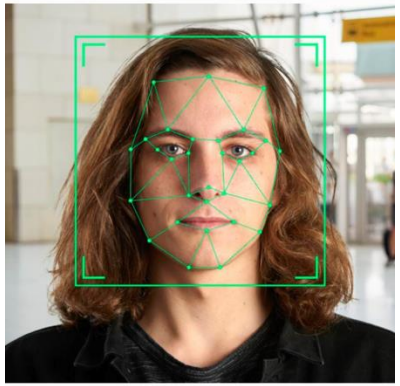
HARVARD
UNIVERSITY

Interpreting AI Systems Through Features, Data, and Model Components: A Tour of Attribution Methods

Shichang (Ray) Zhang

04/08/2025

The AI Advancement



DeepFace
human-level face recognition in 2014 (97.35% accuracy)

Image credit: Forbes 2020

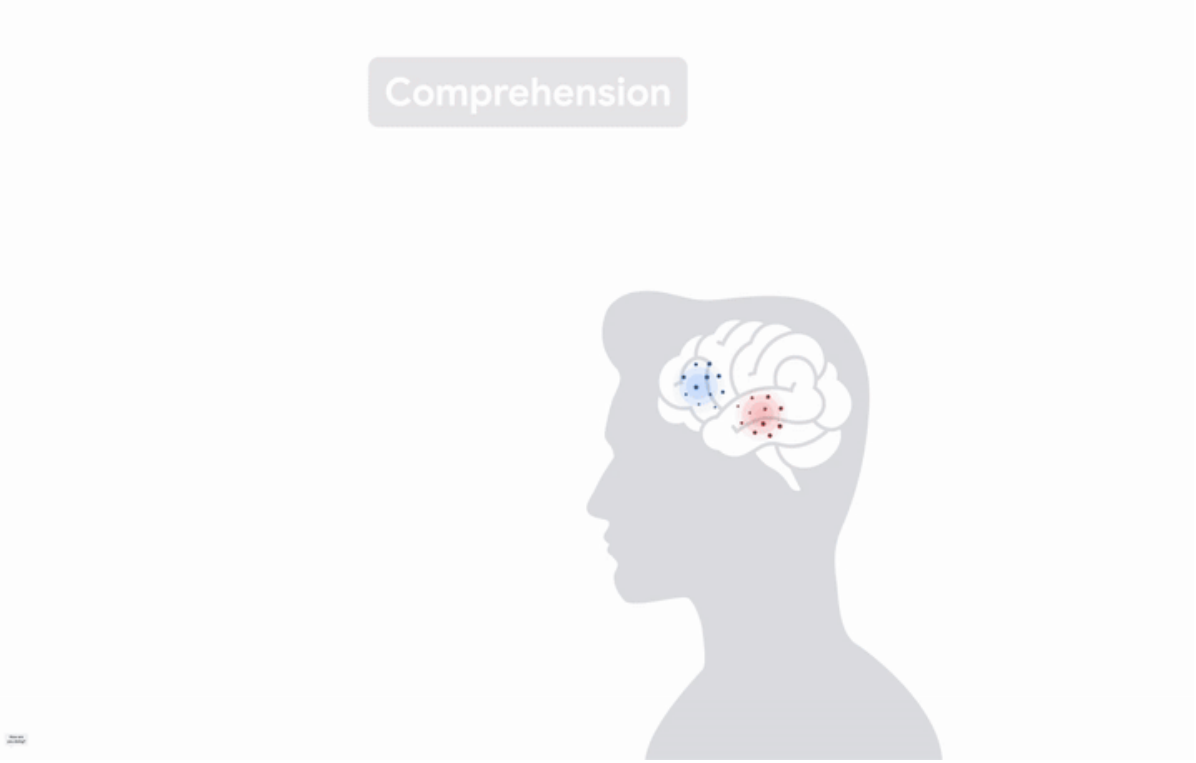
You
When will AI replace human?

ChatGPT
AI is unlikely to completely replace humans. It is designed to assist and enhance human capabilities in specific tasks, but it cannot replicate the full range of human emotions, creativity, or decision-making.

AlphaFold
Accurate prediction of protein structures

[Taigman et al., 2014, Ouyang et al., 2022, Abramson et al., 2024]

The AI Advancement



Neural activity in the human brain aligns linearly with LLM embeddings

[Goldstein et al., 2025]

The AI Advancement

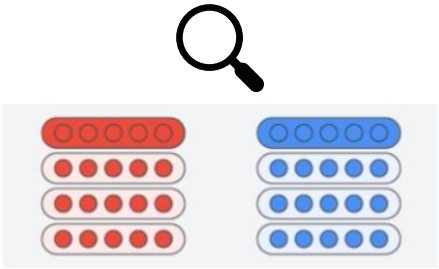
Why?

Neural activity in the human brain aligns linearly with LLM embeddings

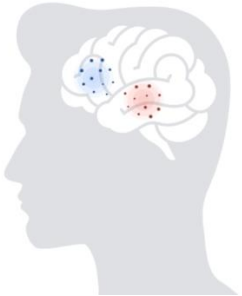
[Goldstein et al., 2025]

The “Why” Question

Why?



Mind of AI



Human Brain

[Taigman et al., 2014, Ouyang et al., 2022, Goldstein et al., 2025]

Why Is The “Why” Question Important?

User
Trust

Data
Insight

Model
Enhancement

What Specific “Why” Questions Can We Ask?

- How is an input processed?
- How does training data shape the model?
- How does each model component influence model behavior?
- Answer: Interpret AI via Attribution
  - Zhang, S., Han, T., Bhalla, U., and Lakkaraju, H. (2025). Building Bridges, Not Walls--Advancing Interpretability by Unifying Feature, Data, and Model Component Attribution. https://arxiv.org/pdf/2501.18887

Outline

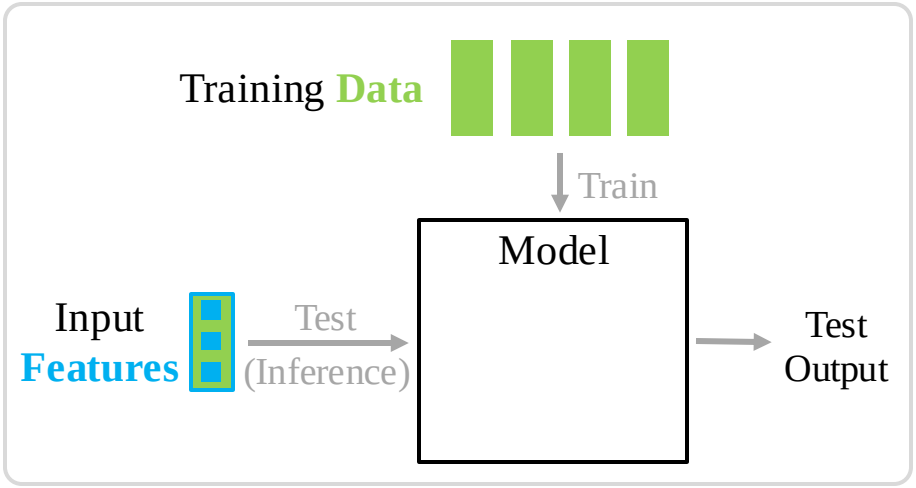
- The Attribution Problem
- Attribution Methods
- A Unified View of Attribution
- Cross-Aspect Innovation and Connections to Other Areas of AI

Outline

- The Attribution Problem
- Attribution Methods
- A Unified View of Attribution
- Cross-Aspect Innovation and Connections to Other Areas of AI

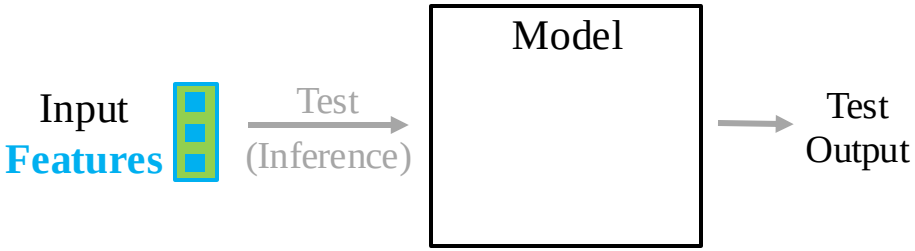
The Attribution Problem

- Consider an abstraction of an AI system
- How to explain the model output? Attribution



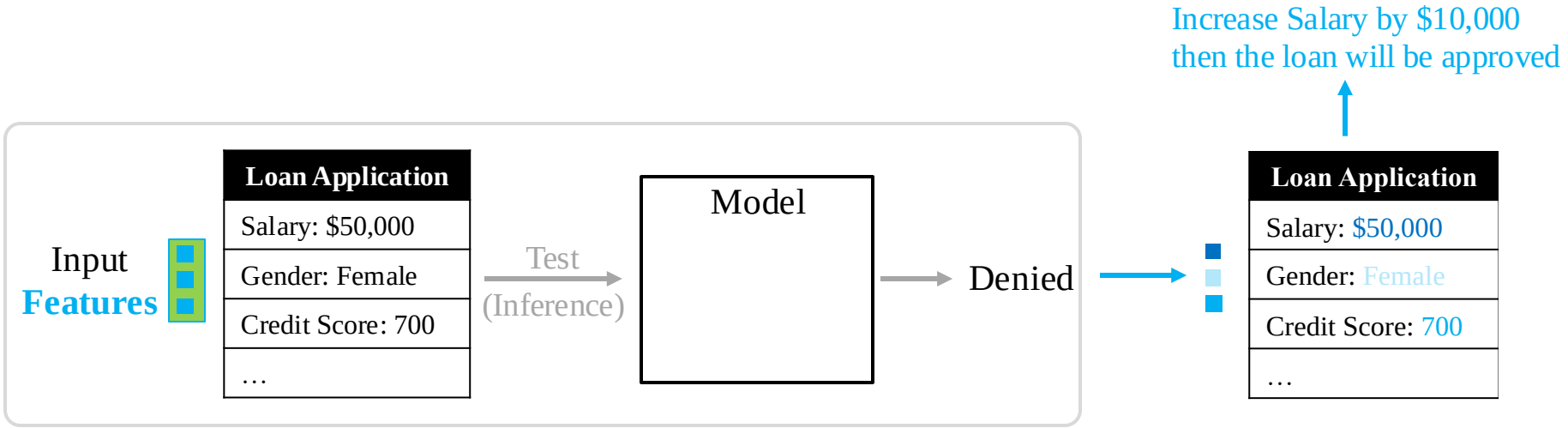
Feature Attribution

- Why this output for these input features?
- Feature attribution quantifies how features influence the model's output
  - Test time, without altering model parameters



Feature Attribution

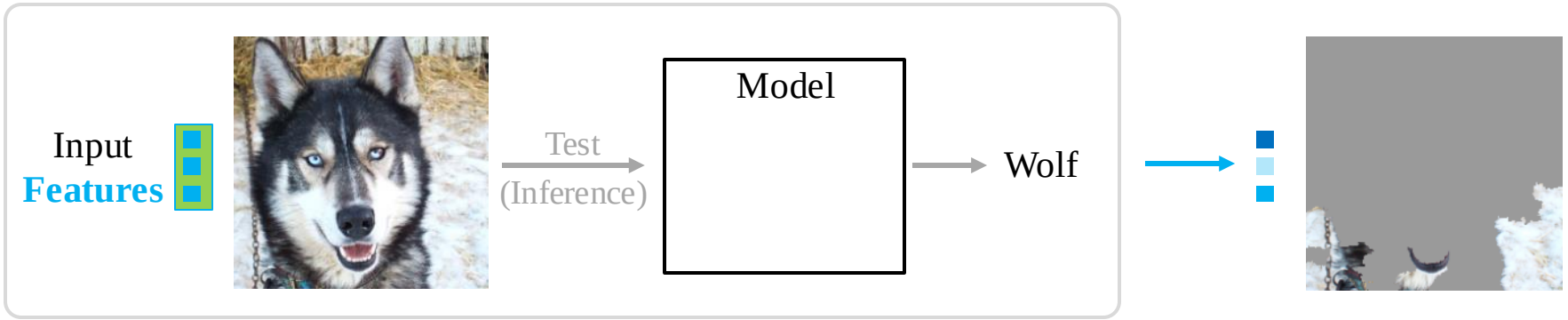
- Justify prediction, gain trust, and provide recourse by counterfactual explanations



