# **Robust Neural Networks Part 4: Intervenability at Inference**





## **Objective** Objective of the Tutorial

### To discuss methodologies that promote robustness in neural networks at inference

- Part 1: Inference in Neural Networks
- Part 2: Explainability at Inference
- Part 3: Uncertainty at Inference
- Part 4: Intervenability at Inference
  - Definitions of Intervenability
    - Causality
    - Privacy
    - Interpretability
    - Prompting
    - Benchmarking
  - Case Study: Intervenability in Interpretability
- Part 5: Conclusions and Future Directions



### **Intervenability** Through the Causal Glass

Assess: The amenability of neural network decisions to human interventions



# "Interventions in data are manipulations that are designed to test for causal factors"

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

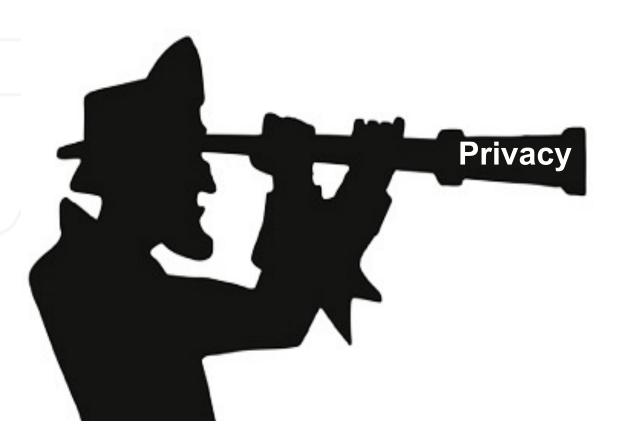


Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.



### **Intervenability** Through the Privacy Glass

**Assure:** The amenability of neural network decisions to human interventions



"Intervenability aims at the possibility for parties involved in any privacy-relevant data processing to interfere with the ongoing or planned data processing"

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



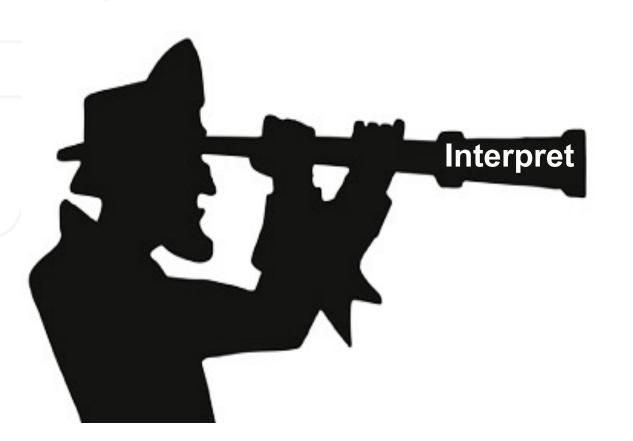
Hansen, M.: Top 10 mistakes in system design from a privacy perspective and privacy protection goals. In: Camenisch, J., Crispo, B., Fischer-Hübner, S., Leenes, R., Russello, G. (eds.) Privacy and Identity Management for Life. IFIP AICT, vol. 375, pp. 14–31. Springer, Heidelberg (2012)



### Intervenability

Through the Interpretability Glass

Interpret: The amenability of neural network decisions to human interventions



"The post-hoc field of explainability, that previously only justified decisions, becomes *active* by being involved in the decision making process and providing limited, but relevant and contextual interventions"

115 of 109 IEEE BigData 2023 Sorrento, Italy [Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

AlRegib, Ghassan, and Mohit Prabhushankar. "Explanatory paradigms in neural networks: Towards relevant and contextual explanations." *IEEE Signal Processing Magazine*39.4 (2022): 59-72.



### Intervenability

### Through the Benchmarking Glass

Verify: The amenability of neural network decisions to human interventions



"... new benchmarks were proposed to specifically test generalization of classification and detection methods with respect to simple algorithmically generated interventions like spatial shifts, blur, changes in brightness or contrast..."

117 of 109 IEEE BigData 2023 Sorrento, Italy [Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.



### **Intervenability** Through the Human Glass

### The amenability of neural network decisions to human interventions



- Assess: Causality
- Assure: Privacy
- Interpret: Interpretability
- Verify: Benchmarking

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. *Proceedings of the IEEE*, *109*(5), 612-634.





