CENG7880
Trustworthy and Responsible AI

Instructor: Sinan Kalkan

(https://ceng.metu.edu.tr/~skalkan)

For course logistics and materials:
https://metu-trai.github.io

MIDDLE EAST TECHNICAL UNIVERSITY
DEPARTMENT OF COMPUTER ENGINEERING

Administrative Notes

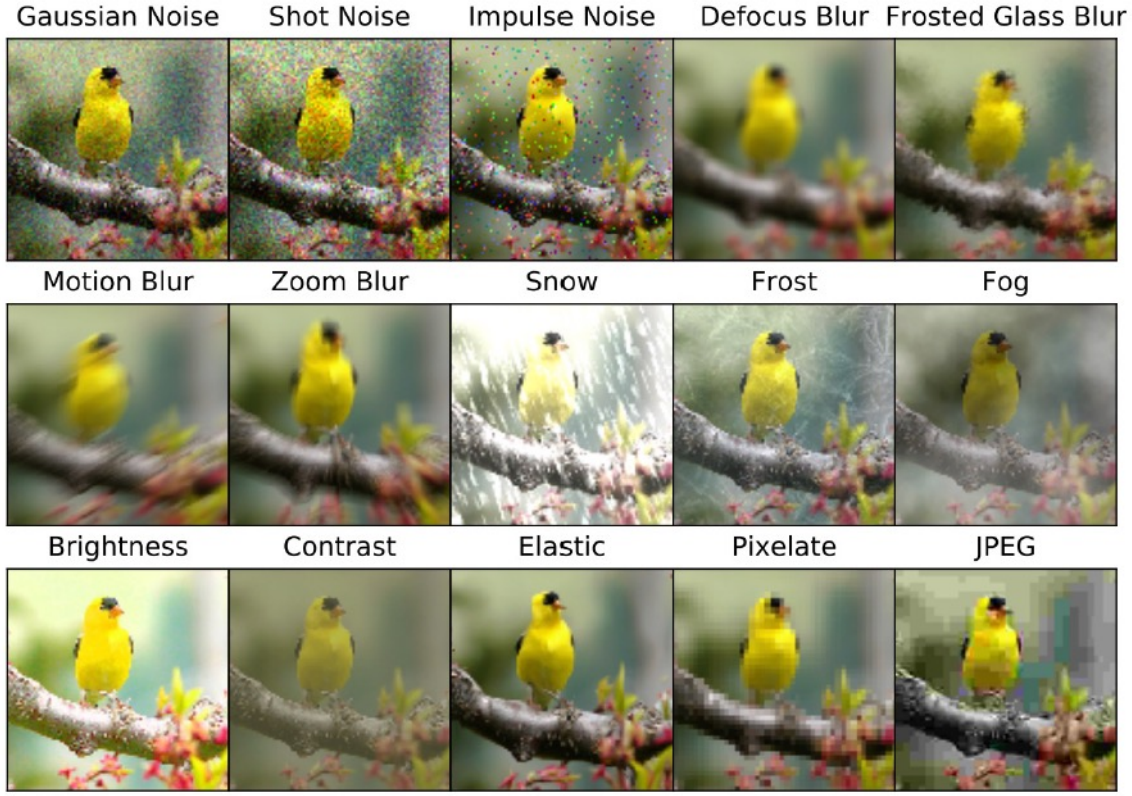
- ~~Selecting papers for the projects~~
  - ~~Deadline: 19 October~~
  - ~~Form: https://forms.gle/A3taWgxoCHumfYfz9~~

Agenda

- Robustness
  - Distribution shifts (this week)
  - Adversarial robustness (next week)
  - Certifying robustness (next next week)
  - Calibrated predictions & uncertainty (next next next week)

Robustness: Agenda
Robustness to Distribution Shifts

- Images
  - Noise, color changes, light changes, day/night, summer/winter, …
- Audio
  - Noise, background, gain level, …
- Text
  - Noise, synonyms, alternative phrasing, ..

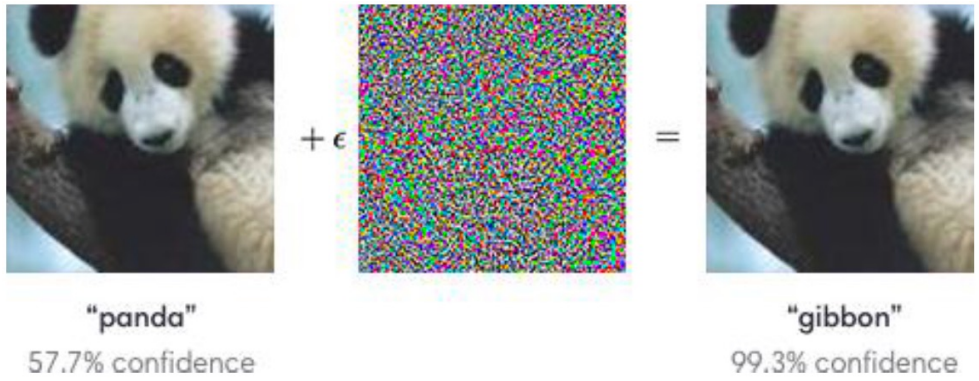


Hendrycks & Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

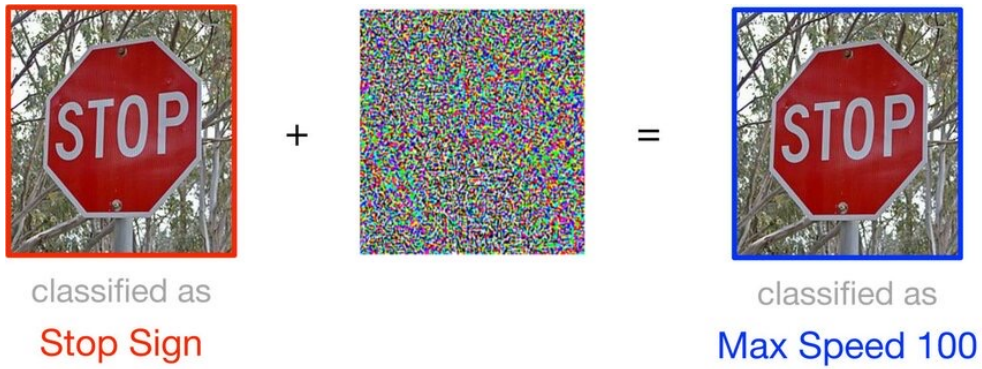
Robustness: Agenda
Adversarial Robustness

- Types of adversarial attacks

- Adversarial training



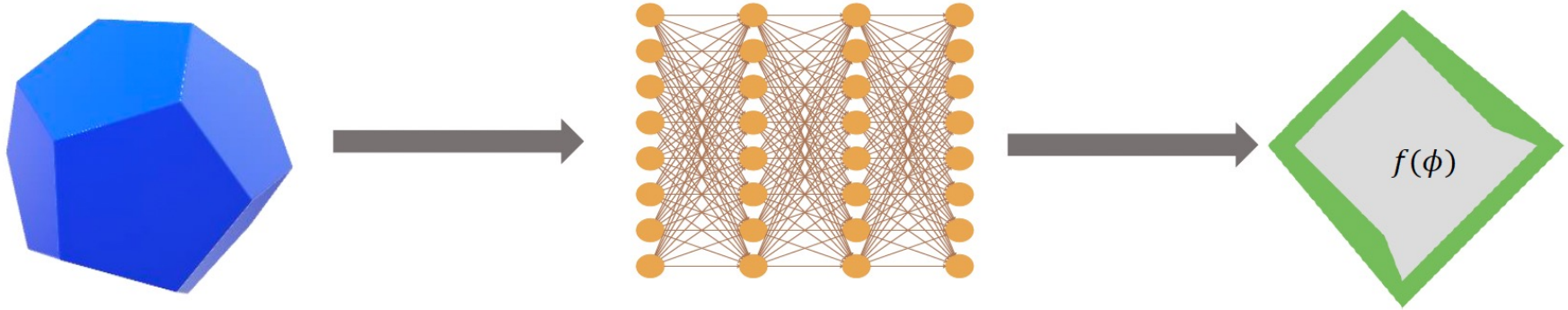
Szegedy et al., Intriguing Properties of Neural Networks, 2014



https://www.altacognita.com/robust-attribution/

Robustness: Agenda
Certifying/Verifying Robustness

Neural network certification: problem statement



Figs: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Robustness: Agenda
Calibrated Predictions & Uncertainty

