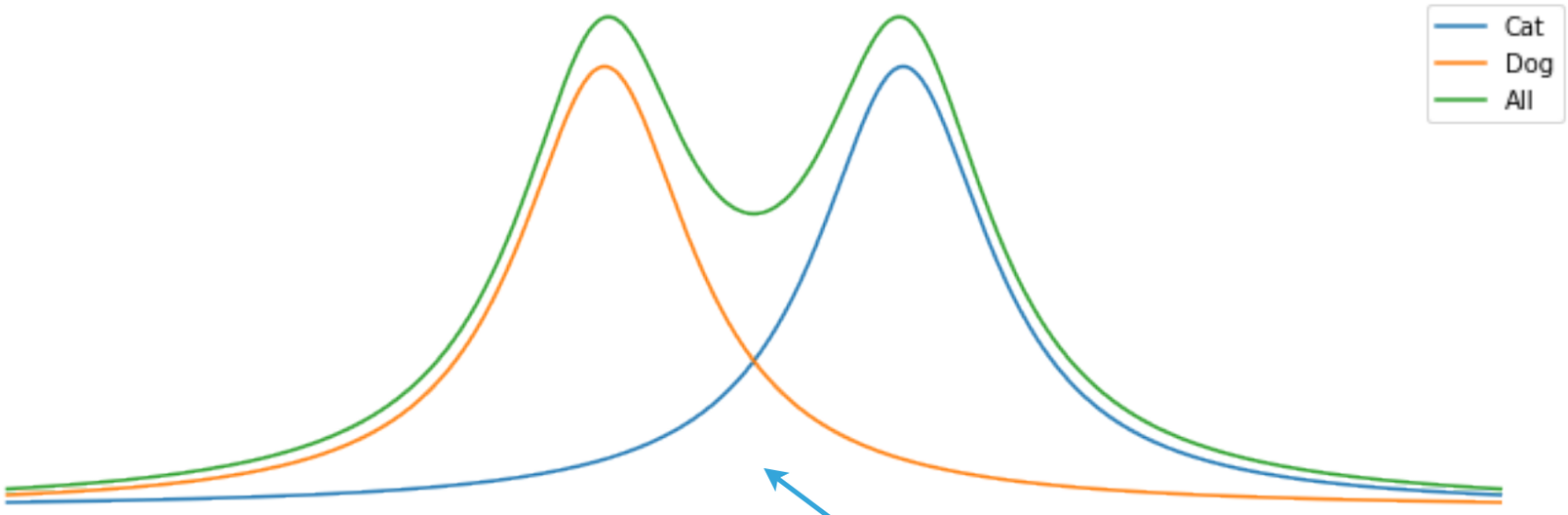
DIFFERENT SHADES OF UNCERTAINTY IN NEURAL NETWORKS

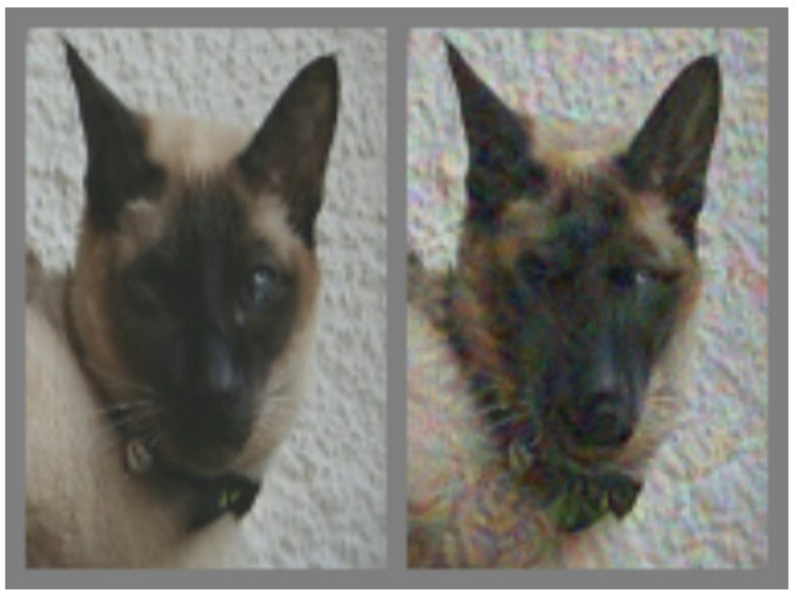
Marcin Możejko

OUTLINE OF THE PRESENTATION

- ***aleatoric error (risk)*** *- what we will probably never know*
- ***epistemic error (uncertainty)*** *- what we do not know but we may know*
- ***out-of-domain error*** *- what we like thinking that we know but we do not know at all*
- ***Bayesian Neural Networks*** *- a quick tour through the Bayesian approach*

ALEATORIC ERROR – WHAT WE WOULD PROBABLY NEVER KNOW





THIS REGION IS RISKY AND
ALWAYS WILL BE (ALEATORIC ERROR)

ALEATORIC ERROR – MAIN ISSUES

- Inaccurate measurement,
- Natural complexity of data,
- Insufficient discriminatory properties of data

You can't distinguish between examples represented by the same data points.

