UNIVERSITY OF CAMBRIDGE

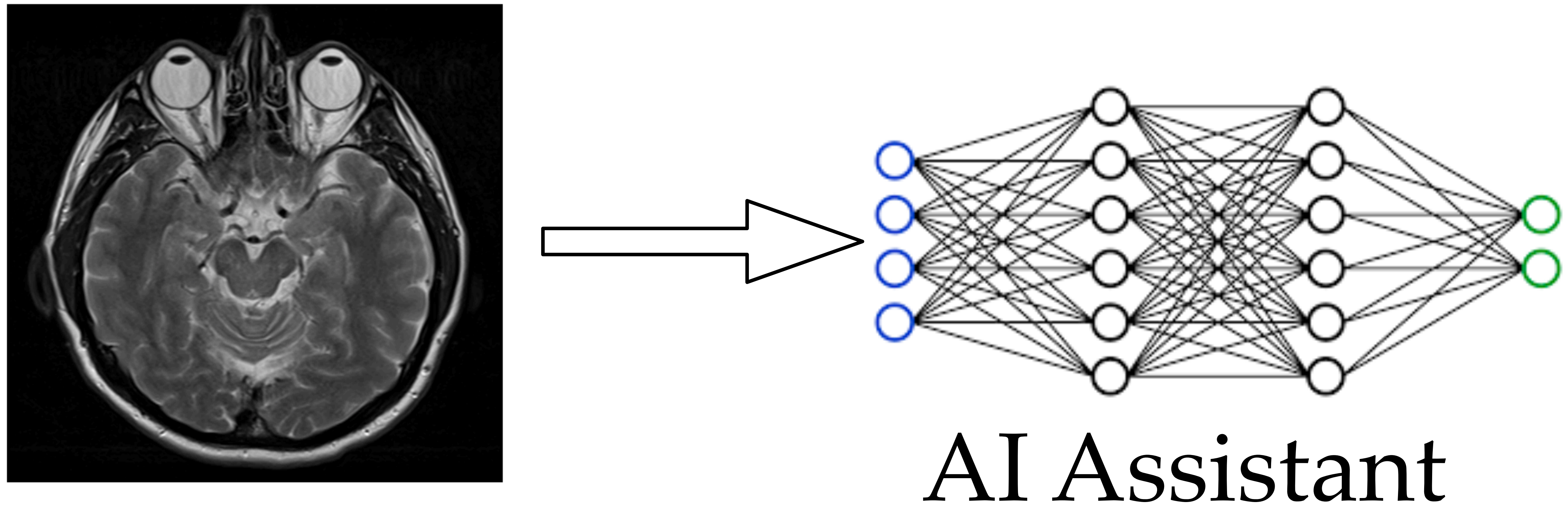
Set Valued Predictions for Human-AI Teams

Varun Babbar

Supervisors: Dr Adrian Weller and Umang Bhatt

*A part of this work was accepted at IJCAI'22 as a conference paper

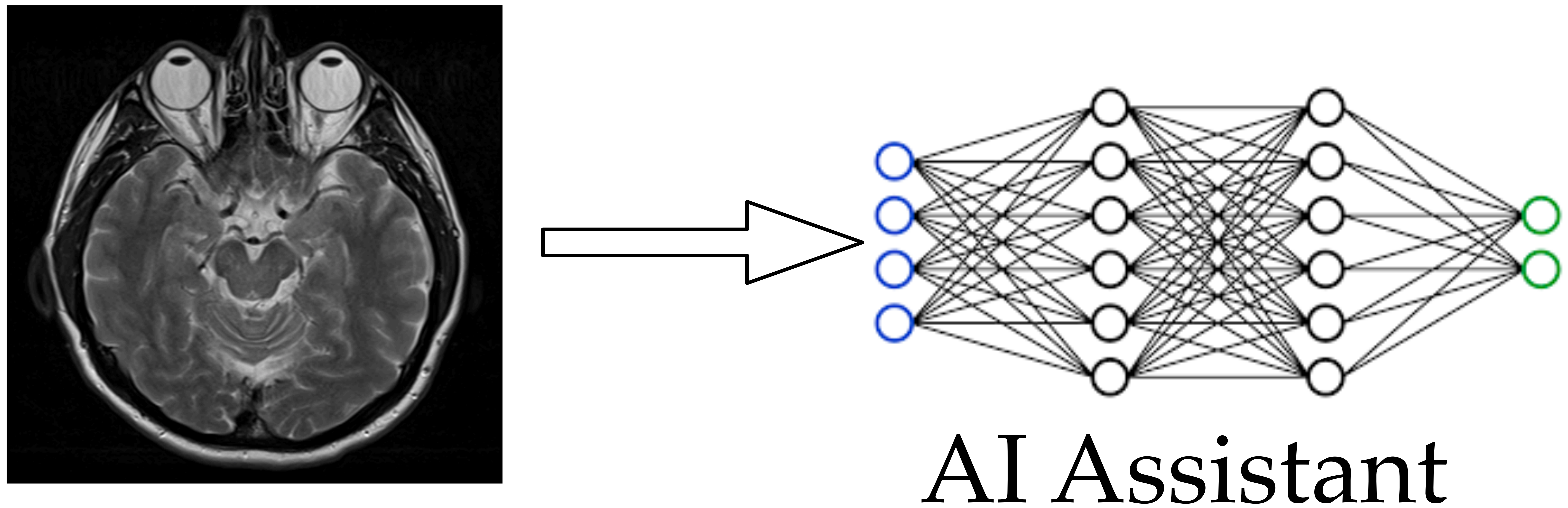
What is a predictive set?



AI Assistant

$\underbrace{\text{Concussion}}_{\text{Most Probable Label}}$

Top-1 Classifier



AI Assistant

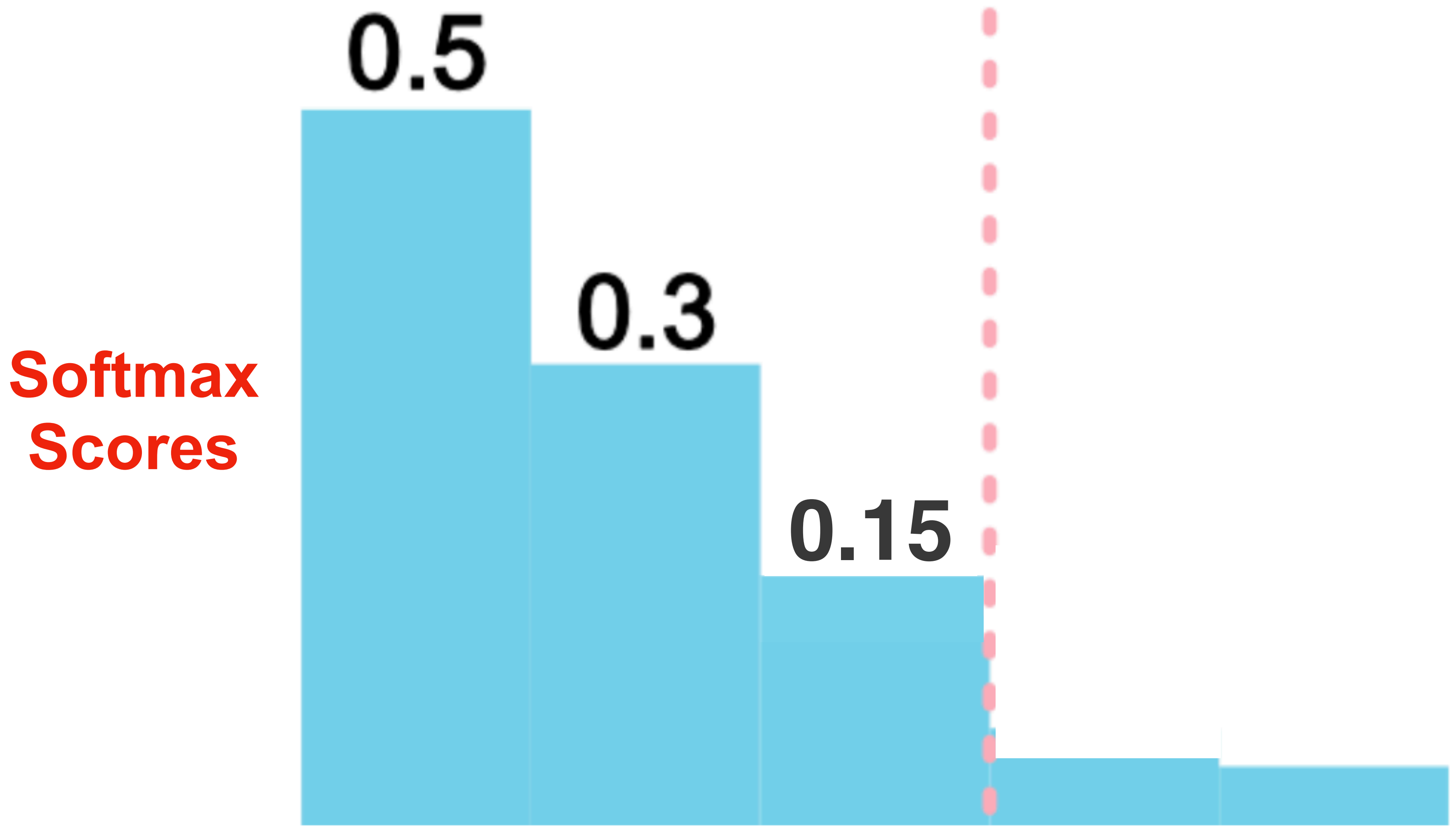
$\underbrace{\{\text{Concussion, Tumour}\}}_{95\,\%\ \text{Confidence Set}}$

Set Valued Classifier

How do we generate calibrated set valued predictions?

