ML4Seismic Partners Meeting 2023

Interventionist Uncertainty in Neural Networks: A Case Study in Prompting

Mohit Prabhushankar, Prithwijit Chowdhury, Mohammad Alotaibi, and Ghassan AlRegib

Prompts allow extracting contextual and relevant information from the model

Segmentation without Prompting

All objects segmented

Segmentation with Prompt

Manual prompting selects only one segment

A naïve view of Prompts: Remove irrelevant (as defined by interpreters) data from input

Segmentation without Prompting

All objects segmented

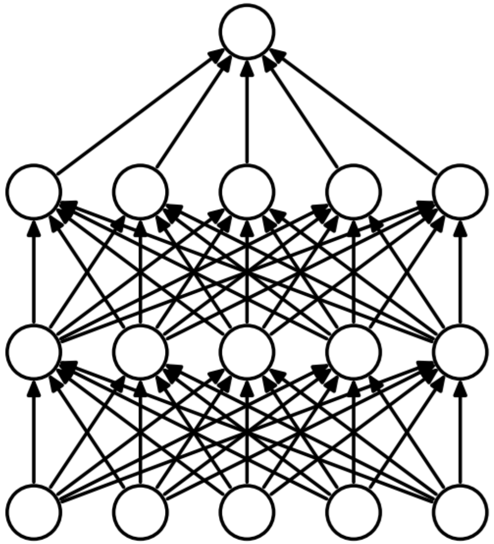
Segmentation with Prompt

Manual prompting selects only one segment

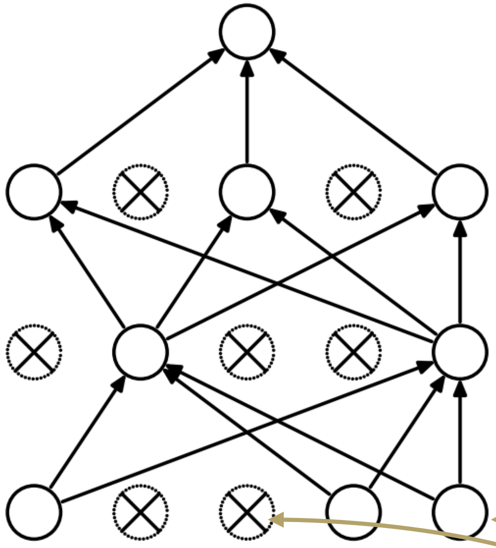
$$x \rightarrow \{x, P\} \rightarrow S_x$$

Prompting at Inference
Extracting Information from Models

A naïve view of Prompts: Remove irrelevant (as defined by interpreters) <u>weights</u> from model



All objects segmented



Manual prompting selects only one segment

Segmentation with Prompt

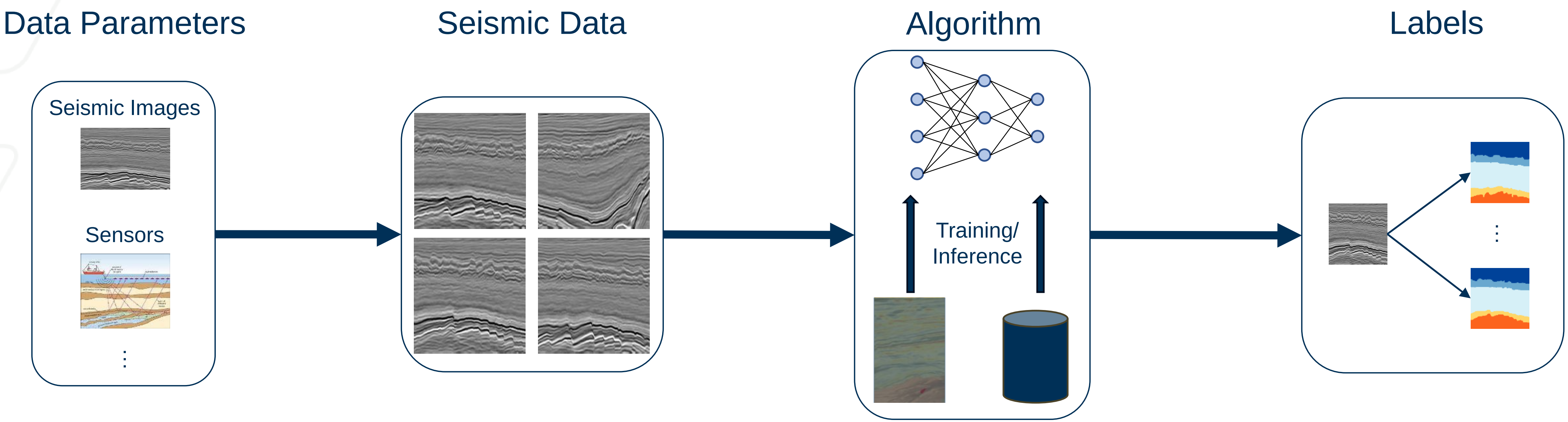
Monte-Carlo Dropout

$$W \rightarrow \{W, P\} \rightarrow W_P$$

Model Uncertainty

Objective: To motivate and quantify Prompts as Uncertainty

Existing Uncertainty Framework: Flow of Information in <u>one direction</u>



Data — Images — Interpretations

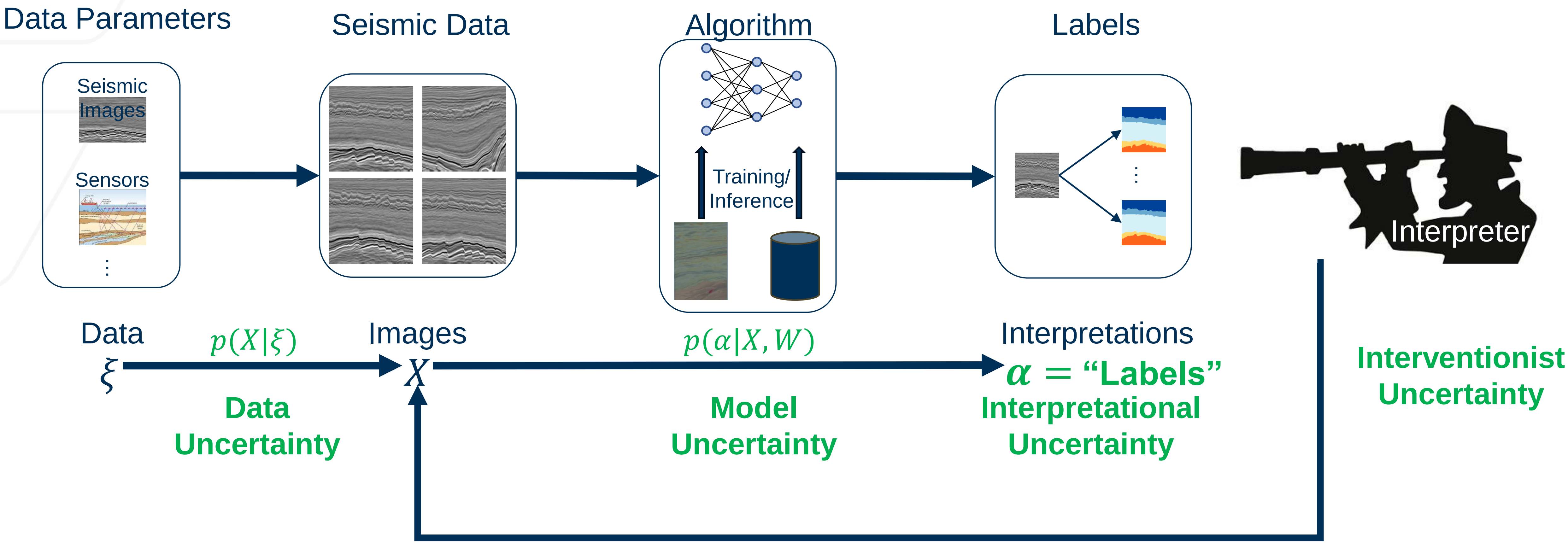
$\xi \xrightarrow{p(X|\xi)} X \xrightarrow{p(\alpha|X,W)} \alpha =$ **"Labels"**

