Inherent Trade-offs with the Local Explanations Paradigm.

Julius Adebayo
MIT
Simons Workshop on Emerging Challenges in Deep Learning
August 8, 2019.

“Explanations”

- **Interpret** means “to explain or to present in understandable terms” to “a human” [Doshi-Velez & Kim, 2017].

"Explanations"

- **Interpret** means "to explain or to present in understandable terms" to "a human" [Doshi-Velez & Kim, 2017].

- "Explanations **[can]** sort themselves into several distinct types corresponding to patterns of causation, content domains, and explanatory stances, all of which have cognitive consequences" [Keil, 2011].

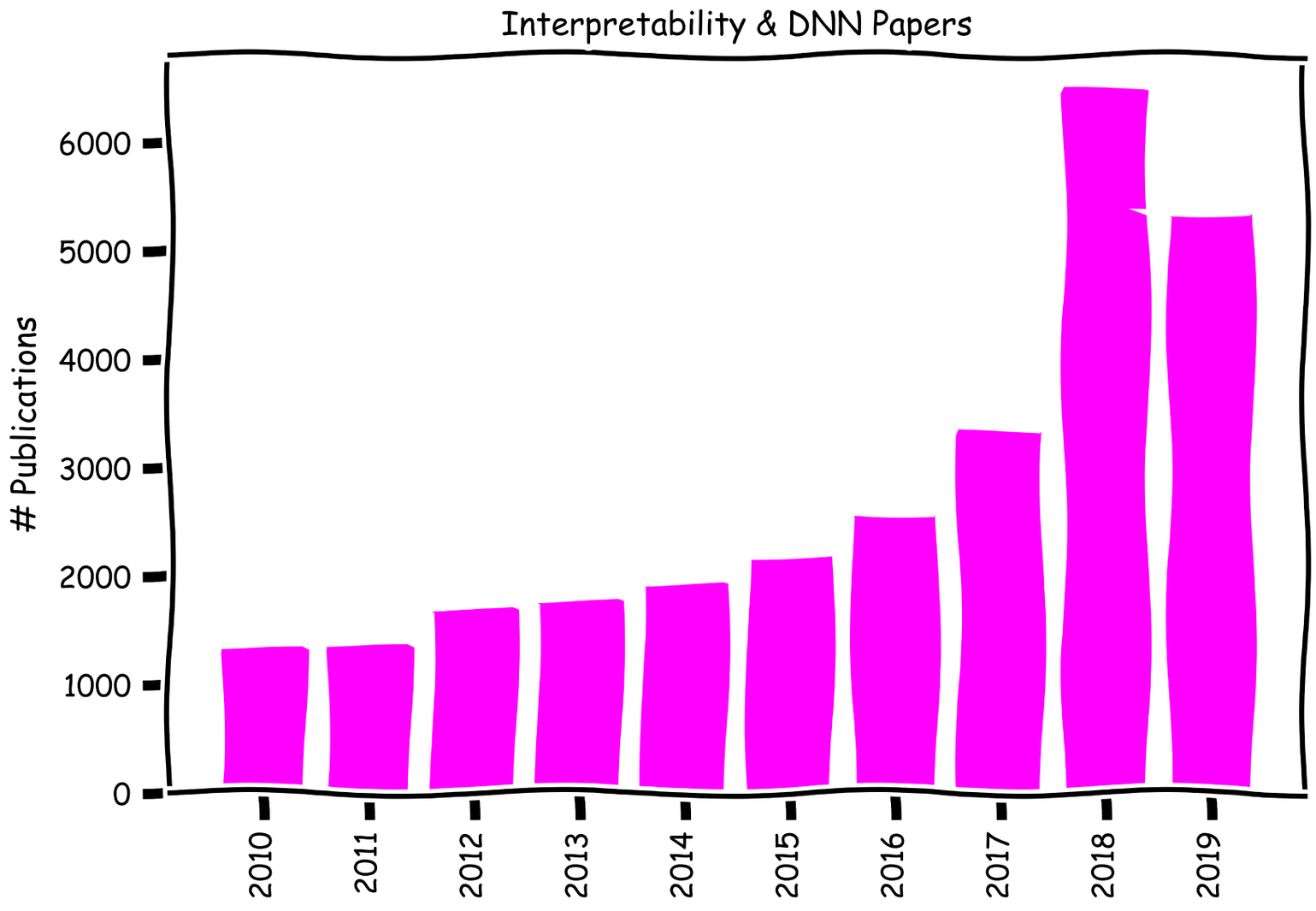
"Explanations"

- **Interpret** means "to explain or to present in understandable terms" to "a human" [Doshi-Velez & Kim, 2017].

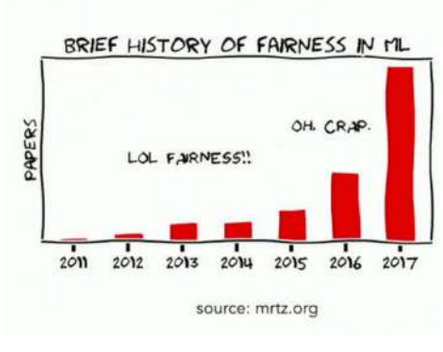
- "Explanations **[can]** sort themselves into several distinct types corresponding to patterns of causation, content domains, and explanatory stances, all of which have cognitive consequences" [Keil, 2011].

- An 'artifact', derived from a 'model', with the goal to provide 'insights' into the 'factors' most 'relevant' to the 'model' for an end-user.

“Interpretability” AND “Neural Network”



Inspired by the ‘Fairness’ version from mrtz.



Some Motivation

[Challenges for Transparency, Weller 2017, & Doshi-Velez & Kim, 2017]

- Developer/Researcher: Model Debugging.
- Safety concerns.
- Ethical concerns.
- Trust: Satiate ‘societal’ need for reasoning to trust an automated system learned from data.

Goals: Model Debugging

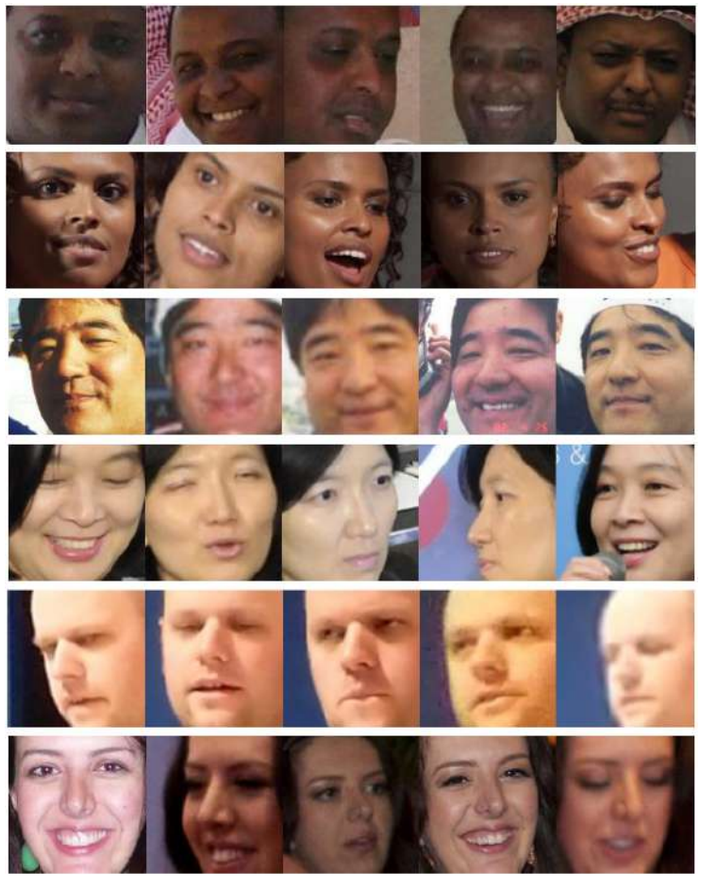
- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

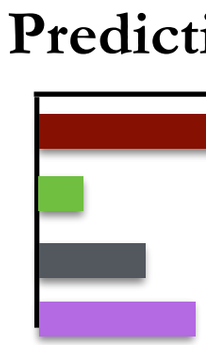
(a) Husky classified as wolf (b) Explanation

[Ribeiro+ 2016]

Systematic Subgroup Errors

[Morales+ 2019]

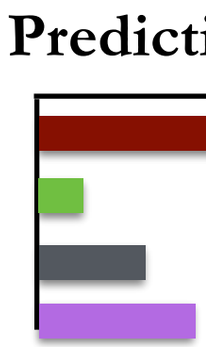




Predictions

Systematic Subgroup Errors

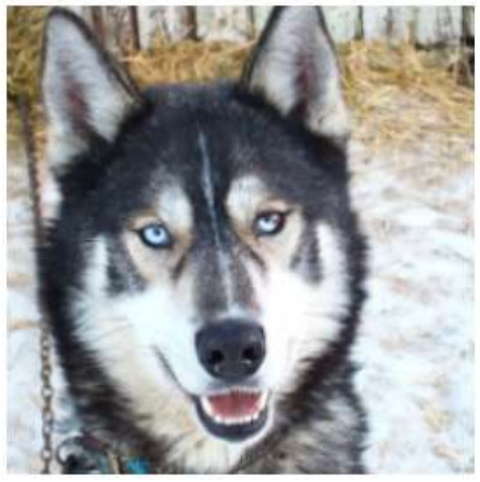
[Morales+ 2019]

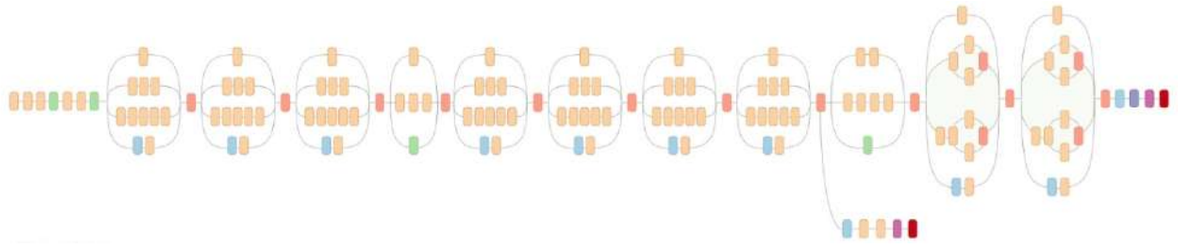


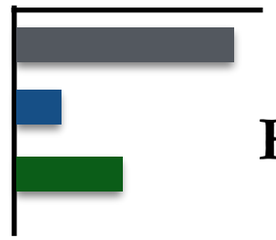
- Reveal subsets (**non-intuitive**) of the data for which the model has bad performance. This can be due to data labeling errors or others.

Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.



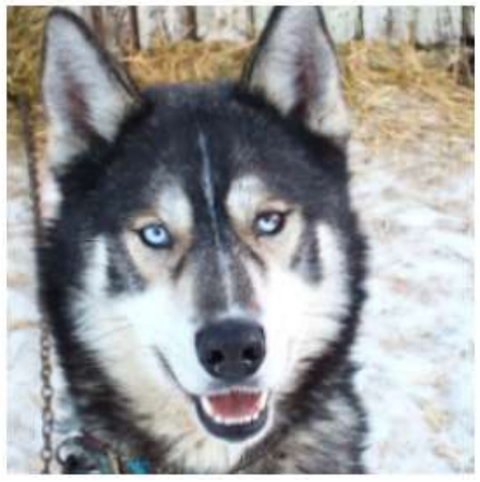


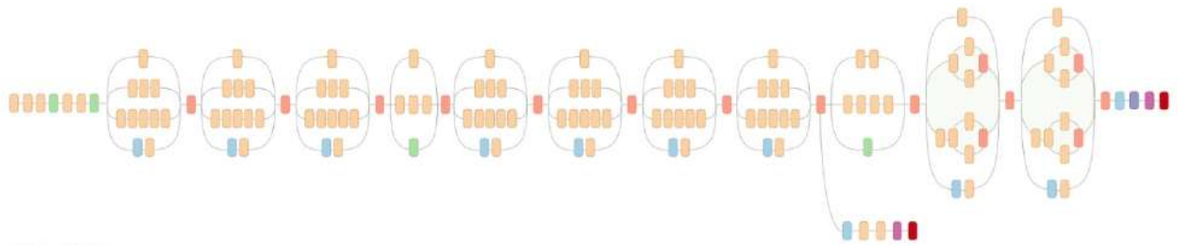


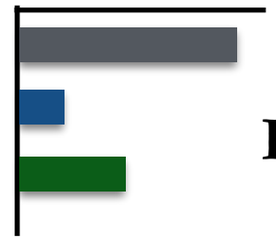
Husky

Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.

