

# Lecture 13 Distribution Shift

IEMS 402 Statistical Learning

Northwestern

# References

<https://hsnamkoong.github.io/assets/html/b9145/index.html>

# Distribution Shift

# Reconsider the ML Theory...

$E_p$

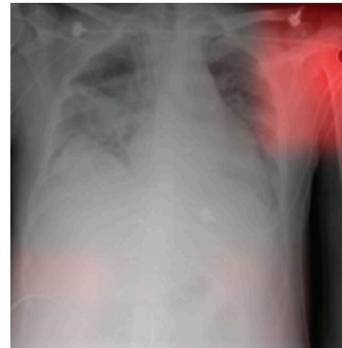
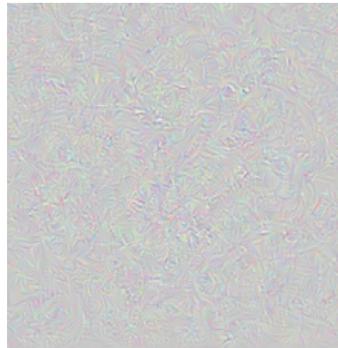
# However...



# Elephant or Cat



# Shortcut learning



Article: Super Bowl 50

Paragraph: "Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had a jersey number 37 in Champ Bowl XXXIV."

Question: "What is the name of the quarterback who was 38 in Super Bowl XXXIII?"

Original Prediction: John Elway

Prediction under adversary: Jeff Dean

Task for DNN	Caption image	Recognise object	Recognise pneumonia	Answer question
Problem	Describes green hillside as grazing sheep	Hallucinates teapot if certain patterns are present	Fails on scans from new hospitals	Changes answer if irrelevant information is added
Shortcut	Uses background to recognise primary object	Uses features irreducible to humans	Looks at hospital token, not lung	Only looks at last sentence and ignores context

# spurious correlation

## Waterbirds

y: waterbird  
a: water background



y: landbird  
a: land background



## Test examples

y: waterbird  
a: land background

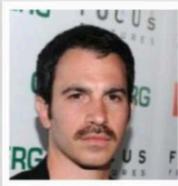


## CelebA

y: blond hair  
a: female



y: dark hair  
a: male



y: blond hair  
a: male



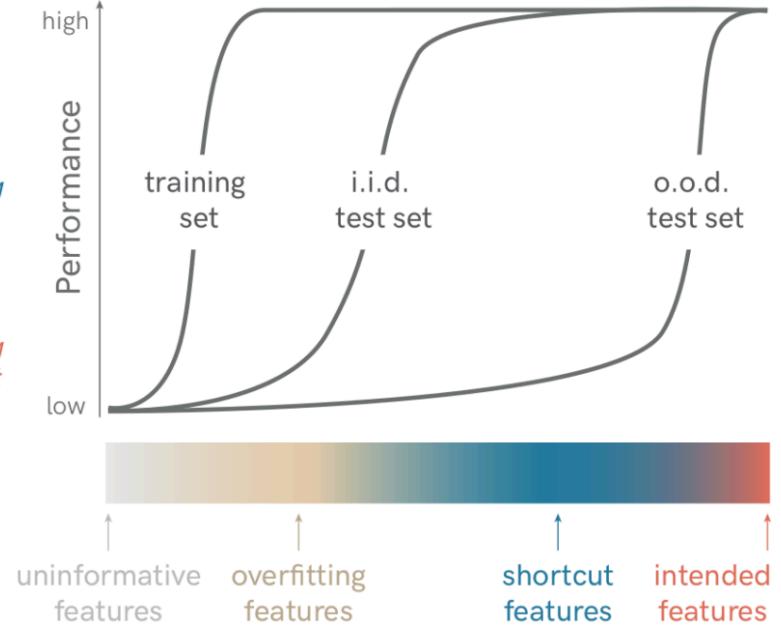
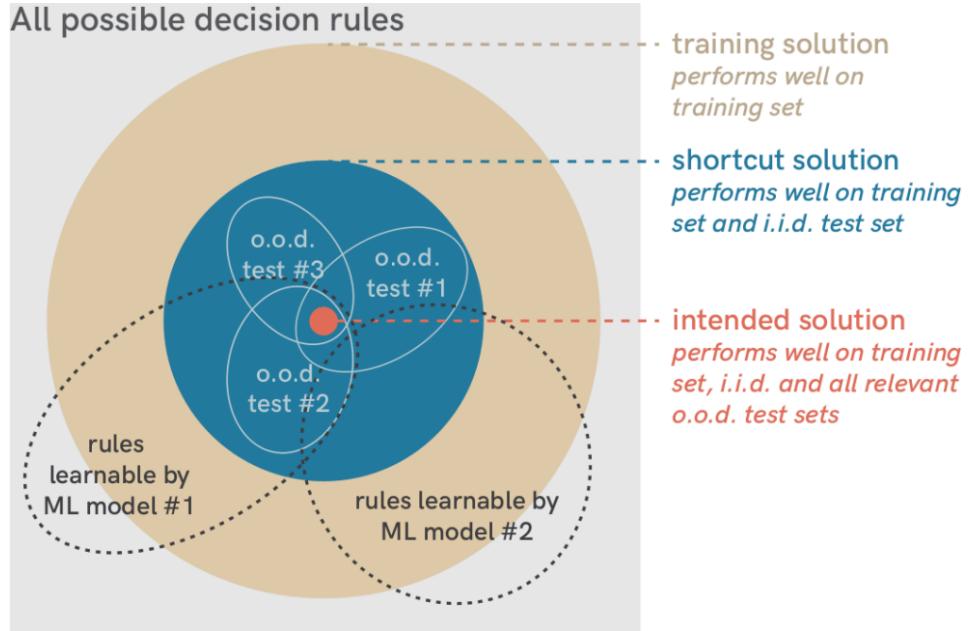
## MultiNLI

y: contradiction  
a: has negation  
(P) The economy could be still better.  
(H) The economy has never been better.

y: entailment  
a: no negation  
(P) Read for Slate's take on Jackson's findings.  
(H) Slate had an opinion on Jackson's findings.

y: entailment  
a: has negation  
(P) There was silence for a moment.  
(H) There was a short period of time where no one spoke.

# From i.i.d to o.o.d



# Importance Weighting

# Importance Weighting

How do we deal with covariate / label shifts?