[Wachter et al, 2018]

Feature Attribution

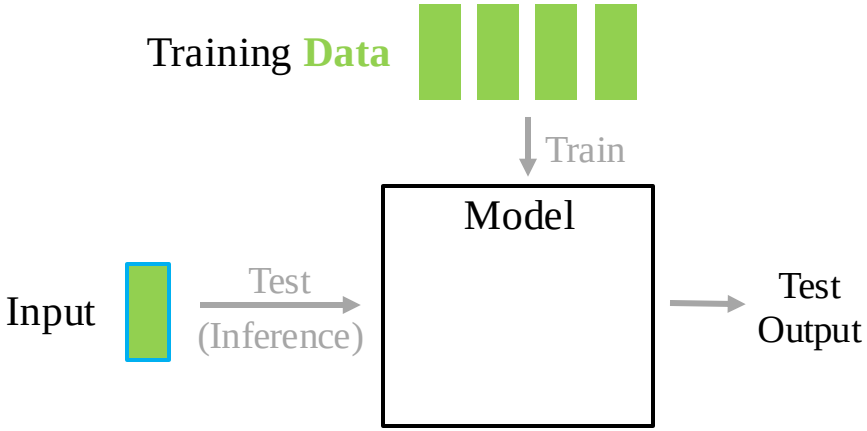
- Justify prediction, gain trust, and provide recourse by counterfactual explanations
- Identify spurious correlations



[Ribeiro et al, 2016]

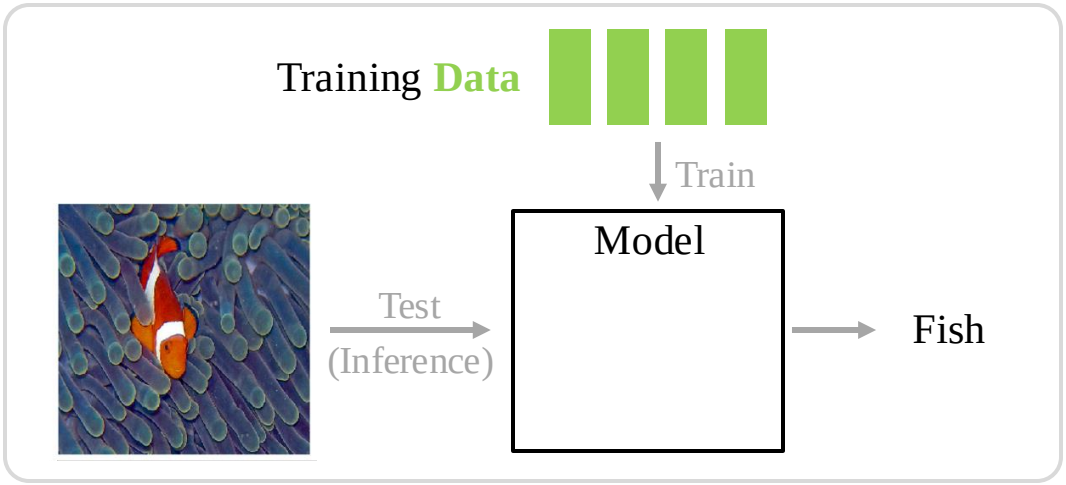
Data Attribution

- Why this output for these training data points?
- Data attribution studies how the training data shape model behavior



Data Attribution

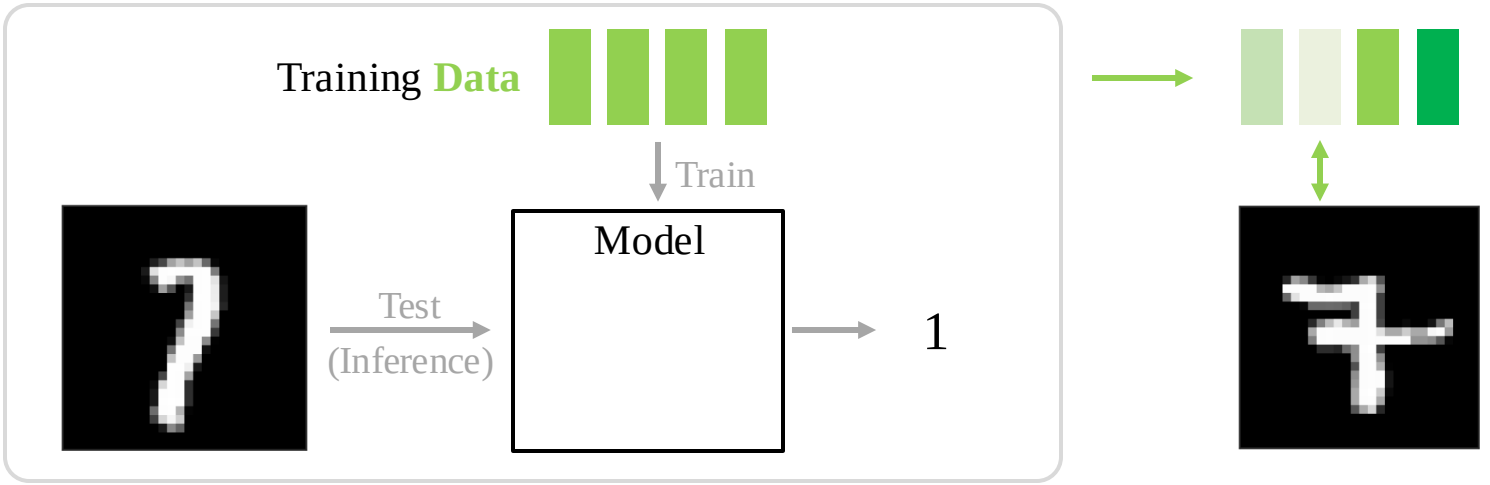
- Characterize training data properties and value



[Koh & Liang, 2017]

Data Attribution

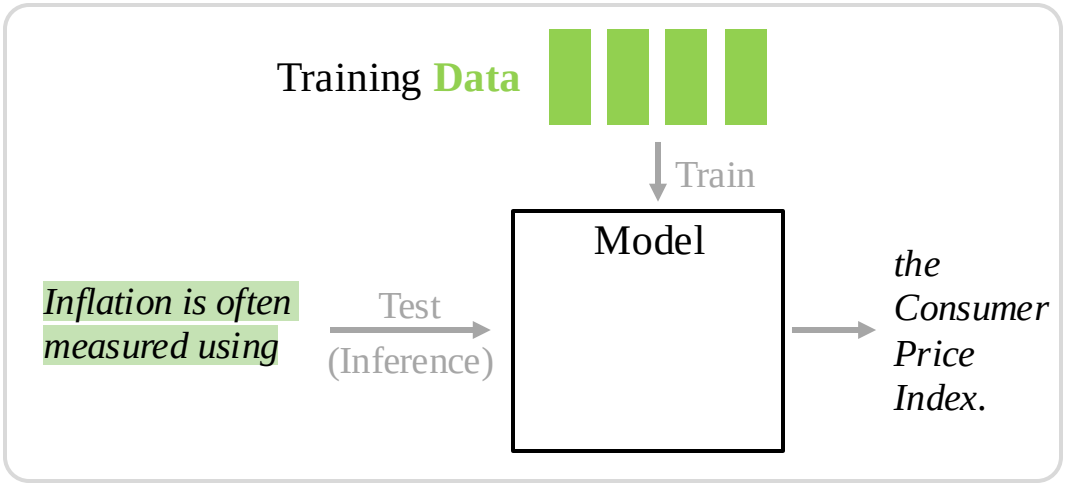
- Characterize training data properties and value
- Justify harmful data



[Koh & Liang, 2017]

Data Attribution

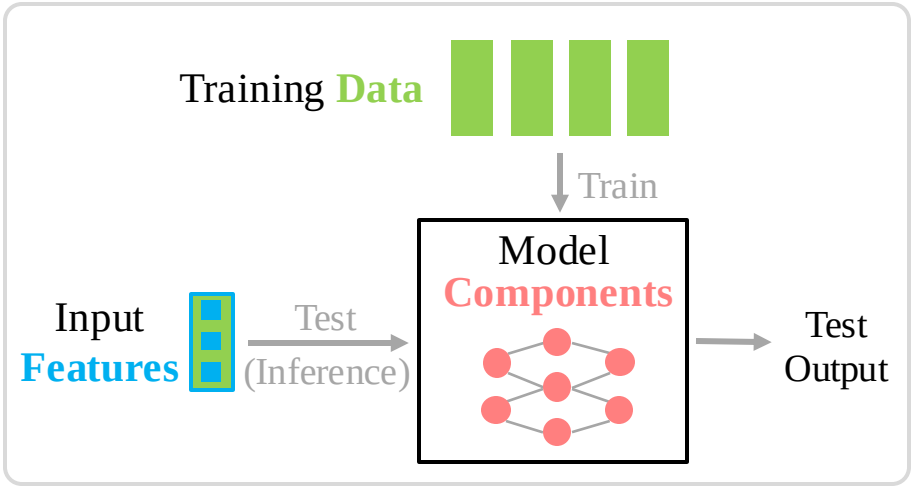
- Characterize training data properties and value
- Justify harmful data



Inflation is most commonly measured using the Consumer Price Index (CPI) or the Producer Price Index (PPI).

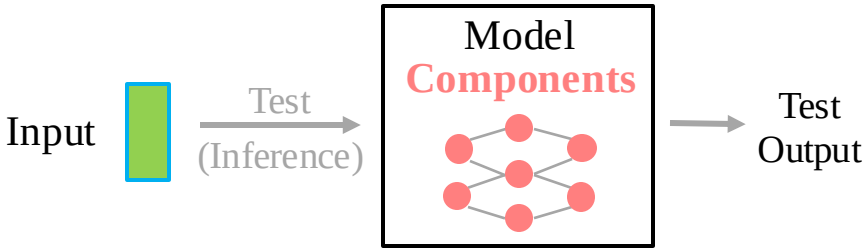
[Grosse et al, 2023]

The Attribution Problem



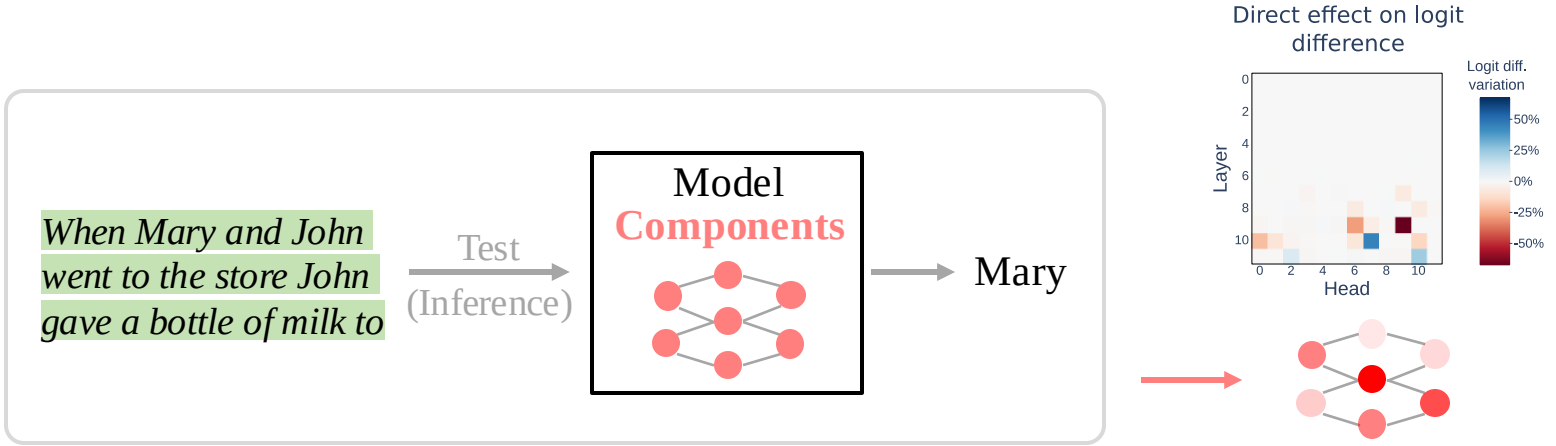
Component Attribution

- Why this output for these model components?
- Component attribution analyzes how model components contribute to its output
  - Components have various definitions, e.g., neurons, attention heads, etc.



