UNIVERSITY OF CAMBRIDGE

Getting a CLUE: A Method for Explaining Uncertainty Estimates

ML-IRL Workshop at ICLR 2020

Javier Antorán, Umang Bhatt, Tameem Adel,

Adrian Weller, and José Miguel Hernández-Lobato

About Us

Javier Antorán
ja666@cam.ac.uk

Umang Bhatt
usb20@cam.ac.uk

Tameem Adel
tah47@cam.ac.uk

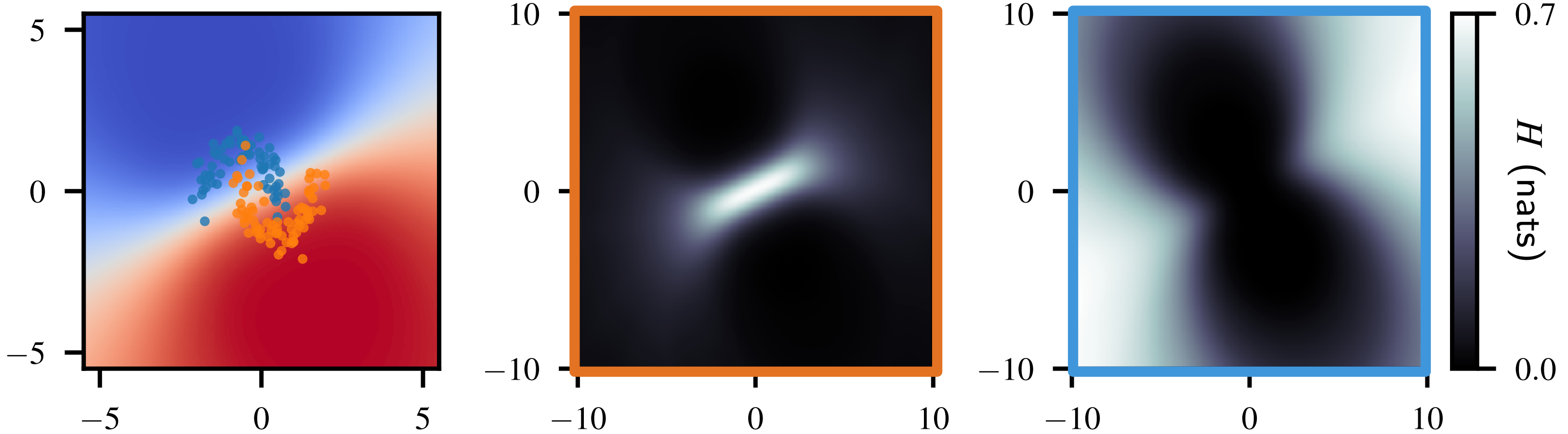
Adrian Weller
aw665@cam.ac.uk

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

Uncertainty in Predictive Models

Is there class overlap in our data?

Have we observed enough data to make confident predictions?



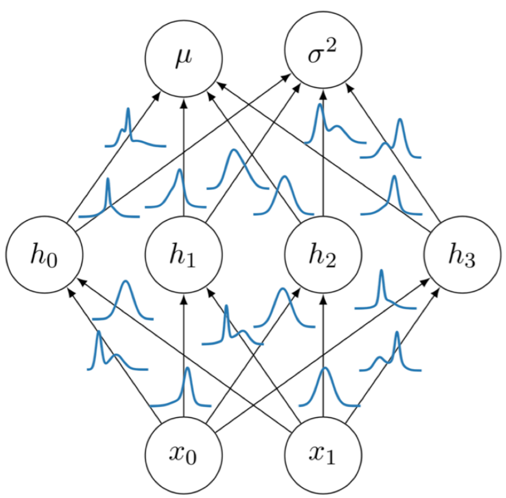
Quantify Uncertainty through Entropy (Classification) or Variance (Regression)