Naive Set Construction Procedure

- Sum all softmax probabilities till we reach the given threshold



An illustration of the naive method. Since the softmax scores are not the true probabilities, the pink threshold does not provide coverage.

Is this really a 95 % Confidence Set?

- Model probabilities are not calibrated
- For 'hard' examples, set sizes will be very large

Generating Predictive Sets

- We want a predictive set that **controls** for some user defined **risk** function with **high probability**

- Distribution-Free $\Rightarrow$ $\{$Input Data Distribution, Model$\}$ = Black Box $\blacksquare$

- <u>All we have to do is learn a threshold τ!</u>

- **Predictive Set:** $\Gamma(X) = \{y : \hat{p}(y \mid x) \geq \tau\}$

Conformal Prediction (CP)[1]

$$\text{FNR} \leq \alpha \equiv P\big(Y \notin \Gamma(X)\big) \leq \alpha$$

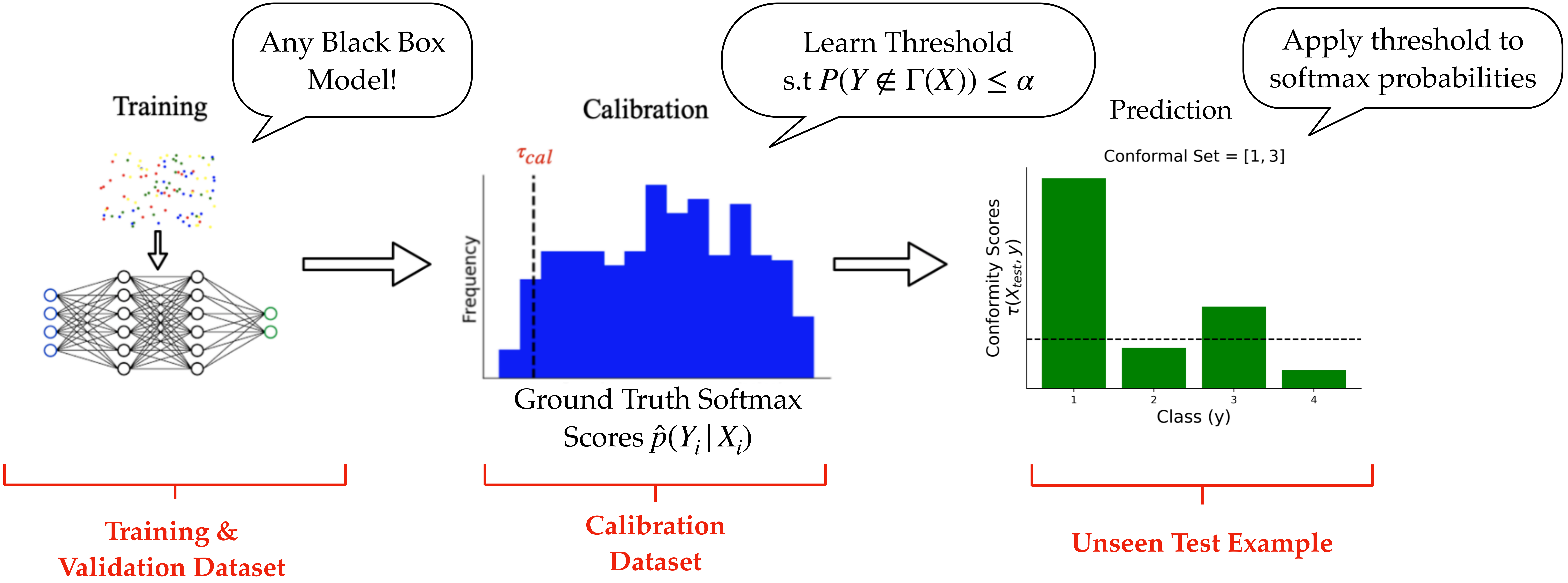
Risk Controlling Prediction Sets (RCPS)[2]

$$P\big(\underbrace{\mathbb{E}[L(Y, \Gamma(X))]}_{\text{Risk}} \leq \alpha\big) \geq 1 - \delta$$

[1] Shafer, G., & Vovk, V. (2008). A tutorial on conformal prediction. *ArXiv, abs/0706.3188*.

[2] Bates, S., Angelopoulos, A., Lei, L., Malik, J., & Jordan, M.I. (2021). Distribution-Free, Risk-Controlling Prediction Sets. *J. ACM, 68*, 43:1-43:34.

Learning the Threshold for Conformal Prediction (CP)



Risk Controlling Prediction Sets (RCPS)

Predictive Set: $\Gamma_\tau(X) = \{y : \hat{p}(y \mid x) \geq \tau\}$

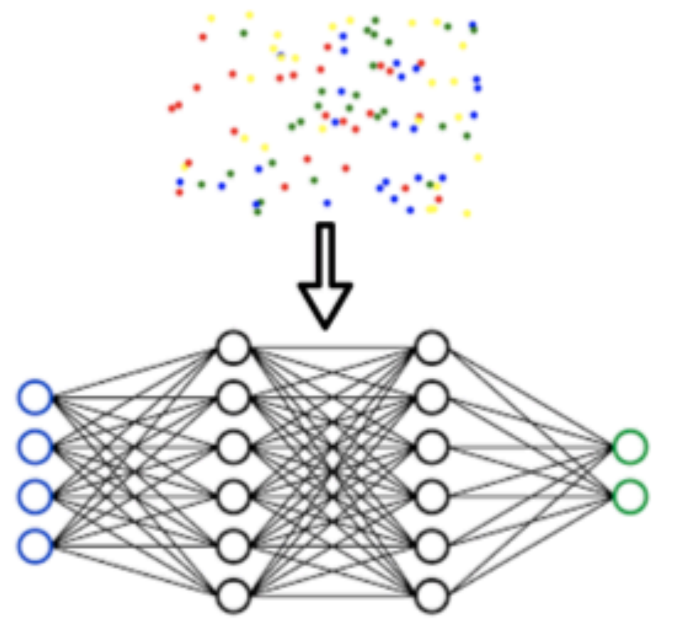
Any Black Box Model!

Compute Risk $R(\tau)$ and $1 - \delta$ UCB $R^+(\tau)$

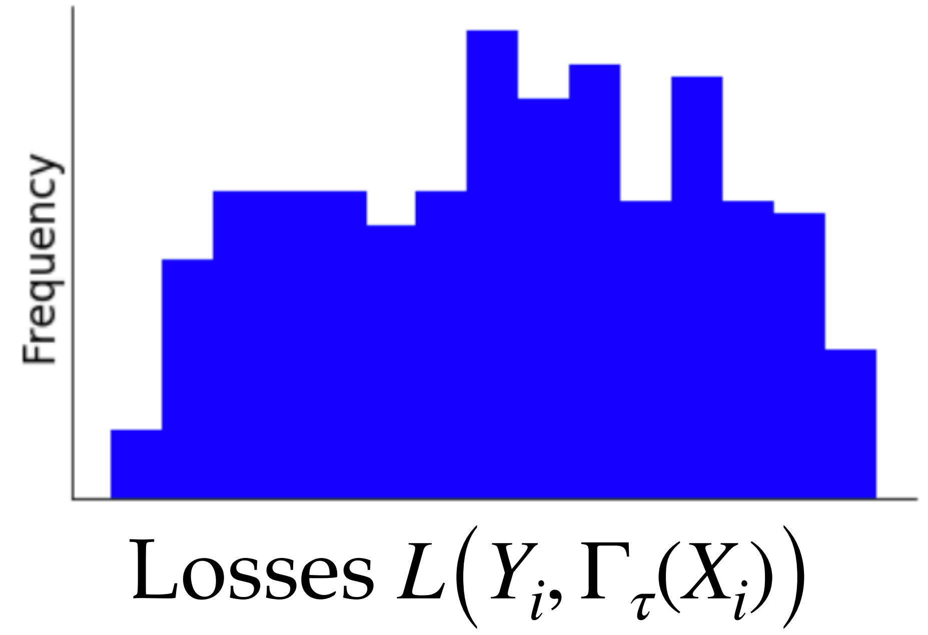
Learn Threshold s.t. $R^+(\tau) \leq \alpha$

