Bayes-TrEx: A Bayesian Sampling Approach to **Tr**ansparency by **Ex**ample

Serena Booth*

Yilun Zhou*

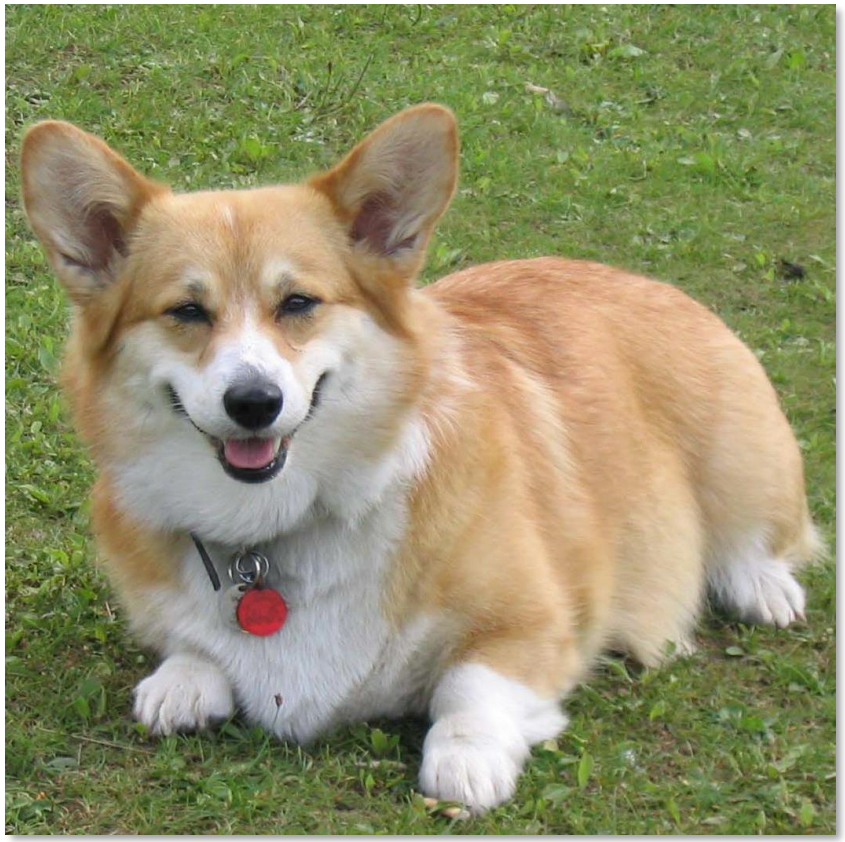
Ankit Shah

Julie Shah

AAAI 2021

Consider a Corgi/Bread Classifier

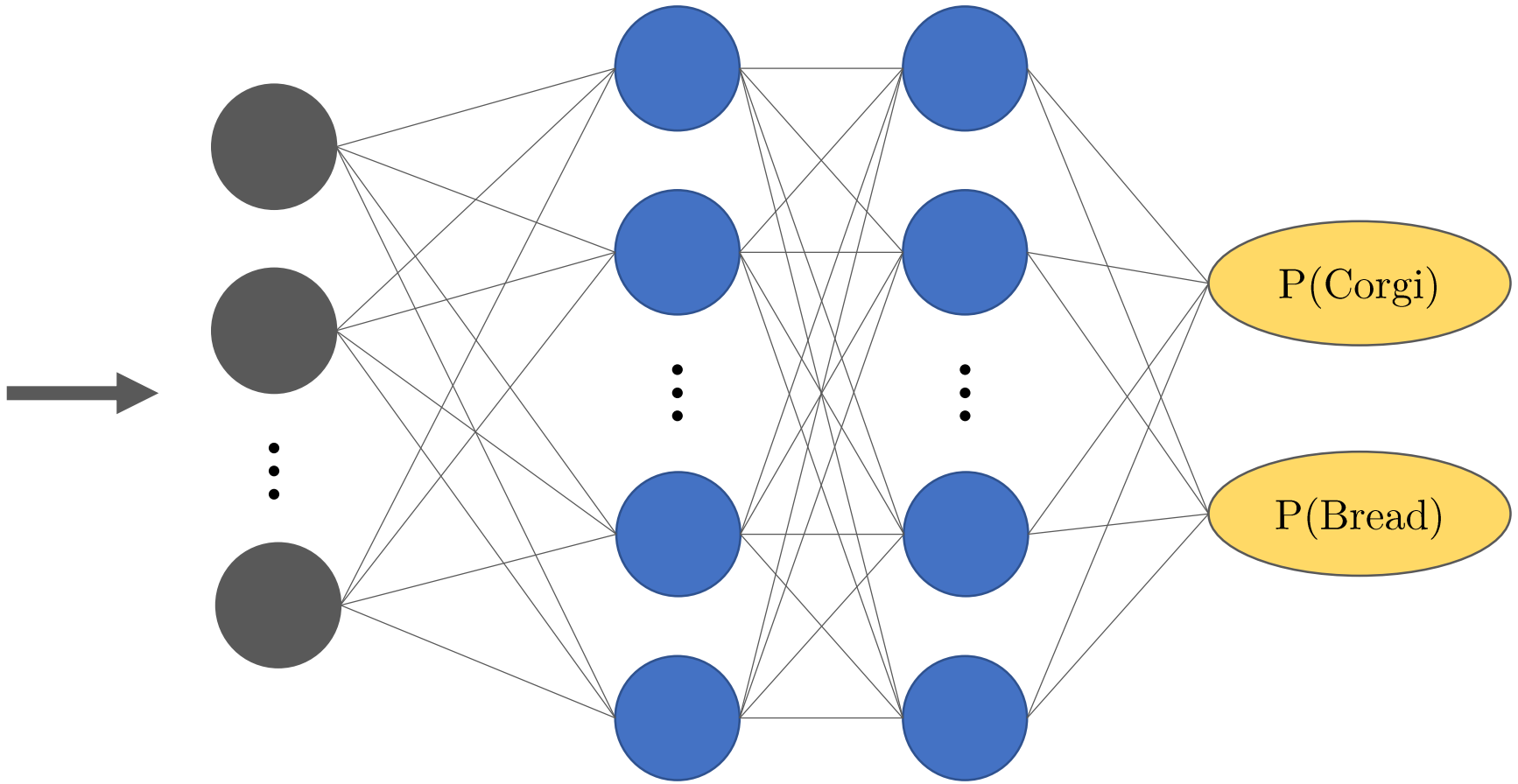
Consider a Corgi/Bread Classifier



Consider a Corgi/Bread Classifier

Consider a Corgi/Bread Classifier

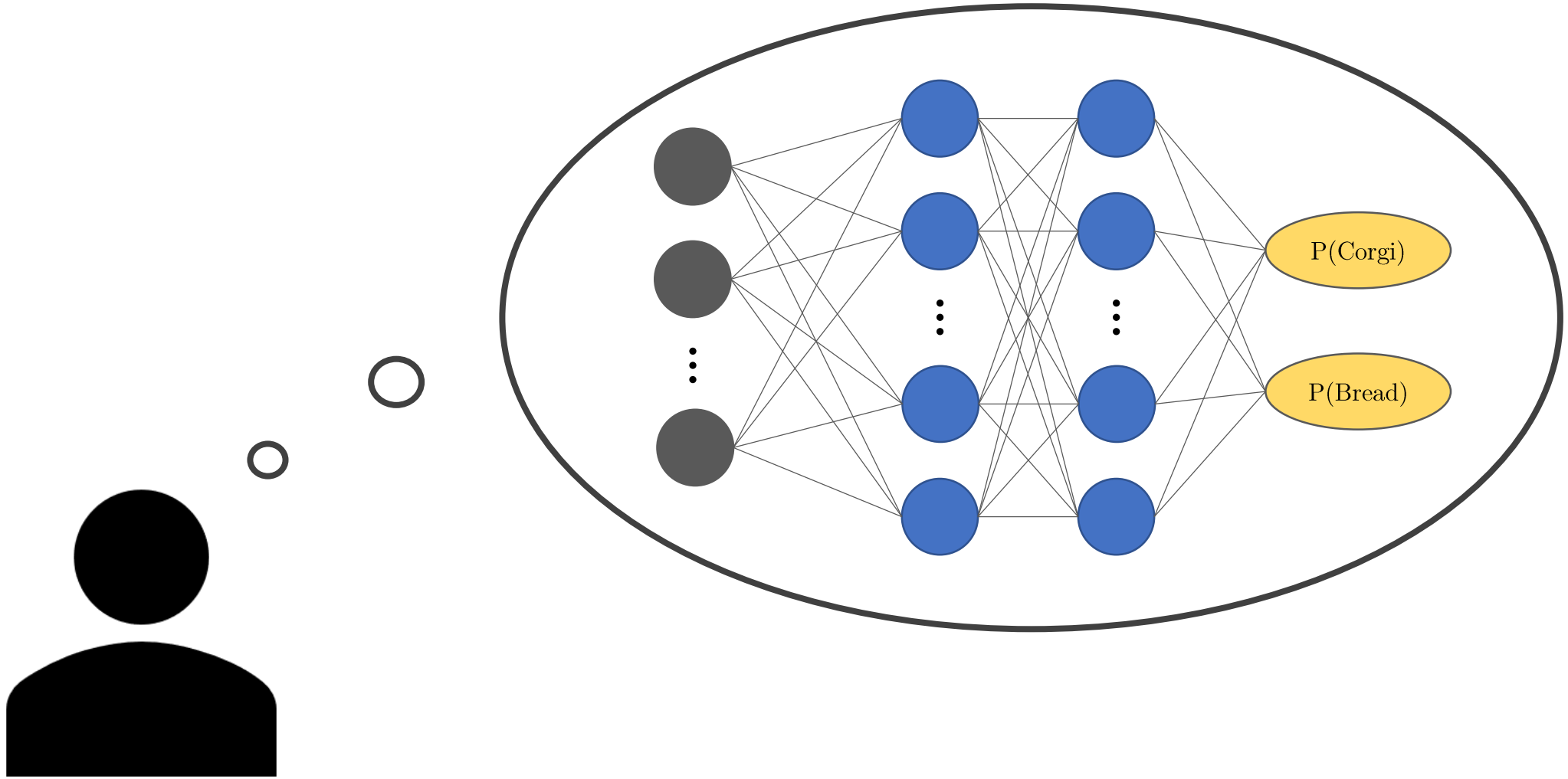
Input Image



Our objective:
Build a *holistic* understanding of the classifier.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Build a *holistic* understanding of the classifier.



Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

1. Develop an understanding of class boundaries.

1. Develop an understanding of class boundaries.

True class

Corgi	Bread	Corgi	Bread
			
51% Corgi, 49% Bread	48% Corgi, 52% Bread	50% Corgi, 50% Bread	45% Corgi, 55% Bread

Classifier's prediction

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

2. Predict classifier behavior on new examples.

2. Predict classifier behavior on new examples.

True class

Corgi

Bread

Corgi

Bread

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

2. Predict classifier behavior on new examples.

True class

Corgi	Bread	Corgi	Bread
			
Corgi	Bread	Bread	Corgi

Classifier's prediction

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

3. Predict classifier behavior on novel class instances.

3. Predict classifier behavior on novel class instances.

True class

Cat	Cake	Croissant	Potatoes
			
Corgi	Corgi	Bread	Bread

Classifier's prediction

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How can we build this understanding?

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

50% Confident? Few images.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Search the test set: ~50% Corgi, ~50% Bread

Test set? Many images.

50% Confident? Few images.

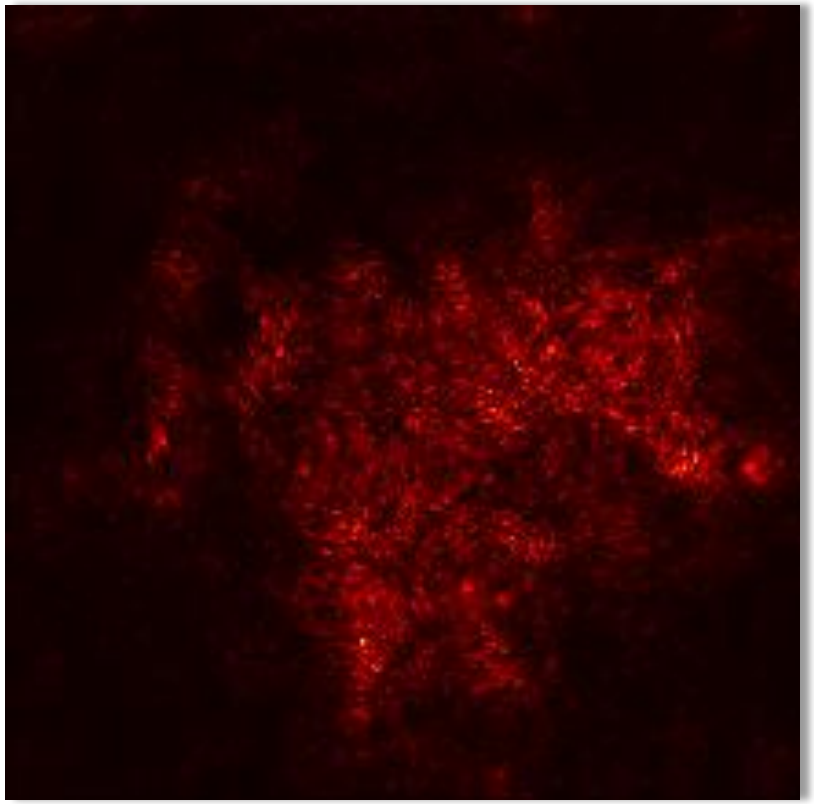
Problem: sparse data

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How can we build a holistic understanding?

Input

Explanation pipeline



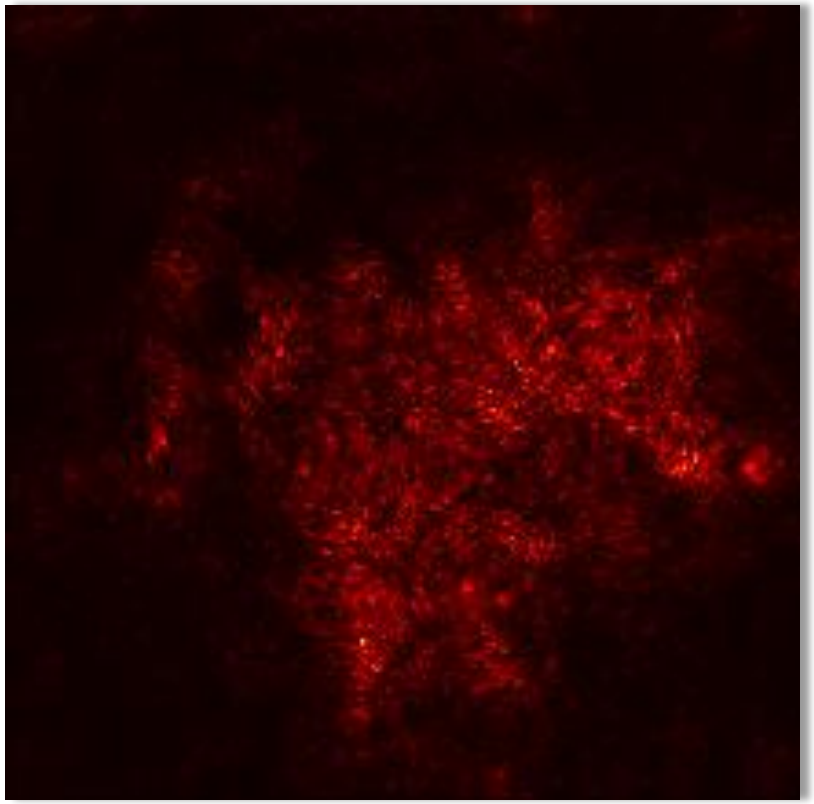
Saliency Map

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How can we build a holistic understanding?

Input

Explanation pipeline



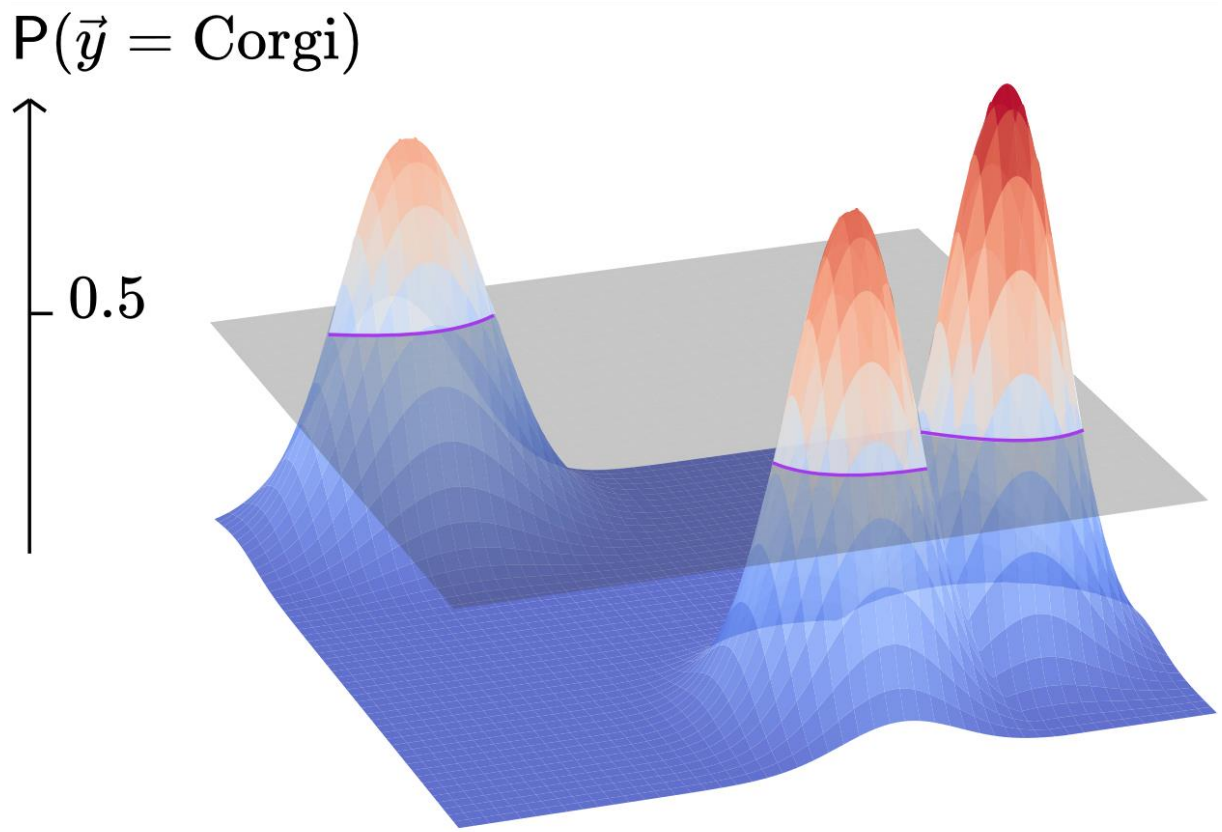
Saliency Map

Problem: where does the test input come from?

