Datasets, bias, and adaption

Bill Freeman, Antonio Torralba, Phillip Isola

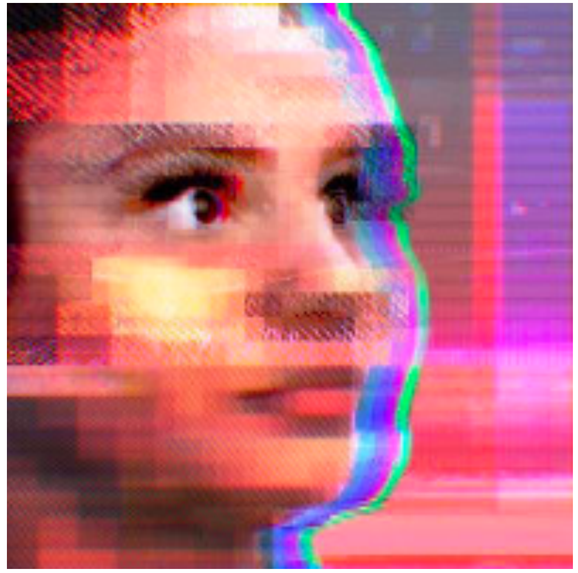
6.819 / 6.869

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.

Microsoft’s Tay chatbot



Chatbot released on twitter.

Learned from interactions with users (?)

Started mimicking offensive language, was shut down.

what is the yellow thing?

Submit

Predicted top-5 answers with confidence:

frisbee

79.844%

surfboard

7.319%

banana

2.844%

lemon

2.438%

surfboards

1.252%

how many trains are in the picture?

Submit

Predicted top-5 answers with confidence:

3

30.233%

5

18.270%

4

17.000%

2

11.343%

6

7.806%

Of number questions (e.g. “how many…”), 26.04% of the time, the answer is 2

Of yes/no questions, 58.83% of the time, the answer is yes

[VQA, Agrawal, Lu, Antol et al., https://arxiv.org/pdf/1505.00468.pdf

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

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

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

Training data

$\mathbf{x}$ $\mathbf{y}$

Test data

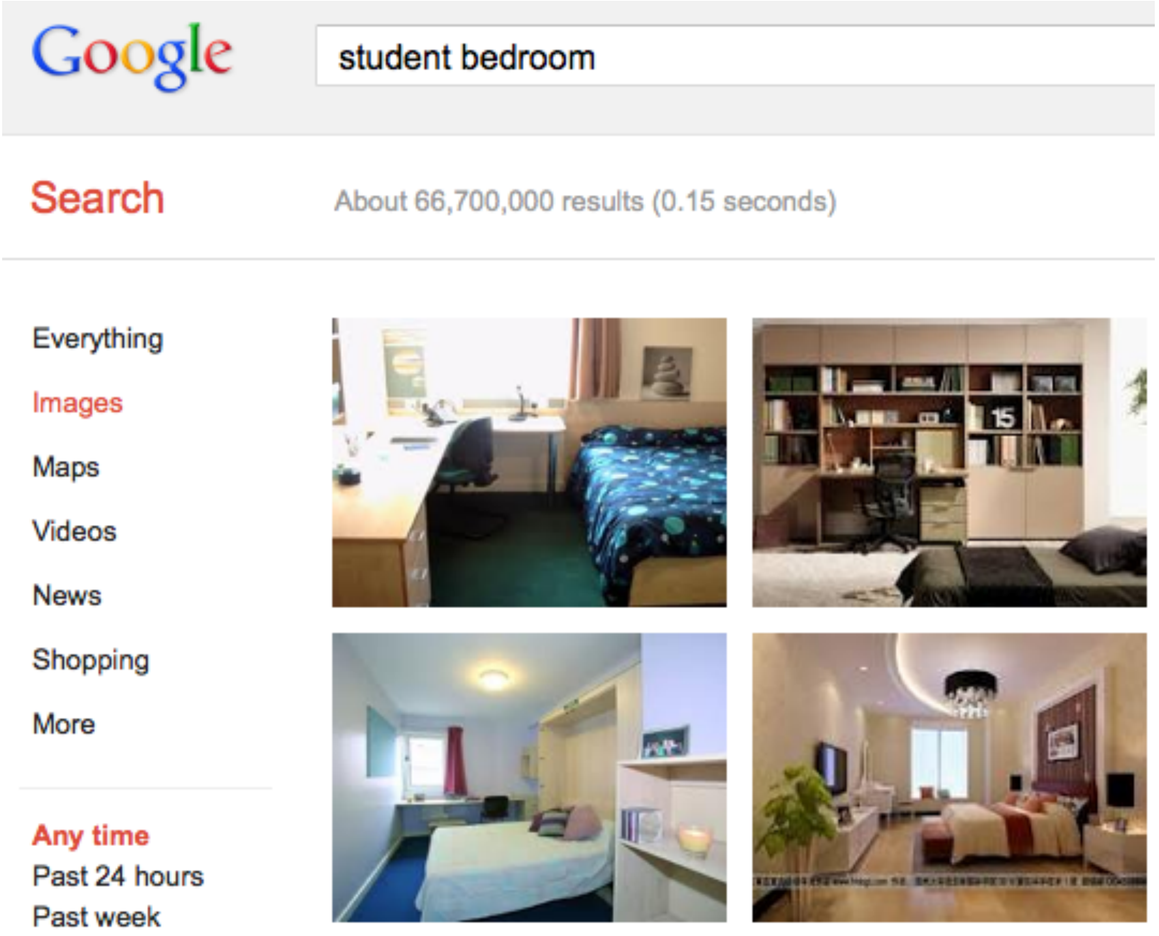
$\mathbf{x}'$

Training data

Test data

Training data

What Google thinks are student bedrooms



Test data

Training data

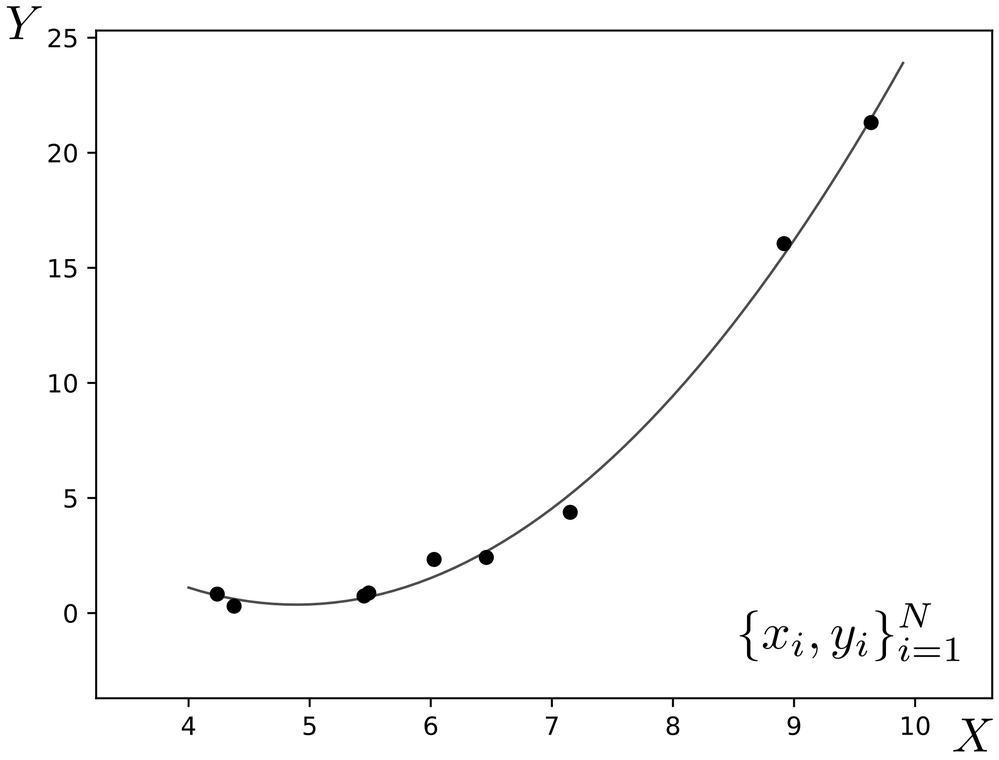
Driving simulator (GTA)

Test data

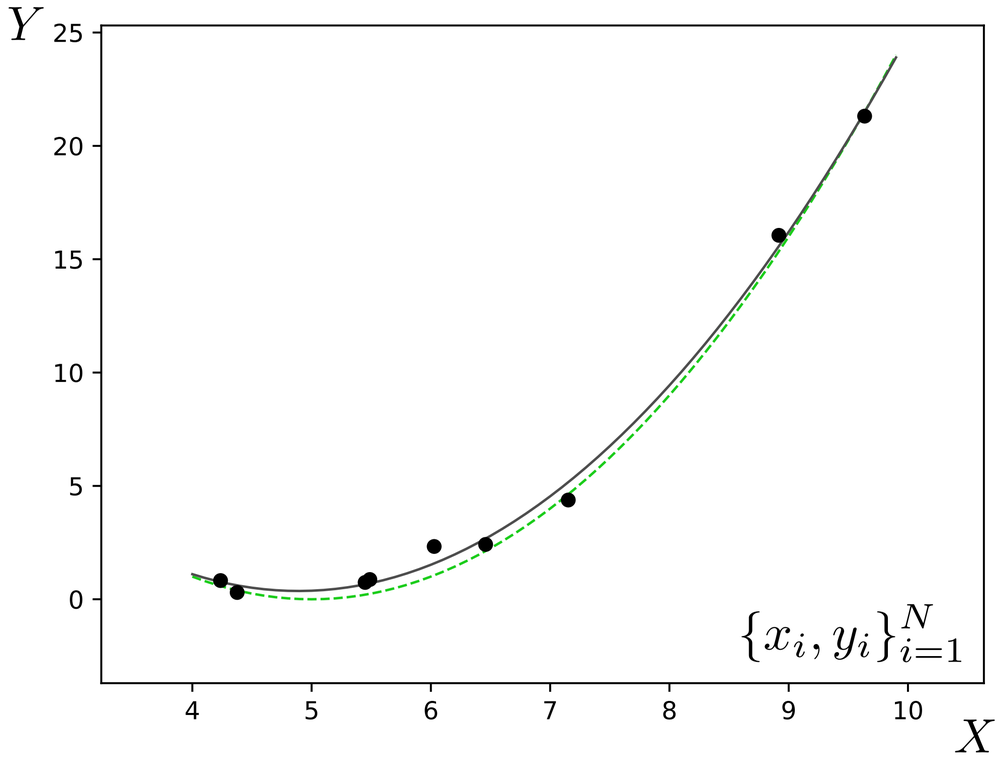
Driving in the real world

Let’s revisit the problem of generalization

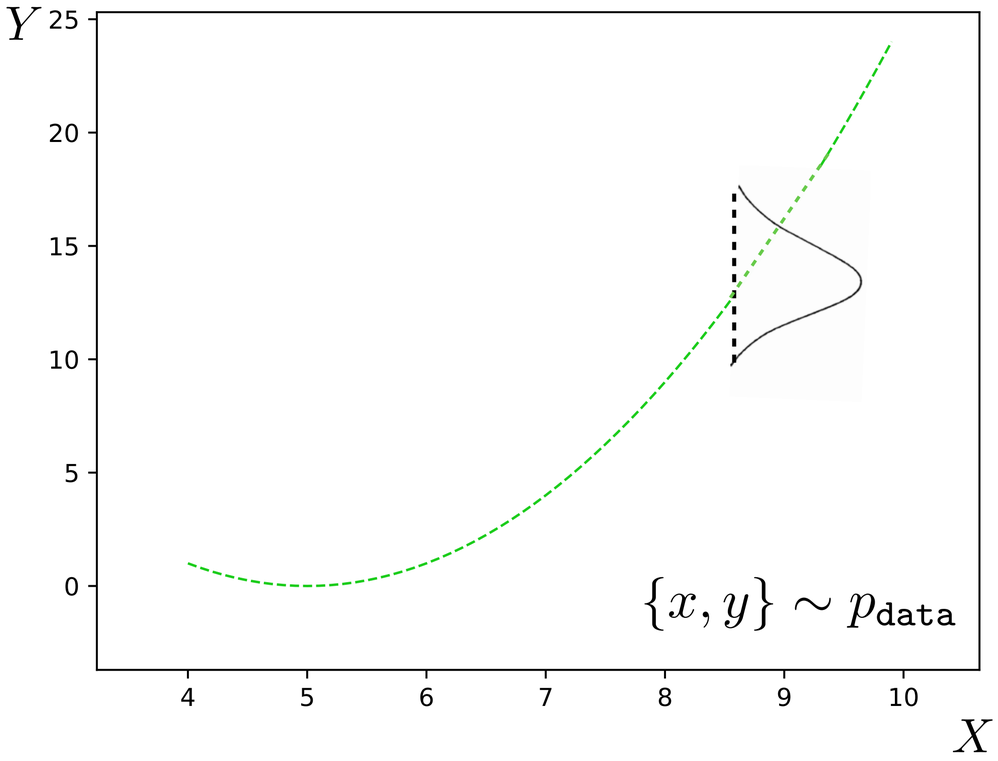
Training data



Training data



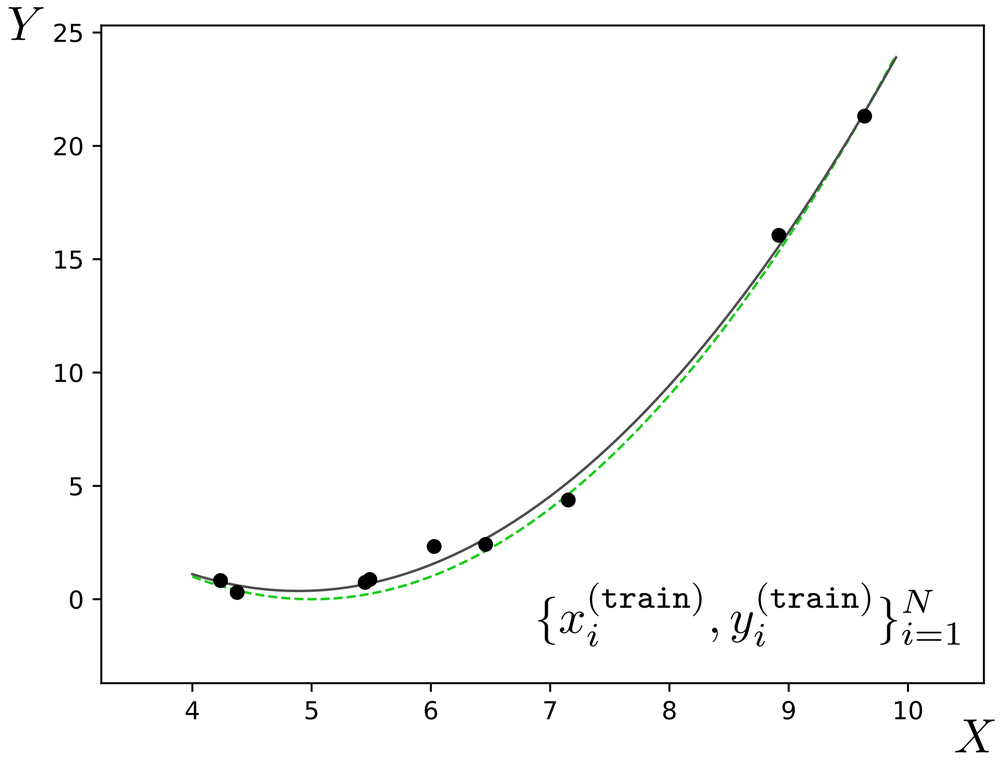
Test data



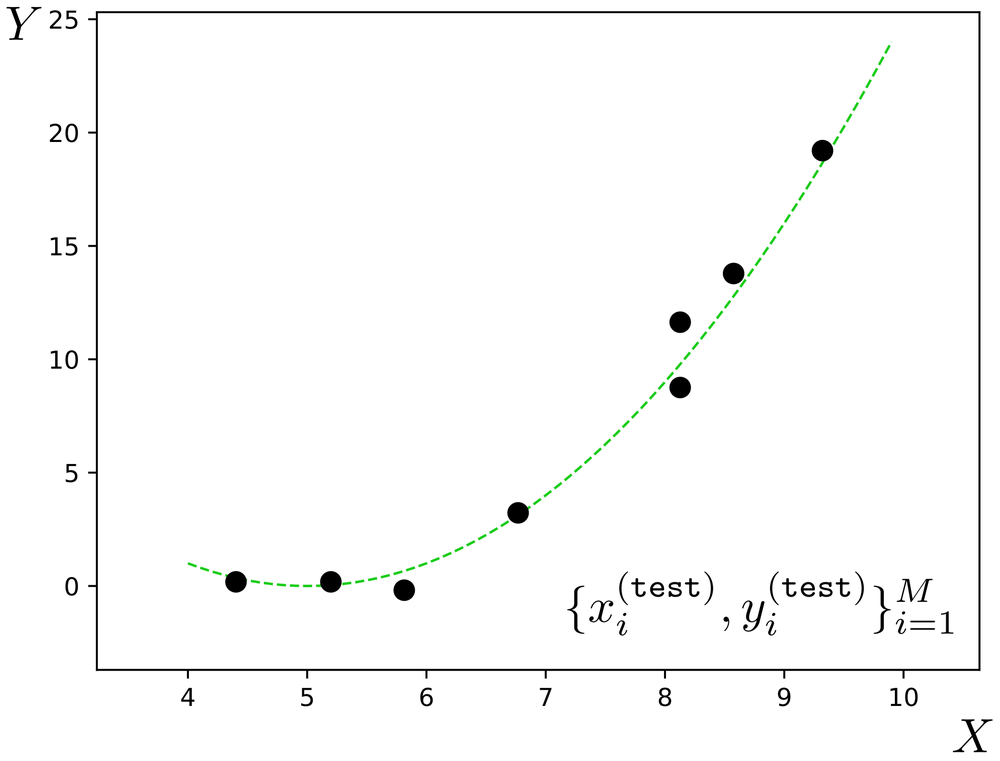
True **data-generating process**

$p_{\texttt{data}}$

Training data



Test data



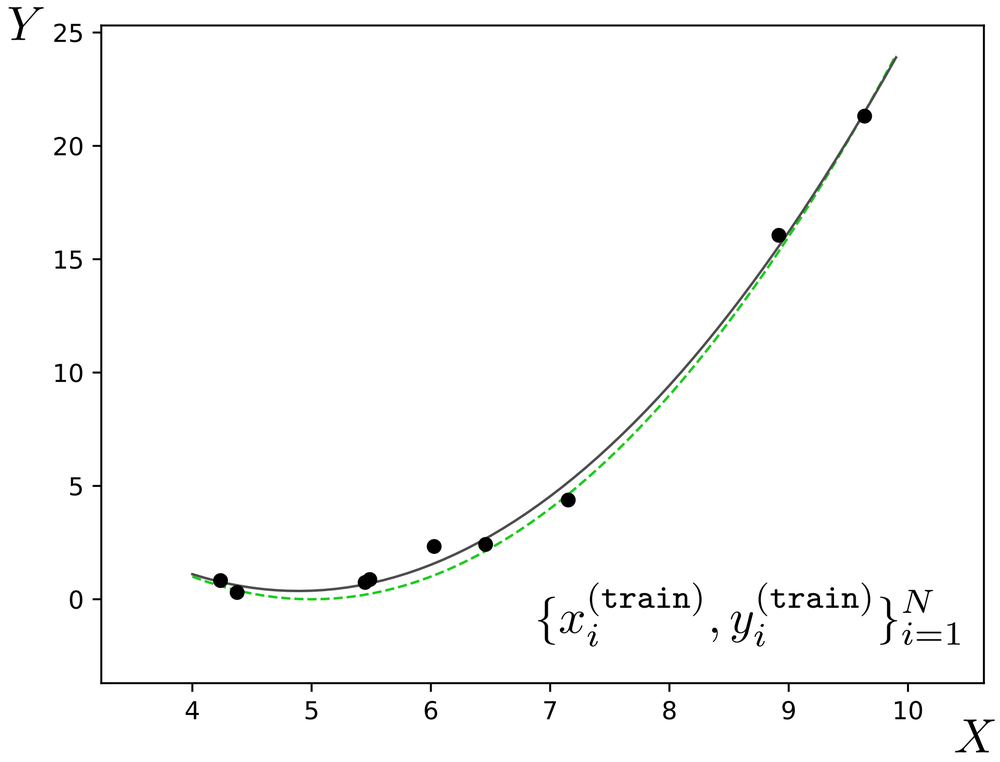
True **data-generating process**

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

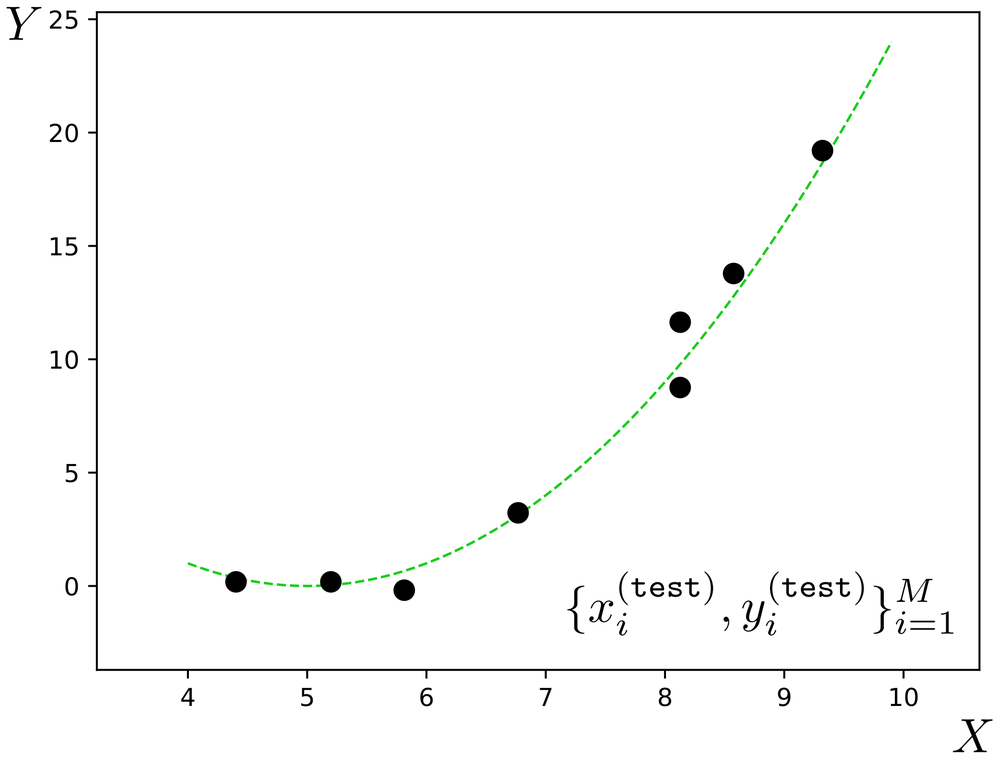
$$\{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