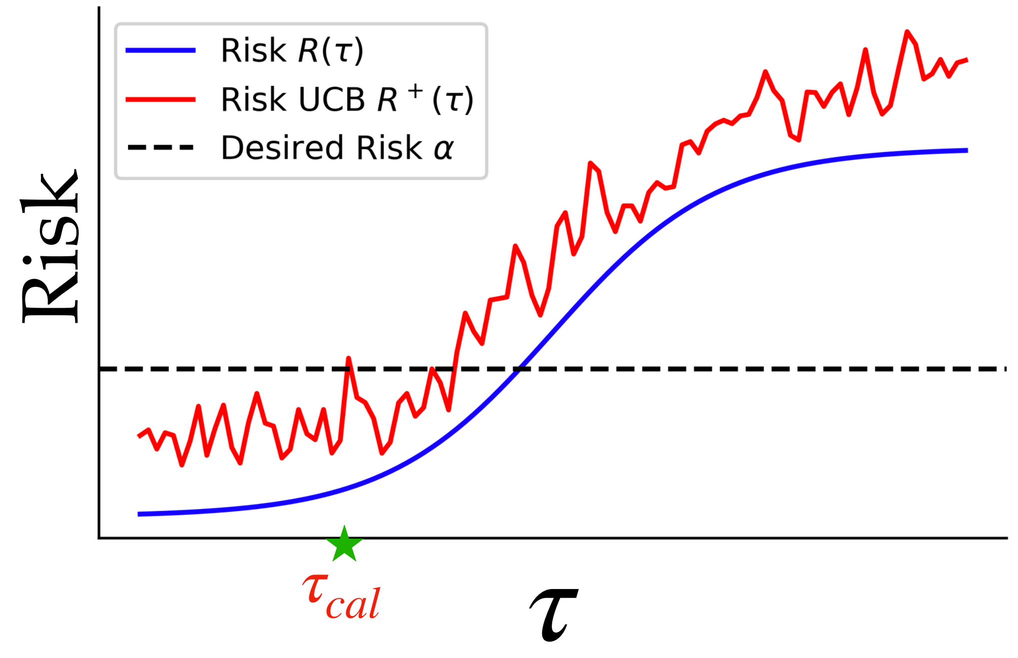
Apply threshold to softmax probabilities