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

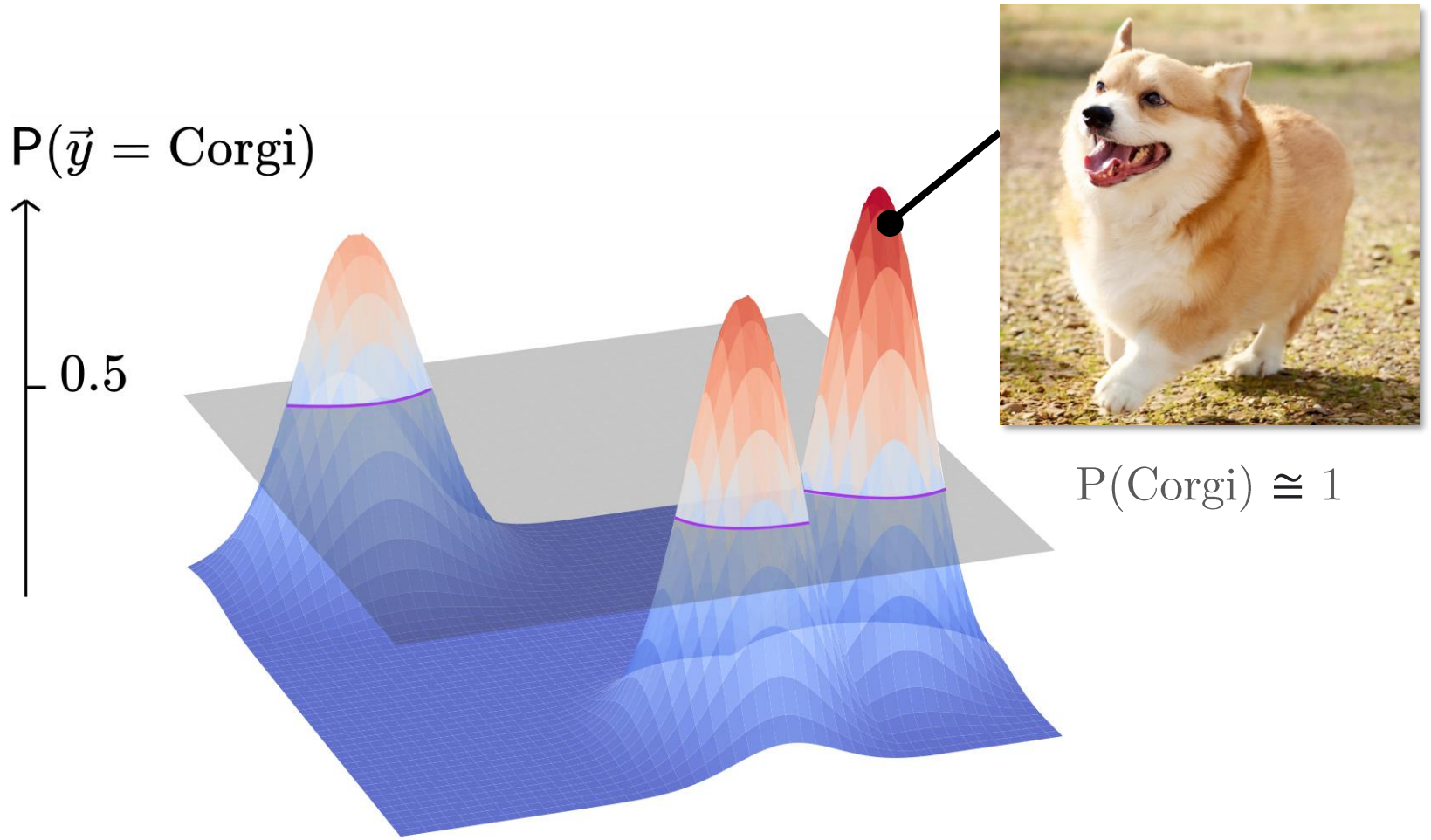
We introduce Bayes-TrEx: an approach to transparency by example

We introduce Bayes-TrEx: an approach to transparency by example



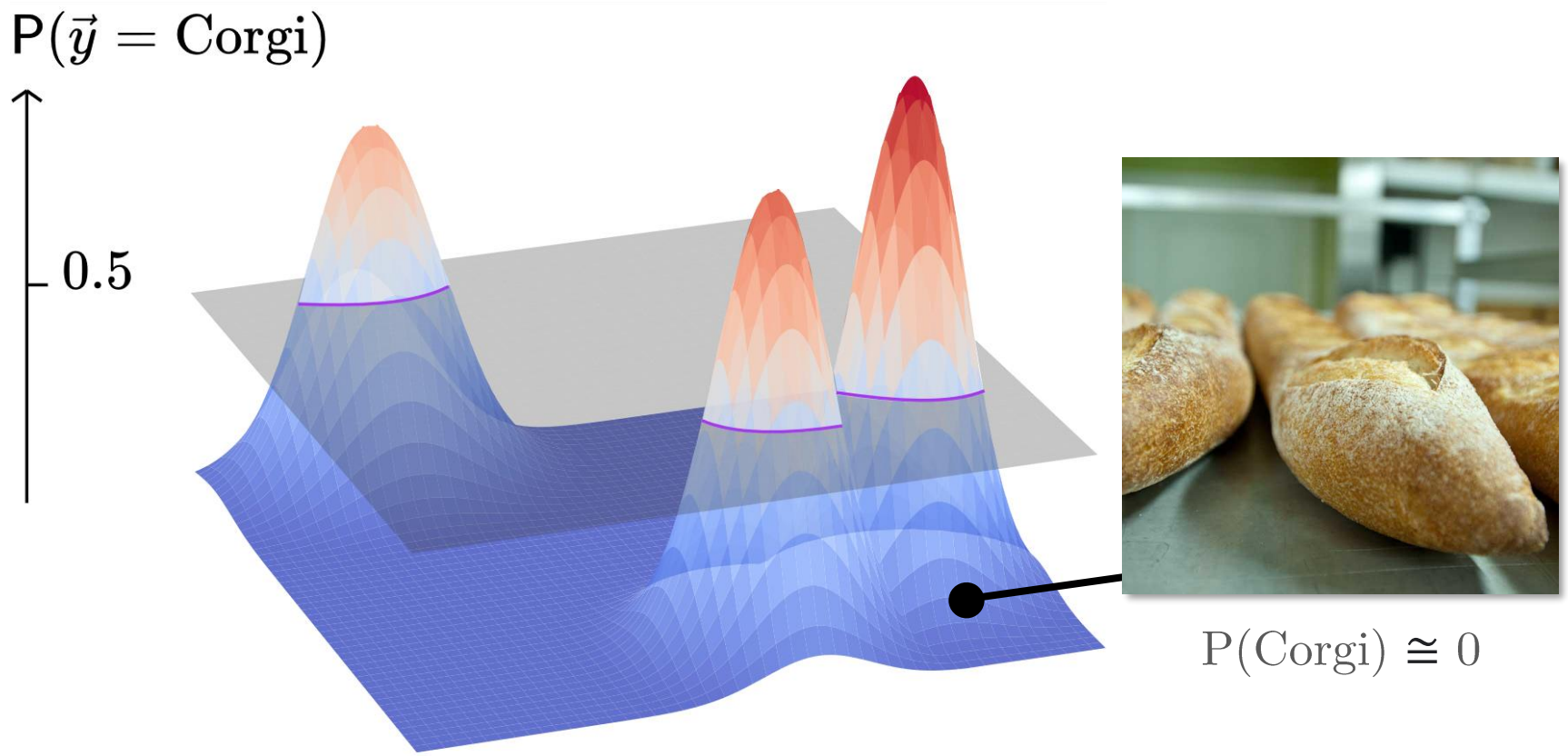
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

A Corgi/Bread Decision Surface



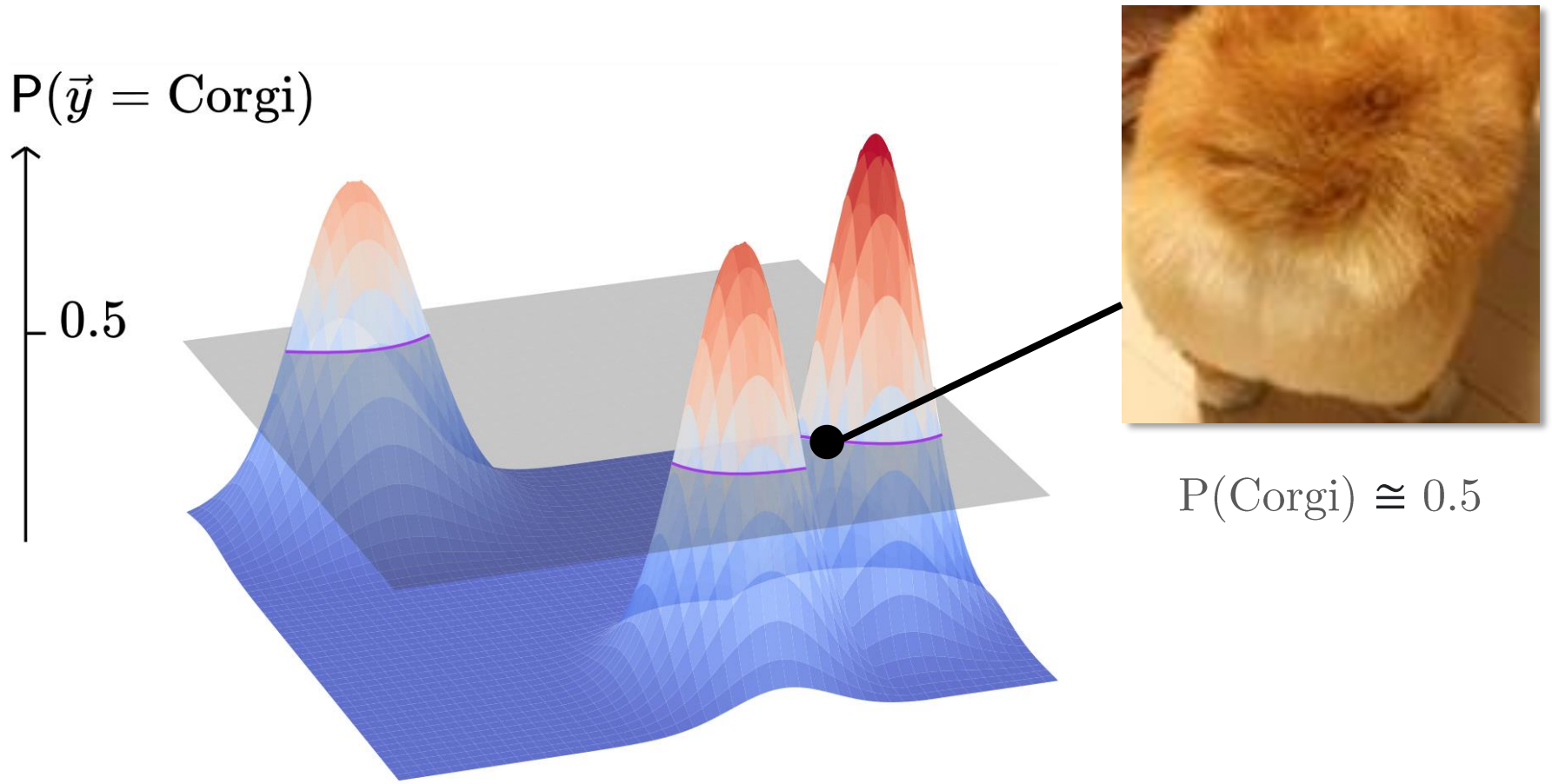
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

A Corgi/Bread Decision Surface



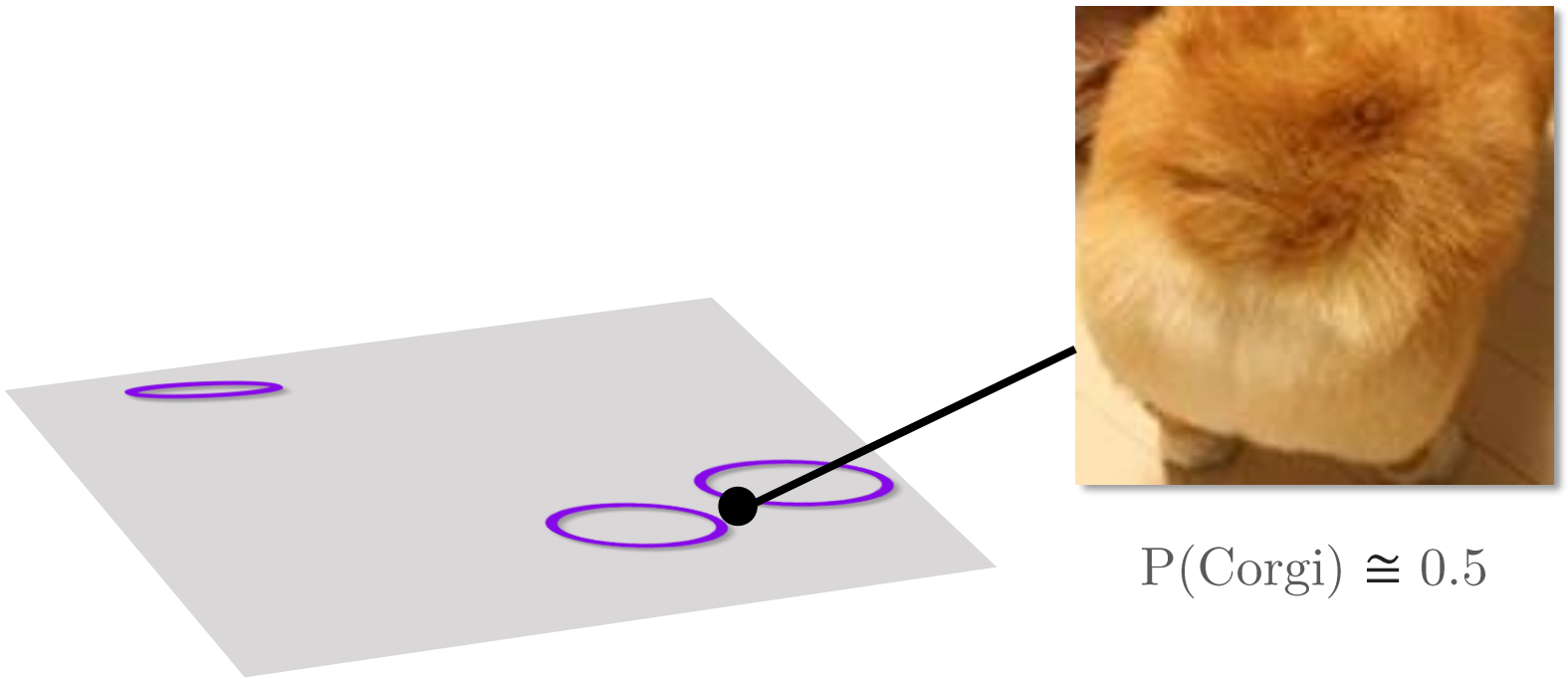
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

A Corgi/Bread Decision Surface



Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

A Corgi/Bread Decision Surface

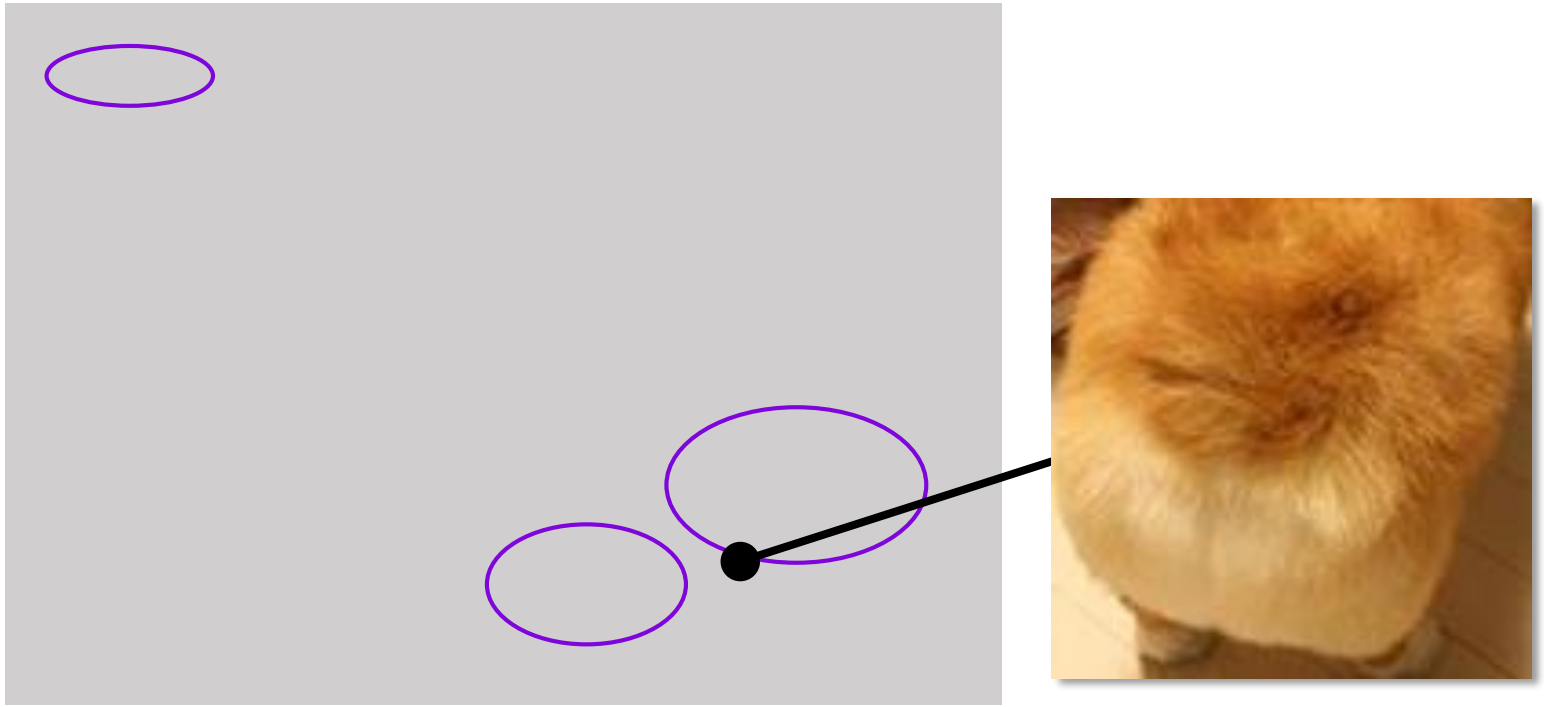


P(Corgi) ≅ 0.5

P(Corgi) = 0.5 Level Set
True Posterior

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We find prediction-matching examples from **p-level sets**

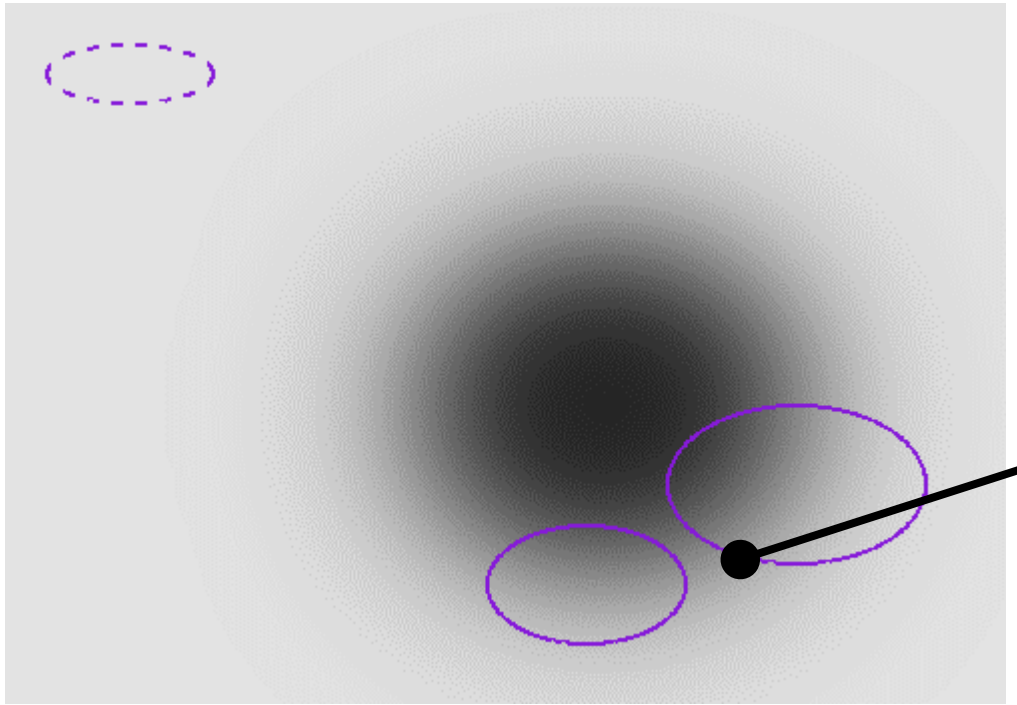


P(Corgi) = 0.5 Level Set
True Posterior

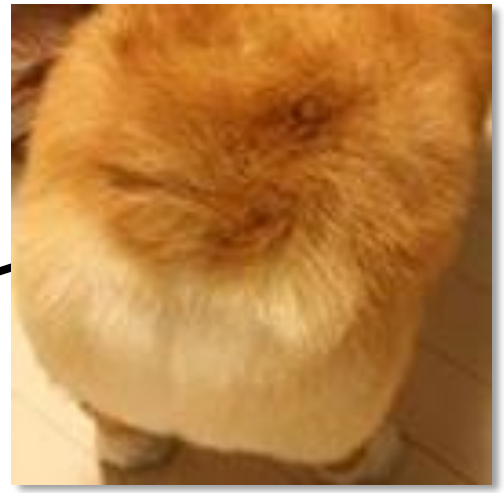
P(Corgi) ≅ 0.5

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We find prediction-matching examples from **p-level sets**