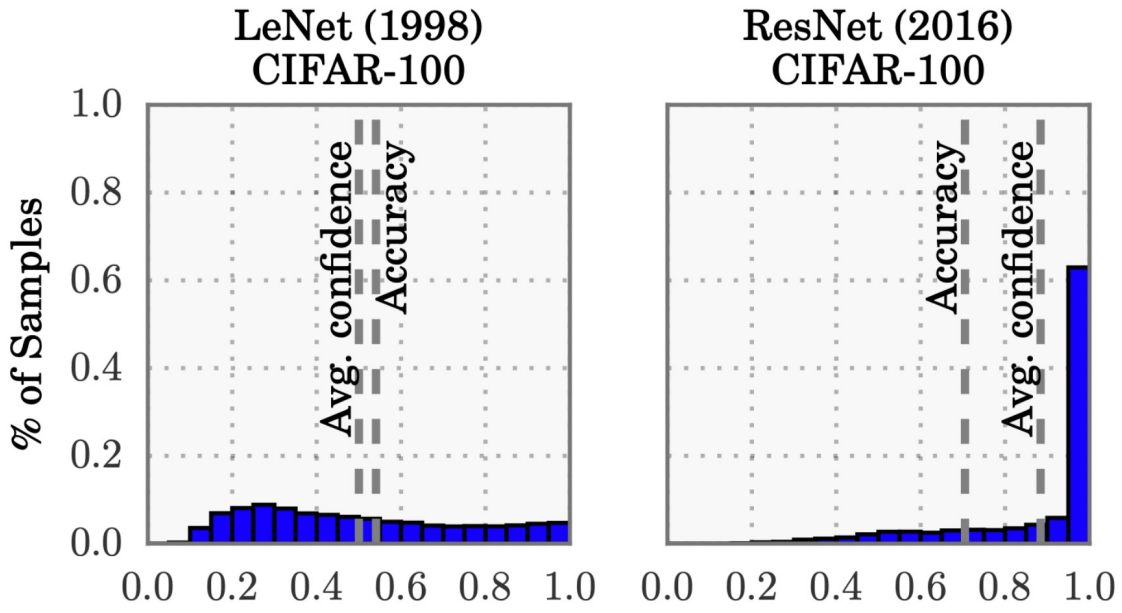


Fig: Uni of Pennsylvania, Trustworthy ML - CIS 7000

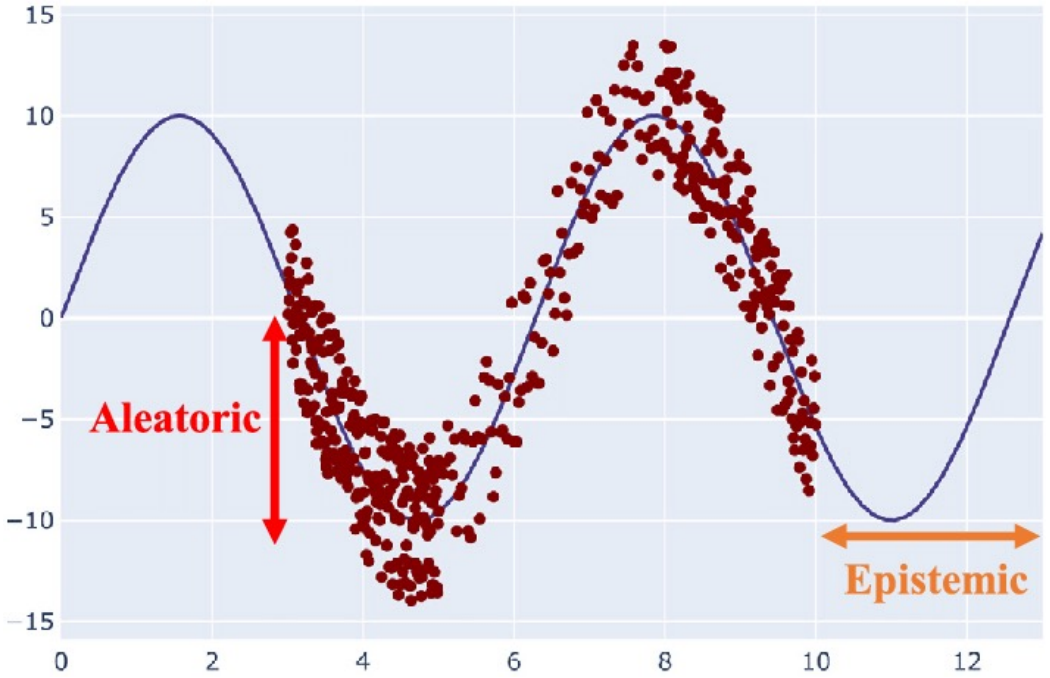


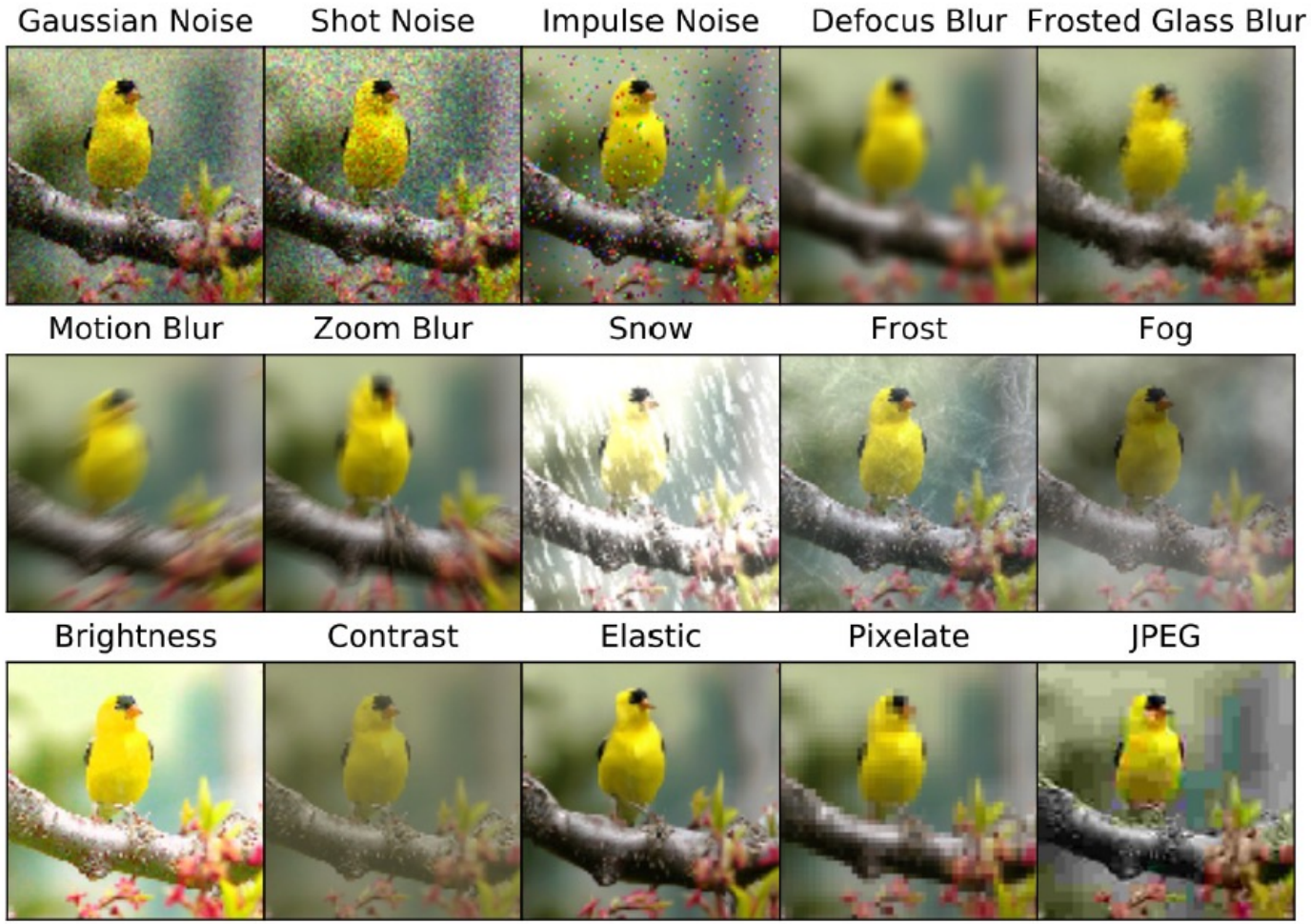
Fig: Abdar et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, 2021

Robustness to Distribution Shift

Disclaimer: The slides are mostly borrowed from:

Uni of Pennsylvania, Trustworthy ML - CIS 7000

Example: Synthetic Perturbations



Hendrycks & Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Example: Synthetic Perturbations

- **Significantly reduces performance**
  - 20% error rate → 80% error rate
- **Data augmentation can help (but not 100% solution)**

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Data Augmentation

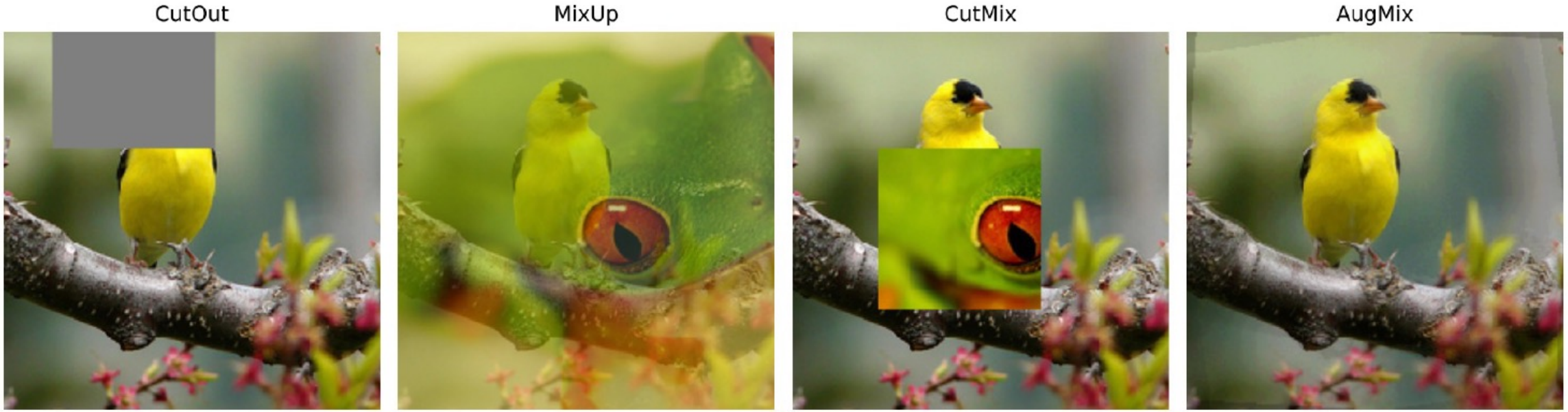


Figure 1: A visual comparison of data augmentation techniques. AUGMIX produces images with variety while preserving much of the image semantics and local statistics.

Hendrycks et al., AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

mixup: Beyond Empirical Risk Minimization, ICLR 2018.

MixUp

$$\begin{aligned} \hat{x} &= \lambda x_i + (1-\lambda) x_j, \\ \hat{y} &= \lambda y_i + (1-\lambda) y_j, \end{aligned}$$

where $\lambda \in [0, 1]$ is a random number

Fig: https://www.kaggle.com/code/kaushal2896/data-augmentation-tutorial-basic-cutout-mixup

AugMix

AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, ICLR 2020.