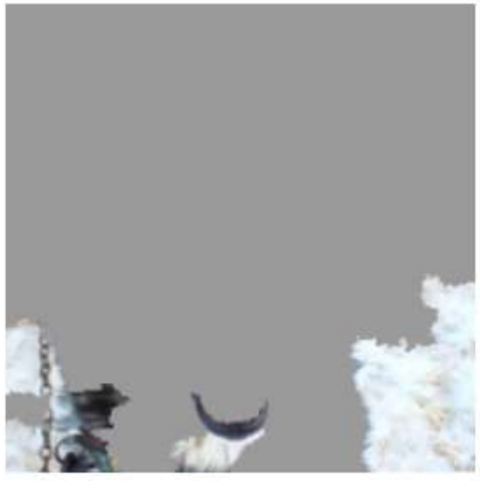






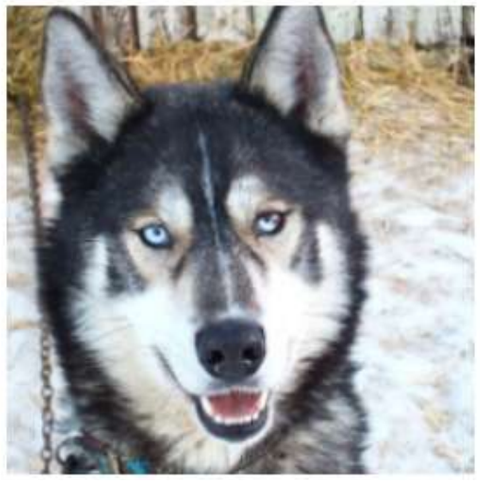
Husky

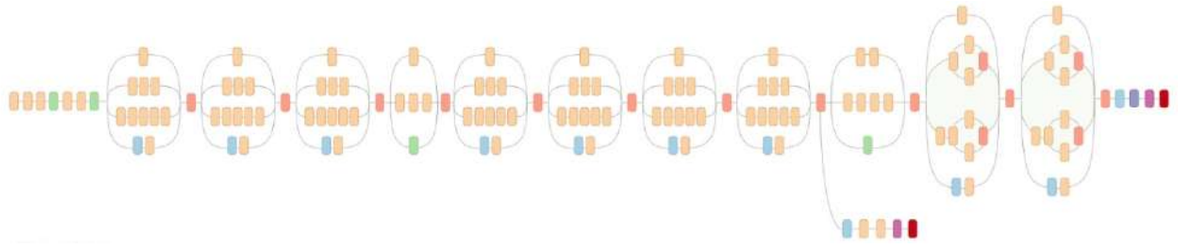
Explanation

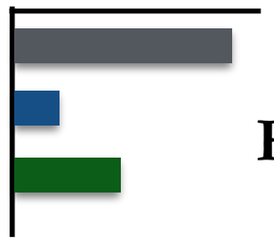


Promise of Explanations

- **Model Debugging**: reveal spurious correlations or the kinds of inputs that a model is most likely to have undesirable performance.



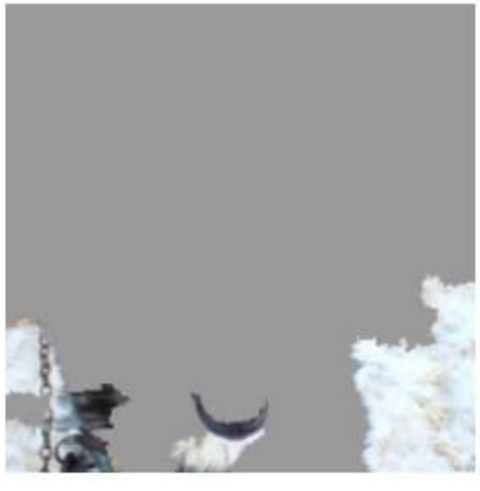




Husky

Fix

Explanation



‘Interpretability’ vs ‘Explainability’

Interpretability: Constrained Model Class.

Falling Rule Lists

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density $\geq$ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age $\geq$ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Table 1: Falling rule list for mammographic mass dataset.

[Wang+ 2015]

‘Interpretability’ vs ‘Explainability’

Interpretability: Constrained Model Class.

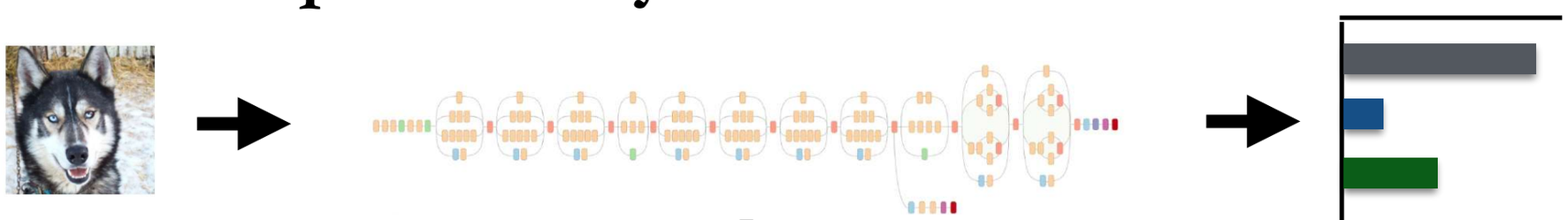
Falling Rule Lists

	Conditions		Probability	Support
IF	IrregularShape AND Age $\geq$ 60	THEN malignancy risk is	85.22%	230
ELSE IF	SpiculatedMargin AND Age $\geq$ 45	THEN malignancy risk is	78.13%	64
ELSE IF	IllDefinedMargin AND Age $\geq$ 60	THEN malignancy risk is	69.23%	39
ELSE IF	IrregularShape	THEN malignancy risk is	63.40%	153
ELSE IF	LobularShape AND Density $\geq$ 2	THEN malignancy risk is	39.68%	63
ELSE IF	RoundShape AND Age $\geq$ 60	THEN malignancy risk is	26.09%	46
ELSE		THEN malignancy risk is	10.38%	366

Table 1: Falling rule list for mammographic mass dataset.

[Wang+ 2015]

Post-Hoc Explainability.



Explanation

[Ribeiro+ 2015]

Focus

- This talk will focus exclusively on post-hoc explanations.
- Post-hoc explanations ‘purport’ to provide flexibility for the model developer/designer.

Focus

- This talk will focus exclusively on post-hoc explanations.

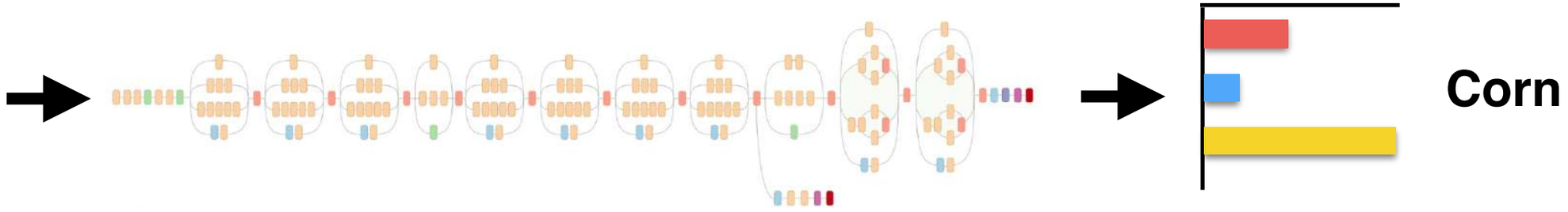
- Post-hoc explanations 'purport' to provide flexibility for the model developer/designer.

Perhaps a questionable thing to do!

Please Stop Explaining Black Box Models for High-Stakes Decisions

Cynthia Rudin
Duke University
cynthia@cs.duke.edu

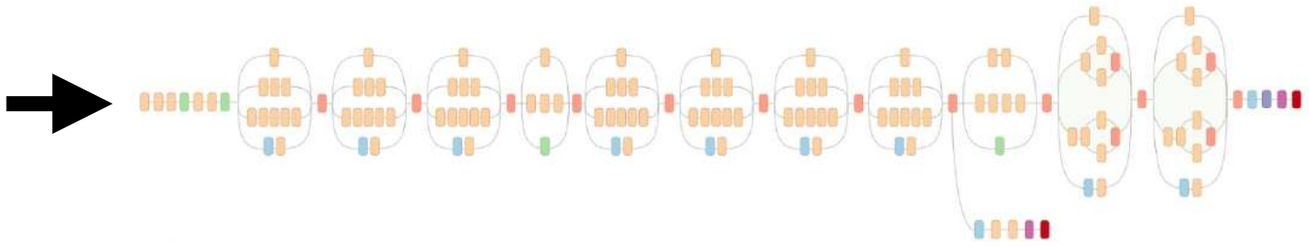
Saliency/Attribution Maps



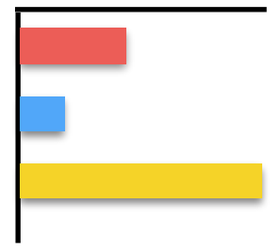
$$S : \mathbb{R}^d \rightarrow \mathbb{R}^C$$

d = input dimension
C = number of output classes

Saliency/Attribution Maps



Predictions

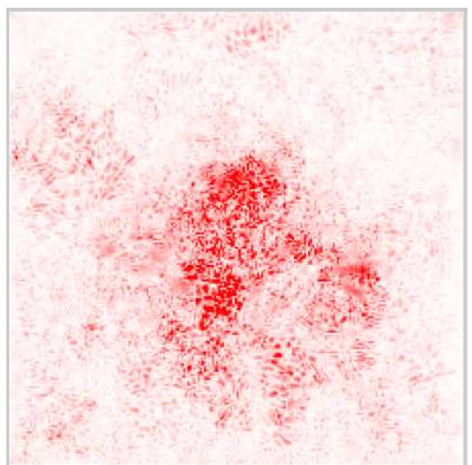


Corn

Explanation

Gradient

$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$

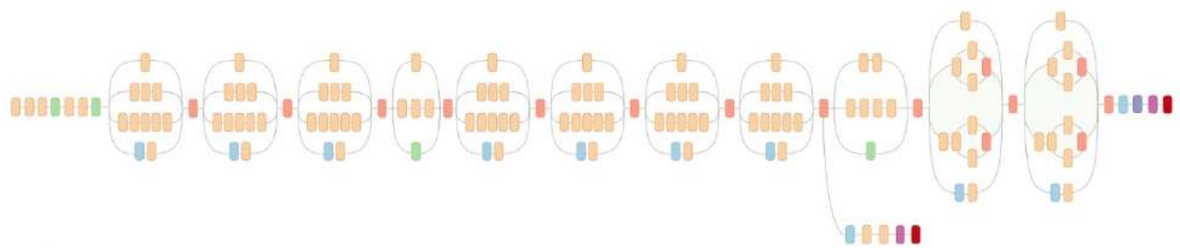


$$E : \mathbb{R}^d \rightarrow \mathbb{R}^d$$

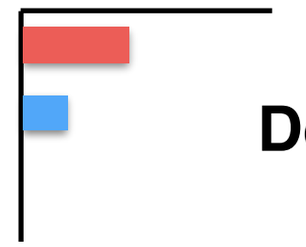
[SVZ'13]

‘Examples’/‘Prototypes’

Test Point



Predictions



Dog

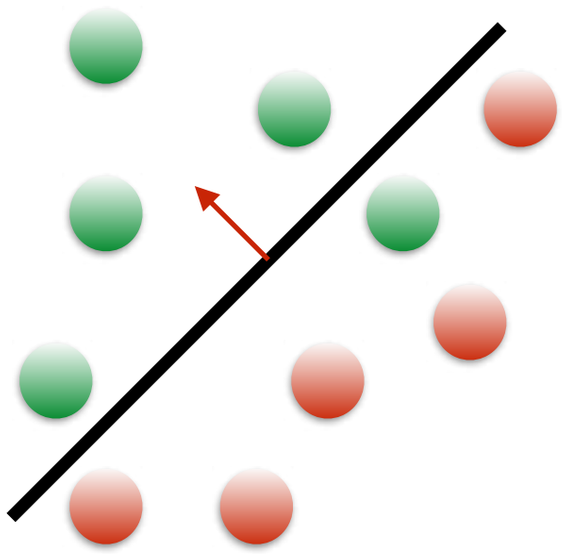
Training Points

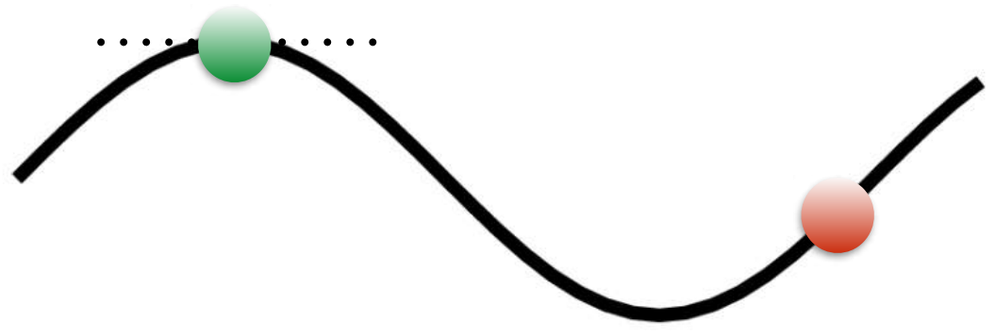
[Koh & Liang 2017, Yeh, 2018, …]

"Local Explanations"

- **Local Explanations:** focus on the behavior of the model around a **single point.**

- **Why is this desirable?**

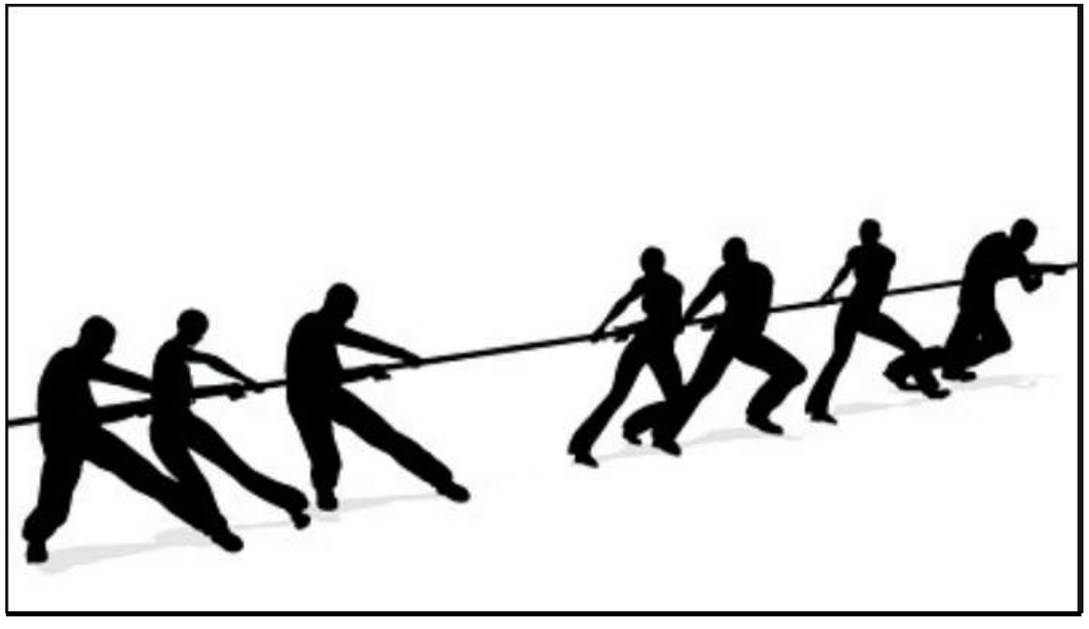




Key Takeaways

- Difficult to assess quality and model fidelity of local explanations.
- Conjecture: local explanations seem to require significant privacy tradeoffs.