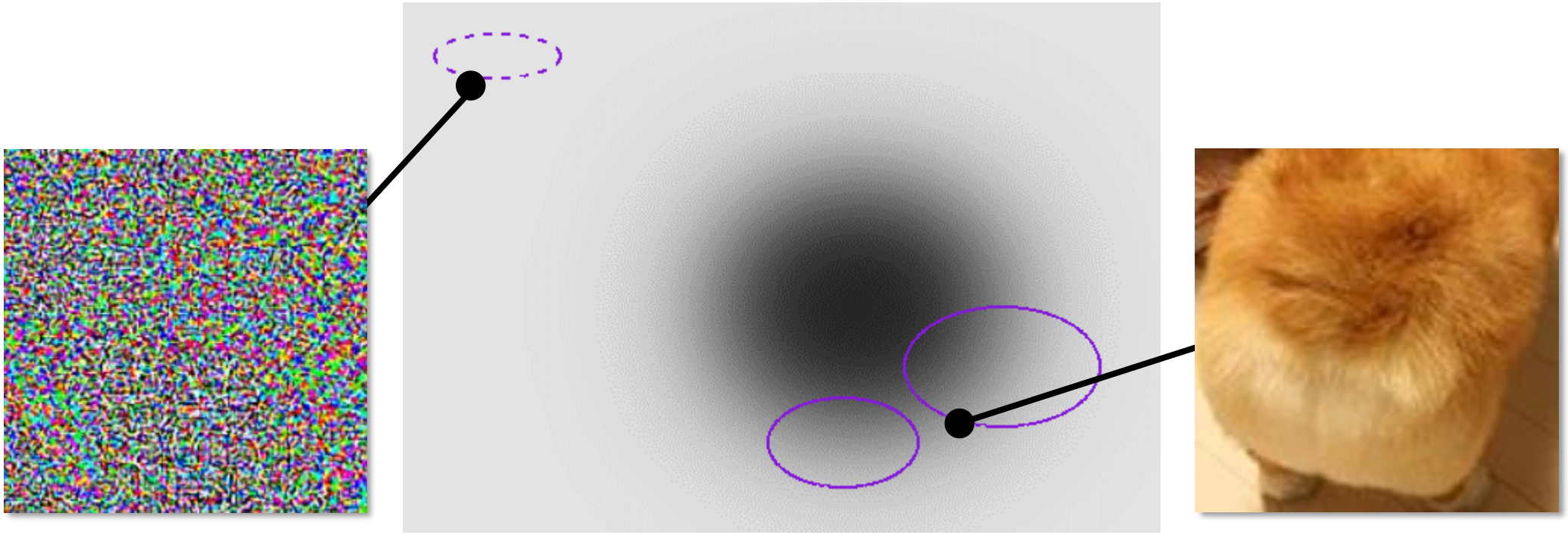
P(Corgi) = 0.5 Level Set
True Posterior



P(Corgi) ≅ 0.5

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We find prediction-matching examples from **p-level sets**



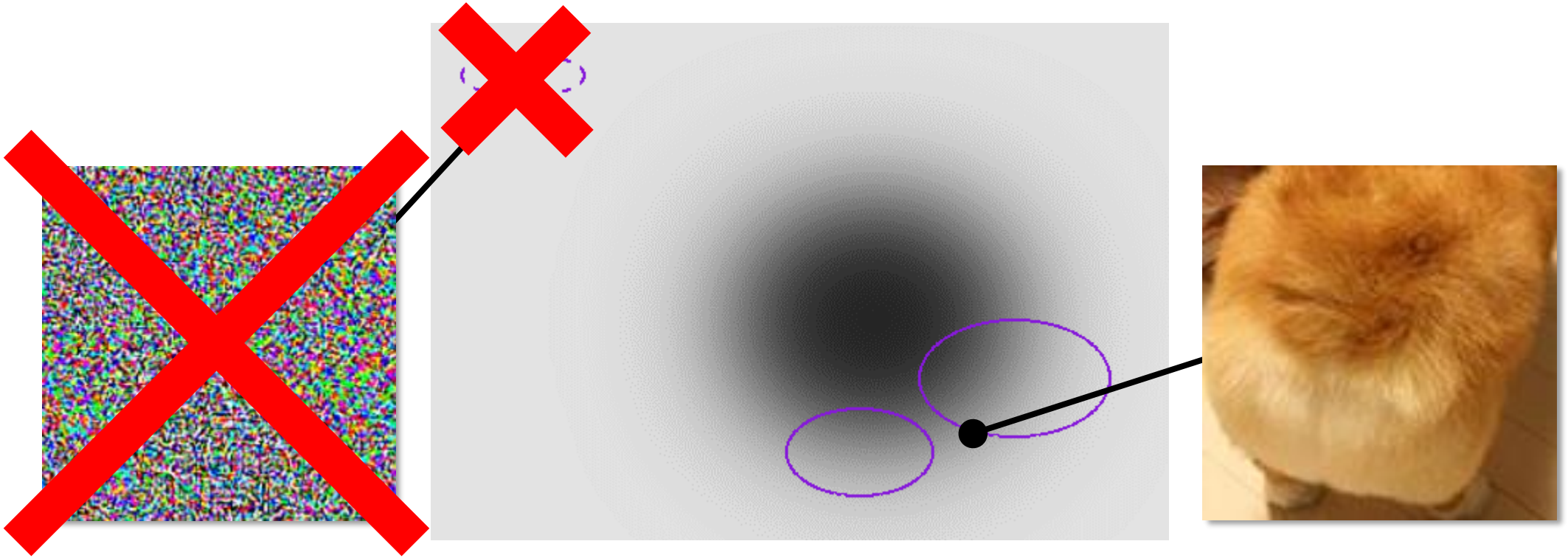
P(Corgi) ≅ 0.5

P(Corgi) = 0.5 Level Set
True Posterior

P(Corgi) ≅ 0.5

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We find prediction-matching examples from **p-level sets**



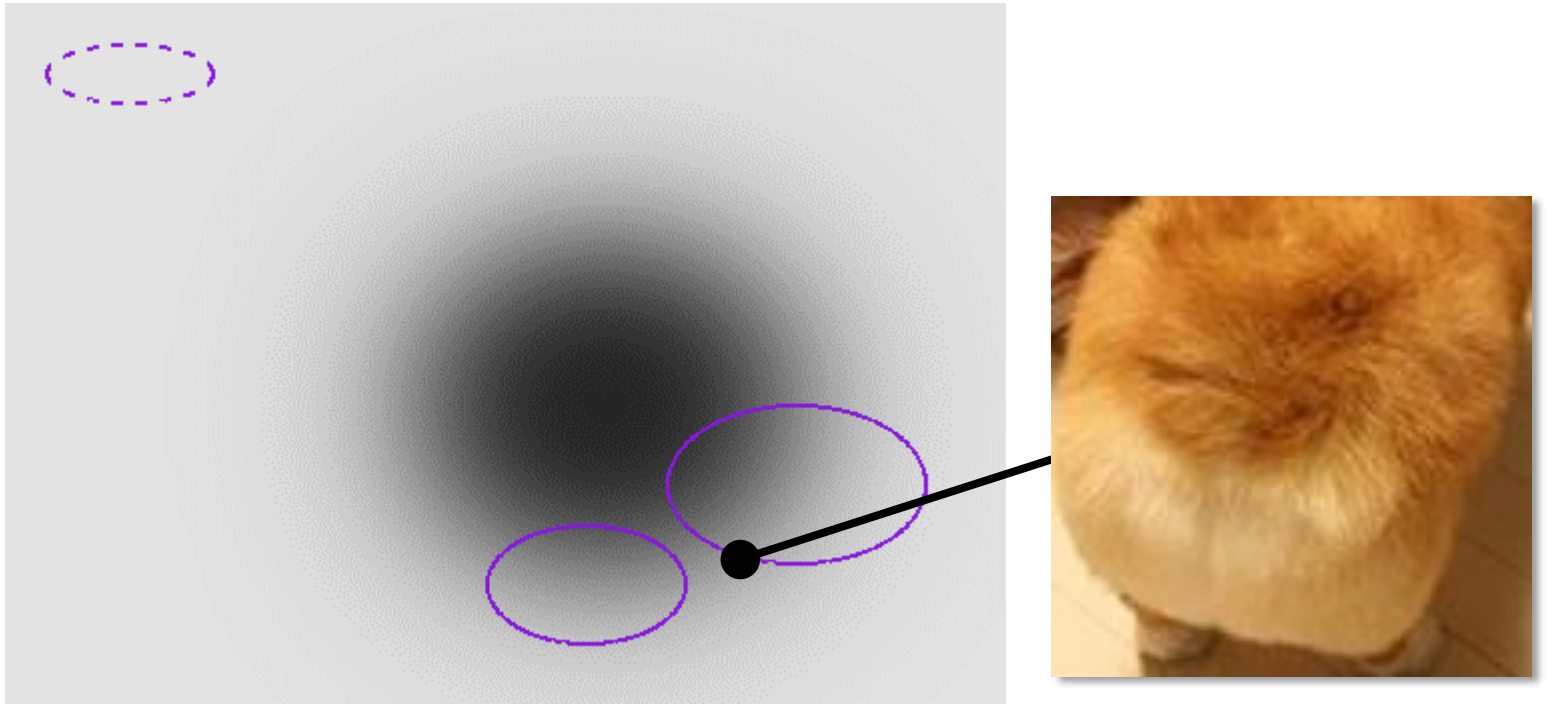
P(Corgi) ≅ 0.5

P(Corgi) = 0.5 Level Set
True Posterior

P(Corgi) ≅ 0.5

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We want to find a *natural* example $\mathbf{x}$ where the classifier $f(\mathbf{x})$ has confidence $\mathbf{p}$

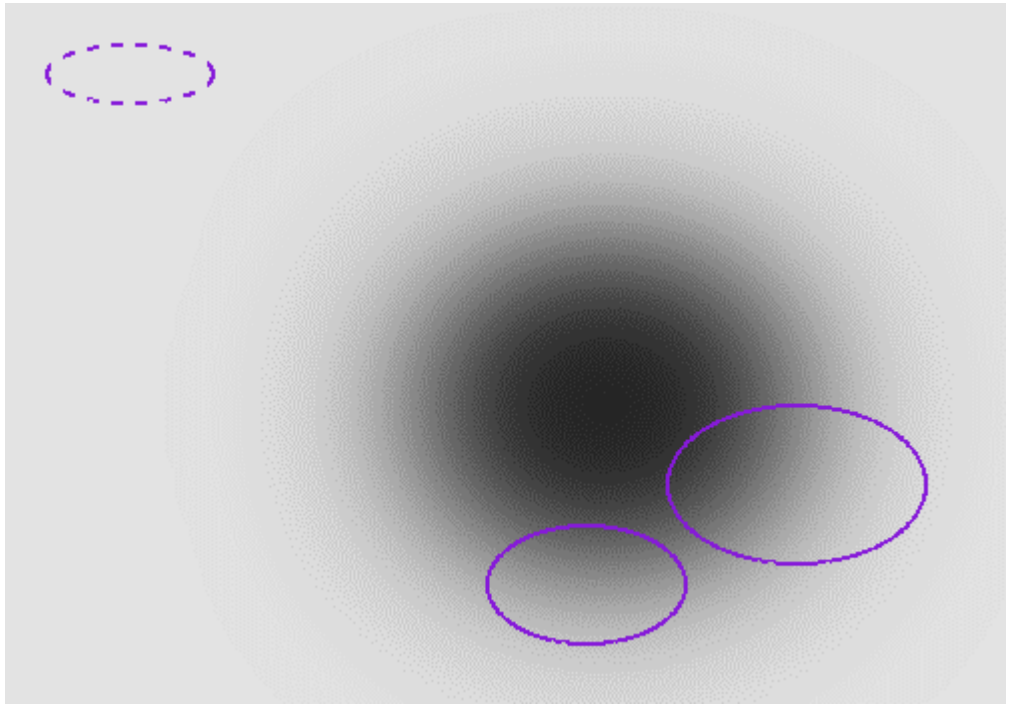


P(Corgi) $\cong$ 0.5

P(Corgi) = 0.5 Level Set
True Posterior

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

We want to find a *natural* example $\mathbf{x}$ where the classifier $f(\mathbf{x})$ has confidence $\mathbf{p}$



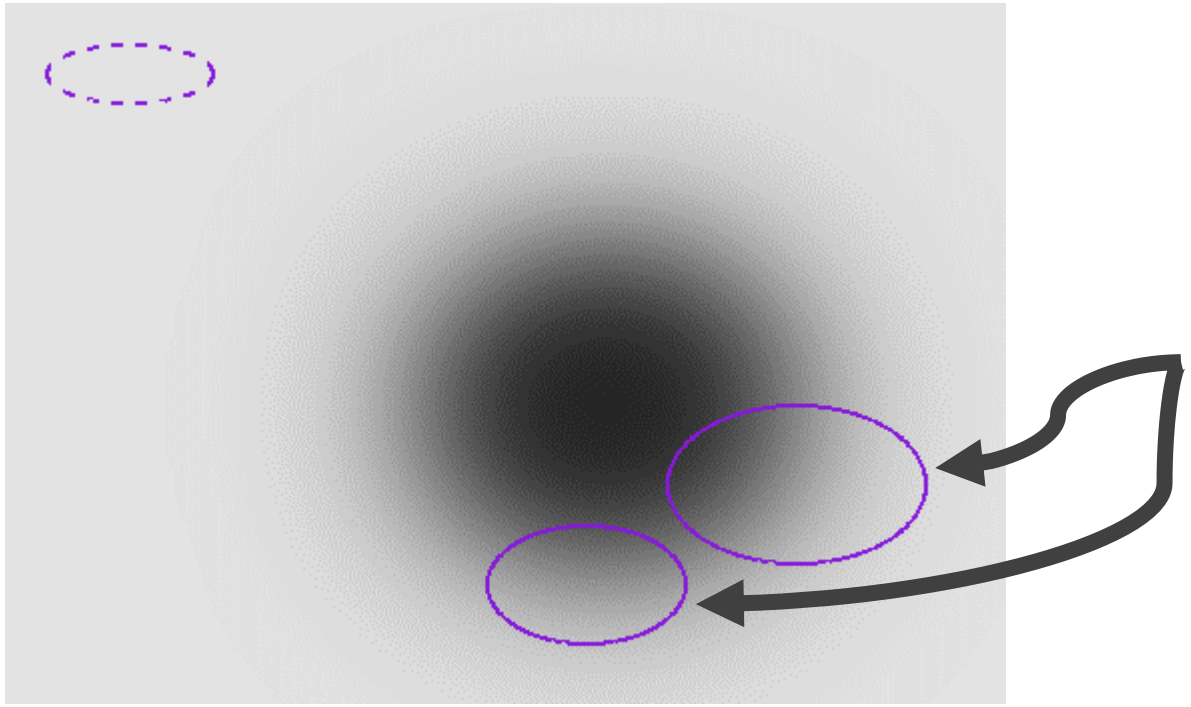
P(Corgi) = 0.5 Level Set
True Posterior

Want to sample from:

$$p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x})\, p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x})$$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Want to sample from: $p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x})\, p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x})$



P(Corgi) = 0.5 Level Set
True Posterior

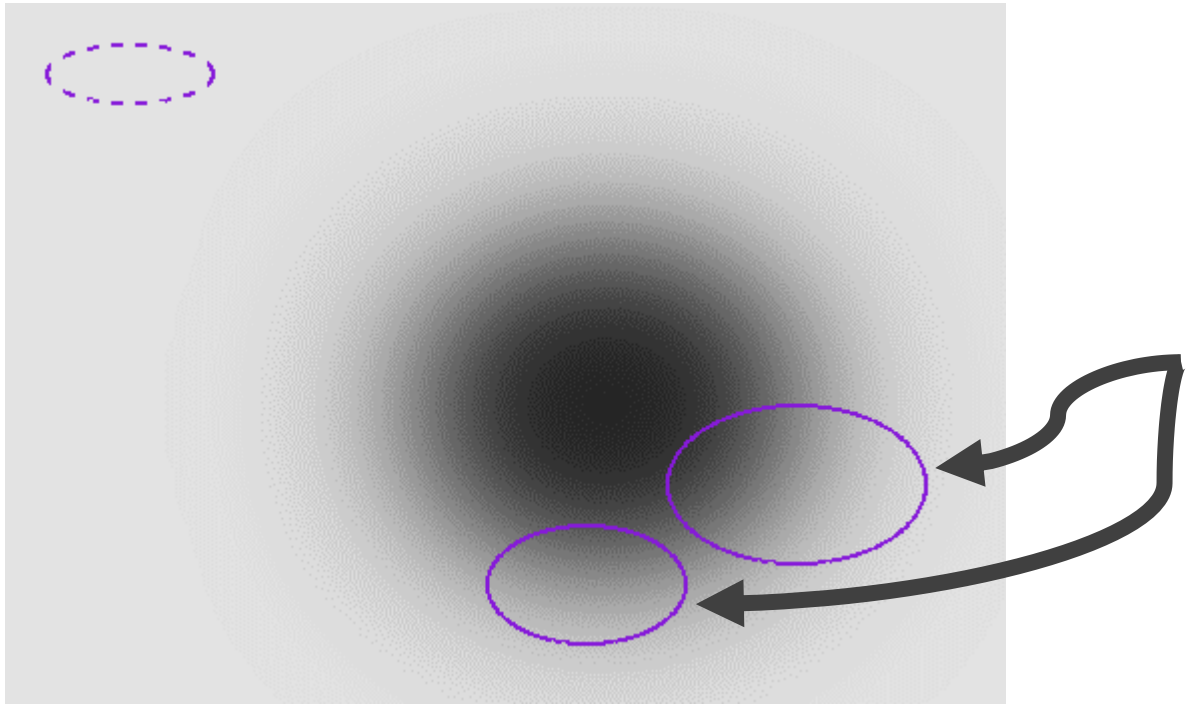
Problem 1:

$$\{\mathbf{x} : f(\mathbf{x}) = \mathbf{p}\}$$

has small or even zero measure

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Want to sample from: $p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x})\, p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x})$



P(Corgi) = 0.5 Level Set
True Posterior

Problem 1:

$\{\mathbf{x} : f(\mathbf{x}) = \mathbf{p}\}$

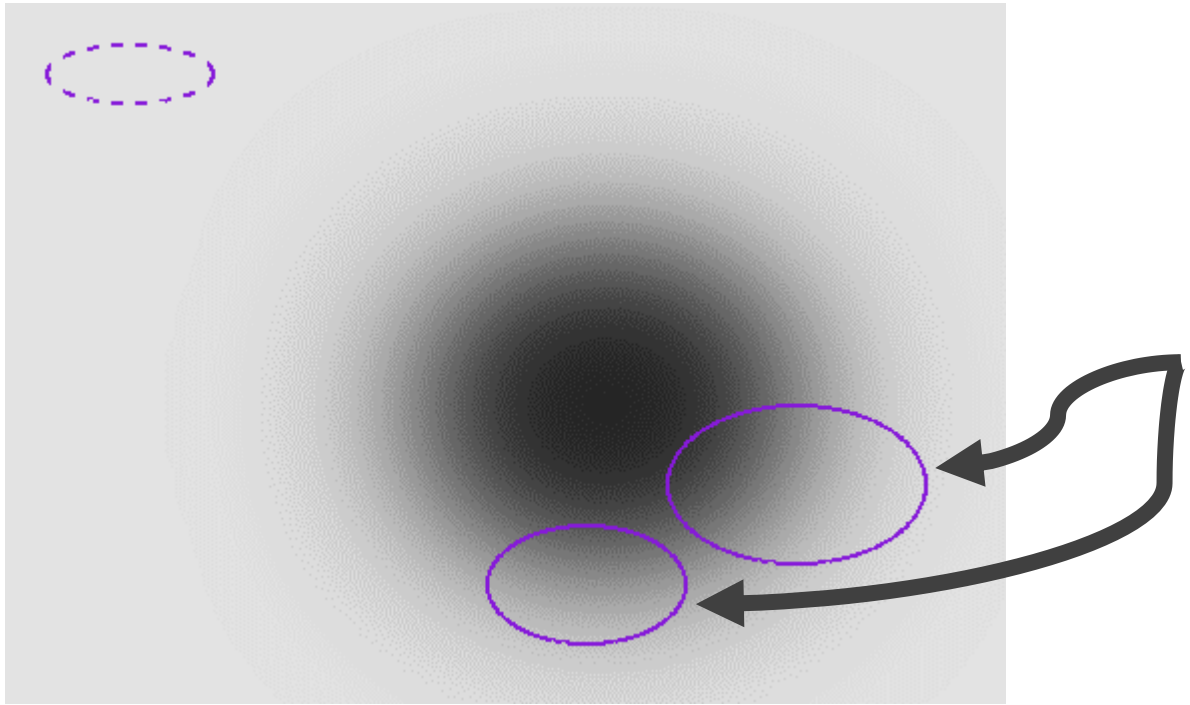
has small or even zero measure

Problem 2:

$\mathbf{x}$ is too high-dimensional

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Want to sample from: ~~$p(\mathbf{x}|f(\mathbf{x}) = \mathbf{p}) \propto p(\mathbf{x})p(f(\mathbf{x}) = \mathbf{p}|\mathbf{x})$~~



P(Corgi) = 0.5 Level Set
True Posterior

Problem 1:

$$\{\mathbf{x} : f(\mathbf{x}) = \mathbf{p}\}$$

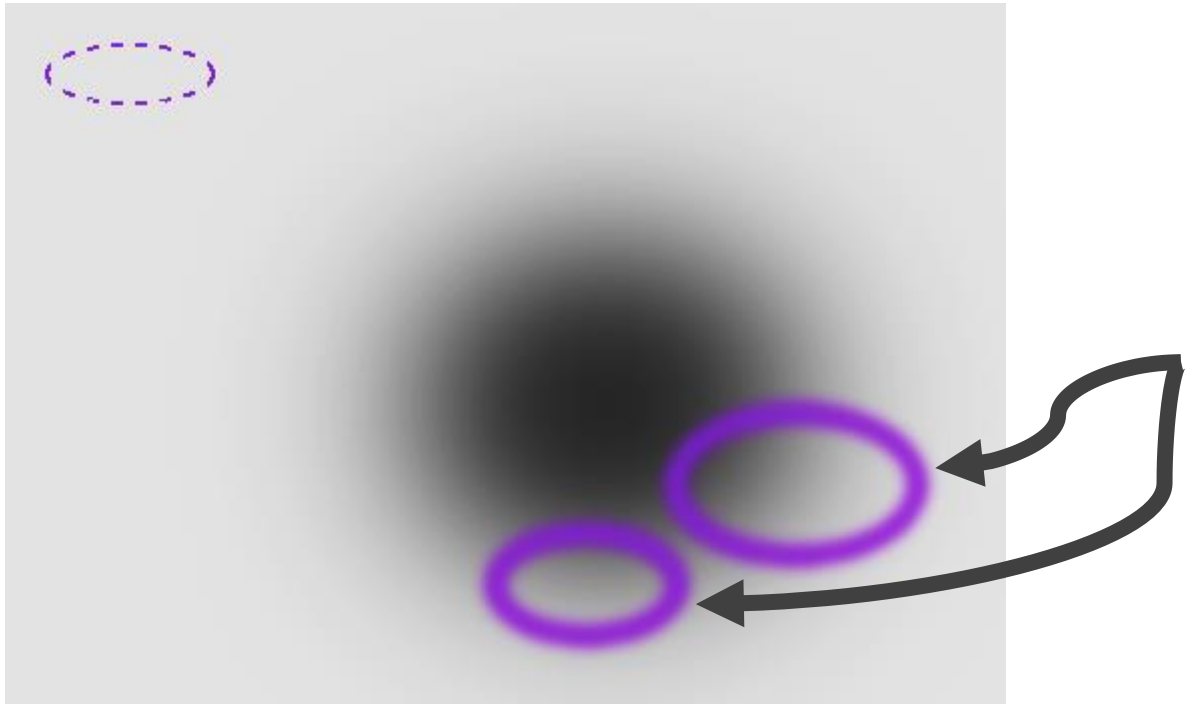
has small or even zero measure

Problem 2:

$\mathbf{x}$ is too high-dimensional

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Make inference tractable

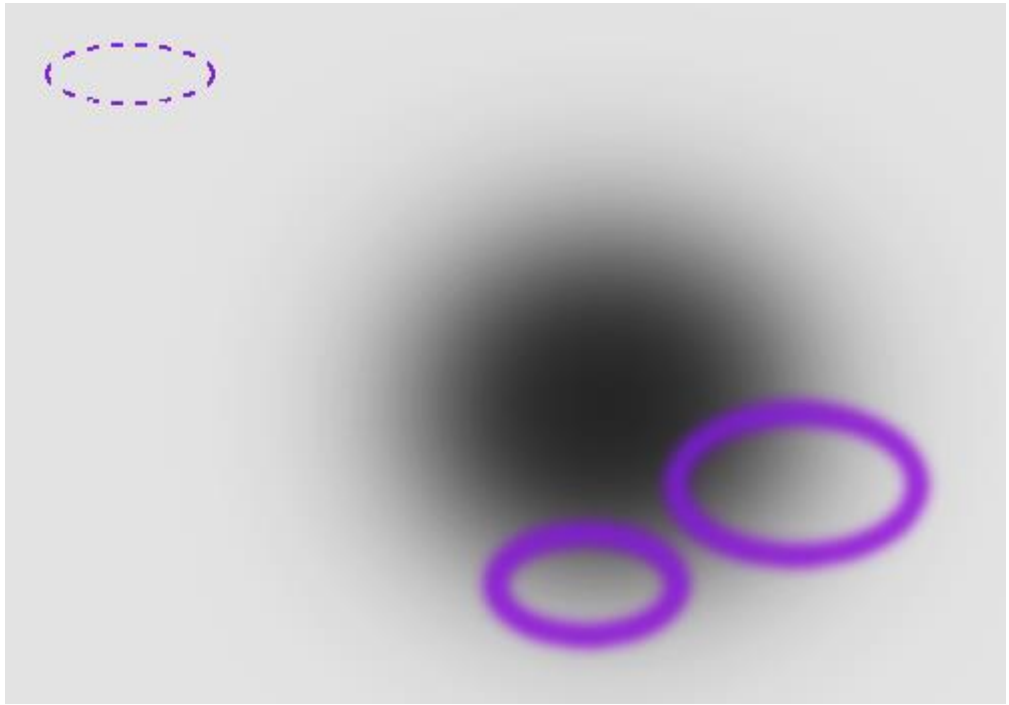


Relax the formulation by "widening" the level set

P(Corgi) = 0.5 Level Set
Relaxed Posterior

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Relaxed Formulation



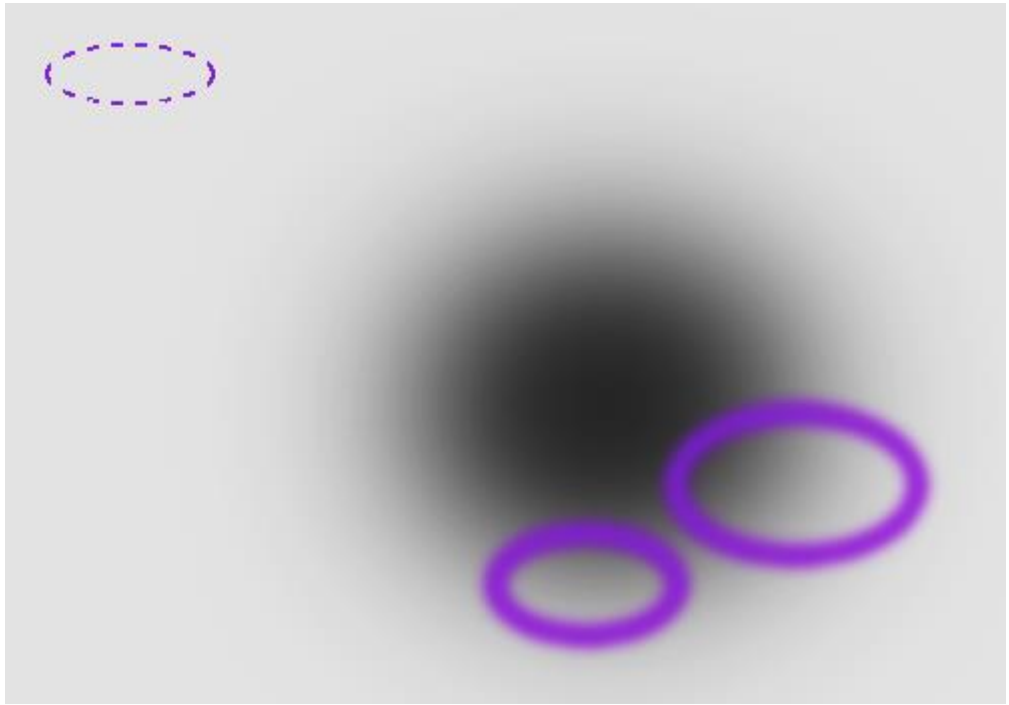
P(Corgi) = 0.5 Level Set
Relaxed Posterior

Introduce a random vector:

$$\mathbf{u}|\mathbf{x} \sim \mathcal{N}\left(f(\mathbf{x}), \sigma^2\right)$$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Relaxed Formulation



P(Corgi) = 0.5 Level Set
Relaxed Posterior

Introduce a random vector:

$$\mathbf{u}|\mathbf{x} \sim \mathcal{N}\left(f(\mathbf{x}), \sigma^2\right)$$

And sample from the new posterior:

$$p(\mathbf{x}|\mathbf{u} = \mathbf{u}^*) \propto p(\mathbf{x})p(\mathbf{u} = \mathbf{u}^*|\mathbf{x})$$

$$\mathbf{u}^* = \mathbf{p}$$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

But, how can we sample an image $\mathbf{x}$?

But, how can we sample an image $\mathbf{x}$?

We sample $\mathbf{z}$ from a latent space Z, instead.

$$\mathbf{x} = g(\mathbf{z})$$

But, how can we sample an image $\mathbf{x}$?

We sample $\mathbf{z}$ from a latent space Z, instead.

$$\mathbf{x} = g(\mathbf{z})$$

$$\mathbf{u}|\mathbf{z} \sim N(f(\mathbf{x}), \sigma^2)$$

$$p(\mathbf{z}|\mathbf{u} = \mathbf{u}^*) \propto p(\mathbf{z})p(\mathbf{u} = \mathbf{u}^*|\mathbf{z})$$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

To use Bayes-TrEx, we need 3 requirements:

To use Bayes-TrEx, we need 3 requirements:

1. A classifier which outputs class probabilities

To use Bayes-TrEx, we need 3 requirements:

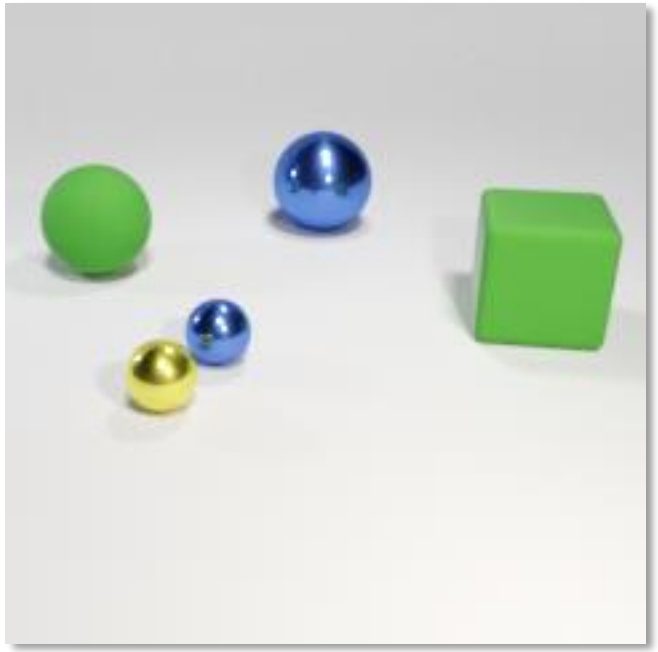
1. A classifier which outputs class probabilities
2. A user-specified confidence target

To use Bayes-TrEx, we need 3 requirements:

1. A classifier which outputs class probabilities
2. A user-specified confidence target
3. A data distribution we can sample from

Experiments

CLEVR

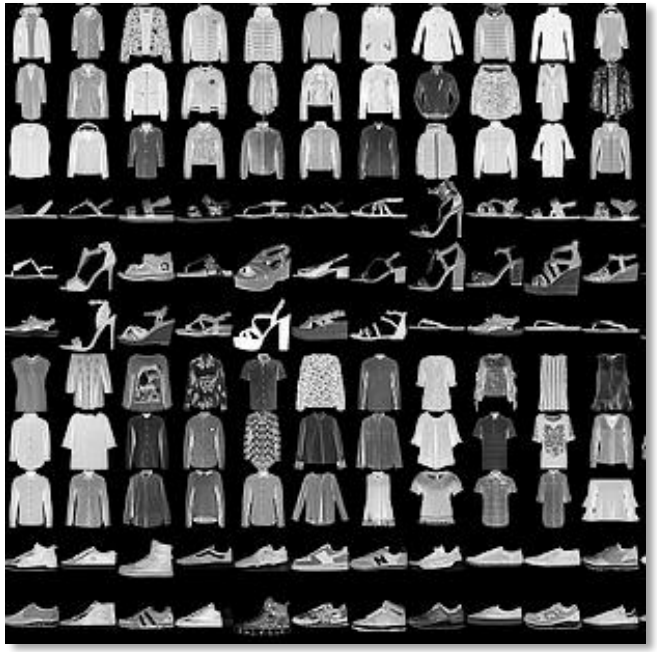


Scene graph

MNIST

VAEs, GANs

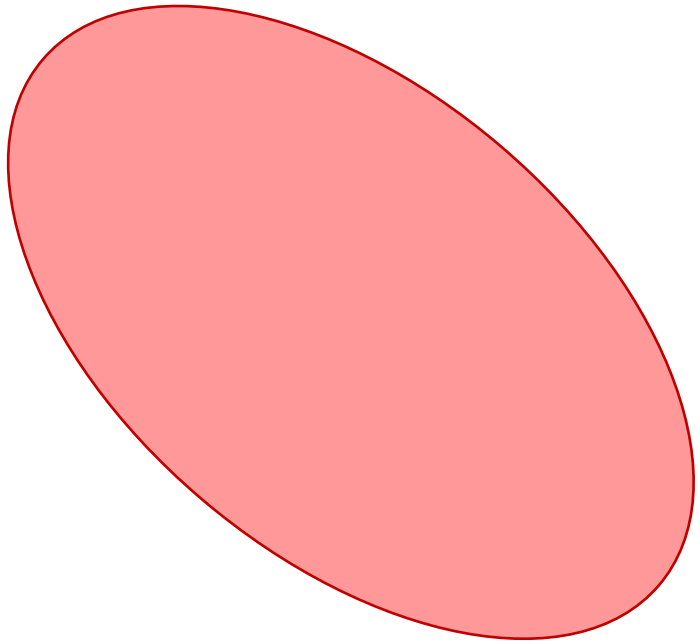
Fashion-MNIST



VAEs, GANs

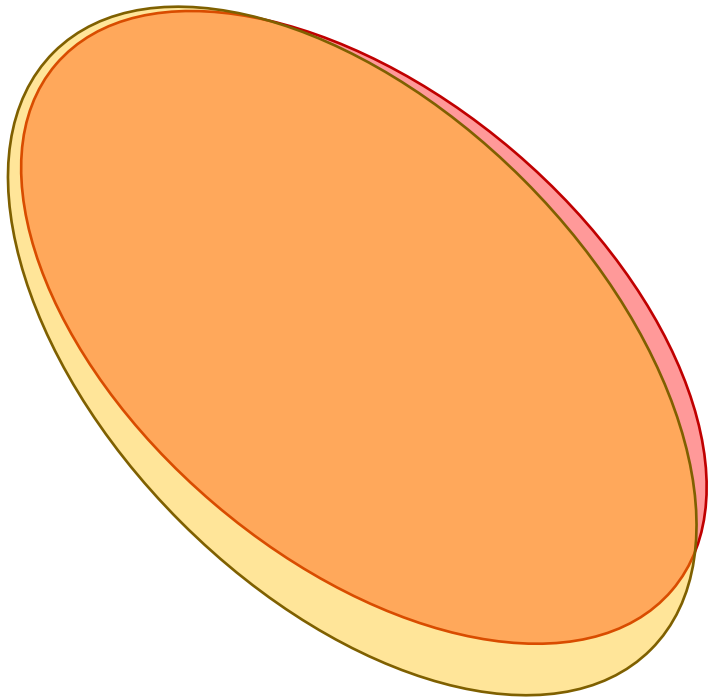
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)



$\mathbb{P}_C$, classifier training distribution, red

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)

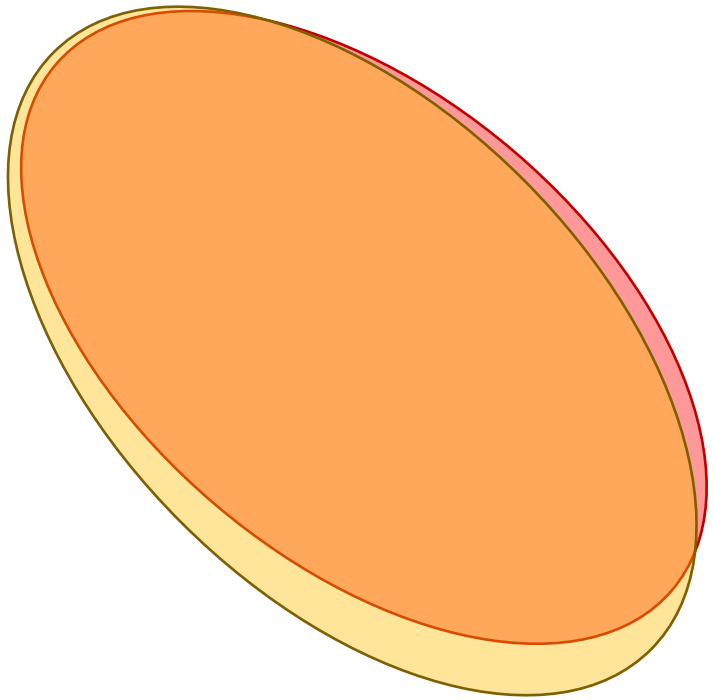


$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)



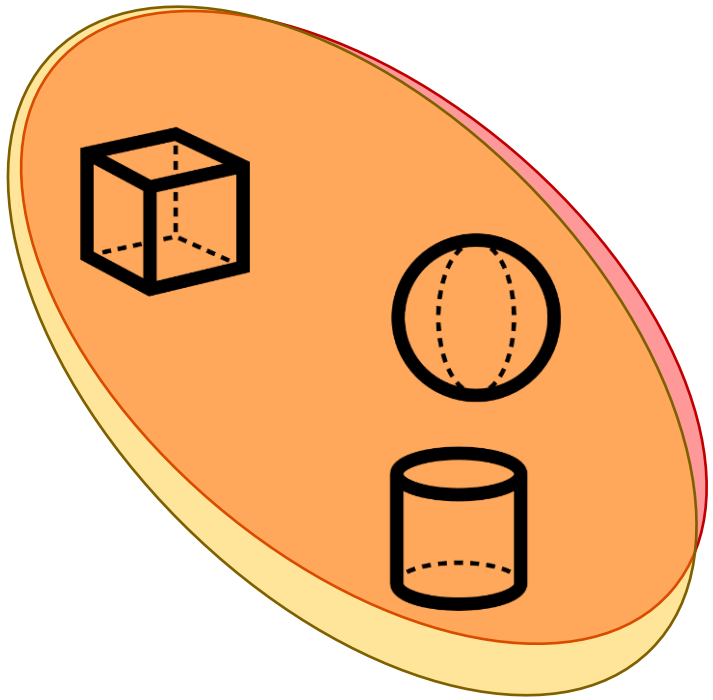
$\mathbb{P}_D \approx \mathbb{P}_C$

$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)



$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

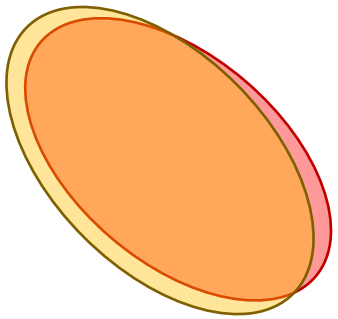
$$\mathbb{P}_D \approx \mathbb{P}_C$$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)

CLEVR:

5 Spheres



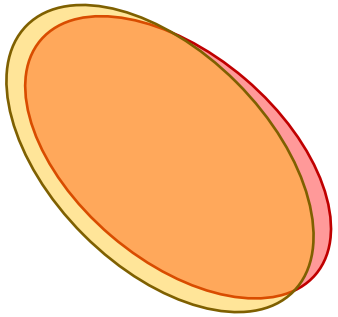
$\mathbb{P}_D \approx \mathbb{P}_C$

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

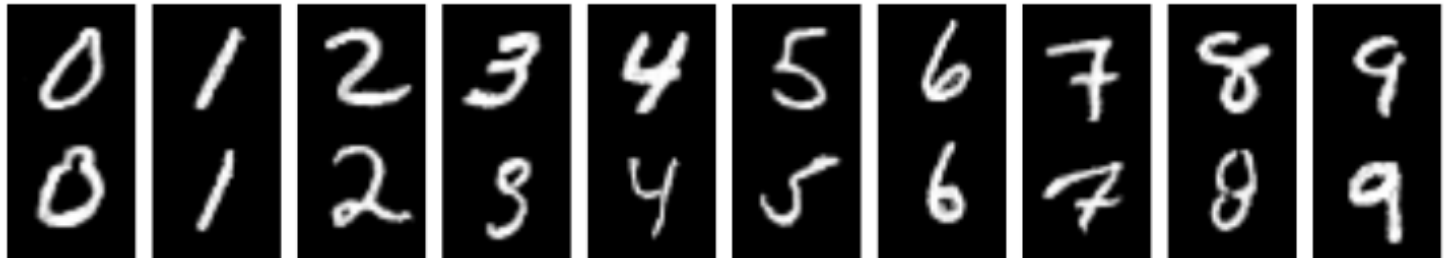
Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)

CLEVR:

5 Spheres



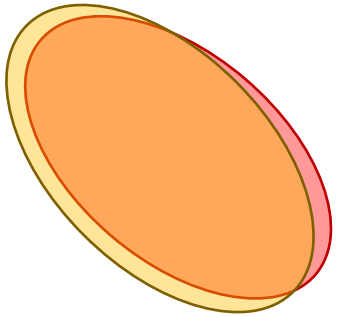
$\mathbb{P}_D \approx \mathbb{P}_C$



MNIST

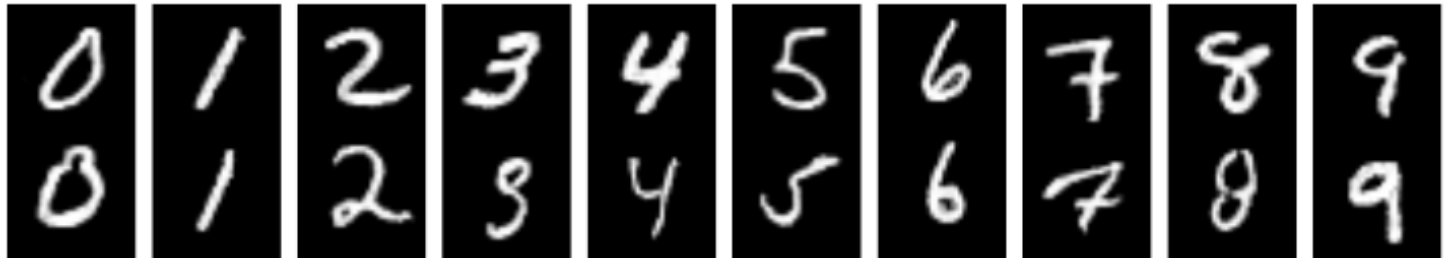
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)



$\mathbb{P}_D \approx \mathbb{P}_C$

CLEVR:
5 Spheres

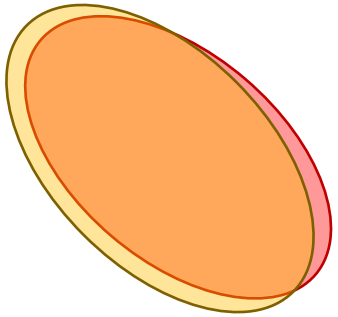


MNIST

Fashion-
MNIST

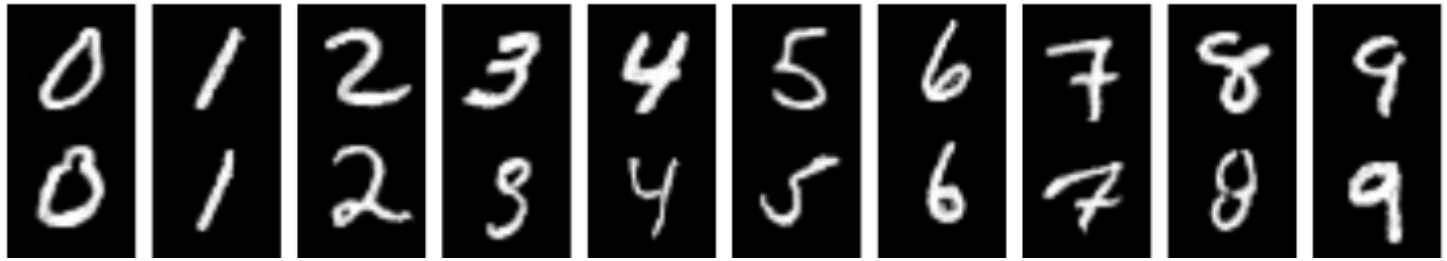
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Smoke Test: High Confidence Examples (p_i=1, $p_{\neg i}$=0)



$\mathbb{P}_D \approx \mathbb{P}_C$

≈0.943 CLEVR: 5 Spheres

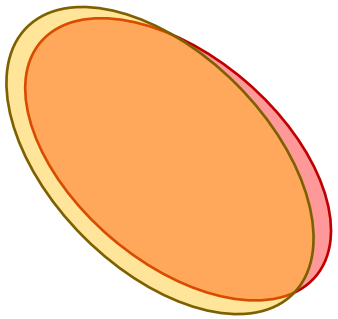


≈0.999 MNIST

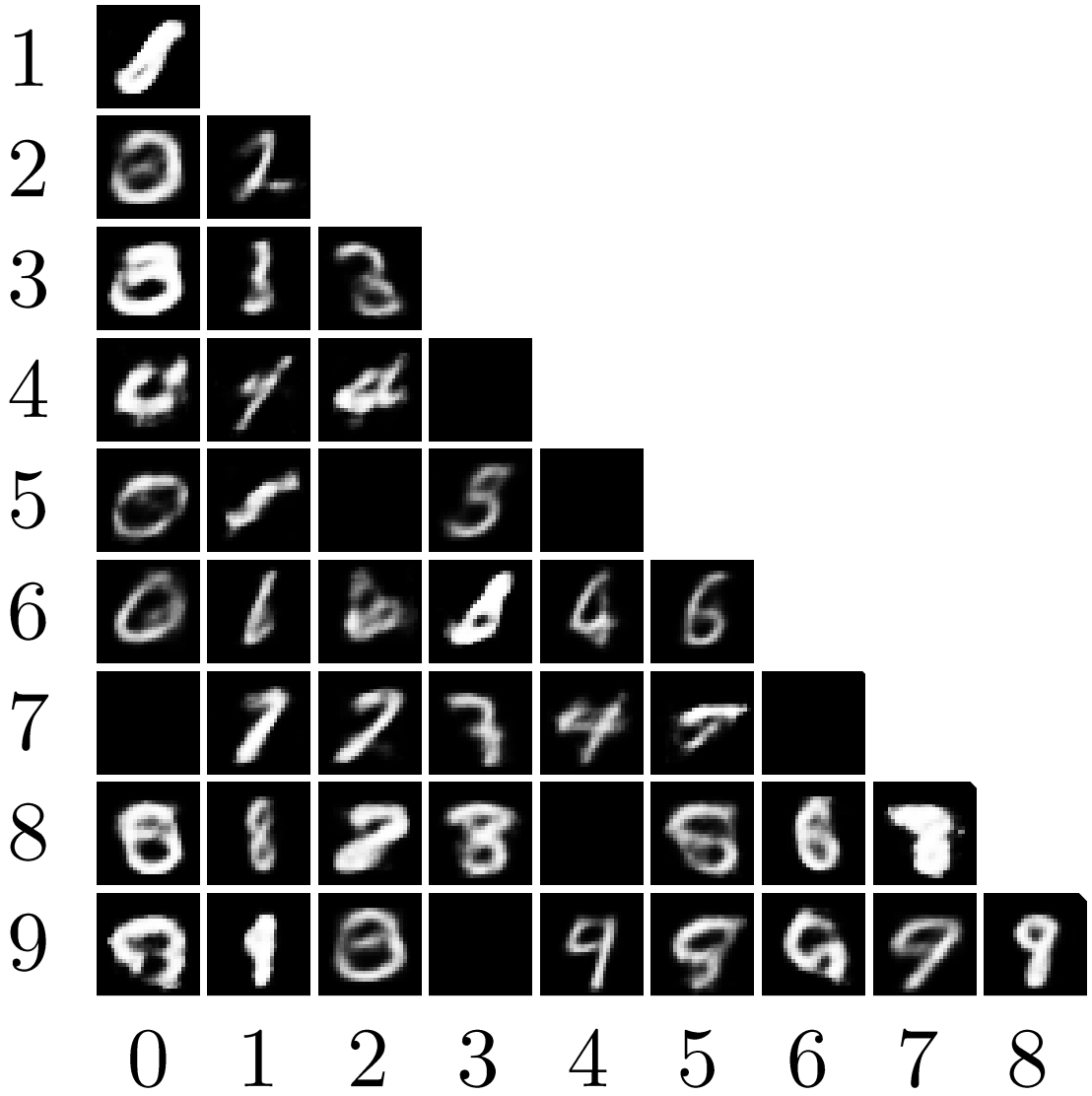
≈0.998 Fashion-MNIST

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Class Boundaries (p_i=0.5, p_j=0.5, $p_{\neg i,\neg j}$=0)

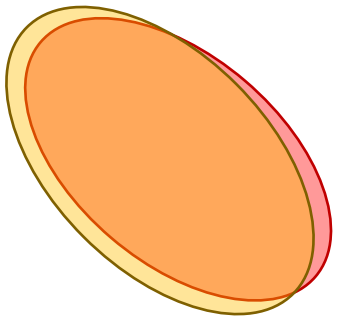


$\mathbb{P}_D \approx \mathbb{P}_C$

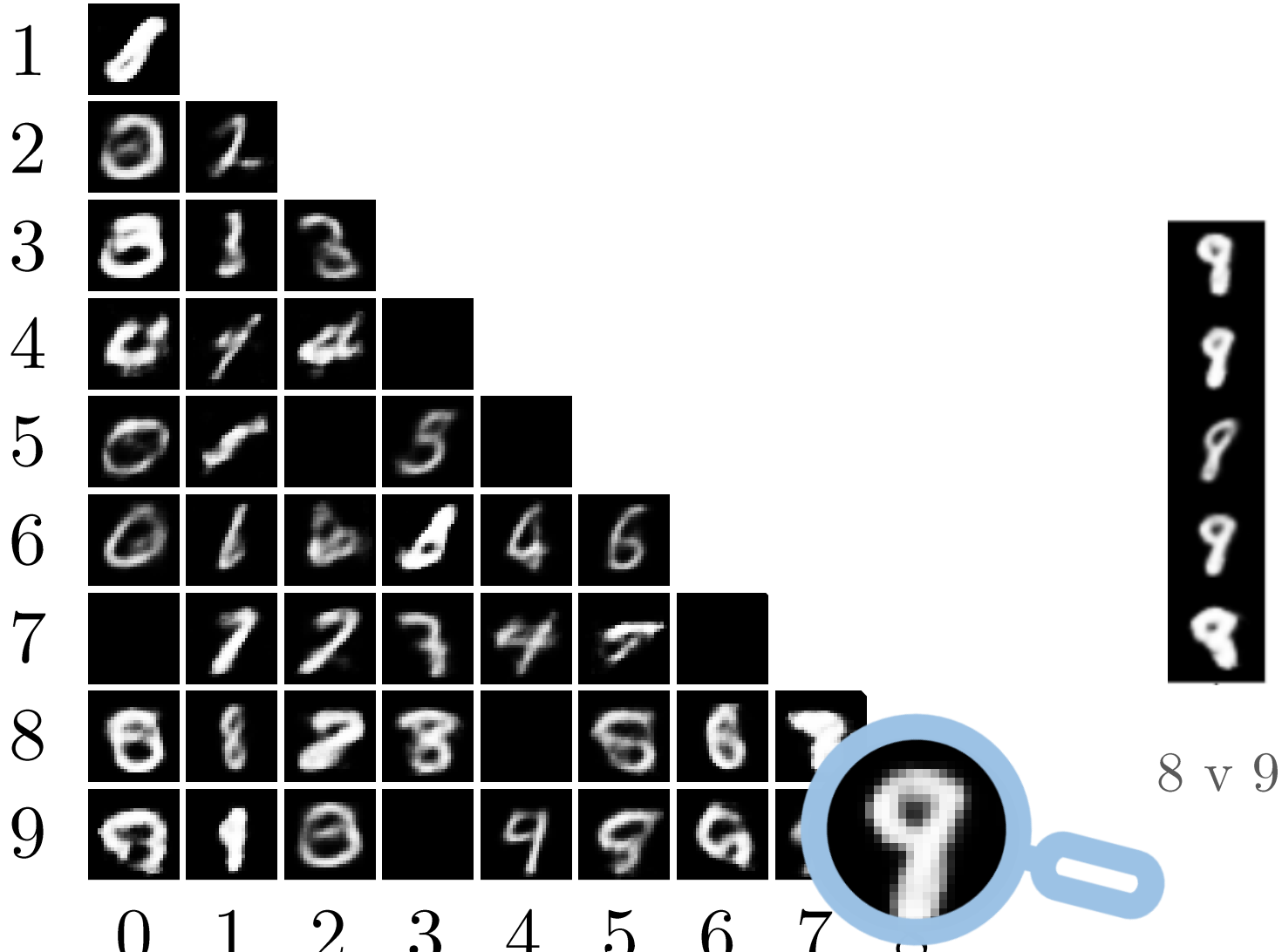


Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Class Boundaries (p_i=0.5, p_j=0.5, $p_{\neg i, \neg j}$=0)



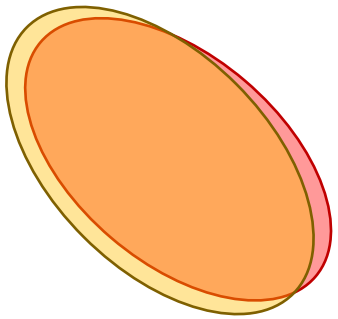
$\mathbb{P}_D \approx \mathbb{P}_C$



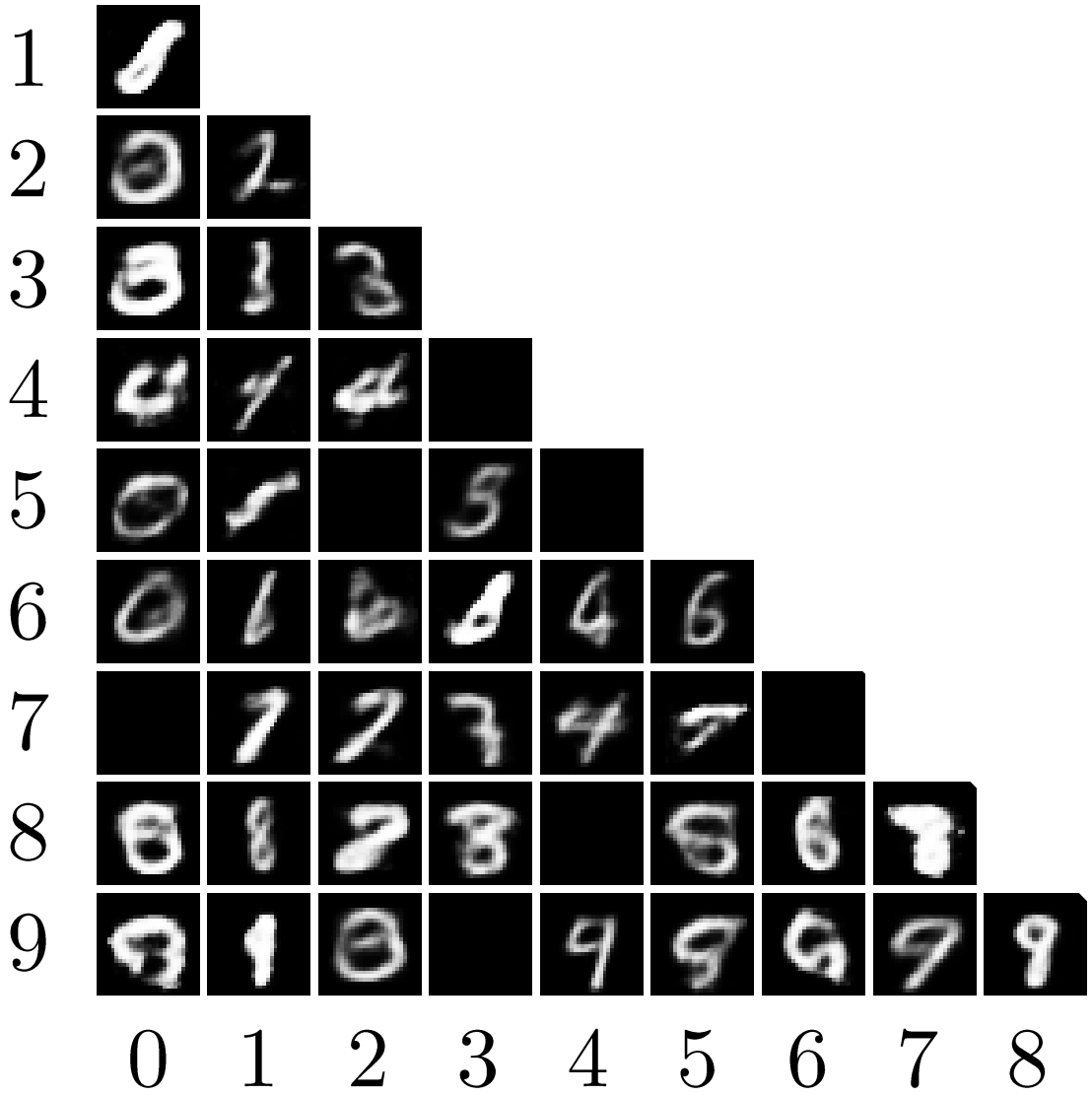
8 v 9

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Class Boundaries (p_i=0.5, p_j=0.5, $p_{\neg i,\neg j}$=0)

