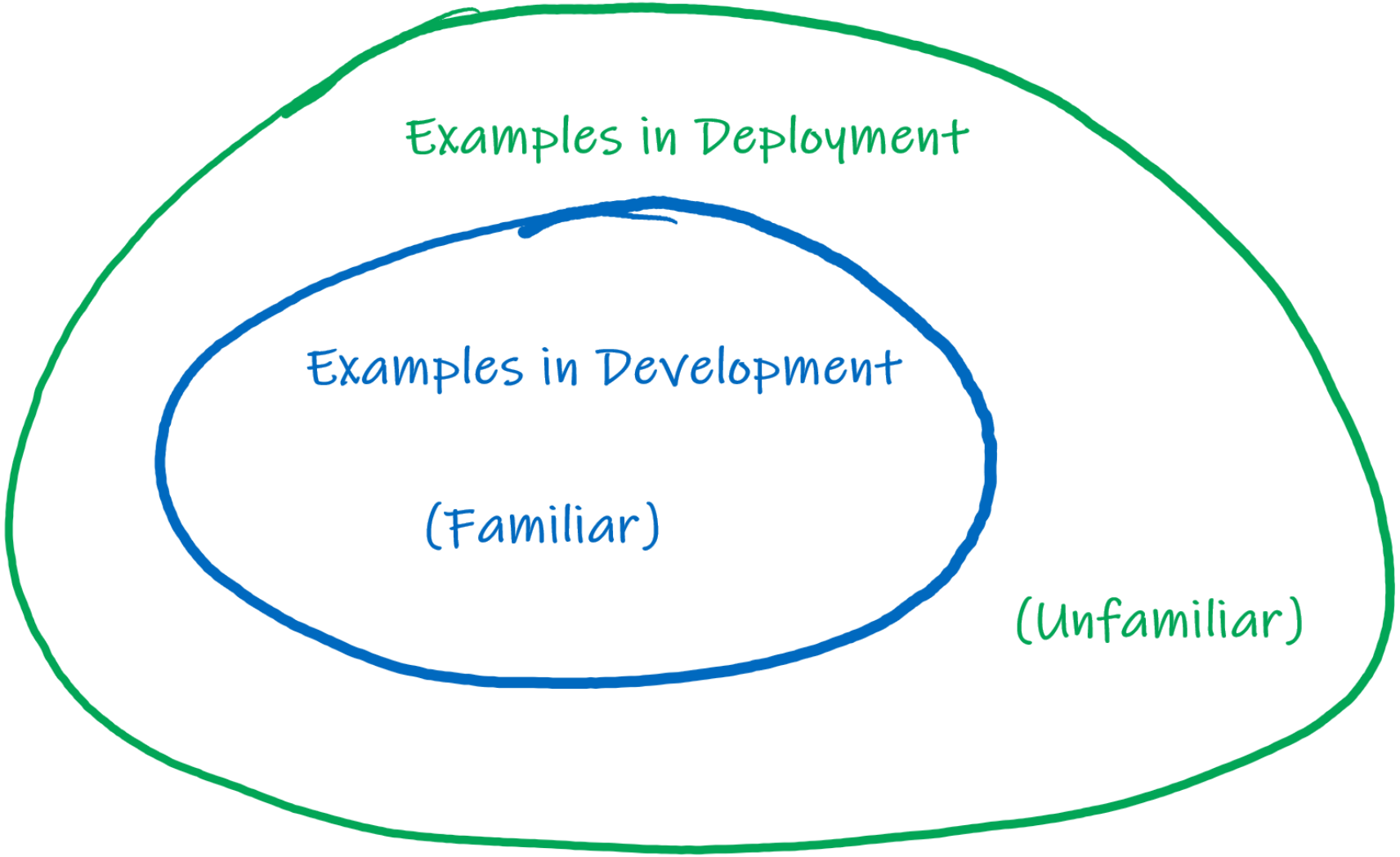
Improving Confidence Estimates for Unfamiliar Examples