**What we have**

$$E_{p_{\text{train}}}[\ell(z; \theta)]$$

Most basic approach: reweight the loss

$$E_{p_{\text{train}}}\left[\frac{p_{\text{test}}(z)}{p_{\text{train}}(z)} \ell(z; \theta)\right] = E_{p_{\text{test}}}[\ell(z; \theta)]$$

Weighted loss over the  
training distribution

**What we want**

$$E_{p_{\text{test}}}[\ell(z; \theta)]$$

① ~~①~~ ~~reweighting~~ ~~data 1~~ ~~data 2~~  
 $\text{loss}(\text{data 1}) + \text{loss}(\text{data 2})$   
resample

(also possible: resample the dataset)

② ~~②~~ ~~data 1~~ ~~data 1~~ ... ~~data 1~~ ~~data 2~~  
in expectation, they are same  
but these values are different.

# Importance weighting

I don't know  $\frac{p_{test}}{p_{train}}$

An alternative algorithm: use a classifier that separates  $p_{train}$  and  $p_{test}$

1. Estimate a classifier  $f(x) \approx \frac{p_{train}(x)}{p_{test}(x) + p_{train}(x)}$

( collect another dataset ·  
(  $x_{train}, 1$  ) (  $x_{test}, 0$  ) )

2. Reweight by  $h(x) = \frac{1}{f(x)} - 1$

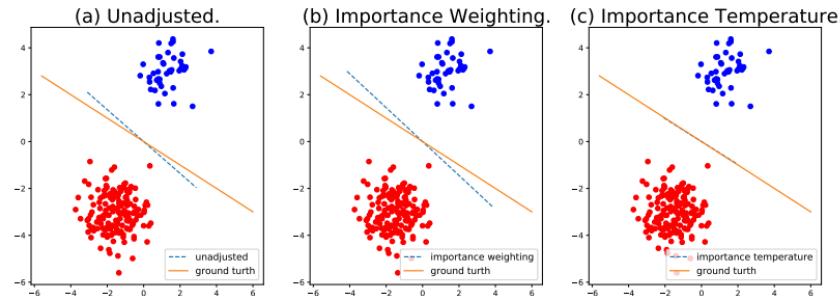
3. Fit a model by minimizing the loss  $h(x)\ell(x, y; \theta)$

Discriminative Learning for Differing  
Training and Test Distributions

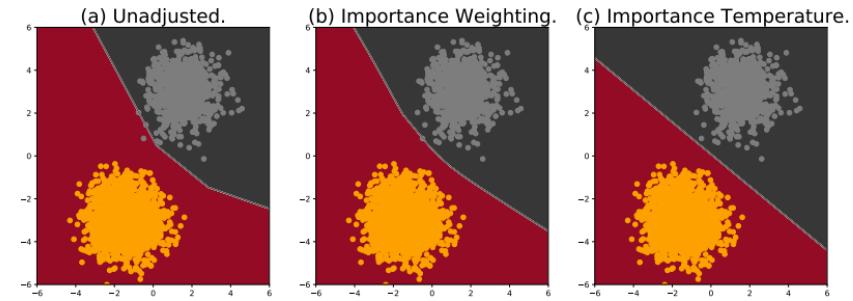
Steffen Bickel  
Michael Brückner  
Tobias Scheffer  
Max Planck Institute for Computer Science, Saarbrücken, Germany

BICKEL@MPI-INF.MPG.DE  
BRÜCKNER@MPI-INF.MPG.DE  
SCHEFFER@MPI-INF.MPG.DE

# Not Working for Over-parameterized Model



(a) Linear Model for Separable Data



(b) Multilayer Perceptron with two hidden layers of size 200

Byrd J, Lipton Z. What is the effect of importance weighting in deep learning? International conference on machine learning. PMLR, 2019: 872-881.



**IPM**

# Background material: integral probability measures

$$E_{P_{\text{test}}} l - E_{P_{\text{train}}} l$$

To state this clearly, we need to first go into some background.

## Definition (IPM):

For two probability distributions  $p$  and  $q$ , the integral probability metric (IPM) for a family of functions  $\mathcal{F}$  is defined as

$$d_{\mathcal{F}}(p, q) = \sup_{f \in \mathcal{F}} |E_p[f(x)] - E_q[f(x)]|$$

*distance between distribution  $P_{\text{train}} / P_{\text{test}}$*

**Intuition:**  $\mathcal{F}$  are ‘test functions’ that can distinguish  $p$  and  $q$

If two have the same function value for all  $\mathcal{F}$ , then they are similar

# IPM and distribution shift

What we want

What we have

Domain distance

$$E_{p_{test}}[\ell(x, y, \theta)] = E_{p_{train}}[\ell(x, y, \theta)] + \Delta$$

From the trivial restatement

$$\Delta = E_{p_{test}}[\ell(x, y, \theta)] - E_{p_{train}}[\ell(x, y, \theta)]$$

$$\bar{E}_{P_{test}} \ell \leq \bar{E}_{P_{train}} \ell + d_F(P_{train}, P_{test})$$

This looks like an IPM! (if  $\ell(x, y, \theta) \in \mathcal{F}$  for all  $\theta$ )

$$\Delta \leq \sup_{f \in \mathcal{F}} E_{p_{test}}[f(x, y)] - E_{p_{train}}[f(x, y)] = d_{\mathcal{F}}(p_{train}, p_{test})$$

**Takeaway:** IPMs bound excess error under transfer

# Example: L1 distance

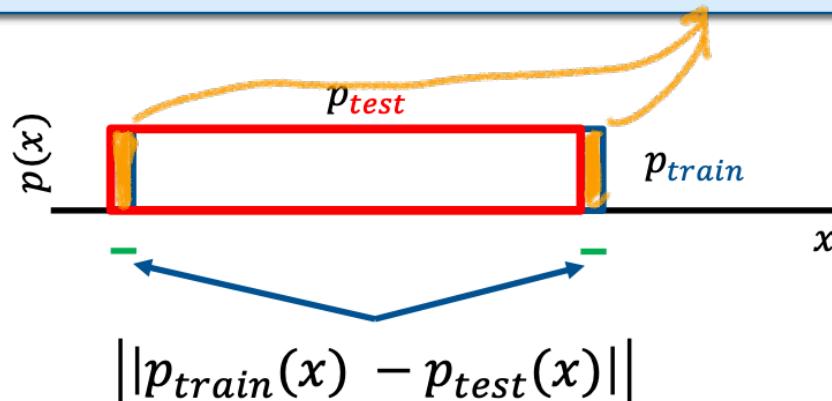
$$d_F(P, Q) = \sup_{\{f \mid -1 \leq f \leq 1\}} | \mathbb{E}_{P(x)} [f(x)] - \mathbb{E}_{Q(x)} [f(x)] | = \sum_x |P(x) - Q(x)|$$

