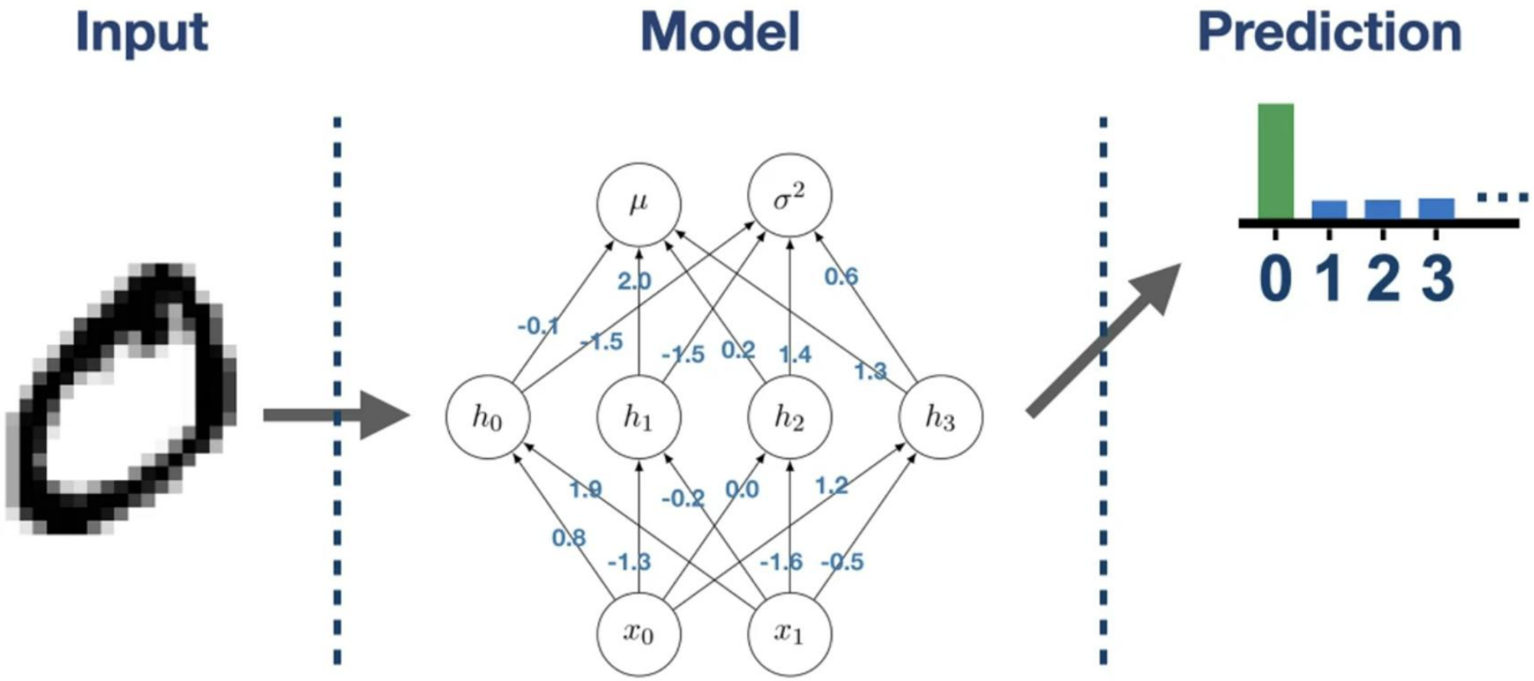
GETTING A CLUE

TANVI SHETTY

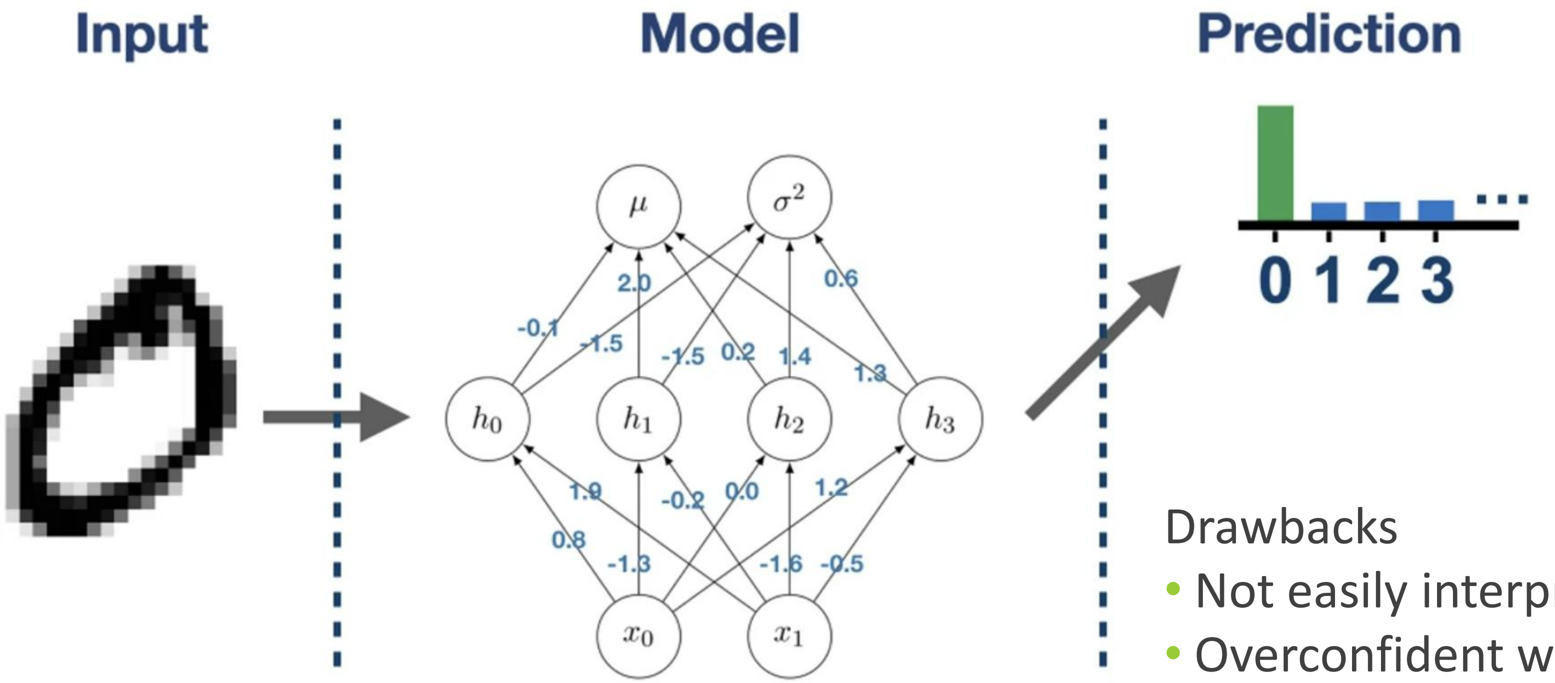
SUPERVISOR: JELLE HÜNTELMANN

Neural Networks