Data Uncertainty | **Model Uncertainty** | **Interpretational Uncertainty**

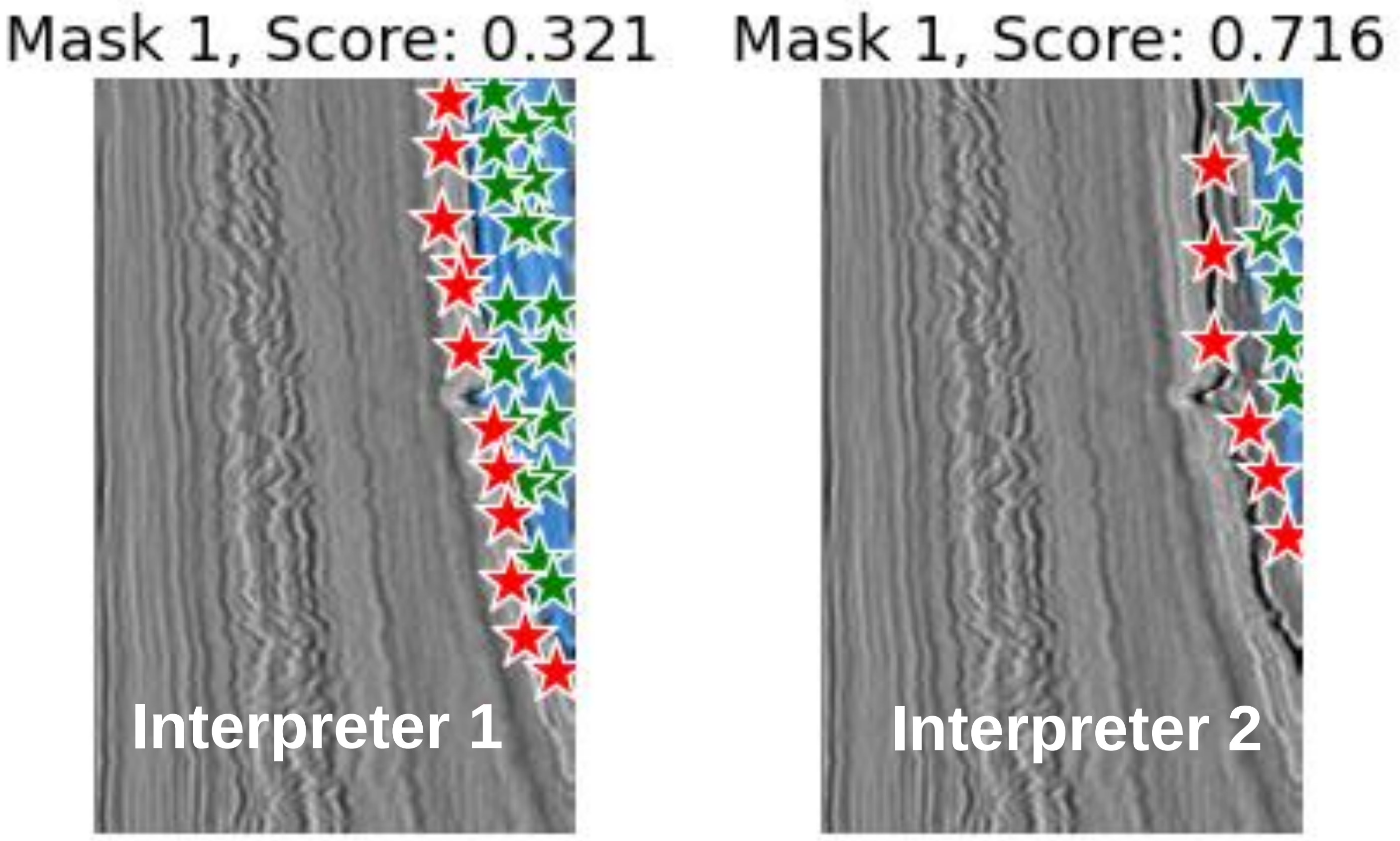
Uncertainty in Deep Learning

Where do Prompts fit in?

Prompts at Inference: Flow of Information is a Loop



Analyzing Prompting: Via objective mean Intersection Over Union (mIOU)



