Are Uncertainty Capabilities of Evidential Deep Learning a Mirage?

Maohao Shen[1*], J. Jon Ryu[1*], Soumya Ghosh[2],

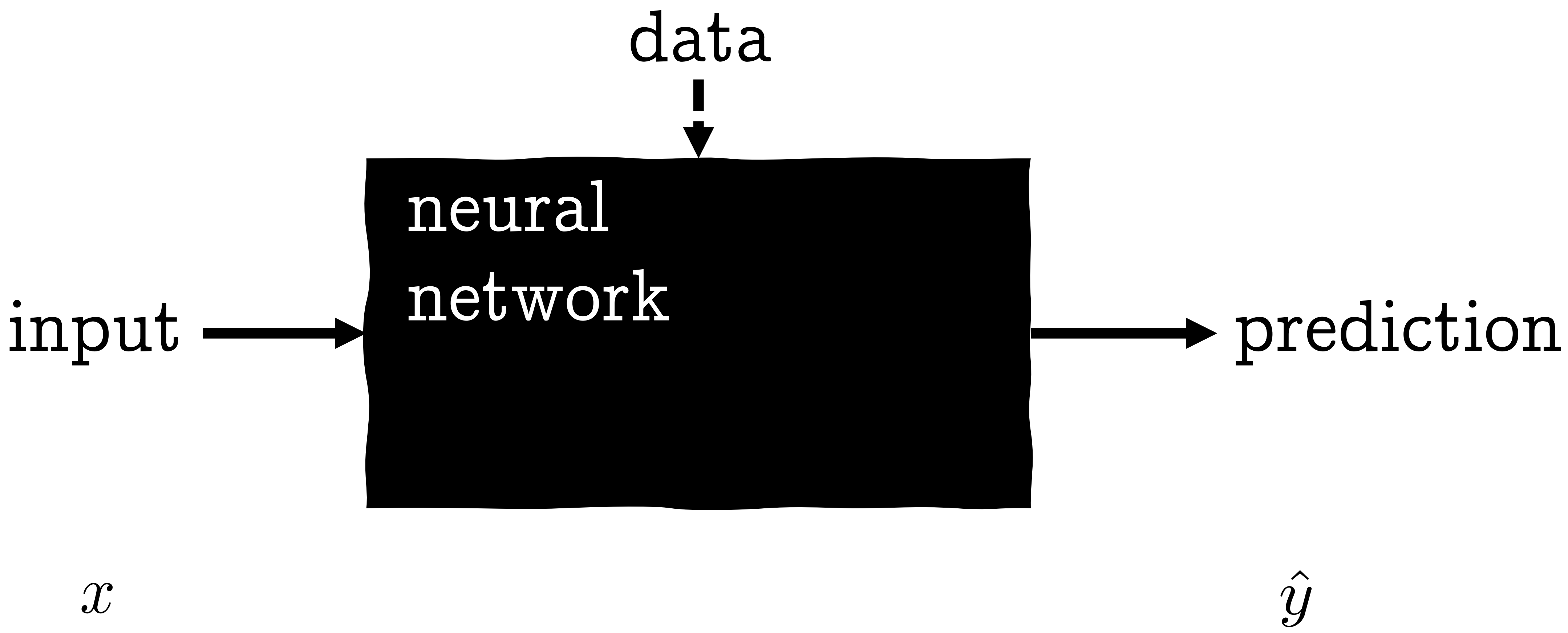
Yuheng Bu[3], Prasanna Sattigeri[2], Subhro Das[2], Gregory W. Wornell[1]

