CMSC 491/691 Robust Machine Learning

Topic 3: OOD/Novelty Detection

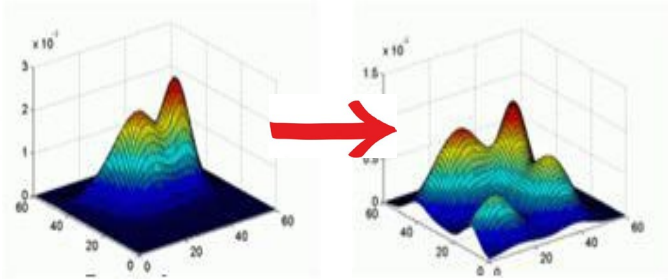
UMBC

Recall

"Adapting/Generalizing to Domain Shift"

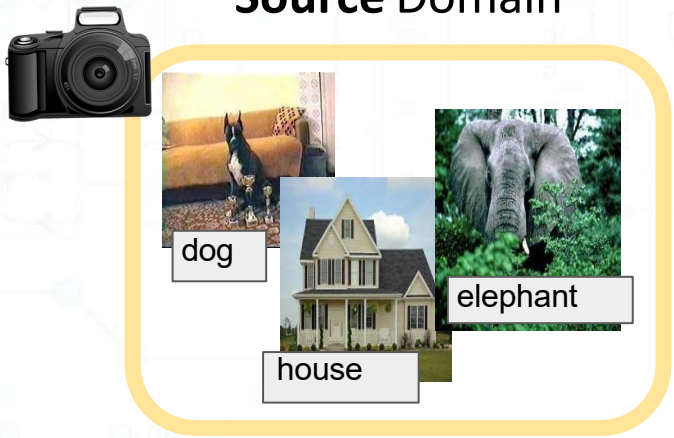
Classical Domain Adaptation

Source (Train)

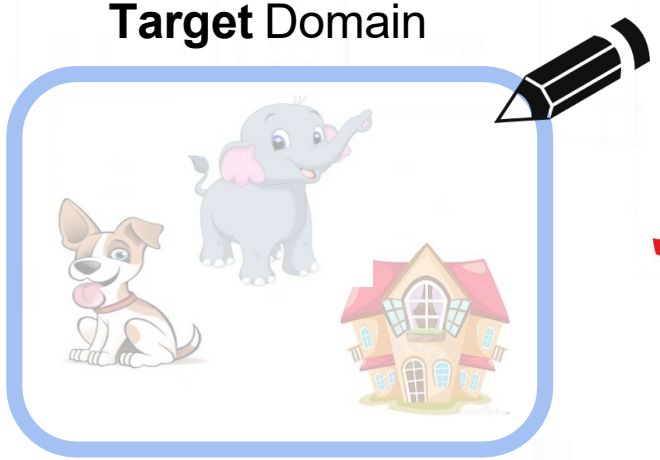


Target (Test)

Labelled **Source** Domain



Unlabelled **Target** Domain

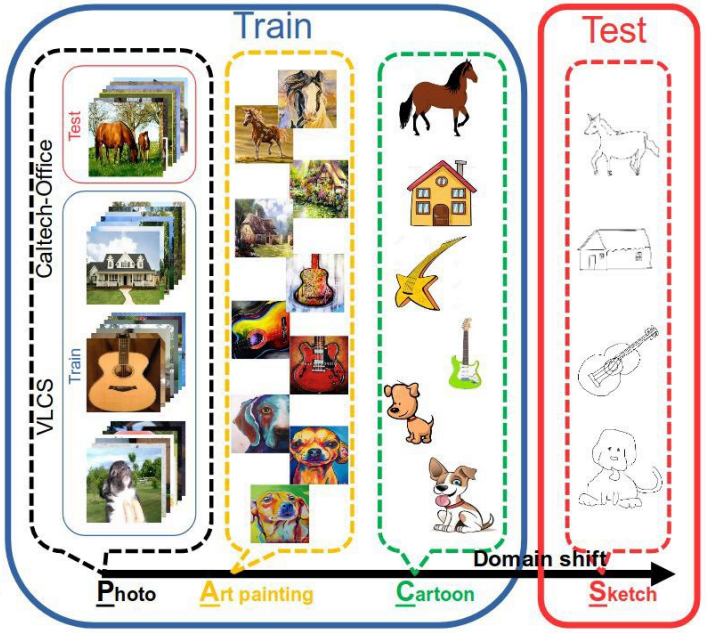


Unsupervised DA, transductive setting

Train

Test

multi-source
Domain
Generalization



[Deeper, Broader and Artier Domain Generalization, ICCV 2017]

