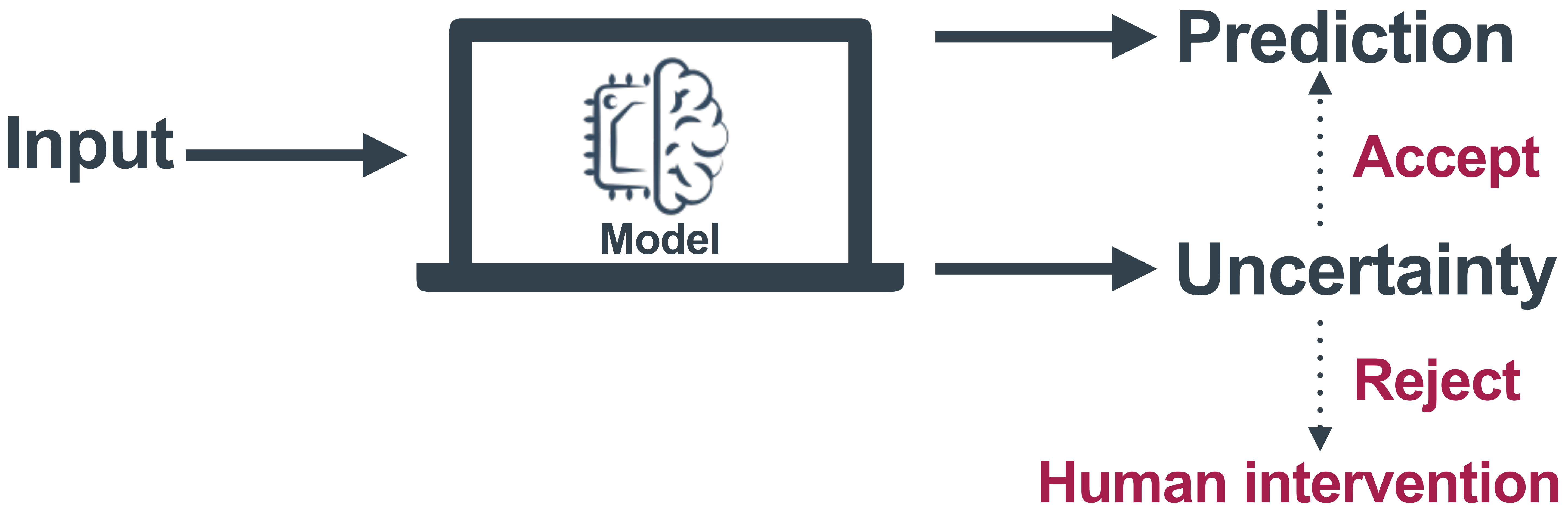
EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Modelling multiple sources of uncertainty

Scenario



AI icon by Icons8

Source of uncertainty

Data uncertainty

Dog

Mop

? ? ? ?

Training

Test

Distributional uncertainty

Training dataset

Test dataset

Figure from https://www.boredpanda.com/dog-food-comparison-bagel-muffin-lookalike-teenybiscuit-karen-zack/ And Photo by Alexander London on Unsplash

Problem:

How to distinguish distributional uncertainty from other uncertainties?

→ How could we distinguish between

- in-domain(distinct) data vs out-of-distribution data
- in-domain(data uncertainty) data vs out-of-distribution data

Previous work:

- Predictive posteriors of regression or classification DNNs
  [Gimpel 2016, Malinin2017, Lee 2018, Liang 2018]
- Bayesian approaches like Monte-Carlo Dropout
  [Gal 2016]
- Non-Bayesian Ensembles like Deep Ensembles
  [Lakshminarayanan 2017]

Common problem?

Solution:

Dirichlet Prior Network with KL- divergence (Malinin 2018)

Our contribution:

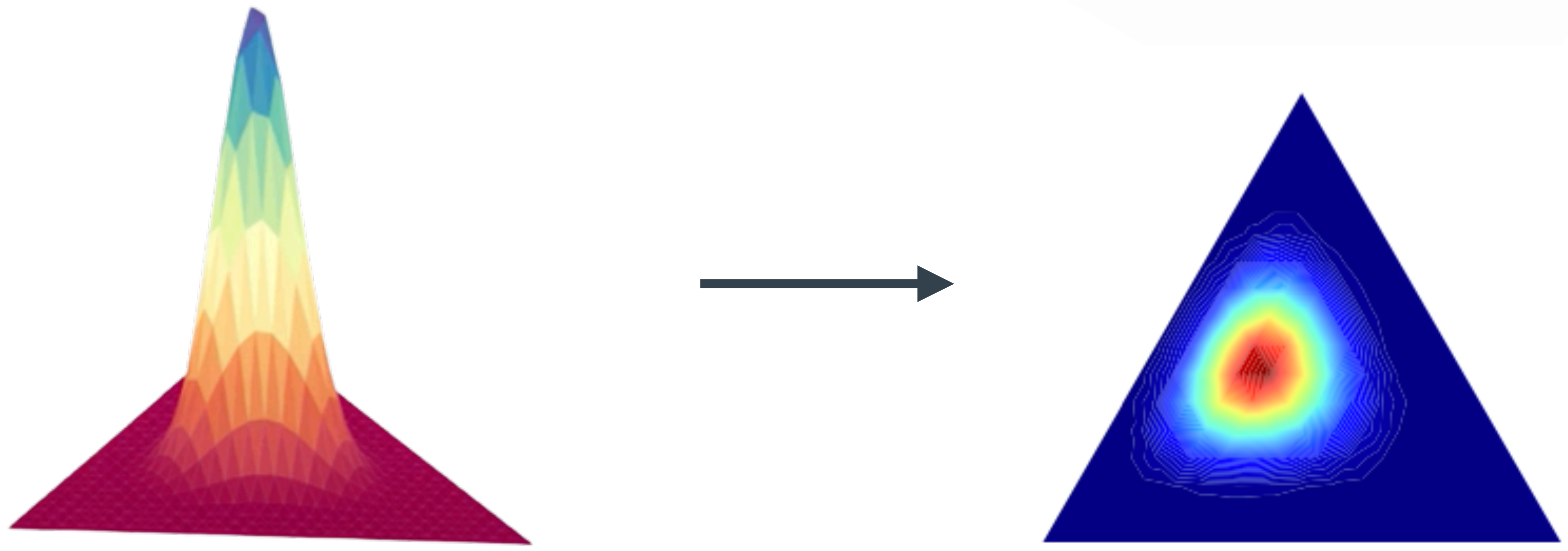
1. Reproduce the paper using DPN with reverse KL-divergence