Interpreter 2 > Interpreter 1

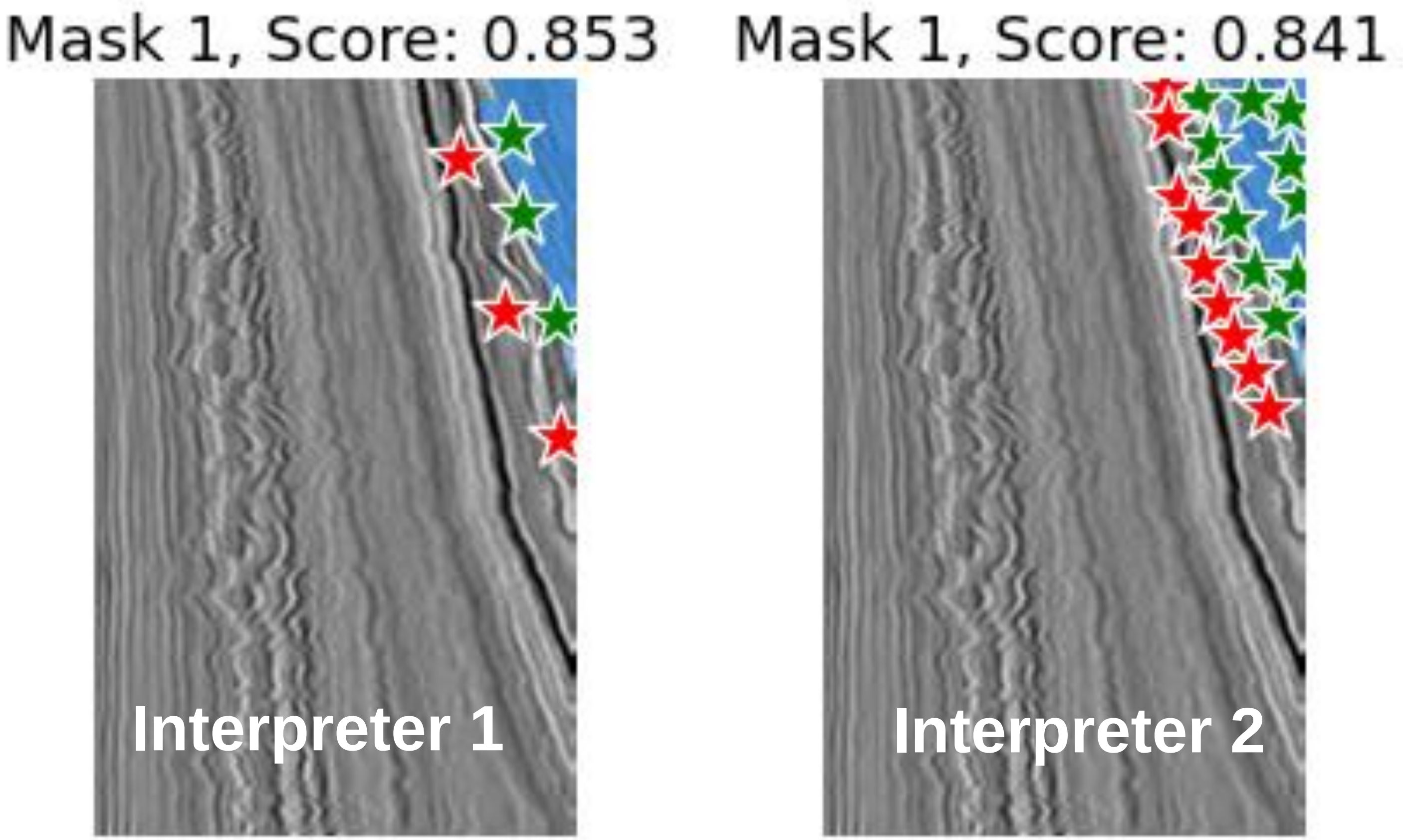
- **Goal:** To delineate region of interest
- **Quantifiable score:** Mean Intersection over Union (mIOU) between prediction and ground truth

$$E(y|S_{x2}) > E(y|S_{x1})$$

Uncertainty in Prompting

Analysis Framework

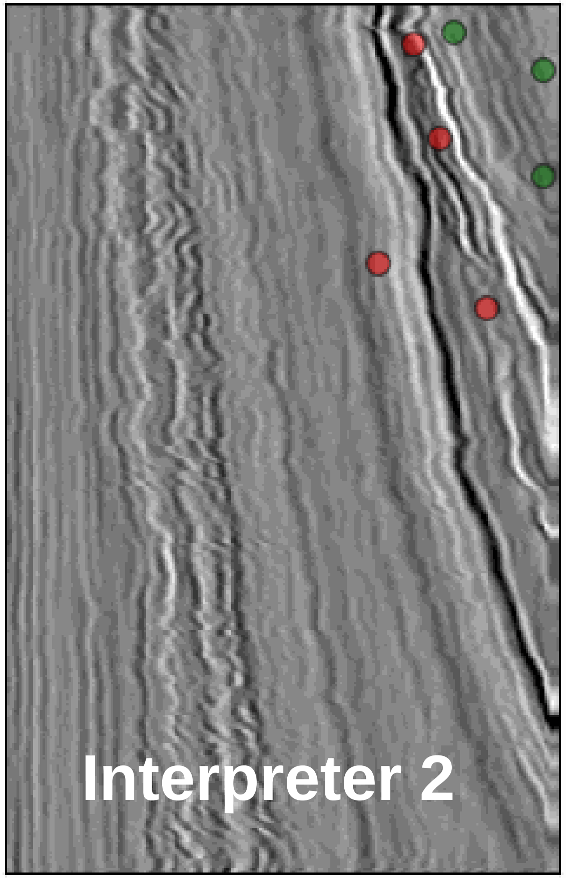
Analyzing Interventionist Uncertainty: Via objective output metrics