Pierre Simon de Laplace (or any reasonable person)

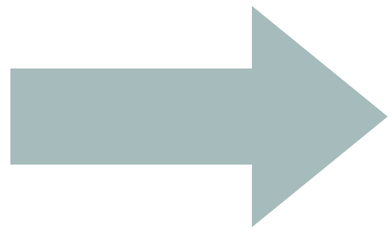
ALEATORIC ERROR – A SIMPLE USEFUL TRICK

Softmax output

0.5 *0.2* *0.3*

trained with categorical cross entropy

it prefers results with low entropy for simple examples

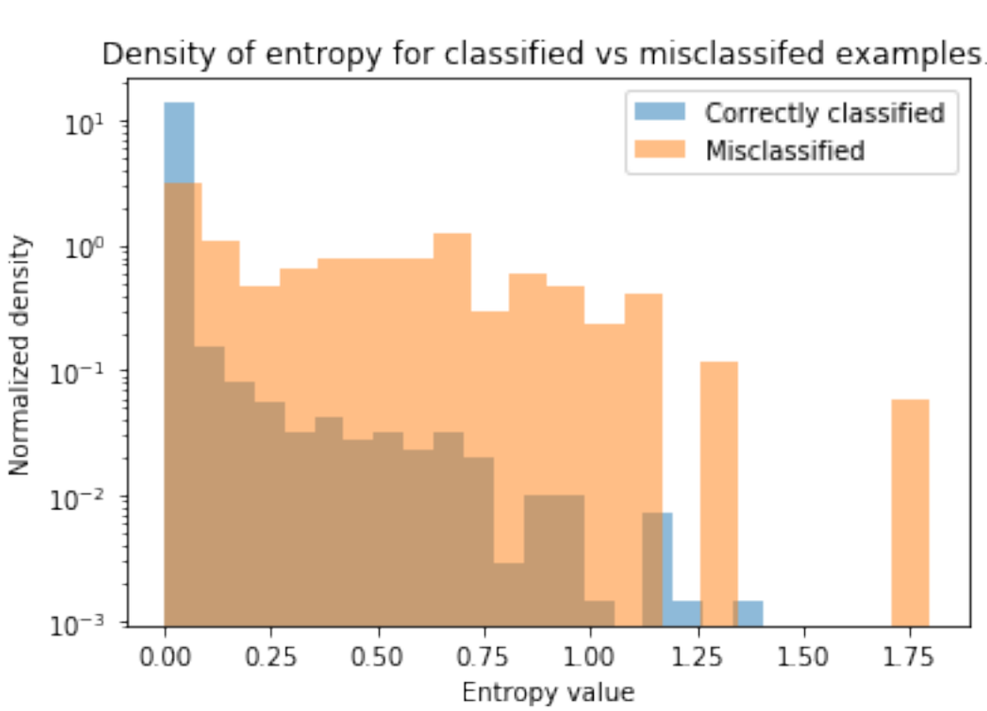


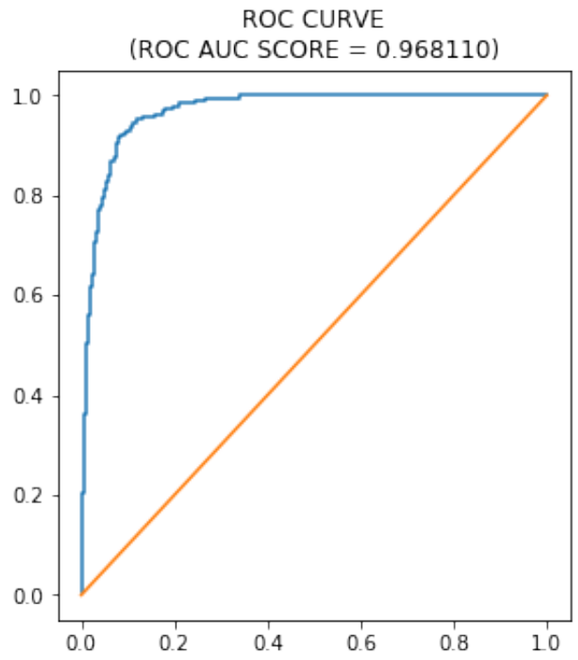
Easy examples

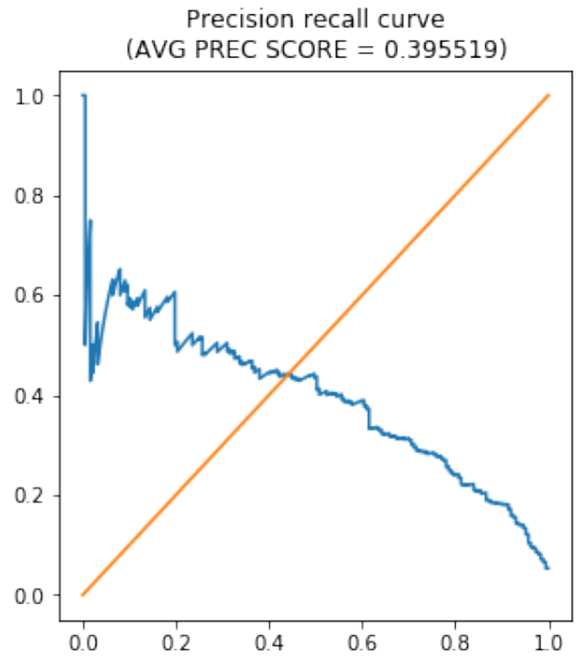
have lower entropies

ALEATORIC ERROR – A SIMPLE USEFUL TRICK

Experiment: difference between entropy distributions of T and F for a simple model trained on MNIST







ALEATORIC ERROR – WHAT IF WE CAN SEND THE HARDEST CASES TO MANUAL ANNOTATION

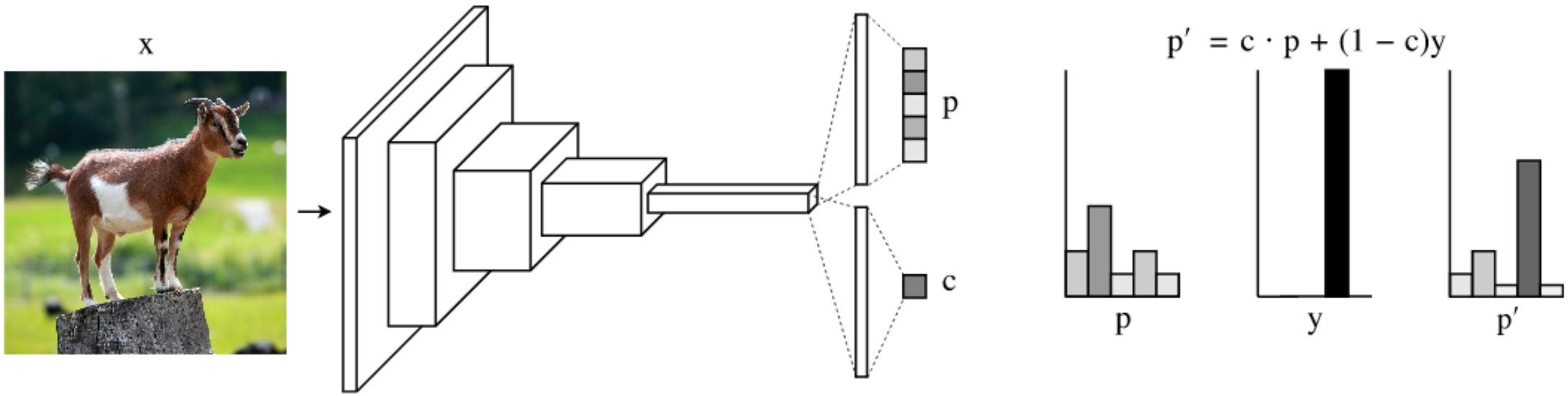


Figure 2. Neural network that has been augmented with a confidence estimation branch. The network receives input x and produces softmax prediction probabilities p and a confidence estimate c. During training, the predictions are modified according to the confidence of the network such that they are closer to the target probability distribution y.

Learning Confidence for Out-of-Distribution Detection in Neural Networks

T. DeVries et al.

ALEATORIC ERROR – HOW TO REDUCE IMPACT OF IRREDUCIBLE ERROR?

Mr Wigner and Mr Heisenberg: who is who?

> You flipped a coin 3 times and you got 3 tails. What is the maximum likelihood estimation of probability of getting a tail?

Frequentists Nightmare

EPISTEMIC ERROR – BOUNDARY EXAMPLES

- **Epistemic error** - the error which decreases when you gather more data points.

