Getting a CLUE: A Method for Explaining Uncertainty Estimates

Javier Antorán, Umang Bhatt, Tameem Adel, Adrian Weller, José Miguel Hernández-Lobato

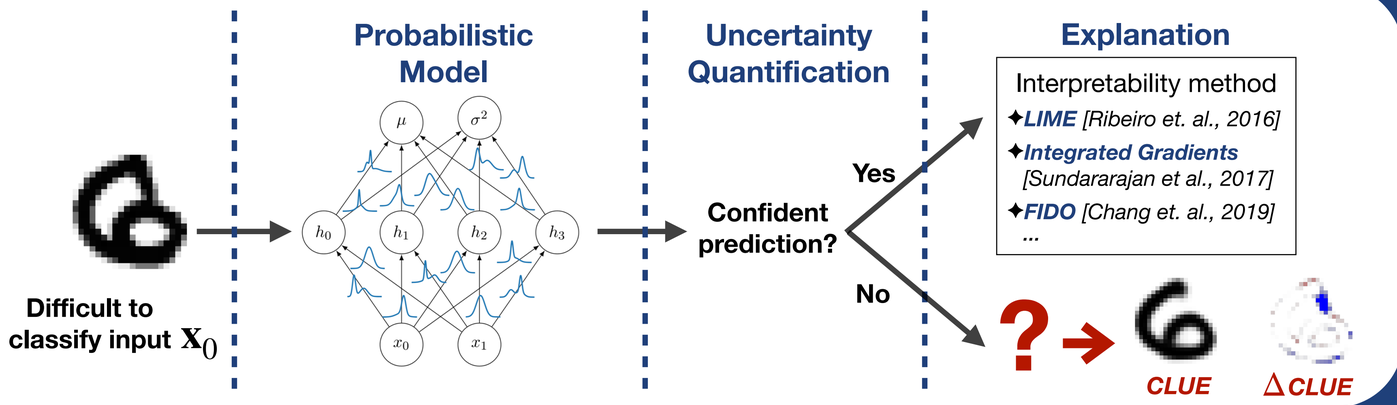
University of Cambridge

Computational and Biological Learning
University of Cambridge

Motivation

- **Critical applications require trustworthy ML**
- **Trustworthy ML requires explainable decisions and calibrated uncertainty**

➔ ***Can we explain why models are uncertain?***

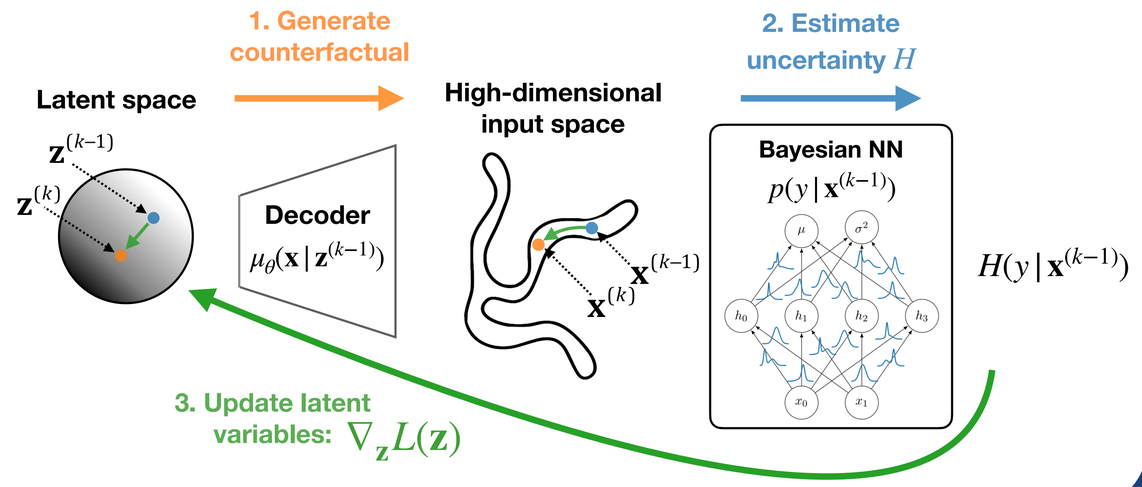


Counterfactual Latent Uncertainty Explanations

"Which input patterns lead my prediction to be uncertain?"

➔ ***CLUE aims to find points in the latent space of a VAE which generate inputs similar to an original observation $\mathbf{x}_0$ but are assigned low uncertainty.***

$$L(\mathbf{z}) = H(y \mid \mu_\theta(\mathbf{x} \mid \mathbf{z})) - d(\mu_\theta(\mathbf{x} \mid \mathbf{z}), \mathbf{x}_0)$$



User Study

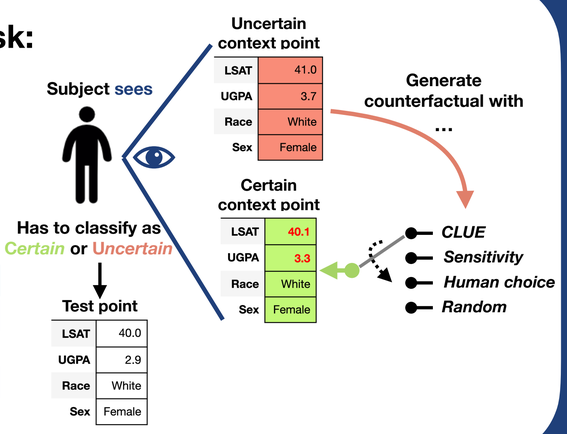
Forward simulation task:

➔ ***Do CLUEs help practitioners predict when their model will be uncertain?***

2 Datasets, 9 questions per dataset:

- **COMPAS** (Criminal recidivism prediction, 7 dim)
- **LAST** (Academic performance prediction, 4 dim)

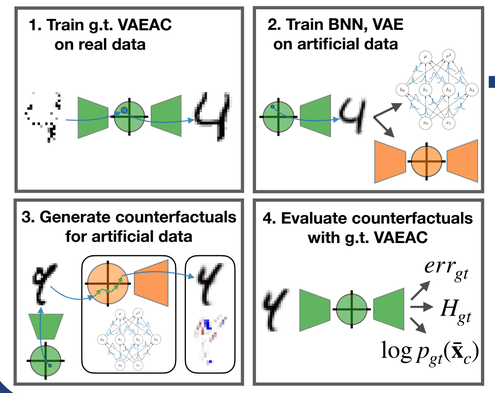
Method	N. participants	Accuracy (%)	Rank p-value
Random	10	61.67	1.47E-05
Sensitivity	10	52.78	2.60E-09
Human	10	62.22	2.34E-05
CLUE	10	**82.22**	-

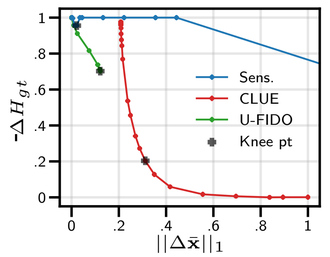


Computational Evaluation

We want to asses if uncertainty is reduced and if counterfactuals are plausible

➔ ***Use conditional VAE (VAEAC) to generate synthetic data: generative process is known***





CLUE provides a flexible tradeoff between plausibility

$$||\Delta\mathbf{x}||_1 = ||\mathbf{x}_{CLUE} - \mathbf{x}_0||_1$$

and uncertainty reduction

$$\Delta H = H(y \mid \mathbf{x}_0) - H(y \mid \mathbf{x}_{CLUE})$$