Taming False Positives in Out-of-Distribution Detection with Human Feedback

AISTATS 2024

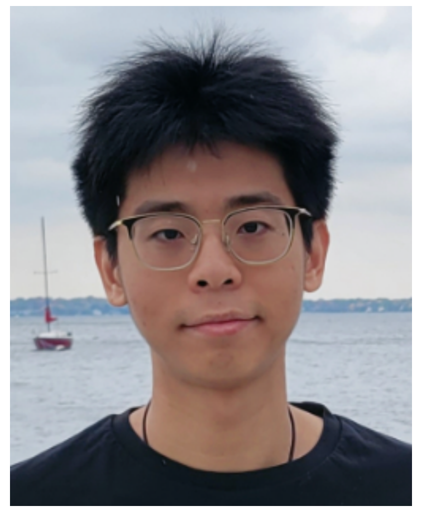
Harit Vishwakarma

CS Ph.D. Student

hvishwakarma@cs.wisc.edu

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

Joint work with



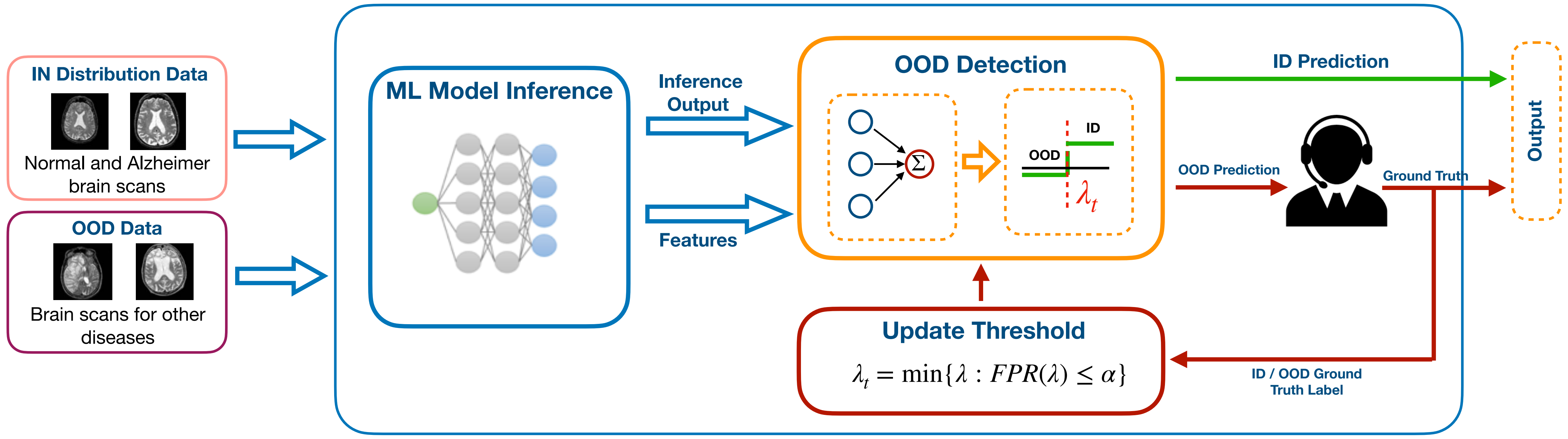
Heguang Lin

Ramya Korlakai Vinayak

TL;DR

- ML models are subject to OOD points after deployment.
- Hard to anticipate all kinds of OOD data and prepare for that.
- Prior works, construct OOD scoring function and set threshold on the scores to achieve 95% TPR
  - We observe, this leads to high FPR.
- We propose to adapt the threshold to maintain FPR below 5% at all times.
  - Use any-time valid confidence sequences to guarantee this.
  - Validate empirically.

ID : Positive
OOD : Negative

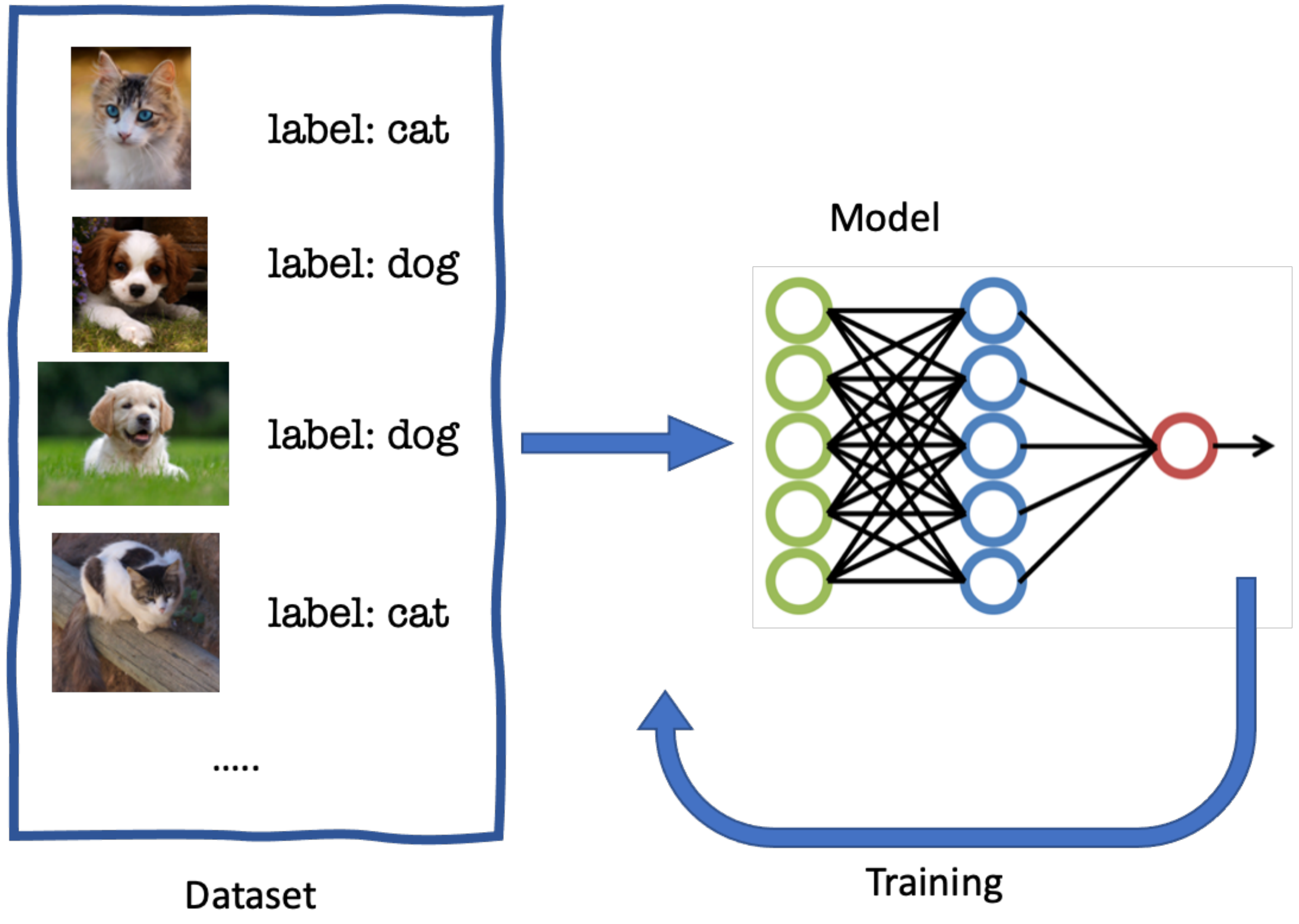


Outline

- Motivation for OOD detection and FPR control
- Our framework for human-in-the-loop OOD detection
- Theoretical guarantees on controlling FPR
- Works well in practice — experiments on synthetic and real scoring functions

Supervised machine learning (ML) Training to Deployment

- Supervised ML models are trained on labeled datasets
- Validation / Model selection on data from same distribution.
- Deploy the model after training and model selection.
- Generalization to unseen data is guaranteed when it is coming iid from the **same distribution as training data**



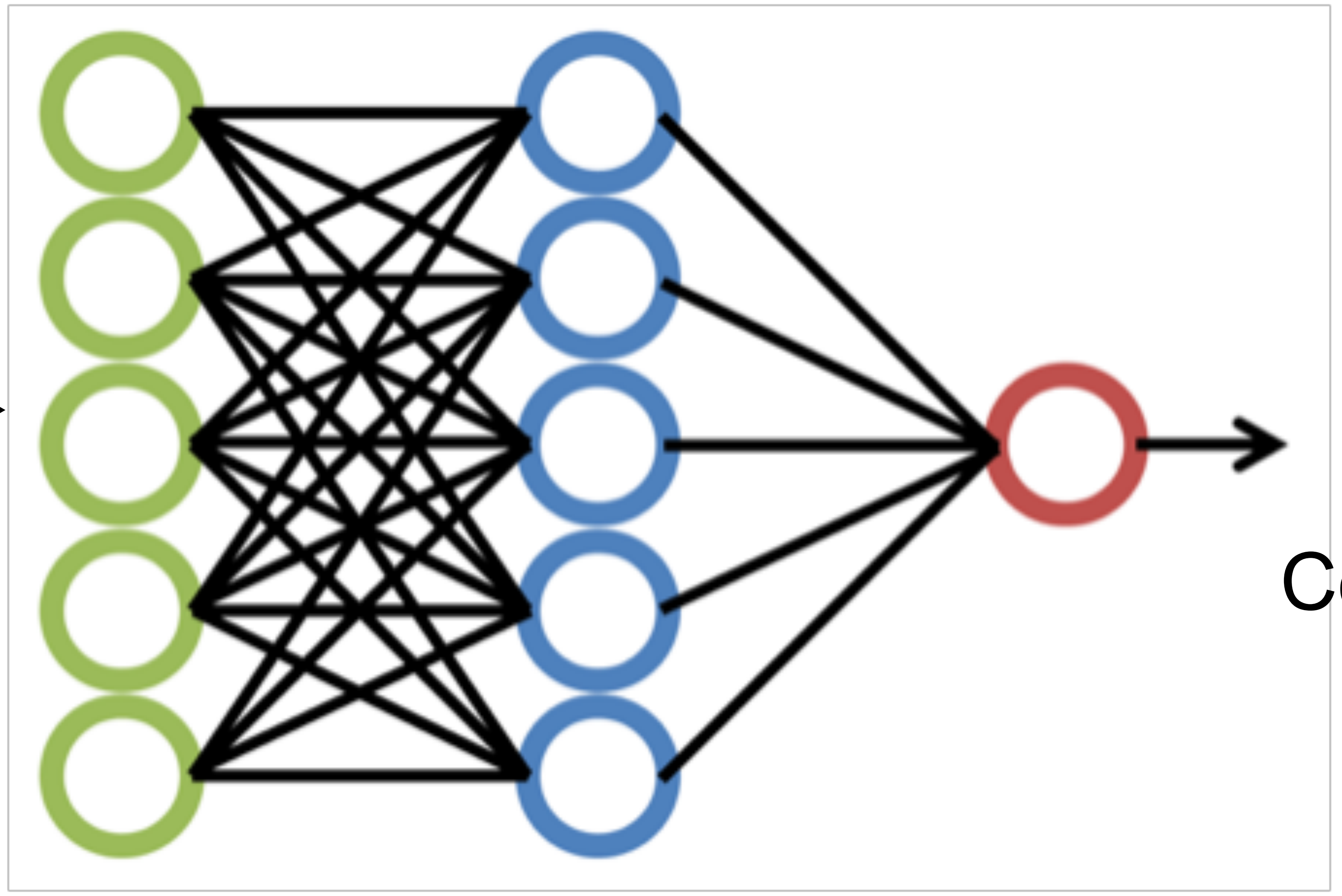
Expectation: Test data matches training data

...

In Distribution (ID)

Usually assume that the test data will come from the same distribution as ID data.

Test Input



Dog

Correct Prediction

Reality: (i) Test data might not match training data

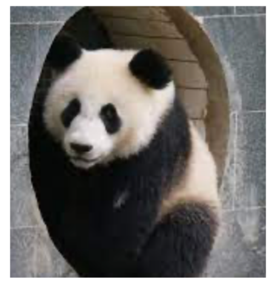
The test data may have samples from different distributions.

$\mathcal{D}_{\text{in}}$: distribution of ID data

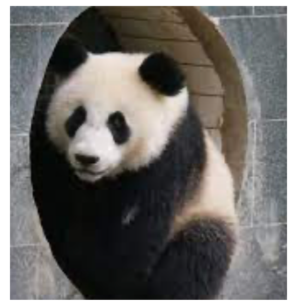
  ...

Expected Test Data

$\mathcal{D}_{\text{ood}}$: distribution of OOD data

  ...

$$x \overset{\text{i.i.d.}}{\sim} (1-\gamma)\,\mathcal{D}_{\text{in}} + \gamma\,\mathcal{D}_{\text{ood}}$$

    ...

Real Test Data

$\gamma \in (0,1)$: OOD fraction

Reality : (ii) Model makes mistakes on OOD points

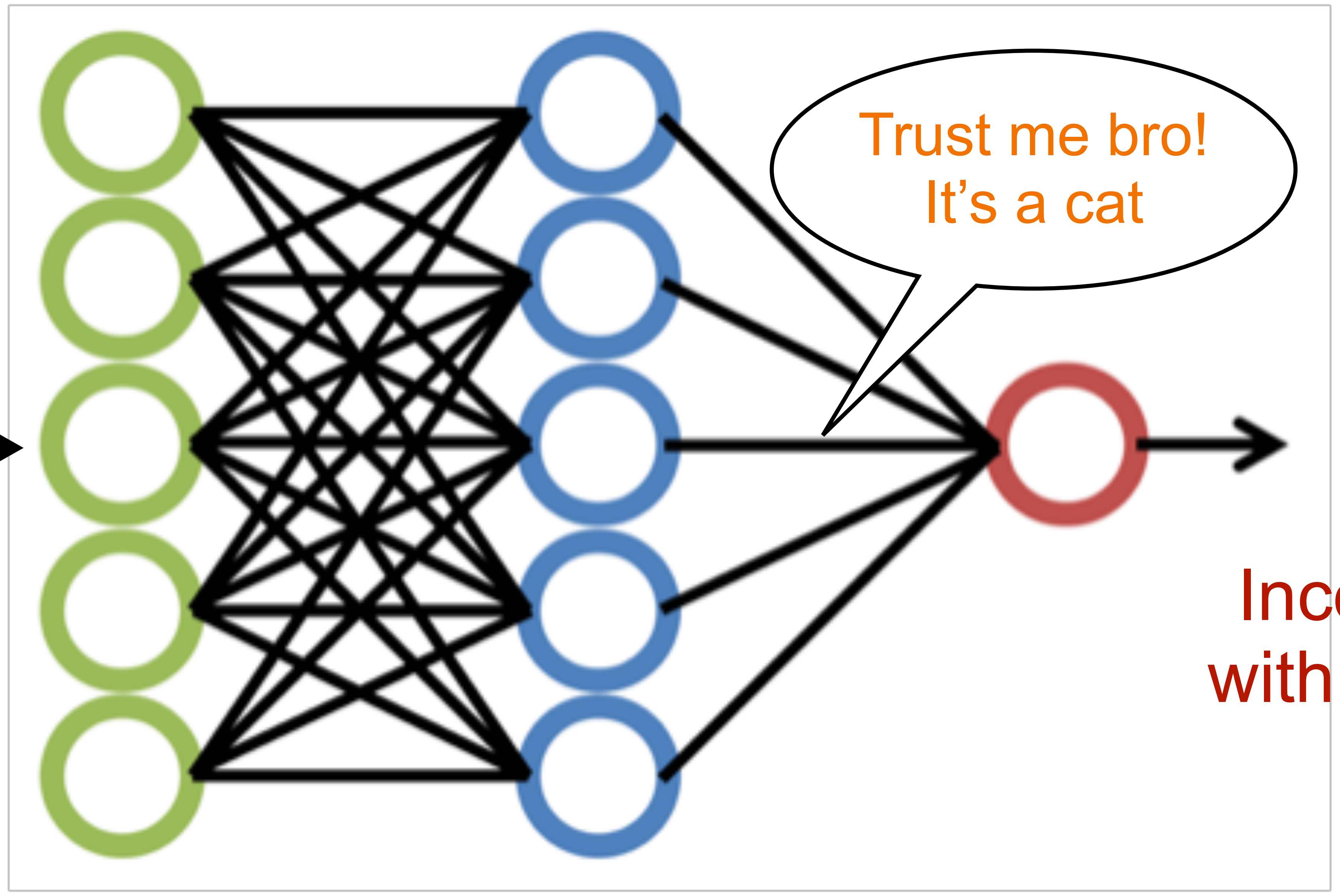
  ...