***Example**: different word meanings in sentiment analysis*

***Phase 1**: Small amount of data:*

***Phase 2**: New datapoints gathered:*

***Phase 3**: Ideal dataset:*

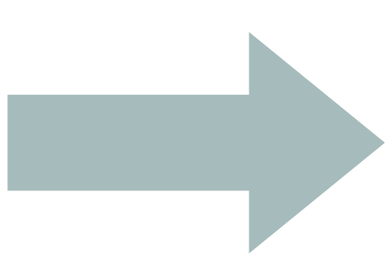
ONE TRICK TO RULE THEM ALL

Softmax output

0.5 0.2 0.3

trained with categorical cross entropy

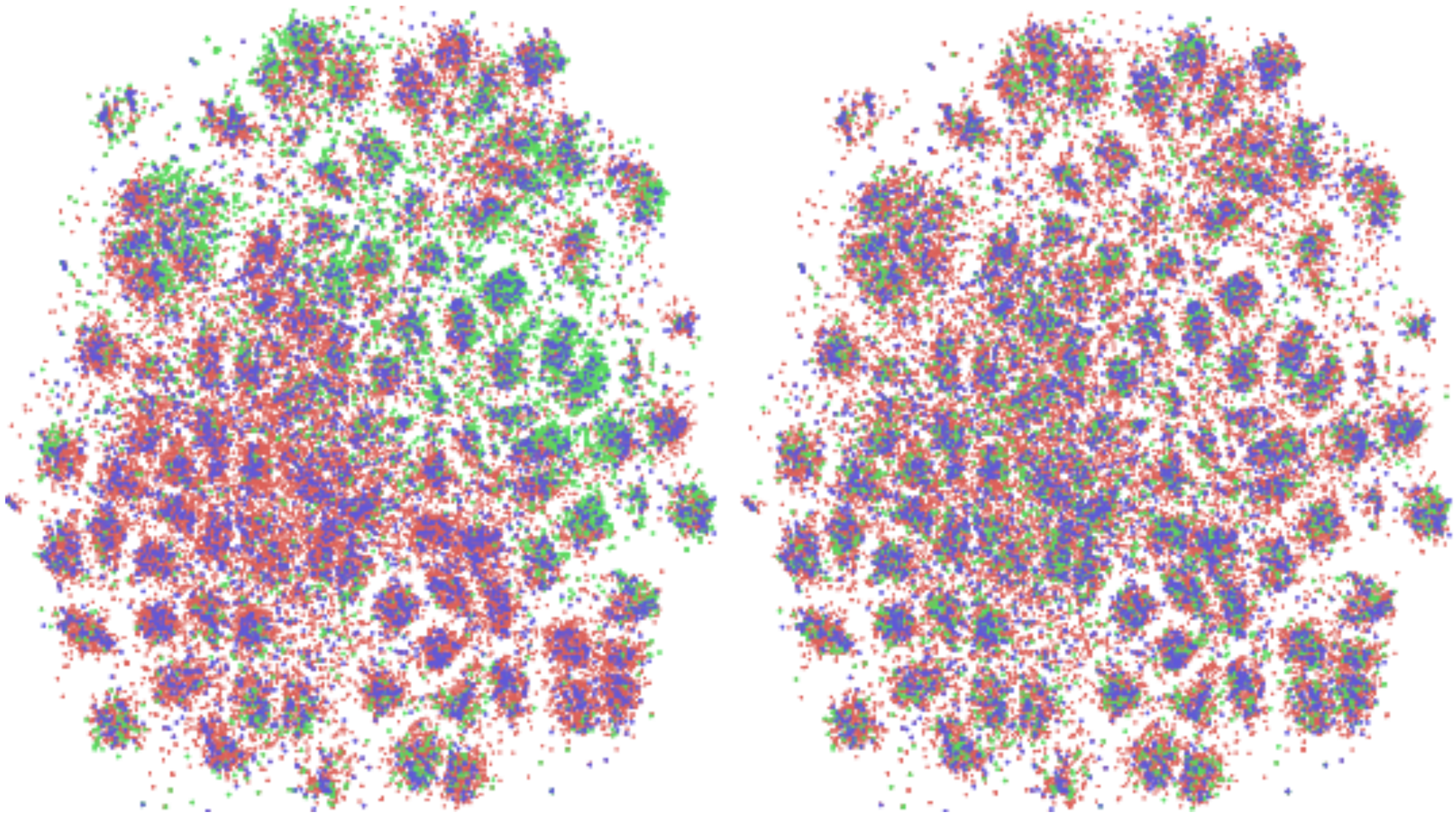
it prefers results with low entropy for simple examples



Cases with lower entropy are the known ones

BUT IT IS NOT AS EASY AS ONE MAY THINK!