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

Training data



Test data

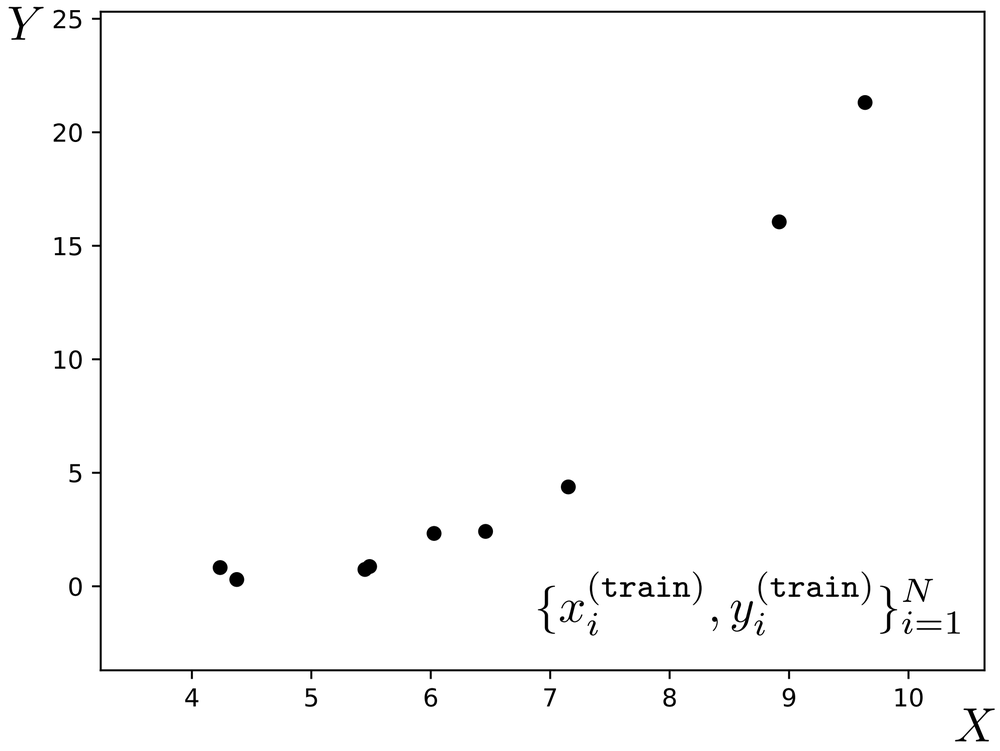


This is a huge assumption!
Almost never true in practice!

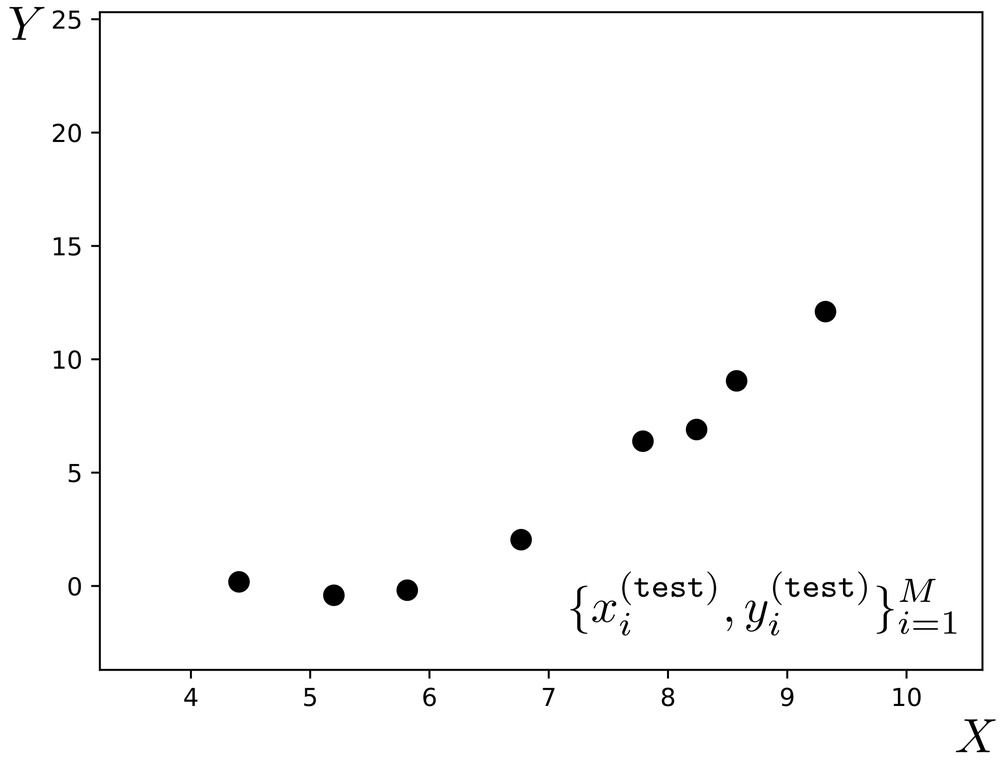
$$\{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} \overset{\texttt{iid}}{\sim} p_{\texttt{data}}$$

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

Training data



Test data

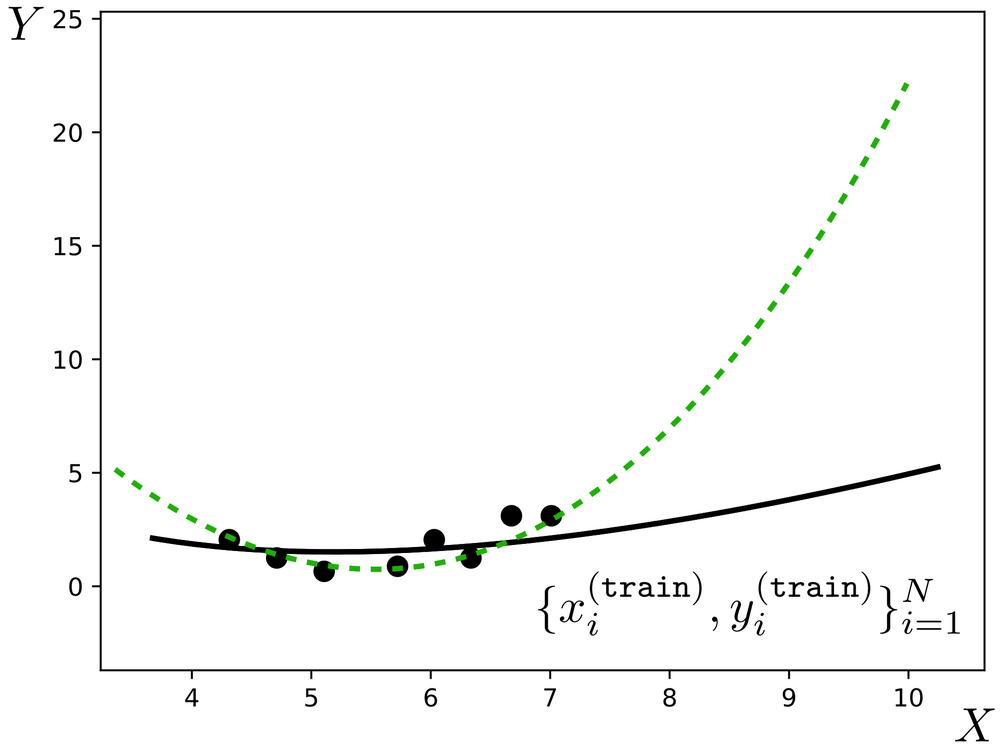


Much more commonly, we have

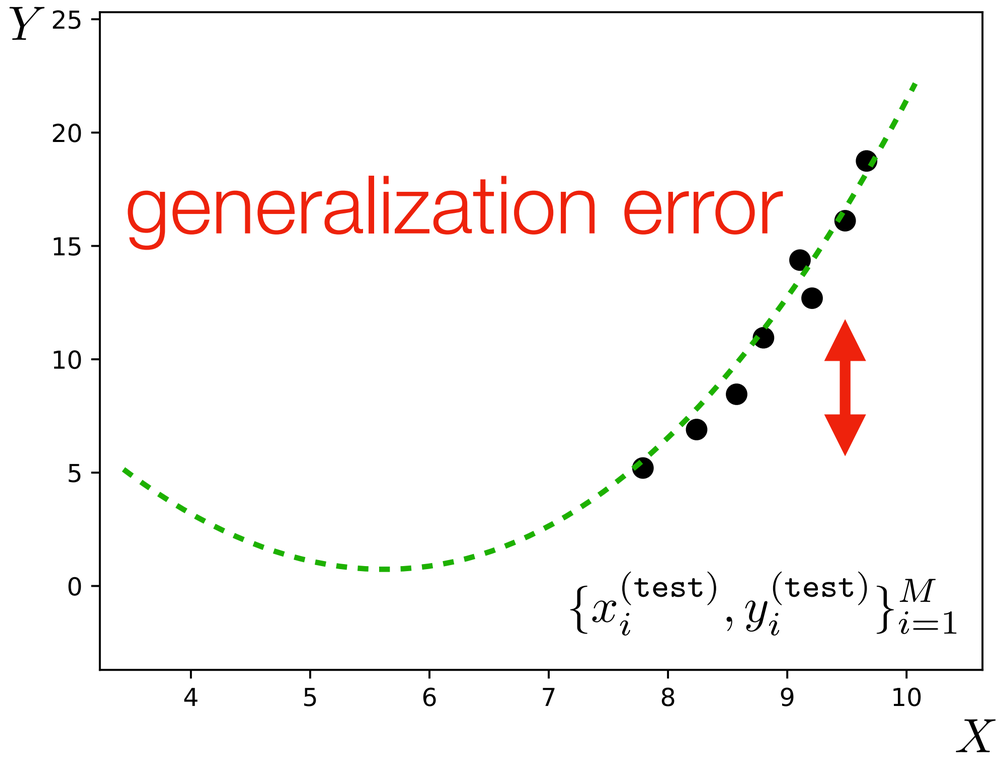
$$p_{\texttt{train}} \neq p_{\texttt{test}}$$

$$\boxed{\begin{aligned} \{x_i^{(\texttt{train})}, y_i^{(\texttt{train})}\} &\overset{\texttt{iid}}{\sim} p_{\texttt{train}} \\ \{x_i^{(\texttt{test})}, y_i^{(\texttt{test})}\} &\overset{\texttt{iid}}{\sim} p_{\texttt{test}} \end{aligned}}$$

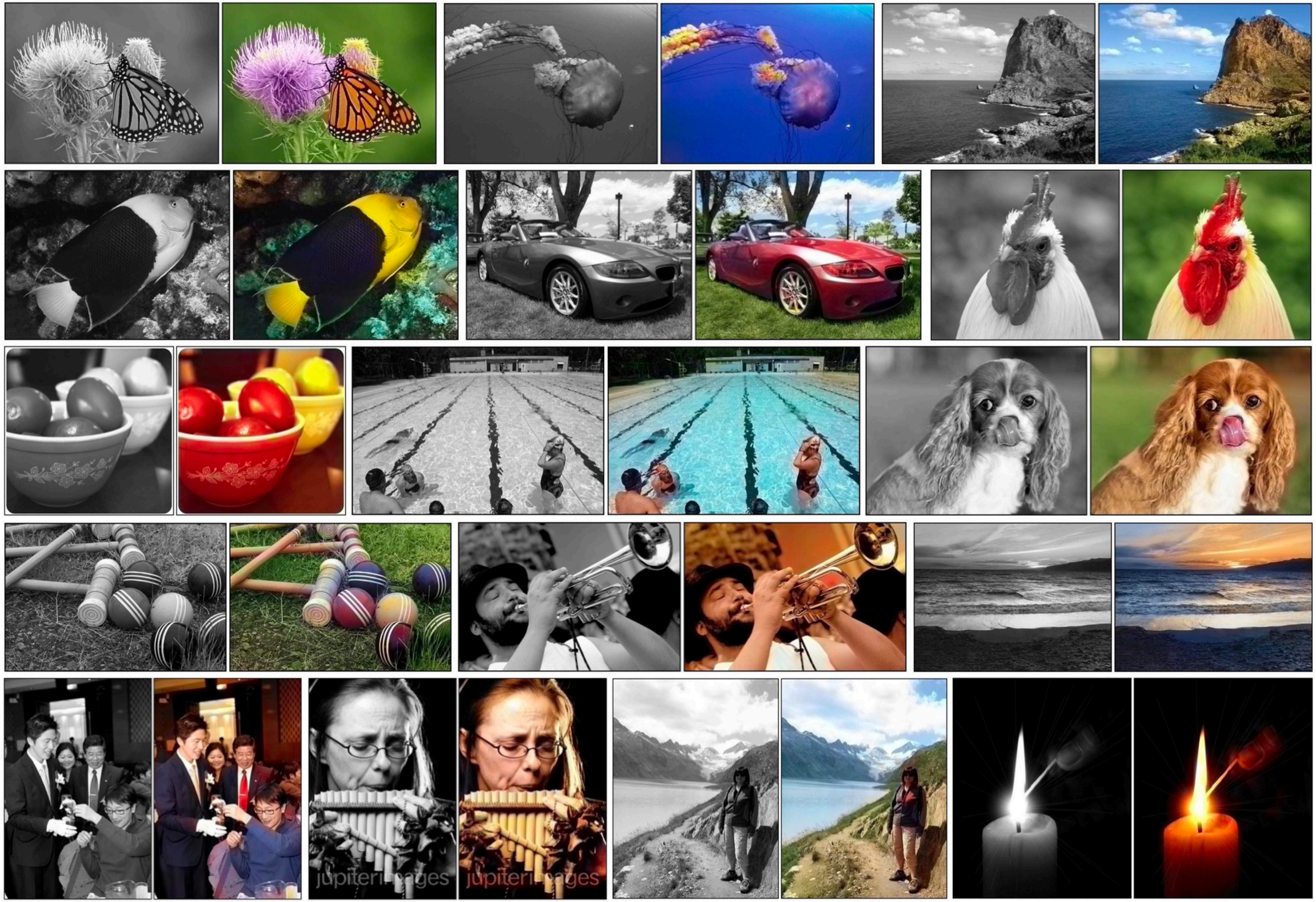
Training data



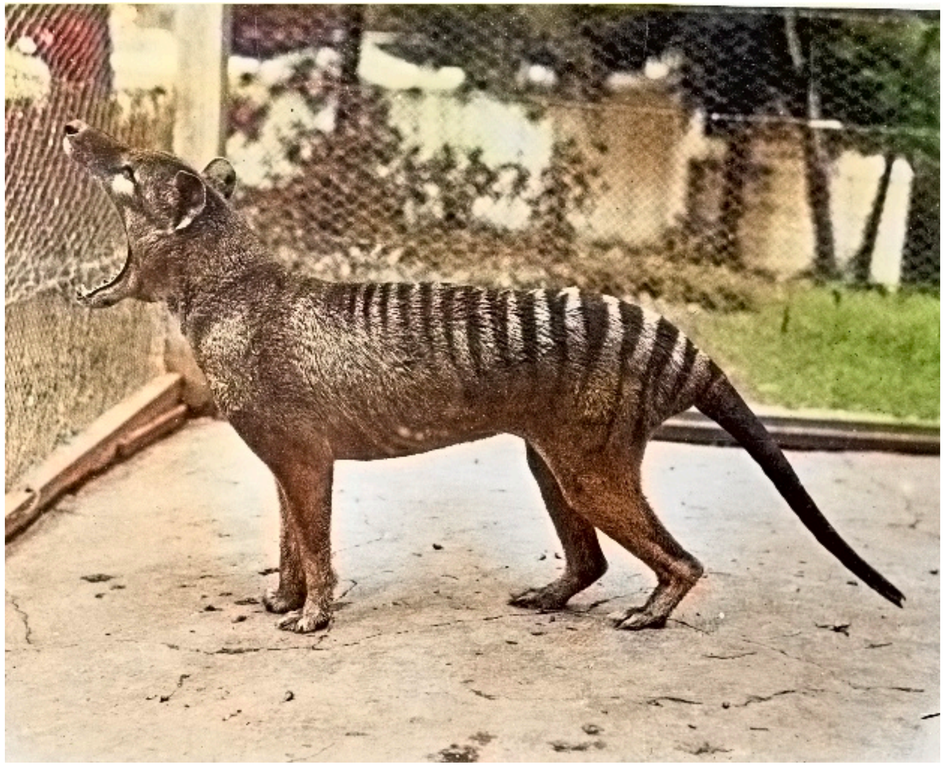
Test data



Our training data did cover the part of the distribution that was tested (**biased data**)


jupiterimages
jupiterimages

u/Rafael_P_S



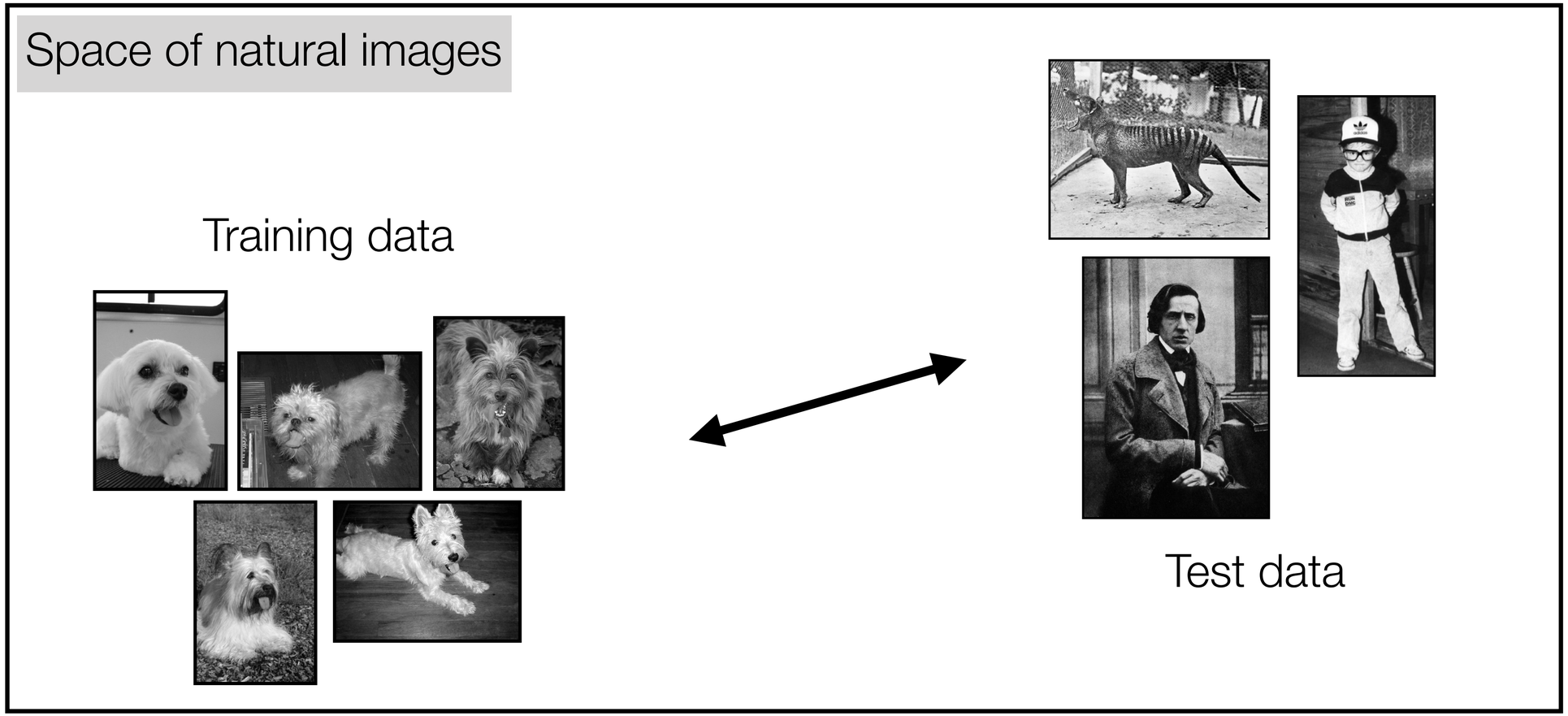
Thylacine

Chopin

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.



