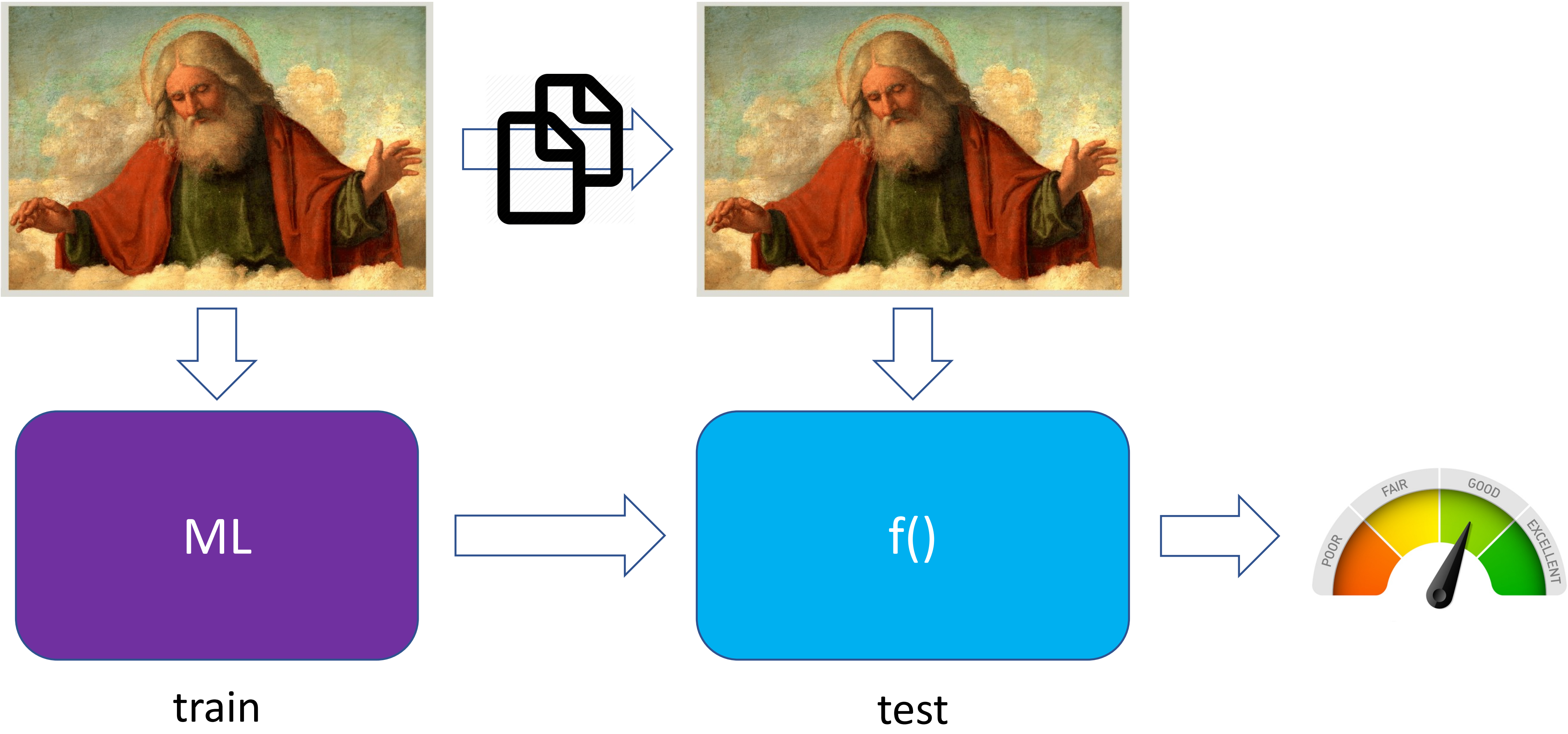
Distribution Shift

Zachary Lipton & Henry Chai

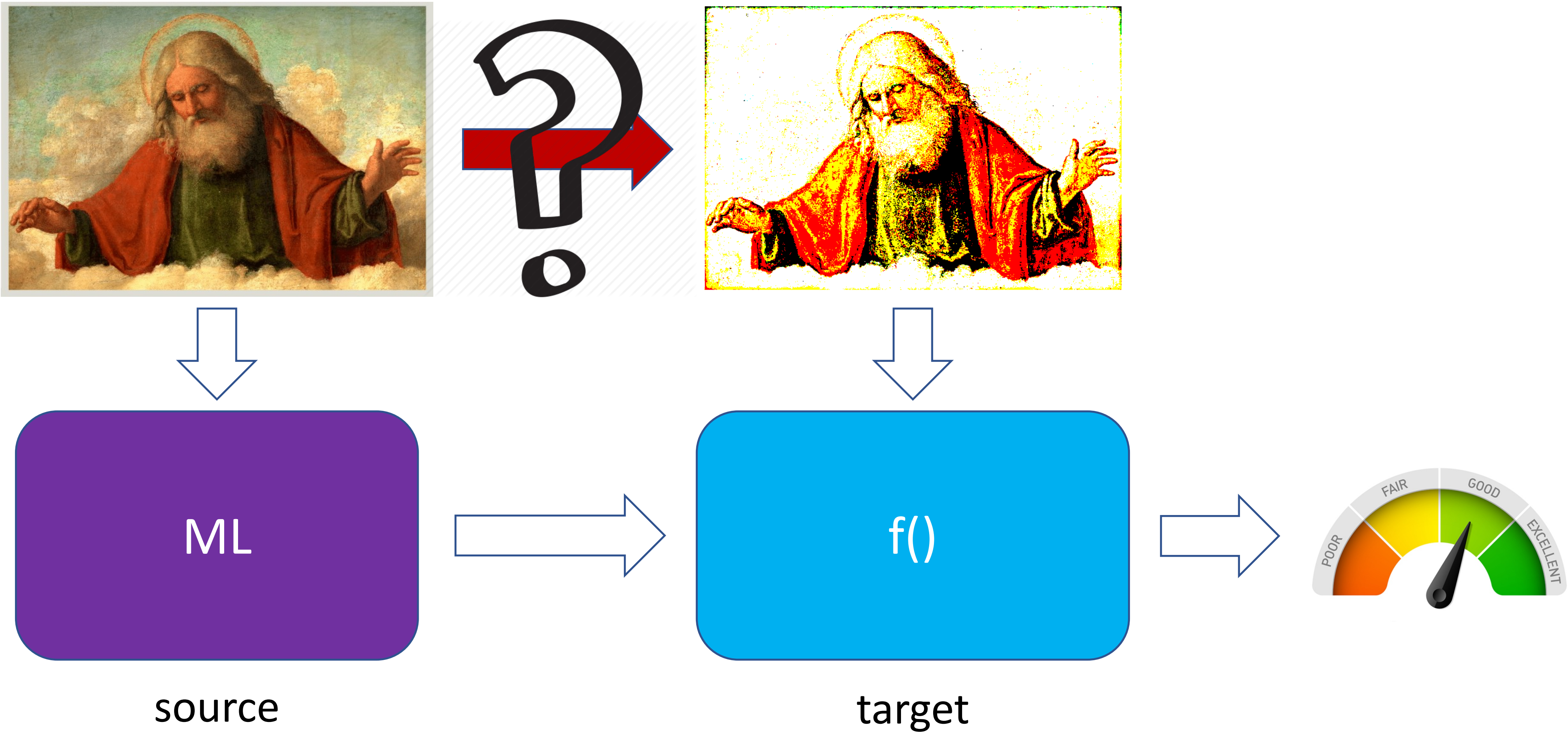
MACHINE LEARNING DEPARTMENT

10701 — November 27th

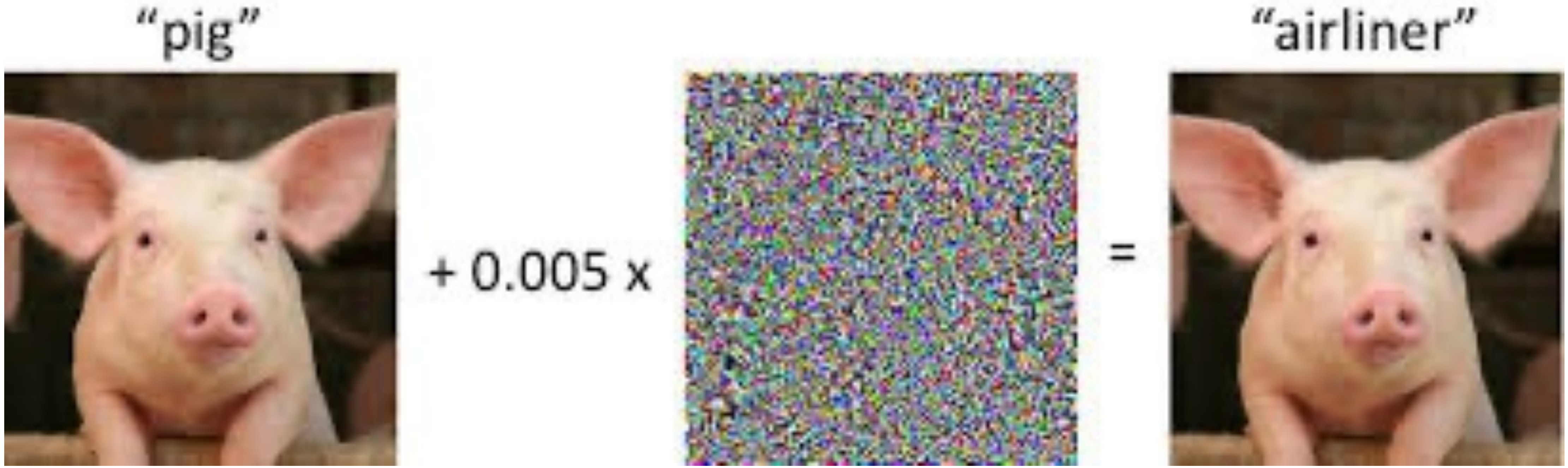
Standard assumptions



Reality



Adversarial Examples



(Szegedy et al 2014 https://arxiv.org/abs/1312.6199)

Targeted vs Untargeted Attacks

- Untargeted: search for a perturbation (under constraint) that maximizes loss

$$\underset{\delta \in \Delta}{\text{maximize}}\, \ell(h_\theta(x+\delta), y)$$

- Targeted: search for a perturbation (under constraint) that maximizes original loss **AND** probability assigned to target class

$$\underset{\delta \in \Delta}{\text{maximize}}(\ell(h_\theta(x+\delta), y) - \ell(h_\theta(x+\delta), y_{\text{target}}))$$

Adversarial Training

1. For each $x, y \in B$, solve the inner maximization problem (i.e., compute an adversarial example)

$$\delta^\star(s) = \underset{\delta \in \Delta(x)}{\operatorname{argmax}}\, \ell(h_\theta(x+\delta)), y)$$

1. Compute the gradient of the empirical adversarial risk, and update θ

$$\theta := \theta - \frac{\alpha}{|B|} \sum_{(x,y)\in B} \nabla_\theta \ell(h_\theta(x + \delta^\star(x))), y).$$

Tutorial (excerpted): https://adversarial-ml-tutorial.org/introduction/
Papers:

1. Original adversarial training paper: https://arxiv.org/abs/1412.6572
2. State of the art (iterated attack): https://arxiv.org/abs/1706.06083

Adversarial Misspellings (Char-Level Attack)