Training

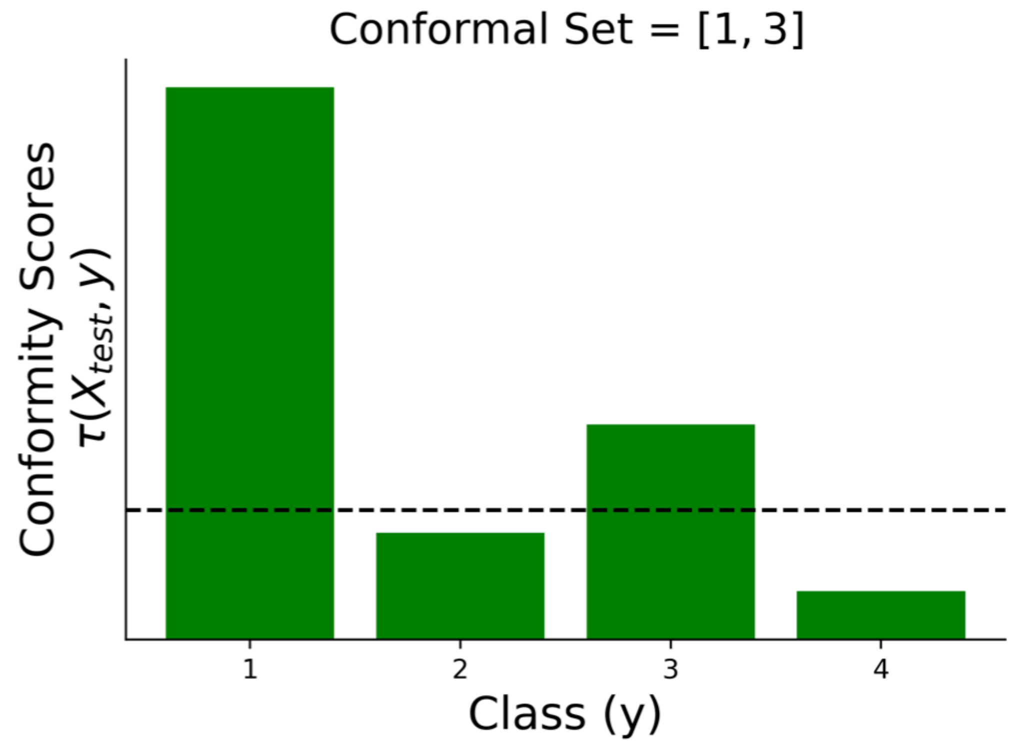


Calibration





Prediction



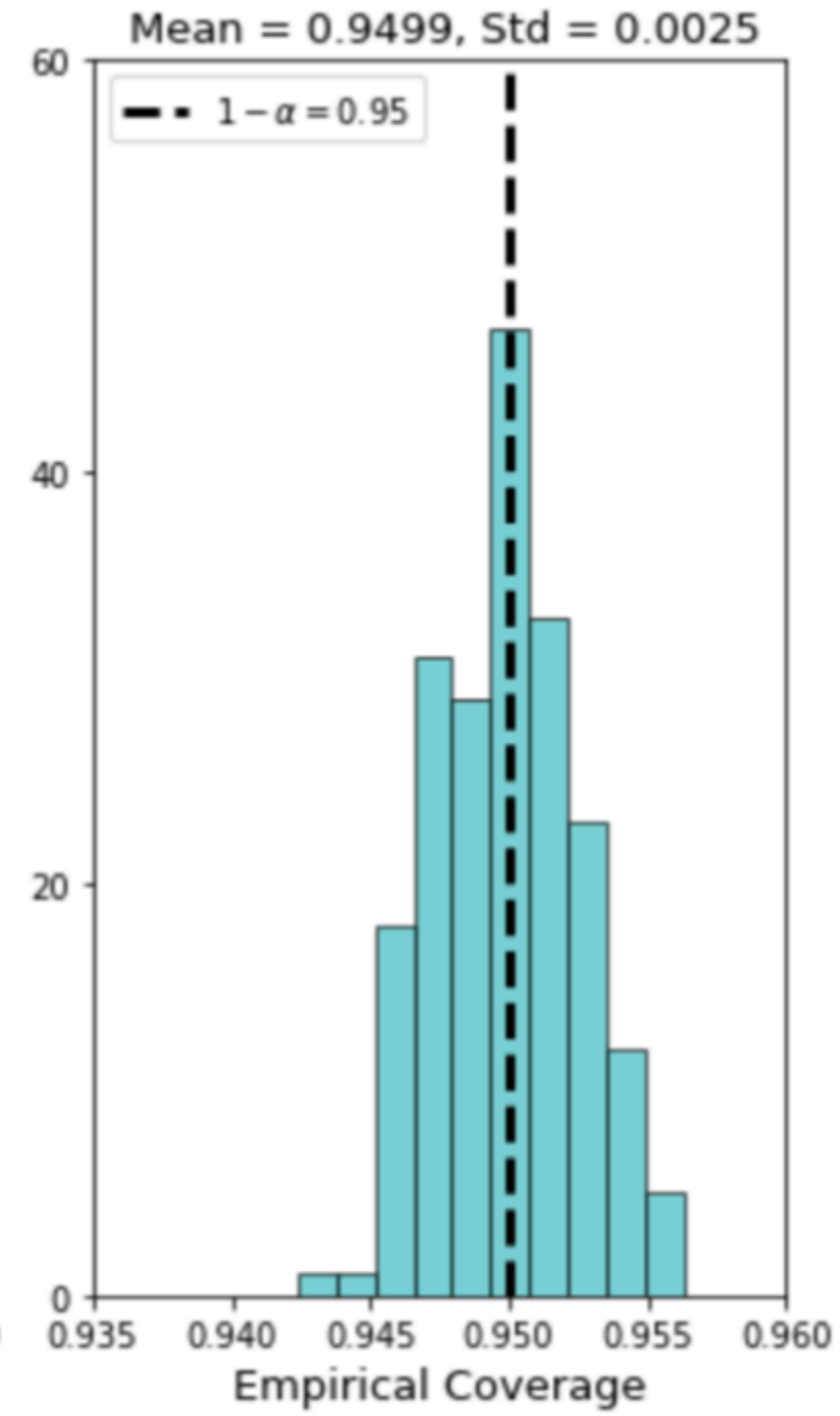
Training & Validation Dataset

Calibration Dataset

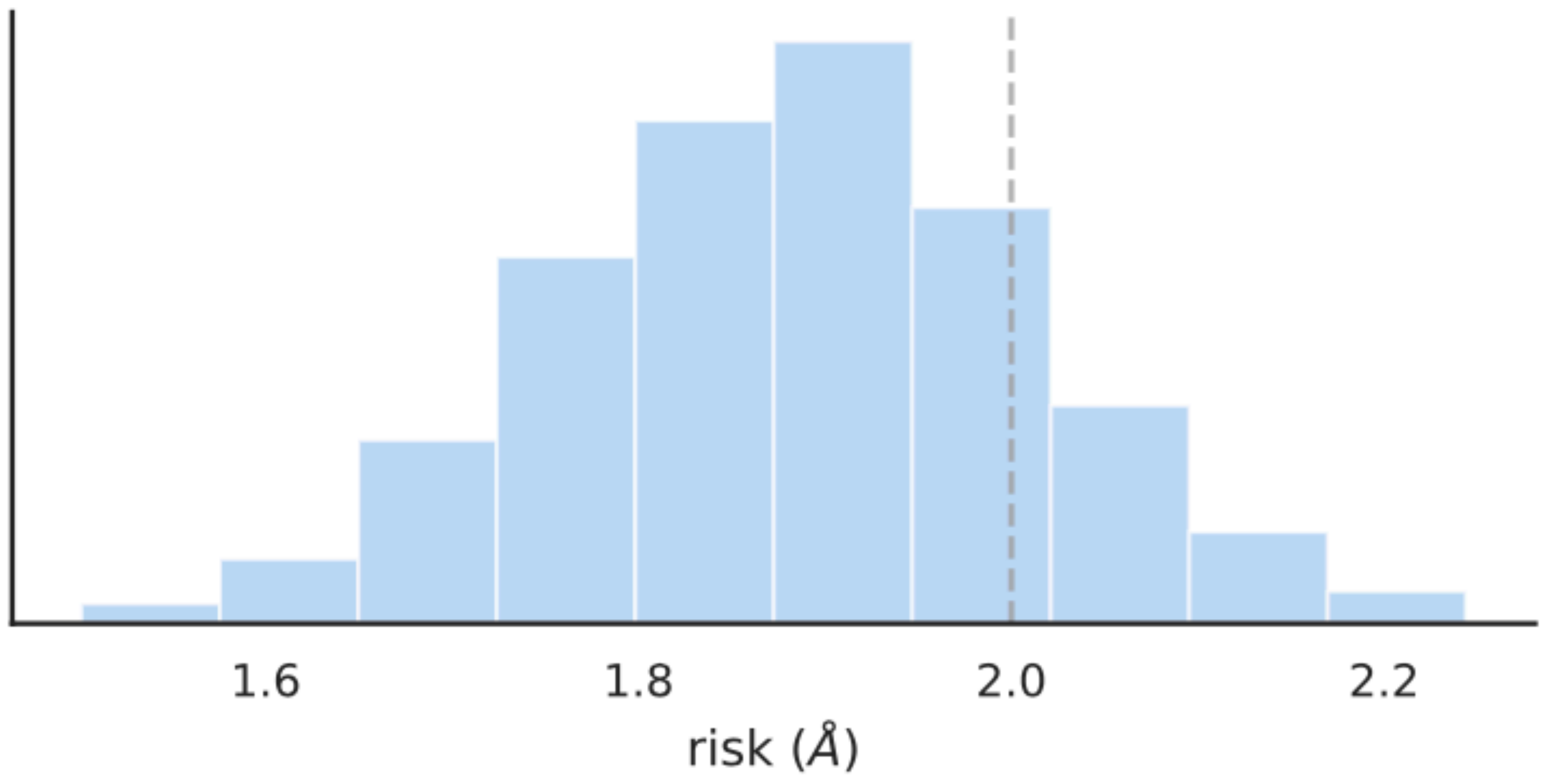
Unseen Test Example

UNIVERSITY OF CAMBRIDGE

CP Set Dist



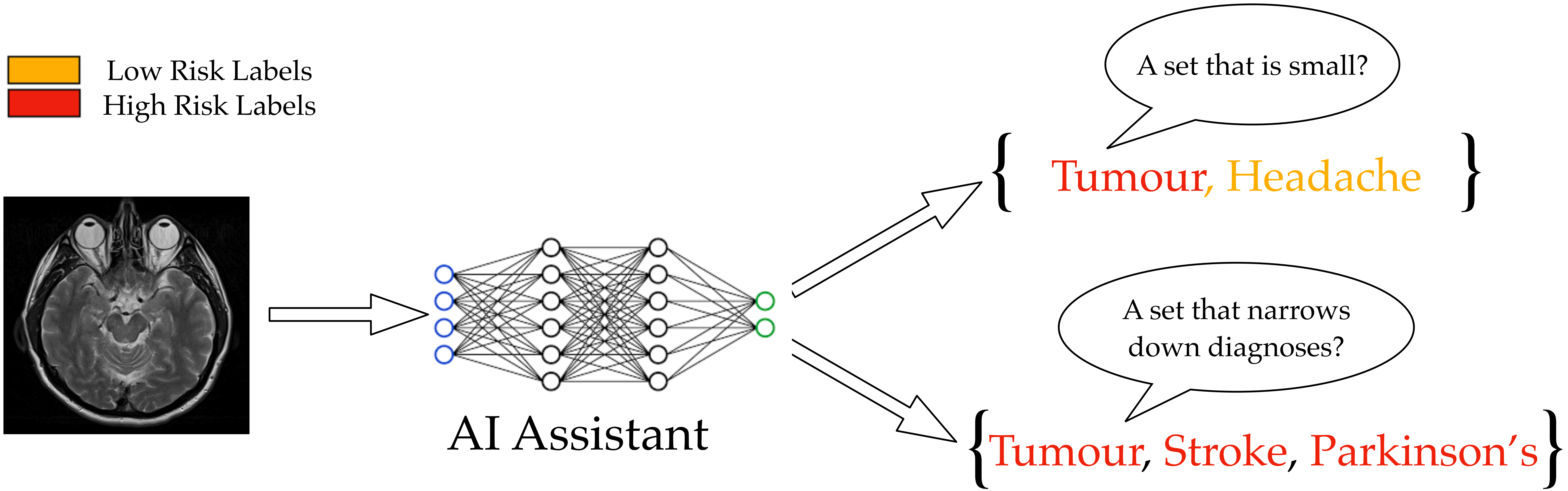
Coverage distribution of CP over 1000 calibration-test splits



Risk Distribution of RCPS over 1000 calibration test splits. With $\delta = 0.1$, there is a 10% chance of violating risk

But what kind of predictive sets should we provide human experts?

What kind of predictive sets should we provide human experts?



Let’s ask a more fundamental question….

Are prediction sets useful in Human-AI teams in the first place?

How Useful are Prediction Sets in Human-AI Teams?

- Are prediction sets better than point predictions?

How Useful are Prediction Sets in Human-AI Teams?

- Are prediction sets better than point predictions? Yes!

A CP Scheme! [3]

- CP sets are perceived to be more useful by humans ✓
- Humans trust CP predictors more than Top-1 classifiers ✓

Metric	Top-1	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.05	0.76 ± 0.05	0.999	0.000
Reported Utility	5.43 ± 0.69	6.94 ± 0.69	**0.003**	1.160
Reported Confidence	7.21 ± 0.55	7.88 ± 0.29	0.082	0.674
Reported Trust in Model	5.87 ± 0.81	8.00 ± 0.69	$< \mathbf{0.001}$	1.487

Table 1: Top-1 vs RAPS ($\alpha = 0.1$)

[3] Angelopoulos, A., Bates, S., Malik, J., & Jordan, M.I. (2021). Uncertainty Sets for Image Classifiers using Conformal Prediction. *ArXiv, abs/2009.14193*.

But we can’t just provide any predictive set!

How Useful are Prediction Sets in Human-AI Teams?

- Can we narrow down properties of set predictions that provide value to human-AI teams?

Yes! (To some extent)

Metric	Top-1 + Random	RAPS	p value	Effect Size
Accuracy	0.72 ± 0.05	0.76 ± 0.05	0.427	0.338
Reported Utility	5.01 ± 0.65	6.94 ± 0.69	**0.003**	1.432
Reported Confidence	7.29 ± 0.47	7.88 ± 0.29	0.082	0.098
Reported Trust in Model	5.73 ± 1.07	8.00 ± 0.69	**0.008**	1.316

Table 2: Top-1 + Random vs RAPS ($\alpha = 0.1$)