LET US SPEED UP THE PROCESS – ACTIVE LEARNING



Active Learning for Convolutional Neural Networks: A Core-Set Approach

O. Sener and S. Savarese

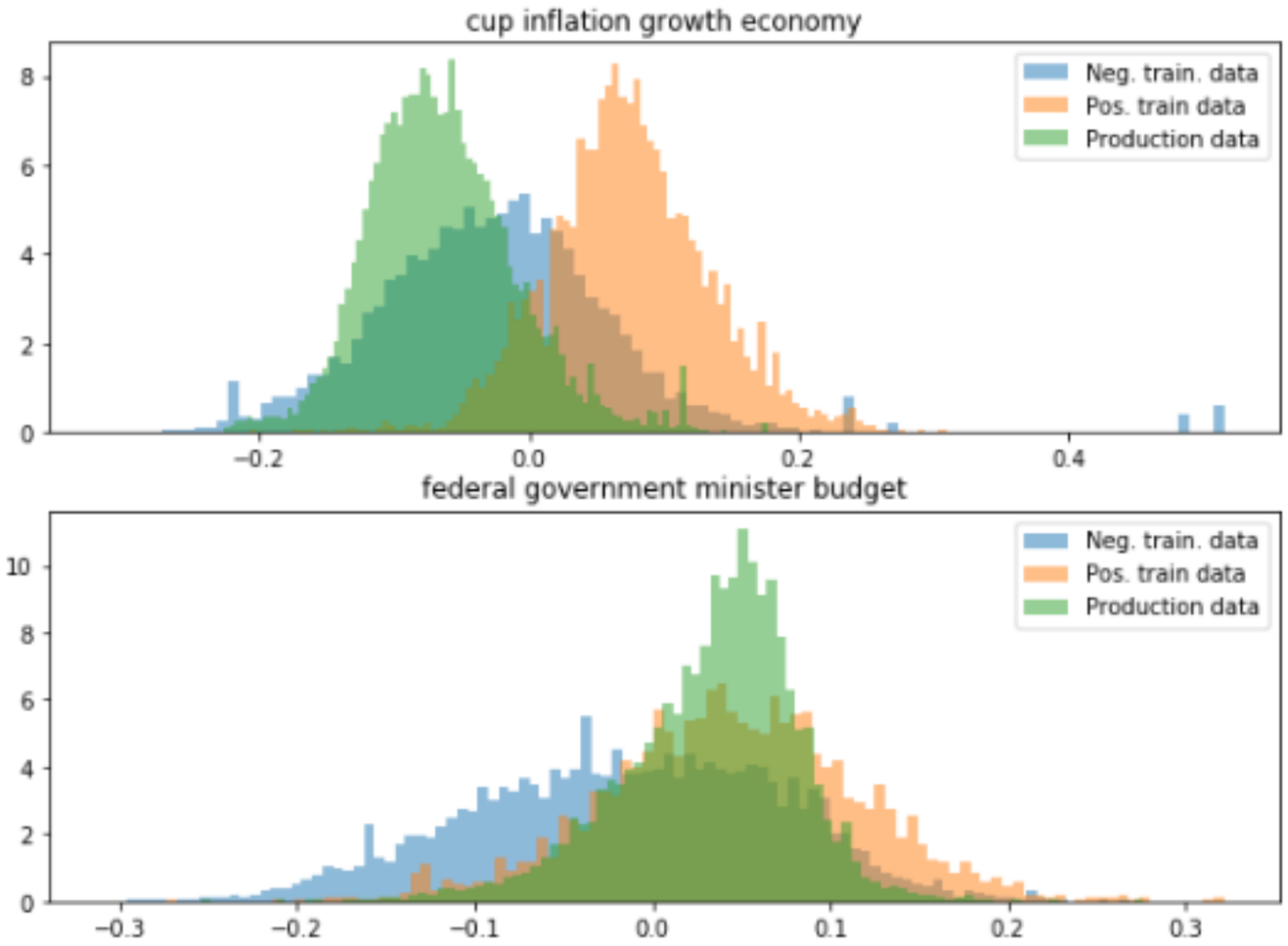
EPISTEMIC ERROR – WAITING FOR THE PERFECT DATASET

OUT-OF-DISTRIBUTION ERROR – WHAT COMES ONLY IN PRODUCTION PHASE

- This error occurs in case when **training distribution** differs from **production distribution**