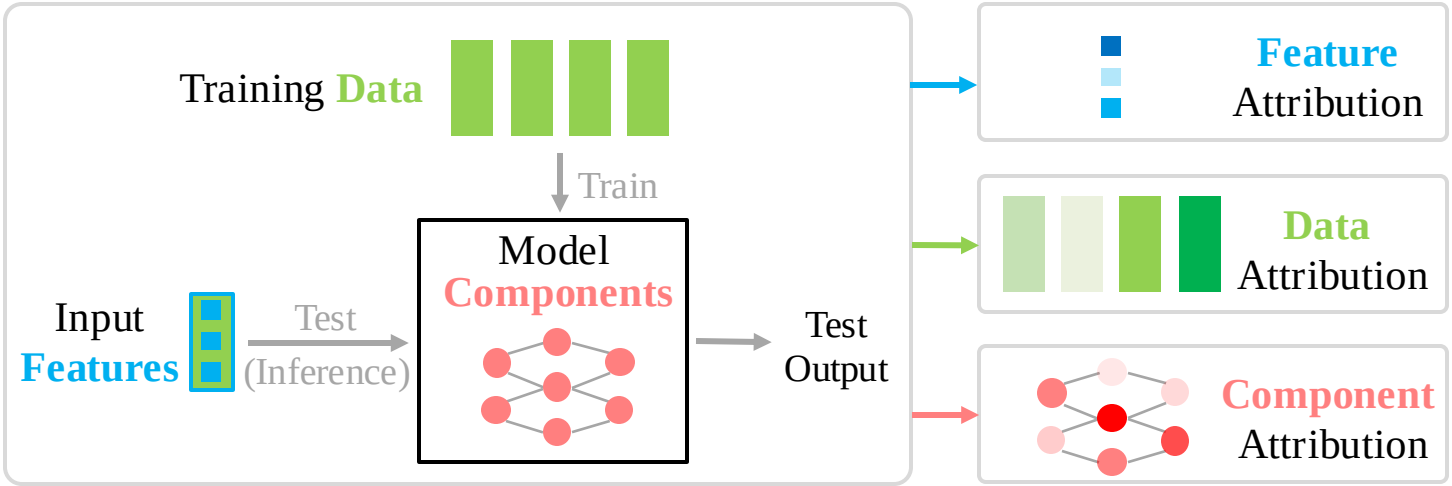
Component Attribution

- Why this output for these model components?
- Component attribution analyzes how model components contribute to its output
  - Components have various definitions, e.g., neurons, attention heads, etc.

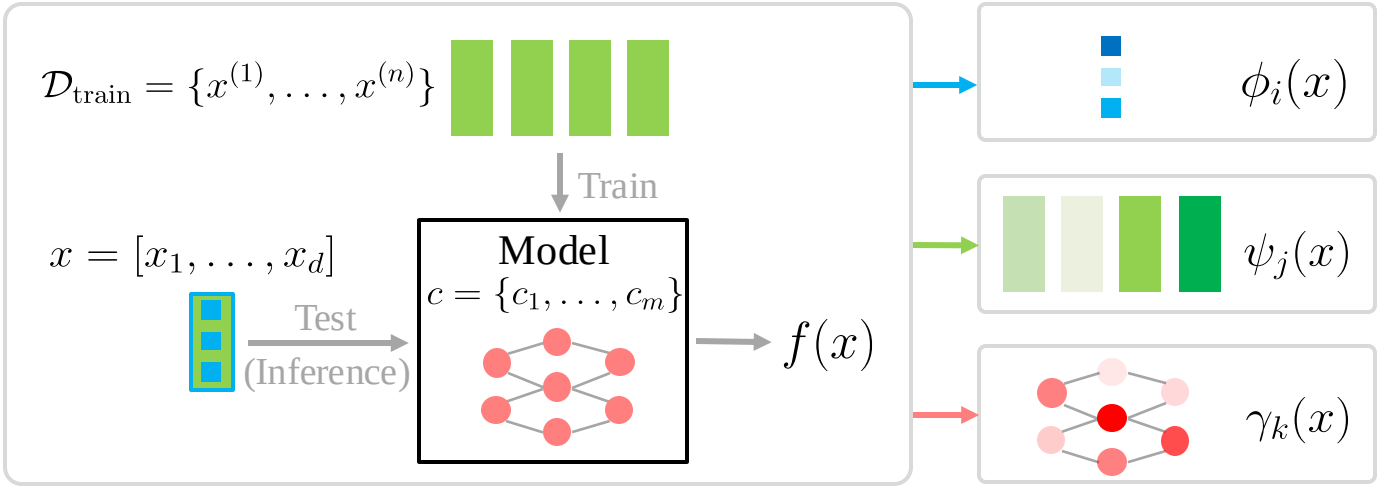


[Wang et al, 2022]

Three Types of Attribution



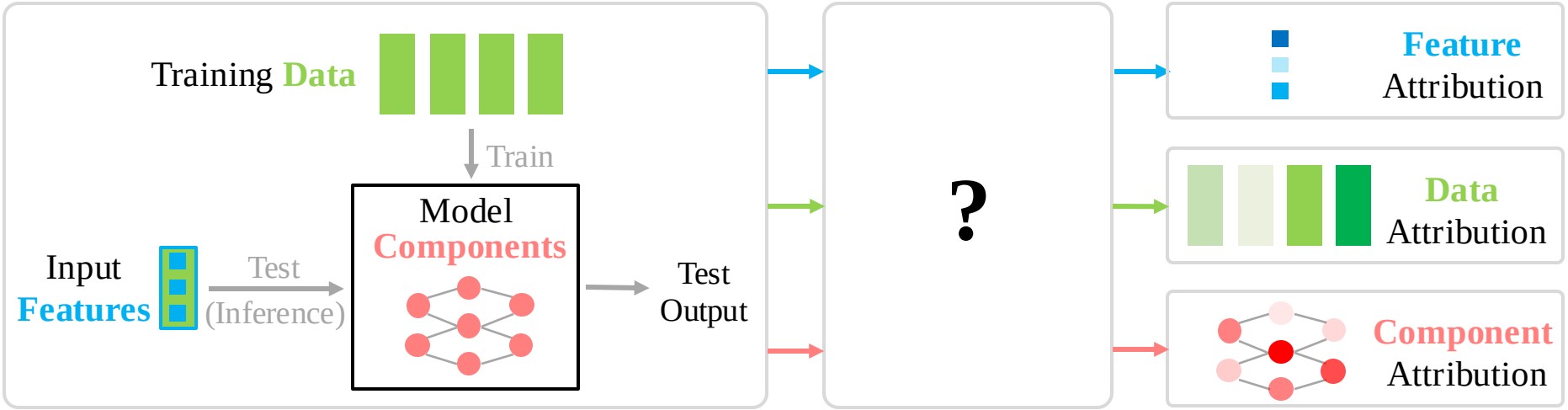
Formalization of The Attribution Problem



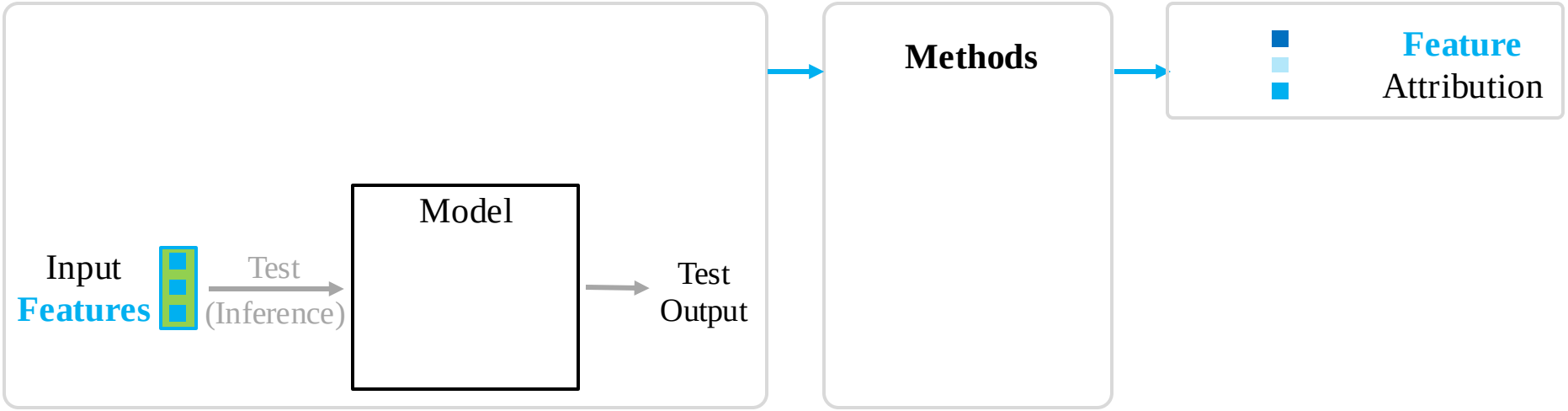
Outline

- The Attribution Problem
- Attribution Methods
- A Unified View of Attribution
- Cross-Aspect Innovation and Connections to Other Areas of AI

How Are These Attribution Results Achieved?



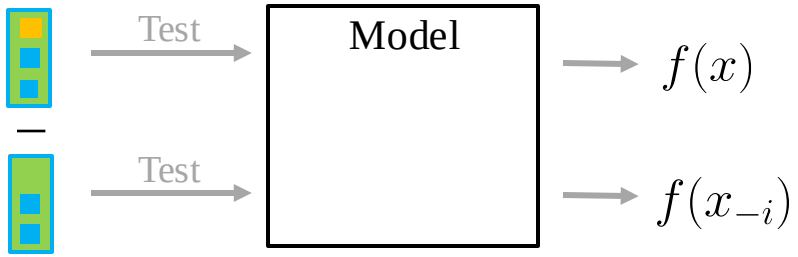
Feature Attribution Methods



Perturbation-Based Feature Attribution

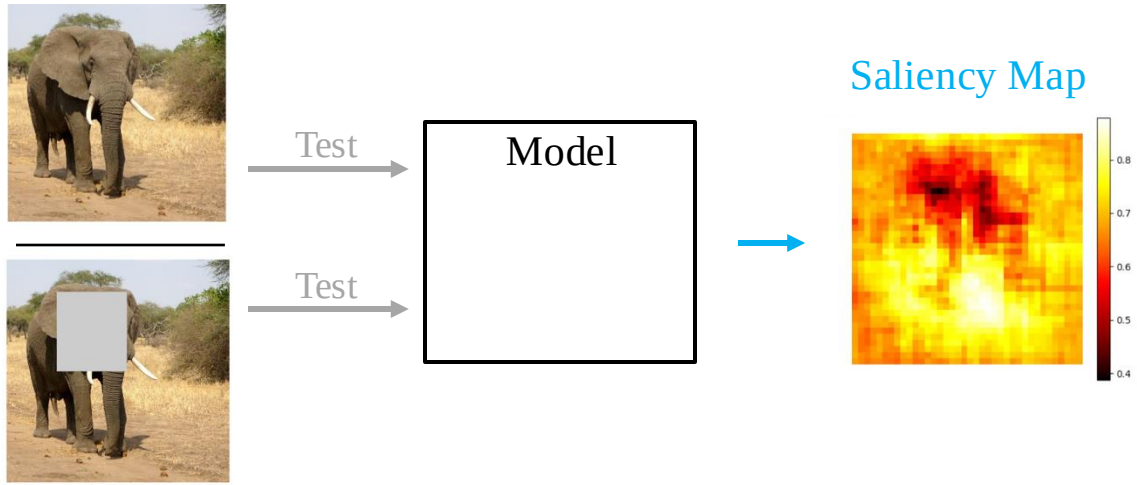
- Direct Perturbation
  - Perturb features and observe output changes

$$f(x) - f(x_{-i}) \rightarrow \phi_i(x)$$



Perturbation-Based Feature Attribution

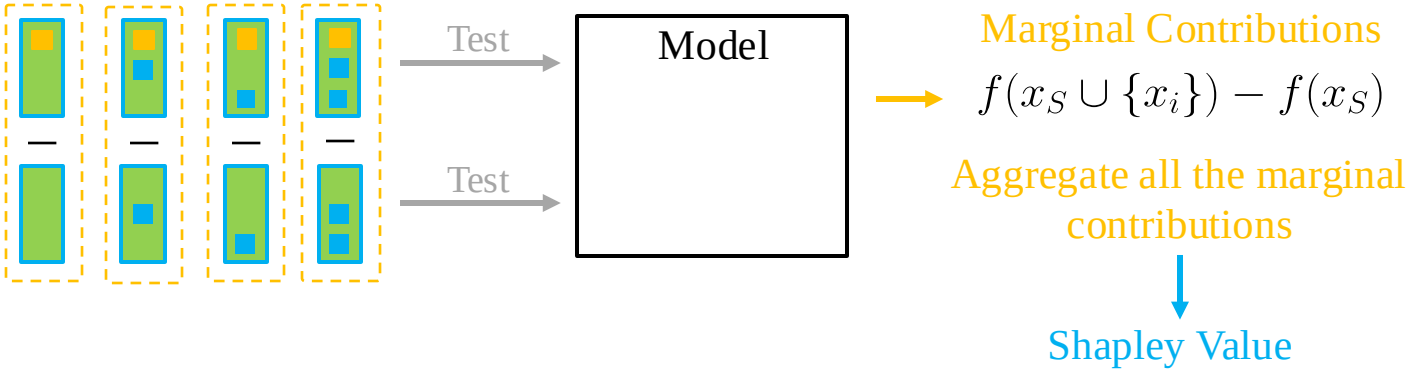
- Direct Perturbation
  - Perturb features and observe output changes
  - Occlusion