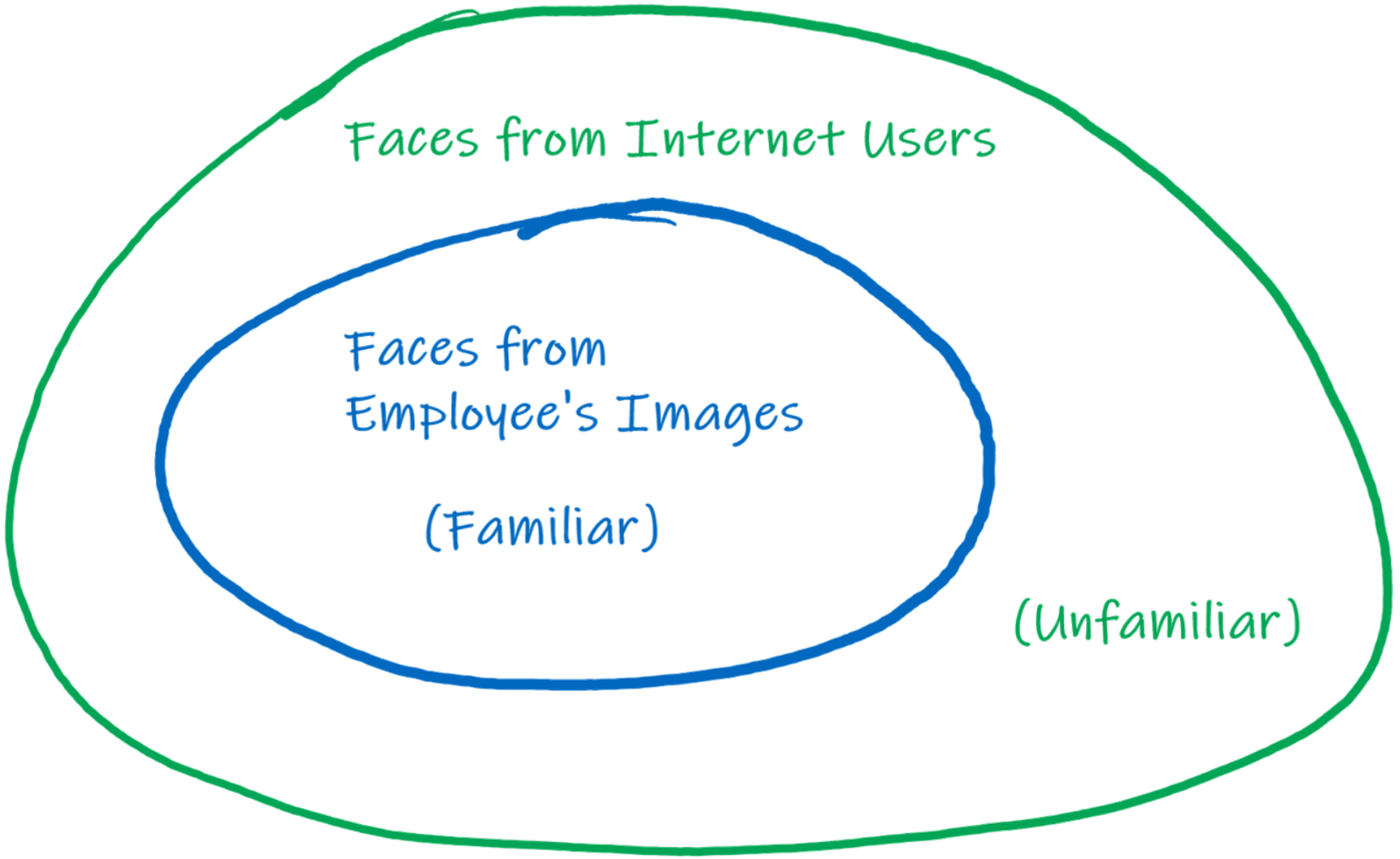
Zhizhong Li, Derek Hoiem
University of Illinois at Urbana-Champaign

CVPR 2020

Practical problem: different types of examples seen in deployment than in development of classifier



Practical problem: different types of examples seen in deployment than in development of classifier



Practical problem cases

Training: 18-59 yrs old

Unfamiliar test:
0-17 and 60+ yrs old

Error among
99% confident:

0.5%

6.0%

[redacted] driver dies in first fatal crash while using autopilot mode

The autopilot sensors on [redacted] failed to distinguish a white tractor-trailer crossing the highway against a bright sky

How well do deep network classifiers perform on unfamiliar samples?

Gender Classification

Familiar (18-59 yrs old)

Unfamiliar (older/younger)

Cat vs. Dog Classification

Familiar (some breeds)

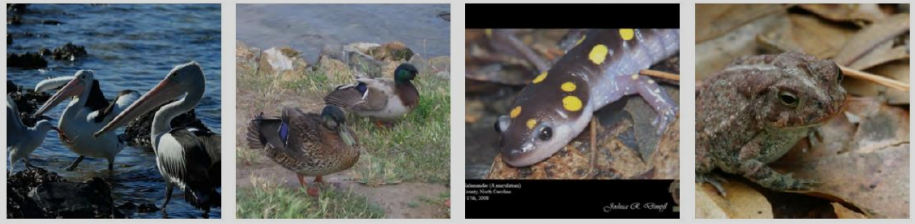
Unfamiliar (other breeds)

Animal* Classification

Familiar (some species)

Unfamiliar (other species)

* Birds, mammals, fishes, and herptiles from ImageNet.

	Training & Validation	Testing ONLY
Gender Classification	 Familiar (18-59 yrs old)	 Unfamiliar (older/younger)
Cat vs. Dog Classification	 Familiar (some breeds)	 Unfamiliar (other breeds)
Animal* Classification	 Familiar (some species)	 Unfamiliar (other species)

* Birds, mammals, fishes, and herptiles from ImageNet.

Gender Classification

Familiar (18-59 yrs old)

Cat vs. Dog Classification

Familiar (some breeds)

Animal* Classification

Familiar (some species)

Not concerned with irrelevant images

* Birds, mammals, fishes, and herptiles from ImageNet.

