PETUUM

Carnegie Mellon University
School of Computer Science

# Probabilistic Graphical Models

## A Brief Overview of Trustworthy Machine Learning

Haohan Wang

Lecture 28, April 29, 2020

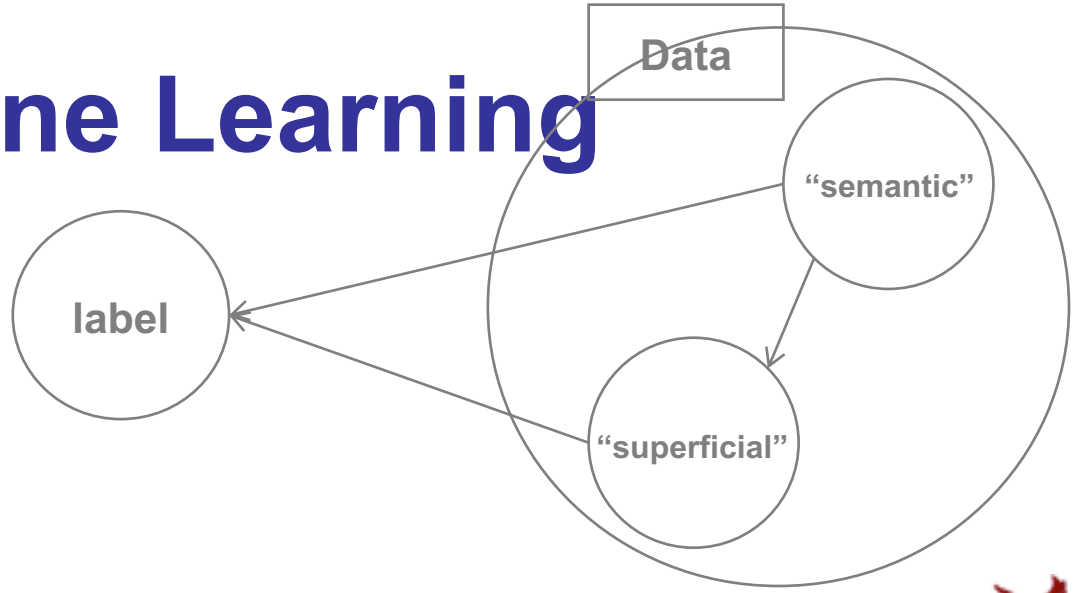

 

# A Brief Overview of Trustworthy Machine Learning

- New Challenges of Modern Machine Learning
  - Empirical Observations
  - Trade-off between Accuracy and Robustness
- Cross-domain Robust Models
  - Generalization Analysis of Domain Adaptation
  - Method Overview – It's mostly about Invariance
  - Domain Generalization and Beyond
- Adversarial Robust Models
  - The Attack vs. Defense Arm Race Highlights
  - Adversarial Training and Its Recent Developments
  - Certified Robustness and Generalization Analysis



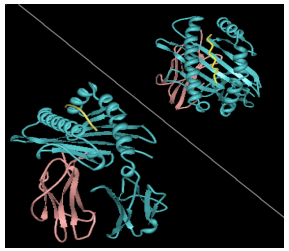

# New Challenges of Modern Machine Learning






# What if I told you…


- We can predict whether you use chopsticks from your genome.
  - HLA-A1 gene
  - (Vilhjálmsson and Nordborg, 2013)

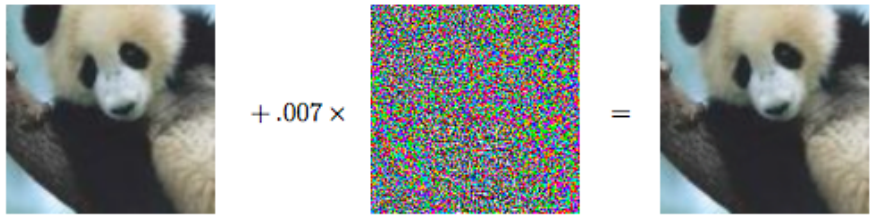


- We can predict whether you speak English or Welsh from your genome
  - (Weale et al 2002)

- Trustworthy statistical models research dates to decades ago…
  - E.g., population stratification in GWAS (Devlin, B. and Roeder, K. 1999)

# Behind the Glory of Deep Learning

- Adversarial example
  - (Szegedy et al 2013)




- CNN's tendency in superficial statistics
  - (Jo and Bengio 2017)
  - (Geirhos et al 2019)

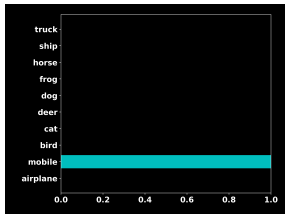
(a) Texture image
81.4% **Indian elephant**
10.3% indri
8.2% black swan

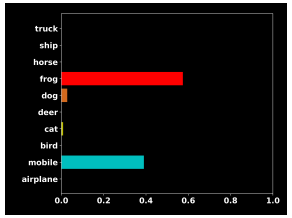
(b) Content image
71.1% **tabby cat**
17.3% grey fox
3.3% Siamese cat

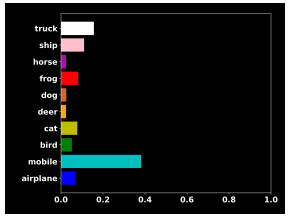
(c) Texture-shape cue conflict
63.9% **Indian elephant**
26.4% indri
9.6% black swan

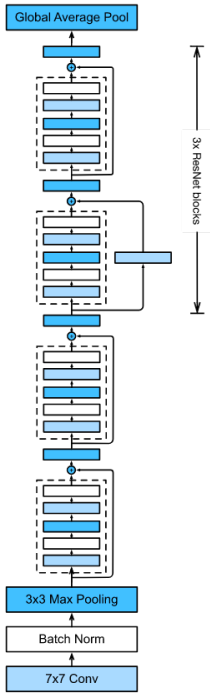
# The Misalignment between CNN and Human

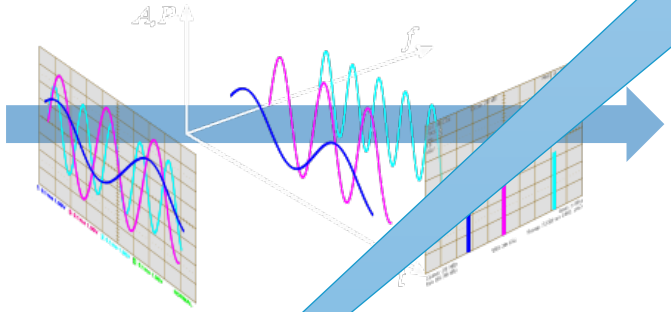
- CNN captures high-frequency information
  - (Wang et al, 2020)



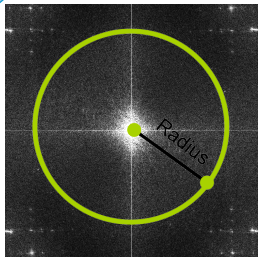



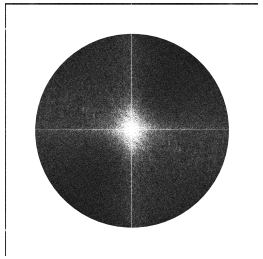



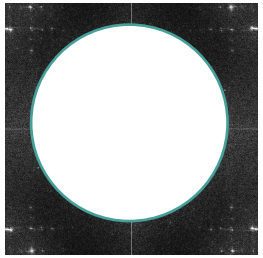



Reconstruction

Reconstruction

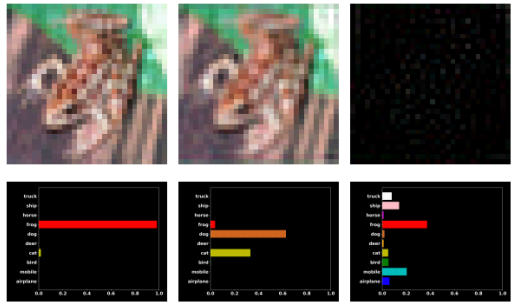


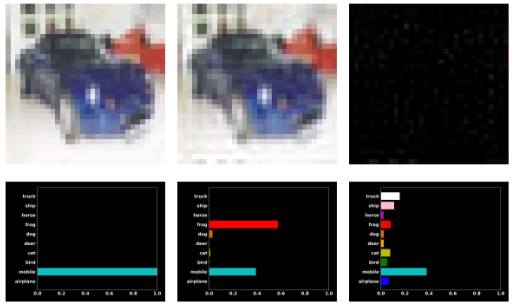


# The Misalignment between CNN and Human

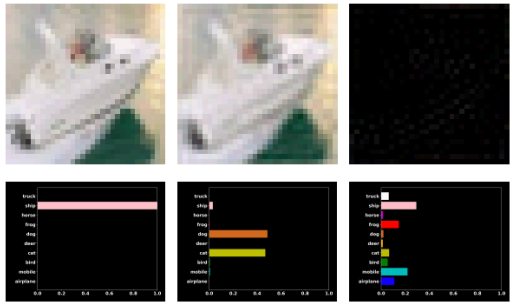
- CNN captures high-frequency information
  - (Wang et al, 2020)