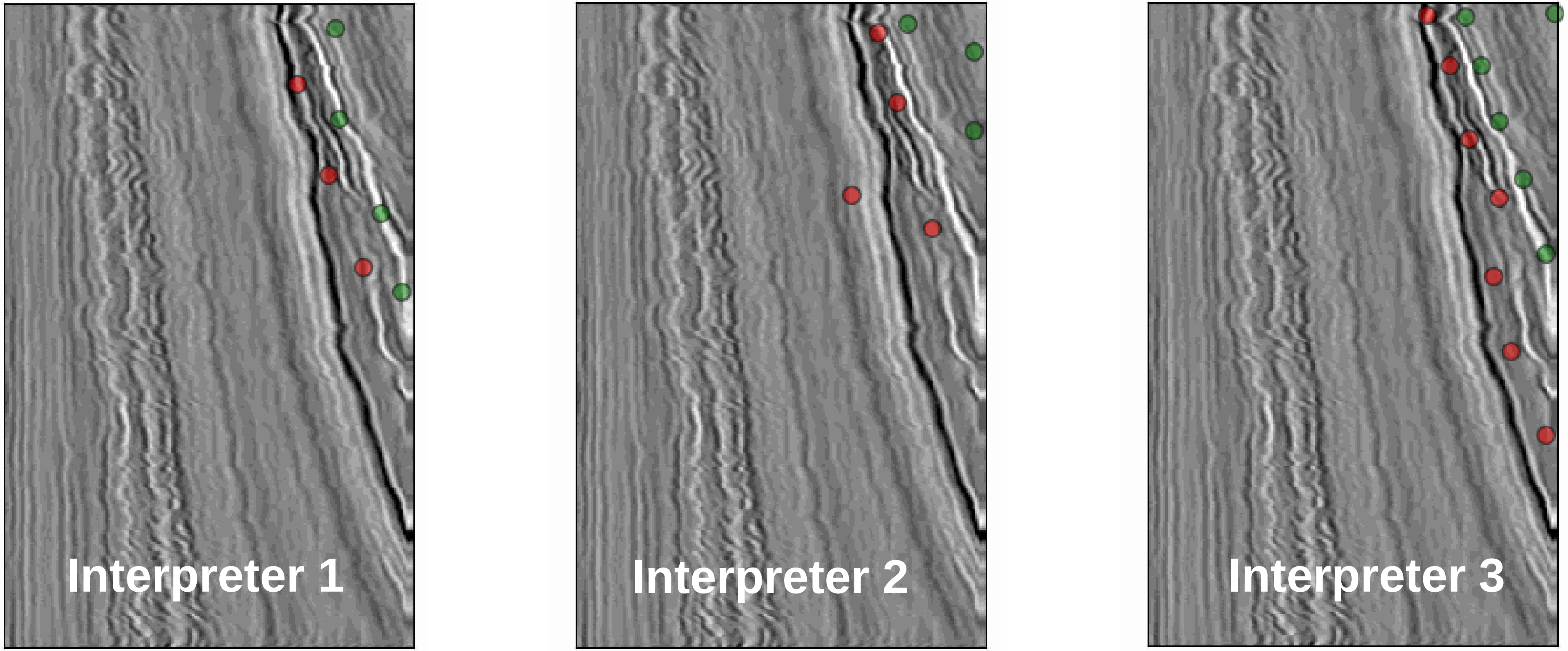
Interpreter 2 ? Interpreter 1

- **Goal:** To delineate region of interest
- **Quantifiable score:** Mean Intersection over Union (mIOU) between prediction and ground truth
  1. **mIOU** between the two interpreters **is the same**
  2. However, the prompts (**numbers and locations**) are different

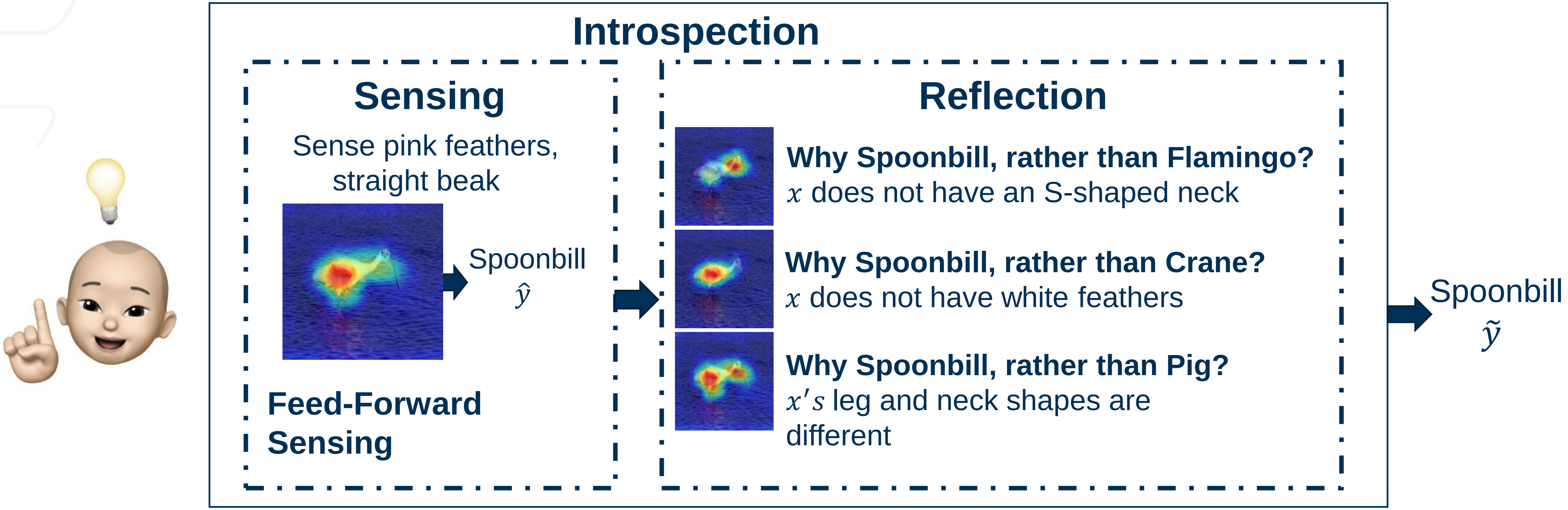
Prompting across Sections: Notice the change in prompts in consecutive sections



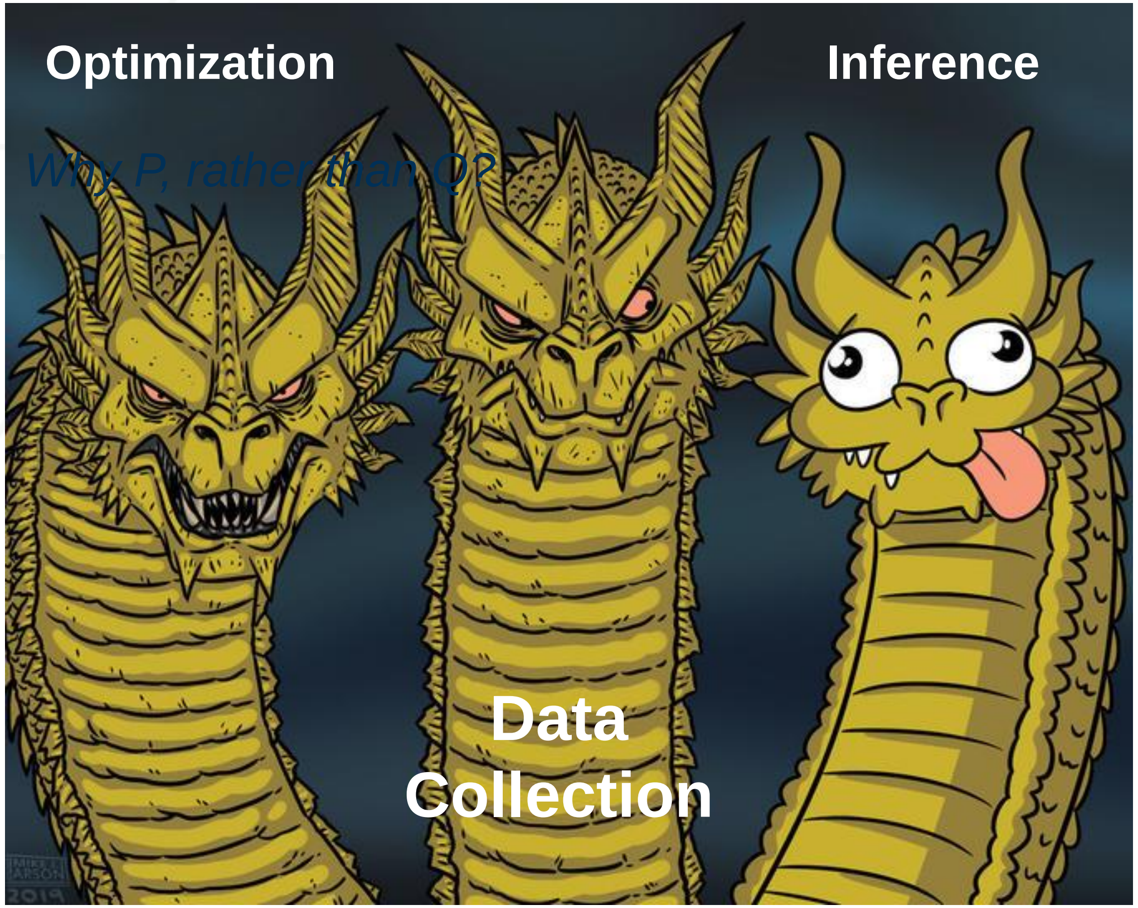
Prompting across Interpreters: Notice the change in prompts between interpreters

