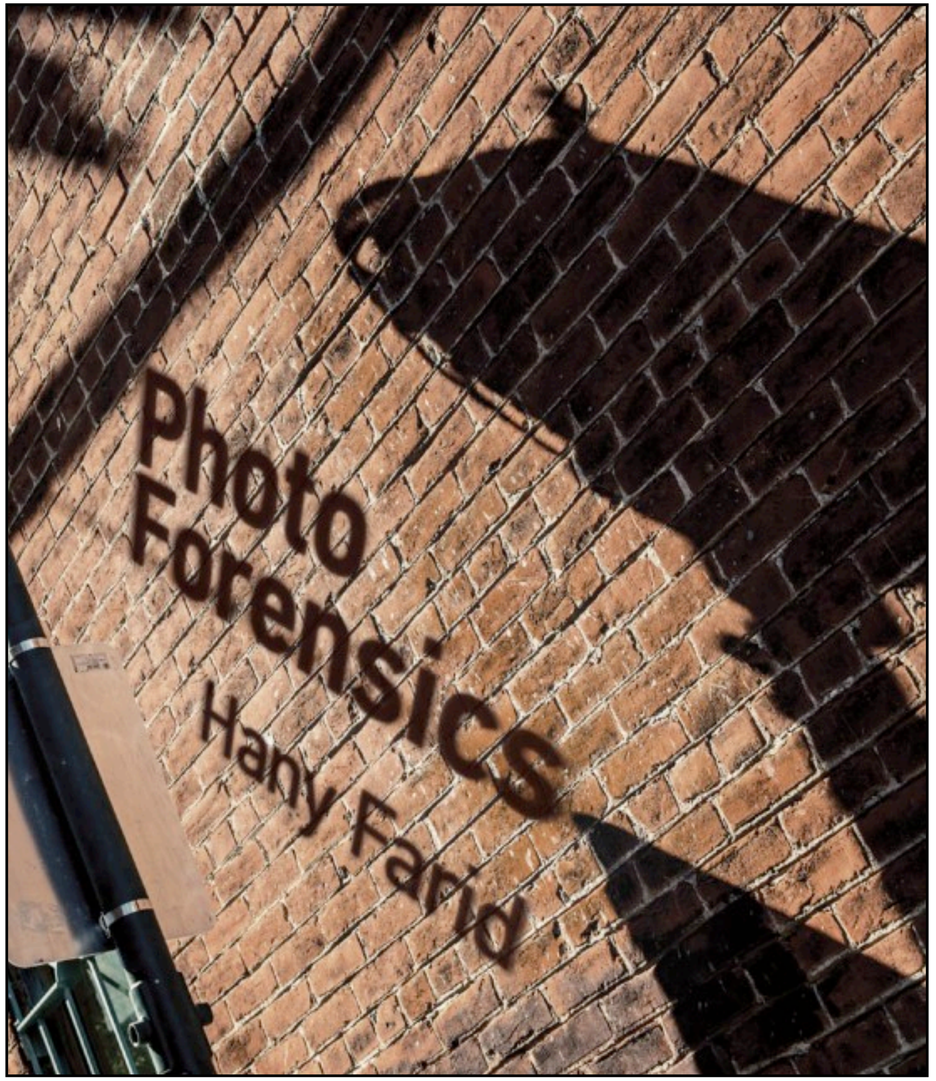
Actress And Anti-War Activist Jane Fonda Speaks to a crowd of Vietnam Veterans as Activist and former Vietnam Vet John Kerry (LEFT) listens and prepares to speak next concerning the war in Vietnam (AP Photo)

Malicious image manipulation

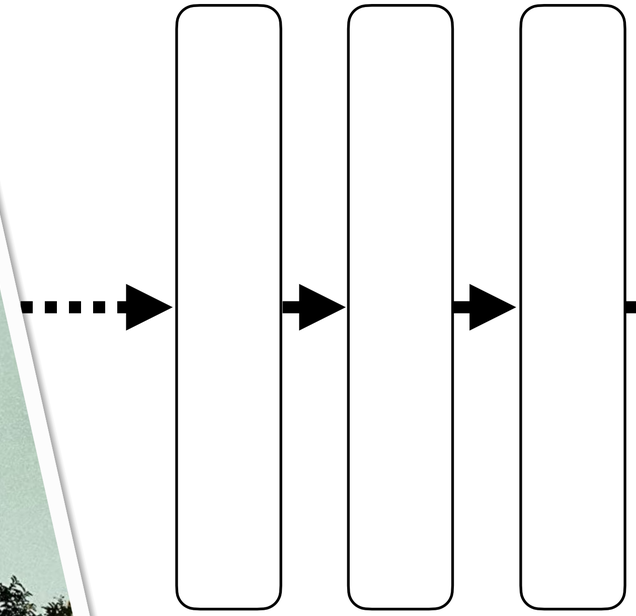
Malicious image manipulation



Detecting fake images



Hard to directly use supervised learning!



"Fake"

Strategy #1: physical models

Self-consistent lighting direction

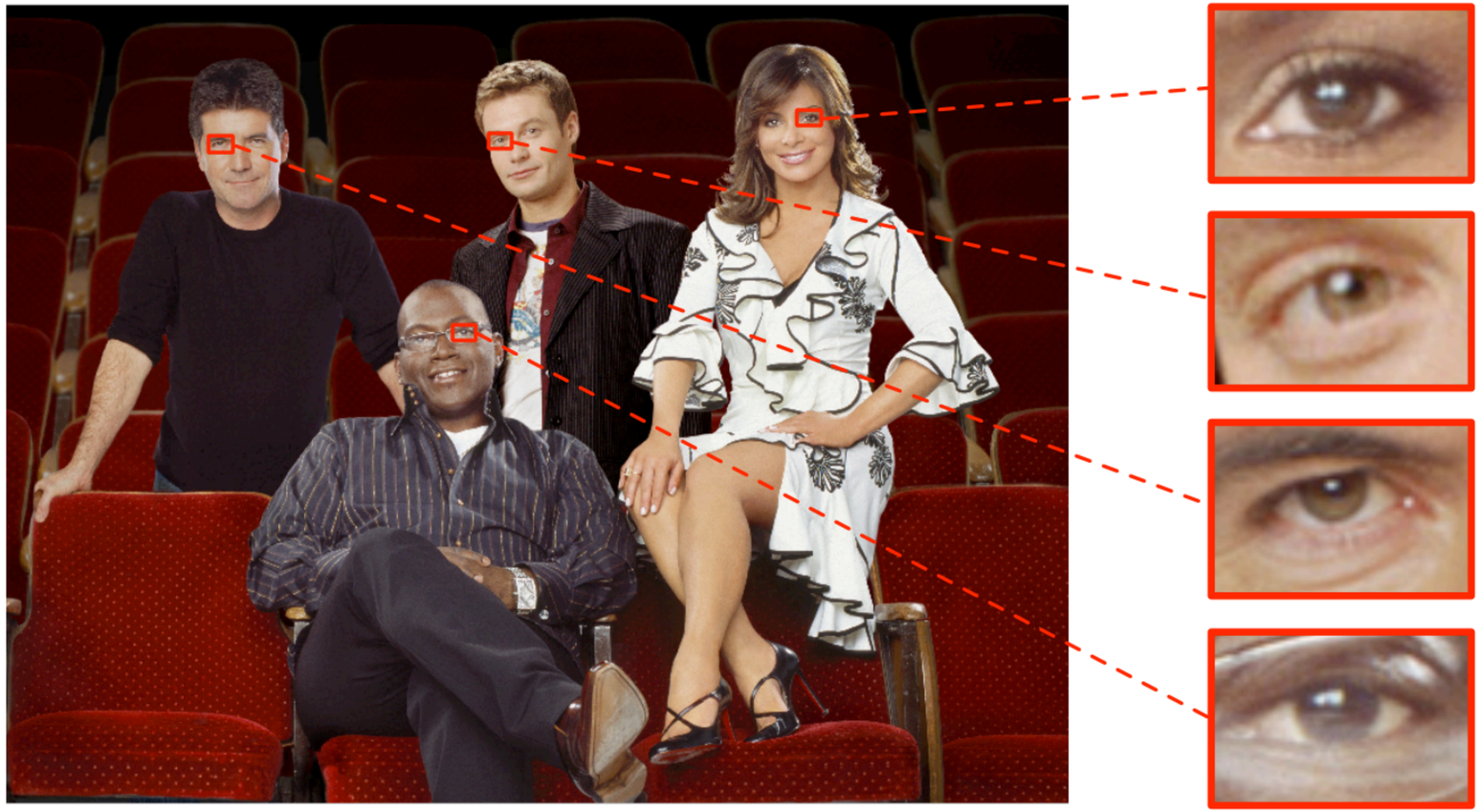
Fake photo

Real photo

[Johnson and Farid, 2005]

Source: S. Lazebnik

Specular reflections



[Johnson and Farid, 2007]

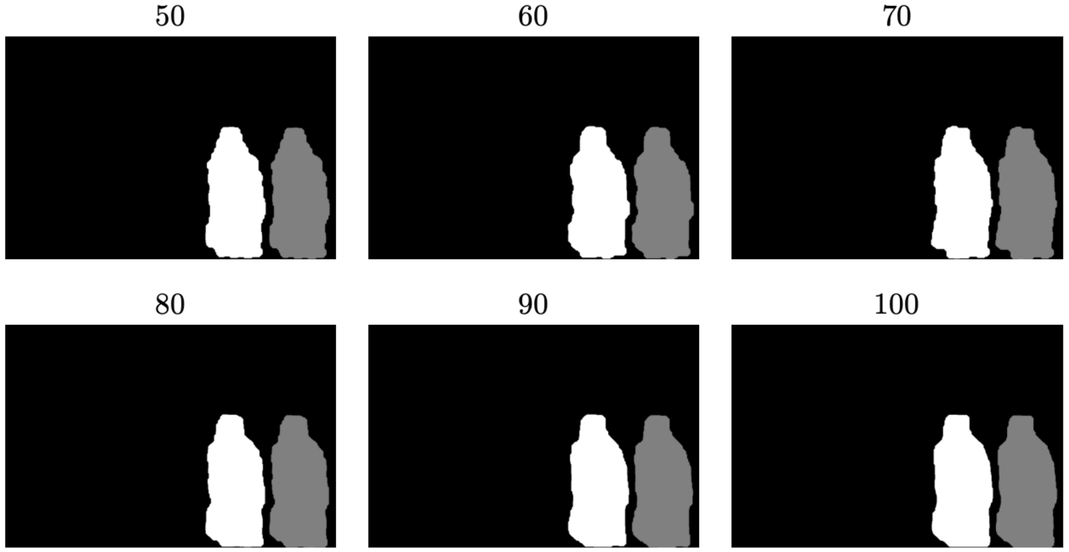
Strategy #2: low-level imaging properties

JPEG artifacts

- Cameras vary in how they do JPEG compression.
- When you quantize a floating point numbers:
  - Some do round(), others do floor() or ceil()
- If a photo seems to have *both* kinds of quantization, it's probably a fake: e.g., a composite from images taken by different cameras!

[Agarwal and Farid, "JPEG Dimples", 2017]

Detecting duplicated image regions



← amount of JPEG compression

- Traditional inpainting methods copy-and-paste image patches.
- Detect near-duplicated patches.
- But sensitive to postprocessing operations, like compression.