2. Experiments on additional dataset that we created

Dirichlet distribution

$$\mathrm{Dir}(\mu|\alpha) = \frac{\Gamma(\sum_c \alpha_c)}{\prod_c \Gamma(\alpha_c)} \prod_c \mu_c^{\alpha_c - 1}$$

$\alpha_c < 1$

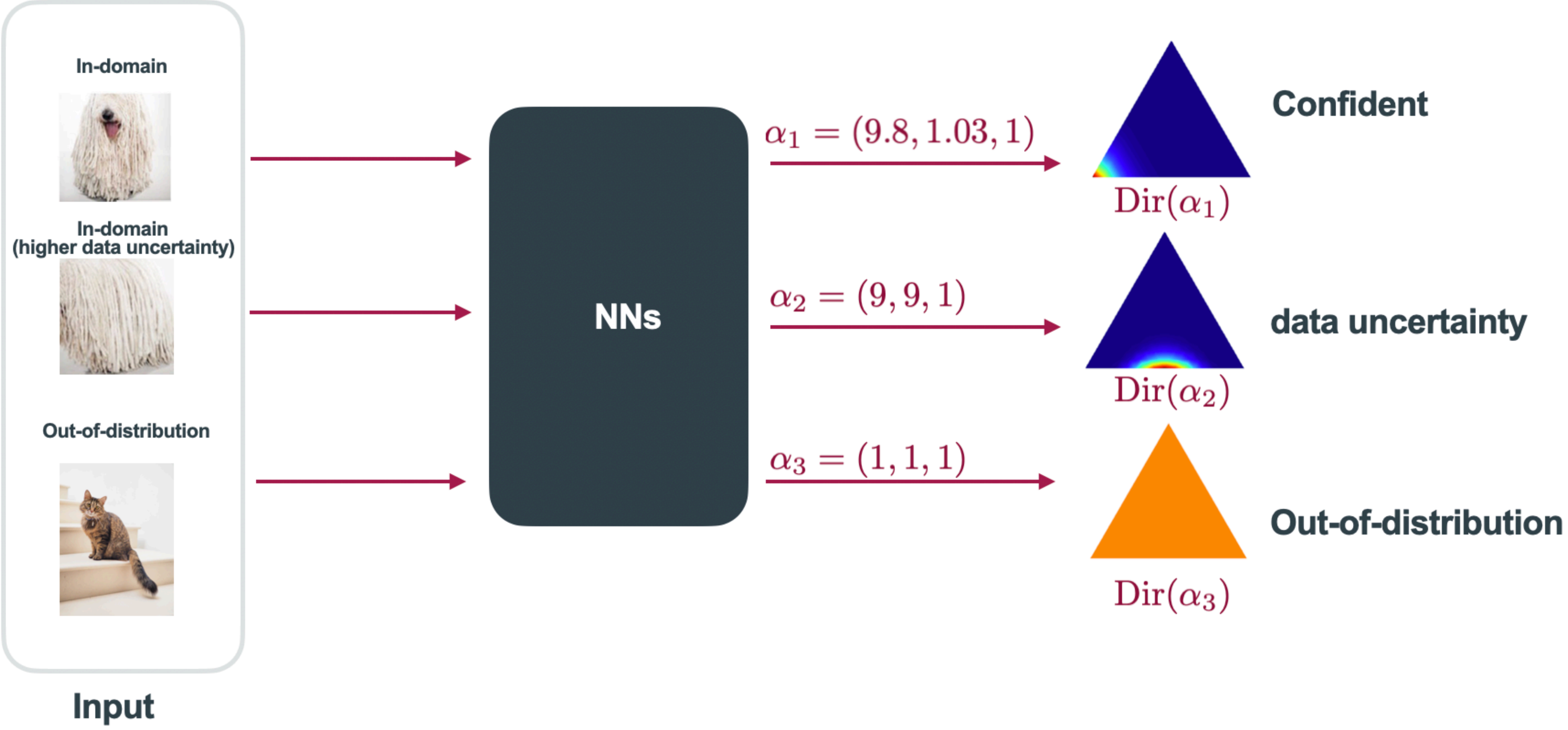


$\alpha_c > 1$

Red figures from https://www.youtube.com/watch?v=DpGkAKdLjdo

Dirichlet Prior Network

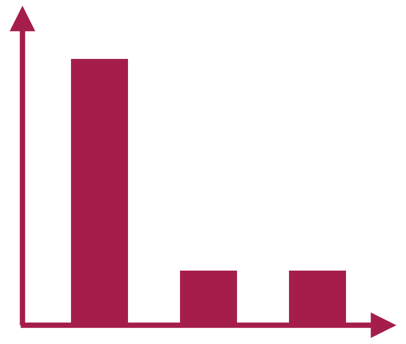
It parameterises Dirichlet distribution as prior to the predictive distribution over simplex



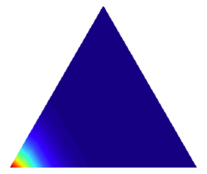
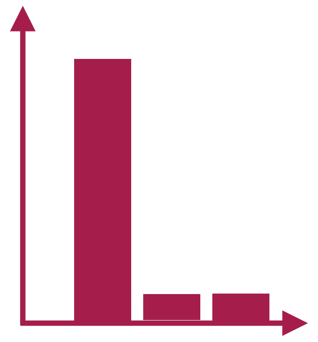
$$P(y = w_c | x^*, \hat{\theta}) = \hat{\mu_c} = \frac{\alpha_c}{\sum \alpha_c}$$