Annotated
Source data

Transductive
Unsupervised
Domain
Adaptation

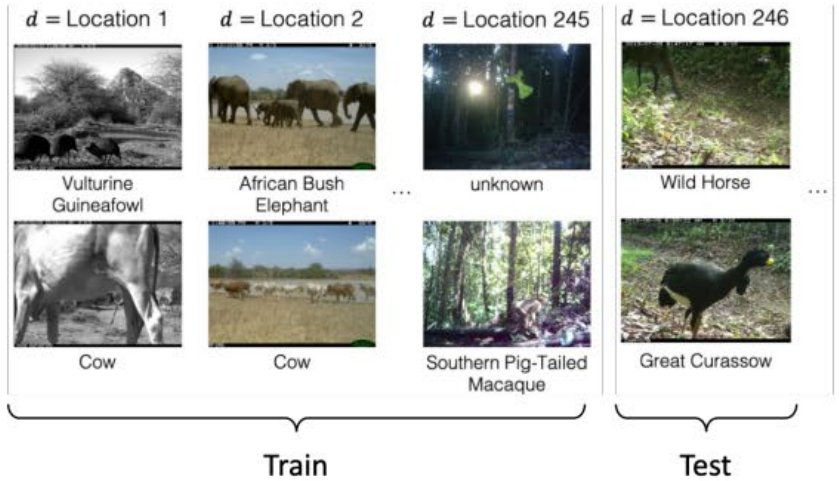
semi-
supervised
DA

Annotated
Target data

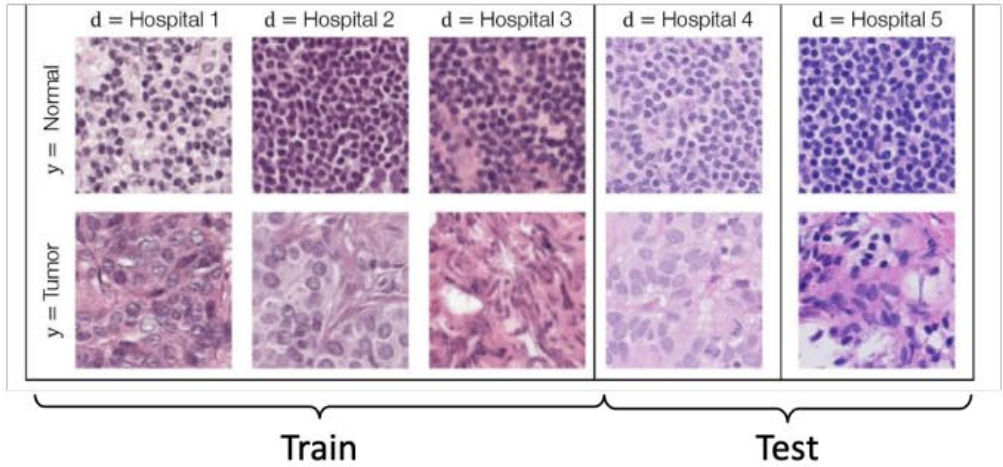
Transfer
Learning

Domain Generalization: Applications

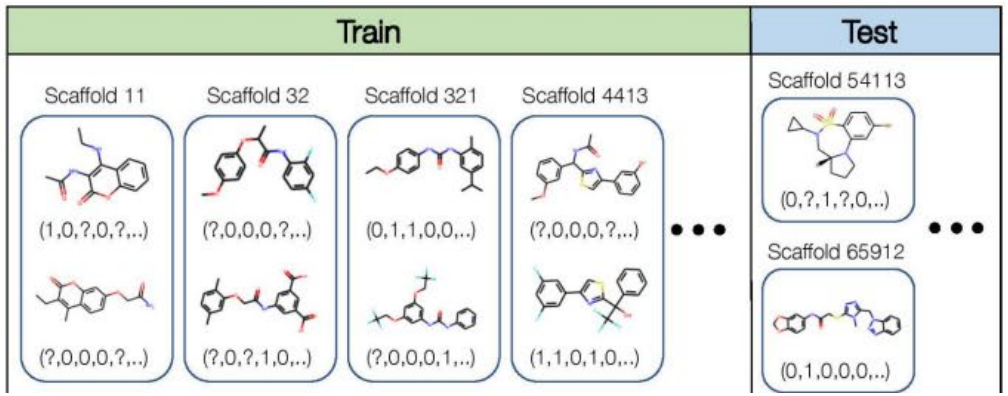
Wildlife recognition



Tissue classification



Molecule property prediction



Code completion

	Repository ID (d)	Source code context (x)	Next tokens (y)
Train	Repository 1	`... from easyrec.gateway import EasyRec <EOL> gateway = EasyRec('tenant','key') <EOL> item_type = gateway. ____`	`get_item_type`
		`... response = gateway.get_other_users() <EOL> get_params = HTTPretty. ____`	`last_request`
	Repository 2	`import numpy as np ... <EOL> if np.linalg.norm(target - prev_target) > far_threshold: <EOL> norm = np. ____`	`linalg`
		`... new_trans = np.zeros((n_beats + max_beats, n_beats) <EOL> new_trans[:n_beats,:n_beats] = np. ____`	`max`
	⋮		
Test	Repository 6,001	`... if e.errno == errno.ENOENT: <EOL> continue <EOL> p = subprocess.Popen () <EOL> stdout = p. ____`	`communicate`
		`... command = shlex.split(command) <EOL> command = map(str, command) <EOL> env = os. ____`	`environ`
	⋮		

Today's Topic: How can we detect OOD samples?

Motivation: When is it safe to use model predictions?