[Popescu and Farid, 2004]

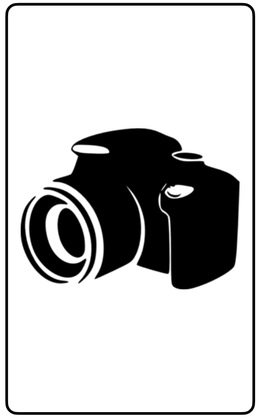
Strategy #3: learned anomaly detection

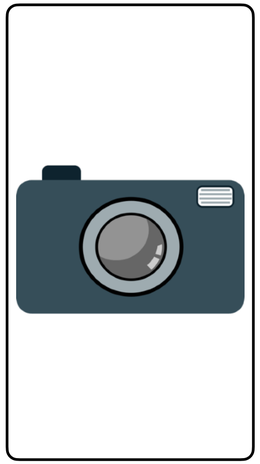
Instead of hand-crafting cues, can we learn to detect “anomalous” images, and flag suspicious images?

[Huh, Liu et al., “Image Splice Detection via Learned Self-Consistency”, 2018]


EXCAVATING
DAEWOO


EXCAVATING
DAEWOO

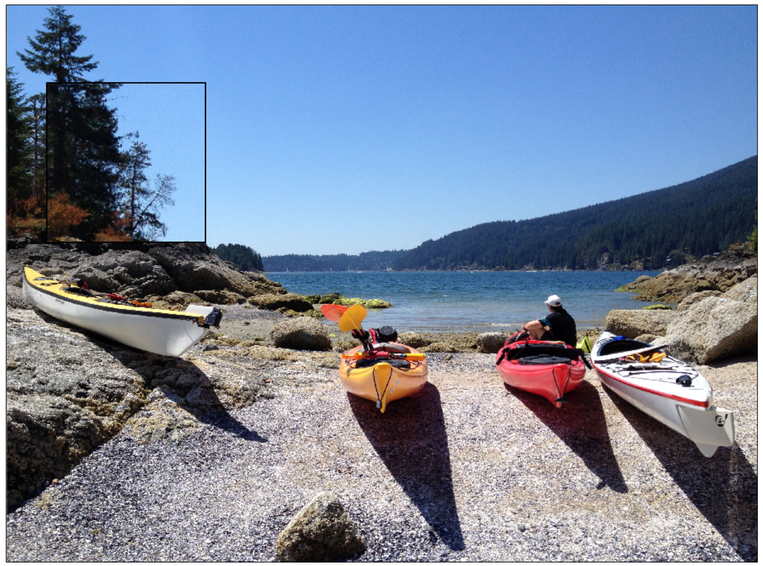




Inconsistent

Consistent

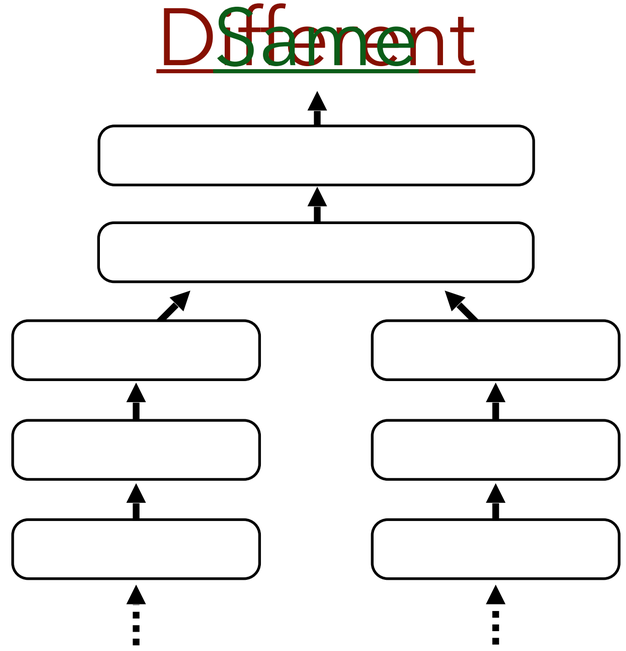
Predicting metadata consistency



```
CameraMake: Apple
CameraModel: iPhone 4s
ColorSpace: sRGB
ExifImageLength: 2448
ExifImageWidth: 3264
Flash: Flash did not fire
FocalLength: 107/2
WhiteBalance: Auto
ExposureTime: 1/2208
                …
```

```
CameraMake: NIKON CORPORATION
CameraModel: NIKON D90
ColorSpace: sRGB
ExifImageLength: 2848
ExifImageWidth: 4288
Flash: Flash did not fire
FocalLength: 18/796
WhiteBalance: Auto
ExposureTime: 1/30
                …
```

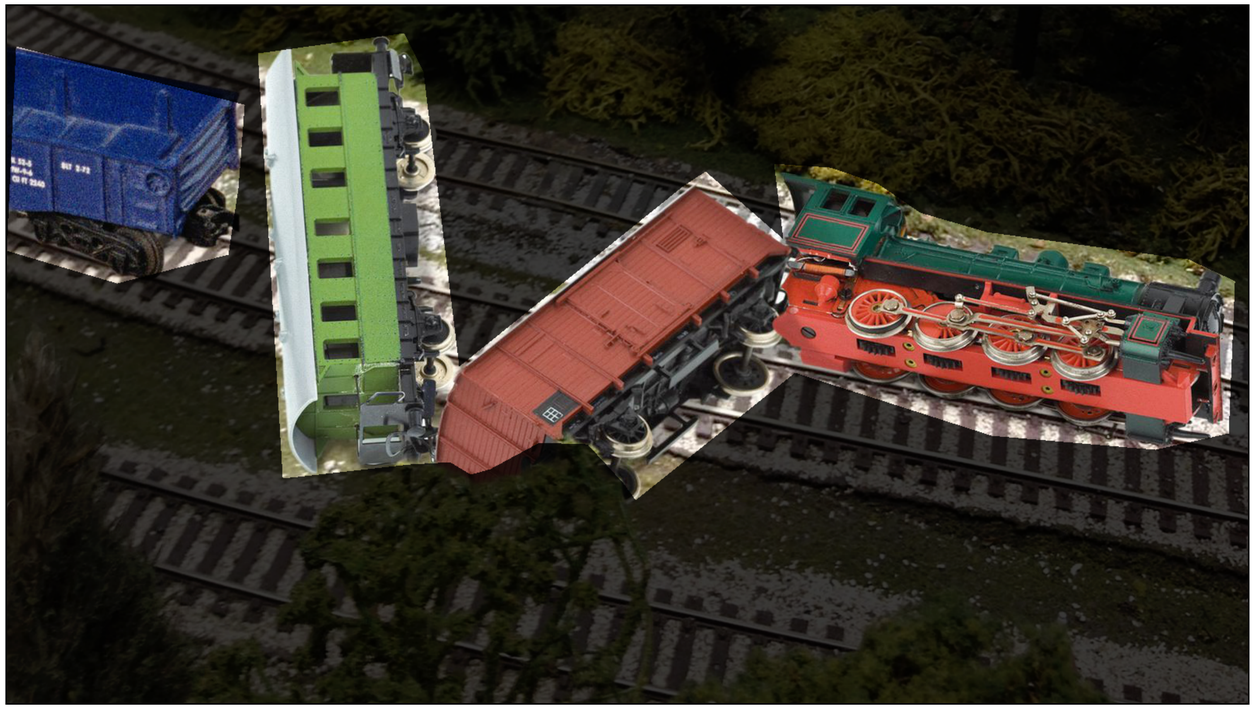
Same white balance?