[Zeiler and Fergus, 2014]

Perturbation-Based Feature Attribution

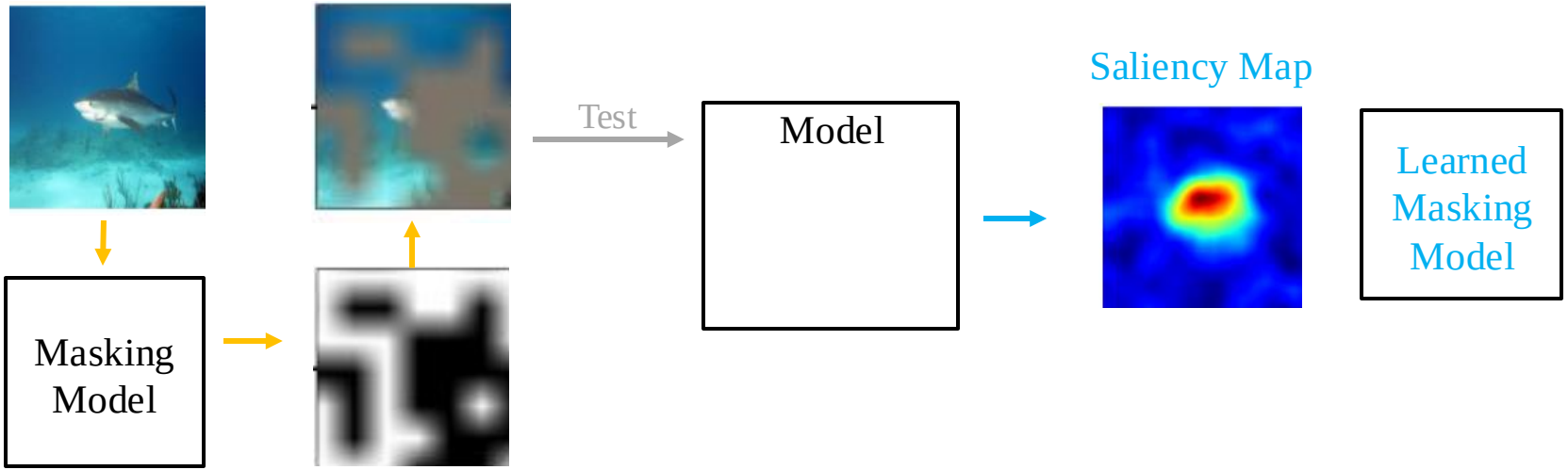
- Feature interactions?
- Game-Theoretic Perturbation
  - Features as players, model processing as a game. Fairly distribute outcome
  - 2^d marginal contributions



[Shapley, 1953]

Perturbation-Based Feature Attribution

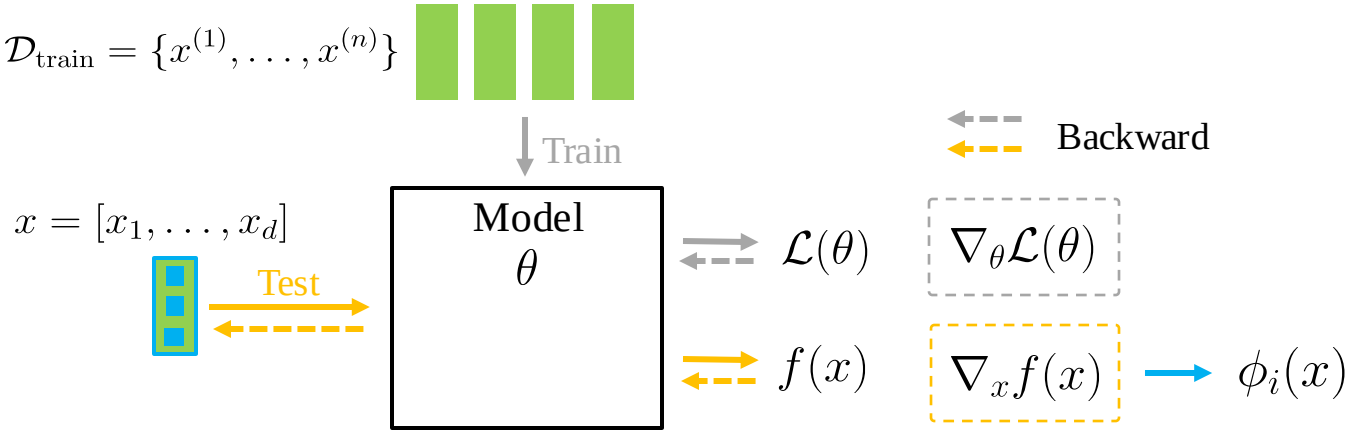
- Perturbation Mask Learning
  - Perturbations can be seen as binary masks
  - Why not make them continuous and learnable?
  - The masking model can be applied to other inputs



[Dabkowski & Gal, 2017, Petsiuk et al., 2018]

Gradient-Based Feature Attribution

- Parameter gradients for model training vs. features gradients for attribution
- Measure output sensitivity to input features; No updates to model parameters



[Dabkowski & Gal, 2017, Petsiuk et al., 2018]

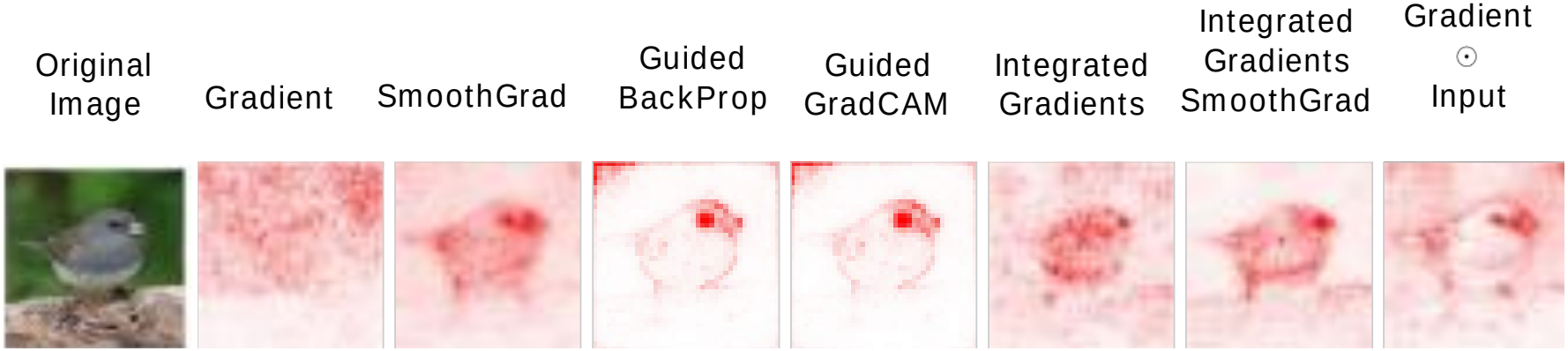
Gradient-Based Feature Attribution

- “Vanilla” Gradients

[Simonyan et al., 2013, Adebayo et al., 2018]

Gradient-Based Feature Attribution

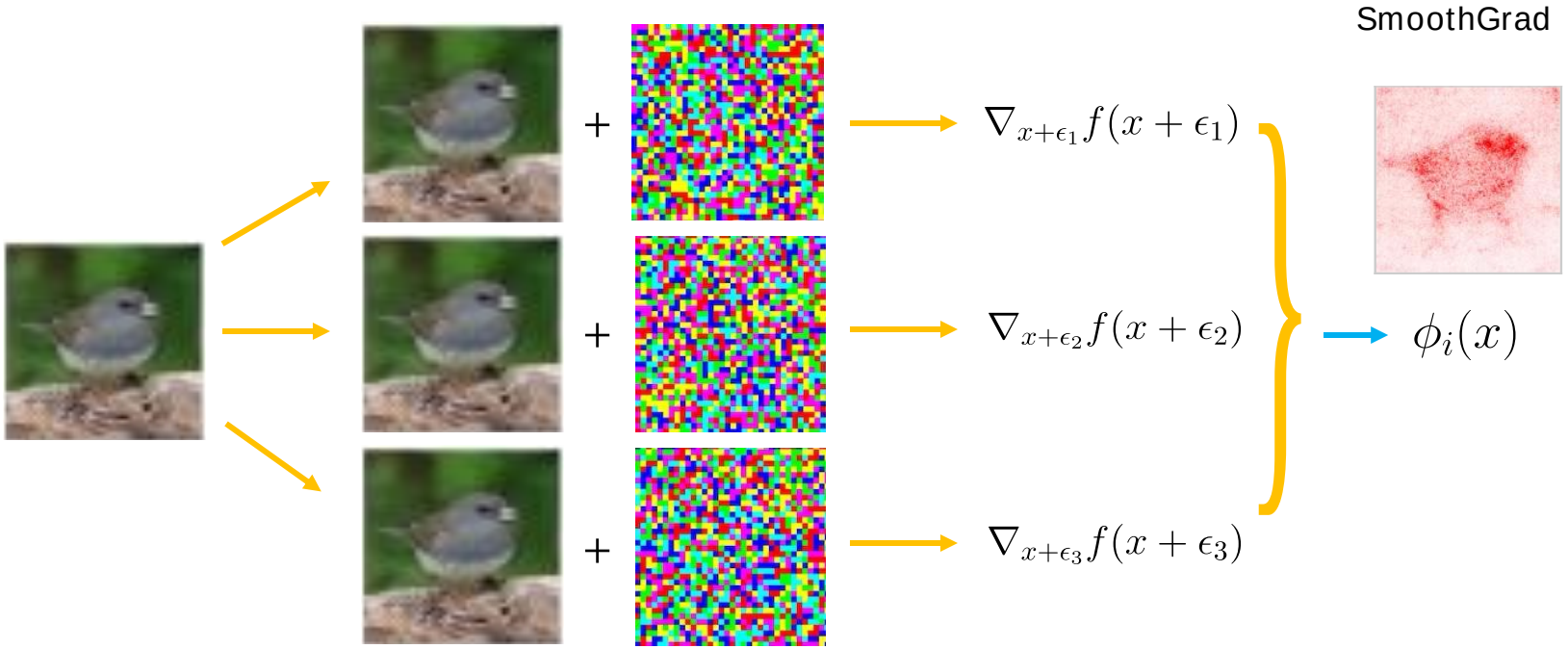
- Variations of gradient-based methods



[Adebayo et al., 2018]

