ML Interpretability: Beyond Feature Importance

Javier Antorán (ja666@cam.ac.uk)
PhD Student, University of Cambridge
@JaviAC7

UNIVERSITY OF CAMBRIDGE

SQRD ROBUST + RESPONSIBLE AI

Recent Trends in ML Research and what they mean for Interpretability

Javier Antorán (ja666@cam.ac.uk)
PhD Student, University of Cambridge
@JaviAC7

Talk Outline

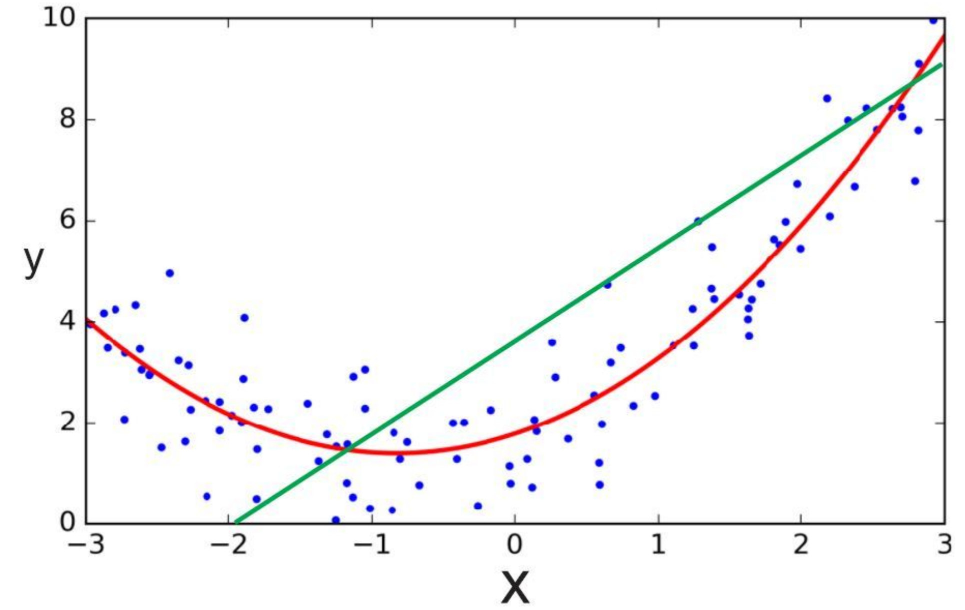
- Existing Approaches to Interpretability
- “Counterfactual” Interpretability
- Uncertainty in ML
- Transparency in ML Systems that Express Uncertainty with CLUE
- Questions / Feedback

Interpretable Data Driven Decision Making

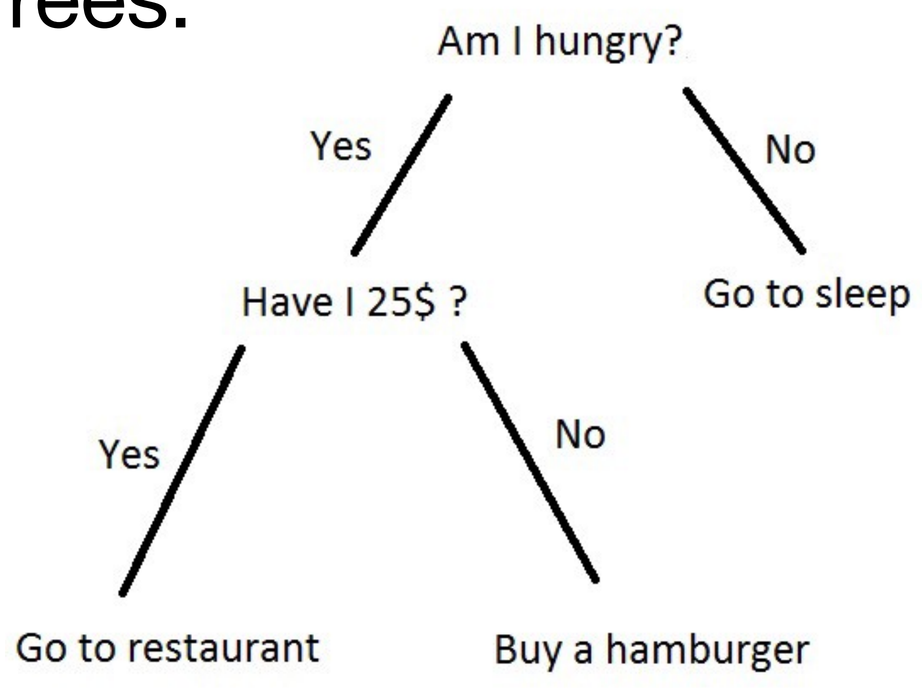
- Generalised Linear Models:

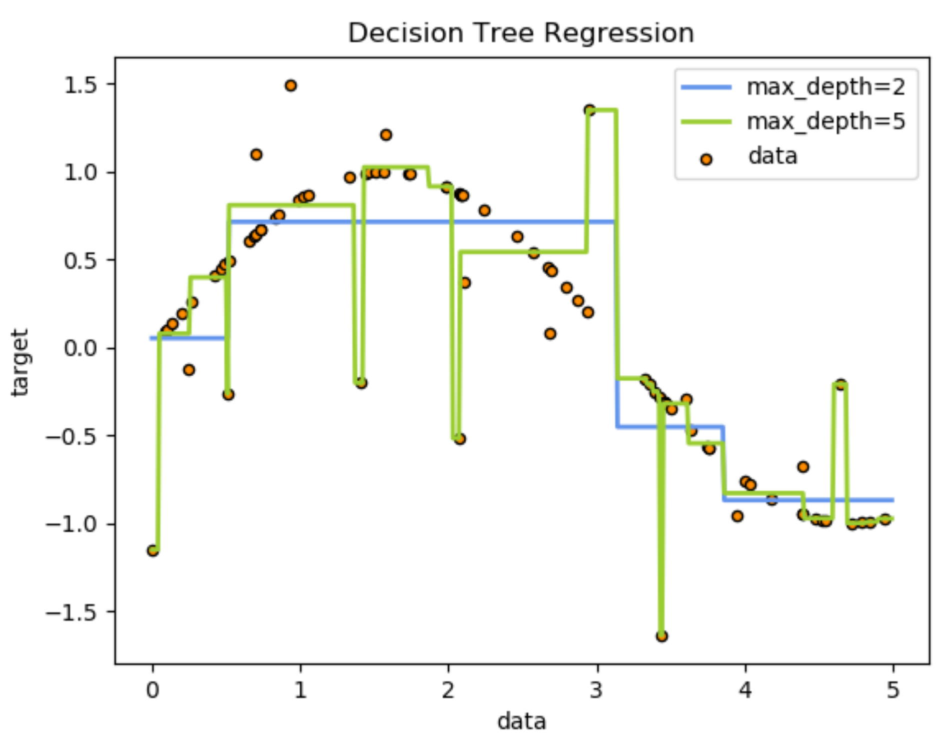
$$\mu = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots$$

  - Importance of x_i is $|w_i|$
  - Polarity of x_i is **sign**(w_i)

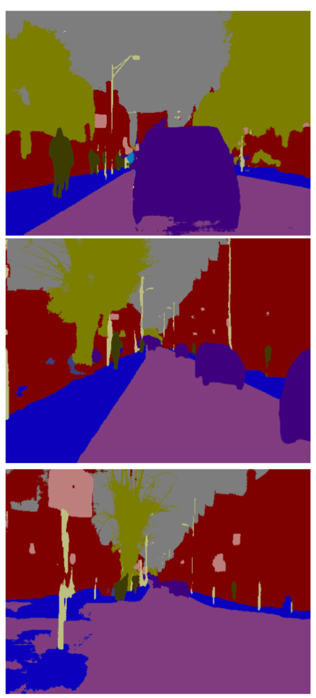


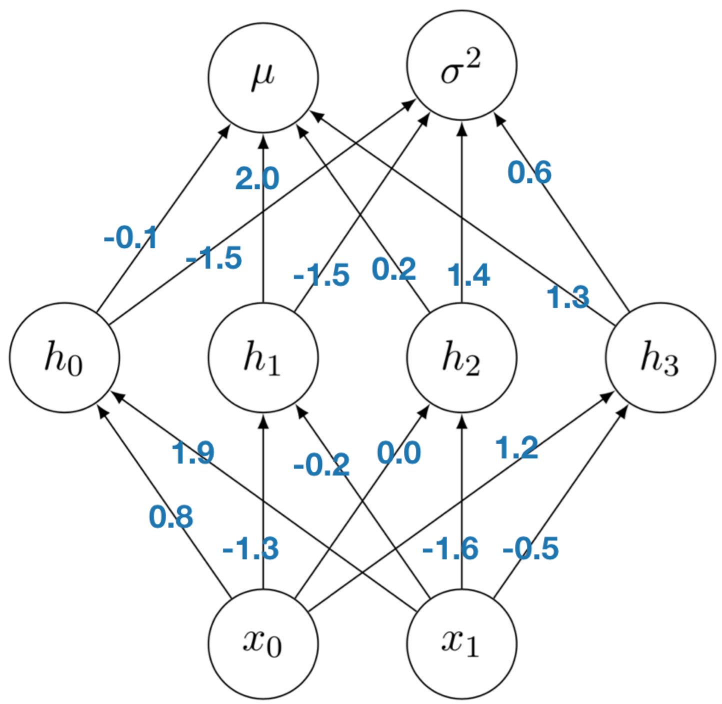
- Decision Trees:





Not Very Interpretable Data Driven Decision Making

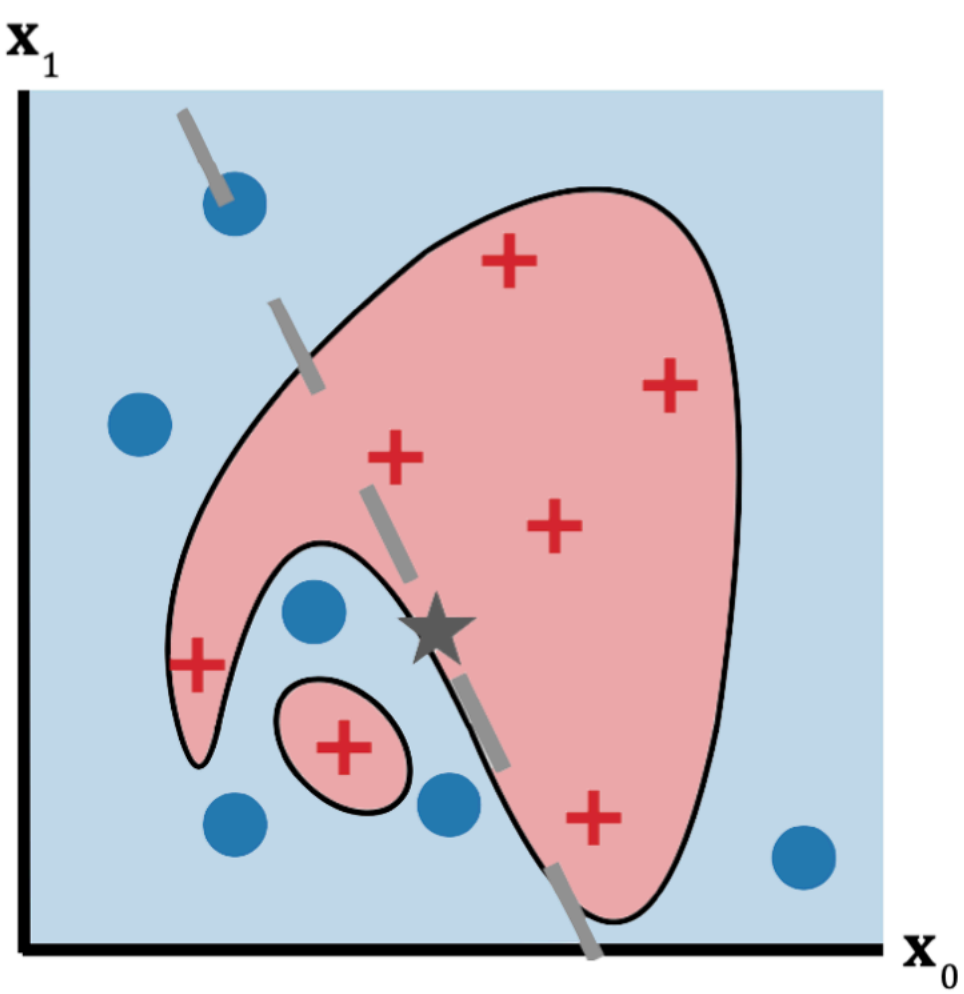




- **Capture non-linear functions (NNs are universal function approximators)**
- **Scale to high dimensional problems**
- **Scale to massive datasets**
- **Simulate complex systems**
- **Etc**