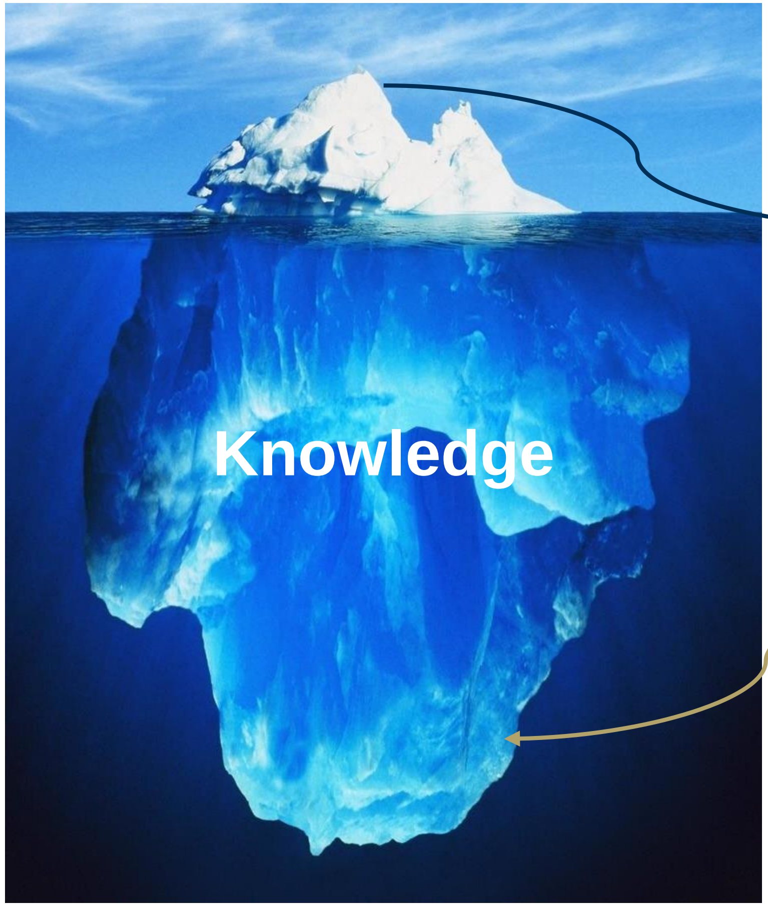


Introspective Learning using Contrastive Questions: Uncertainty resides in contrastive questions

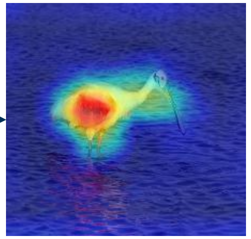


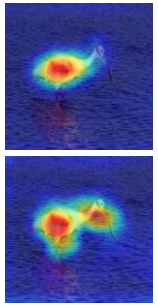
Trained Neural Nets have hidden knowledge. Goal is to prompt it at Inference.





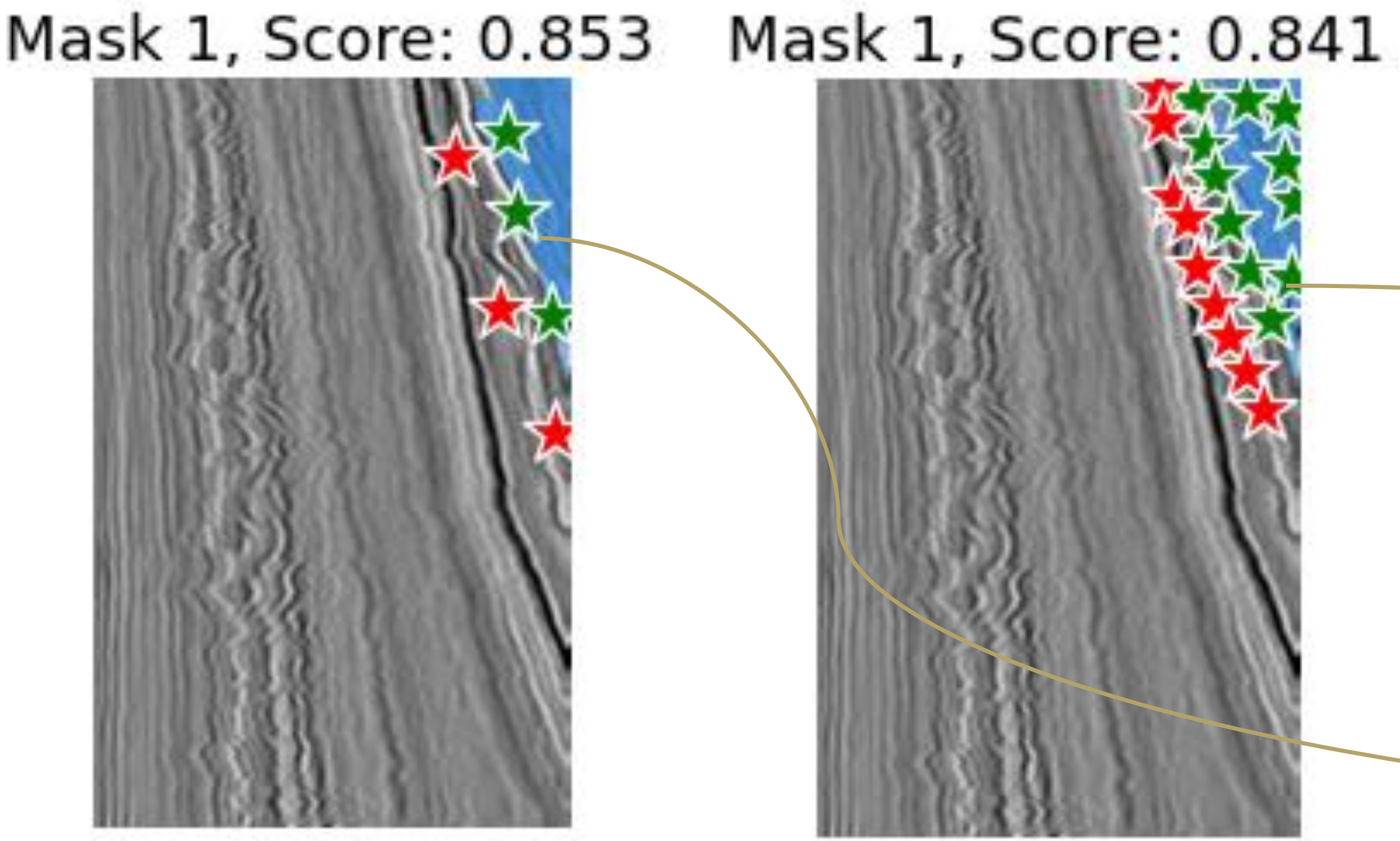
Traditional *Why P?*





What if? Why P, rather than Q?