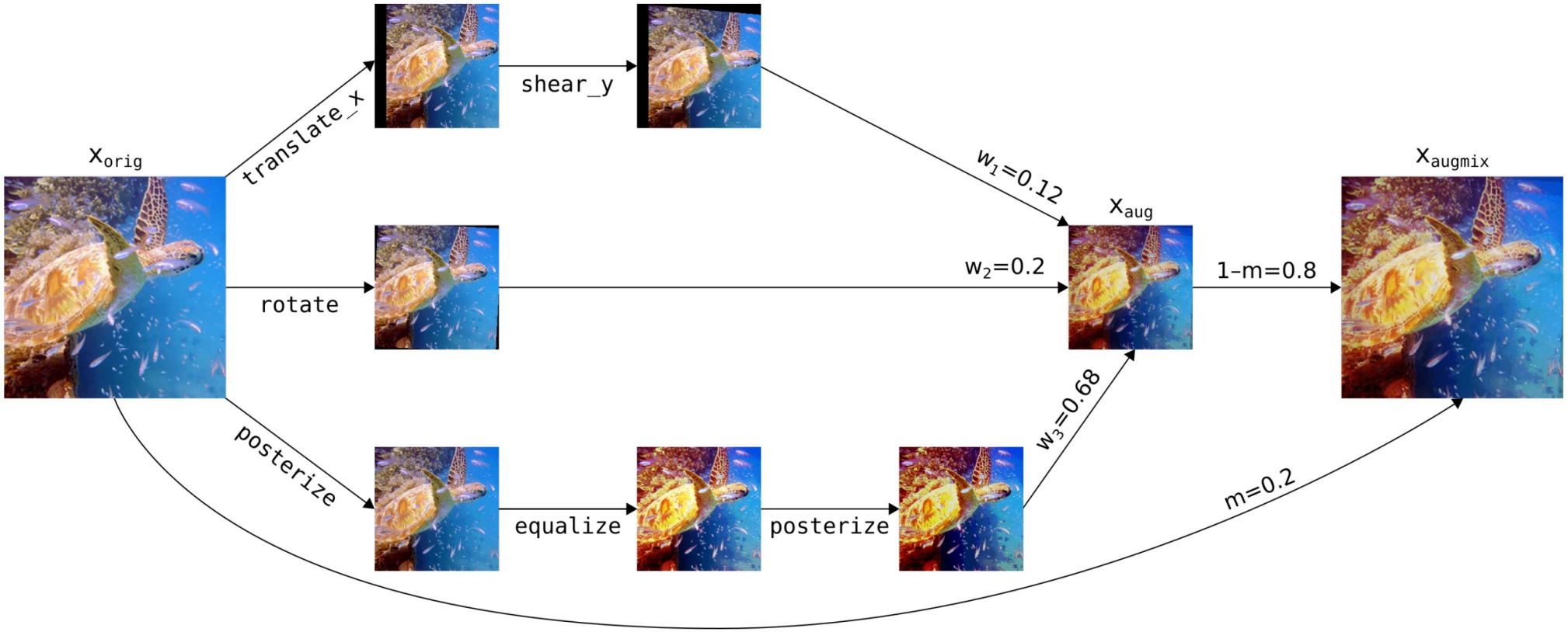
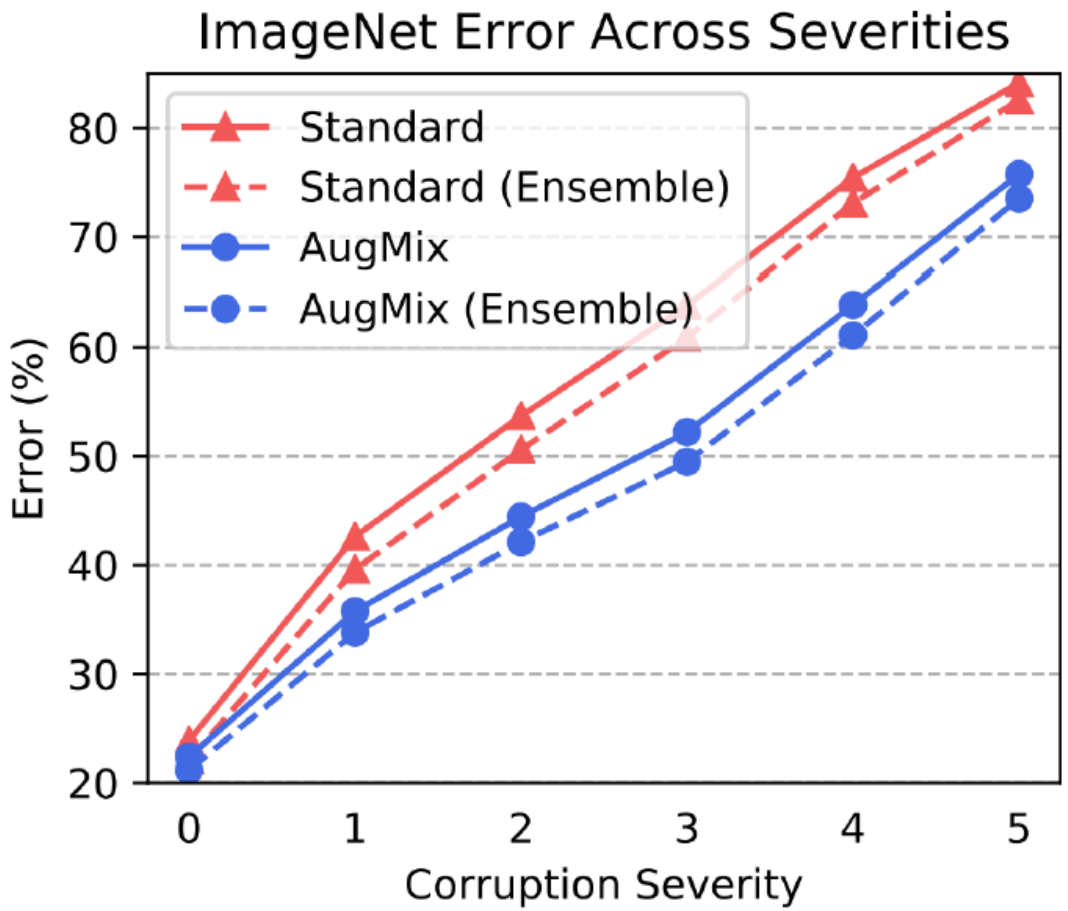


Figure 4: A realization of AUGMIX. Augmentation operations such as translate_x and weights such as m are randomly sampled. Randomly sampled operations and their compositions allow us to explore the semantically equivalent input space around an image. Mixing these images together produces a new image without veering too far from the original.

Data Augmentation



Hendrycks et al., AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

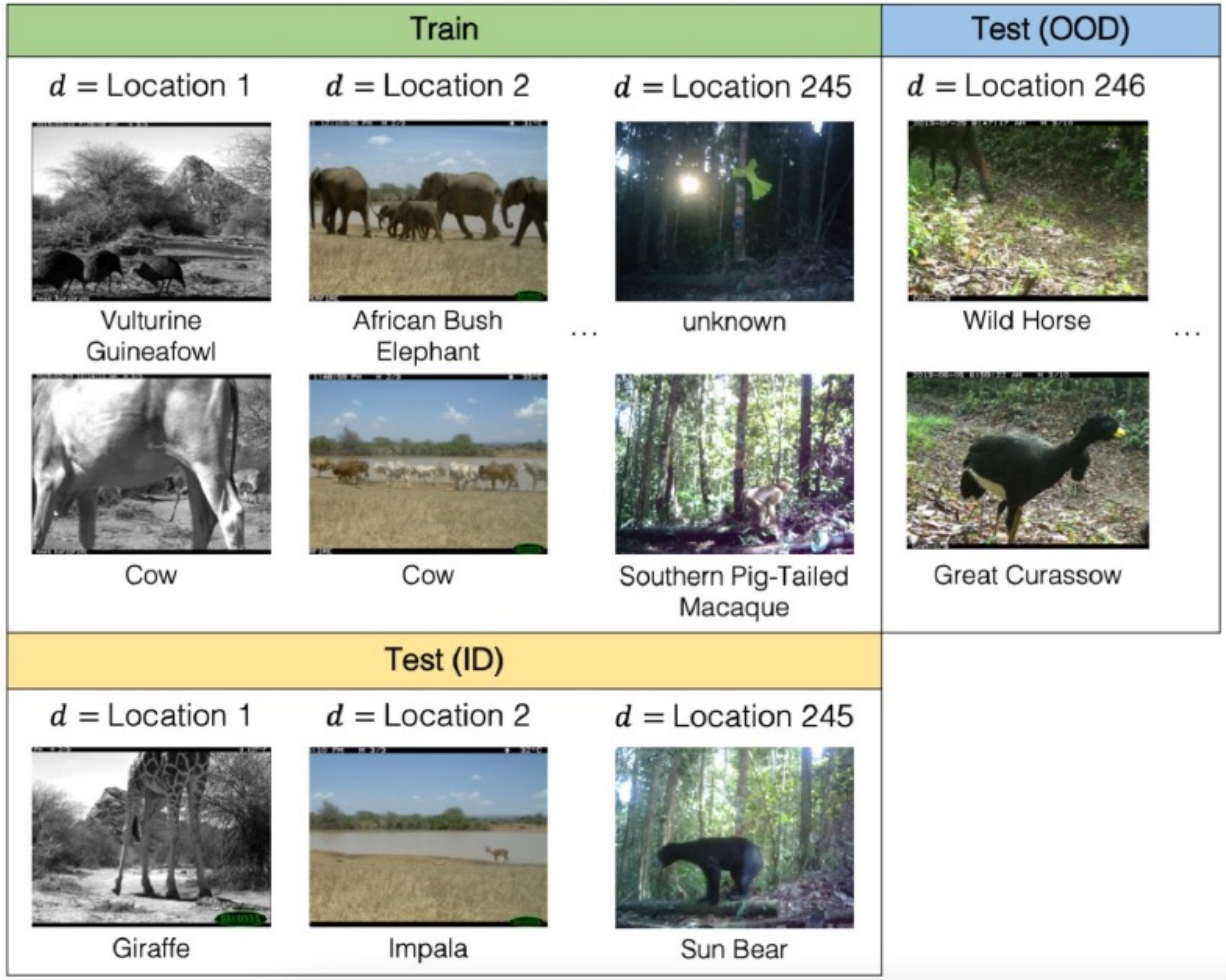
Example: Natural Language Processing

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 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

Jia & Liang, Adversarial Examples for Evaluating Reading Comprehension Systems, 2021

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Example: Real Perturbations



Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

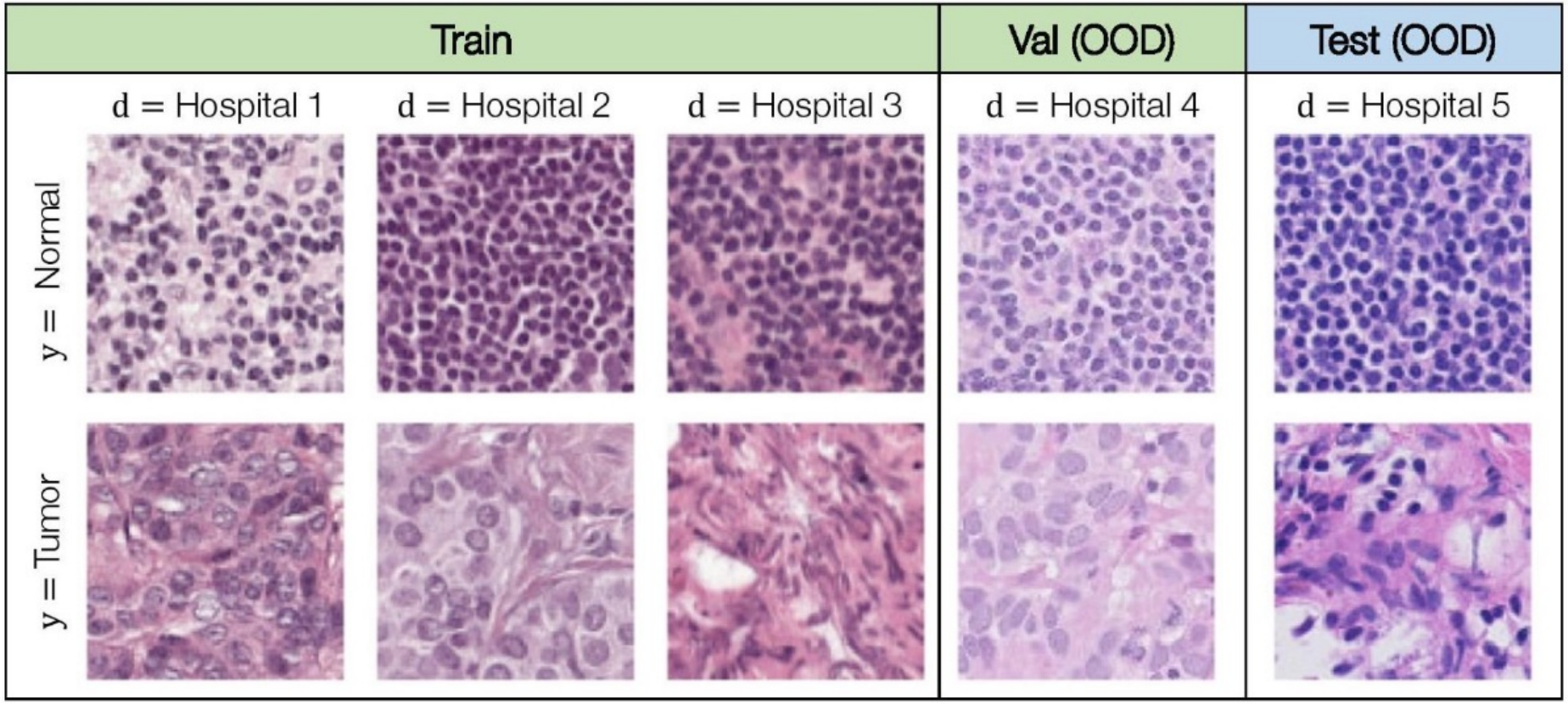
Example: Real Perturbations

	Train			Test	
Satellite Image (x)					
Year / Region (d)	2002 / Americas	2009 / Africa	2012 / Europe	2016 / Americas	2017 / Africa
Building / Land Type (y)	shopping mall	multi-unit residential	road bridge	recreational facility	educational institution

Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Example: Real Perturbations



Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Distribution Shift:
Notation & Problem Formulation

Traditional Supervised Learning

- **Problem setup**
  - Consider a parametric model family $\{ f_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta \}$
  - Consider a loss function $\ell(\theta; x, y)$
  - Consider a data distribution P over $\mathcal{X} \times \mathcal{Y}$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Traditional Supervised Learning

- **Problem setup**
  - Consider a parametric model family $\{ f_\theta : \mathcal{X} \to \mathcal{Y} \mid \theta \in \Theta \}$
  - Consider a loss function $\ell(\theta; x, y)$
  - Consider a data distribution P over $\mathcal{X} \times \mathcal{Y}$

- **Supervised learning problem**
  - Given training dataset $Z \subseteq \mathcal{X} \times \mathcal{Y}$ consisting of i.i.d. samples $(x, y) \sim P$
  - Expected/empirical loss $\mathbb{E}_P[\ell(\theta; x, y)] \approx |Z|^{-1} \sum_{(x,y) \in Z} \ell(\theta; x, y)$
  - Goal is to compute $\hat{\theta} = \min_\theta |Z|^{-1} \sum_{(x,y) \in Z} \ell(\theta; x, y)$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Distribution Shift

- **Distribution shift:** Training and test distributions differ
  - Training set consists of samples $(x_1, y_1), \ldots, (x_n, y_n) \sim P$
  - Test set consists of samples $(x'_1, y'_1), \ldots, (x'_m, y'_m) \sim Q$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Distribution Shift

- **Distribution shift:** Training and test distributions differ
  - Training set consists of samples $(x_1, y_1), \ldots, (x_n, y_n) \sim P$
  - Test set consists of samples $(x'_1, y'_1), \ldots, (x'_m, y'_m) \sim Q$

- **Supervised learning under distribution shift**
  - Given training dataset $Z \subseteq \mathcal{X} \times \mathcal{Y}$ consisting of i.i.d. samples $(x, y) \sim P$
  - Goal is to minimize loss $\mathbb{E}_Q[\ell(\theta; x, y)] \not\approx |Z|^{-1} \sum_{(x,y)\in Z} \ell(\theta; x, y)$
  - Computing $\hat{\theta} = \min_\theta |Z|^{-1} \sum_{(x,y)\in Z} \ell(\theta; x, y)$ may not work

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Aside: Adversarial Robustness

- In adversarial robustness, the goal is to be robust to all perturbations of the form $x' = x + \epsilon$, where ϵ is small but **arbitrary**

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Distribution Shift

- Intuitively, when can we hope to perform well on Q?

- **Impossible in general (what if we swap the labels?)**

- Can we leverage additional information about the shift?
  - Make additional assumptions about shift
  - Leverage additional data
  - Both

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Distributionally Robust Optimization

- **Idea:** Robust to an arbitrary small shift
- **Example:** Small in KL divergence:

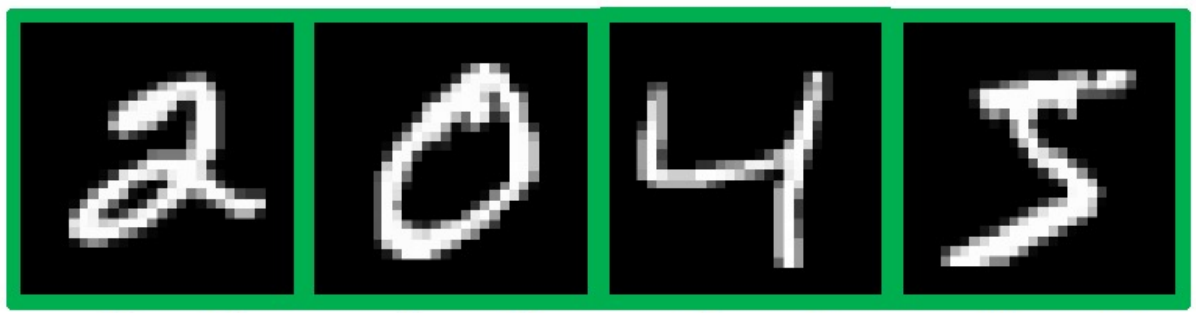
$$\{ Q \mid D_{\mathrm{KL}}(P \parallel Q) \leq \epsilon \}$$

- Very similar to adversarial robustness (covered later)
- We can do much better with a little extra information

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Unsupervised Domain Adaptation

- **Idea:** Use **some** information about the distribution shift
- Consider **unsupervised domain adaptation** setting



$y^* = 2$ $y^* = 0$ $y^* = 4$ $y^* = 5$

labeled training examples

unlabeled test examples
(possibly shifted)

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Unsupervised Domain Adaptation

- Data is easy to collect but labeling costs money
  - **Example:** Data from a different hospital

- Collect data during run time
  - **Example:** Self-driving car

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Types of Distribution Shift: Label Shift

$p(y) \neq q(y)$ but $p(x \mid y) = q(x \mid y)$

Label Shift Assumption

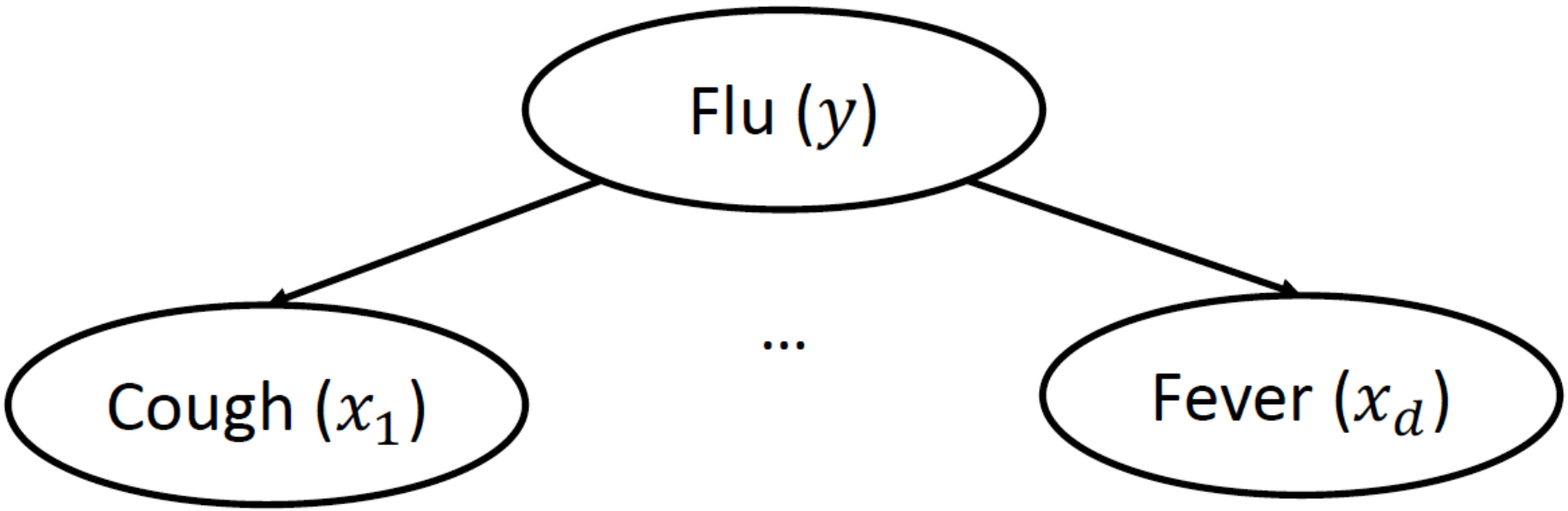
- Let p and q be the density functions for P and Q, respectively

- **Label Shift Assumption:** $p(x \mid y) = q(x \mid y)$
  - But may have $p(y) \neq q(y)$
  - **Intuition:** The rates of labels changes, but the kinds of labels don't

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Label Shift Assumption

- **Example:** Increase in flu cases due to an outbreak
  - x are the symptoms, y is indicator for flu
  - $P(x \mid y)$ is rate of symptoms conditioned on having disease (stays the same)
  - $P(y)$ is rate of flu (can change if there is an outbreak)