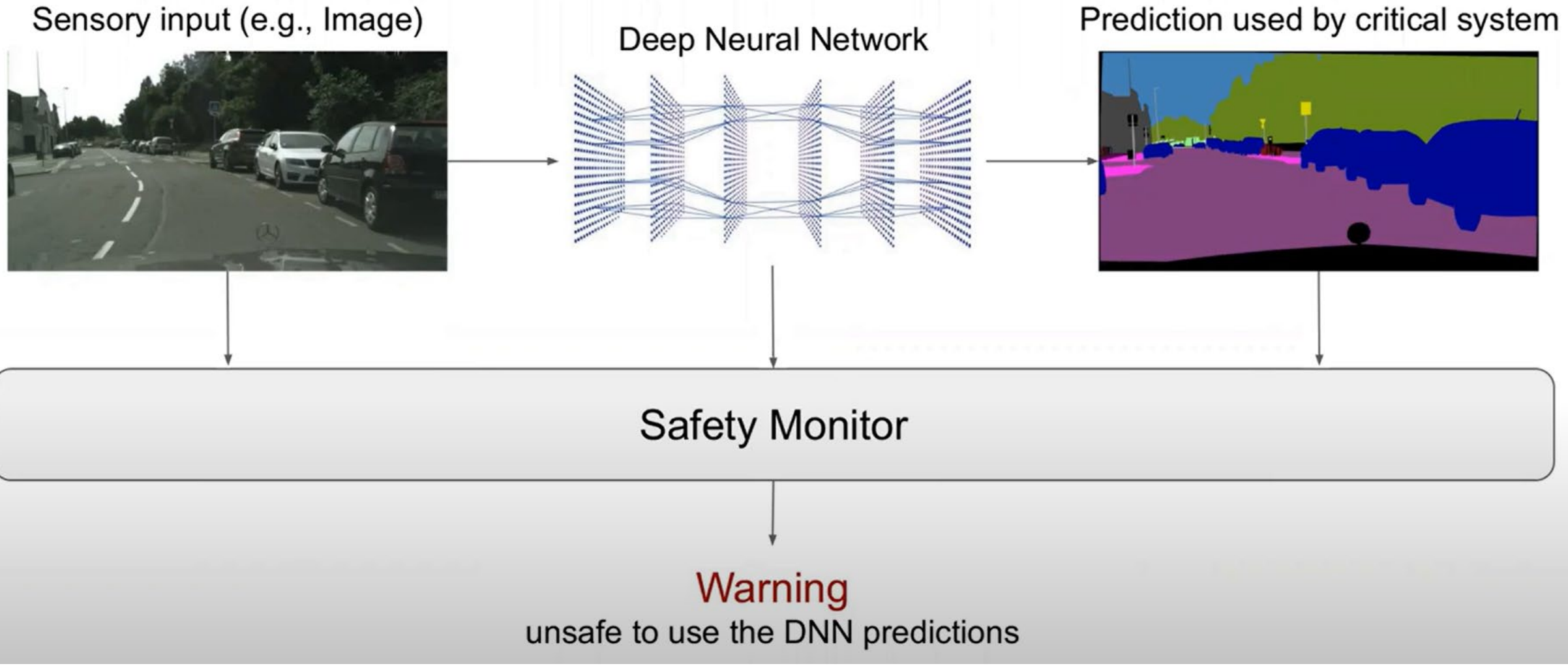


Figure: Guerin AAAI 2023

Ideal world

Independent and identically distributed (IID)

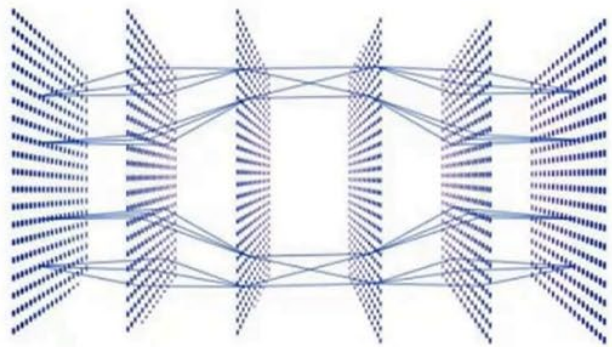
$$p_{TEST}(y,x) = p_{TRAIN}(y,x)$$

Labelled training dataset

Cat Dog Cat Dog

Training process

Neural network classifier



Neural networks are **good for inputs close to their training data** and bad for other inputs.

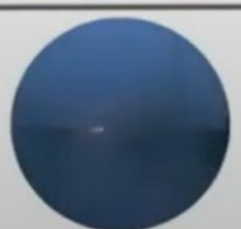
Ideal world

Independent and identically distributed (IID)

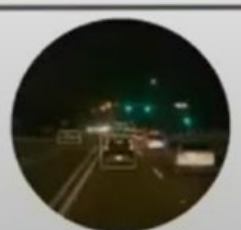
$$p_{TEST}(y,x) = p_{TRAIN}(y,x)$$

Examples of dataset shift:

- ***Covariate shift.*** Distribution of features p(x) changes and p(y|x) is fixed.



Weather



Night

Image credit: [Hendrycks & Dietterich, 2019](); Sun et al, [Waymo Open Dataset]()

Ideal world	Real world
Independent and identically distributed (IID)	Out-of-distribution(OOD)
$p_{TEST}(y,x) = p_{TRAIN}(y,x)$	$p_{TEST}(y,x) \neq p_{TRAIN}(y,x)$

Examples of dataset shift:

- ***Covariate shift.*** Distribution of features p(x) changes and p(y|x) is fixed.

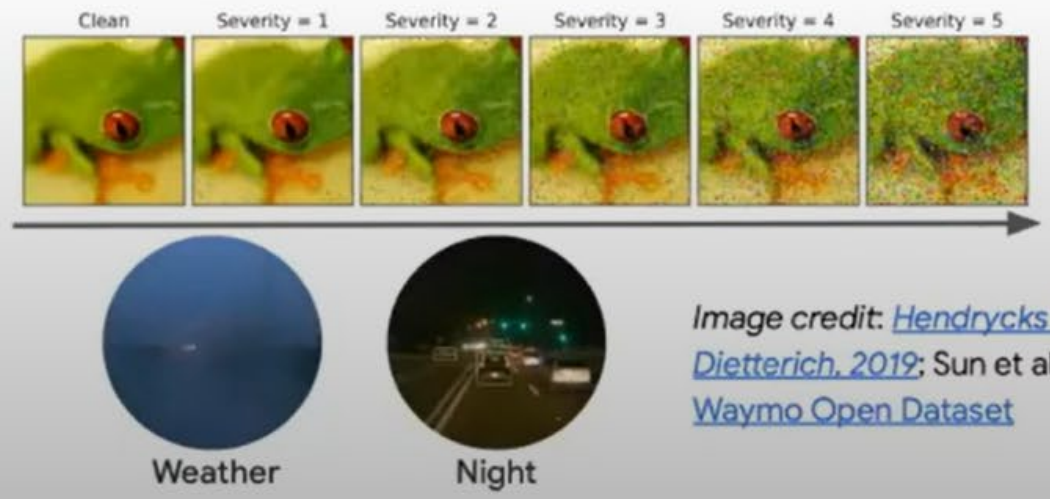


Image credit: Hendrycks & Dietterich, 2019; Sun et al, Waymo Open Dataset

- ***Out-of-distribution detection (open-set recognition/ anomaly detection).*** New classes may appear at test time.