Uncertainty Measures

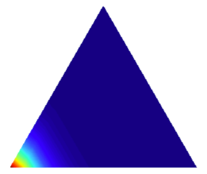
Entropy of y:

$\mathcal{H}[P(y|x^*;\mathcal{D})] \longrightarrow$ $\mathcal{H}[$  $]$

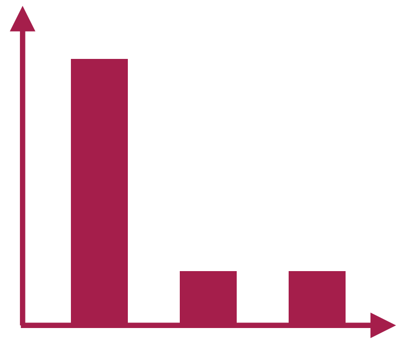
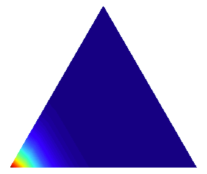
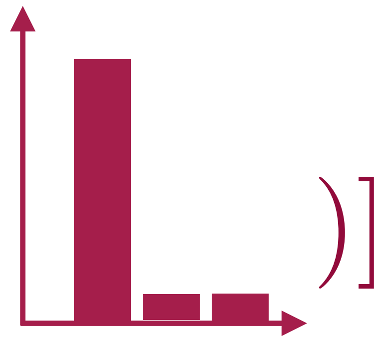
Expected data uncertainty(D.C.):

$\mathbb{E}_{P(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu))] \longrightarrow$ $\mathbb{E}_{P(\mu\sim}$  $_{)}[\mathcal{H}($  $)]$

Entropy of mu:

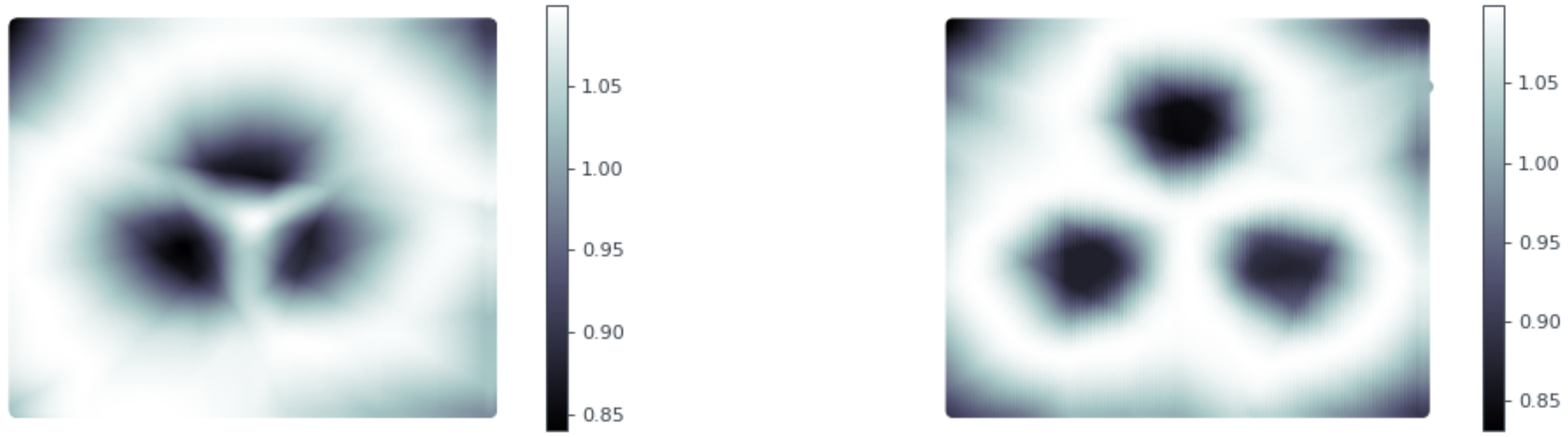
$\mathcal{H}[P(\mu|x^*;\mathcal{D})] \longrightarrow$ $\mathcal{H}[$  $]$

Mutual Information(M.I.):

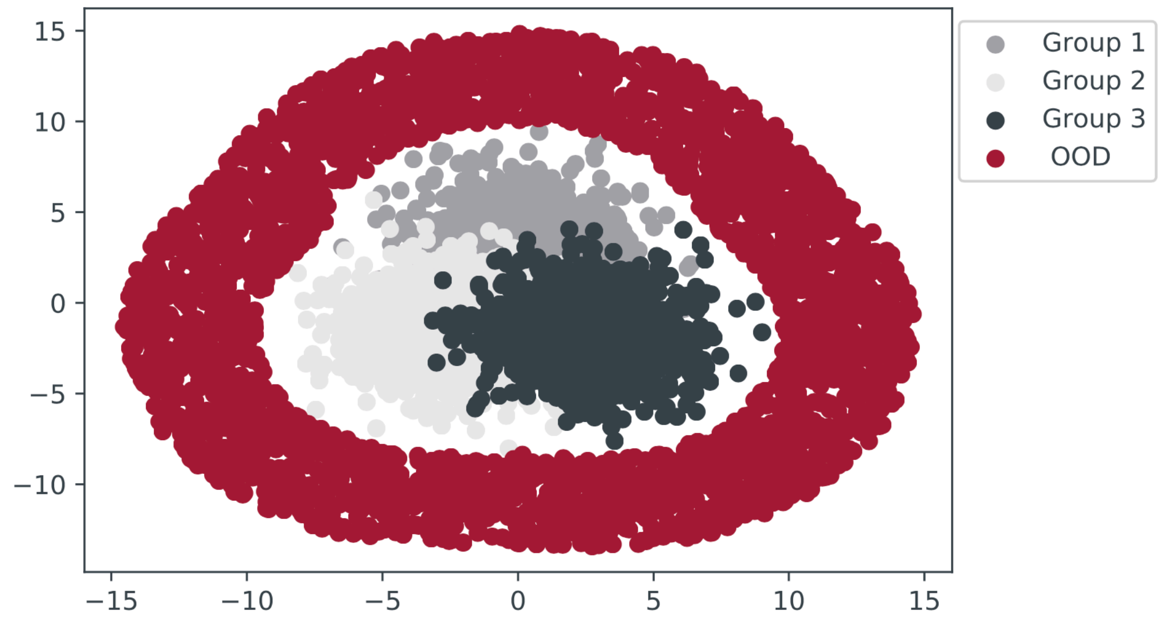
$\mathcal{H}[P(y|x^*;\mathcal{D})] - \mathbb{E}_{P(\mu|x^*;\mathcal{D})}[\mathcal{H}(P(y|\mu))] \longrightarrow$ $\mathcal{H}[$  $] - \mathbb{E}_{P(\mu\sim}$  $_{)}[\mathcal{H}($  $)]$