Our Analysis: Interventionist Uncertainty via Contrastive Questions



Why Prompt P_1, rather than P_2?

Variance Decomposition of Uncertainty under intervention

$$V[y|S_x] = V[E(y|S_x)] + E(V[(y|S_x)])$$

y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
S_x = Intervened Data
$E(Y|S_x)$ = Expectation of class under intervention
$V(Y|S_x)$ = Variance of class under all residuals

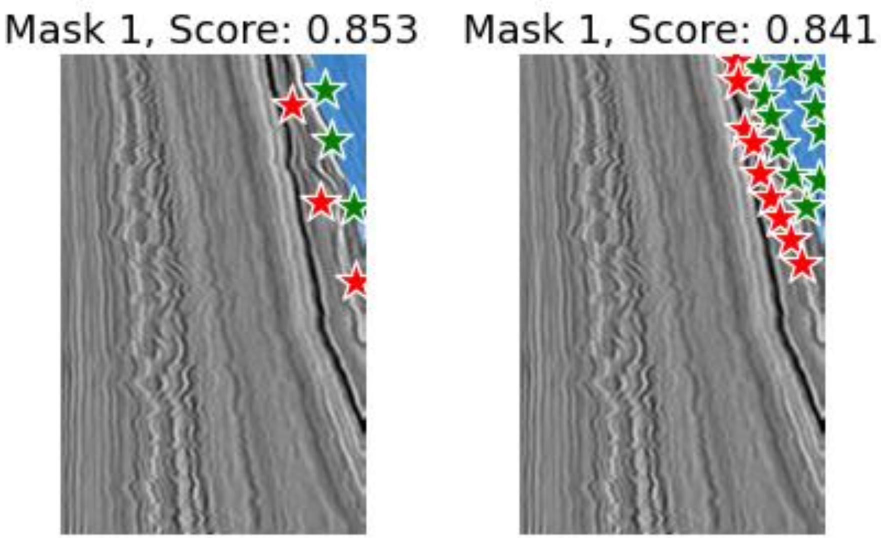
Uncertainty in Prompting

Analysis under Contrastive Framework

Variance Decomposition of Uncertainty under intervention

$$V[y|S_x] = V[E(y|S_x)] + E(V[(y|S_x)])$$

zero



$$E(y|S_{x2}) = E(y|S_{x1})$$

y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
$\mathrm{S_x}$ = Intervened Data
$\mathrm{E(Y|S_x)}$ = Expectation of class under intervention
$\mathrm{V(Y|S_x)}$ = Variance of class under all residuals

Variance Decomposition of Uncertainty under intervention

$$V[y|S_x] = V[E(y|S_x)] + E(V[(y|S_x)])$$

Given an intervention S_x, find alternative interventions $S_x{}'$ that result in non-zero $V[(y|S_x)]$

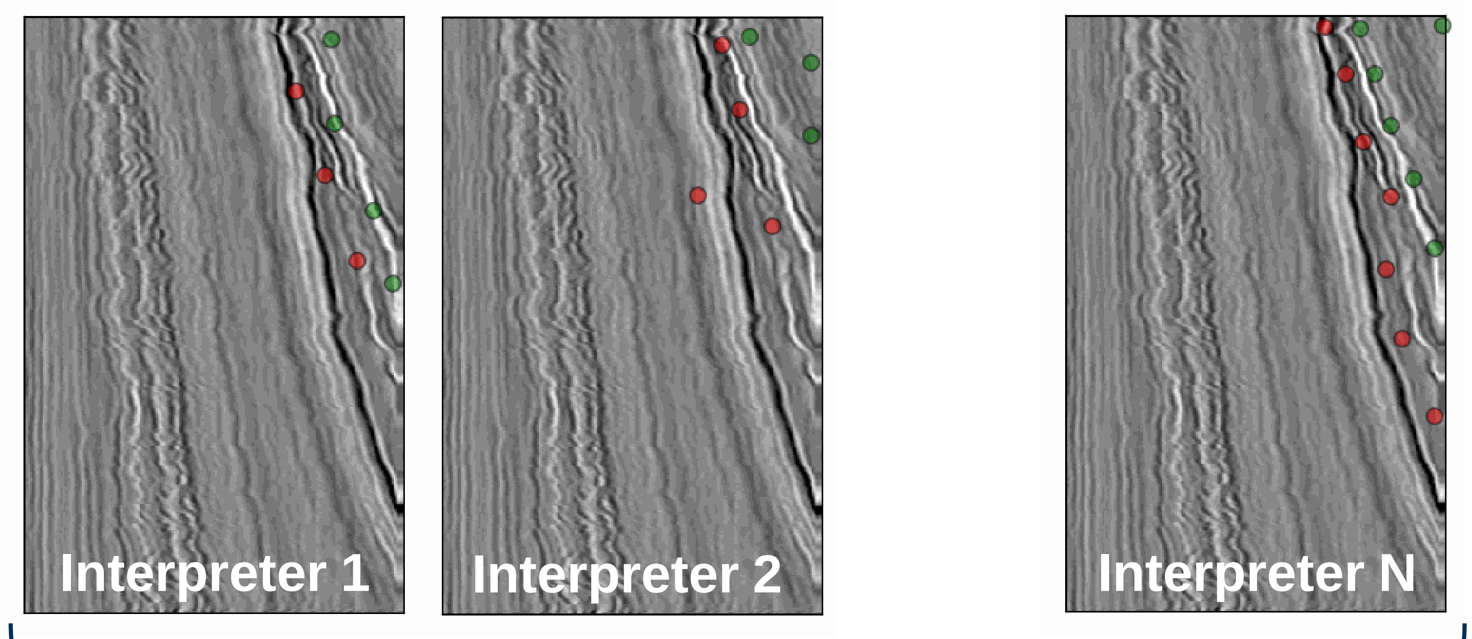
y = Prediction
$V[y]$ = Variance of prediction (Predictive Uncertainty)
$\mathrm{S_x}$ = Intervened Data
$\mathrm{E(Y|S_x)}$ = Expectation of class under intervention
$\mathrm{V(Y|S_x)}$ = Variance of class under all residuals