In Distribution (ID)

Reality

1. May get OOD data at test time.
2. Model can misclassify it as one of the ID classes with high confidence.

OOD Input

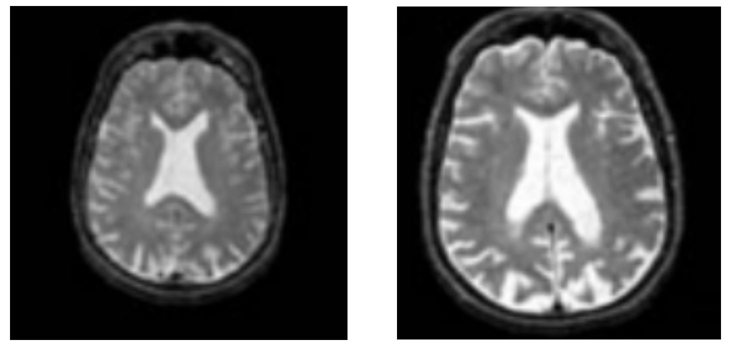


Cat

Incorrect Prediction
with High Confidence

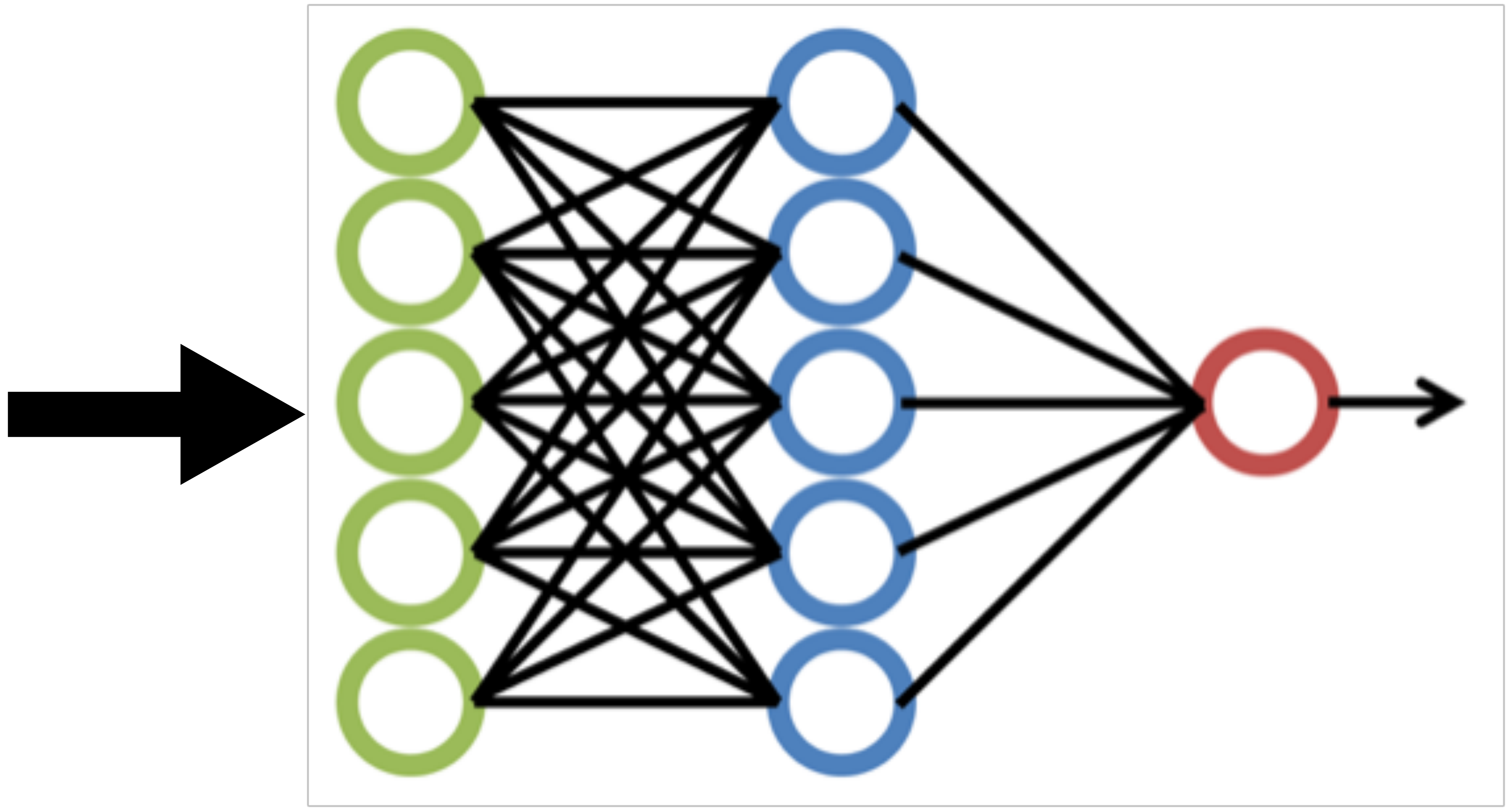
Nguyen et. al, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images ", 2017

A more safety critical example



...

ID: Normal and Alzheimer brain scans



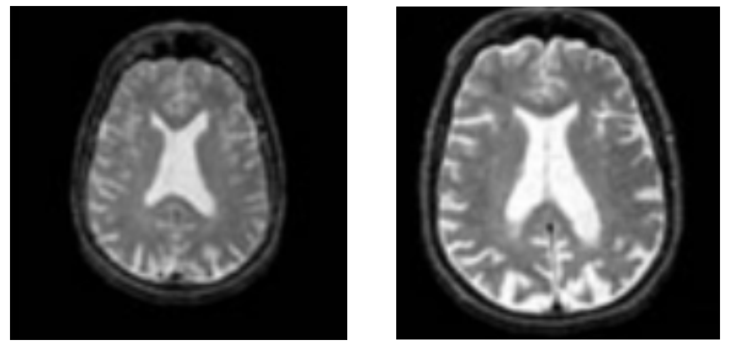
Accurate Predictions
on ID data

ML model to classify brain scans with
Alzheimer vs **Normal** scans

Since it is trained on ID data, assume it is highly accurate on it.

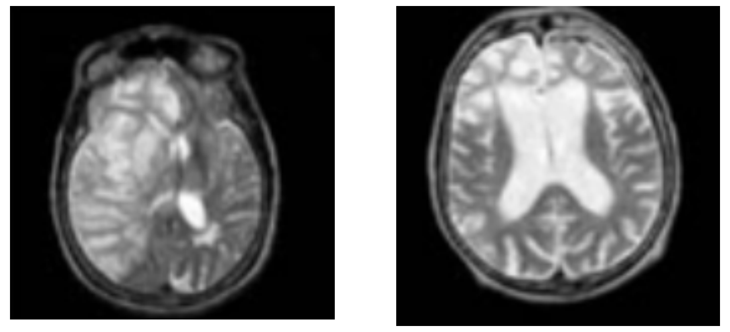
A more safety critical example

ID : Positive
OOD : Negative



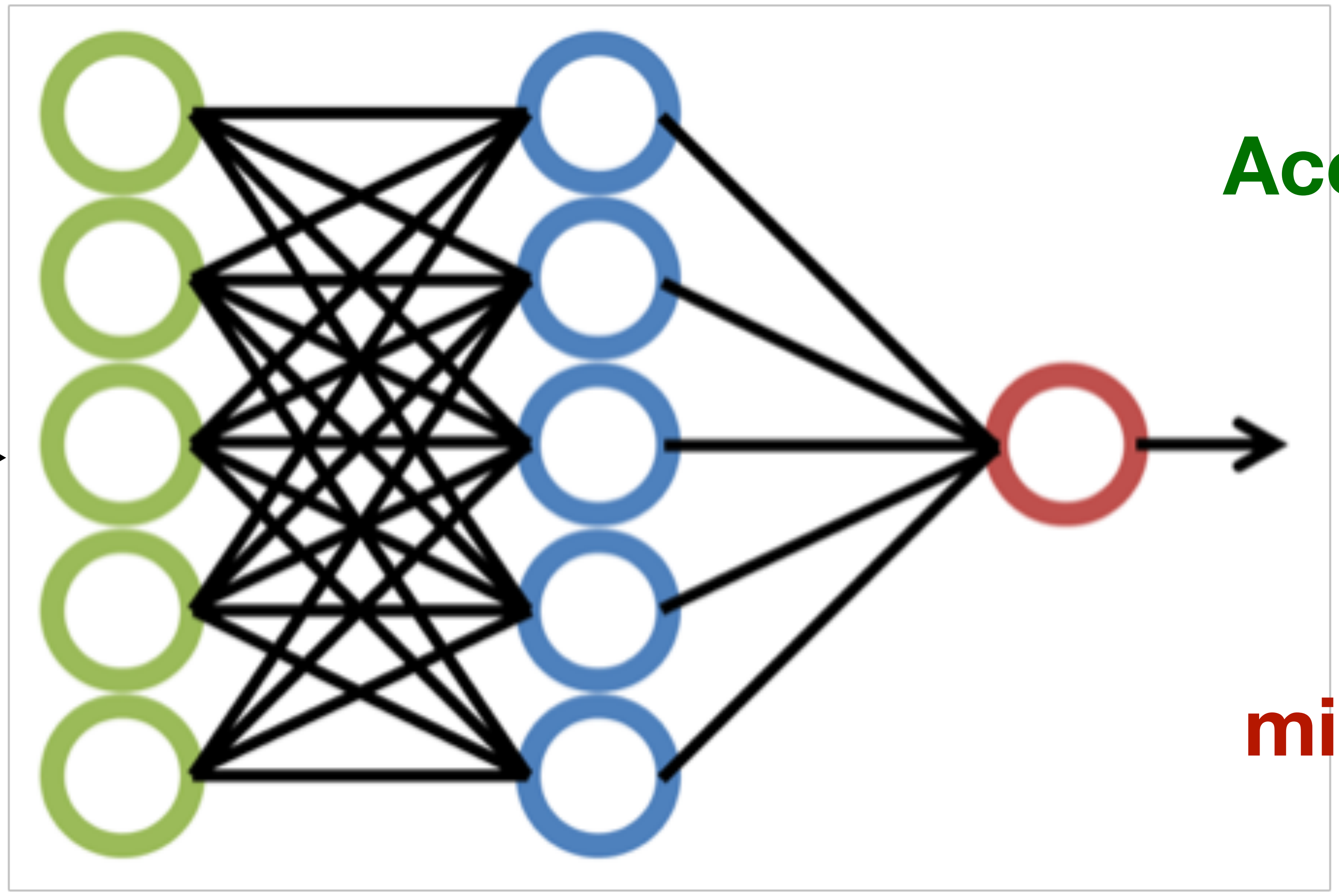
...

ID: Normal and Alzheimer brain scans



...

OOD : Brain scans for other diseases



Accurate Predictions on ID data

Likely to make mistakes on OOD data

It would be catastrophic to misclassify a **scan of other disease (OOD)** as having Alzheimer or as a Normal scan (ID).

OOD misclassified as ID is a False Positive.

Reality of ML model deployment

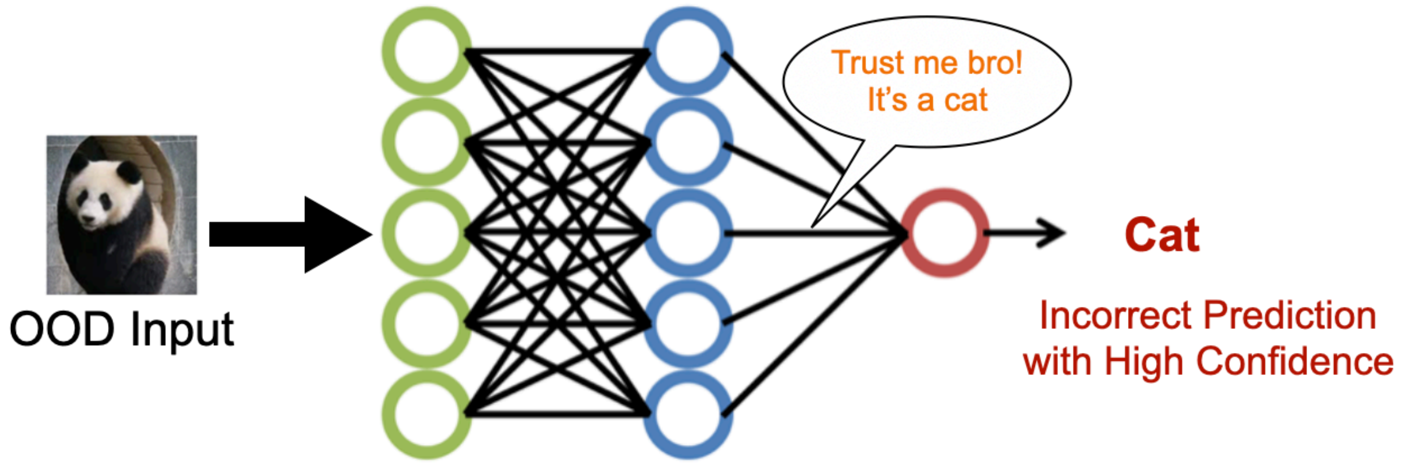
- ML models could be subject to OOD points

...