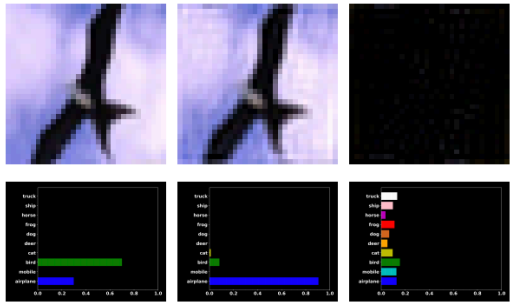
(a) A sample of frog



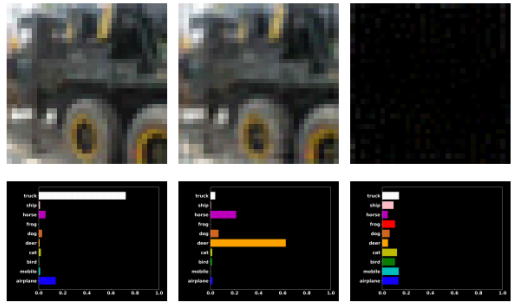
(b) A sample of mobile



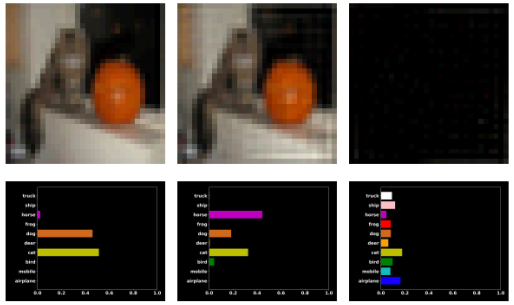
(c) A sample of ship



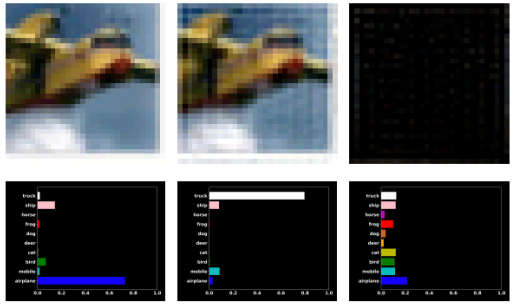
(d) A sample of bird



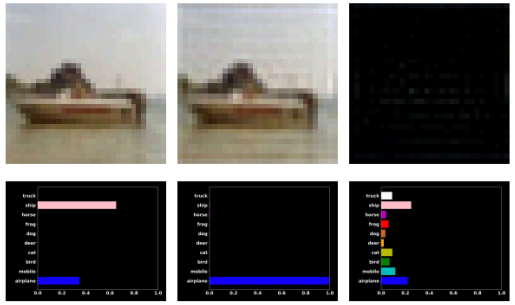
(e) A sample of truck



(f) A sample of cat



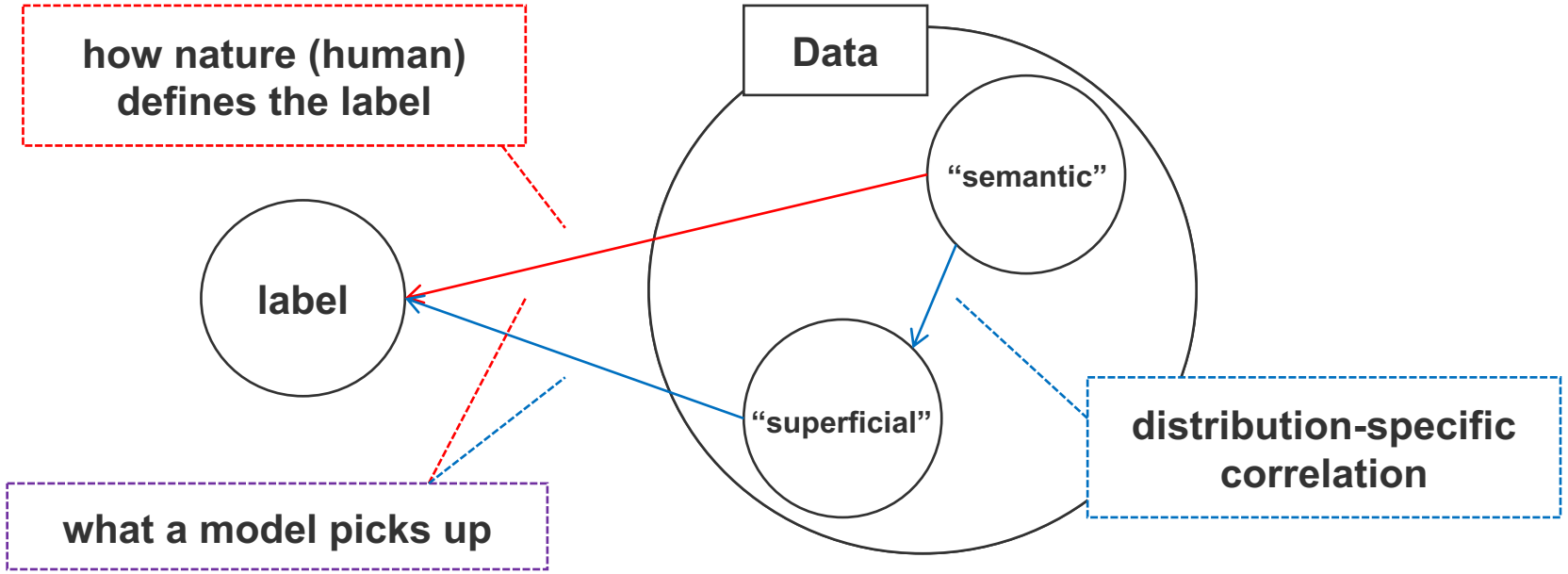
(g) A sample of airplane



(h) A sample of ship

# The New Challenge of Modern Machine Learning

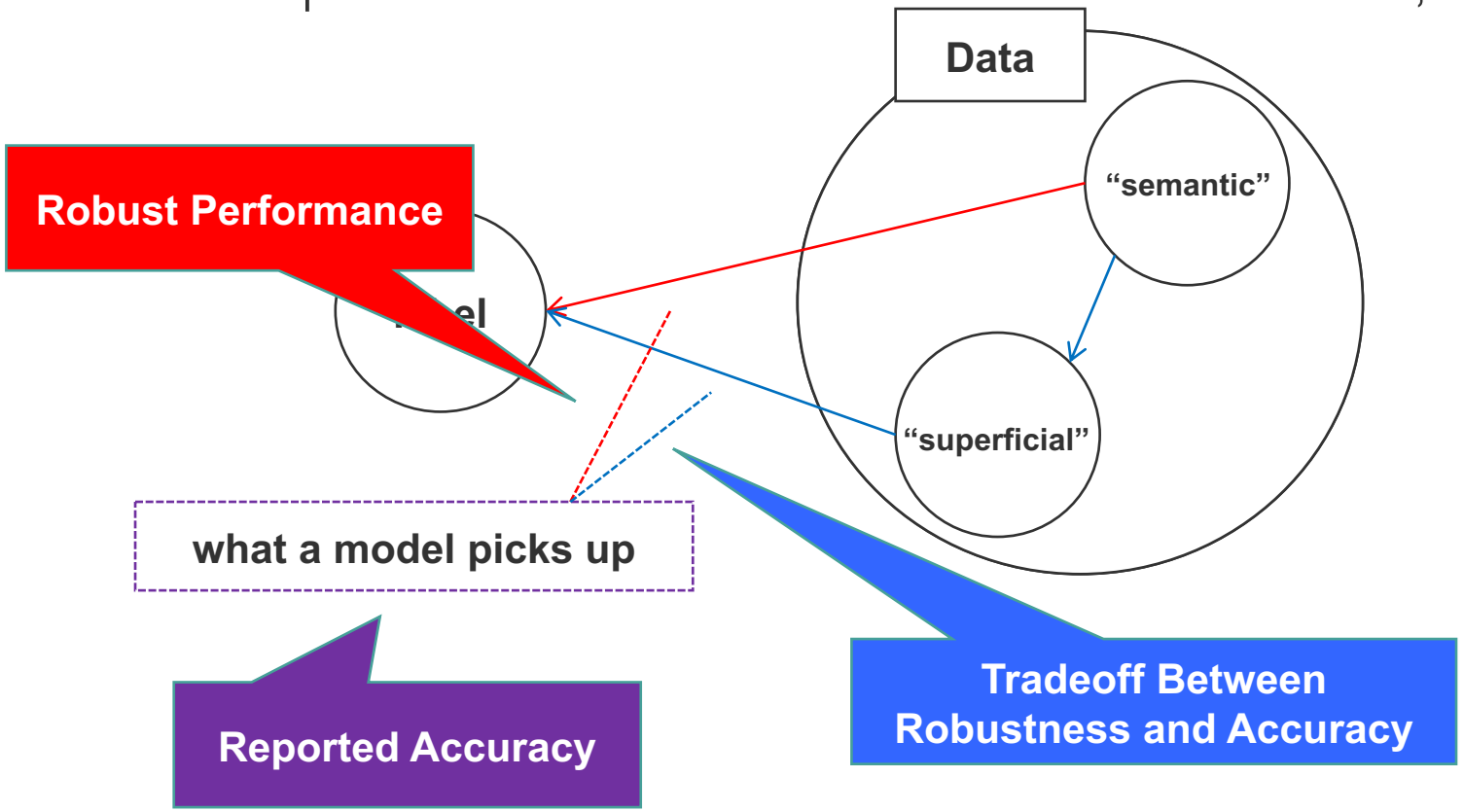
- The misalignment between models and the nature/human
  - (Wang et al, 2020)




**The central problem of today's lecture**

# The Tradeoff between Accuracy and Robustness

- It depends on how one defines "robustness", but roughly




- Formal Discussions
  - Tsipras *et al* 2019
  - Zhang *et al* 2019
  - Yang *et al* 2019
  - Wang *et al* 2020

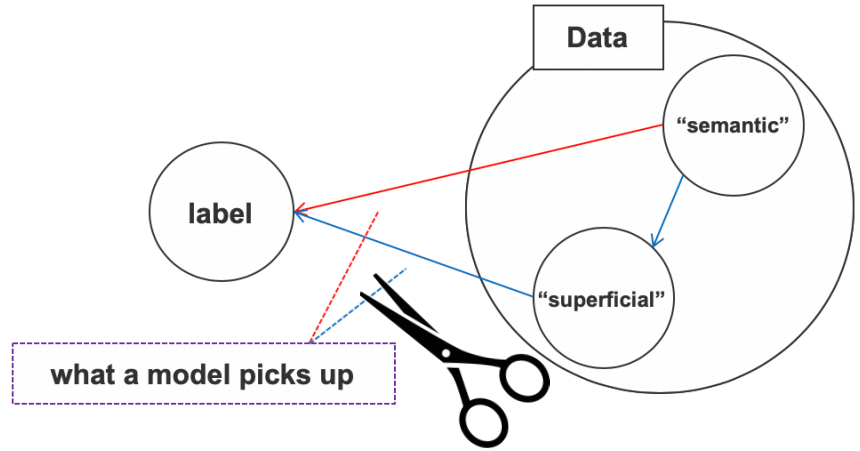
# Why robustness is valued when we have accuracy

- What the current academy is good at

# Why robustness is valued when we have accuracy

- What the world needs

# The Solution




- To summarize it in one sentence
  - Machine learning is about generalization

$$R(h) \leq R_{\text{emp}}(h) + f(|H|, N, \delta)$$

**hypothesis space complexity**

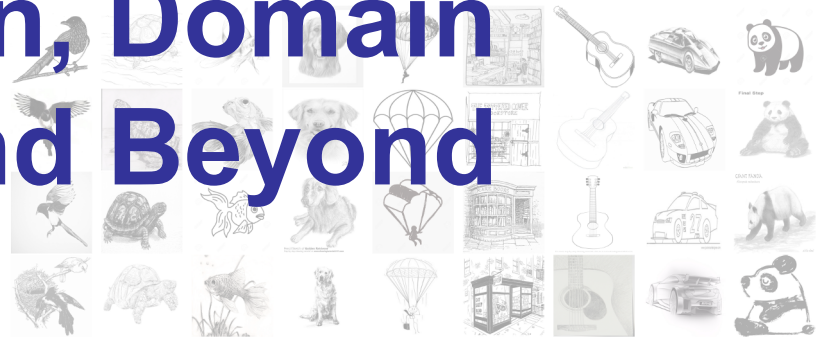
**number of samples**

- Cross-domain Robustness
  - Almost all about hypothesis space (*i.e.*, inductive bias, regularization)
- Adversarial Robustness
  - The dominant solution increases the number of samples

**The central solution of today's lecture**

ImageNet

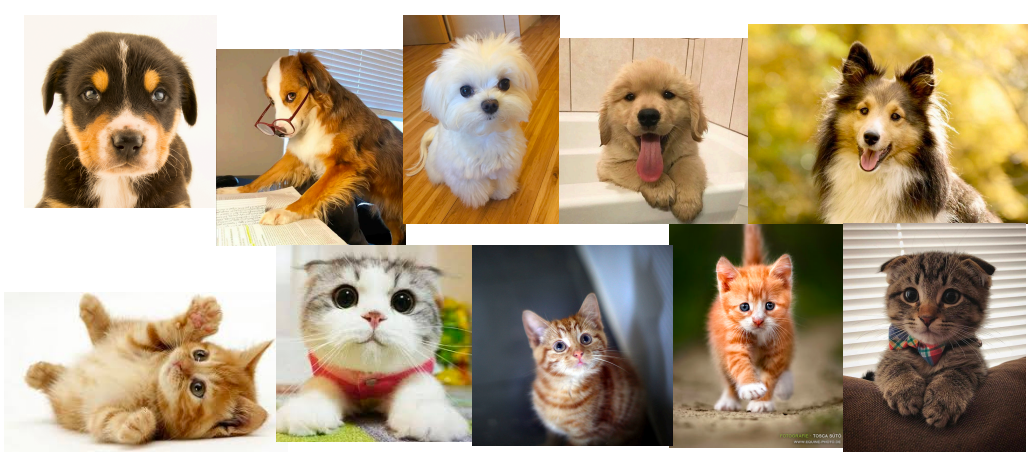
# Cross-Domain Robustness: Domain Adaptation, Domain Generalization, and Beyond



ImageNet-Sketch

# Domain Adaptation

- Model is trained on one distribution/domain, but tested on a “similar but different” distribution/domain
  - Supervised domain adaptation
    - Some labelled samples in the test distribution are available during training
  - Unsupervised domain adaptation
    - Only unlabeled samples in the test distribution are available during training