Local Explanations



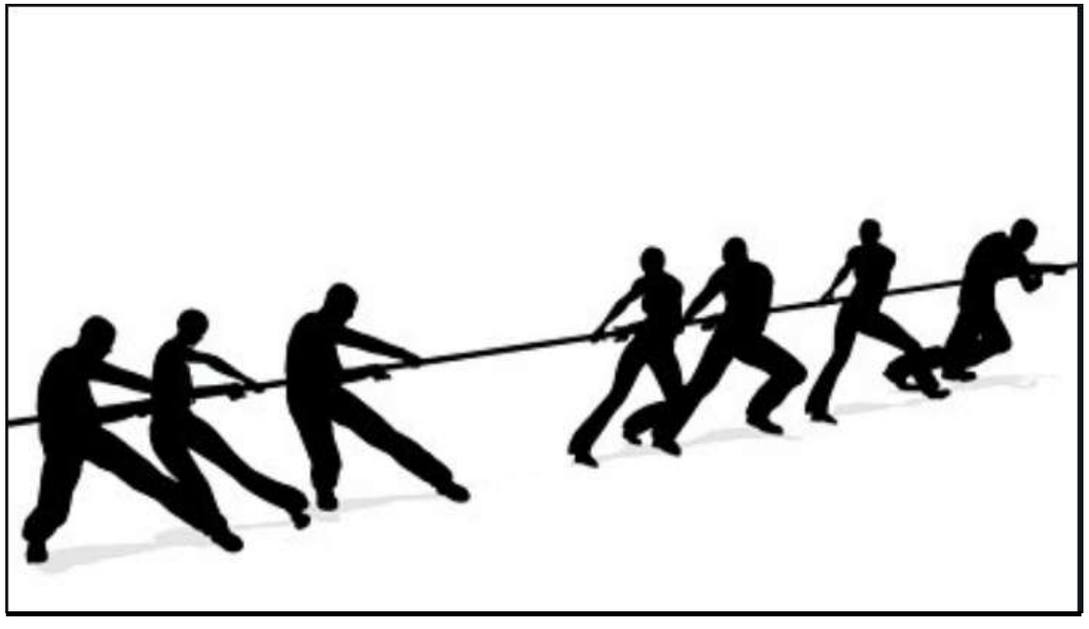
Model & Data Privacy

Key Takeaways

- **Difficult to assess quality and model fidelity of local explanations.**

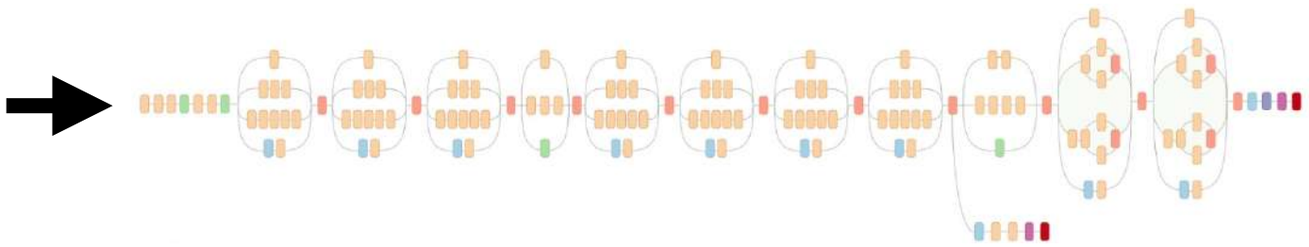
- Conjecture: local explanations seem to require significant privacy tradeoffs.

Local Explanations

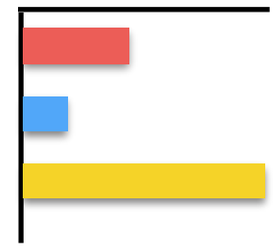


Model & Data Privacy

Gradient/Sensitivity Map



Predictions

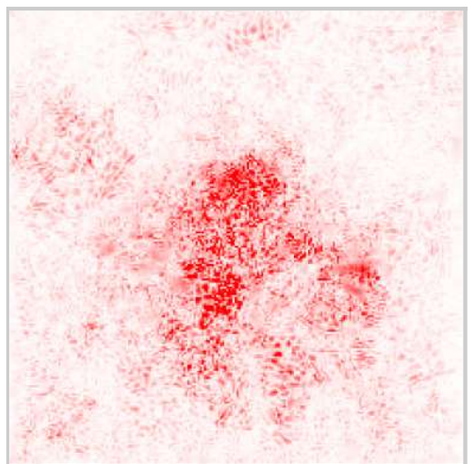


Corn

Explanation

Gradient

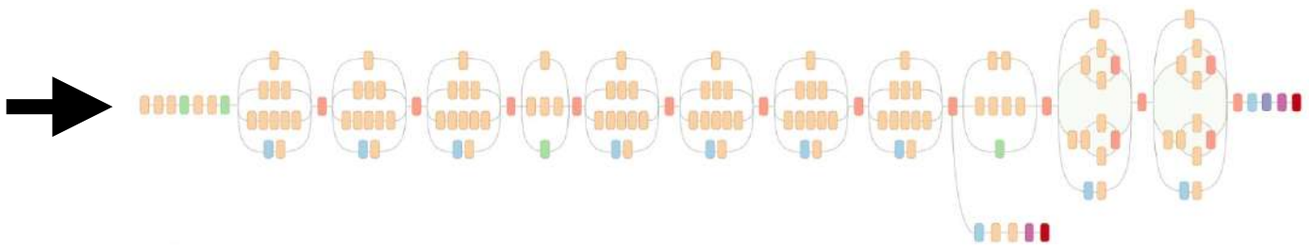
$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$



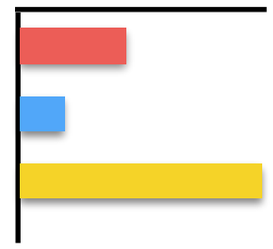
$$E : \mathbb{R}^d \rightarrow \mathbb{R}^d$$

[SVZ'13]

Gradient/Sensitivity Map



Predictions

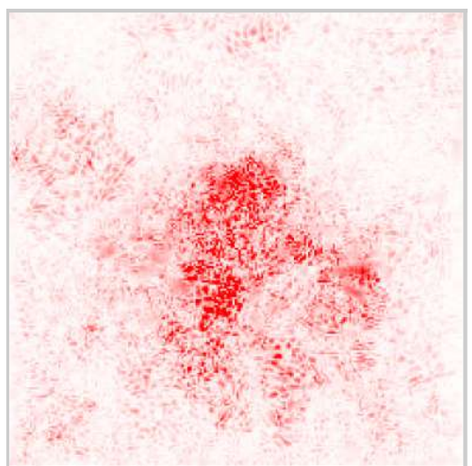


Corn

Explanation

Gradient

$$E_{grad}(x) = \frac{\partial S_i}{\partial x}$$



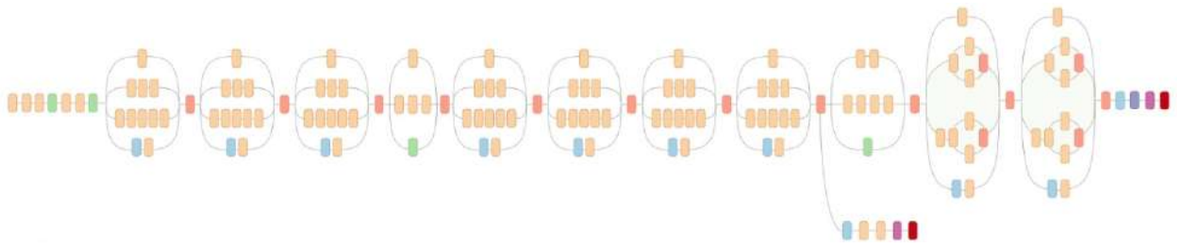
$$E : \mathbb{R}^d \rightarrow \mathbb{R}^d$$

[SVZ'13]

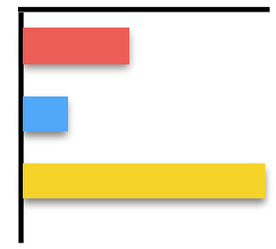
```
self.gradients_node = tf.gradients(y, x)[0]
```

[Google Pair Saliency Codebase]

Smoothgrad



Predictions

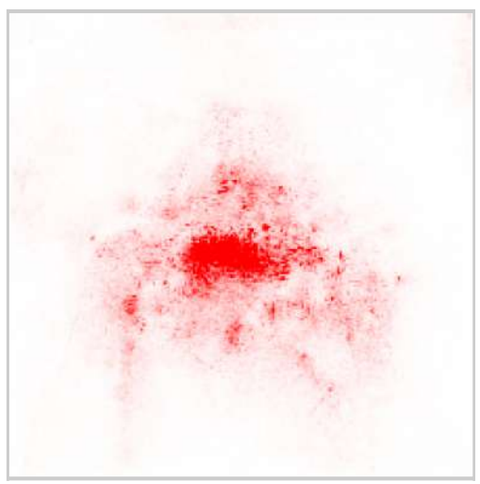


Corn

Explanation

SmoothGrad

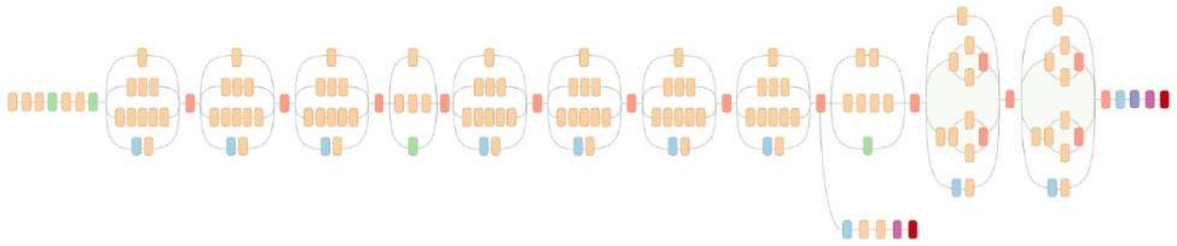
$$E_{\text{sg}}(x) = \frac{1}{N}\sum_{i=1}^{N} E(x + g_i),$$



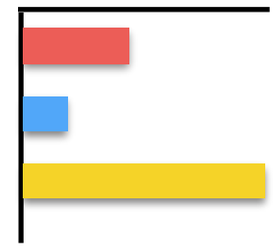
[STKVW'17]

$$E : \mathbb{R}^d \to \mathbb{R}^d$$

Integrated Gradients



Predictions

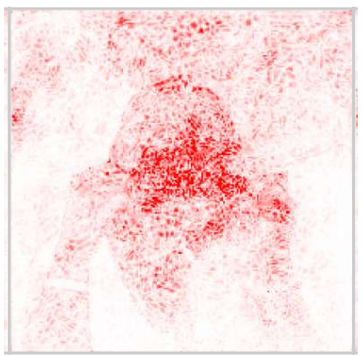


Corn

Explanation

$$E_{\mathrm{IG}}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x})}{\partial x} d\alpha$$