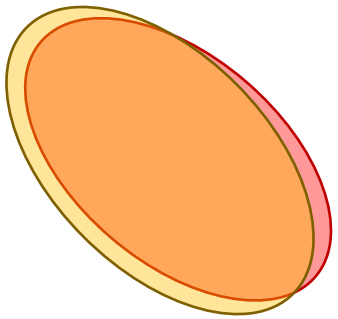


$\mathbb{P}_D \approx \mathbb{P}_C$

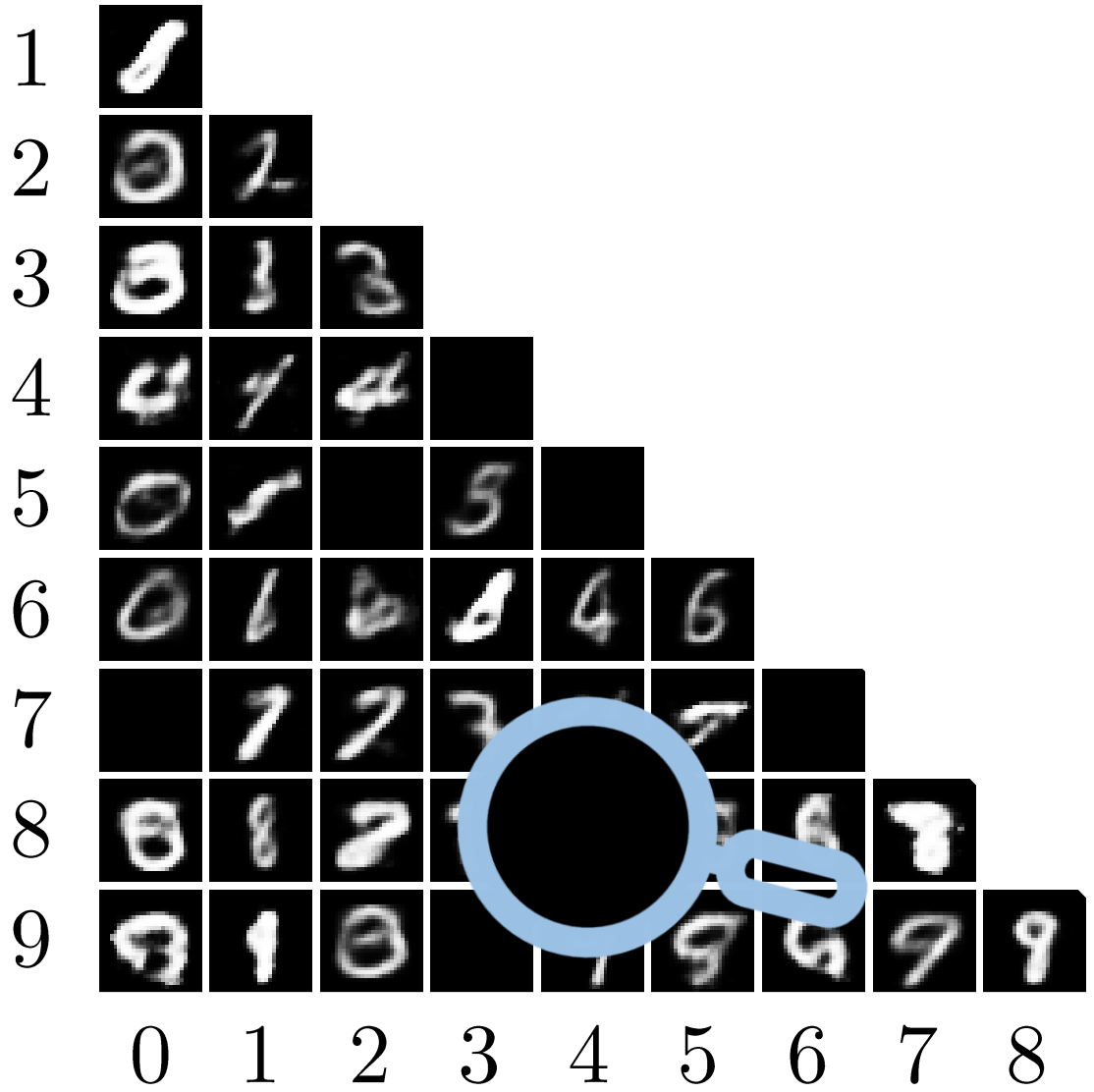


Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Class Boundaries (p_i=0.5, p_j=0.5, $p_{\neg i,\neg j}$=0)

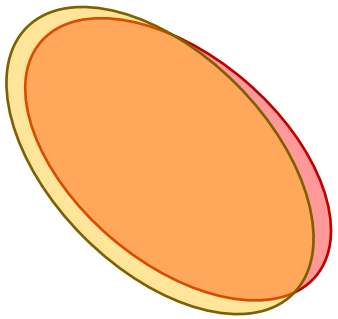


$\mathbb{P}_D \approx \mathbb{P}_C$

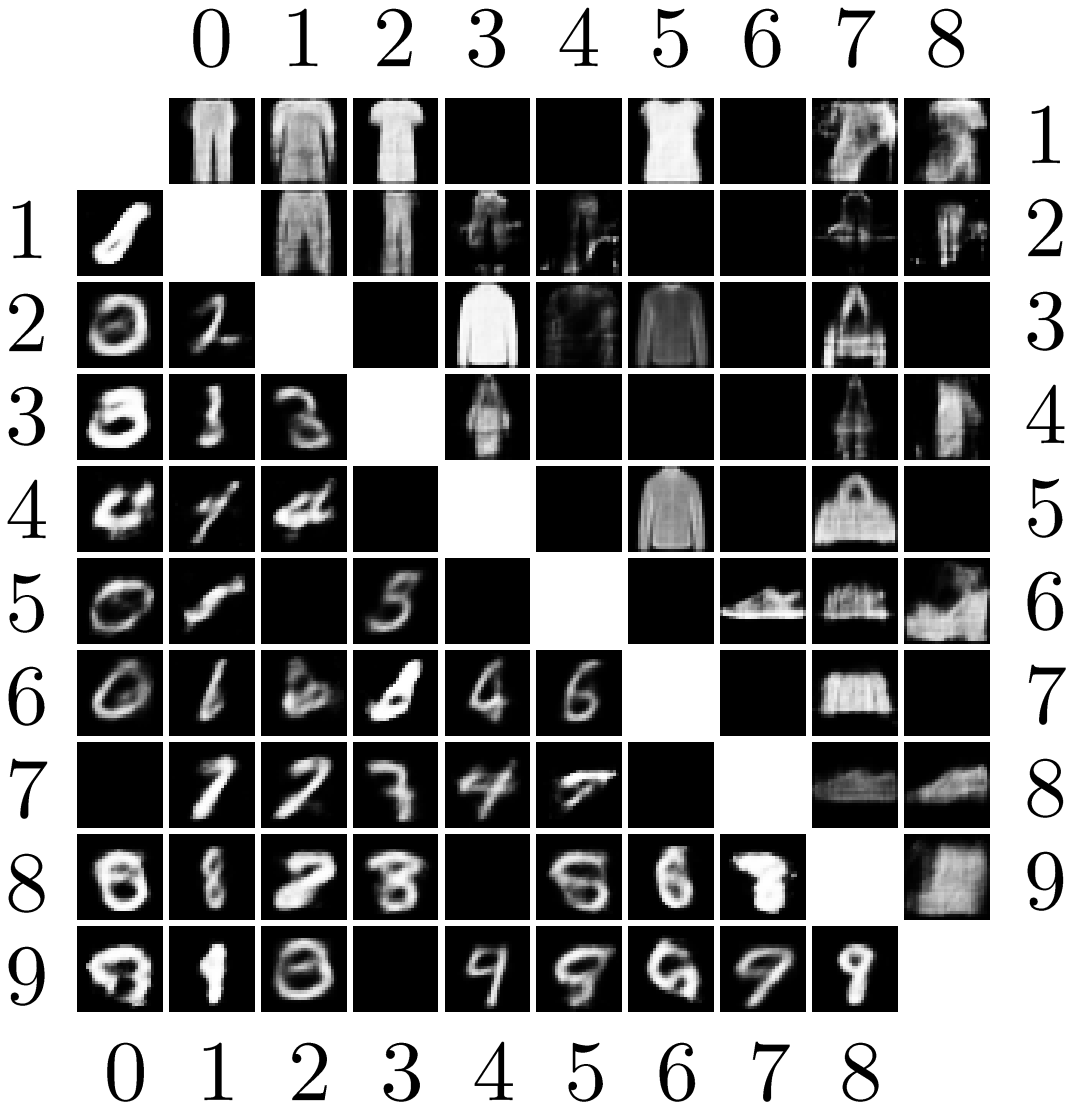


Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Class Boundaries (p_i=0.5, p_j=0.5, $p_{\neg i,\neg j}$=0)

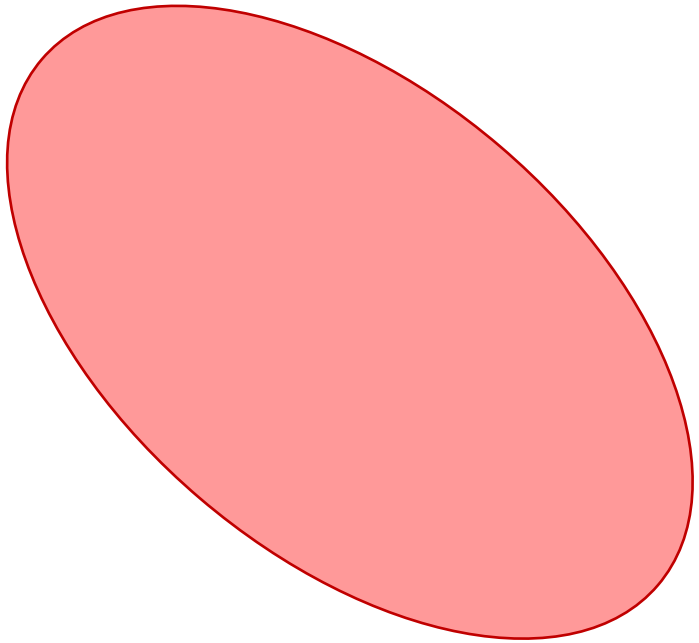


$\mathbb{P}_D \approx \mathbb{P}_C$



Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

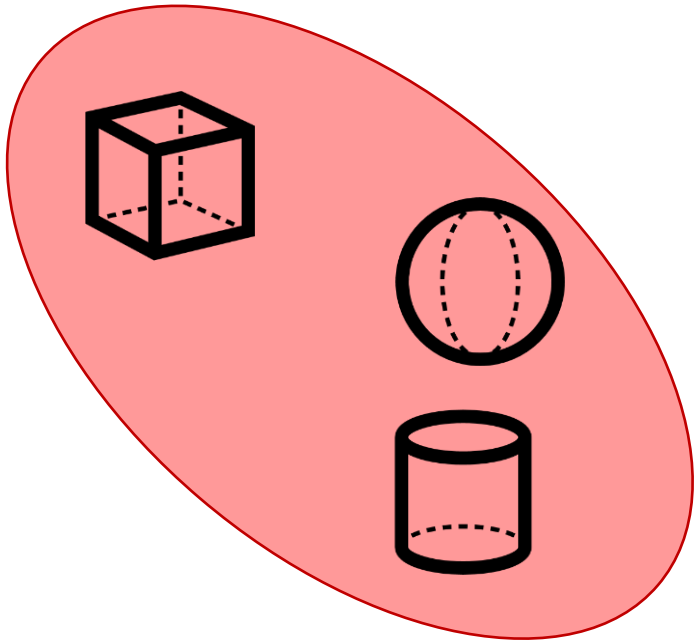
High Confidence Failures



$\mathbb{P}_C$, classifier training distribution, red

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

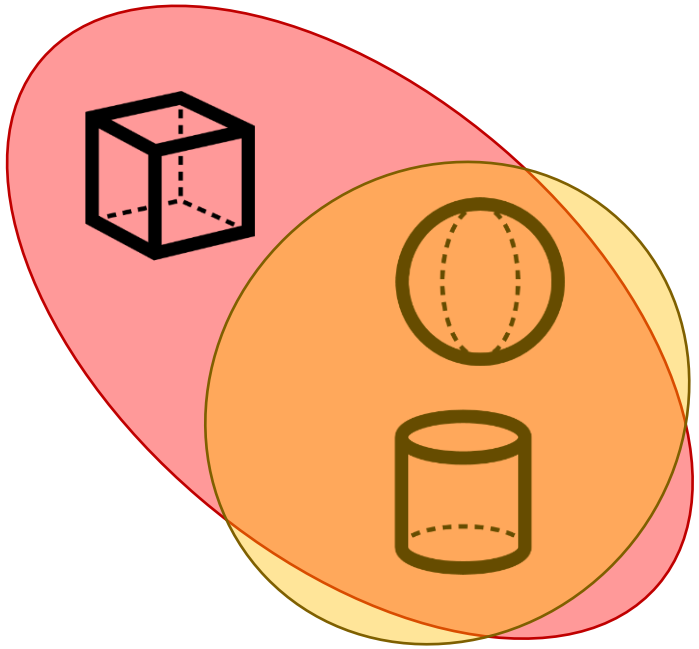
High Confidence Failures



$\mathbb{P}_C$, classifier training distribution, red

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

High Confidence Failures

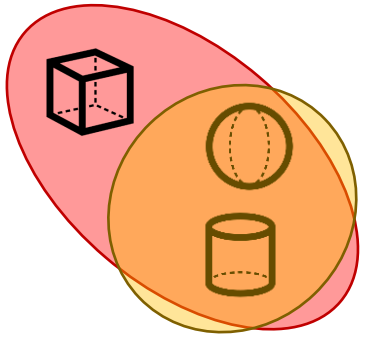


$\mathbb{P}_D \subsetneq \mathbb{P}_C$

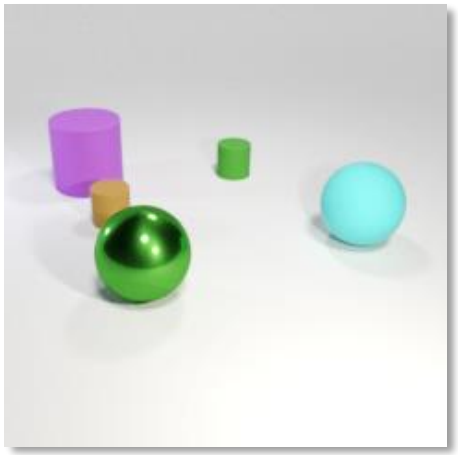
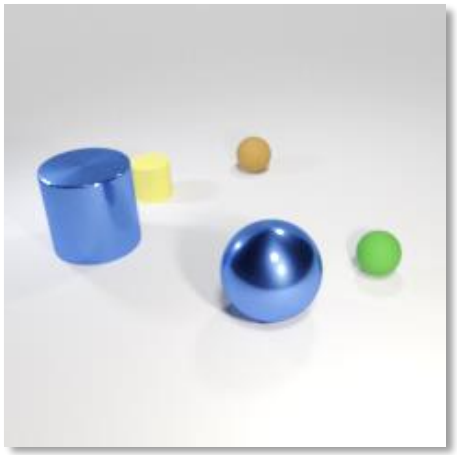
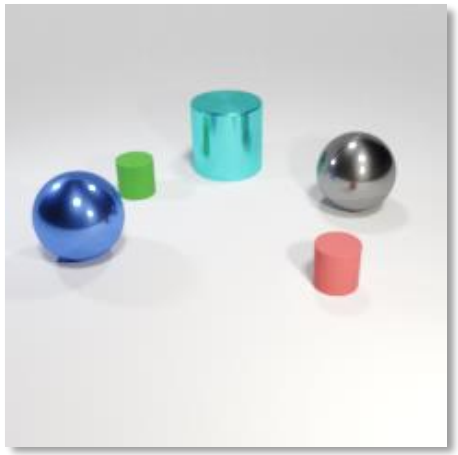
$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021



High Confidence Failures

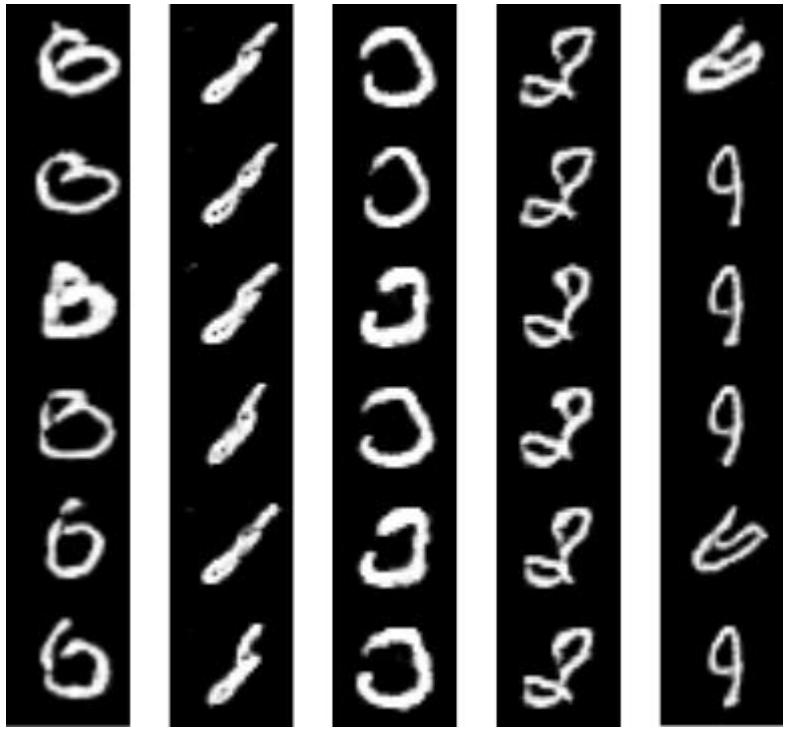




97.2% 96.0% 94.5% 67.3% 93.5%

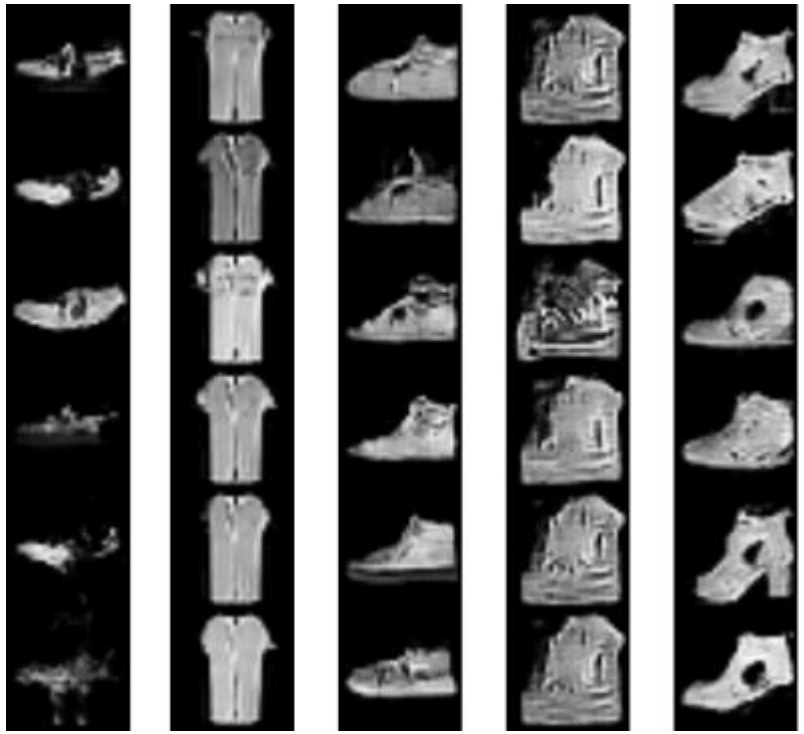
Contains 1 Cube

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

High Confidence Failures



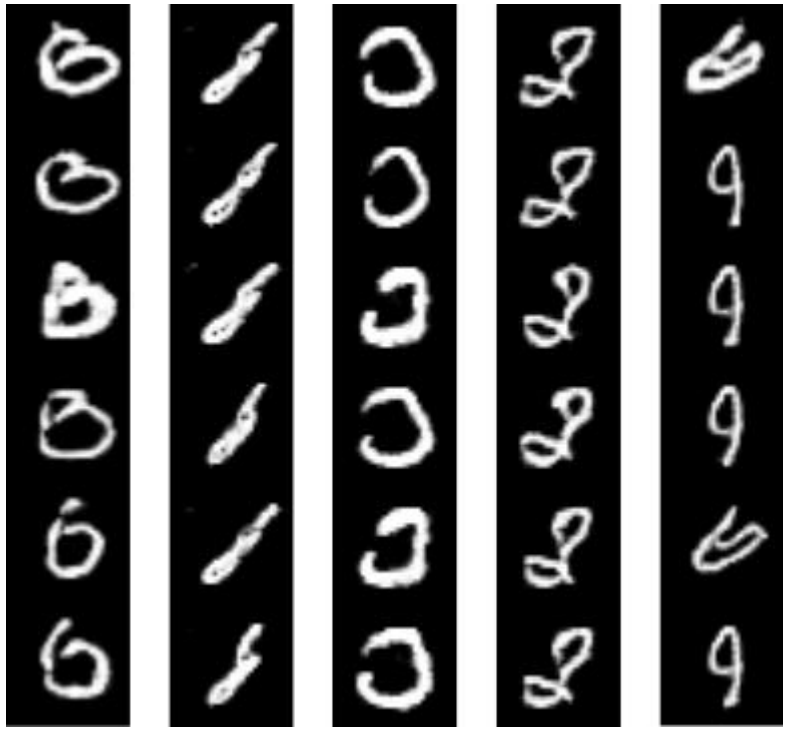
0: 0.981 ± 0.027
1: 0.953 ± 0.028
2: 0.968 ± 0.028
3: 0.969 ± 0.027
4: 0.955 ± 0.030



Sandal: 0.986 ± 0.030
Shirt: 0.938 ± 0.032
Sneaker: 0.969 ± 0.028
Bag: 0.967 ± 0.026
Ankle boot: 0.971 ± 0.027