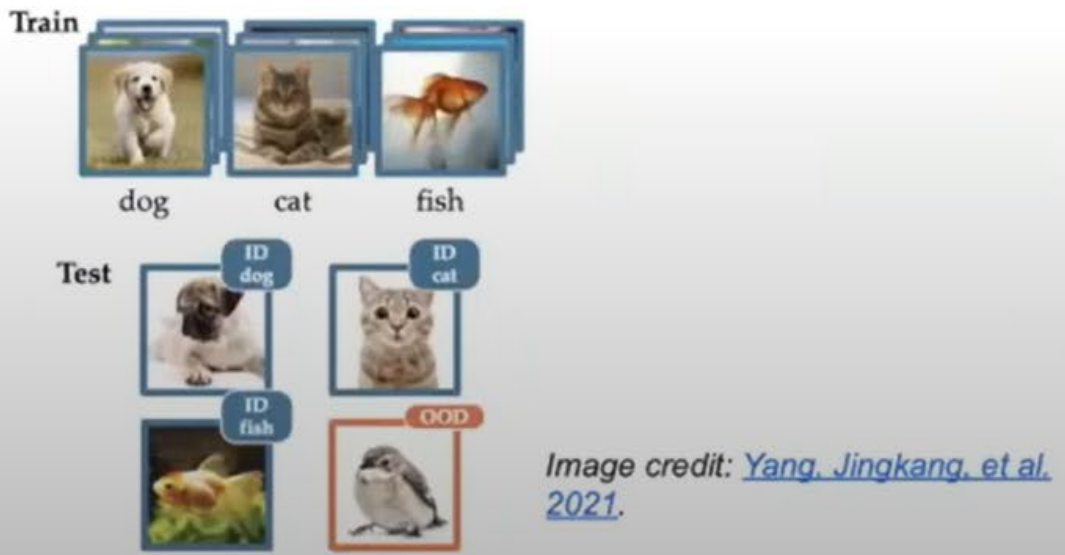


Image credit: Yang, Jingkang, et al. 2021.

Different definitions of Out-of-Distribution data

In-distribution data

Covariate shift

(Changes in image characteristics)

Schorn and Gauerhof (2020), Chen et al. (2020), Cai and Koutsoukos (2020),

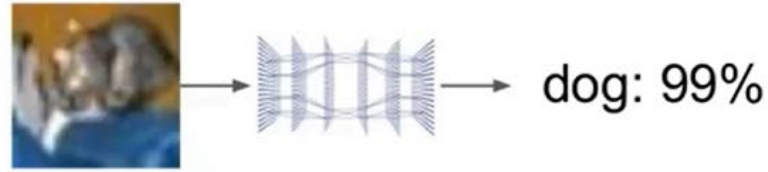
Novelty detection

(Changes in image content)

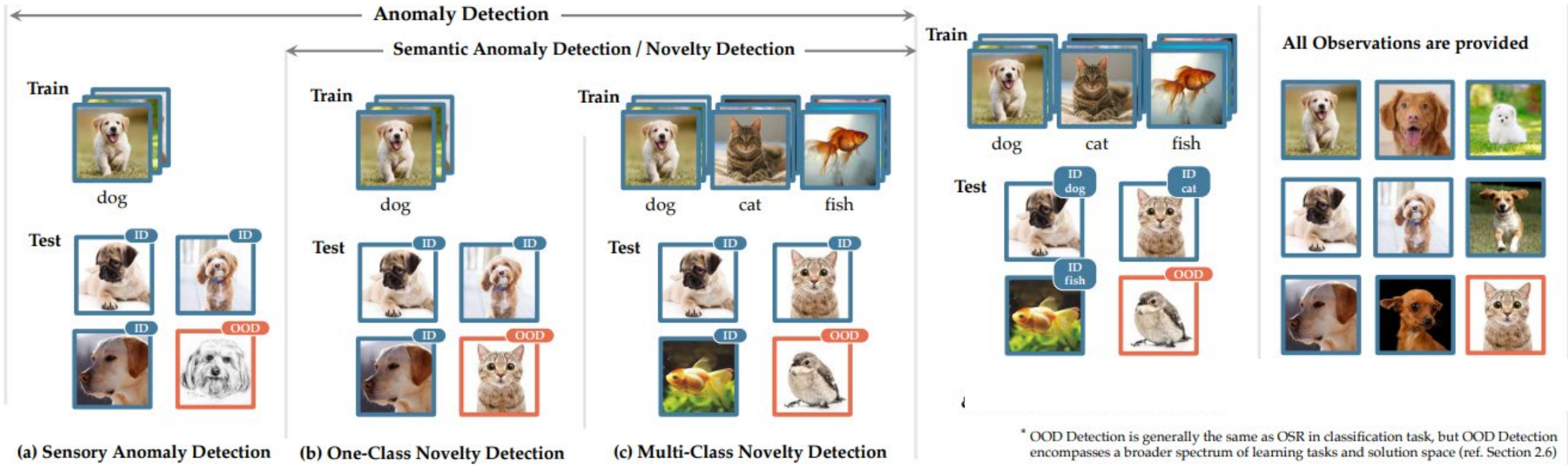
Lee et al. (2018), Liu et al. (2020), Henzinger et al. (2020)

Adversarial attacks

(Imperceptible malicious changes)



Lee et al. (2018), Wang et al. (2019), Kantaros et al. (2021)



* OOD Detection is generally the same as OSR in classification task, but OOD Detection encompasses a broader spectrum of learning tasks and solution space (ref. Section 2.6)

OoD

$(p_{\text{train}}(\mathbf{x}) \neq p_{\text{ood}}(\mathbf{x}))$

Distribution Shift

$(p_{\text{train}}(y|\mathbf{x}) = p_{\text{ood}}(y|\mathbf{x}))$

(perceptually different + semantically similar)

Unseen Categories

$(p_{\text{train}}(y|\mathbf{x}) \neq p_{\text{ood}}(y|\mathbf{x}))$

(semantically different)

Near OoD

(perceptually similar)

Far OoD

(perceptually different)

- *Far-OOD* (CIFAR vs SVHN)
  - Simple statistics can be effective
- *Near-OOD* (CIFAR-100 vs CIFAR-10)
  - Detecting subtle semantic differences
    - pickup truck vs truck
  - Distinguishing semantics vs background
    - fish vs airplane similar background