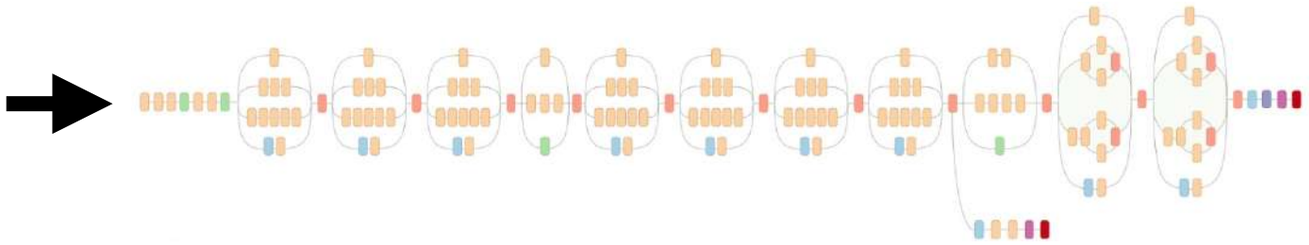
Integrated Gradients



[STY'17]

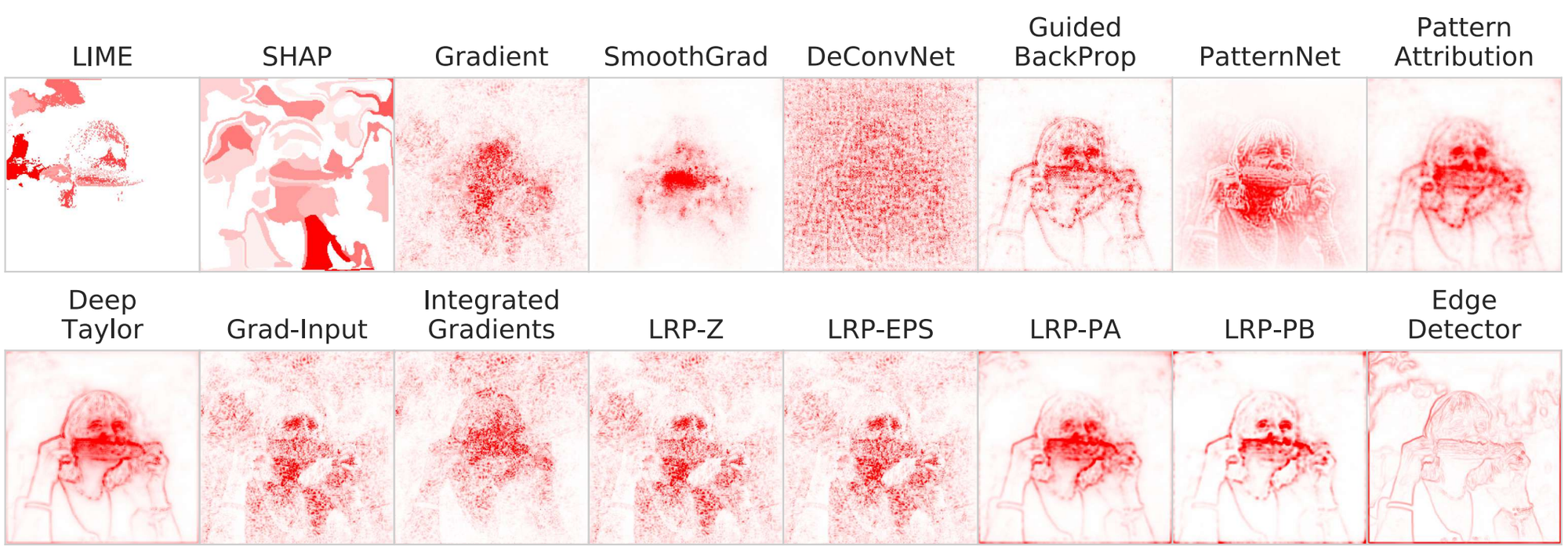
$$E : \mathbb{R}^d \rightarrow \mathbb{R}^d$$

Several More

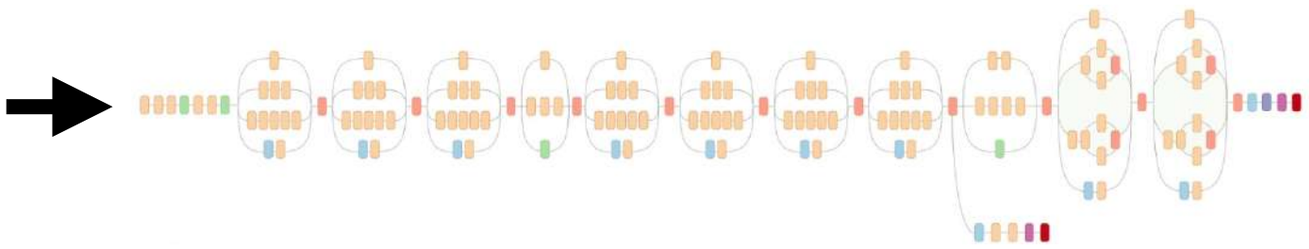


Corn

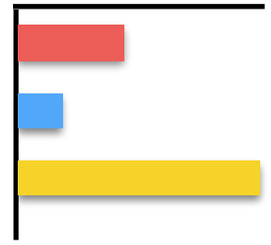
Explanation



Other Learned Kinds



Predictions



Corn

Explanation

[FV'17]

Non-Image Settings

Using attribution to decode binding mechanism in neural network models for chemistry

Kevin McCloskey[a,1], Ankur Taly[a,1], Federico Monti[a,b], Michael P. Brenner[a,c], and Lucy J. Colwell[a,d,1]

[a]Google Research, Mountain View, CA 94043; [b]Institute of Computational Science, Università della Svizzera Italiana, CH-6900 Lugano, Switzerland; [c]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138; and [d]Department of Chemistry, Cambridge University, Cambridge CB2 1EW, United Kingdom

Edited by Michael L. Klein, Institute of Computational Molecular Science, Temple University, Philadelphia, PA, and approved April 29, 2019 (received for review December 4, 2018)

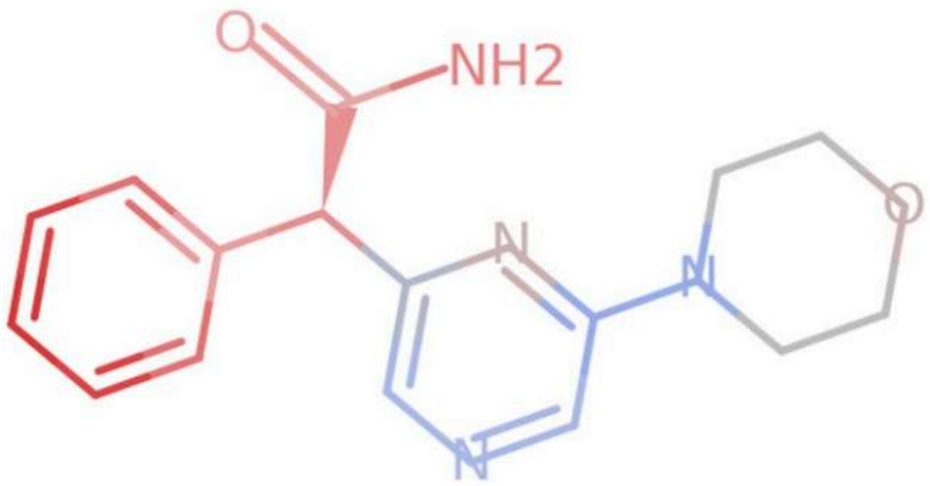
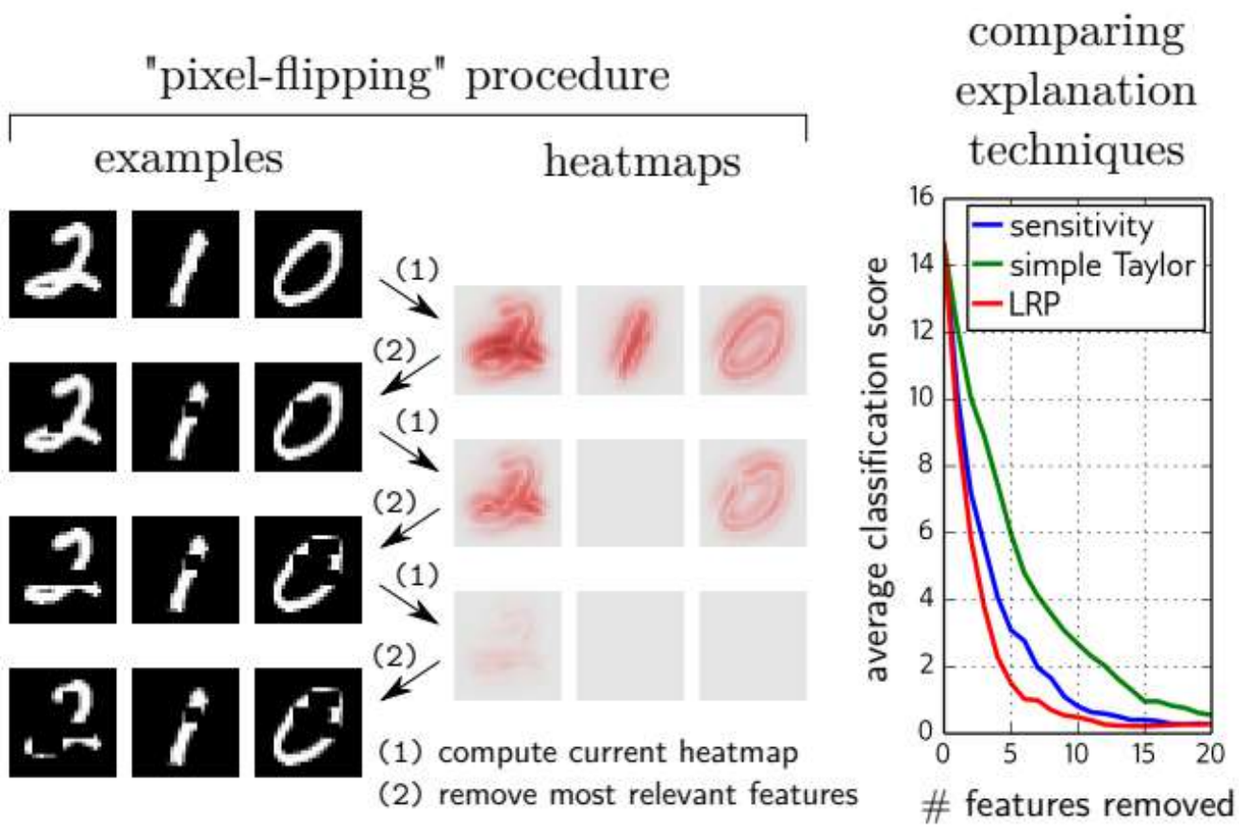


Fig. 1. An example of per-atom model attributions visualized for a molecule. Each atom is colored on a scale from red to blue in proportion to its attribution score, with red being the most positive and blue being the most negative.

Challenge 1: Assessment

- Input Perturbation.
- Localization error in an object localization task.
- Question: **can we design simple tests to 'sanity check' the model fidelity attribution maps?**



[Montavon+ 2017]