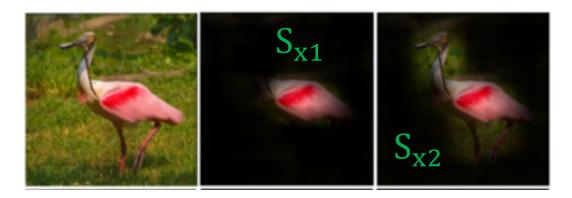
### **Case Study: Intervenability in Interpretability** Explanation Evaluation via Masking

### Common evaluation technique is masking the image and checking for prediction correctness

y = Prediction S<sub>x</sub> = Explanation masked data

 $E(Y|S_x)$  = Expectation of class given  $S_x$ 

2018.



If across N images,  $E(Y|S_{x2}) > E(Y|S_{x1})$ , explanation technique 2 is better than explanation technique 1





[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023] Chattopadhay, Aditya, et al. "Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks." 2018 IEEE winter conference on applications of computer vision (WACV). IEEE,







## **VOICE: Variance of Contrastive Explanations for Quantifying Uncertainty in Interpretability**



Mohit Prabhushankar, PhD Postdoc



Ghassan AlRegib, PhD Professor





**Predictive Uncertainty in Explanations** 

Explanatory techniques have predictive uncertainty Explanation of Prediction Uncertainty of Explanation



Uncertainty in answering Why Bullmastiff?

122 of 109 IEEE BigData 2023 Sorrento, Italy

Why Bullmastiff?

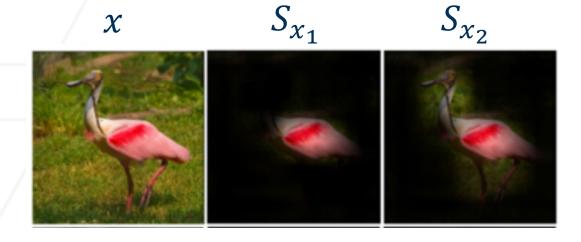
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





### Case Study: Intervenability in Interpretability Predictive Uncertainty

### Uncertainty due to variance in prediction when model is kept constant



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

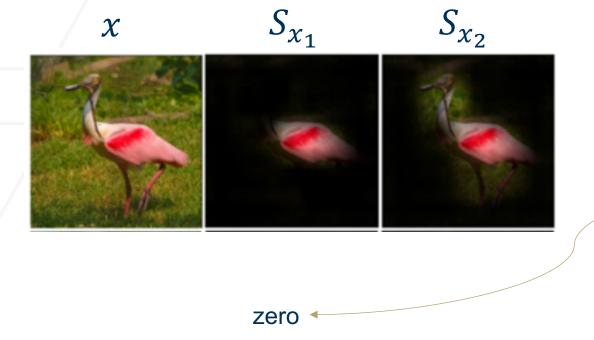
 $\begin{array}{l} y = \mbox{Prediction} \\ V[y] = \mbox{Variance of prediction (Predictive Uncertainty)} \\ S_x = \mbox{Subset of data (Some intervention)} \\ E(Y|S_x) = \mbox{Expectation of class given a subset} \\ V(Y|S_x) = \mbox{Variance of class given all other residuals} \end{array}$ 

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



### **Case Study: Intervenability in Interpretability** Visual Explanations (partially) reduce Predictive Uncertainty

A 'good' explanatory technique is evaluated to have zero  $V[E(y|S_x)]$ 



### Key Observation 1: Visual Explanations are evaluated to partially reduce the predictive uncertainty in a neural network

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

y = Prediction V[y] = Variance of prediction (Predictive Uncertainty)  $S_x = Subset of data (Some intervention)$   $E(Y|S_x) = Expectation of class given a subset$  $V(Y|S_x) = Variance of class given all other residuals$ 

Network evaluations have nothing to do with human Explainability!