Feature Importance: LIME

- We approximate non-linear model Locally with Linear Model



Here explanation is more reliable in x_1 than x_0

$$\mu = f_{NN}(\mathbf{x})$$

$$\mu_{approx} = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_3 \ldots$$

Lime Explanation is $w_1, w_2, w_3 \ldots$

Feature Importance on Images

LIME [Ribeiro et. al., 2016] **SHAP** [Lundberg and Lee, 2017]

- **Per class weight vectors forms a “template images” of positive and negative contributions**
- **Can become meaningless for strongly non-linear functions**

Counterfactual Explanations to the Rescue!

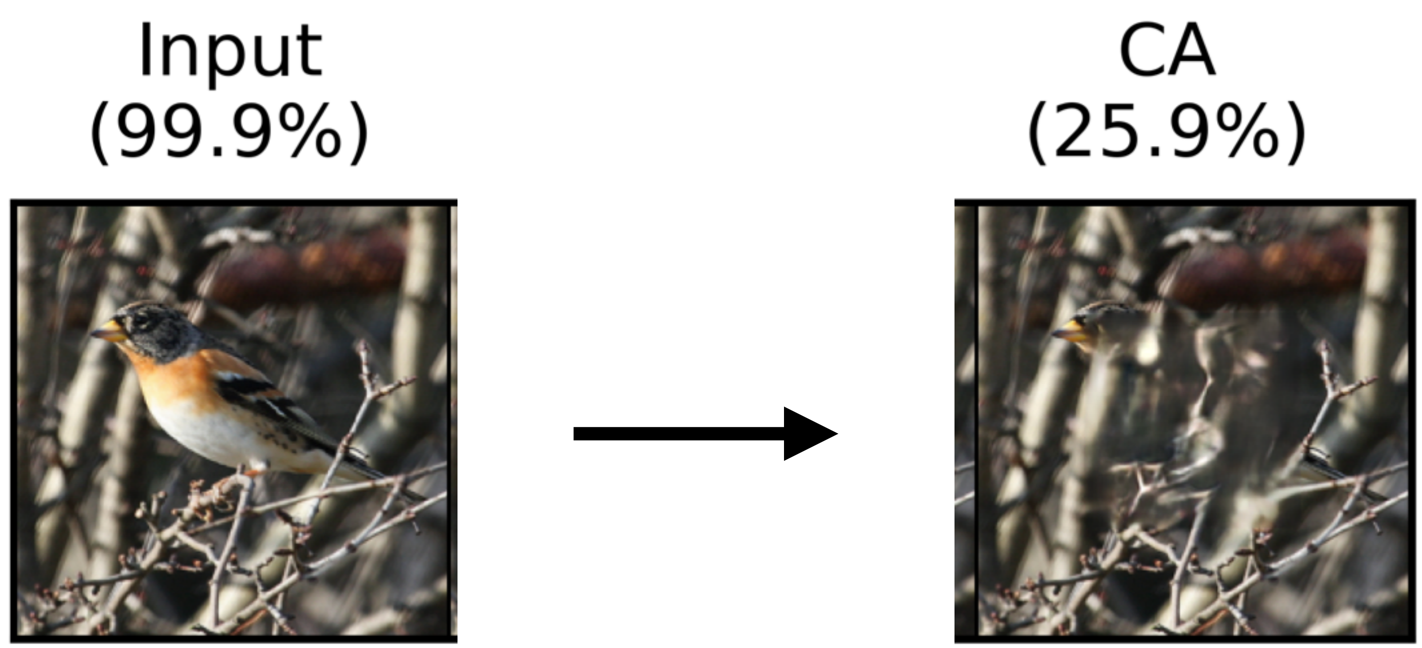
- Counterfactuals capture the notion what would have happened if something had been different

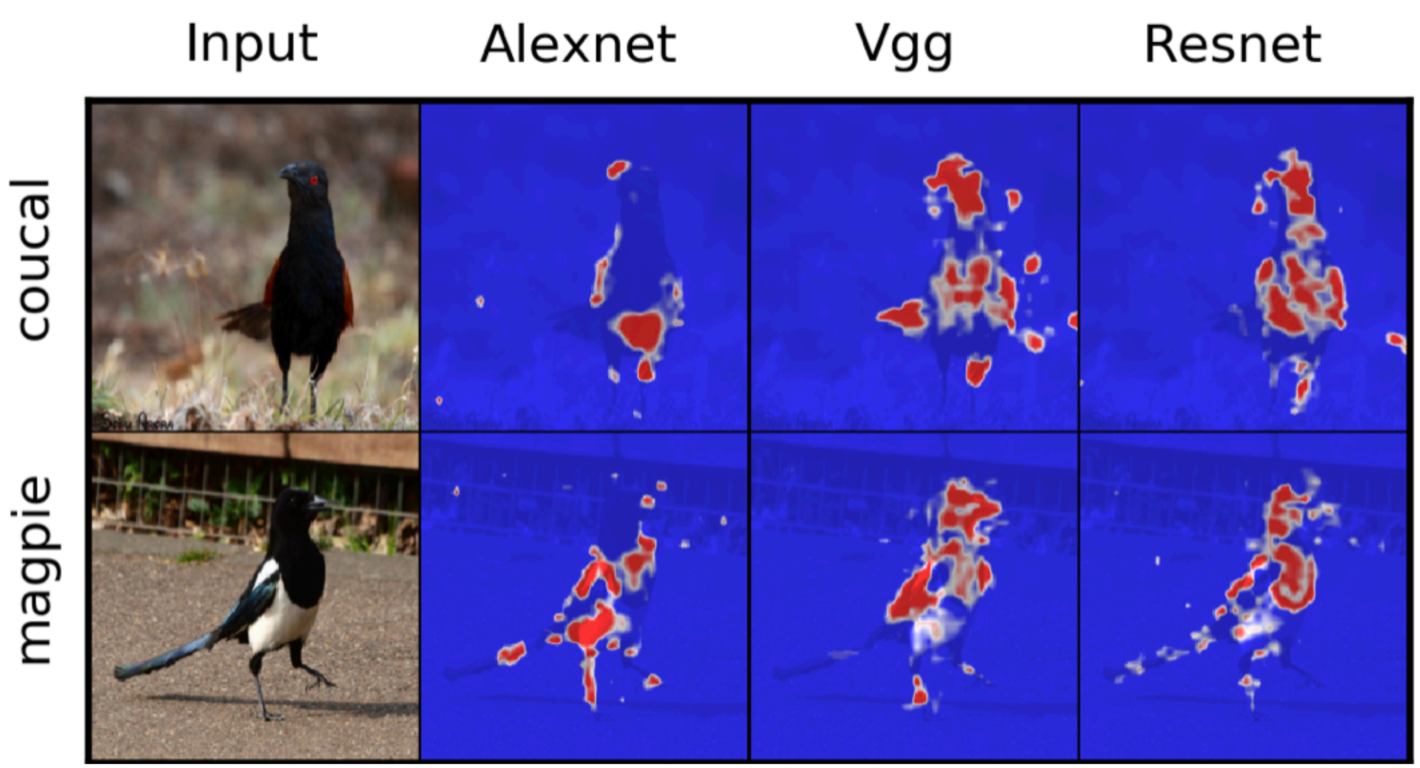
- We can ask a similar question: "**What features would I need to remove such that my model's confidence decreases?**"

    - Or: "**What features would I need to remove such that my model's prediction changes?**"

- This gives a **model-agnostic** question we can answer to provide insights to users. — **Explanation has Clear Meaning**

Counterfactual Explanations for Image Classification





Chang et. al., 2018

Uncertainty in ML

People saying AI will take over the world:

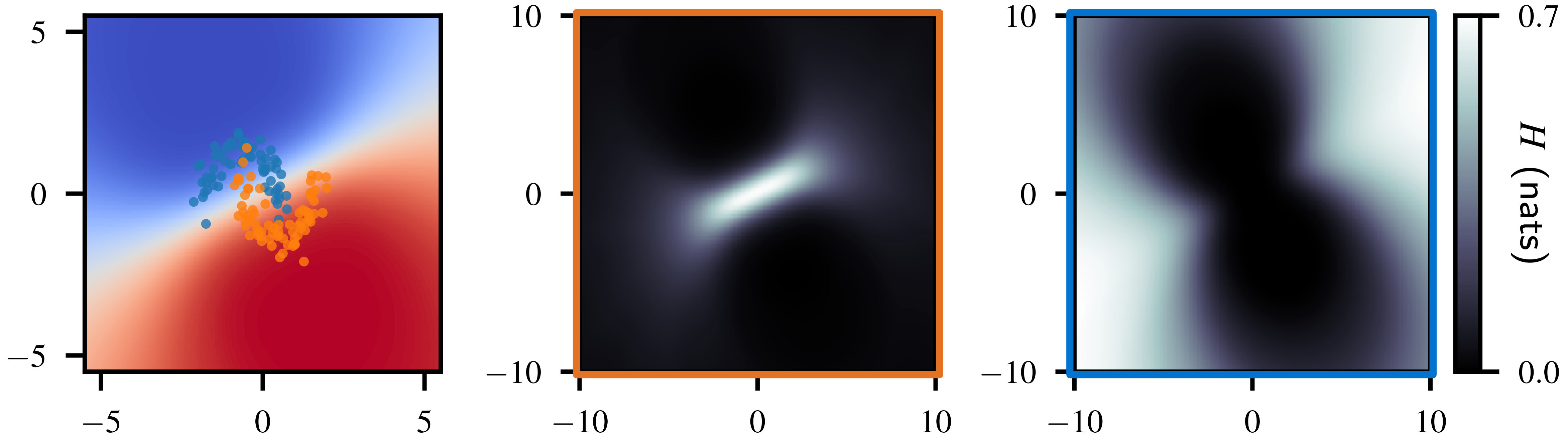
Meanwhile, my Deep Neural Network:

Sources of Uncertainty

Is there class overlap in our data? — Noise (Aleatoric) Uncertainty

Have we observed enough data to make confident predictions? — Model (Epistemic) Uncertainty

Bayesian Linear Regression Example:

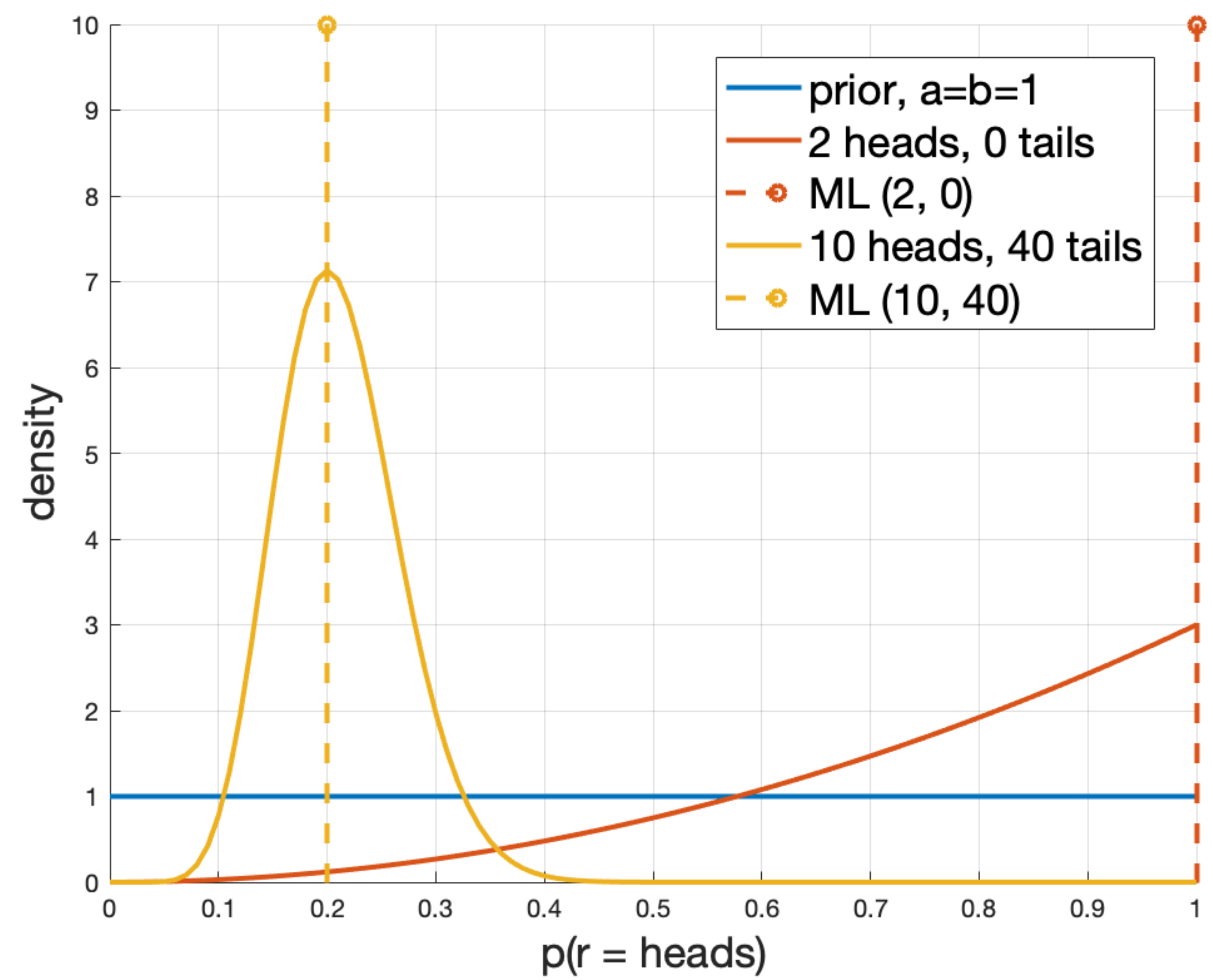


What is Model Uncertainty (1/3)

Likelihood

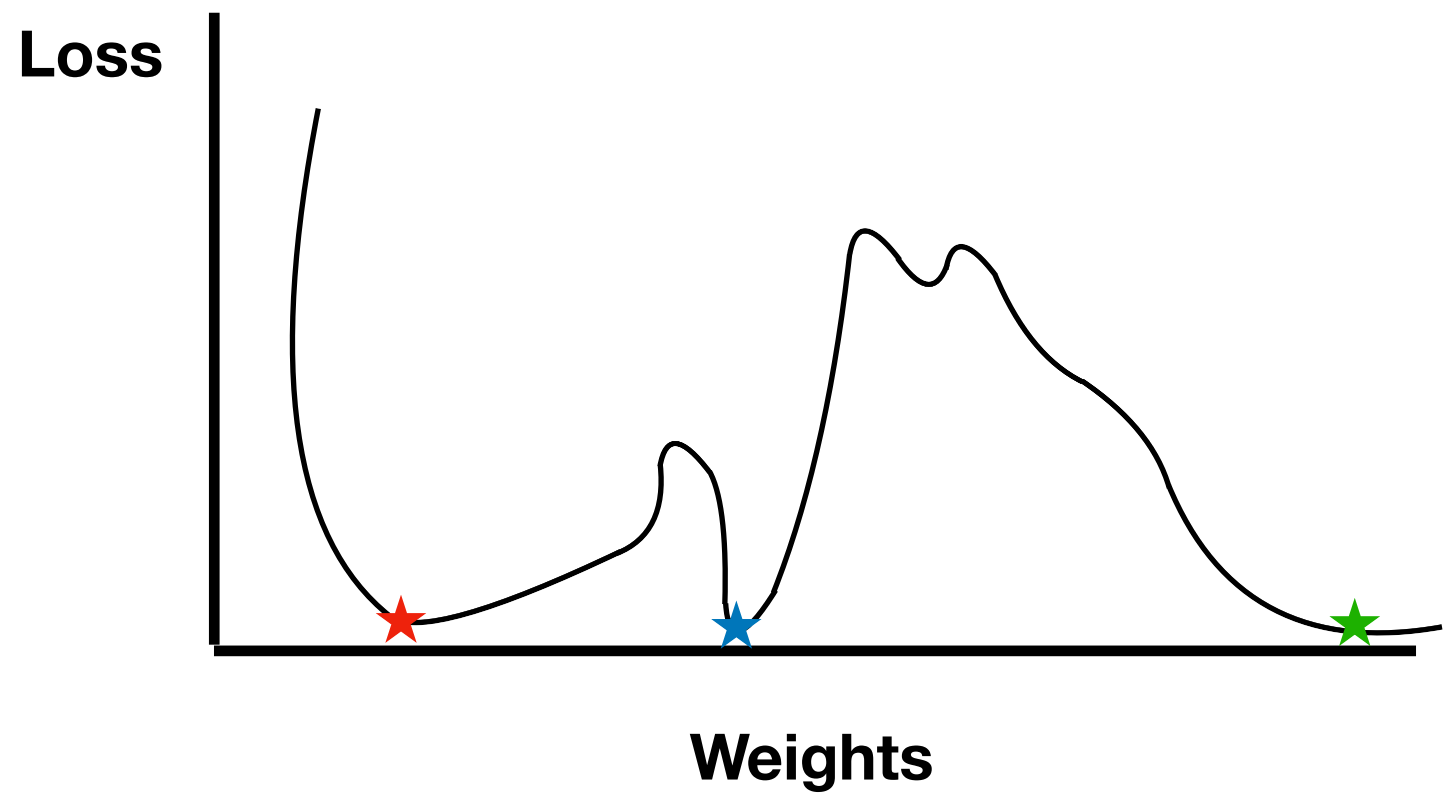
Prior

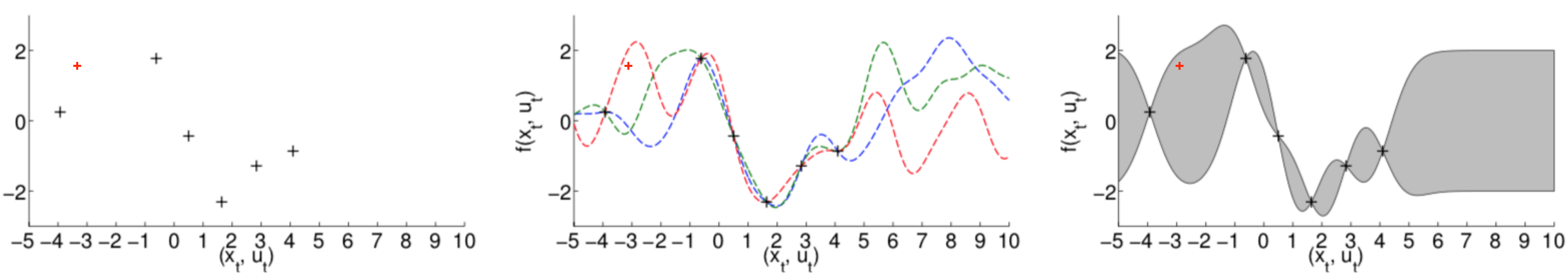
$$p(\mathbf{w}|\mathscr{D}) = \frac{p(\mathscr{D}|\mathbf{w})p(\mathbf{w})}{p(\mathscr{D})}$$



What is Model Uncertainty (2/3)