Source: https://slideslive.com/38953876

Neural Networks



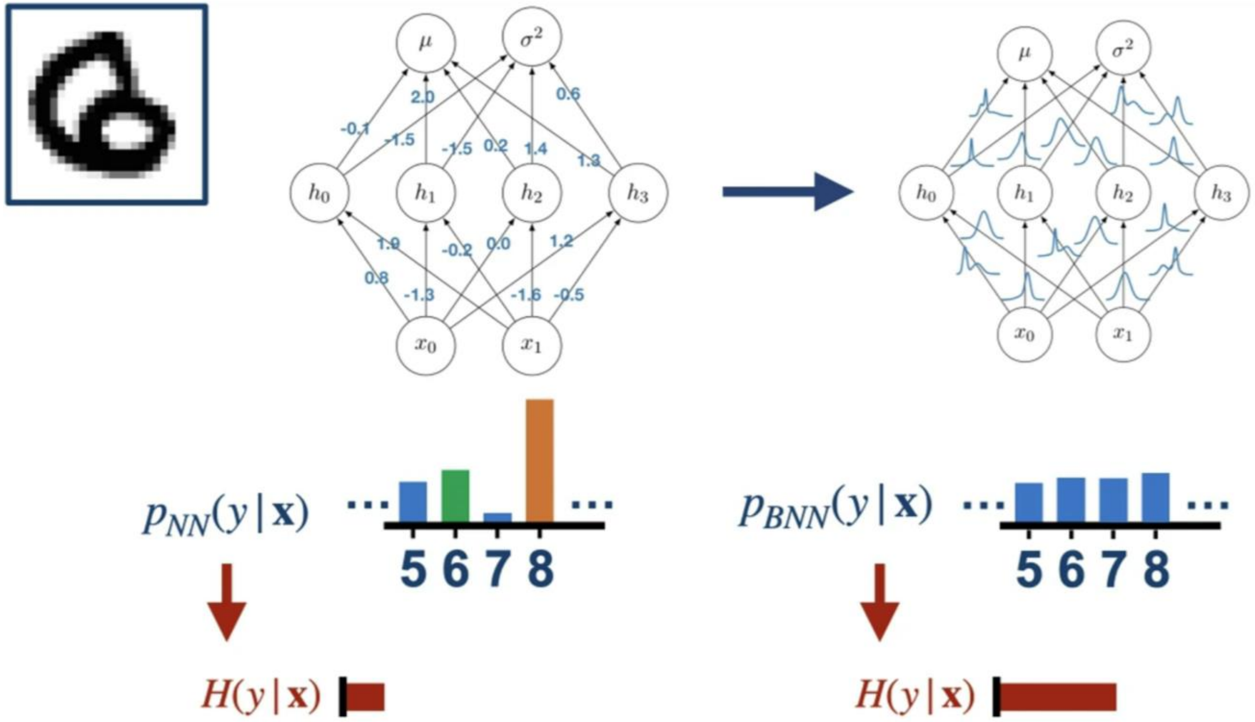
Drawbacks

- Not easily interpretable
- Overconfident wrong predictions