We can now bound test performance in terms of IPMs

$$\mathcal{F} := \{f \mid -1 \leq f(x) \leq 1, \forall x\}, d_F(P, Q) = \sum_x |P(x) - Q(x)|$$

For  $0 \leq \ell(x, y, \theta) \leq 1$  and under covariate shift,

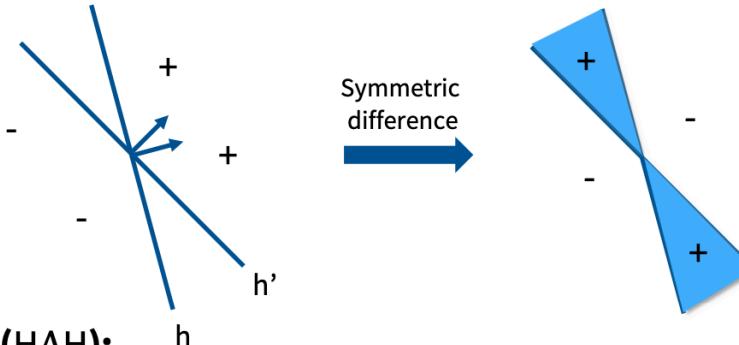
$$E_{p_{test}}[\ell(x, y, \theta)] \leq E_{p_{train}}[\ell(x, y, \theta)] + \|p_{train}(x) - p_{test}(x)\|_1$$



	Reweighting	IPM
Goals	Correct train-test mismatch	Estimate train-test mismatch <i>Can also correct</i>
Assumptions	Overlap $\frac{P_{test}}{P_{train}}$ $\Rightarrow$ if $P_{train}(x) = 0$ then $P_{test}(x) = 0$	Boundedness $d_F(P, P)$
Training	Weighted/modified loss	No change <i>You may use <math>E_{\text{prior}} + d_F(P, P)</math> as loss</i>
Costs	More samples (variance)	Inaccurate models (bias) Curse of dimensionality (next lecture)

## Defining $H\Delta H$ (disagreement)

For a hypothesis class  $\mathcal{H}$ , the  $H\Delta H$  set is defined as the symmetric difference



**Definition ( $H\Delta H$ ):**

For a hypothesis class  $\mathcal{H}$ , the symmetric difference set  $H\Delta H$  is defined as

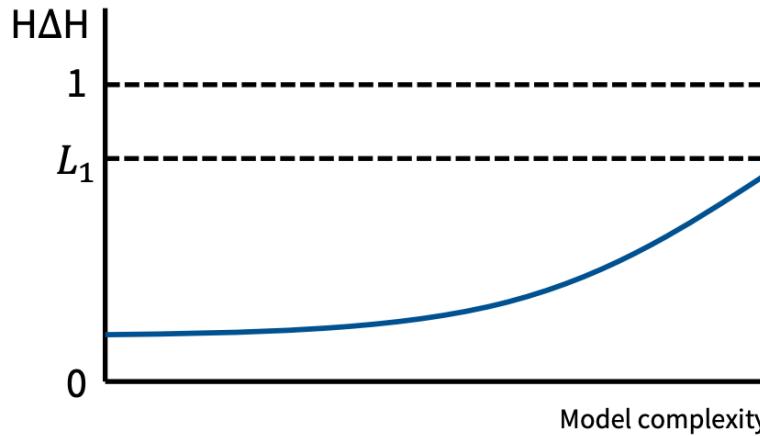
$$H\Delta H := \{g: g(x) = \text{XOR}(h(x), h'(x)) \text{ and } h, h' \in \mathcal{H}\}$$

# Dependency on Hypothesis Space

$$\text{H}\Delta\text{H}: \frac{1}{2} d_{H\Delta H}(p_{train}, p_{test})$$

For a hypothesis class  $\mathcal{H}$ , the H $\Delta$ H-divergence is

$$d_{H\Delta H}(p_{train}, p_{test}) = 2 \sup_{g \in \mathcal{H}} |E_{p_{train}}[g(x)] - E_{p_{test}}[g(x)]|$$



$d_{H\Delta H}$  is upper bounded by the  $L_1$  distance

$d_{H\Delta H}$  increases monotonically with model complexity. If  $H \subset H'$ ,  
 $d_{H\Delta H} \leq d_{H'\Delta H'}$