Gradient-Based Feature Attribution

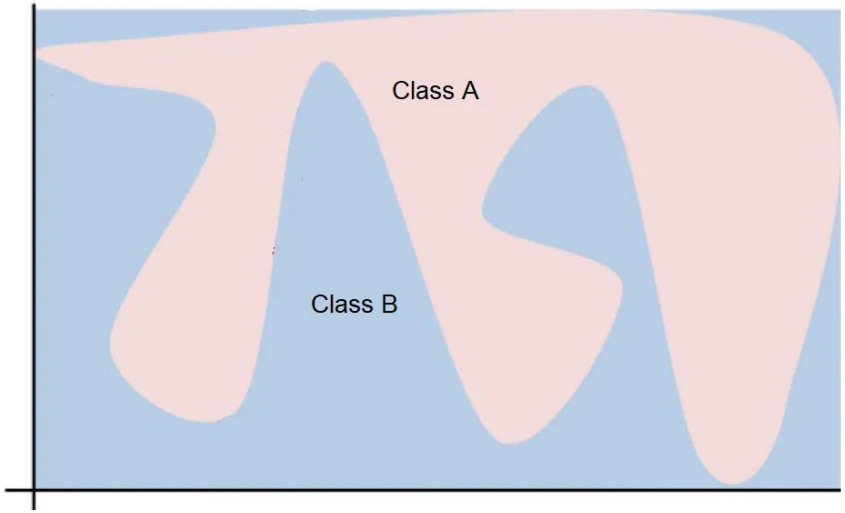
- SmoothGrad
  - Smooth over noisy versions of the input to enhance robustness



[Adebayo et al., 2018, Smilkov et al., 2017]

Linear Approximations for Feature Attribution

- If we cannot understand the complex AI models, what model can we understand?
- Local approximation

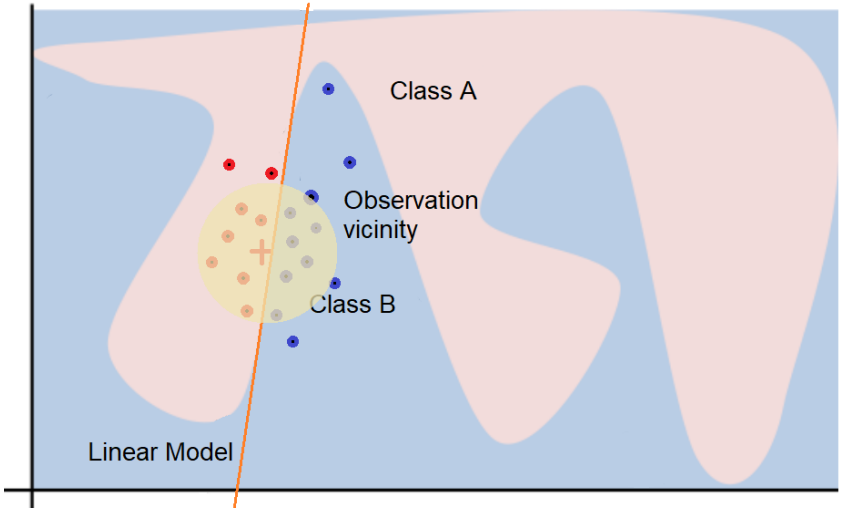


[Ribeiro et al., 2016, Santiago, 2020]

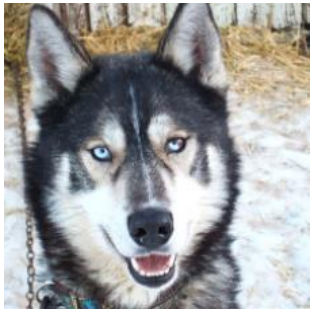
Linear Approximations for Feature Attribution

- If we cannot understand the complex AI models, what model can we understand?
- Local approximation
- We don't even need actual input features, only binary indicators

$$g(x) = w^T x + b \qquad x = \{1, 0\}^d \qquad w \longrightarrow \phi_i(x)$$



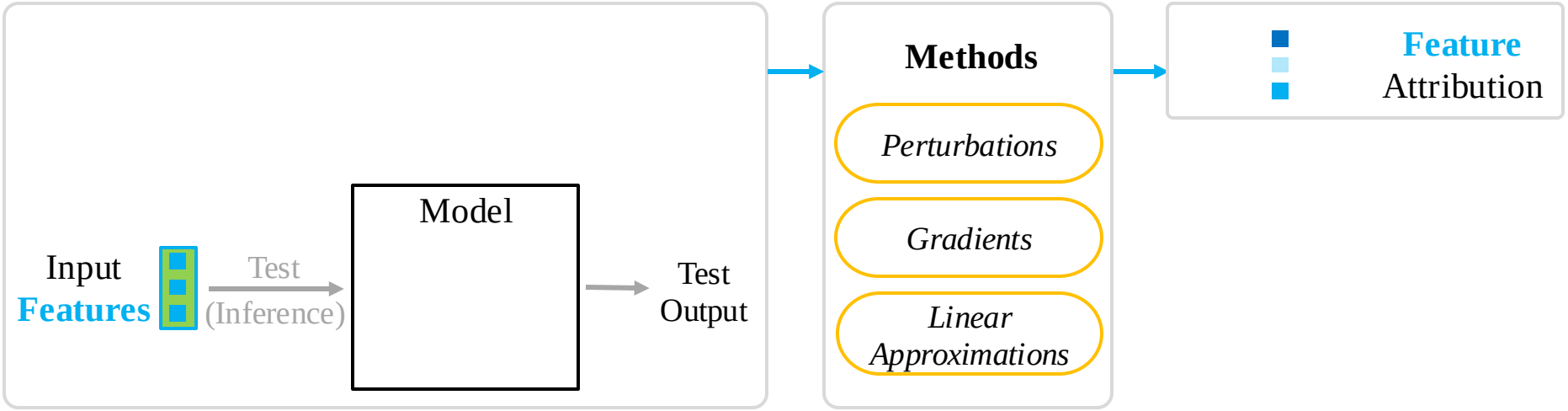
Husky?



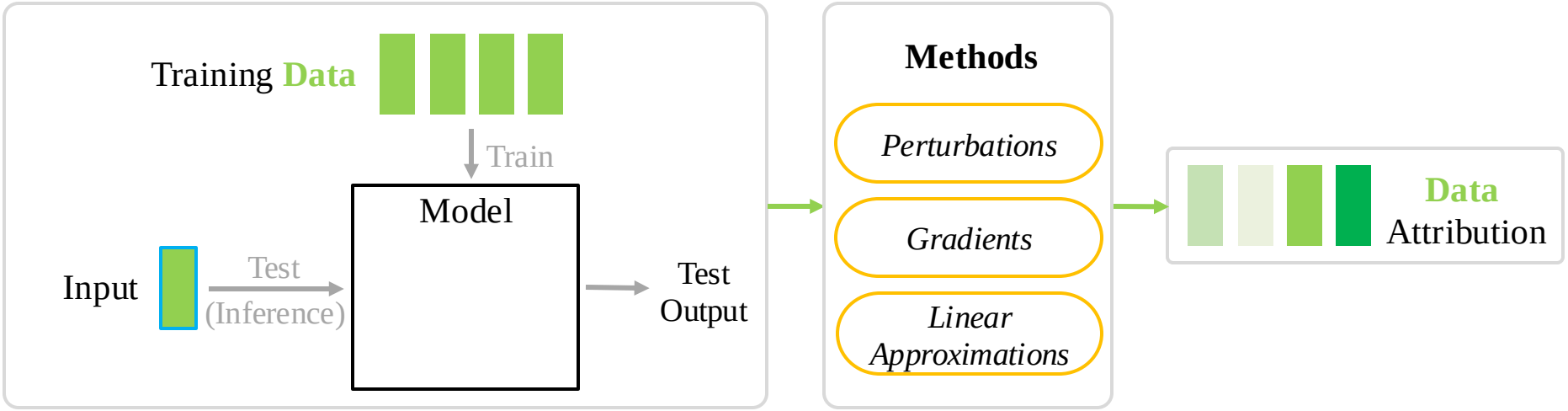
Wolf !

[Ribeiro et al., 2016, Santiago, 2020]

Feature Attribution Methods



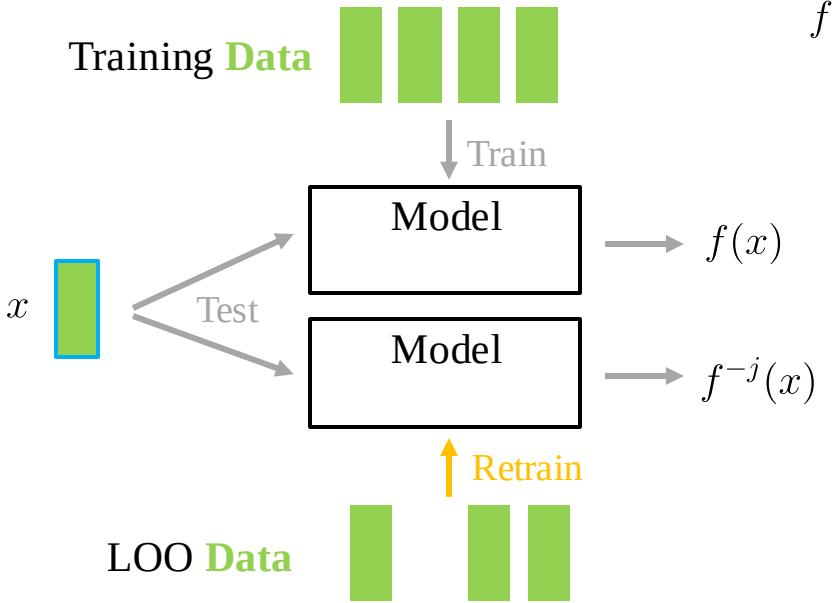
How About Data Attribution?



Perturbation-Based Data Attribution

- Leave-One-Out (Direct Perturbation)
  - Remove one training point at a time, retrain model, and observe output changes
  - Computationally expensive, retrain n times

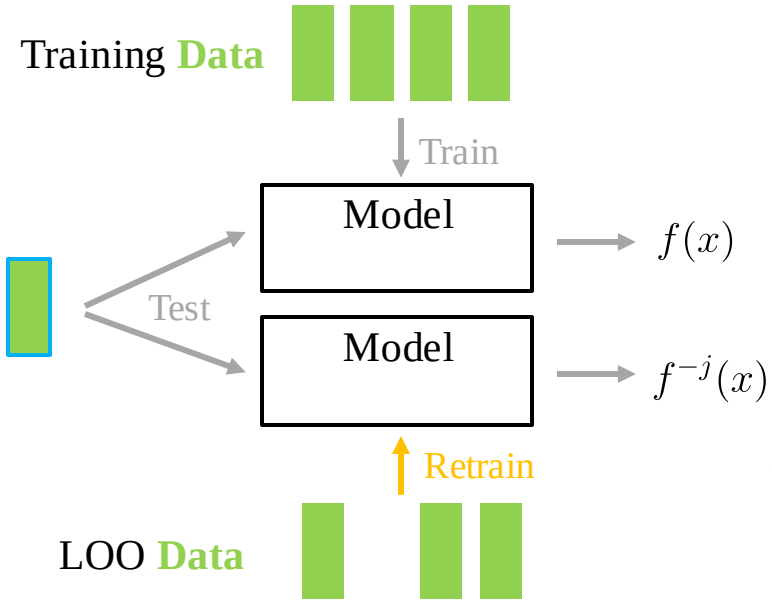
$$f(x) - f^{-j}(x) \rightarrow \psi_j(x)$$

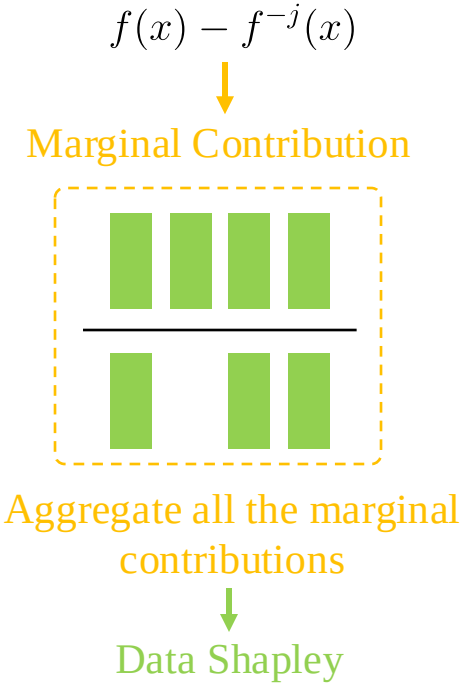


[Cook & Weisberg, 1982]