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.

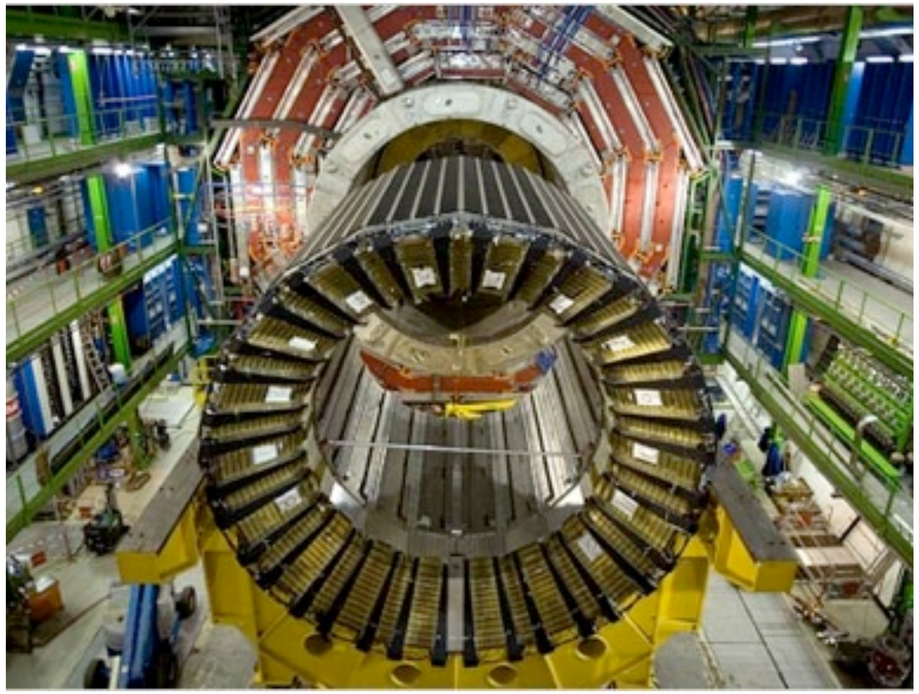
https://www.nytimes.com/2018/02/09/technology/facial-recognition-race-artificial-intelligence.html

How can we collect good data?

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

Crowdsourcing

The value of data



The Large Hadron Collider

$ 10^{10}

Amazon Mechanical Turk

$ 10^{2} - 10^{4}

But can humans collect good data?

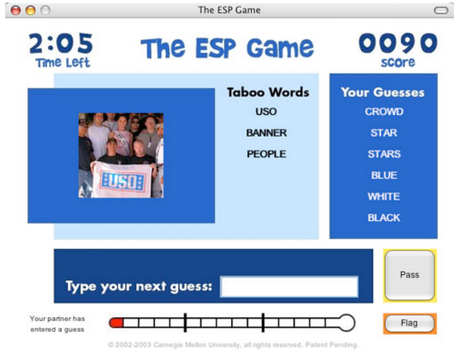

Google
mug
I OWN YOU.
SHARING IS CARING
Mom
/mäm/
n. Absolutely never wrong about anything ever.

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
Artificial Artificial Intelligence

Labeling because it gives you added value

Visipedia

(Belongie, Perona, et al)

Just for labeling

LabelMe

Beware of the human in your loop

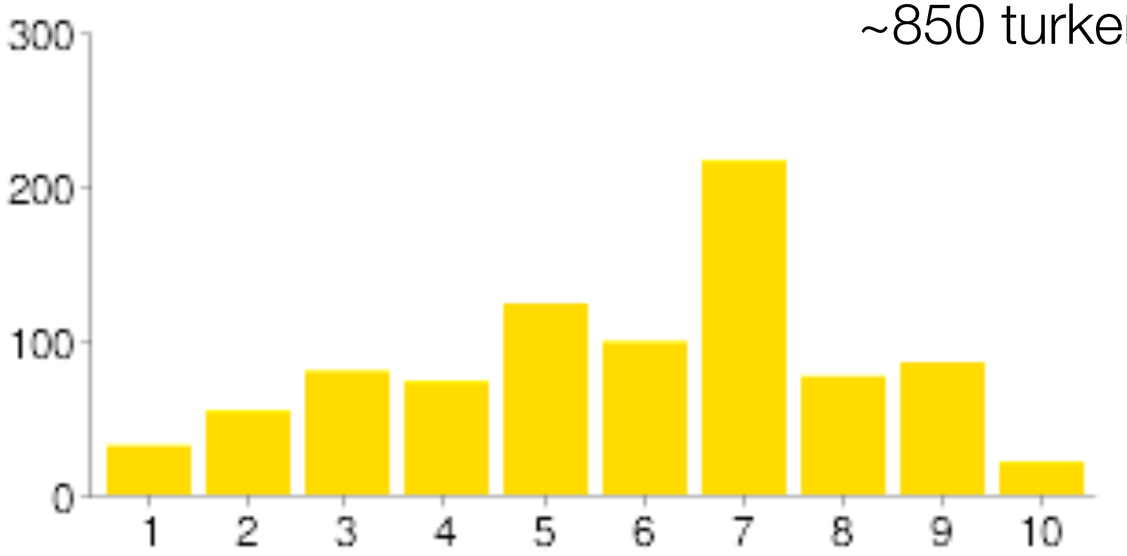
- What do you know about them?
- Will they do the work you pay for?

Let’s check a few simple experiments

People have biases…

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

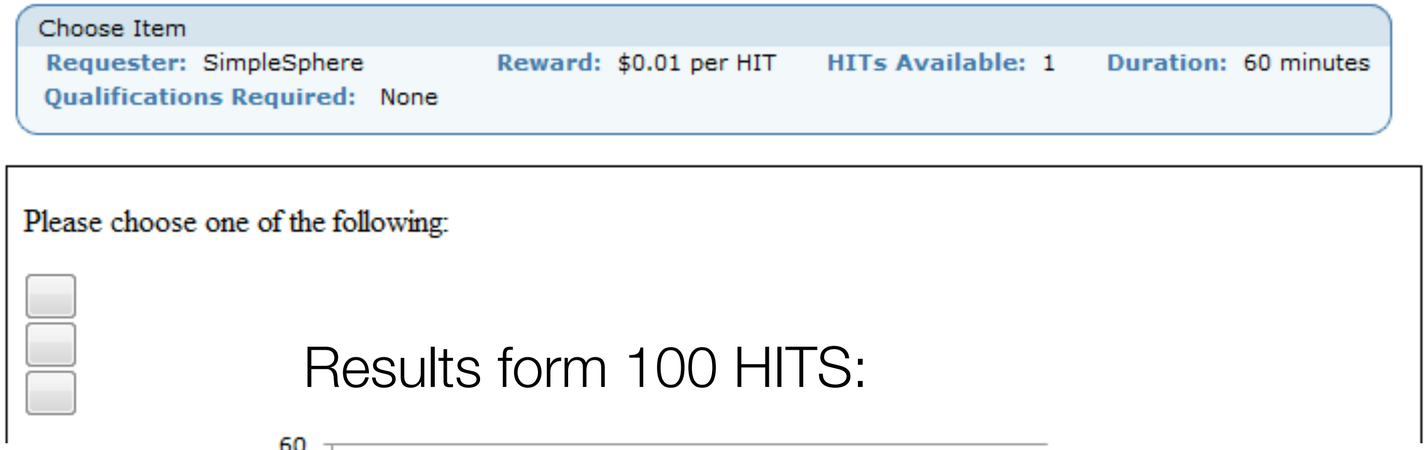
~850 turkers



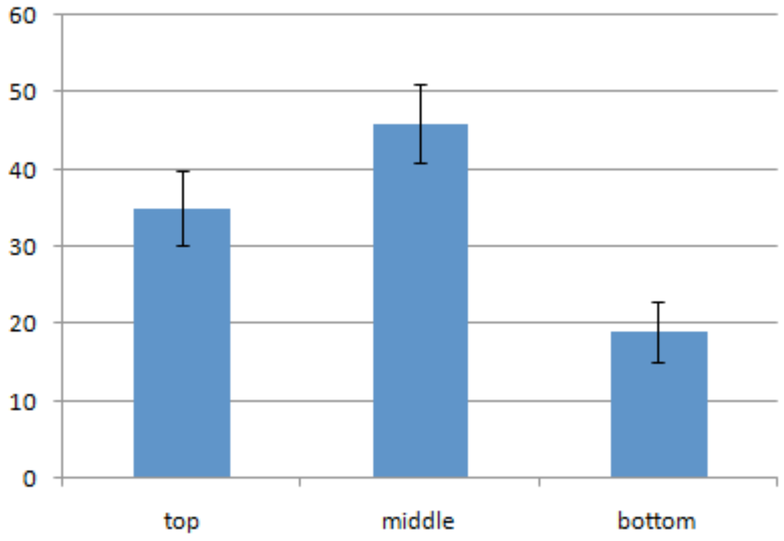
Experiment by Greg Little

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

Do humans have consistent biases?



Results form 100 HITS:



Experiment by Greg Little

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

Do humans do what you ask for?

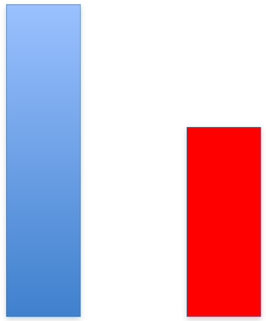
Flip a coin

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

Qualifications Required: None

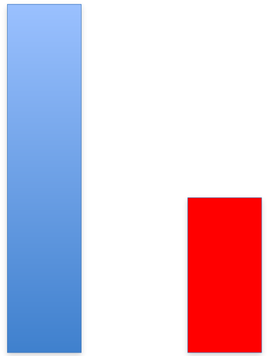
Please flip an actual coin and type either H or T below.

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/deneme/

Are humans reliable even in simple tasks?

Choose the given item.

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

Qualifications Required: None

Please click button B:

B

C

A

Results of 100 HITS:

A: 2

B: 96

C: 2

Experiment by Greg Little

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

Tool went online July 1st, 2005

Labelme.csail.mit.edu

B. Russell, A. Torralba, K. Murphy, W.T. Freeman. IJCV 2008

amazonmechanical turk
Artificial Artificial Intelligence

Your Account | HITs | Qualifications | 56,035 HITs available now

Bryan C Russell | Account Settings | Sign Out | Help

All HITs | HITs Available To You | HITs Assigned To You

Search for HITs containing that pay at least $ 0.00 for which you are qualified GO

Timer: 00:00:13 of 60 minutes

Finished with this HIT? Submit HIT

Let someone else do it? Return HIT

Total Earned: $0.01
Total HITs Submitted: 12

Automatically accept the next HIT

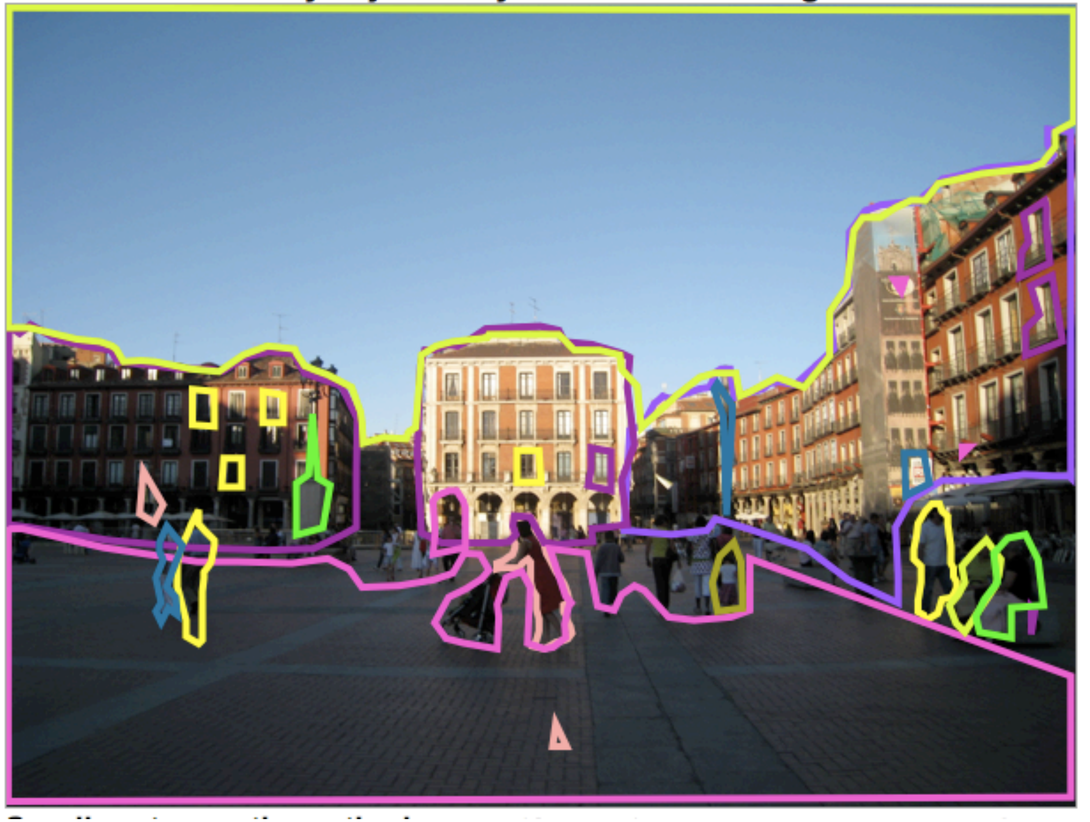
LabelMe: Label objects in this image

Requester: Bryan C Russell

Qualifications Required: None

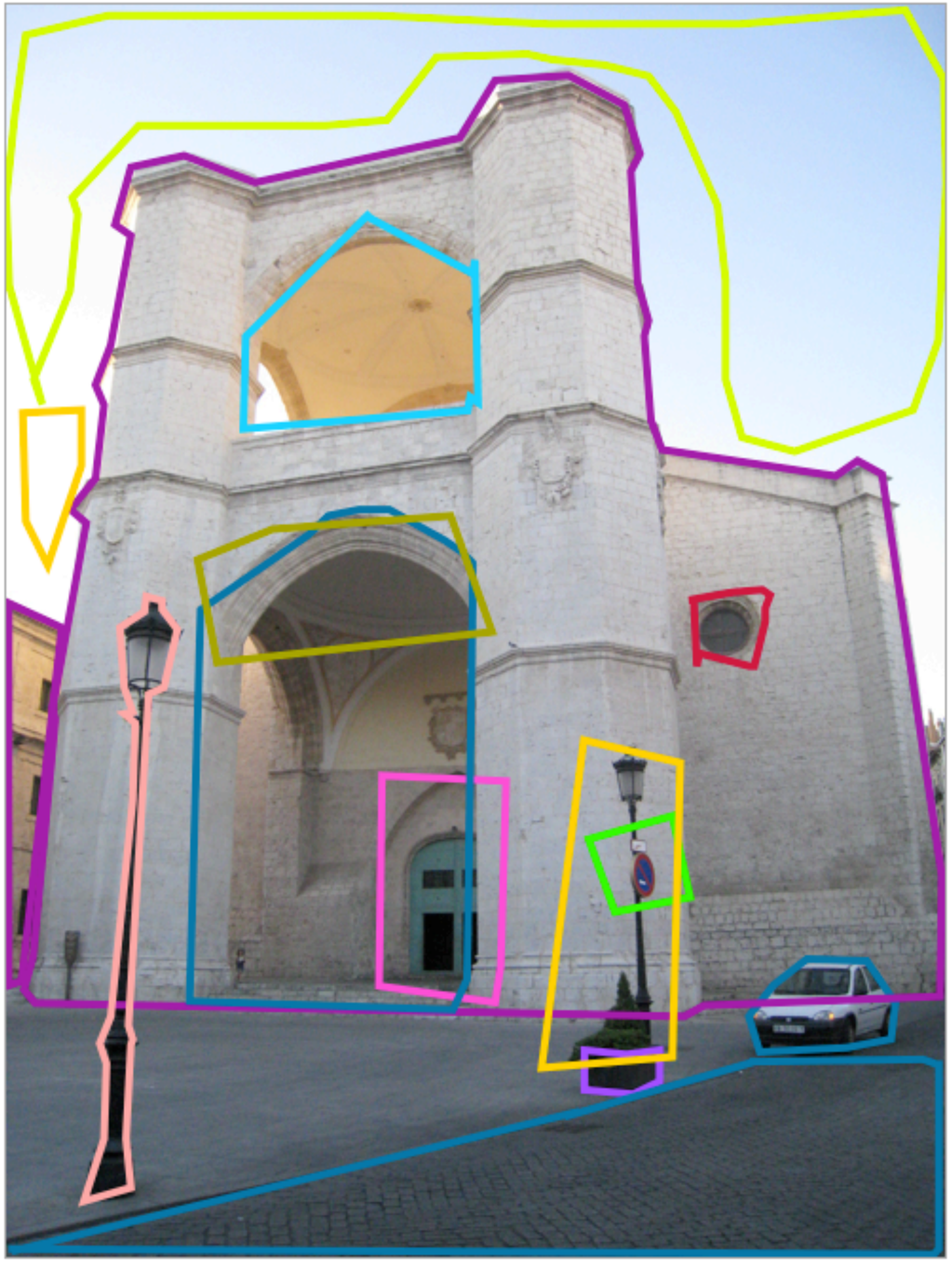
Reward: $0.01 per HIT HITs Available: 269 Duration: 60 minutes

Please label as many objects as you want in this image. Scroll down to see the entire image. Submit HIT



1 cent

Task: Label one object in this image



(IMG, XML)

car
building
building
lampost
planter box
This is a window.
This is the street.
This is a balcony.
door
entrance
Traffic sign
SKY
cloud
arch
street light