[1]MIT

[2]MIT-IBM Watson AI Lab, IBM Research

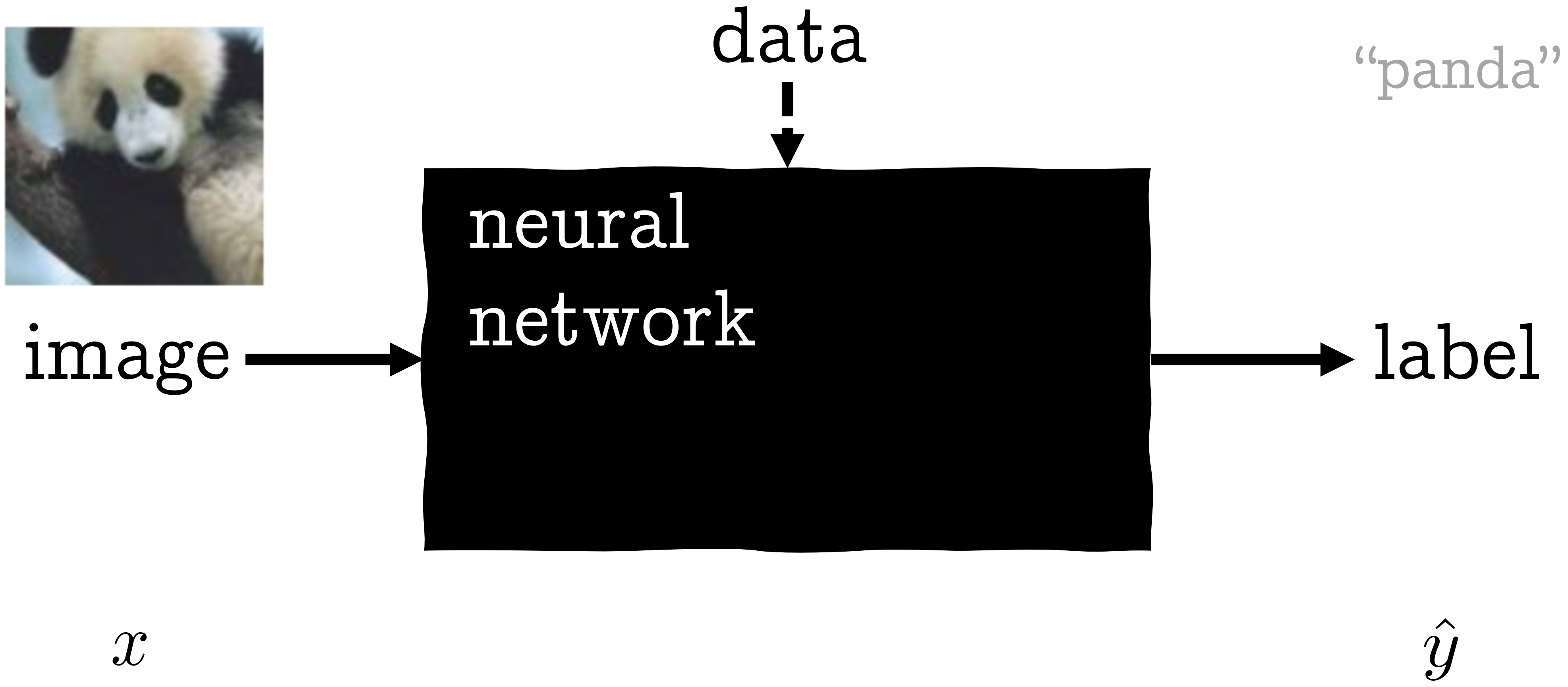
[3]University of Florida

Black-Box Prediction is Unreliable



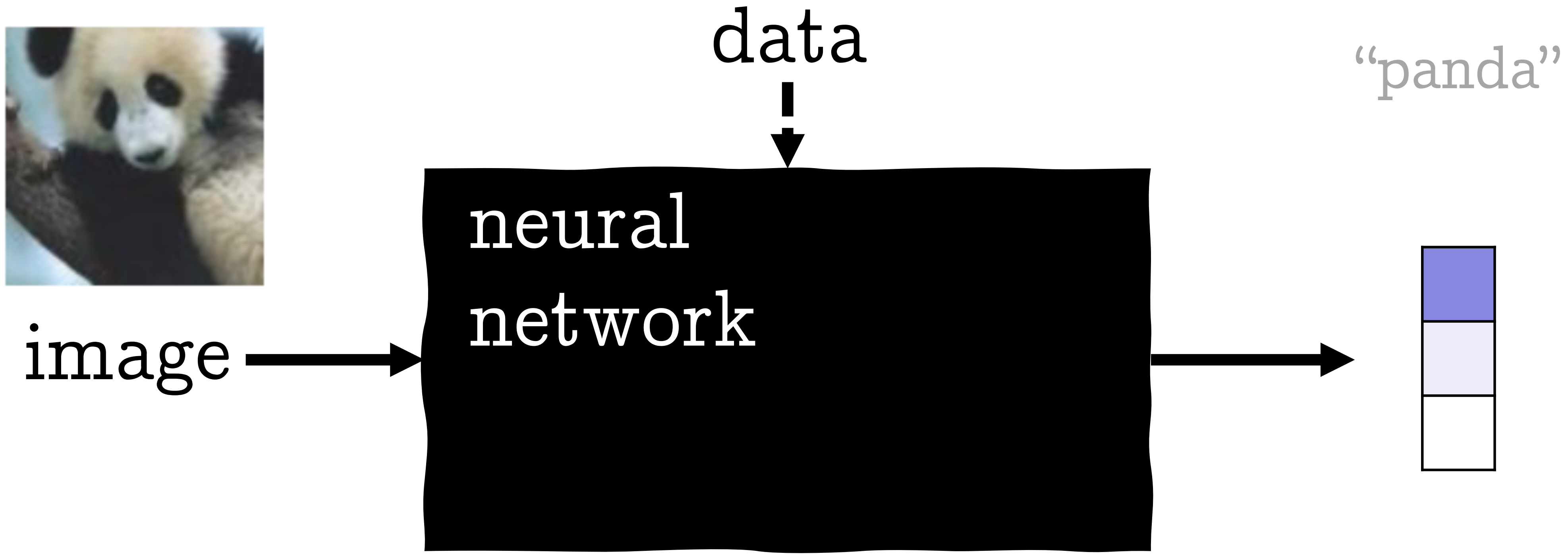
- Neural-network predictors are highly accurate...

Black-Box Prediction is Unreliable



- Neural-network predictors are highly accurate...