Synthetic Experiments

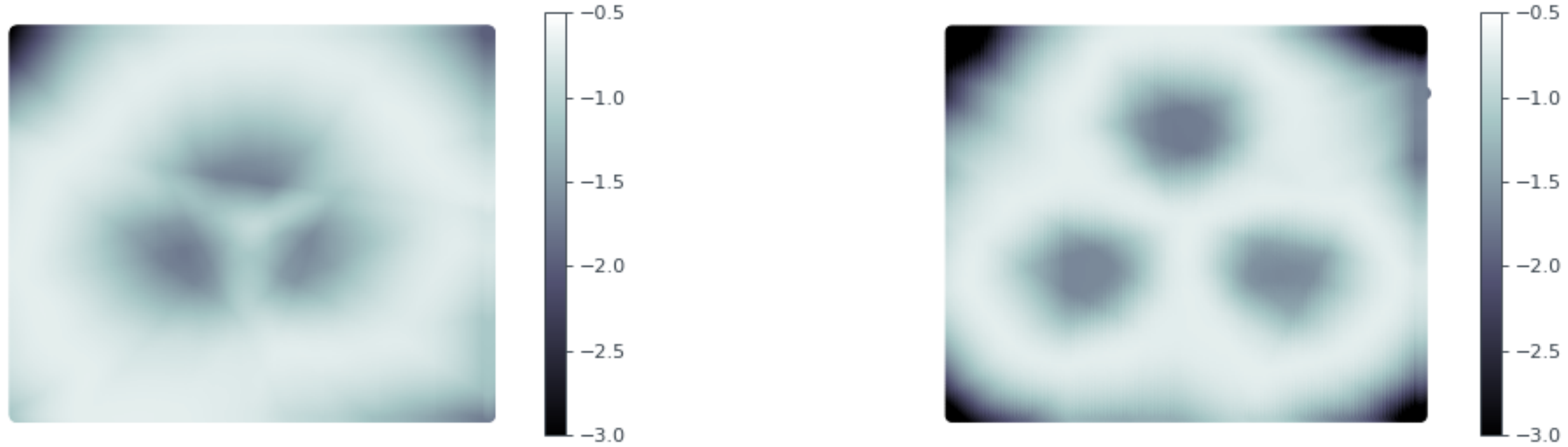
Entropy of y



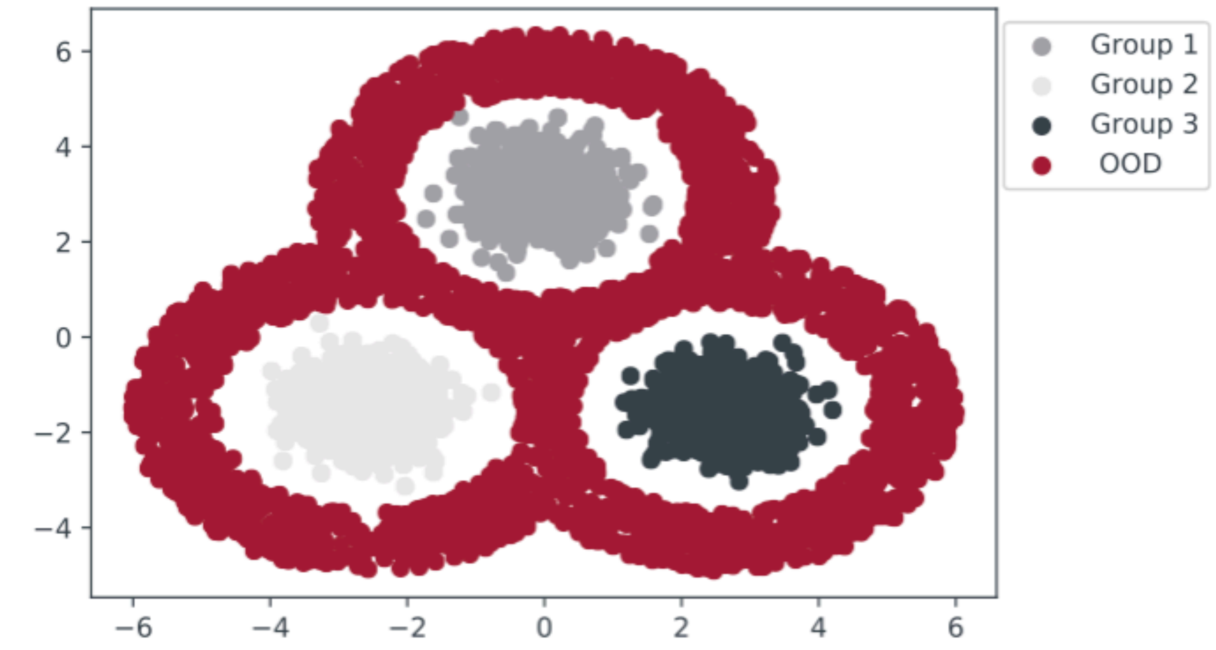
Training Dataset



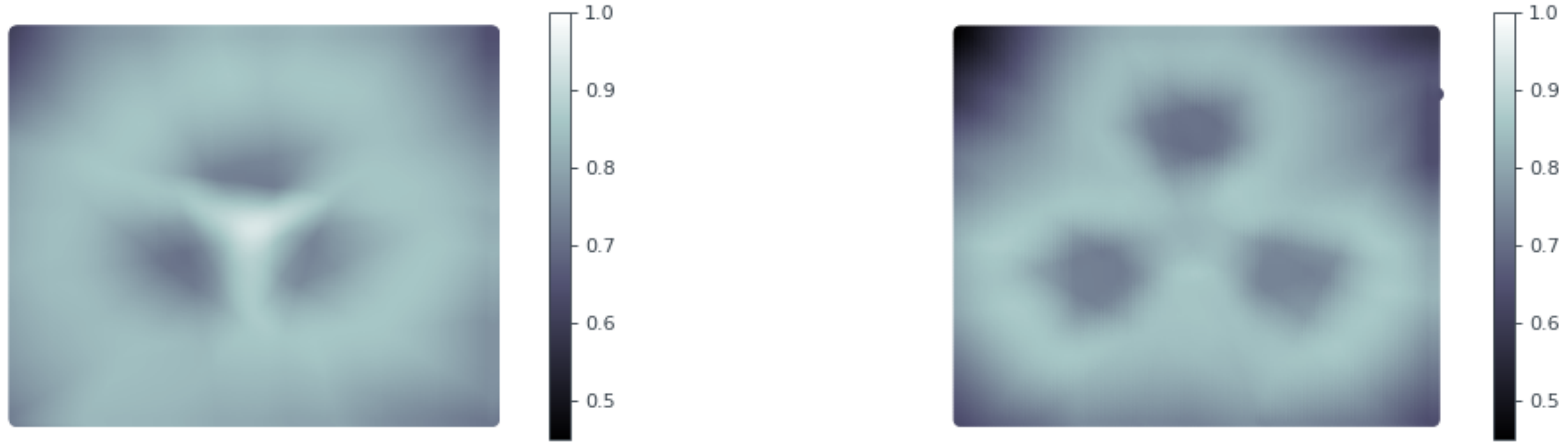
Entropy of mu



Training Dataset

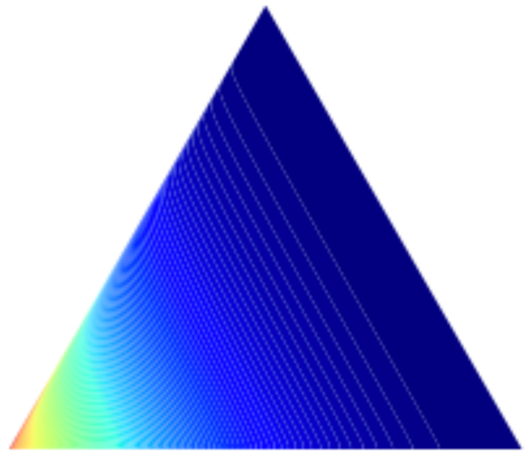


Expected Data Uncertainty

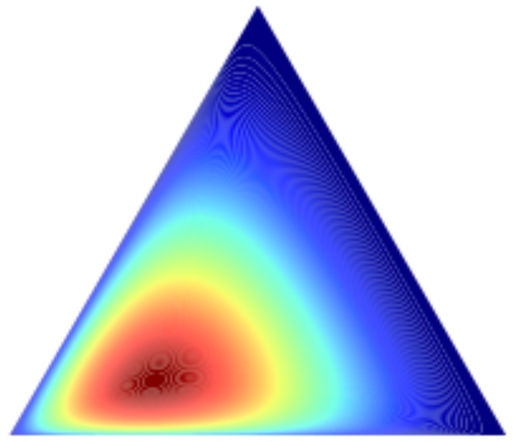


Synthetic Experiments