Sanity Checks for Saliency Maps

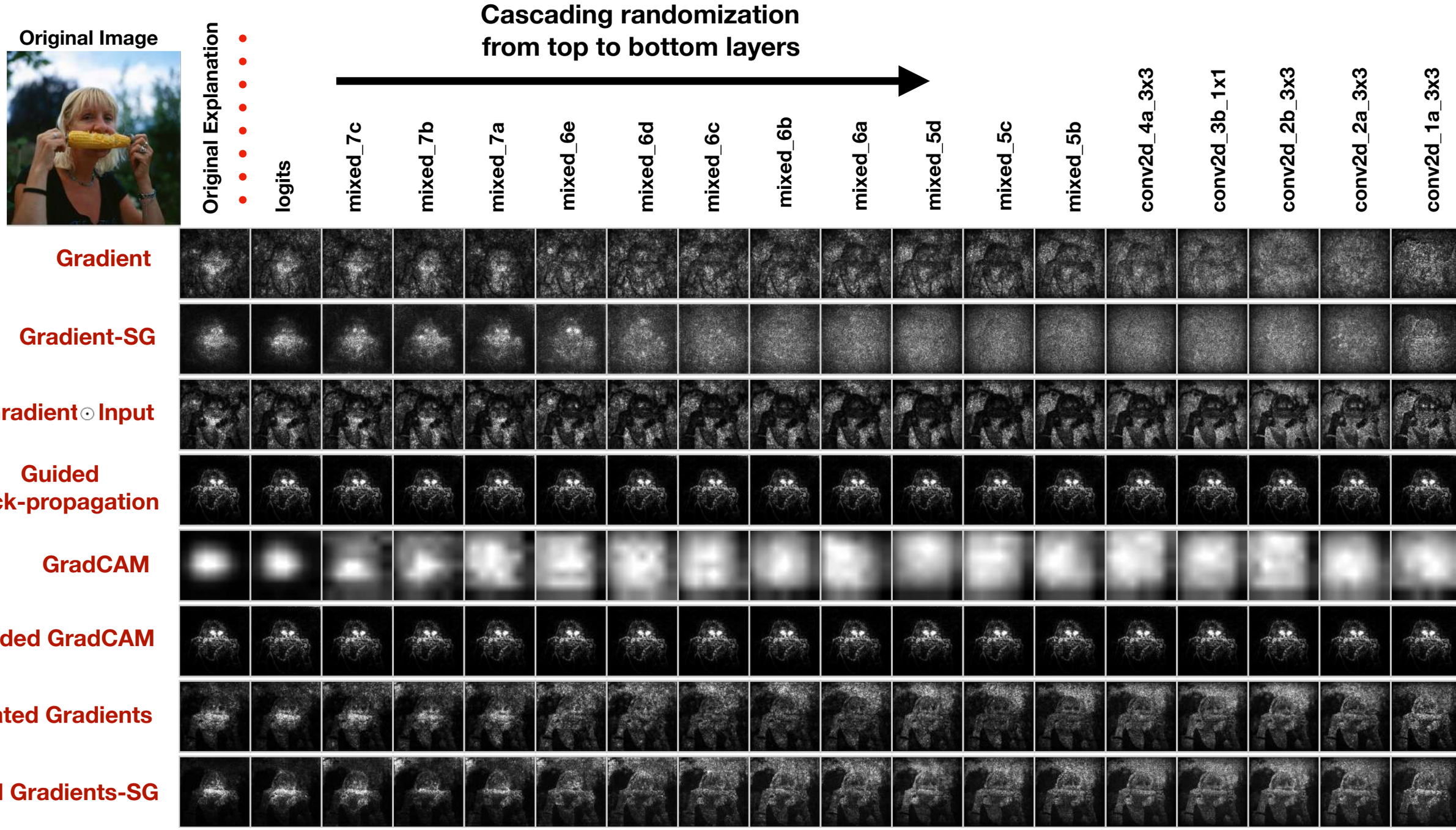
Joint work with





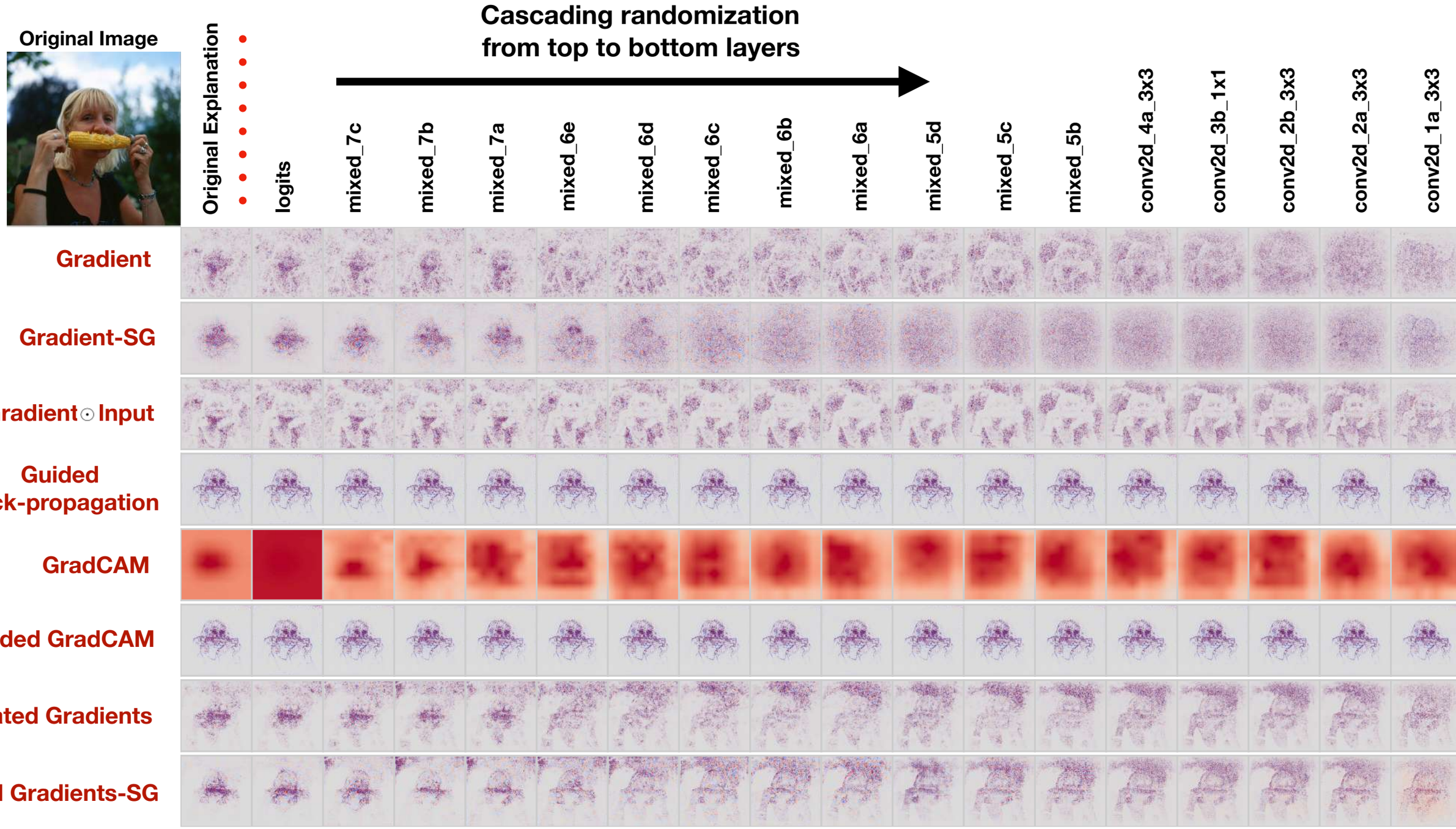
Sanity Check 1: Model Randomization

Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



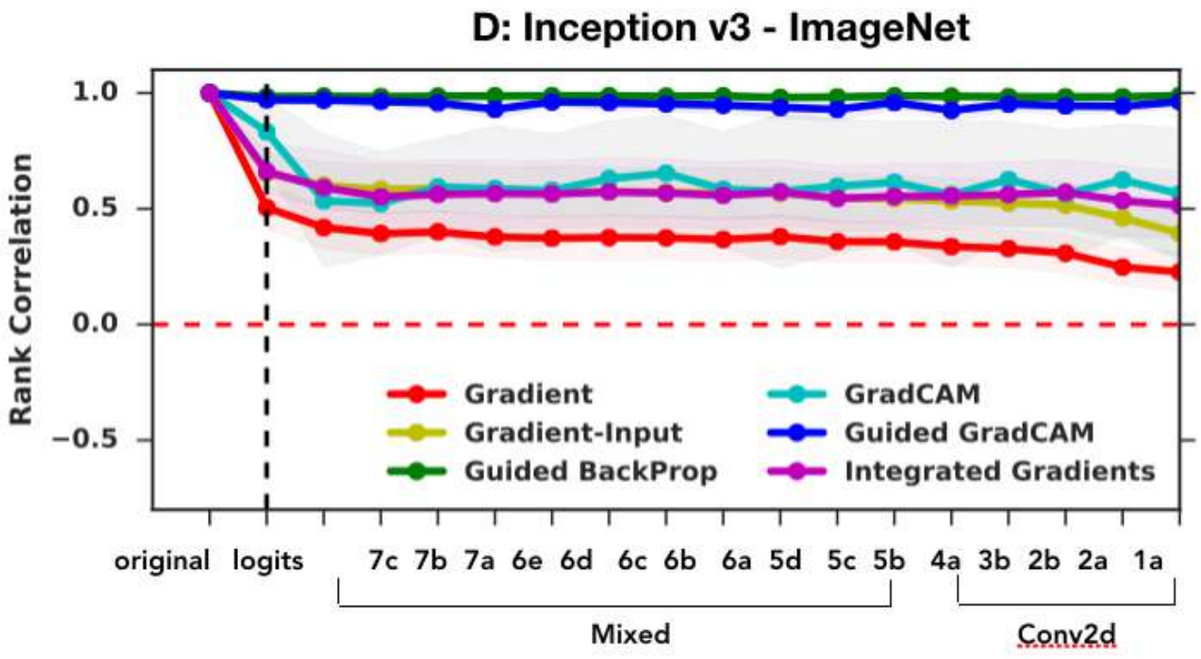
Sanity Check 1: Model Randomization

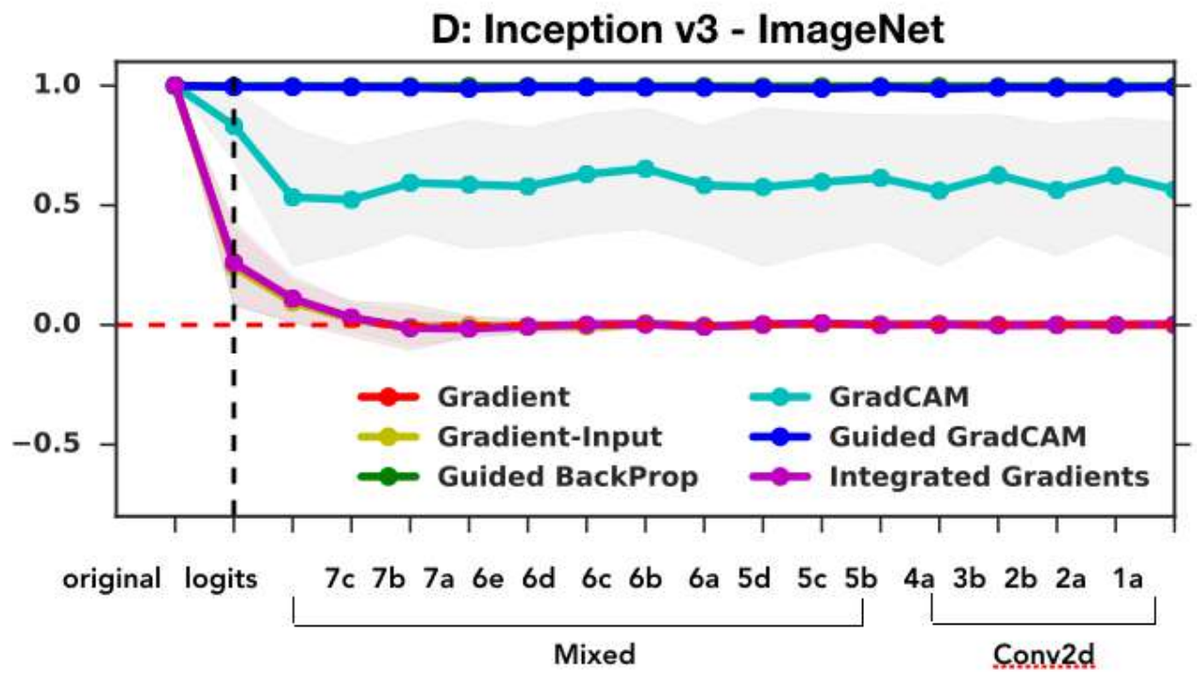
Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.



Sanity Check: Model Randomization

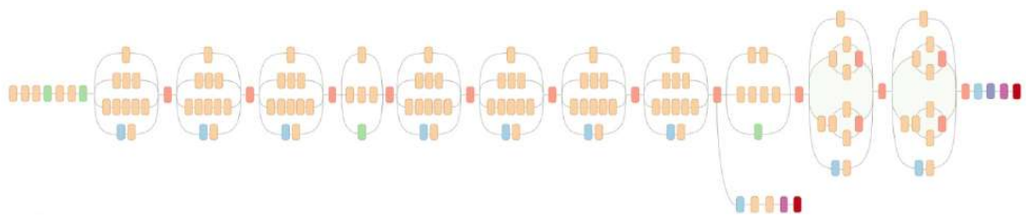
Conjecture: If a model captures higher level class concepts, then saliency maps should change as the model is being randomized.