Black-Box Prediction is Unreliable

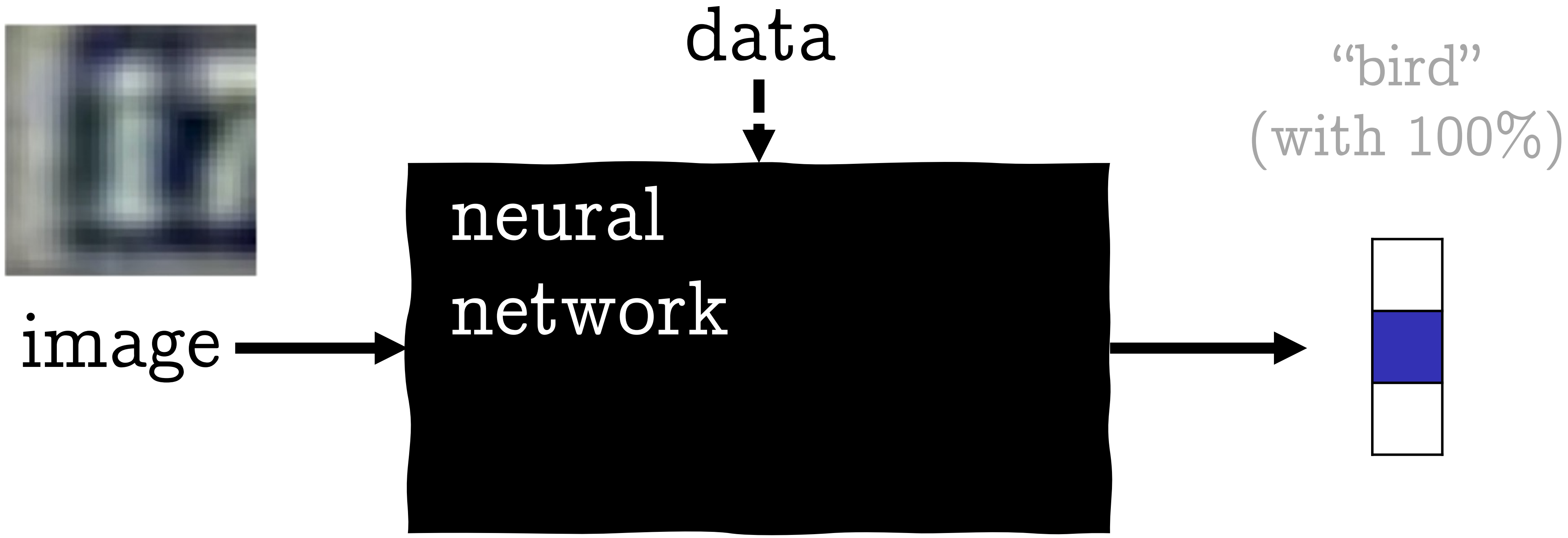


x

$$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$$

- Neural-network predictors are highly accurate...

Black-Box Prediction is Unreliable

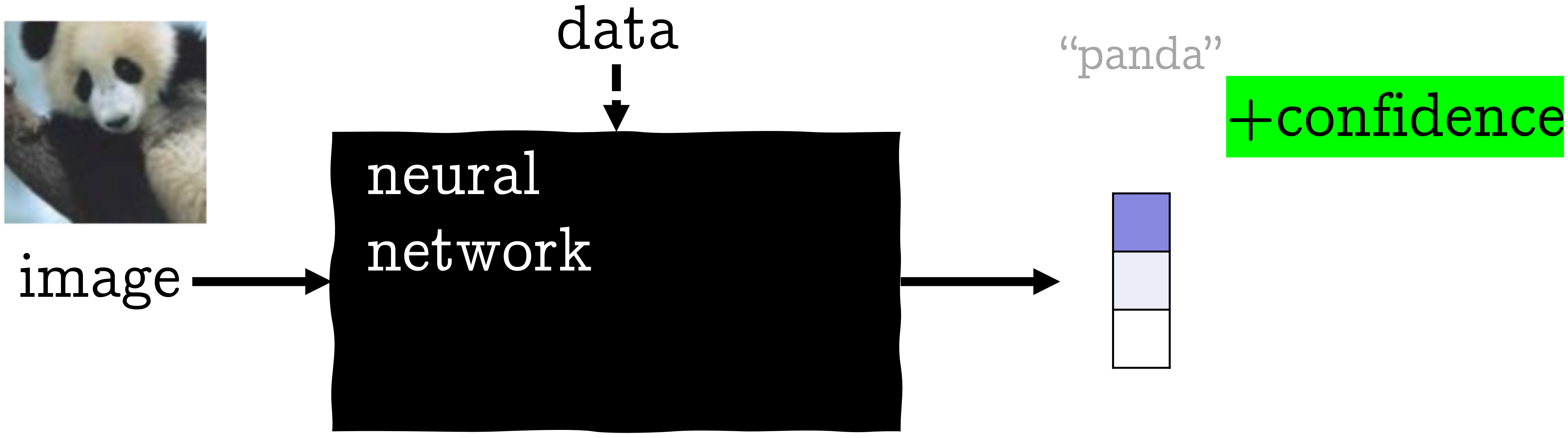


x

$$\boldsymbol{\pi}_{\psi}(x) = (p_{\psi}(y|x))_{y=1}^{C}$$

- Neural-network predictors are highly accurate...
- But often overconfident for out-of-distribution (OOD) data