During Training

During Testing

# How is the generalization changed in UDA

- $\mathcal{H}$-divergence: a measure of two distributions given a hypothesis class
  - $I(h)$: the set of hypotheses that are characteristic functions for the domain
  - $\mathcal{H}$-divergence between the two distributions

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = 2 \sup_{h \in \mathcal{H}} \left| \Pr_{\mathcal{D}} [I(h)] - \Pr_{\mathcal{D}'} [I(h)] \right|$$

  - Can be estimated from finite samples
    - $U$ and $U'$ are samples of size $m$:

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') \leq \hat{d}_{\mathcal{H}}(\mathcal{U}, \mathcal{U}') + 4\sqrt{\frac{d \log(2m) + \log(\frac{2}{\delta})}{m}}$$

    - Estimated from a domain classifier
      - For a symmetric hypothesis class both $h$ and $1$-$h$ are in the hypothesis class

$$\hat{d}_{\mathcal{H}}(\mathcal{U}, \mathcal{U}') = 2\left(1 - \min_{h \in \mathcal{H}} \left[ \frac{1}{m} \sum_{\mathbf{x}: h(\mathbf{x})=0} I[\mathbf{x} \in \mathcal{U}] + \frac{1}{m} \sum_{\mathbf{x}: h(\mathbf{x})=1} I[\mathbf{x} \in \mathcal{U}'] \right]\right)$$



# How is the generalization changed in UDA

- symmetric difference hypothesis space

$$g \in \mathcal{H}\Delta\mathcal{H} \iff g(\mathbf{x}) = h(\mathbf{x}) \oplus h'(\mathbf{x}) \quad \text{for some } h, h' \in \mathcal{H},$$

$$d_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{D}_S, \mathcal{D}_T) = 2 \sup_{h,h' \in \mathcal{H}} \left| \Pr_{x \sim \mathcal{D}_S}\left[h(x) \neq h'(x)\right] - \Pr_{x \sim \mathcal{D}_T}\left[h(x) \neq h'(x)\right] \right|$$

- Generalization analysis
  - (Ben-David et al 2010)

A hypothesis space that is small enough so that the model does not differentiate source and target distribution

$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda$$

$$\lambda = \epsilon_S(h^*) + \epsilon_T(h^*)$$

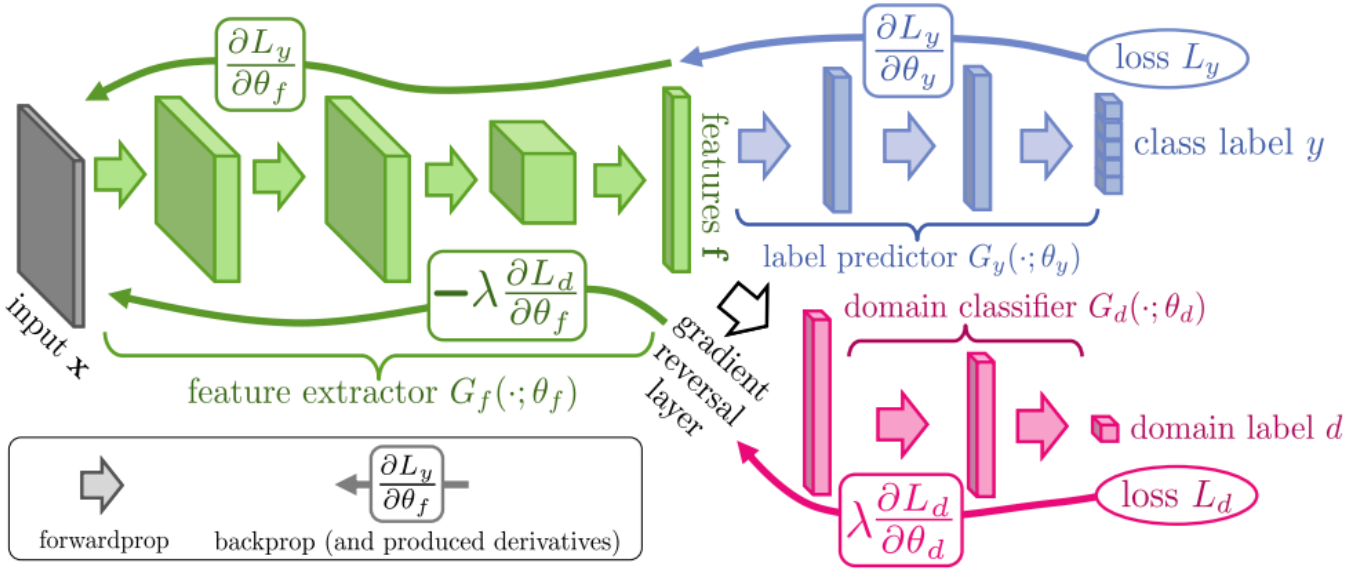
$$h* = \operatorname*{argmin}_{h \in \mathcal{H}} \epsilon_S(h) + \epsilon_T(h)$$

A hypothesis space that is large enough so that we have small errors

# Domain Adversarial Neural Network

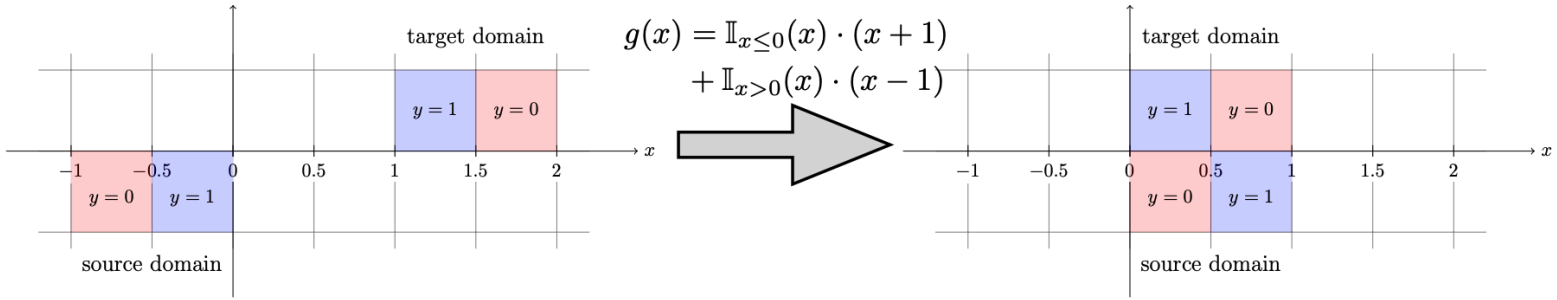
$$\epsilon_T(h) \leq \epsilon_S(h) + \frac{1}{2}\hat{d}_{\mathcal{H}\Delta\mathcal{H}}(\mathcal{U}_S, \mathcal{U}_T) + 4\sqrt{\frac{2d\log(2m') + \log(\frac{2}{\delta})}{m'}} + \lambda$$

- Find a classifier that
  - has small error on the source distribution
  - unable to differentiate source and target distributions
  - (Ganin et al 2016)



# Invariance regularization is not panacea

- Invariant representation and small source risk are not sufficient
  - (Zhao et al, 2019)




  - Labelling function also matters

$$\varepsilon_T(h) \leq \varepsilon_S(h) + d_{\tilde{\mathcal{H}}}(\mathcal{D}_S, \mathcal{D}_T) + \min\{\mathbb{E}_{\mathcal{D}_S}[|f_S - f_T|], \mathbb{E}_{\mathcal{D}_T}[|f_S - f_T|]\}$$

Distance between labelling functions

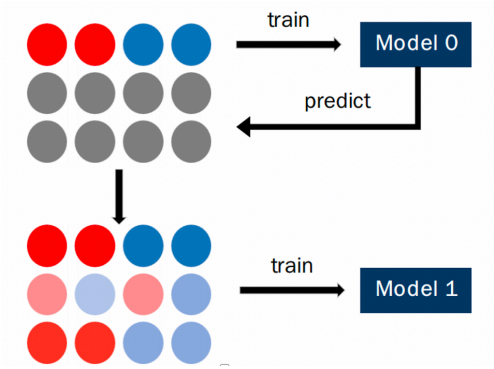
# Domain Adaptation Method Highlights

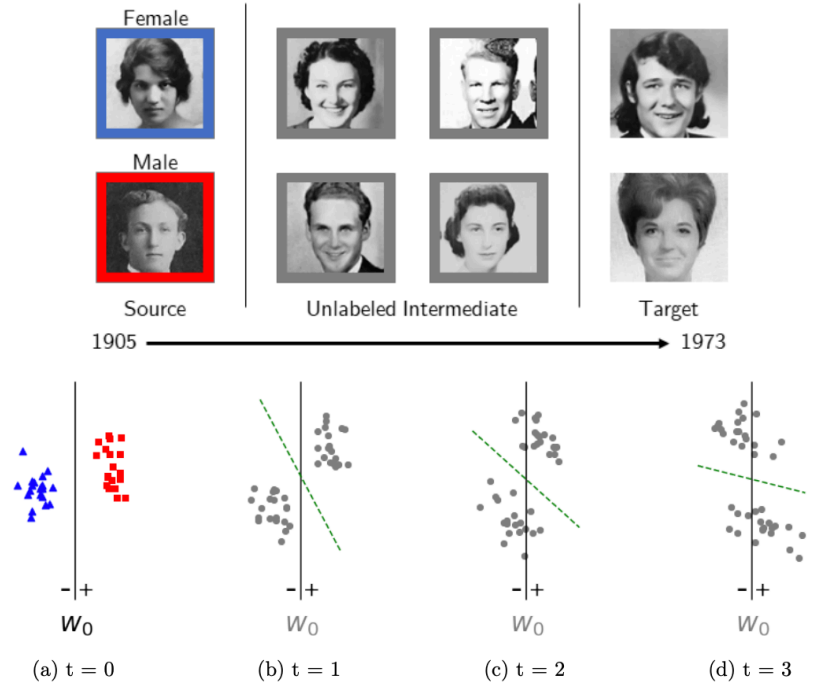
- In the deep learning era:
  - Essentially, to design an inductive bias that learns an invariance representation between source and target distributions
  - How to force the representations to be invariant?
    - Through a distance metric
      - e.g., Maximum Mean Discrepancy (MMD) (Rozantsev et al. 2016)
    - Through an adversarial classifier
      - e.g., Domain Adversarial Neural Network (DANN) (Ganin et al. 2016)
    - Through reconstruction
      - e.g., Deep Reconstruction Classification Network (DRCN) (Ghifary et al. 2016)



# Domain Adaptation Method Highlights

- With the power of samples
  - Self-Training for Gradual Domain Adaptation
  - (Kumar et al. 2020)




  - Practical insights
    - There must be an incentive for the model to update itself
      - Regularization is necessary
      - Label-sharpening  (instead of probabilistic labels) is necessary

Image credit: https://sunprinces.github.io/ML-Assignment3/bonus.html

# Domain Generalization

- Are models trained for domain adaptation good enough for a general real-world deployment?
  - Probably not. --- there might be other distributions that are not considered as target distributions during training.
- Domain Generalization
  - Train with samples from different domains, and test it with a new distribution