***Example**: text classification domain mismatch*

***Explanation:** hidden selection of articles on the client side*

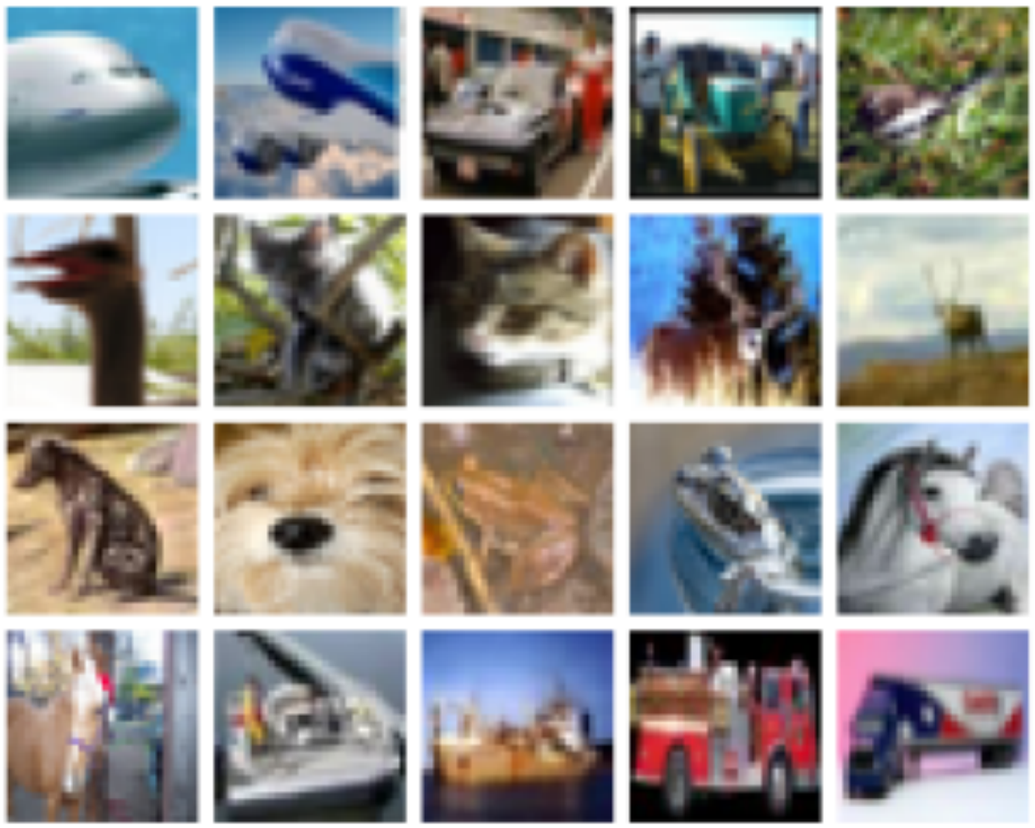


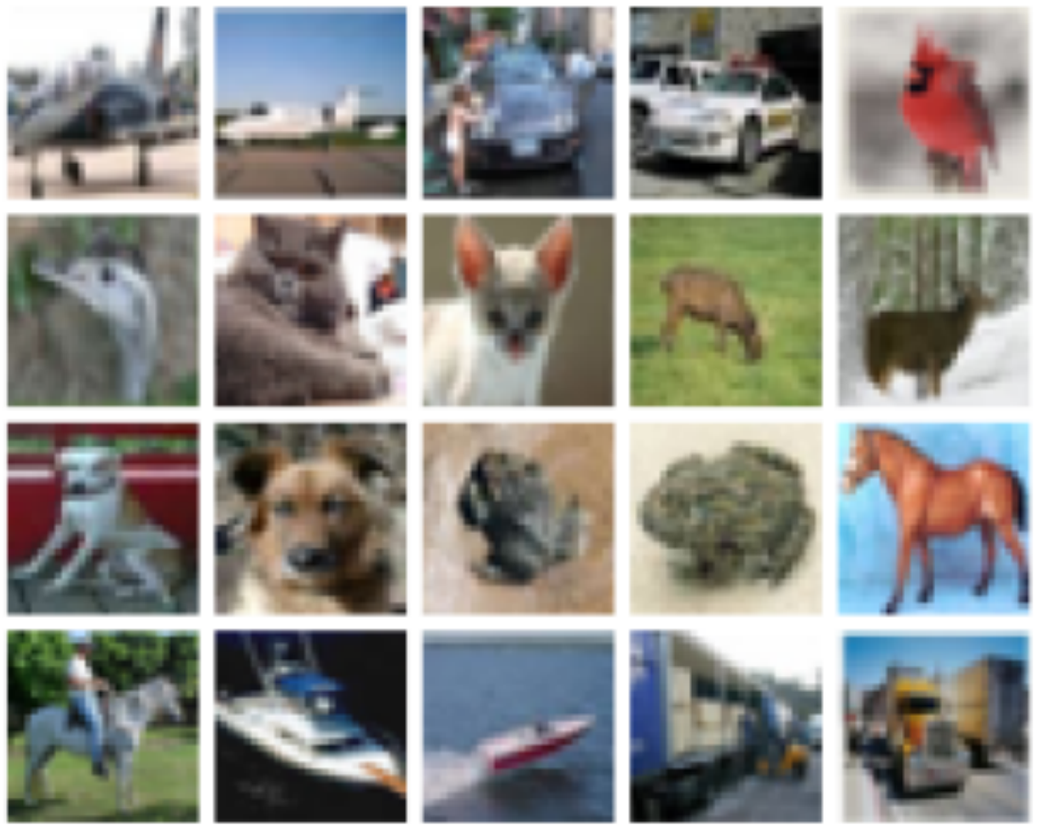
PCA transformation of mean embeddings of both training and production data