Against BERT for sentiment, 1-char attack send error from 90.3%→45.8%

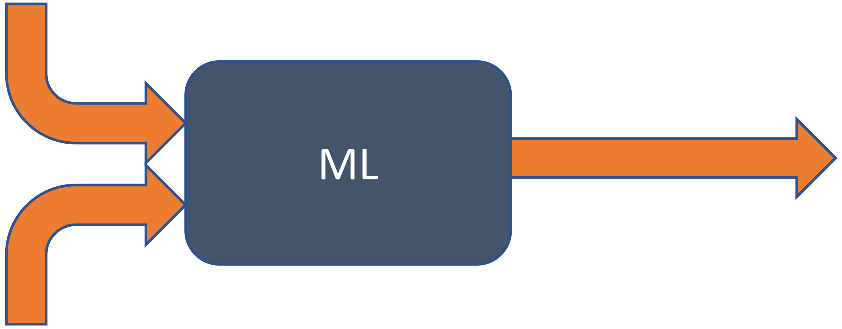
Alteration	Movie Review	Label
Original	A triumph, relentless and beautiful in its downbeat darkness	+
Swap	A triumph, relentless and beuatiful in its downbeat darkness	–
Drop	A triumph, relentless and beautiful in its dwnbeat darkness	–

Combating Adversarial Misspellings with Robust Word Recognition
Danish Pruthi, Bhuwan Dhingra, Z. (ACL 2019)
https://arxiv.org/abs/1905.11268

Training Tasks Can Fail to Represent Reality

E.g., how much *reading* does reading comprehension require?

Which team has won the most Super Bowl titles?



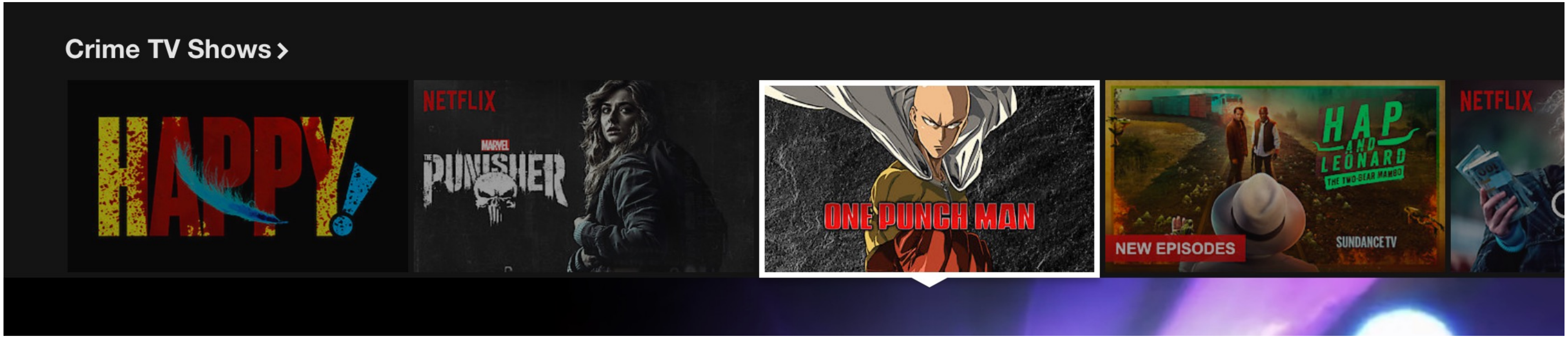
The **Pittsburgh Steelers** have the most Super Bowl championship titles, with six. The **New England Patriots** have the most Super Bowl appearances, with ten. **Charles Haley** and **Tom Brady** both have five Super Bowl rings, which is the record for the most rings won by a single player

The **Pittsburgh Steelers** have the most Super Bowl championship titles, with six. The **New England Patriots** have the most Super Bowl appearances, with ten. **Charles Haley** and **Tom Brady** both have five Super Bowl rings, which is the record for the most rings won by a single player

https://arxiv.org/abs/1808.04926 (Kaushik, Z—EMNLP 2018)

Feedback Loops

- Insidiously, the very deployment of a model can invalidate it
- E.g., recommender system, trained on user behavior, **applied to alter it**