Gender Classification

Familiar (18-59 yrs old)

Unfamiliar (older/younger)

Cat vs. Dog Classification

Familiar (some breeds)

Unfamiliar (other breeds)

Animal* Classification

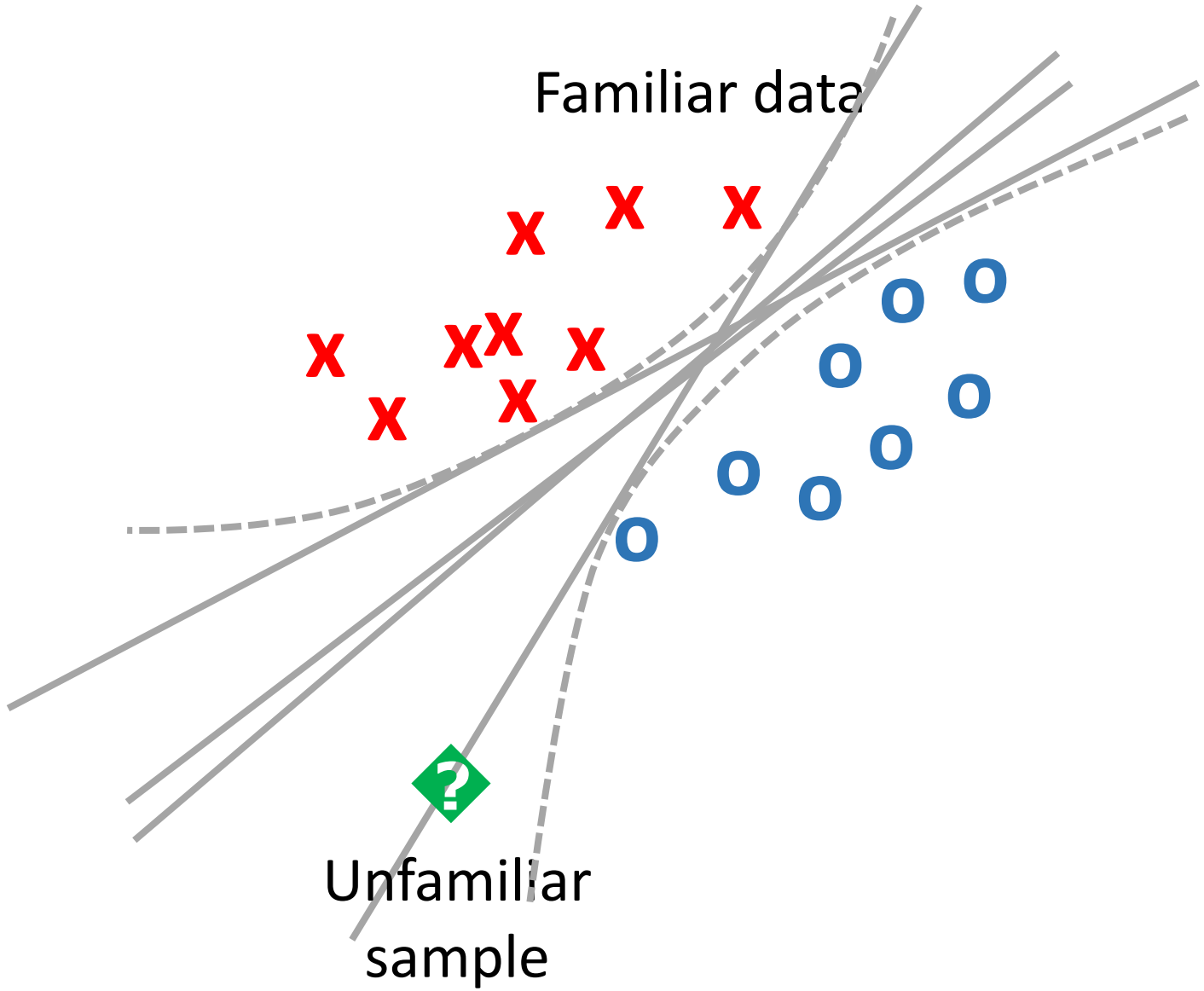
Familiar (some species)

Unfamiliar (other species)

* Birds, mammals, fishes, and herptiles from ImageNet.

Investigated methods

- Ensembles (cf. [1])