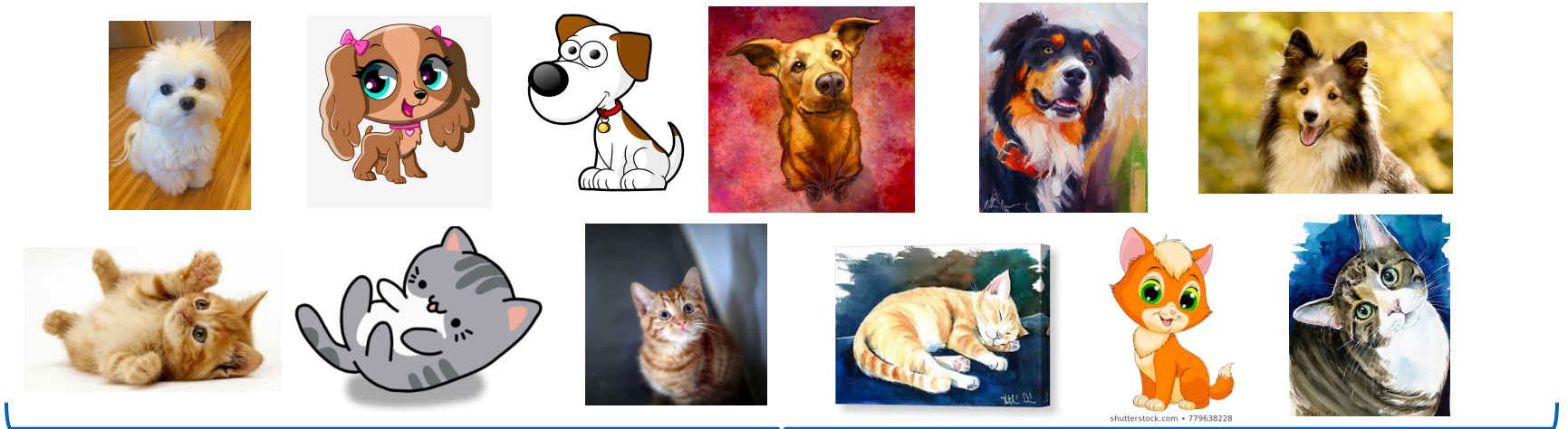
# Domain Generalization Methods Highlights

- Taking advantage of the domain identifications
  - Central assumption:
    - The model that can generalize similarly and well across different training distributions will be able to generalize well to an arbitrary target distribution
  - Method Development
    - Forcing invariance across different distributions
      - e.g., select-additive learning (Wang *et al.*, 2017)
    - Ensemble of classifiers (one for each domain)
      - e.g., best-source forward (Mancini *et al.*, 2018)
    - Meta-learning
      - e.g., iteratively split training domains to virtual training and testing domains (Li et al., 2018)

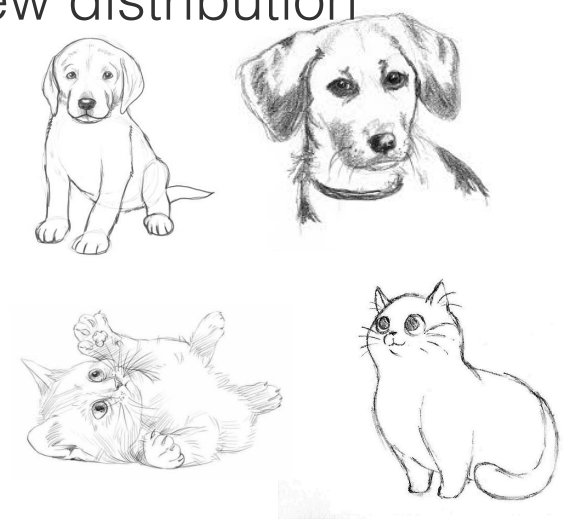


# Beyond Domain Generalization

- Are models trained for domain generalization good enough for a general real-world deployment?
  - Probably not. – the partitions of domains may not always be available in the real world.
- Beyond Domain Generalization
  - Train with samples from some domains, and test it with a new distribution



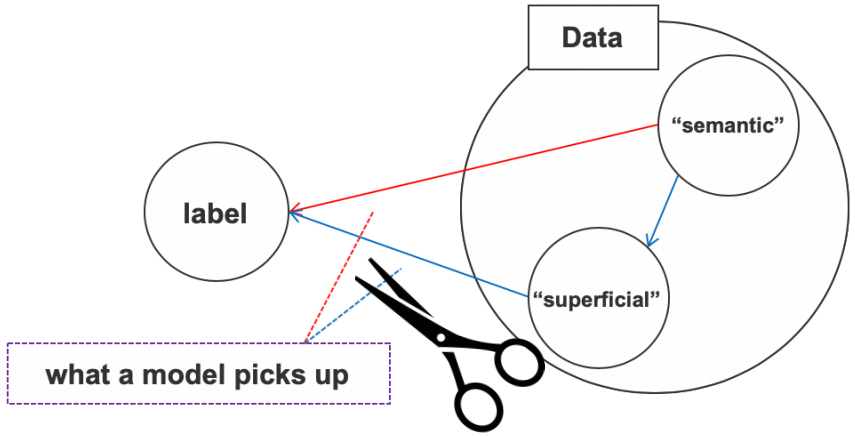
During Training



During Testing

# Beyond Domain Generalization

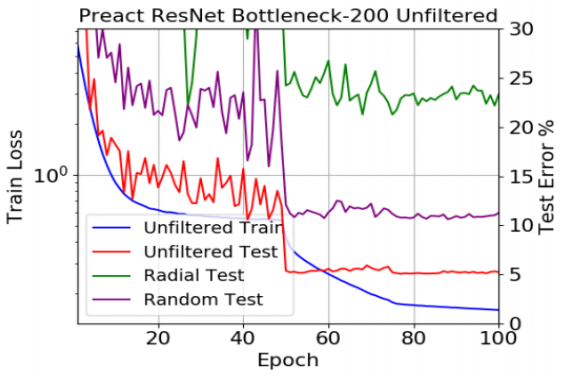
- Now, we probably have a problem that is real enough
  - But how to solve it then?

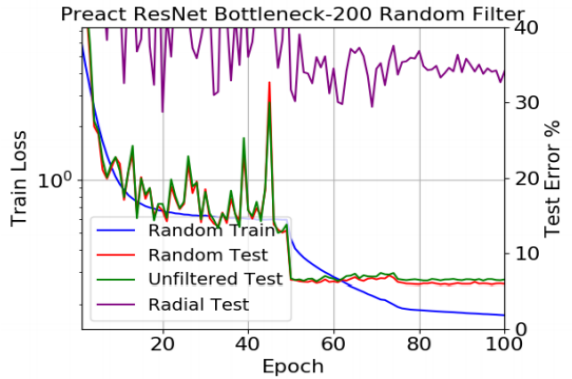


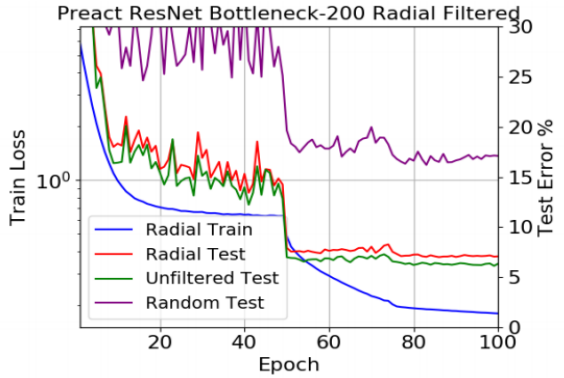

  - Find a hypothesis space that pick less superficial signals
    - Case Studies:
      - Neural GLCM (Wang et al. 2019a)
      - Patch-wise Adversarial Regularization (Wang et al. 2019b)

# Learning Robust Representations by Projecting Superficial Statistics

- The challenge
  - Neural networks learns textural signals
    - (Jo and Bengio 2017)
    - And many examples we have seen



# A Superficial Statistics Learner

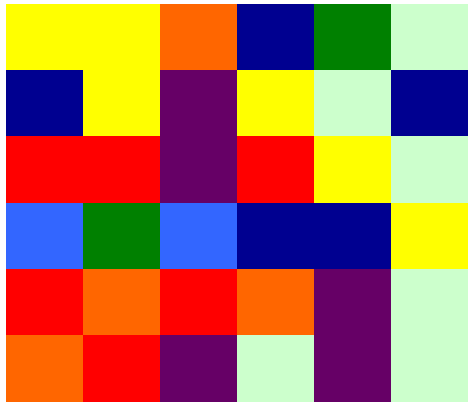
- Existing computer vision techniques
  - SURF (Bay et al., 2006)
  - LBP (He & Wang, 1990)
  - GLCM (Haralick et al., 1973)
- Semantic vs. Textural Experiments
  - Digits (MNIST, SVHN, MNIST-M, USPS)
    - Digit classification vs. Domain classification
  - Rotated MNIST
    - Digit classification vs. Rotation classification
  - FFT Kernelled MNIST
    - Digit classification vs. Kernel classification

| | | LBP | SURF | GLCM |
|---|---|---|---|---|
| Digit | Semantic | 0.179 | 0.563 | 0.164 |
| | Textural | 0.527 | 0.809 | 0.952 |
| Rotated | Semantic | 0.155 | 0.707 | 0.214 |
| | Textural | 0.121 | 0.231 | 0.267 |
| FFT | Semantic | 0.710 | 0.620 | 0.220 |
| | Textural | 0.550 | 0.200 | 0.490 |




# Gray-Level Co-Occurrence Matrix

- Count the number of pixel pairs under a certain direction



Image

| | | | | | |
|---|---|---|---|---|---|
| 1 | 1 | 7 | 5 | 3 | 2 |
| 5 | 1 | 6 | 1 | 2 | 5 |
| 8 | 8 | 6 | 8 | 1 | 2 |
| 4 | 3 | 4 | 5 | 5 | 1 |
| 8 | 7 | 8 | 7 | 6 | 2 |
| 7 | 8 | 6 | 2 | 6 | 2 |

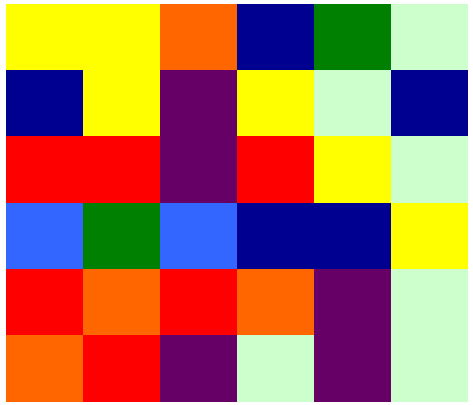
Pixel

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |

GLCM

# Gray-Level Co-Occurrence Matrix

- Count the number of pixel pairs under a certain direction



Image

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 7 | 5 | 3 | 2 | |
| 5 | 1 | 6 | 1 | 2 | 5 | |
| 8 | 8 | 6 | 8 | 1 | 2 | |
| 4 | 3 | 4 | 5 | 5 | 1 | |
| 8 | 7 | 8 | 7 | 6 | 2 | |
| 7 | 8 | 6 | 2 | 6 | 2 | |

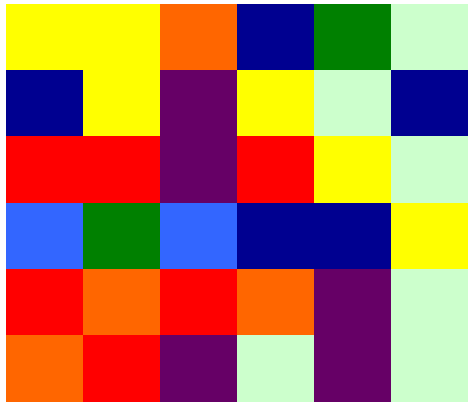
Pixel

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | 3 | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | | |

GLCM

# Gray-Level Co-Occurrence Matrix

- Count the number of pixel pairs under a certain direction



**Image**

| | | | | | | |
|---|---|---|---|---|---|---|
| 1 | 1 | 7 | 5 | 3 | 2 | |
| 5 | 1 | 6 | 1 | 2 | 5 | |
| 8 | 8 | 6 | 8 | 1 | 2 | |
| 4 | 3 | 4 | 5 | 5 | 1 | |
| 8 | 7 | 8 | 7 | 6 | 2 | |
| 7 | 8 | 6 | 2 | 6 | 2 | |