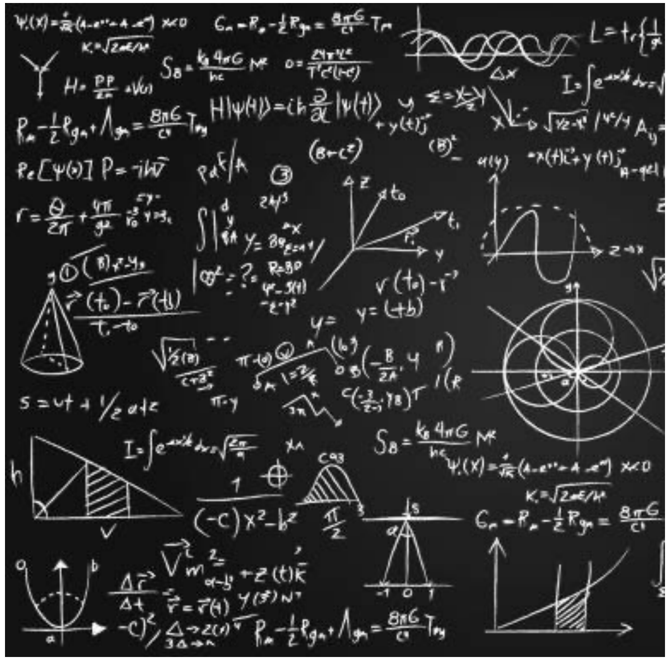
One classifier to rule them all!

Mission Impossible

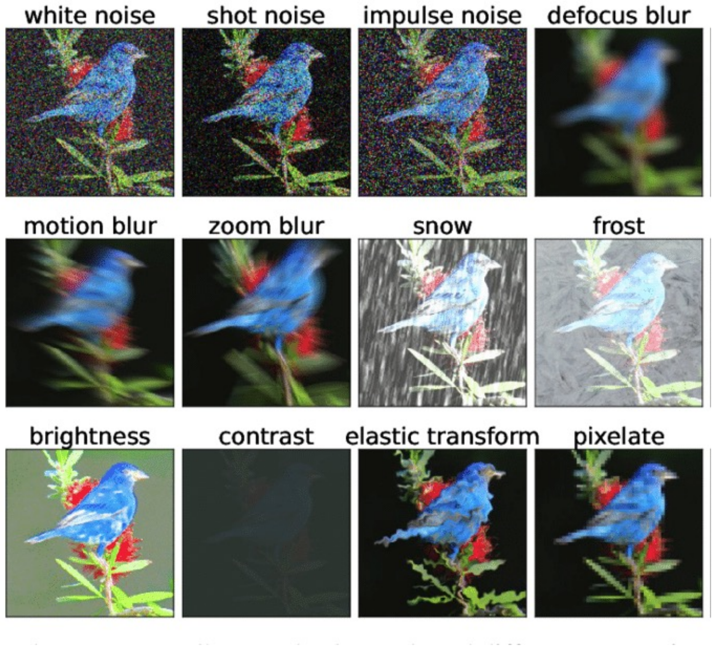
Impossibility absent assumptions

- **No classifier** will work well on **all distributions**
- Guaranteed performance under shift possible w. **strong assumptions**
- Typical: bounded divergence or invariant conditionals, shared support
- Most familiar assumption: covariate shift $p(y|\mathbf{x}) = q(y|\mathbf{x})$
- But when **x** doesn't cause y & absent realizability, $p(y|\mathbf{x})$ *does* change
- Practical benefits under unstated / implicit / murky assumptions?

The Landscape of Distribution Shift



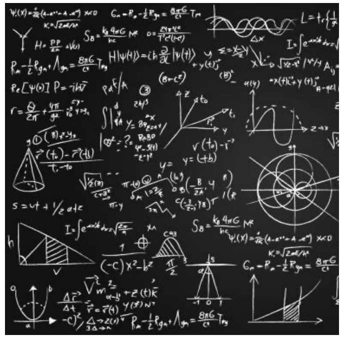
theoretically coherent work
on idealized shift models



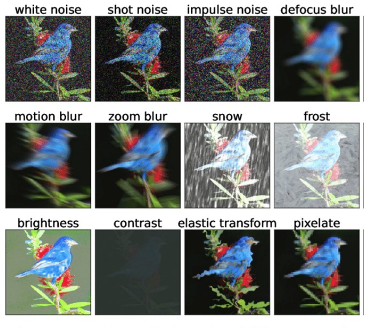
empirical deep learning efforts
benchmark evaluation, heuristics

unpredictable shifts, limited
faithfulness to any assumption

The Landscape of Distribution Shift



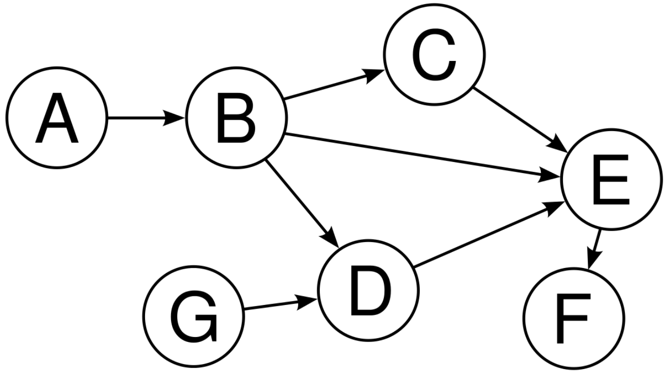
theoretically coherent work
on idealized shift models



empirical deep learning efforts
benchmark evaluation, heuristics

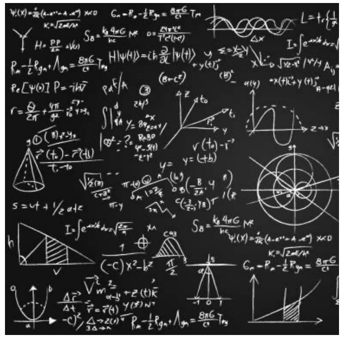
unpredictable shifts, limited
faithfulness to any assumption

Structured Shifts

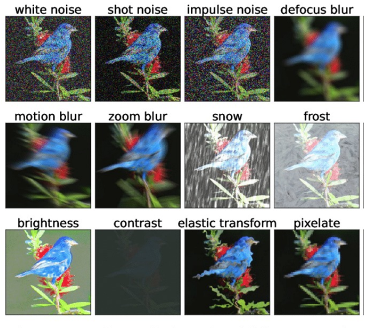


Fuzzy Shifts

The Landscape of Distribution Shift



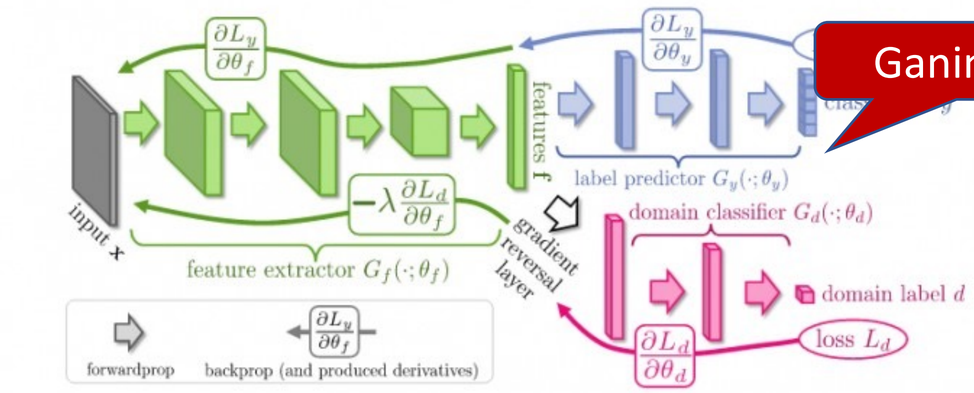
theoretically coherent work
on idealized shift models