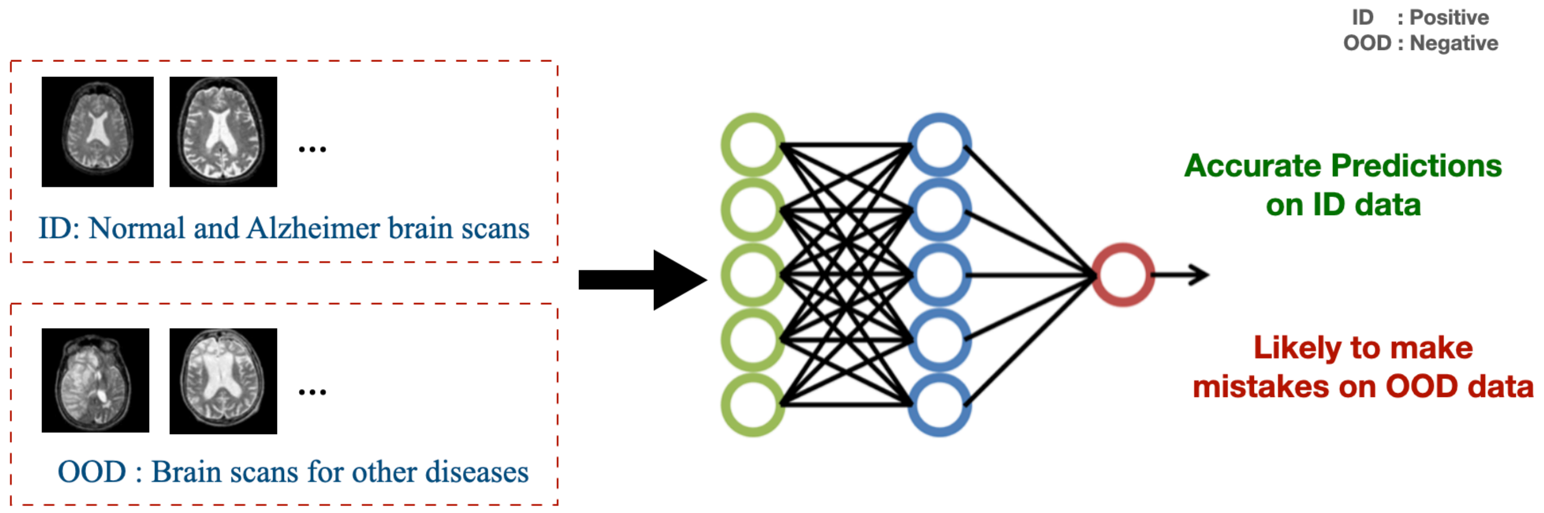
$$x \overset{\text{i.i.d.}}{\sim} (1-\gamma)\ \mathcal{D}_{\text{in}} + \gamma\ \mathcal{D}_{\text{ood}}$$

$\gamma \in (0,1)$: OOD fraction

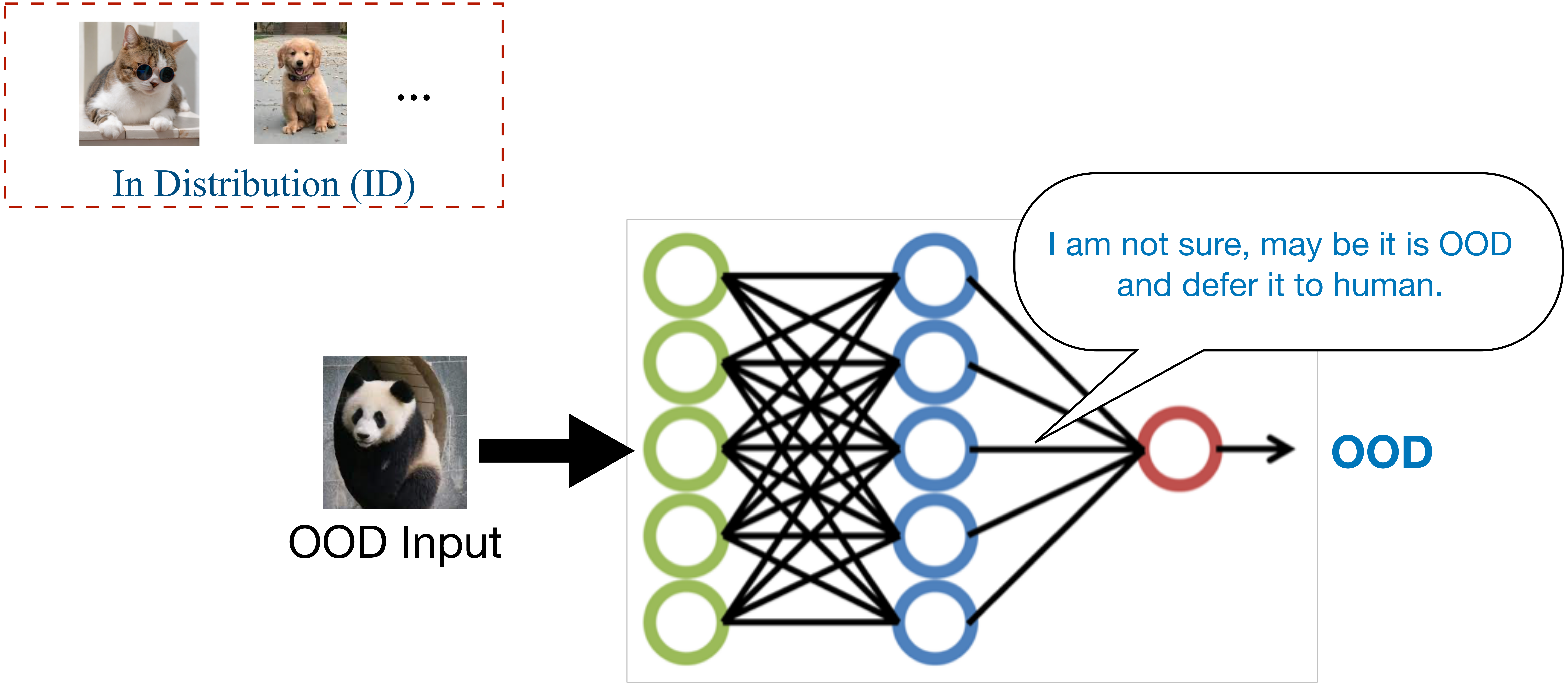
- They can misclassify OOD points as an ID class with high confidence



- The mistakes (false positives) could be serious.



What should we expect on OOD inputs?



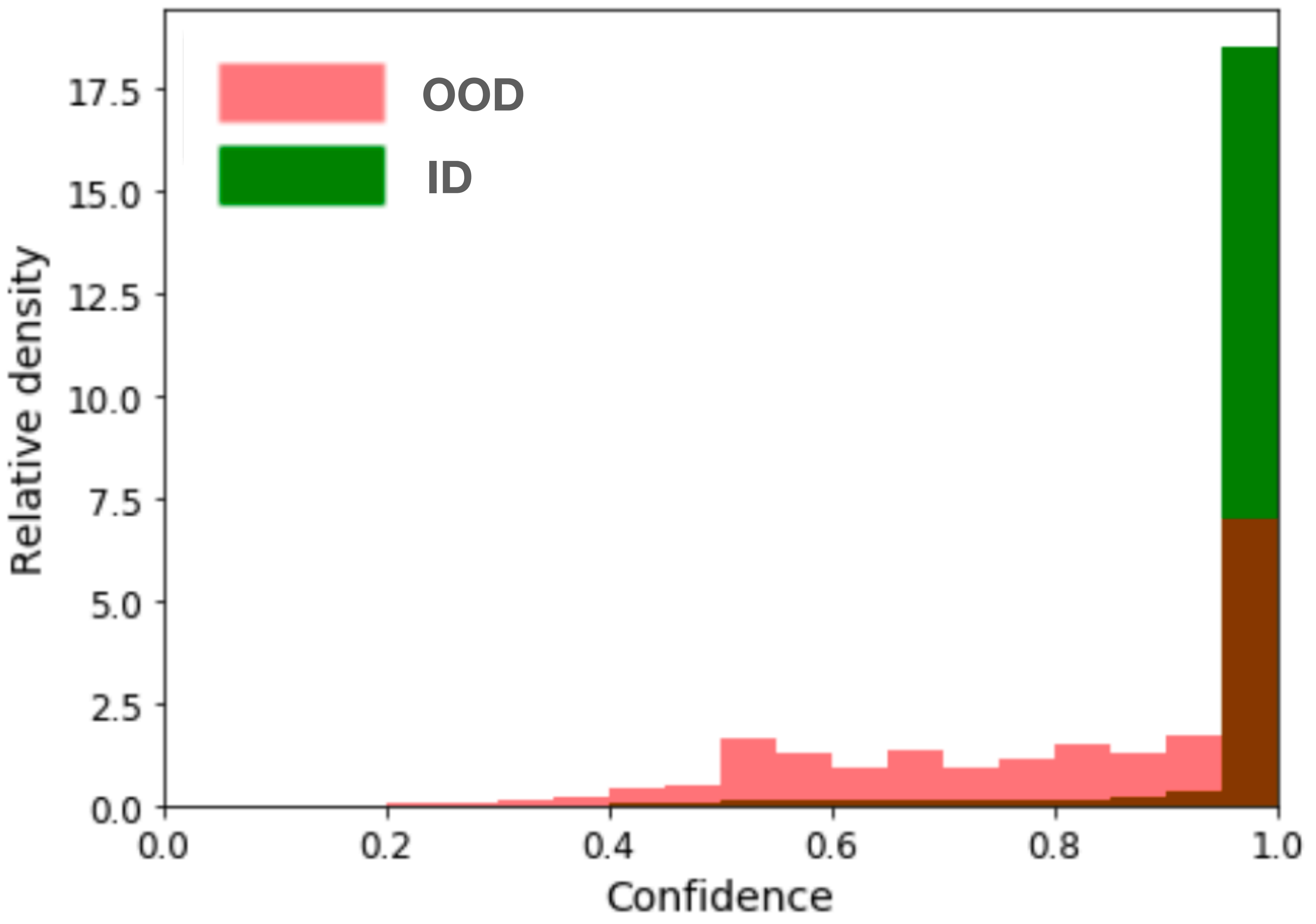
OOD detection with post-hoc methods

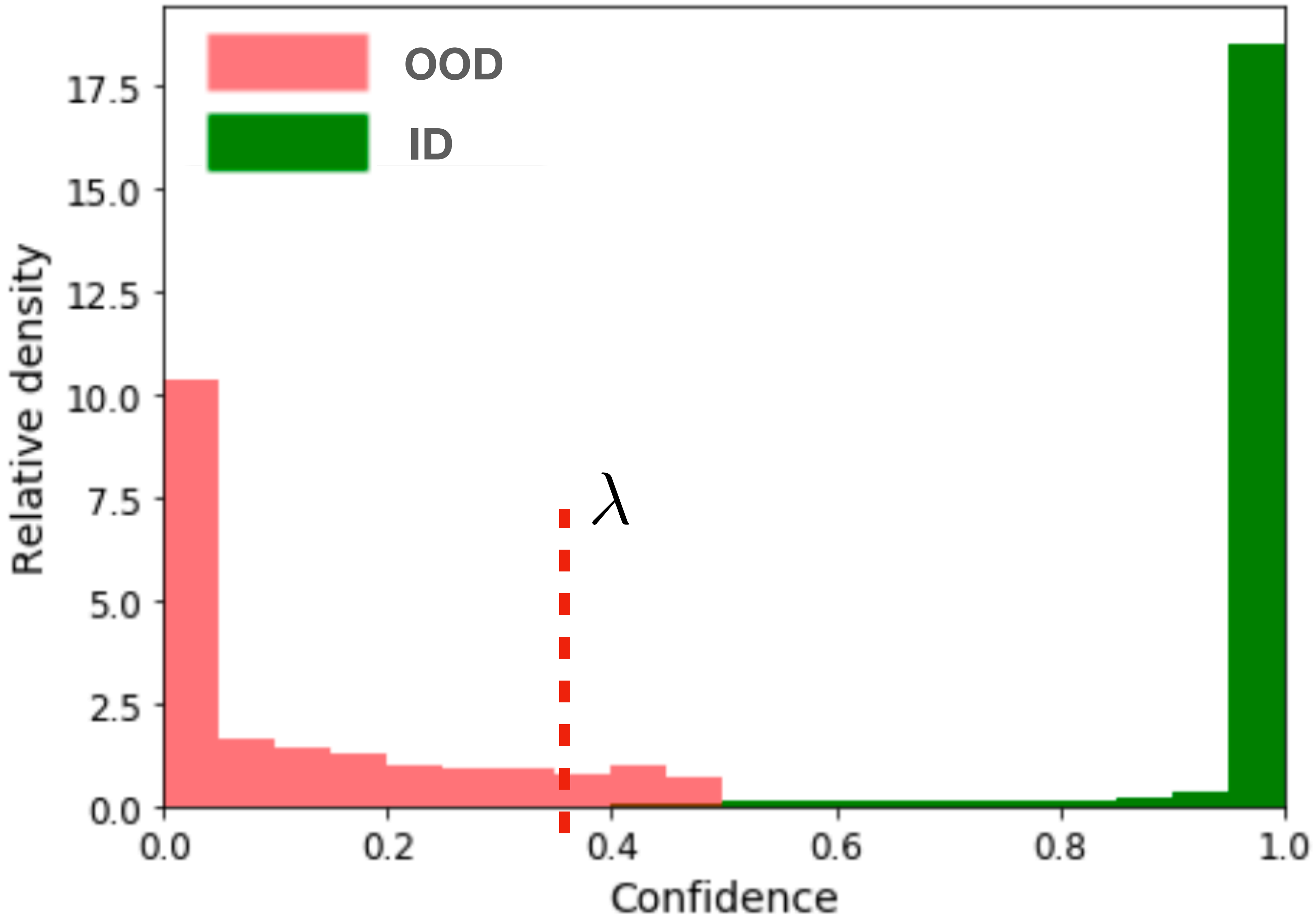
- Scoring function: $g : \mathcal{X} \Rightarrow [\Lambda_{\min}, \Lambda_{\max}] \subset \mathbb{R}$
- Select Threshold λ to achieve 95% TPR.