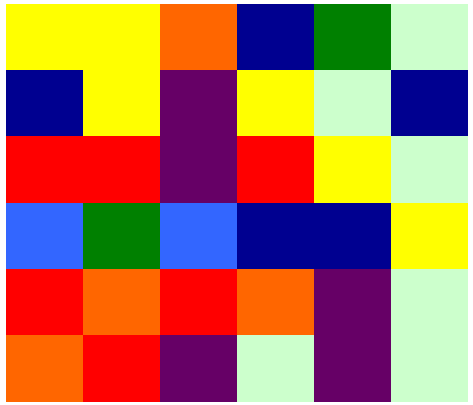
**Pixel**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | |
| 2 | | | | | | | | |
| 3 | | | | | | | | |
| 4 | | | | | | | | |
| 5 | | | | | | | | |
| 6 | | 3 | | | | | | |
| 7 | | | | | | | | |
| 8 | | | | | | | 2 | |

**GLCM**

# Gray-Level Co-Occurrence Matrix

- Count the number of pixel pairs under a certain direction



**Image**

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 7 | 5 | 3 | 2 |
| 5 | 1 | 6 | 1 | 2 | 5 |
| 8 | 8 | 6 | 8 | 1 | 2 |
| 4 | 3 | 4 | 5 | 5 | 1 |
| 8 | 7 | 8 | 7 | 6 | 2 |
| 7 | 8 | 6 | 2 | 6 | 2 |

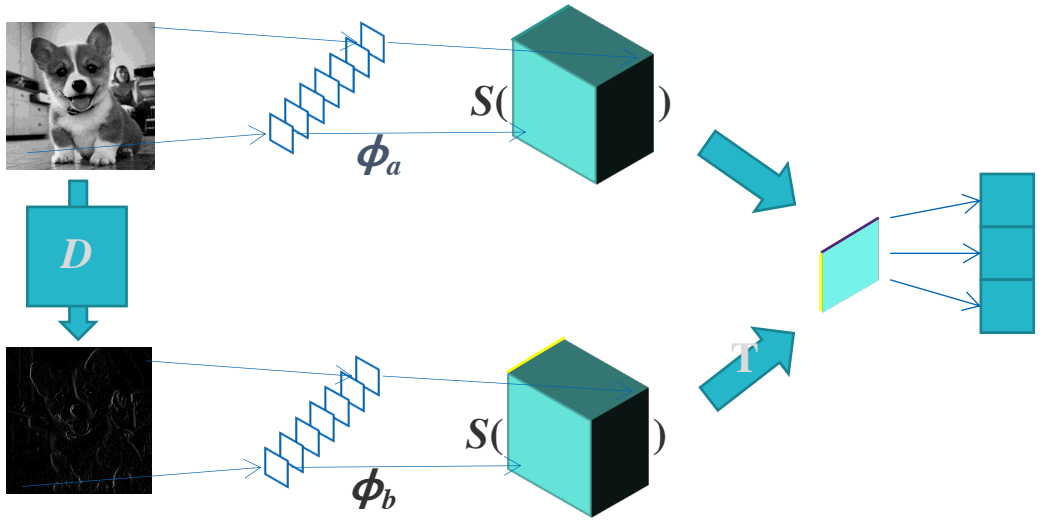
**Pixel**

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 0 | 0 | 0 | 1 | 1 | 0 |
| 2 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 5 | 2 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| 6 | 1 | 3 | 0 | 0 | 0 | 0 | 0 | 1 |
| 7 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 2 |
| 8 | 1 | 0 | 0 | 0 | 0 | 2 | 2 | 1 |

**GLCM**

# Neural Gray-Level Co-Occurrence Matrix

- Built into a neural network
  - Enable end-to-end learning
  - Differentiate weights



# Synthetic Experiments: TESTING HEX Method

- Office data set
  - Webcam (W)
  - Amazon (A)
  - DSLR (D)
- Testing with the Beyond Domain Generalization setting

| Train | Test | Baseline | HEX |
|---|---|---|---|
| A+W | D | 0.405±0.016 | 0.343±0.030 |
| D+W | A | 0.112±0.008 | 0.147±0.004 |
| A+D | W | 0.400±0.016 | 0.378±0.034 |



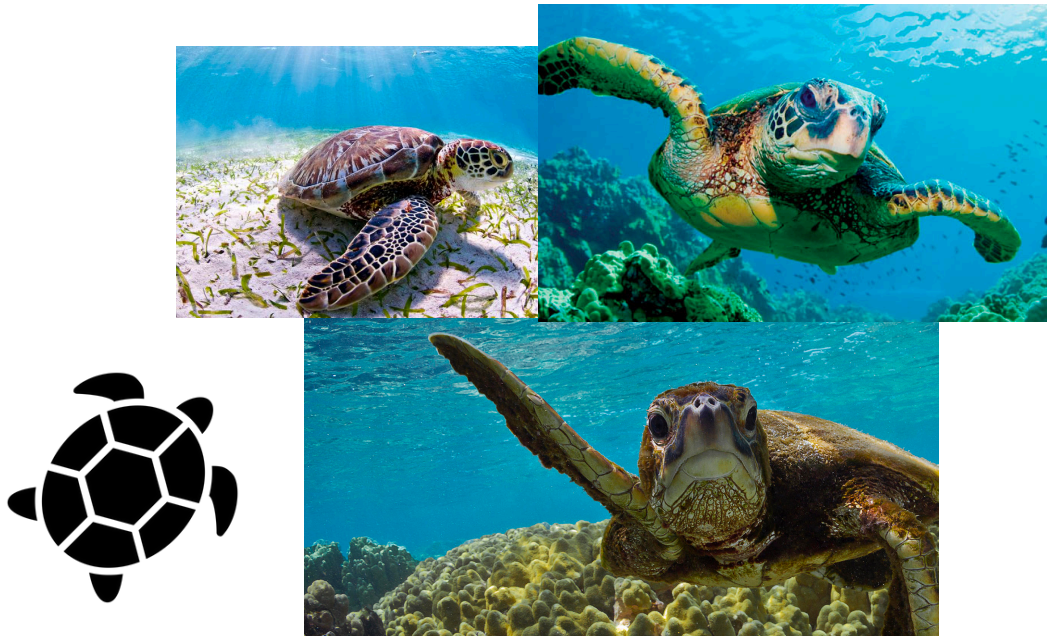

# OFFICE DATA SET: A CLOSER LOOK

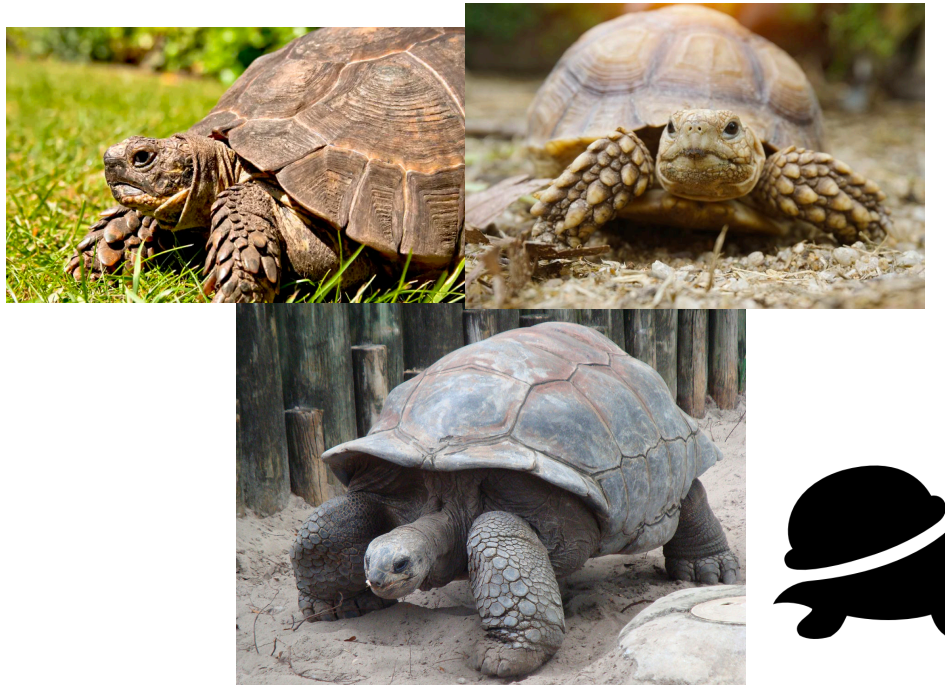
- Distribution similarities between D and W
- Two classifier:
  - C1: classifying object based on object and background
  - C2: classifying object based on object, ignoring background


| Train | Test | Baseline (C1) | HEX (C2) |
|---|---|---|---|
| A+W | D | 0.405±0.016 | 0.343±0.030 |
| D+W | A | 0.112±0.008 | 0.147±0.004 |
| A+D | W | 0.400±0.016 | 0.378±0.034 |

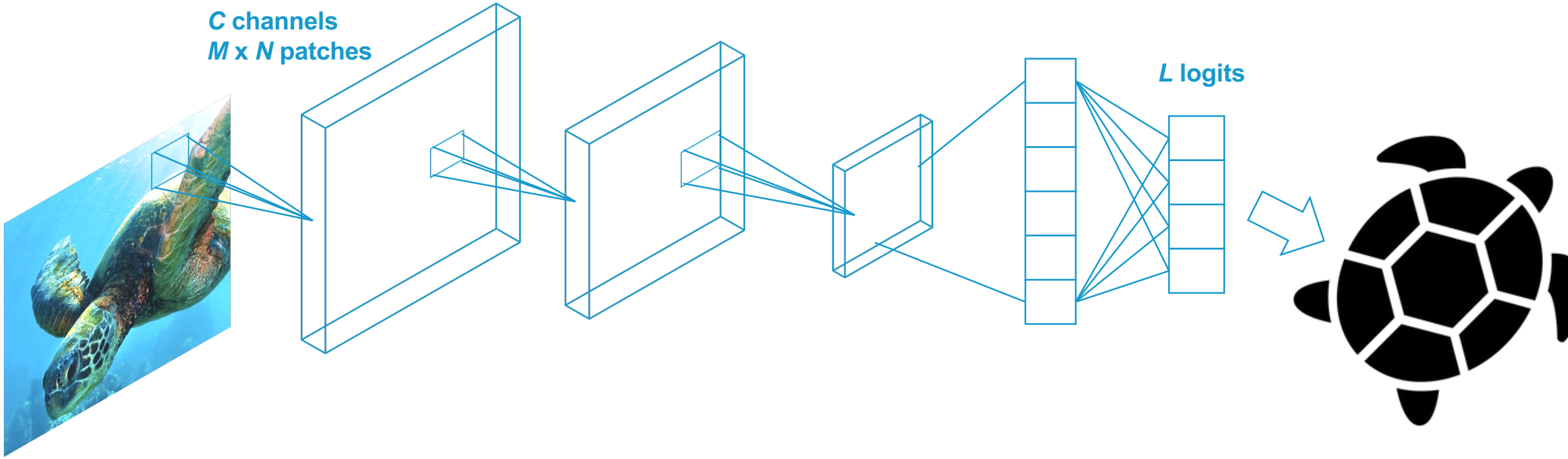
# Learning Robust Presentations by Penalizing the Local prediction power

- The challenge
  - Neural networks can predict through local signals, that do not align well with the annotation of the dataset