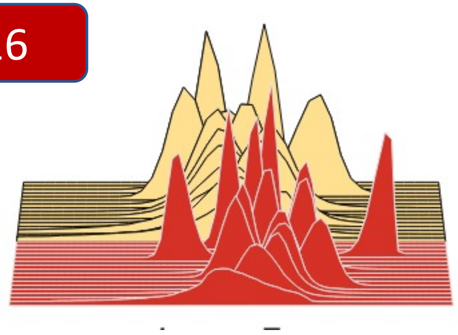
empirical deep learning efforts
benchmark evaluation, heuristics

unpredictable shifts, limited
faithfulness to any assumption

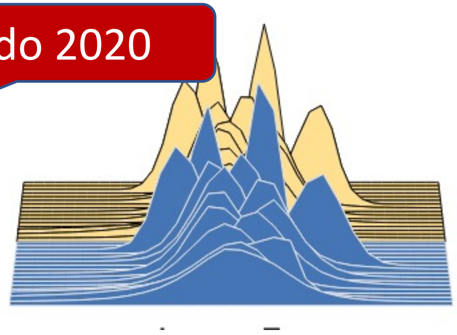
Heuristics like matching representations or adapting batch statistics



Ganin 2016

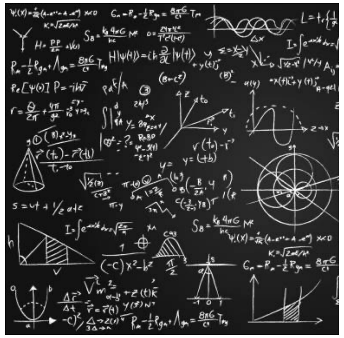


Nado 2020

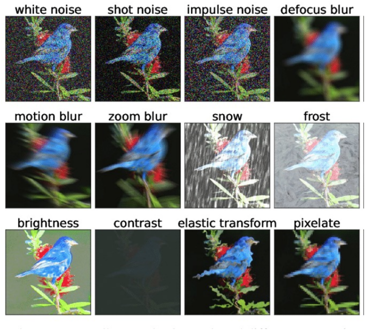


For problems, see: Domain Adaptation with Asymmetrically-Relaxed Distribution Alignment (ICML 2019): (https://arxiv.org/abs/1903.01689)

The Landscape of Distribution Shift

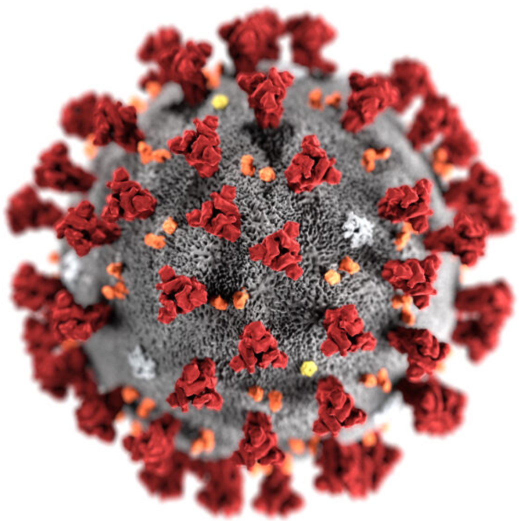


theoretically coherent work
on idealized shift models



empirical deep learning efforts
benchmark evaluation, heuristics

unpredictable shifts, limited
faithfulness to any assumption

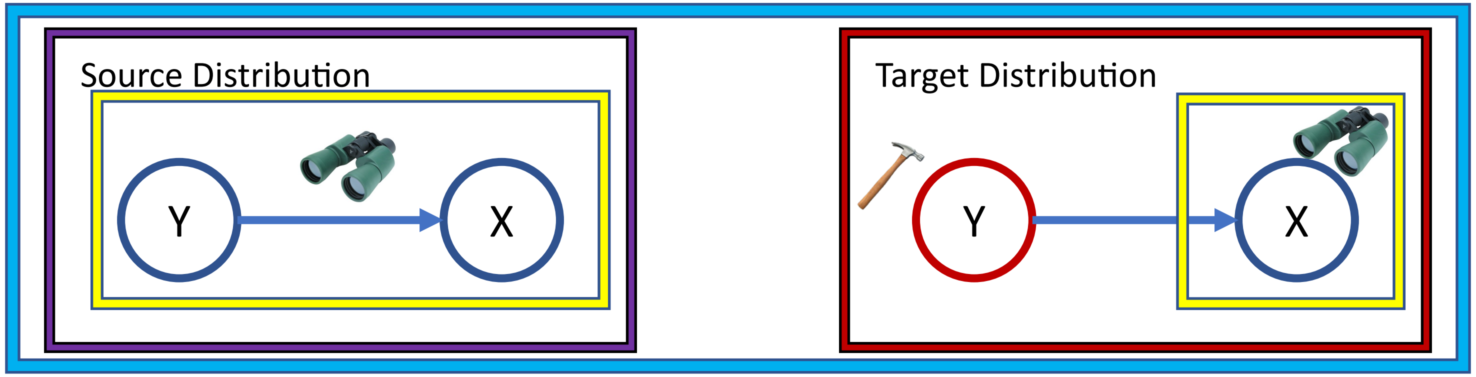


Koh & Sagawa 2020

WILDS

Anatomy of a structured shift problem

- **Domains/Environments—**how many? how much data from each?
- **Structure—**model of data generating process
- **Visibility—** which variables are observed in each environment?
- **Manipulation rules**
  - What parts can/can't change?
  - By what amount?
  - In vs out–of-support?
- **Objective** (what to estimate)
- **Statistical Capabilities**
  - What relationships are estimable (& how well)?

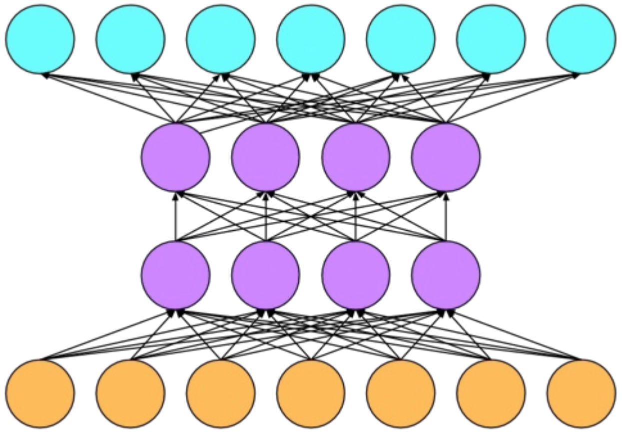


Some Examples of Structured Shift

- Covariate Shift — $P(Y|X)$ invariant, overlapping support $q(x) \subseteq p(x)$
- **Label Shift** — $P(X|Y)$ invariant, overlapping classes $q(y) \subseteq p(y)$
- **PU Learning** — $P(X|Y)$ invariant, + 1 new class: $P(Y=N) = 0$, $Q(Y=N) > 0$
- Open Set Label Shift — $P(Y|X)$ invariant, many prev classes, one new
- Latent Label Shift — $P(X|Y)$ invariant, many domains Q_i, all unlabeled
- Missingness shift — Source data missing at random according to $\mathbf{m}_s$.

