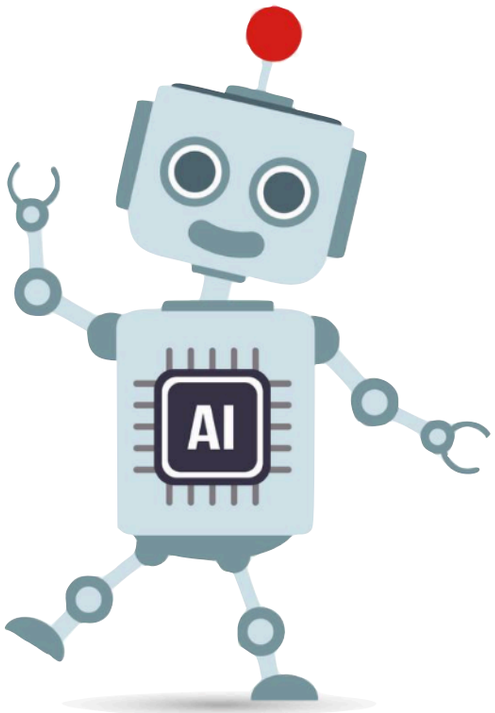
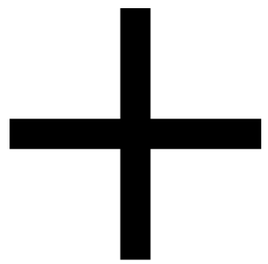
On the Calibration of Learning-to-Defer Systems

Eric Nalisnick

Rajeev Verma

Daniel Barrejón





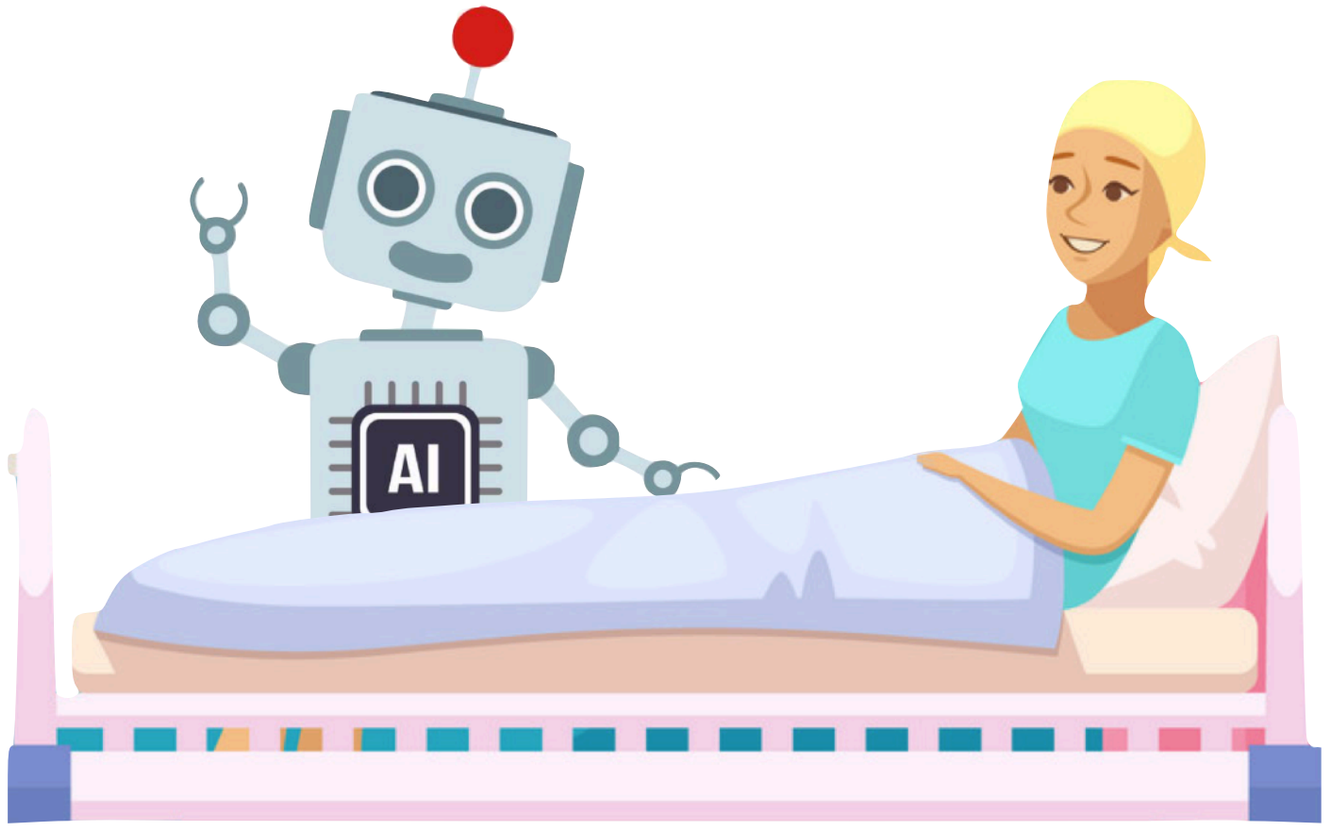
Model

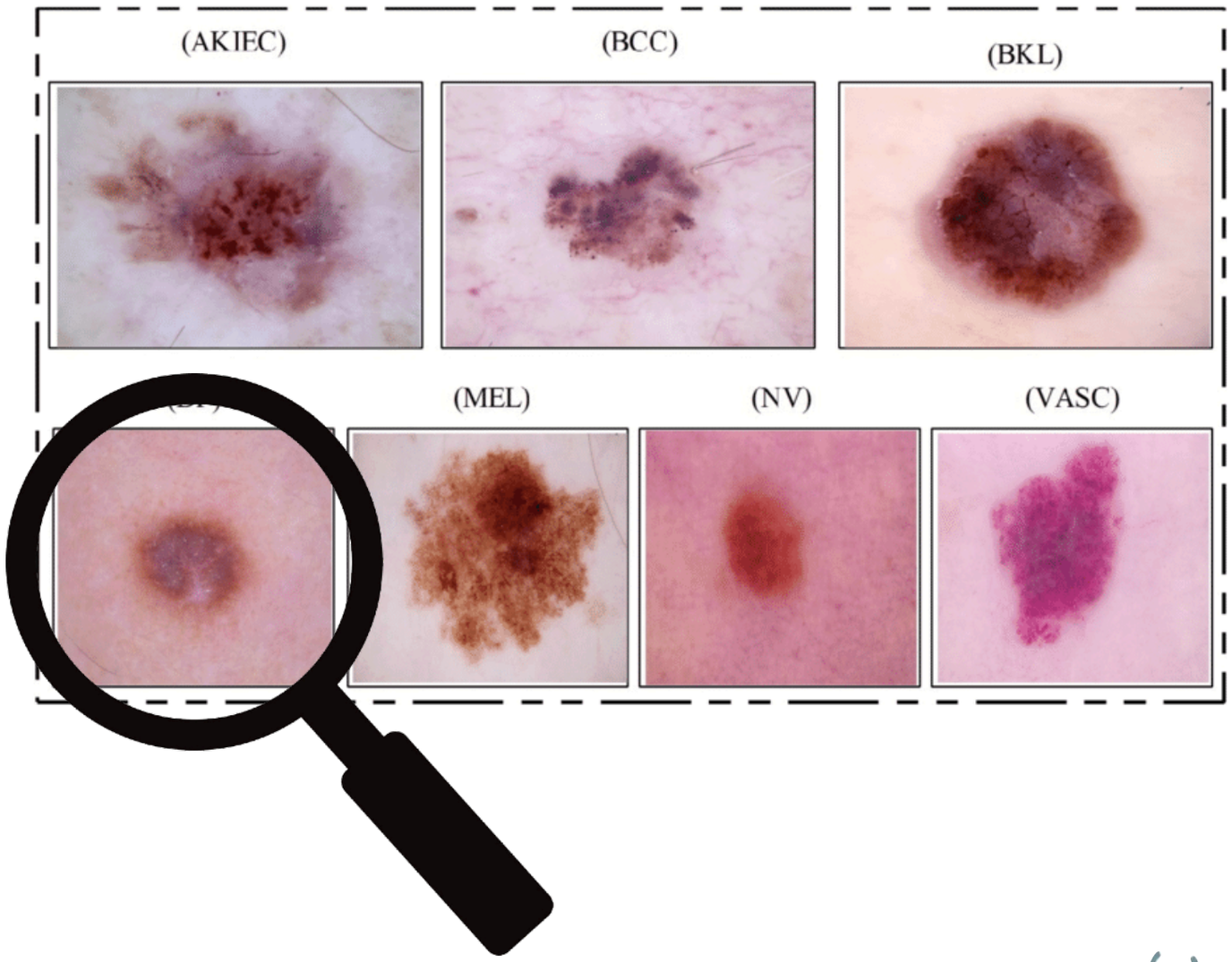
Human Expert

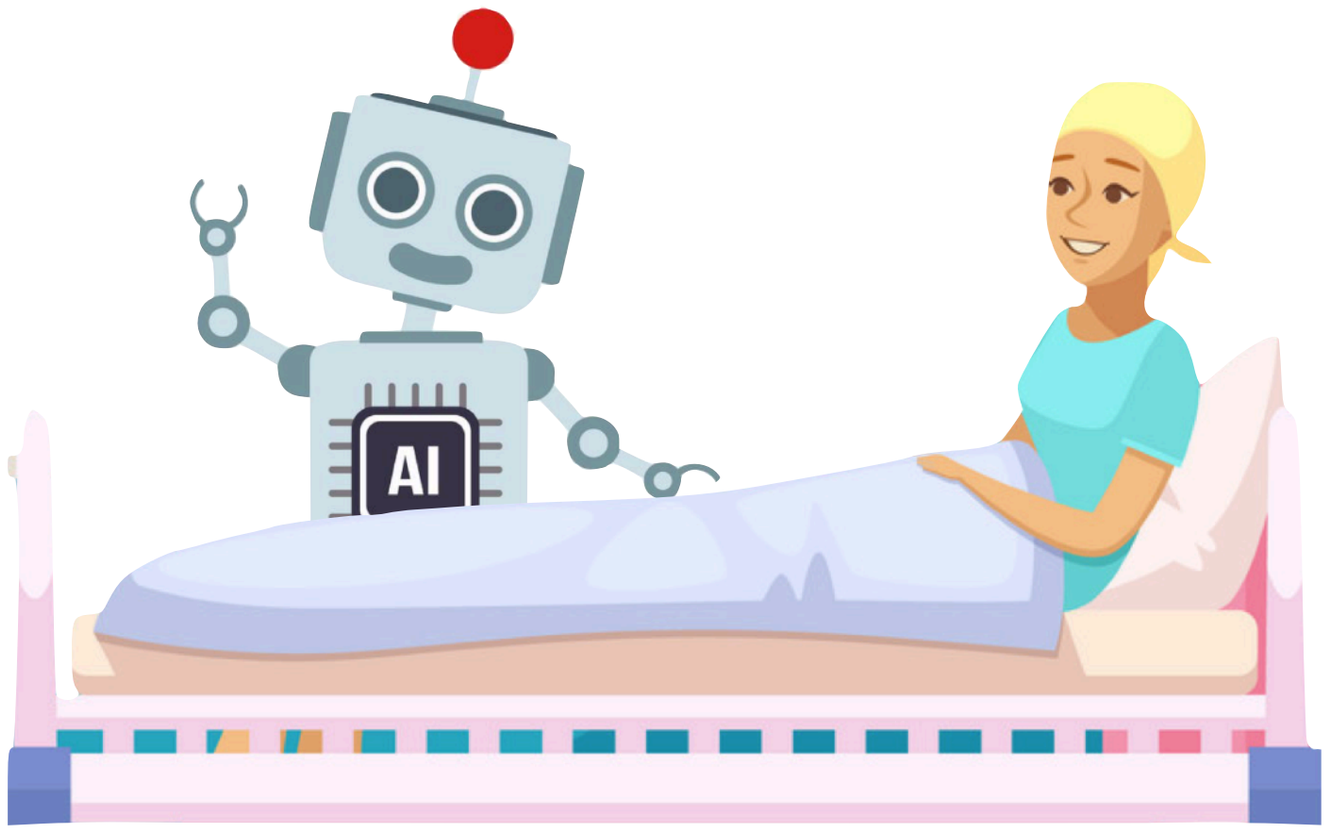
~~Artificial Intelligence~~

Hybrid Intelligence

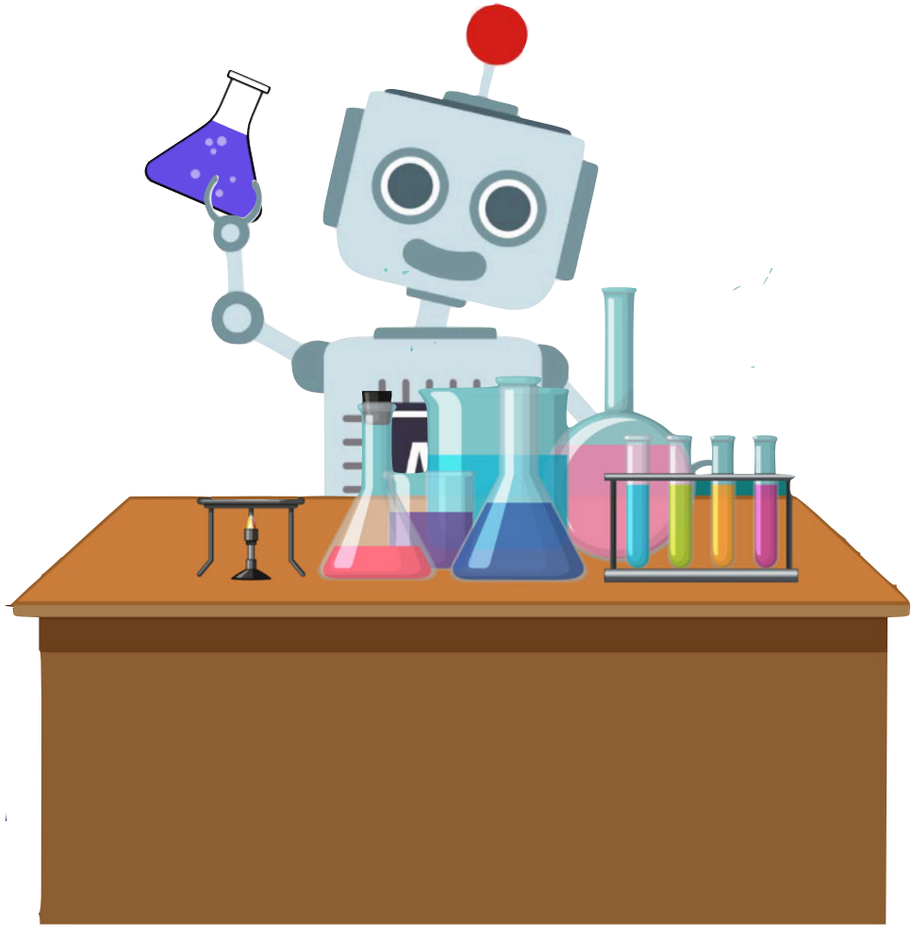
Safe Automation in Healthcare

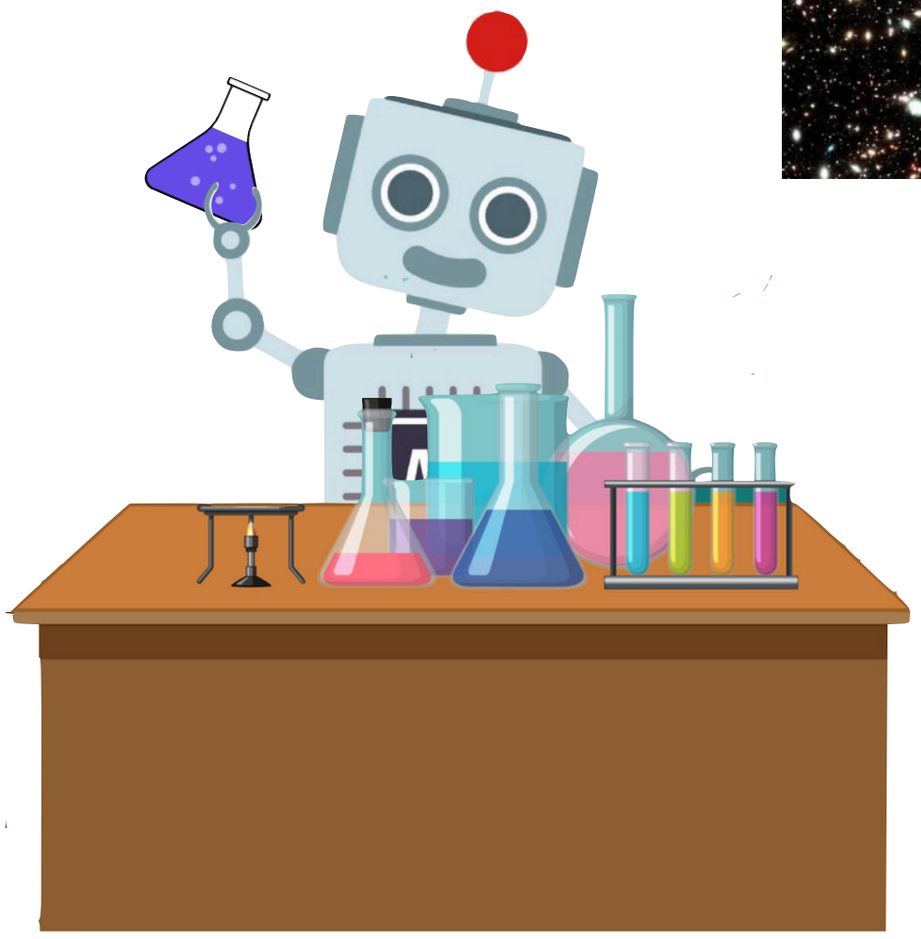


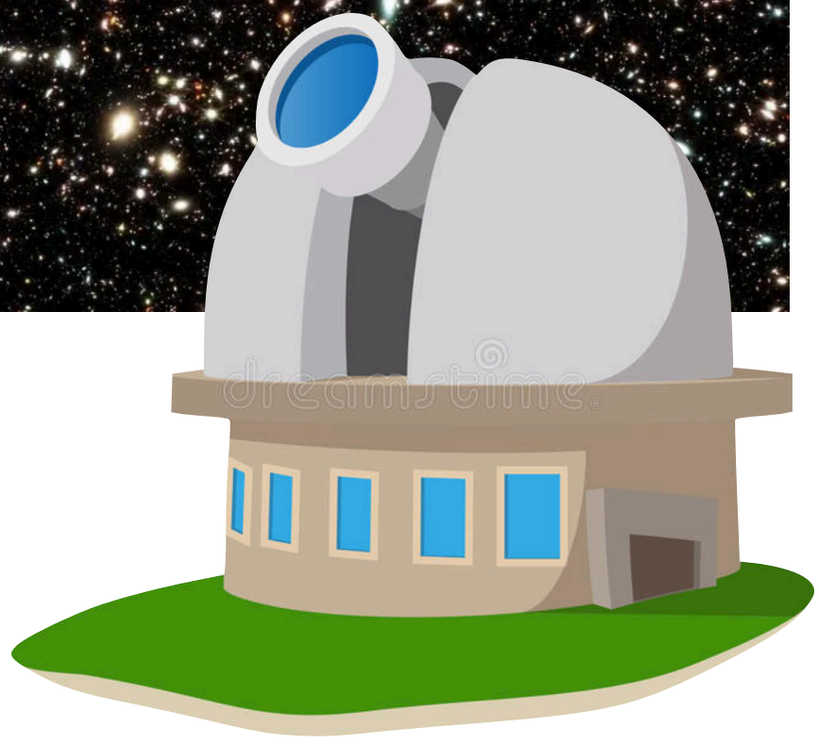

(AKIEC)
(BCC)
(BKL)
(MEL)
(NV)
(VASC)

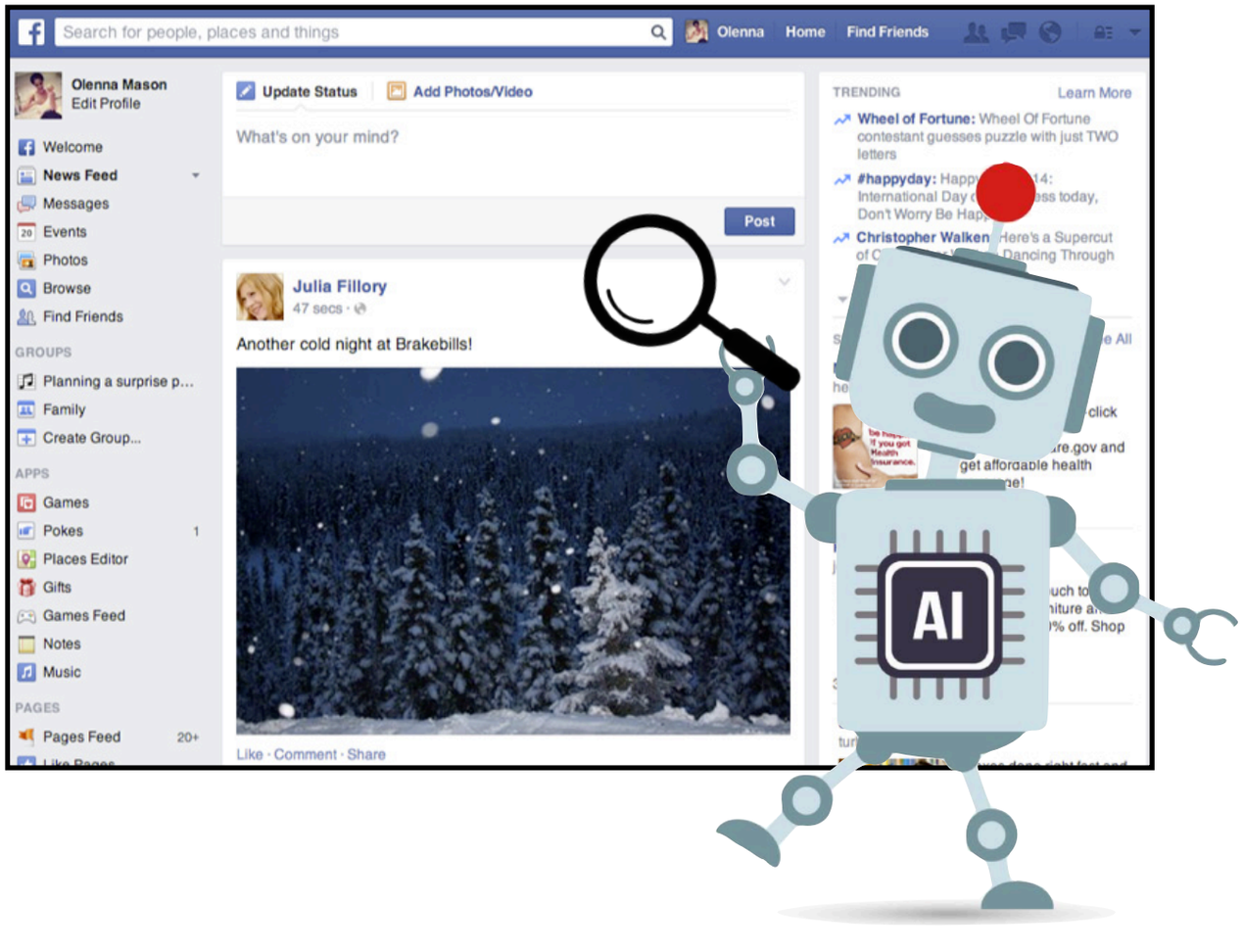

AI

Large-Scale Science

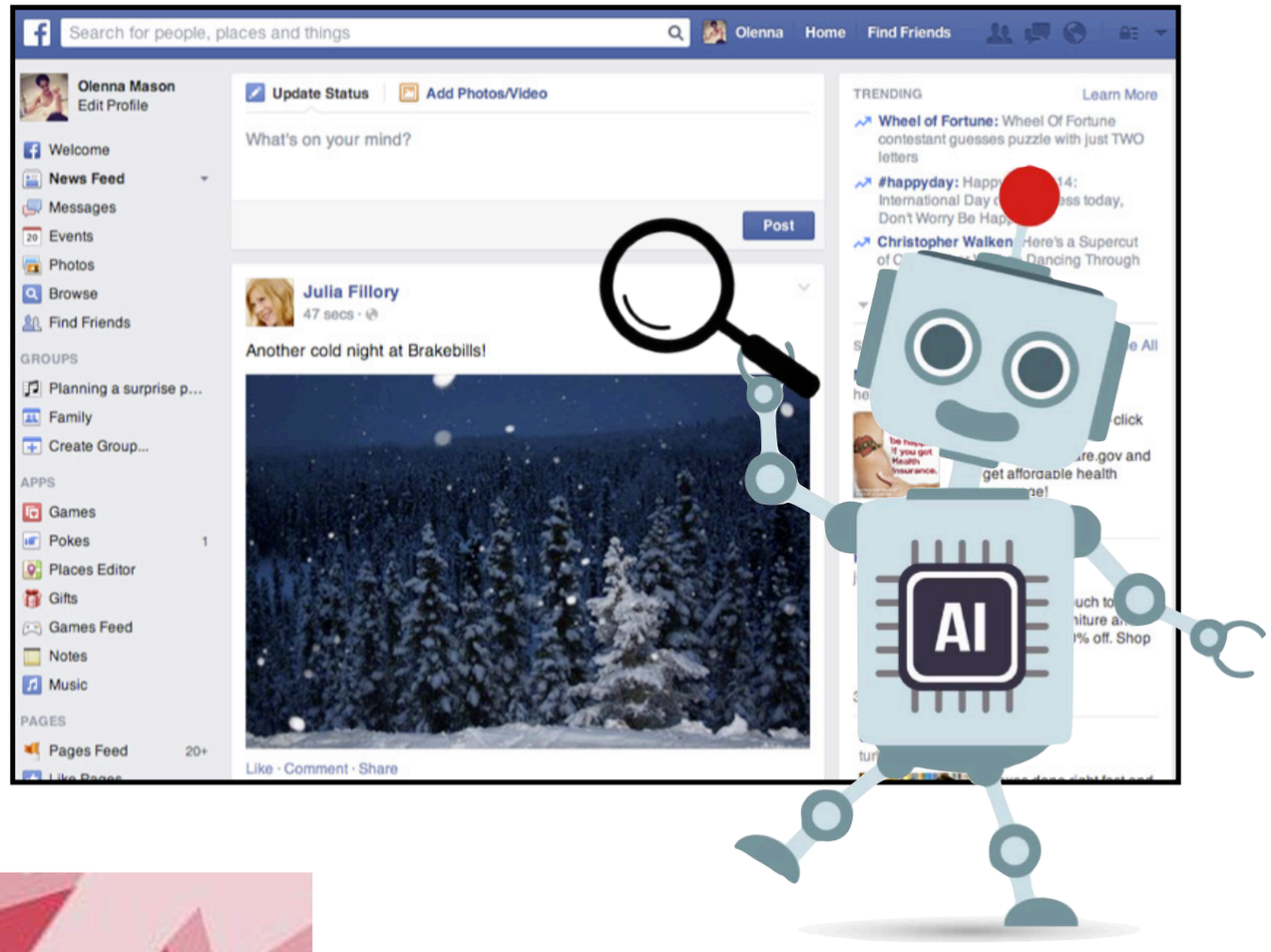




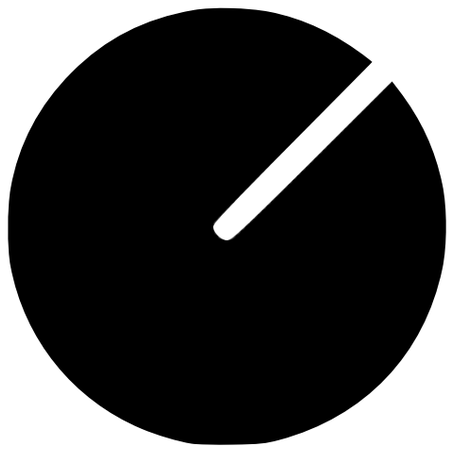


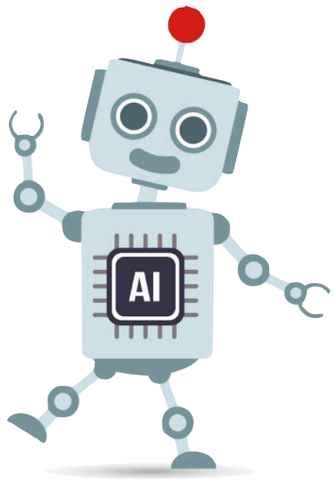


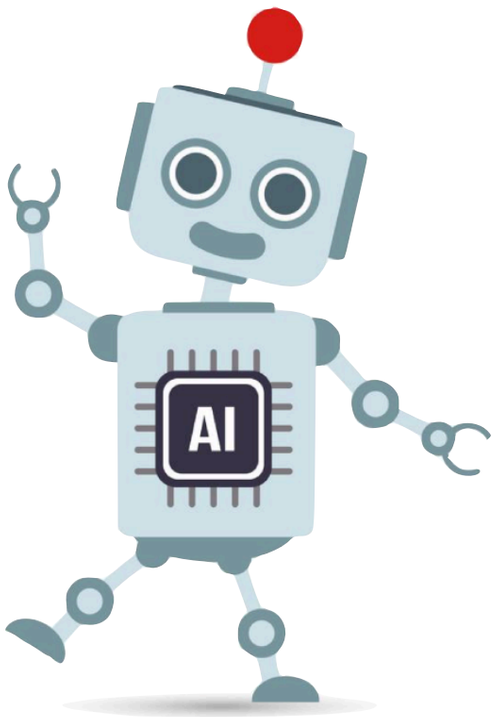
Online Content Moderation


Search for people, places and things
Olenna
Home
Find Friends
Olenna Mason
Edit Profile
Welcome
News Feed
Messages
Events
Photos
Browse
Find Friends
GROUPS
Planning a surprise p...
Family
Create Group...
APPS
Games
Pokes
Places Editor
Gifts
Games Feed
Notes
Music
PAGES
Pages Feed
Update Status
Add Photos/Video
What's on your mind?
Post
Julia Fillory
47 secs
Another cold night at Brakebills!
Like · Comment · Share
TRENDING
Learn More
Wheel of Fortune: Wheel Of Fortune contestant guesses puzzle with just TWO letters
Christopher Walken
AI

Gradual Automation



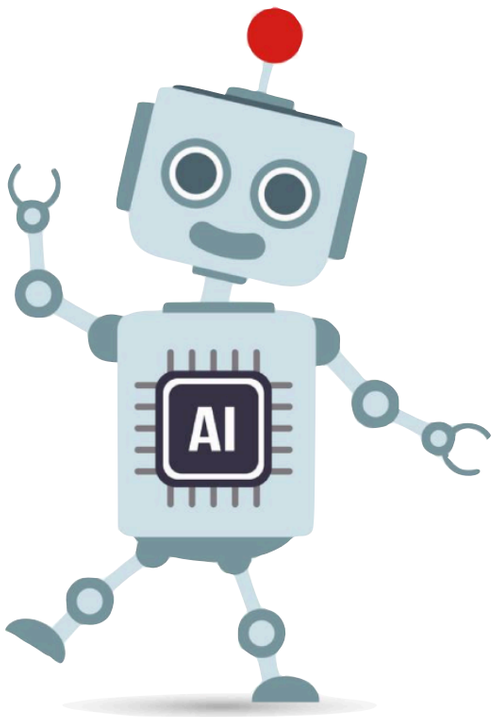




Model

?

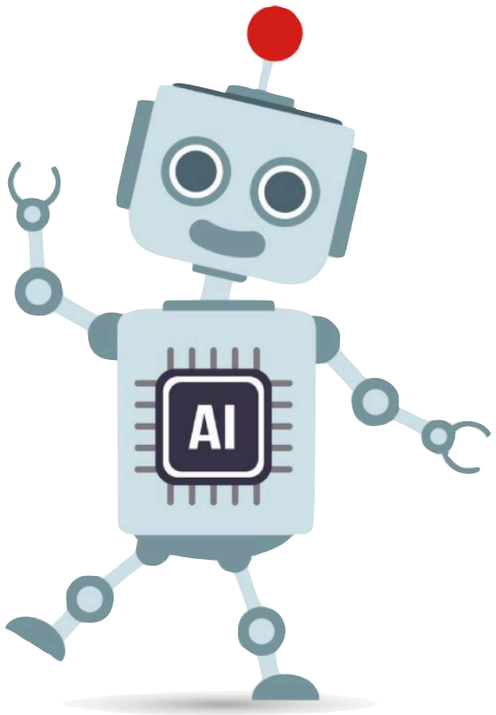
Human Expert



Classifier

?

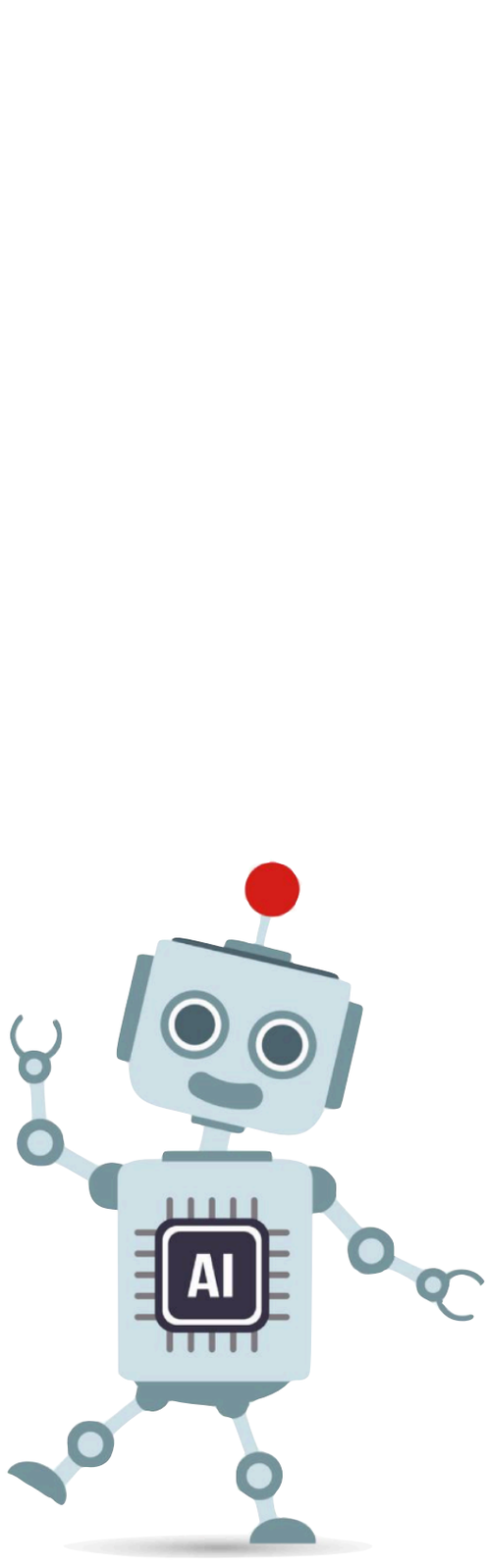
Human Expert



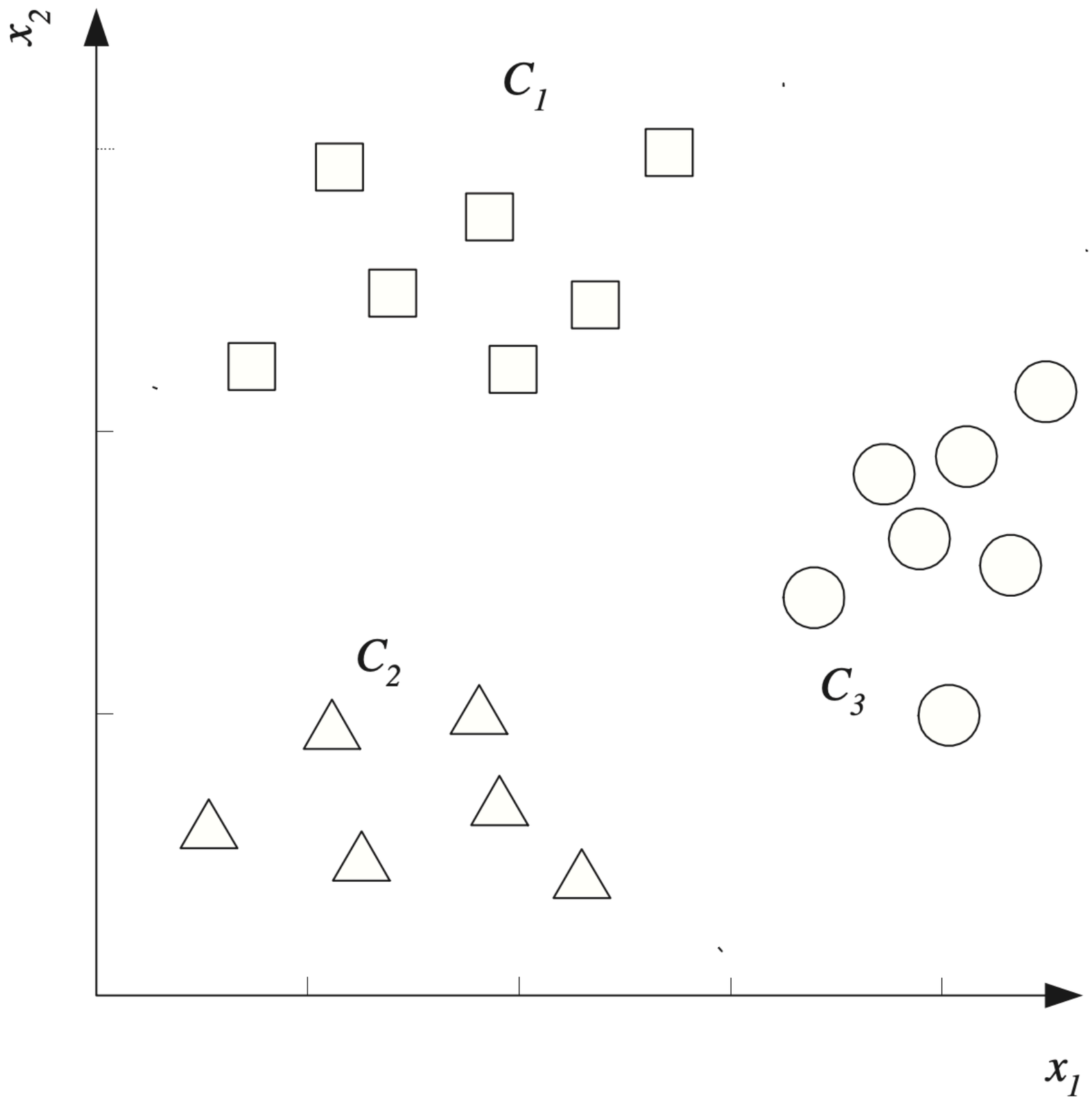
Classifier

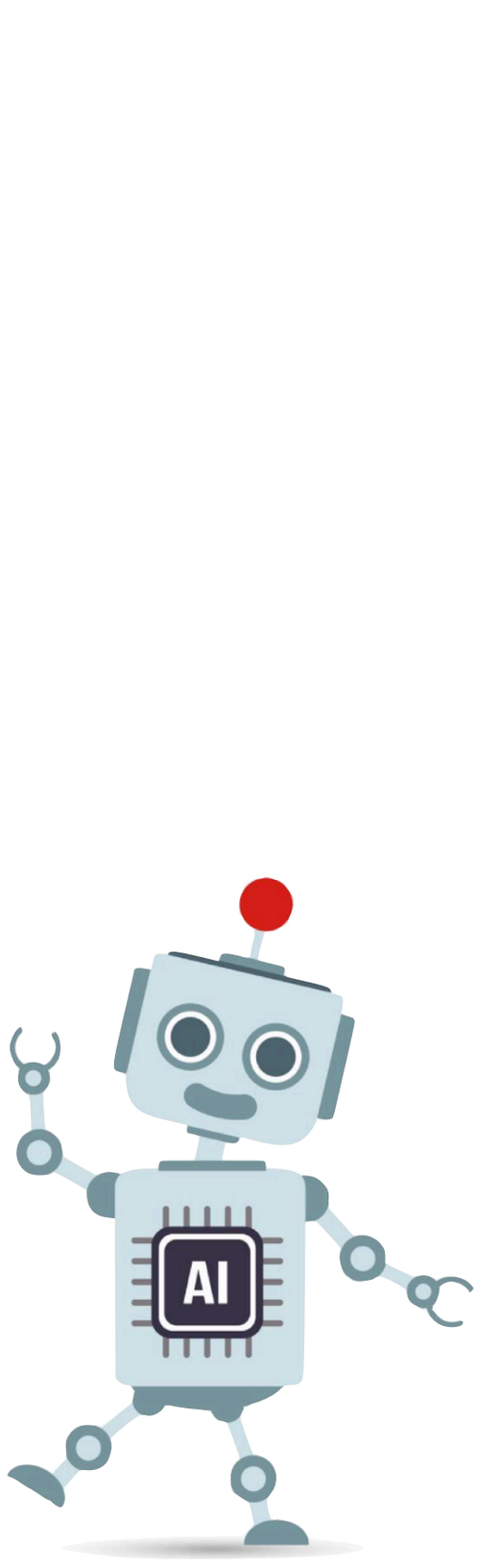
?