Medical Setting

Skeletal Radiograph



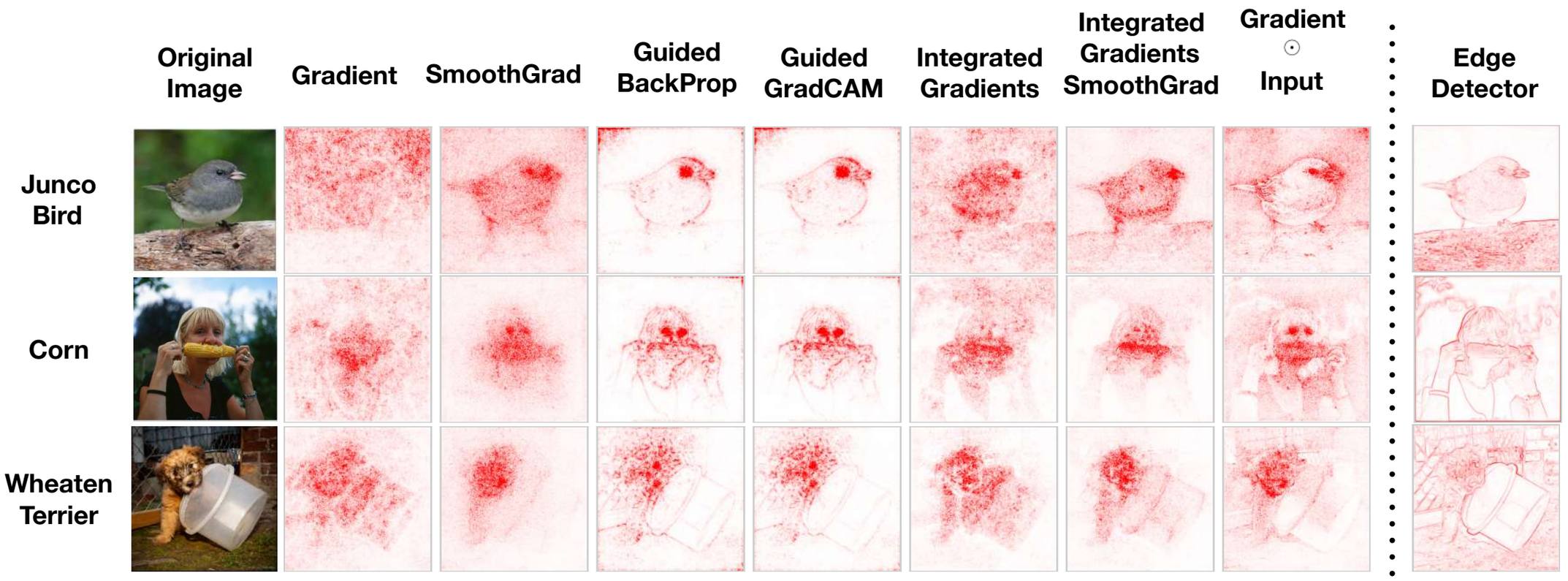
Age

Guided Backpropagation

Analysis & Visual Assessment

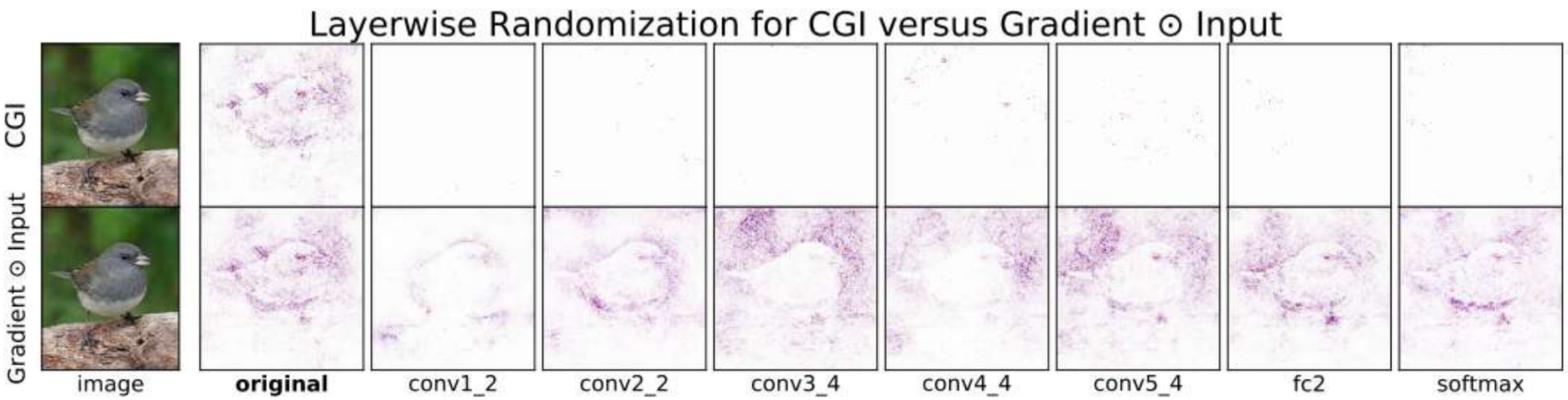
- **Nie et. al.** theoretically analyze gradient, guided backpropagation (GBP), and deconvnet (DCN) on 1-hidden layer random CNN [Nie+ ICML 2018].

- In the limit (conv filters), gradient returns iid Gaussian noise, while GBP and DCN (w/pooling layer) seek to reconstruct the input.



Fix

- **Gupta et. al.** fix this with competition for gradients (CGI).



[Figure from Gupta et. al. 2019.]

Sanity Checks Useful?

- **Sanity Checks are useful for ruling out methods, not selecting them.**

- Some other recent work that aim to assess:

    - Hooker et. al. propose to remove and retrain.

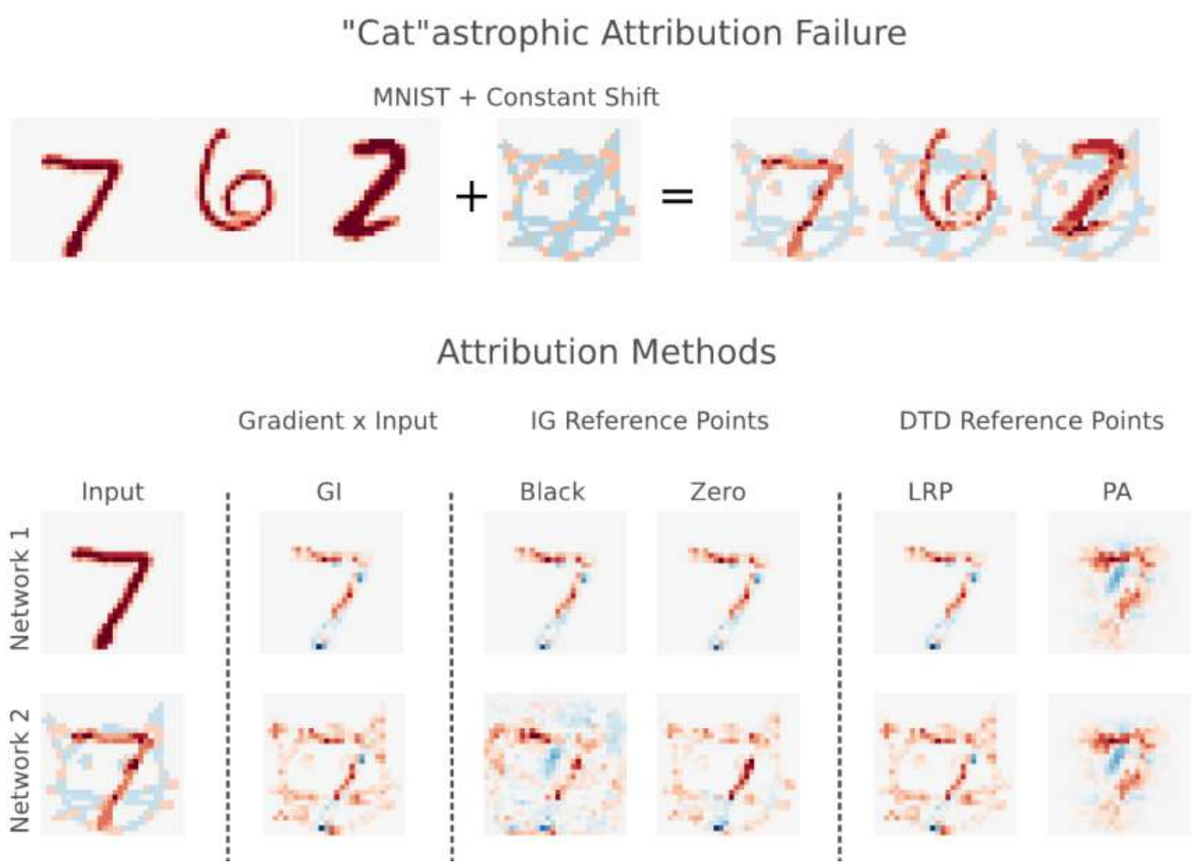
    - Adel et. al. propose FSM which ‘quantifies’ information content of a map.

    - Yang et. al. introduce a benchmark (w/ground truth and other metrics to assess how well a map captures model behavior.

- Interactions with end-users [Collaris et. al.]

Attacks

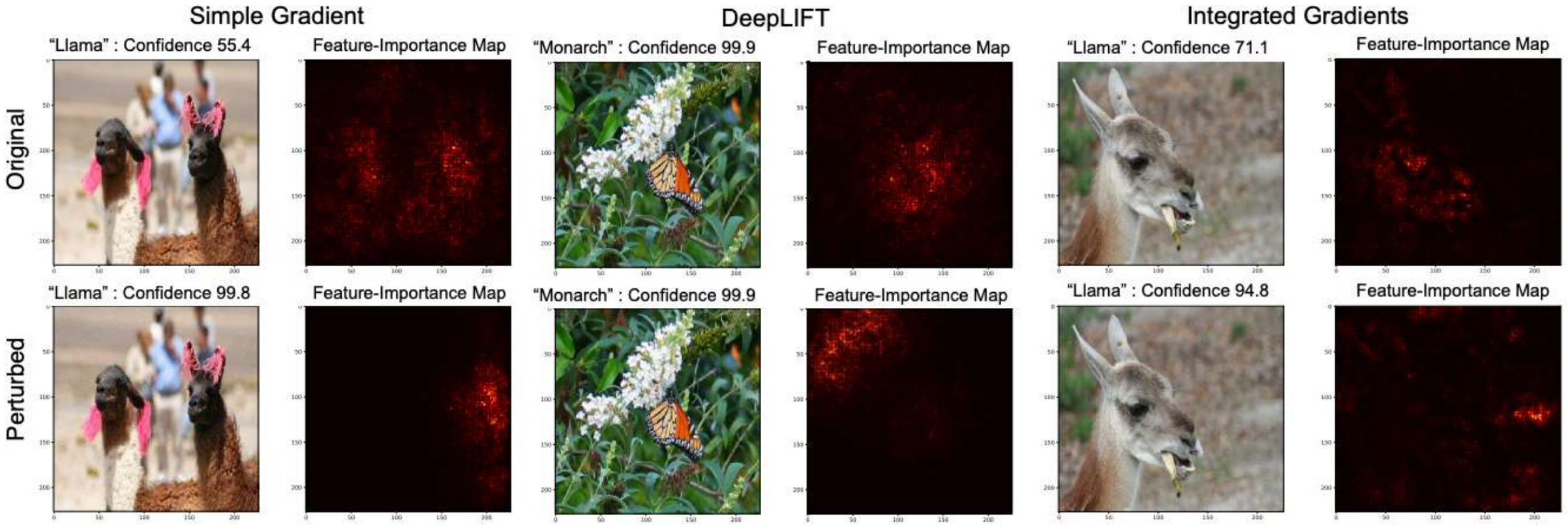
- Mean-shift attack by Kindermans & Hooker et. al.



$$E_{\mathrm{IG}}(x) = (x - \bar{x}) \times \int_0^1 \frac{\partial S(\bar{x} + \alpha(x - \bar{x})}{\partial x} d\alpha$$

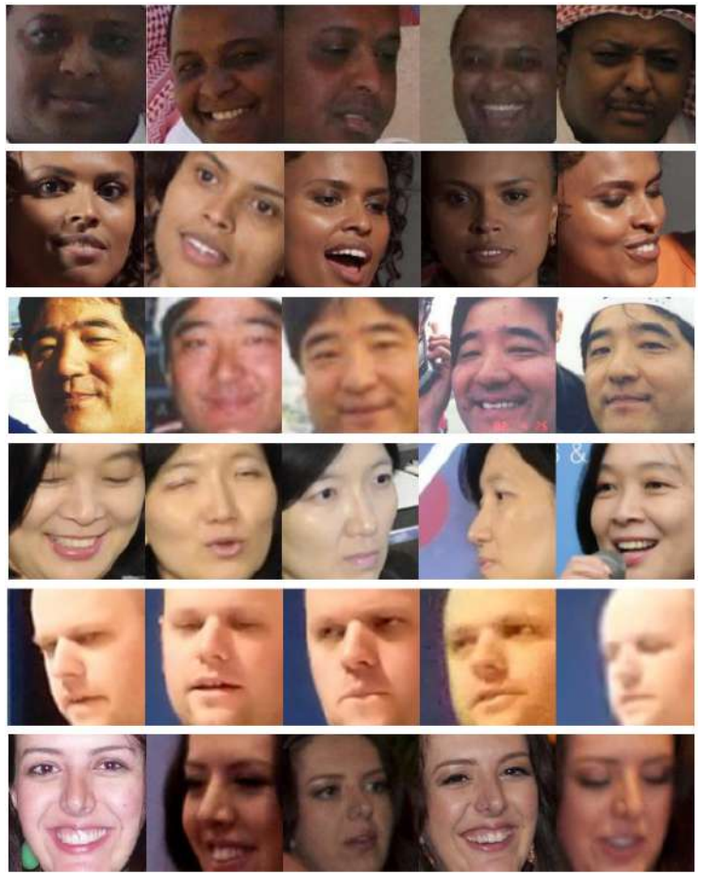
Attacks

- ‘Adversarial’ attack on explanations by Ghorbani et. al (2017).