Two Obstacles to Practicality

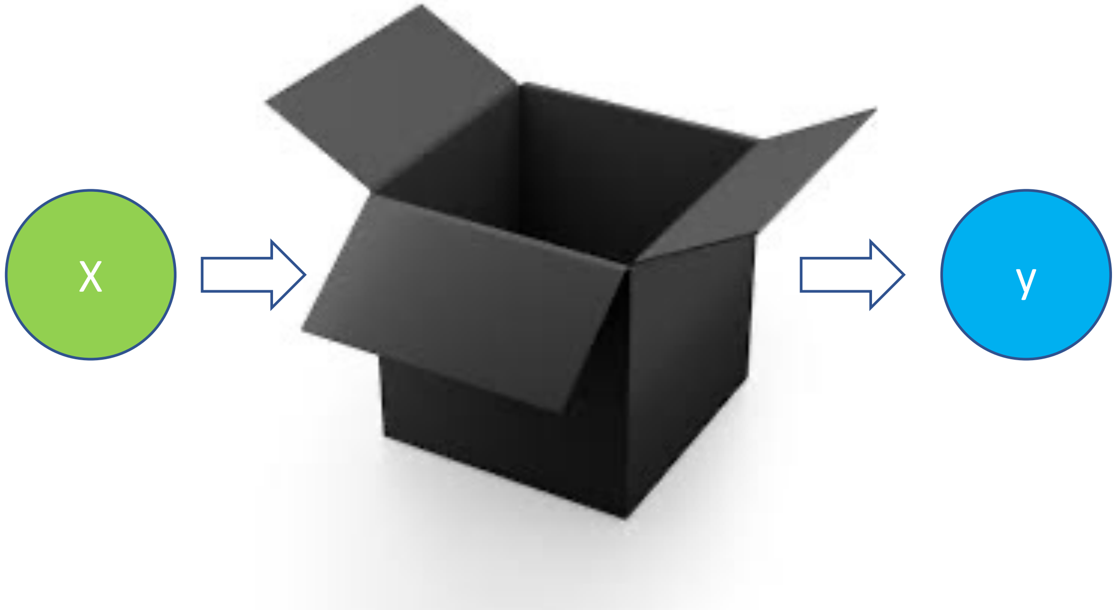
- Identification is nice but we need practical estimators for high-dim data



- Assumptions too rigid, performance under fuzzy violations unknown

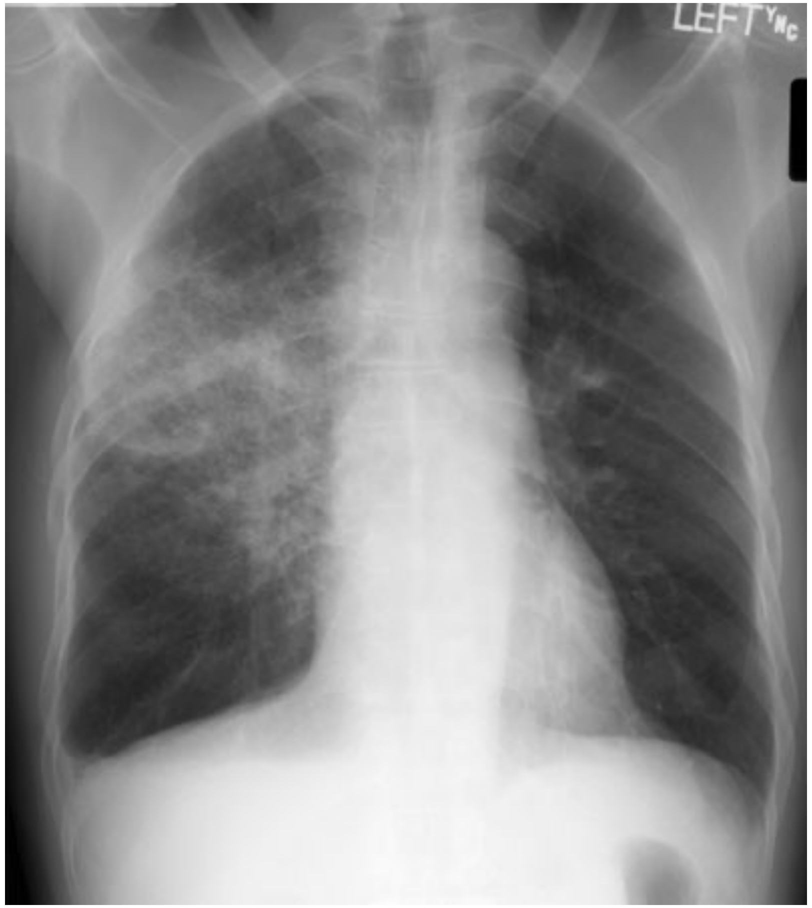
The Move: Leveraging Black Box Predictors

- No theory says we should be able to predict well (even on iid data) w high-dimensional, arbitrarily non-linear data (e.g. images, speech)
- However, we want to show that when it's possible to learn good iid classifiers, we can leverage these black boxes to get target classifiers



Motivation 1: Pneumonia prediction

- August: we train pneumonia predictor *f*
- Prevalence is .05% in population
- We run classifier on training data
  - Model predicts ~.05% positive
- We run it on validation data
  - Model predicts ~.05% positive
- We run it in the wild
  - Model predicts ~.05% positive



Epidemic

- We run classifier in January
- It predicts 5% (vs .05%) positive
- How many ppl *really* have pneumonia?
- **If i.i.d. violated, then why should we trust *f* at all?**

Motivation 2: Image Classification

- Train a classifier to recognize objects with uniform *p*(y)
- Get 70% accuracy, say with balanced errors
- Deploy in wild with some randomly-chosen *q*(y)
- No real-life data distribution will have equal numbers of **axolotl**, **golden retriever**, **mortarboard, ice cream, couch**
- We still get 70% accuracy even though this is an easier problem

The test-Item effect

- Humans can update priors without supervision
  [Zhu, Xiaojin et al. *"Cognitive models of test-item effects in human category learning" (ICML 2010)*]

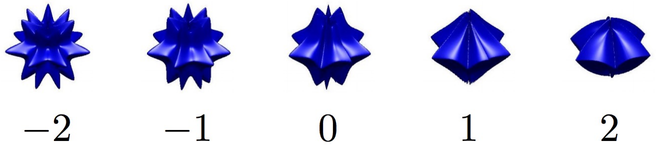


Figure 1. Example stimuli

- Randomly select people
- Show identical training items
- **Finding:** *"one can then manipulate them into classifying some test items in opposite ways, simply depending on what other test items they are asked to classify (without label feedback)"*

Domain Adaptation – Formal Setup

- **Probabilities**
  - Source distribution $p(\mathbf{x},y)$
  - Target distribution $q(\mathbf{x},y)$
- **Data**
  - Training examples $(\mathbf{x}_{1,}\ y_1), \ldots, (\mathbf{x}_n, y_n) \sim p(\mathbf{x},y)$
  - Test examples $(\mathbf{x}'_1, \ldots, \mathbf{x}_m') \sim q(\mathbf{x})$
- **Objective**
  - Predict well on the test distribution, **WITHOUT** seeing any labels $y_i \sim q(y)$

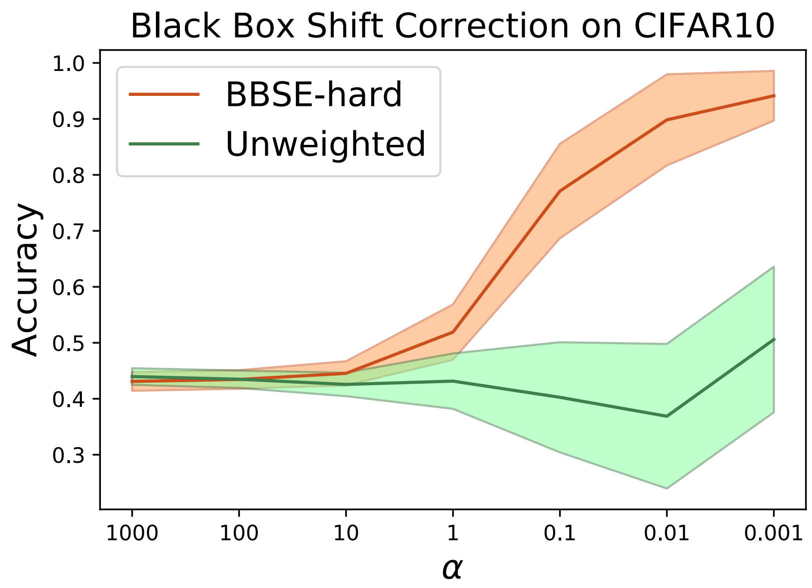
Our goals

- When the distribution $p(f(\mathbf{x}))$ shifts then we know

$$p(\mathbf{x}, y) \neq q(\mathbf{x}, y) \quad \text{because} \quad p(\mathbf{x}) \neq q(\mathbf{x})$$