Human Expert

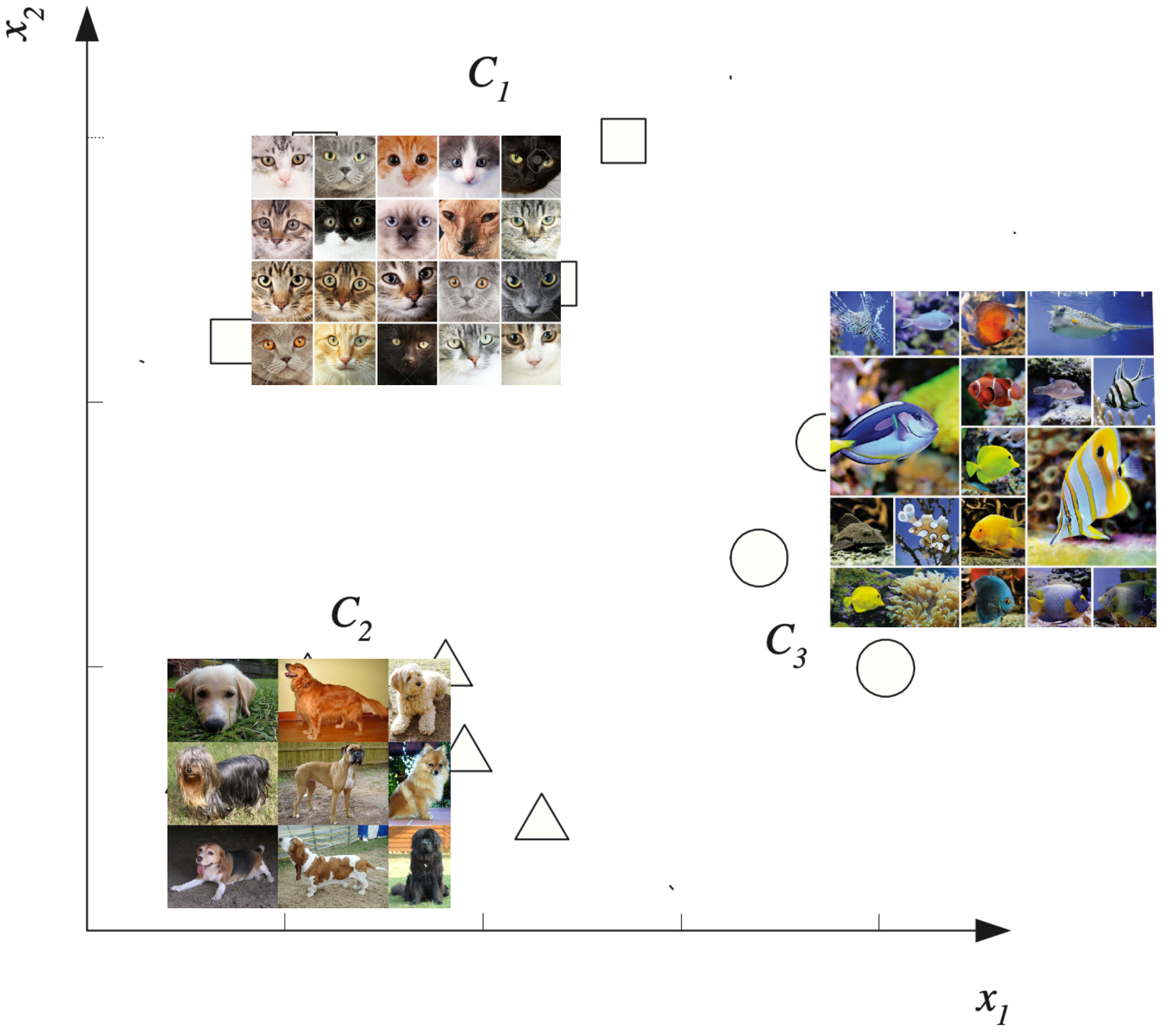


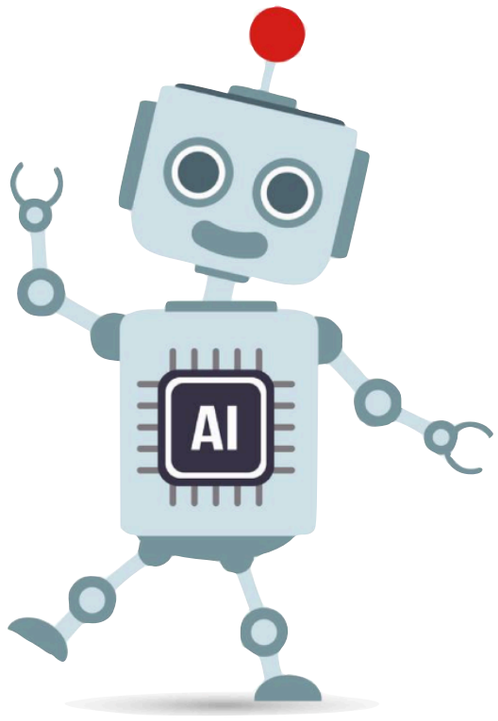
Classifier



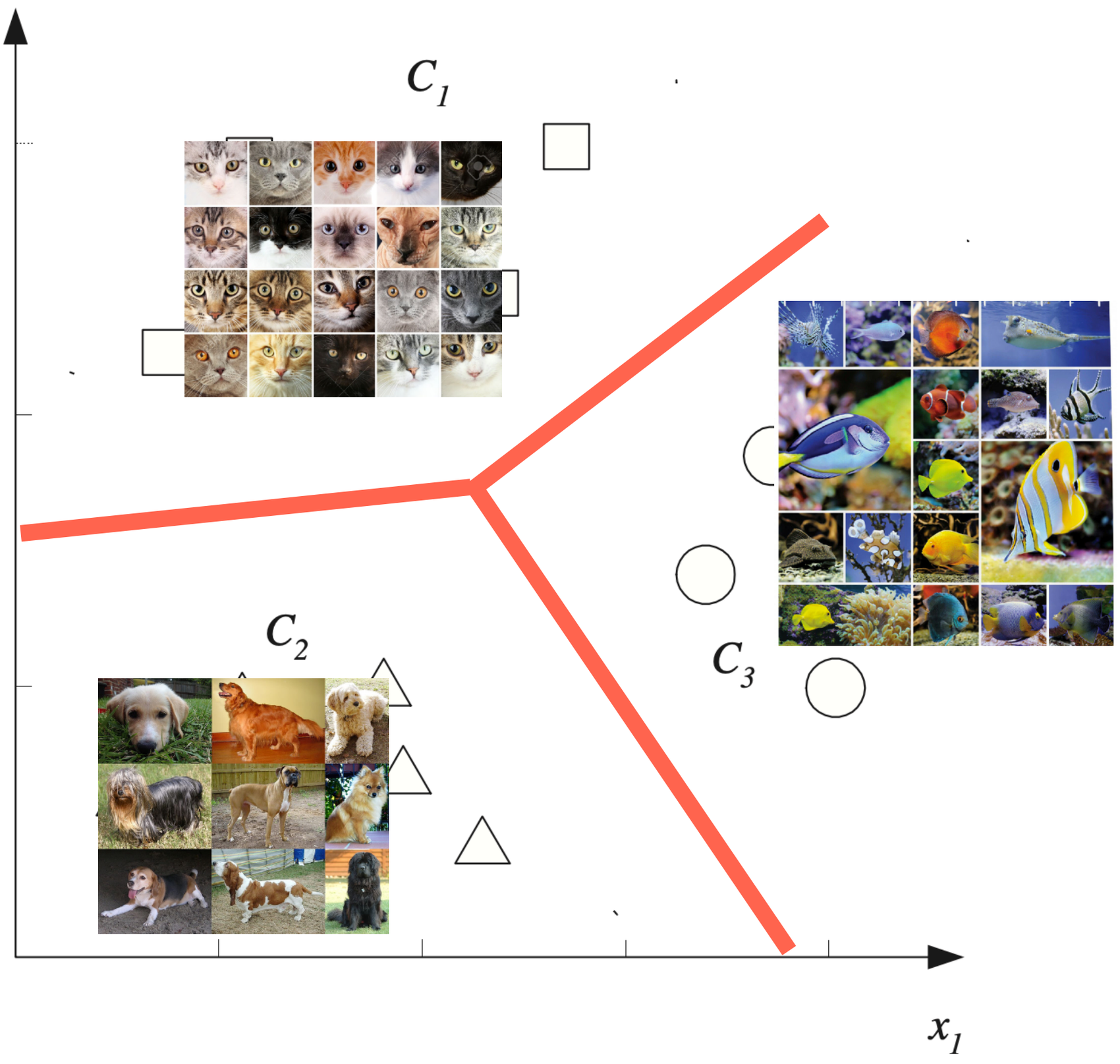


Classifier

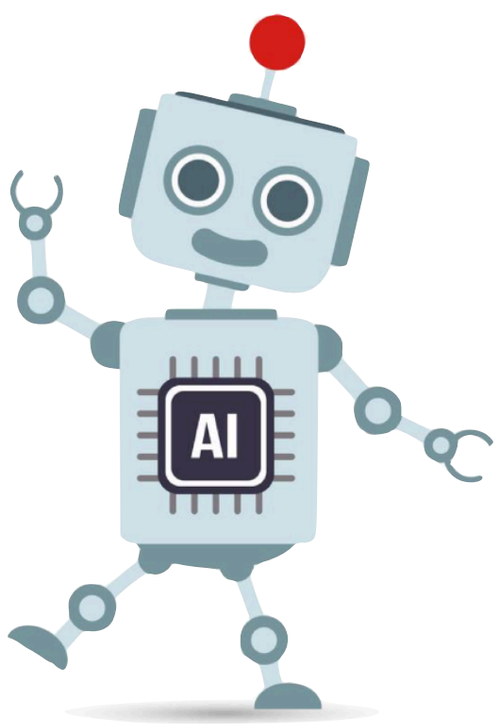




Classifier

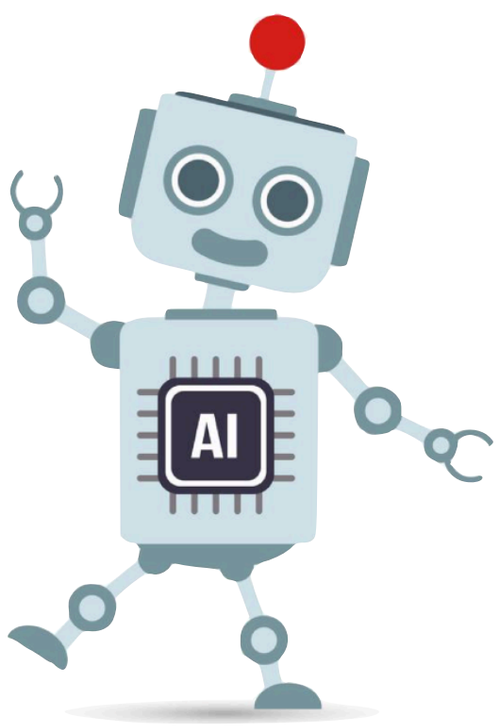


Goal: For input features **x**, predict membership in one of *K* classes.



Classifier

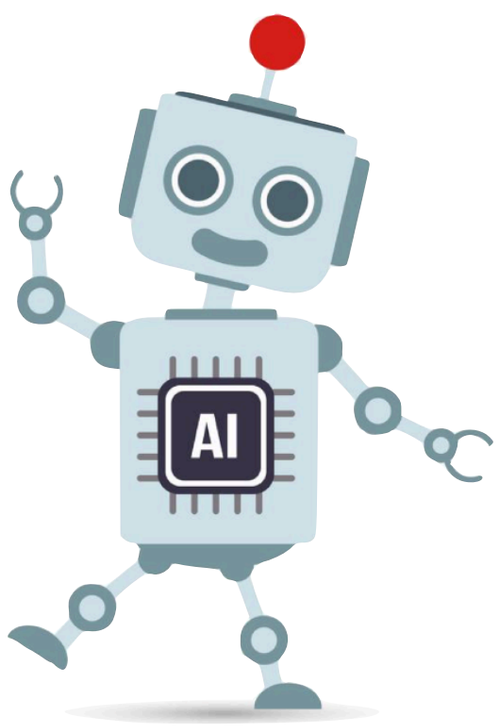
Goal: For input features **x**, predict membership in one of *K* classes.



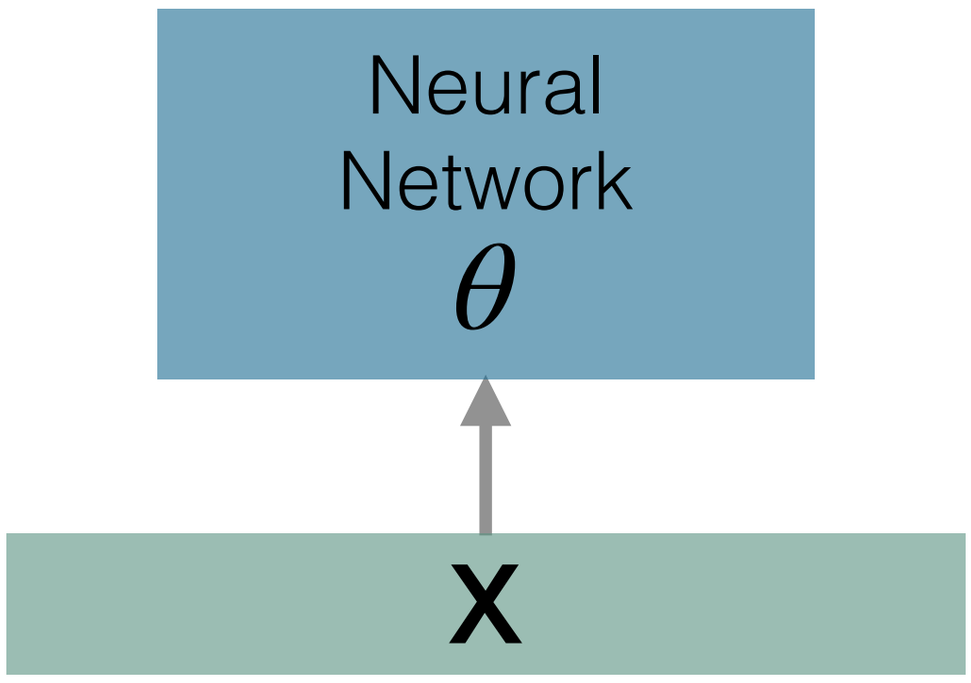
Classifier

X

Goal: For input features **x**, predict membership in one of *K* classes.

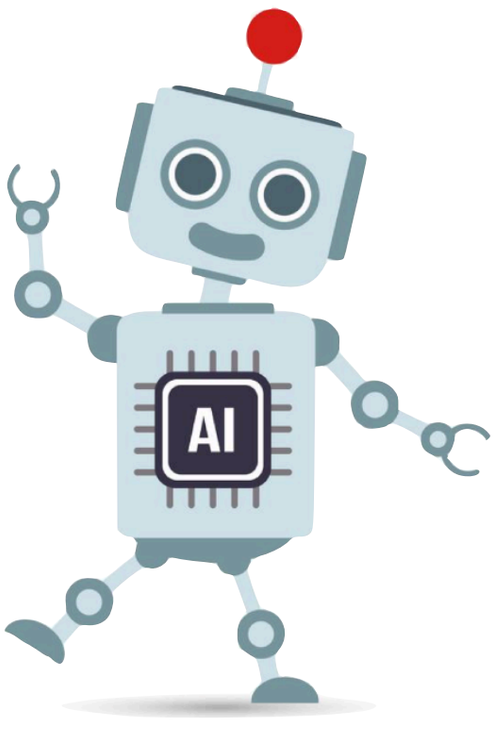


Classifier

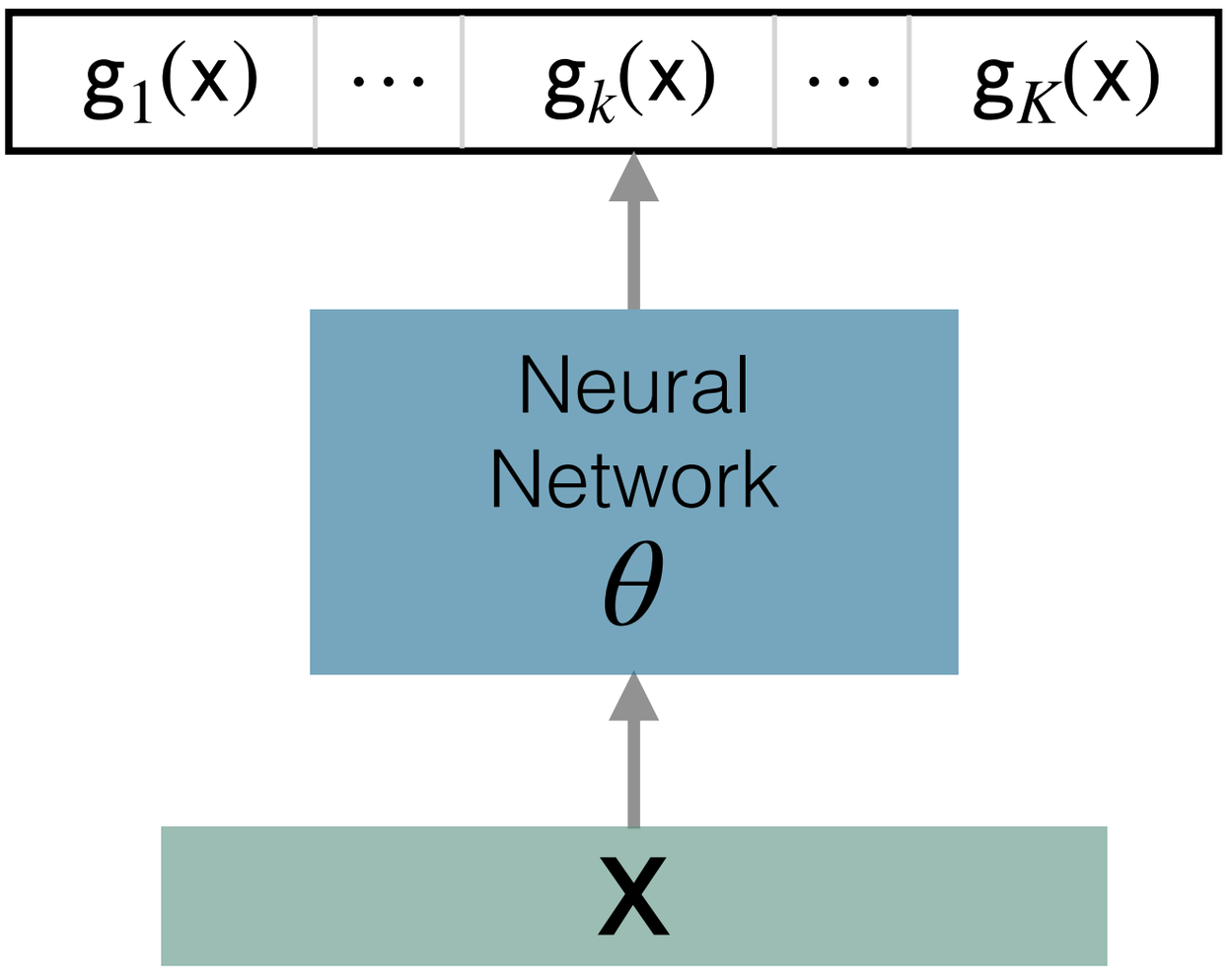


Goal: For input features **x**, predict membership in one of *K* classes.

$$\mathbf{g}_k(\mathbf{x}) \in \mathbb{R}$$

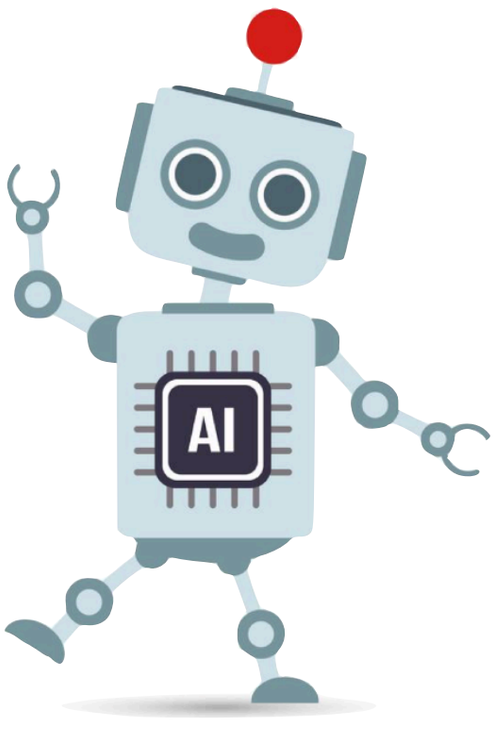


Classifier

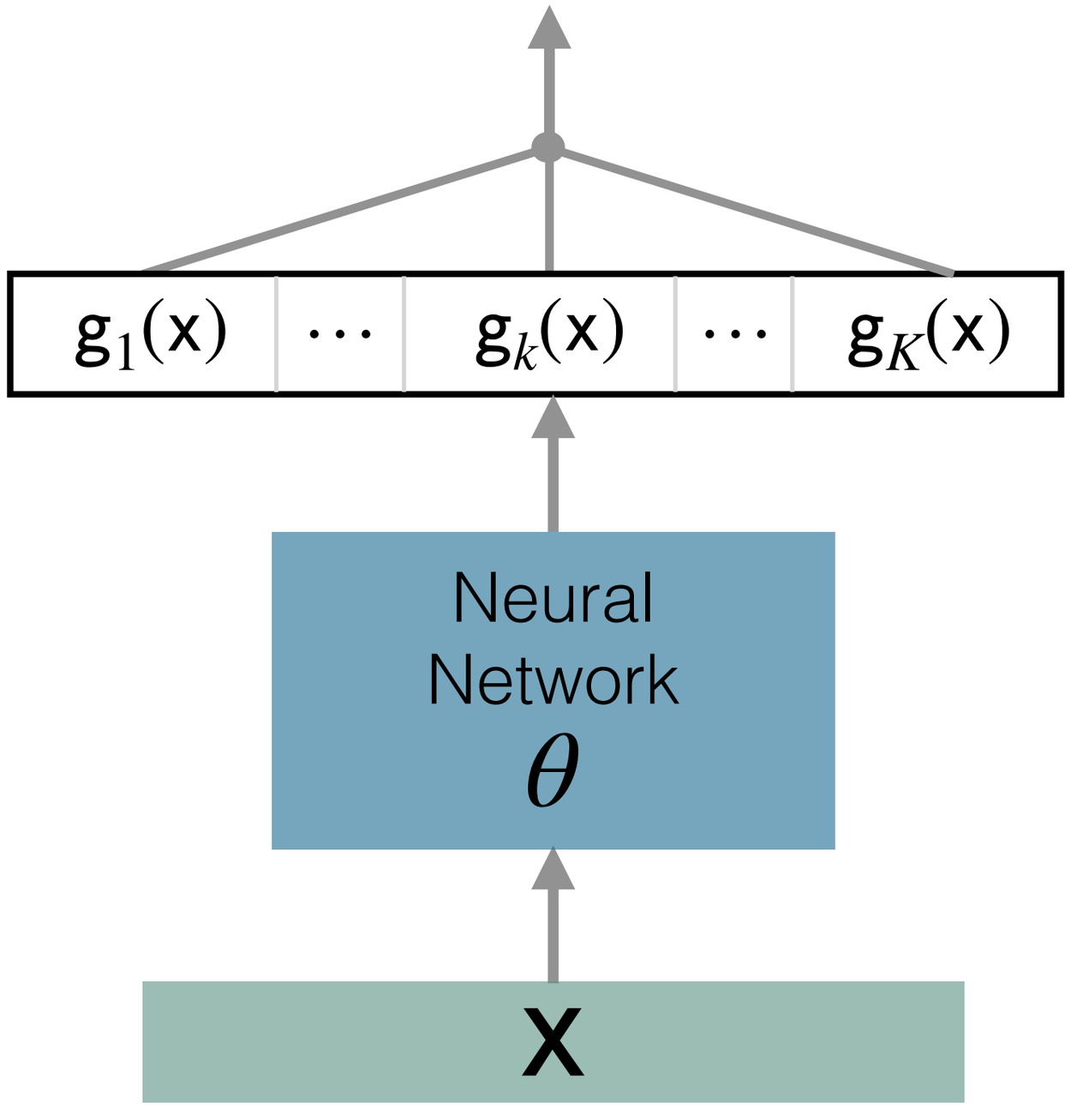


Goal: For input features **x**, predict membership in one of *K* classes.

$$P(y \mid x) = \frac{\exp\{g_y(x)\}}{\sum_{k=1}^{K} \exp\{g_k(x)\}}$$

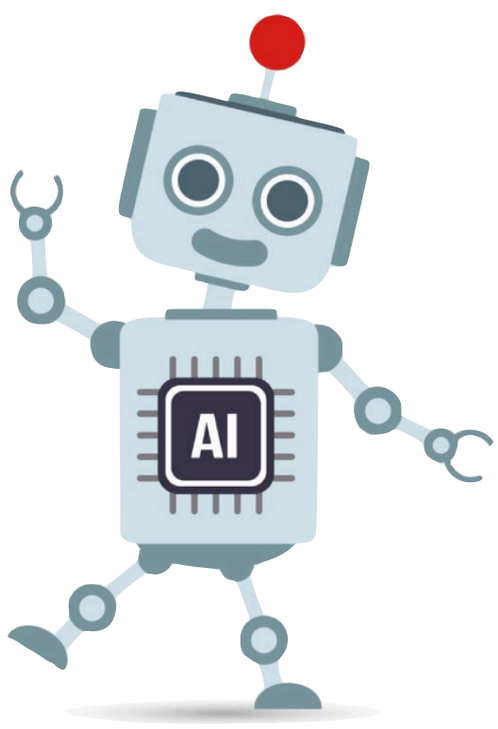


Classifier



Goal: For input features **x**,
predict membership in
one of *K* classes.

Training: Minimize with respect to θ…



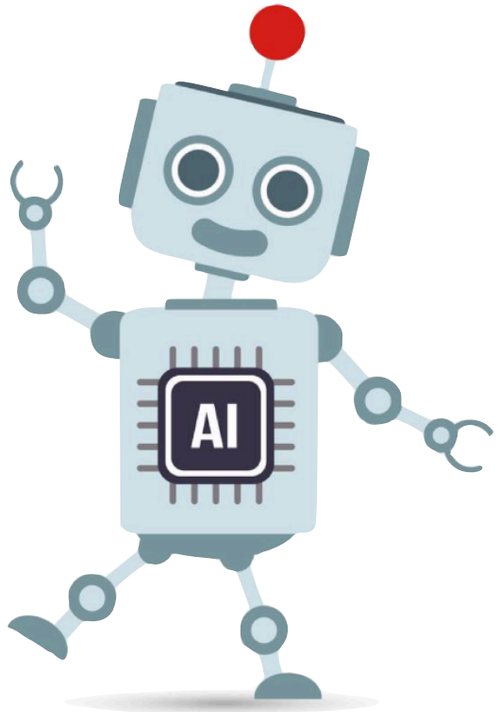
Classifier

$$\ell(\theta; \mathbf{X}, \mathbf{y}) = -\log \left\{ \prod_n \mathsf{P}(\mathsf{y}_n \mid \mathsf{x}_n) \right\}$$

$$= -\sum_n \log \frac{\exp\{\mathbf{g}_{y_n}(\mathsf{x}_n)\}}{\sum_{k=1}^{K} \exp\{\mathbf{g}_k(\mathsf{x}_n)\}}$$

Goal: For input features **x**,
predict membership in
one of *K* classes.

Training: Minimize with respect to θ…



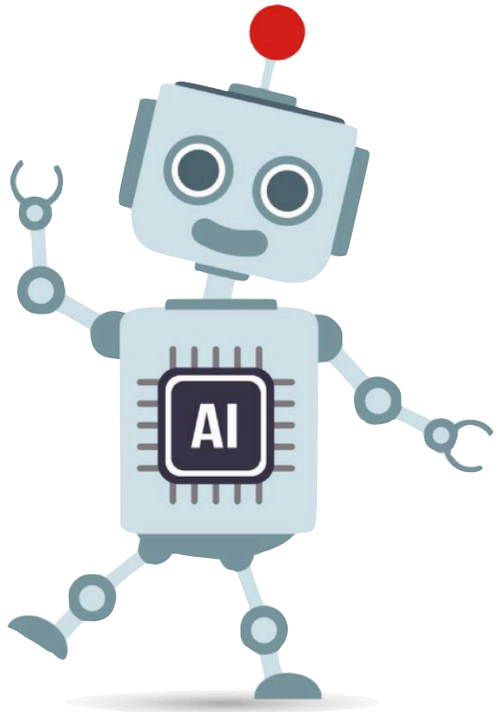
Classifier

$$\ell(\theta; \mathsf{X}, \mathsf{y}) = -\log \left\{ \prod_n \mathsf{P}(\mathsf{y}_n \mid \mathsf{x}_n) \right\}$$

$$= -\sum_n \log \frac{\exp\{\mathsf{g}_{y_n}(\mathsf{x}_n)\}}{\sum_{k=1}^{K} \exp\{\mathsf{g}_k(\mathsf{x}_n)\}}$$

Goal: For input features **x**,
predict membership in
one of *K* classes.

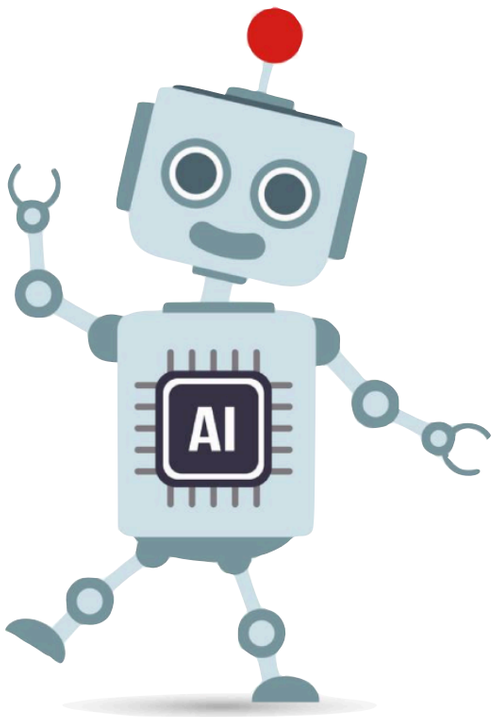
Training: Minimize with respect to θ…



Classifier

$$\ell(\theta; \mathsf{X}, \mathsf{y}) = -\log\left\{\prod_n \mathsf{P}(\mathsf{y}_n \mid \mathsf{x}_n)\right\}$$

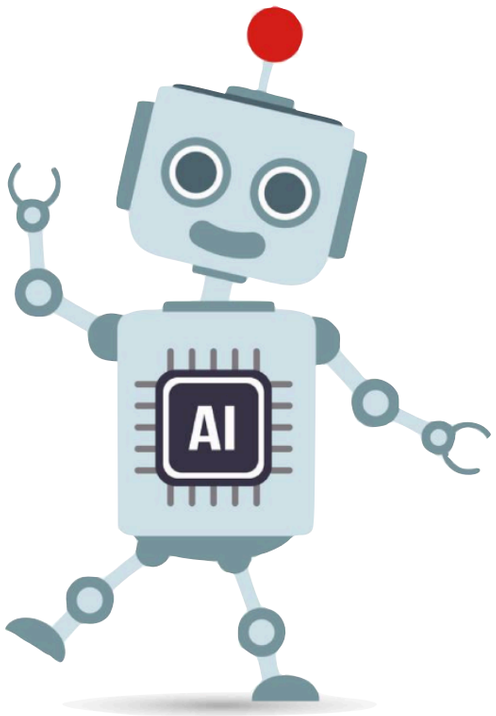
$$= -\sum_n \log \frac{\exp\{\mathbf{g}_{y_n}(\mathsf{x}_n)\}}{\sum_{k=1}^{K} \exp\{\mathbf{g}_k(\mathsf{x}_n)\}}$$



Classifier

?

Human Expert



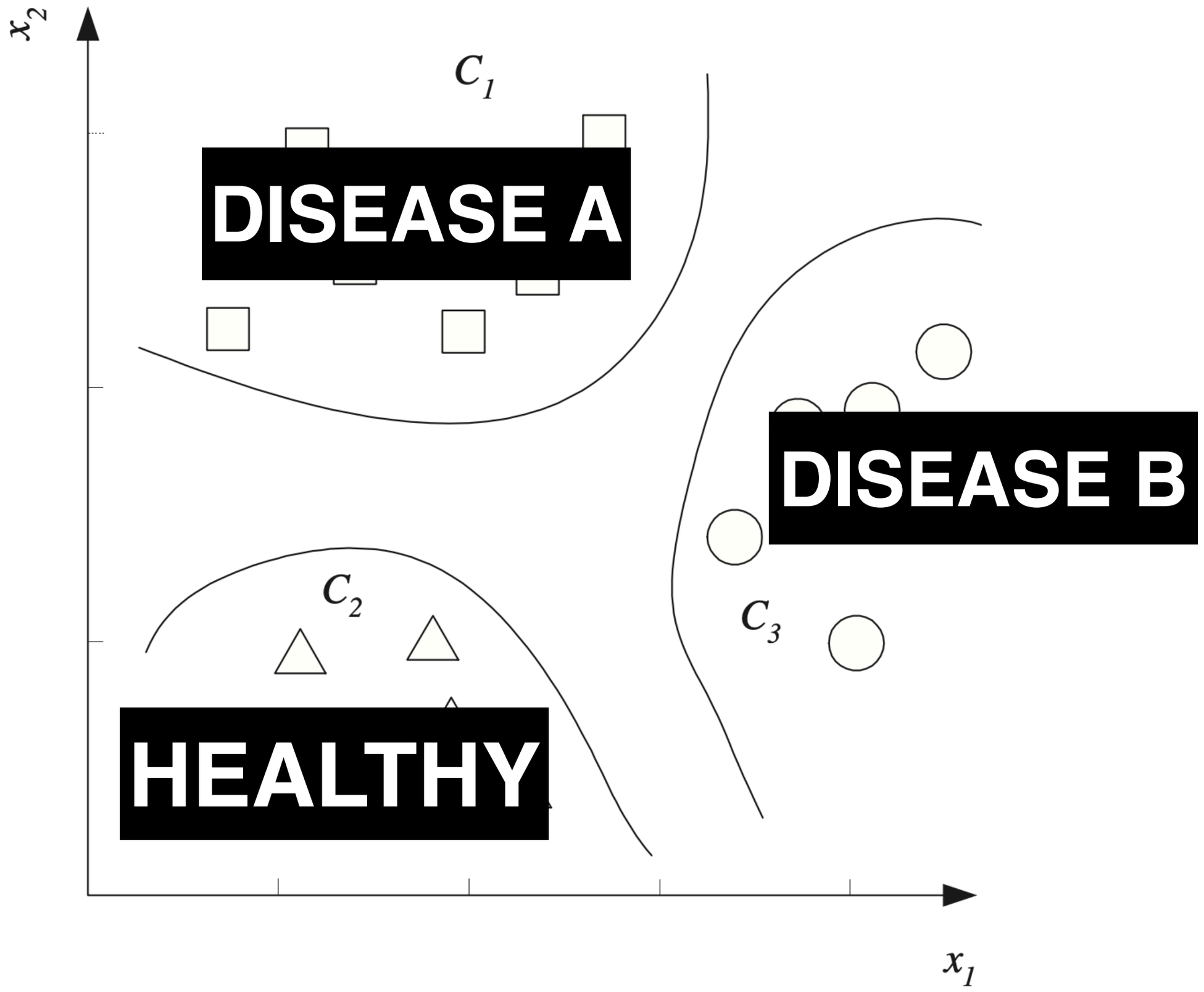
Classifier

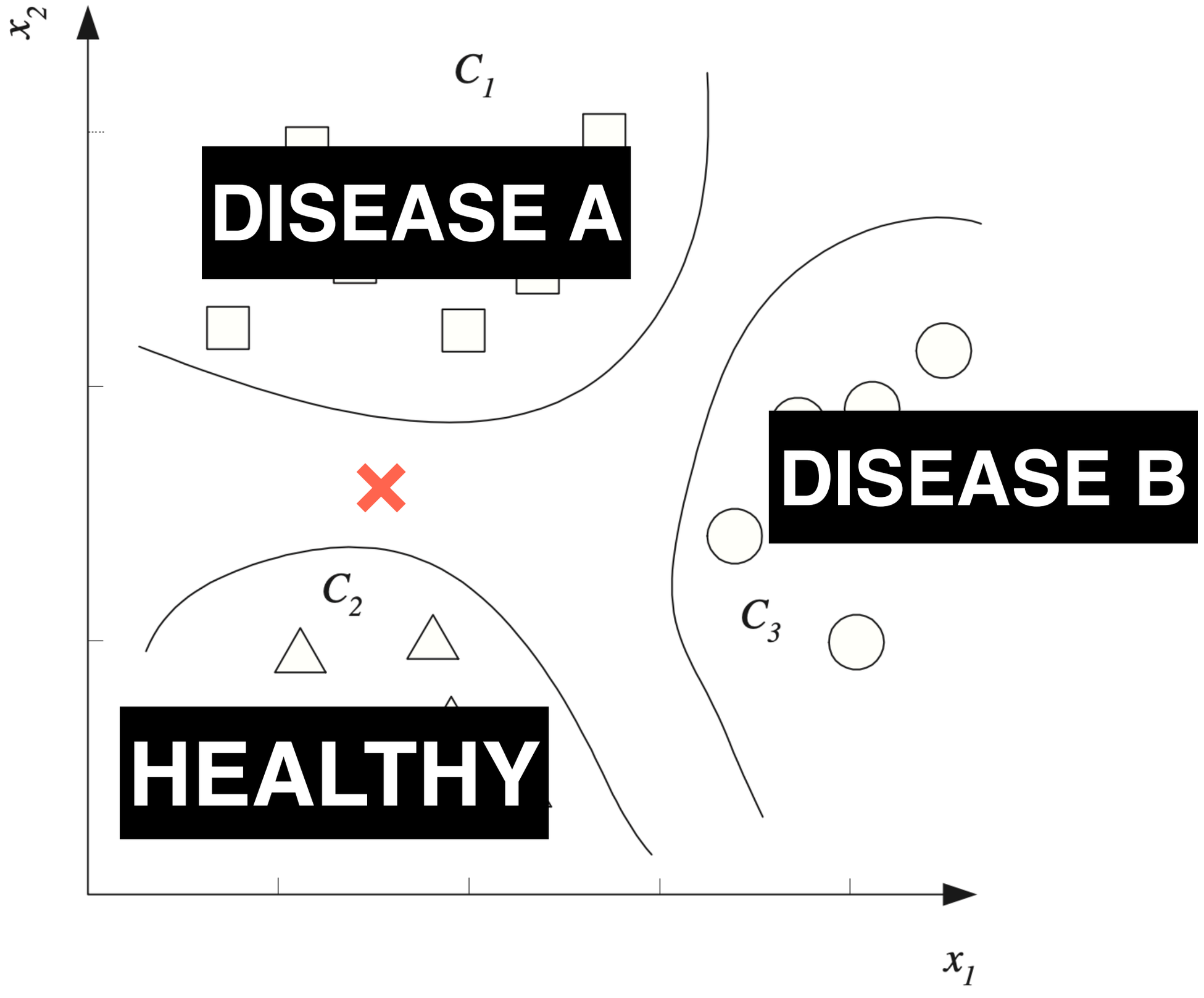
?