Perturbation-Based Data Attribution

- Game-Theoretic Perturbation
  - Capture training data interactions
  - 2^n marginal contributions, each with a retraining

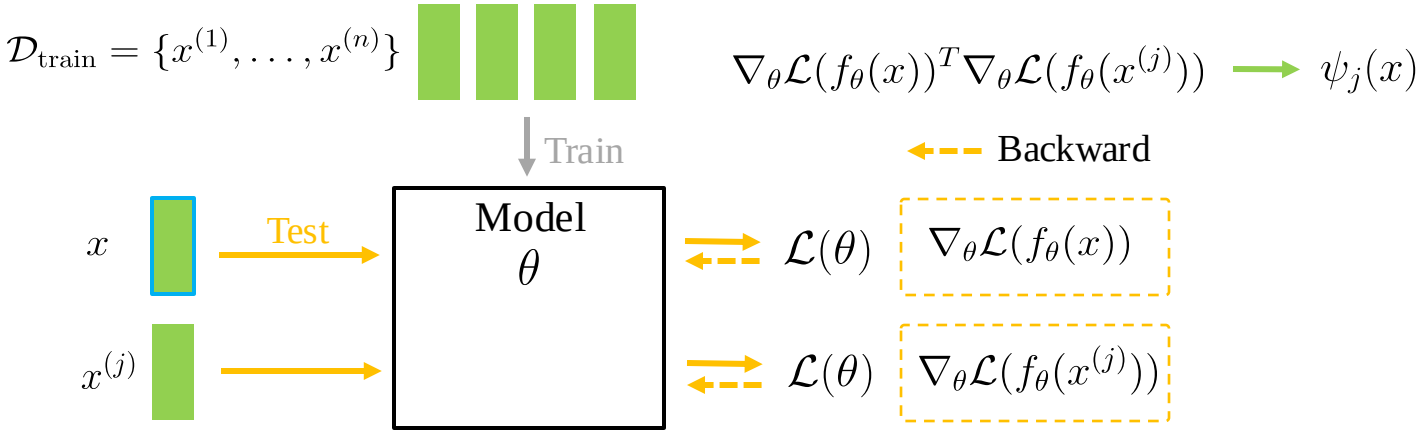




[Ghorbani & Zou, 2019]

Gradient-Based Data Attribution

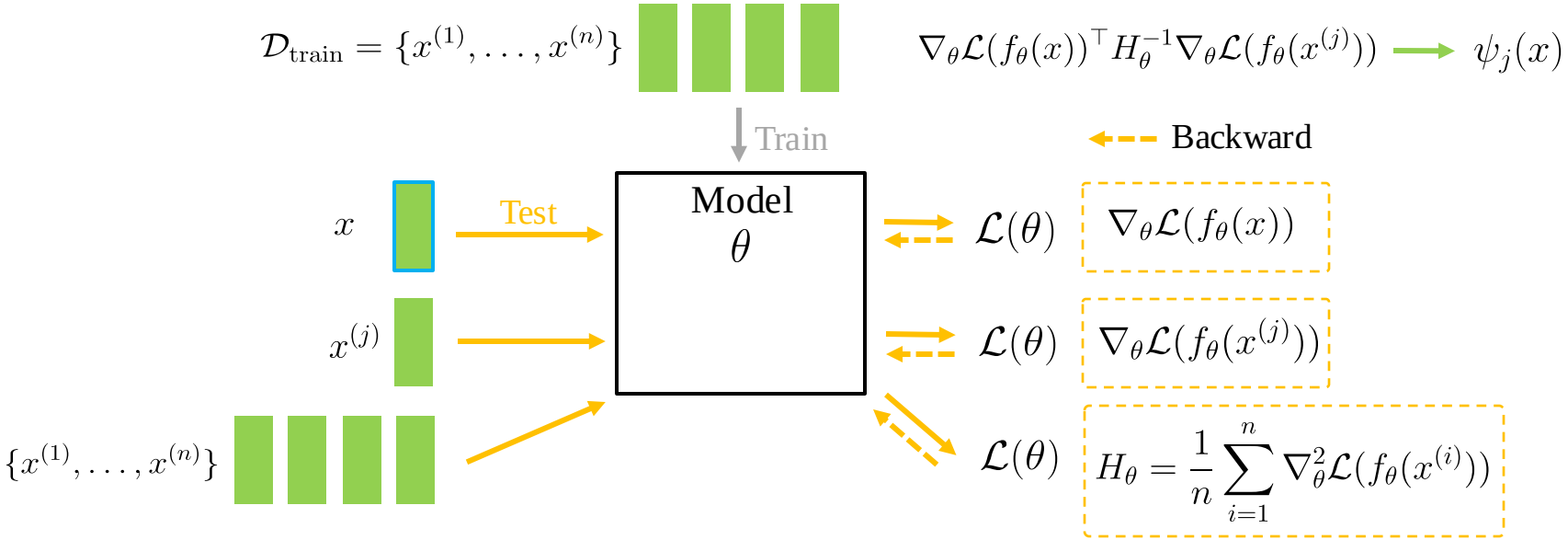
- No retrainings (perturbations)
  - Gradients for measuring similarity between data points
  - Dot products of gradients evaluated at the training and test points



[Charpiat et al., 2019]

Gradient-Based Data Attribution

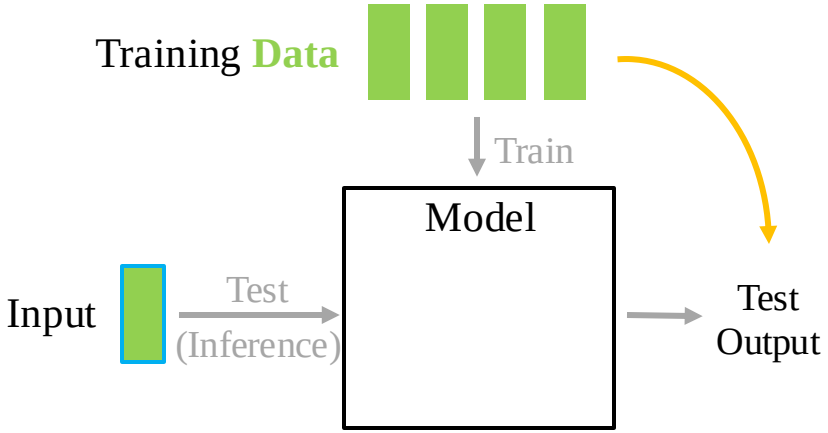
- Influence Functions
  - Approximate LOO by lifting one data point $\theta_{\epsilon,x^{(j)}} = \arg\min_\theta \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(x^{(i)},\theta) + \epsilon\mathcal{L}(x^{(j)},\theta)$
  - Compute Hessian matrix



[Koh & Liang, 2017]

Linear Approximations for Data Attribution

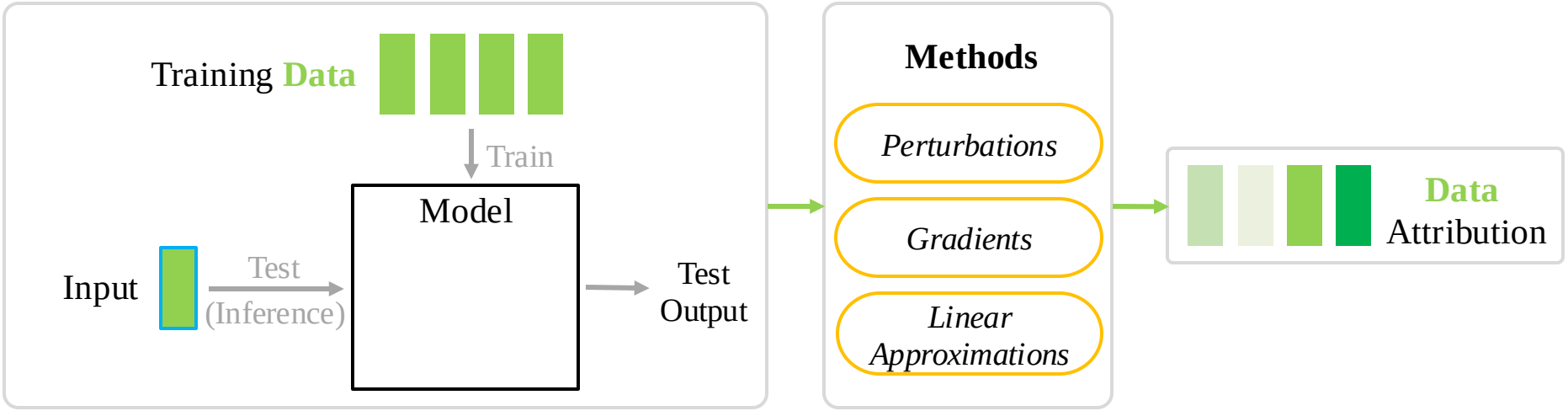
- The Datamodel
  - Skip model training, directly predict model outputs from training data
  - Collect counterfactual data for training



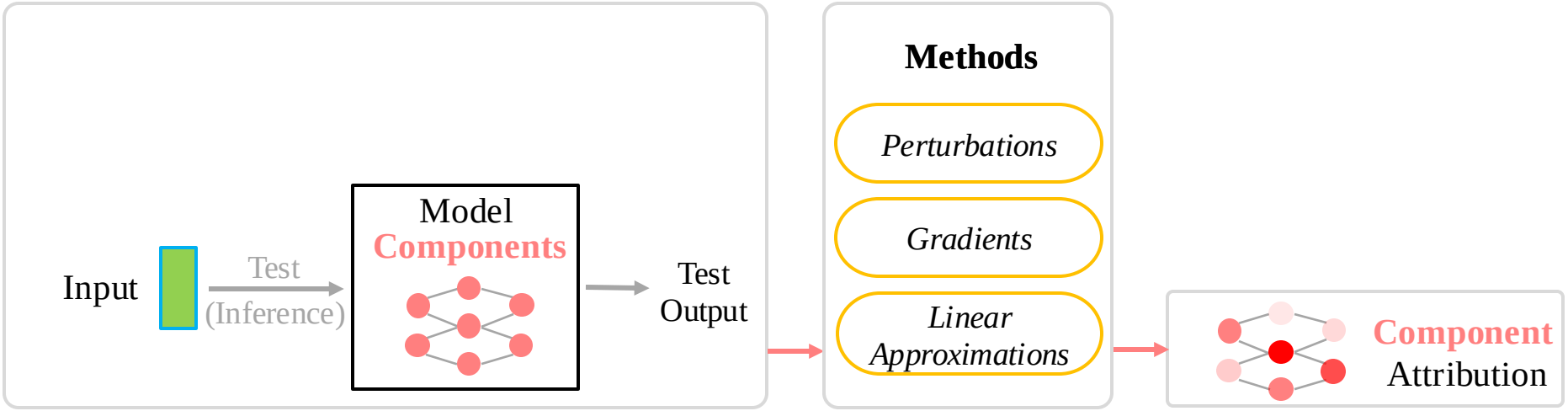
$$g(z) = w^T z + b \qquad z = \{1, 0\}^n \qquad w \rightarrow \psi_j(x)$$

[Ilyas et al., 2022]

Data Attribution Methods



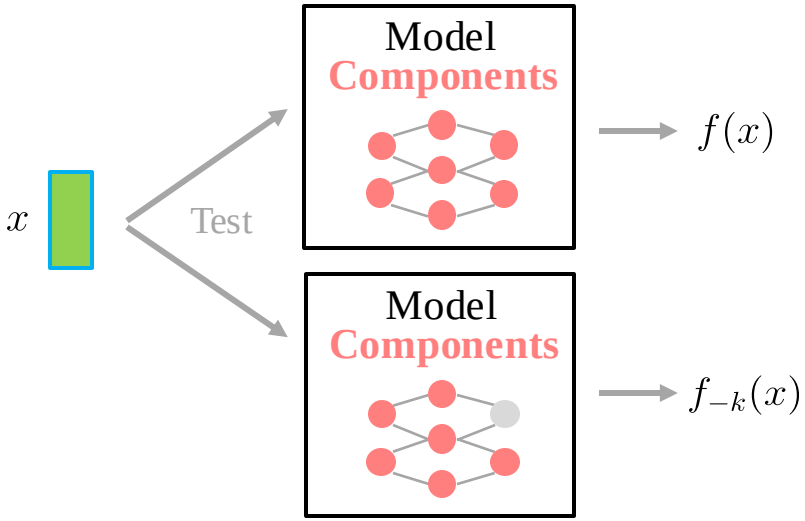
How About Component Attribution?