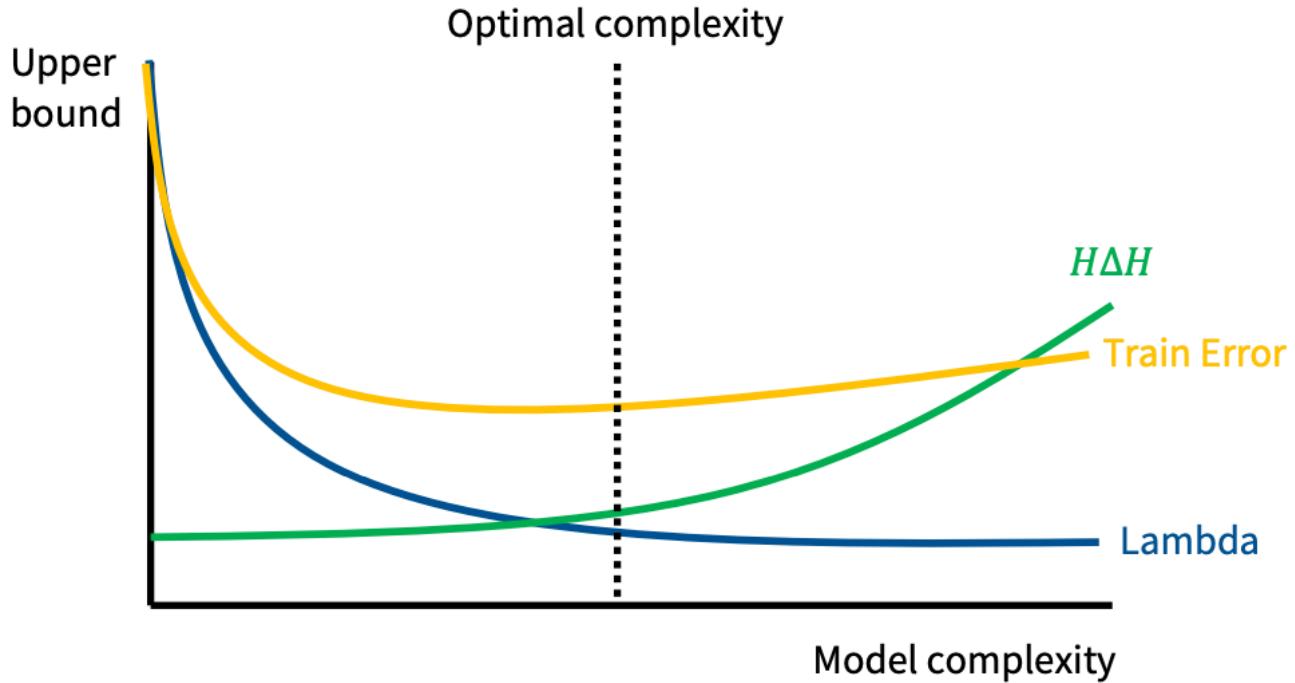
# Another trade-off

Let's walk through the main bound.

$$\begin{aligned} & E_{p_{test}}[\ell(x, y, h)] \quad \text{Different answer on two domains} \\ & \leq E_{p_{train}}[\ell(x, y, h)] + \frac{1}{2}d_{H\Delta H}(p_{train}, p_{test}) + \lambda \quad \text{same answer but} \\ & \quad \quad \quad \text{Both are wrong} \\ & \quad \quad \quad \uparrow \\ \text{Training domain error} & \quad \quad \quad \text{Domain distinguishability} \\ & \quad \quad \quad \uparrow \\ & \quad \quad \quad \text{Minimal error of a classifier on both domains} \\ & \lambda = \inf_{h \in \mathcal{H}} p_{train}(y \neq h(x)) + p_{test}(y \neq h(x)) \end{aligned}$$

**HΔH claim:** Low training domain error + low  $H\Delta H$  divergence + rich  $\mathcal{H}$   
= good generalization to target domain

# Another tradeoff



# Distributionally Robust Optimization

# F-divergence

f-divergence

$$D_f(Q \parallel P) := \int f\left(\frac{dQ}{dP}\right) dP$$

$f(r) = 0$   
 $\rightarrow f$  is a convex function

$f(t) = t \log t$ , then KL-divergence

$f(t) = |t - 1|$   $\rightarrow$  L divergence  
(Total Variation)

$$f(t) = (t - 1)^2$$

$\chi^2$  divergen

# Distributionally Robust Optimization

Empirical Risk  
Minimization

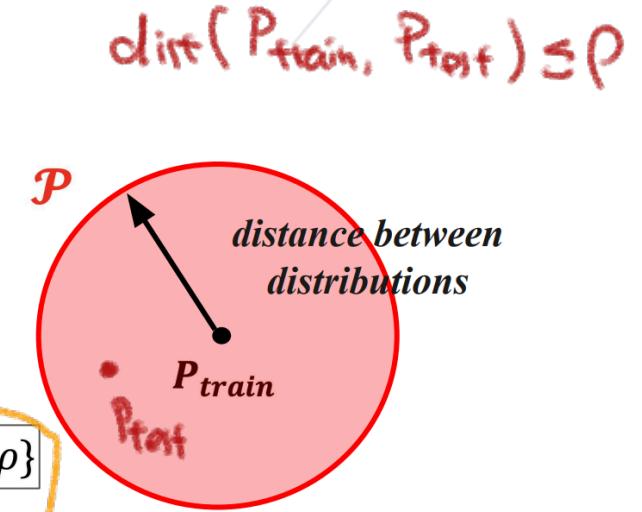
$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}} [\ell(\theta; Z)]$$

$$\mathbb{E}_{Z \sim P_{test}} [\ell(\theta; Z)]$$

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q} [\ell(\theta; Z)]$$

$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$



Instead of minimizing loss over training distribution,  
minimize loss over distributions **near** it

# Generalization of DRO

automatically can be built.

$$\sup_{\substack{d(Q,P) \leq \rho}} \mathbb{E}_{Z \sim Q} [\ell(\theta; Z)] \geq \mathbb{E}_{Z \sim P_{\text{test}}} [\ell(\theta; Z)]$$

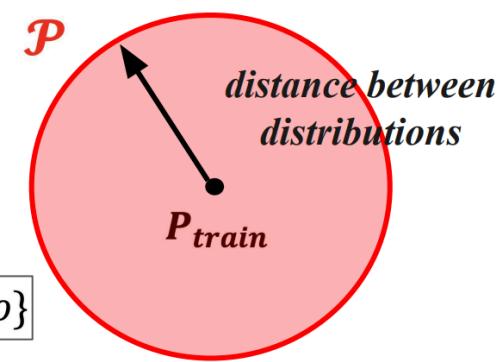
Empirical Risk  
Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{\text{train}}} [\ell(\theta; Z)]$$

DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q} [\ell(\theta; Z)]$$

$$\mathcal{P} = \{Q: \text{Dist}(Q, P_{\text{train}}) \leq \rho\}$$



Instead of minimizing loss over training distribution,  
minimize loss over distributions **near** it

# Duality of DRO

$$R_f(\theta; P) = \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \lambda \mathbb{E}_P \left[ f^* \left( \frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right] + \lambda \rho + \eta \right\}$$

$$f^*(s) := \sup_t \{st - f(t)\}.$$

$$= \sup_{L \geq 0} \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \{ \mathbb{E}_P [L(Z) \ell(\theta; Z)] + \lambda(\rho - \mathbb{E}_P[f(L(Z))]) - \eta(\mathbb{E}_P[L(Z)] - 1) \}$$

$$P-D+I(Q||P) \geq 0$$

$\mathbb{Q}$  is a distribution.

$$L(Z) = \frac{d\mathbb{Q}}{dP} \cdot \mathbb{E}_P \frac{d\mathbb{Q}}{dP} = \mathbb{E}_{\mathbb{Q}}[1] = 1$$

Next step: Solve the sup over L

# Duality of DRO

$$\begin{aligned}
 R_f(\theta; P) &= \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \left\{ \lambda \mathbb{E}_P \left[ f^* \left( \frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right] + \lambda \rho + \eta \right\} & f^*(s) := \sup_t \{st - f(t)\}. \\
 &= \sup_{L \geq 0} \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \{ \mathbb{E}_P [L(Z) \ell(\theta; Z)] + \lambda (\rho - \mathbb{E}_P [f(L(Z))] - \eta (\mathbb{E}_P [L(Z)] - 1)) \} \\
 &= \inf_{\lambda \geq 0, \eta \in \mathbb{R}} \sup_{L \geq 0} \left\{ \lambda \mathbb{E}_P \left[ \frac{L(Z)(\ell(\theta; Z) - \eta)}{\lambda} - f(L(Z)) \right] \right\} + \lambda \rho + \eta. \\
 &= \mathbb{E}_P \left[ f^* \left( \frac{\ell(\theta; Z) - \eta}{\lambda} \right) \right]. \quad \text{↑ reference}
 \end{aligned}$$

$f^*$  (relative)

↳ The loss function:  $f^*$

Duality of  $f$ -divergence DRO, is changing loss function  $\ell$ .