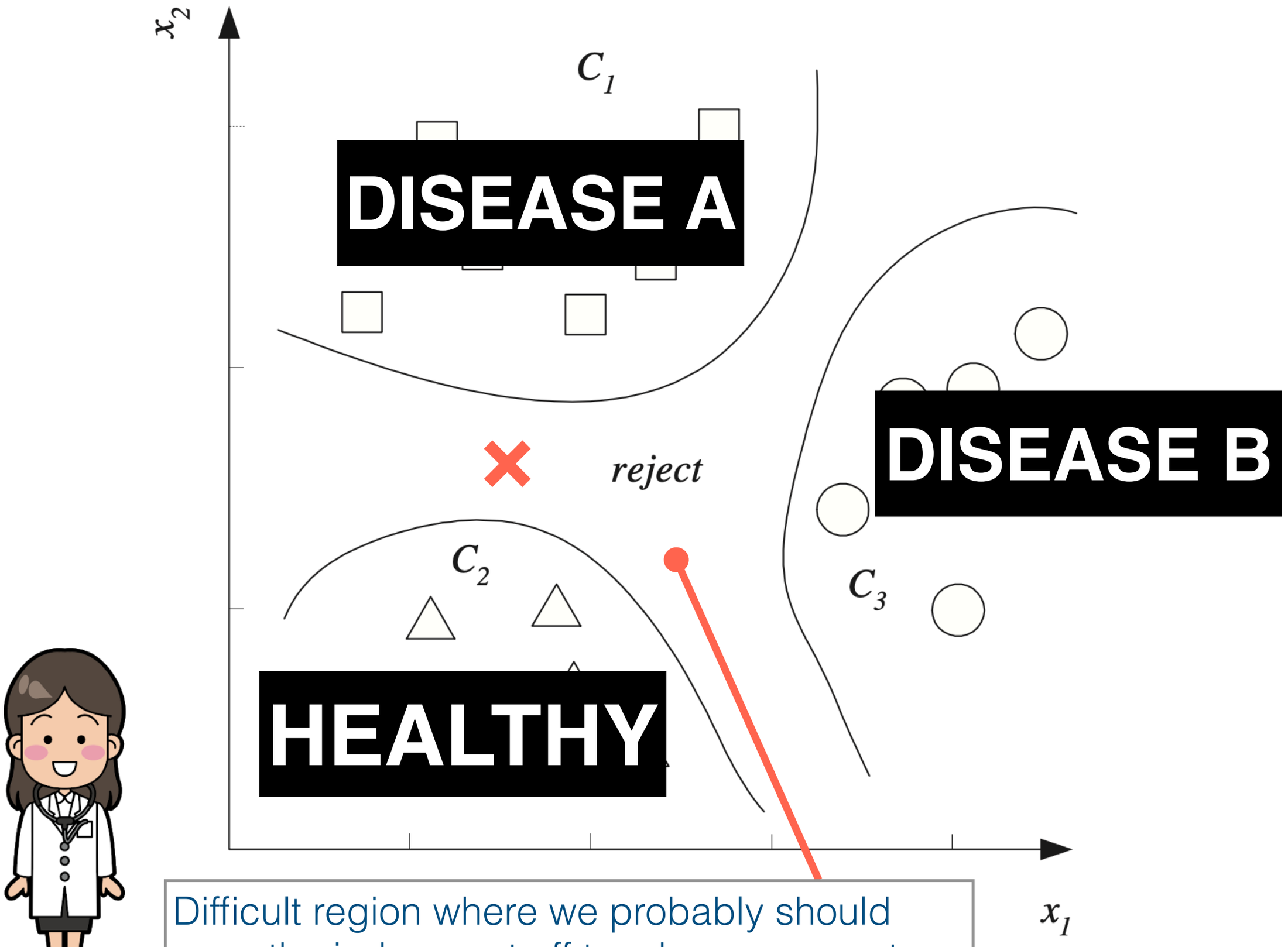
Human Expert

Warm Up

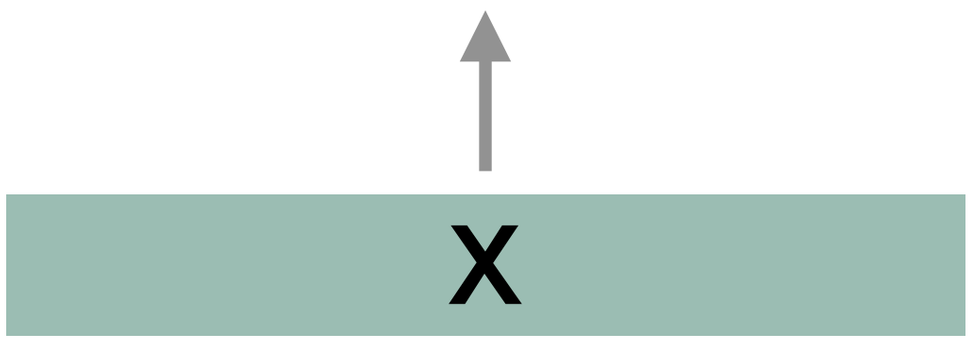
Classification with a Rejection Option

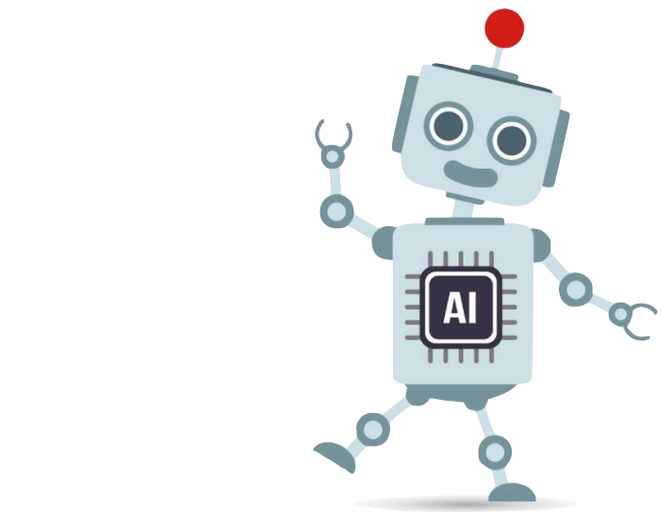

x_2
c_1
DISEASE A
DISEASE B
c_2
c_3
HEALTHY
x_1


x_2
c_1
DISEASE A
DISEASE B
c_2
c_3
HEALTHY
x_1



Difficult region where we probably should pass the judgement off to a human expert

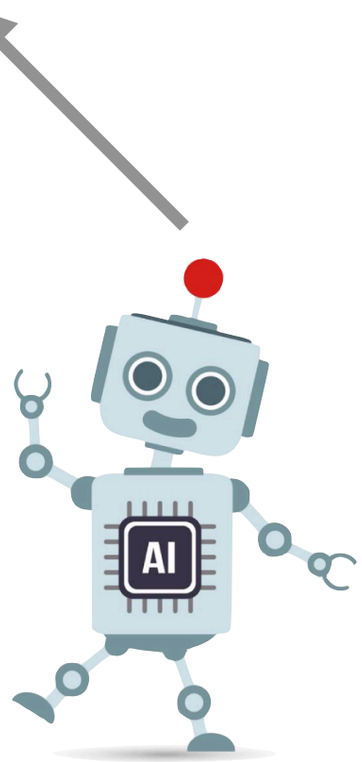

X



Classifier

X

make
prediction

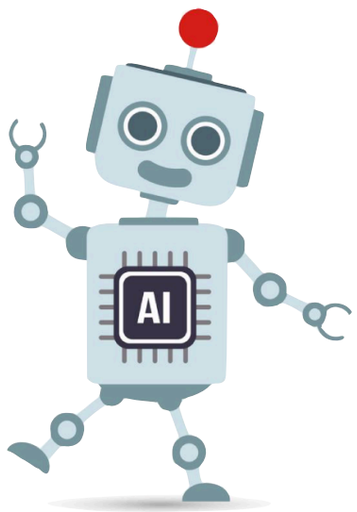


Classifier

X

make prediction

abstain

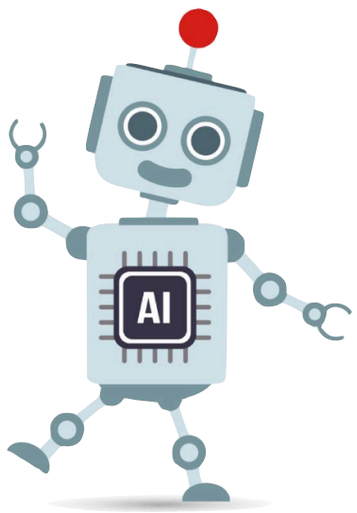


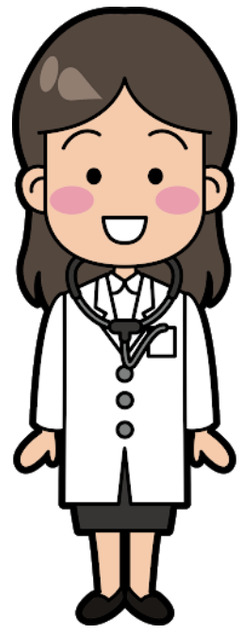
Classifier

X

make prediction

abstain



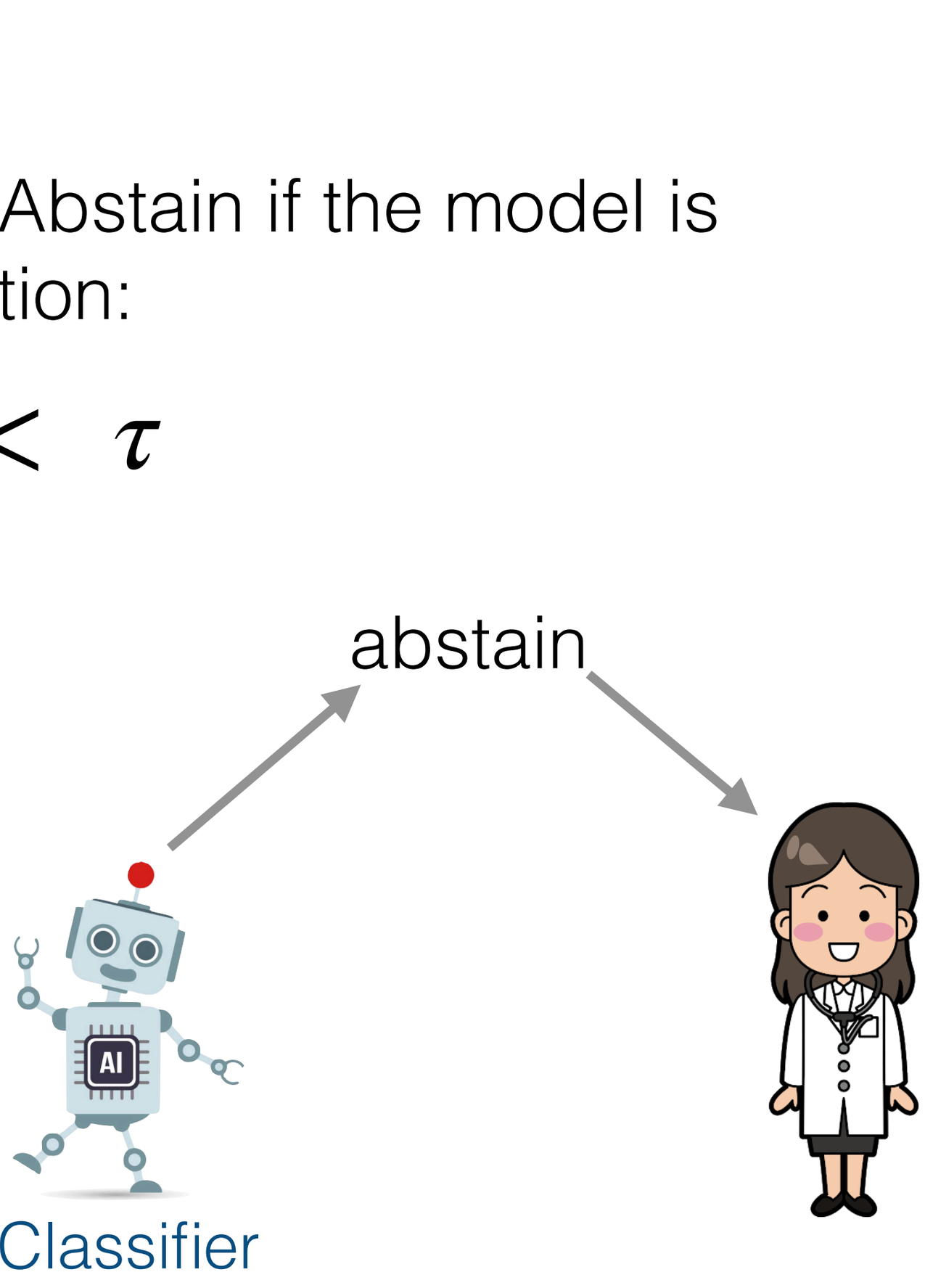


Classifier

X

Score-Based Rejection: Abstain if the model is unconfident in its prediction:

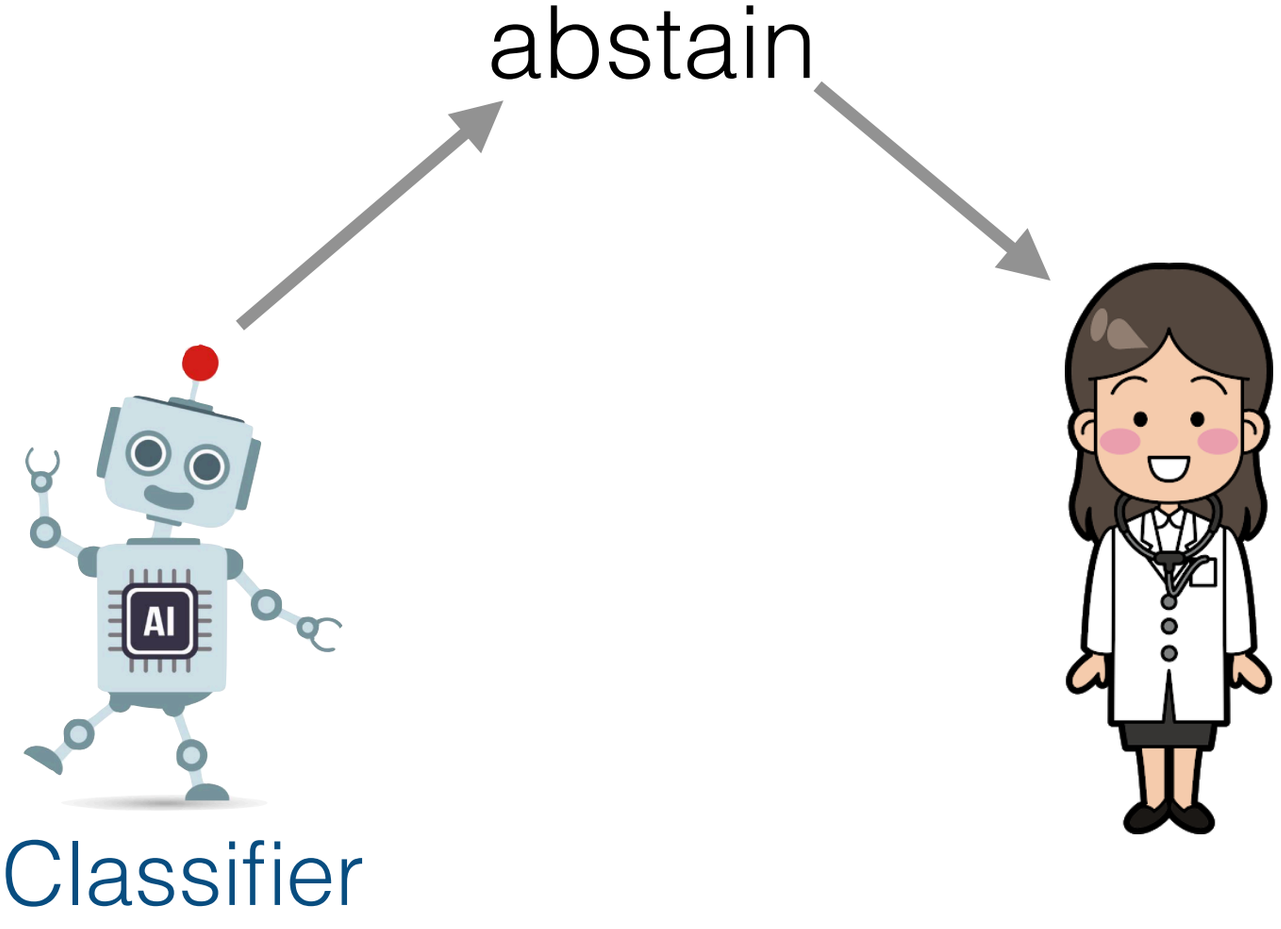
$$\max_{y} \; \mathrm{P}(\mathrm{y} \mid \mathrm{x}_n) \; < \; \tau$$



Score-Based Rejection: Abstain if the model is unconfident in its prediction:

$$\max_{y} \; \mathsf{P}(\mathsf{y} \mid \mathsf{x}_n) \; < \; \tau$$

Human behavior is not modeled!



Challenge: how can we model the human?

If they are a true expert, modeling their decision making— $\mathbb{P}_h(\mathbf{y} \mid \mathbf{x})$—is assumed to be impossible.

Learning to Defer to an Expert

[Madras et al., NeurIPS 2018]

Better Formulation

Model what the human knows,
so we can enable *collaboration*

Better Formulation

Model what the human knows,
so we can enable *collaboration*

Data: $\mathfrak{D} = \left\{ \mathsf{x}_n, \mathsf{y}_n, \mathsf{m}_n \right\}_{n=1}^{N}$

expert predictions

Better Formulation

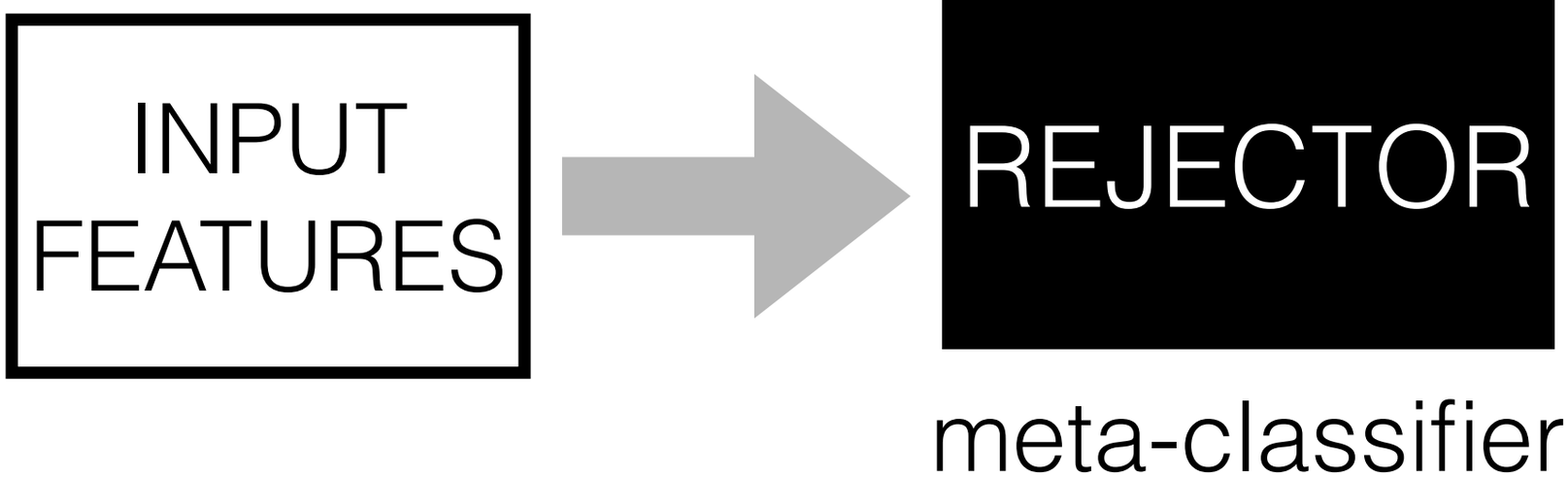
Model what the human knows,
so we can enable *collaboration*

Data: $\mathfrak{D} = \left\{ \mathbf{x}_n, \mathbf{y}_n, \mathbf{m}_n \right\}_{n=1}^{N}$

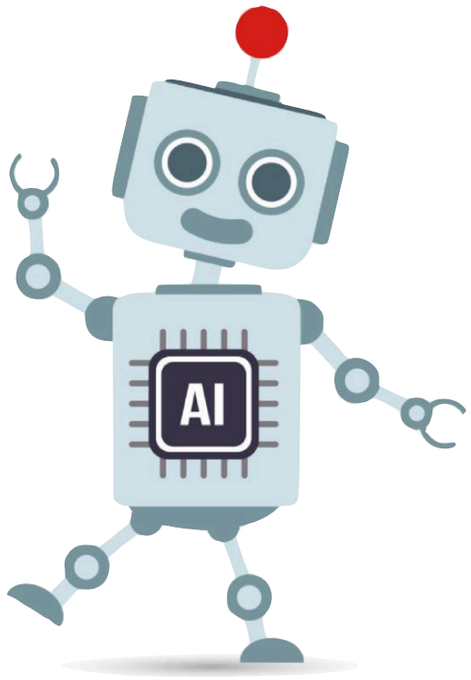
expert predictions

Models: $\mathbf{r}(\mathbf{x})$ Rejector $\quad$ $\mathbf{h}(\mathbf{x})$ Classifier

Learning to Defer



Learning to Defer



Classifier

INPUT FEATURES

REJECTOR

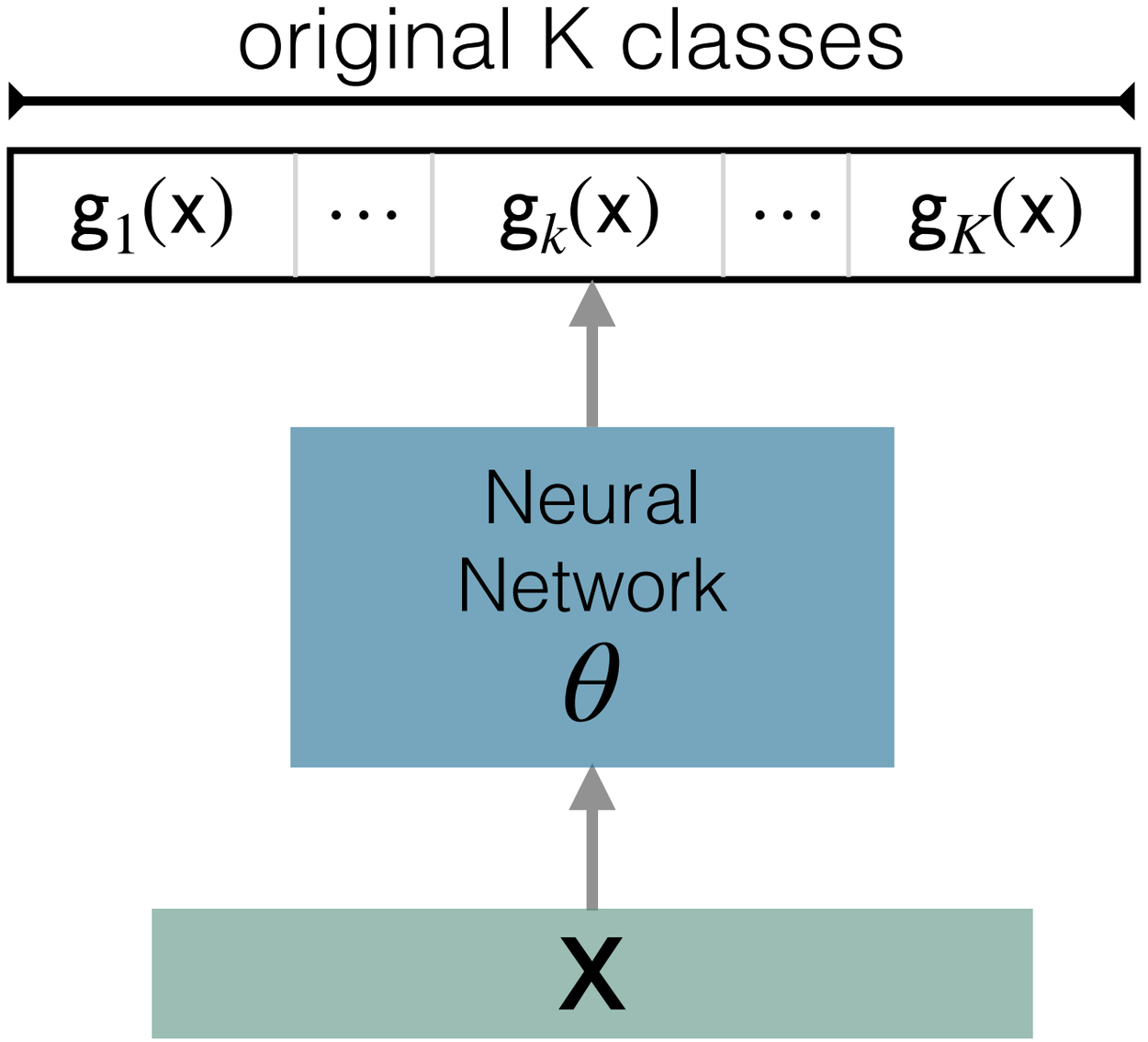
meta-classifier

?

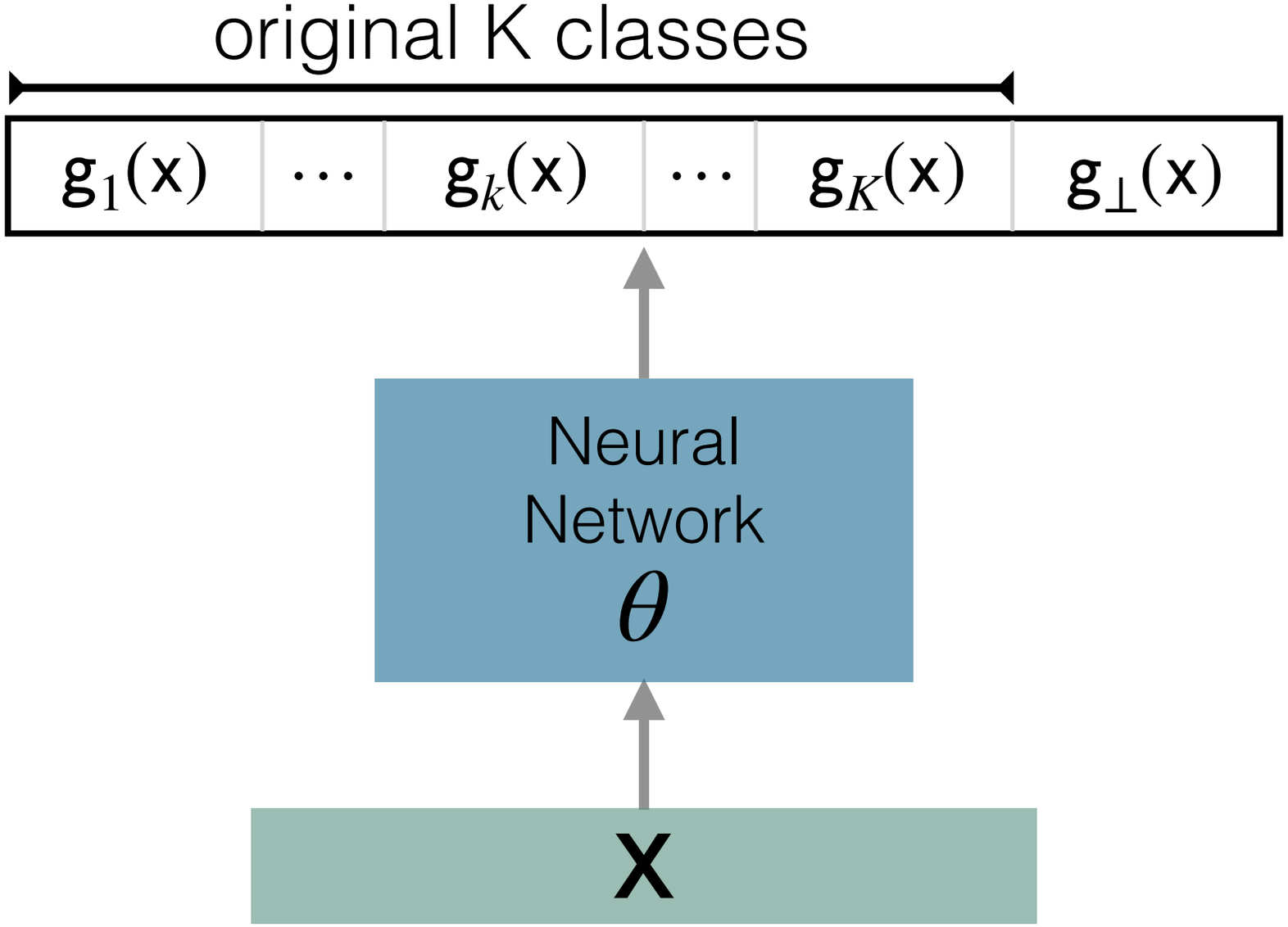
Expert

Softmax Approach [Mozannar & Sontag, ICML 2020]

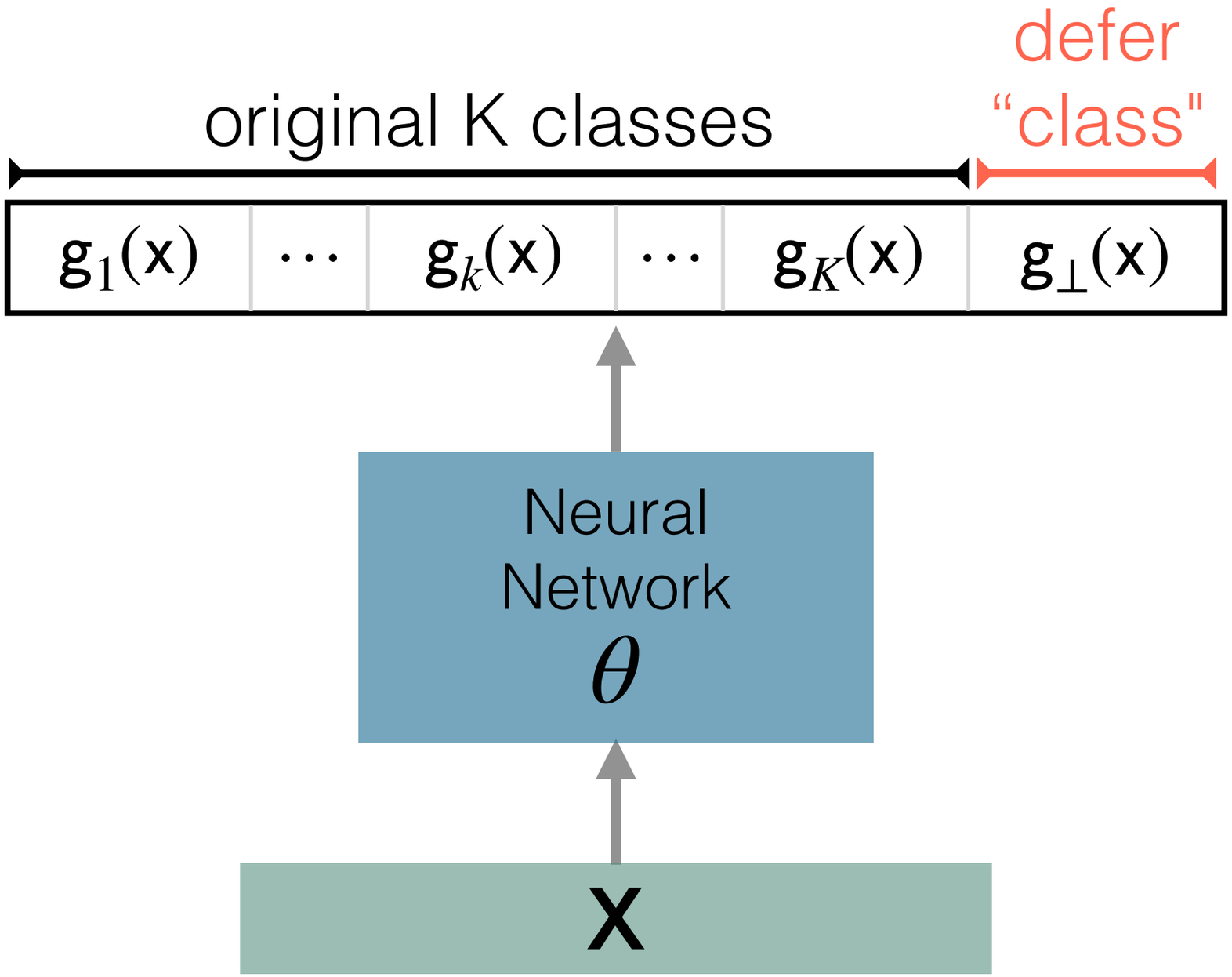
Softmax Approach [Mozannar & Sontag, ICML 2020]



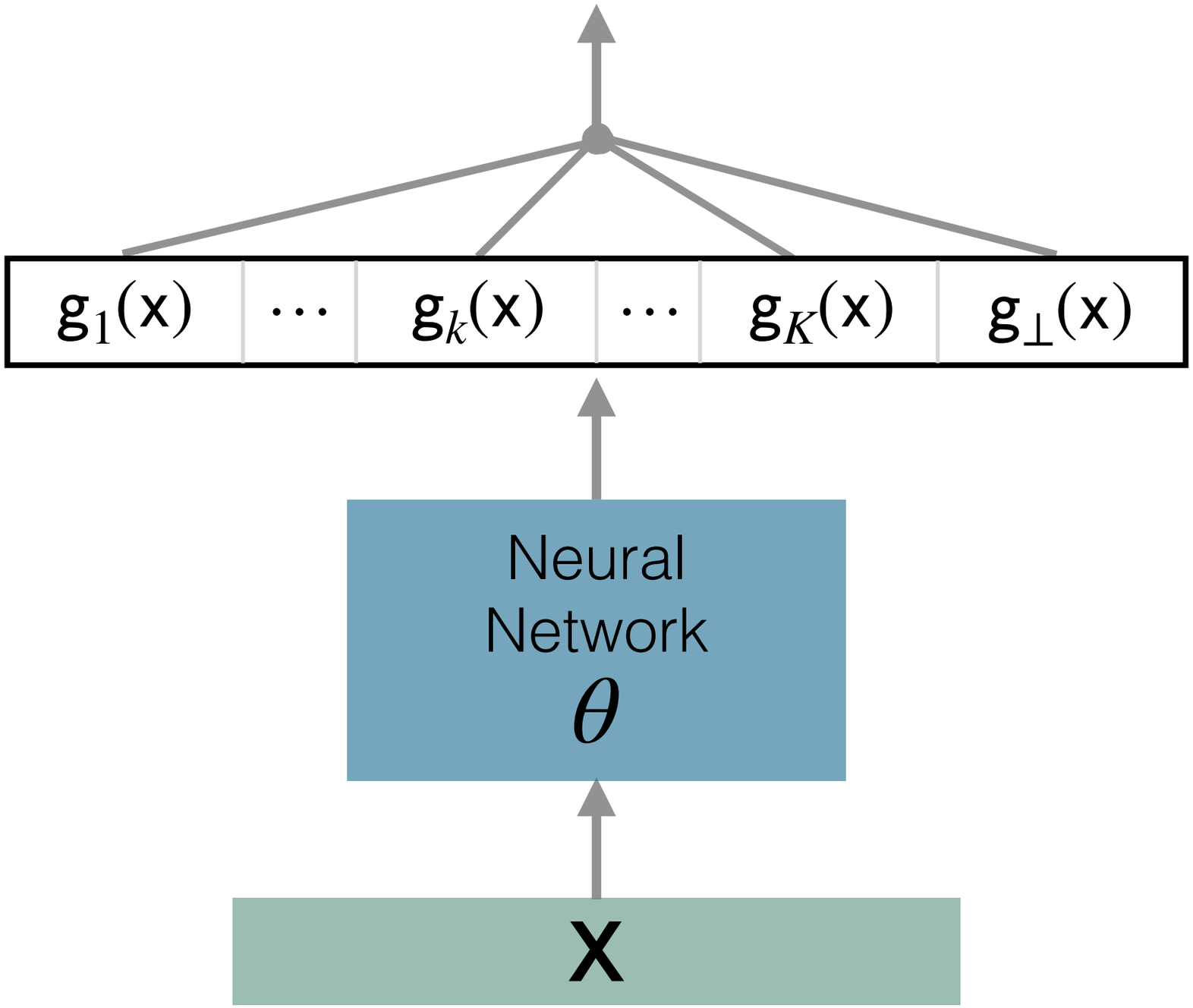
Softmax Approach [Mozannar & Sontag, ICML 2020]



Softmax Approach [Mozannar & Sontag, ICML 2020]