Motivation: Transparency in Deep Learning via Uncertainty

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

ML Practitioner Workflow:

$\mathbf{x}^*$ →



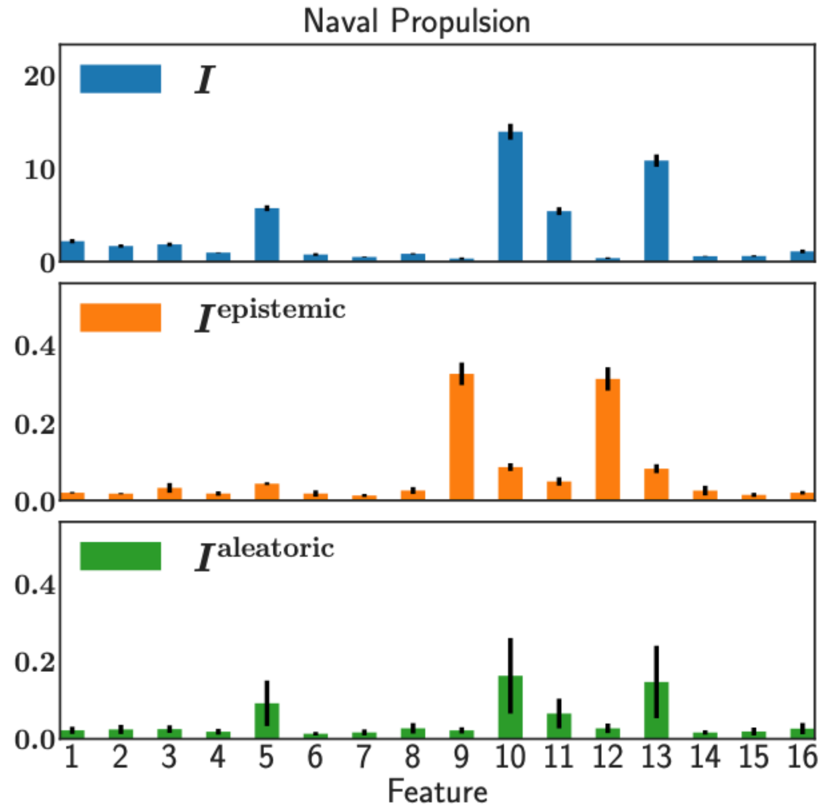
Accept a Certain Prediction → **Get Explanation**

Reject an Uncertain Prediction → ?

Related Work: Uncertainty Sensitivity Analysis

Use gradients of predictive uncertainty w.r.t. inputs

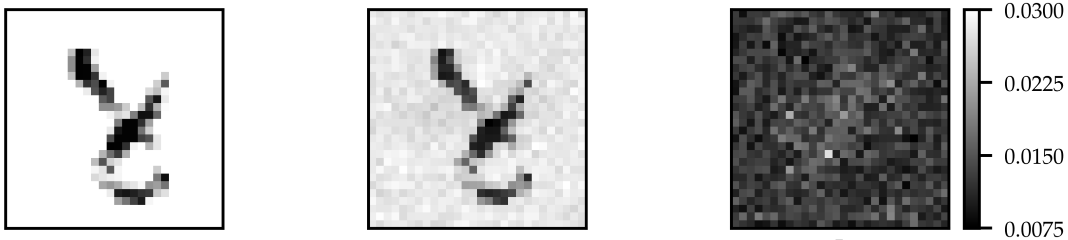
$$I_{i,k} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \left| \frac{\partial f(\mathbf{x}_n^\star)_k}{\partial x_{i,n}^\star} \right|$$



[Depeweg et. al., 2017]