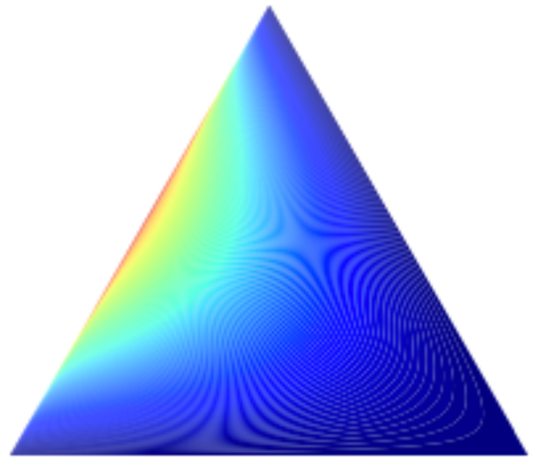
	Points	Entropy of y	Entropy of μ	Data Uncertainty
a	In domain	0.9071	-1.4869	0.7523
b	High class overlap	1.0439	-1.0974	0.8985
c	weak class overlap	1.0402	-1.0914	0.8739
d	Out-of-distribution	1.0951	-0.7129	0.8441



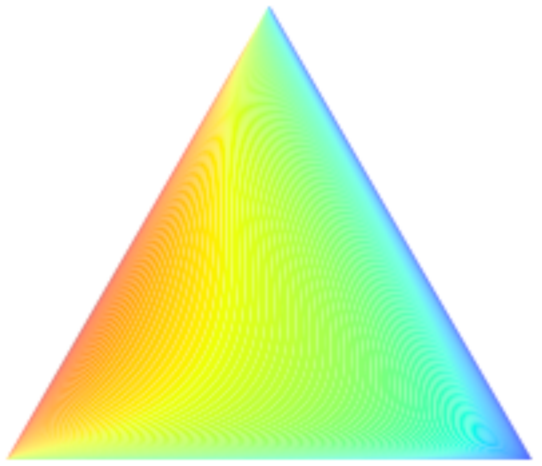
a



b



c

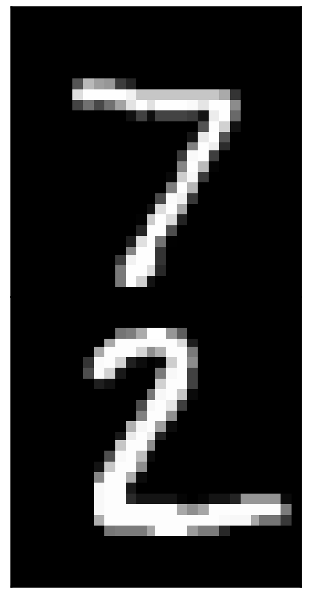


d

Reproduction (MNIST+FashionMNIST)