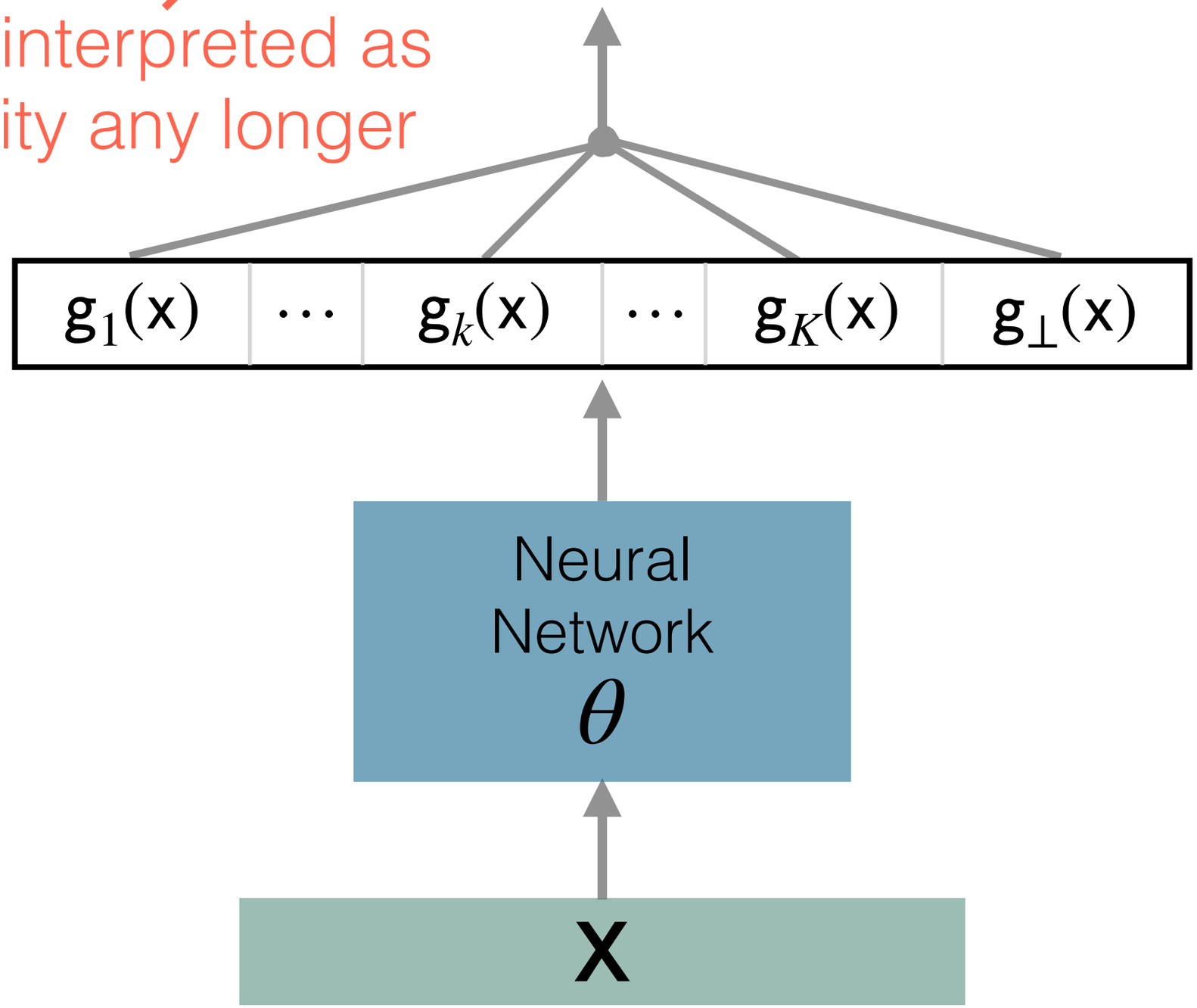


$$p_i(x) = \frac{\exp\{g_i(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$

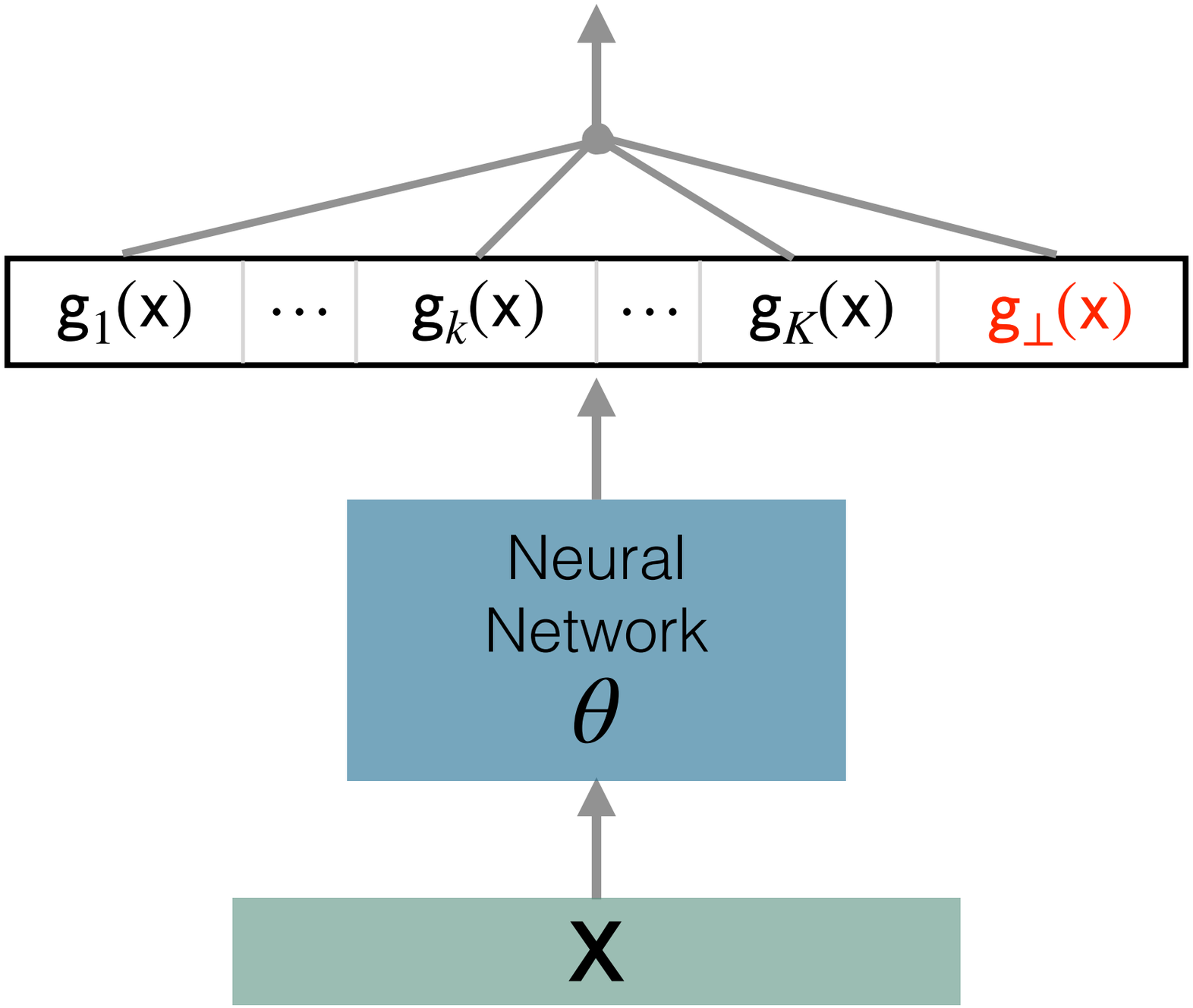


$$p_i(x) = \frac{\exp\{g_i(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$

cannot be interpreted as a probability any longer



$$p_{\perp}(x) = \frac{\exp\{g_{\perp}(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$



[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_{n} \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \;\log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_{n} \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \; + \; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \; \log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_{n} \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \;\log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_n \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \; \log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

only if expert is correct

[Mozannar & Sontag, ICML 2020]

$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_n \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \mathbb{I}[\mathsf{y}_n = \mathsf{m}_n] \; \log \mathsf{p}_{\perp}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

only if expert is correct

Consistency: The minimizers (w.r.t. *g*) correspond to the Bayes optimal classifier and rejector.

Consistency

Classifier and rejector have the following Bayes optimal solutions (under 0-1 loss):

Consistency

Classifier and rejector have the following Bayes optimal solutions (under 0-1 loss):

$$\mathsf{h}^*(\mathbf{x}) = \operatorname*{argmax}_{y} \ \mathbb{P}(\mathbf{y} = y \,|\, \mathbf{x})$$

Consistency

Classifier and rejector have the following Bayes optimal solutions (under 0-1 loss):

$$\mathsf{h}^*(\mathsf{x}) = \operatorname*{argmax}_y \ \mathbb{P}(\mathsf{y} = y \,|\, \mathsf{x})$$

$$\mathsf{r}^*(\mathsf{x}) = \mathbb{I}\left[\mathbb{P}(\mathsf{m} = \mathsf{y} \,|\, \mathsf{x}) \geq \max_y \ \mathbb{P}(\mathsf{y} = y \,|\, \mathsf{x})\right]$$

probability that the expert is correct

Our work

Is this learning-to-defer system calibrated?

Calibration: Is the system a good forecaster?

Calibration: Is the system a good forecaster?

$$\mathbb{P}(\mathbf{y} \mid \mathbf{x})$$

Does the classifier correctly estimate the underlying class probabilities?

Calibration: Is the system a good forecaster?

$$\mathbb{P}(y \mid x)$$

Does the classifier correctly estimate the underlying class probabilities?

$$\mathbb{P}(m = y \mid x)$$

Does the rejector correctly estimate the expert's chance of being correct?

Calibration: Is the system a good forecaster?

$$\mathbb{P}(y \mid x)$$

Does the classifier correctly estimate the underlying class probabilities?

$$\mathbb{P}(m = y \mid x)$$

Does the rejector correctly estimate the expert's chance of being correct?

Calibration: Is the system a good forecaster?

$$\mathbb{P}(y \mid x)$$

Does the classifier correctly estimate the underlying class probabilities?

$$\mathbb{P}(m = y \mid x)$$

Does the rejector correctly estimate the expert's chance of being correct?

Calibration: Is the system a good forecaster?

$$\mathbb{P}(y \mid x)$$

Does the classifier correctly estimate the underlying class probabilities?

$$\mathbb{P}(m = y \mid x)$$

Does the rejector correctly estimate the expert's chance of being correct?

nature medicine

LETTERS

https://doi.org/10.1038/s41591-020-0942-0

Check for updates

Human–computer collaboration for skin cancer recognition

Philipp Tschandl[1,17], Christoph Rinner[2,17], Zoe Apalla[3], Giuseppe Argenziano[4], Noel Codella[5], Allan Halpern[6], Monika Janda[7], Aimilios Lallas[3], Caterina Longo[8,9], Josep Malvehy[10,11], John Paoli[12,13], Susana Puig[10,11], Cliff Rosendahl[14], H. Peter Soyer[15], Iris Zalaudek[16] and Harald Kittler[1]

The rapid increase in telemedicine coupled with recent advances in diagnostic artificial intelligence (AI) create the imperative to consider the opportunities and risks of inserting AI-based support into new paradigms of care. Here we

competitive view of AI is evolving based on studies suggesting that a more promising approach is human–AI cooperation[10–15]. The role of human–computer collaboration in health-care delivery, the appropriate settings in which it can be applied and its impact on the

"The least experienced [physicians] tended to accept AI-based support that contradicted their initial diagnosis even if they were confident."

Recall…

$$p_\perp(x) = \frac{\exp\{g_\perp(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$

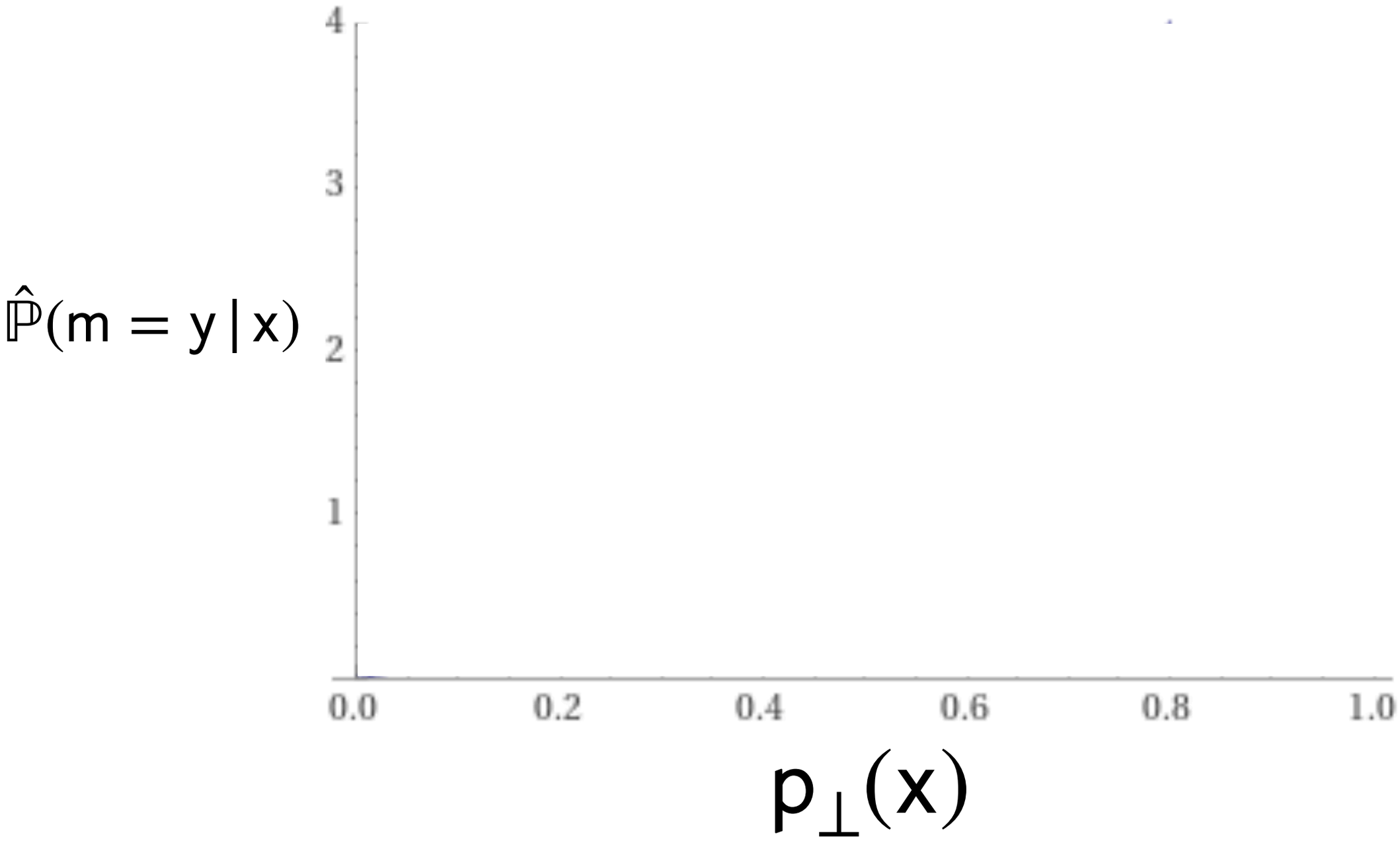
Recall…

$$p_{\perp}(x) = \frac{\exp\{g_{\perp}(x)\}}{\sum_{k=1}^{K+1} \exp\{g_k(x)\}}$$

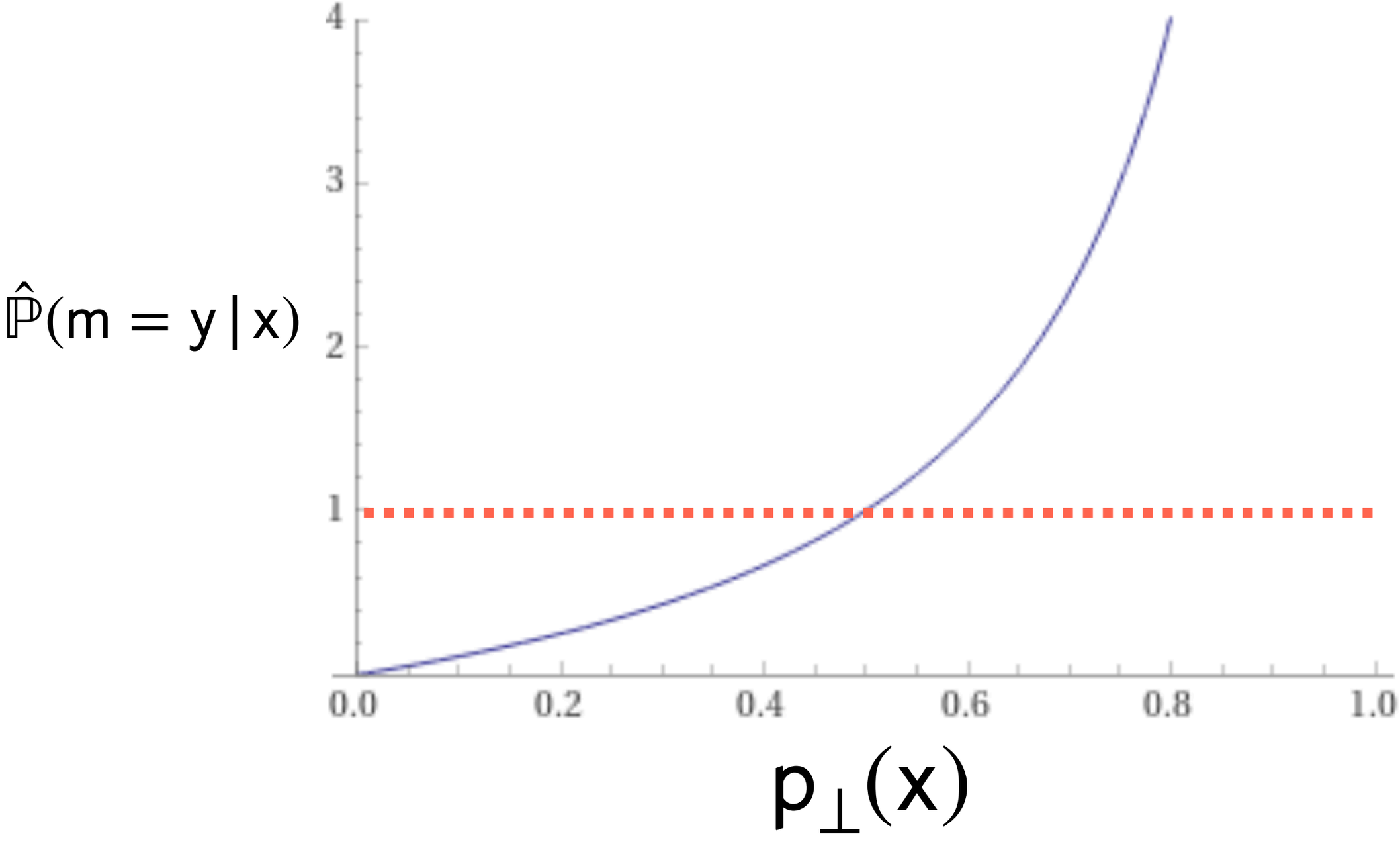
From Mozannar's & Sontag's Theorem #1…

$$\mathbb{P}(m = y \mid x) \;=\; \frac{p^*_{\perp}(x)}{1 - p^*_{\perp}(x)}$$

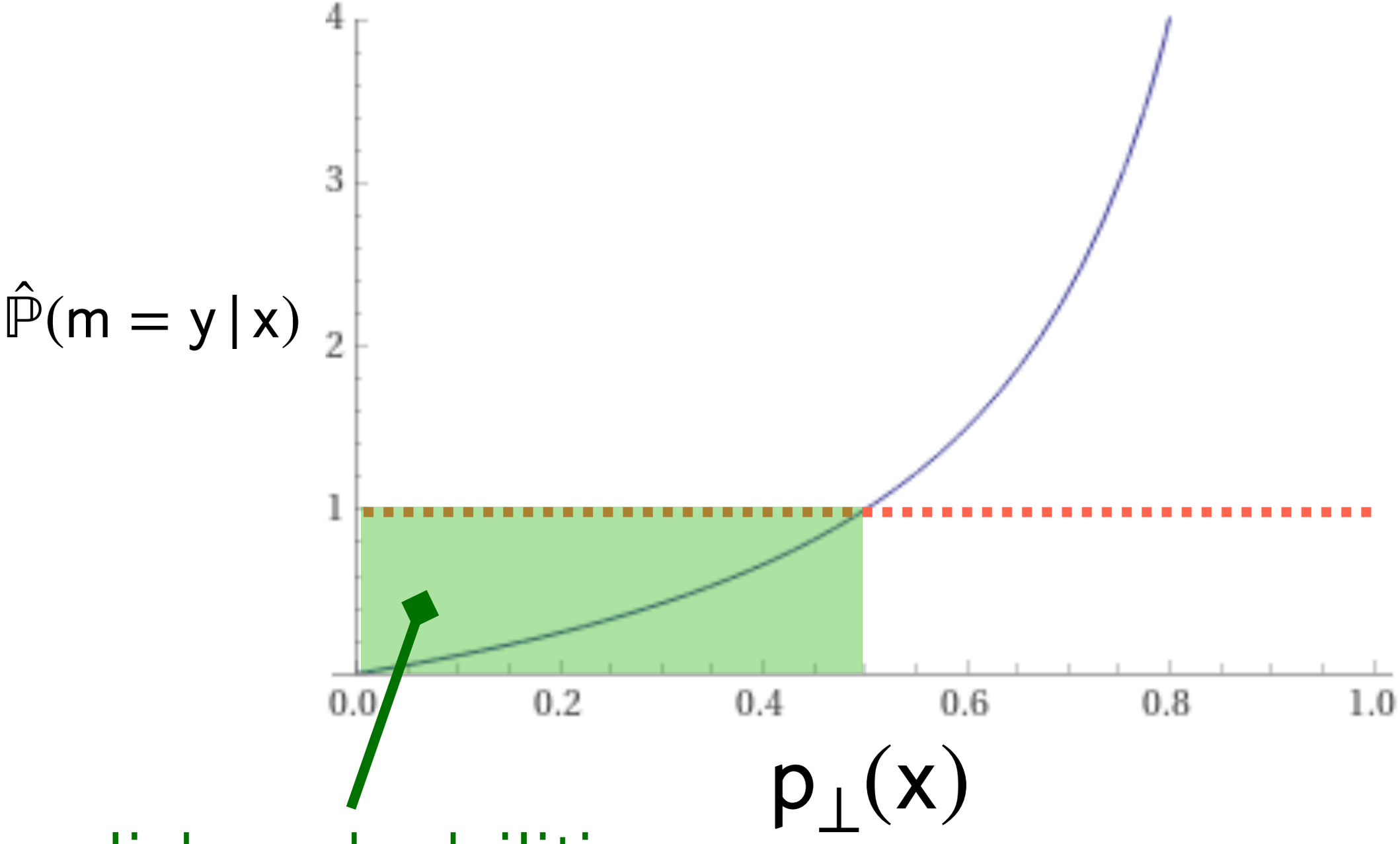
Estimating $\mathbb{P}(m = y \mid x)$



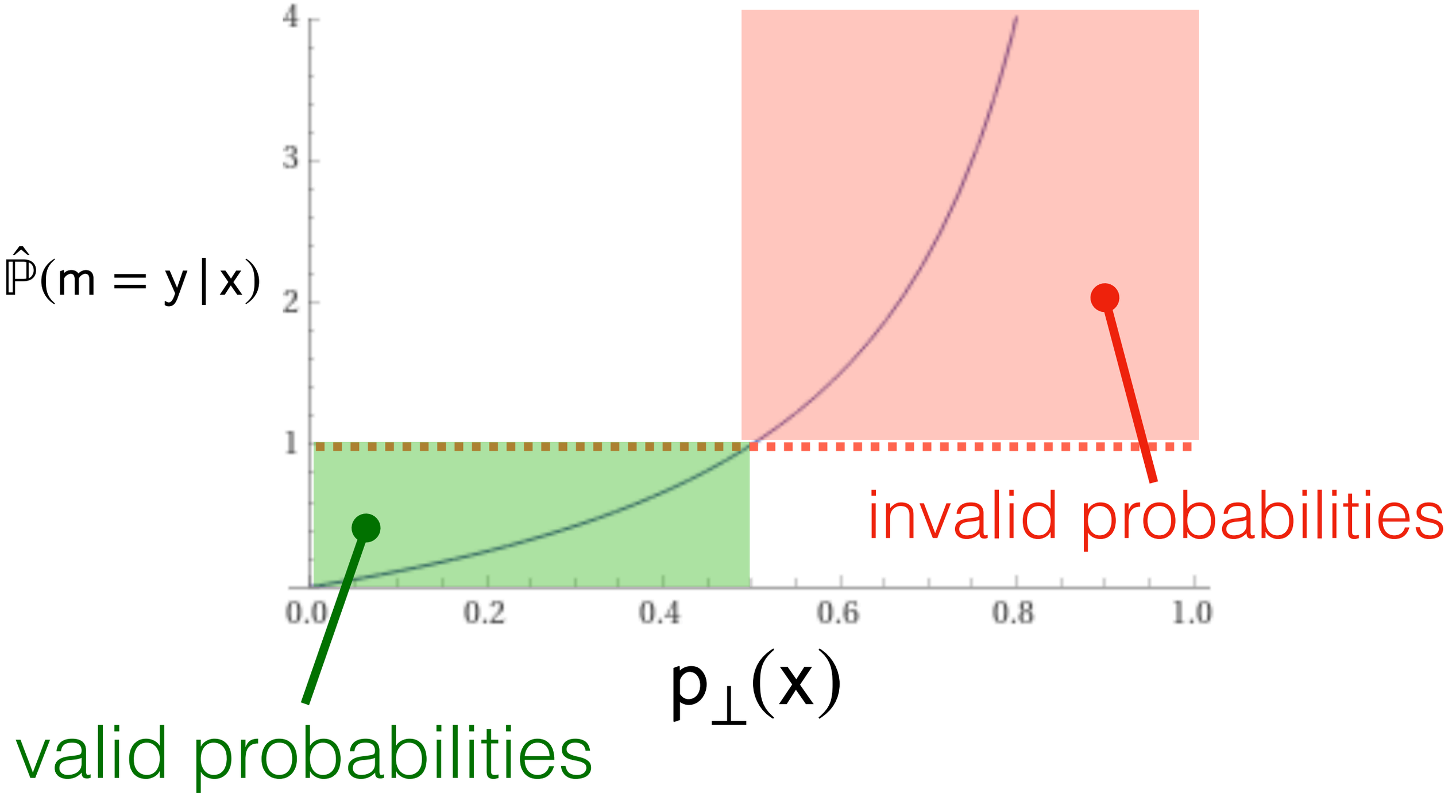
Estimating $\mathbb{P}(m = y \mid x)$



Estimating $\mathbb{P}(m = y \mid x)$



Estimating $\mathbb{P}(m = y \mid x)$

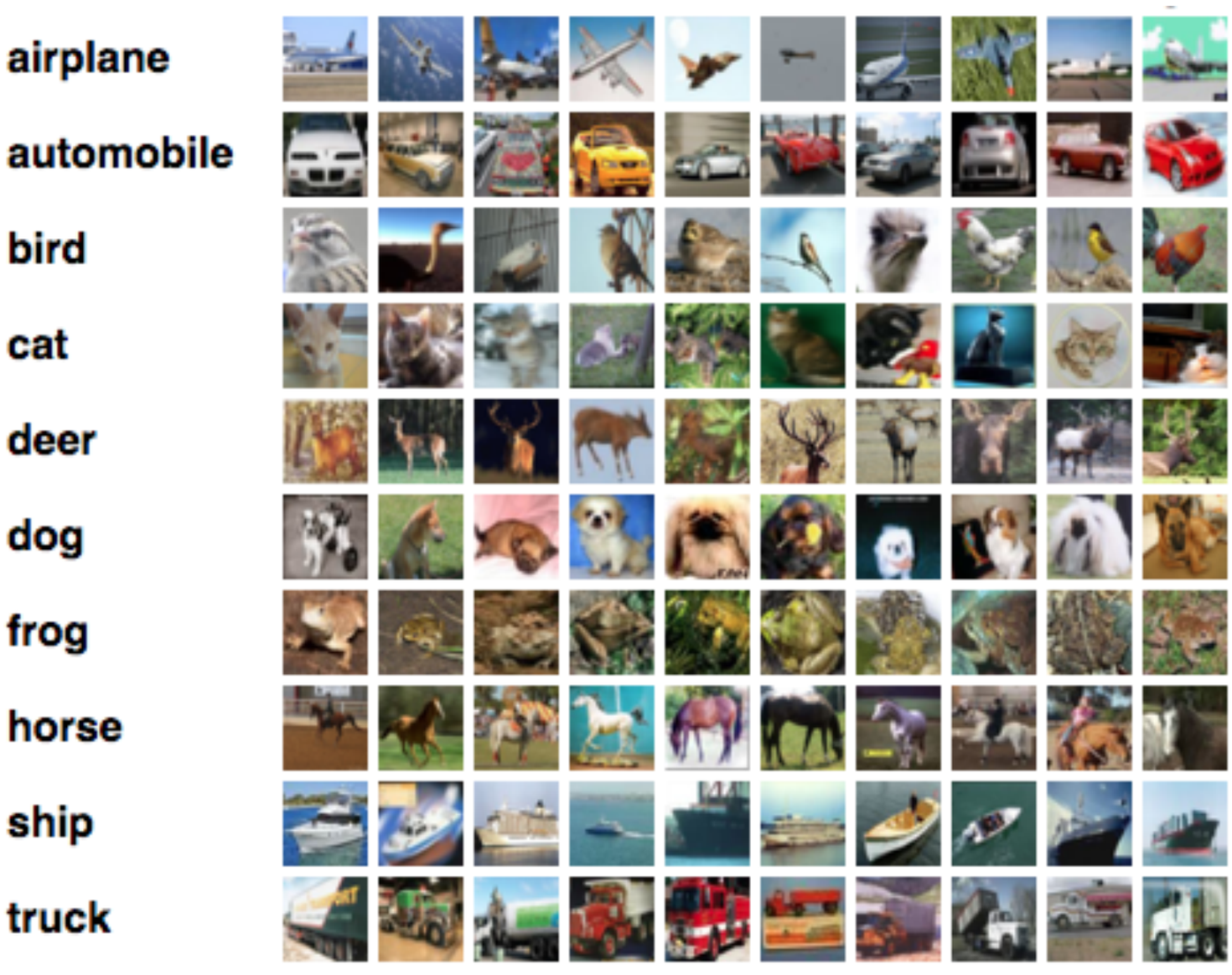


What happens in practice?

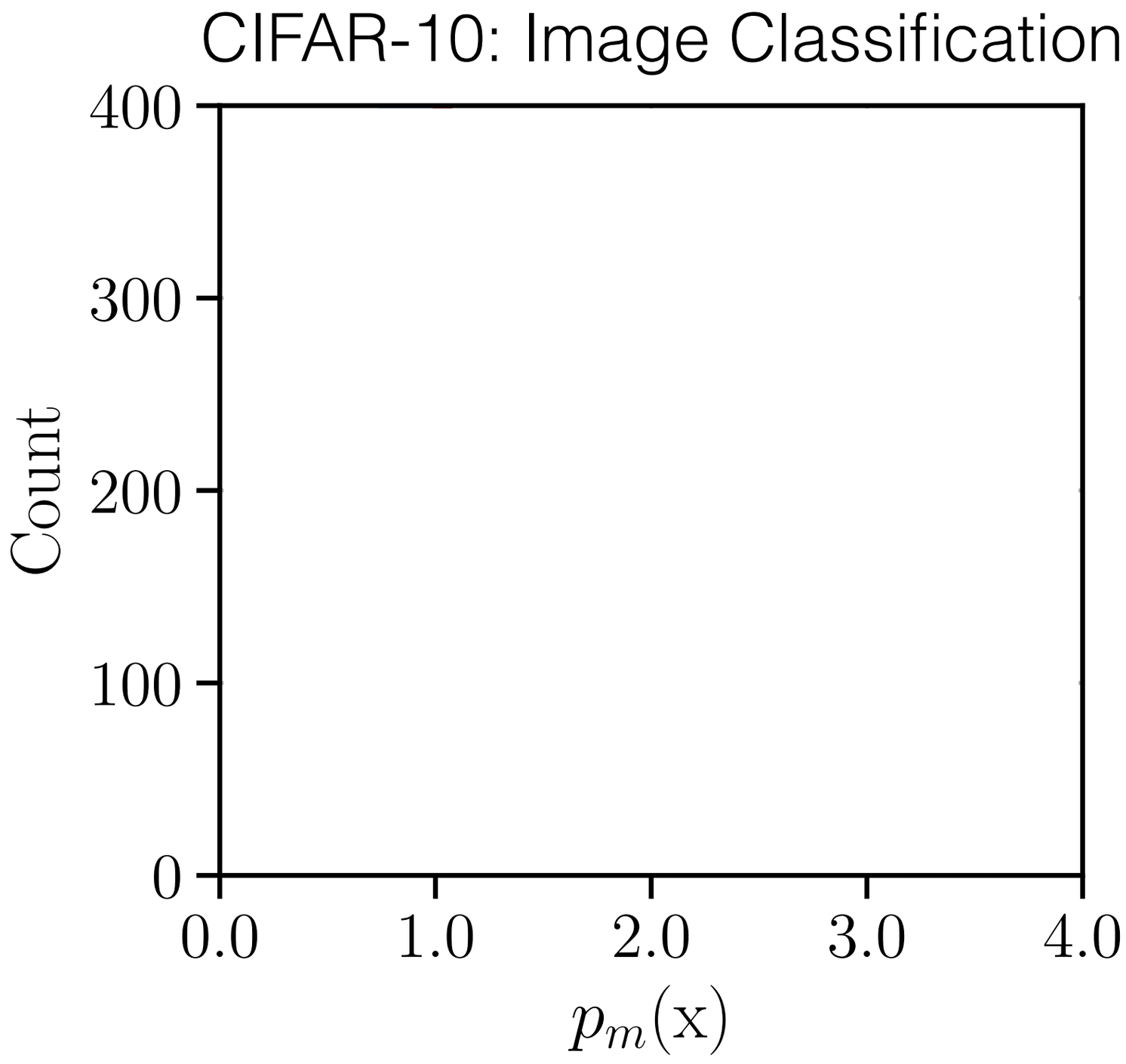
CIFAR-10: Image Classification

What happens in practice?

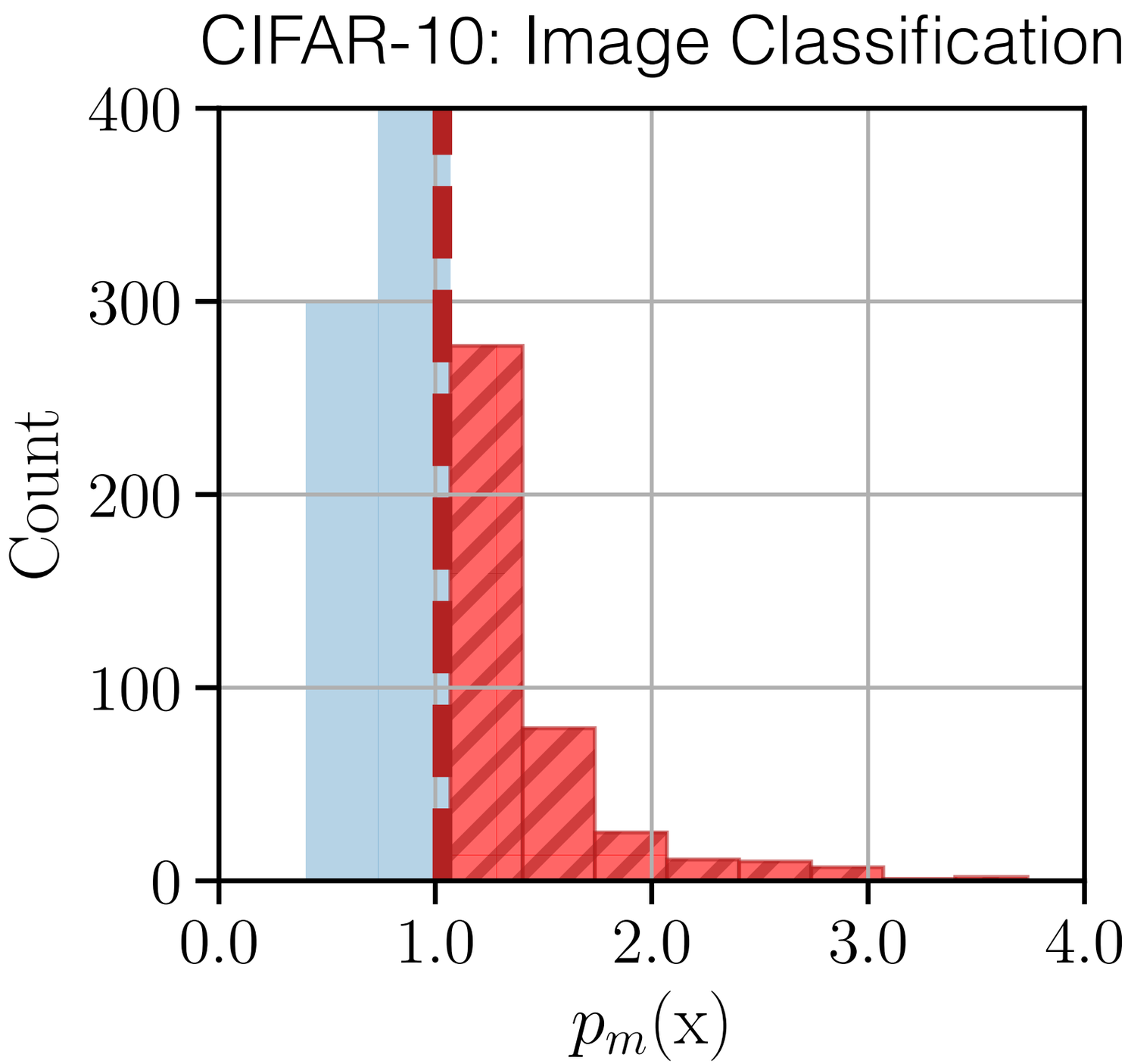
CIFAR-10: Image Classification



What happens in practice?