Do CIFAR-10 Classifiers Generalize to CIFAR-10?

Benjamin Recht et al.

https://arxiv.org/pdf/1806.00451.pdf





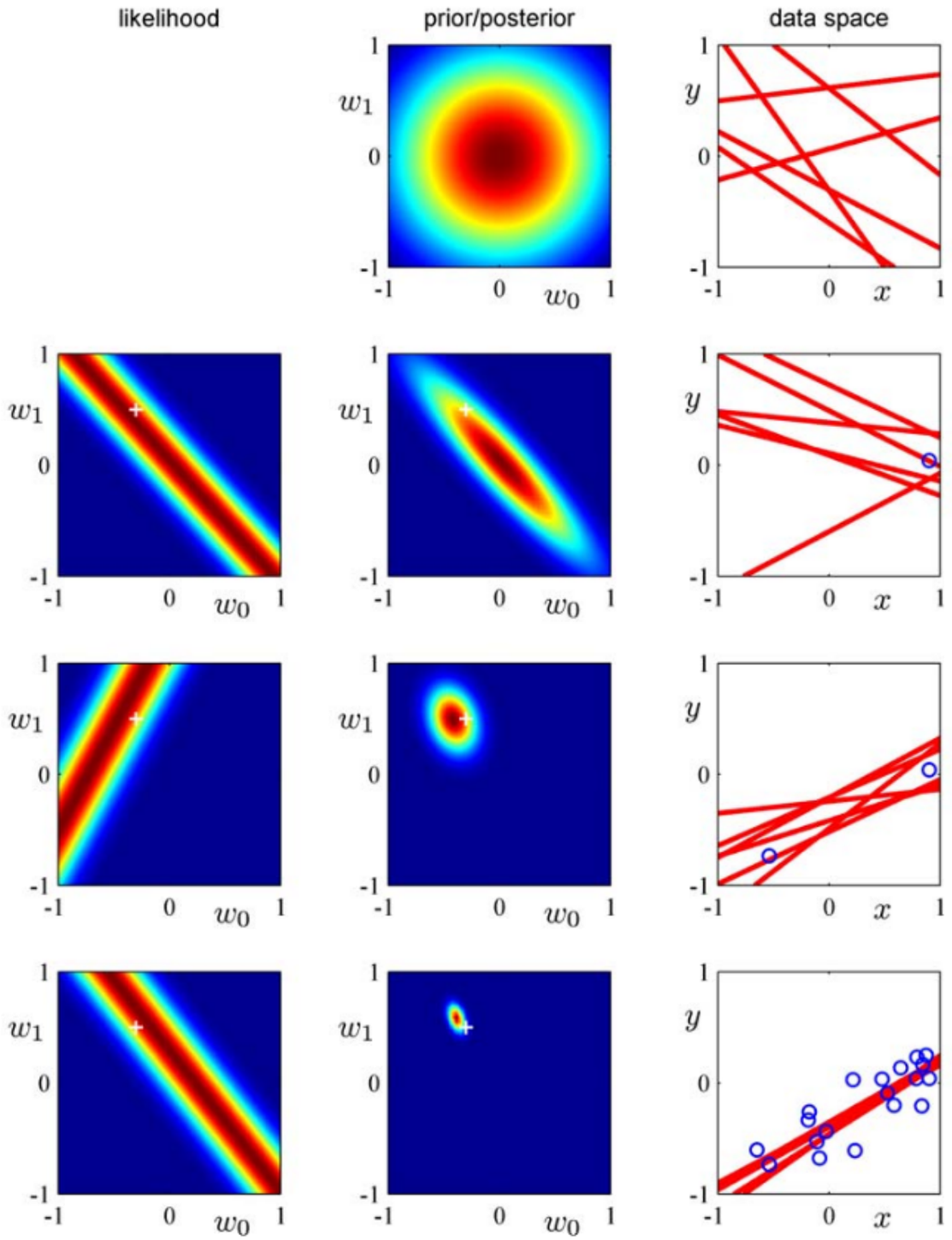
Benjamin Recht et al.