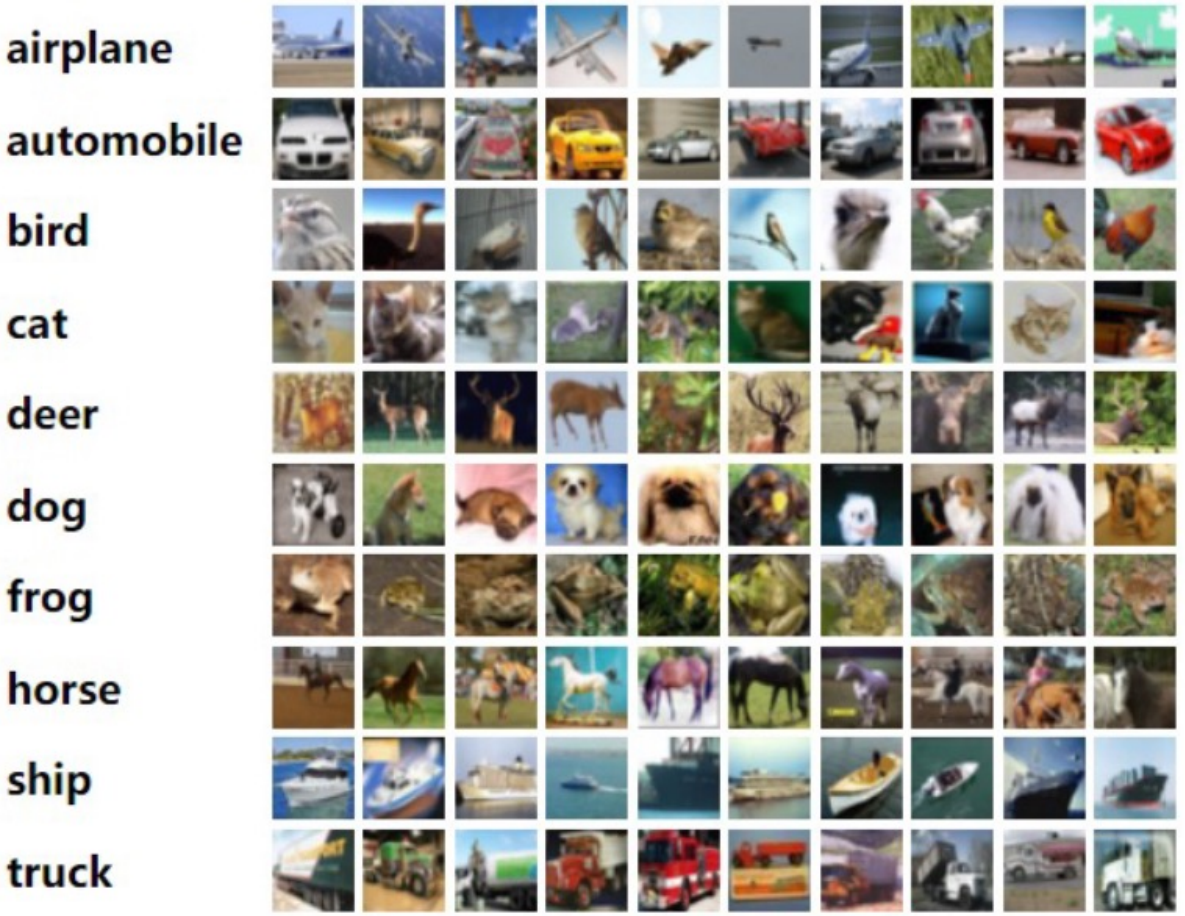


Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Label Shift Assumption

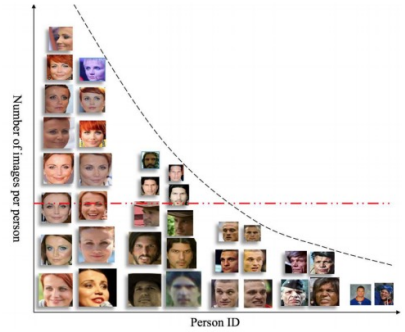
- **Example:** Changes in label distribution
  - x is an image, y is the label
  - $P(x \mid y)$ is the distribution of images of a given label
  - $P(y)$ is rate of that label

- Often, the training labels are balanced, which is a source of label shift

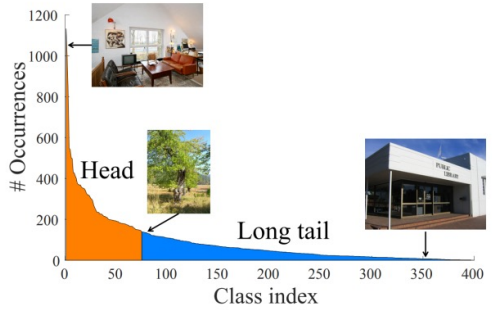


Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

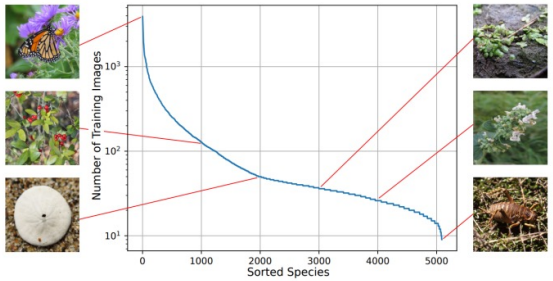
Class Imbalance



Faces [Zhang et al. 2017]



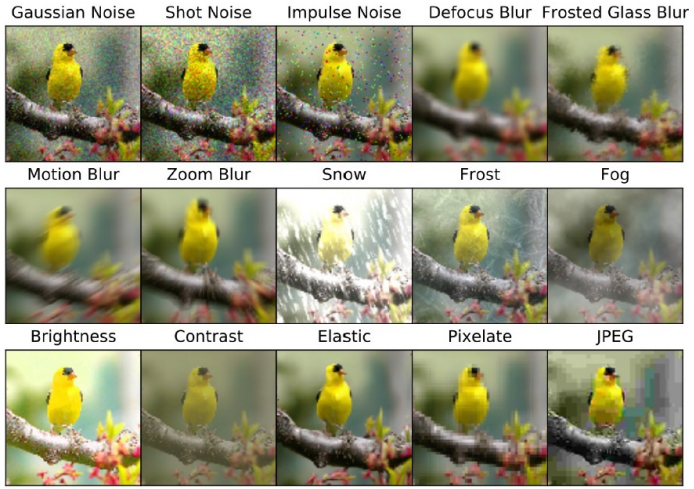
Places [Wang et al. 2017]



Species [Van Horn et al. 2019]

Actions [Zhang et al. 2019]

Source: https://liuziwei7.github.io/papers/longtail_slides.pdf



Hendrycks & Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

Types of Distribution Shift: Covariate Shift

$p(x) \neq q(x)$ but $p(y \mid x) = q(y \mid x)$

Covariate Shift Assumption

- Let p and q be the density functions for P and Q, respectively

- **Covariate Shift Assumption:** $p(y \mid x) = q(y \mid x)$
  - But may have $p(x) \neq q(x)$
  - **Intuition:** The label computation does not change, but the inputs can change

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Covariate Shift Assumption

- Let p and q be the density functions for P and Q, respectively

- **Covariate Shift Assumption:** $p(y \mid x) = q(y \mid x)$
  - But may have $p(x) \neq q(x)$
  - **Intuition:** The label computation does not change, but the inputs can change

- **Examples**
  - $y = \beta^{\top} x + \epsilon$, but $P(x) = N(\mu, \sigma^2)$ while $Q(x) = N(\mu', \sigma'^2)$
  - Daytime vs. nighttime, driving in new city, changes in color/lighting

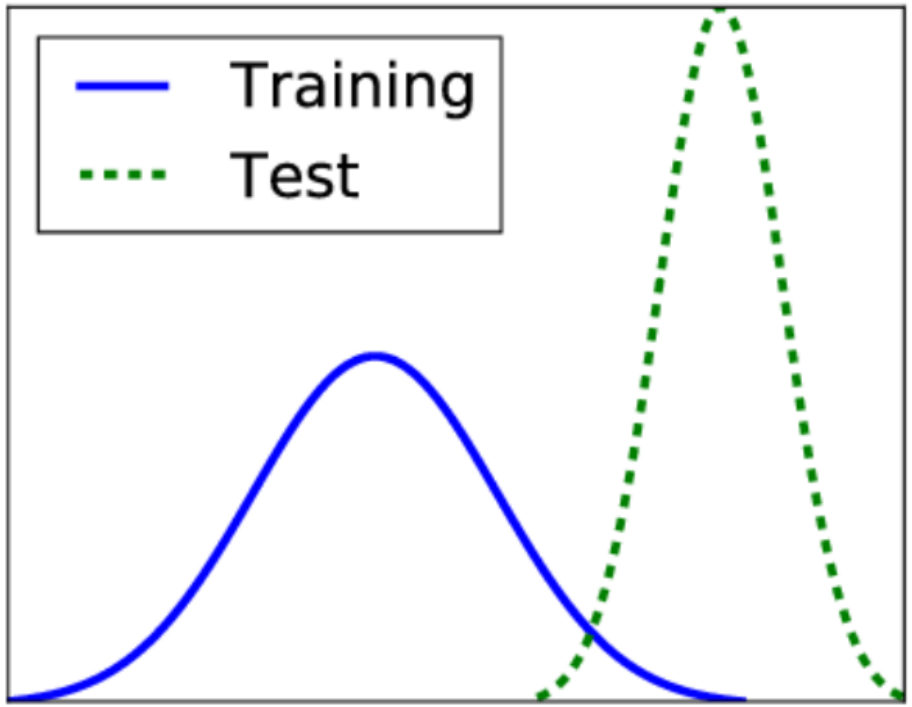
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Covariate Shift Assumption

- $y = \beta^\top x + \epsilon$, but $P(x) = N(\mu, \sigma^2)$ while $Q(x) = N(\mu', \sigma'^2)$

- **Covariate distributions**
  - $P(x) = N(\mu, \sigma^2)$
  - $Q(x) = N(\mu', \sigma'^2)$

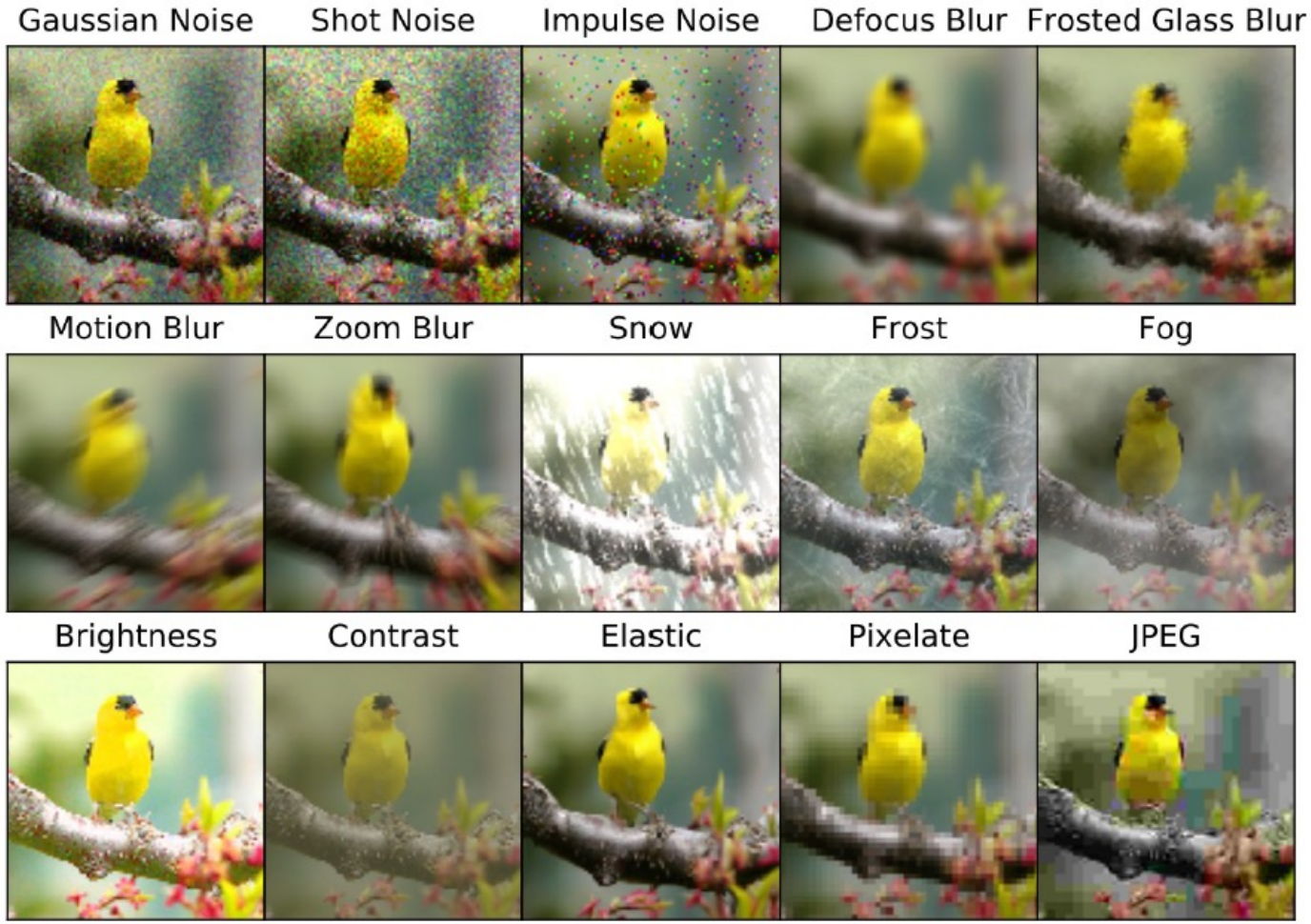
- **Label distribution**
  - $P(y \mid x) = Q(y \mid x) = N(\beta^\top x, \sigma''^2)$
  - I.e., $y = \beta^\top x + \epsilon$, where $\epsilon \sim N(0, \sigma''^2)$