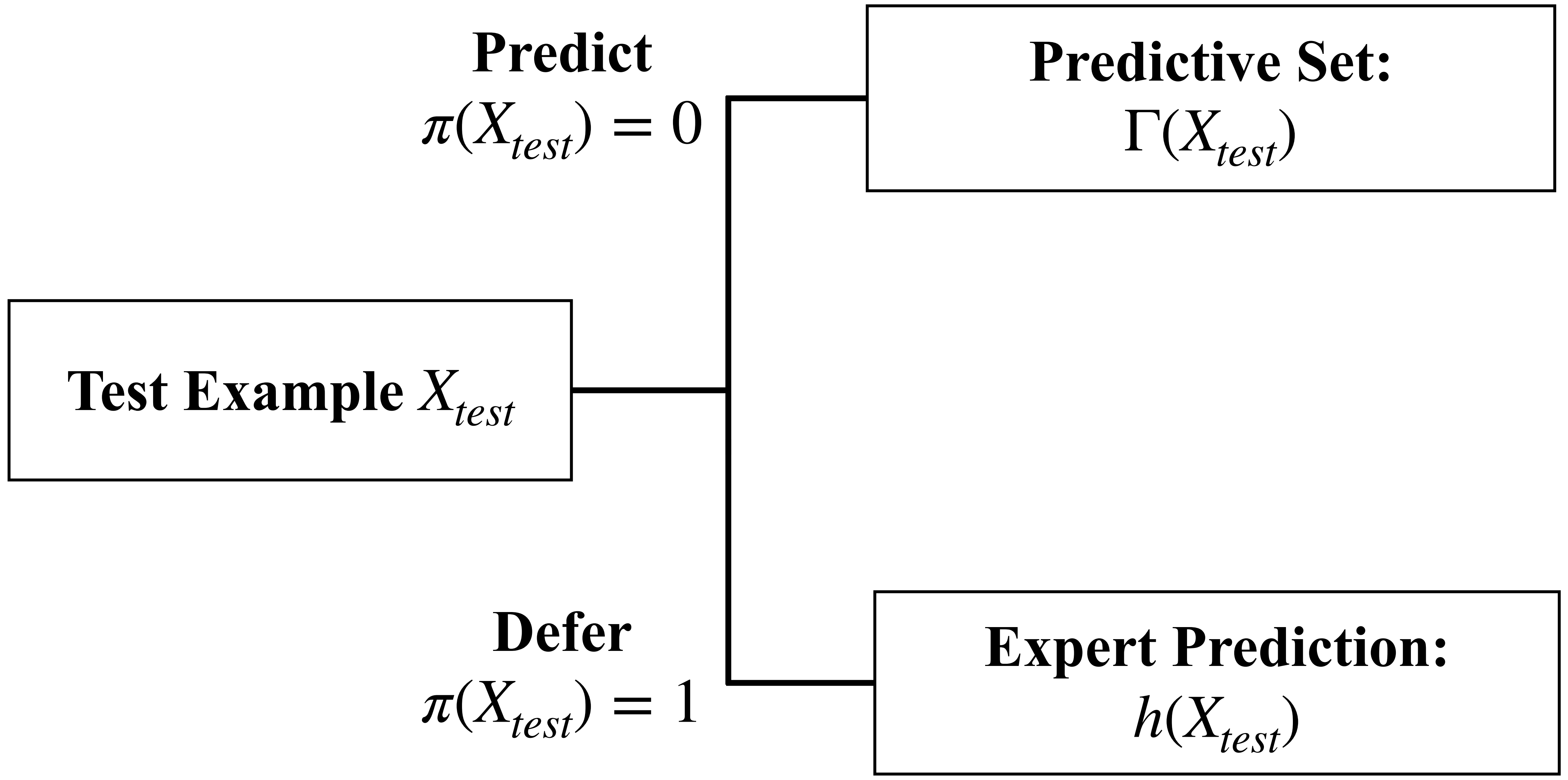
$\Rightarrow$ Prediction sets must accurately reflect model uncertainty

This is a good start……

but let’s improve upon this baseline!

Combining Learning to Defer and Set-Valued Predictions : D-CP

- We need not provide a predictive set for every example!
- Why not leverage the best of the human and the model's abilities? (and provably so!)
- We need to learn a *deferral policy* $\pi(X) \in \{0,1\}$ alongside the classifier!
- We call this scheme D-CP



Empirical Results on 3 CP Schemes

3 Different CP Schemes

Deferral Rate	Team Accuracy	Predictive Set Size of Non-Deferred Examples		
		RAPS	APS	LAC
0	65.18 $\pm$ 0.30	3.75 $\pm$ 0.06	4.61 $\pm$ 0.08	3.26 $\pm$ 0.03
0.1	69.95 $\pm$ 0.31	2.81 $\pm$ 0.05	4.05 $\pm$ 0.06	2.13 $\pm$ 0.04
0.2	72.98 $\pm$ 0.30	2.36 $\pm$ 0.06	2.93 $\pm$ 0.10	2.07 $\pm$ 0.03

Table 3: CIFAR-100: Synthetic Human Expert with 70 % accuracy ($\alpha = 0.1$)

Deferral Rate	Team Accuracy	Predictive Set Size of Non-Deferred Examples		
		RAPS	APS	LAC
0	82.02$\pm$ 0.55	1.91 $\pm$ 0.03	2.83 $\pm$ 0.05	2.47 $\pm$ 0.03
0.1	86.53 $\pm$ 0.68	1.73 $\pm$ 0.08	2.56 $\pm$ 0.07	1.90 $\pm$ 0.04
0.2	89.43 $\pm$ 0.64	1.49 $\pm$ 0.06	2.13 $\pm$ 0.13	1.51 $\pm$ 0.03

Table 4: CIFAR-10H[5]: Real human annotations with 95% accuracy ($\alpha = 0.1$)

- We get lower set sizes on non-deferred examples
- Higher overall team accuracy!
- Win-Win!

[5] Peterson, Joshua C. et al. "Human Uncertainty Makes Classification More Robust." *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019): 9616-9625.

Empirical Results on CIFAR-10H

Human Subject Evaluation of D-CP

Human Subjects benefit from:

- Higher Perceived Utility ✓
- Higher Trust in Model ✓
- Higher Accuracy ✓

Compared to showing CP sets!

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.76 ± 0.08	0.67 ± 0.05	**0.003**	0.832
Reported Utility	7.93 ± 0.39	6.32 ± 0.60	$<$ **0.001**	1.138
Reported Confidence	7.31 ± 0.29	7.28 ± 0.29	0.862	0.046
Reported Trust in Model	8.00 ± 0.45	6.87 ± 0.61	**0.006**	0.754

Table 5: D-RAPS vs RAPS: All Examples
$\alpha = 0.1$, deferral rate $b = 0.2$, CIFAR-100

Human Subject Evaluation of D-CP

$$\textbf{Bias} = \frac{\#\text{ times human is incorrect and their prediction is in the CP set}}{\text{Total Number of Examples}}$$

Lower bias $\Rightarrow$ Human experts are not as influenced by incorrect labels found in the predictive set!

Metric	**D-RAPS**	**RAPS Non-Deferred Examples**
Bias	0.063 ± 0.035	0.189 ± 0.046

Table 6: Human Subject Bias on Non-Deferred Examples CIFAR-100

Why stop at the model? We can also control expert risk!

Dual Risk Control Properties of D-CP

- By combining deferral and set prediction, we can also jointly control for the false negative rate of the **model** and the **expert!** (an extension of [4])

- Define set predictor as:

$$\Gamma(X) = \begin{cases} \emptyset & \pi(X) \geq \lambda_1 \longrightarrow \text{Defer} \\ \{y : \hat{p}(y|x) \geq \lambda_2\} & \text{otherwise} \longrightarrow \text{Predict} \end{cases}$$