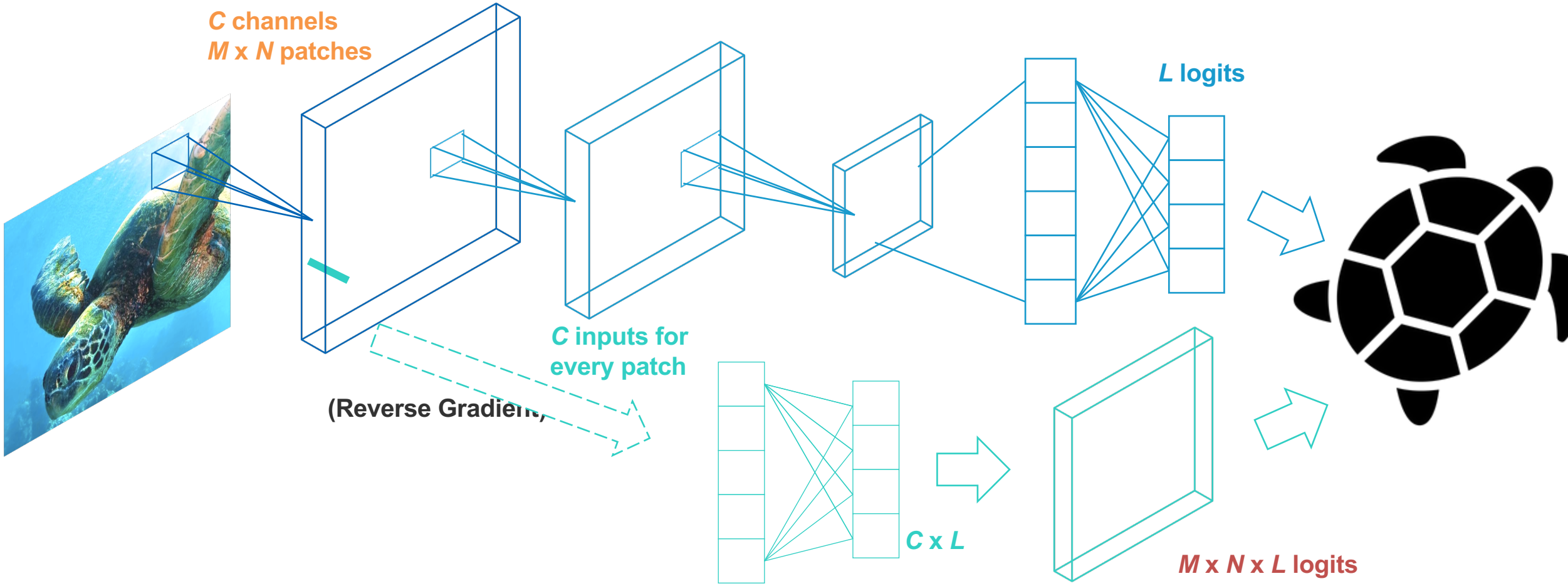


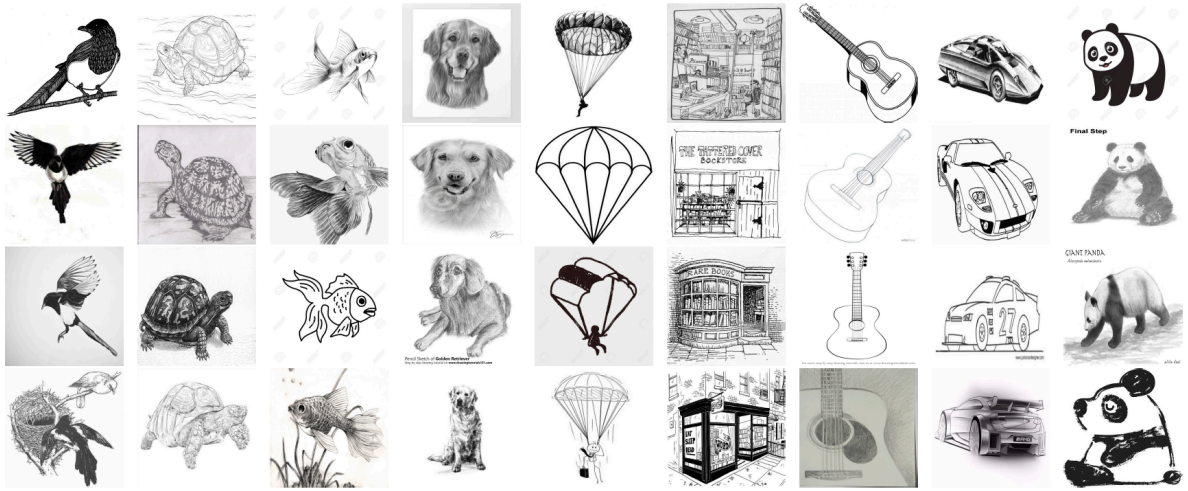
# Patch-wise Adversarial Regularization



# Patch-wise Adversarial Regularization



# Results – ImageNet Sketch

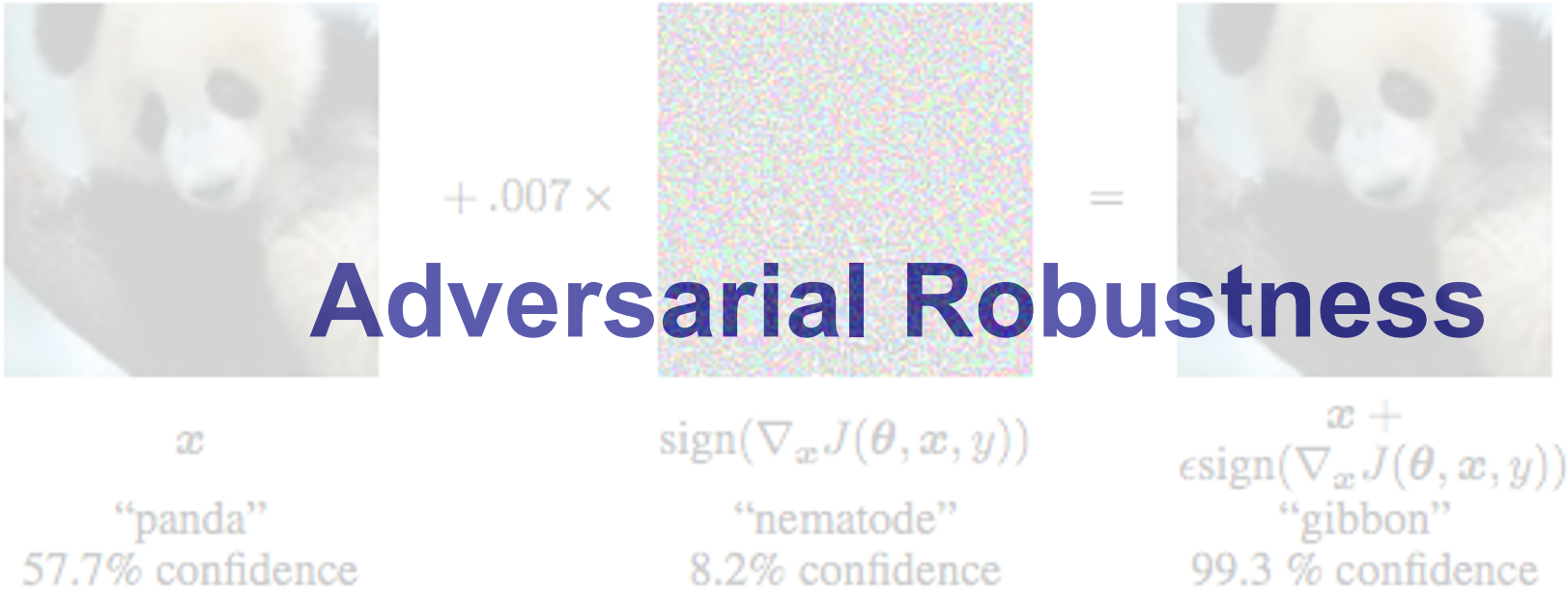


**ImageNet-Sketch Data set**

- ImageNet-Sketch data set consists of 50000 images, 50 images for each of the 1000 ImageNet classes.

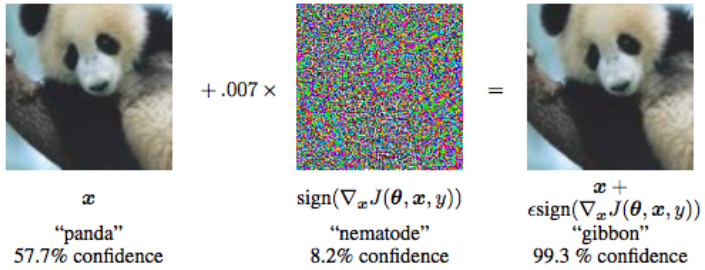
**Results**

| | AlexNet-PAR | | AlexNet | |
|---|---|---|---|---|
| | prediction | confidence | prediction | confidence |
|  | stethoscope | 0.6608 | hook | 0.3903 |
|  | tricycle | 0.9260 | safety pin | 0.5143 |
|  | Afghan hound | 0.8945 | swab (mop) | 0.7379 |
|  | red wine | 0.5999 | goblet | 0.7427 |




# Adversarial Robustness

# What are adversarial examples

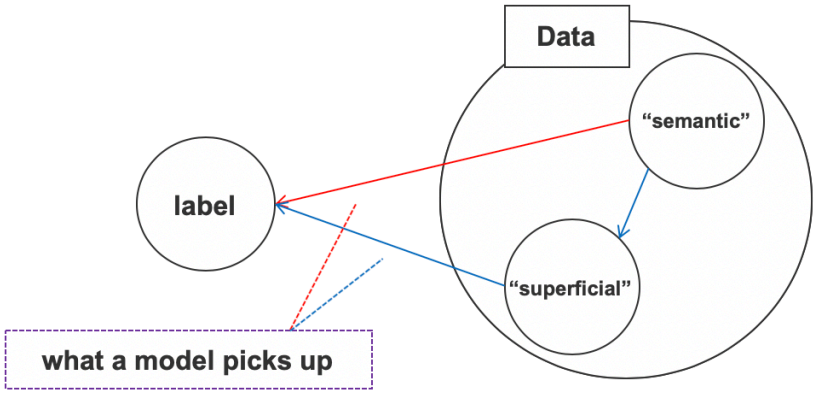



- Recap, with the help of the most popular example
- What are adversarial examples?
  - Samples with carefully crafted patterns added to an original sample, and
    - look identical to the original image to human,
    - but predicted by a model significantly different from the original sample.
    - This sounds quite magic, but not rigorous enough…
  - With sample $x$, model $f(;\boldsymbol{\theta})$, the adversarial example $x'$ is

$$x' = \arg\min_{x': f(x';\theta) \neq f(x;\theta)} d(x, x')$$

    - A result of an optimization, $d(,)$ is a distance
    - Hopefully, this does not sound too magic anymore.
      - How magic it looks to human depends on the choice of $d(,)$

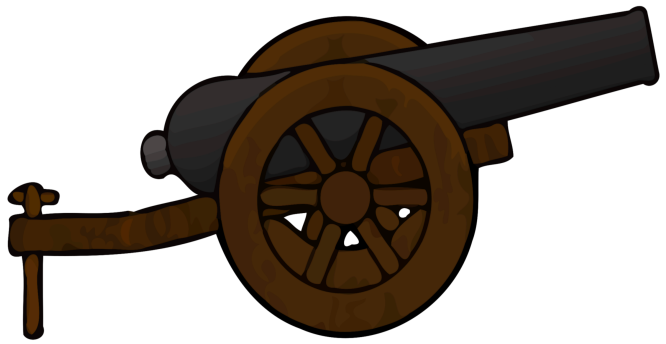
# What are adversarial examples




- Why are there adversarial examples?
  - Adversarial examples are direct outcomes of the fact
    - models are capturing superficial signals
      - (Ilyas *et al*. 2019)
      - (Wang *et al*. 2020)
- Why is adversarial robustness a popular topic
  - Hmm..., one reason might actually be it appears magic
  - The alignment between human and model may play a significant role in the real-world deployment of machine learning
- As a popular topic
  - Tons of efforts made to defend a model against adversarial examples
  - Unfortunately, even more efforts made to generate new adversarial examples…

# The Attack vs. Defense Arm Race Highlights

- Attack Methods
  - FGSM
    - (Goodfellow et al. 2014)
  - PGD
    - (Madry *et al*. 2017)