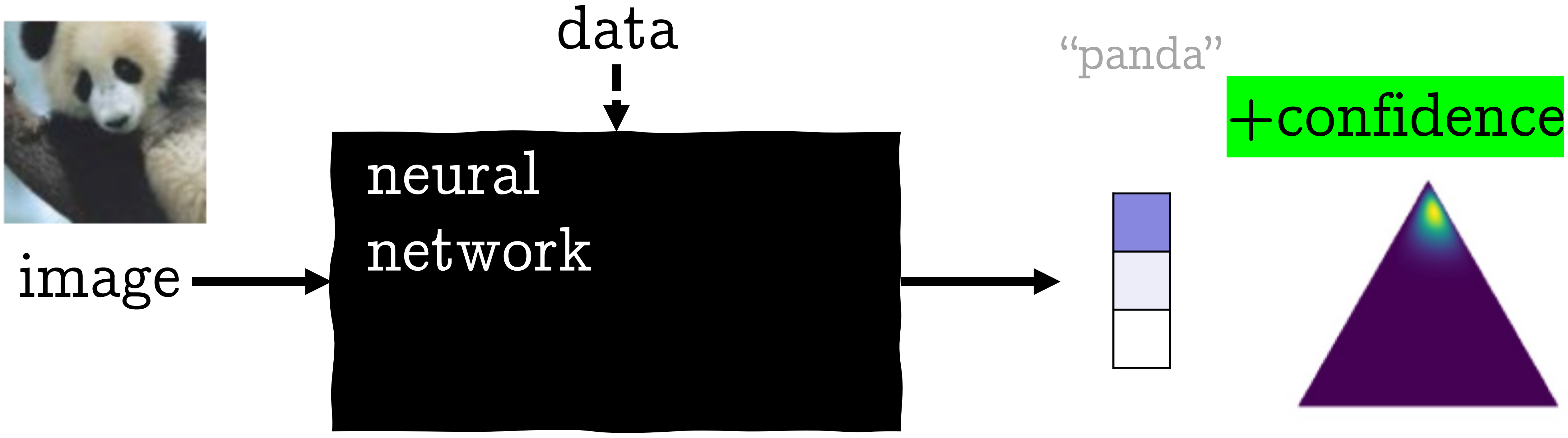
We Need Uncertainty Quantification!



x

$$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$$

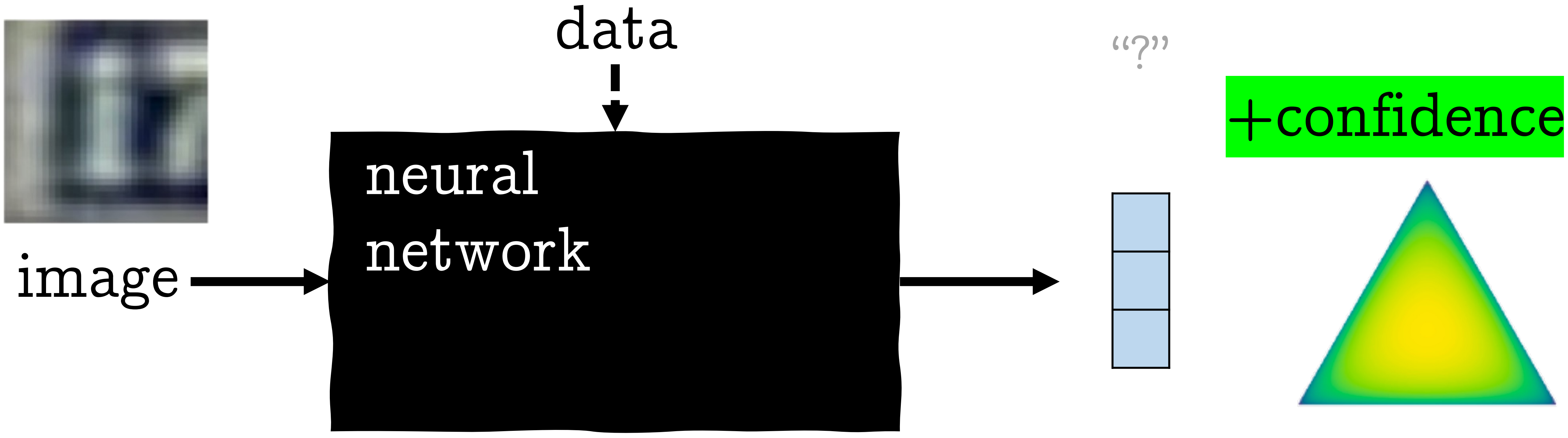
We Need Uncertainty Quantification!



x

$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$ $p_\psi(\boldsymbol{\pi}|x)$

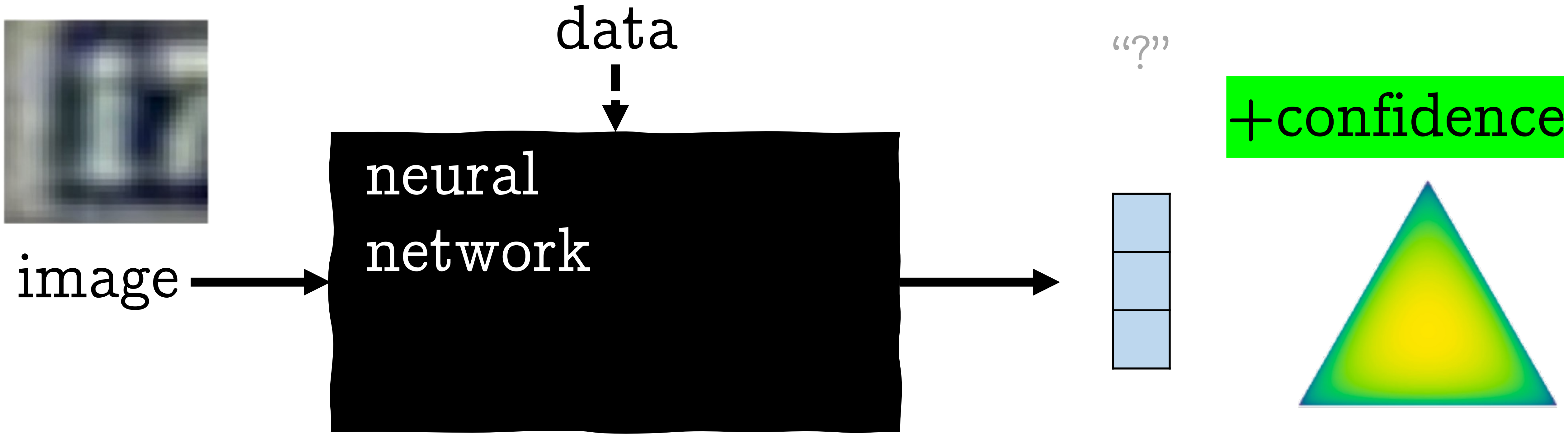
We Need Uncertainty Quantification!