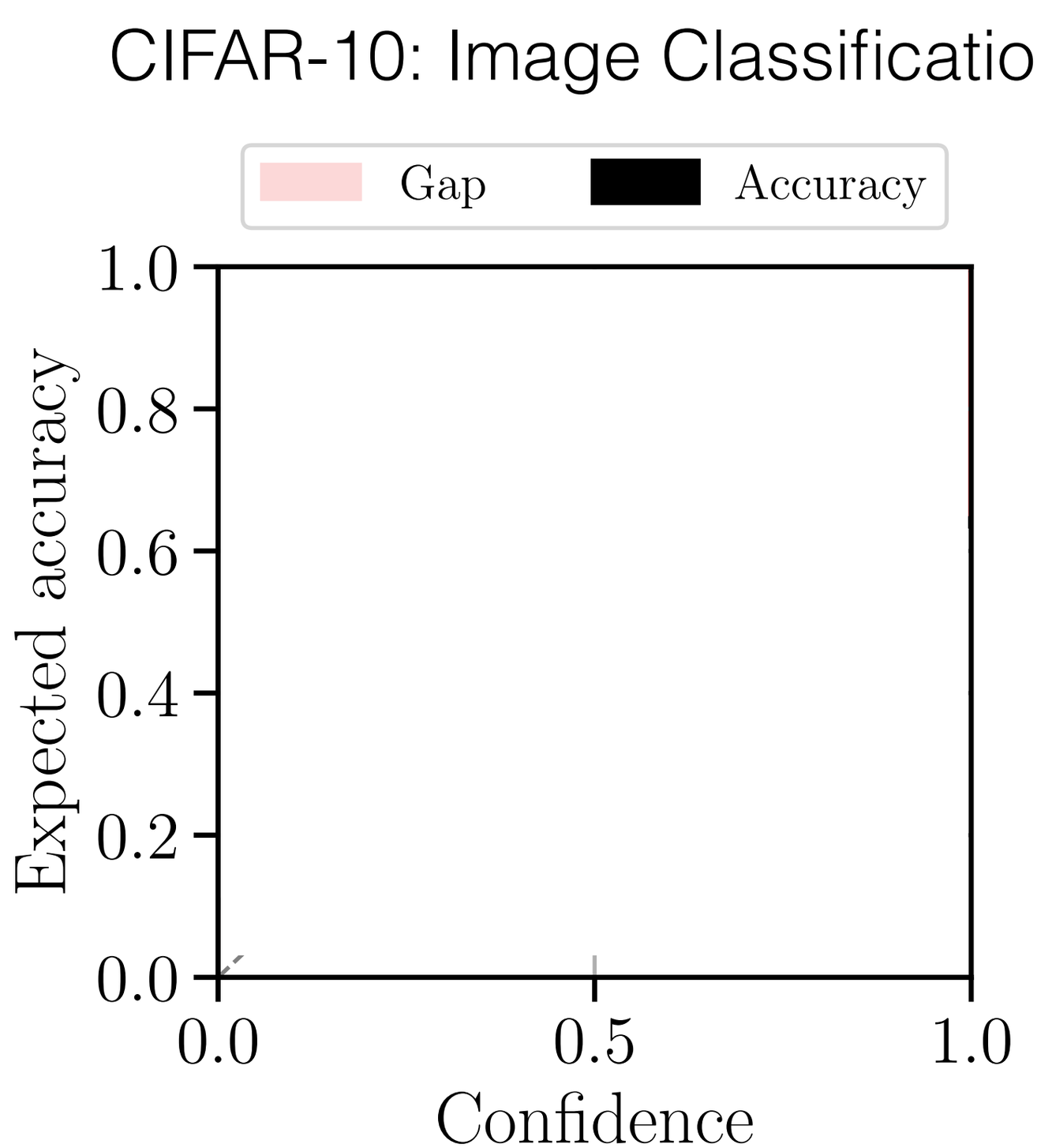
What happens in practice?



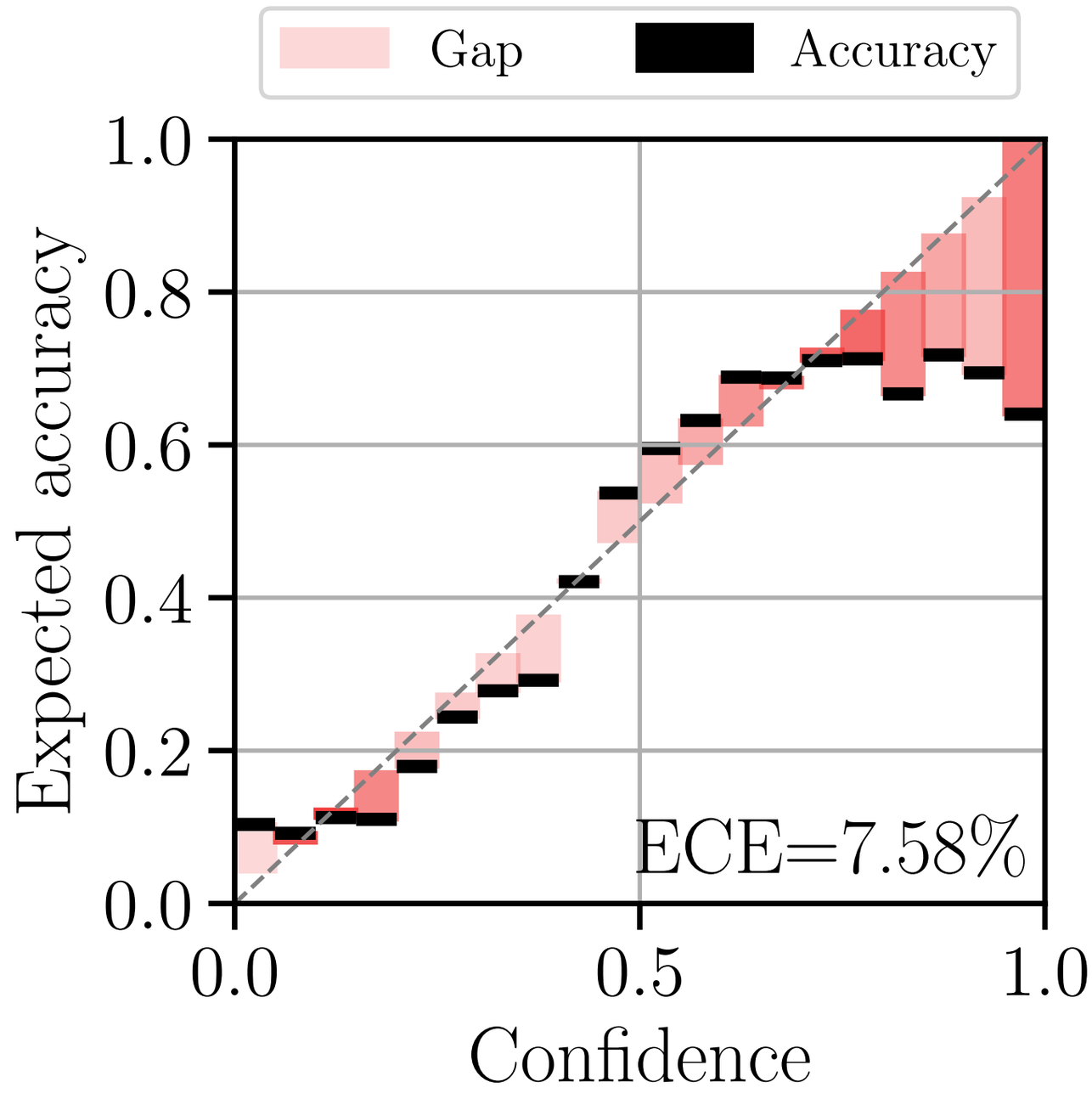
Truncation: $p_m(x) \in (0,1]$

CIFAR-10: Image Classification

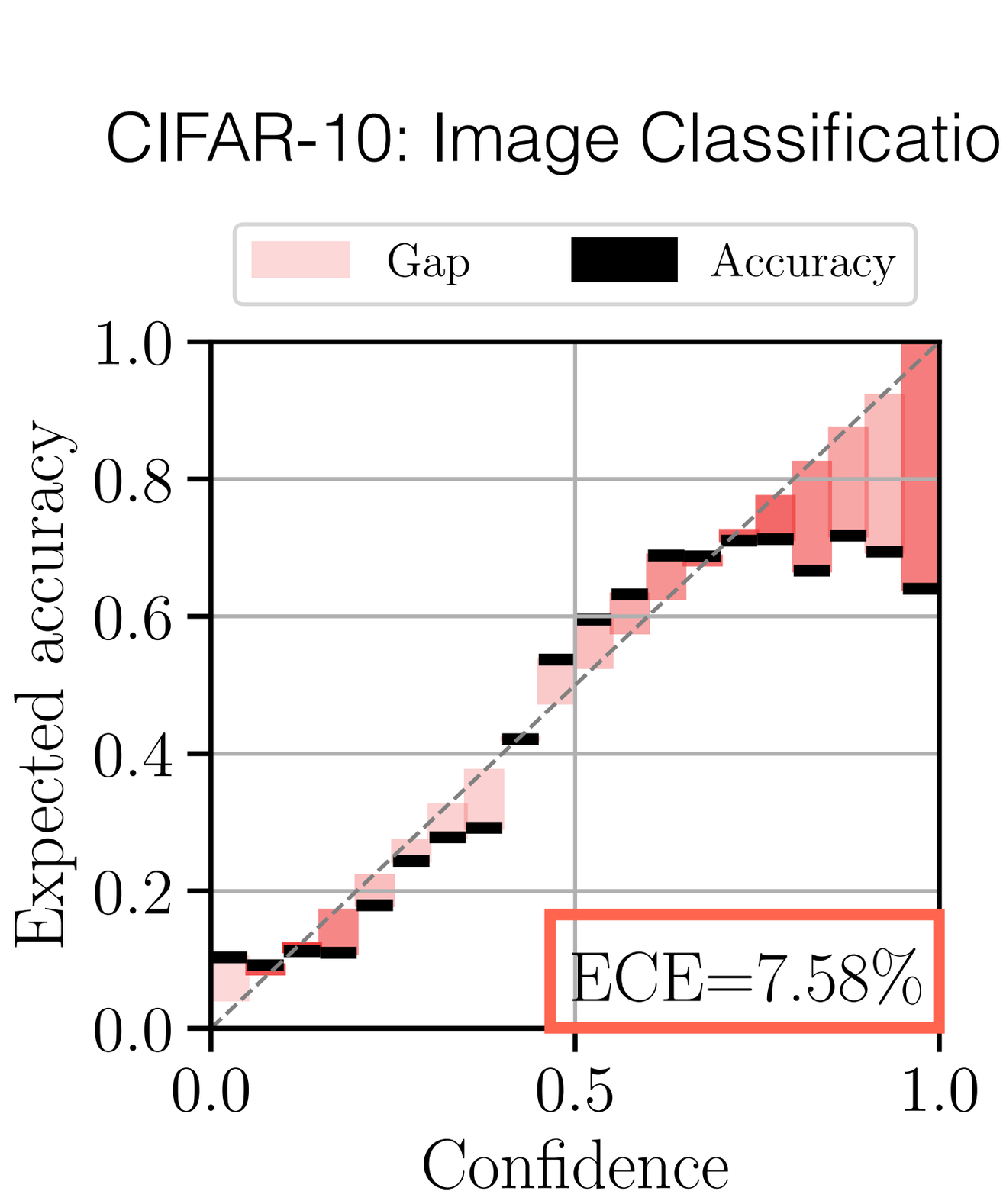
Truncation: $p_m(x) \in (0,1]$



Truncation: $p_m(x) \in (0,1]$

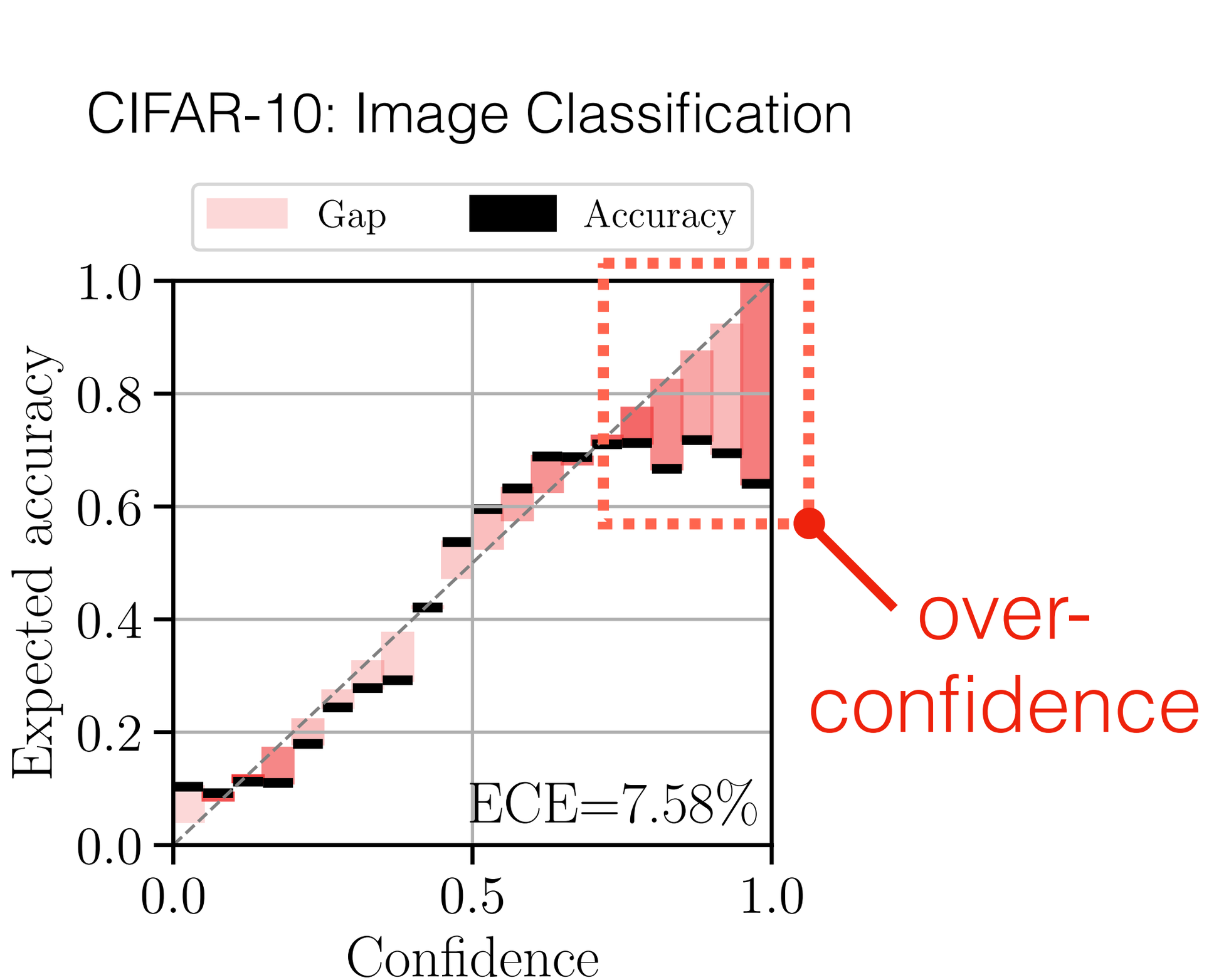


Truncation: $p_m(x) \in (0,1]$



ECE = expected calibration error

Truncation: $p_m(x) \in (0,1]$

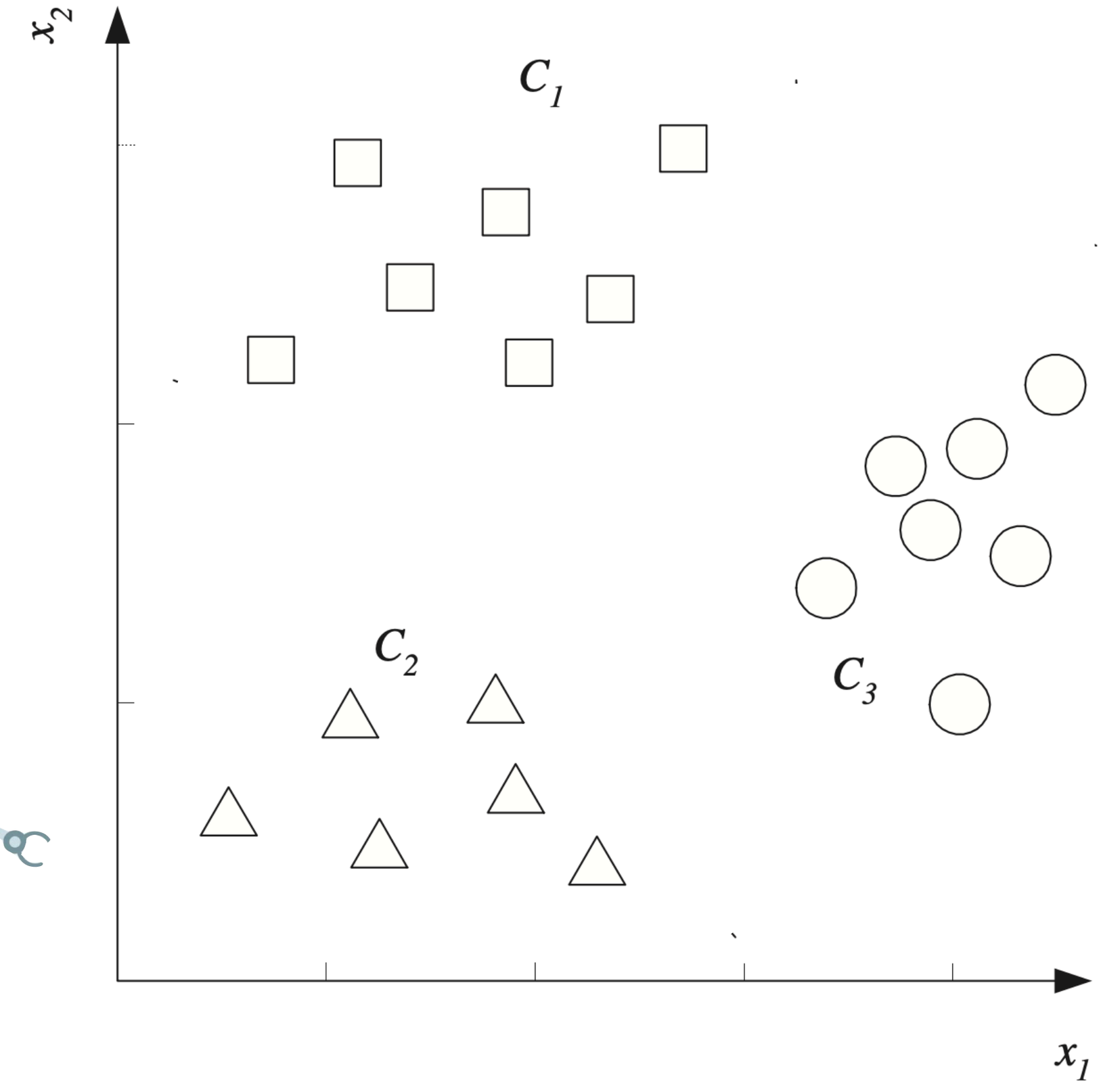


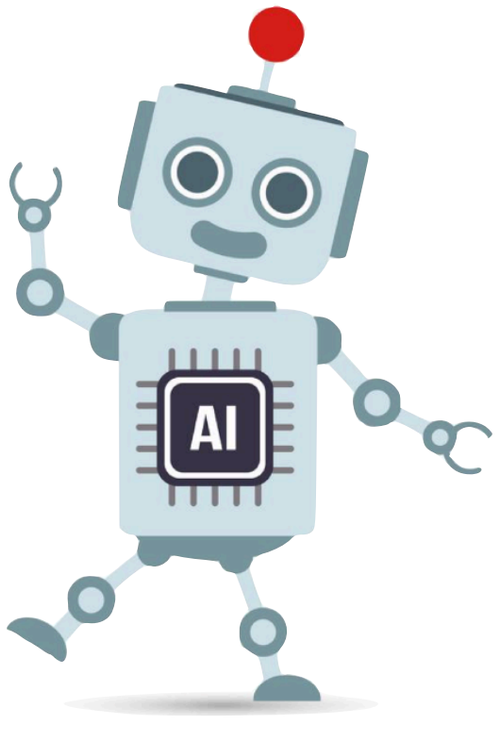
Solution

Learning to Defer using a One-vs-All Classifier

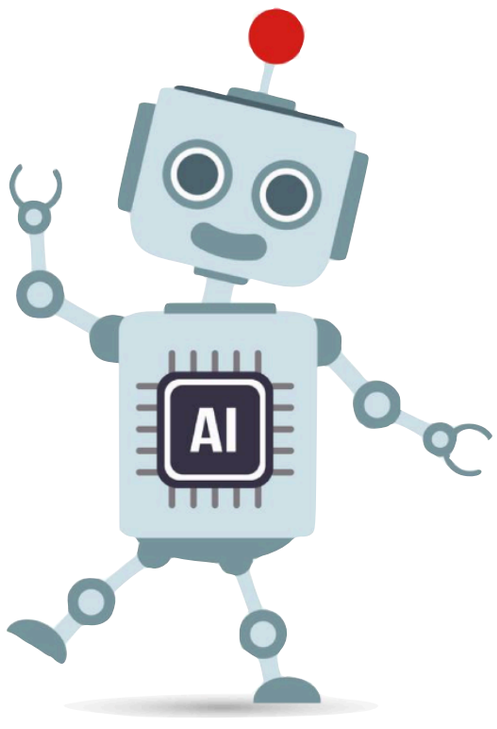
Solution

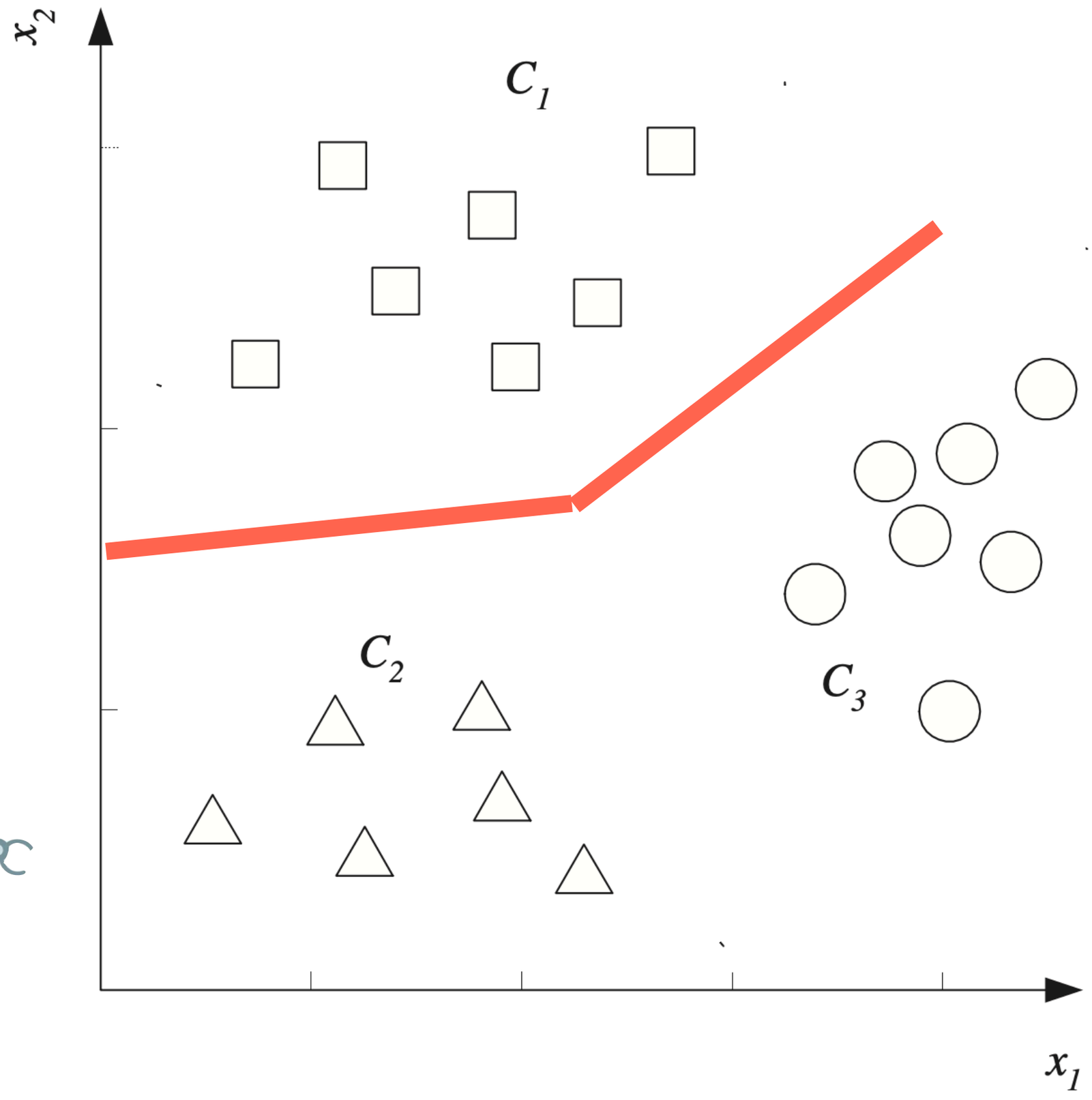
Learning to Defer using a One-vs-All Classifier



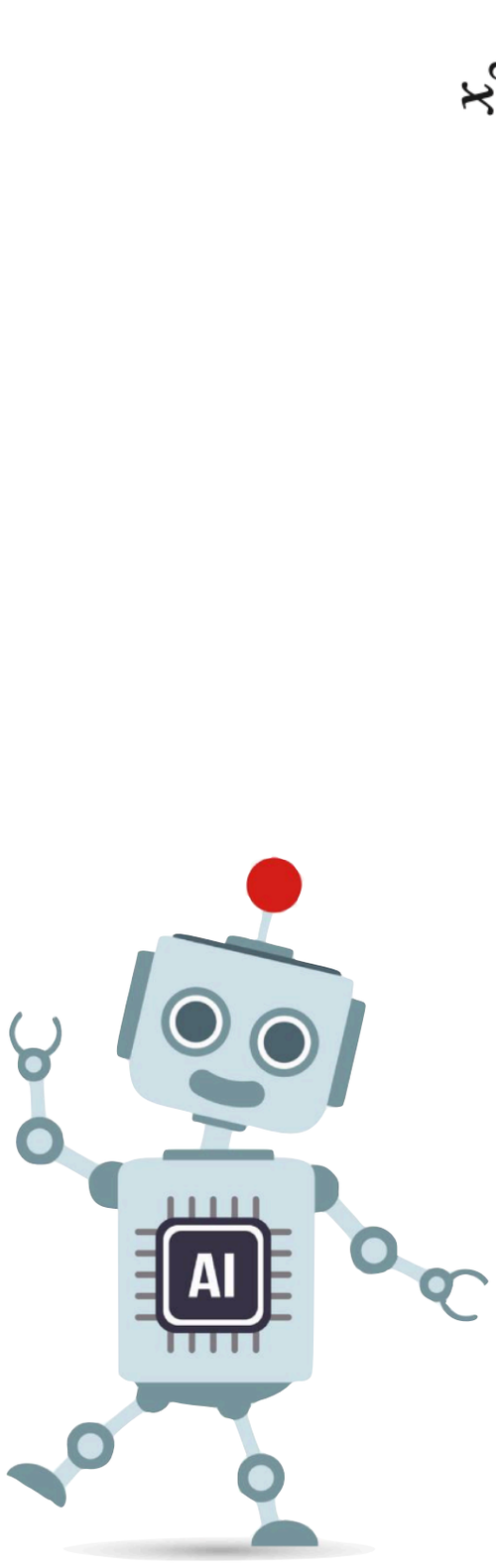


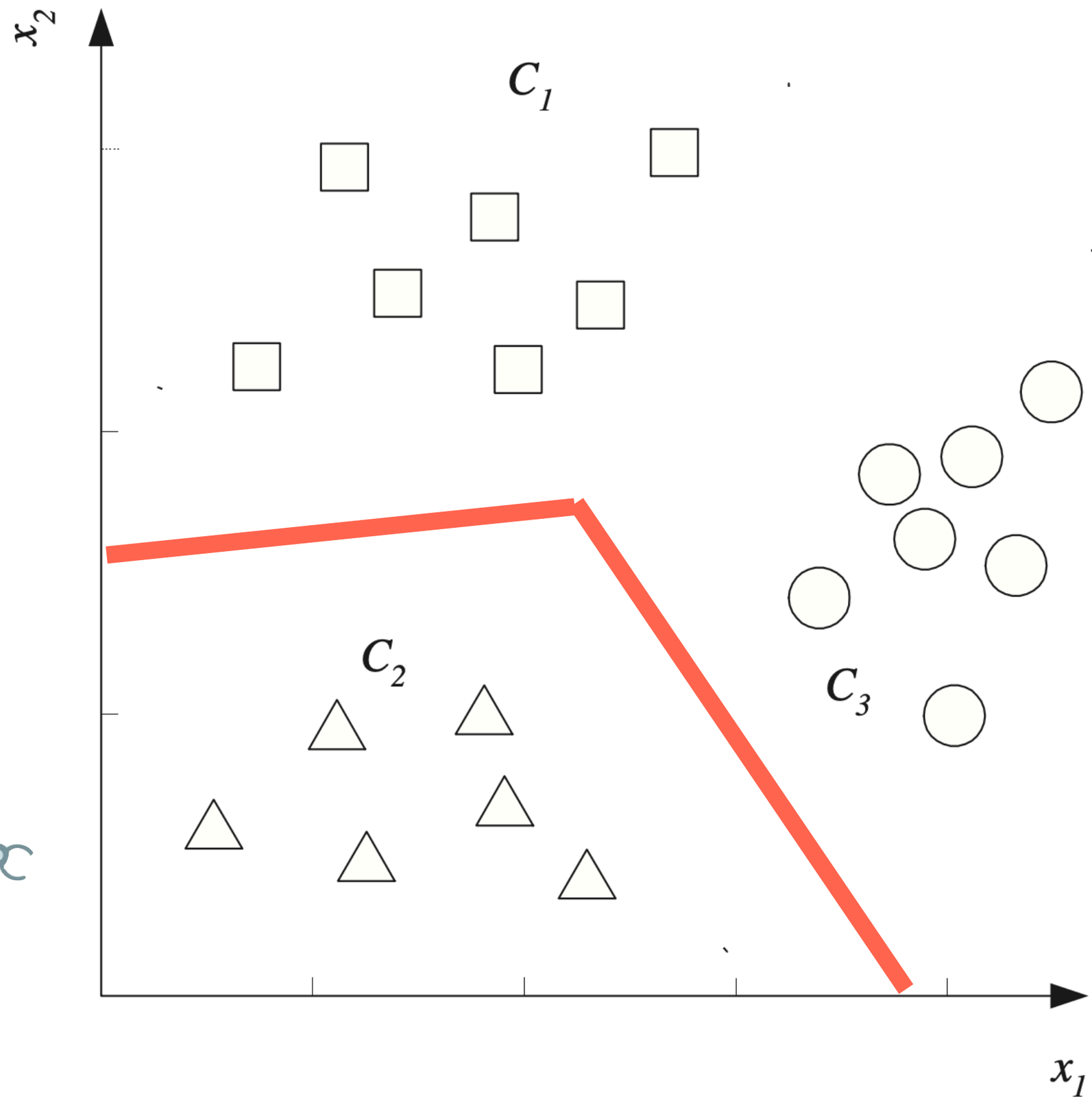
One-vs-All Classifier



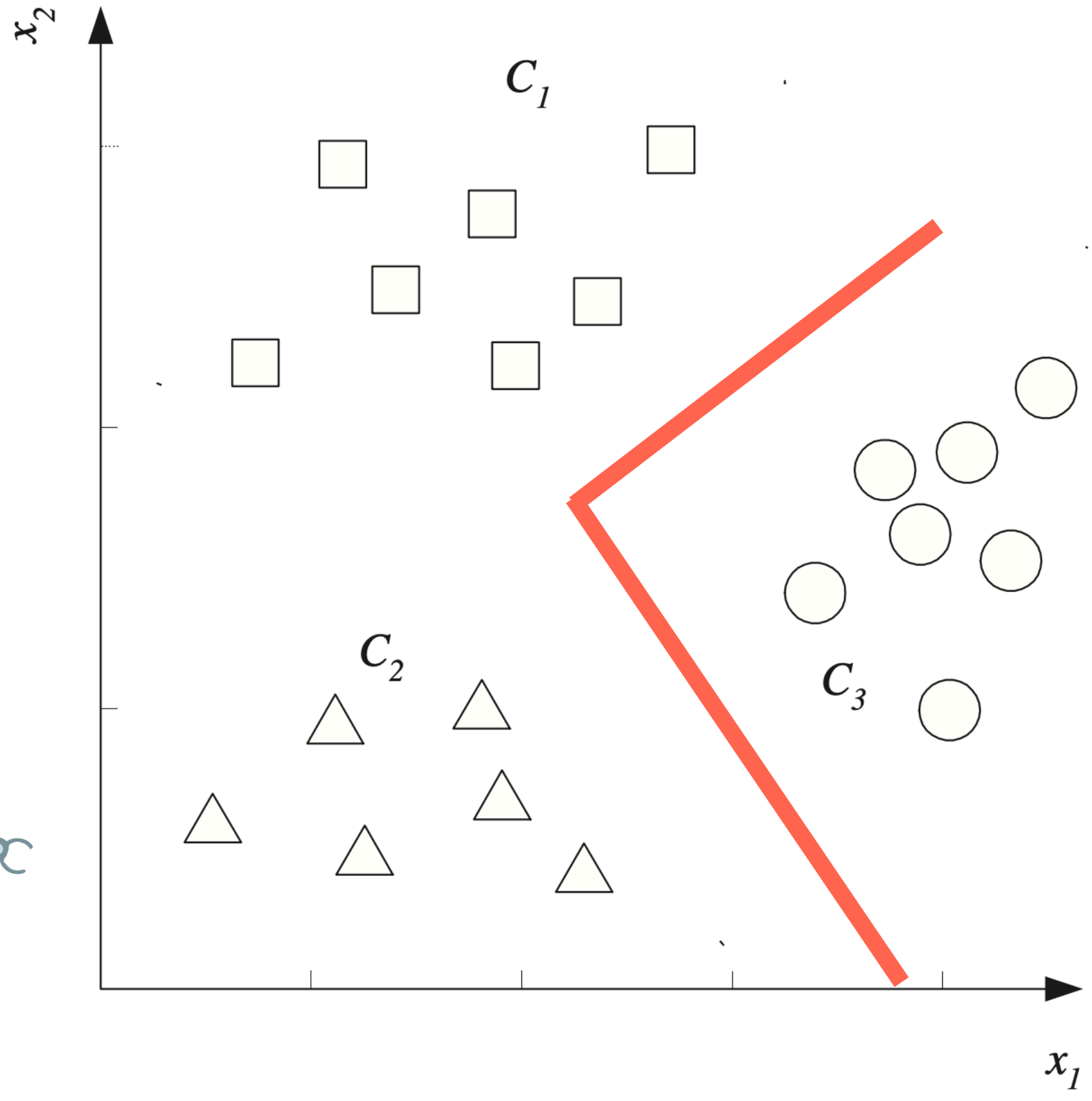


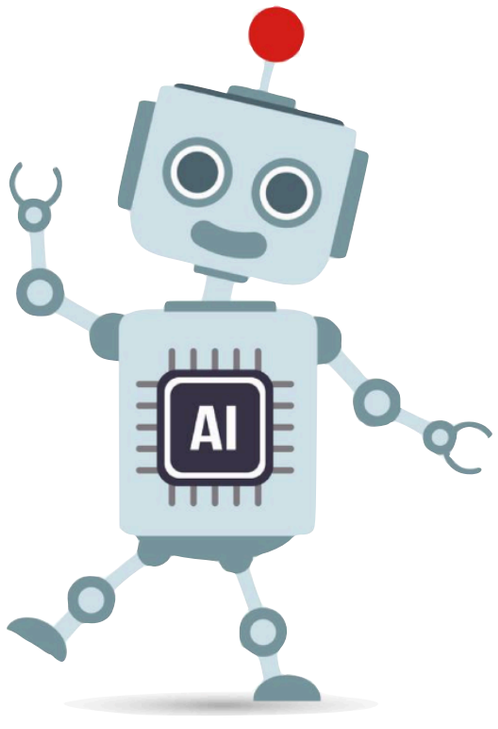
One-vs-All Classifier





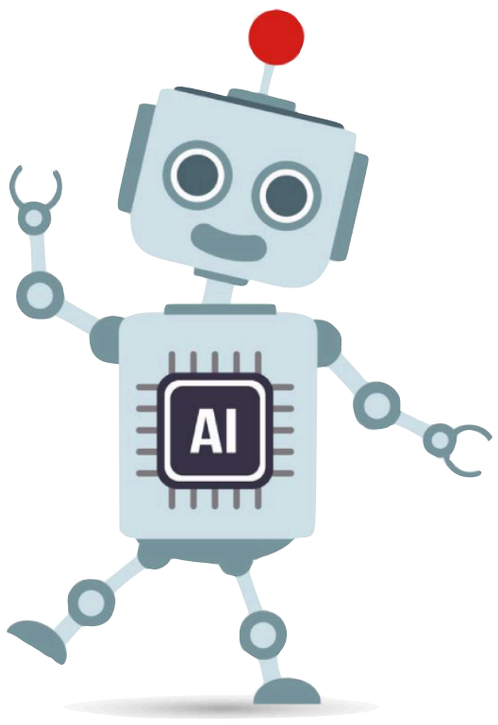
One-vs-All Classifier



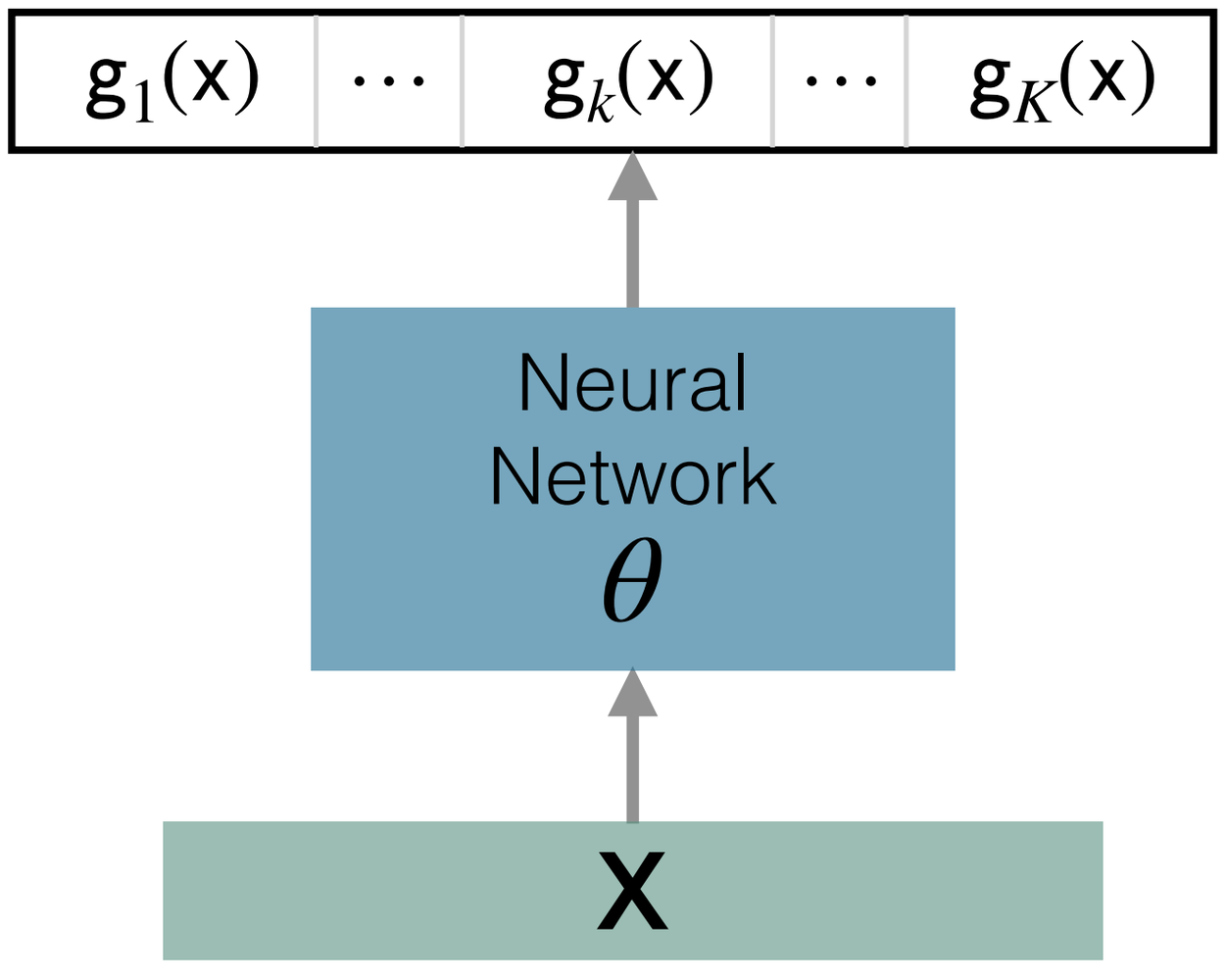


One-vs-All Classifier

Goal: For input features **x**, predict membership in one of *K* classes.