Setup

In-domain Dataset



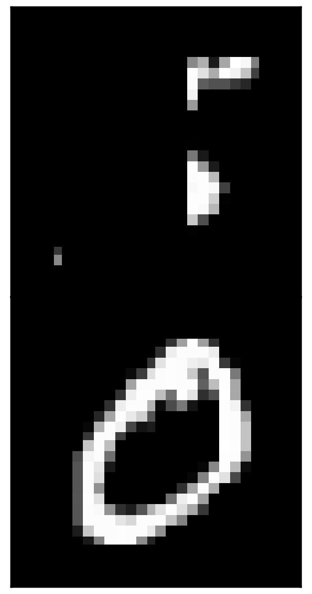
Out-of-distribution Dataset

Results

	AUROC				AUPR			
Detections	**Ent. y**	**Ent.μ**	**M.I.**	**D.U.**	**Ent. y**	**Ent.μ**	**M.I.**	**D.U.**
Misclassification	92.3	89.7	89.9	92.4	33.9	29.2	29.7	34.3
OOD	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
Misclass. vs OOD	100.0	100.0	100.0	99.3	100.0	100.0	100.0	100.0

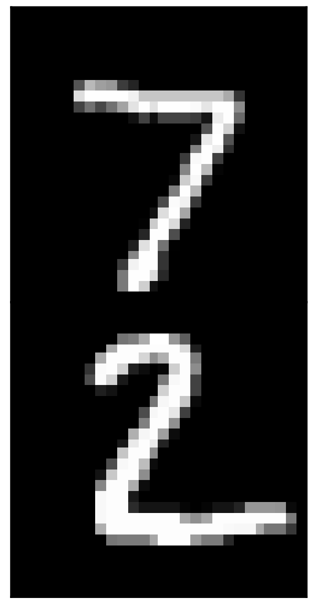