LabelMe iterations

1) Label as many objects as you can

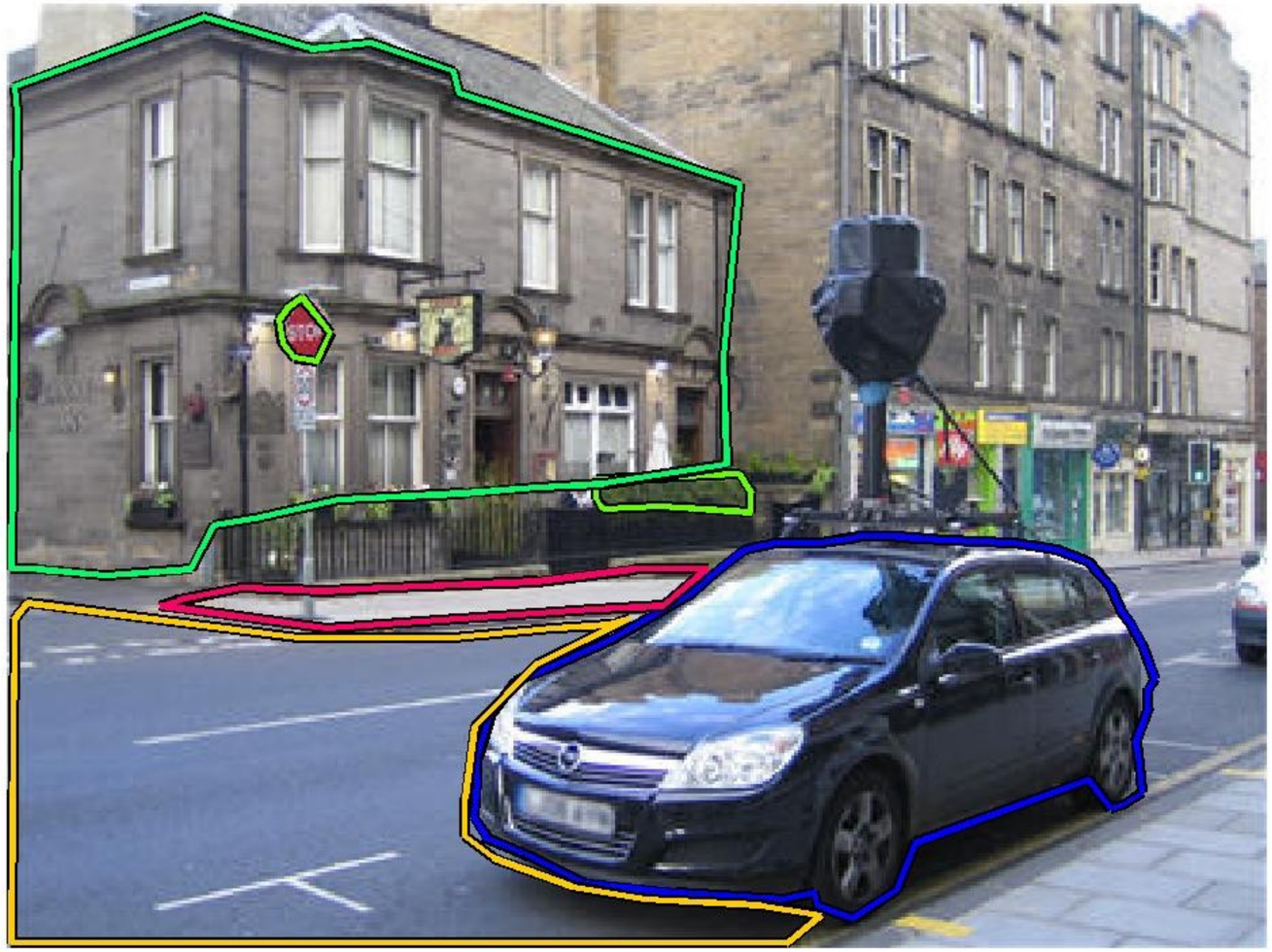
2) Delete any wrong polygon

3) Go to 1


STOP

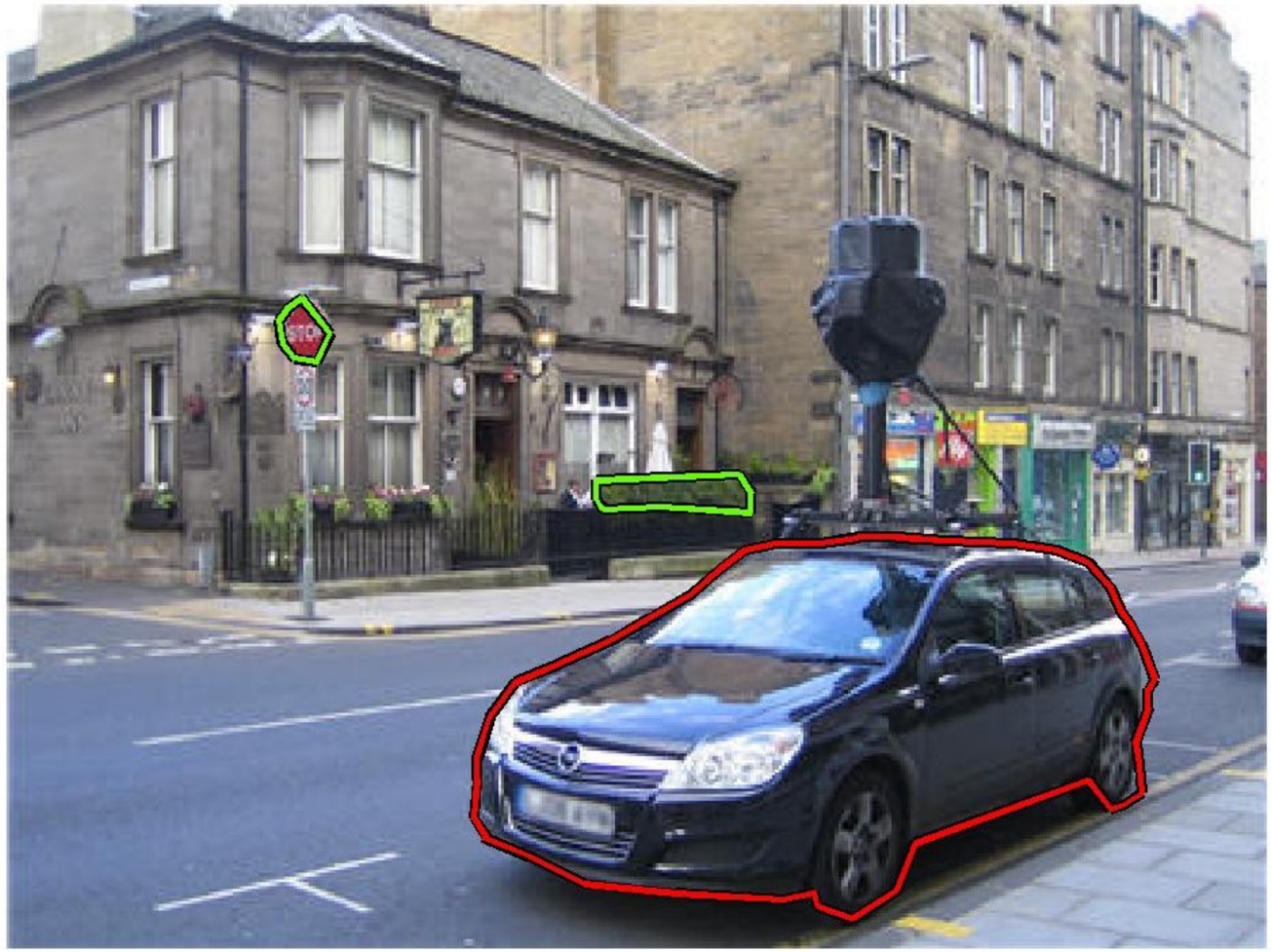
Label some objects

1



Delete any wrong polygons

0



Label some objects

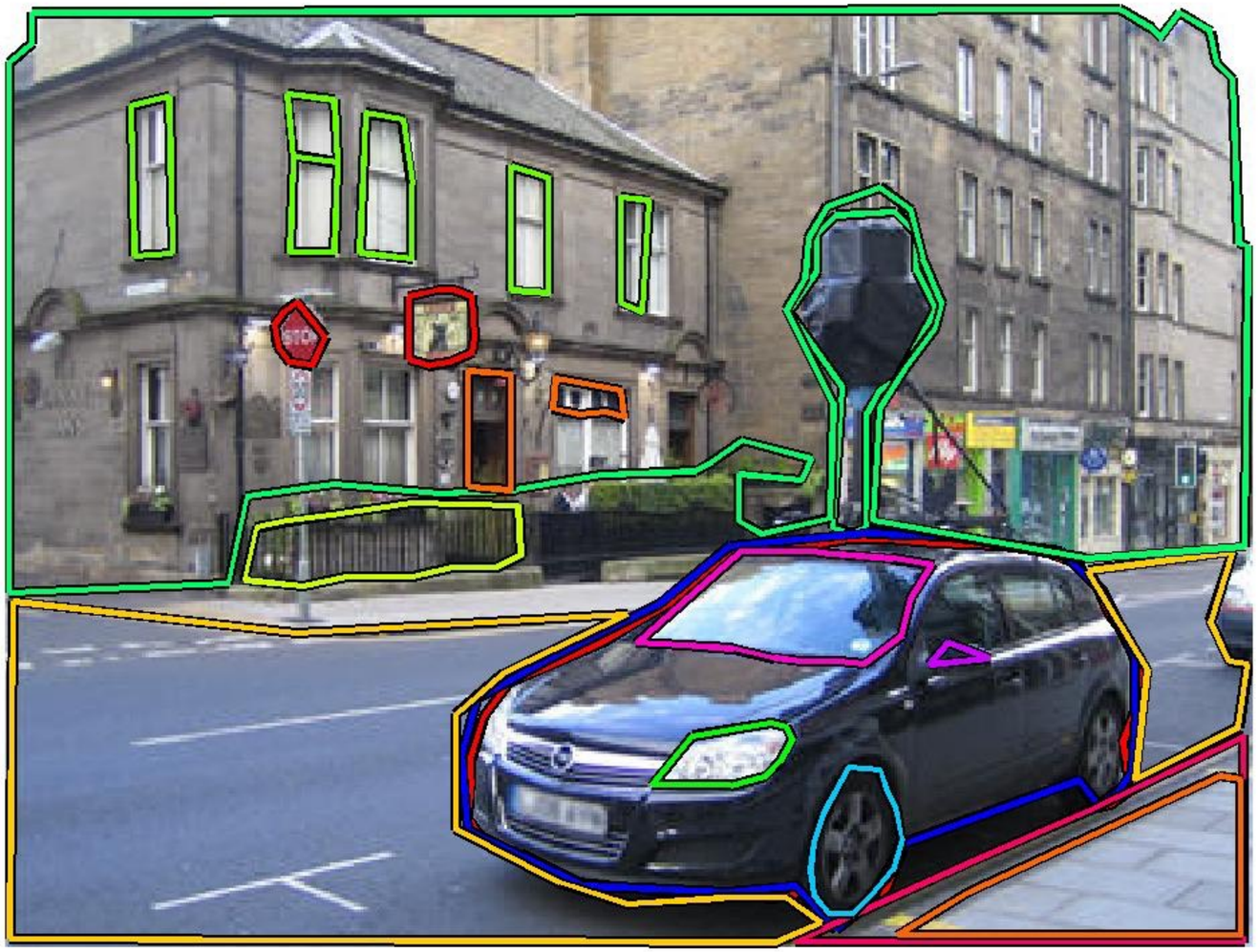
1

Delete any wrong polygons

0

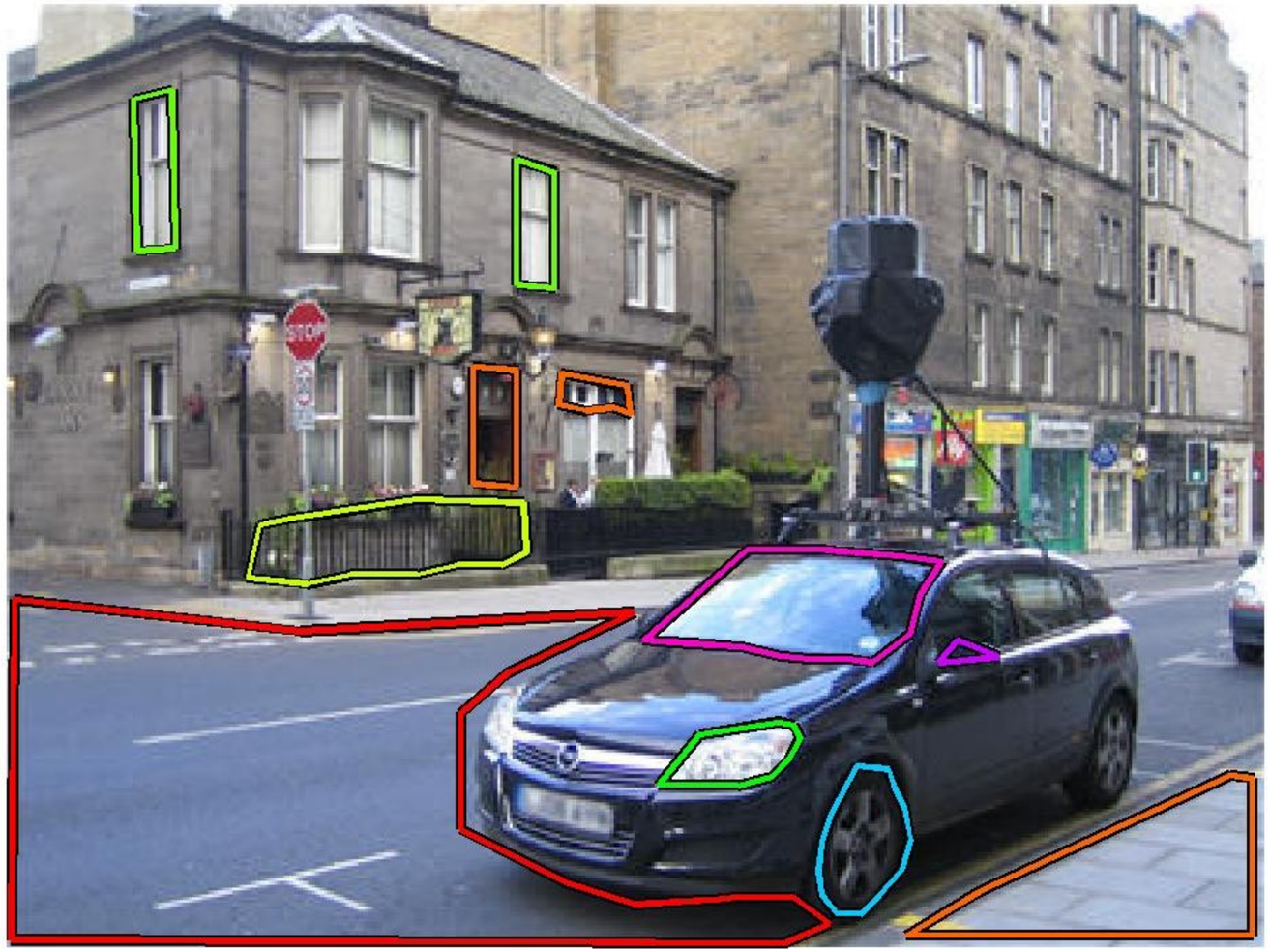
Label some objects

1



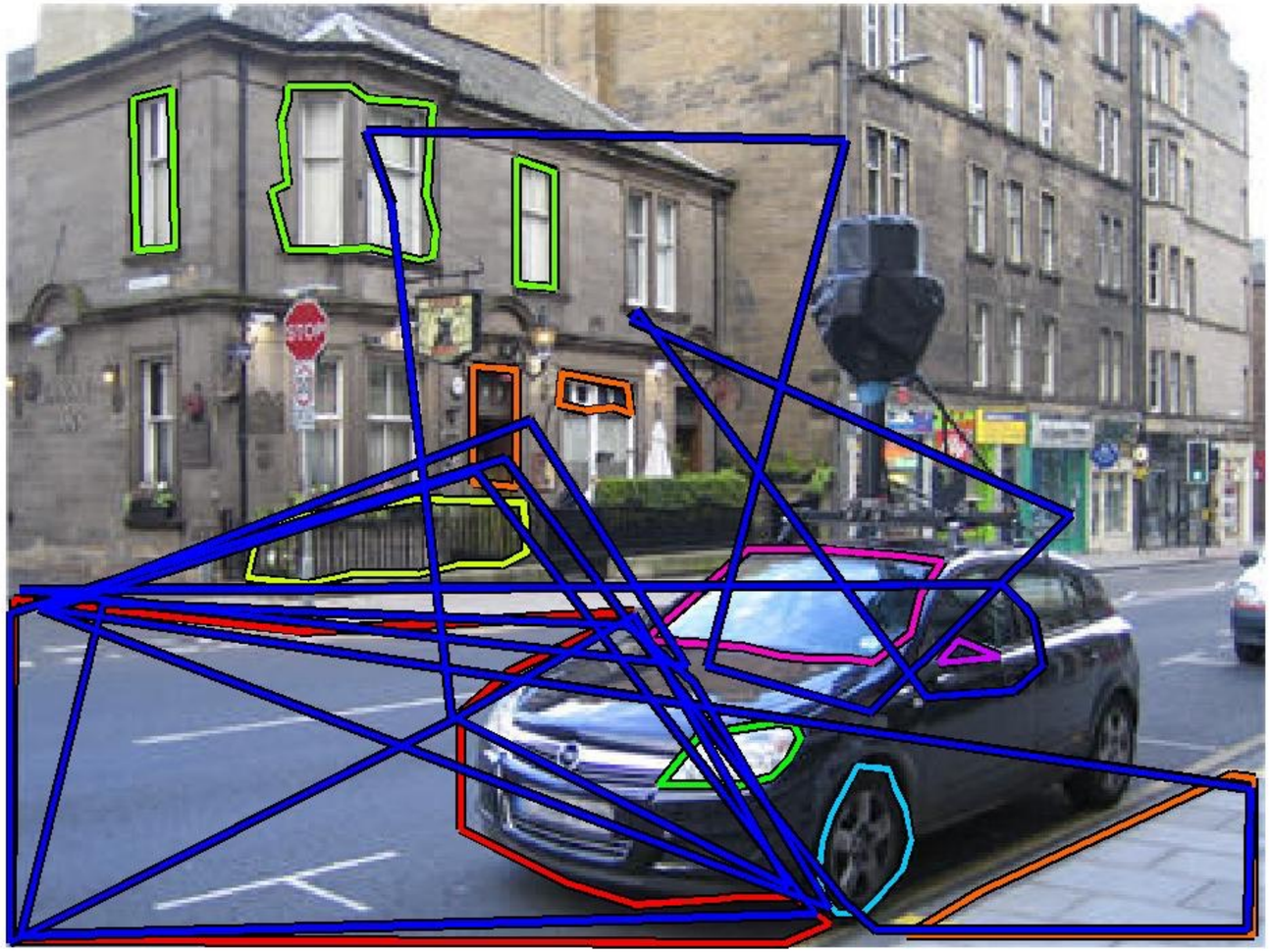
Delete any wrong polygons

0

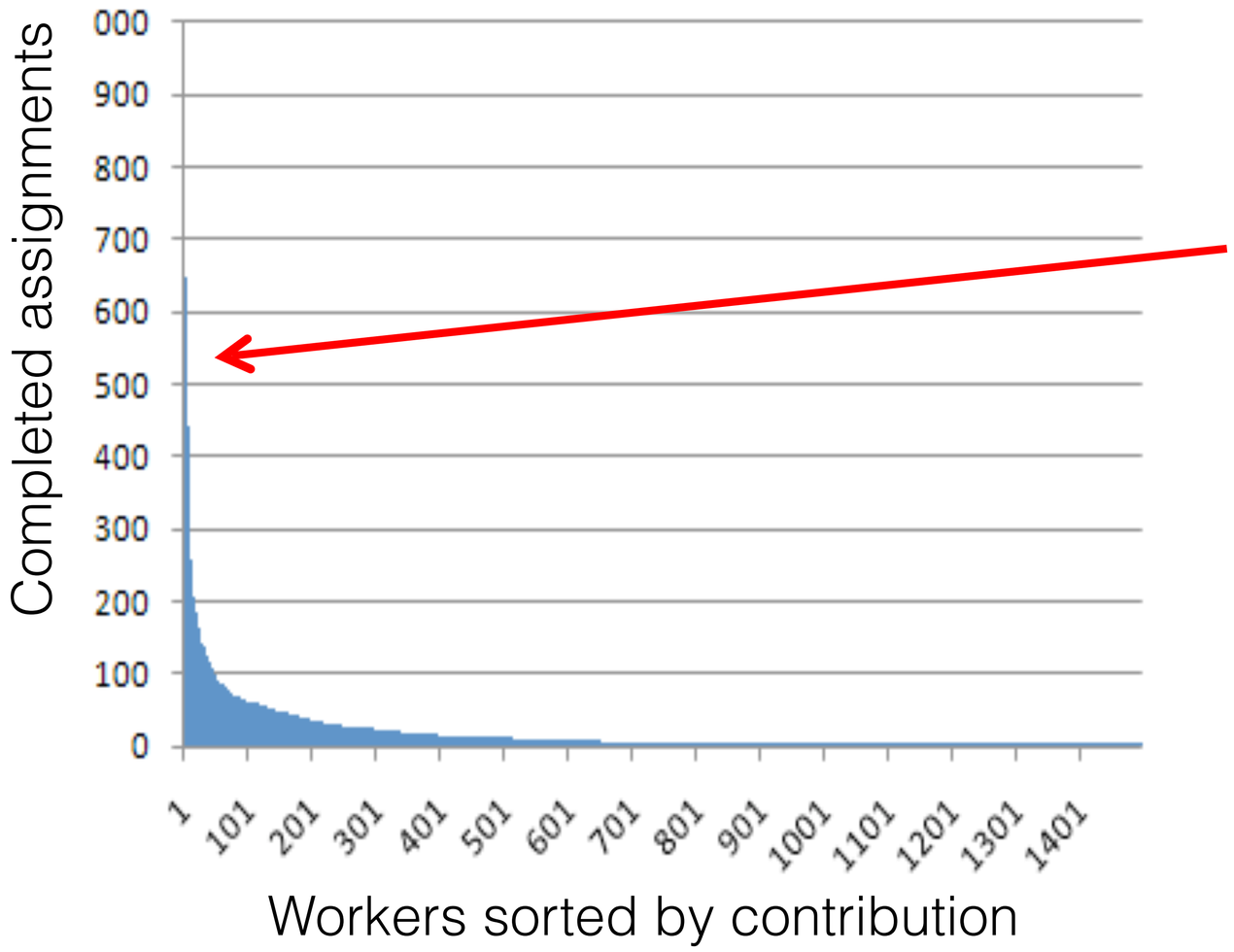


Label some objects

1

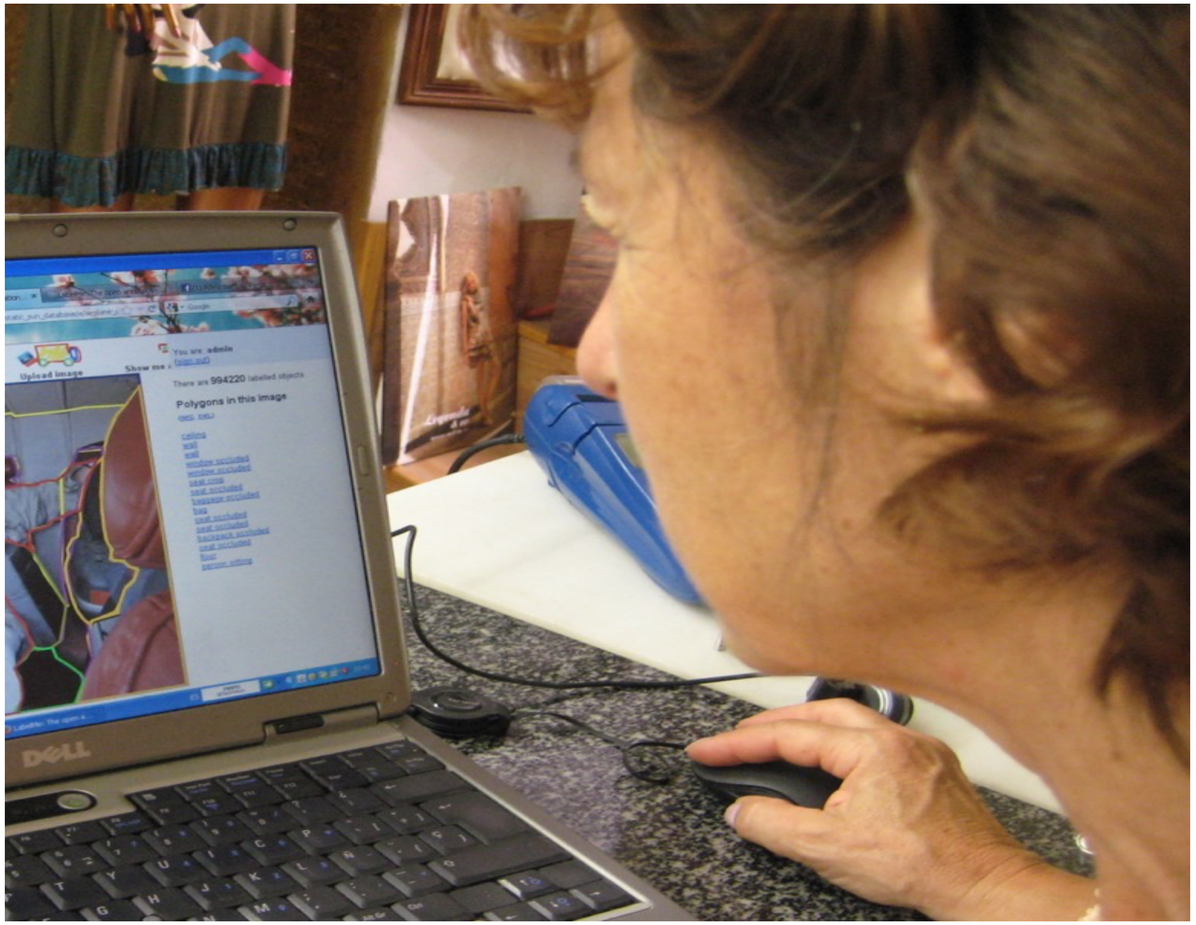


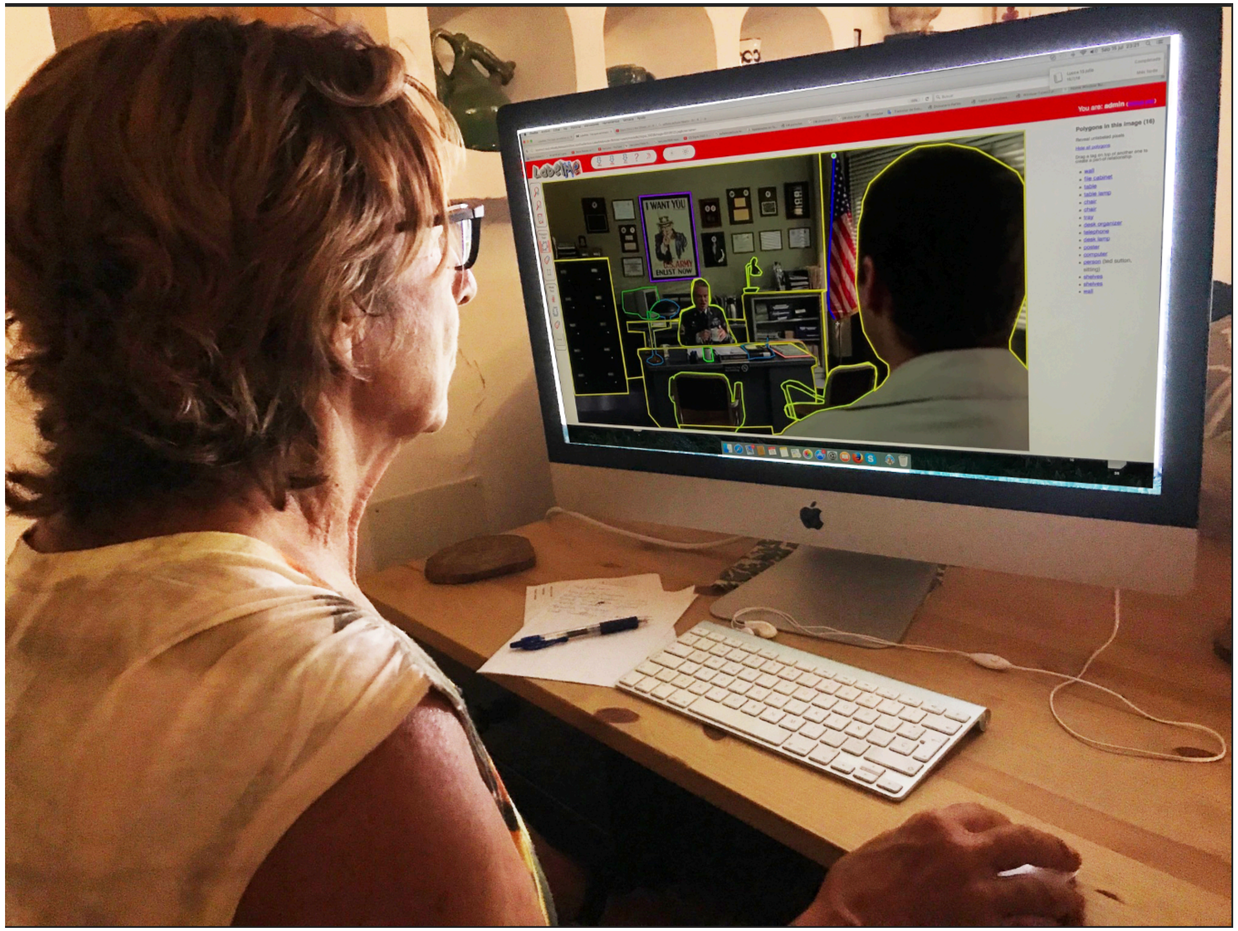
Who does the work?



Let's hire that one!

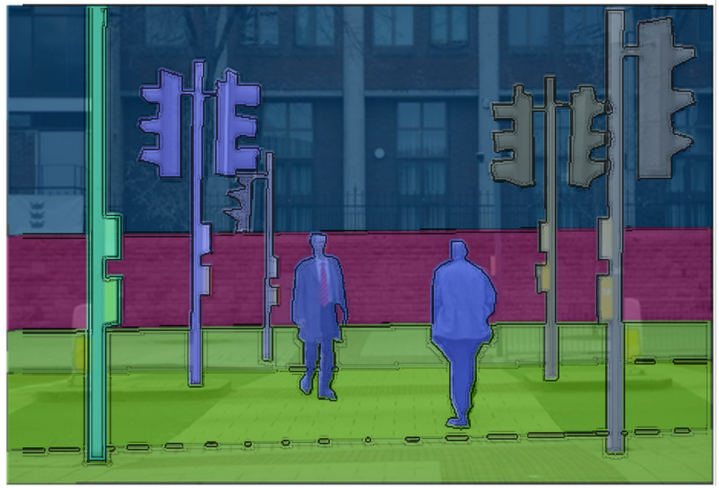
From http://groups.csail.mit.edu/uid/dene


You are: admin
Upload image
There are 994220 labelled objects
Polygons in this image
DELL


I WANT YOU
ENLIST NOW

COCO

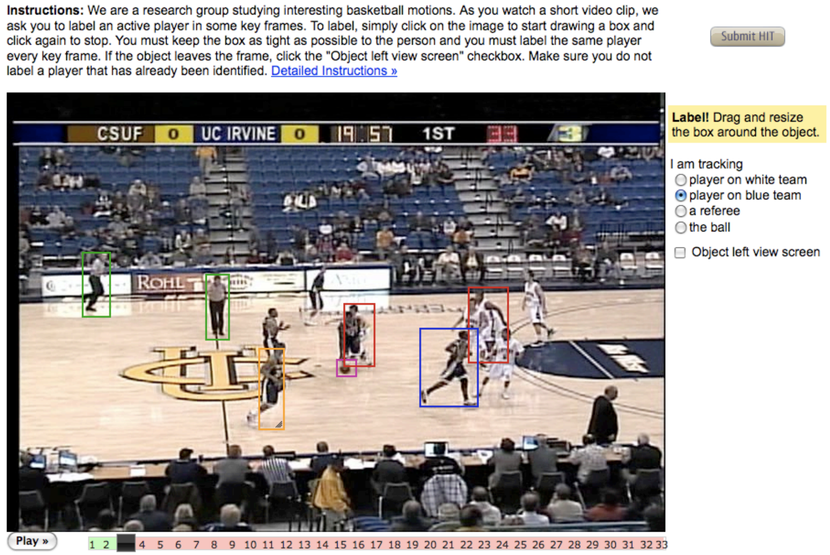
ADE20K



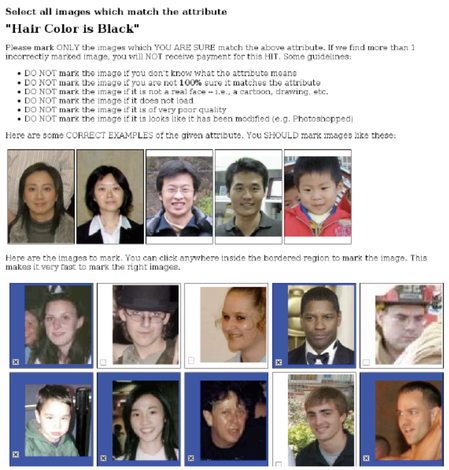
	Images	Obj. inst.	Obj. classes	Part inst.	Part classes	Obj. classes per image
COCO	123,287	886,284	91	0	0	3.5
ImageNet*	476,688	534,309	200	0	0	1.7
NYU Depth V2	1,449	34,064	894	0	0	14.1