- Under distribution shift we would like to
  1. **Detect** that a shift has occurred
  2. **Estimate** the new label distribution $q(y)$
  3. **Correct** the classifier f

- **All without seeing new labels**

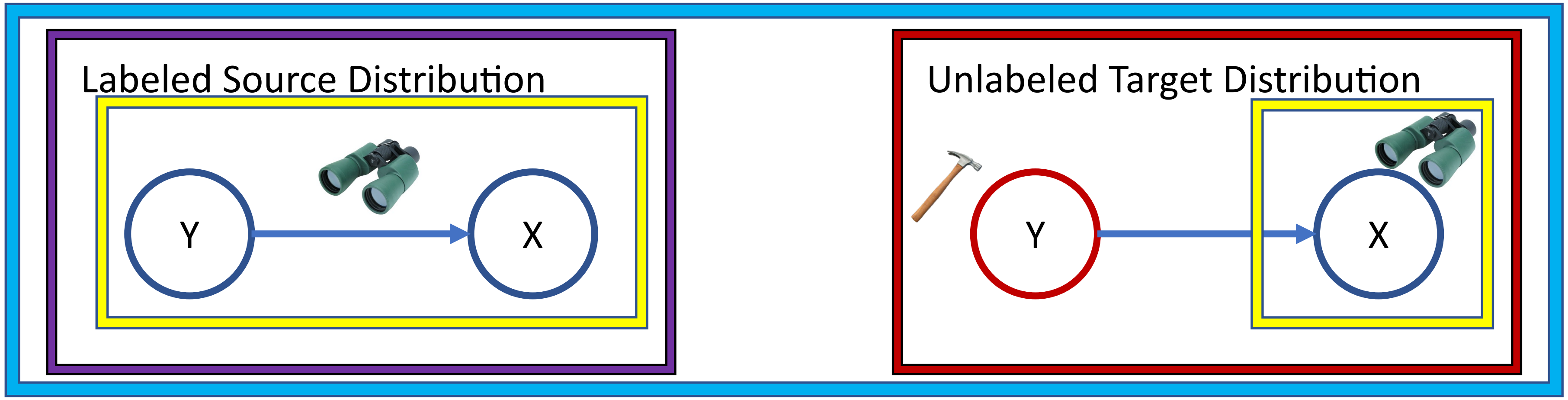


Label Shift (aka Target Shift)

- Assume *p*(**x**,y) changes, but the conditional ***p*(x|y)** is **fixed**

$$q(y, \mathbf{x}) = q(y)p(\mathbf{x}|\mathbf{y})$$

- Corresponds to anticausal assumption, (disease causes symptoms)
- Assumptions: for all y such that q(y) > 0, p(y) > 0



Detecting and Correcting for Label Shift with Black Box Predictors (Z., Wang*, Smola—ICML 2018)*
Schölkopf et al "On Causal and Anticausal Learning" (ICML 2012)

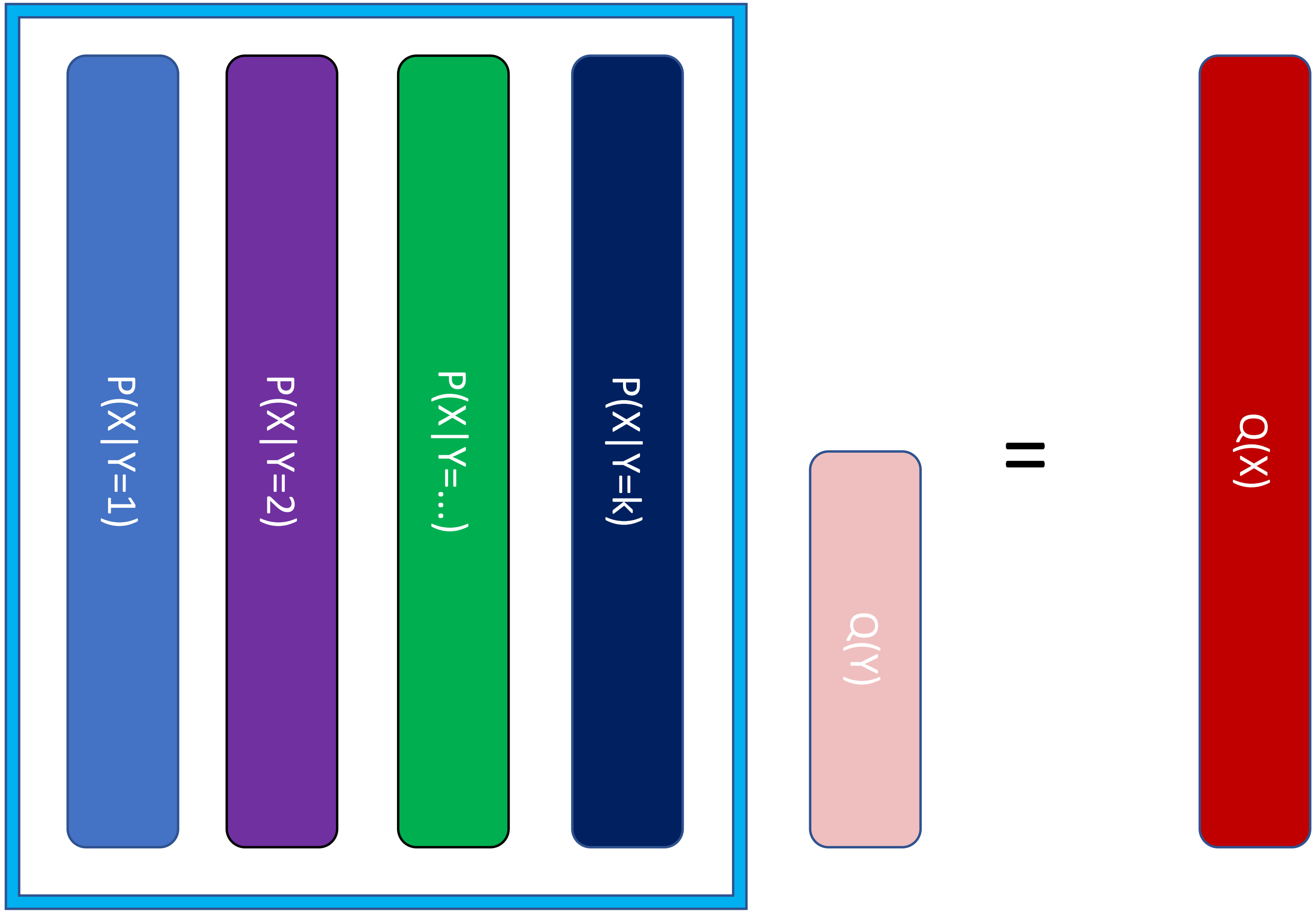
Contrast with Covariate Shift

- Assume that $p(\mathbf{x},y)$ changes, but conditional $p(y|\mathbf{x})$ is **fixed**

$$q(y, \mathbf{x}) = q(\mathbf{x})p(y|\mathbf{x})$$

- *Implicitly assumes that x causes y*
- Appealing because we have samples $x_i \sim p(\mathbf{x})$ and $x_j' \sim q(\mathbf{x})$
- Natural to estimate $q(\mathbf{x})/p(\mathbf{x})$ -> use for importance-weighted ERM
- But symptoms don't causes diseases & pixels don't cause cats!
- Under an epidemic, $p(y|\mathbf{x})$ **should** change!