x

$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$ $p_\psi(\boldsymbol{\pi}|x)$

We Need Uncertainty Quantification!



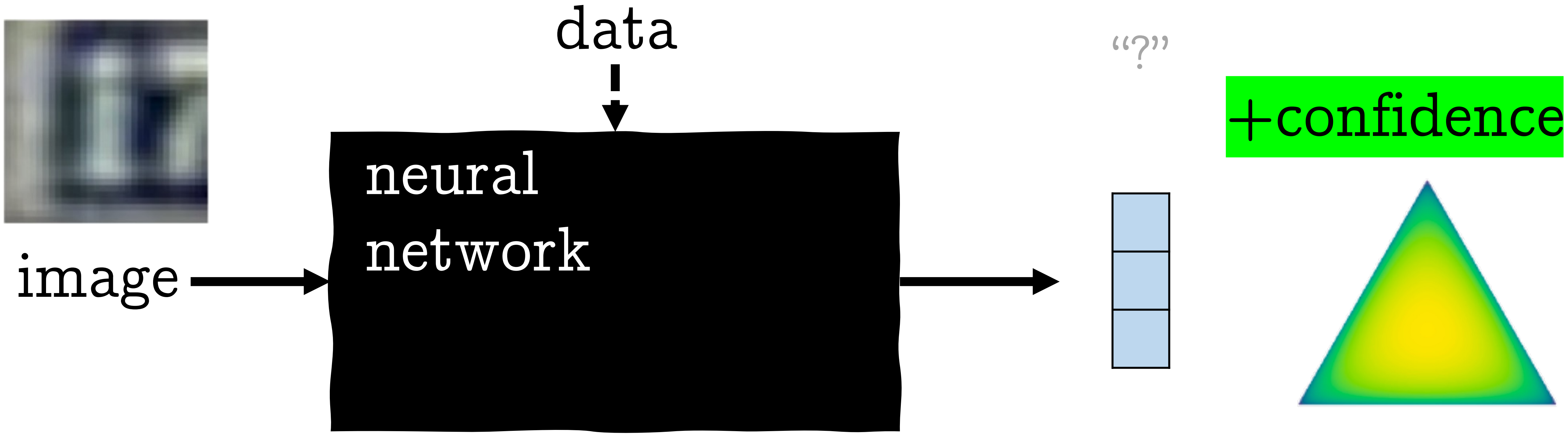
x

$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$ $p_\psi(\boldsymbol{\pi}|x)$

Different ways of inducing $p_\psi(\boldsymbol{\pi}|x)$

- **Bayesian methods**: variational inference, MCMC, Monte Carlo Dropout, ...
- **Frequentist methods**: jackknife, bootstrap, ...
- **Ensemble methods**

We Need Uncertainty Quantification!



x

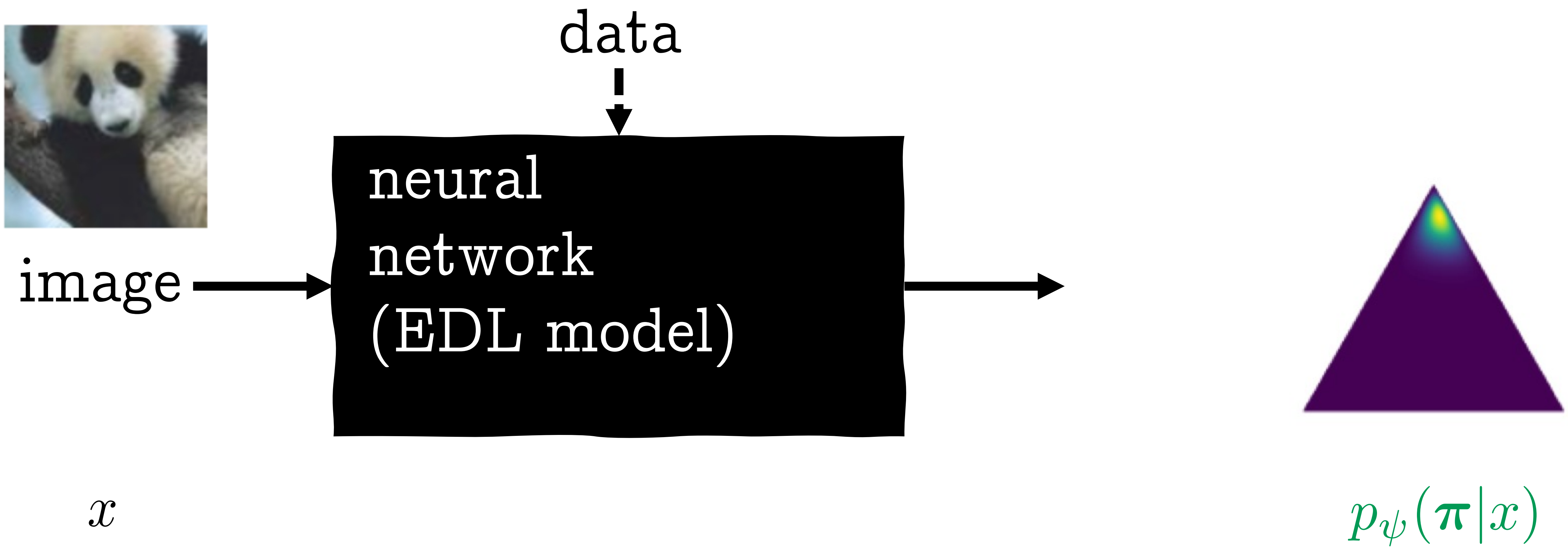
$\boldsymbol{\pi}_\psi(x) = (p_\psi(y|x))_{y=1}^{C}$ $p_\psi(\boldsymbol{\pi}|x)$