Cityscapes	25,000	N/A	30	0	0	N/A
SUN	16,873	313,884	4,479	0	0	9.8
OpenSurfaces	22,214	71,460	160	0	0	N/A
PascalContext	10,103	∼104,398**	540	181,770	40	5.1
ADE20K	22,000	415,099	2,944	171,148	354	10.5

* has only bounding boxes (no pixel-level segmentation). Sparse annotations.
** PascalContext dataset does not have instance segmentation. In order to estimate the number of instances, we find connected components (having at least 150pixels) for each class label.

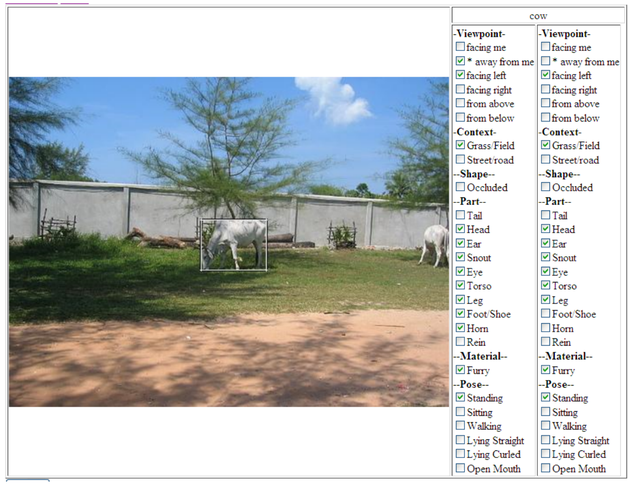
Sorokin, Forsyth, 2008



Carl Vondrick, Deva Ramanan, Don Patterson



N. Kumar, A. C. Berg,
P. N. Belhumeur, and S. K. Nayar, ICCV 2009



Farhadi Endres Hoiem Forsyth CVPR 2008

And many more…

So we can sometimes collect good training data.

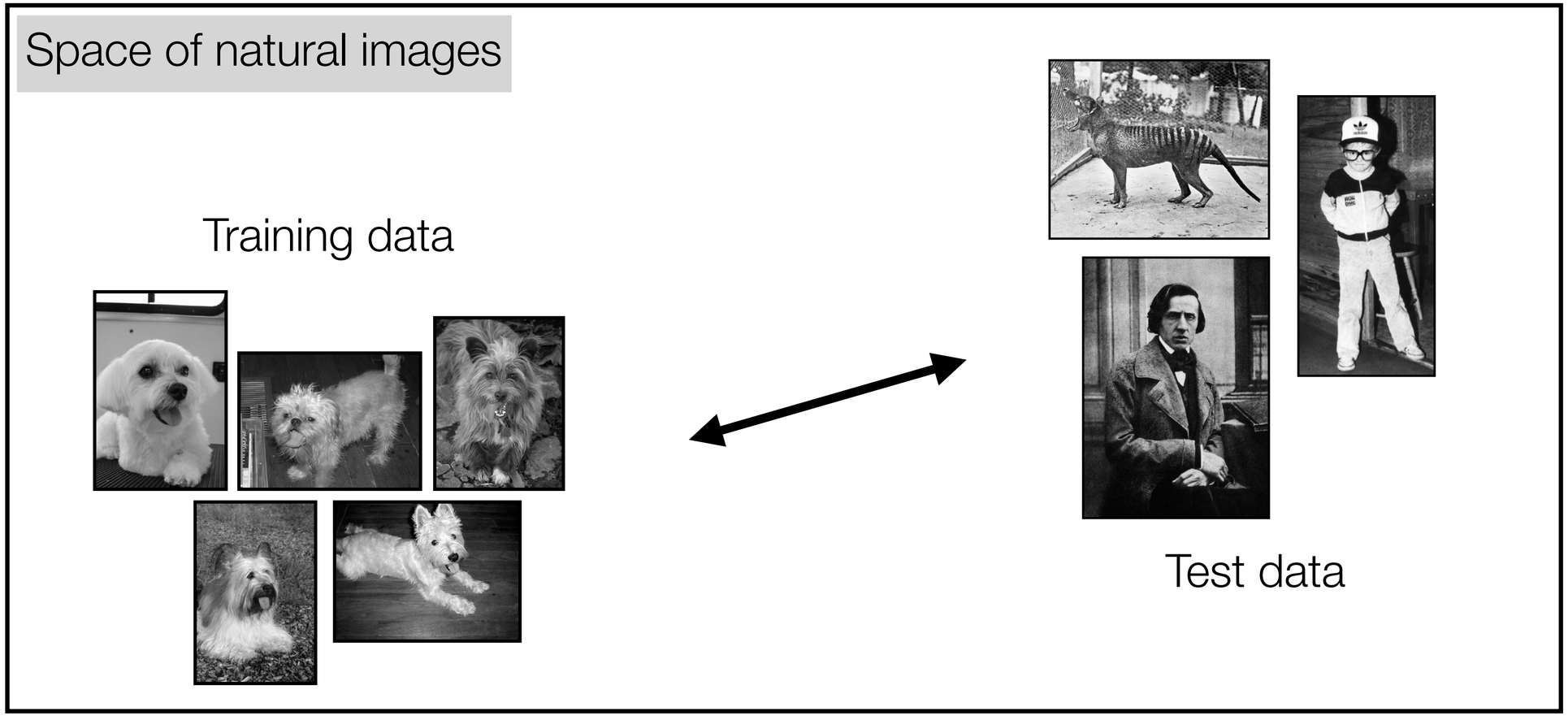
But suppose we messed up. Our test setting doesn’t look like the training data!

How can we bridge the domain gap?

training domain

testing domain
(where we actual use our model)

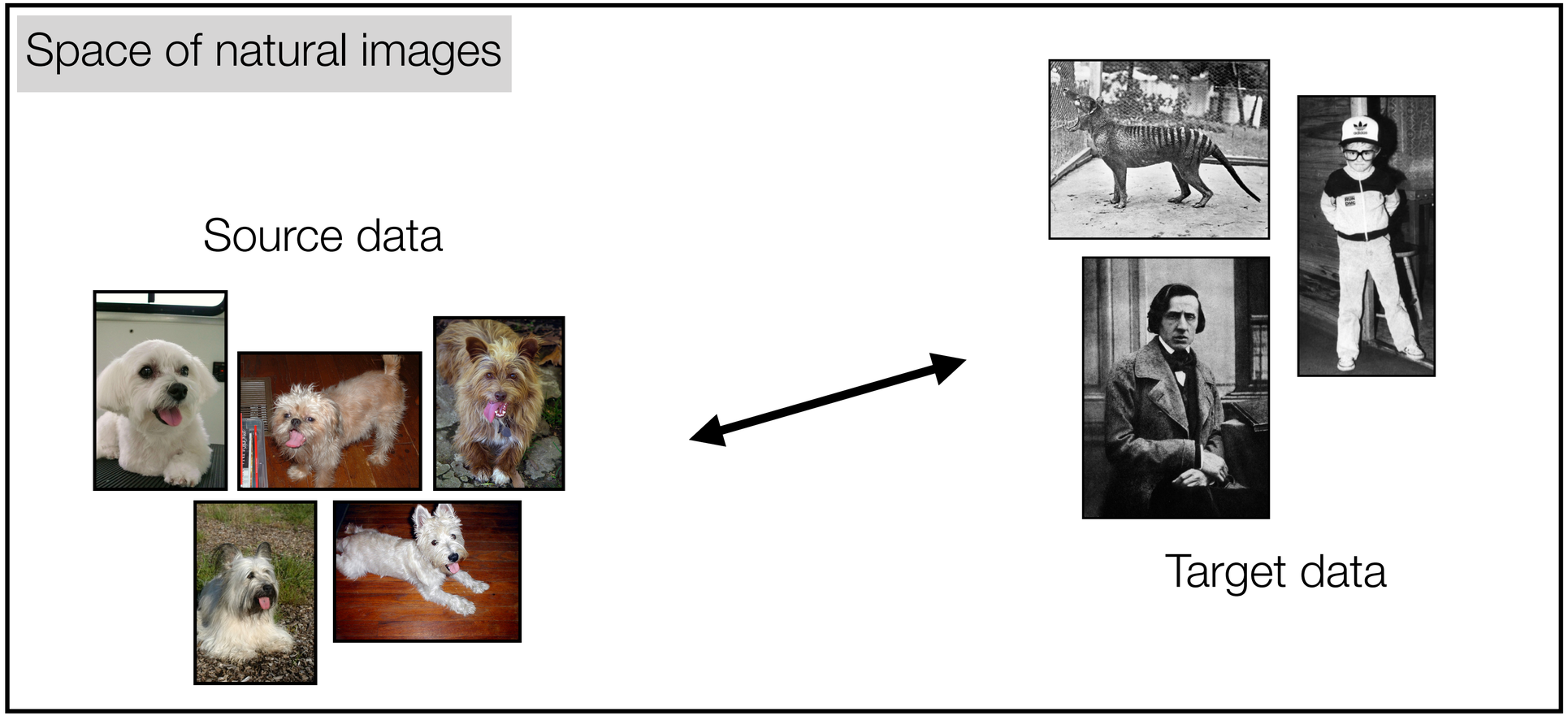
Domain gap between $p_{\texttt{train}}$ and $p_{\texttt{test}}$ will cause us to fail to generalize.