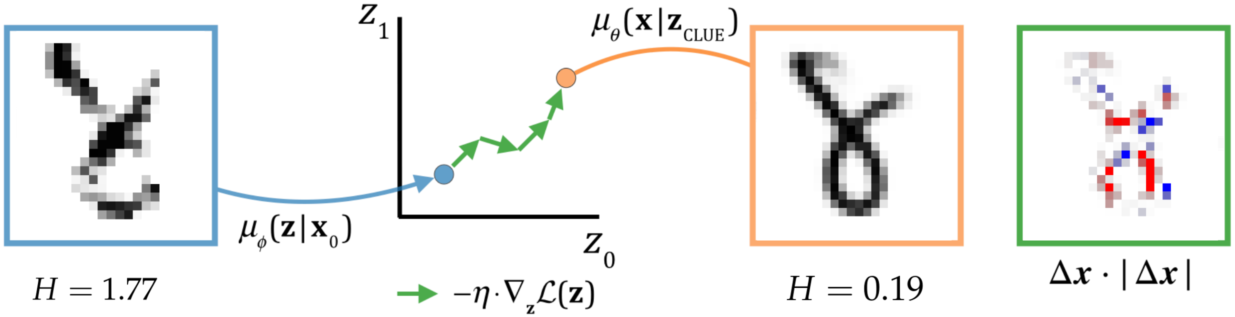
Fixing Sensitivity Analysis

Sensitivity can produce meaningless explanations in high dimensions



$H = 1.77$ $H = 0.125$ I_i

What if we could **constrain our explanations to the data manifold**?

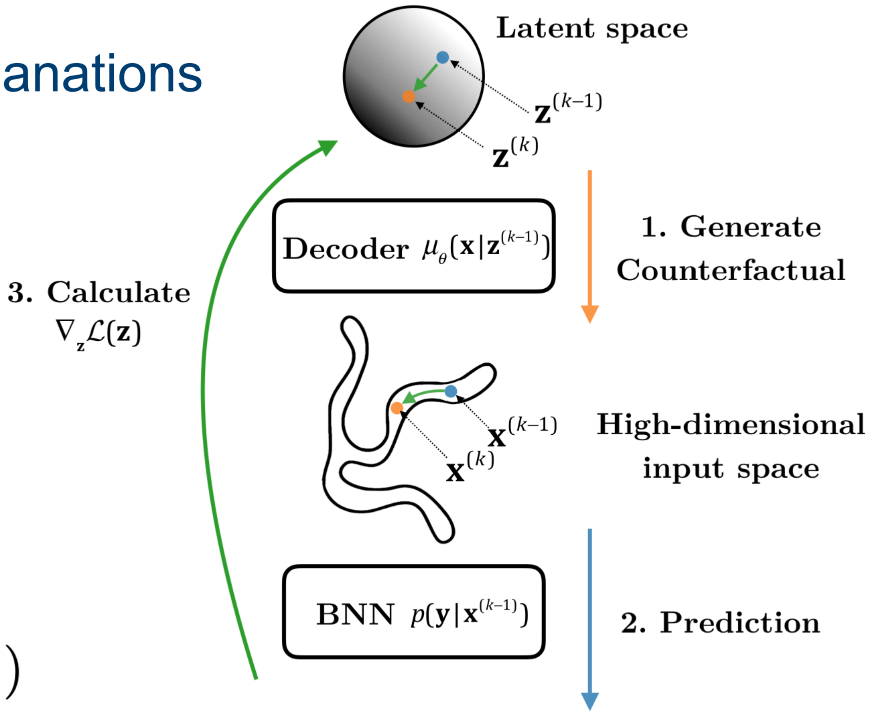


$H = 1.77$ $H = 0.19$ $\Delta x \cdot |\Delta x|$

Getting a CLUE

Use generative model as proxy for the data manifold:

Counterfactual **L**atent **U**ncertainty **E**xplanations



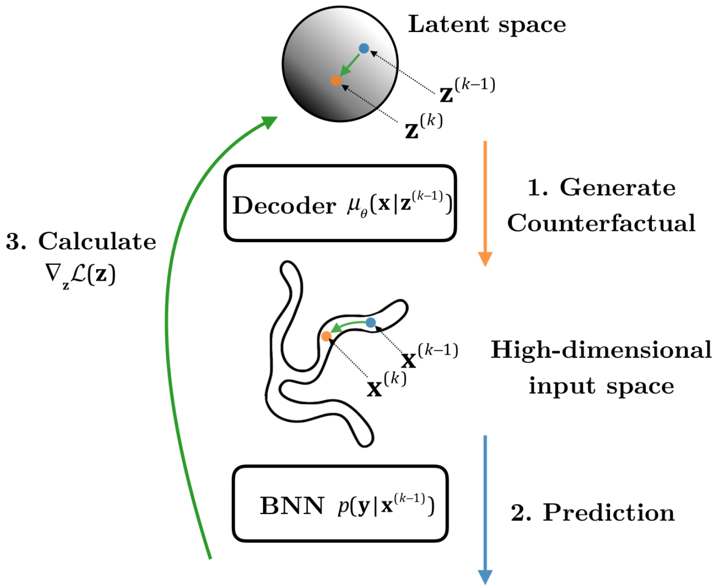
"What is the **smallest change** we need to make to an input such that our model produces more **certain predictions**"

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

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

Getting a CLUE (cont.)

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



Algorithm 1: CLUE

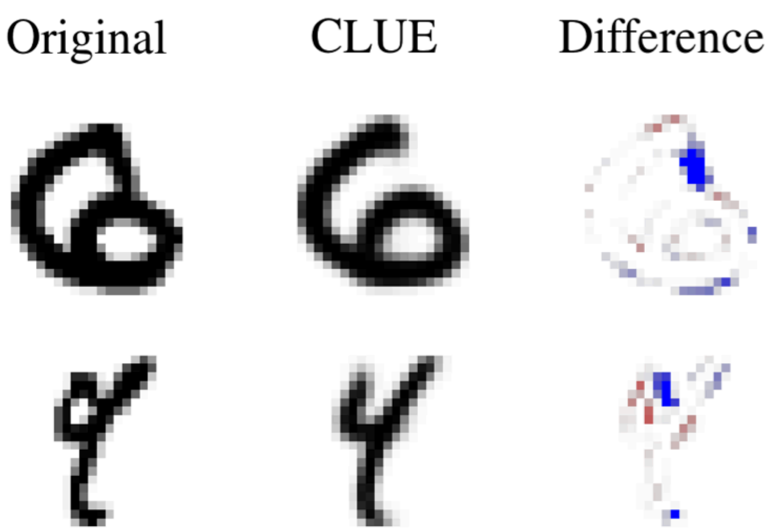
Inputs: original datapoint $\mathbf{x}_0$, distance function $d(\mathbf{x}, \mathbf{x}_0)$, BNN uncertainty estimator H, DGM decoder $\mu_\theta(\cdot)$, DGM encoder $\mu_\phi(\cdot)$

Set initial value of $\mathbf{z} = \mu_\phi(\mathbf{z}|\mathbf{x}_0)$;
while *loss $\mathcal{L}$ is not converged* **do**
    Decode: $\mathbf{x} = \mu_\theta(\mathbf{x}|\mathbf{z})$;
    Use BNN to obtain $H(\mathbf{y}|\mathbf{x})$;
    $\mathcal{L} = H(\mathbf{y}|\mathbf{x}) + d(\mathbf{x}, \mathbf{x}_0)$;
    Update $\mathbf{z}$ with $\nabla_{\mathbf{z}} \mathcal{L}$;
end
Decode explanation: $\mathbf{x}_{\text{CLUE}} = \mu_\theta(\mathbf{x}|\mathbf{z})$;