https://arxiv.org/pdf/1806.00451.pdf

	Original Accuracy	New Accuracy	Gap	Δ Rank
shake_shake_64d_cutout [3, 4]	97.1 [96.8, 97.4]	93.0 [91.8, 94.0]	4.1	0
shake_shake_96d [4]	97.1 [96.7, 97.4]	91.9 [90.7, 93.1]	5.1	-2
shake_shake_64d [4]	97.0 [96.6, 97.3]	91.4 [90.1, 92.6]	5.6	-2
wide_resnet_28_10_cutout [3, 22]	97.0 [96.6, 97.3]	92.0 [90.7, 93.1]	5	+1
shake_drop [21]	96.9 [96.5, 97.2]	92.3 [91.0, 93.4]	4.6	+3
shake_shake_32d [4]	96.6 [96.2, 96.9]	89.8 [88.4, 91.1]	6.8	-2
darc [11]	96.6 [96.2, 96.9]	89.5 [88.1, 90.8]	7.1	-4
resnext_29_4x64d [20]	96.4 [96.0, 96.7]	89.6 [88.2, 90.9]	6.8	-2
pyramidnet_basic_110_270 [6]	96.3 [96.0, 96.7]	90.5 [89.1, 91.7]	5.9	+3
resnext_29_8x64d [20]	96.2 [95.8, 96.6]	90.0 [88.6, 91.2]	6.3	+3
wide_resnet_28_10 [22]	95.9 [95.5, 96.3]	89.7 [88.3, 91.0]	6.2	+2
pyramidnet_basic_110_84 [6]	95.7 [95.3, 96.1]	89.3 [87.8, 90.6]	6.5	0
densenet_BC_100_12 [10]	95.5 [95.1, 95.9]	87.6 [86.1, 89.0]	8	-2
neural_architecture_search [23]	95.4 [95.0, 95.8]	88.8 [87.4, 90.2]	6.6	+1
wide_resnet_tf [22]	95.0 [94.6, 95.4]	88.5 [87.0, 89.9]	6.5	+1

Benjamin Recht et al.

https://arxiv.org/pdf/1806.00451.pdf



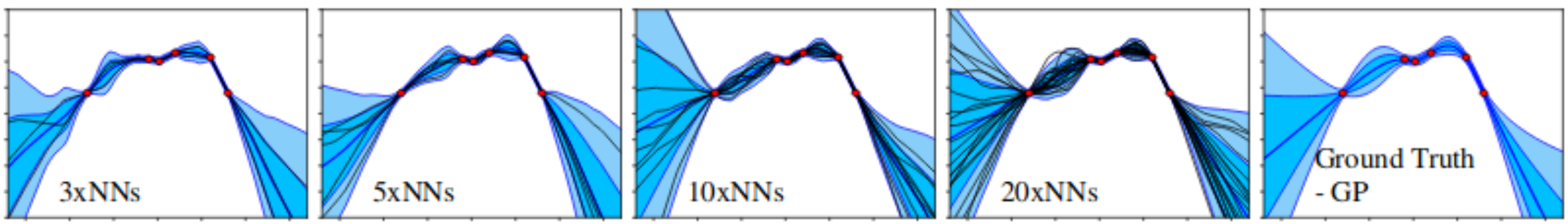
QUICK REMINDER OF BAYESIAN LEARNING

- 1. Prior distribution
- 2, 3, 4 Iterative posterior computation
- **+ sign -** true original distribution
- **Variational learning** - try to find a simpler distribution which approximates the final posterior

"Pattern recognition and machine learning" - Christopher Bishop

BAYESIAN ENSEMBLE SAMPLING

- Although **L2** regularization might be interpreted as maximization of posterior distribution multiple networks trained with this regularization cannot be interpreted as a valid Bayesian sample
- Instead - first sample multiple models from a prior distribution - then penalize the square difference between trained model and original samples.



Uncertainty in Neural Networks: Bayesian Ensembling, T. Pearce et al.

Questions?