Declare "in-distribution" (ID) if $g(x) \geq \lambda$

Declare "out-of-distribution" if $g(x) < \lambda$

Usual Softmax Confidence Scores

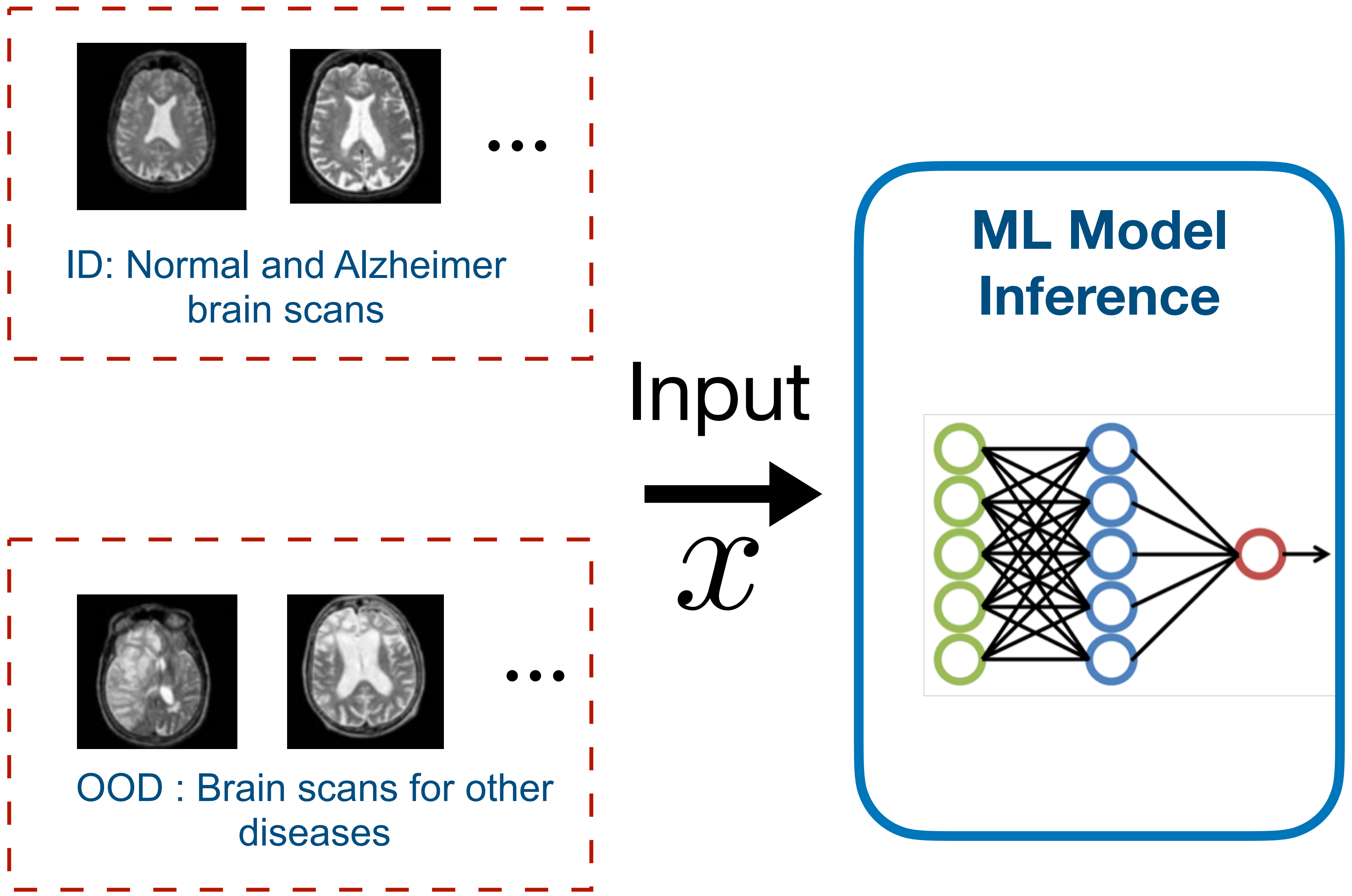


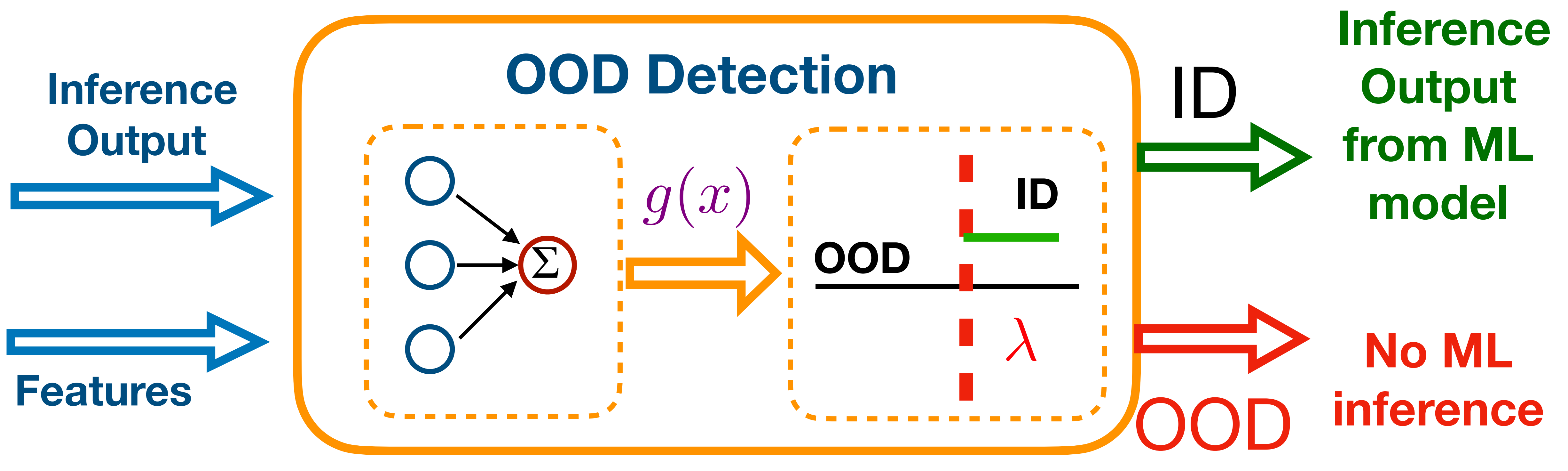


Yang et. al, "Generalized OOD detection: A Survey", 2021

Yang et. al, "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection", 2022

OOD detection with post-hoc methods





$x \overset{\text{i.i.d.}}{\sim} (1-\gamma)\, \mathcal{D}_{\text{in}} + \gamma\, \mathcal{D}_{\text{ood}}$

$\mathcal{D}_{\text{in}}$: distribution of ID data

$\mathcal{D}_{\text{ood}}$: distribution of OOD data

$\gamma \in (0,1)$: OOD fraction

$\text{TPR}(\lambda) := \text{E}_{x\sim\mathcal{D}_{\text{in}}}[\mathbf{1}\{g(x) > \lambda\}]$

- Scoring function: $g : \mathcal{X} \Rightarrow [\Lambda_{\min}, \Lambda_{\max}] \subset \mathbb{R}$
- Select Threshold λ to achieve 95% TPR.

Declare "in-distribution" (ID) if $g(x) \geq \lambda$

Declare "out-of-distribution" if $g(x) < \lambda$

Yang et. al, "Generalized OOD detection: A Survey", 2021

Yang et. al, "OpenOOD: Benchmarking Generalized Out-of-Distribution Detection", 2022

False Positive and True Positive Rates

Scoring function $g : \mathcal{X} \rightarrow [\Lambda_{\min}, \Lambda_{\max}] \subset \mathbb{R}$ Threshold: λ

- **False Positive Rate**

$$\mathrm{FPR}(\lambda) := \mathrm{E}_{x \sim \mathcal{D}_{\mathrm{ood}}} \left[\mathbf{1}\{g(x) > \lambda\} \right]$$

$\mathcal{D}_{\mathrm{ood}}$: distribution of OOD data

Fraction of OOD data that falsely get considered as “ID”

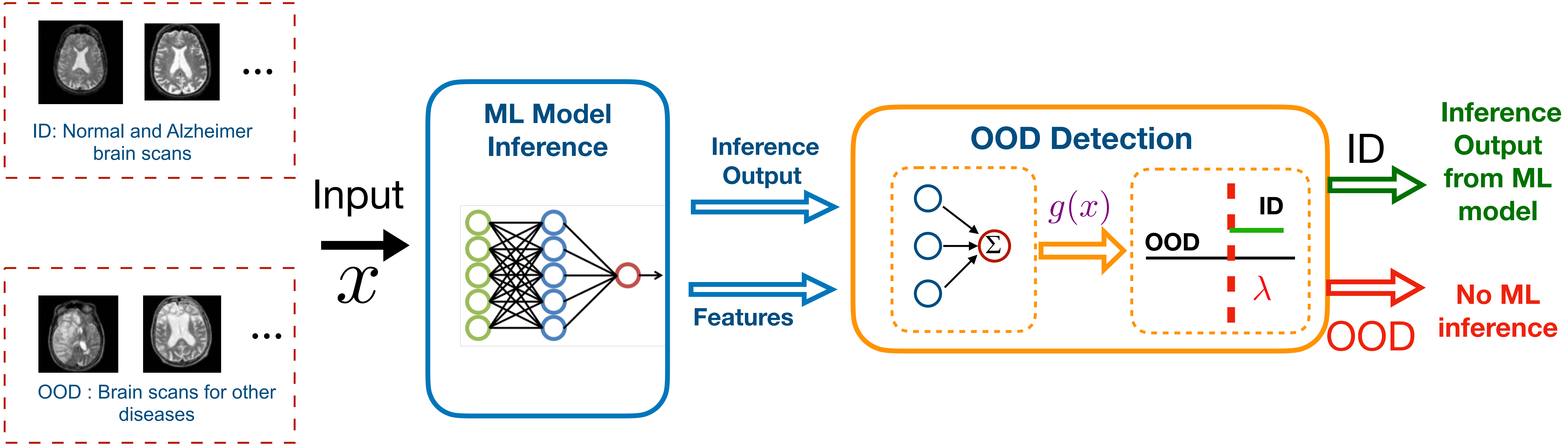
- **True Positive Rate**

$$\mathrm{TPR}(\lambda) := \mathrm{E}_{x \sim \mathcal{D}_{\mathrm{in}}} \left[\mathbf{1}\{g(x) > \lambda\} \right]$$

$\mathcal{D}_{\mathrm{in}}$: distribution of ID data

Fraction of ID data that correctly get considered as “ID”

Safe use in critical applications require guarantees on false positives



$x \overset{\text{i.i.d.}}{\sim} (1-\gamma)\, \mathcal{D}_{\text{in}} + \gamma\, \mathcal{D}_{\text{ood}}$

$\mathcal{D}_{\text{in}}$: distribution of ID data

$\mathcal{D}_{\text{ood}}$: distribution of OOD data

$\gamma \in (0,1)$: OOD fraction

$\text{TPR}(\lambda) := \text{E}_{x\sim\mathcal{D}_{\text{in}}}[\mathbf{1}\{g(x) > \lambda\}]$

It would be catastrophic to misclassify a scan of **other disease (OOD)** as having **Alzheimer or as a Normal scan (ID).**

$$\Pr(\text{declare as "ID"} \,|\, x \text{ is "OOD"}) \leq \alpha$$

$$\text{FPR}(\lambda) := \text{E}_{x\sim\mathcal{D}_{\text{ood}}}[\mathbf{1}\{g(x) > \lambda\}] \leq \alpha$$

Threshold selection and FPR

- Usually, threshold is picked such that 95% of ID data is correctly identified as ID, that is TPR is 95%. **But the FPR at this point can very large**

OpenOOD Full Results (FPR/AUROC/AUPR)

File Edit View Insert Format Data Tools Extensions Help

Share Sign in

100% View only

A1 Method

	A	B	C	D	E	F	G
1	Method	CIFAR-100	TIN	NearOOD	MNIST	SVHN	Texture
2	OpenMax	67.62 / 85.03 / 83.69	64.55 / 86.57 / 85.93	66.09 / 85.80 / 84.81	57.79 / 90.12 / 68.66	71.60 / 84.29 / 65.26	69.18 / 83.15 /
3	MSP	62.01 / 87.11 / 85.92	60.69 / 86.62 / 83.07	61.35 / 86.87 / 84.50	58.59 / 89.91 / 66.95	51.87 / 90.88 / 78.19	59.89 / 88.72 /
4	ODIN	59.09 / 77.68 / 73.24	59.06 / 77.33 / 70.07	59.07 / 77.51 / 71.66	36.23 / 90.91 / 64.74	67.92 / 73.32 / 42.13	51.10 / 80.70 /
5	MDS	81.63 / 66.30 / 63.74	83.76 / 66.79 / 63.28	82.70 / 66.54 / 63.51	0.00 / 99.52 / 99.24	19.69 / 95.78 / 91.00	19.61 / 95.42 /
6	Gram	100 / 39.76 / 59.04	92.43 / 58.11 / 54.72	91.09 / 58.57 / 56.18	76.04 / 77.59 / 43.97	73.21 / 79.28 / 55.46	89.01 / 57.72 /
7	EBO	51.46 / 86.15 / 83.21	45.02 / 88.58 / 86.37	48.24 / 87.36 / 84.79	44.50 / 90.59 / 63.28	44.94 / 88.39 / 66.29	48.32 / 86.85 /
8	GradNorm	82.00 / 54.80 / 52.39	82.07 / 54.75 / 49.54	82.03 / 54.78 / 50.97	77.27 / 59.84 / 20.83	82.38 / 48.96 / 22.78	83.07 / 48.49 /
9	ReAct	53.72 / 86.35 / 83.15	47.00 / 88.90 / 86.53	50.36 / 87.62 / 84.84	50.94 / 88.34 / 50.88	49.23 / 89.50 / 75.36	49.98 / 88.18 /
10	MLS	52.16 / 86.10 / 83.20	49.19 / 86.11 / 80.79	50.67 / 86.11 / 82.00	45.23 / 90.48 / 63.22	44.63 / 88.45 / 66.33	48.63 / 86.86 /
11	KLM	61.99 / 78.71 / 72.88	60.38 / 79.10 / 70.73	61.18 / 78.90 / 71.81	61.49 / 82.36 / 40.65	50.77 / 85.95 / 70.01	59.24 / 83.28 /
12	VIM	55.92 / 87.15 / 86.34	52.00 / 88.90 / 88.63	53.96 / 88.03 / 87.48	63.63 / 87.46 / 60.66	14.41 / 97.22 / 93.76	20.78 / 96.06 /
13	KNN	52.49 / 89.55 / 89.78	46.66 / 91.41 / 92.38	49.58 / 90.48 / 91.08	50.08 / 91.63 / 77.11	33.32 / 95.13 / 92.31	46.01 / 92.77 /
14	DICE	65.98 / 80.25 / 79.23	63.00 / 81.85 / 80.37	64.49 / 81.05 / 79.80	51.26 / 89.65 / 66.27	67.78 / 86.43 / 73.19	67.48 / 80.14 /

Yang et. al, “OpenOOD: Benchmarking Generalized Out-of-Distribution Detection”, 2022

Recap: Main Challenges

- ML models could be **subject to OOD points**
- They can **misclassify OOD points** as an ID class with high confidence
- We **do not have all type of OOD data during training** / development
  - It is observed after deployment
  - It could keep **changing over time**
- Safety critical applications demand **strict control over False Positives** i.e. misclassifying OOD as ID.

Recap: Main Challenges

Focus of prior works

- ML models could be **subject to OOD points**
- They can **misclassify OOD points** as an ID class with high confidence

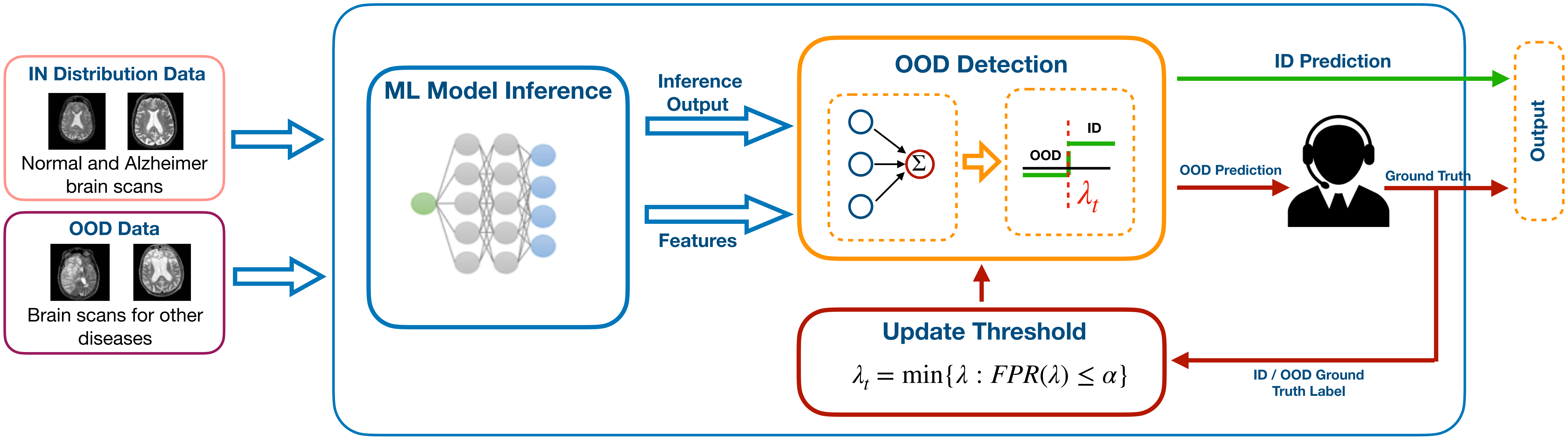
Our work's focus

- We **do not have all type of OOD data during training** / development
  - It is observed after deployment
  - It could keep **changing over time.**
- Safety critical applications demand **strict control over False Positives** i.e. misclassifying OOD as ID.