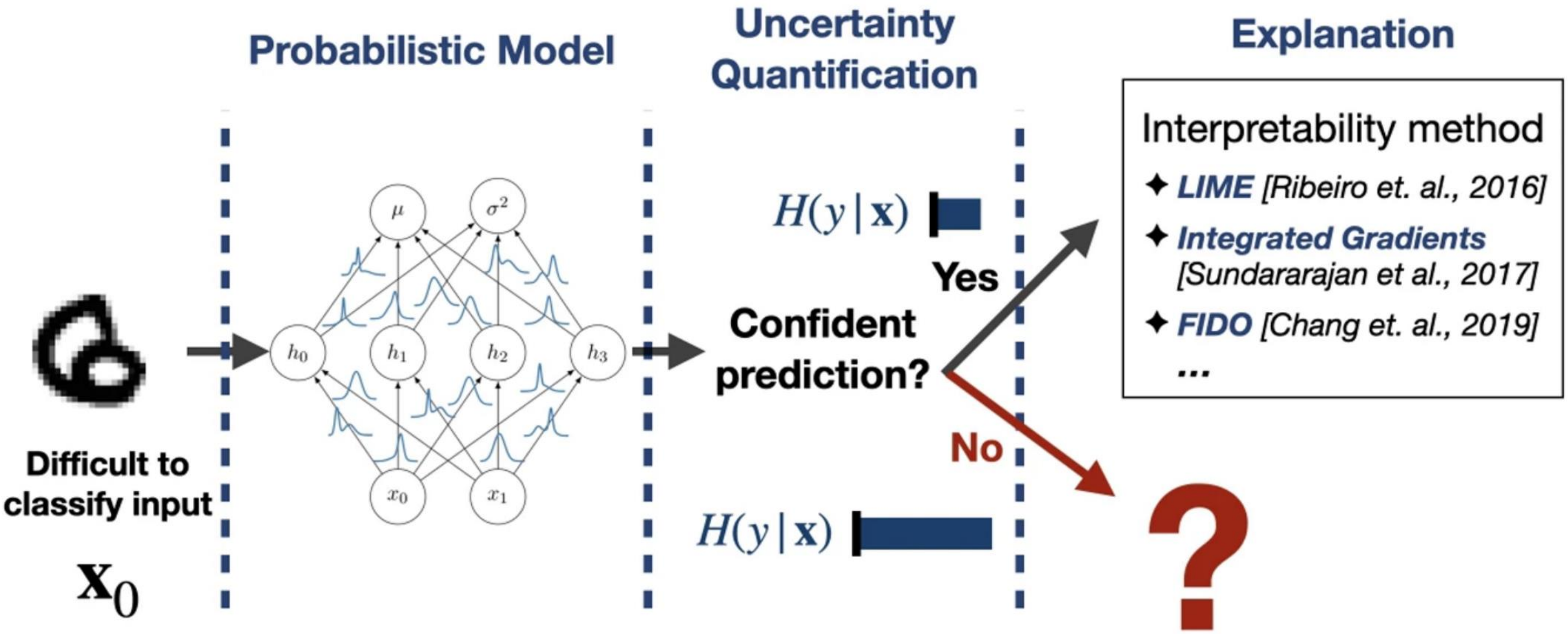
Source: https://slideslive.com/38953876

How to avoid overconfident predictions?

How to capture uncertainty?