Different ways of inducing $p_\psi(\boldsymbol{\pi}|x)$

- Bayesian methods: variational inference, MCMC, Monte Carlo Dropout, ...
- Frequentist methods: jackknife, bootstrap, ...
- Ensemble methods

Computationally Inefficient!

Evidential Deep Learning (EDL)



- Directly train a **single** neural network that outputs $p_\psi(\boldsymbol{\pi}|x)$
- Empirical successes for downstream tasks (e.g., OOD detection)
- Lack of theoretical understanding
- Recent works have reported spurious behaviors

Demystifying EDL Methods

- Q1. What do EDL methods learn as uncertainty?
- Q2. Why are the EDL methods successful?
- Q3. How can we make EDL methods more reliable?

Unifying EDL Objectives: A New Taxonomy

Method (name of loss)
FPriorNet (F-KL loss) [7]
RPriorNet (R-KL loss) [8]
EDL (MSE loss) [10]
Belief Matching (VI loss) [12, 13]
PostNet (UCE loss) [15]
NatPN (UCE loss) [20]

What's the common principle behind all these objectives?

Unifying EDL Objectives: A New Taxonomy

Method (name of loss)
FPriorNet (F-KL loss) [7]
RPriorNet (R-KL loss) [8]
EDL (MSE loss) [10]
Belief Matching (VI loss) [12, 13]
PostNet (UCE loss) [15]
NatPN (UCE loss) [20]

What's the common principle behind all these objectives?

$$\mathcal{L}(\psi) := \underbrace{\mathbb{E}_{p(x,y)}[D(\underbrace{\boxed{p^{(\nu)}(\boldsymbol{\pi}|y)}}_{\text{“fixed” uncertainty target}}, \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{in-distribution objective}} + \underbrace{\gamma_{\mathsf{ood}}\mathbb{E}_{p_{\mathsf{ood}}(x)}[D(p(\boldsymbol{\pi}), \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{OOD objective}}$$

EDL model

Unifying EDL Objectives: A New Taxonomy

Method (name of loss)	likelihood	$D(\cdot,\cdot)$	prior $\boldsymbol{\alpha}_0$	γ_{ood}	$\boldsymbol{\alpha}_\psi(x)$ parameterization
FPriorNet (F-KL loss) [7]	categorical	fwd. KL	$= \mathbb{1}_C$	> 0	direct
RPriorNet (R-KL loss) [8]	categorical	rev. KL	$= \mathbb{1}_C$	> 0	direct
EDL (MSE loss) [10]	Gaussian	rev. KL	$= \mathbb{1}_C$	$= 0$	direct
Belief Matching (VI loss) [12, 13]	categorical	rev. KL	$\in \mathbb{R}^C_{>0}$	$= 0$	direct
PostNet (UCE loss) [15]	categorical	rev. KL	$= \mathbb{1}_C$	$= 0$	density w/ single flow
NatPN (UCE loss) [20]	categorical	rev. KL	$= \mathbb{1}_C$	$= 0$	density w/ multiple flows

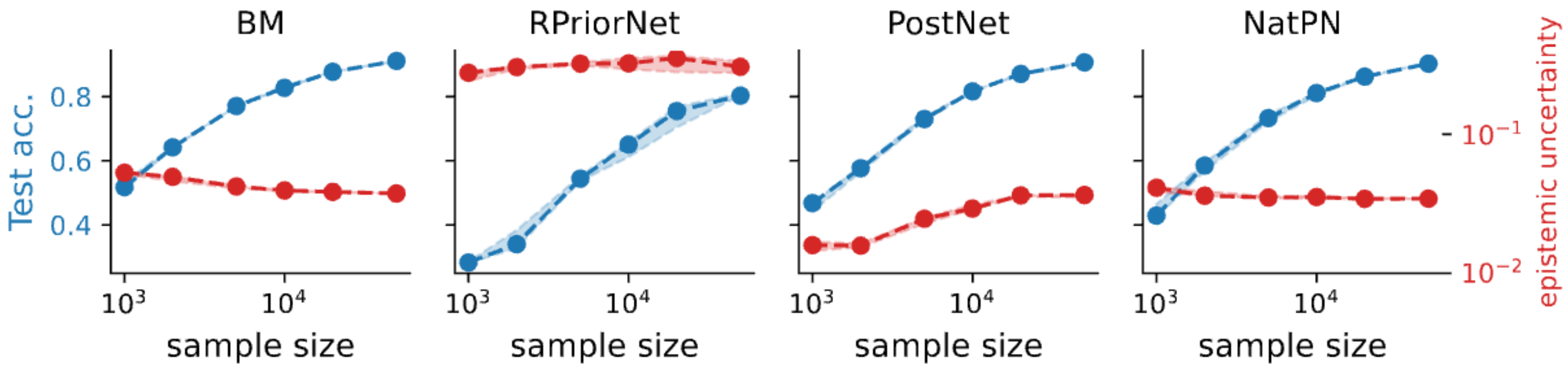
$$\mathcal{L}(\psi) := \underbrace{\mathbb{E}_{p(x,y)}[D(\boxed{p^{(\nu)}(\boldsymbol{\pi}|y)}, \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{in-distribution objective}} + \underbrace{\gamma_{\text{ood}}\mathbb{E}_{p_{\text{ood}}(x)}[D(p(\boldsymbol{\pi}), \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{OOD objective}}$$