Input

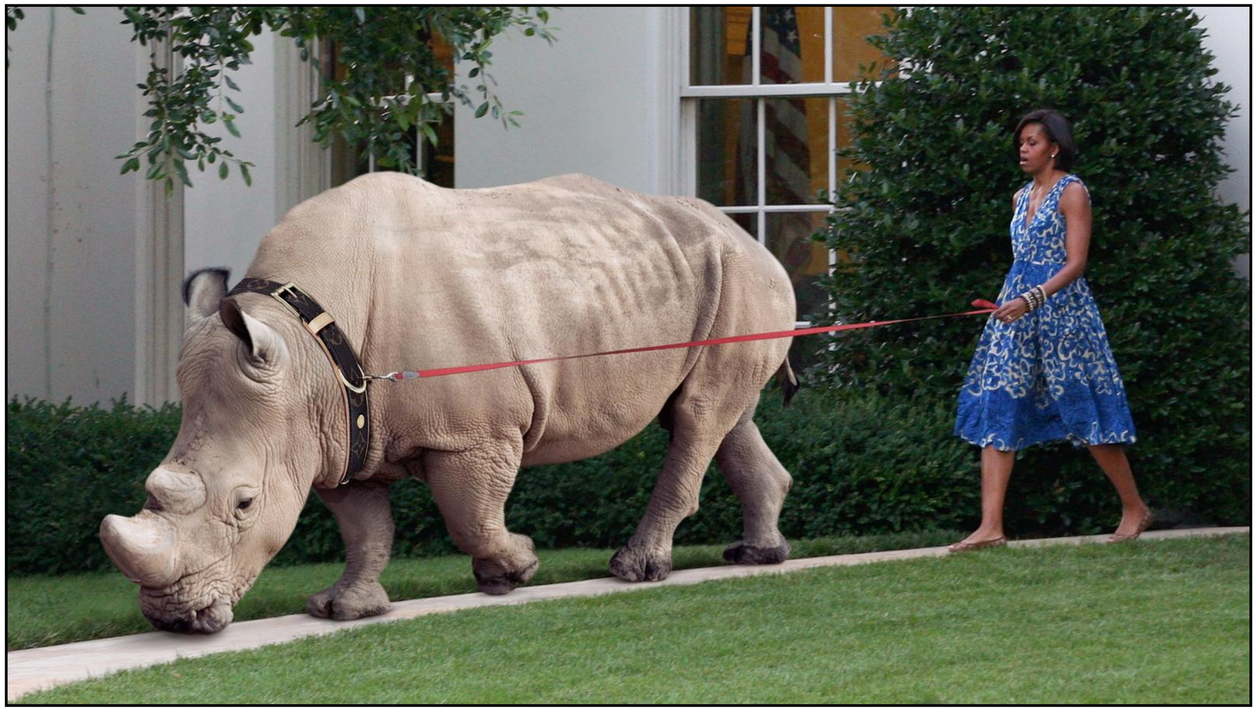
Photo source: TheOnion.com

Prediction



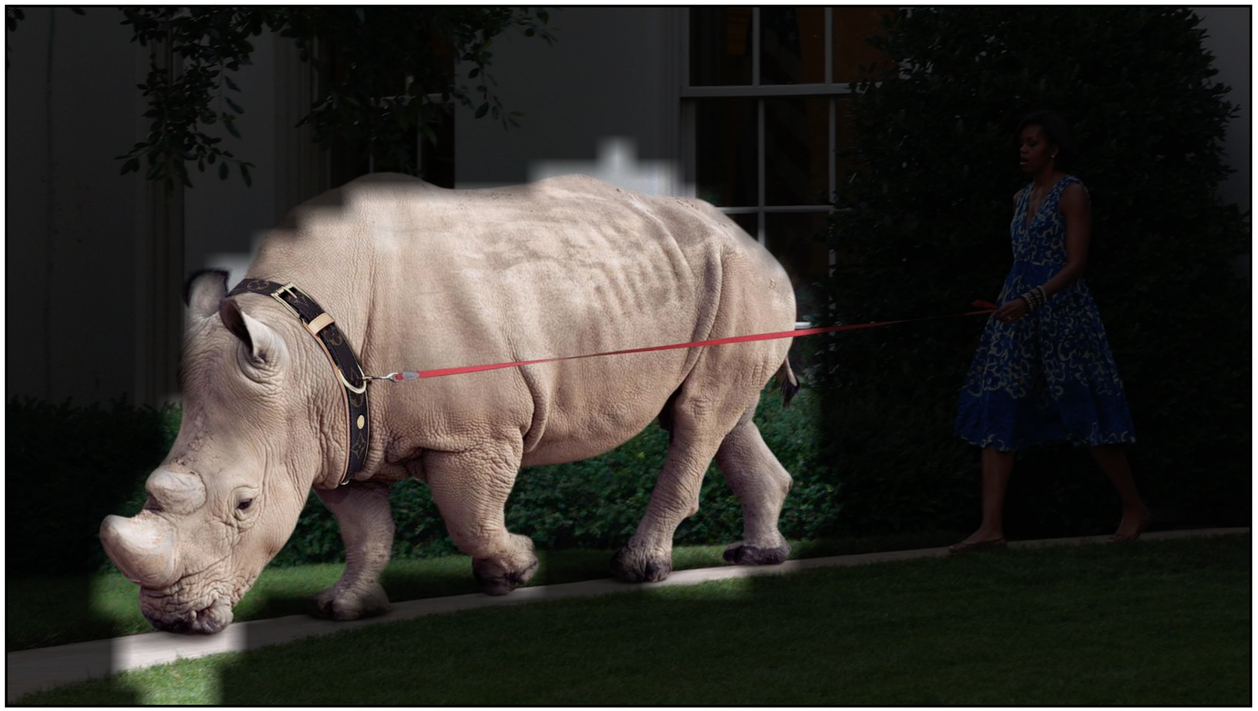
Ground truth

Photo source: TheOnion.com

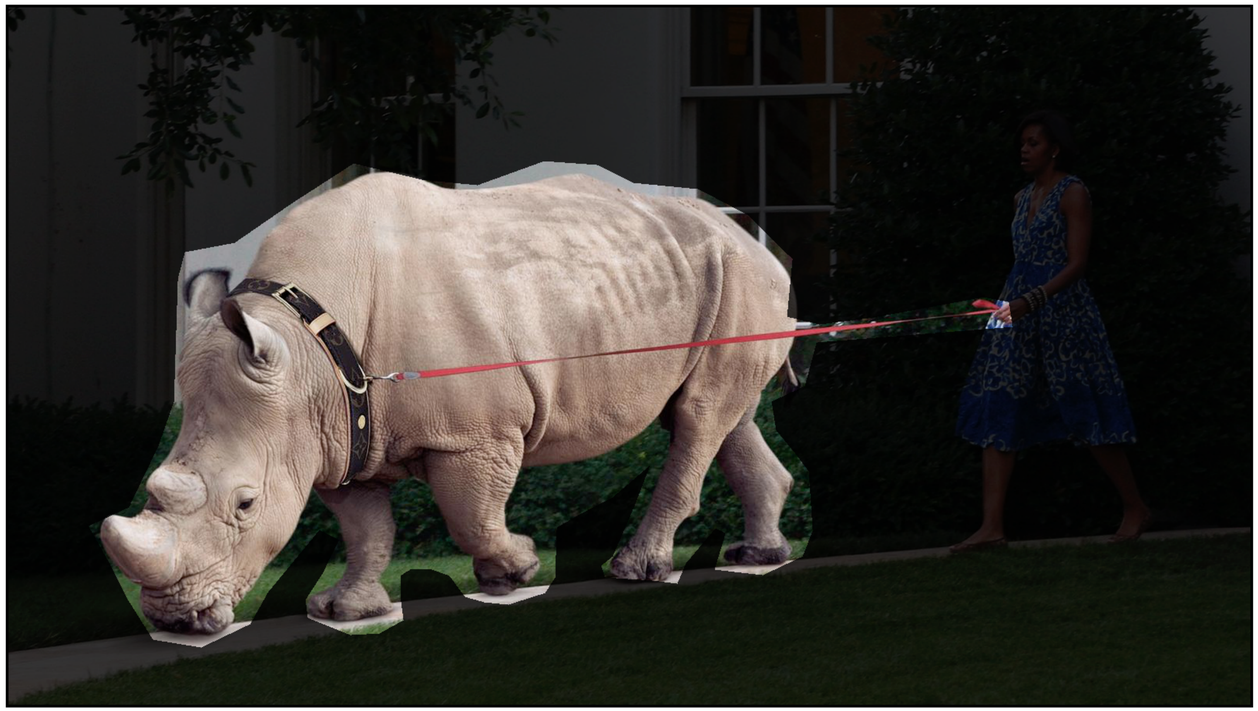


Input

Photo source: TheOnion.com



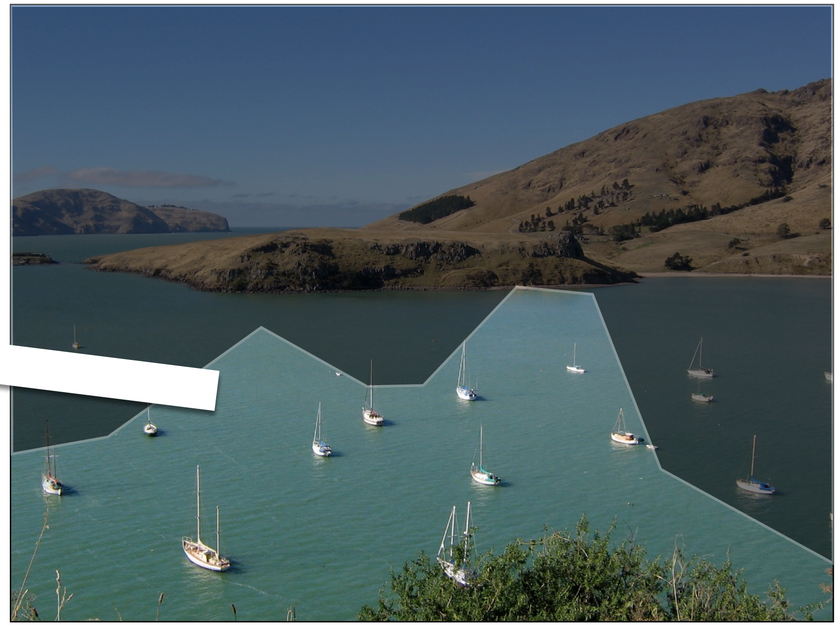
Prediction



Ground truth

Photo source: TheOnion.com

Input



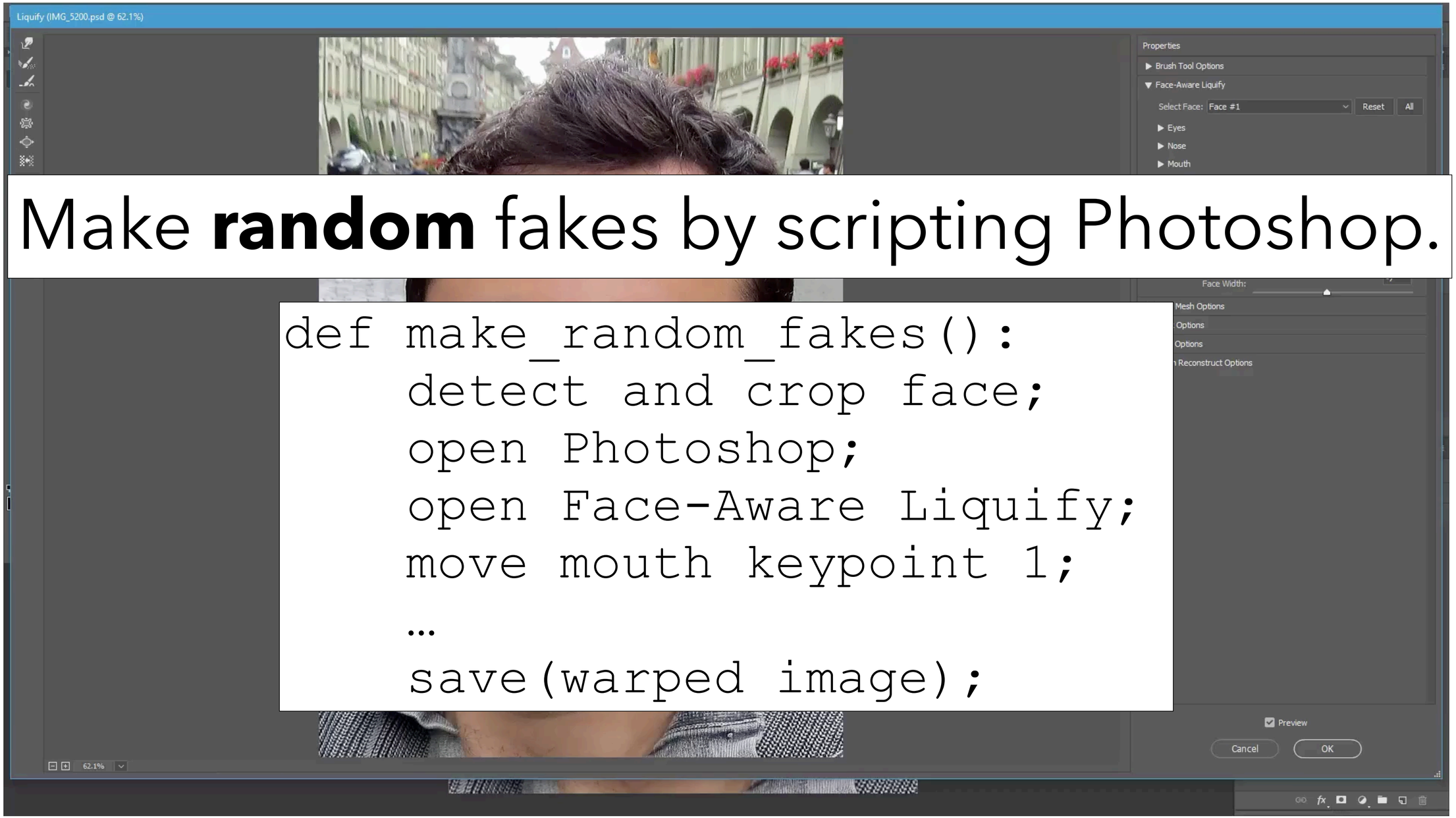
(Hays & Efros 2009)

Prediction

Ground truth

Strategy #4: supervised learning

[Wang et al., “Image Splice Detection via Learned Self-Consistency”, 2018]