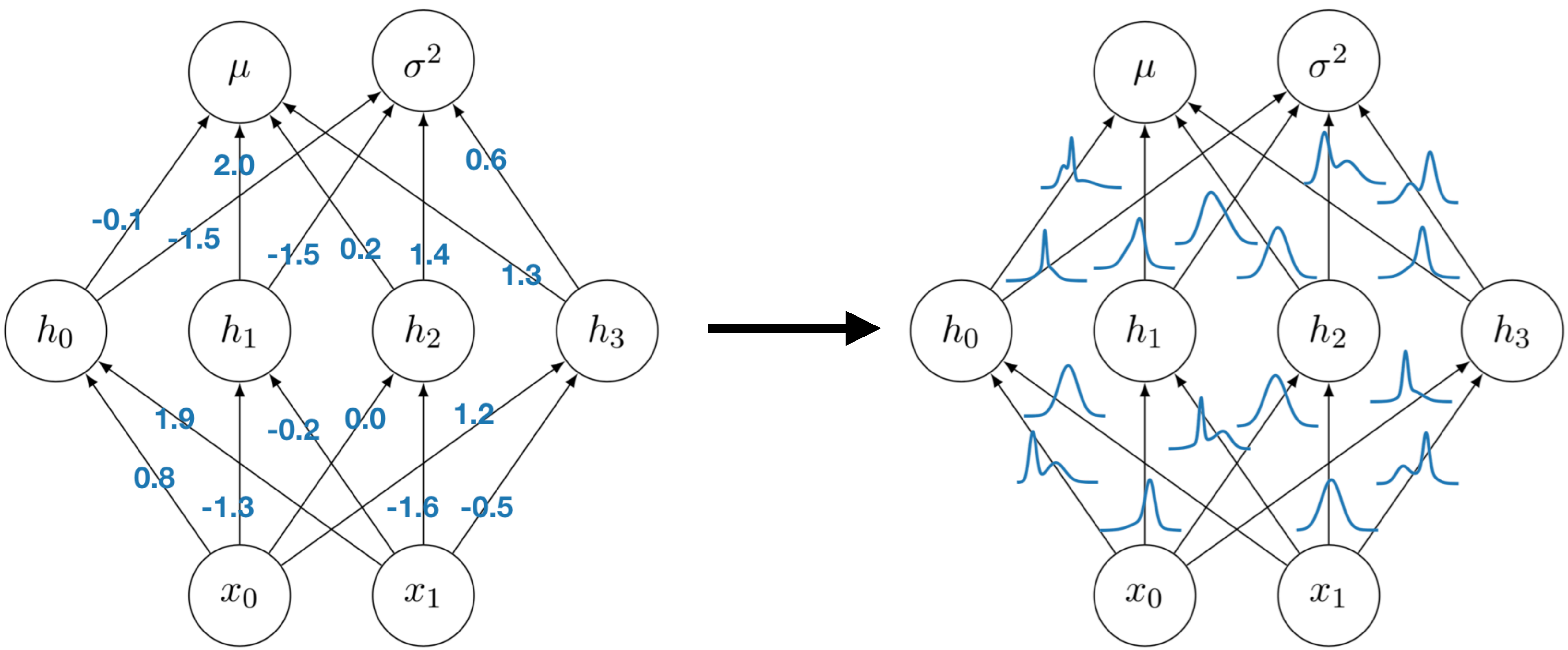
Neural Net Non-Convex Optimisation Landscape:



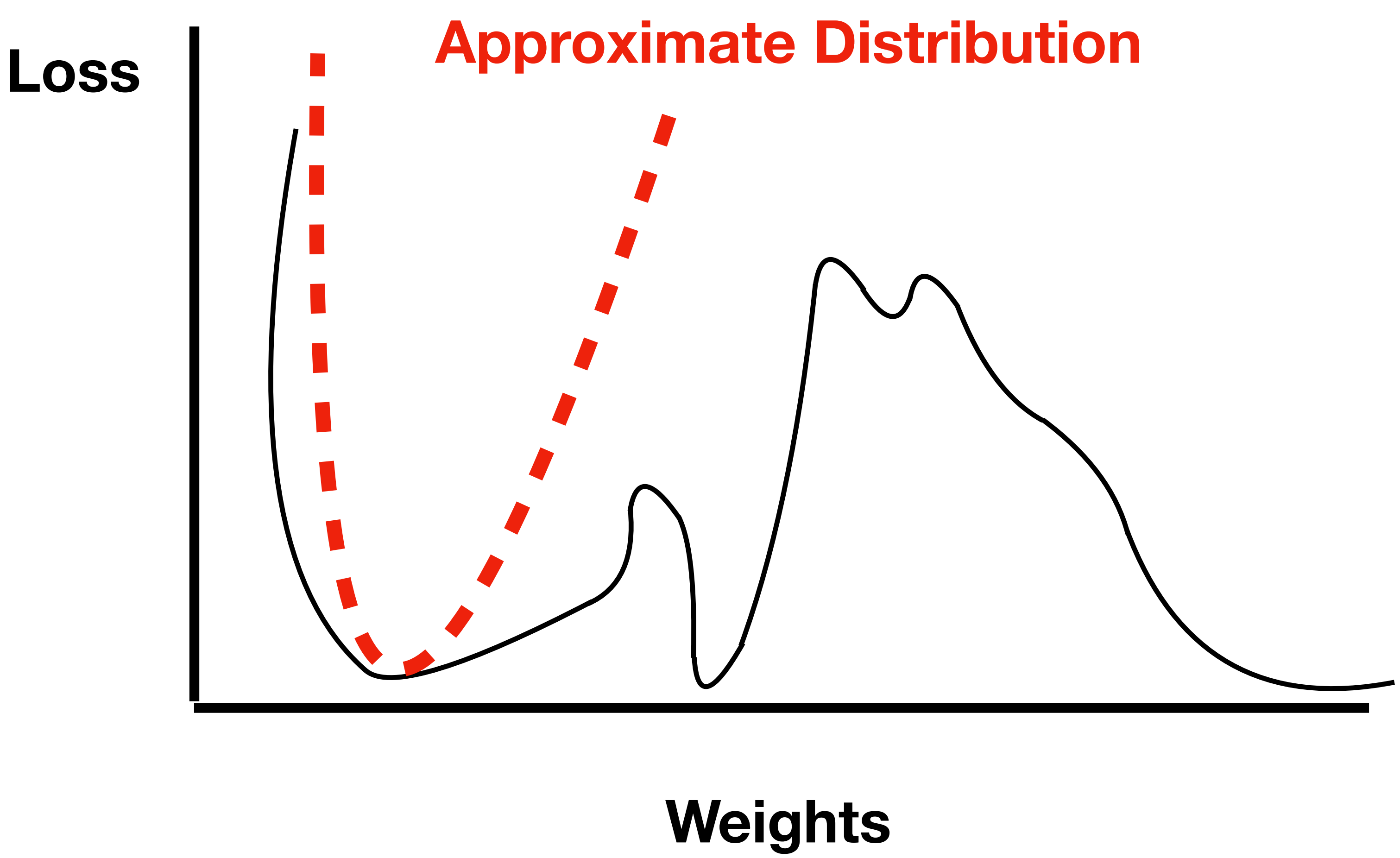


What is Model Uncertainty (3/3)

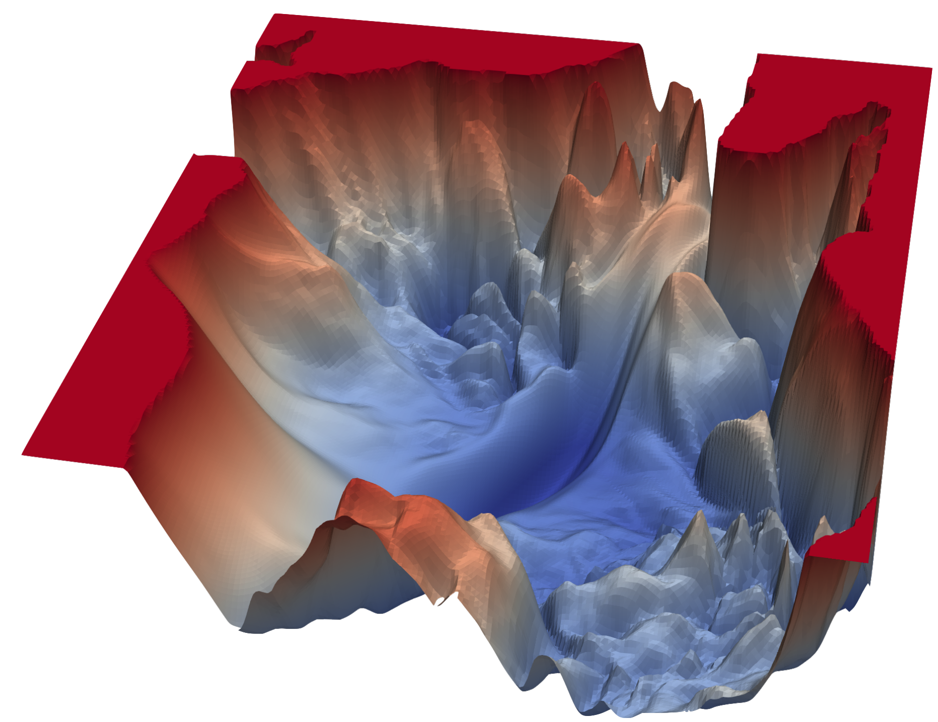
Weights go from single values to probability distributions!