Label Shift Identification



Black Box Shift Estimation (BBSE)

- Consistent estimator with intuitive error bounds
- Accuracy does not depend directly on data dimension
- Exploit black box predictors for dimensionality reduction (d → 1)
  - Much easier than two sample-tests in high-dim spaces (Ramdas et al., 2015)
- Adaptive method
  - For stronger f, we get provably tighter error bounds
  - Lousy (inaccurate, uncalibrated, biased) f → BBSE still consistent

Assumptions

A.1 The *label shift* (also known as *target shift*) assumption

$$p(\boldsymbol{x}|y) = q(\boldsymbol{x}|y) \quad \forall\, x \in \mathcal{X},\ y \in \mathcal{Y}.$$

A.2 For every $y \in \mathcal{Y}$ with $q(y) > 0$ we require $p(y) > 0$.[2]

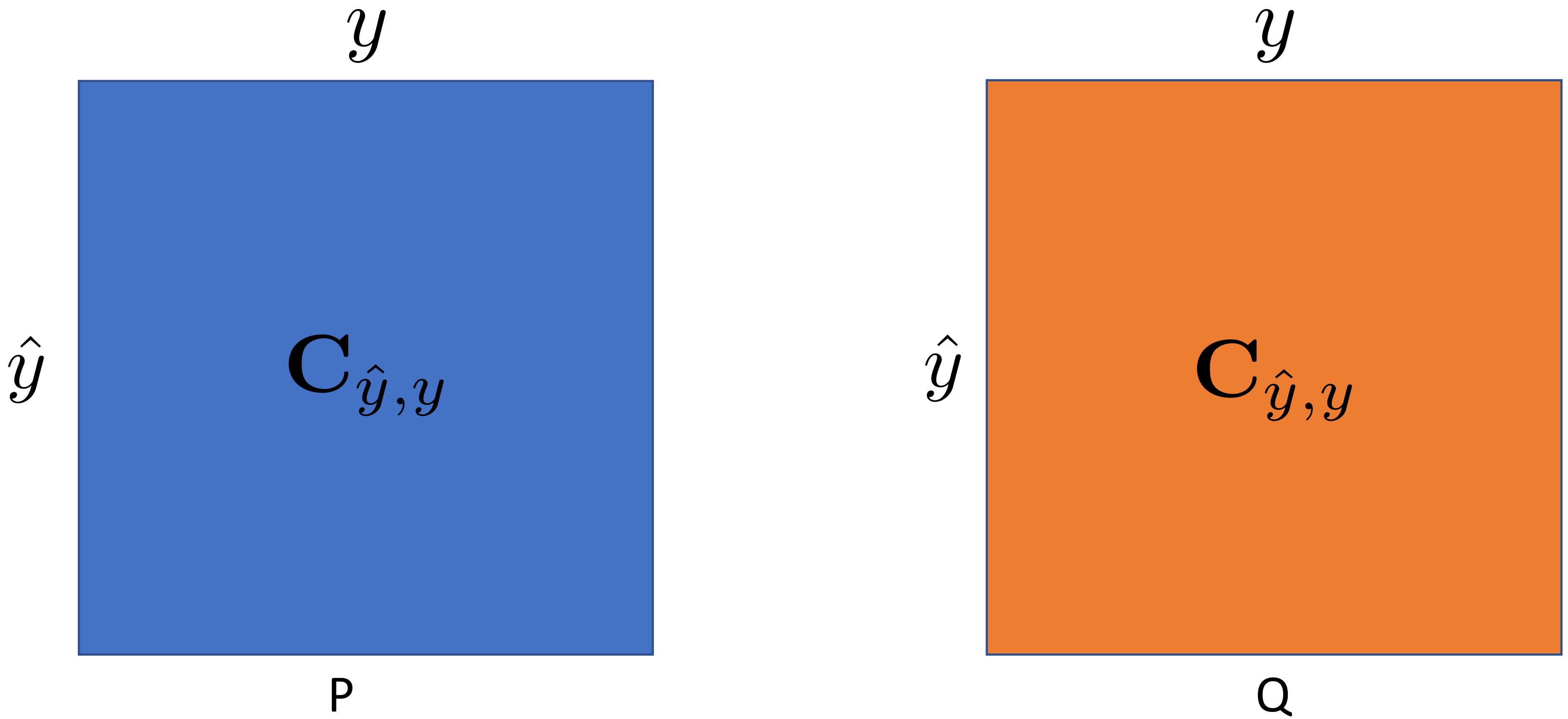
A.3 Access to a black box predictor $f : \mathcal{X} \to \mathcal{Y}$ where the expected confusion matrix $\mathbf{C}_p(f)$ is invertible.

$$\mathbf{C}_P(f) := p(f(\boldsymbol{x}), y) \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$$

- Explanation
  - A.1 – our premise, appropriate under anti-causal learning
  - A.2 – identifiability assumption, can't recognize class y if p(y) = 0
  - A.3 – says our confusion matrix is not degenerate

Confusion matrices

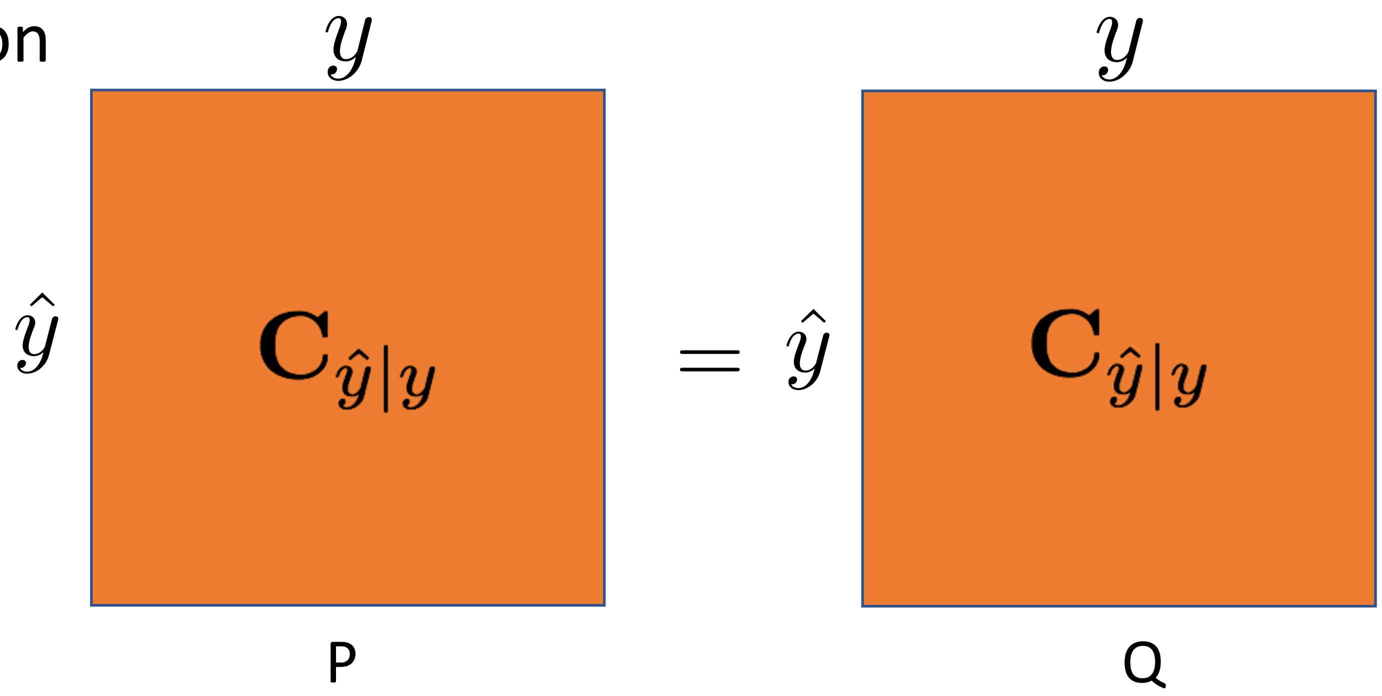
- Let's look at the **expected** confusion matrices