Corrupted MNIST

Setup

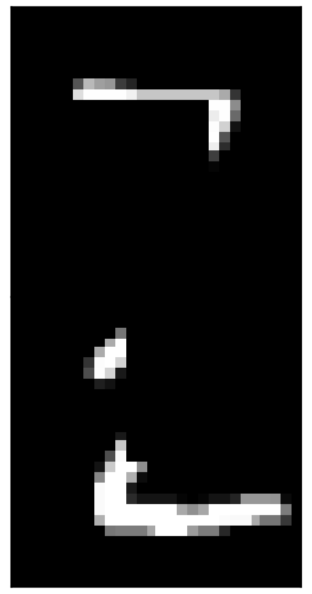


In-domain

Out-of-distribution



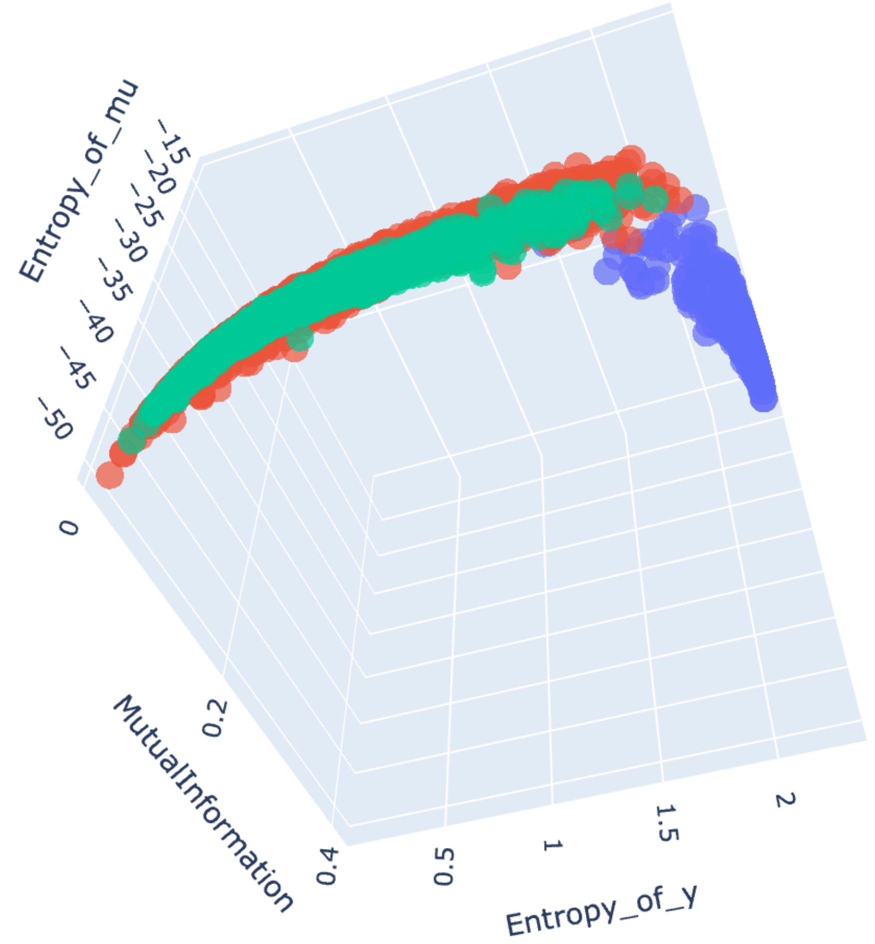
Test set 1



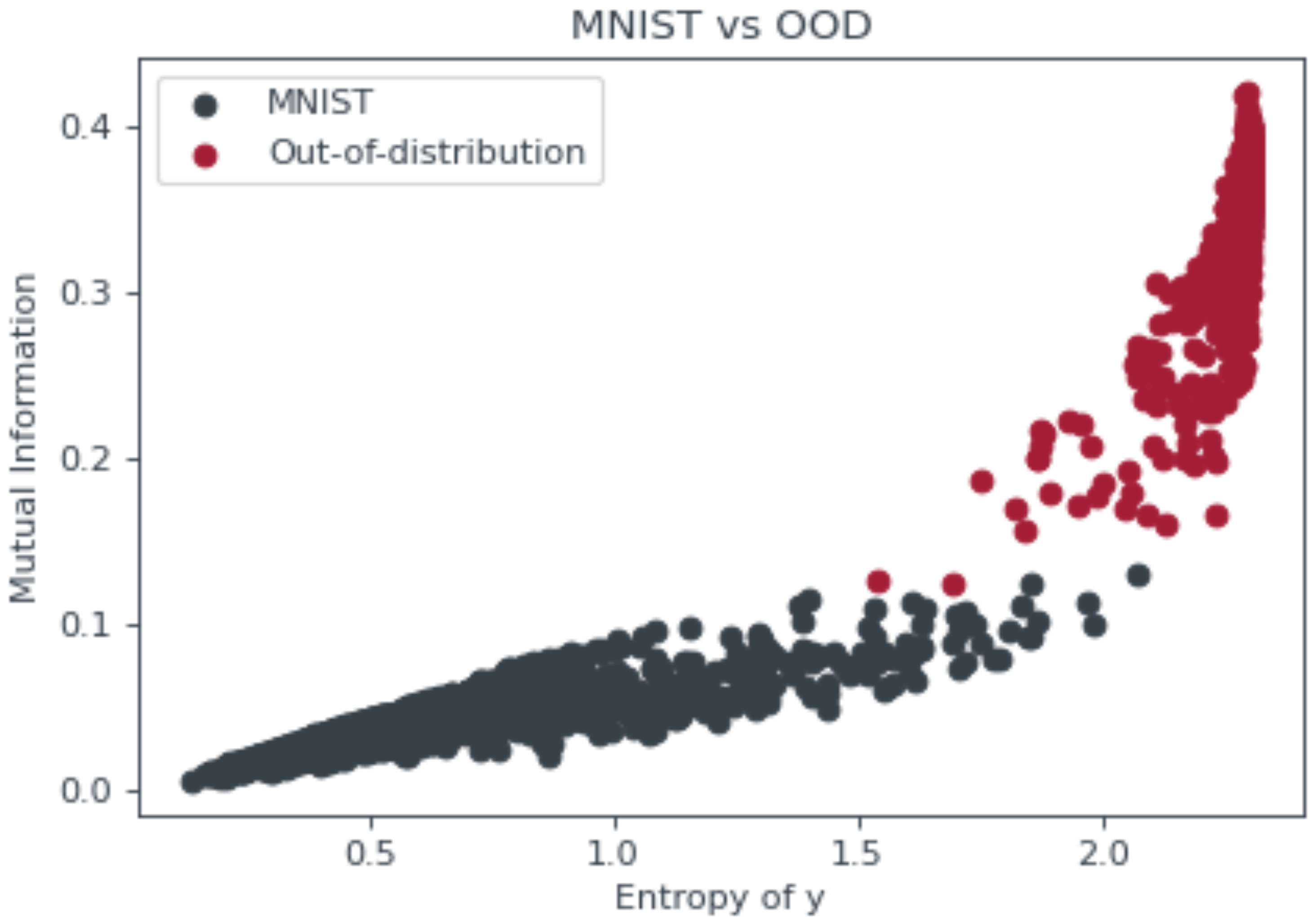
Test set 2

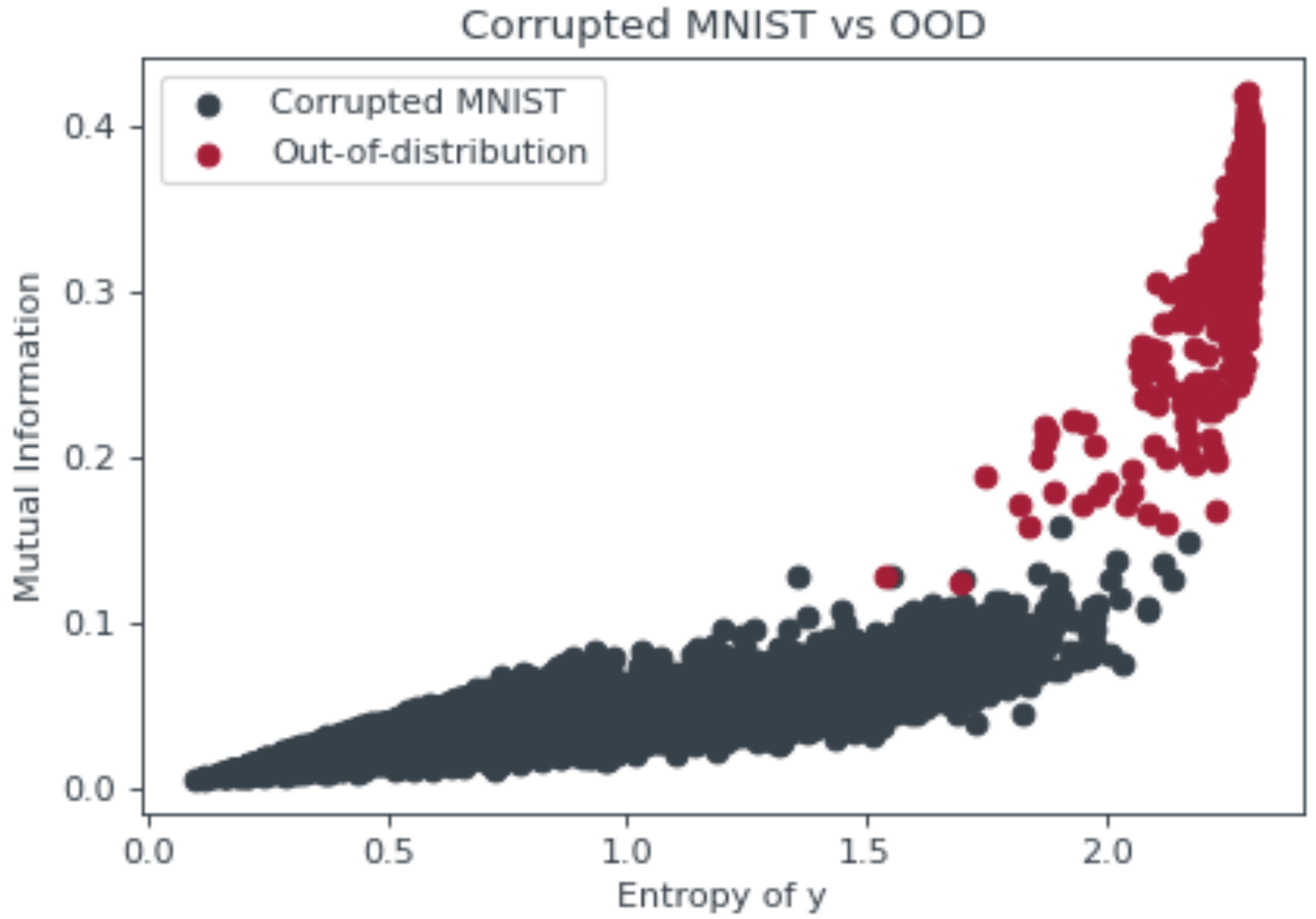
	Dataset	Size	Erased regions%
Training in domain	MNIST	60000	50
Training OOD	FashionMNIST	60000	0
Test set 1	MNIST	10000	0
Test set 2	MNIST	10000	100
Test set OOD	FashionMNIST	10000	0