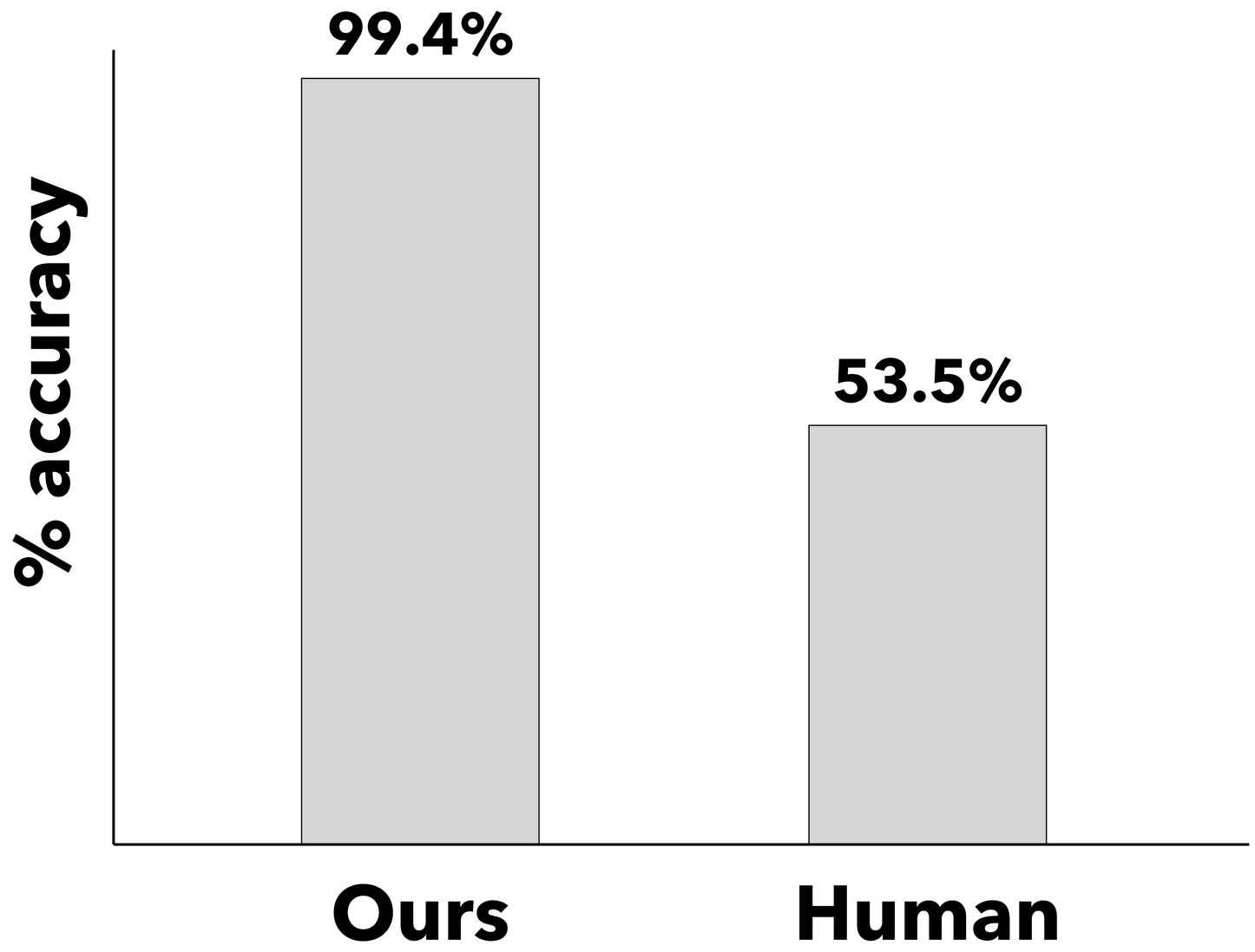
Make **random** fakes by scripting Photoshop.

```
def make_random_fakes():
    detect and crop face;
    open Photoshop;
    open Face-Aware Liquify;
    move mouth keypoint 1;
    …
    save(warped image);
```

Photoshop Face-Aware Liquify tutorial. Source: https://youtu.be/5Qqv_C6iVvQ?t=86

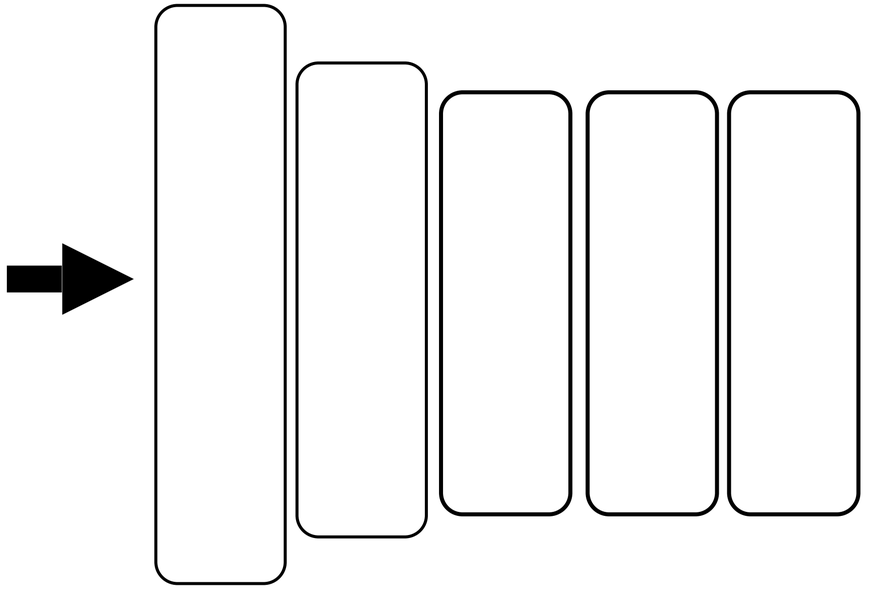
Warp
detector

Real-or-fake classification

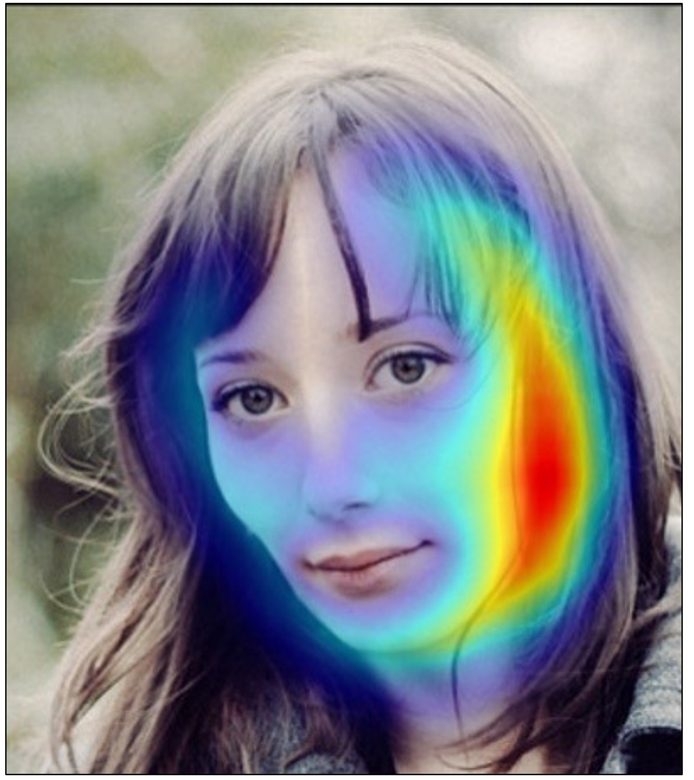


What moved where?

Manipulated Image

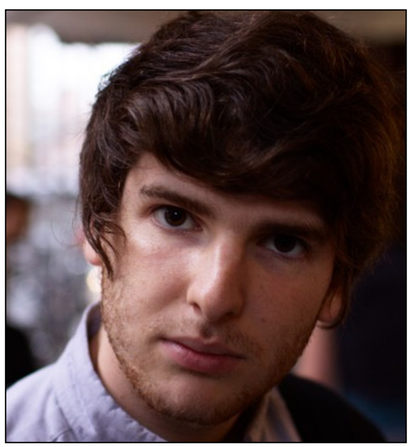
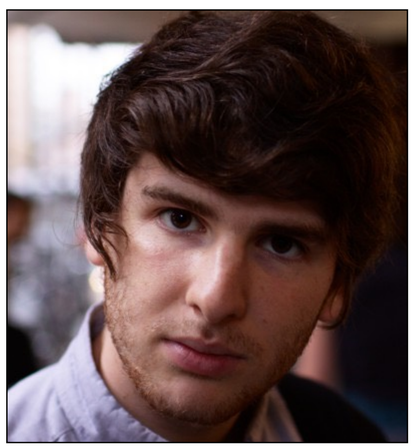


Dilated ResNet



Warp Prediction

What moved where?