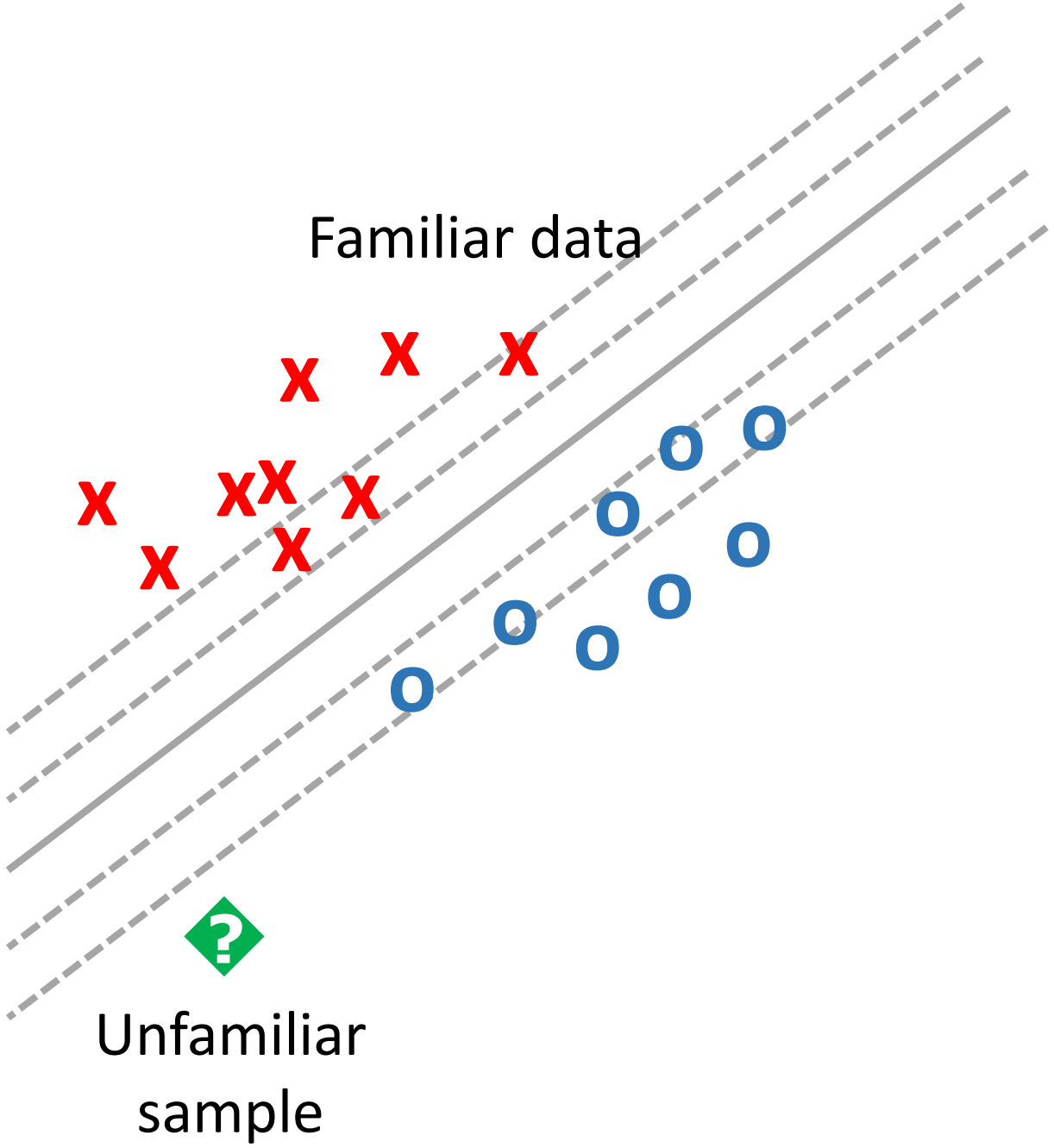
[1] Lakshminarayanan, Balaji, et al. "Simple and scalable predictive uncertainty estimation using deep ensembles." *NIPS* 2017.

Investigated methods

- Ensembles
- Calibration
  - Temperature scaling [2]

$$\mathbf{p}(x) = \mathrm{softmax}(\mathbf{f}(x)/T)$$

  - Use a higher temperature in evaluation
  - Calibrate the temperature in a validation set



[2] Guo, Chuan, et al. "On Calibration of Modern Neural Networks."

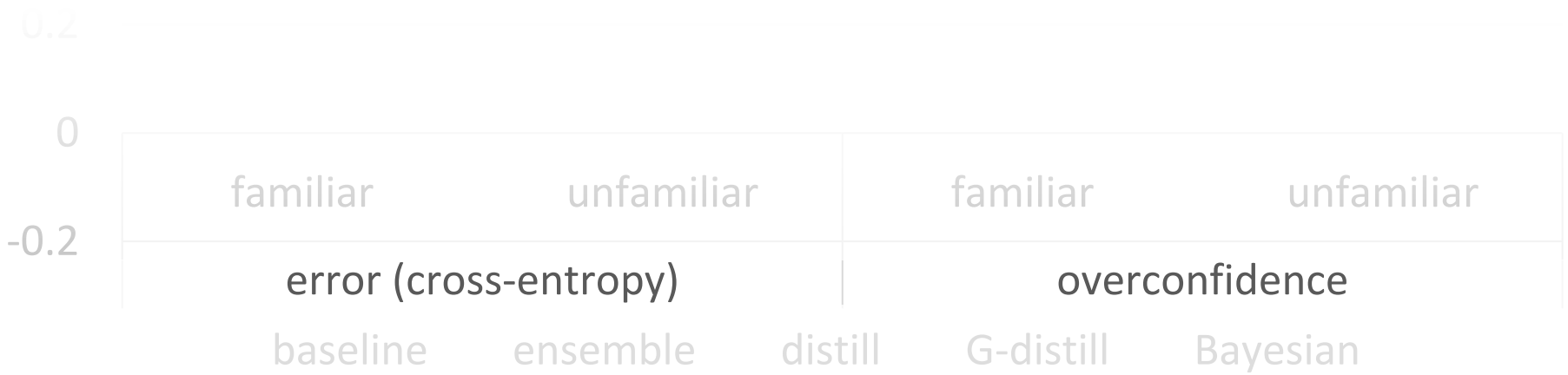
Investigated methods

- Ensembles
- Calibration
- Distillation
- Novelty detection
- Approximate Bayesian methods