Source: https://slideslive.com/38953876



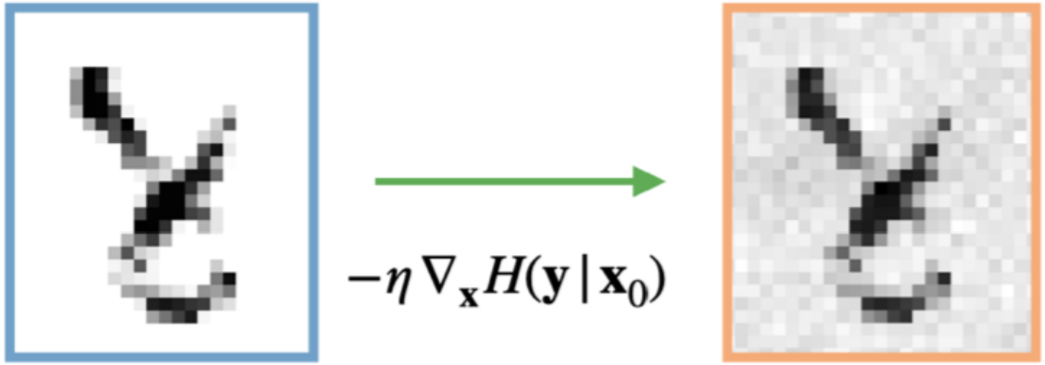
Interpretability method

- ✦ *LIME [Ribeiro et. al., 2016]*
- ✦ *Integrated Gradients [Sundararajan et al., 2017]*
- ✦ *FIDO [Chang et. al., 2019]*
- ...

Why were the predictions uncertain?

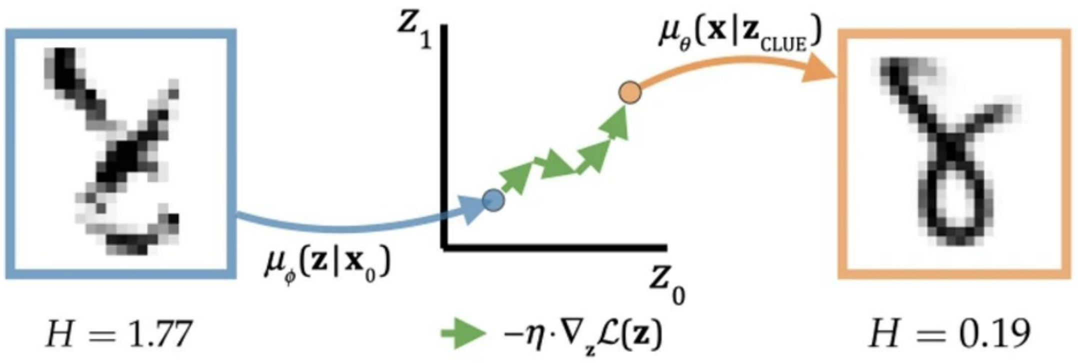
Uncertainty sensitivity analysis

- Does not scale well with high dimensional data.

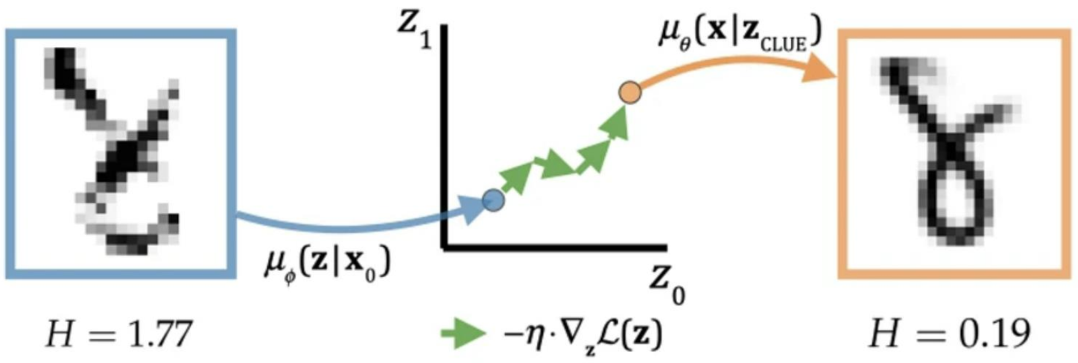

$-\eta \nabla_{\mathbf{x}} H(\mathbf{y} \,|\, \mathbf{x}_0)$

How do we overcome this?