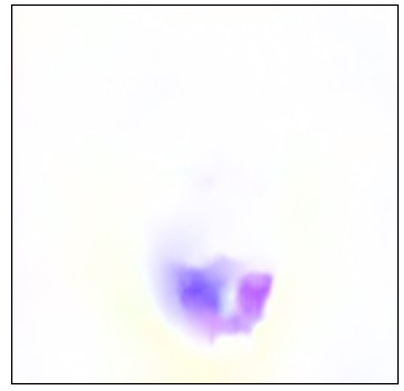





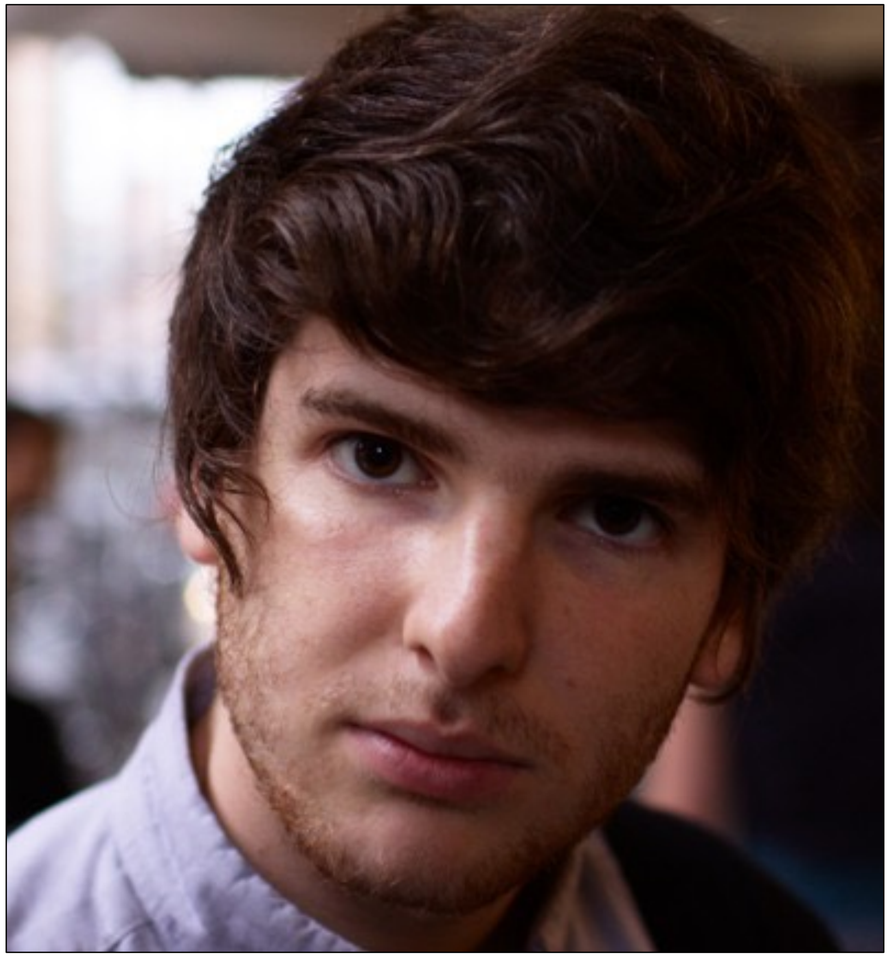
Modified Original Optical Flow





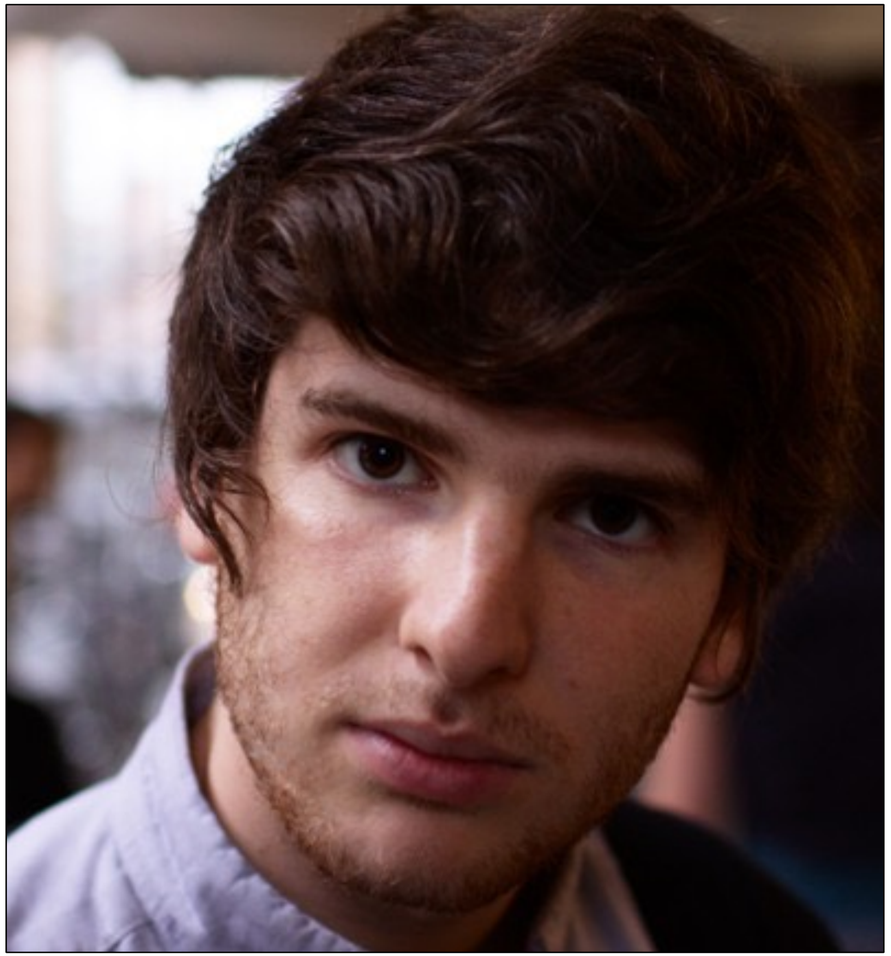



Modified Original Optical Flow

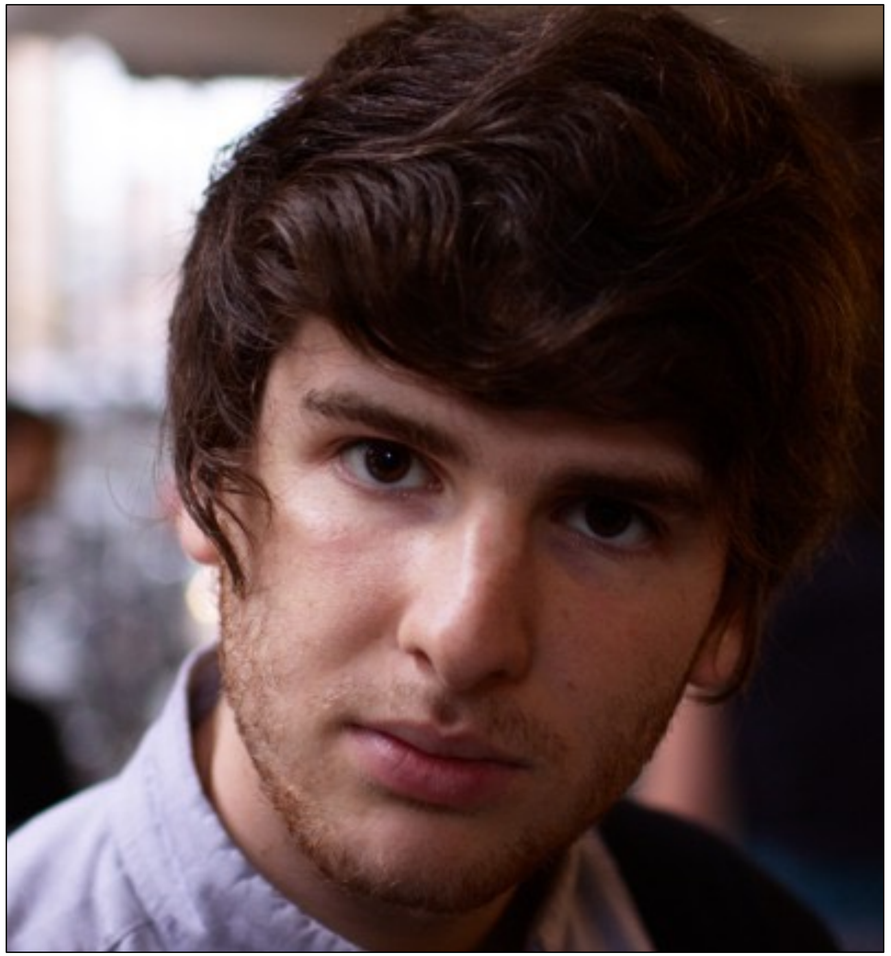


Manipulated Photo

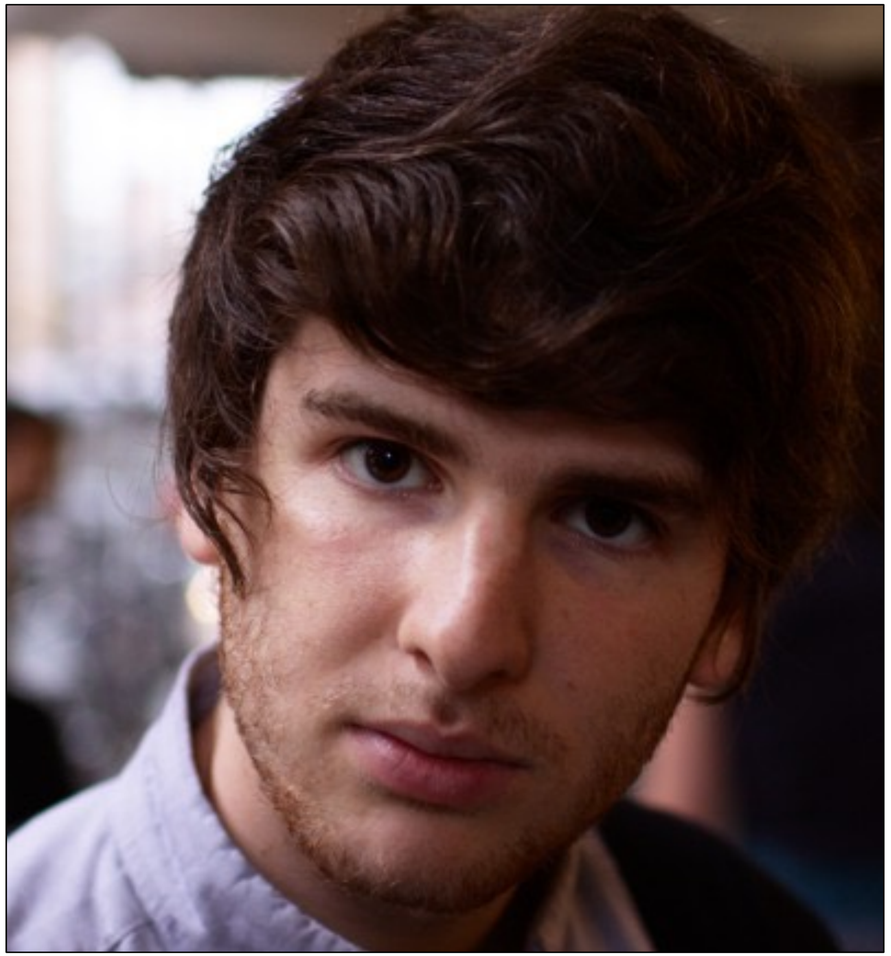
Flow Prediction



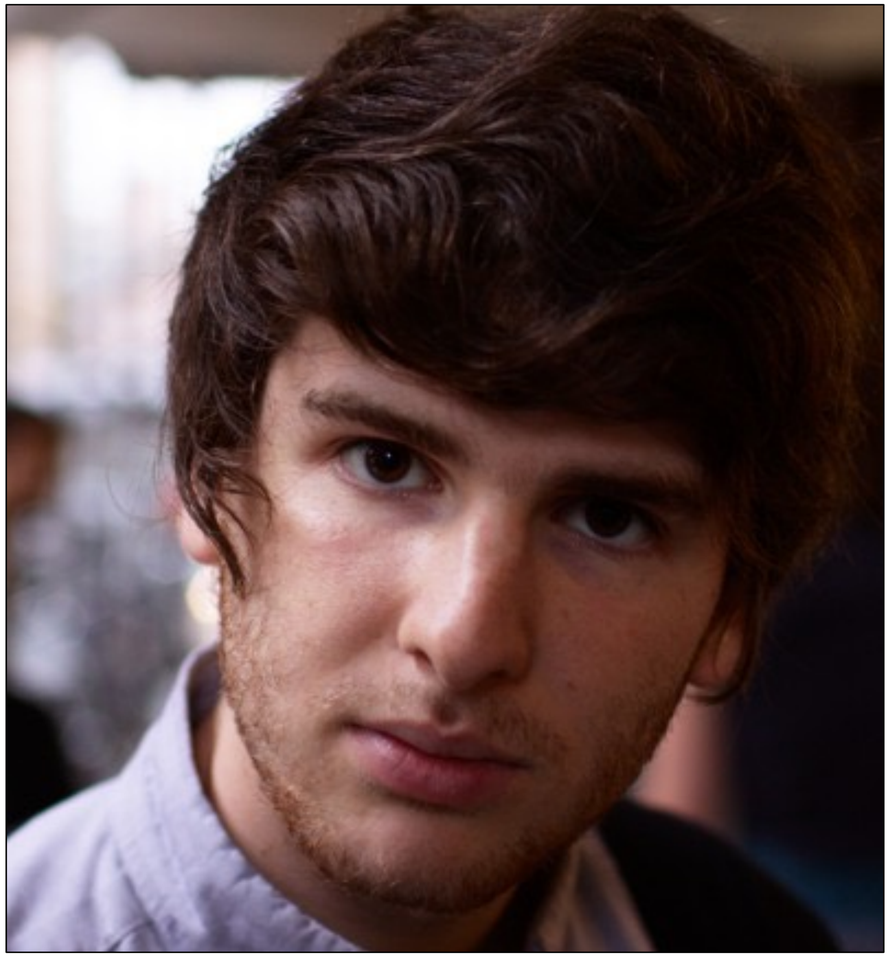
Suggested “Undo”



Original Photo



Manipulated vs. Original



Undo vs. Original

Manipulated Photo

Warp Prediction

Suggested “Undo”

Original Photo

Suggested “Undo”

Manipulated Photo

AI-generated images

“Catholic Pope Francis wearing Balenciaga puffy jacket in drill rap music video, throwing up gang signs with hands, taken using a Canon EOS R camera with a 50mm f/1.8 lens, f/2.2 aperture, shutter speed 1/200s, ISO 100 and natural light, Full Body, Hyper Realistic Photography, Cinematic, Cinema, Hyperdetail, UHD, Color Correction, hdr, color grading, hyper realistic CG animation --ar 4:5 --upbeta --q 2 --v 5.”

AI-generated spam

Jeongsoo Park

Examples from Jeongsoo Park's Facebook feed

Real or fake?