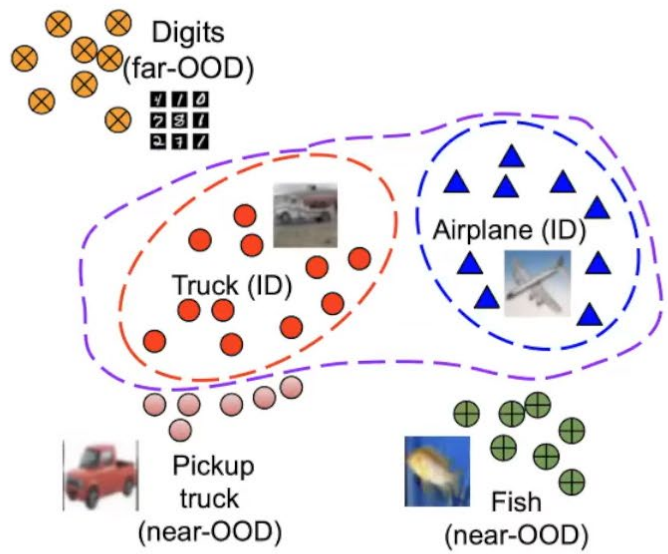
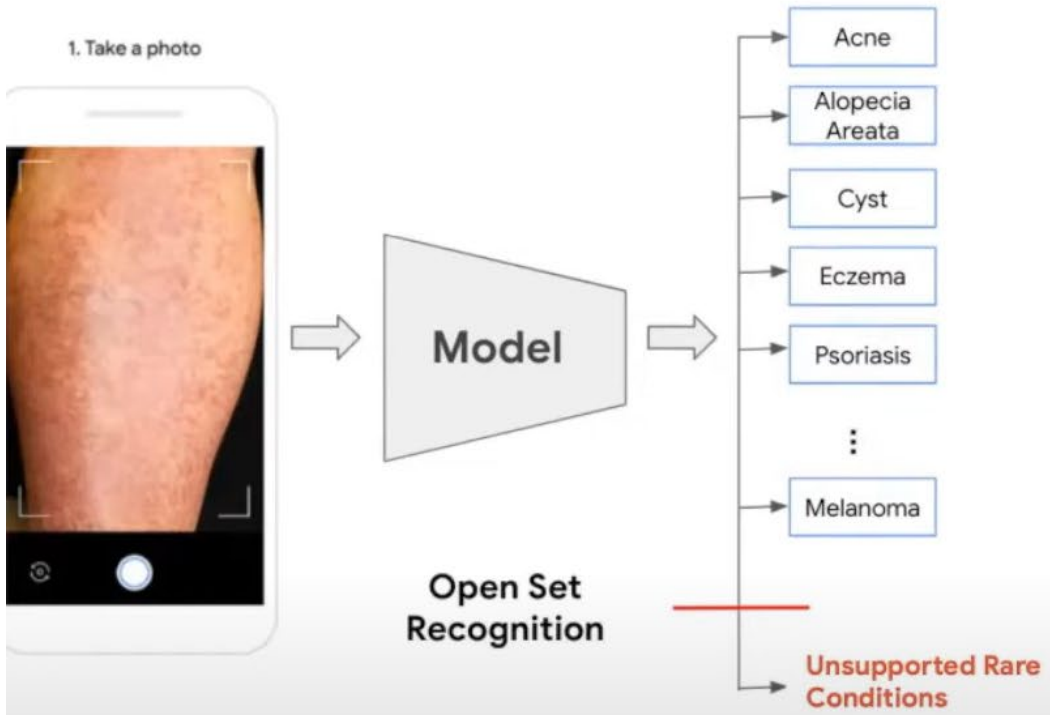


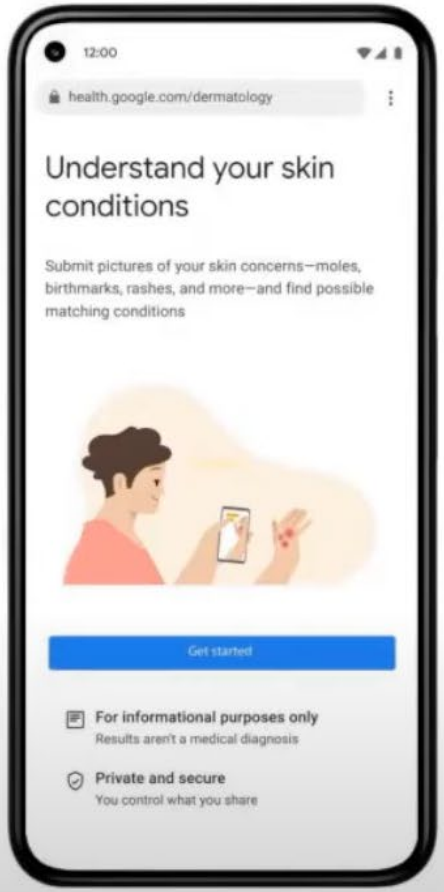
Image source: CIFAR and MNIST datasets

Closed World vs Open World

- **Closed World**
  - Training Data: X_{train} with class labels $Y_{train} = \{y_1 \dots y_n\}$
  - Test Data Data: X_{test} with class labels $Y_{test} \subset Y_{train}$
  - Classes at test-time must be seen during training (no "new" classes)
- **Open World**
  - Unseen Classes in test data
  - **OOD Detection: Detect and flag unseen test classes**