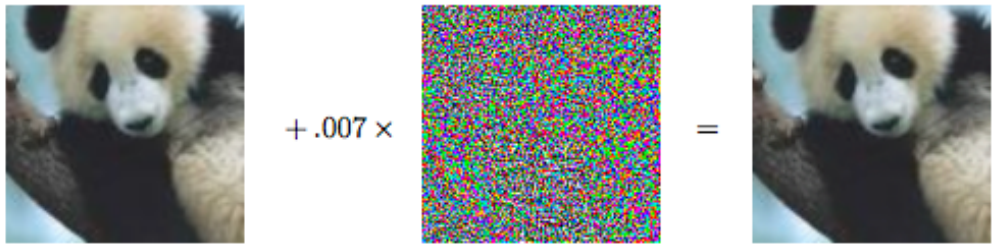


- Defense Methods
  - TRADES
    - (Zhang *et al*. 2019)
  - Adversarial Training
    - (Madry *et al*. 2017)

# The Attack vs. Defense Arm Race Highlights

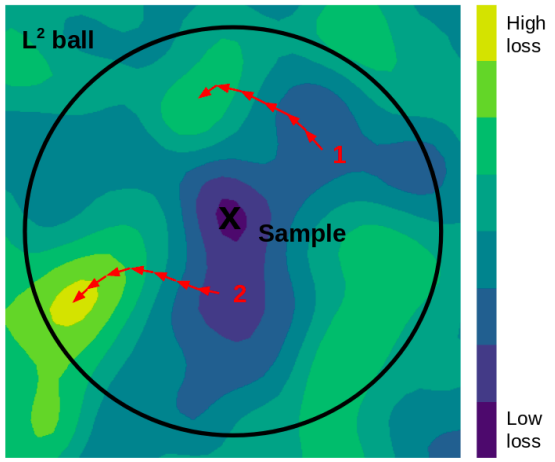
- Fast Gradient Sign Method (FGSM)
  - (Goodfellow et al. 2014)
  - Intuitively: a reverse way of training
    - Training (one epoch) is to update parameters to decrease loss, according to the gradient
    - FGSM is to update the data to increase loss, according to the sign of the gradient

$$x' = x + \epsilon \text{sign}(\nabla_x l(x, y; \theta))$$



# The Attack vs. Defense Arm Race Highlights

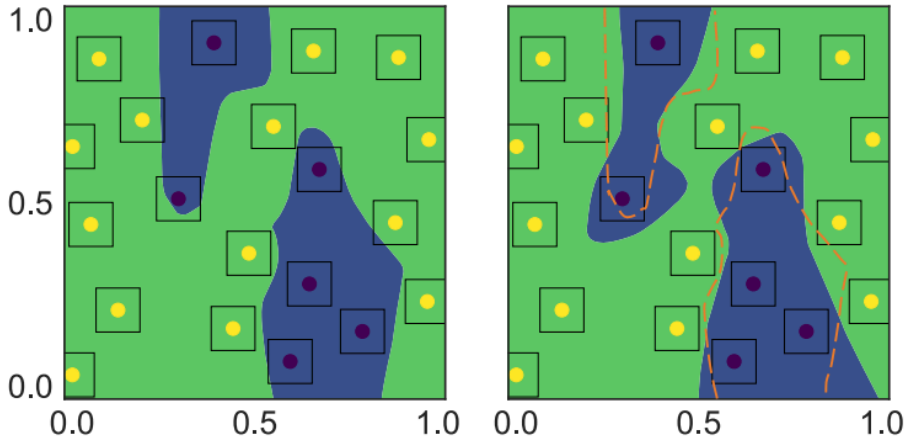
- PGD (Projected Gradient Descent) attack
  - (Madry *et al*. 2017)
  - Fun fact, it's previously invented under the name Basic Iterative Method (BIM)
  - Algorithm: Performing FGSM multiple times, with smaller step size.
    - Start with a random perturbation within the $l_p$ norm ball
    - Perform FGSM: $x' = x + \epsilon \text{sign}(\nabla_x l(x, y; \theta))$
    - Project back to the $l_p$ norm ball
    - Repeat multiple iterations




  - Probably the most powerful attack method nowadays ($l_p$ norm distance)

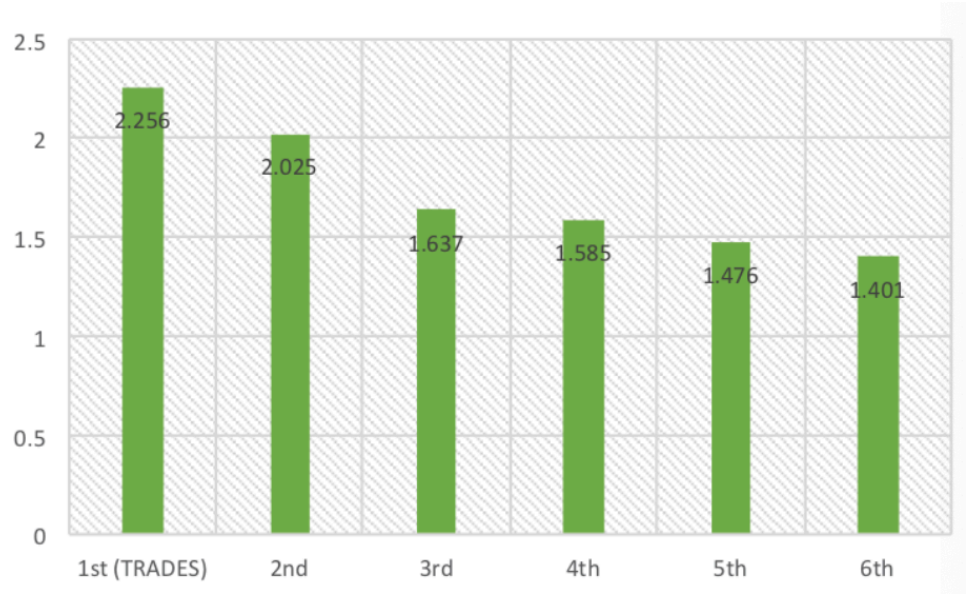
# The Attack vs. Defense Arm Race Highlights

- TRADES Defense
  - (Zhang *et al*. 2019)
  - Take advantage of the worst-case example
    - To identify the worst-case example
      - Apply attack during training to identify x'
      - Regularize the distance of embedding between x and x'
        - KL divergence regularization over softmax



# Competition: NeurIPS 2018 Adversarial Vision Challenge






**The methodology is the foundation of our entry to the NeurIPS 2018 Adversarial Vision Challenge in which we won the 1st place out of ~2,000 submissions, surpassing the runner-up approach by 11.41% in terms of mean $l_2$ perturbation distance**

# The Attack vs. Defense Arm Race Highlights

- Adversarial Training as a Defense
  - (Madry et al. 2017)
  - Summarize in one sentence:
    - augment the training data with generated adversarial examples during training
  - The most popular defense method
    - When integrated with PGD
      - Impressive empirical performance and simplicity

As we we see in next slide.

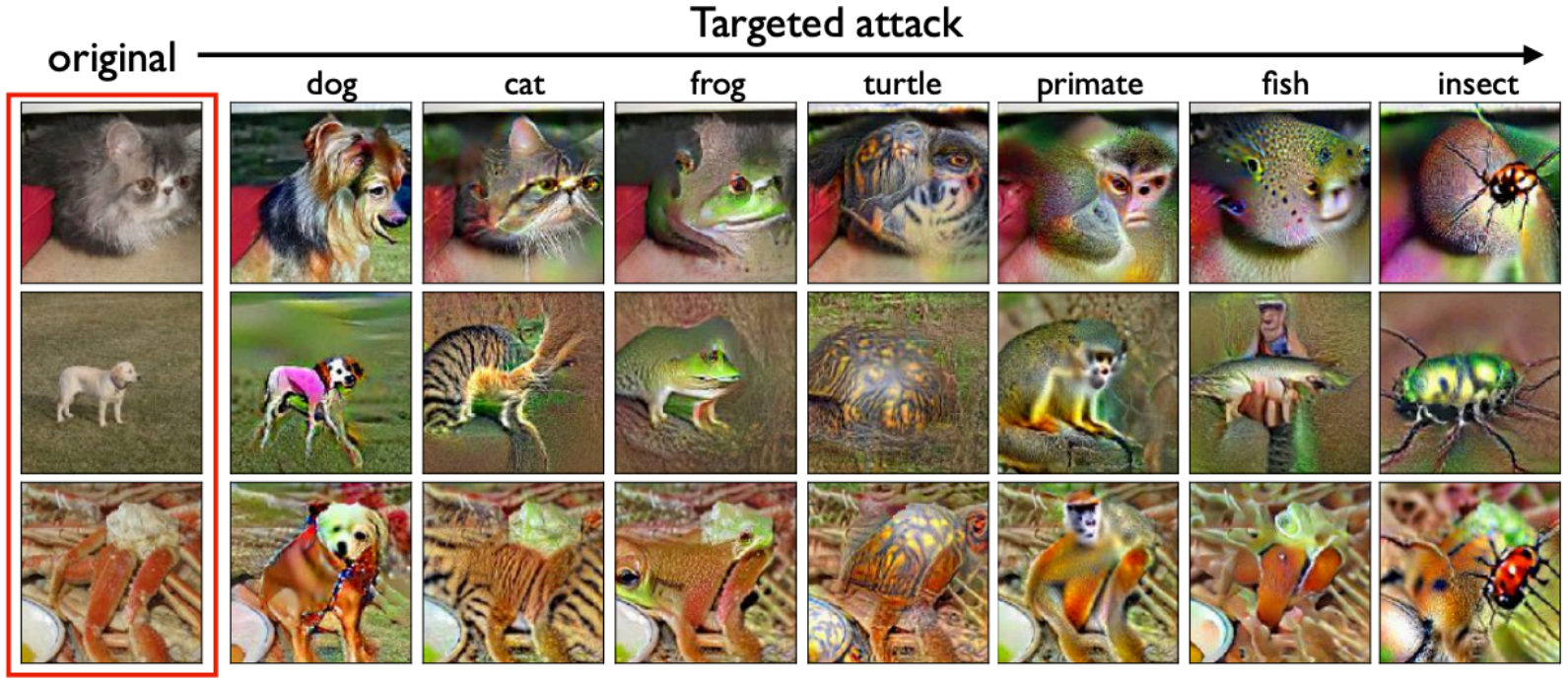
The community is making it easier:

- A faster (than PGD) way to generate the adversarial examples (that are equally powerful to the ones generated by PGD)
  - Adversarial training for free. Shafahi *et al*. 2019
  - Fast is better than free. Wong *et al*. 2020