# Variance Regularization

$\chi^2$  divergence

$$f = (+-1)^2$$

$$\inf_{\begin{array}{l} \lambda > 0 \\ \eta \in \mathbb{R} \end{array}} \mathbb{E}_{IP} \left( \frac{\ell(\theta_i z) - \eta}{\lambda} \right)^2 + \ell(\theta_i z).$$



$$\eta = \mathbb{E}_{IP} \ell$$

$$\lambda \text{Var}(\ell) + \mathbb{E}_{IP} \ell$$

# Generalization of DRO

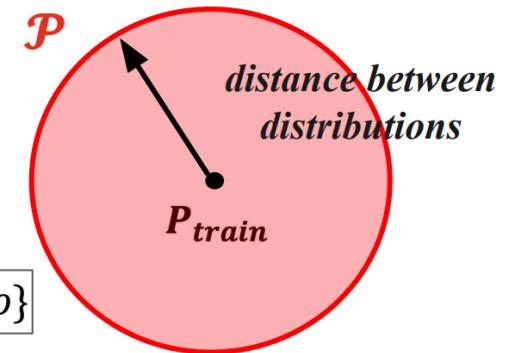
Empirical Risk  
Minimization

$$\min_{\theta \in \Theta} \mathbb{E}_{Z \sim P_{train}} [\ell(\theta; Z)]$$

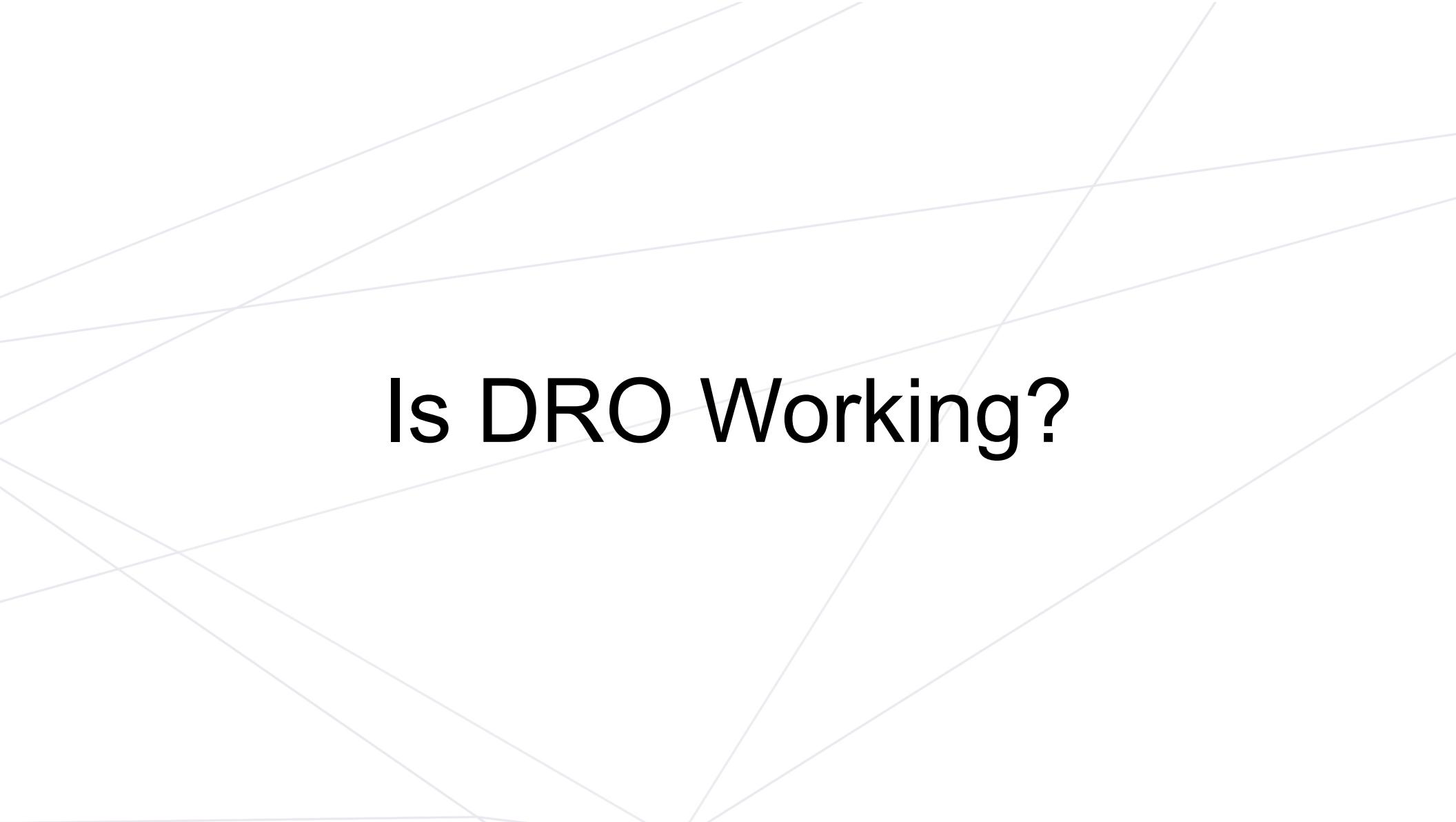
DRO

$$\min_{\theta \in \Theta} \sup_{Q \in \mathcal{P}} \mathbb{E}_{Z \sim Q} [\ell(\theta; Z)]$$