Detector	Result
NoDown (2021)	0.1% AI-Generated Likelihood
HIFI (2023)	36.9% AI-Generated Likelihood
CLIP-ViT (2023)	0.0% AI-Generated Likelihood
GLFF (2024)	100.0% AI-Generated Likelihood
NPR (2024)	10.6% AI-Generated Likelihood
AIDE (2024)	9.5% AI-Generated Likelihood
Effort (2025)	57.2% AI-Generated Likelihood

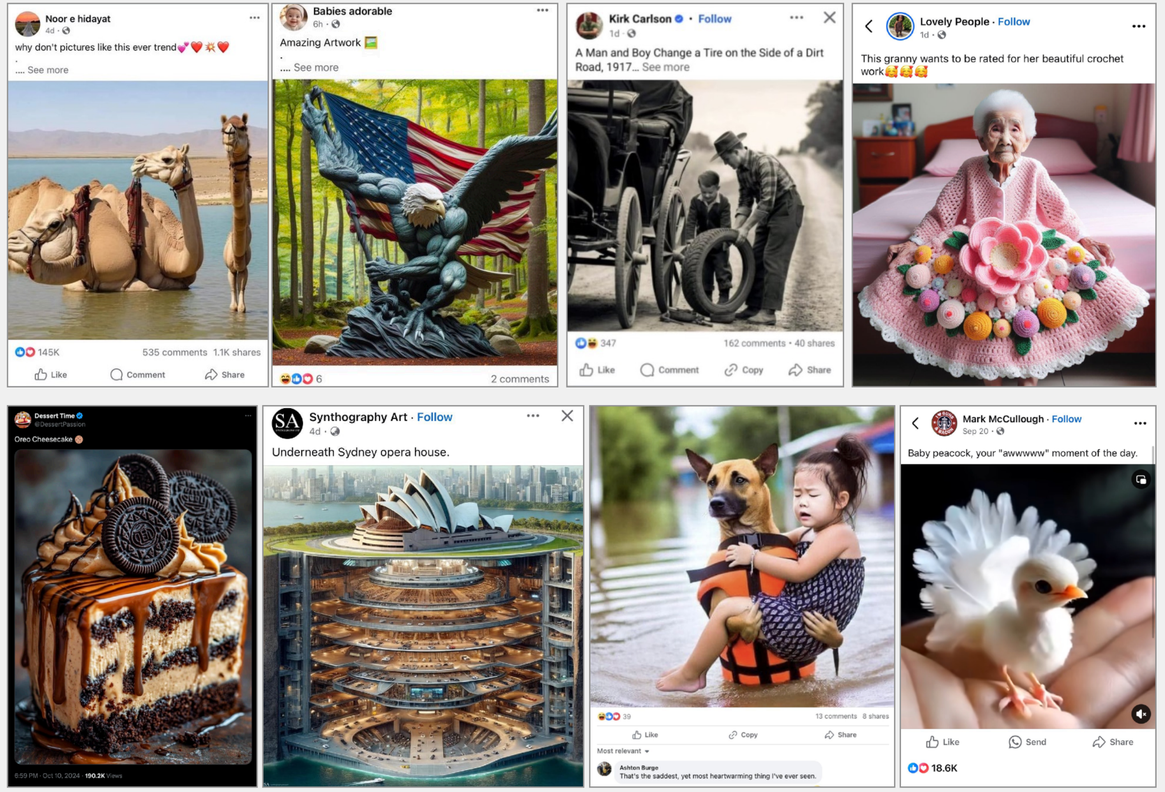
[Li, Zhang, Sun, Qi, Lyu, “DeepFake-o-meter”, 2024]

Downsides of easy image generation

Misinformation and abuse

Raising doubt in real images

AI spam



Source: "Insane Facebook AI slop" @FacebookAIslop

Generalization in AI-generated image detection

Training

ProGAN

DALL·E 2

VQ-GAN

Midjourney v5

Generalization in AI-generated image detection

Training

ProGAN

DALL·E 2

VQ-GAN

Midjourney v5

Test on the *same* generators?
→ Detectors generalize pretty well.

Test on out-of-distribution generators?
→ Not always. But intriguingly there is *some* generalization!

[S. Wang, O. Wang, R. Zhang, A. Owens, A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now", CVPR 2020]