Our Solution

- Framework for **OOD detection with false positive rate control** with human-in-the-loop
- This framework **can work with any scoring functions g**
- **Theoretical guarantees for FPR control** for all time when OOD is not shifting
- **Window based approach** when OOD is shifting

Human-in-the-loop OOD Detection



- Goal: Control FPR and maximize TPR
- Maximize TPR = minimize threshold
- True Positive Rate:

$$\mathrm{TPR}(\lambda) := \mathrm{E}_{x \sim \mathcal{D}_{\text{in}}} \left[\mathbf{1}\{g(x) > \lambda\}\right]$$

Ideal Threshold selection

$$\lambda_t := \arg\min_{\lambda} \lambda$$

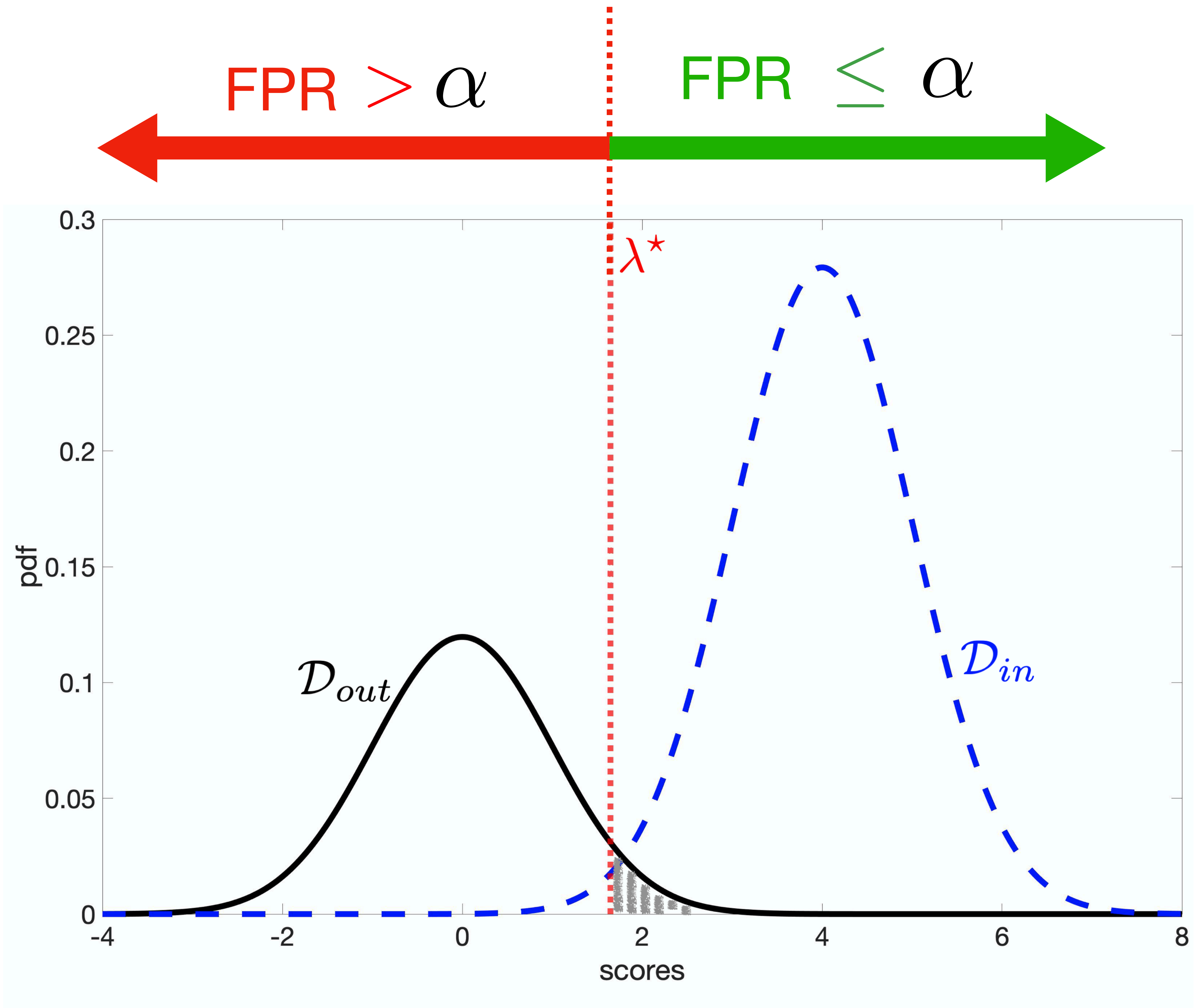
s.t. $\mathrm{FPR}(\lambda) \leq \alpha$

$$\lambda_t := \arg\min_{\lambda} \lambda$$

s.t. $\mathrm{E}_{x\sim\mathcal{D}_{\mathrm{ood}}}\left[\mathbf{1}\{g(x) > \lambda\}\right] \leq \alpha$

$$\lambda^\star := \arg\min_{\lambda} \lambda$$

s.t. $\mathrm{E}_{x\sim\mathcal{D}_{\mathrm{ood}}}\left[\mathbf{1}\{g(x) > \lambda\}\right] \leq \alpha$



$\mathrm{FPR}(\lambda^\star) = \alpha \longleftrightarrow \mathrm{CDF}_{\mathcal{D}_{out}}(\lambda^\star) = 1 - \alpha$

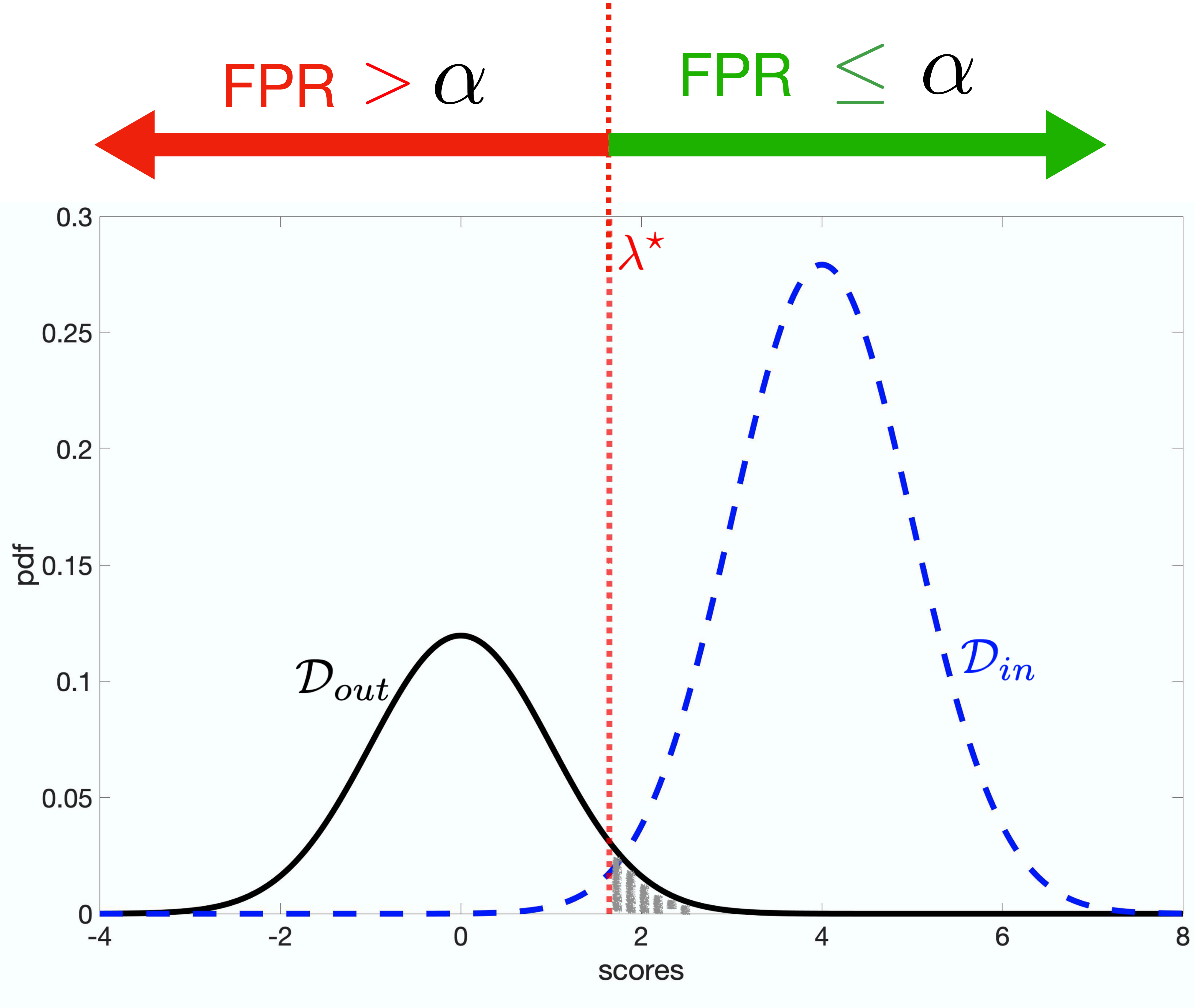
Updating threshold in each round
Idea 1: Using empirical estimate of FPR

$$\lambda_t := \arg\min_{\lambda} \lambda$$

$$\text{s.t.} \quad \widehat{\text{FPR}}(\lambda, t) \le \alpha$$

Estimate of FPR at **all** λ at time t

Not good enough to provide guarantee on FPR since **empirical estimate can sometimes underestimate the true FPR**



$$\text{FPR}(\lambda^\star) = \alpha \longleftrightarrow \text{CDF}_{\mathcal{D}_{out}}(\lambda^\star) = 1 - \alpha$$

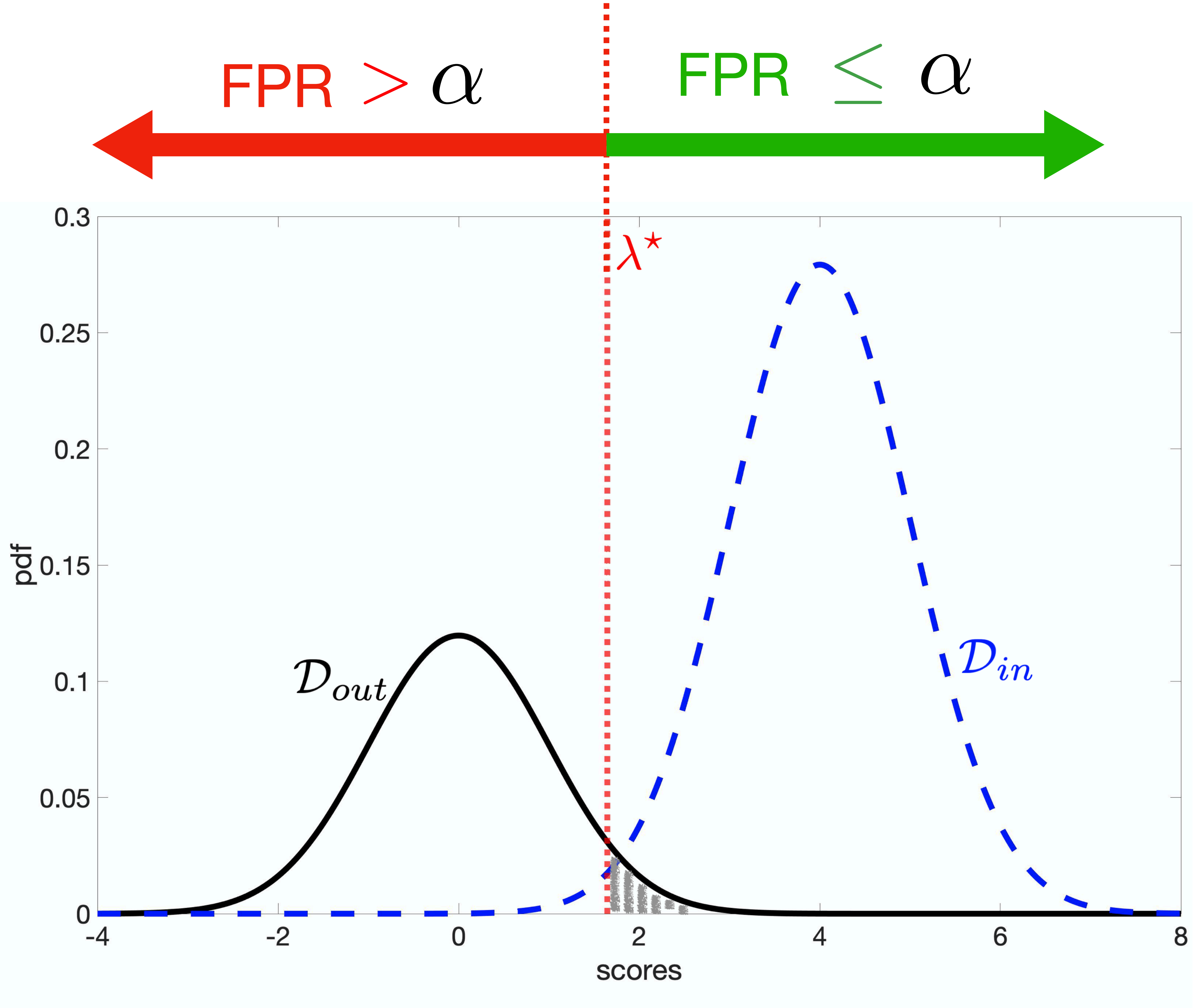
Updating threshold in each round
Idea 2: Empirical estimate with confidence

$$\lambda_t := \arg\min_{\lambda} \lambda$$

$$\text{s.t.} \quad \widehat{\mathrm{FPR}}(\lambda, t) + \boxed{\psi(t, \delta)} \le \alpha$$

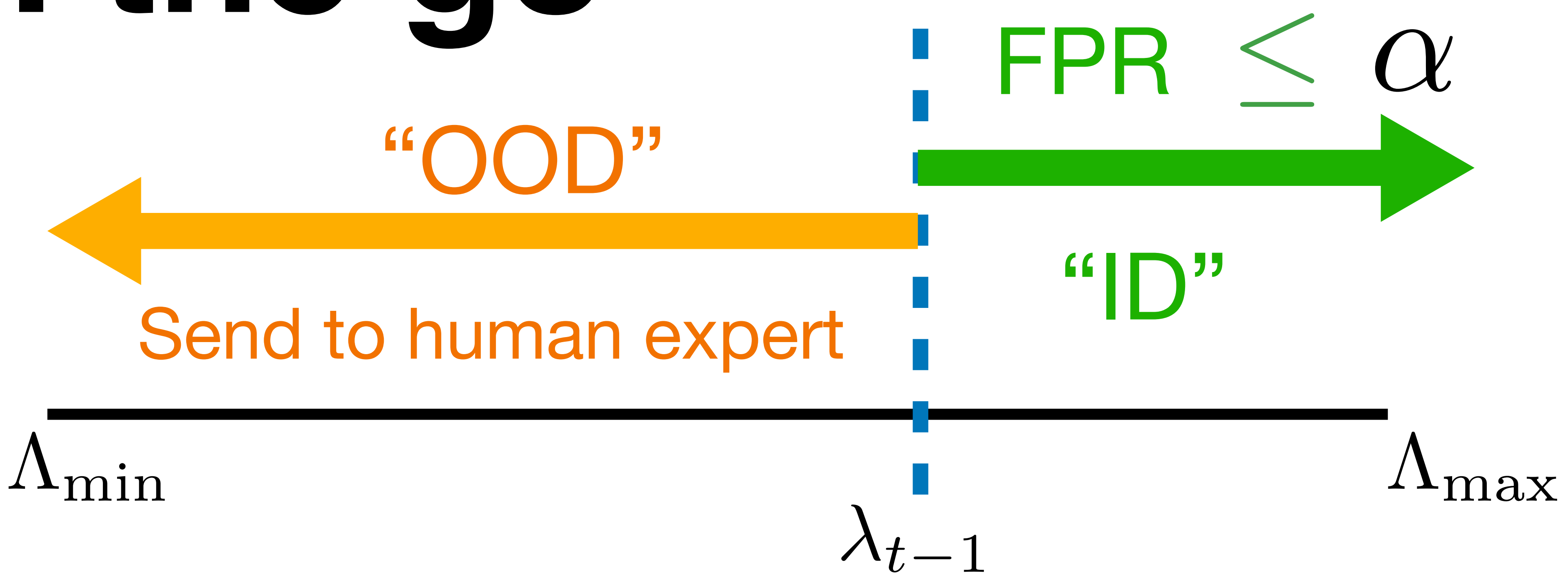
Time-varying confidence interval that is **valid for all time and all λ**