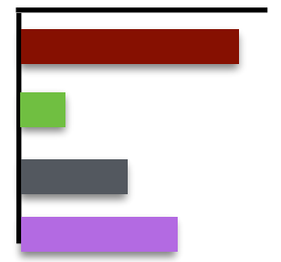


Model Debugging Upshot

[Morales+ 2019]



Predictions



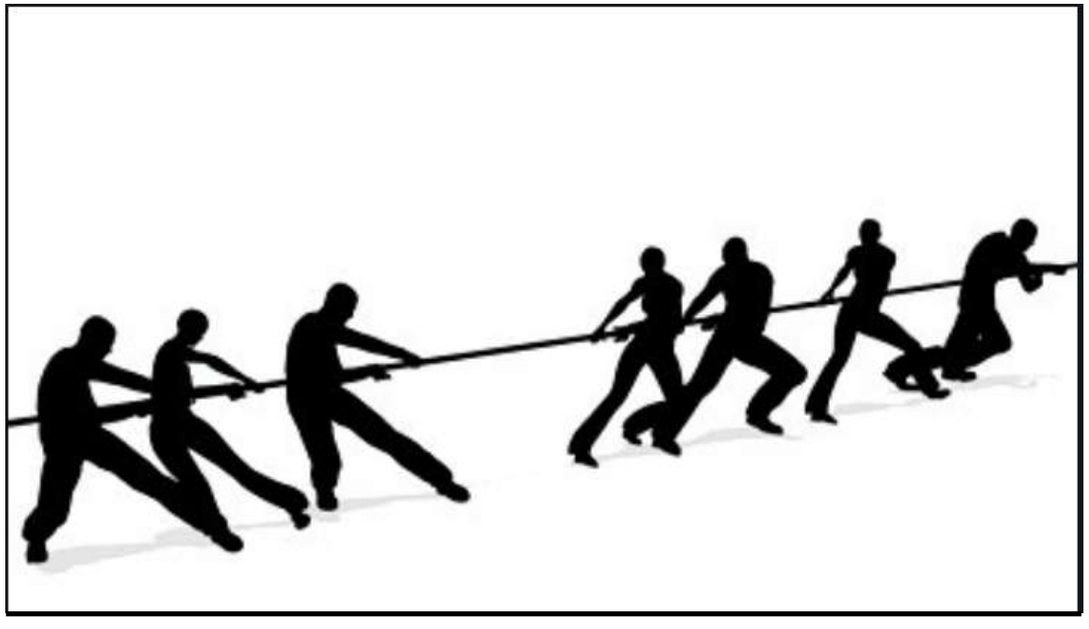
- Reveal subsets (**non-intuitive**) of the data for which the model has bad performance. This can be due to data labeling errors or others.

- Difficulty might not be in finding these subgroup in data and not interpreting them.

Key Takeaways

- Difficult to assess quality and model fidelity of local explanations.
- **Conjecture: local explanations seem to require significant privacy tradeoffs.**

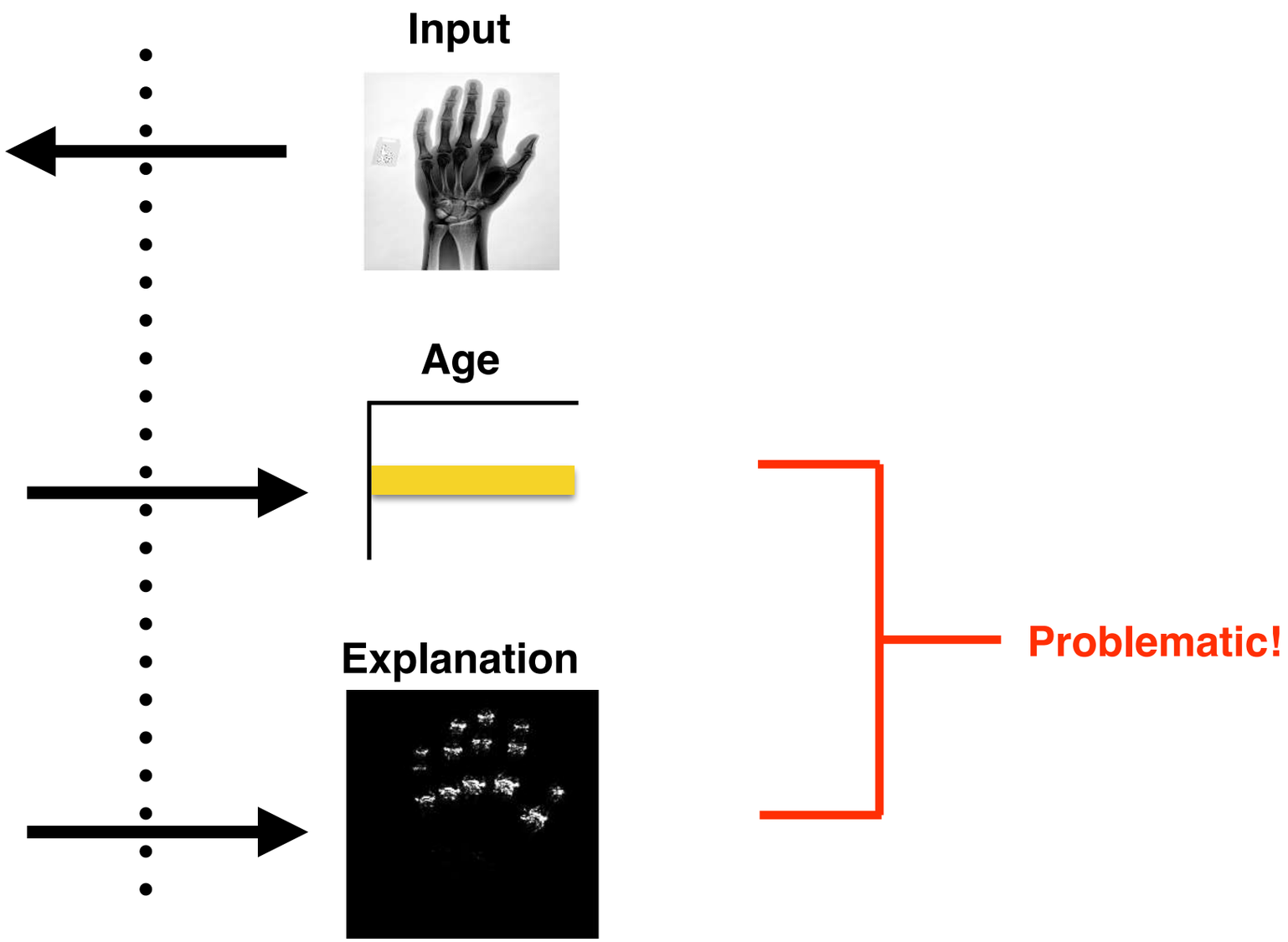
Local Explanations



Model & Data Privacy

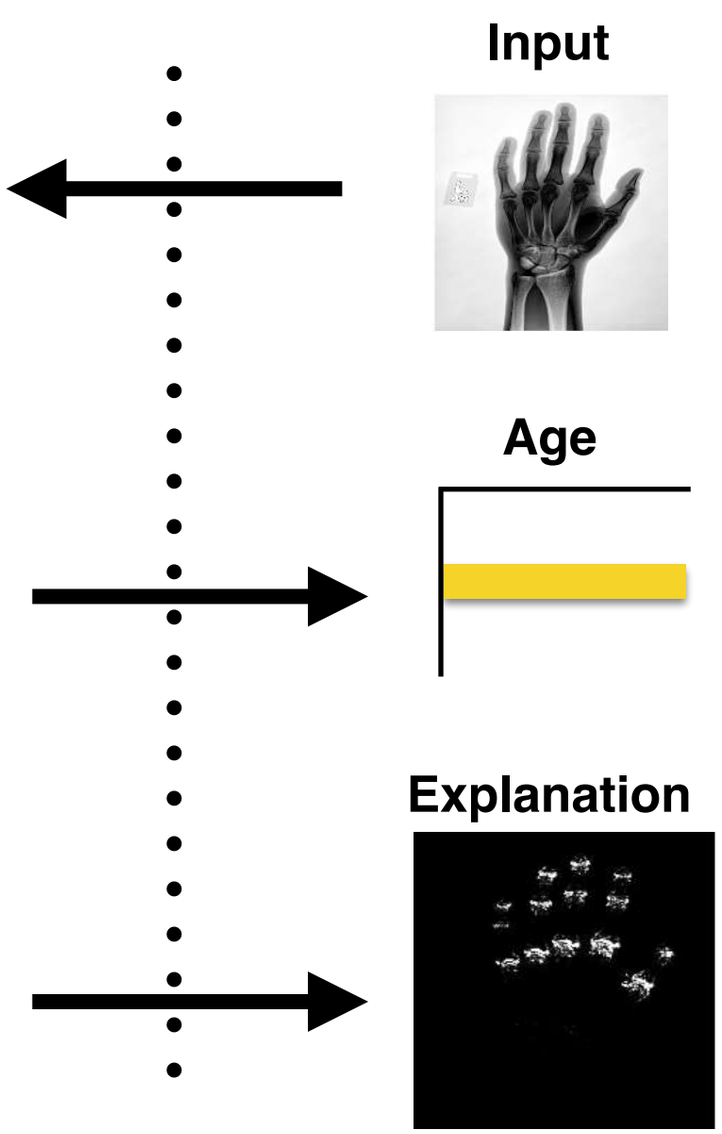
Motivation

Predictive Model



Motivation

Predictive Model



Fundamental Law of Information Recovery

"Overly accurate answers to too many questions will destroy privacy in a spectacular way."

Dwork & Roth 2014.

Model Recovery

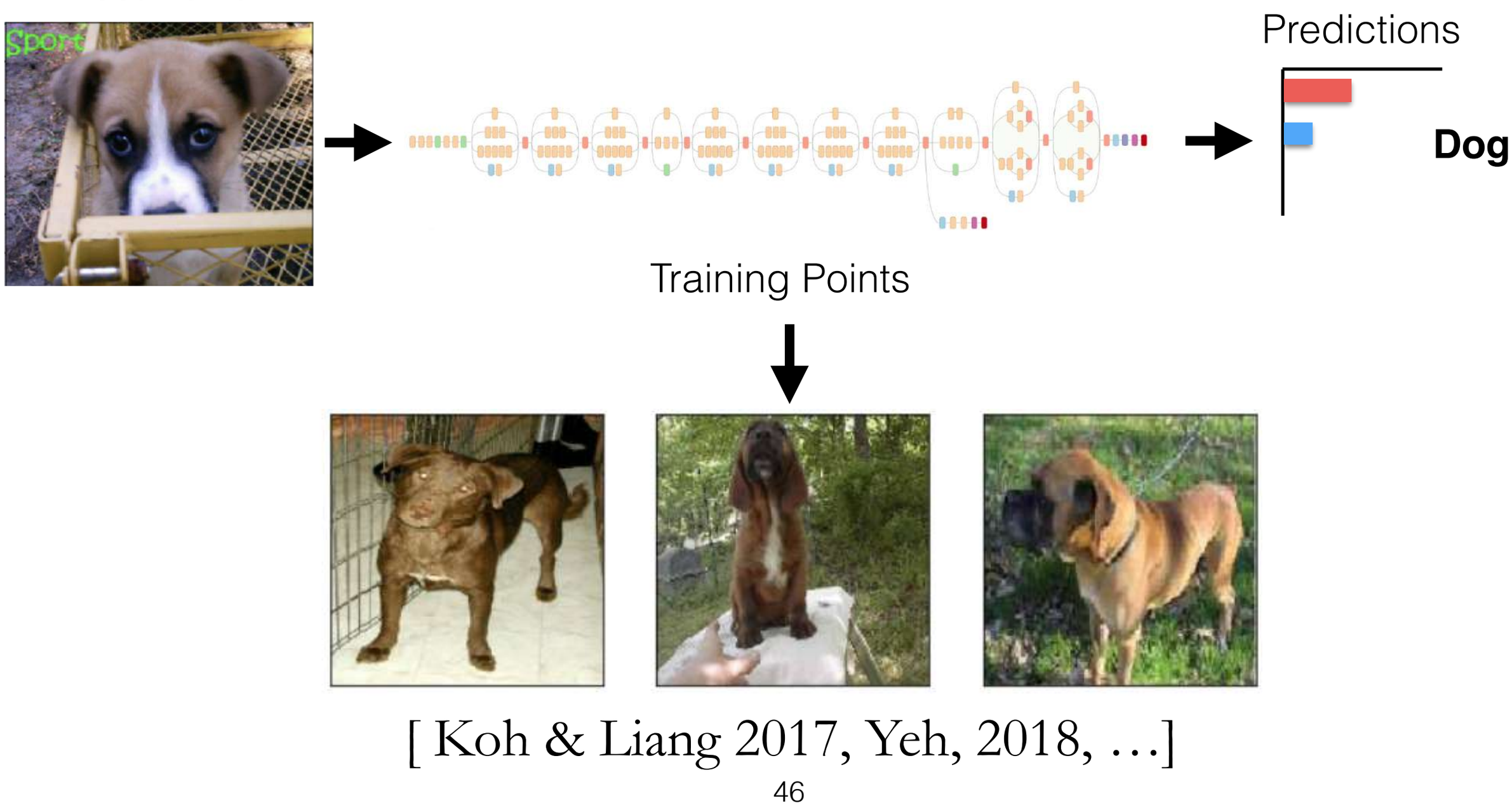
- **Tramer et. al. 2017** recover models through prediction APIs.
- **Milli et. al. 2019** should that one can recover models, (even misspecified ones) with access to local examples.