Approximations



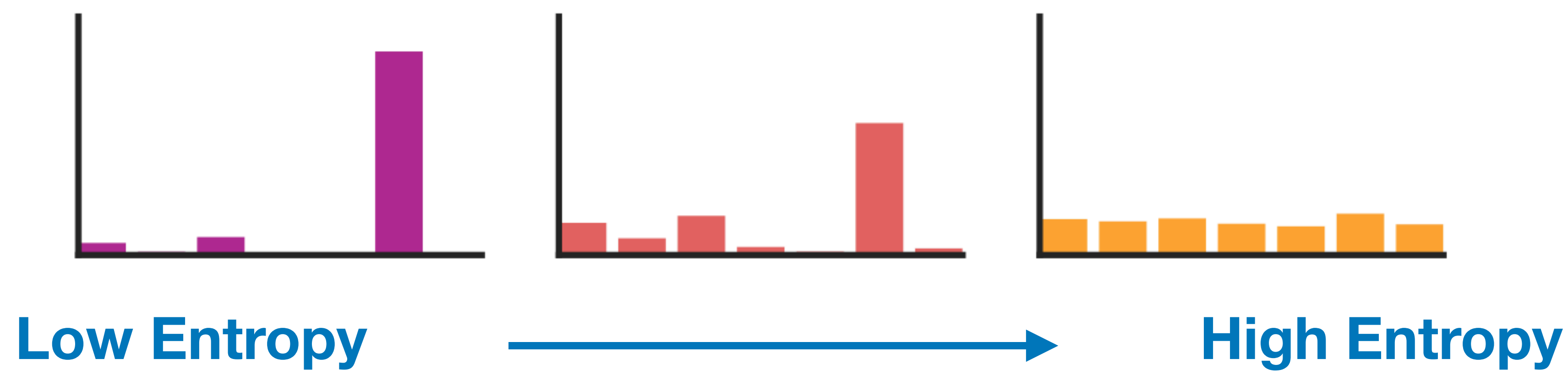
Realistic Loss Landscape



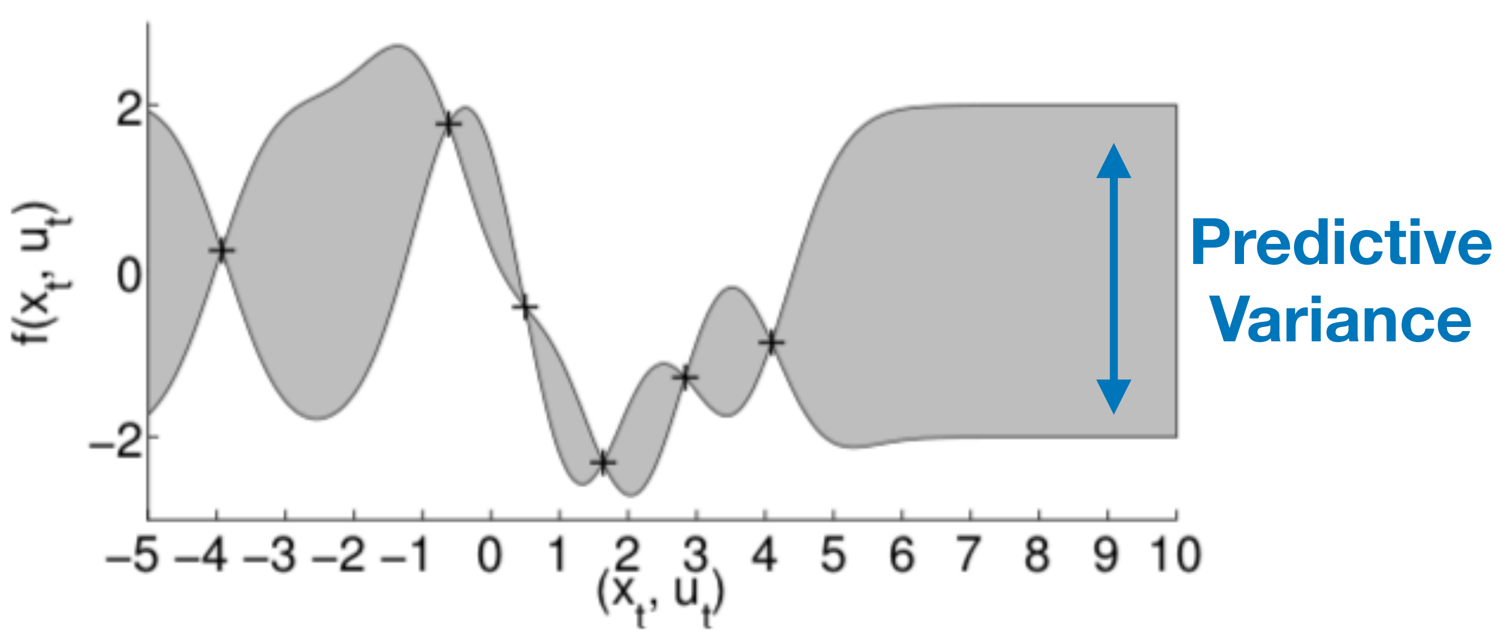
Our Uncertainty Estimates are Almost Always Biased by Our Approximations

Quantifying Uncertainty

- Classification

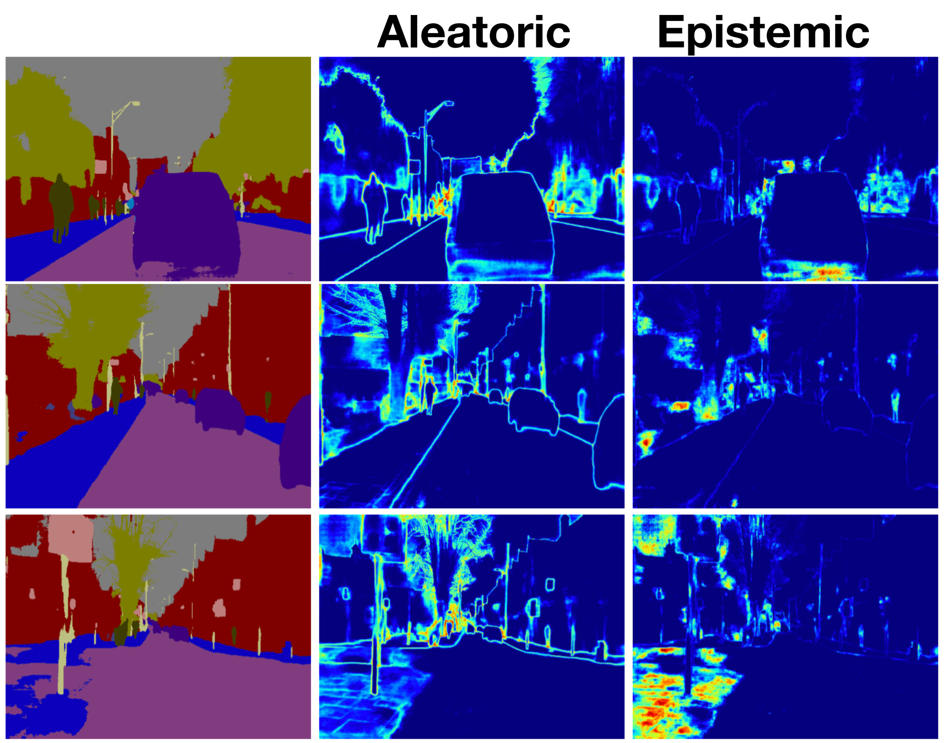


- Regression



Uncertainty in Practise

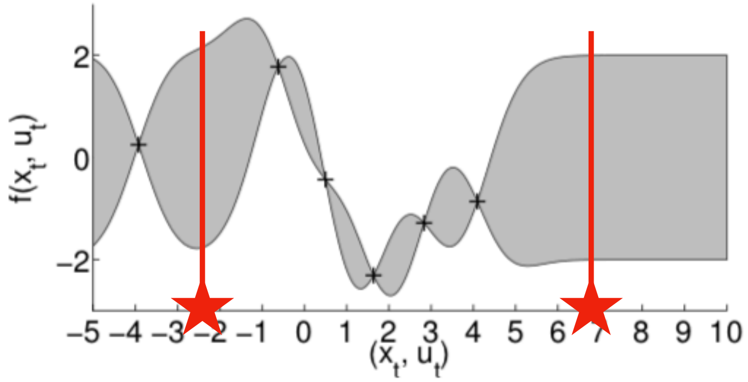
- **Robustness in Critical Applications** (Driving, Medical Diagnosis, etc) — **Reject Option**



Kendall and Gal, 2017

- **Dataset Building, Safety, Fairness** (Identifying disparities in representation of subgroups in data, etc)

- **Active Learning** (Sparse Labels, Drug Discovery)



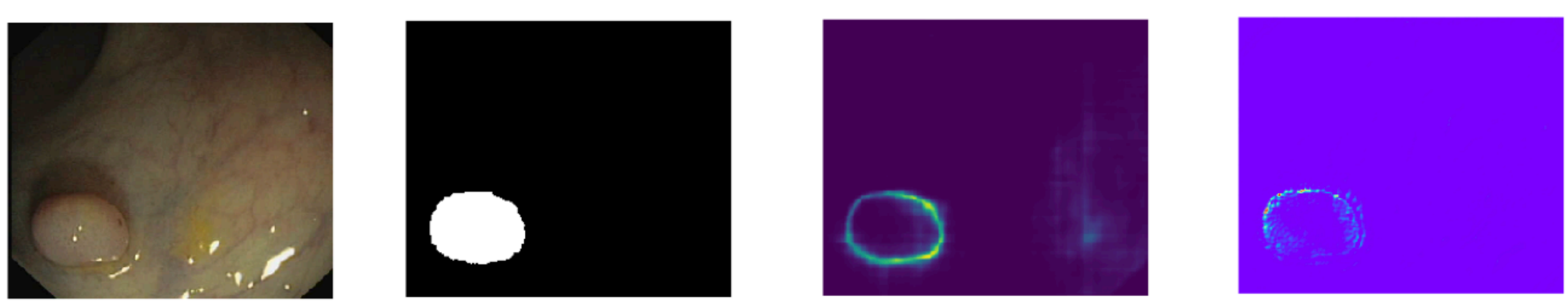
- Try out Bayesian Deep Learning with our **Public Repo**

github.com/JavierAntoran/Bayesian-Neural-Networks

Are Uncertainty Aware Systems Interpretable?

- **Thankfully, Yes!***
- **They are as interpretable as regular ML models****
- **Uncertainty can help users understand prediction in some cases**

Polyp segmentation example:



(a) input image (c) EFCN-8 prediction (e) EFCN-8 uncertainty (g) EFCN-8 interpretability

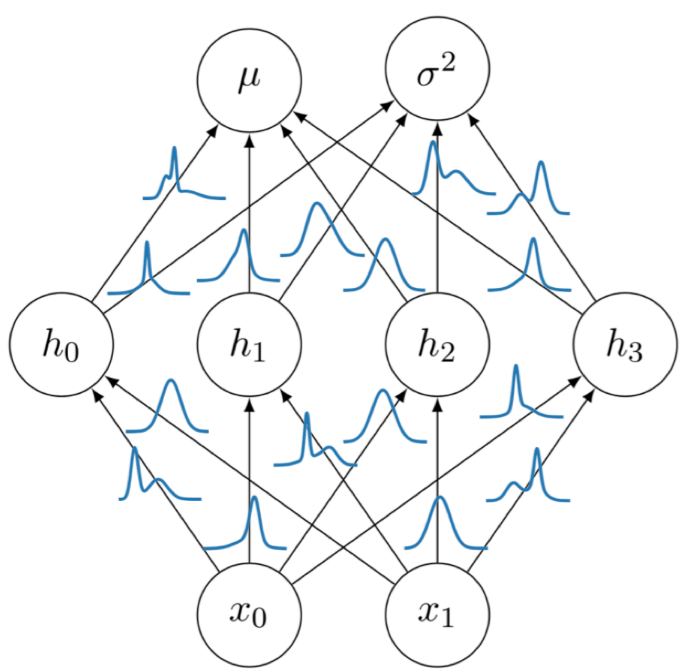
Wickstrøm, et. al., 2019

**But what about when our ML System Doesn't Know the Answer?

- LIME [Ribeiro et al., 2017]
- SHAP [Lundberg et al., 2017]
- Integrated Gradients [Sundararajan et al., 2017]
- FIDO [Chang et al., 2017]

ML User / Practitioner Workflow:

$\mathbf{x}^*$ →

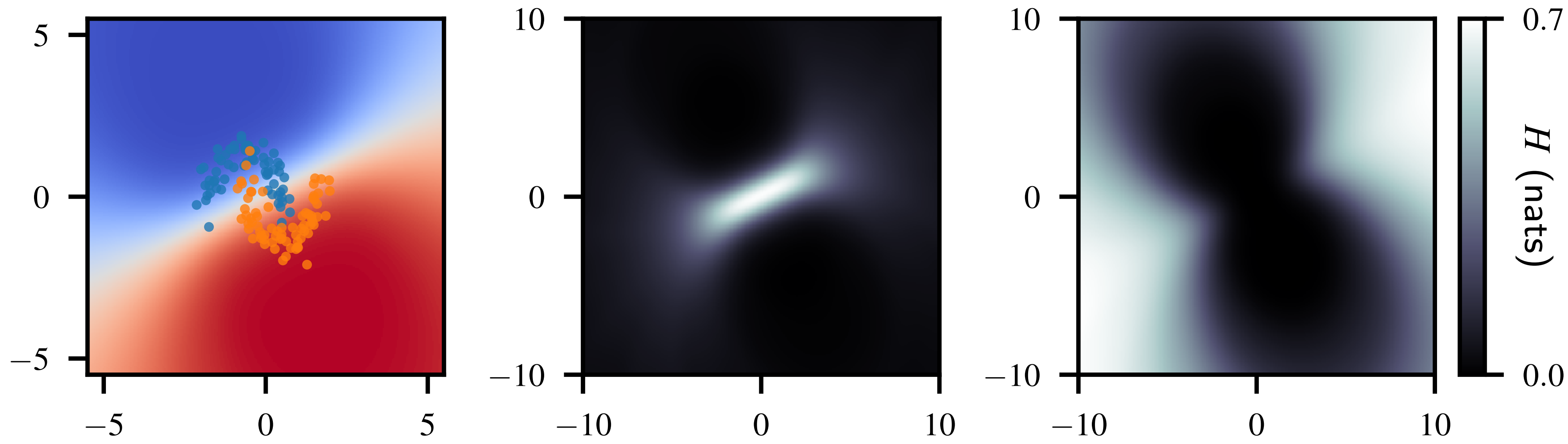


Accept a Certain Prediction → **Get Explanation**

Reject an Uncertain Prediction → ?