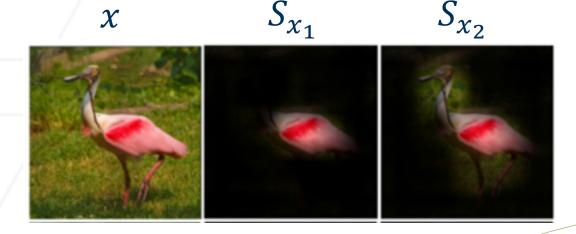
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





Predictive Uncertainty in Explanations is the Residual

All other subsets 'not' chosen by the explanatory technique contributes to uncertainty



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

y = Prediction V[y] = Variance of prediction (Predictive Uncertainty)  $S_x =$  Subset of data (Some intervention)  $E(Y|S_x) =$  Expectation of class given a subset  $V(Y|S_x) =$  Variance of class given all other residuals

### Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision



Sorrento, Italv

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



Predictive Uncertainty in Explanations is the Residual

### All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

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

# The effect of a chosen Interventions can be measured based on all the Interventions that were not chosen

 $\tilde{E}(Y|S_x) = Expectation of class given a subset$  $V(Y|S_x) = Variance of class given all other residuals$ 

Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision

126 of 109 IEEE BigData 2023 Sorrento, Italy [Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





Explanation of Prediction

Predictive Uncertainty in Explanations is the Residual

### All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

Uncertainty of Explanation

Snout is not as highlighted as the jowls in explanation (not as important for decision)

However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

### Key Observation 2: Uncertainty in Explainability occurs due to all combinations of features that the explanation did not attribute to the network's decision



[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





Predictive Uncertainty in Explanations is the Residual

### All other subsets 'not' chosen by the explanatory technique contributes to uncertainty

Snout is not as highlighted as the jowls in explanation (not as important for decision)

### Explanation of Prediction Uncertainty of Explanation

However, snout is an important characteristic that is used to differentiate against other dogs. Hence, there is uncertainty on why this feature is not included in the attribution

### Not chosen features are intractable!



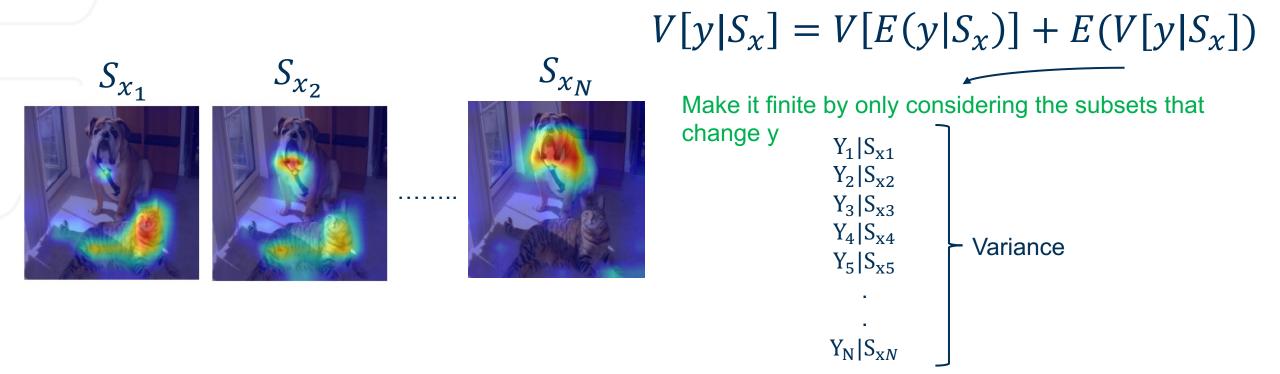
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





Quantifying Interventions in Explainability

Contrastive explanations are an intelligent way of obtaining other subsets





[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



Case Study: Intervenability in Interpretability Quantifying Interventions in Explainability

Uncertainty in Explainability can be used to analyze Explanatory methods and Networks

- Is GradCAM better than GradCAM++?
- Is a SWIN transformer more reliable than VGG-16?

## Need objective quantification of Intervention Residuals



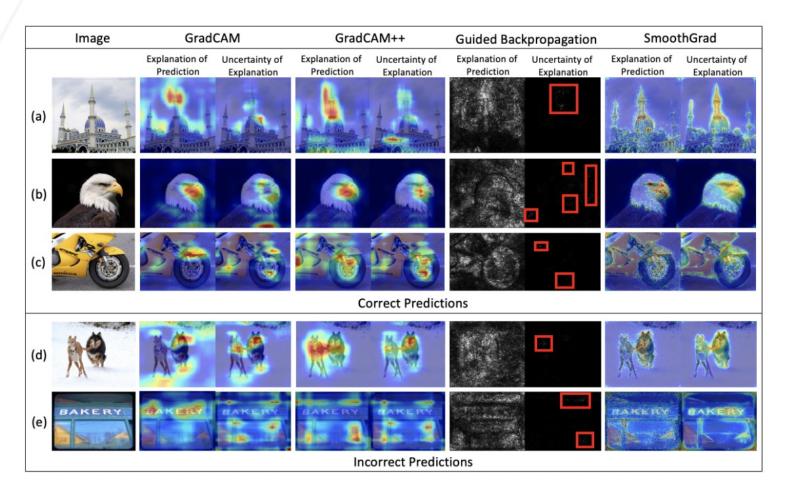
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