- Tune λ_1 and λ_2 to control for risks using calibration dataset

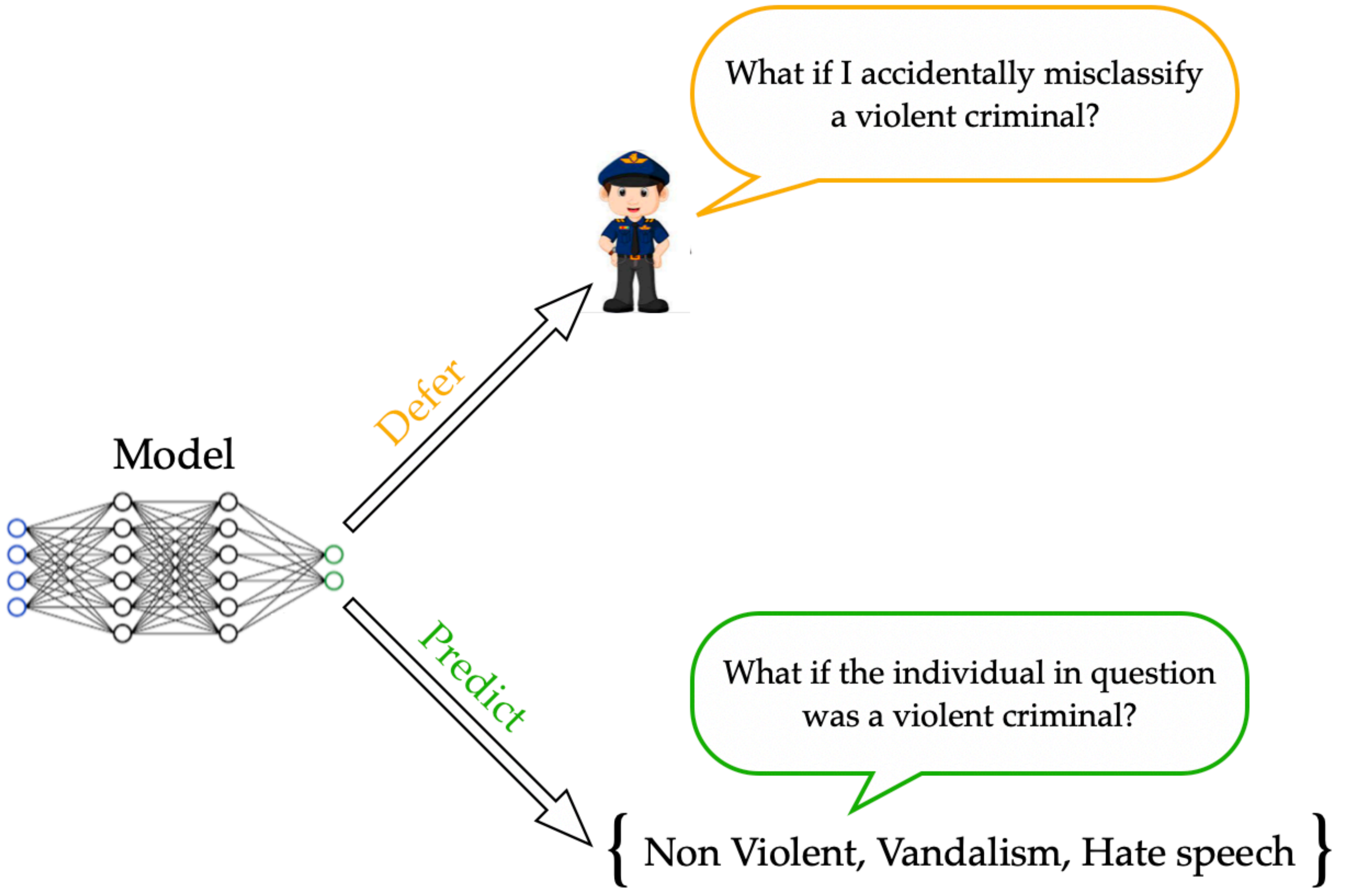


Figure: Illustration of the risks we can control

[4] Angelopoulos, Anastasios Nikolas et al. “Learn then Test: Calibrating Predictive Algorithms to Achieve Risk Control.” *ArXiv* abs/2110.01052 (2021): n. pag.

Dual False Negative Rate Control

- **Synthetic Expert:** 80 % accurate A.
  Acceptable Misclassification Rate: $\alpha_{expert} = 0.1$

- **Classifier:** $\approx$ 60 % accurate (Top-1) B.
  Acceptable FNR: $\alpha_{classifier} = 0.1$

- **Tolerance** $\delta = 0.1$

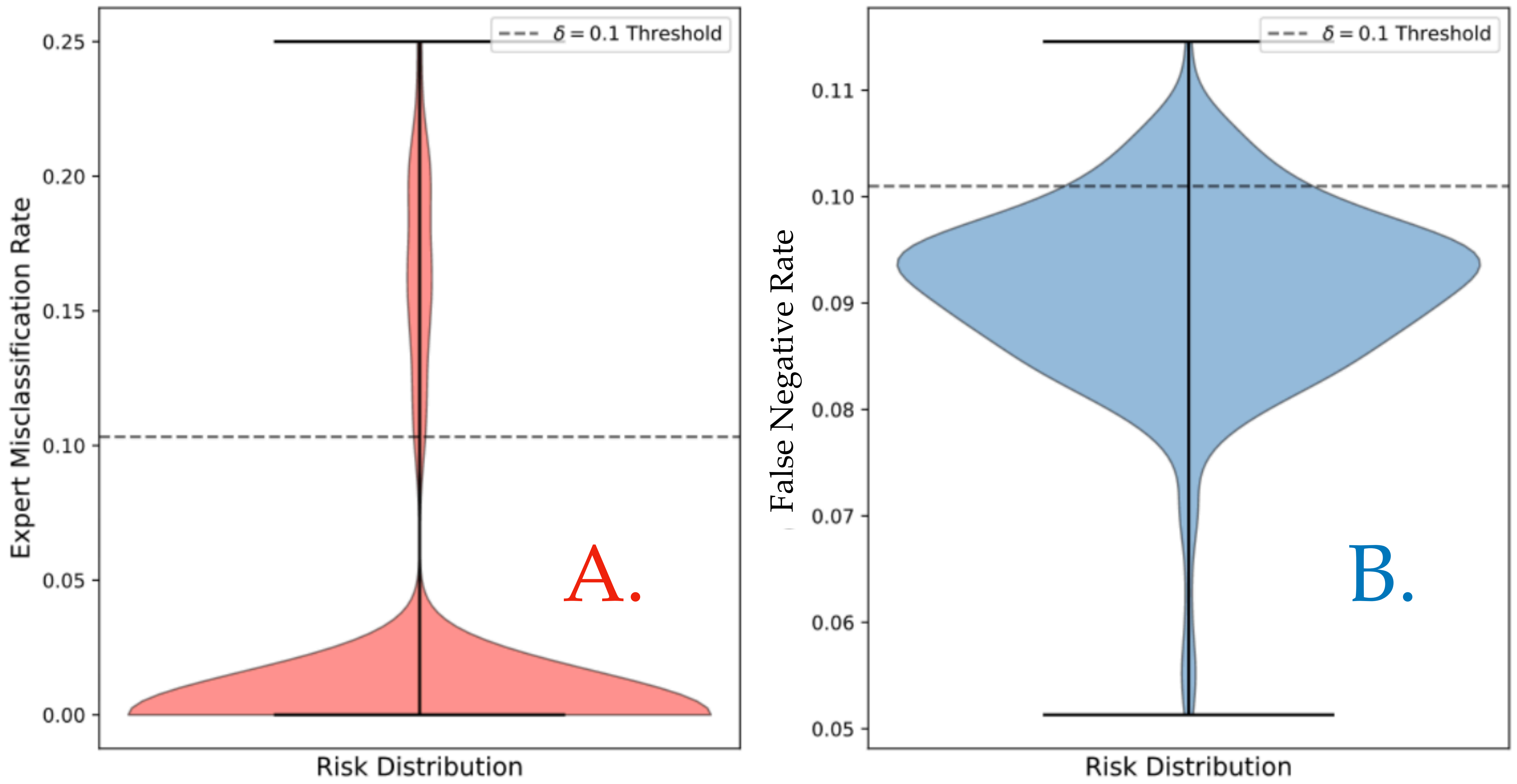


Illustration of dual risk control
1000 validation-calibration splits, CIFAR-100

Dual False Negative Rate Control

We simultaneously guarantee that the expert and set predictor have risk less than 0.1 with high probability ($1 - \delta = 0.9$)!