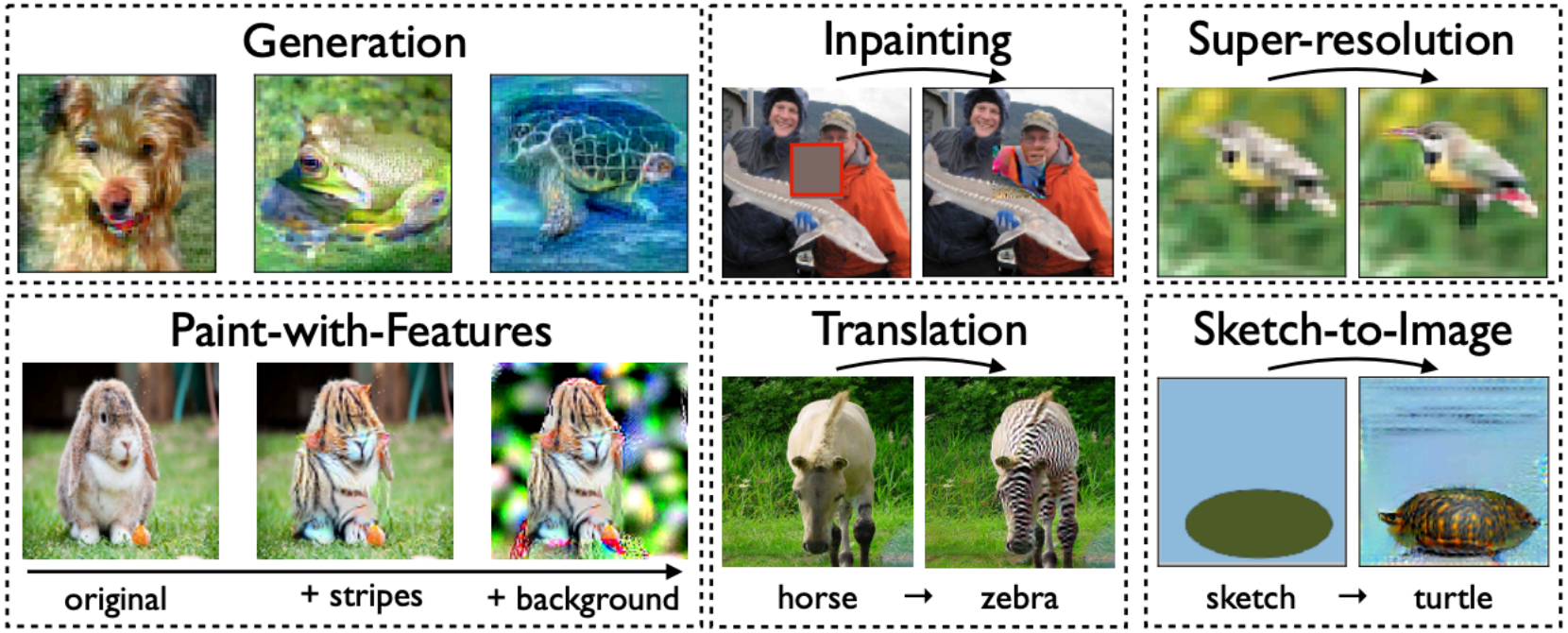
# Perceptually Aligned Representations from Adversarial Training w. PGD

- With PGD + adversarial training, adversarial examples are not that magical any more
  - (Santurkar et al. 2019)



# Perceptually Aligned Representations from Adversarial Training w. PGD

- With PGD + adversarial training, the same model may fulfill multiple tasks
  - (Santurkar et al. 2019)



# The Attack vs. Defense Arm Race

- With the impressive performance of PGD + adversarial training, can we conclude the development of defense methods?
  - Maybe not: It's simple, not not simple enough
  - Definitely not: there might be other attack methods can penetrate the defended model
    - So… is this race endless?
      - Fortunately, the community moves to provably robust models.



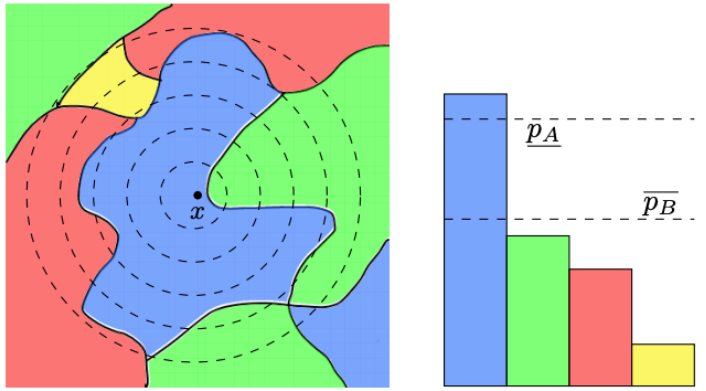
# Certified Robust Models

- Certified Robustness through Randomized Smoothing
  - (Cohen et al. 2019)
  - Randomized smoothing method

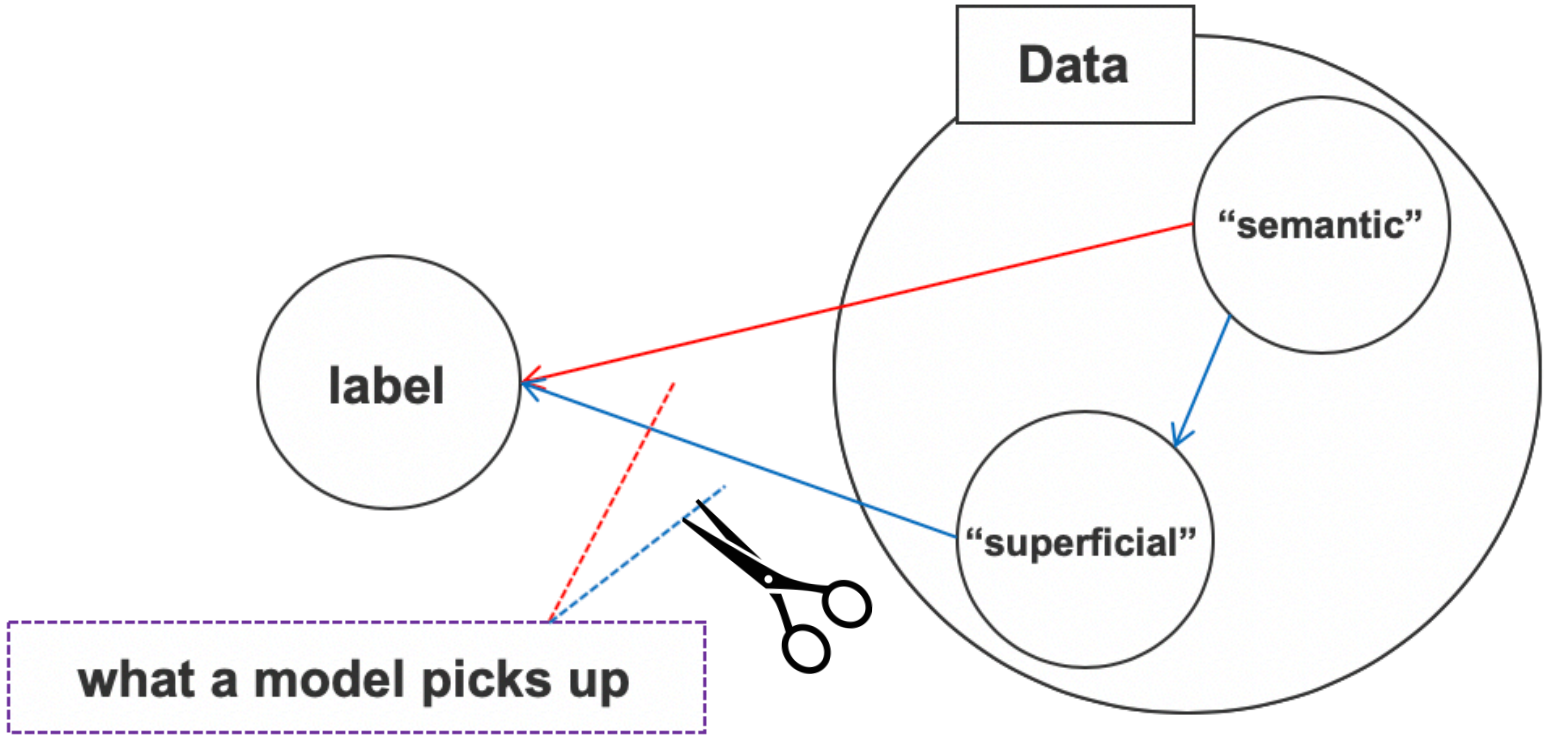
$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\varepsilon) = c)$$

$$\text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2 I)$$

  - Certified Robustness
    - Over $l_2$ norm distance



# Looking into the Future



# A Brief Overview of Trustworthy Machine Learning

- New Challenges of Modern Machine Learning
  - Empirical Observations
  - Trade-off between Accuracy and Robustness
- Cross-domain Robust Models
  - Generalization Analysis of Domain Adaptation
  - Method Overview – It's mostly about Invariance
  - Domain Generalization and Beyond
- Adversarial Robust Models
  - The Attack vs. Defense Arm Race Highlights
  - Adversarial Training and Its Recent Developments
  - Certified Robustness and Generalization Analysis



# References


- Ben-David, Shai, et al. "A theory of learning from different domains." Machine learning 79.1-2 (2010): 151-175.
- Cohen, Jeremy M., et al. "Certified adversarial robustness via randomized smoothing." arXiv preprint arXiv:1902.02918 (2019).
- Devlin, Bernie, and Kathryn Roeder. "Genomic control for association studies." Biometrics 55.4 (1999): 997-1004.
- Ganin, Yaroslav, et al. "Domain-adversarial training of neural networks." The Journal of Machine Learning Research 17.1 (2016): 2096-2030.
- Geirhos, Robert, et al. "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness." arXiv preprint arXiv:1811.12231 (2018).
- Ghifary, Muhammad, et al. "Deep reconstruction-classification networks for unsupervised domain adaptation." European Conference on Computer Vision. Springer, Cham, 2016.
- Goodfellow, Ian J., et al. "Explaining and harnessing adversarial examples." arXiv preprint arXiv:1412.6572 (2014).
- Ilyas, Andrew, et al. "Adversarial examples are not bugs, they are features." Advances in Neural Information Processing Systems. 2019.
- Jo, Jason, and Yoshua Bengio. "Measuring the tendency of CNNs to learn surface statistical regularities." arXiv preprint arXiv:1711.11561 (2017).
- Kumar, Ananya, et al. "Understanding Self-Training for Gradual Domain Adaptation." arXiv preprint arXiv:2002.11361 (2020).





# References


- Li, Da, et al. "Learning to generalize: Meta-learning for domain generalization." Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
- Madry, Aleksander, et al. "Towards deep learning models resistant to adversarial attacks." arXiv preprint arXiv:1706.06083 (2017).
- Mancini, Massimiliano, et al. "Best sources forward: domain generalization through source-specific nets." 2018 25th IEEE International Conference on Image Processing (ICIP). IEEE, 2018.
- Rozantsev, Artem, et al. "Beyond sharing weights for deep domain adaptation." IEEE transactions on pattern analysis and machine intelligence 41.4 (2018): 801-814.
- Santurkar, Shibani, et al. "Image synthesis with a single (robust) classifier." Advances in Neural Information Processing Systems. 2019.
- Shafahi, Ali, et al. "Adversarial training for free!." Advances in Neural Information Processing Systems. 2019.
- Szegedy, Christian, et al. "Intriguing properties of neural networks." arXiv preprint arXiv:1312.6199 (2013).
- Tsipras, Dimitris, et al. "Robustness may be at odds with accuracy." arXiv preprint arXiv:1805.12152 (2018).
- Vilhjálmsson, Bjarni J., and Magnus Nordborg. "The nature of confounding in genome-wide association studies." Nature Reviews Genetics 14.1 (2013): 1-2.
- Wang, Haohan, et al. "Select-additive learning: Improving generalization in multimodal sentiment analysis." 2017 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2017.





# References


- Wang, Haohan, et al. "Learning robust representations by projecting superficial statistics out." ICLR (2019a).
- Wang, Haohan, et al. "Learning robust global representations by penalizing local predictive power." Advances in Neural Information Processing Systems. 2019b.
- Wang, Haohan, et al. "High frequency component helps explain the generalization of convolutional neural networks." CVPR (2020).
- Weale, Michael E., et al. "Armenian Y chromosome haplotypes reveal strong regional structure within a single ethno-national group." Human genetics 109.6 (2001): 659-674.
- Wong, Eric, et al. "Fast is better than free: Revisiting adversarial training." arXiv preprint arXiv:2001.03994 (2020).
- Yang, Fanny, et al. "Invariance-inducing regularization using worst-case transformations suffices to boost accuracy and spatial robustness." Advances in Neural Information Processing Systems. 2019.
- Zhang, Hongyang, et al. "Theoretically principled trade-off between robustness and accuracy." arXiv preprint arXiv:1901.08573 (2019).
- Zhao, Han, et al. "On learning invariant representation for domain adaptation." arXiv preprint arXiv:1901.09453 (2019).