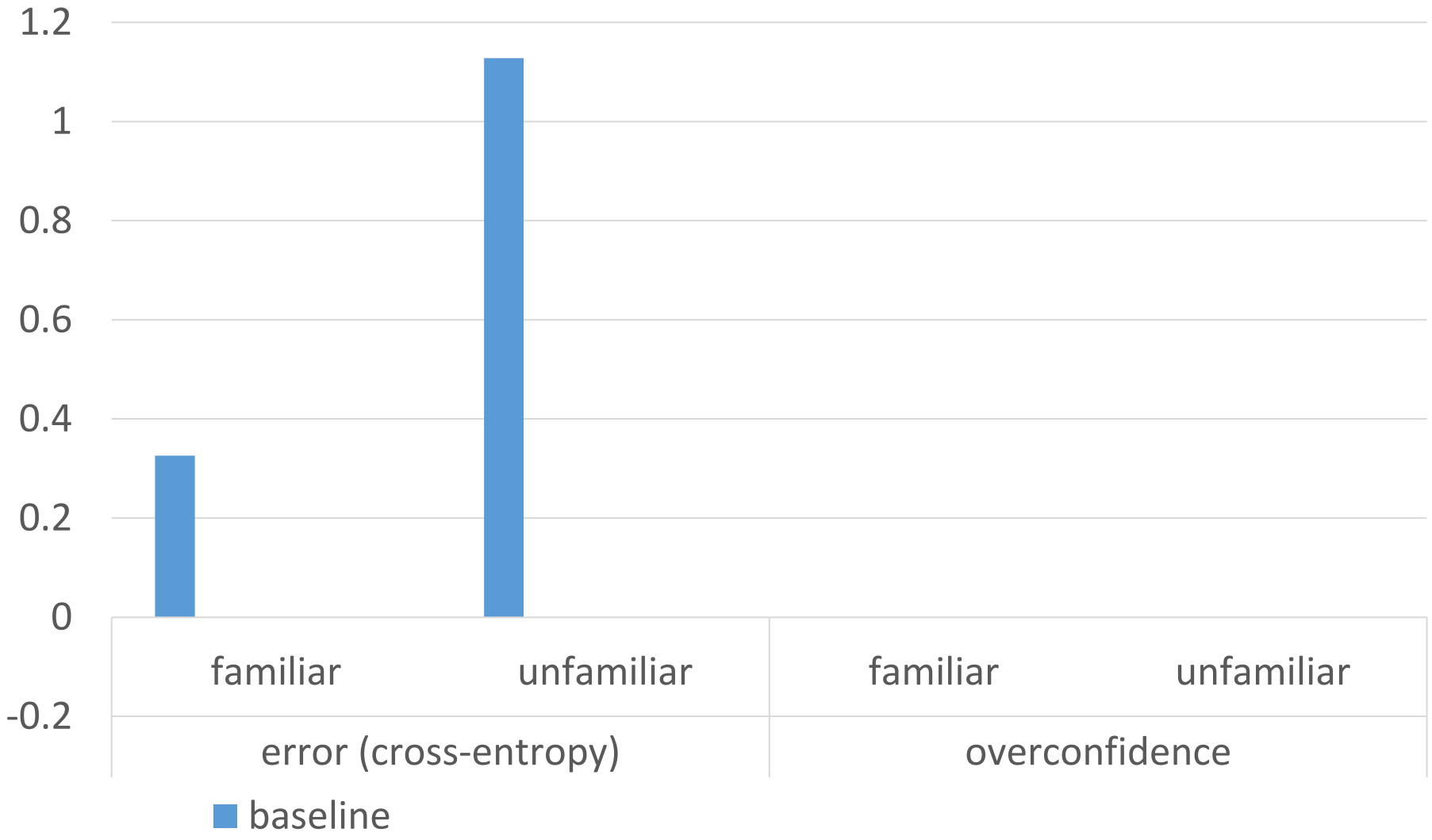
Criteria

- Error
  - Cross-entropy (NLL)