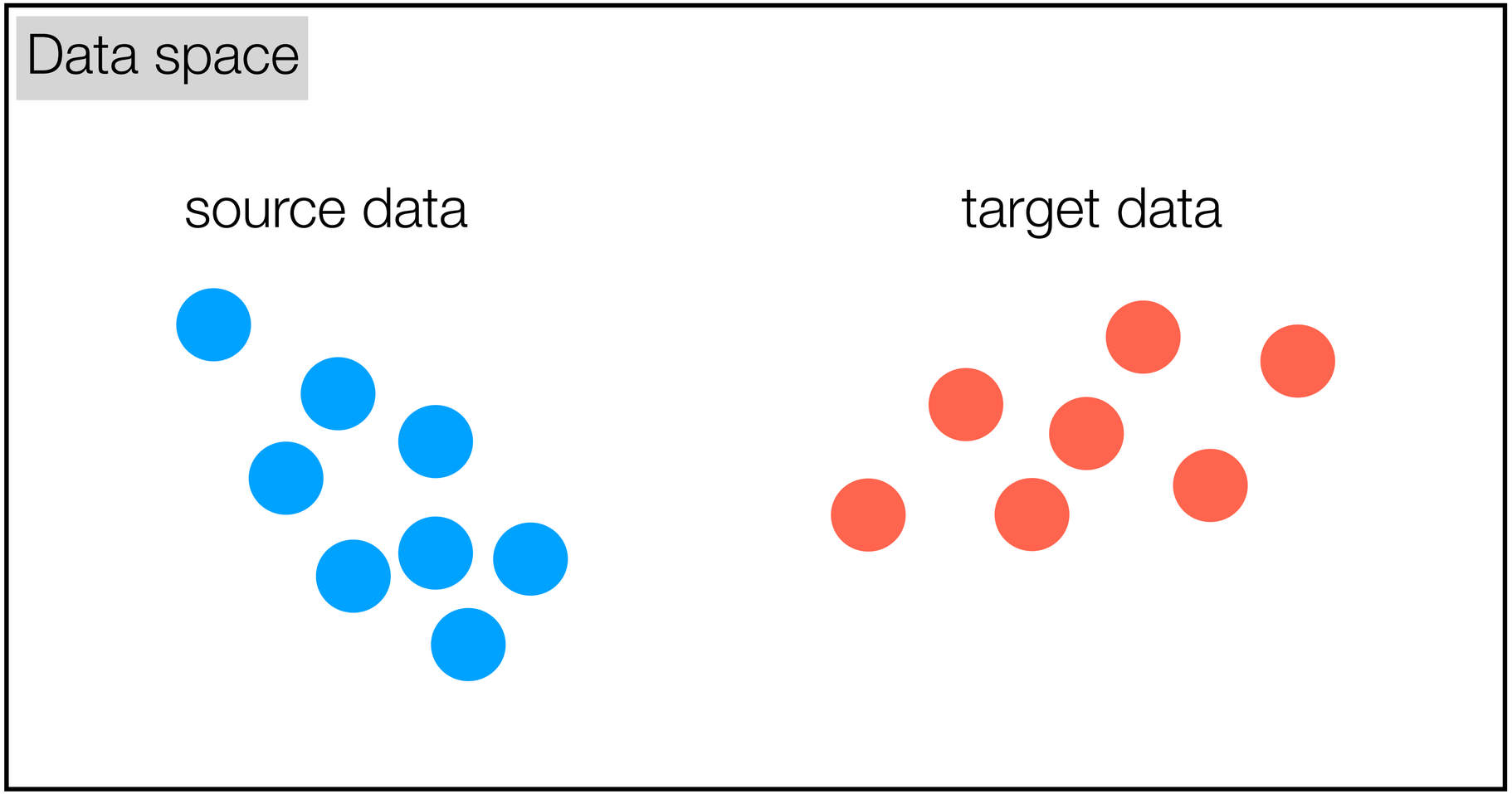
source domain

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.



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

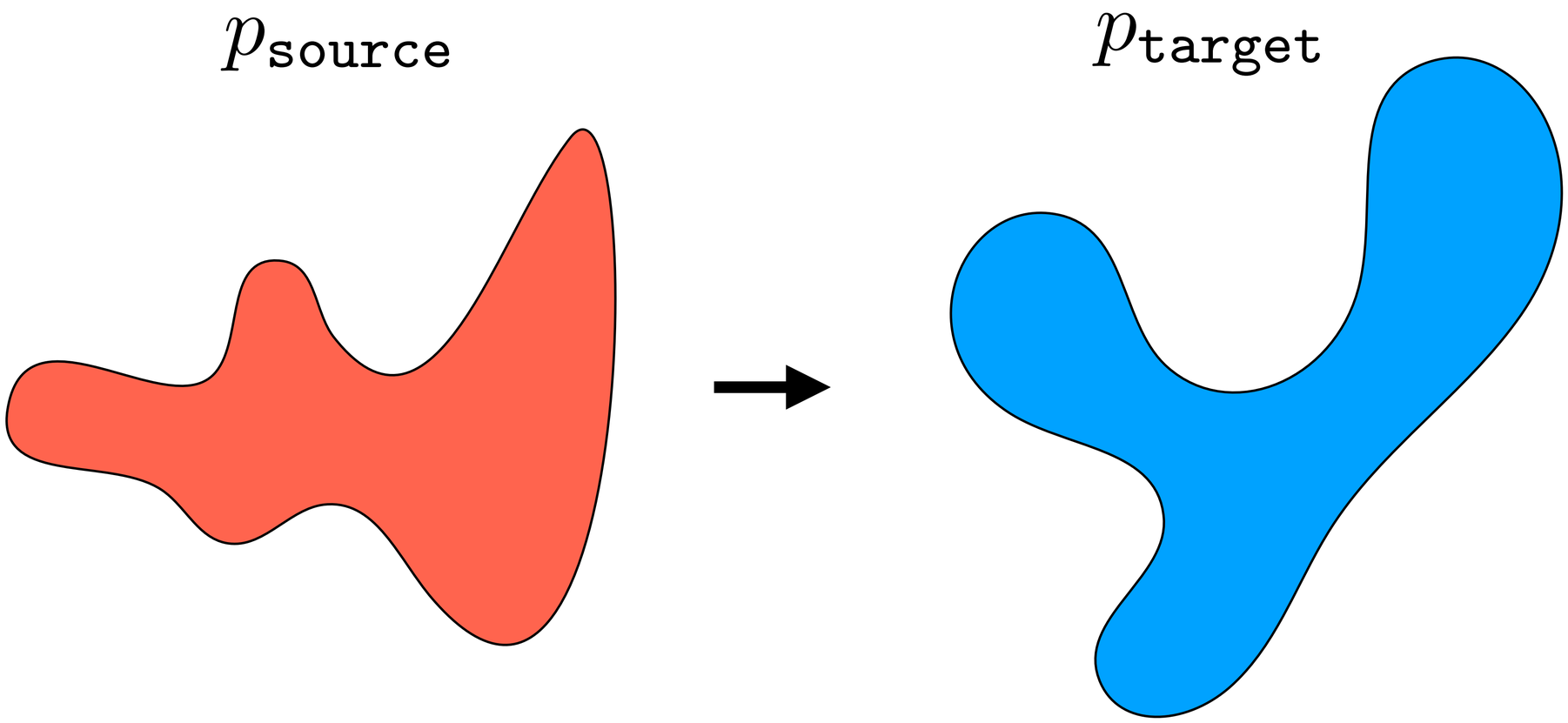


(Or vice versa)

This is called **domain adaptation**

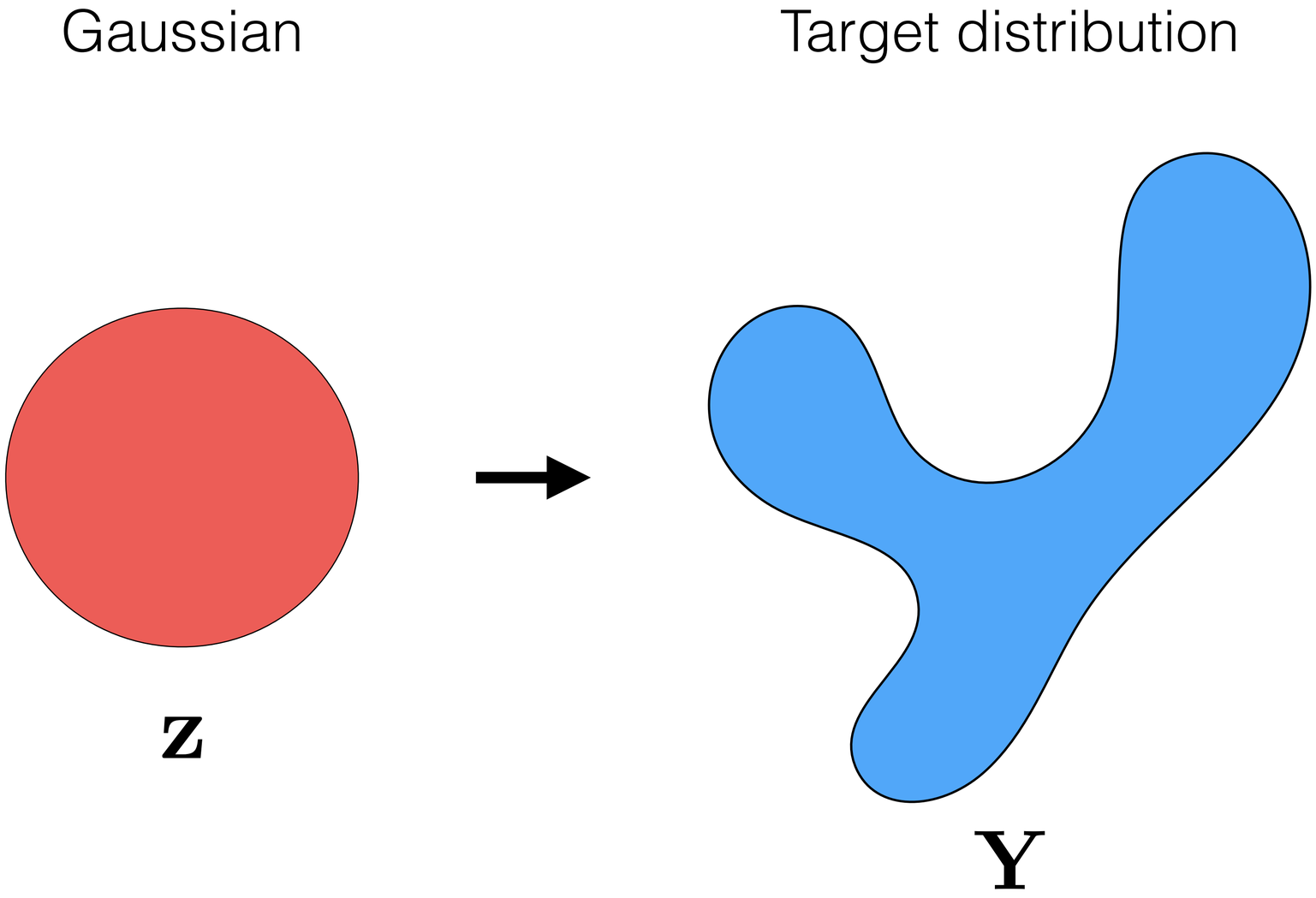
Domain adaptation

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

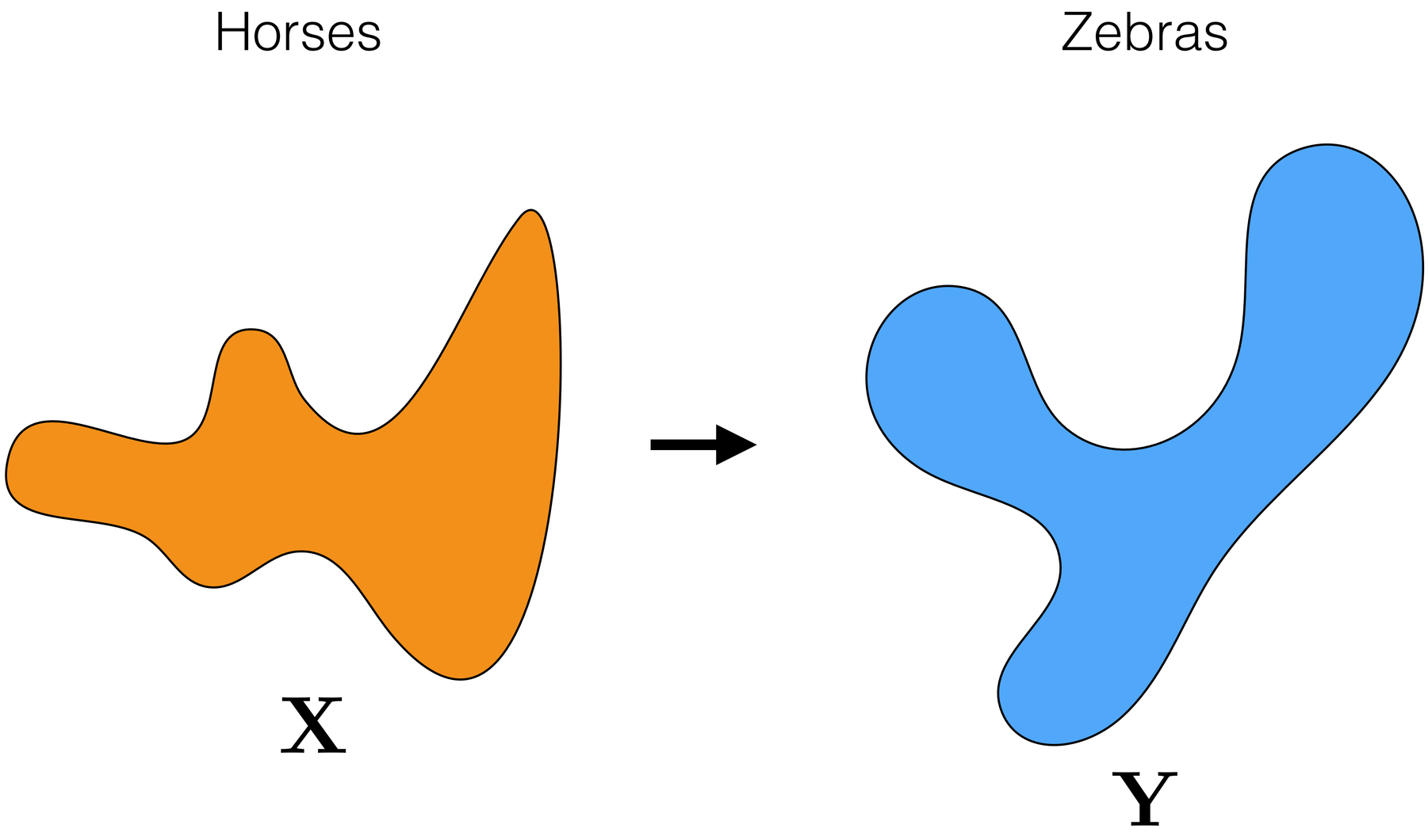


It's a just another distribution mapping problem!