$$\mathcal{P} = \{Q: Dist(Q, P_{train}) \leq \rho\}$$



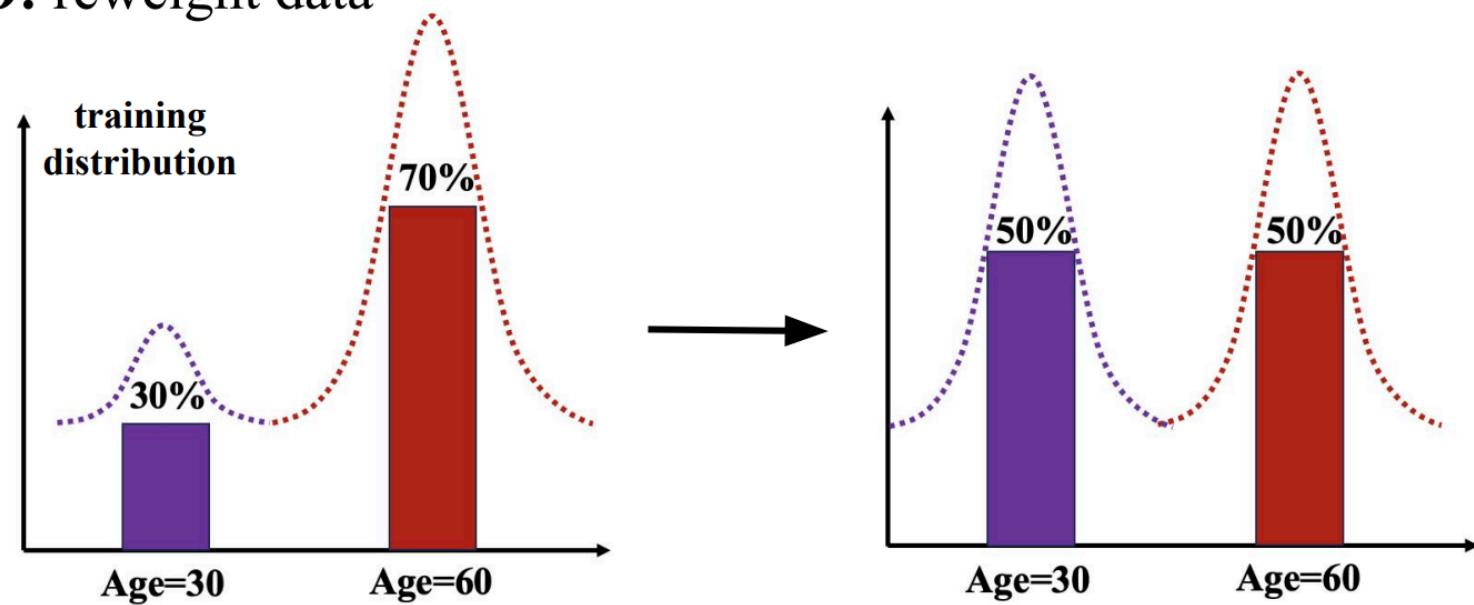
Instead of minimizing loss over training distribution,  
minimize loss over distributions **near** it



# **Is DRO Working?**

# F-divergence DRO only reweighting

*f*-DRO: reweight data



f-DRO

Weight more  
on the data  
have higher loss

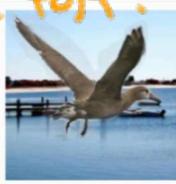
Waterbirds

IW

weight more  
on the data  
opposite in the test.

Common training examples

y: waterbird  
a: water background



y: landbird  
a: land background

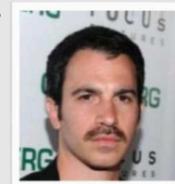


CelebA

y: blond hair  
a: female



y: dark hair  
a: male



MultiNLI

y: contradiction  
a: has negation  
(P) The economy could be still better.  
(H) The economy has never been better.

y: entailment  
a: no negation  
(P) Read for Slate's take on Jackson's findings.  
(H) Slate had an opinion on Jackson's findings.

Weights more on rare data!

Test examples

y: waterbird  
a: land background



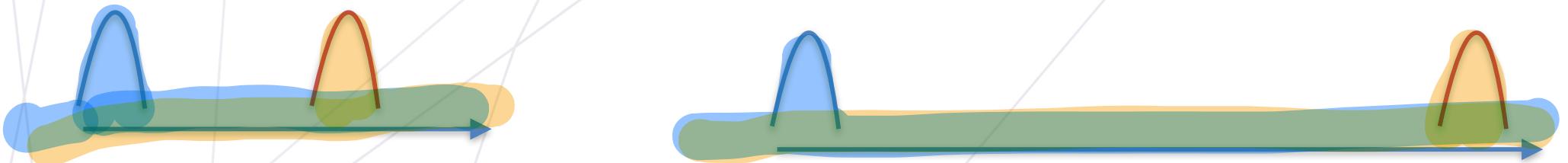
y: blond hair  
a: male



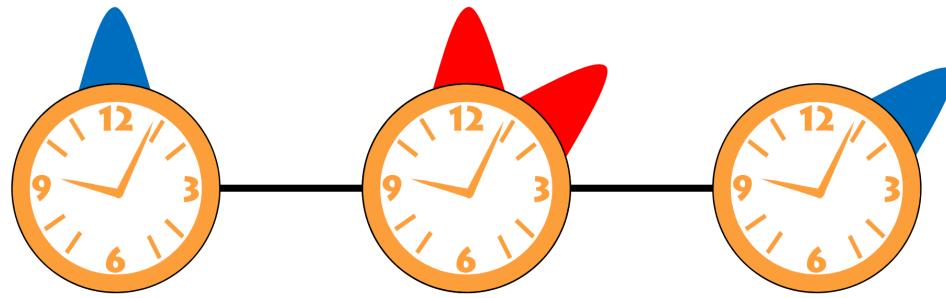
y: entailment  
a: has negation  
(P) There was silence for a moment.  
(H) There was a short period of time where no one spoke.