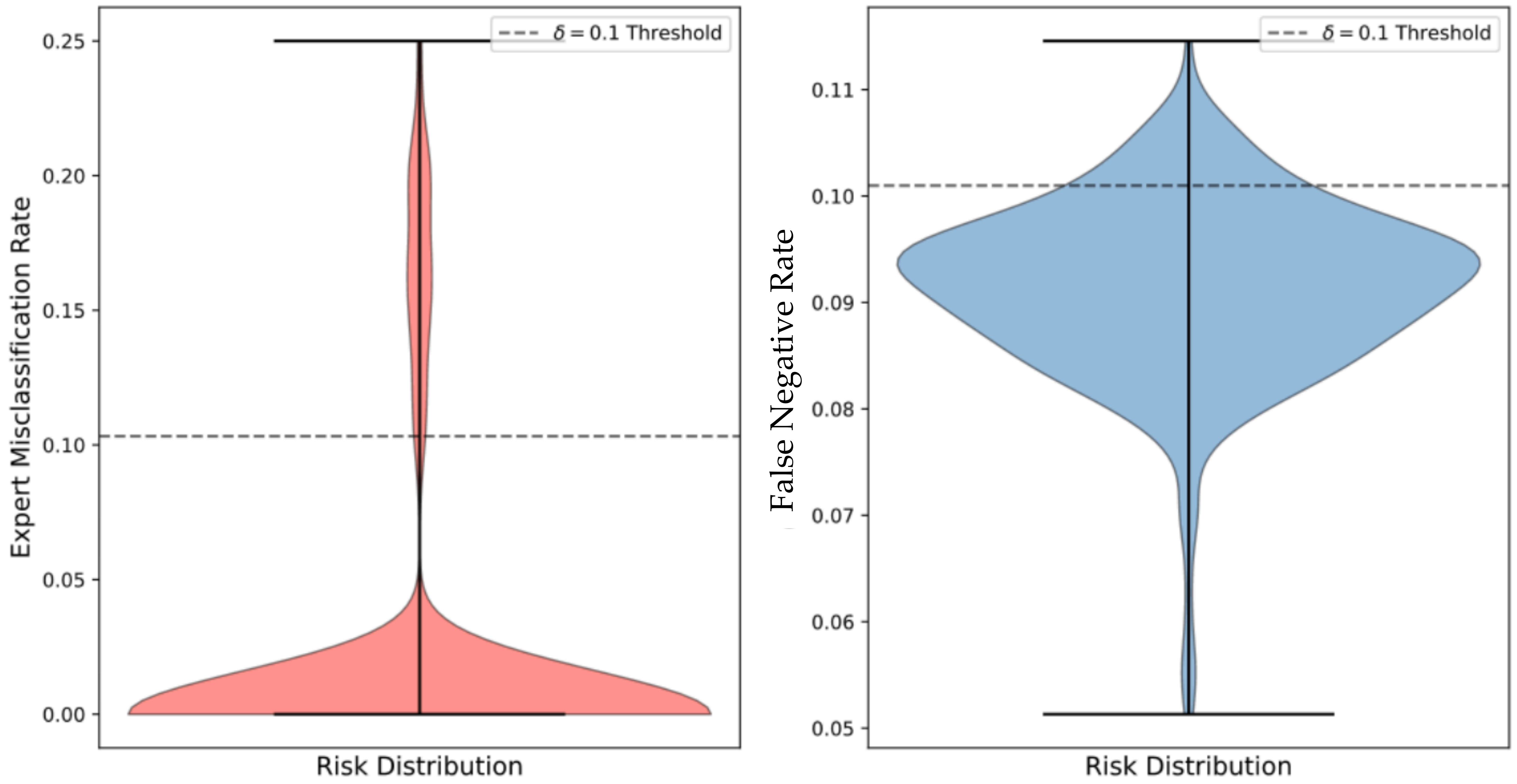


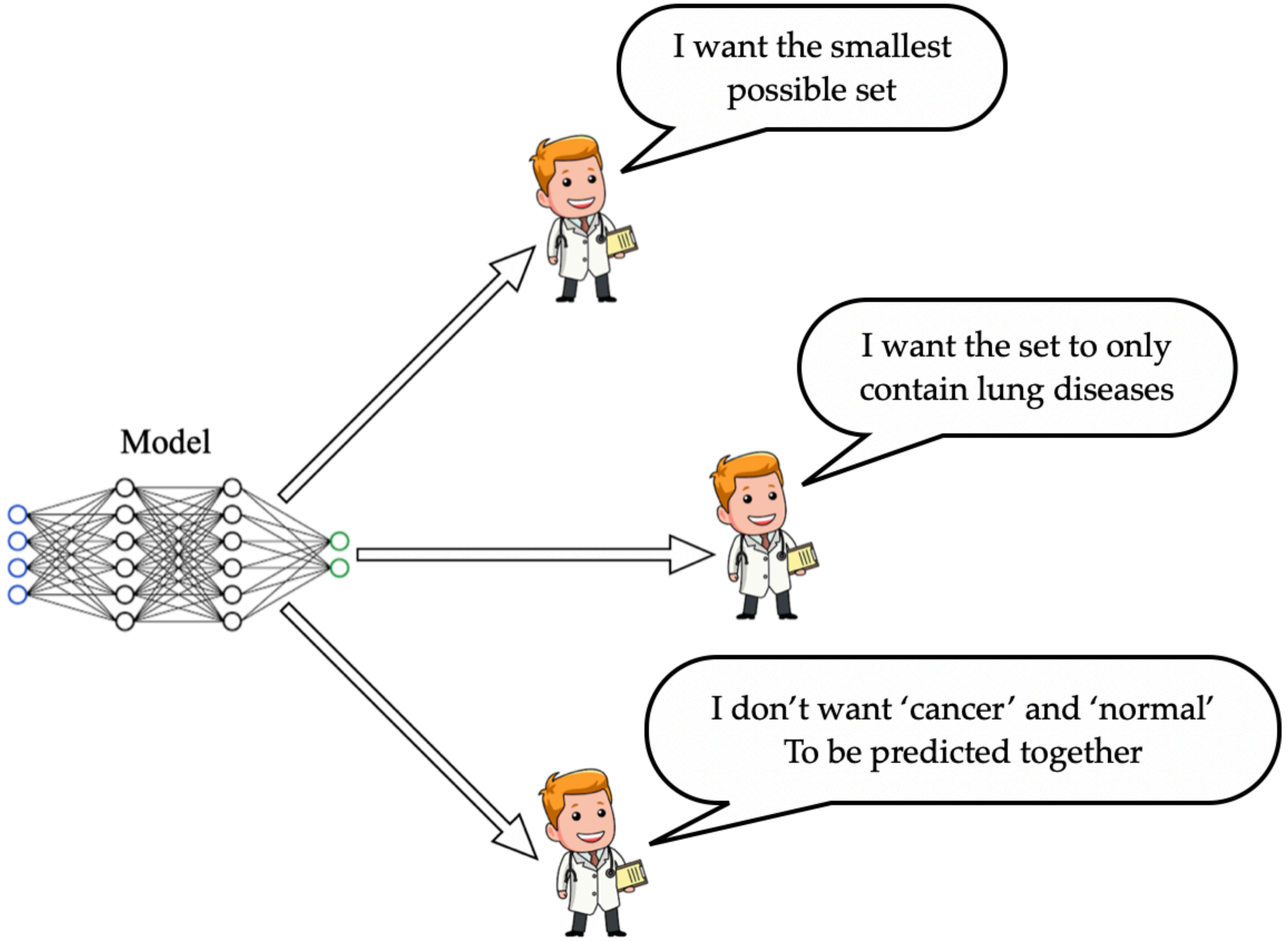
Illustration of dual risk control
1000 validation-calibration splits, CIFAR-100

Some Future Questions to Tackle

- How does the type of risk control impact the utility of the set?
- How does the error tolerance parameter impact the utility of the set?
- Can we control for the risk associated with any (*not necessarily ground truth*) label?
- Can we design better deferral policies that can improve the CP set sizes on non-deferred examples?

Can we shape predictive sets according to a human-specified heuristic?

Generating Similar Sets



- Sometimes it's not feasible to obtain human labels to train a deferral policy
- But we can still generate useful predictive sets if the human provides some form of direction!

Generating Similar Sets

- The human provides a label dissimilarity matrix M where M_{ij} = cost of predicting labels i and j together.
- Define set dissimilarity $\mathscr{D}(S) = \max_{i,j \in S} M_{ij}$
- We can construct predictive sets that reduce $\mathscr{D}(S)$ whilst providing the same risk guarantees!

Dissimilar Set	{Sweet Pepper, Apple, Orange, Tulip} High $\mathscr{D}(S)$
Similar Set	{Sweet Pepper, Apple, Orange} Low $\mathscr{D}(S)$

Both sets provide the same risk guarantees!

A Proof of Concept with Semantically Similar Sets

- Say we want sets that contain semantically similar labels

- Define a label dissimilarity cost matrix M s.t
  $M_{ij} = d(y_i, y_j) = |\text{emb}(y_i) - \text{emb}(y_j)|$

- $\text{emb}(y_i)$ = Word embedding of label y_i

$\{$ Dog, Cat, Ship $\}$ Horse

d(Horse,Dog) $<$ d(Ship,Dog)

d(Horse, Cat) $<$ d(Ship,Cat)

Examples of Semantically Similar Sets: CIFAR-100

Acceptable Label (Minor Penalty)
Ground Truth Label
Undesirable Label (Major Penalty)

RCPS

$\mathcal{S}$ = {Palm Tree, Pine Tree, Forest, Bridge}

$D(\mathcal{S}) = 2.445$

$\mathcal{S}$ = {Sweet Pepper, Apple, Orange, Tulip}

$D(\mathcal{S}) = 2.499$

RCPS-LD

$\mathcal{S}$ = {Palm Tree, Pine Tree, Forest, Willow Tree, Oak Tree}

$D(\mathcal{S}) = 1.001$

$\mathcal{S}$ = {Sweet Pepper, Apple, Orange}

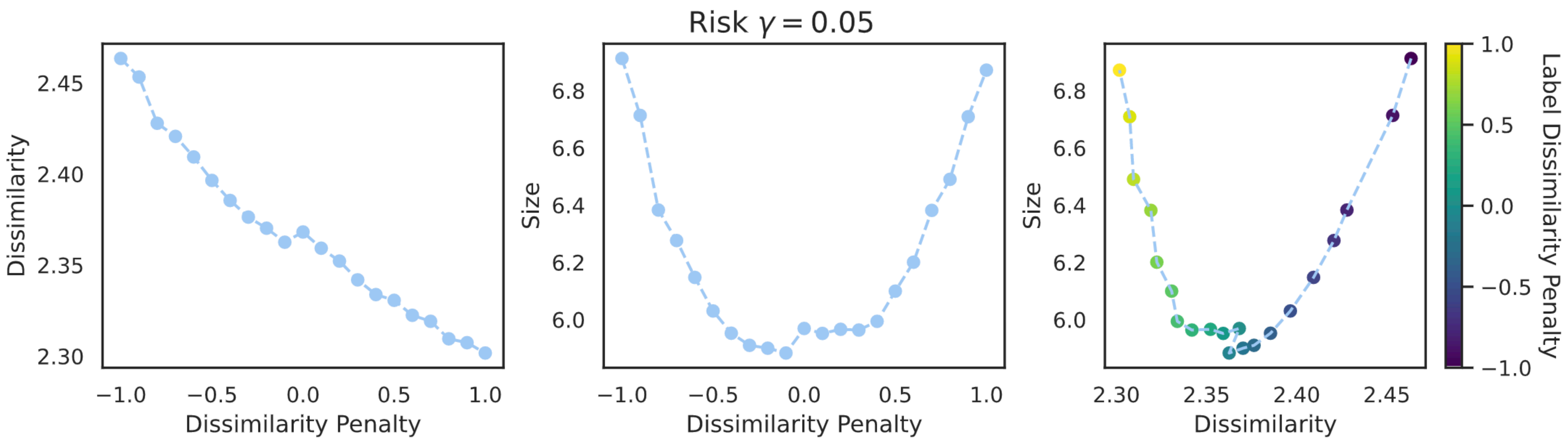
$D(\mathcal{S}) = 1.501$

Both sets provide the same risk guarantees!