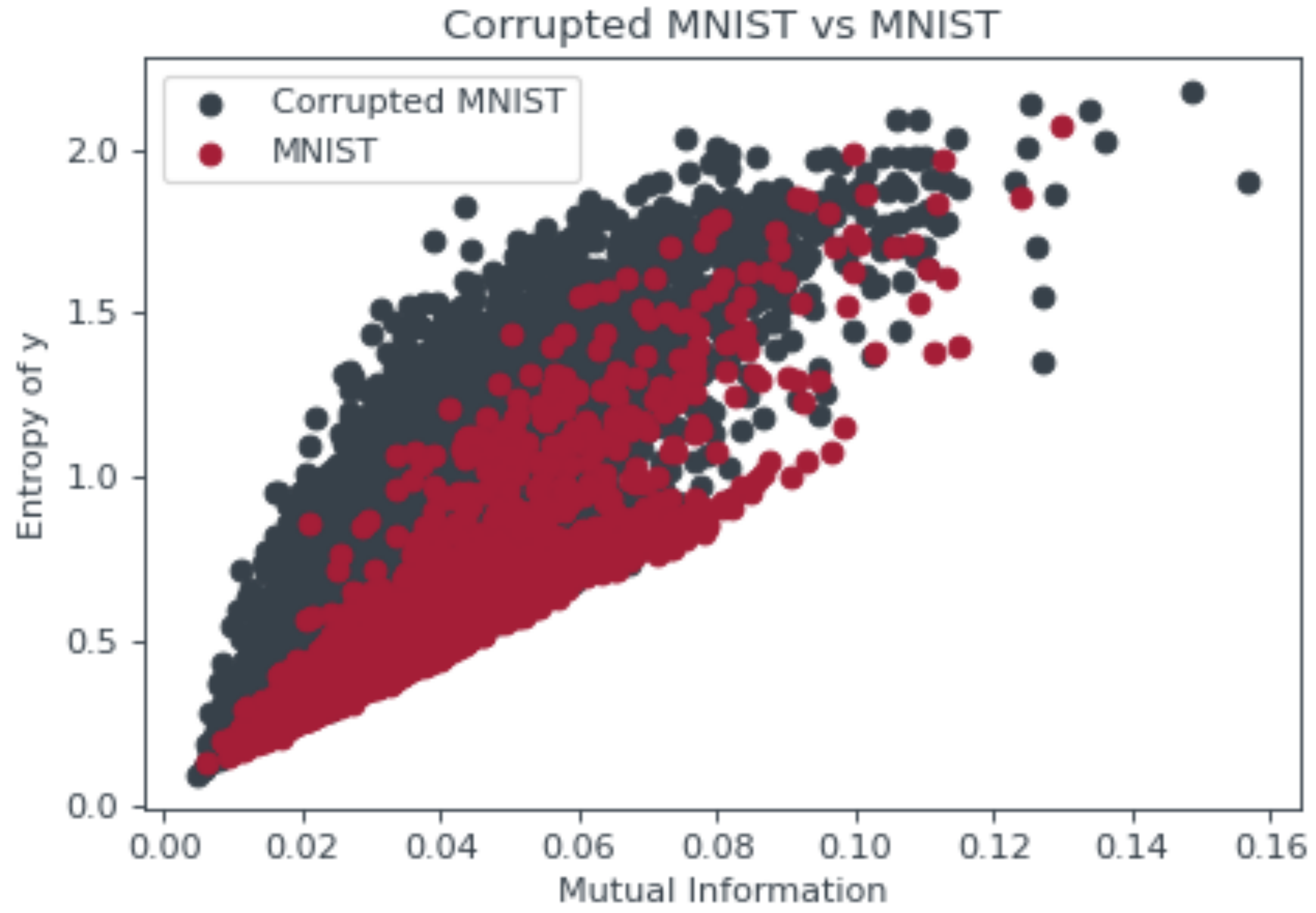
Results



Detections	Sets	AUROC Ent. y	AUROC Ent.μ	AUROC M.I.	AUROC D.U.	AUPR Ent. y	AUPR Ent.μ	AUPR M.I.	AUPR D.U.
In-domain	Corrupted vs MNIST	72.0	67.5	59.2	72.7	73.8	68.0	60.0	74.5
OOD	MNIST	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.9
	Corrupted MNIST	100.0	100.0	100.0	99.9	100.0	100.0	100.0	99.8







Conclusion

1. Dirichlet Prior Network can distinguish distributional uncertainty from others

2. This approach can not clearly distinguish between
    - in-domain data (MNIST)
    - in-domain data with data uncertainty (Corrupted MNIST)

Further work

1. Baseline
2. Out-of-distribution dataset

Thank you for your attention!