High-Stakes Application: Medical Diagnosis





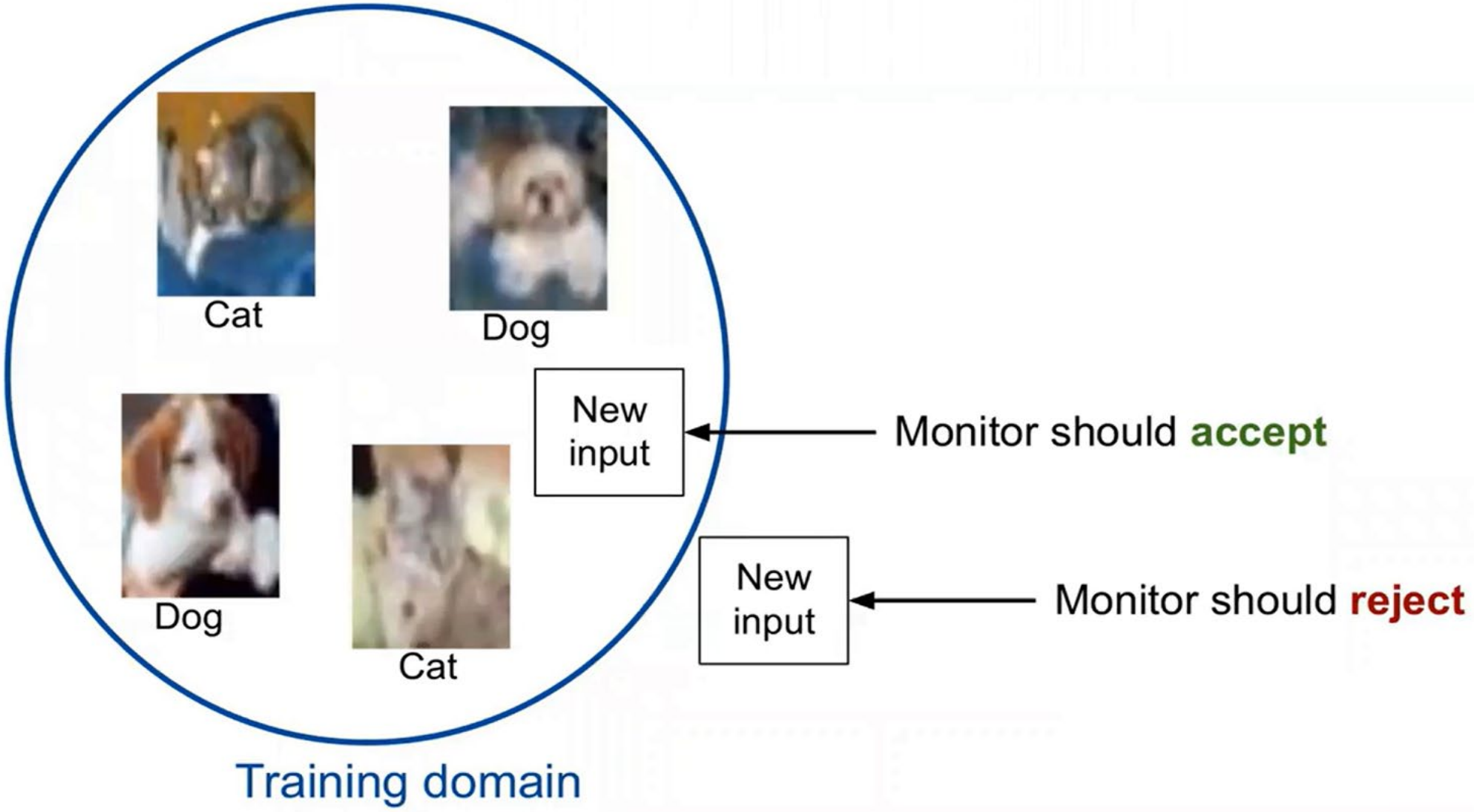
Test input may not belong to one of the K training classes.

Need to be able to say "none-of-the-above".

Image source: https://blog.google/technology/health/ai-dermatology-preview-io-2021/

Source: Jie Ren, Google AI Brain

“Monitor” the Inputs



Two Problem Formulations: OOD and Open Category

Out-of-Distribution Problem

- Training:
  - Data: $(x_1, y_1), \ldots, (x_N, y_N)$ drawn from D_0
  - $y_i \in \{1, \ldots, K\}$
- Testing:
  - Data: Mixture D_m of data from D_0 and D_a
  - $(x, y) \sim D_a$ belong to a **different data set**
- Goal:
  - Given a query x_q, does it belong to D_a or D_0?
  - If from D_a, REJECT as alien
  - Else classify using a classifier trained on D_0 data

Novel Category / Open Set Problem

- Training:
  - Data: $(x_1, y_1), \ldots, (x_N, y_N)$ drawn from D_0
  - $y_i \in \{1, \ldots, K\}$
- Testing:
  - Data: Mixture D_m of data from D_0 and D_a
  - $(x, y) \sim D_a$ belong to **new classes not seen during training ("alien categories")**
- Goal:
  - Given a query x_q, does it belong to D_a or D_0?
  - If from D_a, REJECT as alien
  - Else classify using a classifier trained on D_0 data

Source: Tom Dietterich

Two Problem Formulations: Key Difference: Evaluation Protocol

Out-of-Distribution

- Train on data from domain A
- Test on data from a mix of domain A and domain B
- Example:
  - Train on MNIST
  - Test on MNIST + Fashion-MNIST