Output: Counterfactual example $\mathbf{x}_{\text{CLUE}}$

Showing CLUEs to Users

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



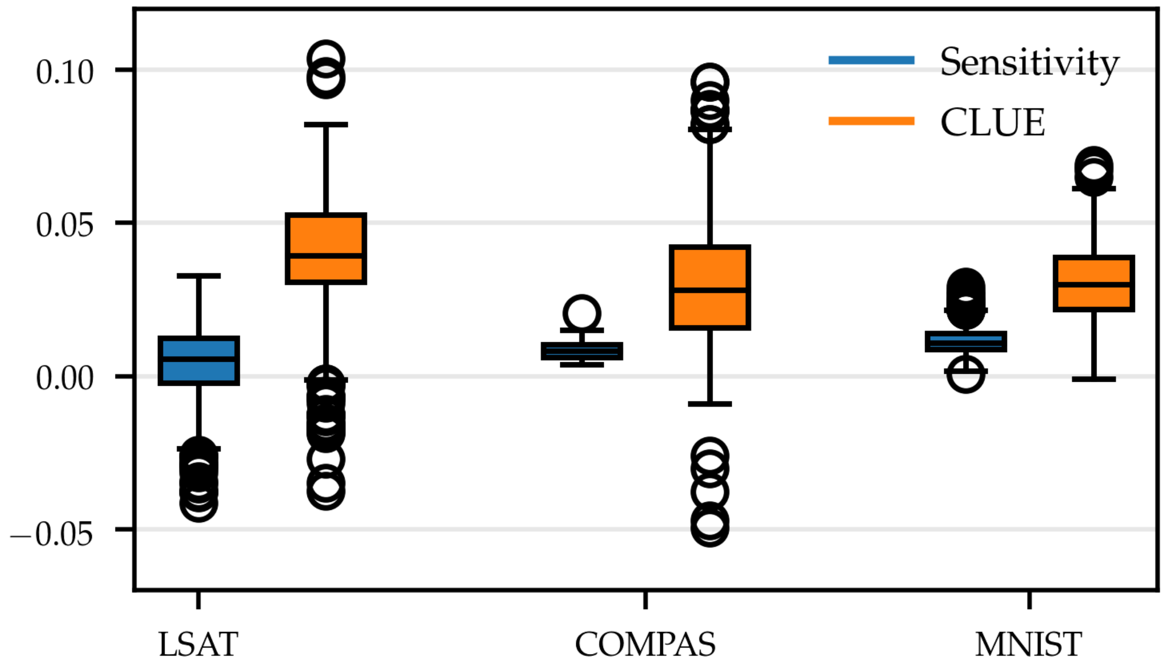
	Person 0		CLUE
AI is uncertain	True	->	False
age	Greater than 45	->	25 - 45
race	African-American		-
sex	Female	->	Male
charge degree	Felony		-
recid before	not recid		-
priors count	1.0		-
days served before	0.0		-

MNIST

COMPAS

Comparing CLUE and Sensitivity

$$f_m = \frac{\Delta_H}{\Delta_{\mathbf{x}}}$$



A Small User Study on COMPAS and LSAT

Here is a set of examples labeled with if the AI has high or low "noise uncertainty." For uncertain points, the corresponding CLUEs for 'noise' uncertainty are shown. Given this information, in subsequent questions, you will be asked to identify if the AI will present "noise uncertainty" on new points. Note that no CLUEs are shown with the questions. Feel free to come back to these context points when answering the questions.

	Person 54		CLUE
AI is uncertain	True	->	False
LSAT	42.0	->	36.8
UGPA	2.6	->	2.9
race	asian		-
sex	female		-

	Person 46
AI is uncertain	False
LSAT	26.0
UGPA	2.8
race	mexican
sex	female

	Person 26		CLUE
AI is uncertain	True	->	False
LSAT	46.0	->	37.9
UGPA	3.1		-
race	black	->	white
sex	male		-

	Person 13
AI is uncertain	False
LSAT	29.0
UGPA	2.3
race	white
sex	male

Will the AI have 'noise uncertainty' for this new point? *

	Person 13
LSAT	33.0
UGPA	3.1
race	mexican
sex	male

- ○ Yes, the AI will be 'noise' uncertain on this point.
- ○ No, the AI will be certain on this point.