## Case Study: Intervenability in Interpretability Quantifying Interventions in Explainability: mIOU

### On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric: Intersection over Union (IoU) between explanation and Uncertainty

Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

### 132 of 109 IE

IEEE BigData 2023 Sorrento. Italv

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.

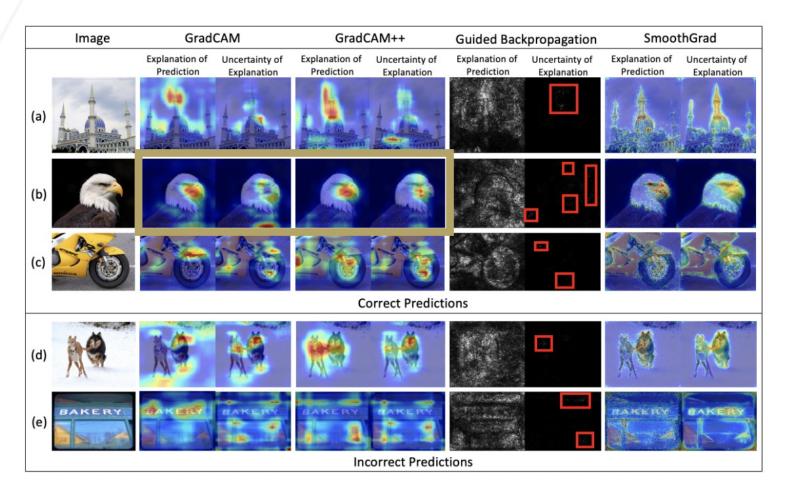
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





## Case Study: Intervenability in Interpretability Quantifying Interventions in Explainability: mIOU

### On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric 1: Intersection over Union (IoU) between explanation and Uncertainty

Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

### 134 of 109

IEEE BigData 2023 Sorrento. Italv

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.

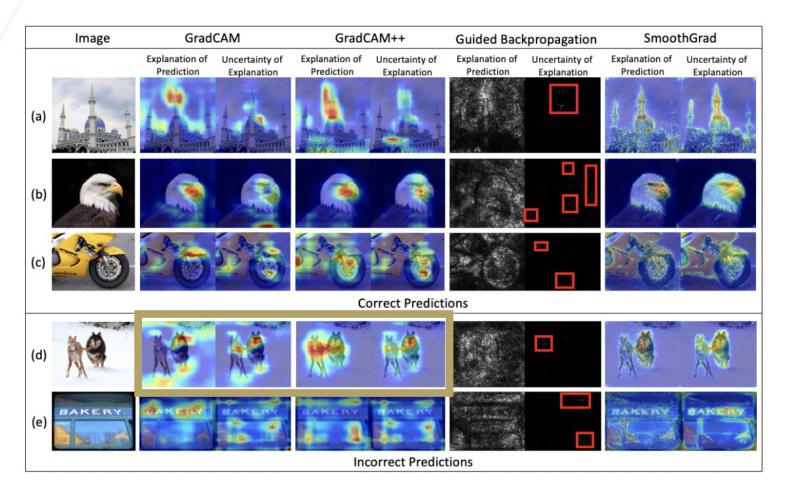
[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]





## Case Study: Intervenability in Interpretability Quantifying Interventions in Explainability: mIOU

### On incorrect predictions, the overlap of explanations and uncertainty is higher



Objective Metric 1: Intersection over Union (IoU) between explanation and Uncertainty

Higher the IoU, higher the uncertainty in explanation (or less trustworthy is the prediction)

### 135 of 109 IEE

IEEE BigData 2023 Sorrento, Italy M. F

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.

[Tutorial@BigData'23] | [Ghassan AlRegib and Mohit Prabhushankar] | [Dec 15, 2023]