Main Idea

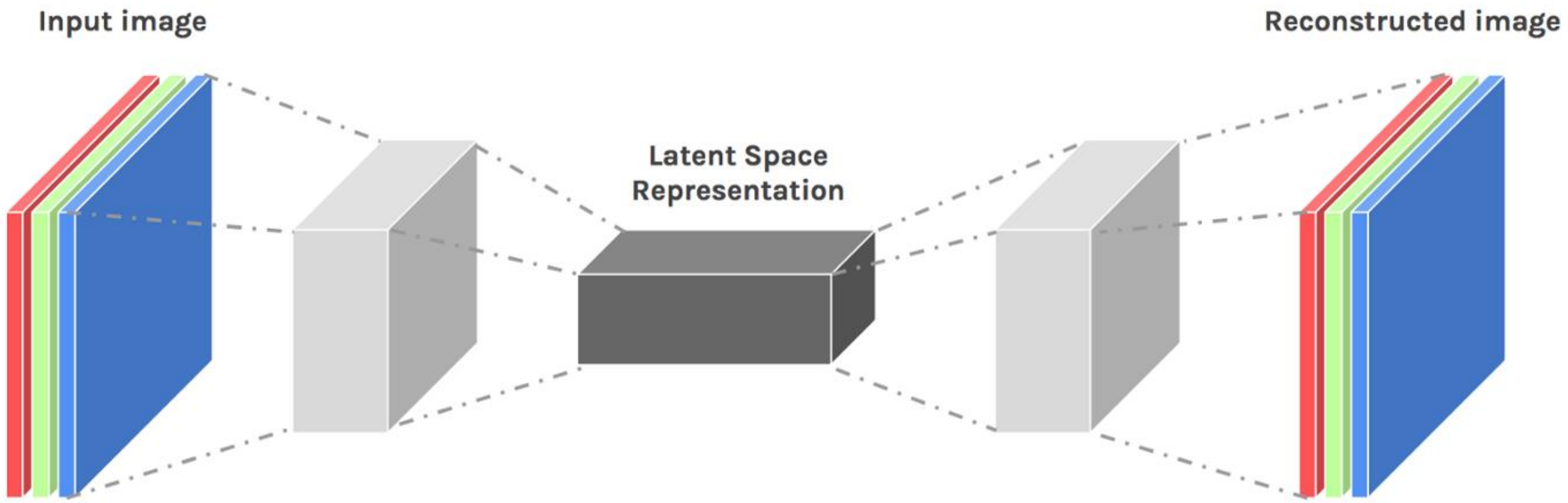

$\mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}})$
z_1
$\mu_\phi(\mathbf{z}|\mathbf{x}_0)$
z_0
$H = 1.77$
$-\eta \cdot \nabla_{\mathbf{z}} \mathcal{L}(\mathbf{z})$
$H = 0.19$

Main Idea


z_1
$\mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}})$
$\mu_\phi(\mathbf{z}|\mathbf{x}_0)$
z_0
$H = 1.77$
$-\eta \cdot \nabla_z \mathcal{L}(\mathbf{z})$
$H = 0.19$

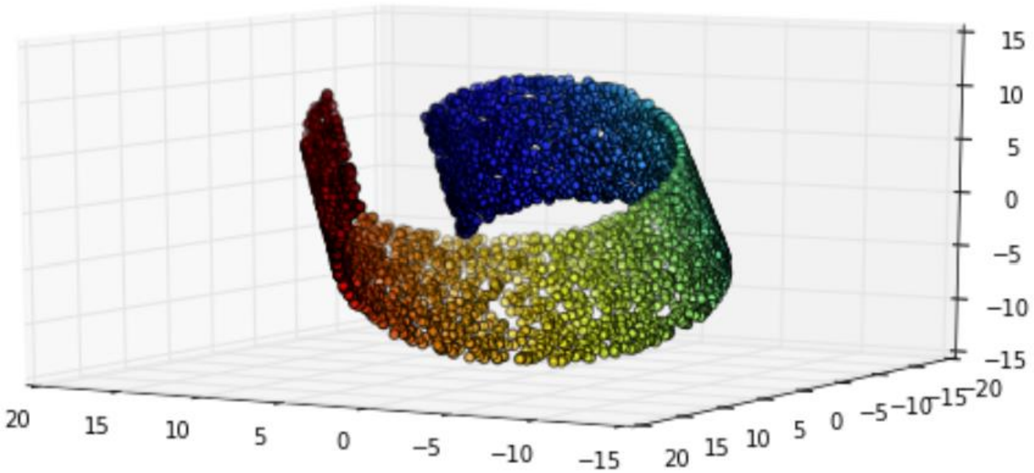
- Encode our input to a latent space.
- Perform some optimization that aims to minimize uncertainty.
- Decode into some resulting input for which our model is more certain.