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

High Confidence Failures

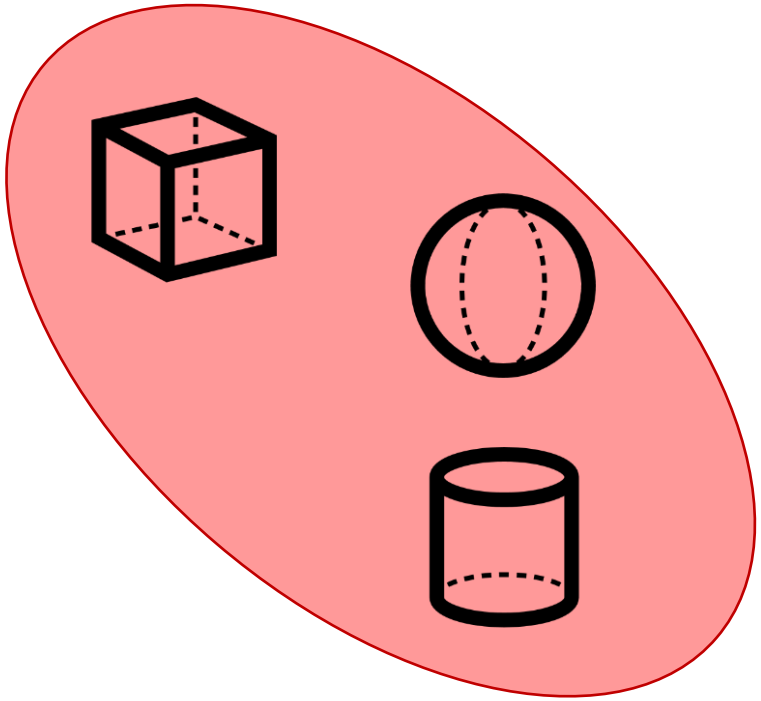


0: 0.981 ± 0.027
1: 0.953 ± 0.028
2: 0.968 ± 0.028
3: 0.969 ± 0.027
4: 0.955 ± 0.030

Sandal: 0.986 ± 0.030
Shirt: 0.938 ± 0.032
Sneaker: 0.969 ± 0.028
Bag: 0.967 ± 0.026
Ankle boot: 0.971 ± 0.027

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

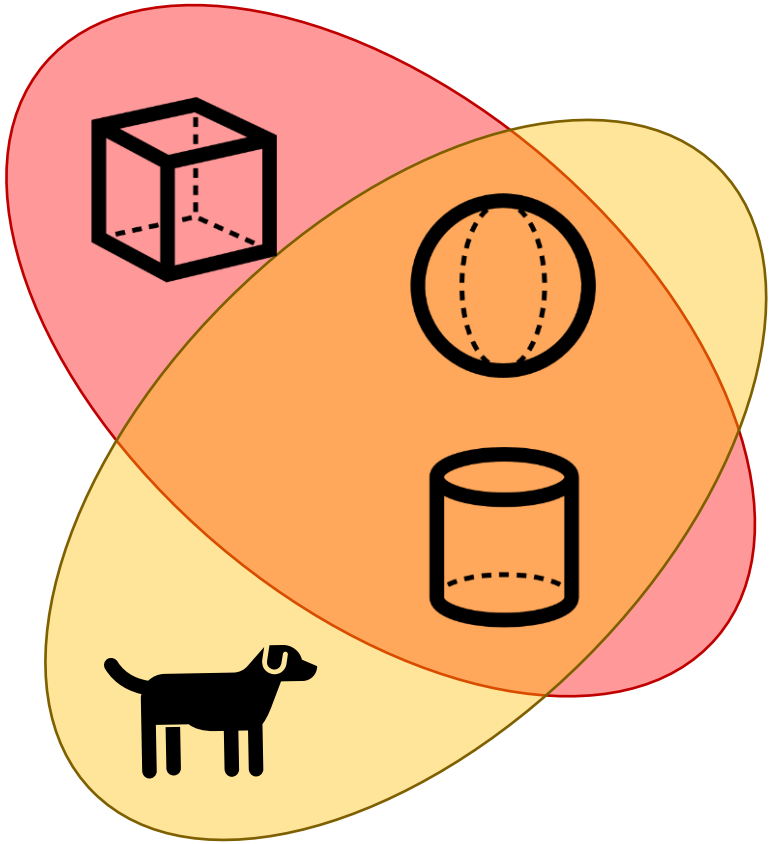
Novel Class Extrapolation



$\mathbb{P}_C$, classifier training distribution, red

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Novel Class Extrapolation

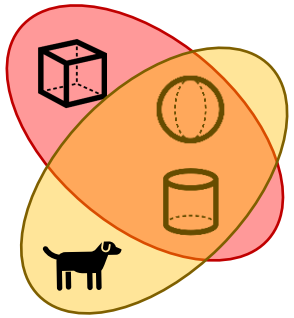


$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

$$\mathbb{P}_D \cap \mathbb{P}_C \neq \mathbb{P}_D \neq \mathbb{P}_C$$

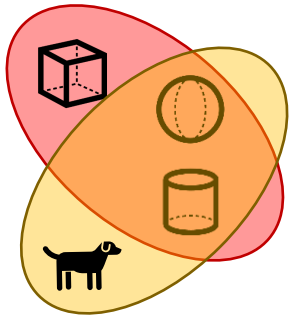
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021



Novel Class Extrapolation

Contains 5 Cubes: 89.3%

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021



Novel Class Extrapolation

89.3%

81.2%

83.5%

90.4%

90.5%

Contains 5 Cubes

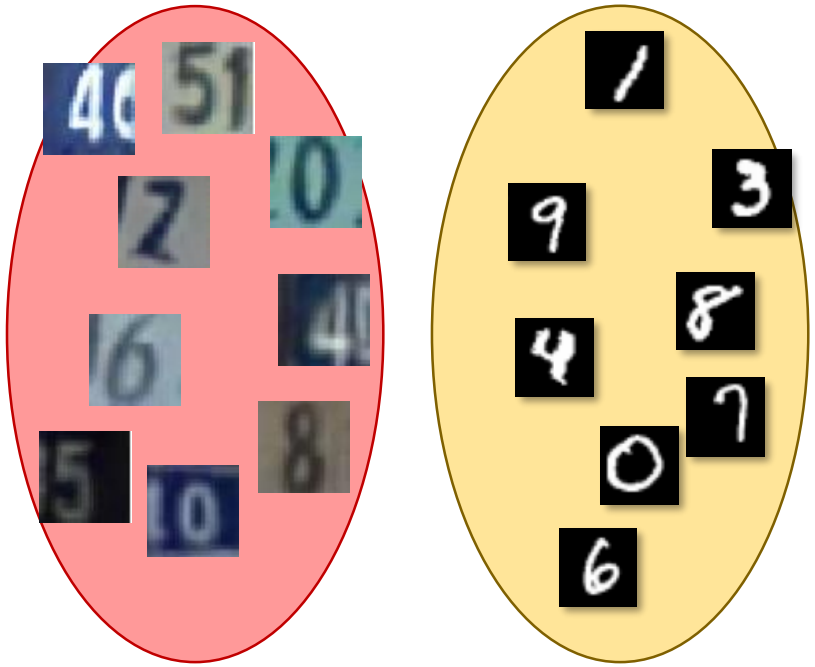
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Domain Adaptation

$\mathbb{P}_C$, classifier training distribution, red

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Domain Adaptation



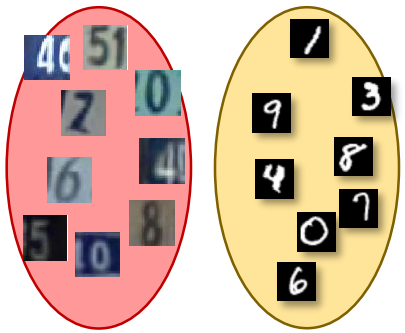
$\mathbb{P}_D \cap \mathbb{P}_C = \varnothing$

$\mathbb{P}_C$, classifier training distribution, red

$\mathbb{P}_D$, Bayes-TrEx's data distribution, yellow

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Domain Adaptation: ADDA

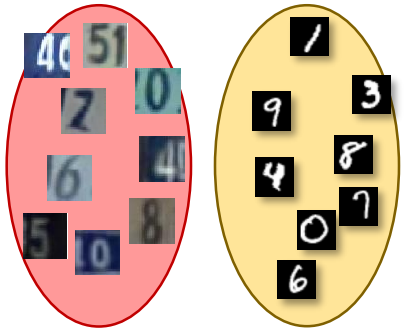


Overall:

- Baseline: 61% accuracy
- ADDA: 71% accuracy

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Domain Adaptation: ADDA

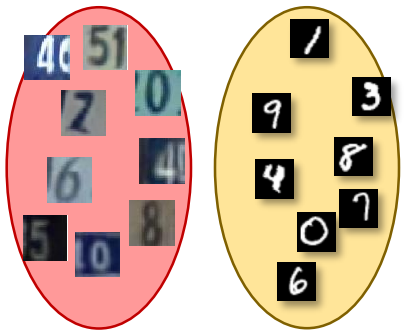


Overall:

- Baseline: 61% accuracy
- **ADDA: 71% accuracy**

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Domain Adaptation: ADDA



Overall:

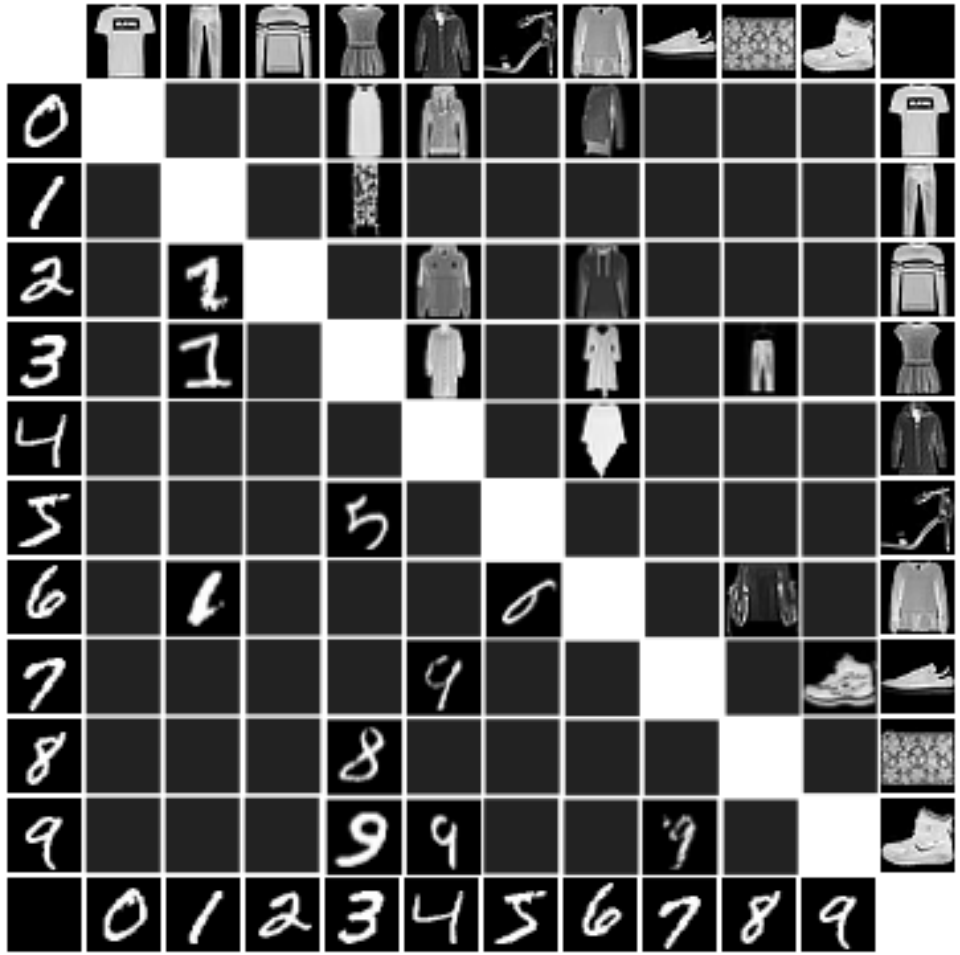
- Baseline: 61% accuracy
- **ADDA: 71% accuracy**

High Confidence Examples:

- **Baseline: 80% accuracy**
- ADDA: 72% accuracy

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How does Bayes-TrEx compare to the test set?

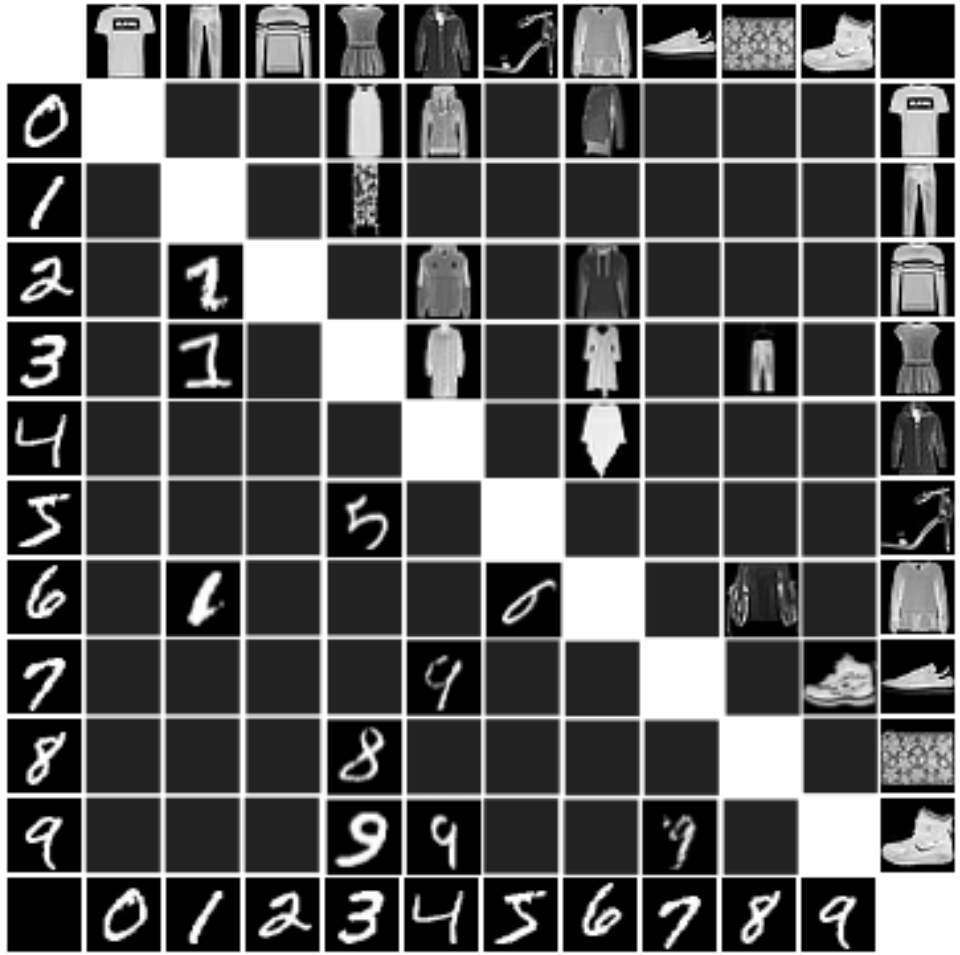


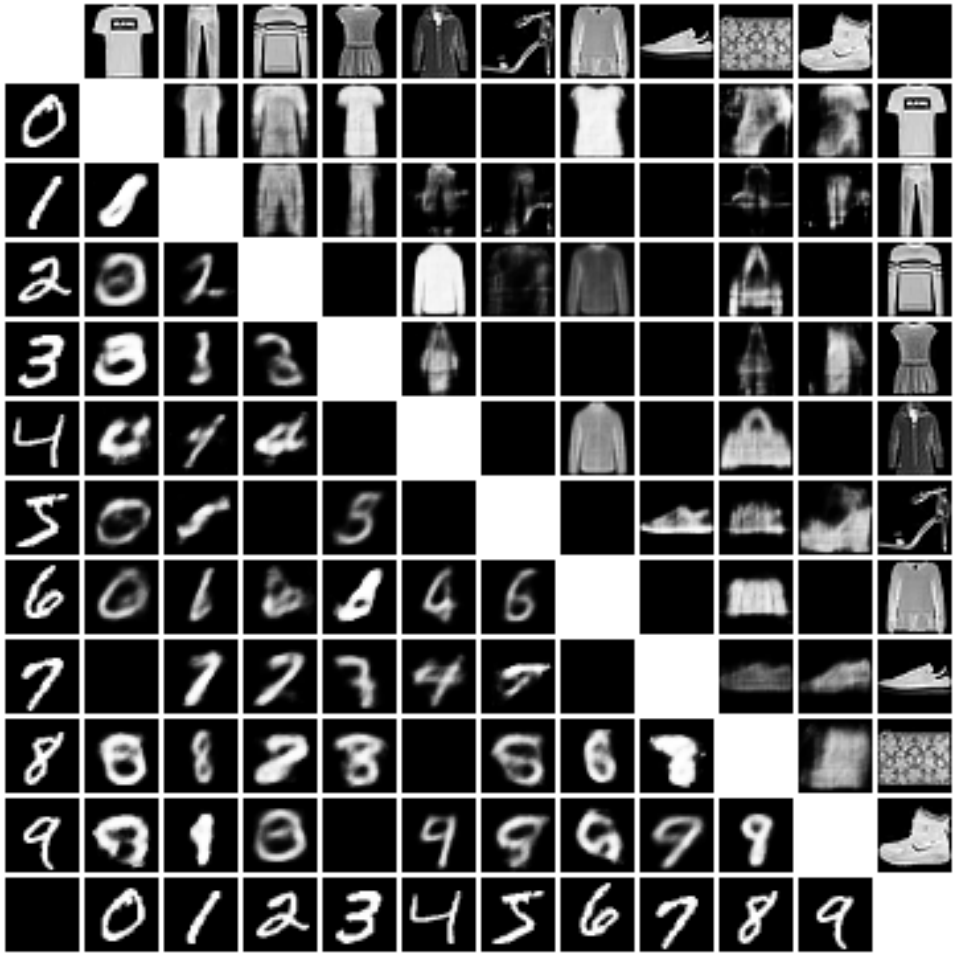
Test set:

22 Ambiguous Pairings

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How does Bayes-TrEx compare to the test set?





Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How does Bayes-TrEx compare to the test set?

Test set exposes more *mislabelings* than true classification failures.

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled

Misclassified

Test set exposes more *mislabelings* than true classification failures.

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled

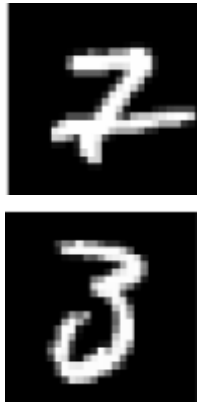
Misclassified

Test set exposes more *mislabelings* than true classification failures.

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled



Misclassified

Test set exposes more *mislabelings* than true classification failures.

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled

Misclassified

Test set exposes more *mislabelings* than true classification failures.

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled	Misclassified
	$\varnothing$

Test set exposes more *mislabelings* than true classification failures.

How does Bayes-TrEx compare to the test set?

Class: 2

Mislabeled	Misclassified
	$\varnothing$

Test set exposes more *mislabelings* than true classification failures.