Guaranteed to approach optimal lambda from the right, so the true FPR is always guaranteed to be below the required rate



$$\mathrm{FPR}(\lambda^\star) = \alpha \longleftrightarrow \mathrm{CDF}_{\mathcal{D}_{out}}(\lambda^\star) = 1 - \alpha$$

Setting threshold on the go



In the beginning, the threshold is set at $\Lambda_{\max}$

At each time t: $x_t \overset{\text{i.i.d.}}{\sim} (1-\gamma)\,\mathcal{D}_{\text{in}} + \gamma\,\mathcal{D}_{\text{ood}}$

- Compute the **score** for the input: $s_t = g(x_t)$
- If $s_t < \lambda_{t-1}$, then predict OOD and send to human expert, get back true label
- If $s_t \geq \lambda_{t-1}$, then predict ID and query human expert for true label with probability p
- Update threshold: $\lambda_t := \arg\min_{\lambda \in \Lambda} \text{ s.t. } \widehat{\text{FPR}}(\lambda, t) + \psi(t, \delta) \leq \alpha$

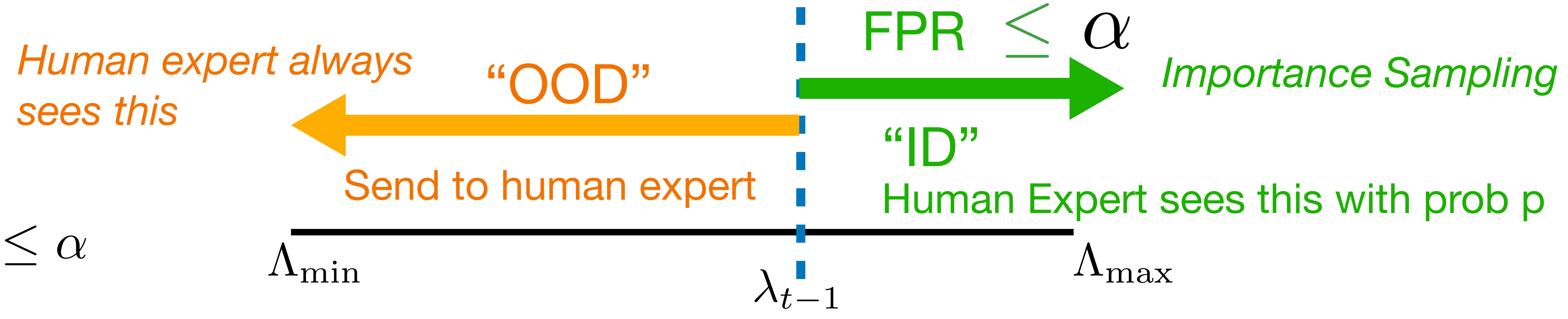
Estimate of FPR at **all** λ at time t

Time-varying confidence interval that is **valid for all time and all** λ

Estimating FPR at all thresholds

$$\lambda_t := \arg\min_{\lambda} \lambda$$

s.t. $\widehat{\mathrm{FPR}}(\lambda, t) + \psi(t, \delta) \leq \alpha$



$$\boxed{\widehat{\mathrm{FPR}}(\lambda, t) = \frac{1}{N_t^{(o)}} \sum_{u \in I_t^{(o)}} Z_u(\lambda)}$$

$$\boxed{\mathrm{E}\left[\widehat{\mathrm{FPR}}(\lambda, t)\right] = \mathrm{FPR}(\lambda, t)}$$ *Unbiased estimate*

$$Z_u(\lambda) := \begin{cases} \mathbf{1}(s_u^{(o)} > \lambda), & \text{if } s_u^{(o)} \leq \hat{\lambda}_{u-1} \\ \frac{1}{p}\mathbf{1}(s_u^{(o)} > \lambda), & \text{w.p. } p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \\ 0, & \text{w.p. } 1-p \text{ if } s_u^{(o)} > \hat{\lambda}_{u-1} \end{cases}$$

- Recall that human expert always sees a point that is declared OOD
- We also ask for human expert to look at ID points with prob p

$$S_t^{(o)} := \left\{ s_1^{(o)}, \cdots, s_{N_t^{(o)}}^{(o)} \right\}$$ "Score" $s := g(x)$

: set of scores for these ood points that are confirmed by human expert.

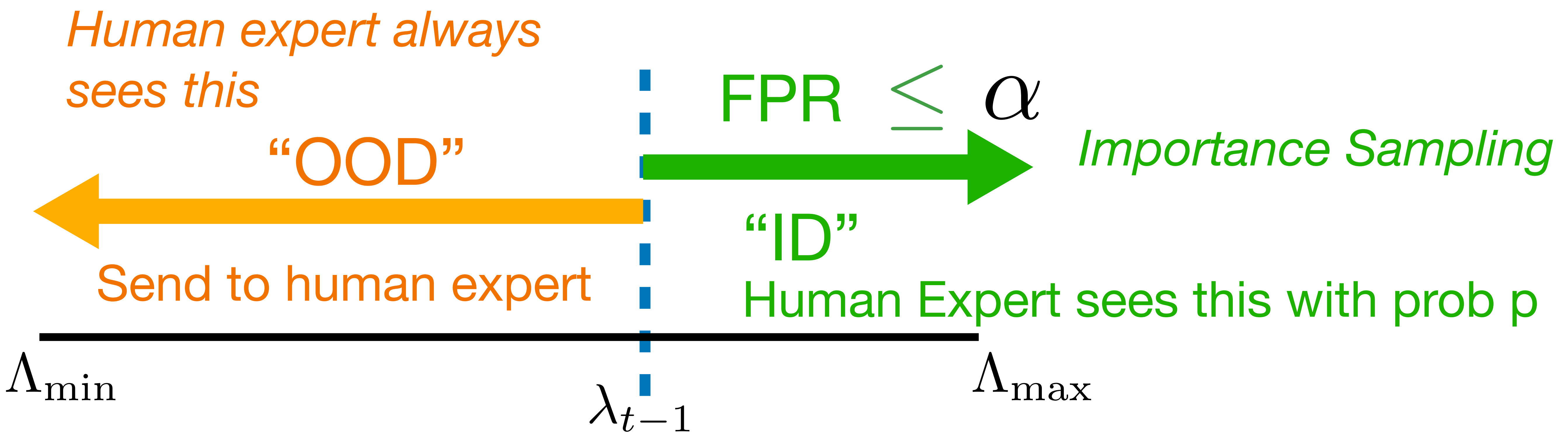
$N_t^{(o)}$: Number of OOD points that are confirmed as OOD from human expert

Valid Time-varying Confidence Intervals

- Law of iterated logarithms (LIL) based bounds for any time valid
- DKW-style bounds for all thresholds — but we do **not** have independent samples

$$\psi(t,\delta) = \sqrt{\frac{3c_t}{N_t^{(o)}}\left[2\log\log\left(\frac{3c_t N_t^{(o)}}{2}\right) + \log\left(\frac{2}{\delta}\frac{|\Lambda_{\max} - \Lambda_{\min}|}{\nu}\right)\right]}$$

$$c_t = 1 - \beta_t + \frac{\beta_t}{p^2} \qquad \beta_t = \frac{N_t^{(o,p)}}{N_t^{(o)}}$$



p : sampling probability when declared "ID"

$N_t^{(o)}$: Number of OOD points that are confirmed as OOD from human expert

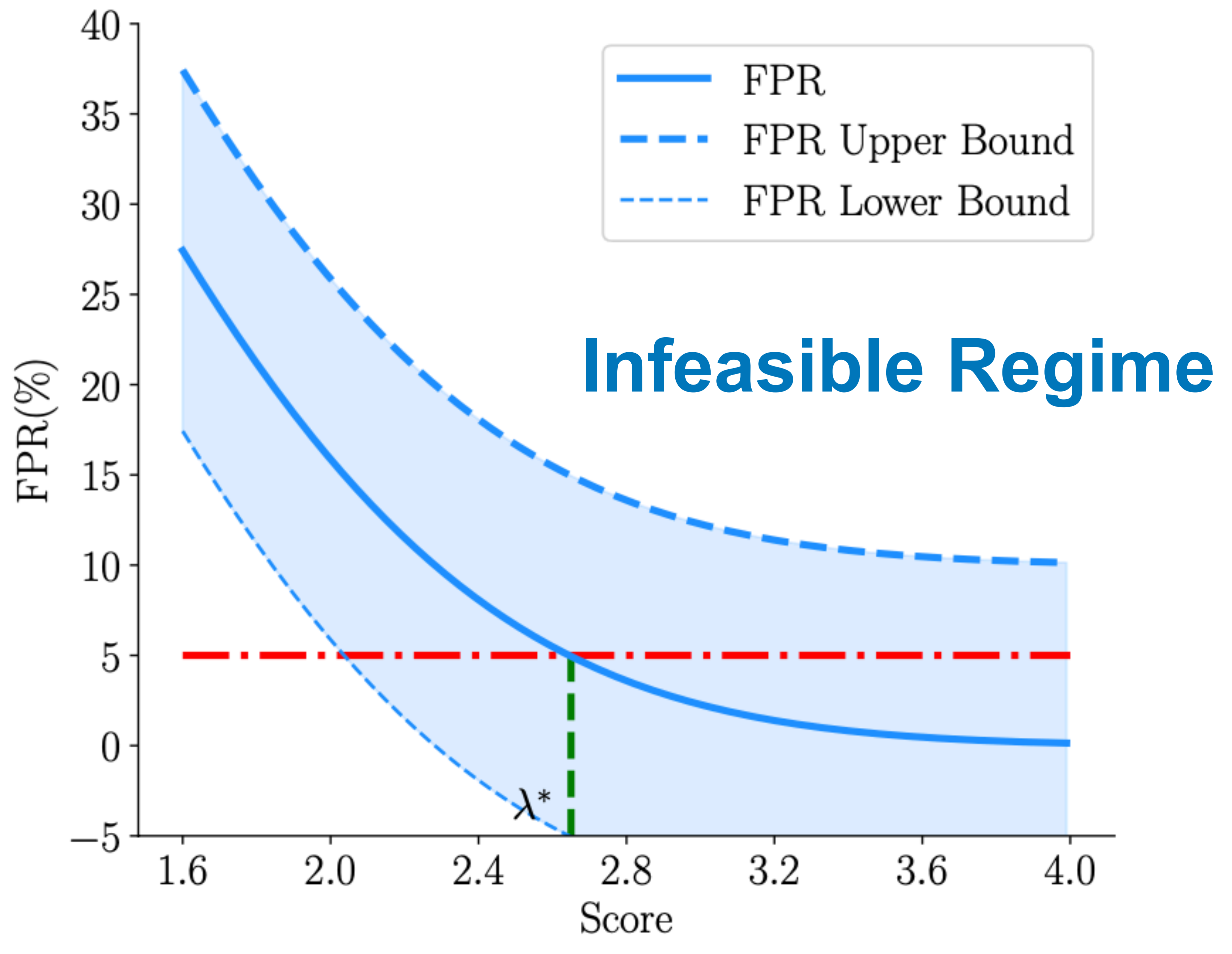
$N_t^{(o,p)}$: Number of points that are importance sampled to get human feedback even when they are declared "ID" by the system

ν : discretization

Khinchine 1924, Jamieson et. al., 2013, Balasubramani 2015, Howard & Ramdas 2022 …..

Illustration of the confidence interval

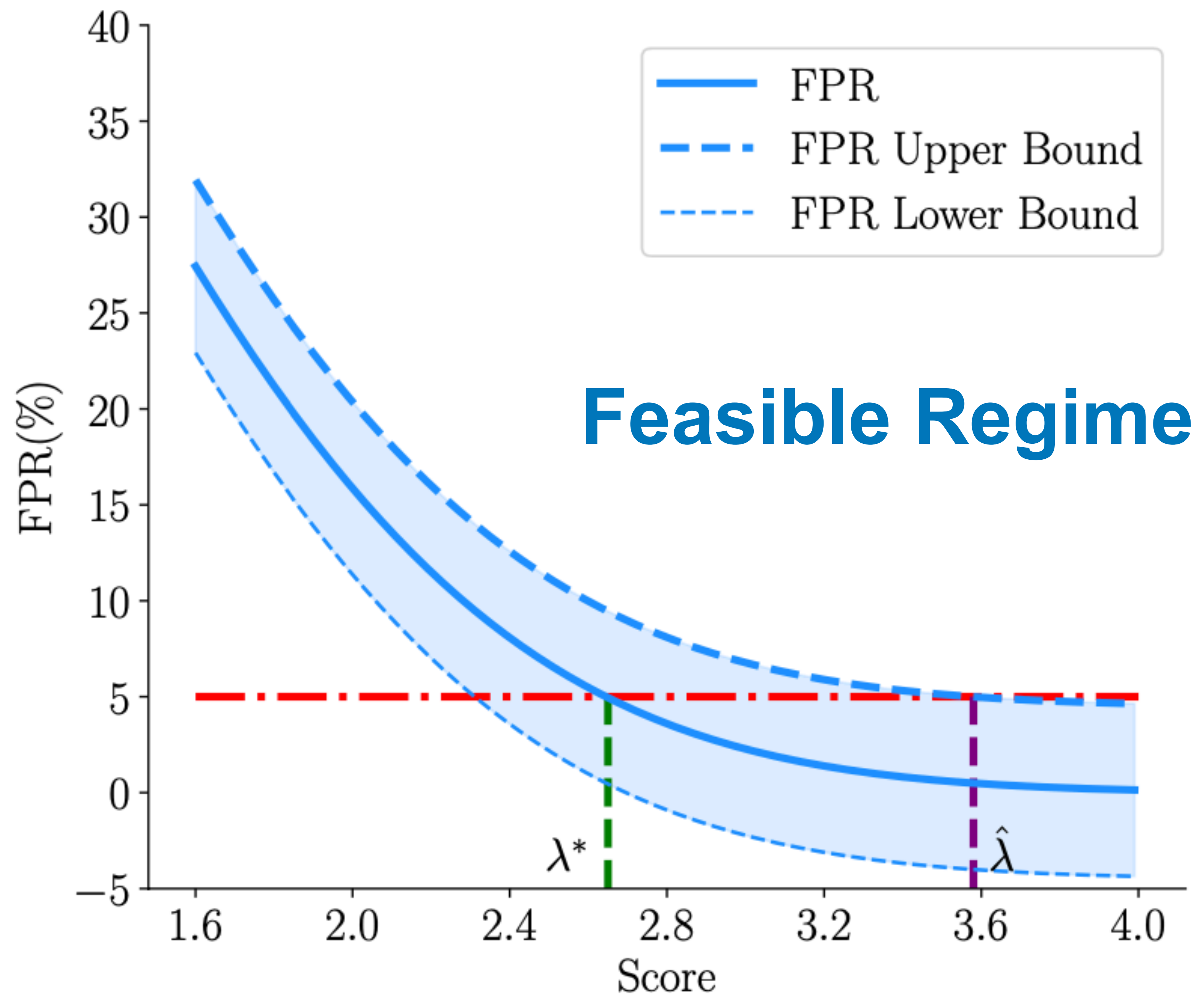
- In the beginning, the threshold is set at $\Lambda_{\max}$
- For first few rounds, the confidence intervals are too wide for a feasible $\lambda_t < \Lambda_{\max}$ to emerge