Image; Glauner et al., 2018

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

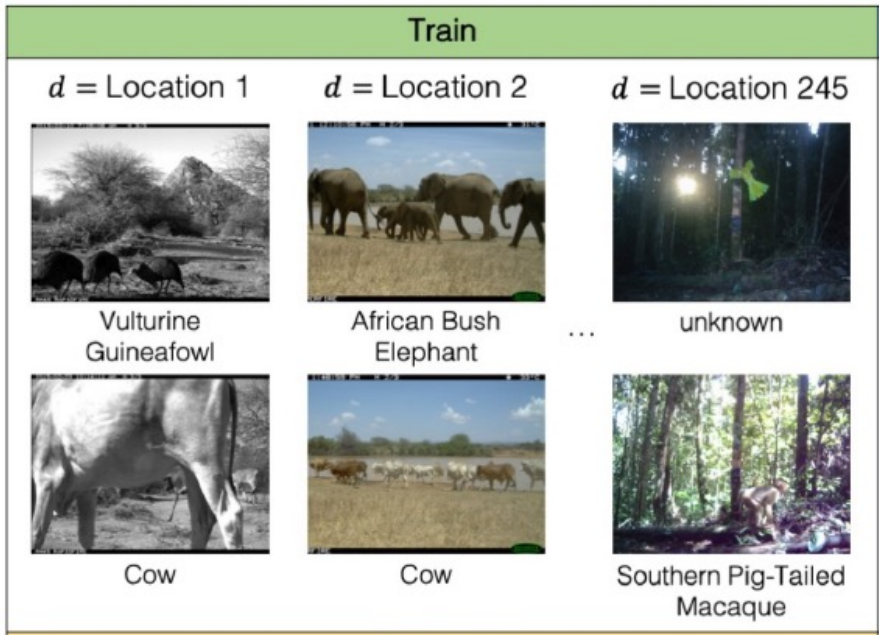
Covariate Shift Assumption



Hendrycks & Dietterich, Benchmarking Neural Network Robustness to Common Corruptions and Perturbations

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

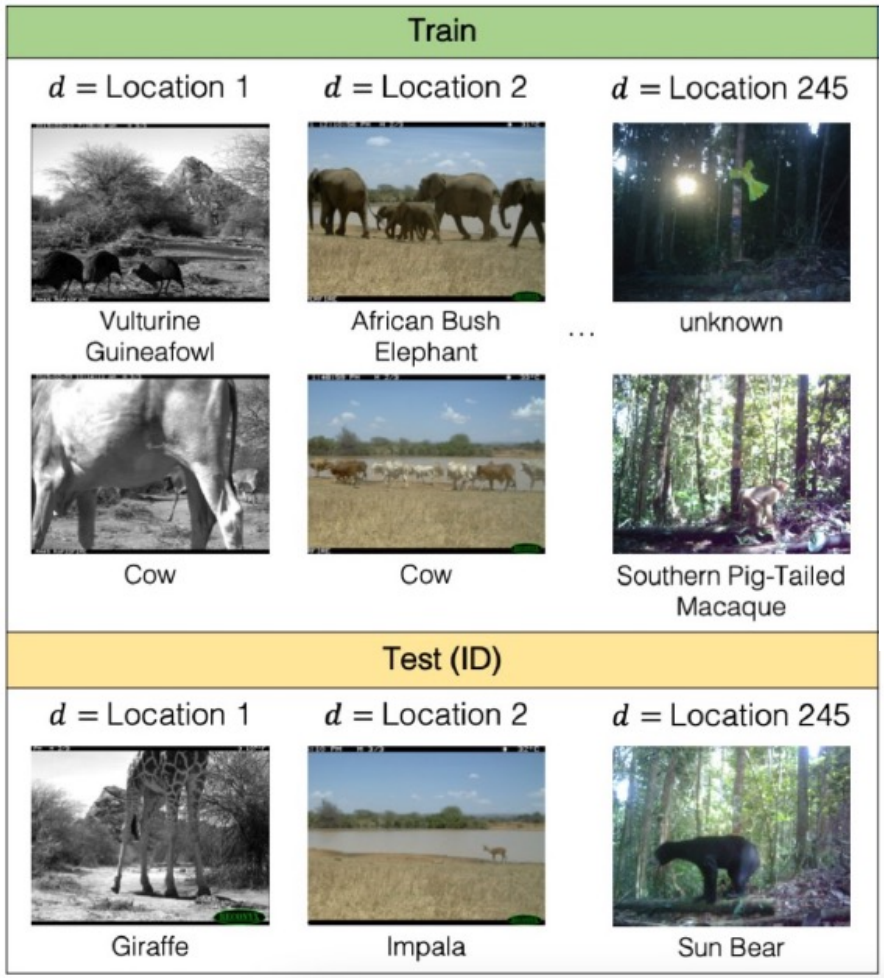
Covariate Shift Assumption



Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

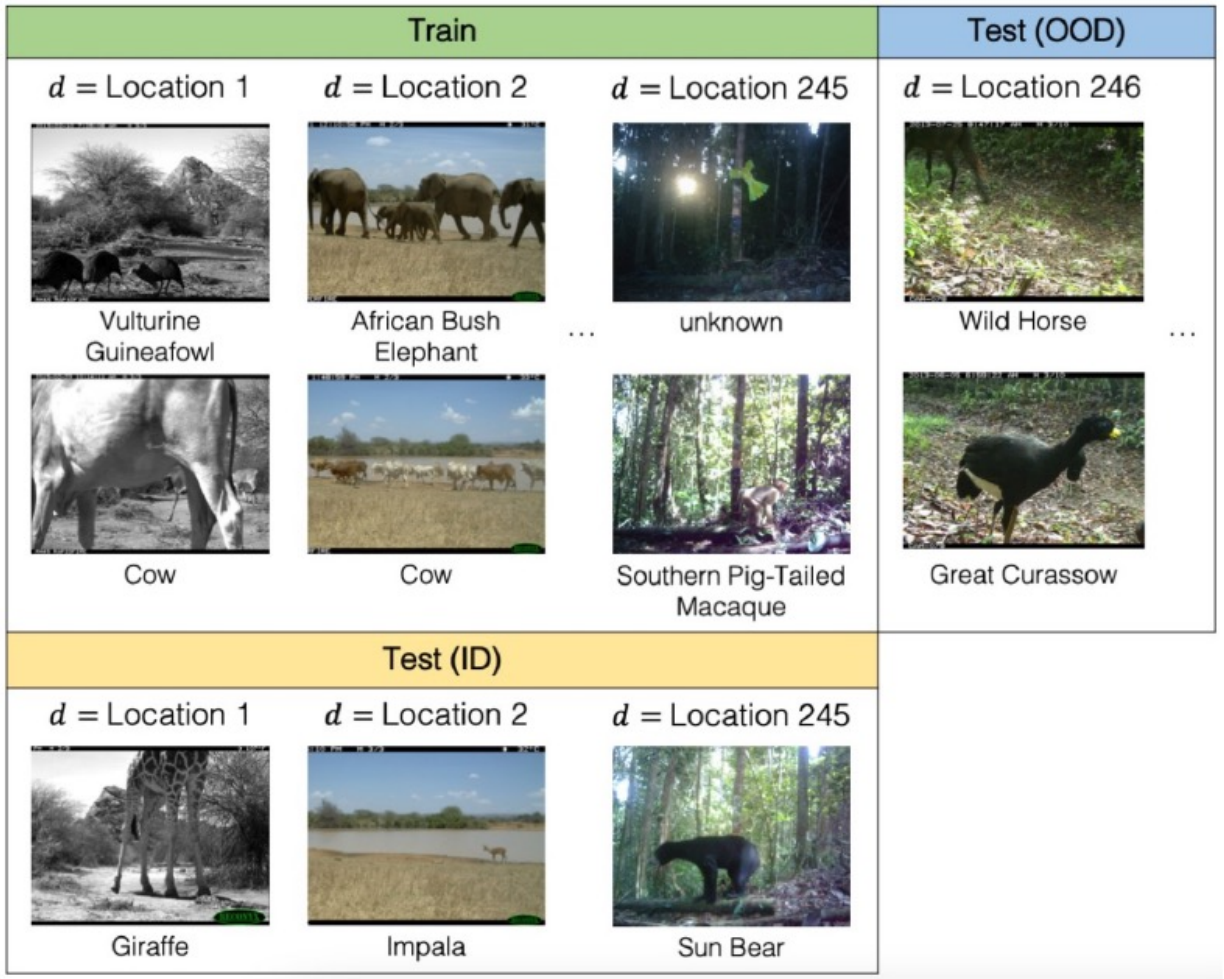
Covariate Shift Assumption



Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Covariate Shift Assumption



Koh et al., WILDS: A Benchmark of in-the-Wild Distribution Shifts, 2020

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Covariate Shift Assumption

- **Computer vision**
  - Daytime vs. nighttime
  - Color shifts, lighting shifts, etc.
  - Driving in a new city
- **Natural language processing**
  - Change in vocabulary frequency over time
  - Regional vocabulary
  - News writing vs. conversational writing
- Covariate shift is pervasive

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Methods for Distributional Robustness

Importance Weighting

Importance Weighting

- Given distributions P and Q, the **importance weight (function)** is

$$w(x, y) = \frac{q(x, y)}{p(x, y)}$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weighting

- Given distributions P and Q, the **importance weight (function)** is

$$w(x, y) = \frac{q(x, y)}{p(x, y)}$$

- **Key property (by definition):**

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)]$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weighting

- Note that

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(\theta; x, y) \cdot q(x, y) \cdot dx \cdot dy$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weighting

- Note that

$$\begin{aligned}
\mathbb{E}_Q[\ell(\theta; x, y)] &= \int_{\mathcal{X}\times\mathcal{Y}} \ell(\theta; x, y) \cdot q(x, y) \cdot dx \cdot dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(\theta; x, y) \cdot \frac{q(x,y)}{p(x,y)} \cdot p(x, y) \cdot dx \cdot dy \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \ell(\theta; x, y) \cdot w(x, y) \cdot p(x, y) \cdot dx \cdot dy \\
&= \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)]
\end{aligned}$$

- We have assumed the support of Q is contained in the support of P!