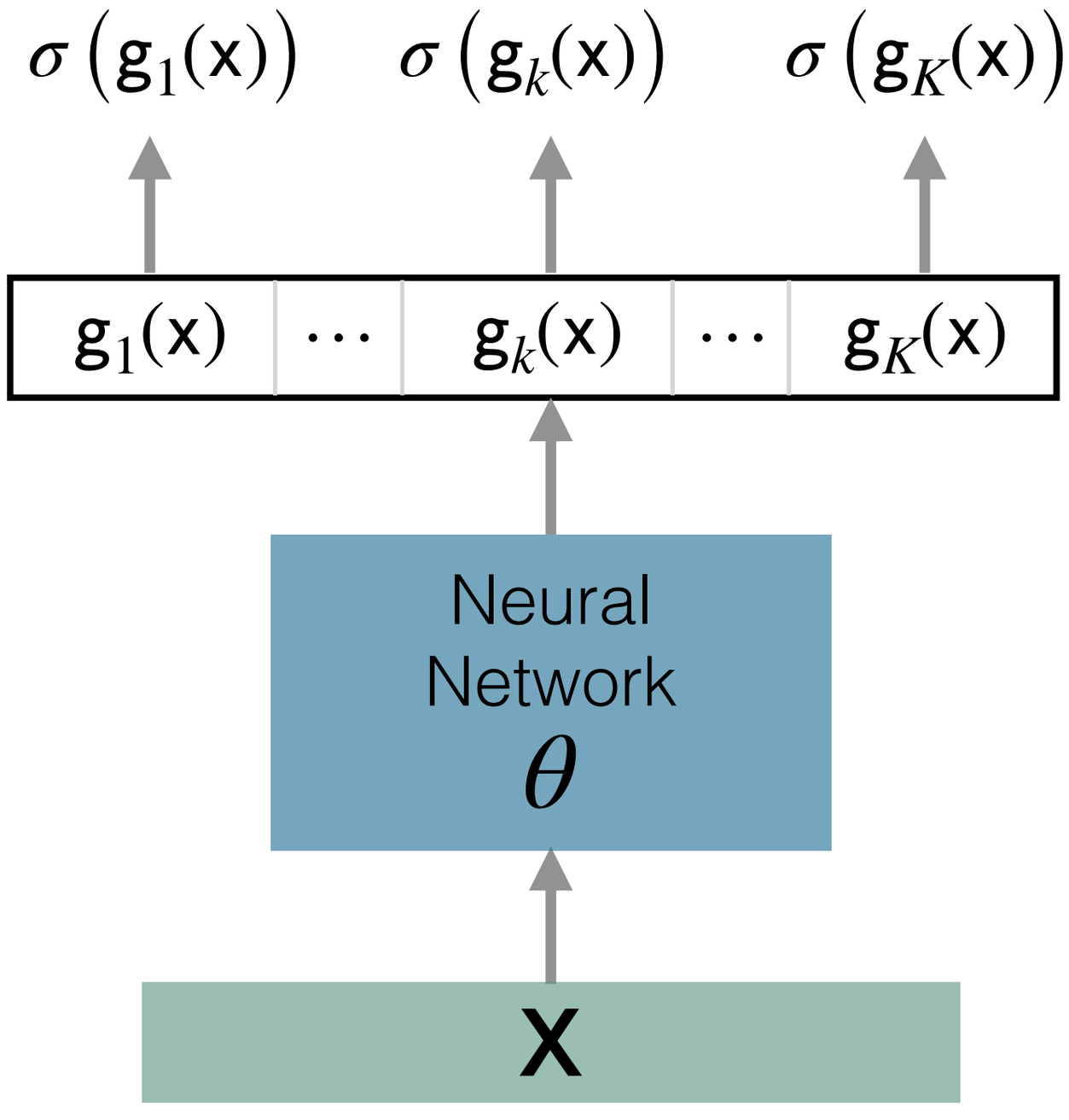


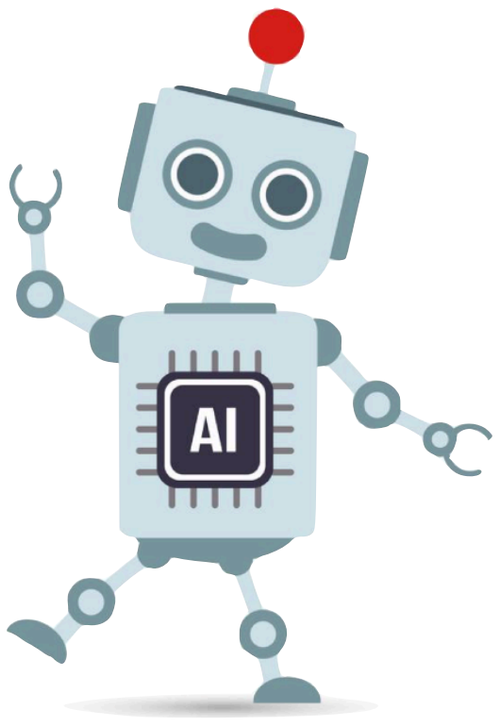
One-vs-All Classifier



Goal: For input features **x**, predict membership in one of *K* classes.

$$\sigma(z) = \frac{1}{1 + \exp\{-z\}}$$

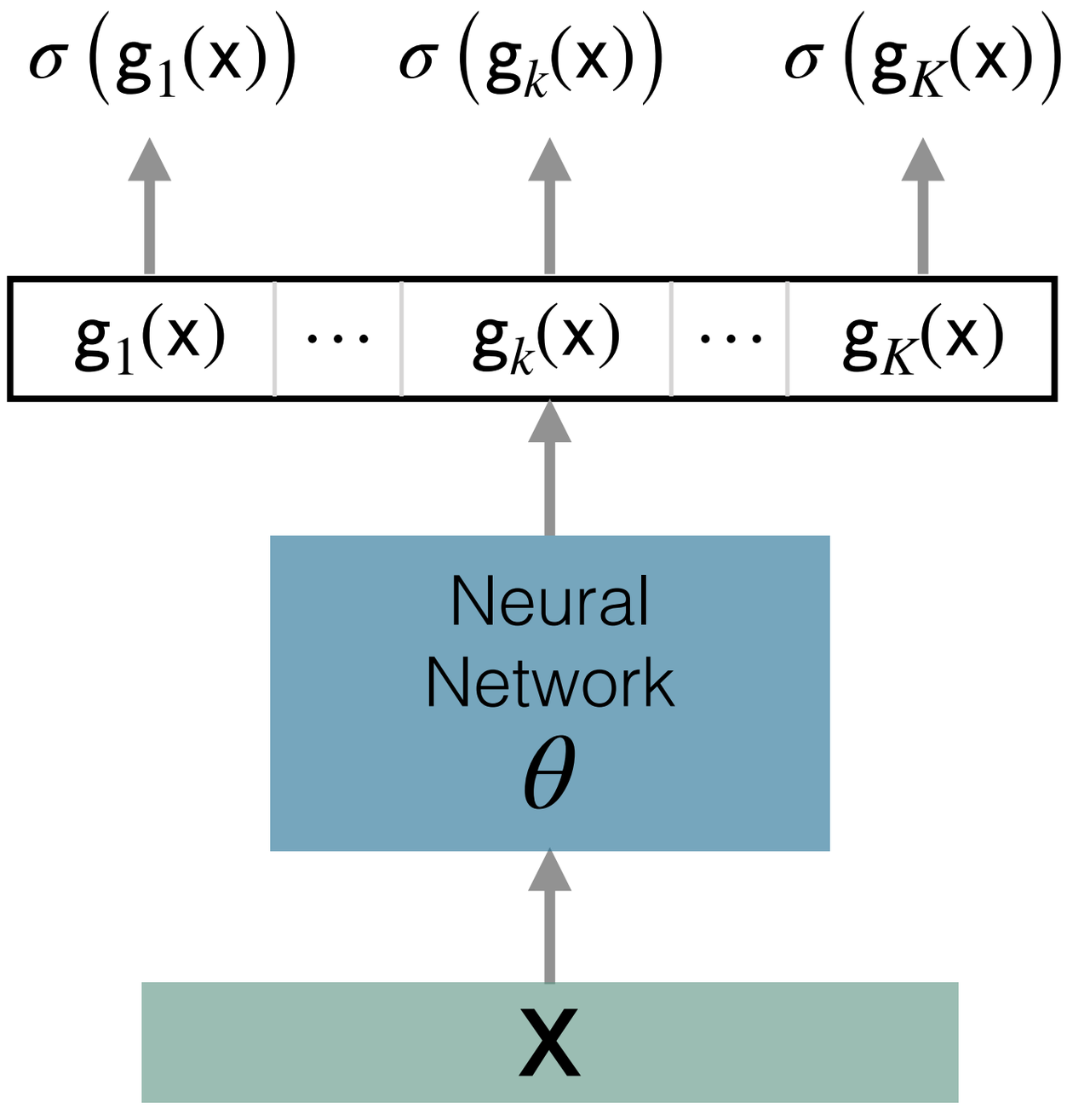


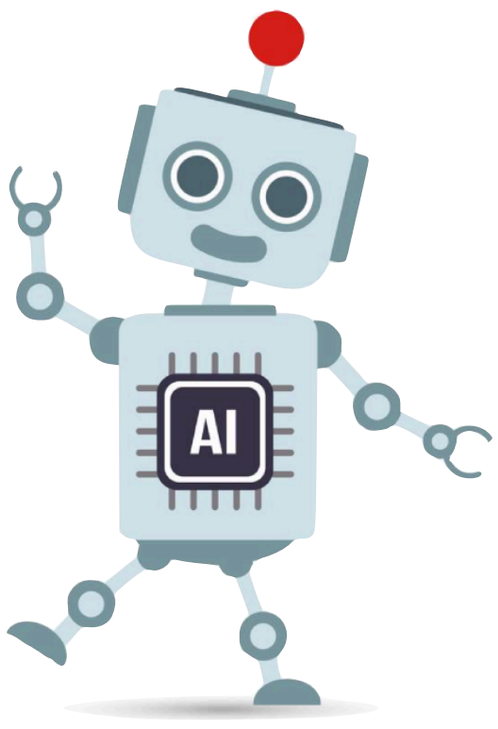


One-vs-All Classifier

Goal: For input features **x**, predict membership in one of *K* classes.

$$P(y \mid x) = \sigma\left(g_y(x)\right)$$

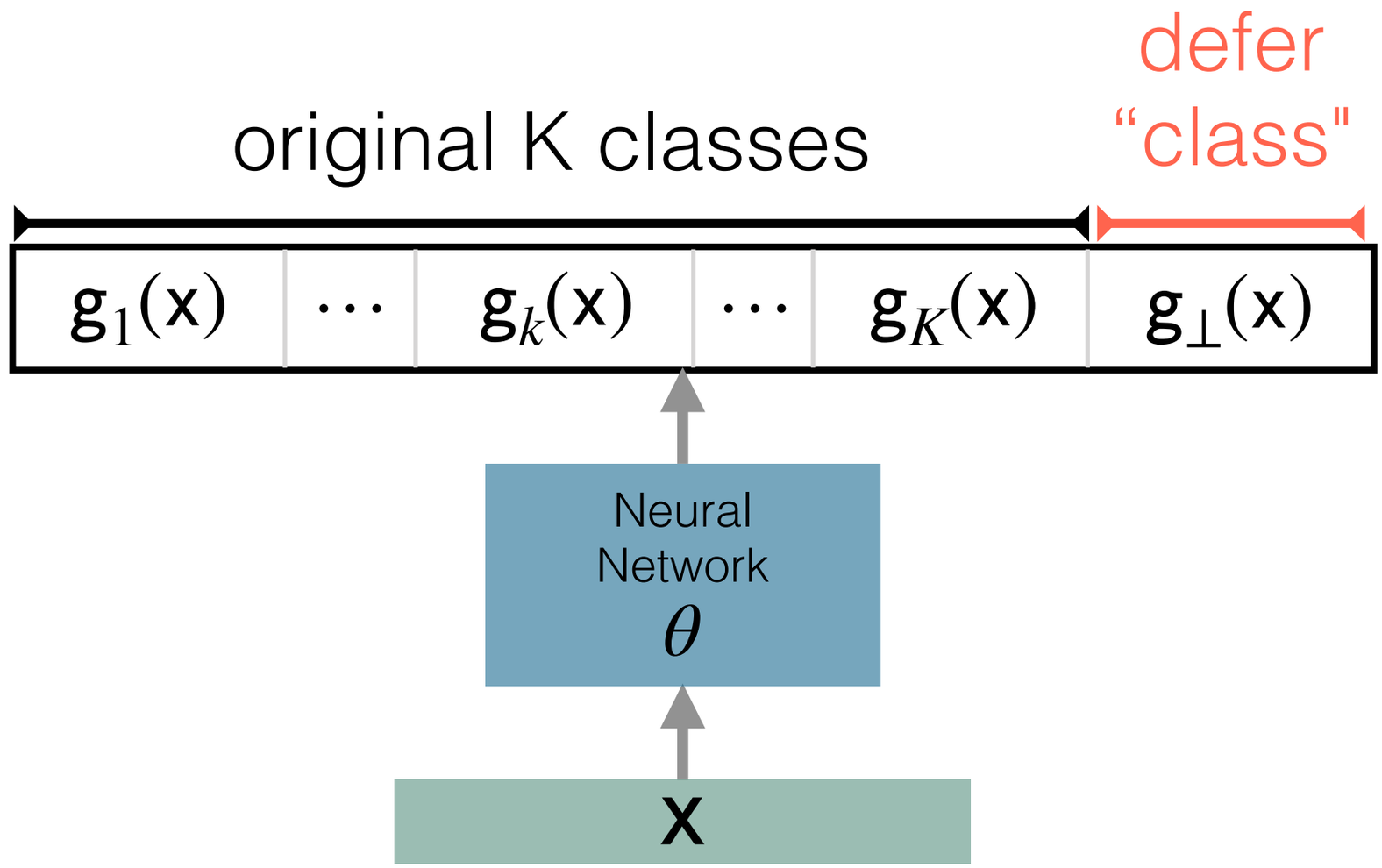




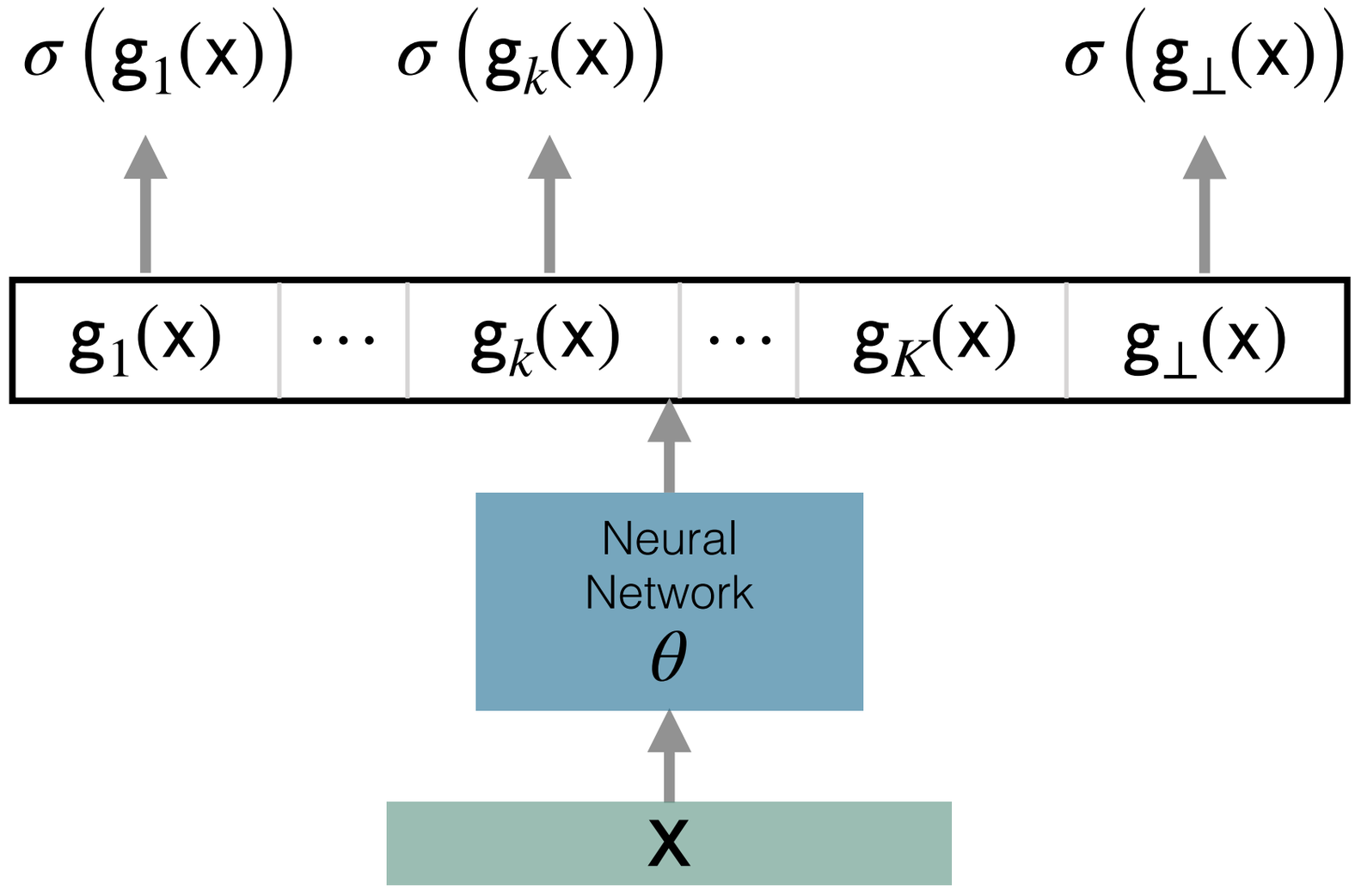
One-vs-All Classifier

Learning-to-Defer: One-vs-All Parameterization

Learning-to-Defer: One-vs-All Parameterization



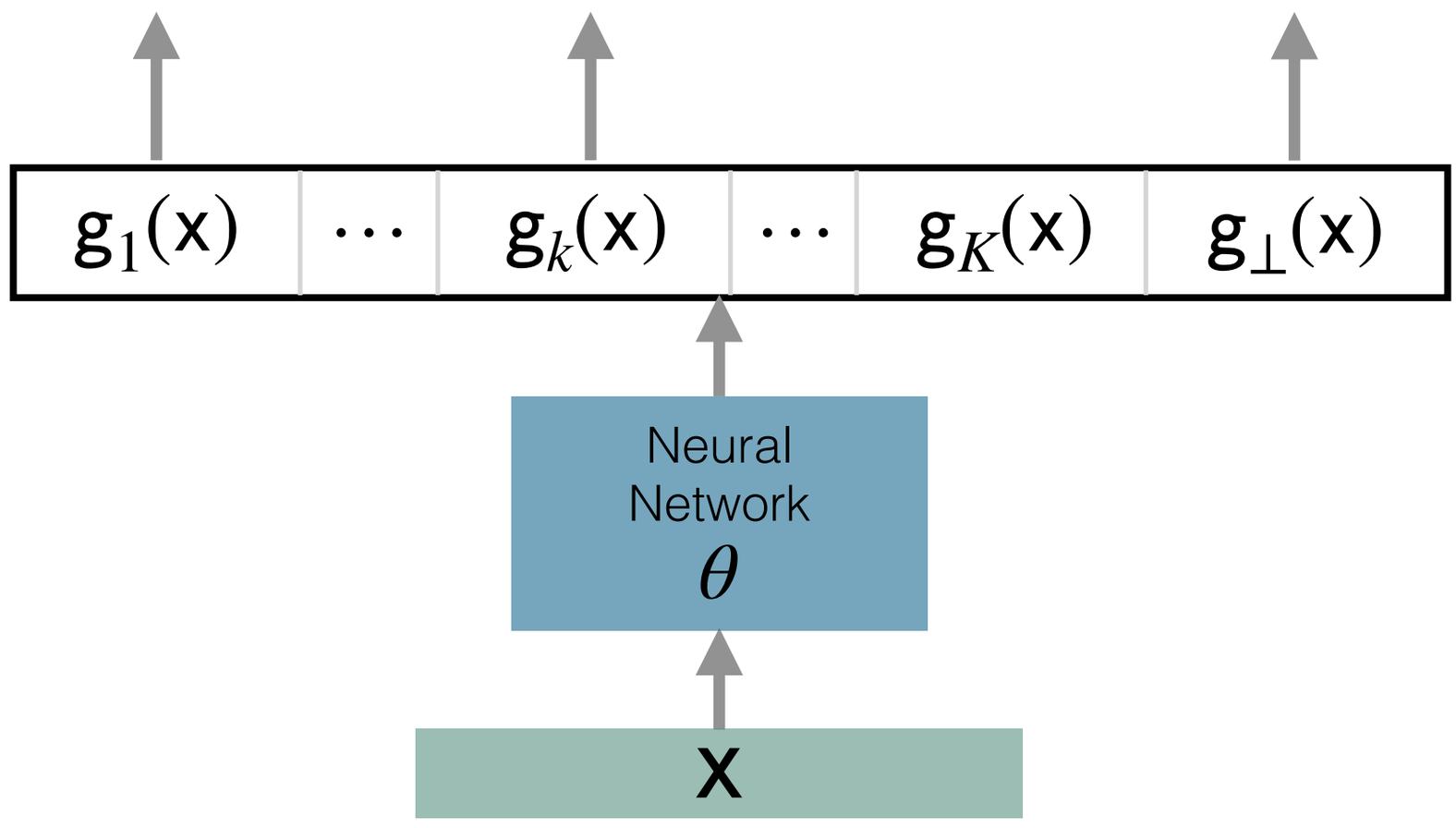
Learning-to-Defer: One-vs-All Parameterization



Learning-to-Defer: One-vs-All Parameterization

$$\mathbb{P}(\mathrm{y} = i \,|\, \mathrm{x}) \approx \sigma\left(\mathrm{g}_i(\mathrm{x})\right)$$

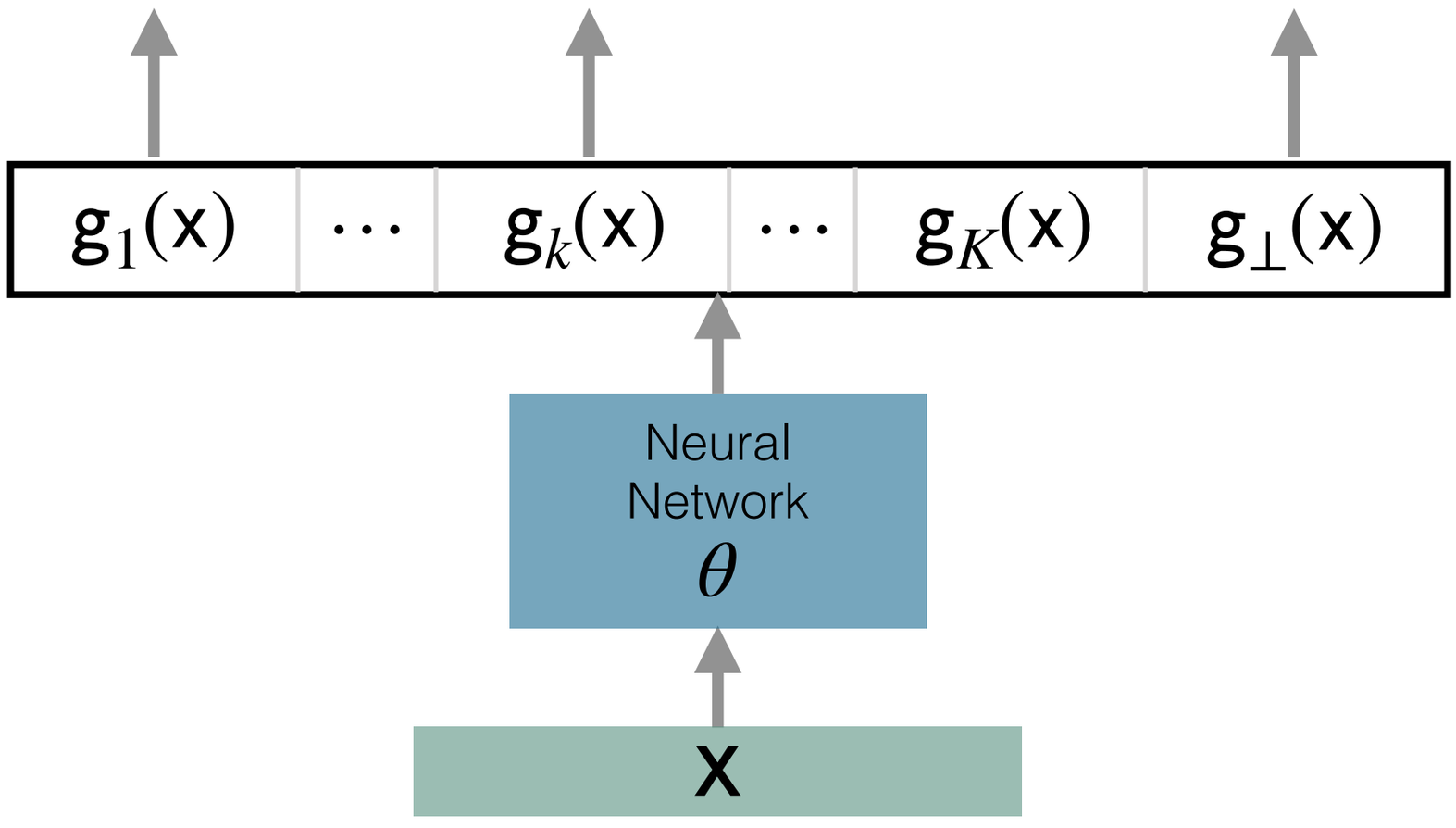
$\sigma\left(\mathrm{g}_1(\mathrm{x})\right)$ $\sigma\left(\mathrm{g}_k(\mathrm{x})\right)$ $\sigma\left(\mathrm{g}_\perp(\mathrm{x})\right)$



Learning-to-Defer: One-vs-All Parameterization

$$\mathbb{P}(\mathsf{m} = \mathsf{y} \,|\, \mathsf{x}) \approx \sigma\left(\mathsf{g}_{\perp}(\mathsf{x})\right)$$

$\sigma\left(\mathsf{g}_1(\mathsf{x})\right)$ $\sigma\left(\mathsf{g}_k(\mathsf{x})\right)$ $\sigma\left(\mathsf{g}_{\perp}(\mathsf{x})\right)$



Learning-to-Defer: One-vs-All Parameterization

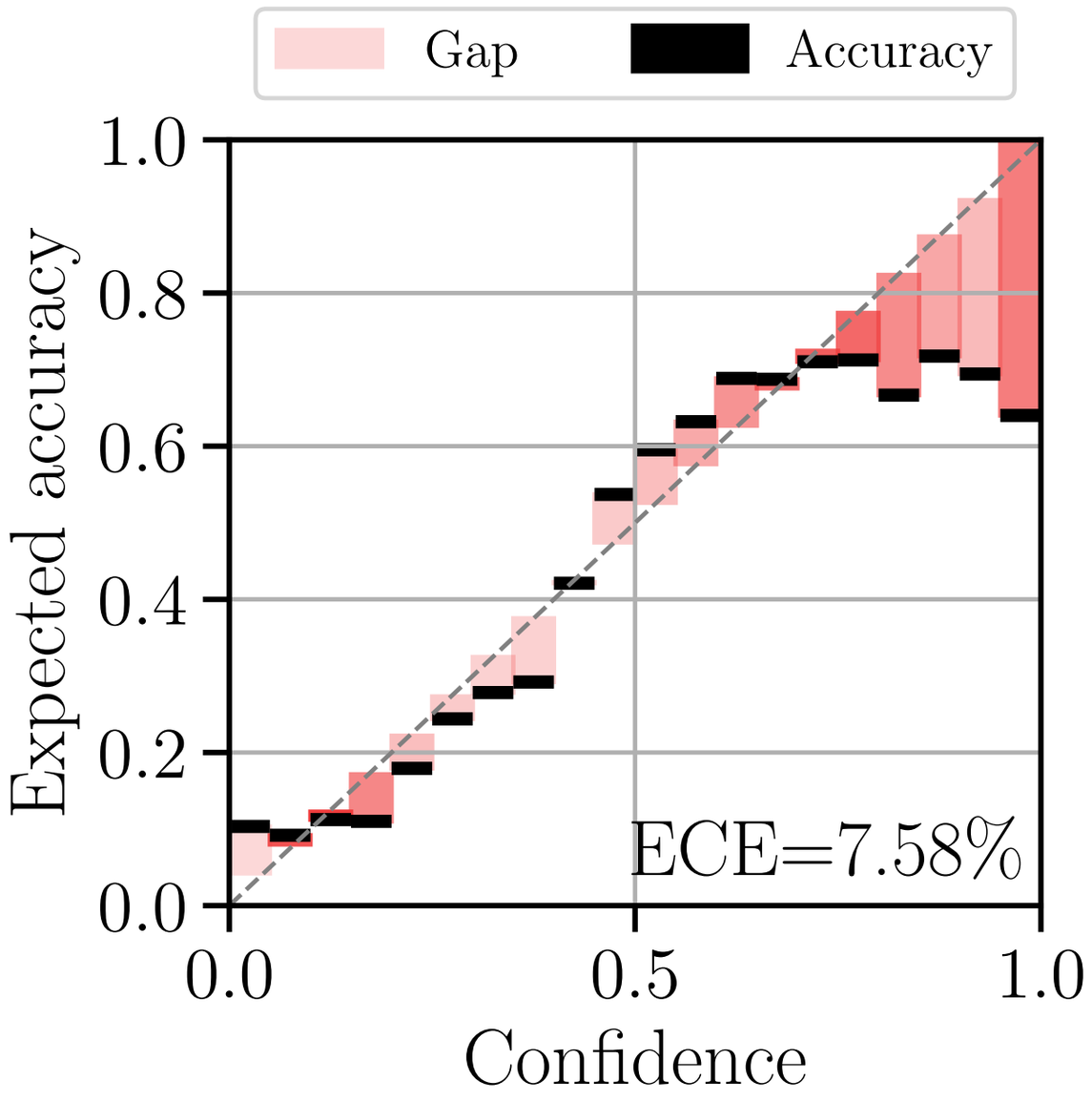
Theorem 4.1 + Corollary 4.2 [Verma & Nalisnick, 2022]**:**
The one-vs-all loss is a *consistent* surrogate for the 0-1 learning-to-defer loss.

Does the one-vs-all loss result in **better calibrated** models in practice?