alternatives

Uncertainty in Prompting
Methodology

Take variance across outputs derived from N prompts

$$V[y|S_x] = V[E(y|S_x)] + E(V[(y|S_x)])$$



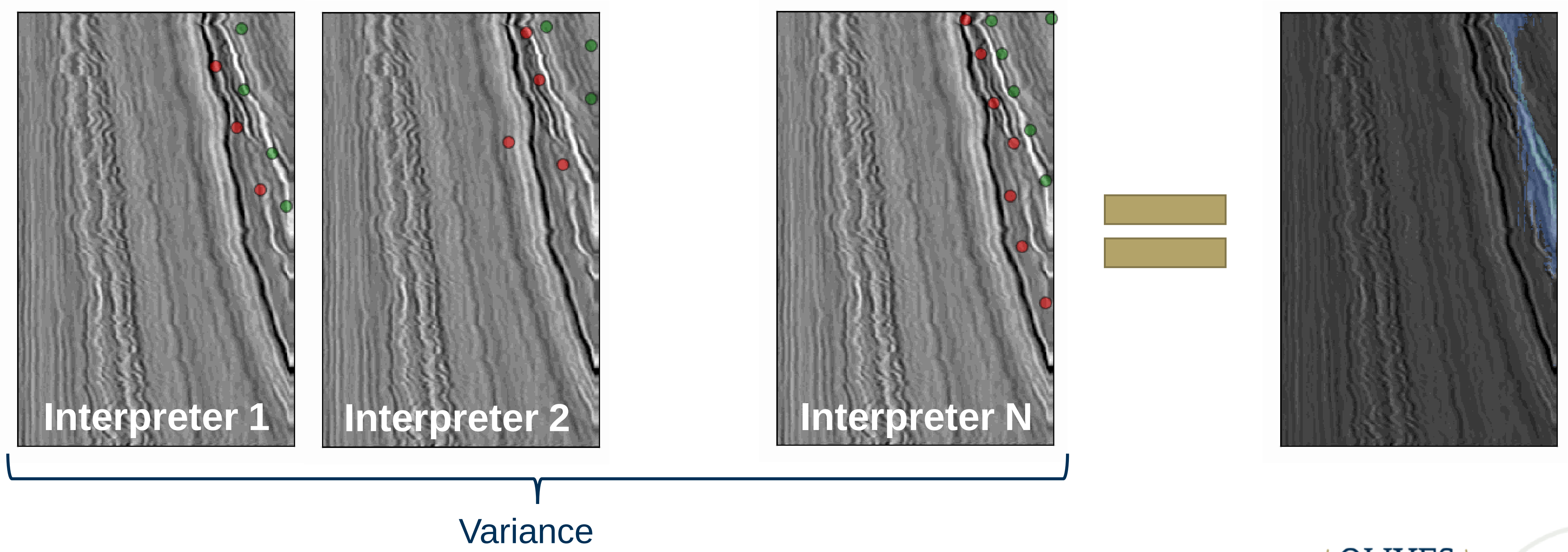
Variance

$$\left.\begin{array}{c} Y_1|S_{x1} \\ Y_2|S_{x2} \\ Y_3|S_{x3} \\ Y_4|S_{x4} \\ Y_5|S_{x5} \\ . \\ . \\ Y_N|S_{xN} \end{array}\right\} \text{Variance}$$

Uncertainty in Prompting
Visual Results

Uncertainty resides near the prompt boundaries

$$V[y|S_x] = V[E(y|S_x)] + E(V[(y|S_x)])$$



Benefits of Analyzing Uncertainty

How to obtain the best prompts?

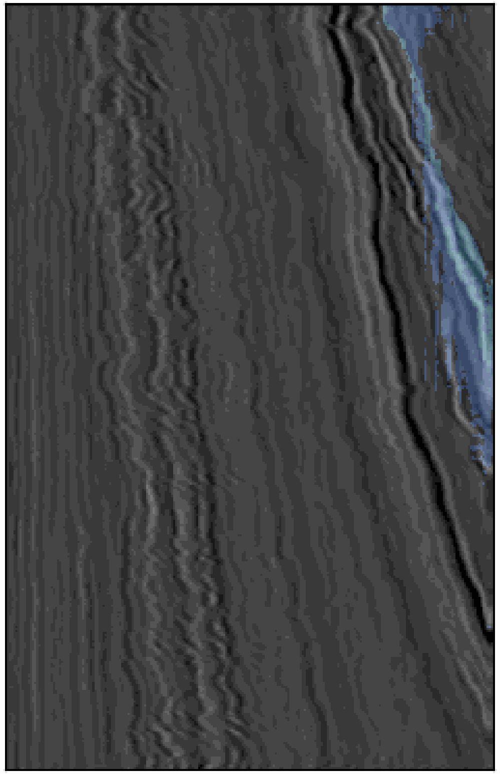
Goal: Among given 7 prompts, choose the best prompts for each section

Best Prompts = **Highest mIOU** against ground truth

Not available at prompting

Best Guess Prompts = Highest mIOU against **uncertainty**

Accuracy(Best prompts, Best Guess Prompts) = 34.66%
(random = 14%)



Benefits of Analyzing Uncertainty

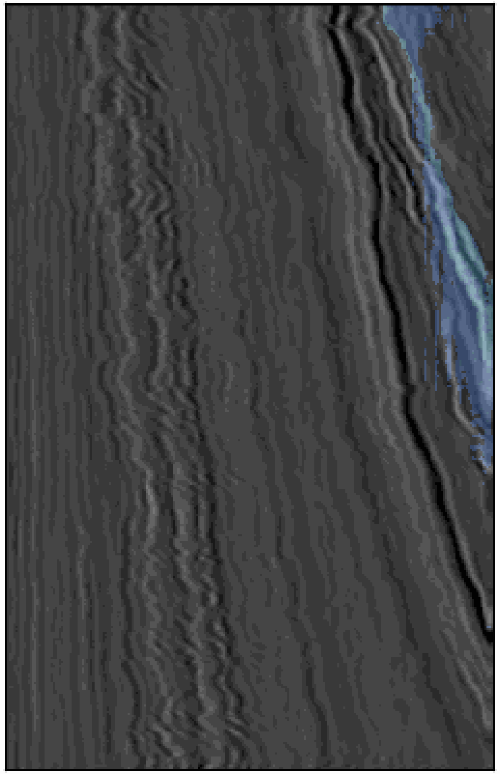
How to estimate Intersection over Union (IoU) without access to Ground Truth?