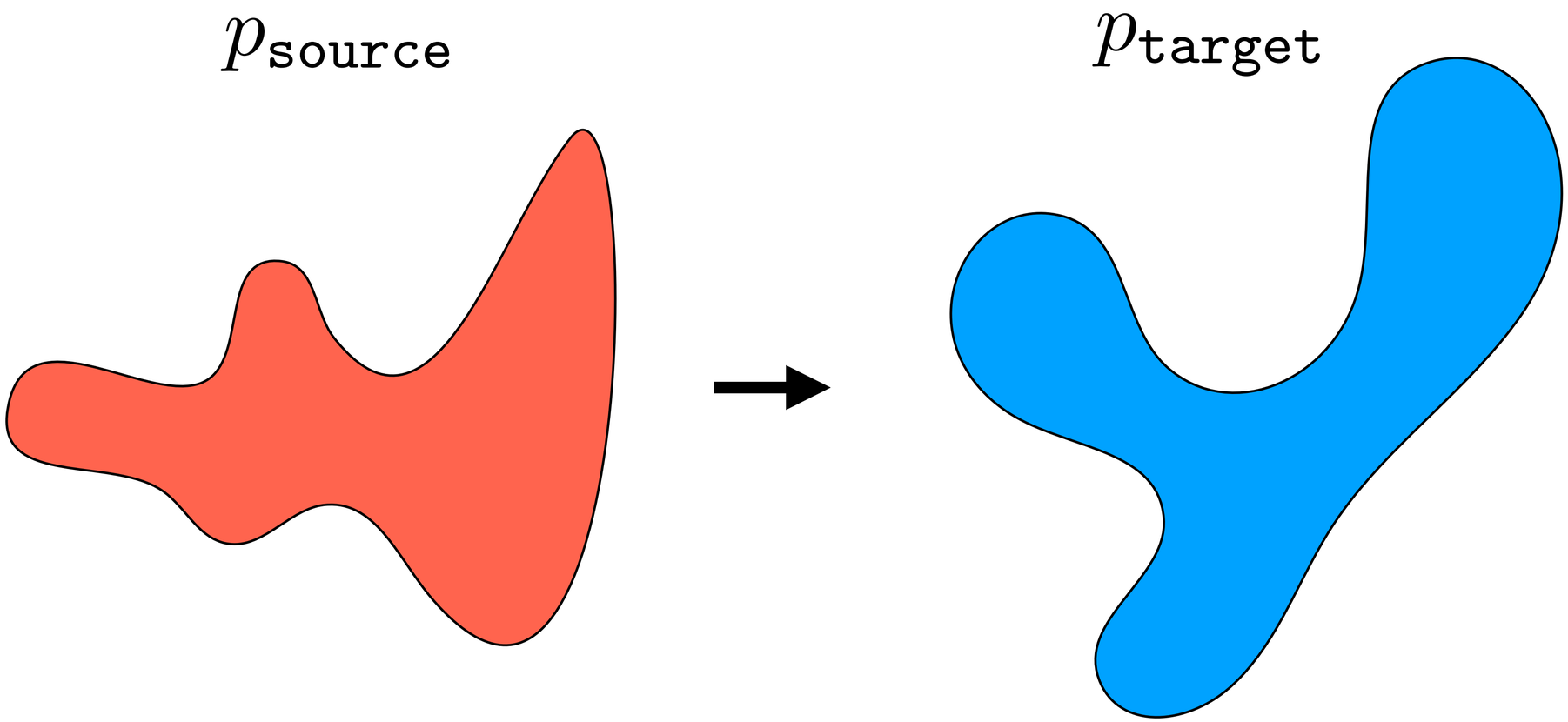
GANs



CycleGAN



Domain adaptation



Sim2real

Simulated data

,

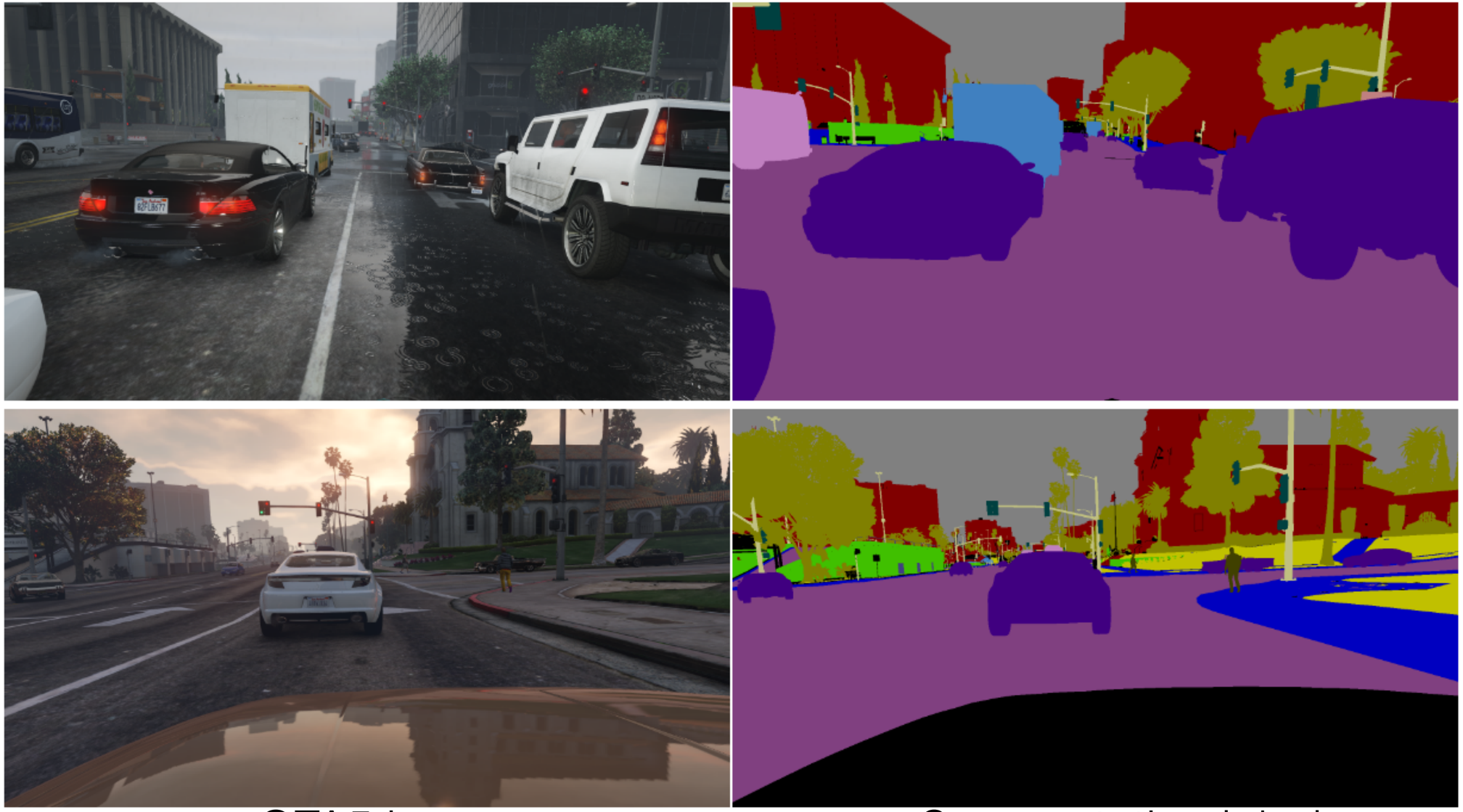
→

Real data

, ?

[Richter*, Vineet* et al. 2016] [Krähenbühl et al. 2018]

Synthetic Data as Supervision

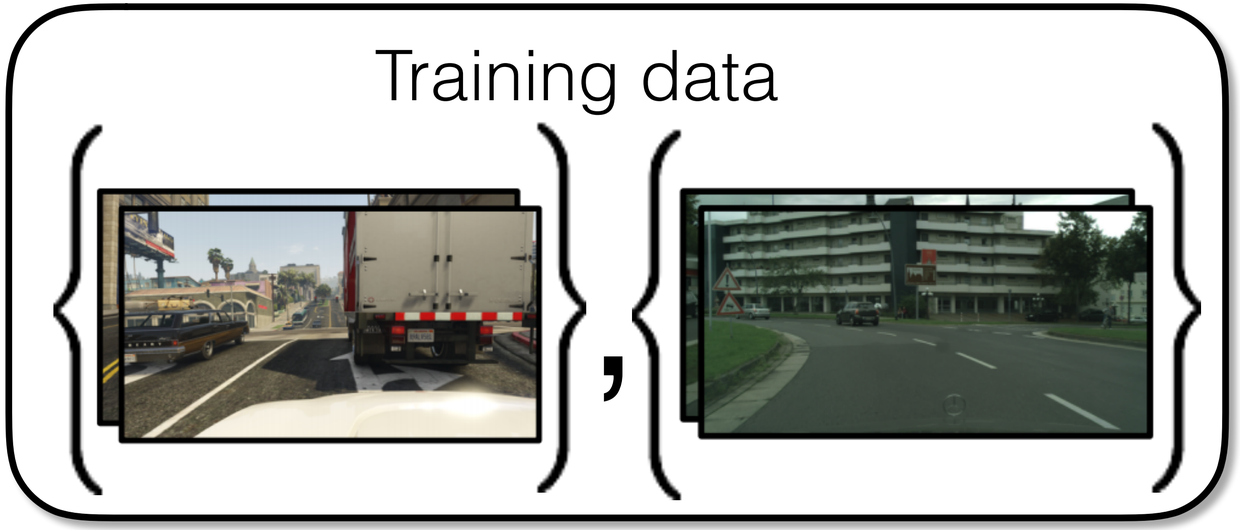


GTA5 images

Segmentation labels

[Richter*, Vineet* et al. 2016] [Krähenbühl et al. 2018]

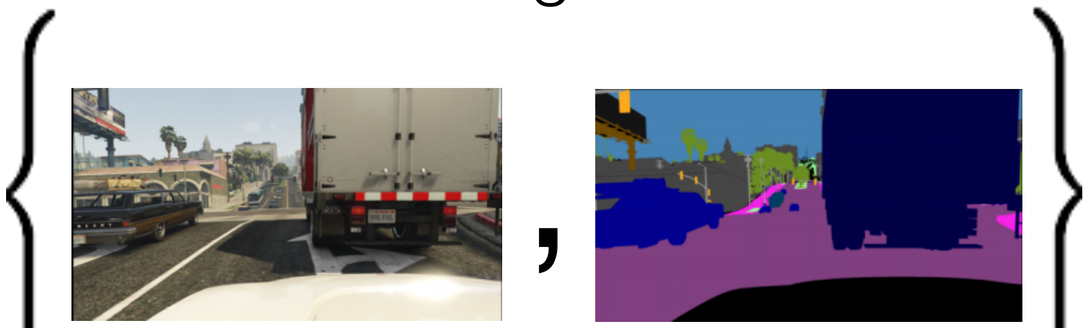
CycleGAN



[Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Darrell, Efros, 2018]

CycleGAN

Training data



[Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Darrell, Efros, 2018]

CycleGAN Semantic segmentation

Training data

[Hoffman, Tzeng, Park, Zhu, Isola, Saenko, Darrell, Efros, 2018]

GTA5 CG Input

Output

Cityscape Input

Output

Domain Adaptation with CycleGAN

Train on GTA5 data

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0

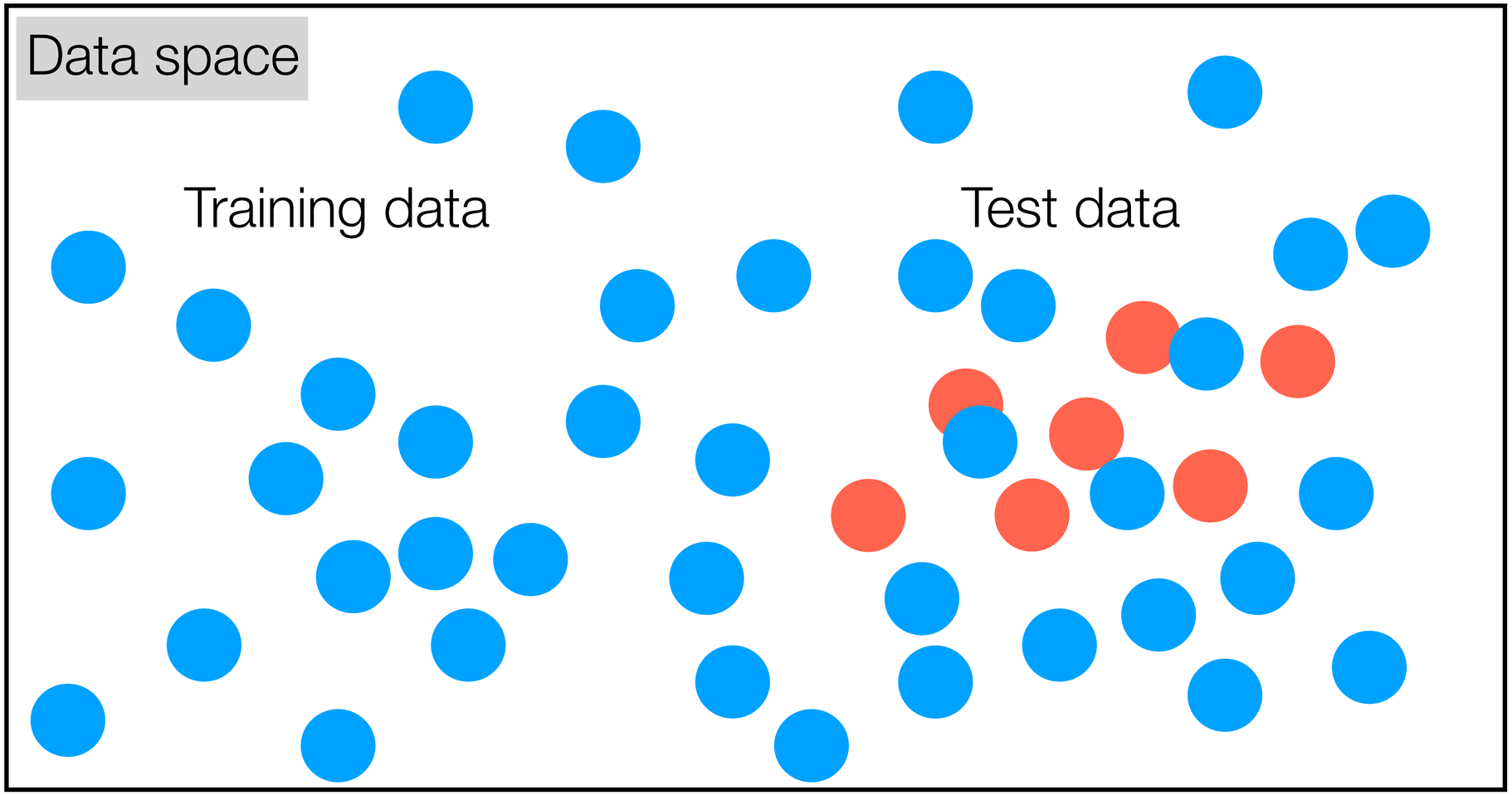
Domain Adaptation with CycleGAN

Train on CycleGAN data

Test on real images

	meanIOU	Per-pixel accuracy
Oracle (Train and test on Real)	60.3	93.1
Train on CG, test on Real	17.9	54.0
Train on CycleGAN, test on Real	**34.8**	**82.8**

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**

Data augmentation

Training data

$\mathbf{x}$ y

"Fish"

"Grizzly"

"Chameleon"

⋮

"Fish" Mirror

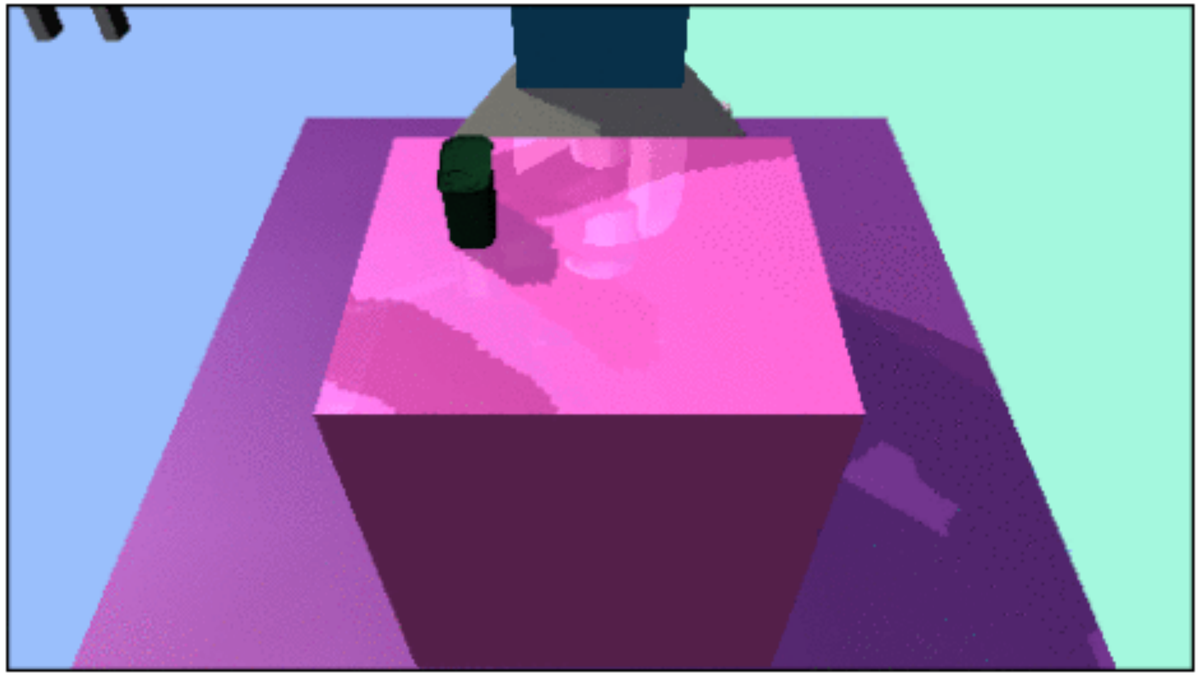
"Fish" Crop

"Fish" Crop

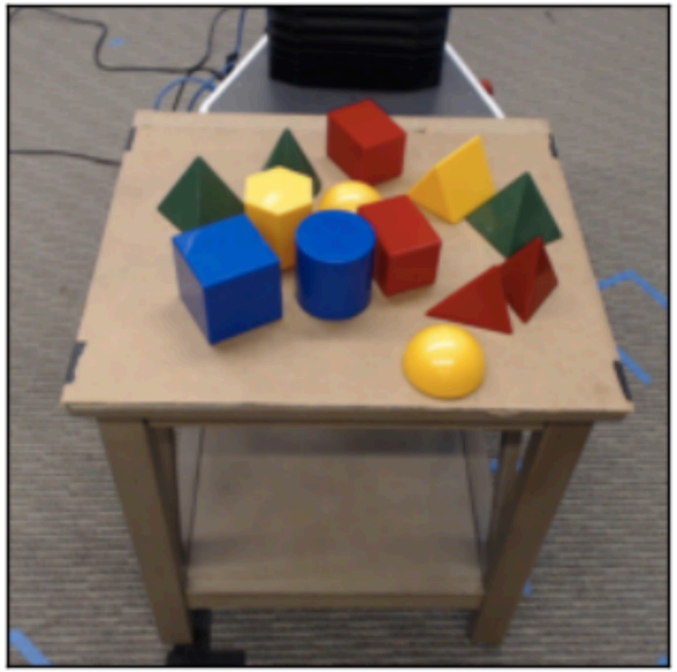
"Fish" Darken

Domain randomization

Training data



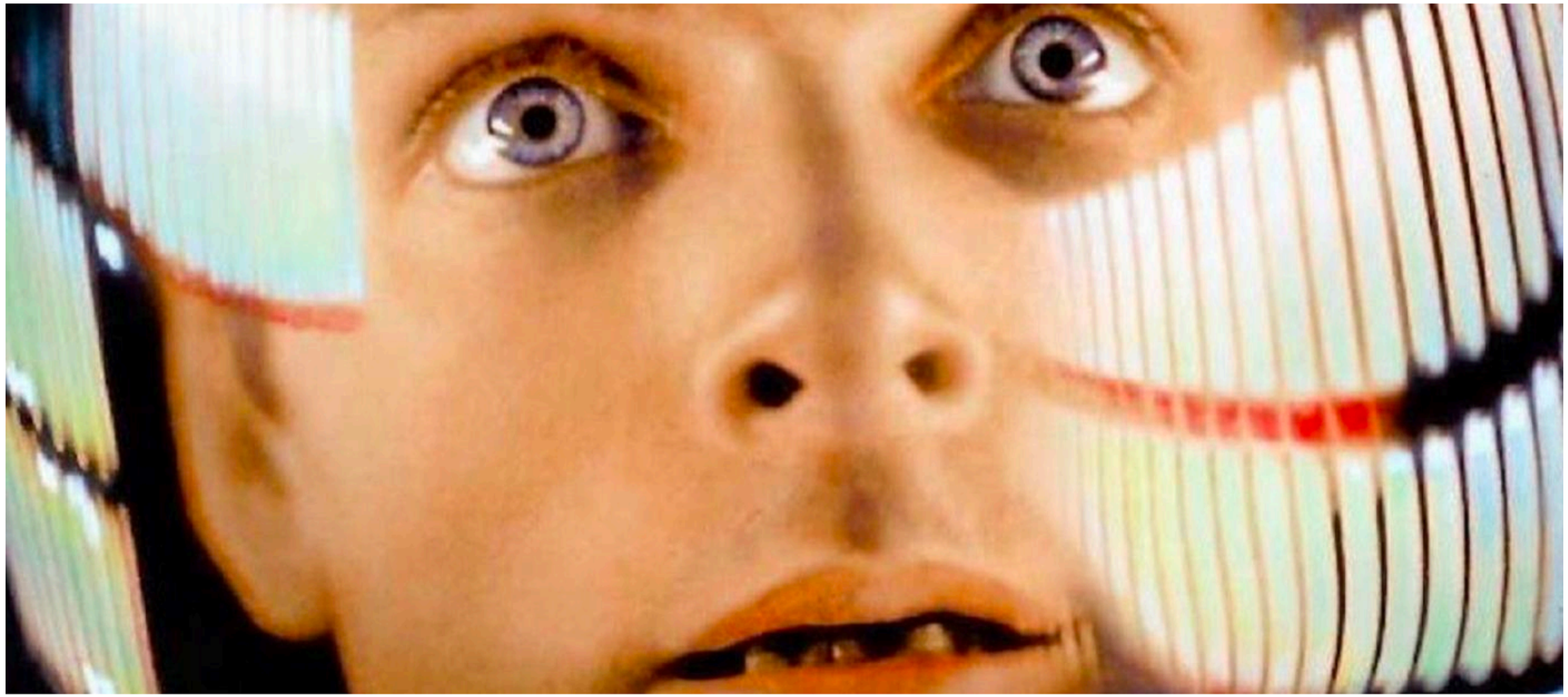
Test data



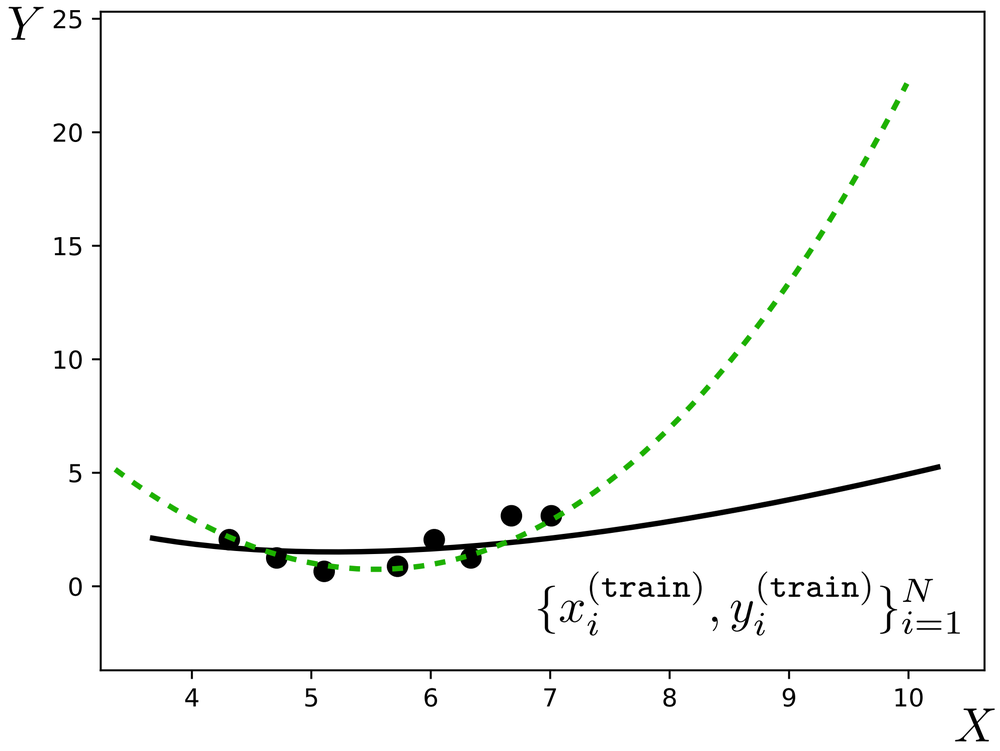
[Sadeghi & Levine 2016]