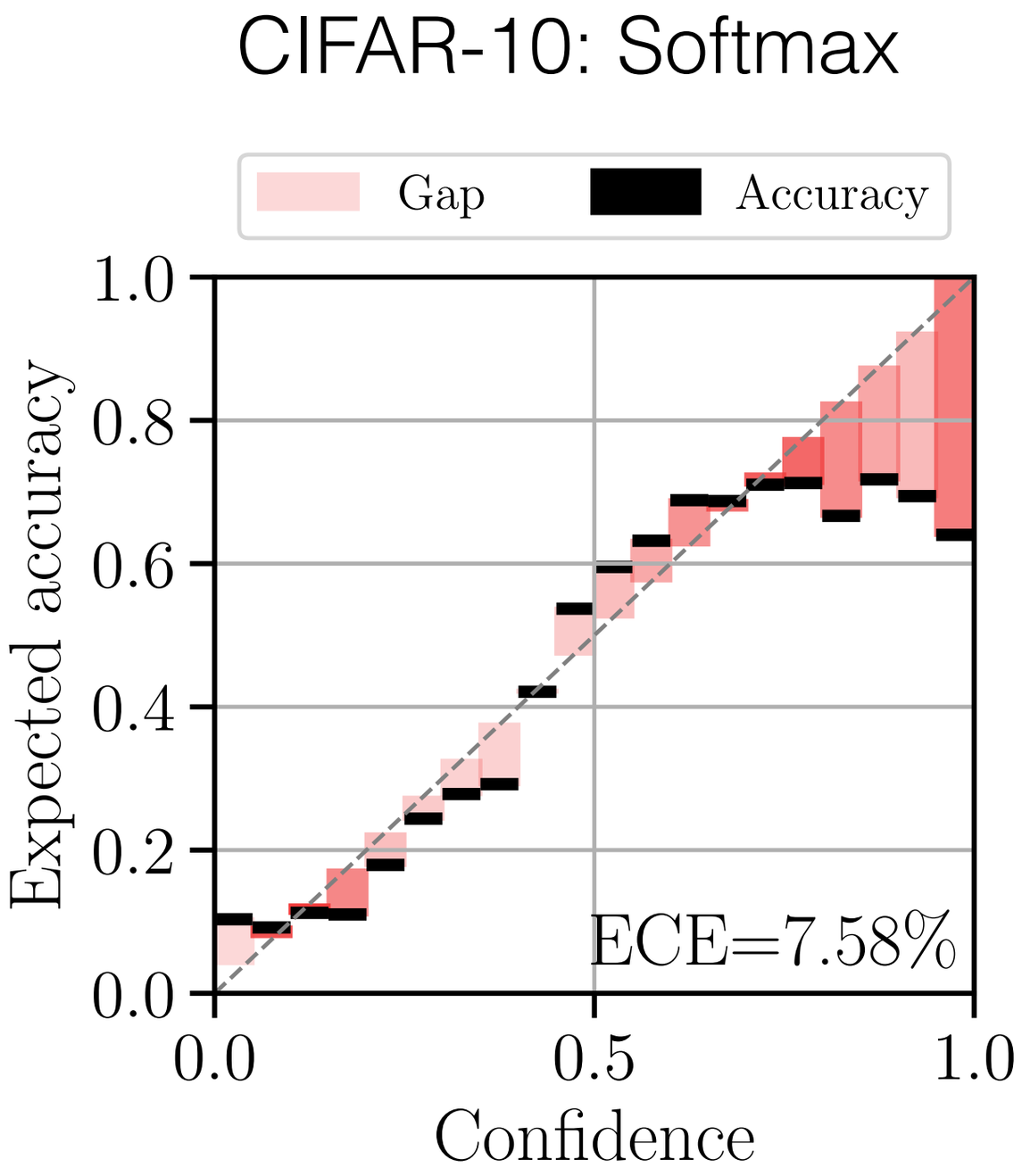
CIFAR-10: Softmax

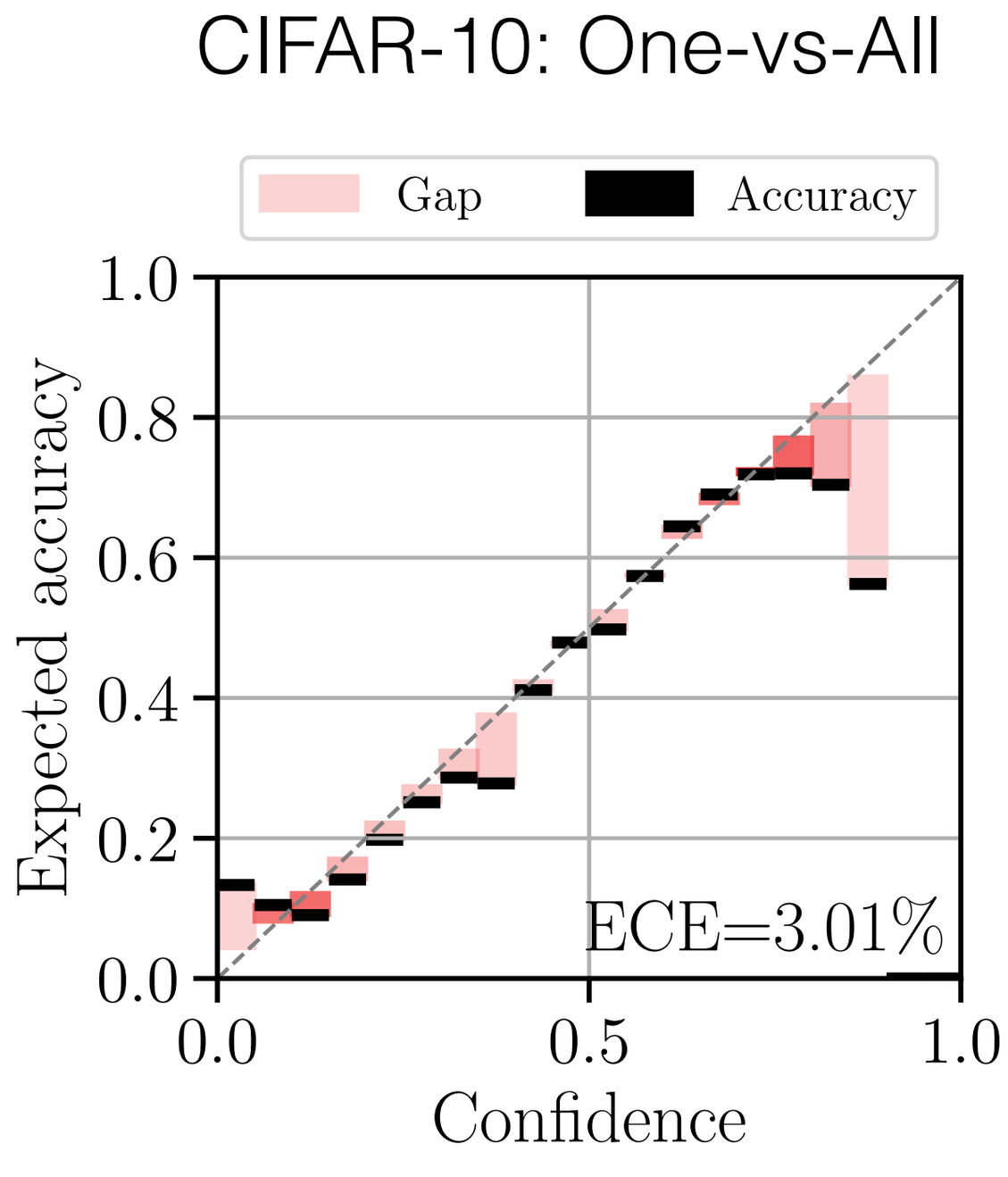
CIFAR-10: One-vs-All

CIFAR-10: Softmax

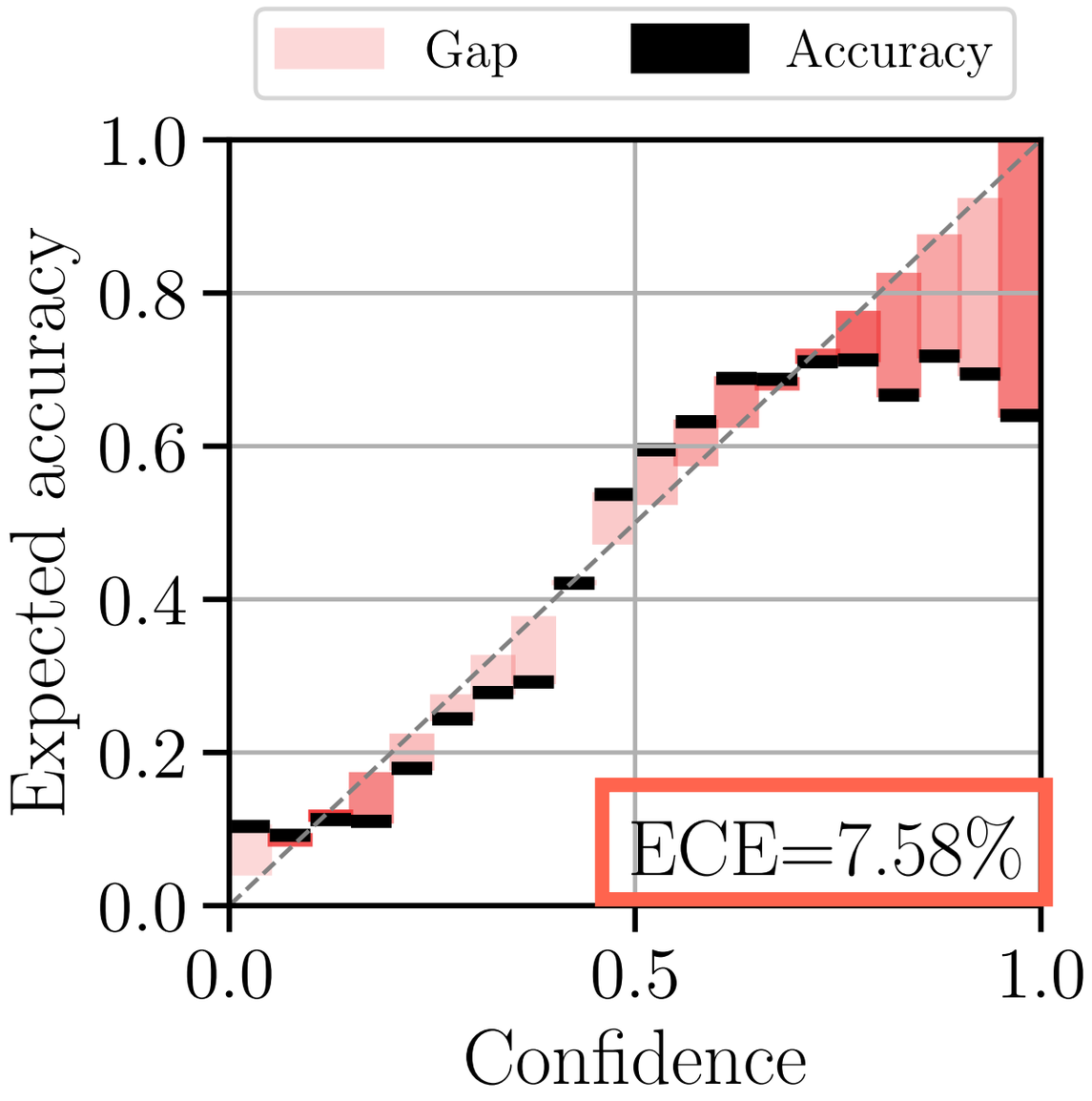


CIFAR-10: One-vs-All

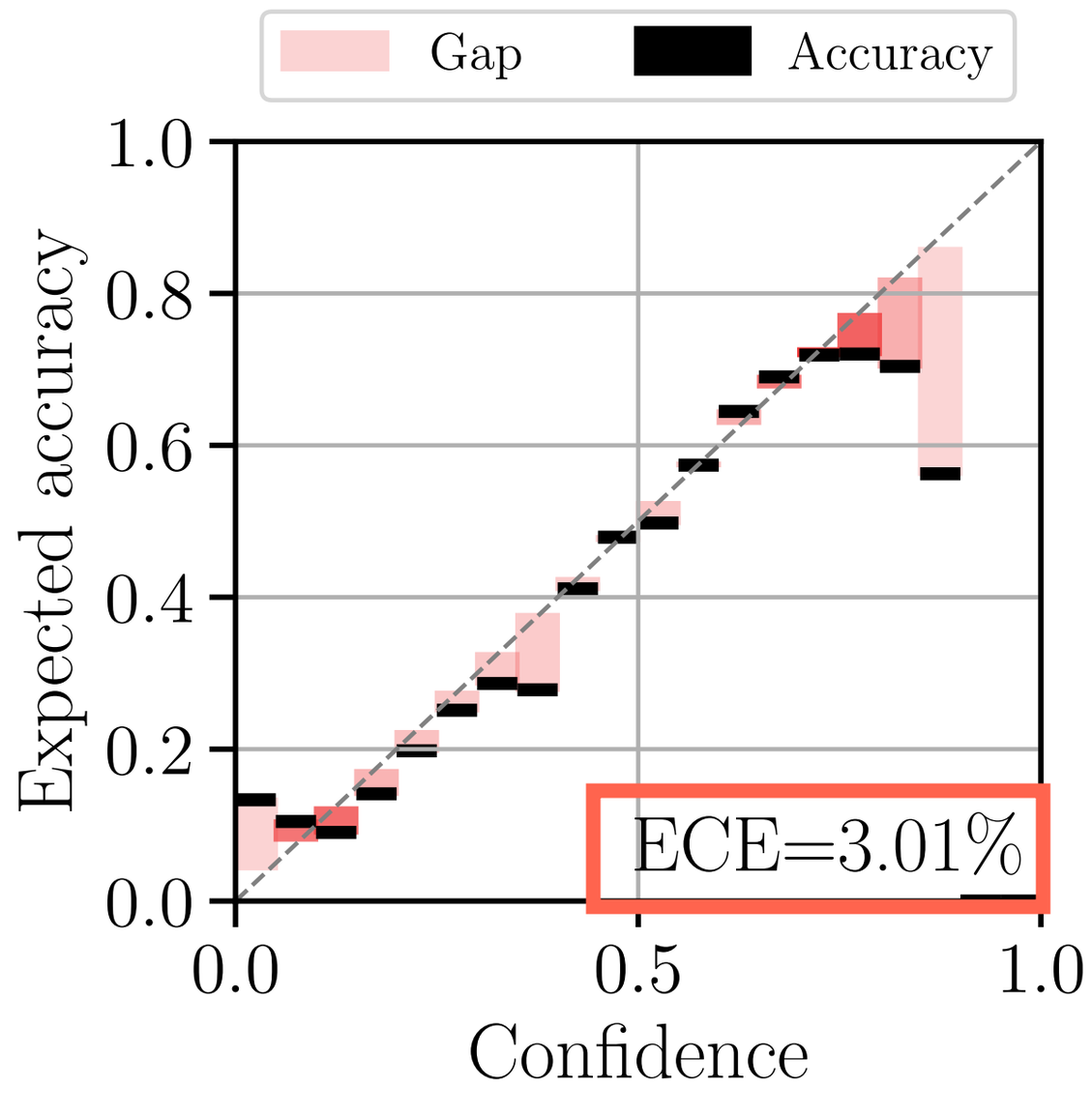

CIFAR-10: Softmax
Gap
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Expected accuracy
ECE=7.58%
0.0
0.5
1.0
Confidence


CIFAR-10: One-vs-All
Gap
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
Expected accuracy
ECE=3.01%
0.0
0.5
1.0
Confidence

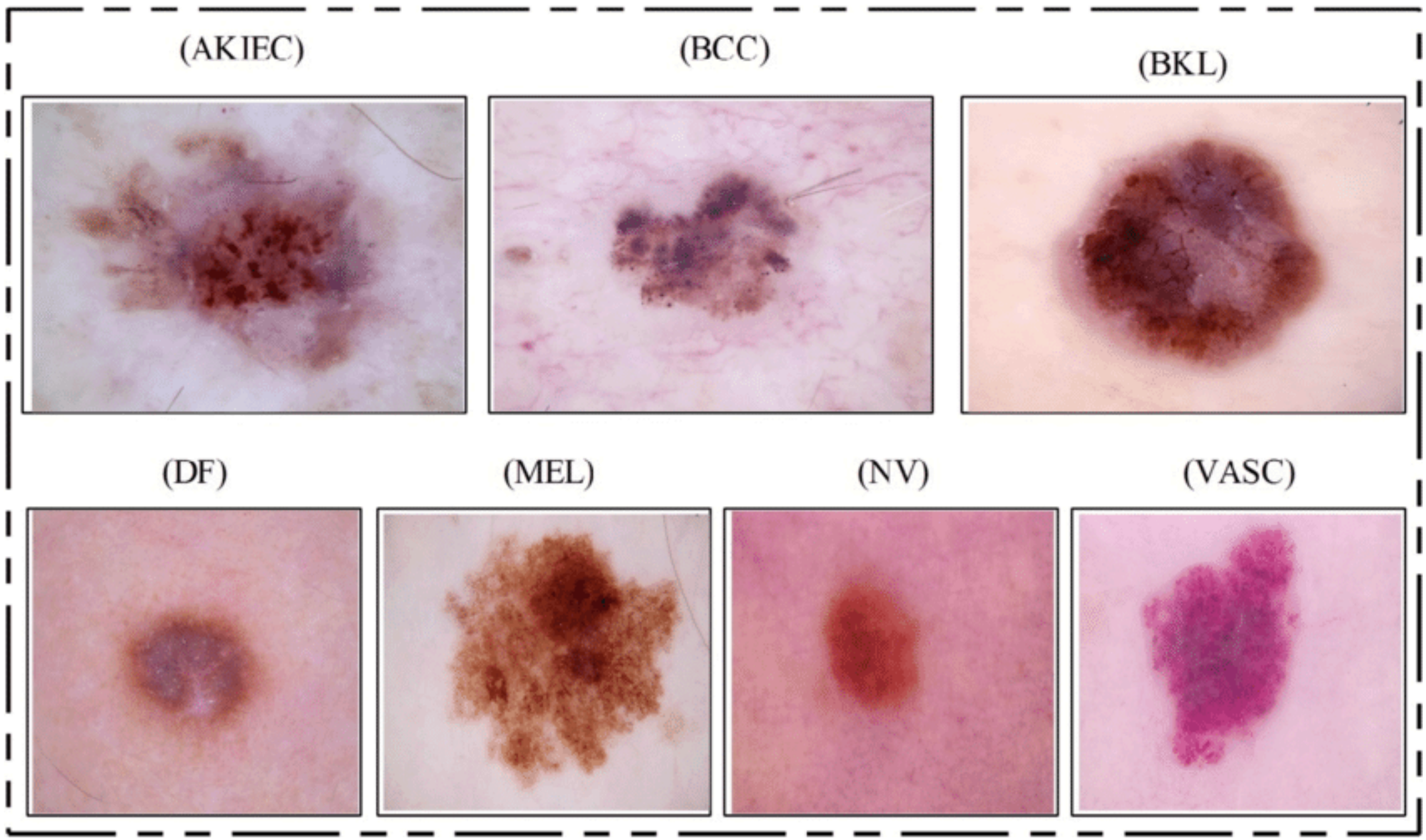
CIFAR-10: Softmax



CIFAR-10: One-vs-All



HAM10000: Skin Lesion Classification



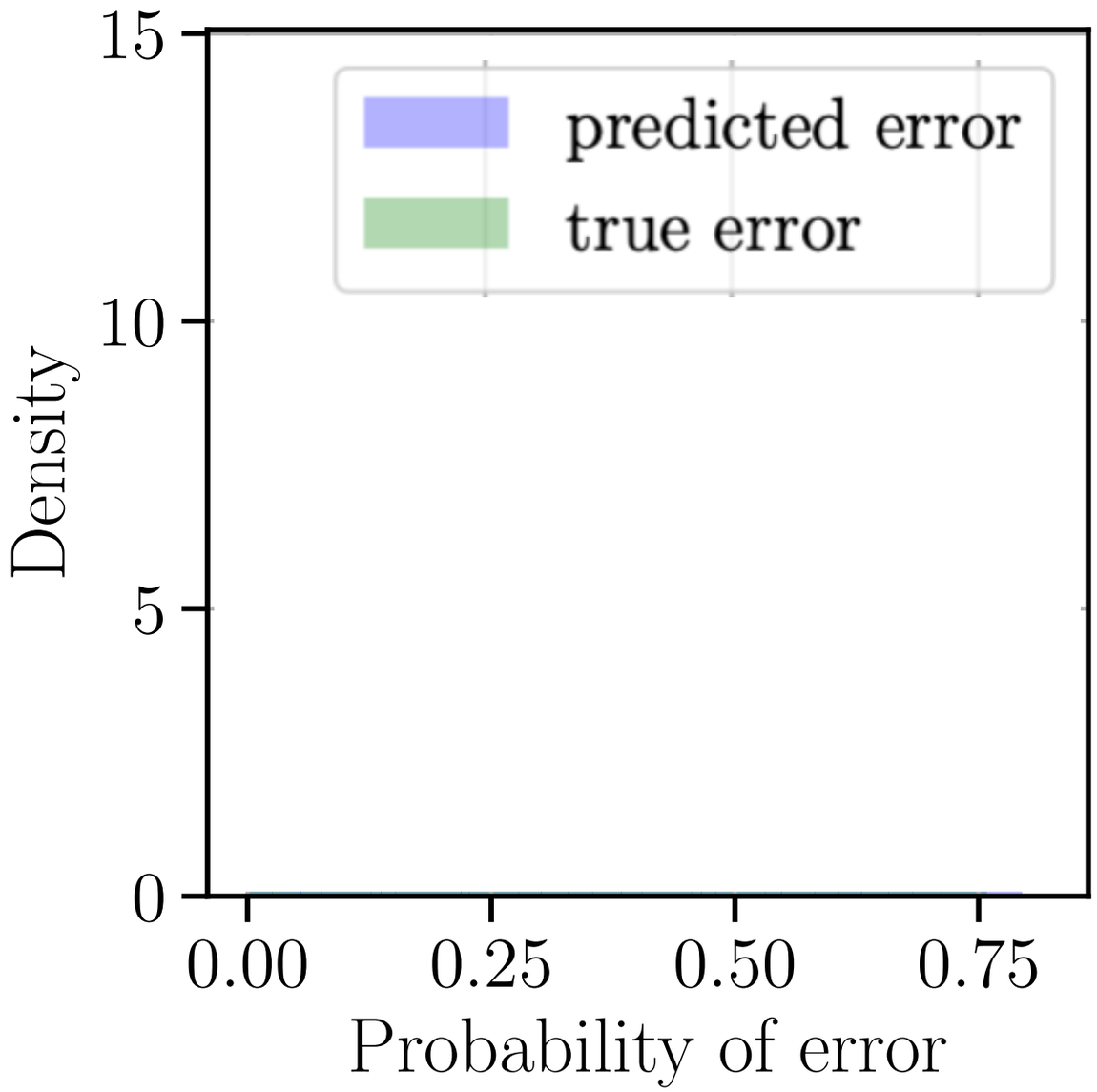
HAM10000: Skin Lesion Classification

Softmax

One-vs-All

HAM10000: Skin Lesion Classification

Softmax

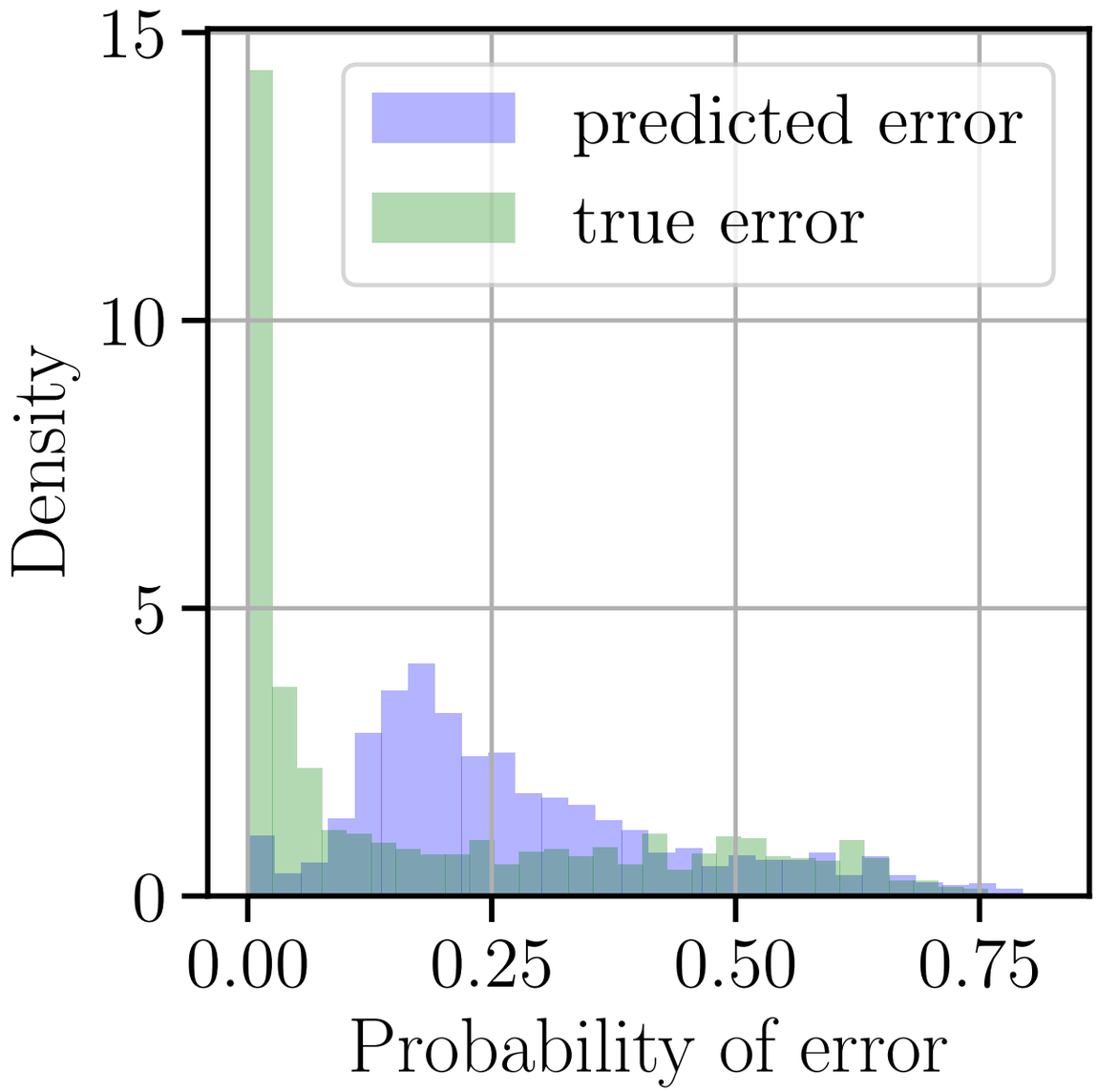


One-vs-All

$1 - P(m = y \mid x)$

HAM10000: Skin Lesion Classification

Softmax

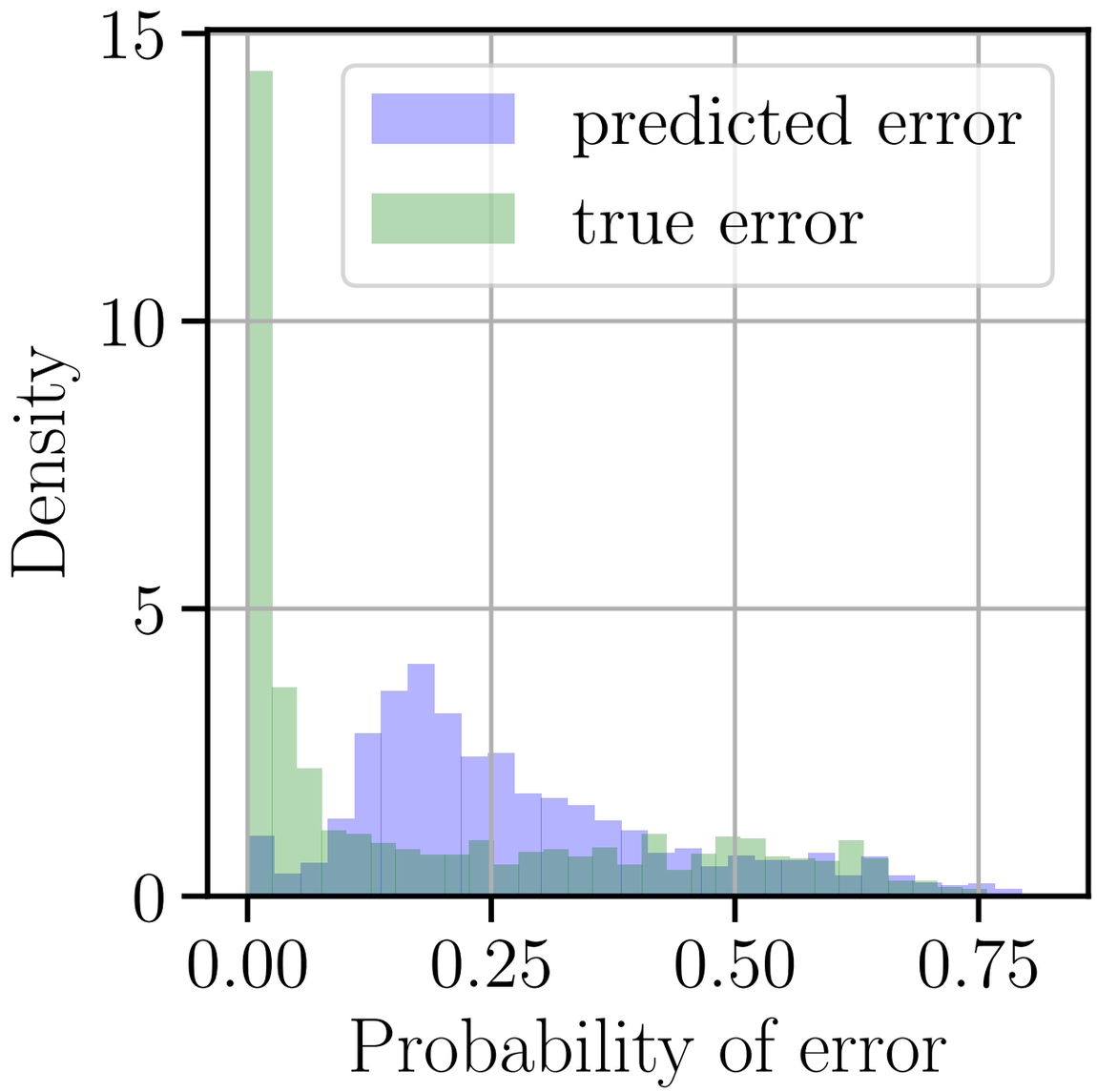


One-vs-All

$$1 - P(m = y \mid x)$$

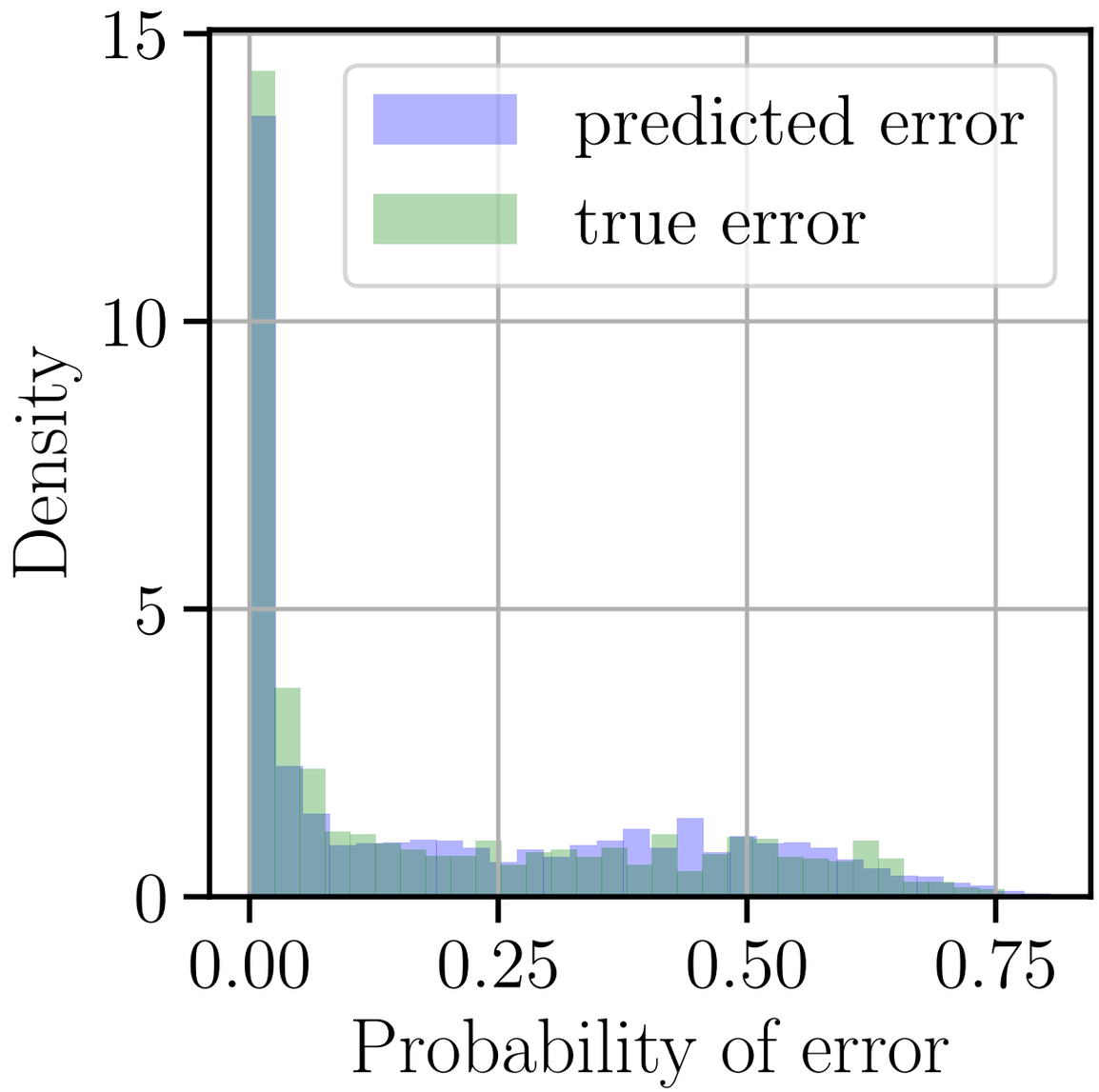
HAM10000: Skin Lesion Classification

Softmax



$1 - P(m = y \mid x)$

One-vs-All



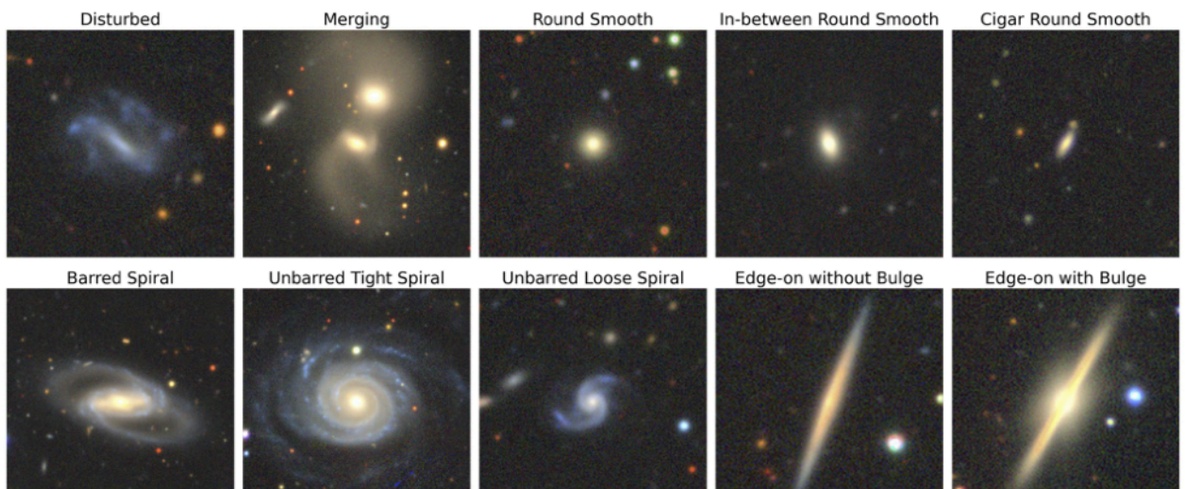
$1 - P(m = y \mid x)$

Does the one-vs-all loss result in **more accurate** models in practice?

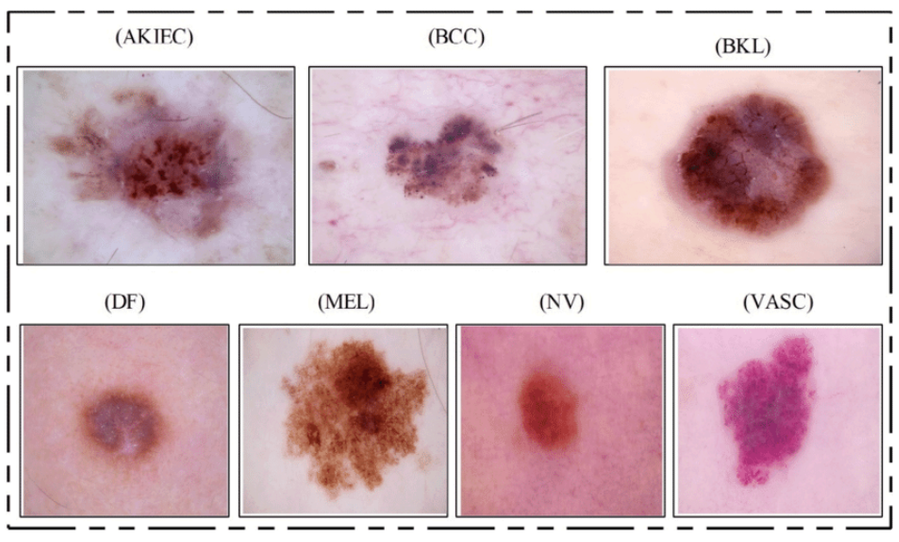
one-vs-all (ours) softmax confidence

score differentiable triage

Galaxy Zoo



HAM10000



Hate Speech

[Davidson et al., ICWSM 2017]

one-vs-all (ours)

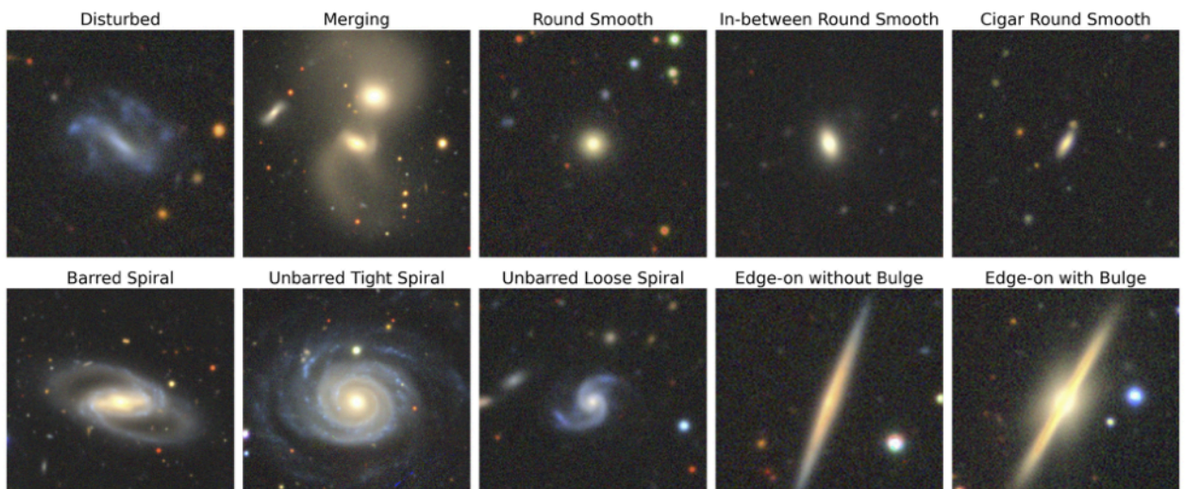
softmax

confidence

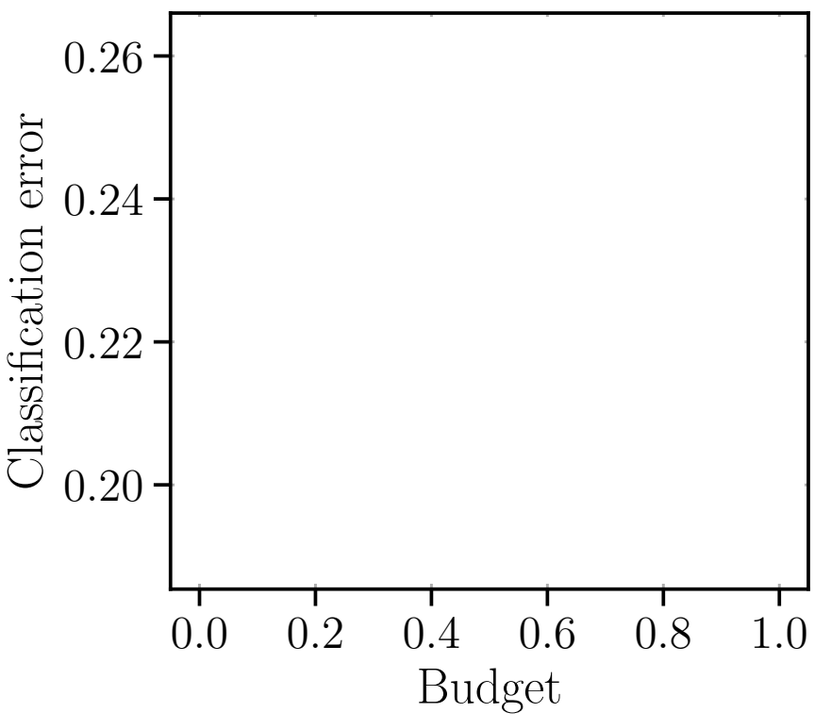
score

differentiable triage

Galaxy Zoo



HAM10000



Hate Speech

[Davidson et al., ICWSM 2017]

one-vs-all (ours)

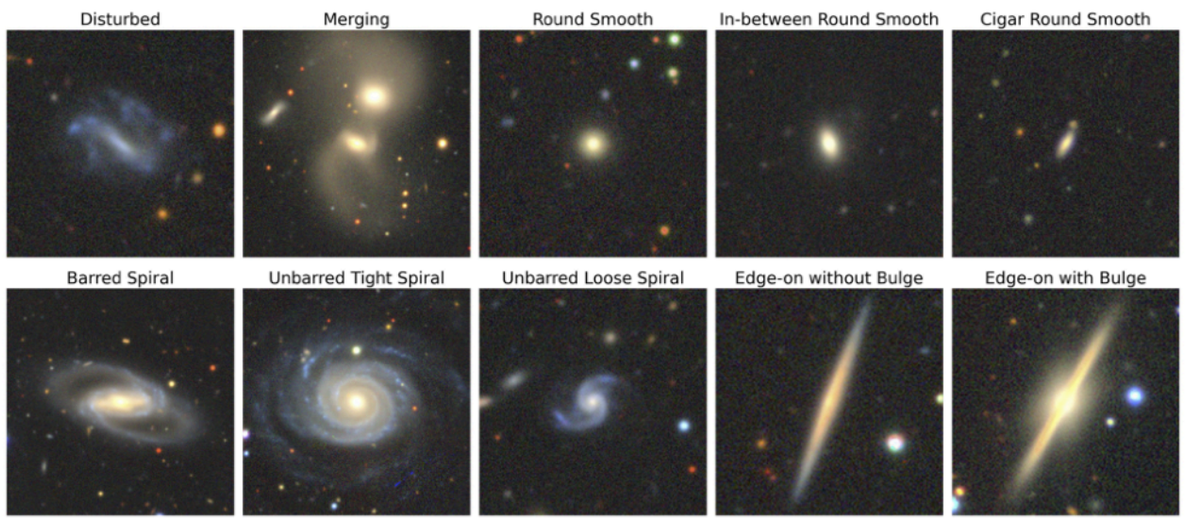
softmax

confidence

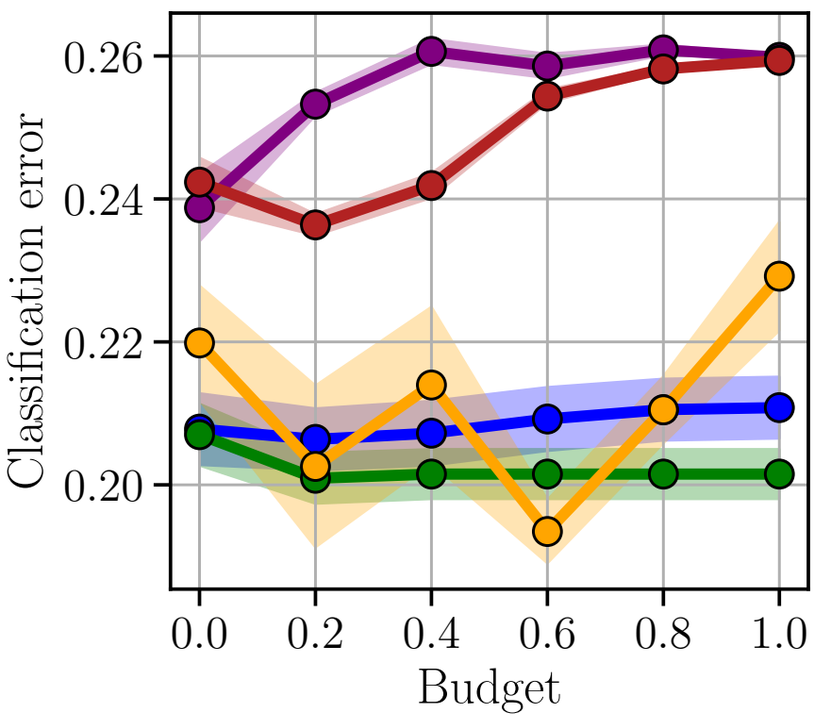
score

differentiable triage

Galaxy Zoo

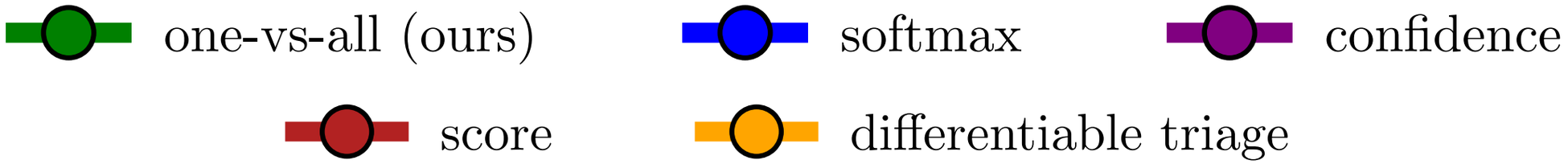


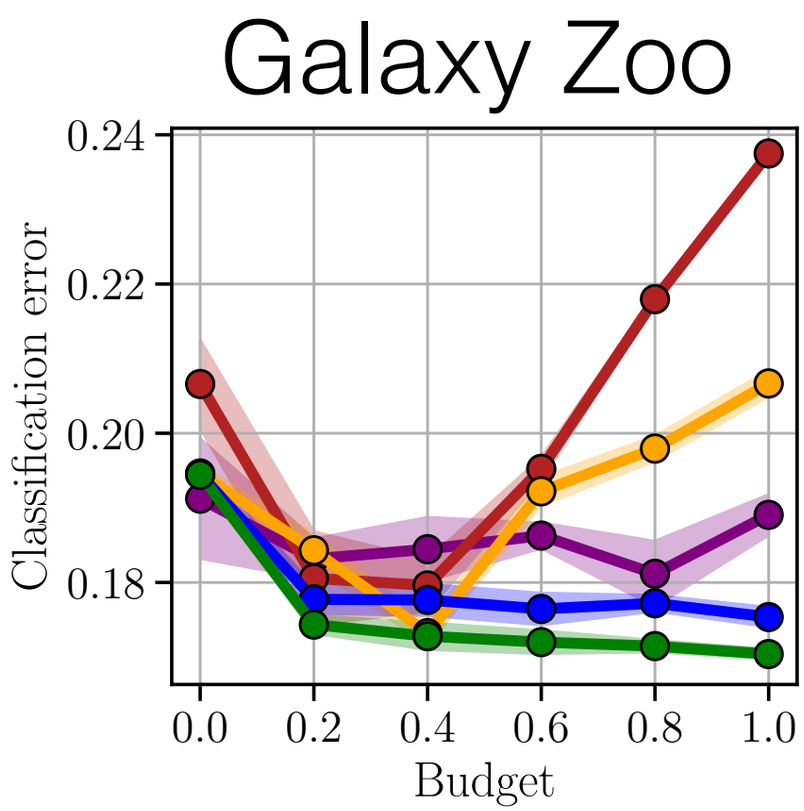
HAM10000

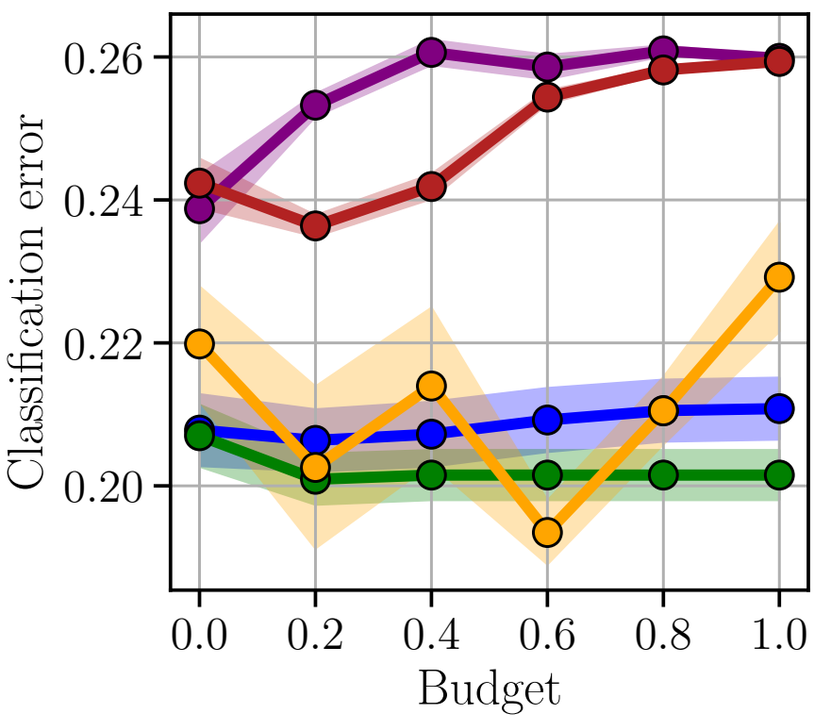


Hate Speech

[Davidson et al., ICWSM 2017]