Theorem 1 (informal). *Assuming the rows of the weight matrix A are linearly independent, our algorithm recovers a functionally equivalent model from $O(h \log h)$ input gradient queries and function evaluations with high probability.*

- **Membership Inference Attacks [Shokri et. al. 2019].**

Privacy 'Harms': Examples & Prototypes

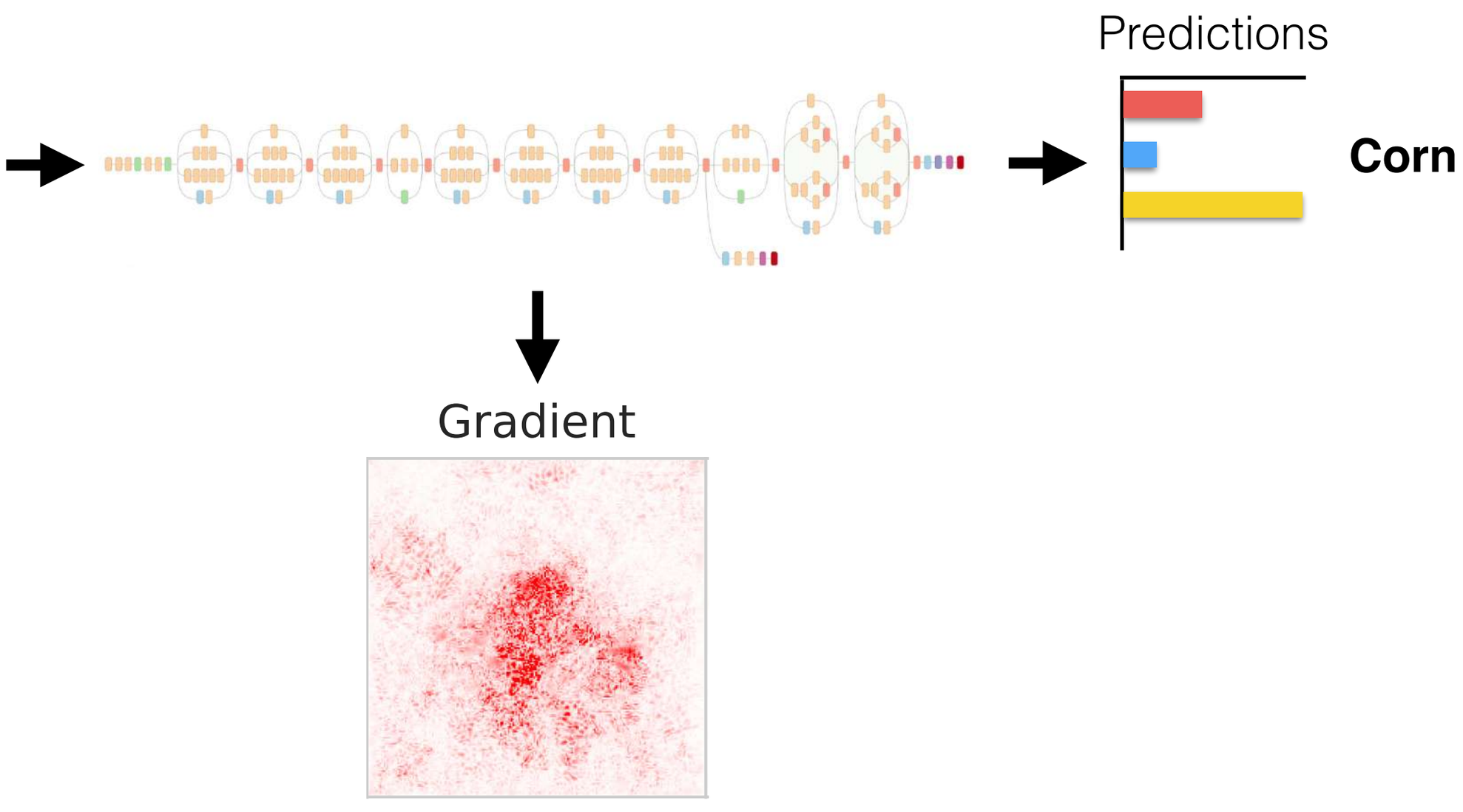
- Membership inference attacks are easier for these, and dataset reconstruction is easier with diverse point selection.



[Koh & Liang 2017, Yeh, 2018, …]

Privacy 'Harms': Maps

- Shokri et. al. also show membership inference attack possible with model learned on the norms of the local explanations.



[SVZ'13]

Can Differential Privacy (DP) Help?

- One is revealing information about exactly the inputs we would like to protect.

Definition 1. A randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta.$$

DP-SGD

- **Abadi et. al. (2016)** introduced a differentially private version of SGD as well as a moments accountant procedure to track the privacy budget.

Algorithm 1 Differentially private SGD (Outline)

Input: Examples $\{x_1, \ldots, x_N\}$, loss function $\mathcal{L}(\theta) = \frac{1}{N}\sum_i \mathcal{L}(\theta, x_i)$. Parameters: learning rate η_t, noise scale σ, group size L, gradient norm bound C.

Initialize θ_0 randomly

for $t \in [T]$ **do**

Take a random sample L_t with sampling probability L/N

Compute gradient

For each $i \in L_t$, compute $\mathbf{g}_t(x_i) \leftarrow \nabla_{\theta_t}\mathcal{L}(\theta_t, x_i)$

Clip gradient

$\bar{\mathbf{g}}_t(x_i) \leftarrow \mathbf{g}_t(x_i) / \max\left(1, \frac{\|\mathbf{g}_t(x_i)\|_2}{C}\right)$

Add noise

$\tilde{\mathbf{g}}_t \leftarrow \frac{1}{L}\left(\sum_i \bar{\mathbf{g}}_t(x_i) + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I})\right)$

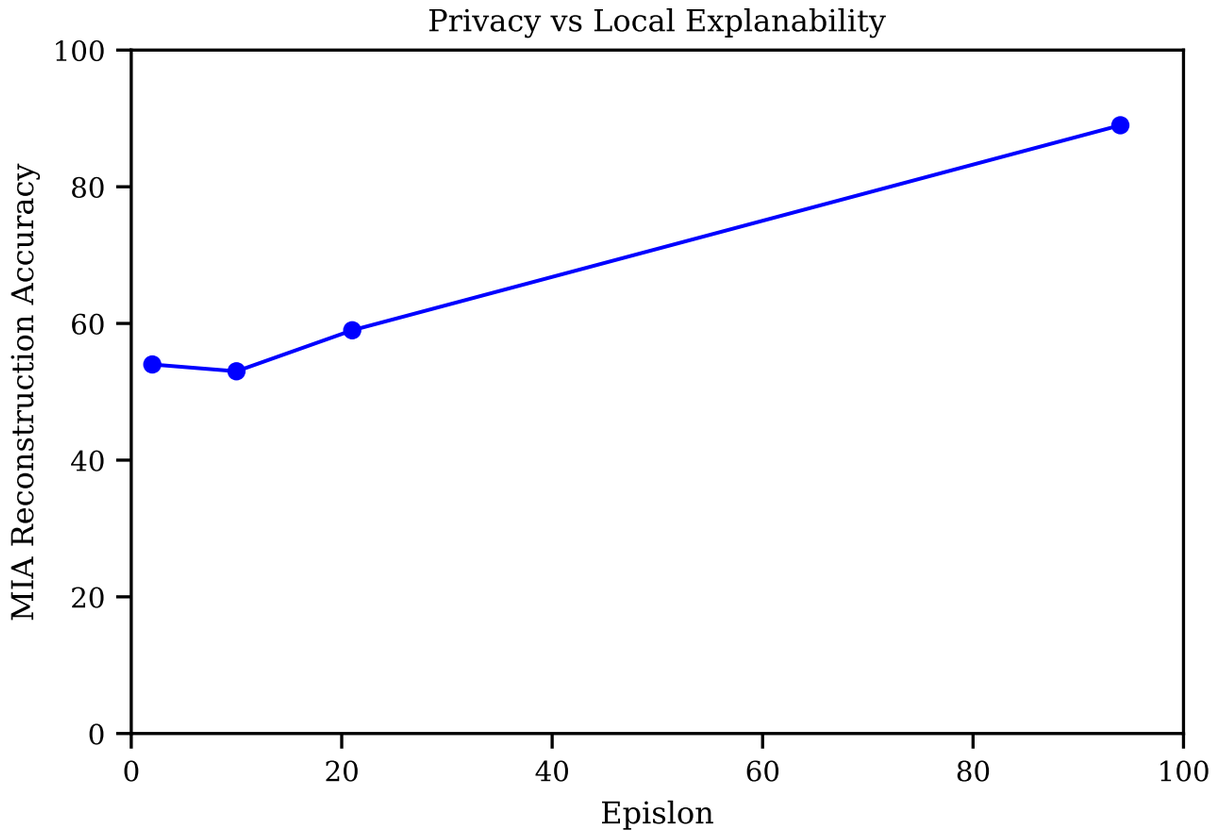
Descent

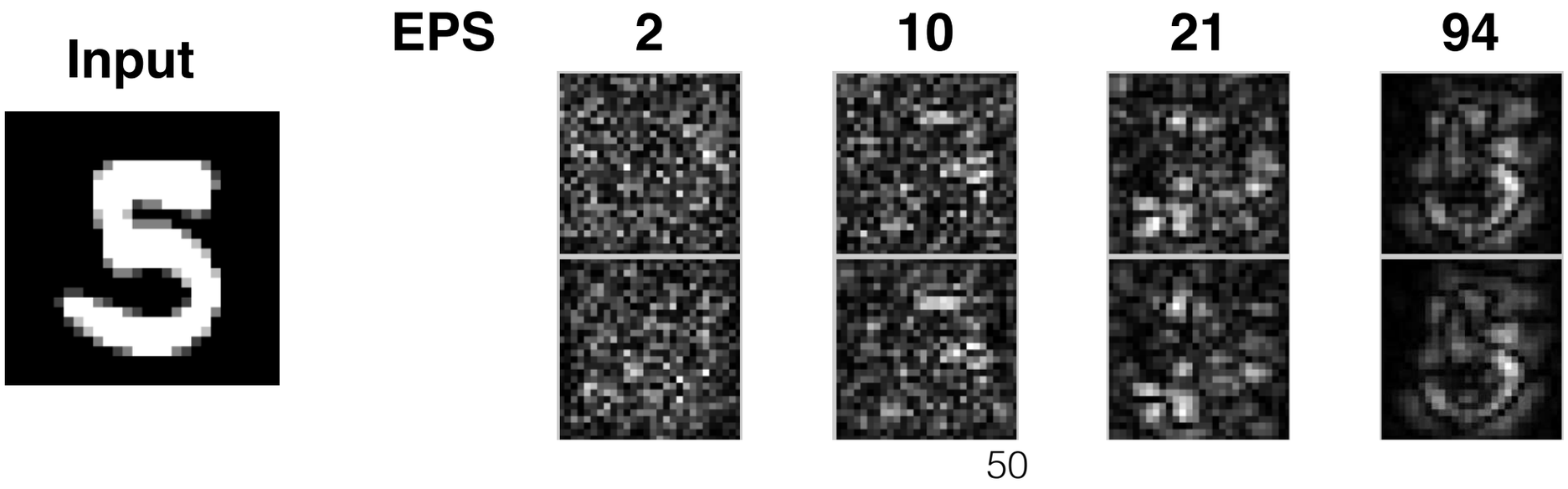
$\theta_{t+1} \leftarrow \theta_t - \eta_t \tilde{\mathbf{g}}_t$

Output θ_T and compute the overall privacy cost (ε, δ) using a privacy accounting method.

Local Explanations & DP?

- **Can DP-trained models help alleviate these concerns?**

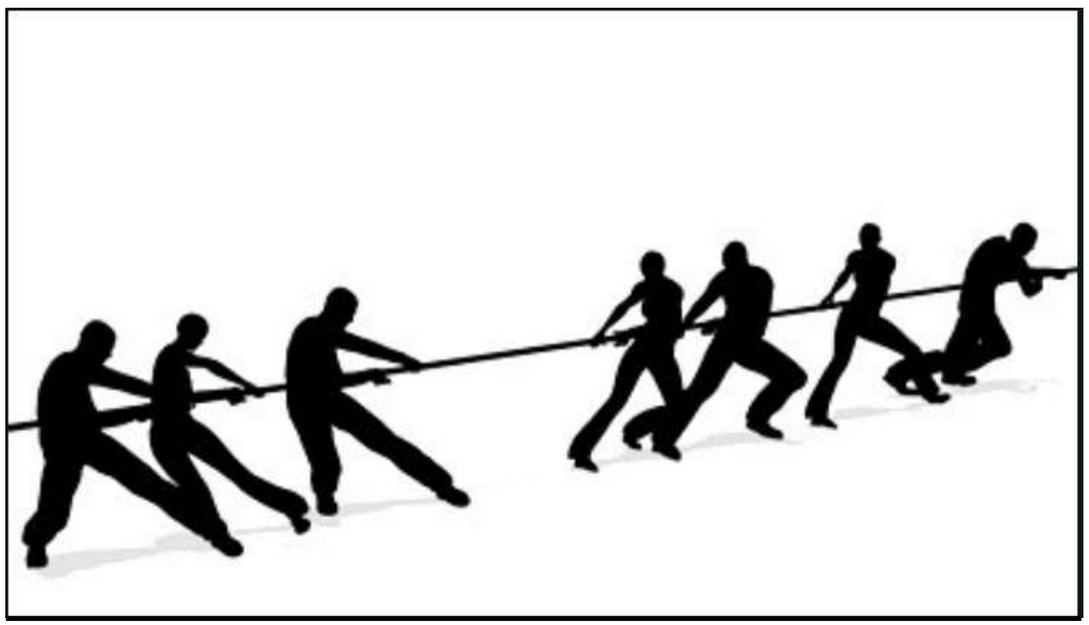




Conclusions

- Difficult to assess quality and model fidelity of local explanations.
- Conjecture: local explanations seem to require significant privacy tradeoffs.
- Perhaps global explanations can help, since it fits the theme of differential privacy?

Local Explanations



Model & Data Privacy