"fixed"
uncertainty target

EDL model

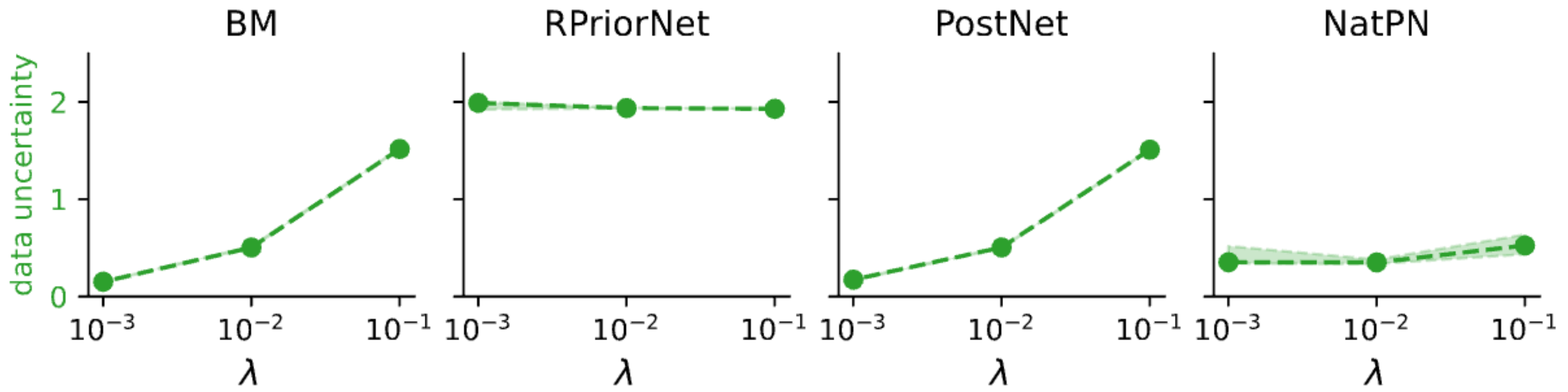
Explain Spurious Behaviors of Learned Uncertainties

Non-vanishing
learned epistemic uncertainty



(a) Epistemic Uncertainty: CIFAR10

Hyperparameter sensitive
learned aleatoric uncertainty



(b) Aleatoric Uncertainty: CIFAR10

Explain Success of EDL Methods

$$\mathcal{L}(\psi) := \underbrace{\mathbb{E}_{p(x,y)}[D(\boxed{p^{(\nu)}(\boldsymbol{\pi}|y)}, \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{in-distribution objective}} + \underbrace{\gamma_{\mathsf{ood}}\mathbb{E}_{p_{\mathsf{ood}}(x)}[D(p(\boldsymbol{\pi}), \boxed{p_\psi(\boldsymbol{\pi}|x)})]}_{\text{OOD objective}}$$