Applying the label shift assumption…

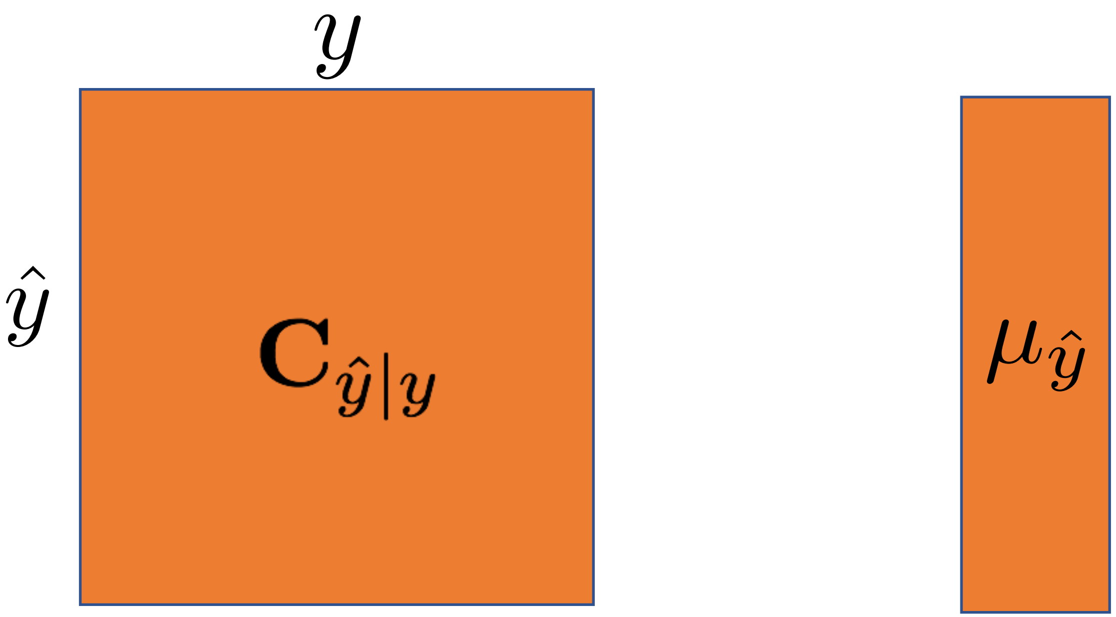
- $\mathbf{C}_{\hat{y}|y}$ - column-normalized is **identical** in under P and Q

- We can estimate confusion matrix on P
- Don't need to observe labels from Q



What do we do with the target data?

- We observe black box predictor outputs on examples $x'_j \sim Q$

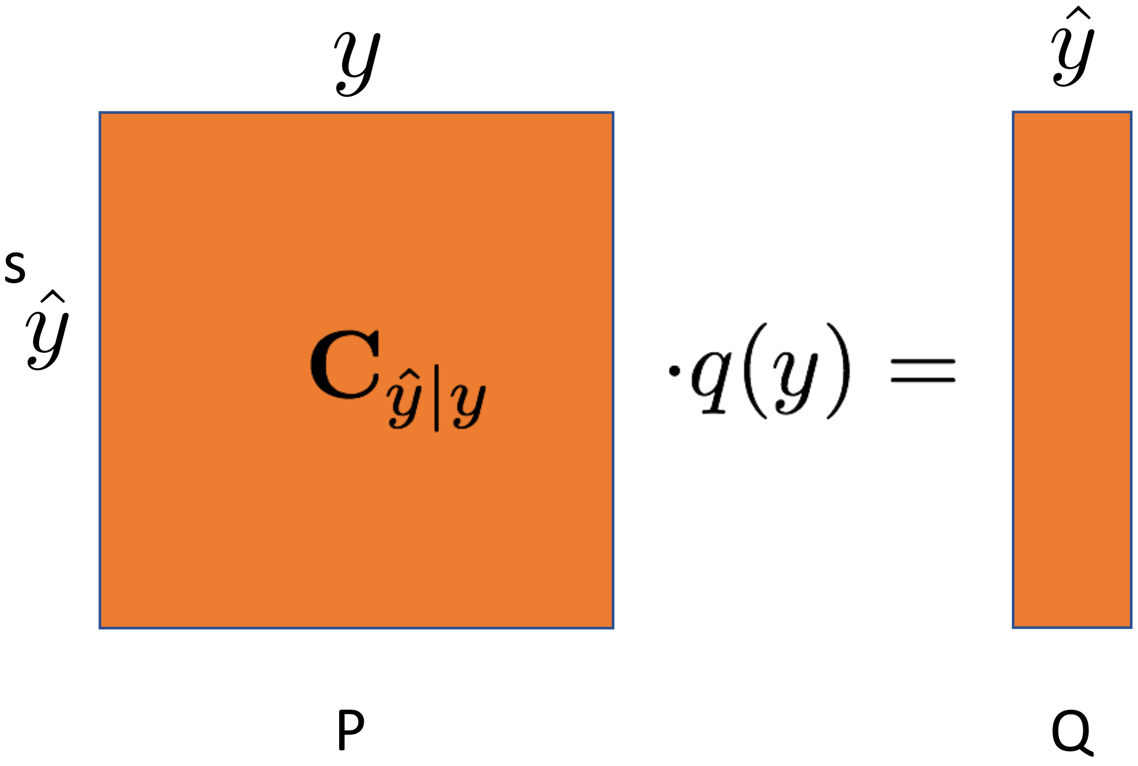


Same on P and Q
Can estimate on P

Can estimate on Q
by running *f*(**x**')

Black box shift estimation

- Because $\mathbf{C}_{\hat{y}|y}$ is same on P and Q, we can solve for q(*y*) by solving a linear system
- We just need:
  1. Confusion matrix converges
  2. Mean (target) output converges
  3. Confusion matrix invertible
- Can solve same system but without normalizing columns to get back importance weights



The estimator

- Gives us a vector of importance weights *q*(y)/*p*(y)

$$\hat{\mathbf{w}} = \hat{C}_{\hat{y},y}^{-1}\hat{\boldsymbol{\mu}}_{\hat{y}}$$

- Either switch with normalized C or multiply element-wise by *p*(*y*) (or its MLE estimate if unknown to get an estimator of *q*(*y*)

$$\hat{\mu}_y = \mathrm{diag}(\hat{\nu}_y)\hat{\mathbf{w}}$$

Consistency

- Easy to show, just need
  1. empirical confusion matrix converges to its expectation
  2. average classifier response (on test data) converges to its expectation
  3. empirical confusion matrix is invertible

- By Strong law of large numbers, as n → ∞

$$\hat{\mathbf{C}}_{\hat{y},y} \longrightarrow \mathbf{C}_{\hat{y},y}$$
$$\hat{\mu}_{\hat{y}} \rightarrow \mu_{\hat{y}}$$

- Can show via Borel-Cantelli lemma that as as n → ∞, probability that empirical confusion matrix is not invertible approaches 0.

Error bound

$$\|\hat{\boldsymbol{w}} - \boldsymbol{w}\|_2^2 \leq \frac{C}{\sigma_{\min}^2} \left(\frac{\|\boldsymbol{w}\|^2 \log n}{n} + \frac{k \log m}{m} \right)$$