Explaining Uncertainty Estimates

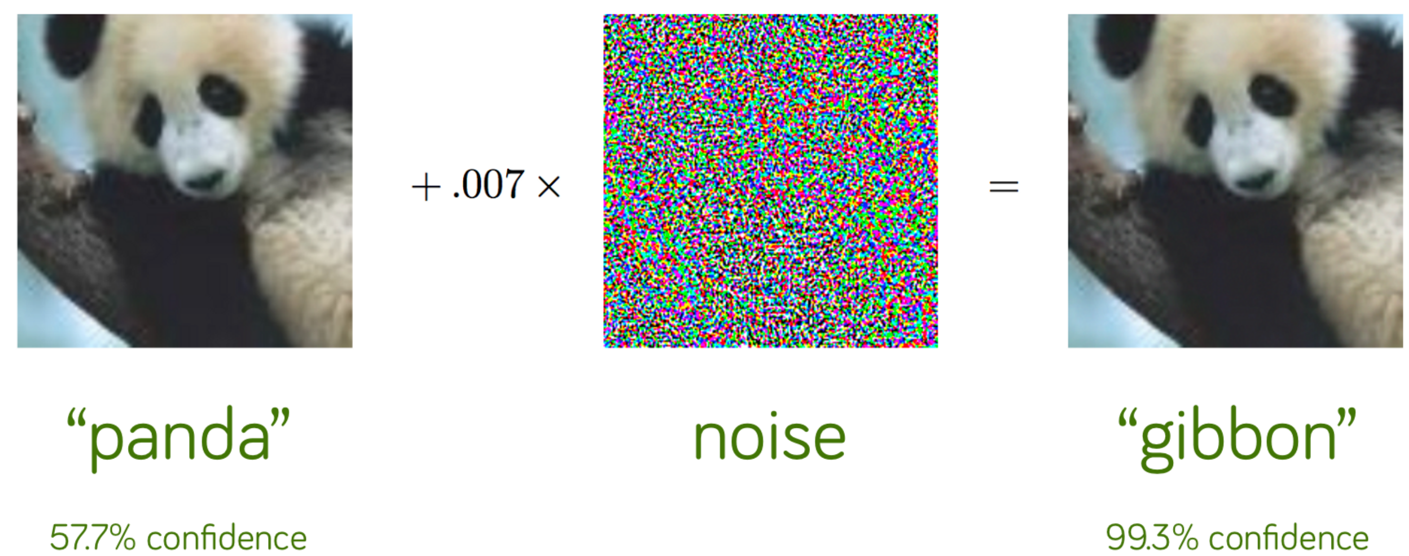
- We would like to highlight the evidence for each class.
- What if there is conflicting evidence (noise) or a lack of evidence for predefined classes (model uncertainty)?
- Recall that Uncertainty Estimates are Non-Linear even for simplest models.



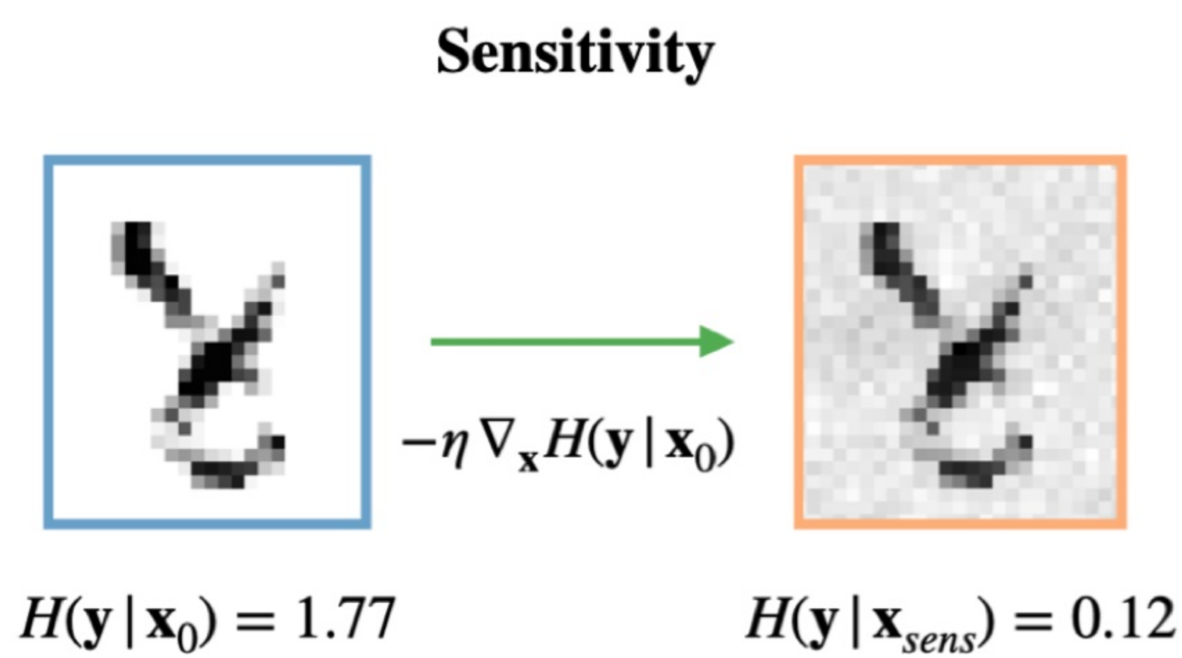
- **Problem is well posed again when using Counterfactuals**

How can we Ensure that Counterfactuals are Relevant?