(a) No feasible solution, in the beginning

Illustration of the confidence interval

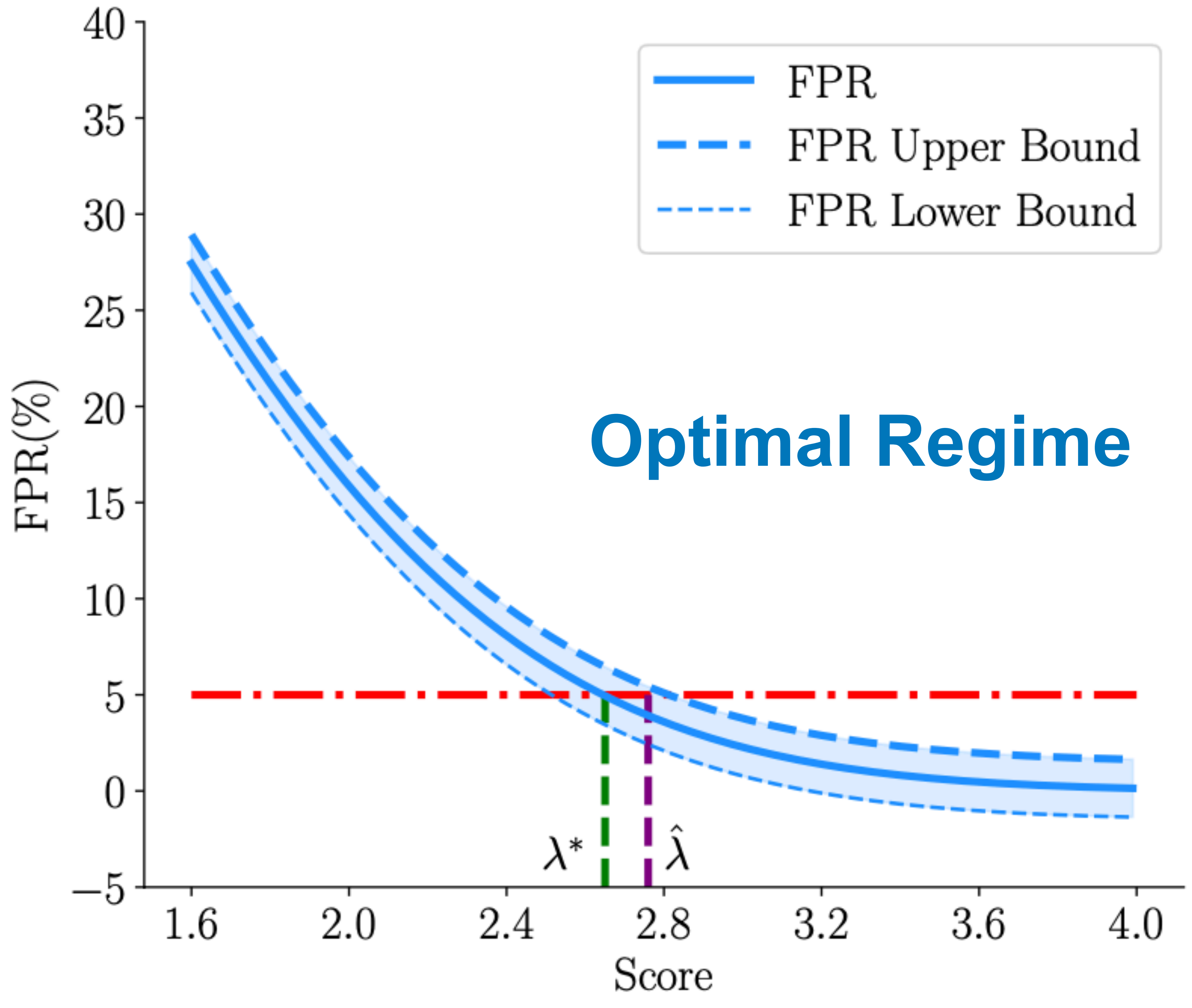
- In the beginning, the threshold is set at $\Lambda_{\max}$
- For first few rounds, the confidence intervals are too wide for a feasible $\lambda_t < \Lambda_{\max}$ to emerge
  - Recall that by construction, $\lambda_t \geq \lambda^*$
- After a while, the confidence intervals get small enough to get a feasible $\lambda_t < \Lambda_{\max}$ to emerge



(b) Feasible solution, after sometime

Illustration of the confidence interval

- In the beginning, the threshold is set at $\Lambda_{\max}$
- For first few rounds, the confidence intervals are too wide for a feasible $\lambda_t < \Lambda_{\max}$ to emerge
  - Recall that by construction, $\lambda_t \geq \lambda^*$
- After a while, the confidence intervals get small enough to get a feasible $\lambda_t < \Lambda_{\max}$ to emerge
- As time progresses, the confidence intervals continue to shrink and the threshold gets closer and closer to the optimal



(c) Near optimal solution, eventually

Theoretical Guarantees

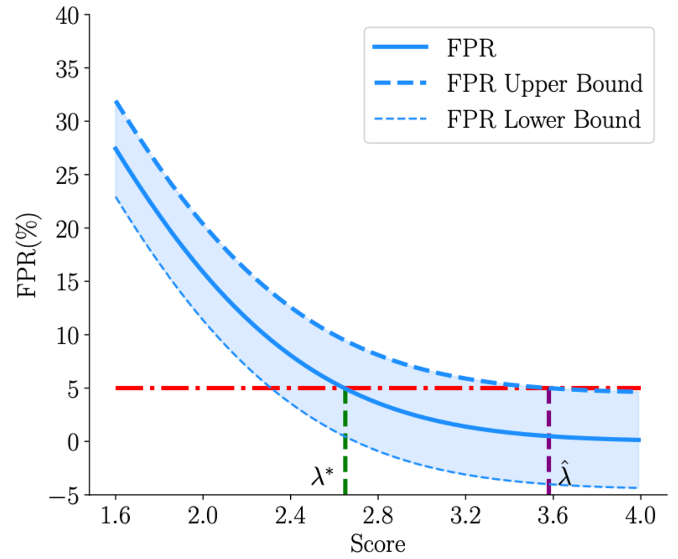
$$\lambda_t := \arg\min_{\lambda} \lambda$$

s.t. $\widehat{\mathrm{FPR}}(\lambda, t) + \psi(t, \delta) \le \alpha$

Under mild conditions, we can provide following guarantees for our procedure with probability $1 - \delta$,

- **FPR is controlled at all times:** for all t, $\mathrm{FPR}(\lambda_t) \le \alpha$
- **Time to reach feasibility:** for all t $\ge T_f := \frac{1}{\gamma\,\alpha^2} \log\left(\frac{1}{\delta}\log\left(\frac{1}{\alpha}\right)\right) + \frac{1}{\gamma^2}\log\left(\frac{1}{\delta}\right)$

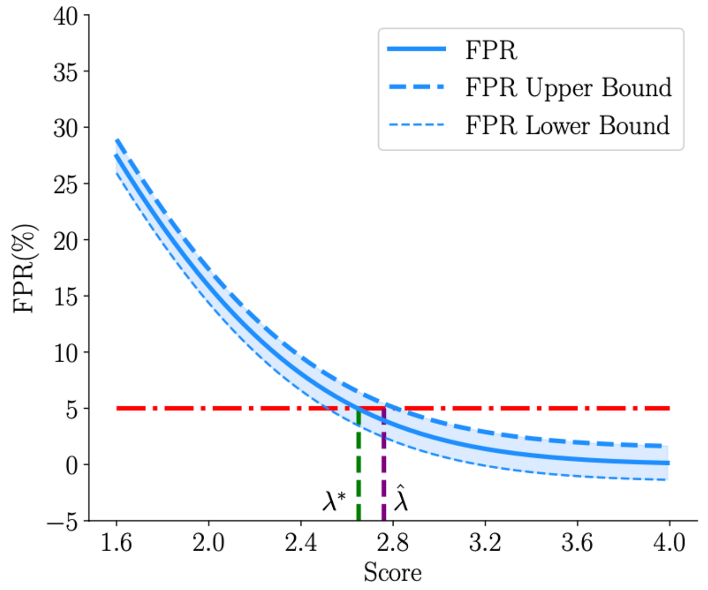
  $\widehat{\mathrm{FPR}}(\lambda_t) + \psi(t, \delta) \le \alpha$ and $\lambda_t < \Lambda_{\max}$



(b) Feasible solution, after sometime

- **Time to reach eta-optimality:** for all t $\ge T_{\eta,\mathrm{opt}} := \frac{1}{\gamma\,\eta^2} \log\left(\frac{1}{\delta}\log\left(\frac{1}{\alpha}\right)\right) + \frac{1}{\gamma^2}\log\left(\frac{1}{\delta}\right)$

  and $\widehat{\mathrm{FPR}}\left(\lambda_{T_{\eta,\mathrm{opt}}}\right) \in \left[\alpha - \frac{\eta}{2},\ \eta\right]$, $\mathrm{FPR}(\lambda^\star) - \mathrm{FPR}(\lambda_t) \le \eta$



(c) Near optimal solution, eventually

Empirical Evaluation

We evaluate our method to verify the following,

Stationary Setting: Distributions do not change.

C1. Compared to **non-adaptive baselines**, our approach achieves lower FPR while maximizing the TPR.

C2. In the stationary setting, **our method satisfies the FPR constraint at all times** and produces high TPR.

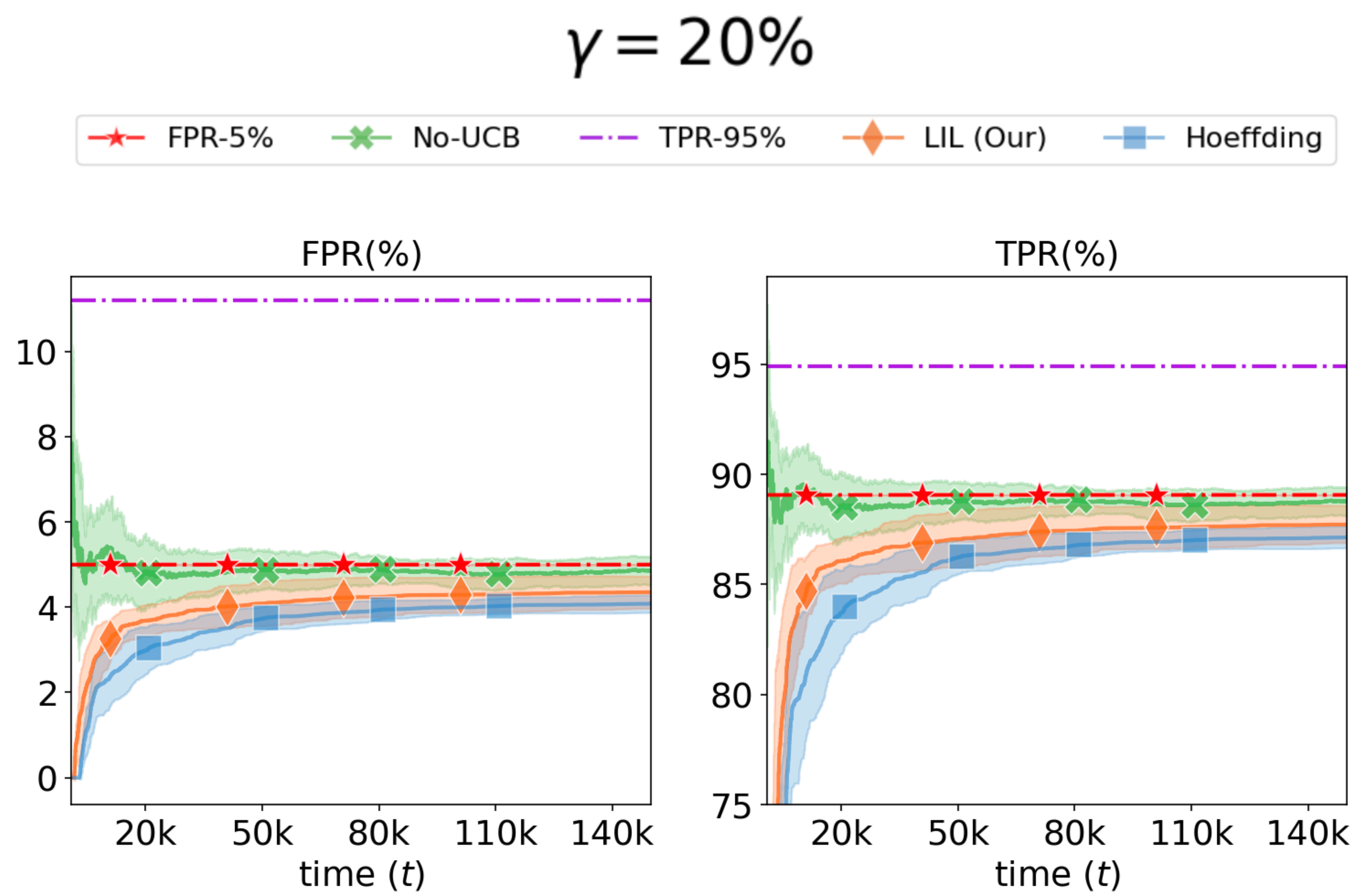
C3. The proposed framework is **compatible with any OOD scoring functions.**

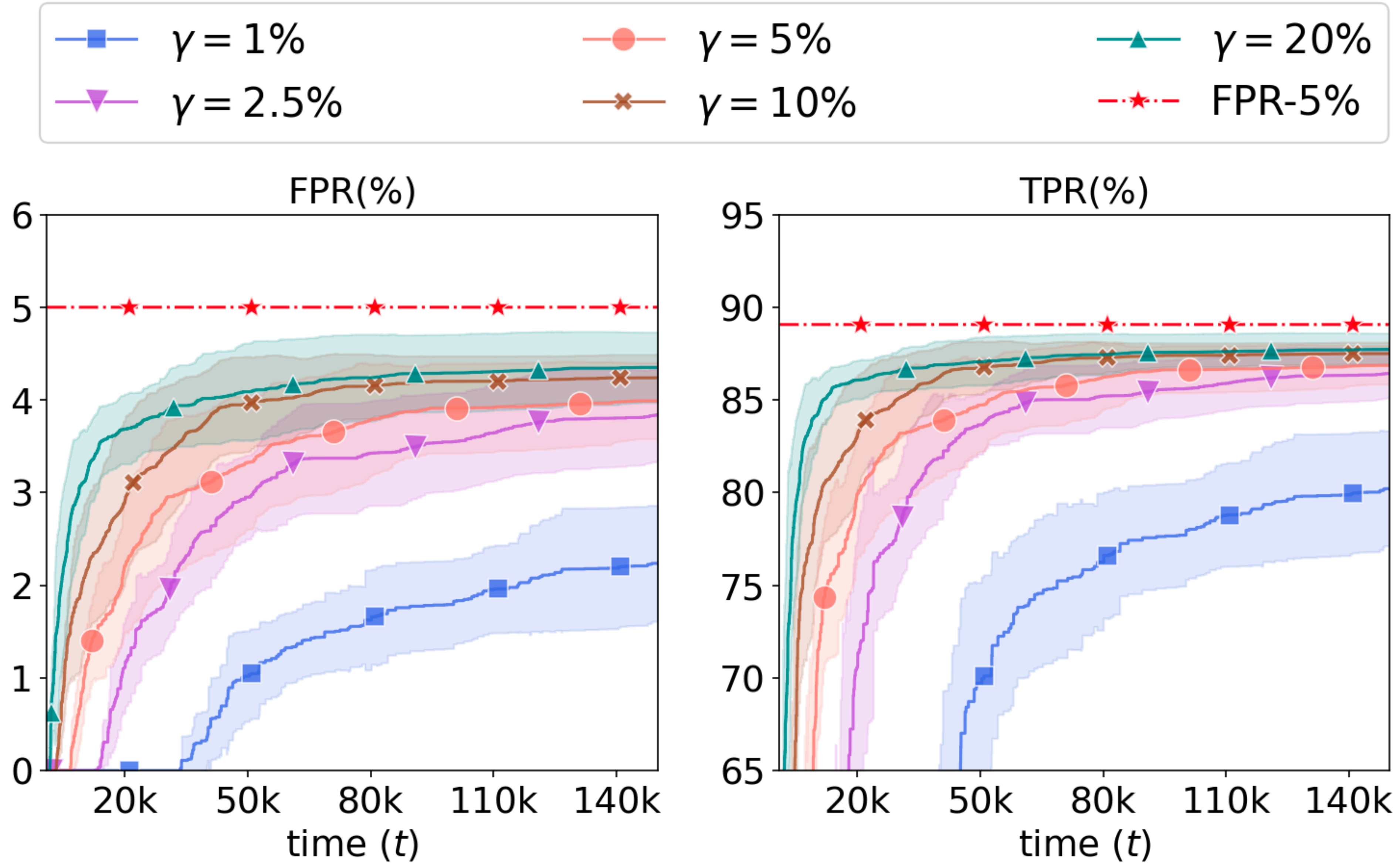
Non-stationary Setting: Distribution(s) shift at some time.