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weighting

- Given distributions P and Q, the **importance weight (function)** is

$$w(x, y) = \frac{q(x, y)}{p(x, y)}$$

- **Key property (by definition):**

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)]$$

- **Key question:** How to compute importance weights?

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

- In the label shift setting, we have

$$
\begin{aligned}
w(x,y) &= \frac{q(x,y)}{p(x,y)} \\
&= \frac{q(x|y)q(y)}{p(x|y)p(y)} \\
&= \frac{q(y)}{p(y)} \\
&:= w(y)
\end{aligned}
$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

- If we know $w(y)$, then we have

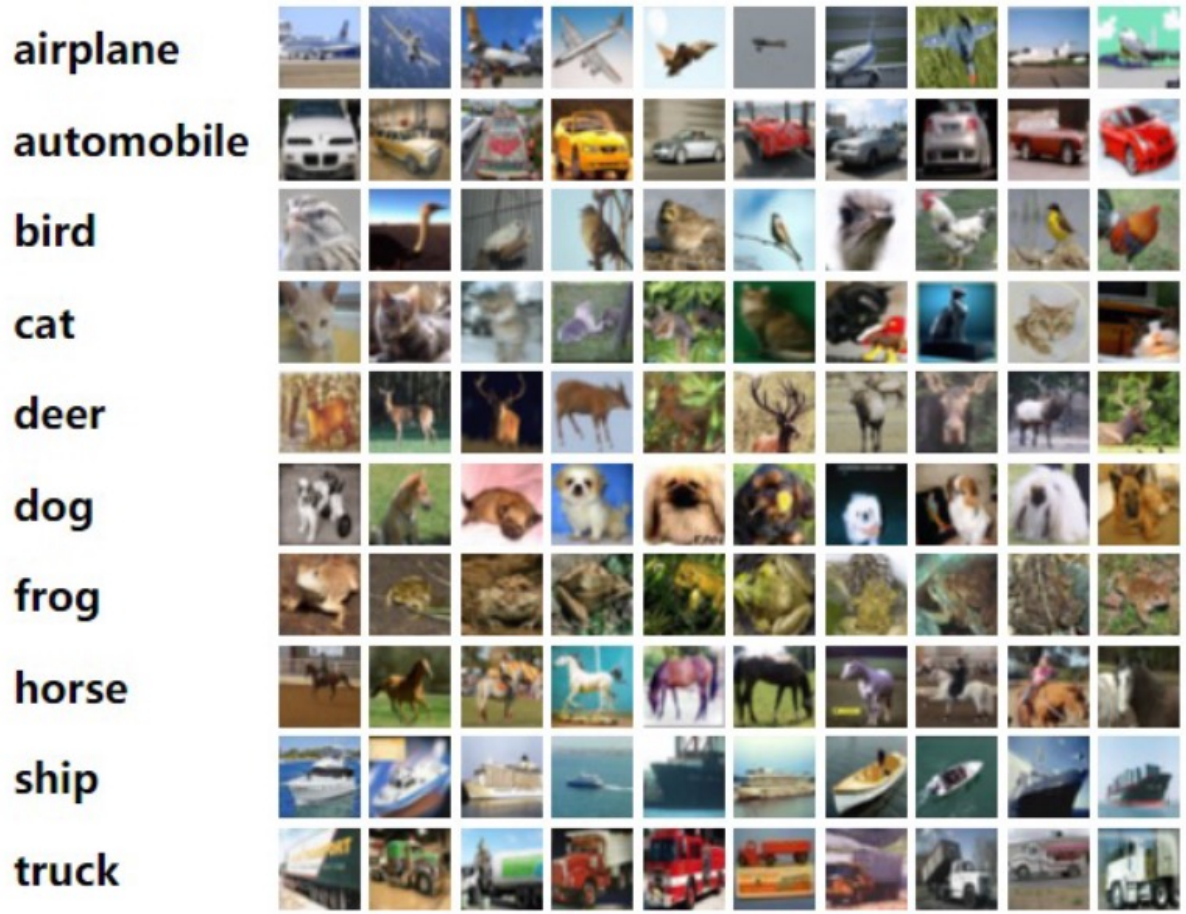
$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(y)]$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Label Shift Assumption

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(y)]$$

- **Training:** $p(y) = \frac{1}{10}$
- **Test:** $q(\text{automobile}) = \frac{1}{2}$ and $q(y) = \frac{1}{18}$ otherwise
- Then, the loss might be



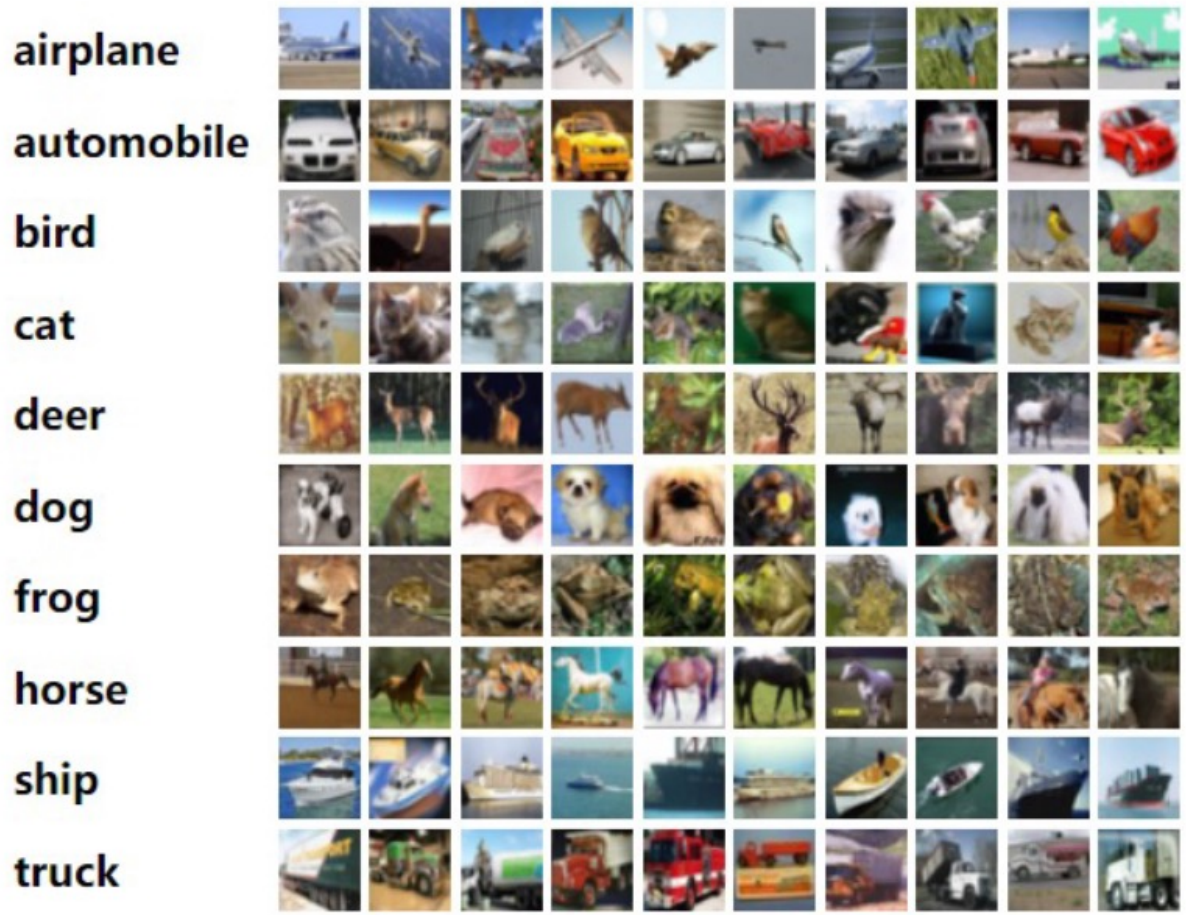
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Label Shift Assumption

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(y)]$$

- **Training:** $p(y) = \frac{1}{10}$
- **Test:** $q(\text{automobile}) = \frac{1}{2}$ and $q(y) = \frac{1}{18}$ otherwise
- Then, the loss might be

$$\frac{10}{18} \cdot \ell(x_1, \text{dog}) + \frac{10}{2} \cdot \ell(x_2, \text{automobile}) + \frac{10}{18} \cdot \ell(x_3, \text{frog})$$



Intuition: Automobile class was underrepresented in the source distribution. We give higher weight to its loss calculation to compensate.