But the bottom sets have semantically similar labels!

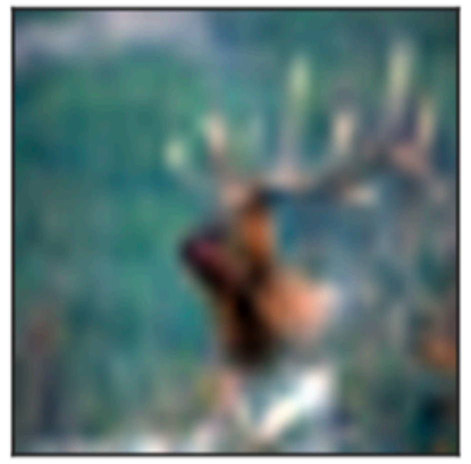
Label Similarity Experiments

- Define label dissimilarity penalty μ
- $\mu > 0 \Rightarrow$ we obtain more similar sets
- $\mu < 0 \Rightarrow$ we obtain more dissimilar sets
- But there is a tradeoff between label similarity / dissimilarity and predictive set size!



Examples of Semantically Similar Sets: CIFAR-10

Acceptable Label (Minor Penalty)
Ground Truth Label
Undesirable Label (Major Penalty)



RCPS	{Horse, Airplane}	{Deer, Bird, Airplane}	{Airplane, Bird}
	$D(\mathcal{S}) = 1.673$	$D(\mathcal{S}) = 1.689$	$D(\mathcal{S}) = 1.677$
RCPS-LD	{Horse, Cat, Dog, Deer}	{Deer, Bird, Frog, Cat}	{Airplane}
	$D(\mathcal{S}) = 1.651$	$D(\mathcal{S}) = 1.536$	$D(\mathcal{S}) = 0$

Some other cool properties of D-CP uncovered

- Humans are negatively influenced by incorrect labels in CP sets - this effect is less pronounced in D-CP sets!

- We can jointly control for the misclassification rate of the human and the false negative rate of the model by learning two thresholds!

Appendix

Appendix: Dual False Negative Rate Control

- **Synthetic Expert:** 80 % accurate
  Acceptable Misclassification Rate: $\alpha_{expert} = 0.1$

- **Classifier:** $\approx$ 60 % accurate (Top-1)
  Acceptable FNR: $\alpha_{classifier} = 0.1$

- **Tolerance** $\delta = 0.1$

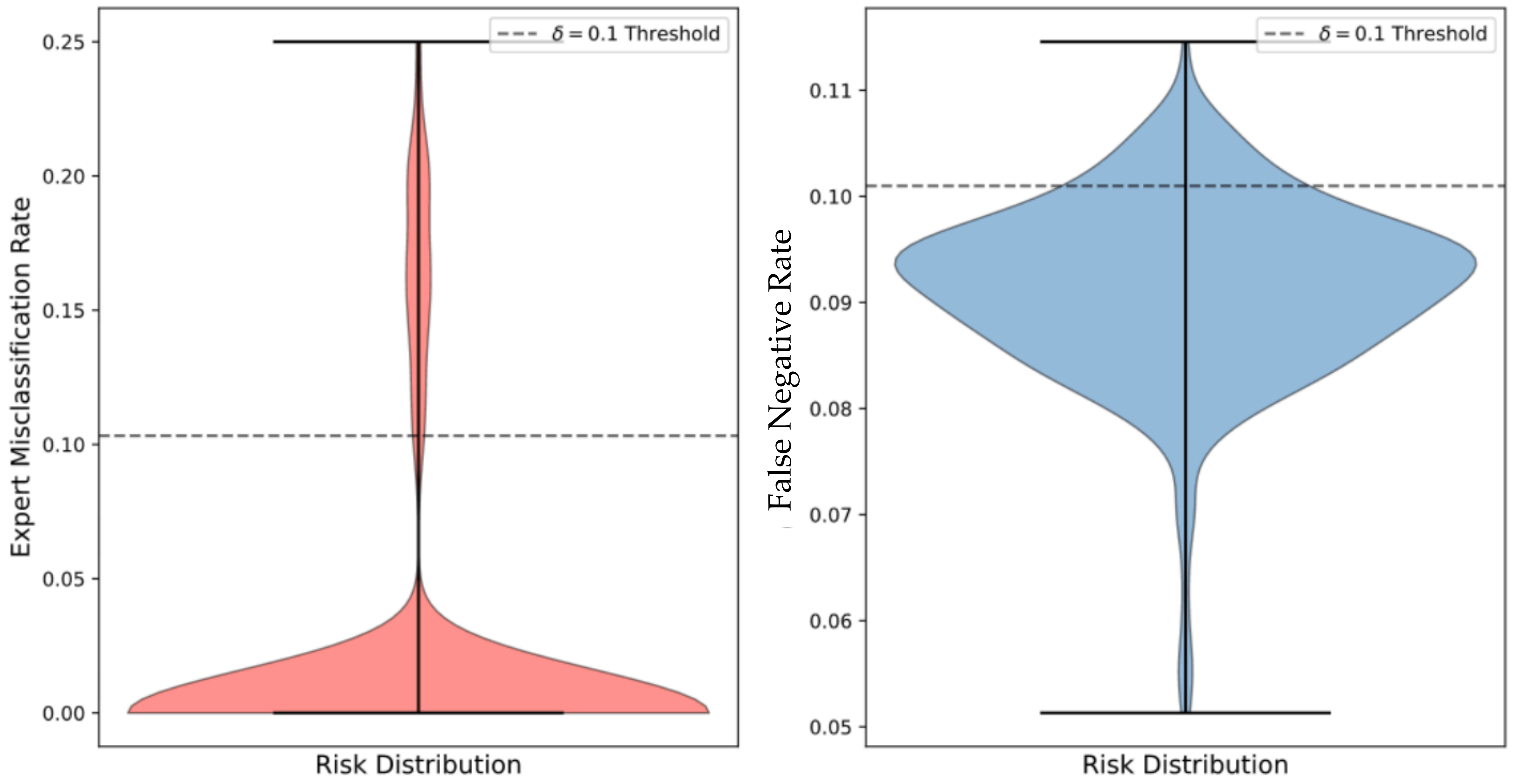


Illustration of dual risk control with a synthetic expert
$\delta = 0.1$, 1000 validation-calibration splits

Appendix: Theoretical Results

- **Theorem 1:** *If a deferral policy $\pi(X)$ defers examples such that the risk on non-deferred examples is lower than before (i.e. $\mathbb{E}[L(Y, \Gamma(X)) \,|\, \pi(X) = 0] \leq \mathbb{E}[L(Y, \Gamma(X))]$), then the prediction set will contain fewer incorrect labels on average*

- **Theorem 2:** *Given any deferral policy $\pi(X)$, set-valued classifier $\Gamma(X)$, and human expert $h(X)$, we can control for the false negative rate of the model on non-deferred examples and expert misclassification rate on deferred examples with high probability, i.e.*

$$P(P(Y \notin \Gamma(X) \,|\, \pi(X) = 0) \leq \alpha_1) \geq 1 - \delta$$
$$P(P(h(X) \notin Y \,|\, \pi(X) = 1) \leq \alpha_1) \geq 1 - \delta$$

for suitably defined α_1, α_2, δ

Appendix: Human Subject Evaluation of D-CP

Metric	D-RAPS	RAPS	p value	Effect Size
Accuracy	0.88 ± 0.05	0.81 ± 0.04	0.058	0.508
Reported Utility	7.93 ± 0.39	6.19 ± 0.62	$< \mathbf{0.001}$	1.211
Reported Confidence	7.78 ± 0.33	7.31 ± 0.34	0.059	0.507

Table 5: D-RAPS vs RAPS: Non-Deferred Examples
$\alpha = 0.1$, deferral rate $b = 0.2$, CIFAR-100

Metric	RAPS	D-RAPS	N	p-value	Effect Size	N_{min}
Accuracy (All)	0.67	0.76	30	**0.003**	0.87	22
Accuracy (Easy)	0.87	0.83	30	0.310	0.27	218
Accuracy (Difficult)	0.55	0.67	30	$< \mathbf{0.001}$	1.04	16

Table 4.10: Accuracy of participants when shown RAPS vs D-RAPS sets on examples stratified by difficulty. N_{min} is the minimum sample size for each group needed for $p \leq 0.05$ with power $1 - \beta = 0.8$ and N is the experimental sample size of each group.