Generalization in AI-generated image detection

Training

ProGAN

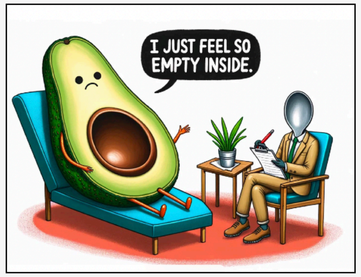
DALL·E 2

VQ-GAN

Midjourney v5

Testing

FLUX.1



DALL·E 3

Imagen v3

Ideogram

github.com/anon123/mycoolgenerator

github.com/bot15/deepfakeface

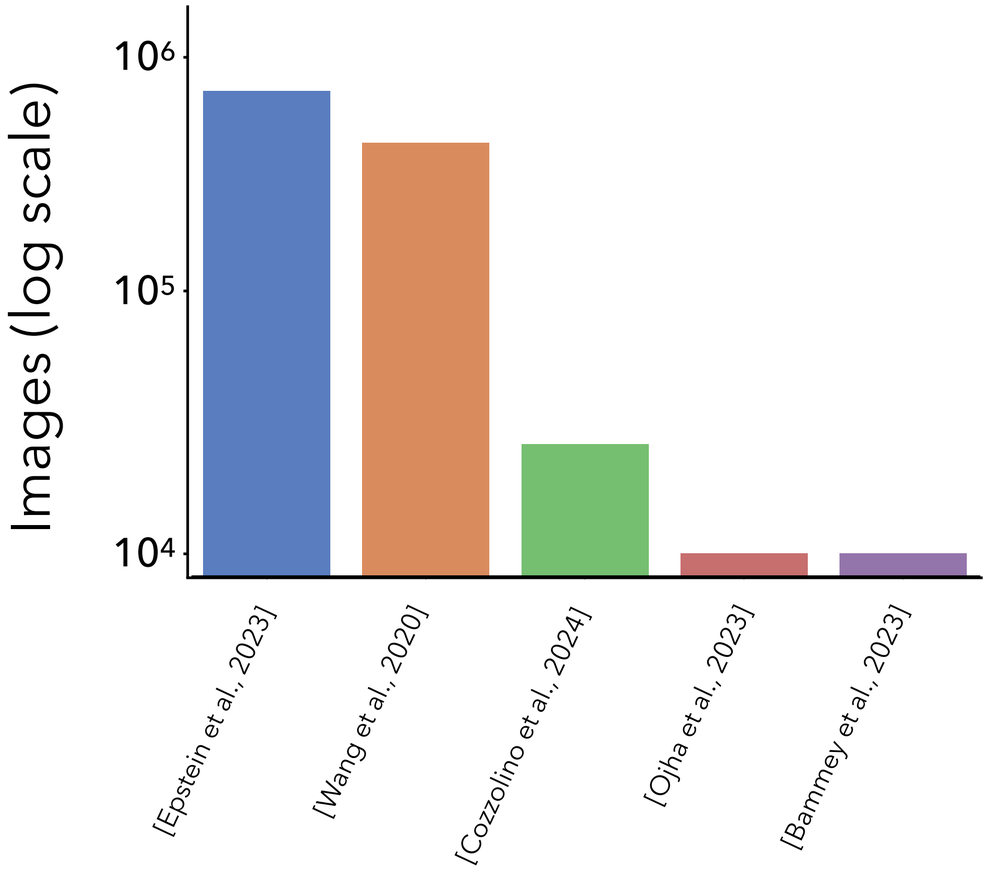
github.com/putin2/electiondeepfake

…

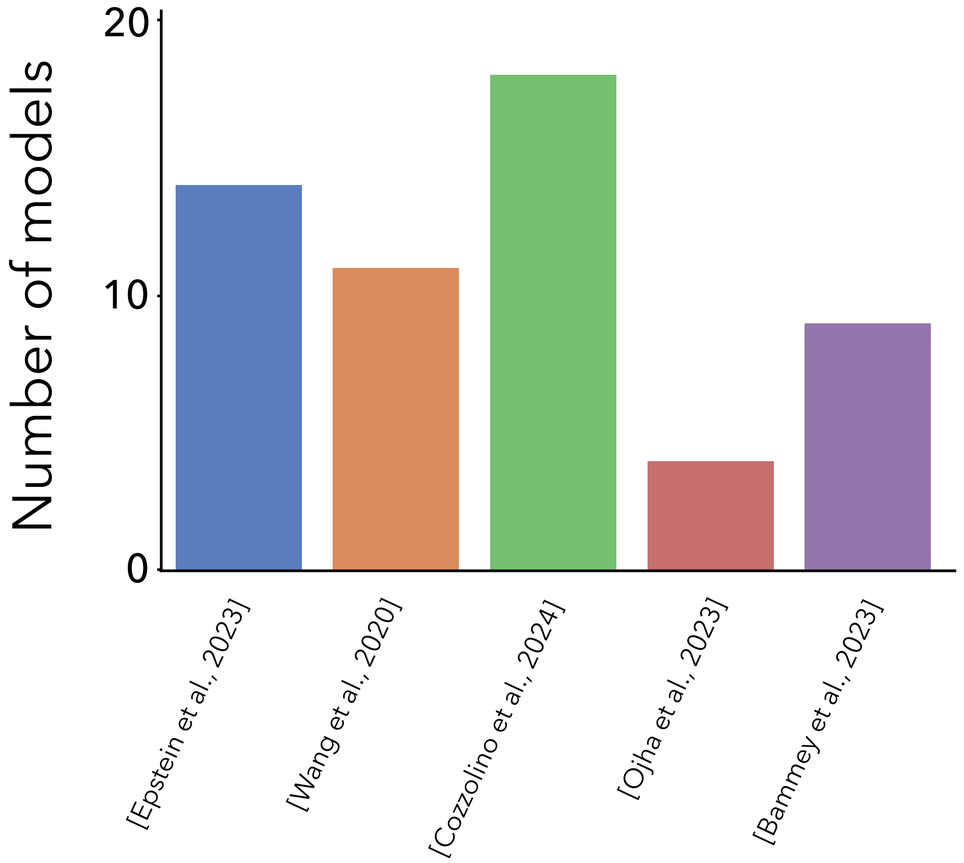
"DIY" models

One problem: the data

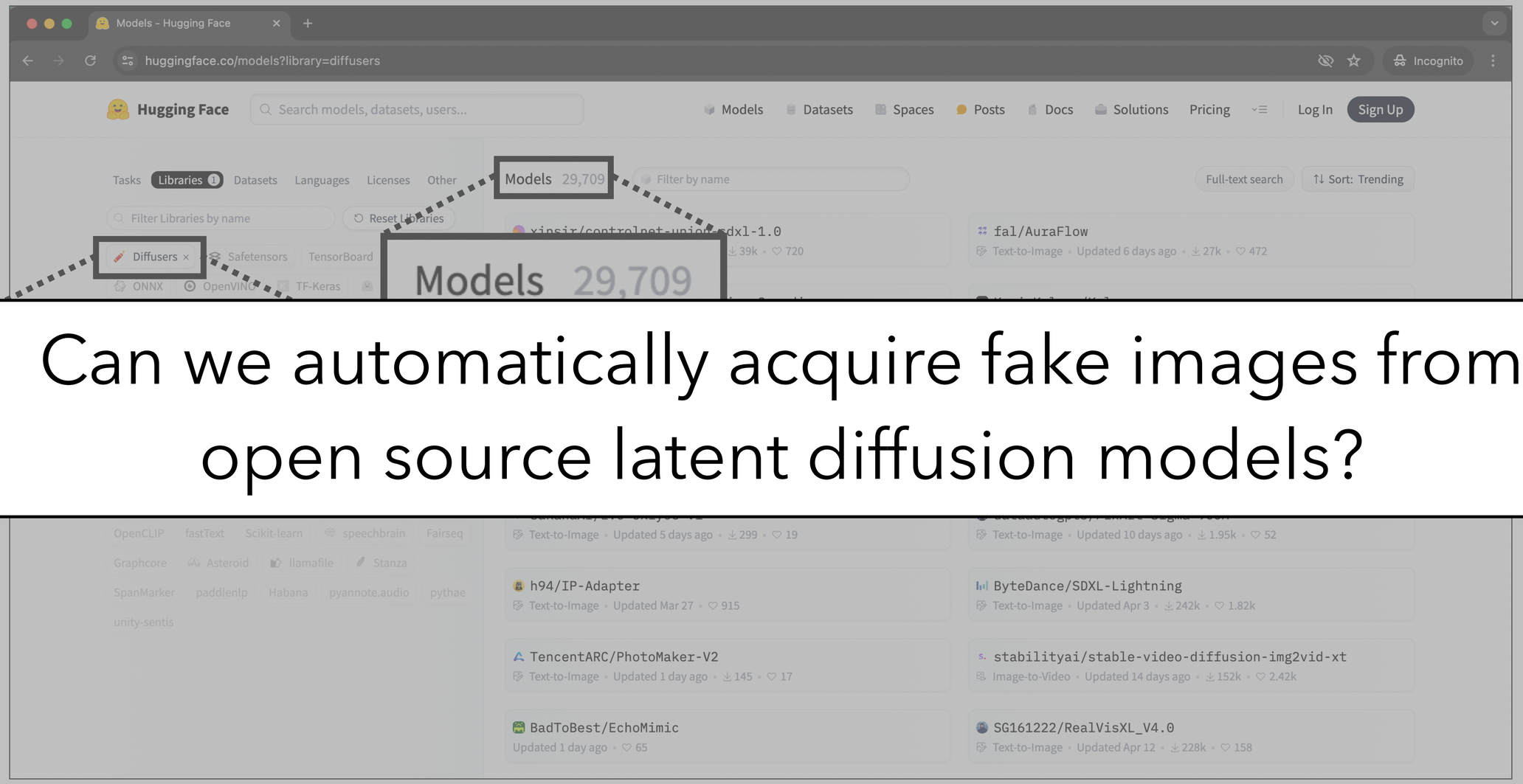
Datasets have lots of images.



But relatively few *models*.



Model sharing communities



Can we automatically acquire fake images from open source latent diffusion models?

[Park & Owens. “Community Forensics: Using Thousands of Generators to Train Fake Image Detectors “]… out soon!

Jeongsoo Park

What are these models?

playgroundai / **playground-v2.5-1024px-aesthetic** like 695 Follow Playground 187

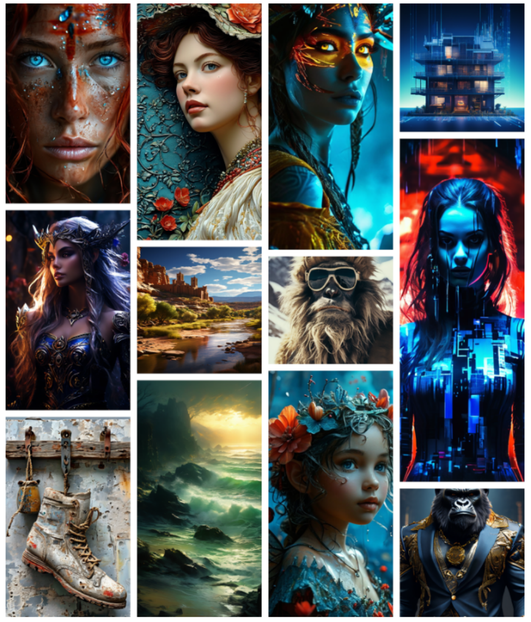
Text-to-Image | Diffusers | Safetensors | StableDiffusionXLPipeline | playground | Inference Endpoints | arxiv:2206.00364 | arxiv:2402.17245 | License: playground-v2dot5-community

Model card | Files | Community 32 | Train | Deploy | Use this model

Edit model card

Playground v2.5 – 1024px Aesthetic Model

This repository contains a model that generates highly aesthetic images of resolution 1024x1024, as well as portrait and landscape aspect ratios. You can use the model with Hugging Face 🧨 Diffusers.



Downloads last month
100,506

Inference API Cold

Text-to-Image | Examples

Your sentence here... | Compute

View Code | Maximize

Model tree for playgroundai/playground-v2.5-1…

Adapters 10 models

Spaces using playgroundai/playground-v2.5… 67

- playgroundai/playground-v2.5
- TIGER-Lab/GenAI-Arena
- Nymbo/image_gen_supaqueue
- ksort/K-Sort-Arena
- ChenoAi/Playground-v1.0
- mukaist/image
- Azure99/blossom-playground-v2.5
- Azure99/playground-v2.5
- allknowingroger/Image-Models-Test348
- poulpy5/Kroqbot2
- cavargas10/Text-Img-UTPL
- AZLABS/Tester

+ 55 Spaces

prompthero / **openjourney-v4** like 1.23k

Text-to-Image | Diffusers | Safetensors | StableDiffusionPipeline | stable-diffusion | Inference Endpoints | License: creativeml-openrail-m

Model card | Files | Community 64 | Train | Deploy | Use this model

Edit model card

Openjourney v4

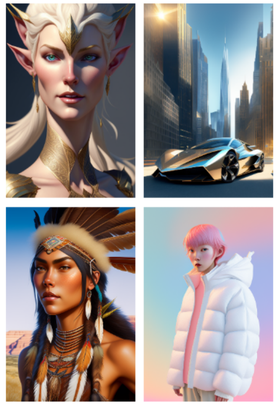
Trained on +124k Midjourney v4 images, by PromptHero

Trained on Stable Diffusion v1.5 using +124000 images, 12400 steps, 4 epochs +32 training hours.

💡 Openjourney-v4 prompts

Pss... "mdjrny-v4 style" is not necessary anymore (yay!)

🎓 **Want to learn how to train Openjourney?** 👉 Join our course 🔥



Downloads last month
44,112

Inference Examples

Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Spaces using prompthero/openjourney-v4 100

- Yntec/ToyWorld
- Yntec/PrintingPress
- Nymbo/image_gen_supaqueue
- phenixrhyder/NSFW-ToyWorld
- Yntec/blitz_diffusion
- Manjushri/OJ-V4-CPU
- John6666/Diffusion80XX4sg
- PeepDaSlan9/B2BMGMT_Diffusion60XX
- phenixrhyder/PrintingPress
- Daniela-C/6x_Image_diffusion
- John6666/PrintingPress4
- John6666/hfd_test_nostopbutton

+ 88 Spaces

What are these models?

MirageML / **fantasy-sword** like 13 Follow Mirage 27

Text-to-Image | Diffusers | Safetensors | English | StableDiffusionPipeline | stable-diffusion | Inference Endpoints

License: creativeml-openrail-m

Model card | Files | Community 4 | Train | Deploy | Use this model

Edit model card

Fantasy Sword on Stable Diffusion via Dreambooth

This the Stable Diffusion model fine-tuned the Fantasy Sword concept taught to Stable Diffusion with Dreambooth. It can be used by modifying the `instance_prompt`: **a photo of fantasy_sword**

Run on Mirage

Run this model and explore text-to-3D on Mirage!

Here are is a sample output for this model:

Downloads last month
15

Inference Examples

Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Space using `MirageML/fantasy-sword` 1

MirageML/fantasy-sword

monadical-labs / **minecraft-skin-generator-sdxl** like 10 Follow Monadical Labs 15

Text-to-Image | Diffusers | Safetensors | English | StableDiffusionXLPipeline | minecraft | skins | gaming | stable diffusion | stable diffusion xl | Inference Endpoints | License: openrail

Model card | Files | Community | Train | Deploy | Use this model

Edit model card

Minecraft Skin Generator XL

Monadical is pleased to announce the official release of the Minecraft Skin Generator XL model. We had previously released the Minecraft Skin Generator model based upon Stable Diffusion 2. This new model offers significant improvements over the last generation of models.

Key Features

1. **Upgrade to Stable Diffusion XL** - Our model is now based upon the Stable Diffusion XL model, which greatly improves the quality of generated skins when compared to previous models.
2. **Transparent Layer Support** - The new model now supports the transparency layer in the hair and helmet section of the skin.

Examples

- 'Kelly Kapoor from the TV show "The Office"'

Downloads last month
16,584

Inference Examples

Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Spaces using `monadical-labs/minecraft-ski…` 11

Nick088/Stable_Diffusion_Finetuned_Minecraft_S…

Nymbo/image_gen_supaqueue

soiz/Stable_Diffusion_Finetuned_Minecraft_Skin_…

Loxera/monadical-labs-minecraft-skin-generator-sdxl

boisterous/minecraft-skin-generator

gege002/monadical-labs-minecraft-skin-generator-sdxl

dadadaadadadad/monadical-labs-minecraft-skin-gene…

K00B404/image_gen_supaqueue_game_assets

ExynoxAI/MineSkin_Magic

issef/test

ThatBlondeGuy/Stable_Diffusion_Finetuned_Mine…

What are these models?

MirageML / **fantasy-sword** like 13 Follow Mirage 27

Text-to-Image | Diffusers | Safetensors | English | StableDiffusionPipeline | stable-diffusion | Inference Endpoints

License: creativeml-openrail-m

Model card | Files | Community 4 | Train | Deploy | Use this model

Edit model card

Fantasy Sword on Stable Diffusion via Dreambooth

This the Stable Diffusion model fine-tuned the Fantasy Sword concept taught to Stable Diffusion with Dreambooth. It can be used by modifying the `instance_prompt`: **a photo of fantasy_sword**

Run on Mirage

Run this model and explore text-to-3D on Mirage!

Here are is a sample output for this model:

Downloads last month
15

Inference Examples

Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Space using `MirageML/fantasy-sword` 1

MirageML/fantasy-sword

monadical-labs / **minecraft-skin-generator-sdxl** like 10 Follow Monadical Labs 15

Text-to-Image | Diffusers | Safetensors | English | StableDiffusionXLPipeline | minecraft | skins | gaming | stable diffusion | stable diffusion xl | Inference Endpoints | License: openrail

Model card | Files | Community | Train | Deploy | Use this model

Edit model card

Minecraft Skin Generator XL

Monadical is pleased to announce the official release of the Minecraft Skin Generator XL model. We had previously released the Minecraft Skin Generator model based upon Stable Diffusion 2. This new model offers significant improvements over the last generation of models.

Key Features

1. **Upgrade to Stable Diffusion XL** - Our model is now based upon the Stable Diffusion XL model, which greatly improves the quality of generated skins when compared to previous models.
2. **Transparent Layer Support** - The new model now supports the transparency layer in the hair and helmet section of the skin.