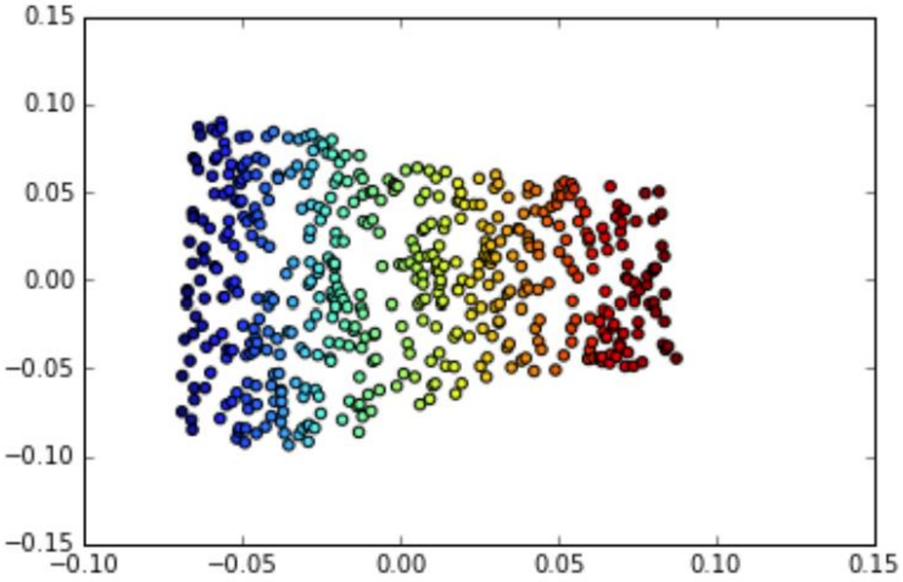
Latent space



Source: https://hackernoon.com/latent-space-visualization-deep-learning-bits-2-bd09a46920df

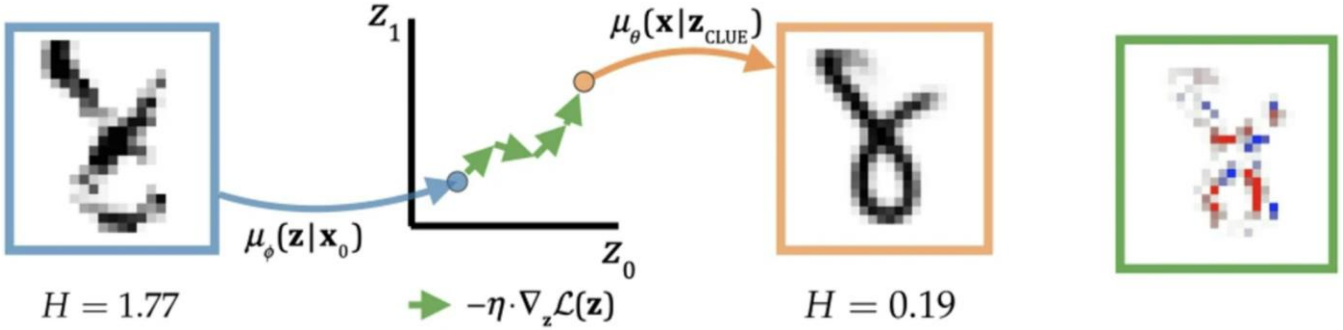
Manifold





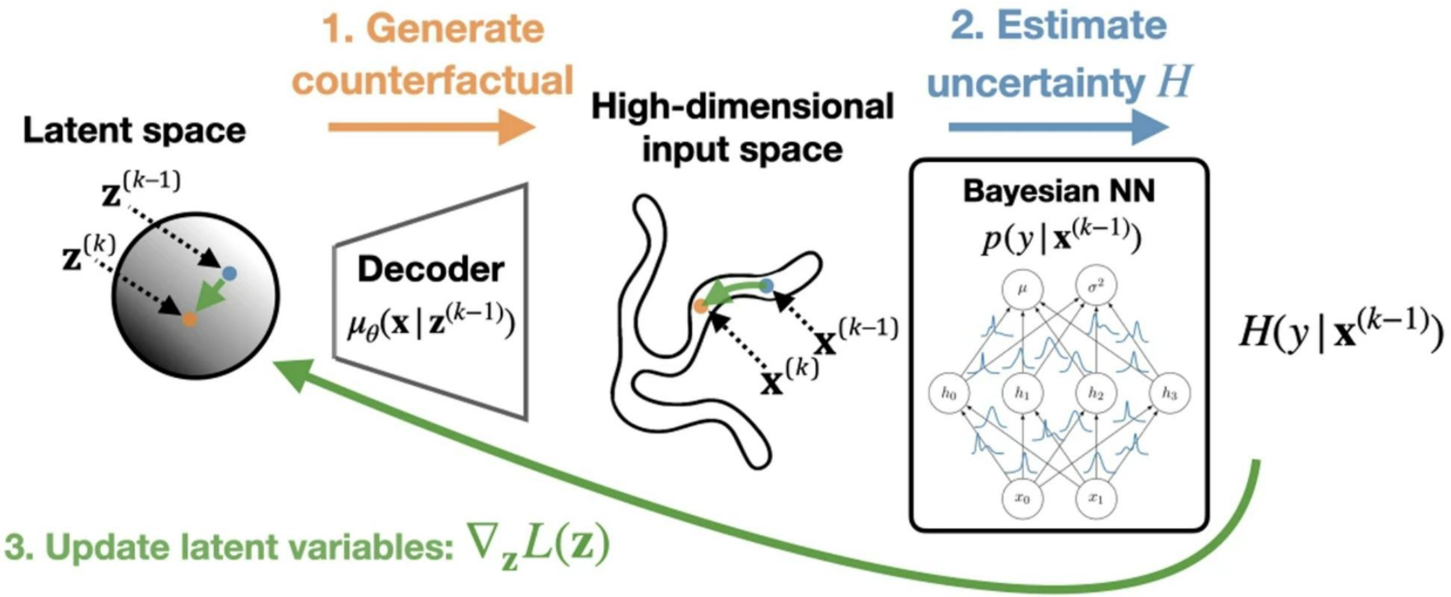
Source: https://datascience.stackexchange.com/a/5698