C4. Our method continues to work with a simple adaption using **window based approach**

Simulations : Stationary Setting (C1, C2)

- ID scores: Gaussian $\mu = 5.5, \ \sigma = 4$
- OOD scores: Gaussian $\mu = -6, \ \sigma = 4$





- Fixed threshold (non-adaptive) methods have high FPR.
- Not using UCB leads to FPR violation.
- With LIL, Hoeffding UCB the FPR constraint is maintained and it converges to optimal TPR over time.

- Convergence is faster with higher OOD fraction.
- It maintains FPR below 5% for all values of γ

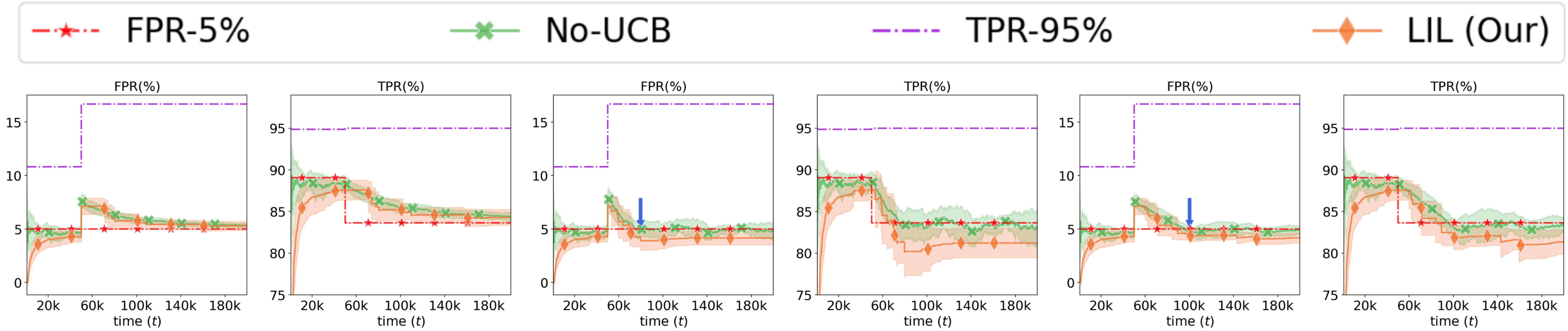
Simulations: Non-stationary Setting (C4)

- ID scores: Gaussian $\mu =$ 5.5, $\sigma = 4$

  $\gamma = 20\%$

- OOD scores: Gaussian $\mu =$ -6, $\sigma = 4$ (till t=50k)
- OOD scores: Gaussian $\mu =$ -5, $\sigma = 4$ (after t=50k)

Only use most recent N_w (window size) samples to compute FPR and confidence intervals.



(a) Distribution shift, no window. (b) Distribution shift, 5k window. (c) Distribution shift, 10k window.

- Our method violates the FPR constraint for a short time and then comes back.
- Non-adaptive methods keep using the initial threshold and incur higher FPR.
- Method without UCB does adapt but takes longer time and has higher variance due to window size.

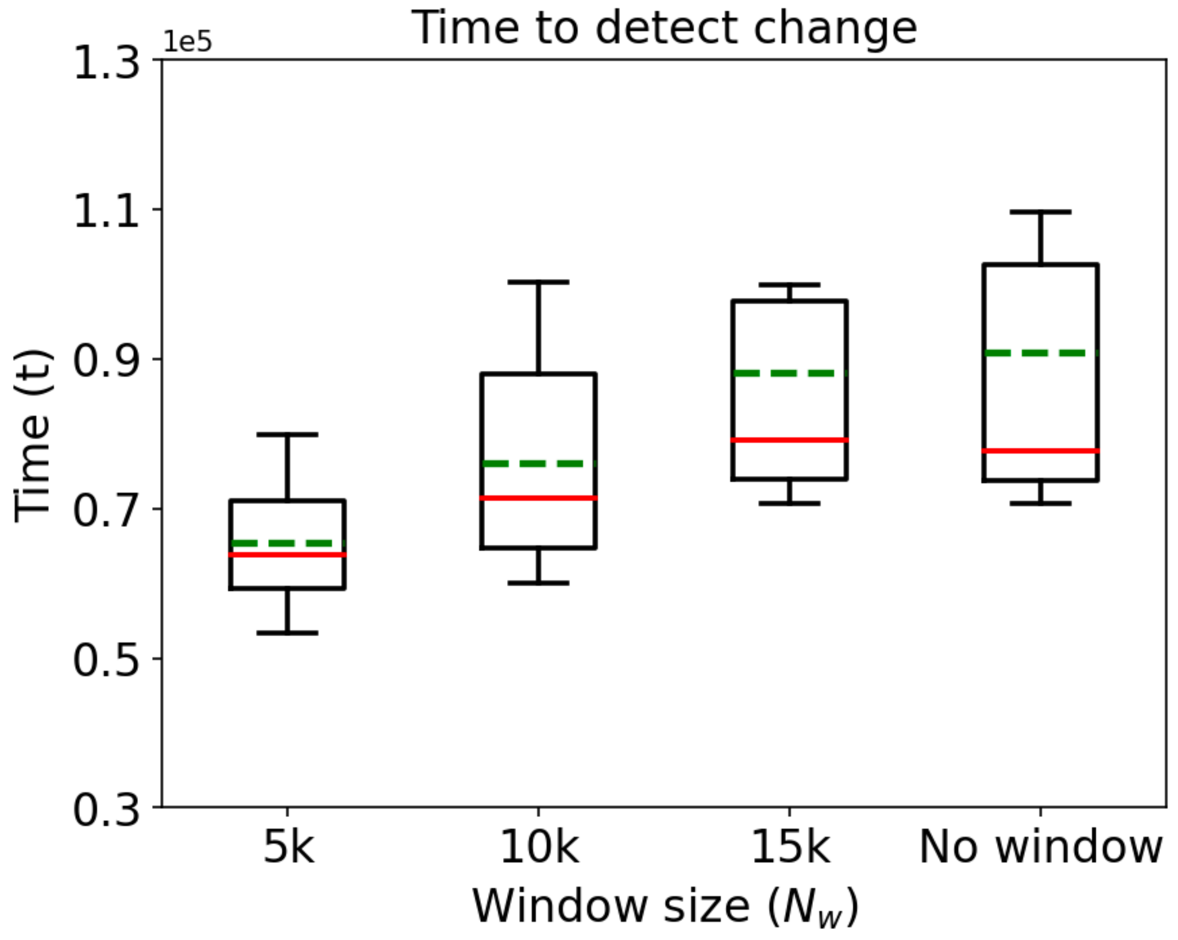
Simulations: Window Size Trade-off

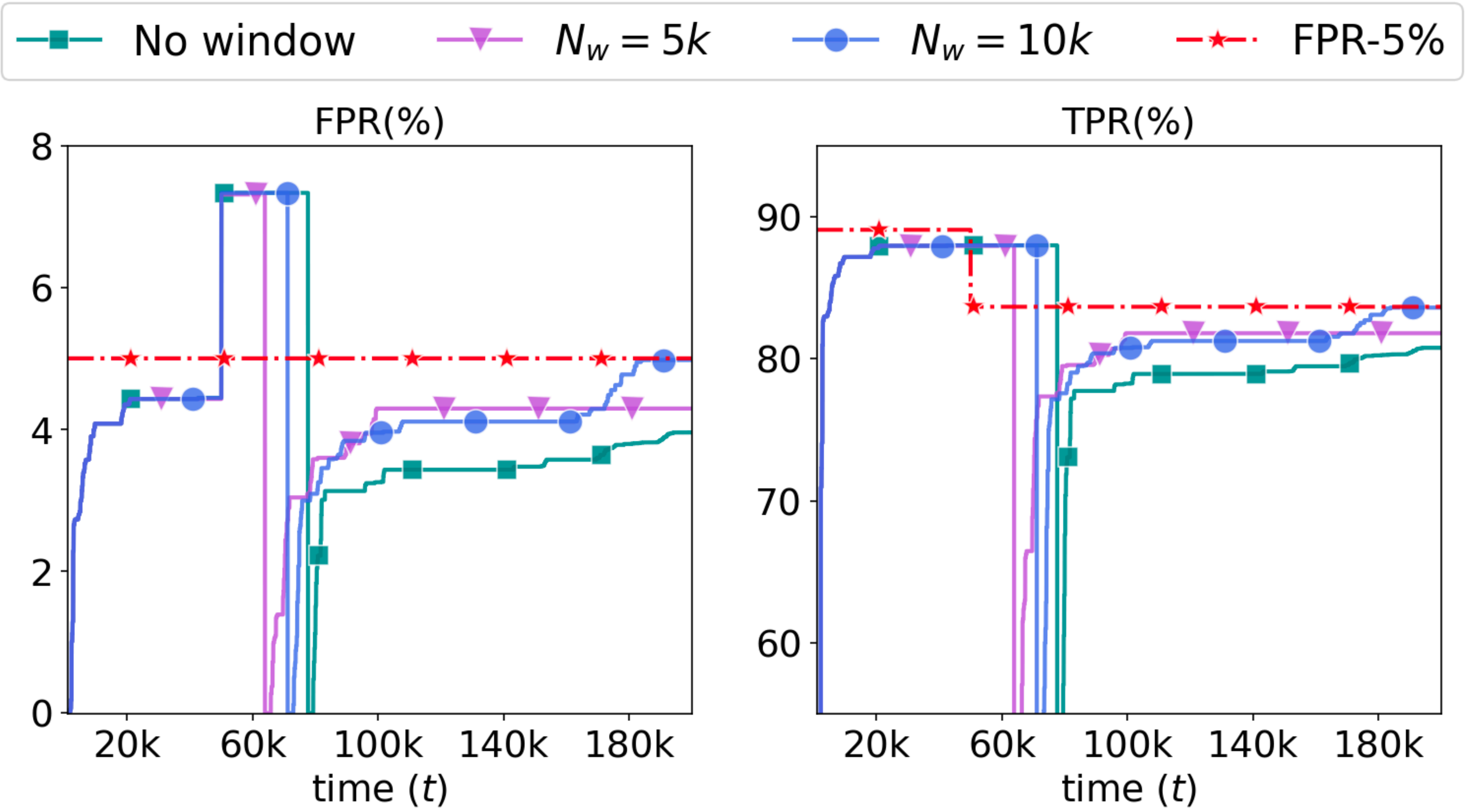
- ID scores: Gaussian $\mu =$ 5.5, $\sigma = 4$

  $\gamma = 20\%$

- OOD scores: Gaussian $\mu =$ -6, $\sigma = 4$ (till t=50k)
- OOD scores: Gaussian $\mu =$ -5, $\sigma = 4$ (after t=50k)

Only use most recent N_w (window size) samples to compute FPR and confidence intervals.





- Shorter window leads to faster change detection but limits optimality.
- With longer window we can reach closer to optimal threshold but it will take long time.

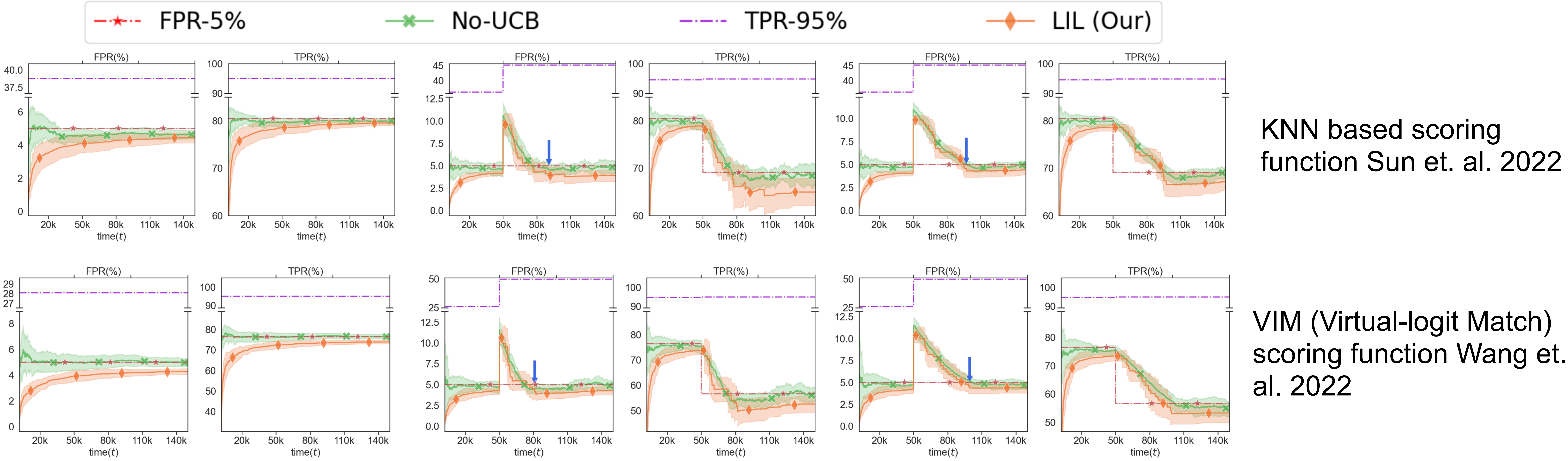
- Conservative approach: restart after detecting change.

Can work with any scoring functions (C3)

- ID: CIFAR-10

$\gamma = 20\%$

- OOD1 : MNIST, SVHN, and Texture (till t=50k)
- OOD2 : TinyImageNet, Places365, CIFAR-100 (after t=50k)



(a) No distribution shift, no window. (b) Distribution shift, 5k window. (c) Distribution shift, 10k window.

- Methods work as expected from simulations.
- The best TPR achievable depends on scoring function and our method approaches it while maintaining FPR guarantee at all times.

Summary

- Framework for human-in-the-loop OOD detection with false positive rate control
- This framework can work with any scoring function
- Guarantees for FPR control for all time when OOD is not shifting
- Windowed approach when OOD is shifting

Thank you!

Questions