Examples

- 'Kelly Kapoor from the TV show "The Office"'

Downloads last month
16,584

Inference Examples

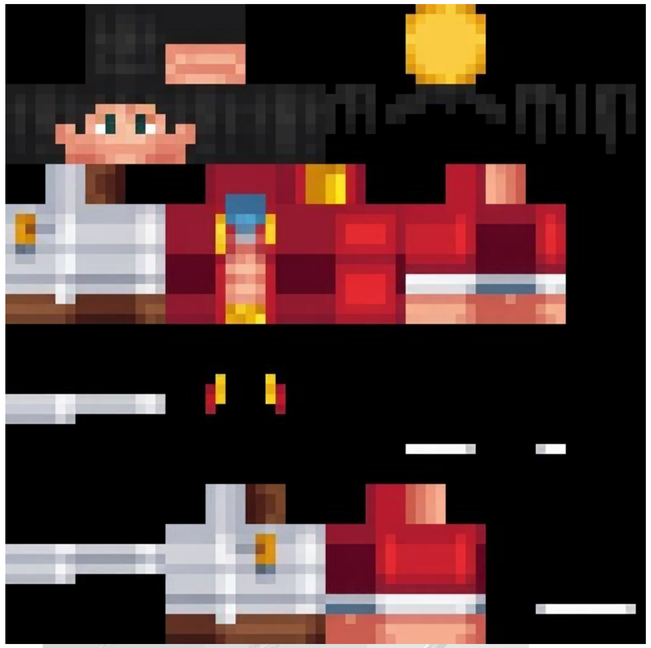
Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Spaces using `monadical-labs/minecraft-ski…` 11

Nick088/Stable_Diffusion_Finetuned_Minecraft_S…
Nymbo/image_gen_supaqueue
soiz/Stable_Diffusion_Finetuned_Minecraft_Skin_…
Loxera/monadical-labs-minecraft-skin-generator-sdxl
boisterous/minecraft-skin-generator
gege002/monadical-labs-minecraft-skin-generator-sdxl
dadadaadadadad/monadical-labs-minecraft-skin-gene…
K00B404/image_gen_supaqueue_game_assets
ExynoxAI/MineSkin_Magic
issef/test
ThatBlondeGuy/Stable_Diffusion_Finetuned_Mine…

What are these models?



monadical-labs / **minecraft-skin-generator-sdxl**

like 10 | Follow Monadical Labs 15

Text-to-Image | Diffusers | Safetensors | English | StableDiffusionXLPipeline | minecraft | skins | gaming | stable diffusion | stable diffusion xl | Inference Endpoints | License: openrail

Model card | Files | Community | Train | Deploy | Use this model

Edit model card

Minecraft Skin Generator XL

Monadical is pleased to announce the official release of the Minecraft Skin Generator XL model. We had previously released the Minecraft Skin Generator model based upon Stable Diffusion 2. This new model offers significant improvements over the last generation of models.

Key Features

1. **Upgrade to Stable Diffusion XL** - Our model is now based upon the Stable Diffusion XL model, which greatly improves the quality of generated skins when compared to previous models.
2. **Transparent Layer Support** - The new model now supports the transparency layer in the hair and helmet section of the skin.

Examples

- 'Kelly Kapoor from the TV show "The Office"'

Downloads last month
16,584

Inference Examples

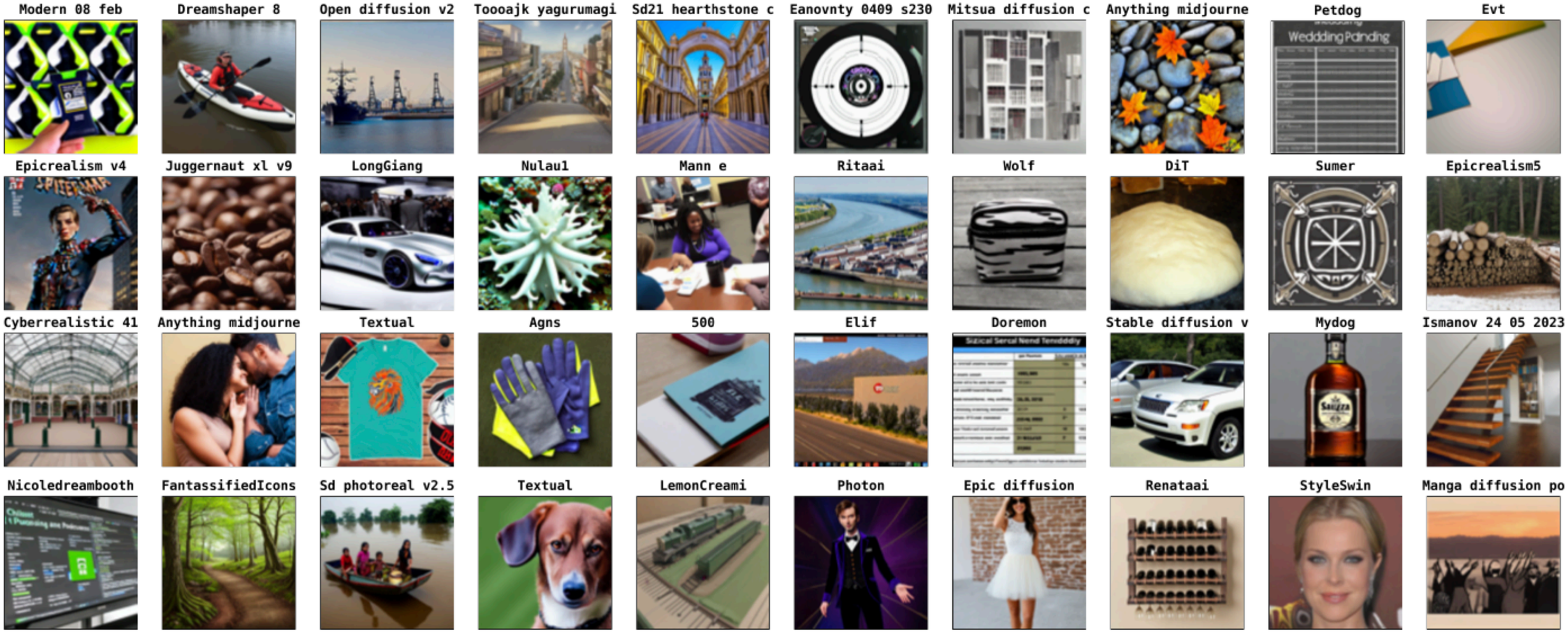
Text-to-Image

This model does not have enough activity to be deployed to Inference API (serverless) yet. Increase its social visibility and check back later, or deploy to Inference Endpoints (dedicated) instead.

Spaces using `monadical-labs/minecraft-ski…` 11

- Nick088/Stable_Diffusion_Finetuned_Minecraft_S…
- Nymbo/image_gen_supaqueue
- soiz/Stable_Diffusion_Finetuned_Minecraft_Skin_…
- Loxera/monadical-labs-minecraft-skin-generator-sdxl
- boisterous/minecraft-skin-generator
- gege002/monadical-labs-minecraft-skin-generator-sdxl
- dadadaadadadad/monadical-labs-minecraft-skin-gene…
- K00B404/image_gen_supaqueue_game_assets
- ExynoxAI/MineSkin_Magic
- issef/test
- ThatBlondeGuy/Stable_Diffusion_Finetuned_Mine…

Automatically sampled images



Sampled images from 4763 LDM text-to-image models from Hugging Face

The *Community Forensics* dataset

Systematically collected diffusion models

4763 LDM text-to-image models from Hugging Face

Other open models


VQ-GAN


CIPS


StyleGAN3


VQ-Diffusion

...

19 models

Commercial models


DALL-E 3


FLUX.1


Midjourney v6.1


Ideogram v2

...

11 models

Real images

from 11 datasets, w/ paired prompts


LAION

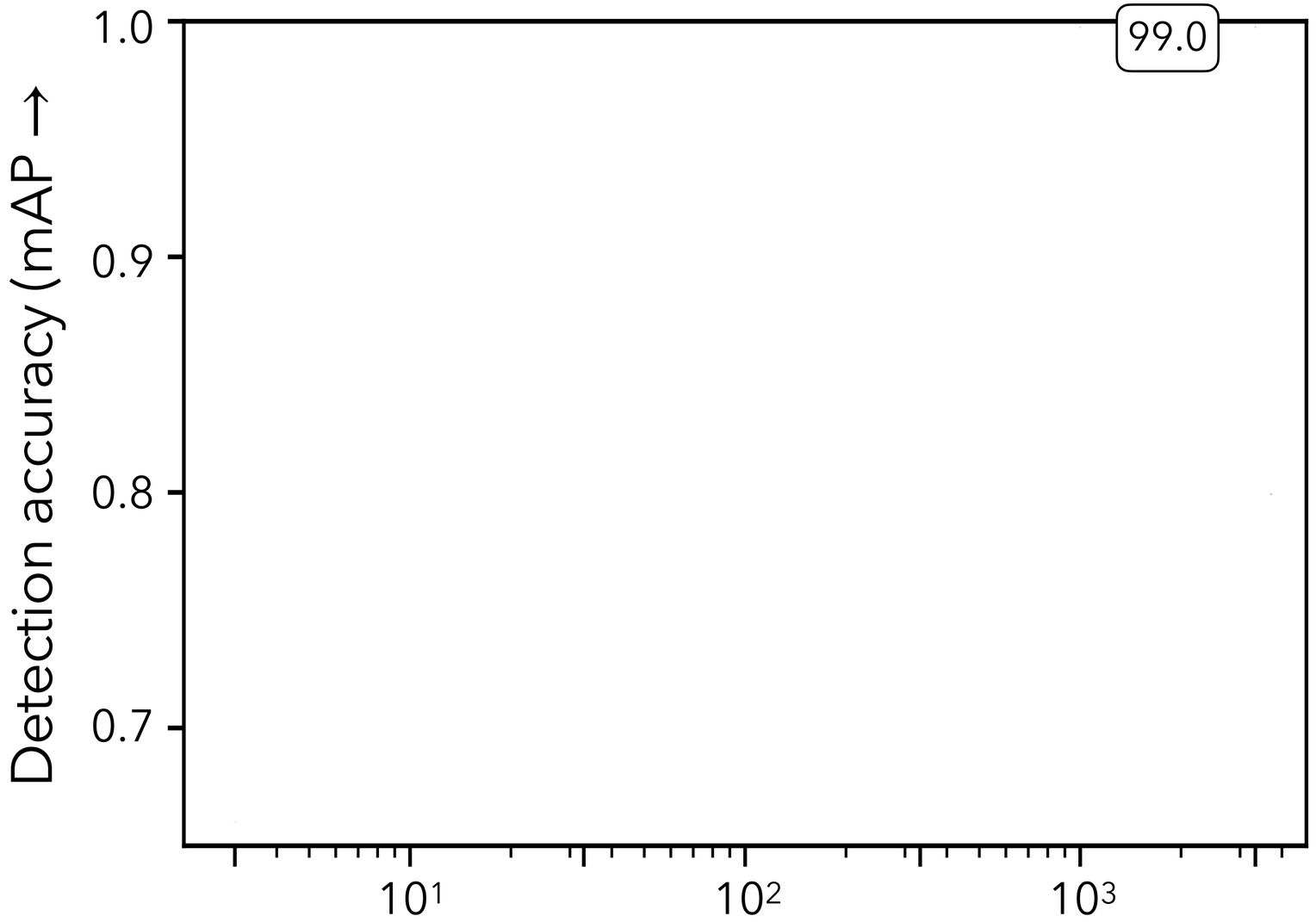

COCO


FFHQ

...

2.8M fake images from 4804 models.

What happens when you train on thousands of generators?



Test on images from:

- Latent Diffusion
- Commercial
- GAN
- Other
- Pixel Diffusion

- 100k images + 10 samples per data point.
- All test models are out-of-distribution.

Challenges on the horizon

- Lots of ways to make fake images.
- If we know what methods were used, it’s much easier.
- But it’s hard to capture all of them!
- False positives are still a huge problem.
- So are postprocessing operations, like cropping and compression.
- Need methods that can handle unseen models.
- Alternative approaches: watermarking, signatures, etc.

Open-ended discussion

- How susceptible are people to fake images?
- Is there any hope of detecting “most” fake images?
- Under what situations might it be important and/or feasible?
- How do we deal with false positives?

Thank you!