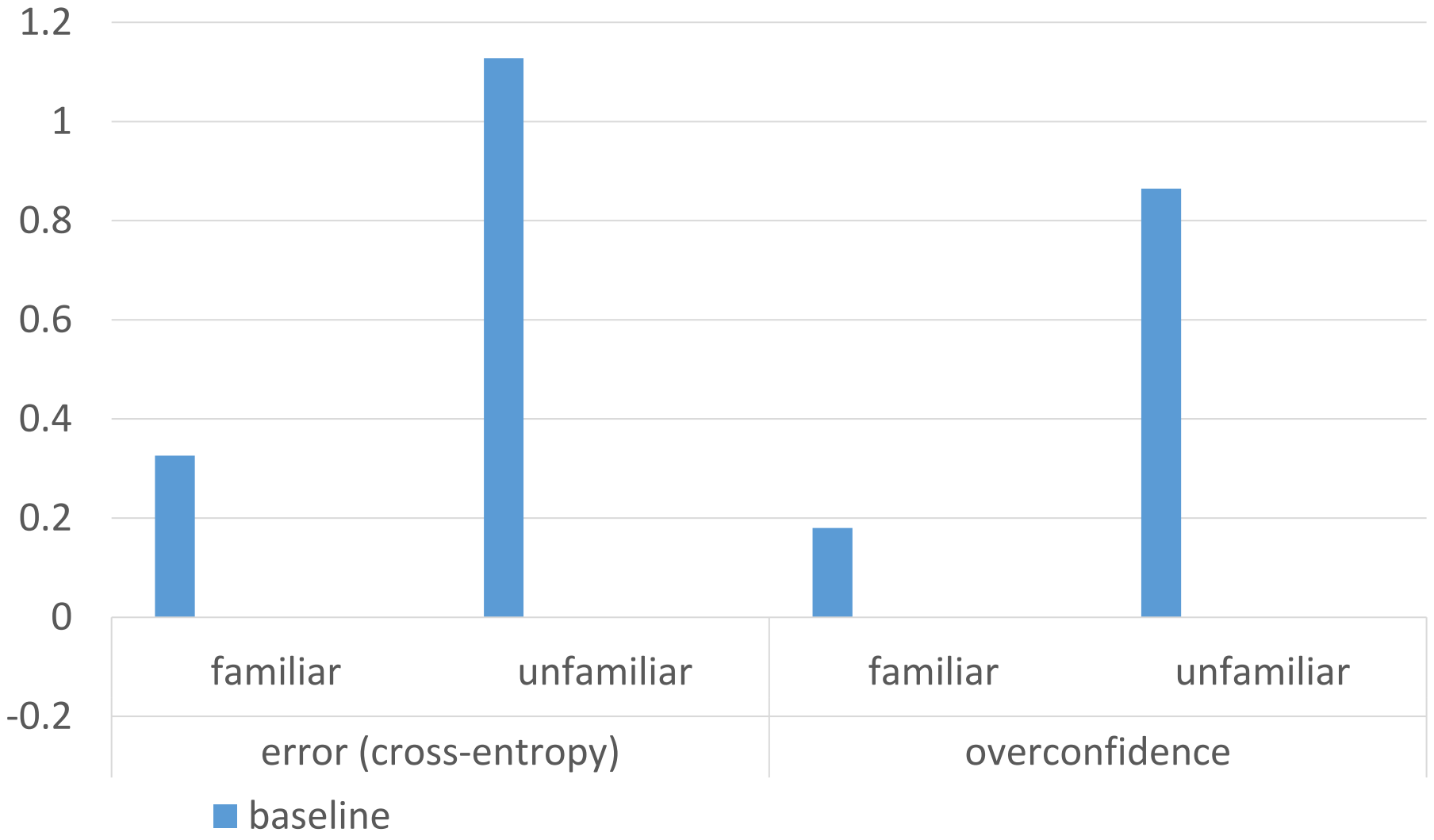
- Overconfidence
  - Cross-entropy minus entropy
    - Positive means entropy too small



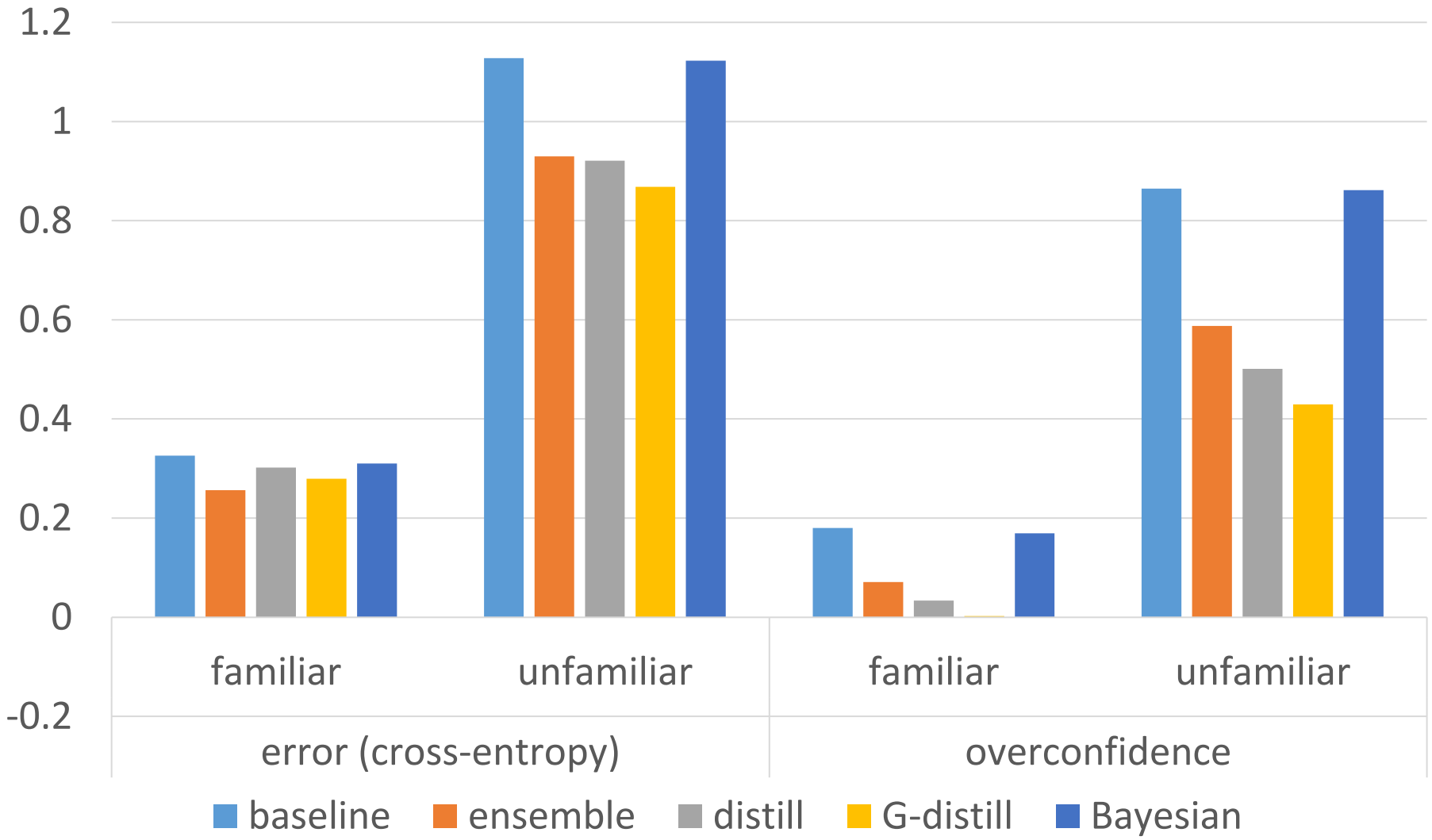
Results (animal classification dataset)