Signal-to-Noise Ratio (SNR) of interventional uncertainty follows IoU

Mean of uncertainty map

$$SNR = \frac{\mu(V(y|S_x))}{\sigma(V(y|S_x))}$$

Standard deviation of uncertainty map



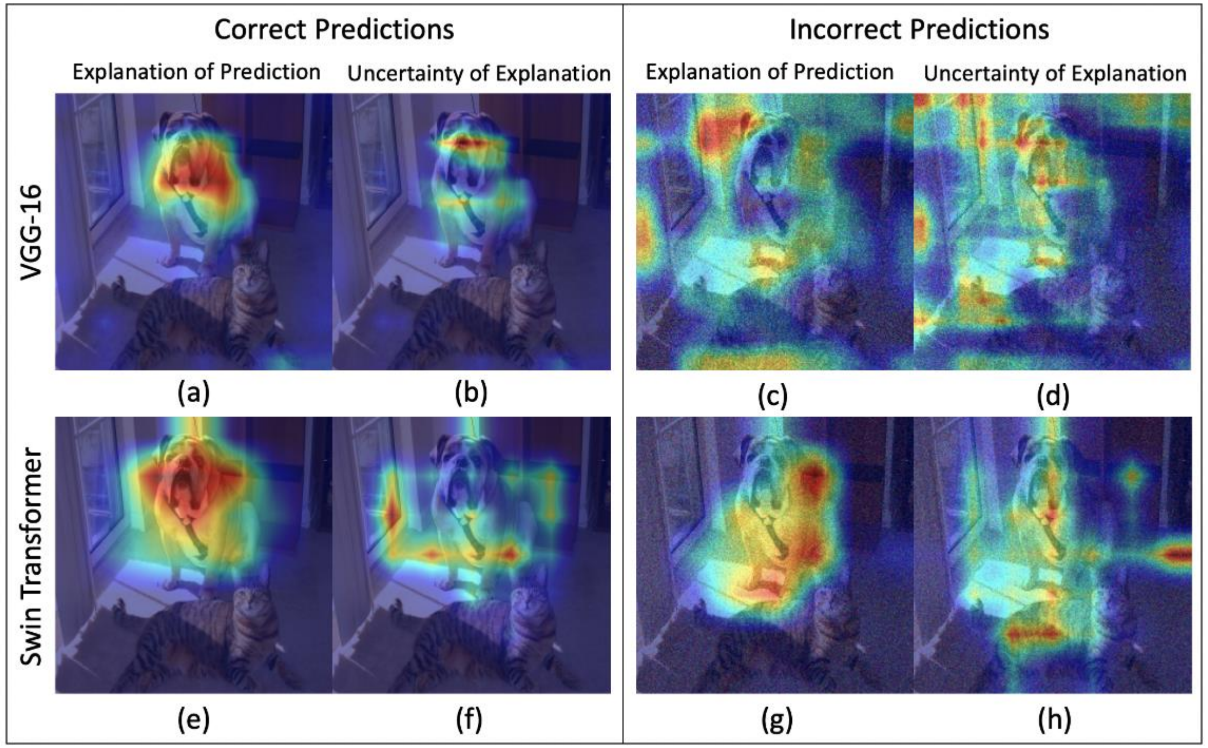
Cosine Similarity(SNR, IoU(Best prompt, GT)) = 0.83

Even without knowing ground truth (GT), we can estimate how well the best prompt will perform

To analyze and quantify Visual Explainability

- **Conclusion 1: Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the explanation)**
- **Conclusion 2: Higher the SNR of uncertainty, more is the dispersal (or less trustworthy is the explanation)**

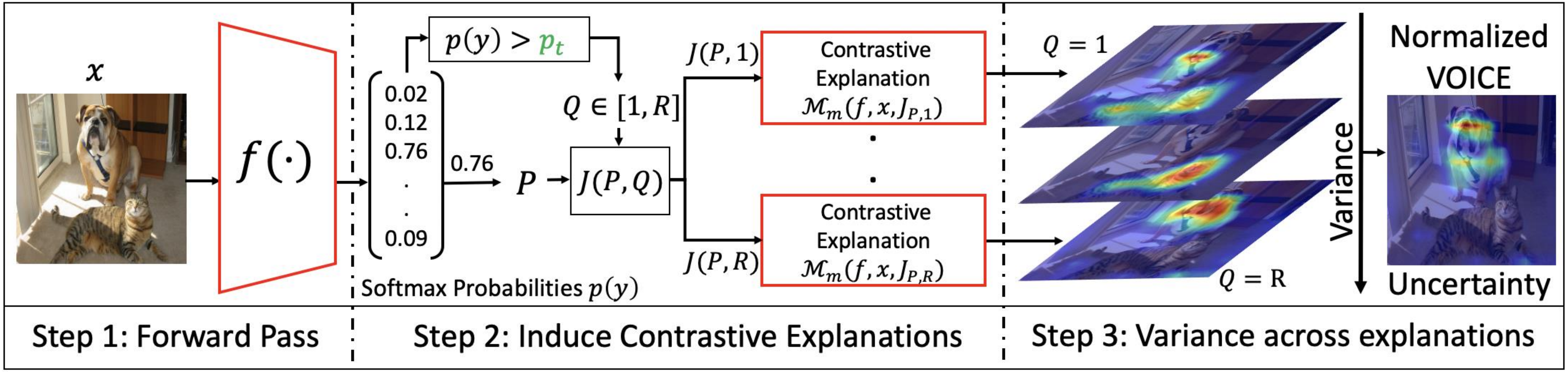
$\boldsymbol{S_{x1}, S_{x2}}$ are explanations from GradCAM and RISE



Other Applications…

VOICE: Variance of Induced Contrastive Explanations to quantify Uncertainty in Neural Network Interpretability

Framework utilizes gradients at Inference



M. Prabhushankar, and G. AlRegib, "VOICE: Variance of Induced Contrastive Explanations to Quantify Uncertainty in Neural Network Interpretability," *Journal of Selected Topics in Signal Processing*, submitted on Aug. 27, 2023.

Conclusion
Key Takeaways

- Interventional Uncertainty provides a framework for analyzing prompts
- Contrastive analysis provides best guess prompts **among available prompts**
  - Answers `*Why Prompt P, rather than Prompt Q?'*
- Contrastive analysis **does not** provide objectively **best prompts**
  - Does not say the exact location of where to prompt
- Future work to include model parameters within the analysis framework

For more OLIVES content, please visit:

GitHub

SCAN ME

Publications

SCAN ME

Tutorials and Short Courses
Completed and Upcoming Tutorials

Latest tutorial delivered at ICIP 2023

SCAN ME

https://alregib.ece.gatech.edu/ieee-icip-2023-tutorial/

Upcoming tutorials/short courses

- **Dec 5 – 7 (Virtual):** 10 hr Short Course on Explainability – Invited by IEEE Signal Processing Society
- **Dec 15-18 (IEEE Big Data 2023):** 3 hr Tutorial on Robustness of Neural Networks
- **Jan 7-8 (WACV 2024):** 3 hr Tutorial on Explainability, Uncertainty, and Intervenability
- **Jan 22 (EI 2024):** 3 hr Tutorial on Explainability, Uncertainty, and Intervenability
- **Feb 21-22 (AAAI 2024):** 3.5 hr Tutorial on Explainability, Uncertainty, and Intervenability