Main Idea


z_1
$\mu_\theta(\mathbf{x}|\mathbf{z}_{\text{CLUE}})$
$\mu_\phi(\mathbf{z}|\mathbf{x}_0)$
z_0
$H = 1.77$
$-\eta\cdot\nabla_z\mathcal{L}(\mathbf{z})$
$H = 0.19$

- Encode our input to a latent space.
- Perform some optimization that aims to minimize uncertainty.
- Decode into some resulting input for which our model is more certain.

Let’s get a CLUE

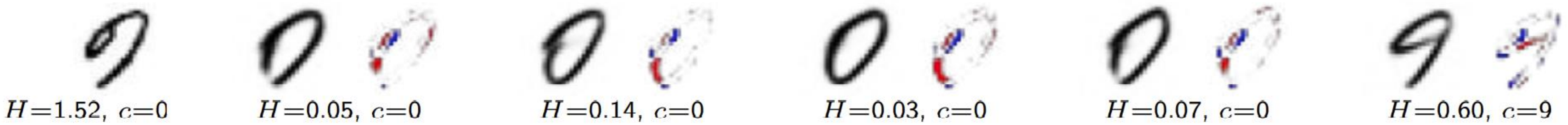

1. Generate counterfactual
2. Estimate uncertainty H
Latent space
$\mathbf{z}^{(k-1)}$
$\mathbf{z}^{(k)}$
Decoder
$\mu_\theta(\mathbf{x} \mid \mathbf{z}^{(k-1)})$
High-dimensional input space
$\mathbf{x}^{(k-1)}$
$\mathbf{x}^{(k)}$
Bayesian NN
$p(y \mid \mathbf{x}^{(k-1)})$
μ
σ^2
h_0
h_1
h_2
h_3
x_0
x_1
$H(y \mid \mathbf{x}^{(k-1)})$
3. Update latent variables: $\nabla_{\mathbf{z}} L(\mathbf{z})$

Optimizing the objective

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

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

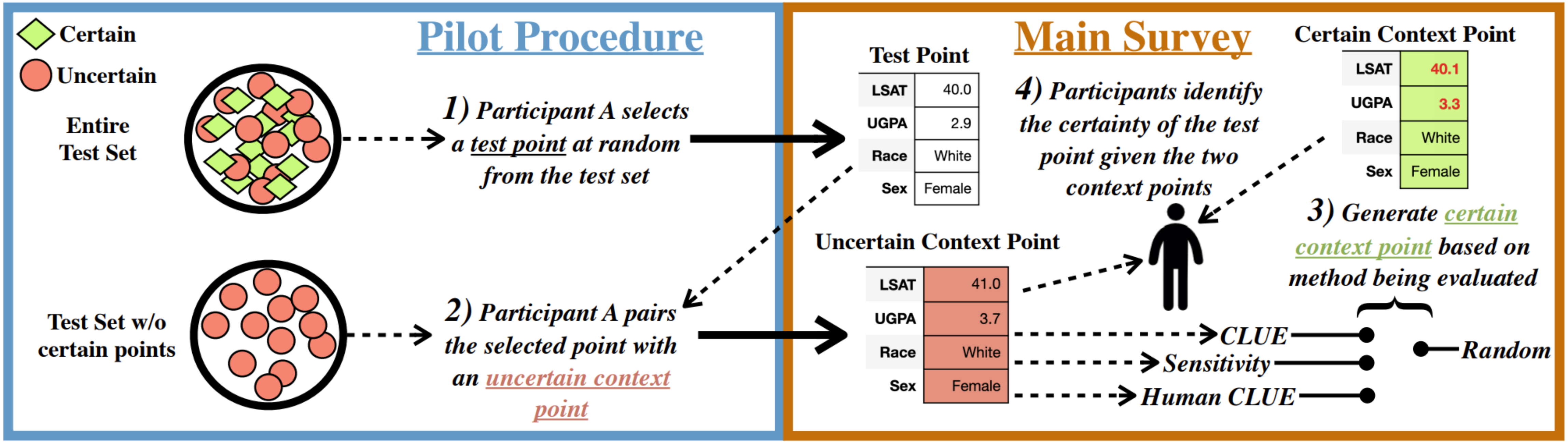
Multiplicity of CLUEs



User Study

1. Show the subject some uncertain context point.
2. Generate a counterfactual explanation for the context point.
3. Show the subject the generated certain context point using either CLUE, Uncertainty Sensitivity, Human choice, Random.
4. The user to has classify a new unseen point as certain or uncertain.