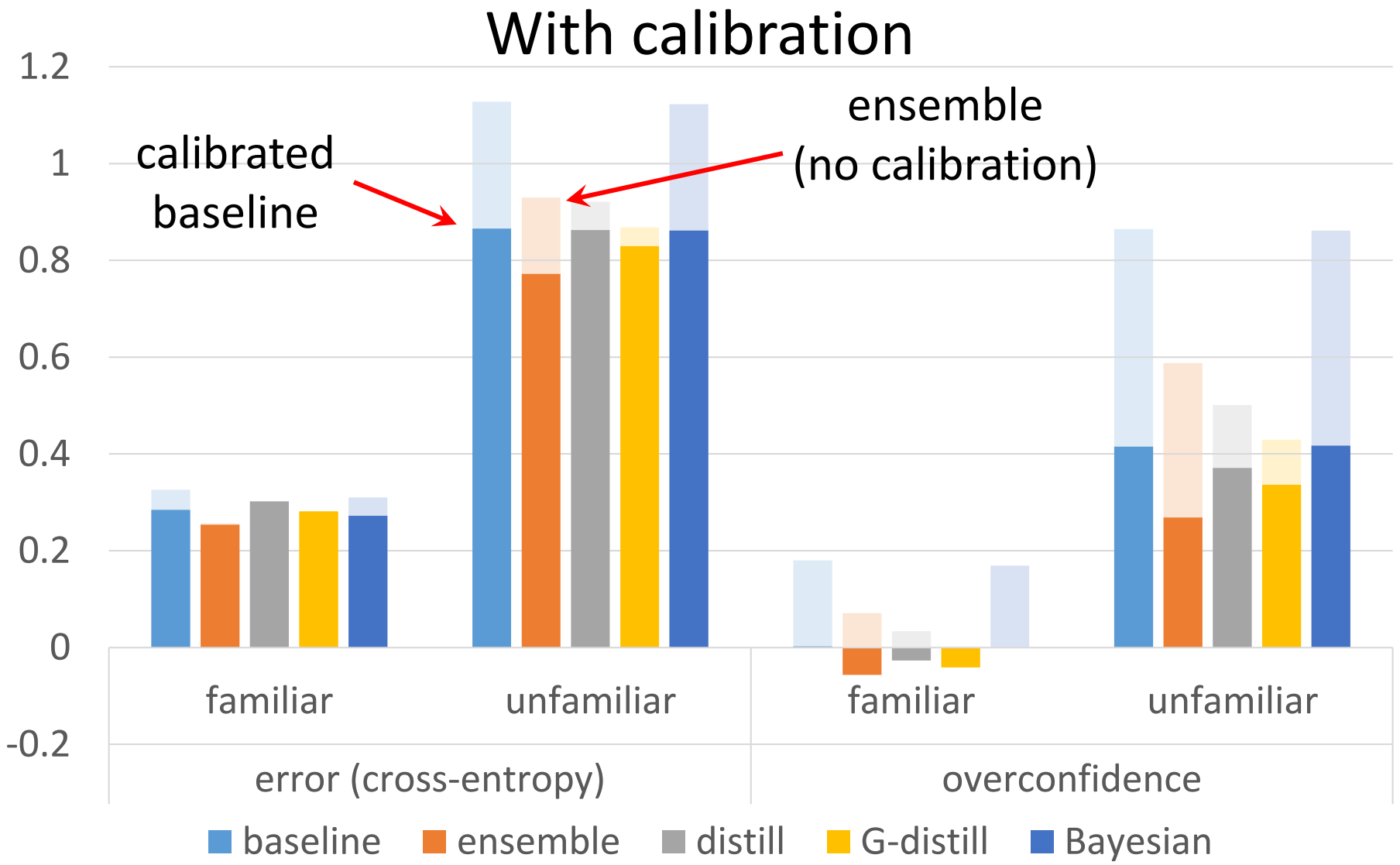
Results (animal classification dataset)



Results (animal classification dataset)