Perturbation-Based Component Attribution

- Casual mediation analysis
  - Perturb (replace with dummy values) components and observe output changes
  - Find the casual components, e.g., subnetworks

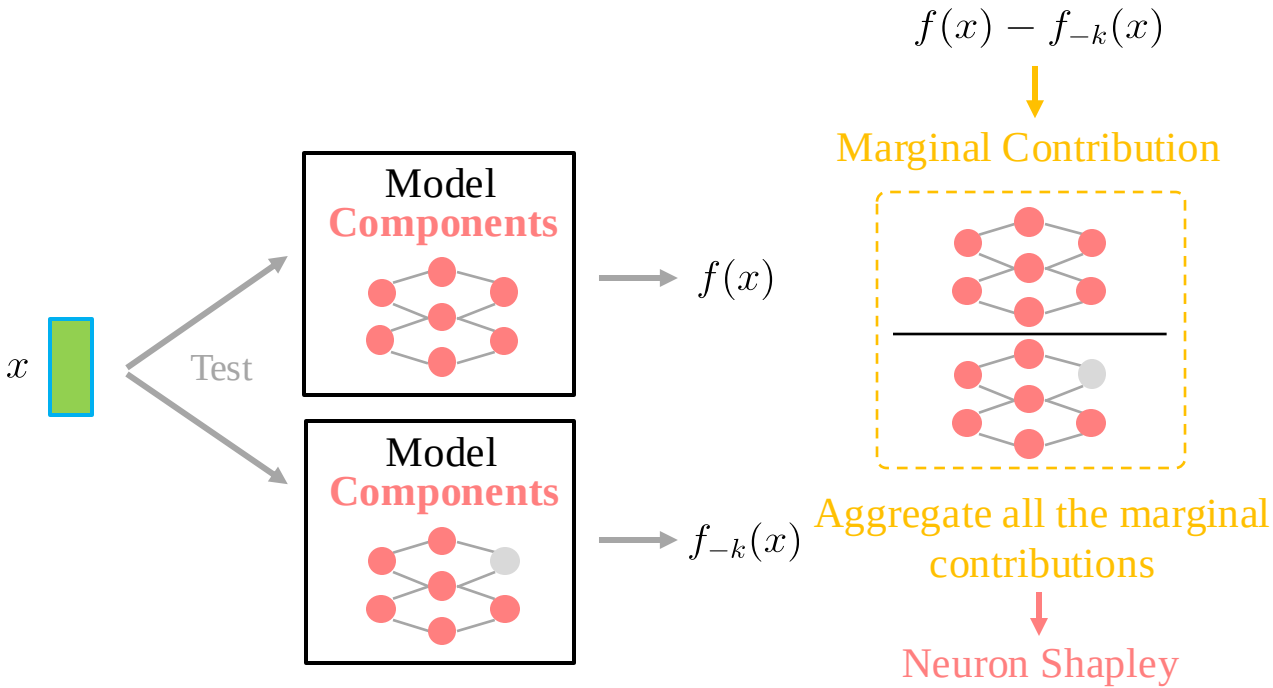
$$f(x) - f_{-k}(x) \rightarrow \gamma_k(x)$$



[Pearl, 2022]

Perturbation-Based Component Attribution

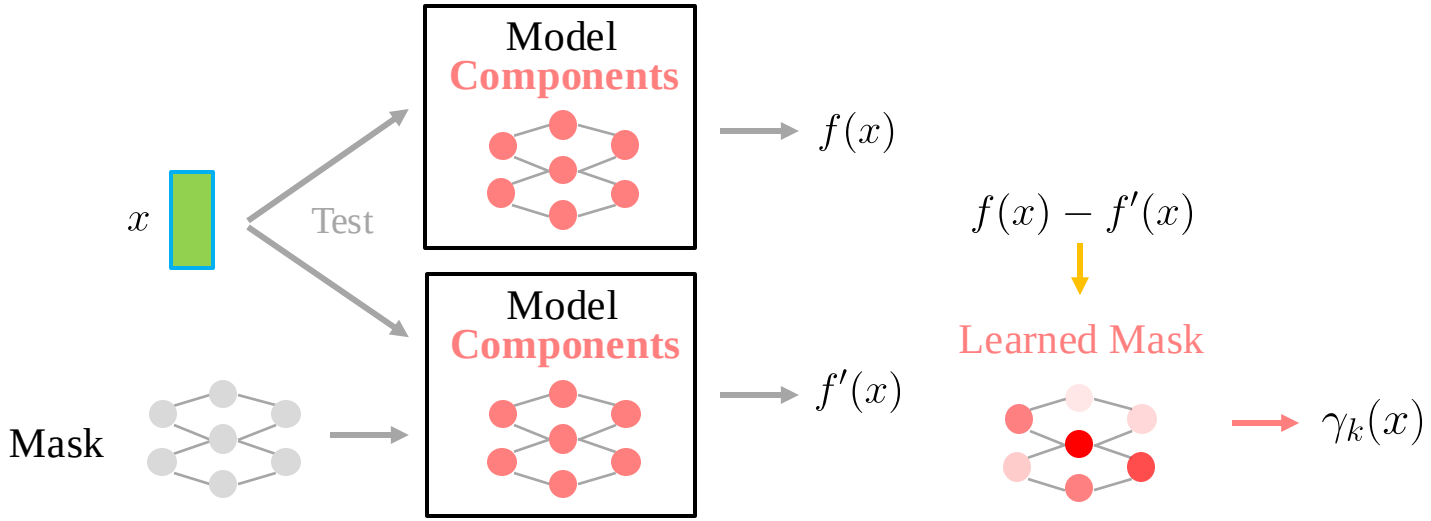
- Game-Theoretic Perturbation
  - Capture component interactions



[Ghorbani & Zou, 2020]

Perturbation-Based Component Attribution

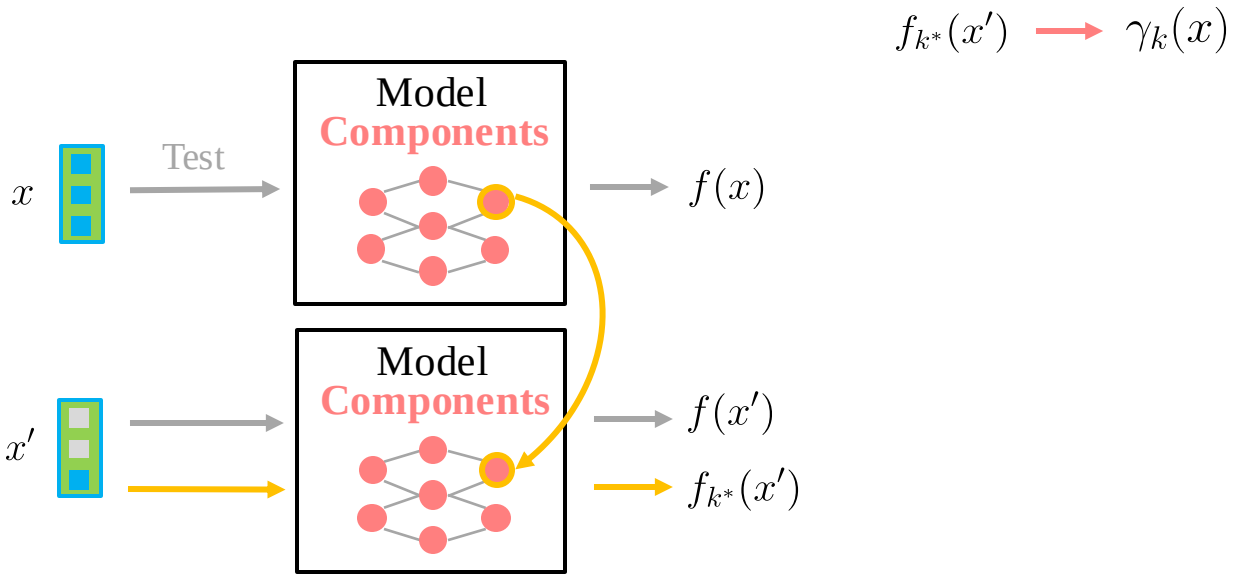
- Mask Learning and Subnetwork Probing
  - Use a learnable mask for selecting neurons



[Cao, 2021]

Perturbation-Based Component Attribution

- Casual mediation analysis
  - Perturb inputs
  - A three-run patching paradigm: original, perturbed, and restored



[Meng et al., 2022]

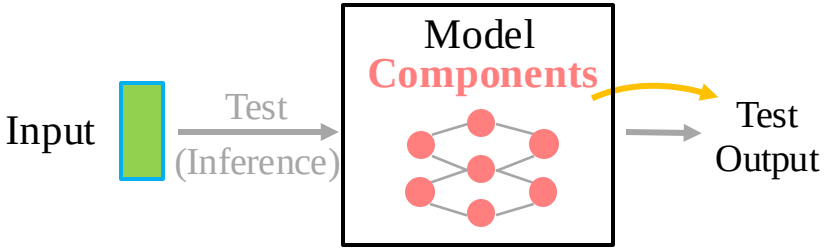
Gradient-Based Component Attribution

- Gradients for approximating the patching paradigm
- An underexplored area

[Nanda, 2023]

Linear Approximations for Component Attribution

- A linear model to directly predict test output with altered components
  - Collect counterfactual data for training



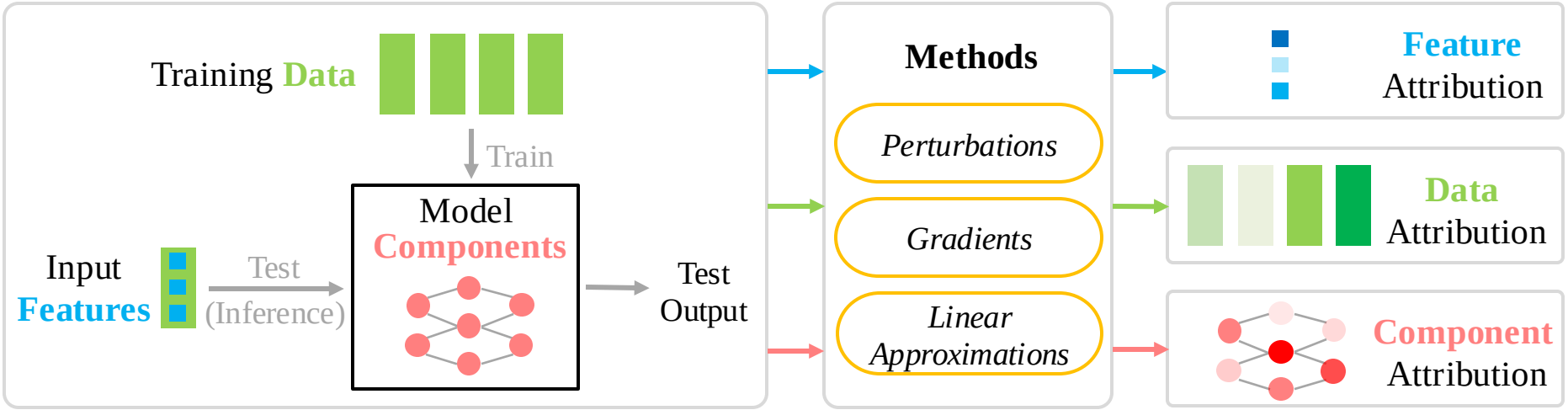
$$g(c) = w^T c + b \qquad c = \{1, 0\}^m \qquad w \rightarrow \gamma_k(x)$$

[Shah et al., 2024]

Outline

- The Attribution Problem
- Attribution Methods
- A Unified View of Attribution
- Cross-Aspect Innovation and Connections to Other Areas of AI

A Unified View of Attribution



Outline

- The Attribution Problem
- Attribution Methods
- A Unified View of Attribution
- Cross-Aspect Innovation and Connections to Other Areas of AI

Cross-Aspect Innovation

- Advanced gradient-based method for component attribution
  - Utilize the Hessian matrix for better approximation with second order information

Cross-Aspect Innovation

- A holistic view of attribute to multiple perspectives
  - A specific model behavior may be explained in terms of features, data, and components jointly

Cross-Aspect Innovation

- A theoretical unification
  - A framework in terms of local function approximation for feature attribution
  - Generalize to data and component attribution?