- Adversarial examples

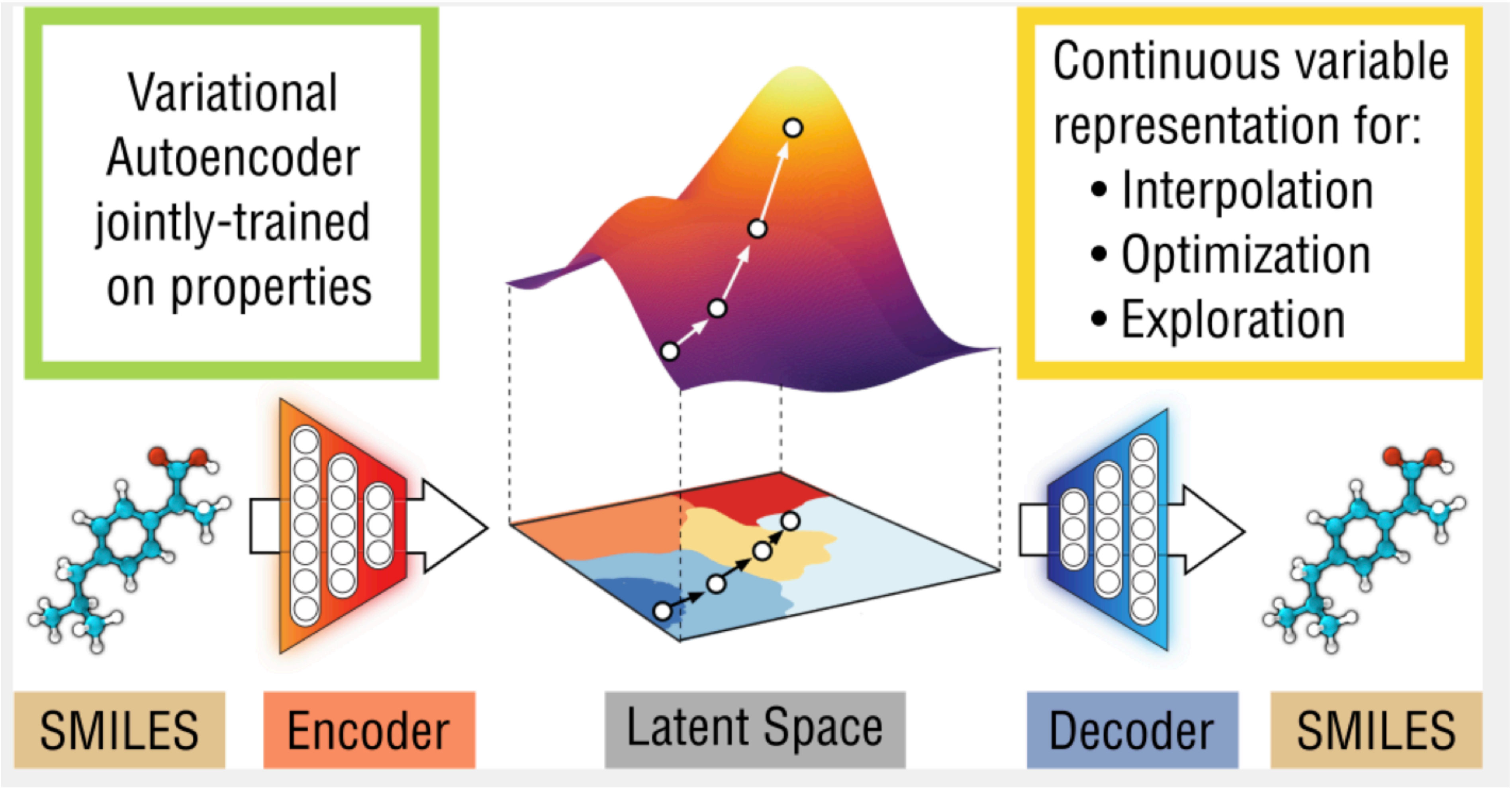


- Adversarial examples for uncertainty



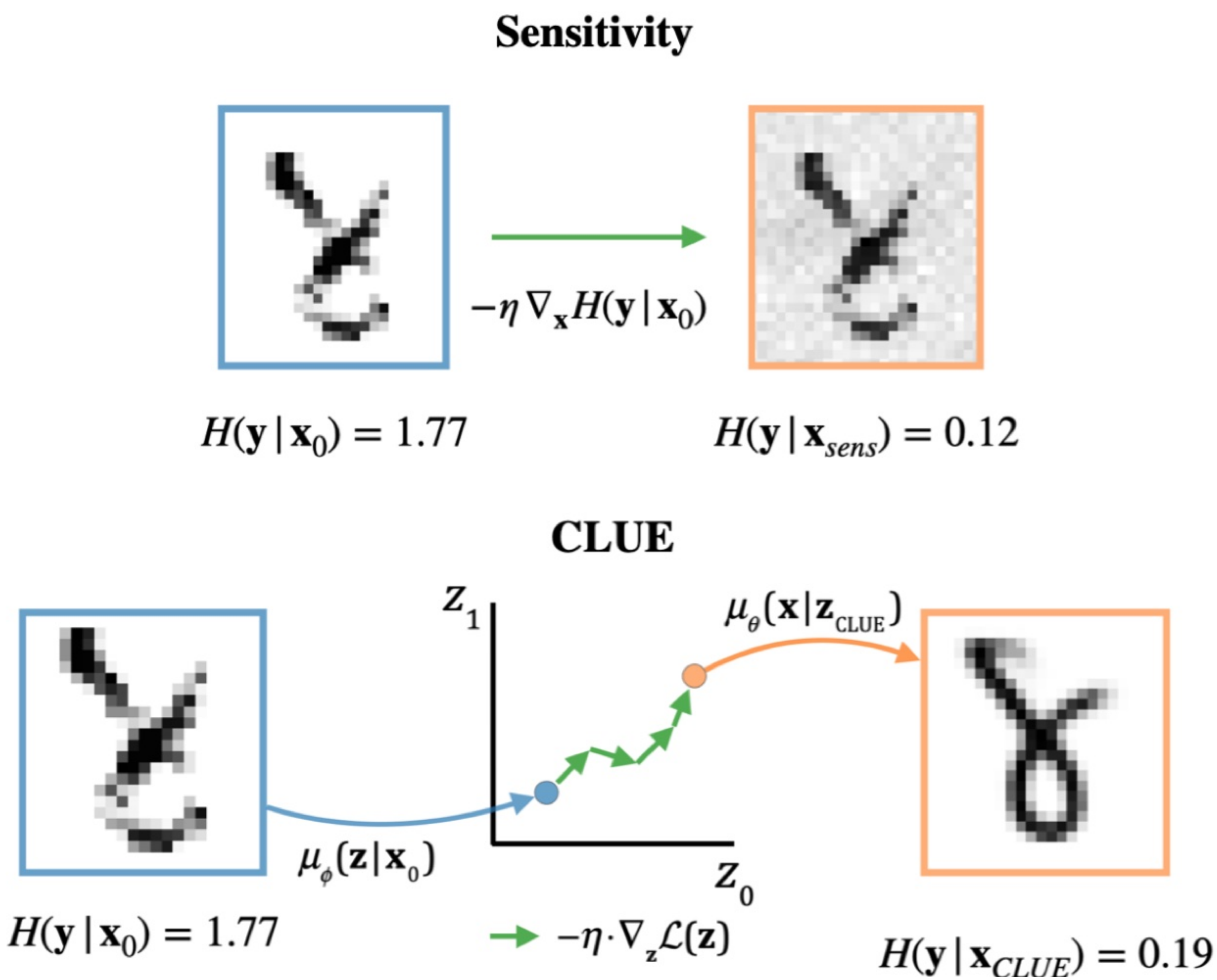
Lets Look at Data Driven Drug Discovery

- **We can** **restrict hypothesis space to manifold captured by generative model**: **this ensures relevant proposals**



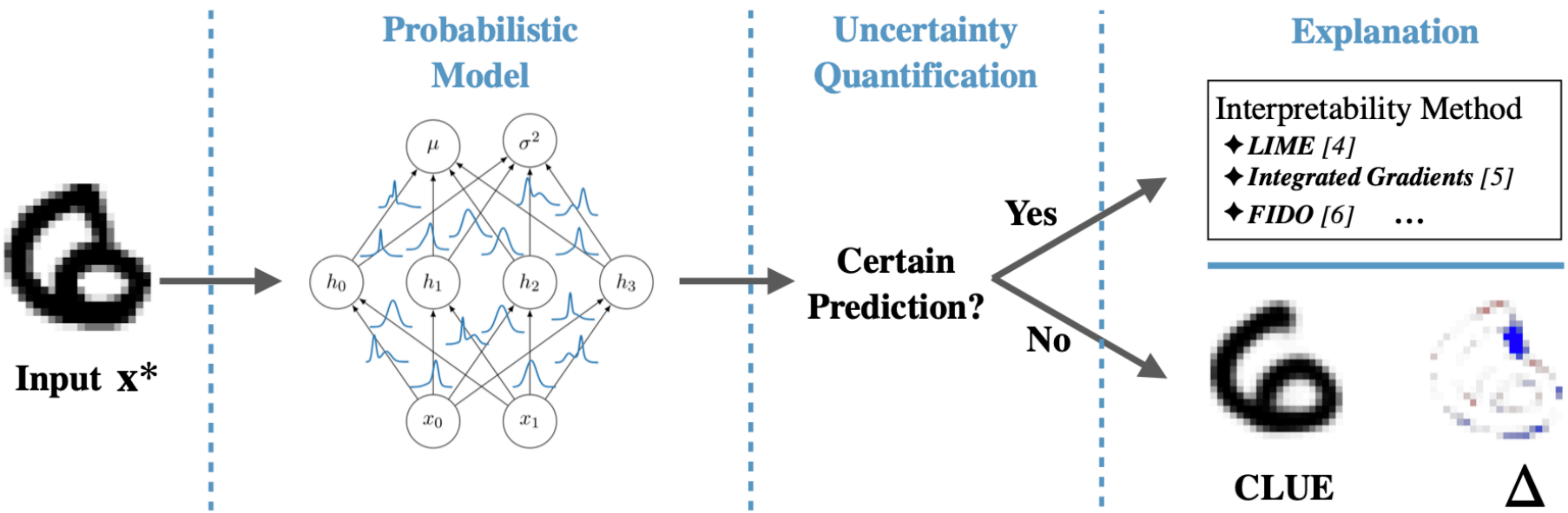
Gómez-Bombarelli et. al., 2018

Lets do the same for Counterfactuals



CLUE: Counterfactual Latent Uncertainty Explanations

"What is the **smallest change** we need to make to an input, **while staying in-distribution**, such that our model produces more **certain predictions?**"



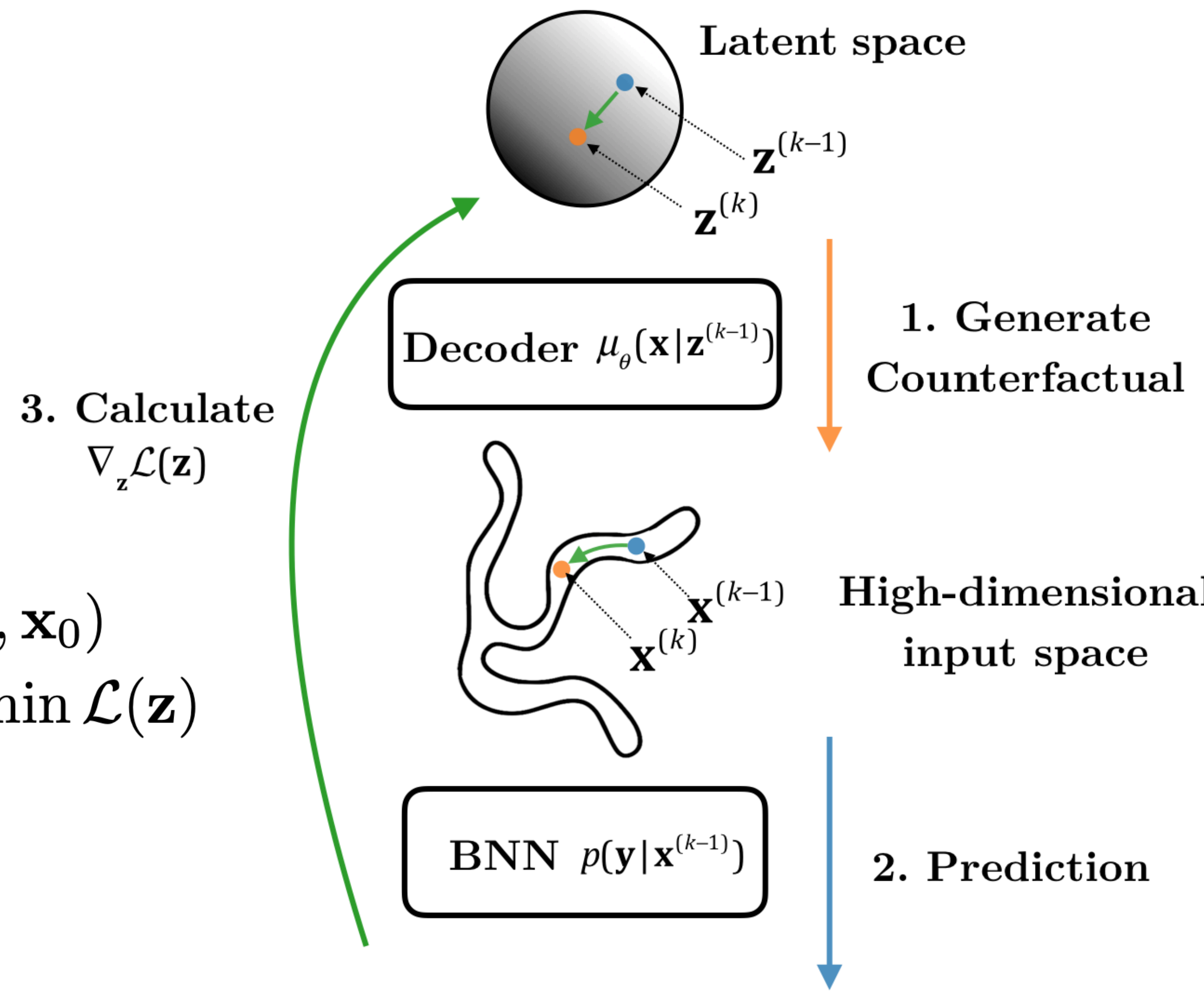
The CLUE Algorithm

Iterative Optimisation:

$$\mathcal{L}(\mathbf{z}) = H(\mathbf{y}|\mu_\theta(\mathbf{x}|\mathbf{z})) + d(\mu_\theta(\mathbf{x}|\mathbf{z}), \mathbf{x}_0)$$