User Study



Pilot Procedure

◇ Certain

● Uncertain

Entire Test Set

Test Set w/o certain points

1) Participant A selects a test point at random from the test set

2) Participant A pairs the selected point with an uncertain context point

Main Survey

Test Point

LSAT	40.0
UGPA	2.9
Race	White
Sex	Female

Certain Context Point

LSAT	40.1
UGPA	3.3
Race	White
Sex	Female

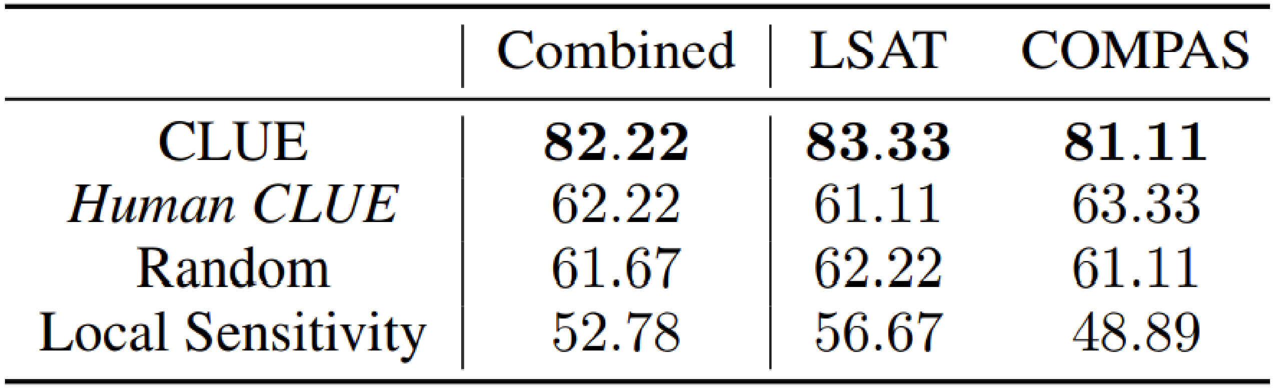
Uncertain Context Point

LSAT	41.0
UGPA	3.7
Race	White
Sex	Female

4) Participants identify the certainty of the test point given the two context points

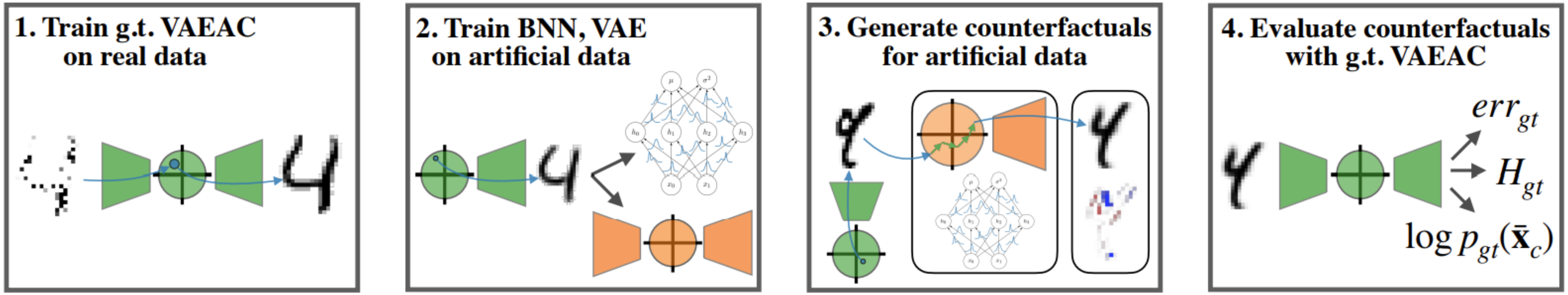
3) Generate certain context point based on method being evaluated

User Study: Results



	Combined	LSAT	COMPAS
CLUE	**82.22**	**83.33**	**81.11**
Human CLUE	62.22	61.11	63.33
Random	61.67	62.22	61.11
Local Sensitivity	52.78	56.67	48.89

Evaluate Counterfactual Explanations


1. Train g.t. VAEAC on real data
2. Train BNN, VAE on artificial data
3. Generate counterfactuals for artificial data
4. Evaluate counterfactuals with g.t. VAEAC
err_{gt}
H_{gt}
$\log p_{gt}(\bar{\mathbf{x}}_c)$

“What is the smallest change that could be made to an input,

while keeping it in distribution,

so that our model becomes certain in its decision for said input?”

Thank you!