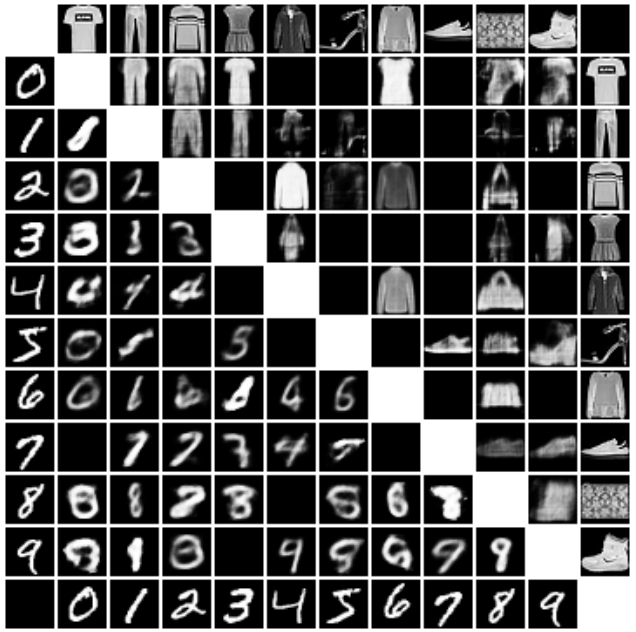
60/84 MNIST

42/93 Fashion-MNIST

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Bayes-TrEx

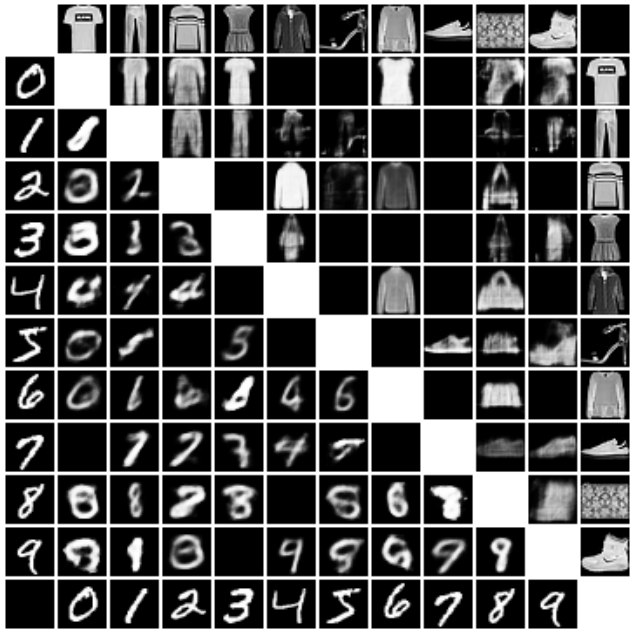
Class Boundaries



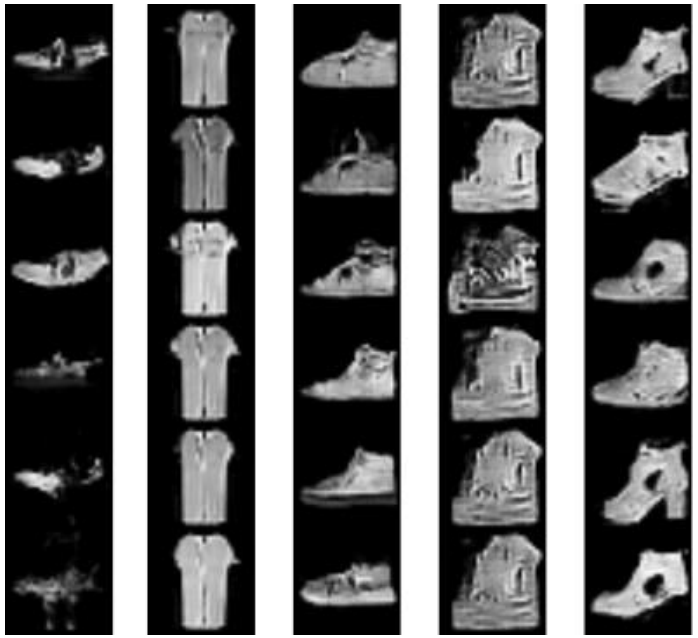
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Bayes-TrEx

Class Boundaries



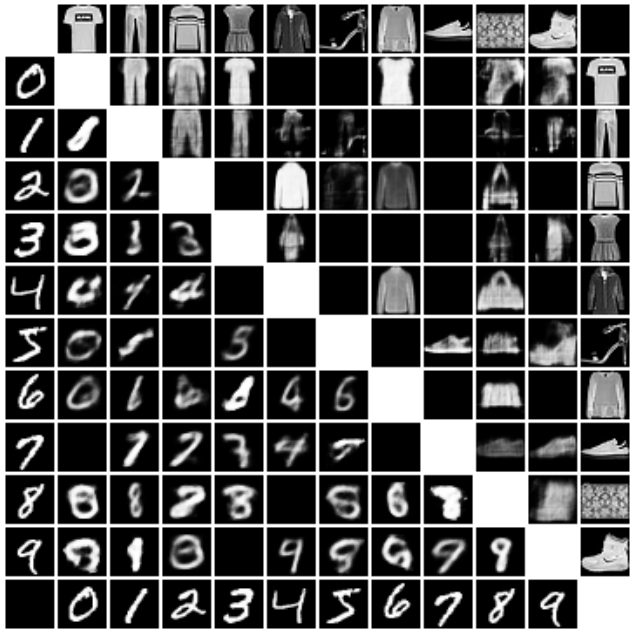
High Confidence Failures



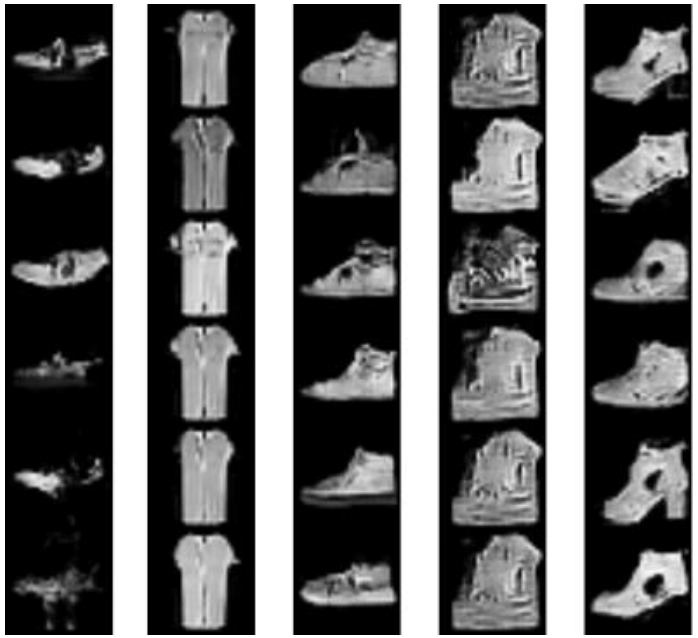
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Bayes-TrEx

Class Boundaries



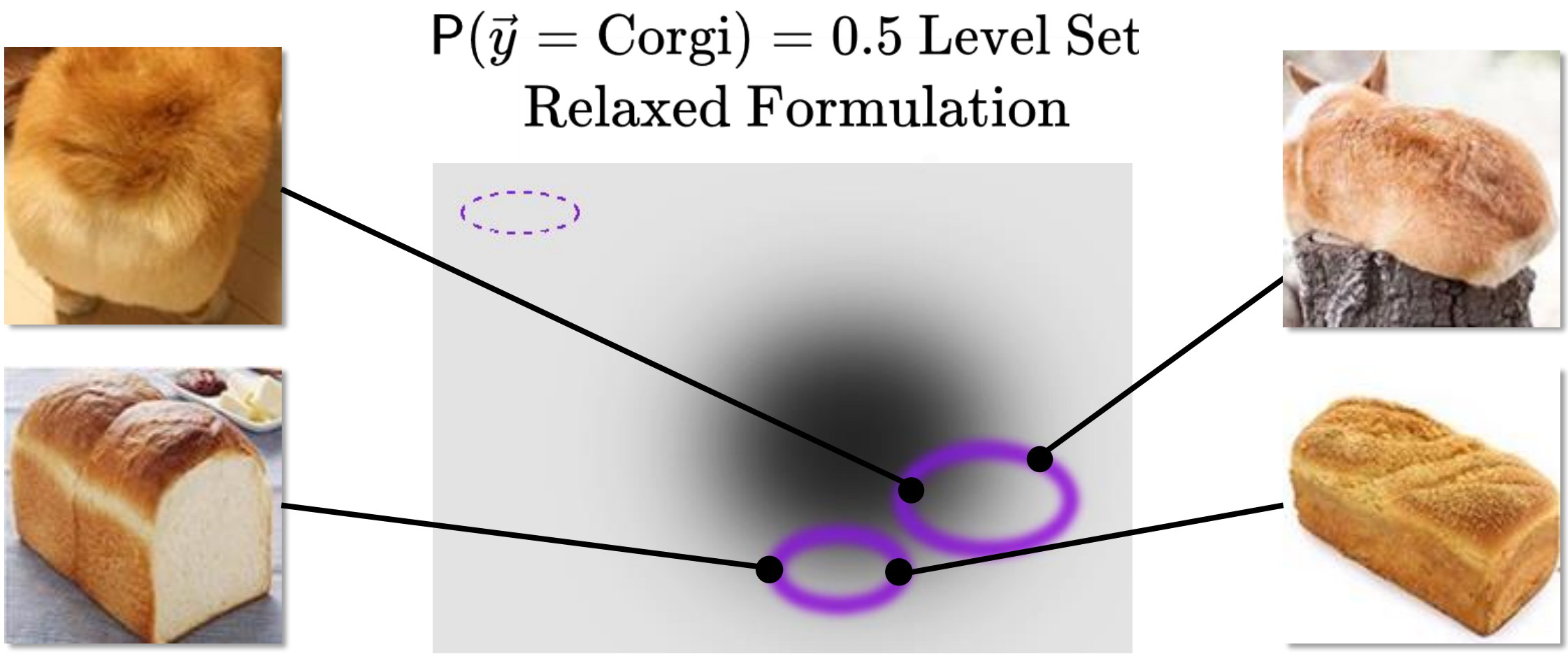
High Confidence Failures



Novel Classes

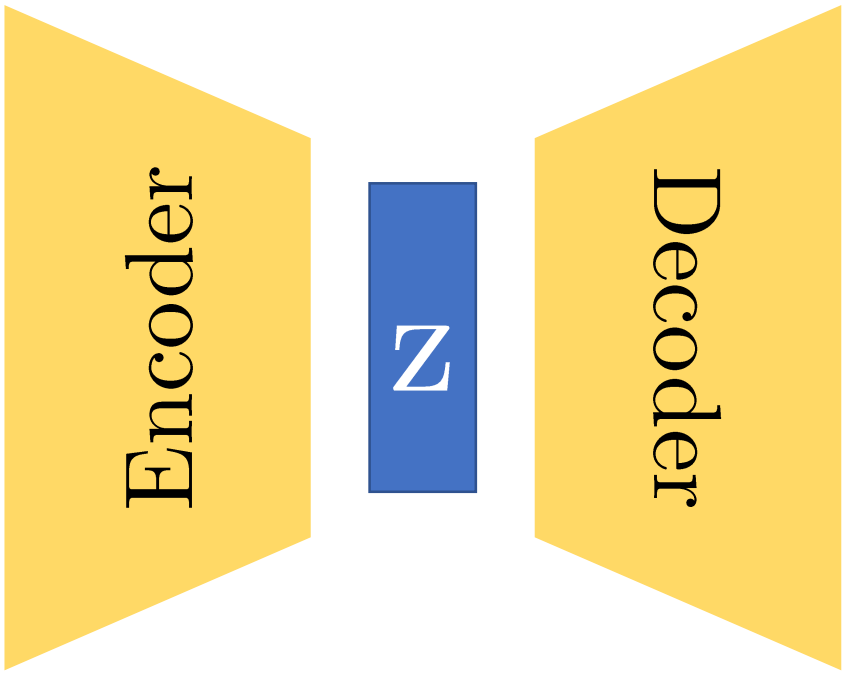
Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Limitation 1: Trends & Coverage Measures



Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

Limitation 2: Latent Space Dimensionality



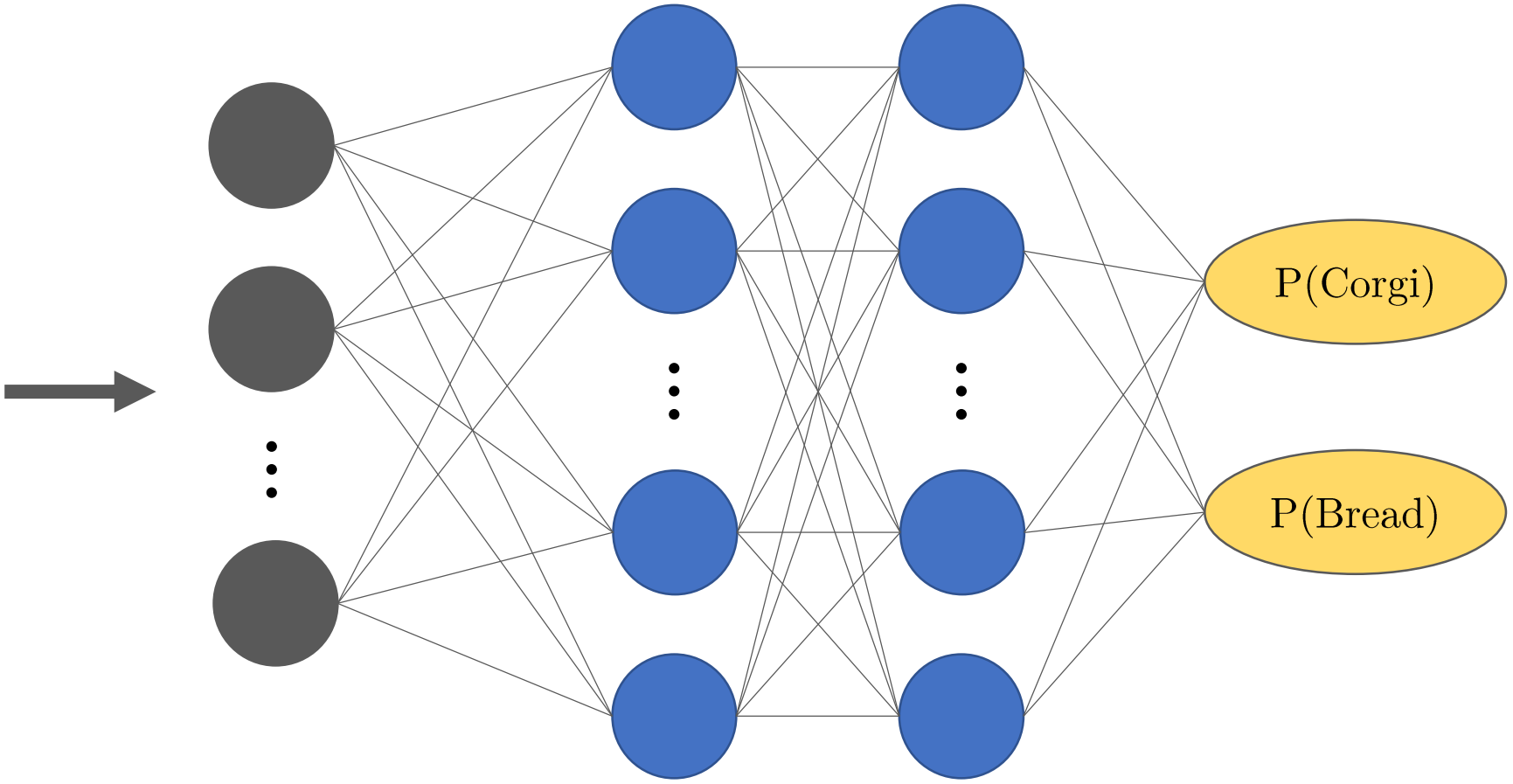
Latent: 10-dim

```
Object 1: {
        shape: {cube, sphere, cylinder}
        color: {cyan, …, green},
        material: {rubber, metal},
        x-coord: [0,1],
        y-coord: [0,1],
        r-coord: [0, 2pi)
},
…,
Object 5: {
        shape: {cube, sphere, cylinder}
        color: {cyan, …, green},
        material: {rubber, metal},
        x-coord: [0,1],
        y-coord: [0,1],
        r-coord: [0, 2pi)
}
```

Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

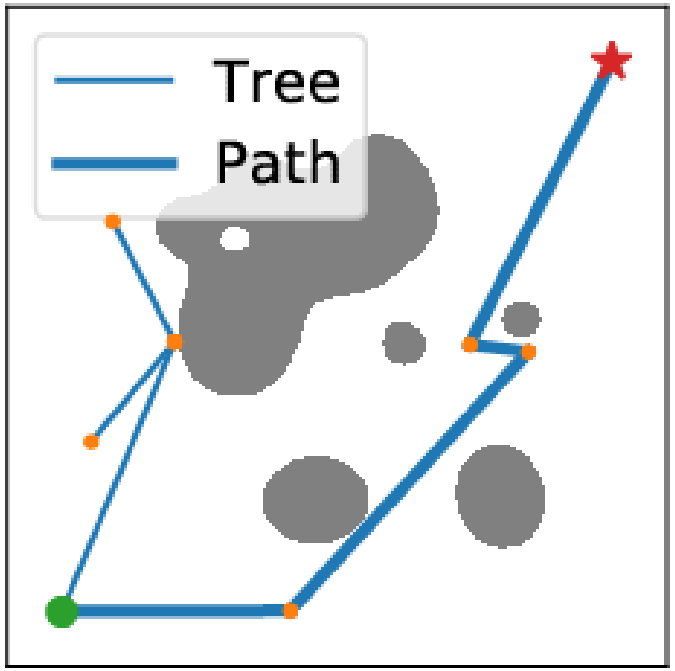
Limitation 3: Transparency for Classification Tasks

Input Image

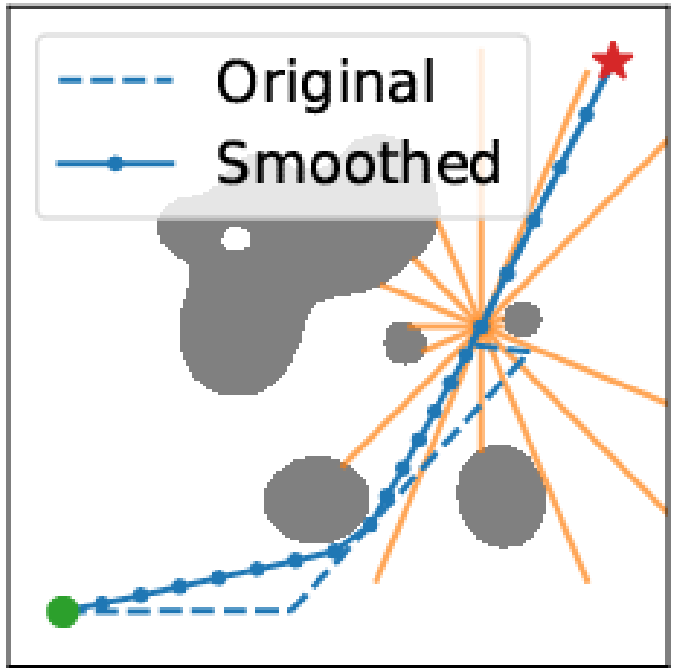


Booth and Zhou et al., Bayes-TrEx, AAAI, 2021

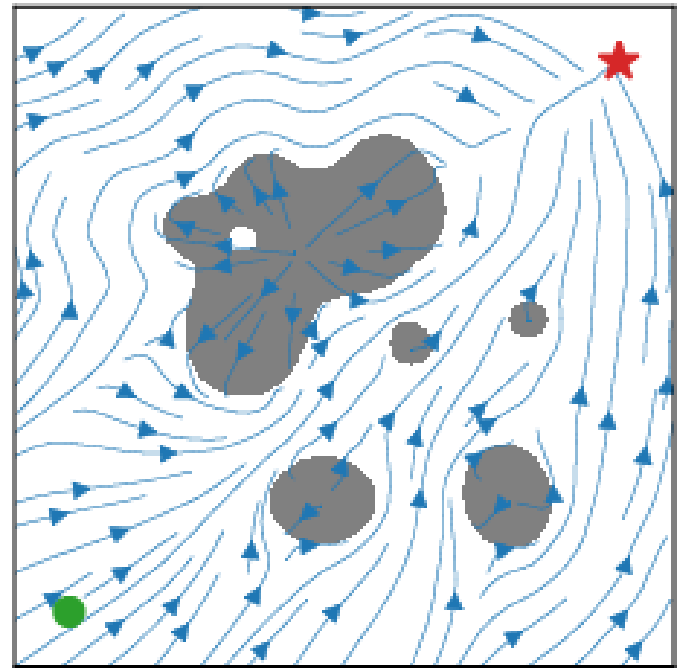
~~Limitation 3: Transparency for Classification Tasks~~



RRT



Smoothing and Lidar



DS Modulation

In Review & On ArXiv (2012.13615):

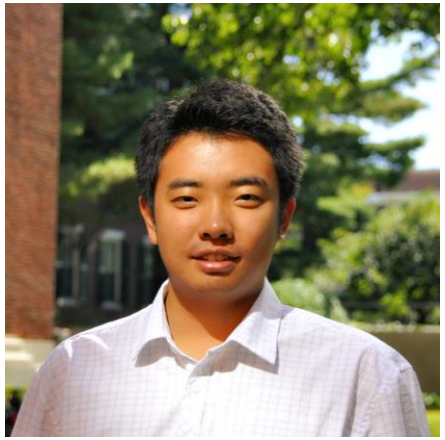
Zhou et al., *RoCUS: Robot Controller Understanding via Sampling*

Bayes-TrEx

Paper, code, & many more experiments: github.com/serenabooth/bayes-trex

Serena Booth*

@SerenaLBooth



Yilun Zhou*

yilun@mit.edu

Ankit Shah

@ankitjs

Julie Shah

@julie_a_shah

Follow up (robotics): arxiv.org/abs/2012.13615