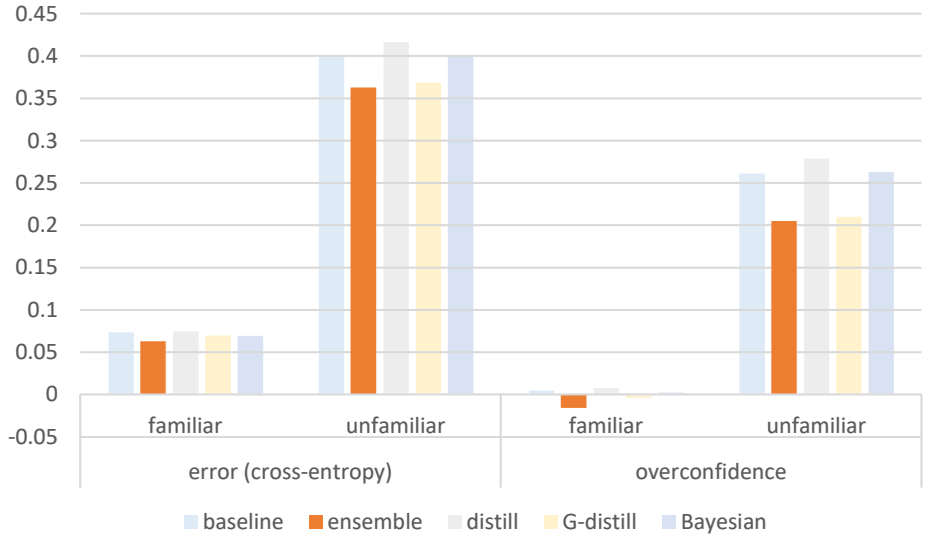


Results (animal classification dataset)

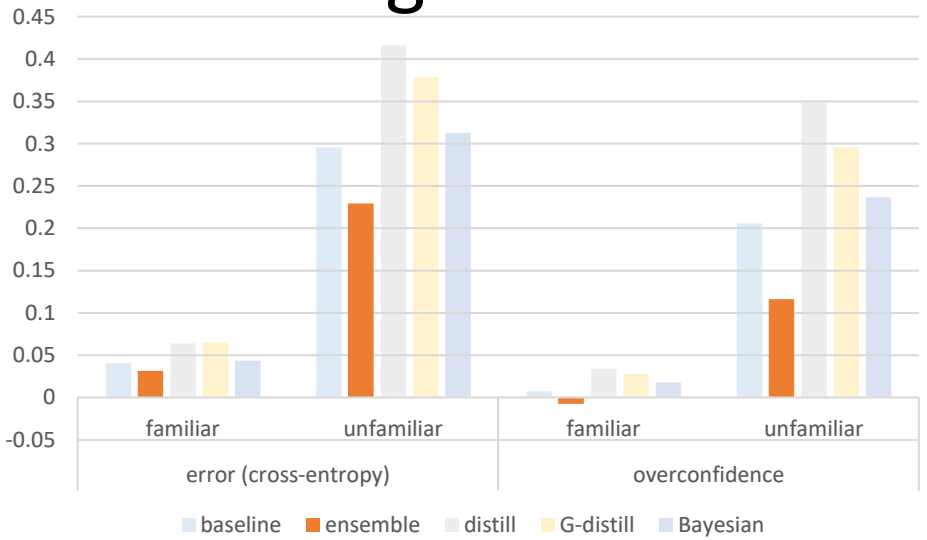


Results (all datasets)

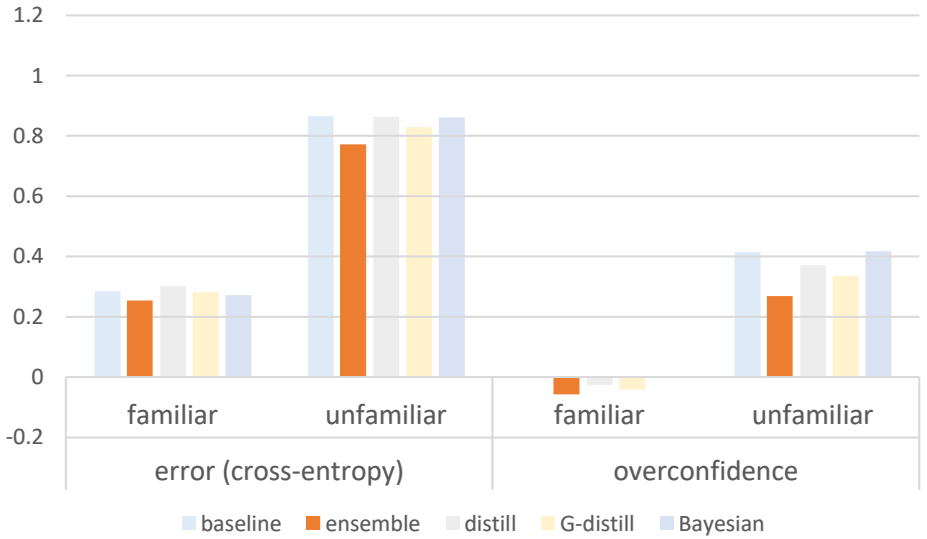
Gender Classification



Cat vs. Dog Classification



Animal Classification



Take-away

We highlight the issue:

- Data underrepresented in training can get confidently misclassified

We propose an experimental methodology to study the issue:

- Familiar / unfamiliar data splits
- Several useful metrics

We find best-performing methods:

- Calibration (T-scaling)
- Calibrated ensembles

~~= 99.9% female~~

= 84.5% female

~~= 99.3% male~~

= 53.7% male