"fixed"
uncertainty target

EDL model

$$\approx$$

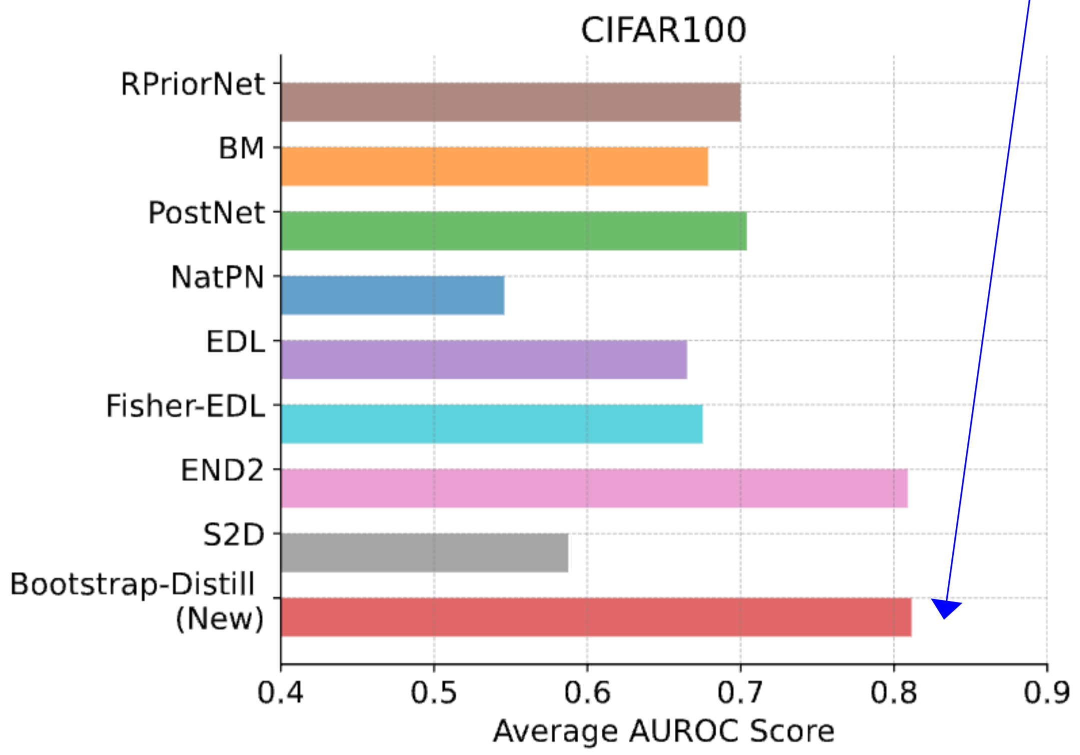
$$\boxed{\underbrace{-\mathbb{E}_{p(x,y)}[\log p_\psi(y|x)] + \tau\{\mathbb{E}_{p(x)}[\max(0, E_\phi(x) - m_{\mathsf{id}})^2]}_{\text{in-distribution objective}} + \underbrace{\mathbb{E}_{p_{\mathsf{ood}}(x)}[\max(0, m_{\mathsf{ood}} - E_\phi(x))^2]\},}_{\text{OOD objective}}}$$

- EDL methods can be better understood as EBM-based OOD detector

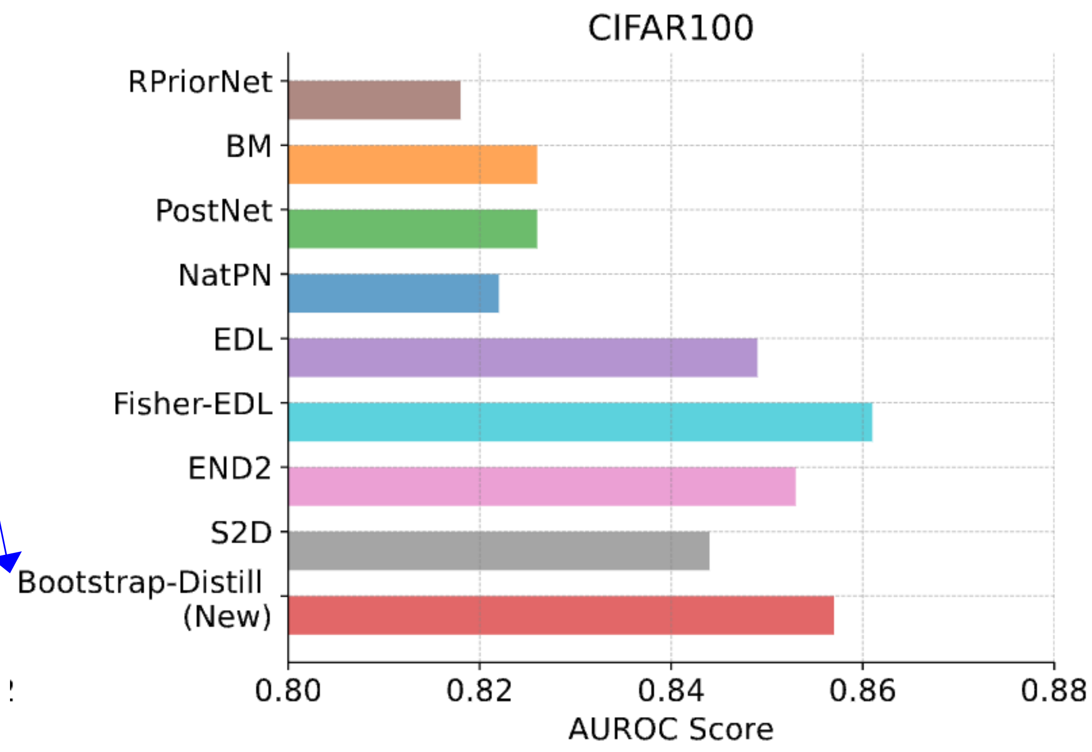
How to Improve EDL?

- All the issues of EDL arise from the ignorance of model stochasticity
- **Conjecture**: Incorporating external stochasticity is the key for meaningful UQ and EDL should be used for "distillation" for fast inference
- Show distilling randomness in Bootstrap can achieve SOTA performance

OOD detection



selective classification



Are Uncertainty Quantification Capabilities of Evidential Deep Learning a Mirage?

Maohao Shen[1,*], **J. Jon Ryu**[1*], **Soumya Ghosh**[2,†]
Yuheng Bu[3], **Prasanna Sattigeri**[2], **Subhro Das**[2], **Gregory W. Wornell**[1]

[1]Department of EECS, MIT, Cambridge, MA 02139
[2]MIT-IBM Watson AI Lab, IBM Research, Cambridge, MA 02142
[3]Department of ECE, University of Florida, Gainesville, FL 32611

{maohao,jongha,gww}@mit.edu,
{ghoshoso,prasanna}@us.ibm.com, subhro.das@ibm.com,
buyuheng@ufl.edu