# What's wrong about f-divergence

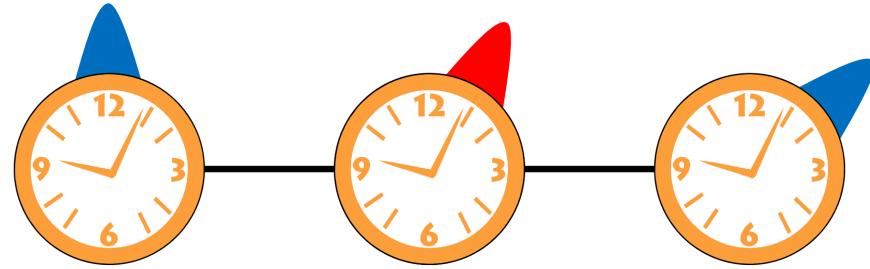
$$D_f(P||Q) = \mathbb{E}_P f\left(\frac{dQ}{dP}\right)$$



# What's wrong about f-divergence



Or



# Over-parameterization?

Standard Regularization		Average Accuracy		Worst-Group Accuracy	
		ERM	DRO	ERM	DRO
		Train	Test	Train	Test
Waterbirds	Train	100.0	100.0	100.0	100.0
	Test	97.3	97.4	60.0	76.9
CelebA	Train	100.0	100.0	99.9	100.0
	Test	94.8	94.7	41.1	41.1
MultiNLI	Train	99.9	99.3	99.9	99.0
	Test	82.5	82.0	65.7	66.4

Strong $\ell_2$ Penalty		Average Accuracy		Worst-Group Accuracy	
		ERM	DRO	ERM	DRO
		Train	Test	Train	Test
Waterbirds	Train	97.6	99.1	35.7	97.5
	Test	95.7	96.6	21.3	84.6
CelebA	Train	95.7	95.0	40.4	93.4
	Test	95.8	93.5	37.8	86.7

model class is  $C_{n-1}$ .  
Strong  $\ell_2$  Penalty

No improv

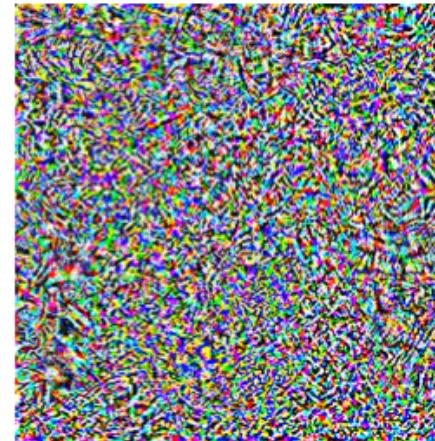
# Adversarial Learning

# adversarial training

“pig”



+ 0.005 x



=



# How to find Adversarial Examples?



$x$   
“panda”  
57.7% confidence

+ .007 ×



$$\text{sign}(\nabla_x J(\theta, x, y))$$

“nematode”  
8.2% confidence

=



$$x + \epsilon \text{sign}(\nabla_x J(\theta, x, y))$$

“gibbon”  
99.3 % confidence

Optimization that maximize the loss

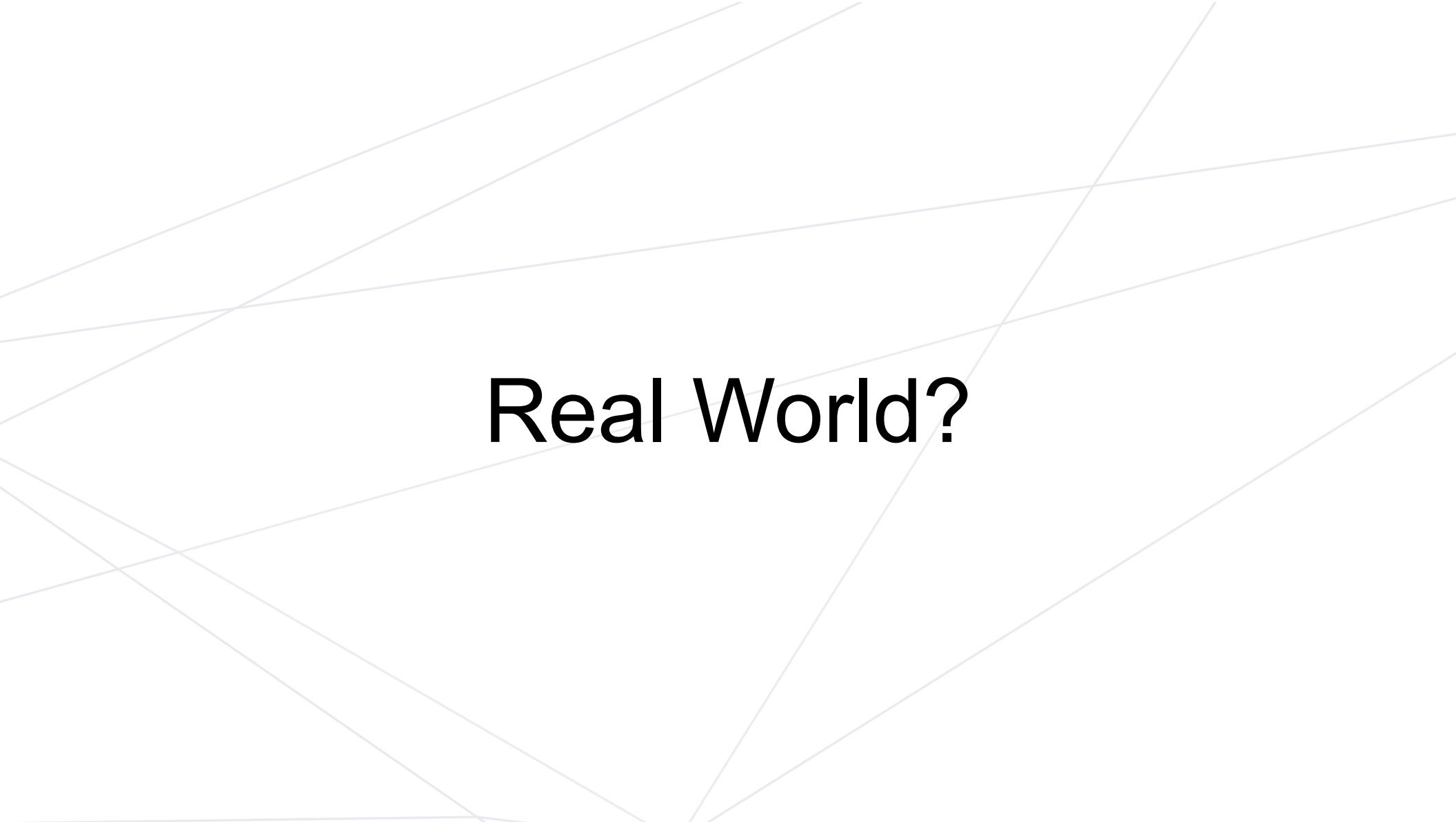
# Adversarial Training

$$\min_{\theta} \rho(\theta), \quad \text{where} \quad \rho(\theta) = \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \max_{\delta \in \mathcal{S}} L(\theta, x + \delta, y) \right].$$