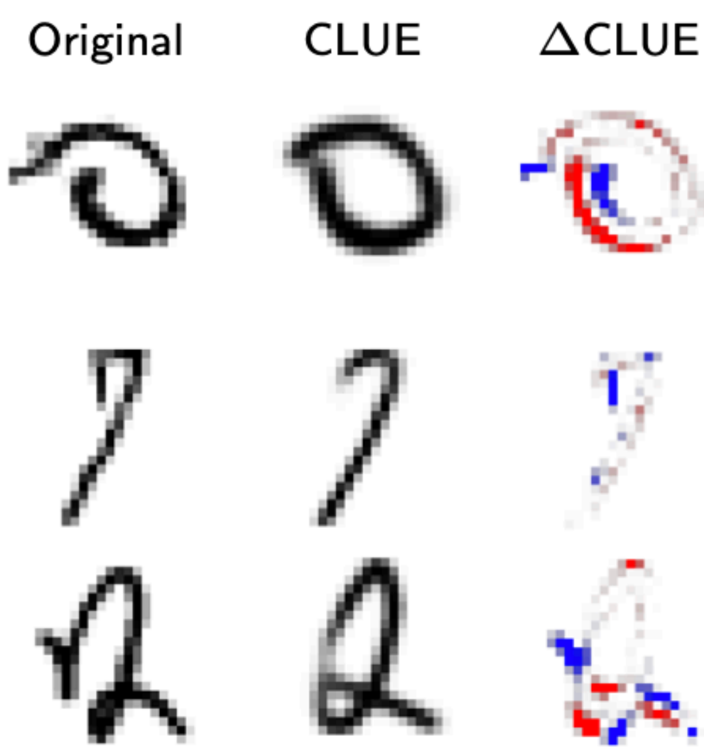
$$\mathbf{x}_{\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}}); \quad \mathbf{z}_{\text{CLUE}} = \arg\min_{\mathbf{z}} \mathcal{L}(\mathbf{z})$$

$$d_x(\mathbf{x}, \mathbf{x}_0) = \|\mathbf{x} - \mathbf{x}_0\|_1$$



Displaying CLUEs to Users

$$\Delta \mathbf{x} = \mathbf{x}_{\mathrm{CLUE}} - \mathbf{x}_0$$



(a) MNIST

	Original
LSAT	41.0
UGPA	3.7
Race	Asian
Sex	Female

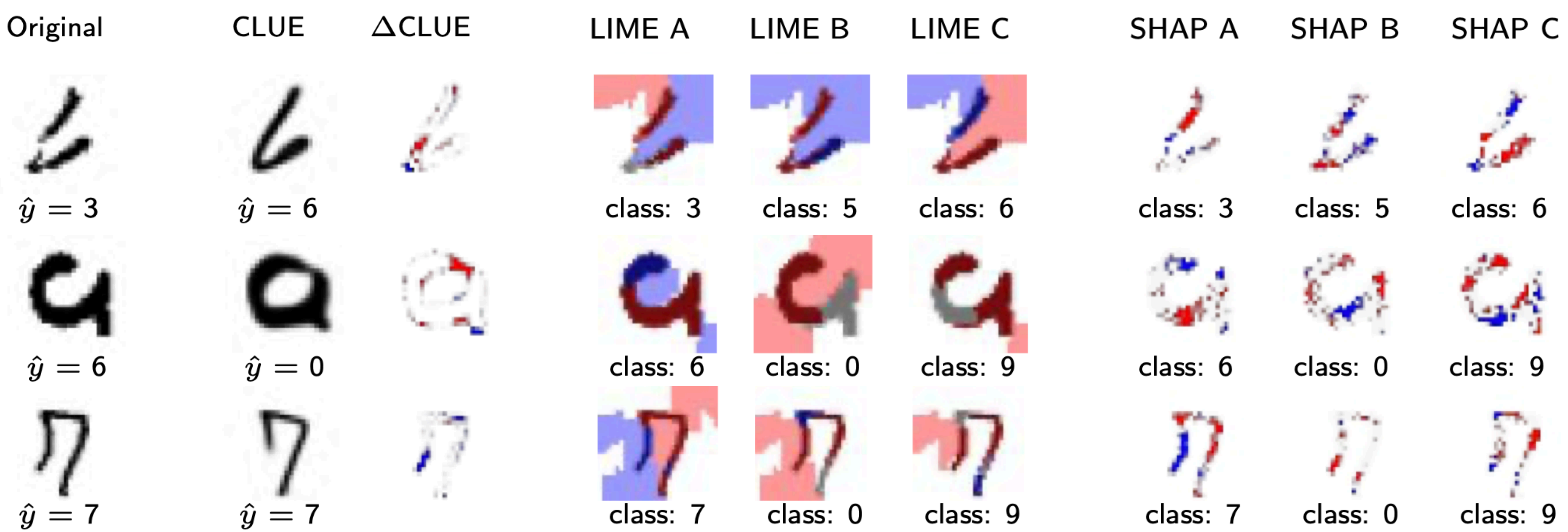
	CLUE
LSAT	41.0
UGPA	3.0
Race	White
Sex	Female

(b) LSAT

Figure 5: Example image and tabular CLUEs.

Comparing CLUE to Feature Importance (LIME / SHAP)

- In high uncertainty scenarios, it is difficult to build an explanation in terms of the provided information (features)
- CLUE's counterfactual nature allows it to add new information



User Study: Setup (1/2)

Human Simulability: Users are shown context examples and are tasked with predicting model behaviour on new datapoint.

	Uncertain
Age	Less than 25
Race	Caucasian
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	Yes
Prior Convictions	1
Days Served	0

	Certain
Age	Less than 25
Race	African-American
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	No
Prior Convictions	0
Days Served	0

	?
Age	Less than 25
Race	Hispanic
Sex	Male
Current Charge	Misdemeanour
Reoffended Before	No
Prior Convictions	0
Days Served	0

User Study: Setup (2/2)

Tasks:

- **COMPAS (Criminal Recidivism Prediction, 7 dim)**
- **LAST (Academic Performance Prediction, 4 dim)**

Users:

- **University Students with ML experience**
- **10 Users per approach, 10 Questions per Dataset**

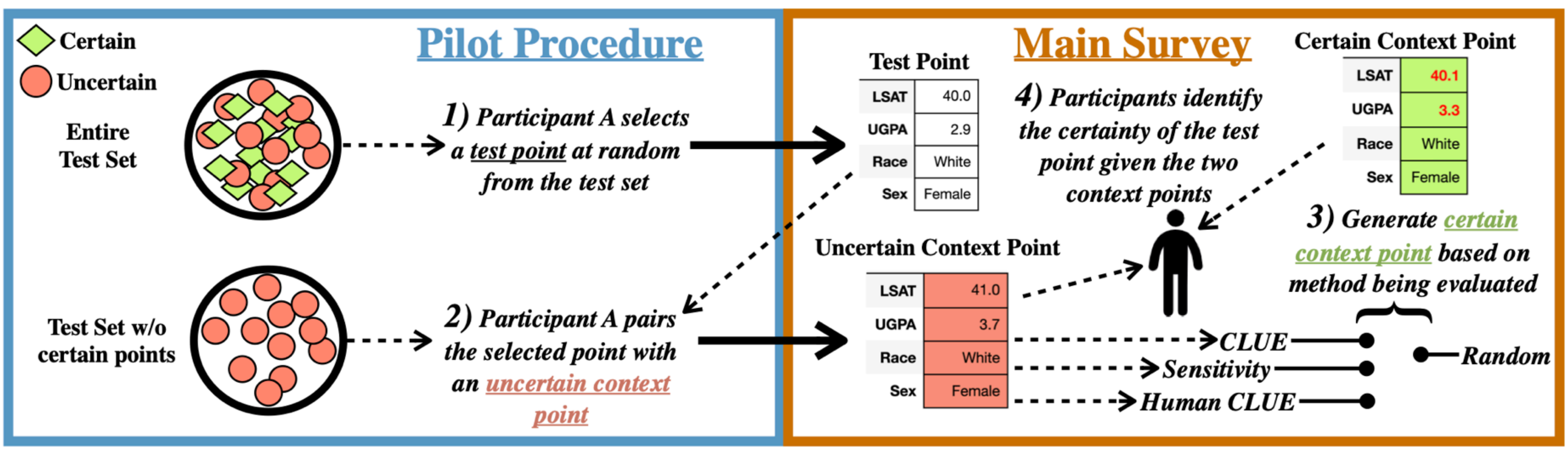


Figure 8: Experimental workflow for our tabular data user study.

User Study: Results

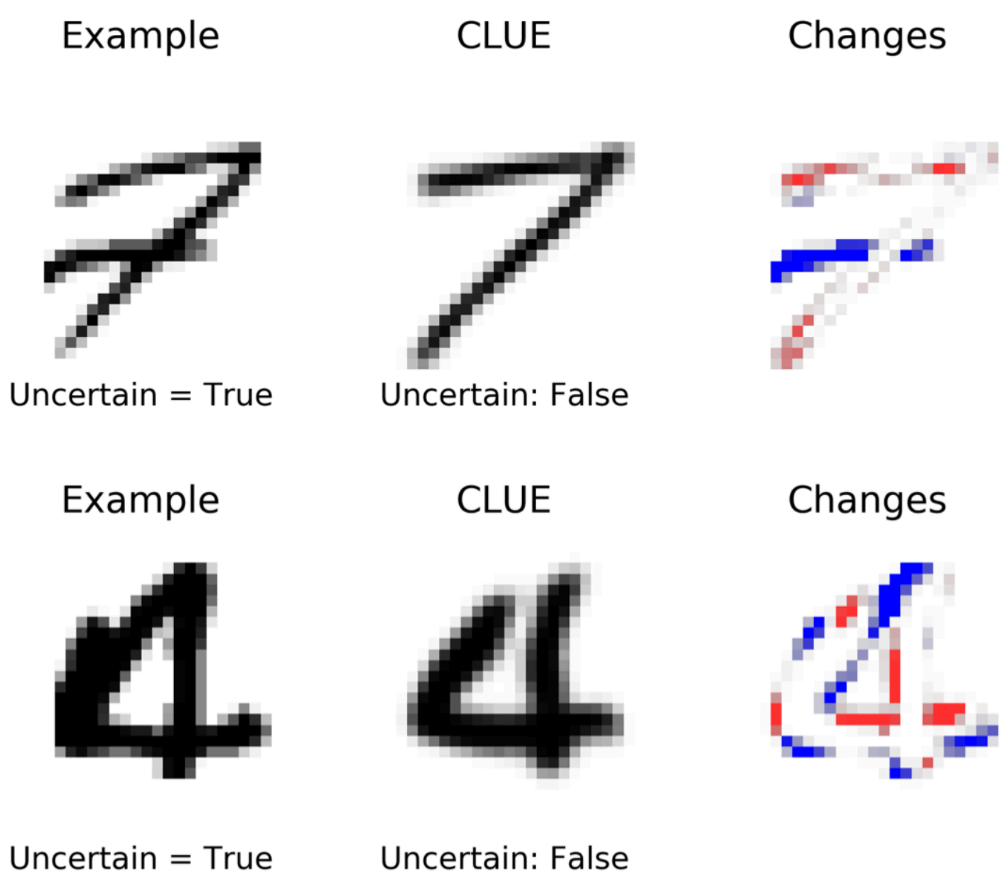
Method	N. participants	Accuracy (%)
Random	10	61.67
Sensitivity	10	52.78
Human	10	62.22
CLUE	10	**82.22**

CLUE's improvement over all other approaches is statistically significant

(Using Nemenyi test for average ranks across test questions)

Now with Images:

We modify the MNIST train set to introduce Out Of Distribution (model) uncertainty.



Method	N. participants	Accuracy
Unc.	5	0.67
CLUE	5	0.88

Thanks to my Collaborators!

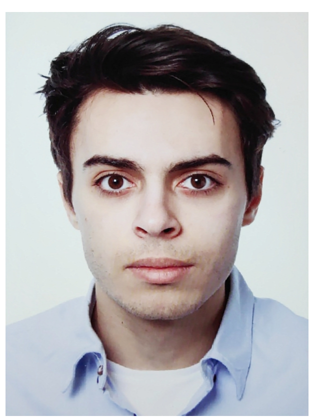
Javier Antorán
ja666@cam.ac.uk

Umang Bhatt
usb20@cam.ac.uk

Tameem Adel
tah47@cam.ac.uk

Adrian Weller
aw665@cam.ac.uk

José Miguel Hernández-Lobato
jmh233@cam.ac.uk



Questions?

Read The Full Paper at:
arxiv.org/abs/2006.06848

See More of my Research (+slides):
javierantoran.github.io/about/

Contact Me:
ja666@cam.ac.uk