Figure 6: A screenshot of a section from the second test variant for LSAT. The top box shows context examples, with CLUEs. The bottom box shows a question asked to the user.

A Small User Study on COMPAS and LSAT

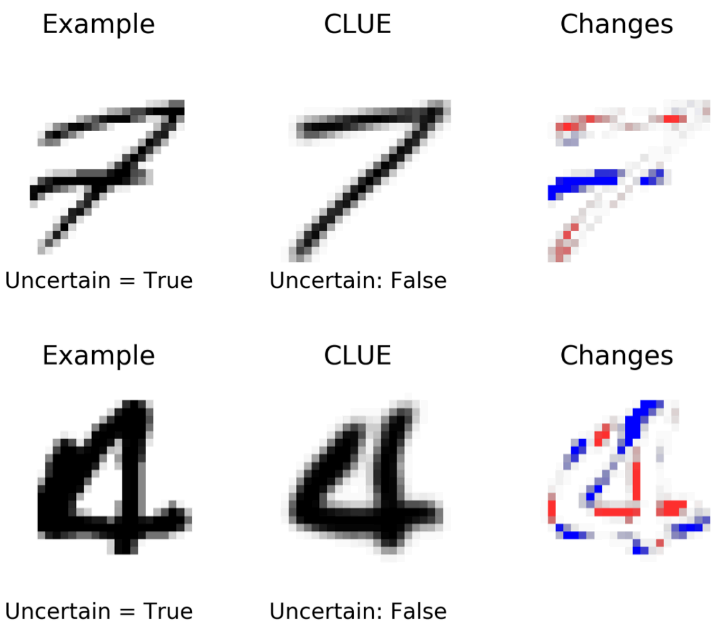
Is CLUE more helpful than just showing uncertainty estimates?

Surveyed	Variant	Sample Size	LSAT Ep. (6)	LSAT Al. (7)	COMPAS Ep. (6)	COMPAS Al. (5)	Total (24)
Prolific	Unc.	10	0.50	0.40	0.53	0.67	0.54
Students	Unc.	8	0.65	0.58	0.56	0.66	0.61
Prolific	CLUE	10	0.60	0.70	0.60	0.40	0.59
Prolific (BS+)	CLUE	9	0.61	0.68	0.54	0.69	0.63
Students	CLUE	7	0.50	0.8	0.67	0.71	0.67

Users are able to predict if a model will be uncertain on new examples more accurately when using CLUE than when shown uncertainty estimates.

A Small User Study on MNIST

We modify the MNIST train set to introduce **O**ut **O**f **D**istribution uncertainty.



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

Summary

- Predictive Uncertainty makes ML systems safer and more reliable
- Sensitivity is not enough to explain Predictive Uncertainty in BNNs
- We introduce CLUE, a method to answer the question:
  "**How should we change an input such that our model produces more certain predictions?**"
- CLUE produces in-distribution explanations which trade-off the amount of change made to inputs and the amount of uncertainty explained away.
- A small user study finds that CLUEs help users understand the sources of a model's uncertainty.

References

- [Antorán et. al., 2020] “Getting a CLUE: A Method for Explaining Uncertainty Estimates”
- [Depeweg et. al., 2017] “Sensitivity Analysis for Predictive Uncertainty in Bayesian Neural Networks”
- [Lundberg et al., 2017] “A Unified Approach to Interpreting Model Predictions”
- [Ribeiro et al., 2017] “"Why Should I Trust You?": Explaining the Predictions of Any Classifier”
- [Sundararajan et al., 2017] “Axiomatic Attribution for Deep Networks”
- [Chang et al., 2017] “Explaining Image Classifiers by Counterfactual Generation”