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

- If we know $w(y)$, then we have

$$\mathbb{E}_Q[\ell(\theta; x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(x, y)] = \mathbb{E}_P[\ell(\theta; x, y) \cdot w(y)]$$

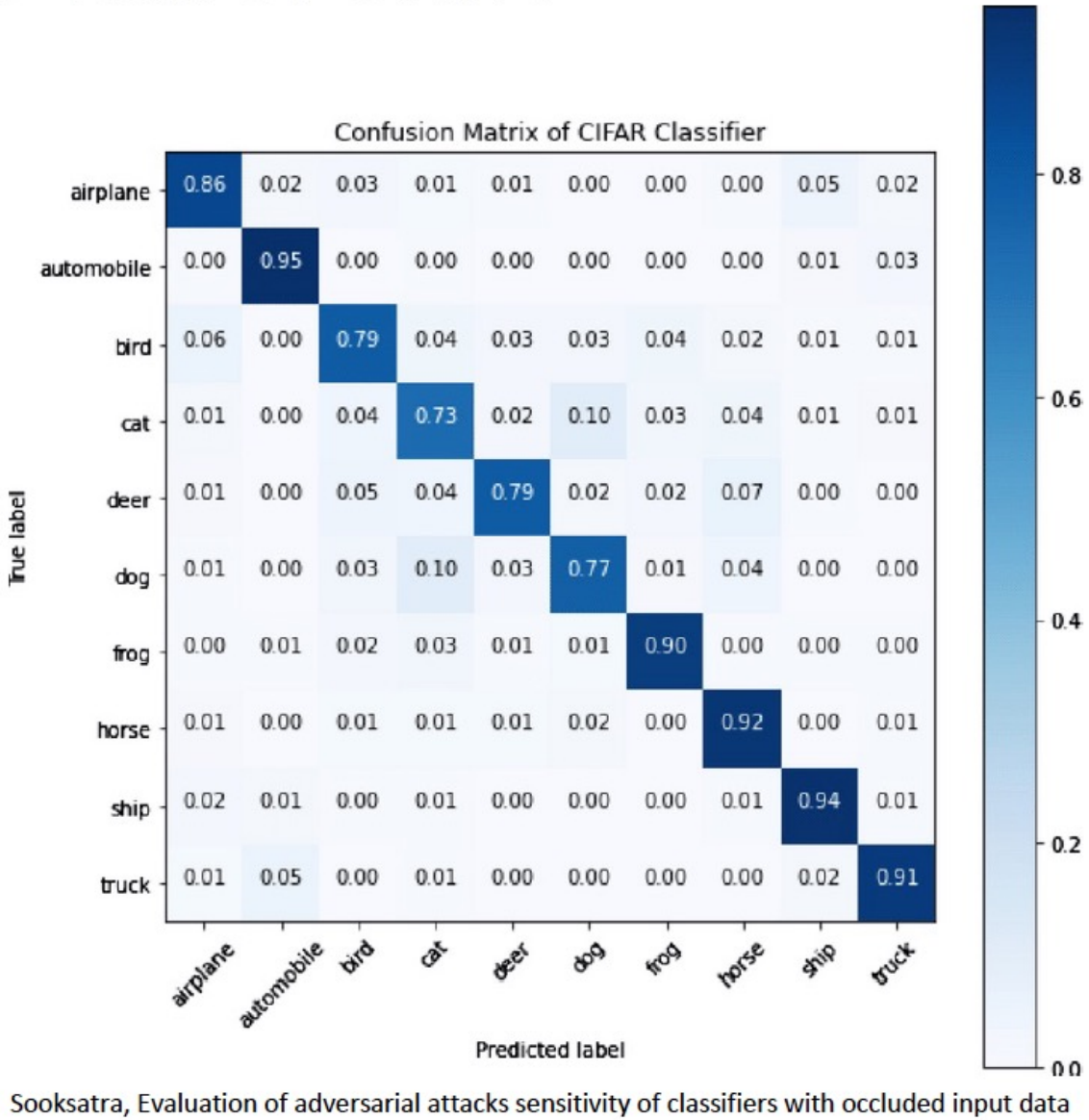
- How do we compute $w(y)$?

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

- Given a classifier $f: \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$, consider the confusion matrix $C \in \mathbb{R}^{K \times K}$ defined by

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j]$$



Sooksatra, Evaluation of adversarial attacks sensitivity of classifiers with occluded input data

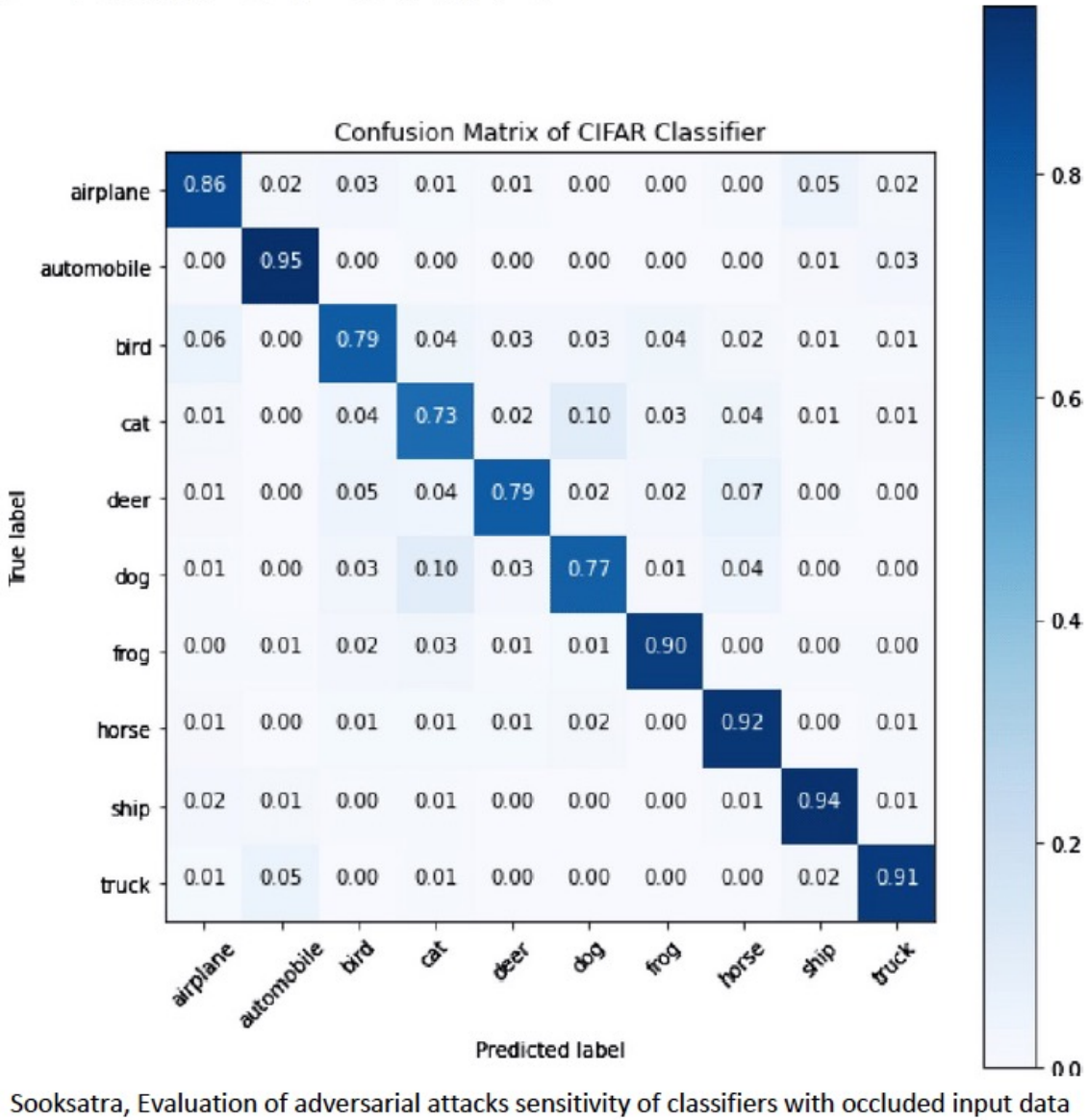
Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

- Given a classifier $f: \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} = \{1, \ldots, K\}$, consider the confusion matrix $C \in \mathbb{R}^{K \times K}$ defined by

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j]$$

- Also, define $p, q \in \mathbb{R}^K$ by
  - $p_i = \mathbb{P}_P[f(x) = i]$
  - $q_i = \mathbb{P}_Q[f(x) = i]$



Sooksatra, Evaluation of adversarial attacks sensitivity of classifiers with occluded input data

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

- Since $f(x)$ only depends on x, we have

$$\mathbb{P}_P[f(x) = i \mid y = j] = \int_X 1(f(x) = i) \cdot p(x \mid y = j) \cdot dx$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

- Since $f(x)$ only depends on x, we have

$$\begin{aligned}\mathbb{P}_P[f(x) = i \mid y = j] &= \int_X 1(f(x) = i) \cdot p(x \mid y = j) \cdot dx \\ &= \int_X 1(f(x) = i) \cdot q(x \mid y = j) \cdot dx \\ &= \mathbb{P}_Q[f(x) = i \mid y = j]\end{aligned}$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$q_i = \sum_{j=1}^{k} \mathbb{P}_Q[f(x) = i \mid y = j] \cdot \mathbb{P}_Q[y = j]$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$\begin{aligned} q_i &= \sum_{j=1}^{k} \mathbb{P}_Q[f(x) = i \mid y = j] \cdot \mathbb{P}_Q[y = j] \\ &= \sum_{j=1}^{k} \mathbb{P}_P[f(x) = i \mid y = j] \cdot \mathbb{P}_Q[y = j] \\ &= \sum_{j=1}^{k} \frac{\mathbb{P}_P[f(x)=i, y=j]}{\mathbb{P}_P[y=j]} \cdot \mathbb{P}_Q[y = j] \\ &= \sum_{j=1}^{k} \mathbb{P}_P[f(x) = i, y = j] \cdot \frac{\mathbb{P}_Q[y=j]}{\mathbb{P}_P[y=j]} \end{aligned}$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$q_i = \sum_{j=1}^{k} \mathbb{P}_P[f(x) = i, y = j] \cdot \frac{\mathbb{P}_Q[y=j]}{\mathbb{P}_P[y=j]}$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$q_i = \sum_{j=1}^{k} \mathbb{P}_P[f(x) = i, y = j] \cdot w(j)$$

$$\begin{bmatrix} q_1 \\ \vdots \\ q_K \end{bmatrix} = \begin{bmatrix} \mathbb{P}_P[f(x) = 1, y = 1] & \cdots & \mathbb{P}_P[f(x) = 1, y = K] \\ \vdots & \ddots & \vdots \\ \mathbb{P}_P[f(x) = K, y = 1] & \cdots & \mathbb{P}_P[f(x) = K, y = K] \end{bmatrix} \begin{bmatrix} w(1) \\ \vdots \\ w(K) \end{bmatrix}$$

$$q = Cw \Rightarrow w = C^{-1}q$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$q_i = \sum_{j=1}^{k} \mathbb{P}_P[f(x) = i, y = j] \cdot w(j)$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Importance Weights for Label Shift

$$C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \quad p_i = \mathbb{P}_P[f(x) = i] \quad q_i = \mathbb{P}_Q[f(x) = i]$$

$$\mathbb{P}_P[f(x) = i \mid y = j] = \mathbb{P}_Q[f(x) = i \mid y = j]$$

- Now, we have

$$w = C^{-1}q$$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000

Supervised Learning with Label Shift

- **Input:** Training dataset Z, unlabeled test dataset X
- **Step 1:** Train f on Z
- **Step 2:** Estimate using the dataset:
  - $C_{ij} = \mathbb{P}_P[f(x) = i, y = j] \approx |Z|^{-1} \sum_{(x,y)\in Z} 1(f(x) = i \wedge y = j)$
  - $q_i = \mathbb{P}_Q[f(x) = i] \approx |X|^{-1} \sum_{x\in X} 1(f(x) = i)$
- **Step 3:** Compute $w = C^{-1}q$
- **Step 4:** Compute $\hat{\theta} = \arg\min_{\theta} \sum_{(x,y)\in Z} \ell(\theta; x, y) \cdot w(y)$

Slide: Uni of Pennsylvania, Trustworthy ML - CIS 7000