<https://arxiv.org/pdf/1706.06083>

# Adversarial Training Can Hurt Generalization

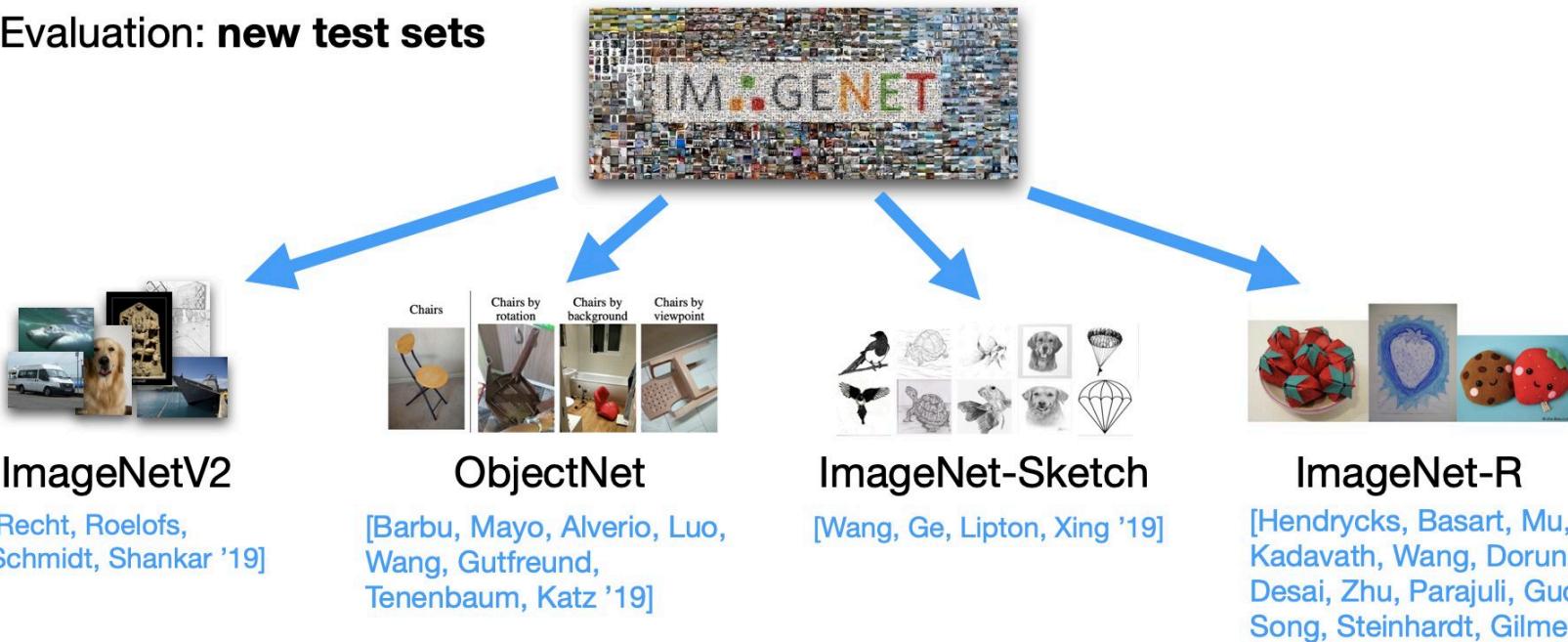
	<b>Standard training</b>	<b>Adversarial training</b>
<b>Robust test</b>	3.5%	45.8%
<b>Robust train</b>	-	100%
<b>Standard test</b>	95.2%	87.3%
<b>Standard train</b>	100%	100%



**Real World?**

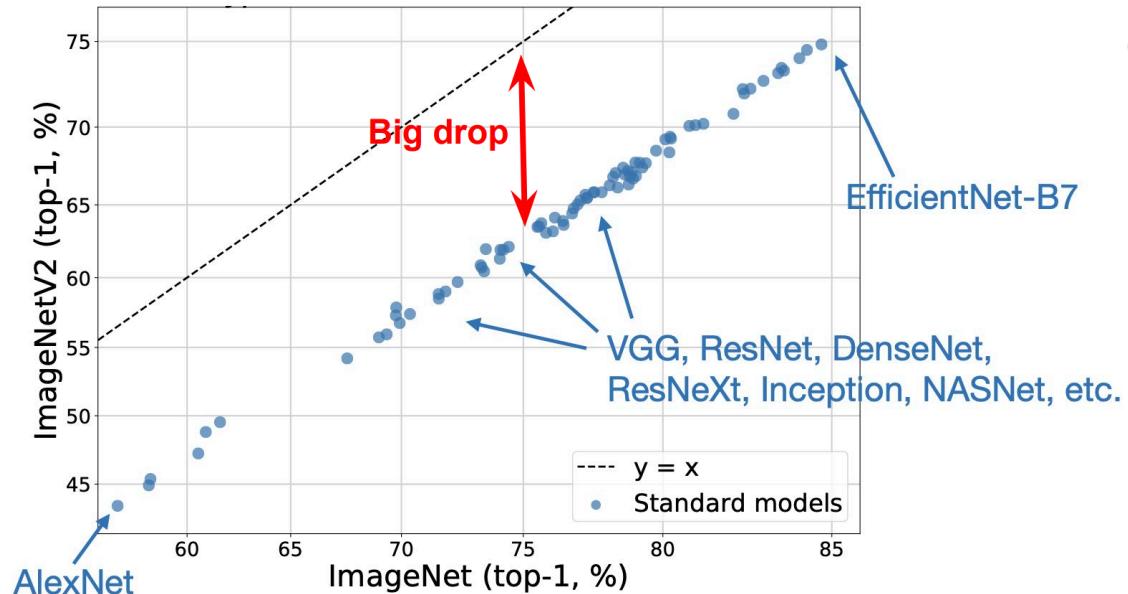
Lots of progress on ImageNet over the past 10 years, but models are still not robust.

Evaluation: **new test sets**



# Agree on the line!

Recht B, Roelofs R, Schmidt L, et al. Do imagenet classifiers generalize to imagenet? [C]// International conference on machine learning. PMLR, 2019: 5389-5400.



[Taori, Dave, Shankar, Carlini, Recht, Schmidt '20]

# Why?