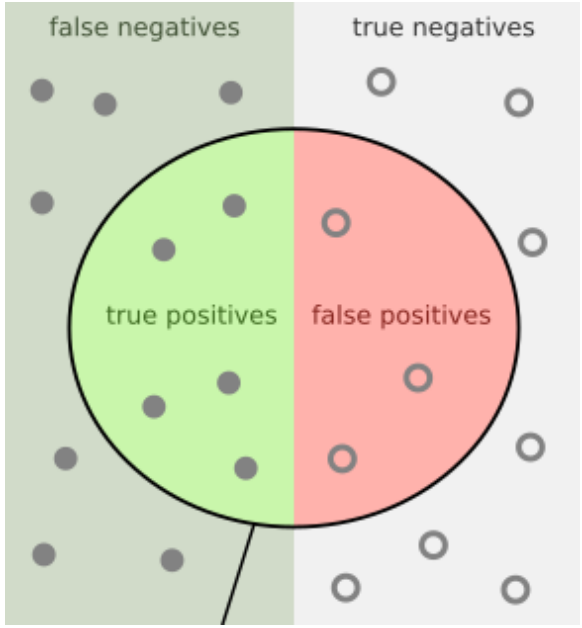
Novel Category

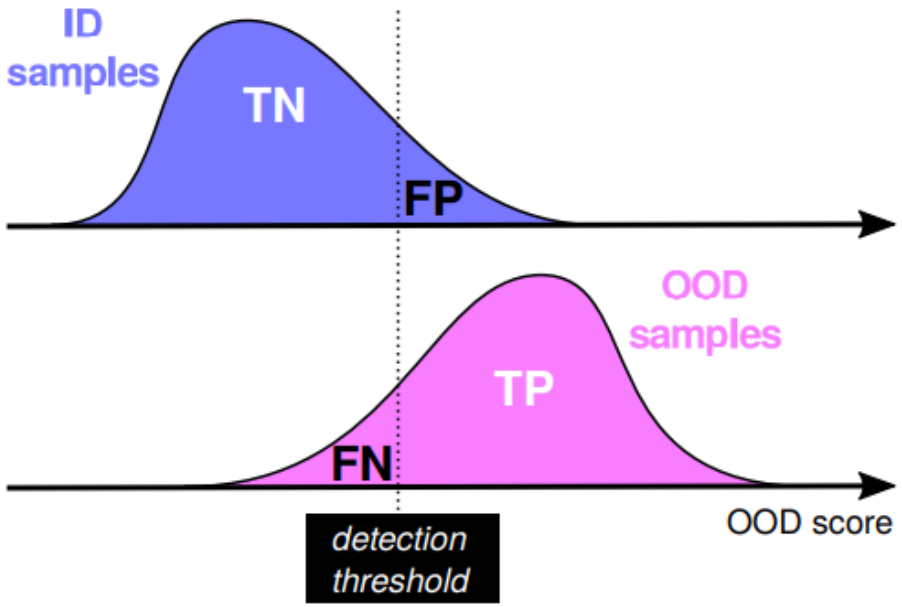
- Divide the classes of domain A into known and unknown
- Train on known classes
- Test on all classes
- Example:
  - Train on MNIST {1,2,3,4,5}
  - Test on MNIST {1,2,3,4,5,**6**,**7**,**8**,**9**,**0**}

Source: Tom Dietterich

Evaluating OOD Detection

- Treat OOD Detection as a binary classification task **(in distribution / out of distribution)**
- 4 types of predictions: TP, TN, FP, FN
- Two types of evaluation metrics:
  - Fixed Threshold Metrics
  - Threshold Independent Metrics

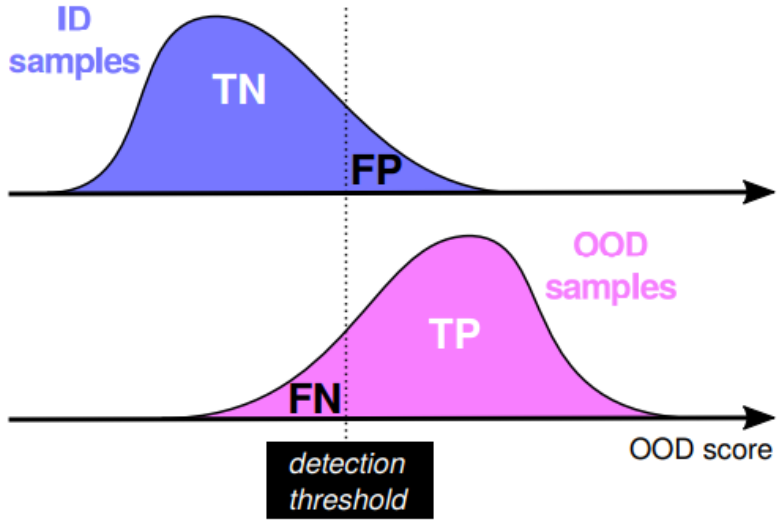




Evaluating OOD Detection

Fixed Threshold Metrics

- Consider performance at a specific operating point
  - Measure "something" that will help us distinguish between ID and OOD
    - Something = Softmax Score (Hendrycks & Gimpel ICLR 2019)
    - Something = Energy (Liu et al. NeurIPS 2020)
  - Decide a threshold to obtain a fixed TPR (usually 95%)

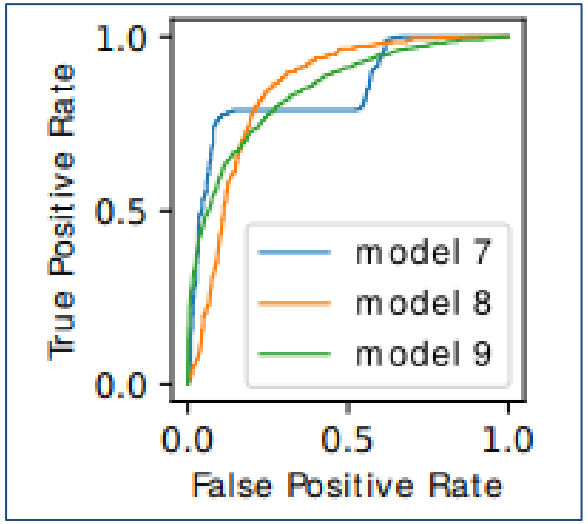


- $\boldsymbol{FPR@TPR}$ $\boldsymbol{(FPR@95)}$
  - Given a fixed TPR what is the FPR?
  - For an ideal OOD Detector, FPR is 0%