Above example is from [Tobin et al. 2017]

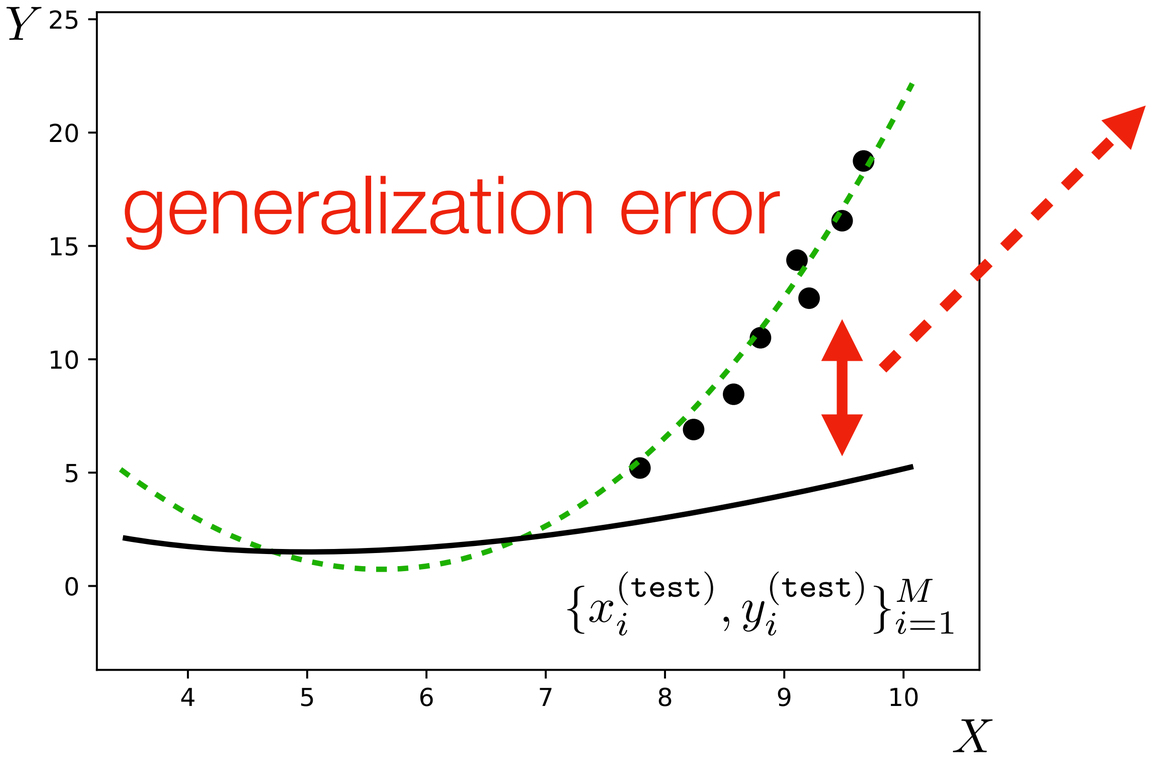
What if we go waaaay outside of the training distribution?



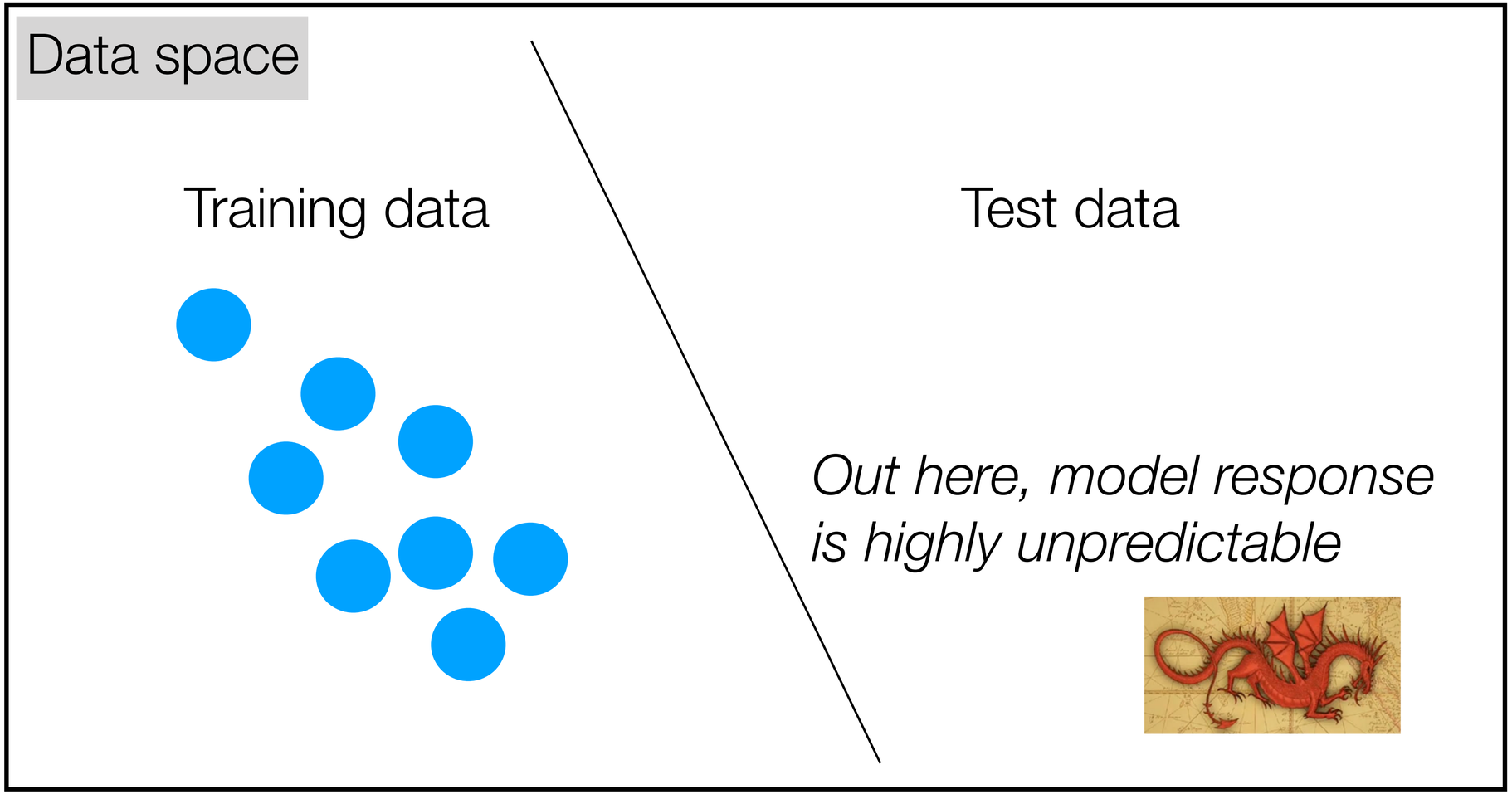
Training data



Test data

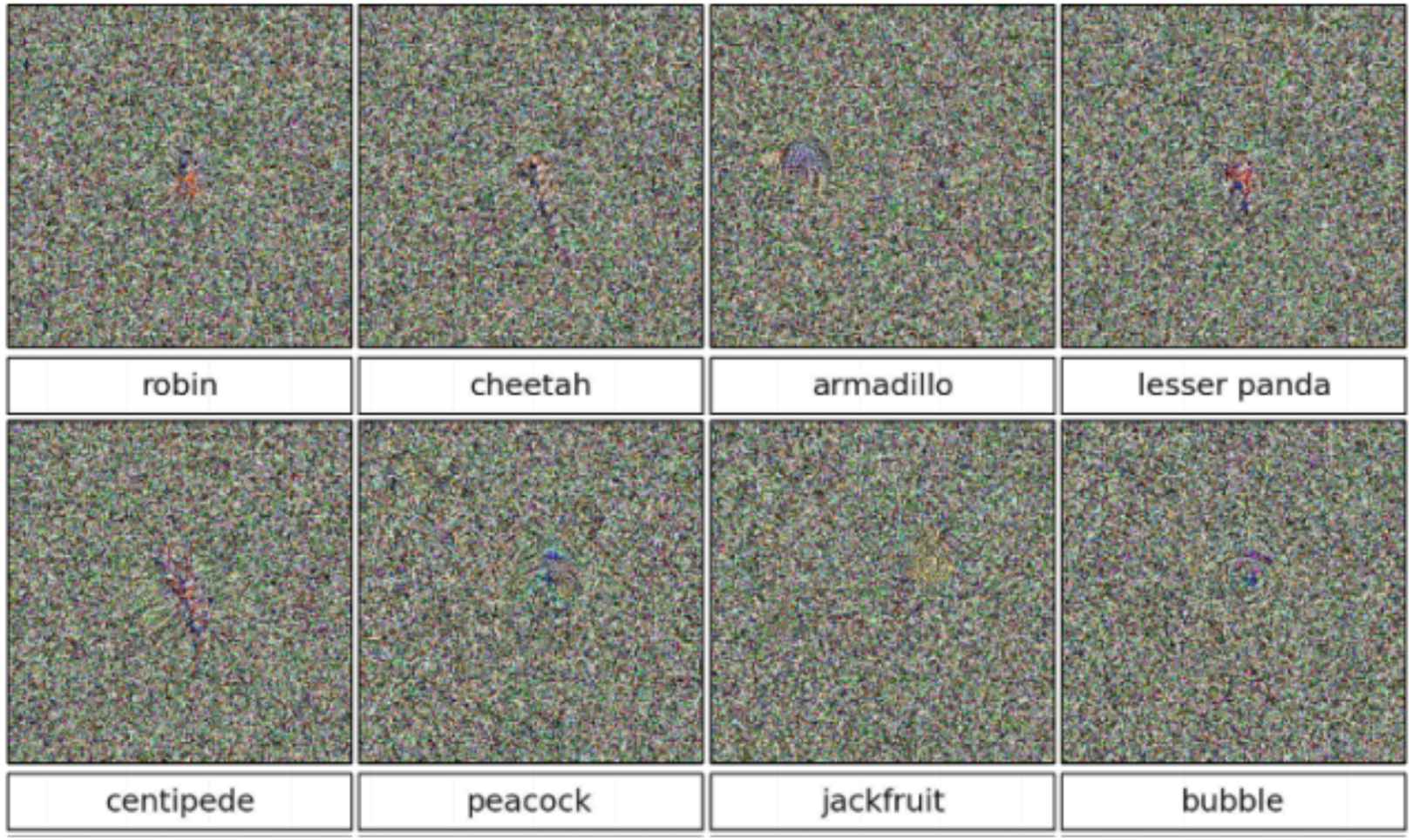


Our training data did cover the part of the distribution that was tested (**biased data**)


Data space
Training data
Test data
Out here, model response is highly unpredictable

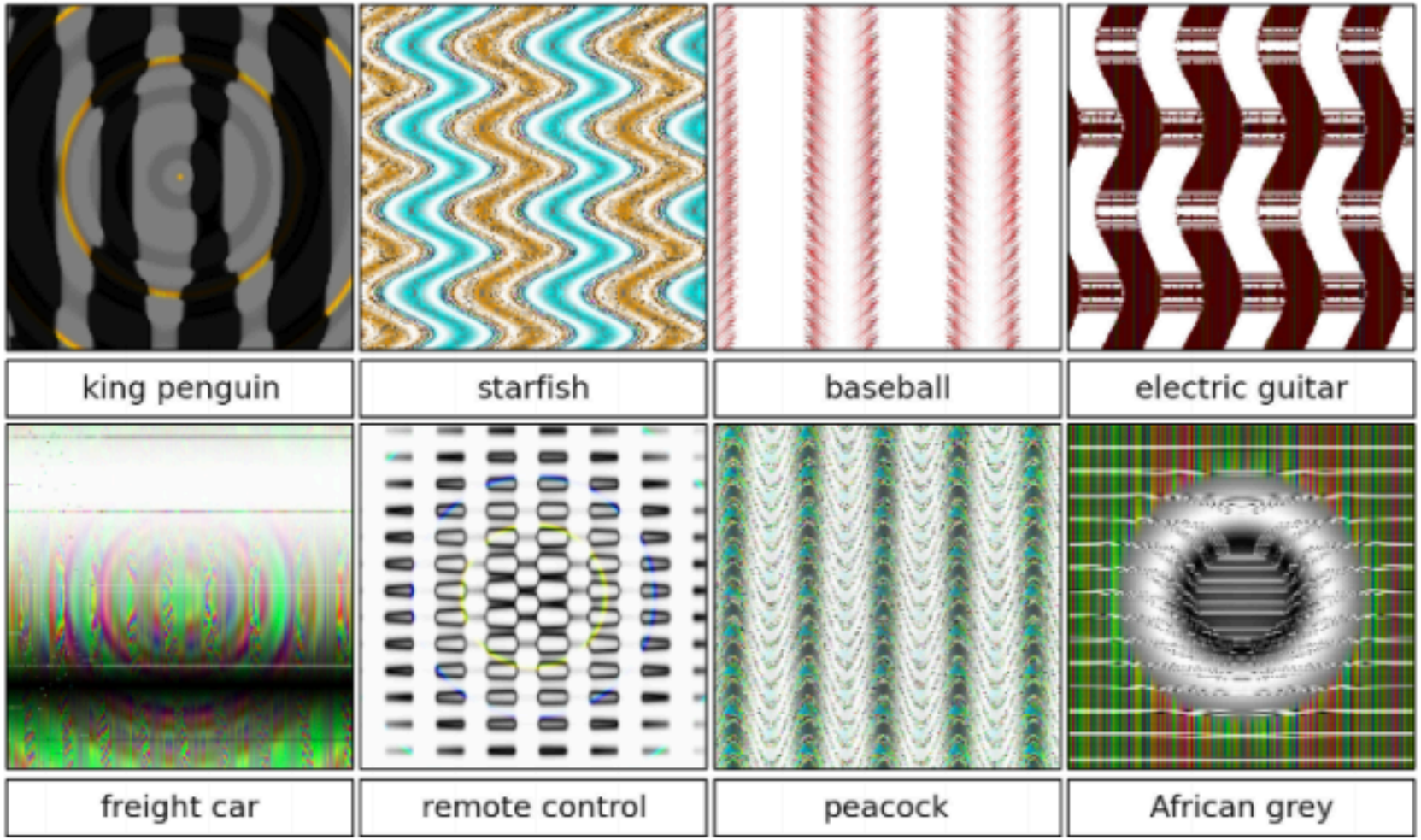
“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”

[Nguyen, Yosinski, and Clune, CVPR 2015]

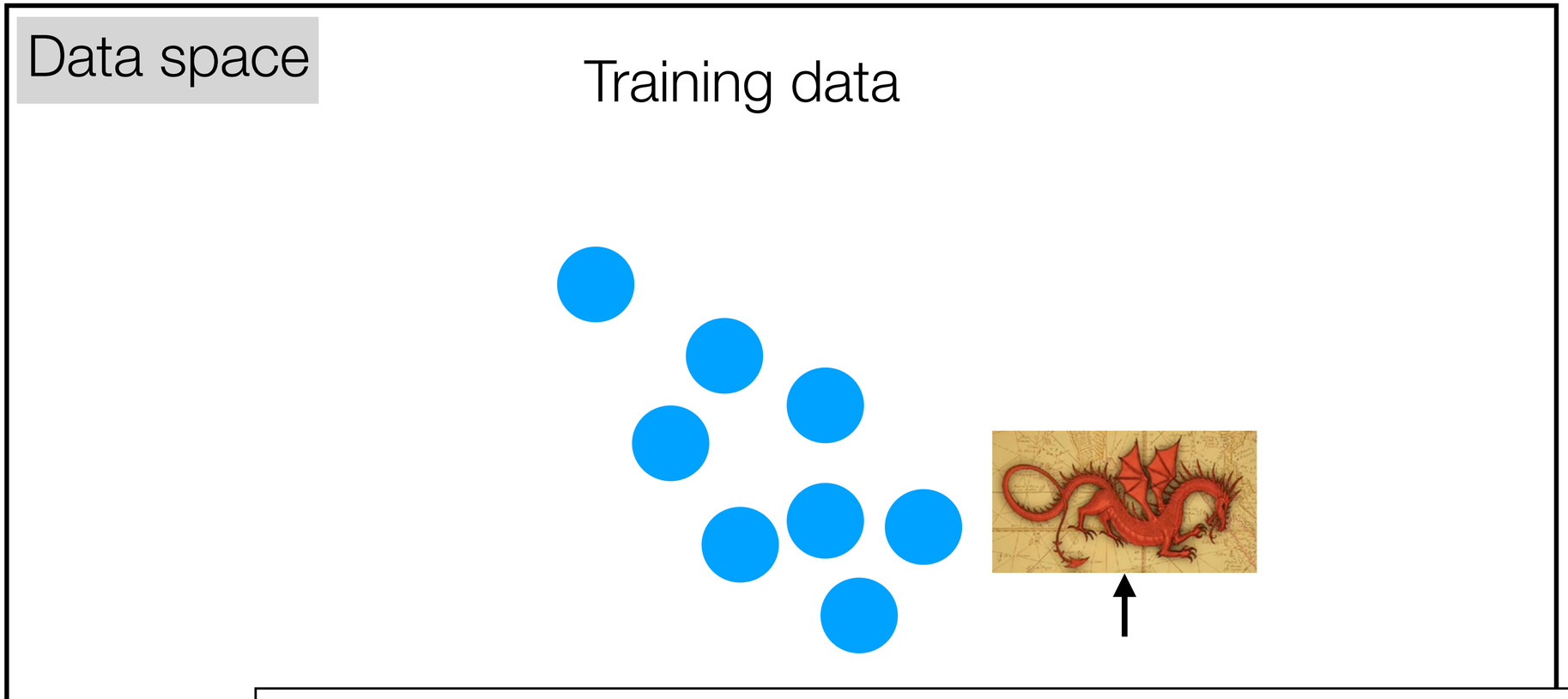


“Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images”

[Nguyen, Yosinski, and Clune, CVPR 2015]



Weirdness of high-dimensional space:

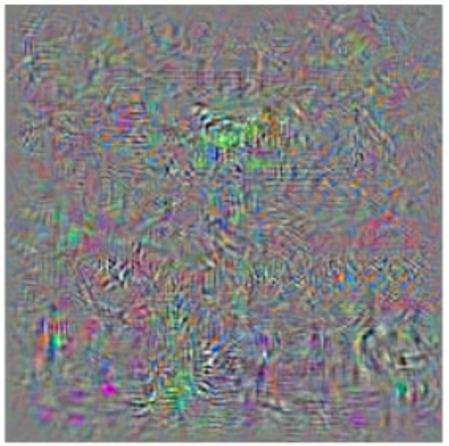


Usually, there are *blind spots* where the model has not fit the distribution well, and behaves unpredictably

Adversarial noise

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

$+$



$=$

$\Downarrow f$ $\qquad$ $\Downarrow f$

y $\qquad$ "School bus" $\qquad$ "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]

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!

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

Anything to worry about?

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

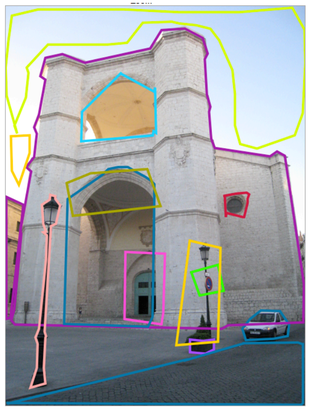
Towards ML You Can Rely On

Prof. Aleksander Madry
madry-lab.ml
6.883

Today: Symposium on Robust, Interpretable Deep Learning Systems
2:30-6:30pm, 46-3002

Anything else to worry about?

- Our datasets are often poorly labeled



- And usually biased (overrepresent certain categories)

- ML method perform beautifully on laboratory data, but often generalize poorly to real-world data

- Can have negative social consequences