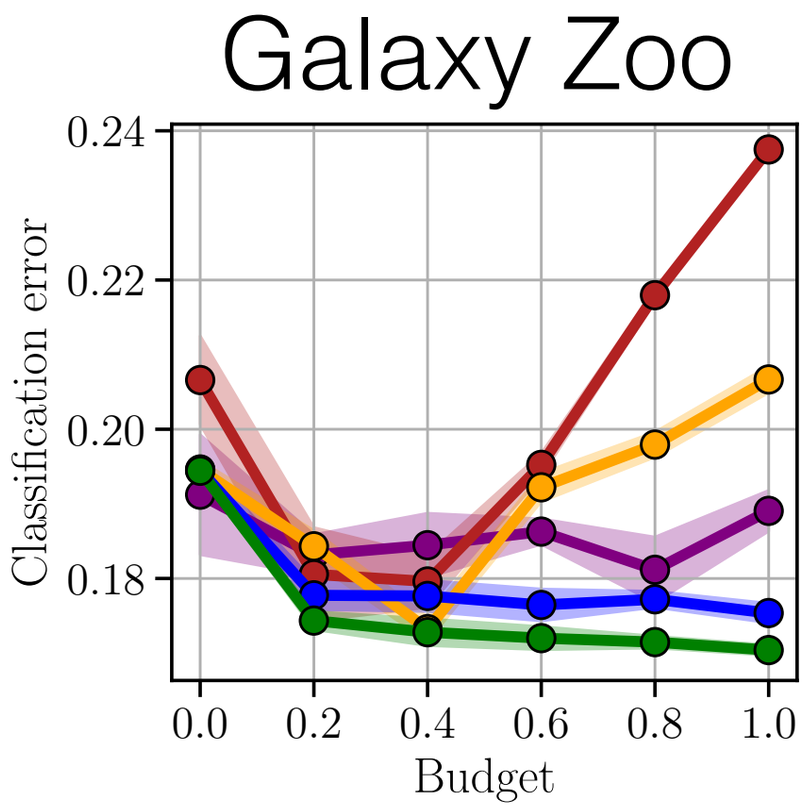


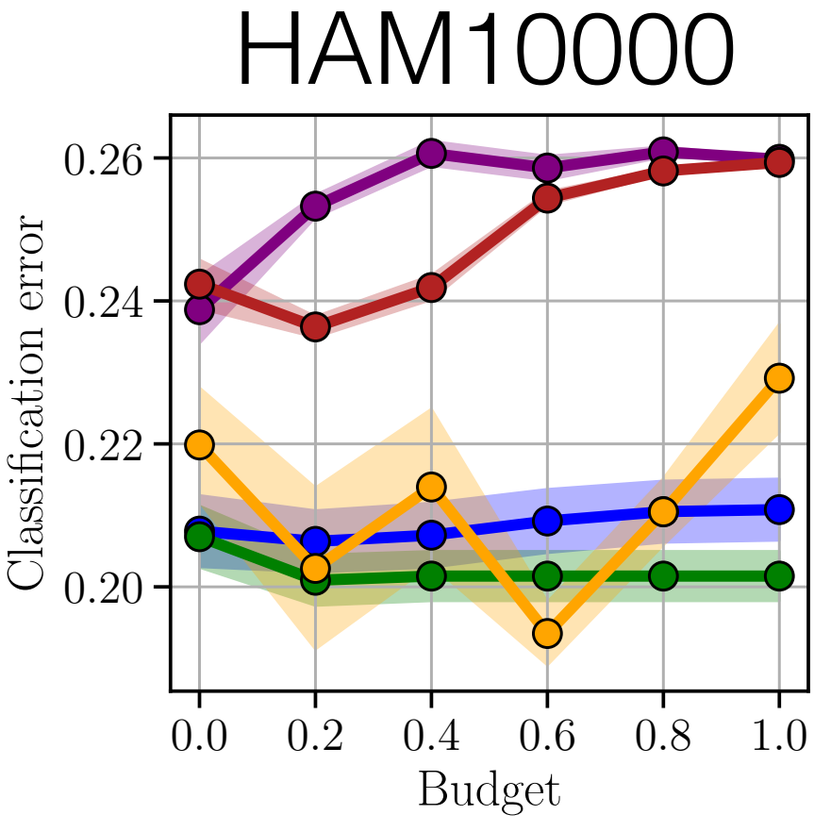


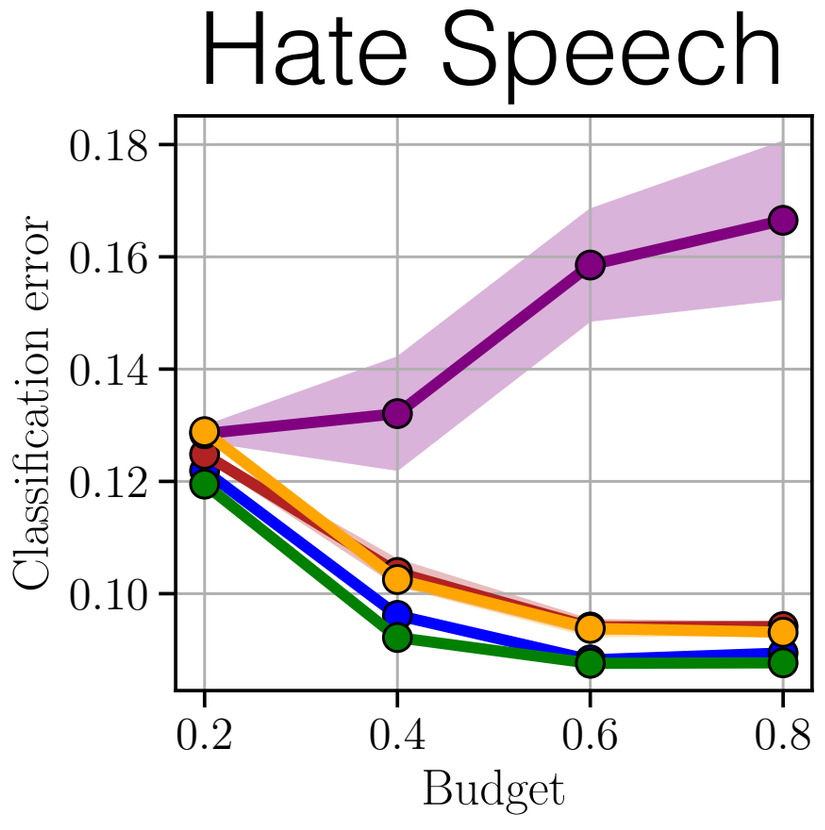
Hate Speech

[Davidson et al., ICWSM 2017]

one-vs-all (ours)
softmax
confidence
score
differentiable triage







one-vs-all (ours)

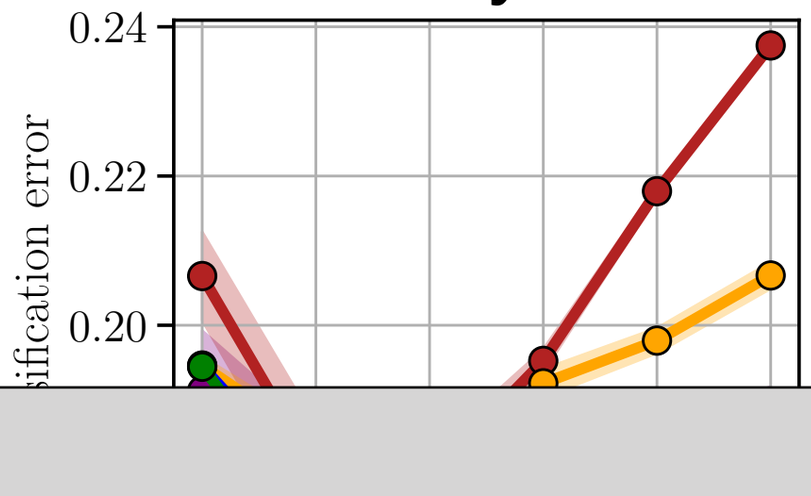
softmax

confidence

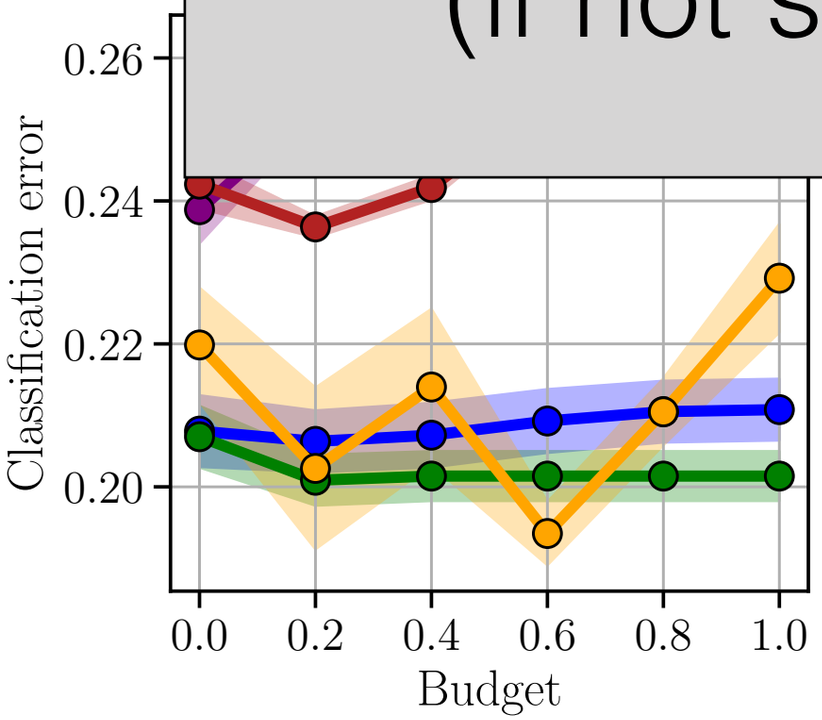
score

differentiable triage

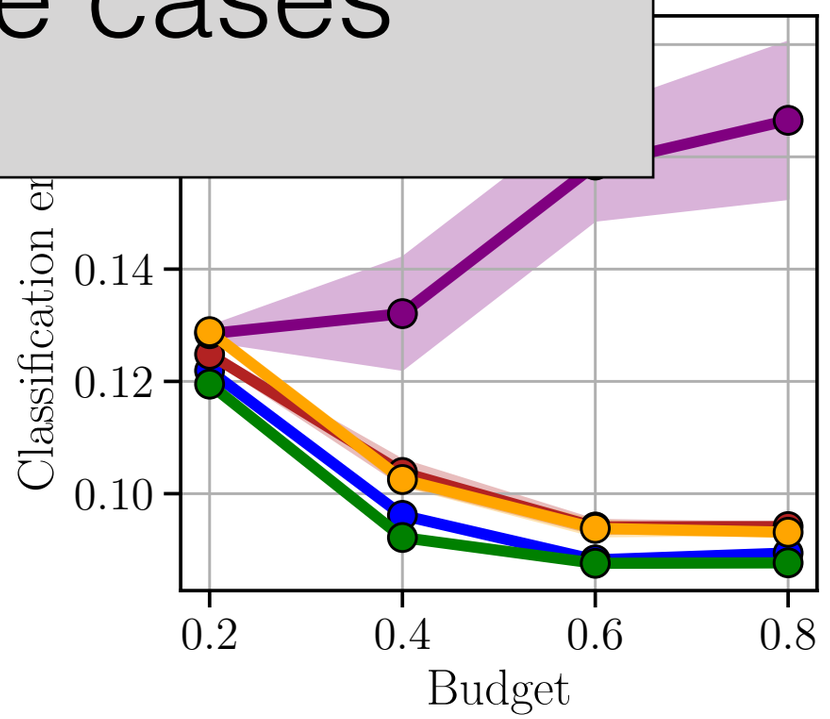
Galaxy Zoo



One-vs-all models are competitive
(if not superior) in all three cases



ech



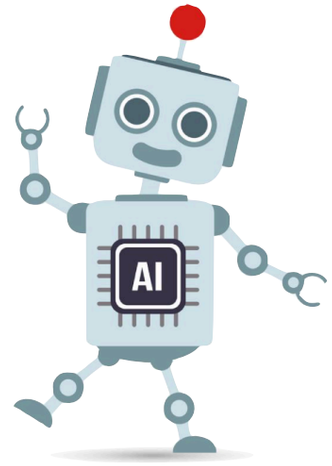
Extension

Learning to Defer to **Multiple** Experts

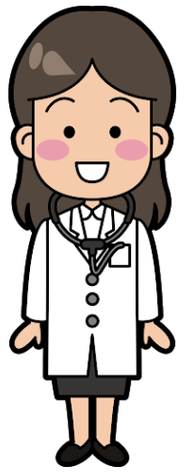
INPUT FEATURES

REJECTOR

meta-classifier



Classifier

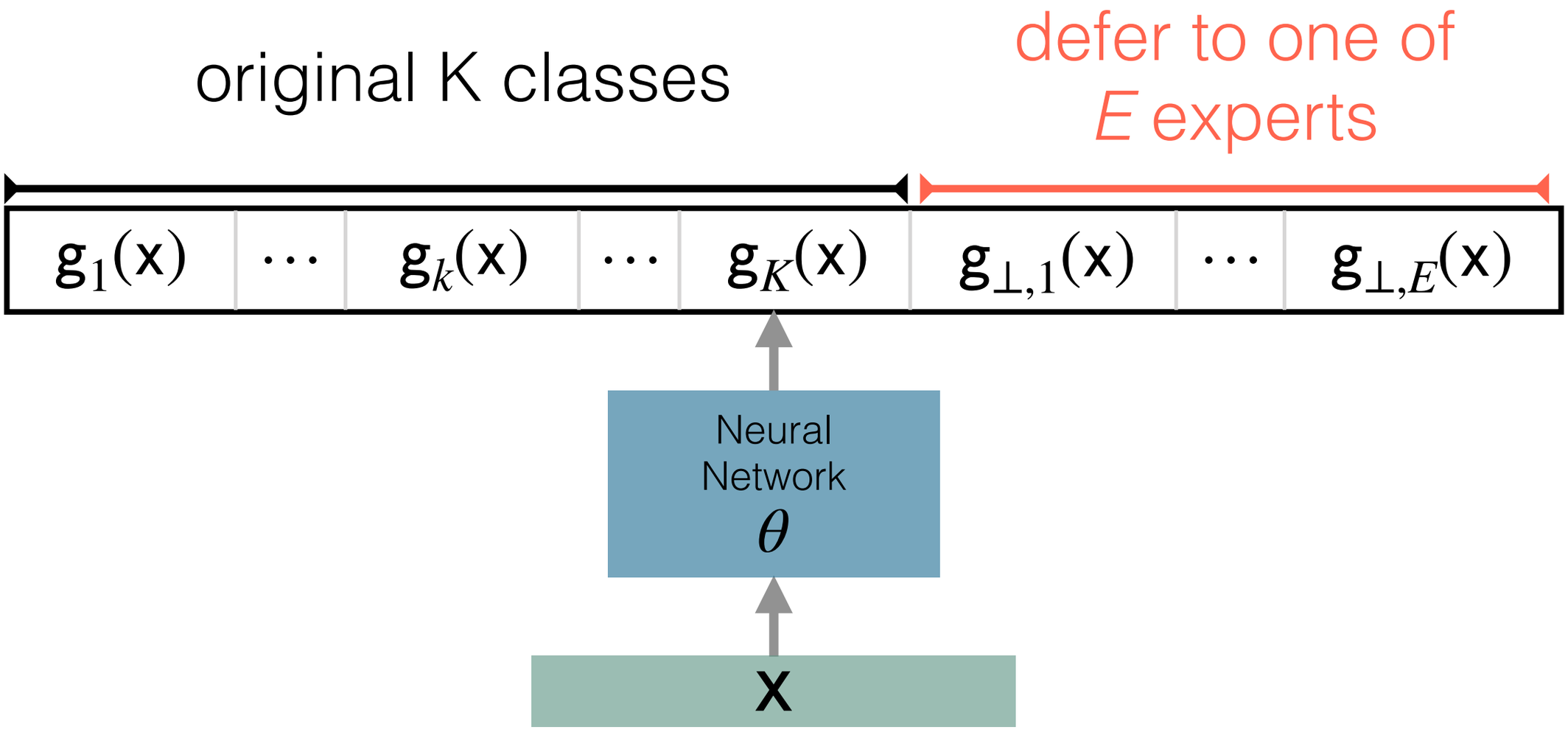


Expert #1

Expert #2

Expert #3

Learning-to-Defer to Multiple Experts



$$\ell(\theta; \mathfrak{D}) =$$

$$-\sum_n \left(\log \mathsf{p}_{y_n}(\mathsf{x}_n) \;+\; \sum_e \mathbb{I}[\mathsf{y}_n = \mathsf{m}_{n,e}] \;\log \mathsf{p}_{\perp,e}(\mathsf{x}_n) \right)$$

classifier loss

rejector loss

only if *e*-th expert is correct

Calibration: Is the system a good forecaster?

$$\mathbb{P}(\mathbf{m}_e = \mathbf{y} \mid \mathbf{x})$$

Does the rejector correctly estimate the *e*-th expert's chance of being correct?

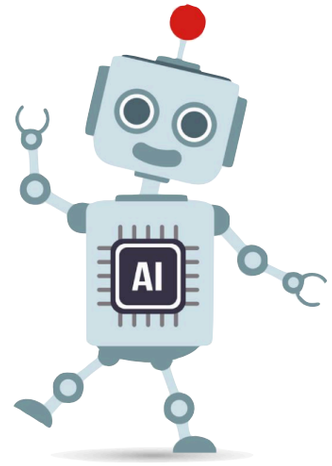
Calibration: Is the system a good forecaster?

$$\mathbb{P}(\mathsf{m}_e = \mathsf{y} \,|\, \mathsf{x}) \;=\; \frac{\mathsf{p}^{*}_{\perp,e}(\mathsf{x})}{1 - \sum_{j=1}^{E} \mathsf{p}^{*}_{\perp,j}(\mathsf{x})}$$

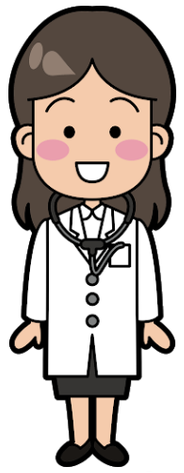
INPUT FEATURES

REJECTOR

meta-classifier



Classifier

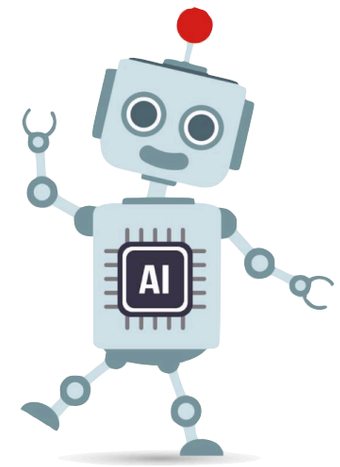


Expert #1

Expert #2

Expert #3

Training: random experts



Classifier

Random #1

INPUT FEATURES

REJECTOR

meta-classifier

Random #2

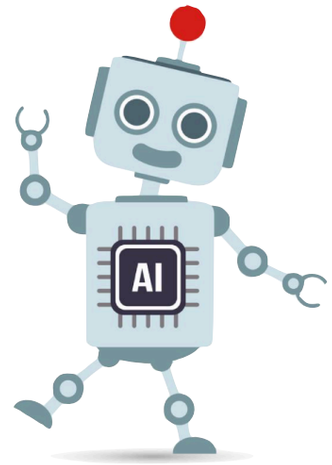
Random #3

Testing: one random

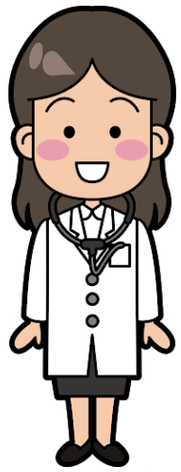
INPUT FEATURES

REJECTOR

meta-classifier



Classifier

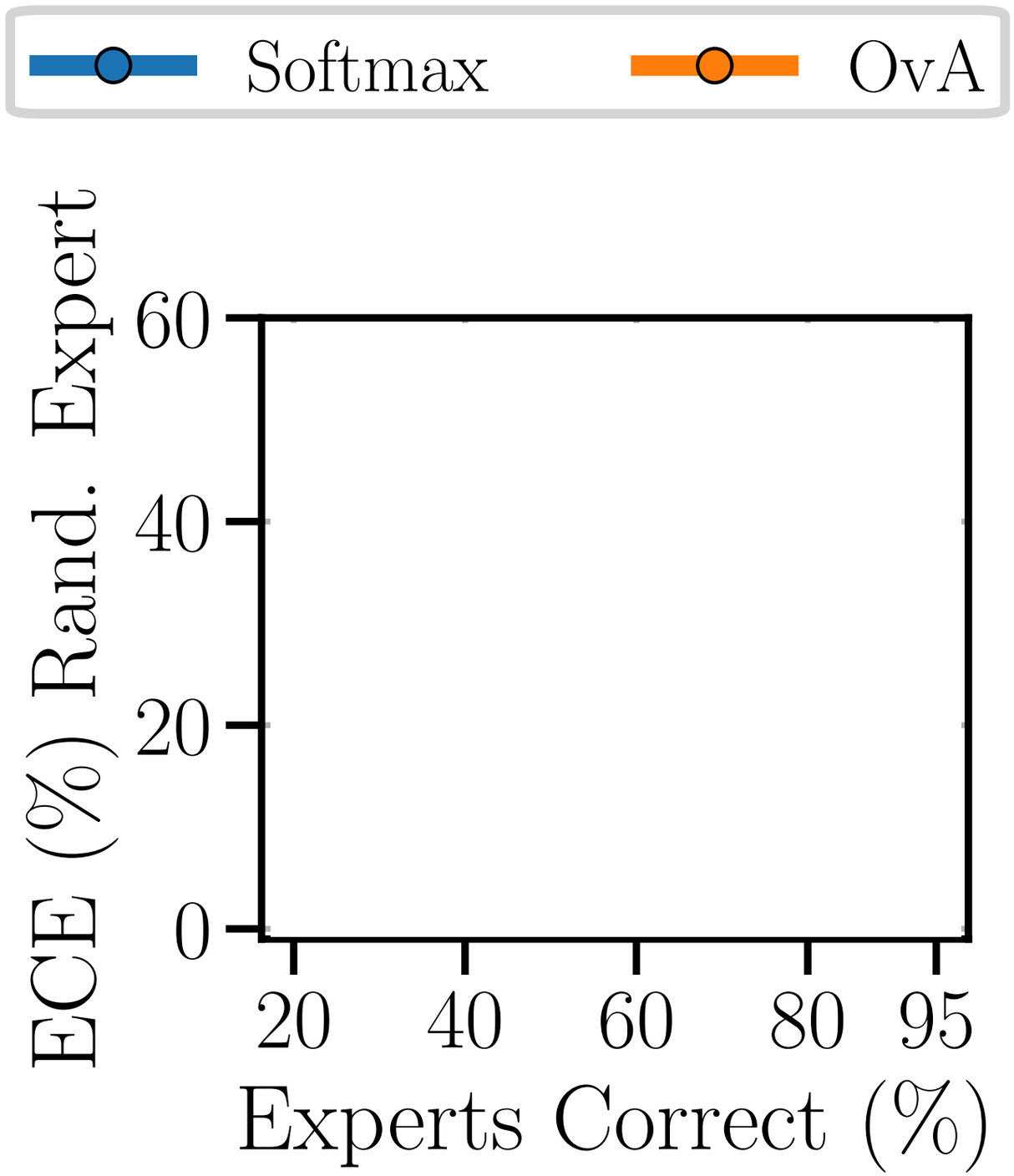


Expert #1

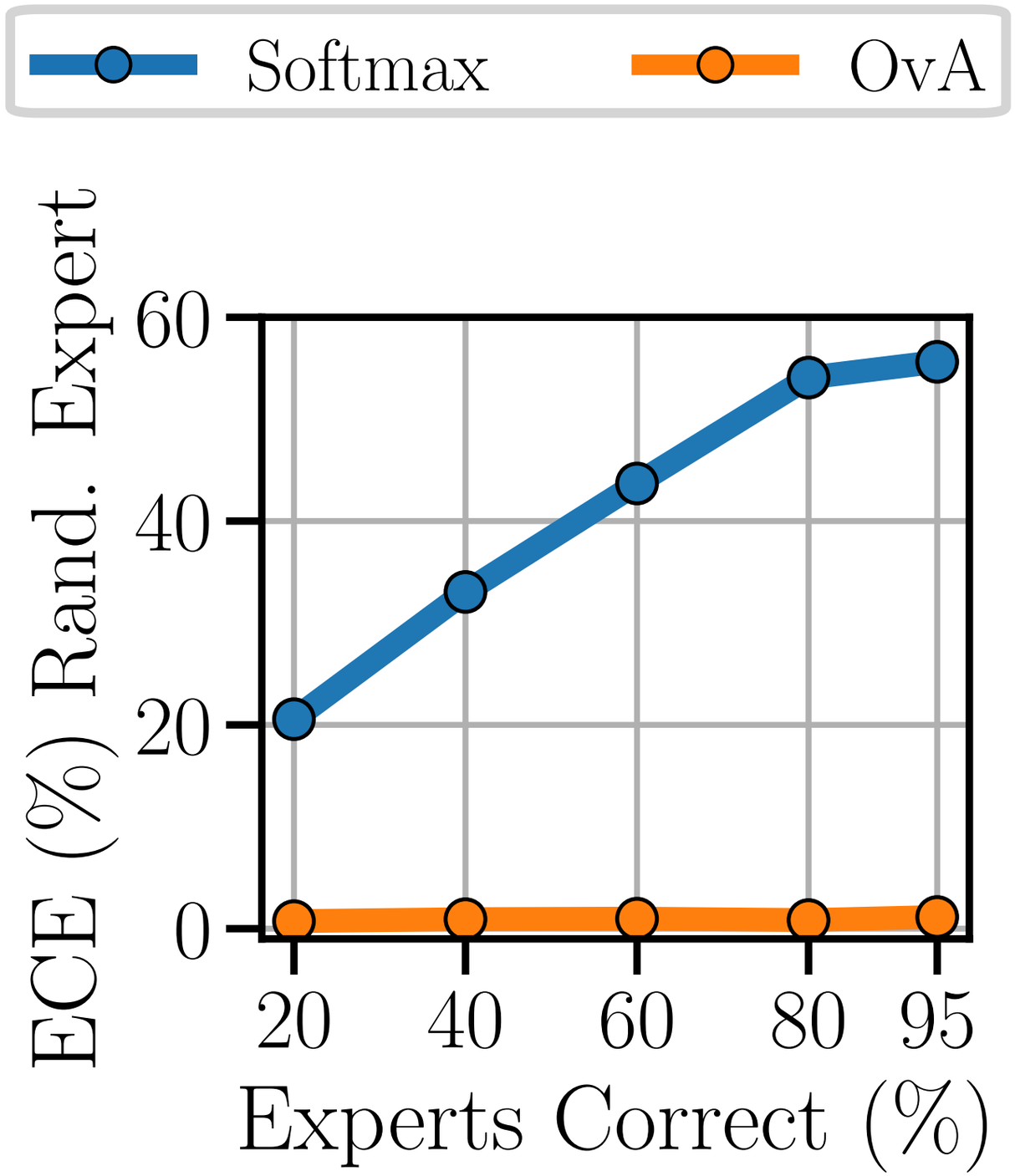
Expert #2

Random #3

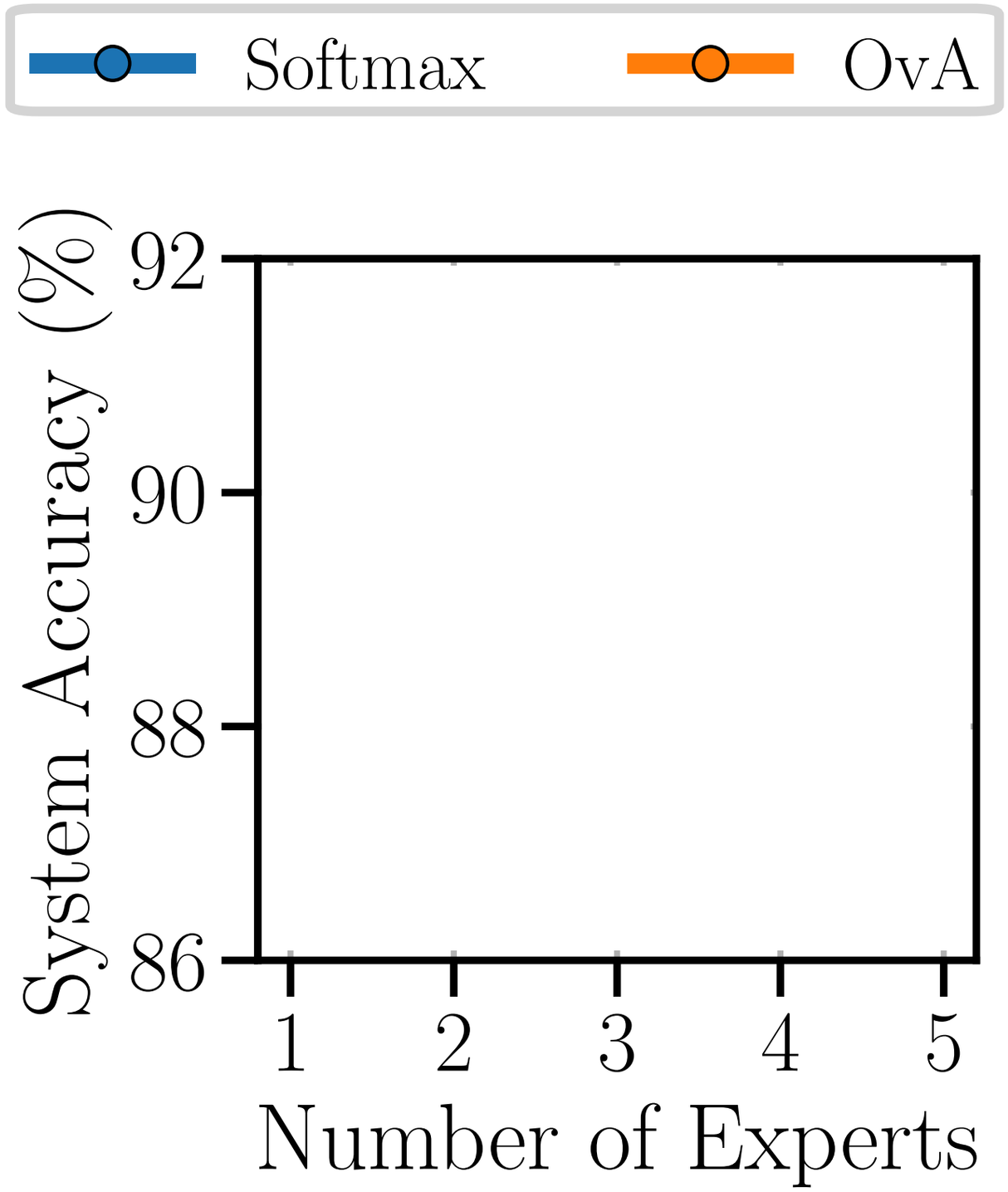
CIFAR-10: Expert Dependence



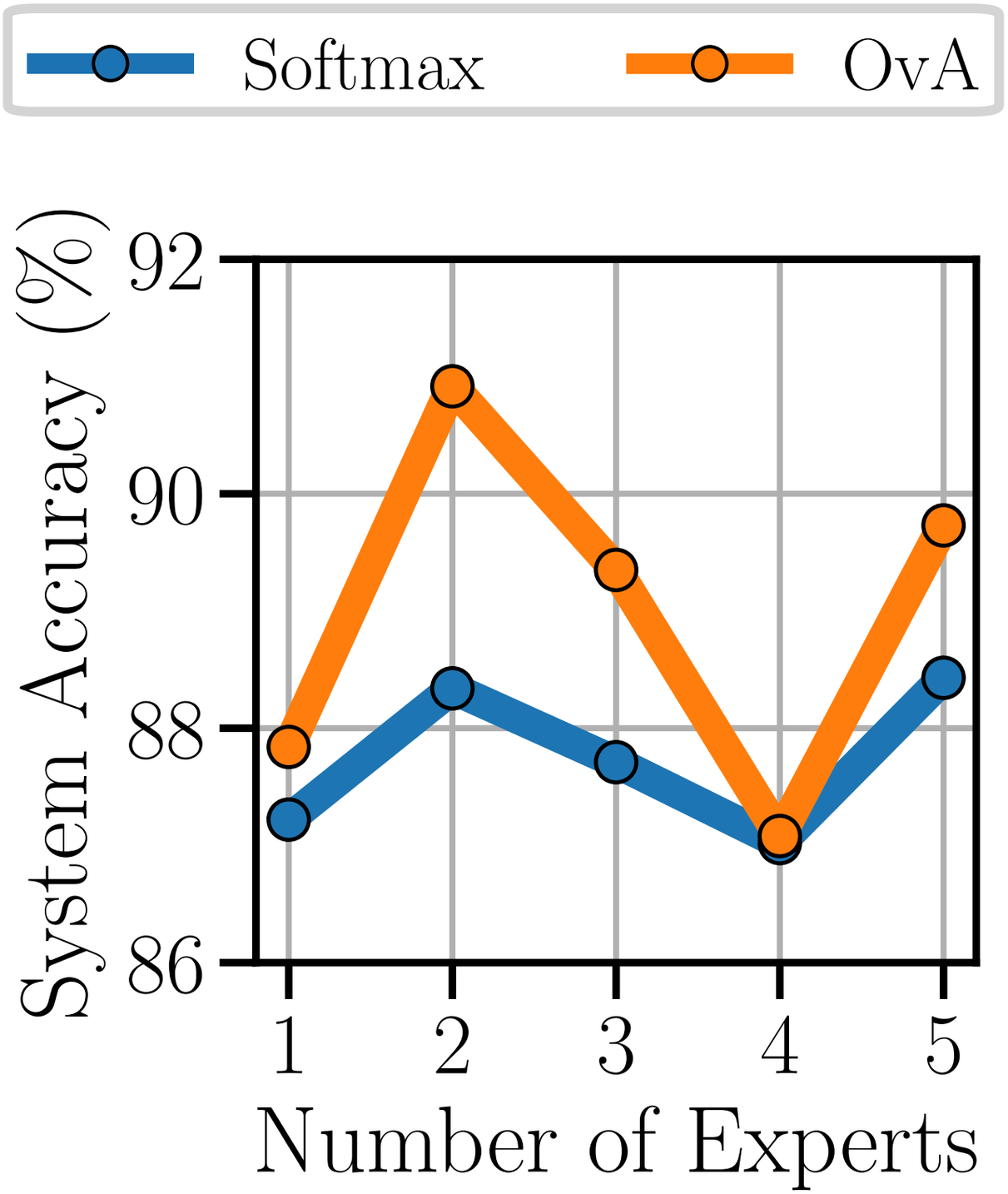
CIFAR-10: Expert Dependence



CIFAR-10: Specialized Experts



CIFAR-10: Specialized Experts



Summary

- ⊗ Softmax-based losses can produce invalid estimators for expert correctness.

- ⊗ One-vs-all formulation retains theoretical and practical benefits while having calibrated estimates of expert correctness.

- ⊗ Problem for softmax gets worse when there are multiple experts.

Future Work

⊗ Heavy dependence on supervised data.

$$\mathfrak{D} = \left\{ \mathsf{x}_n, \mathsf{y}_n, \mathsf{m}_n \right\}_{n=1}^{N}$$

⊗ Calibration will be essential for active learning, experimental design, missing data…

Paper

Code

Funding

Thank you. Questions?