Evaluating OOD Detection

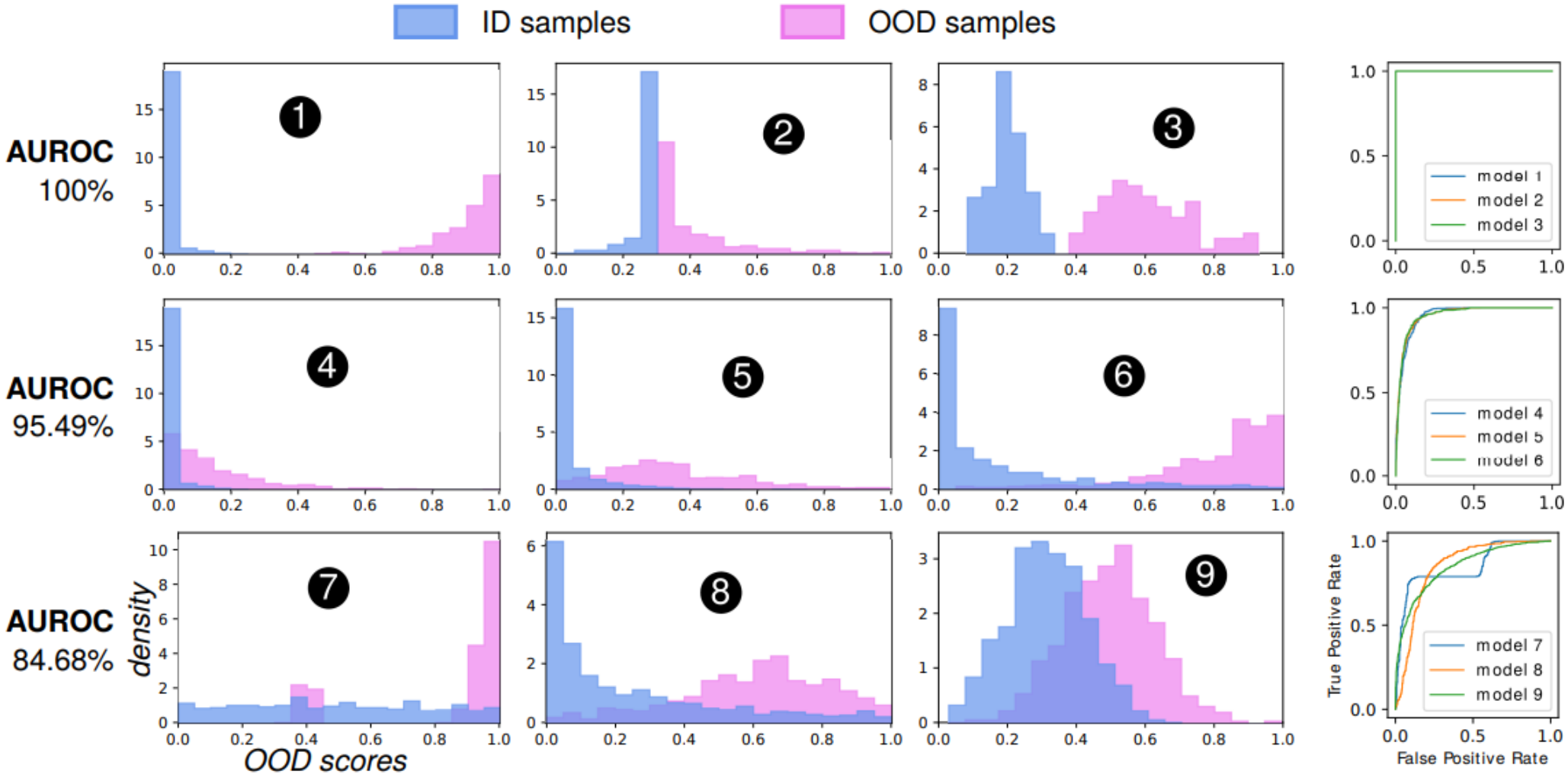
Threshold-Independent Metrics

- Plot the "Receiving Operating Characteristic" (ROC) Curve
  - Plot TPR as a function of FPR
- Measure Area under the ROC Curve **(AUROC)**
- AUROC Interpretation:
  - The probability of a positive sample being assigned a higher score than a negative sample
    - AUROC = 100% Perfect Separation between ID and OOD
    - AUROC = 50% Full Overlap (Detector is random)



Evaluating OOD Detection

Threshold-Independent Metrics



More Applications of OOD Detection

- "Selective Prediction"
  - Enable models to ***abstain*** from making decisions on certain types of inputs
  - when to abstain?
    → OOD Detection!

Post-Abstention: Towards Reliably Re-Attempting the Abstained Instances in QA

ACL 2023

Neeraj Varshney and Chitta Baral
Arizona State University

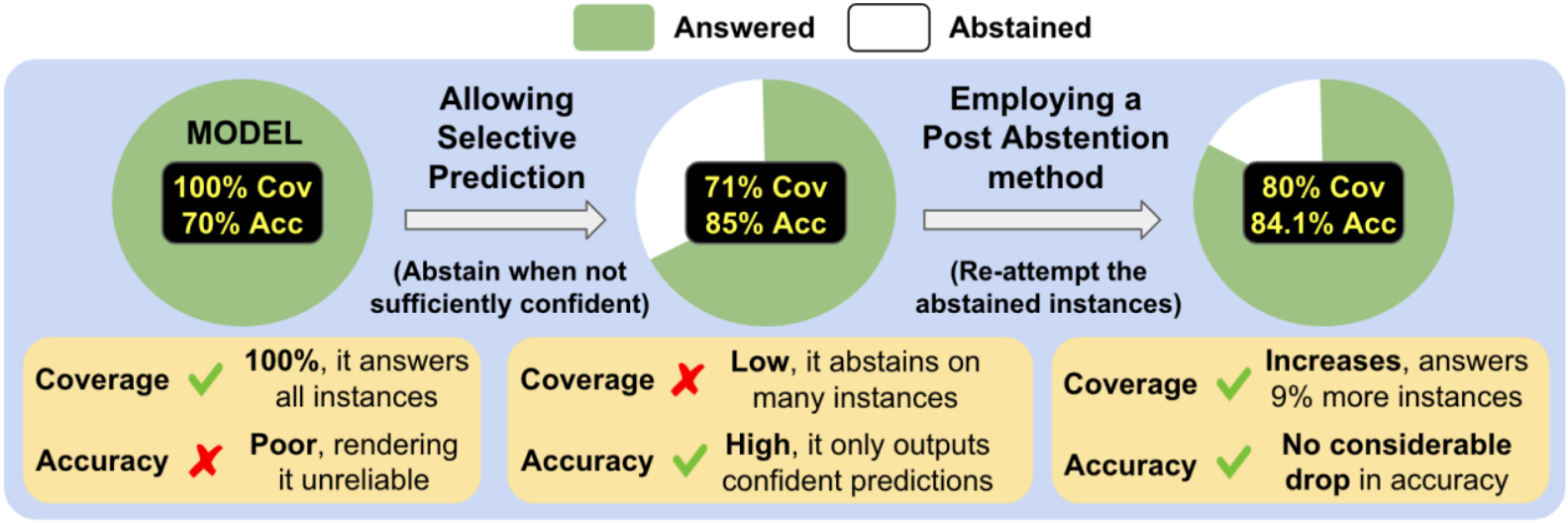


Figure 1: Illustrating the **impact of employing a post-abstention method** on top of selective prediction system. A regular model that has an accuracy of 70% (at coverage 100%) is first enabled with selective prediction ability that increases the accuracy to 85% but drops the coverage to 71%. Then, on employing a post-abstention method to the abstained instances (remaining 29%), coverage increases to 80% without a considerable drop in overall accuracy.

Two Weeks from Now:

Calibration and Uncertainty

(Closely Related to OOD/Novelty Detection)