Table 3. Existing methods perform local function approximation of a black-box model f using the interpretable model class $\mathcal{G}$ of linear models where $g(x) = w^\top x$ over a local neighbourhood $\mathcal{Z}$ around point x based on a loss function ℓ. $\odot$ indicates element-wise multiplication. (Table reproduced from Han et al. (2022)).

Techniques	**Attribution Methods**	**Local Neighborhood $\mathcal{Z}$ around $x^{\{0\}}$**	**Loss Function ℓ**
Perturbations	Occlusion	$x \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Random one-hot vectors	Squared Error
	KernelSHAP	$x^{\{0\}} \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Shapley kernel	Squared Error
Gradients	Vanilla Gradients	$x + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2), \sigma \to 0$	Gradient Matching
	Integrated Gradients	$\xi x;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(0, 1)$	Gradient Matching
	Gradients $\times$ Input	$\xi x;\ \xi(\in \mathbb{R}) \sim \text{Uniform}(a, 1), a \to 1$	Gradient Matching
	SmoothGrad	$x + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Gradient Matching
Linear Approximations	LIME	$x \odot \xi;\ \xi(\in \{0,1\}^d) \sim$ Exponential kernel	Squared Error
	C-LIME	$x + \xi;\ \xi(\in \mathbb{R}^d) \sim \text{Normal}(0, \sigma^2)$	Squared Error

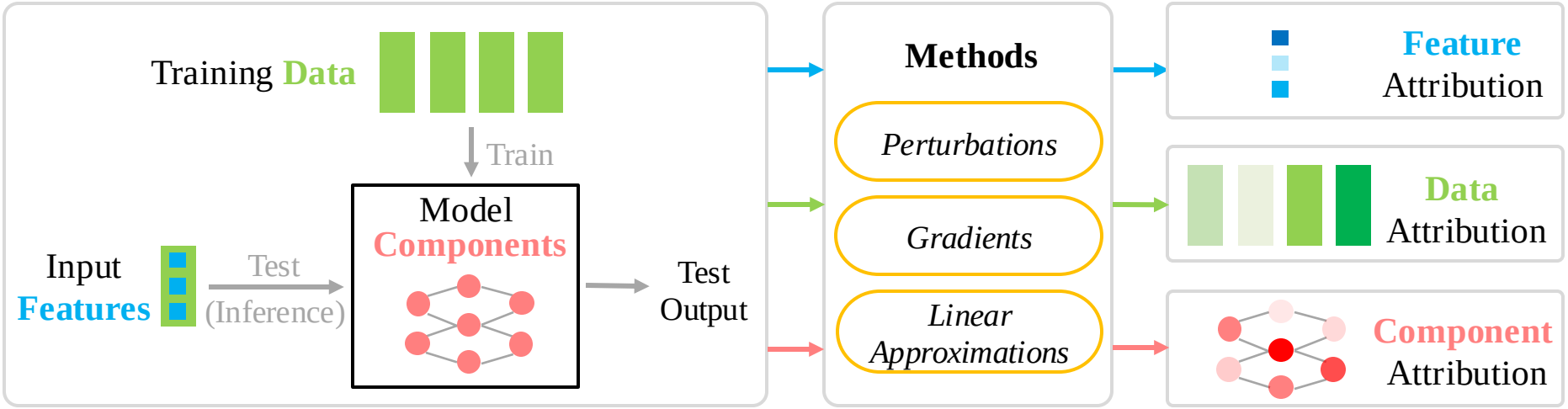
[Han et al., 2022]

Connections to Other Areas of AI

- Model editing
  - Goal: precisely edit model knowledge without retraining
  - Application: correct model mistakes, analogous to fixing bugs in software
  - Connections:
    - Better attribution implies better editing
    - Locate influencial components, spurious feature-label correlations, and problematic training data points can provide insights for editing

Summary

- The attribution problem and three aspects of attribution
- A unified view of attribution
- Cross-aspect innovation and potential theoretical unification
- Connections to model editing



References

- **Zhang, S., Han, T., Bhalla, U., & Lakkaraju, H. (2025). Building Bridges, Not Walls–Advancing Interpretability by Unifying Feature, Data, and Model Component Attribution.**
- Taigman, Y., Yang, M., Ranzato, M. A., & Wolf, L. (2014). "Deepface: Closing the gap to human-level performance in face verification." Computer Vision and Pattern Recognition.
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). "Training language models to follow instructions with human feedback." Advances in Neural Information Processing Systems.
- Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., ... & Jumper, J. M. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature.
- Goldstein, A., Wang, H., Niekerken, L. et al. (2025). A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. Nat Hum Behav.

References

- Wachter, S., Mittelstadt, B., & C Russell. (2018) Counterfactual explanations without opening the black box: automated decisions and the gdpr. Harvard Journal of Law and Technology, 31(2): 841–887
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).
- Koh, P. W., & Liang, P. (2017, July). Understanding black-box predictions via influence functions. In *International conference on machine learning* (pp. 1885-1894). PMLR.
- Grosse, R., Bae, J., Anil, C., Elhage, N., Tamkin, A., Tajdini, A., ... & Bowman, S. R. (2023). Studying large language model generalization with influence functions. *arXiv preprint arXiv:2308.03296*.
- Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.
- Shapley, L. S. (1953). A value for n-person games.
- Dabkowski, P., & Gal, Y. (2017). Real time image saliency for black box classifiers. *Advances in neural information processing systems*, *30*.
- Petsiuk, V., Das, A., & Saenko, K. (2018). Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*.

References

- Simonyan, K., Vedaldi, A., & Zisserman, A. (2013) Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*
- Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018) Sanity checks for saliency maps. *Advances in neural information processing systems*, 31.
- Smilkov, D., Thorat, N., Kim, B., Viégas, F., & Wattenberg, M. (2017). Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*.
- Sundararajan, M., Taly, A., & Yan, Q. (2017, July). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.
- Santiago, F. (2020) Model interpretability — Making your model confesses: LIME. Medium,. https://santiagof.medium.com/model-interpretability-making-your-model-confess-lime-89db7f70a72b
- Cook, R. D., & Weisberg, S. (1982). Residuals and influence in regression.
- Ghorbani, A., & Zou, J. (2019, May). Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning* (pp. 2242-2251). PMLR.
- Charpiat, G., Girard, N., Felardos, L., & Tarabalka, Y. (2019). Input similarity from the neural network perspective. *Advances in Neural Information Processing Systems*, *32*.
- Pruthi, G., Liu, F., Kale, S., & Sundararajan, M. (2020). Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, *33*, 19920-19930.
- Ilyas, A., Park, S. M., Engstrom, L., Leclerc, G., & Madry, A. (2022). Datamodels: Predicting predictions from training data. *arXiv preprint arXiv:2202.00622*.

References

- Pearl, J. (2022). Direct and indirect effects. In *Probabilistic and causal inference: the works of Judea Pearl* (pp. 373-392).
- Ghorbani, A., & Zou, J. Y. (2020). Neuron shapley: Discovering the responsible neurons. *Advances in neural information processing systems*, *33*, 5922-5932.
- Cao, S., Sanh, V., & Rush, A. M. (2021). Low-complexity probing via finding subnetworks. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 960–966
- Meng, K., Bau, D., Andonian, A., & Belinkov, Y. (2022). Locating and editing factual associations in gpt. *Advances in neural information processing systems*, *35*, 17359-17372.
- Nanda, N. (2023). Attribution patching: Activation patching at industrial scale. *URL: https://www. neelnanda. io/mechanistic-interpretability/attribution-patching*.
- Shah, H., Ilyas, A., & Mądry, A. (2024, July). Decomposing and editing predictions by modeling model computation. In *Proceedings of the 41st International Conference on Machine Learning* (pp. 44244-44292).
- Han, T., Srinivas, S., & Lakkaraju, H. (2022). Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations. *Advances in neural information processing systems*, *35*, 5256-5268.

Q & A

Appendix

Notations

Notation	Description
$\mathcal{D}_{\text{train}}$	Training dataset $\{x^{(1)}, \cdots, x^{(n)}\}$
f_θ/ f	Model trained on $\mathcal{D}_{\text{train}}$, parameters θ may be omitted
c	Internal model components $\{c_1, \cdots, c_m\}$, definition is method-specific
x^{test}/x	Model input at test time for inference, superscript "test" may be omitted
$\phi_i(x)$	Attribution score of input feature x_i for model output $f(x)$
$\psi_j(x)$	Attribution score of training data point $x^{(j)}$ for model output $f(x)$
$\gamma_k(x)$	Attribution score of internal model component c_k for model output $f(x)$
g	Attribution function, which provides attribution scores for elements
$\mathcal{L}$	Loss function for training the model f
ℓ	Loss function for learning the attribution function g

Methods Summary

Table 1: A summary of representative feature, data, and component attribution methods classified into three methodological categories demonstrating our unified view.

	Method	Feature Attribution	Data Attribution	Component Attribution
Perturb	Direct	Occlusions [Zeiler and Fergus, 2014] RISE [Petsiuk, 2018]	LOO [Cook and Weisberg, 1982]	Causal Tracing [Meng et al., 2022] Path Patching [Wang et al., 2022] Vig et al. [2020] Bau et al. [2020] ACDC [Conmy et al., 2023]
	Game-Theoretic (Shapley)	SHAP [Lundberg and Lee, 2017]	Data Shapley [Ghorbani and Zou, 2019] TMC Shapley [Ghorbani and Zou, 2019] KNN Shapley [Jia et al., 2019] Beta Shapley [Kwon and Zou, 2022]	Neuron Shapley [Ghorbani and Zou, 2020]
	Game-Theoretic (Others)	STII [Dhamdhere et al., 2019] BII [Patel et al., 2021] Core Value [Yan and Procaccia, 2021] Myerson Value [Chen et al., 2018b] HN Value [Zhang et al., 2022]	Data Banzhaf [Wang and Jia, 2023]	–
	Mask Learning	Dabkowski and Gal [2017] L2X [Chen et al., 2018a]	–	Csordás et al. [2020] Subnetwork Pruning [Cao et al., 2021]
Gradient	First-Order	Vanilla Gradients [Simonyan et al., 2013] Gradient × Input [Shrikumar et al., 2017] SmoothGrad [Smilkov et al., 2017] GBP [Springenberg et al., 2014] Grad-CAM [Selvaraju et al., 2016]	GradDot/GradCos [Pruthi et al., 2020]	Attribution Patching [Nanda, 2023] EAP [Syed et al., 2023]
	Second-Order (Hessian/IF)	Integrated Hessian [Janizek et al., 2021]	IF [Koh and Liang, 2017] FastIF [Guo et al., 2021] Arnoldi IF [Schioppa et al., 2022] EK-FAC [Grosse et al., 2023] RelateIF [Barshan et al., 2020]	–
	Tracing Path	Integrated Grad [Sundararajan et al., 2017]	TracIn [Pruthi et al., 2020] SGD-Influence [Hara et al., 2019] SOURCE [Bae et al., 2024]	Attribution Path Patching [Nanda, 2023]
Linear		LIME [Ribeiro et al., 2016] C-LIME [Agarwal et al., 2021]	Datamodels [Ilyas et al., 2022] TRAK [Park et al., 2023]	COAR [Shah et al., 2024]

Abstract

As AI systems continue to grow in scale and complexity, understanding their behaviors becomes increasingly critical. In this talk, I will explore three complementary perspectives on AI interpretability: feature attribution, which analyzes how input features influence predictions; data attribution, which traces the impact of training data on model behavior; and component attribution, which investigates the role of internal model components such as neurons and layers. I will cover the core ideas and recent advances in each of these areas, including perturbation-based methods, gradient-based approaches, and linear approximation techniques. In addition, I will highlight connections among the three attribution aspects and propose a unified view that reveals their shared foundations. This talk aims to provide a structured overview of the current landscape of attribution methods, equipping researchers and practitioners with practical tools and conceptual frameworks to interpret AI systems from multiple perspectives—and to apply these insights in real-world contexts such as model debugging, editing, and regulation.